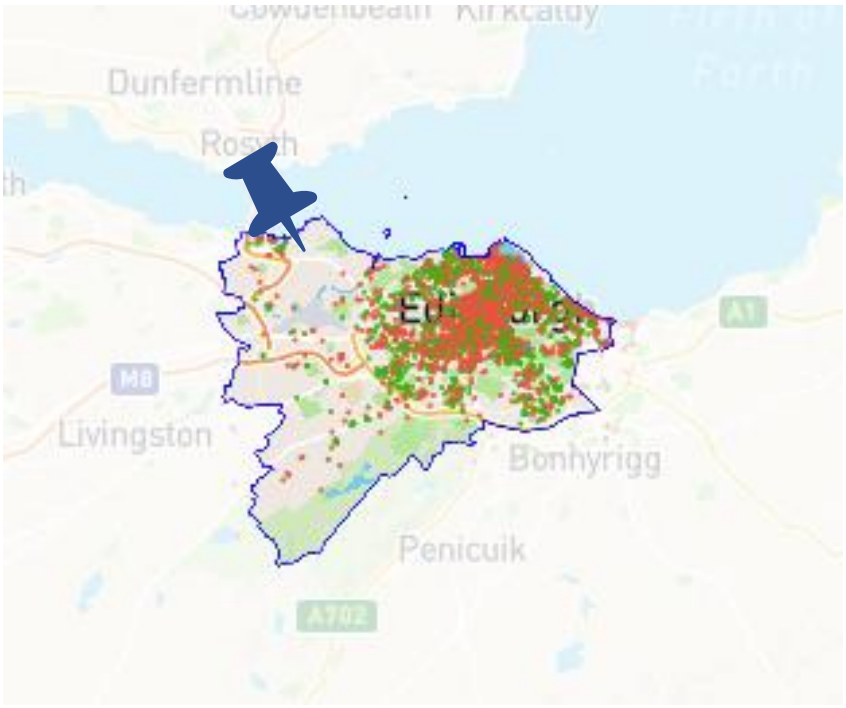
INSTITUTO UNIVERSITÁRIO DE LISBOA

1ºSemestre Ano Letivo 2022/2023

Unidade Curricular de Introdução a Modelos Dinâmicos

Docente Diana Aldea Mendes



Airbnbs em Edinburg

2º ano - LCD – CDB1

Eliane Susso Efraim Gabriel, Nº 103303

Marco Delgado Esperança, Nº 110451

Maria João Ferreira Lourenço, Nº 104716

Umeima Adam Mahomed, Nº 99239

Lisboa, 21 de dezembro de 2022

Índice

Introdução

Entre 2012 e o presente ano, 2022, foram recolhidos dados acerca dos Airbnb's existentes na cidade de Edinburg, na Escócia (Figura 1). Estes dados permitem estudar um conjunto de fatores que têm a possibilidade de influenciar o impacto dos Airbnb's em zonas residenciais, para que os governos possam atuar caso seja necessário no arrendamento destes aos turistas.

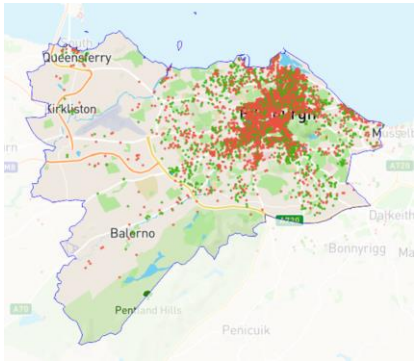


Figura 1-Airbnb's existentes na Escócia

Neste trabalho, o nosso objetivo é o de estudar a influência de um conjunto de variáveis presentes na base de dados, como por exemplo, *neighbourhood*, *room_type*, *latitude*, *longitude*, *availability_365*, entre outras no preço do aluguer dos Airbnb's, pelo que a variável *price* é o nosso target.

Business Understanding

Antes de começar a estudar de forma estatística as variáveis estudamos o seu significado, ou seja, ao que se referem. Depois de uma investigação inicial chegamos às conclusões enunciadas em baixo, é de notar que algumas variáveis estavam vazias, tal como indicado nesses casos:

- ***id***: identificador único de cada Airbnb;
- ***name***: breve descrição com características do Airbnb;
- ***host_id***: identificação numérica única do anfitrão;
- ***host_name***: nome próprio do anfitrião;
- ***neighbourhood_group***: vazio, após a limpeza, mas deve agrupar bairros individuais num conjunto maior de bairros;
- ***neighbourhood***: bairro a que o Airbnb pertence;
- ***latitude***: distância ao Equador, medida ao longo do Meridiano de Greenwich, associada ao Norte e ao Sul, medida em graus e varia de 0º a 90º;

- ***longitude***: distância ao Meridiano de Greenwich medida ao longo do Equador, associada ao Este e Oeste, medida em graus e varia de 0º a 180º;
- ***room_type***: tipos de alojamentos, por exemplo se é apartamento ou hotel, se é o apartamento inteiro ou um quarto privado ou partilhado;
- ***price***: preço do alojamento, provavelmente por noite e em libras (moeda utilizada em Edinburg);
- ***minimum_nights***: número mínimo de noites a ficar alojado no Airbnb;
- ***number_of_reviews***: número de avaliações do alojamento;
- ***last_review***: data da última avaliação;
- ***reviews_per_month***: média das avaliações mensais;
- ***calculated_host_listing_count***: número de anfitriões profissionais, presentes numa listagem ligada a um Airbnb específico;
- ***availability_365***: quantas noites por ano o Airbnb está disponível;
- ***number_of_reviews_ltm***: número de avaliações no último ano/12 meses;
- ***license***: vazio antes e após a limpeza, mas deve estar relacionado com a licença de alugamento necessária a um aluguer correto e legal

De seguida, procedemos à classificação das variáveis, consoante a sua tipologia (Figura 2).

Variável Quantitativa	Variável Qualitativa
• *host_id* • *latitude* • *longitude* • *price* • *minimum_nights* • *number_of_reviews* • *reviews_per_month* • *calculated_host_listing_count* • *availability_365* • *number_of_reviews_ltm* • *id*	• *name* • *host_name* • *neighbourhood* • *room_type* • *last_review*

Figura 2-Classificação das variáveis

Limpeza dos dados

Depois de importar a base de dados, decidimos visualizar com atenção os dados no Excel, uma vez que poderíamos usar o filtro (Figura 3) para perceber quais os valores que as diferentes variáveis poderiam ter, o que possibilitou a perceção de possíveis erros ou incongruências.

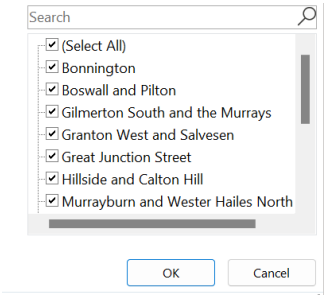


Figura 3- Utilização do filtro no Excel

Desta forma, notamos que a variável *license* não assumia qualquer valor (Figura 4), que quase todas as variáveis tinham alguns valores a *Null* e que a variável *neighbourhood* apresentava valores numéricos (Figura 5) - ao filtrar estes valores percebemos que todos os valores das linhas que estavam nestas consições teriam que ser movidos para a direita porque só assim fariam sentido dado o contexto, pelo que procedemos às alterações necessárias no Excel. Depois de esta estar concluída verificamos que a variável *neighbourhood_group* estava vazia (Figura 4).

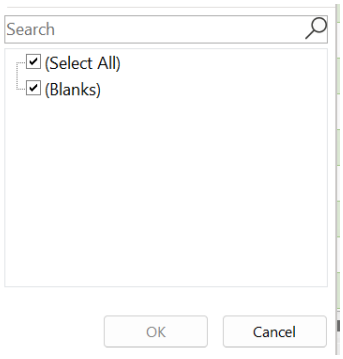


Figura 4-Valores nulos da variável neighbourhood_group e license

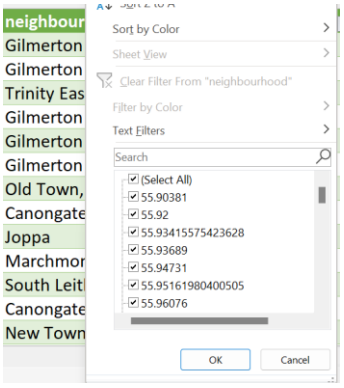


Figura 5-Valores numéricos iniciamente presentes em neighbourhood

De seguida, importamos o ficheiro Edinburg.csv – ficheiro de Excel que inclui a limpeza realizada nesta ferramenta - para o R, onde tivemos alguns problemas inicialmente porque dava erro no ficheiro Excel selecionado. Por isso, decidimos mudar as vírgulas que estavam nas variáveis com valores numéricos por pontos, isto ainda no Excel (Figura 6). Só assim conseguimos importar o ficheiro para o R e resolver os outros problemas ainda existentes.

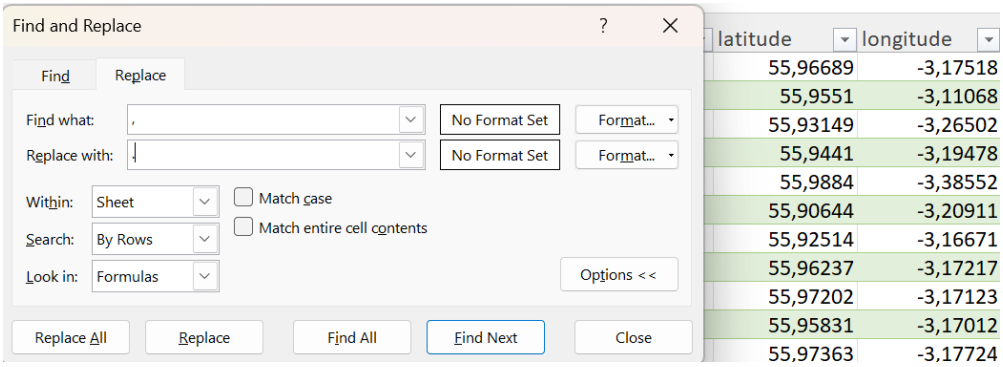


Figura 6-Mudança de vírgulas para pontos

No R, começamos por eliminar as colunas *id* – pois tinha valores NA e não precisamos desses valores para o nosso estudo -, *neighbourhood_group* e *license*, que estavam completamente vazias, através do comando `dados <- dados[-c(1, 5,18)]`

Realizamos o comando `apply(is.na(dados), 2, which)` que nos permitiu descobrir as linhas de cada uma das variáveis que tinham valores NA.

Para que todas as linhas ficassem totalmente preenchidas, decidimos substituir os valores NA pelas respetivas medianas (Figura 7) das seguintes variáveis: *latitude*, *longitude*, *calculated_host_listing_count*, *availability_365*, *number_of_reviews_ltm*, *price*, *minimum_nights*, *number_of_reviews* e *reviews_per_month*.

```
mediana <- median(dados$price,na.rm=TRUE)
mediana

dados$price[which(is.na(dados$price))] <- mediana
summary(is.na(dados$price))
```

Figura 7- Exemplo da substituição dos NA pela mediana

Nas restantes variáveis, ou seja, em *name*, *host_id*, *host_name*, *neighbourhood*, *room_type* e *last_review*, decidimos substituir os valores NA pela moda (Figura 8) destas variáveis, respetivamente.

```
dados$room_type <- replace(dados$room_type, dados$room_type=='', NA)

val1 <- unique(dados$room_type[!is.na(dados$room_type)])
moda1 <- val1[which.max(tabulate(match(dados$room_type, val1)))]
dados$room_type[is.na(dados$room_type)] <- moda1
```

Figura 8-Exemplo da substituição dos NA pela moda

É de salientar que quer no Excel, quer no R, reparamos que em algumas variáveis, como por exemplo, *neighbourhood* e *room_type* existiam valores duplicados, mas considerando o contexto assumimos os mesmos como valores verdadeiros.

Já com vista ao estudo das variáveis *neighbourhood* e *room_type*, ambas categóricas, decidimos transformá-las em variáveis numéricas, tal como explicado em seguida.

Ao analisar a variável *neighbourhood*, ainda no Excel percebemos que tínhamos muitos bairros diferentes, o que tornaria impossível o correto estudo destes. Deste modo, decidimos dividi-los em quatro zonas diferentes, tendo em conta as suas coordenadas - dadas pela longitude e latitude - (Figura 9). É de notar que os números associados a cada umas dessas zonas vão ser inseridos na nova variável *neighborhood_code* criada para o efeito.

```
dados$neighborhood_code = ifelse(dados$latitude>=55.84 & dados$latitude<= 55.95 &
                                   dados$longitude>=-3.444 & dados$longitude<=-3.198, 1,
                                 ifelse(dados$latitude>=55.84 & dados$latitude<= 55.95 &
                                          dados$longitude>=-3.198 & dados$longitude<=-3.05, 2,
                                        ifelse(dados$latitude>=55.95 & dados$latitude<=56.01
                                               & dados$longitude>=-3.444 &
                                                 dados$longitude<=-3.198, 3, 4)))
```

Figura 9- Criação de uma nova variável numérica para estudar o neighbourhood

Relativamente ao *room_type* vimos que havia apenas 4 possibilidades, mas como esta variável é categórica, não a poderíamos estudar mais à frente, uma vez que a correlação e todos os métodos utilizados posteriormente necessitam que esta variável seja numérica, para este efeito criou-se uma variável a que se apelidou de *room_type_codigo* (Figura 10).

```
dados$room_type_codigo = ifelse(dados$room_type=="Entire home/apt", 1,
                                ifelse(dados$room_type== "Hotel room",
                                       2, ifelse(dados$room_type== "Private room", 3, 4)))
```

Figura 10-Criação de uma nova variável numérica para estudar room_type

Estudo das variáveis

A fim de conseguir perceber com quais variáveis trabalhar sobre e quais realmente poderiam mostrar um impacto e explicabilidade sobre a variável alvo, foi necessária uma primeira análise destas mesmas variáveis.

Estatísticas e gráficos representativos

Para fins de comparação, o dataframe inicial (Figura 11) tinha a seguinte estrutura:

```
        id                                              name   host_id host_name neighbourhood_group                neighbourhood latitude
1 31523141 Cosy. comfy. entire house. 2 bedrooms and sleeps 6    942783        Jo                  NA Gilmerton South and the Murrays 55.88566
2 16612439   Old Manse luxury studio apartment near Edinburgh  69412742     Linda                  NA Gilmerton South and the Murrays 55.91416
3 30558669     Large Double Room in Shared Flat with Sea View 115744740     Calum                  NA    Trinity East and The Dudleys 55.98065
4 53955119      Esk Lodge with River View at Kevock Vale Park 436998879   Valerie                  NA Gilmerton South and the Murrays 55.88203
5 5,99E+17              George Avenue - 1 bed shared bathroom 452973676    Robbie                  NA Gilmerton South and the Murrays 55.87941
6 42501468 Private-access room and shower room near Edinburgh 165421613    Sheila                  NA Gilmerton South and the Murrays 55.87221
  longitude       room_type price minimum_nights number_of_reviews last_review reviews_per_month calculated_host_listings_count availability_365
1  -3.08646 Entire home/apt   105             3                45  06/08/2022             1.06                             1              259
2  -3.08876 Entire home/apt    85             3                32  19/09/2020             0.51                             1               70
3  -3.19888    Private room    46             2                87  28/08/2022             2.11                             1              193
4  -3.11861    Private room   140             1                 0                           NA                             3                9
5  -3.15327    Private room    45             1                17  25/08/2022             3.42                             1              356
6  -3.15093    Private room    40             1                30  02/09/2022             1.41                             1              123
  number_of_reviews_ltm license
1                    12      NA
2                     0      NA
3                    49      NA
4                     0      NA
5                    17      NA
6                    28      NA
```

Figura 11- Base de Dados inicial

Como se pode ver na Figura 12, tínhamos variáveis numéricas e categóricas e muitas observações com valores desconhecidos – NA.

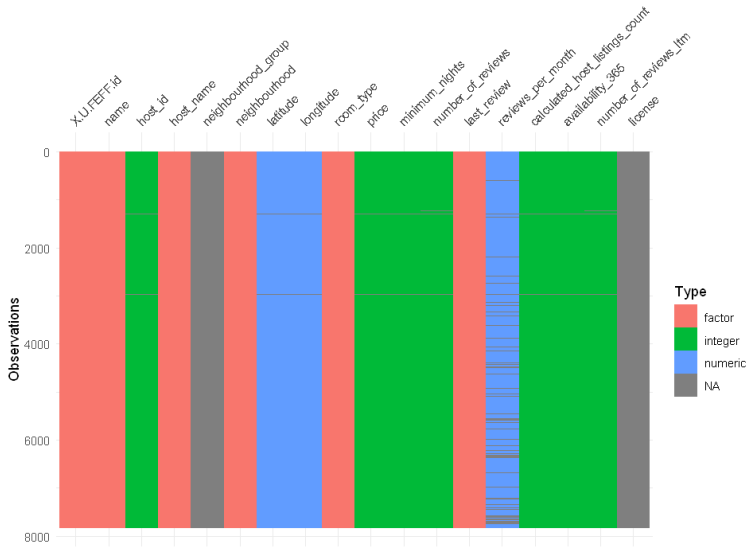


Figura 12-Tipos de dados iniciais

Após a limpeza de dados, quer em Excel, quer em R, a base de dados ficou com a seguinte configuração (Figura 13):

```
                                              name   host_id host_name                neighbourhood latitude longitude       room_type price
1 Cosy. comfy. entire house. 2 bedrooms and sleeps 6    942783        Jo Gilmerton South and the Murrays 55.88566  -3.08646 Entire home/apt   105
2   Old Manse luxury studio apartment near Edinburgh  69412742     Linda Gilmerton South and the Murrays 55.91416  -3.08876 Entire home/apt    85
3     Large Double Room in Shared Flat with Sea View 115744740     Calum    Trinity East and The Dudleys 55.98065  -3.19888    Private room    46
4      Esk Lodge with River View at Kevock Vale Park 436998879   Valerie Gilmerton South and the Murrays 55.88203  -3.11861    Private room   140
5              George Avenue - 1 bed shared bathroom 452973676    Robbie Gilmerton South and the Murrays 55.87941  -3.15327    Private room    45
6 Private-access room and shower room near Edinburgh 165421613    Sheila Gilmerton South and the Murrays 55.87221  -3.15093    Private room    40
  minimum_nights number_of_reviews last_review reviews_per_month calculated_host_listings_count availability_365 number_of_reviews_ltm
1              3                45  06/08/2022             1.06                             1              259                    12
2              3                32  19/09/2020             0.51                             1               70                     0
3              2                87  28/08/2022             2.11                             1              193                    49
4              1                 0  29/08/2022             1.62                             3                9                     0
5              1                17  25/08/2022             3.42                             1              356                    17
6              1                30  02/09/2022             1.41                             1              123                    28
  Neighborhood_code room_type_codigo neighborhood_code
1                 2                1                 2
2                 2                1                 2
3                 3                3                 3
4                 2                3                 2
5                 2                3                 2
6                 2                3                 2
```

Figura 13- Base de dados com alterações feitas

Posteriormente fomos conhecer a estrutura da nossa base de dados já limpa (Figura 14).

```
'data.frame':   7833 obs. of  17 variables:
 $ name                          : Factor w/ 7615 levels "","'A great flat'",..: 2513 5397 4166 3314 3555 5583 3562 2280 3056 5
71 ...
 $ host_id                       : int  942783 69412742 115744740 436998879 452973676 165421613 60423 46498 165635 192586 ...
 $ host_name                     : Factor w/ 2070 levels "","<U+4E00><U+5578>","<U+6606>",..: 930 1143 301 1973 1616 1766 351 7
25 1955 1541 ...
 $ neighbourhood                 : Factor w/ 111 levels "Abbeyhill","Baberton and Juniper Green",..: 45 45 108 45 45 45 83 16 5
8 65 ...
 $ latitude                      : num  55.9 55.9 56 55.9 55.9 ...
 $ longitude                     : num  -3.09 -3.09 -3.2 -3.12 -3.15 ...
 $ room_type                     : Factor w/ 5 levels "","Entire home/apt",..: 2 2 4 4 4 4 2 2 2 4 ...
 $ price                         : num  105 85 46 140 45 40 114 72 51 71 ...
 $ minimum_nights                : num  3 3 2 1 1 1 3 3 4 2 ...
 $ number_of_reviews             : int  45 32 87 0 17 30 432 230 67 45 ...
 $ last_review                   : Factor w/ 702 levels "","01/01/2016",..: 137 416 623 647 543 54 205 367 54 308 ...
 $ reviews_per_month             : num  1.06 0.51 2.11 1.62 3.42 1.41 3.04 1.58 0.67 0.31 ...
 $ calculated_host_listings_count: num  1 1 1 3 1 1 1 1 2 1 ...
 $ availability_365              : num  259 70 193 9 356 123 225 79 187 17 ...
 $ number_of_reviews_ltm         : int  12 0 49 0 17 28 77 2 10 7 ...
 $ neighbourhood_code            : num  2 2 3 2 2 2 4 2 2 2 ...
 $ room_type_codigo              : num  1 1 3 3 3 3 3 1 1 1 3 ...
```

Figura 14-Estrutura dos dados

A nossa base de dados tem 17 variáveis e cada uma tem 7833 observações, estas admitem valores numéricos inteiros ou fracionários e fatores, ou seja, caracteres. Também se pode verificar que não há valores NA em que nenhuma das variáveis, que temos 5 variáveis categóricas - *name*, *host_name*, *neighbourhood*, *last_review*, *room_type*- e as restantes 12 variáveis são numéricas.

Depois, realizamos o comando summary (Figura 15), para obtermos estatísticas acerca dos dados, o que nos permitiu perceber se os valores faziam sentido, uma vez que podemos tirar conclusões acerca da existência de valores extremos/outliers (valores bastante diferentes dos da média), ao observar o mínimo e máximo de cada variável e posteriormente decidir o que podemos fazer com esses casos.

```
                                                                                name
<U+2600><U+2654>Very Nice Room Near Montgomery Street Park<U+2654><U+2600>: 118
¤<U+272A>Trendy Studio Near Montgomery Street Park<U+272A>¤                :   8
<U+2764><U+272F>Bright Studio Near Pilrig Park<U+272F><U+2764>             :   8
Mono Suites - One Bedroom Suite                                            :   7
Ben Cruachan Guesthouse                                                    :   5
Bide Collective                                                            :   5
(Other)                                                                    :7682
   host_id           host_name
Min.   :    46498   Altido : 205
1st Qu.: 22597535   Teodora: 123
Median : 72875578   David  : 112
Mean   :134260847   John   : 100
3rd Qu.:208211166   Alison :  66
Max.   :479045480   Rebecca:  66
                    (Other):7161
                                neighbourhood     latitude
Old Town, Princes Street and Leith Street: 820   Min.   :55.84
Deans Village                            : 430   1st Qu.:55.94
Tollcross                                : 355   Median :55.95
Dalry and Fountainbridge                 : 285   Mean   :55.95
Hillside and Calton Hill                 : 282   3rd Qu.:55.96
New Town West                            : 276   Max.   :56.01
(Other)                                  :5385
  longitude              room_type        price         minimum_nights
Min.   :-3.444                :   0   Min.   :    0.0   Min.   :   0.000
1st Qu.:-3.212   Entire home/apt:5406   1st Qu.:   75.0   1st Qu.:   1.000
Median :-3.193   Hotel room     :  57   Median :  120.0   Median :   2.000
Mean   :-3.198   Private room   :2347   Mean   :  182.9   Mean   :   4.768
3rd Qu.:-3.177   Shared room    :  23   3rd Qu.:  180.0   3rd Qu.:   3.000
Max.   :-3.058                          Max.   :20551.0   Max.   :1000.000

number_of_reviews     last_review   reviews_per_month
Min.   :  0.00    29/08/2022:1322   Min.   : 0.010
1st Qu.:  4.00    11/09/2022: 406   1st Qu.: 0.730
Median : 19.00    28/08/2022: 397   Median : 1.620
Mean   : 60.71    10/09/2022: 271   Mean   : 2.171
3rd Qu.: 75.00    04/09/2022: 262   3rd Qu.: 3.000
Max.   :937.00    30/08/2022: 242   Max.   :57.140
                  (Other)   :4933
```

```
calculated_host_listings_count availability_365 number_of_reviews_ltm
Min.   :  0.000                Min.   :  0.0    Min.   :  0.00
1st Qu.:  1.000                1st Qu.:  2.0    1st Qu.:  1.00
Median :  1.000                Median : 75.0    Median :  7.00
Mean   :  9.259                Mean   :119.1    Mean   : 17.23
3rd Qu.:  4.000                3rd Qu.:225.0    3rd Qu.: 25.00
Max.   :359.000                Max.   :365.0    Max.   :343.00

neighbourhood_code room_type_codigo
Min.   :1.000      Min.   :1.000
1st Qu.:1.000      1st Qu.:1.000
Median :3.000      Median :1.000
Mean   :2.663      Mean   :1.615
3rd Qu.:4.000      3rd Qu.:3.000
Max.   :4.000      Max.   :4.000
```

Figura 15- Estatísticas das variáveis

A variável *host_id* está entre 46498 e 479045480, o anfitrião mais comum é o "Altido" (variável *host_name*), o bairro mais comum é "Old Town, Princess Street and Leith Street" (variável *neighbourhood*), a variável *latitude* está compreendida entre 55.84 e 56.01 e tem média de 55.95, a variável *longitude*, por sua vez, está entre -3.444 e -3.058 e tem média de -3.198. De notar que os valores das variáveis *latitude* e *longitude* estão pouco dispersos, o que vai de acordo com o facto de todos os Airbnb's serem na mesma cidade e em bairros próximos. Quanto ao *room_type* o tipo mais comum é o "Entire home/apt", seguido do "Private Room". A variável *price* está entre 0.0 e 20551.0 e tem média de 182.9. A variável *minimum_nights* tem 0.000 como mínimo, média de 4.768 e máximo de 1000.000, existindo outliers, mas como estes eram casos pontuais e até possíveis decidimos não os alterar. A variável *number_of_review*s varia entre 0.00 e 937.00 e tem mediana de 19.00. A *last_review* mais frequente é em 29/08/2022. As *reviews_per_month* variam entre 0.010 e 57.140, sendo que o primeiro quartil é de 0.730. A *calculated_host_listing_count* varia entre 0.000 e 359.000 e tem média de 9.259. A *availability_365* varia entre 0.0 e 365.0 e o terceiro quartil é de 225.0. O *number_of_reviews_ltm* está entre 0.00 e 343.00 e tem média de 17.23.

Procura de valores omissos

Notamos que o mínimo de algumas variáveis assumia o valor 0, coisa que no contexto do problema não fazia sentido, como na variável *price*, *minimum_nights* e *availability_365*, por isso optamos por substituir estes valores pela respetiva média da variável e obtivemos os resultados ilustrados na Figura 16.

```
     name               host_id           host_name         neighbourhood        latitude       longitude      room_type            price
 Length:7833        Min.   :    46498   Length:7833        Length:7833        Min.   :55.84   Min.   :-3.444   Length:7833        Min.   :    1
 Class :character   1st Qu.: 22597535   Class :character   Class :character   1st Qu.:55.94   1st Qu.:-3.212   Class :character   1st Qu.:   75
 Mode  :character   Median : 72875578   Mode  :character   Mode  :character   Median :55.95   Median :-3.193   Mode  :character   Median :  120
                    Mean   :134260847                                         Mean   :55.95   Mean   :-3.198                      Mean   :  183
                    3rd Qu.:208211166                                         3rd Qu.:55.96   3rd Qu.:-3.177                      3rd Qu.:  180
                    Max.   :479045480                                         Max.   :56.01   Max.   :-3.058                      Max.   :20551
 minimum_nights     number_of_reviews last_review        reviews_per_month calculated_host_listings_count availability_365 number_of_reviews_ltm
 Min.   :   1.000   Min.   :  0.00    Length:7833        Min.   : 0.010    Min.   :  0.000                 Min.   :  1.0    Min.   :  0.00
 1st Qu.:   1.000   1st Qu.:  4.00    Class :character   1st Qu.: 0.730    1st Qu.:  1.000                 1st Qu.: 69.0    1st Qu.:  1.00
 Median :   2.000   Median : 19.00    Mode  :character   Median : 1.620    Median :  1.000                 Median :119.1    Median :  7.00
 Mean   :   4.769   Mean   : 60.71                       Mean   : 2.171    Mean   :  9.259                 Mean   :146.6    Mean   : 17.23
 3rd Qu.:   3.000   3rd Qu.: 75.00                       3rd Qu.: 3.000    3rd Qu.:  4.000                 3rd Qu.:225.0    3rd Qu.: 25.00
 Max.   :1000.000   Max.   :937.00                       Max.   :57.140    Max.   :359.000                 Max.   :365.0    Max.   :343.00
 neighborhood_code room_type_codigo
 Min.   :1.000     Min.   :1.000
 1st Qu.:1.000     1st Qu.:1.000
 Median :3.000     Median :1.000
 Mean   :2.663     Mean   :1.615
 3rd Qu.:4.000     3rd Qu.:3.000
 Max.   :4.000     Max.   :4.000
```

Figura 16-Estatísticas após a substituição pela média onde havia mínimos a zero sem sentido

Procura de outliers

Após a realização do summary, percebemos que existiam outliers em algumas variáveis, mas apenas decidimos tratar dos outliers da variável *price* (Figura 17), porque é a nossa variável target e os outros valores podem eventualmente ser possíveis de ocorrer.

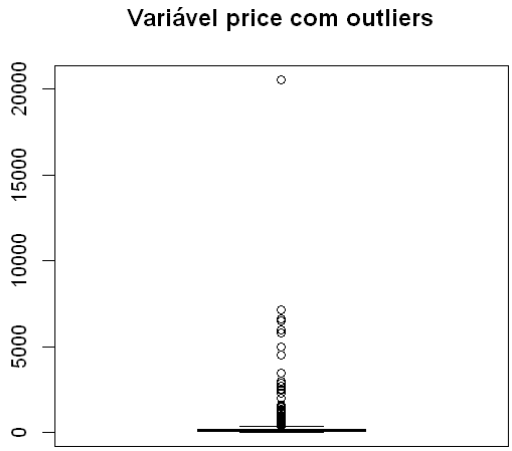


Figura 17-Boxplot da variável price

Correlação e causalidade entre as variáveis

Com vista à continuação do estudo das variáveis presentes na base de dados, criamos uma matriz de correlação de Pearson entre todos os pares de variáveis, com exceção das variáveis eliminadas anteriormente, e das variáveis *name*, *host_name*, *neighbourhood*, *room_type* e *last_review* (Figura 18). É de notar que não estudamos o *name*, porque é uma variável com informação muito dispersa sobre o Airbnb, o *host_name* estudamos através do *host_id*, que fornece um identificador único para cada anfitrião e a *last_review* é uma data de referência, que no máximo tem informação até setembro de 2022, estando por isso, um pouco desatualizada provavelmente. As variáveis *room_type* e *neighbourhood* foram estudadas com as variáveis numéricas criadas para o efeito.

```
> amostra_dados <- dados[,-c(1,3,4,7, 11)]
> head(amostra_dados)
    host_id latitude longitude price minimum_nights number_of_reviews reviews_per_month calculated_host_listings_count availability_365
1    942783 55.88566  -3.08646   105             3               45             1.06                               1              259
2  69412742 55.91416  -3.08876    85             3               32             0.51                               1               70
3 115744740 55.98065  -3.19888    46             2               87             2.11                               1              193
4 436998879 55.88203  -3.11861   140             1                0             1.62                               3                9
5 452973676 55.87941  -3.15327    45             1               17             3.42                               1              356
6 165421613 55.87221  -3.15093    40             1               30             1.41                               1              123
  number_of_reviews_ltm neighborhood_code room_type_codigo
1                    12                2                1
2                     0                2                1
3                    49                3                3
4                     0                2                3
5                    17                2                3
6                    28                2                3
```

Figura 18-Preparação dos dados para a correlação entre variáveis numéricas

De seguida, fizemos a correlação de Pearson, apenas utilizada em variáveis quantitativas e que varia entre -1 e 1 (Figura 19).

```
> cor(amostra_dados, method="pearson")
                                  host_id    latitude    longitude        price minimum_nights number_of_reviews reviews_per_month
host_id                        1.00000000 -0.04927157 -0.033166611 -0.015398148     0.014690321      -0.239699266       0.130009139
latitude                      -0.04927157  1.00000000  0.093693791  0.026777022    -0.039290518       0.032107869       0.041868915
longitude                     -0.03316661  0.09369379  1.000000000  0.002317432    -0.037499289       0.006909257      -0.002479393
price                         -0.01539815  0.02677702  0.002317432  1.000000000     0.016396736      -0.081554886      -0.070916302
minimum_nights                 0.01469032 -0.03929052 -0.037499289  0.016396736     1.000000000      -0.031734476      -0.034417127
number_of_reviews             -0.23969927  0.03210787  0.006909257 -0.081554886    -0.031734476       1.000000000       0.375586952
reviews_per_month              0.13000914  0.04186891 -0.002479393 -0.070916302    -0.034417127       0.375586952       1.000000000
calculated_host_listings_count -0.10685587  0.05697938  0.018154358  0.288701109    -0.005269805      -0.094174434      -0.054731301
availability_365               0.12994721 -0.03344548 -0.018611584  0.062549261     0.065365694      -0.051681841      -0.024517567
number_of_reviews_ltm         -0.08946011  0.03925982 -0.002194308 -0.098022021    -0.037337352       0.703458029       0.537646984
neighborhood_code             -0.07073704  0.69327021  0.383393998  0.060759455    -0.055047198       0.024943893       0.016713357
room_type_codigo              -0.00976794 -0.07326001 -0.005661559 -0.042181287     0.030037247      -0.098347788      -0.004427273
                              calculated_host_listings_count availability_365 number_of_reviews_ltm neighborhood_code room_type_codigo
host_id                                       -0.106855868      0.129947206          -0.089460110       -0.07073704     -0.009767940
latitude                                       0.056979375     -0.033445478           0.039259822        0.69327021     -0.073260009
longitude                                      0.018154358     -0.018611584          -0.002194308        0.38339400     -0.005661559
price                                          0.288701109      0.062549261          -0.098022021        0.06075946     -0.042181287
minimum_nights                                -0.005269805      0.065365694          -0.037337352       -0.05504720      0.030037247
number_of_reviews                             -0.094174434     -0.051681841           0.703458029        0.02494389     -0.098347788
reviews_per_month                             -0.054731301     -0.024517567           0.537646984        0.01671336     -0.004427273
calculated_host_listings_count                 1.000000000      0.101414316          -0.076972843        0.07757799     -0.008848813
availability_365                               0.101414316      1.000000000          -0.028664731       -0.01229440     -0.007746149
number_of_reviews_ltm                         -0.076972843     -0.028664731           1.000000000        0.02950942     -0.115094315
neighborhood_code                              0.077577986     -0.012294398           0.029509417        1.00000000     -0.015438643
room_type_codigo                              -0.008848813     -0.007746149          -0.115094315       -0.01543864      1.000000000
```

Figura 19-Correlação de Pearson

Deste modo, a melhor correlação positiva - quando uma aumenta a outra também aumenta - é entre as variáveis *number_of_reviews_ltm* e *number_of_reviews* com 0.703458029, sendo por isso, uma correlação forte e a pior correlação positiva é de 0.002320913 (correlação fraca) entre o *price* e a *longitude*. A pior correlação negativa - quando uma aumenta, a outra diminui - é de -0.002194308 (correlação fraca) entre a o *number_of_reviews_ltm* e a *longitude* e a melhor correlação negativa é -0.097924374 (correlação fraca).

A variável *calculated_host_listings_count* tem a maior correlação com o preço dos Airbnb's (0.288701109), o que faz sentido considerando que temos anfitriões profissionais, cujo rendimento pode vir da exploração dos mesmos. Apenas as variáveis *number_of_reviews*, *reviews_per_month*, *number_of_reviews_ltm* e room_typecodigo têm correlação negativa com a variável *price*. A variável menos correlada com o preço é a variável *longitude* (0.002317432).

De seguida, criamos uma matriz de correlação com as mesmas 12 variáveis (Figura 20).

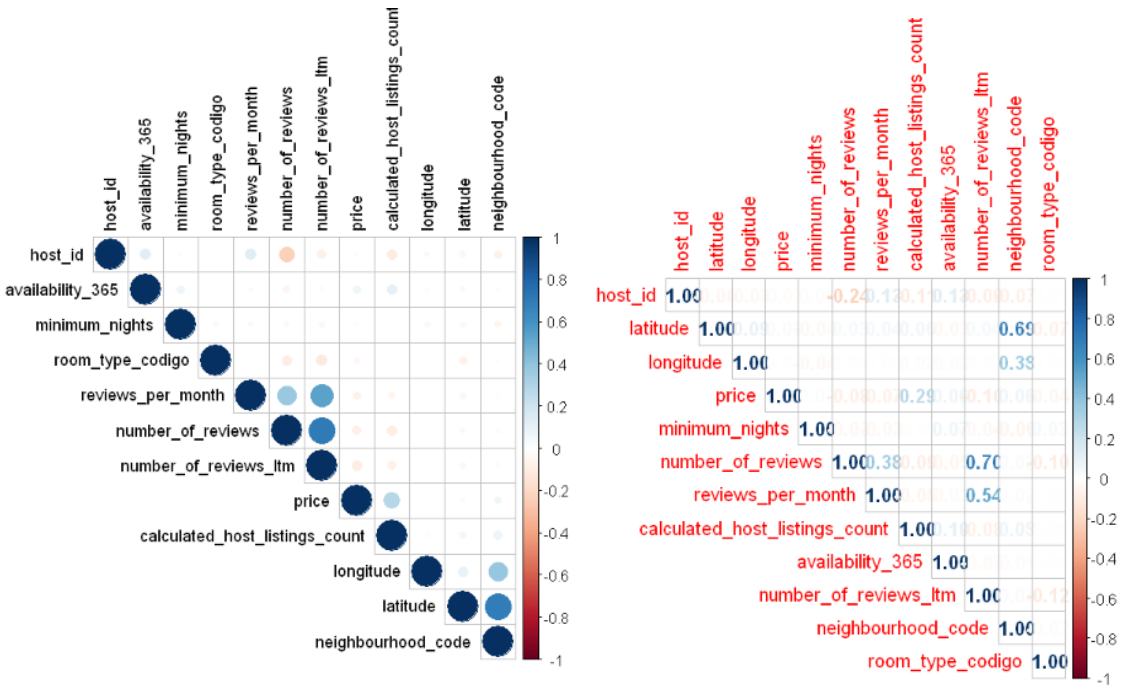


Figura 20-Matriz de correlação entre as variáveis numéricas

De seguida, realizamos a representação gráfica, segundo a qual obtivemos vários gráficos de dispersão entre os pares de variáveis usados anteriormente (Figura 21), utilizando o comando `pairs(amostra_dados)` .

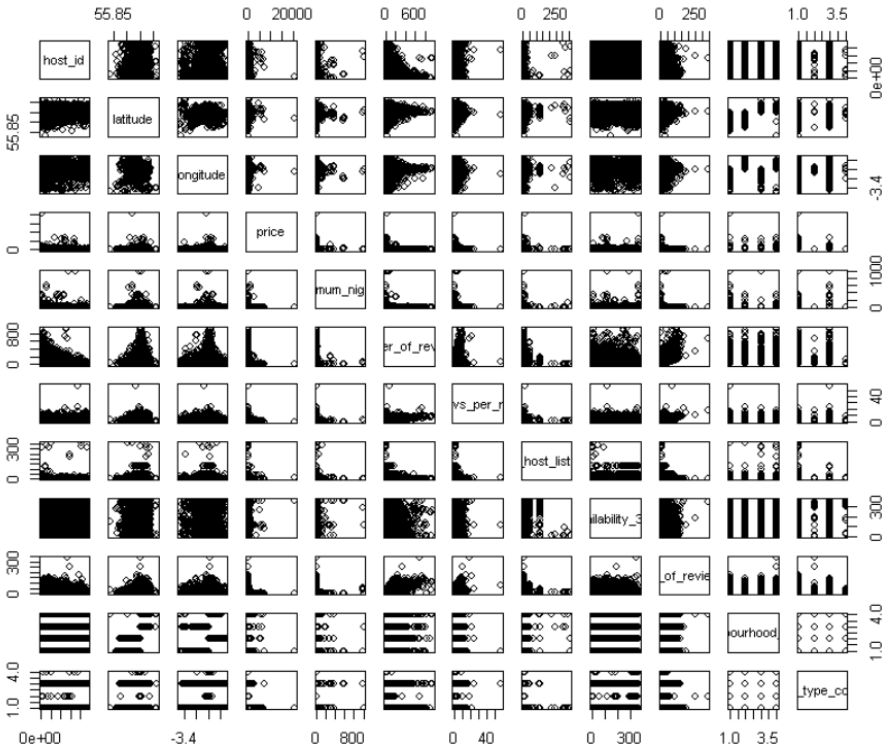


Figura 21-Gráficos de dispersão de pares de variáveis

Concluímos que a variável *price* é assimétrica, que existem padrões lineares e não lineares entre as variáveis e que existem diversos pontos extremos/outliers.

Modelo de regressão linear múltipla

Para encontrar o melhor modelo possível começamos, por usar a biblioteca olsrr, que o faz de forma automática. A Figura 22 mostra o melhor modelo de regressão utilizando o p-value, através do uso da função *ols_step_both_p()*.

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	calculated_host_listings_count	addition	0.083	0.083	101.0760	114667.2486	365.1483
2	number_of_reviews_ltm	addition	0.089	0.089	53.0740	114619.7019	364.0186
3	room_type_codigo	addition	0.092	0.091	34.4830	114601.2073	363.5659
4	neighbourhood_code	addition	0.093	0.093	22.1540	114588.9139	363.2575
5	availability_365	addition	0.094	0.094	15.1830	114581.9514	363.0729
6	latitude	addition	0.095	0.094	12.3870	114579.1565	362.9850
7	longitude	addition	0.095	0.095	8.0700	114574.8354	362.8618

Figura 22-Escolha do modelo de regressão com base no p-value

Por sua vez, a Figura 23 mostra o melhor modelo de regressão utilizando o AIC, através do uso da função *ols_step_both_aic()*.

Stepwise Summary

Variable	Method	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
calculated_host_listings_count	addition	114667.249	1044133193.566	94939834.775	0.08335	0.08323
number_of_reviews_ltm	addition	114619.702	1037549504.347	101523523.994	0.08913	0.08890
room_type_codigo	addition	114601.207	1034838354.505	104234673.836	0.09151	0.09116
neighbourhood_code	addition	114588.914	1032951745.636	106121282.706	0.09316	0.09270
availability_365	addition	114581.951	1031770519.749	107302508.592	0.09420	0.09362
latitude	addition	114579.156	1031139117.470	107933910.871	0.09476	0.09406
longitude	addition	114574.835	1030307343.071	108765685.270	0.09549	0.09468
minimum_nights	addition	114574.594	1030012507.690	109060520.651	0.09574	0.09482

Figura 23-Escolha do modelo de regressão com base no AIC

Tal como podemos observar, os dois modelos não são concordantes, mas decidimos usar o modelo mais completo, neste caso o dado pelo AIC. Para chegar ao melhor modelo e poder compará-lo com os outros modelos realizados, retirámos os outliers da variável *price*, o que deu origem a uma amostra diferente. Amostra esta, que foi utilizada nos restantes modelos, onde transformamos as variáveis, por exemplo, fizemos o logaritmo da variável dependente - *price* -, acrescentamos não linearidade, nas variáveis *latitude*, *longitude*, *minimum_nights*, *calculated_host_listings_count* e *room_type_codigo*, através de polinómios cúbicos. Também criamos funções de não linearidade através da função hiperbólica do tipo 1/x nas variáveis *room_type_codigo*, *minimum_nights*, *latitude* e *longitude*. Além disso, fizemos modelos com interação entre as variáveis *longitude* e

latitude e não linearidade entre outras variáveis. Por fim, criamos diversos modelos, utilizando o método de estimação dos mínimos quadrados com pesos.

Escolha do modelo final

A escolha do modelo final foi realizada, através da Tabela 1, que contém os 12 modelos realizados, com o respetivo erro, AIC e R^2. Em teoria, o melhor modelo é o que tem o menor erro, menor AIC e R^2 elevado. Neste caso, decidimos escolher o modelo 4, pois é o que tem menor erro, sem contar com o modelo 12 - que é uma subamostra - e tem um AIC e R^2 aceitáveis.

numero_modelo	erros	AICs	R_quadrado
1	2.5123347	114574.594	0.09574498
2	0.7059467	79484.443	0.27962478
3	0.5125674	10566.466	0.33300762
4	0.4416521	9443.316	0.42980169
5	0.4897531	10277.779	0.35965073
6	0.4889459	10270.818	0.36043862
7	0.4382143	10503.963	0.43514933
8	0.5712543	9620.451	0.44572571
9	0.5721660	9637.238	NA
10	0.4936788	4934.398	0.89836600
11	0.7248554	7458.910	0.53203729
12	0.3430531	1096.974	0.47167686

Tabela 1-Tabela com erros, AICs e R^2 de todos os modelos criados

Deste modo, o modelo escolhido foi o seguinte (Figura 24):

```
fit4 <- lm(log(price) ~ poly(calculated_host_listings_count, degree=3)
           + number_of_reviews_ltm + poly(room_type_codigo, degree=3)
           + neighbourhood_code + availability_365 + poly(latitude, degree=2)
           +poly(longitude, degree=3) + poly(minimum_nights, degree=3)
           , data=amostra_dados2)
```

Figura 24- Modelo escolhido

Este modelo é constituído pela variável *price* logaritmizada, pelas variáveis *calculated_host_listings_count*, *room_type_codigo*, *minimum_nights*, *longitude* e *latitude*, sob a forma de polinómio, de grau dois ou três e as variáveis *number_of_review_ltm*, *neighbourhood_code* e availability_365 sem qualquer tipo de mudança. Este modelo foi construído com o intuito de corrigir os pressupostos dos resíduos que não se verificavam, neste caso, não se verifica variância constante, ausência de correlação, nem normalidade. Apesar de este modelo não ter conseguido melhorar o modelo nesse sentido, conseguimos obter menor erro, menor AIC e maior R^2 comparativamente ao modelo anterior. Deste modo, foi o modelo matemático explicitado abaixo que apresentou os melhores resultados (Figura 25).

$$log(price) = \beta_0 + \beta_1(calculates\ host\ listings\ count) + \beta_2(calculated\ host\ listings\ count^2)$$
$$+ \beta_3(calculated\ host\ listings\ count^3) + \beta_4(number\ of\ reviews\ ltm) + \beta_5(room\ type\ codigo)$$
$$+ \beta_6(room\ type\ codigo^2) + \beta_7(room\ type\ codigo^3) + \beta_8(neighbourhood\ code)$$
$$+ \beta_9(availability\ 365) + \beta_{10}(latitude) + \beta_{11}(latitude^2) + \beta_{12}(longitude) + \beta_{13}(longitude^2)$$
$$+ \beta_{14}(longitude^3) + \beta_{15}(minimum\ nights) + \beta_{16}(minimum\ nights^2) + \beta_{17}(minimum\ nights^3) + \varepsilon$$

Figura 25-Representação matemática do modelo de regressão

Interpretação do modelo final

```
Call:
lm(formula = log(price) ~ poly(calculated_host_listings_count,
    degree = 3) + number_of_reviews_ltm + poly(room_type_codigo,
    degree = 3) + neighbourhood_code + availability_365 + poly(latitude,
    degree = 2) + poly(longitude, degree = 3) + poly(minimum_nights,
    degree = 3), data = amostra_dados2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0006 -0.2904  0.0029  0.2916  1.5986

Coefficients:
                                              Estimate Std. Error t value
(Intercept)                                  4.542e+00  2.260e-02 200.926
poly(calculated_host_listings_count, degree = 3)1 -4.366e+00  4.724e-01  -9.243
poly(calculated_host_listings_count, degree = 3)2 -1.260e+01  4.714e-01 -26.734
poly(calculated_host_listings_count, degree = 3)3  8.649e-01  4.660e-01   1.856
number_of_reviews_ltm                        -2.296e-03  2.324e-04  -9.877
poly(room_type_codigo, degree = 3)1          -2.752e+01  4.749e-01 -57.950
poly(room_type_codigo, degree = 3)2          -1.300e+00  4.629e-01  -2.809
poly(room_type_codigo, degree = 3)3          -3.319e+00  4.656e-01  -7.127
neighbourhood_code                            1.421e-02  7.782e-03   1.826
availability_365                              8.647e-04  5.238e-05  16.508
poly(latitude, degree = 2)1                  -6.837e-01  7.032e-01  -0.972
poly(latitude, degree = 2)2                  -3.593e+00  5.366e-01  -6.695
poly(longitude, degree = 3)1                 -6.470e-01  5.572e-01  -1.161
poly(longitude, degree = 3)2                 -3.940e+00  5.017e-01  -7.854
poly(longitude, degree = 3)3                 -4.992e+00  4.982e-01 -10.021
poly(minimum_nights, degree = 3)1            -3.304e+00  4.645e-01  -7.113
poly(minimum_nights, degree = 3)2             4.365e+00  4.651e-01   9.385
poly(minimum_nights, degree = 3)3            -1.730e+00  4.661e-01  -3.711
                                             Pr(>|t|)
(Intercept)                                   < 2e-16 ***
poly(calculated_host_listings_count, degree = 3)1  < 2e-16 ***
poly(calculated_host_listings_count, degree = 3)2  < 2e-16 ***
poly(calculated_host_listings_count, degree = 3)3 0.063469 .
number_of_reviews_ltm                         < 2e-16 ***
poly(room_type_codigo, degree = 3)1           < 2e-16 ***
poly(room_type_codigo, degree = 3)2          0.004986 **
poly(room_type_codigo, degree = 3)3          1.12e-12 ***
neighbourhood_code                           0.067901 .
availability_365                              < 2e-16 ***
poly(latitude, degree = 2)1                  0.330956
poly(latitude, degree = 2)2                  2.31e-11 ***
poly(longitude, degree = 3)1                 0.245648
poly(longitude, degree = 3)2                 4.62e-15 ***
poly(longitude, degree = 3)3                  < 2e-16 ***
poly(minimum_nights, degree = 3)1            1.24e-12 ***
poly(minimum_nights, degree = 3)2             < 2e-16 ***
poly(minimum_nights, degree = 3)3            0.000208 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4623 on 7260 degrees of freedom
Multiple R-squared:  0.4298,    Adjusted R-squared:  0.4285
F-statistic: 321.9 on 17 and 7260 DF,  p-value: < 2.2e-16
```

Figura 26-Interpretação do modelo de regressão com os seus coficientes, p-values, valor de estatística de teste e R^2 do modelo

Segundo a Figura 26, o erro residual é de 0.4623, o que é um valor moderado. O R^2 é 0.4298, ou seja, moderado e significa que 42.98% da variação dos preços dos Airbnb conseguem ser explicados pela variação conjunta das variáveis independentes presentes no modelo, as restantes, são de natureza não identificada, ou seja, 57.02% da variação dos preços não é identificada por estas variáveis. Podemos ainda referir o valor da estatística F, como o p-value desta é menor que 0.05, o modelo é adequado globalmente.

Ao interpretar o *summary* do modelo fit4, percebemos que todas as variáveis são estatisticamente significativas, menos a variável latitude e longitude, ambas de grau um, pelo que todas com a exceção destas têm que ser interpretadas.

Podemos concluir, que o valor esperado do preço dos Airbnb quando todas as variáveis estão a zero é de cerca de 93.88 libras (considerando a moeda utilizada em Edinburg). Este valor foi obtido, ao fazer a exponencial do valor do coeficiente do *Intercept.* Relativamente às variáveis com polinómios quadrados - *latitude* - e cúbicos – *calculated_host_listings_count*, *room_type_codigo*, *longitude*, *minimum_nights*- a sua interpretação não é clara, percebemos que depende da variável de grau 1 correspondente e que não apresenta linearidade.

Relativamente à variável number_of_reviews_ltm, para cada aumento de uma unidade na avaliação ao longo do ano, o preço desce em 0.2295865%.

O preço aumenta em 1.420919%, com o incremento de uma unidade em neighbourhood_code e aumenta em 0.08647065% com o incremento de uma unidade em availability_365.

Verificação dos pressupostos

O **primeiro pressuposto**, ou seja, que a média dos resíduos (Figura 27) é nula, é verificado, uma vez que está muito perto de zero.

```
mean(fit4$residuals)  # média nula
```

```
1.84624710649251e-17
```

Figura 27-Verificação do primeiro pressuposto dos resíduos (média dos resíduos nula)

O **segundo pressuposto** (Figura 28) -variância constante, dado pelo teste Breusch-Pagan - não é verificado. Considerando que a hipótese nula é que os erros são homocedásticos, temos de a rejeitar pois p-value é menor que 0.05, pelo que os erros são heterocedásticos.

```
bptest(fit4)   # variância constante
```

```
        studentized Breusch-Pagan test

data:  fit4
BP = 168.77, df = 17, p-value < 2.2e-16
```

Figura 28-Verificação do segundo pressuposto (homocedasticidade)

O **terceiro pressuposto** (Figura 29) - ausência de correlação, dado pelo teste Breusch-Godfrey- não se verifica. É de notar que a hipótese nula é que os resíduos são

independentes. Esta é rejeitada, pois o p-value é menor que 0.05, pelo que os resíduos são dependentes.

```
bgtest(fit4)   # ausência de correlação

        Breusch-Godfrey test for serial correlation of order up to 1

data:  fit4
LM test = 67.613, df = 1, p-value < 2.2e-16
```

Figura 29-Verificação do terceiro pressuposto (ausência de correlação).

O **quarto** e último **pressuposto** (Figura 30) - resíduos normalmente distribuídos, dado pelo teste Jarque Bera - não se verifica. É de notar que a hipótese nula é que os resíduos têm distribuição normal. Esta é rejeitada, pois o p-value é menor que 0.05, pelo que os resíduos não são normalmente distribuídos.

```
jarque.bera.test(fit4$residuals)   # distribuição normal

        Jarque Bera Test

data:  fit4$residuals
X-squared = 12797, df = 2, p-value < 2.2e-16
```

Figura 30-Verificação do quarto pressuposto (resíduos normalmente distribuídos)

Procedemos à representação gráfica (Figura 31) sobre os resíduos do modelo:

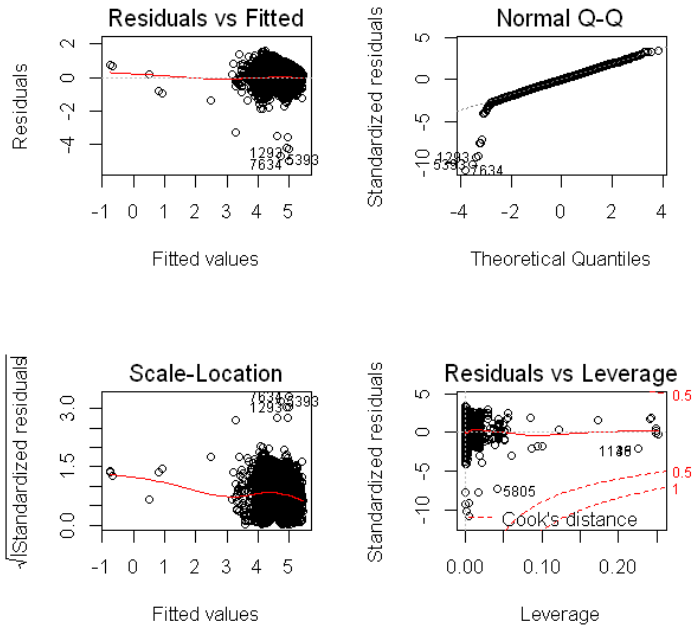


Figura 31-Representação gráfica sobre os resíduos do modelo fit4

O gráfico Residuals vs Fitted apresenta alguma não linearidade, no Normal Q-Q a cauda esquerda está distante, no Scale-Location não há funil e no Residuals vs Leverage existem alguns outleirs.

Previsão

Em seguida, procedemos à previsão *in-sample* e *out-sample*, é de notar que quanto menor o valor do MAPE melhor a precisão do modelo.

Previsão *in-sample* do modelo final

Realizamos a previsão *in-sample* de todos os modelos criados uma vez que todos eram modelos adequados para previsão pela estatística F. No entanto, optamos pelo modelo fit4 (Figura 32), que era o que tinha menor erro de previsão (MAPE), cujo erro era de cerca de 44.17%.

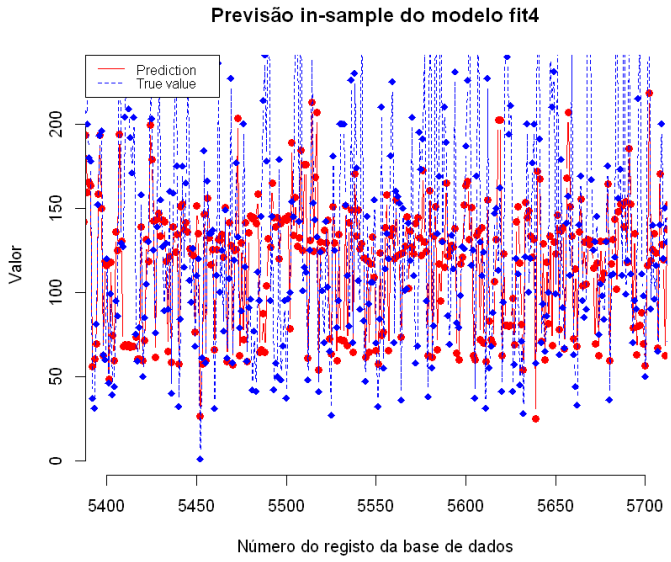


Figura 32-Previsão in-sample do modelo fit4

Previsão *out-sample* do modelo final

Decidimos fazer a previsão out-sample para saber se o modelo fit4 tinha poder preditivo fora da amostra. Para isso dividimos a amostra em dois conjuntos de dados distintos – um de treino, que corresponde a 90% da amostra e um de teste, que corresponde aos restantes 10% da amostra.

Começamos por treinar o modelo de regressão no conjunto de dados de treino, para mais tarde visualizar a possibilidade de previsão no conjunto de teste.

A percentagem correspondente ao erro de previsão out-sample (Figura 33), dado pelo MAPE é de 38.43%, sendo que a *seed* escolhida é de 127, que é menor que a percentagem do erro de previsão *in-sample* de 44.17%. Deste modo, o modelo consegue prever bem fora da amostra, considerando que apresenta um erro de previsão (MAPE) menor.

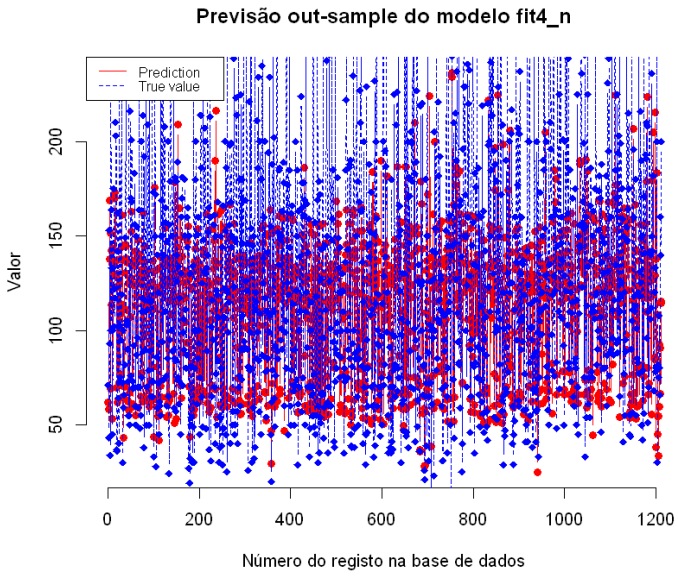


Figura 33-Previsão out-sample do modelo fit4

Subamostra

Considerando que não conseguimos verificar os quatro pressupostos dos resíduos em nenhum dos modelos realizados, sendo que apenas no primeiro conseguimos averiguar dois pressupostos, decidimos realizar uma subamostra, neste caso, restringindo o número de linhas e usando o modelo fit4, que já tínhamos decidido ser o melhor.

A nossa subamostra é composta pelas primeiras 1000 linhas da amostra_dados2, ou seja, já sem os outliers da variável price e foi construída da seguinte forma:

```
amostra_dados3<-amostra_dados2[1:1000,]
```

O modelo fit12 (Figura 34) é o seguinte:

```
fit12 <- lm(log(price) ~ poly(calculated_host_listings_count, degree=3)
          + number_of_reviews_ltm  + poly(room_type_codigo, degree=2) + neighbourhood_code + availability_365
           + poly(latitude, degree=2)  + poly(longitude, degree=3) + poly(minimum_nights, degree=3)
          , data=amostra_dados3)
```

Figura 34-Modelo escolhido

Interpretação do modelo da subamostra

```
Call:
lm(formula = log(price) ~ poly(calculated_host_listings_count,
    degree = 3) + number_of_reviews_ltm + poly(room_type_codigo,
    degree = 2) + neighbourhood_code + availability_365 + poly(latitude,
    degree = 2) + poly(longitude, degree = 3) + poly(minimum_nights,
    degree = 3), data = amostra_dados3)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18867 -0.27417 -0.02341  0.27550  1.39511

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                    4.554e+00  5.515e-02  82.576
poly(calculated_host_listings_count, degree = 3)1  1.868e+00  4.252e-01   4.393
poly(calculated_host_listings_count, degree = 3)2 -1.005e+00  4.247e-01  -2.366
poly(calculated_host_listings_count, degree = 3)3 -6.204e-01  4.252e-01  -1.459
number_of_reviews_ltm                          -2.190e-03  5.405e-04  -4.053
poly(room_type_codigo, degree = 2)1            -1.124e+01  4.403e-01 -25.537
poly(room_type_codigo, degree = 2)2            -4.838e-01  4.172e-01  -1.160
neighbourhood_code                              1.788e-02  1.817e-02   0.984
availability_365                                7.510e-04  1.342e-04   5.597
poly(latitude, degree = 2)1                    -4.361e-01  6.170e-01  -0.707
poly(latitude, degree = 2)2                    -1.012e+00  4.682e-01  -2.161
poly(longitude, degree = 3)1                   -2.858e-01  4.876e-01  -0.586
poly(longitude, degree = 3)2                   -7.059e-01  4.496e-01  -1.570
poly(longitude, degree = 3)3                   -1.487e+00  4.528e-01  -3.284
poly(minimum_nights, degree = 3)1              -1.540e+00  4.193e-01  -3.672
poly(minimum_nights, degree = 3)2               1.431e+00  4.199e-01   3.409
poly(minimum_nights, degree = 3)3              -8.432e-01  4.269e-01  -1.975
                                               Pr(>|t|)
(Intercept)                                     < 2e-16 ***
poly(calculated_host_listings_count, degree = 3)1 1.24e-05 ***
poly(calculated_host_listings_count, degree = 3)2 0.018159 *
poly(calculated_host_listings_count, degree = 3)3 0.144812
number_of_reviews_ltm                          5.46e-05 ***
poly(room_type_codigo, degree = 2)1             < 2e-16 ***
poly(room_type_codigo, degree = 2)2            0.246425
neighbourhood_code                             0.325390
availability_365                               2.83e-08 ***
poly(latitude, degree = 2)1                    0.479881
poly(latitude, degree = 2)2                    0.030930 *
poly(longitude, degree = 3)1                   0.557894
poly(longitude, degree = 3)2                   0.116733
poly(longitude, degree = 3)3                   0.001061 **
poly(minimum_nights, degree = 3)1              0.000253 ***
poly(minimum_nights, degree = 3)2              0.000680 ***
poly(minimum_nights, degree = 3)3              0.048553 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4148 on 983 degrees of freedom
Multiple R-squared:  0.4717,    Adjusted R-squared:  0.4631
F-statistic: 54.85 on 16 and 983 DF,  p-value: < 2.2e-16
```

Figura 35-Interpretação do modelo de regressão com os seus coficientes, p-values, valor de estatística de teste e R^2 do modelo

Segundo a Figura 35, o erro residual é de 0.4148, o que é um valor moderado. O R^2 é 0.4717, ou seja, é moderado e significa que 47.17% da variação dos preços dos Airbnb conseguem ser explicados pela variação conjunta das variáveis independentes presentes no modelo, as restantes, são de natureza não identificada, ou seja, 52.83% da variação dos preços não é identificada por estas variáveis. Podemos ainda referir o valor da estatística F, como o p-value desta é menor que 0.05, o modelo é adequado globalmente.

Ao interpretar o *summary* do modelo fit12, percebemos que a variáveis target (*price*), *calculated_host_listings_count* de grau 1 e 2, *number_of_reviews_ltm*, *room_type_codigo* de grau 1, *availability_365*, *latitude* de grau 2, *longitude* de grau 3 e *minimum_nights* de grau1, 2 e 3 são significativas.

Podemos concluir, que o valor esperado do preço dos Airbnb quando todas as variáveis estão a zero é de cerca de 95 libras. Relativamente às variáveis com polinómios quadrados – *latitude*, *room_type_codigo* - e cúbicos – *longitude*, *minimum_nights*- a sua

interpretação não é clara, percebemos que depende da variável de grau 1 correspondente e que não apresenta linearidade. Relativamente à variável number_of_reviews_ltm, para cada aumento de uma avaliação ao longo do ano, o preço desce em 0.2190350%. O preço aumenta em 0.07510429% com o incremento de uma unidade em availability_365.

Verificação dos pressupostos da subamostra

O **primeiro pressuposto**, ou seja, que a média dos resíduos (Figura 36) é nula, é verificado, uma vez que está muito perto de zero.

```
mean(fit12$residuals)  # média nula
```

```
-1.07238436880541e-18
```

Figura 36-Verificação do primeiro pressuposto (média dos resíduos nula)

O **segundo pressuposto** (Figura 37) -variância constante, dado pelo teste Breusch-Pagan - não é verificado. Considerando que a hipótese nula é que os erros são homocedásticos, temos de a rejeitar pois p-value é menor que 0.05, pelo que os erros são heterocedásticos.

```
bptest(fit12)   # variância constante
```

```
        studentized Breusch-Pagan test

data:  fit12
BP = 40.199, df = 16, p-value = 0.0007282
```

Figura 37-Verificação do segundo pressuposto (homocedasticidade)

O **terceiro pressuposto** (Figura 38) -ausência de correlação, dado pelo teste Breusch-Godfrey- verifica-se. É de notar que a hipótese nula é que os resíduos são independentes. Esta não é rejeitada, pois o p-value é maior que 0.05, pelo que os resíduos são independentes.

```
bgtest(fit12)   # ausência de correlação
```

```
        Breusch-Godfrey test for serial correlation of order up to 1

data:  fit12
LM test = 0.20876, df = 1, p-value = 0.6477
```

Figura 38-Verificação do terceiro pressuposto (ausência de correlação)

O **quarto** e último **pressuposto** (Figura 39) -resíduos normalmente distribuídos, dado pelo teste Jarque Bera – verifica-se. É de notar que a hipótese nula é que os resíduos têm distribuição normal. Esta não é rejeitada, pois o p-value é maior que 0.05, pelo que os resíduos são normalmente distribuídos.

```
jarque.bera.test(fit12$residuals)   # distribuição normal

        Jarque Bera Test

data:  fit12$residuals
X-squared = 1.5268, df = 2, p-value = 0.4661
```

Figura 39-Verificação do quarto pressuposto (resíduos normalmente distribuídos)

A representação gráfica da subamostra (Figura 40) é:

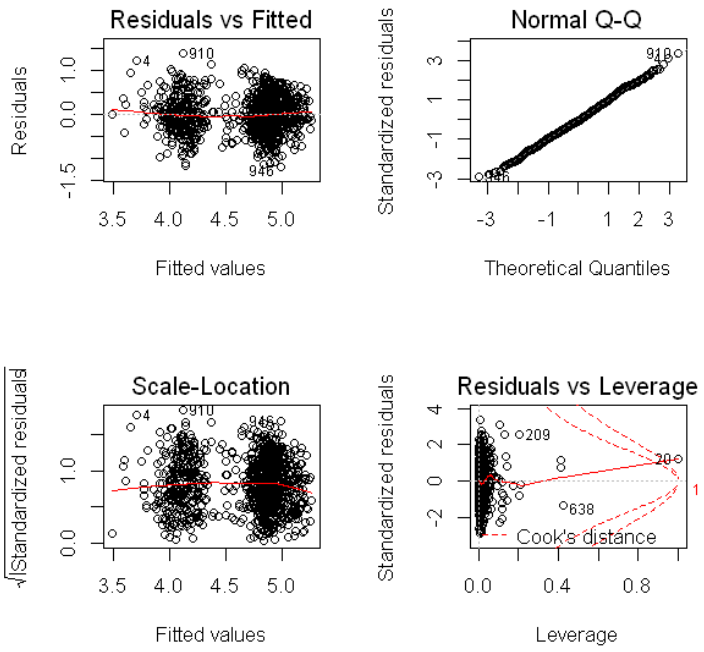


Figura 40-Representação gráfica sobre os resíduos do modelo fit12

O gráfico Residuals vs Fitted apresenta pouca não linearidade, no Normal Q-Q, os pontos estão bem alinhados segundo a reta, o Scale-Location não tem funil e no Residuals vs Leverage nota-se a presença de poucos outliers.

Previsão *in-sample* do modelo da subamostra

A previsão in-sample do modelo fit12 (Figura 41), deu um erro de previsão (MAPE) de 34.30%.

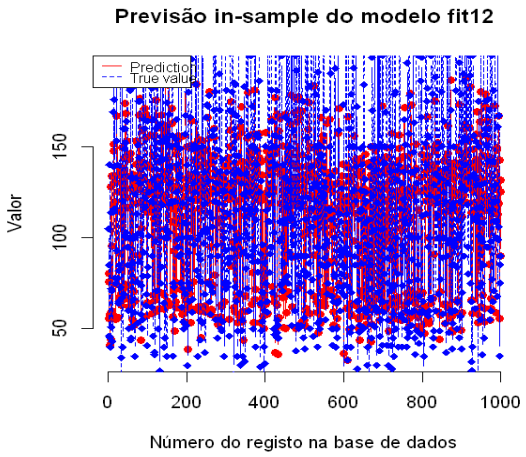


Figura 41-Previsão in-sample do modelo fit12

Previsão *out-sample* do modelo da subamostra

Fizemos previsão out-sample do modelo fit12 (Figura 42) para saber se este tinha poder preditivo, seguindo os mesmos objetivos, passos e metodologias que usamos para o modelo fit4.

A percentagem correspondente ao erro de previsão out-sample, dado pelo MAPE, é de 31.67%, sendo que a *seed* escolhida é de 2, que é menor que a percentagem do erro de previsão *in-sample* de 34.30%. Deste modo, o modelo consegue prever bem fora da amostra, considerando que apresenta um erro de previsão (MAPE) menor.

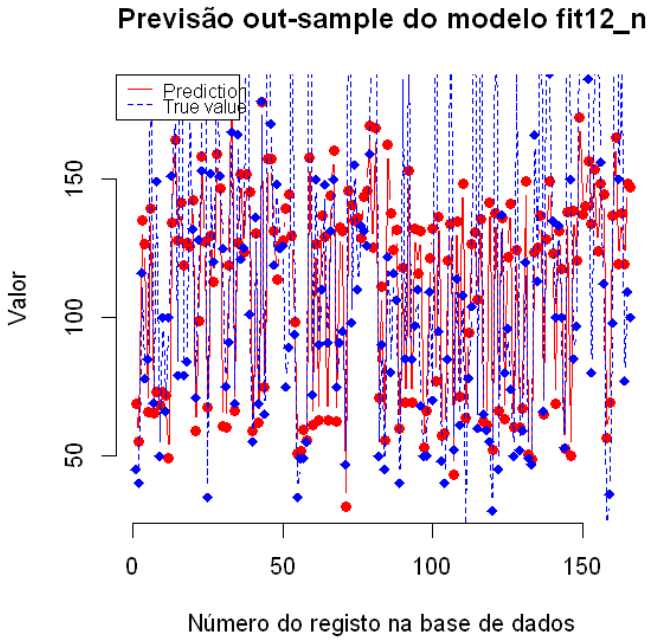


Figura 42-Previsão out-sample do modelo fit12

Conclusão

Considerando as variáveis que estudamos, existe correlação positiva entre o price e latitude, longitude, minimum_nights, calculated_host_listings_count, availability_365 e neighbourhood_code, ou seja, quando um aumenta, o outro aumenta. Por sua vez, o price e host_id, number_of_reviews, reviews_per_month, number_of_reviews_ltm e room_type_codigo, têm correlação negativa, pelo que quando um aumenta, o outro diminui. Através da análise da correlação, a variável calculated_host_listings_count é a que têm a maior correlação com o price. Assim, as variáveis que mais influenciam são calculated_host_listings_count (correlação positiva), number_of_reviews_ltm (correlação negativa) e number_of_reviews (correlação negativa).

Relativamente, ao modelo fit4, considerando os valores de MAPE, este tem uma boa capacidade preditiva in e out-sample, sendo que este melhora para valores fora da amostra. Além disso, o baixo AIC e elevado R^2 tornam este um bom modelo para o projeto.

O modelo fit12, melhora o modelo fit4, na medida é que é possível verificar três dos quatro pressupostos dos resíduos, o AIC, o MAPE e o R^2 melhoram comparativamente aos valores do modelo fit4. E este também é um bom modelo de previsão in e out-sample, melhorando o valor de MAPE, na última.

Consideramos que, no geral, os modelos escolhidos se adequam ao problema proposto, contudo, poderíamos ter feito mais subamostras com o intuito de explorar questões mais específicas a partir do conjunto de dados disponibilizado.

Referências Bibliográficas

insideairbnb. (13 de 9 de 2022). *Edinburgh*. Obtido de Insideairbnb: http://insideairbnb.com/edinburgh

Lauritzen, K. (9 de 9 de 2019). *How to maximize profits on Airbnb? Data-based approach for hosts*. Obtido de towardsdatascience: https://towardsdatascience.com/how-to-maximize-profits-on-airbnb-data-based-approach-for-hosts-beaf08f26941