UNIVERSIDADE FEDERAL DE GOIÁS – UFG

GABRIELA ALVES DE ALMEIDA

RELATÓRIO DA ANÁLISE DOS DADOS ABERTOS DO AIRBNB:

Com os dados da cidade do Rio de Janeiro

SÃO PAULO

2020

UNIVERSIDADE FEDERAL DE GOIÁS – UFG

GABRIELA ALVES DE ALMEIDA

RELATÓRIO DA ANÁLISE DOS DADOS ABERTOS DO AIRBNB:

Com os dados da cidade do Rio de Janeiro

Relatório da análise de dados para a conclusão do Curso de Extensão em *Marketing Analytics* apresentado à Universidade Federal de Goiás – UFG.

Orientadores: Prof. ° Thiago Marques e Prof. Marcos Severo.

SÃO PAULO

2020

SUMÁRIO

1 INTRODUÇÃO

Este relatório tem como objetivo a análise de dados provenientes do *Airbnb*, com dados centrados na cidade do Rio de Janeiro – RJ.

A questão a ser analisada tem como variável de resposta o preço, e para entendê-lo serão analisadas as seguintes variáveis: tipo de quarto, números de *reviews,* bairro e mínimo de noites.

A justificativa para a escolha dessa base de dados se dá em razão do interesse particular da autora em análise de imóveis e por serem a única base de dados com dados brasileiros entre as opções, fazendo com que assim possa garantir uma maior proximidade com o pesquisador e esta base de dados.

2 SOBRE OS DADOS

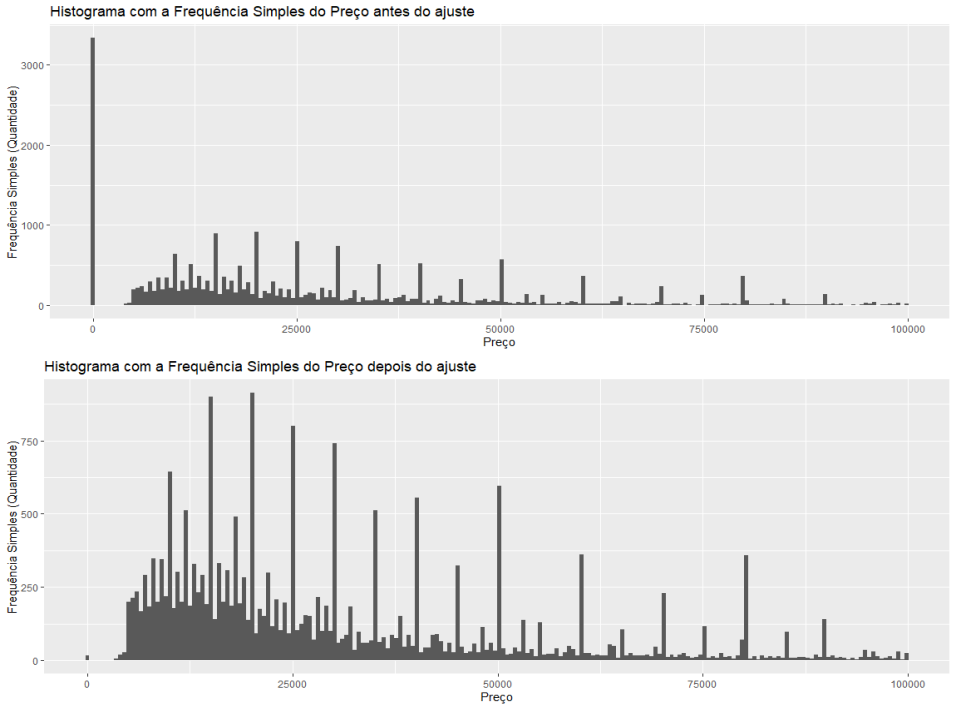
O *dataset* original possui 24.681 *rows* e 74 variáveis, e pode ser encontrado em http://insideairbnb.com/get-the-data.html e conta com os dados da cidade do Rio de Janeiro. Para fim deste relatório algumas variáveis foram removidas, deixando assim o *dataset* modificado com 35 variáveis. E sendo utilizando somente as variáveis citadas na aula 11 do curso de *Marketing Analytics*.

2.1 Análise das variáveis

2.1.1 Preço

A variável resposta deste relatório é a variável preço (*price)* que é classificada como quantitativa contínua. Esta variável possuía vários registros com valores de R$ 1,00, R$ 2,00, tracei então uma linha de corte mínimo em R$ 100,00, isso comprometeu aproximadamente 13,46% dos registros. O corte em R$ 50,00 comprometeria aproximadamente 13,44%, e em R$ 10,00, comprometeria aproximadamente 13,11%. Como a distância entre o corte de R$ 50,00 e de R$ 100,00 é de aproximadamente 0,02 pontos percentuais acho válido manter tal corte, para que não comprometa esta análise, portanto seguiremos utilizando esse preço modificado, no gráfico 1 pode-se observar a diferença nos histogramas.

Gráfico 1 – Comparativo dos histogramas de preço antes e depois do ajuste



Fonte: Elaborado pela autora, 2021.

Ainda sobre o preço, este possui assimetria positiva e é uma curva leptocúrtica, que significa que seus dados estão bem concentrados (no código temos as funções do pacote E1071 com o resultado de assimetria e curtose). Por fim, resumindo as informações de tendência central e dispersão temos:

Tabela 1 – Resumo das estatísticas descritivas da variável preço ajustado.

Mínimo	**1° Quartil**	**Mediana**	**3° Quartil**	**Média**	**Máximo**	**Desvio padrão**	**Coeficiente de Variância**
112,2	14200,0	23500,00	40000,00	29966,2	99900,0	21302,00	71,1%

Fonte: Elaborado pela autora, 2021.

1.1.2 Localização

A variável localização é classificada como qualitativa nominal e esta variável possui 151 bairros diferentes e com isso acaba dificultando o trabalho de análise desta

categoria. Como localização é um elemento altamente importante não podemos deixa-lo de lado e com isso, para fins de análise, foi criada a variável bairros_selecionados, que conta com uma seleção de bairros mais atrativos e turísticos do Rio de Janeiro (Leblon, Ipanema, Lagoa, Gávea, Jardim Botânico, Recreio dos Bandeirantes, Copacabana, Freguesia (Jacarepaguá), Tijuca, Leme, Santa Tereza, Centro, Camorim, Catete, Maracanã e Urca) . Para a seleção destes bairros foram levados os seguintes critérios: sites de turismo, o bairro Centro (que de modo geral, por ser movimentado e próximo a vários lugares é interessante levar em consideração), lugares reconhecidos por atrações turísticas (arquitetura, estádios de futebol, paisagens naturais como cachoeiras e praias).

Esta variável criada é do tipo lógica e devolve *True* se for um bairro selecionado e *False* do contrário, quanto a curtose em ambos os casos esta foi maior que 3, e portanto estas duas distribuições se caracterizam como leptocúrtica, significando que os valores estão com algum grau de concentração e portanto não variam tanto. Quanto a assimetria em ambos os casos ela foi positiva significando que a mediana é menor que a média, e que a maioria dos valores se concentram a esquerda da distribuição, significando uma concentração maior nos valores menores.

Tabela 2 – Resumo das estatísticas descritivas da variável preço ajustado agrupada por bairro selecionado

Bairro Selecionado	Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio padrão	Coeficiente de Variância
Sim	112	15000	24000	40000	30298	99000	20908	69%
Não	116	12000	22000	40000	29172	99000	22199	76,1%

Fonte: Elaborado pela autora, 2021.

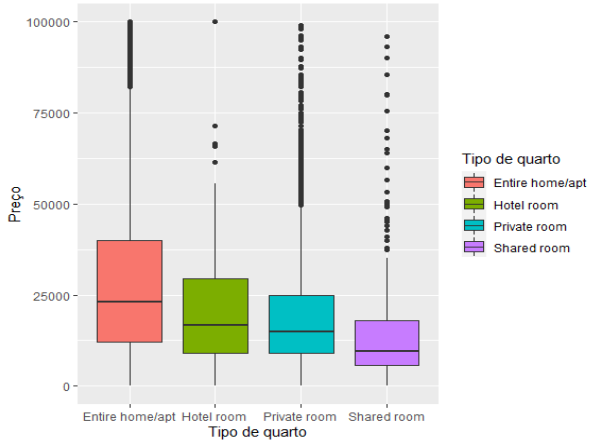
A partir da Tabela 2 pode-se observar que não há uma diferença tão significativa entre as estatísticas de tendência central, mais adiante veremos se essa variável será ou não significativa no modelo de regressão, então por enquanto ela será mantida. Vale ressaltar que tentei diminuir a quantidade de bairros e selecionar apenas os “mais relevantes”, mas as medidas geradas foram muito similares com os

bairros mencionados acima, inclusive no coeficiente de variação houve uma diferença de mais ou menos 1 ponto percentual.

1.1.3 Tipo de quarto

A variável tipo de quarto (room_type) é uma variável qualitativa nominal formada por 4 categorias: casa/apartamento inteiro, quarto de hotel, quarto privado e quarto compartilhado. A seguir podemos observar o seu comportamento.

Gráfico 2 – Boxplot de preço diferenciado por tipo de quarto



Fonte: Elaborado pela autora, 2021.

Segundo o Gráfico 2 a variável preço se comporta de forma crescente em relação à privacidade, quanto maior a privacidade maior o preço que se tende a pagar pelo quarto, ainda que haja *outliers* em todos os tipos de quarto, isto pode ser causado talvez devido a localização onde o diferencial deixa de ser a privacidade e passar ser a localização.

Tabela 3 – Resumo das estatísticas descritivas da variável preço ajustado agrupado por tipo de quarto.

Room Type	Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio Padrão	Coeficiente de Variância
Entire home/apt	112	17700	28000	46400	34301	99900	21739	63,4%
Hotel room	3800	10850	18400	29750	22514	99900	16543	73,5%

Private room	126	9500	15000	25700	20927	99900	16819	80,4%
Shared room	612	6000	10000	18750	16085	95900	16502	103%

Fonte: Elaborado pela autora, 2021.

Em complemento ao Gráfico 2, a tabela apresenta o resumo das estatísticas, corroborando com a hipótese da individualidade e aumento de preço, mas ainda assim há detalhes a serem levados em conta, como os valores mínimos, que mostram que a casa inteira tem um valor menor que um quarto compartilhado. No entanto tanto os quartis, como a média e a mediana colaboram com isto.

Quanto a curtose em todos os casos esta é maior que três, e, portanto, leptocúrtica. Quanto a assimetria, em todos os casos a mesma é positiva, significando concentração dos dados a esquerda da distribuição.

Considerando a variabilidade dos quartos, todas parecem variar bastante visto que os coeficientes de variância são maiores que 25%, o tipo de quarto com preços mais variados é o quarto compartilhado, enquanto a casa inteira ou apartamento parece ser a com menor variabilidade, o Gráfico 2 nos ajuda a visualizar melhor esta variação.

Vale mencionar como hipótese que os valores máximos são iguais em todas as modalidades de quartos, imagino que por uma questão de filtros, visto que se pode filtrar preços tanto do mais barato para o mais caro quanto do mais caro para o mais barato, isto ajudaria também a explicar o valor mínimo baixo de uma casa inteira. Quanto a essa hipótese não vejo uma maneira de comprová-la, mas gostaria de mencioná-la neste momento.

2.1.4 Número de *reviews* por mês

Reviews per month é uma variável quantitativa discreta e se refere ao número de avaliações que um imóvel recebe por mês. Esta variável precisou ser levemente transformada pois quando não havia avaliações a variável apresentava NA como valor, logo os NAs foram substituídos pelo número 0.

Tabela 4 – Medidas resumos da variável número de reviews por mês

Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio padrão	Coeficiente de Variância
0	0	11	44	39.63	922	71,3	180%

Fonte: Elaborado pela autora, 2021.

Esta variável apresenta assimetria positiva e sua forma se dá sendo leptocúrtica, significando que os valores se concentram a esquerda da distribuição e há pouca variabilidade entre eles. A partir da Tabela 4 podemos afirmar que pelo menos 25% da distribuição tem valor 0, isso significando nenhuma avaliação por mês, em contrapartida o desvio padrão e o coeficiente de variância apontam que há muita variação entre os dados dado uma amplitude de 922.

Tabela 5 – Medidas resumos da variável número de *reviews* por mês agrupada por tipo de quarto.

Room Type	Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio Padrão	Coeficiente de Variância
Entire home/apt	0	0	14	60	46,8	922	76.6	164%
Hotel room	0	8,5	36	82	54	317	61	113%
Private room	0	0	6	21	25,1	707	56.9	226%
Shared room	0	0	0	9	7,34	238	18.7	255%

Fonte: Elaborado pela autora, 2021.

Ao comparar o número de *reviews* por mês com os diferentes tipos de quartos a partir da Tabela 5, podemos visualizar que de forma geral, quartos compartilhados têm menos avaliações (pelo menos 50% da amostra não possui avaliações), e conforme aumenta a individualidade a quantidade de avaliações aumentam, os quartos compartilhados também detêm a maior variabilidade entre os tipos de quartos alugados, isso pode ser indicativo de que as pessoas que vão para o Rio de Janeiro estão interessadas em maior privacidade, logo acabam optando por quartos maiores ou até mesmo casas/apartamentos para alugar.

Tabela 6 – Medidas resumos da variável número de reviews por mês agrupada por bairro selecionado.

Bairro Selecionado	Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio padrão	Coeficiente de Variância
Sim	0	0	12	52	43,6	922	74,2	170
Não	0	0	7	28	30,1	852	63,1	209

Fonte: Elaborado pela autora, 2021.

Segundo a Tabela 6, parece haver diferença entre a seleção de bairros em relação ao número de *reviews*, sendo que aqueles que fazem parte da seleção apresentam números maiores. Visto que esta seleção foi baseada em pontos turísticos e famosos faz sentido que os valores sejam maiores para estes bairros, isto pode corroborar com a variável ser interessante para explorar, mas veremos isso de maneira mais detalhada ao traçarmos um modelo de regressão.

2.1.5 Mínimo de noites

A variável mínimo de noites é classificada como uma quantitativa discreta e diz respeito ao número mínimo de noites que uma acomodação tem para poder ser alugada no *Airbnb.*

Tabela 7 – Medidas resumo da variável mínimo de noites.

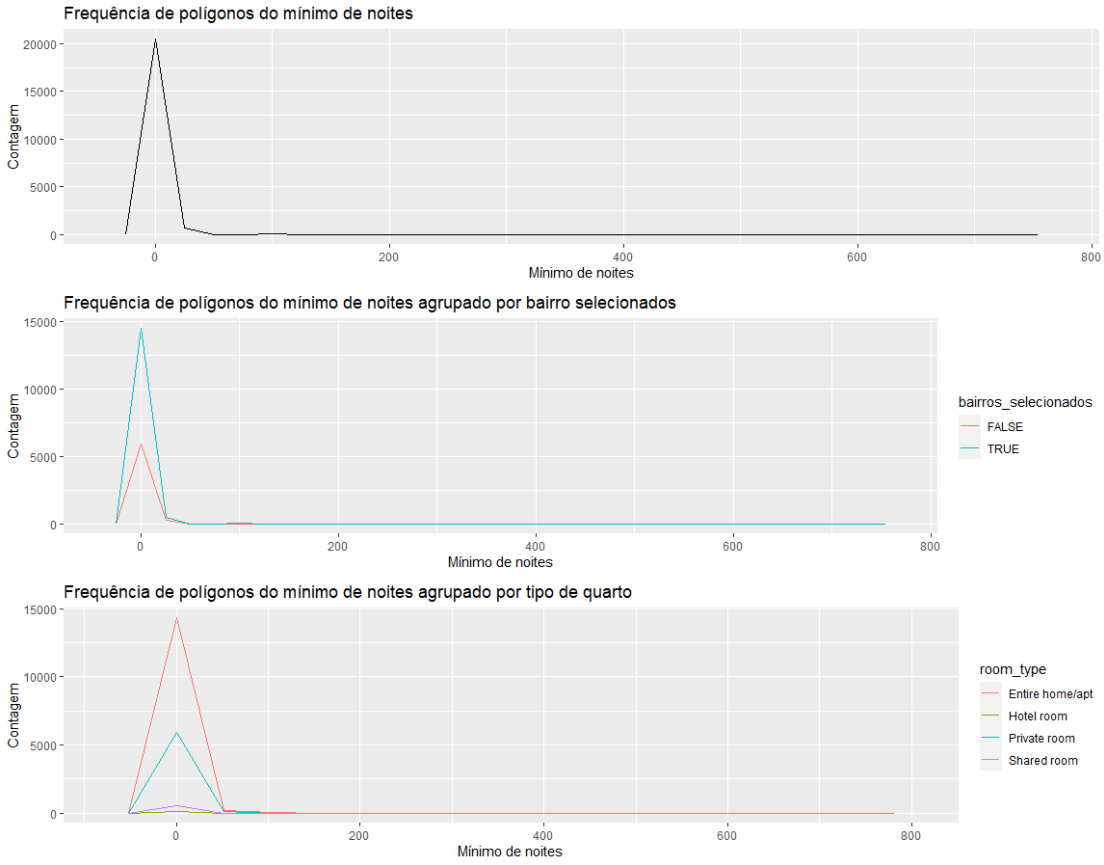
Mínimo	1° Quartil	Mediana	3° Quartil	Média	Máximo	Desvio padrão	Coeficiente de Variância
1	1	2	4	4,485	730	17,4	388%

Fonte: Elaborado pela autora, 2021.

A partir da Tabela 7 pode-se observar que esta variável possui uma grande variabilidade entre os valores, sendo o mínimo de noites da variável igual a 1 nos primeiros 25% da distribuição, mínimo de 2 noites na metade inicial da distribuição e nos primeiros 75% 4 noites. A distorção começa a ser notada quando observado o número máximo de noites, que é de 730 destoando totalmente dos quartis e da média, podemos confirmar isso com o coeficiente de variância que é de 388%. A assimetria desta distribuição é positiva, significando que a concentração de valores está a

esquerda da distribuição, quanto a curtose podemos classificar como leptocúrtica, reafirmando a concentração de valores nos valores iniciais.

Gráfico 3 – Frequência de polígonos da variável mínimo de noites, mínimo de noites agrupado por bairros selecionados e mínimo de noites por tipo de quarto.



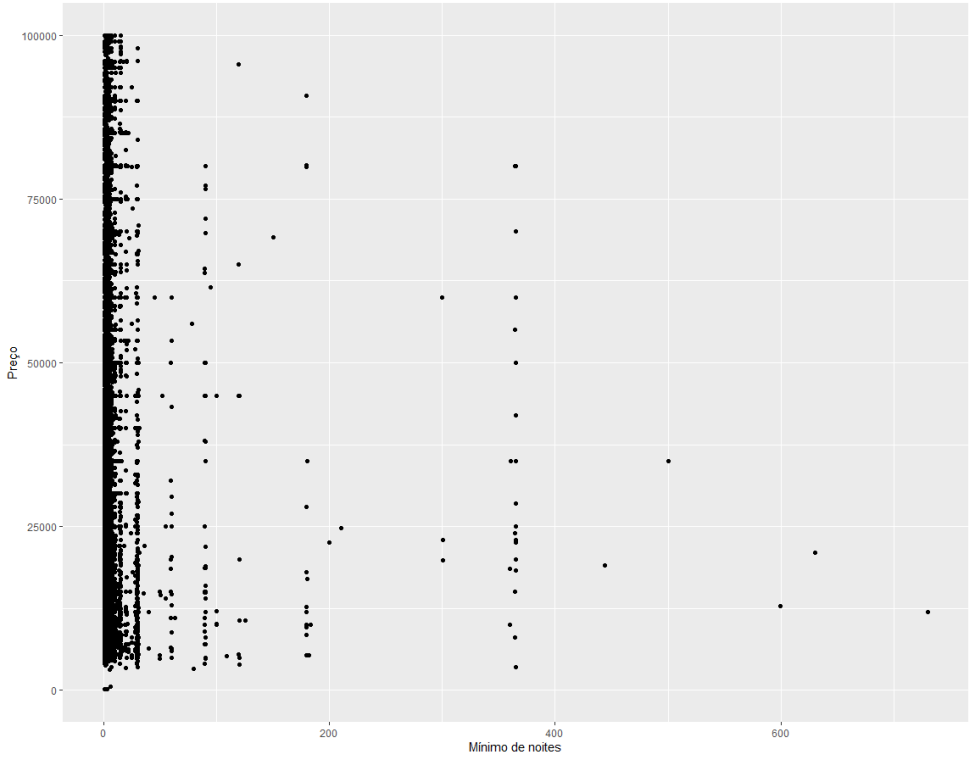
Fonte: Elaborado pela autora, 2021.

O Gráfico 3 mostra primeiramente a frequência de polígonos da variável mínimo de noites sem o agrupamento de outra variável, mostrando o comportamento assimétrico e a concentração de valores no início da distribuição.

Ao agrupar por bairros selecionados pode-se observar que para os bairros selecionados são aqueles com mais *reviews*, isto pode ter sido causado devido a própria formação desta variável que conta com bairros turísticos e de alto padrão, trazendo assim mais interesse para eles, causando um maior número de *reviews* por consequência.

Quanto ao tipo de quarto, a categoria com maior volume de noites mínimas é a casa/apartamento inteiro, seguido de quartos privados, quartos compartilhados e por fim quartos de hotel. Quartos de hotel acabam por não ter (ou quase não ter) mínimo de noites devido ao tipo de negócio (hotel), enquanto este se torna a exceção, o restante parece obedecer a ideia de individualidade inicialmente abordada, conforme aumenta a individualidade, anteriormente o preço tendia a aumentar, mas parece que o requerimento de noites parece aumentar também.

Gráfico 4 – Gráfico de dispersão do preço contra o mínimo de noites.



Fonte: Elaborado pela autora, 2021.

Ao observar o Gráfico 4, podemos analisar graficamente que a variável preço parece não ser muito afetada pelo mínimo de noites, não é possível determinar graficamente qual o tipo de correlação entre estas duas variáveis, então é necessário calculá-la no R, ao fazer isso temos o valor aproximado de 0.0157, o que significa uma correlação positiva fraca. Ainda que esta variável pareça ter pouco impacto em breve veremos como ela se comporta em um modelo de regressão simples e também em um modelo múltiplo.

2.1.6 Disponibilidade em 365 dias

A variável disponibilidade em 365 mostra a disponibilidade de uma acomodação em 365 dias, esta variável tem como característica ser quantitativa discreta. E em comparação com todas as variáveis já mencionadas até então, esta é a única (até o momento) em que sua assimetria é negativa, significando uma concentração maior de valores a direita de sua distribuição, e quanto à curtose também é a única que é platocúrtica, o que significa que seus valores estão bem espalhados. Este comportamento pode ser explicado devido a natureza do negócio envolvendo aluguel de acomodações, pois é possível definir datas específicas em que as acomodações não estão disponíveis, como grandes feriados, e a pessoa dona da acomodação possa querer fazer uso da mesma e ao mesmo tempo vão ter pessoas que vão deixar as acomodações 100% disponíveis, fazendo assim com que a disponibilidade oscile bastante.

Tabela 8 – Medidas resumo da variável disponibilidade em 365 dias.

Mínimo	**1° Quartil**	**Mediana**	**3° Quartil**	**Média**	**Máximo**	**Desvio padrão**	**Coeficiente de Variância**
0	69	183	358	202,6	365	142	70,3%

Fonte: Elaborado pela autora, 2021.

Como vimos pela curtose os valores desta distribuição estão bem espalhados, podemos confirmar isto ao observar a Tabela 8, o valor mínimo da distribuição é 0, mas no primeiro quartil já podemos ver uma grande diferença de 69, a partir da mediana podemos afirmar que metade da distribuição está acima de 183 e abaixo de 183. A variância da distribuição também não é grande comparando as variáveis já analisadas.

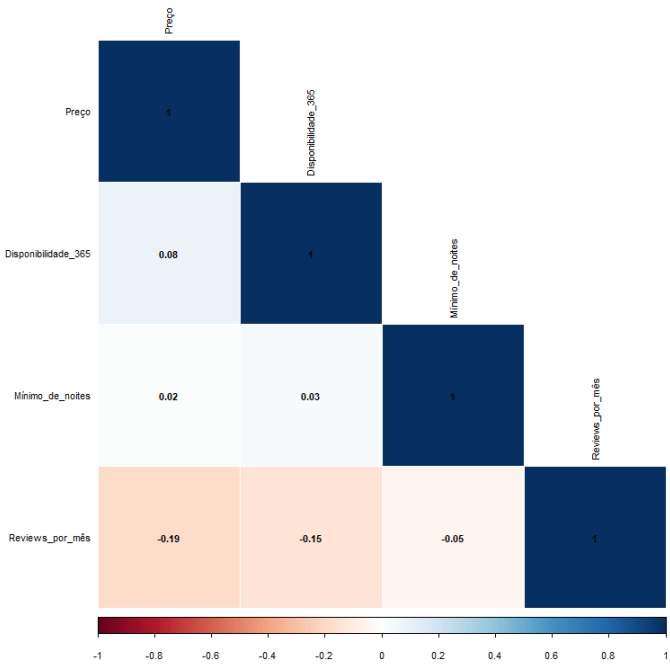
2.2 Correlações

Já vimos o relacionamento dessas variáveis entre elas através de gráficos de dispersão ao longo do relatório, agora podemos ver todas as variáveis contra elas mesmas por meio do cálculo de correlação. A partir do Gráfico 5 podemos ver um

gráfico com as correlações entre três variáveis quantitativas deste relatório: preço, que é a nossa variável resposta, disponibilidade em 365 dias e mínimo de noites.

As correlações das variáveis disponibilidade em 365 dias e mínimo de noites são modo geral levemente positivas, isso significa que quando uma variável se desloca, a outra se descola no mesmo sentido, porém de maneira bem menor. Exemplo: quando a disponibilidade de noites aumenta, o preço tende a aumentar, mas em uma proporção bem menor. Enquanto a maior correlação se dá pela variável *reviews* por mês, porém a mesma é negativa (e ainda leve), isso significa que quanto mais *reviews* por mês o preço tende a ter um comportamento negativo.

Gráfico 5 – Conjunto de correlações envolvendo as variáveis preço, disponibilidade em 365 dias e mínimo de noites.



Fonte: Elaborado pela autora, 2021.

3 MODELOS DE REGRESSÃO LINEARES SIMPLES

A partir deste momento neste relatório, confrontaremos nossa variável dependente (preço) com cada uma das seguintes variáveis independentes: tipo de quarto, localização (a partir da variável criada bairros selecionados), número de reviews e disponibilidade em 365 dias.

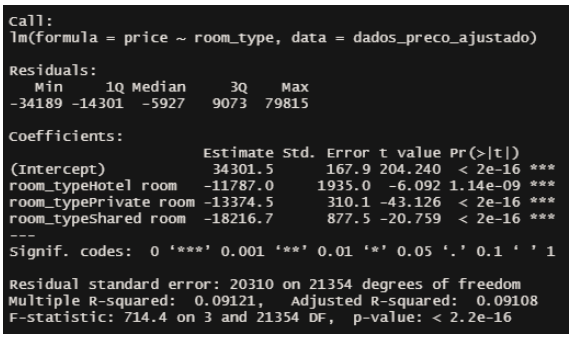
O primeiro método para realizar esta análise é por regressão linear simples de cada variável, com o objetivo de testar a significância de cada variável e assim observar se tais variáveis independentes possuem poder de explicação no preço das locações do *Airbnb* no Rio de Janeiro.

3.1 Tipo de quarto

A partir da Figura 1 podemos observar os resultados da regressão linear simples do preço explicado pelo tipo de quarto. Avaliando a significância do parâmetro tipo de quarto podemos ver que são altamente significativos (p-value < 0.05, inclusive este parâmetro continua sendo significativo a 99% de Nível de Confiança), logo não se aceita a hipótese de que o parâmetro é igual a zero, portanto ele é estatisticamente diferente de zero.

Interpretando os coeficientes temos: quando o tipo de quarto é casa/apartamento inteiro o preço estimado é de R$ 34.301,50, quando é o tipo é hotel o preço estimado cai para R$ 22.514,50, quando é quarto privado cai para R$ 20.927, e por fim quando é um quarto compartilhado o preço cai para R$ 16.084,80. Estes resultados corroboram com a noção mais cedo apresentada que conforme há um aumento na privacidade o preço tende a aumentar, porém embora os coeficientes apresentem alta significância o poder de explicação do modelo não é grande, aproximadamente 9,1%, sendo assim será necessário adicionar mais variáveis para que este modelo fique mais robusto.

Figura 1 – Modelo de regressão linear simples preço vs. tipo de quarto.



```
Call:
lm(formula = price ~ room_type, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-34189 -14301  -5927   9073  79815

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          34301.5      167.9 204.240  < 2e-16 ***
room_typeHotel room  -11787.0     1935.0  -6.092 1.14e-09 ***
room_typePrivate room -13374.5      310.1 -43.126  < 2e-16 ***
room_typeShared room  -18216.7      877.5 -20.759  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20310 on 21354 degrees of freedom
Multiple R-squared:  0.09121,   Adjusted R-squared:  0.09108
F-statistic: 714.4 on 3 and 21354 DF,  p-value: < 2.2e-16
```

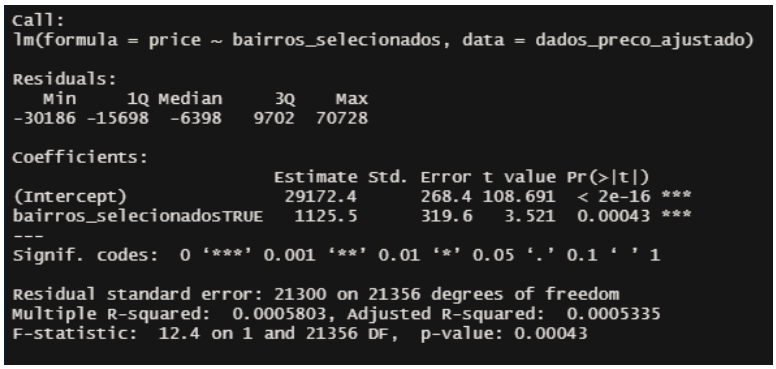
Fonte: Elaborado pela autora, 2021.

3.2 Bairros selecionados

A Figura 2 traz os resultados do modelo de regressão linear simples onde o preço foi a variável dependente e a variável criada anteriormente, bairros selecionados, a variável independente do modelo. Lembrando que esta variável foi criada com base em bairros famosos e turísticos do Rio de Janeiro, e a partir dos resultados da regressão podemos ver que este parâmetro é altamente significativo, inclusive a 99% de Nível de Confiança.

Interpretando os coeficientes temos que quando o bairro não faz parte da seleção que foi feita, o preço estimado é de R$ 29.172,40, e quando o bairro faz parte dessa seleção o preço estimado sobe para R$ 30.297,90. Assim, podemos concluir com este modelo que existe uma relação entre o preço e os bairros selecionados, como estes foram escolhidos com base na relevância turística faz sentido que nesses bairros o preço estimado seja maior, no entanto este modelo possui um baíssimo nível de explicação, significando que este modelo sozinho não consegue explicar o preço.

Figura 2 – Modelo de regressão linear simples preço vs. bairros selecionados



```
Call:
lm(formula = price ~ bairros_selecionados, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-30186 -15698  -6398   9702  70728

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             29172.4      268.4 108.691  < 2e-16 ***
bairros_selecionadosTRUE   1125.5      319.6   3.521  0.00043 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21300 on 21356 degrees of freedom
Multiple R-squared:  0.0005803, Adjusted R-squared:  0.0005335
F-statistic:  12.4 on 1 and 21356 DF,  p-value: 0.00043
```

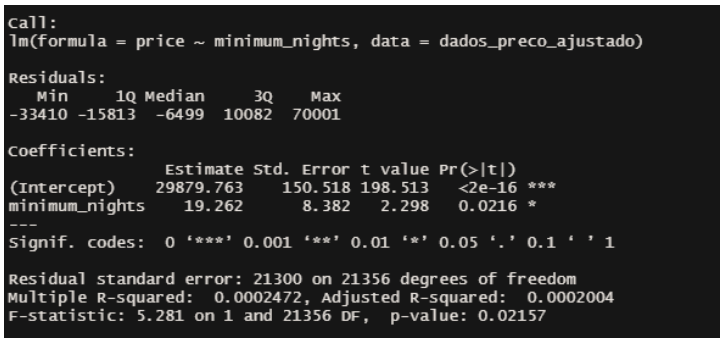
Fonte: Elaborado pela autora, 2021.

3.3 Mínimo de noites

O terceiro modelo nos traz a variável dependente preço e a variável independente mínimo de noites. Começando com a avaliação do modelo temos que este parâmetro não é estatisticamente significativo a 95% de nível de confiança, sendo assim a 95% de nível de confiança podemos dizer que este parâmetro é estatisticamente igual a zero, no entanto ele é significativo a 90% e por isso manteremos essa variável quando formos inserir no modelo de regressão linear múltipla, ainda mais considerando o grau de explicação deste modelo que é bem baixo (0%), logo este modelo acaba não tendo muita utilidade.

Quanto a interpretação temos que o beta é positivo, significando que o adicional de um dia modifica positivamente o preço. Quando o mínimo de noites é zero, o preço estimado é aproximadamente R$ 29.879,76, e para cada noite adicional há um incremento no preço em aproximadamente R$ 19,26.

Figura 3 – Modelo de regressão linear simples preço vs. mínimo de noites



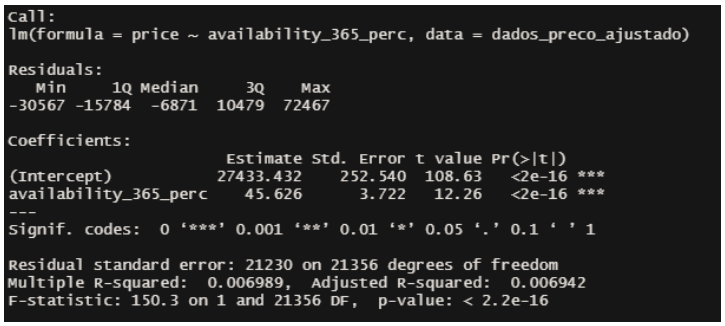
```
Call:
lm(formula = price ~ minimum_nights, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-33410 -15813  -6499  10082  70001

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    29879.763    150.518 198.513   <2e-16 ***
minimum_nights    19.262      8.382   2.298   0.0216 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21300 on 21356 degrees of freedom
Multiple R-squared:  0.0002472, Adjusted R-squared:  0.0002004
F-statistic: 5.281 on 1 and 21356 DF,  p-value: 0.02157
```

Fonte: Elaborado pela autora, 2021.

3.4 Disponibilidade em 365 dias

Para ajudar na análise do modelo, a variável disponibilidade em 365 dias foi modificada, assim ela virou uma variável em percentual pois como a variável tem um limite de 365 dias não faria muito sentido se fossemos estimar no modelo mais dias do que o limite, logo essa variável virou um percentual de ocupação em 365 dias.

Figura 4 – Modelo de regressão linear preço vs. disponibilidade em 365 dias



```
Call:
lm(formula = price ~ availability_365_perc, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-30567 -15784  -6871  10479  72467

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          27433.432    252.540  108.63   <2e-16 ***
availability_365_perc    45.626      3.722   12.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21230 on 21356 degrees of freedom
Multiple R-squared:  0.006989,  Adjusted R-squared:  0.006942
F-statistic: 150.3 on 1 and 21356 DF,  p-value: < 2.2e-16
```

Fonte: Elaborado pela autora, 2021.

Avaliando o modelo a partir da Figura 4 temos que o parâmetro é significativo até em 99% de Nível de Confiança, sendo assim, esta variável é estatisticamente diferente de zero até a 99% de Nível de Confiança. No entanto o poder explicativo deste modelo é baixo (aproximadamente 0%), sendo assim esse modelo não é suficiente para explicar o comportamento do preço baseado na disponibilidade em 365 dias.

Interpretando os coeficientes temos este beta é positivo, logo a adição de 1% de disponibilidade faz o preço sofrer um aumento. Ainda, quando a disponibilidade for 0%, o preço estimado é de aproximadamente R$ 27.433,43, para cada 1% de aumento na variável disponibilidade em 365 dias (em porcentagem) o preço estimado aumenta em aproximadamente R$ 45,63.

3.5 Número de *reviews* por mês

Por fim, a variável número de *reviews* por mês traz um beta estatisticamente significativo, isso significa que o beta é estatisticamente diferente de zero e podemos utilizá-lo em nossa análise. Sendo assim partindo para a interpretação do modelo temos um comportamento negativo, para cada *review* o preço estimado sofre uma redução de aproximadamente R$ 56,45. Quanto ao poder de explicação deste modelo temos aproximadamente 3,5%, sendo assim é interessante incorporar tal variável no modelo completo em vez de usá-la sozinha na tentativa de explicar o preço.

Figura 5 – Modelo de regressão linear simples preço *vs*. *reviews* por mês

```
Call:
lm(formula = price ~ reviews_per_month, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-32077 -15508  -6072   9812  74143

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       32203.075    163.735  196.68   <2e-16 ***
reviews_per_month   -56.449      2.006  -28.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20920 on 21356 degrees of freedom
Multiple R-squared:  0.03574,   Adjusted R-squared:  0.0357
F-statistic: 791.6 on 1 and 21356 DF,  p-value: < 2.2e-16
```

Fonte: Elaborado pela autora, 2021.

4 MODELOS DE REGRESSÃO LINEAR MÚLTIPLA

Agora que vimos os modelos de regressão linear simples para cada variável independente podemos avaliar que em todos os modelos as variáveis sozinhas não eram capazes de ter um grau de explicação grande, o maior foi de aproximadamente 9%. Como essas variáveis sozinhas não conseguiram ter um bom grau de explicação da variável dependente preço, cabe agora testar essas variáveis em conjunto e ver se o modelo de regressão linear múltiplo é significativo e possui um maior grau de capacidade de explicação do preço na base de dados do *Airbnb* no Rio de Janeiro.

Figura 6 – Modelo linear múltiplo com todas as variáveis independentes

```
Call:
lm(formula = price ~ room_type + bairros_selecionados + availability_365_perc +
    minimum_nights + reviews_per_month, data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-37926 -13748  -5152   9236  82585

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              35439.143    355.287  99.748  < 2e-16 ***
room_typeHotel room      -11171.297   1867.229  -5.983 2.23e-09 ***
room_typePrivate room    -15261.151    305.770 -49.910  < 2e-16 ***
room_typeShared room     -21692.628    853.410 -25.419  < 2e-16 ***
bairros_selecionadosTRUE   -359.246    298.869  -1.202   0.2294
availability_365_perc        45.418      3.486  13.030  < 2e-16 ***
minimum_nights              -14.340      7.730  -1.855   0.0636 .
reviews_per_month           -68.586      1.925 -35.636  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19590 on 21350 degrees of freedom
Multiple R-squared:  0.1546,    Adjusted R-squared:  0.1543
F-statistic: 557.8 on 7 and 21350 DF,  p-value: < 2.2e-16
```

Fonte: Elaborado pela autora, 2021.

A partir da Figura 6 podemos avaliar o resultado do modelo de regressão múltipla que tem como variável dependente o preço. Começando pelo teste F, podemos avaliar que o modelo como um todo faz sentido pois o p-value é bem pequeno (p-value < 0.05), portanto rejeita-se a hipótese de que todos os betas estimados são iguais a zero.

Ao testar a significância de cada beta podemos observar que com exceção da variável mínimo de noites e bairros selecionados, todas as outras variáveis são significativas a um nível de confiança de 95%. Analisando o R-ajustado podemos ver que mesmo com mais variáveis este modelo completo também é fraco na capacidade de explicação da variável preço.

Como temos betas que não são significativos a 95% de nível de confiança, podemos dizer que eles são estatisticamente iguais a zero, removeremos estes betas da regressão linear múltipla, e em seguida veremos que a variável mínimo de noites

também afeta os cálculos estimados, pois todos os valores foram distorcidos e ficaram muito próximos a zero, tornando dificultosa e talvez ineficaz a interpretação dos betas do modelo completo, portanto não a faremos.

Seguindo com o modelo de regressão linear múltipla “reduzido”, na Figura 6, composto com as seguintes variáveis dependentes: tipo de quarto, localização, disponibilidade e número de *reviews* por mês, temos um modelo com pelo menos um beta diferente de zero, segundo o teste F, demonstrando que o modelo como um todo faz sentido.

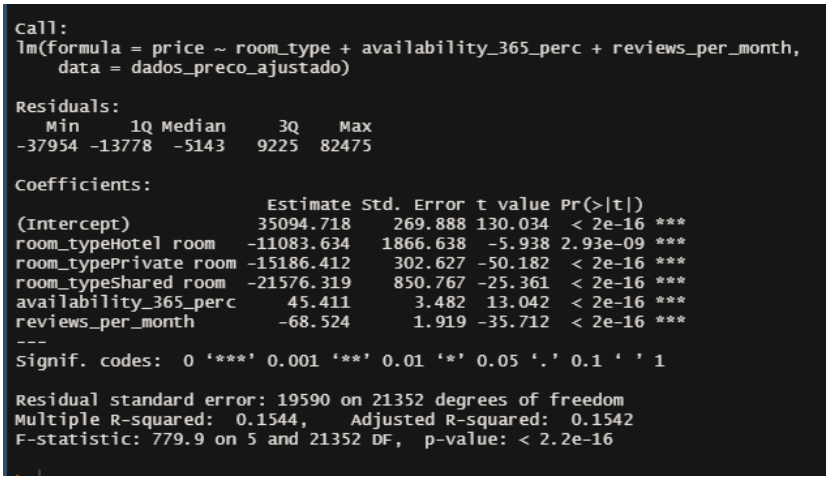
Quanto a significância de cada beta, temos que a um nível de confiança de 95% todas as variáveis são significativas, isto significa que estes betas são estatisticamente diferentes de zero e permite que tais variáveis permaneçam no modelo.

A partir da Figura 7 temos a interpretação de tipo de quartos. Mantendo as demais variáveis constantes em relação ao preço estimado, em relação ao tipo de quarto, temos: o preço estimado de uma casa ou apartamento no Rio de Janeiro é de aproximadamente R$ 35.094,72, quando é um hotel o preço sofre uma redução para aproximadamente R$ 24.011,09, quando é um quarto privado o preço sofre uma redução para aproximadamente R$ 19.908,31, e quando é um quarto compartilhado o preço cai para aproximadamente R$ 13.518,40. A variável tipo de quarto impacta a variável dependente preço de maneira negativa conforme a privacidade diminuiu.

Considerando a variável tipo de quarto sendo apartamento/casa, a variável disponibilidade em 365 (em porcentagem) tem um comportamento positivo, conforme aumenta em 1% a disponibilidade, o preço estimado aumenta em aproximadamente R$ 45,41.

E por fim, a variável *reviews* por mês, considerando a variável tipo de quarto sendo apartamento e casa, esta variável tem um comportamento negativo, para cada *reviews* por mês o preço estimado cai em aproximadamente R$ 68,52.

Figura 7 – Modelo de regressão linear múltipla reduzido.



```
Call:
lm(formula = price ~ room_type + availability_365_perc + reviews_per_month,
    data = dados_preco_ajustado)

Residuals:
   Min     1Q Median     3Q    Max
-37954 -13778  -5143   9225  82475

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             35094.718    269.888 130.034  < 2e-16 ***
room_typeHotel room    -11083.634   1866.638  -5.938 2.93e-09 ***
room_typePrivate room  -15186.412    302.627 -50.182  < 2e-16 ***
room_typeShared room   -21576.319    850.767 -25.361  < 2e-16 ***
availability_365_perc      45.411      3.482  13.042  < 2e-16 ***
reviews_per_month         -68.524      1.919 -35.712  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19590 on 21352 degrees of freedom
Multiple R-squared:  0.1544,    Adjusted R-squared:  0.1542
F-statistic: 779.9 on 5 and 21352 DF,  p-value: < 2.2e-16
```

Fonte: Elaborado pela autora, 2021. 1

Em relação ao modelo completo, este é melhor devido a todos os betas serem significativos, porém devido a exclusão de variáveis o modelo acaba perdendo um pouco de poder de explicação, logo o modelo completo tem mais poder do que este.

5 CONSIDERAÇÕES FINAIS

Este relatório teve como objetivo entender as variáveis independentes que possam explicar a variável dependente preço na base de dados do *Airbnb*, que consiste no aluguel de imóveis no Rio de Janeiro.

As variáveis independentes analisadas nesse relatório foram: tipo de quarto, localização, quantidade de *reviews* por mês, disponibilidade em 365 dias e mínimo de noites. Ao analisar a correlação entre as variáveis quantitativas vimos que as correlações são positivas com as variáveis disponibilidade em 365 dias e mínimo de noites, enquanto a variável *reviews* por mês é negativa, no entanto todas não são tão fortes. O que podemos concluir é que quando há aumentos nas variáveis disponibilidade e mínimo o preço tende a sofrer acréscimos, enquanto quando há um aumento em *reviews* o preço sofre decréscimos, e esta variação negativa tem maior magnitude do que a positiva.

Enquanto os modelos de regressão linear simples mostraram o relacionamento das variáveis separados em relação a variável preço e a significância dos betas separados, com isso vimos que a variável tipo de quarto se comporta de forma negativa conforme a privacidade diminui, isso quer dizer que o preço estimado é maior quando se aluga um hotel em relação a um quarto particular ou um quarto compartilhado. Vimos também que o comportamento da variável bairros selecionados, que foi criada para esta análise, tem um comportamento positivo quando o bairro em questão faz parte da seleção que foi feita anteriormente. Tanto as variáveis mínimo de noites quanto a variável disponibilidade afetam o preço estimado de maneira positiva, no entanto o mínimo de noites não é estatisticamente significativo a 95% de nível de confiança. Por fim, a variável *reviews* por mês assim como foi visto na correlação traz um comportamento negativo para a variável preço, fazendo com que quando a variável *reviews* aumenta, o preço diminui.

Os modelos de regressão linear múltipla mostraram o relacionamento de todas variáveis juntas, fazendo com que pudéssemos observar mudanças dessas variáveis na relação com o preço, como por exemplo a variável bairros selecionados, que quando verdadeira no modelo simples tinha um impacto positivo no preço e quando levada em consideração com outras variáveis o seu comportamento muda. Ainda

que tenha mudado de comportamento, como esta variável não era significativa estatisticamente a 95% de nível de confiança, juntamente com a variável mínimo de noites, estas foram removidas e um novo modelo foi criado. Neste modelo reduzido houve uma queda de poder explicativo do modelo em relação ao antigo de 0.01 pontos percentuais, fazendo com que este modelo novo seja ainda mais interessante, ainda mais que ele consegue fazer sentido como um todo e todos os seus betas são significativos.

Finalmente, com este modelo de regressão linear múltipla reduzido conseguimos verificar que todas as variáveis mantiveram seus relacionamentos com o preço inalterados, sendo assim as variáveis independentes que explicam a variável dependente preço apresentam o seguinte comportamento. Quanto ao tipo de quarto: quanto maior a privacidade maior tende a ser o preço estimado, quanto a disponibilidade: quanto maior a disponibilidade maior tende ser o preço estimado e por fim, quanto a *reviews*: quanto mais *reviews* menor o preço.

REFERÊNCIAS

BRADLEY, Tyler. **Calculating quantiles for groups with dplyr::summarize and purrr::partial**. Disponível em: https://tbradley1013.github.io/2018/10/01/calculating-quantiles-for-groups-with-dplyr-summarize-and-purrr-partial/. Acesso em: 25 jan. 2021.

DINHANI, Renato. **How do I replace NA values with zeros in an R dataframe?** Disponível em: https://stackoverflow.com/questions/8161836/how-do-i-replace-na-values-with-zeros-in-an-r-dataframe. Acesso em: 25 jan. 2021.

ECONÔMICO, Revista Capital. **Melhor lugar para se morar no Rio de Janeiro**. Disponível em: https://revistacapitaleconomico.com.br/melhor-lugar-para-se-morar-no-rio-de-janeiro/. Acesso em: 25 jan. 2021.

INCORPORADORA, Tegra. **Curiosidades sobre os 5 bairros nobres do Rio de Janeiro.** Disponível em: https://www.tegraincorporadora.com.br/blog/mercado/bairros-nobres-do-rio-de-janeiro/. Acesso em: 25 jan. 2021.

LOFT. **Conheça 13 bairros nobres do RJ e o que tem de mais legal em cada um**. Disponível em: https://blog.loft.com.br/bairros-nobres-do-rj/. Acesso em: 25 jan. 2021.

MORE: Mecanismo online para referências, versão 2.0. Florianópolis: UFSC Rexlab, 2013. Disponível em: http://www.more.ufsc.br/. Acesso em: 25 jan 2021.

MOVINGBLOG. **Veja os 8 melhores de bairros para se morar no Rio de Janeiro**. Disponível em: https://blog.movingimoveis.com.br/veja-os-8-melhores-de-bairros-para-se-morar-no-rio-de-janeiro/. Acesso em: 25 jan. 2021.

RIO, Where In. **OS BAIRROS MAIS EXCLUSIVOS DO RIO DE JANEIRO.** Disponível em: https://www.whereinrio.com/pt/blog/os-bairros-mais-exclusivos-do-rio-de-janeiro. Acesso em: 25 jan. 2021.