Desafio Seazone

Sheyla Maria Tavares e Tavares

Análise de dados

Sobre os dados

Os dados a serem analisados são sobre a ocupação e o preço dos anúncios no Airbnb. Foram cedidas duas bases de dados com estas informações, além das características de cada anúncio. A base de dados nomeada como *"desafio_details.csv"* conta com informações de 4.691 diferentes anúncios, enquanto que o arquivo *"desafio_priceav.csv"* traz informações das ocupações nos anos de 2020 e 2021.

Para facilitar as análises, foi realizada a mesclagem dos dois arquivos pela identificação dos anúncios.

```
# ################################################################
#                         DADOS PARA ANALISE                     ##
# ################################################################

price = read.csv2(file = "desafio_priceav.csv", header = T, sep = ",", encoding = "UTF-8"); price = pri
details = read.csv2(file = "desafio_details.csv", header = T, sep = ",", encoding = "UTF-8")

# ----- Mesclando dados
dados = merge(price, details, by.x = "airbnb_listing_id", by.y = "airbnb_listing_id",all = T)
dados[,4] = as.Date(dados[,4]); dados[,5] = as.numeric(dados[,5])
head(dados)
```

```
##   airbnb_listing_id   X.x           booked_on       date price_string occupied
## 1            108658 24825               blank 2020-11-19          300        0
## 2            108658 24826               blank 2020-11-20          300        0
## 3            108658 24827 2020-11-21 00:00:00 2020-11-21          300        1
## 4            108658 24828               blank 2020-11-22          300        0
## 5            108658 24829 2020-11-23 00:00:00 2020-11-23          300        1
## 6            108658 24830               blank 2020-11-24          300        0
##    X.y       suburb                                      ad_name
## 1 2063 Canasvieiras Apartamento aconchegante de frente para o mar
## 2 2063 Canasvieiras Apartamento aconchegante de frente para o mar
## 3 2063 Canasvieiras Apartamento aconchegante de frente para o mar
## 4 2063 Canasvieiras Apartamento aconchegante de frente para o mar
## 5 2063 Canasvieiras Apartamento aconchegante de frente para o mar
## 6 2063 Canasvieiras Apartamento aconchegante de frente para o mar
##   number_of_bedrooms number_of_bathrooms star_rating is_superhost
## 1                2.0                 2.0                   False
## 2                2.0                 2.0                   False
## 3                2.0                 2.0                   False
## 4                2.0                 2.0                   False
## 5                2.0                 2.0                   False
```

```
## 6                2.0                2.0                    False
##   number_of_reviews
## 1               0.0
## 2               0.0
## 3               0.0
## 4               0.0
## 5               0.0
## 6               0.0
```

Conhecendo as Variáveis

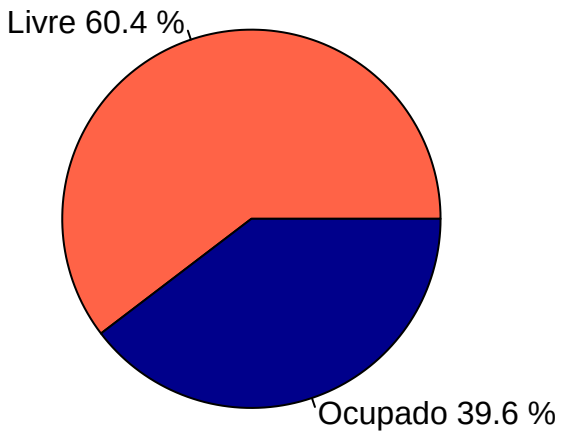
Das variáveis que caracterizam os anúncios nos anos de 2020 e 2021, têm-se que 60,40% dos anúncios estavam livres e 39,60% estavam ocupados. Além disso, apenas 28,50 % são superhosts.

Ao considerar a quantidade de anúnios por bairros, destaca-se o bairro Ingleses com 177.542 listings, seguido por Canasvieira com 92.513. O centro foi o que menos anunciou imóveis com 19.263 listings.

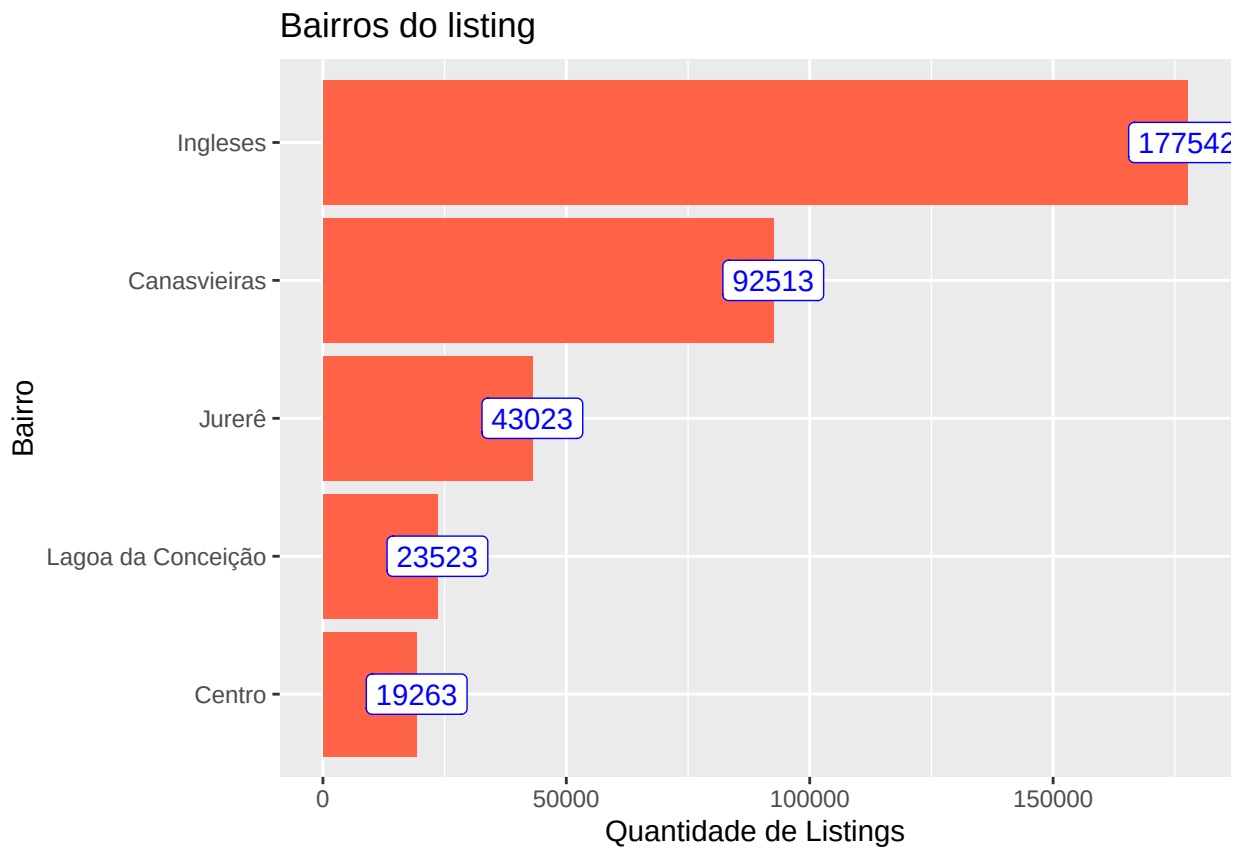
As notas dos anúncios são dispostas de 0 a 5. Com destaque, 45,81% dos listings foram classificados com nota 5. Entretanto, vale ressaltar que há um grande volume de anuncios sem classificação por nota.

Os imóveis anuncioados possuem em sua granpe parte de 1 a 2 banheiros e 1 a 3 quartos. A média de preço gira em torno de 328 reais.

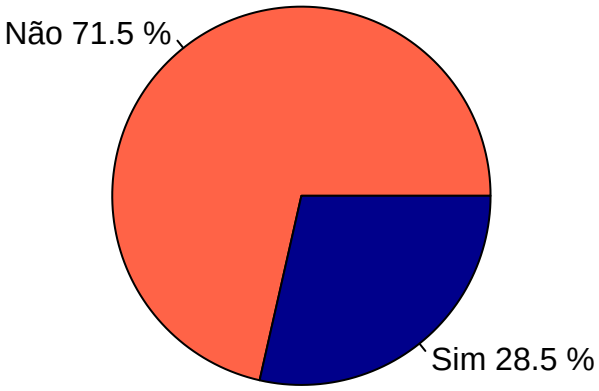
```
# ##################################################################
#                    CONHECENDO AS VARIÁVEIS                      ##
# ##################################################################

# Ocupação
dados = within(dados, {
  occupied <- factor(occupied, labels=c("Livre","Ocupado"))
})

tab1 = table(dados$occupied)
rotulos = paste(row.names(tab1),round(prop.table(tab1)*100,digits = 1),c("%","%"))
pie(tab1,labels = rotulos, col = c("tomato","darkblue"),main = )
```



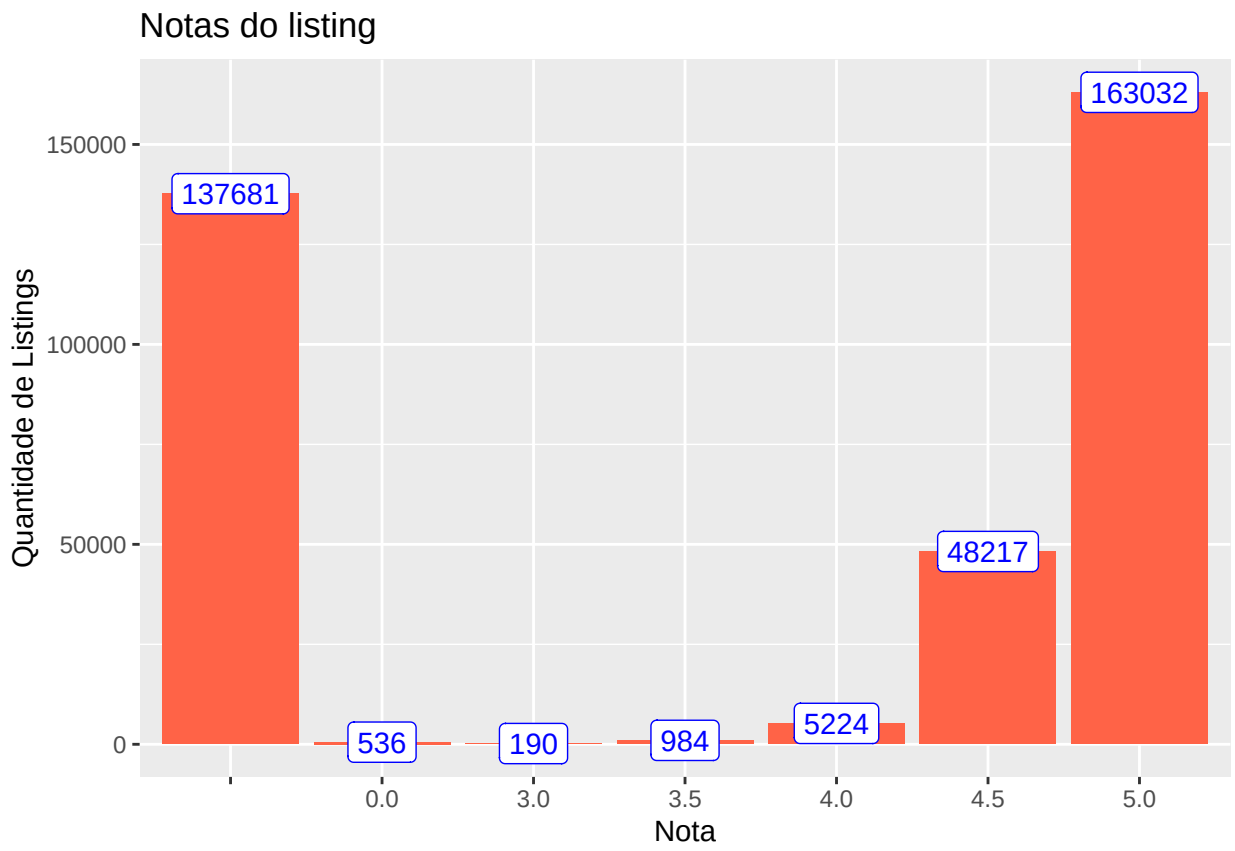

```
# Bairros
tab2 = as.data.frame(table(dados$suburb))
ggplot(tab2, aes(x =fct_reorder(Var1,Freq), y = Freq), fill = Freq) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = Freq),colour="blue") +
  labs(title = "Bairros do listing",
       x = "Bairro",
       y = "Quantidade de Listings")+
  coord_flip()
```



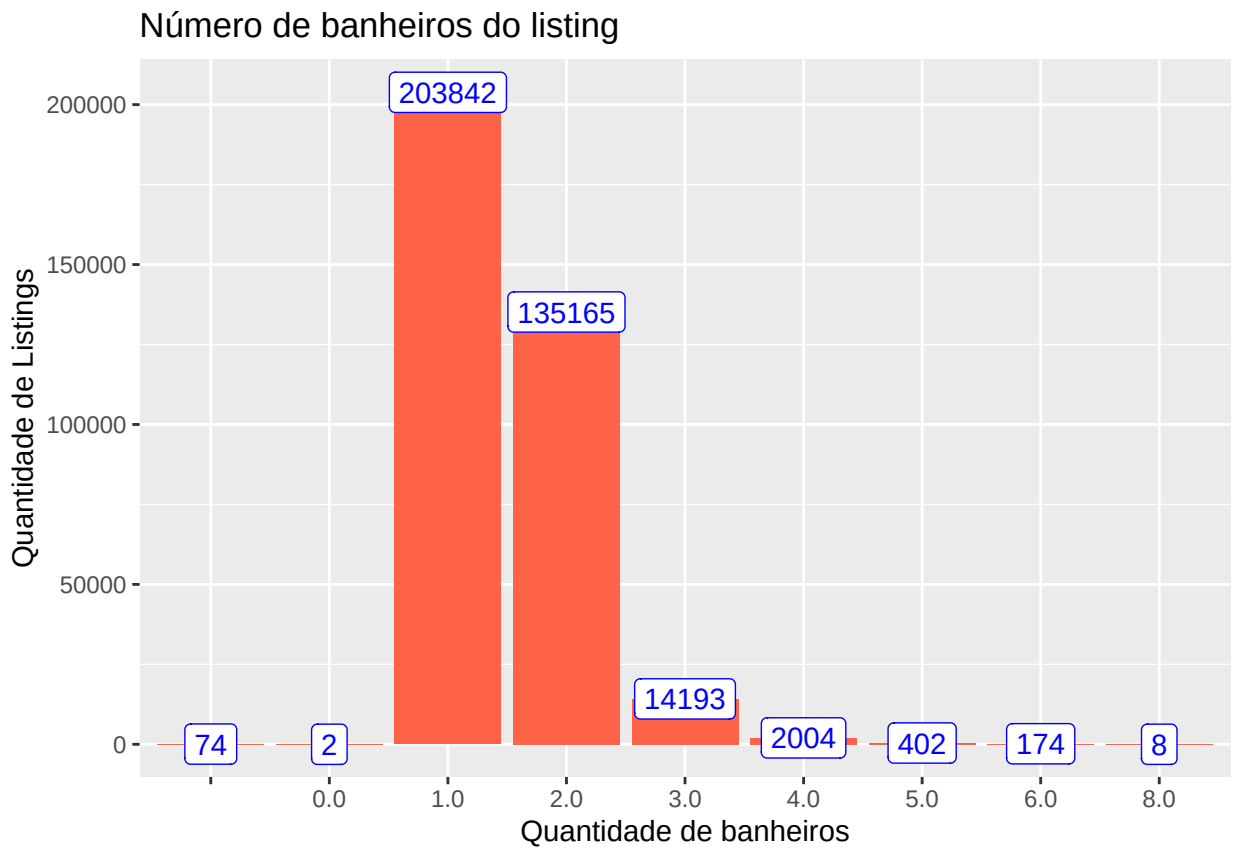

```
# Superhost
tab3 = table(dados$is_superhost)
rot3 = paste(c("Não","Sim"),round(prop.table(tab3)*100,digits = 1),c("%","%"))
pie(tab3,labels = rot3, col = c("tomato","darkblue"),main = )
```



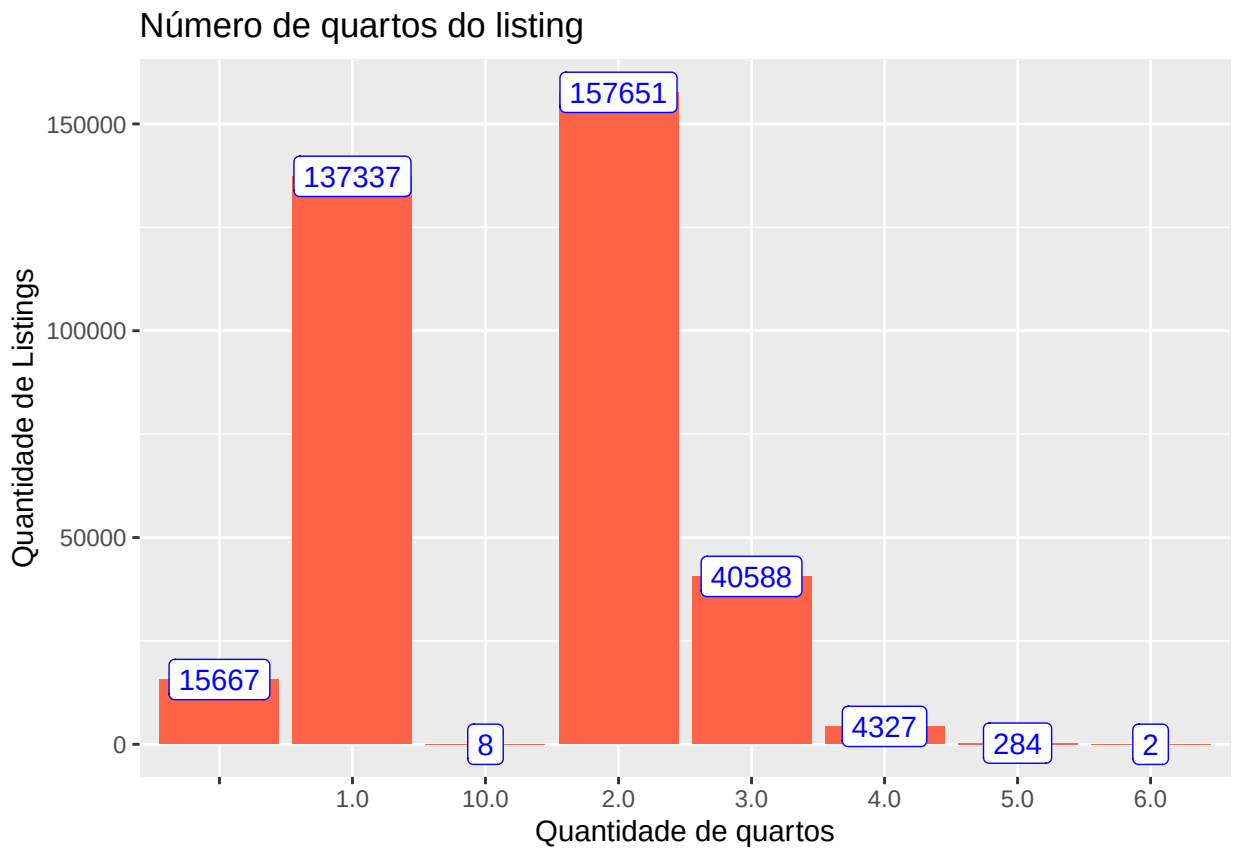

```
# Nota do anuncio
tab4 = as.data.frame(table(dados$star_rating))
ggplot(tab4, aes(x =Var1, y = Freq), fill = Freq) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = Freq),colour="blue") +
  labs(title = "Notas do listing",
       x = "Nota",
       y = "Quantidade de Listings")
```




```
# Nota do Banheiros
tab5 = as.data.frame(table(dados$number_of_bathrooms))
ggplot(tab5, aes(x =Var1, y = Freq), fill = Freq) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = Freq),colour="blue") +
  labs(title = "Número de banheiros do listing",
       x = "Quantidade de banheiros",
       y = "Quantidade de Listings")
```




```
# Nota do Quartos
tab6 = as.data.frame(table(dados$number_of_bedrooms))
ggplot(tab6, aes(x =Var1, y = Freq), fill = Freq) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = Freq),colour="blue") +
  labs(title = "Número de quartos do listing",
       x = "Quantidade de quartos",
       y = "Quantidade de Listings")
```



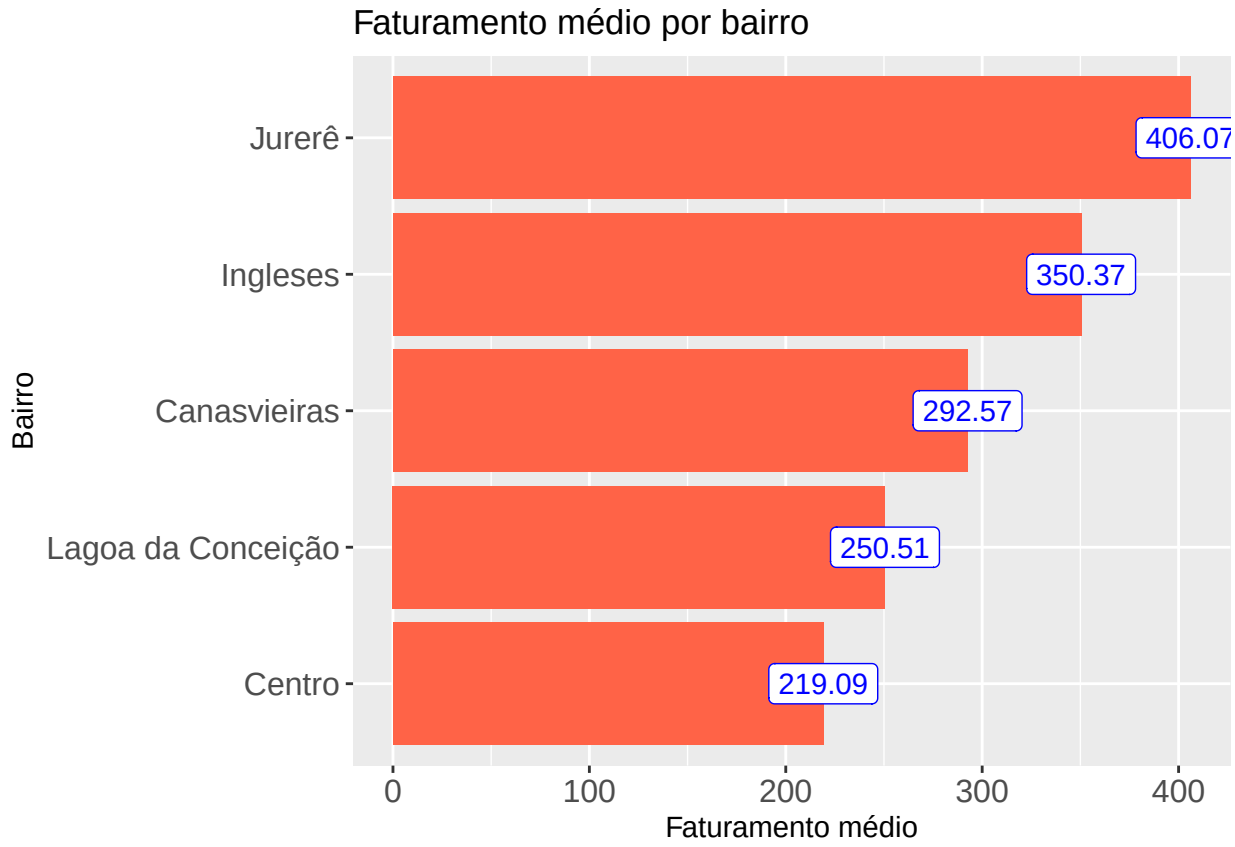

```
# Preços dos anuncios
tab7 = summary(dados$price_string)
est_price = data.frame(x=matrix(tab7),row.names=names(tab7))
print(est_price)
```

```
##                 x
## Min.      41.0000
## 1st Qu.  199.0000
## Median   298.0000
## Mean     328.3722
## 3rd Qu.  418.0000
## Max.    5500.0000
## NA's    1344.0000
```

Estudo do Faturamento

O faturamento do listing é a soma de preços nas datas alugadas. Desse modo, o faturamento médio por bairro mostrou o bairro de Jurerê teve maior faturamento médio com 406,07 reais, seguido pelo bairro Ingleses com 350,37 reais, Canasvieira com 292,57 reais, Lagoa da conceição com 250,51 e, por fim, o Centro com 219,09 reais.

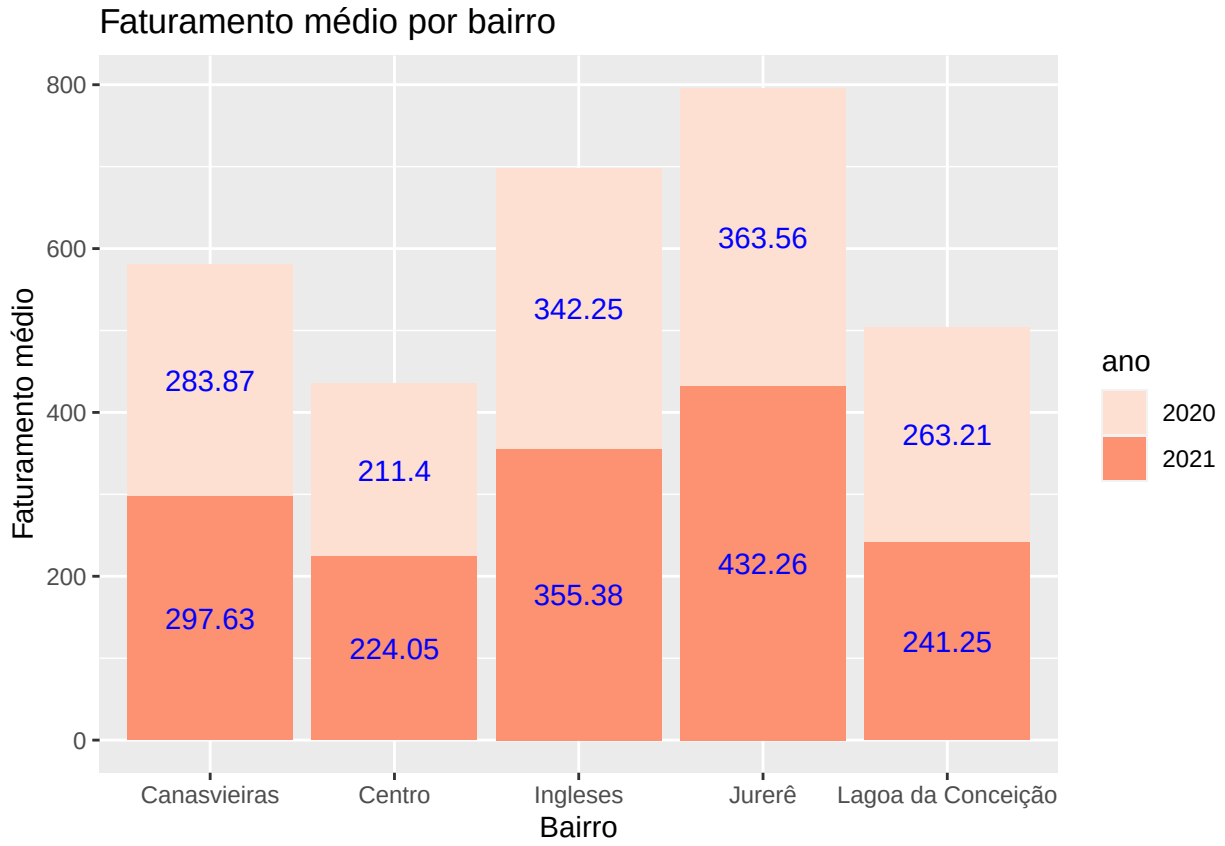
```r
# ###################################
# FATURAMENTO MÉDIO POR BAIRRO    ##
# ###################################

sel2=dados[,c(1,4,5,8)];
sel21 = aggregate(price_string ~ airbnb_listing_id, dados, sum) # faturamento
tab21 = aggregate(price_string ~ suburb, sel2, mean);tab21[,2]=round(tab21[,2],2)

ggplot(tab21, aes(x =fct_reorder(suburb,price_string), y = price_string),) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = price_string),colour="blue") +
  theme(axis.text = element_text(size = 12))+
  labs(title = "Faturamento médio por bairro",
       x = "Bairro",
       y = "Faturamento médio")+
  coord_flip()
```



Como as informações são de anúncios ocorridos em dois anos, pode-se avaliar também o faturamento dos anúncios por bairro durante esses dois anos. Nos dois anos o bairro de Jurerê foi o que teve o menor faturamento médio, assim como, o centro teve o menor faturamento médio nos dois anos.

```r
# ########################################
# FATURAMENTO MÉDIO ANUAL POR BAIRRO    ##
# ########################################

tab22 = aggregate(price_string ~ suburb + year(date), sel2, mean);tab22[,3]=round(tab22[,3],digits = 2)
```

```
tab22 = within(tab22, {ano <- factor(ano, labels=c(2020,2021))})

ggplot(tab22, aes(x = suburb, y = price_string, fill = ano)) +
  geom_col()+
  scale_fill_brewer(palette = "Reds")+
  labs(title = "Faturamento médio por bairro",
       x = "Bairro",
       y = "Faturamento médio")+
  geom_text(aes(label = price_string), position = position_stack(vjust = 0.5),colour = "blue")
```



Por fim, a tabela a seguir demonstra o faturamento médio mensal para os dois anos (2020 e 2021) por bairro de forma mais detalhada.

```
# ##############################################
# FATURAMENTO MÉDIO MENSAL/ANUAL POR BAIRRO    ##
# ##############################################

tab23 = aggregate(price_string ~ suburb + year(date) + month(date,label = T,abbr = F), sel2, mean)

print(tab23)
```

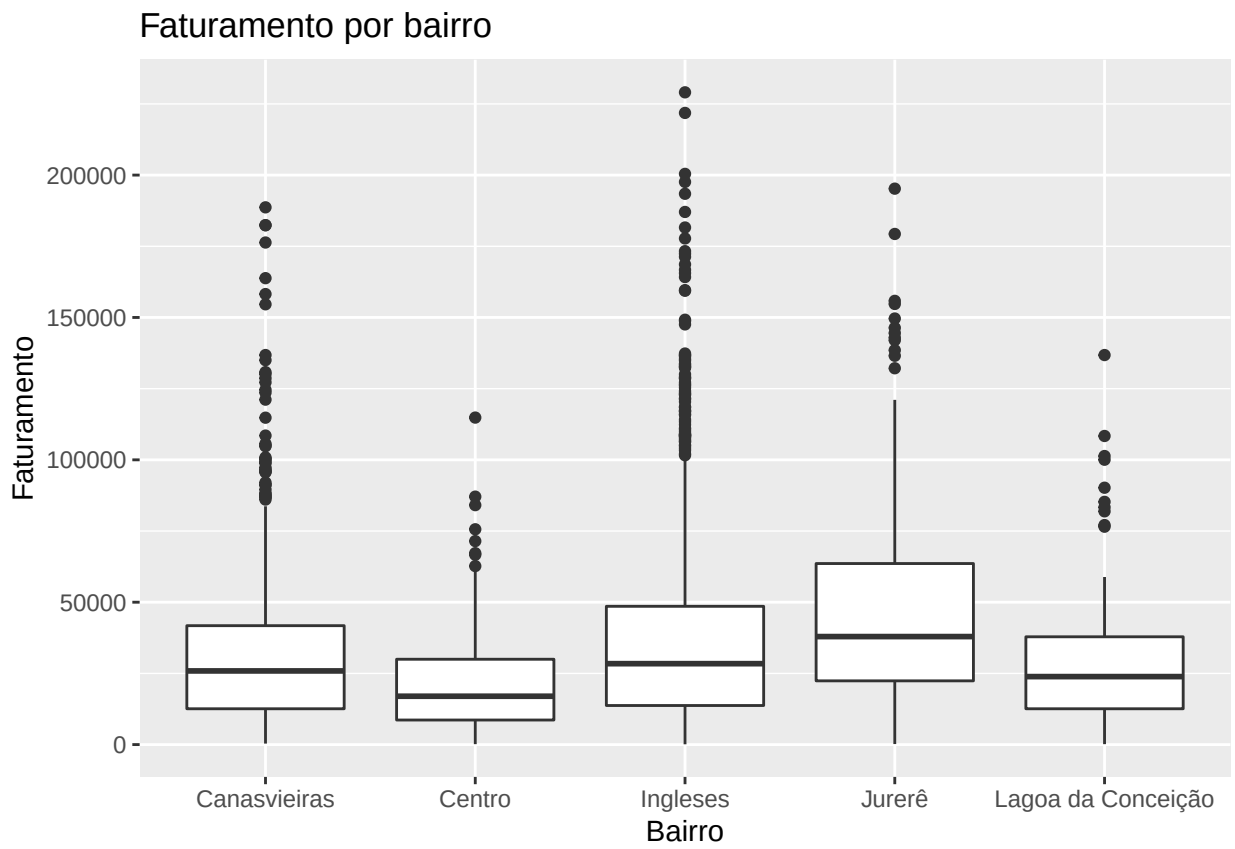
```
##               suburb year(date) month(date, label = T, abbr = F) price_string
## 1        Canasvieiras       2021                          janeiro     334.2837
## 2             Centro        2021                          janeiro     249.9582
## 3           Ingleses        2021                          janeiro     390.4249
```

```
## 4              Jurerê       2021                    janeiro     490.7056
## 5  Lagoa da Conceição       2021                    janeiro     273.6669
## 6        Canasvieiras       2021                  fevereiro     295.4761
## 7              Centro       2021                  fevereiro     217.5211
## 8            Ingleses       2021                  fevereiro     356.3289
## 9              Jurerê       2021                  fevereiro     450.8657
## 10 Lagoa da Conceição       2021                  fevereiro     237.3579
## 11       Canasvieiras       2021                      março     253.7264
## 12             Centro       2021                      março     203.4944
## 13           Ingleses       2021                      março     315.8363
## 14             Jurerê       2021                      março     344.4480
## 15 Lagoa da Conceição       2021                      março     204.3042
## 16       Canasvieiras       2020                   novembro     253.9823
## 17             Centro       2020                   novembro     187.3352
## 18           Ingleses       2020                   novembro     309.2822
## 19             Jurerê       2020                   novembro     323.4782
## 20 Lagoa da Conceição       2020                   novembro     228.2046
## 21       Canasvieiras       2020                   dezembro     304.3809
## 22             Centro       2020                   dezembro     231.1473
## 23           Ingleses       2020                   dezembro     368.8573
## 24             Jurerê       2020                   dezembro     399.8827
## 25 Lagoa da Conceição       2020                   dezembro     292.1460
```

Relações com o Faturamento

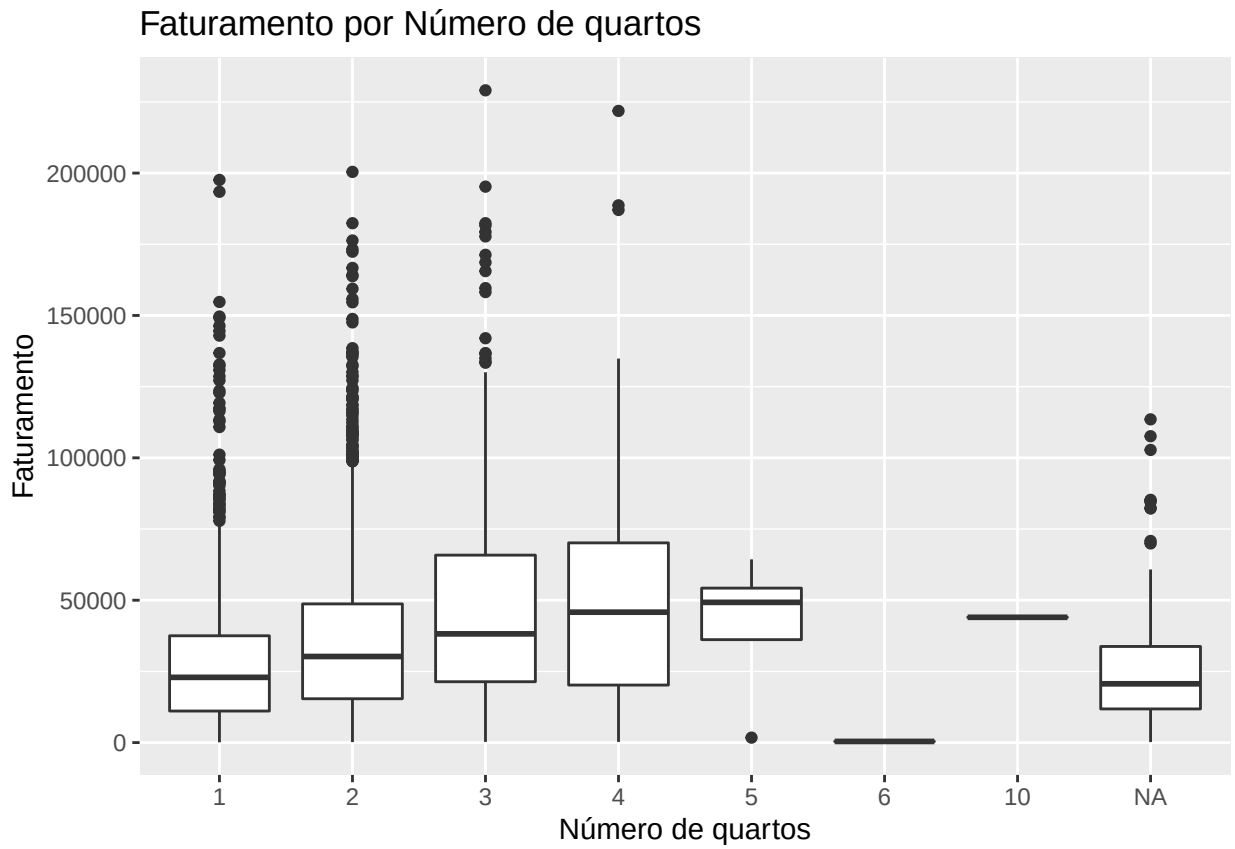
A localização é um fator que tende a influenciar no valor de um imóvel. Apesar de não ser o bairro com maior número de anuncios, o bairro de Jurerê é o que possui o maior faturamento.

```
# ################
# Bairro
# ################
sel3=merge(details,sel21,by = "airbnb_listing_id")

tab31 = aggregate(price_string ~ suburb, sel3, mean)
ggplot(sel3, aes(y = price_string, x = suburb)) +
  labs(title = "Faturamento por bairro",
       x = "Bairro",
       y = "Faturamento")+
  geom_boxplot()
```

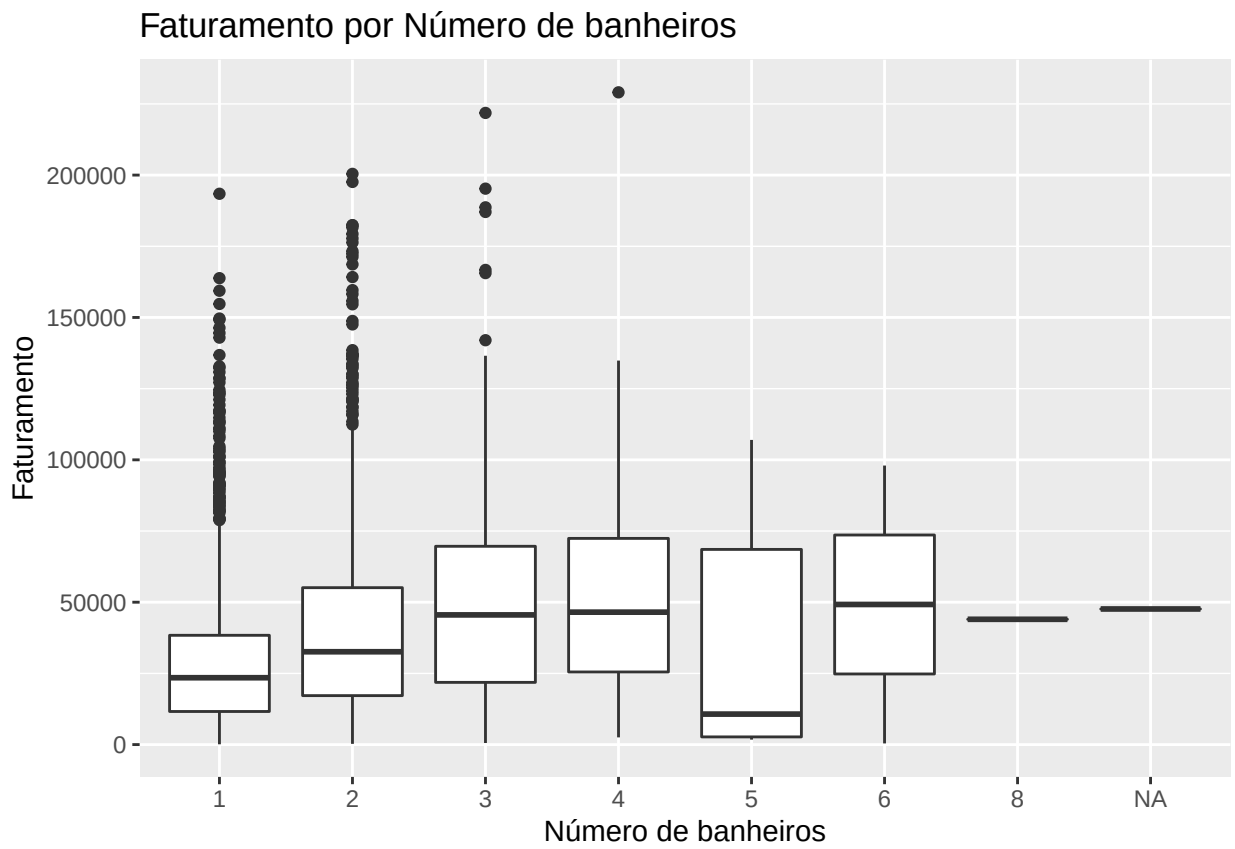


Outro ponto importante é o tamanho do imóvel, quanto maior for, maior é o valor agregado a ele. Desse modo, pode-se notar que quanto maior o número de quartos e banheiros, maior é o faturamento.

```
# ##################
# Número de quartos
# ##################
sel3[,5] = as.factor(as.numeric(sel3[,5]))

ggplot(sel3, aes(y = price_string, x = number_of_bedrooms)) +
  labs(title = "Faturamento por Número de quartos",
       x = "Número de quartos",
       y = "Faturamento")+
  geom_boxplot()
```

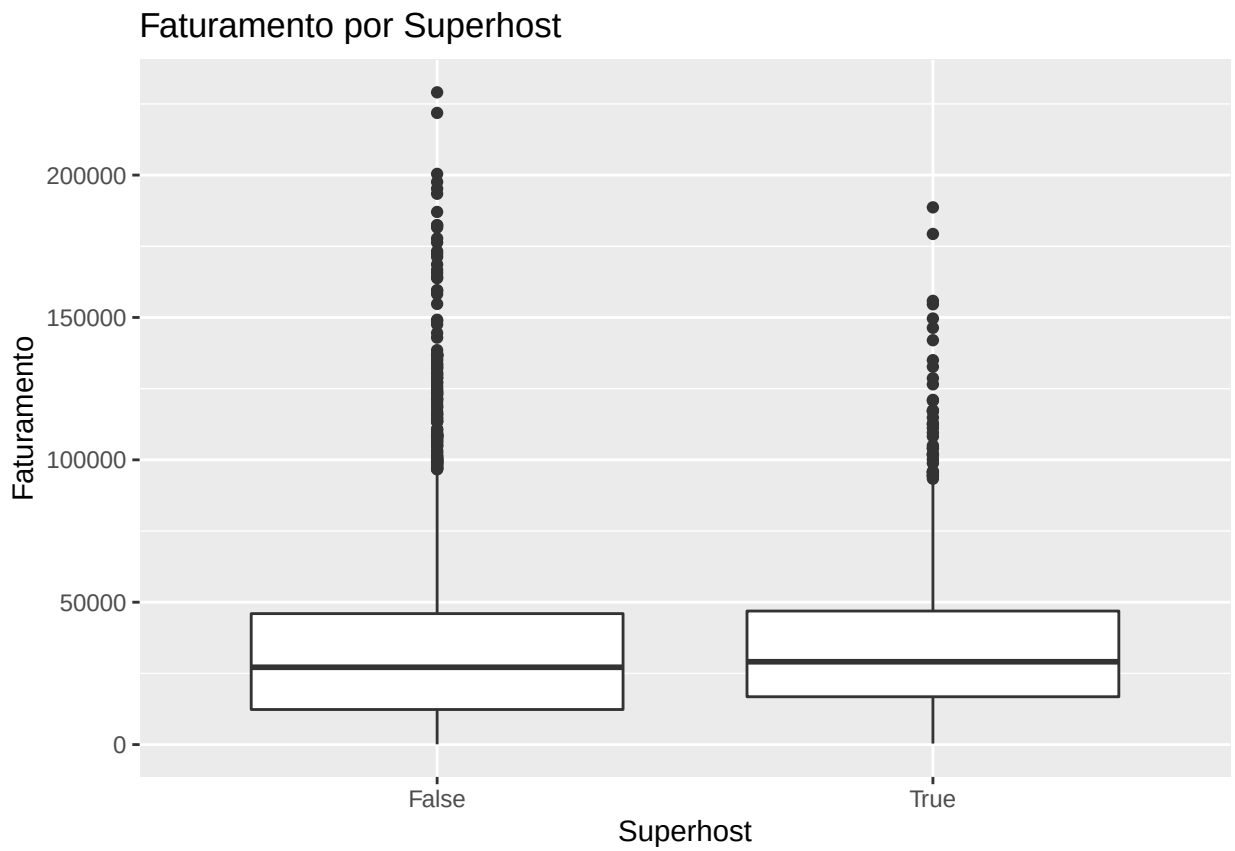



```
# ####################
# Número de banheiros
# ####################

sel3[,6] = as.factor(as.numeric(sel3[,6]))

ggplot(sel3, aes(y = price_string, x = number_of_bathrooms)) +
  labs(title = "Faturamento por Número de banheiros",
       x = "Número de banheiros",
       y = "Faturamento")+
  geom_boxplot()
```



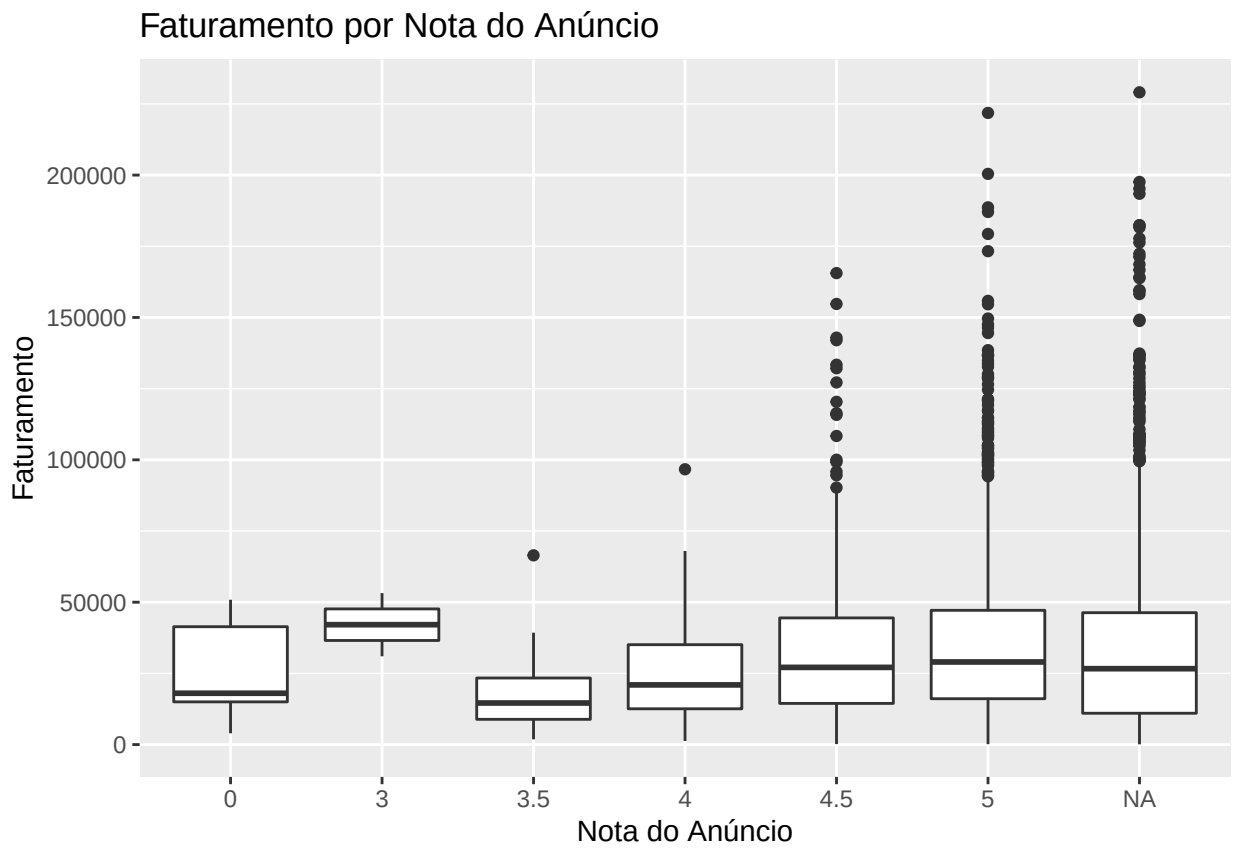
Entre os imóveis que são superhost têm-se que o faturamento é levemente mais alto do que entre aqueles que não são. Para constatar se essa diferença média é significativa, vale realizar algum teste.

```
# #############
# Superhost
# ############

ggplot(sel3, aes(y = price_string, x = is_superhost)) +
  labs(title = "Faturamento por Superhost",
       x = "Superhost",
       y = "Faturamento")+
  geom_boxplot()
```



A nota do anúncio não mostrou ter relação com o seu faturamento.

```
# ###################
# Nota do Anúncio
# ###################
sel3[,7] = as.factor(as.numeric(sel3[,7]))

ggplot(sel3, aes(y = price_string, x = star_rating)) +
  labs(title = "Faturamento por Nota do Anúncio",
       x = "Nota do Anúncio",
       y = "Faturamento")+
  geom_boxplot()
```



Quanto o número de reviews, a correlação mostrou-se muito fraca, ou seja, também não tem impacto relevante sobre o faturamento.

```
# ###################
# Número de Reviews
# ###################
cor(sel3$price_string,as.numeric(sel3$number_of_reviews),use = "pairwise")
```

```
## [1] 0.01422069
```

Antecedencia das reservas

Sobre a antecedência das reservas, têm-se que em média as reservas são realizadas com 32 dias de antecedência.

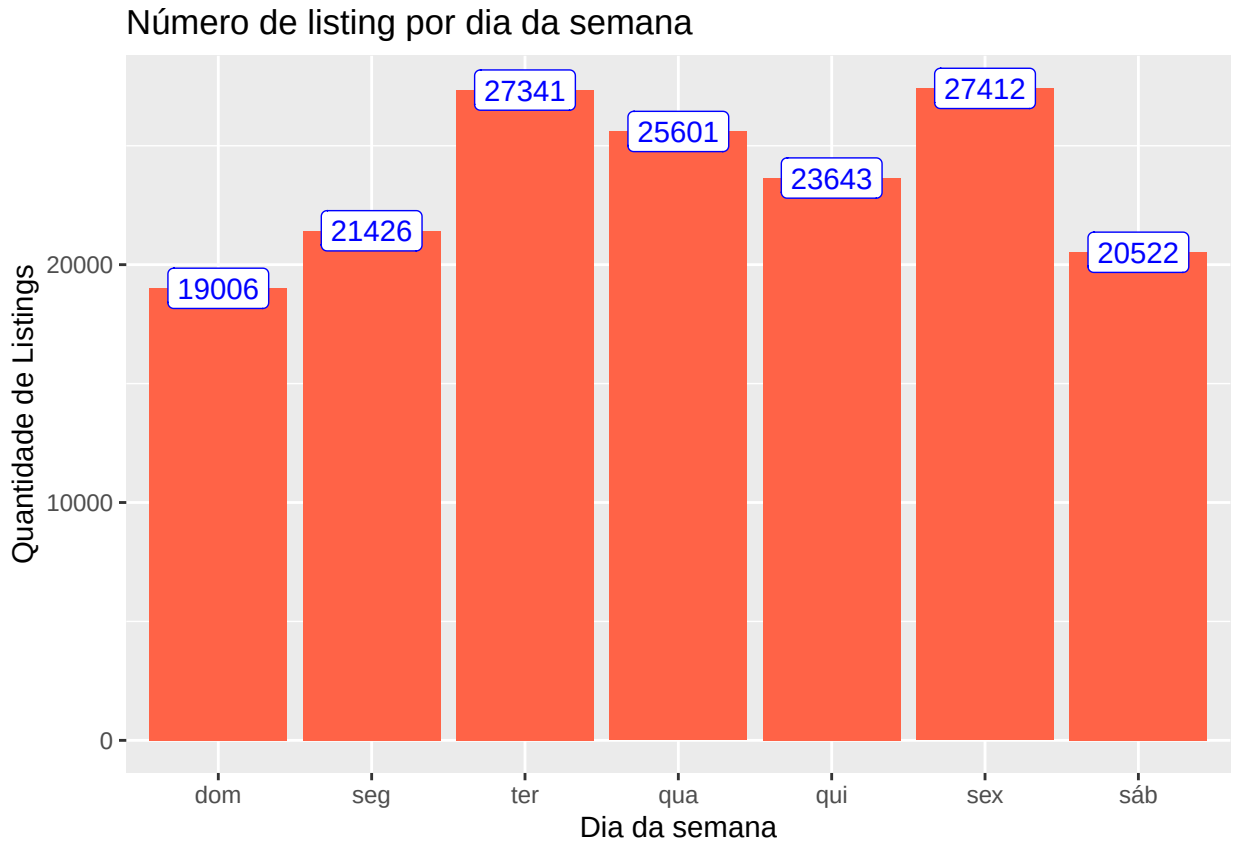
```
df=dados$date  # data alugada
dr=dados$booked_on  # data da reserva
dt=data.frame(dr,df);dt=na.omit(dt)

for(i in 1:length(dt$dr)){if(dt$dr[i]=="blank"){dt$dr[i]=NA}}
dt=na.omit(dt)
dt$dr=as.Date(dt$dr)

antc = dt$df-dt$dr  # dias de antecedência
mean(antc)
```

```
## Time difference of 32.35171 days
```

Antecedencia em fins de semana

Pode-se constatar que também que a reserva é realizada com maior frequência durante a semana, tendo sábado e domingo com os menores quantitativos.

```
# ###############
# Dias da semana
# ###############
ds <- wday(dt$dr,label = T)

tab8 = as.data.frame(table(ds));
ggplot(tab8, aes(x =ds, y = Freq), fill = Freq) +
  geom_col(position = "dodge",fill = "tomato") +
  geom_label(aes(label = Freq),colour="blue") +
  labs(title = "Número de listing por dia da semana",
       x = "Dia da semana",
       y = "Quantidade de Listings")
```



Feedback do Processo seletivo

Sobre o processo seletivo da Seazone, achei muito interessante essa abordagem de solictar a análise de dados relacionados a atividades da empresa por permitir-me mostrar um pouco das minhas habilidades para a vaga

de Analista de dados. Os dados oferecidos são bem construídos, possibilitando assim mais facilidade para manusear e extrair informações.

Desde já agradeço a oportunidade de participar deste processo seletivo de uma empresa que muito tem a acrescentar à minha experiência profissional.

Aguardarei o feedback sobre esta etapa.