Análise Exploratória de Dados

Luiz Carlos Castro Cunha Junior

29/05/2021

Estudo de Caso: Telecom

A base de dados traz informações dos clientes de uma telecom que cancelaram ou não sua conta.Durante a análise deverá ser respondida às seguintes questões de negócio:

- Quantos clientes a base de dados possui? Quantos são mulheres? E de forma relativa, quantas são mulheres?
- Quais são os valores da média, mediana, mínimo, máximo e quartis do tempo de relacionamento?
- Com base na distribuição de frequências do tempo de relacionamento, qual a proporção de clientes que ainda não completaram 1 ano de relacionamento?
- Qual a proporção de clientes que possuem 10 anos de relacionamento?
- Qual o % de clientes que têm 1 produto? E 2 produtos? Utilize a variável Num_de_Produtos.
- Qual o total de clientes que já cancelaram os produtos? E que não cancelaram? Qual a frequência relativa de cada categoria? Considere 1 para o cliente que cancelou e 0 para o cliente que não cancelou.

Carregando os pacotes a serem utilizados

```
library(tidyverse)
library(readxl)
library(rstatix)
library(moments)
library(knitr)
library(rmarkdown)
library(kableExtra)
library(emmeans)
library(htmltools)
```

Sexo	Tempo_relacionamento (anos)	Num_de_Produtos	Cancelou
Feminino	2	1	1
Feminino	1	1	0
Feminino	8	3	1
Feminino	1	2	0
Feminino	2	1	0

```
library(dplyr)
library(questionr)
```

Importando o Conjunto de Dados

```
dataset<-readxl::read_xlsx("Exercícios.xlsx",
                           sheet = "Base de Dados 2")
```

Visualizando o tipo de dados das variáveis

```
str(dataset)
```

```
## tibble [10,000 x 5] (S3: tbl_df/tbl/data.frame)
##  $ ID                          : num [1:10000] 15634602 15647311 15619304 15701354 15737888 ...
##  $ Sexo                        : chr [1:10000] "Feminino" "Feminino" "Feminino" "Feminino" ...
##  $ Tempo_relacionamento (anos): num [1:10000] 2 1 8 1 2 4 10 5 7 9 ...
##  $ Num_de_Produtos             : num [1:10000] 1 1 3 2 1 4 2 2 2 2 ...
##  $ Cancelou                    : num [1:10000] 1 0 1 0 0 1 0 0 0 0 ...
```

visualizando as primeiras linhas do Dataset

```
kbl(dataset[1:5,2:5]) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Perguntas de Negócio

Quantos clientes a base de dados possui? Quantos são mulheres? E de forma relativa, quantas são mulheres?

```
total_registros <- nrow(dataset)
names(total_registros) <-"Total de Clientes"

dfSexo <- data.frame(c(table(dataset$Sexo)[1],
             table(dataset$Sexo)[2],
             total_registros),
           c(prop.table(table(dataset$Sexo))[1] * 100,
             prop.table(table(dataset$Sexo))[2] * 100,
             prop.table(table(total_registros)))
           )
names(dfSexo) <- c("Quantidade", "Frequência")

kbl(dfSexo) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Quais são os valores da média, mediana, mínimo, máximo e quartis do tempo de relacionamento?

	Quantidade	Frequência
Feminino	4543	45.43
Masculino	5457	54.57
Total de Clientes	10000	1.00

	Resultdo
Média	5.0128
Mediana	5.0000
PrimeiroQuartil	3.0000
SegundoQuartil	5.0000
TerceiroQuartil	7.0000
IntervaloInterquartil	4.0000

```
colnames(dataset)[3] <- "TempoRelacionamentoAnos"
indices <- c("Média", "Mediana", "Primerio Quartil", "Segundo Quartil", "Terceiro Quartil", "Intervalo I

dfTempoRelacionalmento <-
               data.frame(
                 c(
                     Média = mean(dataset$TempoRelacionamentoAnos),
                     Mediana =median(dataset$TempoRelacionamentoAnos),
                     PrimeiroQuartil = unname(quantile(dataset$TempoRelacionamentoAnos,0.25)),
                     SegundoQuartil  = unname(quantile(dataset$TempoRelacionamentoAnos, 0.50)),
                     TerceiroQuartil = unname(quantile(dataset$TempoRelacionamentoAnos, 0.75)),
                     IntervaloInterquartil = IQR(dataset$TempoRelacionamentoAnos)
                   ))
colnames(dfTempoRelacionalmento) <-"Resultdo"


kbl(dfTempoRelacionalmento) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

Com base na distribuição de frequências do tempo de relacionamento, qual a proporção de clientes que ainda não completaram 1 ano de relacionamento?

- 4.1 % dos clientes não completaram 1 ano de relacionamento.

Qual a proporção de clientes que possuem 10 anos de relacionamento?

- 4.9 % dos clientes possuem 10 anos de relacionamento.

```
dataset$TempoRelacionamentoAnos<-
  ifelse(dataset$TempoRelacionamentoAnos > 0,paste(dataset$TempoRelacionamentoAnos, " Anos"),"Não compl


dfTabelaFrequenciaTempoRelacionamento <- questionr::freq(dataset$TempoRelacionamentoAnos,
                                          total = TRUE)

colnames(dfTabelaFrequenciaTempoRelacionamento) <- c("Frequência Relativa",
                                                     "Frequência Absoluta")

kbl(dfTabelaFrequenciaTempoRelacionamento[1:2]) %>%
    kable_styling(bootstrap_options = c("striped", "hover"))
```

Qual o % de clientes que têm 1 produto? E 2 produtos?

	Frequência Relativa	Frequência Absoluta
1 Anos	1035	10.3
10 Anos	490	4.9
2 Anos	1048	10.5
3 Anos	1009	10.1
4 Anos	989	9.9
5 Anos	1012	10.1
6 Anos	967	9.7
7 Anos	1028	10.3
8 Anos	1025	10.2
9 Anos	984	9.8
Não completou 1 Ano	413	4.1
Total	10000	100.0

	FrequênciaRelativa	Frequência Absoluta
1 Produtos	5084	50.8
2 Produtos	4590	45.9
3 Produtos	266	2.7
4 Produtos	60	0.6
Total	10000	100.0

Podemos observar na tabela abaixo que 96,7% dos clientes possuem entre 1 e 2 produtos.

Sendo que:

- Para 1 Produto 50.8% dos Clientes
- Para 2 Produtos 45.9% dos Clientes

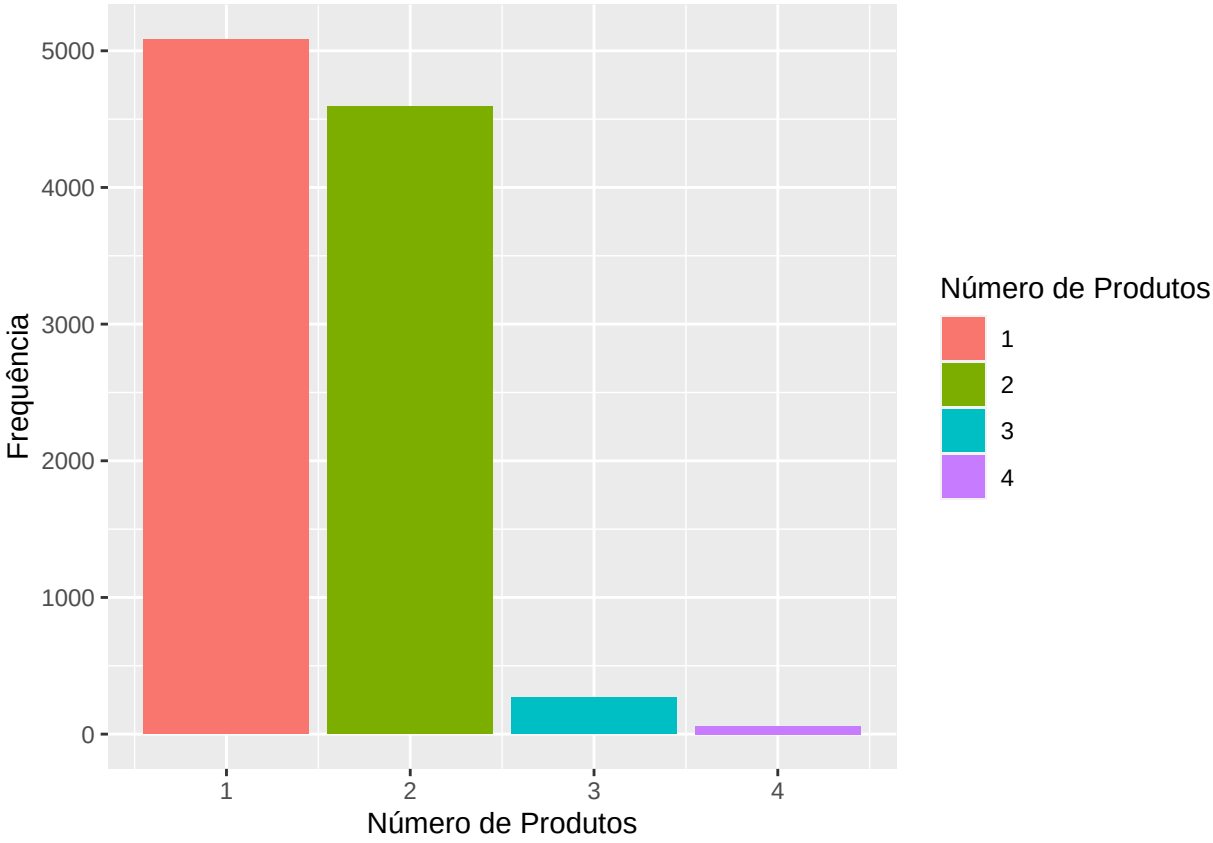
```
dfTabelaFrequenciaClientes <- questionr::freq(dataset$Num_de_Produtos,total = TRUE)

colnames(dfTabelaFrequenciaClientes) <- c("FrequênciaRelativa",
                                          "Frequência Absoluta")
rownames(dfTabelaFrequenciaClientes)[1:4] <-
  paste(rownames(dfTabelaFrequenciaClientes)[1:4], "Produtos")

kbl(dfTabelaFrequenciaClientes[1:2]) %>%
    kable_styling(bootstrap_options = c("striped", "hover"))
```

```
dataset %>%
  ggplot(aes(dataset$Num_de_Produtos, fill = as.factor(dataset$Num_de_Produtos))) +
  geom_bar(stat = "count")  +
  labs(fill = "Número de Produtos")+
  xlab("Número de Produtos") +
  ylab("Frequência")
```

	Frequência_Relativa	Frequência_Absoluta
Cancelou	2037	20.4
Não Cancelou	7963	79.6
Total	10000	100.0



Qual o total de clientes que já cancelaram os produtos? E que não cancelaram?

2037 cancelaram seus produtos e 7963 não cancelaram seus produtos.

Qual a frequência relativa de cada categoria? Considere 1 para o cliente que cancelou e 0 para o cliente que não cancelou.

```
View(ifelse(dataset$Cancelou ==0, "Cancelou",dataset$Cancelou))

dataset$Cancelou <- ifelse(dataset$Cancelou ==0, "Não Cancelou",dataset$Cancelou)
dataset$Cancelou <- ifelse(dataset$Cancelou ==1, "Cancelou",dataset$Cancelou)

dfClientesCancelamento = questionr::freq(dataset$Cancelou, total = TRUE)
names(dfClientesCancelamento)<- c("Frequência_Relativa",
                                  "Frequência_Absoluta")

kbl(dfClientesCancelamento[1:2]) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))
```

```
dataset%>%
  ggplot(aes(Cancelou, fill = as.factor(Cancelou))) +
  geom_bar(stat = "count") +
```

```
  labs(fill = "Cancelamento dos Produtos")+
  ylab("Frequência")
```