**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.

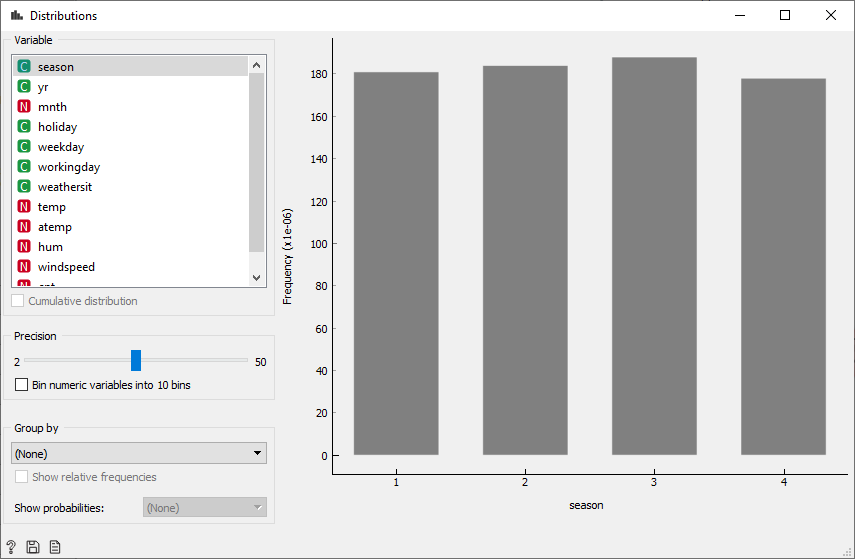
Fonte de dados: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

1. Classifique o tipo de variável

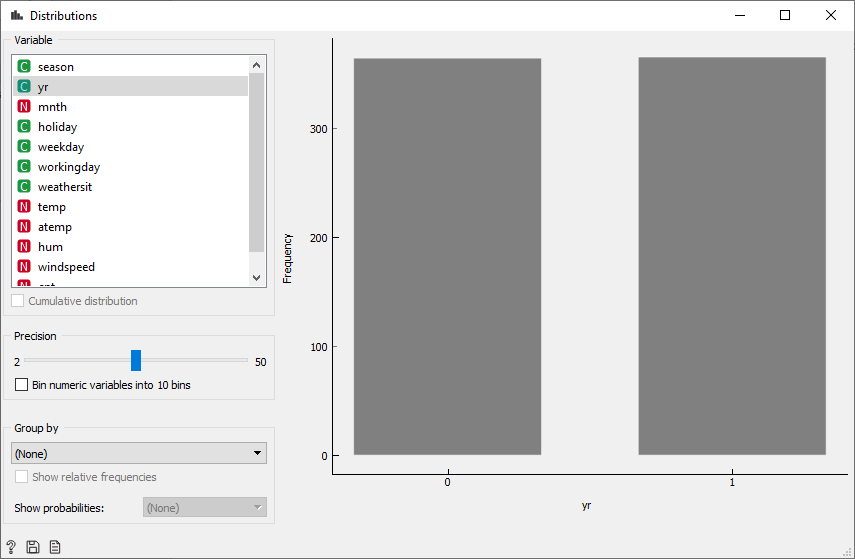
|  |  |  |
| --- | --- | --- |
| **Variável** | **Descrição** | **Tipo de Variável** |
| **instant** | record index | ID |
| **dteday** | date | ID |
| **season** | season (1:springer, 2:summer, 3:fall, 4:winter) | Qualitativa ordinal |
| **yr** | year (0: 2011, 1:2012) | Qualitativa nominal |
| **mnth** | 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:
2. Objetivo do estudo:
3. Apresente as tabelas descritivas e gráficos das variáveis. Interprete os resultados.

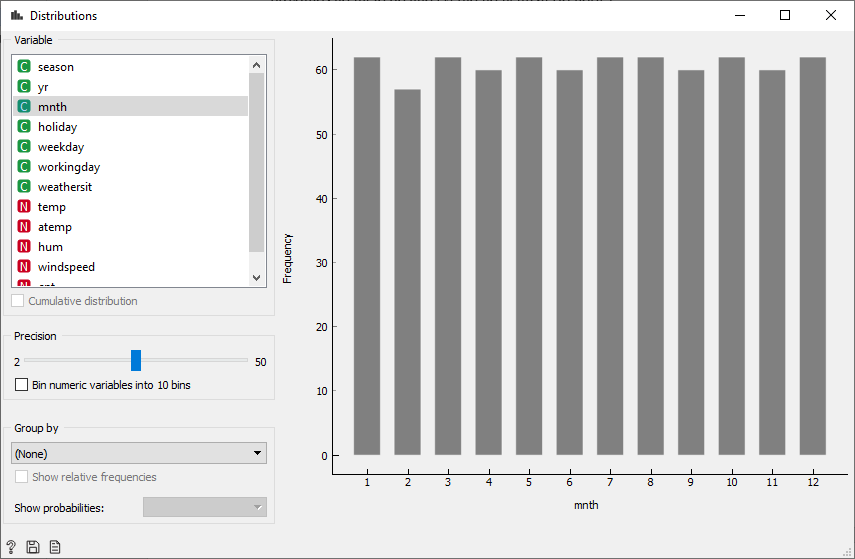
**season (estação do ano):** o gráfico demonstra o número de dias em cada estação do ano, os valores são aproximados.



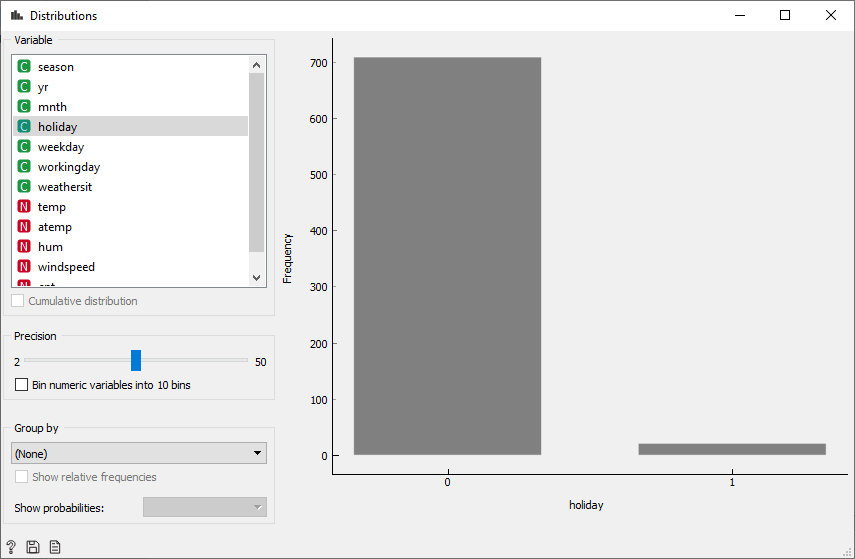
**yr (ano):** o gráfico demonstra o número de dias em cada estação do ano, os valores são aproximados, com apenas 1 a mais no yr=1 por ser ano bissexto.



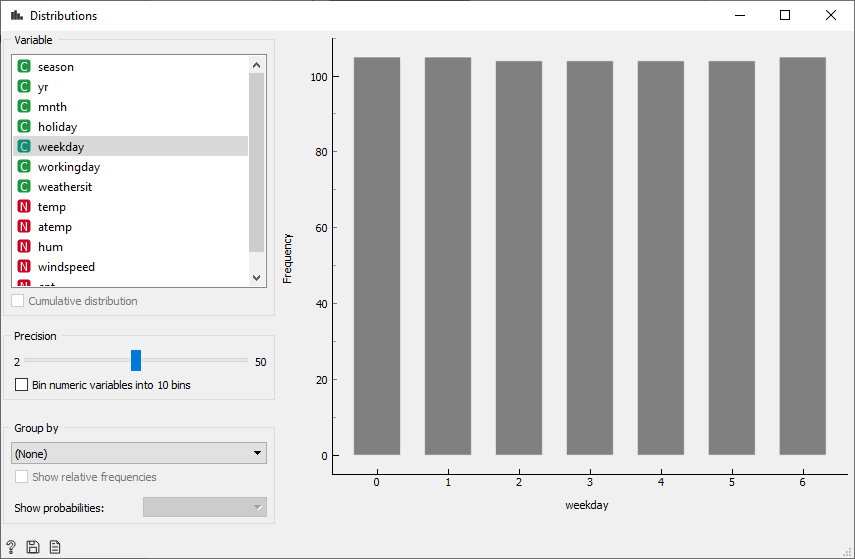
**mnth (mês):** o gráfico demonstra o número de dias em cada mês, os valores são aproximados com o mês 2 um pouco menor do que os outros.



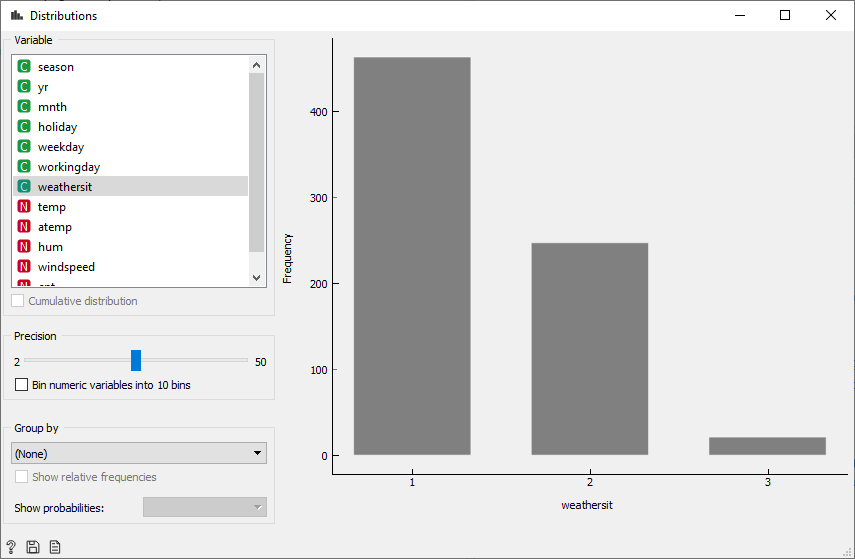
**holiday (feriados):** o gráfico demonstra o número de dias com feriados (holiday=1), a diferença é muito grande para dias sem feriado (holiday=0).



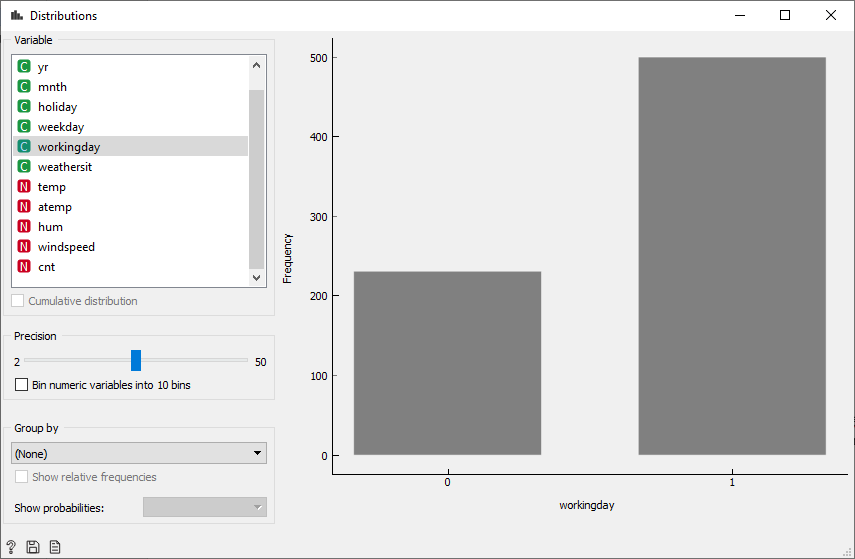
**weekday (dia da semana):** o gráfico demonstra o número de dias em cada dia da semana, os números são quase equivalentes.



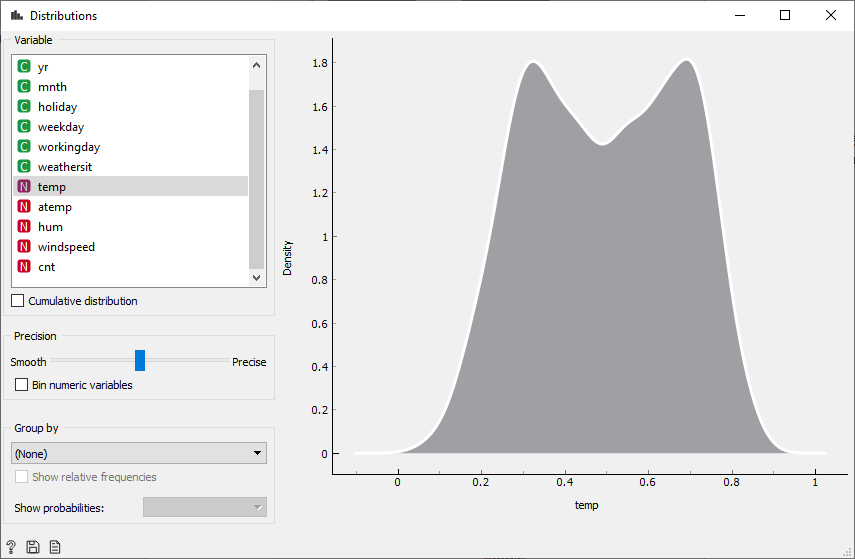
**weathersit (clima):** o gráfico demonstra o número de dias de acordo com o clima (ensolarado=1, nublado=2, chuva ou neve=3).



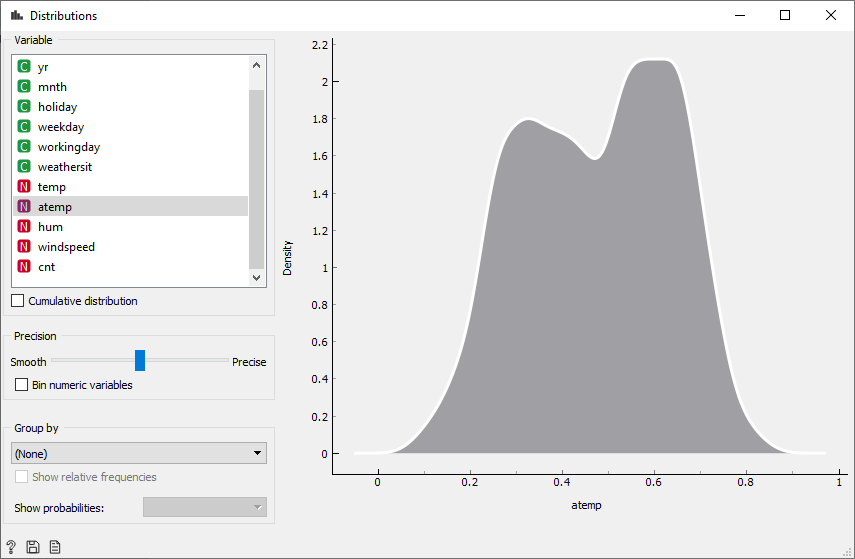
**workday (dia de trabalho):** o gráfico demonstra o número de dias de trabalho (que não não nem finais de semana e nem feriados, workday=1),os dias de trabalho são mais do que o dobro dos dias de sem trabalho.



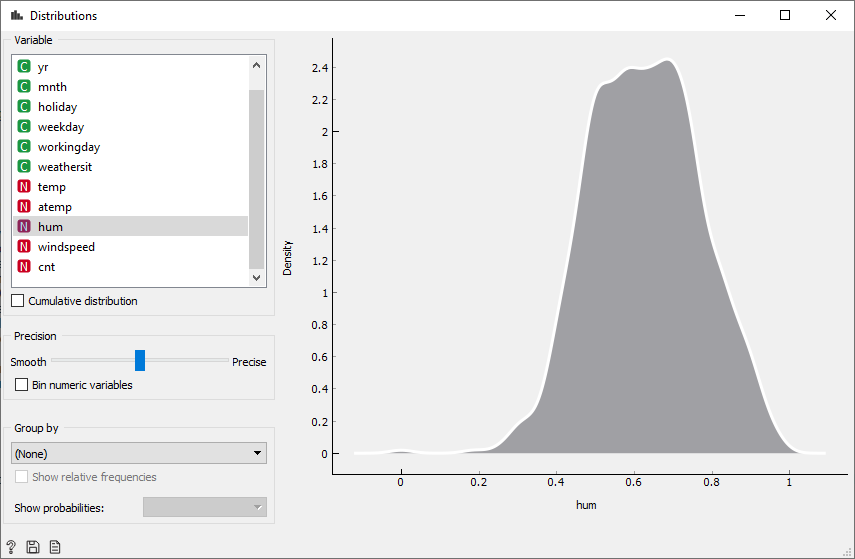
**temp (temperatura):** o gráfico exibe a densidade da distribuição da temperatura, as temperaturas muito baixas e muito altas são menos frequentes.



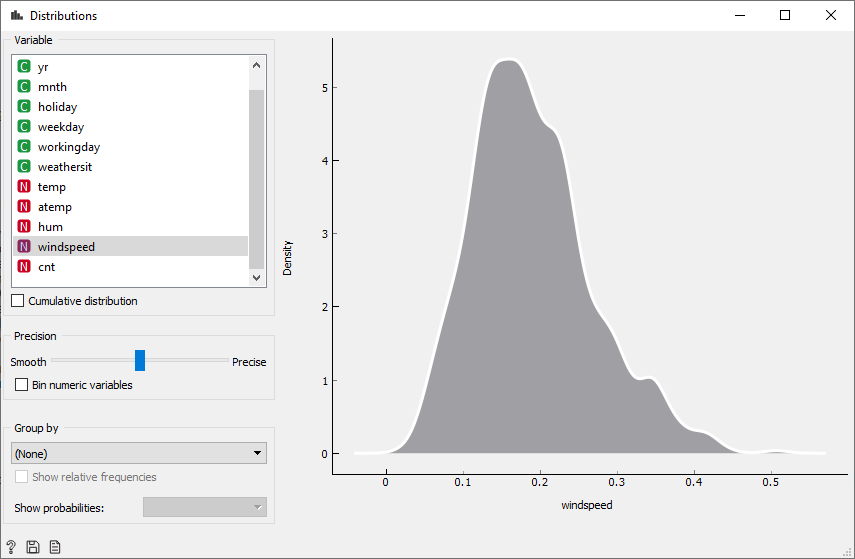
**atemp (sensação termica):** o gráfico exibe a densidade da distribuição da sensação da temperatura, as temperaturas muito baixas e muito altas são menos frequentes.



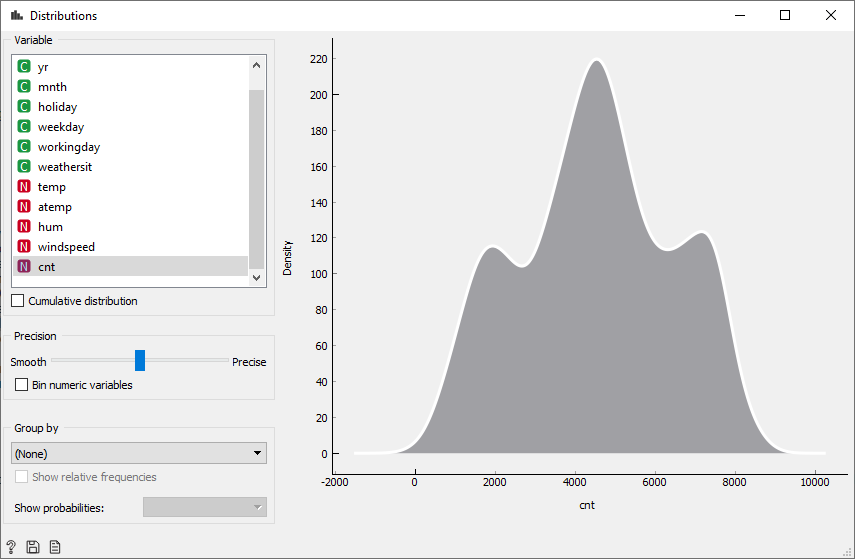
**hum (humidade):** o gráfico exibe a densidade da distribuição da humidade, as humidades muito baixas e muito altas são menos frequentes.



**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. 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. 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 até 100.  
 **- windspeed:** normalização da velocidade do vento. Os valores foram divididos até 67.

1. 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?

**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)

# nomes das variáveis

names(arquivo)

# visualizar primeiras 20 linhas

head(arquivo,20)

## Análise descritiva univariada

# trabalhar com as variáveis

attach(arquivo)

**# selecionar as variáveis quantitativas**

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

**# Medidas resumo**

summary(dadosquant)

**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:** |

# Calcular o desvio padrão da variável CNT

dp<-sd(cnt)

print(dp)

|  |
| --- |
| **Interprete o resultado:** |

# Calcular o coeficiente de variação da variável CNT

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

print(cv)

|  |
| --- |
| **Interprete o resultado:** |

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

matriz$P

matriz$n

#plot

corrplot(matriz$r,p,mat=matriz$P,sig.level=0.005)

corrplot(m$r,p,mat=matriz$P,sig.level=0.005, method=”number”, type=”upper”)

# gráfico de dispersão

plot(cnt~temp)

# pacote correlação

cor(cnt,temp)

# teste estatístico

cor.test(cnt,temp,method= “pearson”)

# 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) |

# Frequência relativa

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

|  |
| --- |
|  |

#Teste Qui-quadrado

|  |
| --- |
| # Entrar com os dados da tabela  x<-matrix(c(, , , , , ,),nc=2);x  chisq.test(x) |

Entrega do exercício no formato apresentação word.

Data de entrega: 10/03/2020

Regina Bernal

03/03/3030