National College of Ireland

Project Submission Sheet – 2020/2021

School of Computing

Student Name : Tanvi Nautiyal

Student ID : 19145381

Programme : Msc Data Analytics

Module : Domain Applications of Predictive Analytic**s**

Lecturer : Vikas Sahni

Submission Due Date: 28/06/2020

Project Title: **Predicting Prices of New York City's AirBnb**

Word Count: 1609

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: **Tanvi Nautiyal**

Date: **28/06/2020**

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not enough to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Prices of New York City's AirBnb

Tanvi Nautiyal
MSc Data Analytics
x19145381
National College of Ireland

Abstract: ***-Airbnb provides inexpensive accommodation for travellers in the form of short-term rentals. The demand for Airbnb has increased over the last decade with the growth of the tourism industry. New York City, one of the most popular and happening city in the world, has generated a huge demand for short term rentals such as Airbnb with an incredibly high tourism rate. Hence the main purpose of this research is to predict Airbnb 's price based on location in New York. With this research it is easy to find out which airbnb are the most common and which are not. Visualizations are done to provide a quick understanding of all the airbnbs present in the New York region. Once all the visualizations have been implemented, machine learning models are executed to predict the price of all the NYC Airbnbs. Machine learning algorithms such as Multiple Linear Regression and Random Forest will be implemented, resulting in the development of a forecasting model. On such models a comparative analysis shall be made, and consideration shall be given to the one that gives the best value.***

Keywords: Airbnb, Visualization, CRISP DM,, Machine Learning, Multiple Linear Regression, Random Forest, Price Prediction, Forecasting Model

Introduction:

People have been moving around more often over the last few years as they used to do before and hence the market for the hospitality sector is growing. Here, short term facilities come to play to address this ever-growing demand. After seeing a sharp increase in tourism domain many hotels have started increasing their rent and not every traveller can afford that. Airbnb, an online marketplace to share experiences and homes where guests seeking accommodation matched to hosts, who are having extra rooms to rent [1].

Frank Yang et al. in [1] states that it is the host who decides the price of the property and how it changes dynamically. However, Airbnb do provide suggestions to these hosts on how they can increase or decrease the rate of the property by comparing with other properties in the same here or during the holiday season.

Airbnb is considered as one of the most promising sector when it comes to hospitality and tourism domains as it is not only providing affordable but luxurious accommodations as well. Since its inception in 2008, Airbnb has been steadily growing in terms of revenue growth and its broad range of service provisions and people are preferring Airbnb's more than any other short-term rentals. Over 100 million Airbnb customers are already existing in many countries making the traditional hospitality industry a major disrupter. Airbnb receives revenue from both hosts and visitors, where 3% of the reservation value goes to the hosts who will arrange the stay and customers are charged between 6-12% of the booking. As a rental ecosystem Airbnb generates loads of data. Along with several other cities, New York is one of Airbnb 's popular markets. Renting an Airbnb in New York City can turn out be very tedious as one has to figure out the best prices, neighbourhood and many other factors while booking one. Since New York is one of the most expensive cities, when choosing Airbnb, one must be careful. The goal of the project was to learn how neighbourhood prices and other attributes differ. The dataset used for this research taken from [5].

Research Goals:

In the following research the dataset has been gone through briefly followed by implementation of machine learning algorithms to create a forecasting model. This research shall focus on:

- Exploratory data analysis on the dataset which includes pre-processing and visualizations.

- Analysing all the independent variables and what impact they have on dependent variable (Price).
- Implementing machine learning techniques once all the exploratory data analysis has been done.
- A comparative analysis is done on the machine learning models and determining the best technique for forecasting.

Ethical Concerns:

Every domain is exposed to ethical breach and Airbnb is no exception here. It is very important to make customers safe and comfortable. Following are certain concerns that comes up in this sector:

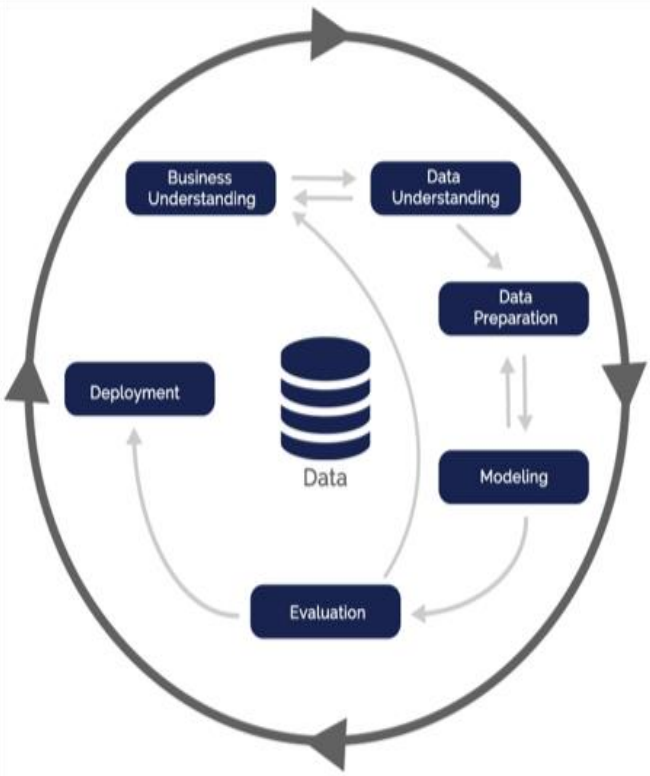
- **Non-discrimination Policy**
  Here people from all backgrounds, no matter how far they have travelled from home, feel welcome and respected and it won't not allow hosts to break local laws or take action that may impose legal liability upon them. Airbnb shall welcome all the guests irrespective of their colour, creed, nationality, gender, marital status.

- **Non-Disclosure Policy:**
  The identity of a customer shall not be revealed by Airbnb Host.

- **Security:**
  Inside theft are one of the most common concern, so to overcome this CCTV cameras are installed in public areas.

- **Hygiene:**
  The room provide to the customers should be neat and clean. Especially the washroom is sanitized properly. The room should be clean thoroughly so that customers won't find any flaws and it should be in proper condition when handed over.

Strategy:



For an Airbnb to stay in this ambitious environment, it is very important to have a systematic plan and work according to that. When it comes to business strategies CRISP DM (CRoss Industry Standard Process for Data Mining) shall be used.

- **Business Understanding**:
  In order to survive in this cutthroat market, it is very important to understand what the customer needs and the booking patterns. This is help levelling up the business.

- **Data understanding:**
  Data shall be loaded into the tool and shall be gone through briefly. Here the main objective is what is expected from the data and what can be achieved out of it.

- **Data Preparation:**
  Includes ELT (extract,transform,load) process. Here all the exploratory data analysis such as data cleaning, converting into numeric or categorical values, visualizations comes.

- **Data Modelling:**
  Once all the data pre-processing is done then machine learning techniques either regression or classification are implemented.

- **Evaluation:**
  Here data is split into test and train data and then machine learning models are implemented. The results obtained here should be correct and are validated in this phase.

- **Deployment:**
  Once the correct results are obtained then it is further deployed on the server for the clients to use.

Preliminary Visualization:

Here visualizations have been performed on Airbnb data to get to know briefly and to have a better understanding about it. The scatterplot shown in Fig 2. displays the latitudes and longitudes of the neighbourhood groups of the AirBnB listings.

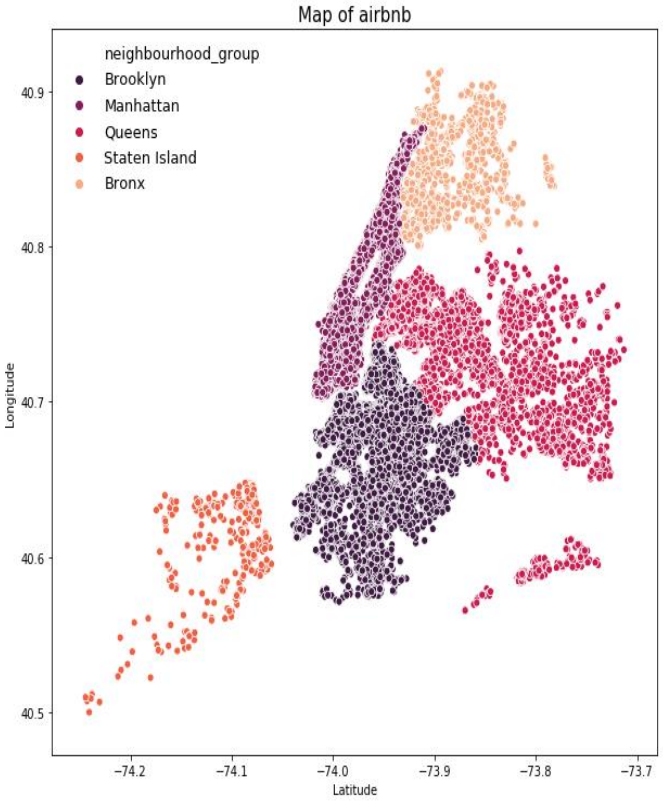


Fig 2

The next fig 3, shows the top 3 neighbourod_groups of Airbnb listings in New York City, where Manhattan is the most popular neighbourhood, having the highest listings followed by Brooklyn and then by Queens.

	Listings	Neighbourhood Group
0	21661	Manhattan
1	20104	Brooklyn
2	5666	Queens

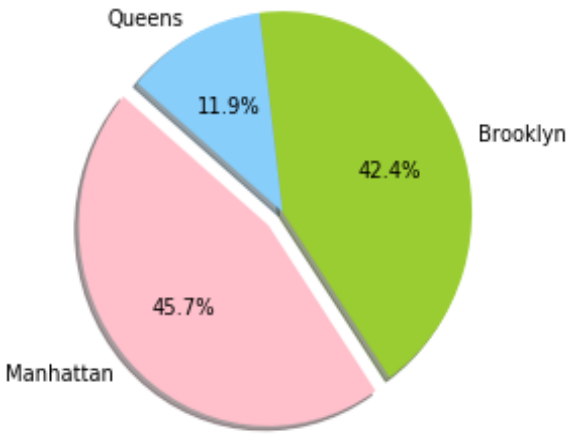


Fig 3

Figure 4 depicts the hosts having the maximum number of Airbnb listings in NY region. Here Michael have the highest number of listings.

	Listings	host_name
0	417	Michael
1	403	David
2	327	Sonder (NYC)
3	294	John
4	279	Alex

```
: sns.barplot(y="Listings", x="host_name", data=top_5_hosts)
  plt.show()
```

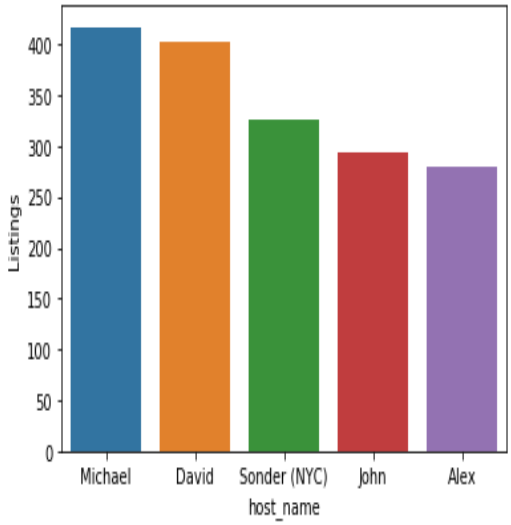


Fig 4

Fig 5 states the most common words used in the name column of the airbnb dataset.

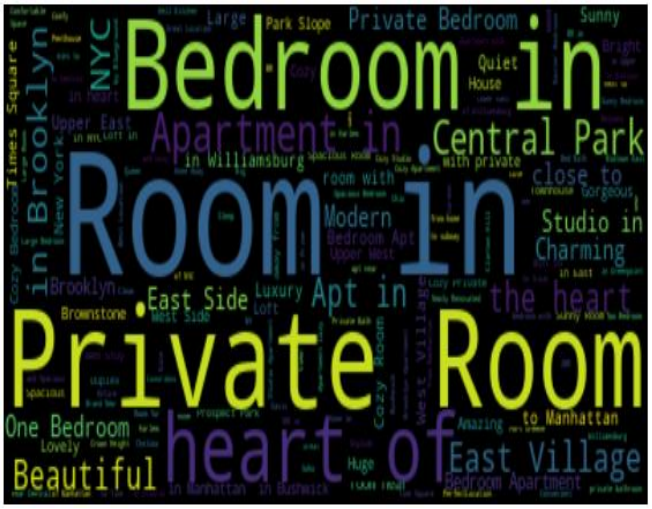


Fig 5

From fig 6, it can be seen the types of rooms Airbnb is providing in New York City. Here Entire home/apt variable have the highest number of count all over the region whereas shared rooms are the lowest.

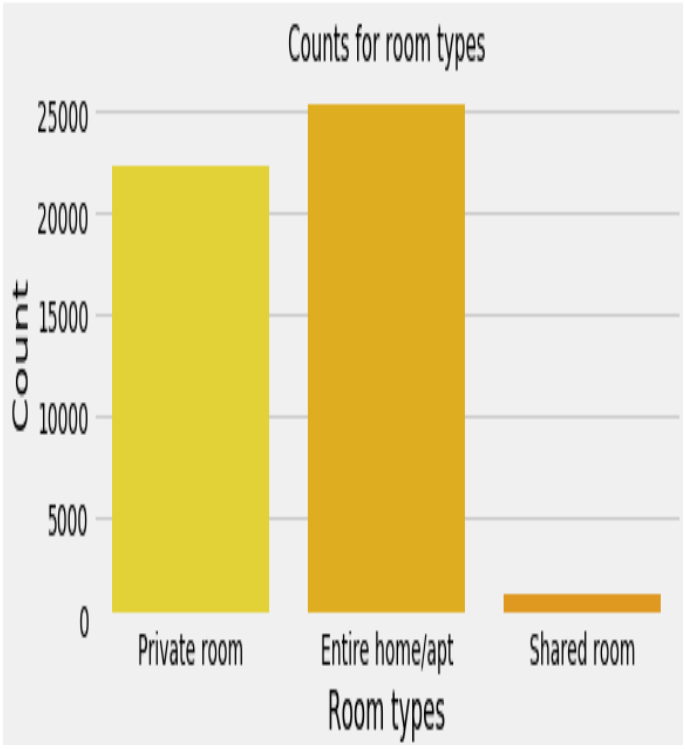


Fig 6

The bar chart in fig 7 and fig 8 states the distribution of the room types (Entire flat and Private room) in the most popular neighbourhood of New York, i.e Manhattan and Brooklyn. Here districts of Manhattan is

shown in the lighter shade, whereas Brooklyn's districts are shown in darker shade. In Fig 7, Williamsburg district of Manhattan have the highest number of Entire home/apt room types, however, Bushwick district of Manhattan are having the minimum number of Entire flat room type.

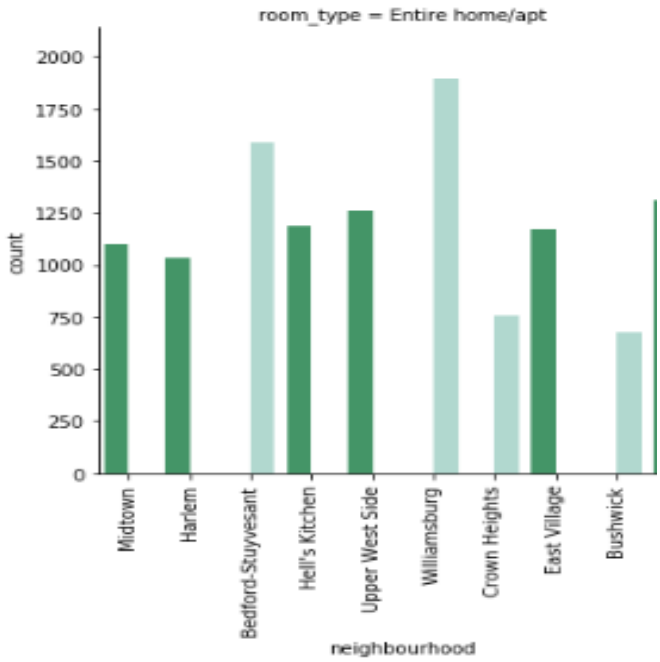


Fig 7

Fig 8, shows the number of Private Room distributed in the districts of Brooklyn and Manhattan. Here, Bedford-Stuyvesant district of Brooklyn having the highest number of private room whereas, Upper East State, district of Manhattan is having the lowest number of private room.

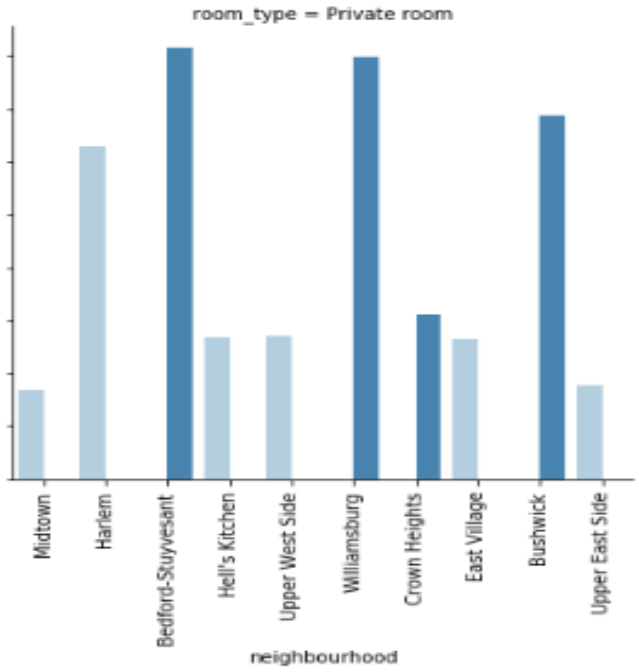


Fig 8

Applicable Techniques:

Pouya Rezazadeh Kalehbasti et al. [3] implemented machine learning, deep learning and natural language processing techniques on the Airbnb dataset and developed a price predicting model through it. In [3], the predicted model will not help the Airbnb hosts but also give an idea about the price to the customers who are looking to book an Airbnb. Here, Ridge Regression, K-means clustering, Support Vector Regression, Neural Network and Gradient Boosting Tree Ensemble techniques are used to create a predictive model.

Tao Yang et al. [4] describes a price prediction paper for holiday rental websites where rental property prices typically used for travel or vacation purposes are forecast using various regression models. Furthermore, they used a multi-scale affinity propagation method, which is essentially the clusters where the unwanted data is extracted from the dataset.

For this research two regression models have been selected and a comparative analysis will be done, and whichever models presents the better result predictive models shall be made from that. Multiple Linear Regression and Random Forest machine learning techniques are used for price prediction.

References:

[1] P. Ye et al., "Customized Regression Model for Airbnb Dynamic Pricing", *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. Available: 10.1145/3219819.3219830 [Accessed 24 June 2020].

[2] I. Rodrigues, "CRISP-DM methodology leader in data mining and big data", *https://towardsdatascience.com/*, 2020. .

[3] P.Rezazadeh, L. Nikolenko and H. Rezaei, "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis",Research Gate, 2019. [Accessed 24 June 2020].

[4] Y. Li, S. Wang, T. Yang, Q. Pan and J. Tang, "Price Recommendation on Vacation Rental Websites", *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 399-407, 2017. Available: 10.1137/1.9781611974973.45 [Accessed 24 June 2020].

[5] "New York City Airbnb Open Data", *Kaggle*, 2019. .