

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](http://insideairbnb.com/) 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.

Final Project Report

Haoyong Lan

OpenRefine original_data Permalink

Open Export Help

Facet / Filter Undo / Redo 11 / 11

48895 rows

Show as: rows records Show: 5 10 25 60 rows

Extensions: Wikidata

Filter:

- Create project
- Text transform on 1 cells in column name: value.trim()
- Text transform on 1531 cells in column name: value.replace(/s+/, '')
- Text transform on 0 cells in column host_name: value.trim()
- Text transform on 0 cells in column neighbourhood_group: value.trim()
- Text transform on 0 cells in column neighbourhood_group: value.replace(/s+/, '')
- Text transform on 0 cells in column neighbourhood: value.trim()
- Text transform on 0 cells in column neighbourhood: value.replace(/s+/, '')
- Text transform on 0 cells in column room_type: value.trim()
- Text transform on 0 cells in column room_type: value.replace(/s+/, '')
- Text transform on 38843 cells in column last_review: value.toDate()

		id	name	host_id	host_name	neighbourhood	neighbourhood	latitude	longitude	room_type	price	minimum_nights
1	2539		Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1
2	2595		SkyIt Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1
3	3647		THE VILLAGE OF HARLEM - NEW YORK I	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3
4	3831		Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1
5	5022		Entire Apt. Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79651	-73.94399	Entire home/apt	80	10
6	5099		Large Cozy 1 BR Apartment in Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/apt	200	3
7	5121		BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45
8	5178		Large Furnished Room Near B'way	8967	Shunschi	Manhattan	HeIt's Kitchen	40.76489	-73.96493	Private room	79	2
9	5203		Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2
10	5238		Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1
11	5295		Beautiful 1br on Upper West Side	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Private room	135	5
12	5441		Central Manhattan/near Broadway	7989	Kate	Manhattan	HeIt's Kitchen	40.76076	-73.98867	Private room	85	2
13	5803		Lovely Room 1, Garden, Best Area, Legal rental	9744	Launie	Brooklyn	South Slope	40.66829	-73.98779	Private room	89	4
14	6021		Wonderful Guest Bedroom in Manhattan for SINGLES	11528	Claudio	Manhattan	Upper West Side	40.79626	-73.96113	Private room	85	2
15	6090		West Village Nest - Superhost	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Entire home/apt	120	90
16	6848		Only 2 stops to Manhattan studio	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Entire home/apt	140	2
17	7097		Perfect for Your Parents + Garden	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Entire home/apt	215	2
18	7322		Chelsea Perfect	18946	Dofl	Manhattan	Chelsea	40.74192	-73.99501	Private room	140	1
19	7726		Hip Historic Brownstone Apartment with Backyard	20960	Adam And Charity Sing	Brooklyn	Crown Heights	40.67592	-73.94694	Entire home/apt	99	3
20	7750		Huge 2 BR Upper East Central Park	17985		Manhattan	East Harlem	40.79665	-73.94872	Entire home/apt	190	7

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.

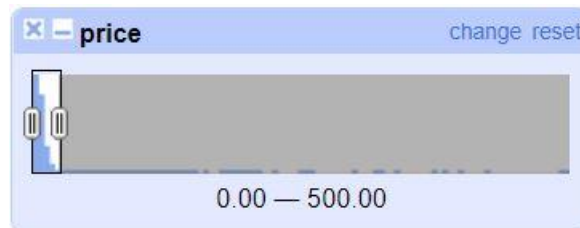


Figure 2. The “price” column numeric facet

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.

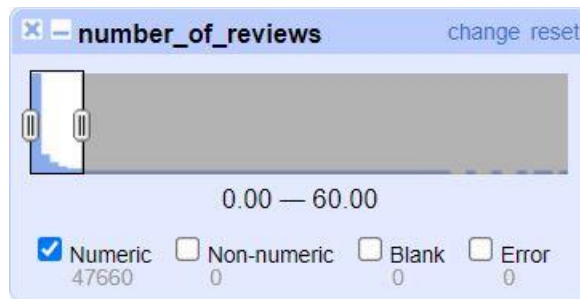


Figure 3. The “number_of_reviews” column numeric facet

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.

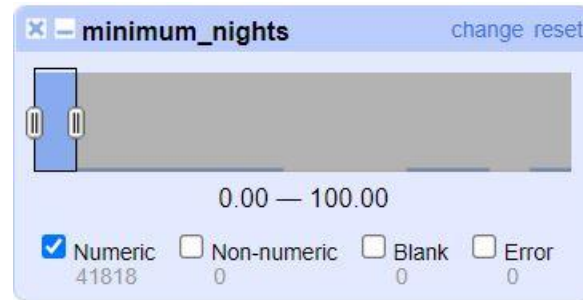


Figure 4. The “minimum_nights” column numeric facet

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.

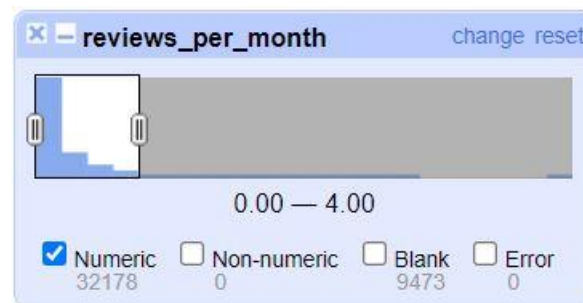


Figure 5. The “reviews_per_month” column numeric facet

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.

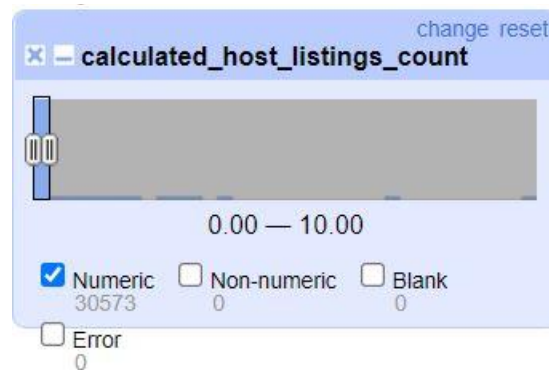


Figure 6. The “calculated_host_listings_count” column numeric facet

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.

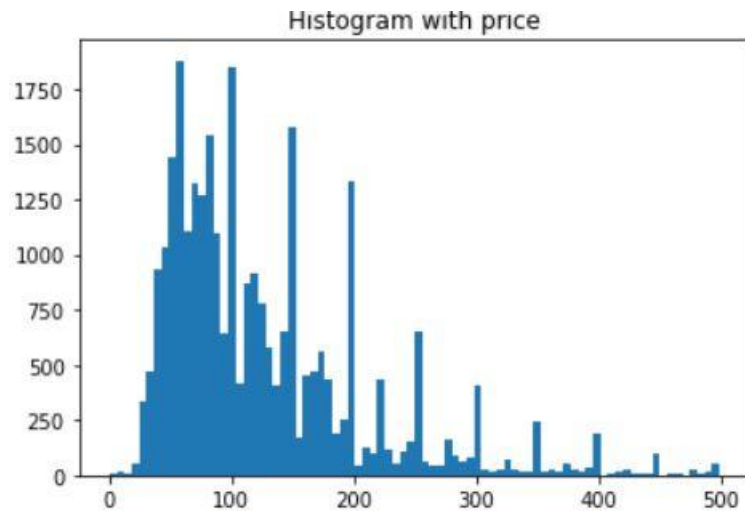


Figure 7. "Price" column histogram

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

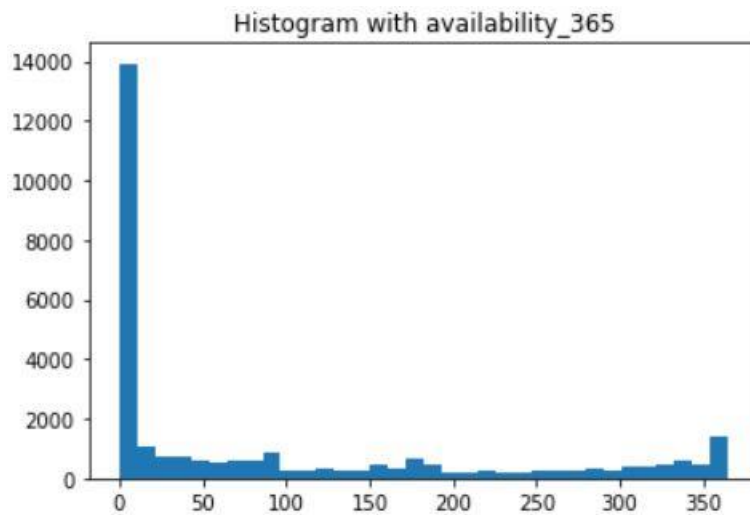


Figure 8. "availability_365" column histogram

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.

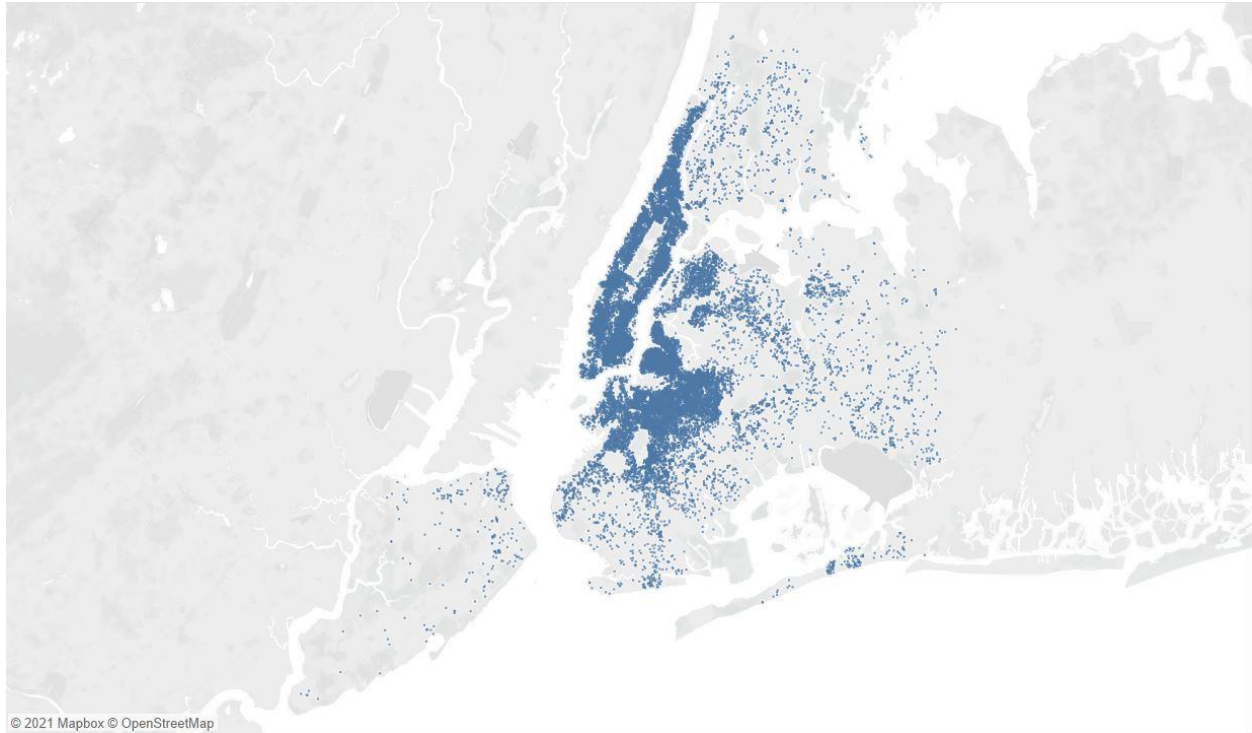


Figure 9. Tableau map visualization with respect to listings price

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>