UNIVERSIDADE CATÓLICA PORTUGUESA

"Análise descritiva e preditiva das vendas de uma operação de retalho."

por

Jorge Miguel Mesquita Moreira

Universidade Católica Portuguesa

Abril de 2022

UNIVERSIDADE CATÓLICA PORTUGUESA

"Análise descritiva e preditiva das vendas de uma operação de retalho."

Trabalho Final na modalidade de Projeto

apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Gestão especializado em Business Analytics

por

Jorge Miguel Mesquita Moreira

sob orientação de

Mário Amorim Lopes

Universidade Católica Portuguesa

Abril de 2022

Agradecimentos

Gostaria de agradecer ao meu professor e orientador Mário Amorim Lopes por toda a disponibilidade e acompanhamento que me deu durante este período, foi sem dúvida uma ajuda imprescindível.

Segundo, gostaria de agradecer ao Eng. Gabriel Gonçalves pela oportunidade que me deu de poder realizar o estágio na sua empresa, bem como todo o acompanhamento e apoio que demonstrou durante todo este período, garantindo que a minha experiência fosse rica em conhecimento, desenvolvendo assim as minhas competências.

Quero também agradecer à minha amiga Vera Silva por toda a orientação e apoio que me deu, especialmente nos momentos mais complicados. Ao Joel Coelho por me elucidar com questões relacionadas com Python.

Por último quero agradecer aos meus pais e irmã por todo o apoio e motivação que me deram este tempo todo, mesmo estando mais ausente.

Resumo

No ano de 2019, a empresa mudou a sua estratégia de obtenção de *Leads*, sendo estas qualquer tipo de contacto fornecido à empresa por parte de outra empresa ou indivíduo. Esta optou por adquirir um número inferior de *Leads*, mas com um valor de ticket médio, ou receita média por venda, mais elevado, indo à procura de oportunidades com "maior qualidade". Em consequência desta alteração surgiu a questão de quais as variáveis que mais influenciavam o resultado da empresa e se o Ticket Médio seria de facto uma dessas variáveis.

Assim sendo, foi estudada a relação entre as várias variáveis com o intuito de obter o peso de cada uma na Receita. Para obter tal relação foi realizada uma análise preditiva, onde foi aplicado o método de Regressão Linear Múltipla. Para além disso foi também efetuada uma análise descritiva, através da criação de *dashboards* no software PowerBI, onde foram explorados e analisados diversos KPIs. Os dados utilizados nas análises foram extraídos do *software* de gestão Bitrix24 e são relativos ao Customer Relationship Management (CRM) da empresa ERP24.

Na análise preditiva foram utilizadas como variáveis de estudo o "Ticket médio", que corresponde à média de receita de cada venda, as "Deals Perdidas", referentes ao número de vendas falhadas, a Taxa de Conversão Leads em Deals, doravante "Taxa de Conversão L-D", relativa à taxa de conversão em vendas, de contactos deixados por empresas ou indivíduos à empresa, e o Tempo de Conversão de Novas Deals em Deals Ganhas, doravante "Tempo de Conversão ND-DG", que representa o tempo que a empresa demora desde que um potencial negócio é introduzido no CRM até que este se concretize em venda. A análise resultou num modelo constituído pelo "Ticket Médio" e o "Tempo de Conversão ND-DG", que conseguiram explicar 55.5 % da variação das Receitas. Concluiu-se que a Receita irá variar cerca de 3.5 por cada unidade adicional do "Ticket Médio" e 7488.16 por cada unidade adicional do "Tempo de Conversão ND-DG".

Em conclusão e após análise do modelo criado podemos afirmar que a variável "Ticket Médio", assim como o "Tempo de conversão ND-DG", tem um impacto considerável na receita da empresa, tendo assim influenciado positivamente os resultados da empresa à alteração de estratégia adotada. Os resultados levam-nos à conclusão que a empresa deve continuar a apostar numa estratégia de aumento do ticket médio e em negócios que demoram mais tempo a converter, pois são aqueles que geram mais receita, o que acaba por ser

contraintuitivo, pois a empresa poderia estar mais preocupada com *quick-wins*, ou seja, negócios de conversão rápida.

Palavras-Chave: Análise Descritiva, Análise Preditiva, Regressão Linear Múltipla, Deals, Leads, Python, PowerBI

Número de Palavras: 9 101

Abstract

In 2019, the company changed its strategy for obtaining Leads, which are any type of contact provided to the company by another company or individual. It chose to acquire a lower number of Leads, but with a higher average ticket value, or average revenue per sale, going in search of opportunities with "higher quality". As a result of this change, the question arose as to which variables most influenced the company's results and if the average ticket was in fact one of those variables.

Therefore, the relationship between several variables was studied in order to obtain the weight of each one in the revenue. To obtain this relationship a predictive analysis was performed, where the Multiple Linear Regression method was applied. In addition, a descriptive analysis was also performed, through the creation of dashboards in PowerBI software, where several KPIs were explored and analyzed. The data used in the analyses were extracted from the Bitrix24 management software and are relative to the Customer Relationship Management (CRM) of the company ERP24.

In the predictive analysis were used as study variables the "Average Ticket", which corresponds to the average revenue of each sale, the "Lost Deals", referring to the number of failed sales, the Leads to Deals Conversion Rate, henceforth "L-D Conversion Rate", concerning the conversion rate into sales, of contacts left by companies or individuals to the company, and the Conversion Time from New Deals to Deals Won, hereinafter "ND-DG Conversion Time", which represents the time it takes from the time a potential deal is entered into the CRM until it becomes a sale. The analysis resulted in a model consisting of the "Average Ticket" and the "ND-DG Conversion Time", which were able to explain 55.5 % of the variation in Revenue. It was concluded that Revenue will vary about 3.5 for each additional unit of the Average Ticket and 7488.16 for each additional unit of the ND-DG Conversion Time.

In conclusion and after analyzing the model created we can state that the variable "Average Ticket", as well as the "ND-DG Conversion Time", has a considerable impact on the company's revenue, thus having positively influenced the company's results to the change of strategy adopted. The results lead us to the conclusion that the company should continue to invest in a strategy to increase the average ticket and in businesses that take longer to convert, because they are the ones that generate more revenue, which turns out to be counter-intuitive because the company could be more concerned with quick-wins, i.e., quick cash conversions.

Keywords: Descriptive Analysis, Predictive Analysis, Multiple Linear Regression, Deals, Leads, Python, PowerBI

Índice

Índice de Gráficos

Índice de Tabelas

Índice de Figuras

1. Introdução

No ano de 2019 a empresa decidiu alterar a sua estratégia de obtenção de *Leads*, que são qualquer contacto fornecido à empresa, como e-mail ou número de telefone. A empresa procurou reduzir o número de *Leads* que adquiria, porém as que adquiria tinham um ticket médio, que é a receita média por venda, maior, o que significa que a empresa optou por procurar oportunidades de "maior qualidade". Esta mudança na estratégia da empresa originou a questão de quais eram as variáveis que mais influenciavam as receitas da empresa, e se o ticket médio era de facto uma delas.

Para esclarecer essa pergunta, foi realizada uma análise preditiva utilizando um método de regressão e foi também realizada uma análise descritiva.

Entre muitos outros, na literatura é possível encontrar dois tipos de métodos para realizar a análise preditiva, nomeadamente séries temporais e Regressão Linear. Uma série temporal é uma sequência de observações temporal ou cronológica de uma variável de interesse, sendo possível obtermos padrões como tendências, ciclos, períodos, observações invulgares ou até mesmo combinações de padrões, através de gráficos (Montgomery et al. 2015). A Regressão Linear visa estimar a relação entre uma variável dependente, ou a prever uma ou mais variáveis independentes, denominadas variáveis preditivas, sendo que caso seja só uma variável é chamada Regressão Linear Simples e, se mais do que uma, Regressão Linear Múltipla (Uyanik & Guler, 2013).

Na análise preditiva foi utilizada a linguagem de programação Python, muito utilizada na análise de dados, tendo sido utilizado o método de Regressão Linear Múltipla. Para a aplicação do método foi utilizado como variável dependente, ou variável a prever, as receitas obtidas. As variáveis preditivas, ou independentes, escolhidas foram o número de "Deals Perdidas", o "Ticket Médio", a "Taxa de Conversão L-D" e o "Tempo de Conversão ND-DG". A primeira foi escolhida com o intuito de avaliar a influência que o acréscimo ou decréscimo das *Deals* consideradas perdidas têm na variação da receita. A segunda corresponde à receita média que cada venda, ou *Deal* ganha, gerou, sendo este um indicador essencial para compreender o quanto a empresa está a auferir em média com cada venda. A terceira variável consiste na proporção entre

o fluxo de Leads e o fluxo de Deals que entraram na empresa mensalmente. Por último, a variável "Tempo de Conversão de ND-DG", utilizada para analisar a influência que o tempo que a empresa demora desde que obtém a oportunidade de negócio até ao momento em que o concretize, com a variação da "Receita".

Para realizar a análise descritiva foi utilizado o *software* da Microsoft, PowerBI, tendo sido criados 3 painéis visuais que contêm métricas, indicadores e informações da empresa, também chamados de *dashboards*, que abordam a situação geral da empresa, assim como abordam de forma mais detalhada as *Deals* e *Leads*. Nesta análise descritiva estão presentes também o "Ticket Médio", "Deals Perdidas", "Taxa de Conversão L-D" e "Tempo de Conversão ND-DG".

Após realizadas as análises foi feita uma discussão de resultados e tiradas as devidas conclusões.

2. Revisão de Literatura

2.1 Métodos preditivos

Ao realizar a análise de negócio, para além da análise descritiva, que está relacionada com a situação atual da empresa e a sua performance, pode ser feita também, de maneira a tornar a análise mais completa, uma análise preditiva. A análise preditiva, como já foi referido anteriormente, é realizada através de modelos preditivos que usam a informação histórica para poder realizar previsões de eventos futuros (Appelbaum et al.,2017).

Existem variáveis que permitem prever com maior facilidade comparativamente a outras, sendo que uma previsão com qualidade está dependente de alguns fatores como, por exemplo, a quantidade de dados que estão disponíveis, o quão similar é o futuro comparativamente com o passado e o quão bem nós compreendemos os fatores que contribuem para essa previsão (Hyndman, R.J., &Athanasopoulos, G., 2018).

Nem todos as variáveis de previsão mudam da mesma forma ou têm a mesma influência no fator que estamos a prever, ou seja, uma boa previsão é aquela que captura os padrões e relações existentes nos dados históricos (Hyndman & Athanasopoulos, 2018).

As situações em que são feitas previsões podem variar bastante, nomeadamente a nível de horizonte temporal, fatores que determinarão o resultado da previsão, tipos de padrão de dados, entre outros. Os métodos de previsão podem também variar desde os mais simples, como o método Naive, que consiste em utilizar a observação mais recente como previsão, até aos métodos mais complexos como redes neurais (Hyndman & Athanasopoulos 2018).

Existem duas possíveis abordagens para a previsão, uma qualitativa e outra quantitativa. A primeira utiliza dados qualitativos, ou seja, dados não numéricos, como opiniões de peritos ou informações sobre algum evento em especial, e é especialmente útil se os dados disponíveis são inadequados a uma análise quantitativa. (Hofmann & Rutschman, 2018). Por outro lado, temos as previsões quantitativas, que são aplicadas quando estão disponíveis dados numéricos do passado e quando é razoável assumir que alguns padrões do

passado irão continuar no futuro (Hyndman & Athanasopoulos, 2018). Os modelos quantitativos podem focar-se em padrões ou mudança de padrões em dados históricos (modelos para séries temporais), ou podem focar-se nas relações específicas entre elementos do sistema, modelos esses chamados causais (Hofmann & Rutschman, 2018).

2.2 Regressão linear e múltipla

O principal objetivo dos métodos de regressão é estimar a relação existente entre variável dependente e variáveis independentes.

Ao usar métodos de Regressão é possível chegar a diversos tipos de conclusões,tais como se existe ou não relação entre a variável dependente e a variável ou variáveis independentes e, caso exista, quão forte é essa relação, ou seja, caso haja alguma variação na variável independente, qual será a variação que irá haver navariável dependente. Podemos também concluir se é possível fazer previsões futuras em relação à variável dependente tendo em conta as variáveis preditivas apresentadas(Alpar, 2003).

Existem vários métodos para estimar os valores dos coeficientes de Regressão, sendo que o mais usado é o Método dos Mínimos Quadrados, que consiste em obter os coeficientes que minimizam a soma dos quadrados das diferenças entre os valores observados e a linha de Regressão (Montgomery et all. 2012).

Um exemplo de um estudo de Regressão Linear Múltipla é o de Ghinea et al. (2016), que tinha como finalidade prever a quantidade de frações de desperdícios sólidos, nomeadamente papel, plástico, metal, vidro, orgânico e outro, usando como variáveis de previsão o número de residentes, a idade da população, a expectativa de vida urbana, e o total de desperdício sólido municipal. Os autores realizaram uma análise de Regressão, em que obtiveram as equações de Regressão. A partir de histogramas que representam a frequência dos residuais obtidos foi possível verificar- se que não existem *outliers*. Também verificaram a existência de alguma correlação entre variáveis através dos valores de *VIF* (variance inflation factor), ou fator de inflação de variância, que mede o quanto a variância de uma variável independente é influenciado pela sua correlação com outras variáveis independentes. Foi possível também verificar uma distribuição normal dos valores através do gráfico de distribuição normal. Os valores de R^2 obtidos em cada modelo foram perto de 100%, o que indica

que as mudanças ocorridas nas variáveis dependentes, neste caso os diferentes tipos de desperdício, são em grande parte explicadas pelos modelos.

Carboneau et al. (2008) também realizaram um estudo que investigava a aplicabilidade de métodos avançados de *Machine Learning* para previsão da procurade uma cadeia de abastecimentos, comparando-os com métodos mais tradicionais de previsão. *Machine Learning* consiste em algoritmos e modelos estatísticos utilizados em sistemas informáticos com o objetivo de realizar tarefas específicas sem precisarem de ser explicitamente programados (Mahesh B., 2018). Os autores utilizaram métodos avançados como redes neuronais, redes neuronais recorrentes e Support Vector Machine, cuja comparação foi feita com os métodos tradicionais Naive, Tendência, Média Movel e Regressão Linear, tendo utilizado como medidor de performance o Mean Absolute Error (MAE). Foram utilizadas duas bases de dados paraeste estudo, nomeadamente uma obtida a partir de uma cadeia de abastecimento simulada, e outra das encomendas da Canadian Foundries. A investigação concluiu que o uso de *Machine Learning* e Regressão Linear Múltipla ofereciam resultados de previsão mais precisos que os métodos mais simples (Naive, Tendência e Médias Móveis), contudo, não se verificou uma performance significativamente melhor dos métodos de *Machine Learning* comparativamente com o método de Regressão Linear.

Bon e NG (2017) apresentaram uma investigação cujo objetivo era a otimizaçãodas vendas de Panadol 650mg de um Centro Universitário de Saúde, através de métodos de previsão. Foram utilizados 10 métodos de previsão, nomeadamente o método de Médias Móveis Único, Alisamento Exponencial Único, Médias Móveis Duplo, Alisamento Exponencial Duplo, Regressão, Holt-Winter´s Aditivo, Aditivo Sazonal, Holt-Winter´s Multiplicativo, Sazonal Multiplicativo e ARIMA (Autoregressive Integrated Moving Average) . Os autores utilizaram como método de avaliação de performance o Root Mean Square Error (RMSE). Após realizados todos os cálculos chegou-se à conclusão de que o método de previsão com melhor performance dentro dos escolhidos foi o método de Regressão.

Um outro exemplo de aplicação de Regressão Linear e Múltipla foi o de Lassen et al. (2014) que demonstraram como os dados recolhidos da rede social Twitter podem ser usados para prever as vendas trimestrais dos iPhones, usando modelos de Regressão Linear e Múltipla. Os autores desenvolveram um modelo de Regressãolinear que relacionava os tweets associados aos iPhones (variável independente) e as vendas trimestrais dos iPhones, que demonstrou uma forte

relação entre as variáveis, vindo a mostrar-se ainda mais forte quando os autores acrescentaram uma segunda variável, nomeadamente o sentimento nos tweets, tornando este modelo num modelode Regressão Linear Múltipla.

2.3 Séries Temporais

Para além dos métodos preditivos de Regressão Linear Simples e Regressão Linear Múltipla, existem também outros métodos, como os de Alisamento Exponencial de Séries Temporais, sendo estes também amplamente usados nas áreas de análises de dados. Os métodos de Alisamento Exponencial são métodos de previsão baseados em médias ponderadas de observações históricas, cujo peso de cada observação decresce exponencialmente com idade das observações, ou seja, quanto mais antigo for, menor será o seu peso na previsão (Batselier & Vanhoucke, 2017).

Uma série temporal é uma sequência de observações temporal ou cronológica de uma variável de interesse, sendo que a partir dos gráficos de séries temporais é possível obtermos padrões como tendências, ciclos, períodos, observações invulgares ou até mesmo combinações de padrões (Montgomery et al. 2015). Estes tipos de padrões são fundamentais, pois diversos métodos de previsão de séries temporais usam-nos para fazer os cálculos das previsões. Os métodos de previsão de séries temporais mais simples apenas usam a informação disponível nas observações para fazer previsões, isto é, dão bastante ênfase à tendência ou à sazonalidade, ignorando os fatores que podem ter afetado o comportamento das variáveis observadas, tais como iniciativas de marketing, atividadedos competidores, mudanças nas condições económicas, entre outros. (Hyndman & Athanasopoulos, 2021).

Hyndman e Athanasopoulos (2021) apresentam uma definição para cada um destes padrões. Estes definem tendência como sendo um crescimento ou decréscimo de longo prazo nos dados fornecidos, sendo que este acréscimo ou decréscimo não precisam de ser lineares, apenas tem de se verificar uma tendência ao longo do tempo. Definem também sazonalidade como sendo um padrão de período fixo que ocorre quando as séries temporais são afetadas por algum tipo de fator sazonal, que pode ser a altura do ano, nomeadamente as estações, que podem ter influência na procura de certos produtos ou serviços (chocolates no inverno e gelados no verão, por exemplo), altura da semana, sendo o fim-de-semana um fator a ter em conta, ou até mesmo a altura do dia. Por fim,

os autores referem que há um ciclo quando as observações apresentam subidas e descidas sem frequência fixa.

Existem então diversos métodos preditivos, sendo que, para avaliar qual seria omais adequado a uma determinada previsão, deve ser usado algum tipo de comparação entre eles, nomeadamente um termo relacionado com os erros de cada método, podendo ser, por exemplo, o MAPE (Mean Absolute Percentage Error), MdAPE (Median Absolute Percentage Error), MASE (Mean Absolute Scaled Error), entre outros (Hyndman & Koehler, 2006).

Os métodos que serão abordados serão os métodos Naive, Médias Móveis Simples, Alisamento Exponencial Simples, Holt e, por último, os métodos Holt Winters Aditivo e Holt Winter Multiplicativo.

O método Naive é o mais simples de todos, sendo que os valores das previsões serão iguais ao último valor dos dados fornecidos. Este método assume que o último dado observado é a melhor estimativa do que pode ser o futuro (Hyndman & Athanasopoulos, 2021).

As previsões obtidas a partir do método de Médias Móveis Simples são obtidas calculando a média de *n* observações, em horizonte rolante, sendo que todos os dadosusados são igualmente ponderados, ou seja, têm o mesmo peso no cálculo dos valores previstos (Hansun, 2013).

O método de Alisamento Exponencial Simples é um método cujas previsões são calculadas a partir das médias ponderadas das observações passadas, sendo que quanto mais antigas forem, menos peso terão no cálculo das previsões. O parâmetro alfa segue uma exponencial negativa, o que significa que o peso das observações mais antigas decai a um ritmo exponencial. Este método é indicado para prever dados que não apresentem tendência ou sazonalidade, visto que estes fatores não são tidos emconta quando as previsões são calculadas.

Nirmala, Harjadi e Awaluddin (2021) realizaram uma investigação que consistiu numa análise ao padrão de vendas dos produtos Power Bumi da empresa PT.Zamrud Bumi Indonésia durante o período da pandemia Covid-19 e fazer previsões das mesmas utilizando dois métodos, sendo estes posteriormente comparados para determinar qual o melhor. Os autores utilizaram dois métodos de previsão, nomeadamente o método de Alisamento Exponencial e o método de Tendência Mínimos Quadrados, sendo que ao aplicar o AE estes usaram 3 valores de alfa (0.1, 0.5 e 0,9). Foram calculados 3 tipos de erro, sendo eles o MAD (Mean Absolute Deviation), MSE (Mean Square Error)

e MAPE (Mean Absolute Percentage Error), cujo objetivo é a minimização dos mesmos. Após se analisarem os resultados, conclui-seque o melhor método seria o método de Alisamento Exponencial, do qual alfa era 0.9 pois era o que apresentava melhores resultados no sentido da minimização dos erros calculados. Aumentando a complexidade dos métodos, temos então o método Holt, que é semelhante ao método de Alisamento Exponencial Simples, porém, o Holt acrescenta o parâmetro tendência no cálculo da previsão.

Por último, os métodos Holt-Winters Aditivo e Multiplicativo, para além de usarem a tendência, como é o caso do método Holt, incluem também a sazonalidade como parâmetro no cálculo das previsões (Hyndman & Athanasopoulos, 2021). Um exemplo da aplicação destes métodos é o estudo realizado por Kotsialos et al. (2005) que aborda o problema de fazer previsões a longo prazo que alguns métodos preditivos têm. Estes utilizaram o método Holt-Winters com uma extrapolação mais conservadora da tendência ao utilizarem a raiz quadrada do índice *m* na equação de previsão Ft+t=(Lt + bt√m)St+m , e usaram também o método Feedforward Multilayer Neural Networks (FMNN). Os resultados das previsões foram bastante satisfatórios, tendo o método FMNN o melhor desempenho, porém apenas por uma ligeira diferença comparativamente com o método de Holt-Winters que é bastante maissimples.

Uma outra utilização dos métodos de previsão em cima abordados foi de Nazime Aafthanorhan (2014) que, com o objetivo de fazer uma previsão da população da Malásia, utilizaram quatro métodos preditivos, nomeadamente Alisamento Exponencial único, Duplo Alisamento Exponencial, Holt´s (Brown) e Adaptative Response Rate Exponential Smoothing (ARRES). Os autores utilizaram dados de 1957 a 2013, recolhidos a partir do Departamento de Estatistica da Malásia. A medida de avaliação de performance escolhida foi o MSE, tendo o método de Holt´s obtido o valor mais baixo, logo, foi o método com melhor performance e, como tal, o escolhido para fazer a previsão.

2.4 Data Collection

A recolha dos dados é uma parte bastante importante no processo de análise de dados, pois é a fase de recolha dos dados que irão ser analisados posteriormente, logo, é essencial que estes sejam de qualidade, para não correr o risco se obterem resultados enviesados.

Granello e Wheaton (2004) abordaram o tema da recolha de dados online,

nomeadamente sobre a forma de e-mails ou inquéritos online, apresentando os benefícios e limitações deste tipo de recolha de dados e também um procedimento detalhado, incluindo estratégias para gerir tais limitações para os investigadores que queiram fazer os seus inquéritos online. Os benefícios apontados pelos autores foram os custos reduzidos, facilidade de introdução dos dados, flexibilidade do formato e facilidade de acesso a diferentes populações, enquanto as limitações referidas foram a dificuldade em obter uma amostra representativa, a baixa taxa de resposta e problemas tecnológicos.

Outro exemplo de estudos relacionados com recolha de dados é o de Ramakrishnan et al. (2012), que realizaram uma investigação empírica em que abordam os fatores que influenciam as estratégias de recolha de dados de BI. Os autores desenvolveram e testaram um modelo teórico que tinha como objetivo relacionar os fatores externos com a escolha de implementar ou não BI e as estratégiasde recolha de dados a usar. Os resultados da investigação revelaram uma relação positiva entre objetivos de implementação de BI, como a consistência e a transformação organizacional, e a seleção de uma estratégia abrangente de recolha de dados. Também Hox e Boeije (2005) escreveram um artigo chamado "Data Collection, Primary vs Secundary" que aborda as vantagens e desvantagens da recolha de dados primária e secundária, sendo que a primeira consiste nos dados recolhidos para um determinado objetivo de estudo, e a segunda consiste na reutilização desses mesmo dados para outro estudo ou investigação. Os autores abordam as principais estratégias de recolha de dados primária e estratégias para encontrar dados secundários úteis, os problemas associados à recuperação desses dados e metodologias para avaliar a qualidade desses dados secundários.

Kevin D.Frick (2009) descreveu os métodos básicos de estimar micro-custos focados na recolha de dado quantitativos, sendo que os métodos de recolha de dados utilizados foram as bases de dados administrativas em instalações singulares, dados administrativos da seguradora, formulários aplicados através de múltiplas configurações, painel de peritos, inquéritos ou entrevistas a um ou mais tipos deprestadores, revisão de fichas de pacientes, observação direta, assistentes pessoas digitais, programa de registos e dados diários. Os autores utilizaram os exemplos da literatura apresentada no artigo para demonstrar as etapas dos métodos de recolha dedados para um estudo de micro-custos.

2.5 Data Visualization

A Visualização de Dados é uma componente importante e cada vez mais utilizada para análise de dados e negócios, oferecendo ferramentas como *dashboards* que têm tido uma procura crescente ao longo dos anos. Isto são algumas conclusões tiradas por Diamond e Mattia (2015) no seu artigo "Data visualization: an exploratory study into software tools used by businesses" onde exploraram as várias ferramentas de Visualização de Dados usadas nos negócios. Os autores afirmam também que é imperativo que haja um foco na Visualização de Dados nas tomadas de decisões estratégicas de negócio na educação empresarial atual. Sadiku et al. (2016) abordam o tema Visualização de Dados no seu artigo "Data Visualization". Estes concluem que a Visualização de Dados é um processo de representar dados de forma gráfica ou em imagem de maneira clara, efetiva e eficiente, sendo esta uma ferramenta bastante poderosa quando o objetivo é a análise e interpretação de uma grande quantidade de dados.

Dentro do mesmo tema, Gorodov e Gubarev (2013) escreveram um artigo sobre a visualização e representação de dados em *Big Data*, abordando alguns problemas que estes possam ter e abordagens para os evitar. Para além disso, os autores fazem uma revisão dos métodos existentes de visualização de dados aplicados a *Big Data*, sendo que no final do artigo fornecem uma classificação para esses métodos. O tema foi também abordado por Fiaz et al. (2016), que escreveram um artigo focado em *Big Data* e técnicas de Visualização de Dados. Estes indicam que a Visualização de Dados é bastante vantajosa no que toca a análise de dados, pois tem a vantagem de oferecer uma "linguagem" comum entre os cientistas de dados e os executivos da empresa, ou seja, uma boa Visualização de Dados faz com que uma pessoa que tenha pouco conhecimento técnico no que toca a análise de dados possa facilmente interpretar os resultados e conclusões da análise através da visualização dos dados apresentados. Os autores referem também que pela Visualização de Dados as empresas conseguem controlar e analisar o valor real da grande quantidade de dados disponíveis, acelerando a compreensão dos mesmos e permitindo uma tomada de decisão mais rápida por parte dos executivos.

3. Metodologia

3.1 Método Regressão Linear

O modelo utilizado para análise preditiva foi o modelo de Regressão Linear Múltipla, criado por Francis Galton por volta do ano de 1877, ano em que este apresentou a primeira linha de regressão (Stanton, J. M., 2001). O principal foco dos métodos de regressão é estimar a relação existente entre variável dependente e variáveis independentes, sendo que nos casos em que só exista uma variável dependente a regressão é chamada Regressão Linear Simples, enquanto se tivermais do que uma variável dependente, então trata-se de uma Regressão LinearMúltipla (Uyanik & Guler, 2013).

A Regressão Linear Simples procura ajustar as observações a uma reta dada pela expressão $y = b_0 + b_1x + \varepsilon$, sendo y a variável dependente, b_0 e b_1 os coeficientes de Regressão da equação, x a variável independente e ε o erro, que consiste no desvio em relação à linha reta subjacente do modelo (Hyndman & Athanasopoulos, 2021).

A Regressão Linear Múltipla considera mais do que um regressor através da expressão $y = b_0 + b_1x_1 + b_2x_2 \ldots + b_kx_k$, em que y é a variável a ser prevista e x_1, x_2 e x_k são as variáveis independentes ou preditivas. Os coeficientes b_1, b_2 e b_k medem a influência que cada variável preditiva tem em relação à variável dependente y, tendo em conta o efeito de que as restantes variáveis preditivas têm em relação ao modelo (Hyndman & Athanasopoulos, 2021), mais concretamente, é a variação da média da distribuição da variável dependente resultante da variação de 1 unidade de x, quando todas as outras variáveis se mantêm constantes (Montgomery et all. 2012).

O primeiro passo para a aplicação deste método é, como em todos os outros modelos, a preparação e limpeza dos dados. Seguidamente, no sentido de avaliar se existe algum tipo de relação ou dependência entre as variáveis independentes, o que pode levar a resultados enviesados, é necessário sujeitar os dados a certos critérios ou pré-requisitos, nomeadamente análise de Multicolinearidade, Normalidade de Residuais, Homoscedasticidade e Autocorrelação de Residuais (Olive, D. J., 2017). O primeiro consiste na

dependência linear entre variáveis, cujo objetivo para uma Regressão Linear bem aplicada é que não haja multicolinearidade entre variáveis,sendo que o indicador utilizado para esta avaliação é o Variance Inflation Factor (*VIF*), em que geralmente a partir do valor 10 considera-se que há uma alta correlação entre as variáveis, ou uma matriz de correlação. O segundo, tal como o nome indica, consiste em verificar se os residuais, que são a diferença entre os valores obtidos pelo modelo e os valores reais, seguem uma distribuição normal em que, para tal avaliação, os dados são dispostos num *scatter plot*. Quanto à verificação da homoscedasticidade, esta é feita através da disposição dos residuais num *scatter plot* e consiste numa variância constante dos residuais, que é o pretendido para realizar uma boa regressão e, caso isto não se verifique, significa que estamos perante um caso de heteroscedasticidade. Quanto à Autocorrelação de Residuais, esta verifica-se quando há uma relação de dependência entre os residuais, ou seja, os residuais não são independentes uns dos outros, e tal pode ser verificado através de um gráfico que disponha os residuais em função do tempo.

Após verificados os critérios e calculados os Mínimos Múltiplos Quadrados, é necessário analisar os valores obtidos pelo modelo no sentido de verificar se as variáveis independentes produziram um modelo eficaz, ou seja, que explique bem a variação da variável dependente. Um dos primeiros indicadores a observar é o *R- squared*, ou R^2, que nos diz a percentagem da variação da variável dependente que é explicada pelo nosso modelo, por exemplo um R^2 de 0,7 significa que a variação de yé explicada em 70% pelo nosso modelo, sendo que a diferença deste para o R^2 ajustado é que o último tem em conta o número de variáveis usadas. O *F-statistic* é umindicador que verifica se o grupo de variáveis usadas são estatisticamente significativas ao comparar o modelo linear produzido com um cuja influência das variáveis independentes na dependente é 0. O indicador Prob (*F-Statistic*) mede a probabilidade da hipótese nula, ou seja, de as variáveis independentes não terem influência na variável dependente. Indicadores importantes a analisar são o Akaike´s Information Criteria (AIC) e Bayesian Information Criteria (BIC), que avaliam a eficácia do modelo no processo de Regressão Linear. Passando para os resultados relacionados com as variáveis, o indicador mais importante a observar é o *p-value* que representa a probabilidade de a variável independente não ter influência na variável dependente, sendo este valor comparado com valores referência, sendo o mais comum 5%, ou seja, as variáveis que obtiverem *p-values* superiores a este valor de referência não têm uma influência estatisticamente significativa na estimativa da variável dependente.

3.2 Dados

Para a realização desta análise foram utilizados dados retirados do software de gestão Bitrix24 relativos à empresa ERP24, tendo sido extraídas quatro tabelas, nomeadamente Leads, Deals, Companies e DealsStages cujos dados estão situados no espaço temporal de 2017 a 2022. As *Leads* correspondem a informações de contacto fornecidos à empresa, como o número telefónico ou email, a troco de alguma ofertade valor como por exemplo informações sobre um determinado produto ou serviço, sendo que esta tabela contém cerca de 5133 linhas e 47 colunas. As *Deals* correspondem à fase seguinte das *Leads*, e são definidas como oportunidades denegócio para a empresa, tendo a tabela Deals 3009 linhas e 82 colunas. As tabelas Companies, DealsStages são consideradas tabelas índice, ou seja, contêm valores únicos de cada elemento, atuando como um dicionário, sendo que a primeira corresponde às empresas registadas no *CRM* e a segunda às fases das *Deals* e respetivos funis de venda. Um funil de venda é definido como sendo um modelo que representa o percurso de um cliente, ou *Deals*, desde o primeiro contacto até à venda final.

3.3 Tratamento de Dados para Análise Preditiva

Para a análise preditiva foram extraídas as tabelas Deals e Leads. A primeira é constituída por 3009 linhas e 170 colunas, porém nem todos os dados são relevantes, visto que dos 14 funis de vendas, representados na coluna "CATEGORY_ID", apenas são utilizados para efeitos de controlo de vendas pela empresa 3 deles, nomeadamente o "Bitrix Sales", o "Bitrix FT Licenças" e o "Vendas Clientes Bitrix", ficando assim a base de dados, ou também designado *dataset*, reduzido a 1379 linhas. As colunas utilizadas foram "OPPORTUNITY", que corresponde ao valor que a empresa poderá ganhar com o cliente; "STAGE_ID", que contém o índice da fase da *Deal* associada e as colunas de data "DATE_CREATE" e "CLOSEDATE", que se referem à data de criação da *Deal* e data de encerramento, respetivamente. Apesar de haver 170 colunas, uma grande parte destas não pode ser utilizada para análise visto que contémsó, ou quase só, valores nulos.

Na primeira tabela criaram-se as colunas "*Deals* Ganhas" e "*Deals*

Perdidas", ambas binárias, em que o 1 representa uma *Deal* ganha no caso da primeira ou *Deal* perdida no caso da segunda e 0 o contrário, tendo como referência a coluna "STAGE_ID". Para além destas foram também geradas as colunas "*Deals* Contagem", que contém apenas valores "1" e servirá para a contagem de *Deals* durante o período especificado. Foi também criada a coluna "Deals sem Lead", que irá ser usada à posteriori para obter o valor total das *Leads*, isto porque existe o problema de várias *Deals* terem entrado diretamente no sistema, ou seja, não são precedentes de nenhuma *Lead*, para resolver isto, foi criada então esta coluna para mais tarde obter o número de *Leads* total caso a introdução de dados no sistema seguisse o procedimento normal. A receita obtida pela empresa foi calculada numa coluna denominada "Receita", contendo os proveitos que a empresa obteve com a *Deal*, retirados da coluna OPPORTUNITY, mas apenas contendo valores nas *Deals* que foram ganhas, ou seja, apenas nas vendas efetuadas. Foram ainda geradas no mesmo *dataset* as colunas "Tempo diff", calculando a diferença em meses desde a criação da *Deal* ("DATE_CREATE") até o encerramento da mesma ("CLOSEDATE"), e a coluna "Tempo de Conversão ND-DG", semelhante à coluna "Tempo diff", porém apenas apresenta valores nas Deals consideradas ganhas. Por último, foram adicionadas as colunas "year" e month", correspondendo ao ano e mês da coluna "DATE_CREATE", que serviram para posterior agrupamento dos dados em valores mensais e criação de um índice de data.

A tabela Leads contém 5133 linhas e 59 colunas, sendo que as colunas utilizadas nesta análise foram "DATE_CREATE" e "ID".

Seguidamente, foram criados 3 datasets, o "g", "t" e "l". No primeiro agruparam-se as colunas "Receita"," Deals Perdidas", "Deals Contagem", "Deals sem Lead", "year" e "month" por somatório mensal. No segundo por média mensal das colunas "Receita", "Tempo de Conversão ND-DG", "year" e "month", sendo que a coluna obtida pelo agrupamento mensal médio da "Receita" foi nomeada de "Ticket Médio", ou seja, o valor médio de receita que cada *Deal* ganha gera. Por fim, no *dataset* "l" agruparam-se por contagem mensal as colunas "ID", "year" e "month" da tabela Leads, em que a contagem da coluna "ID" originou a coluna que contém o número de *Leads*, que foi depois somada com a coluna "Deals sem Lead" para obter então o valor total de *Leads*. Esta soma foi realizada pois havia meses com mais *Deals* do que *Leads*, o que não é suposto, visto que em situações normais as *Deals* são precedidas pelas *Leads*. O *dataset* "g" foi o escolhido para agregar as variáveis a utilizarno modelo.

Posteriormente, foi gerada no *dataset* "g" a coluna "Leads", que é o

somatório da coluna de contagem "ID" com a coluna "Deals sem Lead", como anteriormente mencionado e "Taxa de Conversão L-D", ou Taxa de Conversão *Lead* para *Deal*, resultante da divisão entre o total de *Deals*, ou seja, "Deals Contagem" pela coluna "Leads" calculada anteriormente. Foram anexadas a este *dataset* g as colunas "Ticket Médio" e "Tempo de Conversão ND-DG" provenientes do *dataset* "t". Por fim, o *dataset* "g" ficou apenas com as colunas "Receita", "Deals Perdidas", "Taxa de Conversão L-D", "Ticket Médio" e "Tempo de Conversão ND-DG", tendo sido as restantes colunas removidas.

Para a aplicação do método de Regressão Linear, foram eliminadas as colunas que continham valores nulos, para uma análise mais eficaz, ficando o *dataset* reduzido a 33 valores, agrupado por valores mensais.

O objectivo da aplicação do modelo é a estimativa das receitas, pelo que foi escolhida como variável a prever, ou dependente, a Receita, para desta forma ser possível fazer uma previsão da receita que a empresa gera consoante a evolução de certas variáveis, ou fatores. Para variáveis independentes foram utilizadas as variáveis "Deals Perdidas", com o objetivo de analisar a influência que o aumento ou decréscimo do número de *Deals* que foram perdidas tem na receita; "Taxa de Conversão L-D", que permite explorar as variações da Receita com a variação do fluxo de *Leads* e *Deals* mensais; "Ticket Médio", para comparar o valor médio de cada vendacom a Receita total desse mês; e, por último, "Tempo de Conversão ND-DG", quemede o tempo médio que a empresa demora a fechar uma *Deal* desde o momento em que esta é criada.

3.4 PowerBI

Esta análise foi realizada com o intuito descritivo, com o objetivo de obter informações em relação ao estado da empresa, tendo sido utilizadas as tabelas Leads, Deals, Companies e DealsStages. Os dados foram filtrados, ficando apenas os correspondentes aos dados utilizados para aplicação do método de Regressão Linear Múltipla utilizados em Python, ou seja, os que se encontram situados nos 33 meses.

A nível de limpeza das tabelas, foram eliminadas todas as colunas que não foram necessárias para a criação dos visuais das *dashboards*, à exceção de certas colunas índice usadas para ligar as tabelas. Na tabela "Companies" ficaram as colunas "ID", "Localidade" e "Atividade". Na tabela "Deals" ficaram as colunas "CATEGORY_ID", "ID", "COMPANY_ID", "CONTACT_ID", "STAGE_ID",

"OPPORTUNITY", "DATE_CREATE", "LEAD_ID", "CLOSEDATE", "ASSIGNED_BY_ID" e "E-Categoria de cliente". Na DealsStages ficaram as colunas "stages.STATUS_ID", "pipeline.Name", "pipeline.ID" e "stages.Name". Por fim, na tabela *Leads* ficaram as colunas "ID"," COMPANY_ID"," CONTACT_ID","SOURCE_ID","STATUS_ID", "ASSIGNED_BY_ID", "DATE_CREATE", "Tamanho da Empresa", "Área de Trabalho" e "Setor Empresarial".

Visto que as tabelas não possuíam colunas com certos *KPIs* ou variáveis que se pretendia utilizar nas *dashboards*, e várias colunas apresentavam os valores da tabela como índices e não como nomes, foi preciso proceder ao cálculo dessas colunas ou medidas. Foram então criadas na tabela *Deals* as medidas "Deals ganhas contagem", "Deals perdidas contagem", "Taxa de Conversão L-D" e "Ticket Médio", sendo que a primeira corresponde ao total de *Deals* ganhas, a segunda ao total de *Deals* perdidas, a terceira à Taxa de Conversão *Leads* para *Deals* (total de *Deals* / total de *Leads*) e a quarta à receita média de cada *Deal* ganha. Foram também criadas nesta tabela as colunas "Diferença de tempo M", "Receita" e "Responsável", sendo que a primeira é referente à diferença em meses desde a criação da *Deal* até ao fecho da mesma, a segunda à receita obtida pela empresa com as *Deals* ganhas e a última ao responsável de cada *Deal*, tendo sido calculada a partir da coluna "ASSIGNED_BY", onde os valores numéricos foram convertidos para os respetivos nomes dos responsáveis. Na tabela Leads foram criadas as medidas "Leads Ganhas total", "Leads Perdidas" e "Total Leads", sendo que a primeira corresponde ao total de *Leads* acrescidas das *Deals* sem LEAD_ID, ou seja, que não foram precedentes de nenhuma *Lead*, a segunda corresponde ao total de *Leads* perdidas e a terceira ao Total de *Leads*, calculada a partir do somatório do total de *Leads* com as *Deals* sem LEAD_ID. A nível de colunas, foram criadas a "Área de Trabalho nome", "Estado", "Setor Empresarial nome", "Fontenome" e "Tamanho da Empresa nome", sendo uma conversão das colunas "Área de Trabalho", "STATUS_ID", "Setor Empresarial", "SOURCE_ID" e "Tamanho da Empresa", respetivamente, convertendo os valores numéricos nos respetivos nomes.

3.5 Estatística Descritiva

Tabela 1-Estatística descritiva das variáveis

Variáveis	Média	Mediana	Desvio Padrão	25%	75%	Min	Máx
Receita	16494.63	13 128.69	15882.27	2 723.4	25495.60	563.31	54975.05
Deals Perdidas	26.91	16.00	31.40	8.00	35.00	0.00	167.00
Taxa de Conversão L-D (%)	63.02	75.00	30.27	51.43	84.75	0.28	100.00
Ticket Médio	2300.86	1 654.25	2046.43	1 188.00	3185.95	187.77	10420.45
Tempo de Conversão ND-DG (meses)	1.26	1.00	1.17	0.50	1.75	0.00	4.28

Os dados acima apresentados permitem-nos tirar algumas conclusões relativamente às variáveis e à sua estrutura. Podemos observar que entre o período de 2017 e 2022, a empresa obteve uma receita média mensal de 16494.63€. Olhando para os resultados da variável *Deals* Perdidas, podemos verificar que no pior cenário a empresa apresentou um valor de mensal de *Deals* Perdidas de 167, no melhor cenário não teve nenhuma, tendo um registo mensal médio de 26.91 durante o período analisado. Analisando os valores respetivos à "Taxa de Conversão L-D" concluímos que a empresa conseguiu converter mensalmente, em média, 63% das suas *Leads* em *Deals*, sendo que no melhor dos casos esta taxa chegou a 100%. Em relação aos valores da variável "Ticket médio", podemos verificar que as *Deals* tiveram uma média mensal de 2300€ e 50% das vendas obtiveram um valor superior a 1 654.25€. Quanto ao "Tempo de Conversão ND-DG", é possível afirmar que a empresa demora em média 1,26 meses a fechar a *Deal*, ou seja, desde o momento em que a *Deal* é criada até que a venda se concretize definitivamente, sendo que em 50% das vendas foram concretizadas em menos de 1 mês. De salientar que os valores obtidos foram resultantes de cálculos mensais, e não individuais.

3.6 Verificação da Correlação entre as variáveis

Primeiramente, foi realizada uma análise à multicolinearidade das variáveis, para verificar o nível de correlação que existe entre as variáveis, e se este é estatisticamente significativo.

Tabela 2 - Matriz de correlação

	Receita	Deals Perdidas	Ticket Médio	Tempo de Conversão ND-DG	Taxa de Conversão L-D
Receita	1.0***	-0.02	0.61***	0.59***	0.53***
Deals Perdidas	-0.02	1.0***	-0.08	0.14	0.34*
Ticket Médio	0.61***	-0.08	1.0***	0.19	0.22
Tempo de Conversão ND-DG	0.59***	0.14	0.19	1.0***	0.4**
Taxa de Conversão L-D	0.53***	0.34*	0.22	0.4**	1.0***

Analisando os resultados obtidos na tabela acima, cujos valores representam o nível de correlação entre as variáveis, e os asteriscos o nível de significância (* = 0.01, **= 0.05 e ***=0.1) é possível verificar que existe correlação significativa entre as variáveis "Taxa de Conversão L-D" e "Tempo de Conversão ND-DG", sendo que a variável "Ticket Médio" não apresenta correlação significativa com nenhuma das variáveis e as "Deals Perdidas" correlaciona-se com a "Taxa de Conversão L-D". Tendo em conta estes resultados, foi construído um modelo que contém as três variáveis que não apresentam correlação significativa entre elas, sendo então o "Ticket Médio", o "Tempo de Conversão ND-DG" e as "Deals Perdidas". Posto isto, a hipótese a analisar é que a variação destas três variáveis tem influência na variação da Receita.

Para verificar se o modelo e as variáveis são bons indicadores de previsão da Receita, eobter os respetivos coeficientes de Regressão, foi utilizado como método de estimaçãoo método dos Mínimos Quadrados.

4. Resultados da Estimação

4.1 Modelo 1

Tabela 3 - Resultados da aplicação do método MMQ no Modelo 1

Dep. Variable:	Receita	R-squared:	0.565
Model:	OLS	Adj. R-squared:	0.505
Method:	Least Squares	F-statistic:	9.516
Date:	Tue, 12 Apr 2022	Prob (F-statistic):	0.000318
Time:	22:11:26	Log-Likelihood:	-277.24
No. Observations:	26	AIC:	562.5
Df Residuals:	22	BIC:	567.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1953.5541	5191.161	-0.376	0.710	-1.27e+04	8812.255
Tempo de Conversão ND-DG	6512.2476	1842.955	3.534	0.002	2690.193	1.03e+04
Ticket Médio	5.4618	1.598	3.418	0.002	2.148	8.775
Deals Perdidas	-41.4464	67.752	-0.612	0.547	-181.954	99.062

Omnibus:	0.178	Durbin-Watson:	1.371
Prob(Omnibus):	0.915	Jarque-Bera (JB):	0.387
Skew:	-0.072	Prob(JB):	0.824
Kurtosis:	2.419	Cond. No.	6.23e+03

Os resultados desta estimação mostram que 56,5% (R^2), ou 50,5% (Adj. R^2) da variação da Receita é explicada pelo modelo, sendo este modelo significante, F = 9.516, *p-value* = 0.000318. Olhando para o contributo individual, podemos verificar que tanto o Tempo de Conversão ND-DG ($t = 3.534$, *p-value* = 0.002) como o "Ticketmédio" ($t = 3.418$, *p-value* = 0.002) são boas variáveis preditivas, visto que os seus *p- values* são inferiores ao valor de referência 0.05. Porém, as Deals Perdidas mostraram ser um mau indicador de previsão, visto que o seu *p-value* foi de 0.547, bastante acimado valor de referência, o que levou à exclusão desta variável do modelo.

Posto isto, o modelo passou a ser constituído apenas pelo *Ticket* Médio e o Tempo de Conversão ND-DG, tendo sido feita novamente uma estimação utilizando o método dos Mínimos Quadrados, obtendo os seguintes resultados:

4.2 Modelo 2

Tabela 4 - Resultados da aplicação do método MQ no Modelo 2

OLS Regression Results

Dep. Variable:	Receita	R-squared:	0.557
Model:	OLS	Adj. R-squared:	0.519
Method:	Least Squares	F-statistic:	14.48
Date:	Tue, 12 Apr 2022	Prob (F-statistic):	8.50e-05
Time:	22:12:39	Log-Likelihood:	-277.46
No. Observations:	26	AIC:	560.9
Df Residuals:	23	BIC:	564.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3360.2213	4590.354	-0.732	0.472	-1.29e+04	6135.649
Ticket Médio	5.5438	1.570	3.530	0.002	2.295	8.792
Tempo de Conversão ND-DG	6431.7972	1813.077	3.547	0.002	2681.162	1.02e+04

Omnibus:	0.221	Durbin-Watson:	1.446
Prob(Omnibus):	0.896	Jarque-Bera (JB):	0.422
Skew:	0.018	Prob(JB):	0.810
Kurtosis:	2.377	Cond. No.	5.64e+03

Segundo os resultados da estimação, 55.7% (R^2), ou 51.9% (Adj. R^2), da variação da Receita é explicada pelas variáveis Ticket Médio e Tempo de Conversão ND-DG. Este modelo, segundo o teste ANOVA, revelou-se significativo, apresentando um *F-statistic* de 14.48 e um *p-value* bastante baixo de 0.000085. A nível individual pode-se verificar que o Ticket Médio é uma boa variável preditiva (t = 3.530, *p-value* = 0.002), assim como o Tempo de Conversão ND-DG (t = 3.547, *p-value* = 0.002). Os resultados desta estimação revelam que a Receita irá variar em mais 5.5 por cada 1 Euro adicional de ticket médio e 6431.8 por cada 1 unidade adicional de Tempo de Conversão ND-DG, logo, podemos afirmar que a "Receita" irá variar no mesmo sentido que o "Ticket Médio" e "Tempo de Conversão ND-DG". É possível verificar que tanto o AIC como o BIC reduziram do primeiro modelo para este, o que significa que houve um melhoramento do modelo.

4.2.1 Análise dos Pressupostos

Após escolhido o modelo, foi feita uma verificação para determinar se cumpria com os pressupostos do Regressão Linear Múltipla, nomeadamente Multicolinearidade, Normalidade dos Residuais, Homoscedasticidade e Autocorrelação dos residuais. Os resultados obtidos foram os seguintes:

- Multicolinearidade – Para verificar este pressuposto, para além da tabela mostrada anteriormente com as correlações e significâncias das variáveis, foram calculados também os valores *VIF*, cujos resultados podem ser vistos na seguinte tabela:

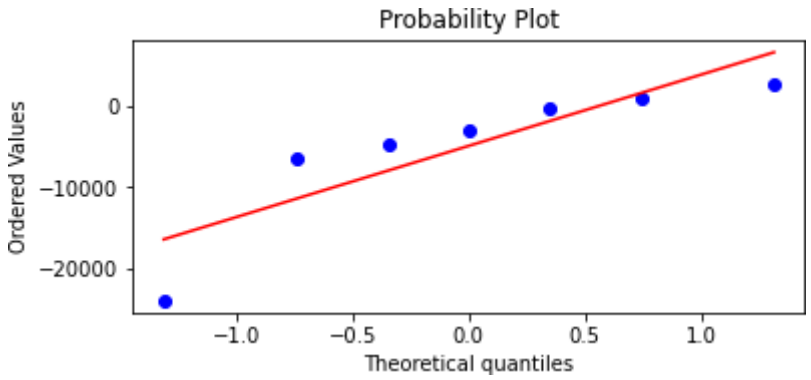
Tabela 5 - Valores VIF do modelo 2

	Ticket Médio	Tempo de Conversão ND-DG
vif	1.91	1.91

Tendo em conta os resultados obtidos, sendo que 10 é o valor de referência, no qual considera-se que existe um elevado nível de Multicolinearidade, podemos verificar que nenhum dos valores se aproxima desse valor de referência, sendo assim opressuposto é cumprido.

- Normalidade dos Residuais: Para verificação deste pressuposto, foi utilizado um *scatter plot* onde foram dispostos os valores dos residuais e uma linha de tendência, tendo sido obtido o seguinte gráfico:

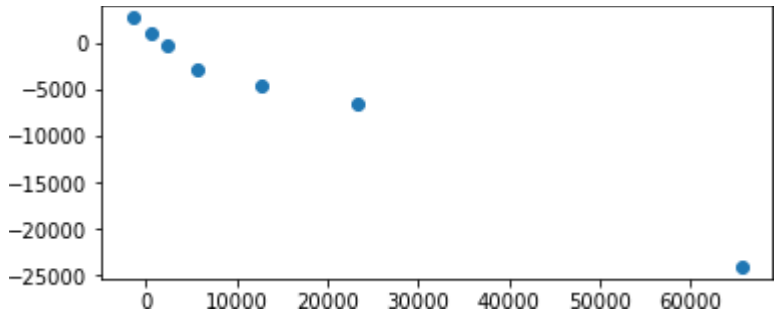
Gráfico 1- Scatter plot para análise da normalidade dos residuais do modelo 2



Analisando os gráficos, verifica-se que os residuais se situam próximos da linha da tendência, logo, podemos concluir que há normalidade dos mesmos, estando assim cumprido o pressuposto.

- Homoscedasticidade: Este pressuposto foi verificado através da disposição dos residuais num *scatter plot*:

Gráfico 2 - Scatter plot para análise da homoscedasticidade do modelo 2



Os resultados apresentam heteroscedasticidade, pois a variância segue um padrão decrescente, logo, não é constante, o que viola o pressuposto da homoscedasticidade.

Porém, para se poder afirmar que há de facto heteroscedasticidade, devido a haver poucos dados, foi realizada uma nova aplicação do método, com a divisão de dados treinos e teste a ser feita com a proporção 70:30 respetivamente, aumentando assim os dados de teste.

4.3 Novos resultados

Tabela 6- Resultados da aplicação do método MQ no Modelo 2 com a nova divisão

OLS Regression Results

Dep. Variable:	Receita	R-squared:	0.596
Model:	OLS	Adj. R-squared:	0.555
Method:	Least Squares	F-statistic:	14.74
Date:	Mon, 25 Apr 2022	Prob (F-statistic):	0.000117
Time:	22:44:12	Log-Likelihood:	-242.84
No. Observations:	23	AIC:	491.7
Df Residuals:	20	BIC:	495.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1781.7316	4305.943	-0.414	0.683	-1.08e+04	7200.308
Ticket Médio	3.3518	1.628	2.059	0.053	-0.044	6.748
Tempo de Conversão ND-DG	7488.1615	1674.236	4.473	0.000	3995.766	1.1e+04

Omnibus:	1.538	Durbin-Watson:	1.734
Prob(Omnibus):	0.464	Jarque-Bera (JB):	0.944
Skew:	0.017	Prob(JB):	0.624
Kurtosis:	2.008	Cond. No.	5.18e+03

Correndo novamente a regressão, foram obtidos os resultados representados na Tabela 6. O modelo explica 59,6% (R2), ou 55,5% (R-ajustado) da variação da Receita, sendo estes resultados marginalmente melhores do que os obtidos com a divisão anterior. Quanto aos *p-values* das variáveis, evidencia-se uma subida do *p-value* do Ticket Médio, obtendo um valor de 0.053, porém, apesar desta subida, mantém-se abaixo dos 0.1, logo ainda é estatisticamente significativa, assim como o Tempo de Conversão ND-DG que apresenta um *p-value* de 0. De a cordo com os testes ANOVA, o *F-statistic* subiu ligeiramente, de 14.48 da divisão anterior, para 14.74. Comparando também os valores de AIC e BIC podemos verificar que estes diminuíram em relação à divisão anterior. Analisando os coeficientes obtidos, podemos concluir que a Receita irá variar cerca de 3.5 por cada unidade adicional do Ticket Médio e 7488.16 por cada unidade adicional do Tempo de Conversão ND-DG. O facto de o coeficiente do Tempo de Conversão ND-DG ser positivo é devido aos maiores projetos e vendas da empresa serem também os mais complexos e que necessitam de mais tempo de resolução, logo as maiores vendas estão também associadas a Tempos de Conversão maiores.

4.3.1 Análise de pressupostos

Foi realizada uma nova análise de pressupostos para avaliar o modelo:

- Multicolinearidade:

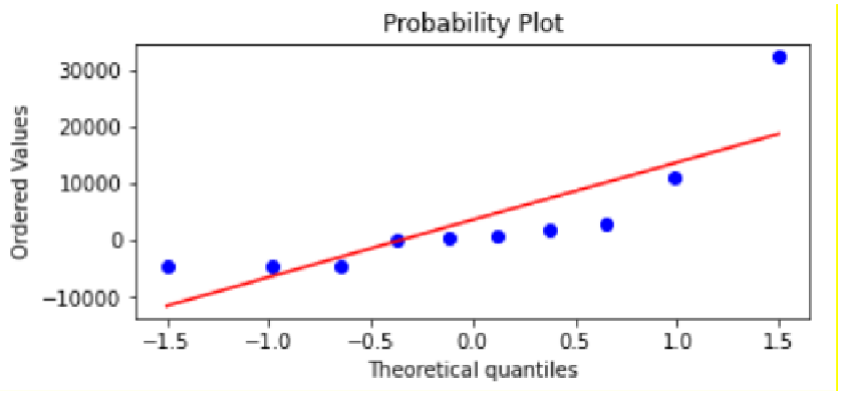
Tabela 7 - Valores VIF do modelo 2 com a nova divisão

	Ticket Médio	Tempo de Conversão ND-DG
vif	2.05	2.05

De acordo com os resultados, sendo ambos os valores inferiores a 10, o pressuposto écumprido, ou seja, não existe multicolinearidade.

- Normalidade dos Residuais:

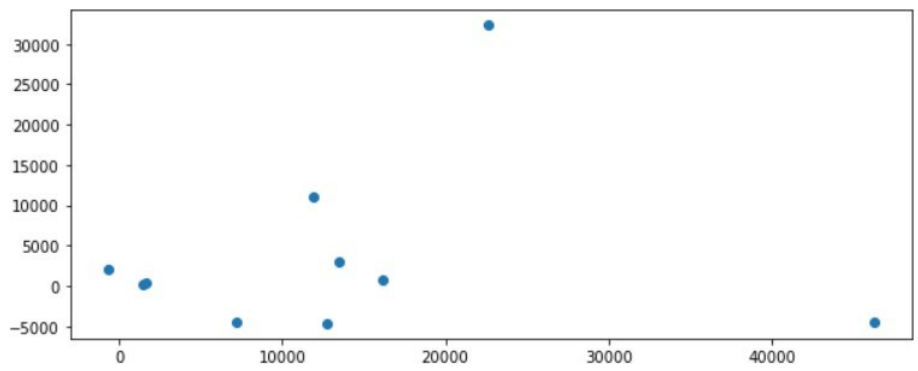
Gráfico 3- Scatter plot para análise da normalidade dos residuais do modelo 2 com a nova divisão



Como é possível verificar, os residuais situam-se em torno da linha de tendência, logo, o pressuposto da normalidade dos residuais é cumprido.

- Homoscedasticidade:

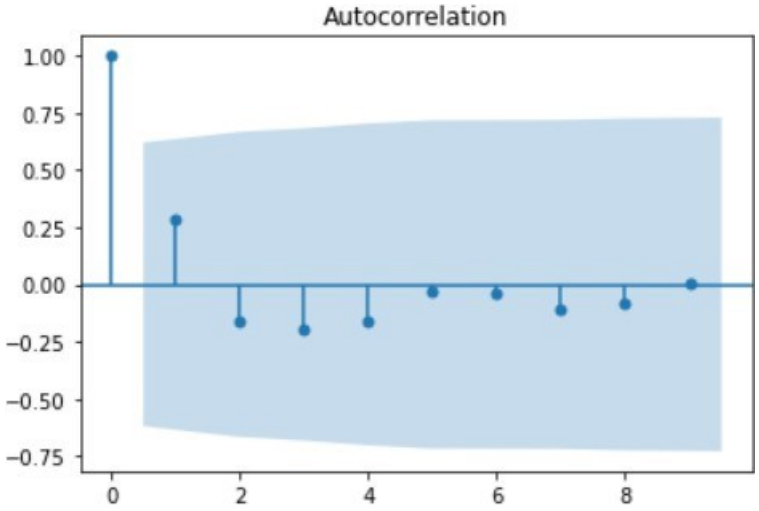
Gráfico 4 - Scatter plot para análise da homoscedasticidade do modelo 2 com a nova divisão



Analisando o gráfico em cima, podemos verificar que os residuais estão dispersos, sem qualquer tendência crescente ou decrescente como no modelo anterior, não apresentando nenhum padrão, podendo-se verificar que existe uma variância dos erros relativamente constante. Podemos assim concluir que existe homoscedasticidade, estando este pressuposto cumprido.

- Autocorrelação dos residuais: Relativamente a este pressuposto, foi utilizadoum gráfico ACF (Auto-Correlation Function):

Gráfico 5 - Gráfico ACG para análise da autocorrelação dos residuais do modelo 2 da nova divisão



Visto que nenhum dos residuais passou os limites a azul, concluímos que nenhuma das autocorrelações são significativamente diferentes de 0, logo, este pressuposto foi também cumprido.

4.3.2 Conclusão dos Pressupostos

Realizada a análise, conclui-se que o modelo cumpre com todos os pressupostos do método deRegressão Linear Múltipla, logo, este modelo é considerado válido.

5. PowerBI Wireframe

Antes da construção das *dashboards* foi criado um *wireframe*, ou um esboço, de como estas iriam ser estruturadas, estando o resultado final disposto nas imagens em baixo. O objectivo deste *wireframe* foi o de validar, junto dos decisores, quais os KPIs a apresentar, assim como a organização dos mesmos por ordem de importância, antesda implementação. Desta forma, alterações podem ser feitas sem um grande acréscimo de tempo.

Figura 1- PowerBI Wireframe Geral

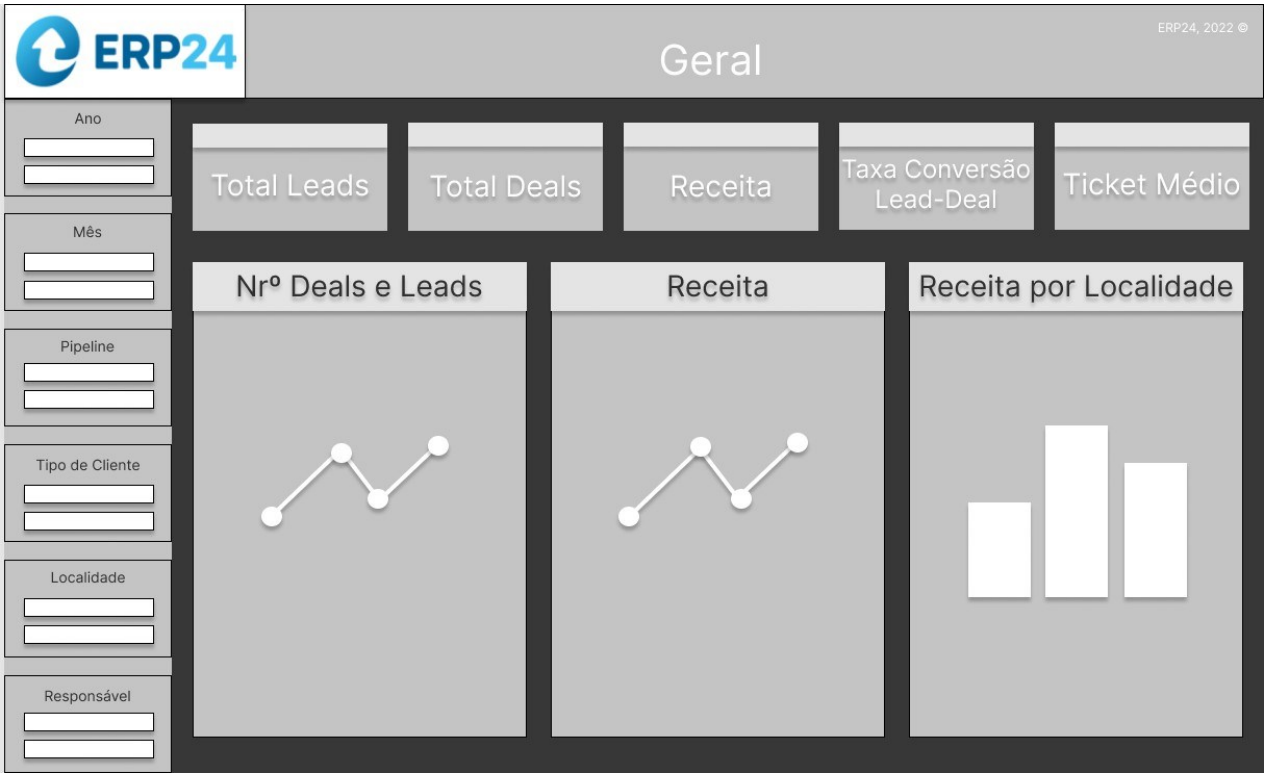


Figura 2- PowerBI Wireframe Deals

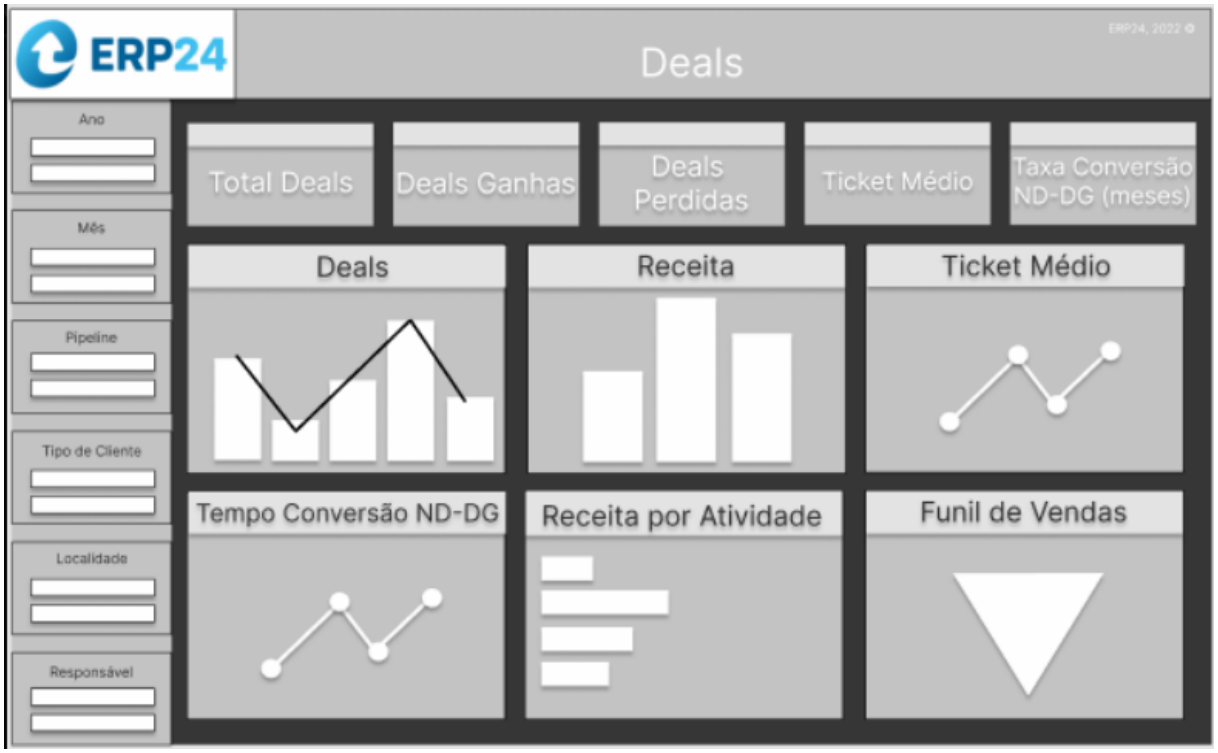
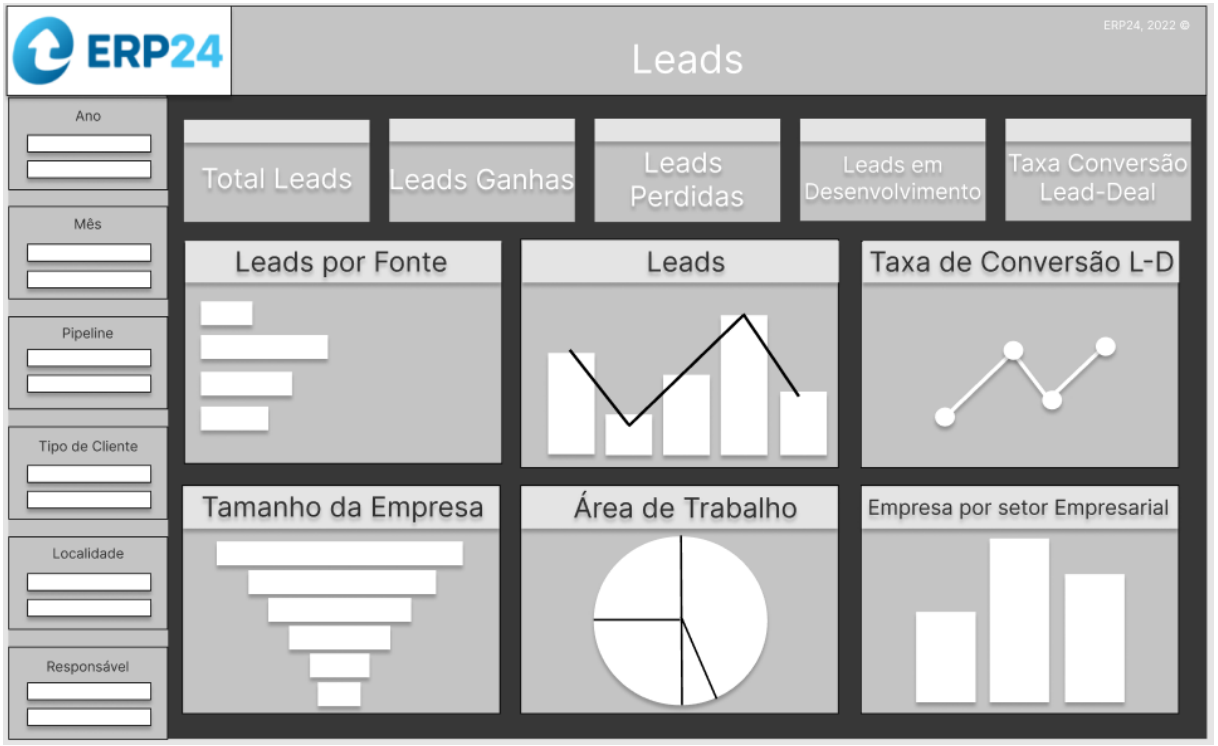


Figura 3 - PowerBI Wireframe Leads



5.1 PowerBI Dashboards

Após criação dos *wireframes* e aprovação, foram criados os seguintes *dashboards*:

Figura 4- PowerBI dasboard Geral

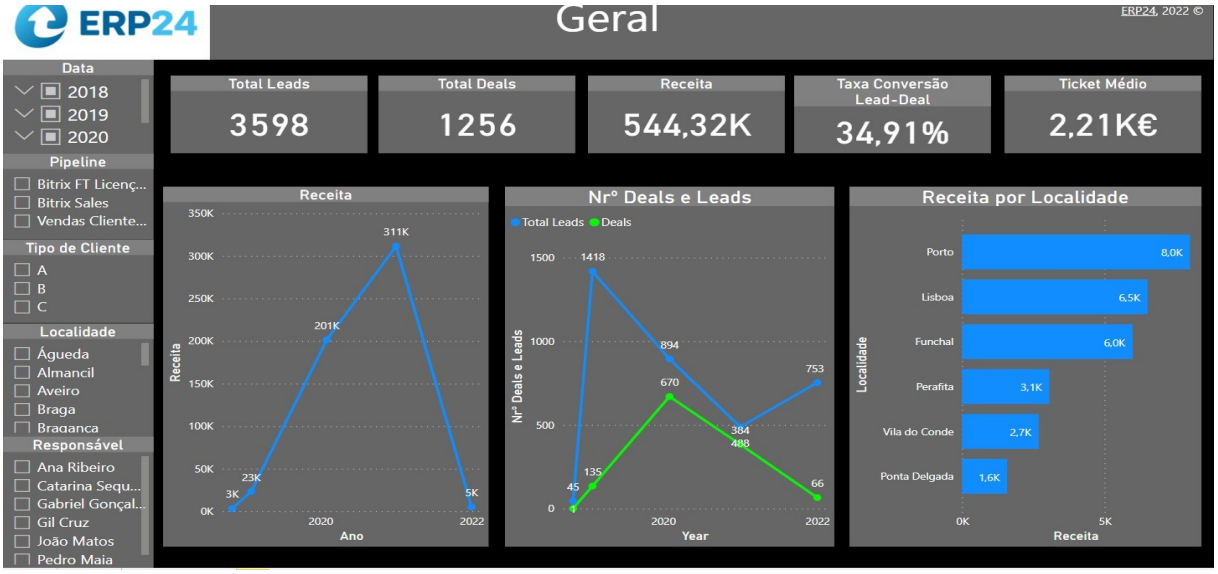


Figura 5 - PowerBI dashboard Leads

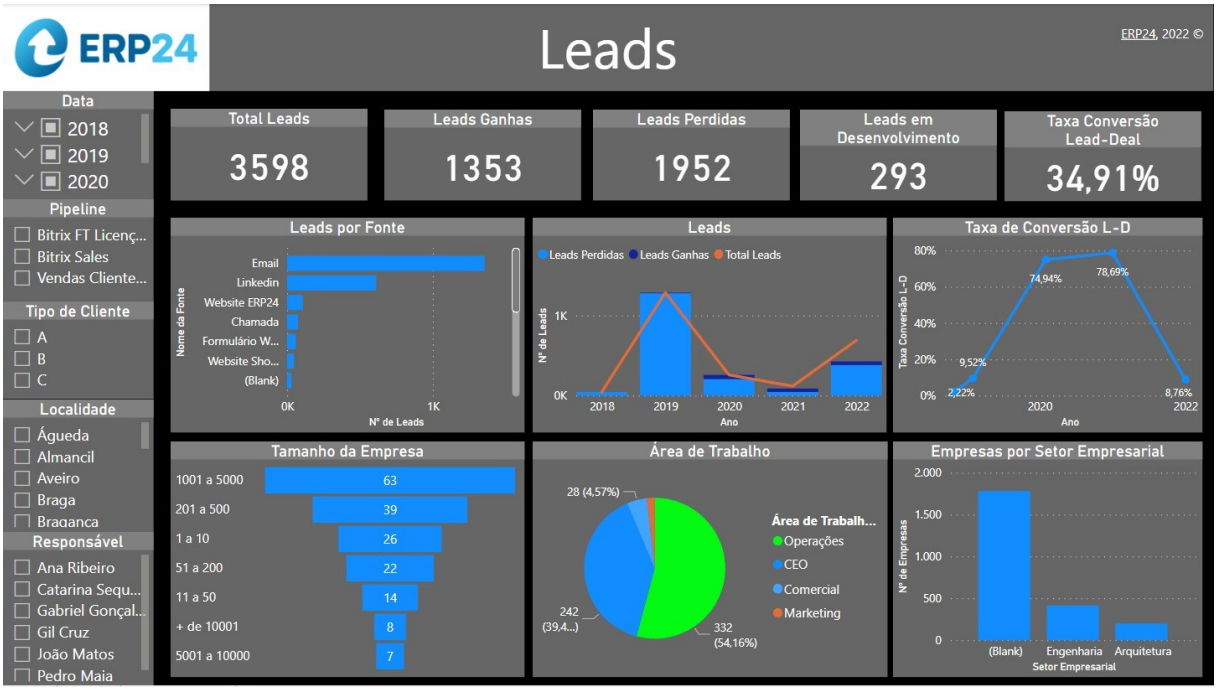


Figura 6 - PowerBI dashboard Deals

5.1.1 Geral

Este *dashboard* foi criada com o intuito de ser uma compilação dos KPIs mais importantes dos dois *dashboards* seguintes, Deals e Leads. Como é possível verificar, do lado esquerdo estão presentes 5 filtros, nomeadamente "Data", "Pipeline", "Tipo de Cliente", "Localidade" e "Responsável", estando estes presentes em todos os *dashboards*. Este *dashboard* é constituída por 5 cartões de informação, onde foram inseridos os indicadores que achei mais relevantes para o panorama geral da empresa, nomeadamente "Total Leads", "Total de Deals", "Receita", "Taxa de Conversão Lead-Deal" e "Ticket Médio". Para uma melhor perceção do histórico, foram criados dois gráficos de continuidade, sendo o primeiro a evolução da Receita nos últimos anos e o segundo a evolução do número de *Leads* e *Deals*. Por fim, foi também criado um gráfico de barras representando a distribuição da receita de empresa a nível geográfico, para uma melhor compreensão dos locais onde a empresa está a tirar mais proveitos. Analisandoos resultados obtidos podemos verificar que a receita da empresa tem vindo a aumentar consistentemente ao longo dos anos, sendo que o valor de 2022 encontra-se baixo devido a ainda não estar o ano todo completo. É possível também concluir que a empresa conseguiu converter 34,91% das *Leads* em *Deals*, sendo que, em média, cada venda gerou 2212 euros. Verifica-se que houve um

enorme crescimento das *Leads* em 2019 e das *Deals* em 2020, que ocorreram devido ao Covid-19, pois as empresas sentiram necessidade de desenvolver os seus meios de comunicação e trabalho digital devido ao teletrabalho, sendo que a discrepância do ano entre as duas variáveis deve- se ao tempo de conversão das *Leads* em *Deals.* Quanto à distribuição por localidade, podemos concluir que a região do Porto é aquela que gera mais receita à empresa.

5.1.2 Leads

Neste *dashboard* foi feita uma análise às *Leads*, também ele constituída por 5 cartões de informação no topo e gráficos em baixo. Os cartões de informação são "Total Leads", "Leads Ganhas", "Leads Perdidas", "Leads em Desenvolvimento" e "Taxa de Conversão Lead-Deal", apresentando assim o panorama geral das *Leads* e os seus diferentes estados, assim como o rácio entre *Deals* e *Leads* na empresa. A nível de representações gráficas, foram criados dois gráficos de continuidade, nomeadamente "Leads" e "Taxa de Conversão L-D", onde é possível observar a evolução histórica do total de *Leads* e respetiva proporção de *Leads* ganhas e perdidas, assim como a evolução da "Taxa de Conversão Lead-Deal" ao longo dos anos. Foram também criados quatro gráficos categóricos, nomeadamente "Leads por Fonte", "Tamanho da Empresa", "Área de Trabalho" e "Empresas por Setor Empresarial", com o objetivo de ter uma melhor perceção da origem das *Leads*, o tamanho das empresas com que a empresa lida, o departamento da pessoa que originou a *Lead* e ainda o setor em que a empresa se insere. De salientar que as variáveis "Tamanho da Empresa", "Área de Trabalho" e "Setor Empresarial" foram criadas recentemente, o que faz com que haja uma grande quantidade de valores nulos, logo foram excluídos dos gráficos para poder ter uma melhor visão dos resultados, tendo permanecido o gráfico "Empresas por Setor Empresarial" com os valores nulos representados para não dar uma imagem errónea de que a ERP24 apenas trabalha com empresas dos setores da "Engenharia" e "Arquitetura".

Analisando os resultados, verifica-se um decréscimo de *Leads* desde 2019 até 2021, sendo que a partir deste ano estas começaram a aumentar. A "Taxa de Conversão L-D" tem vindo a aumentar ao longo dos anos, atingindo em 2021 um valor de 78,69%, ou seja, a empresa conseguiu converter 78,69% das *Leads* em *Deals*, sendo este um valor bastante positivo. Podemos concluir que a maior fonte de *Leads* foi via e-mail, com uma diferença considerável para os restantes, sendo o Linkedin a segunda fonte que mais *Leads* originou. A nível de empresas,

constata-se que mais *Leads* geraram têm entre 1001 a 5000 funcionários e mais de 90% dos representantes de cada empresa ou vêm do departamento de operações ou são o CEO da respetiva empresa.

5.1.3 Deals

Este *dashboard* foi criado com o intuito de analisar as vendas da empresa, sendo também ele constituído por cinco cartões de informação na parte superior e seis gráficos na parte inferior. Os cartões de informação contêm os valores do total de *Deals*, assim como das *Deals* ganhas e perdidas, o que nos permite ter uma perceção do nível de conversão das *Deals* em vendas, assim como do nível de perda *Deals*. Para além disso podemos ver o valor médio que cada *Deal* gera à empresa, assim como a média do tempo que essas *Deals* demoraram a converter-se em vendas. A nível de gráficos a *dashboard* contém quatro contínuos, que nos permitem ter uma visão da evolução dos últimos anos dos indicadores mencionados nos cartões de informação. Para além destes, estão presentes dois gráficos categóricos, que revelam quanto é que cada atividade está a gerar de receita para a empresa, assim como o estado do funil de vendas, mais concretamente o número de *Deals* presentes em cada fase do funil, sendo que este gráfico permite ter uma noção da taxa de conversão ao longo das fasesdas *Deals*.

Analisando os resultados obtidos é possível afirmar que o *ticket* médio sofreu um decréscimo desde 2018 até 2020, tendo depois aumentado em 2021, o que pode ser comprovado pelo aumento da receita que vemos no gráfico de barras e pelo decréscimo de *Deals* a partir de 2020, ou seja, visto que as *Deals* diminuíram e as receitas aumentaram, o valor médio de cada *Deal*, ou *ticket* médio, aumentou também. Verifica-se que a Segurança – Sistemas de Informação e Controlo de Acessos foi a atividade que mais receita gerou para a empresa. Quanto ao tempo de conversão de novas *Deals* em vendas, verifica-se uma tendência de decréscimo desde 2019 atéao presente ano, sendo que entre 2020 e 2021 este indicador manteve-se constante.

6. Conclusão

Foi realizada uma análise preditiva em que foram utilizadas diversas variáveis para a construção do modelo de regressão que tinha como objetivo encontrar as variáveis que mais influenciam na variação da receita da empresa. Devido à limitação de dados não foi possível calcular outras variáveis que poderiam ser úteis à análise, visto que uma grande parte das colunas extraídas continham poucos ou até mesmo nenhuns valores, aliado ao facto de existirem poucas observações para correr a regressão.

A análise realizada leva à conclusão que dentro das variáveis analisadas, aquelas que mais impactam a variação da receita da empresa são o "Ticket Médio" e o "Tempo de Conversão ND-DG", que explicam 55.5% da sua variação. Os resultados mostram que que a receita aumenta 3.35 unidades por cada unidade de "Ticket Médio" adicional e 7488.16 unidades por cada unidade adicional de "Tempo de Conversão ND-DG". Isto significa que a empresa deve procurar obter negócios com Ticket Médio mais elevado e aqueles cujo tempo de conversão é maior, pois são os que originam maior receita. A evidência de que um tempo de conversão maior resulta numa maior receita é contra- intuitivo, mas tem importantes implicações para a empresa: o esforço adicional que é despendido em converter Leads maiores parece justificar-se, pois resulta em receitas superiores. Isto sugere que uma estratégia apenas apostada em converter vendas rápidas (quick wins) não será, no contexto deste sector de actividade, necessariamente a mais indicada. Esta observação é especialmente relevante se os recursos humanos afectos ao esforço comercial forem escassos, implicando isso um elevado custo de oportunidade em cada lead.

Estes resultados vão de encontro ao objetivo da análise que era analisar que variáveis tinham mais impacto sobre as vendas da empresa e se o "Ticket Médio" era uma delas, o que acaba por apoiar a estratégia da empresa de procurar reduzir o número de Leads, aumentando o Ticket Médio de cada venda.

7. Bibliografia

Appelbaum, D., Gogan, A., Vasarhelyi, M. & Zhaoki, Y. (2017). Impact of business analytics and enterprise systems on managerial accounting. International Journal of Accounting Information Systems, 25, 29-44. http://dx.doi.org/10.1016/j.accinf.2017.03.003

Batselier, J., & Vanhoucke, M. (2017). Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. International journal of project management, 35(1), 28-43. http://dx.doi.org/10.1016/j.ijproman.2016.10.003

Bon, A. T., & Ng, T. K. (2017). An optimization of inventory demand forecasting in university healthcare centre. IOP Conference Series: Materials Science and Engineering (Vol. 166, No. 1, p. 012035). IOP Publishing

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140-1154. http://dx.doi.org/10.1016/j.ejor.2006.12.004.

Diamond, M., & Mattia, A. (2017). Data visualization: An exploratory study into the software tools used by businesses. Journal of Instructional Pedagogies, 18.

Fiaz, A. S., Asha, N., Sumathi, D., & Navaz, A. S. (2016). Data visualization: enhancing big data more adaptable and valuable. International Journal of Applied Engineering Research, 11(4), 2801-2804.

Frick, K. D. (2009). Micro-costing quantity data collection methods. Medical care, 47(7 Suppl 1), 76- 81. https://doi.org/10.1097/mlr.0b013e31819bc064

Ghinea, C., Drăgoi, E. N., Comăniţă, E. D., Gavrilescu, M., Câmpean, T., Curteanu, S. & Gavrilescu, M. (2016). Forecasting municipal solid waste generation using prognostic tools and regression analysis. Journal of environmental management, 182, 80-93. http://dx.doi.org/10.1016/j.jenvman.2016.07.026

Gorodov, E. Y. E., & Gubarev, V. V. E. (2013). Analytical review of

data visualization methods in application to big data. Journal of Electrical and Computer Engineering, 2013. https://doi.org/10.1155/2013/969458

Granello, D. H., & Wheaton, J. E. (2004). Online data collection: Strategies for research. Journal of Counseling & Development, 82(4), 387-393. https://doi.org/10.1002/j.1556-6678.2004.tb00325.x

Hansun, S. (2013). A new approach of moving average method in time series analysis. 2013 conference on new media studies, 1-4. IEEE. http://dx.doi.org/10.1109/conmedia.2013.6708545

Hofmann, E., & Rutschmann, E. (2018). Big data analytics and demand forecasting in supply chains: a conceptual analysis. The International Journal of Logistics Management,29(2),739-766. http://dx.doi.org/10.1108/IJLM-04-2017-0088

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary vs. secondary. Encyclopedia of social measurement, 1(1), 593-599.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. Kotsialos, A., Papageorgiou, M., & Poulimenos, A. (2005). Long-term sales forecasting using holt– winters and neural network methods. Journal of Forecasting, 24(5), 353-368. https://doi.org/10.1002/for.943

Lassen, N. B., Madsen, R., & Vatrapu, R. (2014). Predicting iphone sales from iphone tweets. 2014 IEEE 18th International Enterprise Distributed Object Computing Conference (pp. 81-90). IEEE. http://dx.doi.org/10.1109/EDOC.2014.20

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9, 381-386. http://dx.doi.org/10.21275/ART20203995

Montgomery, D. C., Peck, A.E., Vining, G.G. (2012). Introduction to Linear Regression Analysis. Wiley Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.

Nazim, A., & Afthanorhan, A. (2014). A comparison between single

exponential smoothing (SES), double exponential smoothing (DES), holt's (brown) and adaptive response rate exponential smoothing (ARRES) techniques in forecasting Malaysia population. Global Journal of Mathematical Analysis, 2(4), 276-280. https://doi.org/ 10.14419/gjma.v2i4.3253

Nirmala, V. W., Harjadi, D., & Awaluddin, R. (2021). Sales Forecasting by Using Exponential Smoothing Method and Trend Method to Optimize Product Sales in PT. Zamrud Bumi Indonesia During the Covid-19 Pandemic. International Journal of Engineering, Science and Information Technology, 1(4), 59-64. https://doi.org/10.52088/ijesty.v1i4.169

Olive, D. J. (2017). Multiple linear regression. In Linear regression (pp. 17-83). Springer, Cham. https://doi.org/10.1007/978-3-319-55252-1_2

Ramakrishnan, T., Jones, M. C., & Sidorova, A. (2012). Factors influencing business intelligence (BI) data collection strategies: An empirical investigation. Decision support systems, 52(2), 486-496. https://doi.org/10.1016/j.dss.2011.10.009

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. International Journal of Engineering Research And Advanced Technology (IJERAT), 2(12), 11-16.

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. Journal of Statistics Education, 9(3). https://doi.org/10.1080/10691898.2001.11910537

Uyanik, G. & Guler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences, 106, 234-240. https://doi.org/10.1016/j.sbspro.2013.12.027

8. Apêndice

```
#Importar dados e bibliotecas

import pandas as pd
from matplotlib import pyplot as plt
import scipy as sp
import seaborn as sns
from statsmodels.formula.api import ols
import statsmodels.api as sm
import statsmodels.tsa.api as smt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from statsmodels.stats.outliers_influence import
variance_inflation_factor
from sklearn import linear_model
import json
import requests
import numpy as np
from scipy.stats import pearsonr
import numpy.ma as ma
from datetime import datetime
from sklearn.metrics import mean_squared_error

d = open('Deals1.json',encoding="utf8" )
j = json.load(d)

r = open('Leads1.json',encoding="utf8" )
l = json.load(r)

#Criar datafranes df e df2
df1 = pd.DataFrame.from_dict(j)
df2=pd.DataFrame.from_dict(l)


import warnings
warnings.filterwarnings("ignore")

#Mostrar colunas todas
pd.options.display.max_columns = None

#Extração dos dados apenas de Deals que pertençam aos funis de vendas 2 ,
32 e 73


df = df1[(df1['CATEGORY_ID'] == "2") | (df1['CATEGORY_ID'] == "32") |
(df1['CATEGORY_ID'] == "73")]

#Substituição dos valores NaN da coluna LEAD_ID por "n.d."

df['LEAD_ID']=df['LEAD_ID'].fillna('n.d.')
```

```
#Alterar o tipo de dados das colunas


df["OPPORTUNITY"] = df["OPPORTUNITY"].astype(float)
df['DATE_CREATE'] = pd.to_datetime(df['DATE_CREATE'],utc=True)
df['CLOSEDATE'] = pd.to_datetime(df['CLOSEDATE'],utc=True)


# Criação das colunas "Deals Ganhas", "Deals Perdidas", "Deals
Contagem","Deals sem Lead","Tempo diff","Tempo de Conversão ND-DG" e
"Receita"

df['Deals Ganhas'] = ["1" if  x == "C2:WON" or x == "C32:WON" or  x
=="C73:WON"  else '0' for x in df['STAGE_ID']]
df['Deals Perdidas'] = ["1" if x == 'C2:LOSE' or x == "C2:APOLOGY" or x
== "C2:3" or x == "C2:12" or x == "C2:10" or x == "C2:15" or x == "C2:4"
or x =="C32:APOLOGY" or x =="C32:10" or x =="C32:11" or x == "C32:LOSE"
or x =="C73:LOSE" or x== "C73:7" or x =="C73:APOLOGY" or x == "C73:8"
else '0' for x in df['STAGE_ID']]
df['Deals Contagem'] = ["1" if x =="2" or x == "32" or x == "73"  else
'0' for x in df['CATEGORY_ID']]
df['Retornado'] = ["1" if x =="Y"  else '0' for x in
df['IS_RETURN_CUSTOMER']]

deals_sem_lead = []

for _,x in df.iterrows():
    if x['Deals Contagem'] == "0":
        deals_sem_lead.append("0")
    else:
        if x['LEAD_ID'] == "n.d.":
            deals_sem_lead.append("1")
        else:
            deals_sem_lead.append("0")

df['Deals sem Lead'] = deals_sem_lead


df['Receita'] = [ "n.d." if x['Deals Ganhas'] == "0" else
x['OPPORTUNITY'] for _, x in df.iterrows()]


from datetime import datetime

def diff_month(d1, d2):
    return (d1.year - d2.year) * 12 + d1.month - d2.month
```

```
df['Tempo diff'] = [ 0 if diff_month(x['CLOSEDATE'], x['DATE_CREATE']) <=
0 else diff_month(x['CLOSEDATE'], x['DATE_CREATE']) for _, x in
df.iterrows()]
df['Tempo de Conversão ND-DG']= [ "n.d." if x['Deals Ganhas'] == "0" else
x['Tempo diff'] for _, x in df.iterrows()]




#Substituição dos valores "n.d." por NaN

df['Tempo de Conversão ND-DG'].replace("n.d.",np.NaN, inplace=True)
df['Receita'].replace("n.d.",np.NaN, inplace=True)

#Conversão das colunas "Deals Ganhas", "Deals Perdias","Deals Contagem" e
"Deals sem Lead"

df['Deals Ganhas'] = df['Deals Ganhas'].apply(pd.to_numeric)
df['Deals Perdidas'] = df['Deals Perdidas'].apply(pd.to_numeric)
df['Deals Contagem'] = df['Deals Contagem'].apply(pd.to_numeric)
df['Deals sem Lead'] = df['Deals sem Lead'].apply(pd.to_numeric)
df['Retornado'] = df['Retornado'].apply(pd.to_numeric)


#Criar datasets test_df e test2_Df


df['DATE_CREATE'] = pd.to_datetime(df['DATE_CREATE'],utc=True)
df['year'] = pd.DatetimeIndex(df['DATE_CREATE']).year
df['month'] = pd.DatetimeIndex(df['DATE_CREATE']).month

test_df = df[['Receita',"Deals Perdidas" ,"Deals Contagem",'Deals
Ganhas','Deals sem Lead','year', 'month','Retornado']]
test2_df = df[['Receita', 'Tempo de Conversão ND-DG','year','month' ]]

#Criação de um dataframe g correspondente ao dataframe test_df agrupadas
por soma, mês e ano
#Criação da coluna "date" constituída pelas colunas "year" e "month", e
utilização da mesma como índice

g = test_df.groupby([(test_df.year), (test_df.month)],
as_index=False).sum()
g['date'] = pd.to_datetime(g[['year', 'month']].assign(DAY=1))
g.drop(['year', 'month'], inplace=True, axis=1)
g.reset_index(drop=True)
g = g.set_index('date')
```

```
#Criação um dataframe t constituído pelas colunas do dataframe test2_df
agrupadas pela média, mês e ano

t = test2_df.groupby([(test2_df.year), (test2_df.month)],
as_index=False).mean()

#Criação da coluna "date" constituída pelas colunas "year" e "month", e
utilização da mesma como índice

t['date'] = pd.to_datetime(t[['year', 'month']].assign(DAY=1))
t.drop(['year', 'month'], inplace=True, axis=1)
t.reset_index(drop=True)
t = t.set_index('date')

# Criação da coluna Ticket Médio correspondente á coluna Receita (que foi
agrupada pela média), e posterior remoção da coluna "Receita"
t['Ticket Médio'] = t['Receita']
t.drop(['Receita'], axis = 1, inplace = True)

#Converter coluna "DATE_CREATE" para datetime


df2['DATE_CREATE'] = pd.to_datetime(df2['DATE_CREATE'],utc=True)



#Criação das colunas "year" e "month" e dataframe test_df2 constituído
pelas colunas "ID", "year" e "month" do dataframe df2


df2['year'] = pd.DatetimeIndex(df2['DATE_CREATE']).year
df2['month'] = pd.DatetimeIndex(df2['DATE_CREATE']).month
test_df2 = df2[['ID', 'year', 'month']]

#Criação do dataframe l agrupando o dataframe test_df2 fazendo contagem
por mês e ano

l = test_df2.groupby([(test_df2.year), (test_df2.month)],
as_index=False).count()

#Criação da coluna "date" constituída pelas colunas "year" e "month", e
utilização da mesma como índice
l['date'] = pd.to_datetime(l[['year', 'month']].assign(DAY=1))
l.drop(['year', 'month'], inplace=True, axis=1)
l.reset_index(drop=True)
l = l.set_index('date')

#Criação da coluna Ticket Médio e Tempo de Conversão ND-DG no dataset g
correspondentes aos mesmos do dataset t
```

```
g = g.join(pd.Series(t['Ticket Médio']).rename('Ticket Médio'),
how='right')
g = g.join(pd.Series(t['Tempo de Conversão ND-DG']).rename('Tempo de
Conversão ND-DG'), how='right')

#Criação da coluna leads no dataset g correspondente á coluna ID no
dataset t

g = g.join(pd.Series(l['ID']).rename('leads'), how='right')

# Preenchimento dos valores nulos das Deals sem Lead com 0, para
realização do cálculo a seguir

g['Deals sem Lead'] = g['Deals sem Lead'].fillna(0)


#Criação da coluna Deals no dataset g, correspondente á soma das Deals
sem Lead e leads

g['Leads'] = g['Deals sem Lead'] + g['leads']

#Criação da coluna Taxa de Conversão, correspondente á conversão Leads
para Deals

Taxa_de_Conversão_L_D = []
for _,x in g.iterrows():
    if x['Deals Contagem'] == "0":
        Taxa_de_Conversão_L_D.append("0")
    else:
        Taxa_de_Conversão_L_D.append((x['Deals
Contagem']/x['Leads'])*100)


g['Taxa de Conversão L-D'] =   Taxa_de_Conversão_L_D


#Remoção dos valores nulos do dataset g

g = g.dropna( axis = 0)

#Representar graficamente os dados


g['Receita'].plot(title="Opportunity data")

#Formatar colunas para apresentarem apenas 2 casas decimais

pd.options.display.float_format = '{:.2f}'.format
```

```
#Remoção das colunas que não serão usadas como variáveis no modelo

g.drop([ 'Deals Contagem','Deals sem Lead',
'Leads','leads','Retornado','Deals Ganhas'], axis = 1, inplace = True)

#Matriz de correlação com a significância em asteriscos:


rho = g.corr()
pval = g.corr(method=lambda x, y: pearsonr(x, y)[1]) - np.eye(*rho.shape)
p = pval.applymap(lambda x: ''.join(['*' for t in [0.01,0.05,0.1] if
x<=t]))
rho.round(2).astype(str) + p

                                                    #MODELO 1

#Definição do y e x do modelo 1

x = g[['Ticket Médio','Tempo de Conversão ND-DG','Deals Perdidas']]
y = g['Receita']

# Dividir a amostra em treino e teste, sendo que 80% foi usada para
treino
x_train, x_test, y_train, y_test = train_test_split(x, y,
                                                    test_size=0.20
,shuffle = False , random_state=0)

#Estimação dos coeficientes usando método dos Mínimos Múltiplos Quadrados

x_with_constant = sm.add_constant(x_train)
model = sm.OLS(y_train, x_with_constant)

#Coeficientes resultantes do método

results = model.fit()
results.params

#Print dos resultados

print(results.summary())

                                                    #MODELO 2

#Definição do y e x do modelo 2

x2 = g[['Ticket Médio','Tempo de Conversão ND-DG']]
y2 = g['Receita']
```

```
# Dividir a amostra em treino e teste, sendo que 80% foi usada para
treino e 20% para teste
x2_train, x2_test, y2_train, y2_test = train_test_split(x2, y2,
                                                        test_size=0.20,
shuffle = False, random_state = 0)

#Estimação dos coeficientes usando método dos Mínimos Múltiplos Quadrados

x2_with_constant = sm.add_constant(x2_train)
model2 = sm.OLS(y2_train, x2_with_constant)

#Coeficientes resultantes do método

results2 = model2.fit()
results2.params

#Print dos resultados

print(results2.summary())

#Adicionar constante a x2_test

x2_test = sm.add_constant(x2_test)

# Variável y_pred como sendo a previsão do método usando a variável
x2_test

y2_pred = results2.predict(x2_test)

# Cálculo dos residuais

residual2 = y2_test - y2_pred

#Verificar multicolinearidade
vif2 = [variance_inflation_factor(x2_train.values, i) for i in
range(x2_train.shape[1])]
pd.DataFrame({'vif': vif2[0:]}, index=x2_train.columns).T

#Disposição dos residuais num scatter plot para verificação da
normalidade

fig, ax = plt.subplots(figsize=(6,2.5))
_, (__, ___, r) = sp.stats.probplot(residual2, plot=ax, fit=True)

#Verificar Homoscedasticidade


fig, ax = plt.subplots(figsize=(10,4))
_ = ax.scatter(y2_pred, residual2)
```

```
                                                    #Modelo 2 com nova
divisão

# Dividir a amostra em treino e teste, sendo que 70% foi usada para
treino e 30% para teste
x2_train, x2_test, y2_train, y2_test = train_test_split(x2, y2,
                                                        test_size=0.30,
shuffle = False, random_state = 0)

#Estimação dos coeficientes usando método dos Mínimos Múltiplos Quadrados

x2_with_constant = sm.add_constant(x2_train)
model2 = sm.OLS(y2_train, x2_with_constant)

#Coeficientes resultantes do método

results2 = model2.fit()
results2.params

#Print dos resultados

print(results2.summary())

#Adicionar constante a x2_test

x2_test = sm.add_constant(x2_test)

# Variável y_pred como sendo a previsão do método usando a variável
x2_test

y2_pred = results2.predict(x2_test)

# Cálculo dos residuais

residual2 = y2_test - y2_pred

#Verificar multicolinearidade
vif2 = [variance_inflation_factor(x2_train.values, i) for i in
range(x2_train.shape[1])]
pd.DataFrame({'vif': vif2[0:]}, index=x2_train.columns).T

#Disposição dos residuais num scatter plot para verificação da
normalidade

fig, ax = plt.subplots(figsize=(6,2.5))
_, (__, ___, r) = sp.stats.probplot(residual2, plot=ax, fit=True)

#Verificar Homoscedasticidade
```

```
fig, ax = plt.subplots(figsize=(10,4))
_ = ax.scatter(y2_pred, residual2)

#Verificar autocorrelação dos residuais


acf2 = smt.graphics.plot_acf(residual2, alpha=0.05)
acf2.show()

# Criação e aplicação do método de Regressão Linear



reg2 = linear_model.LinearRegression()
reg2.fit(x2,y2)

# Disposição dos coeficientes do método

reg2.coef_

# Disposição da constante de interseção

reg2.intercept_

g.head(100)

#Previsão

reg2.predict([[417.25,0.25]])

                                                    #Estatística Descritiva

#Receita

#Média
print("Média ", g['Receita'].mean())

#Mediana
print( "Mediana -", g['Receita'].median())

#Desvio Padrão

print("Desvio Padrão -", g['Receita'].std())

#25%

print("25% -", g['Receita'].quantile(0.25))
```

```
#75%

print ("75% -", g['Receita'].quantile(0.75))

#Min

print("Min -", g['Receita'].min())

#Máx

print("Máx -", g['Receita'].max())



#Ticket Médio

#Média
print("Média -", g['Ticket Médio'].mean())

#Mediana
print( "Mediana -", g['Ticket Médio'].median())

#Desvio Padrão

print("Desvio Padrão -", g['Ticket Médio'].std())

#25%

print("25% -", g['Ticket Médio'].quantile(0.25))

#75%

print ("75% -", g['Ticket Médio'].quantile(0.75))

#Min

print("Min -", g['Ticket Médio'].min())

#Máx

print("Máx -", g['Ticket Médio'].max())

#Taxa de Conversão L-D

#Média
print("Média ", g['Taxa de Conversão L-D'].mean())

#Mediana
```

```
print( "Mediana -", g['Taxa de Conversão L-D'].median())

#Desvio Padrão

print("Desvio Padrão -", g['Taxa de Conversão L-D'].std())

#25%

print("25% -", g['Taxa de Conversão L-D'].quantile(0.25))

#75%

print ("75% -", g['Taxa de Conversão L-D'].quantile(0.75))

#Min

print("Min -", g['Taxa de Conversão L-D'].min())

#Máx

print("Máx -", g['Taxa de Conversão L-D'].max())

#Tempo de Conversão ND-DG

#Média
print("Média ", g['Tempo de Conversão ND-DG'].mean())

#Mediana
print( "Mediana -", g['Tempo de Conversão ND-DG'].median())

#Desvio Padrão

print("Desvio Padrão -", g['Tempo de Conversão ND-DG'].std())

#25%

print("25% -", g['Tempo de Conversão ND-DG'].quantile(0.25))

#75%

print ("75% -", g['Tempo de Conversão ND-DG'].quantile(0.75))

#Min

print("Min -", g['Tempo de Conversão ND-DG'].min())

#Máx

print("Máx -", g['Tempo de Conversão ND-DG'].max())
```

```
#Deals Perdidas

#Média
print("Média ", g['Deals Perdidas'].mean())

#Mediana
print( "Mediana -", g['Deals Perdidas'].median())

#Desvio Padrão

print("Desvio Padrão -", g['Deals Perdidas'].std())

#25%

print("25% -", g['Deals Perdidas'].quantile(0.25))

#75%

print ("75% -", g['Deals Perdidas'].quantile(0.75))

#Min

print("Min -", g['Deals Perdidas'].min())

#Máx

print("Máx -", g['Deals Perdidas'].max())
```