Visualization of New York Airbnb Data

Heying Yu, Liu Wu, Jiayao Wang

**Link to the project repository:**

<https://github.com/HeyingYu/Data-Visualization-Project>

**Link to the project dataset:**

<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

**Background and Motivation:**

For many years, economic researchers have been interested in the survey of contributing and dominant factors towards the price of housing as well as that of hotels. In recent years, as Airbnb introduced a new way of travel accommodation, predicting and investigating Airbnb prices started to draw our attention.

In this project, we are trying to visualize features of the Airbnb dataset such as the location, convenience, availability and necessary metrics of the room and use the visualization to gain more insights in making predictions of the price. Moreover, we are interested in Machine Learning and we think it would be beneficial if we can use visualization to guide or aid our machine learning model building or if we can use machine learning tools to make predictions or clustering and generate a new form or feature for visualization.

In class, we learned about geographic visualization such as Cartograms and Choropleth Maps and we plan to apply the knowledge and skills learned in class into this project to visualize the distribution of the Airbnb houses and to gain insights into how the geographics will affect the price of Airbnb.

**Objectives:**

We would like to gain insights into the contributing factors of Airbnb housing to its price. In order to achieve the goal, we are planning on using different forms of visualization to try to analyze the data in different ways. This will help us to develop our skill for data visualization and analytics, to have a better understanding of how the price is determined and what are the important factors of housing price, as well as to learn more visualization techniques such as encoding Cartograms and Choropleth Maps as well as using different APIs to help to build our project.

**Data:**

We are using the New York Airbnb dataset to develop a model that can predict the price of the Airbnb rooms given specific attributes. One other major goal we have is to determine which are some features that have the most impact on the price. The link of the dataset is<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>. This dataset includes features such as the location of the room (neighborhood, latitude, longitude, etc.) and the number of reviews of the room.

**Data Processing:**

In order to perform the analysis, we will first clean the data by removing data points that have missing values or corrupted values. Currently, the dataset has about 50,000 entries and each entry has 16 attributes. We will perform feature selection to remove unrelated features such as hostname, and the host\_id. We will spend about a week to clean and preprocess the data before we start our visualization and analysis.

**Must-Have Features:**

The following features are absolutely necessary or highly possible related to our goal for the project:

1. Availability
2. Neighborhood
3. Latitude
4. Longitude
5. room type
6. Minimum nights
7. Price
8. id

**Optional Features:**

The following features are optional:

1. Number of reviews
2. Reviews per month
3. Calculated host list

**Project Schedule:**

**Week 1, October 25th**

Discuss the data we are going to use and the general plan to visualize them. Finish the proposal. Download the dataset and set the check point and finish the proposal

October 25th: Team Formation

November 1st: Proposal Submission

**Week 2, November 3rd**

Trying to design and visualize the following groups of attributes and predicts what we can get from visualization:

1. Latitude and longitude with the price (Liu)

2. Neighborhood and price (Heying)

3. Availability and Price (Jiayao)

4. Room Type and Price (Liu)

5. Minimum Night and the Price (Heying)

6. Pick some optional features that might lead better result.(Jiayao)

November 8th Update

**Week 3, November 10th**

Finish visualizing the data, make assessment and modify the poorly designed graph, add more groups if necessary.

**Week 4, November 17th**

Design and make interaction between each pair of groups.

Optional: We are considering training the data in machine learning program. If it is applicable, we might finish the simple training and draw decision boundary this week and next week

November 18th Prototype

**Week 5, November 24th**

Draw Conclusion, write a report about what we find after visualizing the data. What advice we can give to the Airbnb house owner.

**Week 6, December 1st**

Final modification(All members)

Prepare for the presentation. (All members)

December 2rd Presentation

**Week 7, December 8**

Wrap up the project and get it ready for submission.

November 9th: final project submission

**Visualization Design:**

**Design 1:** We will use a scatterplot matrix in order to explore the relationships between all possible pairwise combinations of variables, including price, availability, neighborhood, room type and minimum nights.

According to the result scatter plot matrix, the following questions will be considered:

1. Is there a covariance between any pair of attributes?
2. Which pair has the strongest relationship?

The interaction with the scatterplot matrix is designed as follows:

If you select a point in one scatter plot, the corresponding point will be selected in all other scatter plots. This interaction may be convenient for future research on a specific object, such as outliers.

**Design 2:** To make a correlation between different attributes visible, parallel coordinates could be used instead with the same five variables described above.

The way to interpret a parallel coordinates is as follows:

1. If two neighboring axes have high positive correlation, the line segments are mostly parallel.
2. If two axes have a high negative correlation, the line segments mostly cross over each other at a single spot between the axes.
3. The pattern in between uncorrelated axes is a mix of crossing angles.

Besides, the parallel correlation also provides an overview of all attributes, finding the range of individual attributes, selecting a range of items and outlier detection.

**Design 3:** Another alternative design is to use a heatmap, where the amount of variables is narrowed down to three--price, availability, and neighborhood. To be more specific, the keys are neighborhood (x-axis) and availability (y-axis), and the quantitative value attribute is the price of the room in its neighborhood, with its availability. The heatmap could use a diverging red-green colormap. The benefit of the heatmap is that visually encoding price with color using small area marks is very compact so that providing overviews with high information density.

**Conclusion:** the final design could be some amalgamation of the previous three design. The specific content depends on the results of the previous three alternative designs. First, using a scatterplot matrix and parallel coordinates, we pick two keys that have the most impact on price, since our goal is to predict prices for rooms with the help of different variables. Finally, we will show the visual encoding of how price is affected by two keys using heatmap.