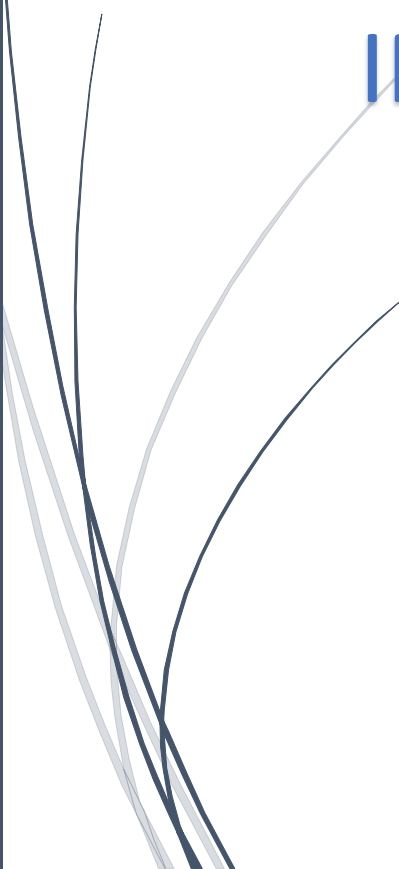




12/14/2020

# FINAL PROJECT – STAT 663

## AIRBNB LISTING RENTAL PRICE PREDICTION IN NEW YORK CITY



# REPORT

## 1. Problem Introduction & Data set

In this project, we would like to target the scientific question of predicting Airbnb listing prices in New York city using other information about a listing such as location (GPS coordinates, neighbourhood), number of reviews, number of listings its host has, whether it is shared space, its availability throughout the year, etc. Our project objective is to figure out the major factors that influence the price of a listing and whether there exists any interesting patterns that allow us to know more about the renting market in New York city. We plan to achieve this goal by applying statistical learning methods that we have learned throughout the course to predict listing prices and explore the important features that have significant effects.

The data set we use to explore can be directly downloaded on the Kaggle website (Kaggle, 2019). Each observation contains detailed information about an Airbnb listing in New York city in 2019. There are 48895 observations and 16 features, including: **ID** (listing ID), **name** (name of the listing), **host ID** (ID of the host), **host\_name** (name of the host), **neighbourhood\_group** (the borough in New York), **neighbourhood** (the area of the listing), **latitude** (latitude coordinate of the listing), **longitude** (longitude coordinate of the listing), **room\_type** (type of the listing, e.g., a private room or an entire apartment), **price** (price in dollars), **minimum\_nights** (number of minimum nights that customers have to spend), **number\_of\_reviews** (total number of reviews of the listing), **last\_review** (when the latest review was posted), **reviews\_per\_month** (average monthly number of reviews of the listing), **calculated\_host\_listings\_count** (number of listings the host owns), **availability\_365** (number of available days through the year 2019 that people could book the listing). Figure 19 (in the Appendix) displays an example of how the original data set looks like.

## 2. Data preprocessing

Before visualizing the data set and fitting models, we need to transform and convert some variables into other formats. The data preprocessing process is executed in details as follows:

- **neighbourhood\_group** & **room\_type**: these features are converted from character to factors. The number of levels of *neighbourhood\_group* and *room\_type* is 5 and 2 respectively.
- **last\_review**: this feature has a date format. To convert it from date to numeric in a simple manner, we decide to keep the year number only.

Also, we exclude four features, including: **id**, **host\_id**, **host\_name** and **neighbourhood**. We determine that *id*, *host\_id* and *host\_name* of a listing are irrelevant for predicting price. Even though host name and id may be good indicators for price prediction (as some hosts appear to have better reputations over the others), we believe they are not good for generalizing to listings from new/unseen hosts and they are not interesting to learn from. We also exclude the **neighbourhood** categorical variable due to its numerous levels (as many as 221) with many of them having less than 4 observations, which may make the model prone to overfitting.

When fitting models, we perform variable transformation for **name** and **minimum\_night**.

- **name**: we convert it from character to numeric (unigram and bigram). This process is explained in details in Model 4 of section 4.

- **minimum\_night**: this feature is transformed from numeric to a factor with 2 levels (“short” and “long”). The motivation for this is explained in Model 2 of section 4.

### 3. Data visualization

We perform exploratory data analysis using histograms, density plots, scatter plots, correlation matrix, and box plots to understand more about the data set. Our prediction target is price, so it is very important to see the distribution of this variable. From Figure 1, we can see that the distribution of price is extremely skewed to the left with only 2.1% of listings having rental price per night over \$500. From our subjective understanding, Airbnb listings with over \$500 per night are very luxurious places or they could simply be data crawling errors, scams, or the results of someone wanting to test the limit of the Airbnb system (there are 6 listings with price as much as \$10000/night). If we are to fit a statistical learning model on this kind of data set with so many outliers, the result could become very catastrophic. **As a result, we decide to drop these outliers which consist of all listings with over \$500 per night.** This also fits nicely with our project goal which is to focus on learning what factors contribute to the rental price of the majority (97.9% in the data set) of Airbnb listings, rather than trying to make a model fit well on the extremely pricey listings (the other 2.1%) which seems to be an extremely difficult task.

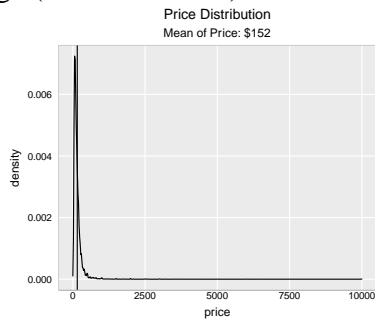


Figure 1: Distribution of Price

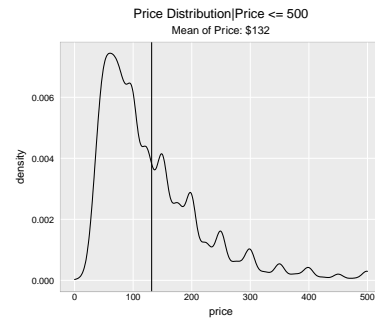


Figure 2: Distribution of Price (Price <= 500)



Figure 3: Price and Neighbourhood\_group

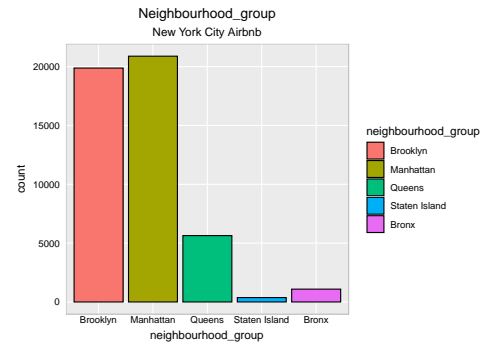


Figure 4: Number of listings in each neighbourhood\_group

Next, we would like to visualize the relationship between target and predictor variables. As we know, one of the most important factors that contributes to the rental price of one listing is the current market price. This means that when a host puts their listing on Airbnb, they would have to make comparisons between the prices of all listings in the same neighborhood or within a close proximity to determine the price for their own listing. Therefore, *neighbourhood\_group* is a useful feature that will help us understand more about the distribution of the target variable. We visualize the *neighbourhood\_group* vs *price* distribution in Figure 3. According to the figure, Manhattan and Brooklyn tend to have higher prices than the other boroughs. In addition, on Figure 4, the listings in Manhattan and Brooklyn are the most densely populated areas in New York.

We continue to perform data visualization between *longitude/latitude* and *price*, which describes how listing price differs according to geographical location. Figure 5 displays the *longitude* and *latitude* of all listings in New York. The left plot shows all listings divided into each

borough, while the right plot separates all observations according to 4 price groups. We would like to compare these two plots to discover patterns about the relationship between location and price. It can be observed that Manhattan (**areas around longitude -70.4 and latitude 40.75**) has more listings in the range \$200-\$300 and over \$300, while other boroughs have most listings less than \$200. This shows that there exists a relationship between geographical locations and price that we could exploit later when fitting statistical learning models.

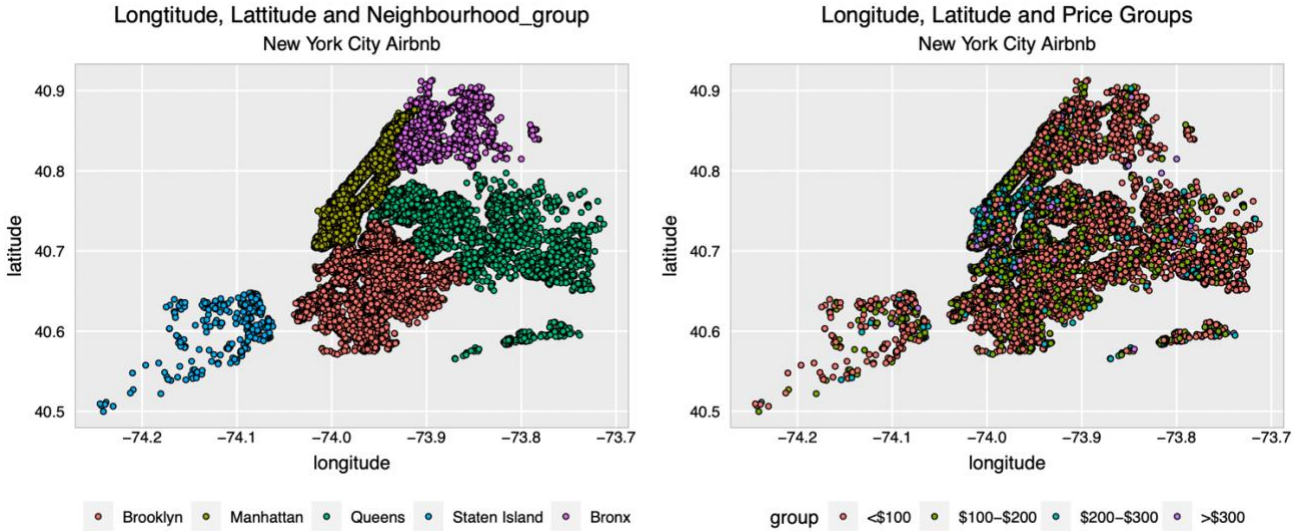


Figure 5: Latitude, Longitude and Price

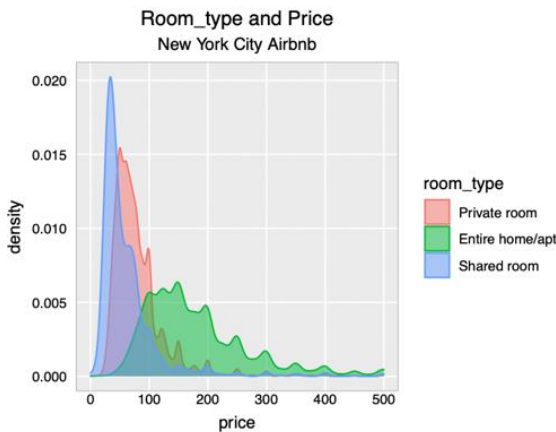


Figure 6: Price and Room\_Type

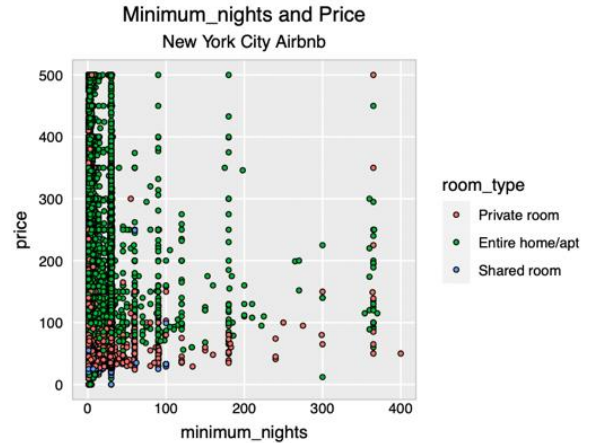


Figure 7: Price and Minimum\_nights

Next, we would like to explore how price changes according to the types of room being offered at an Airbnb listing (variable “*room\_type*”). From Figure 6, it can be seen that the cost for an entire home/apartment is the highest, while a shared room is the best solution for tenants who want to save money. It is very clear to conclude that the *room\_type* feature definitely has a strong effect on rental price. Apart from room type, when a tenant is looking for a listing, they also need to check the requirement about the minimum number of days they need to rent. In our data set, this requirement is represented as the “*minimum\_nights*” variable. According to Figure 7, it can be observed that listings with lower number of *minimum\_nights* tend to charge customers with higher prices. In other words, the price for a short-term lease is mostly higher than that of a long-term lease requirement. This shows that *minimum\_nights* definitely has a relationship with price that would allow for more accurate prediction when fitting models.

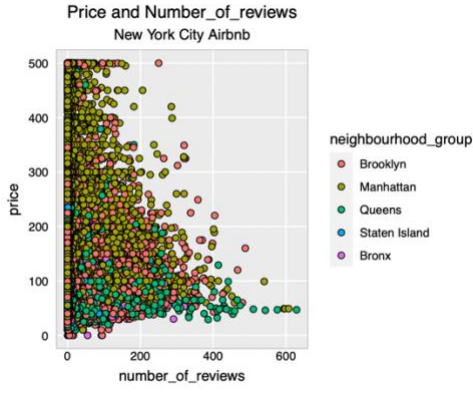


Figure 8: Price and Number\_of\_reviews



Figure 9: Price and Reviews\_per\_month

Next, we would like to explore the *number\_of\_reviews* and *reviews\_per\_month* variables because we think they are important factors that contribute to the price. Now, look at our data set, Figure 8 and 9 show that the number of reviews and monthly number of reviews tend to be higher for the listings in a low-price group. This makes sense because when a listing price is lower, it will attract more tenants and thus more reviews. Another feature we want to explore is *availability\_365*, which is the number of available days through the year 2019 that customers can book at a listing. It can be seen from Figure 10 that the higher the prices are, the more available days the listings have. This makes sense since more expensive listings tend to have fewer tenants throughout the year.

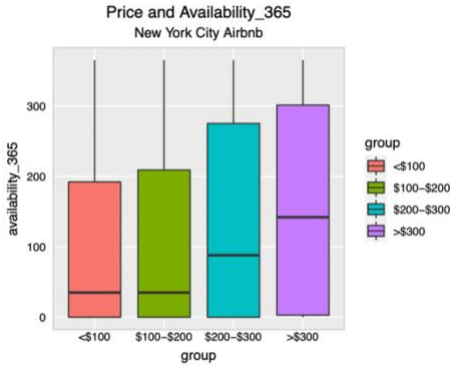


Figure 10: Price and Availability\_365

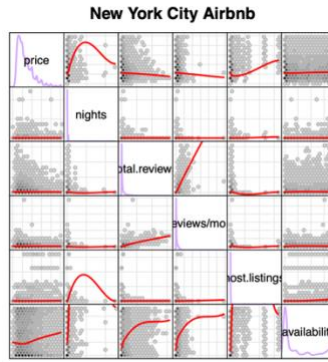


Figure 11: Scatter plot matrix

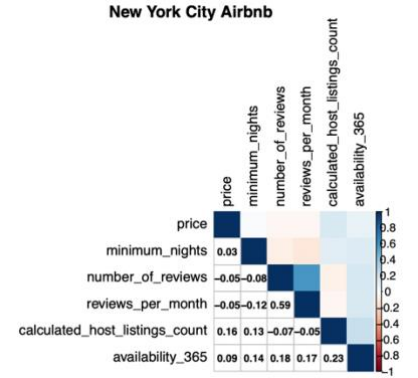


Figure 12: Correlation matrix

Finally, we would like to use the scatterplot matrix with hexagon binning and smooth lines, and correlation matrix to visualize the bivariate relationship between pairs of variables. According to Figure 11, there exists a highly correlated relationship between *number\_of\_reviews* and *reviews\_per\_month*. Also, the correlation matrix (Figure 12) shows that the correlation coefficient between these two variables is very high (0.59). This is an indicator for us to explore whether there is an interaction effect between number of reviews and monthly number of reviews.

#### 4. Proposed Model

In this section, we investigate different statistical learning approaches, namely multiple linear regression and random forest, to predict the price of listings as well as to determine the main contributing factors to these listing expenses. We start with an elementary baseline linear regression model that uses most of the provided variables. Upon analysis of the baseline model, we further perform additional techniques such as incorporating interaction between different terms, transforming textual features into numeric values, and categorizing numeric features into factors so as to better assist these statistical learning methods in modeling the relationship between the provided features with the Airbnb listing price.

##### 4.1. Multiple linear regression



**Model 1 (baseline):** Price ~ neighbourhood\_group + latitude + longitude + room\_type + minimum\_nights + number\_of\_reviews + last\_review + reviews\_per\_month + calculated\_host\_listings\_count + availability\_365.

This is a baseline model that simply uses most of the features provided in the dataset in their original format (excluding *listing\_id*, *name*, *host\_id*, *host\_name* and *neighbourhood*). The summary output of this baseline model is shown in Figure 13.

**The 10-fold cross validation RMSE is 67.3157.** Very fortunately, from the output, we can see that all of the variables appear to be significant for the linear regression model in making price prediction. This is parallel with our observations from Section 3 that these variables should have a major effect on the listing price. As we seek to reduce the RMSE even more, we employ additional techniques to dive deeper into some variables which we believe would help predicting price more accurately.

```
Call:
lm(formula = price ~ ., data = Train1)

Residuals:
    Min       1Q   Median       3Q      Max
-210.65  -39.14  -11.32   20.57   440.09

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.118e+04  1.067e+03  -19.840 < 2e-16 ***
neighbourhood_groupManhattan  3.771e+01  1.220e+00   30.917 < 2e-16 ***
neighbourhood_groupQueens    1.609e+01  1.479e+00   10.879 < 2e-16 ***
neighbourhood_groupStaten Island -8.895e+01  4.380e+00  -20.306 < 2e-16 ***
neighbourhood_groupBronx     1.218e+01  2.895e+00    4.207 2.59e-05 ***
latitude        -9.367e+01  1.041e+01   -8.994 < 2e-16 ***
longitude       -3.390e+02  1.200e+01  -28.237 < 2e-16 ***
room_typeEntire home/apt     8.821e+01  7.196e-01   122.577 < 2e-16 ***
room_typeShared room       -2.745e+01  2.261e+00  -12.145 < 2e-16 ***
minimum_nights  -3.002e-01  1.787e-02  -16.798 < 2e-16 ***
number_of_reviews -9.157e-02  9.667e-03   -9.473 < 2e-16 ***
last_review     -8.666e-03  4.591e-04  -18.878 < 2e-16 ***
reviews_per_month  9.778e-01  2.789e-01    3.506 0.000455 ***
calculated_host_listings_count  6.294e-02  1.118e-02    5.632 1.80e-08 ***
availability_365   9.087e-02  2.843e-03   31.966 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67.28 on 38265 degrees of freedom
Multiple R-squared:  0.4169,    Adjusted R-squared:  0.4166
F-statistic: 1954 on 14 and 38265 DF,  p-value: < 2.2e-16
```

Figure 13: Summary output of Model 1

There are four ideas to improve the baseline model accuracy:

**1. minimum\_nights:** According to Figure 20 (in the Appendix), the distribution of *minimum\_nights* is highly skewed to the left and concentrating within the range from 1 to 10 (only 13.5% of the listings require customers over 10 minimum nights). If we directly use the original *minimum\_nights* numbers, then rare observations (for example, there is only one listing requiring at least 1000 nights) will become an outlier which can have an influence on the model output and accuracy. Hence, it would be better to convert this feature from numeric data to a factor with 2 levels: the first level (“short”) indicates a short-term lease which applies for all listings with minimum nights less than or equal to 10; the second level (“long”) describes a long-time lease and applies to all listings where customers have to spend at least 11 nights.

**2. reviews\_per\_month & num\_of\_reviews:** Both these features should have an impact on price prediction since they describe whether a listing is attractive or judged as bad by many people. This is clearly shown in the summary output above where we can see both these variables having very significant p-values. However, since we find from Figure 12 that these two variables are highly correlated, we suspect that by further exploring their relationship (such as taking a division between *num\_of\_reviews* by *reviews\_per\_month* to attain the number of months a listing has been put on), we could facilitate more improvement on our linear regression model.

**3. name:** We make an assumption that listing name also helps in price prediction for listings on Airbnb. This stems from the observation that listing name possibly contains keywords that describe particular features about the listing that are not covered by the other variables (e.g., “spacious”,

“sunny”, “times square”). Therefore, we propose to convert listing name from its textual form to numeric values so that we can fit the linear regression model and study its effect on price prediction. The idea is to extract the most frequent keywords and count the occurrence of them in the listing names.

**4. Latitude and longitude:** We believe geographical features should have a major effect on prices. For example, listings at Manhattan are much more expensive than most of the other places. Geographical locations are made up of two variables - *latitude* and *longitude*. This means that using any of the 2 features alone is not sufficient to model geographical locations. Therefore, we propose to add an interaction term for latitude and longitude so that the model can better capture these geographical patterns.

We will employ these 4 ideas in the following models below.

#### **Model 2: Baseline + Convert minimum\_nights (variable transformation)**

Here, we will implement the first idea for model 2, where we combine the baseline model and the transformation of *minimum\_nights* feature. Now *minimum\_nights* is converted from its numeric format to a factor with 2 levels (“short” and “long”) denoting whether a listing booking requirement is short-term or long-term. The summary output of this model is shown in Figure 21 (in the Appendix).

**The 10-fold cross validation RMSE is 66.6676**, which is lower than the baseline model. This shows that by denoting *minimum\_nights* as short-term or long-term, we are able to effectively reduce the RMSE and produce a more accurate price prediction model. This can be intuitively explained as following: when a host puts their listing booking requirement as short-term on the lease market, it means they offer greater flexibility for tenants but at the cost of a higher renting expense. Therefore, it makes sense for the short/long-term requirement to have an impact on rental price.

#### **Model 3: Baseline + Convert minimum\_night + reviews\_per\_month/number of reviews (interaction effect)**

We would like to investigate the interaction effect between *reviews\_per\_month* and *number\_of\_reviews* in model 3 because these two variables have a strong correlation as demonstrated in Figure 12. Instead of using a multiplication term for these 2 variables, we divide the *number\_of\_reviews* by the *reviews\_per\_month* variable to get the number of months a listing has been publicly put on the Airbnb platform. Our assumption is that the longer a listing has been available for renting on Airbnb, the more reliable and well-received it is for many customers (since otherwise, the host would no longer be able to rent it out to tenants). The summary output is shown in Figure 22 (in the Appendix) after adding: **month = reviews\_per\_month/number\_of\_reviews**.

**The 10-fold cross validation RMSE is 66.66339**, which is only slightly lower than the previous model. Nonetheless, the p-value of *month* is significant, which shows that our assumption about its relationship with listing price is correct. However, the *reviews\_per\_month* variable is now no longer significant as its p-value has become higher. This is not a problem as explained in the ISLR book, page 89 (An Introduction to Statistical Learning with Application in R, 2015) where it says that as long as we have an interaction variable with a high significance level, it is not that necessary for the constituent term to be significant as well. Since the cross-validation RMSE of this model is still slightly better than that of the previous model, we decide to keep this interaction term.

#### Model 4: Model 3 + Convert name (variable transformation)

In the next model, we would like to investigate the effect of the name of listings on price. As explained above, the motivation for using listing name is that a host usually shows the main characteristics and distinctive features of a house in its name to attract customers. For example, when a host mentions “Times Square” in the name, they want to emphasize that their house has a good location which is in close proximity to one of the most famous places in New York. Other examples are “sunny”, “spacious”, and “renovated” which are attractive features that customers might be looking for. As we can see, all of the mentioned examples are important factors to determine the rental pricing. Therefore, it is necessary for us to analyze this feature to extract meaningful insights from it. However, it is not easy to perform this investigation if the name is displayed as textual data, so we need to convert it into numeric format.

The idea to perform this conversion is to create a set of the most frequent keywords, represented as 1-word keyword (unigram) or 2-word keyword (bigram), then record the number of occurrences of these keywords in a listing name. We will examine 3 approaches: (a) using unigram keywords, (b) using bigram keywords, and (c) using both unigram and bigram keywords.

**a. Unigram keywords:** from the listing names in the training set, we create a set  $S$  consisting of the top 500 most frequent unigram keywords. We then iterate over each listing name in the whole dataset, count the number of unigram keywords it has in its name and use that number as its feature rather than using its textual name. We show in Figure 14 the word cloud of this set  $S$ . We also show the summary output of this model in Figure 23 (in the Appendix).

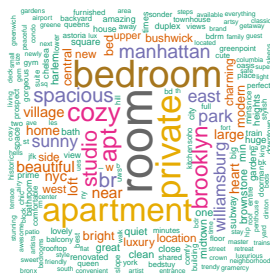


Figure 14: Unigram Wordcloud

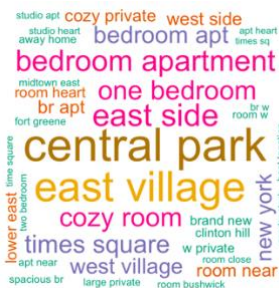


Figure 15: Bigram Wordcloud

**The 10-fold cross validation RMSE of this model is 66.64549**, which is a bit smaller than the previous model. This shows that utilizing unigrams of listing names could be a good direction to tackle but there is still more work to be done in order to make it really effective in predicting price. We also see that its p-value (the one corresponding to variable *unigram*) is significant (0.03). We continue to use bigram instead of unigram in the next model.

**b. Bigram keywords:** similarly to the above approach but this time, we use the top 100 bigram keywords (consisting of every two adjacent words) in the listing names. We show in Figure 15 the top 100 most frequent bigram keywords. We also show the summary output in Figure 24 (in the Appendix). **The 10-fold cross validation RMSE is 66.55733**, which is better than the unigram approach. Its p-value, the one corresponding to the *bigram* variable, is also significant ( $<2e-16$ ). This potentially shows that bigram is more effective than unigram in describing particular features of the listings. Most examples we can see from the word cloud belong to locational features (e.g., “central park”, “east village”, “midtown east”, “time square”), room features (e.g., “spacious br”, “large private”, “brand new”) that are not described using any of the other variables in the data set. Therefore, explicitly using these bigrams helps our linear regression model to better predict model listing price.



**c. Using both unigram and bigram:** we now try using both *unigram* and *bigram* keywords. We show the summary output in Figure 25 (in the Appendix). **The 10-fold cross validation RMSE is 66.46523**, which is smaller than that when using unigram or bigram alone. Both p-values of the *unigram* and *bigram* variables are also significant. Therefore, we decide to employ both *unigram* and *bigram* for our final model.

#### Model 5: Model 4c + latitude:longitude (interaction effect)

**The 10-fold cross validation RMSE is 66.23307**, which further decreases from the model above. This shows that adding the interaction term for *latitude* and *longitude* really helps making price prediction more accurate. Figure 26 (in the Appendix) shows the summary output, from which we can see that the interaction term is important since it has a very significant p-value. Here is the summary of results of all of our models above.

Model	Residual standard error	Multiple R-squared	Training RMSE	Cross-validation RMSE
Model 1 (Baseline)	67.28	0.4169	67.26961	67.3157
Model 2 = Baseline + Transform minimum_night	66.66	0.4276	66.64892	66.6676
Model 3 = Model 2 + number_of_reviews/reviews_per_month	66.66	0.4277	66.6434	66.66339
Model 4a = Model 3 + unigram	66.63	0.4281	66.62011	66.64549
Model 4b = Model 3 + bigram	66.55	0.4295	66.53595	66.55733
Model 4c = Model 3 + unigram + bigram	66.45	0.4312	66.43863	66.46523
Model 5 = Model 4c + longitude x latitude	66.22	0.4352	66.20204	66.23307

Table 1: The summary of results of all linear regression models

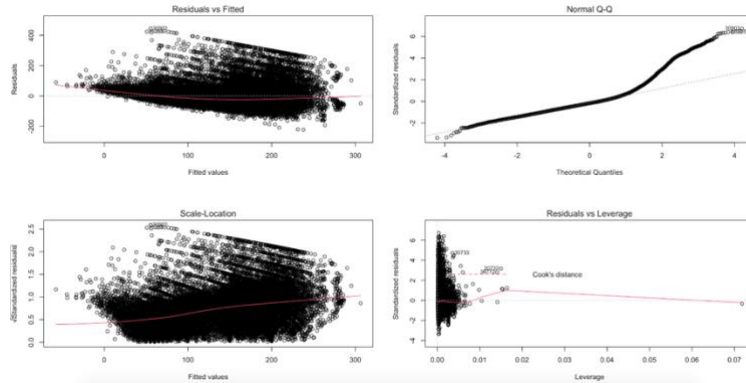


Figure 16: Diagnostic Plots of Model 5

Next, we would like to use four diagnostic plots of model 5 to check the linear model assumptions. These plots are presented in Figure 16. For the **Residuals vs Fitted** plot, we can see that the red line is very close to the horizontal line at 0. It indicates our model does not violate the mean of zero assumption. Also, the residuals do not appear to have any particular patterns, so we can confirm that there is a linear relationship between the predictors and the target. Another assumption we have to verify is the normality assumption that the residuals are normally distributed. According to the **Normal Q-Q** plot, it can be observed that the residuals deviate from the straight line. Additionally, the right standardized residuals tail is thicker than the right theoretical tail, which shows that the distribution of residuals does not comply with the normal distribution. For the third plot, **Scale-Location**, the red line is not actually straight, so we cannot confirm that the residuals have the same variance. The final plot, **Residuals vs Leverage**, is used to identify high leverage points that can have an influence on the model accuracy. Here, this plot of model 5 shows that there is only one point having high leverage, but it is still outside the Cook's

distance. Hence, there might be no influential cases in model 5. In summary, even though model 5 is our most optimal model, it still violates some linear model assumptions. Therefore, we would like to use a non-parametric method to predict the rental price in the next section.

#### 4.2. Random forest

Section 4.1. above shows that we can improve the baseline model by adding interaction effect and variable transformation as follows: (1) minimum\_nights transformation, (2) interaction effect of month (month = number\_of\_reviews/reviews\_per\_month), (3) unigram and bigram, (4) interaction effect of longitude x latitude. Therefore, we continue to apply these features for our random forest. For the hyperparameters, we choose the number of predictors at each split is 2 and the number of trees is 100. Figure 17 shows the summary output of this model and Figure 18 displays the important features from this approach. We can see that *room\_type* is the most important variable, which also matches with our linear regression models because the p-value of *room\_type* is always significant from the linear regression model 1 to model 5. Comparing the test RMSE between random forest and the final linear regression model, it can be seen that random forest outperforms the other one when using the same subset of features.

Model	Test RMSE
Linear Regression (Model 5)	66.76745
Random Forest	61.6417

Table 2: Result Comparison between Linear Regression - Model 5 and Random Forest Regression

```
Call:
randomForest(formula = price ~ . + latitude:longitude, data = Train6,
             mtry = 2, ntree = 100, importance = TRUE)
Type of random forest: regression
Number of trees: 100
No. of variables tried at each split: 2

Mean of squared residuals: 3760.97
% Var explained: 51.53
```

Figure 17: Summary Output of Random Forest

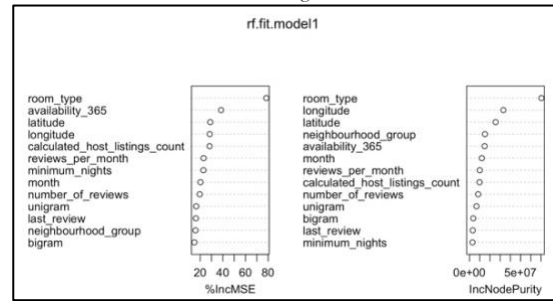


Figure 18: Variable Importance Plot

#### 5. Conclusion

In conclusion, based on both the random forest model and the final linear regression model, we would like conclude that the important factors which can determine the rental price in Airbnb New York market are:

- **neighbourhood\_group, longitude, latitude:** these features represent the location of a listing. For a place having top-rated tourist attractions like New York, visitors tend to book a place near landmarks or famous sites. Therefore, it's reasonable to confirm that location is one of the most influential components to determine prices.
- **room\_type, minimum\_nights:** this is true because the more flexibility the listing offers, the higher payment tenants have to pay.
- **number\_of\_reviews:** this is important because it reflects not only how well-received the listing is but also the reliability.

Even though the random forest approach produces a better prediction result than the linear regression model, we still choose to put more effort into investigating the linear models. The reason is that we prefer inference to prediction. Our main goal is to understand the relationship between price and predictors and study what factors contribute to the rental market. Hence, we choose a restrictive model like linear regression because it would be easily interpretable. That is a trade-off between prediction accuracy and interpretability (An Introduction to Statistical Learning with Application in R, 2015).

# REFERENCES

- (2015). In G. James, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning with Application in R* (p. 89). Springer Texts in Statistics.
- (2015). In G. James, D. Witten, T. Hastie, & R. Tibshirani, *An Introduction to Statistical Learning with Application in R* (p. 24). Springer Texts in Statistics.
- (2019). Retrieved from Kaggle: <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

## ADDITIONAL FIGURES

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude
1	2539 Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237
2	2595 Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377
3	3647 THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190
4	3831 Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976
5	5022 Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399
6	5099 Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.97500
7	5121 BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596
8	5178 Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493
9	5203 Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723
10	5238 Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037

room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
Private room	149	1	9	2018-10-19	0.21	6	365
Entire home/apt	225	1	45	2019-05-21	0.38	2	355
Private room	150	3	0		NA	1	365
Entire home/apt	89	1	270	2019-07-05	4.64	1	194
Entire home/apt	80	10	9	2018-11-19	0.10	1	0
Entire home/apt	200	3	74	2019-06-22	0.59	1	129
Private room	60	45	49	2017-10-05	0.40	1	0
Private room	79	2	430	2019-06-24	3.47	1	220
Private room	79	2	118	2017-07-21	0.99	1	0
Entire home/apt	150	1	160	2019-06-09	1.33	4	188

Figure 19: An example of how the original data set looks like

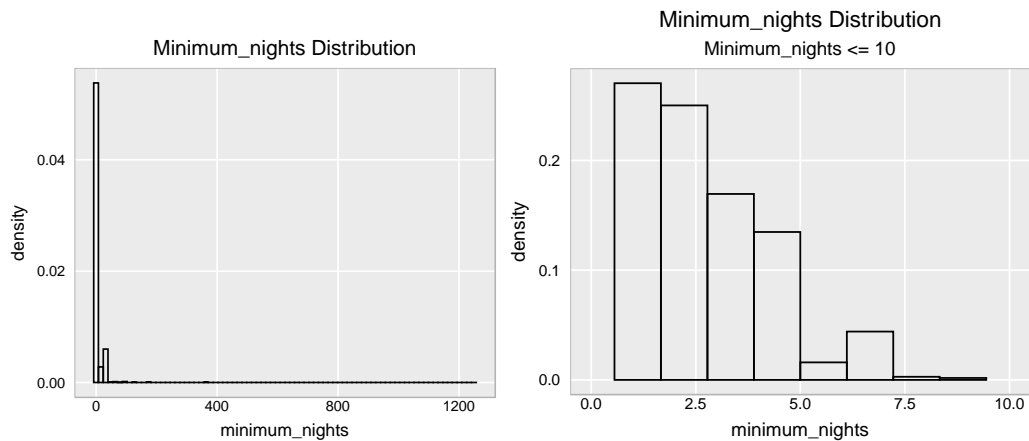


Figure 20: Distribution of minimum\_nights

```

Call:
lm(formula = price ~ ., data = Train2)

Residuals:
    Min       1Q   Median       3Q      Max
-221.24  -39.08  -11.01   21.21  442.61

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.189e+04  1.058e+03  -20.698 < 2e-16 ***
neighbourhood_groupManhattan  3.779e+01  1.209e+00  31.271 < 2e-16 ***
neighbourhood_groupQueens    1.590e+01  1.466e+00  10.852 < 2e-16 ***
neighbourhood_groupStaten Island -9.155e+01  4.341e+00  -21.089 < 2e-16 ***
neighbourhood_groupBronx     9.401e+00  2.870e+00   3.276 0.00105 **
latitude        -8.658e+01  1.032e+01  -8.388 < 2e-16 ***
longitude       -3.443e+02  1.189e+01 -28.950 < 2e-16 ***
room_typeEntire home/apt     8.920e+01  7.137e-01  124.973 < 2e-16 ***
room_typeShared room        -2.772e+01  2.240e+00  -12.376 < 2e-16 ***
minimum_nightsshort         3.493e+01  1.103e+00  31.678 < 2e-16 ***
number_of_reviews    -9.957e-02  9.582e-03  -10.392 < 2e-16 ***
last_review        -9.646e-03  4.563e-04  -21.139 < 2e-16 ***
reviews_per_month     2.039e-01  2.778e-01   0.734 0.46308
calculated_host_listings_count 1.222e-01  1.128e-02  10.836 < 2e-16 ***
availability_365       1.056e-01  2.869e-03  36.812 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.66 on 38265 degrees of freedom
Multiple R-squared:  0.4276,    Adjusted R-squared:  0.4274
F-statistic: 2042 on 14 and 38265 DF,  p-value: < 2.2e-16

```

Figure 21: Summary Output of Model 2

```

Call:
lm(formula = price ~ ., data = Train3)

Residuals:
    Min       1Q   Median       3Q      Max
-221.37  -39.13  -11.00   21.21  443.00

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.202e+04  1.059e+03  -20.795 < 2e-16 ***
neighbourhood_groupManhattan  3.775e+01  1.209e+00  31.237 < 2e-16 ***
neighbourhood_groupQueens    1.589e+01  1.465e+00  10.841 < 2e-16 ***
neighbourhood_groupStaten Island -9.191e+01  4.343e+00  -21.162 < 2e-16 ***
neighbourhood_groupBronx     9.211e+00  2.870e+00   3.209 0.00133 **
latitude        -8.600e+01  1.032e+01  -8.331 < 2e-16 ***
longitude       -3.457e+02  1.191e+01 -29.037 < 2e-16 ***
room_typeEntire home/apt     8.930e+01  7.148e-01  124.925 < 2e-16 ***
room_typeShared room        -2.784e+01  2.240e+00  -12.428 < 2e-16 ***
minimum_nightsshort         3.486e+01  1.103e+00  31.599 < 2e-16 ***
number_of_reviews    -7.948e-02  1.247e-02  -6.374 1.86e-10 ***
last_review        -8.818e-03  5.624e-04  -15.680 < 2e-16 ***
reviews_per_month     -2.763e-01  3.370e-01  -0.820 0.41220
calculated_host_listings_count 1.200e-01  1.131e-02  10.609 < 2e-16 ***
availability_365       1.058e-01  2.870e-03  36.862 < 2e-16 ***
month           -5.880e-02  2.335e-02  -2.518 0.01181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.66 on 38264 degrees of freedom
Multiple R-squared:  0.4277,    Adjusted R-squared:  0.4274
F-statistic: 1906 on 15 and 38264 DF,  p-value: < 2.2e-16

```

Figure 22: Summary Output of Model 3

```

Call:
lm(formula = price ~ ., data = Train4)

Residuals:
    Min       1Q   Median       3Q      Max
-223.52  -39.00  -11.04   21.12  443.41

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.156e+04  1.062e+03  -20.301 < 2e-16 ***
neighbourhood_groupManhattan  3.770e+01  1.208e+00  31.200 < 2e-16 ***
neighbourhood_groupQueens    1.570e+01  1.465e+00  10.717 < 2e-16 ***
neighbourhood_groupStaten Island -9.086e+01  4.346e+00  -20.904 < 2e-16 ***
neighbourhood_groupBronx     9.777e+00  2.872e+00   3.405 0.000663 ***
latitude       -8.744e+01  1.032e+01  -8.470 < 2e-16 ***
longitude      -3.403e+02  1.195e+01  -28.479 < 2e-16 ***
room_typeEntire home/apt     8.921e+01  7.148e-01  124.804 < 2e-16 ***
room_typeShared room        -2.733e+01  2.242e+00  -12.192 < 2e-16 ***
minimum_nightsshort    3.513e+01  1.104e+00  31.822 < 2e-16 ***
number_of_reviews   -8.042e-02  1.247e-02  -6.451 1.13e-10 ***
last_review        -8.986e-03  5.631e-04  -15.958 < 2e-16 ***
reviews_per_month    -3.188e-01  3.369e-01  -0.946 0.344059
calculated_host_listings_count  1.132e-01  1.138e-02   9.947 < 2e-16 ***
availability_365      1.065e-01  2.872e-03  37.088 < 2e-16 ***
month             -5.044e-02  2.340e-02  -2.156 0.031122 *
unigram           1.150e+00  2.222e-01   5.173 2.32e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.63 on 38263 degrees of freedom
Multiple R-squared:  0.4281, Adjusted R-squared:  0.4278
F-statistic: 1790 on 16 and 38263 DF, p-value: < 2.2e-16

```

Figure 23: Summary Output of Model 4a (unigram)

```

Call:
lm(formula = price ~ ., data = Train5)

Residuals:
    Min       1Q   Median       3Q      Max
-216.12  -39.20  -10.89   21.50  443.78

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.243e+04  1.058e+03  -21.209 < 2e-16 ***
neighbourhood_groupManhattan  3.882e+01  1.210e+00  32.067 < 2e-16 ***
neighbourhood_groupQueens    1.619e+01  1.463e+00  11.066 < 2e-16 ***
neighbourhood_groupStaten Island -9.372e+01  4.339e+00  -21.599 < 2e-16 ***
neighbourhood_groupBronx     9.083e+00  2.866e+00   3.169 0.00153 **
latitude       -8.748e+01  1.031e+01  -8.487 < 2e-16 ***
longitude      -3.522e+02  1.190e+01  -29.594 < 2e-16 ***
room_typeEntire home/apt     8.843e+01  7.180e-01  123.162 < 2e-16 ***
room_typeShared room        -2.967e+01  2.243e+00  -13.231 < 2e-16 ***
minimum_nightsshort    3.469e+01  1.101e+00  31.490 < 2e-16 ***
number_of_reviews   -7.778e-02  1.245e-02  -6.247 4.22e-10 ***
last_review        -8.622e-03  5.618e-04  -15.348 < 2e-16 ***
reviews_per_month    -2.612e-01  3.364e-01  -0.776 0.43748
calculated_host_listings_count  1.144e-01  1.130e-02  10.124 < 2e-16 ***
availability_365      1.040e-01  2.870e-03  36.234 < 2e-16 ***
month             -6.822e-02  2.333e-02  -2.924 0.00346 **
bigram          -4.454e+00  4.004e-01  -11.122 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.55 on 38263 degrees of freedom
Multiple R-squared:  0.4295, Adjusted R-squared:  0.4293
F-statistic: 1800 on 16 and 38263 DF, p-value: < 2.2e-16

```

Figure 24: Summary Output of Model 4b (bigram)



```

Call:
lm(formula = price ~ ., data = Train6)

Residuals:
    Min       1Q   Median       3Q      Max
-217.61  -39.00  -10.91   21.17  445.04

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.159e+04  1.059e+03  -20.384 < 2e-16 ***
neighbourhood_groupManhattan  3.914e+01  1.209e+00   32.369 < 2e-16 ***
neighbourhood_groupQueens    1.592e+01  1.462e+00   10.891 < 2e-16 ***
neighbourhood_groupStaten Island -9.214e+01  4.335e+00  -21.252 < 2e-16 ***
neighbourhood_groupBronx     1.029e+01  2.864e+00   3.593 0.000327 ***
latitude        -9.132e+01  1.030e+01   -8.867 < 2e-16 ***
longitude       -3.428e+02  1.192e+01  -28.765 < 2e-16 ***
room_typeEntire home/apt     8.786e+01  7.189e-01  122.209 < 2e-16 ***
room_typeShared room        -2.930e+01  2.240e+00  -13.082 < 2e-16 ***
minimum_nightsshort         3.523e+01  1.101e+00   31.994 < 2e-16 ***
number_of_reviews    -7.916e-02  1.243e-02   -6.367 1.95e-10 ***
last_review        -8.915e-03  5.616e-04  -15.874 < 2e-16 ***
reviews_per_month    -3.497e-01  3.360e-01   -1.041 0.297982
calculated_host_listings_count  9.699e-02  1.141e-02    8.502 < 2e-16 ***
availability_365       1.049e-01  2.867e-03   36.588 < 2e-16 ***
month              -5.353e-02  2.334e-02   -2.294 0.021819 *
bigram             -6.325e+00  4.372e-01  -14.468 < 2e-16 ***
unigram            2.566e+00  2.423e-01   10.591 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.45 on 38262 degrees of freedom
Multiple R-squared:  0.4312,    Adjusted R-squared:  0.4309
F-statistic: 1706 on 17 and 38262 DF,  p-value: < 2.2e-16

```

Figure 25: Summary Output of Model 4c (unigram+bigram)

```

Call:
lm(formula = price ~ . + latitude:longitude, data = Train6)

Residuals:
    Min       1Q   Median       3Q      Max
-222.78  -39.01  -11.28   21.11  444.62

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.093e+06  4.903e+05  16.508 < 2e-16 ***
neighbourhood_groupManhattan  3.035e+01  1.317e+00   23.052 < 2e-16 ***
neighbourhood_groupQueens    1.775e+01  1.461e+00   12.154 < 2e-16 ***
neighbourhood_groupStaten Island -5.330e+01  4.916e+00  -10.841 < 2e-16 ***
neighbourhood_groupBronx     2.788e+01  3.045e+00   9.154 < 2e-16 ***
latitude       -1.995e+05  1.205e+04  -16.559 < 2e-16 ***
longitude       1.094e+05  6.632e+03   16.500 < 2e-16 ***
room_typeEntire home/apt     8.753e+01  7.167e-01  122.142 < 2e-16 ***
room_typeShared room        -2.877e+01  2.232e+00  -12.890 < 2e-16 ***
minimum_nightsshort         3.555e+01  1.097e+00   32.399 < 2e-16 ***
number_of_reviews    -7.784e-02  1.239e-02   -6.283 3.36e-10 ***
last_review        -8.700e-03  5.598e-04  -15.542 < 2e-16 ***
reviews_per_month    -3.724e-01  3.348e-01   -1.112 0.26609
calculated_host_listings_count  1.005e-01  1.137e-02    8.842 < 2e-16 ***
availability_365       1.055e-01  2.857e-03   36.925 < 2e-16 ***
month              -6.089e-02  2.326e-02   -2.618 0.00885 **
bigram           -6.317e+00  4.356e-01  -14.503 < 2e-16 ***
unigram           2.447e+00  2.416e-01   10.129 < 2e-16 ***
latitude:longitude    -2.698e+03  1.630e+02  -16.552 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.22 on 38261 degrees of freedom
Multiple R-squared:  0.4352,    Adjusted R-squared:  0.435
F-statistic: 1638 on 18 and 38261 DF,  p-value: < 2.2e-16

```

Figure 26: Summary Output of Model 5