|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | | | |  | |
|  | escola  superior  de tecnologia  e gestão  POLITÉCNICO  DO PORTO | | | | |  | |
|  | | | CTeSP |  |  | | |
| Curso Técnico Superior Profissional  Desenvolvimento para a Web e Dispositivos Moveis | | |  |
|  | | | Relatório Projeto Final de Oficina de Estatística  Daniel Faria – 8200113  Adélio Carneiro - 8200107  21/2022 | | | |  |
|
|
|  | | | | | | | |

Índice

[Introdução 3](#_Toc94968595)

[Variáveis em Estudo 3](#_Toc94968596)

[Filial 3](#_Toc94968597)

[Funcionários 3](#_Toc94968598)

[Faturação 4](#_Toc94968599)

[Despesa 4](#_Toc94968600)

[Avaliação 4](#_Toc94968601)

[Comparações entre os grupos 5](#_Toc94968602)

[Regressão Linear 6](#_Toc94968603)

[Faturação - Despesa 6](#_Toc94968604)

[Faturação - Funcionários 7](#_Toc94968605)

[Estimação 8](#_Toc94968606)

[Teste a Normalidade das variáveis 9](#_Toc94968607)

[Médias entre grupos 10](#_Toc94968608)

[Regressão Linear 11](#_Toc94968609)

[Conclusão 12](#_Toc94968610)

# Introdução

No âmbito da disciplina de Oficina de Estatística foi requisitado o estudo e analise de 20 registos de faturação mensal de várias empresas pertencentes a uma filial.

Este trabalho ira ter dois objetivos principais:

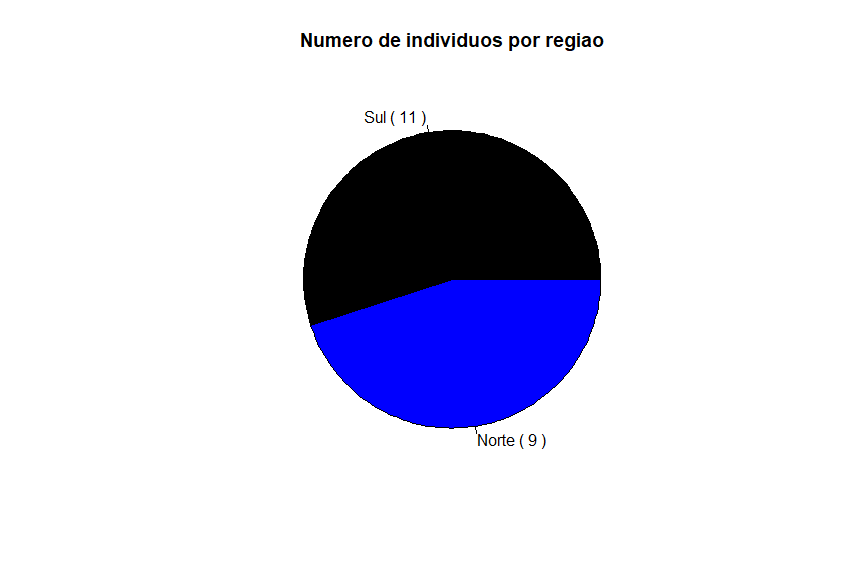
O primeiro sendo uma análise exploratória dos dados da amostra. E a segunda ira basear-se na estimação de parâmetros e na previsão com base no modelo Regressão Linear Simples.

# Variáveis em Estudo

## Filial

Qualitativa nominal

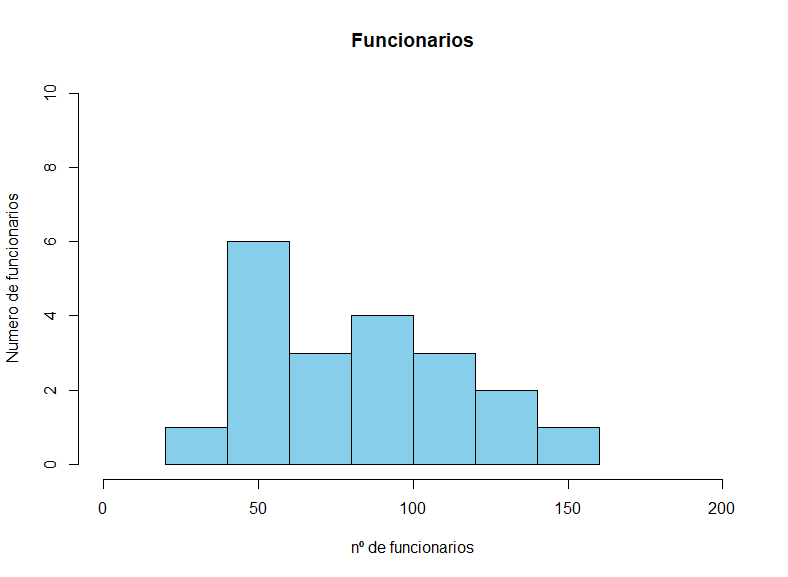
Esta variável identifica de qual filial e o registo mensal da empresa, podendo ter 2 valores diferentes: “Norte” e “Sul”.

No gráfico podemos ver que na amostra que nos foi dada 11 dos 20 registos foram da filial do Sul, sendo o restante da filial do Norte.

## Funcionários

Quantitativa continua

A variável funcionários e o número de funcionários de cada empresa na altura do registo mensal.

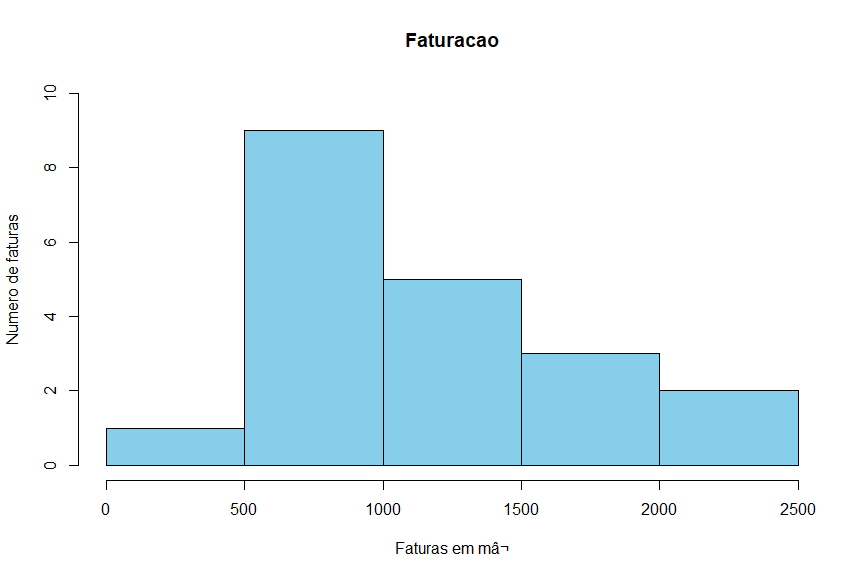


Podemos ver neste gráfico a concentração da variável, tendo todas as empresas entre 28 funcionários e 156. Sendo a mediana 75, e a média de 81.85 funcionários.

## Faturação

Quantitativa continua

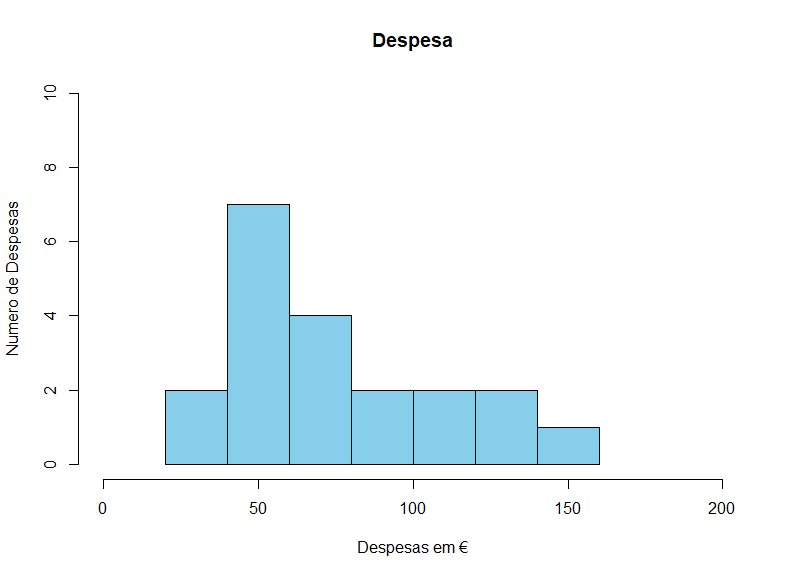
Esta variável tem o valor da faturação da empresa mensalmente, medida em milhares de euros.

Neste gráfico podemos ver que as faturações tem uma maior concentração na classe de 500 a 1000, tendo uma valor mediano de 980.8. E um valor mínimo de 320 e um valor máximo de 2292.3.

## Despesa

Quantitativa continua

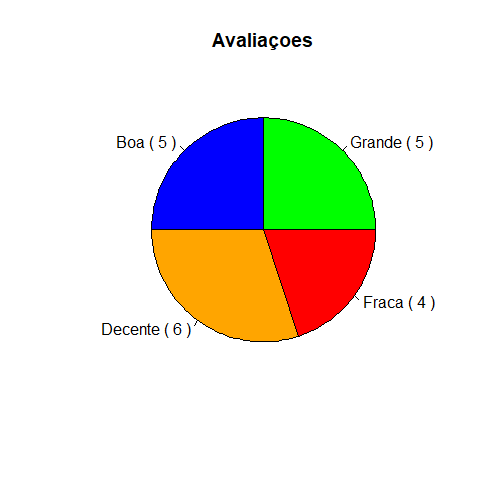
A variável despesa é os valores da despesa de cada empresa mensalmente, medida em milhares de euros.

Neste histograma podemos ver uma maior concentração na segunda classe, que vai de 40 a 60, sendo cada classe de 20.

## Avaliação

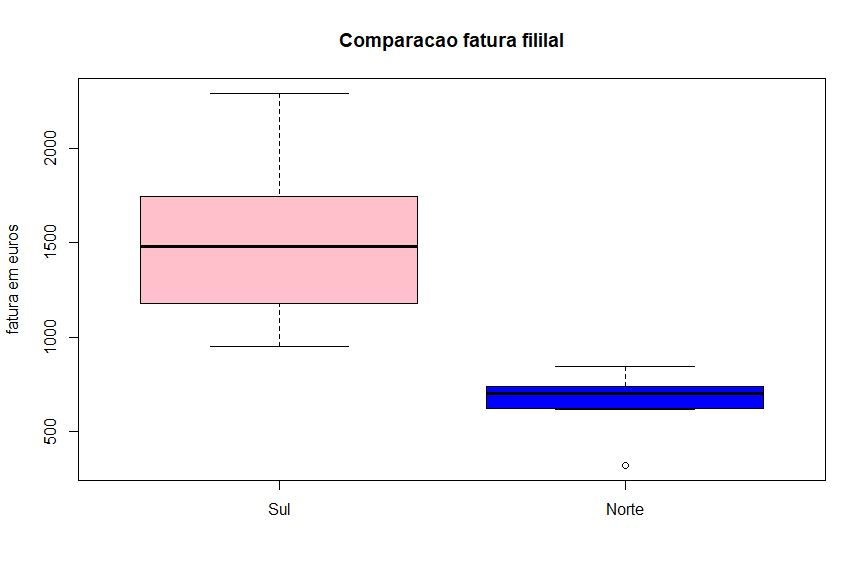
Qualitativa ordinal

Esta variável ordinal, pode ter 4 valores distintos, de 1 a 4. Sendo cada número correspondente a uma avaliação: “Poupança Grande”, “Poupança Boa”, “Poupança Decente”, “Poupança Fraca”.



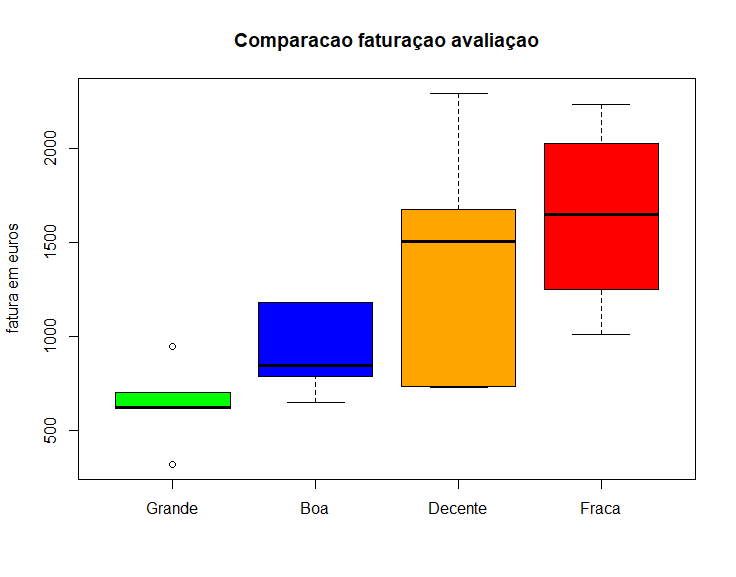
Neste piechart podemos ver a distribuição das avaliações das empresas na amostra, sendo a mais predominante uma poupança decente. Mesmo assim as outras avaliações estão distribuídas de forma igual tendo pouca diferença entre elas.

# Comparações entre os grupos



Nesta comparação entre Norte e Sul pela faturação de cada grupo, podemos ver que os valores são muito mais altos no Sul, sendo que o valor máximo do Norte e menor que a faturação mais pequena do Sul, podemos também ver uma maior dispersão de valores no Sul, sendo que os valores do Norte são menos dispersos, dando a aparência díspar dos dois gráficos.

Também podemos ver um outlier no grupo Norte, uma fatura de 320 m€.

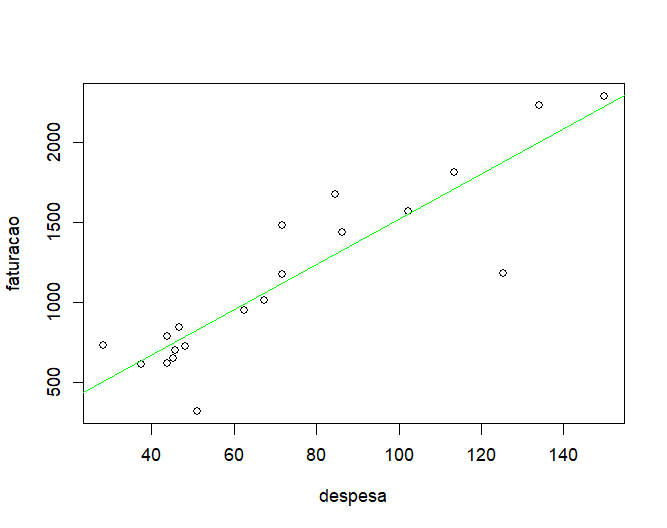


Neste gráfico de extremos e quartis podemos ver a distribuição das faturas pela avaliação mensal da empresa.

Podemos tirar a conclusão de que quanto maior a poupança menor é a faturação, também podemos ver que temos dois outliers na poupança “Grande”, sendo eles 320 m€ e 948.8 m€.

# Regressão Linear

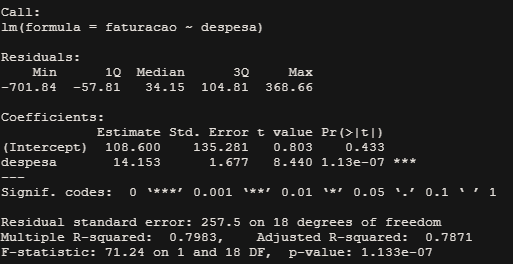
## Faturação - Despesa



Neste diagrama de dispersão podemos ver a correlação entres duas variáveis, nomeadamente a faturação no eixo do y, e a despesa no eixo do x.

Em cima podemos ver a função y = a + b \* x

Sendo o coeficiente de correlação 0.89, podemos dizer que estas duas variáveis tem uma correlação positiva forte. Enquanto isso não implica causalidade entre as duas variáveis com uma correlação forte podemos usá-la para estimar valores com baixo nível de erro.



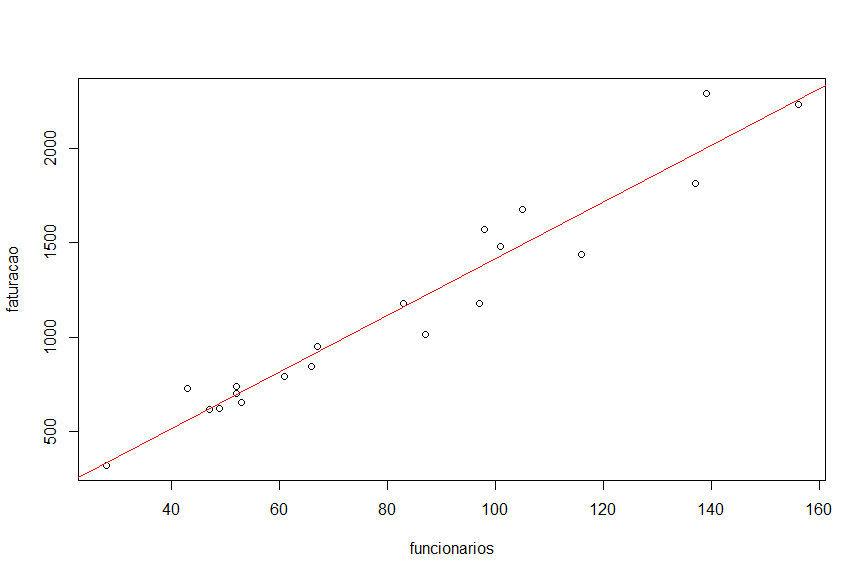
Nesta imagem podemos ver os coeficientes do modelo de regressão linear sendo o intercept estimado quando a faturação for 108.6 e a despesa 14.1. O que significa que a faturação estimada e 108.6 quando a despesa é 0, e por cada valor da despesa a ser adicionada sobe 14.1.

Podemos inferir também pelo p –value ser menor que o valor de significância, que este modelo pode apresentar estimações corretas, sendo que podemos ter o risco de rejeitar a hipótese nula.

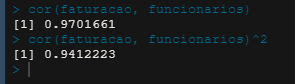


Segundo o coeficiente de determinação podemos concluir que 80% da variância da faturação e explicada pela variância da despesa.

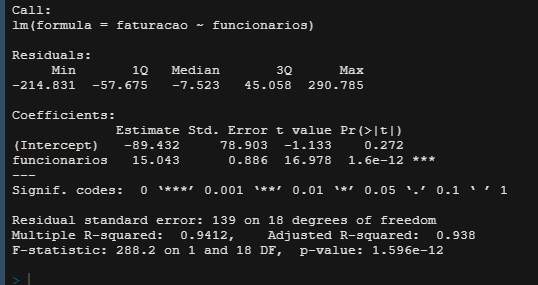
## Faturação - Funcionários



Neste diagrama de dispersão podemos ver a correlação entres duas variáveis, nomeadamente a faturação no eixo do y, e a funcionários no eixo do x.



Com este coeficiente de correlação podemos dizer que estas duas variáveis mantem uma correlação positiva forte. E pelo coeficiente de determinação podemos dizer que 94 porcento da variação da faturação e devido a variância da variável funcionários.



Com este gráfico podemos tirar algumas inferências do modelo, uma delas sendo que se uma empresa não tiver funcionários ira em base perder 89 m€ por mês. Também podemos tirar que por cada funcionário contratado a empresa de acordo com o modelo ira aumentar a fatura por 15 m€ mensais.

Podemos inferir também pelo p –value ser menor que o valor de significância, que este modelo pode apresentar estimações corretas, sendo que podemos ter o risco de rejeitar a hipótese nula.

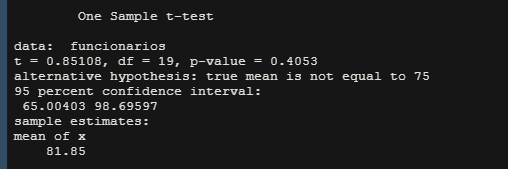
# Estimação

3.1.1. Sabe-se que a filial do Norte é mais recente. A proporção de registos para esta filial é de 21%. Calcule a probabilidade de se considerarem 9 registos mensais da filial N numa amostra de 20, por recurso à distribuição Binomial. Conclua acerca da representatividade da amostra em estudo em relação à localização da filial;



Como podemos ver pelo resultado da função dbinom, a probabilidade de 9 registos mensais retirados aleatoriamente ser da filial Norte é de a volta de 1 porcento. O que significa que a amostra que nos foi dada no enunciado do trabalho não é representativa da população.

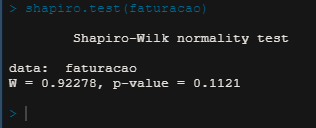
3.1.2. O gestor considera que o número médio de funcionários é de 75 por loja. Verifique se o gestor tem razão.



Utilizando o teste de media á variável podemos tirar conclusões sobre a media da população só com os dados da amostra. Utilizando a media dada pelo gestor, podemos fazer um teste a ver se podemos aceitar a media como correta.

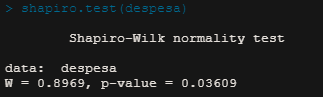
Sendo que o p-value e maior do que o valor de significância não podemos rejeitar a hipótese nula, o que significa que não podemos utilizar 75 como media da população.

# Teste a Normalidade das variáveis

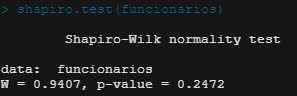


O nosso primeiro teste á normalidade foi utilizando a variável faturação, aonde o p-value é de 0.11, o que nos obriga a rejeitar a hipótese nula, levando isso a conclusão de que a variável não segue uma distribuição é normal.

Mas de acordo com o que foi proposto com a professora, iremos proceder como se a variável fosse normal.



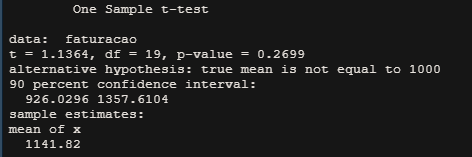
Neste teste á normalidade o valor de p e menor que o valor de significância o que nos leva a não rejeitar a hipótese nula, o que em conclusão diz nos que a variável segue uma distribuição normal.



Neste teste Shapiro-Wilk para testar se a variável segue uma distribuição normal, podemos inferir devido ao valor do p-value sendo maior que o valor de significância atribuído ao teste que esta variável não segue uma distribuição normal.

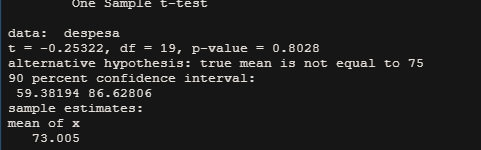
Como a primeira variável testada, mesmo sendo ela não normal vamos proceder como fosse.

3.1.4. Estime outros valores médios populacionais com base nos dados e apresente intervalos de confiança a um grau de confiança de 90%;



Na variável faturação com um nível de confiança de 90 porcento e uma média de 1000, podemos aceitar 1000 como media populacional, sendo que o p-value e maior que o nível de significância, o que nos leva a não rejeitar a hipótese nula, sendo a hipótese nula que a média é igual.

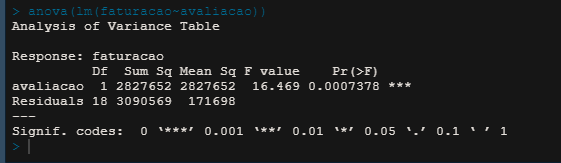
Os intervalos de confiança são de 926.02 a 1357.61.



Na variável despesa com um nível de confiança de 90 porcento e uma média de 75, podemos dizer que 75 e uma média que pode representar a população, sendo que o p-value e superior ao valor da significância levando isso a não rejeitar a hipótese nula.

Os intervalos de confiança são 59.38 a 86.62.

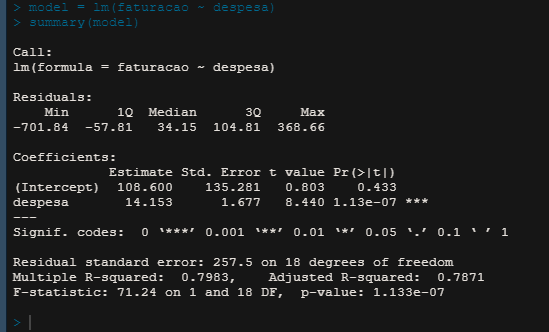
# Médias entre grupos



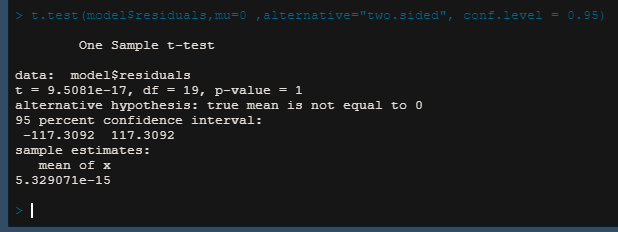
Nesta imagem podemos ver o resultado de um teste de grupos da faturação em conta da avaliação, podendo dizer que as medias dos grupos não são iguais, desde que o p-value e negativo, sendo que rejeitamos a hipótese nula, sendo a hipótese nula que as medias dos grupos são iguais.

# Regressão Linear

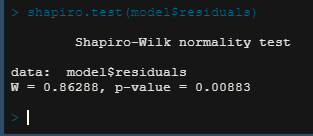
Significância dos coeficientes de regressão e de determinação:

As duas variáveis têm uma correlação sendo que o p-value e menor que o valor de significância por isso rejeitamos a hipótese nula, sendo a hipótese nula que os valores não tem correlação.

Teste à Normalidade e a média dos erros do modelo:



O p-value é 1. O que é esperado.



No teste de shapiro-wilk que fizemos aos resíduos do modelo podemos ver que o p-value é menor que o valor de significância, o que nos leva a rejeitar a hipótese nula. Sendo assim podemos dizer que os erros não seguem uma distribuição normal.

3.2.3. Sabe-se que se contratou para o próximo mês uma Despesa em publicidade de 87.8 m€. Com base na reta ajustada preveja o valor da faturação esperada nesse mês.

Utilizando o modelo podemos concluir que com uma despesa de 87.8m€ a empresa ira ter uma faturação de 1351.2334.

3.2.4. Preveja o valor da Despesa em Marketing que se deve contratar para atingir o objetivo de 3000 m€ em faturação.

Utilizando o modelo linear podemos concluir que 204.29 m€ iria ser o valor necessário para chegar a 3000 m€, com isso também podemos dizer como o valor 3000 esta fora do modelo previsto, que o valor da despesa pode estar incorreto.

3.2.5. Faça previsão de valores “in the box”, “out of the box” de outras regressões;

Este e uma previsão out the box, porque o valor de despesa esta fora dos intervalos de confiança.

Sendo o valor de y 120, de acordo com o modelo linear o valor de x será 1806.96

Esta e uma previsão in the box, porque o valor da faturação estra dentro dos valores de confiança.

Com x a valer 950, y equivale a 59.45.

# Conclusão

Concluímos todos os objetivos propostos no enunciado do trabalho, mesmo assim podíamos ter feito mais, como áreas que podíamos ter aprofundado mais tais como as comparações entre grupos que só completamos duas, porém, poderíamos ter feito mais comparações como entre a variável funcionário ou despesa com as duas variáveis qualitativas, filial e avaliação.

Outras comparações que poderíamos ter abordado são as várias correlações entre as nossas variáveis.

A única dificuldade que encontramos foi no teste de médias de grupos, onde foi necessário fazer testes com variáveis com mais de dois níveis, temos depois descoberto que deveríamos ter utilizado o teste “anova”.

Apos este trabalho podemos dizer que compreendemos melhor a utilidade que estatística pode trazer no nosso futuro como profissionais. Podemos dizer que estamos orgulhosos com o que foi conseguido e com o processo até ao resultado final.