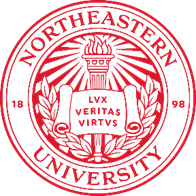
**CAPSTONE PROJECT DRAFT REPORT**

****

**Northeastern University, Boston**

**College of Professional Studies**

**ALY6140 70584 Analytics Systems Technology**

**Professor:** Prof.Richard He

**Date of Submission:** 12/02/2021

**Project Members:**

Sakshi Grover

Durga Bhanu Nayak

Rohith Mallula

Abhay Somani

**INTRODUCTION**

The era that we all are living in is being driven by the data. And it is being generated in enormous volume and speed. We can collect the data that is required using various tools and methodologies and use it to extract meaningful results and can make well informed decisions. And to share our findings and insights to those who need to take data driven decisions, data visualization comes to the picture. A compelling and meaningful visualization of data will help the audience or stakeholders to capture the gist of the analysis.

The purpose of this draft is to explore a public dataset from Airbnb for Amsterdam location using basic data exploratory techniques. Cleaning of dataset by dropping unused variables, converting the variables as per the required data types, checking for null values etc. are some of the processes that are included in the data cleaning process. After the data cleanup we will try to understand the cleaned data more by applying the visualization techniques and plot graphs to answer our questions to the dataset.

**Why did we choose Airbnb Amsterdam?**

Airbnb is an online platform actively working in the short-term rental property sector. It facilitates people around the globe to rent out space for temporary accommodations, which presents an excellent choice for those who are commuting to new cities and looking for short stays. We have chosen this dataset to understand and analyze the current statuses of the company's data with respect to its available features and prices provided to its customers. This dataset lists the company's activities in Amsterdam for the year 2021. As it includes various categorical (such as amenities of the listed accommodations, neighborhood overview, location, hostname, host response time, etc.) as well as numerical variables (rating/ review, pricing, number of bedrooms, number of bathrooms, dates of availability, and many more) which makes it apt for performing and presenting our analysis on this dataset.

**EXPLORATORY DATA ANALYSIS**

1. **Data Extraction**

According to Maina, S. (2021, June 8). *What is Data Extraction,* “Dataextraction involves pulling data from different sources and converting it into a useful format for further processing or analysis. We might need to combine data that is available in multiple file formats such as JSON, XML, CSV and SQL.” Our dataset for Airbnb is available on http://insideairbnb.com/get-the-data.html is available in csv format.

1. Loading python libraries

Pandas and NumPy are libraries used for the data analysis

Seaborn and matplotlib are the libraries used for the data visualizations

Capstone\_group10 utility file is being used to invoke the self-defined functions

1. Reading the csv file

For reading the data set, we have created a method called rd\_csv\_file(path) in our utility file which takes the path of the file as an argument and returns a dataframe after reading the csv file. The first line contains the variable names. Overall, our data set has 16116 entries and 74 variables excluding the header.

1. **Data Cleanup**

As a part of data clean up strategies, we have done the following steps for the chosen data set.

### Converting the price variable to float

### In the given dataset, the price variable is a string value which we have converted to float for the purpose of descriptive analysis and data modelling based on prices.

1. Handling the missing values

Variables such as bathrooms, bedrooms, beds, review\_scores\_rating, review\_scores\_accuracy and many others have a lot of null values present in them. Since such fields are numeric values. So, we have also replaced the null values wherever present in the numerical variables from the dataset as 0.

1. Choosing the relevant columns

There are 74 variables as part of this dataset out of which we have chosen only a few as part of our study that made sense to us for the exploratory analysis and predictive modelling. The filtered variables are below along with the reason why they have been chosen as part of this study

* ***'Id’: 'name'* (**For unique identification)
* ***'Neighbourhood\_cleansed'* (**For understanding the neighborhood of the given Airbnb)
* ***'Latitude’, ‘longitude' (***For understanding diversity of various types of Airbnb in various locations)
* ***'Property\_type','room\_type','accommodates','bathrooms','bedrooms','beds','amenities'*** (For understanding the type of given listing it is and analyze the price of the listing based on these variables
* ***'Price' (***For understanding the price based on other variables for various Airbnb and predicting future prices)
* ***'Minimum\_nights','Number\_of\_reviews','review\_scores\_rating','review\_scores\_accuracy','review\_scores\_cleanliness','review\_scores\_checkin','review\_scores\_communication','review\_scores\_location','review\_scores\_value’ (***For understanding the feedback of various Airbnb in the Netherlands in the given dataset based on these variables.)

1. Converting date to correct datetime format

The date columns in the dataset are in string format. I have tried to modify the column to correct date time format for our analysis.

**DESCRIPTIVE ANALYSIS**

Descriptive analysis is all about converting the raw data into a form that will make it easy to understand & interpret which means rearranging, ordering, and manipulating data to provide insightful information about the given dataset.

As part of this dataset, we have performed various aggregations to understand the dataset better

**Calculating top 5 listings with highest estimated revenue**

We have tried aggregate the estimate revenue for every listing\_id to to understand the top 5 highest revenue listings in Amsterdam which shown below in Table 1

*Table 1: Top five highest revenue listings*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | listing\_id | number\_of\_reviews | minimum\_nights | accommodates | bedrooms | beds | estimated\_revenue |
| 8008 | 19183967 | 103 | 300 | 4 | 1.0 | 2.0 | 4511400.0 |
| 7818 | 18816109 | 24 | 1001 | 4 | 2.0 | 1.0 | 2906904.0 |
| 2846 | 7332860 | 333 | 180 | 1 | 1.0 | 1.0 | 2697300.0 |
| 2868 | 7382264 | 321 | 180 | 2 | 1.0 | 1.0 | 2600100.0 |
| 352 | 853645 | 70 | 30 | 4 | 3.0 | 2.0 | 1680000.0 |

**Top 5 highest revenue listings with minimum\_nights 7 or less and then 4 minimum\_nights or less**

As part of this tabulation, we have aggregated the listings based on minimum nights and then relate the estimate\_revenue with respect to the minimum nights. Here are top five listings with highest estimate\_revenue with minimum number of nights 7 or less. We have done the same for minimum number of nights 4 or less

*Table 2: Top five listings with highest estimate revenue listings with minimum\_nights 7 or less*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | listing\_id | number\_of\_reviews | minimum\_nights | accommodates | bedrooms | beds | estimated\_revenue |
| 353 | 854328 | 330 | 7 | 4 | 3.0 | 3.0 | 678608.0 |
| 154 | 510836 | 110 | 7 | 7 | 5.0 | 8.0 | 561330.0 |
| 320 | 785432 | 804 | 2 | 4 | 2.0 | 3.0 | 325600.0 |
| 448 | 1044452 | 286 | 7 | 2 | 1.0 | 1.0 | 300300.0 |
| 666 | 1601408 | 224 | 4 | 4 | 2.0 | 2.0 | 273600.0 |

*Table 3: Top five listings with highest estimate revenue listings with minimum\_nights 4 or less*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | listing\_id | number\_of\_reviews | minimum\_nights | accommodates | bedrooms | beds | estimated\_revenue |
| 320 | 785432 | 804 | 2 | 4 | 2.0 | 3.0 | 325600.0 |
| 666 | 1601408 | 224 | 4 | 4 | 2.0 | 2.0 | 273600.0 |
| 290 | 754613 | 278 | 4 | 4 | 1.0 | 3.0 | 259200.0 |
| 4825 | 12596322 | 173 | 3 | 16 | 5.0 | 17.0 | 239184.0 |
| 105 | 327285 | 535 | 3 | 4 | 1.0 | 4.0 | 237540.0 |

**Understanding the correlation between Estimated revenue and Minimum nights**

Before understanding this correlation, we removed the minimum\_nights of 1000 and 1001, which we are assuming are the outliers in the dataset for this variable, as these many numbers of minimum\_nights are huge for any listing and are rare possibilities. Hence, we are considering them as the outliers.

*Table 4: Correlation between minimum nights and estimated revenue (filtering min night 1000,1001 as outliers)*

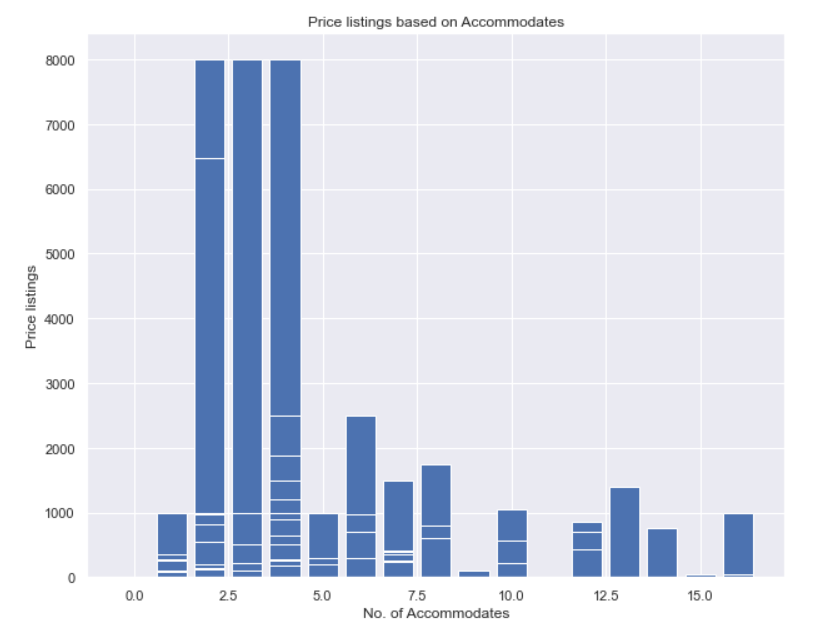
|  |  |  |
| --- | --- | --- |
|  | minimum\_nights | estimated\_revenue |
| minimum\_nights | 1.000000 | 0.087199 |
| estimated\_revenue | 0.087199 | 1.000000 |

The correlation between “minimum\_nights'' and “estimated\_revenue” is 0.087. This correlation indicates that the relation between the two variables is positive but doesn’t have that much impact i.e there is little association between the minimum nights at the given listing and estimated revenue of that listing.

**DATA VISUALIZATION**

After data extraction and its cleanup, we are now moving ahead to explore the data and preparing visuals to present our findings based on it.

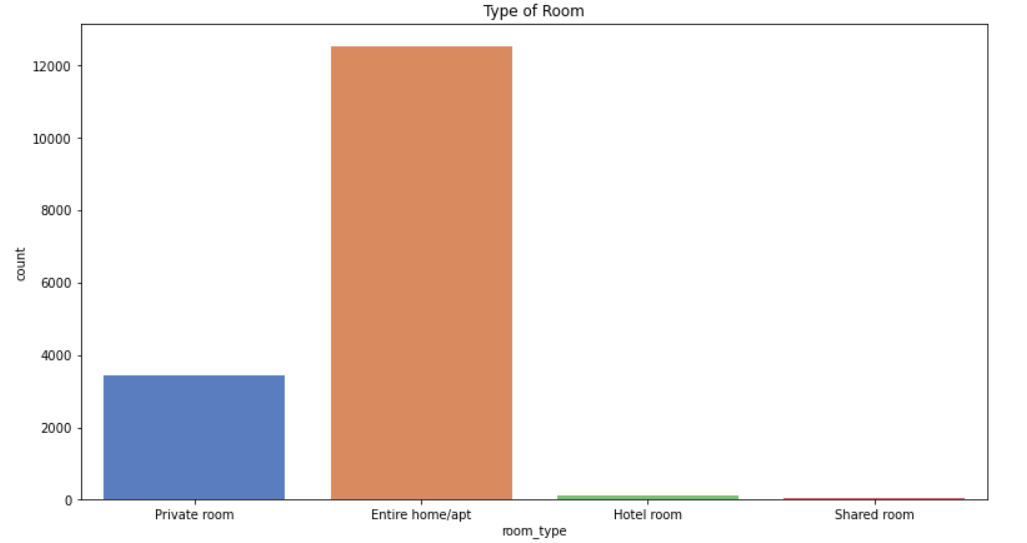
While data exploration, we first wanted to check how the Airbnb booking prices are getting affected with respect to the number of people visiting Amsterdam. Fig. 1 is portrayed to display the price range among the number of people renting out the properties in Amsterdam via Airbnb. According to the figure, we can see that most of the price is listed for couples and small families with under 5 people. This also suggests that as a higher number of people as a group are visiting Amsterdam, they are comfortable renting out cheaper properties and mostly focused on sightseeing.

****

*Fig. 1: Price listings compared to number of accommodates*

After getting the information about the price variations corresponding to the people groups who are making their bookings with Airbnb, we are now interested to check the type of property they have made the bookings for in the year of 2021 for their stay.

Here, Fig. 2 gives us an idea about the property/room type that has been preferred by people to make bookings for their stay in Amsterdam. We can clearly observe that instead of choosing hotel rooms or shared rooms, most of the people have chosen an entire apartment to stay for their visit in Amsterdam. People were also interested in booking private rooms rather than booking shared rooms. As we have seen in Fig.1, most of the bookings have been made either by couples or small families or small groups of people (< 5), so it might have been suitable for them to book a private room or apartment instead of sharing with strangers.

****

*Fig. 2: Frequency of property type booked via Airbnb in 2021 for Amsterdam*

As we are performing the analysis for the whole year of bookings through Airbnb in Amsterdam, a logical question arises, that how the overall bookings are being compared across each month and in which month was the profitable one for Airbnb. Using Fig.3, we have put across multiple charts displaying the number of bookings, revenue generated, and average booking price across each month. Here, we can observe, people did love to visit Amsterdam in the late Summer and Fall season (July till October). Most of the bookings have happened around this time and that has a positive correlation with the revenue generated. Also, we can see that the average booking price around this season is comparably lower than other months. However, we can also observe that around Winters and Summers, property booking prices are higher on average which might result in lower bookings around such time.

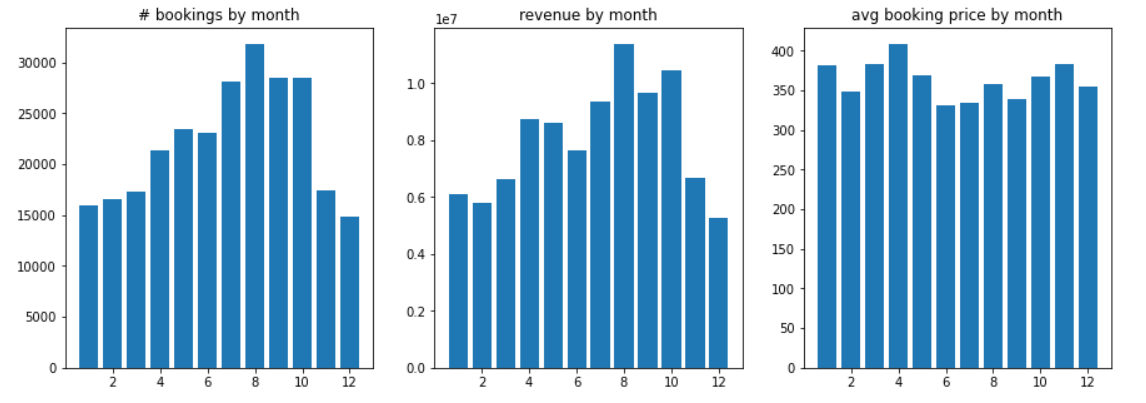
*****Fig. 3: Monthly property booking stats by Airbnb*

Fig. 4 helps us understand the distribution of property types booked in Amsterdam through Airbnb in the year of 2021. Here the properties are being located using the location coordinates across a map. As we can see, most of the people had opted to book an entire apartment followed by private rooms. And a very few bookings can be seen for the hotel rooms or the shared rooms. As per the geographical area, we can see that, in most of the central and northern parts of Amsterdam booking an entire apartment/ home seems to be more convenient than booking other room types.

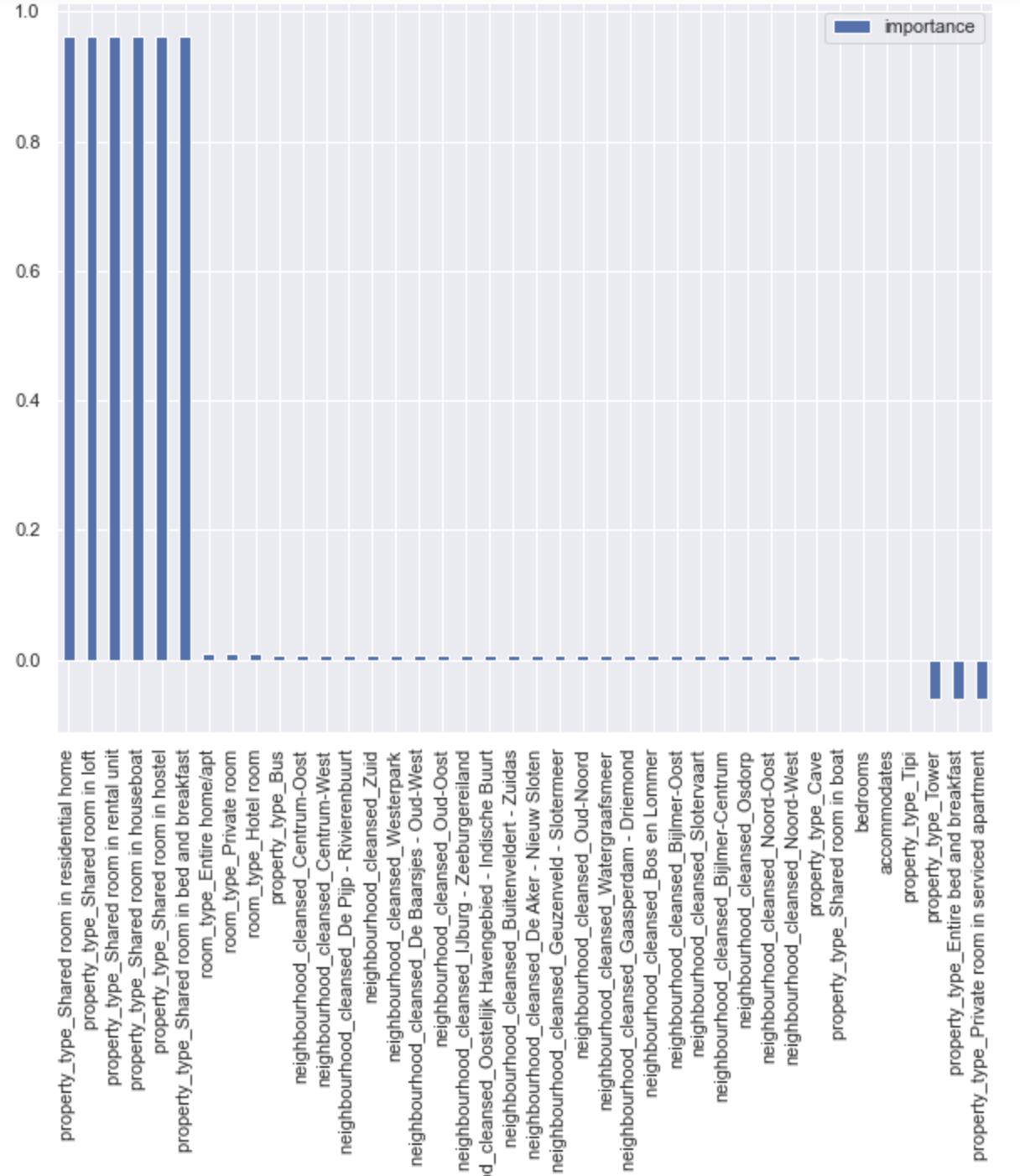
****

*Fig. 4: Property types distributed across Amsterdam*

**PREDICTIVE MODEL**

The idea behind our project is to predict the listing price of Airbnb. Which means the price of an Airbnb listing per night? While our initial data has a little over 70 features it is very important for us to initially identify and understand what the most important features among those are. We have decided to normalize and encode certain character features like neighborhood\_cleansed, property\_type and room\_type. After using the min\_max\_scalar, we created appropriate dummy variables. This final data frame comprises our independent features while the price remains as our dependent variable that we try to fit out models on.

Computing a **linear regression** to find the top features of our model yielded in the following graph. This can be further fine-tuned if the neighborhood and property type is factorized.



We use the sklearn' preprocessing library and split the data into training and test data with the train\_test\_split*()* function. After that, we can utilize various CART (classification and regression models) like decision trees, random forests, and gradient boosting to predict the price of a new listing.

We use R-squared and MSE (Mean Square Error) to monitor our model’s performance and compare them against each other.

**CONCLUSION**

For the dataset of Airbnb Amsterdam data, so far, we have performed exploratory data analysis using Python programming language. While doing so, we imported the dataset using the Pandas library and tried to analyze it. In the process, we came across various variables which do not have any data or are missing a lot of values. We cleaned the dataset using various strategies, removing the unnecessary columns and imputing values wherever required. As a part of Descriptive analysis, we tried to understand the relationship between minimum nights the listing was booked and estimated revenue of that listing. While performing various data visualizations over the cleaned data, we found that most of the bookings were made by groups of people ranging from 2 to 5. Also, most of the bookings happened around the month of August where people have preferred to rent an entire house than a shared place or a hotel room.

After performing our basic data exploration, we have decided to implement the predictive modelling algorithms for price prediction. Up until now, we have applied the linear regression model to find out the top features of our model.

As a team we are now actively working on implementing the rest of the models mentioned in our proposal using named K-nearest neighbors, Decision Tree and Gradient Boosting (xGBoosting) over our dataset to understand other factors in booking the listings given in the dataset.

**BIBLIOGRAPHY**

1. Maina, S. (2021, June 8). *What is Data Extraction? A Python Guide to Real-World Datasets*. Medium. https://towardsdatascience.com/what-is-data-extraction-python-review-of-json-xml-apis-sql-and-csv-formats-a5470afc27b6
2. Rawat, A. S. (n.d.). What is Descriptive Analysis? - Types and Advantages | Analytics Steps. Unknown. https://www.analyticssteps.com/blogs/overview-descriptive-analysis
3. *sklearn.linear\_model.LinearRegression*. (2007). Scikit-Learn. Retrieved 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html
4. *Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data*. (n.d.). Tableau. https://www.tableau.com/learn/articles/what-is-data-cleaning