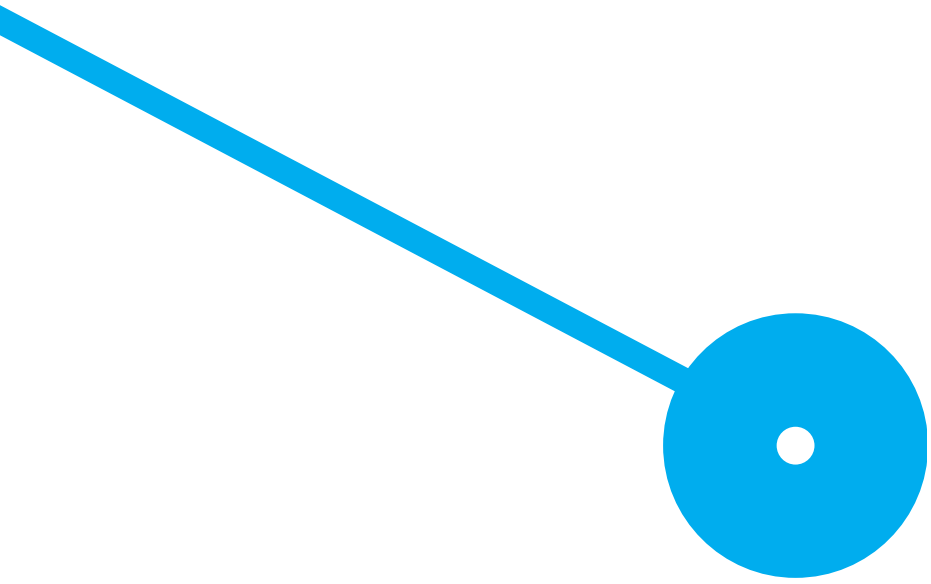
ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

CTeSP
CURSO TÉCNICO SUPERIOR PROFISSIONAL
DESENVOLVIMENTO PARA A WEB E DISPOSITIVOS MÓVEIS

Indústria Automóvel

Maria Dias
Tiago Costa

2021/2022



1 Introdução

Recolheram-se dados relativos a 21 **testes** executados em automóveis. Identificou-se o **condutor** que executou o teste com as letras A ou B, o **peso** de cada viatura testada, em kg, a **distância** que cada automóvel percorreu com 10 litros de gasóleo, a distância de travagem, o **consumo** médio de litros aos 100km e a **avaliação** de desempenho ecológico que varia entre 1 e 4 (1 – desempenho muito bom; 4 – desempenho muito fraco).

O "**teste**" é uma variável quantitativa discreta finita, porque realizou-se no máximo 21 testes.

O "**condutor**" é uma variável qualitativa nominal (binominal), por essa razão, foi codificada. O "0" corresponde a "A" e o "1" corresponde a "B".

O "**peso**", a "**distancia**" e o "**consumo**" são variáveis quantitativas contínuas.

A "**avaliacao**" é uma variável qualitativa ordinal, por essa razão, foi também, codificada. O "1" corresponde a "desempenho muito bom", o "2" corresponde a "desempenho bom", o "3" corresponde a "desempenho fraco" e o "4" corresponde a "desempenho muito fraco".

Com este tema deveremos:

- Estatística Descritiva
    - Caracterizar a população e a amostra;
    - Realizar representações gráficas de variáveis individuais e de cruzamento de variáveis;
    - Identificar os "outliers", se houver, da amostra;
    - Comparar grupos (p.ex. variável numérica vs variável nominal);
    - Representar os diagramas de dispersão;
    - Determinar dos coeficientes de regressão e da equação da reta associada;
    - Interpretar os coeficientes de correlação e de determinação;
    - Analisar o ajustamento do modelo;
    - Contextualizar a análise dos resultados;

- Estatística Indutiva
    - Identificar as distribuições teóricas e aproximações;

- Calcular as probabilidades;
- Estimar as médias populacionais e Intervalos de Confiança;
- Comparar as médias de grupos;
- Representar os diagramas de dispersão;
- Determinar dos coeficientes de regressão e da equação da reta associada;
- Significância dos coeficientes de correlação e de determinação;
- Analisar o ajustamento do modelo;
- Previsão;

2 Estatística Descritiva

2.1 Identifique a população e a amostra

Recolheram-se dados relativos a 21 **testes** executados em automóveis. Identificou-se o **condutor** que executou o teste com as letras A ou B, o **peso** de cada viatura testada, em kg, a **distância** que cada automóvel percorreu com 10 litros de gasóleo, a distância de travagem, o **consumo** médio de litros aos 100km e a **avaliação** de desempenho ecológico que varia entre 1 e 4 (1 – desempenho muito bom; 4 – desempenho muito fraco).

2.2 Caracterize as variáveis em estudo

O "**teste**" é uma variável quantitativa discreta finita, porque realizou-se no máximo 21 testes.

O "**condutor**" é uma variável qualitativa nominal (binominal), por essa razão, foi codificada. O "0" corresponde a "A" e o "1" corresponde a "B".

```
condutor=c(1, 0, 0, 0, 0,
       1, 0, 1, 0, 0,
       1, 1, 1, 0, 1,
       0, 0, 1, 0, 1,
       0)
```

Figura 1: variável "condutor"

O “**peso**”, a “**distancia**” e o “**consumo**” são variáveis quantitativas contínuas.

```
peso=c(2035, 1730, 1180, 1530, 1750,
       1940, 1460, 1960, 1850, 1533,
       1760, 1650, 2050, 1710, 1897,
       1380, 1490, 1820, 1510, 1950,
       1570)
```

Figura 2: variável "peso"

```
distancia=c(138.9, 181.8, 243.9, 204.1, 178.6,
       163.9, 217.4, 163.9, 175.4, 198.2,
       178.6, 188.7, 137.0, 185.2, 166.7,
       212.8, 221.7, 172.4, 204.1, 151.5,
       196.1)
```

Figura 3: variável "distancia"

```
consumo=c(11, 6.3, 5.1, 6.3, 6.5,
       8.1, 5.3, 7.9, 7.3, 7.9,
       6.7, 6.1, 9.8, 5.9, 8.2,
       5.9, 5.2, 6.1, 5.4, 7.5,
       5.9)
```

Figura 4: variável "consumo"

A “**avaliacao**” é uma variável qualitativa ordinal, por essa razão, foi também, codificada. O “1” corresponde a “desempenho muito bom”, o “2” corresponde a “desempenho bom”, o “3” corresponde a “desempenho fraco” e o “4” corresponde a “desempenho muito fraco”.

```
avaliacao=c(4, 2, 1, 2, 3,
       4, 1, 4, 3, 3,
       3, 2, 4, 3, 4,
       2, 1, 3, 1, 4,
       2)
```

Figura 5: variável "avaliacao"

2.3 Utilize representações gráficas

```r
#Tabelas
f_g<-table(condutor) #fa do condutor
f_g #n de elementos de cada grupo (fa do condutor)
fr_g<-prop.table(f_g) #tabelas de forma proporcional da fr da fa do condutor
fr_g #fr por condutor
```

Figura 6: script tabelas condutor

Neste gráfico circular podemos observar o número de testes realizados, em automóveis, por condutor.

Tendo o condutor A realizado 12 testes e o condutor B realizado 9 testes. Com isto, podemos concluir que, o condutor A realizou mais 3 testes do que o condutor B.

```r
nomes_c<-c("Condutor A", "Condutor B") #legendas
cores<-c("purple","skyblue") #cores do grafico
rotulo<-paste(nomes_c,"(",paste(f_g),")",sep=" ") #dados (legendas, n. elementos)
pie(f_g, main="Numero de testes por condutor",labels=rotulo,col=cores) #grafico circular
```

Figura 7: script gráfico circular "Número de Testes realizados por condutor"

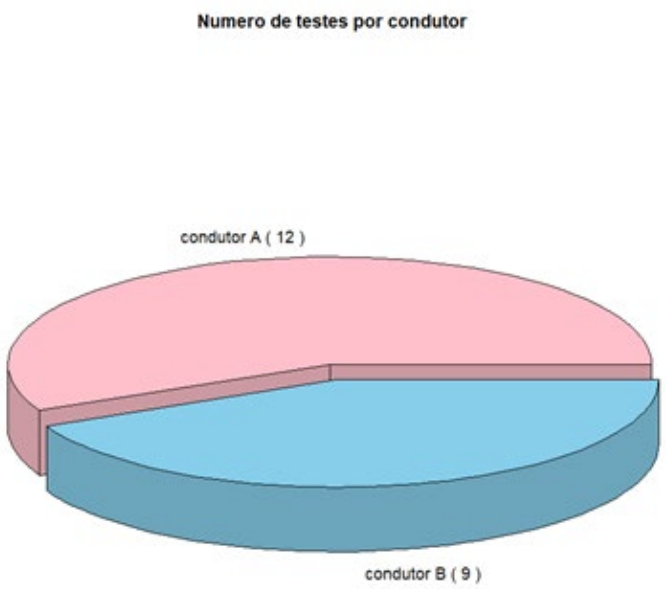


Figura 8: gráfico circular "Número de Testes realizados por condutor"

Neste gráfico circular podemos observar a percentagem de testes realizados, em automóveis, por condutor.

Tendo o condutor A realizado 57% dos 21 testes nos automóveis, enquanto o condutor B realizou 43% dos 21 testes nos automóveis. Com isto, podemos concluir que, o condutor A realizou mais de 14% dos 21 testes do que o condutor B.

```
nomes_c<-c("condutor A","condutor B") #legendas
cores<-c("pink","skyblue") #cores do grafico
rotulo<-paste(nomes_c,"(",paste(round(100*fr_g), "%"),")",sep=" ") #dados (legendas, n. elementos)
pie(fr_g, main="Percentagem de Testes realizados por condutor",labels=rotulo,col=cores) #grafico circular
```

Figura 9: script gráfico circular "Percentagem de Testes realizados por condutor"

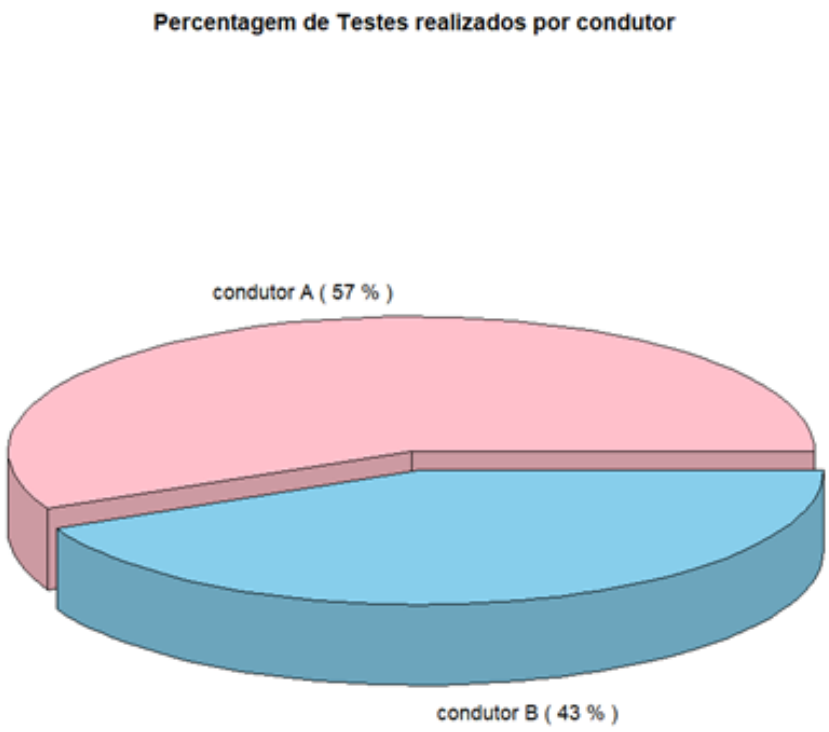


Figura 10: gráfico circular "Percentagem de Testes realizados por condutor"

Neste histograma podemos observar a distribuição dos testes realizados por distância, com 10 litros de gasóleo.

Tendo a distância no intervalo [170, 190] km o maior número de testes (7 testes), tendo as distâncias nos intervalos [150, 170] km, [210, 230] km e [230, 250] km o menor número de testes (1 teste).

A distância que cada automóvel percorreu com 10 litros de gasóleo, está entre o intervalo [130, 250] km.

```
hd=hist(distancia,main="Numero de Testes por Distancia", xlab="Distancia (Km)",ylab="N. de Testes",col = "skyblue", ylim = c(0,10))
legend("topright", legend = c("Testes por Distancia"), fill = c("skyblue"), bty = "n")
summary(distancia)
freq_rel=hd$counts
text(locator(n=7), paste(round(freq_rel)))
```

Figura 11: script histograma "Distribuição dos testes por distância, com 10 litros de gasóleo"

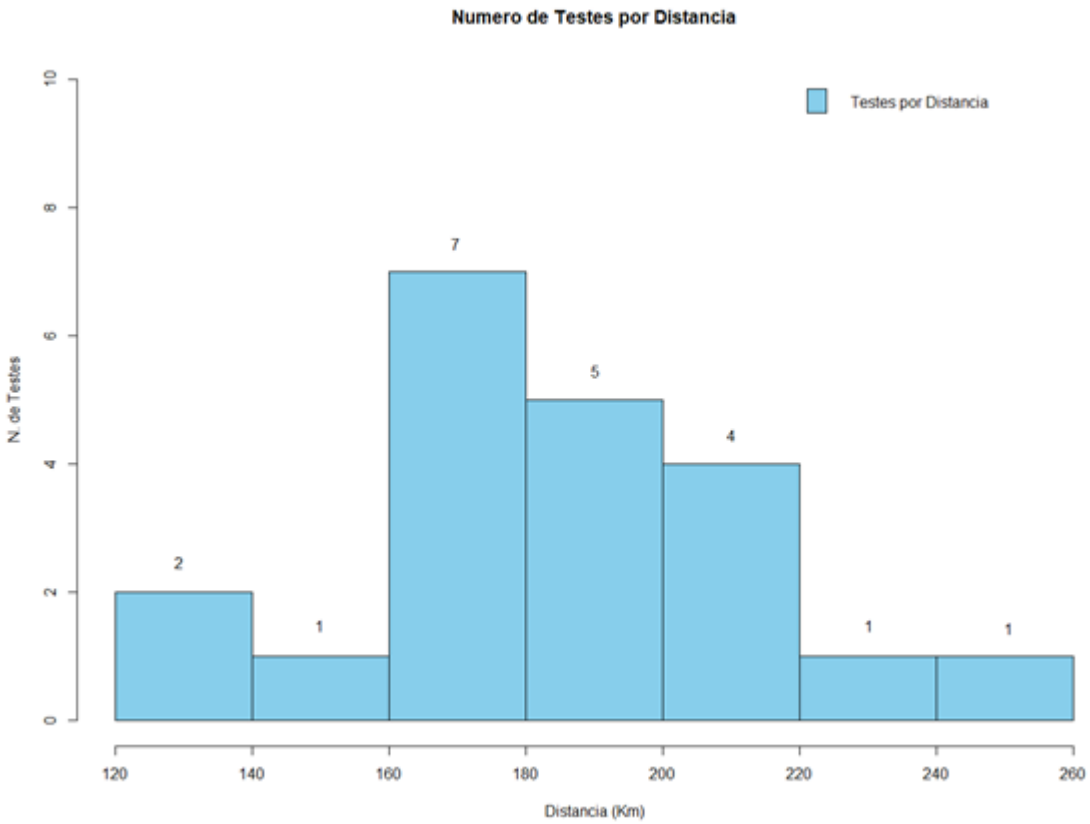


Figura 12: histograma "Distribuição dos testes por distância, com 10 litros de gasóleo"

Neste histograma podemos observar a distribuição dos testes realizados por peso do automóvel.

Tendo o peso nos intervalos [1500, 1700] kg e [1700, 1900] kg o maior número de testes (6 testes), tendo as distâncias nos intervalos [1100, 1300] kg e [1300, 1500] kg o menor número de testes (1 teste).

O peso de cada automóvel, está entre o intervalo [1100, 2100] kg.

```
hp=hist(peso, main="Numero de Testes por Peso", xlab="Peso (Kg)",ylab="N. de Testes" ,col = "skyblue", ylim = c(0,8))
legend("topright", legend = c("Testes por Peso"), fill = c("skyblue"), bty = "n")
summary(peso)
freq_abs=hp$counts
text(locator(6), paste(round(freq_abs)))
```

Figura 13: script histograma "Distribuição dos testes por peso"

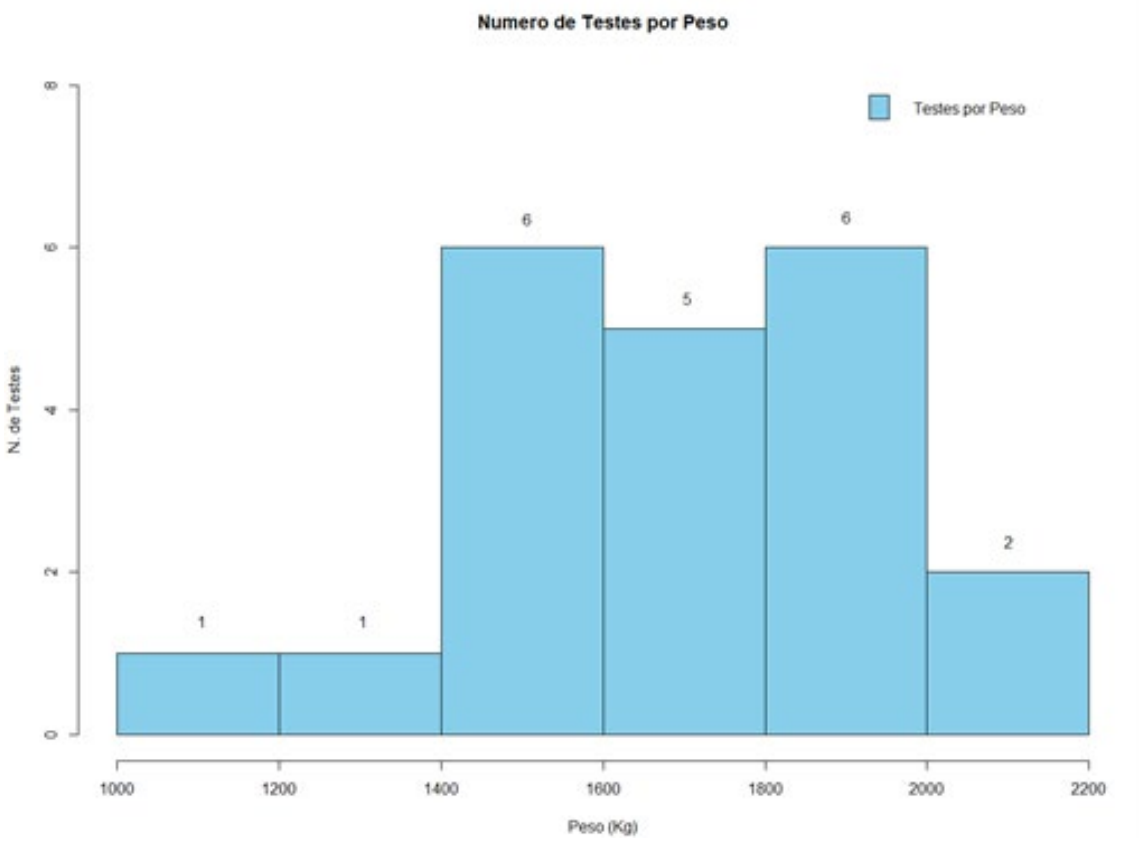


Figura 14: histograma "Distribuição dos testes por peso"

2.4 Identifique comparações entre os grupos A e B

Neste diagrama de extremos e quartis podemos observar a comparação da distância, com 10 litros de gasóleo, entre o condutor A e o condutor B.

O condutor A tem maior distribuição de dados (distância) do que o condutor B.

O condutor A tem o 3º quartil maior do que o 3º quartil do condutor B, enquanto o condutor A tem o 1º quartil menor do que o 1º quartil do condutor B.

O condutor A está no intervalo [175, 244] km, enquanto o condutor B está no intervalo [135, 190] km.

Como no condutor A, a mediana e a média têm aproximadamente o mesmo valor, então a sua simetria é aproximada.

Como no condutor B, a mediana e a média têm aproximadamente o mesmo valor, então a sua simetria é aproximada.

```r
boxplot(distancia ~ condutor, main = "Comparacao da Distancia por condutor", ylab="Distancia (Km)", xlab="Condutor", names=c("A","B"),
col=c("purple","skyblue"))
#segundo esta amostra o genero feminino tem o ordenado atual menor que dos homens
legend("topright", legend = c("Condutor A","Condutor B"), fill = c("purple","skyblue"), bty = "n")
IQR(distancia) #da intervalo interquartil (no ordenado atual, a diferenca entre o 1 qartil e o 3 quartil sao 798.85)
tapply(distancia,condutor,summary)# Para interpretar os valores do boxplot (qt + proximo a media e a mediana estao, mais normais estao
#(assimetria); o min da mulher e menos de metade do min do homem; 1 quartil e 3 quartil(50%);)
```

Figura 15: script do diagrama de extremos e quartis "Comparação da distância por condutor"

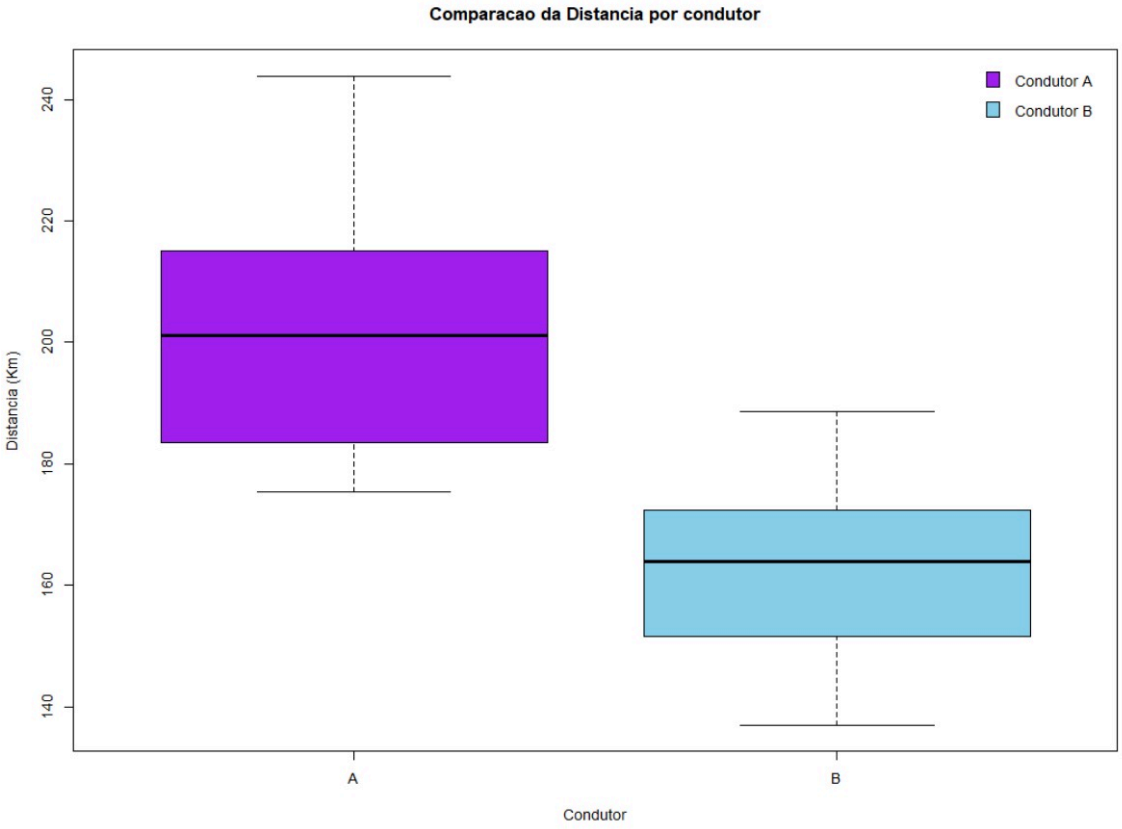


Figura 16: diagrama de extremos e quartis "Comparação da distância por condutor"

Neste diagrama de extremos e quartis podemos observar a comparação do peso entre o condutor A e o condutor B.

O condutor A tem maior distribuição de dados (peso) do que o condutor B.

O condutor A tem o 3º quartil maior do que o 3º quartil do condutor B.

O condutor A tem o 1º quartil maior do que o 1º quartil do condutor B.

O condutor A está no intervalo [1180, 1850] kg, enquanto o condutor B está no intervalo [1650, 2050] kg.

Como no condutor A, a mediana é menor que a média, então a sua assimetria é positiva.

Como no condutor B, a mediana é maior que a média, então a sua assimetria é negativa.

```
boxplot(peso ~ condutor ,main = "Comparacao do Peso do veiculo por condutor", ylab="Peso (Kg)", xlab="Condutor", names=c("A","B"),
col=c("purple","skyblue"), xlim=c(0,4))
legend("topright", legend = c("Condutor A","Condutor B"), fill = c("purple","skyblue"), bty = "n")

tapply(peso,condutor,summary)# Para interpretar os valores do boxplot (qt + proximo a media e a mediana estao, mais normais estao (assimetria);
```

Figura 17: script do diagrama de extremos e quartis "Comparação do peso por condutor"

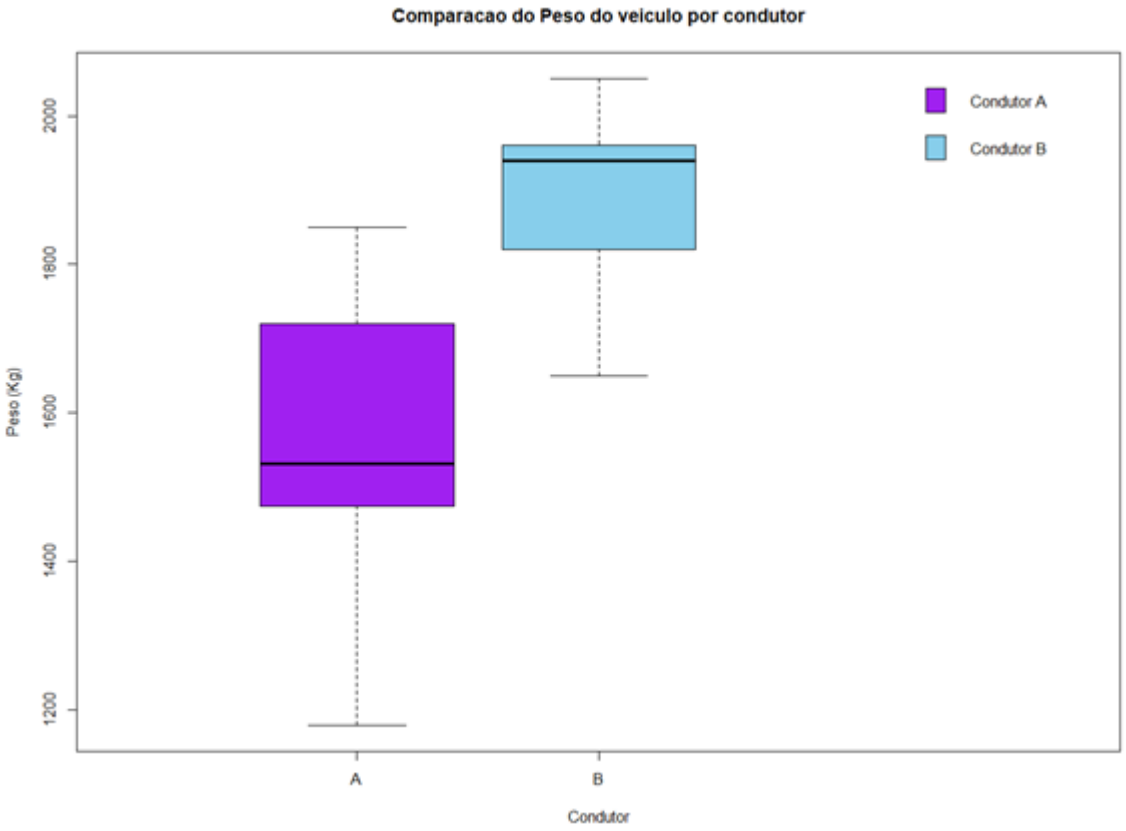


Figura 18:diagrama de extremos e quartis "Comparação do peso por condutor"

Neste diagrama de extremos e quartis podemos observar a comparação do consumo médio de litros aos 100km entre o condutor A e o condutor B.

O condutor B tem um "outlier" presente nos 11 litros/100km.

O condutor A tem menor distribuição de dados (consumo) do que o condutor B.

O condutor A tem o 3º quartil menor do que o 3º quartil do condutor B.

O condutor A tem o 1º quartil menor do que o 1º quartil do condutor B.

O condutor A está no intervalo [5, 8] litros/100km, enquanto o condutor B está no intervalo [6, 11] litros/100km.

Como no condutor A, a mediana é menor que a média, então a sua assimetria é positiva.

Como no condutor B, a mediana é menor que a média, então a sua assimetria é positiva.

```
boxplot(consumo ~ condutor, main = "Comparacao do consumo por condutor", ylab="consumo medio de litros aos 100km",
xlab="Condutor", names=c("condutor A","condutor B"),col=c("purple","skyblue"))
legend("topright", legend = c("Condutor A","Condutor B"), fill = c("purple","skyblue"), bty = "n")
IQR(consumo)
tapply(consumo,condutor,summary)
```

Figura 19:script do diagrama de extremos e quartis "Comparação do consumo por condutor"

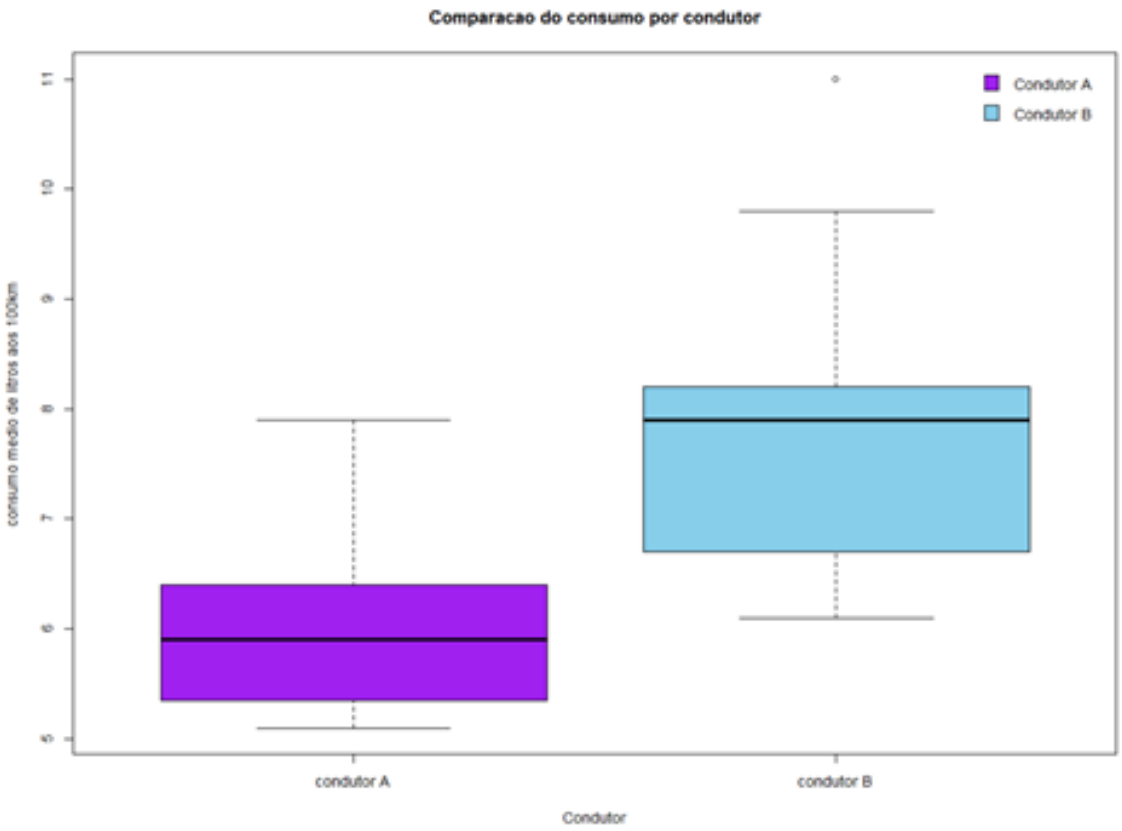


Figura 20: diagrama de extremos e quartis "Comparação do consumo por condutor"

2.5 Identifique e compare outros grupos

Neste diagrama de extremos e quartis podemos observar a comparação da distância, com 10 litros de gasóleo, entre a avaliação ("desempenho muito bom", "desempenho bom", "desempenho fraco" e "desempenho muito fraco").

A avaliação “desempenho muito fraco” tem maior distribuição de dados (distância), enquanto a avaliação “desempenho fraco” tem menor distribuição de dados.

A avaliação “desempenho fraco” tem o 3º quartil maior, enquanto a avaliação “desempenho muito fraco” tem o 3º quartil menor.

A avaliação “desempenho bom” tem o 1º quartil maior, enquanto a avaliação “desempenho muito fraco” tem o 1º quartil menor.

A avaliação “desempenho muito bom” está no intervalo [204, 244] km.

A avaliação “desempenho bom” está no intervalo [182, 213] km.

A avaliação “desempenho fraco” está no intervalo [172, 198] km.

A avaliação “desempenho muito fraco” está no intervalo [137, 167] km.

Como na avaliação “desempenho muito bom”, a mediana é menor que a média, então a sua assimetria é positiva.

Como na avaliação “desempenho bom”, a mediana e a média têm aproximadamente o mesmo valor, então a sua simetria é aproximada.

Como na avaliação “desempenho fraco”, a mediana é menor que a média, então a sua assimetria é positiva.

Como na avaliação “desempenho muito fraco”, a mediana é maior que a média, então a sua assimetria é negativa.

```
boxplot(distancia ~ avaliacao, main = "Comparacao da distancia por avaliacao", ylab="Distancia (Km)", xlab="Avaliacao",
names=c("1","2","3","4"),col=c("purple","skyblue","darkseagreen","coral"))
legend("topright", legend = c("1 - desempenho muito bom", "2 - desempenho bom", "3 - desempenho fraco","4 - desempenho muito fraco"),
fill = c("purple","skyblue","darkseagreen","coral"), bty = "n")
IQR(distancia)
tapply(distancia,avaliacao,summary)
```

Figura 21: script do diagrama de extremos e quartis “Comparação da distância, com 10 litros de gasóleo, por avaliação”

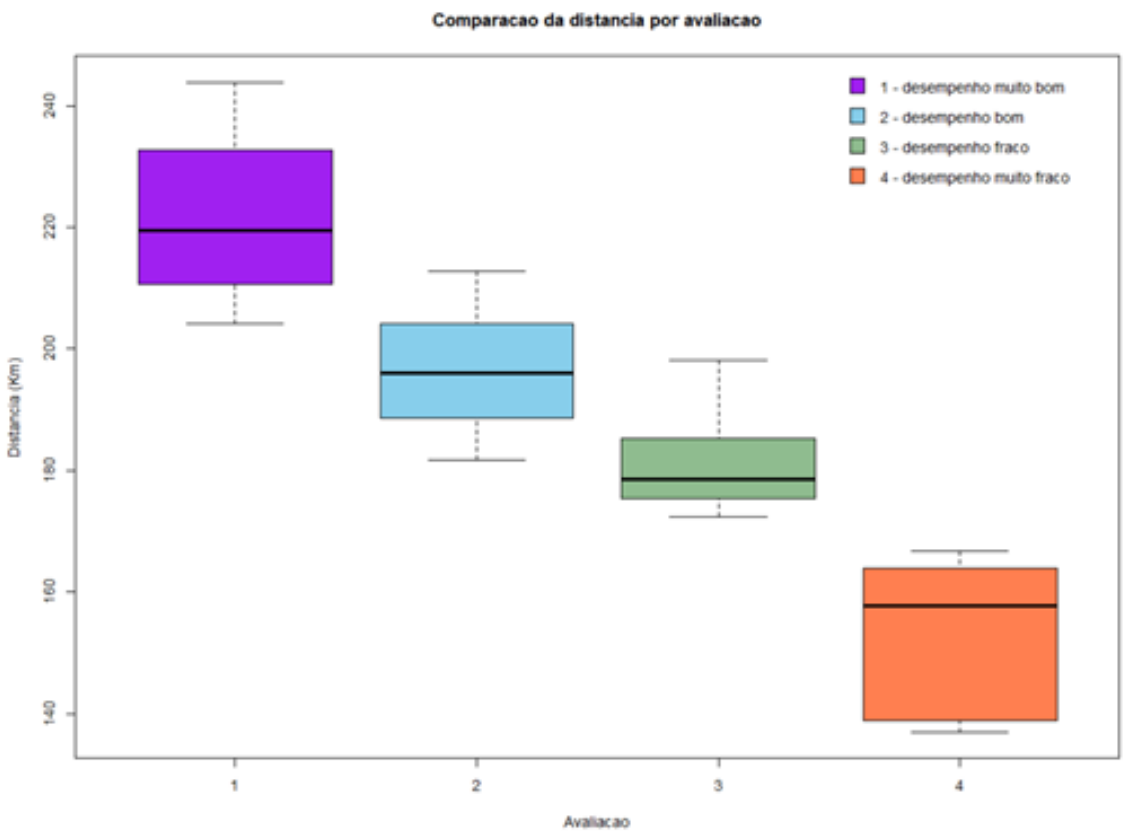


Figura 22: diagrama de extremos e quartis "Comparação da distância por avaliação"

Neste diagrama de extremos e quartis podemos observar a comparação do peso entre a avaliação ("desempenho muito bom", "desempenho bom", "desempenho fraco" e "desempenho muito fraco").

A avaliação "desempenho fraco" tem um "outlier" presente nos 1530kg.

A avaliação "desempenho muito bom" tem maior distribuição de dados (peso), enquanto a avaliação "desempenho muito fraco" tem menor distribuição de dados.

A avaliação "desempenho bom" tem o 3º quartil maior, enquanto a avaliação "desempenho muito bom" tem o 3º quartil menor.

A avaliação "desempenho bom" tem o 1º quartil maior, enquanto a avaliação "desempenho fraco" tem o 1º quartil menor.

A avaliação "desempenho muito bom" está no intervalo [1180, 1510] kg.

A avaliação "desempenho bom" está no intervalo [1380, 1730] kg.

A avaliação "desempenho fraco" está no intervalo [1533, 1850] kg.

A avaliação "desempenho muito fraco" está no intervalo [1897, 2050] kg.

Como na avaliação “desempenho muito bom”, a mediana é maior que a média, então a sua assimetria é negativa.

Como na avaliação “desempenho bom”, a mediana e a média têm aproximadamente o mesmo valor, então a sua simetria é aproximada.

Como na avaliação “desempenho fraco”, a mediana é maior que a média, então a sua assimetria é negativa.

Como na avaliação “desempenho muito fraco”, a mediana é menor que a média, então a sua assimetria é positiva.

```
boxplot(peso ~ avaliacao, main = "Comparacao do peso por avaliacao", ylab="Peso (Kg)", xlab="Avaliacao", names=c("1","2","3","4"),
col=c("purple","skyblue","darkseagreen","coral"))
legend("topleft", legend = c("1 - desempenho muito bom", "2 - desempenho bom", "3 - desempenho fraco","4 - desempenho muito fraco"),
 fill = c("purple","skyblue","darkseagreen","coral"), bty = "n")
IQR(peso)
tapply(peso,avaliacao,summary)
```

Figura 23: script do diagrama de extremos e quartis “Comparação do peso por avaliação”

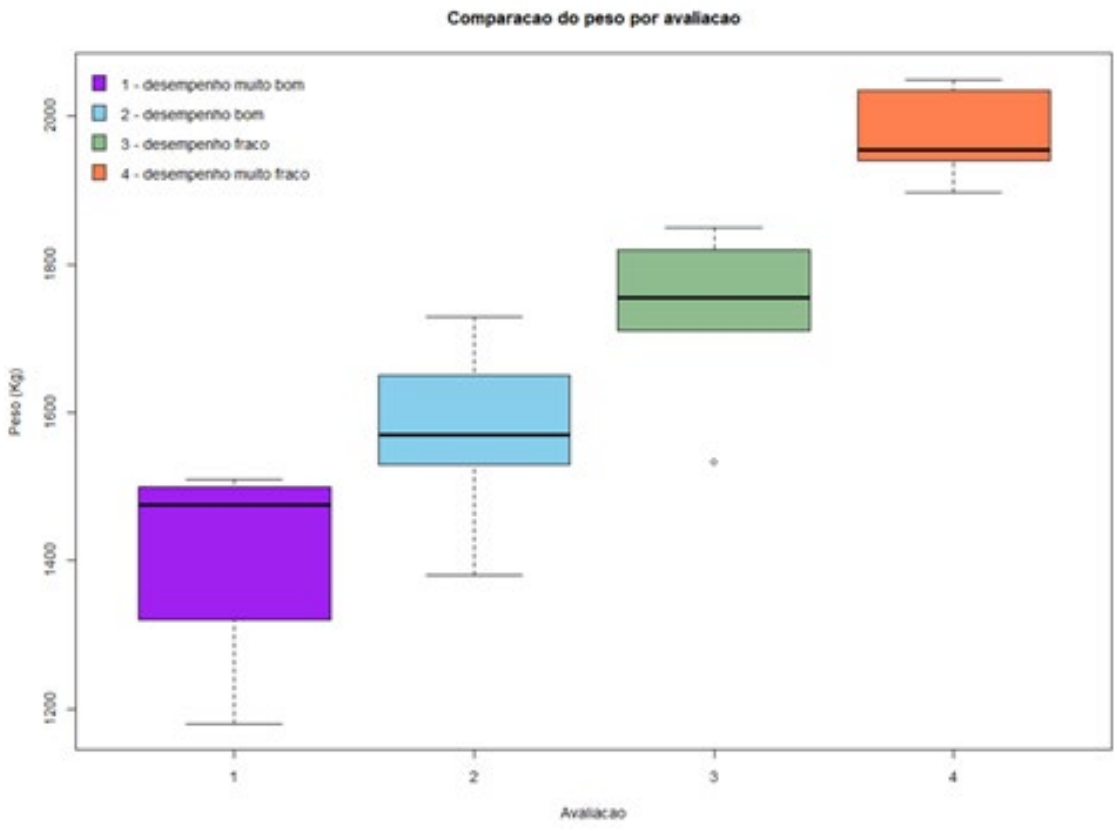


Figura 24: diagrama de extremos e quartis “Comparação do peso por avaliação”

Neste diagrama de extremos e quartis podemos observar a comparação do consumo médio de litros aos 100km entre a avaliação (“desempenho muito bom”, “desempenho bom”, “desempenho fraco” e “desempenho muito fraco”).

A avaliação “desempenho muito fraco” tem maior distribuição de dados (consumo), enquanto a avaliação “desempenho muito bom” tem menor distribuição de dados.

A avaliação “desempenho muito fraco” tem o 3º quartil maior, enquanto a avaliação “desempenho bom” tem o 3º quartil menor.

A avaliação “desempenho muito fraco” tem o 1º quartil maior, enquanto a avaliação “desempenho bom” tem o 1º quartil menor.

A avaliação “desempenho muito bom” está no intervalo [5,1; 5,4] litros/100km.

A avaliação “desempenho bom” está no intervalo [5,9; 6,3] litros/100km.

A avaliação “desempenho fraco” está no intervalo [5,9; 7,9] litros/100km.

A avaliação “desempenho muito fraco” está no intervalo [7,5; 11] litros/100km.

Como na avaliação “desempenho muito bom”, a mediana e a média têm o mesmo valor, então a sua simetria é pura.

Como na avaliação “desempenho bom”, a mediana e a média têm o mesmo valor, então a sua simetria é pura.

Como na avaliação “desempenho fraco”, a mediana é menor que a média, então a sua assimetria é positiva.

Como na avaliação “desempenho muito fraco”, a mediana é menor que a média, então a sua assimetria é positiva.

```
boxplot(consumo ~ avaliacao, main = "Comparacao do consumo por avaliacao", ylab="Consumo Medio de Litros aos 100km", xlab="Avaliacao",
 names=c("1","2","3","4"),col=c("purple","skyblue","darkseagreen","coral"))
legend("topleft", legend = c("1 - desempenho muito bom", "2 - desempenho bom", "3 - desempenho fraco","4 - desempenho muito fraco"),
 fill = c("purple","skyblue","darkseagreen","coral"), bty = "n")
IQR(consumo)
tapply(consumo,avaliacao,summary)
```

Figura 25: script do diagrama de extremos e quartis “Comparação do consumo por avaliação”

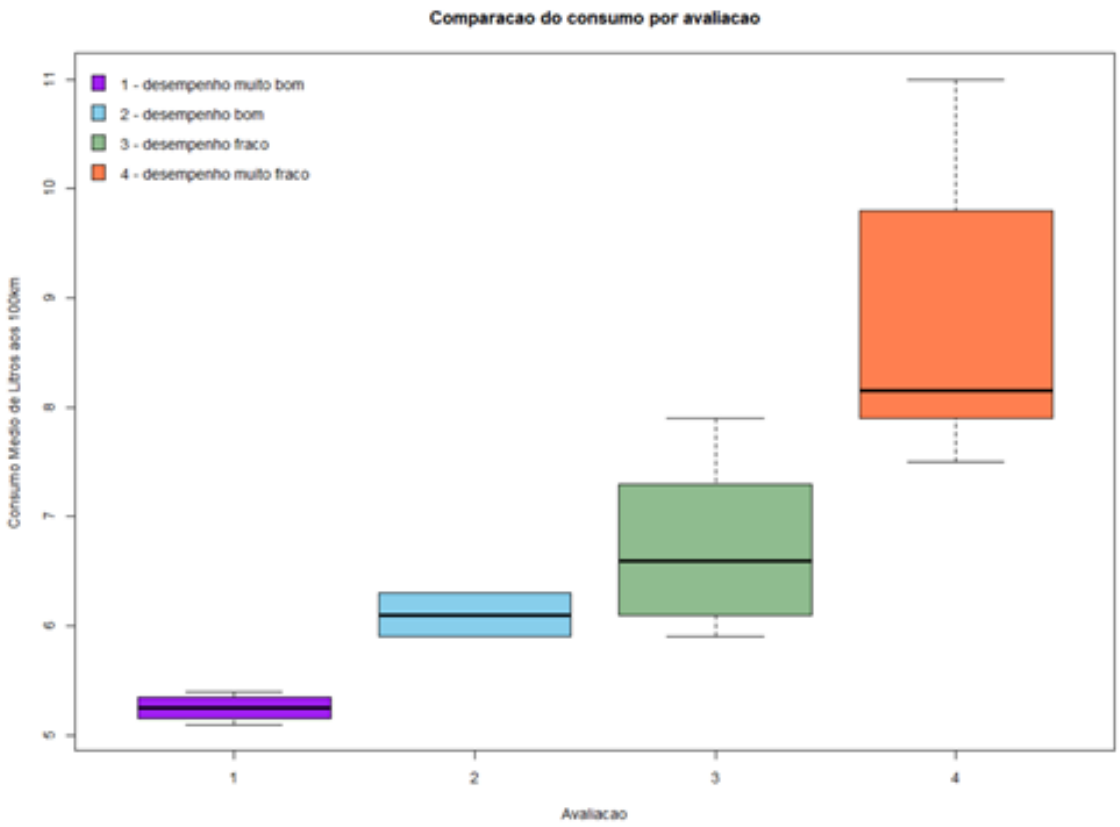


Figura 26: diagrama de extremos e quartis "Comparação do consumo por avaliação"

2.6 Regressão Linear

2.6.1 Represente o diagrama de dispersão entre as variáveis Peso e Distância. Conclua acerca da correlação aparente mediante a observação da orientação da nuvem de pontos. Verifique pelo valor do coeficiente de correlação.

Neste diagrama de dispersão podemos observar a regressão linear entre o Peso e a Distância.

Podemos concluir que, como o valor de "r" está muito longe de 1, por isso existe uma relação de correlação forte e negativa entre as variáveis "peso" e "distância".

```
plot(peso,distancia , pch = 19, col = "lightslateblue", main = "Distancia vs Peso",
 xlab = "Peso (Kg)", ylab = "Distancia (Km)")
cor(peso, distancia) # r= -0.9772903
cor(peso, distancia)^2
```

Figura 27: script do diagrama de dispersão "Peso vs Distância"

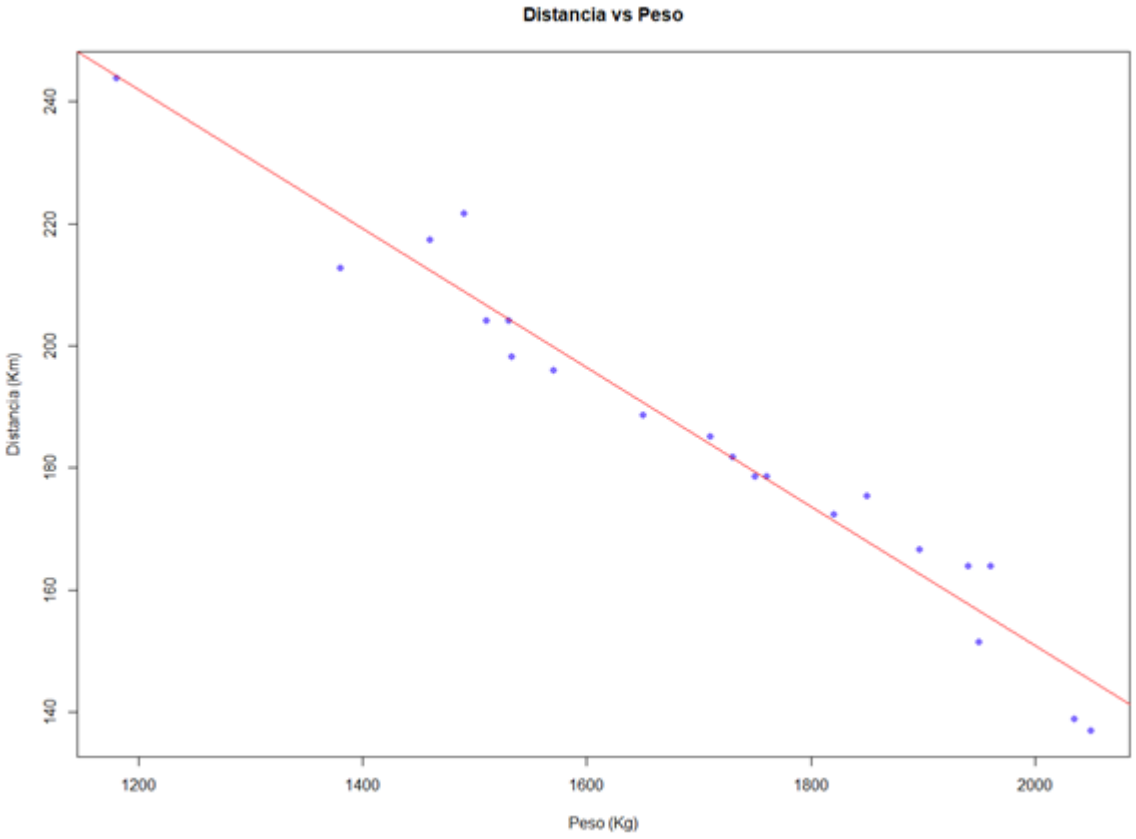


Figura 28: digrama de dispersão "Peso vs Distância"

2.6.2 Ajuste, pelo método dos mínimos quadrados, a reta de regressão e interprete os valores obtidos para os coeficientes;

Podemos concluir que, com o valor de “model”, se a distância for nula (zero), o peso é de 3253kg. E por cada 1km que se acrescenta á distância, retira 8kg no peso.

```
model = lm(peso ~ distancia) #linear simples #dependente~independente
model #peso=3253.986+-8.395*distancia
abline(model, col="red") #desenhar a reta de regressao linear simples
summary(model) #sumario com as estimativas dos coeficientes
```

Figura 29: script da reta de regressão linear

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;

- o “intercept” (=0.0000000000000002) tem um significado porque é diferente de zero;
- a “distancia” (=0.000000000000029) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\,(= 3253.986) + B1\,(= -8.395) * X\,(= distancia) + erro$

```
> summary(model) # sumario com as estimativas dos coeficientes,

Call:
lm(formula = peso ~ distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-87.610 -32.200  -4.706  31.006  97.103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3253.9856    77.9632   41.74  < 2e-16 ***
distancia     -8.3946     0.4176  -20.10  2.9e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.8 on 19 degrees of freedom
Multiple R-squared:  0.9551,    Adjusted R-squared:  0.9527
F-statistic: 404.1 on 1 and 19 DF,  p-value: 2.901e-14
```

Figura 30: console do "summary(model)"

2.6.3 Interprete a qualidade do ajustamento por referência ao coeficiente de determinação.

Podemos concluir que, com o valor de “r^2”, que 95,50% da variância do peso é explicada pela variância da distância.

```
cor(peso,distancia)^2  # r^2 = 0.9550963
```

Figura 31: script do coeficiente de determinação

Podemos estimar com este cálculo que, um automóvel, com 10 litros de gasóleo, se percorresse 100km, o peso do automóvel seria de 2414.486kg.

```
3253.986+-8.395*100 # = 2414.486
```

Figura 32: cálculos estimativas

Podemos estimar com este cálculo que, um automóvel, com 10 litros de gasóleo, se percorresse 200km, o peso do automóvel seria de 1574.986kg.

```
3253.986+-8.395*200 # = 1574.986
```

Figura 33: cálculos estimativas

Podemos estimar com este cálculo que, se o peso do automóvel fosse 100kg, o automóvel, com 10 litros de gasóleo, percorria 375.6982km.

```
(100-3253.986)/-8.395 # = 375.6982
```

Figura 34: cálculos estimativas

2.6.4 Analise outras correlações.

Neste diagrama de dispersão podemos observar a regressão linear entre a Distância e o Consumo.

Podemos concluir que, como o valor de “r” está muito longe de 1, por isso existe uma relação de correlação moderada e negativa entre as variáveis “distancia” e “consumo”.

```
plot(distancia,consumo , pch = 19, col = "lightslateblue", main = "Distancia vs Consumo",
xlab = "Distancia (Km)",ylab = "Consumo Medio (l/100Km)")
cor(distancia, consumo) # r= -0.8343691
cor(distancia, consumo)^2 # r^2 = 0.6961718
```

Figura 35: script do diagrama de dispersão "Distância vs Consumo"

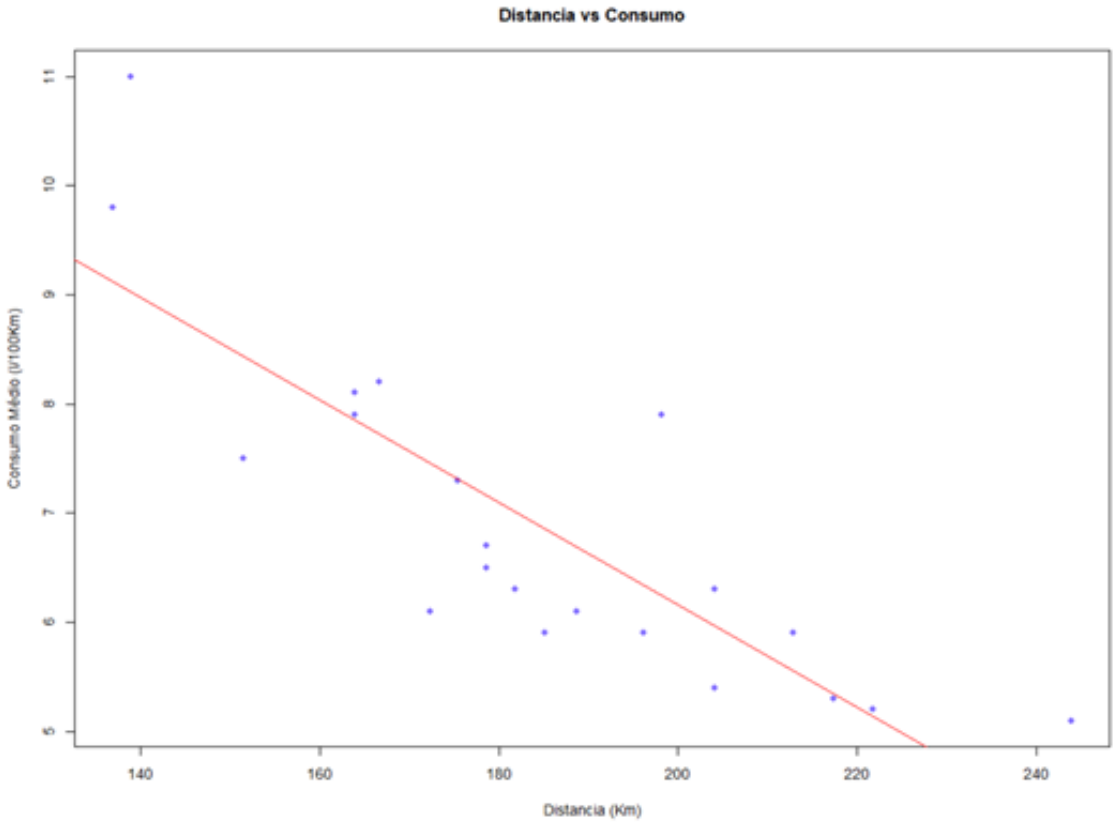


Figura 36: digrama de dispersão "Distância vs Consumo"

Podemos concluir que, com o valor de “model”, se o consumo for nulo (zero), a distância é de 286.61km. E por cada 1 litro/100km que se acrescenta ao consumo, retira 14.81km na distância.

```
model = lm(distancia ~ consumo) #linear simples #dependente~independente
model #distancia=286.61+-14.81*consumo
abline(model, col="red") #desenhar a reta de regressao linear simples
summary(model) #sumario com as estimativas dos coeficientes
```

Figura 37: script da reta de regressão linear

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;
- o “intercept” (=0.000000000000184) tem um significado porque é diferente de zero;

- o “consumo” (=0.00000258) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\ (= 286.61) + B1\ (= -14.81) * X\ (= consumo) + erro$

```
> summary(model)

Call:
lm(formula = distancia ~ consumo)

Residuals:
    Min      1Q  Median      3Q     Max
-24.069  -8.813  -3.130  10.764  32.798

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  286.611     15.791  18.151 1.84e-13 ***
consumo      -14.806      2.244  -6.598 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 19 degrees of freedom
Multiple R-squared:  0.6962,    Adjusted R-squared:  0.6802
F-statistic: 43.54 on 1 and 19 DF,  p-value: 2.581e-06
```

Figura 38: console do "summary(model)"

Podemos concluir que, com o valor de “r^2”, que 69,61% da variância da distância é explicada pela variância do consumo.

```
cor(consumo,distancia)^2  # r^2 = 0.6961718
```

Figura 39: script do coeficiente de determinação

Podemos estimar com este cálculo que, um automóvel que consumisse em média 5 litros/100km, percorria, com 10 litros de gasóleo, 212.56km.

```
286.61+-14.81*5 # = 212.56
```

Figura 40: cálculos estimativas

Podemos estimar com este cálculo que, um automóvel que consumisse em média 8 litros/100km, percorria, com 10 litros de gasóleo, 168.13km.

```
286.61+-14.81*8 # = 168.13
```

Figura 41: cálculos estimativas

Podemos estimar com este cálculo que, um automóvel que consumisse em média 19.01485 litros/100km, percorria, com 10 litros de gasóleo, 5km.

```
(5-286.61)/-14.81 # = 19.01485
```

Figura 42: cálculos estimativas

Neste diagrama de dispersão podemos observar a regressão linear entre o Peso e o Consumo.

Podemos concluir que, como o valor de "r" está longe de 1, por isso existe uma relação de correlação moderada e positiva entre as variáveis "peso" e "consumo".

```
plot(peso,consumo , pch = 19, col = "lightslateblue", main = "Peso vs Consumo",
xlab = "Peso (Kg)", ylab = "Consumo Medio (l/100Km)")
cor(peso, consumo) # r= -0.8343691
cor(peso, consumo)^2 # r^2 = 0.6961718
```

Figura 43: script do diagrama de dispersão "Peso vs Consumo"

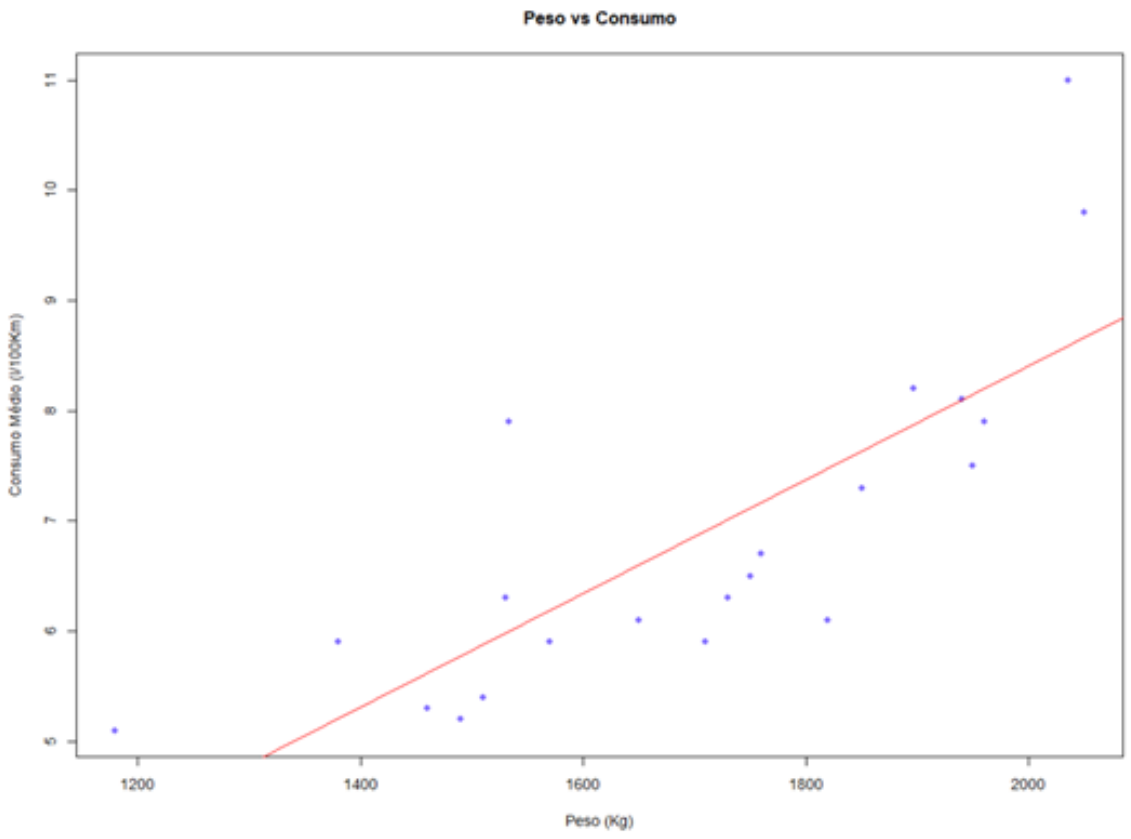


Figura 44: digrama de dispersão "Peso vs Consumo"

Podemos concluir que, com o valor de “model”, se o consumo for nulo (zero), o peso é de 878.2kg. E por cada 1 litro/100km que se acrescenta ao consumo, acrescenta 119.9kg no peso.

```
model = lm(peso ~ consumo) #linear simples #dependente~independente
model #peso=878.2+119.9 * consumo
abline(model, col="red") #desenhar a reta de regressao linear simples
summary(model) #sumario com as estimativas dos coeficientes
```

Figura 45: script da reta de regressão linear

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;
- o “intercept” (=0.0000144) tem um significado porque é diferente de zero;
- o “consumo” (=0.0000235) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\ (= 878.2) + B1\ (= 119.9) * X\ (= consumo) + erro$

```
> summary(model)

Call:
lm(formula = peso ~ consumo)

Residuals:
    Min      1Q  Median      3Q     Max
-309.66  -53.64   35.66   96.46  210.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   878.19     151.94   5.780 1.44e-05 ***
consumo       119.90      21.59   5.553 2.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148 on 19 degrees of freedom
Multiple R-squared:  0.6188,    Adjusted R-squared:  0.5987
F-statistic: 30.84 on 1 and 19 DF,  p-value: 2.347e-05
```

Figura 46: console do "summary(model)"

Podemos concluir que, com o valor de “r^2”, que 61,87% da variância do peso é explicada pela variância do consumo.

```
cor(consumo,peso)^2  # r^2 = 0.6187556
```

Figura 47: script do coeficiente de determinação

Podemos estimar com este cálculo que, um automóvel que consumisse em média 5 litros/100km, pesava 1477.77kg.

```
878.2+119.9 *5 # = 1477.77
```

Figura 48: cálculos estimativas

Podemos estimar com este cálculo que, um automóvel que consumisse em média 8 litros/100km, pesava 1837.4kg.

```
878.2+119.9 *8 # = 1837.4
```

Figura 49: cálculos estimativas

Podemos estimar com este cálculo que, um automóvel que consumisse em média -7.282736 litros/100km, pesava 5kg.

```
(5-878.2)/119.9 # = -7.282736
```

Figura 50: cálculos estimativas

3 Inferência Estatística

3.1 Estimação

3.1.1 Sabe-se que a proporção de testes a automóveis executados pelo condutor B é de 36%. Calcule a probabilidade de se considerarem 8 classificações B numa amostra de 21 testes por recurso à distribuição Binomial. Conclua acerca da representatividade da amostra em estudo face à proporção dos condutores

Podemos concluir que há 17,35% de probalidade de considerarem 8 condutores B numa amostra de 21 testes.

```
dbinom (8, 21, 0.36) #P(condutor B = 8 em 21, 0.36) = 17.35%
```

Figura 51: cálculo da proporção

3.1.2 Teste a Normalidade das variáveis;

Como p-value de “peso” é 0.6065, e é menor que 0.05, então não rejeita-se a hipótese nula.

```
shapiro.test(peso) # = 0.6065
```

Figura 52: teste a normalidade da variável "peso"

Como p-value de “distancia” é 0.9803, e é menor que 0.05, então não rejeita-se a hipótese nula.

```
shapiro.test(distancia) # = 0.9803
```

Figura 53: teste a normalidade da variável "distância"

Como p-value de “consumo” é 0.02075, e é maior que 0.05, então rejeita-se a hipótese nula.

```
shapiro.test(consumo) # = 0.02075
```

Figura 54: teste a normalidade da variável "consumo"

3.1.3 O gestor considera que o peso médio dos automóveis desta indústria é de 1700 Kg. Verifique se o gestor tem razão.

Como 1700 é menor que 1808.977, então o gestor tem razão.

```
t.test(peso,
       alternative="two.sided",
       mu=1700, #media do peso
       conf.level=0.95)
```

Figura 55: t-test peso

```
        One Sample t-test

data:  peso
t = 0.051367, df = 20, p-value = 0.9595
alternative hypothesis: true mean is not equal to 1700
95 percent confidence interval:
 1596.261 1808.977
sample estimates:
mean of x
 1702.619
```

Figura 56: console t-test peso

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

Como 180 é menor que 195.0425, então está dentro do intervalo de confiança.

```
t.test(distancia,
       alternative="two.sided",
       mu=180, #media da distancia
       conf.level=0.90)
```

Figura 57: t-test distancia

```
data:  distancia
t = 0.80944, df = 20, p-value = 0.4278
alternative hypothesis: true mean is not equal to 180
90 percent confidence interval:
 174.5670 195.0425
sample estimates:
mean of x
 184.8048
```

Figura 58: console t-test distancia

Como 6.3 é menor que 7.453134, então está dentro do intervalo de confiança.

```
t.test(consumo,
       alternative="two.sided",
       mu=6.3, #media do consumo
       conf.level=0.90)
```

Figura 59: t-test consumo

```
        One Sample t-test

data:  consumo
t = 1.7225, df = 20, p-value = 0.1004
alternative hypothesis: true mean is not equal to 6.3
90 percent confidence interval:
 6.299247 7.453134
sample estimates:
mean of x
  6.87619
```

Figura 60: console t-test consumo

3.1.5 Compare médias entre grupos.

Como p-value é 0.0001381, e é menor que 0.05, então rejeita-se a hipótese nula. A diferença entre as médias na população é positiva (201.6083 - 162.4000 = 39.2083), então é o grupo 0 (condutor A) com maior distância.

```r
t.test(distancia ~ condutor,
       alternative="two.sided",
       conf.level=0.95)
```

Figura 61: t-test distancia e condutor

```
        Welch Two Sample t-test

data:  distancia by condutor
t = 4.7785, df = 18.597, p-value = 0.0001381
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 22.00945 56.40722
sample estimates:
mean in group 0 mean in group 1
       201.6083        162.4000
```

Figura 62: console t-test distancia e condutor

Como p-value é 0.0000913, e é menor que 0.05, então rejeita-se a hipótese nula. A diferença entre as médias na população é negativa (1557.750 - 1895.778 = -338.028), então é o grupo 1 (condutor B) com maior peso.

```r
t.test(peso ~ condutor,
       alternative="two.sided",
       conf.level=0.95)
```

Figura 63: t-test peso e condutor

```
        Welch Two Sample t-test

data:  peso by condutor
t = -4.9393, df = 18.981, p-value = 9.13e-05
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -481.2754 -194.7801
sample estimates:
mean in group 0 mean in group 1
       1557.750        1895.778
```

Figura 64: console t-test peso e condutor

Como p-value é 0.01, e é menor que 0.05, então rejeita-se a hipótese nula. A diferença entre as médias na população é negativa (6.083333 - 7.933333 = -1.85), então é o grupo 1 (condutor B) com maior consumo.

```
t.test(consumo ~ condutor,
       alternative="two.sided",
       conf.level=0.95)
```

Figura 65: t-test consumo e condutor

```
        Welch Two Sample t-test

data:  consumo by condutor
t = -3.0922, df = 11.243, p-value = 0.01
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -3.1633278 -0.5366722
sample estimates:
mean in group 0 mean in group 1
       6.083333        7.933333
```

Figura 66: console t-test consumo e condutor

3.2 Regressão Linear

3.2.1 Teste a significância dos coeficientes de regressão e de determinação

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;
- o “intercept” (=0.0000144) tem um significado porque é diferente de zero;

- o “consumo” (=0.0000235) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\ (= 878.2) + B1\ (= 119.9) * X\ (= consumo) + erro$

```
> summary(model)

Call:
lm(formula = peso ~ consumo)

Residuals:
    Min      1Q  Median      3Q     Max
-309.66  -53.64   35.66   96.46  210.44

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   878.19     151.94   5.780 1.44e-05 ***
consumo       119.90      21.59   5.553 2.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 148 on 19 degrees of freedom
Multiple R-squared:  0.6188,    Adjusted R-squared:  0.5987
F-statistic: 30.84 on 1 and 19 DF,  p-value: 2.347e-05
```

Figura 67: console summary(model)

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;
- o “intercept” (=0.0000000000000002) tem um significado porque é diferente de zero;
- a “distancia” (=0.000000000000029) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\ (= 3253.986) + B1\ (= -8.395) * X\ (= distancia) + erro$

```
> summary(model)

Call:
lm(formula = peso ~ distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-87.610 -32.200  -4.706  31.006  97.103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3253.9856    77.9632   41.74  < 2e-16 ***
distancia     -8.3946     0.4176  -20.10  2.9e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.8 on 19 degrees of freedom
Multiple R-squared:  0.9551,    Adjusted R-squared:  0.9527
F-statistic: 404.1 on 1 and 19 DF,  p-value: 2.901e-14
```

Figura 68: console summary(model)

Podemos concluir que, com o valor de “summary(model)”:

- “t value” (p-value) rejeita ou não rejeita, dependendo do valor de alpha; se for menor rejeita-se a hipótese nula;
- o “intercept” (=0.000000000000184) tem um significado porque é diferente de zero;
- o “consumo” (=0.00000258) é menor que 0.05, então rejeita-se a hipótese nula;
- “residuals” são os erros do modelo do $y = B0\ (= 286.61) + B1\ (= -14.81) * X\ (= consumo) + erro$

```
> summary(model)

Call:
lm(formula = distancia ~ consumo)

Residuals:
    Min      1Q  Median      3Q     Max
-24.069  -8.813  -3.130  10.764  32.798

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  286.611     15.791  18.151 1.84e-13 ***
consumo      -14.806      2.244  -6.598 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 19 degrees of freedom
Multiple R-squared:  0.6962,    Adjusted R-squared:  0.6802
F-statistic: 43.54 on 1 and 19 DF,  p-value: 2.581e-06
```

Figura 69: console summary(model)

3.2.2 Teste a Normalidade e a média dos erros do modelo. Conclua acerca deste pressuposto do modelo de regressão

Como p-value dos erros do “model” é 0.1006, e é maior que 0.05, então rejeita-se a hipótese nula.

```
> model = lm(peso ~ consumo)
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.9232, p-value = 0.1006
```

Figura 70: shapiro test

Como p-value é 1, e é maior que 0.05, então não rejeita-se a hipótese nula.

```
> t.test(model$residuals,
+         alternative="two.sided",
+         mu=0,
+         conf.level=0.95)

         One Sample t-test

data:  model$residuals
t = -9.6634e-17, df = 20, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -65.67061  65.67061
sample estimates:
    mean of x
-3.042251e-15
```

Figura 71: t test

Como p-value dos erros do “model” é 0.7301, e é maior que 0.05, então rejeita-se a hipótese nula.

```
> model = lm(peso ~ distancia)
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.96987, p-value = 0.7301
```

Figura 72: shapiro test

Como p-value é 1, e é maior que 0.05, então não rejeita-se a hipótese nula.

```
> t.test(model$residuals,
+        alternative="two.sided",
+        mu=0,
+        conf.level=0.95)

        One Sample t-test

data:  model$residuals
t = -3.1704e-16, df = 20, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -22.53773  22.53773
sample estimates:
    mean of x
-3.425501e-15
```

Figura 73: t test~

Como p-value dos erros do “model” é 0.3575, e é maior que 0.05, então rejeita-se a hipótese nula.

```
> model = lm(distancia ~ consumo)
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.95111, p-value = 0.3575
```

Figura 74: shapiro test

Como p-value é 1, e é maior que 0.05, então não rejeita-se a hipótese nula.

```
> t.test(model$residuals,
+        alternative="two.sided",
+        mu=0,
+        conf.level=0.95)

        One Sample t-test

data:  model$residuals
t = -9.3825e-17, df = 20, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -6.825048  6.825048
sample estimates:
    mean of x
-3.069841e-16
```

Figura 75: t test

3.2.3 Sabe-se que o Peso de um determinado automóvel é de 1798 kg. Com base na reta ajustada preveja o valor da Distância percorrida com 10 litros de combustível.

```
model = lm(distancia ~ peso) #linear simples #dependente~independente
predict(model)
model #ver modelo;peso=378.5198+-0.1138*distancia
abline(model, col="red") #desenhar a reta de regressao linear simples
```

Figura 76: script previsao

Podemos concluir que com 10 litros de gasóleo e com 1798kg, o automóvel percorria entre o intervalo [160, 180] km.

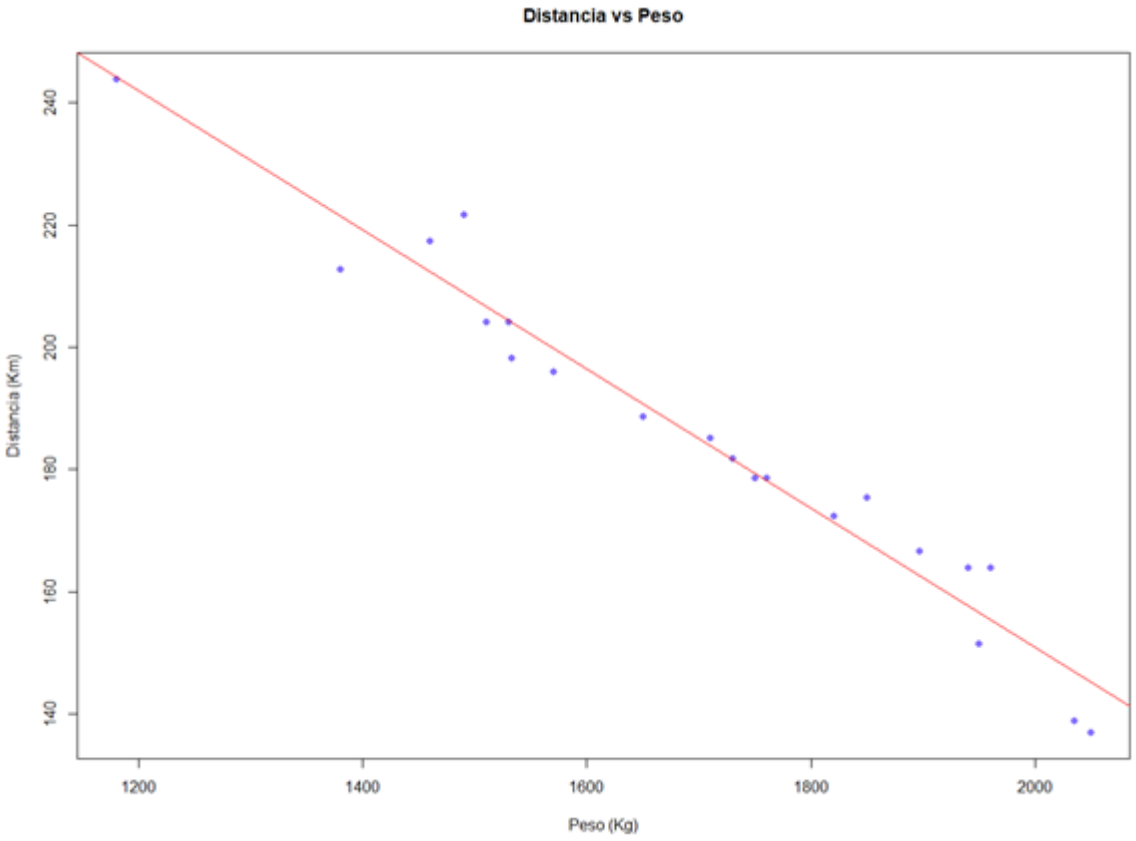


Figura 77: diagrama de dispersão "peso vs distância"

3.2.4 Preveja o valor que o Peso de um automóvel não deverá atingir para que a distância percorrida com 10 litros de gasóleo seja superior a 200km

Podemos concluir que com 10 litros de gasóleo e percorrer mais de 200km, o automóvel não poderia pesar entre o intervalo [1500, 2000[kg.

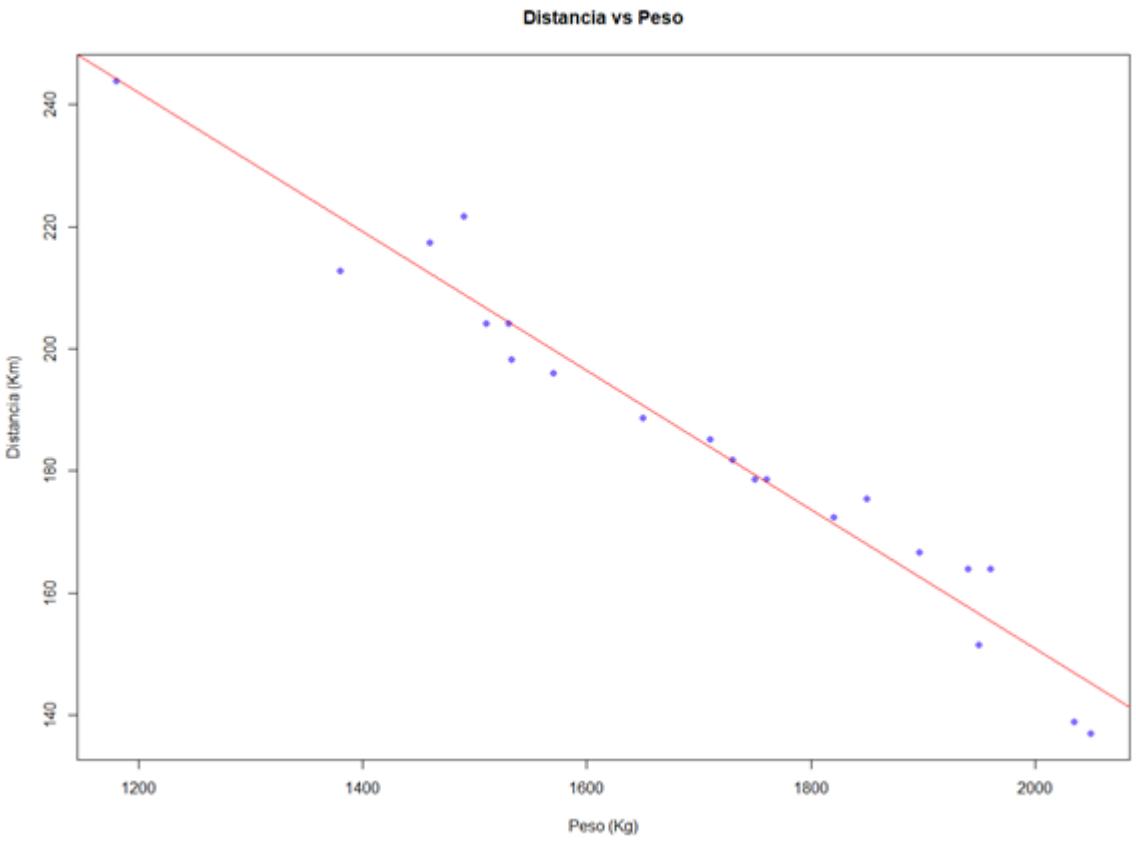


Figura 78: diagrama de dispersão "peso vs distância"

3.2.5 Faça previsão de valores “in the box” e “out of the box” de outras regressões

Podemos concluir que com o consumo médio de 5 litros/100km, o automóvel percorria, com 10 litros de gasóleo, entre o intervalo [220, 240[km.

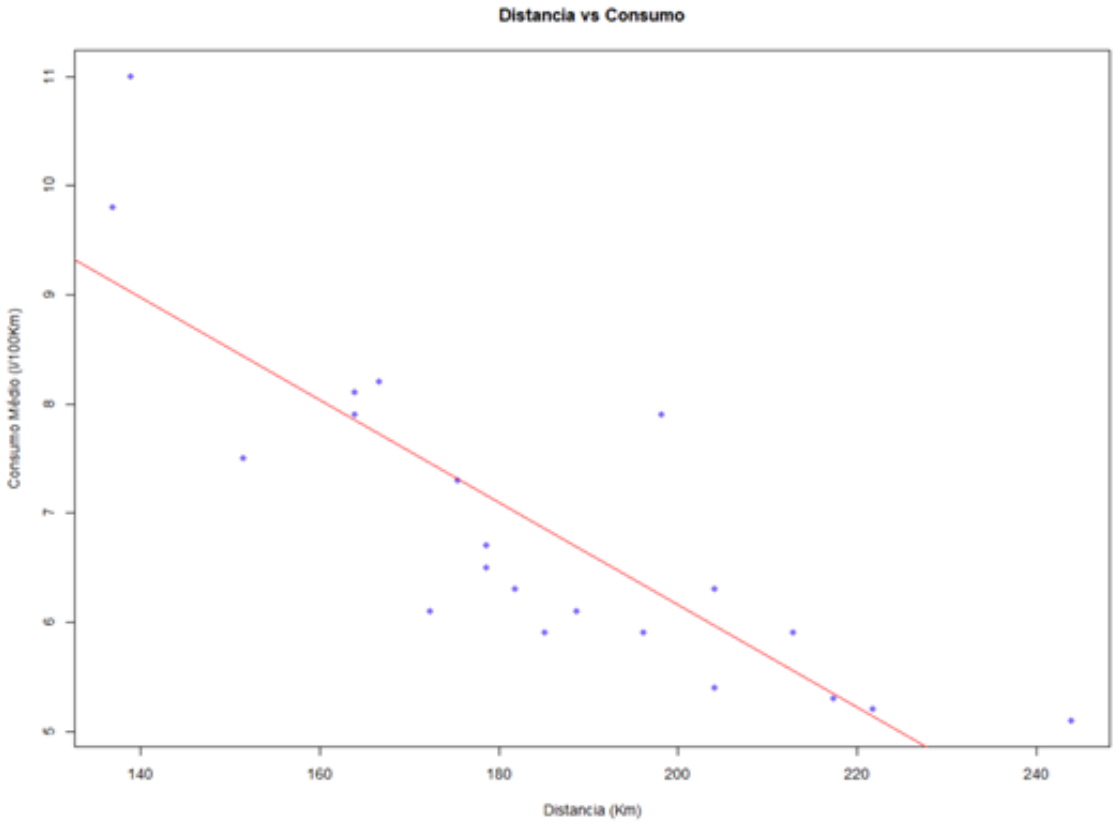


Figura 79: diagrama de dispersão "distância vs consumo"

Podemos concluir que com o consumo médio de 5 litros/100km, o automóvel pesaria entre o intervalo]1200, 1500] kg.

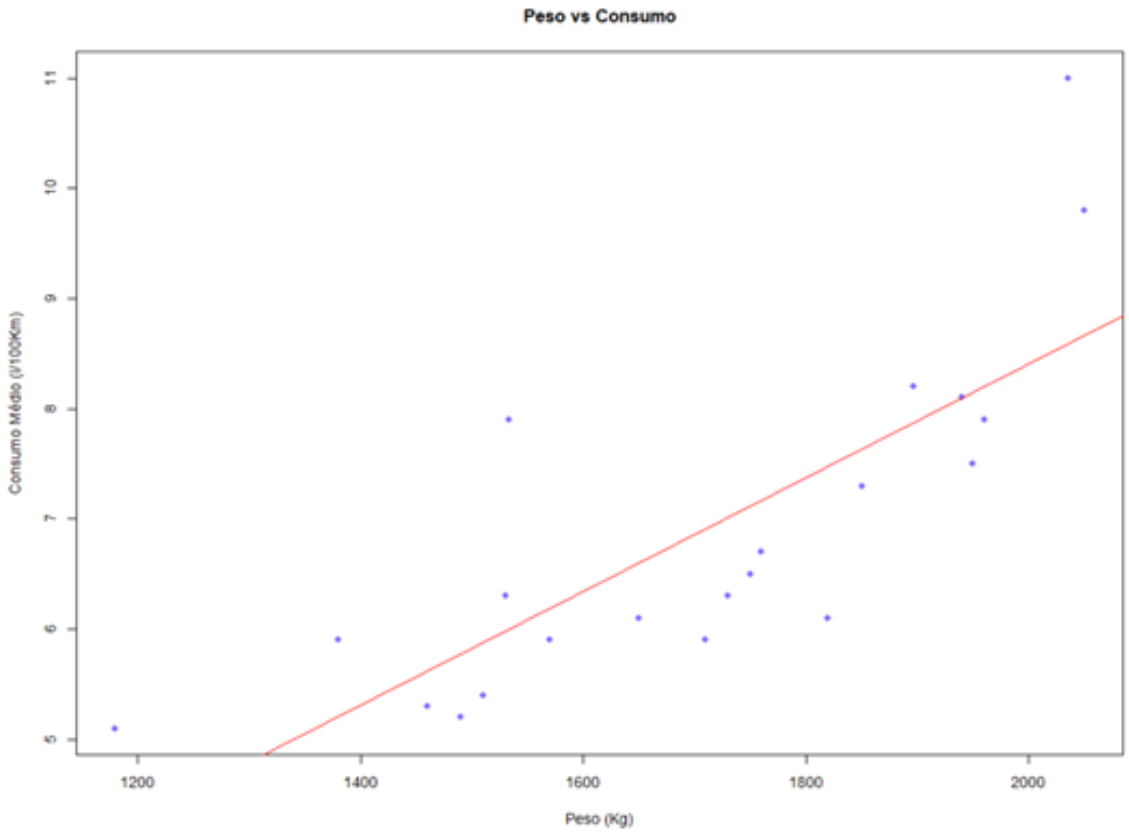


Figura 80: diagrama de dispersão "peso vs consumo"

4 Conclusão

Podemos concluir com este trabalho que, obtivemos conhecimento acerca do R Studio. Onde tivemos, por exemplo, de comparar diversas variáveis.

Todos os objetivos que indicámos na Introdução, foram realizados.

Concluímos também, que o conhecimento que adquirimos, é a base de grandes profissões como “Big Data Engineer”.

Índice

5 Índice de Imagens