ISCTE-IUL
Licenciatura de Ciência de Dados

TODO Título

TODO Subtítulo

André Plancha, 105289
<Andre_Plancha@iscte-iul.pt>
Tomás Ribeiro, 105220
<tfroo1@iscte-iul.pt>
Afonso Silva, 105208
<agsos@iscte-iul.pt>
Rui Chaves, 104914
<rfpcs1@iscte-iul.pt>

20/12/2022
Versão 0.0.2

```
df <- read.csv(here("data", "listings.csv"))
shape <- st_read(here("data", "SF Planning Neighborhood Groups Map"))
tmap_mode("plot")
shape_plot <- shape %>%
    ggplot() + geom_sf() + theme(legend.position = "bottom")
```

A base de dados que nos foi disponibilizada vem do projeto, fundado por Murray Cox com a missão de "[...] fornecer dados e defesa sobre o impacto do Airbnb em comunidades residenciais" [2].

A base de dados contém 6629 entradas, e cada uma delas representa um registo de um anúncio para o aluguer de um alojamento disponível no Airbnb, em São Francisco, Califórnia. Cada alojamento contém informação sobre o seu preço, localização, hospedeiro, o tipo de alojamento, as *reviews* do alojamento, e licensa do alojamento.

```
data.frame(row.names = colnames(df), type = sapply(df,
    class)) %>%
    showT()
```

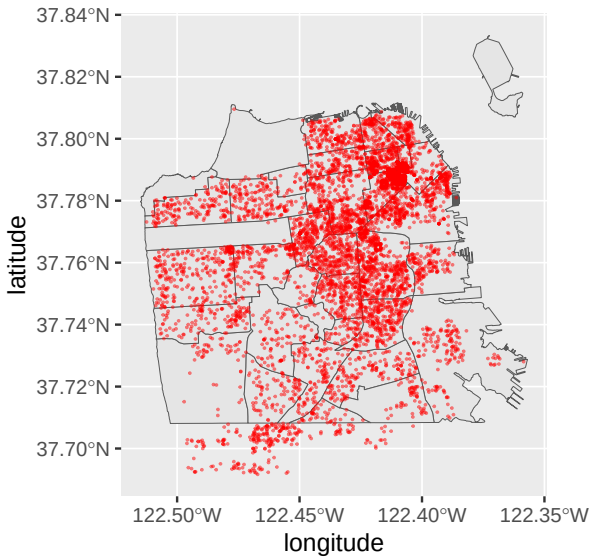
	type
id	numeric
name	character
host_id	integer
host_name	character
neighbourhood_group	logical
neighbourhood	character
latitude	numeric
longitude	numeric
room_type	character
price	integer
minimum_nights	integer
number_of_reviews	integer
last_review	character
reviews_per_month	numeric
calculated_host_listings_count	integer
availability_365	integer
number_of_reviews_ltm	integer
license	character

De forma a perceber melhor a base dados, o *Airbnb* disponibiliza de um "dicionário de dados" [1] que explica o significado de cada uma das variáveis:

- **id**: Número que representa um identificador único do anúncio;
- **name**: Título do anúncio;
- **host_id**: Identificador único da conta do hospedeiro;
- **host_name**: Nome da conta do hospedeiro (Normalmente este campo inclui apenas o primeiro nome ou nome da instituição hospedeira);
- **neighbourhood_group**: Este campo encontra-se vazio e não inclui descrição no dicionário;
- **neighbourhood**: Embora este campo não inclua descrição no dicionário, nesta base de dados este campo representa os bairros de São Francisco como definido pelo Departamento de Planeamento da cidade (os bairros de São Francisco não contém fronteiras oficiais e dependem da fonte (`tldrify.com/19p8`), logo a definição das fronteiras definidas pelo Airbnb tiveram de ser determinadas; mais à frente será demonstrado as fronteiras);
- **latitude/longitude**: Coordenadas geográficas do alojamento;
- **room_type**: Tipo de alojamento, entre "Quarto privado", "Quarto partilhado", "Quarto de hotel", e "Casa/Apartamento inteiro";
- **price**: Preço do alojamento por noite em USD;
- **minimum_nights**: Número mínimo de noites que o hospedeiro exige para alugar o alojamento;
- **number_of_reviews**: Número total de *reviews* que o alojamento tem desde o seu registo no Airbnb;
- **last_review**: Data da última *review* que o alojamento recebeu;
- **reviews_per_month**: Número médio de *reviews* que o alojamento recebe por mês;
- **calculated_host_listings_count**: Número de alojamentos que o hospedeiro tem disponíveis em São Francisco;
- **availability_365**: Número de dias que o alojamento está disponível por ano.
- **number_of_reviews_ltm**: Número de *reviews* que o alojamento recebeu nos últimos 12 meses;
- **license**: A licença/autorização/número de registo do alojamento.

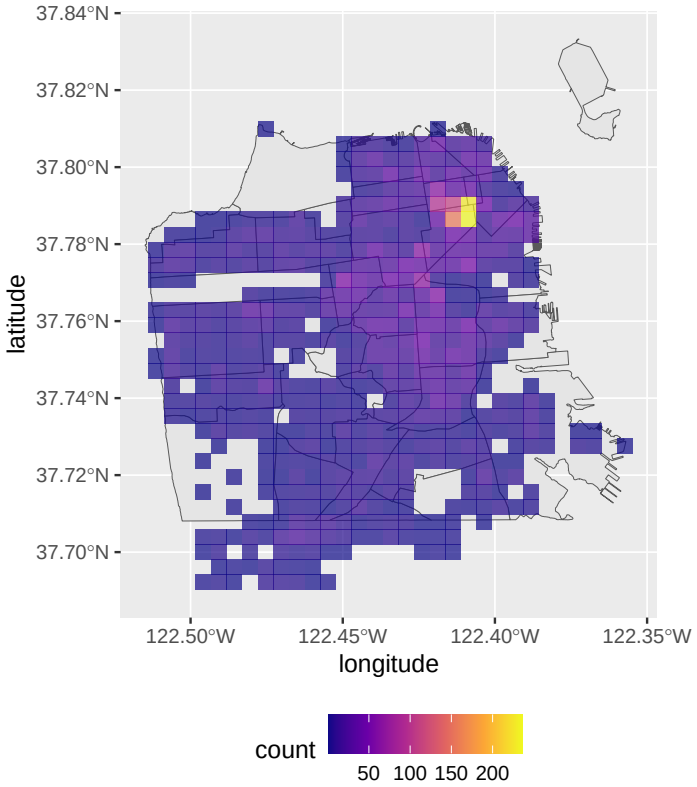
Para o nosso objetivo, algumas colunas não vão ser úteis, devido à sua naturaza. Estas são o id, name, as categorias que referem informações sobre o hóspede (estas colunas conseguem justificar valores atípicos, principalmente em termos de preço; e.g. Um preço extremamente alto pode acontecer devido a um hotel de luxo na cidade. Estes problemas vão ser discutidos mais à frente.), a disponibilidade do alojamento durante o ano, e a licença do alojamento. Como o nosso objetivo será prever o preço esperado baseado na localização do apartamento, não vamos também utilizar variáveis associadas aos hóspedes, como o número de *reviews* e o número de alojamentos que o hóspede tem disponíveis em São Francisco. Cada registo contém as coordenadas geográficas e se as representármos graficamente, podemos verificar que grande parte dos alojamentos se encontram concentrados a nordeste da cidade, principalmente em *Downtown/Civic Center*. Apesar disso, também existem muitos alojamentos no resto da cidade.

```
shape_plot + geom_point(data = df, aes(y = latitude,
    x = longitude), alpha = 0.5, color = "red",
    size = 0.1)
```



Inesperadamente, o mapa mostra alguns pontos de alojamento fora da cidade, mas julgamos que não vá interferir com as nossas análises, devido ao correto agrupamento (demonstrado mais à frente) e à proximidade da cidade. Embora a razão nos seja desconhecida, acreditamos que o próprio Airbnb agrupa desta forma esses locais devido à sua proximidade com a cidade.
A concentração torna-se mais óbvia quando visualizamos o mapa de calor.

```
rast <- (shape_plot + stat_bin2d(data = df,
    aes(x = longitude, y = latitude), alpha = 0.7,
    bins = 30, linejoin = "round") + scale_fill_viridis_c(option = "C"))
rast
```



O mapa claramente demonstra a concentração de alojamentos na zona clara, mas também consegue-se observar uma grande quantidade, embora mais dispersos, na zona central. Este gráfico demonstra uma possibilidade de agrupar os alojamentos nestas zonas.

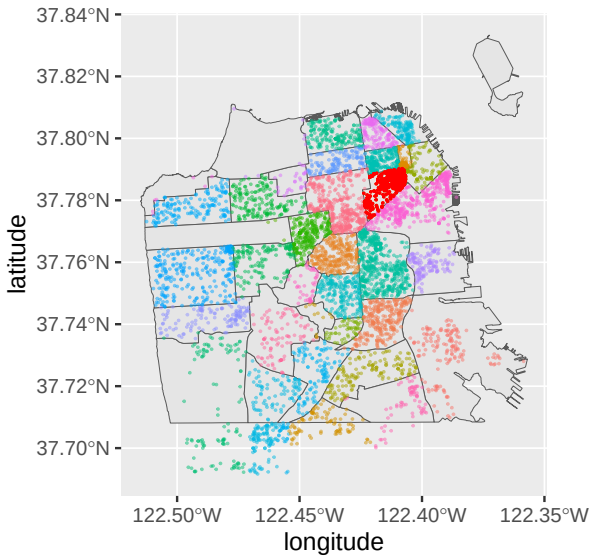
```
df %>%
    group_by(neighbourhood) %>%
    summarise(n = n(), freq = n/nrow(df)) %>%
    arrange(-n) %>%
    head(8) %>%
    showT()
```

neighbourhood	n	freq
Downtown/Civic Center	745	0.1123850
Mission	558	0.0841756
South of Market	450	0.0678835
Western Addition	418	0.0630563
Nob Hill	328	0.0494796
Outer Sunset	281	0.0423895
Bernal Heights	280	0.0422386
Haight Ashbury	276	0.0416352

A tabela mostra que grande parte dos alojamentos listados estão localizados no distrito de *Downtown/Civic Center* e *Mission*.

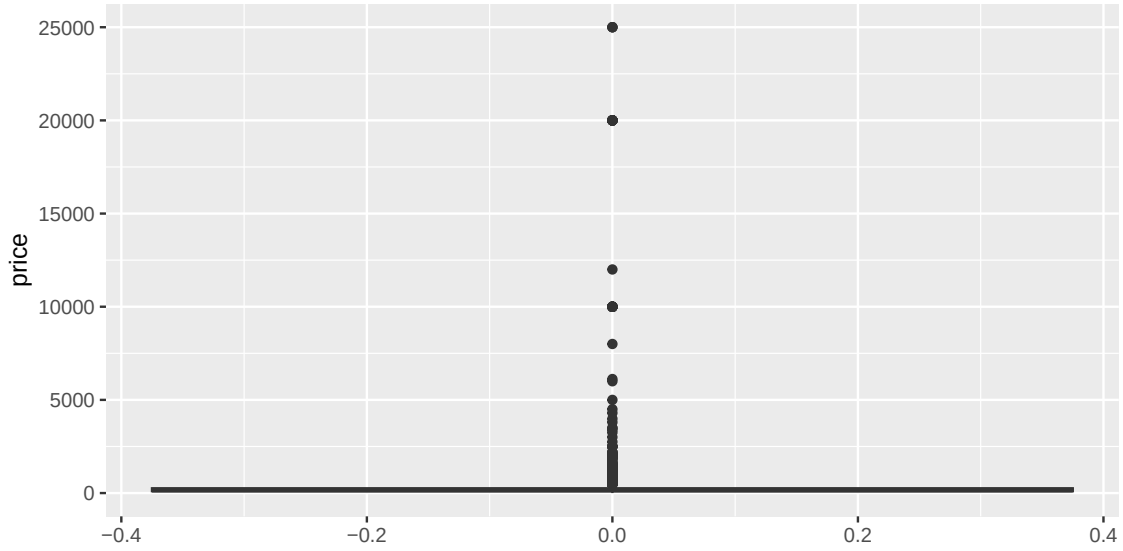
```
shape_plot + geom_point(data = df, aes(y = latitude,
    x = longitude, color = neighbourhood),
    alpha = 0.5, size = 0.1) + geom_point(data = (df %>%
```

```
    filter(neighbourhood == "Downtown/Civic Center")),
    aes(y = latitude, x = longitude), color = "red",
    alpha = 1, size = 0.1) + theme(legend.position = "none")
```



Esta figura demonstra que os bairros estão em conformidade com a definição do Departamento de Planejamento da cidade. Mostra também a posição do distrito *Downtown/Civic Center* a vermelho, através do mapa de calor.

```
ggplot(data = df, aes(price)) + geom_boxplot() +
    coord_flip()
```



Neste *boxplot* do preço conseguimos notar imediatamente a existência de muitos valores atípicos, que equivalem a preços muito altos tendo em conta a média de preços dos registos que é de 303.465 USD. Estes preços vão sem dúvida interferir com as nossas análises. Estes preços conseguem ser explicados quando analisamos a sua fonte.

```
df %>%
    select(name, price) %>%
    arrange(-price) %>%
    head(7) %>%
    showT(T)
```

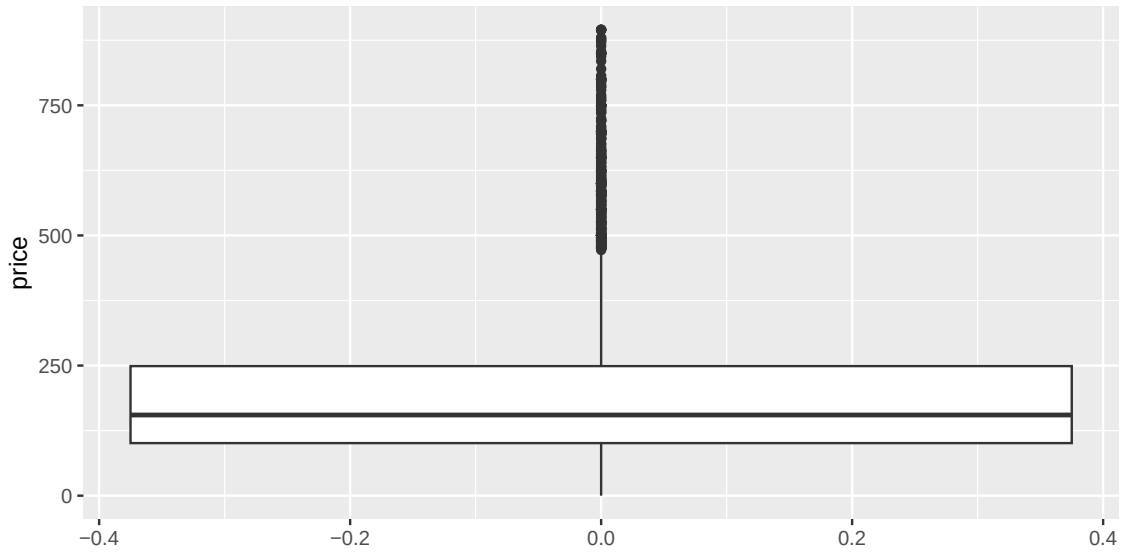
name	price
Harbor Court Hotel, Bay View King Room	25000
Harbor Court Hotel, Bay View Queen Room	25000
Hotel Griffon by the Bay, queen bedded room	25000
1-Bedroom Suite with One Bed and One Sofabed at Fairmont San Francisco by Suiteness	20000
Suite plus Connecting Room with Three Beds and One Sofabed at Fairmont San Francisco by Suiteness	20000
Suite plus Connecting Room with Two Beds and One Sofabed at Fairmont San Francisco by Suiteness	20000
1-Bedroom Suite with One Bed at Fairmont San Francisco by Suiteness	20000

Assim, estes preços equivalem a alojamentos de luxo, que pela sua natureza terão de ser tratadas de forma diferente quando for feita a modelação, uma vez que não são comparáveis com o resto dos alojamentos.

Como tentativa de mitigar estes valores muito altos (*outliers*), provavelmente será necessário fazer uma transformação logarítmica do objetivo, de forma a reduzir a influência destes valores no modelo.

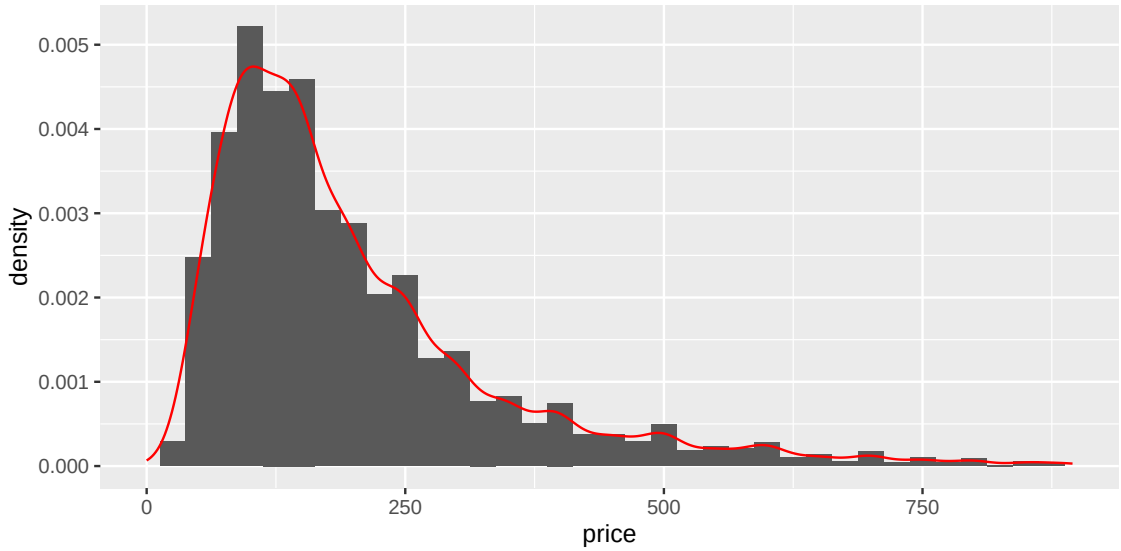
```
upper_limit <- quantile(df$price, 0.975) +
    20
ggplot(data = df, aes(price)) + geom_boxplot() +
    coord_flip() + xlim(0, upper_limit)

## Warning: Removed 158 rows containing non-finite values (`stat_boxplot()`).
```



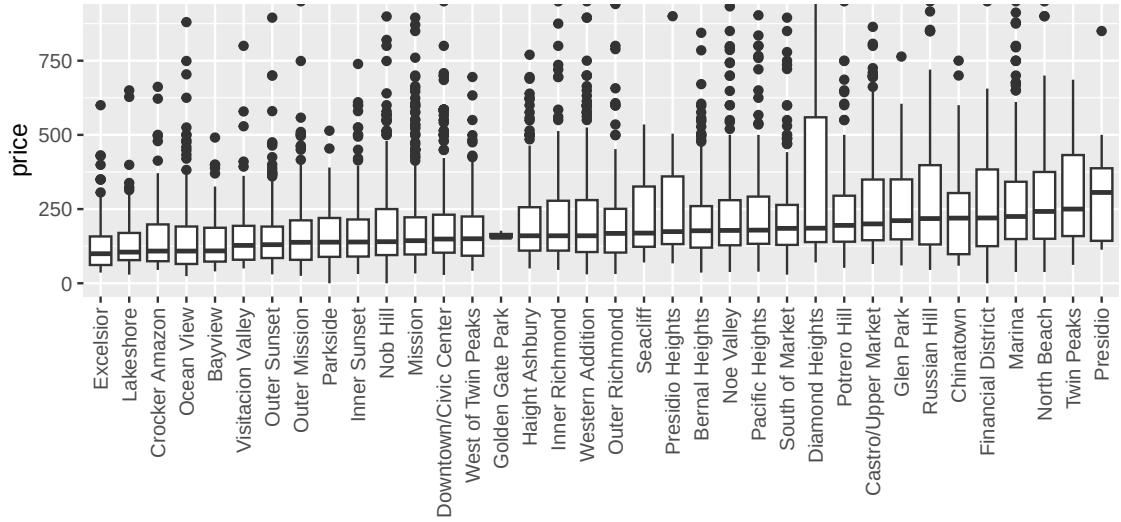
Este *boxplot* mostra que a maioria dos alojamentos tem preços entre 103 e 254 USD. Este facto torna-se ainda mais evidente quando analisamos a distribuição dos preços.
O gráfico mostra também que há muitos alojamentos fora destes limites, podendo ser valores atípicos também, embora não tão extremos como aqueles vistos anteriormente. No entanto, à primeira vista estes não devem ser valores atípicos, devido à sua quantidade, mesmo quando comparado com o número de registos.

```
ggplot(data = df, aes(price)) + geom_histogram(binwidth = 25,
    aes(y = ..density..)) + geom_density(color = "red") +
    xlim(0, upper_limit)
```



A distribuição de preços apresentada demonstra que a maioria dos alojamentos se encontram no limite mostrado anteriormente, e a distribuição parece aproximar-se de uma distribuição χ^2_k, com um pequeno grau de liberdade. Curiosamente, o gráfico mostra que os preços parecem aumentar algumas vezes a cada 50 USD, o que pode ser devido ao facto de que os hospedeiros escolhem preços redondos, como 50, 100, 175, etc. Este fenómeno parece ser mais visível nos 250 e nos 500.

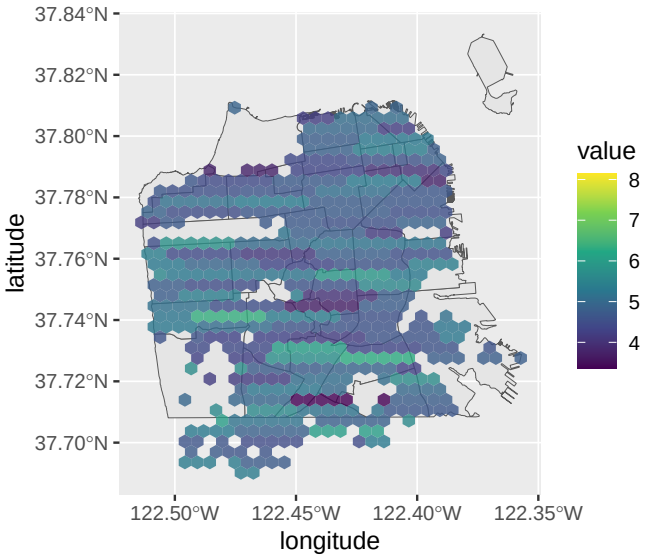
```
df %>%
    ggplot(aes(y = price, x = forcats::fct_reorder(neighbourhood,
        price, .fun = median))) + geom_boxplot() +
    coord_cartesian(ylim = c(0, upper_limit)) +
    theme(axis.text.x = element_text(angle = 90,
        vjust = 0.5, hjust = 1)) + labs(x = "")
```



Os *boxplots* mostram que os preços dos alojamentos não variam bastante de acordo com o bairro sendo que o ponto médio não varia bastante entre bairros, excepto os bairros *Twin Peaks*, e *Presidio*. No entanto, o gráfico mostra uma grande variância dos preços em todos os bairros, exceto no bairro Golden Gate Park.

```
shape_plot + stat_summary_hex(data = df,
    aes(x = longitude, y = latitude, z = log(price)),
    alpha = 0.8) + scale_fill_viridis_c() +
    theme(legend.position = "right")

## Warning: Removed 3 rows containing non-finite values (`stat_summary_hex()`).
```



Este gráfico demonstra a variariadade presente nos vários distritos, e da mesma forma não é possível descrever nenhum padrão *a priori* para os preços.

```
df %>%
    group_by(room_type) %>%
    summarise(n = n(), freq = n()/nrow(.),
        averagePrice = mean(price), sd = sd(price),
        min = min(price), max = max(price)) %>%
    showT()
```

room_type	n	freq	averagePrice	sd	min	max
Entire home/apt	4243	0.6400664	275.8965	406.8473	33	12000
Hotel room	65	0.0098054	266.2154	211.0756	0	1220
Private room	2239	0.3377583	360.1474	1972.6562	0	25000
Shared room	82	0.0123699	211.7683	1101.4410	25	10000

A tabela mostra que a maioria dos alojamentos são apartamentos ou casas inteiras, enquanto que os quartos privados são menos frequentes. Mostra também a pequena quantidade de quartos de hotéis e de alojamentos partilhados, sendo estes apenas 2% dos registos. Isto pode levar a que seja necessário o uso de alguma técnica de *oversampling* para os quartos de hotel e para alojamentos partilhados, de forma a aumentar a quantidade de registos destes tipos de alojamentos.

A tabela também expõe que os quartos privados são os mais caros em média, enquanto que os alojamentos partilhados são os mais baratos. Esta observação é esperada na parte que diz respeito aos quartos partilhados, no entanto é supreendente que os quartos privados sejam mais caros em média que as casas inteiras e os quartos de hotel. Isto pode ser porque a diferença entre ”quarto privado”e ”quarto de hotel”pode ser confusa, tanto que os hoteis de alto preço notados em cima estão caracterizados como ”quartos privados”. Isto pode explicar também o grande desvio padrão das variáveis.

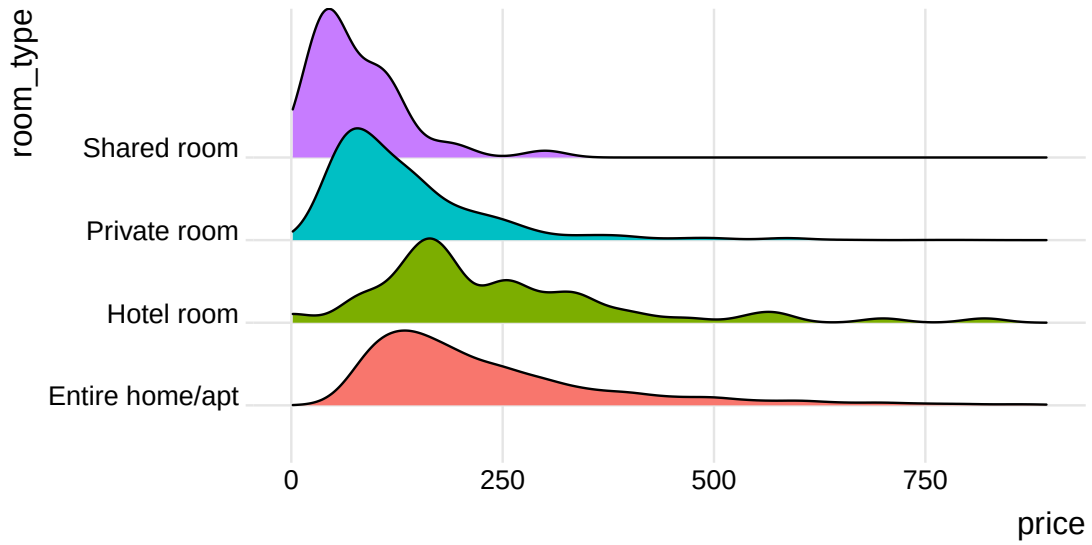
```
df %>%
    group_by(room_type) %>%
    filter(price < upper_limit, price > 0) %>%
    summarise(n = n(), freq = n()/nrow(.),
        averagePrice = mean(price), sd = sd(price),
        min = min(price), max = max(price)) %>%
    showT()
```

room_type	n	freq	averagePrice	sd	min	max
Entire home/apt	4131	0.6391769	233.9990	147.74938	33	880
Hotel room	61	0.0094383	248.9180	151.65776	72	820
Private room	2191	0.3390067	135.2793	95.85747	24	800
Shared room	80	0.0123782	78.3125	55.54006	25	300

Se excluirmos os apartamentos de luxo, conseguimos observar valores mais esperados; quartos de hoteis serem os mais caros com preços aproximados aos das casas inteiras, e os alojamentos partilhados serem os mais baratos. Os preços médios dos quartos privados desceram significativamente. Isto pode ser explicado pela classificação de alojamentos de luxo como ”quartos privados”. Deste modo suspeitamos que, sem os quartos de luxo, esta nova categoria identifica-se mais com albergues.

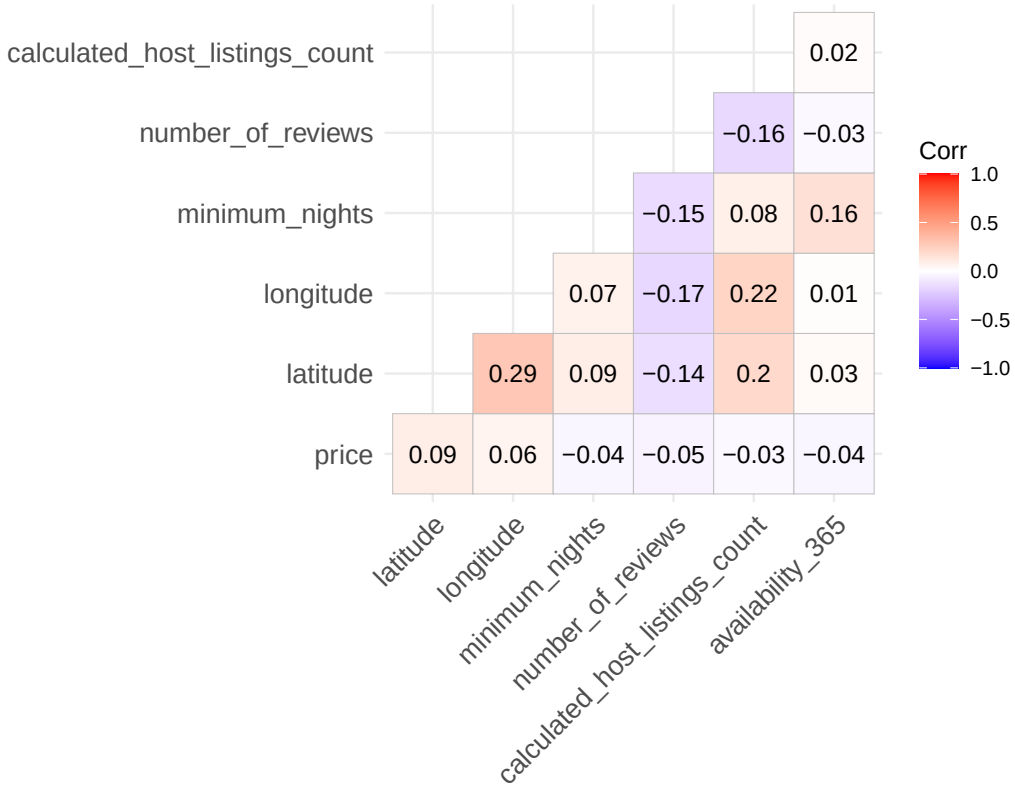
```r
ggplot(df, aes(x = price, y = room_type,
    fill = room_type)) + geom_density_ridges() +
    xlim(0, upper_limit) + theme_ridges() +
    theme(legend.position = "none")

## Picking joint bandwidth of 24.6
```



O diagrama apresentado demonstra as distribuições dos preços por tipo de alojamento. Este corrobora que os quartos de hóteis são os mais caros, mas parece demonstrar também que a maioria dos outros grupos se encontram com preços semelhantes, com uma diferença na densidade na cauda do gráfico, explicando assim a alta média de apartamentos inteiros. desta forma, a maioria dos alojamentos inteiros estão de acordo com quartos privados, mas existem mais alojamentos inteiros mais caros.

```r
df %>%
    select(price, latitude, longitude, minimum_nights,
        number_of_reviews, calculated_host_listings_count,
        availability_365) %>%
    cor(use = "complete.obs") %>%
    ggcorrplot(lab = TRUE, type = "lower")
```



O gráfico de correlações não mostra também nenhuma correlação significante entre o preço e as outras variáveis numéricas analisaveis. Mostra também uma pouca correlação entre as variáveis independentes.

```
recipeUpSample <- recipe(price ~ neighbourhood +
    latitude + longitude + room_type + minimum_nights +
    number_of_reviews + calculated_host_listings_count +
    availability_365, data = df) %>%
    step_upsample(room_type, skip = FALSE,
        over_ratio = 0.363) %>%
    step_dummy(neighbourhood, room_type) %>%
    prep()
recipeNoUpSample <- recipe(price ~ neighbourhood +
    latitude + longitude + room_type + minimum_nights +
    number_of_reviews + calculated_host_listings_count +
    availability_365, data = df) %>%
    step_dummy(neighbourhood, room_type) %>%
    prep()
dfAll <- bake(recipeUpSample, df %>%
    filter(price > 0))
dfAll %>%
    group_by(isHotel = room_type_Hotel.room,
        isPrivateRoom = room_type_Private.room,
        isSharedRoom = room_type_Shared.room) %>%
    summarise(n = n(), freq = n/nrow(dfAll),
        averagePrice = mean(price)) %>%
```

```
    showT()
```

isHotel	isPrivateRoom	isSharedRoom	n	freq	averagePrice
0	0	0	4243	0.4437820	271.0295
0	0	1	1540	0.1610710	212.9039
0	1	0	2238	0.2340759	391.4486
1	0	0	1540	0.1610710	274.1903

A quantidade de hotéis e alojoamentos partilhados foram aumentada para 1540 (ainda inclui desequilíbrio, mas devido à realmente pequena quantidade destes decidimos não equilibrar demasiado), e as nossas variáveis categóricas foram transformadas em *dummies*, de forma a tornar possível a sua análise.

Para analisar e comparar modelos, vai ser usado *in-sample* e analisado as estatísticas do coeficiente de determinação R^2, e o *MAPE*. A estatística F vai ser insignificante pare o nosso modelo, devido à grande quantidade de observações, levando a uma rejeição da hipótese nula constante. Para comparar modelos, vai ser observado o *AIC* de cada, quando apropriado. Para verificar os pressupostos, vai ser analisado se a média dos resíduos padrozinados é praticamente zero, o teste de *Breusch-Pagan* para verificar a homocedasticidade, o teste de *Breusch-Godfrey* para verificar a autocorrelação, e o teste de *Jarque-Bera* para verificar a normalidade dos resíduos. Em alguns modelos vai ser também analisado o gráficos de resíduos.

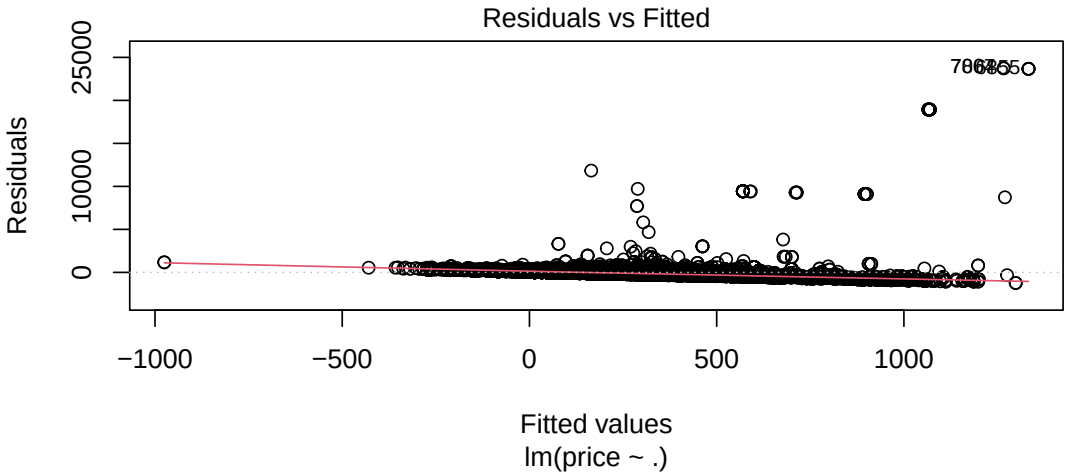
Em cada um destes testes, se o valor p for menor que 0.05, então rejeitamos H_0. No teste *Breusch-Pagan*, H_0 é que os resíduos são homocedásticos, e no teste *Breusch-Godfrey*, H_0 é que os resíduos não são autocorrelacionados. No teste *Jarque-Bera*, H_0 é que os resíduos são normais.

O nosso primeiro modelo linear foi um modelo que diretamente relaciona as variáveis todas com o preço.

```
fit0 <- lm(price ~ ., dfAll)
fit0 %>%
    glance() %>%
    select(R2 = r.squared, AIC = AIC) %>%
    mutate(MAPE = MAPE(dfAll$price, predict(fit0,
        dfAll)), `Breusch-Pagan` = validP(bptest(fit0)$p.value,
        FALSE), `Breusch-Godfrey` = validP(bgtest(fit0)$p.value,
        FALSE), `Jarque-Bera` = validP(jarque.bera.test(fit0$residuals)$p.value,
        TRUE)) %>%
    t() %>%
    showT()
```

R2	0.04582814
AIC	161374.8
MAPE	3.782943
Breusch-Pagan	p = 1.57675882618306e-50
Breusch-Godfrey	p > 0.05
Jarque-Bera	p < 0.05

```r
plot(fit0, 1)
```



Referências

[1] `tinyurl.com/DataDictAirbnb`. Accessed: 24/11/2022.

[2] About Inside Airbnb. `insideairbnb.com/about.html`. Acessado: 24/11/2022.