ANLY 506-91

Course Project Status Report

New York City Airbnb Data

Prepared by:

Srikanth Goli

Ramaswamy Iyer

Abinaya Karunagaran

Abisheik Mani

Shuang Tao

Pradeep Paladugula

- **What is the problem you are trying to solve or question you are trying to answer?**

The data set contains 48,000 Airbnb listings in New York City. It also contains other metrics such as information about hosts, geographical location, type of rental, availability, reviews and price. The goal of this project is to predict the price of Airbnb rentals in New York City based on the available data.

Link to the data: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

- **Relevant background information including any relevant literature you have/will use**

Airbnb is an online ecommerce company serving the industry of lodging and hospitality where people can list, discover and book unique rooms around the world. Whether travellers want an apartment for a night, a castle for a week, or a villa for a month, San Francisco based Airbnb began operations in 2008 and currently has thousands of employees across the globe supporting property rentals. The company does not own any real estate but simply acts as a broker and makes money in the form of commissions from each of these rentals.

- **The overall process you will follow for the entire project**

The group follows the process for data acquisition, data cleaning, and exploratory data analysis on the Airbnb New York data. After that we will perform regression modelling on cleaned data, and do analysis on modelling results. We will also compare different modelling results and find the most appropriate one as our predicting model.

- **A description of relevant, interesting exploratory data analysis**

We will perform exploratory data analysis tools such as graphical visualization (scatter plot, box plot, historical plot, ggplot, hierarchical clustering), transformation for key variables in our dataset, to clean for outliers and meet the model assumptions. After pre-processing the data, we will use PCA and backward variable selection techniques to choose the independent variables.

- **A description of the methods/techniques/tools/algorithms you have/will use to complete the project**

The group has performed data cleansing on the dataset which includes handling bad data, doing transformations on the variables converting to appropriate data types for performing analysis further downstream. After completing the transformations, exploratory analysis is performed on the Airbnb dataset to discover interesting insights in the NY housing / rental market. Some of the EDA analysis include correlations/covariance between variables, visualizing the variables through multiple plots, etc.

After completing the EDA portion, the team is going to work on performing regression analysis by utilizing predefined linear regression models which are available in the R language. In the linear regression model, we are going to test statistical significance between variables from the dataset. After completing the regression analysis, a detailed write-up documentation is created along with the PPT presentation summarizing the insights the team discovered along the process in this project.

- **Description of challenges working on this project**

Although there is a rich collection of data in the NYC Airbnb dataset for 2019 listings, the lack of historical data makes it difficult to compare past trends or patterns. Hence the baseline for the analysis would have to be determined from our dataset based on common statistical measures such as median price by neighbourhood group in NYC or from any other open source references on the internet (such as prevailing short-term rental rates prior to 2019). Also, there are no listing dates available within the dataset to further explore rental trends by season or specific months of the year etc. Lastly, certain variables were excluded as they were not relevant to the scope of our analysis (such as last reviewed date) or were kept out for privacy reasons as best practices always recommend not using names or other confidential information. Overall, this project has helped the team so far in

gaining interesting insights about the short-term rental landscape in one of the busiest cities as well as using appropriate exploratory data analysis techniques while dealing with different data formats.

- **Discussion of the parts of the project that have been completed**

Once the project team zeroed in on the NYC Airbnb dataset, we did some basic research on the short-term rental market and Airbnb's role. Then we did some pre-processing steps to get a better understanding of the data formats and types used in the dataset as well as cleaning the dataset to handle missing and inaccurate data, if any. As part of the exploratory analysis we have looked at price and availability trends by neighbourhood group, top hosts, room types etc. to name a few. We intend to use heat maps to further explore the geographic distribution of key rental attributes. It will be followed by the use of appropriate regression modelling to predict the price of rental listing based on several factors.

- **Discussion of the parts of the project that remain to be completed**

We observed through Exploratory data analysis how geographical distribution of key rental attributes works with key variables such as price and availability trends by neighbourhood group, top hosts, room types. The question which still remains to be answered is if the linear regression model we found is a fit for the Airbnb data. If not, we can proceed ahead with a logistic regression model analysis in case it will give us more accuracy from our results and more of a statistically significant model. The next few weeks should have expectations for a rest of the paper including a final paper writing/detailed report and correction if any is needed.

- **A discussion of how you will finish the final project report and presentation**

We will be splitting up into subgroups, few people will work on the modelling and rest of them on eda analysis. Depending on the type of analysis and the type of response variable we will determine

the type of regression model we are going to choose. We plan to maintain APA format for graphs and tables. If we had missed out any statistical analysis or any reports that will be included in the final report. When it comes to power point presentation, all the statistical/graphical analysis performed on the data set won't be included, we planned to use only analysis (ex, graphs) that strongly supports our project work, elaborate full subject knowledge and helps to conclude our experiment on Airbnb data.