1) Análise exploratória dos dados:

Utilizando a linguagem R, abri os dados e me deparei com uma base de 16 variáveis e 48864 observações. Para saber como a estrutura dos dados está organizada, eu apliquei o comando *str()* que nos fornece uma visão geral concisa das classes e estruturas de dados contidas no objeto **base** criado:

```
'data.frame':     48894 obs. of  16 variables:
 $ id                            : int  2595 3647 3831 5022 5099 5121 5178
5203 5238 5295 ...
 $ nome                          : chr  "Skylit Midtown Castle" "THE VILLAGE
OF HARLEM....NEW YORK !" "Cozy Entire Floor of Brownstone" "Entire Apt:
Spacious Studio/Loft by central park" ...
 $ host_id                       : int  2845 4632 4869 7192 7322 7356 8967
7490 7549 7702 ...
 $ host_name                     : chr  "Jennifer" "Elisabeth" "LisaRoxanne"
"Laura" ...
 $ bairro_group                  : chr  "Manhattan" "Manhattan" "Brooklyn"
"Manhattan" ...
 $ bairro                        : chr  "Midtown" "Harlem" "Clinton Hill"
"East Harlem" ...
 $ latitude                      : num  40.8 40.8 40.7 40.8 40.7 ...
 $ longitude                     : num  -74 -73.9 -74 -73.9 -74 ...
 $ room_type                     : chr  "Entire home/apt" "Private room"
"Entire home/apt" "Entire home/apt" ...
 $ price                         : int  225 150 89 80 200 60 79 79 150 135 ...
 $ minimo_noites                 : int  1 3 1 10 3 45 2 2 1 5 ...
 $ numero_de_reviews             : int  45 0 270 9 74 49 430 118 160 53 ...
 $ ultima_review                 : chr  "2019-05-21" "" "2019-07-05"
"2018-11-19" ...
 $ reviews_por_mes               : num  0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99
1.33 0.43 ...
 $ calculado_host_listings_count: int  2 1 1 1 1 1 1 1 4 1 ...
 $ disponibilidade_365           : int  355 365 194 0 129 0 220 0 188 6 ...
```

Para facilitar futuras análises, preciso transformar alguns dados antes de prosseguir. Dentre eles:

1. Transformar as variáveis categóricas em factor: **bairro, bairro_group** e **room_type**
2. Transformar a variável **reviews_por_mês** no tipo date.

```
'data.frame':     48894 obs. of  16 variables:
 $ id                            : int  2595 3647 3831 5022 5099 5121 5178
5203 5238 5295 ...
 $ nome                          : chr  "Skylit Midtown Castle" "THE VILLAGE
OF HARLEM....NEW YORK !" "Cozy Entire Floor of Brownstone" "Entire Apt:
Spacious Studio/Loft by central park" ...
```

```
 $ host_id                       : int  2845 4632 4869 7192 7322 7356 8967
7490 7549 7702 ...
 $ host_name                     : chr  "Jennifer" "Elisabeth" "LisaRoxanne"
"Laura" ...
 $ bairro_group                  : Factor w/ 5 levels "Bronx","Brooklyn",..: 3
3 2 3 3 2 3 3 3 3 ...
 $ bairro                        : Factor w/ 221 levels "Allerton","Arden
Heights",..: 128 95 42 62 138 14 96 203 36 203 ...
 $ latitude                      : num  40.8 40.8 40.7 40.8 40.7 ...
 $ longitude                     : num  -74 -73.9 -74 -73.9 -74 ...
 $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 1
2 1 1 1 2 2 2 1 1 ...
 $ price                         : int  225 150 89 80 200 60 79 79 150 135 ...
 $ minimo_noites                 : int  1 3 1 10 3 45 2 2 1 5 ...
 $ numero_de_reviews             : int  45 0 270 9 74 49 430 118 160 53 ...
 $ ultima_review                 : Date, format: "2019-05-21" NA "2019-07-05"
"2018-11-19" ...
 $ reviews_por_mes               : num  0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99
1.33 0.43 ...
 $ calculado_host_listings_count: int  2 1 1 1 1 1 1 4 1 ...
 $ disponibilidade_365           : int  355 365 194 0 129 0 220 0 188 6 ...
```

Por último, é necessário saber se existe um anúncio repetido para o mesmo imóvel. Para isso, filtrei as observações que contam com a mesma latitude, longitude, bairro e área. Assim, saberemos, pelas coordenadas, se há mais de um anúncio para o mesmo imóvel. Porém, após fazer isso, **percebi que não é possível saber se o anúncio se trata do mesmo imóvel. Pois, filtrando as observações restringidas pelas variáveis supracitadas, o resultado são 23 imóveis repetidos 47 vezes no total (como consta a imagem abaixo). Mas os resultados poderiam estar indicando apartamentos diferentes em um mesmo edifício. Para saber se há repetição de imóveis, seria necessário uma variável que identificasse o apartamento/casa especificamente. Então, a partir de agora, devemos supor que cada anúncio se trata da locação de um imóvel único.**

	bairro	bairro_group	latitude	longitude	n
1	Bedford-Stuyvesant	Brooklyn	40.67825	-73.92346	2
2	Bedford-Stuyvesant	Brooklyn	40.68398	-73.94101	2
3	Chelsea	Manhattan	40.74913	-73.99575	2
4	East Village	Manhattan	40.72145	-73.97881	2
5	East Village	Manhattan	40.72504	-73.98327	2
6	Hell's Kitchen	Manhattan	40.75584	-73.99559	2
7	Hell's Kitchen	Manhattan	40.75888	-73.99077	2
8	Hell's Kitchen	Manhattan	40.76914	-73.98757	2
9	Midtown	Manhattan	40.75368	-73.97358	2
10	Midtown	Manhattan	40.75414	-73.96595	2
11	Murray Hill	Manhattan	40.74882	-73.97788	2
12	Nolita	Manhattan	40.72347	-73.99302	2
13	Ridgewood	Queens	40.70125	-73.91051	2
14	SoHo	Manhattan	40.72607	-74.00166	2
15	SoHo	Manhattan	40.72741	-74.00178	2
16	Theater District	Manhattan	40.76122	-73.98583	2
17	Upper East Side	Manhattan	40.76989	-73.94961	2
18	Upper West Side	Manhattan	40.77874	-73.98437	2
19	Williamsburg	Brooklyn	40.70818	-73.94952	2
20	Williamsburg	Brooklyn	40.71145	-73.95302	2
21	Williamsburg	Brooklyn	40.71232	-73.94220	3
22	Williamsburg	Brooklyn	40.71353	-73.96216	2
23	Williamsburg	Brooklyn	40.71603	-73.96417	2

Dito isso, podemos seguir com nossa análise. Primordialmente, decidi realizar uma análise de estatística sumária sobre algumas variáveis:

```
   longitude        latitude
Min.   :-74.24   Min.   :40.50
1st Qu.:-73.98   1st Qu.:40.69
Median :-73.96   Median :40.72
Mean   :-73.95   Mean   :40.73
3rd Qu.:-73.94   3rd Qu.:40.76
Max.   :-73.71   Max.   :40.91
```

Como podemos ver, tanto a latitude quanto a longitude não variam tanto, posto que os anúncios são todos dentro de Nova York.

```
          room_type
Entire home/apt:25409
Private room   :22325
Shared room    : 1160
```

No que tange ao tipo de espaço dos anúncios, temos:

1. 25409 anúncios para apartamento ou casa inteira.
2. 22325 anúncios para apenas um quarto privado.
3. 1160 anúncios para a locação de quarto compartilhado.

```
     price
Min.   :    0.0
1st Qu.:   69.0
Median :  106.0
Mean   :  152.7
3rd Qu.:  175.0
Max.   :10000.0
```

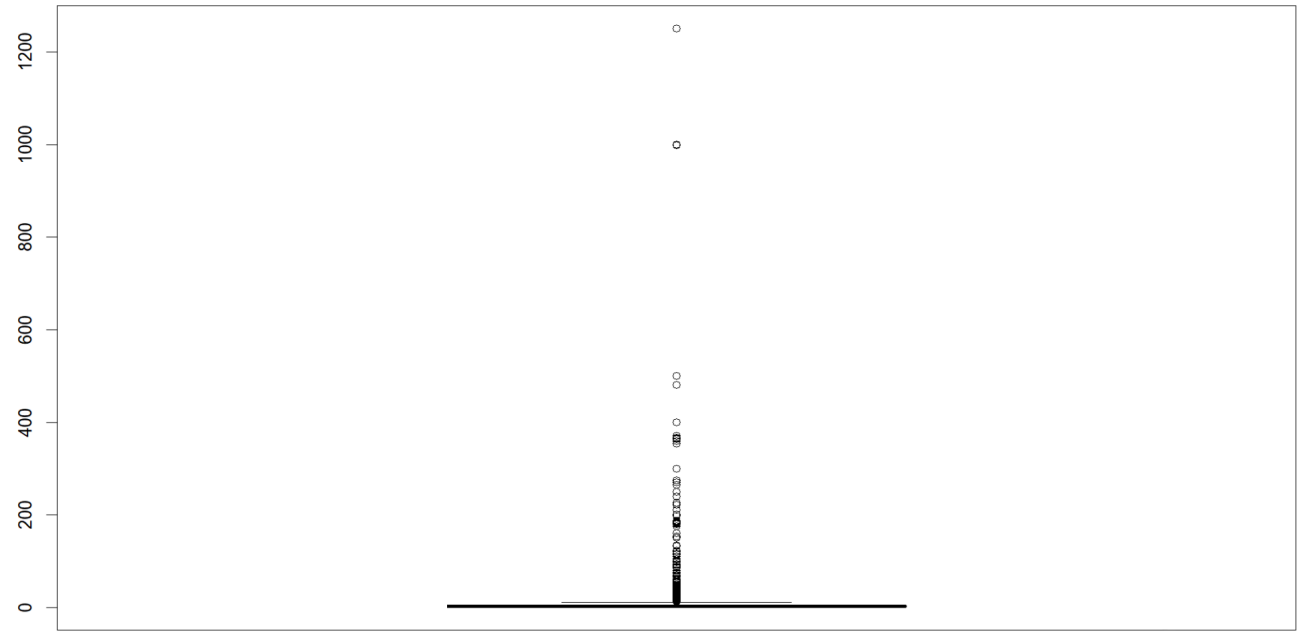
Sobre o preço cobrado por noite hospedada, notamos que o mínimo cobrado é zero dólares (provavelmente o anunciante não dispôs o preço na publicação), o primeiro quadrante (ou 25%) temos valores até 69 dólares por noite. A mediana (ou 50%) apresenta valores até 106 dólares, enquanto que 75% dos valores estão entre 0 e 175 dólares por noite. É interessante notar, que 25% dos valores estão entre 175 e 10.000 dólares por noite, provavelmente explicado pela minoria dos locais de luxo. **No preço, é relevante verificarmos algumas medidas de dispersão:**

1. Seu desvio padrão é de 240 dólares. **(Em média, o preço varia 240 dólares ao redor de sua média - essa medida está um pouco alta devido aos prováveis hotéis de luxo que cobram caro e pode atrapalhar nas futuras análises estatísticas)**
2. Sua variância é de 57675 dólares. **(240 elevado ao quadrado)**

```
minimo_noites
Min.   :   1.00
1st Qu.:   1.00
Median :   3.00
Mean   :   7.03
3rd Qu.:   5.00
Max.   :1250.00
```

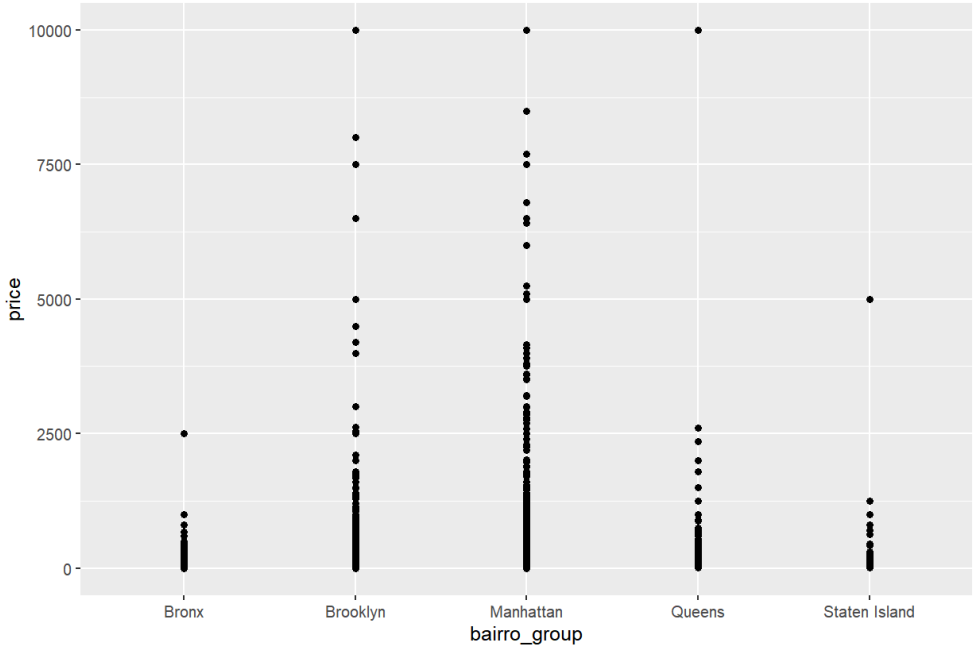
Já na variável que representa a quantidade mínima de noites hospedadas, 75% dos anúncios exigem, pelo menos, 5 noites hospedadas. **Evidenciado no gráfico abaixo, são poucos valores acima**

disso. E as observações mais extremas podem ser tratadas como outliers.



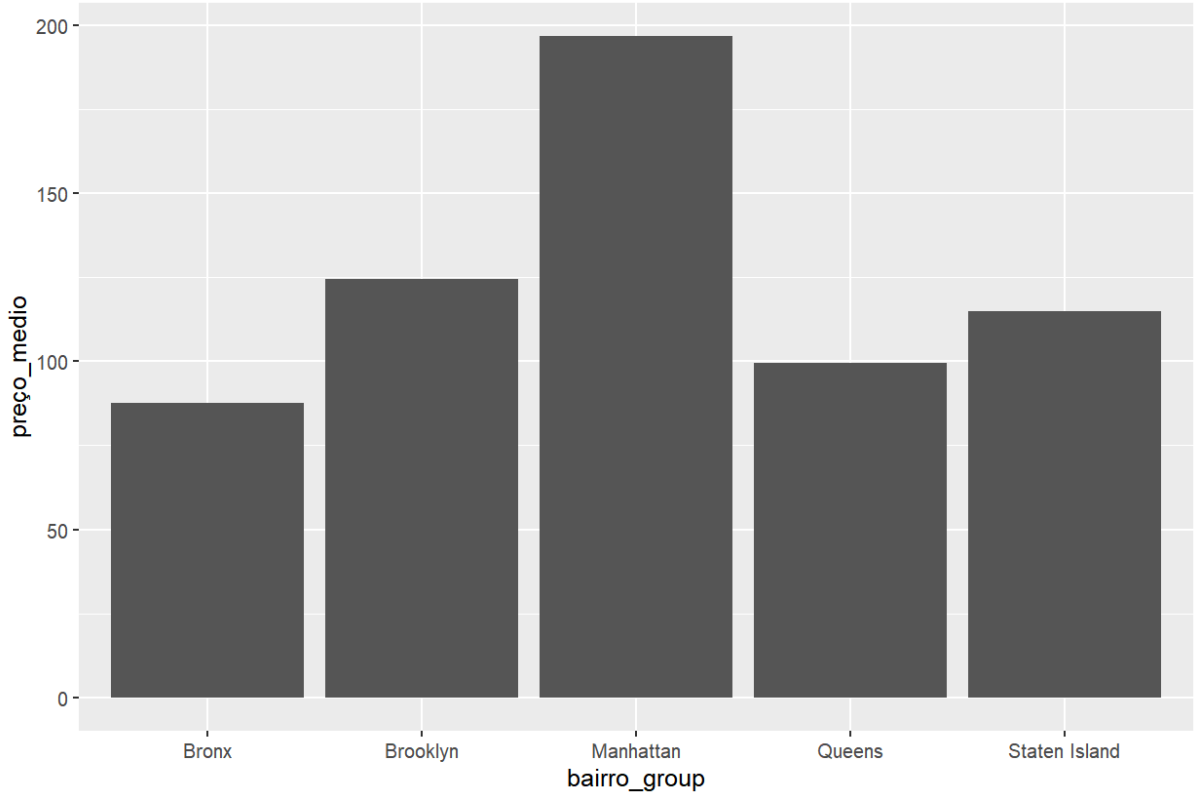
Respondendo a pergunta simples: qual bairro é mais caro?

Para responder essa pergunta, eu plotei 2 gráficos que nos mostram quais bairros são mais caros e quais são baratos.



Esse gráfico simples de ponto evidencia que o bairro Manhattan tem mais anúncios localizados ao longo da distribuição de preços. Portanto,

Manhattan é o mais caro. Em segundo lugar temos o **Brooklyn.**
Claro que essa é uma análise visual. Contudo, plotei um gráfico que mostra a média dos preços em cada bairro para termos a confirmação:



Com esse gráfico concluímos que **Manhattan é o mais caro!**

Por último, realizei uma regressão para saber **o quanto o bairro influencia na formação dos preços:**

```
Call:
lm(formula = price ~ bairro_group, data = base)

Residuals:
   Min     1Q Median     3Q    Max
-196.9  -74.4  -36.9   22.5 9900.5

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 87.497      7.168  12.207  < 2e-16 ***
bairro_groupBrooklyn        36.885      7.360   5.012 5.41e-07 ***
bairro_groupManhattan      109.379      7.346  14.890  < 2e-16 ***
bairro_groupQueens          12.021      7.827   1.536   0.1246
bairro_groupStaten Island   27.316     14.200   1.924   0.0544 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 236.8 on 48889 degrees of freedom
Multiple R-squared:  0.02822,   Adjusted R-squared:  0.02815
F-statistic:   355 on 4 and 48889 DF,  p-value: < 2.2e-16
```

Nos resultados da regressão podemos tirar algumas conclusões:

1. **Por estar situado em Manhattan, há uma valorização, em média, de 109 dólares por noite na hospedagem.**
2. **Se o imóvel estiver no Brooklyn, há um acréscimo de 36 dólares.**
3. **O valor de locação médio que independe da locação é 87 dólares (intercepto).**

Essas conclusões podem ser precisas porque são **estatisticamente significativas,** ao passo que foi testada com um número infinitesimalmente pequeno.
Enquanto que os coeficientes nas regiões do Queens e Staten Island foram testadas a um nível superior a 5%.
Outra conclusão que pode ser retirada a princípio, é o indicador do R-ajustado, **que nos mostra o quanto da variável independente explica a dependente (no nosso caso, o quanto o bairro explica o preço),** evidencia um percentual de 2% **- um número muito baixo! Isso indica que há muitos outros fatores que explicam o preço e que não estão sendo contabilizados em nosso cálculo.**

Agora, cabe analisar a veracidade do nosso modelo: **Será que podemos confiar nesses resultados?**

Primeiro irei analisar a normalidade nos resíduos. É imprescindível que a distribuição dos resíduos do nosso modelo seja normal, caso contrário, não é possível confiar nos testes de hipóteses aplicados no modelo.

```
        Jarque Bera Test

data:  res1
X-squared = 771529509, df = 2, p-value < 2.2e-16
```

Usando o teste de Jarque Bera, conclui-se que o resíduo está longe de ser normal, pois o p-valor está muito abaixo de 0,05 **- que é o limite mínimo de confiança para aceitarmos a hipótese nula de normalidade dos resíduos.**

Há diversos métodos na literatura de como corrigir a normalidade. Um deles é a exclusão de outliers do data-frame. **Recordando de análises anteriores, a variável que mede o preço tem alguns outliers - que são as locações com preços exorbitantes. Para uma análise de regressão, que trabalha com médias, é necessário excluir esses outliers para que nosso modelo tenha mais robustez.**

Modelo sem outliers:

```
Call:
lm(formula = price ~ bairro_group, data = base2)

Residuals:
    Min      1Q  Median      3Q     Max
-174.86  -61.52  -24.86   30.90  743.48

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   84.444      3.102  27.222  < 2e-16 ***
bairro_groupBrooklyn          32.079      3.185  10.071  < 2e-16 ***
bairro_groupManhattan         90.416      3.180  28.432  < 2e-16 ***
bairro_groupQueens             9.661      3.388   2.852  0.00435 **
bairro_groupStaten Island     11.705      6.160   1.900  0.05741 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102.4 on 48501 degrees of freedom
Multiple R-squared:  0.09429,   Adjusted R-squared:  0.09421
F-statistic:  1262 on 4 and 48501 DF,  p-value: < 2.2e-16
```

O nosso modelo sem outliers teve melhores resultados!!!

1. **Agora o coeficiente que mede o acréscimo do preço por ser situado no Queens é estatisticamente significativo a 0,1%!!**
2. **Notamos também que o imóvel estar no Staten Island tem um acréscimo de 11 dólares estatisticamente significativo a 5%!**
3. **Por último, notamos que o R ajustado aumentou (embora ainda pequeno) para 9%!**

Segundo vamos analisar a homocedasticidade dos resíduos. Um dos axiomas dos modelos de regressão linear, é que o resíduo do modelo deve apresentar
variância constante, **caso contrário, nosso modelo pode ser ineficiente.**

```
        studentized Breusch-Pagan test

data:  reg2
BP = 610.97, df = 4, p-value < 2.2e-16
```

O teste de BP indica que o modelo **é heterocedástico.**

Para corrigir a heterocedasticidade, nós usaremos um método que vai levar dois coelhos numa cajadada só! Porque um terceiro axioma que o modelo

deve seguir, é o de que os resíduos não podem ser autocorrelacionados. Ou seja, **um determinado preço de um imóvel não pode depender de outro preço de outro imóvel.** Para verificar se existe autocorrelação nos resíduos, nós usamos o teste de Durbin-Watson, **mas irei pular essa parte, já que o comando para corrigir a heterocedasticidade e a autocorrelação é o mesmo. Então, corrigindo um, necessariamente corrige o outro!**

```
t test of coefficients:

                          Estimate Std. Error t value  Pr(>|t|)
(Intercept)                84.4435     2.2037 38.3186 < 2.2e-16 ***
bairro_groupBrooklyn       32.0792     2.2907 14.0043 < 2.2e-16 ***
bairro_groupManhattan      90.4162     2.3555 38.3850 < 2.2e-16 ***
bairro_groupQueens          9.6613     2.3883  4.0453 5.234e-05 ***
bairro_groupStaten Island  11.7051     4.8978  2.3899   0.01686 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Por fim, temos um modelo ainda melhor que o inicial!

1. **Agora o coeficiente de Queens sendo estatisticamente significativo a 0%**
2. **E o de Staten Island sendo significativo a 1%**

Nos resultados finais, temos:

1. Brooklyn representa um acréscimo de, em média, 32 dólares no preço de locação por noite.
2. Manhattan representa um acréscimo de, em média, 90 dólares.
3. Queens representa 9 dólares.
4. E Staten Island 11 dólares.
5. Também, vale ressaltar, o preço médio que independe da localização (podemos dizer o preço padrão mínimo médio) é 84 dólares.

2.A) onde seria mais indicada a compra?

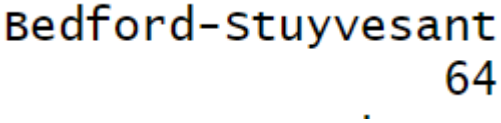
Estudando economia nós aprendemos uma coisa: **tudo depende!** Nesse caso, para a compra de um imóvel e futura locação, os critérios que devem ser analisados são: **número de avaliações mensal, preço e a localidade.**

Iremos considerar o número de avaliações mensais seguindo pela lógica de que **mais avaliações é um indicador de mais locações, por consequência, mais demanda!**
Já o preço, é indicado selecionar um imóvel em que o preço não seja exorbitante (a demanda será baixa) e nem pequeno demais (retorno baixo do investimento), **mas que esteja na média.**
No que tange a localidade, esse é **um indicador importante para a regulagem de preços, por isso deve ser levado em conta.**

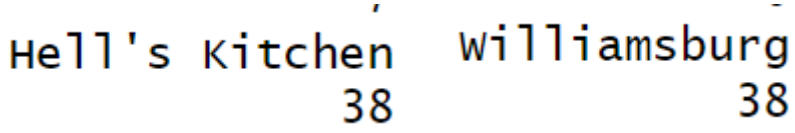
Para isso, decidi filtrar os dados que:

1. estão entre os 10% mais avaliados mensalmente
2. estejam na mediana de preço por noite (106 dólares)
3. esteja localizado nos dois bairros mais demandados/ofertados (para maior liquidez

Após aplicar todas as restrições, cheguei numa tabela com 456 observações. Para saber qual área é a mais indicada para investir, eu **filtrei qual a localização mais ofertada:**



Pelas minhas análises, e seguindo a teoria básica da economia que oferta é igual a demanda, **a localização mais indicada para investir é Bedford-Stuyvesant, no Brooklyn!!**



Em segundo lugar, e empatado com o terceiro, temos: **Hell's kitchen em Manhattan** e **Williamsburg no Brooklyn.**

2.b) O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

Para isso, devemos modelar mais uma regressão linear tendo o **preço** como **variável dependente,** e o número de noites mínimo e disponibilidade como **independentes.**

```
Call:
lm(formula = price ~ minimo_noites + disponibilidade_365, data = base2)

Residuals:
    Min      1Q  Median      3Q     Max
-159.97  -68.99  -29.21   36.54  729.66

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.281e+02  6.473e-01  197.96   <2e-16 ***
minimo_noites     3.771e-02  2.417e-02    1.56    0.119
disponibilidade_365 9.217e-02  3.735e-03   24.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 106.9 on 48503 degrees of freedom
Multiple R-squared:  0.01294,   Adjusted R-squared:  0.0129
F-statistic: 317.9 on 2 and 48503 DF,  p-value: < 2.2e-16
```

De acordo com o resultado, **apenas a disponibilidade de reserva.** Mesmo corrigindo a heterocedasticidade e a autocorrelação, os resultados não mudam. A única variável que apresentou ter **influência sobre o preço foi a disponibilidade de locação: para o acréscimo de 1 dia de locação, o preço varia 0.092168 dólares.**
O resultado tende a estar certo quando nós buscamos uma explicação racional: **Na verdade, quanto mais o mínimo de noites é necessário, mais desvalorizado a locação é.** Pois, não há necessidade de cobrar caro por **muitas** noites. **Análises futuras irão comprovar isso…**

2.c) Existem padrões no nome dos anúncios de aluguéis mais caros?

Para responder essa pergunta, eu utilizei a base filtrada com os maiores preços. Após isso verifiquei quais palavras mais repetem nos anúncios. As 10 primeiras são:

```
private     private       111
room           room       108
apartment apartment        68
bedroom     bedroom        60
studio       studio        54
brooklyn   brooklyn        53
the             the        53
cozy           cozy        49
apt             apt        45
```

A liderança fica com “private”. O que faz total sentido, pois todos preferem

alugar algo com privacidade. Por isso, é mais caro.

3) Previsão de preços

Para modelar uma regressão linear múltipla seguindo o método do Mínimos Quadrados Ordinários, eu utilizei as seguintes variáveis:

```
Call:
lm(formula = price ~ minimo_noites + disponibilidade_365 + host_id +
    bairro_group + bairro + reviews_por_mes + room_type, data = base3)

Residuals:
    Min      1Q  Median      3Q     Max
-233.79  -39.38  -11.17   17.56  744.11
```

Como resultado, eu obtive:

```
Residuals:
    Min      1Q  Median      3Q     Max
-187.89  -43.66  -12.48   19.30  765.65

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                1.235e+02  2.936e+00   42.050  < 2e-16 ***
minimo_noites             -3.533e-01  2.469e-02  -14.311  < 2e-16 ***
disponibilidade_365        1.084e-01  3.371e-03   32.171  < 2e-16 ***
host_id                    8.542e-08  5.884e-09   14.516  < 2e-16 ***
bairro_groupBrooklyn       3.207e+01  2.874e+00   11.162  < 2e-16 ***
bairro_groupManhattan      7.195e+01  2.874e+00   25.031  < 2e-16 ***
bairro_groupQueens         1.249e+01  3.037e+00    4.111 3.94e-05 ***
bairro_groupStaten Island -6.447e+00  5.415e+00   -1.190    0.234
reviews_por_mes           -2.500e+00  2.666e-01   -9.379  < 2e-16 ***
room_typePrivate room     -9.823e+01  8.612e-01 -114.060  < 2e-16 ***
room_typeShared room      -1.287e+02  2.909e+00  -44.263  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.29 on 38635 degrees of freedom
  (9860 observations deleted due to missingness)
Multiple R-squared:  0.3386,    Adjusted R-squared:  0.3385
F-statistic:  1978 on 10 and 38635 DF,  p-value: < 2.2e-16
```

Para simplificar a visualização, eu omiti o resultado das áreas, pois são mais de 200 e ficaria ruim de visualizar. Porém, para a previsão, eu vou utilizá-las.

Fazendo uma análise rápida sobre os coeficientes, verificamos que apenas uma variável não foi estatisticamente significativa. Ou seja, podemos confiar mais nos resultados. A interpretação é a seguinte: **Na coluna ‘Estimate’**

estão os valores dos coeficientes, e seus sinais, indicam se a variável em questão, influencia os preços de forma positiva ou negativa.
Por exemplo: a variável “minimo-noites” apresenta sinal negativo, ou seja, a cada noite acrescentada, o preço desvaloriza, em média, -3 dólares.

Analisando o R-ajustado:

1) **R-ajustado**

```
Residual standard error: 78.3 on 38422 degrees of freedom
  (9860 observations deleted due to missingness)
Multiple R-squared:  0.4045,    Adjusted R-squared:  0.401
F-statistic:   117 on 223 and 38422 DF,  p-value: < 2.2e-16
```

Após adicionar ao modelo a variável “bairro”, o R-ajustado nos mostrou ser de 40%. Isso é um pouco baixo.

Considerações: O modelo estimado não é o melhor para fazer a previsão de preços, deveríamos ajustar mais, como:

1. Corrigir os axiomas que foram infligidos, como: não-normalidade nos resíduos, heterocedasticidade e autocorrelação.
   Para isso, deveríamos transformar os dados em logaritmo, excluir mais outliers, etc.
2. Aplicar modelos mais robustos, como os do pacote “robuste”.

Porém, por falta de tempo, não consegui refinar mais o modelo…
Sobre a medida de performance, eu escolhi o próprio R-ajustado. **Eu tentei tirar o Erro Quadrático Médio, porém surgiram alguns problemas que seriam muito difíceis de consertar e o tempo não deixaria.**

4) Sugestão de preço:

A partir dessa nova observação, utilizando o meu modelo de previsão, o preço sugerido é **294 dólares por noite, com uma tolerância de 64 dólares para baixo e para cima.**