Análise Exploratória de Dados

Luiz Carlos Castro Cunha Junior

29/05/2021

Estudo de Caso: Imóveis

De acordo com a localização de um imóvel, sabe-se que o valor do mesmo pode variar substancialmente. Na base de dados disponibilizada são fornecidas informações sobre o valor do imóvel (R$ mil) por m2, a distância para estação de metrô (km), a idade e a região.

- Faça a distribuição de frequências da variável idade.
- Faça a distribuição de frequências da variável região.
- Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável distância ao metrô? Interprete os valores.
- Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável valor do imóvel (R$ mil) por m²? Interprete os valores.

Carregando os pacotes a serem utilizados

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

Idade_imovel	Região	Distancia_metro_Km	Mil_reais_m2
3. Acima de 25 anos	Norte	1.083595	7.58
2. 10 a 25 anos	Sul	1.396946	8.44
2. 10 a 25 anos	Sul	1.544789	9.46
2. 10 a 25 anos	Norte	1.544789	10.96
1. Até 10 anos	Norte	1.456010	8.62

	Frequência Absoluta	Frequência Relativa
1. Até 10 anos	109	26.4
2. 10 a 25 anos	187	45.3
3. Acima de 25 anos	117	28.3
Total	413	100.0

Importando o Conjunto de Dados

```
dataset<-readxl::read_xlsx("Exercícios.xlsx",
                           sheet = "Base de Dados 3")
```

Visualizando o tipo de dados das variáveis

```
str(dataset)
```

```
## tibble [413 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Id_Imovel         : num [1:413] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Idade_imovel      : chr [1:413] "3. Acima de 25 anos" "2. 10 a 25 anos" "2. 10 a 25 anos" "2. 10 a
##  $ Região            : chr [1:413] "Norte" "Sul" "Sul" "Norte" ...
##  $ Distancia_metro_Km: num [1:413] 1.08 1.4 1.54 1.54 1.46 ...
##  $ Mil_reais_m2      : num [1:413] 7.58 8.44 9.46 10.96 8.62 ...
```

Visualizando as primeiras linhas do Dataset

```
kbl(dataset[1:5,2:5]) %>%
  kable_paper()
```

Perguntas de Negócio

Faça a distribuição de frequências da variável idade.

```
dfFrequenciaIdade <- questionr::freq(dataset$Idade_imovel, total = TRUE)
colnames(dfFrequenciaIdade) <- c("Frequência Absoluta", "Frequência Relativa")

kbl(dfFrequenciaIdade[, 1:2]) %>%
  kable_paper()
```

Faça a distribuição de frequências da variável região.

```
dfFrequenciaIdade <- questionr::freq(dataset$Região, total = TRUE)
colnames(dfFrequenciaIdade) <- c("Frequência Absoluta", "Frequência Relativa")

kbl(dfFrequenciaIdade[, 1:2]) %>%
  kable_paper()
```

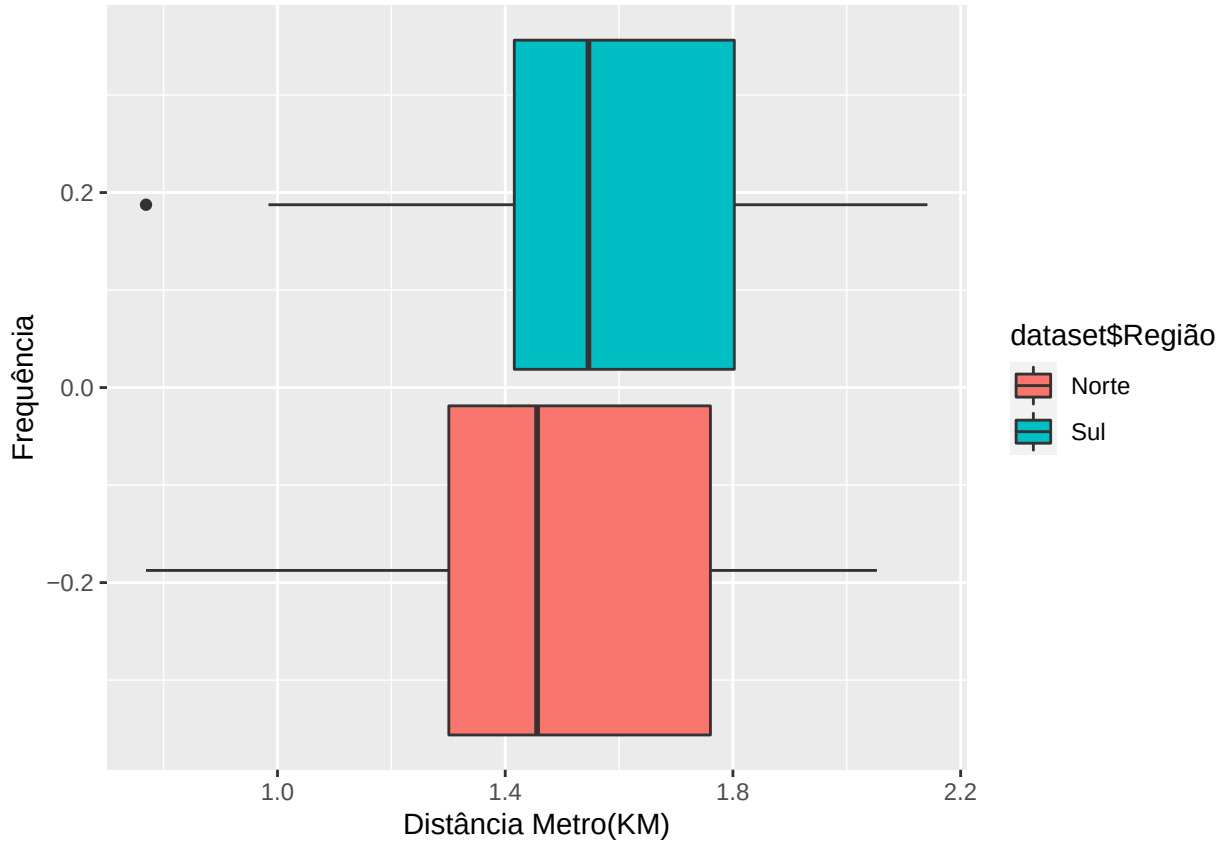
Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável distância ao metrô? Interprete os valores.

	Frequência Absoluta	Frequência Relativa
Norte	155	37.5
Sul	258	62.5
Total	413	100.0

	Valores
Minimo	0.7690434
Máximo	2.1416361
Média	1.5599941
Mediana	1.5124546
PrimeiroQuartil	1.3828009
SegundoQuartil	1.5124546
TerceiroQuartil	1.7770231
Amplitude	1.3725927
LimiteInferior	0.7914677
IntervaloInterQuartil	0.3942222

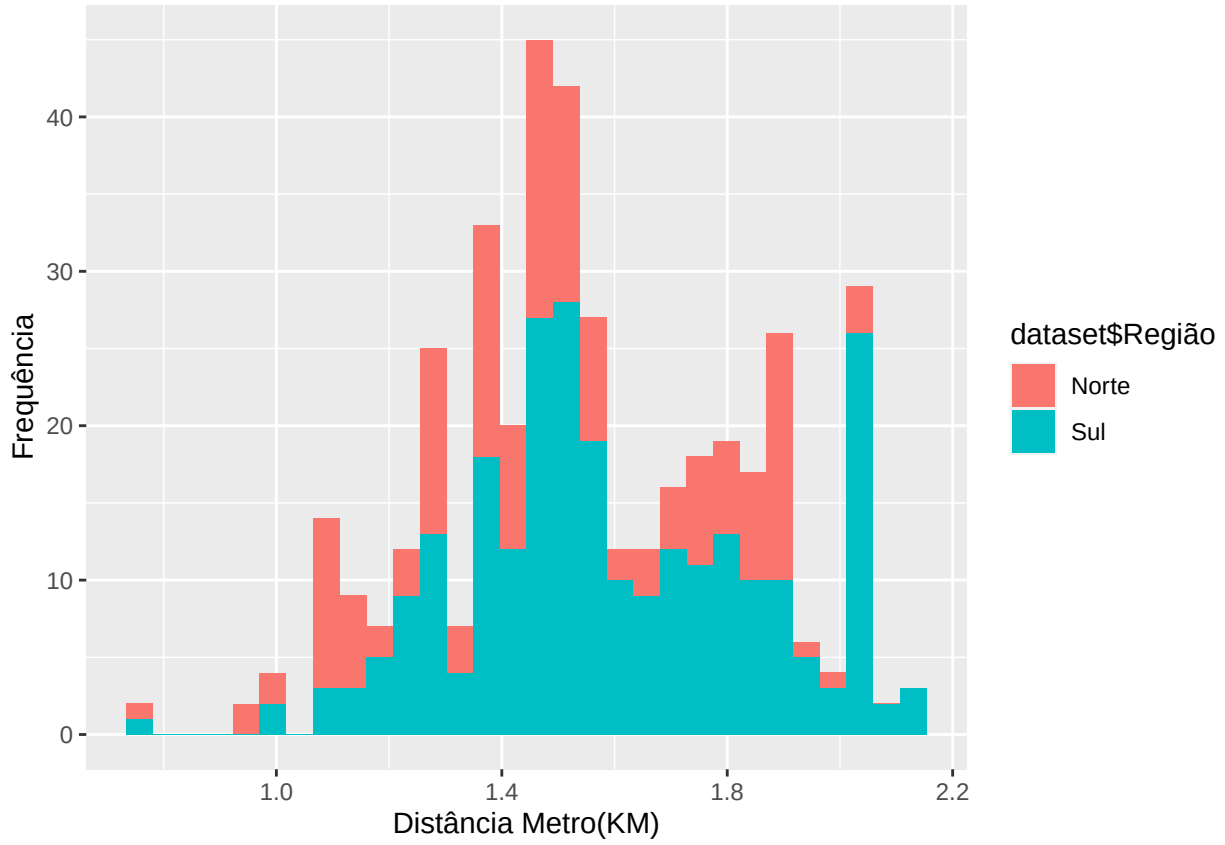
Podemos observar que a localização dos imóveis estão variando aproximadamente 1.3 KM de distância em relação ao metrô.

```
dataset %>%
  ggplot(aes(dataset$Distancia_metro_Km, fill = dataset$Região)) +
  geom_boxplot(bins = 30)+
  xlab("Distância Metro(KM)")+
  ylab("Frequência")
```



	Valores
Minimo	1.52000
Máximo	15.66000
Média	7.55753
Mediana	7.68000
PrimeiroQuartil	5.54000
SegundoQuartil	7.68000
TerceiroQuartil	9.32000
Amplitude	14.14000
LimiteInferior	-0.13000
IntervaloInterQuartil	3.78000

```
dataset %>%
  ggplot(aes(dataset$Distancia_metro_Km, fill = dataset$Região)) +
  geom_histogram(bins = 30)+
  xlab("Distância Metro(KM)")+
  ylab("Frequência")
```

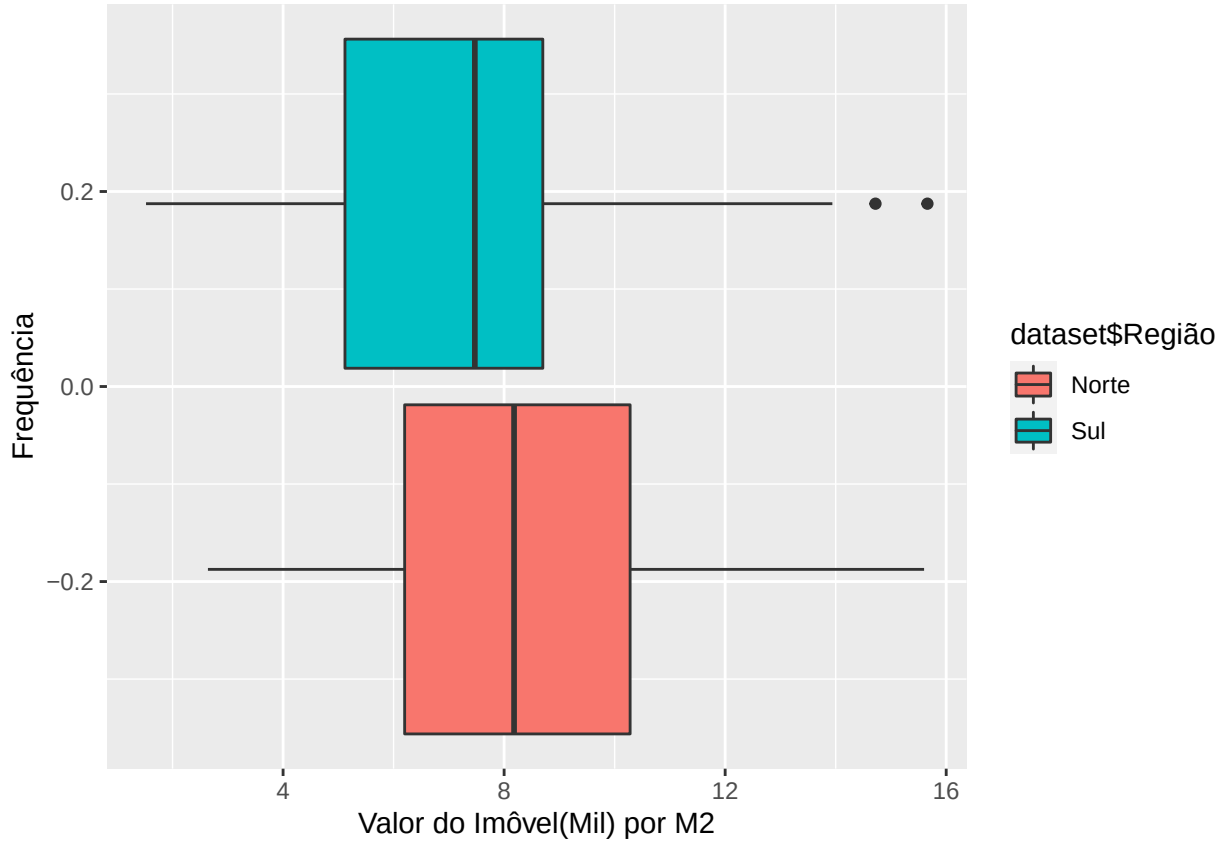


Qual o valor do mínimo, máximo, mediana, Q1 e Q3 da variável valor do imóvel (R$ mil) por m²? Interprete os valores.

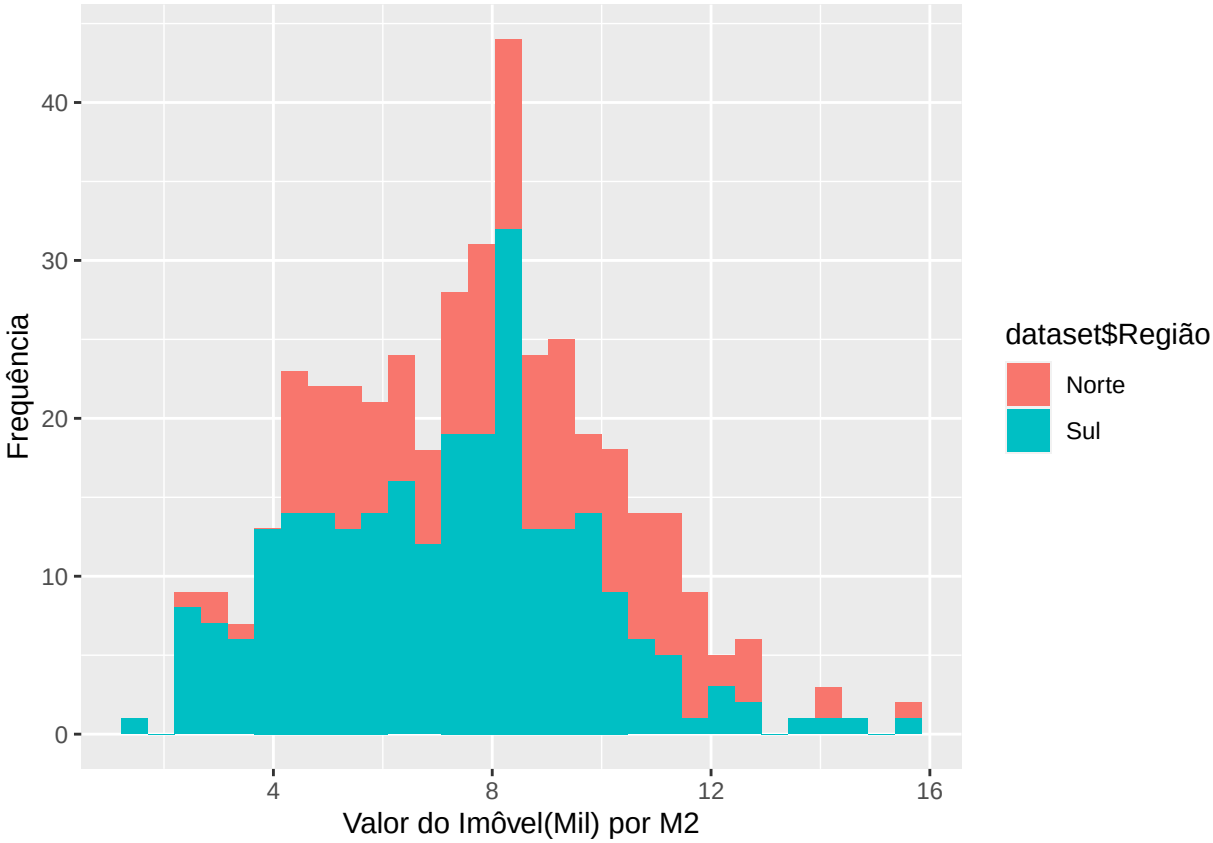
De acordo com os valores apresentados abaixo podemos observar que o valor dos imóveis estão variando a aproxidamente R$ 14.140,00.

Sendo que o imóvel mais barato está custando R$ 1.520,00 e o mais caro R$ 15.660,00

```
dataset %>%
  ggplot(aes(dataset$Mil_reais_m2, fill = dataset$Região)) +
  geom_boxplot(bins = 30)+
  xlab("Valor do Imôvel(Mil) por M2")+
  ylab("Frequência")
```




```
dataset %>%
  ggplot(aes(dataset$Mil_reais_m2, fill = dataset$Região)) +
  geom_histogram(bins = 30)+
  xlab("Valor do Imôvel(Mil) por M2")+
  ylab("Frequência")
```


Frequência
40
30
20
10
0
4
8
12
16
Valor do Imôvel(Mil) por M2
dataset$Região
Norte
Sul