AIRBNB LISTING PRICE PREDICTION FOR SEATTLE CITY

Aashirwad Kumar*
Integrated Msc in Mathematics and Computing
Department of Mathematics
Birla Institute of Technology,Mesra
Ranchi, Jharkhand-835215
imh10004.18@bitmesra.ac.in

Anmol Sharma
Integrated Msc in Mathematics and Computing
Department of Mathematics
Birla Institute of Technology,Mesra
Ranchi, Jharkhand-835215
imh10057.18@bitmesra.ac.in

Mentor : Ankit Tewari
Artificial Intelligence Engineer
Knowledge Engineering and Machine Learning Group
ankit.tewari@estudiant.upc.edu

July 2, 2019

ABSTRACT

The search for an ideal accommodation on travelling has been a real issue in recent years.AirBnB has revolutionized the way people think of finding places to stay. It has allowed people to open up their homes to visitors and stay at far more interesting places than the same old drab hotels. We using Data analytic tools and Data visualizations techniques have tried to provide eventful insights into data set. And general findings from data sets have been used by machine learning process to generate a prediction on price.

1 Introduction

The Open Source Seattle AirBnB Csv file is being added to the project as an input file. This contains three Csv files namely i)Listings ii) Calendar iii) Reviews . Our system will first process the data set employing data cleaning tools of python on it.Then employing the Data analysis ,visualization techniques we would try to analyze the features affecting the price of the listings .This project's one of major aim is predicting price employing KNN Regression model and Linear Regression model.

The major task can be broken down as 1) Data Cleaning 2) Data Processing 3)Analysis of features of data set 4) Answering important questions like How Neighbourhood affects price ? Which is the best time of year to visit Seattle ? What are good Reviews for a listing ? How is the price distribution and availability throughout the year? Which kind of property is more commomn and are there specific locations which are favourable for a particular type of listing ?

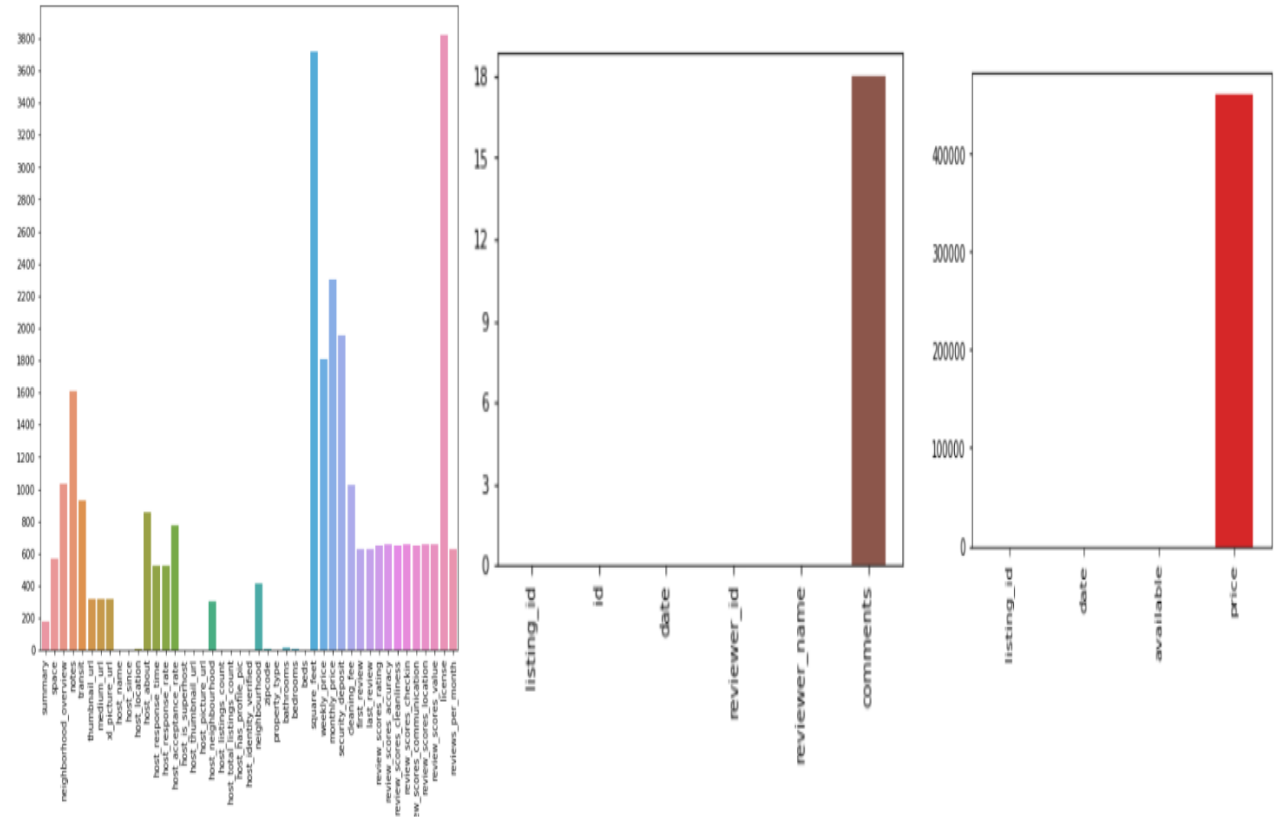
2 Data set

The Listings Data set has complete information on location ,coordinates , reviews , price , availability of each Host(listing). Rows : 0 to 3817 Data columns (total 92 columns) Calendar has day to day values listed and daily stats of listings. Rows:0 to 1393569 Data columns (total 4 columns) Reviews has review from various customers listed. Rows: 0 to 84848 Data columns (total 6 columns)

We then applying methods of Pandas and Numpy try visualizing the amount of NULL VALUES in all the three data sets and then remove the rows with missing values.

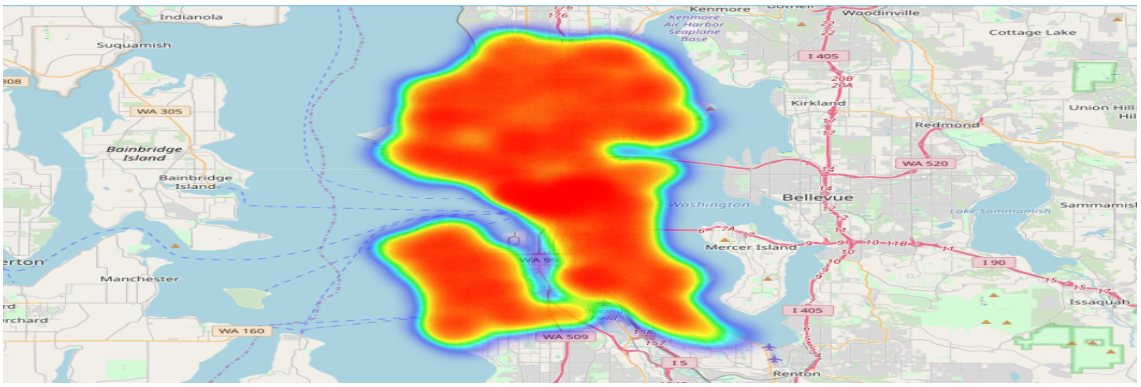
*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

2.1 We have visualized the no of null values in Listings data set ,Calendar data set and Reviews data set. Similarly we have visualized Null Values for two other data sets in our project.



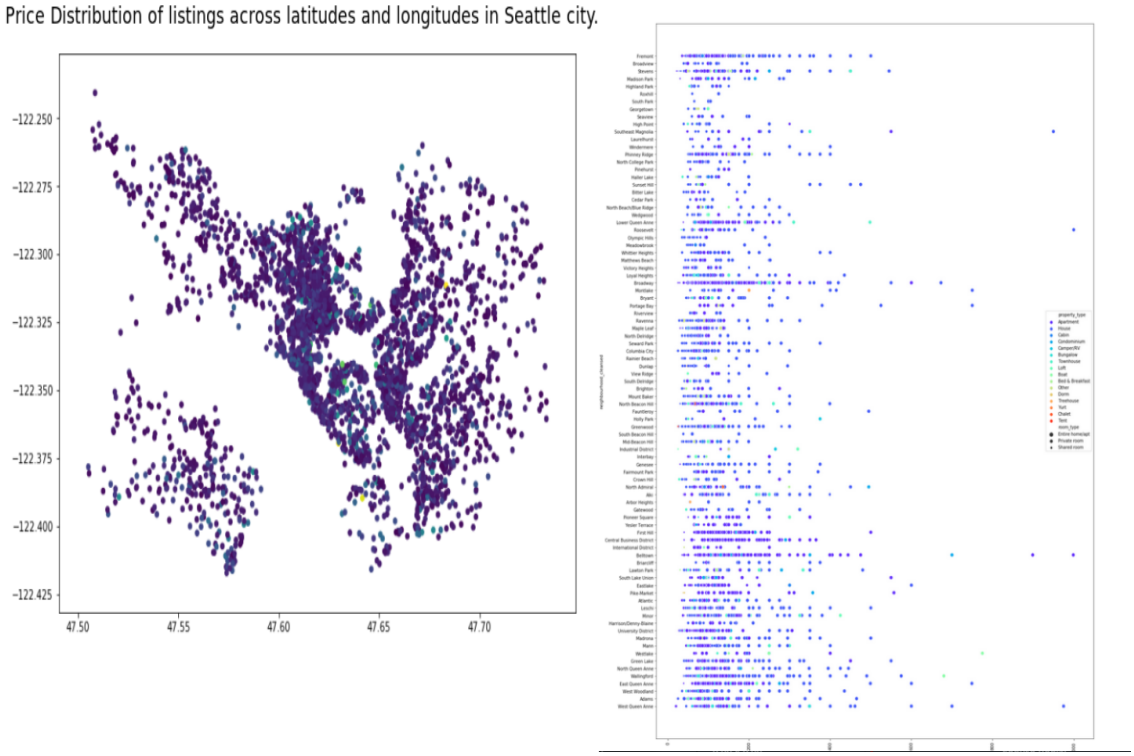
3 Features and Analysis of Dataset drawing important inferences

We have analyzed the listing data set and tried to answer a no of questions for convenience of tourists in selecting a perfect time and perfect kind of for them. Here we are presenting few important inferences from the project. Distribution of listings over the heatmap folium of Seattle City.



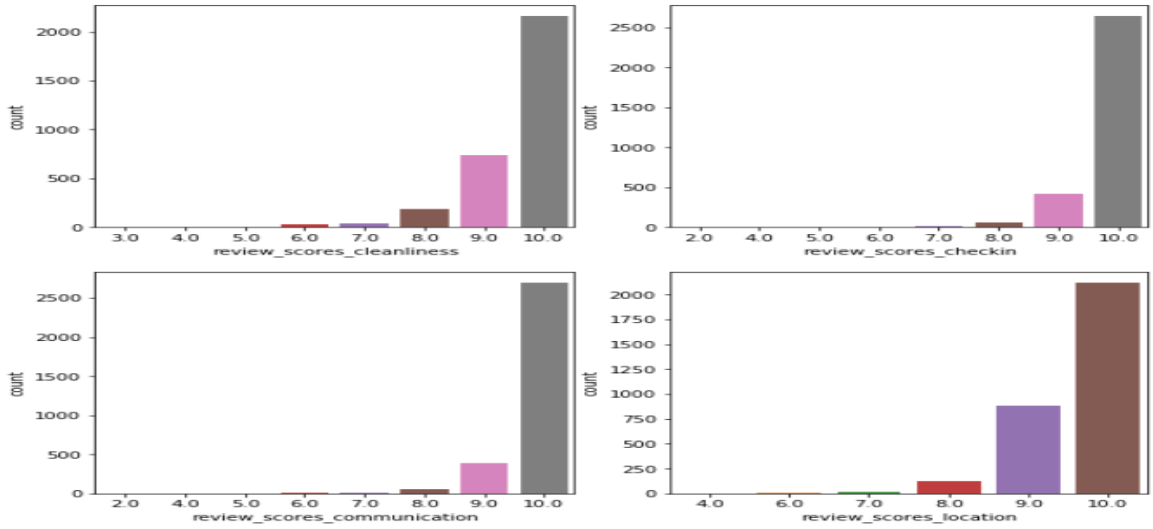
3.1 Relation of Neighbourhood with Price Plotted on MAP.

We have tried mapping coordinates and colour variations showing the changing prices over the various areas of Seattle City showing how the prices vary across the city. Also we have a scatter plot showing relation between Neighbourhood,Price,Property Type and Room type.



3.2 Analyzing Review's of listings

We have plotted reviews of various customers to listings and scores which shows us a trend about reviews on AirBnB's data sets that very few listings are reviewed below 8 .

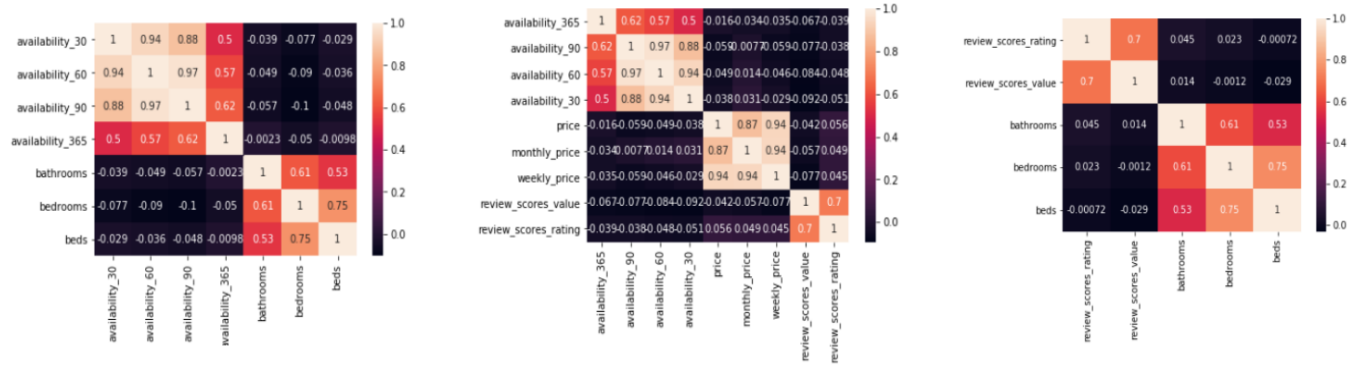


3.3 Availability Vs various other variables co relation using heat-maps.

The heat-maps showing co relation between availability over the i)month ii) two months iii) three months and iv) year with

1. Price
2. Property type
3. Room type and room features like beds, bathrooms and bedrooms.
4. Reviews.

We see that there is very unlikely co-relation between them which is surprising but this means that availability of listings is not affected by these features,



3.4 Analysis of changes in variables over the time.

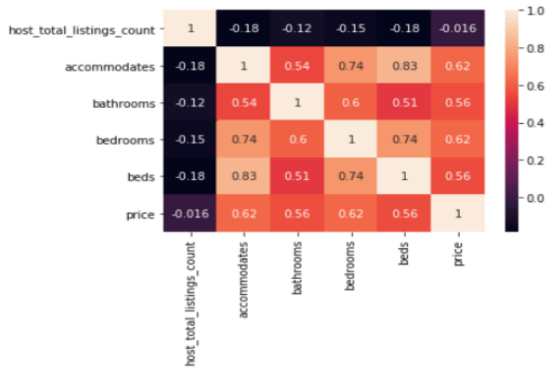
We plot the change in average no of listings ,average price and average availability over the time. We see that there is drop in price of listings in 2015 in January ,June,July,August which coincides with larger no of listings in those months.

4 Methods and Experiments

We have employed KNN Regression functions and Linear Regression model to predict Price of listings by features supporting use of regression model and having significant co relation with the price of the listing dataset. This co-relation is checked by using corr method in python and heat-maps.

4.1 Heat-Map

A heat map is a two-dimensional representation of data in which values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex data sets. There can be many ways to display heat maps, but they all share one thing in common – they use color to communicate relationships between data values that would be would be much harder to understand if presented numerically in a spreadsheet.



	host_total_listings_count	accommodates	bathrooms	bedrooms	beds	price
host_total_listings_count	1.000000	-0.176626	-0.124907	-0.149510	-0.183053	-0.015655
accommodates	-0.176626	1.000000	0.544817	0.742773	0.830499	0.624500
bathrooms	-0.124907	0.544817	1.000000	0.603602	0.507761	0.564460
bedrooms	-0.149510	0.742773	0.603602	1.000000	0.742577	0.623789
beds	-0.183053	0.830499	0.507761	0.742577	1.000000	0.558749
price	-0.015655	0.624500	0.564460	0.623789	0.558749	1.000000

We see that

- Bedrooms
- Bathrooms
- Beds
- Accomodates

Have very strong co relation with price. So we will use these features in prediction of price employing KNN method build from scratch and Linear regression to see the dependencies of predicted price with coefficients of variables taken into account while prediction.

4.2 KNN Regression

Pr(Y=j|X=x0)=1K iN0I(yi=j)

K Nearest Neighbors is a simple algorithm that uses all available cases and use those to predict for new and unseen cases used. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

We Divide the data set into train and test. With test size of 40 percent of data set

Getting minimum RMSE for testing data. For optimal value of K.

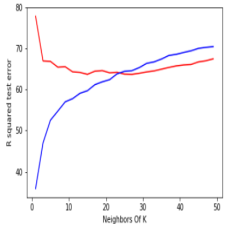
Defining a Predict Function. In this funtion we calculate the nearest distance and get the corresponding values of y train for value of K and get nearest neighbors value means of y train for more accuracy in predicted value and new values are so predicted.

4.2.1 Root Mean Square Error

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}}$$

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

Graph between RMSE and K value:



4.3 Linear Regression

Linear Regression is a Linear approach for modeling the relationship between Target Variable(Dependent Variable) and Other Variables(Independent Variable)

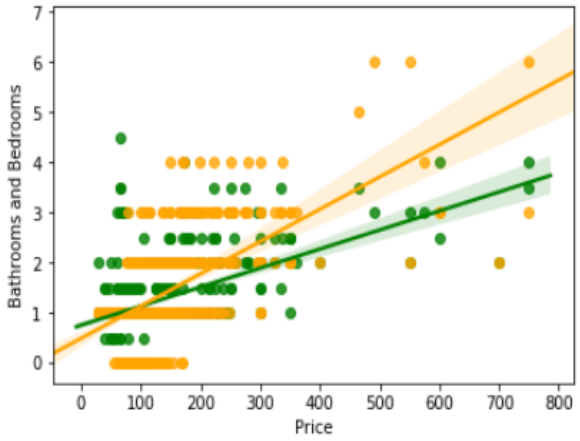
Mathematically We Take Account Linear Regression as:-

f(X)=0 + 1X1 + 2X2 + ...+ pXp

Where, 0=Intercept 1,2....,p=Coefficients X1,X2....,Xp=Independent Variable

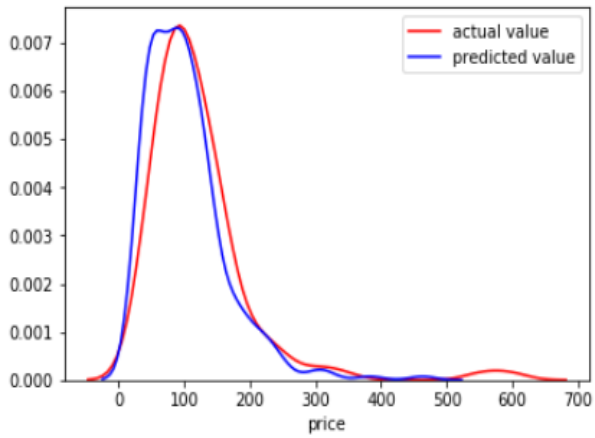
Value of Coefficients Defines how properlly the Independent Variables are Correlated to the Dependent Variable. i.e., How much our funtion will get effected for one unit increase in the Dependent Variable With Respect To The Value Of Coefficient.

So after predicting the values we employ Linear Regression to get values of coefficients of dependent variables. Here in the figure we show the regression plot of most important variables with price.



4.4 Results

We predicted the price of Listings using KNN regression machine learning algorithm and employed various methods like RMSE value and Distribution plot between ytest and predicted value to check for accuracy of our our which for randomly varying values of K comes pretty accurate. Here in the figure we show the distribution plot of actual data and predicted value.



5 Conclusion and Future Works

This project shows the import co relations between various amenities and features of listings in Seattle. After employing analysis visualization and co relation methods we outlined the variables necessary for price prediction We have predicted the price for listings by using machine learning method of KNN and Linear Regression. RMSE value helps in best possible implementation of KNN algorithm.The accuracy of predicted Price is judged by using distribution plots. Though KNN Regression seems pretty accurate when using few regression variables for price prediction but we can employ various advance machine learning algorithms also for price prediction which would thence increase the accuracy thus further reducing variance. In future we can employ algorithms to also take into account classification qualitative data to get more precise analysis and prediction.

6 Acknowledgment

We would sincerely thank our mentor Mr. Ankit Tewari for providing support to us throughout this project through his insightful views and hepling us understand key concepts of KNN and Linear Regression.

7 References

https://jakevdp.github.io/PythonDataScienceHandbook/04.13-geographic-data-with-basemap.html

https://stackoverflow.com/questions/17682216/scatter-plot-and-color-mapping-in-python

https://www.statisticshowto.datasciencecentral.com

https://statisticsbyjim.com/regression/

https://seaborn.pydata.org/generated/

https://stackoverflow.com

7.1 GITHUB LINK FOR PROJECT :

https://github.com/aashirwad01/Project_knn.git