**Análise Preditiva E Data Mining**

**Turma 41BDT**

Trabalho em Grupo 1

**Habilidades desenvolvidas: Definição do problema. Análise exploratória dos dados.**

**Análise de correlação e associação.**

Use os dados *Bike\_Sharing.xlsx* para construir as análises descritivas, correlação, associação e modelo preditivo para previsão do número de bikes alugadas por mês.

Descrição:

Os sistemas de compartilhamento de bicicletas são uma nova geração de aluguel de bicicletas tradicional, onde todo o processo de associação, locação e devolução tornou-se automático. Através destes sistemas, o usuário pode facilmente alugar uma bicicleta a partir de uma determinada posição e retornar em outra posição. Atualmente, existem cerca de 500 programas de compartilhamento de bicicletas em todo o mundo, compostos por mais de 500 mil bicicletas. Hoje, existe um grande interesse nesses sistemas devido ao seu importante papel no trânsito, questões ambientais e de saúde.

Fonte de dados: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

1. Classifique o tipo de variável

|  |  |  |
| --- | --- | --- |
| **Variável** | **Descrição** | **Tipo de Variável** |
| **instant** | record index | ID |
| **dteday** | date | ID |
| **season** | season (1:springer, 2:summer, 3:fall, 4:winter) | Qualitativa ordinal |
| **yr** | year (0: 2011, 1:2012) | Qualitativa nominal |
| **mnth** | month ( 1 to 12) | Quantitativa discreta |
| **holiday** | weather day is holiday or not | Qualitativa nominal |
| **weekday** | day of the week | Qualitativa ordinal |
| **workingday** | if day is neither weekend nor holiday is 1, otherwise is 0. | Qualitativa nominal |
| **weathersit** | 1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | Qualitativa nominal |
| **temp** | Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) | Quantitativa contínua |
| **atemp** | Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale) | Quantitativa contínua |
| **hum** | Normalized humidity. The values are divided to 100 (max) | Quantitativa contínua |
| **windspeed** | Normalized wind speed. The values are divided to 67 (max) | Quantitativa contínua |
| **casual** | count of casual users | Quantitativa discreta |
| **registered** | count of registered users | Quantitativa discreta |
| **cnt** | count of total rental bikes including both casual and registered | Quantitativa discreta |

1. Definição do problema:

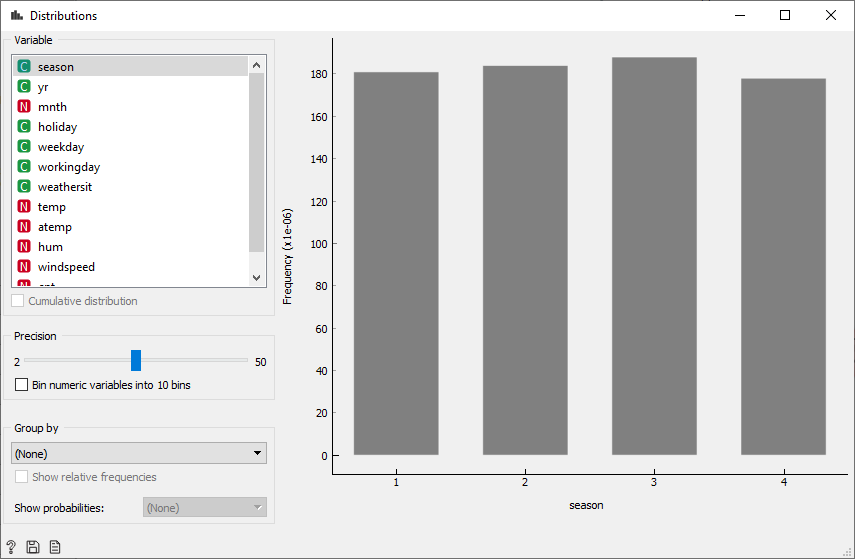
Além das interessantes aplicações do mundo real dos sistemas de compartilhamento de bicicletas, as características dos dados gerados por esses sistemas os tornam atraentes para a pesquisa. Ao contrário de outros serviços de transporte, como ônibus ou metrô, a duração da viagem, a partida e a posição de chegada são explicitamente registradas nesses sistemas. Esse recurso transforma o sistema de compartilhamento de bicicletas em uma rede virtual de sensores que pode ser usada para detectar a mobilidade na cidade. Portanto, espera-se que a maioria dos eventos importantes na cidade possa ser detectada através do monitoramento desses dados.

1. Objetivo do estudo:

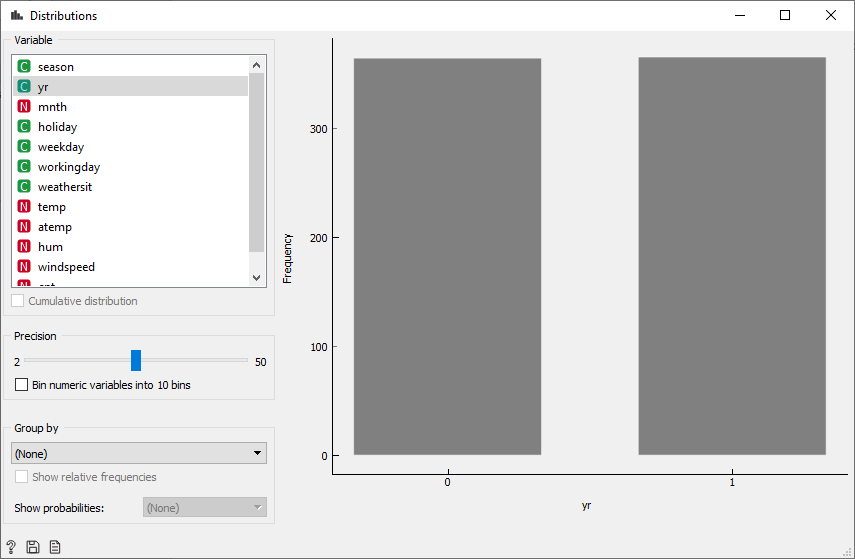
Para onde vão os pilotos da Capital Bikeshare? Quando eles montam? Até onde eles vão? Quais estações são mais populares? Em que dias da semana a maioria dos passeios são realizados? Ouvimos todas essas perguntas desde o lançamento em 2010 e temos o prazer de fornecer os dados que mostram as respostas da nossa primeira viagem até hoje.

1. Apresente as tabelas descritivas e gráficos das variáveis. Interprete os resultados.

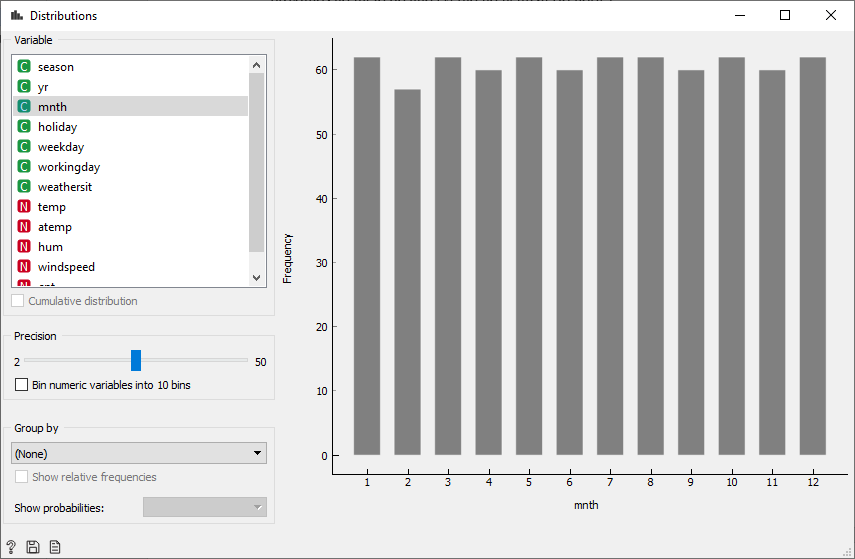
**season (estação do ano):** o gráfico demonstra o número de dias em cada estação do ano, os valores são aproximados.



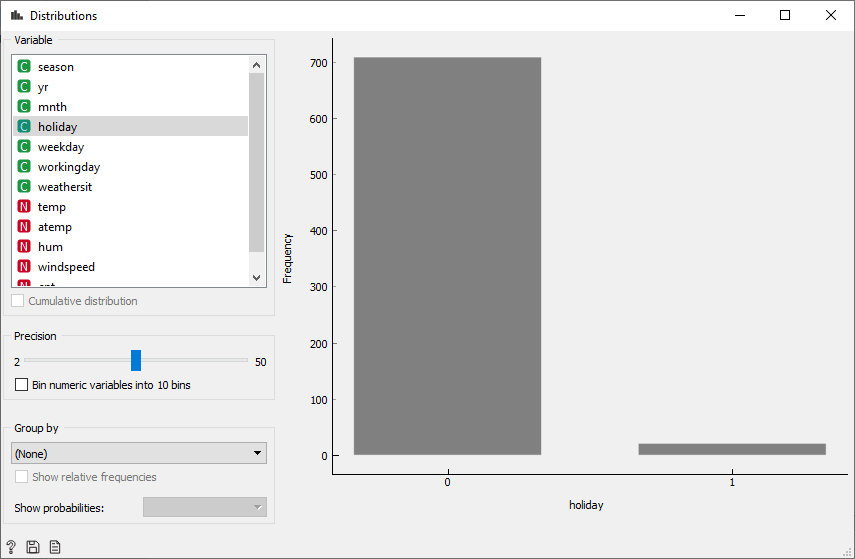
**yr (ano):** o gráfico demonstra o número de dias em cada estação do ano, os valores são aproximados, com apenas 1 a mais no yr=1 por ser ano bissexto.



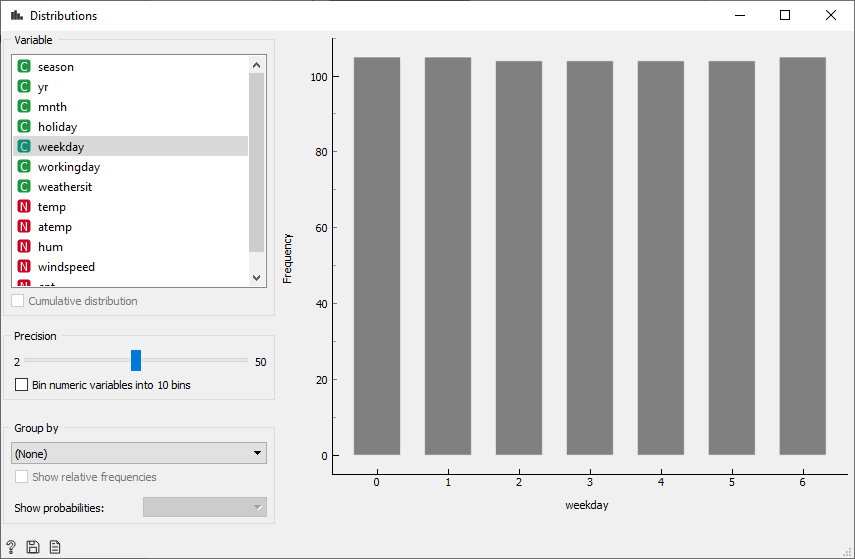
**mnth (mês):** o gráfico demonstra o número de dias em cada mês, os valores são aproximados com o mês 2 um pouco menor do que os outros.



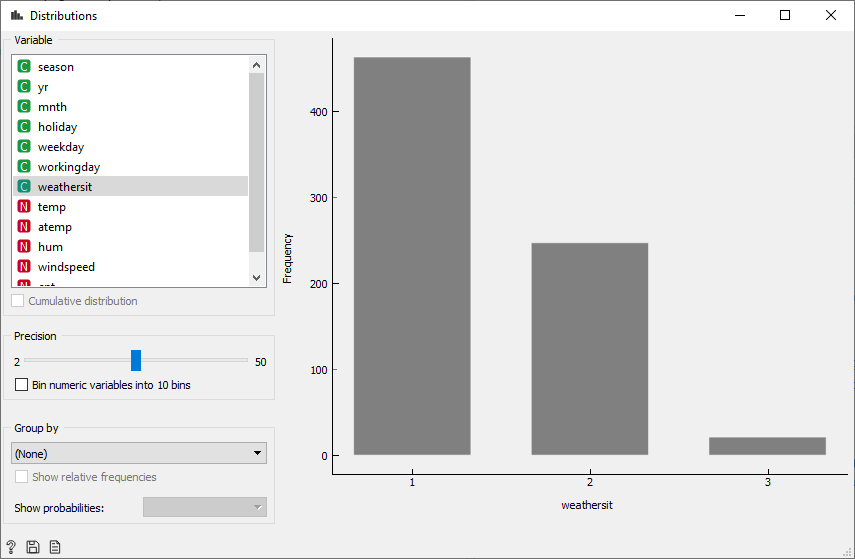
**holiday (feriados):** o gráfico demonstra o número de dias com feriados (holiday=1), a diferença é muito grande para dias sem feriado (holiday=0).



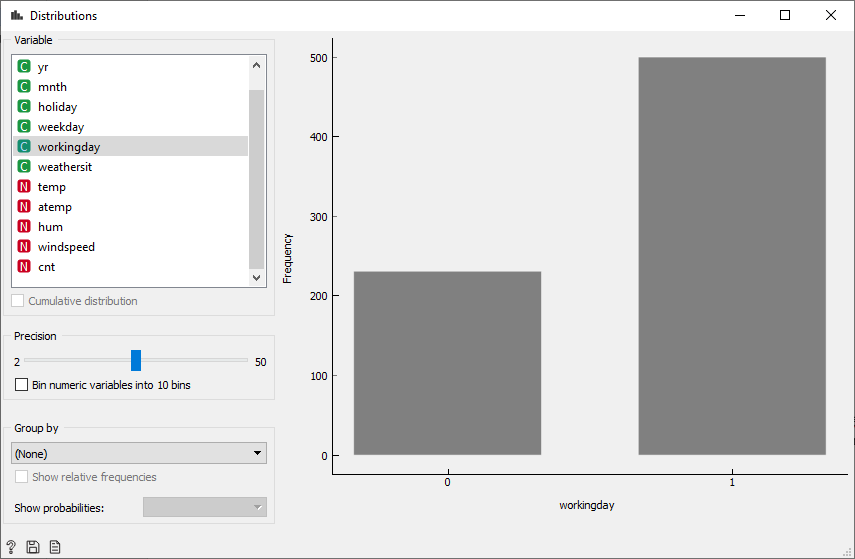
**weekday (dia da semana):** o gráfico demonstra o número de dias em cada dia da semana, os números são quase equivalentes.



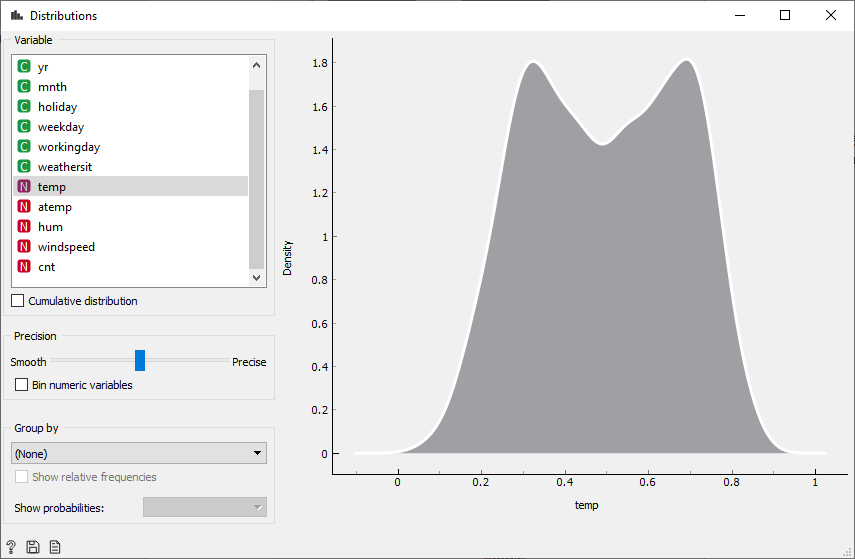
**weathersit (clima):** o gráfico demonstra o número de dias de acordo com o clima (ensolarado=1, nublado=2, chuva ou neve=3).



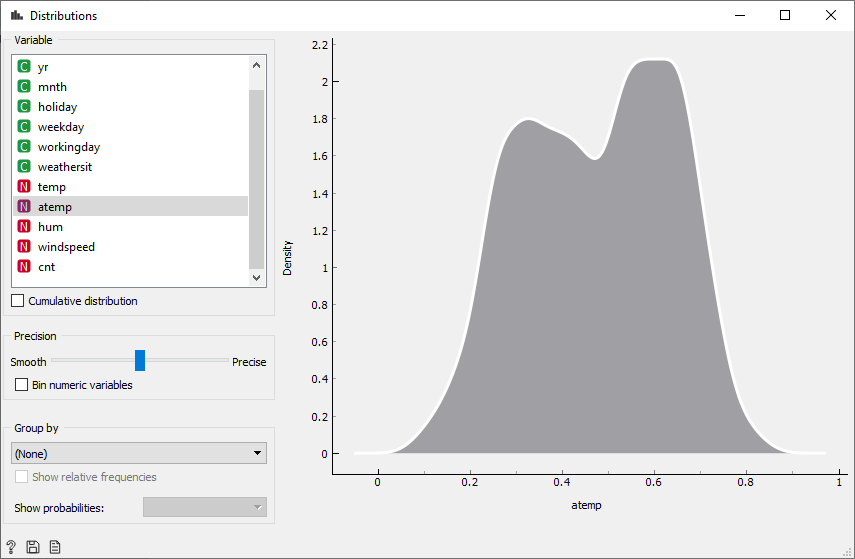
**workday (dia de trabalho):** o gráfico demonstra o número de dias de trabalho (que não não nem finais de semana e nem feriados, workday=1),os dias de trabalho são mais do que o dobro dos dias de sem trabalho.



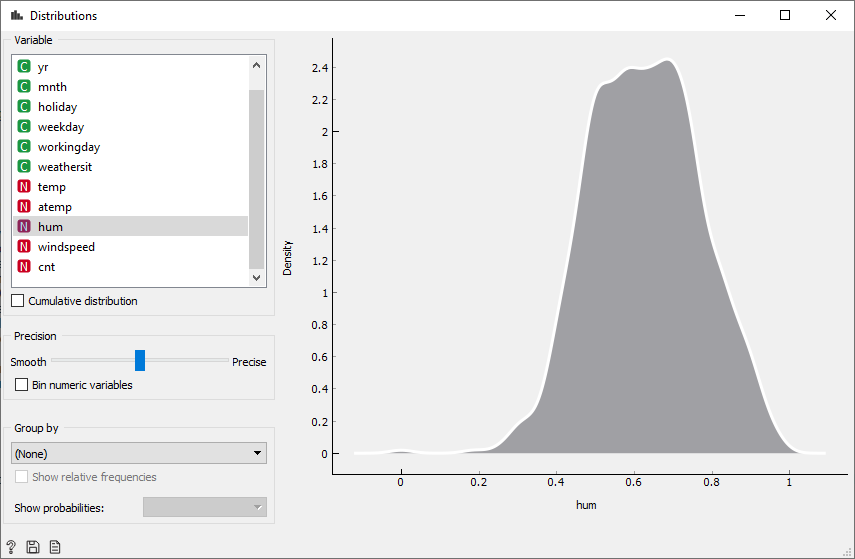
**temp (temperatura):** o gráfico exibe a densidade da distribuição da temperatura, as temperaturas muito baixas e muito altas são menos frequentes.



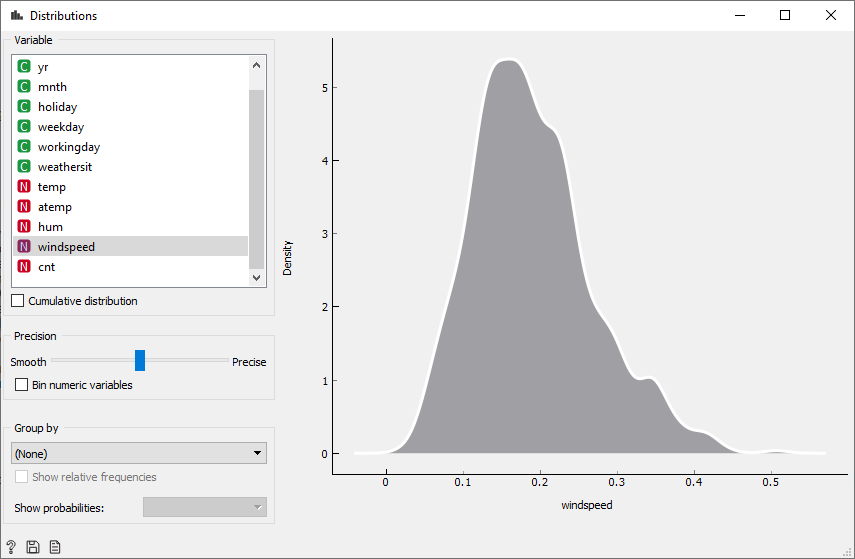
**atemp (sensação termica):** o gráfico exibe a densidade da distribuição da sensação da temperatura, as temperaturas muito baixas e muito altas são menos frequentes.



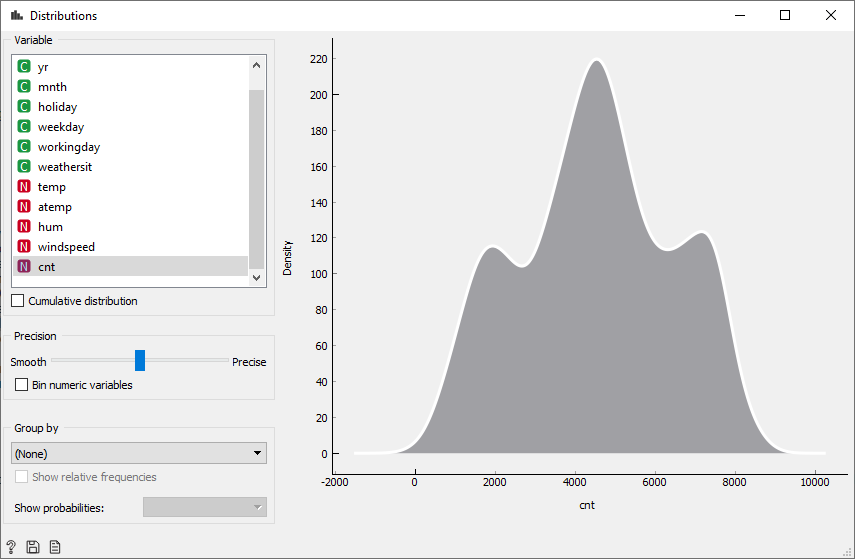
**hum (humidade):** o gráfico exibe a densidade da distribuição da humidade, as humidades muito baixas e muito altas são menos frequentes.



**windspeed (velocidade do vento):** o gráfico exibe a densidade da velocidade do vento, as velocidades muito baixas e muito altas são menos frequentes.



**cnt (total de bicicletas alugadas):** o gráfico exibe a densidade de bicicletas alugadas.



1. Quais variáveis foram normalizadas? Por quê? Apresente a fórmula utilizada.

As variáveis que normalizadas foram:

**- temp:** normalização da temperatura em graus Celsius utilizando o mínimo e máximo.

Os valores foram obtidos com a seguinte formula:

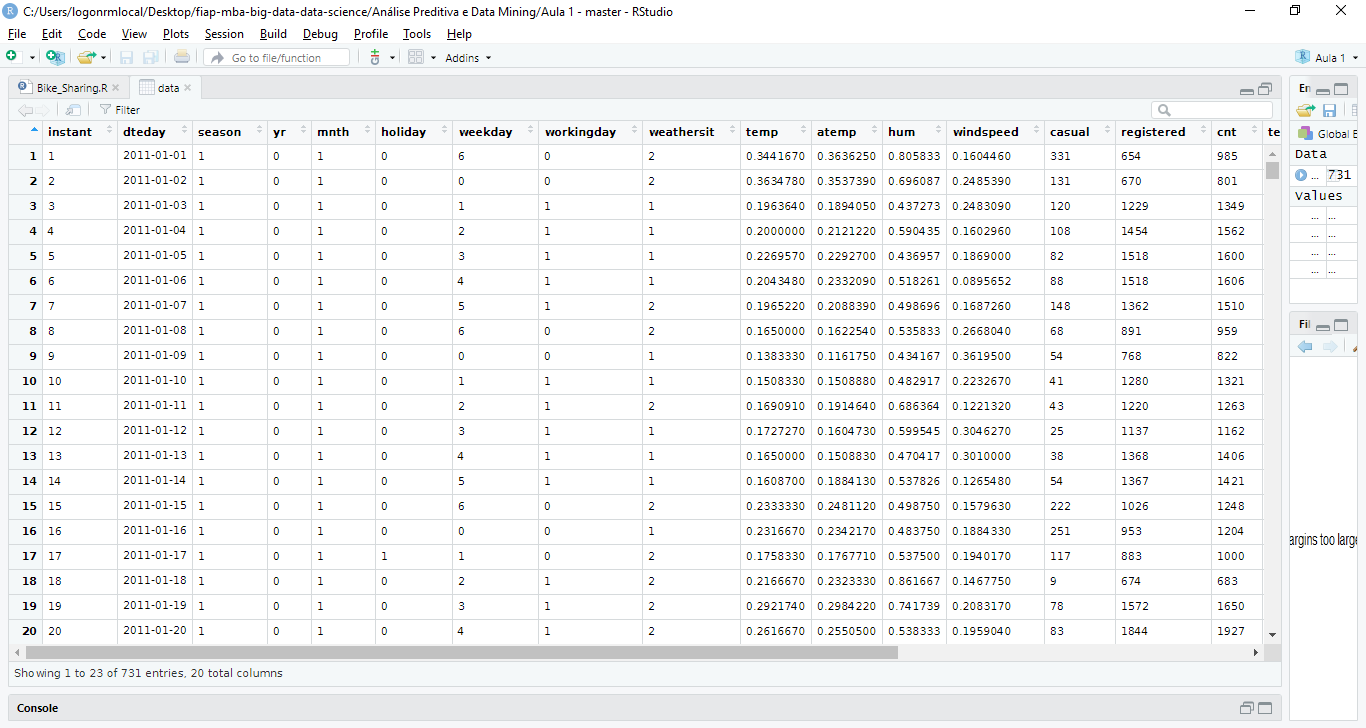
(t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39  
 **- atemp:** normalização da temperatura em graus Celsius utilizando o mínimo e máximo.   
Os valores foram obtidos com a seguinte formula:

(t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50  
 **- hum:** normalização da humidade. Os valores foram divididos por 100.  
 **- windspeed:** normalização da velocidade do vento. Os valores foram divididos por 67.

Utilize o programa R para criar as variáveis originais das variáveis normalizadas e faça o histograma da variável original e normalizada. Compare as duas variáveis (original e padronizada), qual a diferença entre elas?

**Leitura da base:**  
data <- read.csv(file = 'Bike\_Sharing.csv')  
  
**# Visualização da base**

View(data)



**- temp:** temperatura em graus Celsius, foi normalizada utilizando o mínimo (-8) e máximo (39).

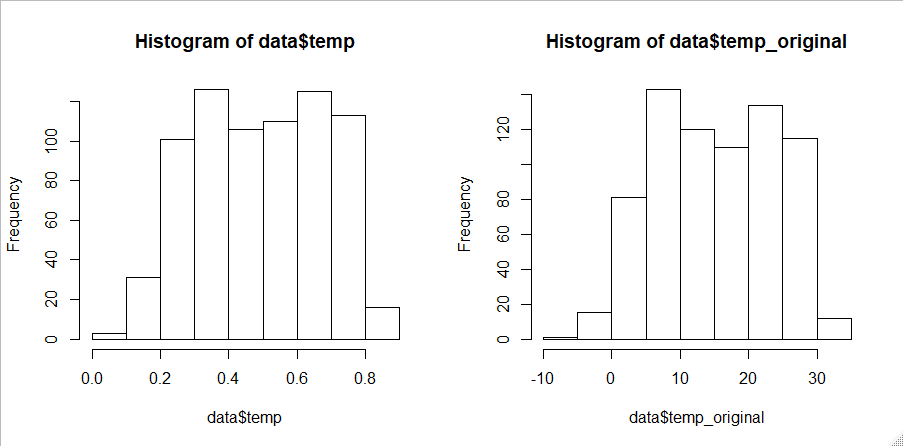
temp\_max = 39 **# temperatura original máxima**

temp\_min = -8 **# temperatura original mínima**

**# formula para voltar a variável original**  
data$temp\_original = data$temp \* (temp\_max - temp\_min) + temp\_min

**# histograma da variável normalizada “temp”**  
hist(data$temp)  
  
**# histograma da variável original “temp\_original”**

hist(data$temp\_original)



Os histogramas são parecidos, o que mudou foi a escala.

**- atemp:** sensação da temperatura em graus Celsius, foi normalizadas utilizando o mínimo   
(-16) e máximo (50).

atemp\_max = 50  **# sensação de temperatura original máxima**

atemp\_min = -16 **# sensação de temperatura original mínima**

**# formula para voltar a variável original**

data$atemp\_original = data$atemp \* (atemp\_max - atemp\_min) + atemp\_min

**# para plotar lado a lado**

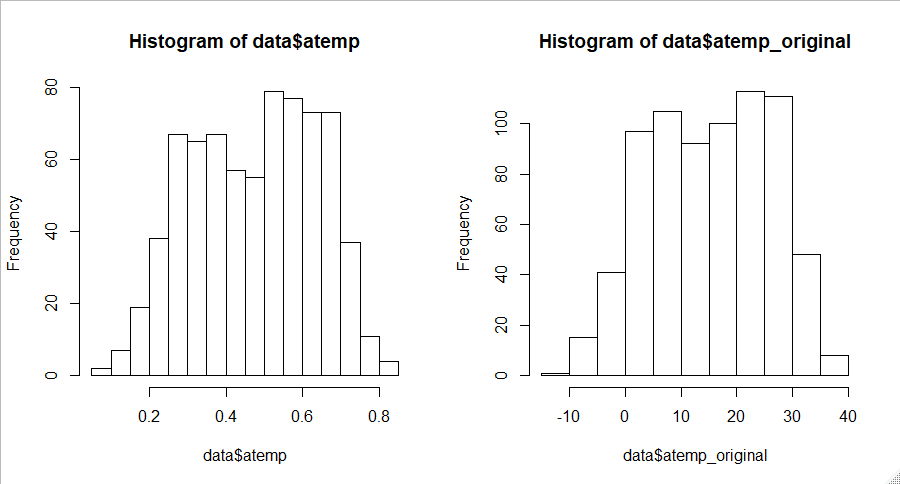
par(mfrow=c(1,2))

**# histograma da variável normalizada “atemp”**

hist(data$atemp)

**# histograma da variável original “atemp\_original”**

hist(data$atemp\_original)



Os histogramas são parecidos, somente o número de colunas foi diferente pelo padrão da função, o que mudou foi a escala.

**- hum:** humidade, os valores foram divididos por 100.

# formula para voltar a variável original

data$hum\_original = data$hum \* 100

# para plotar lado a lado

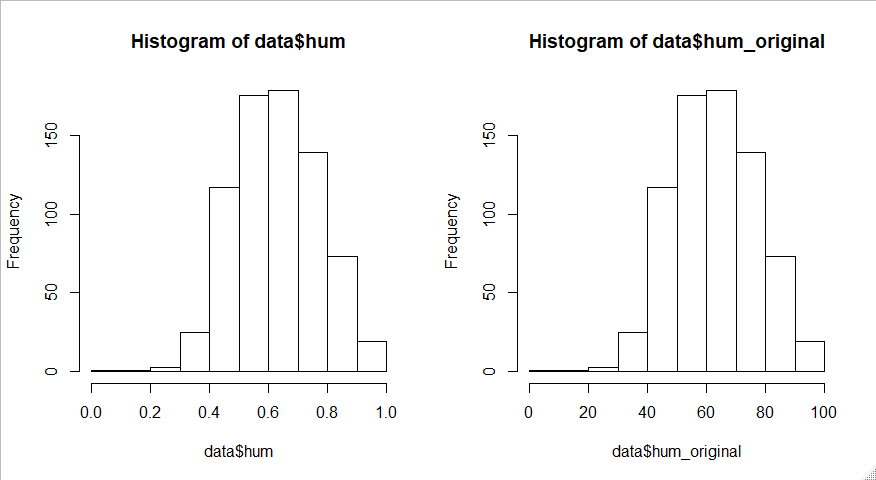
par(mfrow=c(1,2))

# histograma da variável normalizada “hum”

hist(data$hum)

# histograma da variável original “hum\_original”

hist(data$hum\_original)



Os histogramas são idênticos pois o valor foi apenas multiplicado por 100, apenas a escala foi alterada.

**- windspeed:** velocidade do vento, os valores foram divididos por 67.

# formula para voltar a variável original

data$windspeed\_original = data$windspeed \* 67

# para plotar lado a lado

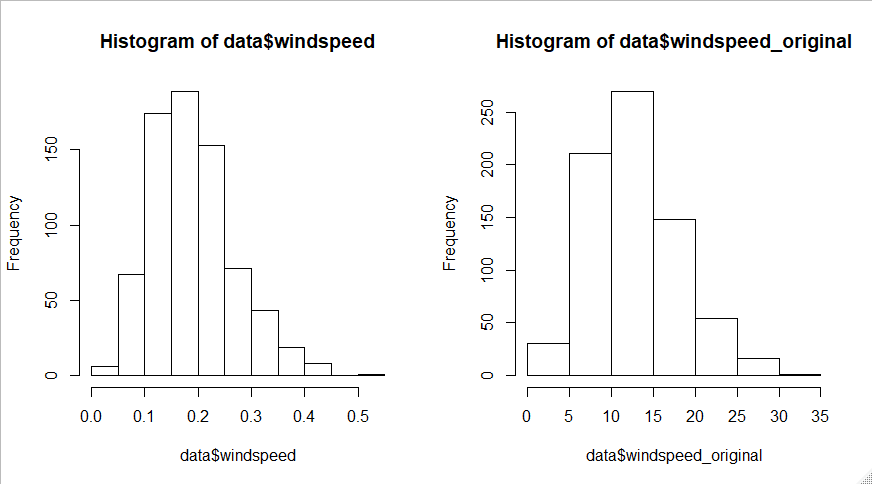
par(mfrow=c(1,2))

# histograma da variável normalizada “windspeed”

hist(data$windspeed)

# histograma da variável original “windspeed\_original”

hist(data$windspeed\_original)



Os histogramas são idênticos pois o valor foi apenas multiplicado por 67, apenas a escala foi alterada.

**Atividade prática no R**

## Copiar o arquivo “Bike\_sharing.csv” no diretório de trabalho

# Abrir o programa R

## Programando no R

# indicar diretório de trabalho

setwd("**F:/BDT/41BDT/TRAB1**")

## Importando arquivo

# utilizando a função read.table para ler o arquivo CSV

arquivo<-read.table(file="Bike\_Sharing.csv", header=T,sep=",")

# dimensão da base de dados (tabela)

dim(arquivo)

[1] 731 16

# nomes das variáveis

names(arquivo)

[1] "instant" "dteday" "season" "yr" "mnth" "holiday" "weekday"

[8] "workingday" "weathersit" "temp" "atemp" "hum" "windspeed" "casual"

[15] "registered" "cnt"

# visualizar primeiras 20 linhas

head(arquivo,20)

instant dteday season yr mnth holiday weekday workingday weathersit temp atemp hum

1 1 2011-01-01 1 0 1 0 6 0 2 0.344167 0.363625 0.805833

2 2 2011-01-02 1 0 1 0 0 0 2 0.363478 0.353739 0.696087

3 3 2011-01-03 1 0 1 0 1 1 1 0.196364 0.189405 0.437273

4 4 2011-01-04 1 0 1 0 2 1 1 0.200000 0.212122 0.590435

5 5 2011-01-05 1 0 1 0 3 1 1 0.226957 0.229270 0.436957

6 6 2011-01-06 1 0 1 0 4 1 1 0.204348 0.233209 0.518261

7 7 2011-01-07 1 0 1 0 5 1 2 0.196522 0.208839 0.498696

8 8 2011-01-08 1 0 1 0 6 0 2 0.165000 0.162254 0.535833

9 9 2011-01-09 1 0 1 0 0 0 1 0.138333 0.116175 0.434167

10 10 2011-01-10 1 0 1 0 1 1 1 0.150833 0.150888 0.482917

11 11 2011-01-11 1 0 1 0 2 1 2 0.169091 0.191464 0.686364

12 12 2011-01-12 1 0 1 0 3 1 1 0.172727 0.160473 0.599545

13 13 2011-01-13 1 0 1 0 4 1 1 0.165000 0.150883 0.470417

14 14 2011-01-14 1 0 1 0 5 1 1 0.160870 0.188413 0.537826

15 15 2011-01-15 1 0 1 0 6 0 2 0.233333 0.248112 0.498750

16 16 2011-01-16 1 0 1 0 0 0 1 0.231667 0.234217 0.483750

17 17 2011-01-17 1 0 1 1 1 0 2 0.175833 0.176771 0.537500

18 18 2011-01-18 1 0 1 0 2 1 2 0.216667 0.232333 0.861667

19 19 2011-01-19 1 0 1 0 3 1 2 0.292174 0.298422 0.741739

20 20 2011-01-20 1 0 1 0 4 1 2 0.261667 0.255050 0.538333

windspeed casual registered cnt

1 0.1604460 331 654 985

2 0.2485390 131 670 801

3 0.2483090 120 1229 1349

4 0.1602960 108 1454 1562

5 0.1869000 82 1518 1600

6 0.0895652 88 1518 1606

7 0.1687260 148 1362 1510

8 0.2668040 68 891 959

9 0.3619500 54 768 822

10 0.2232670 41 1280 1321

11 0.1221320 43 1220 1263

12 0.3046270 25 1137 1162

13 0.3010000 38 1368 1406

14 0.1265480 54 1367 1421

15 0.1579630 222 1026 1248

16 0.1884330 251 953 1204

17 0.1940170 117 883 1000

18 0.1467750 9 674 683

19 0.2083170 78 1572 1650

20 0.1959040 83 1844 1927

## Análise descritiva univariada

# trabalhar com as variáveis

attach(arquivo)

**# selecionar as variáveis quantitativas**dadosquant=subset(arquivo,select=c(cnt,temp,weekday))  
  
**# Medidas resumo**

summary(dadosquant)

cnt temp weekday

Min. : 22 Min. :0.05913 Min. :0.000

1st Qu.:3152 1st Qu.:0.33708 1st Qu.:1.000

Median :4548 Median :0.49833 Median :3.000

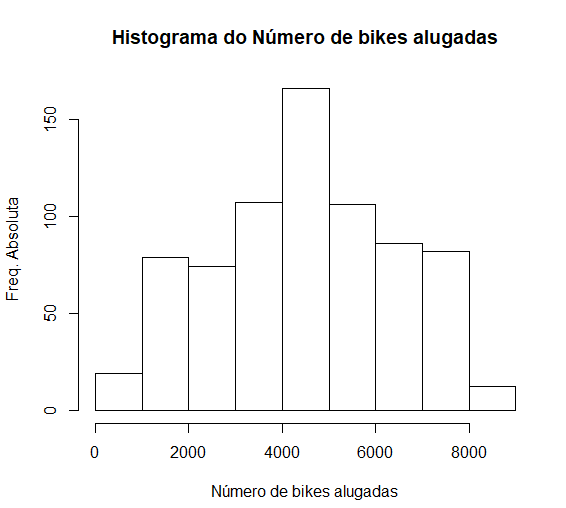
Mean :4504 Mean :0.49538 Mean :2.997

3rd Qu.:5956 3rd Qu.:0.65542 3rd Qu.:5.000

Max. :8714 Max. :0.86167 Max. :6.000

**#Análise da variável CNT**

hist(cnt, xlab="Número de bikes alugadas", ylab="Freq. Absoluta", main="Histograma do Número de bikes alugadas")



# Calcular a média da variável CNT

media<-mean(cnt)

print(media)

|  |
| --- |
| **Interprete o resultado:** [1] 4504.349  A média de bicicletas alugadas por dia é de **4.504,349**. |

# Calcular o desvio padrão da variável CNT

dp<-sd(cnt)

print(dp)

|  |
| --- |
| **Interprete o resultado:** [1] 1937.211  A variação na média de bicicletas alugadas por dia é de **1.937,211**. |

# Calcular o coeficiente de variação da variável CNT

cv<-sd(cnt)/mean(cnt)

print(cv)

|  |
| --- |
| **Interprete o resultado:** [1] 0.4300758  O coeficiente de variação demonstra que o número de bicicletas alugadas por dia varia em **43%**. |

Atividade extra classe:

https://www.bing.com/videos/search?q=correla%c3%a7%c3%a3o+de+pearson+no+r&ru=%2fsearch%3fq%3dcorrela%25C3%25A7%25C3%25A3o%2bde%2bpearson%2bno%2br%26FORM%3dQSRE3&view=detail&mmscn=vwrc&mid=C23FC80A167927CA87F5C23FC80A167927CA87F5&FORM=WRVORC

**# selecionar as variáveis quantitativas**

dadosquant=subset(arquivo,select=c( , , ))

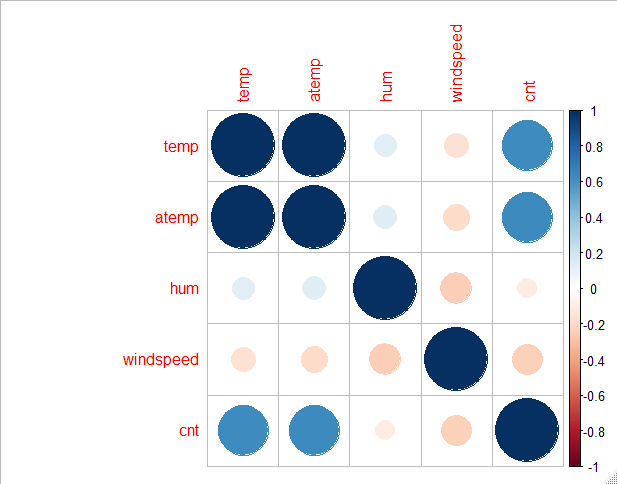
matriz<-cor(dadosquant)

install.package(“corrplo”)

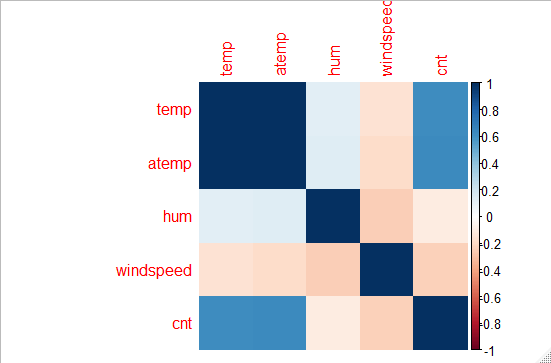
library(corrplot)

corrplot

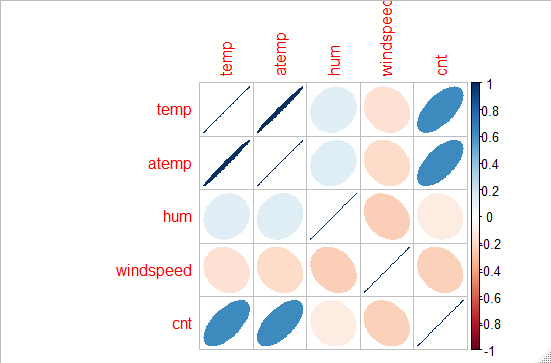
corrplot(matriz)



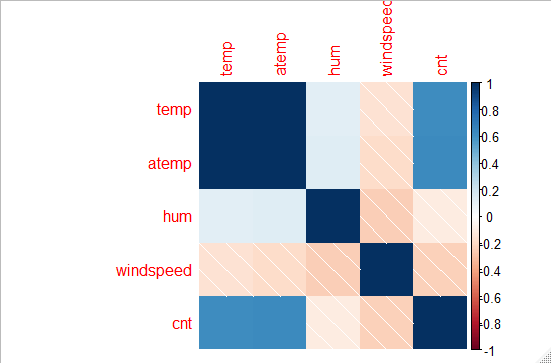
corrplot(matriz, method = “color”)



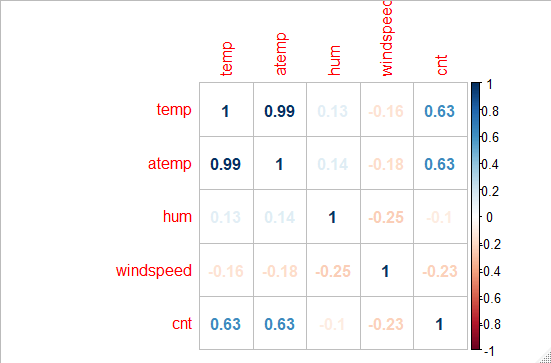
corrplot(matriz, method = “ellipse”)



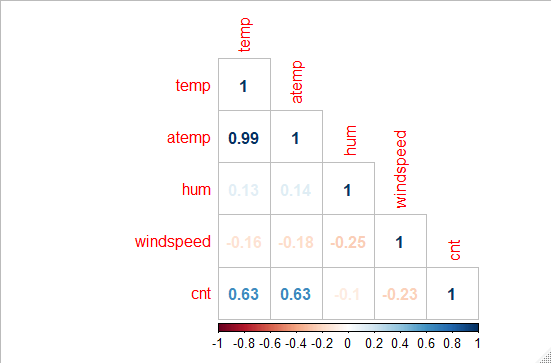
corrplot(matriz, method = “shade”)



corrplot(matriz, method = “number”)



corrplot(matriz, type=”lower”, method = “number”)



# p-value

install.package(Hmisc)

libray(Hmisc)

matriz<-rcorr(as.matrix(dadosquant))

matriz$r

temp atemp hum windspeed cnt

temp 1.0000000 0.9917016 0.1269629 -0.1579441 0.6274940

atemp 0.9917016 1.0000000 0.1399881 -0.1836430 0.6310657

hum 0.1269629 0.1399881 1.0000000 -0.2484891 -0.1006586

windspeed -0.1579441 -0.1836430 -0.2484891 1.0000000 -0.2345450

cnt 0.6274940 0.6310657 -0.1006586 -0.2345450 1.0000000

matriz$P

temp atemp hum windspeed cnt

temp 0.0000 0.0006 0.0000 0.0000

atemp 0.0000 0.0001 0.0000 0.0000

hum 0.0006 0.0001 0.0000 0.0065

windspeed 0.0000 0.0000 0.0000 0.0000

cnt 0.0000 0.0000 0.0065 0.0000

matriz$n

temp atemp hum windspeed cnt

temp 731 731 731 731 731

atemp 731 731 731 731 731

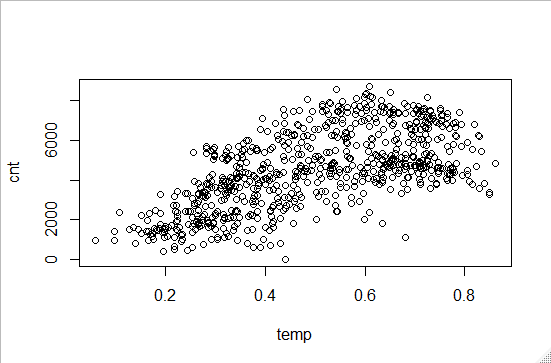
hum 731 731 731 731 731

windspeed 731 731 731 731 731

cnt 731 731 731 731 731

# gráfico de dispersão

plot(cnt~temp)



# pacote correlação

cor(cnt,temp)

[1] 0.627494

# teste estatístico

cor.test(cnt,temp,method= “pearson”)

Pearson's product-moment correlation

data: cnt and temp

t = 21.759, df = 729, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5814369 0.6695422

sample estimates:

cor

0.627494

# teste qui-quadrado

arquivo$faixacnt <- cut(cnt

, breaks = c(22, 892, 1762, 2632, 3502, 4372, 5242, 6112, 6982, 7851, 8720)

, right = F

, labels = c('[22 a 892]','(892 a 1762]', '(1762 a 2632]', '(2632 a 3502]','(3502 a 4372]', '(4372 a 5242]','(5242 a 6112]', '(6112 a 6982]', '(6982 a 7851]', '(7851 a 8720]'))

3) Faça a tabela bivariada e comente os resultados.

|  |
| --- |
| # Frequência absoluta  table(arquivo$faixacnt, holiday)  holiday  0 1  [22 a 892] 13 0  (892 a 1762] 57 5  (1762 a 2632] 68 3  (2632 a 3502] 69 4  (3502 a 4372] 121 1  (4372 a 5242] 136 1  (5242 a 6112] 77 4  (6112 a 6982] 75 2  (6982 a 7851] 77 1  (7851 a 8720] 17 0  Distribuição por número de alugueis em feriados não feriados. |

# Frequência relativa

prop.table(table(arquivo$faixacnt,holiday"))

|  |
| --- |
| holiday  0 1  [22 a 892] 0.017783858 0.000000000  (892 a 1762] 0.077975376 0.006839945  (1762 a 2632] 0.093023256 0.004103967  (2632 a 3502] 0.094391245 0.005471956  (3502 a 4372] 0.165526676 0.001367989  (4372 a 5242] 0.186046512 0.001367989  (5242 a 6112] 0.105335157 0.005471956  (6112 a 6982] 0.102599179 0.002735978  (6982 a 7851] 0.105335157 0.001367989  (7851 a 8720] 0.023255814 0.000000000  Distribuição percentual do número de alugueis em feriados não feriados. |

#Teste Qui-quadrado

|  |
| --- |
| # Entrar com os dados da tabela  x<-matrix(c(temp, atemp, hum, windspeed, casual, registered, cnt), nc=2);x  chisq.test(x)  Pearson's Chi-squared test  data: x  X-squared = 703918, df = 2558, p-value < 2.2e-16  Há uma boa correlação entre as variáveis. |

Entrega do exercício no formato apresentação word.

Data de entrega: 10/03/2020

Regina Bernal

03/03/3030