Project: NLP analysis of Airbnb Reviews

Goals:

The objecting of our project to predict ratings of Airbnb listings in New York Area using Sentiment analysis of actual reviews written by users who have used Airbnb service in New York area. We are trying to come-up with a model, which can predict the overall rating of listings based on the score generated by our sentiment analysis engine. The other interesting things to see here would be if the reviews scores could predict the frequency of booking of a listing.

About Data:

The data obtained is very rich in both variables and observations. We are using two different types of datasets namely listings and reviews. The listings datasets have different variables related to Airbnb listings and Reviews datasets have different variable related to reviews written against the listings.

Feature Engineering:

The review dataset have few variables available whereas the listings have a large number of variables. Since the project needs data from both the datasets, it is important to select feature data variables from both datasets, which would help us make better predictions.

Some of the important features in the Listings datasets are:

Name	description
id	Unique identifier for each listing
host_id	Unique identifier for each host
host_since	Joining data of host
host_response_rate	Response time of host
host_acceptance_rate	Host acceptance rate for each booking
neighbourhood_cleansed	Neighborhood of listing
city	City where listing is located
property_type	Type of Listings
price	Price of each listing
number_of_reviews	Number of reviews listed
first_review	When was the first review written for each listing
last_review	When was the last review written for each listing
review_scores_rating	Overall listing of review out of 100.
review_scores_accuracy	Overall listing accuracy score out of 10
review_scores_cleanliness	Overall listing cleanliness score out of 10
review_scores_checkin	Overall listing check-in score out of 10
review_scores_location	Overall listing location score out of 10
review_scores_communication	Overall communication score out of 10
reviews_per_month	Average review per month

Some of the important variables in the Reviews datasets are:

Name	Description
listing_id	Unique Id for each listing
comments	Reviews for each listing

Data cleaning:

Raw data obtained for this project has a very well-defined structure. We have taken different approaches to clean to data. Apart from converting the data variables into the desired format for data analysis, we have taken steps to remove NA values from important variables.

Data Analysis:

We performed some basic data analysis first to answer some of the basic questions.

How many Reviews users have written by the users:

```
[1] 664323
```

How many Listings are listed in New York Area:

```
[1] 40753
```

How many Neighborhood have listings in New York Area:

```
[1] 217
```

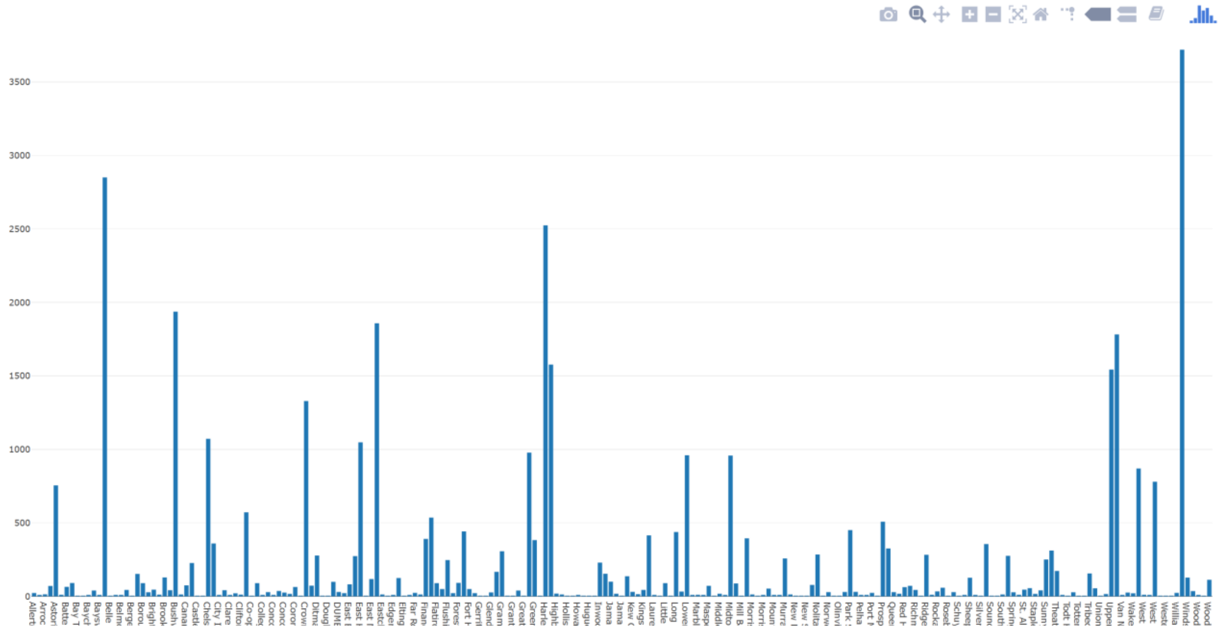
Analysis

- From above analysis, we can see that there are ~**40k** unique listings listed on Airbnb in New York area and have ~**664k** reviews against them. These listings are distributed over **217** different neighborhood.

Let us explore the distribution of Listings and Reviews in New York City

Part 1: Listings

Let us see how the listings are distributed by neighborhoods.



Analysis:

Above graph shows the same data in different formats. I have used plotly to make the graph interactive so that it becomes analyze large categories of data.

From the above charts, we can conclude there is a large variation in distribution in listings over neighborhoods.

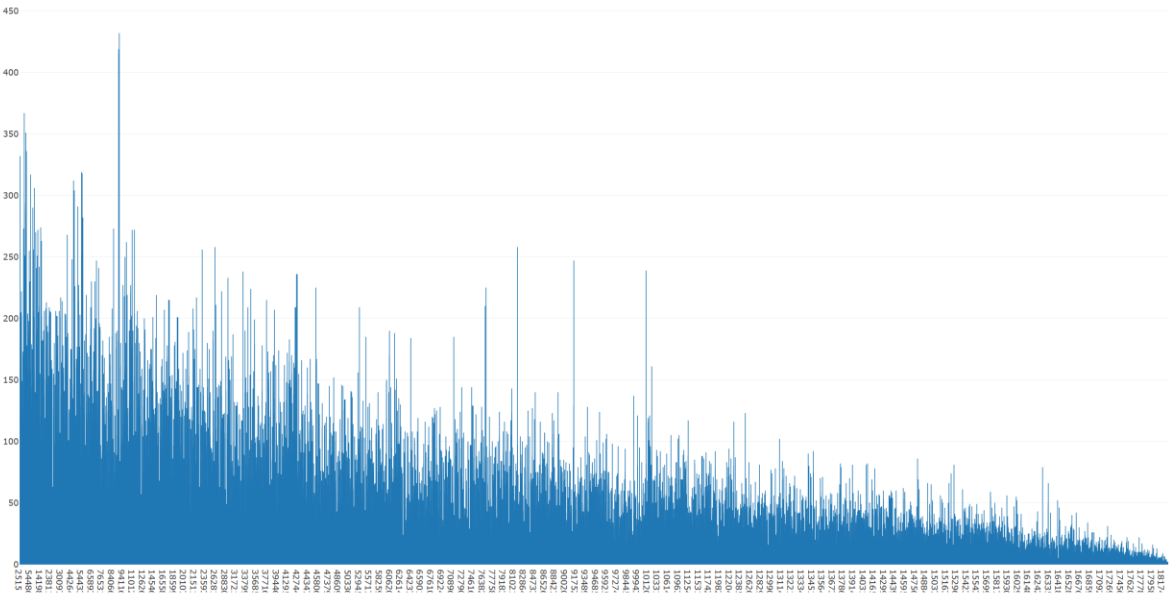
```
  >>Min. 1st Qu.  Median    Mean   3rd Qu.    Max.

       1.0     6.0       24.0       187.8     100.0     3719.0
```

With the use, of the summary function we can see the overview of the listings distribution. Interpretation of the result:

- One or more neighborhoods have 1 listing.
- One or more neighborhoods have 3719 listings.
- 25% of the neighborhoods have less than or equal to 6 listings.
- 50% of the neighborhoods have less than or equal to 24 listings.
- Average listings distributed over neighborhoods is 189.
- 75% of the neighborhoods have less than or equal to 100 listings.

Part 2: Reviews:

Let us see how many reviews each listing have against them:



```
Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
 1.00    3.00    8.00      21.12      25.00      432.00
```

Analysis:

Above graph shows the same data in different formats. I have used plotly to make the graph interactive so that it becomes analyze large categories of data.

From the above charts, we can conclude there is a large variation in distribution in a total number of reviews for each listing.

With the use, of the summary function we can see the overview of the listings distribution. Interpretation of the result:

- One or more listings have 1 review.
- One or more neighborhood has 432 reviews.
- 25% of the neighborhoods have less than or equal to 3 reviews.
- 50% of the neighborhoods have less than or equal to 8 reviews.
- An average number of reviews per listing is 21.
- 75% of the neighborhoods have less than or equal to 25 reviews.

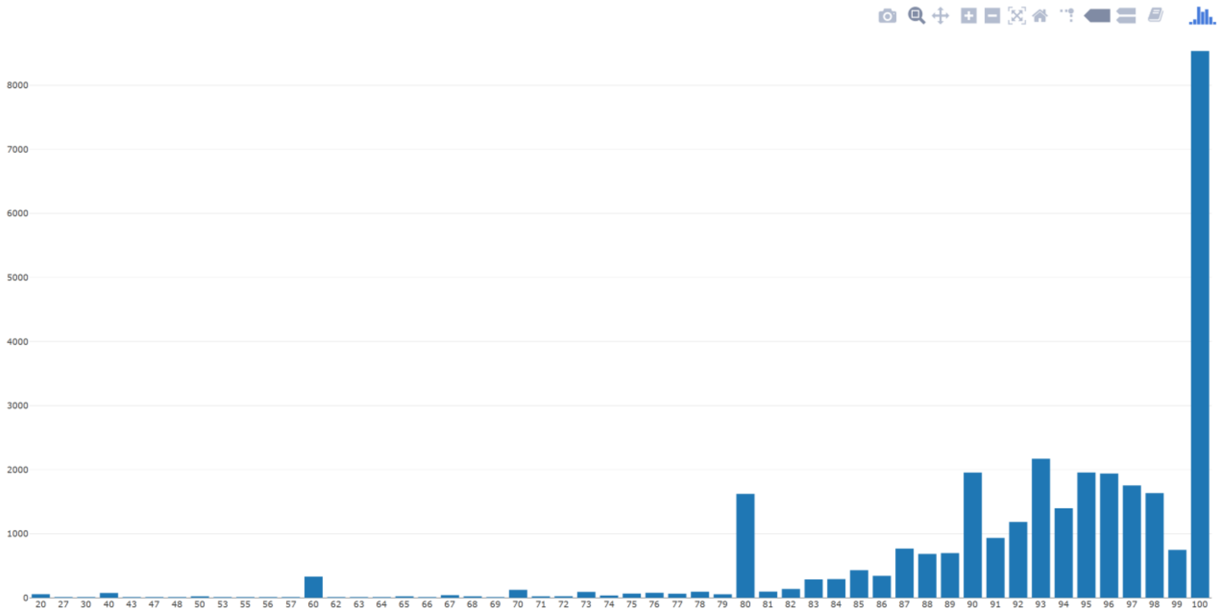
Part 3: Ratings:

Let us explore how the rating scores are distributed.

>>Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
20	**90**	**95**	**93**	**100**	**100**

With the use, of summary function we can see the overview of the listings distribution. Interpretation of the result:

- One or more listings has been rated with a score of 20.
- One or more listings has been rated with a score of 100.
- 25% of the listings have been rated with a score less than or equal to 90.
- 50% of the listings have been rated with a score of 95 or less.
- Average rating per review is 93.
- 25% of the listings have been rated with a score of 100 or less.

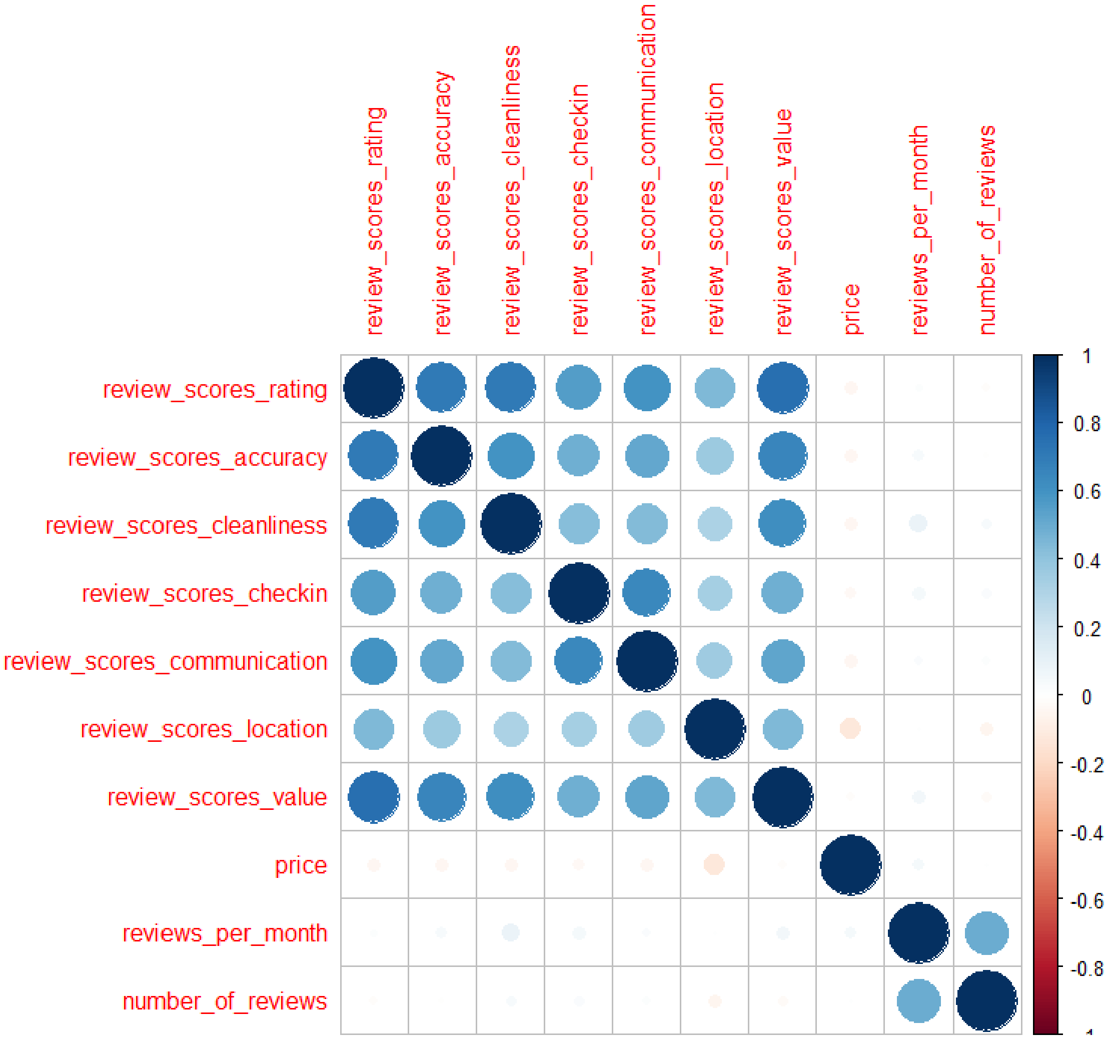


Analysis:

Above graph shows the same data in different formats. I have used plotly to make the graph interactive so that it becomes analyze large categories of data.

From the above charts, we can conclude very few listings are poorly rated and most of the listings rated with a score of 90 and more.

Feature Correlation:



The above correlation plot describes the correlation between some of the feature variables from the listings dataset. Some of the interesting conclusion, which can be derived from the plot, are:

- No is no correlation between price and variable like price, number of reviews, reviews per month and review scores rating.
- There is a significant correlation between different types of review_score variable.

Next Steps:

- Performing correlation analysis with scores obtained from Sentiment analysis.
- Using Random forest model to predict the reviews scores.