# Análise sobre as informações públicas do Airbnb.

## Caso de estudo: Rio de Janeiro

Base de dados: Anúncios de hospedagem do site Airbnb existentes na cidade do Rio de Janeiro, em 21/05/2019. Fonte: http://insideairbnb.com/rio-de-janeiro/# A base foi limitada a anúncios de bairros que possuam ao menos 50 acomodações anunciadas para evitar distorções de casos isolados.

## 1. Quais os bairros com mais apartamentos anunciados?

```
bairrosMaisApt = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(totalApartamentos = n_distinct(id))%>%
  arrange(-totalApartamentos)

bairrosMaisApt
```

```
## # A tibble: 37 x 2
##    bairro                   totalApartamentos
##    <chr>                                <int>
##  1 Copacabana                            8957
##  2 Barra da Tijuca                       5848
##  3 Ipanema                               3021
##  4 Recreio dos Bandeirantes              1941
##  5 Botafogo                              1792
##  6 Leblon                                1662
##  7 Santa Teresa                          1110
##  8 Flamengo                               996
##  9 Tijuca                                 809
## 10 Laranjeiras                            765
## # ... with 27 more rows
```

```
bairrosMaisAptGeral = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(totalAnuncios = n_distinct(id),
            precoMedio=mean(price, na.rm=TRUE),
            reviewsMedia=mean(number_of_reviews),
            reviewsMediaMes=mean(reviews_per_month, na.rm=TRUE)) %>%
  arrange(-totalAnuncios)

bairrosMaisAptGeral
```

```
## # A tibble: 37 x 5
##    bairro             totalAnuncios precoMedio reviewsMedia reviewsMediaMes
##    <chr>                      <int>      <dbl>        <dbl>           <dbl>
##  1 Copacabana                  8957       492.        12.0            0.704
##  2 Barra da Tijuca             5848      1018.         4.48           0.560
##  3 Ipanema                     3021       665.        15.5            0.795
##  4 Recreio dos Bande~          1941       773.         2.41           0.483
```

```
##  5 Botafogo                    1792       401.         6.70           0.527
##  6 Leblon                      1662       764.        10.9            0.647
##  7 Santa Teresa                1110       456.         7.97           0.537
##  8 Flamengo                     996       423.         6.03           0.480
##  9 Tijuca                       809       465.         2.47           0.439
## 10 Laranjeiras                  765       492.         4.89           0.435
## # ... with 27 more rows
```

## 2. Quais os bairros mais caros?

```
bairrosPreco = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(precoMedio = mean(price)) %>%
  arrange(-precoMedio)

bairrosPreco
```

```
## # A tibble: 37 x 2
##    bairro                   precoMedio
##    <chr>                         <dbl>
##  1 Joá                           2782.
##  2 São Conrado                   1548.
##  3 Barra da Tijuca               1018.
##  4 Lagoa                          967.
##  5 Barra de Guaratiba             919.
##  6 Jardim Botânico                866.
##  7 Recreio dos Bandeirantes       773.
##  8 Leblon                         764.
##  9 Gávea                          760.
## 10 Andaraí                        713.
## # ... with 27 more rows
```

```
bairrosPrecoGeral = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(precoMedio=mean(price, na.rm=TRUE),
            totalAnuncios = n_distinct(id),
            reviewsMedia=mean(number_of_reviews),
            reviewsMediaMes=mean(reviews_per_month, na.rm=TRUE)) %>%
  arrange(-precoMedio)

bairrosPrecoGeral
```

```
## # A tibble: 37 x 5
##    bairro             precoMedio totalAnuncios reviewsMedia reviewsMediaMes
##    <chr>                   <dbl>         <int>        <dbl>           <dbl>
##  1 Joá                     2782.            93         2.86           0.463
##  2 São Conrado             1548.           286         1.41           0.304
##  3 Barra da Tijuca         1018.          5848         4.48           0.560
##  4 Lagoa                    967.           395         5.91           0.419
##  5 Barra de Guaratiba       919.            88         3.17           0.358
##  6 Jardim Botânico          866.           347         2.53           0.317
```

```
##  7 Recreio dos Bande~       773.          1941         2.41           0.483
##  8 Leblon                   764.          1662        10.9            0.647
##  9 Gávea                    760.           338         3.63           0.318
## 10 Andaraí                  713.            97         1.32           0.38
## # ... with 27 more rows
```

### 3. Quais os bairros mais baratos?

```
bairrosPreco = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(precoMedio = mean(price)) %>%
  arrange(precoMedio)

bairrosPreco
```

```
## # A tibble: 37 x 2
##    bairro            precoMedio
##    <chr>                  <dbl>
##  1 Estacio                 184.
##  2 Méier                   258.
##  3 Centro                  263.
##  4 Catete                  285.
##  5 Lapa                    288.
##  6 Grajaú                  297.
##  7 Engenho Novo            297.
##  8 São Cristóvão           343.
##  9 Rio Comprido            353.
## 10 Praça da Bandeira       398.
## # ... with 27 more rows
```

```
bairrosPrecoGeral = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(precoMedio=mean(price, na.rm=TRUE),
            totalAnuncios = n_distinct(id),
            reviewsMedia=mean(number_of_reviews),
            reviewsMediaMes=mean(reviews_per_month, na.rm=TRUE)) %>%
  arrange(precoMedio)

bairrosPrecoGeral
```

```
## # A tibble: 37 x 5
##    bairro            precoMedio totalAnuncios reviewsMedia reviewsMediaMes
##    <chr>                  <dbl>         <int>        <dbl>           <dbl>
##  1 Estacio                 184.            82        4.21           0.465
##  2 Méier                   258.            59        0.475          0.19
##  3 Centro                  263.           582        8.47           0.791
##  4 Catete                  285.           333        5.77           0.531
##  5 Lapa                    288.           484        7.95           0.738
##  6 Grajaú                  297.           111        0.946          0.306
##  7 Engenho Novo            297.            53        0.509          0.254
##  8 São Cristóvão           343.           130        5.49           0.493
```

```
##  9 Rio Comprido            353.           118        2.14           0.336
## 10 Praça da Bandeira       398.            77        2.14           0.564
## # ... with 27 more rows
```

## 4. Quais os bairros com maior taxa de ocupação?

Segundo o projeto "Inside Airbnb", a base de dados foi populada a partir dos dados públicos capturados do site. Como não existe abertamente no site os dados sobre os aluguéis realizados de cada hospedagem anunciada, a taxa de ocupação dos imóveis pode ser estimada a partir da quantidade de reviews de cada anúncio. De acordo com o projeto estima-se que em 50% dos casos os usuários deixam reviews, assim quanto maior a quantidade de reviews mais podemos considerar que o anúncio foi reservado pelo site.

```
bairrosMaisReservas = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(mediaReviews = mean(number_of_reviews)) %>%
  arrange(-mediaReviews)

bairrosMaisReservas
```

```
## # A tibble: 37 x 2
##    bairro       mediaReviews
##    <chr>               <dbl>
##  1 Ipanema             15.5
##  2 Copacabana          12.0
##  3 Leblon              10.9
##  4 Urca                 9.60
##  5 Leme                 9.05
##  6 Centro               8.47
##  7 Vidigal              8.45
##  8 Santa Teresa         7.97
##  9 Lapa                 7.95
## 10 Glória               7.88
## # ... with 27 more rows
```

```
bairrosMaisReservasGeral = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(reviewsMedia=mean(number_of_reviews),
            totalAnuncios = n_distinct(id),
            precoMedio=mean(price, na.rm=TRUE),
            reviewsMediaMes=mean(reviews_per_month, na.rm=TRUE)) %>%
  arrange(-reviewsMedia)

bairrosMaisReservasGeral
```

```
## # A tibble: 37 x 5
##    bairro       reviewsMedia totalAnuncios precoMedio reviewsMediaMes
##    <chr>               <dbl>         <int>      <dbl>           <dbl>
##  1 Ipanema             15.5           3021       665.           0.795
##  2 Copacabana          12.0           8957       492.           0.704
##  3 Leblon              10.9           1662       764.           0.647
##  4 Urca                 9.60           164       651.           0.724
##  5 Leme                 9.05           646       478.           0.622
```

```
##  6 Centro               8.47          582       263.          0.791
##  7 Vidigal              8.45          187       474.          0.492
##  8 Santa Teresa         7.97         1110       456.          0.537
##  9 Lapa                 7.95          484       288.          0.738
## 10 Glória               7.88          357       444.          0.515
## # ... with 27 more rows
```

## 5. Quais os bairros com maior frequência de ocupação mensal?

Se observarmos apenas a média de reviews, podemos não considerar situações onde ocorreu uma alta ocupação esporádica em determinado bairro, mas que não corresponda uma frequência habitual de ocupação. Assim, cabe tambpem observar a média mensal de ocupação de cada bairro.

```
bairrosMaisReservasMensal = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(reviewsMediaMes = mean(reviews_per_month, na.rm=TRUE)) %>%
  arrange(-reviewsMediaMes)

bairrosMaisReservasMensal
```

```
## # A tibble: 37 x 2
##    bairro            reviewsMediaMes
##    <chr>                       <dbl>
##  1 Rocha                       0.879
##  2 Ipanema                     0.795
##  3 Centro                      0.791
##  4 Lapa                        0.738
##  5 Urca                        0.724
##  6 Copacabana                  0.704
##  7 Leblon                      0.647
##  8 Leme                        0.622
##  9 Praça da Bandeira           0.564
## 10 Barra da Tijuca             0.560
## # ... with 27 more rows
```

```
bairrosMaisReservasMensalGeral = airbnb.rio %>%
  group_by(bairro) %>%
  summarise(reviewsMediaMes=mean(reviews_per_month, na.rm=TRUE),
            totalAnuncios = n_distinct(id),
            precoMedio=mean(price, na.rm=TRUE),
            reviewsMedia=mean(number_of_reviews)) %>%
  arrange(-reviewsMediaMes)

bairrosMaisReservasMensalGeral
```

```
## # A tibble: 37 x 5
##    bairro            reviewsMediaMes totalAnuncios precoMedio reviewsMedia
##    <chr>                       <dbl>         <int>      <dbl>        <dbl>
##  1 Rocha                       0.879            58       423.         1.97
##  2 Ipanema                     0.795          3021       665.        15.5
##  3 Centro                      0.791           582       263.         8.47
##  4 Lapa                        0.738           484       288.         7.95
```

```
##  5 Urca                      0.724          164       651.         9.60
##  6 Copacabana                0.704         8957       492.        12.0
##  7 Leblon                    0.647         1662       764.        10.9
##  8 Leme                      0.622          646       478.         9.05
##  9 Praça da Bandeira         0.564           77       398.         2.14
## 10 Barra da Tijuca           0.560         5848      1018.         4.48
## # ... with 27 more rows
```

### 6. Qual o tipo de acomodação mais anunciado e o mais procurado (de acordo com quantidade de reviews)?

```
tipoAptMaisProcurado = airbnb.rio %>%
  group_by(room_type) %>%
  summarise(reviewsMediaMes = mean(reviews_per_month, na.rm=TRUE),
            totalAnuncios = n_distinct(id),
            precoMedio=mean(price, na.rm=TRUE)) %>%
  arrange(-reviewsMediaMes)

tipoAptMaisProcurado
```

```
## # A tibble: 3 x 4
##   room_type       reviewsMediaMes totalAnuncios precoMedio
##   <chr>                     <dbl>         <int>      <dbl>
## 1 Entire home/apt           0.648         23971       787.
## 2 Private room              0.543          8464       262.
## 3 Shared room               0.363           717       213.
```

### 7. Quais os apartamentos com maior número de reviews, onde fica e qual seu valor/dia?

```
aptMaisOcupado = airbnb.rio %>%
  select(bairro, listing_url, number_of_reviews, price) %>%
  arrange(-number_of_reviews) %>%
  rename(nReviews = number_of_reviews)
aptMaisOcupado
```

```
## # A tibble: 33,152 x 4
##    bairro       listing_url                              nReviews price
##    <chr>        <chr>                                       <dbl> <dbl>
##  1 Lagoa        https://www.airbnb.com/rooms/273463           338   306
##  2 Copacabana   https://www.airbnb.com/rooms/996602           337   200
##  3 Ipanema      https://www.airbnb.com/rooms/223073           334    94
##  4 Copacabana   https://www.airbnb.com/rooms/672835           327   322
##  5 Copacabana   https://www.airbnb.com/rooms/10730455         312    69
##  6 Ipanema      https://www.airbnb.com/rooms/70080            306   367
##  7 Vidigal      https://www.airbnb.com/rooms/494903           304   200
##  8 Copacabana   https://www.airbnb.com/rooms/35764            295   220
##  9 Santa Teresa https://www.airbnb.com/rooms/219250           277   159
## 10 Copacabana   https://www.airbnb.com/rooms/2092178          275   232
## # ... with 33,142 more rows
```

## 8. Quais os apartamentos com maior frequência de ocupação mensal, onde ficam e qual seu valor/dia?

```
aptMaisOcupadoMes = airbnb.rio %>%
  select(bairro, listing_url, reviews_per_month, price) %>%
  arrange(-reviews_per_month)%>%
  rename(nReviewsMes = reviews_per_month)
aptMaisOcupadoMes
```

```
## # A tibble: 33,152 x 4
##    bairro       listing_url                              nReviewsMes price
##    <chr>        <chr>                                          <dbl> <dbl>
##  1 Santa Teresa https://www.airbnb.com/rooms/273753           10       49
##  2 Copacabana   https://www.airbnb.com/rooms/29751978          9.89   106
##  3 Centro       https://www.airbnb.com/rooms/32165301          8.39   139
##  4 Centro       https://www.airbnb.com/rooms/27253710          8.38    73
##  5 Lapa         https://www.airbnb.com/rooms/19505341          8.21   131
##  6 Copacabana   https://www.airbnb.com/rooms/10730455          7.77    69
##  7 Botafogo     https://www.airbnb.com/rooms/27286419          7.31   131
##  8 Catete       https://www.airbnb.com/rooms/32179928          7.28   163
##  9 Centro       https://www.airbnb.com/rooms/32500735          7.11   118
## 10 Copacabana   https://www.airbnb.com/rooms/17865987          7.04    86
## # ... with 33,142 more rows
```

## 9. Quantos anuncios tem os 5 maiores anunciantes?

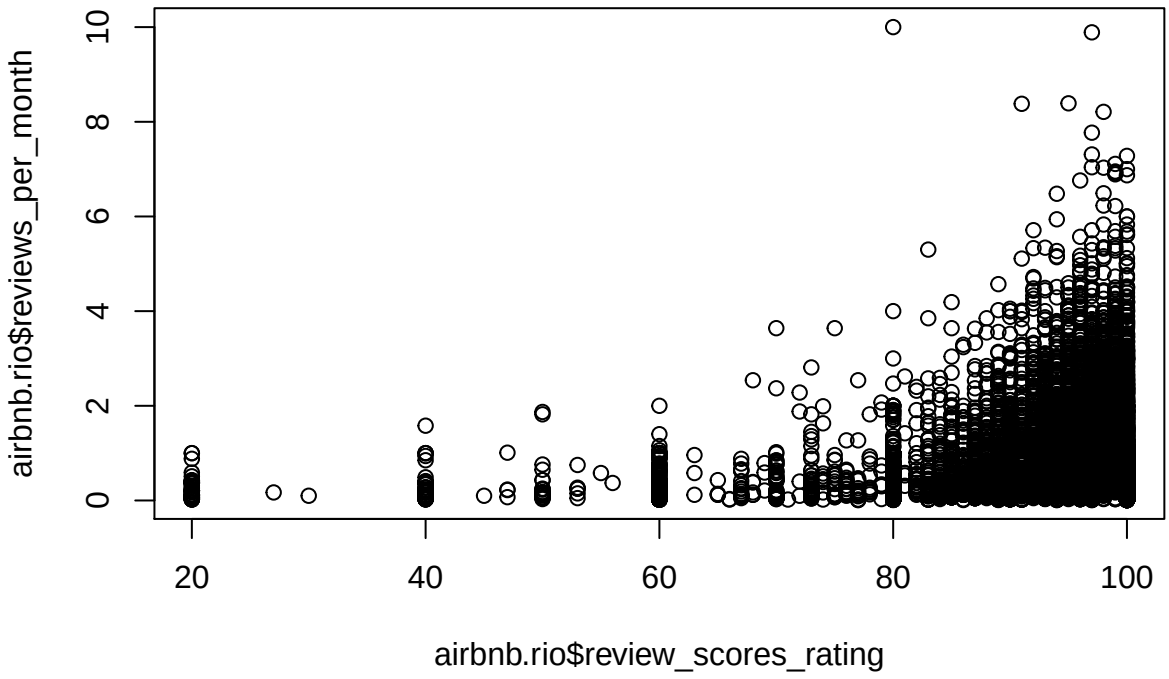
```
totalPorAnunciante = airbnb.rio %>%
  group_by(host_id, host_url) %>%
  summarise(totalAnuncios = n_distinct(id))%>%
  arrange(-totalAnuncios)%>%
  head(n = 5)
totalPorAnunciante
```

```
## # A tibble: 5 x 3
## # Groups:   host_id [5]
##    host_id host_url                                    totalAnuncios
##      <dbl> <chr>                                               <int>
## 1 91654021 https://www.airbnb.com/users/show/91654021          250
## 2 81876389 https://www.airbnb.com/users/show/81876389          238
## 3 31275569 https://www.airbnb.com/users/show/31275569          123
## 4 22805631 https://www.airbnb.com/users/show/22805631           74
## 5  1982737 https://www.airbnb.com/users/show/1982737            63
```

## 10. Modelo de Regressao Linear

- Qual a relação entre as avaliações recebidas e taxa de ocupação mensal da acomodação? Gráfico de Dispersão

```r
plot(airbnb.rio$review_scores_rating, airbnb.rio$reviews_per_month)
```




Coeficiente de Correção Linear

```r
cor(!is.na(airbnb.rio$review_scores_rating), !is.na(airbnb.rio$reviews_per_month))
```

```
## [1] 0.9456286
```

Teste de Hipótese

```r
cor.test(airbnb.rio$review_scores_rating, airbnb.rio$reviews_per_month)
```

```
##
##  Pearson's product-moment correlation
##
## data:  airbnb.rio$review_scores_rating and airbnb.rio$reviews_per_month
## t = 9.7588, df = 17357, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05905956 0.08864958
## sample estimates:
##        cor
## 0.07387083
```

### Ajustes do Modelo de Regressao Linear

```
model = lm(reviews_per_month~review_scores_rating, data=airbnb.rio)
model
```

```
##
## Call:
## lm(formula = reviews_per_month ~ review_scores_rating, data = airbnb.rio)
##
## Coefficients:
##          (Intercept)  review_scores_rating
##             0.022427              0.006541
```

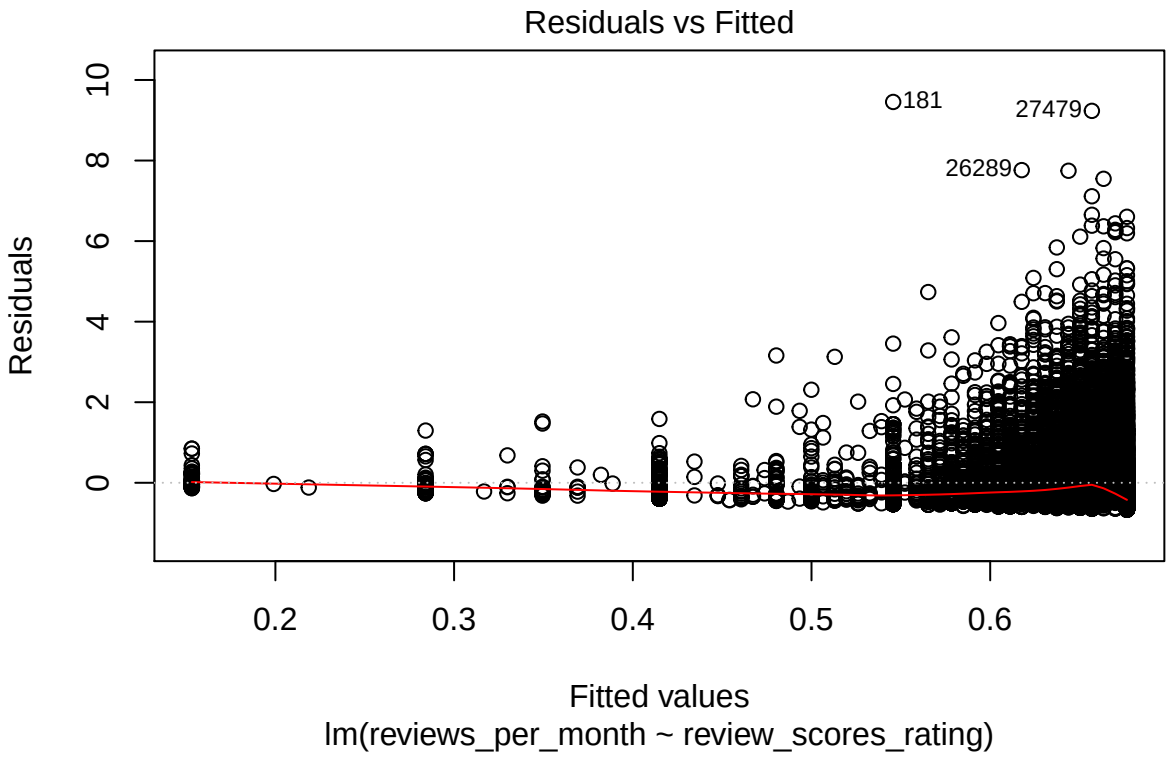
```
summary(model)
```

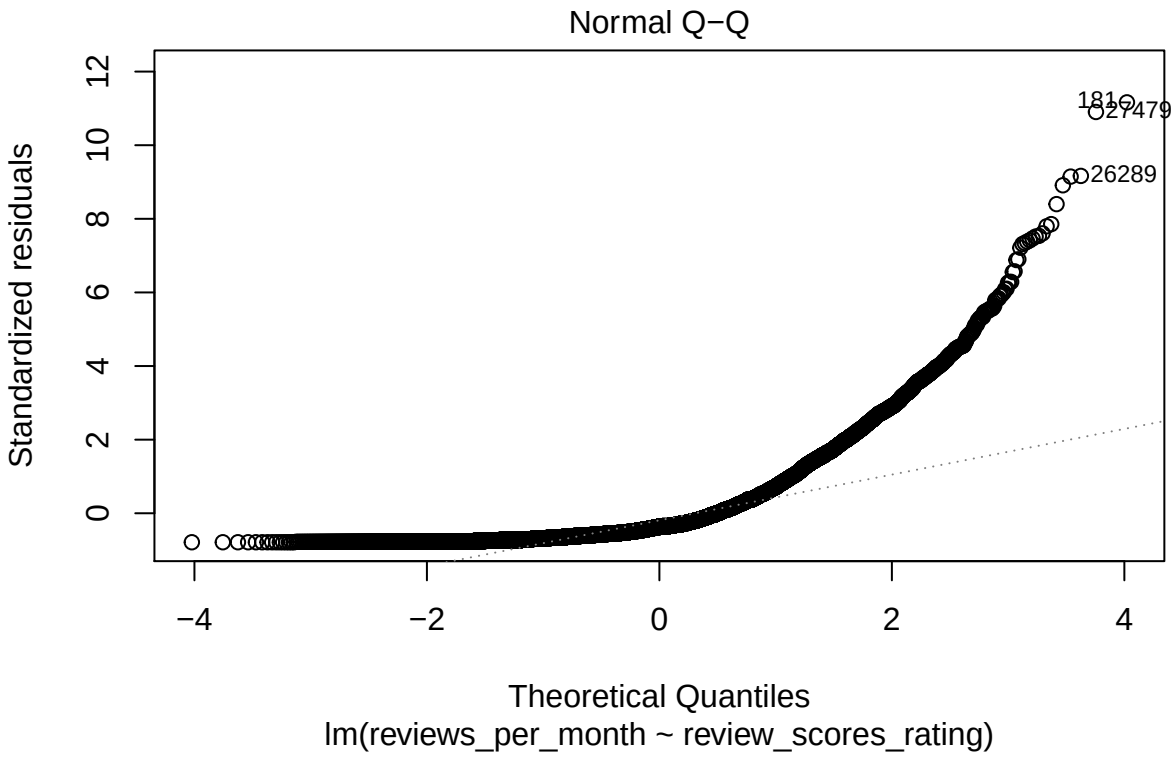
```
##
## Call:
## lm(formula = reviews_per_month ~ review_scores_rating, data = airbnb.rio)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6665 -0.5165 -0.3107  0.1927  9.4543
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.0224269  0.0635896   0.353    0.724
## review_scores_rating 0.0065409  0.0006703   9.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8469 on 17357 degrees of freedom
##   (15793 observations deleted due to missingness)
## Multiple R-squared:  0.005457,   Adjusted R-squared:  0.0054
## F-statistic: 95.24 on 1 and 17357 DF,  p-value: < 2.2e-16
```

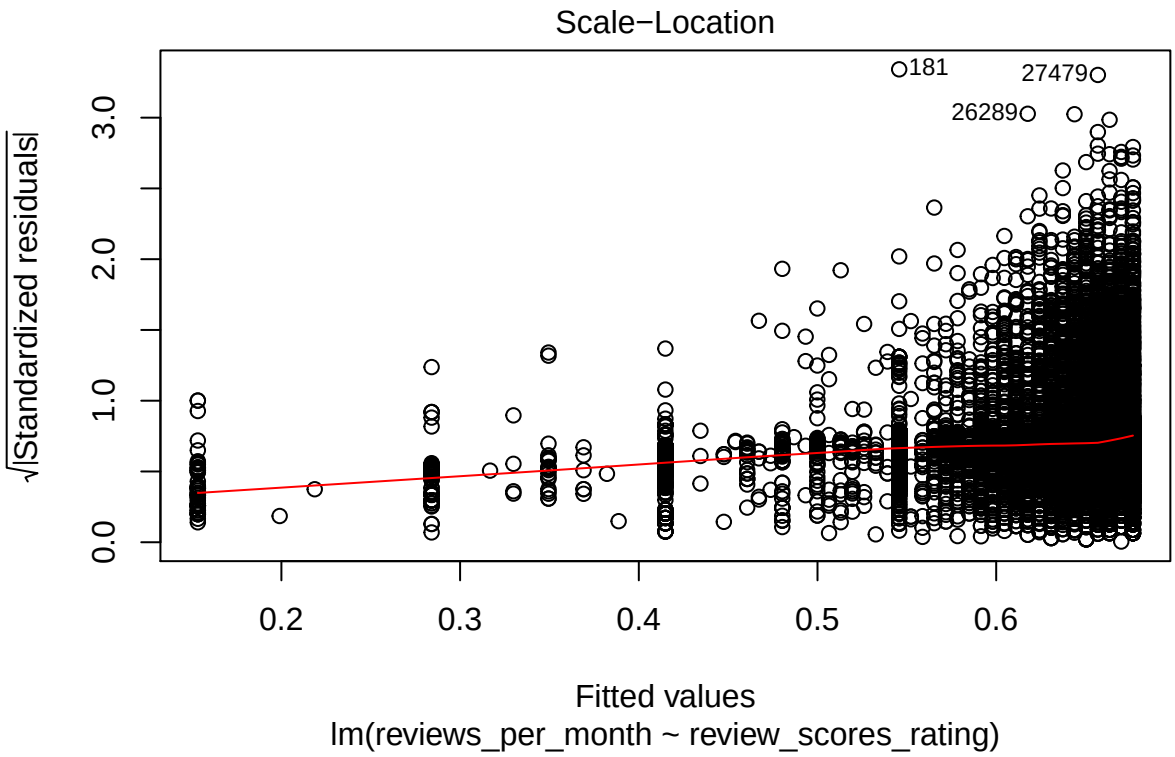
```
anova(model)
```

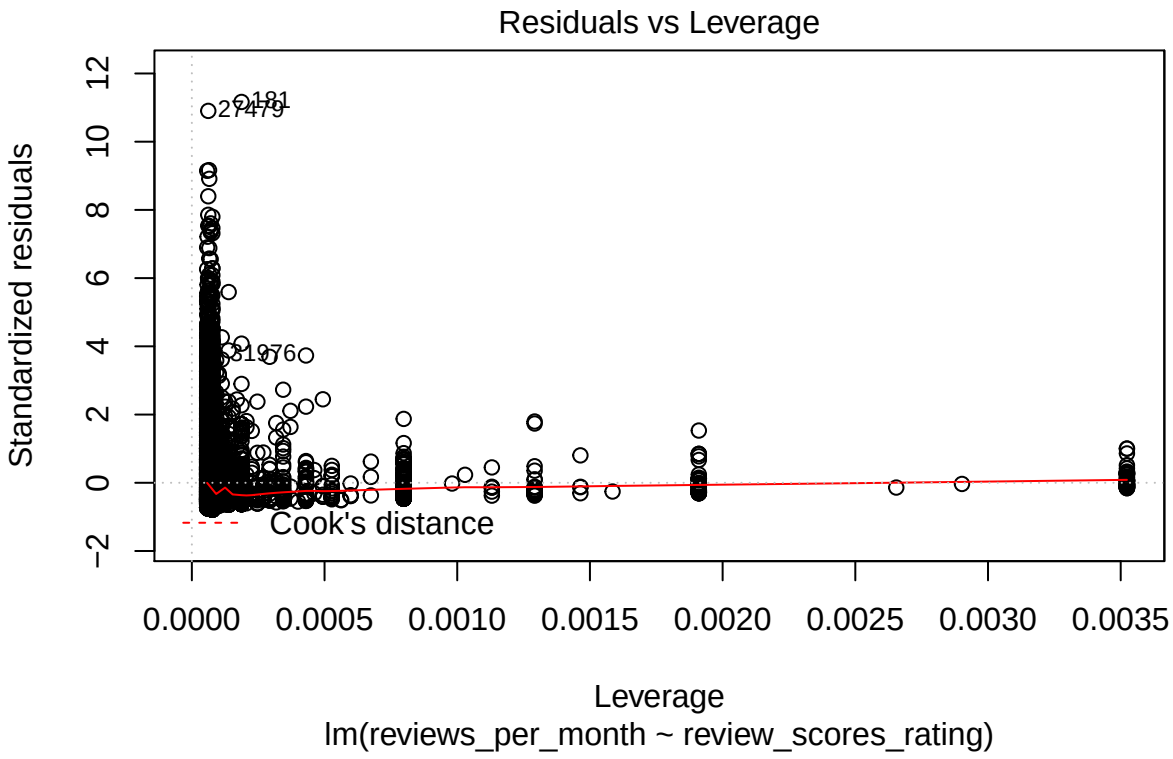
```
## Analysis of Variance Table
##
## Response: reviews_per_month
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## review_scores_rating     1    68.3  68.306  95.235 < 2.2e-16 ***
## Residuals            17357 12449.1   0.717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model)
```


Residuals vs Fitted
Residuals
0
2
4
6
8
10
181
27479
26289
0.2
0.3
0.4
0.5
0.6
Fitted values
lm(reviews_per_month ~ review_scores_rating)


Normal Q−Q
Standardized residuals
0
2
4
6
8
10
12
1810
27479
26289
−4
−2
0
2
4
Theoretical Quantiles
lm(reviews_per_month ~ review_scores_rating)

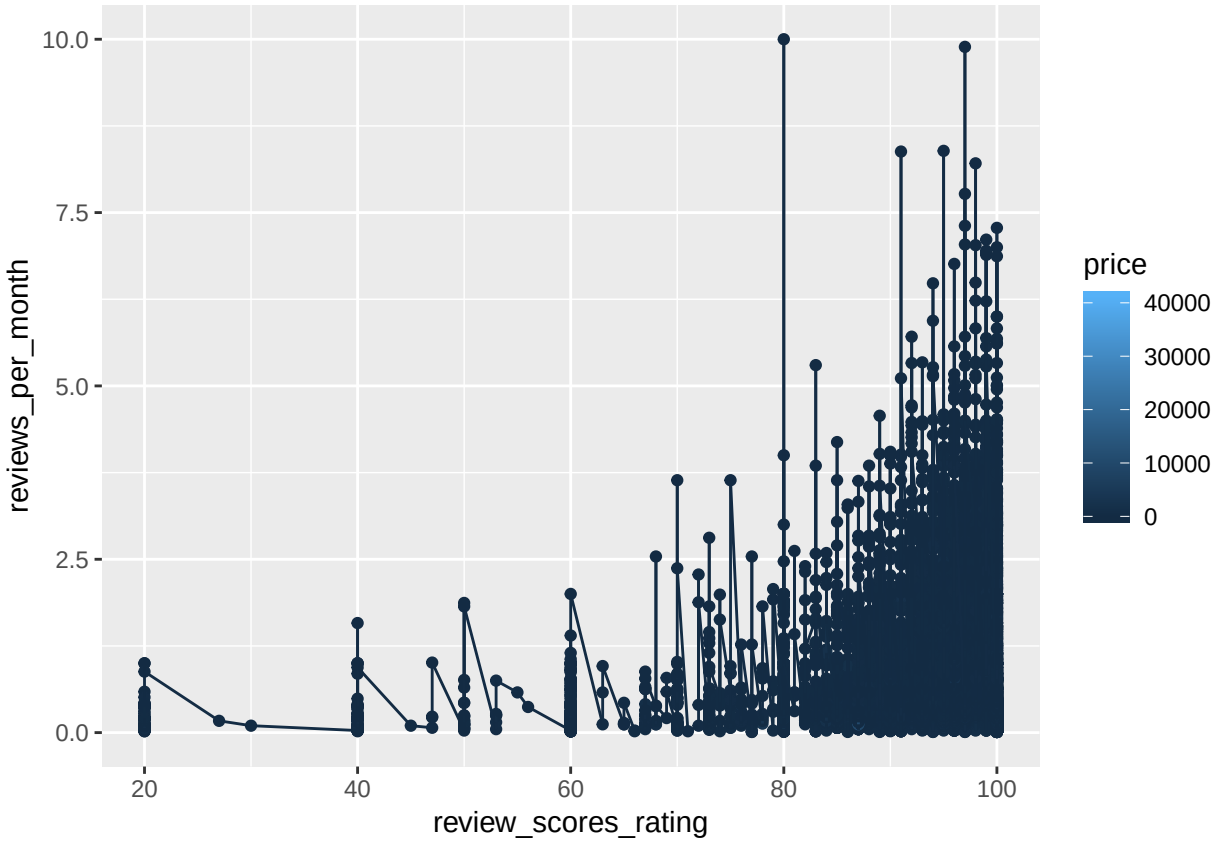

Scale−Location
181
27479
26289
√|Standardized residuals|
0.0
1.0
2.0
3.0
0.2
0.3
0.4
0.5
0.6
Fitted values
lm(reviews_per_month ~ review_scores_rating)




```
ggplot(airbnb.rio, aes(x=review_scores_rating,y=reviews_per_month,color=price))+ geom_point() + geom_li
```

```
## Warning: Removed 15793 rows containing missing values (geom_point).
```

```
## Warning: Removed 15791 rows containing missing values (geom_path).
```


10.0
7.5
5.0
2.5
0.0
reviews_per_month
20
40
60
80
100
review_scores_rating
price
40000
30000
20000
10000
0