Process Book

Overview and Motivation

The goal of the Airbnb data visualization is first to show users a geographic distribution of Airbnb properties in New York city and second to allow people to explore relationship among Airbnb's attributes including price, location, yearly availability, minimum nights and reviews per month. What motivates us to do this data visualization is that we are eager to figure out if there exists an attribute correlated with the price attribute of Airbnb.

Related Work

A web site provides us the dataset of Airbnb, including all needed information to find out more about geographical availability, necessary metrics to make predictions and draw conclusions. Inspired by the scatterplot matrix visualization discussed in class, we think scatterplot is a great way to present all possible pairwise combinations of attributes (price, yearly availability, minimum nights and reviews per month).

Questions

The questions we are trying to answer are if there exists an attribute which shows strong correlation with price attribute of Airbnb and what it would be. At first, with data too big, it is very hard to find a connection among all attributes. The only thing we notice is that the price appears to be high at certain region of longitude and latitude. Thus we think it is possible that the price is affected by the location of Airbnb. So we paid more attention on the geographic distribution of Airbnb at these locations. Finally, we find out that most of the Airbnb properties are located nearby the metric system of New York city. After that, we were trying to figure out if there exists a relationship between price and other attributes.

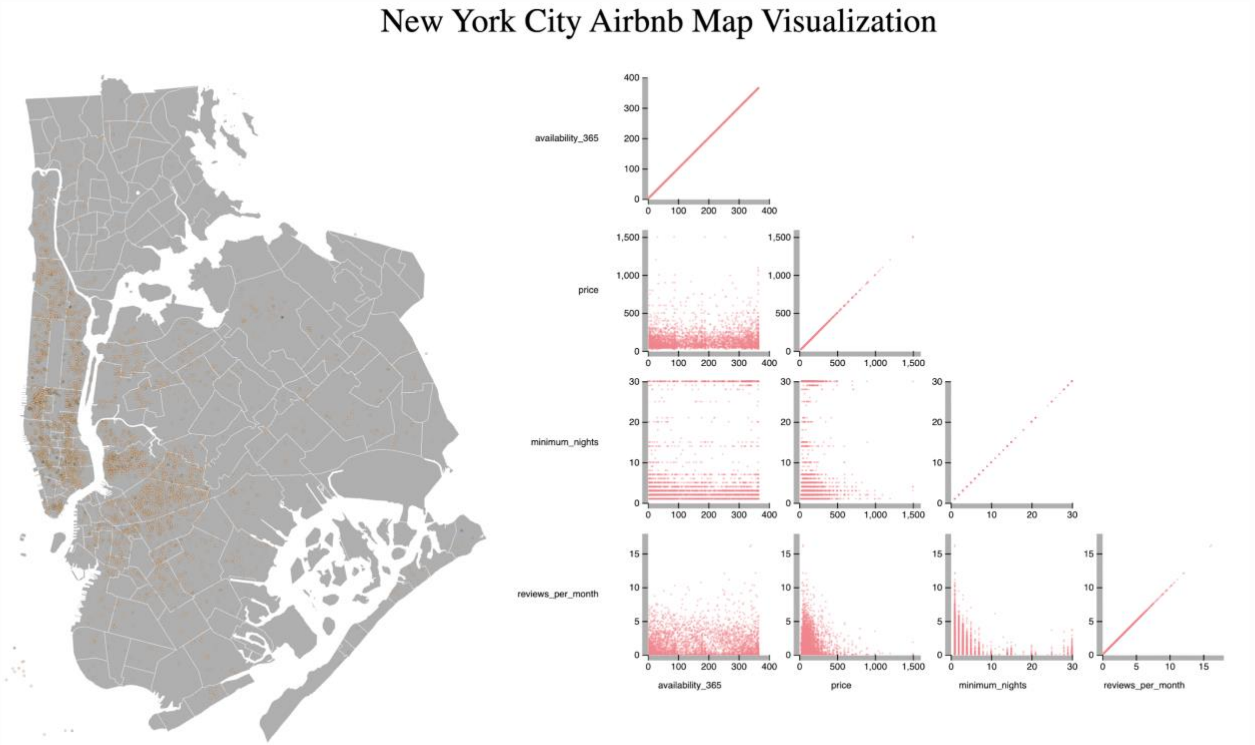
Data

Since the Airbnb dataset is very large, about 50000 objects altogether, it's important to do data preprocessing and cleaning. First of all, we deleted the irrelevant features, such as host ids. Then we removed the noises, like removing data entries with yearly availability larger than 365 days. Due to the very broad range of prices value, the outliers are defined as prices larger than

95% and then deleted. Finally, to decrease the size of dataset, we shuffled the data randomly and pick 12.5% out of the result.

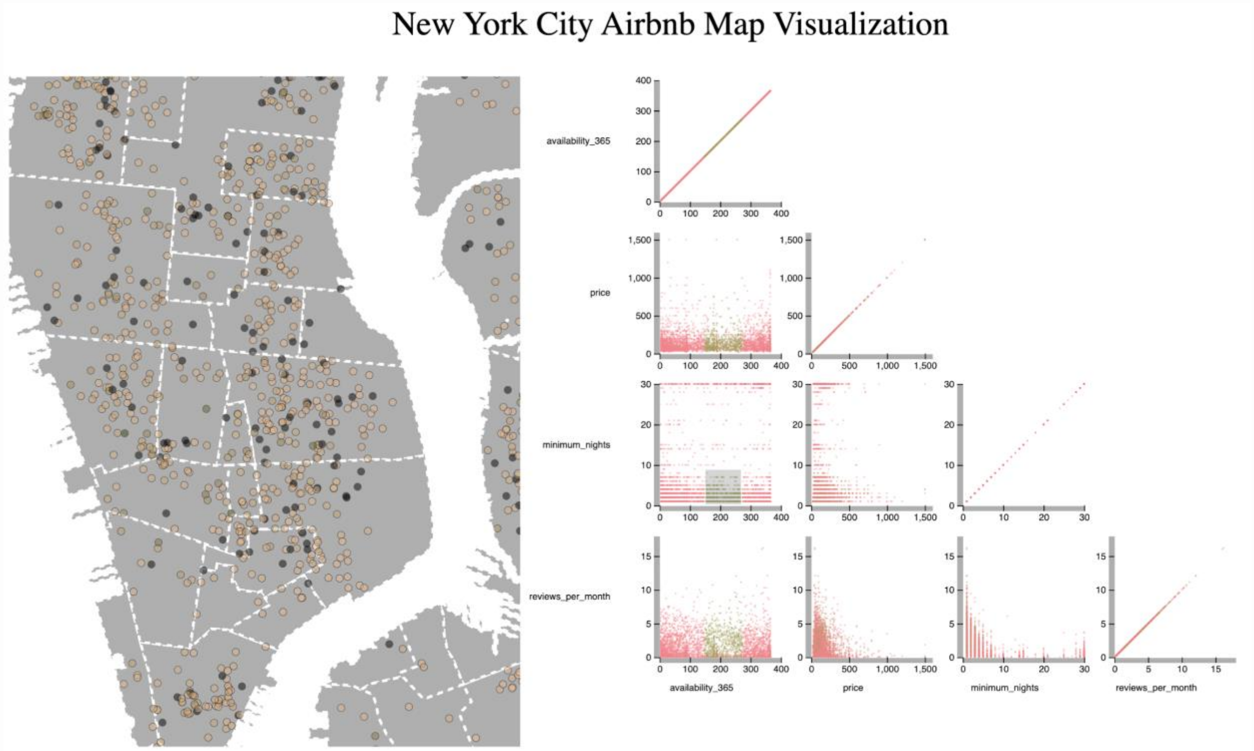
Implementation

The intent of our project is to visualize the New York City Airbnb data and see if there are some correlations between each of the Airbnb features. We implemented mainly two visualizations: the map and the scatter plot as shown below. This visualization is intended to help the user to look at the data encoded with geographical information in more detail and be able to find trends that exist in the data.



The map visualization consists of all the neighborhoods of New York City (except Staten Island which we removed from the dataset since it does not contain many data points but the area taken lots of space). Each individual dot stands for one Airbnb data and the color is encoded according to the price (the higher the price, the darker the color).

The scatter plot is consisted of 10 subplots which depict the correlation between each pair of the four attributes: availability, price, reviews per month, and minimum night requirements.


New York City Airbnb Map Visualization

The map visualization has a zooming interaction feature: if the user double-click on an area on the map, the map will zoom in and display that area in more detail; if the user double-click again, the map will zoom out and display the original map.

The scatter plot has a brushing interaction feature: if the user brushes on one of the scatter plot, the brushed dots and corresponding dots in other scatter plot will be colored green. The corresponding data points on the map will also be colored black.

Exploratory Data Analysis

Our team initially used Pandas Jupyter Notebook to load and preprocess data. We used Pandas data frame to load in and analyze the data. We also used libraries such as matplotdib to generate correlation plots between each of the attributes. This helped to determine the importance of each feature and helped our decision to remove or combine certain features and remove outlier data. After that, we used the D3 map visualization and projected all data points onto the New York City map according to their latitude and longitude. We found that the data is clustered in the downtown area and distributed along the New York City metro line. However, the relationship between geographical attributes and the price is not very clear. Therefore, we decided to add a scatter plot to display other features such as availability and minimum night requirements along with the price attributes and try to find trends in the data.

Design Evolution

We first decided to color the Airbnb data points according to the range their price is in and used four colors: red, green, yellow, blue. However, although these colors are distinct from each other and is easy to detect on the graph, it is hard to see the trend of the price. Moreover, data with a price of 50 dollars apart can also be within the same range, which makes the graph lost some information. Therefore, we decided to use a consistent color scale to color the data.

As for the plot that depicts the relationships between certain features. We first used boxplots to display the quantile information of the selected data on the map. However, that will not give enough information to the user. Therefore, we decided to switch to the scatterplot, which is better at rendering relationships between certain features.

As for the interaction of the data, we first used zoom together with the brush on our map and display the brushed data on the scatter plot. However, this prevents the user to select data in nonconsecutive places to display on the scatter plot and this will cause the program to run really slow. Therefore, we decided to use the double click zoom feature and we decided to display all of the data in the scatterplot and let the user decide which part of the data they would like to see on the map.

These changes cause a slight deviation from our original proposal but we think it is a better way to render the data.

Analysis

After displaying all the data points on the map, we found that the New York City has most Airbnb housing in the downtown area near central parks and Time Square. Moreover, the most expensive data are gathered around the financial district and there are less number of data points in that district. We also found through the visualization that the distribution of the Airbnb is along the New York City metro lines. This is a convincing correlation since Airbnb is most for travelers and, since travelers uses the New York City metro lines, the demand for Airbnb that located near the metro lines is large.

We also find through the scatter plot that there are correlations between the number of reviews per month and minimum night requirements: the lower the requirements, the more review the house gets. Moreover, the lower the price, the more reviews the house gets. There are not many correlations between availability and other features.

The strongest relationship between the dataset features is still the longitude/latitude and the price, which indicates that location is the most essential factor that contributes to the price of an Airbnb. However, we can see from the map and the scatter plot that location is not the only contributing factor.

Future Work

Currently, the data visualization works well but there is still a lot of room for improvement. One thing is that the Airbnb data takes a long time to load and to display. We are planning to work on the optimization of the program. Moreover, if we have more time, we can add more interactive features to our visualization to further assist the user to make visual queries.