Springboard--Data Science Career Track
Airbnb price in New York City
Capstone Project 1 Milestone Report
By Kin Sun
April 2020

Problem

With so many Airbnb available in NYC, some people may wonder if they are charging a fair price for their rental listings. The 48,000-samples dataset I am planning to use--which I will obtain from Kaggle[1]-- contains some features that can be used to predict the price per night of a listing in NYC for the year of 2019. Prospective clients are any individuals, property management companies, existing Airbnb hosts, etc—in other words, anyone who just wants to use their properties for rental purposes on Airbnb. These clients would be able to use the results of this project to find out how much they can reasonably charge per night of stay.

Data Description

Feature	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	price in dollars, per night
minimum_nights	amount of nights minimum

[1] https://www.kaggle.com/dgomonov/new-york-city-airbnb-

number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Data Wrangling

1 - Dropping features

The dataset has some features that serve no purpose for the analysis. For example, id and host_id are identifiers of the listing and the host.. Another two features that will be removed are name and host_name. Lastly the last_review feature, which is the date of the last review left to the listing, will also be removed.

2 - MIssing Variables

By utilizing Panda's ".isnull().sum()," the following result was generated:

```
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

There are 10,052 null values for the feature "reviews_per_month."

Below, rows with null values were selected:

	number_of_reviews	reviews_per_month
2	0	NaN
19	0	NaN
26	0	NaN
36	0	NaN
38	0	NaN
...	...	...
48890	0	NaN
48891	0	NaN
48892	0	NaN
48893	0	NaN
48894	0	NaN

10052 rows × 2 columns

By summing the "number_of_reviews" reveals that 0 "number_of_reviews" is the reason for "reviews_per_month" to be a null value. One option to deal with these null values is to remove those rows. However, the size is over 20% of the dataset. Luckily, there is a better and wiser way to deal with the null values here, and that is to simply replace them with 0's. This option is viable since 0 number of reviews should have an average of 0 reviews per month.

Another issue was that the dataset contains 0 for price, showed below:

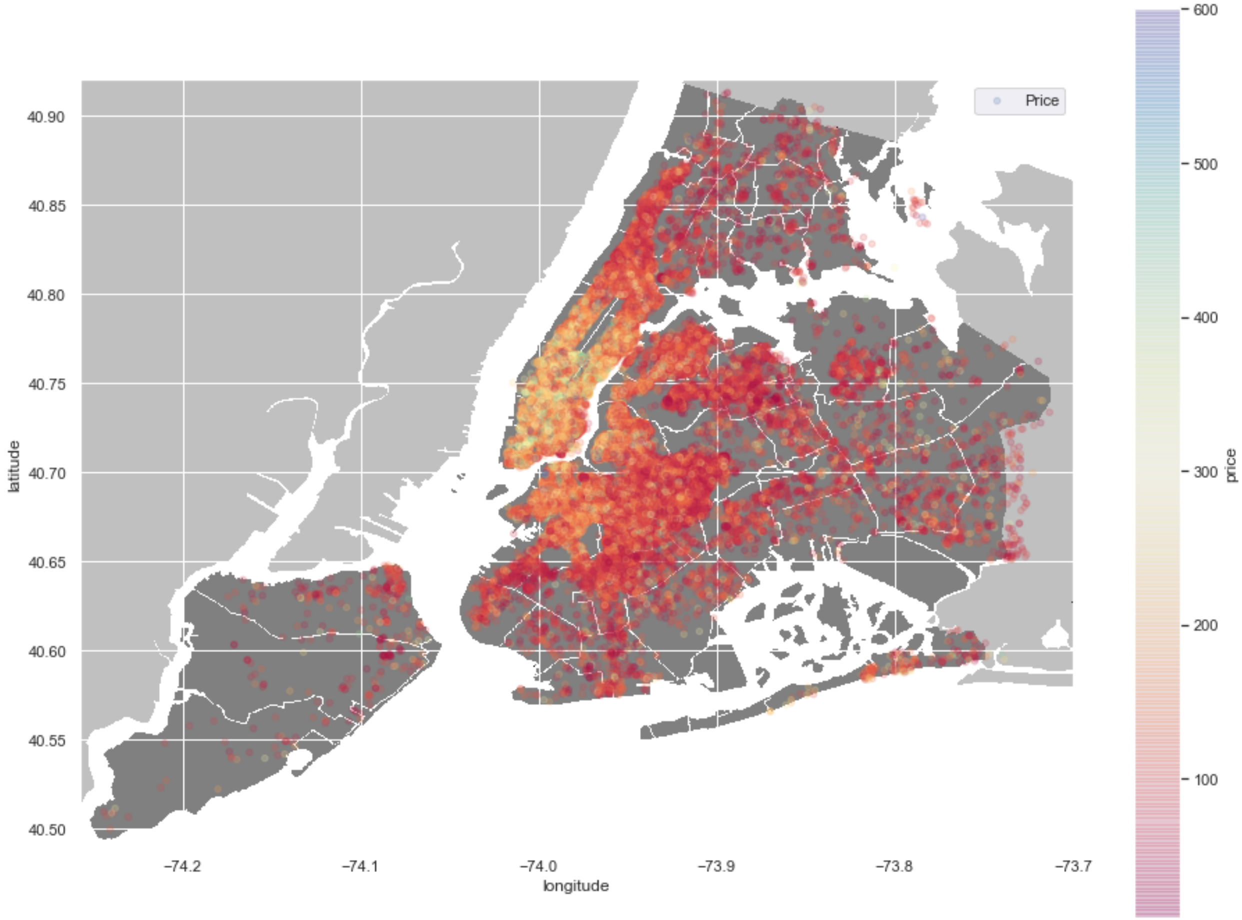
```
raw_data.describe(include='all')
```

	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
count	48895	48895	48895.000000	48895.000000	48895	48895.000000
unique	5	221	NaN	NaN	3	NaN
top	Manhattan	Williamsburg	NaN	NaN	Entire home/apt	NaN
freq	21661	3920	NaN	NaN	25409	NaN
mean	NaN	NaN	40.728949	-73.952170	NaN	152.720687
std	NaN	NaN	0.054530	0.046157	NaN	240.154170
min	NaN	NaN	40.499790	-74.244420	NaN	0.000000
25%	NaN	NaN	40.690100	-73.983070	NaN	69.000000
50%	NaN	NaN	40.723070	-73.955680	NaN	106.000000
75%	NaN	NaN	40.763115	-73.936275	NaN	175.000000
max	NaN	NaN	40.913060	-73.712990	NaN	10000.000000

This is an issue because the problem we are interested in solving is how much people can charge on their listings. There are only 11 of those instances and they will be removed from our final data.

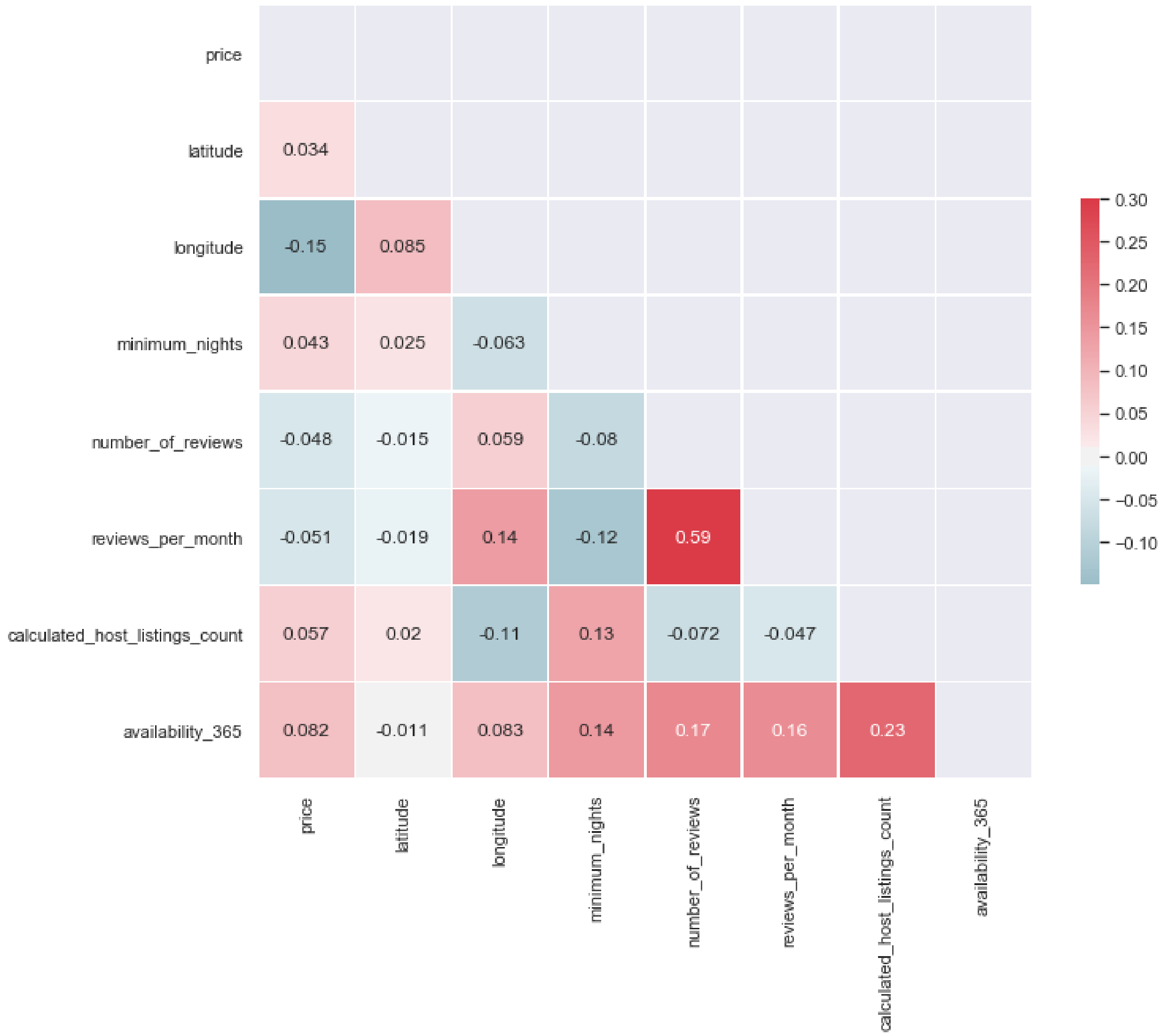
<u>Exploratory Data Analysis</u>

Below heatmap shedded lights on what would otherwise not have been observed just by looking at numbers:



Dataset was slightly modified to exclude any price higher than $600.00, which affected roughly 778 rows of data. The reason to do that is to prevent distortion on the heatmap caused by abnormally high price listings. What the heat map shows is that the closer to downtown Manhattan, the more expensive the Airbnb listing prices are. Note that we also observe the higher price in Brooklyn as well, as it gets closer to Manhattan.

Below is a heatmap for correlation coefficient of the numerical features:



We see that the correlations are not strong between any of those variables, with exception of "reviews_per_month" and "number_of_reviews," but it is still not very strong. This is a good thing and it will be interesting when doing feature selection later on.

Next Steps

The next tasks to continue working on the project are:

- Data Preprocessing
- Model Selection