Trabalho de Estatística 01 - Análise Exploratória de Dados (AED)

Pedro Manoel (119210706), Lucas Lima (119210396), Felipe Oliveira (119210148)

05/12/2022

Sumário

1. Contexto

Uma loja de varejo localizada nos Estados Unidos (EUA), decidiu realizar um estudo estatístico sobre as suas vendas no período de 2015 a 2018, com o objetivo de utilizar esse estudo como apoio na tomada de decisões executivas/comerciáis para o ano de 2019. Com isso, ela elencou alguns dados especificos que gostaria de ter no referido estudo, bem como algumas hipóteses que ela gostaria que fosse verificada e documentada.

Os dados específicos elencados pela loja foram:

- Qual é o total de vendas por dia, mês e ano
- Qual é a média e o desvio padrão das vendas por dia, mês e ano
- Qual é o tipo de envio mais utilizado nas vendas por mês e ano
- Quais as dez cidades e estados que possuem o maior valor de vendas e qual foi a categoria de produto mais vendida

As hipóteses levantadas pela loja foram:

- (1) As vendas mensais da categoria Technology são maiores que as da categoria Office Supplies
- (2) O valor das vendas mensais no ano de 2018 pode ser considerada proviniente de uma distribuição Normal
- (3) O estado onde a venda foi realizada afeta o seu valor

2. Levantamentos do estudo estatístico

2.1. População

- **Qual:** Todas as vendas da super loja nos anos de 2015 a 2018
- **Tipo:** Finita

2.2. Variáveis

- Data da venda (Dia, Mês, Ano)
  - **Tipo:** Qualitativa Ordinal
- Tipo de envio
  - **Tipo:** Qualitativa Nominal
- Estado da venda
  - **Tipo:** Qualitativa Nominal
- Cidade da venda
  - **Tipo:** Qualitativa Nominal
- Categoria do produto
  - **Tipo:** Quantitativa Nominal
- Valor da venda
  - **Tipo:** Quantitativa Contínua

2.3. Objetivos

- **Geral:** Tem-se como objetivo geral a utilização do estudo estatístico para a tomada de decisões executivas/comerciáis para o ano de 2019 da loja.
- **Específicos:** Tem-se como objetivos específicos o levantamento de dados elencados pela loja e a verificação e documentação de hipóteses tembém levantados pela loja.

2.4. Foi necessário a coleta de uma amostra?

Não. O presente estudo foi realizado com a população por completa, ou seja, utilizou-se de um censo no lugar de uma amostra. Dessa forma, toda e qualquer informação/dado gerado pelo presente estudo extraído da população é um parâmetro.

Contudo, para os testes de hipóteses assumimos que os dados são provenientes de uma amostra e que para todo teste de hipótese temos como nível de significância 0.05

3. Tratamentos da base de dados

3.1. Instalando bibliotecas/pacotes necessários

```
library(lubridate)
library(dplyr)
library(DataExplorer)
library(ggplot2)
library(nortest)
```

Paleta de cores

```
Backgroundv="#282a36"
Current_Line="#44475a"
Foreground="#f8f8f2"
Comment="#6272a4"
Cyan="#8be9fd"
Green="#50fa7b"
Orange="#ffb86c"
Pink="#ff79c6"
Purple="#bd93f9"
Red="#ff5555"
Yellow="#f1fa8c"
```

3.2. Carregando base de dados

```
dados = read.csv("datasets/data.csv")
```

3.3. Visualizando informações sobre a base de dados

Número de linhas e colunas

```r
dim(dados)
```

```
[1] 9800   18
```

Como pode ser visualizado acima temos uma população de tamanho 9800 e 18 varíaveis

Nome das colunas

```r
names(dados)
```

```
 [1] "Row.ID"        "Order.ID"      "Order.Date"    "Ship.Date"
 [5] "Ship.Mode"     "Customer.ID"   "Customer.Name" "Segment"
 [9] "Country"       "City"          "State"         "Postal.Code"
[13] "Region"        "Product.ID"    "Category"      "Sub.Category"
[17] "Product.Name"  "Sales"
```

Dessas colunas(variáveis) vamos utilizar somente

- **Order.Date** (Data da venda)
- **Ship.Mode** (Tipo de envio)
- **State** (Estado da venda)
- **City** (Cidade da venda)
- **Category** (Categoria do produto)
- **Sales** (Valor da venda)

3.4. Manipulando base de dados para o estudo

Selecionandos as variáveis de interesse

```r
dados =
  dados %>%
  select(Order.Date, Ship.Mode, State, City, Category, Sales)

knitr::kable(head(dados), caption="Primeiras linhas da base de dados")
```

Table 1: Primeiras linhas da base de dados

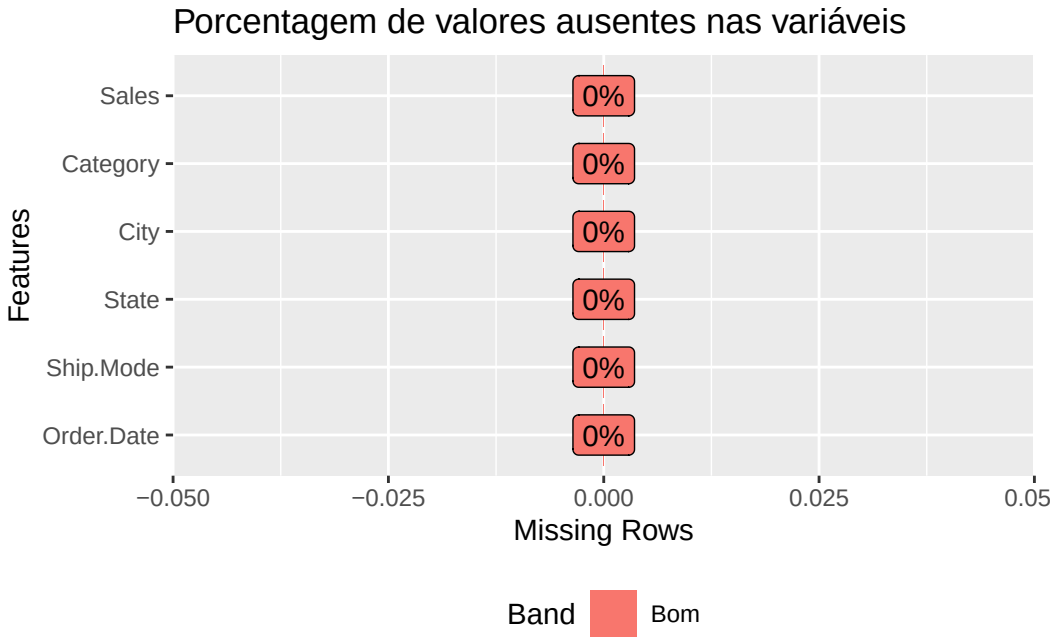
Order.Date	Ship.Mode	State	City	Category	Sales
08/11/2017	Second Class	Kentucky	Henderson	Furniture	261.9600
08/11/2017	Second Class	Kentucky	Henderson	Furniture	731.9400
12/06/2017	Second Class	California	Los Angeles	Office Supplies	14.6200
11/10/2016	Standard Class	Florida	Fort Lauderdale	Furniture	957.5775
11/10/2016	Standard Class	Florida	Fort Lauderdale	Office Supplies	22.3680
09/06/2015	Standard Class	California	Los Angeles	Furniture	48.8600

Buscando por valores ausentes (NAs, zeros, etc) com o pacote "DataExplorer"

Detalhes

- O parâmetro **group** da função categoriza a variável de acordo com os limites superiores estipulados

```
plot_missing(dados,
             title="Porcentagem de valores ausentes nas variáveis",
             group = list(Bom = 0.05, Ok = 0.4, Ruim = 0.8, Remover = 1))
```



Como pode ser visualizado acima não existe valores ausentes na base de dados para as variáveis utilizadas.

Transformação dos dados

Para melhorar o processamento e facilitar nas operações estátisticas, devemos separar a coluna **Order.Data** em três novas colunas **Day**, **Month** e **Year**

```
dados =
  dados %>%
  mutate(
    Day = day(dmy(Order.Date)),
    Month = month(dmy(Order.Date), label = TRUE),
    Year = year(dmy(Order.Date)),
    Order.Date = NULL
  )

knitr::kable(head(dados), caption="Primeiras linhas da base de dados")
```

Table 2: Primeiras linhas da base de dados

Ship.Mode	State	City	Category	Sales	Day	Month	Year
Second Class	Kentucky	Henderson	Furniture	261.9600	8	nov	2017
Second Class	Kentucky	Henderson	Furniture	731.9400	8	nov	2017
Second Class	California	Los Angeles	Office Supplies	14.6200	12	jun	2017
Standard Class	Florida	Fort Lauderdale	Furniture	957.5775	11	out	2016
Standard Class	Florida	Fort Lauderdale	Office Supplies	22.3680	11	out	2016
Standard Class	California	Los Angeles	Furniture	48.8600	9	jun	2015

4. Calculando o total de vendas por dia, mês e ano

4.1. Por dia

```
soma_dia =
  dados %>%
  group_by(Day) %>%
  summarise(Sum = sum(Sales))
```

Tabela

```
knitr::kable(soma_dia, caption="Soma das vendas por dia")
```

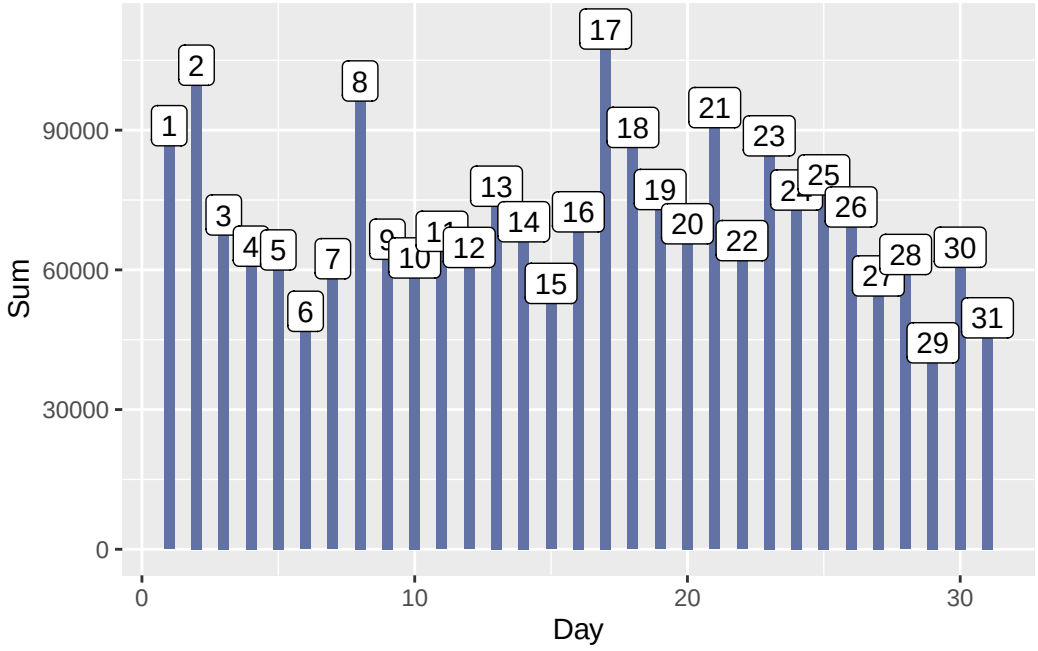
Table 1: Soma das vendas por dia

Day	Sum
1	90957.98
2	103951.16
3	71788.69
4	65094.95
5	64277.08
6	51099.14
7	62393.95
8	100560.35
9	66582.37
10	62563.78
11	68274.31
12	64819.37
13	77964.33
14	70330.92
15	57091.81
16	72464.27
17	111719.93

Day	Sum
18	90587.30
19	77253.76
20	69898.69
21	94842.87
22	66279.49
23	88779.63
24	77146.85
25	80400.94
26	73507.62
27	58700.33
28	63308.70
29	44334.43
30	64858.27
31	49703.52

Gráfico

```
ggplot(soma_dia, aes(x=Day, y=Sum, label=Day)) +
geom_bar(stat = "identity", width = 0.4, fill = Comment) +
geom_label()
```



4.2. Por mês

```
soma_mes =
  dados %>%
  group_by(Month) %>%
  summarise(Sum = sum(Sales))
```

Tabela

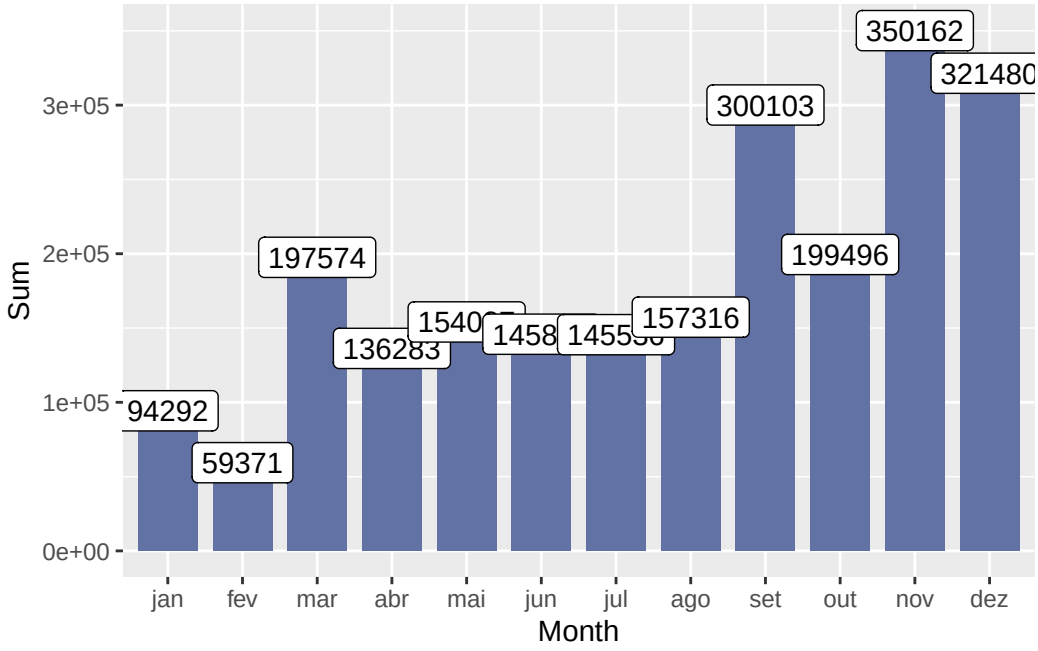
```
knitr::kable(soma_mes, caption="Soma das vendas por mês")
```

Table 2: Soma das vendas por mês

Month	Sum
jan	94291.63
fev	59371.12
mar	197573.59
abr	136283.00
mai	154086.72
jun	145837.52
jul	145535.69
ago	157315.93
set	300103.41
out	199496.29
nov	350161.71
dez	321480.17

Gráfico

```
ggplot(soma_mes, aes(x=Month, y=Sum, label=round(Sum))) +
geom_bar(stat = "identity", width = 0.8, fill = Comment) +
geom_label()
```



4.3. Por ano

```
soma_ano =
  dados %>%
  group_by(Year) %>%
  summarise(Sum = sum(Sales))
```

Tabela

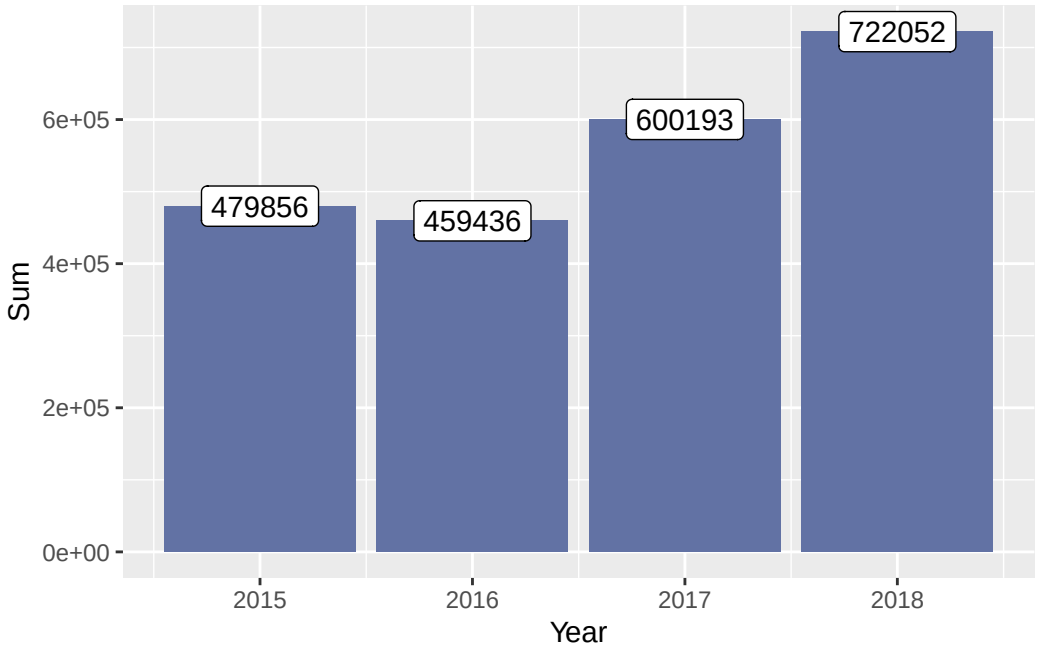
```
knitr::kable(soma_ano, caption="Soma das vendas por ano")
```

Table 3: Soma das vendas por ano

Year	Sum
2015	479856.2
2016	459436.0
2017	600192.6
2018	722052.0

Gráfico

```
ggplot(soma_ano, aes(x=Year, y=Sum, label=round(Sum))) +
geom_bar(stat = "identity", fill = Comment) +
geom_label()
```



5. Calculando a média e o desvio padrão das vendas por dia, mês e ano

5.1. Média

5.1.1. Por dia

```
media_dia =
  dados %>%
  group_by(Day) %>%
  summarise(Mean = mean(Sales))
```

Tabela

```
knitr::kable(media_dia, caption="Média das vendas por dia")
```

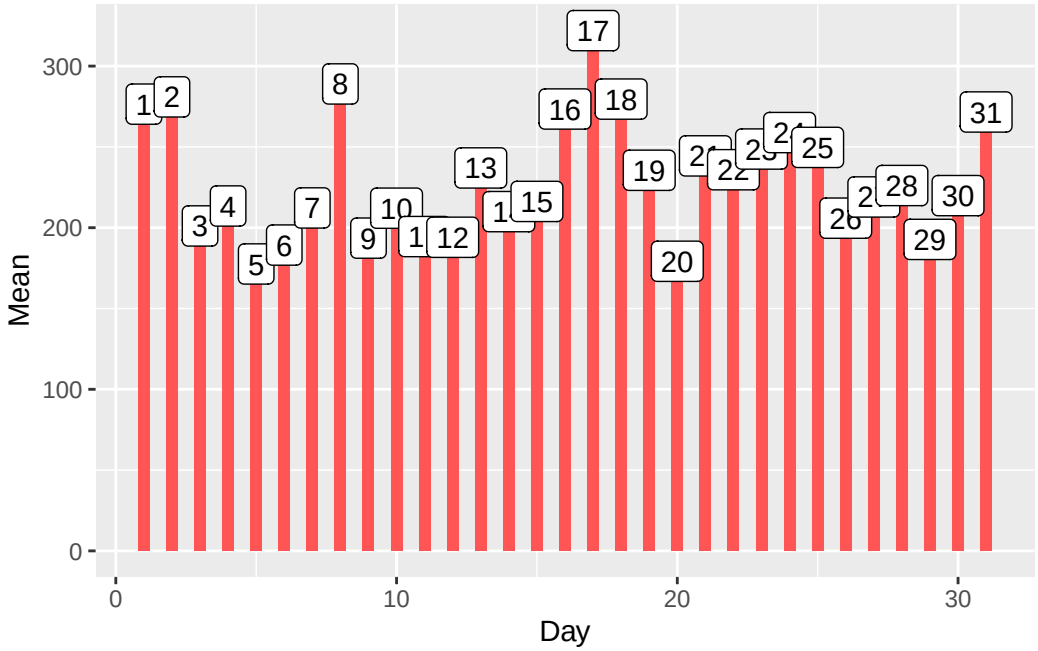
Table 1: Média das vendas por dia

Day	Mean
1	276.4680
2	280.9491
3	201.0888
4	213.4261
5	177.5610
6	189.2561
7	212.2243
8	288.9665
9	193.5534
10	212.0806
11	194.5137
12	194.0700
13	236.9736
14	209.9430
15	216.2568

Day	Mean
16	273.4501
17	321.9595
18	279.5904
19	235.5298
20	179.2274
21	244.4404
22	235.8701
23	248.6824
24	258.8821
25	249.6924
26	205.9037
27	219.0311
28	226.1025
29	192.7584
30	219.8585
31	271.6040

Gráfico

```
ggplot(media_dia, aes(x=Day, y=Mean, label=Day)) +
geom_bar(stat = "identity", width = 0.4, fill = Red) +
geom_label()
```



5.1.2. Por mês

```
media_mes =
  dados %>%
  group_by(Month) %>%
  summarise(Mean = mean(Sales))
```

Tabela

```
knitr::kable(media_mes, caption="Média das vendas por mês")
```

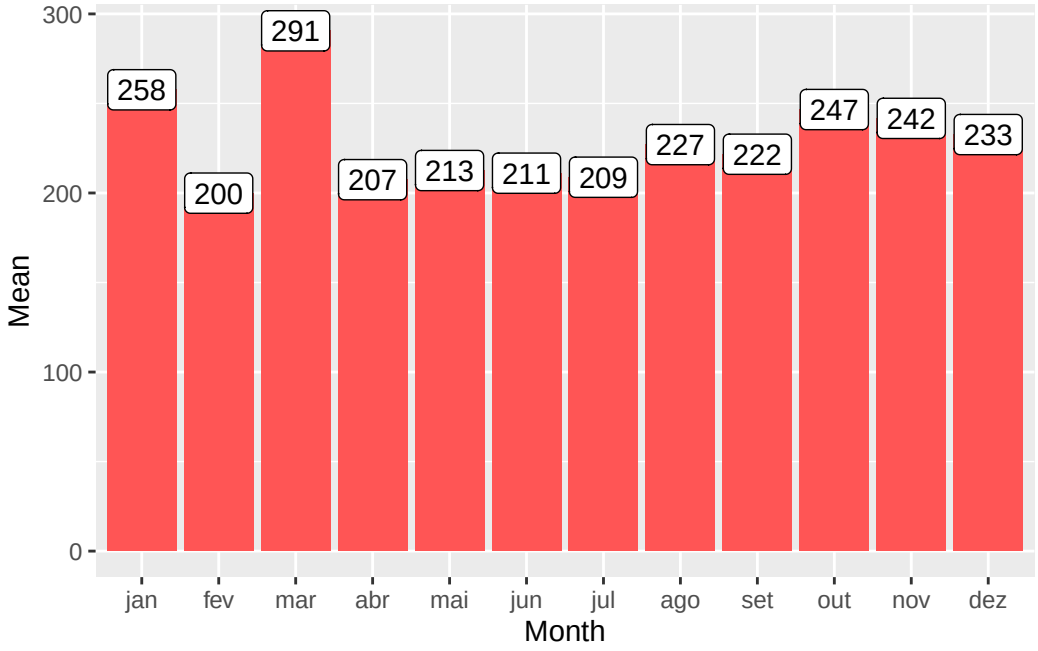
Table 2: Média das vendas por mês

Month	Mean
jan	257.6274
fev	199.9027
mar	290.5494
abr	207.4323
mai	212.5334
jun	211.0529
jul	208.8030

Month	Mean
ago	227.0071
set	221.6421
out	246.5962
nov	241.6575
dez	232.6195

Gráfico

```
ggplot(media_mes, aes(x=Month, y=Mean, label=round(Mean))) +
geom_bar(stat = 'identity', fill = Red) +
geom_label()
```



5.1.3. Por ano

```
media_ano =
  dados %>%
  group_by(Year) %>%
  summarise(Mean = mean(Sales))
```

Tabela

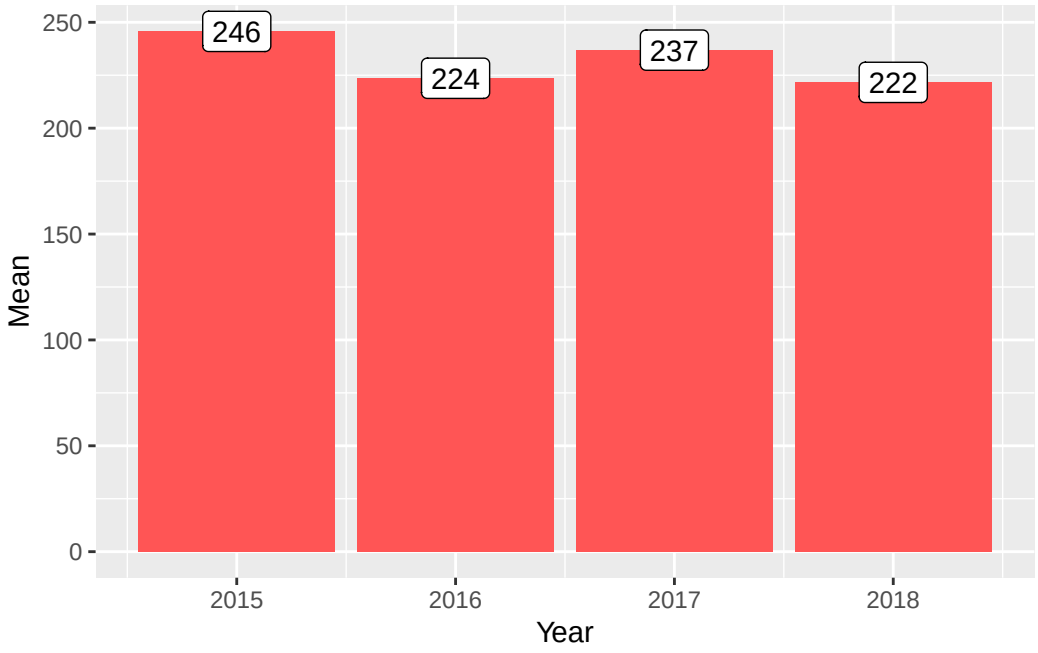
```
knitr::kable(head(media_ano), caption="Média das vendas por ano")
```

Table 3: Média das vendas por ano

Year	Mean
2015	245.7021
2016	223.5698
2017	236.8558
2018	221.6243

Gráfico

```
ggplot(media_ano, aes(x=Year, y=Mean, label=round(Mean))) +
geom_bar(stat = 'identity', fill = Red) +
geom_label()
```



5.2. Desvio padrão

5.2.1. Por dia

```
sd_dia =
  dados %>%
  group_by(Day) %>%
  summarise(Sd = sd(Sales))
```

Tabela

```
knitr::kable(sd_dia, caption="Desvio padrão das vendas por dia")
```

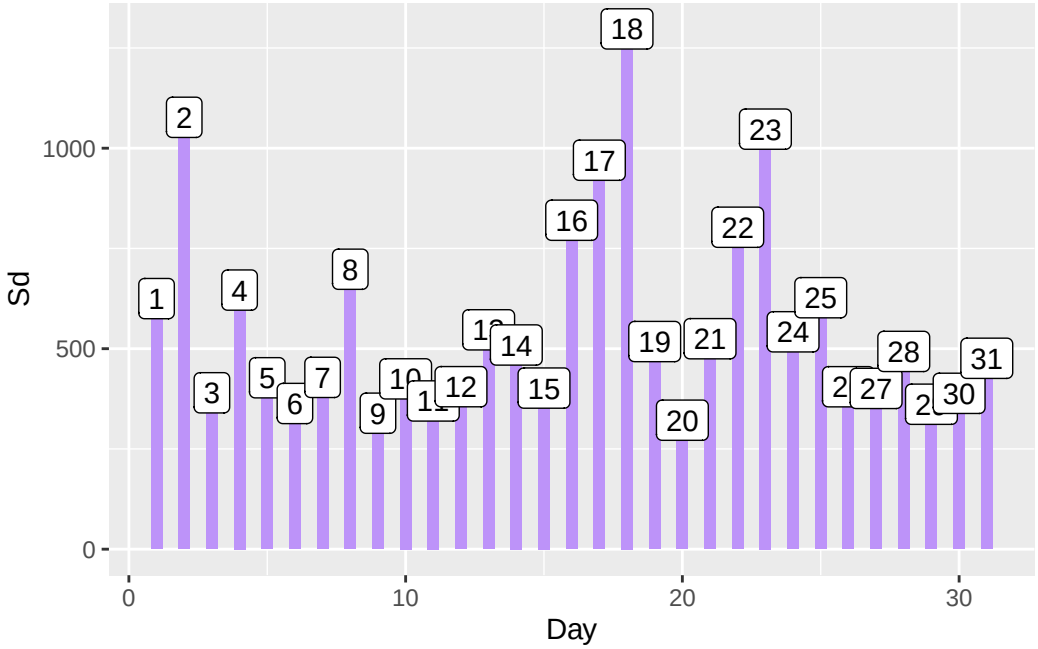
Table 4: Desvio padrão das vendas por dia

Day	Sd
1	624.9360
2	1076.9867
3	390.0055
4	646.1560
5	427.4781
6	363.3357
7	428.4560
8	698.1524
9	339.1020
10	424.7524
11	369.6554
12	405.2767
13	547.0640
14	509.2610
15	401.7606
16	820.3548
17	970.0272
18	1297.4940
19	517.8450
20	321.6795
21	524.8960
22	802.8406
23	1047.5200
24	542.1035
25	627.1615

Day	Sd
26	405.1301
27	400.9207
28	493.1791
29	361.4720
30	388.6582
31	476.0632

Gráfico

```
ggplot(sd_dia, aes(x=Day, y=Sd, label=Day)) +
geom_bar(stat = "identity", width = 0.4, fill = Purple) +
geom_label()
```



5.2.2. Por mês

```
sd_mes =
  dados %>%
  group_by(Month) %>%
  summarise(Sd = sd(Sales))
```

Tabela

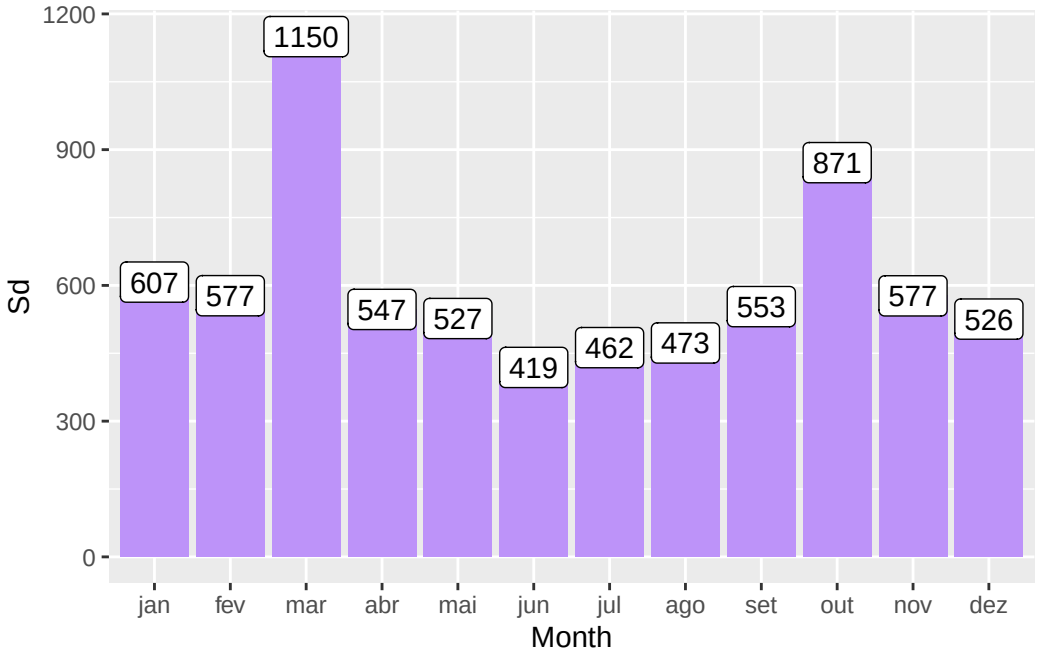
```
knitr::kable(sd_mes, caption="Desvio padrão das vendas por mês")
```

Table 5: Desvio padrão das vendas por mês

Month	Sd
jan	607.3252
fev	577.2335
mar	1149.8230
abr	546.8830
mai	526.8144
jun	418.7019
jul	462.4673
ago	472.7973
set	553.1230
out	871.2893
nov	576.9116
dez	525.5582

Gráfico

```
ggplot(sd_mes, aes(x=Month, y=Sd, label=round(Sd))) +
geom_bar(stat = "identity", fill = Purple) +
geom_label()
```



5.2.3. Por ano

```
sd_ano =
  dados %>%
  group_by(Year) %>%
  summarise(Sd = sd(Sales))
```

Tabela

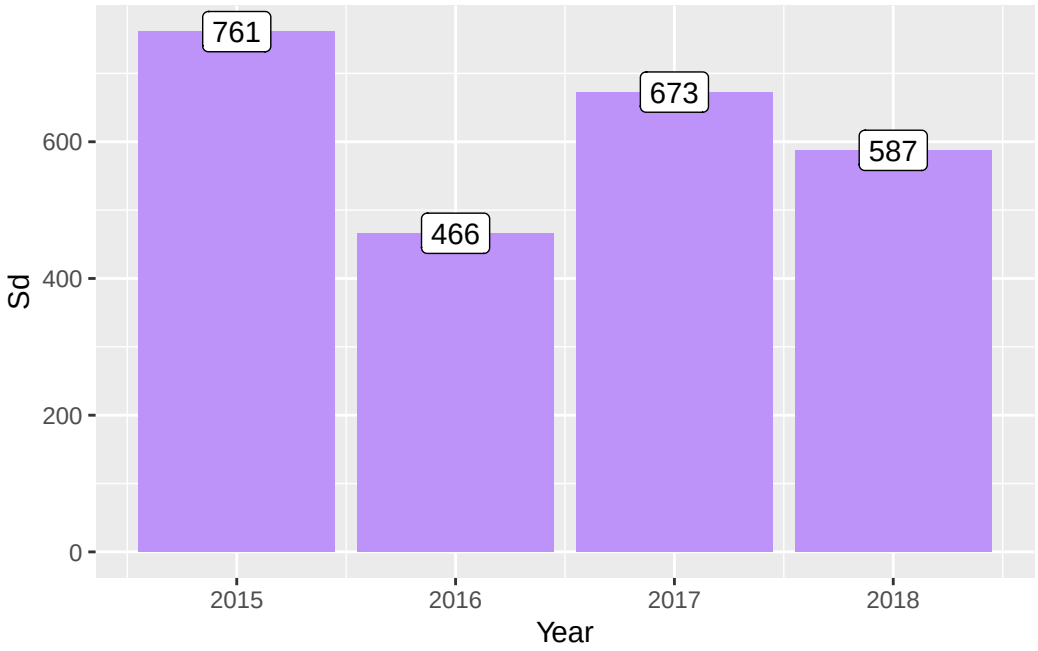
```
knitr::kable(sd_ano, caption="Desvio padrão das vendas por ano")
```

Table 6: Desvio padrão das vendas por ano

Year	Sd
2015	761.1221
2016	466.1155
2017	672.6523
2018	587.4377

Gráfico

```
ggplot(sd_ano, aes(x=Year, y=Sd, label=round(Sd))) +
geom_bar(stat = "identity", fill = Purple) +
geom_label()
```



6. Verificando o tipo de envio mais utilizado nas vendas por mês e ano

6.1. Por mês

Total de envios de cada tipo por mês

```
envio_mes =
  dados %>%
  group_by(Month, Ship.Mode) %>%
  summarise(Count = n(), .groups = 'drop_last')

knitr::kable(envio_mes, caption="Total de envios de cada tipo por mês")
```

Table 1: Total de envios de cada tipo por mês

Month	Ship.Mode	Count
jan	First Class	40
jan	Same Day	4
jan	Second Class	81
jan	Standard Class	241
fev	First Class	46
fev	Same Day	14
fev	Second Class	62
fev	Standard Class	175
mar	First Class	100
mar	Same Day	54
mar	Second Class	152
mar	Standard Class	374
abr	First Class	104
abr	Same Day	18
abr	Second Class	111
abr	Standard Class	424
mai	First Class	117

Month	Ship.Mode	Count
mai	Same Day	47
mai	Second Class	109
mai	Standard Class	452
jun	First Class	127
jun	Same Day	29
jun	Second Class	136
jun	Standard Class	399
jul	First Class	98
jul	Same Day	45
jul	Second Class	134
jul	Standard Class	420
ago	First Class	106
ago	Same Day	34
ago	Second Class	135
ago	Standard Class	418
set	First Class	205
set	Same Day	84
set	Second Class	254
set	Standard Class	811
out	First Class	140
out	Same Day	65
out	Second Class	148
out	Standard Class	456
nov	First Class	204
nov	Same Day	72
nov	Second Class	297
nov	Standard Class	876
dez	First Class	214
dez	Same Day	72
dez	Second Class	283
dez	Standard Class	813

Envio mais utilizado por mês

```
envio_mes =
  dados %>%
  group_by(Month, Ship.Mode) %>%
  summarise(Count = n(), .groups = 'drop_last') %>%
  filter(Count == max(Count))
```

```r
knitr::kable(envio_mes, caption="Envio mais utilizado por mês")
```

Table 2: Envio mais utilizado por mês

Month	Ship.Mode	Count
jan	Standard Class	241
fev	Standard Class	175
mar	Standard Class	374
abr	Standard Class	424
mai	Standard Class	452
jun	Standard Class	399
jul	Standard Class	420
ago	Standard Class	418
set	Standard Class	811
out	Standard Class	456
nov	Standard Class	876
dez	Standard Class	813

6.2. Por ano

Total de envios de cada tipo por ano

```r
envio_ano =
  dados %>%
  group_by(Year, Ship.Mode) %>%
  summarise(Count = n(),.groups = 'drop_last')

knitr::kable(envio_ano, caption="Total de envios de cada tipo por ano")
```

Table 3: Total de envios de cada tipo por ano

Year	Ship.Mode	Count
2015	First Class	282
2015	Same Day	90
2015	Second Class	374
2015	Standard Class	1207
2016	First Class	272
2016	Same Day	108
2016	Second Class	406

Year	Ship.Mode	Count
2016	Standard Class	1269
2017	First Class	381
2017	Same Day	156
2017	Second Class	480
2017	Standard Class	1517
2018	First Class	566
2018	Same Day	184
2018	Second Class	642
2018	Standard Class	1866

Envio mais utilizado por ano

```
envio_ano =
  dados %>%
  group_by(Year, Ship.Mode) %>%
  summarise(Count = n(), .groups = 'drop_last') %>%
  filter(Count == max(Count))

knitr::kable(envio_ano, caption="Envio mais utilizado por ano")
```

Table 4: Envio mais utilizado por ano

Year	Ship.Mode	Count
2015	Standard Class	1207
2016	Standard Class	1269
2017	Standard Class	1517
2018	Standard Class	1866

7. Verificando as dez cidades e estados que possuem o maior valor de vendas e qual foi a categoria de produto mais vendida

7.1. Por estado

```
compras_estado =
  dados %>%
  group_by(State, Category) %>%
  summarise(Sum=sum(Sales), .groups = 'drop_last') %>%
  filter(Sum == max(Sum)) %>%
  arrange(desc(Sum)) %>%
  head(10)

knitr::kable(compras_estado, caption="10 estados com maiores vendas")
```

Table 1: 10 estados com maiores vendas

State	Category	Sum
California	Technology	154684.18
New York	Technology	126902.69
Texas	Technology	64656.27
Washington	Technology	50536.71
Florida	Technology	46968.04
Pennsylvania	Technology	42064.07
Michigan	Office Supplies	37688.20
Ohio	Technology	34550.55
Illinois	Technology	31637.88
Georgia	Office Supplies	26397.78

7.2. Por cidade

```
compras_cidade =
  dados %>%
  group_by(City, Category) %>%
  summarise(Sum=sum(Sales), .groups = 'drop_last') %>%
  filter(Sum == max(Sum)) %>%
  arrange(desc(Sum)) %>%
  head(10)

knitr::kable(compras_cidade, caption="10 cidades com maiores vendas")
```

Table 2: 10 cidades com maiores vendas

City	Category	Sum
New York City	Technology	108980.14
Los Angeles	Technology	72249.71
Seattle	Technology	42506.37
San Francisco	Office Supplies	41918.45
Philadelphia	Technology	41772.12
Jacksonville	Technology	29188.35
Houston	Technology	24951.98
Chicago	Technology	21838.30
San Diego	Furniture	20031.24
Newark	Technology	18485.69

8. Verificando as hipóteses leventadas

8.1. 1º Hipótese

> As vendas mensais da categoria Technology são maiores que as da categoria Office Supplies

Assumindo que

μ_1 É a média de vendas mensais da categoria Technology

μ_2 É a média de vendas mensais da categoria Office Supplies

Temos:

$h0 : \mu_1 = \mu_2$ (Hipótese nula)

$h1 : \mu_1 > \mu_2$ (Hipótese alternativa)

Testando a normalidade das vendas mensais da categoria Technology

```
vendas_cat_tec =
  (
    dados %>%
    filter(Category == "Technology") %>%
    group_by(Month) %>%
    summarise(Sales = sum(Sales))
  )$Sales

shapiro.test(vendas_cat_tec)
```

```
    Shapiro-Wilk normality test

data:  vendas_cat_tec
W = 0.94431, p-value = 0.5559
```

Como p-value = 0.5559 e 0.5559 > 0.05 então há evidência estátistica para comprovar que as vendas mensais da categoria Technology seguem uma distribuição normal

Testando a normalidade das vendas mensais da categoria Office Supplies

```
vendas_cat_off =
  (
    dados %>%
    filter(Category == "Office Supplies") %>%
    group_by(Month) %>%
    summarise(Sales = sum(Sales))
  )$Sales

shapiro.test(vendas_cat_off)
```

```
    Shapiro-Wilk normality test

data:  vendas_cat_off
W = 0.86555, p-value = 0.0574
```

Como p-value = 0.0574 e 0.0574 > 0.05 então há evidência estátistica para comprovar que as vendas mensais da categoria Office Supplies seguem uma distribuição normal

Como ambas as vendas seguem uma distribuição normal podemos aplicar o test t de Student

Executando o teste t de Student

```
result = t.test(vendas_cat_tec, vendas_cat_off, alternative = "greater")
```

Utilizamos a função t.test do R, especificando a opção "alternative = 'greater' " para indicar que estamos interessado em testar a hipótese de que a média das vendas mensais da Categoria Technology é maior que as da Categoria Office Supplies. O resultado da função incluira o p-valor, que é uma medida da evidência contra a hipótese nula. Se o p-valor for menor do que o nível de significância estabelecido previamente (0.05), podemos concluir com um nível de confiança adequado que as vendas da Categoria Technology são maiores que as vendas da Categoria Office Supplies. Se o p-valor for maior do que o nível de significância, não podemos rejeitar a hipótese nula e devemos concluir que as vendas das duas categorias são iguais.

Observando o resultado e definindo uma conclusão

```
result
```

```
    Welch Two Sample t-test

data:  vendas_cat_tec and vendas_cat_off
t = 0.83259, df = 21.284, p-value = 0.2072
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -10835.18       Inf
sample estimates:
mean of x mean of y
 68954.66  58785.19
```

Resultados:

O resultado do teste mostra que o p-valor é de 0.2072, o que significa que não há evidências suficientes para rejeitar a hipótese nula. Isso é reforçado pelo intervalo de confiança de 95% que inclui o valor zero (-10835,18 a Inf). Com base nesse resultado, não podemos concluir com certeza que a média das vendas mensais da categoria Technology é maior que a média das vendas mensais da categoria Office Supplies.

8.2. 2º Hipótese

> O valor das vendas mensais no ano de 2018 pode ser considerada proveniente de uma distribuição Normal.

Assumindo que X é o valor das vendas mensais, temos:

$h0 : X = N\left(\mu,\ \sigma^2\right)$ (Hipótese nula)

$h1 : X \neq N\left(\mu,\ \sigma^2\right)$ (Hipótese alternativa)

Calculando o valor das vendas mensais do ano de 2018

```r
vendas_mes_2018 =
  dados %>%
  filter(Year == '2018') %>%
  group_by(Month) %>%
  summarise(Sum = sum(Sales))

knitr::kable(vendas_mes_2018, caption="Vendas mensais em 2018")
```

Table 1: Vendas mensais em 2018

Month	Sum
jan	43476.47
fev	19921.00
mar	58863.41
abr	35541.91
mai	43825.98
jun	48190.73
jul	44825.10
ago	62837.85
set	86152.89
out	77448.13
nov	117938.15
dez	83030.39

Executando o teste Shapiro-Wilk

```
shapiro.test(vendas_mes_2018$Sum)
```

```
    Shapiro-Wilk normality test

data:  vendas_mes_2018$Sum
W = 0.949, p-value = 0.6224
```

Observando o resultado e definindo uma conclusão

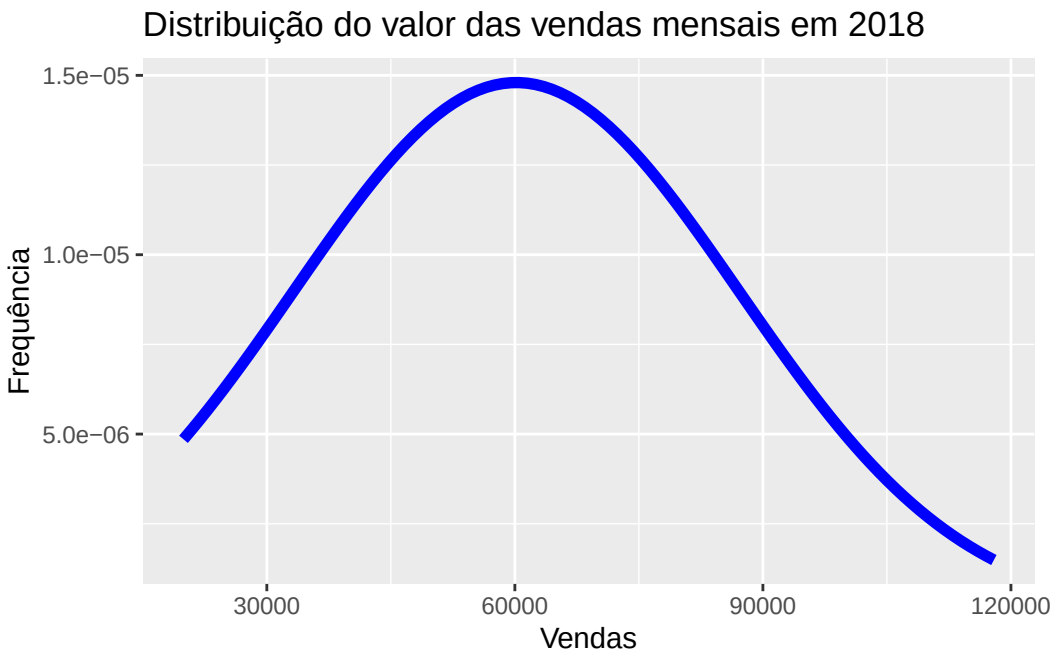
O resultado do teste mostra que o p-valor é de 0.6224, o que significa que não há evidências suficientes para rejeitar a hipótese nula de que os valores das vendas mensais em 2018 seguem uma distribuição Normal. Isso é reforçado pelo valor W de 0.949, que é um indicador da aproximação da amostra à distribuição Normal. Valores próximos de 1 indicam que a amostra está mais próxima da distribuição Normal.

Com base nesse resultado, podemos concluir que os valores das vendas mensais em 2018 seguem uma distribuição Normal.

Visualizando o gráfico de distribuição das vendas mensais em 2018

```
vendas = vendas_mes_2018$Sum

mu <- mean(vendas)
sigma <- sd(vendas)
```

```r
x <- seq(min(vendas), max(vendas), length.out = 100)
y <- dnorm(x, mean = mu, sd = sigma)

df <- data.frame(x = x, y = y)



ggplot(df, aes(x = x, y = y), xlab("sdfl")) +
  geom_line(colour = "blue", linewidth = 2) +
  labs(
    title = "Distribuição do valor das vendas mensais em 2018",
    x = "Vendas", y = "Frequência"
  )
```



8.3. 3º Hipótese

O estado onde a venda foi realizada afeta o seu valor

Assuminfo que, $\mu_1, \mu_2, ..., \mu_n$ são os valores médios das vendas para cada tipo de envio, temos:

$h0 : \mu_1 = \mu_2 = ... = \mu_n$ (Hipótese nula)

$h1$: Pelo menos uma das μ_i é diferente de outra μ_j (Hipótese alternativa)

Para fazer um teste de hipótese de que o estado afeta o valor das vendas em R, vamos utilizar uma análise de variância (ANOVA). A ANOVA é um método estatístico que permite comparar a média de uma variável dependente (no nosso caso, o valor da venda) entre dois ou mais grupos de uma variável independente (no nosso caso, o estado onde ocorreu a venda).

Transformando o estado em uma variável categórica

```
dados$Ship.Mode <- as.factor(dados$Ship.Mode)
```

Ajustando um modelo de ANOVA

Temos como variável dependente "Sales" e a variável independente é "Ship.Mode"

```
modelo <- aov(Sales ~ Ship.Mode, data = dados)
```

Executando o teste de comparação múltipla de Tukey

```
TukeyHSD(modelo)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sales ~ Ship.Mode, data = dados)

$Ship.Mode
                                diff       lwr      upr     p adj
Same Day-First Class         2.521124 -78.39946 83.44171 0.9998162
Second Class-First Class     6.319919 -49.27912 61.91896 0.9913313
Standard Class-First Class  -1.378163 -47.96559 45.20926 0.9998426
Second Class-Same Day        3.798796 -74.83884 82.43643 0.9993170
Standard Class-Same Day     -3.899287 -76.44591 68.64733 0.9990606
Standard Class-Second Class -7.698083 -50.19660 34.80043 0.9665870
```

Observando o resultado e definindo uma conclusão

O resultado mostra as diferenças entre os valores médios das vendas para cada par de tipos de envio (por exemplo, "Same Day" versus "First Class") e os intervalos de confiança para essas diferenças. O p-valor ajustado mostra a probabilidade de que os resultados observados sejam devidos ao acaso, dado que a hipótese nula é verdadeira.

No caso deste exemplo, o p-valor ajustado para todas as comparações é maior do que o nível de significância que escolhemos (0.05). Isso significa que não podemos rejeitar a hipótese nula,

ou seja, não podemos concluir que há uma diferença significativa entre os valores médios das vendas para os diferentes tipos de envio.

9. Conclusão

Por meio da análise , foi possível identificar que houve um crescimento de aproximadamente 250 mil vendas durante os 4 anos. Há um maior número de vendas nos meses de Novembro e Dezembro, devido a uma grande quantidade de datas comemorativas e feriados nos EUA durante esse período. Além disso, Março é o mês onde tem a maior variabilidade e média de vendas mensais durante os anos de 2015 a 2018. Portanto é fundamental que a loja, busque estratégias para que o número de vendas se mantenha constante, a fim de mitigar os riscos.

Além do mais, dentre os dez estados que possuem a maior receita de vendas na loja, oito possuem tecnologia como a principal categoria de produto. Dito isso, é necessário traçar planejamentos de marketing e disposição dos produtos dentro da loja para maximizar o faturamento dessa categoria.

Ja no que diz respeito às hipóteses, foi possível identificar que não podemos concluir com certeza que as vendas mensais da categoria Technology é maior que as da categoria Office Supplies, e concluir que as vendas mensais no ano de 2018 seguem uma distribuição normal e que não podemos concluir que O estado onde a venda foi realizada afeta o seu valor.

10. Referências

SOUSA, Nuno. Visualização de dados e testes de hipóteses com R: uma breve abordagem prática. 2018

WICKHAM, Hadley; CHANG, Winston; WICKHAM, Maintainer Hadley. Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version, v. 2, n. 1, p. 1-189, 2016.

BOLFARINE, Heleno; SANDOVAL, Mônica Carneiro. Introdução à inferência estatística. SBM, 2001.