IBM DATA SCIENCE CERTIFICATION
CAPSTONE PROJECT

-Srinath Ramachandran

New York Airbnb House Price Prediction

Introduction:

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. It therefore, becomes very important to serve the customers with the best possible pricing based on the area of the house.

Keeping these points in mind, following business questions can be answered with the help of this analysis.

What can we learn about different hosts and areas? What can we learn from predictions? (ex: locations, prices, reviews, etc.)? What can we learn about the impact of nearby venues to the house location, do they impact pricing and so on.

In short, I am going to predict the price of a housing listed on Airbnb using all the information available with me. Also, I would be leveraging Foursquare API for getting the list of venues in a radius of 500 meters of a particular Airbnb house.

Data:

The data was obtained from Kaggle https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data. This dataset describes the listing activity and metrics in NYC, NY for 2019. Also, it includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Data Description:

Number of Columns: 16

1. idlisting ID
2. namename of the listing
3. host_idhost ID
4. host_namename of the host
5. neighbourhood_grouplocation
6. neighbourhoodarea
7. latitudelatitude coordinates
8. longitudelongitude coordinates
9. room_typelisting space type
10. priceprice in dollars
11. minimum_nightsamount of nights minimum
12. number_of_reviewsnumber of reviews
13. last_reviewlatest review
14. reviews_per_monthnumber of reviews per month
15. calculated_host_listings_countamount of listing per host

-Srinath Ramachandran

16. availability_365number of days when listing is available for booking

Number of Rows: 48,895

Methodology:

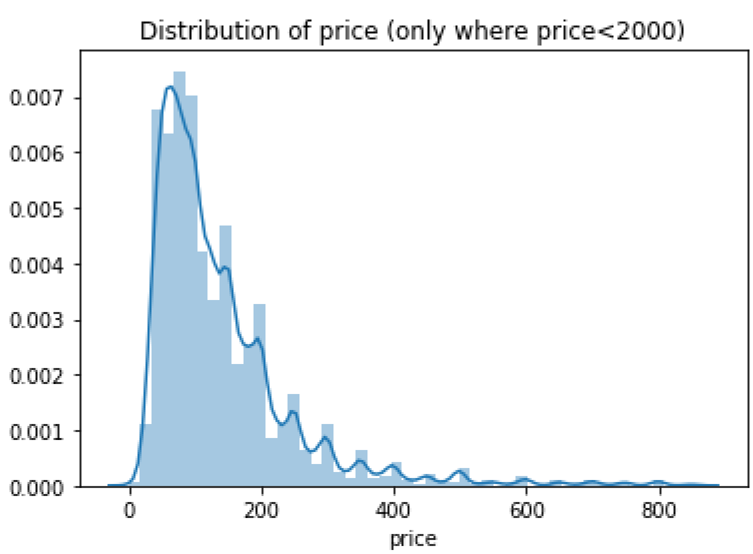
Exploratory Data Analysis:

- **To find out the numerical statistics of the dataset:**

	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

From this chart, it can be found out there are many outliers that need to be handled before proceeding ahead.

- **Histogram showing Distribution of Price:**

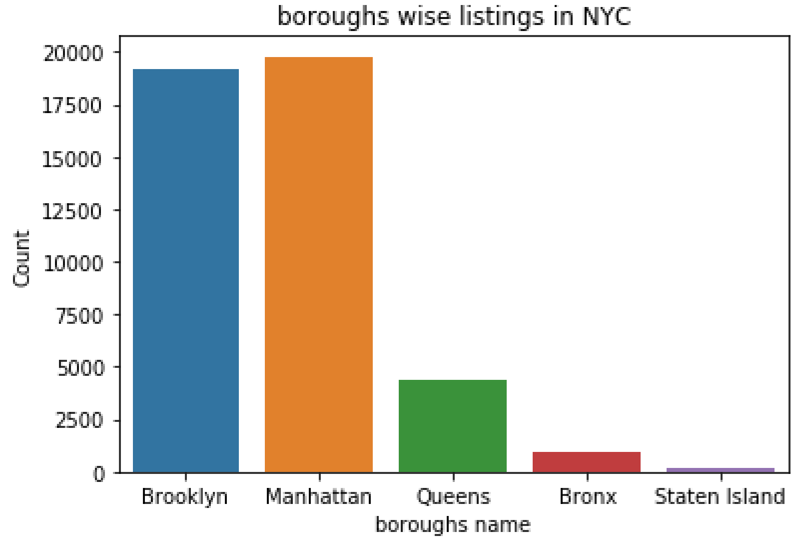


From the above histogram, it can be observed that price of most listings are between $10 to $200.

IBM DATA SCIENCE CERTIFICATION
CAPSTONE PROJECT

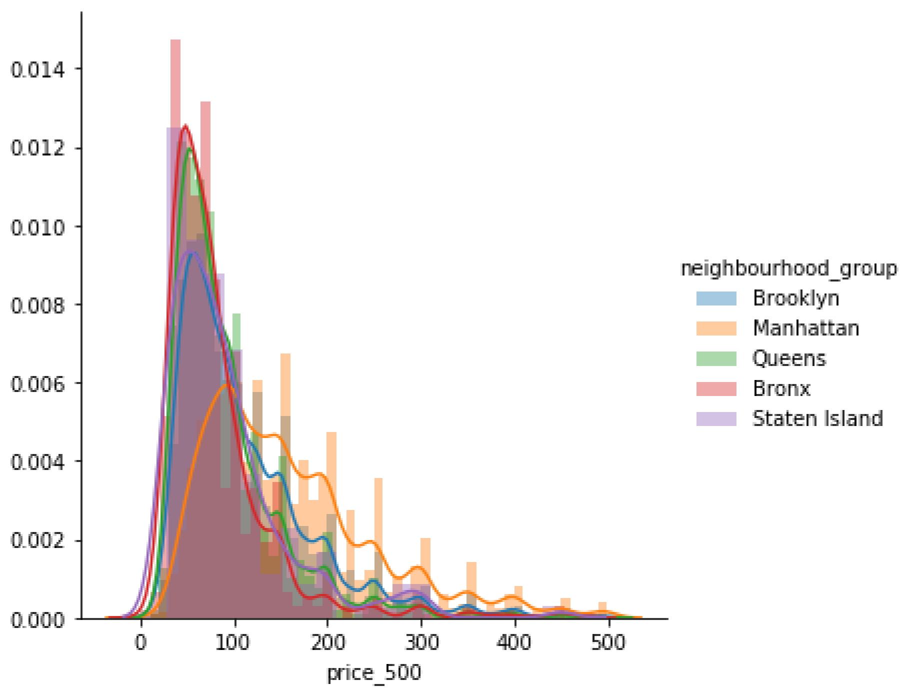
-Srinath Ramachandran

- **Borough-wise listings in NYC:**



From the above chart, it can be clearly observed that Brooklyn and Manhattan dominate the borough-wise listings in NYC.

- **Graph comparing the prices(<500) in different boroughs.**



The price of rooms in Manhattan and Brooklyn is more than the price of rooms in other boroughs.

After some exploratory analysis, I tried to count the number of neighbourhood venues (within 500m range) using Foursquare API.

IBM DATA SCIENCE CERTIFICATION CAPSTONE PROJECT

-Srinath Ramachandran

NOTE: I was able to find out only the number of venues in the neighborhood of houses in Bronx borough. My further analysis is thus only on Bronx borough. However, this idea can be extended to other boroughs as well.

Results:

- **Dataframe after including the number of venues.**

	index	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	Number of Venues(500m range)
0	207	Bronx	Highbridge	40.83075	-73.93058	Private room	45	1	138	1.45	3	323	35
1	260	Bronx	Clason Point	40.81309	-73.85514	Private room	90	2	0	0.00	7	349	11
2	261	Bronx	Eastchester	40.88057	-73.83572	Entire home/apt	105	2	38	0.50	13	365	13
3	309	Bronx	Kingsbridge	40.87207	-73.90193	Entire home/apt	90	30	4	0.35	2	346	25
4	484	Bronx	University Heights	40.85811	-73.90675	Private room	37	4	117	1.21	1	232	17
5	645	Bronx	Kingsbridge	40.86790	-73.90023	Private room	42	2	108	1.36	2	302	31
6	966	Bronx	Spuyten Duyvil	40.87991	-73.91673	Entire home/apt	120	2	47	1.22	1	318	7
7	1060	Bronx	Mott Haven	40.81128	-73.92399	Private room	49	1	23	0.27	1	333	20
8	1069	Bronx	Longwood	40.81611	-73.89909	Entire home/apt	100	5	82	0.96	1	63	21
9	1167	Bronx	Allerton	40.86870	-73.85240	Private room	35	7	2	0.17	1	90	13
10	1228	Bronx	Concourse	40.82822	-73.92439	Entire home/apt	250	3	119	1.41	1	339	70
11	1724	Bronx	Port Morris	40.80461	-73.92276	Private room	60	3	86	1.13	2	1	26
12	1749	Bronx	Fieldston	40.88757	-73.90522	Entire home/apt	60	1	25	0.67	1	311	36
13	2198	Bronx	Concourse	40.81906	-73.92806	Private room	85	2	11	0.18	2	363	41
14	2411	Bronx	Port Morris	40.80904	-73.93037	Private room	65	2	64	0.87	1	307	22
15	2498	Bronx	Mott Haven	40.81291	-73.90772	Private room	60	2	147	2.02	1	213	29
16	2587	Bronx	Kingsbridge	40.88166	-73.91103	Private room	90	3	0	0.00	1	353	27
17	2752	Bronx	Port Morris	40.80011	-73.91330	Private room	60	21	19	0.28	2	178	10
18	2930	Bronx	Williamsbridge	40.88296	-73.86264	Private room	50	1	19	0.83	1	311	16
19	2983	Bronx	Soundview	40.82138	-73.87603	Private room	45	3	53	0.82	1	249	13
20	2999	Bronx	Mount Eden	40.84367	-73.91718	Private room	43	2	11	0.16	1	338	35
21	3050	Bronx	Co-op City	40.86317	-73.82494	Private room	75	2	32	0.46	13	363	27

For the purpose of Price Prediction, I used Linear Regression Algorithm and I would like to talk about how I was able to reduce the RMSE from around $80 to around $40. The route which I took is as follows:

1. **Linear Regression(Without Standardizing):**
   The RMSE value for this model was $84.00
2. **Linear Regression(With Standardized Data):**
   The RMSE value for this model was $82.60
3. **Linear Regression (After Outlier removal and identification of significant attributes):**
   The RMSE value for this model was $41.75

Notable Observations:

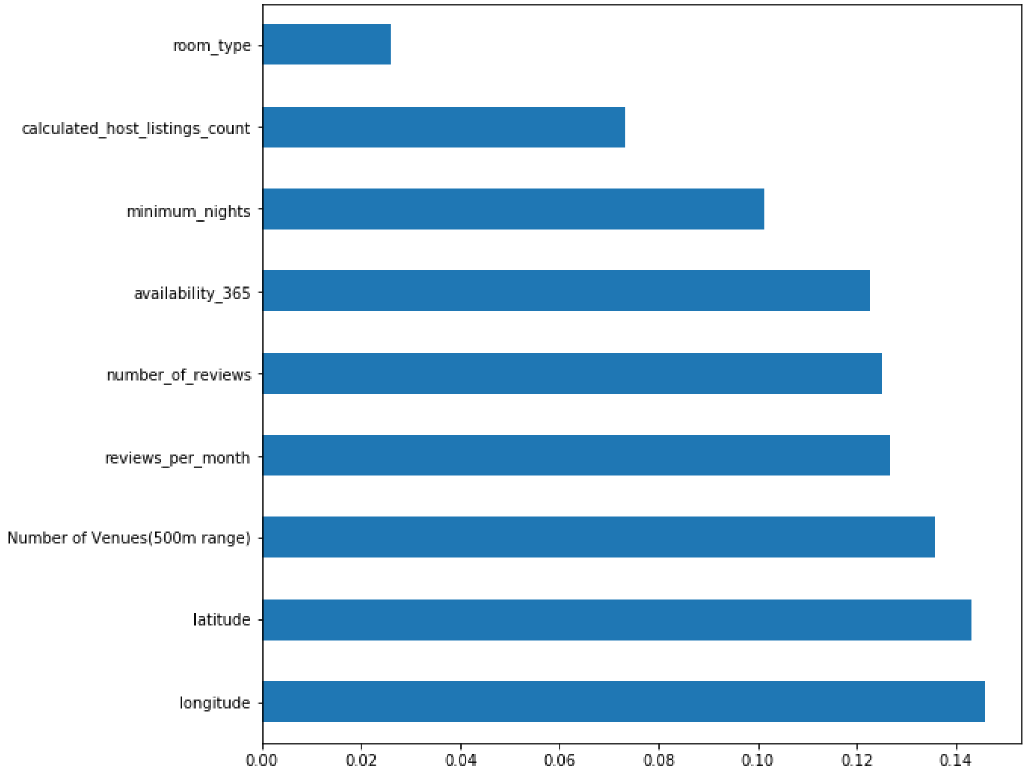
Here, I would like to talk about one of the notable observation which I saw while working on this project. Initially I had various geographical and numeric attributes in the dataset. However, it was mentioned in the project requirement to use Foursquare API so that we can work on more data. I therefore tried to find out the number of venues(within 500m radius) of the provided latitude and longitude of the Airbnb listing.

By incorporating this, I was soon able to find the impact of this attribute in my Linear Regression Model.

The following graph below shows the significant attributes impacting the prediction of prices.

IBM DATA SCIENCE CERTIFICATION CAPSTONE PROJECT

-Srinath Ramachandran



From this graph, it is pretty evident that Number of venues is indeed important in predicting the price of houses in a given locality.

Conclusion:

The final error score was $41.75 and this error was just using the data of Bronx borough. This error is remarkable given the small amount of training data that was used. Also, the idea of incorporating neighborhood data using Foursquare API was mindboggling. I could see the impact of neighborhoods on the price prediction of Airbnb listings.

Future Recommendations:

The error can be further reduced if more boroughs were included as this would have increased the training data. Also, different models like Random Forest Regression, Gradient Boosting Regression could be used on more training data.