Project Report

Anonymous Author(s)

ABSTRACT

Airbnb is a unique idea of lodging where a person offers their home to some other person for staying rather than letting them reside in a hotel. Founded by Brian Chesky, Nathan Blecharczkya, and Joe Gebbia, Airbnb is a "sharing economy" that makes sharing effortless, entertaining, and secure. In this project, critical analyses were made on Airbnb properties, their owners, and the neighborhoods using open-source, reliable Airbnb data that consists of information about property types, facilities provided by the owner with respect to the property, and many other important factors. For the project, datasets from New York City and London were used because these cities are popular all around the world and are considered to be tourist and vacation hot-spots, thereby consisting of significant number of active Airbnbs.

In this project report, a wide range of data processing and data cleaning techniques were used post which three data mining models were built: A K-Prototype clustering model to find most preferred neighborhoods, a linear regression model with multiple variables to predict prices of listings, and a multivariate regression model that helped identify factors that would boost reservations.

KEYWORDS

Airbnb, Listings, Review, Calendar, Regression, Clustering, Data Cleaning, Boosting Reservation, Price Prediction, New York, London, SQL Server Management System

1 INTRODUCTION

A pioneer in the lodging industry, Airbnb works on a unique business model of ensuring smooth business between the owners of the property and the customers willing to stay in that property. With a market value of almost $84 billion, surely Airbnb is ruling the market and this makes studying the Airbnb data an interesting topic for data exploration. In this project, the Airbnb dataset [1] was taken into consideration to understand the trends of most popular neighborhoods in different cities and compare them to analyze people's selection criteria for these neighborhoods based on number of listings, price, and reviews. A large amount of data was required to analyze the information mentioned above. Two cities, New York City, USA, and London, UK were chosen for this purpose, with a total data set of around 43 million tuples.

The motivation to select the Airbnb dataset for analysis was to understand the trends of different variables in the dataset and provide meaningful analysis, such as Boosting Reservations and Predicting Prices for the properties based on the location in New York City and London. Airbnb owners sometimes face issues while getting an ample number of reservations. One of the analyses provided in this project aims at helping these owners provide some insights as to what features of the property make it an attractive choice for people selecting it. While one analysis focused on the producer's issue in the Airbnb ecosystem, the second analysis focused on the consumer's issue of helping customers gain an insight into the price expected (predicted) in a particular neighborhood. This analysis will help keep a check on overpricing of the property and help the property owners decide the tariff of the property per night.

In the course of the project, the data went through several steps such as pre-processing which includes Data Collection and Data Cleaning, preparing a schema for storing the data in the database, loading the data into the database (SSMS), and preparing analyses on the clean, accurate data. The steps mentioned have been explained in detail in the Design and Implementation section.

Section 2 will elaborate on the goals of the project. Section 3 will emphasize on the existing analysis done on Airbnb data. Section 4 will explain the design of the algorithms used. Section 5 discusses the legal and ethical considerations related to the project, Section 6 covers the learnings from the project. Section 7 presents the current progress and future scope of the project.

2 PROJECT GOALS

The following applications are the goals of this project:

(1) ***Boosting Reservations***: Provides an analysis of how various factors influence the number of bookings. This was done by looking at the listings with the most and most minor bookings and analyzing how different factors affect the number of reservations.[4] This can serve hosts in determining how to improve bookings, retain clients, and enhance income. To discover how multiple variables impact the number of bookings, Multivariate Regression was be applied in this evaluation.

(2) ***Price Prediction***: The best areas in each city were identified by weighing location review scores, the price range of available listings in the region, number of listings in each neighborhood, number of reviews for each neighborhood and a few more. This data was then used to predict the price of a listing depending on its location and other factors such as host acceptance rate, host response rate, accommodates, number of amenities, beds and a few more. This application involved the usage of Classification to categorize listings and Prediction to forecast pricing. With this, customers will be able to acquire a pricing estimate, and new hosts will be able to establish prices for their properties.[3]

The ultimate goal of this project is to support current or potential owners of Airbnb-registered residences by providing practical approaches and visualizations. These will help improve their sales practices and hence, increase their income.

3 RELATED WORK

This section reviews other people's research work that are related to the objectives and analysis of this paper.

(1) In the paper by Yang Li et al. [5], the authors have put forward Multi-Scale Affinity Propagation (MSAP), which predicts a reasonable price of an Airbnb. For developing this system they have initially developed a clustering model that

assembles houses on the basis of its landmark and facilities provided. Subsequently, for predicting reasonable prices, a Linear Regression model with Normal Noise (LRNN) is used which is further authenticated with an improvement on renting reviews.

(2) The paper by Zu et al. [8] used Airbnb data from New York city to build prediction models using various natural language processing and machine learning algorithms in order to predict prices of listings and compared their performance. Algorithms like Deep neural network (DNN), random forest, XGBoost, linear regression, generalized additive model (GAM) and bootstrap aggregation(Bagging) were used. GAM AND DNN were outperformed by bagging and machine learning models. It was determined that Bagging would be the best model to use for price prediction.

(3) The model presented in the paper by Abdar et al. [2] is called Crowd Preference Mining Model (CPMM) aims to detect a competent type of property, rooms, and stars according to user behavior by maneuvering a Rating Matching Rate (RMRate) on three significant aspects of an Airbnb that are property-type, room-type and star-rating. They observed that an Apartment, an entire house, and Bed Breakfast were the important property types all over the world excluding the USA and Spain. They also noticed that the entire apartment or house was not so famous in the UK. Also, people in the USA, Germany, Australia, UK, and Russia were more concerned about 5-star accommodation than other countries.

(4) In the research paper, Dhillon et al. [3] conducted a thorough study of the Airbnb pricing data and discovered several trends regarding common and uncommon property listing types, price changes over time, and review patterns. This information was used to evaluate the performance of three machine learning models, namely random forest, linear, and logistic regression, in predicting listing prices, based on their room-mean-square-error (RMSE) values. It was concluded that the random forest model produced the best results

4 DESIGN AND IMPLEMENTATION

4.1 Data Flow Diagram

The Dataflow diagram in Figure (1) illustrates the sequence of steps followed in order to implement the goals of this paper. These steps have been described in detail in the sections below.

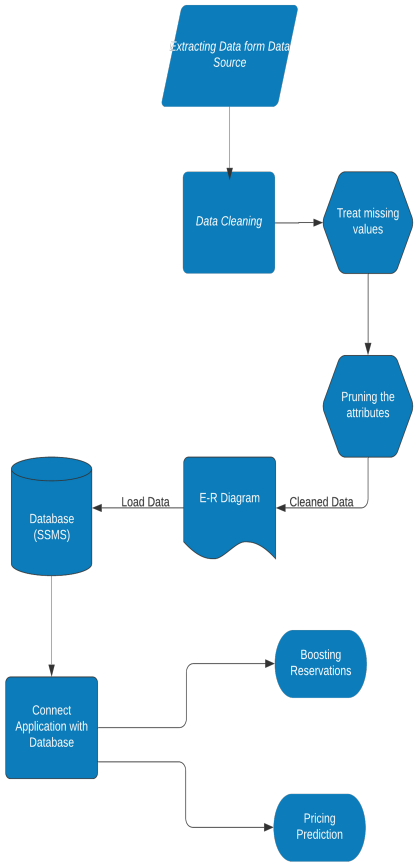

Figure 1: Data Flow Diagram

4.2 Data Collection

The data was collected from the open-source Airbnb dataset of different cities. Data from 2 cities - New York City and London, was selected from the Inside Airbnb dataset. The data collected for each city was in CSV format and was primarily divided into three datasets, namely Listings, Reviews, and Calendar. These datasets mentioned property details such as id, URL, description of the property, the id of the host, and their names. The description for these datasets is given below:

(1) ***Calendar***: The Calendar table (1) consists of seven attributes which are as follows: 1. listing_id: The primary key from the listings table that recognizes each listing uniquely. 2. date: The date for which the booking has been made. 3. available: Whether the property is available for that date or not. 4. price: The per day tariff for the property. 5. adjusted_price: The adjusted price in case of any discounts. 6. minimum_nights: The minimum number of nights for which the property has to be booked. 7. maximum_nights: The maximum number of nights for which the property can be booked. This table consists of 13,467,706 tuples for New York City and 27,329,674 tuples for London

(2) ***Reviews***: The Reviews table (2) consists of six attributes which are as follows: 1. listing_id: The primary key from the listings table that recognizes each listing uniquely 2. id: The unique id of the review 3. date: Date on which the review was written 4. reviewer_id: The unique id of the reviewer 5. reviewer_name: The name of the person giving the review 6. comments: The comments given by the reviewed listings. This table consists of 836,586 tuples for New York city and 1,129,280 tuples for London.

(3) ***Listings***: There are total seventy four attributes in the Listings table (3). For representation purpose, some of the attributes that will be used for the analysis have been defined

ListingID	*Date*	*Available*	*Price*	*Adj_Price*	*MinNights*	*MaxNights*
9357	2021-04-09	f	$150.00	$150.00	30	1125
2595	2021-04-09	t	$150.00	$150.00	30	1125
2595	2021-04-10	t	$150.00	$150.00	30	1125
48728383	2021-04-06	t	$102.00	$102.00	31	365
48728383	2021-04-07	t	$102.00	$102.00	31	365
48728383	2021-04-08	t	$102.00	$102.00	31	365
18127	2021-10-27	t	$170.00	$170.00	30	30
206957	2021-08-31	t	$95.00	$95.00	30	365
206957	2021-09-01	t	$95.00	$95.00	30	365

Table 1: Calendar Sample Data

ListingID	*ID*	*Date*	*ReviewerID*	*ReviewerName*	*Comments*
2595	19176	2009-12-05	53267	Cate	Great Experience.
2595	19760	2009-12-10	38960	Anita	I've stayed with my friend at the Midtown...
5803	33387264	2015-05-27	32274754	Abby	Laurie emailed us a helpful booklet...
3831	145980375	2017-04-21	29683774	Rog	Nice place to stay, away from hectic...
3831	148154612	2017-04-30	76155255	Paul	We had a nice stay. Lisa was very helpful...
3831	149662395	2017-05-05	111051745	Chris	We had a lovely time at Lisa's place. Our...
62925	220688548	2017-12-23	15822971	Michael	Awesome! I highly recommend this place.
62925	341296084	2018-10-26	129359069	Helen	Great place with a friendly neighborhood.
4746171	49497614	2015-09-19	18179156	Karil	The host cancelled the reservation 33 days...

Table 2: Review Sample Data

below: 1. id: The unique id for each listing 2. host_id: The unique id of the host 3. amenities: The list of amenities available at the property 4. price: The per day tariff of the property 5. accommodates: The number of people that can be accommodated at the property 6. number_of_reviews: The number of reviews written for each listing 7. host_is_superhost: True/False based on whether the host is a super host 8. beds: The number of beds available at the property 9. neighbourhood_group_cleansed: The neighborhood in which the property is located. This table consists of 36,905 tuples for New York city and 66,856 tuples for London.

ID	*HostID*	*Amenities*	*Price*	*Accommodates*	*Superhost*	*Beds*	*Neighborhood*
2595	2845	["Refrigerator", "Essentials",...	$ 150.00	2	f	1	Manhattan
3831	4869	["Bath tub", "Smoke Alarm",...	$ 76.00	3	f	3	Brooklyn
5121	7356	["Kitchen", "Heating",...	$ 60.00	2	f	1	Brooklyn
5136	7378	["Kitchen", "Table TV",...	$ 175.00	4	f	2	Brooklyn
10962	9744	["Essentials", "Dryer", "Patio",...	$ 92.00	2	t	0	Brooklyn
12192	46978	["Heating", "Shampoo", "TV",...	$ 40.00	2	f	1	Manhattan
12937	50124	["Essentials", "Luggage Dropoff",...	$ 130.00	4	t	2	Queens
18127	69829	["Cable TV", "Refrigerator",...	$ 176.00	2	f	1	Manhattan
42882	185978	["Heating", "Washer",...	$ 70.00	2	t	1	Staten Island

Table 3: Listings Sample Data

4.3 Data Cleaning and Preparation

After selecting the dataset from the source, the data had to be cleaned and prepared for further analysis.[7] This step is known as ***Data Cleaning and Preparation*** . For the Data Cleaning process, pandas library in python was used to read the data tables and process them. In the dataset selected, it had been observed that multiple tables in the dataset contained different attribute names for the same attributes. Naming the identical attributes uniformly across all tables would be an important step provided, data from different cities would be merged in queries for analysis. Elimination of redundant data was also conducted. After a preliminary inspection of the datasets, it had been observed that the data contained fewer null values. The numeric null values were replaced with 0 and "N/A" was substituted for null values for attributes in the text or string format. Furthermore, attributes were selected from each table in the dataset which would accurately help in the analyses decided on the Airbnb dataset. After finalizing the questions that needed to be answered through analyses to be done on the dataset, those attributes which won't help in these analyses like listing_url, scrape_id, host_thumbnail and host_image, were removed and a copy of new data tables were formed. The old version of the data was preserved in case of data loss during processing. It was made sure that the type of each attribute was formatted according to the requirements and the type of the values restricted for each attribute remain the same throughout the data. Since multiple cities from different countries were being dealt with, so for simplicity purposes, the prices of all the properties were converted in USD. Prices not mentioned in USD were converted using the conversion rate on the day of calculation.

4.4 ER Model

Once the data was prepared, the ***ER Model*** was made. The ER diagram in Figure (2) explains the database schema and the relationships between the tables in the dataset. From the main data source, three main data sets were taken for exploring the database and analyzing it. The entities of the ER model were Listings, Review, and Calendar data tables. The relationship between these tables was established by selecting primary and secondary keys that could be key associations between the entities. The relationship between the tables was a one-to-many relationship between the Listings and Review table with multiple reviews for one listing_id. Also, the Listings table shares a one-to-many relationship with the Calendar table (1: M).

4.5 Database Hosting

The next step was ***Database hosting*** where using python code a connection was established with the SQL Server i.e., SQL Server Management Studio(SSMS). SSMS was chosen because it is an open source SQL Server which is known to handle high performance workloads well. After the tables were cleaned, they were uploaded into an SQL database for further exploration. This could be done by converting the excel sheets into text files and using appropriate tools such as the Bulk Insert command or the BCP Tool. There are also tools that allow us to import data from excel sheets directly into an SQL Server like the SQL Server Import and Export Wizard. This was the method used to upload the preprocessed tables into the Secure SQL Management Server. Following which, certain queries were run to extract the data needed to generate some useful visualizations and data required for further analysis.

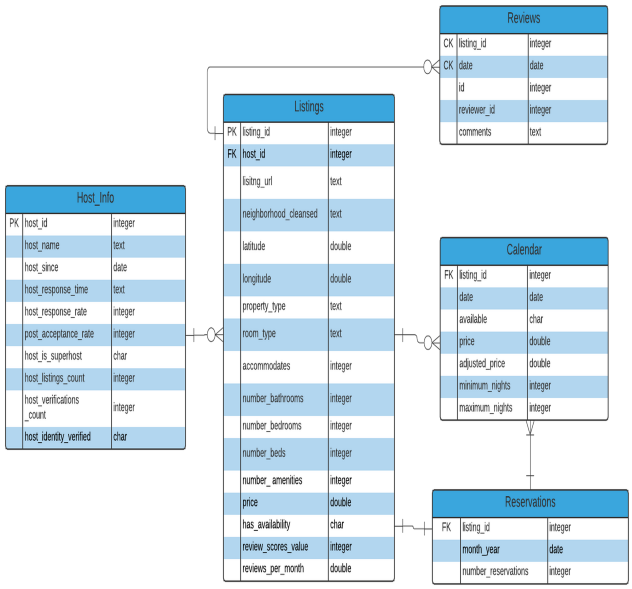

Figure 2: ER Diagram

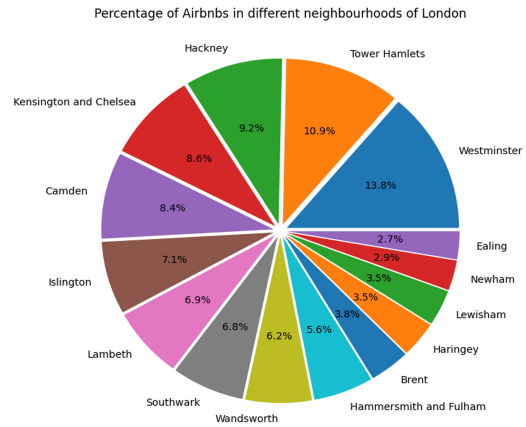

Figure 3: Top 15 Neighbourhoods of London with highest Listings

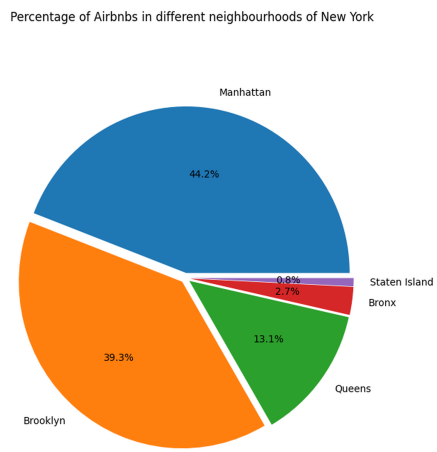

Figure 4: Top 5 Neighbourhoods of New York with highest Listings

4.6 Exploratory Analysis

Tables were joined and queries were run on them to extract some insightful information such as number of listings per neighborhood, number of active properties for each property type, understanding the relation between accommodates and prices, and so on. Using this data, visualizations were created, some of which are described below:

(1) The number of listings for different neighborhoods in each city were analyzed in order to study the popularity of these neighborhoods. The pie charts in Figure (3) and (4) represent the top 15 neighborhoods that have highest number of listings in London and the top 5 in New York City.
(2) The demand of different types of properties in both the cities was studied to determine their impact on the price of a listing. The bar graphs in Figure (5) and (6) depict top 10 types of properties and their respective number of listings in London and New York City.

4.7 Data Mining

The steps taken to implement the objectives of this paper have been described below:

(1) ***Boosting Reservations*** : To boost reservations, factors such as price, amenities, the host response rate and some related attributes were selected on the basis of their correlation with the number of reservations. Multivariate regression determines how multiple independent variables affect a dependent variable. This regression technique has been used to determine how the selected factors affect the number of reservations made. This relationship can be leveraged to boost the number of bookings made. The prepared model, predicted the number of reservations for different combinations of factors. It was observed that the three main factors that influenced reservations were amenities, accommodation and price. For the purpose of training the model, the dataset was randomly divided into training and testing datasets, with 70% of the data being considered as the training data and 30% of it as the testing data.
(2) ***Pricing Prediction*** : When determining the price, various factors such as the preference of the neighborhood, the type of building, and facilities must be taken into account. The

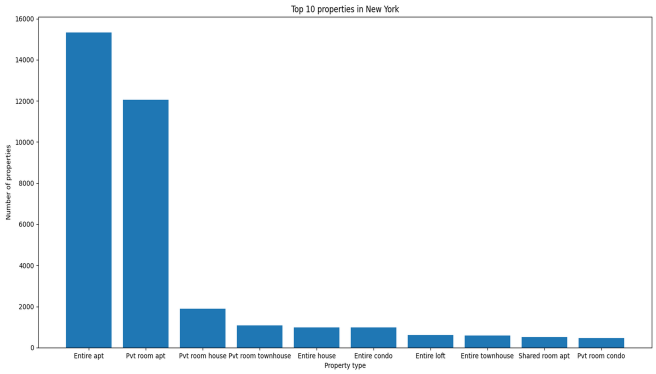

Figure 5: Top 10 property types of New York

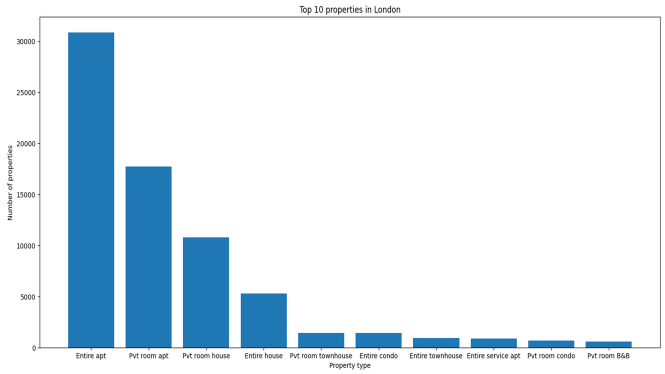

Figure 6: Top 10 property types of London

popularity of the area is likely to have a significant impact on the price. There was, however, no factor in the data that explicitly reflected the popularity of a neighborhood. To calculate this statistic, a clustering model was created that took into account variables like the number of listings in the area, the number of reviews for those listings, the pricing and the review ratings of those listings to divide individual listings into different clusters. Since the attributes considered consisted of both numerical and categorical data, K-means clustering algorithm could not be used as, it is restricted to numerical data. An algorithm called K-Prototype was used for clustering instead, which deals with mixed data-types (numerical and categorical). These clusters indicated the preference of neighborhoods and were introduced as a new attribute to the Listings database. This variable was combined with accommodates, beds, number of bedrooms, host acceptance rate and amount of amenities provided to generate a regression model that predicted the listing price. To accomplish so, the dataset was divided into training and test sets. 70% of the data was utilized for training, and 30% was used for testing.

4.8 System Architecture

Figure 7 depicts the architecture model used to implement the applications of this project. The clean and processed data was first uploaded into the database and the tables were then manipulated using SQL queries such as Join, Group By, and other aggregation functions. Some of factors from the datasets were provided as input to data mining algorithms to get the desired outputs.

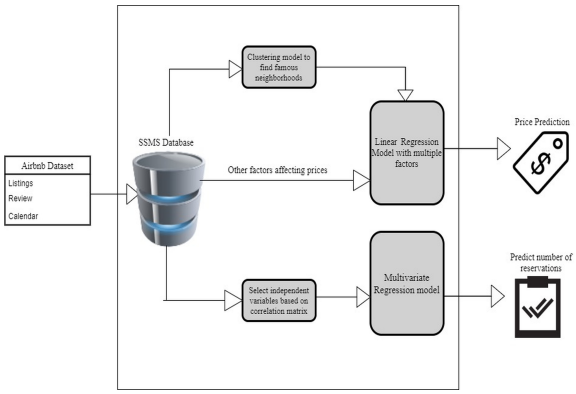

Figure 7: System Architecture

5 LEGAL AND ETHICAL CONSIDERATIONS

The ACM Code of Ethics and Professional Conduct was diligently followed throughout the entire project and every step in the project was carefully checked to make sure it doesn't violate the code. The data was collected from an open source, which[6] clearly mentioned in the Disclaimer that no private information whatsoever was used in the data provided. The project also followed the general ethical principle of contributing to the society by providing important analysis for both the elements in the Airbnb model, the property owners as well as the customers. While the team performed some important analyses from the data, it was also ensured that the future scope of the project was explained in one of the sections of the report. This will help carry forward the research and will help other data analysts with some ideas. Throughout the project, the team's analyses were free of bias on any basis. The notable prediction models as well as originality in the visualizations, data cleaning techniques and other work done in this project highlights the Professional Ethics of providing a high quality product.

6 LESSONS LEARNED

The project provided excellent exposure to handle large amount of data and helped us understand the steps of converting data to knowledge by actually putting us in the situation where we had to deal with some very basic, but important things and some critical things like providing useful analyses. We learnt how to clean the data, the steps involved in that part, like finding missing values, merging columns with same variables, maintaining consistency of the attributes throughout the data set and many more. We also realised that data cleaning was a time taking step, that demands meticulousless, as any erroneous values can heavily impact the results obtained at a later stage. We closely went through the data

set and noted a list of questions that we answered throughout the project in the form of different analyses. Helping property owners boost reservations and predicting prices based on neighborhood were the type of analyses done on the dataset. In this, we studied a number of topics such as K-Prototype Clustering and Multivariate Regression. Also, we learned that while considering data this huge, many complications such as loading the data in SSMS and time taken for this step, using the prediction models on data that is not clean in some instances yeilding poor results and many such scenarios come into picture which needed to be handled. However, an important lesson of solving the issues effectively and efficiently as and when they come was learned through the means of this project.

7 CURRENT STATUS & FUTURE WORK

7.1 Current Status

Upto this point the following steps have been completed :

- ***Data Cleaning and Integration*** According to the scope of the project, data was prepared and necessary cleaning was done on all the datasets. Some of the cleaning steps were replacing null values with '0' or 'NA', removing columns that were not needed and handling inconsistencies in the data. The tables were then uploaded into an SQL Server (SSMS) and queries were performed on the tables to study the required attributes and create some visualizations.
- ***K − Prototype Clustering***: To find the famous neighborhoods in each city, the neighborhood were divided into clusters based on factors such as, number of reviews, price, review scores and a few more. Since the attributes used consisted of both numerical and categorical data, K-Prototype algorithm was used for clustering the neighborhoods. The dataframe was converted into a numpy array in order to be used as an input for the k-prototype algorithm.
- ***Multivariate Regression***: To predict the price of an Airbnb according to user preference, multivariate regression model was used. Various attributes like famous neighborhoods that were found using K-Prototype Clustering, demand of different house types, prices, ratings and a few more were used to train the model. To help boost the number of reservations, factors such as price and the number of amenities available at the property were considered. A multivariate regression model took these factors as the inputs to predict the number of reservations.

7.2 Future Work

For the applications that have been developed, the future work that can be incorporated is:

- ***Natural Language Processing−*** More insightful results can be obtained if comments received from the users of Airbnb can be taken into consideration by the developed systems. For that, it is necessary to perform Natural Language Processing on each comment given by the reviewer. This can be done by using either of the two approaches: a rule based approach or by developing a machine learning algorithm.
- ***Developing an UI−*** To take the project onto the next level, a user friendly UI should be developed. A User Interface or UI is important because it gives a user the benefit of interacting with the product to fulfill their purpose. To build an effective UI, JavaFX Scene Builder can be used where one can drag an UI component and drop it in their Content panel without worrying about the code. According to the business demand, UI of this application can be developed and hence this work has been left for future.

8 CONCLUSIONS

For this project Airbnb datasets for two cities- London and New York City were selected. These datasets were then thoroughly studied and explored and then a number of visualizations were made to understand the trends for types of properties, famous neighborhoods, prices and reservations. Based on the analysis and visualizations, two applications were further developed. Firstly, for identifying famous neighborhoods, a clustering model was built based on several factors such as number of listings, number of reviews, maximum price, average rating and a few more. Further, a multivariate regression model was built by considering the preference of neighborhoods along with acceptance, response rate of a host, accommodates, beds and a few more factors to predict prices of various listings. The second application was developed to help boost the number of reservations using multivariate regression. Factors such as the price and the number of amenities available at the property were observed to affect the amount of reservations the most.

REFERENCES

[1] 2021. Inside Airbnb - Adding data to the debate. http://insideairbnb.com/get-the-data.html.

[2] Moloud Abdar, Kuan-Hua Lai, and Neil Y. Yen. 2017. Crowd Preference Mining and Analysis Based on Regional Characteristics on Airbnb. In *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*. 1–6. https://doi.org/10.1109/CYBConf.2017.7985771

[3] Jasleen Dhillon, Nandana Priyanka Eluri, Damanpreet Kaur, Aafreen Chhipa, Ashwin Gadupudi, Rajeswari Cherupulli Eravi, and Matin Pirouz. 2021. Analysis of Airbnb Prices using Machine Learning Techniques. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 0297–0303. https://doi.org/10.1109/CCWC51732.2021.9376144

[4] Donghun Lee, Woochang Hyun, Jeongwoo Ryu, Woo Jung Lee, Wonjong Rhee, and Bongwon Suh. 2015. An Analysis of Social Features Associated with Room Sales of Airbnb. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work amp; Social Computing* (Vancouver, BC, Canada) *(CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 219–222. https://doi.org/10.1145/2685553.2699011

[5] Yang Li, Quan Pan, Tao Yang, and Lantian Guo. 2016. Reasonable price recommendation on Airbnb using Multi-Scale clustering. In *2016 35th Chinese Control Conference (CCC)*. 7038–7041. https://doi.org/10.1109/ChiCC.2016.7554467

[6] Erika Rawes and Tyler Lacoma. 2021. What is Airbnb? What to know before becoming a guest or host.

[7] Hao Yan and Xing-chun Diao. 2008. The Design and Implementation of Data Cleaning Knowledge Modeling. In *2008 International Symposium on Knowledge Acquisition and Modeling*. 177–179. https://doi.org/10.1109/KAM.2008.114

[8] Ang Zhu, Rong Li, and Zehao Xie. 2020. Machine Learning Prediction of New York Airbnb Prices. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*. 1–5. https://doi.org/10.1109/AI4I49448.2020.00007