Gustavo Santos Follow
Jul 23, 2020 · 5 min read · ✨ · Listen

Insights from Airbnb dataset

. . .

Airbnb is a company from San Francisco, CA in the USA. Their core business is the rental market, where they act on the PaaS (Platform as a Service) model, offering a platform where the users can list their houses, apartments or even a single room for rental.

That is a successful business mo[illegible]en growing in many segments (transportation, streaming servic[illegible]nerating data that can help us to

In this project, I used Python language to explore the "**New York City Airbnb Open Data**" (link for download) and extract as many insights as possible.

Some of the questions I was seeking answer for are:

- What can we learn about different hosts and areas?
- What can we learn about best locations, price average, rent length, reviews, etc
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

. . .

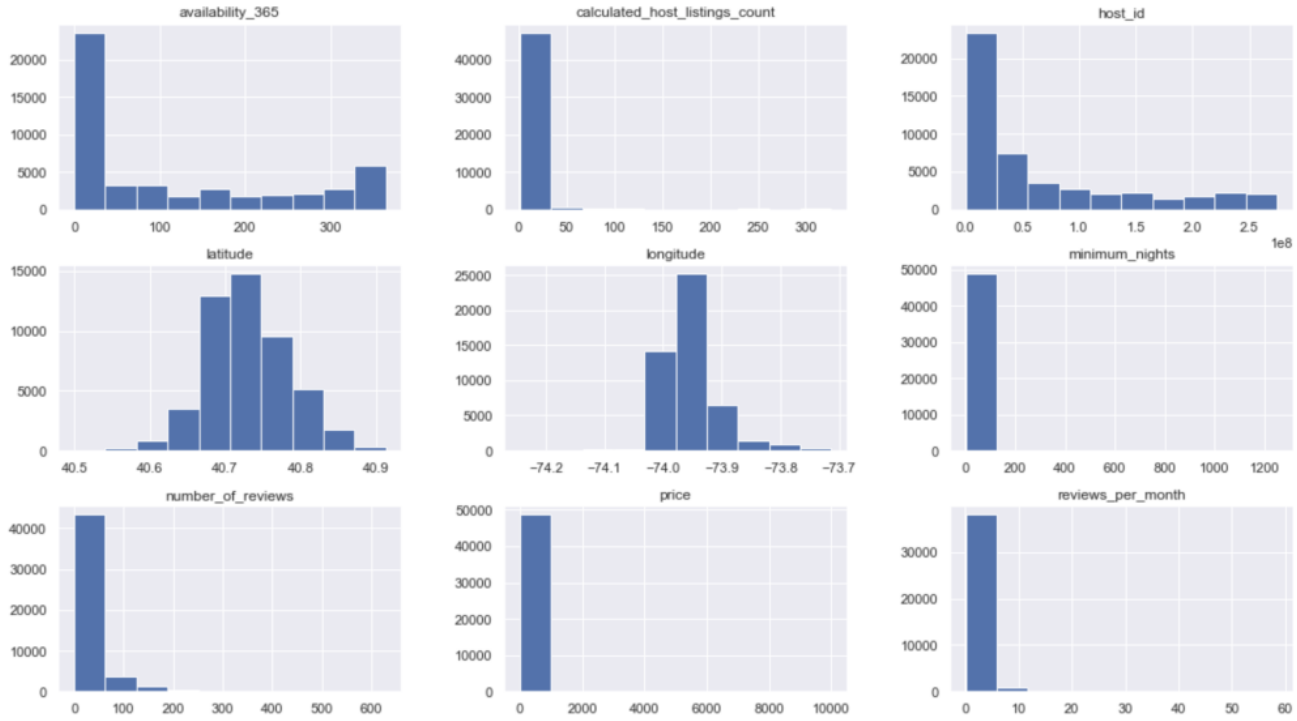
The Dataset

This dataset has 16 columns *('id', 'name', 'host_id', 'host_name', 'neighbourhood_group', 'neighbourhood', 'latitude', 'longitude', 'room_type', 'price', 'minimum_nights', 'number_of_reviews', 'last_review', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365')* and 48.895 observations.

The initial insights extracted after using the *.describe( )* function were:

- Host Name 'Michael is the top one, with 417 appearences;
- Most of the properties are in the Manhattan island (21.6k out of 48.8k = 44%);
- More than 50% of the rentals are for 'Entire home or apt', not just a room;
- On average, the rate is $ 152 dollars and that does not seem to vary too much, since 75% of the obervations are under $ 175 dollars;
- Minumum night average is 7 nights. People apparently like to stay for a whole week in NYC;

In an exploration phase, it is interesting to create histograms, so you can see how the data is distributed.



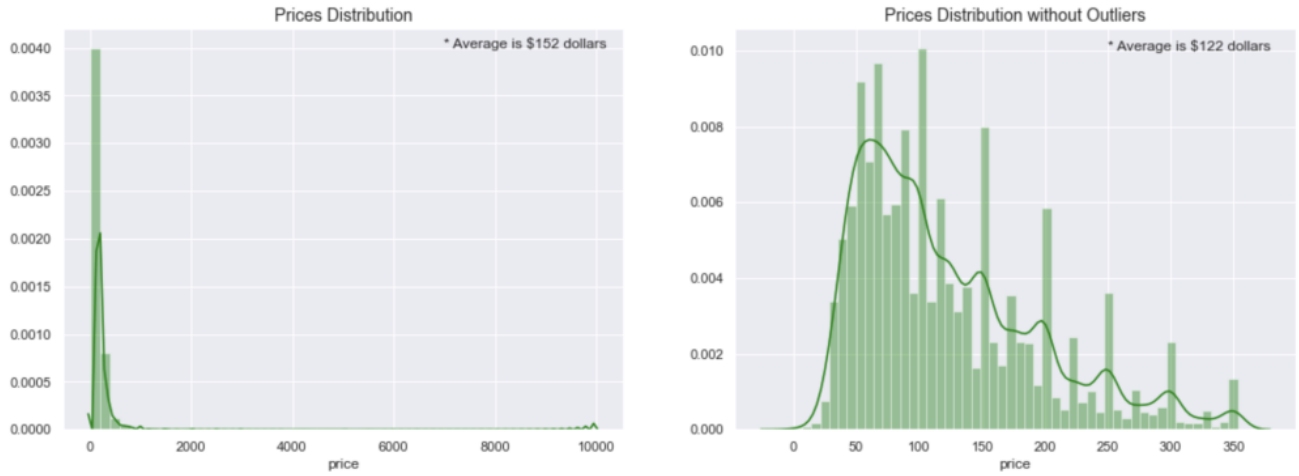
How long, How much and Where

Next step was to figure out how long people stay, how much do they pay and the most wanted locations.

As result, I found out that people seek for a week long stay in NYC, but this can vary from 2 to 7 days. Also, there is a good portion of the rentals on the 30 days average. With further exploration, I could conclude that the top 10 hosts focus their rentals on business over tourists, thus, keeping their properties rented for an average of 30 days.

The average expense is between $120 to $150 dollars per night, what was a good deal compared to hotels on similar locations in 2019. Additionally, from this variable is also possible to see how the business trips influence the prices in New York, as there were some observations with prices on the $200's and

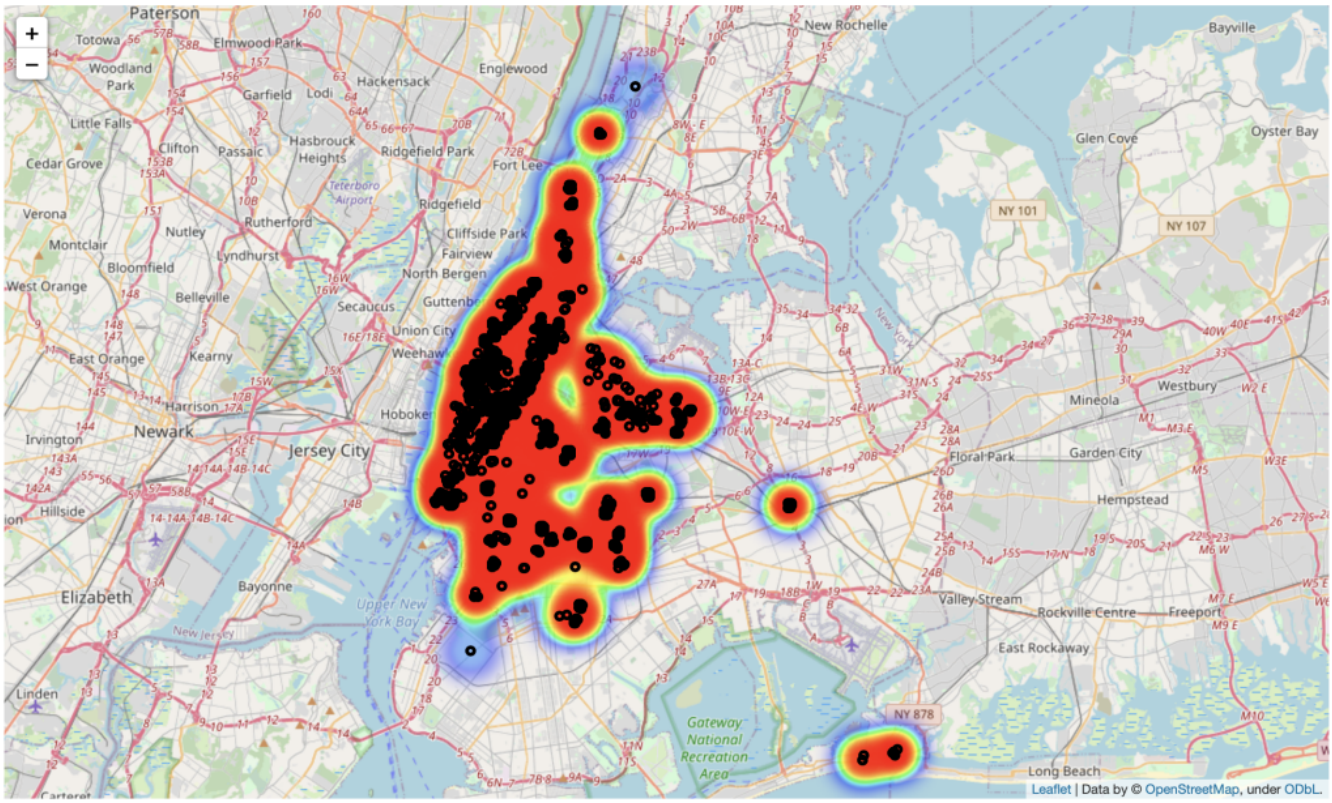
even $300+.



Furthermore, on the location matter, **85%** of the listings are from Manhattan (44%) and Brooklyn (41%, probably given its proximity to the island). And this is not a surprise, actually. Manhattan is the economic heart of the city, concentrating the skyscrapers with its countless offices and world class famous touristic attractions such as the *Times Square, 5th Ave or the Central Park.*

Millions of people land in NYC every year for business or pleasure and they aim to stay as close as possible to Manhattan in order to take advantage of the location, avoiding long subway or car trips and saving time. Naturally, following the offer vs. demand rule, the prices are likely to go up.

Here is a Heat Map created using Folium Map package to demonstrate where are the real estates and the prices (being red the more expensives).



The Top 10 hosts

After ending the visualizations of the entire dataset, I have created this subset with the top 10 hosts, to see what they had different than the others that could make them successful.

It is interesting to notice that the top 10 hosts represent 2,6% of the total rentals, in a sample of 37.500 hosts. That is something!

Location: the first and obvious reason was found right in the first graphic, where I could see that all of them have properties located in Manhattan and/ or Brooklyn. It makes a lot of sense connecting to the location findings.

That group charges over the average prices. Again, this is a consequence of many options they have in the most expensive ground in the NYC. Location is usually a key factor when it regards real estate market.

They probably focus on business rentals over tourism rentals. This conclusion comes from the graphic that shows the top 10 with an average rental length of 30 days against 7 days if we consider the entire dataset. 30 days look much more like a business rental than vacation. Furthermore, after a quick check on the Internet, I saw that Sonder, Vida, Blueground are all companies from the rental market in NYC, so they have a staff and a supporting structure for their clients (features commonly required by other businesses).

Natural Language Processing (NLP)

To conclude the project, I performed a NLP using the package *nlkt* to extract some insights from the reviews made by the users. I have created a word cloud and word frequency count to see the most frequent words. *Bedroom, cozy, Brooklyn, Manhattan, private, Studio* were among the top used words.

Then I moved to the analysis of Bigrams and Trigrams — *two / three words used together and in a sequence within a text* — as it brings us more understanding of how the words are used in a context. The result showed that the users are renting private rooms, 1 or 2 bedroom apartments. They like to stay on the Upper East side, near the 5th Ave, Central Park to take advantage of that well located area. Reviews mentioning subjective qualities, such as *cozy* or *home away from home* concepts appeared among the top frequent, demonstrating the clients are worried about cleanliness and maintenance of the place.

. . .

Wrap up

Data Science is more than only predictions. It is about extracting good information and knowledge from the data.

In this project I was able to perform many tasks commonly used for data exploration and visualization, extracting insights from the Airbnb dataset to help companies to make better and data driven decisions.

To check the Jupyter Notebook, please go to my GitHub repo: https://github.com/gurezende/NYC-Airbnb

```
if data:
    data.science()
```

Gus

Sign up for Data Science Gustavo Santos

By gustavorsantos

Data Science posts with tutorials and code snippets. Take a look.

Your email

Get this newsletter

By signing up, you will create a Medium account if you don't already have one. Review our Privacy Policy for more information about our privacy practices.