|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| UPE_13 | UNIVERSIDADE DE PERNAMBUCO  **Escola Politécnica de Pernambuco**  **Coordenação Setorial de Pós-graduação e Pesquisa** | |  |  | | --- | --- | |  |  | |

**LAÍS INGRID SOARES DE SOUZA VIDOTO**

**FATORES QUE INFLUENCIARAM O CONSUMO DE COMBUSTÍVEL**

Atividade da disciplina de Estatística Computacional, apresentada ao Curso de Pós-graduação em Ciência de Dados e Analytics, da Escola Politécnica de Pernambuco da Universidade de Pernambuco.

Profa. Dra. Roberta Andrade de Araújo Fagundes

Recife, PE

2019

**Resumo**

Foi solicitado que escolhêssemos uma base de dados, e que por sua vez, fosse respondido algumas questões referente a base de dados escolhida utilizando RStudio, questões que pelos quais são referente a estatística computacional, e que serão respondidas ao longo das páginas como foi solicitado.

**Descrição da base:** é uma base responsável por realizar uma análise dos carros, em relação ao seu consumo de gasolina.

**Atributos de sua base:**

|  |  |  |  |
| --- | --- | --- | --- |
| Nome campo | Tipo | Descrição | Utilização |
| Mpg | contínuo | Consumo de combustível do ciclo urbano em milhas por galão | Será considerada |
| Cilindros | discreto com vários valores | Força mecânica que ajuda o deslocamento do carro | Será considerada |
| Deslocamento | contínuo | Mudança de um lugar para outro | Será considerada |
| **cavalo-vapor** | **contínuo** | **Força do motor** | **Não será considerada** |
| Peso | contínuo | Peso total do carro | Será considerada |
| Aceleração | contínuo | Velocidade máxima do carro | Será considerada |
| Ano modelo | discreto com vários valores | Ano de fabricação do modelo do carro | Será considerada |
| Origem | discreto com vários valores | Mesma montadora de origem | Será considerada |
| **Nome do carro** | **string** | **Como o carro será chamado** | **Não será considerada** |

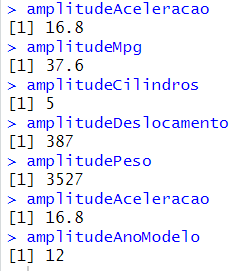
Analise realizada na base de dados, escolhida abaixo:

http://archive.ics.uci.edu/ml/datasets/Auto+MPG

**Conclusões**

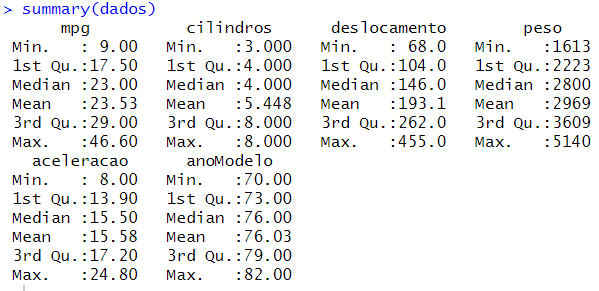
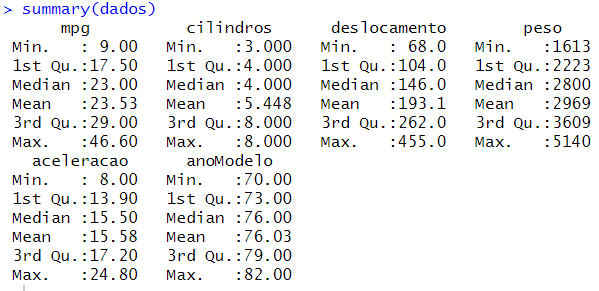
* **Identificar a amplitude de cada uma das variáveis**

Abaixo está a diferença entre o maior e o menor valor de cada campo para mostra a dispersão deles.



* **Calcular média, moda e mediana para cada uma das variáveis numéricas**

Com essa informação, podemos saber ponto médio de cada campo

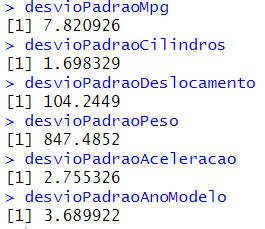


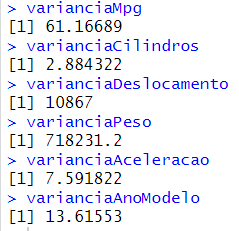
* **Calcular variância e desvio padrão para cada uma das variáveis**

Variância->valor que mostra o quanto esse valor está distante da media. Quanto maior for esse valor, mais distante da media e quanto menor, mais próximo o valor esta da media.

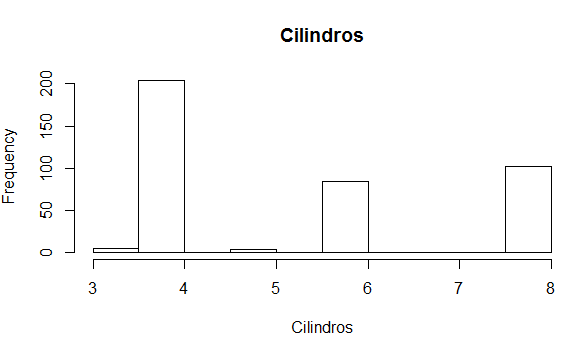
Desvio Padrão-> quanto for menor tem mais próximo estará do valor médio e o esperado

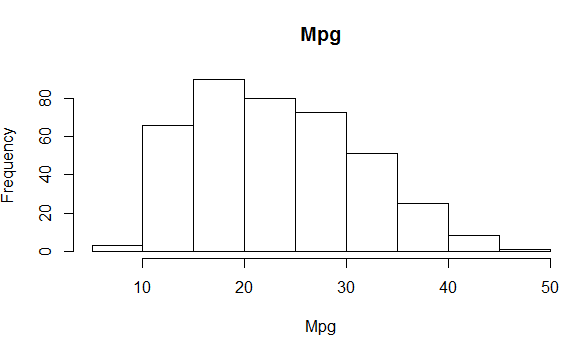
quanto for maior tem mais dados espalhados com uma grande variação de valores

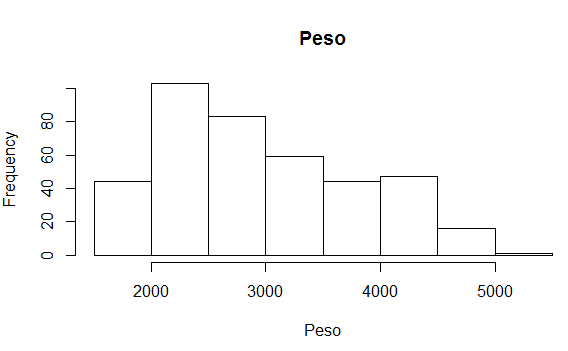


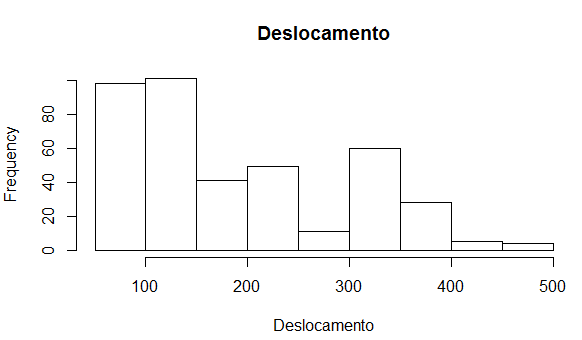


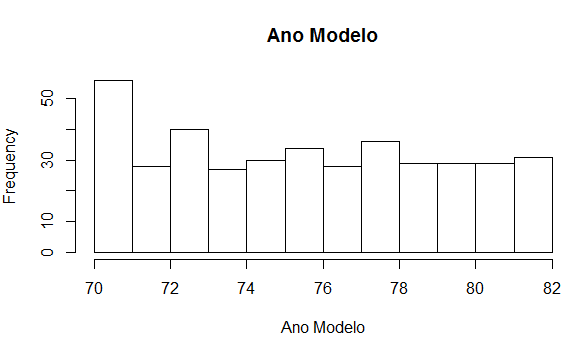
* **Plotar o melhor gráfico para cada uma das variáveis;**

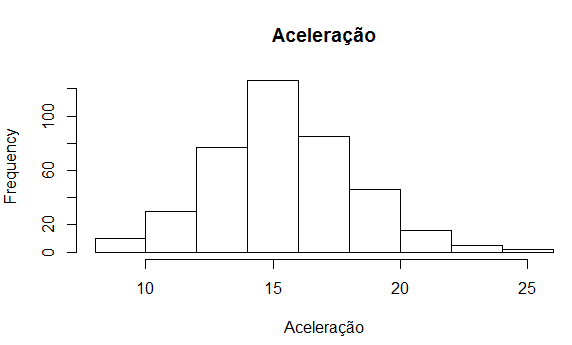
****



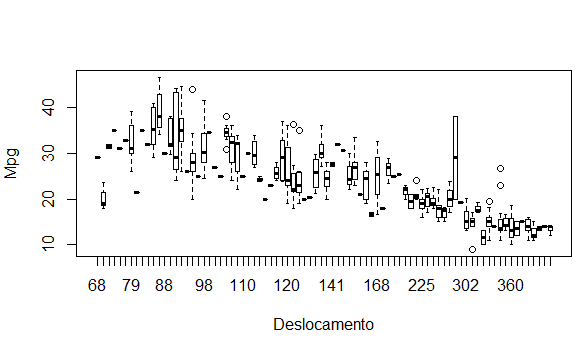
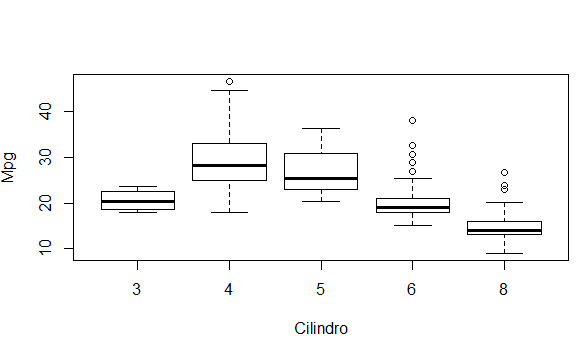


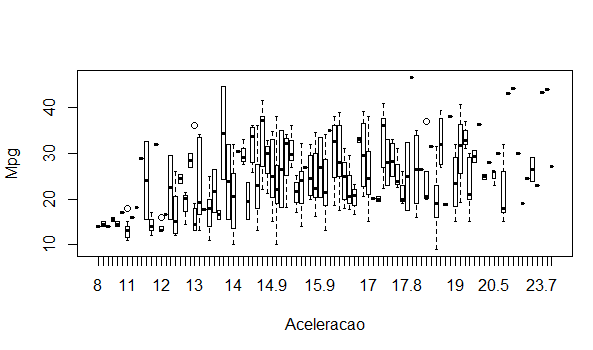
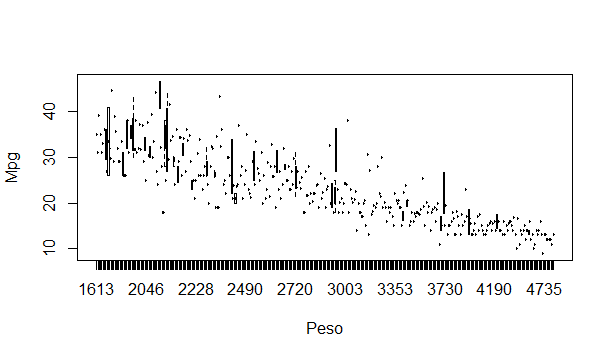


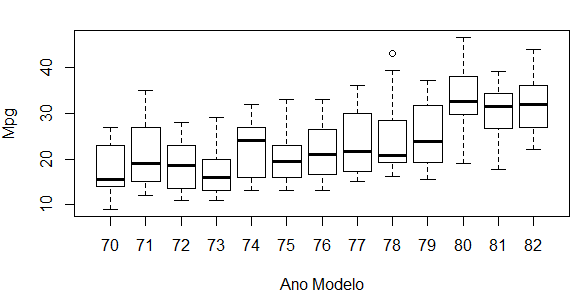




* **Plotar o boxplot para todas as variáveis quantitativas**

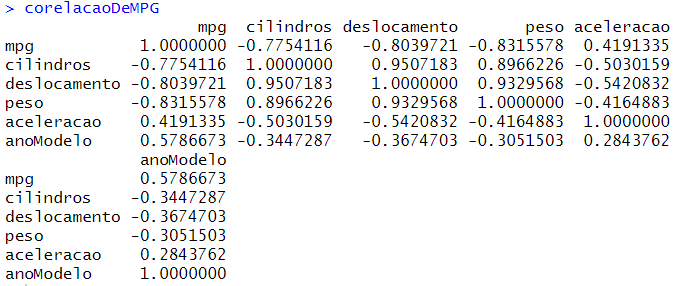
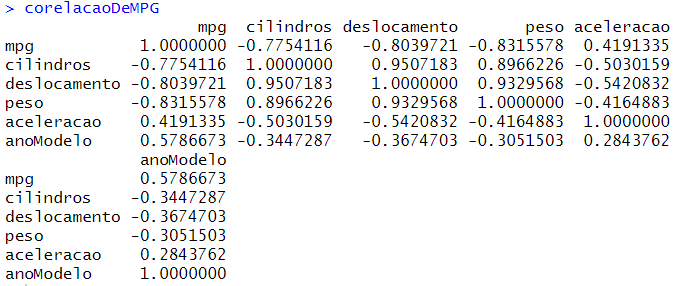
****

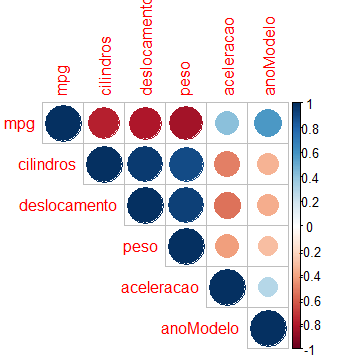




* **Calcular a correlação e plotar o gráfico de dispersão**

Com os dados abaixo, podemos perceber que, cilindros, deslocamento e peso possuem maior correlação ao campo mpg,





* **Formular um problema de regressão e análisa-lo como também discuti-lo**

Modelo de Regressão 1 - comparação do mpg com todos os dados

Conclusões após análise dos dados

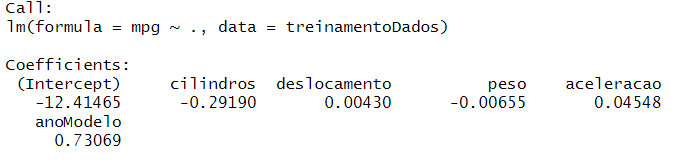
Referente a Interpretação de B

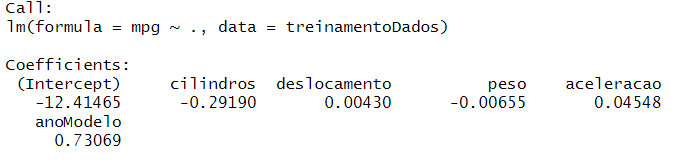
Aumentando a quantidade de cilindros(X),o consumo de gasolina mpg(Y) diminue, em media, 2,60 por litro, e com 1 litro em X, Y será de 17,04, já que sua amplitude é de 16,78.

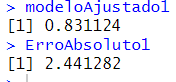
y = a + (b \* x)→ (16,78 + (0,26 \* 1) )→16,78 + (0,26)→17,04 litros

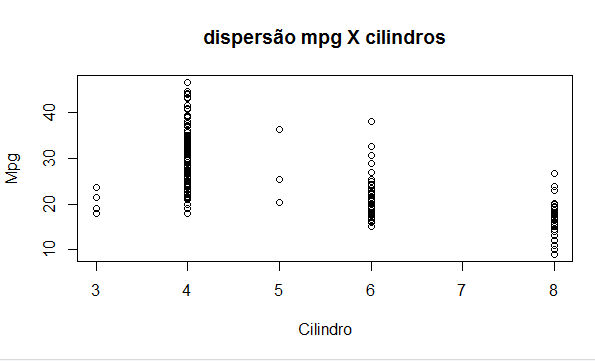
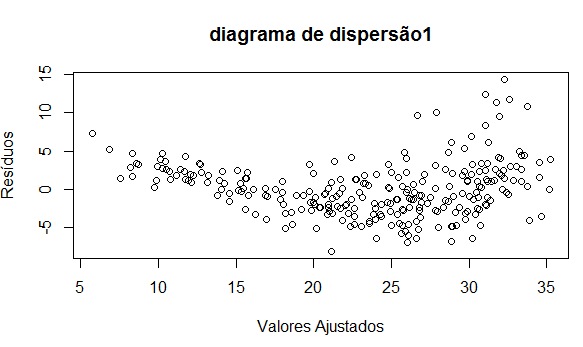
Maiores valores para x, são dos atributos: cilindros, aceleração e ano Modelo segundo á coleta de dados da api.

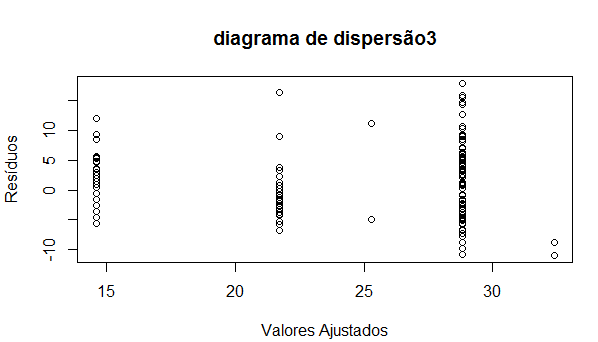
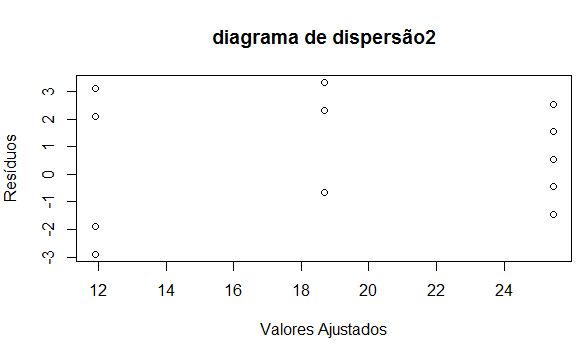
O Y é o atributo: Mpg consumo de gasolina



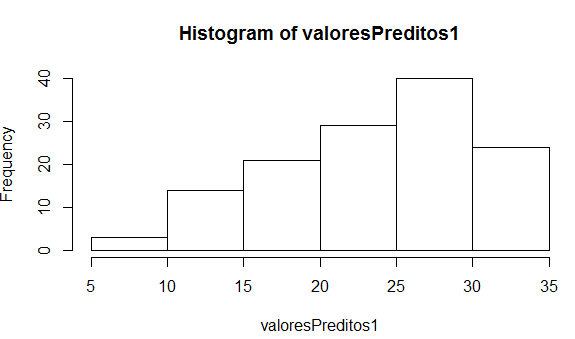


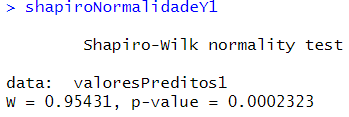




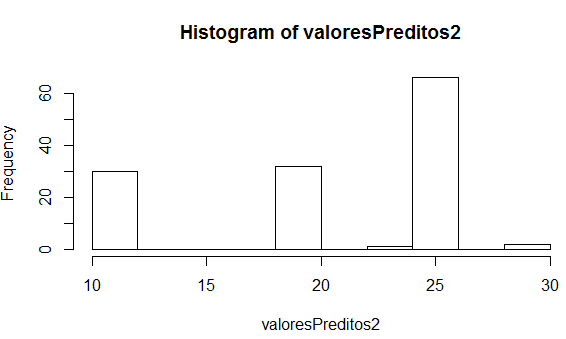


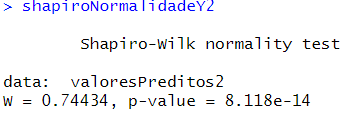
* **Realizar teste de normalidade usando boxplot, histograma e shapiro.test, discutir os resultados;**

Teste de Normalidade1 – A Distribuição dos dados de p-value é abaixo de 0.05, então distribuição dos dados não normais.

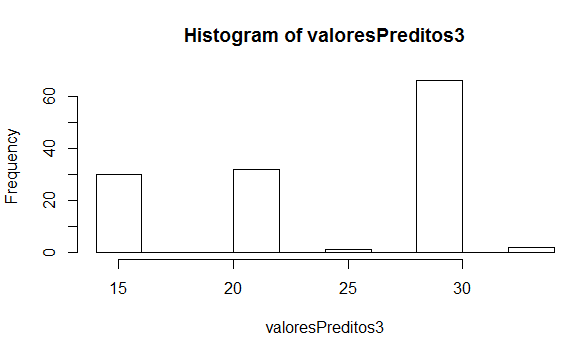
****

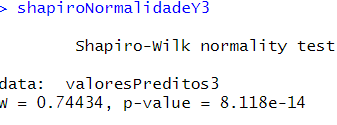
Teste de Normalidade2 – A Distribuição dos dados de p-value é acima de 0.05, então distribuição dos dados normais.

****

****

Teste de Normalidade3 – A Distribuição dos dados de p-value é acima de 0.05, então distribuição dos dados normais.

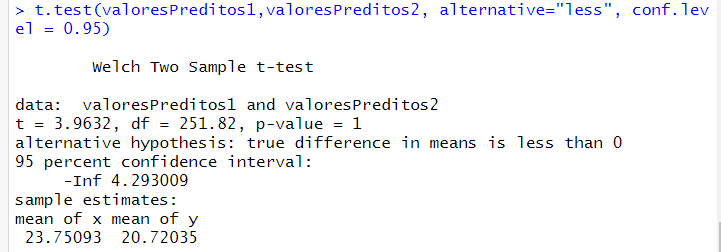




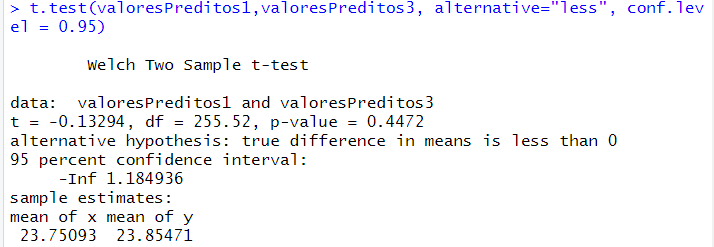
Segundo os dados acima, o teste de normalidade nos informa que teste 2 e 3 são dados normais, e que apenas para o teste 1 que são dados não normais, porque estamos usando teste tipo Shapiro e que seu tipo de comparador é o p-value.

* **Formular teste de hipótese e Anova, discutir os resultados**

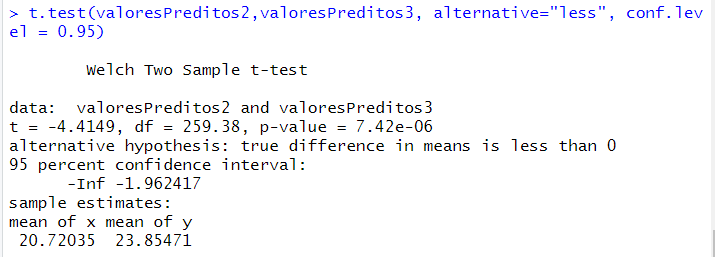
Comparação de hipóteses 1 e 2 - p-value é maior que 0.05, então as médias dos dados são diferentes



Comparação de hipóteses 1 e 3 - p-value é maior que 0.05, então as médias dos dados são diferentes

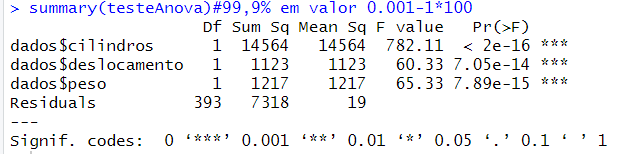
****

Comparação de hipóteses 2 e 3 - p-value é maior que 0.05, então as médias dos dados são diferentes



Teste Anova

Podemos concluir que, existe diferença de desempenho entre as classes abaixo, está com nível de confiança de 99%.



**Código Fonte em R**

#-------------------------------------------------------------#

# Leitura de Arquivos txt

#-------------------------------------------------------------#

getwd()

setwd("C:/Users/UpePos/Desktop/Nova pasta/upe/05 - Estatistica Computacional")

leituraArquivo = read.table("auto-mpg.txt", header = TRUE)

colnames(leituraArquivo) = c("mpg", "cilindros", "deslocamento", "cavaloVapor", "peso", "aceleracao", "anoModelo","origem", "nomeCarro" )

#-------------------------------------------------------------#

# trazendo apenas as colunas que eu queiro

#-------------------------------------------------------------#

dados = cbind.data.frame(leituraArquivo$mpg,leituraArquivo$cilindro,leituraArquivo$deslocamento,leituraArquivo$peso,leituraArquivo$aceleracao,leituraArquivo$anoModelo)

colnames(dados) = c("mpg", "cilindros", "deslocamento", "peso", "aceleracao", "anoModelo")

#-------------------------------------------------------------#

# amplitude de cada variavel quantitativa

#-------------------------------------------------------------#

amplitudeMpg = ( range(dados$mpg)[2] - range(dados$mpg)[1] )

amplitudeCilindros = ( range(dados$cilindros)[2] - range(dados$cilindros)[1] )

amplitudeDeslocamento = ( range(dados$deslocamento)[2] - range(dados$deslocamento)[1] )

amplitudePeso = ( range(dados$peso)[2] - range(dados$peso)[1] )

amplitudeAceleracao = ( range(dados$aceleracao)[2] - range(dados$aceleracao)[1] )

amplitudeAnoModelo = ( range(dados$anoModelo)[2] - range(dados$anoModelo)[1] )

#-------------------------------------------------------------#

# min,max,mediana,media dos dados

#-------------------------------------------------------------#

summary(dados)

#-------------------------------------------------------------#

# variançia

#-------------------------------------------------------------#

var(dados)

varianciaMpg = var(dados$mpg)

varianciaCilindros = var(dados$cilindros)

varianciaDeslocamento = var(dados$deslocamento)

varianciaPeso = var(dados$peso)

varianciaAceleracao = var(dados$aceleracao)

varianciaAnoModelo = var(dados$anoModelo)

#-------------------------------------------------------------#

# desvio padrao

#-------------------------------------------------------------#

desvioPadraoMpg = sd(dados$mpg)

desvioPadraoCilindros = sd(dados$cilindros)

desvioPadraoDeslocamento = sd(dados$deslocamento)

desvioPadraoPeso = sd(dados$peso)

desvioPadraoAceleracao = sd(dados$aceleracao)

desvioPadraoAnoModelo = sd(dados$anoModelo)

#-------------------------------------------------------------#

# o melhor grafico para cada variavel quantitativa

#-------------------------------------------------------------#

plot(dados$cilindros, dados$mpg, main = "consumo combustivel X Qtd. Cilindros", type="o", col="blue",lwd=1, xlab="Cilindro", ylab="Mpg")

hist(dados$mpg, main="Mpg", xlab = "Mpg")

hist(dados$cilindros, main="Cilindros", xlab = "Cilindros")

hist(dados$deslocamento, main="Deslocamento", xlab = "Deslocamento")

hist(dados$peso, main="Peso", xlab = "Peso")

hist(dados$aceleracao, main="Aceleração", xlab = "Aceleração")

hist(dados$anoModelo, main="Ano Modelo", xlab = "Ano Modelo")

#-------------------------------------------------------------#

# graficos das variaveis quantitativas

#-------------------------------------------------------------#

boxplot(dados$mpg ~ dados$cilindros, xlab = "Cilindro", ylab = "Mpg")

boxplot(dados$mpg ~ dados$deslocamento, xlab = "Deslocamento", ylab = "Mpg")

boxplot(dados$mpg ~ dados$peso, xlab = "Peso", ylab = "Mpg")

boxplot(dados$mpg ~ dados$aceleracao, xlab = "Aceleracao", ylab = "Mpg")

boxplot(dados$mpg ~ dados$anoModelo, xlab = "Ano Modelo", ylab = "Mpg")

#-------------------------------------------------------------#

# Correlação

#-------------------------------------------------------------#

library(corrplot)

corelacaoDeMPG = cor(dados)

corrplot(corelacaoDeMPG, method="circle", type = "upper")

#-------------------------------------------------------------#

# partição dos dados em teste e treinamento

#-------------------------------------------------------------#

library(lattice)

library(ggplot2)

library(caret)

indice = createDataPartition(dados$mpg, p=2/3, list = FALSE)

treinamentoDados = dados[indice,]

testeDados = dados[-indice,]

#-------------------------------------------------------------#

# Modelo de Regressão 1 - comparação do mpg com todos os dados

#-------------------------------------------------------------#

modelo1 = lm(mpg ~ ., data = treinamentoDados)

valoresPreditos1 = predict(modelo1, newdata = data.frame(testeDados) )

modeloAjustado1 = R2(valoresPreditos1, testeDados$mpg)

ErroAbsoluto1 = MAE(valoresPreditos1, testeDados$mpg)

x=dados$cilindros[1]

Yestimado = modelo1$coefficients[1] + modelo1$coefficients[2]\*x #-> para deslocamento

plot( dados$cilindros,dados$mpg, main = "dispersão mpg X cilindros", xlab ="Cilindro", ylab = "Mpg" )

abline(modelo1)

plot(fitted(modelo1), main = "diagrama de dispersão1", residuals(modelo1), xlab="Valores Ajustados", ylab="Resíduos")

#-------------------------------------------------------------#

# Modelo de Regressão 2 - 3 maiores

#-------------------------------------------------------------#

modelo2 = lm(mpg ~ cilindros,aceleracao,anoModelo, data = treinamentoDados)

valoresPreditos2 = predict(modelo2, newdata = data.frame(testeDados) )

modeloAjustado2 = R2(valoresPreditos2, testeDados$mpg)

ErroAbsoluto2 = MAE(valoresPreditos2, testeDados$mpg)

plot(dados$mpg, dados$cilindros,main = "D")

abline(modelo2)

plot(fitted(modelo2), main = "diagrama de dispersão2", residuals(modelo2), xlab="Valores Ajustados", ylab="Resíduos")

#-------------------------------------------------------------#

# Modelo de Regressão 3 - o maior

#-------------------------------------------------------------#

modelo3 = lm(mpg ~ cilindros, data = treinamentoDados)

valoresPreditos3 = predict(modelo3, newdata = data.frame(testeDados) )

modeloAjustado3 = R2(valoresPreditos3, testeDados$mpg)

ErroAbsoluto3 = MAE(valoresPreditos3, testeDados$mpg)

plot(dados$mpg, dados$cilindros)

abline(modelo3)

plot(fitted(modelo3), main = "diagrama de dispersão3", residuals(modelo3), xlab="Valores Ajustados", ylab="Resíduos")

#-------------------------------------------------------------#

# teste de normalidade

#-------------------------------------------------------------#

shapiroNormalidadeY1 = shapiro.test(valoresPreditos1)

ks.test(valoresPreditos1, "pnorm", mean(valoresPreditos1), sd(valoresPreditos1))

hist(valoresPreditos1)

shapiroNormalidadeY2 = shapiro.test(valoresPreditos2)

ks.test(valoresPreditos2, "pnorm", mean(valoresPreditos2), sd(valoresPreditos2))

hist(valoresPreditos2)

shapiroNormalidadeY3 = shapiro.test(valoresPreditos3)

ks.test(valoresPreditos3, "pnorm", mean(valoresPreditos3), sd(valoresPreditos3))

hist(valoresPreditos3)

#-------------------------------------------------------------#

# teste de hipotese

#-------------------------------------------------------------#

t.test(valoresPreditos1,valoresPreditos2, alternative="less", conf.level = 0.95)

t.test(valoresPreditos1,valoresPreditos3, alternative="less", conf.level = 0.95)

t.test(valoresPreditos2,valoresPreditos3, alternative="less", conf.level = 0.95)

#-------------------------------------------------------------#

# teste de Anova

#-------------------------------------------------------------#

testeAnova = aov(dados$mpg ~ dados$cilindros + dados$deslocamento + dados$peso)

summary(testeAnova)#99,9% em valor 0.001-1\*100