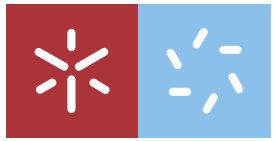
Previsão E-commerce: Indicadores de Desempenho por Canal

Ana Patrícia Graça Gonçalves

UMinho | 2019

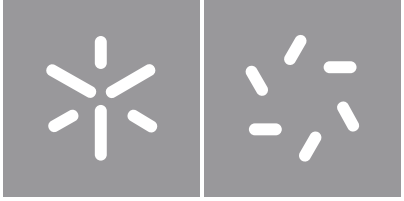


Universidade do Minho
Escola de Ciências

Ana Patrícia Graça Gonçalves

Previsão E-commerce: Indicadores de Desempenho por Canal

outubro de 2019



Universidade do Minho
Escola de Ciências

Ana Patrícia Graça Gonçalves

Previsão E-commerce: Indicadores de Desempenho por Canal

Tese de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Ana Paula Conceição Amorim
e da
Dr.ª Liliana Sofia Oliveira Martins

outubro de 2019

Agradecimentos

Agradeço a todas as pessoas que me acompanharam na vida e tornaram possível a chegada a este momento. Um especial agradecimento aos meus pais e á minha irmã pelo apoio incondicional, e acreditarem que sou capaz de encontrar todas as respostas aos meus problemas. Este trabalho só foi possível com a colaboração da Overcube que disponibilizou os dados e todas as condições para este projeto e, por isso, agradeço a todas as pessoas com quem entrei em contacto, naquela organização e pelo apoio que me deram.

Destaco a Dra. Liliana Martins que me ajudou a integrar com os restantes membros de trabalho e pela disponibilidade diária para qualquer esclarecimento, pela paciência e pela partilha dos seus conhecimentos acerca de aparelhos que nunca tinha entrado em contacto. Agradeço também à professora Ana Paula Amorim pelo seu acompanhamento, disponibilidade e atenção aos pormenores que demonstrou ao longo do estágio, assim como todos os restantes docentes do mestrado pelos ensinamentos e paciência nesta etapa.

Quero agradecer à minha tia Adelaide por me ter acompanhado nesta etapa, e me ter ajudado sempre que possível. Ao Diogo Macedo, um obrigada especial, por todo o incentivo e compreensão ao longo destes meses. Não me posso esquecer dos meus amigos e colegas que me apoiaram em momentos de maior dificuldade e com quem partilhei estes últimos dois anos.

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.

Resumo

Com o avanço da internet, muitas empresas adotaram o comércio eletrónico, a partir de uma ferramenta básica de comunicação num mercado interativo de produtos e serviços. Iniciou-se uma análise do modelo de negócio da empresa Overcube para compreender melhor os principais indicadores de desempenho deste negócio. Esta análise serviu para depois prever mensalmente os resultados por canal. O objetivo é o desenvolvimento de modelos de previsão para o custo e a receita em tempo real.
No presente estudo, a pedido pela Overcube, foram tratados os dados disponíveis em duas bases de dados. Estas base de dados, Facebook Ads e Google Ads, foram devidamente divididas e tratadas de modo a permitir uma análise mais eficiente, para a previsão do custo e da receita. Os KPIs ("*Key Performance Indicator*") de previsão podem ajudar essas empresas de comércio electrónico a terem a percepção do impacto das suas ações para garantir que as metas de negócios sejam alcançadas.
Os dois canais utilizados foram: Facebook e o Google Adwords. Estes canais fornecem métricas relacionadas com o número de cliques e as impressões do cliente, mas a Overcube tem como principais KPIs os cliques (*clicks*), impressões (*impressions*), alcance (*reach*), conversões (*conversions*), adições ao carrinho (*add to cart*), número de resultados (*results*), custo por resultado (CPR), compromisso (*engagement*), CTR, CPC, CPA, ROAS, CR, CPM e o AOV. Estas são as métricas importantes usadas pela Overcube para indicar o seu progresso em direção a uma meta de negócios definida.
O principal objetivo desta tese foi melhorar a previsão de KPIs usando dados históricos. A abordagem metodológica usada consistiu na análise da regressão múltipla e na análise de séries temporais (modelo ARIMA).

O objetivo deste trabalho, foi adquirir conhecimento suficiente do negócio, para compreender quais os indicadores principais de performance. E, desta forma, prever por canal quais seriam os resultados esperados. Pretende-se desenvolver modelos de previsão para o custo e a receita em tempo real.

O principal objetivo desta tese foi melhorar a previsão de KPIs usando dados históricos. A abordagem metodológica usada consistiu na análise da regressão múltipla e na análise de séries temporais (modelo ARIMA). Através da análise das bases de dados, verificou-se qua a nível das séries temporais a informação era escassa e a amostra era de dimensão reduzida. Dos modelos finais de previsão obtidos, os que apresentam melhores resultados são os da regressão múltipla.

Abstract

With the advancement of the internet, many companies have adopted e-commerce from a basic communication tool in an interactive market for products and services. An analysis of Overcube's business model was started to better understand the key performance indicators of this business. This analysis then served to view monthly results by channel. The objective of this internship is to develop real-time cost and revenue forecasting models. In the present study, an Overcube requests, were available at two databases. These databases, Facebook Ads and Google Ads, have been properly split and treated to allow more efficient analysis for cost and revenue prediction. Forecasting *Key Performance Indicator* (KPIs) can help these e-commerce companies realize the impact of their actions to ensure that business goals are reached. The two channels used were: Facebook and Google Adwords. These channels provide metrics related to the numeber of clicks and impressions, but Overcube's primary KPIs are *clicks*, *impressions*,*range*, *conversions*, *add to cart*, number of results, cost per result (CPR), *engagement*, CTR, CPC, CPA, ROAS, CR, CPM and AOV. These are the important metrics used by Overcube to indicate the progress in direction to a defined business goal. The main objective of this thesis was to improve KPI prediction using historical data. The methodological approach used consisted on a multiple regression analysis and time series analysis (ARIMA model). The purpose of this internship was to acquire sufficient knowledge of the business to understand the main performance indicators. And, in this way, predict monthly and by channel what would be the expected results. It is intended to develop forecast models for real-time cost and revenue. The main objective of this thesis was to improve KPI prediction using historical data.

The methodological approach used, consisted of multiple regression analysis and time series analysis (ARIMA model).Trough the database analysis, it was found that at the time series level, information was scarce and the sample was small sized. From the final prediction models obtained, the ones with the best results are the multiple regression.

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

KPI Key Performance Indicator.
Spend Custo do Facebook.
C.V. Conversion Value Receita do Facebook.
Cost VF Custo do Google.
T.C.V. Total Conversion Value Receita do Google.
CPR Cost per Result.
CTR Taxa de Cliques.
CPC Custo por Clique.
CPA Custo por Aquisição.
ROAS Retorno Gasto com Publicidade.
CR Taxa de Conversão.
CPM Custo por Mil.
AOV Valor Médio do Pedido.
ACF Função de Autocorrelação.
PACF Função de Autocorrelação Parcial.
RLS Regressão Linear Simples.
RLM Regressão Linear Múltipla.
SSE standard error of estimate.
FA Facebook Adwords.
GA Google Adwords.
VIF Variance Inflation Factor.
ARIMA Autoregressive Integrated Moving Average.
ACP Análise de Componentes Principais.
AIQ Amplitude Interquartil.
DSA Dynamic Search Ads.
LRT Teste da Razão de Verosimilhança.

1. Introdução

Com o avanço da internet, muitas empresas em todo o mundo tentam abraçar o comércio eletrónico, a partir de uma ferramenta básica de comunicação num mercado interativo de produtos e serviços. KPI é a sigla que corresponde a "Key Performance Indicator", uma técnica de gestão conhecida em português como Indicador-chave de Desempenho e são medidas quantificáveis para compreender se os objetivos estão a ser atingidos. Consequentemente, esses indicadores determinam se é preciso tomar atitudes diferentes que melhorem os resultados atuais. Os indicadores-chave de desempenho só devem ser alterados se os objetivos primários de uma empresa também sofrer alteração. Existem outras medidas que servem de base para a constituição de um indicador, chama-se de métricas, normalmente associadas ao comportamento do usuário e sem uma meta definida.

Ao rastrear e medir esses indicadores, a recomendação de melhoria operacional pode ser feita com base nos dados reais. Existem muitos KPIs onde alguns dos quais são tráfego do website, taxa de conversão, vendas e receita. No campo dos KPIs, muita pesquisa foi feita para encontrar fatores que os afeta e como terão impacto nas vendas. No entanto, para previsão de KPIs, a maioria da pesquisa concentrou-se apenas na previsão do custo e da receita das vendas.

A previsão dos KPIs ajudará os profissionais de marketing de comércio eletrónico a recuperar *insights* o mais rápido possível, o que é importante no mundo competitivo e concorrido do comércio eletrónico. E, desta forma, prever mensalmente por canal quais seriam os resultados esperados, pretende-se desenvolver ferramentas para que a empresa as possa aplicar em tempo real.

Modelos de previsão são aplicados em áreas de finanças, econômicas, meteorológicas, produção de energia, sociologia entre outras. Entre as diversas técnicas de previsão existem as que são baseadas em modelos estatísticos e matemáticos. Estes modelos utilizam dados históricos que são estudados a fim de se identificar os seus comportamentos e padrões para que sejam traçadas projeções futuras com base nos mesmos. Pretende-se identificar os Modelos de previsão mais adequados, procedendo-se ao seu ajustamento e comparação em termos da sua capacidade explicativa e preditiva.

1.1. Descrição da Empresa

A Overcube é um Marketplace Global com o objetivo de promover uma nova inspiradora forma de comprar online. Formada por várias empresas do sector do calçado, a

plataforma empresarial visa o desenvolvimento do sector ao nível das vendas e da qualidade.

Esta plataforma associa um conjunto de fabricantes nacionais, tendo em vista explorar o mundo online para chegar o mais longe possível, a cada cliente. Reforça a capacidade estratégica e a capacidade de venda e distribuição das empresas, foi lançada a 28 de Março de 2018, em São João de Ponte, Guimarães.

1.2. Canais em Estudo

No marketing tradicional, anúncios em jornais e revistas são uma maneira comum de divulgar um negócio e promover vendas. A nível do marketing digital, temos um elemento similar : os Ads, ou seja, anúncios online. O Facebook Ads e Google Adwords, são conhecidos como exemplos de ”mídia paga”, são canais usados para divulgar ads. Os recurso que eles oferecem e os resultados que nos garantem são os que lhe dão pódio graças á sua abrangência. A Overcube tem como principais KPIs as seguintes variáveis.

(1) Custo (*spend*);
(2) Cliques (*clicks*);
(3) Impressões (*impressions*): A métrica de impressões não é necessariamente uma indicação do desempenho do anúncio, mas mostra quantas pessoas realmente o veem;
(4) Alcance (*reach*);
(5) Receita (*Conversion Value*);
(6) Conversões (*Conversions*);
(7) Adições ao Carrinho (*Add to Cart*);
(8) Resultados (*Results*): Este é o número de vezes que o anúncio alcançou o objetivo definido. É uma métrica interessante para compararmos o desempenho entre campanhas semelhantes;
(9) Compromisso *("engagement")*: Grau de compromisso dos usuários em relação à sua marca. É o número de comentários, *retweets*, menções, *likes*, favoritos e todos os sinais sociais que se traduzem em interações com a sua marca;
(10) Custo por Resultado *("Cost per Result - CPR")*: Indica o custo médio por resultado do anúncio;

$$CPR = \frac{Receita}{Resultados}$$

(11) Taxa de Cliques *("CTR")*: mede o número de cliques que são feitos em relação ao número de impressões. Quanto mais elevado for o CTR mais eficaz é a campanha;

$$CTR = \frac{NumeroCliques}{NumeroImpressoes} \times 100$$

(12) Custo por Clique *("CPC")*: Custo por clique, será cobrado cada vez que o anúncio receber um clique;

$$CPC = \frac{CustoTotal}{NumeroCliques}$$

(13) Custo por Ação/Aquisição *("CPA")*: Indica quanto será cobrado por cada conversão (podendo ser uma compra, uma inscrição, um download, etc);

$$CPA = \frac{CustoTotal}{NumeroConversoes}$$

(14) Retorno no gasto com publicidade *("ROAS")*: mede quanto dinheiro em receita se recebe por cada euro gasto em publicidade, busca conhecer o retorno sobre o investimento publicitário;

$$ROAS = \frac{Receita}{CustoPublicitario}$$

(15) Taxa de Conversão *("CR")*: proporção de visitas (cliques) que resultaram em uma transação;

$$TaxadeConversao = \frac{Transaçoes}{NumeroCliques} \times 100$$

(16) Custo por Mil *("CPM")*: Representa o gasto gerado a cada mil impressões do anúncio);

$$CPM = \frac{CustoTotal \times 1000}{TotalImpressoe}$$

(17) AOV *("Valor Médio do Pedido")*: Valor médio dos itens comprados em uma visita);

$$AOV = \frac{Receita}{Transacoes}$$

1.3. Objetivos

Previsão, ou em inglês, *forecasting*, que se refere ao "ato ou efeito de prever, antevisão, presciência....", pode ser definida como uma sequência de passos que o tomador de decisões realiza, seja implícita ou explicitamente, para antever satisfatoriamente um valor futuro. As previsões desempenham um papel cada vez mais importante em uma empresa moderna, pois elas podem ser usadas para prever vendas, por exemplo. Assim, tanto o processo de previsão como o tipo de modelos a serem usados desempenham um papel cada vez mais importante na função do processo de previsão.

Um conjunto alargado de métodos e modelos permitem fazer muitas vezes, para um mesmo problema, abordagens completamente distintas.

Neste trabalho temos como objetivo principal identificar os KPIs de maior importância a nível da variável resposta, onde neste estudo será o custo e a receita da empresa.

Verificar empiricamente quais os melhores KPIs que nos ajudam a investigar quais os modelos que fornecem a melhor qualidade de previsão do custo e receita, como as suas limitações e se os mesmos têm aplicações práticas.

2. Breve descrição das Metodologias

A metodologia de pesquisa foi desenvolvida em 5 grandes etapas:

- Definição do Problema de Previsão;
- Definição dos Modelos de Previsão a serem desenvolvidos e hipóteses a serem testadas;
- Construção das bases de dados;
- Desenvolvimento dos modelos de previsão;
- Comparação dos modelos de previsão.

O objetivo geral desta dissertação é avaliar modelos quantitativos de previsão baseados em análise de séries temporais e em métodos causais. Para cumpri-lo foram desenvolvidas alternativas de modelos de previsão em cada método e aquela alternativa com a melhor acurácia foi eleita para ser testada e comparada com os demais métodos.

FIGURA 1. Modelos de previsão avaliados.

(A) **ARIMA ou Método de Box-Jenkings [6] [8]:**

O método Autoregressivos ou Box-Jenkins utiliza um ferramental matemático mais complexo. Basicamente tem-se a utilização de técnicas de autoregressão, ou seja, regressões com base no tempo, médias móveis com o objetivo de suavizar e identificar sazonalidades e diferenciação buscando a incorporação de processos não estacionários. O modelo ARIMA, portanto, é um caso geral dos modelos propostos por Box e Jenkins (1976), que é apropriado para descrever séries não estacionárias. Ou seja, séries em que a média não é constante no período de análise, nas quais os parâmetros quase sempre são pequenos, apresentando tendência e ou sazonalidade. A construção dos modelos Box-Jenkins é baseada num ciclo iterativo, no qual a escolha do modelo é feita com base nos próprios dados. São três as etapas para construção do modelo:

- *Identificação*: consiste em descobrir qual dentre as várias versões dos modelos de Box-Jenkins, sejam eles sazonais ou não, descreve o comportamento da série. A identificação do modelo a ser estimado ocorre pelo comportamento das funções de autocorrelações (ACF) e das funções de autocorrelações parciais (PACF).
- *Estimação*: consiste em estimar os parâmetros ϕ e Φ do componente autoregressivo, os parâmetros θ e Θ do componente de médias móveis e a variância de ϵt.
- *Verificação*: consiste em avaliar se o modelo estimado é adequado para descrever o comportamento dos dados.

Caso o modelo não seja adequado, o ciclo é repetido, voltando-se à fase de identificação. Um procedimento utilizado é identificar não só um único modelo, mas alguns modelos que serão, então, estimados e verificados. Quando se obtém um modelo satisfatório, passa-se para a última etapa da metodologia de Box-Jenkins, que constitui o objetivo principal da metodologia: realizar previsões.

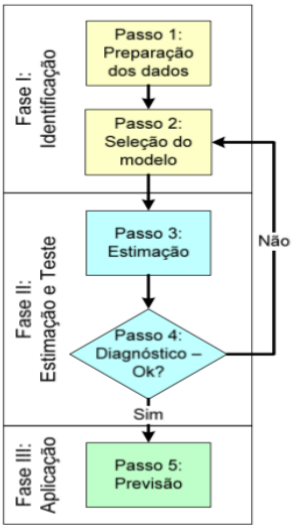


FIGURA 2. Esquema da metodologia: Box-Jenkings.

(B) **Regressão Linear Múltipla [7] [13]**:

Os dois tipos de regressão mais utilizados: as regressões lineares simples com apenas uma variável explicativa e as regressões lineares múltiplas, com mais de uma variável explicativa. A análise de regressão é uma técnica estatística que se ocupa do estudo da dependência de uma variável (dependente) em relação a uma ou mais variáveis (independentes ou explicativas). O objetivo principal deste modelo é estimar e/ou prever a média ou o valor médio da variável dependente em relação aos valores conhecidos (ou fixos) das variáveis independentes. A análise de regressão é um dos modelos mais usados, sobretudo para fazer previsões. Se num modelo de regressão linear simples (RLS) introduzirmos mais variáveis

explicativas passamos a ter um modelo de regressão linear múltipla (RLM):

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n$$

- Y_i a variável explicada ou dependente (aleatória), representa o que o modelo tentará prever;
- β_0, β_1, ... , β_p são designados por parâmetros ou coeficientes de regressão desconhecidos do modelo;
- x_1, x_2, ... , x_p são as p variáveis explicativas (independentes) medidas sem erro (não aleatória);
- e_i variável aleatória residual na qual se procura incluir todas as influências no comportamento da variável Y_i que não podem ser explicadas linearmente pelo comportamento da variável X_i, é a diferença entre a variável resposta observada Y_i e a variável resposta prevista $\hat{Y}_i$.

Interpretação dos coeficientes:

- β_0 - representa o valor esperado da variável dependente Y quando as variáveis explicativas são simultaneamente iguais a zero;
- β_j - representa a variação do valor esperado de Y por cada incremento unitário em x_j quando se mantêm constantes as restantes variáveis explicativas.

Um método de estimação dos coeficientes de regressão β é o método dos mínimos quadrados que consiste em minimizar a soma de quadrados dos erros aleatórios: $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})]^2$

Obtém-se o estimador dos mínimos quadrados:

$\hat{\beta} = (X^T X)^{-1} X^T Y$

Este método determina uma reta que minimiza as diferenças entre os valores estimados pela reta e os pontos da amostra. Assim o valor de β_0 é o ponto de partida sobre o qual os outros fatores têm influência. Já os demais (β_1, $\cdots$, β_n) são os coeficientes das variáveis independentes, ou, seja, exprimem o grau de influência que cada uma das variáveis explicativas exerce sobre o modelo.

A análise de regressão, tem três propósitos gerais:

- modelar a relação entre a variável dependentes (Y) e uma ou mais variáveis independentes;
- mensurar o erro ao usar a relação que prediz a variável dependente; e
- medir o grau de associação entre a variável dependente e as independentes.

Para que esses propósitos sejam alcançados, uma série de testes estatísticos em relação ao ajuste e significância deve ser analisada. Entre eles destacam-se:

- erro padrão de estimação (*standard error of estimate* ou SSE): mede a dispersão entre os valores originais em relação aos valores ajustados. O valor desta estatística deve ser pequeno, próxima a zero.

- coeficiente de determinação ($\mathbb{R}$) e o coeficiente de determinação ajustado ($\mathbb{R}^2$) : o primeiro mede a quantidade de variabilidade nos dados explicada ou considerada pelo modelo de regressão; enquanto o segundo mede a proporção de variação na variável dependente (Y), que é explicada pela relação com as variáveis independentes (X);
- teste de significância dos coeficientes de regressão (Teste t): testa a relação linear entre Y e os Xs, ou seja, verifica se as variáveis Xs explicam a variabilidade de Y.
- análise de variância (ANOVA): testa a significância geral da regressão, ou seja, confirma se há relação estatística significante entre a variável dependente e uma ou mais variáveis explicativas.

Os valores $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}$ designam-se por valores estimados ou valores preditos de y_i. As quantidades $e_i = y_i$ - $\hat{y}_i$ são designados por resíduos.

Assim, com base nos propósitos gerais da anállise de regressão, nos testes estatísticos necessários e nas hipóteses que devem ser verdadeiras, a Figura 3 apresenta uma representação esquemática do processo de moldagem da regressão múltipla.

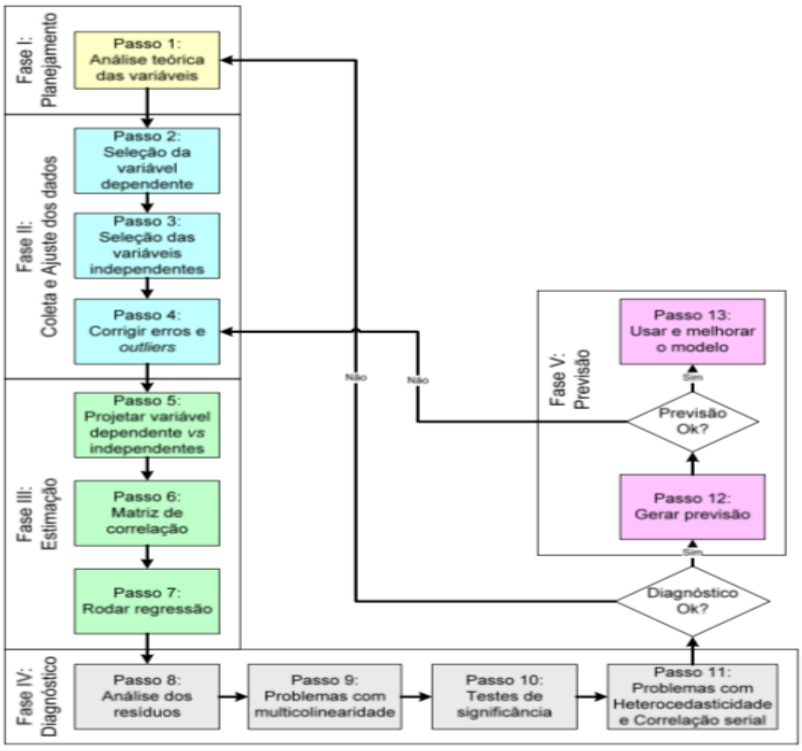


FIGURA 3. Esquema da metodologia: Regressão Múltipla.

(1) **ACP - Análise de Componentes Principais [10][12]**:

Seja X um vetor aleatório p-dimensional com vetor valor médio μ (também p dimensional) e matriz de covariância populacional Σ, quadrada, de ordem pxp, simétrica e definida positiva.

Com a análise em componentes principais pretende-se explicar a estrutura da variância ou covariância através de combinações lineares das variáveis aleatórias originais. Essas combinações lineares das p variáveis aleatórias originais são independentes entre si.

Portanto, a partir de p variáveis altamente correlacionadas, obtemos p variáveis independentes. Para explicar toda a variabilidade do sistema, são necessárias p componentes principais. No entanto, grande parte dessa variabilidade pode ser expressa através de um número k $<$ p componentes.

Esta técnica é mais utilizada como um meio do que como um fim. Como as p variáveis que obtemos são independentes entre si, podem ser utilizadas em, por exemplo, estudos de regressão múltipla.

Álgebricamente os componentes principais representam combinações lineares das p variáveis que constituem X..

Geometricamente, considerando X_1,X_2, ..., X_p um sistema de p eixos coordenados em que as observações estão descritas, os componentes principais representam rotações destes eixos, representando as direções de máxima variabilidade e formando um sistema ortogonal.

Os componentes principais depende apenas de X_1,X_2, ..., X_p e de σ.

Sabemos já que σ é simétrica e semi-definida positiva, pelo que os seus valores próprios são reais e não negativos, $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0$.

Sejam e_1,e_2,...,X_p representantes de norma 1 dos p vetores próprios associados a cada um dos valores próprios, respetivamente.

- Obtenção dos componentes principais a partir de Σ

  Os componentes principais. $Y_1, Y_2, \cdots, Y_p$ são tais que

  $$Y_i = e_i'X = \epsilon_{i1}X_1 + \epsilon_{i2}X_2 + \cdots + \epsilon_{ip}X_p, i = 1, \cdots, p$$

  ou seja,

  $$Y_1 = e_1'X = \epsilon_{11}X_1 + \epsilon_{21}X_2 + \cdots + \epsilon_{p1}X_p$$

  $$Y_2 = e_2'X = \epsilon_{21}X_1 + \epsilon_{22}X_2 + \cdots + \epsilon_{p2}X_p$$

  $$\cdots$$

  $$Y_k = e_p'X = \epsilon_{p1}X_1 + \epsilon_{p2}X_2 + \cdots + \epsilon_{pp}X_p$$

  Nestas condições,

  1 $\text{Var}(Y_i) = e_i'\Sigma e_i = e_i'\lambda_i e_i = \lambda_i e_i'e_i = \lambda_i$;
  2 $\text{Cov}(Y_i, Y_k) = e_i'\Sigma e_k = e_i'\lambda_i e_k = \lambda_i e_i'e_i = 0$
  3 Mais, $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = Var(X_1) + Var(X_2) + \cdots + Var(X_p) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_p)$

  De facto, da decomposição ortogonal de Σ, sabendo que P é ortogonal, isto é, $P' = P^{-1}, \Sigma = PDP'$, tem-se, $tr(\Sigma) = tr(PDP') = tr(D) = \sum_{i=1}^{p} \lambda_i$
  Assim,

  – a percentagem de variação total explicada pelo k componente principal é $\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$.

- Obtenção dos componentes principais a partir de ρ.
  Seja ρ a matriz de correlação de X. A matriz de correlação não é mais que a matriz de covariâncias de $Z = V^{\frac{-1}{2}}(X - \mu)$ onde $V^{\frac{-1}{2}}$ é a matriz diagonal com os elementos da diagonal iguais a $\frac{1}{\sqrt{\sigma_{ii}}}$. Assim, os componentes principais são dados pelos vetores próprios normalizados associados aos valores próprios de $\rho : Y_i = e_i'Z = e_i'V^{\frac{-1}{2}}(X - \mu), i = 1, \cdots, p$.
  Verifica-se ainda que:

  1 $\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$;
  2 $\sum_{i=1}^{p} \lambda_i = p$;
  3 $\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}$.

  A proporção da variância total explicada pela i-ésima componente principal é $\frac{\lambda_i}{p}$.

(2) **Teste da Razão de Verosimilhança [15]**:

O teste da razão de verossimilhança para a significância dos p coeficientes das variáveis independentes do modelo é realizado da mesma forma que no modelo de regressão logística simples. A estatística teste G é dada por :

$D = -2\ln[\frac{Verosimilhançadomodeloajustado}{Verosimilhançadomodelosaturado}]$ ou seja: $D = -2\ln(L_S) + 2\ln(L_c)$

em que, L_S é a verossimilhança do modelo sem a covariável e L_C é a verossimilhança do modelo com a covariável.

No caso da regressão múltipla, temos o interesse em saber se pelo menos uma variável é significativa para o modelo. Sob a hipótese nula, os p coeficientes são iguais a zero , assim, a estatística G tem distribuição Qui-Quadrado com p graus de liberdade. Nesse caso L_C é a verossimilhança do modelo com as p variáveis explicativas e L_S é a verossimilhança do modelo apenas com o intercepto.

(3) **Teste de Wald [15]**:

O teste de Wald tem como objetivo testar a significância de cada coeficiente dentro do modelo obtido, ou seja se o coeficiente é diferente de zero. Deste modo, o teste de Wald averigua se uma determinada variável independente apresenta uma relação estatísticamente significativa com a variável dependente. Assim, pretende-se testar:

$H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0, j = 0, \cdots, p.$

A estatística de teste é dada por

$W_j = \frac{\hat{\beta_j}}{\sigma(\hat{\beta})}$

De forma equivalente, o teste de Wald também pode ser obtido pela multiplicação dos seguintes vetores:

$W = \hat{\beta}^T(X_T V X)\hat{\beta}$

Com distribuição Qui-quadrado e p+1 g.l. sob a hipótese que cada um dos p+1 coeficientes é igual a zero.

(4) **Critério de Informação de Akaike [15]**:

AIC, ou critério de Akaike, é uma ferramenta para seleção de modelos, pois oferece uma medida relativa quanto à qualidade do ajuste de um modelo estatístico. Este não se apresenta na forma de um teste de um modelo no sentido usual de testar uma hipótese nula, ou seja, o AIC não pode indicar nada sobre o quão bem o modelo ajusta os dados num sentido absoluto.

Na sua forma geral, AIC é dado por:

$AIC = 2K - 2\ln(L)$

onde k é o número de parâmetros no modelo estatístico, e L é o valor maximizado da função de verossimilhança para o modelo estimado.

Dado um conjunto de modelos candidatos, o modelo preferível é aquele com o valor mínimo de AIC. Este indicador será desenvolvido na aplicação de R em modelos de regressão logística múltipla.

(5) **Teste F-parcial na ANOVA [15]**:
Em problemas de regressão linear múltipla, certos testes de hipóteses sobre os parâmetros do modelo são úteis para verificar a "adequabilidade"do modelo. O teste para significância da regressão é um teste para determinar se há uma relação linear entre a variável resposta e algumas das variáveis regressoras. Consideremos as hipóteses:
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs $H_1 : \beta_j \neq 0$ para qualquer $j = 1, \cdots, p$

Se rejeitamos H_0, temos que pelo menos uma variável explicativa contribui significativamente para o modelo. Prova-se que $\frac{SQR}{\sigma^2}$ e $\frac{SQE}{\sigma^2}$ são independentes e seguem distribuição qui-quadrado, nomeadamente
$\frac{SQR}{\sigma^2} \sim X^2_p$ e $\frac{SQE}{\sigma^2} \sim X^2_{n-p-1}$.
donde a estatística F_0 dada por
$F_0 = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} = \frac{QMR}{QME} \sim F_{p;n-p-1}$.

Portanto, rejeitamos H_0 se $F_0 > F_{(1-\alpha;p;n-p-1)}$ e $p - valor = P[F_{p;n-p-1} > F_0 | H_0 Verdadeira] < \alpha$ em que α é o nível de significância considerado. Geralmente adotamos α=5%. A tabela da ANOVA com a estatística F_0 é dada por:

Tabela 1. Tabela anova.

Fonte	Soma de Quadrados	GL	Quadrado Médio	F_0
Regressão	SQR	p	$QMR = \frac{SQR}{p}$	$F_0 = \frac{QMR}{QME}$
Erro (Resíduo)	SQE	n-p-1	$QME = \frac{SQE}{n-p-1}$	
Total	SQT	n-1		

(6) **No software R serão utilizados testes variados para validar pressupostos, nomeadamente:**:

- Teste da Razão de Verosimilhança [15];
- Teste F-parcial na ANOVA [15];
- Teste de Kolmogorov-Smirnov com correção de Lilliefors;
- Teste da tendência do Cox-Stuart;
- Teste t-student;
- Teste do Shapiro Wilk;
- Teste da sazonalidade de Kruskal-Wallis.

3. Bases de Dados

Os dados utilizados neste trabalho foram fornecidos pelo Facebook Ads e o Google Ads. Através da ferramenta *Python* extraiu-se as base de dados, onde se obteve os KPIs/variáveis que é necessário para conseguir prever um modelo para cada canal, que se ajuste melhor aos dados.

3.1. Facebook Ads

Este tipo de canal, Facebook Ads, é a alternativa mais usada no que se refere a *"paid social"*, ou seja, anúncios em redes sociais. Embora existam serviços de ads em várias outras redes, o Facebook ainda é considerado o canal de maior retorno e de maior alcance [2]. O Facebook Ads (FA) permite criar anúncios com texto e foto, que podem redirecionar para a *fanpage* da Overcube ou diretamente para o site da empresa. Com o FA conseguimos obter informação sobre o sexo do utilizador, localização (cidade, estado, país), idade, status de relacionamento, profissão e até mesmo interesses pessoais através de padrões usados. Para o canal Facebook Ads, na tabela seguinte são expressas, de um modo resumido, as diferentes tipos de variáveis, assim como breves descrições das mesmas.

Tabela 2. Descrição das variáveis do Facebook Ads.

	Variáveis	Descrição
	•Mês •Ano •Mercado	• Período e Mercado ativo
V.Resposta	•Custo (*Spend*) •Receita (*Conversion Value*)	•Custo total dos anúncios •Receita total proveniente das transações
KPIs	•Cliques(*Clicks*) •Impressões(*Impressions*) •Alcance(*reach*) •Conversões/Transações(*Conversions*) •Adições ao Carrinho(*Add to Cart*) •Resultados(*Results*) •Compromisso(*engagement*) •Custo por Resultado(*Cost per Result*) •Taxa de Cliques(*CTR*) •Custo por Clique(*CPC*) •Custo por Ação(*CPA*) •Retorno no gasto com publicidade(*ROAS*) •Taxa de Conversão(*CR*) •Custo por Mil(*CPM*)	•Número de cliques que são feitos em relação ao número de impressões •Número de vezes que um anúncio foi exibido •O número de pessoas que viram os anúncios pelo menos uma vez. •Número Total de aquisições/valor acumulado no site •O número de eventos de adição ao carrinho monitorizados pelo site e atribuídos aos anúncios. •O número de vezes que o anúncio alcançou um resultado. •Grau de compromisso dos usuários em relação à sua marca •Investimento médio por resultado do anúncio •Número de cliques que são feitos em relação ao número de impressões •Custo por clique, valor cobrado cada vez que o anúncio receber um clique •Custo por Aquisição, valor cobrado por cada conversão •Retorno no gasto da publicidade, valor em receita se recebe por cada euro gasto em publicidade •Proporção de visitas que resultaram em uma transação •Custo por Mil Impressões

Os dados em estudo dizem respeito ao estudo das campanhas do Facebook em função de algumas das variáveis presentes na base de dados. Interessa salientar que a base de dados original contém 26 variáveis e estuda 1237 campanhas num total, depois de agregar os 6 mercados (Portugal, Alemanha, Canada, Espanha, US, UK). As análises do presente projeto incidem sobre uma base de dados com 19 variáveis selecionadas que se passa a descrever.

- *Conversion Value* (Receita) é o valor total das conversões de compras no site;
- *Spend* (Gasto Total) é o custo de cada anúncio;
- *Clicks* (Cliques) é o número de cliques que o anúncio obteve;
- *Impressions* (Impressões) é o número de vezes que os anúncios foram apresentados;
- *Reach* (Alcance) é o número de pessoas que viram os anúncios pelo menos uma vez;
- *Conversions* (Conversões) é uma ação que uma pessoa exerce no site, como efetuar um pagamento, registar-se, adicionar um item ao carrinho de compras ou ver uma página específica;
- *Add to Cart* (Adições ao carrinho no site) é o número de eventos de adição ao carrinho monitorizados pelo pixel no site e atribuídos aos anúncios;
- *Results* (Resultados) é o número de vezes que o anúncio alcançou um resultado, com base no objetivo e nas definições selecionadas;
- *Engagement* (Interação com a publicação) é o número total de ações que as pessoas executam com base nos anúncios;
- *Cost per Result* (Custo por Resultado) é o custo médio por resultado dos anúncios;
- *CTR* (Taxa de cliques) é a percentagem de vezes que as pessoas viram o anúncio e clicaram na ligação;
- *CPC* (Custo por clique) é o custo médio por cada clique na ligação;
- *CPA* (Custo por Aquisição) permite pagar apenas pelas ações que as pessoas tomam devido ao anúncio.
- *ROAS* (Retorno dos Gastos do Anúncio) é o retorno total dos gastos do anúncio (ROAS) resultante de compras no site. Isto baseia-se no valor de todas as conversões registadas pelo pixel do Facebook no site e atribuídas aos anúncios;
- *CR* (Taxa de Conversão) é a proporção de visitas que resultaram em uma transação/conversão;
- *CPM* (Custo por Mil Impressões) é um indicador comum utilizado pelo sector da publicidade online para medir a rentabilidade de uma campanha de anúncios.
- *Mês* mês onde está inserida a campanha;
- *Ano* ano onde está inserida a campanha;
- *Mercado* mercado onde está inserida a campanha;

3.2. Google Ads

Outro tipo de canal para o E-commerce é o Google Ads, que se tornou sinónimo de *"paid search"*, ou *"busca paga"*. Existem outros mecanismos de busca, que também oferecem este tipo de serviços, mas a grande vantagem do Google Adwords (GA), comparado aos concorrentes, é o seu alcance. [5]. Basicamente qualquer usuário de internet no mundo utiliza o Google diariamente. O GA permite criar anúncios de texto redirecionando para o site da Overcube ou até para a página do Facebook da Overcube. Quando se cria um

anúncio, é escolhida uma palavra identificadora do anúncio que a nível das buscas vai permitir que este seja resultado de busca e assim exibido nas posições de topo. Para o Google Ads, na tabela seguinte, serão expressas, de um modo resumido, os diferentes tipos de variáveis, assim como uma breve descrição das mesmas.

Tabela 3. Descrição de variáveis do Google Ads.

	Variáveis	Descrição
	•Mês •Ano •Mercado	• Período ativo •Mercado (Portugal, Alemanha, Canada, Espanha, Estados Unidos, Irlanda, Reino Unido)
V. Resposta	•Custo(*Cost VF*) •Receita(*Total Conv. Value*)	•Custo total dos anúncios •Receita total proveniente das transações
KPIs	•Conversões(*Conversions*) •Cliques(*Clicks*) •Impressões(*Impressions*) •Taxa de Cliques(*CTR*) •Custo por Clique (*CPC*) •Custo por Aquisição (*CPA*) •Retorno Gasto com Publicidade (*ROAS*) •Taxa de Conversão (*CR*) •Valor Médio do Pedido (*AOV*)	•Número Total de aquisições/valor acumulado no site. •Numero Total de Cliques. •Número de vezes que um anúncio foi exibido. •Número de cliques que são feitos em relação ao número de impressões. •Valor cobrado, porque cada vez que o anúncio receber um clique. •Custo por Aquisição, valor cobrado por cada conversão. •Retorno no gasto da publicidade, valor em receita se recebe por cada euro gasto em publicidade •Proporção de visitas que resultaram em uma transação •Valor médio dos itens comprados em uma visita

Os dados em estudo dizem respeito ao estudo das campanhas do Google Ads em função de algumas das variáveis presentes na base de dados. Acrescenta-se que a base de dados original contém 16 variáveis e estuda 2276 campanhas. As conclusões retiradas do presente projeto incidem sobre uma base de dados com 11 variáveis selecionadas, que se passam a descrever:

- *Cost VF* (Custo Total) é o custo de cada anúncio;
- *Total Conv. Value* (Receita) é o valor total das conversões de compras no site;
- *Campaign ID* número de identificação da Campanha;
- *Ad type final*, tipo de anúncio final;
- *Conversions* (Conversões) é uma ação que uma pessoa exerce no site, como efetuar um pagamento, registar-se, adicionar um item ao carrinho de compras ou ver uma página específica;
- *Clicks* (Cliques) é o número de cliques que o anúncio obteve;
- *Impressions* (Impressões) é o número de vezes que os anúncios foram apresentados;
- *CTR* (Taxa de cliques) é a percentagem de vezes que as pessoas viram o anúncio e clicaram na ligação;
- *CPC* (Custo por clique) é o custo médio por cada clique na ligação;
- *CPA* (Custo por Aquisição), permite pagar apenas pelas ações que as pessoas tomam devido ao anúncio.
- *ROAS* (Retorno dos Gastos do Anúncio) é o retorno total dos gastos do anúncio (ROAS) resultante de compras no site. Isto baseia-se no valor de todas as conversões registadas pelo pixel do Facebook no teu site e atribuídas aos anúncios;
- *CR* (Taxa de Conversão) é a proporção de visitas que resultaram em uma transação/conversão

- *AOV* (Valor médio do Pedido) é utilizada para acompanhar o desempenho de promoções e outras mudanças no valor, como oferecer um produto ou serviço.
- *Mercado* Mercado/País em questão;
- *Month of Year* mês;
- *Year* ano;

3.3. Variáveis em Estudo

Designam-se por variáveis as qualidades, propriedades ou características de objectos, de pessoas ou de situações que são analisadas num estudo podendo assumir diferentes valores para exprimir graus, quantidades e diferenças. No decorrer do trabalho, as diferentes variáveis foram abordadas de formas diferenciadas, conforme os objetivos do estudo, tendo-se optado pela seguinte categorização:

3.3.1. Variáveis Independentes.

Uma variável independente ou explicativa é aquela a partir da qual se pretende medir os efeitos, sobre uma variável resposta designada como variável dependente. Para os dois canais, são apresentadas a seguir as variáveis consideradas:

- Facebook Ads:
  Cliques (*"Clicks"*); **Impressões** (*"Impressions"*); **Alcance** (*"Reach"*); **Adições ao Carrinho** (*"Add to Cart"*); **Conversões/Transações** (*"Conversions"*); **Compromisso** (*"Engagement"*); **Resultados** (*"Results"*); **Custo por Resultado** (*"Cost per Result"*); **Taxa de Cliques** (*"CTR"*); **Custo por Clique** (*"CPC"*); **Custo por Aquisição** (*"CPA"*); **Retorno Gasto com Publicidade** (*"ROAS"*);**Taxa de Conversão** (*"CR"*); **Custo por Mil** (*"CPM"*); **Mês** e **Ano**.

- Google Ads:
  Cliques (*"Clicks"*); **Impressões** (*"Impressions"*); **Transações** (*"Conversions"*); **Taxa de Cliques** (*"CTR"*); **Custo por Clique** (*"CPC"*); **Custo por Aquisição** (*"CPA"*); **Retorno Gasto com Publicidade** (*"ROAS"*);**Taxa de Conversão** (*"CR"*); **Valor Médio do Pedido** (*"AOV"*); **Ano** (*"Year"*) e **Month** (*"Mês"*)

3.3.2. Variáveis Dependentes.

As variáveis dependentes também designadas como variáveis explicadas são aquelas que sofrem o efeito esperado da variável independente e representam o comportamento, a resposta ou o resultado observado que é devido à presença da variável independente. Isto

é, a variável dependente é aquela que o pesquisador tem interesse em compreender, explicar ou prever. Desta forma, foram consideradas neste estudo como potenciais variáveis dependentes as seguintes:

- **Custo**: designada por *Spend* no canal Facebook Ads e designada por *Cost VF* no canal Google Ads.
- **Receita**: designada por *Conversion Value* no canal Facebook Ads e designada por *Total Conversion Value* no canal Google Ads.

4. Análise Exploratória

A partir dos dados fornecidos, será desenvolvida uma análise exploratória dos dados em estudo para uma melhor percepção do assunto em causa.

4.1. Análise - Facebook Ads

Da análise à base de dados do Facebook verificou-se a existência de 8 valores omissos em número não relevante para 5 das 16 variáveis. Foram obtidos valores outliers, mas que não se vai optar pela remoção destes, pois são valores significativos para a previsão do modelo. O cálculo dos quartis e amplitude interquartil das variáveis do Facebook Ads encontram-se registados na Tabela 4:

Tabela 4. Quartis - FA.

Variáveis	1º Quartil	Mediana	3º Quartil	AIQ
Spend	42.33	175.40	753.6	711.27
Clicks	186	1009	3912	3726
Impressions	9414	38796	154978	145564
Reach	4789	15776	55341	50552
Conversion Value	0	47.86	581.56	581.56
Conversions	0	1	7	7
Add to Cart	1	9	85	84
Results	3	59	758	755
Engagement	262	1405	5621	5359
Cost per Result	0.036	0.1711	37.18	37.144
CTR	0.017	0.028	0.041	0.024
CPC	0.0928	0.1703	0.3735	0.2807
CPA	0	14.44	82.19	82.19
ROAS	0	0.15	1.13	1.13
CR	0	0	0.0031	0.0031
CPM	2.45	4.25	8.26	5.81

Para a variável *Spend*, o primeiro quartil é de 42.33€, ou seja, significa que 25% dos dados são menores que esse valor, assim como outros 25% são superiores a 711.27€(terceiro quartil). Da *Conversion Value*, pode ver-se que os valores já são diferentes, 25% das campanhas têm uma receita superior a 581.56€bastante menos significativos como o custo.

Em seguida, numa perspectiva geral, fez-se uma pequena pesquisa dos dados, para perceber melhor o funcionamento da empresa. O maior número de campanhas é na época de verão (Agosto, Junho), embora o mês de Novembro de 2018 esteja equiparado ao mesmo número de campanhas do mês de Junho de 2018. É nos meses de Abril e Maio que existe um menor número de campanhas.

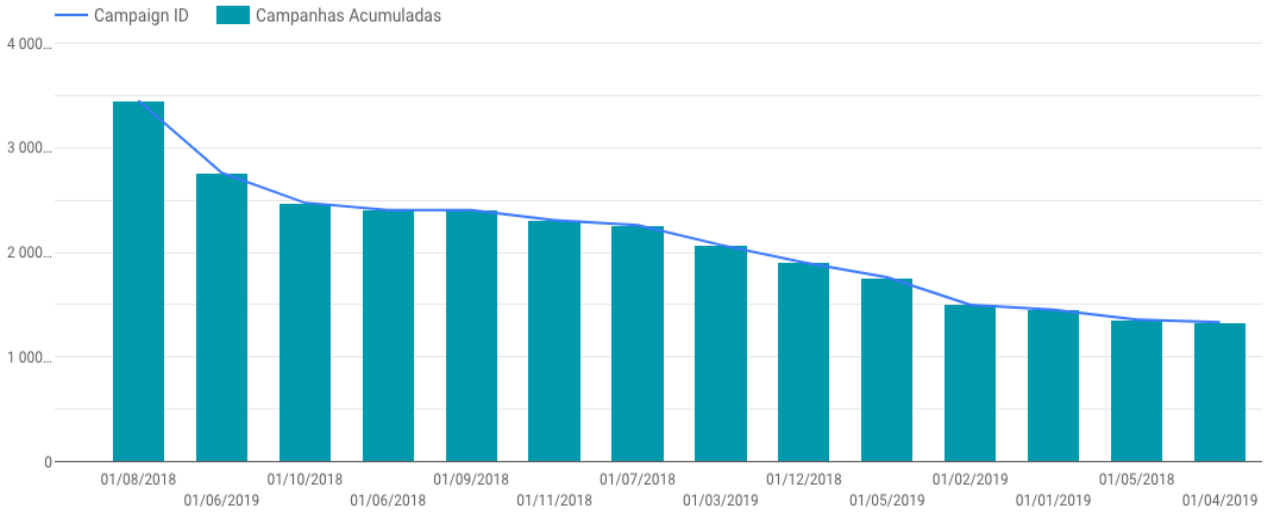


FIGURA 4. Gráfico de barras por campanhas por patamar.

No ano de 2018, nos meses de Maio, Junho, Julho, Agosto e Setembro, o custo de campanhas foi ligeiramente maior que a receita. Nos restantes meses, tanto do ano 2018 como o ano de 2019, a receita foi sempre maior. A receita atingiu o seu valor máximo no mês de Novembro de 2018 e o seu valor mínimo no mês de Maio de 2018, e o custo atingiu o seu valor máximo no mês de Dezembro de 2018 e atingiu o seu valor mínimo no mês de Setembro do mesmo ano.

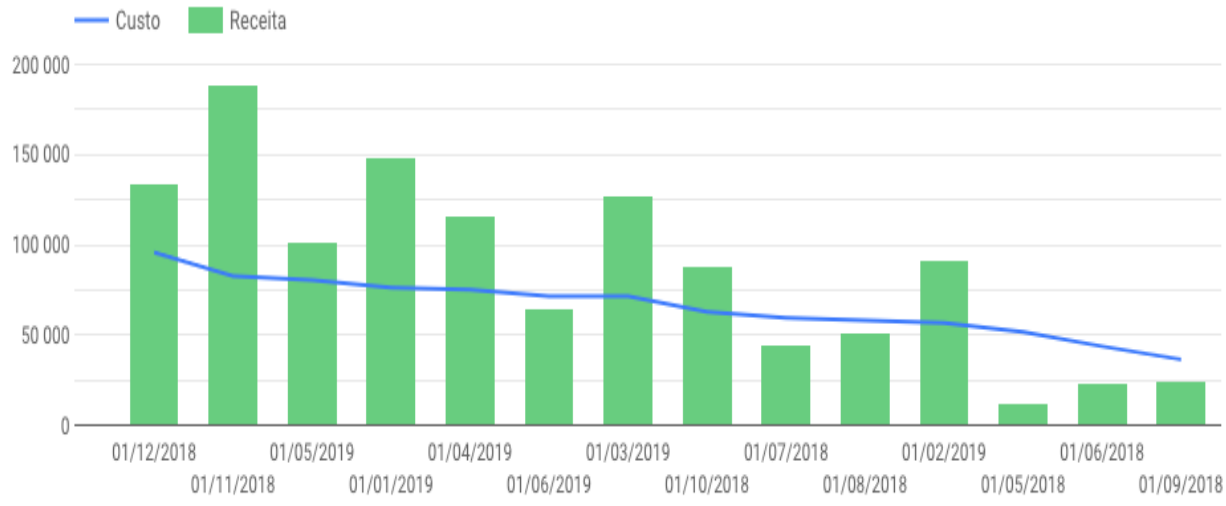


FIGURA 5. Gráfico de barras para as variáveis custo e a receita.

A receita é maior quando o objetivo final da campanha é *"Conversões"*, *"Remarketing"*, e em seguida *"Tráfego"*, mas o custo é aproximadamente igual quando o objetivo final é *"Tráfego"* e *"Conversões"* apesar da diferença entre receitas destes dois ser significativa, mais de 20000€.

O objetivo Tráfego foi concebido para direcionar as pessoas para o site ou app.

Com o objetivo Tráfego, podemos criar anúncios que:

- enviam pessoas para um destino dentro ou fora do Facebook (Cliques para o Site)
- aumentam o número de pessoas que acedem à app móvel ou para computador (Interação com a App)

Contudo, é de reparar que quando o objetivo é *"Remarketing"* os valores da receita vão até acima dos 700000€, o que é bastante positivo, quando apenas se gasta 400000€.

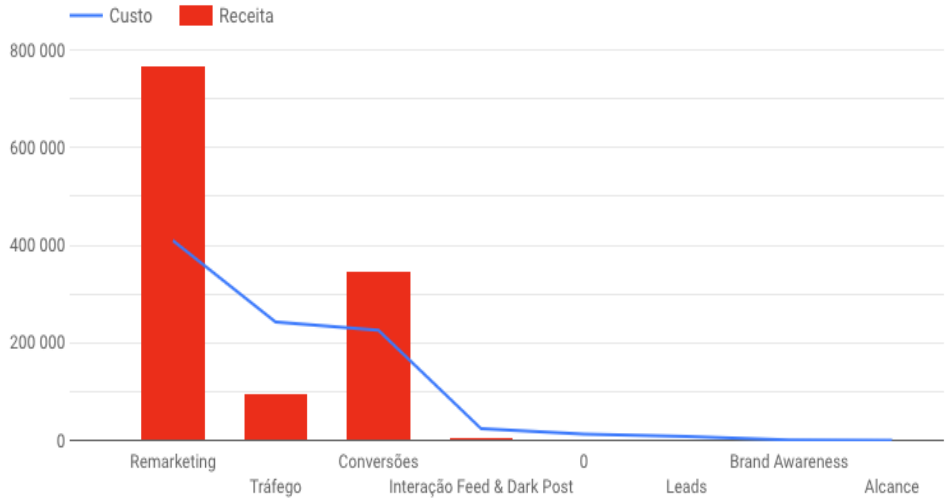


FIGURA 6. Gráfico do objetivo final de cada campanha para as variáveis custo e a receita.

Existe mais campanhas em que o objetivo final é o "Tráfego"e o "Interação Feed & Dark Post".
Quando se tem *Objetivo final=0* significa que existiram 33 campanhas lançadas em que não tinham objetivo.

Tabela 5. Dimensão do objetivo final.

Objetivo final	Total
0	33
Alcance	7
Brand Awareness	6
Conversões	261
Interação Feed & Dark Post	268
Leads	16
Remarketing	340
Tráfego	306
Total	1237

O mês onde existe mais campanhas é em Agosto, onde se pode ver na Tabela 6, comprova que provavelmente mais objetivos terá de cumprir, e então o custo será maior.

Tabela 6. Dimensão do tipo de campanhas por mês no ano 2018.

Mês	Junho18	Julho18	Agosto18	Setembro18	Outubro18	Novembro18	Dezembro18
Nº de campanhas	101	95	145	101	104	97	80

Tabela 7. Dimensão do tipo de campanhas por mês no ano 2019.

Mês	Janeiro19	Fevereiro19	Março19	Abril19	Maio19	Junho19
Nº de campanhas	61	63	87	56	74	116

Depois de analisar os dados, aos olhos do mercado e da empresa, concluiu-se diferentes comportamentos do custo e da receita ao longo do tempo.

Tabela 8. Resumo das medidas estatísticas das variáveis do canal Facebook Ads.

Variáveis	n	média	desvio padrão	min	25%	50%	75%	max
spend	1237	745.6	1381.4	0	42.33	175.4	753.6	10806
Conversion Value	1237	985.3	2850.8	0	0	47.8	581.5	36783.1
clicks	1237	5133.2	13284.22	0	186	1009	3912	135779
impressions	1237	184908	455051	6	9414	38796	154978	5.697628e+06
reach	1237	60826.1	124266	5	4789	15776	55341	923137
Conversions	1237	13.2	39.7	0	0	1	7	486
Add to Cart	1237	116.0	298.6	0	1	9	85	3994
Results	1237	2280.6	8420.7	0	3	59	758	121090
Engagement	1237	6235.6	14204.3	0	262	1405	5621	151530
Cost per Result	1236	39.9	105.6	0	0.033	0.171	37.18	1396.5
CTR	1237	0.031	0.019	0	0.01	0.028	0.041	0.166
CPC	1236	0.30	0.40	0	0.09	0.17	0.3735	4.54
CPA	1236	74.14	148.28	0	0	14.44	82.19	1396.5
ROAS	1236	1.24	6.7	0	0	0.15	1.13	162.19
CR	1233	0.002	0.006	0	0	0	0.003	0.071
CPM	1237	6.81	7.54	0	2.45	4.25	8.26	79.92

A mediana e a média medem a tendência central. Mas os valores atípicos, chamados de outliers, podem afetar a mediana menos do que afetam a média. Se os dados forem simétricos, a média e a mediana são semelhantes.

Os valores da mediana são bastante menores que os valores da média, daí, mais uma vez, mostram que os dados não são simétricos, e que têm sim assimetria á direita. A média das variáveis respostas (*Spend* e *Conversion Value*) é de 745.6 euros e de 985.3 euros por resultado.

O desvio padrão é uma medida que indica a dispersão dos dados dentro de uma amostra com relação à média. Um desvio padrão grande significa que os valores amostrais estão bem distribuídos em torno da média, enquanto que um desvio padrão pequeno indica que eles estão condensados próximos da média, ou seja, quanto menor o desvio padrão, mais homogénea é a amostra. Facilmente, como a Tabela 8, o desvio padrão da receita varia mais que o custo.

A análise de correlação tem por objetivo apenas medir o grau de relacionamento das variáveis. Sendo as variáveis quantitativas, utilizou-se o método do coeficiente de correlação linear de Pearson para quantificar o grau de associação entre pares de variáveis.

Aqui usou-se a correlação com o método de "Pearson", porque mede o grau da correlação linear entre duas variáveis quantitativas. Por isso, é possível que exista uma relação significativa mesmo que os coeficientes de correlação sejam 0.

A matriz diz que a variável *Spend* está muito relacionada positivamente (acima dos 0.80) com a variável *"Add to Cart"*, e com as variáveis *clicks*, *impressions* e *engagement* (acima dos 0.70) , e de certo modo, quando uma aumenta, o *Spend* aumenta moderadamente.

Com a variável *"Conversion Value"*, a situação é bastante diferente, está significativamente correlacionada com a variável *"Conversions"* e *"Add to Cart"*, ou seja, quando *"Add to Cart"* aumenta, a variável resposta também aumenta. De resto, como se pode concluir, as correlações não são muito fortes, Figura 7.

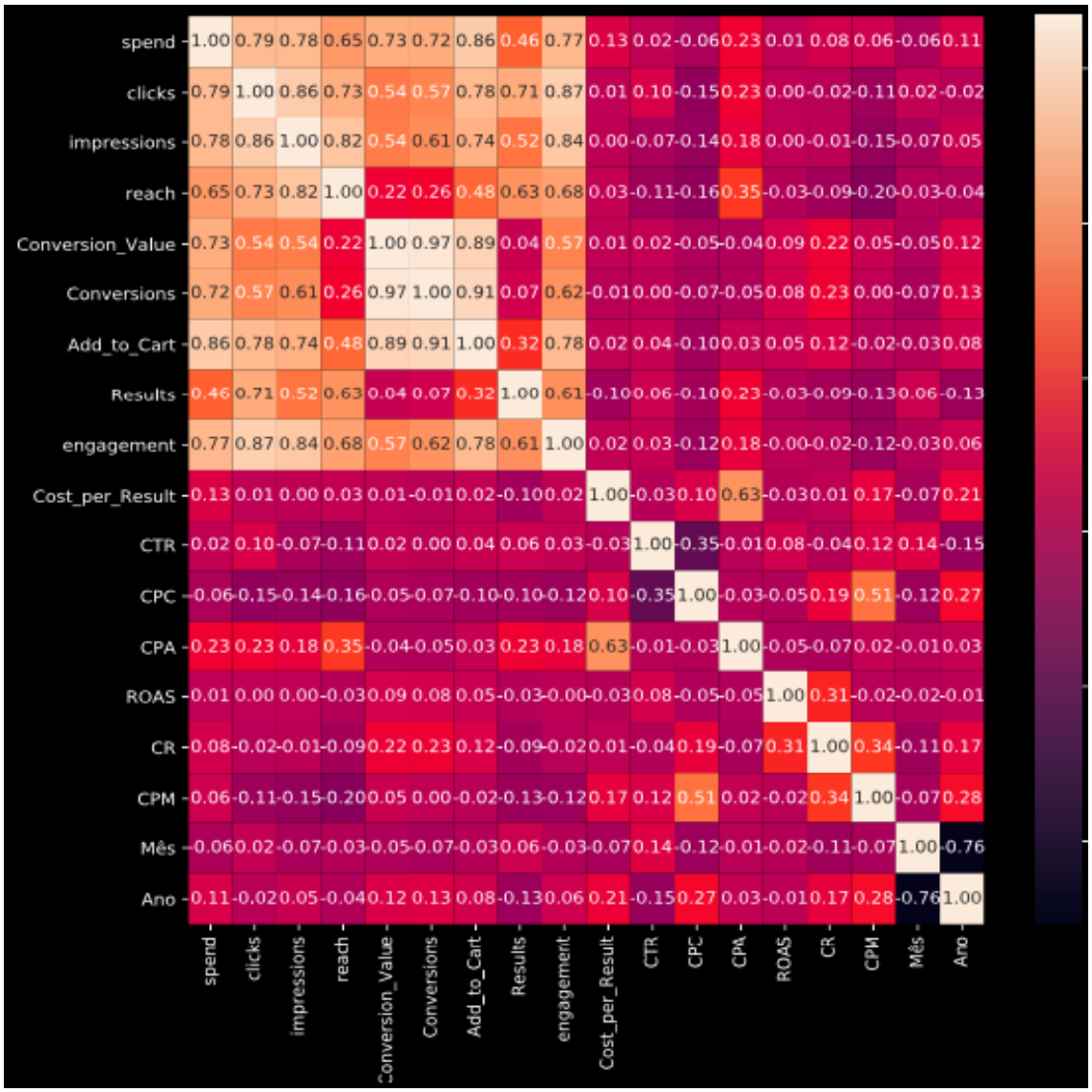


FIGURA 7. Cálculo do coeficiente de correlação linear de Pearson para as variáveis do canal Facebook Ads.

Através dos gráficos de dispersão, apresentados a seguir, consegue-se determinar o tipo de relação e a força da relação dos dados, é um complemento à análise da matriz de correlação. Inicialmente, analisa-se o comportamento de todas as variáveis com a variável *Spend*.

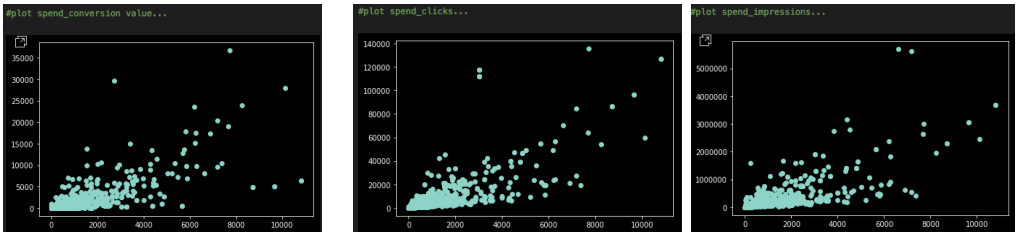


FIGURA 8. Spend vs C.Value (esq.); Spend vs Clicks (no meio); Spend vs Impressions (dir.).

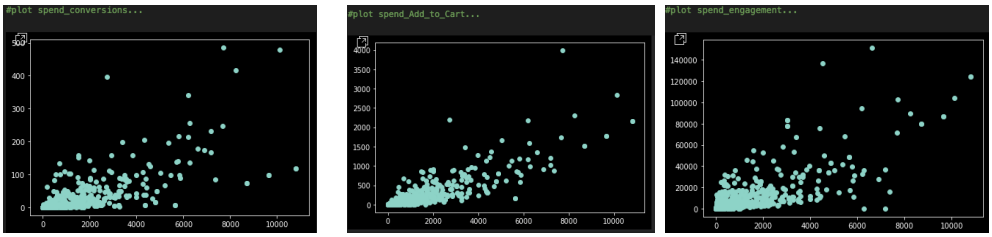


FIGURA 9. Spend vs Conversions (esq.); Spend vs Add to Cart (no meio); Spend vs Engagement (dir.).

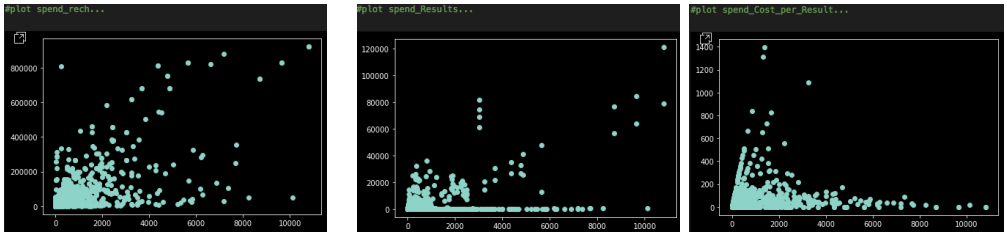


FIGURA 10. Spend vs Reach (esq.);Spend vs Results (no meio); Spend vs Cost per Result (dir.).

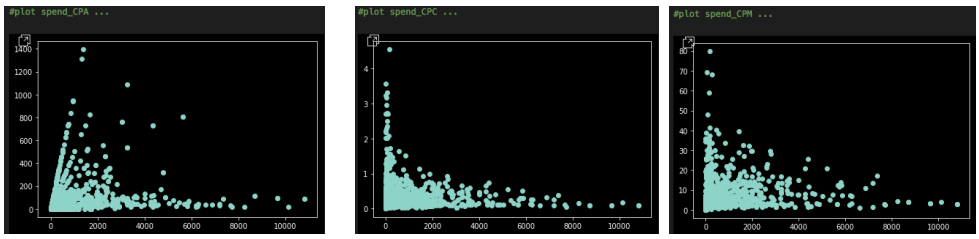


FIGURA 11. Spend vs CPA (esq.); Spend vs CPC (no meio); CPM (dir.).

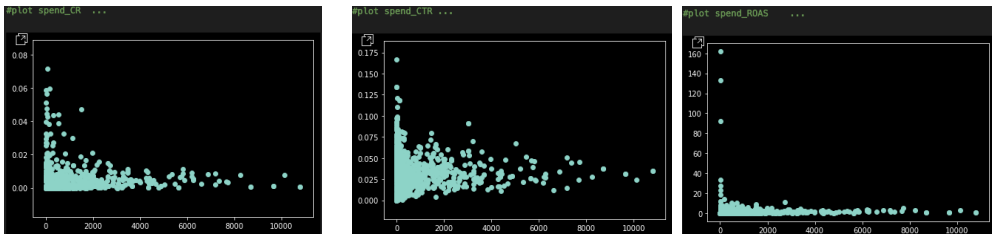


FIGURA 12. Spend vs CR (esq.); Spend vs CTR (no meio); Spend vs ROAS (dir.).

A variável *Spend* tem uma relação linear com as variáveis *Clicks*, *Impressions* e *Add to Cart*, apesar de existirem alguns pontos outliers, como se pode ver nos gráficos.

Em seguida procedeu-se a uma análise idêntica para se observar o comportamento de todas as variáveis com a outra variável resposta, (*Conversion Value*).

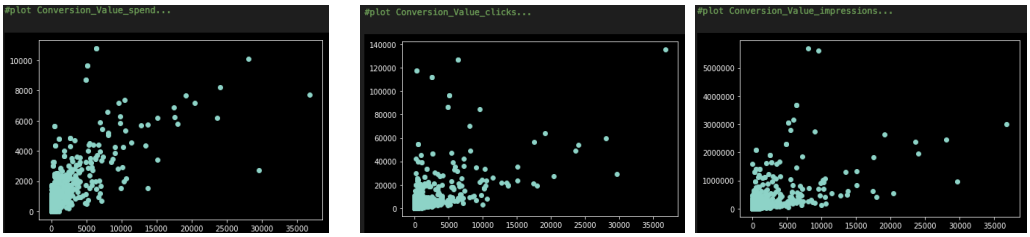


FIGURA 13. Conversion Value vs Spend (esq.); Conversion Value vs Clicks (no meio); Conversion Value vs Impressions (dir.).

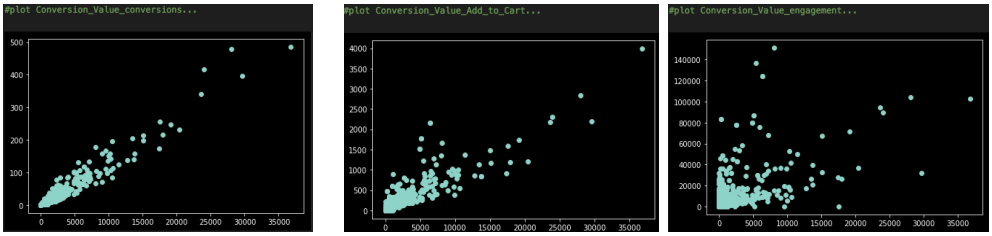


FIGURA 14. Conversion Value vs Conversions (esq.); Conversion Value vs Add to Cart (no meio);Conversion Value vs Engagement (dir.).

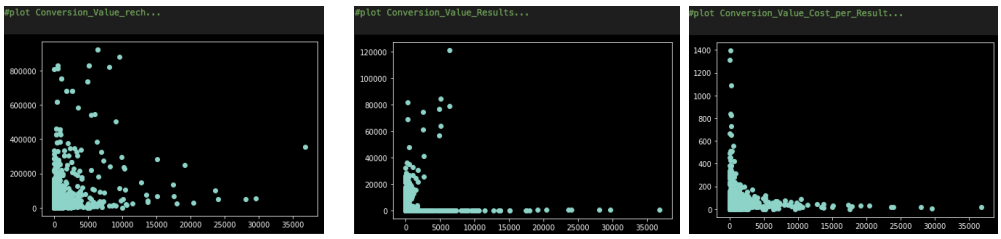


FIGURA 15. Conversion Value vs Reach (esq.); Conversion Value vs Results (no meio);Conversion Value vs Cost per Result (dir.).

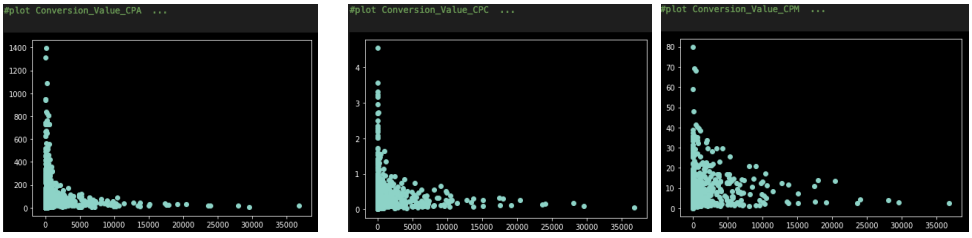


FIGURA 16. Conversion Value vs CPA (esq.);Conversion Value vs CPC (no meio);Conversion Value vs CPM (dir.).

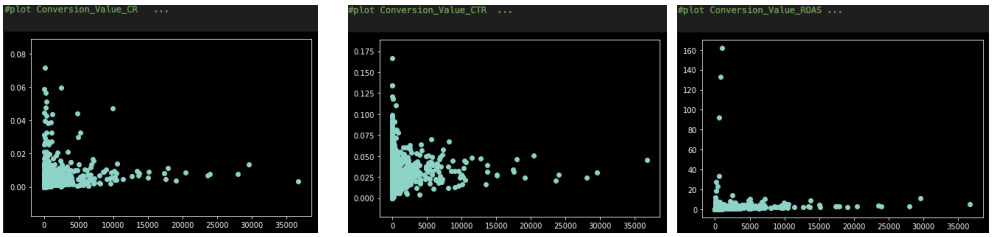


FIGURA 17. Conversion Value vs CR (esq.);Conversion Value vs CTR (no meio);Conversion Value vs ROAS (dir.).

A variável *Conversion Value* tem uma relação linear apenas com as variáveis *Conversions* e *Add to Cart* segundo os gráficos da Figura 14. A variável *Conversions* é linear á variável resposta porque esta é constituída pela mesma, ou seja, é o total das conversões, logo iria ser correlacionada.

Em seguida, analisamos os gráficos da função de densidade de probabilidade.

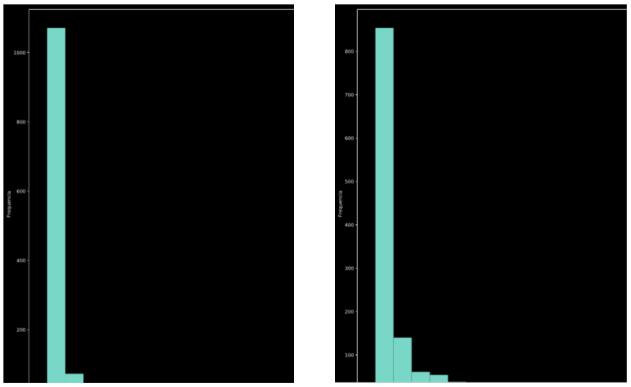


FIGURA 18. Histogramas das variáveis respostas (esq.: Conv.Value; dir.: Spend.)

Histograma é uma representação gráfica da distribuição de frequências de um conjunto de dados quantitativos.
Praticamente quase todos os histogramas são assimétricos e com apenas um pico. A frequência decresce bruscamente em um dos lados de forma gradual no outro, produzindo uma "cauda"mais longa em um dos lados. A média localiza-se fora do meio da faixa de variação. Quando a assimetria é à esquerda a mediana é superior à média, o que não é o caso, pois em todas as variáveis temos assimetria à esquerda, os valores rondam mais perto de zero.

O diagrama de extremos e quartis ou caixa de bigodes é uma ferramenta gráfica que permite visualizar a distribuição e valores discrepantes (outliers) dos dados, fornecendo assim um meio complementar para desenvolver uma perspectiva sobre o carácter dos dados. Aqui pode-se ver que existem vários *outliers*, mas esses outliers são os maiores valores de custo, isso significa que existem muitos valores perto do valor zero, onde o custo é mínimo, e estes valores *outliers* são os de maior valor relativamente á frequência em que se obteve os tais custos mínimos, como mostra o histograma da variável *Spend*.

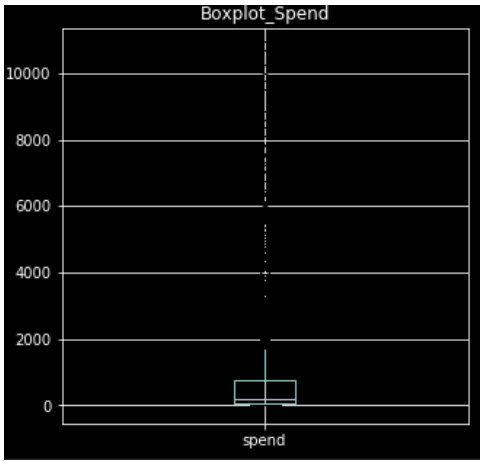


FIGURA 19. Diagrama de extremos e quartis e histograma da variável *Spend*.

Para a variável *Conversion Value* os valores são maiores que os valores da variável *Spend*, como se pode ver pela escala do histograma, apesar de ele se concentrar mais perto do zero. Existem alguns *outliers*, Figura 20, a frequência de valores já não é tão discrepante, daí existirem menos *outliers* nesta variável resposta que na outra variável resposta e também neste caso, são os valores considerados grandes.

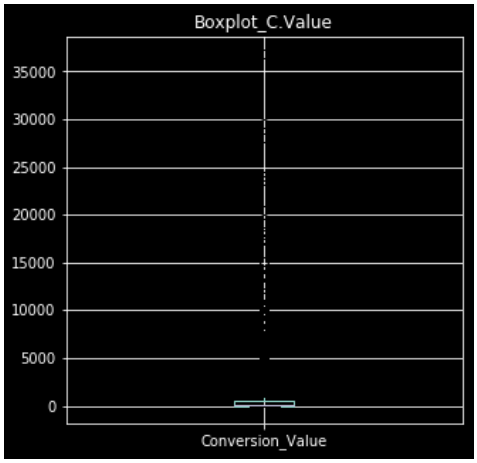


FIGURA 20. Diagrama de extremos e quartis e histograma da variável *Conversion Value*.

No primeiro gráfico da Figura 21 temos uma análise por mês relativamente à variável *Spend*, facilmente se pode ver que os dados são assimétricos positivos, porque a linha da mediana está próxima do primeiro quartil.

Cerca de 75% dos dados rondam os valores mínimos (perto de zero). Todos os diagramas de caixa são consideravelmente "achatados", o que indica uma baixa variabilidade e desvio-padrão.

Nos meses de Dezembro de 2018, Abril de 2019 e Maio de 2019 revelam ter uma variabilidade da variável *Spend* mais alta, tendo o mês de Dezembro o valor mediano alto.

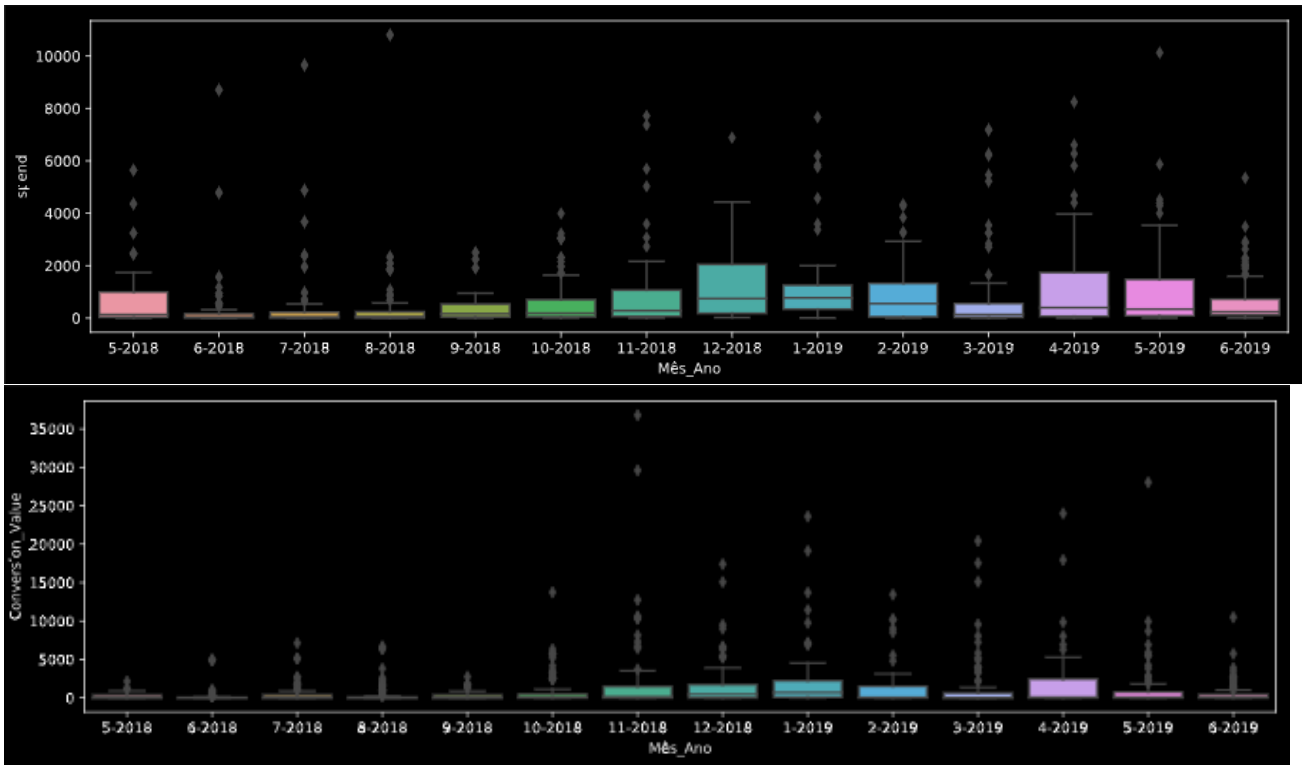


FIGURA 21. Diagrama de extremos e quartis das variáveis resposta em relação ao mês.

A variável *Conversion Value*, como está descrito no segundo gráfico da Figura 21, mostra que é no mês de Abril de 2019 e de Janeiro de 2019, que os valores parecem ter uma maior variabilidade, apesar de não ser igual à variabilidade da variável *Spend*. Os valores são maiores, mais diversificados e aleatórios que o custo.

4.2. Análise - Google Ads

Em segundo lugar, prossegue-se a uma análise exploratória dos dados do Google Ads. Novamente, verificou-se se existem valores omissos e concluiu-se a sua não existência nas 11 variáveis selecionadas, realça-se contudo a existência de muitos valores nulos, nomeadamentes nas medidas estatísticas do primeiro quartil e mediana. Obteve-se, na Tabela 9, uma análise dos quartis.

Tabela 9. Quartis - GA.

Variáveis	1º Quartil	Mediana	3º Quartil	AIQ
Cost VF	41.36	10.53	74.23	72.87
Conversions	0	0	1	1
Total Conv Value	0	0	95	95
Clicks	9	54	317	308
Impressions	139.75	1019.5	6469.75	6330
CTR	0.012	0.084	0.194	0.181
CPC	0.126	0.210	0.341	0.215
CPA	0	0	4.71	4.71
ROAS	0	0	1.17	1.17
CR	0	0	0.005	0.005
AOV	0	0	59.5	59.5

A variável *Cost VF*, tem como primeiro quartil 41.36€, ou seja, significa que 25% dos dados são menores que esse valor, assim como outros 25% são superiores a 74.23€(terceiro quartil). Da variável *Total Conv. Value*, pode-se ver que os valores já são diferentes, 25% das campanhas têm uma receita superior a 0€. A mediana é 0 e o valor do terceiro quartil é de 95€, ou seja, apenas cerca de 25% da totalidade dos dados são superiores ao valor do terceiro quartil.

Mais uma vez, não se procede à remoção dos valores que são considerados outliers, pois esses são os valores iniciais que é consequência do arranque da empresa.

Quando se trata da plataforma Google Ads, o maior número de campanhas é nos meses de Abril e Março do ano atual (2019). O mês de Dezembro e de Novembro de 2018 estão equiparados, existe um elevado número de campanhas, e nos restantes meses a decadência destes é mais significativa.

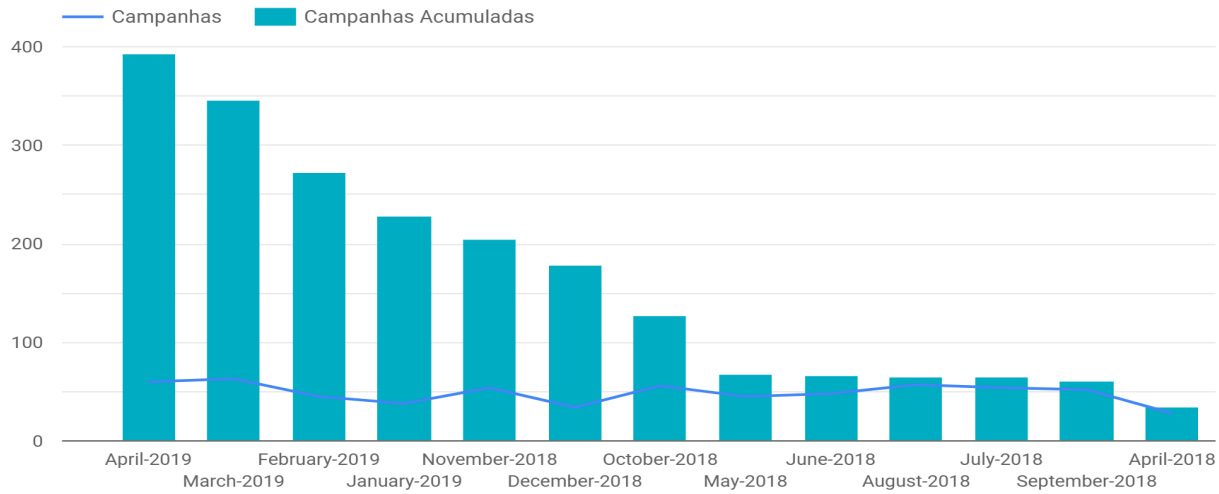


FIGURA 22. Gráfico de barras por campanhas por patamar.

No ano de 2018, os meses de Novembro e de Outubro, foram os que tiveram maior receita, e nos meses de Maio e de Abril a receita é menor mas é devido ao arranque da empresa, assim como aconteceu com o FA. Mas apesar do arranque da empresa, os meses com maior valor de receita são de 2018. No ano de 2019, os meses de Janeiro e de Março foram os de maior receita, e no mês de Fevereiro o custo é de 16357 euros. O pico de maior custo é do mês de Outubro de 2018, cerca de 34991 euros.

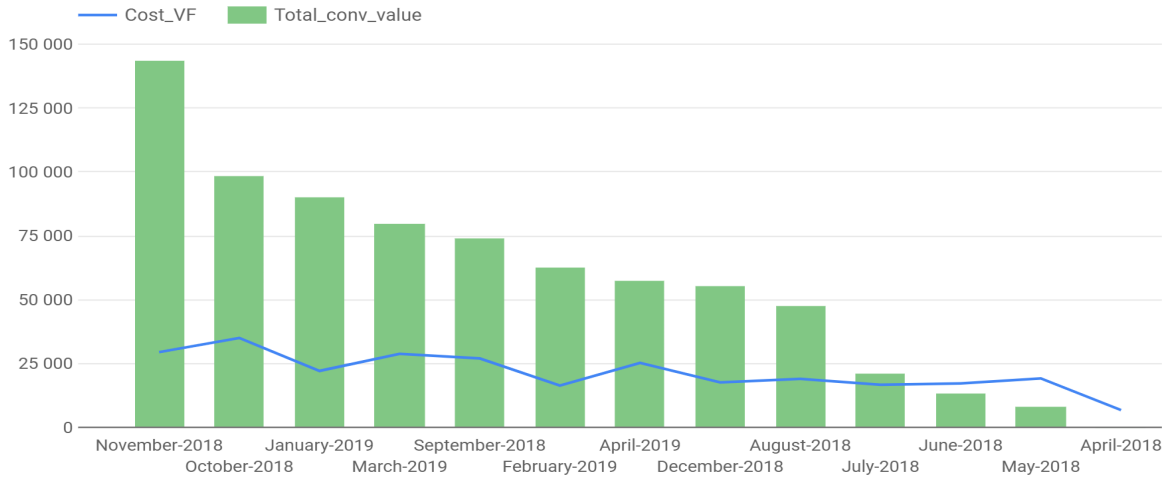


FIGURA 23. Gráfico de barras para as variáveis custo e a receita.

No Google Ads existem vários tipos de campanhas, nomeadamente do tipo Search, Shopping, DSA e Remarketing. Nos últimos anos, o GA tornou-se uma ferramenta bastante procurada por empresas de e-commerce e para as empresas que buscam mais conhecimento de marca ou aumento de vendas por meio de anúncios na internet. Contudo, empresas com diversos produtos ou com muitas campanhas diferentes no ar acabam gastando muito tempo desenvolvendo diariamente novas segmentações, criando e detalhando cada anúncio. Este probelma tem uma solução, criando vários canais diferentes:

- **DSA** - sigla de *Dynamic Search Ads*, pode ser traduzido para o português como "Anúncios Dinâmicos". Essa funcionalidade do *Google Ads* funciona da seguinte maneira: o anunciante seleciona se deseja utilizar todo seu website ou apenas algumas páginas específicas para determinado público da campanha.
- **Shopping** - permite a promoção do inventário por parte de retailers, bem como o aumento do número de visitas aos respetivos websites e lojas físicas. Os anúncios *Google Shopping* consistem numa foto do produto em questão com o respetivo título, preço e nome da loja (entre outros). Tendo em conta que o utilizador será impactado com esta informação antes de clicar no anúncio, será possível obter visitas mais qualificadas. Por exemplo, se o preço for demasiado elevado face à expectativa do utilizador, este à partida não clicará no anúncio.
- **Search** - A Rede de Pesquisa da Google é um grupo de Websites e aplicações relacionados com a pesquisa onde os seus anúncios podem ser apresentados. Quando anuncia na Rede de Pesquisa da Google, o seu anúncio pode aparecer junto a resultados da pesquisa quando alguém realiza pesquisas com termos relacionados com uma das suas palavras-chave.
- **Remarketing** - As campanhas de remarketing são usadas para exibir anúncios para pessoas que visitaram seu site ou usaram a app. Essas campanhas fornecem configurações e relatórios extras especificamente para alcançar visitantes e usuários anteriores.

A receita é maior quando o tipo de campanha é *Search* e o custo mais baixo é quando o tipo de campanha é *DSA*.

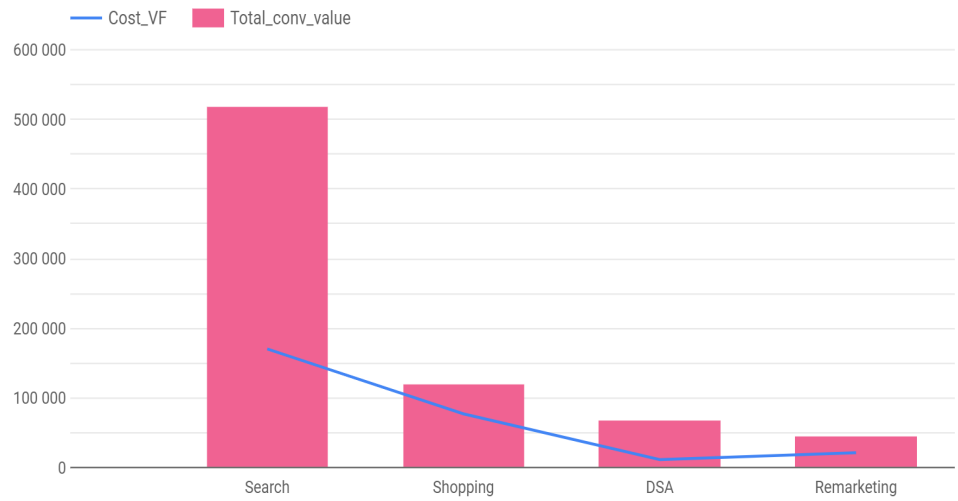


FIGURA 24. Gráfico de barras por tipo de campanha para as variáveis custo e receita.

Existem mais campanhas do tipo *Search*, e existem apenas 44 campanhas de *Remarketing*.

TABELA 10. Dimensão do tipo de campanha.

Tipo de Campanha	Total
Search	1545
Shopping	631
DSA	56
Remarketing	44
Total	2276

Para o ano de 2018, o mês onde existiu mais campanhas foi o mês de Novembro e foi no mês de Abril que em 2019 se destacou com maior número, como se pode ver nas seguintes tabelas. No ínicio do ano de 2018, com o arranque da imprensa, a quantidade de campanhas partilhadas é menor, como é compreensível.

TABELA 11. Dimensão do tipo de campanha por mês do ano 2018.

Mês	Abril18	Maio18	Junho18	Julho18	Agosto18	Setembro18	Outubro18	Novembro18	Dezembro18
Número de Campanhas	35	68	66	65	69	69	132	211	183

TABELA 12. Dimensão do tipo de campanha por mês do ano 2019.

Mês	Janeiro19	Fevereiro19	Março19	Abril19
Número de Campanhas	263	306	388	421

Em estatística, a análise exploratória de dados é uma abordagem à análise de conjuntos de dados de modo a resumir as suas características principais, frequentemente com métodos visuais. Na Tabela 13 será apresentada uma breve descrição das variáveis.

TABELA 13. Resumo das medidas estatísticas das variáveis do canal Google Ads.

Variáveis	count	mean	std	min	25%	50%	75%	max
Cost VF	2276	123.12	307.18	0.01	1.38	10.52	74.23	4849.98
Total Conv. Value	2276	330.85	1327.69	0	0	0	95	22987
clicks	2276	435.69	1108.97	1	9	54	317	10224
Conversions	2276	4.02	15.42	0	0	0	1	303
impressions	2276	19667	87275	1	139.75	1019.5	6469.75	2.152053e+06
CTR	2276	0.128	0.154	0.0007	0.012	0.084	0.194	2
CPC	2276	0.2621	0.2380	0	0.126	0.210	0.341	3.375
CPA	2276	21.67	75.40	0	0	0	4.71	1022.36
ROAS	2276	15.71	247.77	0	0	0	1.179	11300
CR	2276	0.009	0.068	0	0	0	0.005	2
AOV	2276	28.55	46.14	0	0	0	59.5	388

Mais uma vez, estes dados não são considerados simétricos, pois os valores da mediana e da média não são semelhantes.

Os valores da mediana são bastante menores que os valores da média, o que apresentam ter uma assimetria á direita. A média das variáveis respostas (*Cost VF* e *Total Conv.Value*) é de 123.12 euros e de 330.85 euros.

Como já se sabe, o desvio-padrão é uma medida que indica a dispersão dos dados, os valores do desvio-padrão são maiores, relativamente ao valor da média. A base de dados é bastante heterogénea e o valor do desvio padrão da receita é maior que a do custo.

A matriz relata que a variável *Cost VF* está relacionada muito positivamente com a variável *Clicks*, *Impressions* e a variável *Conversions*, quando uma aumenta, o *Cost VF* aumenta também.

Com a variável *Total Conv.Value*, o mesmo acontece, neste caso, as variáveis mais correlacionadas são igualmente as variáveis *Clicks* e *Conversions*, com valores de 0.71 e 0.97.

Num todo, as correlações não são muito fortes, as variáveis como o *CTR*, *CPC*, *CPA*, *ROAS*, *CR* e *AOV*, não têm um valor de correlação significativo com nenhuma variável, apesar que as variáveis *CR* e *ROAS* estão correlacionadas significativamente com um valor de 0.72.

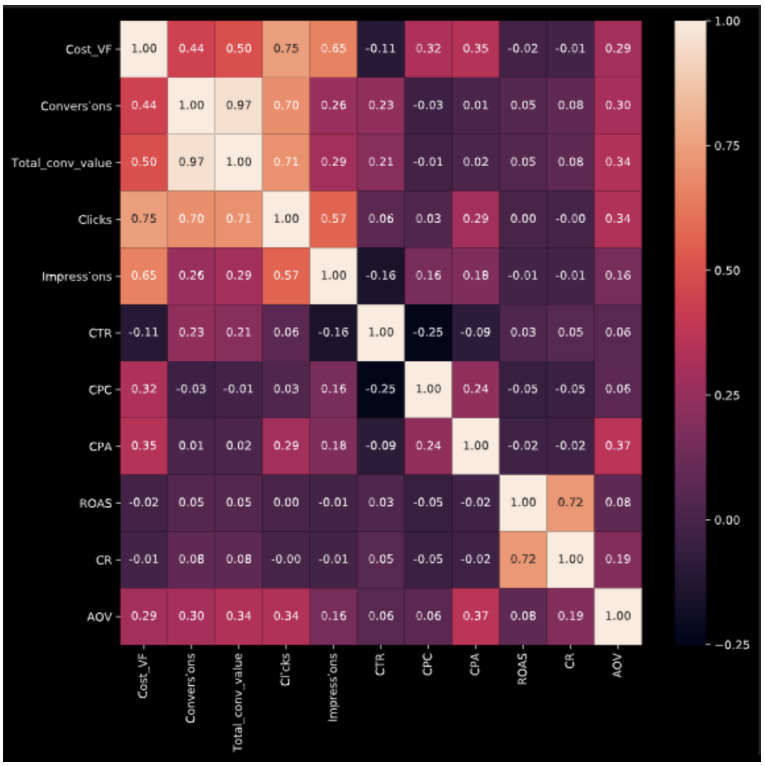


FIGURA 25. Cálculo do coeficiente de correlação linear de Pearson para as variáveis do canal Google Ads.

Em seguida, apresenta-se os gráficos de dispersão para perceber melhor a relação das variáveis umas com as outras. Serão apresentadas as relações das variáveis *Cost VF* e *Total Conv. Value* com as restantes variáveis.

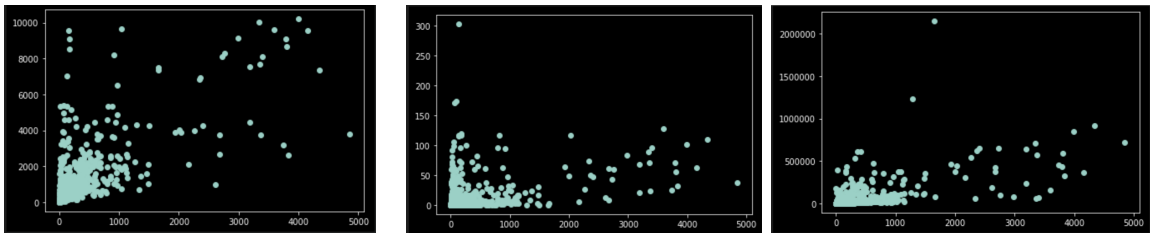


FIGURA 26. Cost VF vs Clicks (esq.); Cost VF vs Conversions (no meio); Cost VF vs Impressions (dir.).

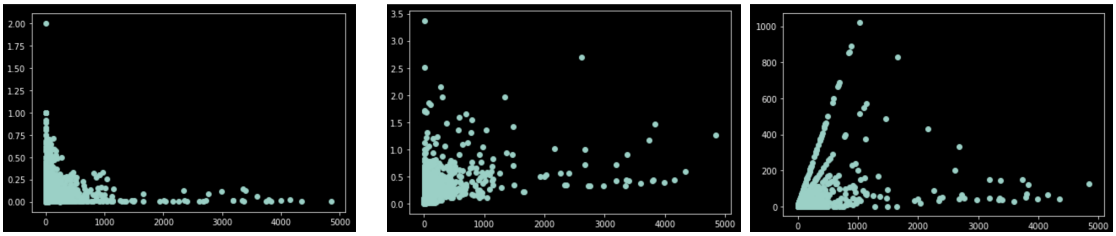


FIGURA 27. Cost VF vs CTR (esq.); Cost VF vs CPC (no meio); Cost VF vs CPA (dir.).

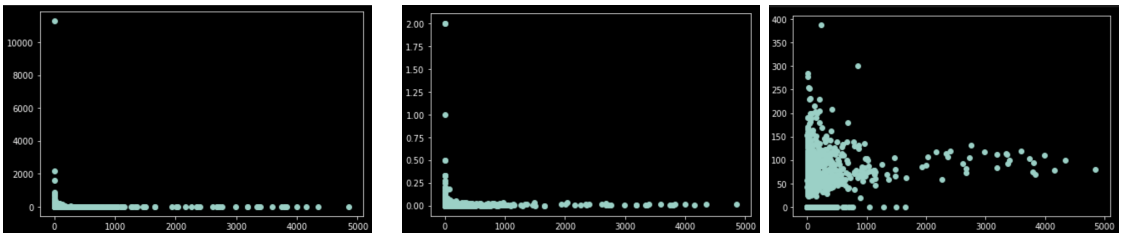


FIGURA 28. Cost VF vs ROAS (esq.); Cost VF vs CR (no meio); Cost VF vs AOV (dir.).

A variável *Cost VF* tem uma relação linear com as 3 variáveis apresentadas nos gráficos da Figura 26 (*Clicks*, *Conversions* e *Impressions*), mas pelos gráficos da Figura 27, as variáveis *CPC* e *CPA* têm alguma correlação com esta variável.

Agora, será apresentada a relação da variável *Total Conv. Value* com as restantes variáveis.

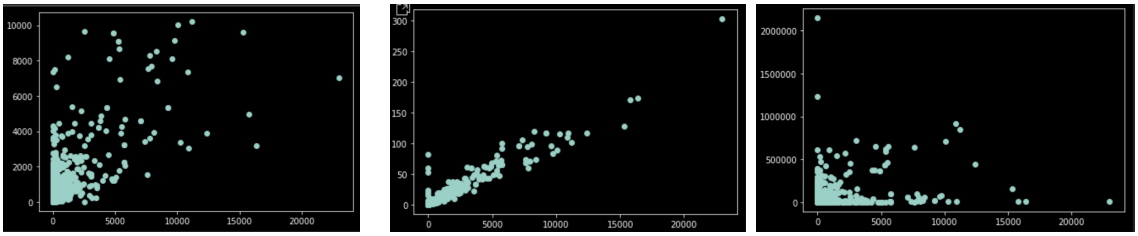


FIGURA 29. Total Conv. Value vs Clicks (esq.); Total Conv. Value vs Conversions (no meio); Total Conv. Value vs Impressions (dir.).

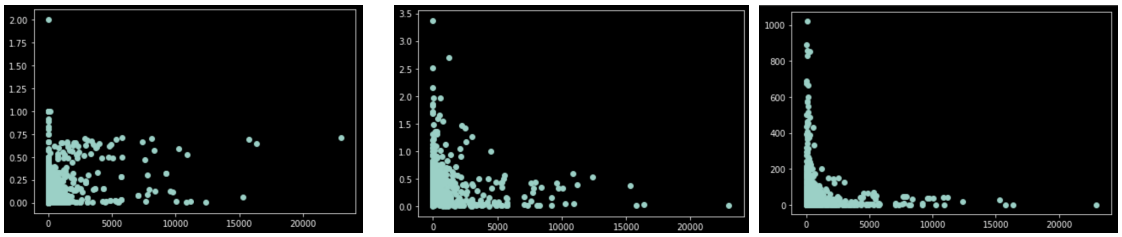


FIGURA 30. Total Conv. Value vs CTR (esq.); Total Conv. Value vs CPC (no meio); Total Conv. Value vs CPA (dir.).

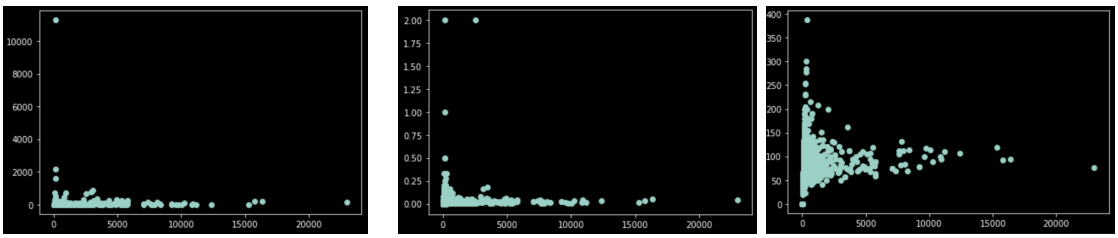


FIGURA 31. Total Conv. Value vs ROAS (esq.); Total Conv. Value vs CR (no meio); Total Conv. Value vs AOV (dir.).

A variável *Total Conv. Value* tem uma relação linear apenas com as variáveis *Conversions* e *Clicks*, segundo os gráficos da Figura 29. A variável *Conversions* é também linear

á variável resposta.

Agora analisando as duas variáveis respostas, no gráfico abaixo, repara-se que as duas variáveis não são muito associadas, ou seja, não são fortemente correlacionadas (0.50), o que se pode supor que provavelmente uma não depende da outra.

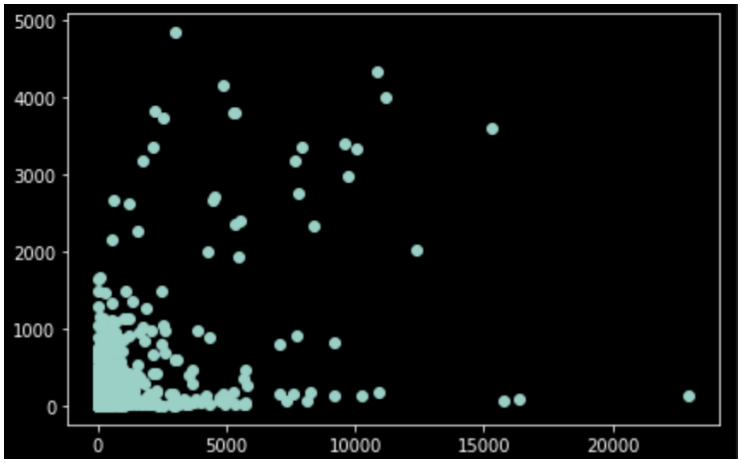


FIGURA 32. Total Conv. Value - CostVF.

Analisou-se os gráficos da função de densidade de probabilidade.

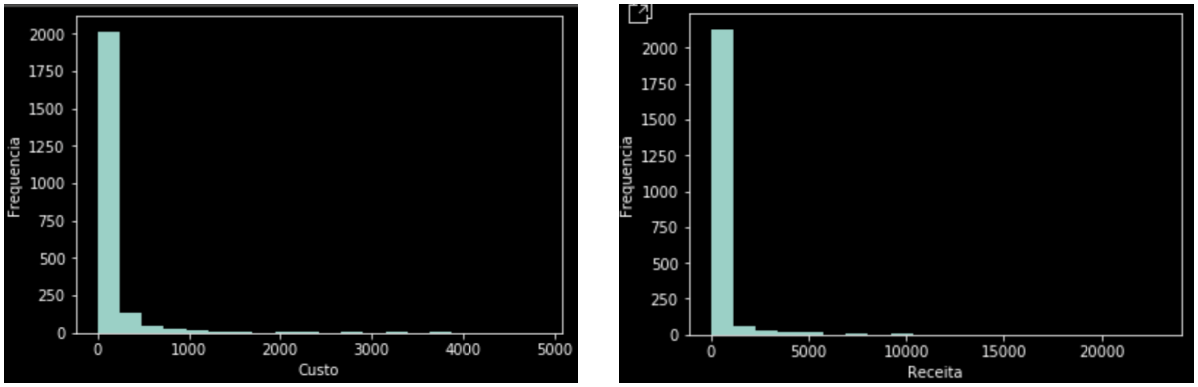


FIGURA 33. Histogramas das variáveis respostas (dir.: Total Conv.Value; esq.: Cost VF).

Os valores da média são superiores aos valores da mediana, o que causa uma curva de distribuição assimetrica positiva, como é conhecido. Diz-se que a assimétria é positiva quando predominam os valores mais altos das observações, isto é, a distribuição ou curva de frequência tem uma "cauda" mais longa à direita da ordenada (frequência) máxima do que à esquerda, e mais uma vez, como podemos ver na Figura 33, em todas as variáveis temos assimetria à direita, os valores aproximam-se mais de zero.

Serão apresentados os gráficos das caixas de bigodes das variáveis respostas, estes apresentam como outliers os valores mais elevados desta base de dados, que são também de grande significância para a mesma. Numa distribuição assimétrica positiva, a tendência é que hajam desvios positivos muito maiores do que negativos.
Pode-se concluir, por exemplo, que o *Cost VF* apresenta maior variabilidade que o *Total Conv. Value*.

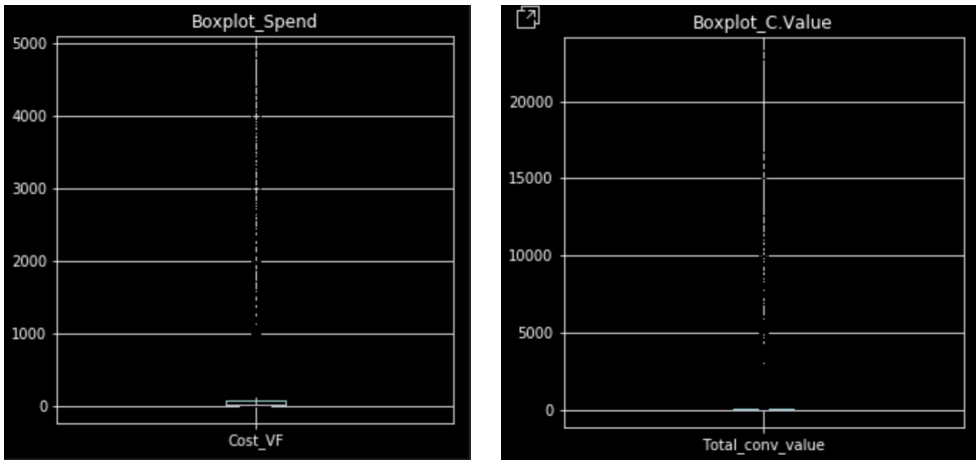


FIGURA 34. Diagrama de extremos e quartis *Cost VF e Total Conv. Value*.

Pela escala das duas caixas de bigodes, entende-se que os valores da variável *Total Conv. Value* são bastante mais elevados que os valores da variável *Cost VF*, o que é um

sinal positivo.

Mais uma vez, os chamados de *outliers*, também neste caso, são os valores considerados grandes.

No gráfico da Figura 35 temos uma análise por mês relativamente à variável *Cost VF*, facilmente podemos ver que os dados são assimétricos positivos, porque a linha da mediana está muito próxima do primeiro quartil.

Cerca de 75% dos dados rondam os valores mínimos (perto de zero), para os meses de Abril19, Fevereiro19, Janeiro19 e Março19. Os diagramas de caixa indicam uma baixa variabilidade e desvio-padrão.

Os meses de Abril, Junho, Maio e Setembro do ano de 2018 são os que apresentam uma variabilidade maior desta variável resposta.

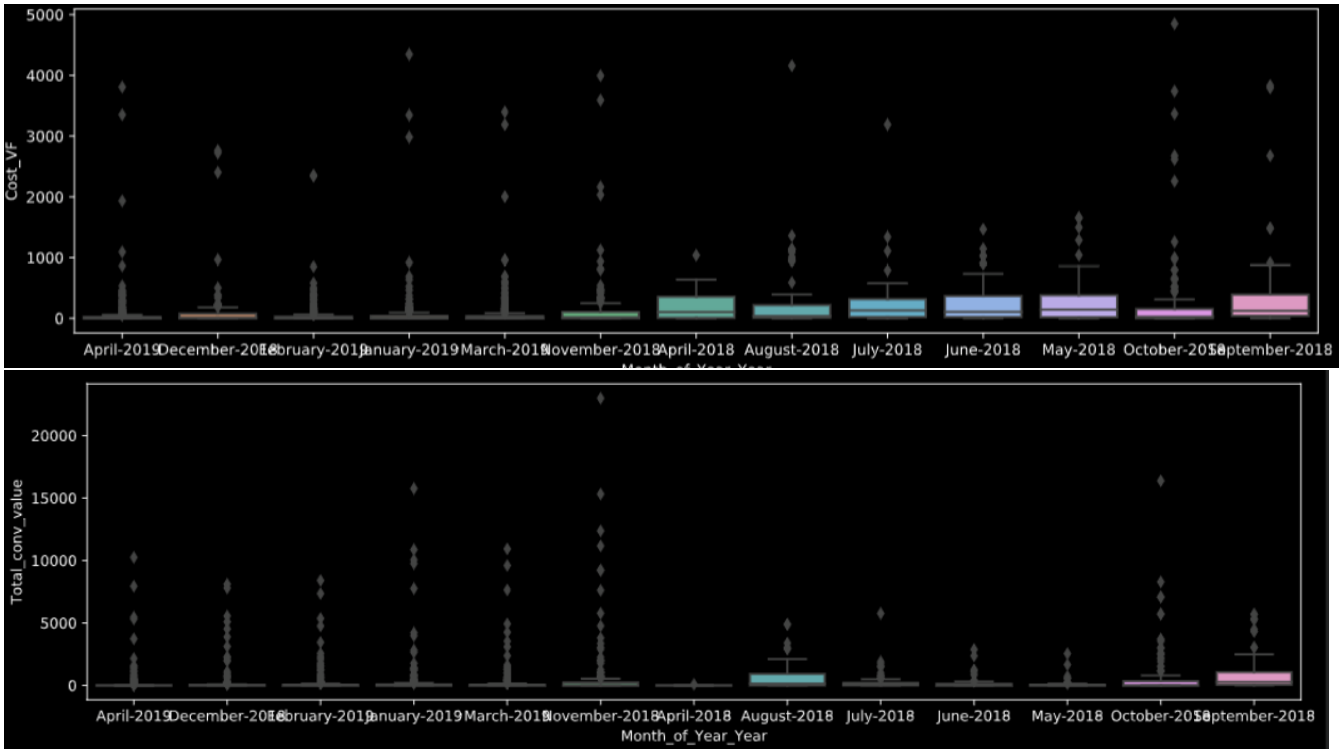


FIGURA 35. Diagrama de extremos e quartis das Variáveis respostas relativamente ao mês.

Para variável *Total Conv. Value*, os meses de maior variabilidade são os meses de Agosto, Outubro e Setembro de 2018.

5. Resultados

Neste capítulo serão implementadas duas técnicas mencionadas anteriormente no capítulo 2.

5.1. Análise de Regressão Múltipla

Os resultados do modelo de regressão múltipla para a previsão do custo e da receita a ser recebido pela empresa Overcube para os canais Facebook e Google Ads são realizados nos capítulos a seguir e seguem as fases descritas no esquema da Figura 3.

5.1.1. Planeamento.

Como descrito no capítulo 4.1., para o Facebook Ads, os valores do custo e da receita são bastante variados. Essa oscilação gera dificuldades para a empresa e até mesmo para os agentes de mercado tomarem a melhor decisão sobre momento de venda de determinado produto, disponibilizando para o cliente um determinado tipo de campanha com o objetivo de que a venda do produto seja um sucesso. Como descrito no capítulo 3.1, a base de dados original é constituída por 26 variáveis e 1237 campanhas, que após uma análise detalhada resumiu-se a 19 variáveis importantes para prever o custo e a receita da empresa.

Para o Google Ads, a abordagem é semelhante. Da base de dados de 2276 observações, com 16 variáveis, foram consideradas e analisadas 11. A previsão do preço de custo e de receita serão efectuadas com base nas variáveis selecionadas para cada situação através da modelação de equações de regressão linear múltipla.

O trabalho realizado divide-se em duas fases. Na primeira fase procedeu-se à seleção das variáveis dependentes (variáveis respostas) tendo em consideração as condições e interesses da empresa. Na segunda fase, prosseguiu-se com a modelação das equações de regressão.

Sempre que possível os dados padronizados, pois a padronização contribui para a estabilização dos processos, assegurando a melhor forma de execução à das variáveis em diferentes escalas.

5.1.2. Seleção e ajuste dos dados - Facebook Ads.

Na procura do melhor modelo de regressão que modele a relação das variáveis dependentes com as variáveis independentes específicas, foram construídos diferentes modelos de modo a encontrar por fim a melhor solução.

É de referenciar que foi feita uma análise exploratória mais aprofundada, de modo a obter as base de dados BD1 e BD2 de acordo a que os dados sejam mais homogéneos e não foram utilizados os valores pertencentes ao arranque da empresa, por estes dificultarem assim o ajuste do modelo. Sendo que para a base de dados BD1, utilizou-se os valores acima do primeiro quartil (43.33) da base de dados original, e para a BD2, os valores acima do segundo quartil (47.92). Para cada uma das variáveis dependentes,*Spend* e *Conversion Value*, as variáveis explicativas serão referidas a seguir:

- ***Spend* - BD1** - Add to Cart, Reach, Results, Cost per Result, CPA, CPC e CPM;
- ***Conversion Value* - BD2** - Cost per Result, ROAS, Conversions, Add to Cart e o Engagement.

Esta escolha de variáveis é feita, não pela associação/correlação da variável resposta com elas, mas sim pela sua forma de obter segundo a sua fórmula, como é o caso das variáveis CPA, CPC e CPM para a variável *Spend*, e a variável ROAS para a variável *Conversion Value*. Um dos fatores importantes, é o seu significado, por exemplo, as variáveis como o *engagement* e *Conversions* são indicadores de maior receita, por isso estão incluídas como importantes para montar o seu modelo.

A fim de solucionar problemas como o da variância não constante e não normalidade dos erros, realizou-se uma transformação nos dados. Apesar de ser possível, em muitos casos, selecionar empiricamente a transformação adequada, apresentar-se-á aqui apenas a técnica mais formal e objetiva, usando dois métodos:

- **Transformação de Box-Cox**: Quando a distribuição normal não se adequa aos dados, muitas vezes é útil aplicar a transformação de Box-Cox para obtermos a normalidade. Considerando X_1, ..., X_n os dados originais, a transformação de Box-Cox consiste em encontrar um λ tal que os dados transformados Y_1, ..., Y_n se aproximem de uma distribuição normal. Esta transformação é dada por:

$$y_i(\lambda) = \begin{cases} \ln(X_i), & \text{se} \quad \lambda = 0 \\ \frac{X_i^{\lambda}-1}{\lambda}, & \text{se} \quad \lambda \neq 0 \end{cases} \tag{1}$$

  As variáveis que sofreram transformação *Box-Cox* foram as variáveis *Spend*, *Add to Cart*, CPC, CPM, ROAS e *Conversion*, com a ajuda da função *Box-Cox.lambda()* do pacote *forecast* no software R, onde o valor de λ é escolhido de forma a maximizar a log-likelihood de um modelo linear ajustado para a tal variável.
- **Transformação de Yeo-Johnson**: Como a transformação de Box-Cox é válida apenas para valores positivos de y, melhorou-se esta tranformação e embora seja possível efetuar uma troca de parâmetros, em caso de valores negativos para utilização da transformação de Box-Cox, existe o inconveniente de tal ação afetar a teoria que suporta a definição do intervalo de confiança de λ. Então, graças a

Yeo & Johnson (2000), elaborou-se uma nova família de transformação de dados, válida tanto para valores positivos como para valores negativos da variável x. Esta transformação é dada pela seguinte família de transformações:

$$y^{(\lambda)} = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \text{se} \quad y \geq 0, \lambda \neq 0 \\ \ln(y+1), & \text{se} \quad y \geq 0, \lambda = 0 \\ \frac{-((-x+1)^{2-\lambda}-1)}{2-\lambda}, & \text{se} \quad y < 0, \lambda \neq 2 \\ -\ln(-y+1), & \text{se} \quad y < 0, \lambda = 2 \end{cases} \tag{2}$$

A função *preProcess()* do pacote *caret* no software R dá uma estimativa a partir dos dados e pode ser aplicada a qualquer conjunto de dados com as mesmas variáveis. Usando o método do *YeoJohnson*, este calcula a verosimilhança perfilhada do parâmetro λ, para as seguintes variáveis - *Conversion Value*, *Reach*, *Results*, *Cost per Result*, CPA e Engagement.

É apresentada uma matriz de gráficos de dispersão, *Scatterplot*, onde surge os histogramas das variáveis na diagonal e respectivos diagramas de dispersão na parte inferior da matriz. Da relação das variáveis escolhidas para a modelação das variáveis respostas, estas são todas minimamente correlacionadas. Como se pode observar, o número de variáveis explicativas é menor no caso da variável receita.

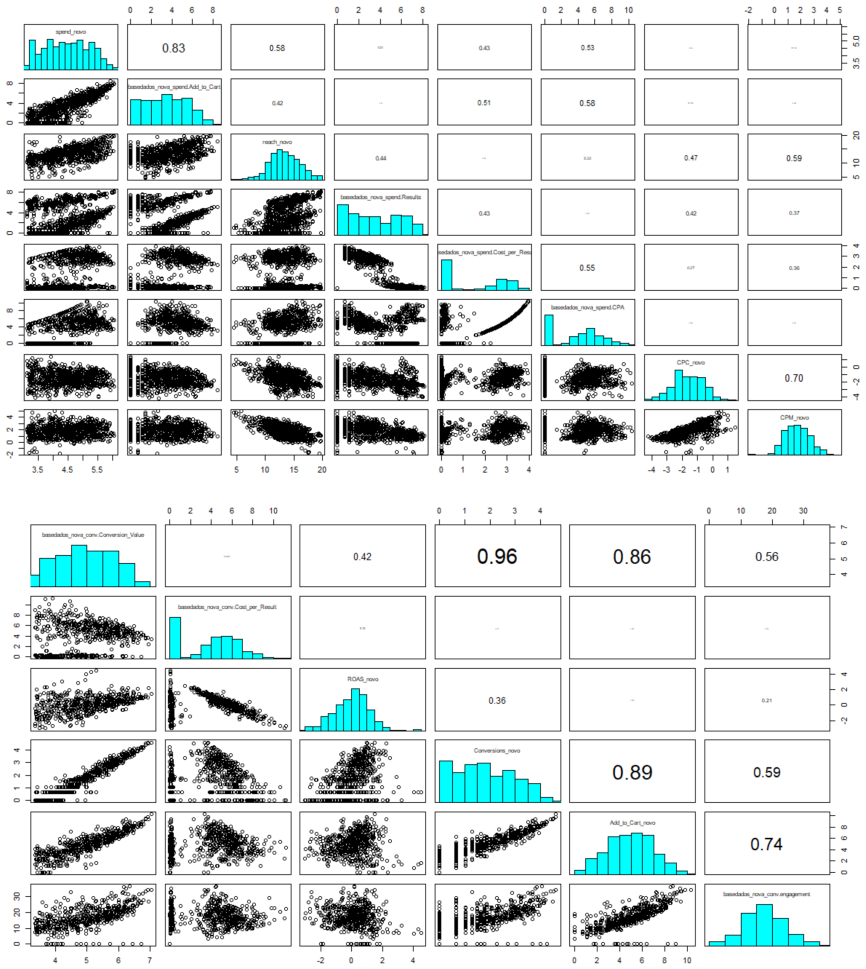


FIGURA 36. Scatterplot para a variável *Spend* (em cima) e para a variável *Conversion Value* (em baixo).

A análise prosseguiu com uma Análise de Componentes Principais.

A Análise de Componentes Principais (ACP) ou *Principal Component Analysis* (PCA) é uma técnica de análise multivariada que usa transformações ortogonais de um conjunto de variáveis, possivelmente correlacionadas para um outro conjunto de variáveis que são linearmente não correlacionadas, chamadas de componentes principais.

É possível encontrar um meio de condensar a informação contida em várias variáveis originais num conjunto menor de variáveis estatísticas (componentes) com uma perda mínima de informação. Com a função *prcomp()* do pacote *stats* no software R, foi efetuada a ACP e aqui utilizou-se a matriz de covariância para a variável *Spend*, ou seja, a BD1, e a matriz de correlação para a variável *Conversion Value*, ou seja, a BD2, isto acontece porque a variância das variáveis de cada base de dados, no caso da BD1 são bastante semelhantes, mas da BD2 o mesmo não acontece, por isso é necessário usar a matriz de correlação.

Tabela 14. Resumo da metodologia ACP para a variável Spend.

COMPONENTE	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Desvio Padrão	3.69	3.00	1.81	1.48	0.97	0.81	0.46	0.21
Proporção da Variância	0.45	0.29	0.10	0.07	0.03	0.02	0.007	0.001
Proporção Acumulada	0.45	0.75	0.86	0.93	0.96	0.991	0.998	1

Tabela 15. Resumo da metodologia ACP para a variável Conversion Value.

COMPONENTE	PC1	PC2	PC3	PC4	PC5	PC6
Desvio Padrão	1.836	1.140	0.990	0.497	0.238	0.190
Proporção da Variância	0.562	0.216	0.164	0.041	0.009	0.006
Proporção Acumulada	0.562	0.778	0.9433	0.984	0.993	1

Para a Tabela 14, tabela da *Spend*, os resultados relativos ao desvio padrão são os valores singulares da matriz de dados centrada. Nas matrizes de covariância os valores singulares coincidem com os valores próprios, e os vetores próprios coincidem com os vetores singulares.

O número de componentes a escolher, quando é usada a matriz de covariância, é o valor da média dos seus valores próprios, o que dá um valor de 1.56, e escolhemos até à terceira componente pois o valor do desvio padrão é superior à média até essa componente. Pode-se notar que as 3 primeiras componentes são capazes de explicar mais de 86% da variabilidade das amostras.

Para a Tabela 15, tabela da *Conversion Value*, segundo *Kaiser*, um dos métodos muito utilizados, assim como na ACP, é o método de retenção com base nos λ's de *Kaiser*. É com base na porcentagem de contribuição da variabilidade total de cada componente que é realizada a escolha do modelo de k componentes. Um critério muito utilizado na retenção de fatores é o de *Kaiser* que afirma que os componentes com $\lambda_i > 1$ representam parcela suficiente da variação total dos dados. Então, sendo assim, para a variável *Conversion Value*, escolhe-se até à segunda componente, e estas duas são capazes de explicar cerca de 78% da variabilidade.

Tabela 16. Pesos das variáveis em cada componente para a variável resposta Spend.

Loadings	PC1	PC2	PC3
spend	0.157	-	-
Add to Cart	0.429	0.151	0.233
reach	0.518	-0.458	0.555
Results	0.205	-0.646	-0.635
Cost per Result	0.154	0.333	-
CPA	0.669	0.414	-0.442
CPC	-	0.170	-
CPM	-	0.198	-0.147

TABELA 17. Pesos das variáveis em cada componente para a variável resposta Conversion Value.

Loadings	PC1	PC2	
Conversion Value	-	-	
Cost per Result	-	0.994	
ROAS	-	-	
Conversions	0.111	-	-
Add to Cart	0.225	-	–
engagement	0.964	-	-

Na Tabela 16, temos os pesos de cada variável nas componentes, quando a variável resposta é *Spend*. Facilmente se ve que a variável *CPA* e *Reach* têm um grande peso na componente 1 e na componente 2, é a variável *Results*, com uma associação negativa.

Para a variável *Spend*, se optássemos por utilizar essas 3 componentes, estaríamos reduzindo o número de 8 variáveis originais para 3 variáveis latentes, perdendo menos de 14% da informação acerca da variabilidade dos dados.

Na Tabela 17, apresentam-se os pesos de cada componente, e segundo essa, para a componente 1, a variável *engagement* é a mais "pesada"nessa componente, assim como a variável *Cost per Resut* para a componente 2. Para a variável *Conversion Value*, reduz-se de 6 variáveis para apenas duas, perdendo assim, menos de 22% da informação acerca da variabilidade dos dados.

A matriz de loadings de cada variável nas componentes principais ao ser multiplicada pela matriz original de dados fornece a matriz de contagens (scores) de cada caso em relação às componentes principais.
É essa nova matriz de scores que vai ser a nova base de dados para encontrar o melhor modelo de regressão.

Os valores assim obtidos, denominados scores de cada observação na CP, podem obter-se através da função *prcomp()* solicitando que seja exibido no objeto de saída da função a componente que é relativa ao valor dos *scores*.

De seguida, apresentarei os modelos para as duas variáveis respostas serão apresentadas a seguir.

As equações 1 e 2 apresentam um modelo da previsão das variáveis respostas Custo e da Receita.

Spend (Modelo inicial):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \beta_3 SCORES.PC3 + \epsilon \tag{3}$$

onde:

- Y : Custo (variável dependente);
- β_0 : Constante;
- β_i : Coeficientes das variáveis independentes, $i = 1, \cdots, 3$;

- ϵ : erro aleatório.

Conversion Value (Modelo inicial):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{4}$$

onde:

- Y : Receita (variável dependente);
- β_0 : Constante;
- β_i : Coeficientes das variáveis independentes, $i = 1, \cdots, 2$;
- ϵ : erro aleatório.

Analisando as variáveis ao longo do tempo, percebe-se que tanto a variável dependente como as independentes apresentam algum padrão de comportamento.

A primeira componente é a única que tem alguma associação com a variável resposta *Spend*, cerca de 0.157, nada significativo. Com a variável *Conversion Value*, a situação é diferente, nenhuma das componentes tem tendência de elevação da receita, mas juntamente com as restantes variáveis assegurou a melhor previsão.

Essa característica pode ser causada pelo efeito da inflação e ou por outras variações temporais.

5.1.3. Estimação - Facebook Ads.

Após a seleção do melhor modelo para a receita e o custo será utilizado o *teste da Razão de Verosimilhança (LRT)* [14] para a estimação dos coeficientes envolvidos no modelo e o *Critério de Informação de Akaike [15] (AIC)* para validar o modelo. A LRT compara dois modelos aninhados, testando se os parâmetros aninhados do modelo mais complexo diferem significativamente do valor nulo. Um modelo mais simples (com menos parâmetros) é aninhado em outro, mais complexo (com mais parâmetros), se o modelo complexo for reduzido para o mais simples pela retirada de um dos parâmetros. [15]

O critério de Akaike é uma ferramenta para seleção de modelos, pois oferece uma medida relativa do goodness-of-fit (qualidade do ajuste) de um modelo estatístico. AIC não fornece um teste de um modelo no sentido usual de testar uma hipótese nula, ou seja, ele não pode dizer nada sobre o quão bem o modelo ajusta os dados em um sentido absoluto. [15]

Numa fase inicial, foi construído o modelo com todas as variáveis que, porventura, seriam as adequadas para ingressar no modelo por indicação da Overcube. Numa segunda fase, foi formulado o modelo através de um método iterativo, efetuado através da função **drop1()** em ambiente R, na qual se retira a variável com valor de prova superior, uma a uma. Os modelos finais obtidos foram os seguintes:

Spend (Modelo drop1):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC3 + \epsilon \tag{5}$$

Conversion Value (Modelo drop1):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{6}$$

A variável *Conversion Value* preservou as duas variáveis, enquanto que a variável *Spend*, com o *drop1()*, retirou-se a variável SCORES.PC2.
As Tabela 18 e 19 indicam os valores das estimativas dos coeficientes do modelo obtido pelo método iterativo.

TABELA 18. Coeficientes do modelo de equação 5, pelo teste da razão de verosimilhança.

Variáveis	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de Prova
(Intercept)	4.461978	0.014724	303.040	<2e-16
SCORES.PC1	0.156585	0.003980	39.343	<2e-16
SCORES.PC3	0.071692	0.008112	8.838	<2e-16

TABELA 19. Coeficientes do modelo de equação 6 pelo teste da razão de verosimilhança.

Variáveis	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de Prova
(Intercept)	4.945108	0.007880	627.58	<2e-16
SCORES.PC1	-0.467767	0.004295	-108.92	<2e-16
SCORES.PC2	0.148345	0.006917	21.45	<2e-16

Tal como no método apresentado anteriormente, no *Critério de Informação de Akaike*, o modelo inicial ajustado é o modelo completo. As variáveis preditoras vão sendo removidas enquanto o valor do AIC do modelo for superior ao valor do AIC do mesmo modelo sem uma das covariáveis, é uma forma de se validar o modelo. Este modelo é obtido, em ambiente R, a partir da função **step()** disponível na biblioteca *stats*. Os modelos finais obtidos foram os seguintes:

Spend (Modelo step):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \beta_3 SCORES.PC3 + \epsilon \tag{7}$$

Conversion Value (Modelo step):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{8}$$

Não se retirou nenhuma variável, em nenhuma das variáveis respostas.
As Tabela 20 e 21 indicam os valores das estimativas dos coeficientes do modelo obtido pelo método iterativo.

Tabela 20. Coeficientes do modelo de equação 7, pelo Critério do AIC.

Variável	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de prova
(Intercept)	4.461978	0.014708	303.376	<2e-16
SCORES.PC1	0.156585	0.003976	39.386	<2e-16
SCORES.PC2	-0.008539	0.004894	-1.745	0.0813
SCORES.PC3	0.071692	0.008103	8.847	<2e-16

Tabela 21. Coeficientes do modelo de equação 8, pelo Critério do AIC.

Variáveis	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de Prova
(Intercept)	4.945108	0.007880	627.58	<2e-16
SCORES.PC1	-0.467767	0.004295	-108.92	<2e-16
SCORES.PC2	0.148345	0.006917	21.45	<2e-16

5.1.4. Qualidade do ajustamento - Facebook Ads.

De forma a comparar os dois modelos obtidos, das duas técnicas, de cada uma das variáveis resposta (*Spend* e *Conversion Value*), com vista à obtenção do melhor modelo que se ajusta aos dados em estudo, efetuou-se um teste F-parcial na ANOVA [16] sob as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_i = 0$$

$$vs$$

$$H_1 : \exists \beta_i \neq 0, i = 1, ..., p$$

Em ambiente R o *output* devolvido para a *Spend* foi o seguinte:

```
Analysis of Variance Table

Model 1: juntartudo.spend_novo ~ (SCORES_quarta.PC1 + SCORES_quarta.PC2 +
    SCORES_quarta.PC3) - SCORES_quarta.PC2
Model 2: juntartudo.spend_novo ~ SCORES_quarta.PC1 + SCORES_quarta.PC2 +
    SCORES_quarta.PC3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    921 184.50
2    920 183.89  1   0.60865 3.0451 0.08131 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURA 37. Teste F-parcial ANOVA para a variável Spend.

A um nível de significância de 5%, não se rejeita a hipótese nula. Foi obtido um valor de prova de 0.081, ou seja, há evidência estatísticas para afirmar que o melhor modelo é aquele que se obtém pelo método da *Máxima Verosimilhança*. Considera-se o modelo obtido por *Máxima Verosimilhança* e o modelo obtido pelo *Critério de Inf. de Akaike* como modelo 1 e modelo 2, respetivamente.

Para a variável *Conversion Value*, o *output* devolvido foi o seguinte:

```
Analysis of Variance Table

Model 1: juntartudo_conv.basedados_nova_conv.Conversion_Value ~ SCORES_uma.PC1 +
    SCORES_uma.PC2
Model 2: juntartudo_conv.basedados_nova_conv.Conversion_Value ~ SCORES_uma.PC1 +
    SCORES_uma.PC2
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1    613 23.445
2    613 23.445  0        0
```

FIGURA 38. Teste F parcial ANOVA para a variável Conversion Value.

Como mostra a Figura 38, foi obtido um valor de prova igual a 0 , logo os modelos são iguais.

Após os resultados do teste F-parcial na ANOVA, foram efetuados vários testes de qualidade de ajustamento, como se observa na Tabela 22. Com a análise dos valores obtidos, comparando os dois modelos para a variável resposta *Spend*, o que tem menor AIC (*Critério de Informação de Akaike*) e menor BIC (*Critério Bayesiano*) é o modelo que se obtém pelo método da *Máxima Verosimilhança* onde comprova que todos os coeficientes são significativos e o modelo explica cerca de 64%.

Para a variável resposta *Conversion Value*, os modelos são completamente iguais. Ambos os modelos explicam cerca de 96% da variabilidade dos dados.

Através das Tabela 18 e Tabela 19 pode-se observar o output as estimativas dos parâmetros, e segundo o teste t utilizado estas estimativas são realmente diferentes de zero, ou seja as variáveis Xs explicam a variabilidade de Y.

TABELA 22. Indicadores de qualidade do ajustamento.

	Modelo drop_spend	Modelo step_spend
R_a^2	0.6376	0.6384
AIC	1141.558	1140.505
BIC	1160.873	1164.649
	Modelo drop_Conversion Value	Modelo step_Conversion Value
R_a^2	0.9525	0.9525
AIC	-257.31	-257.31
BIC	-239.62	-239.62

5.1.5. Análise de diagnóstico - Facebook Ads.

Concluída a escolha dos modelos, deve-se agora avaliar os pressupostos que garantem a validação dos mesmos.

A verificação do pressuposto de normalidade dos resíduos é realizada através de testes que examinam se a série apresenta distribuição próxima à distribuição normal. Para isto, são formuladas as seguintes hipóteses:

H_0 :Os resíduos seguem uma distribuição normal.

vs

H_1 : Os resíduos não seguem uma distribuição normal.

Realizou-se uma análise por meio do teste de *Kolmogorov-Smirnov* com a correção de *Lilliefors* para testar a normalidade, acrescido da explicação gráfica por meio de um histograma de resíduos e do gráfico de normalidade dos resíduos.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod_acp_ori_transf22$residuals
D = 0.029776, p-value = 0.05146
```

FIGURA 39. Teste da normalidade dos resíduos com o teste KS Lilliefors para o modelo da variável resposta Spend.

Pode-se ver na Figura 39 que foi obtido um valor de prova maior que 0.05 no teste de *Kolmogorov-Smirnov* com a correção de *Lilliefors*, logo para um nível de significância de 5% não se rejeita H_0, ou seja, há evidências estatísticas para afirmar os resíduos do modelo cuja a variável resposta é *Spend*; seguem uma distribuição normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod_acp_ori_transf2$residuals
D = 0.037571, p-value = 0.038
```

FIGURA 40. Teste da normalidade dos resíduos com o teste KS Lilliefors para o modelo da variável resposta Conversion Value.

Para a variável *Conversion Value*, na Figura 40, foi obtido um valor de prova maior que 0.01, ou seja, a um nível de 1% não rejeito a hipótese de evidenciar que os resíduos seguem normalidade.

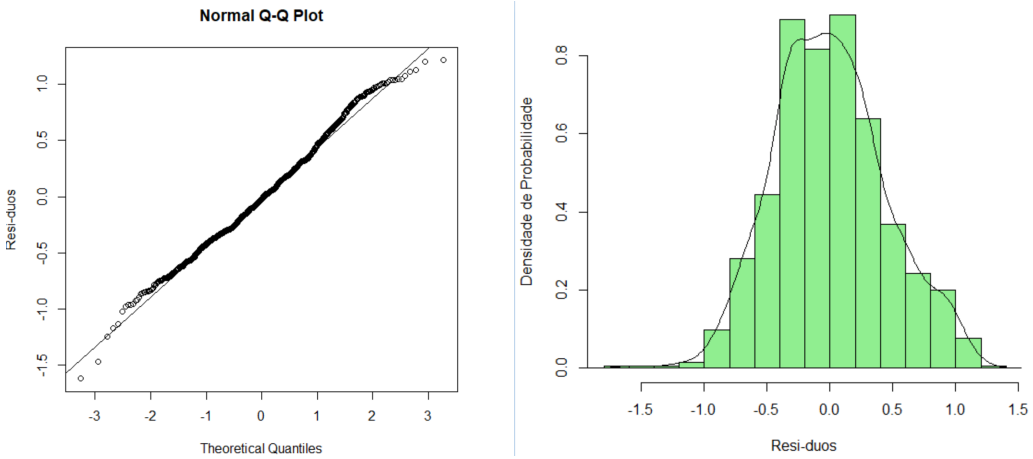


FIGURA 41. Histograma e gráfico da normalidade dos resíduos da variável resposta Spend.

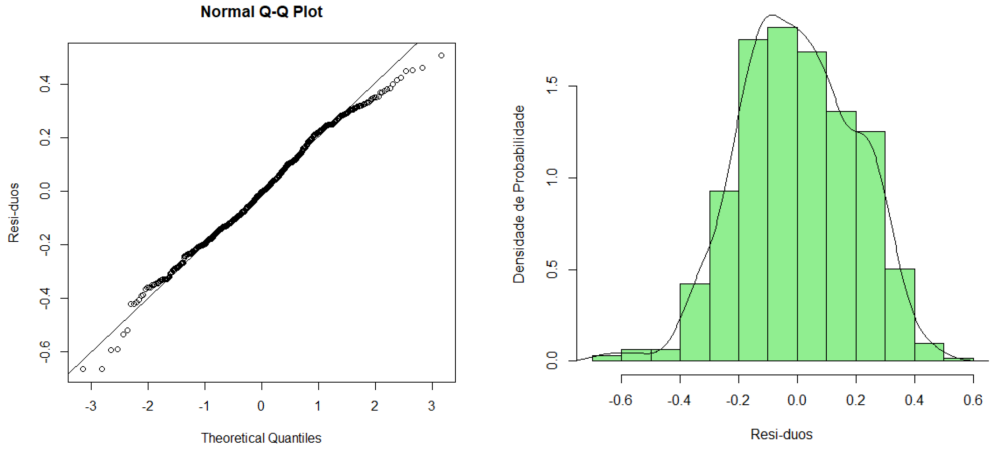


FIGURA 42. Histograma e gráfico da normalidade dos resíduos da variável resposta Conversion Value.

Repare-se nas Figura 41 e 42, no gráfico à esquerda estão apresentados os gráficos da normalidade dos resíduos juntamente com o histograma dos mesmos, no qual se observa

que os resíduos apresentam distribuição normal, visto que os dados se distribuem ao longo da reta.

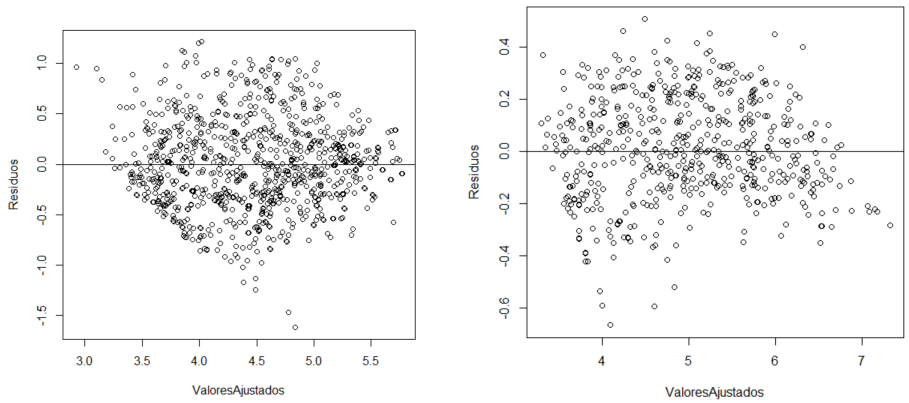


FIGURA 43. Resíduos vs valores estimados para a variável resposta Spend (esq.) e variável resposta Conversion Value(dir.).

Para o diagnóstico de heteroscedasticidade, tentou-se encontrar alguma tendência no gráfico. Por isso, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento ou tendência, temos indícios de que a variância dos resíduos é homoscedástica. Já a presença de "funil"é um indicativo da presença de heteroscedasticidade. Na Figura 43 está representado o gráfico dos resíduos contra os valores estimados para a variável *Spend* e para a variável *Conversion Value*, no qual se observa que a variância dos resíduos é constante ao longo de toda a amostra.

Para análise do pressuposto de multicolinearidade faz-se uso dos coeficientes *Tolerance* ou *VIF* (*Variance Inflaction Factor*), o segundo é calculado a partir do inverso do primeiro. A análise de VIF é feita da seguinte forma:

- Até 1 - sem multicolinearidade;
- De 1 até 10 - com multicolinearidade aceitável;
- Acima de 10 - com multicolinearidade problemática.

TABELA 23. VIF (Inflação de variância).

(A) Variável resposta Spend.

Variável	VIF
SCORES.PC1	1
SCORES.PC3	1

(B) Variável resposta Conversion Value.

Variável	VIF
SCORES.PC1	1
SCORES.PC2	1

Para ambas as variáveis, podemos concluir que as variáveis explicativas têm valor de VIF igual a 1, por isso não apresentam multicolinearidade.

5.1.6. Seleção e ajuste dos dados - Google Ads.

Com a base de dados do Google Ads, a dificuldade foi maior, o que resultou numa análise mais aprofundada. Aqui mais uma vez, apenas usei as variáveis que tinham maior associação com cada variável dependente.

Após uma breve análise exploratória optou-se por não considerar os valores correspondentes ao arranque da empresa. Sendo que para a base de dados BD3, utilizou-se os valores acima do terceiro quartil (74.23) da base de dados original, e para a BD2, os valores acima da média (330.9).

- **Cost VF - BD3** - Cost VF, CTR, CPC, CPA e ROAS;
- **Total Conv. Value - BD4** - Total Conv. Value, Impressions e CTR.

As variáveis como CTR, CPC, CPA são custos associados a cada campanha, por isso estão ligadas ao cálculo do melhor modelo quando se trata da variável *Cost VF*. Para a variável impressions, esta é considerada uma variável de "entrada de dinheiro", no geral, está associada à variável receita.

Com o objetivo de usar a melhor transformação dos dados foi aplicada a função *bestNormalize()* do pacote *bestNormalize* no software R. Esta transformação consiste em executar um conjunto de transformações normalizadas com base no p-value de Pearson. [3]. Atualmente, esta função estima algumas transformações como *Yeo-Johnson*, a transformação *Box-Cox* (se os dados forem positivos), a transformação log_{10} (x + a), entre outras. Usei esta função do bestNormalize para as duas bases de dados (BD3 e BD4), e obtive a seguinte tabela:

Tabela 24. Resultados do output de aplicação do bestNormalize para as variáveis respostas do canal Google Ads.

(a) BD3 - CostVF

Variável	Transformação	Estatística estimada de normalidade
CostVF	orderNorm	1.12
CTR	orderNorm	1.25
CPC	Yeo-Johnson (λ = -2.43)	0.95
CPA	orderNorm	2.92
ROAS	orderNorm	3.33
CR	orderNorm	2.85

(b) BD4 - Total Conv. Value

Variável	Transformação	Estatística estimada de normalidade
Total Conv. Value	orderNorm	1.12
impressions	Box-Cox (λ = -0.04)	1.09
CTR	orderNorm	1.27

Conhecendo a transformação de *Box-Cox* (Equação 1) e de *Yeo-Johnson* (Equação 2), agora com a transformação de normalização do *Ordered Quantile*, *orderNorm()*, o procedimento é baseado numa classificação pela qual os valores de um vetor são mapeados

para o seu percentil e, em seguida, mapeados para o mesmo percentil da distribuição normal, sem a presença de laços, o que garante essencialmente que a transformação seja uma distribuição uniforme.

Na Tabela 24, temos descritas, para as duas base de dados (BD3 e BD4) as transformações utilizadas para cada variável, com a função *bestNormalize*.

Elaborou-se o *Scatterplot* de cada base de dados, e verificou-se que na variável resposta BD3, a variável CR ainda é correlacionada com ROAS (0.84), assim como na variável resposta BD4, a CR é correlacionada com impressions (0.71).

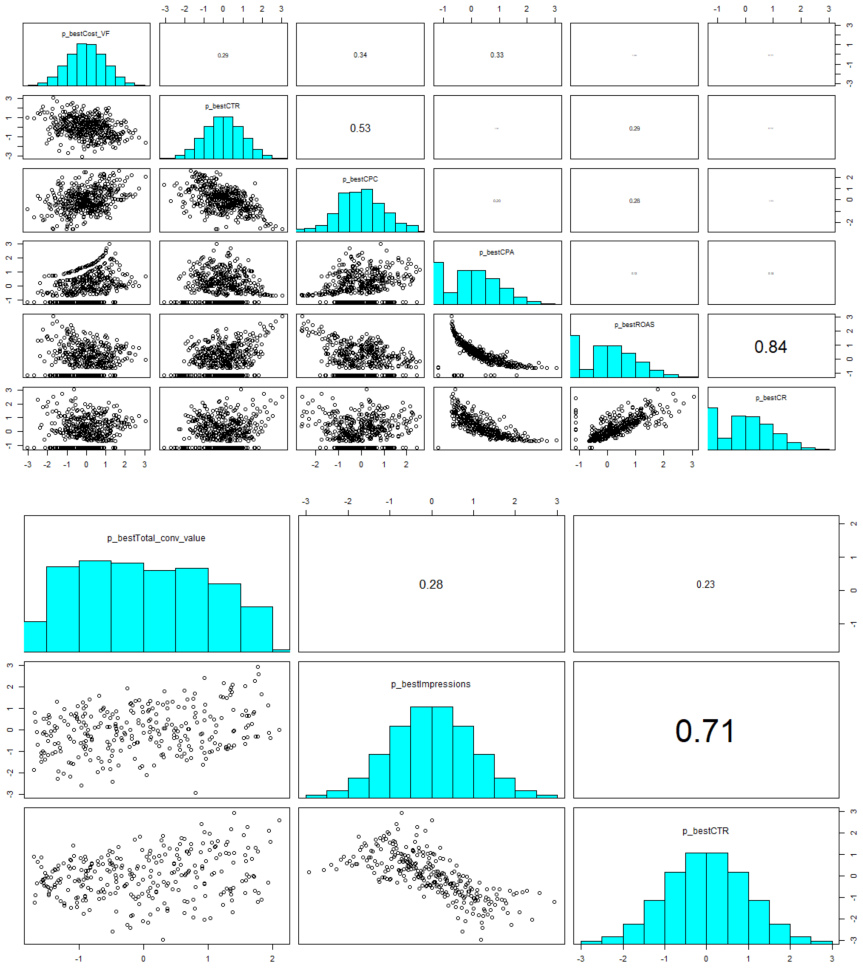


FIGURA 44. Scatterplot para as variáveis respostas BD3 (em cima) e BD4 (em baixo).

Mais uma vez, também para uma melhor modelação do custo e da receita do *Google Ads*, foi elaborada uma ACP.

Através da função *prcomp()* do pacote *stats* no software R, foi efetuada a ACP. Para ambas as variáveis, *CostVF* e *Total Conv.Value*, foi usada a matriz de covariância, pois a variância das variáveis em questão nas duas base de dados (BD3 e BD4) são próximas, não são consideradas discrepantes.

O número de componentes principais torna-se o novo número de variáveis que serão consideradas em diante na análise. As as primeiras componentes são as mais importantes, já que explicam a maior parte da variação total.

Tabela 25. Resumo da metodologia ACP para a variável CostVF.

COMPONENTE	PC1	PC2	PC3	PC4	PC5	PC6
Desvio Padrão	1.419	1.318	0.906	0.770	0.678	0.301
Proporção da Variância	0.352	0.303	0.143	0.103	0.080	0.015
Proporção Acumulada	0.352	0.656	0.799	0.903	0.984	1

Tabela 26. Resumo da metodologia ACP para a variável Total Conv. Value.

COMPONENTE	PC1	PC2	PC3
Desvio Padrão	1.311	1.065	0.363
Proporção da Variância	0.575	0.380	0.044
Proporção Acumulada	0.575	0.955	1

Quando se utiliza a matriz de covariância para extração, as componentes são influenciadas pelas variáveis de maior variância.

Para descobrir quantas componentes escolher, e sabendo que foi usada a matriz de covariância em vez da matriz de correlação, olhamos para a média dos seus valores próprios, que dá um valor de 0.891 para *CostVF*, e de 0.911 para a variável *Total Conv. Value*, ou seja, para a modelação e previsão do custo escolhemos até à terceira componente, e para a receita, até à segunda, pois, o valor é superior à média calculada. Pode-se notar que as três primeiras componentes são capazes de explicar cerca de 80% da variabilidade das amostras, para a BD3, e as duas primeiras explicam 95% da variabilidade da amostra da BD4.

Tabela 27. Pesos das variáveis em cada componente para a variável Cost VF.

Loadings	PC1	PC2	PC3
Cost VF	0.365	0.421	0.381
CTR	-0.546	-	0.456
CPA	0.159	0.399	0.625
ROAS	-0.413	0.512	-0.203
CR	-0.267	0.596	-0.351

Tabela 28. Pesos das variáveis em cada componente para a variável Total Conv. Value.

Loadings	PC1	PC2
Total Conv. value	-	0.927
Impressions	0.714	0.230
CTR	-0.698	0.297

Nas tabelas anteriores, temos os *loadings* que são os pesos pelo qual cada variável normalizada original deve ser multiplicada para se obter a pontuação de componente.

Nas Tabela 27 e 28 estão representados os pesos de cada variável em cada componente. A Tabela 27 refere-se à BD3, ou seja, o custo, e consta que apesar de negativa a variável *CTR* tem um peso de -0.546 na componente 1, e a variável *CPA*, um peso de 0.62 positivo na componente 3. A variável custo, tem um peso de 0.42 na componente 2.

A Tabela 28 trata a BD4, e explica que a variável resposta, receita, tem um peso enorme de 0.927 na componente 2, e *impressions* e *CTR* na componente 1 (0.71 e 0.69).

À BD3 reduzimos de 6 variáveis para 3, e à BD4 de 3 para duas, perdendo assim apenas 5% da informação.

Os valores assim obtidos, denominados *scores* de cada observação na CP, podem obter-se através da função *prcomp()* solicitando que seja exibido no objeto de saída da função a componente que é relativa ao valor dos *scores*. Com os scores, a análise continua com o estudo dos modelos.

As equações 1 e 2 apresentam um modelo esquemático da previsão do *Cost VF* e da *Total Conv. Value*.

Cost VF (Modelo inicial):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \beta_3 SCORES.PC3 + \epsilon \tag{9}$$

onde:

- Y : Custo (variável dependente)
- β_0 : Constante;
- β_i : Constantes das variáveis independentes, $i = 1, \cdots, 3$
- ϵ : erro aleatório.

Total Conv. Value (Modelo inicial):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{10}$$

onde:

- Y : Receita (variável dependente)
- β_0 : Constante;
- β_i : Constantes das variáveis independentes, $i = 1, \cdots, 2$
- ϵ : erro aleatório.

5.1.7. Estimação - Google Ads.

Em seguida, foram aplicados novamente o *teste da Razão de Verosimilhança* e o *Critério de Informação de Akaike (AIC).*. Os modelos finais obtidos foram os seguintes:

Cost VF (Modelo drop1):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \beta_3 SCORES.PC3 + \epsilon \tag{11}$$

Total Conv. Value (Modelo drop1):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{12}$$

Não foi necessário retirar nenhuma variável, para ambos os modelos, segundo o teste da Razão de Verosimilhaça.

Tabela 29. Coeficientes do modelo da equação 11, pelo teste da razão de verosimilhança.

Variáveis	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de Prova
(Intercept)	5.210e-07	2.723e-02	0.00	1
SCORES.PC1	-3.652e-01	1.925e-02	-18.97	<2e-16
SCORES.PC2	4.214e-01	2.125e-02	19.83	<2e-16
SCORES.PC3	-3.811e-01	3.046e-02	-12.51	<2e-16

Tabela 30. Coeficientes do modelo da equação 12 pelo teste da razão de verosimilhança.

Variáveis	$\hat{\beta}$	Erro Padrão	Estatística de Teste	Valor de Prova
(Intercept)	1.421e-05	8.405e-03	0.002	0.999
SCORES.PC1	-6.536e-02	6.443e-03	-10.145	<2e-16
SCORES.PC2	9.202e-01	7.865e-03	116.998	<2e-16

Tal como foi feito para as base de dados anterios, também será testado o método do AIC. Este modelo é obtido, em ambiente R, a partir da função **step()** disponível na biblioteca *stats*. Os modelos finais obtidos foram os seguintes:

Cost VF (Modelo step):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \beta_3 SCORES.PC3 + \epsilon \tag{13}$$

Total Conv. Value (Modelo step):

$$Y = \beta_0 + \beta_1 SCORES.PC1 + \beta_2 SCORES.PC2 + \epsilon \tag{14}$$

Podemos concluir que os modelos são iguais, para ambos os métodos, não é necessário retirar nenhuma variável. Os valores dos coeficientes obtidos pelo *Critério de Inf. de Akaike* são iguais aos coeficientes do *teste da Razão de Verosimilhaça*, por isso o capítulo "Qualidade do ajustamento"não será necessária para o *Google Ads*, porque obtiveram-se resultados iguais.

Através das Tabela 29 e Tabela 30 pode-se observar o output as estimativas dos parâmetros, e segundo o teste t utilizado estas estimativas são realmente diferentes de zero, ou seja as variáveis Xs explicam a variabilidade de Y.

TABELA 31. Indicadores de qualidade do ajustamento.

	Modelo drop_CostVF	Modelo step_CostVF
R_a^2	0.6742	0.6742
AIC	759.205	759.205
BIC	779.62	779.62
	Modelo drop_Total Conv. Value	Modelo step_Total Conv. Value
R_a^2	0.9813	0.9813
AIC	-341.63	-341.63
BIC	-326.77	-326.77

5.1.8. Análise de diagnóstico - Google Ads.

Concluída a escolha dos modelos, deve-se agora avaliar os pressupostos que garantem a validação dos mesmos.

A verificação do pressuposto de normalidade dos resíduos é realizada através de testes que examinam se a série apresenta distribuição próxima à distribuição normal. Para isto, são formuladas as seguintes hipóteses:

H_0 :Os resíduos seguem uma distribuição normal.

vs

H_1 : Os resíduos não seguem uma distribuição normal.

Realizou-se novamente uma análise por meio do teste de *Kolmogorov-Smirnov* com a correção de *Lilliefors* para testar a normalidade, acrescido da explicação gráfica por meio de um histograma de resíduos e do gráfico de normalidade dos resíduos.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod_t$residuals
D = 0.028556, p-value = 0.5193
```

FIGURA 45. Teste da normalidade dos resíduos com o teste KS Lilliefors para o modelo da variável resposta CostVF.

Podemos ver pela Figura 45 que foi obtido um valor de prova maior que 0.05, logo para um nível de significância de 5% não se rejeita H_0, ou seja, há evidências estatísticas

para afirmar que os resíduos do modelo cuja a variável resposta é *Cost VF* seguem uma distribuição normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  mod_t$residuals
D = 0.048854, p-value = 0.07701
```

FIGURA 46. Teste da normalidade dos resíduos com o teste KS Lilliefors para o modelo da variável resposta Total Conv. Value.

Para a variável *Total Conv. Value*, na Figura 46, foi obtido um valor de prova maior que 0.05, ou seja, a um nível de 5% não rejeito a hipótese de evidenciar que os resíduos seguem normalidade.

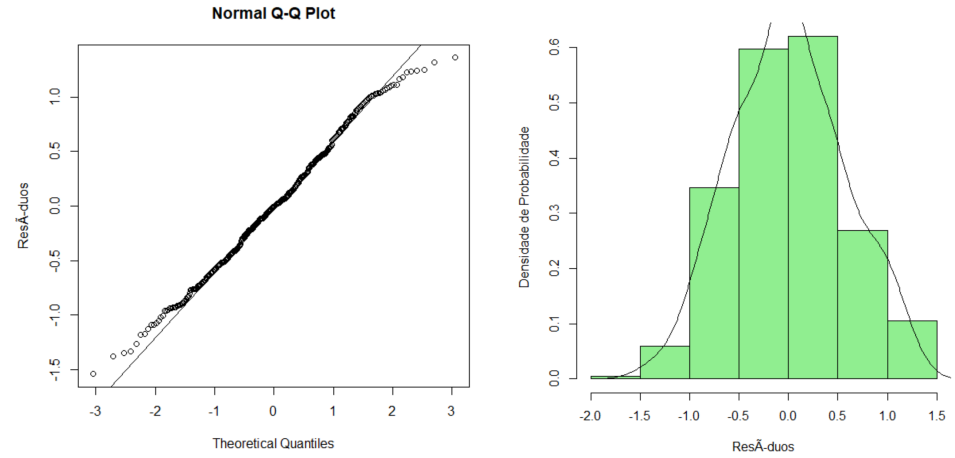


FIGURA 47. Histograma e gráfico da normalidade dos resíduos da ariável resposta Cost VF.

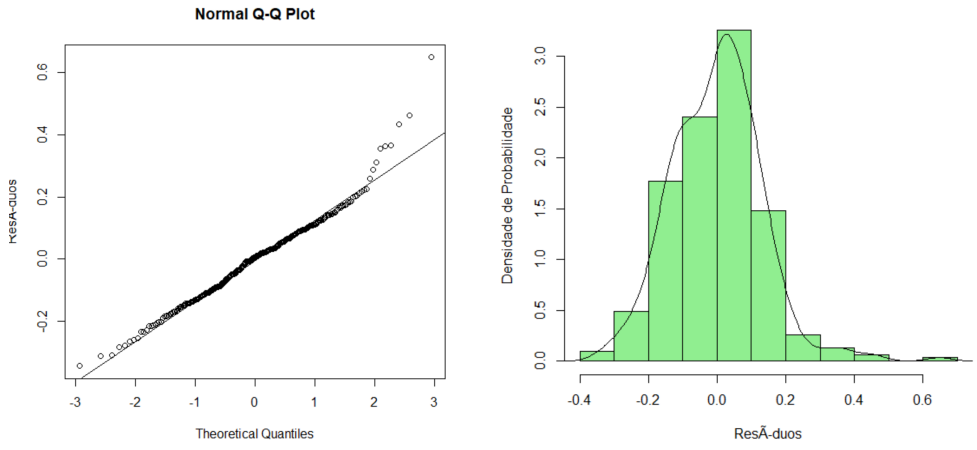


FIGURA 48. Histograma e gráfico da normalidade dos resíduos da variável resposta Total Conv. Value.

Observa-se que os resíduos apresentam distribuição normal, Figura 47 e 48, os dados distribuem-se ao longo da reta.

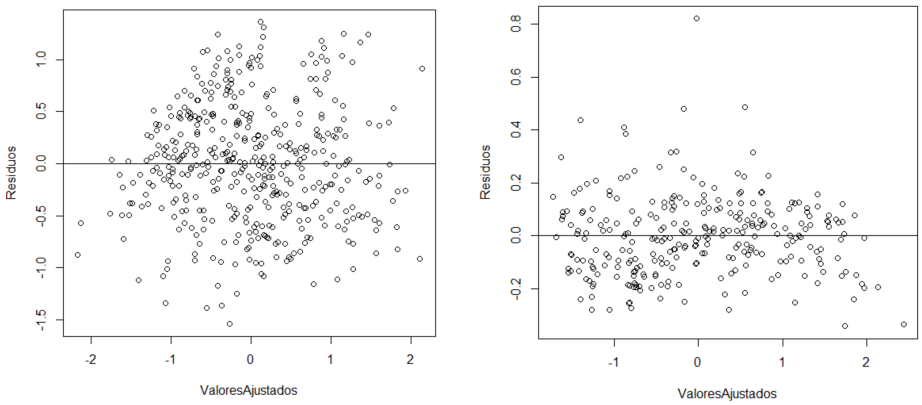


FIGURA 49. Resíduos vs valores estimados para a variável resposta Cost VF (esq.) e variável resposta Total Conv. Value(dir.).

Na Figura 49 está representada o gráfico dos resíduos contra os valores estimados para a variável *Cost VF* e para a variável *Total Conv. Value*, no qual se observa que a variância dos resíduos é constante ao longo de toda a amostra.

Para análise do pressuposto de multicolinearidade faz-se uso dos coeficientes *Tolerance* ou *VIF* (*Variance Inflaction Factor*), o segundo é calculado a partir do inverso do primeiro. A análise de VIF é feita da seguinte forma:

- Até 1 - sem multicolinearidade;
- De 1 até 10 - com multicolinearidade aceitável;
- Acima de 10 - com multicolinearidade problemática.

TABELA 32. VIF (Inflação de variância.)

(A) Variável resposta Cost VF.

Variável	VIF
SCORES.PC1	1
SCORES.PC2	1
SCORES.PC3	1

(B) Variável resposta Total Conv. Value.

Variável	VIF
SCORES.PC1	1
SCORES.PC2	1

Para ambas as variáveis, podemos concluir que as variáveis explicativas têm valor de VIF igual a 1, por isso não apresentam multicolinearidade.

5.2. Análise de Séries Temporais - ARIMA

Os resultados do modelo ARIMA, para a previsão do custo e da receita a serem recebidos pela empresa Overcube, são realizados nos capítulos a seguir e seguem as fases descritas no esquema da Figura 2. Esta metodologia foi escolhida para uma análise ao longo do tempo de acordo com o solicitado pela empresa.

5.2.1. Identificação - Facebook Ads.

Nesta primeira fase, chamada de *Identificação*, tem como objetivo tornar os dados da série estacionários, pois esta característica é preponderante para que se possa modelar o processo ARIMA. Para que isso seja possível, esta fase está dividida nos seguintes passos:

(1) Passo 1 - Preparação dos dados - para atingir a estacionariedade dos dados utiliza-se os seguintes procedimentos: ***(a)***: projeção dos dados da série em gráficos, para verificar a existência de algum padrão; ***(b)***: se necessário, fazer ajustes e/ou transformações matemáticas nos dados da série (como por exemplo, a logaritimização), estabilizando, assim, a variância; ***(c)***: usar a Função de Autocorrelação (ACF) e a Função Parcial de Autocorrelação (PACF), que são as principais ferramentas de identificação e diagnóstico da análise ARIMA, para verificar a existência de algum padrão nos dados da série; ***(d)***: usar a diferenciação dos dados para obter estacionariedade.

(2) Passo 2 - Seleção do modelo - os dados e os vseus respectivos ACF e PACF são examinados para identificar modelos potenciais.

Como o Passo 1 indica, em primeiro lugar, analisou-se a existência de algum padrão nos dados originais.

Geralmente, as séries são não-estacionárias, e os modelos ARMA correspondem a processos estacionários, por isso deve-se operar sobre os dados originais um conjunto de trans formações de modo a que a série resultante possa ser descrita pelos modelos acima referidos. É de realçar que os dados aqui foram transformados em semanas, pois tentou-se representar por mês mas as observações seriam poucas, pois a empresa não tem dados suficientes.
Por isso, essa divisão ficou assim :

- para a série da variável *Spend* - 58 semanas (utilizou-se a mesma base de dados na BD1, mas em semanas);
- para a série da variável *Conv. Value* - 58 semanas (utilizou-se a mesma base de dados na BD2, mas em semanas);

Para uma melhor previsão, os dados (em semanas) sofreram uma transformação, a transformação de Box-Cox (Equação 1) , onde os λ's encontrados têm valor de 0.8 para a BD1 e de 0.3 para BD2.

A estacionariedade pode ser discutida a partir da representação gráfica da série temporal (Figura 50) que deve evidenciar uma média e variância constantes, pressupostos

para a estacionariedade de um série temporal. A representação gráfica da série selecionada, para a variável *Spend* e para a variável *Conversion Value*:

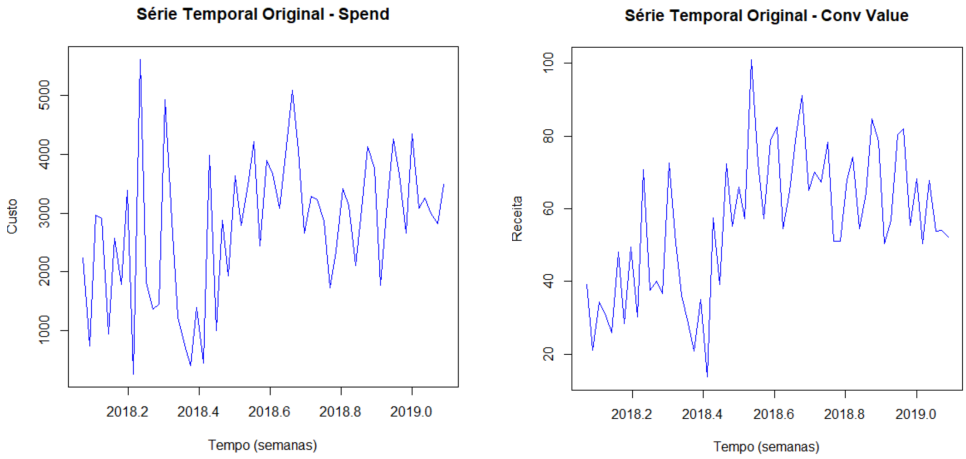


FIGURA 50. Série temporal original para as variáveis Spend e Conversion Value.

Com o uso de transformações, a estacionarização das séries originais, tem como finalidade a estabilização da variância e neutralização da tendência e a respectiva eliminação de movimentos de caráter peródico. Com respeito à estabilização da variância, foi comparado o gráfico da série temporal original como o gráfico da série temporal sujeita a uma transformação lnX_t. Das análises dos gráficos podemos concluir que a transformação não é significativa no sentido de melhorar o comportamento das séries pelo que a opção será de trabalhar com os dados originais.

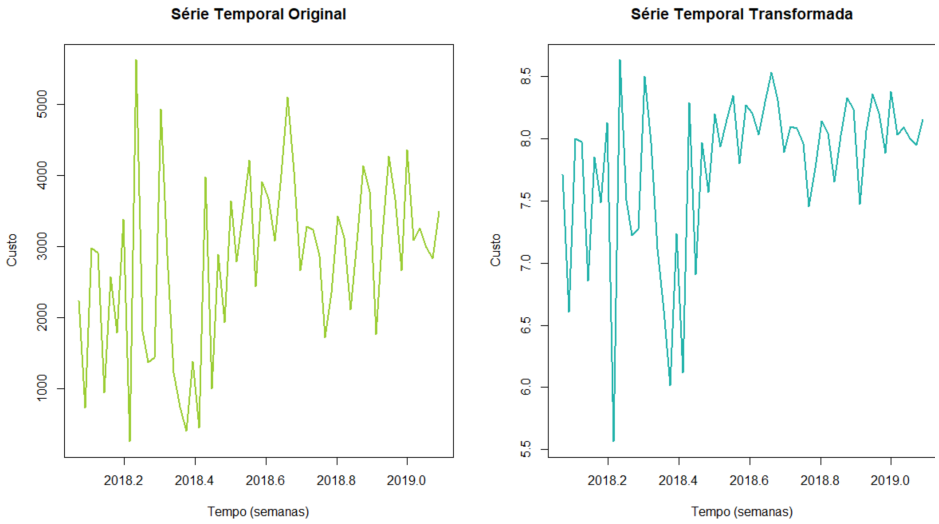


FIGURA 51. Série Temporal original e transformada para a variável Spend.

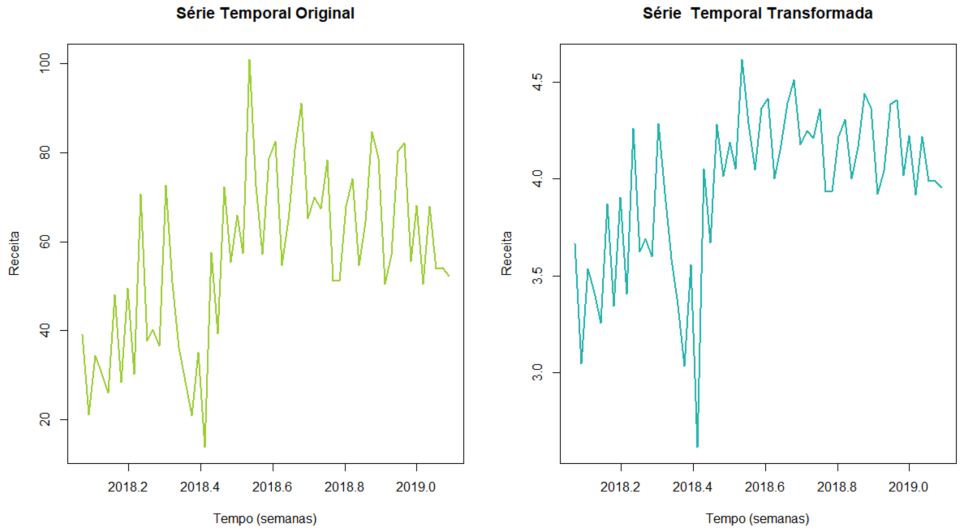


FIGURA 52. Série temporal original e transformada para a variável C. Value.

O estudo da estacionariedade da série pode ser complementado através do comportamento da função de autocorrelação (FAC) e da função de autocorrelação parcial (FACP). A análise dos gráficos e da Tabela 33a, os valores da correlações respetivas:

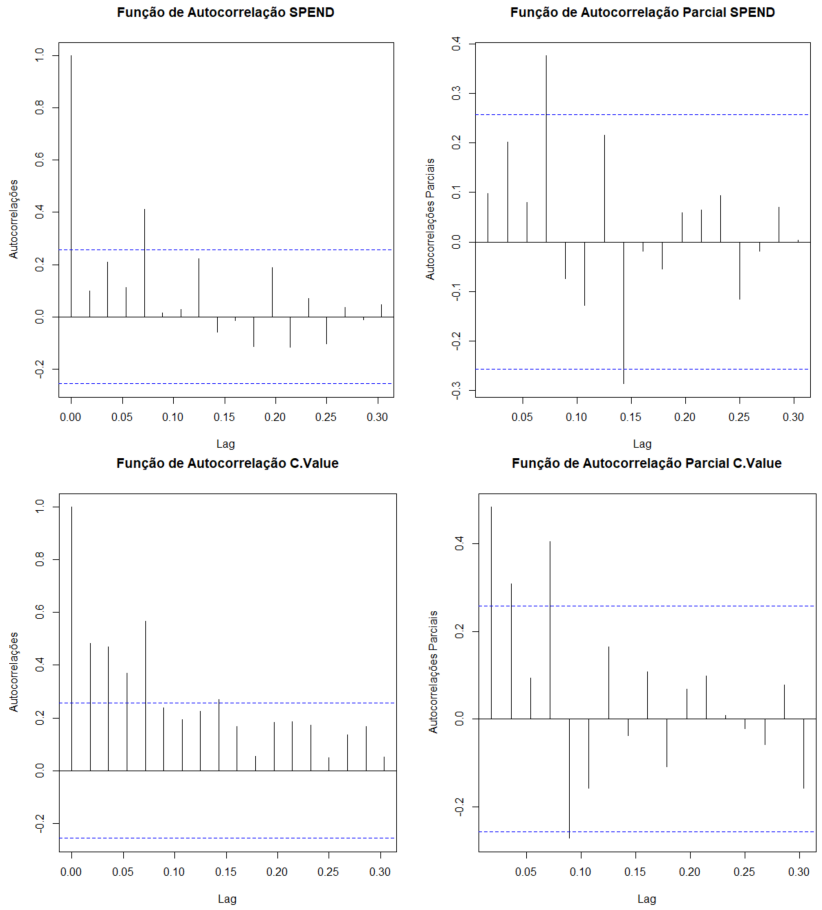


FIGURA 53. FAC e FACP das variáveis Spend (em cima) e Conversion Value (em baixo).

TABELA 33. Valores das correlações obtidas pelas funções de autocorrelação e autocorrelação parcial para as variáveis Spend e Conversion Value.

(A) Série da variável resposta Spend.

	FAC	FACP
$\rho_{0.0179}$	0.098	0.098
$\rho_{0.0357}$	0.210	0.202
$\rho_{0.0536}$	0.111	0.079
$\rho_{0.0714}$	0.411	0.376
$\rho_{0.0893}$	0.015	-0.074
$\rho_{0.1071}$	0.029	-0.128
$\rho_{0.1250}$	0.222	0.215
$\rho_{0.1429}$	-0.058	-0.287
$\rho_{0.1607}$	-0.013	-0.018
$\rho_{0.1786}$	-0.115	-0.054
$\rho_{0.1964}$	0.188	0.059
$\rho_{0.2143}$	-0.116	0.064
$\rho_{0.2321}$	0.069	0.094
$\rho_{0.2500}$	-0.103	-0.116
$\rho_{0.2679}$	0.035	-0.018
$\rho_{0.2857}$	-0.012	0.069
$\rho_{0.3036}$	0.045	0.004

(B) Série da variável resposta Conversion Value.

	FAC	FACP
$\rho_{0.0179}$	0.484	0.484
$\rho_{0.0357}$	0.470	0.308
$\rho_{0.0536}$	0.371	0.093
$\rho_{0.0714}$	0.567	0.406
$\rho_{0.0893}$	0.238	-0.272
$\rho_{0.1071}$	0.193	-0.158
$\rho_{0.1250}$	0.225	0.165
$\rho_{0.1429}$	0.269	-0.037
$\rho_{0.1607}$	0.168	0.108
$\rho_{0.1786}$	0.053	-0.109
$\rho_{0.1964}$	0.183	0.068
$\rho_{0.2143}$	0.185	0.098
$\rho_{0.2321}$	0.172	0.008
$\rho_{0.2500}$	0.048	-0.022
$\rho_{0.2679}$	0.137	-0.058
$\rho_{0.2857}$	0.167	0.078
$\rho_{0.3036}$	0.051	-0.157

A estacionaridade pode ser estudada a partir do comportamento das correlações ao longo do tempo, como tal, uma vez que, para a variável *Spend*, a não-estacionaridade está associada a uma função de autocorrelação com algumas oscilações à medida que o k aumenta tornam-se menores, uma situação que deve ser verificada a nível dos gráficos da FAC e FACP. Para a variável *Conversion Value*, podemos ver que apenas cinco apresentam ter uma correlação significativa, ou seja, a linha azul a tracejado delimita os valores estatisticamente significativos.

Os valores acima da linha tracejada correspondem às correlações significamente diferentes de zero, e os valores abaixo da linha tracejada azul são as correlações muito próximas de zero, e que podem ser assumidas como nulas ou inexistentes.

Retomando a análise dos gráficos das séries temporais, conclui-se que o pressuposto de estacionariedade da média não falha. As séries temporais apresentam um comportamento linear ligeiramente positivo para as duas variáveis, comportamento este que é eviddenciado nos gráficos (Figura 54) obtidos pelo cálculo das matrizes do modelo de regressão linear.

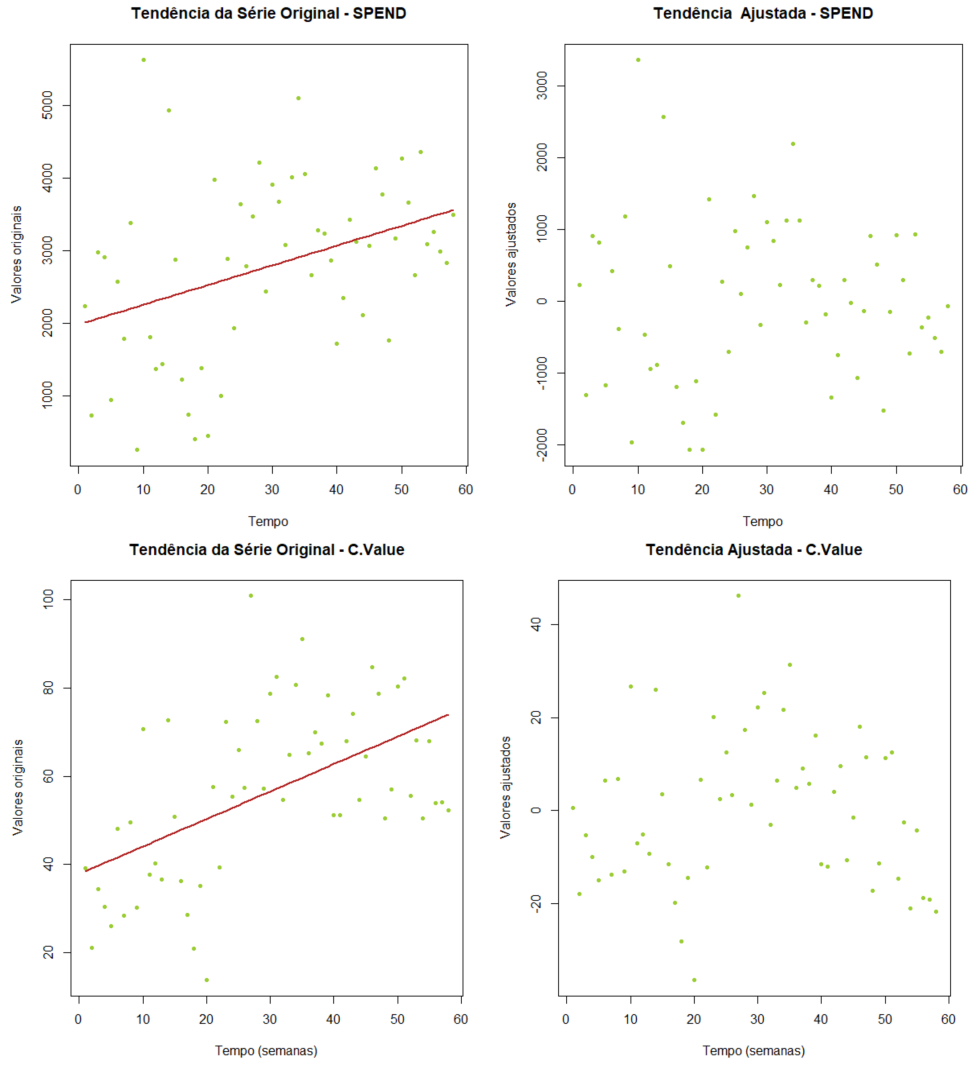


FIGURA 54. Tendência das séries originais das variáveis Spend (em cima) e Conversion Value (em baixo).

Sabe-se que *"a tendência ("trend") de uma série temporal identifica a inclinação, positiva ou negativa, que certas séries apresentam ao longo do tempo. Esta tendência ou inclinação pode ser consequência do facto dos valores observados dependerem de uma componente determinística que é função monótona do tempo linear ou não linear (...)"*.

A série temporal da variável resposta *Spend* não possui tendência, mas a série temporal da *Conversion Value* possui tendência, como se pode observar nos gráficos da Figura 54. Acrescenta-se que os gráficos acima ilustrados, referentes à tendência da série temporal, foram obtidos a partir do cálculo e operações com matrizes do modelo de regressão indicado.

O teste de *Cox-Stuart* foi utilizado para averiguar se a componente de tendência estava presente na série. Este teste é definido como um teste pouco potente (poder igual a 0,78), mas muito robusto para a análise de tendências de séries.

```
        Cox Stuart test

data:  serie_spend
statistic = 20, n = 29, p-value = 0.06143
alternative hypothesis: non randomness
```

```
        Cox Stuart test

data:  serie_conv
statistic = 24, n = 29, p-value = 0.0005461
alternative hypothesis: non randomness
```

FIGURA 55. Teste da tendência do Cox-Stuart - Spend (esq.) e C.Value (dir.).

O valor do p-valor da série da variável *Spend* é maior que 0.05, logo ao nível de 5% de significância temos evidência estatística para continuar a acreditar que esta série não possui tendência. Contudo para a variável *Conversion Value*, a situação é diferente, esta possui tendência a um nível de 5%.

Para a variável *Conversion Value*, vai ser preciso aplicar a diferença entre amostras adjacentes da série temporal, gerando uma nova série, chamada de série diferenciada.

Das conclusões retiradas pelas análises dos gráficos acima ilustrados e pelo comportamento da série temporal, conclui-se que a série da variável *Spend* não apresenta tendência, contudo esta série ao ser diferenciada resulta num melhor modelo. Para a série da *Conversion Value* foi realizada uma diferenciação de segunda ordem, resultando que os resíduos apresentam o comportamento desejado, ou seja, distribuídos em torno de uma média igual a zero. Os gráficos das séries temporais com diferenciação de ordem 1 e 2, para as variáveis *Spend* e *Conversion Value* são apresentados respetivamente a seguir:

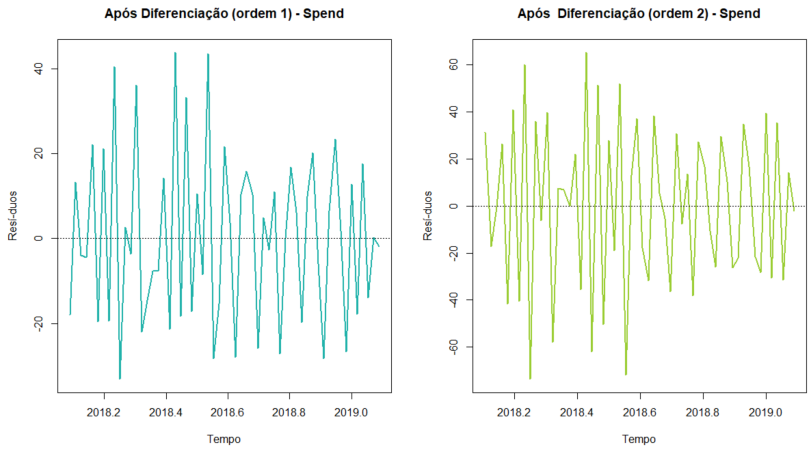


FIGURA 56. Representação da série temporal após uma e duas diferenciações para a variável Spend.

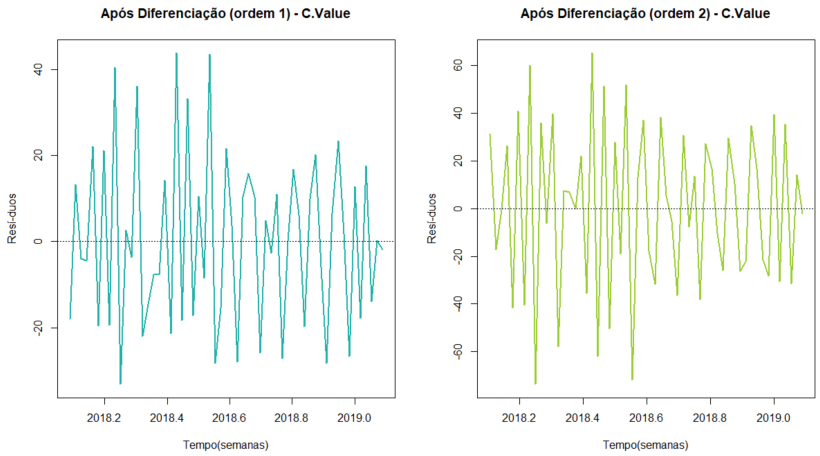


FIGURA 57. Representação da série temporal após uma e duas diferenciações para a variável Conversion Value.

Para a análise da sazonalidade foi escolhido o teste do *Kruskal-Wallis*, cujo o *output* dos resultados foi o seguinte:

TABELA 34. Teste da sazonalidade de Kruskal-Wallis.

Teste Kruskal-Wallis	Estatística de Teste	P-valor
Spend	57	0.4751
Conversion Value	57	0.4751

Ambos os p-valores presentam valores elevados, pelo que a hipótese nula de não sazonalidade não é rejeitada.

Tendo já verificado que a variância dos dados é constante, ou seja, traduz uma inexistência de sazonalidade, complementa-se então a conclusão retirada anteriormente pelo cálculo do período das séries temporais que tomam os valores de 0.038 e de 0.071, respetivamente, para a *Spend* e *Conversion Value*. Como o valor do período é menor que o tamanho dos nossos dados, assume-se, então, que estes dados têm periodicidade, ou seja, existe parte sazonal no modelo a selecionar.

O código devolve o seguinte periodograma para as variáveis:

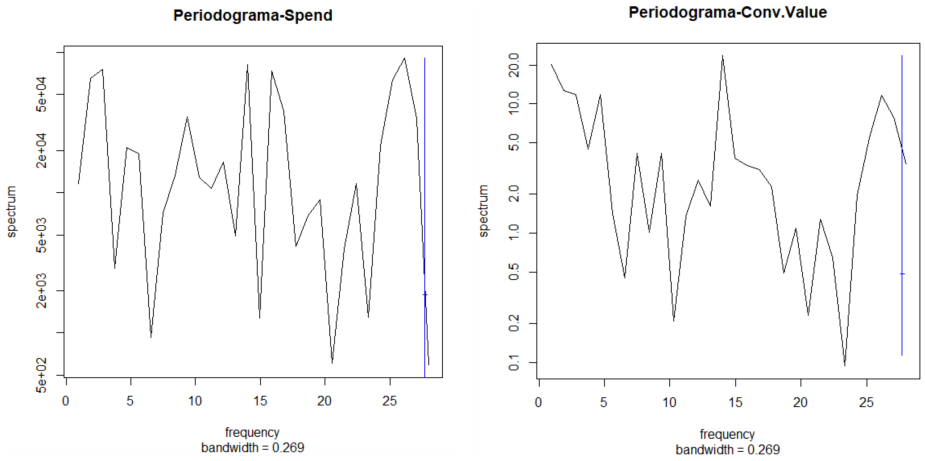


FIGURA 58. Periodograma para as variáveis Spend e Conversion Value.

Interessa saber se a série temporal diferenciada apresenta sinais de sazonalidade (ou periocidade). O periodograma mede as contribuições para a variância total de uma série de componentes periódicos, de uma determinada frequência. Se o periodograma apresenta um "pico" nalguma frequência, isto indica que esta é mais importante na série, que o resto.

Observando os periodogramas das duas séries em estudo, é possível afirmar que existe periocidade das mesmas, isto é, existe um "padrão" nas oscilações.

5.2.2. Estimação e teste - Facebook Ads.

Aqui nesta fase, os coeficientes *(p,d,q)* do modelo ARIMA são determinados e testados quanto á estacionariedade. Para isso, esta fase também é dividida em dois passos, como segue:

(1) Passo 3 - Estimação - todas as estatísticas dos coeficientes são geradas, tais como: ***(a)***: erro padrão para cada coeficiente; ***(b)***: estatística dos dados; ***(c)***: testes de significância e ***(d)***: variância dos resíduos.

(2) Passo 4 - Diagnóstico - utilizando-se os coeficientes e as estatísticas geradas no passo anterior, analisa-se a validade do modelo e, até mesmo, a possibilidade de melhoria deste. Para isso, os seguintes aspetos devem ser considerados: ***(a)***: significância estatística dos coeficientes; ***(b)***: análise da ACF e da PACF, para verificar se há alguma orientação de modelos puramente AR ou MA; ***(c)***: verificar se poderia ter mais de um modelo plausível e determinar qual deles possui menor soma dos erros quadrados (o que será escolhido) e ***(d)***: análise dos resíduos, para se ter a certeza de que não há mais nenhum padrão a ser considerado. Caso o diagnóstico do modelo não seja adequado, deve-se voltar ao Passo 2.

Como os pressupostos da estacionariedade da série temporal foram validados, vão ser testados modelos *ARIMA* de forma a decifrar quais os valores de p e q da parte regular

do modelo. Os critérios a ter em atenção na seleção do modelo adequado são o valor do *Critério de Informação de Akaike* (AIC) e o erro padrão dos coeficientes do modelo. Relativamente ao AIC, seleciona-se o modelo que detenha o menor valor em comparação com os restantes testados, o erro padrão dos coeficientes determina a significância dos coeficientes do modelo, isto é, para os coeficientes obtidos serem significativos, metade do valor em módulo desses coeficientes deve ser maior que o respectivo erro.

Foram testados vários modelos, até encontrar o melhor cujo os coeficientes fossem considerados significativos e com melhor valor de AIC.

TABELA 35. Modelo selecionado para a variável Spend.

ARIMA(p,d,q)	AIC	Coeficientes significativos?
ARIMA(4,1,2)	967.41	SIM

TABELA 36. Modelo selecionado para a variável C.Value.

ARIMA(p,d,q)	AIC	Coeficientes significativos?
ARIMA(1,2,0)	522.35	SIM

De seguida, como nas Tabela 35 e 36 indicam, todas as análises a efetuar incidirão sobre o modelo $ARIMA(4,1,2)$ para a variável *Spend* e sobre o modelo $ARIMA(1,2,0)$ para a variável *Conversion Value*.

A equação dos modelos são os seguintes:

- **Spend**: $(1-\phi B-\phi 2B^2-\phi 3B^3-\phi 4B^4)(1-B)^1X_t=(1-\theta B-\theta 2B^2)\epsilon_t$
- **Conversion Value**: $(1-\phi B)(1-B)^2X_t=0$

Os valores e as conclusões assumidos anteriormente são retirados do output da função arima():

```
> ml_spend

Call:
arima(x = serie_spend, order = c(4, 1, 2))

Coefficients:
          ar1      ar2      ar3      ar4     ma1     ma2
      -2.0539  -1.9481  -1.4382  -0.6488  1.3812  0.4948
s.e.   0.1302   0.2557   0.2187   0.0984  0.1592  0.1743

sigma^2 estimated as 1017057:  log likelihood = -476.71,  aic = 967.41
```

```
> ml_conv

Call:
arima(x = serie_conv, order = c(1, 2, 0))

Coefficients:
          ar1
      -0.6924
s.e.   0.0940

sigma^2 estimated as 605.9:  log likelihood = -259.18,  aic = 522.35
```

FIGURA 59. Output dos modelos ARIMA(4,1,2) (esq.) e ARIMA(1,2,0) (dir.).

Para segunda parte da metodologia de *Box-Jenkins*, com base na formulação de testes de hipóteses, foi estudada a adequação do modelo estimado à série temporal em estudo através dos resíduos, que deverão ter um comportamento análogo ao de um ruído branco. Então, deve-se comprovar se as seguintes hipóteses básicas relativas aos resíduos da série cumprem:

(1) Média zero, $E[\epsilon_t]=0$;
(2) Variância constante (resíduos homocedásticos);

(3) Independência, não apresentam autocorrelações distintas de zero para nenhum $lag \quad k$;
(4) Gaussianidade/Normalidade dos resíduos, $\epsilon_t \sim N(\mu, \sigma^2)$;

A partir do teste $t - Student$, foi realizada a inferência sobre o valor médio dos erros ser nulo:

$$H_0 : E[\epsilon_t] = 0$$

$$vs$$

$$H_1 = E[\epsilon_t] \neq 0$$

Para um nível de significância de 5% aceita-se a média nula dos resíduos para as duas variáveis, *Spend* e *Conversion Value*, isto porque o p-valores obtidos são maiores que o nível de significância. As conclusões retiradas a partir do output do teste $t - Student$ para as duas variáveis indica que o valor médio dos erros para ambas as séries pode ser considerado nulo.

```
> t.test(ml_spend$residuals)

        One Sample t-test

data:  ml_spend$residuals
t = 0.38496, df = 57, p-value = 0.7017
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -213.9142  315.7369
sample estimates:
mean of x
 50.91137
```

```
> t.test(ml_conv$residuals)

        One Sample t-test

data:  ml_conv$residuals
t = 0.10715, df = 57, p-value = 0.915
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.071400  6.757853
sample estimates:
mean of x
0.3432267
```

FIGURA 60. Teste t-Student para a análise dos resíduos dos modelos selecionados para as duas variáveis.

Para a análise da homocedasticidade dos resíduos foram construídos os seguintes gráficos.

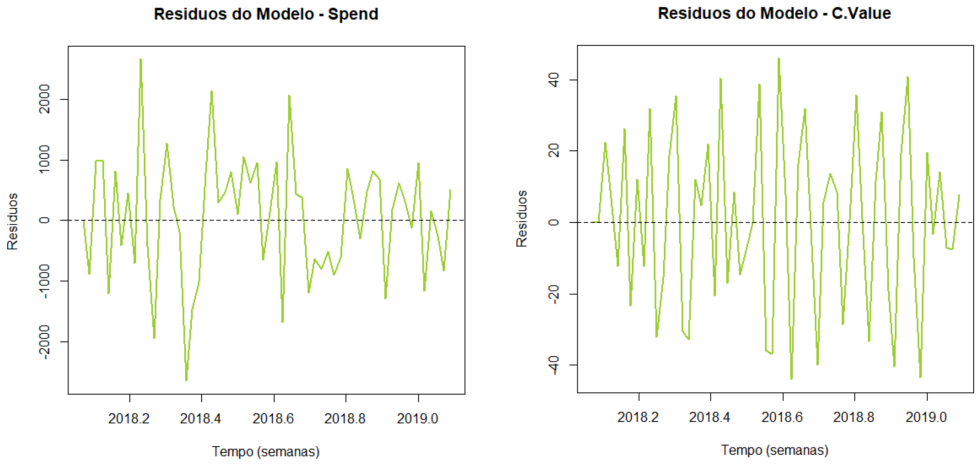


FIGURA 61. Gráfico da homocedasticidade para as variáveis Spend (esq.) e Conversion Value (dir.).

Os resíduos distribuem-se em torno da reta horizontal $y = 0$, validando assim o pressuposto da variância constante/resíduos homocedásticos.

A independência dos resíduos é estudada a partir da função de autocorrelação (ACF) e da função de autocorrelação parcial ($PACF$) dos resíduos. Os gráficos correspondentes ao estudo das correlações (ACF) e ($PACF$) são apresentados a seguir.

As correlações são representadas no gráfico e na tabela a seguir apresentados:

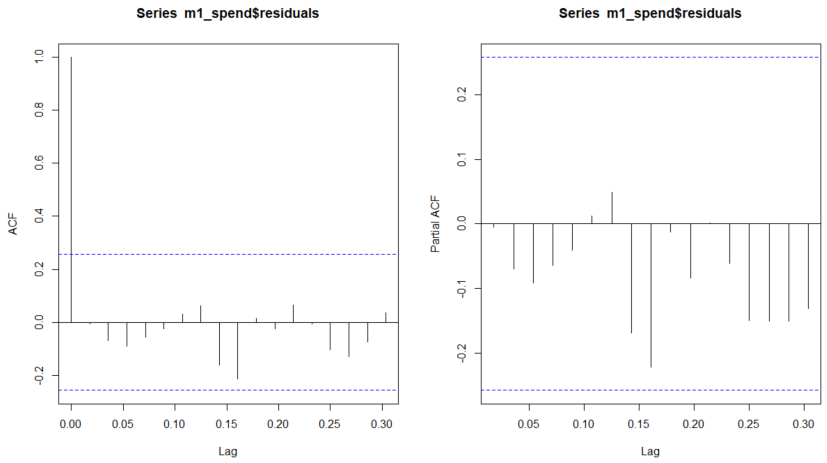


FIGURA 62. Funções de autocorrelações e autocorrelações parciais para a variável Spend.

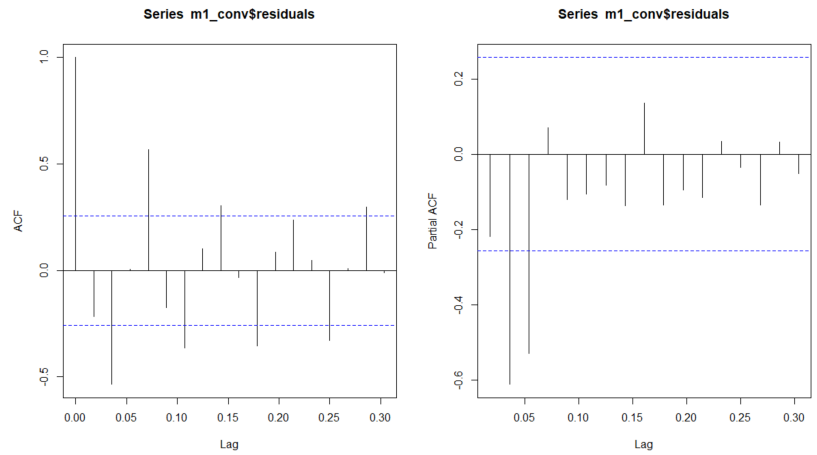


FIGURA 63. Funções de autocorrelações e autocorrelações parciais para a variável C.Value.

TABELA 37. Valores das correlações obtidas pelas funções de autocorrelação e autocorrelação parcial.

(A) Série da variável resposta Spend.

	FAC	FACP
$\rho_{0.0179}$	-0.005	-0.005
$\rho_{0.0357}$	-0.070	-0.070
$\rho_{0.0536}$	-0.090	-0.091
$\rho_{0.0714}$	-0.057	-0.064
$\rho_{0.0893}$	-0.026	-0.041
$\rho_{0.1071}$	0.030	0.012
$\rho_{0.1250}$	0.063	0.049
$\rho_{0.1429}$	-0.160	-0.168
$\rho_{0.1607}$	-0.213	-0.221
$\rho_{0.1786}$	0.015	-0.012
$\rho_{0.1964}$	-0.024	-0.084
$\rho_{0.2143}$	0.064	0.001
$\rho_{0.2321}$	-0.006	-0.061
$\rho_{0.2500}$	-0.105	-0.149
$\rho_{0.2679}$	-0.129	-0.150
$\rho_{0.2857}$	-0.074	-0.150
$\rho_{0.3036}$	0.035	-0.131

(B) Série da variável resposta Conversion Value.

	FAC	FACP
$\rho_{0.0179}$	-0.218	-0.218
$\rho_{0.0357}$	-0.536	-0.612
$\rho_{0.0536}$	0.006	-0.529
$\rho_{0.0714}$	0.568	0.070
$\rho_{0.0893}$	-0.175	-0.120
$\rho_{0.1071}$	-0.363	-0.106
$\rho_{0.1250}$	0.103	-0.082
$\rho_{0.1429}$	0.305	-0.136
$\rho_{0.1607}$	-0.033	0.137
$\rho_{0.1786}$	-0.354	-0.134
$\rho_{0.1964}$	0.087	-0.095
$\rho_{0.2143}$	0.237	-0.115
$\rho_{0.2321}$	0.048	0.034
$\rho_{0.2500}$	-0.330	-0.034
$\rho_{0.2679}$	0.009	-0.135
$\rho_{0.2857}$	0.298	0.032
$\rho_{0.3036}$	-0.010	-0.052

As correlações dos resíduos obtidas pelas funções *ACF* e *PACF* e a análise dos gráficos das respetivas funções de autocorrelações mostram evidências de correlações significativas, sendo que se assume a proximidade das mesmas a zero, validando então o pressuposto da independência dos resíduos. A variável *Spend*, no gráfico ACF, tem um grande pico no *lag 1* que diminui depois a partir do segundo, para a *Conversion Value*, o gráfico ACF, tem um grande pico na primeira defasagem seguido por uma onda decrescente que alterna entre correlações positivas e negativas.

Foram elaborados um gráfico QQ, um histograma com a curva da distribuição Normal sobreposta a seguir ilustrados, para o *Spend* e para a *Conversion Value*:

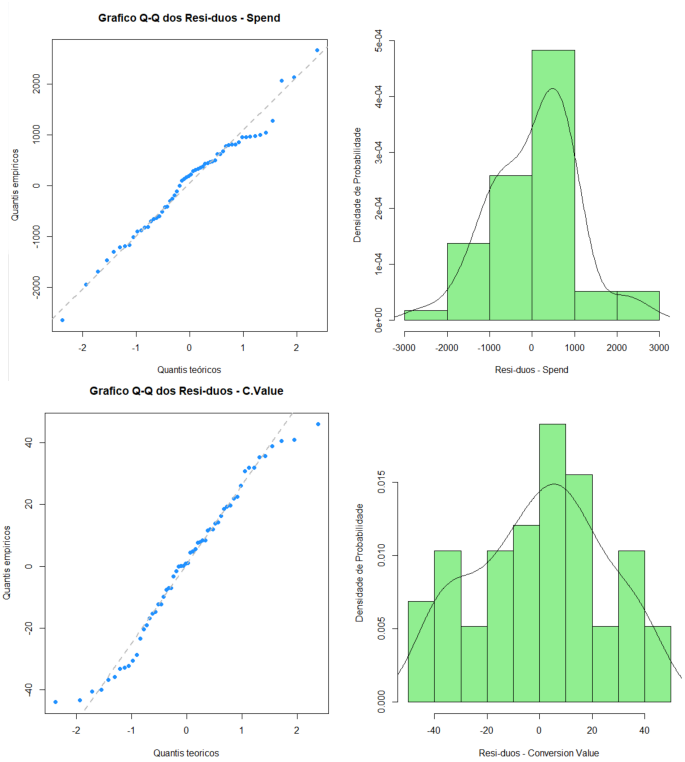


FIGURA 64. Gráfico QQ e histograma dos resíduos.

No caso dos gráficos Q-Q verifica-se que grande parte dos dados estão distribuídos em torno da reta, o que valida a normalidade dos resíduos da variável *Spend* e *Conversion Value*. No caso dos histogramas, verifica-se que os resíduos têm um comportamento gaussiano visto que assumem um comportamento semelhante ao da curva sobreposta que é respeitante à distribuição Normal dos resíduos.

Para concluir a análise dos resíduos recorreu-se ao teste de*Shapiro-Wilk*.

$$H_0 : \epsilon_t \quad seguem \quad uma \quad distribuição \quad normal \quad N(\mu, \sigma^2)$$

$$vs$$

$$H_1 = \epsilon_t \quad não \quad seguem \quad uma \quad distribuição \quad normal \quad N(\mu, \sigma^2)$$

Os testes referidos foram efetuados novamente na ferramenta *R*, e as conclusões são retiradas a partir dos seguintes outputs (Figura 65):

```
        Shapiro-Wilk normality test

data:  ml_spend$residuals
W = 0.98151, p-value = 0.5189
```

```
        Shapiro-Wilk normality test

data:  ml_conv$residuals
W = 0.97102, p-value = 0.179
```

FIGURA 65. Teste do Shapiro Wilk para as variáveis Spend (esq.) e C. Value (dir.).

5.2.3. Identificação - Google Ads.

Como foi feito com a plataforma Facebook Ads, será modelado o processo ARIMA igualmente com os dados da Google Ads. Os mesmo passos falados no capítulo 5.2.1. (Passo 1 e 2) serão repetidos para os dados do Google Ads.

Analisou-se a existência de algum padrão nos dados originais, assim como refere o Passo 1. Novamente estes dados foram transformados em semanas, pois obtivemos o mesmo problema de apenas considerar por mês, o tamanho da série seria demasiado curta. Obteve-se a seguinte divisão:

- para a série da variável *Cost VF* - 53 semanas (utilizou-se a mesma base de dados na BD3, mas em semanas);
- para a série da variável *Total Conv. Value* - 47 semanas (utilizou-se a mesma base de dados na BD4, mas em semanas);

Para uma melhor previsão, estes dados (em semanas) também sofreram uma transformação, a transformação de Box-Cox (Equação 1) , onde os λs encontrados têm valor de -0.05 para a BD3 e de 0.25 para BD4.

As séries são não-estacionárias, e os modelos ARMA correspondem a processos estacionários, por isso devem operar-se sobre os dados originais um conjunto de transformações de modo a que a série resultante possa ser descrita pelos modelos acima referidos. A estacionariedade ou a falta desta pode ser observada a partir da representação gráfica da própria série temporal, seja a variável *Cost VF* ou para a *Total Conv. Value*, como apresentada na Figura 66

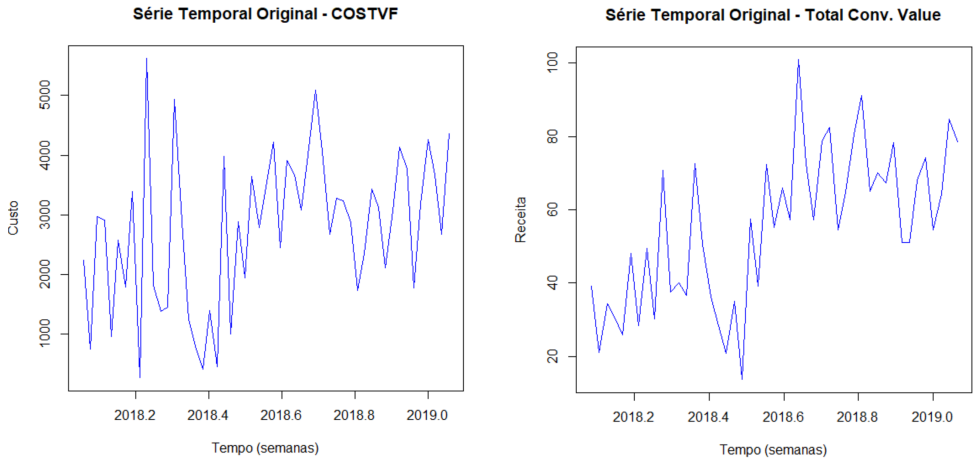


FIGURA 66. Série temporal original para as variáveis Cost VF e Total Conv. Value.

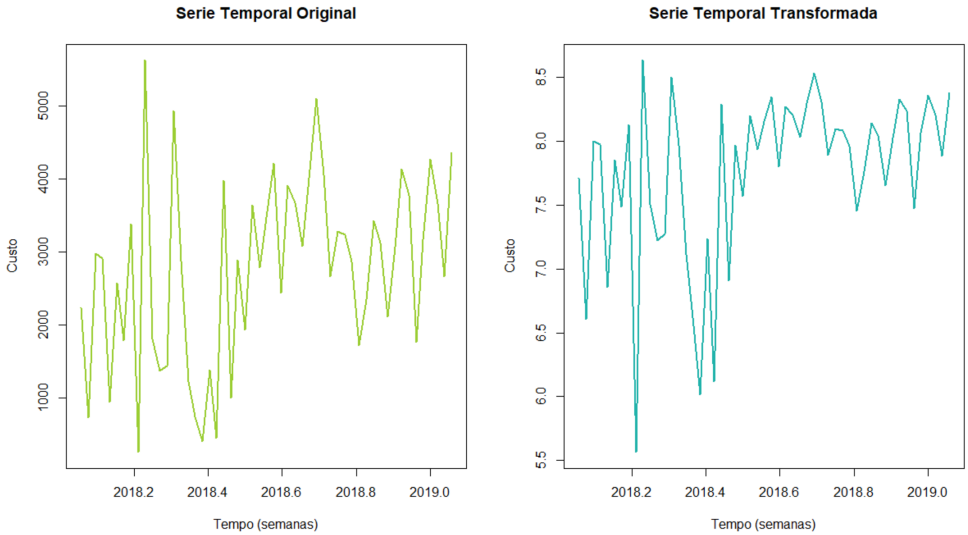


FIGURA 67. Série temporal original e transformada para a variável Cost VF.

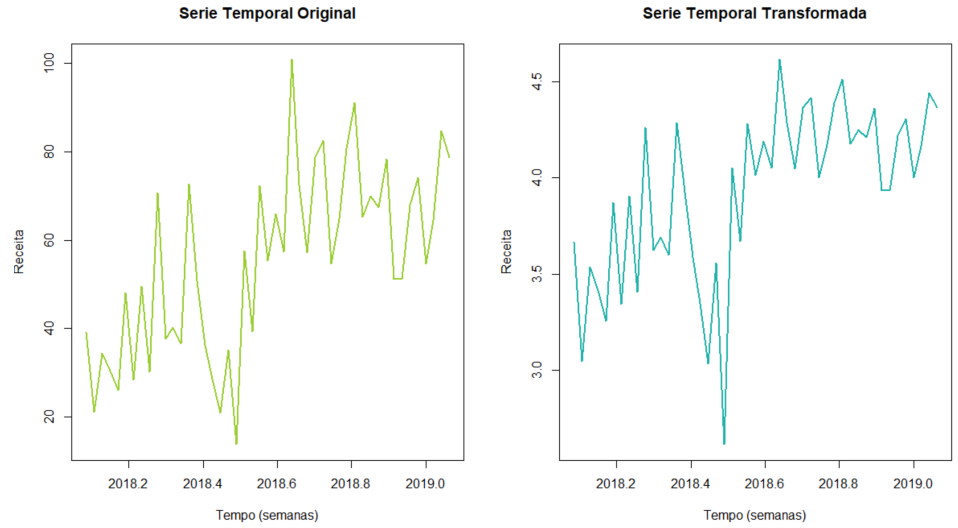


FIGURA 68. Série temporal original e transformada para a variável Total Conv. Value.

Repara-se que não é necessário estabilizar a variância, nem neutralizar a tendência, ou seja, não é necessário recorrer a qualquer transformação dos dados. Utilizou-se a transformação $ln(X_t)$, e esta ainda apresentar algumas diferenças, mas não foram significativas para continuar com a série logaritmizada. Ambas séries (*Cost VF* e *Total Conv. Value*) prosseguem com a série original, como se pode ver pelos seguintes gráficos:

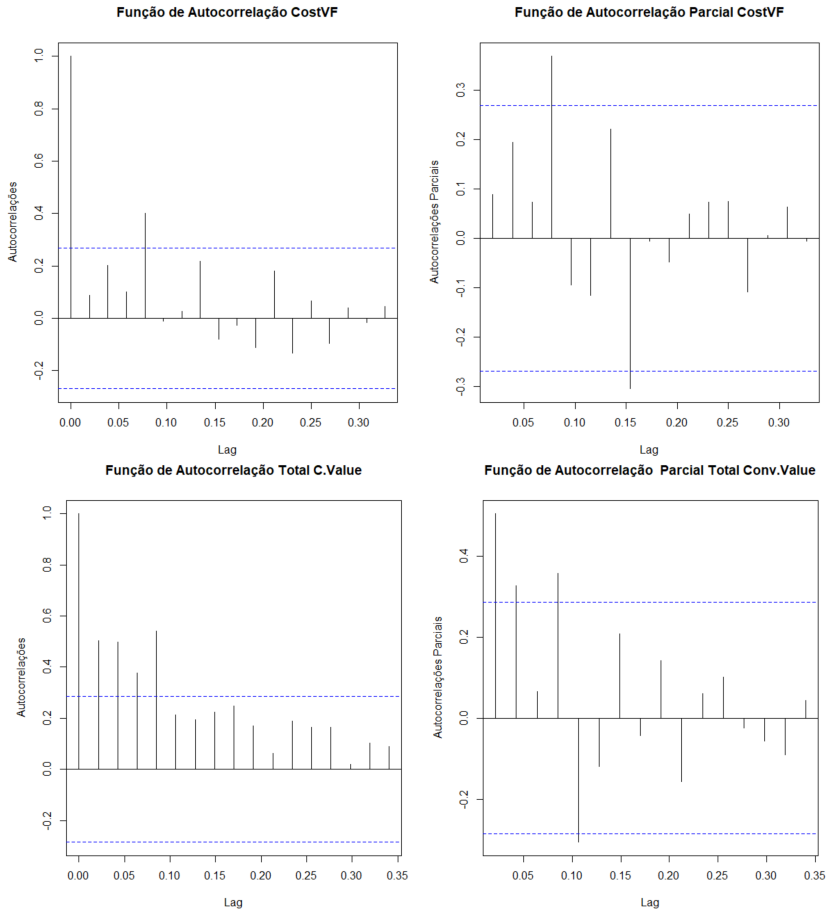


FIGURA 69. FAC e FACP para as variáveis CostVF (em cima) e Total Conv. Value (em baixo).

Em seguida será observado o comportamento da função de autocorrelação (FAC) e/ou da função de autocorrelação parcial (FACP), nas seguintes tabelas e gráficos, respetivamente.

TABELA 38. Valores das correlações obtidas pelas funções de autocorrelação e autocorrelação parcial.

(A) Série da variável resposta Cost VF.

	FAC	FACP
$\rho_{0.0192}$	0.089	0.089
$\rho_{0.0385}$	0.201	0.194
$\rho_{0.0577}$	0.102	0.073
$\rho_{0.0769}$	0.401	0.369
$\rho_{0.0962}$	-0.010	-0.095
$\rho_{0.1154}$	0.027	-0.116
$\rho_{0.1346}$	0.218	0.222
$\rho_{0.1538}$	-0.080	-0.305
$\rho_{0.1731}$	-0.027	-0.006
$\rho_{0.1923}$	-0.112	-0.049
$\rho_{0.2115}$	0.182	0.049
$\rho_{0.2308}$	-0.132	0.074
$\rho_{0.2500}$	0.066	0.074
$\rho_{0.2692}$	-0.094	-0.108
$\rho_{0.2885}$	0.039	0.005
$\rho_{0.3077}$	-0.017	0.063
$\rho_{0.3269}$	0.046	-0.006

(B) Série da variável resposta Total Conv. Value.

	FAC	FACP
$\rho_{0.0213}$	0.505	0.505
$\rho_{0.0426}$	0.498	0.327
$\rho_{0.0638}$	0.378	0.066
$\rho_{0.0851}$	0.542	0.357
$\rho_{0.1064}$	0.213	-0.306
$\rho_{0.1277}$	0.194	-0.120
$\rho_{0.1489}$	0.225	0.209
$\rho_{0.1702}$	0.249	-0.042
$\rho_{0.1915}$	0.171	0.142
$\rho_{0.2128}$	0.062	-0.156
$\rho_{0.2340}$	0.190	0.061
$\rho_{0.2553}$	0.164	0.102
$\rho_{0.2766}$	0.164	-0.024
$\rho_{0.2979}$	0.018	-0.056
$\rho_{0.3191}$	0.102	-0.090
$\rho_{0.3404}$	0.088	0.043

Foi analisado o comportamento das correlações. Para a variável *Cost VF* encontrou-se uma correlação significativa, e para a variável *Total Conv. Value* encontraram-se bastantes, pelos menos até ao oitavo lag e depois deixam de o ser.

Valores acima do tracejado correspondem às correlações signifcamente diferentes de zero, e os valores abaixo do tracejado são as correlações muito próximas de zero, e que podem ser assumidas como nulas.

De modo a prosseguir a análise das séries temporais, conclui-se que o pressuposto de estacionariedade da média não falha.

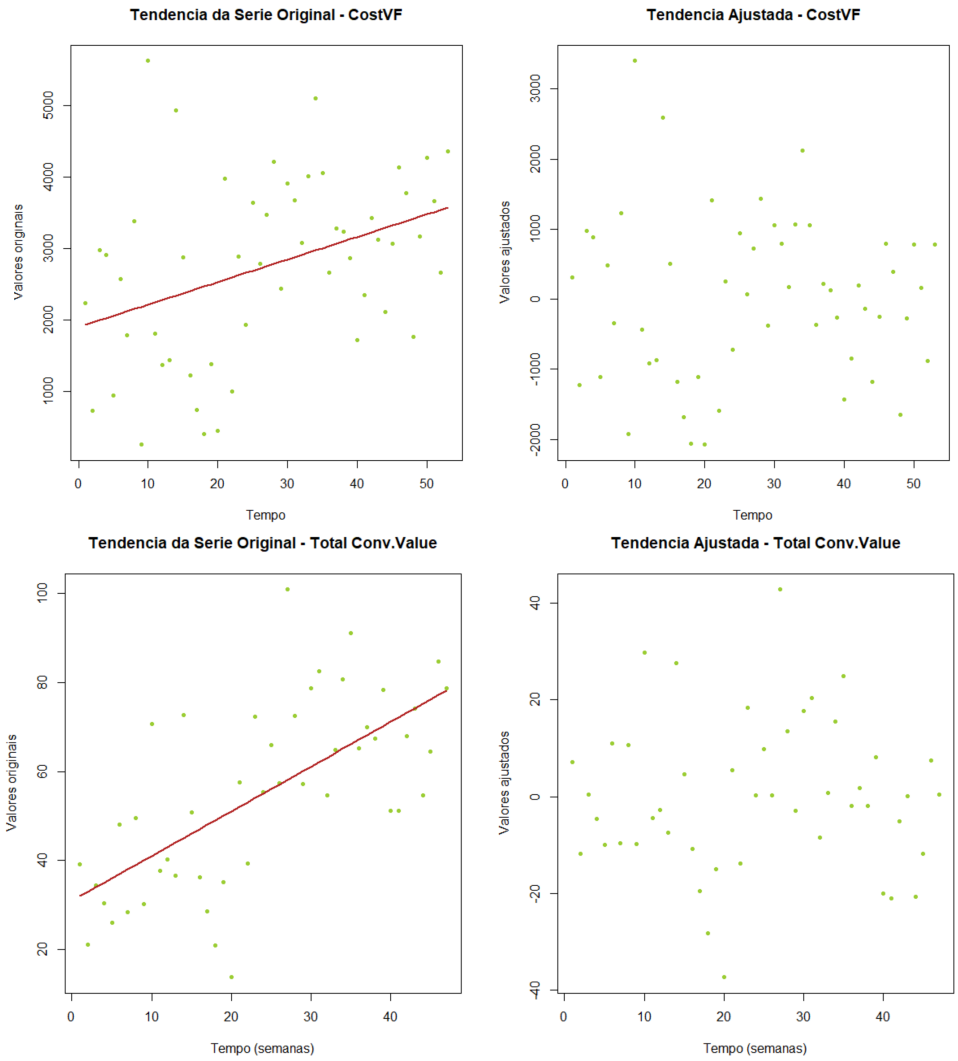


FIGURA 70. Tendência das séries originais.

Ambas as séries temporais apresentam tendência. Sabe-se que a série temporal tem um comportamento linear ligeiramente positivo para as duas variáveis, comportamento este que é evidente a partir dos gráficos obtidos pelo cálculo das matrizes do modelo de regressão linear, como mostra os gráficos da Figura 70.

Utilizou-se o teste de *Cox-Stuart* para complementar este estudo e analisar se realmente existe a componente tendência.

```
        Cox Stuart test

data:  serie_spend
statistic = 22, n = 26, p-value = 0.0005335
alternative hypothesis: non randomness
```

```
        Cox Stuart test

data:  serie_conv
statistic = 22, n = 23, p-value = 5.722e-06
alternative hypothesis: non randomness
```

FIGURA 71. Teste da tendência do Cox-Stuart - Cost VF (esq.) e Total Conv.Value (dir.).

Os valores do p-valor, de ambas as séries, são menores que 0.05, logo ao nível de 5% de significância temos evidência estatística para continuar a acreditar que estas possuem tendência.

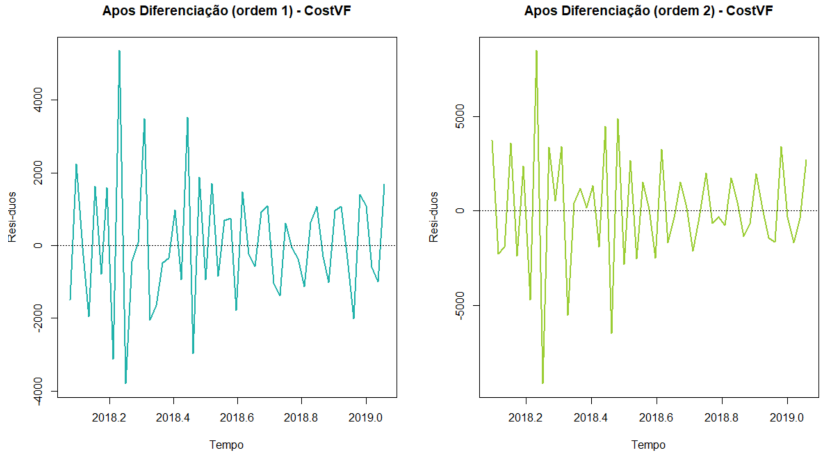


FIGURA 72. Representação da série temporal após uma e duas diferenciações para a variável Cost VF.

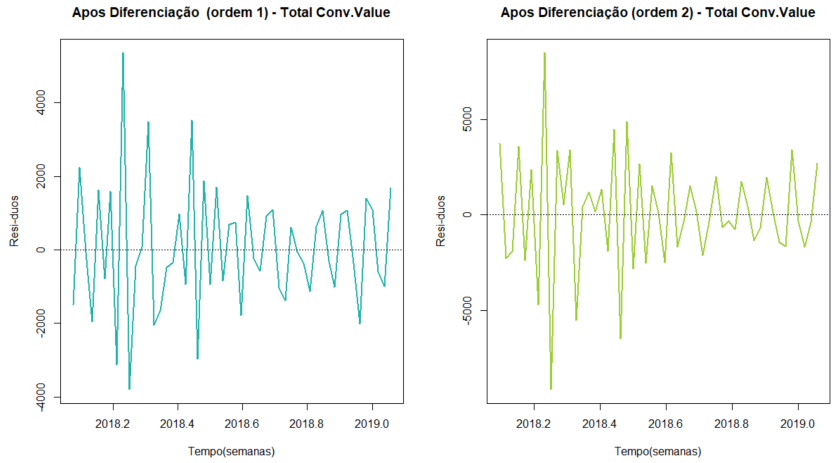


FIGURA 73. Representação da série temporal após uma e duas diferenciações para a variável Total Conv. Value.

Diferenciou-se as duas séries, pois resultou em melhores modelos. Fez-se a diferenciaçao de ordem 1 e de ordem 2 para as duas séries, mas apenas se vai utilizar a ordem 1 respetivamente, pois é aquela que melhor se enquadra nas duas séries.

Através do teste do *Kruskal-Wallis*, testou-se se as séries têm indícios de possuir componente sazonal, e o output dos resultados foi o seguinte:

TABELA 39. Teste da sazonalidade de Kruskal-Wallis.

Teste Kruskal-Wallis	Estatística de Teste	P-valor
CostVF	57	0.4751
Total Conv. Value	57	0.4751

Os resultados informam que ambos os p-valores são maiores que 0.05, logo ao nível de 5% de significância, ou seja, não rejeito a hipótese nula de que não existe sazonalidade determinística.
Prosseguiu-se com o estudo dos modelos de série temporal, calculou-se o período das séries temporais. Estes tomam o valor de 0.074 para *Cost VF*, e de 1.021 para *Total Conv. Value*. Como o valor do período é menor que o tamanho dos nossos dados assume-se então que estes dados têm periodicidade, ou seja, existe parte sazonal no modelo a selecionar. O código devolve o seguinte periodograma para as variáveis:

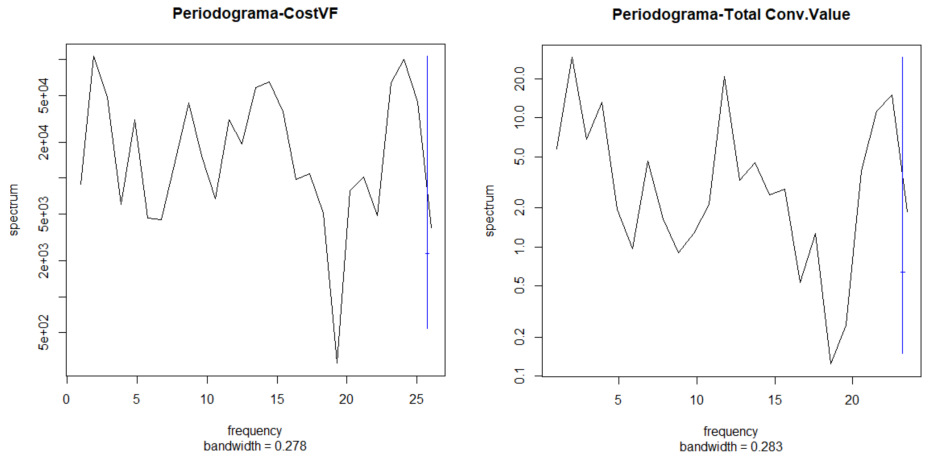


FIGURA 74. Periodograma das variáveis Cost VF e Total Conv. Value.

Observando o periodograma da série em estudo, é possível afirmar que existe periocidade da série temporal, isto é, existe um "padrão" nas oscilações. Logo, a parte sazonal do modelo é não nula.

5.2.4. Estimação e teste - Google Ads.

Neste capítulo irão ser retratados os Passos 3 e 4 (como descrito no capítulo 5.2.2.), onde irão ser determinados e testados os coeficientes *(p,d,q)* do modelo ARIMA que melhor se ajusta. Para selecionar os coeficientes, estes irão ser testados de acordo com os valores de AIC e o respectivo erro padrão dos mesmos e estes sejam considerados significativos.

TABELA 40. Modelo selecionado para a variável Cost VF.

ARIMA(p,d,q)	AIC	Coeficientes significativos?
ARIMA(2,1,0)	902.96	SIM

TABELA 41. Modelo selecionado para a variável Total Conv.Value.

ARIMA(p,d,q)	AIC	Coeficientes significativos?
ARIMA(3,1,0)	389.29	SIM

De seguida, como nas Tabela 40 e 41 indicam, todas as análises a efetuar incidirão sobre o modelo $ARIMA(2,1,0)$ para a variável *CostVF* e sobre o modelo $ARIMA(3,1,0)$ para a variável *Total Conv. Value*.

A equação dos modelos são os seguintes:

- **Cost VF**: $(1-\phi B-\phi 2B^2)(1-B)^1X_t=0$
- **Total Conv. Value**: $(1-\phi B-\phi 2B^2-\phi 3B^3)(1-B)^2X_t=0$

Os valores e conclusões assumidos anteriormente são retiradas do output da função arima():

```
> ml_costv

Call:
arima(x = serie_spend, order = c(2, 1, 0))

Coefficients:
          ar1      ar2
      -0.7254  -0.2843
s.e.   0.1323   0.1334

sigma^2 estimated as 1795499:  log likelihood = -448.48,  aic = 902.96
```

```
> ml_tconv

Call:
arima(x = serie_conv, order = c(3, 1, 0))

Coefficients:
          ar1      ar2      ar3
      -0.7178  -0.4678  -0.4292
s.e.   0.1320   0.1568   0.1316

sigma^2 estimated as 227.9:  log likelihood = -190.65,  aic = 389.29
```

FIGURA 75. Output dos modelos ARIMA(2,1,0)(esq) e ARIMA(3,1,0) (dir).

Na segunda parte da metodologia de Box-Jenkins, irá ser analisado o comportamento dos resíduos do modelo, e ve se os mesmos se comportam como um ruído branco. A partir de um teste $t-Student$, querem-se validar as seguintes hipóteses:

$$H_0 : E[\epsilon_t] = 0$$

$$vs$$

$$H_1 = E[\epsilon_t] \neq 0$$

Para um nível de significância de 5% aceita-se a média nula dos resíduos para as duas variáveis, *CostVF* e *Totsl Conv. Value*, isto porque o p-valores obtidos são maiores que o nível de significância. As conclusões são retiradas a partir do seguinte output:

```
> t.test(ml_spend$residuals)

        One Sample t-test

data:  ml_spend$residuals
t = 0.35805, df = 52, p-value = 0.7218
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1167327  0.1674378
sample estimates:
 mean of x
0.02535252
```

```
> t.test(ml_conv$residuals)

        One Sample t-test

data:  ml_conv$residuals
t = 0.82255, df = 46, p-value = 0.415
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.9702804  2.3112355
sample estimates:
mean of x
0.6704775
```

FIGURA 76. Teste t-Student para a análise dos resíduos dos modelos selecionados para as duas variáveis respostas.

A homocedasticidade dos resíduos é estudada a partir do seguinte gráfico:

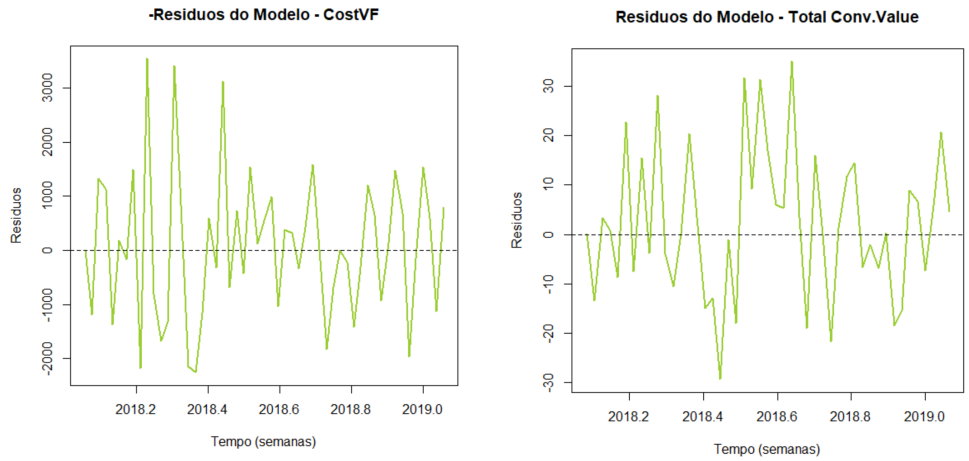


FIGURA 77. Gráfico da homocedasticidade para as variáveis respostas Cost VF (esq.) e Total Conv. Value (dir.).

Verificou-se então que os resíduos distribuem-se em torno da reta $y = 0$, validando assim o pressuposto da variância constante/resíduos homocedásticos.

A independência dos resíduos é estudada a partir da função de autocorrelação (ACF) e pela função de autocorrelação parcial ($PACF$) dos resíduos. As correlações são representadas no gráfico e tabela a seguir apresentados:

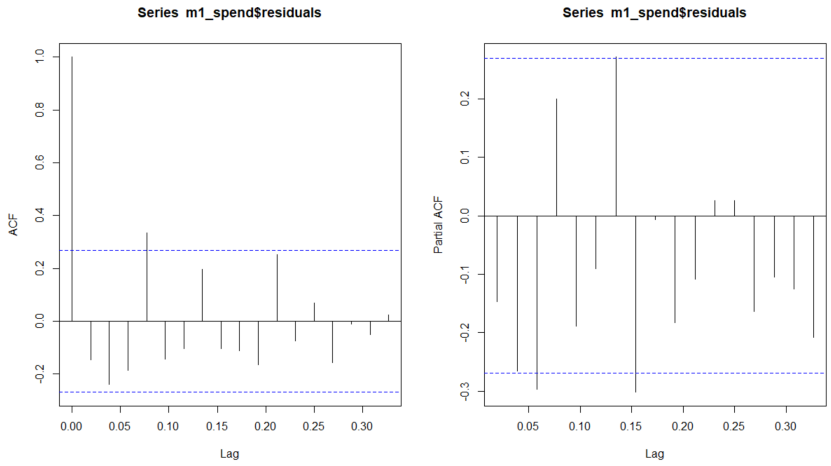


FIGURA 78. Funções de autocorrelações e autocorrelações parciais para a variável Cost VF.

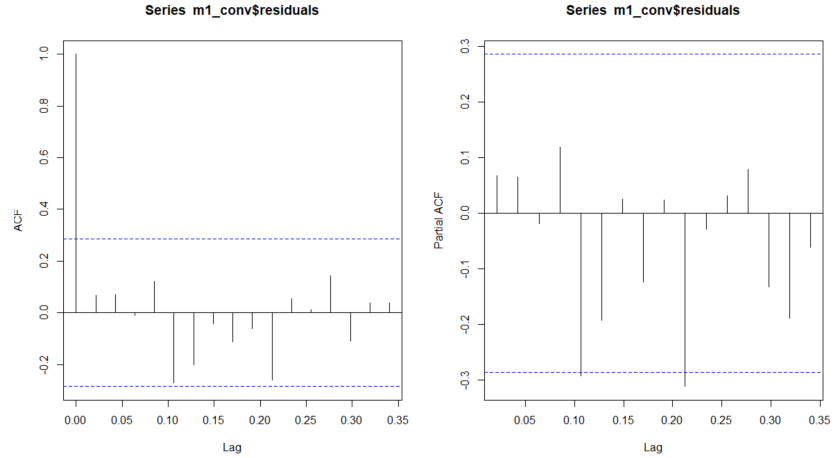


FIGURA 79. Funções de autocorrelações e autocorrelações parciais para a variável Total Conv.Value.

TABELA 42. Valores das correlações obtidas pelas funções de autocorrelação e autocorrelação parcial para as variáveis CostVF e Total Conversion Value.

(A) Série da variável resposta Cost VF.

	FAC	FACP
$\rho_{0.0192}$	-0.146	-0.146
$\rho_{0.0385}$	-0.239	-0.266
$\rho_{0.0577}$	-0.186	-0.296
$\rho_{0.0769}$	0.335	0.200
$\rho_{0.0962}$	-0.142	-0.189
$\rho_{0.1154}$	-0.104	-0.090
$\rho_{0.1346}$	0.197	0.271
$\rho_{0.1538}$	-0.103	-0.302
$\rho_{0.1731}$	-0.111	-0.007
$\rho_{0.1923}$	-0.163	-0.183
$\rho_{0.2115}$	0.251	-0.108
$\rho_{0.2308}$	-0.074	0.026
$\rho_{0.2500}$	0.069	0.026
$\rho_{0.2692}$	-0.155	-0.163
$\rho_{0.2885}$	-0.010	-0.105
$\rho_{0.3077}$	-0.050	-0.125
$\rho_{0.3269}$	0.024	-0.208

(B) Série da variável resposta Total Conv. Value.

	FAC	FACP
$\rho_{0.0213}$	0.067	0.067
$\rho_{0.0426}$	0.069	0.065
$\rho_{0.0638}$	-0.009	-0.018
$\rho_{0.0851}$	0.121	0.119
$\rho_{0.1064}$	-0.272	-0.292
$\rho_{0.1277}$	-0.203	-0.192
$\rho_{0.1489}$	-0.043	0.025
$\rho_{0.1702}$	-0.112	-0.123
$\rho_{0.1915}$	-0.062	0.024
$\rho_{0.2128}$	-0.260	-0.311
$\rho_{0.2340}$	0.053	-0.029
$\rho_{0.2553}$	0.012	0.031
$\rho_{0.2766}$	0.144	0.079
$\rho_{0.2979}$	-0.109	-0.131
$\rho_{0.3191}$	0.038	-0.188
$\rho_{0.3404}$	0.038	-0.061

As correlações dos resíduos obtidas pelas funções ACF e PACF e a análise dos gráficos das respectivas funções de autocorrelações mostram evidências de correlações significativas. Para a variável *Cost VF*, no gráfico ACF, tem um grande pico no lag 1 que diminui depois a partir do segundo, mas volta a crescer no lag 5, onde chega a ser significativo. Para a *Total Conv. Value*, o gráfico ACF tem um grande pico no lag 1 seguido por uma onda decrescente que alterna entre correlações positivas e negativas.

Foram elaborados um gráfico QQ e um histograma com a curva da distribuição Normal para *CostVF* e *Total Conv. Value*.

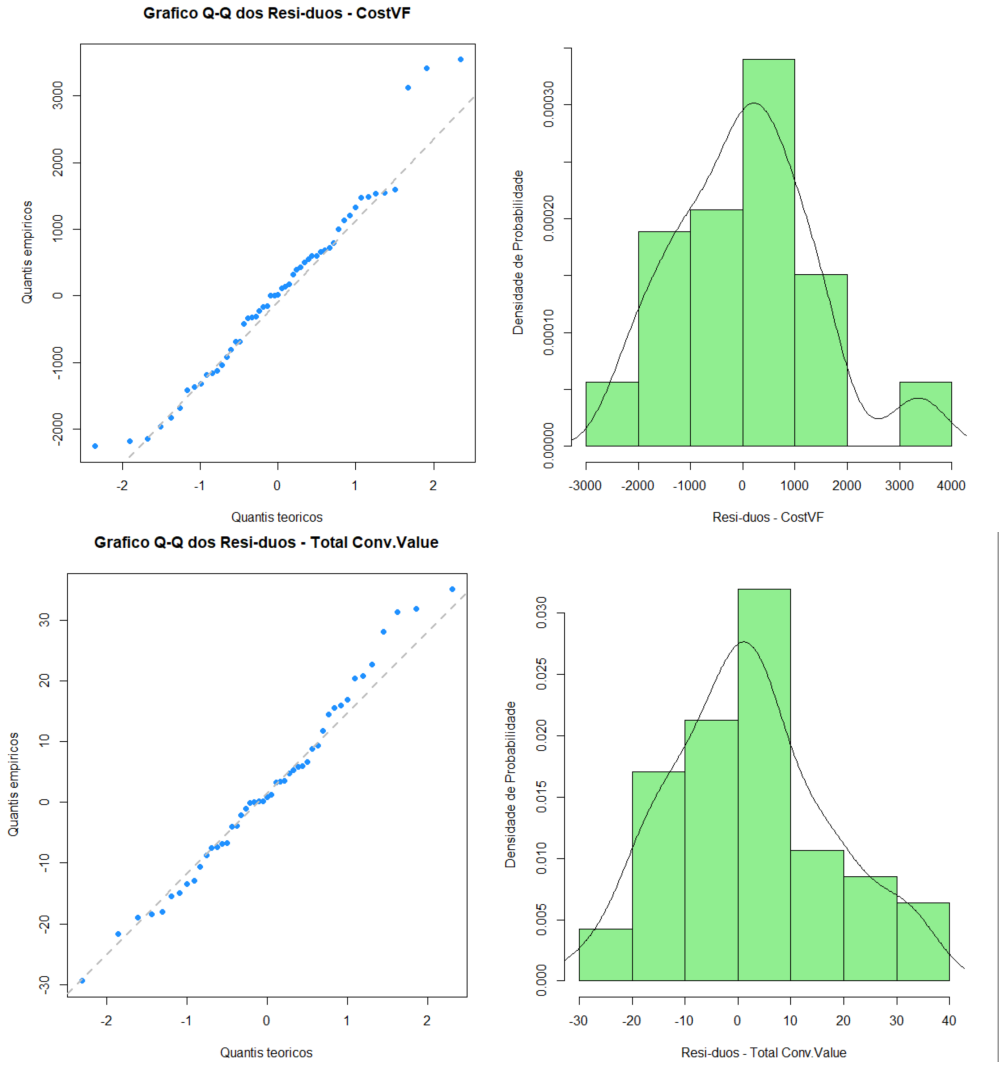


FIGURA 80. Gráfico QQ e histograma dos resíduos.

No caso dos gráficos Q-Q verifica-se que grande parte dos dados estão distribuídos em torno da reta, o que valida a normalidade dos resíduos da variável *Cost VF* e *Total Conv. Value*. Os histogramas assumem um comportamento semelhante ao da curva sobreposta que é respeitante à distribuição Normal dos resíduos.
Para concluir a análise dos resíduos, recorreu-se ao teste de Shapiro-Wilk. As hipóteses a testar são as seguintes:

$$H_0 : \epsilon_t \quad seguem \quad uma \quad distribuição \quad normal \quad N(\mu, \sigma^2)$$

$$vs$$

$$H_1 = \epsilon_t \quad não \quad seguem \quad uma \quad distribuição \quad normal \quad N(\mu, \sigma^2)$$

As conclusões são retiradas a partir dos seguintes outputs (Figura 81):

FIGURA 81. Teste do Shapiro-Wilk para as variáveis Cost VF (esq.) e Total Conv. Value (dir.).

Para um nível de significância $\alpha = 5\%$, aceita-se a hipótese nula, para ambas as variáveis, parecem ter resíduos gaussianos, ou seja, é possível verificar que os valores do p-valor aceitam a normalidade dos resíduos.

6. Previsão

As previsões desempenham um papel cada vez mais importante numa empresa moderna, pois elas são usadas para programação de produção, orçamento de capital e para a alocação de recursos em projetos. Uma aplicação importante do modelo de regressão é estimar valores da variável resposta para um valor específico dos estimadores.
Para tal finalidade, deve-se construir intervalos de confiança e de previsão para as estimativas.
A previsão está no centro da função de planejamento das organizações, por auxiliar nas tomadas de decisões. Além disso, têm uma série de benefícios que a previsão pode gerar para as organizações, tais como:

(1) melhoria da informação estratégica;
(2) melhoria das informações de marketing;
(3) melhoria das informações financeiras;
(4) etc.

6.1. Previsão - Análise de Regressão Múltipla

Foi utilizado o comando *predict()* do software R para obter as previsões, onde foram retirados apenas 6 exemplos destas primeiras, para ambas as variáveis respostas e obteve-se os seguintes resultados.

6.1.1. Facebook Ads.

Neste capítulo será apresentada as previsões do custo e da receita do Facebook Ads, pelo método de Regressão Múltipla. Os valores do coeficiente de determinação ($\mathbb{R}$) e do coeficiente de determinação ajustado ($\mathbb{R}^2$ para a variável *Spend* são ambos 0.64 e para a *Conversion Value* são ambos de 0.95. O $\mathbb{R}$ mede a quantidade de variabilidade nos dados explicada pelo modelo de regressão e o $\mathbb{R}^2$ mede a proporção de variação na variável dependente que é explicada pelas variáveis independentes. Ambos os valores são significativamente altos, o que torna um modelo ótimo. Para o valor de SSE, são ambos valores pequenos, muito próximos de zero, o que representa terem uma pequena dispersão no que toca aos valores originais em relação aos valores ajustados.

Analisou-se a a variância, ou seja, testou-se a significância geral da regressão, e confirmou-se que existe relação estatística significante entre a variável dependente e as

variáveis explicativas, para os dois modelos, com valores de estatística F de 813 para *Spend* e de 6162 para *Conversion Value*.

Os modelos de regressão adequados no ajuste dos dados serão representados a seguir:

Modelo de regressão múltipla para a variável resposta Spend:

$$Y = 4.46 + 0.1565 SCORES.PC1 + 0.071 SCORES.PC3 \tag{15}$$

Modelo de regressão múltipla para a variável resposta Conversion Value:

$$Y = 4.94 - 0.467 SCORES.PC1 + 0.148 SCORES.PC2 \tag{16}$$

Nestas equações representadas a cima encontram-se a relação entre as variáveis e os modelos a serem usados para fazer a previsão.

Nas Tabela 43 e 44 está explícito que os valores previstos como os originais são muito semelhantes, assim como os intervalos de confiança de cada valor são precisos e os valores originais devem estar contidos nele.

Tabela 43. Previsão da regressão múltipla da variável resposta Spend.

N da campanha	Valor original	Valores previstos	Intervalo (95%)
1	4.264368	4.370190	[4.339363;4.401017]
2	4.985368	4.707832	[4.675939;4.739725]
3	4.736381	4.467704	[4.434532; 4.500876]
4	4.536631	4.485566	[4.454403; 4.516728]
5	3.971548	4.138115	[4.093791; 4.182439]
6	5.674412	5.767338	[5.691559 ; 5.843117]

Tabela 44. Previsão da regressão múltipla para a variável resposta Conversion Value.

N da campanha	Valor original	Valores previstos	Intervalo (95%)
1	5.688831	5.733104	[5.701748;5.764460]
2	5.216497	5.107654	[5.091005; 5.124304]
3	4.570308	4.650077	[4.631220;4.668934]
4	4.486500	4.518775	[4.500593;4.536957]
5	3.468896	3.575030	[3.545017;3.605043]
6	4.877417	4.953928	[4.921718;4.986138]

6.1.2. Google Ads.

As previsões do custo e da receita do Google Ads tiveram o auxílio de um outro comando do *software R*. Como foi utilizado a transformação do bestNormalize nas variáveis utilizadas, tive de fazer o inverso dessa transformação quando descobri os valores preditos para cada uma das variáveis respostas.
Utilizei uma segunda vez a função *predict()*, mas com o argumento $inverse = TRUE$, e assim para as variáveis que foram transformadas por cada tranformação que lhes foi atribuída, é-lhes aplicado o inverso dessa fórmula:

- pelo *orderNorm* : $g(x)=\Phi^{-1} \cdot (\frac{(rank(x)-0.5)}{(length(x))})$; como é o caso das variáveis *Cost VF*, *Total Conv. Value*, *CTR*, *CPA*, *ROAS* e *CR*;
- pela *Transformação de Yeo-Johnson*, Equação 2; como é o caso da variável CPC;
- pela *Transformação de Box-Cox*, Equação 1; como é o caso da variável impressions.

Os valores do coeficiente de determinação ($\mathbb{R}$) e do coeficiente de determinação ajustado ($\mathbb{R}^2$ para a variável *CostVF* são ambos 0.67 e para a *Total Conv. Value* são ambos de 0.98, o que se pode considerar valores relativamente significativos.
Para o valor de SSE, são ambos valores pequenos, muito próximos de zero, o que representa terem uma pequena dispersão no que toca aos valores originais em relação aos valores ajustados.

Novamente para os dados do Google Ads, analisou-se a a variância, ou seja, testou-se a significância geral da regressão, e confirmou-se que existe relação estatística significante entre a variável dependente e as variáveis explicativas, para os dois modelos, com valores de estatística F de 303.1 para *CostVF* e de 8405 para *Total Conv. Value*.

Os modelos de regressão adequados no ajuste dos dados serão representados a seguir:

Modelo de regressão múltipla para a variável resposta Cost VF:

(17)
$$Y = (5.21e{-}07){-}(3.65e{-}01)SCORES.PC1{+}(4.21e{-}01)SCORES.PC2{-}(3.81e{-}0)1SCORES.PC3$$

Modelo de regressão múltipla para a variável resposta Total Conv. Value:

(18) $$Y = (1.42e-05) - (6.53e-02)SCORES.PC1 + (9.20e-01)SCORES.PC2$$

Nestas equações representadas a cima encontram-se a relação entre as variáveis e os modelos a serem usados para fazer a previsão.

Sendo assim, prosseguiu-se com a análise das próximas tabelas (Tabela 45 e 46), onde estão referidos os valores originais e previstos que os modelos de cada variável resposta proporcionou.

Tabela 45. Previsão da regressão múltipla para a variável resposta Cost VF.

N da campanha	Valor original	Valores previstos	Intervalo (95%)
1	386.20	386.1059	[358.7023 ;400.3337]
2	292.03	342.5587	[329.3481;370.1194]
3	410.90	420.6234	[394.6159 ; 440.1502]
4	283.76	285.2616	[256.5132; 305.0368]
5	414.95	299.7745	[283.4367; 305.5232]
6	371.41	266.6002	[247.8632 ; 284.0331]

Tabela 46. Previsão regressão múltipla para a variável resposta Total Conv. Value.

N da campanha	Valor original	Valores previstos	Intervalo (95%)
1	1361	1359.1894	[1318.9097 ;1391.9013]
2	4258	4759.5904	[4497.7243; 4873.5280]
3	12373	11032.7398	[10915.4291;11210.5091]
4	1055	865.3911	[848.2758 ; 894.1025]
5	507	491.9424	[480.0312;495.8572]
6	464	468.6379	[460.5853 ;474.3847]

6.2. Previsão - Séries Temporais

Como indica o esquema da Figura 2, as Séries Temporais têm, nesta última fase que se realizam, as previsões, usando o modelo resultante do Passo 4. Esta fase é composta por um único passo (Passo 5) que é a previsão propriamente dita.
A previsão pode não ser tão precisa quanto o intervalo de predição sugere, pois, a modelagem matemática pode ser muito complexa para permitir que uma incerteza adicional seja incluída no modelo.

A partir do modelo adequado, divulgado anteriormente, serão efetuadas previsões para as semanas futuras, mas de modo a verificar se as previsões se aproximam da realidade, serão ignorados as útimas 5 semanas de estudo e calculadas as previsões para essas semanas e respetivos desvios padrão, recorrendo à função *predict()* em ambiente R.

Como os dados foram transformados pela transformação de *Box-Cox*, é necessário fazer o inverso dessa transformação, para que psteriormente, sejam comparados os valores originais da série com os previstos de forma a obter uma precisão na previsão do custo e da receita das campanhas dos dois canais (*Facebook Ads* e *Google Ads*), sendo o ideal, estas previsões estarem dentro do intervalo de 95% de confiança dos desvios padrão dos valores previstos.

6.2.1. Facebook Ads.

Como referido, serão retiradas 5 semanas da série temporal, ou seja, as observações das últimas cinco semanas serão retiradas, fazendo com que as duas séries temporais (tanto do custo como da receita) perfaçam um total de 53 observações. As equações dos modelos obtidos para cada uma das variáveis serão representadas a seguir:

Modelo de séries temporais para a variável resposta Spend:

(19)
$$(1-(-2.054)B-(-1.948)B^2-(-1.438)B^3-(-0.648)B^4)(1-B)^1X_t = (1-1.38\,B-0.494B^2)\epsilon_t$$

Modelo de séries temporais para a variável resposta Conversion Value:

(20)
$$(1-(-0.692)B)(1-B)^2X_t = 0$$

É utilizado o modelo *ARIMA(4,1,2)* com 53 observações para a variável *Spend* e o modelo *ARIMA(1,2,0)* para a variável *Conversion Value*, e desta forma procede-se às previsões para as últimos cinco semanas.

Tabela 47. Valores previstos obtidos em ambiente R para as previsões dos últimos 5 meses da série temporal da variável resposta Spend.

Nº da Observação	Valores previstos (transformado λ=0.8)	Erro padrão	Intervalo de Predição	Valor original
54	24880.75	6320	(11834.33;42578.93)	17404.4
55	24526.01	5077	(6133.70;36799.80)	18600.6
56	19955.98	4909	(5611.55; 36968.15)	16731.48
57	20213.13	4751.80	(8904.41; 42601.70)	15613.83
58	26268	4432	(6596.34;46678.32)	20290.34

TABELA 48. Valores previstos obtidos em ambiente R para as previsões dos últimos 5 meses da série temporal da variável resposta Conversion Value.

Nº da Observação	Valores previstos (transformado λ=0.3)	Erro padrão	Intervalo de Predição	Valor original
54	7155	199278.2	(-222500;161474.70)	10545
55	10385.92	79234.78	(-21510.9 ;78712)	27040
56	10231.23	27893.04	(-41829.70; 671882)	13043
57	16652.44	6167.72	(-499.49 ; 285075)	13191
58	13142.5	1351.75	(12.22;105119)	11800

O erro aqui é elevado, para as duas séries estudadas, assim como os seus intervalos de confiança. Os valores previstos e originais não são semelhantes assim como os intervalos de confiança, pelo que estes modelos em termos de previsão não funcionam bem e os intervalos apresentam uma aplitude demsiado grande. Uma série com poucos dados dificulta uma boa estimação e previsão.

Vamos representar graficamente esta previsão.

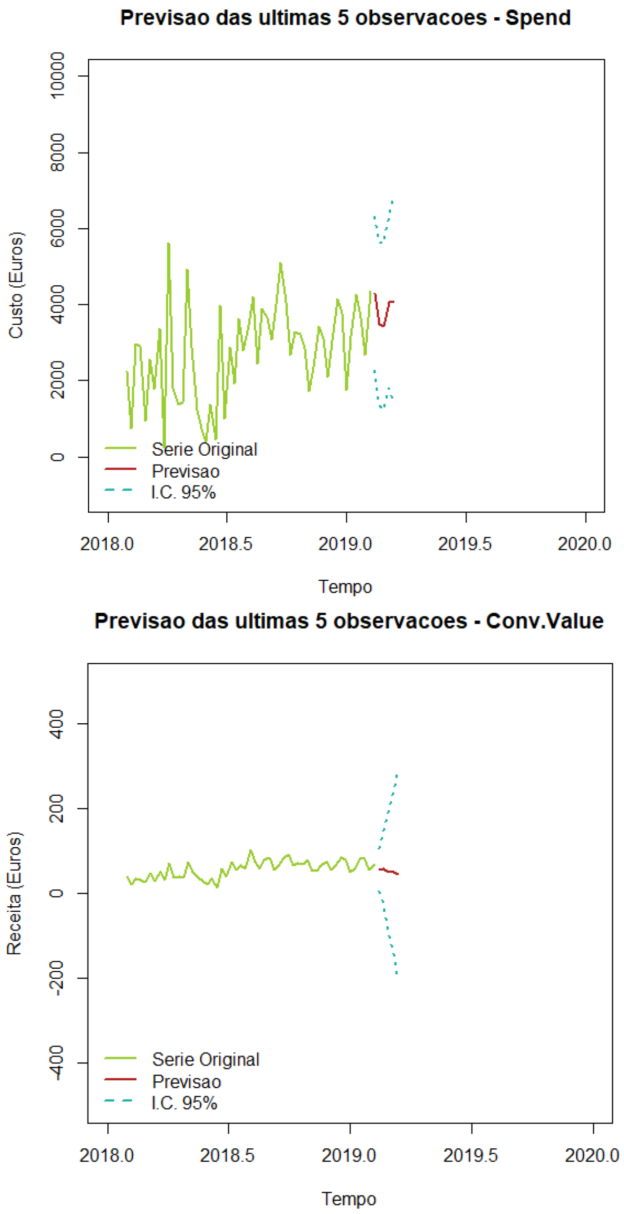


FIGURA 82. Gráfico da Previsão das últimas 5 observações para as variáveis respostas Spend (em cima) e Conversion Value (em baixo).

6.2.2. Google Ads.

Como foi feito para o Facebook Ads, será elaborado igual, neste capítulo, para o Google Ads; serão retiradas 5 semanas da série temporal, ou seja, as observações das últimas cinco semanas serão retiradas, fazendo com que as duas séries temporais (tanto do custo como da receita) perfaçam um total de 53 observações. As equações dos modelos obtidos para cada uma das variáveis serão representadas a seguir:

Modelo de séries temporais para a variável resposta Cost VF:

$$(1 - (-0.7254)B - (-0.2843)B^2)(1 - B)^1 X_t = 0 \tag{21}$$

Modelo de séries temporais para a variável resposta Total Conv. Value:

$$(1-(-0.7178)B-(-0.4678)B^2-(-0.4292)B^3)(1-B)^2X_t=0 \tag{22}$$

É utilizado o modelo *ARIMA(2,1,0)* com 53 observações para a variável *Cost VF* e para o modelo *ARIMA(3,1,0)* da variável *Total Conv. Value* com 47 observações , e desta forma procede-se às previsões para as últimos cinco semanas.

TABELA 49. Valores previstos obtidos em ambiente R para as previsões dos últimos 5 meses da série temporal da variável Cost VF.

Nº da Observação	Valores previstos (transformado λ=-0.05)	Erro padrão	Intervalo de Predição	Valor original
49	6745.73	165	(1510.32:33440.37)	2486.09
50	4880.28	1.71	(1041.73;26018.80)	4970.38
51	9082.69	1.76	(1720.34; 55773.60)	2713.30
52	6958.84	1.91	(1071.66;55773.60)	12856.75
53	6242.37	1.99	(892.07;55773.60)	2631.02

TABELA 50. Valores previstos obtidos em ambiente R para as previsões dos últimos 5 meses da série temporal da variável Total Conv. Value.

Nº da Observação	Valores previstos (transformado λ=0.25)	Erro padrão	Intervalo de Predição	Valor original
43	13680.6	148.35	(1236.57;60602.67)	8467
44	30359.5	226.63	(3232.10 ;126656)	9319
45	19256.66	275.74	(1096.93; 100613.36)	5018
46	15456.08	266.39	(647.80;88543.44)	5290
47	13807.56	364.69	(318.64;86427.36)	8394

Os valores previstos estão mais próximos dos originais quando se trata da variável *Cost VF*, mas têm um intervalo de confiança enorme, e o erro é considerado pequeno. Para a variável *Total Conv. Value* os resultados são bastante negativos. Não se obteve boas estimativas de valores previsto.

Nas figuras seguintes estão apresentadas graficamente esta previsão.

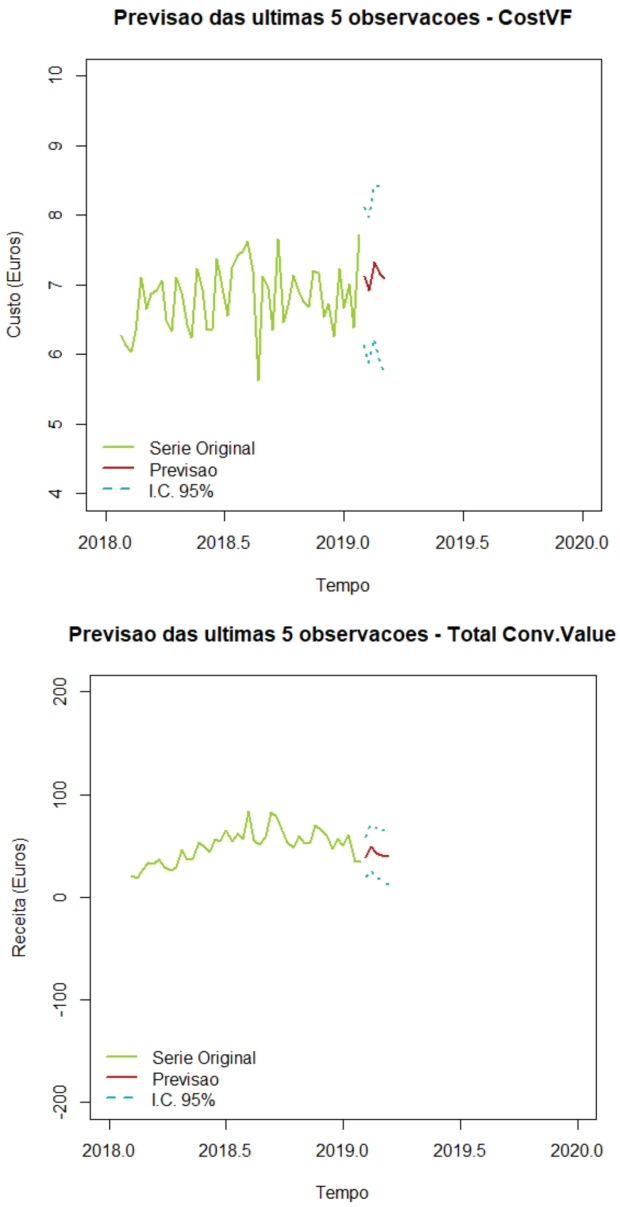


FIGURA 83. Gráfico da previsão das últimas 5 observações para as variáveis respostas Cost VF (em cima) e Total Conv. Value (em baixo).

7. Conclusão

O objetivo principal, como mencionado na introdução deste trabalho, foi o de comparar a análise de regressão múltipla com as séries temporais para a previsão do custo e da receita das campanhas associadas aos dois canais, Facebook Ads e Google Ads, com dados reais de uma plataforma de venda digital (Overcube). Na fase exploratória dos dados verificou-se a existência de muitos valores nulos associados ao facto da empresa ser recente e corresponderem ao início da sua atividade. Assim, foi decidido eliminar parte desses valores, que correspondiam ao *burn in* do modelo de negócio associado à venda por canais.

O objetivo deste trabalho consiste em avaliar a precisão pelos métodos de regressão múltipla e séries temporais, aplicados a dados de venda via canais de redes sociais através da implementação de campanhas diferenciadas. A nível dos diferentes tipos de campanhas, destacam-se as do tipo *Remarketing* para o Facebook Ads e as do tipo *Search* para o Google Ads. Estas campanhas estão associadas a maior número de vendas, e consequentemente, maior valor de receita para a empresa.

Da análise detalhada das variáveis iniciais, resultaram diferentes modelos de previsão a nível da regressão múltipla e das séries temporais para os diferentes canais estudados. Os resultados do estudo comparativo, sugerem que os modelos de regressão múltipla são mais adequados para prever o custo e a receita associados aos diferentes tipos de campanhas. Estes modelos, são mais eficazes na explicação das variáveis resposta (Custo e Receita). A análise de regressão foi realizada para as variáveis Spend e Conversion Value.

Os valores de $R^2_{ajustado}$ do Facebook Ads são de 92% e de 64%, e para o Google Ads (*CostVF* e *Total Conv.Value*) são de 98% e de 68%, respetivamente.

A nível da análise das séries temporais, o modelo ARIMA selecionado apresenta para todos os casos um padrão idêntico a nível dos erros e dos valores preditos. É de salientar que o número reduzido de observações por semana condicionou a previsão das variáveis resposta custo e receita.

Na comparação dos modelos obtidos destacam-se as seguintes conclusões: ambos os tipos de modelos servem para realizar previsões destacando-se, contudo, o modelo de regressão múltipla como sendo o que melhor predita as variáveis respostas, pois foi onde se obteve melhores resultados, como descrito no capítulo 6.1.

Ambas as metodologias são adequadas no contexto do trabalho, mas contudo, a análise de séries temporais fica condicionada pelo número reduzido de observações, devido ao facto da empresa ser nova. Uma actualização da base de dados permitiria um aumento de registos o que iria contribuir para uma melhor qualidade da previsão associada às séries.

8. Bibliografia

[1] Sen, Ashish & Srivastava, M.(2012). *Regressions analysis: theory, methods, and applications.*. Springer Science and Business Media.

[2] Pedro Santos. *Facebook Ads ou Google Adwords: qual é o melhor para seu negócio?, 2018.*.
`https://rockcontent.com/blog/facebook-ads-ou-google-adwords/`

[3] *Package 'bestNormalize' , 2019.*
`https://cran.r-project.org/web/packages/bestNormalize/bestNormalize.pdf`

[4] Wheelwright, Steven & Makridakis, S. H.R.J. (1998) *Forecasting: methods and applications.* John Wiley & Sons.

[5] RockContent *Guia completo do Google Analytics.*
`https://materiais.rockcontent.com/guia-google-analytics`

[6] Tsay, R. (2010) *Analysis pf Financial Time Series.* John Wiley and Sons, 3ªed.

[7] Fahrmeir, L. and Kneib, Thomas & Lang, S.M.B. (2013) *Regression: models, methods and applications.* . Springer Science & Business Media

[8] Box. G & Jenkings, G.R.G. (2013) *Times series analysis: forecasting and control.* John Wiley and Sons.

[9] Chatfield, C. (2004) *The analysis of time series: an introduction.* Chapman and Hall/-CRC, 5ªed.

[10] Jackson, J.E. (2005) *A user's guide to princiapl components.* New York, Wiley.

[11] Metcalfe, Andrew V. & Cowpertwait, P.S. (2009) *Introductorry time series with R.* Springer.

[12] Jolliffe, I.T. (2002) *Principal Component Analysis, 2ªed.* New York, Springer-Verlag

[13] João Pedro Bento Clemente da Silva *Modelos de Regressão Linear e Logística utilizando o software R.*

[14] Diogo Borges Provete, Fernando Rodrigues da Silva, Thiago Gonçalves Souza *Livro Estatística aplicada à ecologia usando o R .*

[15] *Análise de Variância (Teste F) - Medidas de Associação.*
`http://www.portalaction.com.br/analise-de-regressao/24-analise-de-variancia-te`