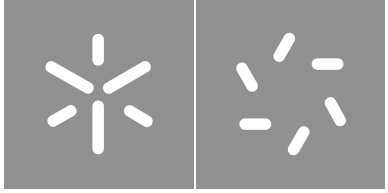
Universidade do Minho
Escola de Ciências

José Alberto Silva Castro

Modelos para a Quantificação e Valorização de Segmentos de Mercados

Outubro de 2023

Modelos para a Quantificação e Valorização de Segmentos de Mercados
José Castro
UMinho | 2023



Universidade do Minho
Escola de Ciências

José Alberto Silva Castro

Modelos para a Quantificação e Valorização de Segmentos de Mercados

Dissertação de Mestrado
Mestrado em Estatística para Ciência de Dados

Trabalho efetuado sob a orientação de:
Professora Doutora Marta Ferreira
Arquiteto Pedro Araújo

Outubro de 2023

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Agradecimentos

Gostaria de aproveitar esta oportunidade para expressar a minha sincera gratidão a todas as pessoas que contribuíram para o sucesso da minha jornada académica e da elaboração desta tese.

Um agradecimento à Jofebar, pela oportunidade da realização deste projeto pela total disponibilidade de acesso aos dados e ferramentas, fazendo com que fosse possível dar o meu contributo.

À professora Marta Ferreira, pela excelente orientação e apoio constante, partilhando os seus conhecimentos, que tiveram um papel fundamental para a realização deste trabalho.

Ao meu orientador, Pedro Araújo, o meu muito obrigado por toda orientação, sabedoria e paciência durante este processo.

Aos meus colegas de trabalho que me fizeram sentir em casa, onde a amizade e a colaboração tornaram os desafios do dia a dia mais leves e as conquistas mais significativas.

À minha namorada, obrigado por estares ao meu lado, pelo amor e apoio incondicional em cada etapa deste caminho. A sua presença tornou tudo mais significativo.

A todos os meus amigos, que me acompanharam durante este percurso e por tudo o que fizeram por mim.

Por último, queria agradecer à minha família, em especial aos meus pais e ao meu irmão, por todo o esforço que fizeram para que pudesse cumprir este sonho, por todo o apoio e paciência, estando sempre presentes.

Obrigado a todos pela oportunidade de crescer e aprender ao longo deste percurso!

José Castro

"A gratidão é a virtude das almas nobres."

Esopo

Resumo

Com o passar do tempo e com o avanço das tecnologias, novas pesquisas na área da segmentação de mercado foram realizadas. O uso de *Machine Learning* tornou-se, atualmente, uma ferramenta importante para a resolução desses problemas.

Esta dissertação tem como objetivo a criação de modelos de quantificação e valorização de segmentos de mercado em Portugal, Espanha e Estados Unidos, recorrendo a métodos de *Machine Learning* para desenvolver modelos analíticos que proporcionem informações cruciais para compreender e prever o comportamento de mercado nesses três países.

Estes métodos utilizados baseiam-se em árvores de regressão, uma vez que a variável de interesse é uma variável contínua que diz respeito à faturação da empresa em cada localidade. Foram escolhidos os métodos *boosting*, *bagging* e *random forest* para a previsão de faturação.

Todo o trabalho realizado no âmbito da dissertação foi realizado com o *software R* e representado graficamente com a ajuda do *software Power BI*, ferramenta útil para a criação de *Dashboards*. O tratamento dos dados foi também trabalhado com o recurso ao *Excel*.

Os resultados da previsão obtidos dos métodos de *Machine Learning* permitem à empresa ter um maior conhecimento sobre quais as localidades mais importantes e que variáveis influenciam diretamente a faturação. Estes resultados serão aplicados na empresa e irão sustentar a tomada de decisão.

Palavras-Chave: *Machine Learning*; *boosting*; *bagging*; *random forest*;

Abstract

Over time with the advancement of technologies, new research in the field of market segmentation has been conducted. The use of Machine Learning has currently become an important tool for solving these problems.

This dissertation aims to create models for quantifying and valuing market segments in Portugal, Spain, and the United States. We use Machine Learning methods to develop analytical models which provide crucial information to understand and predict market behavior in these three countries.

The methods used are based on regression trees. The variable of interest is a continuous variable related to the company's revenue in each location. The "boosting," "bagging," and "random forest" methods were chosen for revenue forecasting.

All the work carried out in the context of the dissertation was done using the R software and graphically represented with the help of the Power BI software, a useful tool for creating dashboards. Data processing was also done using Excel.

The forecasting results obtained from Machine Learning methods enable the company to gain a better understanding of the most important locations and the variables that directly influence revenue. These results will be applied in the company and will support decision-making.

Keywords: *Machine Learning*; *boosting*; *bagging*; *random forest*;

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

GAM	*Generalized Additive Model*
SQT	*Soma dos Quadrados Totais*
SQE	*Soma dos Quadrados dos Erros*
SQR	*Soma dos Quadrados da Regressão*
MSE	*Mean Squared Error*
SVM	*Support Vector Machine*
INE	*Instituto Nacional de Estatística*
GPEARI	*Gabinete de Planeamento, Estratégia, Avaliação e Relações Internacionais*
VIF	*Variance Inflation Factor*
RMSE	*Root Mean Square Error*
OOB	*Out Of Bag*

Capítulo 1

Introdução

1.1 Enquadramento Teórico

No contexto empresarial contemporâneo, a aplicação da Ciência de Dados e, mais especificamente, dos modelos de *Machine Learning*, emerge como um catalisador significativo no aprimoramento da tomada de decisões estratégicas. A era do *Machine Learning* trouxe consigo um conjunto de ferramentas e abordagens que permitem às organizações explorar *insights* valiosos a partir de vastas quantidades de dados, gerando impacto substancial numa ampla variedade de setores. Este trabalho tem como objetivo central investigar a aplicabilidade e os benefícios desses avanços tecnológicos num contexto específico, cujo mercado alvo é restrito e singular.

A Jofebar opera num nicho de mercado altamente específico, caracterizado por peculiaridades únicas que demandam uma abordagem cuidadosamente adaptada. A comercialização de janelas, o produto central da organização, está intrinsecamente relacionada com a procura variável de segmentos de trabalho distintos. A diversidade destes segmentos, cada um com as suas próprias características e exigências, representa um desafio complexo em termos de gestão de recursos e estratégia de negócios.

Neste contexto, a presente dissertação assume um papel crucial ao propor o desenvolvimento de um modelo de *Machine Learning* destinado à quantificação e valorização desses segmentos de trabalho. O intuito é capacitar a empresa a tomar decisões mais informadas e eficientes na alocação de recursos, na otimização de processos e na personalização da oferta de produtos.

A ascensão do *Machine Learning*, que se materializa na aplicação de algoritmos para analisar e extrair *insights* de dados brutos, representa um marco paradigmático nas abordagens tradicionais de resolução de problemas empresariais. A capacidade de identificar padrões, prever tendências e tomar decisões baseadas em dados com precisão nunca antes vista está a redefinir a forma como as empresas operam e prosperam. Neste contexto, este estudo visa demonstrar como a adoção dessa tecnologia pode traduzir-se em vantagens competitivas tangíveis para uma empresa que enfrenta os desafios de um mercado especializado, como o das janelas.

A dissertação está estruturada em conformidade com uma metodologia rigorosa, que abrange desde a coleta e tratamento de dados até à implementação e validação do modelo proposto. Cada etapa é delineada com o intuito de garantir a solidez das conclusões e a aplicabilidade prática das descobertas.

Por conseguinte, o presente trabalho visa contribuir não apenas para o avanço do conhecimento teórico no campo da Ciência de Dados e *Machine Learning*, mas também para a implementação prática de soluções inteligentes num contexto empresarial altamente específico e desafiador. A relevância deste estudo reside na capacidade de gerar impacto direto na eficácia operacional e na competitividade da empresa em questão, abrindo portas para novas abordagens estratégicas e inovadoras.

Nos capítulos subsequentes, serão explorados em detalhe os fundamentos teóricos do *Machine Learning*, a metodologia de pesquisa adotada, os resultados obtidos e as suas implicações práticas. No final, espera-se consolidar a compreensão do papel transformador que a era do *Machine Learning* desempenha nas organizações, tornando-se uma ferramenta poderosa para a quantificação e valorização de segmentos de trabalho num contexto empresarial peculiar.

1.2 Objetivos

A presente dissertação tem como objetivo central a criação de um modelo que visa a quantificação e valorização de segmentos de trabalho. Para alcançar com êxito este objetivo, é imperativo desdobrá-lo em metas específicas e definir uma sequência de etapas meticulosamente elaboradas, as quais compreendem as seguintes tarefas:

- Realizar uma análise minuciosa dos mercados territoriais, a fim de estimar a sua dimensão hipotética e o seu potencial relativo para um nicho de especialização particular;
- Construir um modelo preditivo, cujo propósito é calcular o fluxo de receitas potenciais, tendo como base variáveis previamente identificadas com uma correlação substancial e empiricamente comprovada com o nicho de mercado em foco.

1.3 Local de Estágio

A presente dissertação foi desenvolvida durante a realização de um estágio na Jofebar, no âmbito do Mestrado em Estatística para Ciência de Dados, da Escola de Ciências, da Universidade do Minho.

A Jofebar é uma companhia portuguesa estabelecida em 1986 cuja qualidade dos produtos e serviços na área de fachadas, estruturas e caixilharias conduziu a um exponencial crescimento e a uma crescente implementação no mercado de trabalho. É, desde 2004, uma fabricante especializada de sistemas de janelas minimalistas e responsável, em 2008, pela criação da marca suíça *panoramah!®*. A partir de 2011 inicia a verticalização da sua atividade, no contexto do grupo de empresas de que faz parte, passando a integrar a transformação de vidro e ajustando o seu modelo de negócio para se colocar como empresa de referência, no seu nicho de mercado, a nível mundial.

Com uma equipa excecionalmente motivada e experiente, a Jofebar tem tido um crescimento sustentável e encontra-se presente, diretamente ou através da sua rede de parceiros especializados, em mais de 50 países.

1.4 Estrutura do Documento

Esta dissertação encontra-se dividida em seis capítulos.

No capítulo 1 é introduzido o tema do trabalho, os objetivos, os *softwares* utilizados e a instituição onde foi realizado o estudo.

O capítulo 2 é relativo ao estado de arte. Neste capítulo é elaborado um mapeamento com parte da produção académica sobre a segmentação de mercado mais tradicional e a segmentação de mercado com recurso ao *Machine Learning*.

No capítulo 3 é apresentada a base de dados utilizada e apresenta-se uma análise descritiva dos dados.

A descrição das metodologias é apresentada no capítulo 4. Serão abordadas todas as metodologias utilizadas para elaboração deste projeto.

No capítulo 5 serão apresentados os resultados obtidos.

Por fim, e depois de uma análise global do estudo, surge, no capítulo 6, uma síntese das conclusões retiradas, assim como sugestões de trabalho futuro.

1.5 Software Utilizado

Ao longo do estudo foram utilizados três *softwares*:

- ***Microsoft Office Excel*** - ferramenta de suporte para a criação das bases de dados utilizadas;

- ***Power BI*** - ferramenta de apoio na análise exploratória dos dados e na criação de uma *Dashboard*;
- ***R*** - ferramenta de apoio na análise exploratória dos dados e na validação dos resultados das metodologias [1];

Capítulo 2

Estado de Arte

A presente dissertação visa aprofundar o conhecimento sobre segmentação de mercado e as suas aplicações práticas no mundo empresarial, com a incorporação de técnicas de *Machine Learning*. Neste capítulo são apresentados alguns conceitos-chave, tais como, critérios tradicionais de segmentação, incluindo demográficos, geográficos, psicográficos e comportamentais, e a utilização de algoritmos de *Machine Learning* para identificar e caracterizar segmentos de forma mais precisa e eficiente, combatendo as limitações dos critérios tradicionais.

2.1 Segmentação de Mercado

O conceito de segmentação de mercado aparece em 1956 com o trabalho pioneiro de *Wendell R. Smith*, mas só nos anos 70 é que se torna uma das questões principais no *marketing*. De modo a entender a importância da segmentação é necessário perceber no que consiste. A segmentação de mercado envolve a visualização de um mercado heterogéneo como vários mercados menores e homogéneos, em resposta a preferências diferentes, atribuíveis aos desejos dos clientes por satisfações mais precisas das diversas necessidades [2]. Esta é uma definição que até aos dias de hoje se vem a mostrar muito precisa.

Atualmente, as empresas têm recorrido com maior frequência à segmentação de mercado, uma vez que desempenha um papel fundamental na estratégia de *marketing* com vista em ajudar nas tomadas de decisões [3]. É essencial e indispensável para o funcionamento e objetivos das grandes empresas o recurso à segmentação do

mercado, sendo isso o principal condutor do negócio [4]. Compreender os diferentes segmentos de mercado e identificar as suas características distintas, com nichos de mercados muito específicos, é essencial para direcionar os esforços do *marketing*, melhorar na inovação de produtos e serviços personalizados e alcançar uma vantagem competitiva entre empresas [5].

Apesar de tudo, há sempre algumas questões importantes de se colocar: é uma mais valia para as empresas? Se sim, porque é que existe erro, já que, por vezes, a segmentação falha? E se falha, o que é que pode ser feito para reduzir esse erro?

2.1.1 A segmentação é uma mais valia?

Com o objetivo de responder e elucidar as questões apresentadas anteriormente, rememora-se a definição de segmentação de mercado como ponto de partida. A noção de segmentação de mercado é amplamente estabelecida e reconhecida como um conceito fundamental. Parte-se do pressuposto que os clientes demonstram uma heterogeneidade nas preferências e nos comportamentos de compra [6], sendo que essa variabilidade é geralmente explicada pelos diferentes produtos e características [7]. Para uma empresa, torna-se impensável e é simplesmente irrealista satisfazer todas as necessidades dos clientes no mercado. Assim, para lidar com essa heterogeneidade, recorre-se à segmentação de mercado para equilibrar a variabilidade das necessidades dos clientes com as limitações dos recursos disponíveis. Para a empresa haverá uma vantagem competitiva, até que os principais concorrentes diretos copiem ou segmentem o seu mercado. [8].

A Segmentação, Mercado Alvo e Posicionamento (*Segmentation, Targeting and Positioning - STP*) são conceitos amplamente utilizados neste campo. É uma técnica estratégica de *marketing* para o mercado, onde esta auxilia a compreender o quão eficientemente a empresa está a planear diversas atividades de *marketing* para competir nesse mercado e a analisar a forma como se relacionam com o mercado em geral [9].

De acordo com o modelo representado na figura 2.1, podemos ver as três etapas relatadas anteriormente.

A etapa de segmentação inicia-se pelo processo de seleção das variáveis mais relevantes que possam identificar e diferenciar os diversos grupos de consumidores. Nesse sentido, as empresas devem procurar compreender quais são essas variáveis-chave que melhor representam as características e necessidades dos diferentes segmentos. Uma seleção criteriosa e fundamentada é essencial para garantir uma segmentação eficaz [10].

Uma vez definidas as variáveis, o próximo passo é construir um perfil detalhado de cada segmento. Isso envolve analisar e compreender os comportamentos, preferências, motivações e características demográficas dos consumidores em cada segmento. Essa compreensão aprofundada é fundamental para adequar as estratégias

Figura 2.1: Segmentação, Mercado Alvo e Posicionamento

de *marketing* às necessidades e desejos específicos de cada grupo de consumidores [11].

É importante ressaltar que a segmentação de mercado não é um processo estático, mas sim dinâmico [12]. À medida que o mercado evolui e os comportamentos dos consumidores mudam, as empresas precisam de se adaptar e ajustar as suas estratégias de segmentação. A análise contínua do mercado, a monitorização das tendências e o *feedback* dos consumidores são essenciais para garantir uma segmentação atualizada e relevante ao longo do tempo.

Desta forma, a segmentação de mercado é um processo contínuo que requer uma constante adaptação por parte das empresas para atender às mudanças do mercado. A capacidade de se manter atualizado e ágil na segmentação é um fator-chave para o sucesso no mercado altamente competitivo de hoje [11].

A segmentação pode ser dividida em quatro pontos importantes [13]:

- Segmentação Geográfica;
- Segmentação Demográfica;
- Segmentação Comportamental;
- Segmentação Psicográfica;

Passando para o *targeting*, conforme o próprio nome sugere, ocorre a seleção do público-alvo. Nessa etapa, é fundamental optar pela estratégia de *marketing* a ser empregada. Torna-se crucial identificar os grupos de consumidores e direcionar toda a estratégia de segmentação a esses clientes específicos. É importante ressaltar que esses potenciais clientes podem apresentar diferenças significativas em diversos aspetos, o que enfatiza a necessidade de definir com precisão cada segmento, a fim de atender plenamente às necessidades de todos os clientes envolvidos [14].

A próxima etapa crucial é o posicionamento, um processo complexo que visa compreender as perceções dos consumidores em relação ao produto ou serviço e estabelecer uma posição adequada no mercado. Torna-se imperativo adentrar profundamente na mente dos consumidores, a fim de discernir a maneira mais eficaz de se destacar perante a concorrência, apresentando uma proposta única e cativante. É por meio da diferenciação, ressaltando características exclusivas e benefícios superiores em comparação aos concorrentes, ou por meio de uma liderança em termos de custo, oferecendo produtos de excelência a preços mais acessíveis, que é possível sobressair diante dos concorrentes mais diretos [11].

Fundamentalmente, este processo traz-nos dois problemas: as empresas acreditam que estão a segmentar o mercado quando, inequivocamente, não o estão; e a segmentação que estão a usar não trará resultados previstos [3].

O primeiro problema surge, em parte, pela imprecisão no uso da linguagem de segmentação. É muito associado à segmentação o agrupamento de clientes, então surge aí o problema, muitas vezes são mal agrupados, isto é, clientes com necessidades diferentes agrupados no mesmo grupo. Para se tornar mais fácil de perceber, um exemplo disso é quando nos bancos decidem dividir os clientes pelo critério "volume/tamanho de negócio"[15]. Esta abordagem, baseada exclusivamente em critérios financeiros, pode negligenciar outros aspetos relevantes, como as necessidades específicas de cada cliente, o setor em que atuam, as preferências em relação aos serviços bancários e os comportamentos de compra. Assim, essa segmentação que se baseia unicamente em critérios de tamanho ou volume de negócios pode não ser eficaz na identificação de grupos de clientes com necessidades semelhantes. No entanto, nos últimos anos tem havido progressos consideráveis, uma vez que os bancos têm procurado desenvolver as suas bases de dados de forma a permitir a implementação de segmentos com base nas necessidades dos clientes [16].

O segundo problema prende-se com o facto de que os profissionais de *marketing* que seguem a segmentação prescrita, por vezes, não conseguem gerar uma solução de segmentação utilizável. Neste sentido, a aparente simplicidade do processo STP em três etapas esconde algumas das dificuldades subjacentes. Por exemplo, existem situações em que as empresas utilizam a análise de *clusters*, uma abordagem amplamente reconhecida [17] para segmentar o mercado, agrupando clientes com necessidades idênticas. No entanto, ocorre que nem sempre isso produz resultados satisfatórios, uma vez que os grupos resultantes podem não apresentar necessidades idênticas entre si.

2.1.2 O porquê da segmentação não ter sempre sucesso?

Conforme mencionado anteriormente, quando ocorre uma falha na segmentação, refere-se ao fato de que, ao seguir o processo de segmentação, não se consegue gerar

uma solução viável para implementação. Isso implica que não foi possível utilizar a segmentação para desenvolver uma estratégia de *marketing* adequada.

Em primeiro lugar, é essencial compreender o princípio subjacente à segmentação. Uma compreensão inadequada desse conceito pode levar a fracassos significativos [3]. Resumidamente, algumas empresas encaram a segmentação como uma maneira conveniente de dividir os seus mercados, focando-se apenas em benefícios operacionais. No entanto, essa abordagem negligencia a satisfação do cliente e a capacidade de atender às suas necessidades. Tomando novamente como exemplo o setor bancário, embora as estratégias adotadas possam resultar em ganhos operacionais, elas falham em proporcionar uma experiência satisfatória ao cliente. Consequentemente, desenvolver uma estratégia de *marketing* eficaz para atingir segmentos específicos torna-se impossível.

Um segundo problema está relacionado com a excessiva concentração nos detalhes do processo de segmentação e o esquecimento dos principais objetivos que o levaram a segmentar o mercado. Acaba-se "muitas vezes" por se dar demasiada importância na coleta e analise de dados, que se perde o foco no objetivo do estudo. Este problema surge pela dificuldade que as empresas encontram em tornar eficaz a segmentação. Na prática, o que acontece é que as empresas vão apenas focar-se na parte económica, comparando as vendas em diferentes segmentos com a contribuição financeira para a empresa, e perder todos os outros fatores igualmente importantes, tal como a satisfação do cliente. É importante perceber que análise da segmentação precisa de ter um horizonte de tempo mais longo do que apenas uma semana ou um mês [3]. Deve existir uma análise clara das necessidades e comportamentos dos clientes, examinar também os ambientes competitivos e comerciais mais amplos, e deve resultar em programas de *marketing* consistentes. Não poderá haver dúvidas sobre o que se vende e para quem se dirige a marca.

Académicos e o mundo empresarial envolvidos em projetos de segmentação podem recorrer à literatura em busca de orientação. Embora a literatura académica concentre-se extensivamente no uso de diferentes abordagens estatísticas na pesquisa de segmentação, há muito pouca orientação prática sobre os inúmeros problemas e armadilhas. Para aqueles que têm pouco conhecimento estatístico, essas armadilhas podem ter consequências graves [18]. A facilidade de acesso a *softwares* estatísticos com as diversas bibliotecas disponíveis, a análise multivariada tornou-se uma ferramenta acessível para os gestores de empresas, surgindo aí o problema. Embora o acesso a tudo seja facilitado, existe toda uma complexidade por detrás do que aparenta ser fácil. A necessidade de validar os resultados estatísticos com testes e procedimentos adequados é apenas uma das áreas que podem ser ignoradas. Supondo que esse problema é resolvido, a empresa apresenta alguém especializado para o trabalho, é fundamental a recolha dos dados. Estes devem manter-se sempre atualizados e com fontes seguras. Resumidamente, o problema da estatística é que

vemos algumas empresas a recorrerem à análise estatística sem conhecimento prévio.

2.1.3 O que pode ser feito para reduzir o erro?

Seguir um plano para todo o projeto de segmentação só traz benefícios. O planeamento encoraja a definição de objetivos claros para que se estabeleça, desde o início, o que se deseja alcançar com a segmentação. Recorrer a alguém com conhecimentos estatísticos torna-se uma mais-valia, uma vez que, para além do tratamento de dados, consegue desenvolver modelos que definam a segmentação, tendo por base *softwares* estatísticos. A obtenção dos dados é algo muito sério. É preciso questionar e considerar como e quando foram coletados, uma vez que os dados são a parte vital da segmentação. O método de obtenção dos dados deve ser apropriado e robusto, e com uma atualização constante ao longo do tempo. As análises da segmentação devem servir para poder conhecer melhor o cliente, perceber as suas necessidades e comportamentos, para que ao longo do tempo se possam tomar as decisões de *marketing* mais corretas. Planear a abordagem que se deve ter é importante, nomeadamente, pensar em três pontos cruciais: antes, durante e o pós. Uma atualização constante fará com que o projeto se fortaleça ao longo do tempo [19].

Em suma, apesar de todas as preocupações descritas, a segmentação de mercados é fundamental para o crescimento de uma empresa. É algo que nos dias de hoje tem sido cada vez mais usado, trazendo inúmeros benefícios. Contudo, com o avanço da tecnologia, surge uma nova abordagem, a segmentação com *machine learning*. Este método vem colmatar a segmentação mais tradicional, uma vez que esta é menos precisa que o uso de *machine learning*.

2.2 *Machine Learning*

Machine Learning é uma técnica de Inteligência Artificial que permite que um computador possa aprender a partir de dados, sem ter sido explicitamente programado para tal. Esta abordagem permite que o sistema possa identificar padrões e relações complexas nos dados de forma automatizada, tornando possível a construção de modelos preditivos mais precisos e eficientes [20].

"Machine learning is about extracting knowledge from data" - Yann LeCun

Uma das principais vantagens do *Machine Learning* é a capacidade de lidar com grandes volumes de dados. Com o aumento do volume de dados disponíveis, a capacidade de analisar e extrair informações relevantes desses dados torna-se cada vez mais importante. O *Machine Learning* permite que as empresas possam utilizar

esses dados para obter *insights*[1] valiosos sobre os seus clientes e melhorar a tomada de decisões [21].

2.2.1 Tipos de Aprendizagem em Machine Learning

O *Machine Learning* pode ser dividido em três categorias principais: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Cada uma dessas técnicas possui as suas próprias características e aplicações [22].

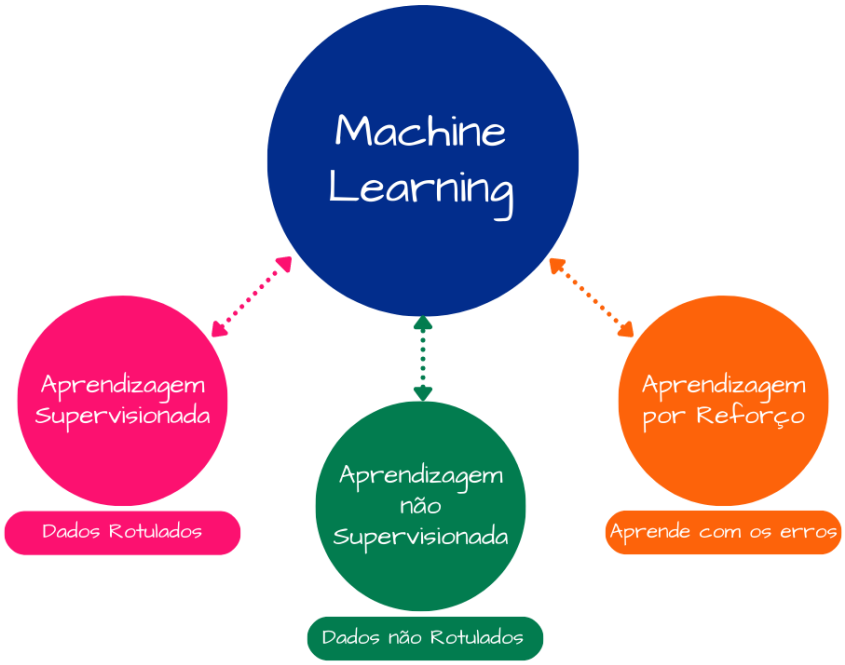


Figura 2.2: Tipos de Aprendizagem em Machine Learning.

Aprendizagem Supervisionada

Na aprendizagem supervisionada, para cada observação da medição dos preditores x_i, i = 1, ..., n, há uma medição de resposta associada y_i [23]. É desejável ajustar um modelo que relacione a variável resposta às variáveis preditoras, com o objetivo de prever com precisão a resposta para futuras observações (previsão) ou entender melhor a relação entre a resposta e os preditores (inferência). Muitos métodos estatísticos clássicos, como regressão linear e regressão logística, bem como abordagens mais modernas, como *GAM*, *boosting* e máquinas de vetor de suporte, operam no domínio de aprendizagem supervisionada [24].

[1] Informações valiosas e perspicazes que podem ser obtidas por meio da análise de dados dos clientes usando *Machine Learning*

Na Figura 2.3 podemos ver diferentes tipos de ajustes do modelo aos dados. No modelo com *underfitting*, o problema está associado à falta de capacidade do modelo na representação dos dados. No modelo com *overfitting*, o problema está associado à sua perda de capacidade de generalização, uma vez que este se ajusta muito aos dados. No modelo ideal, tal como o nome indica, é o modelo perfeito para os dados, o modelo não tem os problemas vistos anteriormente.

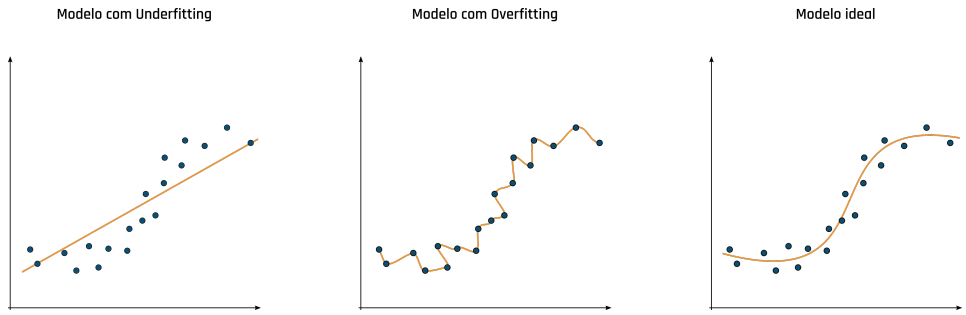


Figura 2.3: Diferentes Ajustes do Modelo aos Dados (Underfitting, Overfitting e Ideal).

Aprendizagem Não Supervisionada

A aprendizagem não supervisionada descreve uma situação um pouco mais desafiadora na qual, para cada observação i = 1, ..., n, observamos um vetor de medidas xi, mas não há uma resposta associada y_i. Não é possível ajustar um modelo de regressão linear, pois não há variável de resposta para prever. Nesse cenário, estamos a trabalhar, de certa forma, às cegas. A situação é denominada não supervisionada porque não temos uma variável de resposta que possa supervisionar a nossa análise [23]. Na Figura 2.4 podemos ver o fluxo de execução da aprendizagem não supervisionada, que recorre a inferências a partir de um conjunto de dados que não foi rotulado, categorizado ou classificado anteriormente.

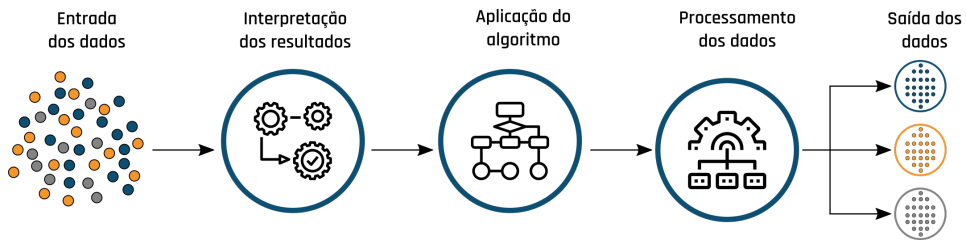


Figura 2.4: Fluxo de Execução da Aprendizagem Não Supervisionada.

Podemos tentar compreender as relações entre as variáveis ou entre as observações. Uma ferramenta estatística de aprendizagem que podemos usar nesse contexto é a análise de *cluster*. O objetivo da análise de *cluster* é determinar, com base em x_1, ..., x_n, se as observações se enquadram em grupos relativamente distintos [24]. Por exemplo, num estudo de segmentação de mercado, podemos observar várias características (variáveis) de potenciais clientes, tais como: código postal, renda familiar e hábitos de compra. Podemos acreditar que os clientes se enquadram em diferentes grupos, como "gastadores" *versus* "não gastadores". Quando a informação sobre os padrões de gastos de cada cliente está disponível, então é possível realizar uma análise supervisionada, caso contrário estaríamos numa análise não supervisionada. No entanto, essa informação não está disponível, ou seja, não sabemos se cada potencial cliente é um "gastador" ou não. Nesse cenário, podemos tentar a análise de *cluster* nos clientes com base nas variáveis medidas, a fim de identificar grupos distintos de potenciais clientes. A identificação desses grupos pode ser de interesse uma vez que esses podem diferir em relação a alguma propriedade de interesse, por exemplo, hábitos de gastos.

Aprendizagem por Reforço

A aprendizagem por reforço corresponde a algoritmos supervisionados nos quais um agente interage com o ambiente e aprende a maximizar a recompensa máxima [25]. Estes algoritmos são frequentemente utilizados como sistemas de recomendação para configurar campanhas, publicidade digital e aumentar o rendimento, promovendo diferentes categorias de produtos e retalho, entre outros.

Com o objetivo de perceber melhor o processo, é importante perceber alguns conceitos chave:

- **Agente** - Entidade, podendo ser tanto um *software* quanto a combinação com *hardware*, que tomará decisões no ambiente, interagindo com ele, tomando determinadas ações e recebendo as recompensas correspondentes;
- **Ambiente** - É a materialização (ou simulação) do problema a ser resolvido. Podendo ser real ou virtual, este será o espaço no qual o agente realizará suas ações.
- **Estado** - É como o sistema, agente e ambiente, encontra-se num determinado instante. Sempre que o agente realiza uma ação, o ambiente fornece um novo estado e uma recompensa correspondente.
- **Recompensa** - Norteia as ações do agente. Se a ação tomada pelo agente num determinado estado for a desejada, a recompensa é positiva, caso contrário, ela é negativa ou nula.

Na figura 2.5, de forma a percebermos melhor como funciona todo o processo, é ilustrado o modelo. O processo começa no instante t (tempo), o agente seleciona uma ação, muda para um novo estado, recebe uma recompensa e então o ciclo é reiniciado para o próximo instante (t+1) .

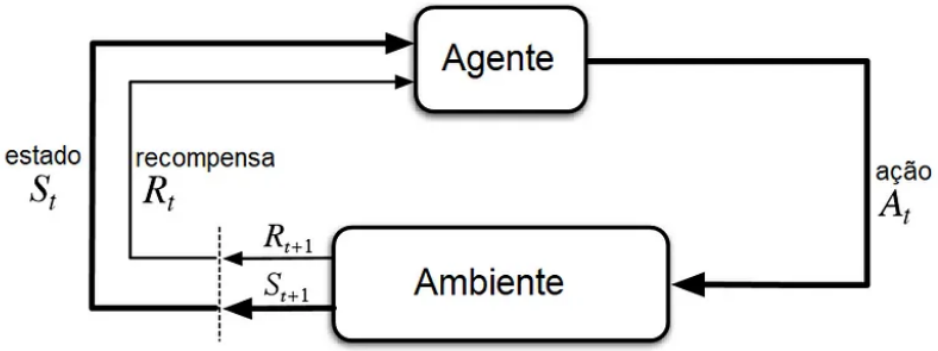


Figura 2.5: Aprendizagem por Reforço.

Na aprendizagem por reforço, o modelo aprende e tenta maximizar a recompensa acumulada, usando um esquema de sobrevivência dos mais aptos, com uma busca organizada aleatória para encontrar a melhor solução para um problema [26].

2.2.2 Problemas Resolvidos com *Machine Learning*

Com o intuito de explorar o potencial do *machine learning* no auxilio às empresas, são identificados três problemas frequentes e os estudos relativos a esses temas para perceber o comportamento do consumidor, segmentar o mercado e, por último, as previsões.

Comportamento do consumidor

O comportamento do consumidor refere-se ao estudo de como os clientes, tanto indivíduos quanto organizações, satisfazem as suas necessidades e desejos ao escolher, comprar, usar e descartar bens, ideias e serviços. Por outras palavras, refere-se à decisão tomada pelos clientes durante o processo de compra e aos fatores que podem influenciar essa decisão [27]. Esses fatores podem ser culturais, sociais, psicológicos, entre outros.

Até aos dias de hoje, foram realizados vários estudos para perceber o comportamento do consumidor. Esses artigos analisaram diversas facetas do comportamento do consumidor em diferentes contextos, como compras *online*, turismo, classificação de marcas, avaliações de usuários, respostas a campanhas e sensibilidades ao preço [28]. Destacam a importância de fatores como a confiança, *design*, segurança, influência social e perceção de valor na tomada de decisão dos consumidores. Além

disso, exploraram o uso de modelos e análises de dados para compreender os padrões de comportamento e prever tendências futuras.

Segmentação do Mercado

A segmentação de mercado é uma das principais estratégias de *marketing*. Como referido anteriormente, o objetivo é identificar os grupos de consumidores para de seguida planear a estratégia de *marketing* para cada segmento [29]. Os artigos analisados nesta área abordaram várias temáticas relativas ao mercado, como preferências de hotéis, segmentação de clientes com base no retalho, comportamento do consumidor nas redes sociais, padrões de interesse no comércio, entre outras [28].

Previsão

A previsão de *marketing* é uma análise que projeta as tendências, características e números futuros no mercado-alvo. Isso permite realizar antecipadamente uma investigação dessas variáveis económicas, por meio de pesquisas de mercado, utilizando métodos de previsão de *marketing* [30]. Desta forma, as empresas podem preparar-se adequadamente, tomar decisões informadas e desenvolver estratégias alinhadas com as dinâmicas do mercado e as necessidades dos clientes. A maioria das previsões apresentadas nos artigos consultados está relacionada com a previsão de preços de mercado, padrões de compra em segmentos de negócios, classificação ou preços de produtos, ou diferença de preços em leilões [28].

Capítulo 3

Caso de Estudo

Neste capítulo é apresentada a base de dados referente ao estudo, assim como uma análise da mesma.

A base de dados original contém 2840 registos e 11 variáveis. O procedimento inicial consistiu em eliminar algumas variáveis não relevantes para o objeto de estudo, portanto, as colunas "Morada", "Localidade", "Código Postal" e "Entidade" foram removidas.

3.1 Caracterização da Base de Dados

Após a etapa de limpeza de dados, a base de dados está adequadamente preparada para a realização de uma análise exploratória de dados. A base de dados abrange vendas em mais de setenta países diferentes, distribuídos globalmente, sendo as variáveis utilizadas:

- **Número**: representa o número da obra associado;
- **Concelho**: representa os concelhos de Portugal e as cidades nos Estados Unidos da América, nas quais foram realizadas as vendas;
- **Distrito**: representa os distritos mais as ilhas de Portugal, em Espanha e em Inglaterra representam as suas cidades, e nos Estados Unidos da América representam os estados nas quais foram realizadas as vendas;

- **Região**: representa as regiões de Portugal (Nut3), as regiões de Espanha, os condados dos estados Unidos da América e os condados de Inglaterra, nas quais foram realizadas as vendas;
- **País**: representa o País na qual foi realizada a venda;
- **Data**: representa a data na qual foi realizada a venda, dia/mês/ano. Começa em 2015 e vai até dezembro de 2022;
- **Valor**: representa o valor da faturação da venda.

Esta base de dados apenas é referente a dados internos da empresa, neste caso, apenas é referente à faturação de cada obra em cada país. Através destes dados, são construidas novas variáveis:

- **Média**: representa a média dos valores de venda em cada país;
- **Projetos**: representa o número total de projetos realizados em cada país;

Ainda nesta dissertação, no capítulo 5, serão abordadas novas variáveis, variáveis externas à empresa, que serão úteis para a construção do modelo.

3.2 Análise Exploratória dos Dados

A Tabela 3.1 apresenta os dez países com maior faturação. Estes dados reportam o volume total de vendas realizadas entre 2015 e 2022, agregado por país. Portugal ocupa a posição cimeira, tendo registado o maior número de projetos realizados. De seguida, surge Espanha que, a par de Portugal, se destaca dos restantes países. As Maldivas e Angola destacam-se pela sua média elevada, tendo um baixo número de vendas.

País	Valor	Media	Projetos
Portugal	57 163 511	113 645	503
Espanha	37 418 159	88 043	425
Estados Unidos	18 406 592	131 476	140
México	13 301 098	158 346	84
Inglaterra	11 151 876	24 349	458
Suíça	8 670 763	52 871	164
Israel	8 626 794	45 404	190
Maldivas	6 945 000	2 315 000	3
Zimbabué	5 932 680	5 932 680	1
Canadá	5 633 704	61 235	92

Tabela 3.1: Os 10 países com maior faturação, entre 2015 e 2022.

A Tabela 3.2 apresenta os cinco concelhos de Portugal com maior faturação. É em Loulé, com menos de metade dos projetos do Porto, que a faturação total e a média por projeto são as mais altas. De seguida, surge Lisboa com o segundo valor mais alto.

Concelho	Valor	Média	Projetos
Loulé	13 430 059	282.872	46
Cascais	9 771 176	128.579	50
Lisboa	9 453 532	139.812	64
Porto	6 755 203	59.011	115
Vila Nova de Gaia	3 536 499	46.305	31

Tabela 3.2: Os 5 concelhos de Portugal com maior faturação, entre 2015 e 2022.

Neste caso, a figura 3.1 representa a faturação total, por ano, desde 2015 a 2022. Revela um aumento na faturação ao longo dos anos, com uma queda de quatro milhões, no ano de 2020 em relação ao ano de 2019. É em 2022, com uma faturação de quase cinquenta milhões, o melhor ano para a empresa. Destes sete anos, 2016 revela-se como o pior ano, com cerca de dezassete milhões.

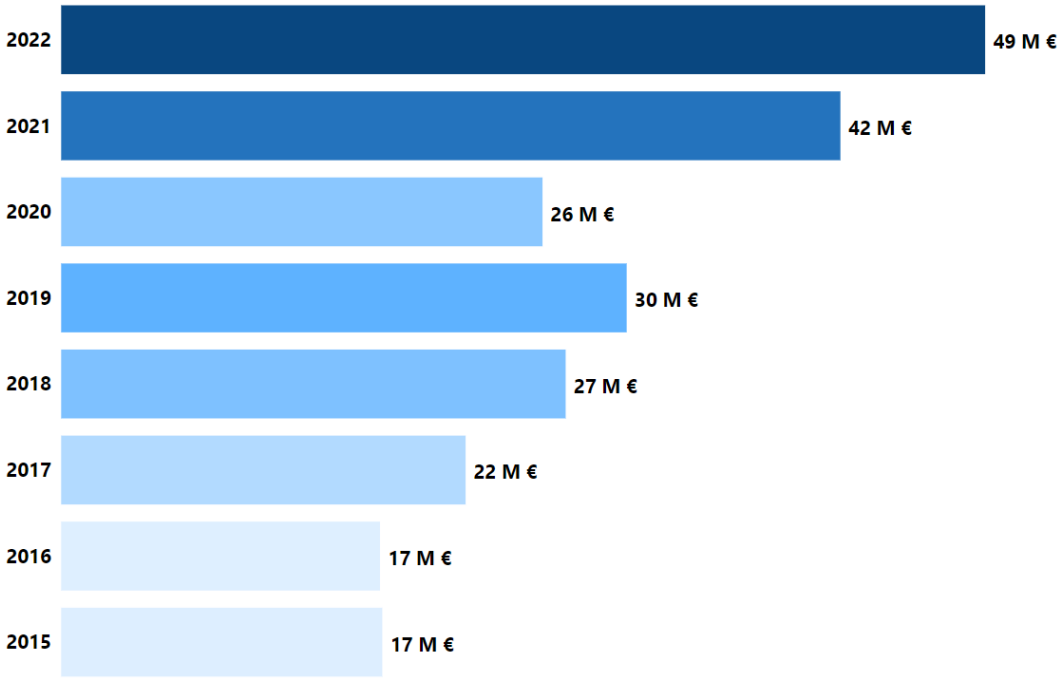


Figura 3.1: Faturação Total por Ano, desde 2015 a 2022

A figura 3.2 representa o número de projetos realizados, por ano, desde 2015 a 2022. Apesar de não ser o ano com maior faturação, 2021 é ano com o maior número de vendas (490 vendas). O ano de 2016, com 253 vendas, é o ano com menos vendas realizadas. Tal como na figura 3.1, há um aumento de vendas ao longo do tempo.

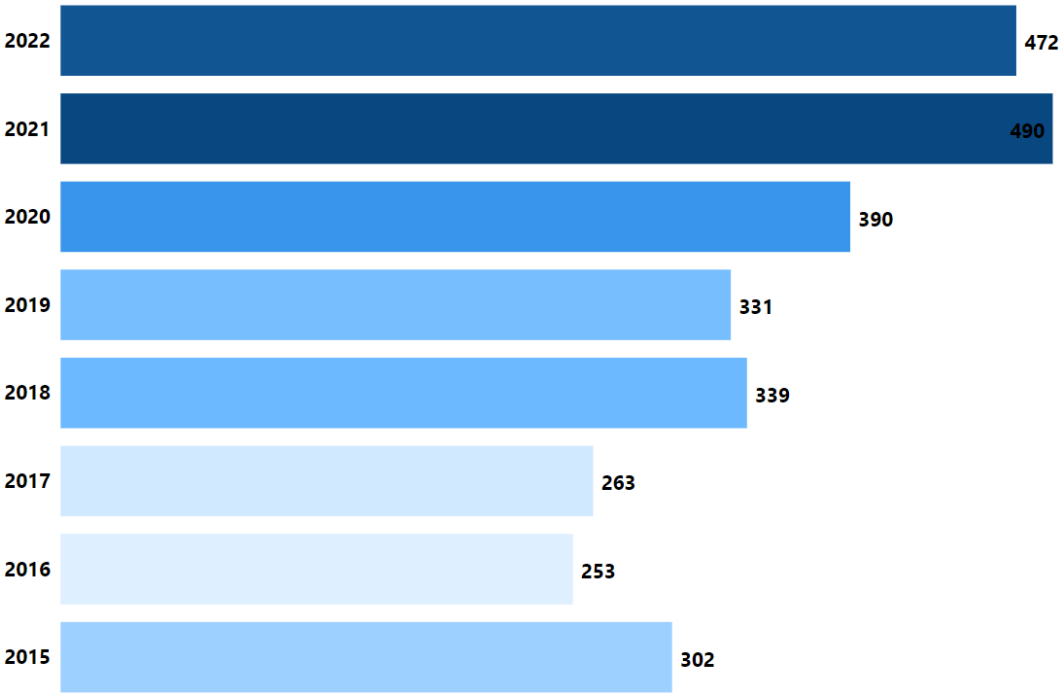


Figura 3.2: Total de Projetos por ano, desde 2015 a 2022

A Figura 3.3 ilustra os oito países que apresentam os maiores valores de vendas (em milhares). Zimbabué e as Maldivas, apesar de um valor alto de faturação, apresentam um número de vendas muito reduzido, deste modo, faria pouco sentido estar representado neste tipo de gráfico. É possível ver neste gráfico vários *outliers*.

A Figura 3.4 ilustra, tal como na Figura 3.3, os oito países com maiores valores de vendas (em milhares), com a diferença do limite do gráfico que foi estabelecido em 400 mil, apenas para uma melhor visualização dos dados. Neste gráfico fica mais fácil analisar as medianas nestes países. México e Estados Unidos apresentam os valores da mediana mais altos, enquanto que Israel e Inglaterra concentram-se muito em valores mais baixos.

A Figura 3.5 ilustra o mapa de Portugal, dividido por concelhos, sendo que a intensidade do azul indica o volume de vendas em cada concelho. Destacam-se no mapa Loulé, Lisboa, Porto e Cascais. A Madeira possui apenas duas regiões com vendas, enquanto que os Açores não têm nenhuma venda. Os concelhos marcados a cinzento não apresentam qualquer venda.

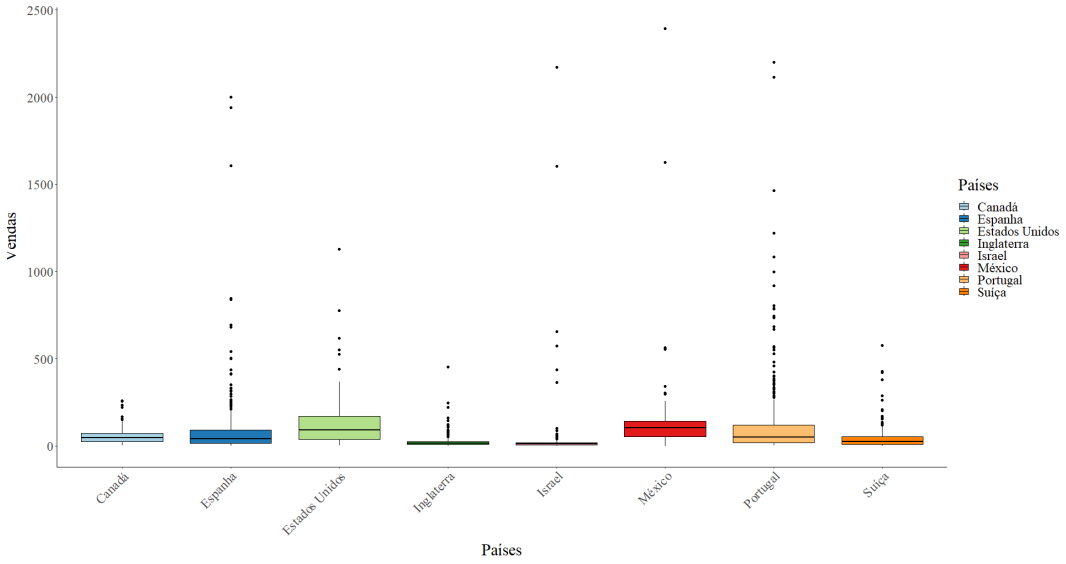


Figura 3.3: *Boxplot* relativo ao valor das vendas, em milhares, de 2015 a 2022, nos 8 países com valor de projetos mais altos (com a exceção de Zimbabué e Maldivas).

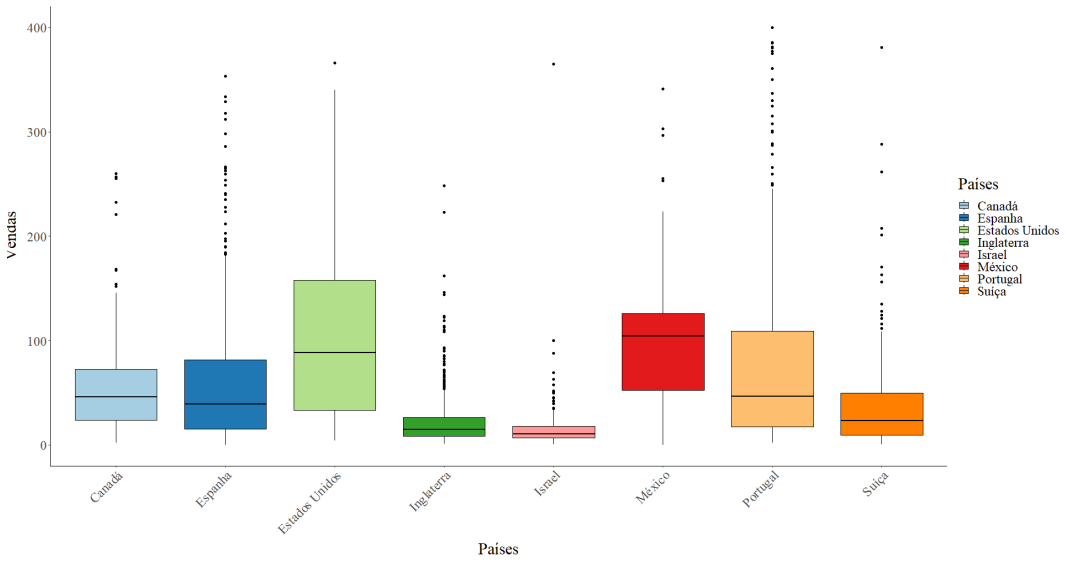


Figura 3.4: *Boxplot*, com limite máximo em 400 mil, relativo ao valor das vendas, de 2015 a 2022, nos 8 países com valor de projetos mais altos (com a exceção de Zimbabué e Maldivas).

Depois de analisar o mapa de Portugal, é importante analisar o gráfico de Espanha, o segundo país com maior faturação. A Figura 3.6 representa a faturação total de 2015 a 2022, separado por províncias, em Espanha. A Comunidade de Madrid, apresenta a maior faturação total, seguido das Ilhas Baleares. As regiões marcadas a cinzento não apresentam qualquer venda.

Por último, temos os Estados Unidos, o terceiro país com maior faturação e o

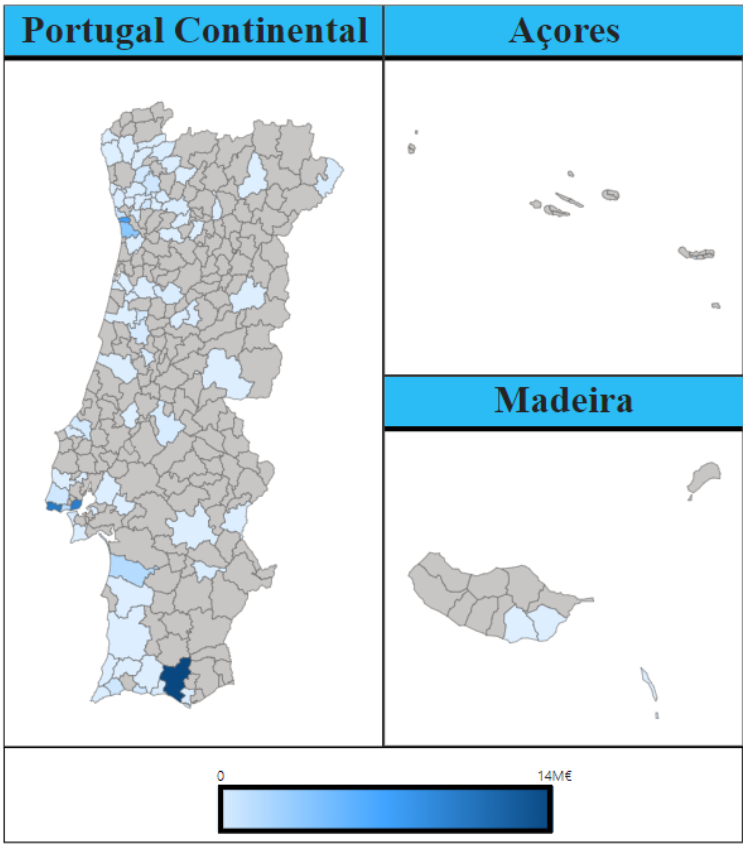


Figura 3.5: Mapa de Portugal, relativo à Faturação Total, por Concelho, desde 2015 a 2022.

sexto país com o maior número de projetos. A Figura 4.1 representa o mapa dos estados dos Estados Unidos. Nesta análise visual podemos realçar três estados: estado de Nova Iorque, estado do Texas e o estado da Califórnia. Os estados marcados a cinzento não apresentam qualquer venda.

Série Temporal

A Figura 3.8 contém quatro gráficos: a distribuição dos dados; os resíduos; a sazonalidade; e a tendência.

No primeiro gráfico, podemos analisar de forma mais detalhada a evolução da faturação total entre janeiro de 2015 e dezembro de 2022. A análise destes dados é fundamental para compreender a dinâmica do mercado e tomar decisões informadas relacionadas à estratégia de vendas.

Vendo a representação gráfica da sazonalidade, podemos concluir que existe sazonalidade. Os meses de agosto e dezembro apresentam os maiores picos de quebra, enquanto que o mês de fevereiro apresenta o maior pico de vendas.

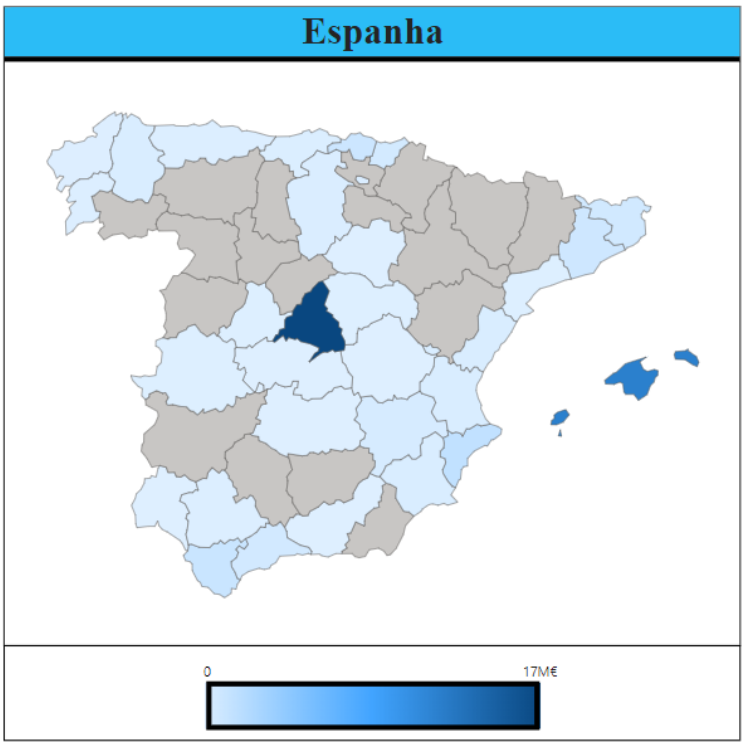


Figura 3.6: Mapa de Espanha, relativo à Faturação Total, por províncias, desde 2015 a 2022.

Relativamente à tendência, conclui-se que existe um aumento de 2015 a 2018 e um ligeiro decaimento de 2018 a 2020. De 2020 a 2022, este valor volta a crescer. No geral, de 2015 a 2022, há um aumento considerável na faturação.

A sazonalidade era de esperar, uma vez que esta acontece nos meses em que a empresa se encontra fechada para férias.

3.3 *Power BI*

No atual cenário empresarial, a capacidade de acessar, interpretar e agir com base em dados é fundamental para o sucesso e para a tomada de decisões. Os dados, quando apresentados de maneira clara e acessível, têm o poder de revelar *insights* valiosos e fornecer uma visão aprofundada do desempenho da empresa. É neste contexto que a criação de uma dashboard no *Power BI* desempenha um papel essencial.

Com o objetivo de apresentar os resultados à empresa, de forma a fornecer uma visão global e interativa dos dados, bem como aprimorar a eficiência operacional e a agilidade na resposta a desafios em constante evolução, foi criada a *dashboard* visível em 3.9 e 3.10. Esta ferramenta tornou-se um ativo valioso para a empresa,

Figura 3.7: Mapa dos Estados Unidos, relativo à Faturação Total, por Estados, desde 2015 a 2022.

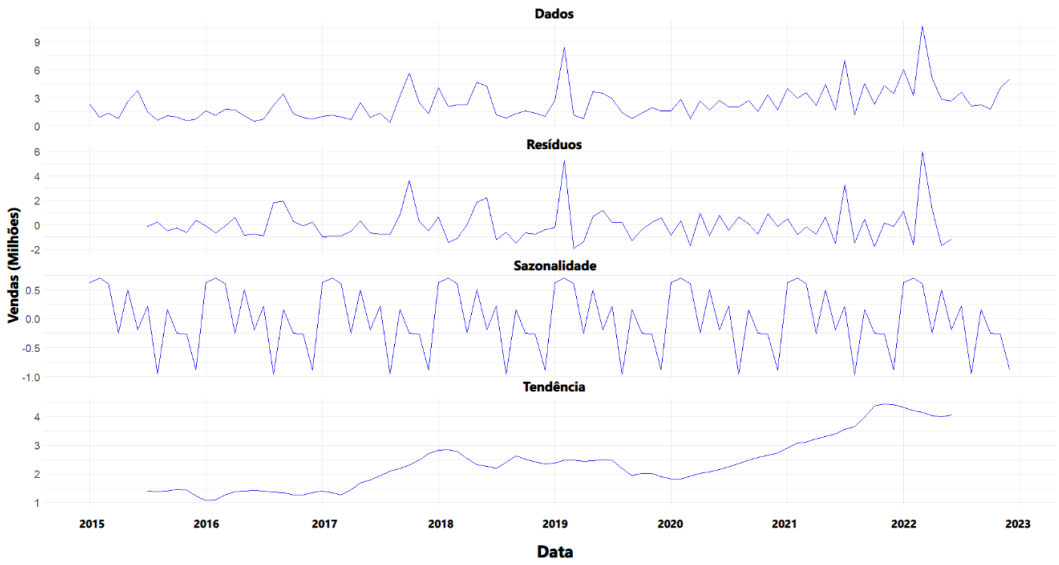


Figura 3.8: Sazonalidade, Tendência e Resíduos, de janeiro de 2015 a dezembro de 2022, com dados relativos à faturação total, agrupadas por mês.

capacitando as equipas de obter informações adicionais que impulsionam a tomada de decisões estratégicas e operacionais.

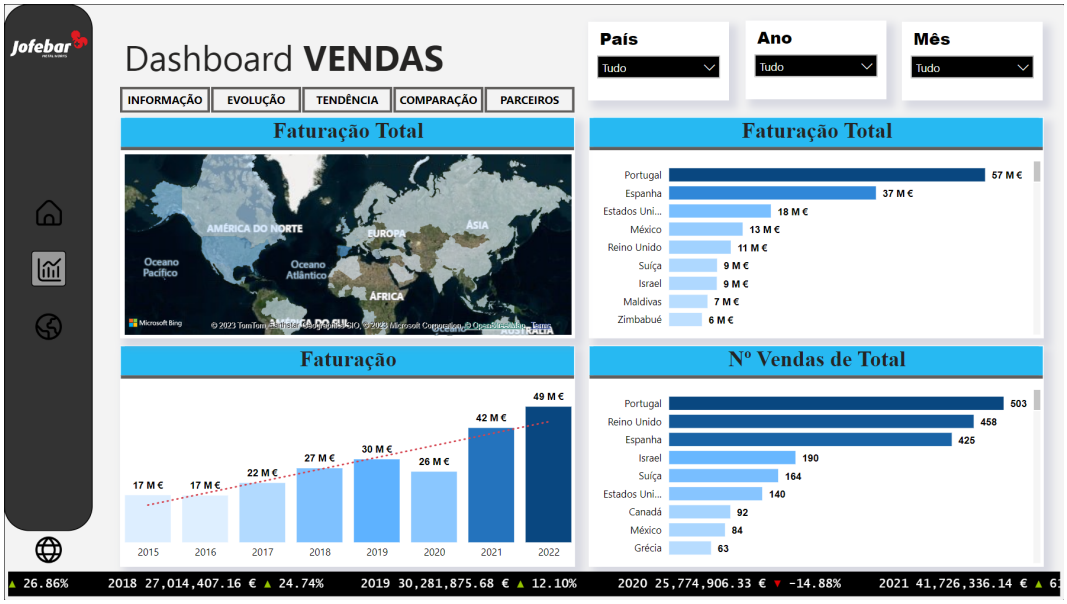


Figura 3.9: *Dashboard* Relativa à Faturação Total da Empresa

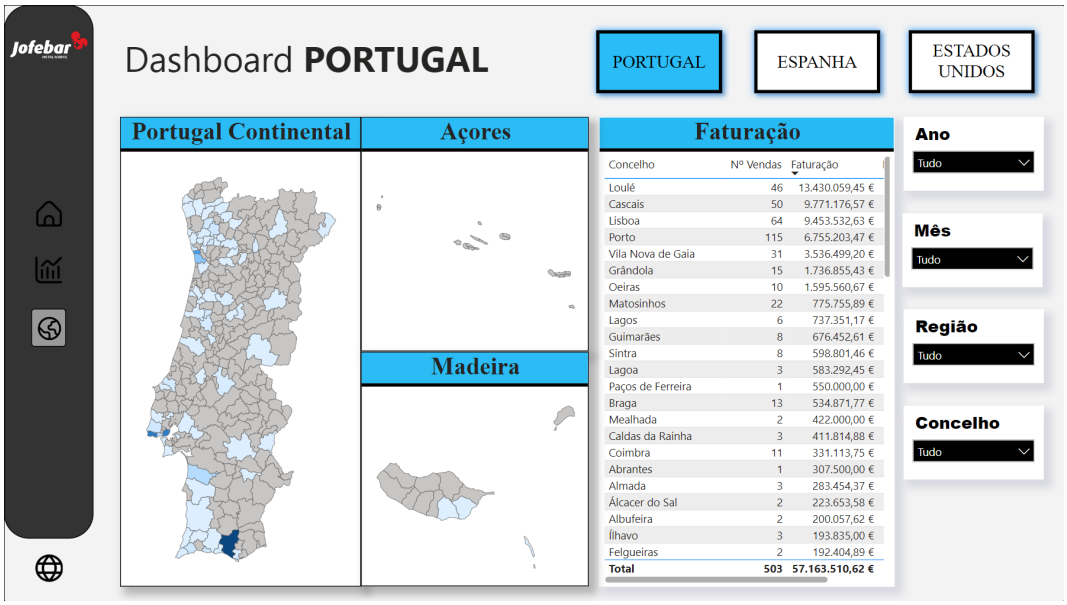


Figura 3.10: *Dashboard* Relativa à Faturação em Portugal

A figura 3.9 mostra, num modo geral, a faturação total. É ainda possível segmentar os dados por país, ano e mês.

A figura 3.10 mostra a faturação total, por concelho, em Portugal. É ainda possível segmentar os dados por região (NUT 3), concelho, ano e mês. Esta análise pode ainda ser efetuada em Espanha e nos Estados Unidos.

Capítulo 4

Metodologia Estatística

Neste capítulo, serão apresentados as técnicas e métodos estatísticos que serão empregados para atingir os objetivos deste estudo. Especificamente, este capítulo vai se concentrar na aplicação de quatro abordagens essenciais: a regressão linear e algumas técnicas de *Machine Learning*, nomeadamente, *random forest*, *bagging* e o *boosting*.

4.1 Regressão Linear

Neste capítulo será abordado o conceito de regressão linear, uma abordagem simples para aprendizagem supervisionada. A regressão linear é uma ferramenta útil para prever resultados quantitativos. É uma técnica antiga, discutida em inúmeros recursos didáticos. Embora possa parecer menos interessante em comparação a algumas abordagens de estatísticas mais modernas, ainda é um método estatístico valioso e amplamente utilizado. Além disso, serve como uma base sólida para compreender abordagens mais avançadas, portanto, a importância de compreender bem a regressão linear antes de explorar métodos de *Machine Learning* não pode ser subestimada.

Neste capítulo, revisaremos algumas das ideias fundamentais que sustentam o modelo de regressão linear, bem como a abordagem de mínimos quadrados, que é amplamente utilizada para ajustar esse modelo.

Todos os modelos são aproximações da realidade. Na escolha do melhor modelo, pelo princípio da parcimónia, é preferível o mais simples de todos os modelos considerados adequados. Num modelo estatístico, a relação entre a variável resposta e preditores não é necessariamente uma relação causa efeito, isto é, existe apenas uma associação.

O modelo matemático da regressão linear dá-se por,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, ..., n \tag{4.1}$$

em que:

- Y_i é a variável resposta i;
- $\beta_j, j = 1, \cdots, p$ são os parâmetros (coeficientes de regressão) a determinar;
- x_{ij} é o valor conhecido da variável independente na observação i;
- $\varepsilon_i \sim N(0, \sigma^2)$ é o erro da observação i.

Na forma matricial,

$$\vec{\boldsymbol{Y}} = \boldsymbol{x}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}, \quad \vec{\boldsymbol{\varepsilon}} \sim MN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \tag{4.2}$$

onde

$$\vec{\boldsymbol{Y}} = [Y_i]_{n\times 1}, \tag{4.3}$$

$$\boldsymbol{x} = [x_{ij}]_{n\times(p+1)} \tag{4.4}$$

com

$$x_{i0} = 1 \ ; \ \vec{\boldsymbol{\varepsilon}} = [\varepsilon_i]_{n\times 1} \tag{4.5}$$

Os três pressupostos estatísticos envolvidos na especificação do modelo são:

- **Independência** - ε_i's são independentes entre si;
- **Homocedasticidade** - ε_i's tem variância igual a σ^2;
- **Normalidade** - ε_i's tem distribuição normal.

Os parâmetros β_j da regressão são desconhecidos e têm de ser estimados. Para tal, recorre-se ao método mais popular, método dos mínimos quadrados [31]. Este minimiza a soma dos quadrados dos resíduos. O estimador $\hat{\beta}$ é dado por:

$$\hat{\vec{\beta}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{Y}. \tag{4.6}$$

4.1.1 Coeficiente de Determinação

O Coeficiente de Determinação (R^2) é uma medida de ajustamento e toma valores entre zero e um, indicando o quanto o modelo consegue explicar os valores observados. Um valor de R^2 mais alto, resulta num modelo mais explicativo e que melhor se ajusta à amostra. Por exemplo, $R^2 = 1$, significa que todas as observações estão sobre a reta da regressão.

O coeficiente de determinação é o quadrado do coeficiente de correlação empírico,

$$R = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2\Sigma(Y_i - \bar{Y})^2}} \tag{4.7}$$

em que $\bar{X}$ e $\bar{Y}$ são as médias dos X_i e Y_i, respetivamente.

Este varia de -1 a 1 e quanto maior for o número em módulo, mais forte é a correlação, por exemplo, R = 1, representa uma correlação forte e positiva; R = -1, representa uma correlação forte e negativa; as correlações mais fracas têm valores próximos de 0.

Uma outra forma de calcular R^2 é usando as somas dos quadrados totais, a soma dos quadrados dos resíduos e a soma dos quadrados da regressão (SQT, SQE e SQR, respetivamente).

A medida da variabilidade total do conjunto de observações de Y é expressa em função da soma dos quadrados dos desvios totais.

$$SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{4.8}$$

A variabilidade total pode ser dividida em duas partes, a primeira parte referente à soma dos quadrados dos erros (SQE) e a segunda referente à soma dos quadrados da regressão (SQR).

$$SQE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4.9}$$

$$SQR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \tag{4.10}$$

O R^2 é definido como a proporção entre a variação explicada pela regressão e a variação total,

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

O R^2 tem a capacidade de encontrar a probabilidade de acontecimentos futuros dentro dos resultados previstos, isto é, se forem adicionadas novas amostras, o coeficiente mostrará a probabilidade de um novo ponto cair na linha estimada pela regressão [32].

4.1.2 Fator de Inflação da Variância (VIF)

Considerando que as variáveis estão centralizadas e padronizadas, podemos expressar a matriz de correlação inversa como $R = (X^T X)^{-1}$. Nessa matriz, os elementos da diagonal são conhecidos como VIF e quantificam o aumento na variância devido à presença de multicolinearidade.

O VIF é definido como:

$$VIF_j = \frac{1}{1 - R_j^2} \qquad j = 1, 2, \cdots, p \tag{4.11}$$

em que:

- p representa o número das variáveis preditoras;
- R_j^2 representa o coeficiente de correlação múltipla, resultante da regressão de X_j nos outros $p - 1$.

Quando o valor de R_j^2 se aproxima de 1 indica uma forte correlação entre a variável X_j e as outras variáveis. Como consequência, $1 - R_j^2$ aproxima-se de zero, o que resulta num VIF elevado, sugerindo que esta variável está envolvida em multicolinearidade. Um VIF máximo acima de 10 indica que a multicolinearidade pode estar a afetar as estimativas de mínimos quadrados.

4.1.3 Método de Seleção dos Modelos

A abordagem *stepwise* envolve a inclusão e/ou exclusão iterativa de variáveis preditoras no modelo de previsão. Isso é feito com o objetivo de identificar o conjunto de variáveis mais adequado no conjunto de dados, resultando num modelo de alto desempenho que minimiza o erro de predição.

A seleção por *backward* começa com todos os preditores no modelo (modelo completo), remove iterativamente os preditores menos contributivos e para quando se tem um modelo em que todos os preditores são estatisticamente significativos.

Na seleção *forward*, o processo começa sem nenhum preditor no modelo e, em seguida, adiciona de forma iterativa os preditores que contribuem mais significativamente para o modelo. Esse processo é interrompido quando a melhoria não é mais estatisticamente significativa.

A abordagem *stepwise* é uma combinação de seleção para frente e para trás. Inicia-se sem nenhum preditor e, em seguida, acrescenta de forma sequencial os preditores que contribuem significativamente (semelhante à seleção para frente). Após a inclusão de cada nova variável, são removidas aquelas que não oferecem mais aprimoramento no ajuste do modelo, num processo semelhante à seleção para trás.

Na escolha do modelo final, usando o método de seleção dos modelos, o Critério Informação de *Akaike* (AIC) desempenha um papel fundamental.

Critério de Informação de *Akaike*

O AIC é uma métrica para avaliar a qualidade do ajuste de modelos estatísticos. Num cenário de Regressão Linear Múltipla com k variáveis independentes, é definido como:

$$AIC = n \cdot ln(\frac{SQE_k}{n}) + 2(k+1) \tag{4.12}$$

O AIC mede dois aspetos importantes num modelo: a qualidade do ajustamento e a sua simplicidade. Pela equação 4.12, a primeira parcela mede o ajustamento, isto é, quanto menor o valor, melhor é o ajustamento. Relativamente à segunda parcela, esta mede a complexidade do modelo, sendo que o modelo que envolva o mínimo de parâmetros possíveis a serem estimados é o melhor modelo, com $k+1$ o número de parâmetros. Um modelo é considerado melhor que o outro se tiver um AIC menor.

4.2 Métodos Baseados em Árvores

Os métodos baseados em árvores representam uma classe poderosa de técnicas de *Machine Learning* e análise estatística. Estes são amplamente utilizados para resolver problemas de classificação, regressão e tomada de decisões numa variedade de domínios, incluindo ciência de dados, engenharia, medicina e finanças.

Os métodos baseados em árvores envolvem a criação de estruturas de árvore que dividem os dados em regiões cada vez mais específicas. Essas árvores consistem em nós e em ramos, onde cada nó representa uma decisão ou teste sobre um atributo dos dados, e cada ramo leva a uma ramificação subsequente [33].

- **Árvores de Decisão:** Uma das formas mais comuns de métodos baseados em árvores são as árvores de decisão. Estas são usadas para tomar decisões binárias, como "Sim"ou "Não", e são frequentemente usadas em problemas de classificação. A árvore começa com um nó raiz que contém a pergunta principal

e, em seguida, divide-se em nós filhos que representam respostas às perguntas subsequentes.

- **Árvores de Regressão:** Enquanto as árvores de decisão são usadas para problemas de classificação, as árvores de regressão são aplicadas quando a variável de destino é contínua. Estas ajudam a prever valores numéricos, dividindo o espaço de recursos em intervalos e atribuindo valores de saída a cada intervalo.
- **Construção de Árvores:** A construção de árvores envolve a seleção de atributos e pontos de divisão ideais para minimizar a impureza ou o erro. Isto é geralmente feito usando critérios como o *Gini impurity* para classificação e a redução do erro quadrático médio (MSE) para regressão.

As principais vantagens de usar estas técnicas são:

- **Interpretabilidade:** As árvores são altamente interpretáveis, o que significa que podemos entender facilmente como tomam decisões com base nas regras nas bifurcações da árvore.
- **Lidar com Dados Não Lineares:** Árvores podem lidar com relacionamentos não lineares entre variáveis, tornando-os úteis em muitos cenários do mundo real.
- **Escalabilidade:** São eficientes para conjuntos de dados grandes e podem ser acelerados usando técnicas como o *Random Forest*.

Apesar de apresentar vantagens, tem sempre a parte dos desafios para resolver que são:

- ***Overfitting:*** Árvores individuais podem ser propensas a *overfitting*, ou seja, ajustarem-se muito aos dados de treino e não generalizarem bem para novos dados.
- **Baixa Robustez:** Árvores únicas podem ser sensíveis a pequenas variações nos dados de treino, resultando em diferentes árvores para conjuntos de dados semelhantes.

Para abordar os desafios mencionados, foram desenvolvidas soluções mais avançadas:

- ***Random Forest:*** Este método cria múltiplas árvores independentes e combina as suas previsões, reduzindo o *overfitting* e aumentando a robustez.
- ***Boosting:*** É uma técnica que melhora o desempenho do modelo, dando mais peso às observações mal classificadas, gerando árvores ponderadas.

A preparação adequada dos dados desempenha um papel crucial na construção de modelos de *Machine Learning* robustos e eficazes. Antes de qualquer algoritmo de *Machine Learning* ser aplicado, é imperativo submeter os dados a um processo minucioso de limpeza e transformação. Esta etapa, muitas vezes subestimada, é fundamental para garantir que o modelo extraia informações valiosas e generalize bem para dados não vistos.

Um aspeto fundamental da preparação de dados é a identificação e tratamento de valores ausentes, *outliers* e erros. Essas anomalias podem distorcer a capacidade do modelo de aprender relações significativas nos dados. Portanto, a eliminação cuidadosa desses valores é crucial para manter a integridade dos dados.

Além disso, a codificação adequada de variáveis categóricas é essencial. Muitos algoritmos de *Machine Learning* exigem que todas as variáveis sejam numéricas. Portanto, é necessário converter variáveis categóricas em variáveis numéricas.

A normalização ou padronização de características também é uma etapa importante. A escala das variáveis pode variar significativamente, e algoritmos como regressão logística ou SVM (*support vector machine*) são sensíveis a essa disparidade. Portanto, é comum aplicar técnicas que garantam que todas as variáveis tenham a mesma influência no modelo.

No entanto, a preparação dos dados é apenas metade da equação. A forma como os dados são divididos em conjuntos de treino e em conjuntos de teste desempenham um papel crucial na avaliação do desempenho do modelo.

O conjunto de treino é usado para alimentar o modelo e permitir que ele aprenda com os dados. É aqui que o modelo é ajustado às peculiaridades dos dados de treino. Por outro lado, o conjunto de teste é reservado exclusivamente para avaliar o quão bem o modelo generaliza para dados não vistos. Isso é fundamental para verificar se o modelo está a capturar padrões genuínos ou simplesmente a memorizar os dados de treino.

A divisão dos dados deve ser feita de maneira aleatória e estratificada, especialmente quando lidamos com tarefas de classificação desequilibradas. Essa abordagem garante que as características das classes sejam preservadas em ambos os conjuntos.

Validação Cruzada

A validação cruzada é uma técnica usada para avaliar o desempenho de um modelo de *Machine Learning* de forma mais robusta e precisa. Ela ajuda a estimar o quão bem um modelo generaliza para dados não vistos, o que é crucial para evitar a super estimação do desempenho do modelo.

A ideia central da validação cruzada é dividir o conjunto de dados em várias partes, geralmente chamadas de *fold*, onde uma parte é usada como conjunto de teste e o restante como conjunto de treino. O processo é repetido várias vezes, com cada *fold* atuando como conjunto de teste numa iteração diferente. No final, as

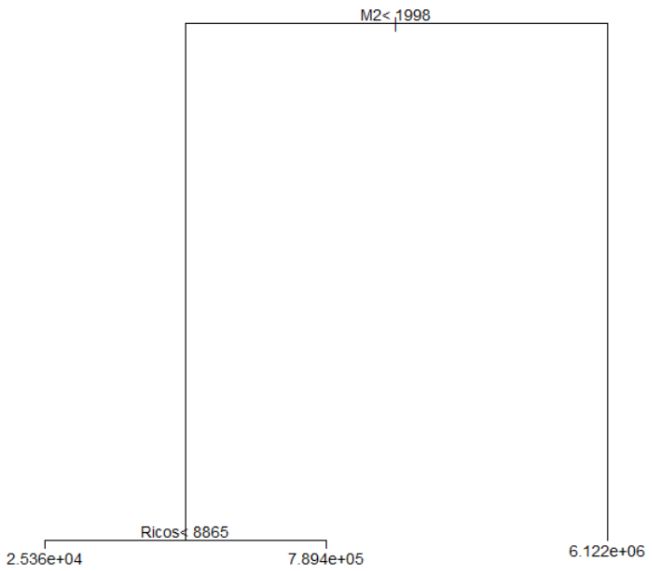


Figura 4.1: Exemplo de uma árvore de decisão.

métricas de desempenho são agregadas e usadas para avaliar o modelo de forma mais abrangente.

A validação cruzada é crucial por várias razões:

- **Estimativa mais confiável:** Em vez de depender de uma única divisão de treino/teste, a validação cruzada utiliza múltiplas divisões, o que resulta numa estimativa mais confiável do desempenho do modelo.
- **Avaliação justa:** Garante que o modelo seja testado em diferentes conjuntos de dados, evitando viés na avaliação.
- **Deteção de *Overfitting*:** Ajuda a identificar se o modelo está super ajustado (*overfitting*) aos dados de treino ou se é capaz de generalizar bem para dados não vistos.

Sem a validação cruzada, existe o risco de criar um modelo que funciona bem apenas nos dados de treino, mas falha em fornecer boas previsões para novos dados.

Quando a validação cruzada é utilizada, é obtido uma avaliação mais realista e precisa do desempenho do modelo, sendo considerado múltiplas divisões de treino/teste. Isso permite tirar conclusões com uma maior confiança, relativamente aos dados que temos.

Medidas de Importância de Variáveis

Em análises de regressão, é essencial compreender a contribuição relativa das variáveis independentes para o modelo. As medidas de importância de variáveis desempenham um papel crucial nesse contexto, permitindo avaliar quais preditores têm maior influência nas previsões e, assim, auxiliando na interpretação dos resultados. Uma dessas medidas é o SQE, que nos fornece uma visão valiosa do impacto das variáveis num modelo de regressão. Variáveis que resultam numa redução substancial do SQE, quando incluídas no modelo, são consideradas mais importantes, pois têm um impacto significativo nas previsões.

No entanto, a interpretação das medidas de importância de variáveis requer cuidado, pois o SQE não considera automaticamente a multicolinearidade (correlação entre preditores) e outros fatores que podem afetar a relevância das variáveis. Portanto, enquanto o SQE fornece uma visão valiosa da importância das variáveis em modelos de regressão, é importante considerá-lo em conjunto com outras técnicas de seleção e avaliação de características para obter uma imagem completa da contribuição de cada variável na modelação.

4.2.1 *Bagging*

Aumentar a precisão dos modelos de *Machine Learning* é uma procura constante na área da ciência de dados e na inteligência artificial. Uma técnica que se tem destacado nesse contexto é o *bagging*, que utiliza uma abordagem conhecida como *bootstrap* para aprimorar o desempenho dos modelos. Para entender completamente o *bagging* é essencial compreender em que consiste o *bootstrap*.

O *bootstrap* é uma técnica estatística revolucionária que transformou a forma como abordamos a inferência estatística e a validação de modelos. O conceito central baseia-se na reamostragem, que nos permite extrair informações valiosas de um conjunto de dados, de forma mais robusta e eficaz. Funciona criando múltiplas amostras, chamadas de *bootstraps*, a partir de um conjunto de dados original, com a particularidade de que cada amostra é obtida através de reamostragem com reposição. Por outras palavras, o *bootstrap* permite que os dados sejam selecionados aleatoriamente para formar uma nova amostra, com a possibilidade de que a mesma observação possa ser escolhida mais que uma vez.

Esse processo permite estimar a distribuição da estatística de interesse diretamente a partir dos nossos dados originais. Ele é especialmente valioso quando não é possível obter mais dados ou quando não queremos fazer suposições rígidas sobre a distribuição dos dados.

É então que surge o *Bagging*, ou seja, método de agregação de amostras *bootstrap*. Como relatado anteriormente, este é um procedimento com o intuito de reduzir a variância de um método de aprendizagem estatística. Dado um conjunto de n

observações independentes $Z_1, \cdots, Z_n$, cada um com variância σ^2, com a variância da média $\bar{Z}$ das observações é dada por σ^2/n. Por outras palavras, a média de um conjunto de observações reduz a variância. Portanto, uma maneira natural de reduzir a variância e, assim, aumentar a precisão das previsões de um método estatístico é criar vários conjuntos de treino a partir da população, construir um modelo de previsão separado usando cada conjunto de treino e média das previsões resultantes. Poderíamos calcular $\hat{f}^1(x), \hat{f}^2(x), \cdots, \hat{f}^B(x)$ usando B conjuntos de treino separados e medi-los para obter um único modelo de aprendizagem estatístico com baixa variância, dado por:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{4.13}$$

Apesar de tudo, não é pratico, uma vez que normalmente não temos acesso a múltiplos conjuntos de treino. Em vez disso, podemos aplicar a técnica de *bootstrap*, amostragens repetidas de um único conjunto de dados de treino. Nessa abordagem, geramos B conjuntos de dados de treino *bootstrap* diferentes. Em seguida, treinamos o nosso método no b-ésimo conjunto de treino para obter $\hat{f}^{*b}(x)$ e, finalmente, fazemos a média de todas as previsões para obter

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x). \tag{4.14}$$

Isto definde o modelo *Bagging*.

4.2.2 *Random Forest*

Assim como no *bagging*, construímos um número de árvores de decisão em amostras de treino *bootstrap*. No entanto, ao construir essas árvores de decisão, cada vez que uma divisão numa árvore é considerada, uma amostra aleatória de m preditores é escolhida como candidatos para a divisão a partir do conjunto completo de p preditores. A divisão só pode usar um desses m preditores. Uma nova amostra de m preditores é retirada a cada divisão e, normalmente, escolhemos $m \approx \sqrt{p}$, ou seja, o número de preditores considerados em cada divisão é aproximadamente igual à raiz quadrada do número total de preditores.

Por outras palavras, ao construir uma floresta aleatória, em cada divisão na árvore, o algoritmo não está autorizado a considerar a maioria dos preditores disponíveis. Supondo que existe um preditor muito forte no conjunto de dados, juntamente com vários outros preditores moderadamente fortes, na maioria ou em todas as árvores do modelo *bagging*, o forte preditor é usado na primeira divisão. Como resultado, todas as árvores parecerão bastante semelhantes entre si. Portanto, as previsões serão altamente correlacionadas. Fazer a média de quantidades altamente correlacionadas não resulta numa redução tão grande na variância quanto fazer a

média de quantidades não correlacionadas. Em particular, isto significa que o *bagging* não levará a uma redução substancial na variância em relação a uma única árvore nesse cenário.

O *random forest* supera esse problema ao forçar cada divisão a considerar apenas um subconjunto dos preditores. Portanto, em média, $(p - m)/p$ das divisões nem sequer considerarão o forte preditor, dando assim mais oportunidades para outros preditores. Podemos pensar neste processo como uma descorrelação das árvores, tornando a média das árvores resultantes menos variável e, portanto, mais confiável.

A principal diferença entre o *bagging* e o *random forest* é a escolha do tamanho do subconjunto de preditores m. Por exemplo, se uma floresta aleatória for construída usando $m \approx \sqrt{p}$, isso equivale simplesmente ao *bagging*. Usar um valor pequeno de m ao construir uma floresta aleatória geralmente será útil quando temos um grande número de preditores correlacionados.

4.2.3 *Boosting*

O *Boosting* é mais uma abordagem com o intuito de melhorar as previsões resultantes de uma árvore de decisão. Assim como o *bagging*, o *boosting* é uma abordagem geral que pode ser aplicada a muitos métodos de aprendizagem estatística para regressão ou classificação.

Lembrando que o *bagging* envolve a criação de múltiplas cópias do conjunto de treino original usando o método de *bootstrap*, ajustando uma árvore de decisão separada para cada cópia e depois combinando todas as árvores para criar um único modelo preditivo. Notavelmente, cada árvore é construída num conjunto de dados de *bootstrap*, independente das outras árvores. O *boosting* funciona de maneira semelhante, exceto que as árvores são cultivadas sequencialmente, isto é, cada árvore é cultivada usando informações das árvores cultivadas anteriormente. O *boosting* não envolve amostragem de *bootstrap*, em vez disso, cada árvore é ajustada a uma versão modificada do conjunto de dados original.

Primeiro, considere o cenário de regressão. Assim como o *bagging*, o *boosting* envolve a combinação de um grande número de árvores de decisão, $\hat{f}^1, \cdots, \hat{f}^B$.

Ao contrário de ajustar uma única árvore de decisão grande aos dados, o que equivale a ajustar os dados de forma intensa e potencialmente causar *overfitting*, a abordagem do *boosting* aprende de forma gradual. Dado o modelo atual, ajustamos uma árvore de decisão aos resíduos do modelo. Ou seja, ajustamos uma árvore usando os resíduos atuais, em vez do resultado Y, como resposta. Em seguida, adicionamos essa nova árvore de decisão à função ajustada para atualizar os resíduos. Cada uma dessas árvores pode ser relativamente pequena, com apenas alguns nós terminais. Ao ajustar árvores pequenas aos resíduos, melhoramos gradualmente $\hat{f}$ nas áreas em que não se sai bem. Em geral, abordagens de aprendizagem estatística que aprendem lentamente tendem a sair-se bem. Note que no *boosting*, ao contrário

do *bagging*, a construção de cada árvore depende fortemente das árvores que já foram desenvolvidas.

Capítulo 5

Resultados

O capítulo dos resultados é a parte central desta investigação, onde os frutos de um extenso trabalho de análise são revelados. Neste capítulo, são apresentados os resultados obtidos a partir das análises e dos modelos desenvolvidos ao longo deste estudo. Estes resultados fornecem *insights* cruciais sobre os padrões, tendências e conclusões alcançadas com base nos dados coletados e nas metodologias empregadas.

Como recordado no Capítulo 3 e 4, a base de dados e a metodologia que será utilizada, respetivamente, terão o seu papel desempenhado neste capítulo. Será revista a variável relativa à faturação (variável resposta), tais como todas as variáveis explicativas/preditoras.

Para a criação deste modelo, vão ser testados os modelos lineares múltiplos e os métodos de *Machine Learning* (*Bagging*, *Boosting* e *Random Forest*).

Nos três países que vão ser analisados, Portugal, Espanha e Estados Unidos, as variáveis explicativas são variáveis externas à empresa. Essas variáveis têm uma componente mais ligada à riqueza pessoal, não estando ligada a outro tipo de fatores. A escolha destes países está ligada ao grande faturamento existente, desde 2015 a 2022, e ao grande potencial que estes países têm para este nicho de mercado.

5.1 Portugal

Portugal é o país com maior faturamento e número de vendas. País na qual a empresa está sediada, sendo esse um dos grandes fatores para tal faturamento. Em Portugal, tal como podemos comprovar pela tabela 3.2 e pela figura 3.5, há 3 regiões

em que a faturação é mais elevada, são essas a zona do Porto, Lisboa e Loulé (Quinta do Lago). Sendo o produto direcionado para um tipo de cliente muito específico, um cliente que tenha uma alta riqueza, as variáveis utilizadas em Portugal tentaram explicar isso. A análise será feita por concelho.

As variáveis utilizadas para este estudo são:

- ***População*** - Número de Habitantes;
- ***Densidade*** - Densidade Populacional;
- ***Área*** - Área em Quilómetros;
- ***Rendimentos*** - Rendimento Médio Anual;
- ***Gini*** - Índice de Gini;
- ***Rend35k*** - Número de Pessoas com rendimento Anual superior a 35 Mil Euros;
- ***CarrosHab*** - Número de Carros de Luxo[1] por Habitante;
- ***CarrosRend*** - Número de Carros de Luxo por 100 Mil de Rendimento;
- ***M2*** - Preço do Metro Quadrado para Construção;
- ***Campos*** - Número de Campos de Golfe;
- ***Fogos*** - Número de Novos Fogos Licenciados (T4 ou superior);
- ***P90P10*** - Indicador de Desigualdade na Distribuição do Rendimento. Número de Vezes que o Percentil 90 é Superior ao Percentil 10.

As variáveis foram coletadas do INE (Instituo Nacional de Estatística), Jornal Expresso, Idealista e GPEARI (Gabinete de Planeamento, Estratégia, Avaliação e Relações Internacionais).

5.1.1 Regressão Linear

O modelo inicial, com todas as variáveis, usado no R é:

$$\begin{aligned} Valor \sim Popula\tilde{c}ao + Densidade + \acute{A}rea + Rendimentos + Gini + \\ Rend35k + CarrosHab + CarrosRend + M2 + Campos + \\ Fogos + P90P10. \end{aligned} \tag{5.1}$$

[1] São considerados todos os carros das marcas: *Maserati*, *Lamborghini*, *Bentley*, *Ferrari*, *Jaguar*, *Porsche* e *Aston Martin*.

A partir daqui, existem alguns pressupostos que o modelo deve seguir. Um dos pressupostos assume que o modelo não poderá ter colinearidade. A multicolinearidade pode dificultar a interpretação dos coeficientes e levar a resultados instáveis. Para tal, é necessário medir a multicolinearidade entre as variáveis explicativas.

Correlação das Variáveis

A correlação pode ser calculada usando três coeficientes: coeficiente de *Pearson*, coeficiente de *Spearman* e coeficiente de *Kendall*. Como se pretende calcular a correlação entre variáveis quantitativas, o coeficiente de *Pearson* é o mais adequado. O coeficiente de correlação entre todas as variáveis explicativas foi calculado e está representado na figura 5.1.

A matriz de correlação está representada em tons que varia de vermelho a azul, sendo o vermelho a cor que representa a correlação negativa -1 e o azul escuro a correlação positiva $+1$. Quando o valor é próximo de 0, afirma-se que não há correlação linear, sendo representado pela cor branca.

Um dos pressupostos do modelo linear é a não-correlação entre variáveis explicativas. Assim sendo, é necessário selecionar as variáveis mais importantes e que não apresentem correlação elevada, eliminando as variáveis fortemente correlacionadas.

A figura 5.1 é relativa às correlações do modelo para Espanha. Definiu-se o valor $\pm 0,80$ como valor de referência, eliminando todos os valores, com valor absoluto, maior que $0,80$. A variável removida foi ***Rend35k***.

VIF (*Variance Inflation Factor*) do Modelo

O VIF é uma estatística usada na análise de regressão para medir a multicolinearidade entre variáveis independentes (também conhecidas como variáveis explicativas) num modelo estatístico. A multicolinearidade ocorre quando duas ou mais variáveis independentes num modelo de regressão estão altamente correlacionadas, o que pode causar problemas na interpretação dos coeficientes de regressão e na precisão das previsões.

O VIF quantifica o grau de multicolinearidade, examinando como a variância de um coeficiente de regressão é aumentada devido à multicolinearidade. Em termos simples, quanto maior o VIF para uma variável independente específica, maior é a probabilidade de multicolinearidade.

Em geral, um VIF maior do que 5 ou 10 é considerado como um sinal de multicolinearidade problemática, o que pode exigir a remoção de variáveis do modelo. O objetivo é ter VIFs baixos para todas as variáveis independentes, o que indica que elas contribuem de forma única e não estão altamente correlacionadas entre si.

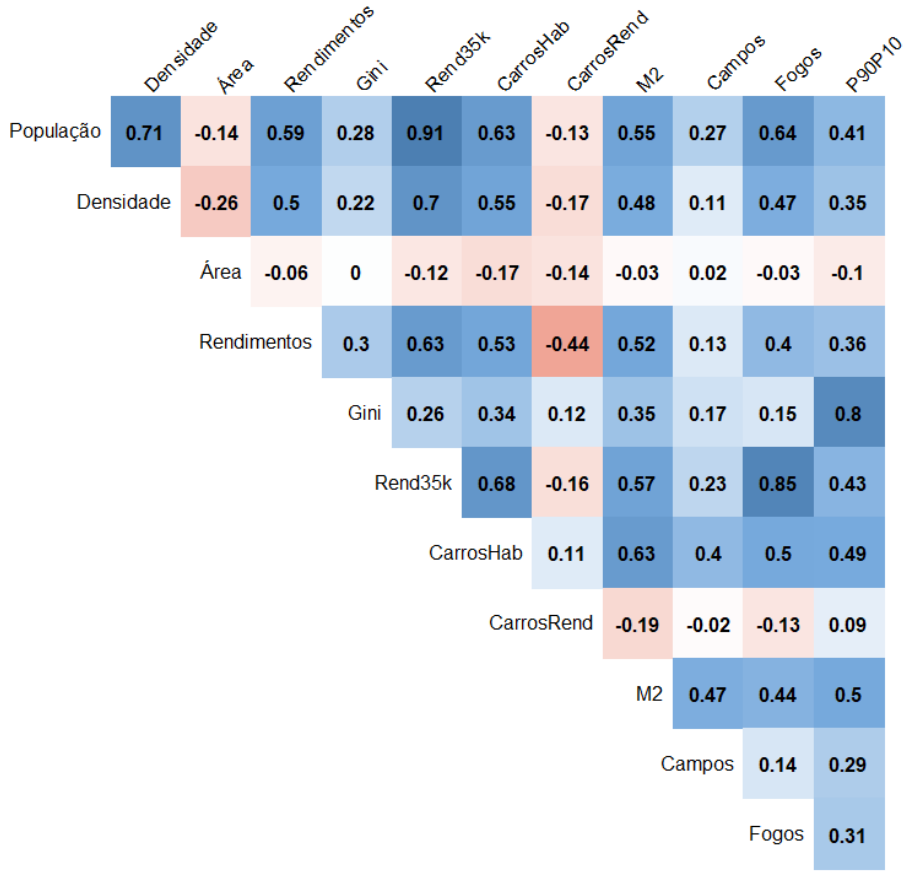


Figura 5.1: Correlação entre as variáveis explicativas

População	Densidade	Área	Rendimentos	Gini	CarrosHab
3.24	2.41	1.19	2.65	2.98	3.08

CarrosRend	M2	Campos	Fogos	P90P10
1.87	2.41	2.12	1.87	3.54

Tabela 5.1: VIF do Modelo Inicial

Os valores do VIF para todas as variáveis explicativas são menores que 5, indicando que a multicolinearidade entre essas variáveis não é um problema significativo. Portanto, podemos afirmar que essas variáveis explicativas podem ser usadas no modelo de regressão linear sem preocupações de multicolinearidade.

Depois dos VIF calculados, passamos à criação do novo modelo para cada estado.

Análise do Modelo Final

O modelo criado, depois de retirar a variável ***Rend35k***, usado no R é:

$$Valor \sim Popula\tilde{c}ao + Densidade + \acute{A}rea + Rendimentos + Gini + CarrosHab + CarrosRend + Campos + P90P10 + M2 + Fogos. \quad (5.2)$$

Depois do modelo criado, e com o auxílio do R, usamos o comando *summary(modelo)*.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	5.928e+05	5.215e+05	1.137	0,2567	
População	-8.979e-01	1.282e+00	-0,700	0,4844	
Densidade	1.579e+02	7.789e+01	2.027	0,0436	*
Área	1.530e+02	1.651e+02	0.926	0,3550	
Rendimentos	-3.090e+01	2.656e+01	-1.163	0,2457	
Gini	-6.645e+05	1.134e+06	-0,586	0,5583	
CarrosHab	4.470e+04	5.577e+03	8.015	3,52e-14	***
CarrosRend	-3.209e+05	1.421e+05	-2.258	0,0248	*
M2	-2.143e+02	1.288e+02	-1.663	0,0974	.
Campos	6.454e+05	5.679e+04	11.364	<2e-16	***
Fogos	5.049e+03	8.573e+02	5.889	1,17e-08	***
P90P10	4.096e+03	5.870e+04	0.070	0,9444	

Tabela 5.2: *Summary(modelo)*

Com este comando podemos ver os resultados na tabela 5.2 sobre as estimativas para os coeficientes do modelo, o erro padrão, o valor de t (razão entre a estimativa do coeficiente e o seu erro padrão) e um valor p associado ao valor t (indica a probabilidade de observar um valor t tão extremo quanto o valor calculado, sob a hipótese nula de que o coeficiente é igual a zero). As variáveis com "*" são as que tem um *P-Value* abaixo de 0,05. Quanto menor for esse valor, mais "*" a variável tem (máximo de três).

Para obter o melhor modelo de regressão linear, realiza-se um processo de eliminação sequencial das variáveis, usando o método *backward*, com base no AIC. Isso resulta num modelo final composto pelas seguintes variáveis: (Ver Anexo A).

$$Valor \sim Densidade + Rendimentos + CarrosHab + CarrosRend + M2 + Campos + Fogos \quad (5.3)$$

Na tabela 5.3 e na equação 5.3 podemos ver o modelo final, nas quais todas as variáveis tem *P-Value* abaixo de 0.05.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	6.553e+05	4.858e+05	1.349	0.17845	
Densidade	1.170e+02	6.667e+01	1.755	0.08048	.
Rendimentos	-4.062e+01	2.479e+01	-1.638	0.10255	
CarrosHab	4.427e+04	5.532e+03	8.003	3.66e-14	***
CarrosRend	-3.768e+05	1.343e+05	-2.804	0.00541	**
M2	-2.239e+02	1.239e+02	-1.807	0.07187	.
Campos	6.442e+05	5.524e+04	11.662	< 2e-16	***
Fogos	4.912e+03	7.631e+02	6.437	5.57e-10	***

Tabela 5.3: *Summary(modelo)* Final

O R^2 (coeficiente de determinação múltipla) e o R^2 ajustado também foram calculados. O R^2 tem um valor de 0.6777 enquanto que o R^2 ajustado tem um valor de 0.6718. Aproximadamente 68% da variabilidade do modelo foi explicada pelas variáveis independentes.

Depois das variáveis estarem escolhidas, é importante perceber se o modelo segue todos os outros pressupostos para um modelo linear. Com o recurso ao *software* R, é possível criar um conjunto de gráficos *Residuals vs Fitted*, *Normal Q-Q*, *Scale-Location* e *Residuals vs Leverage*, a partir do comando *plot(modelo)*, em que o modelo é dado em 5.3, com o objetivo de analisar os resíduos.

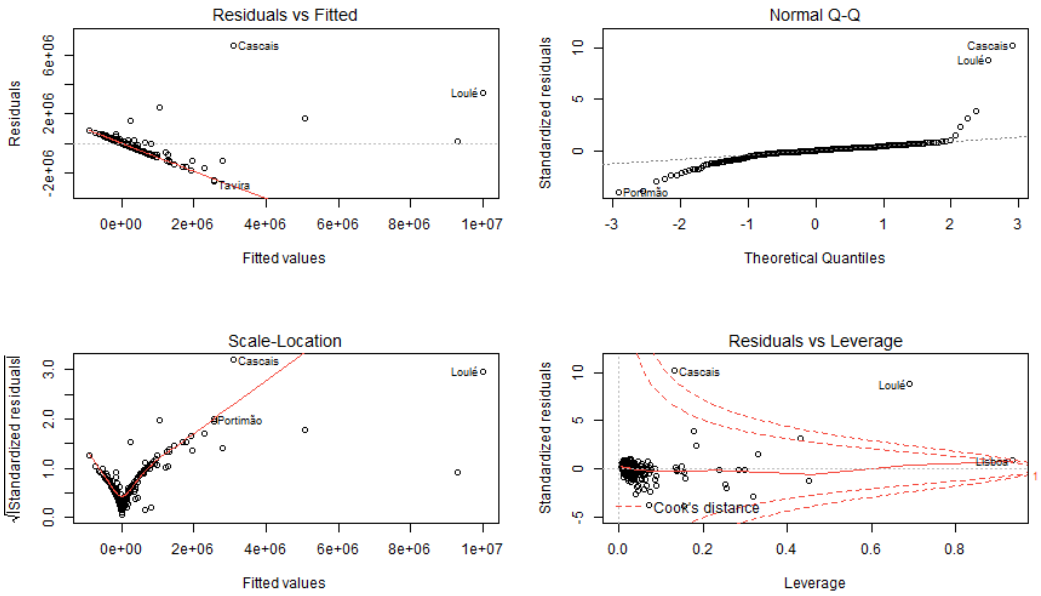


Figura 5.2: Análise de Resíduos

Na figura 5.2 é possível visualizar os quatros gráficos anteriormente mencionados. Relativamente ao gráfico *Residuals vs Fitted* podemos ver que os valores não se

encontram em torno do 0, de forma aleatória, com média igual a zero. A homocedasticidade não foi atendida.

Em relação ao gráfico *Normal Q-Q* podemos ver que os pontos não seguem uma reta na diagonal, e as caudas apresentam muitos desvios daquela que seria a reta. Conclui-se assim que os resíduos não seguem uma distribuição normal.

No gráfico *Scale-Location*, era de esperar ver os pontos em torno de uma linha horizontal. Mais uma vez, não acontece.

Por último, o gráfico *Residuals vs Leverage* ajuda-nos a ver que pontos surgem como pontos influentes e pontos alavanca. Cascais e Loulé surgem como pontos influentes e Lisboa como ponto alavanca.

Foi também utilizado o teste de *Shapiro-Wilk* para testar a normalidade. A hipótese nula deste teste é que os resíduos seguem uma distribuição normal. O *P-Value* foi inferior a $2.2e-16$, rejeitando-se assim a hipótese nula, confirmando-se mais uma vez que o modelo não segue uma distribuição normal.

Em resumo, a utilização de modelos lineares pode não ser a melhor escolha quando estamos diante de dados que não seguem uma relação linear e não atendem à suposição de normalidade dos resíduos.

5.1.2 Modelos Baseados em Árvores

De forma a ultrapassar os problemas da regressão linear (5.1.1), a análise estatística, neste trabalho, recorre à aplicação de modelos de *Machine Learning*, especificamente os modelos de *Boosting*, *Bagging* e *Random Forest*, que foram desenvolvidos com o auxílio da linguagem de programação *R*. Nestes modelos, o uso de variáveis que estão correlacionadas não precisam necessariamente de serem excluídas. A variável *Rend35k* eliminada no modelo de regressão linear, nestes modelos de *Machine Learning* é utilizada.

Para estes métodos foram utilizadas duas bibliotecas do *Software R*, *library (gbm)* e *library (randomForest)*. O processo inicia-se com a divisão da base de dados original em dois conjuntos distintos: os dados de treino (80% dos dados escolhidos de forma aleatória) e os dados de teste (os restantes 20%), separados no *R* com os comandos:

$$\begin{aligned} &training.samples = Dados\$Valor \;\%>\% \\ &\quad createDataPartition(p = 0.8,\ list = FALSE) \\ &train.data = Dados[training.samples,] \\ &test.data = Dados[-training.samples,] \end{aligned} \tag{5.4}$$

Boosting

Os modelos criados com o método *Boosting* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$\begin{aligned} model\_boosting = gbm(Valor \sim Popula\tilde{c}ao + Densidade + \acute{A}rea + \\ Rendimentos + Gini + M2 + CarrosHab + CarrosRend + \\ Rend35k + Campos + Fogos + P90P10,\ data = train.data, \\ distribution = "gaussian",\ n.trees = 5000,\ shrinkage = 0.1, \\ interaction.depth = 3,\ n.minobsinnode = 10) \end{aligned} \tag{5.5}$$

A equação 5.5 representa o modelo *boosting*, criado com a função **gbm()**. É usada uma distribuição "*gaussian*"(***distribution = "gaussian"***) uma vez que é um problema de regressão. O argumento ***n.trees = 5000*** indica que são construidas 5000 árvores. O ***shrinkage = 0.1*** controla o tamanho das atualizações dos pesos a cada iteração do *boosting*. A opção ***interaction.depth=3*** limita a profundidade de cada árvore do modelo. O ***n.minobsinnode = 10*** define o número mínimo de observações num nó de uma árvore final.

$$predict\_boosting = model\_boosting\ \%>\%\ predict(test.data) \tag{5.6}$$

A equação 5.6 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.5) para prever a base de dados de teste. Para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_boosting, test.data$Valor)
[1]  438 998.1
```

Com *Cross-Validation*

$$\begin{aligned} model\_boosting\_cv = train(Valor \sim Popula\tilde{c}ao + Densidade + Gini + \\ Rendimentos + P90P10 + CarrosHab + CarrosRend + Rend35k + \\ Campos + \acute{A}rea + Fogos + M2,\ method = "xgbTree", \\ data = train.data,\ trControl = trainControl("cv", number = 10)) \end{aligned} \tag{5.7}$$

A equação 5.7 representa o modelo *boosting* com *cross-validation*, criado com a função **train()**. Este é um modelo com menos parâmetros em relação ao modelo em 5.5. O método utilizado é o *xgbTree* (***method = "xgbTree"***). O argumento ***trControl = trainControl("cv", number = 10)*** usa a função **trainControl()** com o argumento ***"cv"*** (para função **train()** usar *cross-validation*) e o argumento ***number = 10*** para definir o número de vezes que o modelo será treinado.

$$predict\_boosting\_cv = model\_boosting\_cv\ \%>\%\ predict(test.data) \tag{5.8}$$

A equação 5.8 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.7) para prever a base de dados de teste. Para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_boosting_cv, test.data$Valor)
[1]  83 433.67
```

Bagging

Os modelos criados com o método *Bagging* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$\begin{aligned} model\_bagging = bagging(&Valor \sim Popula\tilde{c}ão + Densidade + Área + \\ &Rendimentos + Gini + M2 + CarrosHab + CarrosRend + \\ &Rend35k + Campos + Fogos + P90P10,\ data = train.data, \\ &nbagg = 100,\ coob = TRUE,\ shrinkage = 0.1, \\ &control = rpart.control(minsplit = 2, cp = 0)) \end{aligned} \tag{5.9}$$

A equação 5.9 representa o modelo *bagging*, criado com a função **bagging()**. A novidade neste modelo são os argumentos: ***nbagg = 100***; que indica o número de modelos de decisão que são treinados e combinados usando o *bagging*; ***coob = TRUE***, que indica se o erro OOB (medida do desempenho do modelo usando os dados que não foram selecionados em cada sub amostra do *bootstrap*) será calculado; ***control = rpart.control(minsplit = 2, cp = 0)***, que especifica as configurações de controlo para o algoritmo que é usado internamente no *bagging*, sendo o parâmetro ***minsplit = 2*** responsável por definir o número mínimo de observações num nó para que este seja dividido e o parâmetro ***cp = 0*** o que controla a complexidade do modelo.

$$predict\_bagging = model\_bagging\ \%>\%\ predict(test.data) \tag{5.10}$$

A equação 5.10 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.9) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_bagging, test.data$Valor)
[1]  76 437.34
```

Com *Cross-Validation*

$$\begin{aligned} model\_bagging\_cv = train(&Valor \sim Popula\tilde{c}ao + Densidade + Gini + \\ &Rendimentos + P90P10 + CarrosHab + CarrosRend + Rend35k + \\ &Campos + \acute{A}rea + Fogos + M2,\ method = "treebag", \\ &data = train.data,\ trControl = trainControl("cv", number = 10)) \\ &nbagg = 100,\ control = rpart.control(minsplit = 2, cp = 0)) \end{aligned} \tag{5.11}$$

A equação 5.11 representa o modelo *bagging*, criado com a função **train()**. Este modelo utiliza argumentos já referenciados anteriormente.

$$predict\_bagging\_cv = model\_bagging\_cv\ \%>\%\ predict(test.data) \tag{5.12}$$

A equação 5.12 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.11) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_bagging_cv, test.data$Valor)
[1]  92 666.03
```

Random Forest

O modelo criado com o método *Random Forest* foi apenas feito com *Cross-Validation*.

$$model\_rf = train(Valor \sim População + Densidade + Gini + Rendimentos + P90P10 + CarrosHab + CarrosRend + Rend35k + Campos + Área + Fogos + M2,\ data = train.data,\ method = "rf",\ trControl = trainControl("cv", number = 10)) \tag{5.13}$$

A equação 5.13 representa o modelo *random forest*, criado com a função **train()**. A única diferença nos argumentos é o ***method = "rf"***, que utiliza o método *random forest*.

$$predict\_rf = model\_rf\ \%>\%\ predict(test.data) \tag{5.14}$$

A equação 5.14 cria uma variável relativa à previsão, em que usa o modelo *random forest* (5.13) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_rf, test.data$Valor)
[1] 70 477.85
```

5.1.3 Modelo Final

Depois dos modelos criados, é o momento de escolher o melhor modelo para as previsões.

Boosting		***Bagging***		***Random Forest***
Sem CV	Com CV	Sem CV	Com CV	Com Cv
438 998.1	83 433.67	76 437.34	92 666.03	70 477.85

Tabela 5.4: RMSE dos Modelos de *Machine Learning* criados

Após uma análise dos valores de RMSE constatou-se que o modelo de *random forest* exibe o menor valor em comparação aos outros modelos. Passa, então, a ser fundamental analisar a importância das variáveis do modelo final.

A tabela 5.5 fornece informações importantes sobre quais as variáveis que têm maior peso na tomada de decisões do modelo. É importante entender as características que podem ser mais relevantes ao analisar e interpretar os resultados do modelo e ao tomar decisões com base nas previsões. Variáveis com importâncias mais elevadas merecem uma atenção especial, enquanto aquelas com importâncias muito baixas podem não ser significativas para o modelo.

Neste caso, a variável *CarrosHab* tem importância de 100, isto é, tem um grande impacto nas previsões. As variáveis *Campos*, *M2* e *Gini* têm uma importância ainda

Variáveis	Overall
CarrosHab	100
Campos	40.6756
M2	26.7909
Gini	22.9463
Rend35k	4.4849
P90P10	4.4607
População	3.9418
Rendimentos	2.6284
Área	1.4590
CarrosRend	0.1855
Densidade	0.3687
Fogos	0

Tabela 5.5: Importância das Variáveis no Modelo de Random Forest

considerável, ao contrário das restantes variáveis que têm valor de importância perto do 0.

Depois da analise de importância das variáveis do modelo, passamos para a previsão das vendas, em Portugal Continental, por concelhos.

5.1.4 Discussão de Resultados

Após analisarmos a figura 5.3, identificamos que quatro municípios destacam-se com previsões de faturação mais elevadas: Loulé, Lisboa, Cascais e Porto. Além disso, notamos que outros municípios também apresentam previsões de faturação consideráveis, tais como Grândola, Sintra, Faro, Lagos e Albufeira.

5.2 Espanha

Depois de uma análise feita a Portugal, passamos para Espanha. É o país vizinho, com a segunda maior faturação, sendo esse também um dos grandes fatores para tal faturamento. A analise será feita por províncias, sendo a Comunidade de Madrid e as Ilhas Baleares os locais com maior faturação.

As variáveis utilizadas para este estudo são:

- ***M2*** - Preço do Metro Quadrado para Construção;
- ***Campos*** - Número de Campos de Golfe;
- ***Área*** - Área em Quilómetros;
- ***População*** - Número de Habitantes;
- ***Rendimento*** - Rendimento Médio por Pessoa;

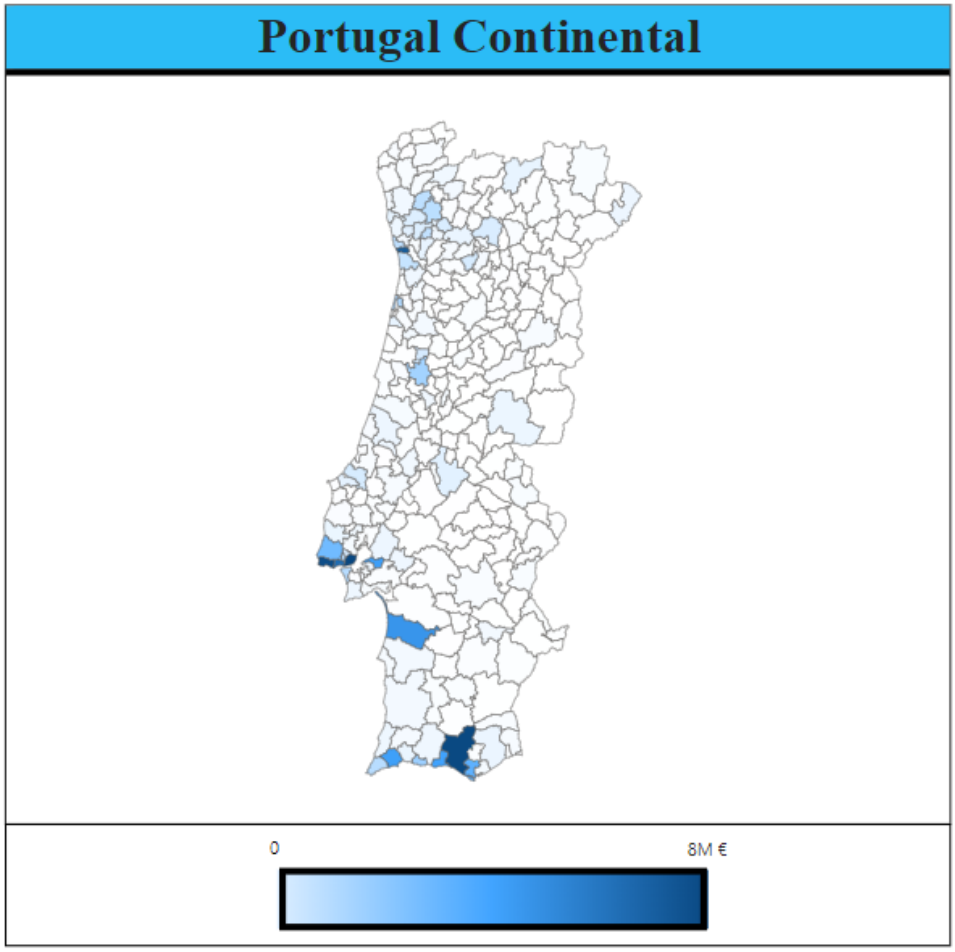


Figura 5.3: Previsões de Faturação em Portugal Continental, por concelho.

- ***Gini*** - Índice de Gini;
- ***Rend_Ricos*** - Percentagem da População com Rendimento Per Capita Superior a 200% da Mediana;
- ***Carros*** - Número de Carros Vendidos;
- ***CarrosHab*** - Número de Carros por Habitante;
- ***PIB_PC*** - PIB Per Capita;
- ***Carros_M_PIB*** - Número de Carros Vendidos por Milhão de PIB;

As variáveis foram coletadas do INE (Instituo Nacional de Estatística), Idealista, *Statista* e *DadosMundias.com*.

5.2.1 Regressão Linear

O modelo inicial, com todas as variáveis, usado no R é:

$$\begin{aligned} Valor \sim M2 + Campos + Área + População + \\ Rendimento + Gini + Rend\_Ricos + PIB\_PC + \\ Carros\_Hab + Carros + Carros\_M\_PIB \end{aligned} \tag{5.15}$$

Correlação das Variáveis

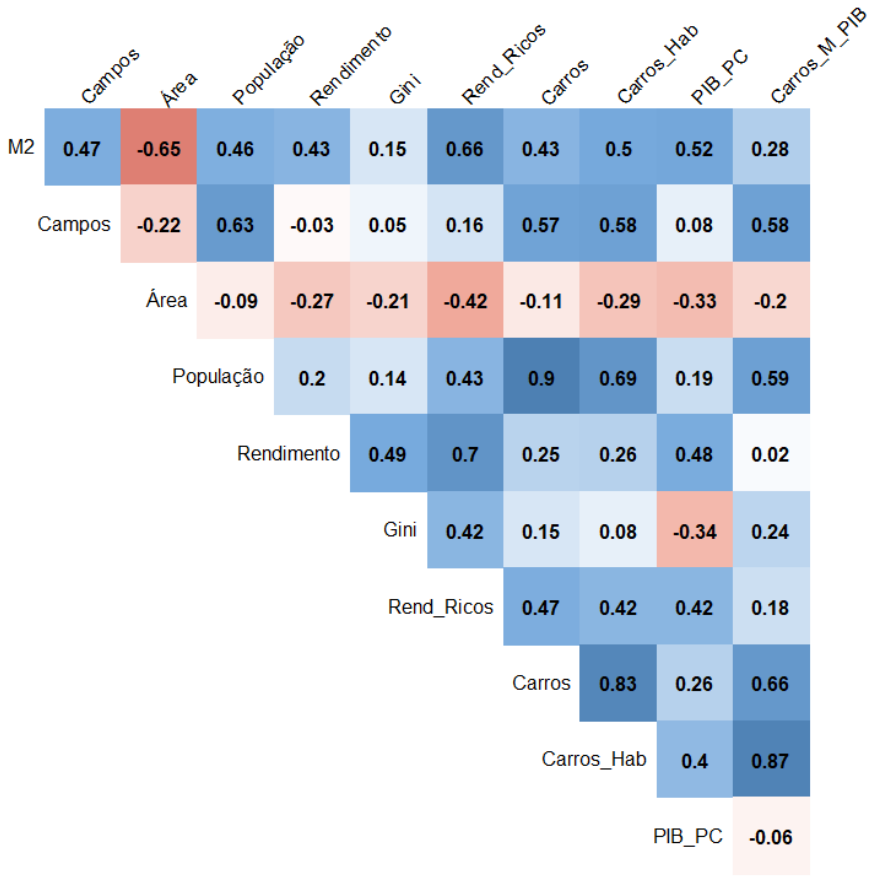


Figura 5.4: Correlação entre as variáveis explicativas

A figura 5.4 é relativa às correlações do modelo para Espanha. Definiu-se o valor $\pm 0,80$ como valor de referência, eliminando todos os valores, com valor absoluto, maior que $0,80$. As variáveis removidas foram ***Carros*** e ***Carros_M_PIB***.

VIF (*Variance Inflation Factor*) do Modelo

M2	Campos	Área	População	Rendimento
5.225815	3.165217	2.198980	2.569989	4.426096

Gini	Rend_Ricos	PIB_PC	Carros_Hab
4.226202	3.750527	5.126112	2.873971

Tabela 5.6: VIF do Modelo Inicial

Os valores do VIF, para todas as variáveis explicativas, são menores que 10, indicando que a multicolinearidade entre essas variáveis não é um problema significativo. Portanto, podemos afirmar que essas variáveis explicativas podem ser usadas no modelo de regressão linear, sem preocupações de multicolinearidade.

Depois dos VIF calculados, passamos à criação do novo modelo para cada estado.

Análise do Modelo Final

O modelo criado, depois de retirar as variáveis ***Campos***, ***Carros***, ***Carros_M_PIB***, usado no R é:

$$\begin{aligned} Valor \sim M2 + \text{Área} + População + \\ Rendimento + Rend\_Ricos + PIB\_PC + \\ Carros\_Hab + Gini \end{aligned} \tag{5.16}$$

Depois do modelo criado, e com o auxílio do R, usamos o comando *summary(modelo)*.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	-4.423e+06	3.524e+06	-1.255	0.2163	
M2	1.307e+03	1.076e+03	1.214	0.2315	
Área	1.454e+02	6.647e+01	2.188	0.0343	*
População	-1.886e-01	3.164e-01	-0.596	0.5542	
Rendimento	2.635e+01	2.120e+02	0.124	0.9017	
Gini	-4.464e+04	8.894e+04	-0.502	0.6183	
Rend_Ricos	4.713e+04	3.629e+04	1.299	0.2012	
PIB_PC	-1.307e+02	1.135e+02	-1.151	0.2563	
Carros_Hab	4.154e+08	7.714e+07	5.385	3.02e-06	***

Tabela 5.7: *Summary(modelo)*

Para obter o melhor modelo de regressão linear, realiza-se um processo de eliminação sequencial das variáveis, usando o método *backward*, com base no AIC. Isso

resulta num modelo final composto pelas seguintes variáveis: (Ver Anexo B)

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-5.919e+06	1.525e+06	-3.881 0.000324	***
M2	2.054e+03	7.055e+02	2.911 0.005486	**
Area	1.399e+02	5.742e+01	2.437 0.018627	*
PIB_PC	-9.370e+01	5.798e+01	-1.616 0.112796	
Carros_Hab	4.175e+08	5.188e+07	8.046 2.18e-10	***

Tabela 5.8: *Summary(modelo)* Final

$$Valor \sim M2 + Area + PIB\_PC + Carros\_Hab \tag{5.17}$$

Na tabela 5.8 e na equação 5.17 podemos ver o modelo final, nas quais todas as variáveis tem *P-Value* abaixo de 0.05.

O R^2 (coeficiente de determinação múltipla) e o R^2 ajustado também foram calculados. O R^2 tem um valor de 0.6777 enquanto que o R^2 ajustado tem um valor de 0.6718. Aproximadamente 68% da variabilidade do modelo foi explicada pelas variáveis independentes.

Depois das variáveis estarem escolhidas, é importante perceber se o modelo segue todos os outros pressupostos para um modelo linear.

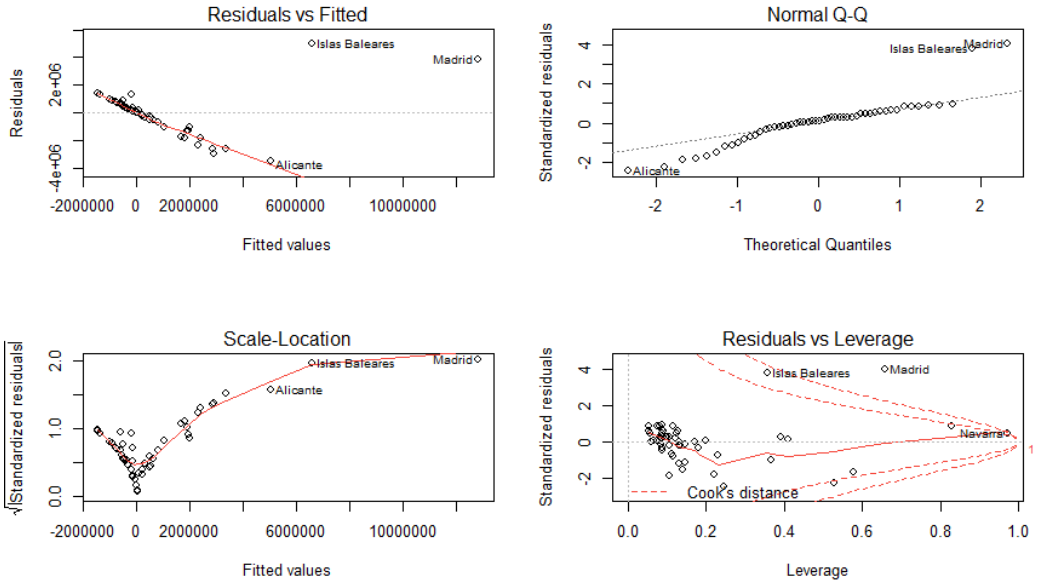


Figura 5.5: Análise de Resíduos

Na figura 5.5 é possível visualizar os quatros gráficos da análise de resíduos, já mencionados anteriormente.

Relativamente ao gráfico *Residuals vs Fitted* podemos ver que os valores não se encontram em torno do 0, de forma aleatória, com média igual a zero. A homocedasticidade não foi atendida.

Em relação ao gráfico *Normal Q-Q* podemos ver que os pontos não seguem uma reta na diagonal, e as caudas apresentam muitos desvios daquela que seria a reta. Conclui-se, assim, que os resíduos não seguem uma distribuição normal.

No gráfico *Scale-Location* era de esperar ver os pontos em torno de uma linha horizontal. Mais uma vez, não acontece.

Por último, o gráfico *Scale-Location* ajuda-nos a ver que pontos surgem como pontos influentes e pontos alavanca. Madrid e as Islas Baleares surgem como pontos influentes e Navarra como ponto alavanca.

Foi também utilizado o teste de *Shapiro-Wilk* para testar a normalidade. Neste caso, o *P-Value* foi inferior a 0.003143, rejeitando-se assim a hipótese nula, confirmando-se mais uma vez que o modelo não segue uma distribuição normal.

5.2.2 Modelos Baseados em Árvores

Para Espanha, tal como em Portugal, recorreu-se à aplicação de modelos de *Machine Learning*, especificamente os modelos de *Boosting*, *Bagging* e *Random Forest*, que foram desenvolvidos com o auxílio da linguagem de programação *R*.

O código utilizado na linguagem *R*, para a construção destes modelos, mantém-se igual aos modelos vistos anteriormente. As únicas diferenças vistas são as variáveis explicativas utilizadas, uma vez que a variável resposta mantém-se igual.

O processo inicia-se com a divisão da base de dados original em dois conjuntos distintos: os dados de treino (80% dos dados escolhidos de forma aleatória) e os dados de teste (os restantes 20%), separados no R com os comandos:

$$\begin{aligned} &training.samples = Dados\$Valor \ \%>\% \\ &\quad createDataPartition(p = 0.8,\ list = FALSE) \\ &train.data = Dados[training.samples,] \\ &test.data = Dados[-training.samples,] \end{aligned} \tag{5.18}$$

Boosting

Os modelos criados com o método *Boosting* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$model\_boosting = gbm(Valor \sim M2 + Campos + \text{Área} + População + Rendimento + Gini + Rend\_Ricos + PIB\_PC + Carros\_Hab + Carros + Carros\_M\_PIB,\ data = train.data,\ distribution = "gaussian",\ n.trees = 5000,\ shrinkage = 0.1,\ interaction.depth = 3,\ n.minobsinnode = 10) \tag{5.19}$$

A equação 5.19 representa o modelo *boosting*, criado com a função **gbm()**. Os parâmetros utilizados são iguais ao modelo criado para Portugal (5.5).

$$predict\_boosting = model\_boosting\ \%>\%\ predict(test.data) \tag{5.20}$$

A equação 5.20 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.19) para prever a base de dados de teste. Para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_boosting, test.data$Valor)
[1]  3 942 295
```

Com *Cross-Validation*

$$model\_boosting\_cv = train(Valor \sim M2 + Campos + \text{Área} + População + Rendimento + Gini + Rend\_Ricos + PIB\_PC + Carros\_Hab + Carros + Carros\_M\_PIB,\ method = "xgbTree",\ data = train.data,\ trControl = trainControl("cv", number = 10)) \tag{5.21}$$

A equação 5.21 representa o modelo *boosting* com *cross-validation*, criado com a função **train()**. Os parâmetros deste modelo mantém-se iguais ao modelo para Portugal (5.7), uma vez que, dentro deste modelo, estes parâmetros resultam no menor RMSE.

$$predict\_boosting\_cv = model\_boosting\_cv\ \%>\%\ predict(test.data) \tag{5.22}$$

A equação 5.22 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.21) para prever a base de dados de teste. Para avaliarmos a sua eficácia,

calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_boosting_cv, test.data$Valor)
[1]  452 029.7
```

Bagging

Os modelos criados com o método *Bagging* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$\begin{aligned} model\_bagging = bagging(&Valor \sim M2 + Campos + \\ &Área + População + Rendimento + Gini + Rend\_Ricos + \\ &PIB\_PC + Carros\_Hab + Carros + Carros\_M\_PIB, \\ &data = train.data,\ nbagg = 100,\ coob = TRUE, \\ &shrinkage = 0.1,\ control = rpart.control(minsplit = 2, cp = 0)) \end{aligned} \tag{5.23}$$

A equação 5.23 representa o modelo *bagging*, criado com a função **bagging()**. Novamente o modelo criado mantém-se com os mesmo parâmetros.

$$predict\_bagging = model\_bagging\ \%>\%\ predict(test.data) \tag{5.24}$$

A equação 5.24 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.23) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_bagging, test.data$Valor)
[1]  508 272.5
```

Com *Cross-Validation*

$$model\_bagging\_cv = train(Valor \sim M2 + Campos + \text{Área} + PIB\_PC + População + Rendimento + Gini + Rend\_Ricos + Carros\_Hab + Carros + Carros\_M\_PIB,\ method = "treebag",\ data = train.data,\ trControl = trainControl("cv", number = 10))\ nbagg = 100,\ control = rpart.control(minsplit = 2, cp = 0)) \tag{5.25}$$

A equação 5.25 representa o modelo *bagging*, criado com a função **train()**. Este modelo volta a ter os parâmetros iguais, à exceção do parâmetro ***nbagg***, que tem o valor de 100, em vez de 200, como referido no modelo 5.11.

$$predict\_bagging\_cv = model\_bagging\_cv\ \%>\%\ predict(test.data) \tag{5.26}$$

A equação 5.26 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.25) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_bagging_cv, test.data$Valor)
[1] 321 048.3
```

Random Forest

O modelo criado com o método *Random Forest* foi apenas feito com *Cross-Validation*.

$$model\_rf = train(Valor \sim M2 + Campos + \text{Área} + PIB\_PC + População + Rendimento + Gini + Rend\_Ricos + Carros\_Hab + Carros + Carros\_M\_PIB,\ data = train.data,\ method = "rf",\ trControl = trainControl("cv", number = 10)) \tag{5.27}$$

A equação 5.27 representa o modelo *random forest*, criado com a função **train()**. Os parâmetros do modelo mantém-se iguais aos do modelo de Portugal.

$$predict\_rf = model\_rf\ \%>\%\ predict(test.data) \tag{5.28}$$

A equação 5.28 cria uma variável relativa à previsão, em que usa o modelo *random forest* (5.27) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

```
RMSE(predict_rf, test.data$Valor)
[1] 293 269.5
```

5.2.3 Modelo Final

Depois dos modelos criados, é o momento de escolher o melhor modelo para as previsões.

Boosting		***Bagging***		***Random Forest***
Sem CV	Com CV	Sem CV	Com CV	Com CV
3 942 295	452 029.7	508 272.5	321 048.3	293 269.5

Tabela 5.9: RMSE dos Modelos de *Machine Learning* criados

Após uma análise dos valores de RMSE constatou-se que o modelo de *random forest* exibe o menor valor em comparação aos outros modelos. Denotar que em comparação aos modelos criados para Portugal, os valores são bastante diferentes. Passa, então, a ser fundamental analisar a importância das variáveis do modelo final.

Variáveis	Overall
Carros_Hab	100
Carros_M_PIB	93.639
Carros	75.174
Campos	73.789
Rend_Ricos	65.591
População	53.183
M2	45.840
PIB_PC	13.540
Gini	9.209
Rendimento	8.737
Área	0

Tabela 5.10: Importância das Variáveis no Modelo de Random Forest

Para Espanha, a variável relativa ao número de carros por habitante tem uma importância de 100, enquanto que a variável relativa à área tem uma importância de 0.

Depois do modelo criado e feita uma analise da importância das variáveis do modelo, passamos para a previsão das vendas, em Espanha, por províncias.

5.2.4 Discussão de Resultados

Após analisarmos a figura 5.6, identificamos duas províncias que se destacam com previsões de faturação mais elevadas: Comunidade de Madrid e as Ilhas Baleares.

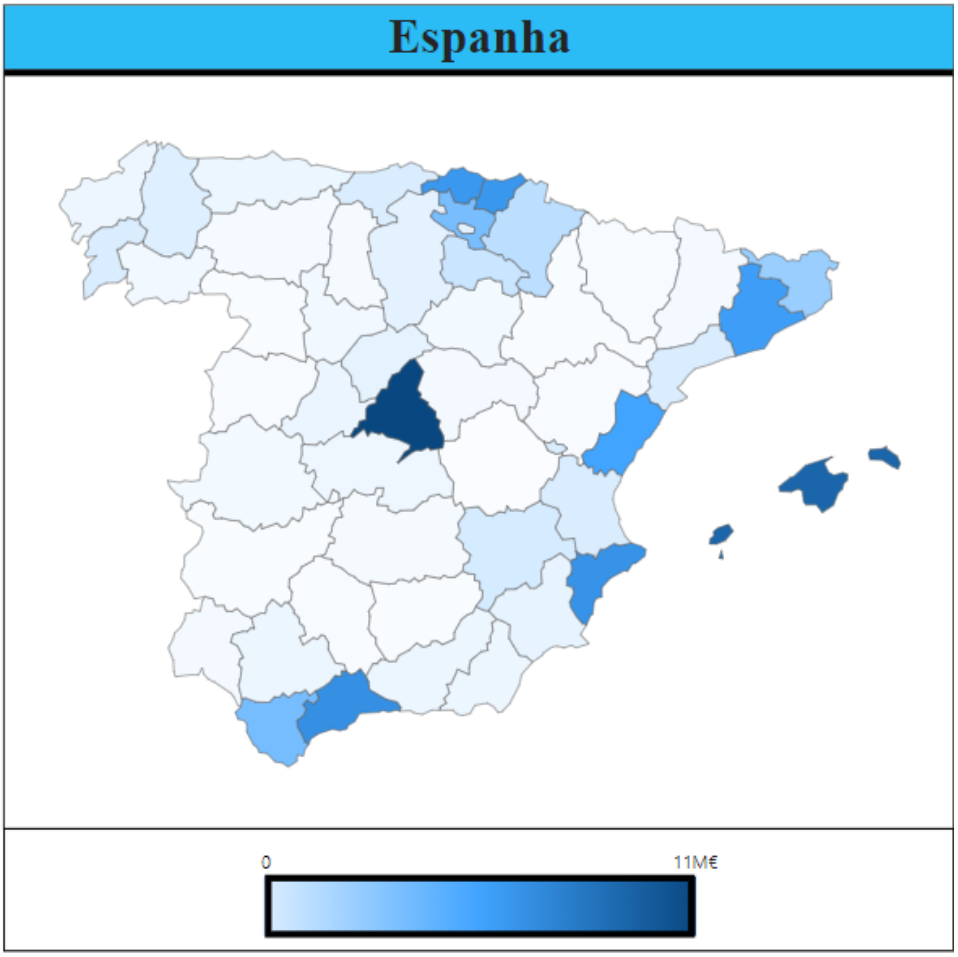


Figura 5.6: Previsões de Faturação em Espanha, por províncias.

Além disso, notamos que outras províncias também apresentam previsões de faturação consideráveis, tais como Biscaia, Guipúscoa, Barcelona, Málaga, Alicante e Castellón. Alguns destes locais suscitam o interesse da empresa, uma vez que não são locais com muitas vendas.

Ao examinarmos a tabela 5.10, fica evidente que a variável mais significativa para essa previsão é o número de carros por habitante. Isto sugere que, nestes locais, esta variável possui valores mais elevados. Esta conclusão baseia-se na combinação dos resultados da Tabela 5.8, referente ao modelo criado com regressão linear, com a importância da variável, uma vez que o coeficiente da variável *CarrosHab* é positivo.

5.3 Estados Unidos

Depois de uma análise feita a Portugal e Espanha, os países com maior faturação, passamos para os Estados Unidos. É o terceiro país com maior faturação e o país com o maior potencial, tanto pela sua riqueza mas também pela sua dimensão.

Sendo o país constituído por cinquenta estados, a análise será feita apenas para três estados: Califórnia, Nova Iorque e Flórida. Dentro de cada estado, será analisado os valores por *Counties*(condados). A escolha destes estados foi feita pela empresa. Acredita-se que estes três estados possam ser os estados com maior potencial.

As variáveis utilizadas para este estudo são:

- ***Total*** - Número de Agregados Familiares;
- ***Perc200*** - Percentagem de Agregados Familiares que ganha acima de 200 000$ por ano;
- ***Num200*** - Número de Agregados Familiares que Ganha acima de 200 000$ por Ano;
- ***MeanSal*** - Média do Salário Anual;
- ***DoisM*** - Número de Casas com Valor acima de 2 Milhões de Dólares;
- ***Car4__1P*** - Número de Casas com 1 Morador que tem 4 ou Mais Carros;
- ***Car4__2P*** - Número de Casas com 2 Moradores que tem 4 ou Mais Carros;
- ***CasaSazonal*** - Número de Casas usadas para Férias;
- ***Perc9Rooms*** - Percentagem de Casas com Mais de 9 Compartimentos;
- ***Num9Rooms*** - Número de Casas com Mais de 9 Compartimentos;
- ***Perc5Bedrooms*** - Percentagem de Casas com Mais de 5 Quartos;
- ***Num5Bedrooms*** - Número de Casas com Mais de 5 Quartos;

As variáveis foram coletadas do Departamento dos Censos dos Estados Unidos (*United States Census Bureau*).

5.3.1 Regressão Linear

O modelo inicial, com todas as variáveis, para os três estados, usado no R é:

$$\begin{aligned} Valor \sim Total + Perc200 + Num200 + DoisM + MeanSal + \\ Car4_1P + Car4_2P + CasaSazonal + Perc9Rooms + \\ Num9Rooms + Perc5Bedrooms + Num5Bedrooms \end{aligned} \tag{5.29}$$

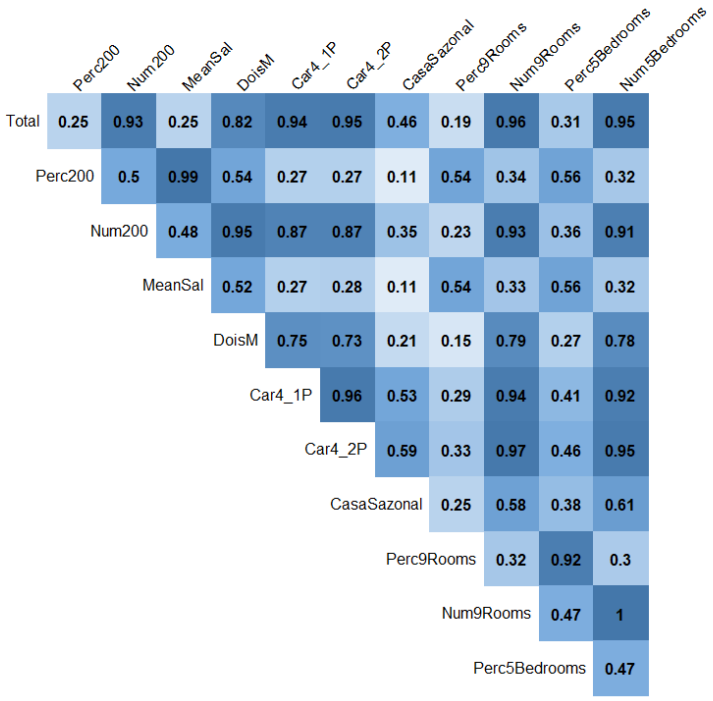


Figura 5.7: Correlação entre as Variáveis Explicativas no Estado da Califórnia

Correlação das Variáveis

A figura 5.7 é relativa às correlações do modelo para a Califórnia. Definiu-se o valor $\pm 0,80$ como valor de referência, eliminando todos os valores, com valor absoluto maior que $0,80$. As variáveis removidas foram ***Total***, ***Num200***, ***MeanSal***, ***Car4__1P***, ***Car4__2P***, ***Perc9Rooms*** e ***Num9Rooms***.

A figura 5.8 é relativa às correlações do modelo para a Flórida. Definiu-se o valor $\pm 0,80$ como valor de referência, eliminando todos os valores, com valor absoluto maior que $0,80$. As variáveis removidas foram ***Total***, ***Num200***, ***MeanSal***, ***Perc9Rooms*** e ***Num9Rooms***.

A figura 5.9 é relativa às correlações do modelo para a Nova Iorque. Definiu-se o valor $\pm 0,80$ como valor de referência, eliminando todos os valores, com valor absoluto maior que $0,80$. As variáveis removidas foram ***Num200***, ***MeanSal***, ***Perc9Rooms***, ***Num9Rooms*** e ***Num5Bedrooms***.

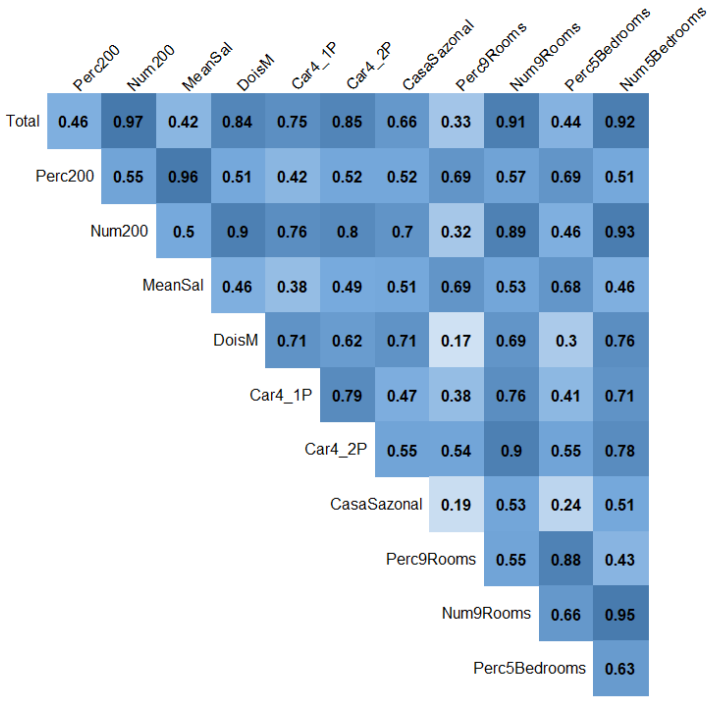


Figura 5.8: Correlação entre as Variáveis Explicativas no Estado da Flórida

VIF (*Variance Inflation Factor*) do Modelo

Perc200	DoisM	CasaSazonal	Perc5Bedrooms	Num5Bedrooms
2.81	6.77	2.43	2.46	8.51

Tabela 5.11: VIF do Modelo Inicial da Califórnia

Perc200	DoisM	Car4_1P	Car4_2P	CasaSazonal
3.71	3.05	4.48	5.52	1.83

Perc5Bedrooms	Num5Bedrooms
2.24	6.01

Tabela 5.12: VIF do Modelo Inicial da Flórida

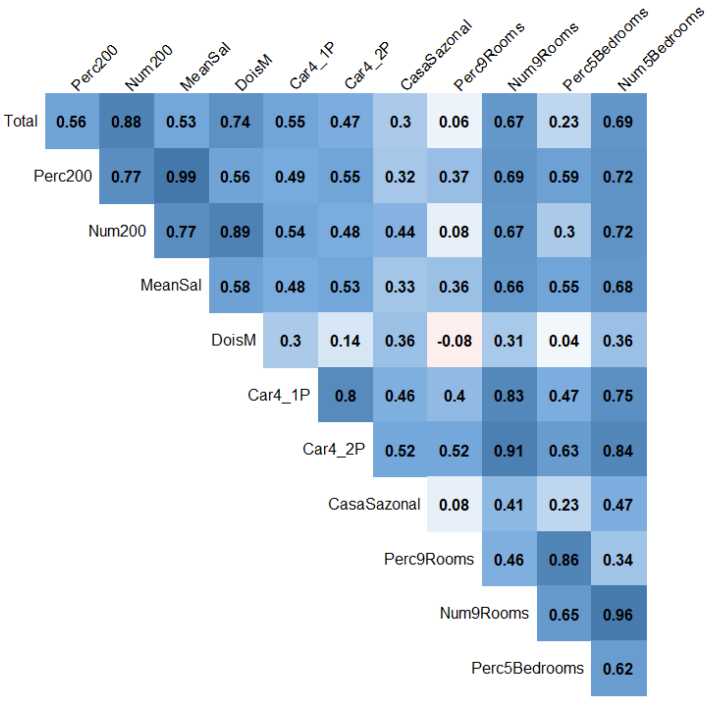


Figura 5.9: Correlação entre as Variáveis Explicativas no Estado de Nova Iorque

Total	Perc200	DoisM	Car4_1P	Car4_2P	CasaSazonal
3.71	3.05	4.48	3.26	5.52	1.8

Perc5Bedrooms
2.24

Tabela 5.13: VIF do Modelo Inicial de Nova Iorque

Os valores do VIF para todas as variáveis explicativas, nos três estados apresentados, são menores que 10, indicando que a multicolinearidade entre essas variáveis não é um problema significativo. Portanto, podemos afirmar que essas variáveis explicativas podem ser usadas no modelo de regressão linear sem preocupações de multicolinearidade.

Depois dos VIF calculados, passamos à criação do novo modelo para cada estado.

Análise do Modelo Final para a Califórnia

O modelo criado para a Califórnia, depois de retirar as variáveis ***Total***, ***Num200***, ***MeanSal***, ***Car4_1P***, ***Car4_2P***, ***Perc9Rooms*** e ***Num9Rooms***, usado no R é:

$$Valor \sim Perc200 + DoisM + CasaSazonal + Perc5Bedrooms + Num5Bedrooms \tag{5.30}$$

Depois do modelo criado, e com o auxílio do R, usamos o comando *summary(modelo)*.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	1.151e+05	4.874e+04	2.362	0.0220	*
Perc200	-1.341e+06	5.509e+05	-2.434	0.0184	*
DoisM	1.494e+01	3.315e+00	4.506	3.78e-05	***
CasaSazonal	-2.811e+00	4.976e+00	-0.565	0.5746	
Perc5Bedrooms	-1.981e+06	1.689e+06	-1.172	0.2463	
Num5Bedrooms	7.736e+00	3.445e+00	2.246	0.0290	*

Tabela 5.14: *Summary(modelo)* para a Califórnia

Para obter o melhor modelo de regressão linear, realiza-se um processo de eliminação sequencial das variáveis, usando o método *backward*, com base no AIC. Isso resulta num modelo final composto pelas seguintes variáveis: (Ver Anexo C)

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	2.877e+04	3.423e+04	0.841	0.4044	
DoisM	8.274e+00	1.681e+00	4.922	8.75e-06	***
Car4_2P	8.091e+01	1.051e+01	7.701	3.37e-10	***
CasaSazonal	-8.532e+00	3.593e+00	-2.374	0.0212	*
Perc5Bedrooms	-5.149e+06	1.008e+06	-5.110	4.50e-06	***

Tabela 5.15: *Summary(modelo)* Final para a Califórnia

$$Valor \sim DoisM + Car4\_2P + CasaSazonal + Perc5Bedrooms \tag{5.31}$$

Na tabela 5.15 e na equação 5.31 podemos ver o modelo final, nas quais todas as variáveis tem *P-Value* abaixo de 0.05.

O R^2 (coeficiente de determinação múltipla) e o R^2 ajustado também foram calculados. O R^2 tem um valor de 0.8087 enquanto que o R^2 ajustado tem um valor de 0.7981. Aproximadamente 81% da variabilidade do modelo foi explicada pelas variáveis independentes.

Depois das variáveis estarem escolhidas, é importante perceber se o modelo segue todos os outros pressupostos para um modelo linear.

Na figura 5.10 é possível visualizar os quatros gráficos da análise de resíduos, já mencionados anteriormente.

Relativamente ao gráfico *Residuals vs Fitted* podemos ver que os valores não se encontram em torno do 0, de forma aleatória, com média igual a zero. A homocedasticidade não foi atendida.

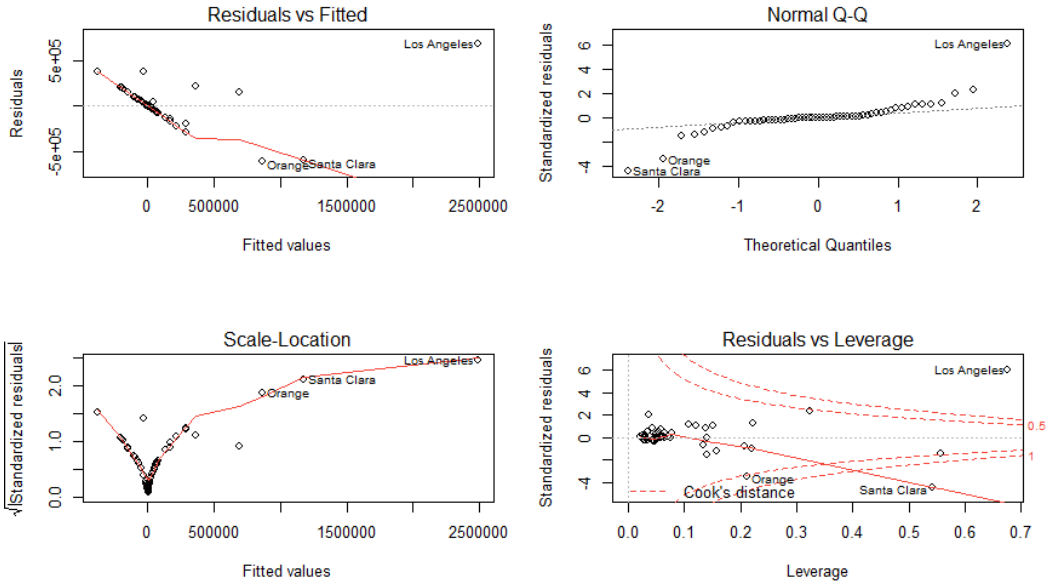


Figura 5.10: Análise de Resíduos para o Estado da Califórnia

Em relação ao gráfico *Normal Q-Q* podemos ver que os pontos não seguem uma reta na diagonal, e as caudas apresentam muitos desvios daquela que seria a reta. Conclui-se, assim, que os resíduos não seguem uma distribuição normal.

No gráfico *Scale-Location* era de esperar ver os pontos em torno de uma linha horizontal. Mais uma vez, não acontece.

Por último, o gráfico *Scale-Location* ajuda-nos a ver que pontos surgem como pontos influentes e pontos alavanca. Los Angeles e Santa Clara surgem como pontos influentes.

Foi também utilizado o teste de *Shapiro-Wilk* para testar a normalidade. Neste caso, o *P-Value* foi $6.195e-07$, rejeitando-se assim a hipótese nula, confirmando-se mais uma vez que o modelo não segue uma distribuição normal.

Análise do Modelo Final para a Flórida

O modelo criado para a Flórida, depois de retirar as variáveis ***Total***, ***Num200***, ***MeanSal***, ***Perc9Rooms*** e ***Num9Rooms***, usado no R é:

$$Valor \sim Perc200 + DoisM + Car4_1P + Car4_2P + CasaSazonal + Perc5Bedrooms + Num5Bedrooms \tag{5.32}$$

Depois do modelo criado, e com o auxílio do R, usamos o comando *summary(modelo)*.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	8.915e+03	9.064e+03	0.984	0.3294	
Perc200	-2.114e+05	2.420e+05	-0.874	0.3858	
DoisM	2.348e+01	4.762e+00	4.931	7.01e-06	***
Car4_1P	4.592e+01	2.957e+01	1.553	0.1259	
Car4_2P	-1.521e+01	8.574e+00	-1.774	0.0812	.
CasaSazonal	-1.194e+00	4.585e-01	-2.604	0.0116	*
Perc5Bedrooms	-1.110e+05	4.293e+05	-0.259	0.7969	
Num5Bedrooms	3.032e+00	1.375e+00	2.206	0.0313	*

Tabela 5.16: *Summary(modelo)* para a Flórida

Para obter o melhor modelo de regressão linear, realiza-se um processo de eliminação sequencial das variáveis, usando o método *backward*, com base no AIC. Isso resulta num modelo final composto pelas seguintes variáveis: (Ver Anexo D)

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	9.590e+03	8.612e+03	1.113	0.2700	
Perc200	-2.549e+05	1.726e+05	-1.476	0.1450	
DoisM	2.402e+01	4.258e+00	5.640	4.87e-07	***
Car4_1P	4.552e+01	2.930e+01	1.553	0.1256	
Car4_2P	-1.523e+01	8.507e+00	-1.790	0.0784	.
CasaSazonal	-1.177e+00	4.502e-01	-2.614	0.0113	*
Num5Bedrooms	2.827e+00	1.112e+00	2.542	0.0136	*

Tabela 5.17: *Summary(modelo)* Final para a Flórida

$$Valor \sim Perc200 + DoisM + Car4\_1P + Car4\_2P + CasaSazonal + Num5Bedrooms \tag{5.33}$$

Na tabela 5.17 e na equação 5.33 podemos ver o modelo final, nas quais todas as variáveis tem *P-Value* abaixo de 0.05.

O R^2 (coeficiente de determinação múltipla) e o R^2 ajustado também foram calculados. O R^2 tem um valor de 0.6912 enquanto que o R^2 ajustado tem um valor de 0.6816. Aproximadamente 69% da variabilidade do modelo foi explicada pelas variáveis independentes.

Depois das variáveis estarem escolhidas, é importante perceber se o modelo segue todos os outros pressupostos para um modelo linear.

Na figura 5.11 é possível visualizar os quatros gráficos da análise de resíduos, já mencionados anteriormente.

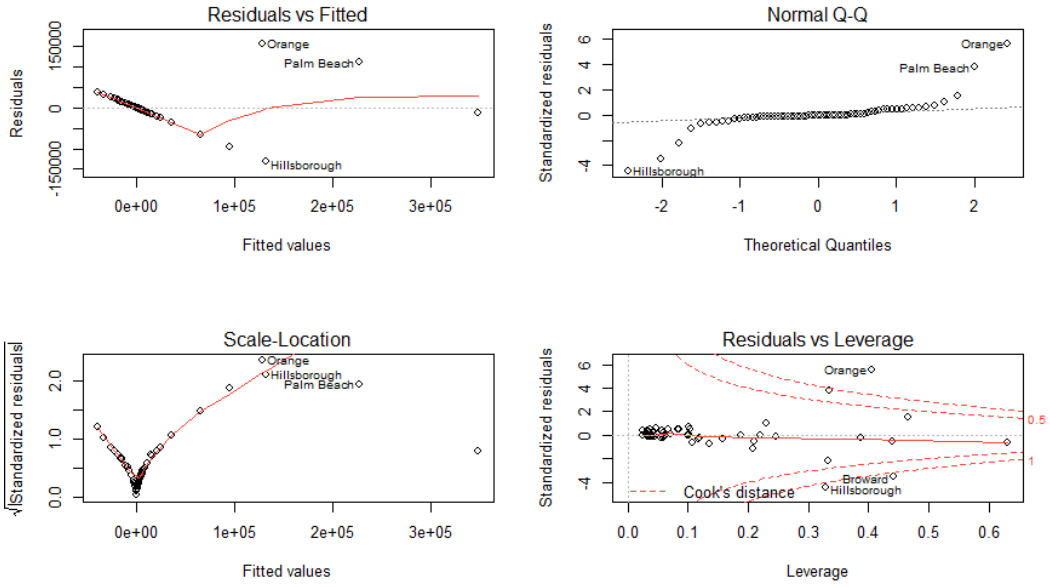


Figura 5.11: Análise de Resíduos para o Estado da Flórida

Relativamente ao gráfico *Residuals vs Fitted* podemos ver que os valores não se encontram em torno do 0, de forma aleatória, com média igual a zero. A homocedasticidade não foi atendida.

Em relação ao gráfico *Normal Q-Q* podemos ver que os pontos não seguem uma reta na diagonal, e as caudas apresentam muitos desvios daquela que seria a reta. Conclui-se, assim, que os resíduos não seguem uma distribuição normal.

No gráfico *Scale-Location* era de esperar ver os pontos em torno de uma linha horizontal. Mais uma vez, não acontece.

Por último, o gráfico *Scale-Location* ajuda-nos a ver que pontos surgem como pontos influentes e pontos alavanca. Orange, Broward e Hillsborough surgem como pontos influentes.

Foi também utilizado o teste de *Shapiro-Wilk* para testar a normalidade. Neste caso, o *P-Value* foi $1.464e-11$, rejeitando-se assim a hipótese nula, confirmando-se mais uma vez que o modelo não segue uma distribuição normal.

Análise do Modelo Final para Nova Iorque

O modelo criado para Nova Iorque, depois de retirar as variáveis ***Num200***, ***MeanSal***, ***Perc9Rooms***, ***Num9Rooms*** e ***Num5Bedrooms***, é dado por:

$$\begin{aligned} Valor \sim Total + Perc200 + Car4_2P + Car4_1P + \\ DoisM + CasaSazonal + Perc5Bedrooms \end{aligned} \tag{5.34}$$

Depois do modelo criado, e com o auxílio do R, usamos o comando *summary(modelo)*.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	-1.930e+05	4.248e+04	-4.543	3.16e-05	***
Total	-2.414e-01	1.967e-01	-1.227	0.2252	
Perc200	6.194e+05	6.421e+05	0.965	0.3390	
DoisM	6.849e+00	6.405e+00	1.069	0.2897	
Car4_1P	1.573e+02	1.567e+02	1.004	0.3199	
Car4_2P	1.162e+02	5.300e+01	2.192	0.0327	*
CasaSazonal	4.968e+01	5.479e+00	9.067	1.95e-12	***
Perc5Bedrooms	-2.128e+06	1.131e+06	-1.882	0.0652	.

Tabela 5.18: *Summary(modelo)* para Nova Iorque

Para obter o melhor modelo de regressão linear, realiza-se um processo de eliminação sequencial das variáveis, usando o método *backward*, com base no AIC. Isso resulta num modelo final composto pelas seguintes variáveis: (Ver Anexo E)

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	-1.842e+05	3.296e+04	-5.590	6.68e-07	***
DoisM	5.316e+00	3.257e+00	1.632	0.108126	
Car4_2P	1.306e+02	3.344e+01	3.906	0.000251	***
CasaSazonal	5.099e+01	5.099e+00	10.000	3.76e-14	***
Perc5Bedrooms	-1.484e+06	9.801e+05	-1.514	0.135515	

Tabela 5.19: *Summary(modelo)* Final para Nova Iorque

$$Valor \sim DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms \tag{5.35}$$

Na tabela 5.19 e na equação 5.35 podemos ver o modelo final, nas quais todas as variáveis tem *P-Value* abaixo de 0.05.

O R^2 (coeficiente de determinação múltipla) e o R^2 ajustado também foram calculados. O R^2 tem um valor de 0.8086 enquanto que o R^2 ajustado tem um valor de 0.8021. Aproximadamente 81% da variabilidade do modelo foi explicada pelas variáveis independentes.

Depois das variáveis estarem escolhidas, é importante perceber se o modelo segue todos os outros pressupostos para um modelo linear.

Na figura 5.12 é possível visualizar os quatros gráficos da análise de resíduos, já mencionados anteriormente.

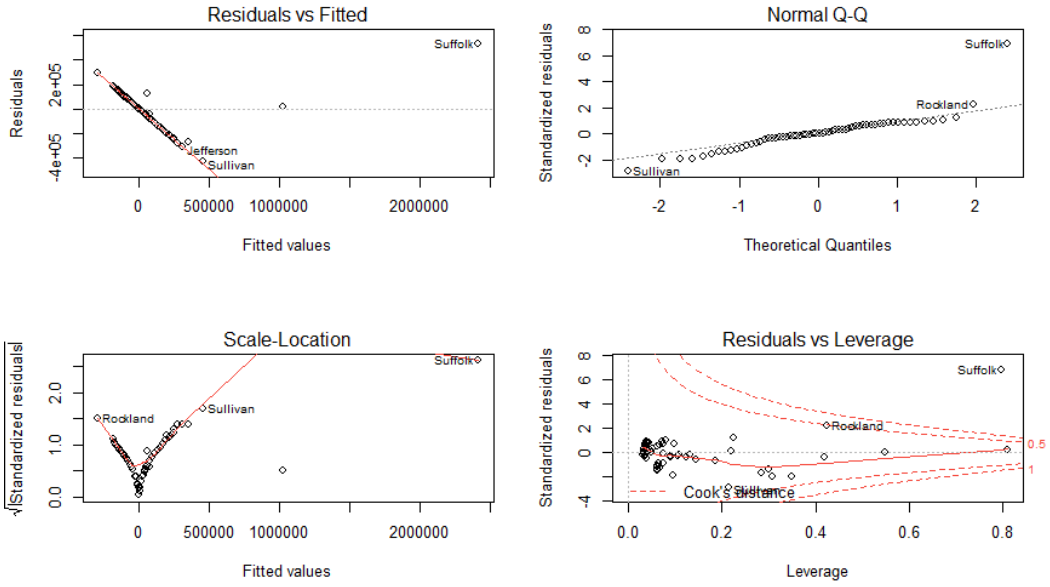


Figura 5.12: Análise de Resíduos para o Estado de Nova Iorque

Relativamente ao gráfico *Residuals vs Fitted* podemos ver que os valores não se encontram em torno do 0, de forma aleatória, com média igual a zero. A homocedasticidade não foi atendida.

Em relação ao gráfico *Normal Q-Q* podemos ver que os pontos não seguem uma reta na diagonal, e as caudas apresentam muitos desvios daquela que seria a reta. Conclui-se, assim, que os resíduos não seguem uma distribuição normal.

No gráfico *Scale-Location* era de esperar ver os pontos em torno de uma linha horizontal. Mais uma vez, não acontece.

Por último, o gráfico *Scale-Location* ajuda-nos a ver que pontos surgem como pontos influentes e pontos alavanca. Suffolk surge como ponto influente.

Foi também utilizado o teste de *Shapiro-Wilk* para testar a normalidade. Neste caso, o *P-Value* foi inferior a 0.02779, rejeitando-se assim a hipótese nula, confirmando-se mais uma vez que o modelo não segue uma distribuição normal.

Uma forma de combater estes problemas é a utilização de modelos de *Machine Learning*.

5.3.2 Modelos Baseados em Árvores

Novamente recorreu-se à aplicação de modelos de *Machine Learning*, especificamente os modelos de *Boosting*, *Bagging* e *Random Forest*, que foram desenvolvidos com o auxílio da linguagem de programação *R*.

Dado que o objetivo consiste em realizar previsões para os estados da Califórnia, Flórida e Nova Iorque, empregou-se a mesma estrutura e código, sendo a única discrepância os valores contidos na base de dados.

O processo inicia-se com a divisão da base de dados original em dois conjuntos distintos: os dados de treino (80% dos dados escolhidos de forma aleatória) e os dados de teste (os restantes 20%), separados no R com os comandos:

$$\begin{aligned} &training.samples = Dados\$Valor\ \%>\% \\ &\quad createDataPartition(p = 0.8,\ list = FALSE) \\ &train.data = Dados[training.samples,] \\ &test.data = Dados[-training.samples,] \end{aligned} \tag{5.36}$$

Boosting

Os modelos criados com o método *Boosting* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$\begin{aligned} model\_boosting = gbm(&Valor \sim Total + Perc200 + Num200 + MeanSal + \\ &DoisM + Car4_1P + Car4_2P + CasaSazonal + Perc9Rooms + \\ &Num9Rooms + Perc5Bedrooms + Num5Bedrooms, \\ &data = train.data,\ distribution = "gaussian",\ n.trees = 5000, \\ &shrinkage = 0.1,\ interaction.depth = 3,\ n.minobsinnode = 10) \end{aligned} \tag{5.37}$$

A equação 5.37 representa o modelo *boosting*, criado com a função **gbm()**. Os parâmetros mantém-se iguais ao modelos anteriormente criados.

$$predict\_boosting = model\_boosting\ \%>\%\ predict(test.data) \tag{5.38}$$

A equação 5.38 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.37) para prever a base de dados de teste. Para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

Califórnia	**Flórida**	**Nova Iorque**
159 341.5	84 411.25	1 045 112

Tabela 5.20: RMSE dos Estados da Califórnia, Flórida e Nova Iorque

Com *Cross-Validation*

$$\begin{aligned} model\_boosting\_cv = train(&Valor \sim Total + Perc200 + Num200 + MeanSal + \\ &DoisM + Car4_1P + Perc5Bedrooms + CasaSazonal + Perc9Rooms + \\ &Num9Rooms + Car4_2P + Num5Bedrooms,\ method = "xgbTree", \\ &data = train.data,\ trControl = trainControl("cv", number = 10)) \end{aligned} \tag{5.39}$$

A equação 5.39 representa o modelo *boosting* com *cross-validation*, criado com a função **train()**. O modelo mantém os mesmos parâmetros que os modelos já criados.

$$predict\_boosting\_cv = model\_boosting\_cv\ \%>\%\ predict(test.data) \tag{5.40}$$

A equação 5.40 cria uma variável relativa à previsão, em que usa o modelo *boosting* (5.39) para prever a base de dados de teste. Para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

Califórnia	**Flórida**	**Nova Iorque**
25 290.26	87 805.96	272 260.7

Tabela 5.21: RMSE dos Estados da Califórnia, Flórida e Nova Iorque

Bagging

Os modelos criados com o método *Bagging* podem ser utilizados sem *Cross-Validation* ou com *Cross-Validation*.

Sem *Cross-Validation*

$$\begin{aligned} model\_bagging = bagging(&Valor \sim Total + Perc200 + Num200 + MeanSal + \\ &DoisM + Car4_1P + Car4_2P + CasaSazonal + Perc9Rooms + \\ &Num9Rooms + Perc5Bedrooms + Num5Bedrooms, \\ &data = train.data,\ nbagg = 100,\ coob = TRUE, \\ &shrinkage = 0.1,\ control = rpart.control(minsplit = 2, cp = 0)) \end{aligned} \tag{5.41}$$

A equação 5.41 representa o modelo *bagging*, criado com a função **bagging()**. Novamente o modelo criado mantém-se com os mesmo parâmetros.

$$predict\_bagging = model\_bagging\ \%>\%\ predict(test.data) \tag{5.42}$$

A equação 5.42 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.41) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

Califórnia	**Flórida**	**Nova Iorque**
5 563.946	78 058.5	191 530.4

Tabela 5.22: RMSE dos Estados da Califórnia, Flórida e Nova Iorque

Com *Cross-Validation*

$$model\_bagging\_cv = train(Valor \sim Total + Perc9Rooms + Num200 + DoisM + MeanSal + Car4_1P + Car4_2P + CasaSazonal + Perc5Bedrooms + Num9Rooms + Perc200 + Num5Bedrooms,\ method = "treebag",\ data = train.data,\ trControl = trainControl("cv", number = 10))\ nbagg = 100,\ control = rpart.control(minsplit = 2, cp = 0)) \tag{5.43}$$

A equação 5.43 representa o modelo *bagging*, criado com a função **train()**. Este modelo volta a ter os parâmetros iguais ao modelo de Espanha.

$$predict\_bagging\_cv = model\_bagging\_cv\ \%>\%\ predict(test.data) \tag{5.44}$$

A equação 5.44 cria uma variável relativa à previsão, em que usa o modelo *bagging* (5.43) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

Califórnia	**Flórida**	**Nova Iorque**
9 004.442	76 482.32	168 953.8

Tabela 5.23: RMSE dos Estados da Califórnia, Flórida e Nova Iorque

Random Forest

O modelo criado com o método *Random Forest* foi apenas realizado com *Cross-Validation*.

$$model\_rf = train(Valor \sim Total + Perc9Rooms + Num200 + DoisM + MeanSal + Car4_1P + Car4_2P + CasaSazonal + Perc5Bedrooms + Num9Rooms + Perc200 + Num5Bedrooms,\ data = train.data,\ method = "rf",\ trControl = trainControl("cv", number = 10)) \tag{5.45}$$

A equação 5.45 representa o modelo *random forest*, criado com a função **train()**. Os parâmetros do modelo mantém-se novamente iguais.

$$predict\_rf = model\_rf\ \%>\%\ predict(test.data) \tag{5.46}$$

A equação 5.46 cria uma variável relativa à previsão, em que usa o modelo *random forest* (5.45) para prever a base de dados de teste. Novamente, para avaliarmos a sua eficácia, calculamos o RMSE (Raiz do Erro Quadrático Médio).

Califórnia	**Flórida**	**Nova Iorque**
12 926.77	65 946.28	204 910.7

Tabela 5.24: RMSE dos Estados da Califórnia, Flórida e Nova Iorque

5.3.3 Modelo Final

Depois dos modelos criados, é o momento de escolher o melhor modelo para as previsões. Foram, então, calculados os valores do RMSE de cada modelo para os três estados.

Califórnia

Boosting		**Bagging**		***Random Forest***
Sem CV	Com CV	Sem CV	Com CV	Com CV
159 341.5	25 290.26	9 004.442	5 563.946	12 926.77

Tabela 5.25: RMSE dos Modelos de *Machine Learning* criados para a Califórnia

Flórida

Boosting		**Bagging**		***Random Forest***
Sem CV	Com CV	Sem CV	Com CV	Com CV
84 411.25	87 805.96	78 058.5	76 482.32	65 946.28

Tabela 5.26: RMSE dos Modelos de *Machine Learning* criados para a Flórida

Nova Iorque

Boosting		**Bagging**		***Random Forest***
Sem CV	Com CV	Sem CV	Com CV	Com CV
1 045 112	272 260.7	191 530.4	168 953.8	53 125.66

Tabela 5.27: RMSE dos Modelos de *Machine Learning* criados para a Nova Iorque

Após uma análise dos valores de RMSE constatou-se que o modelo de *bagging* com *cross-validation* exibe o menor valor em comparação aos outros modelos, no estado da Califórnia. Relativamente ao estado da Flórida e de Nova Iorque, o modelo de *random forest* apresenta o menor RMSE. De um modo geral, o método *random forest* apresenta ser um método mais seguro, com um RMSE mais baixo em relação

à maioria. Os valores RMSE variam muito em comparação aos outros países e até entre estados. Apesar de alguns RMSE serem "baixos", os valores de faturação também são menores em comparação a Portugal e Espanha, o que acaba por ser um erro um pouco alto.

Depois desta análise, passa, então, a ser fundamental analisar a importância das variáveis do modelo final, isto é, o modelo que apresenta o menor RMSE em cada estado, antes de avançar para a previsão final.

Variáveis	Importância
DoisM	100.000
Total	93.386
Num200	87.798
Perc200	66.391
MeanSal	55.741
Car4_1P	48.056
Car4_2P	32.501
CasaSazonal	16.903
Num9Rooms	13.741
Num5Bedrooms	10.186
Perc9Rooms	4.289
Perc5Bedrooms	0.000

Tabela 5.28: Importância das Variáveis para o Modelo de *Bagging* com Validação Cruzada na Califórnia

No estado da Califórnia podemos ver um conjunto de variáveis importantes para a criação do modelo, sendo a variável *DoisM* a variável com mais importância (100).

Variáveis	Importância
DoisM	100.0000
Num200	73.5149
Num9Rooms	58.6721
Car4_2P	54.0063
Car4_1P	46.9057
Num5Bedrooms	45.5442
Total	40.4807
Perc200	3.4103
MeanSal	1.6121
CasaSazonal	1.5547
Perc5Bedrooms	0.2797
Perc9Rooms	0.0000

Tabela 5.29: Importância das Variáveis para o Modelo Random Forest na Flórida

No estado da Flórida podemos ver um conjunto de variáveis importantes para a criação do modelo, sendo a variável *DoisM*, novamente, a variável com mais importância (100).

Variáveis	Importância
CasaSazonal	100.000
Car4_2P	81.032
Car4_1P	59.255
Num5Bedrooms	58.173
Num9Rooms	41.195
Num200	12.830
MeanSal	11.709
Perc5Bedrooms	10.884
DoisM	9.371
Perc200	3.691
Total	1.560
Perc9Rooms	0.000

Tabela 5.30: Importância das Variáveis para o Modelo de Bagging com Validação Cruzada em Nova Iorque

No estado de Nova Iorque podemos ver um conjunto de variáveis importantes para a criação do modelo, sendo a variável *CasaSazonal* a variável com mais importância (100). Neste modelo, a variável *DoisM*, que anteriormente era uma variável com importância 100, tem um importância de 9.

Depois dos modelos criados e feita uma analise da importância das variáveis de cada modelo, passamos para a previsão das vendas, nos Estados Unidos, por estados da Califórnia, Flórida e Nova Iorque.

5.3.4 Discussão de Resultados

A análise nos Estados Unidos acaba por ser uma análise ainda muito inicial, com um erro considerável, uma vez que o modelo tem ainda poucas variáveis de qualidade.

A figura 5.13 requer uma análise diferente, aquando comparado aos mapas anteriormente comentados. Nesta figura podemos analisar os três estados estudados.

Relativamente ao estado da Califórnia, tem uma escala que vai de zero a três milhões, uma vez que é o estado em que o valor da faturação é mais alto. Há principalmente duas zonas em que se espera que a faturação seja mais alta, são os condados à volta de São Francisco e os condados à volta de Los Angeles. Los Angeles acaba por ser o local na Califórnia com a previsão mais alta.

Relativamente ao estado da Flórida, tem uma escala menor, que vai de zero a quinhentos mil, uma vez que, relativamente à Califórnia, a faturação foi bem inferior. Acaba por ser um estado analisado uma vez que há um potencial enorme. Há principalmente duas zonas em que se espera que a faturação seja mais alta, são os

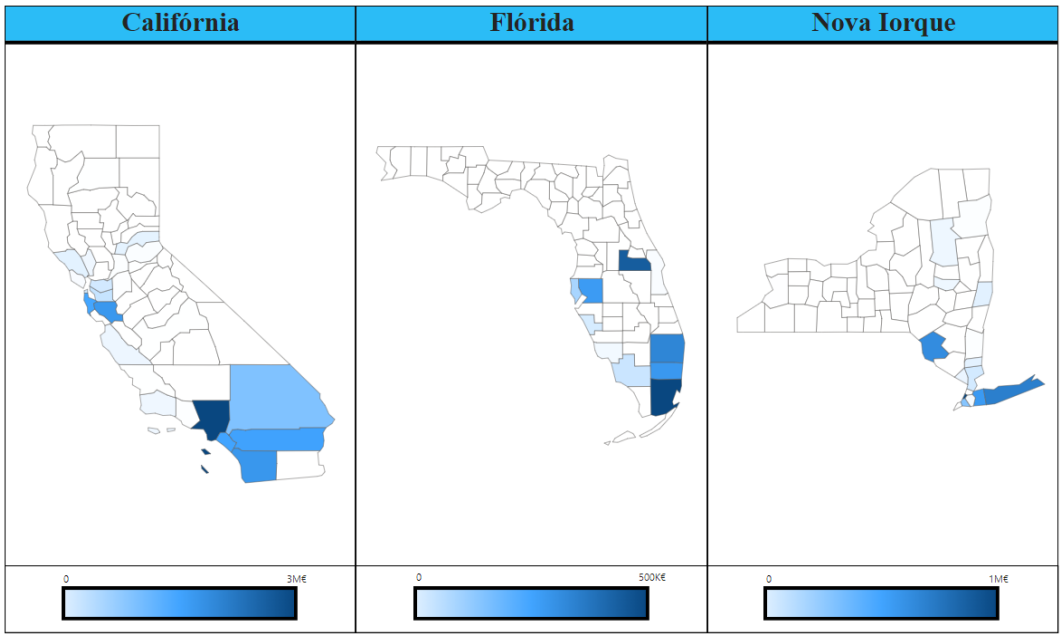


Figura 5.13: Previsões de Faturação no Estado da Califórnia, Flórida e Nova Iorque, por Condados.

condados de Miami-Dade a Palm Beach e o condado de Orange. Miami-Dade acaba por ser o local na Flórida com a previsão mais alta. Existe ainda uma previsão alta nos condados a Oeste.

Relativamente ao estado de Nova Iorque, tem uma escala que vai de zero a um milhão. Há principalmente uma grande zona em que se espera que a faturação seja mais alta, na zona de Long Island. É a zona mais rica de Nova Iorque, sendo esse um grande fator para tal faturação. Existe, ainda, o condado de Sullivan que apresenta uma previsão também interessante.

Capítulo 6

Conclusões e Trabalho Futuro

Uma análise abrangente deste estudo revela que, de acordo com as diretrizes da literatura, a área da previsão da procura é altamente dinâmica. Cada decisão tomada ao longo do projeto pode exercer um impacto substancial nos resultados. Esta dissertação concentra-se especificamente numa parte limitada da previsão, investigando-a por meio de técnicas de *Machine Learning*.

Esta dissertação concentrou-se, primordialmente, numa parte específica do processo de previsão da procura, onde foram adotadas técnicas de *Machine Learning* para estimar o potencial de faturação em contextos territoriais. O objetivo consistia em construir um modelo preditivo, cujo propósito é calcular o fluxo de receitas potenciais, tendo como base variáveis previamente identificadas com uma correlação substancial e empiricamente comprovada com o nicho de mercado em foco.

No decorrer da dissertação, revisitamos conceitos essenciais da segmentação de mercado e sublinhamos a crescente relevância do *Machine Learning* nas estratégias empresariais. Dentro do vasto campo do *Machine Learning*, concentramo-nos na exploração de técnicas como *boosting*, *bagging* e *random forest*. Paralelamente, efetuamos uma análise aprofundada dos modelos lineares múltiplos, realçando a sua possível aplicabilidade, particularmente em contextos empresariais específicos.

A empresa disponibilizou um conjunto de dados que abrange a faturação dos projetos desde 2015 até 2022, segmentada por países. Para a construção dos nossos modelos, integramos variáveis externas à organização, visando avaliar o seu impacto nas previsões de faturação. Com o auxílio da ferramenta *Power BI*, desenvolveu-se uma *dashboard* destinada a proporcionar uma representação gráfica interativa dos

dados, a qual foi posteriormente utilizada na empresa com o propósito de apresentar as informações de forma mais acessível e elucidativa.

As análises foram feitas a Portugal, Espanha e Estados Unidos. Portugal é o local onde a empresa está estabelecida e onde está concentrada a maioria das operações. Considera-se Portugal como a base sólida para realizar e expandir os projetos. Quanto a Espanha, país vizinho, a proximidade geográfica cria uma sinergia natural que permite estender o sucesso de Portugal para esse mercado. Além disso, Espanha é o segundo país em termos de faturação, o que aumenta ainda mais o seu apelo como uma fonte significativa de oportunidades de negócios. Por fim, a escolha de incluir os Estados Unidos na análise baseia-se na ideia de ser o mercado com o maior potencial de crescimento.

Em Portugal, apesar do RMSE do modelo, as previsões alinham-se com as expectativas da empresa. Isto leva à conclusão de que o modelo foi adequadamente ajustado.

No caso de Espanha, o RMSE apresenta um valor mais elevado em comparação com o modelo de Portugal. No entanto, o modelo revela resultados de interesse que, mais uma vez, se alinham com as expectativas da empresa.

Quanto aos três estados dos Estados Unidos, a interpretação do RMSE associado deve ser distinta. Em comparação com Portugal e Espanha, a faturação e números de vendas nesses estados são significativamente mais baixos. Portanto, é plausível que o modelo não seja capaz de oferecer previsões altamente precisas nesse contexto.

Relativamente aos objetivos centrais, verificou-se que, de um modo geral, os modelos criados ostentam uma boa capacidade para explicar as variações na faturação. Contudo, a necessidade de refinação é indiscutível. A expansão das variáveis consideradas, incorporando fatores adicionais de relevância, surge como um imperativo. Da mesma forma, a ampliação do volume de dados de vendas disponíveis representa uma etapa crucial para fortalecer a capacidade preditiva dos modelos.

Em suma, esta investigação estabelece as bases para futuras pesquisas, com o intuito de aperfeiçoar a compreensão e a precisão das previsões de faturação, mediante a incorporação de variáveis suplementares e a adoção de abordagens mais avançadas. Este desenvolvimento contribuirá de forma substancial para o progresso contínuo do campo de análise de dados e previsões financeiras.

Referências

[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. [Citado na página 4]

[2] Wendell R Smith. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1):3–8, 1956. [Citado na página 5]

[3] Sally Dibb. Market segmentation: strategies for success. *Marketing Intelligence & Planning*, 16(7):394–406, 1998. [Citado nas páginas 5, 8 e 9]

[4] Gary L Lilien and Arvind Rangaswamy. *Marketing engineering: computer-assisted marketing analysis and planning*. DecisionPro, 2004. [Citado na página 6]

[5] Yoram Wind. Issues and advances in segmentation research. *Journal of marketing research*, 15(3):317–337, 1978. [Citado na página 6]

[6] Paul E Green. A new approach to market segmentation. *Business Horizons*, 20(1):61–73, 1977. [Citado na página 6]

[7] Vijay Mahajan and Arun K Jain. An approach to normative segmentation. *Journal of Marketing Research*, 15(3):338–345, 1978. [Citado na página 6]

[8] Tony McBurnie and David Clutterbuck. Give your company the marketing edge. *Weidenfeld and Nicolson, London*, 1988. [Citado na página 6]

[9] Sunghae Jun. Technology marketing using pca, som, and stp strategy modeling. *International Journal of Computer Science Issues (IJCSI)*, 8(1):87, 2011. [Citado na página 6]

[10] Kh Khaled Kalam. Market segmentation, targeting and positioning strategy adaptation for the global business of vodafone telecommunication company. *International Journal of Research and Innovation in Social Science*, 4(6):427–430, 2020. [Citado na página 6]

[11] L Moutinho et al. Segmentation, targeting, positioning and strategic marketing. *Strategic management in tourism*, pages 121–166, 2000. [Citado nas páginas 7 e 8]

[12] Thomas Reutterer, Andreas Mild, Martin Natter, and Alfred Taudes. A dynamic segmentation approach for targeting and customizing direct marketing campaigns. *Journal of interactive Marketing*, 20(3-4):43–57, 2006. [Citado na página 7]

[13] TP Beane and DM Ennis. Market segmentation: a review. *European journal of marketing*, 21(5):20–42, 1987. [Citado na página 7]

[14] Syed Saad Andaleeb. Market segmentation, targeting, and positioning. In *Strategic Marketing Management in Asia*, pages 179–207. Emerald Group Publishing Limited, 2016. [Citado na página 7]

[15] Neil A Morgan and Nigel F Piercy. Market-led quality. *Industrial Marketing Management*, 21(2):111–118, 1992. [Citado na página 8]

[16] Maureen Meadows and Sally Dibb. Assessing the implementation of market segmentation in retail financial services. *International Journal of Service Industry Management*, 9(3):266–285, 1998. [Citado na página 8]

[17] Jean-Marie Choffray. *A methodology for investigating the nature of the industrial adoption process and the differences in perceptions and evaluation criteria among decision participants.* PhD thesis, Massachusetts Institute of Technology, 1977. [Citado na página 8]

[18] W.L. Wilkie, J.B. Cohen, and Marketing Science Institute. *An Overview of Market Segmentation: Behavioral Concepts and Research Approaches.* Marketing Science Institute. Research program. Working paper report. Marketing Science Institute, 1977. [Citado na página 9]

[19] Robert M Morgan and Shelby D Hunt. The commitment-trust theory of relationship marketing. *Journal of marketing*, 58(3):20–38, 1994. [Citado na página 10]

[20] Ethem Alpaydin. *Introduction to machine learning.* MIT press, 2020. [Citado na página 10]

[21] Tom M Mitchell. Artificial neural networks. *Machine learning*, 45(81):127, 1997. [Citado na página 11]

[22] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016. [Citado na página 11]

[23] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009. [Citado nas páginas 11 e 12]

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. [Citado nas páginas 11 e 13]

[25] Yawen Li, Weifeng Jiang, Liu Yang, and Tian Wu. On neural networks and learning systems for business computing. *Neurocomputing*, 275:1150–1159, 2018. [Citado na página 13]

[26] M'hamed Outanoute, Mohamed Baslam, and Belaid Bouikhalene. Genetic algorithm learning of nash equilibrium: application on price-qos competition in telecommunications market. *Journal of Electronic Commerce in Organizations (JECO)*, 13(3):1–14, 2015. [Citado na página 14]

[27] Roger D Blackwell, Paul W Miniard, James F Engel, et al. *Comportamiento del consumidor*. Thomson Mexico City, México, 2002. [Citado na página 14]

[28] Vannessa Duarte, Sergio Zuniga-Jara, and Sergio Contreras. Machine learning and marketing: A systematic literature review. *IEEE Access*, 2022. [Citado nas páginas 14 e 15]

[29] A Caroline Tynan and Jennifer Drayton. Market segmentation. *Journal of marketing management*, 2(3):301–335, 1987. [Citado na página 15]

[30] J Scott Armstrong, Roderick J Brodie, and Shelby H McIntyre. Forecasting methods for marketing: Review of empirical research. *International Journal of Forecasting*, 3(3-4):355–376, 1987. [Citado na página 15]

[31] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012. [Citado na página 26]

[32] Bento Murteira, Carlos Silva Ribeiro, João Andrade e Silva, and Carlos Pimenta. *Introdução à estatística*. McGraw-Hill, 2007. [Citado na página 28]

[33] Daniela Witten and Gareth James. *An introduction to statistical learning with applications in R*. springer publication, 2013. [Citado na página 29]

Anexo A

Portugal

Método de Regressão Linear Múltipla para Portugal:

```
> modelo = lm(Dados$Valor ~ Populacao + Densidade + Area +
    Rendimentos + Gini + CarrosHab + CarrosRend + M2 + Campos +
    Fogos + P90P10, data = Dados)
> summary(modelo)

Call:
lm(formula = Dados$Valor ~ Populacao + Densidade + Area +
    Rendimentos +
     Gini + CarrosHab + CarrosRend + M2 + Campos +
     Fogos + P90P10, data = Dados)

Residuals:
      Min        1Q    Median        3Q       Max
-2575668   -165776     38957    227001   6670945

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.928e+05  5.215e+05   1.137   0.2567
Populacao   -8.979e-01  1.282e+00  -0.700   0.4844
Densidade    1.579e+02  7.789e+01   2.027   0.0436 *
Area         1.530e+02  1.651e+02   0.926   0.3550
Rendimentos -3.090e+01  2.656e+01  -1.163   0.2457
Gini        -6.645e+05  1.134e+06  -0.586   0.5583
CarrosHab    4.470e+04  5.577e+03   8.015 3.52e-14 ***
```

```
CarrosRend   -3.209e+05  1.421e+05  -2.258   0.0248 *
M2           -2.143e+02  1.288e+02  -1.663   0.0974 .
Campos        6.454e+05  5.679e+04  11.364  < 2e-16 ***
Fogos         5.049e+03  8.573e+02   5.889 1.17e-08 ***
P90P10        4.096e+03  5.870e+04   0.070   0.9444
---

Residual standard error: 702000 on 265 degrees of freedom
Multiple R-squared:  0.6901,  Adjusted R-squared:  0.6761
F-statistic: 49.18 on 12 and 265 DF,  p-value: < 2.2e-16
```

Depois do modelo inicial, recorreu-se ao método de AIC com o objetivo de encontrar o melhor modelo:

```
> step(modelo, dir = "backward")
Start:  AIC=7497.35
Dados$Valor ~ Populacao + Densidade + Area + Rendimentos + Gini +
    CarrosHab + CarrosRend + M2 + Campos + Fogos +
    P90P10

               Df  Sum of Sq         RSS    AIC
- P90P10        1 2.3995e+09 1.3058e+14 7495.4
- Gini          1 1.6924e+11 1.3075e+14 7495.7
- Populacao     1 2.4158e+11 1.3082e+14 7495.9
- Area          1 4.2295e+11 1.3100e+14 7496.2
- Rendimentos   1 6.6701e+11 1.3124e+14 7496.8
<none>                       1.3058e+14 7497.4
- M2            1 1.3633e+12 1.3194e+14 7498.2
- Densidade     1 2.0253e+12 1.3260e+14 7499.6
- CarrosRend    1 2.5119e+12 1.3309e+14 7500.6
- Fogos         1 1.7088e+13 1.4766e+14 7529.5
- CarrosHab     1 3.1656e+13 1.6223e+14 7555.7
- Campos        1 6.3628e+13 1.9421e+14 7605.7

Step:  AIC=7495.36
Dados$Valor ~ Populacao + Densidade + Area + Rendimentos + Gini +
    CarrosHab + CarrosRend + M2 + Campos + Fogos

               Df  Sum of Sq         RSS    AIC
- Populacao     1 2.4149e+11 1.3082e+14 7493.9
- Gini          1 3.2382e+11 1.3090e+14 7494.0
- Area          1 4.2097e+11 1.3100e+14 7494.3
- Rendimentos   1 6.6710e+11 1.3125e+14 7494.8
<none>                       1.3058e+14 7495.4
- M2            1 1.3786e+12 1.3196e+14 7496.3
- Densidade     1 2.0379e+12 1.3262e+14 7497.7
- CarrosRend    1 2.5116e+12 1.3309e+14 7498.7
- Fogos         1 1.7474e+13 1.4805e+14 7528.3
```

```
- CarrosHab     1 3.1765e+13 1.6234e+14 7553.9
- Campos        1 6.4030e+13 1.9461e+14 7604.3

Step:  AIC=7493.87
Dados$Valor ~ Densidade + Area + Rendimentos + Gini + CarrosHab +
    CarrosRend + M2 + Campos + Fogos

               Df  Sum of Sq        RSS    AIC
- Gini          1 3.3732e+11 1.3116e+14 7492.6
- Area          1 4.2210e+11 1.3124e+14 7492.8
<none>                       1.3082e+14 7493.9
- Rendimentos   1 9.7624e+11 1.3180e+14 7493.9
- M2            1 1.4106e+12 1.3223e+14 7494.9
- Densidade     1 1.8344e+12 1.3266e+14 7495.7
- CarrosRend    1 2.7280e+12 1.3355e+14 7497.6
- Fogos         1 1.9150e+13 1.4997e+14 7529.8
- CarrosHab     1 3.1589e+13 1.6241e+14 7552.0
- Campos        1 6.4499e+13 1.9532e+14 7603.3

Step:  AIC=7492.59
Dados$Valor ~ Densidade + Area + Rendimentos + CarrosHab +
    CarrosRend +
    M2 + Campos + Fogos

               Df  Sum of Sq        RSS    AIC
- Area          1 3.6538e+11 1.3152e+14 7491.4
<none>                       1.3116e+14 7492.6
- Rendimentos   1 1.3213e+12 1.3248e+14 7493.4
- M2            1 1.7385e+12 1.3290e+14 7494.2
- Densidade     1 1.7793e+12 1.3294e+14 7494.3
- CarrosRend    1 3.5079e+12 1.3467e+14 7497.9
- Fogos         1 1.9560e+13 1.5072e+14 7529.2
- CarrosHab     1 3.1613e+13 1.6277e+14 7550.6
- Campos        1 6.4371e+13 1.9553e+14 7601.6

Step:  AIC=7491.36
Dados$Valor ~ Densidade + Rendimentos + CarrosHab + CarrosRend +
    M2 + Campos + Fogos

               Df  Sum of Sq        RSS    AIC
<none>                       1.3152e+14 7491.4
- Rendimentos   1 1.3121e+12 1.3284e+14 7492.1
- Densidade     1 1.5051e+12 1.3303e+14 7492.5
- M2            1 1.5966e+12 1.3312e+14 7492.7
- CarrosRend    1 3.8453e+12 1.3537e+14 7497.4
- Fogos         1 2.0259e+13 1.5178e+14 7529.2
- CarrosHab     1 3.1312e+13 1.6284e+14 7548.7
- Campos        1 6.6493e+13 1.9802e+14 7603.1

```

```
Call:
lm(formula = Dados$Valor ~ Densidade + Rendimentos + CarrosHab +
    CarrosRend + M2 + Campos + Fogos, data = Dados)

Coefficients:
(Intercept)    Densidade  Rendimentos    CarrosHab   CarrosRend
              M2       Campos           Fogos
  6.553e+05    1.170e+02   -4.062e+01    4.427e+04   -3.768e+05
      -2.239e+02    6.442e+05       4.912e+03

> modelofinal = lm(Dados$Valor ~ Densidade + Rendimentos +
    CarrosHab + CarrosRend + M2 + Campos + Fogos, data = Dados)
> summary(modelofinal)

Call:
lm(formula = Dados$Valor ~ Densidade + Rendimentos + CarrosHab +
    CarrosRend + M2 + Campos + Fogos, data = Dados)

Residuals:
     Min       1Q   Median       3Q      Max
-2572690  -195779    51752   234021  6667366

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.553e+05  4.858e+05   1.349  0.17845
Densidade    1.170e+02  6.667e+01   1.755  0.08048 .
Rendimentos -4.062e+01  2.479e+01  -1.638  0.10255
CarrosHab    4.427e+04  5.532e+03   8.003 3.66e-14 ***
CarrosRend  -3.768e+05  1.343e+05  -2.804  0.00541 **
M2          -2.239e+02  1.239e+02  -1.807  0.07187 .
Campos       6.442e+05  5.524e+04  11.662  < 2e-16 ***
Fogos        4.912e+03  7.631e+02   6.437 5.57e-10 ***
---


Residual standard error: 699200 on 269 degrees of freedom
Multiple R-squared:  0.6879,  Adjusted R-squared:  0.6786
F-statistic:  74.1 on 8 and 269 DF,  p-value: < 2.2e-16
```

Anexo B

Espanha

Método de Regressão Linear Múltipla para Espanha:

```
> modelo = lm(Valor ~ M2 + Campos_Area + Area + Populacao +
    Rendimento + Gini + Rend_Ricos + PIB_PC + Carros_Hab, data =
    Dados)
> summary(modelo)

Call:
lm(formula = Valor ~ M2 + Campos_Area + Area + Populacao +
    Rendimento +
     Gini + Rend_Ricos + PIB_PC + Carros_Hab, data = Dados)

Residuals:
     Min       1Q   Median       3Q      Max
-3487161  -545431   138739   716159  4987835

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.423e+06  3.524e+06  -1.255   0.2163
M2           1.307e+03  1.076e+03   1.214   0.2315
Campos_Area  1.945e+08  2.582e+08   0.754   0.4553
Area         1.454e+02  6.647e+01   2.188   0.0343 *
Populacao   -1.886e-01  3.164e-01  -0.596   0.5542
Rendimento   2.635e+01  2.120e+02   0.124   0.9017
Gini        -4.464e+04  8.894e+04  -0.502   0.6183
Rend_Ricos   4.713e+04  3.629e+04   1.299   0.2012
```

```
PIB_PC       -1.307e+02  1.135e+02  -1.151   0.2563
Carros_Hab    4.154e+08  7.714e+07   5.385 3.02e-06 ***
---


Residual standard error: 1633000 on 42 degrees of freedom
Multiple R-squared:  0.7155,  Adjusted R-squared:  0.6545
F-statistic: 11.74 on 9 and 42 DF,  p-value: 5.891e-09
```

Depois do modelo inicial, recorreu-se ao método de AIC com o objetivo de encontrar o melhor modelo:

```
> step(modelo, dir = "backward")
Start:  AIC=1496.72
Valor ~ M2 + Campos_Area + Area + Populacao + Rendimento + Gini +
    Rend_Ricos + PIB_PC + Carros_Hab

               Df  Sum of Sq        RSS    AIC
- Rendimento    1 4.1205e+10 1.1206e+14 1494.7
- Gini          1 6.7192e+11 1.1269e+14 1495.0
- Populacao     1 9.4820e+11 1.1296e+14 1495.2
- Campos_Area   1 1.5144e+12 1.1353e+14 1495.4
- PIB_PC        1 3.5321e+12 1.1555e+14 1496.3
- M2            1 3.9310e+12 1.1595e+14 1496.5
<none>                      1.1202e+14 1496.7
- Rend_Ricos    1 4.4975e+12 1.1651e+14 1496.8
- Area          1 1.2768e+13 1.2478e+14 1500.3
- Carros_Hab    1 7.7346e+13 1.8936e+14 1522.0

Step:  AIC=1494.74
Valor ~ M2 + Campos_Area + Area + Populacao + Gini + Rend_Ricos +
    PIB_PC + Carros_Hab

               Df  Sum of Sq        RSS    AIC
- Gini          1 8.9739e+11 1.1295e+14 1493.2
- Populacao     1 9.8136e+11 1.1304e+14 1493.2
- Campos_Area   1 1.5829e+12 1.1364e+14 1493.5
- M2            1 3.8918e+12 1.1595e+14 1494.5
<none>                      1.1206e+14 1494.7
- Rend_Ricos    1 5.3920e+12 1.1745e+14 1495.2
- PIB_PC        1 5.8019e+12 1.1786e+14 1495.4
- Area          1 1.3743e+13 1.2580e+14 1498.8
- Carros_Hab    1 7.8607e+13 1.9066e+14 1520.4

Step:  AIC=1493.15
Valor ~ M2 + Campos_Area + Area + Populacao + Rend_Ricos + PIB_PC
    +
    Carros_Hab
```

```
36
37                Df  Sum of Sq        RSS    AIC
38 - Populacao     1 8.5188e+11 1.1381e+14 1491.5
39 - Campos_Area   1 1.5547e+12 1.1451e+14 1491.9
40 - M2            1 4.1208e+12 1.1708e+14 1493.0
41 <none>                       1.1295e+14 1493.2
42 - Rend_Ricos    1 4.6299e+12 1.1758e+14 1493.2
43 - PIB_PC        1 5.3077e+12 1.1826e+14 1493.5
44 - Area          1 1.5550e+13 1.2850e+14 1497.9
45 - Carros_Hab    1 7.7716e+13 1.9067e+14 1518.4
46
47 Step:  AIC=1491.54
48 Valor ~ M2 + Campos_Area + Area + Rend_Ricos + PIB_PC + Carros_Hab
49
50                Df  Sum of Sq        RSS    AIC
51 - Campos_Area   1 1.2692e+12 1.1508e+14 1490.1
52 - M2            1 3.6048e+12 1.1741e+14 1491.2
53 - Rend_Ricos    1 3.9990e+12 1.1781e+14 1491.3
54 <none>                       1.1381e+14 1491.5
55 - PIB_PC        1 4.6603e+12 1.1847e+14 1491.6
56 - Area          1 1.5030e+13 1.2884e+14 1496.0
57 - Carros_Hab    1 9.6394e+13 2.1020e+14 1521.5
58
59 Step:  AIC=1490.12
60 Valor ~ M2 + Area + Rend_Ricos + PIB_PC + Carros_Hab
61
62               Df  Sum of Sq        RSS    AIC
63 - Rend_Ricos   1 2.8105e+12 1.1789e+14 1489.4
64 <none>                      1.1508e+14 1490.1
65 - PIB_PC       1 7.3318e+12 1.2241e+14 1491.3
66 - M2           1 1.1310e+13 1.2639e+14 1493.0
67 - Area         1 1.4676e+13 1.2975e+14 1494.4
68 - Carros_Hab   1 1.5551e+14 2.7058e+14 1532.6
69
70 Step:  AIC=1489.37
71 Valor ~ M2 + Area + PIB_PC + Carros_Hab
72
73               Df  Sum of Sq        RSS    AIC
74 <none>                      1.1789e+14 1489.4
75 - PIB_PC       1 6.5498e+12 1.2444e+14 1490.2
76 - Area         1 1.4902e+13 1.3279e+14 1493.6
77 - M2           1 2.1261e+13 1.3915e+14 1496.0
78 - Carros_Hab   1 1.6237e+14 2.8026e+14 1532.4
79
80 Call:
81 lm(formula = Valor ~ M2 + Area + PIB_PC + Carros_Hab, data = Dados
       )
82
83 Coefficients:
```

```
(Intercept)            M2          Area        PIB_PC    Carros_Hab
 -5.919e+06     2.054e+03     1.400e+02    -9.370e+01     4.175e+08

> modelofinal = lm(Valor ~ M2 + Area + PIB_PC + Carros_Hab, data =
    Dados)
> summary(modelofinal)

Call:
lm(formula = Valor ~ M2 + Area + PIB_PC + Carros_Hab, data = Dados
    )

Residuals:
     Min       1Q   Median       3Q      Max
-3716623  -590749   123355   783775  5405950

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.919e+06  1.525e+06  -3.881 0.000324 ***
M2           2.054e+03  7.055e+02   2.911 0.005486 **
Area         1.399e+02  5.742e+01   2.437 0.018627 *
PIB_PC      -9.370e+01  5.798e+01  -1.616 0.112796
Carros_Hab   4.175e+08  5.188e+07   8.046 2.18e-10 ***
---


Residual standard error: 1584000 on 47 degrees of freedom
Multiple R-squared:  0.7006,  Adjusted R-squared:  0.6751
F-statistic: 27.49 on 4 and 47 DF,  p-value: 8.599e-12
```

Anexo C

Califórnia

Método de Regressão Linear Múltipla para a Califórnia:

```
> modelo = lm(Valor ~ Perc200 + DoisM + CasaSazonal +
    Perc5Bedrooms + Num5Bedrooms, data = Dados)
> summary(modelo)

Call:
lm(formula = Valor ~ Perc200 + DoisM + CasaSazonal + Perc5Bedrooms
    +
    Num5Bedrooms, data = Dados)

Residuals:
    Min      1Q  Median      3Q     Max
-616733  -53867   -6584   50324  685554

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.151e+05  4.874e+04   2.362   0.0220 *
Perc200       -1.341e+06  5.509e+05  -2.434   0.0184 *
DoisM          1.494e+01  3.315e+00   4.506 3.78e-05 ***
CasaSazonal   -2.811e+00  4.976e+00  -0.565   0.5746
Perc5Bedrooms -1.981e+06  1.689e+06  -1.172   0.2463
Num5Bedrooms   7.736e+00  3.445e+00   2.246   0.0290 *
---


```

```
Residual standard error: 198700 on 52 degrees of freedom
Multiple R-squared:  0.8142,  Adjusted R-squared:  0.7963
F-statistic: 45.57 on 5 and 52 DF,  p-value: < 2.2e-16
```

Depois do modelo inicial, recorreu-se ao método de AIC com o objetivo de encontrar o melhor modelo:

```
> step(modelo, dir = "backward")
Start:  AIC=1420.83
Valor ~ Perc200 + DoisM + CasaSazonal + Perc5Bedrooms +
    Num5Bedrooms

                Df  Sum of Sq        RSS    AIC
- CasaSazonal    1 1.2604e+10 2.0662e+12 1419.2
- Perc5Bedrooms  1 5.4292e+10 2.1079e+12 1420.3
<none>                       2.0536e+12 1420.8
- Num5Bedrooms   1 1.9915e+11 2.2528e+12 1424.2
- Perc200        1 2.3402e+11 2.2877e+12 1425.1
- DoisM          1 8.0171e+11 2.8553e+12 1438.0

Step:  AIC=1419.19
Valor ~ Perc200 + DoisM + Perc5Bedrooms + Num5Bedrooms

                Df  Sum of Sq        RSS    AIC
- Perc5Bedrooms  1 4.8046e+10 2.1143e+12 1418.5
<none>                       2.0662e+12 1419.2
- Num5Bedrooms   1 2.6505e+11 2.3313e+12 1424.2
- Perc200        1 2.7826e+11 2.3445e+12 1424.5
- DoisM          1 1.3123e+12 3.3786e+12 1445.7

Step:  AIC=1418.52
Valor ~ Perc200 + DoisM + Num5Bedrooms

               Df  Sum of Sq        RSS    AIC
<none>                      2.1143e+12 1418.5
- Num5Bedrooms  1 2.3537e+11 2.3497e+12 1422.6
- Perc200       1 8.2239e+11 2.9367e+12 1435.6
- DoisM         1 2.2861e+12 4.4004e+12 1459.0

Call:
lm(formula = Valor ~ Perc200 + DoisM + Num5Bedrooms, data = Dados)

Coefficients:
 (Intercept)       Perc200          DoisM  Num5Bedrooms
   1.069e+05    -1.811e+06     1.764e+01     4.664e+00

> modelofinal = lm(Valor ~ DoisM + Car4_2P + CasaSazonal +
    Perc5Bedrooms, data = Dados)
```

```
> summary(modelofinal)

Call:
lm(formula = Valor ~ DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms
    ,
    data = Dados)

Residuals:
    Min      1Q  Median      3Q     Max
-388684  -28891    4811   65025  486990

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.877e+04  3.423e+04   0.841   0.4044
DoisM          8.274e+00  1.681e+00   4.922 8.75e-06 ***
Car4_2P        8.091e+01  1.051e+01   7.701 3.37e-10 ***
CasaSazonal   -8.532e+00  3.593e+00  -2.374   0.0212 *
Perc5Bedrooms -5.149e+06  1.008e+06  -5.110 4.50e-06 ***
---


Residual standard error: 163200 on 53 degrees of freedom
Multiple R-squared:  0.8723,  Adjusted R-squared:  0.8627
F-statistic: 90.55 on 4 and 53 DF,  p-value: < 2.2e-16
```

Anexo D

Flórida

Método de Regressão Linear Múltipla para a Flórida:

```
> modelo = lm(Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P +
    CasaSazonal + Perc5Bedrooms + Num5Bedrooms, data = Dados)
> summary(modelo)

Call:
lm(formula = Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P +
    CasaSazonal +
    Perc5Bedrooms + Num5Bedrooms, data = Dados)

Residuals:
    Min      1Q  Median      3Q     Max
-132365   -5408   -1897    6012  156090

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.915e+03  9.064e+03   0.984   0.3294
Perc200       -2.114e+05  2.420e+05  -0.874   0.3858
DoisM          2.348e+01  4.762e+00   4.931 7.01e-06 ***
Car4_1P        4.592e+01  2.957e+01   1.553   0.1259
Car4_2P       -1.521e+01  8.574e+00  -1.774   0.0812 .
CasaSazonal   -1.194e+00  4.585e-01  -2.604   0.0116 *
Perc5Bedrooms -1.110e+05  4.293e+05  -0.259   0.7969
Num5Bedrooms   3.032e+00  1.375e+00   2.206   0.0313 *
---
```

```
23
24
25 Residual standard error: 36260 on 59 degrees of freedom
26 Multiple R-squared:  0.7363,  Adjusted R-squared:  0.705
27 F-statistic: 23.54 on 7 and 59 DF,  p-value: 6.575e-15
```

Depois do modelo inicial, recorreu-se ao método de AIC com o objetivo de encontrar o melhor modelo:

```
1  > step(modelo, dir = "backward")
2  Start:  AIC=1414.27
3  Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P + CasaSazonal +
       Perc5Bedrooms +
4       Num5Bedrooms
5
6                   Df  Sum of Sq        RSS    AIC
7  - Perc5Bedrooms   1 8.7905e+07 7.7653e+10 1412.3
8  - Perc200         1 1.0037e+09 7.8569e+10 1413.1
9  <none>                         7.7565e+10 1414.3
10 - Car4_1P         1 3.1694e+09 8.0735e+10 1415.0
11 - Car4_2P         1 4.1394e+09 8.1705e+10 1415.8
12 - Num5Bedrooms    1 6.3982e+09 8.3964e+10 1417.6
13 - CasaSazonal     1 8.9123e+09 8.6478e+10 1419.6
14 - DoisM           1 3.1961e+10 1.0953e+11 1435.4
15
16 Step:  AIC=1412.35
17 Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P + CasaSazonal +
       Num5Bedrooms
18
19                 Df  Sum of Sq        RSS    AIC
20 <none>                        7.7653e+10 1412.3
21 - Perc200        1 2.8215e+09 8.0475e+10 1412.7
22 - Car4_1P        1 3.1229e+09 8.0776e+10 1413.0
23 - Car4_2P        1 4.1485e+09 8.1802e+10 1413.8
24 - Num5Bedrooms   1 8.3633e+09 8.6017e+10 1417.2
25 - CasaSazonal    1 8.8431e+09 8.6496e+10 1417.6
26 - DoisM          1 4.1165e+10 1.1882e+11 1438.8
27
28 Call:
29 lm(formula = Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P +
       CasaSazonal +
30      Num5Bedrooms, data = Dados)
31
32 Coefficients:
33  (Intercept)        Perc200          DoisM        Car4_1P
        Car4_2P    CasaSazonal   Num5Bedrooms
34    9.590e+03     -2.549e+05      2.402e+01      4.552e+01     -1.523e
          +01     -1.177e+00      2.827e+00
```

```
35
36 > modelofinal = lm(Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P +
       CasaSazonal + Num5Bedrooms, data = Dados)
37 > summary(modelofinal)
38
39 Call:
40 lm(formula = Valor ~ Perc200 + DoisM + Car4_1P + Car4_2P +
       CasaSazonal +
41     Num5Bedrooms, data = Dados)
42
43 Residuals:
44     Min      1Q  Median      3Q     Max
45 -129989   -5436   -1770    6404  157826
46
47 Coefficients:
48                Estimate Std. Error t value Pr(>|t|)
49 (Intercept)   9.590e+03  8.612e+03   1.113   0.2700
50 Perc200      -2.549e+05  1.726e+05  -1.476   0.1450
51 DoisM         2.402e+01  4.258e+00   5.640 4.87e-07 ***
52 Car4_1P       4.552e+01  2.930e+01   1.553   0.1256
53 Car4_2P      -1.523e+01  8.507e+00  -1.790   0.0784 .
54 CasaSazonal  -1.177e+00  4.502e-01  -2.614   0.0113 *
55 Num5Bedrooms  2.827e+00  1.112e+00   2.542   0.0136 *
56 ---
57
58 Residual standard error: 35980 on 60 degrees of freedom
59 Multiple R-squared:  0.736, Adjusted R-squared:  0.7096
60 F-statistic: 27.88 on 6 and 60 DF,  p-value: 1.221e-15
```

Anexo E

Nova Iorque

Método de Regressão Linear Múltipla para Nova Iorque:

```
> modelo = lm(Valor ~ Total + Perc200 + DoisM + Car4_1P + Car4_2P
    + CasaSazonal + Perc5Bedrooms, data = Dados)
> summary(modelo)

Call:
lm(formula = Valor ~ Total + Perc200 + DoisM + Car4_1P + Car4_2P +
    CasaSazonal + Perc5Bedrooms, data = Dados)

Residuals:
    Min      1Q  Median      3Q     Max
-431778  -72985    6917  112652  531247

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.930e+05  4.248e+04  -4.543 3.16e-05 ***
Total         -2.414e-01  1.967e-01  -1.227   0.2252
Perc200        6.194e+05  6.421e+05   0.965   0.3390
DoisM          6.849e+00  6.405e+00   1.069   0.2897
Car4_1P        1.573e+02  1.567e+02   1.004   0.3199
Car4_2P        1.162e+02  5.300e+01   2.192   0.0327 *
CasaSazonal    4.968e+01  5.479e+00   9.067 1.95e-12 ***
Perc5Bedrooms -2.128e+06  1.131e+06  -1.882   0.0652 .
---

```

```
24
25 Residual standard error: 170700 on 54 degrees of freedom
26 Multiple R-squared:  0.834, Adjusted R-squared:  0.8125
27 F-statistic: 38.77 on 7 and 54 DF,  p-value: < 2.2e-16
```

Depois do modelo inicial, recorreu-se ao método de AIC com o objetivo de encontrar o melhor modelo:

```
1  > step(modelo, dir = "backward")
2  Start:  AIC=1501.37
3  Valor ~ Total + Perc200 + DoisM + Car4_1P + Car4_2P + CasaSazonal
       +
4      Perc5Bedrooms
5
6                  Df  Sum of Sq        RSS    AIC
7  - Perc200        1 2.7122e+10 1.6012e+12 1500.4
8  - Car4_1P        1 2.9374e+10 1.6034e+12 1500.5
9  - DoisM          1 3.3327e+10 1.6074e+12 1500.7
10 - Total          1 4.3874e+10 1.6179e+12 1501.1
11 <none>                        1.5741e+12 1501.4
12 - Perc5Bedrooms  1 1.0327e+11 1.6773e+12 1503.3
13 - Car4_2P        1 1.4008e+11 1.7141e+12 1504.7
14 - CasaSazonal    1 2.3965e+12 3.9706e+12 1556.7
15
16 Step:  AIC=1500.43
17 Valor ~ Total + DoisM + Car4_1P + Car4_2P + CasaSazonal +
       Perc5Bedrooms
18
19                 Df  Sum of Sq        RSS    AIC
20 - Car4_1P        1 2.4150e+10 1.6253e+12 1499.4
21 - Total          1 5.0401e+10 1.6516e+12 1500.3
22 <none>                        1.6012e+12 1500.4
23 - Perc5Bedrooms  1 7.6243e+10 1.6774e+12 1501.3
24 - DoisM          1 1.0006e+11 1.7013e+12 1502.2
25 - Car4_2P        1 1.9595e+11 1.7971e+12 1505.6
26 - CasaSazonal    1 2.3809e+12 3.9821e+12 1554.9
27
28 Step:  AIC=1499.36
29 Valor ~ Total + DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms
30
31                 Df  Sum of Sq        RSS    AIC
32 - Total          1 4.2601e+10 1.6679e+12 1499.0
33 <none>                        1.6253e+12 1499.4
34 - Perc5Bedrooms  1 8.0610e+10 1.7060e+12 1500.4
35 - DoisM          1 1.1059e+11 1.7359e+12 1501.4
36 - Car4_2P        1 4.4385e+11 2.0692e+12 1512.3
37 - CasaSazonal    1 2.3873e+12 4.0126e+12 1553.4
38
```

```
Step:  AIC=1498.96
Valor ~ DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms

                Df  Sum of Sq        RSS    AIC
<none>                        1.6679e+12 1499.0
- Perc5Bedrooms  1 6.7087e+10 1.7350e+12 1499.4
- DoisM          1 7.7968e+10 1.7459e+12 1499.8
- Car4_2P        1 4.4640e+11 2.1143e+12 1511.7
- CasaSazonal    1 2.9261e+12 4.5940e+12 1559.8

Call:
lm(formula = Valor ~ DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms
    ,
    data = Dados)

Coefficients:
  (Intercept)          DoisM        Car4_2P    CasaSazonal
      Perc5Bedrooms
   -1.842e+05      5.316e+00      1.306e+02      5.099e+01
       -1.484e+06

> modelofinal = lm(Valor ~ DoisM + Car4_2P + CasaSazonal +
    Perc5Bedrooms, data = Dados)
> summary(modelofinal)

Call:
lm(formula = Valor ~ DoisM + Car4_2P + CasaSazonal + Perc5Bedrooms
    ,
    data = Dados)

Residuals:
    Min      1Q  Median      3Q     Max
-463938  -92219   20211  121016  545344

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.842e+05  3.296e+04  -5.590 6.68e-07 ***
DoisM          5.316e+00  3.257e+00   1.632 0.108126
Car4_2P        1.306e+02  3.344e+01   3.906 0.000251 ***
CasaSazonal    5.099e+01  5.099e+00  10.000 3.76e-14 ***
Perc5Bedrooms -1.484e+06  9.801e+05  -1.514 0.135515
---


Residual standard error: 171100 on 57 degrees of freedom
Multiple R-squared:  0.8241,  Adjusted R-squared:  0.8118
F-statistic: 66.78 on 4 and 57 DF,  p-value: < 2.2e-16
```