Mohammad Javad Ranjbar
810101173

Project

Statistical Inference, Fall 2022

Contents

Note:

- In the R code, I used a couple of libraries that are not installed by default. If you want to run the code, please check and uncomment the install.pakges(“”) in my code.
- The dataset needed some preprocessing and cleaning. which I have done In the first project phase. I will use the previous code that I implemented for cleaning this dataset in the first phase of this project. I have explained this code thoroughly in the first report. However, I will put the first phase report in this project’s file too.

Question 1:

A) I choose the review neighborhood group and instant bookable variables for this test.
The neighborhood group variable has 5 levels:

```
> levels(df$neighbourhood.group)
[1] "Bronx"          "Brooklyn"       "Manhattan"      "Queens"         "Staten Island"
```

The instant bookable variable has 2 levels:

```
> levels(df$instant_bookable)
[1] "False" "True"
```

Now we take 500 samples from our dataset:

```
> sub_df<-df[sample(nrow(df), 500), ]
> addmargins(table(sub_df$review.rate.number,sub_df$instant_bookable))

      False True Sum
  1      17   28  45
  2      65   60 125
  3      52   52 104
  4      65   52 117
  5      56   53 109
  Sum   255  245 500
```

Our sample is like the below table:

	False	True	Total
Bronx	17	28	45
Brooklyn	65	60	125
Manhattan	52	52	104
Queens	65	52	117
Staten Island	56	53	109
Total	255	245	500

Now we need to calculate the $\hat{p}$ for each level:

$$\hat{p}_{\text{Bronx}} = \frac{28}{58}, \hat{p}_{\text{Brooklyn}} = \frac{60}{125}, \hat{p}_{\text{Manhattan}} = \frac{52}{104}, \hat{p}_{\text{Queens}} = \frac{52}{117}, \hat{p}_{\text{Staten Island}} = \frac{53}{109}$$

Now for each level we check the conditions:

1. Independence:
   a. within groups: each sample (house) us independent within each group
      i. the houses are assigned randomly
      ii. The number of houses sampled are less than 10 percent of houses in NYC. n < 10% of population
   b. between groups: houses in different area are not dependent (non-paired)
2. Sample size/skew: Each sample meets the success-failure condition
   For each sample we check the above condition:

$$\hat{p}_{\text{Bronx}} * n_1 > 10, n_1(1 - \hat{p}_{\text{Bronx}}) > 10$$

$$\hat{p}_{\text{Brooklyn}} * \text{n}_2 > 10, \text{n}_2\ (1 - \hat{p}_{\text{Brooklyn}}) > 10$$
$$\hat{p}_{\text{Manhattan}} * \text{n}_3 > 10, \text{n}_3\ (1 - \hat{p}_{\text{Manhattan}}) > 10$$
$$\hat{p}_{\text{Queens}} * \text{n}_4 > 10, \text{n}_4\ (1 - \hat{p}_{\text{Queens}}) > 10$$
$$\hat{p}_{\text{Staten Island}} * \text{n}_5 > 10, \text{n}_5\ (1 - \hat{p}_{\text{Staten Island}}) > 10$$

All the above conditions are met.

Therefore, we can use confidence interval. Hypothesis would look like this:

H_0 (nothing going on): neghborhood group and instant bookablity are independent.

H_A (something going on): neghborhood group and instant bookablity are dependent.

Since we want to calculate the 95% CI $z^* = 1.96$. Now we calculate the CI for each pair:

Bronx-Brooklyn:

$\hat{p}_{\text{Bronx}} = \frac{28}{58}, \hat{p}_{\text{Brooklyn}} = \frac{60}{125}, \hat{p}_{\text{Manhattan}} = \frac{52}{104}, \hat{p}_{\text{Queens}} = \frac{52}{117}, \hat{p}_{\text{Staten Island}} = \frac{53}{109}$

$$\hat{\text{p}}_{\text{Bronx}} - \hat{\text{p}}_{\text{Brooklyn}} \pm z^{**}\text{SE} = \frac{28}{58} - \frac{60}{125} + 1.96* \sqrt{\frac{\frac{28}{58}(1\text{-}\frac{28}{58})}{58} + \frac{\frac{60}{125}(1\text{-}\frac{60}{125})}{125}} = (\text{-}0.024, 0.30)$$

$$\hat{\text{p}}_{\text{Bronx}} - \hat{p}_{\text{Manhattan}} \pm z^{**}\text{SE} = \frac{28}{58} - \frac{52}{104} + 1.96* \sqrt{\frac{\frac{28}{58}(1\text{-}\frac{28}{58})}{58} + \frac{\frac{52}{104}(1\text{-}\frac{52}{104})}{104}} = (\text{-}0.028, 0.31)$$

$$\hat{\text{p}}_{\text{Bronx}} - \hat{p}_{\text{Queens}} \pm z^{**}\text{SE} = \frac{28}{58} - \frac{52}{117} + 1.96* \sqrt{\frac{\frac{28}{58}(1\text{-}\frac{28}{58})}{58} + \frac{\frac{52}{117}(1\text{-}\frac{52}{117})}{117}} = (\text{-}0.025, 0.31)$$

$$\hat{\text{p}}_{\text{Bronx}} - \hat{p}_{\text{Staten Island}} \pm z^{**}\text{SE} = \frac{28}{58} - \frac{53}{109} + 1.96* \sqrt{\frac{\frac{28}{58}(1\text{-}\frac{28}{58})}{58} + \frac{\frac{53}{109}(1\text{-}\frac{53}{109})}{109}} = (\text{-}0.027, 0.31)$$

$$\hat{p}_{\text{Brooklyn}} - \hat{p}_{\text{Manhattan}} \pm z^{**}\text{SE} = \frac{60}{125} - \frac{52}{104} + 1.96* \sqrt{\frac{\frac{60}{125}(1\text{-}\frac{60}{125})}{125} + \frac{\frac{52}{104}(1\text{-}\frac{52}{104})}{104}} = (\text{-}0.028, 0.31)$$

$$\hat{p}_{\text{Brooklyn}} - \hat{p}_{\text{Queens}} \pm z^{**}\text{SE} = \frac{60}{125} - \frac{52}{117} + 1.96* \sqrt{\frac{\frac{60}{125}(1\text{-}\frac{60}{125})}{125} + \frac{\frac{52}{117}(1\text{-}\frac{52}{117})}{117}} = (\text{-}0.026, 0.31)$$

$$\hat{p}_{\text{Brooklyn}} - \hat{p}_{\text{Staten Island}} \pm z^{**}\text{SE} = \frac{60}{125} - \frac{53}{109} + 1.96^{*}\sqrt{\frac{\frac{60}{125}(1\text{-}\frac{60}{125})}{125} + \frac{\frac{60}{125}(1\text{-}\frac{60}{125})}{125}} = (\text{-}0.027, 0.31)$$

$$\hat{p}_{\text{Manhattan}} - \hat{p}_{\text{Queens}} \pm z^{**}\text{SE} = \frac{52}{104} - \frac{52}{117} + 1.96^{*}\sqrt{\frac{\frac{52}{104}(1\text{-}\frac{52}{104})}{104} + \frac{\frac{52}{117}(1\text{-}\frac{52}{117})}{117}} = (\text{-}0.025, 0.31)$$

$$\hat{p}_{\text{Manhattan}} - \hat{p}_{\text{Staten Island}} \pm z^{**}\text{SE} = \frac{52}{104} - \frac{53}{109} + 1.96^{*}\sqrt{\frac{\frac{52}{104}(1\text{-}\frac{52}{104})}{104} + \frac{\frac{53}{109}(1\text{-}\frac{53}{109})}{109}} = (\text{-}0.27, 0.31)$$

$$\hat{p}_{\text{Queens}} - \hat{p}_{\text{Staten Island}} \pm z^{**}\text{SE} = \frac{52}{117} - \frac{53}{109} + 1.96^{*}\sqrt{\frac{\frac{52}{117}(1\text{-}\frac{52}{117})}{117} + \frac{\frac{53}{109}(1\text{-}\frac{53}{109})}{109}} = (\text{-}0.27, 0.31)$$

Because 0 is in all the confidence intervals. It means that instance book ability is independent of the neighborhood. And we can not reject the H_0.

b)

First, we need to check for to see if the conditions of test are met:

- Independence: Sampled observations are independent.
  - The houses have been assigned randomly
  - The number of samples is less than 10% all the houses, n < 10% of population
  - Each sample contributes to only one cell.
- Sample size: Each particular scenario have at least 5 expected cases

Therefore we can use the Chi-square test:

I both calculated the test with the built in function and without the built-in function

Our original table is:

	False	True	Sum
1	1406	1336	2742
2	3439	3284	6723
3	3450	3402	6852
4	3448	3390	6838
5	3395	3407	6802
Sum	15138	14819	29957

We calculate the expected values:

```
          False      True        Sum
1     1385.599  1356.401   2742.000
2     3397.295  3325.705   6723.000
3     3462.482  3389.518   6852.000
4     3455.408  3382.592   6838.000
5     3437.216  3364.784   6802.000
Sum 15138.000 14819.000 29957.000
```

Now we can calculate the $\chi = 2.81$:

```
> X<-0
> #calculating X2
> for (i in 1:5)
+ {
+   for (j in 1:2)
+   {
+     X=X +((tbl2[i,j]-tbl3[i,j])^2)/tbl3[i,j]
+   }
+ }
> X
[1] 2.813366
```

p-value = 0.589(

```
> pchisq(X,degree_f,lower.tail = FALSE)
[1] 0.5895277
```

) which is bigger than α and therefore we can not reject the H_0. Meaning that neighborhood group and instant bootability are independent.

Also, the built in function will give the same result:

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 2.8134, df = 4, p-value = 0.5895
```

p-value = 0.589 which is bigger than α and therefore we can not reject the H_0. Meaning that neighborhood group and instant bootability are independent.

Question 2:

I choose the instant bookable variable for this test

This variable is distributed like the below table:

Instance bookability:	True	False
Population	15138	14819

In R:

```
> table(df$instant_bookable)

False  True
15138 14819
```

Now we take a sample with a population of 15 and start the test:

```
> set.seed(120)
> SP<-sample(df$host_identity_verified,15)
> table(SP)
SP
unconfirmed    verified
          7           8
```

Our sample success rate is $\frac{8}{15} = 0.53$

Hypothesis would be as followed:

$$H_0\!: p = 0.5$$

$$H_A\!: \mathrm{p} > 0.5$$

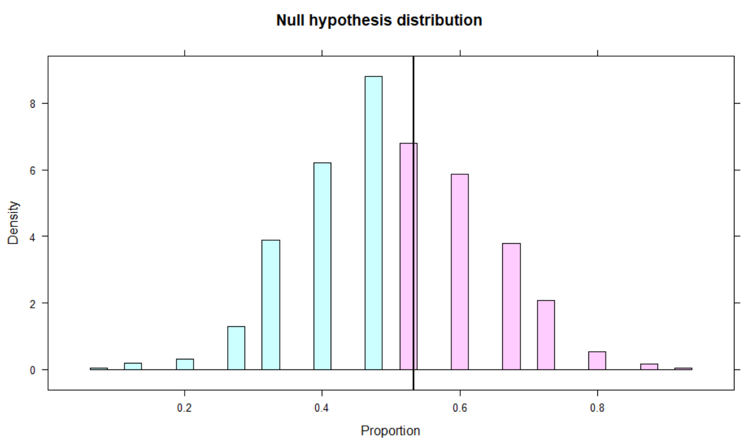
$$\mathrm{n} = 15, \hat{p} = 0.53$$

Now we need to check for the conditions:

1. Independence: we can assume the houses selected are independent.

2. Sample size / skew: 15 × 0.5 = 7.5 → not met distribution of sample proportions cannot be assumed to be nearly normal

Now we do a simulation:




```
> paste("One-sided p-value is", pvalue)
[1] "One-sided p-value is 0.482"
```

The p-value is bigger than α. Therefore, we cannot reject H_0.

Question 3:

a) I chose the cancellation policy variable. Which, has 3 levels. The population at each level is shown in the below table:

cancellation policy	Flexible	Moderate	Strict
Population	10009	10035	9913

In R:

```
> table(df$cancellation_policy)

flexible moderate   strict
   10009    10035     9913
```

Now we take two samples with a size of 100, the first sample is random. (I add the line Set.seed(5900) so every time someone runs the code the results are the same as the results of this report.)

```
> sub_df<-df[sample(nrow(df), 100), ]
> random_sample<-sub_df$cancellation_policy
> table(random_sample)
random_sample
flexible moderate   strict
      27       31       42
```

And the second sample with a size of 100, has a bias for each group. Each of the members of groups flexible, moderate, and strict have 0.1, 0.3, and 0.6 chances to be selected, respectively.

```
> biased_sample <- sample(levels(df$cancellation_policy),100,prob = c(0.6,0.1,0.3),replace = T)
> table(biased_sample)
biased_sample
flexible moderate   strict
      60        6       34
```

So our table would be as followed:

cancellation policy	Flexible	Moderate	Strict
Population	10009	10035	9913
Random sample	27	31	42
Biased sample	60	6	34

- Now for the random sample, we have the below hypothesis:

H_0 (Nothing going on): The policies selected are a simple random sample from the population. The observed counts of policies from various groups follow the same distribution in the population.

H_A (Sothing going on): The policies selected are not a simple random sample from the population. The observed counts of policies from various groups do not follow the same distribution in the population.

Now we check for if the conditions are met in the Chi-square Test:

1. Independence: Sampled observations must be independent.
    - The sample are assigned randomly.
    - The sample size is 100 which is less than 10% data.
    - Each sample only contributes to one cell in the table.
2. Sample size: Each particular scenario has at least 5 samples. Also, the excepted number of samples is 30 which is bigger than 5.

With the verification of the conditions. We can use Chi-square Test to test our hypothesis.

We now need to calculate the expected number of samples based on the percent of each group in the population.

cancellation policy	Flexible	Moderate	Strict
% in population	$\frac{10009}{29957} = 33.41\%$	$\frac{10035}{29957} = 33.49\%$	$\frac{9913}{29957} = 33.0\%$
Random sample	27	31	42
Expected	33	34	33

$$\chi^2 = \frac{(27-33)^2}{33} + \frac{(31-34)^2}{34} + \frac{(42-33)^2}{33} = 3.81016$$

$$\mathrm{df} = \mathrm{k} - 1 = 3 - 1 = 2$$

```
> Percent_population<-as.numeric(table(df$cancellation_policy))/sum(as.numeric(table(df$cancellation_policy)))
> #calculating the expcted values
> expctec<-round(Percent_population*100)
> expctec[2]<-expctec[2]+1
> Percent_population<-expctec/100
> expctec
[1] 33 34 33
> #calcualting the x2
> X<-0
> epoch<-length(random_sample)
> for (i in 1:epoch)
+ {
+   X<- X+((random_sample[i]-expctec[i])^2)/expctec[i]
+ }
> X
[1] 3.81016
> degree_f<-length(random_sample)-1
```

$p - value = 0.16($
```
> pchisq(X,degree_f,lower.tail = FALSE)
[1] 0.1488107
```
$)$

p-value is bigger than $\alpha = 0.05$. Therefore, Therefore, There is no evidence to reject H_0.

Now if we use the Chi-square Test we will get the same result:

```
> chisq.test(random_sample,p=Percent_population)

        Chi-squared test for given probabilities

data:  random_sample
X-squared = 3.8102, df = 2, p-value = 0.1488
```

p-value is bigger than $\alpha = 0.05$. Therefore, There is no evidence to reject H_0.

- Biased sample

We have the below hypothesis:

H_0 (Nothing going on): The policies selected are a simple random sample from the population. The observed counts of policies from various groups follow the same distribution in the population.

H_A (Sothing going on): The policies selected are not a simple random sample from the population. The observed counts of policies from various groups do not follow the same distribution in the population.

Now we check for if the conditions are met in the Chi-square Test:

3. Independence: Sampled observations must be independent.
    - The sample are assigned randomly.

- The sample size is 100 which is less than 10% data.
- Each sample only contributes to one cell in the table.

4. Sample size: Each particular scenario has at least 5 samples. Also, the excepted number of samples is 30 which is bigger than 5.

With the verification of the conditions. We can use Chi-square Test to test our hypothesis.

We now need to calculate the expected number of samples based on the percent of each group in the population.

cancellation policy	Flexible	Moderate	Strict
% in population	$\frac{10009}{29957} = 33.41\%$	$\frac{10035}{29957} = 33.49\%$	$\frac{9913}{29957} = 33.0\%$
Biased sample	60	6	34
Expected	33	34	33

$$\chi^2 = \frac{(60-33)^2}{33} + \frac{(6-34)^2}{34} + \frac{(34-33)^2}{33} = 45.18004$$

$$\mathrm{df} = \mathrm{k} - 1 = 3 - 1 = 2$$

```
> biased_sample<-as.numeric(table(biased_sample))
> Percent_population<-as.numeric(table(df$cancellation_policy))/sum(as.numeric(table(df$cancellation_policy)))
> #calculating the expcted values
> expctec<-round(Percent_population*100)
> expctec[2]<-expctec[2]+1
> Percent_population<-expctec/100
> expctec
[1] 33 34 33
> #calcualting the X2
> X<-0
> epoch<-length(biased_sample)
> for (i in 1:epoch)
+ {
+   X<- X+((biased_sample[i]-expctec[i])^2)/expctec[i]
+ }
> X
[1] 45.18004
> degree_f<-length(biased_sample)-1
```

$p - value \approx 0($
```
> pchisq(X,degree_f,lower.tail = FALSE)
[1] 1.546251e-10
```
$)$

p-value is smaller than $\alpha = 0.05$. Therefore, Therefore, There is a strong evidence to reject H_0. Which means, there is something going on. The observed counts of policies from various groups do not follow the same distribution in the population.

Now if we use the Chi-square Test we will get the same result:

```
> chisq.test(biased_sample,p=Percent_population)

        Chi-squared test for given probabilities

data:  biased_sample
X-squared = 45.18, df = 2, p-value = 1.546e-10
```

p-value is smaller than $\alpha = 0.05$. Therefore, Therefore, There is a strong evidence to reject H_0. Which means, there is something going on. The observed counts of policies from various groups do not follow the same distribution in the population.

B) I choose the review rate number as the second variable:

And I take 200 samples for this test. Which has the below distribution in each group:

Review rate number	1	2	3	4	5
Population	24	44	53	40	39

In R:

```
> set.seed(5900)
> sub_df <- df[sample(nrow(df), 200), ]
> table(sub_df$review.rate.number)

 1  2  3  4  5
24 44 53 40 39
```

Our two-way table would be as followed:

In R:

	Flexible	Moderate	Strict	Total
1	12	7	5	24
2	14	12	18	44
3	18	22	13	53
4	9	12	19	40
5	10	12	17	39
Total	10009	10035	9913	29957

```
> tbl <- table(sub_df$review.rate.number,sub_df$cancellation_policy)
> tbl2<-addmargins(table(sub_df$review.rate.number,sub_df$cancellation_policy))
> tbl2

      flexible moderate strict Sum
  1         12        7      5  24
  2         14       12     18  44
  3         18       22     13  53
  4          9       12     19  40
  5         10       12     17  39
  Sum       63       65     72 200
```

We want to check the relationship between two categorical variables. In order to use the Chi-square test of independence we need to check if our variables met the conditions needed for the test:

1. Independence: Sampled observations should be independent.
   - Random sample/assignment
   - if sampling without replacement, n < 10% of population
   - each case only contributes to one cell in the table.
2. Sample size: Each particular scenario (i.e. cell) must have at least 5 expected cases.

We need to calculate the expected counts in two-way tables:

$$E_{1,\text{Flexible}} = \frac{24 * 63}{200} = 7.560 \text{ , } E_{1,\text{Moderate}} = \frac{24 * 65}{200} = 7.800, E_{1,\text{Strict}} = \frac{24 * 72}{200} = 8.640$$

$$E_{2,\text{Flexible}} = \frac{44 * 63}{200} = 13.860 \text{ , } E_{2,\text{Moderate}} = \frac{44 * 65}{200} = 14.300 \text{ , } E_{2,\text{Strict}} = \frac{44 * 72}{200} = 15.840$$

$$E_{3,\text{Flexible}} = \frac{53 * 63}{200} = 16.695, E_{3,\text{Moderate}} = \frac{53 * 65}{200} = 17.225, E_{3,\text{Strict}} = \frac{53 * 72}{200} = 19.080$$

$$E_{4,\text{Flexible}} = \frac{40 * 63}{200} = 12.600, E_{4,\text{Moderate}} = \frac{40 * 65}{200} = 13.000, E_{4,\text{Strict}} = \frac{40 * 72}{200} = 14.400$$

$$E_{5,\text{Flexible}} = \frac{39 * 63}{200} = 12.285, E_{5,\text{Moderate}} = \frac{39 * 65}{200} = 12.675, E_{5,\text{Strict}} = \frac{39 * 72}{200} = 14.040$$

The two-way table with expected values respective to each cell is shown below:

	Flexible	Moderate	Strict	Total
1	12 (7.560)	7 (7.800)	5 (8.640)	24
2	14 (13.860)	12 (14.300)	18 (15.840)	44
3	18 (16.695)	22 (17.225)	13 (19.080)	53
4	9 (12.600)	12 (13.000)	19 (14.400)	40
5	10 (12.285)	12 (12.675)	17 (14.040)	39
Total	10009	10035	9913	29957

I need to mention that for calculating the expected values in the above table, I did not round the numbers.

In R:

```
> #calcuating expcted values
> for (i in 1:6)
+ {
+   for (j in 1:4)
+   {
+     tbl3[i,j]<-tbl2[i,4]*tbl2[6,j]/tbl2[6,4]
+
+   }
+ }
> tbl3

     flexible moderate  strict     Sum
  1     7.560    7.800   8.640  24.000
  2    13.860   14.300  15.840  44.000
  3    16.695   17.225  19.080  53.000
  4    12.600   13.000  14.400  40.000
  5    12.285   12.675  14.040  39.000
  Sum  63.000   65.000  72.000 200.000
```

$$\chi^2 = \frac{(12-7.560)^2}{7.560} + \frac{(7-7.800)^2}{7.800} + \frac{(5-8.640)^2}{8.640} + \frac{(14-13.860)^2}{13.860} + \frac{(12-14.300)^2}{14.300} + \frac{(18-15.840)^2}{15.840} + \frac{(18-16.695)^2}{16.695} + \frac{(22-17.225)^2}{17.225} + \frac{(13-19.080)^2}{17.225} + \frac{(9-12.600)^2}{17.225} + \frac{(12-13.000)^2}{17.225} + \frac{(19-14.400)^2}{17.225} + \frac{(10-12.285)^2}{17.225} + \frac{(12-12.675)^2}{17.225} + \frac{(17-14.040)^2}{17.225} = 11.91216$$

$$df = (R-1)*(C-1) = (4-1)*(3-1) = 6$$

In R:

```
> X<-0
> #calculating X2
> for (i in 1:6)
+ {
+   for (j in 1:4)
+   {
+     X=X+ ((tbl2[i,j]-tbl3[i,j])^2)/tbl3[i,j]
+   }
+ }
> X
[1] 11.91216
```

And we calculate the p-value:

```
> pchisq(X,degree_f,lower.tail = FALSE)
[1] 0.1551662
```

p-value = 0.1551. With a p-value greater than 5%, Since p-value is high, we fail to reject H_0. Therefore, we conclude that there is not enough evidence in the data to suggest that the cancelation policty is dependent on the rating levels of the houses.

The build in function gives us the same result:

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 11.912, df = 8, p-value = 0.1552
```

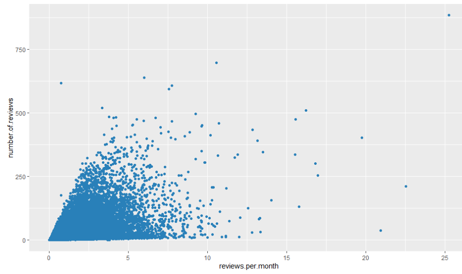
p-value = 0.1551. With a p-value greater than 5%, Since p-value is high, we fail to reject H_0. Therefore, we conclude that there is not enough evidence in the data to suggest that the cancelation policy is dependent on the rating levels of the houses.

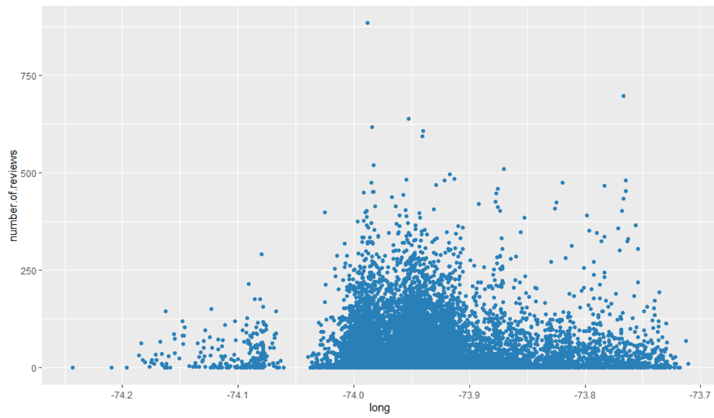
Question 4:

One of the variables that should have an obvious relationship with other variables is the number of reviews.

Number of reviews is clearly dependent on reviews per month. Any house with more reviews per month will have more Number of reviews eventually.

Also, since some neighborhoods of New York are more attractive for tourists such as Queens and Manhattan. Variables such as longitude could have a relationship with number of reviews. Meaning that, houses located at more visited areas will have more number of reviews overall. I should mention that since the distribution of houses in NYC is normal. The assumption of exact linear relation between number of reviews and longitude is not correct. However, we keep this this variable for the sake of answering this question.





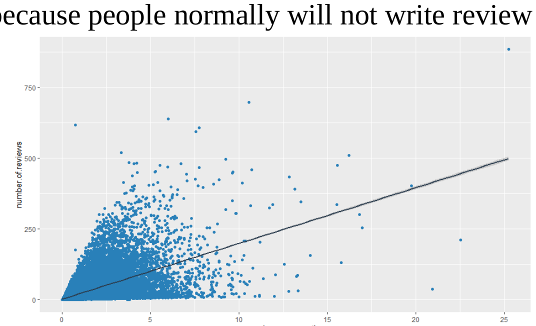
A) It is understandable that with more reviews per month, the total number of reviews should also increase. In other words, these two variables have an obvious linear relationship. Therefore, with the number of reviews as the response variable, the reviews per month variable would provide the best predictor.

B) For the each explanatory variable we check the conditions:

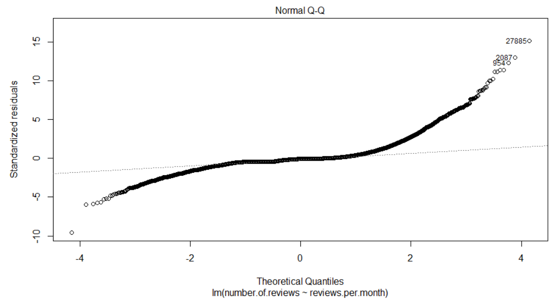
a.

1- The reviews per month variable:

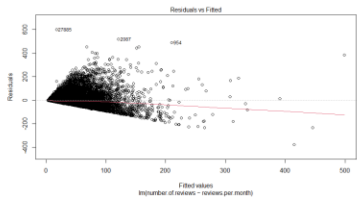
i. Linearity: the relationship between reviews per month and the number of reviews, in theory should be linear. As the reviews per month increases the total number of reviews increases too.
The scatter plot shows a similar pattern of linear relationship between this two variable. However, the scatter plot is mostly saturated around low numbers which for makes sense because people normally will not write review that much.

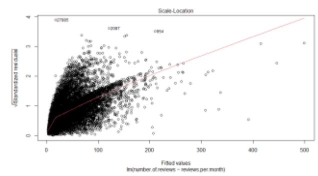


ii. Nearly normal residuals: the residuals are close to normal. However, our data has some outliers which will skew any of our results.



iii. Constant variability: we do not met this condition for this variable. Because, variability of points around the least squares line is not roughly constant. It was obvious that this will not be correct because we had so many outliers.

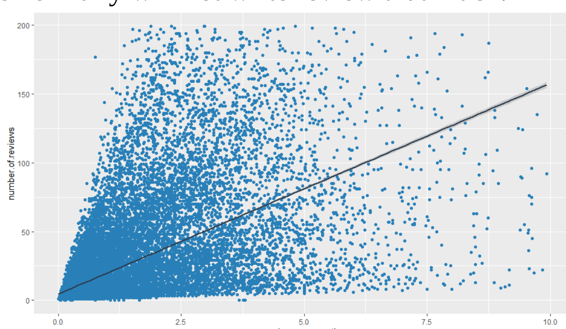




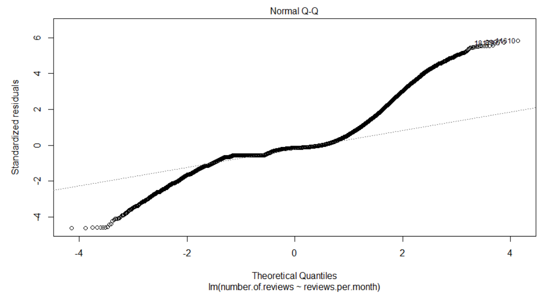
We can check these conditions with deleting the outliers:

We delete outlier variables in reviews per month. Here, we consider houses with more than 10 reviews per month and more than 200 total number of reviews as outlier.

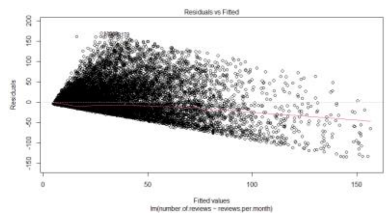
i. Linearity: the relationship between reviews per month and the number of reviews, in theory should be linear. As the reviews per month increases the total number of reviews increases too.
The scatter plot shows a similar pattern of linear relationship between this two variable. However, the scatter plot is mostly saturated around low numbers which for makes sense because people normally will not write review that much.

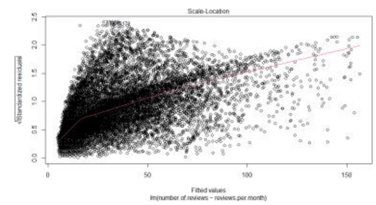


i. Nearly normal residuals: the residuals are kind of close to normal. (If we take the condition rules strictly this condition is not met.)



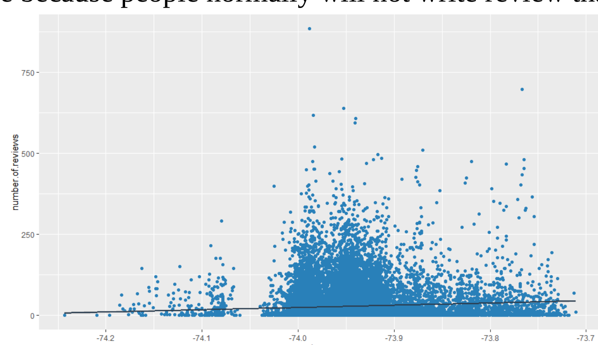
ii. Constant variability: With the deleting the outliers the variably is almost constant in our data.



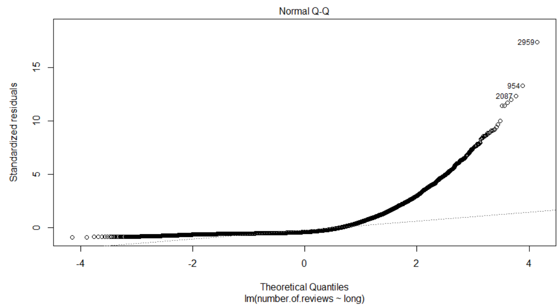


2- The longitude:

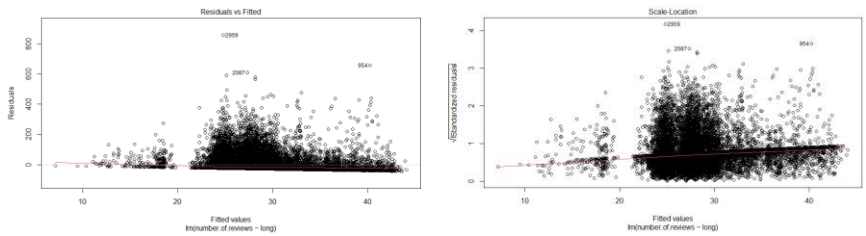
i. Linearity: As we move toward the center of NYC, and more popular area, houses in these area get more visitor and as the result more total number of reviews. However, as I mentioned, the distribution of hoses are close to normal distribution. Therefore, a approximated positive linear relationship is established up to the middle point of the data, and after that we have approximated negative linear relationship between varaibles.
The scatter plot shows a similar pattern of linear relationship between this two variable. However, the scatter plot is mostly saturated around low numbers which for makes sense because people normally will not write review that much.



ii. Nearly normal residuals: the residuals are close to normal. However, our data has some outliers which will skew any of our results.



iii. Constant variability: This condition is not met for this variable. Because, variability of points around the least squares line is not roughly constant. It was obvious that this will not be correct because we had so many outliers.

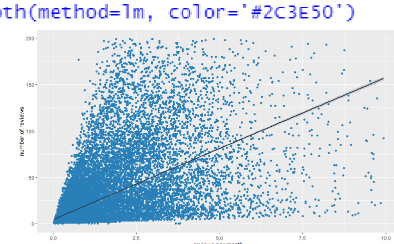


We can again delete parts of data that forbids us from having a linear relationship. But, for this variable eliminating parts of data does not make any sense because the data is definitely correct (in the other variables, that many review per month could be incorrect data which I explained about this in the first parts of the project). Therefore, we will not do this.

b.

1- The reviews per month variable: we calculate and plot the least square line.

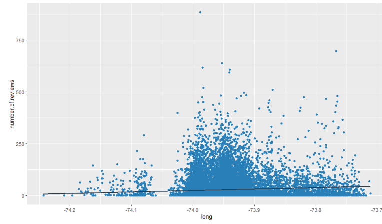
```
> ggplot(df, aes(x=reviews.per.month, y=number.of.reviews)) +
+   geom_point(color='#2980B9') +
+   geom_smooth(method=lm, color='#2C3E50')
```




```
> my_model_rev <- lm( number.of.reviews~reviews.per.month, data = df)
> summary(my_model_rev)

Call:
lm(formula = number.of.reviews ~ reviews.per.month, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-377.77  -17.12   -2.73    4.19  600.88 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.3501     0.3002   7.829 5.09e-15 ***
reviews.per.month  19.6953     0.1518 129.708  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.64 on 29955 degrees of freedom
Multiple R-squared:  0.3597,	Adjusted R-squared:  0.3596 
F-statistic: 1.682e+04 on 1 and 29955 DF,  p-value: < 2.2e-16
```

2- Longitude: we calculate plot the least square line.

```
> ggplot(df, aes(x=long, y=number.of.reviews)) +
+   geom_point(color='#2980B9') +
+   geom_smooth(method=lm, color='#2C3E50')
```



c.

1- The reviews per month variable:
The slope= 19.69526 and intercept= 2.350147

$$\text{number\_of\_revs} = 19.69*\text{revs\_per\_month} + 2.35$$

The slope meaning is with adding one reviews per month the number of reviews increases by 69.47656.
The intercept means a listing with zero reviews per month has 2.35 reviews.

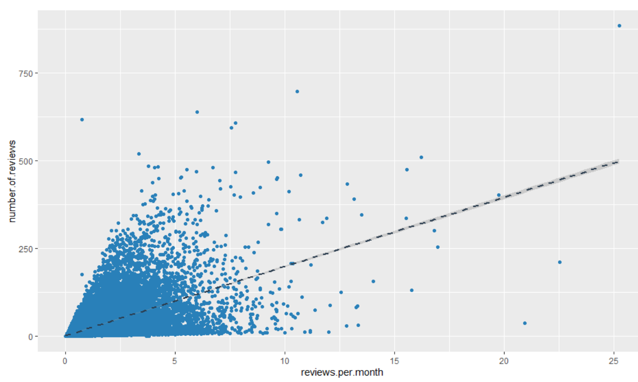
2- Longitude variable:
The slope= 69.47656 and intercept= 5165.287

$$\text{number\_of\_revs} = 69.47656\ *\text{Longitude}\ + 5165.287$$

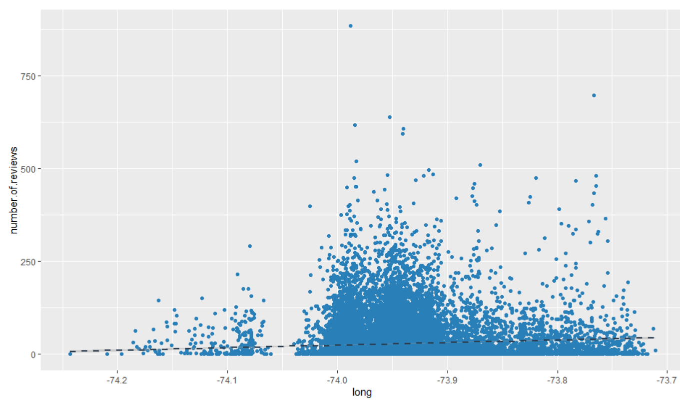
The slope meaning is with increase of one degree in longitude the number of reviews increases by 69.47656.
The intercept means a listing located in zero Longitude has 5165.287 reviews.

d.

1- The reviews per month variable:



2- Longitude variable:



C) Considering both of these variables did not very perform well.

However, the reviews per month variable is the more significant predictor. It predicts the pattern of the number of reviews correctly and is linearly depended.

D)

The model for the Review per month variable is:

```
> summary(my_model_rev)

Call:
lm(formula = number.of.reviews ~ reviews.per.month, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-377.77  -17.12   -2.73    4.19  600.88 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.3501     0.3002   7.829 5.09e-15 ***
reviews.per.month  19.6953     0.1518 129.708  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.64 on 29955 degrees of freedom
Multiple R-squared:  0.3597,	Adjusted R-squared:  0.3596 
F-statistic: 1.682e+04 on 1 and 29955 DF,  p-value: < 2.2e-16
```

It is shown that adjusted R-squared=0.3596 And the p-value=0.

Therefore, we conclude that: The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

For the longitude variable:

```
> summary(my_model_long)

Call:
lm(formula = number.of.reviews ~ long, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-43.62 -25.10 -19.64   3.00 859.18 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5165.287    428.647   12.05   <2e-16 ***
long          69.477      5.796   11.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.42 on 29955 degrees of freedom
Multiple R-squared:  0.004773,	Adjusted R-squared:  0.00474 
F-statistic: 143.7 on 1 and 29955 DF,  p-value: < 2.2e-16
```

It is shown that adjusted R-squared=0.04 And the p-value=0.

Therefore, we conclude that: The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

The first model has a higher adjusted R2. This means the review per month variable accounts for more percent of variably in the response variable. Therefore, it is the better predictor.

E) The best predictor is a predictor that has high value of adjusted R2 and the explanatory variable be a significant predictor of the response variable.

In this dataset the review per month variable is the best predictor for the number of reviews that we could get.

F) We make two predictor one with review per month variable and one with longitude.

We split the data into two parts of train and test

```
> set.seed(5900)
> sub_df <- df[sample(nrow(df), 100), ]
> train<-sub_df[sample(nrow(sub_df), 90), ]
> test<-sub_df[-sample(nrow(sub_df), 90), ]
```

a. Our hypotheses will look like this:

H_0: $\beta_1 = 0$: Nothing going on. The review per month variable is not a significant predictor of the response variable, i.e. no relationship → slope of the relationship is 0.

$H_A: \beta_1 \neq 0$: *something going on*. The review per month variable is a significant predictor of the response variable, i.e. relationship → slope of the relationship is different than 0.

```
> my_model <- lm(number.of.reviews ~ reviews.per.month , data = train)
> summary(my_model)

Call:
lm(formula = number.of.reviews ~ reviews.per.month, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-84.980 -15.758  -3.625   4.690 221.480 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)          4.632      4.741   0.977    0.331    
reviews.per.month   14.835      2.195   6.760 1.45e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.52 on 88 degrees of freedom
Multiple R-squared:  0.3418,	Adjusted R-squared:  0.3343 
F-statistic: 45.69 on 1 and 88 DF,  p-value: 1.45e-09
```

p-value=0.

Therefore, we conclude that the data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

For the longitude variable we have:

```
> my_model2 <- lm(number.of.reviews ~ long , data = train)
> summary(my_model2)

Call:
lm(formula = number.of.reviews ~ long, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.873 -23.309 -10.483   5.055 212.128 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23484.37    6641.81   3.536 0.000651 ***
long          317.21      89.81   3.532 0.000659 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.82 on 88 degrees of freedom
Multiple R-squared:  0.1242,	Adjusted R-squared:  0.1142 
F-statistic: 12.48 on 1 and 88 DF,  p-value: 0.0006591
```

$H_0: \beta_1 = 0$: Nothing going on. The longitude variable is not a significant predictor of the response variable, i.e. no relationship → slope of the relationship is 0.

$H_A: \beta_1 \neq 0$: *something going on*. The longitude variable is a significant predictor of the response variable, i.e. relationship → slope of the relationship is different than 0.

Therefore, we conclude that the data provide convincing evidence that the slope is significantly different than 0, i.e. the longitude variable is a significant predictor of the response variable.

b.

```
> confint(my_model)
                       2.5 %   97.5 %
(Intercept)        -4.790233 14.05373
reviews.per.month 10.473302 19.19606
> confint(my_model2)
                 2.5 %     97.5 %
(Intercept) 10285.1578 36683.5757
long          138.7333   495.6869
```

c. The predicted value of test are as followed:

```
> predict(my_model,test)
    25672     23743     29230     26894      8320     28490     14557      9034     15326     28262
 5.818522 35.932920 98.980304 15.757757 38.751509 12.049087 15.757757 23.175096 53.586187 12.049087
> predict(my_model2,test)
    25672     23743     29230     26894      8320     28490     14557      9034     15326     28262
30.447945  6.891923 11.872121 14.961748 32.534871  8.398671 33.442409 11.152055 13.841996 30.517731
```

d. Our success rate for rate for the models are as followed:

For rev per month:

```
> data.frame(RMSE= RMSE(predict(my_model,test), test$number.of.reviews),
+            R2= R2(predict(my_model,test), test$number.of.reviews),
+            MAE= MAE(predict(my_model,test), test$number.of.reviews))
      RMSE          R2      MAE
1 39.74197 0.004095572 28.14099
```

For longitude:

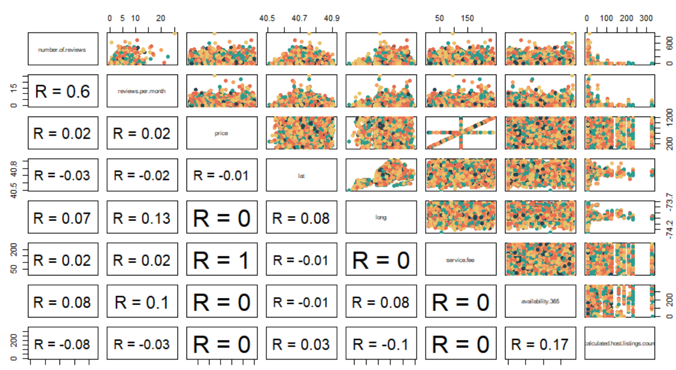
```
> data.frame(RMSE= RMSE(predict(my_model2,test), test$number.of.reviews),
+            R2= R2(predict(my_model2,test), test$number.of.reviews),
+            MAE= MAE(predict(my_model2,test), test$number.of.reviews))
     RMSE        R2      MAE
1 34.0806 0.0774928 22.25453
```

Question 5:

a) As predicted in the last question the reviews per month has the most correlation with number of reviews.

The other variable that has the most correlation with number of reviews is availability.365 which has a close to 0 correlation and normally we will not use it. But, for the sake of training multiple linear regression we use this variable too.

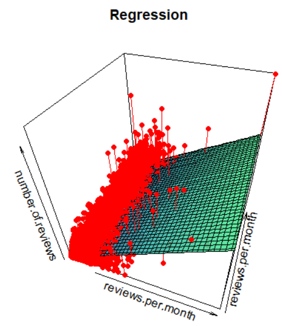
I explained the relationship between longitude and number of reviews in the last question. It has a correlation close to availability.365 too. And both of them have almost zero correlation to number of reviews. Here, for the sake of simplicity we only use the availability.365 feature.



b)

```
> my_model <- lm(number.of.reviews ~ reviews.per.month+availability.365 , data = df)
> summary(my_model)

Call:
lm(formula = number.of.reviews ~ reviews.per.month + availability.365,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-376.20  -16.13   -2.92    4.14  598.56

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.190543   0.366637   3.247  0.00117 **
reviews.per.month 19.613049   0.152502 128.609  < 2e-16 ***
availability.365   0.009814   0.001783   5.504 3.75e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.62 on 29954 degrees of freedom
Multiple R-squared:  0.3603,    Adjusted R-squared:  0.3603
F-statistic:  8435 on 2 and 29954 DF,  p-value: < 2.2e-16
```

c) The model follows the patter of number of the reviews and reviews per month. However, as it is shown in below figure. We have lots of error but with considering the linearity of our model. It performs well enough.



d) We split the data into two groups of train and test with the ratio of 0.8:

```
> sub_df<-df[,c('number.of.reviews','reviews.per.month')]
> train <- sub_df[sample(nrow(sub_df),floor(nrow(sub_df)*0.8)),]
> test <- sub_df[-sample(nrow(sub_df),floor(nrow(sub_df)*0.8)),]
```

Now we train our model:

```
> my_model <- lm(number.of.reviews ~., data = train )
> summary(my_model)

Call:
lm(formula = number.of.reviews ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-377.80  -17.02   -2.65    4.20  600.98 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         2.2410     0.3310    6.77 1.32e-11 ***
reviews.per.month  19.7017     0.1674  117.71  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.2 on 23963 degrees of freedom
Multiple R-squared:  0.3664,	Adjusted R-squared:  0.3663 
F-statistic: 1.386e+04 on 1 and 23963 DF,  p-value: < 2.2e-16
```

We calculate the RMSE, MAE and R squared value for train and test:

```
> predictions <- my_model %>% predict(train)
> data.frame(RMSE= RMSE(predictions, train$number.of.reviews),
+ R2= R2(predictions, train$number.of.reviews),
+ MAE= MAE(predictions, train$number.of.reviews))
      RMSE        R2      MAE
1 39.19584 0.3663711 21.59933
> predictions <- my_model %>% predict(test)
> data.frame(RMSE= RMSE(predictions, test$number.of.reviews),
+            R2= R2(predictions, test$number.of.reviews),
+            MAE= MAE(predictions, test$number.of.reviews))
      RMSE        R2     MAE
1 40.89616 0.3567291 22.4497
```

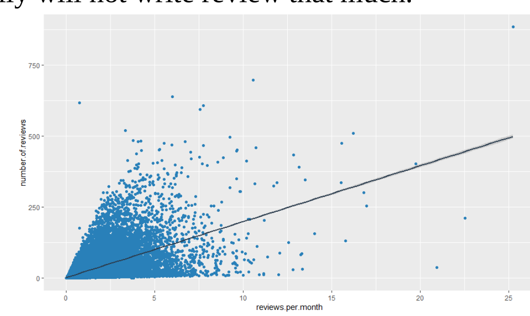
e)

```
> #E)
> set.seed(123)
> train.control <- trainControl(method = "cv", number = 5)
> # Train the model
> model <- train(number.of.reviews ~., data = train, method = "lm",
+                trControl = train.control)
> # Summarize the results
> print(model)
Linear Regression 

23965 samples
    1 predictor

No pre-processing
Resampling: Cross-Validated (5 fold) 
Summary of sample sizes: 19171, 19174, 19171, 19172, 19172 
Resampling results:

  RMSE     Rsquared   MAE     
  39.1847  0.3668594  21.60242

Tuning parameter 'intercept' was held constant at a value of TRUE
```

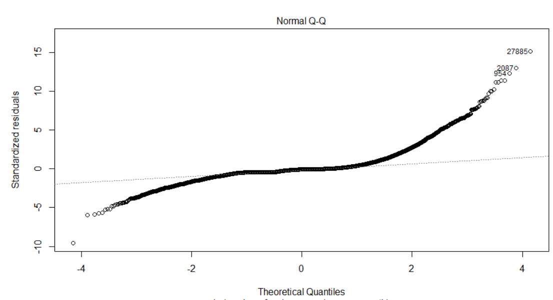
As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance. Therefore, models with lower RMSE are better. The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures accuracy for continuous variables. Therefore, models with lower MAE are better. R-Squared is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. Therefore, models with higher R-Squared are better.

f) I explained the relationship between these two variable in the question number 4, However I repeat it here:

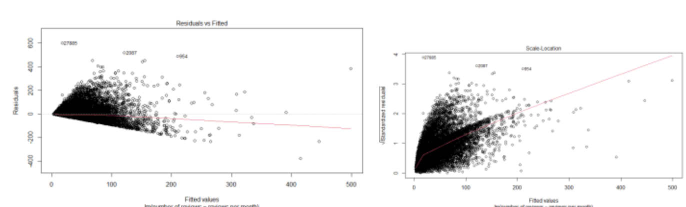
   i. Linearity: the relationship between reviews per month and the number of reviews, in theory should be linear. As the reviews per month increases the total number of reviews increases too.
   The scatter plot shows a similar pattern of linear relationship between this two variable. However, the scatter plot is mostly saturated around low numbers which for makes sense because people normally will not write review that much.



   ii. Nearly normal residuals: the residuals are close to normal. However, our data has some outliers which will skew any of our results.



   iii. Constant variability: we do not met this condition for this variable. Because, variability of points around the least squares line is not roughly constant. It was obvious that this will not be correct because we had so many outliers.



One of the conditions is not met. Therefore, the model is not reliable. Although, I should mention that the model is a linear model. And for this data this model performs good enough.

g) *R*-squared is the proportion of variability in *y* explained by the model. In part B R-squared=36.03. which means model B explains 36.03 percent of variability of the number of reviews. In part D R-squared=36.28. which means model D explains 36.28 percent of variability of the number of reviews.

```
> summary(fit)

Call:
lm(formula = z ~ x + y)

Residuals:
    Min      1Q  Median      3Q     Max 
-376.20  -16.13   -2.92    4.14  598.56 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.190543   0.366637   3.247  0.00117 ** 
x           19.613049   0.152502 128.609  < 2e-16 ***
y            0.009814   0.001783   5.504 3.75e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.62 on 29954 degrees of freedom
Multiple R-squared:  0.3603,	Adjusted R-squared:  0.3603 
F-statistic:  8435 on 2 and 29954 DF,  p-value: < 2.2e-16

> print(model)
Linear Regression 

23965 samples
    1 predictor

No pre-processing
Resampling: Cross-Validated (5 fold) 
Summary of sample sizes: 19173, 19172, 19172, 19171, 19172 
Resampling results:

  RMSE      Rsquared   MAE     
  39.49906  0.3628545  21.61435

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Question 6:

I chose the instant bookable variable as my response variable. This variable mostly depends on the owner decision. Therefore, other variable in this dataset might not be as effective on this response variable. However, I suspect variables such as room type can be a good explanatory variable. (for example hotel rooms have a processor for accepting new guests)

Categorical and numerical variable that I chose for this tests are:

Price, room.type, neighbourhood.group, cancellation_policy and host_identity_verified. Between these variables the room.type and cancellation_policy might have an effect on instant bootability of the house.

a) Our logistic regression:

```
> my_model <- glm(instant_bookable ~ price + neighbourhood.group
+                 +room.type+cancellation_policy+host_identity_verified, data = df,family ="binomial" )
> summary(my_model)

Call:
glm(formula = instant_bookable ~ price + neighbourhood.group + 
    room.type + cancellation_policy + host_identity_verified, 
    family = "binomial", data = df)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-1.265  -1.166  -1.132   1.188   1.225  

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)  
(Intercept)                   3.427e-04  4.966e-02   0.007   0.9945  
price                         8.187e-07  3.497e-05   0.023   0.9813  
neighbourhood.group2          3.670e-03  4.535e-02   0.081   0.9355  
neighbourhood.group3          3.523e-02  4.523e-02   0.779   0.4360  
neighbourhood.group4          3.205e-02  4.524e-02   0.708   0.4787  
neighbourhood.group5          5.321e-02  4.527e-02   1.175   0.2399  
room.typeHotel room           5.446e-02  3.294e-01   0.165   0.8687  
room.typePrivate room        -4.467e-02  2.345e-02  -1.905   0.0568 .
room.typeShared room          1.404e-01  7.962e-02   1.764   0.0778 .
cancellation_policymoderate  -6.702e-02  2.827e-02  -2.371   0.0178 *
cancellation_policystrict    -4.479e-02  2.835e-02  -1.580   0.1142  
host_identity_verifiedverified  7.942e-03  2.313e-02   0.343   0.7313  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41526  on 29956  degrees of freedom
Residual deviance: 41509  on 29945  degrees of freedom
AIC: 41533

Number of Fisher Scoring iterations: 3
```

The interpretation of each of intercept and slope are as followed:

- Intercept: The log odds of instance bootability for a house with price of 0 and other variable having the reference value.
- Slope: The interpretation for a slope is the change in log odds ratio per unit change in the predictor.
  - Price: For a unit increase in price (1$ more expensive) how much will the log odds ratio change.
  - Room type: When the other predictors are held constant this is the log odds ratio between the contrast of room type value and the reference value of room type (entire home). In other word the model predicts the chance of instance bookablity for each room type is X% higher than entire home.
  - Cancelation policy: When the other predictors are held constant this is the log odds ratio between the contrast of cancelation policy value and the reference value of cancelation policy (flexible). In other word the model predicts the chance of instance bookability for each cancelation policy is X% higher than flexible policy.
  - Host identity verification: When the other predictors are held constant this is the log odds ratio between the contrast of host identity value and the reference value(not verified). In other word the model predicts the chance of instance bookablity for each verified owners is X% higher than not verified.
  - Host identity verification: When the other predictors are held constant this is the log odds ratio between the contrast of neighborhood group value and the reference value(group 1). In other word the model predicts the chance of instance bookablity for each group is X% higher than group 1.

```
> exp(coef(my_model))
                (Intercept)                       price          neighbourhood.group2
                  1.0003428                   1.0000008                     1.0036767
       neighbourhood.group3        neighbourhood.group4          neighbourhood.group5
                  1.0358574                   1.0325695                     1.0546543
         room.typeHotel room       room.typePrivate room          room.typeShared room
                  1.0559721                   0.9563129                     1.1507755
 cancellation_policymoderate    cancellation_policystrict host_identity_verifiedverified
                  0.9351798                   0.9561998                     1.0079735
```

b) We split the data into two groups of train and test with the ratio of 0.8:

```
> sub_df<-df[,c('instant_bookable','cancellation_policy', 'price'
+                ,'room.type',"neighbourhood.group","host_identity_verified")]
> train <- sub_df[sample(nrow(sub_df),floor(nrow(sub_df)*0.8)),]
> test <- sub_df[-sample(nrow(sub_df),floor(nrow(sub_df)*0.8)),]
```

Now we train our model using the train data:

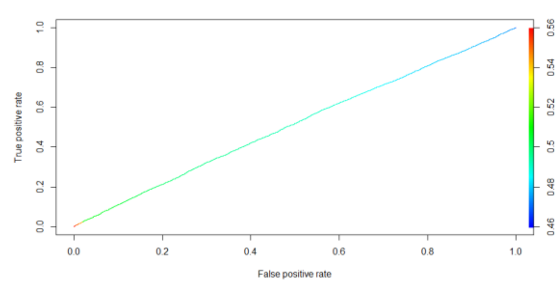
```
> my_model <- glm(instant_bookable ~., data = train,family ="binomial" )
> summary(my_model)

Call:
glm(formula = instant_bookable ~ ., family = "binomial", data = train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.279  -1.165  -1.136   1.189   1.240

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   1.516e-02  5.550e-02   0.273   0.7847
cancellation_policymoderate  -4.252e-02  3.162e-02  -1.345   0.1787
cancellation_policystrict    -3.249e-02  3.171e-02  -1.025   0.3055
price                        -1.341e-05  3.913e-05  -0.343   0.7319
room.typeHotel room          -1.162e-01  3.441e-01  -0.338   0.7357
room.typePrivate room        -6.047e-02  2.623e-02  -2.306   0.0211 *
room.typeShared room          1.980e-01  8.904e-02   2.223   0.0262 *
neighbourhood.group2          3.837e-03  5.067e-02   0.076   0.9396
neighbourhood.group3          2.377e-02  5.060e-02   0.470   0.6386
neighbourhood.group4          1.276e-02  5.056e-02   0.252   0.8008
neighbourhood.group5          2.161e-02  5.057e-02   0.427   0.6691
host_identity_verifiedverified -1.790e-03  2.586e-02  -0.069   0.9448
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33218  on 23964  degrees of freedom
Residual deviance: 33203  on 23953  degrees of freedom
AIC: 33227

Number of Fisher Scoring iterations: 3
```

Now we can make a prediction based on our train data, the confusion matrix with threshold of 0.5 is as followed:

```
> pred_resp <- predict(my_model,type="response")
> table(train$instant_bookable, (pred_resp > 0.5)*1, dnn=c("Truth","Predicted"))
       Predicted
Truth      0    1
  False 9587 2565
  True  9174 2639
```

Now we can plot the roc curve for the train data:

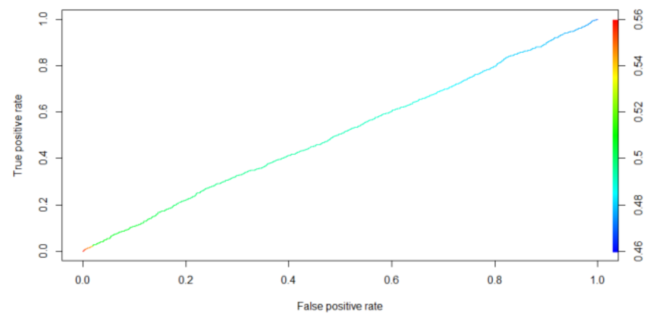
```
> pred <- prediction(pred_glm0_train, train$instant_bookable)
> perf <- performance(pred, "tpr", "fpr")
> plot(perf, colorize=TRUE)
```



And The AUC= 0.5127846 (`> unlist(slot(performance(pred, "auc"), "y.values"))` `[1] 0.5127846`)

We repeat the above steps for the test data too:

```
> pred_glm0_test<- predict(my_model, newdata = test, type="response")
> pred <- prediction(pred_glm0_test, test$instant_bookable)
> perf <- performance(pred, "tpr", "fpr")
> plot(perf, colorize=TRUE)
```



The AUC= 0.5073914 (`> unlist(slot(performance(pred, "auc"), "y.values"))` `[1] 0.5073914`)

The ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

Here, the curve shows, the separation of two classes is very hard for our model. Our AUC is around 0.5 which means, the model has low to no class separation capacity.

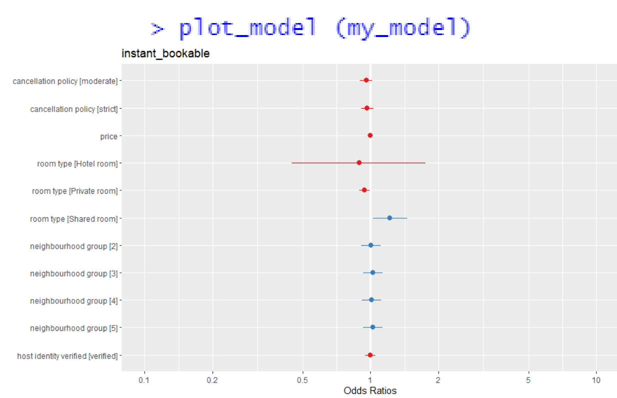
The acquired result was anticipated because as I said in the first parts of this question the instance bootability is really based on the owner decision.

c)

The plot below shows odds ratio with their confidence interval:

- OR > 1 means greater odds of association with the exposure and outcome.
- OR = 1 means there is no association between exposure and outcome.
- OR < 1 means there is a lower odds of association between the exposure and outcome.

The only variable that has a significant impact on instant bootability is room type.



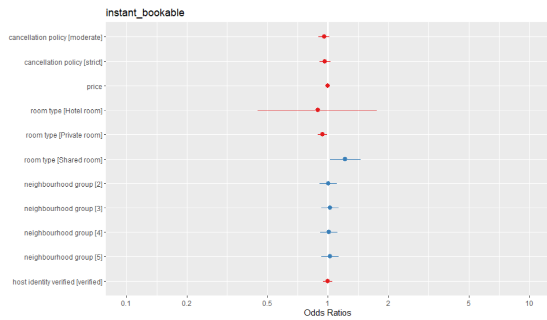
d) Our confidence intervals are as followed:

```
> exp(confint(my_model))
Waiting for profiling to be done...
                                 2.5 %    97.5 %
(Intercept)                  0.9106104 1.1319580
cancellation_policymoderate  0.9007868 1.0196371
cancellation_policystrict    0.9096962 1.0301008
price                        0.9999099 1.0000633
room.typeHotel room          0.4486321 1.7511145
room.typePrivate room        0.8941583 0.9909705
room.typeShared room         1.0240796 1.4521265
neighbourhood.group2         0.9089432 1.1086888
neighbourhood.group3         0.9273761 1.1308447
neighbourhood.group4         0.9172991 1.1183701
neighbourhood.group5         0.9254301 1.1283472
host_identity_verifiedverified 0.9488834 1.0501041
```

Question 7:

a) The room type variable had the most significant role in the prediction. Based on the p-values and the Odds ratio. Which makes sense, some listings such as hotels have procedure for booking and they do not have instant bootability available. However, as I mentioned in the question 6. The model is not very powerful, because our classes are distribuend almost evenly and it is not simply separable with our features.

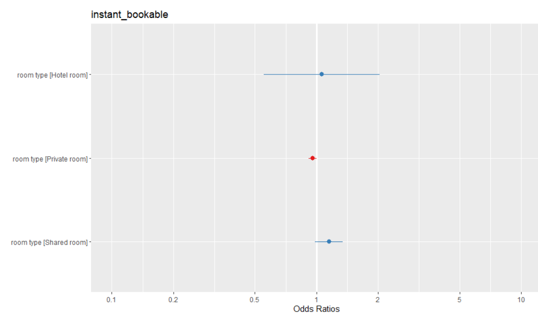
b)



The cancellation policy has OR < 1 which means, there is a lower odds of association between the exposure and outcome. Again, this variable has almost no effect on the response variable.

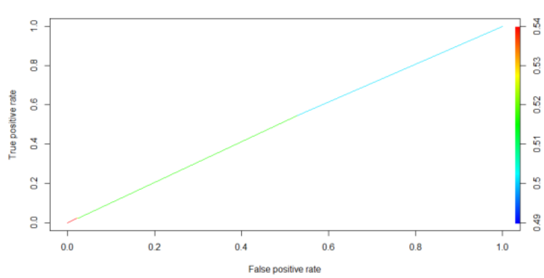
c) In this model the room type has a significant impact on instant bootability.

```
> my_model <- glm(instant_bookable ~ room.type, data = df,family ="binomial" )
> summary(my_model)

Call:
glm(formula = instant_bookable ~ room.type, family = "binomial",
    data = df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.236  -1.176  -1.157   1.179   1.198

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.003961   0.015987  -0.248   0.8043
room.typeHotel room  0.058029   0.329306   0.176   0.8601
room.typePrivate room -0.045137  0.023446  -1.925   0.0542 .
room.typeShared room  0.140330   0.079594   1.763   0.0779 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41526  on 29956  degrees of freedom
Residual deviance: 41518  on 29953  degrees of freedom
AIC: 41526

Number of Fisher Scoring iterations: 3
```



The room type is significant. But it still can not perform the task of classification of the response variable very well.

d) Because this model is still not powerful, with any threshold the result will be the same:



As you can see in the above curve, threshold does not really play a role in the classification of our data.