Exploratory Data Analysis on Airbnb Datasets

By

Maniketh Aley(manik1a)

Sri Valli Guduru(gudur1s)

Vasistinitha Pamireddy(pamir1v)

ITC-510 Software and Data Modeling

05/12/2022

Exploratory Data Analysis on Airbnb Datasets of New York

OVERVIEW:

Airbnb, Inc. is an American corporation that offers an online marketplace for lodging, principally homestays for vacation rentals, and tourism activities. The platform, which has a San Francisco, California, base, is accessed online and through a mobile app. None of the listed properties are owned by Airbnb; rather, it makes money by taking a cut of each booking. The business was established in 2008. The term AirBedandBreakfast.com was originally referred to as Airbnb.

Since 2008, Airbnb has enabled travelers and hosts to experience more distinctive and customized travel. This dataset provides information about homestay listings in New York City as part of the Airbnb Inside program. This New York dataset includes Listings with complete descriptions and the overall rating of the reviews. Reviews with individual reviewer identifiers and thorough comments. Calendar with listing identifiers, price, and availability for each day.

In order to get insights on pricing variations for various room kinds, discrepancies between prices and reviews, the price range of new and old structures, etc., we have conducted a variety of analyses on this dataset.

DATASET:

We obtained the New York Airbnb datasets from the Airbnb website and additional datasets from Kaggle. The first dataset that we obtained from the Airbnb website did not fully contain the data we needed for our analysis, so we obtained a second dataset from Kaggle. The second dataset included helpful columns like construction year, price, availability 365, review rate number, and so on.

The Airbnb Hotel name is the feature shared by the two datasets. In order to obtain the columns, we needed, we merged the two datasets with this feature. So, there are 21465 rows and 18 features/columns in our final dataset.

Below are the columns in the final dataset:

1. id
2. name
3. host_id
4. City
5. host_name
6. minimum_nights
7. number_of_reviews
8. last_review

9. calculated_host_listings_count
10. availability_365
11. Construction year
12. cancellation_policy
13. country
14. instant_bookable
15. price
16. review rate number
17. room type
18. service_fee

EDA:

For a few columns, we conducted exploratory data analysis to gain some knowledge about the N ew York Airbnb's. Additionally, we found details about price variances, rooms that are available 365 days a year, and price differences depending on the year a building was constructed. A correl ation between many attributes was also discovered. Following are some statistics, graphs, and co nclusions:

Descriptive Statistics:

Statistics explain numerical characteristics. It offers a column's count, mean, minimum, and maximum data. The findings are as follows:

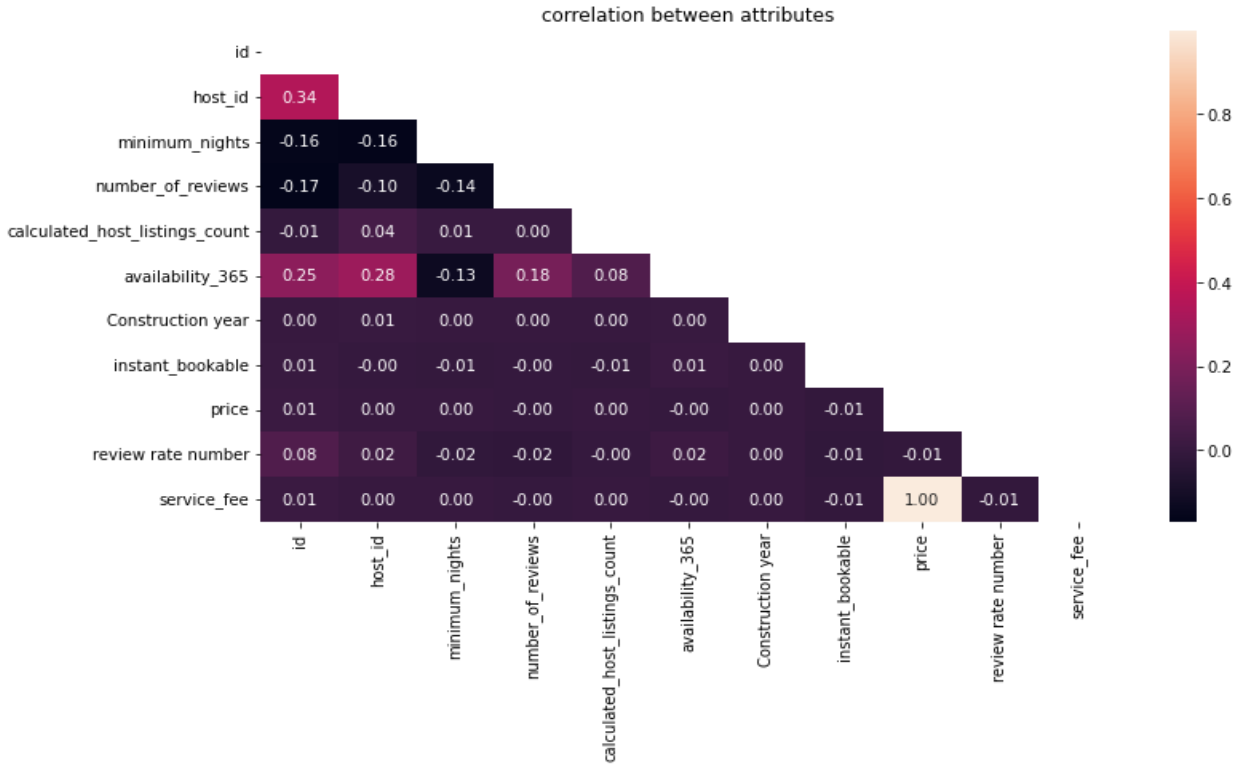
	id	host_id	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365	Construction year	price
count	2.146600e+04	2.146600e+04	21466.000000	21466.000000	21466.000000	21466.000000	21466.000000	21466.000000
mean	7.515327e+16	1.109003e+08	17.368956	36.088465	1.612364	106.255660	2012.490264	621.462219
std	2.038813e+17	1.368000e+08	28.753720	62.456322	5.330946	128.918063	5.763490	333.813198
min	2.539000e+03	2.438000e+03	1.000000	1.000000	1.000000	0.000000	2003.000000	50.000000
25%	1.157362e+07	9.774817e+06	2.000000	3.000000	1.000000	0.000000	2008.000000	331.000000
50%	2.805056e+07	4.283208e+07	7.000000	11.000000	1.000000	35.000000	2012.000000	617.500000
75%	4.850341e+07	1.727280e+08	30.000000	40.000000	1.000000	219.000000	2017.000000	911.000000
max	7.067124e+17	4.772572e+08	1250.000000	1229.000000	453.000000	365.000000	2022.000000	1200.000000

review rate number	service_fee
21466.000000	21466.000000
3.044536	124.294885
1.403689	66.767238
1.000000	10.000000
2.000000	66.000000
3.000000	123.500000
4.000000	182.000000
5.000000	240.000000

Descriptive statistics insights:

The minimum and maximum nights reserved for a hotel were 1 and 1250, respectively. The average number of hotel rooms available each year is 106.25, while average room rates are 621.46 and average service charges appear to be 124$.

Correlation Matrix:



Analysis from Correlation:

It is clear from the graph that the service fee and the price have a strong relationship. Additionally, hotel prices rise as the number of rooms available declines.

Hotels With Most Expensive Rooms:

	name	price
14768	Live, Work, Stay in Prime Midtown~Elevator~Lau...	1200.0
16108	Newly Renovated Brownstone in Clinton Hill	1200.0
13372	Spacious 1 bedroom apt with garden access	1200.0
18529	Lovely rental room in New York	1200.0
7950	Cozy room in the best area! Affordable stay in...	1200.0

The top 5 Highest Price Hotel Rooms are Live, Work, Stay, Newly Renovated Brownstone, Spacious 1 bedroom apt, Lovely rental room, and Cozy room in the best area, and All these hotels seem to have the same price value for highest room.

Least Expensive Room Hotels:

	name	price
113	Modern Brooklyn Apt., August sublet	50.0
515	UES 1Huge BR avail in sweet 2BR apt, I host :)	50.0
19258	Greenpoint By The Park	50.0
12894	Amazing beautiful apartment, cool area, fair p...	50.0
8008	Cozy 1 bedroom in the heart of E. Village	50.0

The Least priced Hotels are Modern Brooklyn Apt, UES 1Huge BR aval, Greenpoint. Amazing, beautiful apartment, Cozy 1 bedroom. The least expensive room price is 50 dollars.

Least Busiest Hotel Rooms

	name	availability_365
9011	Cozy private room, 10 mins away from Central P...	365
18033	Alnez's Cove	365
17952	Spacious rooms in heart of third avenue in bronx.	365
17970	HOTEL IN THE BRONX 7 DAYS HOTEL	365
19290	Centrally located Harlem 1-Bedroom	365

Cozy private rooms, Alnez’s cove, Spacious rooms in heart of the third avenue, Hotel in the Bronx 7 days Hotel, Centrally located Harlem 1 -Bedroom are the least busiest roomed hotels.

Busiest Hotel Rooms

	name	availability_365
21465	Just Blocks to Grove PATH and JC Med Ctr	0
6209	Tribeca Luxury 2000sf Loft	0
11460	UES WITH VIEW OF EAST RIVER	0
6211	BIG GORGEOUS 1 bd 2 br in PRIME of Chelsea	0
6212	Spacious room in Prospect Park	0

Just Blocks to Grove PATH and JC med Ctr, Tribeca Luxury 200sf Loft, UES With View, Big Gorgeous 1bd and 2br, Spacious room are the top 5 busiest hotel rooms.

Recently Constructed Hotels

	name	Construction year
1533	Room Bedford Heart of Williamsburg	2022.0
11736	Mi casa es tu casa! Cozy & comfortable!	2022.0
11768	Perfect one-bedroom in the best location possi...	2022.0
1605	Art-Packed, One-Of-A-Kind Triplex	2022.0
11770	Brand new Luxury apartment in a Luxury building	2022.0

The above 5 hotel rooms are newly constructed in 2022.

Top Rated Airbnb's

:

	name	review rate number
21465	Just Blocks to Grove PATH and JC Med Ctr	5.0
13461	Nice 1 Bedroom in Jamaica Queens	5.0
4320	Williamsburg Two Bedroom/Two Bath	5.0
4319	Cool, cozy urban pad	5.0
13446	Todo esta cercano , trasportación y comida	5.0

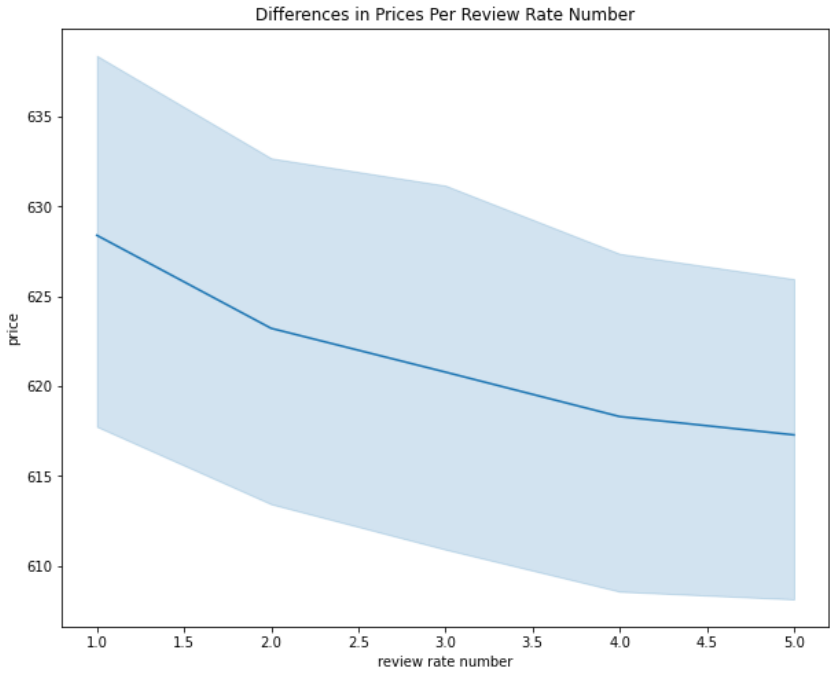
The top-rated Airbnb's are with 5-star ratings and they are Just Blocks to Grove PATH and Nice 1 Bedroom, Williamsburg Two Bedroom, Coo,, cozy urban pad, Todo esta cercano.

Least Rated Airbnb's

	name	review rate number
9689	Cozy Brooklyn Apt (Near JFK/Manhattan/Times Sqr)	1.0
3508	Charming, Central Park Studio Apt	1.0
3507	Large bedroom in an extremely large apartment	1.0
7087	The Cozy Cole room at The Harlem Flophouse	1.0
15901	Private room in a two bedroom apartment room#2	1.0

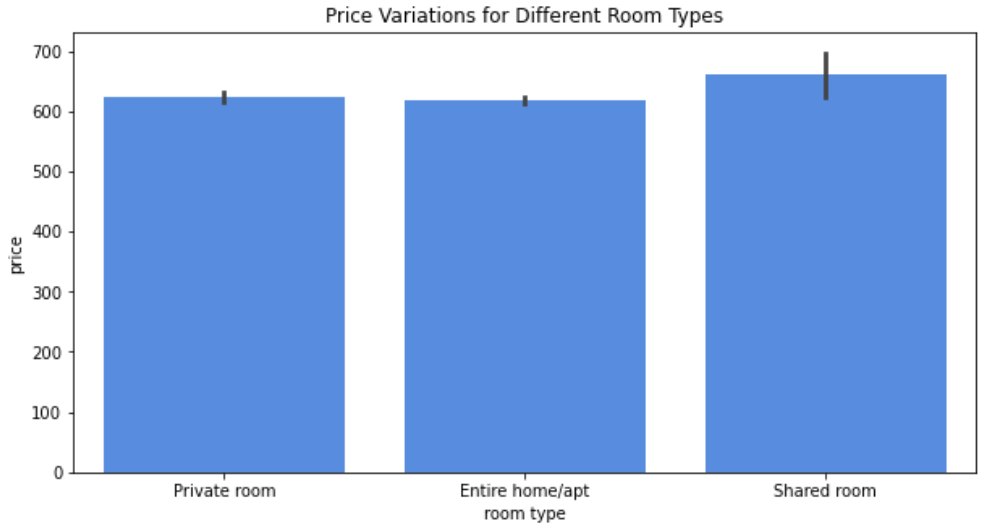
The above are the least Rated Airbnb's.

Analysis of Prices vs Review Rate:



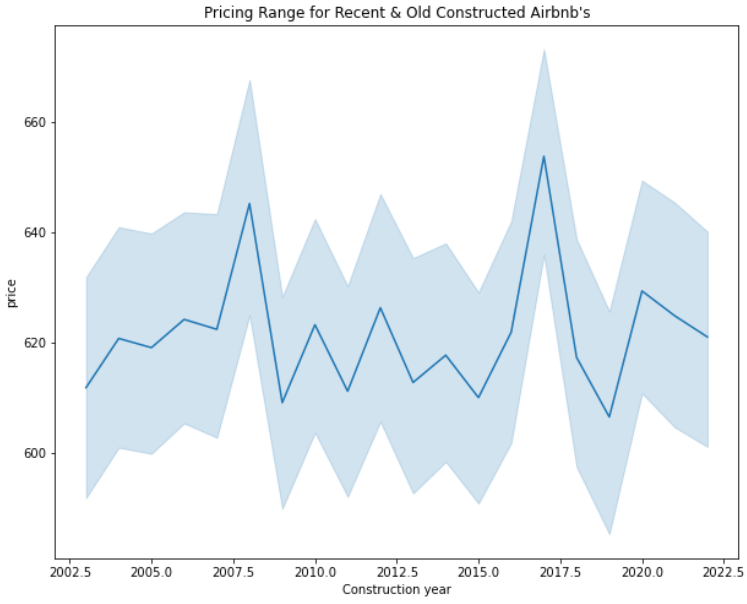
The above graph shows the prices vs review rate. We can observe that the most reviewed hotels have affordable prices.

Analysis of Price Variations Vs Room Types



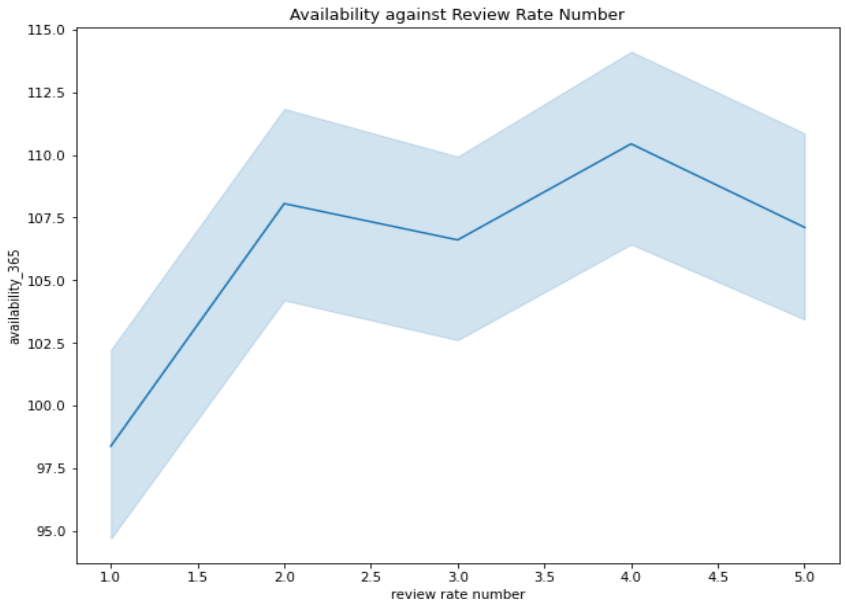
Shared rooms have the highest prices whereas private rooms and Entire homes/apt seem to maintain almost same prices

Analysis of Pricing Recent & Old Constructed Airbnb's



The pricing of hotels seems to be varied based on the current market conditions and there seems to be high prices in the year 2017.

Analysis of Availability Vs Review Rate Number



The hotels with 4 ratings seem to be available for fewer days in a year and 1 rating is highly available. This shows that the customers seem to choose the hotels based on the ratings.

CONCLUSION

Both the data sets from Airbnb and Kaggle have qualities that allowed us to develop insights about hotels in New York. When compared to private rooms, shared room options seem to be more expensive. Additionally, ratings and availability have a strong correlation. There were also multiple hotels with the same high-priced rooms. Depending on the localities within New York the service charge is again varied.

One can use our EDA to choose a hotel in New York City that provides service. We can also get insights into how the service fee varies from most expensive to cheapest. Our EDA can also be extended to compare hotels in different cities.

APPENDIX A

Initial Data Identification:

This New York dataset includes the following Airbnb activity:

Listings with complete descriptions and the overall rating of the reviews. Reviews with individual reviewer identifiers and thorough comments. Calendar with listing identifiers, price, and availability for each day.

We have taken the dataset from the below link of Kaggle and Airbnb website, and we will be using the same for the project.

https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata
http://insideairbnb.com/get-the-data

The above attachment consists of the data for AirBnB analysis.

The dataset contains below attributes and the description for the each one is:

ID:	Unique identifier for each row
NAME:	Listing name.
host_identity_verifed:	Verified by Airbnb
host_name:	The person presenting the property
neighbourhood group:	Boroughs
neighbourhood:	Neighbourhood of the borough
lat:	Latitude
long:	Longitude
instant_bookable:	Whether you can book immediately
cancellation_policy:	Kind of cancellation policy
room_type:	Kind of property
Construction year:	In which year it was built?
Price:	Rental price
service fee:	Airbnb profit
Minimum nights:	Minimum amount of stay
Number of reviews:	How many people have qualified the property?
last review:	Last time that has been qualified.
reviews per month:	Average number of reviews per month.
review rate:	Total average of reviews.
calculated host listings count:	Amount of guests.
availability 365:	number of days the property is available in the year.

We have formed our final data frame out of 2 CSV files.

The details of the CSV files from the New York data are as follows:
Listings.NY.csv
Airbnb_Open_Data.csv

DATA CLEANING AND DATA WRANGLING:

When we saw the combined data frames, several of the characteristics were missing values. This indicates that the data must be processed prior to analysis.

DATA CLEANING:

1) **Removal of some data:** We had to find the features with the most missing data. Such characteristics were ineffective for analysis and should be eliminated from the dataset. As a result, we estimated which characteristics had missing values of more than 50% and eliminated them from the dataset.
2) **Drop Values:** The missing values were denoted by NaN and also found many duplicate entries in the data frame. Hence, we eliminated such entries using dropna() and drop_duplicates() methods. We also encountered unnecessary columns in the data and removed them using drop ().

DATA WRANGLING:

While working on the dataset it seems some of the data was unformatted which we identified and transformed into the required format. For Instance, we had a "last review" column containing improper date format. So, we changed the date to the required American date format.

RegEx operations were performed on the "price" and "service fee" columns to eliminate special characters from the data and changed the columns into required floating point values using the typecasting method. Furthermore, we had to rename a few label names to fit our requirements.

After data cleaning and Data Merging, we were able to retrieve the final dataset on which we performed our EDA.

DATA MERGING OF listings_NY.CSV WITH Airbnb_Open_Data.CSV:

We have merged listings_NY.csv data with Airbnb_Open_Data.csv based on Name to get the following columns/insights of AirbnbNY.csv into the data frame:

- id
- name
- host_id
- City
- host_name
- minimum_nights

- number_of_reviews
- last_review
- calculated_host_listings_count
- availability_365
- Construction year
- cancellation_policy
- country
- instant_bookable
- price
- review rate number
- room type
- service_fee

APPENDIX B

Maniketh Aley

I have contributed to projects working on data cleaning and wrangling where I used functions like filtering out missing values or NaN values, removing unnecessary columns, and dropping out duplicate rows. Additionally, engaged in performing RegEx operations and was also responsible for generating line graphs and performing statistical analysis in the project. I was also part of the initial data collection and in the writing of the report.

- I learned about using filtering data depending on the requirements needed for data cleaning and wrangling.
- I learned how to generate line graphs.
- I learned different ways of sorting and performing operations on datasets.
- Engaged in learning RegEx operations and downloading the filtered datasets

Sri Valli Guduru

I was first involved in the data cleaning process, which comprised converting the unformatted data into the necessary format, eliminating special characters, and typecasting. After a thorough analysis of the datasets, a new column feature called "City" was added. I also grabbed a portion of the code that produces line graphs for EDA analysis and correlation matrices for statistical analysis.

- I learned how to generate line graphs from the data
- I figured out which data is valuable for analysis and also learned to format and insert important data.

- I learned to write reusable and generic functions in python which helps us code Faster.
- I learned to collaborate with the team and was also involved in dividing the tasks among the team members.

Vasistinitha Pamireddy

I have contributed to the data sets from different sources like Kaggle.com and the insideAirbnb.com site in New York City. I was involved in identifying and understanding the dataset to see what metrics we could use for analysis and later merged two datasets into a single file. I also took part in writing a few of the sections of the report. I also took part in the code that helps in generating bar graphs

- I learned how to collect data from different sources.
- I learned how to merge two files into a single file.
- I also learned how to extract common features between 2 CSV files to bring meaningful data.

APPENDIX C

- https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata
- http://insideairbnb.com/get-the-data
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
- https://seaborn.pydata.org/tutorial.html
- https://www.w3schools.com/python/pandas/pandas_cleaning.asp
- https://www.w3schools.com/python/python_regex.asp
- https://www.w3schools.com/python/matplotlib_plotting.asp