New York City Airbnb Data Analysis

ENG 498 IM/IMO Final Project

Spring 2021

Haoyong Lan

Project Repository: https://github.com/haoyonglan/ENG-498-Final_Project

Introduction

In recent years, short-term rentals have become popular choices for many people. Airbnb is one of the major companies that provide online rental listing services. According to Jamie (2021), the number of listings on Airbnb has increased more than twice over the prior four years. Compared with traditional hotels, short-term rentals provide more bedrooms and larger living space (Jamie, 2021). Thus, short-term rentals are good alternatives to traditional hotels for travelers.

Data Description

The dataset is originally from Inside Airbnb which can be found at http://insideairbnb.com/. The data file has information about listings and hosts in New York City for 2019. The dataset contains many interesting columns such as price, room_type, latitude, longitude, and neighbourhood_group. And these columns worth further exploration to find more insights. For instance, there could be a potential relationship between listing price and location. As the listing location gets closer to the downtown, the listing price could also increase. The dataset has 48895 rows and 16 columns.

Data Cleaning

Almost every dataset needs data cleaning before it can be analyzed. I loaded the dataset into the OpenRefine to perform the initial data cleaning. Then I trimmed leading and trailing whitespaces on every text column. I also collapsed consecutive whitespaces on every text column. And I found that "last_review" column contains dates, but its data are in text type, so I transformed its data type from text to date.

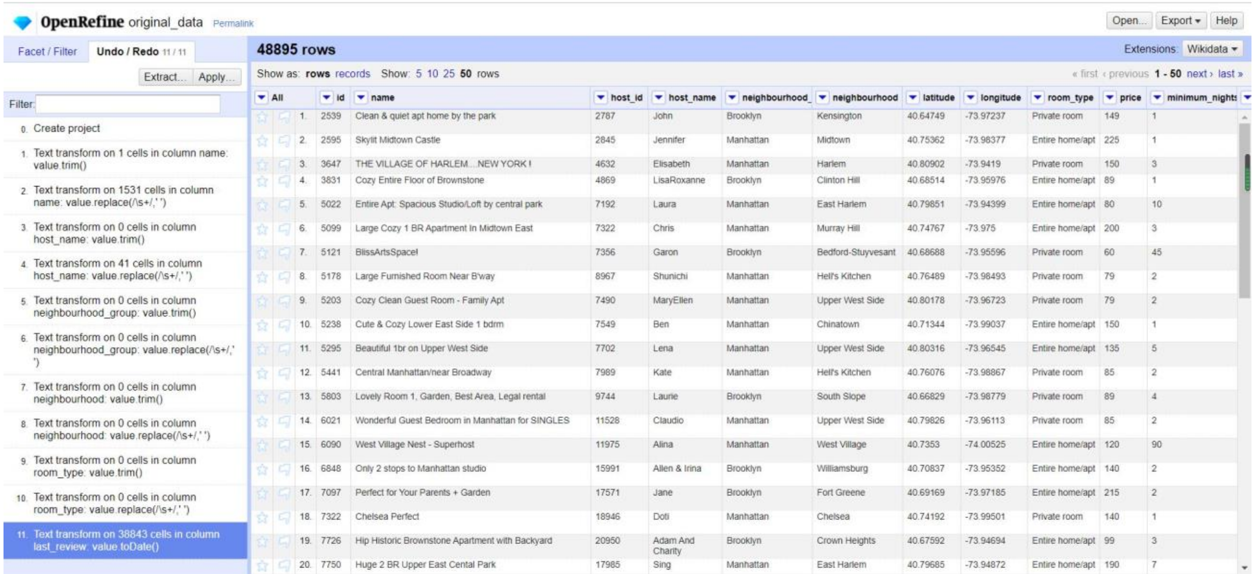

Figure 1. The OpenRefine data cleaning operation history

As seen from Figure 1, 1531 rows of column “name” have consecutive whitespaces and 41 rows of column “host_name” have consecutive whitespaces. After finishing data transformation, I transformed the data type of “price” from text into number and did numeric facet on the “price” column.

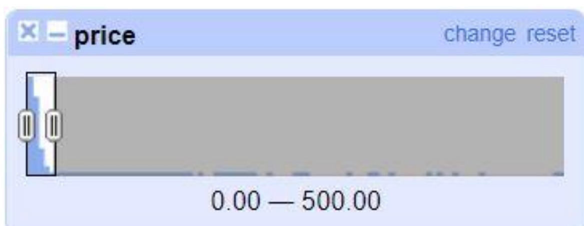

Figure 2. The “price” column numeric facet

The Figure 2 shows that most listing prices are between 0 and 500, so I matched the rows of prices that are between 0 and 500.

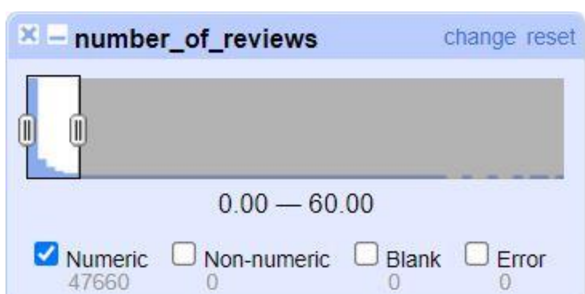

Figure 3. The “number_of_reviews” column numeric facet

According to Figure 3, most listings’ number of reviews are below 60, so I matched the rows of number of reviews which are below 60.

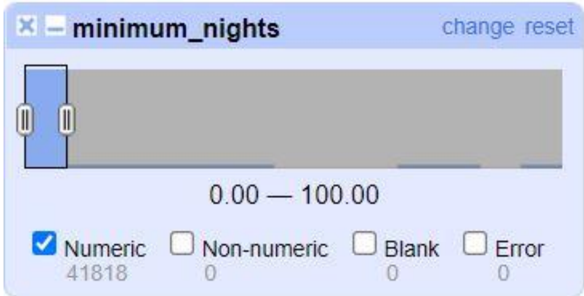

Figure 4. The “minimum_nights” column numeric facet

The Figure 4 shows that most listings’ minimum nights are below 100, so I matched the rows of minimum nights which are between 0 and 100.

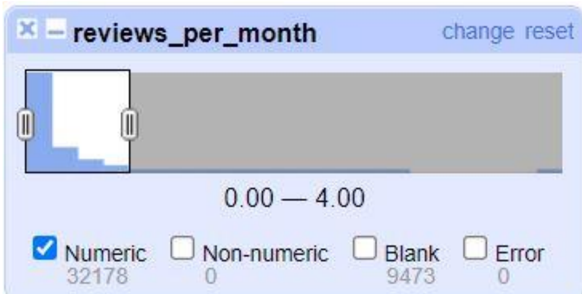

Figure 5. The “reviews_per_month” column numeric facet

As seen from Figure 5, most of its values are under 4 and it has 9473 empty values. Thus, I decided to include non-empty values which are below 4.

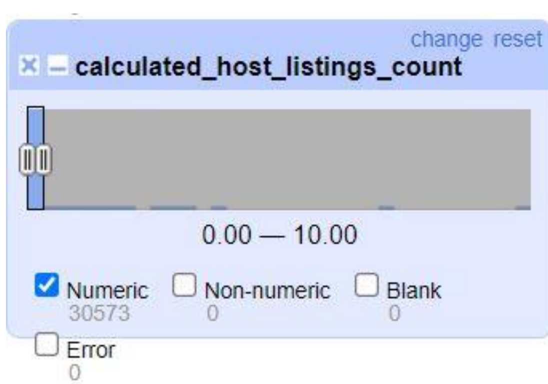

Figure 6. The “calculated_host_listings_count” column numeric facet

The Figure 6 displays that its most values are below 10, so I matched its rows of values which are between 0 and 10.

Conclusion

After finishing the data cleaning by OpenRefine, I loaded the cleaned data into the Python editor and did some data visualizations.

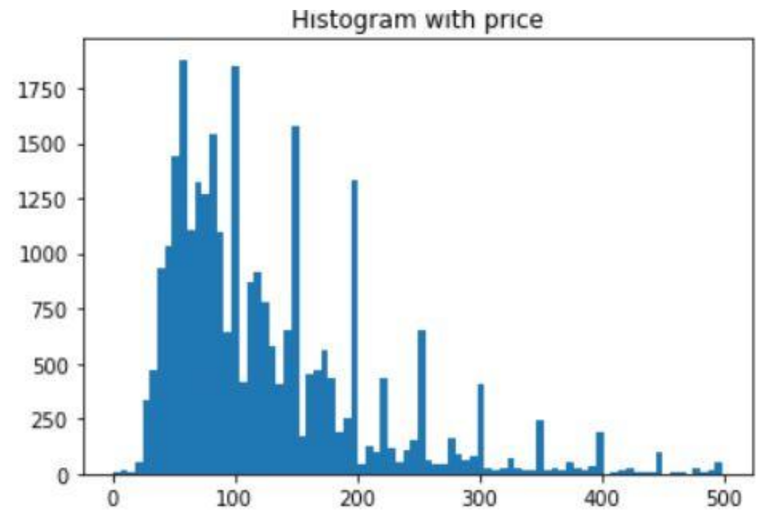

Figure 7. "Price" column histogram

As seen from Figure 7, most prices are between 50 and 100.

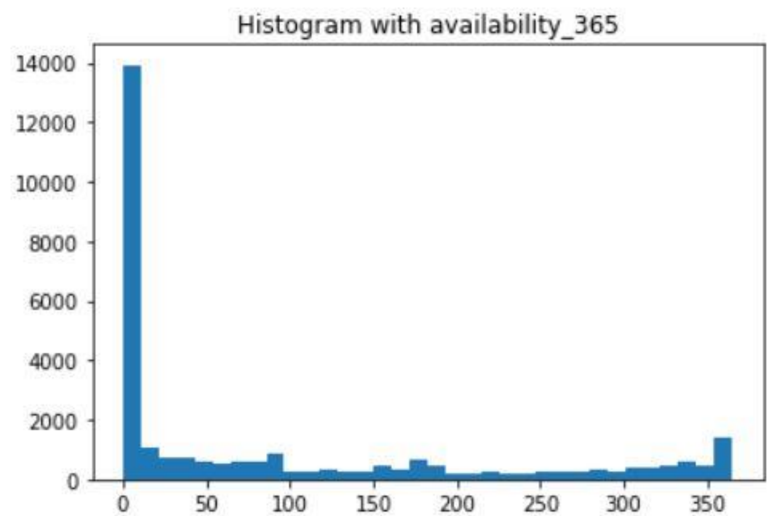

Figure 8. "availability_365" column histogram

The above figure shows that most listings are only available for 1 day throughout the year. To find the relationship between listings location and price, I loaded the cleaned data into the Tableau to create the map visualization with listings price.

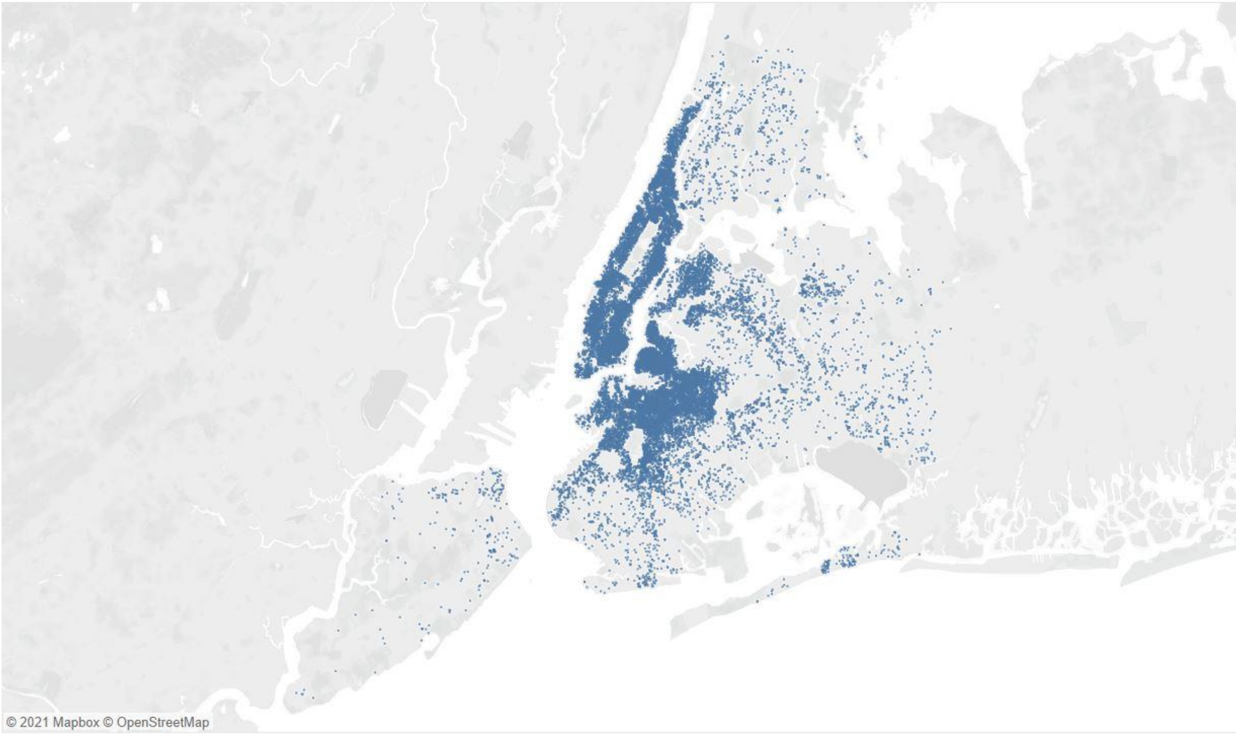

Figure 9. Tableau map visualization with respect to listings price

As seen from Figure 9, each bubble represents one listing location. As the bubble size grows, its price also increases. So, it shows a few clusters which have many expensive listings, such as Manhattan.

References

Gomonov, D. (2019, August 12). New York City Airbnb Open Data. Retrieved from kaggle: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Lane, J. (2021, March 26). Lost Supply? How has the COVID-19 Pandemic Impacted the Supply of Short-term Rentals on Airbnb? Retrieved from AirDNA: https://www.airdna.co/blog/covid-19-pandemic-impacted-the-supply-of-strs