Unidade Curricular de Introdução de Modelos Dinâmicos

Trabalho Final

1º semestre

Allan Kardec da Silva Rodrigues Nº103380

Diogo Alexandre Alonso de Freitas Nº104841

João Francisco Marques Gonçalves da Silva Botas Nº104782

Pedro Brígido Machado Nº98601

Turma CDB1

Índice

Introdução

Inside Airbnb é um projeto que fornece dados sobre o impacto do Airbnb em comunidades residenciais. A base de dados usada neste projeto é das diferentes localizações dos Airbnbs localizadas em Sevilla, onde nos são apresentadas várias variáveis que serão alvos de estudo e análise, a fim de entender quais as que mais influenciam o preço de cada Airbnb. Antes de começarmos a tirar conclusões, é necessário referir/explicar cada uma das variáveis e desenvolver um Data understanding para prosseguir a uma fase seguinte.

Data understanding das variáveis

- **Id** - O identificador único da Airbnb para a lista.
- **Name** - Nome da respetiva Airbnb.
- **Host_id** - O identificador único da Airbnb para o anfitrião/utilizador.
- **Host_name** - Nome do anfitrião. Normalmente apenas o(s) nome(s) próprio(s).
- **Neighbourhood_group** – O grupo de bairros geocodificado utilizando a latitude e longitude contra os bairros.
- **Neighbourhood** – O bairro como geocodificado utilizando a latitude e longitude.
- **Latitude e Longitude** - Utiliza a projecção do Sistema Geodésico Mundial (WGS84) para latitude e longitude.
- **Room_type** – Tipos de quartos existentes em cada hotel. Existem, respetivamente 4 tipos:
    - **Entire home/apt** – Um espaço inteiro para o consumidor que o alugar;
    - **Hotel room** – Quarto de um hotel;
    - **Private rooms** – Espaço partilhada com outros consumidores, mas possui um quarto privado para cada uma das pessoas;
    - **Shared rooms** - Espaço partilhada com outros consumidores, não possuindo quartos privados.
- **Price** - Preço diário de cada Airbnb em moeda local (euros).
- **Minimum_nights** - número mínimo de noites de estadia para cada Airbnb (as regras do calendário podem ser diferentes).
- **Number_of_reviews** - O número de avaliações que cada Airbnb tem.
- **Last_review** - A data da última avaliação do Airbnb.
- **Reviews_per_month** - O número de avaliações que a listagem tem ao longo da vida da mesma.
- **calculated_host_listings_count** - O número Airbnb que o anfitrião tem na atual lista, na geografia da cidade/região.
- **Availability_365** - A disponibilidade da listagem x dias no futuro, conforme determinado pelo calendário. Notar que uma Airbnb pode não estar disponível porque foi reservada por um convidado ou bloqueada pelo anfitrião.
- **Number_of_reviews_ltm** - O número de revisões que cada Airbnb tem (nos últimos 12 meses).
- **License** - O número de licença/permissão/registo.

Ao lado, na *figura 1*, podemos visionar a classificação de todas as variáveis apresentadas anteriormente de forma a compreender melhor cada uma delas, podendo assim manipulá-las corretamente a fim de podermos chegar ao objetivo pretendido.

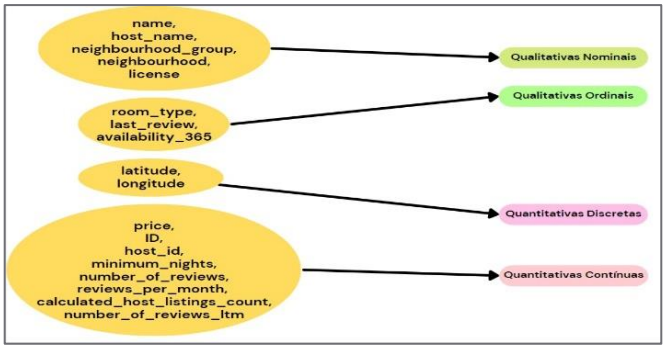


Figura 1 *- classificação das variáveis das bases de dados*

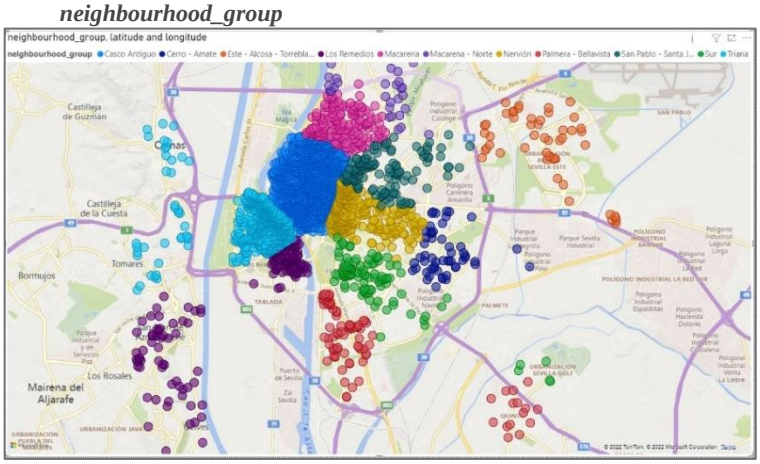
Problemas/questões a responder/resolver

Este trabalho tem como objetivo descobrir quais variáveis têm uma maior influência no nosso variável target ("price"). Para isso, o grupo começará com a hipótese de que todas as variáveis presentes na base de dados que nos foi atribuída, possuem alguma influência e, à medida que o grupo avança, iremos visionar quais destas se destacam mais e menos.

Escolha das variáveis para estudo

As variáveis escolhidas para se trabalhar foram: todas as variáveis numéricas (menos o **id**) e 2 variáveis categóricas, sendo estas as variáveis room_type e o neighbourhood_group. As restantes das variáveis não foram usadas para se estudar/aplicar no modelo, pois, o grupo apercebeu-se que estas seriam irrelevantes para o modelo, não tendo quase nenhuma, ou mesmo nenhuma. correlação com a nosso variável target **price**. Exemplo disso é a variável **id** que é diferente para todas as linhas da base de dados, diferenciando as linhas. Em relação às variáveis categóricas, as 2 escolhidas foram as que pareciam ser mais marcantes no modelo. A variável **room_type** refere qual dos 4 tipos de quarto esse Airbnb possui, podendo influenciar o preço.

Figura 2 *- Localizações dos diferentes Airbnb, relacionado com o* ***neighbourhood_group***



Já a variável **neighbourhood_group** revela a localização do Airbnb, estando estas localizações "codificadas" e agrupadas. O grupo acredita que esta variável possui alguma correlação com a latitude e a longitude, pois, ambos estão relacionados à localização dos Airbnb.

A *figura 2*, mostrada ao lado, comprova aquilo que foi referido anteriormente. Esta foi realizada num programa intitulado de PowerBI. Os pontos marcados no mapa representam os diferentes Airbnb existentes na base de dados, estando estes marcados com a sua respetiva latitude e longitude. Em seguida, estes foram agrupados por neighbourhood_group e, analisando a *figura 2* é notório a existência de

"comunidades" todas agrupadas em cores, dando a entender que existe alguma relação entre estas variáveis.

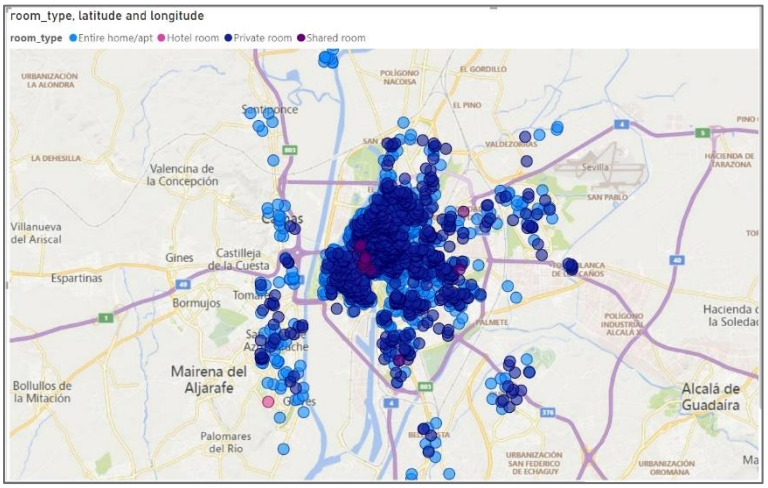


Figura 3 *- Localizações dos diferentes* tipos de quartos.

Passando para a análise da próxima variável categórica que nos chamou à atenção, o **room_type**, o grupo já tendo criado uma variável dummy para inserir esta variável no modelo, podendo assim diferenciar os tipos de quartos diferentes, o grupo tentou descobrir se existia alguma relação na localização dos diferentes quartos. Ao observar a *figura 3*, descobrimos que não existe um padrão em concreto.

Importação da base de dados

Antes de prosseguirmos para o próximo tópico, é necessário referir que, caso esta base de dados seja aberta no Excel (após referir que estes têm de ser divididos por vírgulas) aparecem **6512 linhas** (*figura 2*);

6510	7,23001E+17	Charming 2 Be
6511	7,23121E+17	Un lugar de fá
6512	7,23888E+17	San Felipe - V
6513		
6514		
6515		

Figura 4 *- nº dados excel*

Após a leitura dos dados no R através do comando:

sevilla <— read.csv("listings.csv")

e guardados os dados na variável **sevilla**, fomos contar o número de linhas existentes no ficheiro utilizando **nrow(sevilla),** este devolve-nos o valor de **6494 linhas**;

Por algum motivo, o R não contabiliza 17 linhas (ter em atenção que a primeira linha do excel são o nome das colunas, então este não deve ser contabilizado), portanto, fomos estudar mais a fundo sobre o que realmente estava a acontecer. Na *tabela 1* temos as linhas que não foram contabilizadas pelo R.

Analisando estas linhas, é notório que estas possuam colunas deslocadas. De seguida, fomos ao site visionar informação adicional sobre esta base de dados e descobrimos o seguinte:

	A	B	C	D	E	F	G	H	I	J	K
1	id	name	host_id	host_name	neighbourhood_gro	neighbourhood	latitude	longitude	room_t	price	minimum_nights
1105		106039508		Los Remedios	Los Remedios	37.37848	-6.00011	Entire home/apt	55	2	166
1312		20234432		Casco Antiguo	San Bartolomé	37.38875	-5.9853	Entire home/apt	54	2	220
1840		173856414		Casco Antiguo	Encarnación, Regina	37.39563	-5.99197	Entire home/apt	120	3	33
3062		273366117		Macarena - Norte	Bda. Pino Montano	37.42588	-5.96973	Private room	17	1	33
3277		54863577		Los Remedios	Los Remedios	37.37898	-5.99844	Entire home/apt	50	1	102
3524		325898609		Casco Antiguo	San Bartolomé	37.38633	-5.98744	Entire home/apt	129	2	10
4080		131041944		Casco Antiguo	Alfalfa	37.39226	-5.99383	Entire home/apt	55	2	1
4113		223219051		Casco Antiguo	Santa Cruz	37.3872	-5.99386	Entire home/apt	93	1	169
4197		170512874		Casco Antiguo	San Vicente	37.39519	-5.9968	Entire home/apt	286	1	9
5330		419943077		Nervión	Nervión	37.38227	-5.96368	Entire home/apt	60	2	17
5389		44289797		Este - Alcosa - Torreblanca	Colores, Entreparques	37.39554	-5.80406	Entire home/apt	264	2	0
6018		406815931		Casco Antiguo	Alfalfa	37.39052	-5.99487	Entire home/apt	115	2	2
6170		2314391		Casco Antiguo	Feria	37.39991840000001	-5.989340511639411	Entire home/apt	106	2	1
6395		209110979		Casco Antiguo	Feria	37.39781918343608	-5.991300707404811	Entire home/apt	79	2	0

Tabela1 – Linhas deslocadas no excel

Visualizando a *Tabela 3*, é possível verificar que é referido que esta base de dados possui apenas **6494 linhas**. Então, o grupo optou por ignorar estas 17 linhas, pois, não só são poucas linhas, como também estas estão "a mais" se tivermos em atenção ao que é referido pelo próprio *website* que disponibilizou a base de dados.

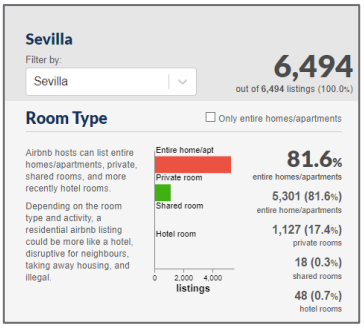


Figura 5 *- nº de linhas no site de onde os dados foram retirados*

Visualização e tratamento de dados

Começamos por importar o dataset que nos foi atribuído, dado por "listings.csv". Em baixo é possível verificar uma tabela com as 5 primeiras linhas, podendo assim obter alguma contextualização. (ter atenção que não aparecem todas as variáveis)

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews
32347	Explore Cultural Sights from a Family-Friendly Apartment	139939	Alejandro	Casco Antiguo	San Vicente	37.39358	-5.99975	Entire home/apt	99	2	168
49287	BEAUTIFUL APARTMENT IN SEVILLE	224697	Walter	Casco Antiguo	San Lorenzo	37.39898	-5.99533	Entire home/apt	75	3	42
94187	(2) ROOM + PRIVATE BATHROOM. CASA DEL BUEN VIAJE	503692	Margot	Casco Antiguo	San BartolomÃ©	37.38816	-5.98537	Private room	79	2	86
108236	Sunny apt in heart of seville!!	560040	Pepe	Casco Antiguo	San Lorenzo	37.39794	-5.99795	Entire home/apt	84	2	160
108568	TERRACE ALAMEDA. WIFI GARAGE DOWNTOWN	589600	Miguel	Casco Antiguo	San Lorenzo	37.39941	-5.99379	Entire home/apt	85	3	103

Tabela2 – *Linhas da base de dados*

Visualização das estatísticas descritivas das variáveis/ limpeza dos dados

Figura 6 *– Summary das nossas variáveis*

```
   latitude       longitude                room_type        price
Min.   :37.30   Min.   :-6.047   Entire home/apt:5301   Min.   :   0.0
1st Qu.:37.39   1st Qu.:-5.998   Hotel room     :  48   1st Qu.:  68.0
Median :37.39   Median :-5.992   Private room   :1127   Median :  99.0
Mean   :37.39   Mean   :-5.992   Shared room    :  18   Mean   : 147.4
3rd Qu.:37.40   3rd Qu.:-5.987                          3rd Qu.: 145.0
Max.   :37.46   Max.   :-5.804                          Max.   :9036.0

minimum_nights    number_of_reviews reviews_per_month
Min.   :  1.000   Min.   :  0.00    Min.   : 0.010
1st Qu.:  1.000   1st Qu.:  4.00    1st Qu.: 0.510
Median :  2.000   Median : 21.00    Median : 1.230
Mean   :  3.844   Mean   : 60.64    Mean   : 1.712
3rd Qu.:  2.000   3rd Qu.: 75.00    3rd Qu.: 2.490
Max.   :380.000   Max.   :963.00    Max.   :12.210
                                    NA's   :739
calculated_host_listings_count availability_365 number_of_reviews_ltm
Min.   :  1.00                 Min.   :  0.0    Min.   :  0.0
1st Qu.:  1.00                 1st Qu.: 70.0    1st Qu.:  1.0
Median :  4.00                 Median :144.0    Median :  8.0
Mean   : 13.86                 Mean   :165.1    Mean   : 16.7
3rd Qu.: 15.00                 3rd Qu.:275.0    3rd Qu.: 25.0
Max.   :125.00                 Max.   :365.0    Max.   :172.0
```

Observemos as medidas descritivas das variáveis que nós adotámos como sendo as mais interessantes para estudar em função da variável dependente, o preço. No output da *Tabela 3*

foi observada a existência de valores omissos na variável **reviews_per_month.** De forma a entender por que motivo existiam dados em falta, começámos por estudar as respetivas linhas que possuíam dados omissos:

```
In [9]: sevilla[which(is.na(sevilla$reviews_per_month)),]
```

executed in 2.30s, finished 22:48:57 2022-11-29

host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month
3258098	Apartamentos Reservaloen	Casco Antiguo	Museo	37.39195	-6.000250	Entire home/apt	120	1	0		NA
3258098	Apartamentos Reservaloen	Casco Antiguo	Museo	37.39250	-5.998050	Entire home/apt	122	1	0		NA
3258098	Apartamentos Reservaloen	Casco Antiguo	Alfalfa	37.39133	-5.991100	Entire home/apt	124	1	0		NA
3258098	Apartamentos Reservaloen	Casco Antiguo	Arenal	37.38719	-5.995560	Entire home/apt	416	1	0		NA
4697283	Caridad	Casco Antiguo	San Gil	37.40109	-5.989670	Entire home/apt	80	2	0		NA
4830830	Alejandro	Casco Antiguo	Feria	37.39661	-5.993990	Private room	59	2	0		NA
4830830	Alejandro	Casco Antiguo	San Lorenzo	37.39839	-5.994360	Entire home/apt	120	2	0		NA

***Tabela 3** – Algumas linhas da base de dados que possuem dados omissos*

Analisando os valores, é notório que as linhas que possuem dados omissos na coluna **reviews_per_month,** não possuem nenhuma avaliação (**number_of_reviews**), nem nenhuma última avaliação (last_review). Devido a isso, o grupo optou por substituir os valores omissos por 0, já que esses Airbnb não possuem avaliações.

Após isso, fomos avaliar as restantes variáveis numéricas representadas na *figura 4*, de forma a investigar outliers e valores incoerentes de algumas delas.

Nas variáveis **reviews_per_month** (fora os NA's), **number_of_reviews_ltm**, **number_of_reviews** e **availability_365** não houve nenhuma alteração na base de dados, após o estudo separado das mesmas.

Em relação à variável **minimum_nights**, o mesmo não ocorreu. Após uma breve análise na variável **minimum_nights**, foi identificado um Airbnb que possuía 380 noites como **minimum_nights**, algo que logo ao princípio se concluiu como sendo um erro. Após algumas pesquisas no site do airbnb pelo quarto que apresentava 380 noites como **minimum_nights**, o grupo acabou por descobrir que afinal se tratava de um dado real.

Por esta variável possuir dados muito dispersos, e estes serem todos verdadeiros, o grupo decidiu criar uma coluna com os valores logaritmizados desta variável, com

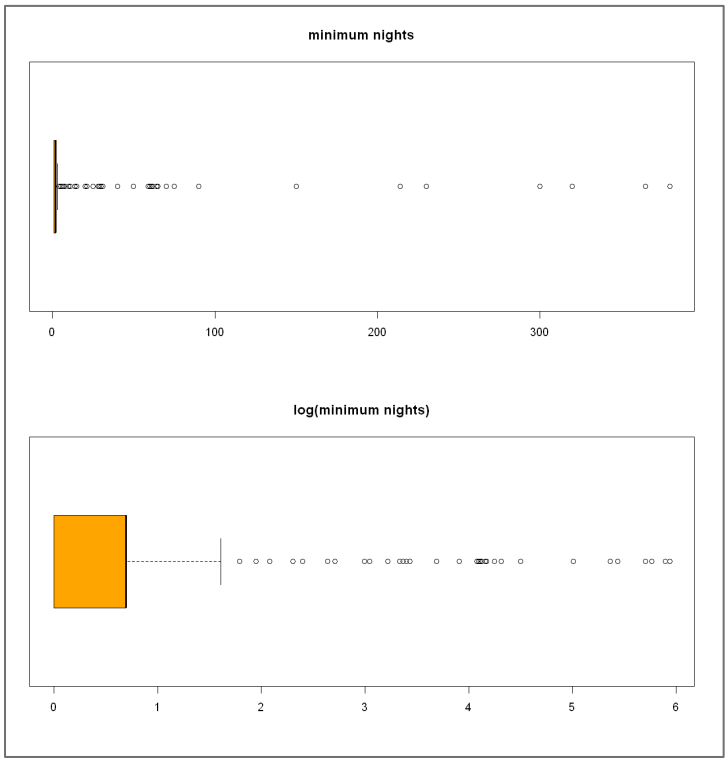


Figura 7** – Boxplot da variável **minimum_nights

o objetivo de diminuir a grande dispersão existente nesta variável (outliers) *figura 7*.

Outra variável que o grupo optou por logaritmizar foi a nosso variável target, o **price**. Após estudar esta variável com especial atenção, tal e qual como a variável anterior, através de um boxplot (*figura 8*) visionou-se a existência de outliers, já esperados pelo summary (sevilla) da *figura 6*. Com isso, tornava-se difícil a perceção da distribuição dos valores, devido à escala apresentada. Esta vasta distribuição dos preços, possivelmente, é originada pela existência de Airbnb de luxo.

Importante referir também que, na análise destes valores, foi identificada uma linha (linha 3978, descrita no código) que possuía o preço como sendo 0, mas também outras variáveis como **last_review**, **reviews_per_month** ou **number_of_reviews**. Assim, foi decidido que se retirava esta linha da base de dados, sendo criado o subset **sevilla1**, sem esta linha apenas.

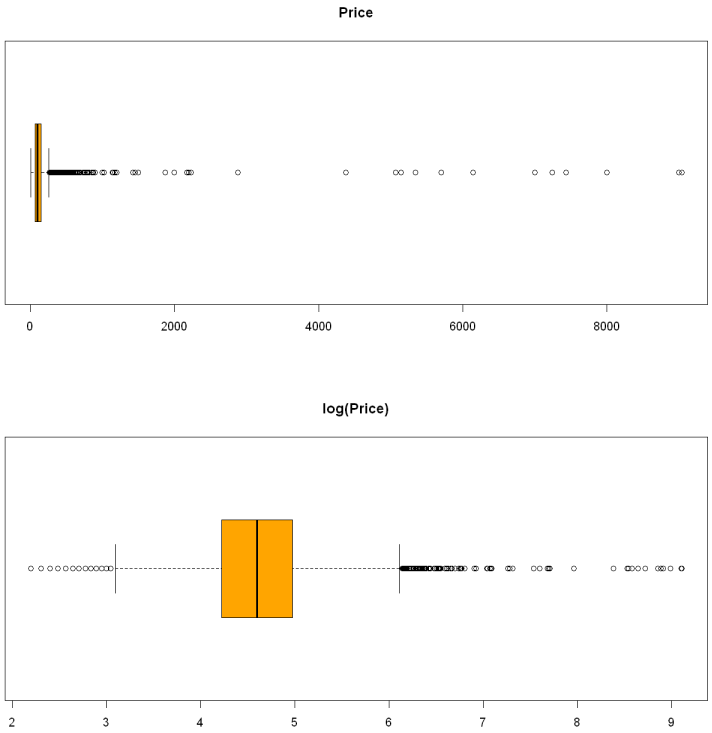


*Figura 8 - boxplot da variável **price***

Variáveis numéricas

Assim, após a análise descritiva dos dados, o grupo efetuou todas as mudanças referidas anteriormente, com o objetivo de deixar as variáveis numéricas da melhor forma para conseguirem ter uma maior relevância e impacto no modelo que o grupo irá apresentar já à frente.

De seguida, mostraremos uma imagem com todos boxplots das variáveis numéricas:

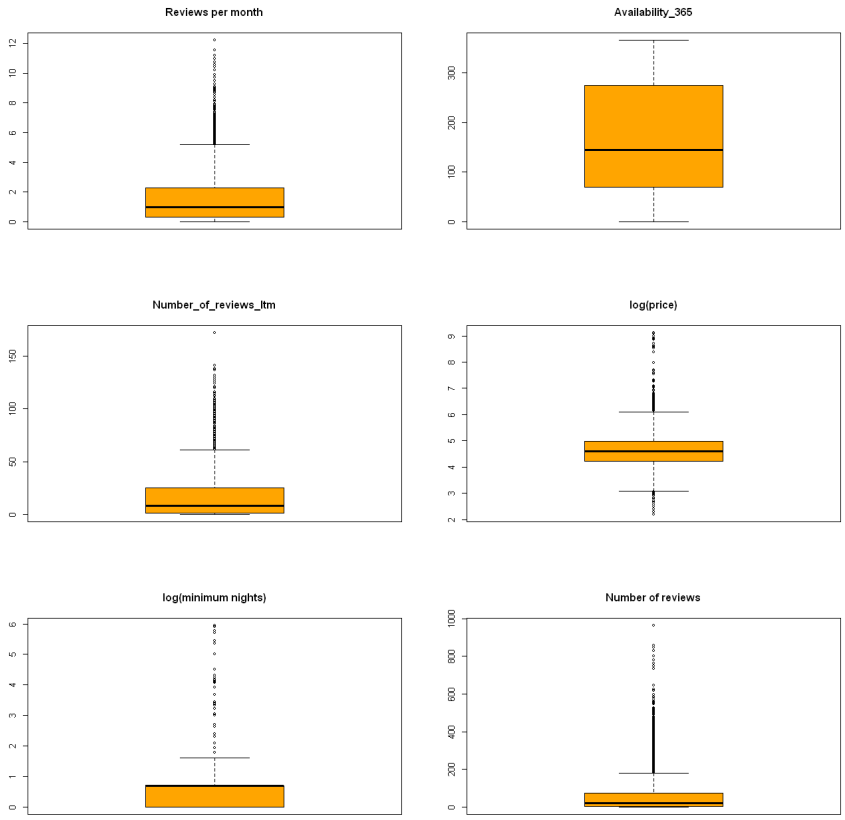


Figura 9 – boxplot das variáveis numéricas

Variáveis dummy

Já finalizado o estudo das variáveis numéricas, passamos então para as restantes, as categóricas. Após uma breve análise, originalmente, o grupo optou por converter apenas uma variável desta base de dados em variável dummy, sendo esta a variável **room_type,** já comentada anteriormente. Esta variável é do tipo qualitativa ordinal e varia entre 4 categorias (**Entire home/apt, Hotel room, Private rooms, Shared rooms**). Então, originámos uma nova coluna intitulada de **dummy_room_type.** De forma a realizar este processo, aplicámos o seguinte código em **R:**

sevilla$dummy_room_type <— as.factor(sevilla$room_type)

Apenas no momento da criação de modelos, que será apresentado mais à frente (modelo 5, concretamente), decidimos criar mais uma dummy no **neighbourhood_group** porque sentimos a necessidade de estudar se o local poderia influenciar o preço do imóvel.

Casco Antiguo: 4116

Cerro - Amate: 54 **Este - Alcosa - Torreblanca:** 53 **Los Remedios:** 158 **Macarena:** 187 **Macarena - Norte:** 70 **Nervión:** 318 **Palmera - Bellavista:** 101 **San Pablo - Santa Justa:** 130 **Sur:** 107 **Triana:** 724

*Figura 10 – variável dummy **neighbourhood_group***

Observando a *figura 10*, observamos uma grande disparidade entre os grupos nos quais os airbnb estavam alocados e, por isso, criado uma variável dummy com dois grupos, um só com Casco Antiguo, e outro com os restantes.

Nota: para a figura, foi utilizado um subset criado posteriormente no código (**sevilla2**)

Correlação

De seguida, realizou-se a análise da correlação entre as variáveis numéricas da base de dados, com as variáveis price e minimum_nights logaritmizadas. Em baixo, é possível verificar que não existe, de forma geral, uma grande relação entre as variáveis; existindo apenas uma relação forte entre as variáveis **number_of_reviews** e **reviews_per_month; number_of_reviews** e **reviews_per_month_ltm.** Veremos a matriz de correlação, na tabela seguinte:

	latitude	longitude	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	log_price	log_minimum_nights
latitude	1.00	0.08	0.05	0.03	0.00	-0.06	0.03	-0.02	0.02
longitude	0.08	1.00	0.00	-0.04	-0.04	0.02	-0.05	-0.10	0.05
number_of_reviews	0.05	0.00	1.00	0.64	-0.12	-0.08	0.67	-0.18	0.02
reviews_per_month	0.03	-0.04	0.64	1.00	-0.07	-0.09	0.80	-0.19	-0.07
calculated_host_listings_count	0.00	-0.04	-0.12	-0.07	1.00	0.05	-0.06	0.18	-0.14
availability_365	-0.06	0.02	-0.08	-0.09	0.05	1.00	-0.07	0.11	-0.03
number_of_reviews_ltm	0.03	-0.05	0.67	0.80	-0.06	-0.07	1.00	-0.15	-0.06
log_price	-0.02	-0.10	-0.18	-0.19	0.18	0.11	-0.15	1.00	-0.21
log_minimum_nights	0.02	0.05	0.02	-0.07	-0.14	-0.03	-0.06	-0.21	1.00

***Tabela 4** – Tabela com a corelação entre as diferentes variáveis numéricas*

Após a visualização da matriz de correlação, os dados serão transpostos para um corrplot, recorrendo à biblioteca corrplot do R.

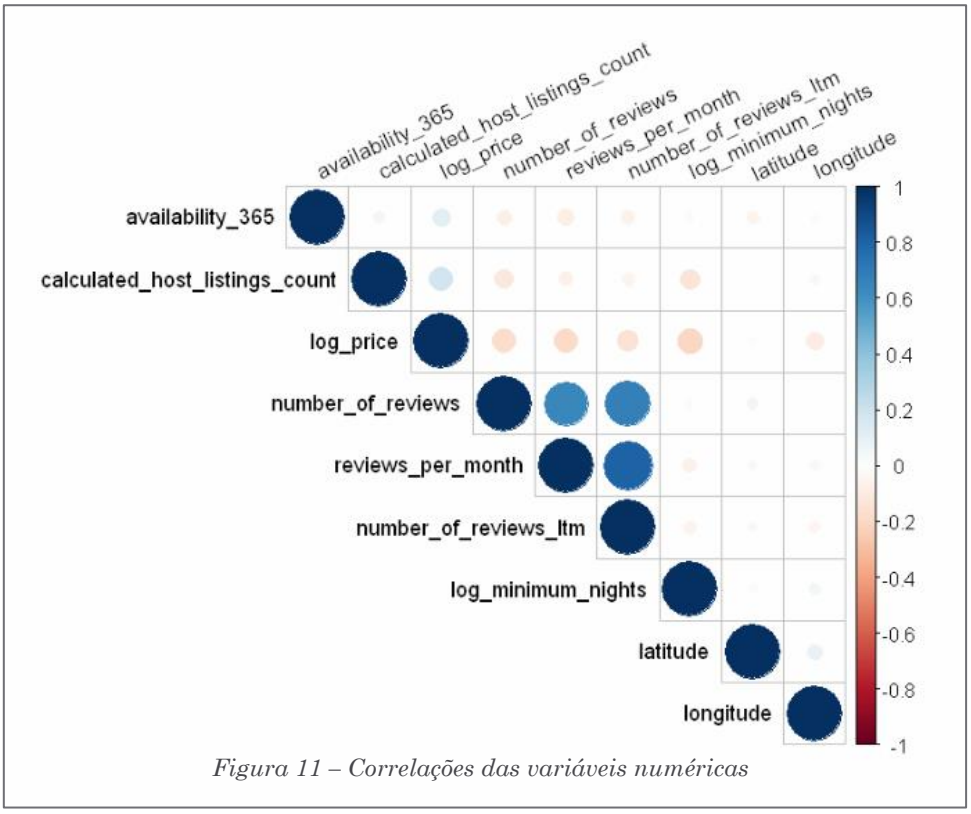


Figura 11 – Correlações das variáveis numéricas

Execução de modelos

Inicialmente o grupo começou por fazer um *baselinemodel*, que possuía todas as variáveis da base de dados pretendidas sem logaritmização de variáveis (pretendia-se verificar as variáveis originais, sem alterações), tendo-o intitulado de **baselinemodel**. Abaixo podemos visionar o modelo em si:

```
baselinemodel <- lm(price ~ dummy_room_type + latitude + longitude + minimum_nights + number_of_reviews + reviews_per_month + calculated_host_listings_count + availability_365 + number_of_reviews_ltm, data = sevilla1)
```

Ao analisarmos o output do *baselinemodel*, o grupo reparou que este modelo estava muito longe de ser um modelo ideal, pois, este modelo possui um R^2 de 0.01803 (um valor muito baixo) e apenas 4 variáveis significativas a respeitar o nível de significância de 5% (0.05 código de insignificância: *).

```
Residual standard error: 363 on 6481 degrees of freedom
Multiple R-squared:  0.01803,   Adjusted R-squared:  0.01636
F-statistic: 10.82 on 11 and 6481 DF,  p-value: < 2.2e-16
```

Figura 12 – summary do baselinemodel

De seguida, iremos descrever todos os processos que o grupo realizou até chegar ao modelo final, referindo o número do modelo onde houve a mudança, sendo que os novos modelos possuem as mudanças realizadas anteriormente.

Para então o grupo conseguir encontrar um modelo que cumprisse minimamente o objetivo, o grupo começou por fazer várias experiências e testes com as variáveis.

Inicialmente, foi utilizada a função **stepAIC**, função que devolve a melhor versão do modelo dado como input, através do menor valor do critério de informação akaike, e reparámos que as variáveis **Latitude** e **Longitude** não tinham uma grande influência em qualquer versão do modelo. Consequentemente, o grupo optou por retirar as respetivas variáveis, referidas anteriormente, no modelo seguinte (**Modelo1**). Já que a localização do Airbnb em si não tinha quase nenhuma influencia no preço, o grupo decidiu criar a variável dummy da neighbourhood_group, já referida anteriormente, de forma a estudar se a localidade do airbnb, se colocado em "grupos", influenciava o preço.

Passando para a próxima alteração, o grupo simplesmente retirou os outliers da variável **price**. Desta forma, foram deixados os valores mais coesos (**Modelo2**), tendo obtido um novo subset sem os outliers. Esta nova base de dados foi intitulada de **sevilla2**, sendo retirados cerca de 400 linhas de **sevilla1**.

Após realizar todas as mudanças referidas anteriormente e criar/executar alguns modelos, o grupo chegou à etapa onde seria necessário estudar a não linearidade das variáveis presentes. O primeiro modelo onde isto foi aplicado foi no **modelo6.** Antes de prosseguirmos, iremos apenas resumir as pequenas alterações que foram oficialmente realizadas neste, e nos modelos anteriores, tais como:

1. adição de variáveis logaritmizadas, já criadas nas medidas descritivas (**price** e **minimum_nights**) – **Modelo3**;
2. criação de interações entre variáveis com forte correlação (**reviews_per_month** e **number_of_reviews**) – **Modelo4**;
3. adição e alteração de variáveis dummy (adicionada variável **dummy_ng**, do **neighbourhood_group**, em um novo subset intitulado de **sevilla3)** – **Modelo5**

Nesta etapa do **Modelo6**, começámos por verificar a não linearidade que, eventualmente, poderia existir nas variáveis de estudo. Após algum estudo, o grupo reparou numa pequena não linearidade em **host_listings_count**, como identificado na figura seguinte.

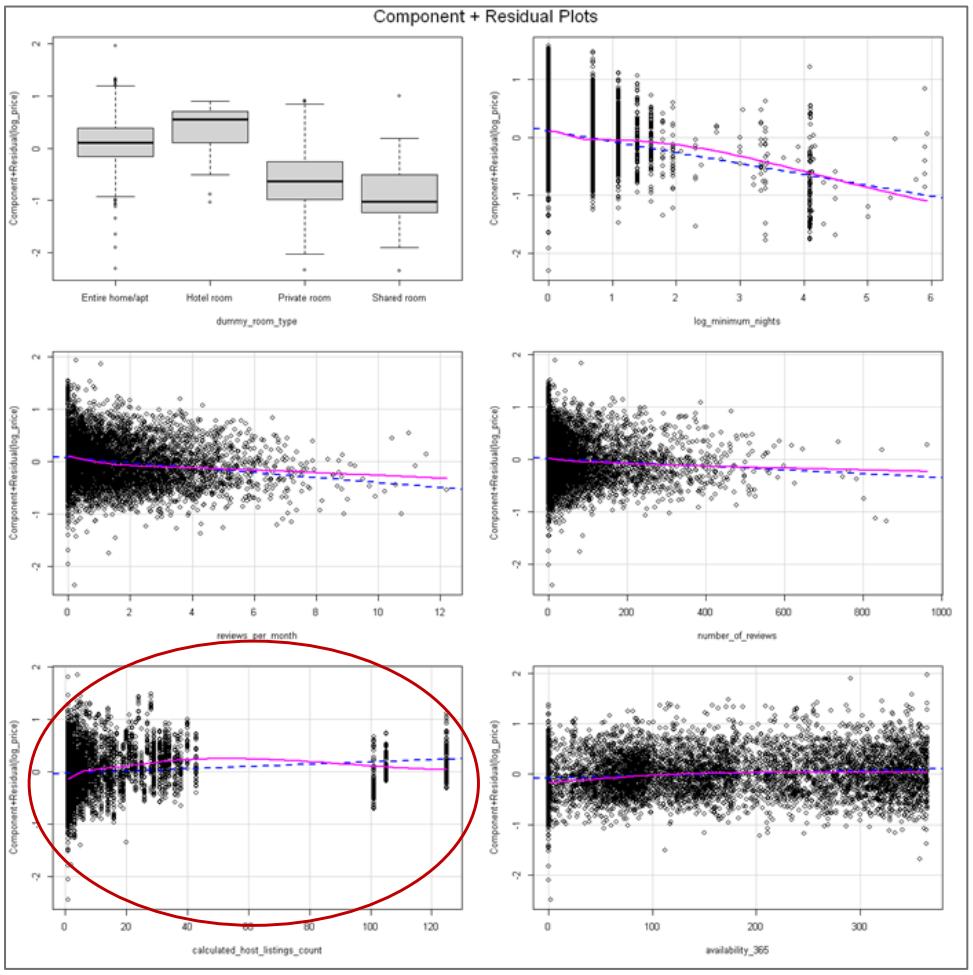


*Figura 13 – não linearidade em **host_listings_count***

Assim, surgiu a possibilidade de fazer uma não linearidade, que foi dividida em três modelos semelhantes (modelo 6, modelo 6.1, modelo 6.2) que só mudavam e aumentavam um grau polinomial desta variável. Destes 3 modelos executados, optou-se pelo modelo 6.1 pois, ao aumentar o grau da variável, o grupo reparou que a partir do 4º grau para a frente (Temos o modelo 6.2 como exemplo) já não existia uma melhoria significativo no modelo, sendo possível concluir isso através da visualização do p-value e do R^2, assumindo assim o modelo 6.1 como o melhor destes três (o de 3º grau).

Demonstra-se na figura seguinte o modelo em questão.

modl6_1= lm(formula = log_price ~ dummy_room_type + log_minimum_nights + reviews_per_month*number_of_reviews+poly(calculated_host_listings_count,degree=3,raw =TRUE)+availability_365+dummy_ng, data = sevilla3)

```
Residual standard error: 0.4095 on 6005 degrees of freedom
Multiple R-squared:  0.4353,    Adjusted R-squared:  0.4341
F-statistic: 385.7 on 12 and 6005 DF,  p-value: < 2.2e-16
```

Figura 14 – summary do modelo6.1

Fazendo uma análise breve do output dado pela fórmula, retirou-se um valor de 0.4353 de R^2 e uma rejeição da estatística F, com p-value de *2.2e-16*. Quanto aos pressupostos dos resíduos, observou-se que o 1º seria o único a ser respeitado, dado que a média dos resíduos é praticamente nula.

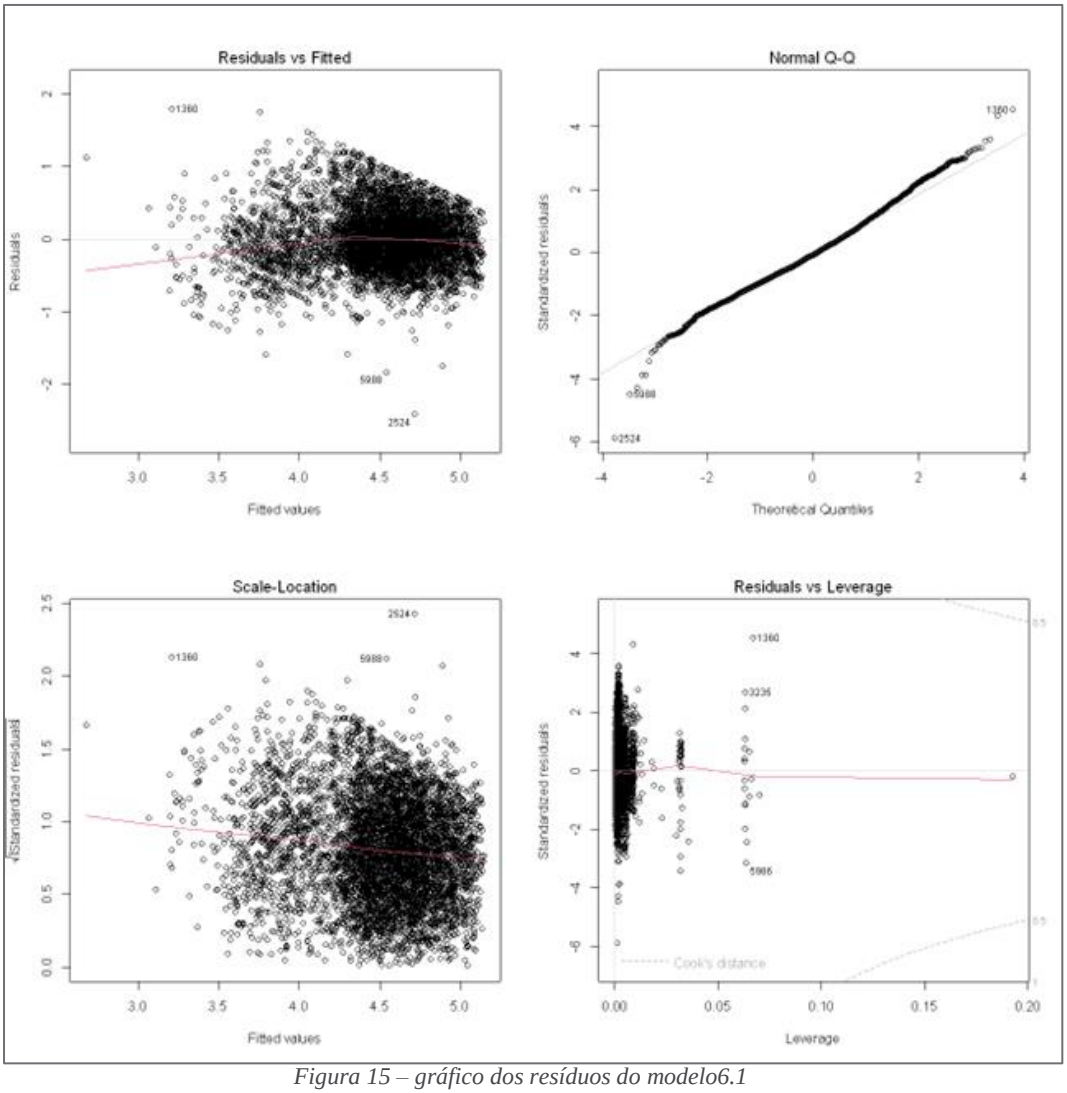


Figura 15 – gráfico dos resíduos do modelo6.1

Após fazer uma reflexão, vimos que o valor de R^2 obtido neste último modelo não estava a ter uma variação considerável, tal como o valor do critério de informação akaike (AIC), que será mostrado e estudado mais à frente, face aos modelos anteriores. Também se verificou que

os pressupostos dos resíduos não se verificavam, exceto o 1º, o da média ser nula, o que nos levou a tomar uma decisão.

O grupo aplicou uma regressão ponderada, com pesos, com o objetivo de tentar minimizar a soma dos resíduos quadrados ponderados, para produzir resíduos padronizados com uma variância constante, por outras palavras, garantir a **homoscedastidade**.

Apresentação do modelo final

O modelo final é composto por: Quase todas as variáveis da base de dados, estando apenas 2 variáveis excluídas (Latitude e Longitude). Em relação a variáveis logaritmizadas, apenas duas se encontram nesse estado, sendo essas: as variáveis **price** e **minimum nights** (O que obrigou o grupo a retirar todas as linhas que possuíam algum zero nestas duas variáveis). Foi aplicado uma interação entre as variáveis **number_of_reviews** e **reviews per month,** devido à alta correlação entre estas. Somado a isto tudo, foram criadas 2 variáveis dummy na variáveis **room_type** e **neighbourhood_group**. Finalizando, para este novo modelo, o grupo aplicou o peso tendo em conta os resíduos do modelo 6.1, caracterizado como o melhor modelo que não possuía uma alteração dos pesos.

$$\text{Pesos} = \frac{1}{\wedge residuals(6.1.)^2}$$

Em baixo podemos visionar a fórmula do modelo 10 (o modelo final):

```
h <- sqrt(modl6_1$residuals^2)

modl10 = lm(formula = log_price ~ dummy_room_type + log_minimum_nights +
reviews_per_month*number_of_reviews +
poly(calculated_host_listings_count,degree=3,raw=TRUE)+ availability_365 + dummy_ng,
data = sevilla3, weights = 1/h)
```

Na *figura 16* é possível visualizar alguma informação sobre os valores estimados para cada variável do modelo 10, tendo usado a função summary do **R**.

```
Coefficients:
                                                                Estimate
(Intercept)                                                    4.520e+00
dummy_room_typeHotel room                                      2.423e-01
dummy_room_typePrivate room                                   -6.603e-01
dummy_room_typeShared room                                    -1.057e+00
log_minimum_nights                                            -1.674e-01
reviews_per_month                                             -4.726e-02
number_of_reviews                                             -8.436e-04
poly(calculated_host_listings_count, degree = 3, raw = TRUE)1  1.981e-02
poly(calculated_host_listings_count, degree = 3, raw = TRUE)2 -4.043e-04
poly(calculated_host_listings_count, degree = 3, raw = TRUE)3  2.125e-06
availability_365                                               4.378e-04
dummy_ng                                                       1.871e-01
reviews_per_month:number_of_reviews                            1.039e-04
```

Figura 16 – summary do modelo10 (valores estimados)

Dado o **summary** das variáveis do modelo, conseguimos observar que todas as variáveis são significativas para um nível de significância de 5% (devido ao número excessivo de variáveis, o grupo optou por não colocar uma imagem com os respetivos p-values, mas esta informação pode ser confirmada através da visualização do código jupyter). Em primeiro lugar, observámos que este modelo de regressão interceta o eixo do O_y, da variável **log_price**, em **4.52** o que significa que, se todas as variáveis forem 0, o valor que passa na reta de regressão será, exatamente, **4.52**. Passemos agora para a **análise detalhada** do modelo em questão.

Antes de analisar os valores estimados das variáveis é importante referir que na expressão polinomial do **calculated_host_listings_count**, não é fácil de interpretar os valores correspondentes ao grau e, por isso, assumimos a não linearidade da variável.

Comecemos a analisar a variável dummy do room_type e os seus valores estimados. Em relação aos **quartos serem casas/apartamentos** (Entire home/apt), retira-se que os **quartos de hotel** (Hotel Room) **aumentam** o preço em **24.23%**, os **quartos privados** (Private Room) **diminuam** o preço em **66.03%** e os **quartos partilhados** (Shared Room) **diminuam** o preço em **105.7%**. Estas conclusões foram realizadas utilizando a comparação entre variável dependente logaritmizada (preço) e variável independente linear (a dummy).

Na **variável minimum_nights**, que se encontra logaritmizada, por cada 1% acrescido ao número de noites mínimas de um imóvel, o **preço decresce 0.16%**, em média. Isto foi concluído devido ao facto da variável dependente (**price**) e da variável independente (**minimum_nights**) estarem ambas logaritmizadas. Sendo assim, o -0.16% é uma medida constante de elasticidade do **price** (X) em relação a **minimum_nights** (Y) - (variações relativas são as mesmas, para quaisquer valores).

Quanto à variável **reviews_per_month**, por cada unidade adicional nesta variável, há um **decréscimo relativo de 4.762%** no preço do airbnb. Tem esta interpretação pelo facto de a variável dependente estar logaritmizada e **review_per_month** não. Por isso a comparação exponencial entre as duas variáveis, em semelhança à comparação da dummy do room_type descrita acima.

A variável **number_of_reviews** possui a mesma interpretação que a anterior e, por isso, para cada review adicional, há um **decréscimo de 0.08436%** na variável **price**.

Na variável **availability_365**, para um aumento de cada unidade da disponibilidade do airbnb, existe um **aumento relativo de cerca de 0.04378%** no nosso target.

Por último, na outra **dummy** criada por duas subdivisões, no **neighbourhood_group**, concluímos que em relação aos imóveis da região de Casco Antiguo, os **restantes grupos de vizinhanças** (*figura 2*) **aumentam** o preço em **18.71%**.

```
Residual standard error: 0.5668 on 6005 degrees of freedom
Multiple R-squared:  0.8769,    Adjusted R-squared:  0.8767
F-statistic:  3565 on 12 and 6005 DF,  p-value: < 2.2e-16
```

Figura 17 – summary do modelo10 (Estatisticas)

Visualizando as medidas do modelo, em geral, é possível verificar que este é quem possui o maior $\boldsymbol{R^2}$, quando comparado aos outros, e, somado a isso, tem os erros dos resíduos não demasiado elevados, embora tenha mais erros que outros. Ainda é verificada a estatística-F, visto que rejeitamos a H_0 (p-value<0.05), o que significa que o modelo poderá ser bom para previsão (será verificado em concreto na fase seguinte).

Em seguida, serão vistos os pressupostos dos resíduos, a fim de validar os testes adjacentes a esta etapa.

Verificação dos pressupostos

De seguida, foi feita a verificação dos pressupostos dos resíduos através da validação dos testes Breusch-Pagan, Breusch-Godfrey, e Jarque Bera, com as funções das bibliotecas do R instaladas.

1. pressuposto: H_0-> a média dos resíduos é zero; $E(\epsilon)$=0. É verificado uma vez que a média dos resíduos é baixa, aproximando-se muito de zero.

mean(modl10$residuals)
0.00292726235989713

2. pressuposto: H_0-> variância constante; $Var(\epsilon i)$=σ2ϵ=σ2. Como o p-value é praticamente 0 e é inferior a 0.05, tomando 5% como nível de significância, rejeitamos a hipótese nula, então a variância não é constante. O gráfico assim deve apresentar uma estrutura em funil, pois há uma violação da homocedasticidade da variância no modelo em questão. (teste **Breusch-Pagan**)

```
        studentized Breusch-Pagan test

data:  modl10
BP = 139414, df = 12, p-value < 2.2e-16
```

*Figura 18 – teste **Breusch-Pagan** modelo10*

3. pressuposto: H_0 -> ausência de autocorrelação; $Cov(\epsilon i, \epsilon j)$=0,($i \neq j$) . O p-value é inferior a 0.05, logo, rejeita-se a hipótese nula, significando que os resíduos estão autocorrelacionados. (teste **Breusch-Godfrey**)

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

data:  modl10
LM test = 125.75, df = 1, p-value < 2.2e-16
```

*Figura 19 – teste **Breusch-Godfrey** modelo10*

4. pressuposto: H_0-> distribuição de probabilidades ϵ é normal; $\epsilon \approx N(0, \sigma)$. O 4º pressuposto, relativo ao teste de Jarque-Bera, também não é verificado, uma vez que o p-value é inferior a 0.05, que obriga à rejeição da hipótese nula. Assim, os resíduos não têm distribuição normal. (teste **Jarque Bera**)

```
        Jarque Bera Test

data:  modl10$residuals
X-squared = 148.91, df = 2, p-value < 2.2e-16
```

*Figura 19 – teste **Jarque Bera** modelo10*

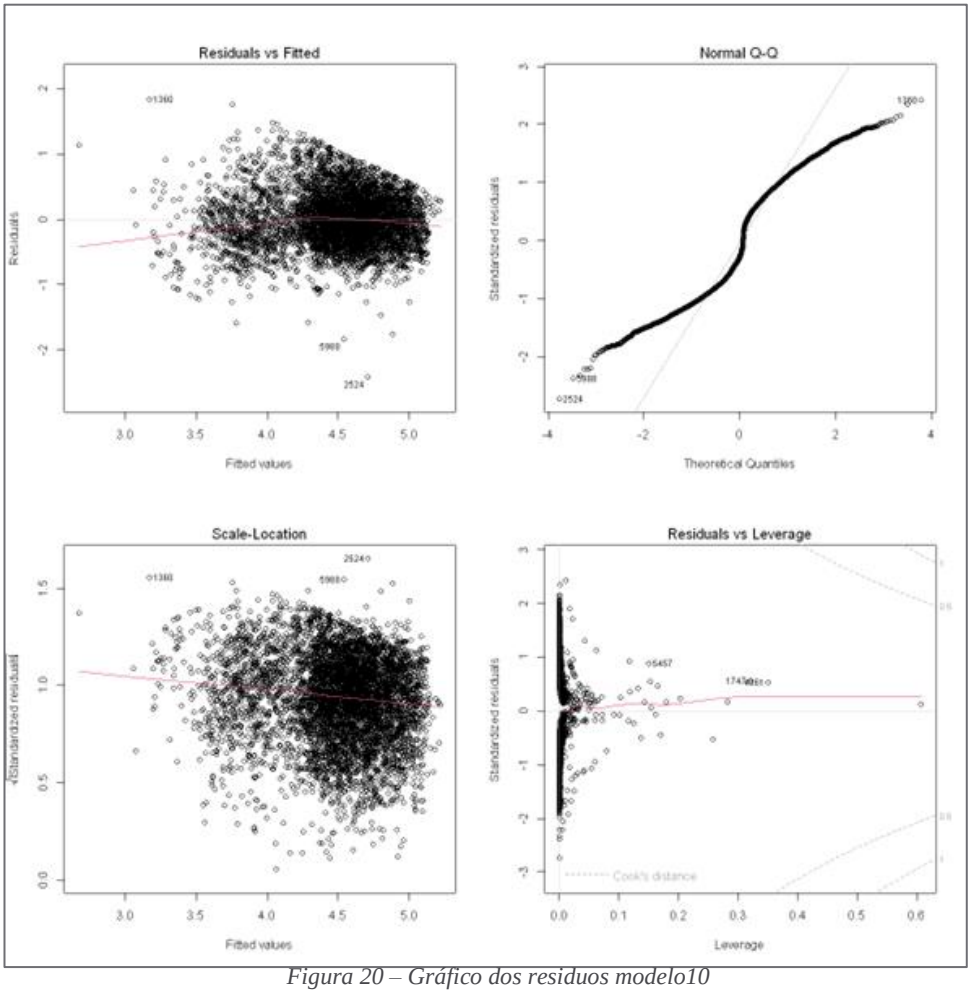


Figura 20 – Gráfico dos resíduos modelo10

Visualizando a *Figura 20*, onde consta a distribuição dos resíduos, é possível verificar que estes se encontram bastante concentrados (estando estes representados pelas bolinhas pretas), formados pela nuvem de pontos pretos.

Relativamente ao segundo gráfico, a curva dispersa da normal, logo, podemos assim concluir que os resíduos não seguem uma distribuição normal.

Para além destes dois, vemos que a variância não é constante e que o gráfico demonstra um "funil", observando para a raiz dos resíduos estandardizados. No entanto, vê-se a não existência de outliers severos para fora da Distância de Cook, ponto positivo do modelo em análise.

Discussão acerca dos modelos realizados, com base no AIC e *R*2

Modelos	AICs	r_squared
baseline model	94985.0405	0.018
modelo 1	94980.2386	0.0175
modelo 2	62689.945	0.2522
modelo 3	6931.425	0.3768
modelo 4	6915.4395	0.3787
modelo 5	6653.6063	0.4051
modelo 6	6434.7013	0.4267
modelo 6.1.	6346.7089	0.4353
modelo 6.2.	6348.4855	0.4353
modelo 7	7115.4756	0.4512
modelo 8	6420.7937	0.4458
modelo 9	6284.4024	0.4246
modelo 10	902.2874	0.8769

***Tabela 5** – Tabela com as estatísticas dos diferentes modelos*

Ao observar a tabela, verificamos que o modelo que se destaca é claramente o modelo 10, com o menor akaike (AIC) de todos os modelos, atingindo o valor de 902.2874. Isto confirma que, de todos os modelos, o modelo 10 é o modelo com melhor qualidade e simplicidade. Em relação ao R^2, o modelo 10 é o modelo que apresenta o valor mais elevado.

Apesar de ser indiscutível que o modelo 10 é o melhor, teremos de dar especial atenção aos modelos 6.1 e 8, pois apresentam valores consistentes das duas colunas mais à direita da tabela. Também será tomado em atenção o modelo 7, que foi o que obteve maior R^2, excluindo o modelo 10. Este último modelo (7), também foi produzido com a utilização de pesos, nomeadamente com pesos referentes à quantidade de linhas da variável dependente a mostrar, o logaritmo do **price**.

Em seguida, serão realizadas previsões in-sample e out-sample para alguns modelos, de forma a identificar se existem casos de overfitting e underfitting, ou se o modelo consegue validar o conjunto de treino, fazendo previsões com o conjunto de teste.

Previsão in-sample

Nesta parte iremos fazer a análise e a previsão in-sample dos modelos, com base no MAPE (erros de previsão de cada um deles), e colocados numa tabela para uma melhor interpretação das previsões.

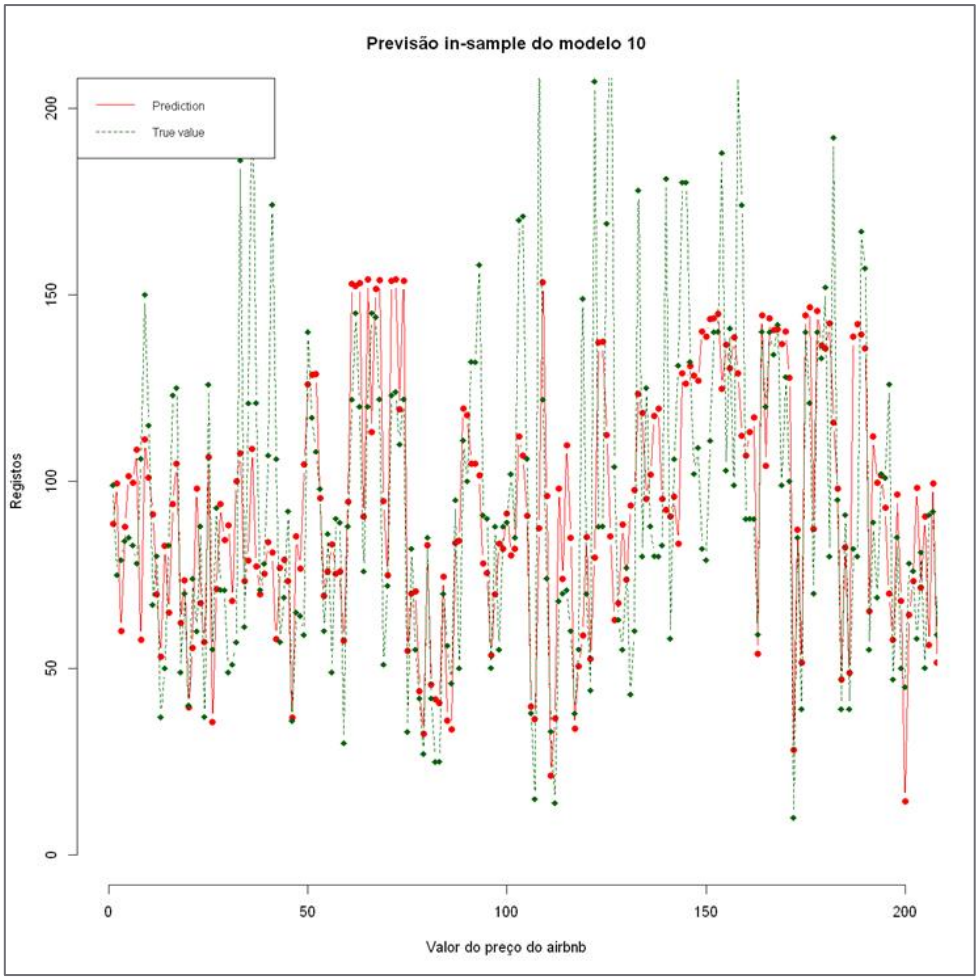
Modelos	AICs	r_squared	MAPE In-Sample	MAPE In-Sample(%)
baseline model	94985.0405	0.018	0.9104	91.04
modelo 1	94980.2386	0.0175	0.9116	91.16
modelo 2	62689.945	0.2522	0.412	41.2
modelo 3	6931.425	0.3768	0.3555	35.55
modelo 4	6915.4395	0.3787	0.3545	35.45
modelo 5	6653.6063	0.4051	0.3464	34.64
modelo 6	6434.7013	0.4267	0.3375	33.75
modelo 6.1.	6346.7089	0.4353	0.334	33.4
modelo 6.2.	6348.4855	0.4353	0.3339	33.39
modelo 7	7115.4756	0.4512	0.3319	33.19
modelo 8	6420.7937	0.4458	0.3338	33.38
modelo 9	6284.4024	0.4246	0.3342	33.42
modelo 10	902.2874	0.8769	0.3325	33.25

***Tabela 6** – Os valores MAPE In-Sample dos diferentes modelos*

Se analisarmos a tabela, é possível verificar que o modelo 10 não é o modelo com o menor valor do MAPE, sendo o modelo 7 que ocupa esse espaço. Apesar disso, o grupo optou por escolher o modelo 10 como o modelo final pois este possui o maior R^2 e o menor akaike (AIC), tal como exposto num tópico anterior.

De seguida, iremos visionar um gráfico que mostra o modelo 10 a prever 200 valores da base de dados (**sevilla3**).

Figura 21 *– Previsão in-sample do modelo10*



Previsão out-sample

A previsão *out-sample* foi realizada para perceber se os modelos de regressão têm poder preditivo fora da amostra. Deste modo, dividiu-se a amostra em dois conjuntos de dados – os de **treino (90%)** e os de **teste (10%)**. Para realizar esta divisão, o grupo usou a biblioteca *caTools*, através do comando splitRatio. Após isso, foram escritos os modelos novamente e adicionados os pesos, aos que necessário, para executar as previsões e guardados numa tabela as informações extraídas.

Em baixo, é possível visualizar a percentagem do erro de previsão out-sample (MAPE) dos modelos que o grupo considerou como **mais importantes em etapas anteriores**, sendo estes os modelos 6.1., 7, 8 e 10.

Modelos1	AICs	r_squared	MAPE in-sample	MAPE in-sample(%)	MAPE out-sample	MAPE out-sample(%)
modelo 6.1.	6346.7089	0.4353	0.334	33.4	0.3323	33.23
modelo 7	7115.4756	0.4512	0.3319	33.19	0.3311	33.11
modelo 8	6420.7937	0.4458	0.3338	33.38	0.3323	33.23
modelo 10	902.2874	0.8769	0.3325	33.25	0.3308	33.08

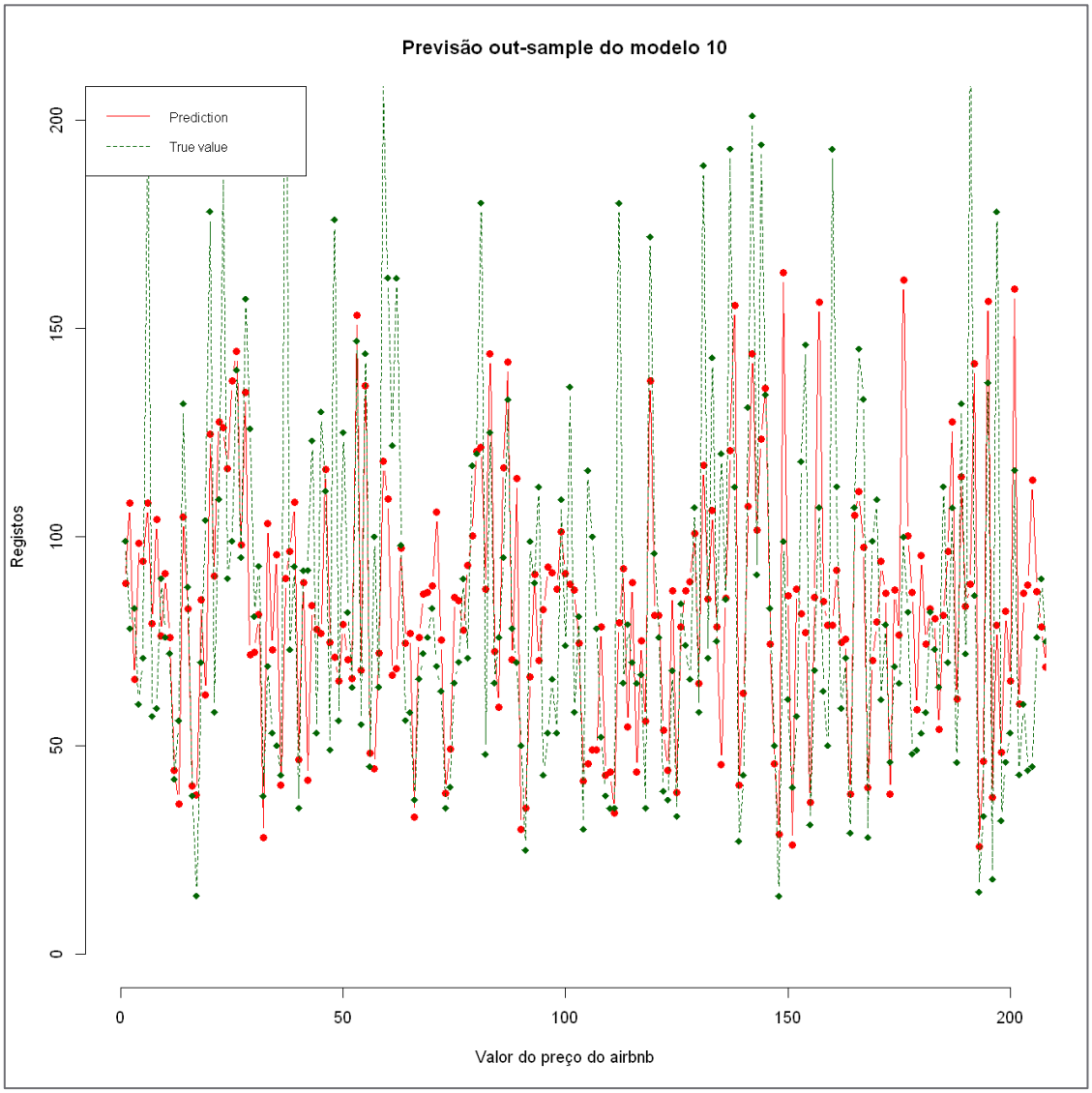
Tabela 7 *– Os valores MAPE para os melhores modelos*

Observando com atenção, é possível verificar que os modelos, no âmbito geral, não variam muito, tomando valores de MAPE in-sample e out-sample bastante semelhantes, quando comparado entre modelos e previsões.

Apesar disso, o modelo com **menor percentagem de erro é o modelo10**, considerando-o como modelo final para a previsão out-sample (ter em atenção que o grupo definiu uma seed inicial, de forma aleatória). Comparativamente ao erro de previsão in-sample, o **modelo 10** possui um erro de previsão out-sample inferior, mas não significativamente menor (aproximadamente **0.20%**). Assim podendo concluir que este modelo tem um **bom poder preditivo fora da amostra**, utilizando o conjunto de treino para testar o modelo e o conjunto de teste para prevê-lo, mas também que não há casos nem de underfitting, nem de overfitting.

Dito isto, será mostrado o gráfico da previsão fora da amostra para este modelo, a prever 200 linhas da base de dados (**sevilla3**).

Figura 22 *– Previsão out-sample do modelo10*



Conclusões

Concluindo, este trabalho permitiu ao grupo aprender vários tópicos sobre Airbnb, já que foi necessário realizar pesquisas para entender as várias variáveis e saber como as iriamos aplicar neste trabalho.

Em relação à nossa pergunta, sobre quais variáveis influenciavam mais o preço dos Airbnb, apesar de ser um pouco difícil de se interpretar valores, podemos concluir que todas as variáveis influenciam o preço, quer seja de forma negativa ou positiva, exceto a **latitude**, a **longitude** e as **number_of_reviews_ltm**.

Para chegarmos a esta conclusão, e para ser possível dar valores em concreto, o grupo necessitou de realizar vários modelos para conseguir chegar a uma resposta devida, tendo chegado ao modelo 10, o melhor modelo. O grupo pôde afirmar que o modelo 10 foi o melhor graças à realização de diversos testes de modelos, comparando os demais. O grupo percebeu que o modelo 10 destacava-se como sendo o melhor modelo, pelo valor do seu R^2(valor mais elevado ≈ 0.8769) e do seu AIC- akaike criteria information (valor mais baixo ≈ 902.3). No entanto, não devemos excluir o desempenho do modelo 6.1., utilizado nos pesos do modelo 10, que se mostrou ter um bom equilíbrio entre os valores de R^2 e AIC face aos restantes modelos, mesmo tendo sido feito sem pesos. Já os pressupostos dos resíduos, mantiveram-se iguais para todos os modelos, obedecendo sempre ao primeiro pressuposto apenas. Quanto às previsões realizadas, é de notar que, entre os modelos testados, o que se mostrou o melhor modelo para previsão fora da amostra foi o modelo 10, por uma margem pequena, considerando o valor do MAPE.

Por todas as razões descritas acima, escolheu-se o modelo 10 como o modelo final do nosso estudo das variáveis da base de dados do airbnb, tendo a variável **price** como target. Este foi o modelo que contribuiu para a conclusão retirada anteriormente, pois permitiu fazer uma sensibilização dos valores obtidos e perceber que, face ao nosso target, existem valores que, quando interpretados, podem fazer sentido. Alguns destes exemplos são: os quartos partilhados e os quartos privados diminuem o preço face às casas; o aumento do mínimo de noites diminui o preço em uma quantidade significativamente pequena, ou seja, quanto mais noites tiver o airbnb, por regra, mais barato é; quanto maior o número de reviews, menor o preço do airbnb.

Assim, consideramos este modelo como o principal, escrevendo-o algebricamente da seguinte forma (figura):

$$
\begin{aligned}
\log(\text{price}) = & 4.52 \\
& + 2.423\text{e}^{-1} \times \text{dummy\_room\_typeHotel room} \\
& - 6.603\text{e}^{-1} \times \text{dummy\_room\_typePrivate room} \\
& - 1.057\text{e}^{-1} \times \text{dummy\_room\_typeShared room} \\
& - 1.674\text{e}^{-1} \times \log(\text{minimum\_nights}) \\
& + 1.039\text{e}^{-4} \times \text{reviews\_per\_month} \times \text{number\_of\_reviews} \\
& + 1.981\text{e}^{-2} \times \text{calculated\_host\_listings\_count} \\
& - 4.043\text{e}^{-4} \times \text{calculated\_host\_listings\_count}^2 \\
& + 2.125\text{e}^{-6} \times \text{calculated\_host\_listings\_count}^3 \\
& + 4.378\text{e}^{-4} \times \text{availability\_365} \\
& + 1.871\text{e}^{-1} \times \text{dummy\_ng}
\end{aligned}
$$

Figura 22 *– Modelo10 escrito da forma algébrica*

Webgrafia

Mendes, D., 2022, Moodle: IMAD, RegressionsCoefficients.pdf

Mendes, D., 2022, Moodle: IMAD, scripts das semanas

AirBnbB, https://www.airbnb.pt/rooms/1017062?_set_bev_on_new_domain=1669485936_MjhlNjg2MzY4NmYx&source_impression_id=p3_1669485937_Ed%2FeGQfdidZIcDja&guests=1&adults=1#availability-calendar

HowToProgram, 2016, https://howtoprogram.xyz/2016/10/08/write-csv-file-in-r/