# Coursera Data Science Capstone Project

## Airbnbs in Staten Island

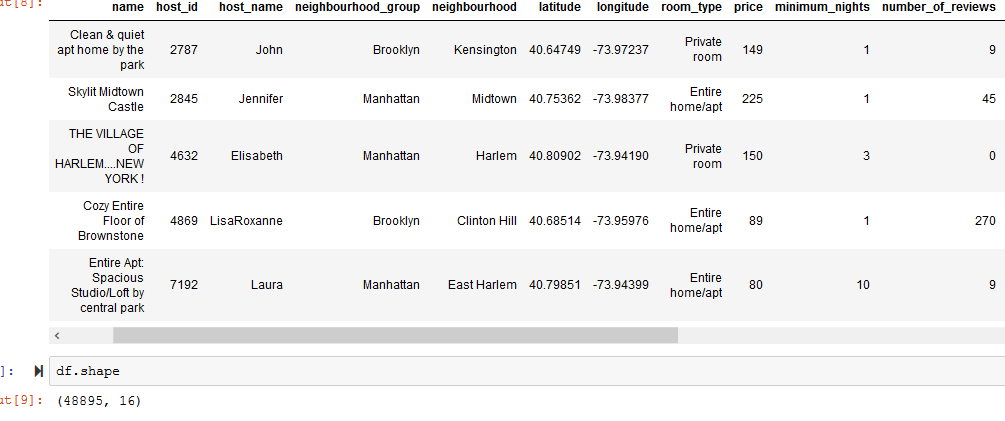
**Introduction/Business Problem**:

As a topic for my capstone project I choose to analyse the Airbnb market in New York. I will try to find the connection between prices and some features as minimum nights, venues nearby the flat and the number of reviews the flat takes. It will help the investors to decide weather buy a flat with lower airbnb prices or higher prices. They can consider the number of venues in the district and so on. I think 90$ is the border in thiss case, to consider a good or bad price. Investors should invest in flats that they can list on good prices later.

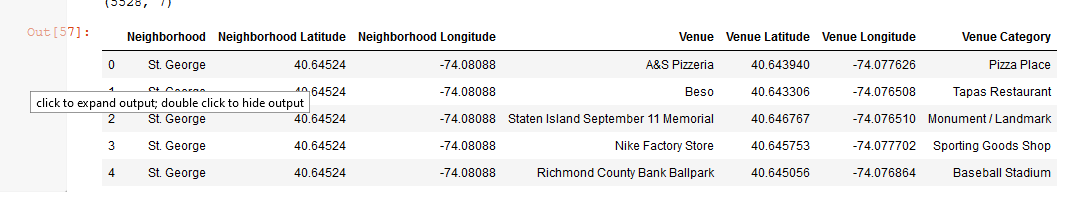
**Data**:

The data that I will be using is coming from some different sources. Firstly from Kaggle I got a .csv file that contains information about the Airbnb data. I will use the Foursquare API to obtain data about the venues, etc.

Here is a bit of the Airbnb data:



From Foursquare API:



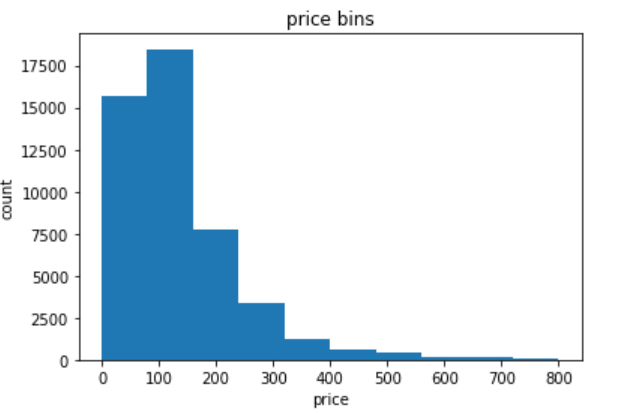
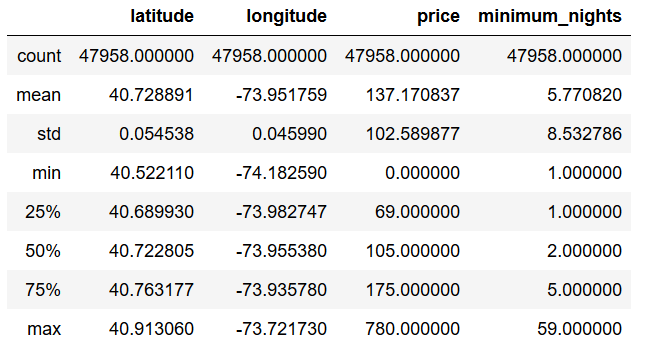
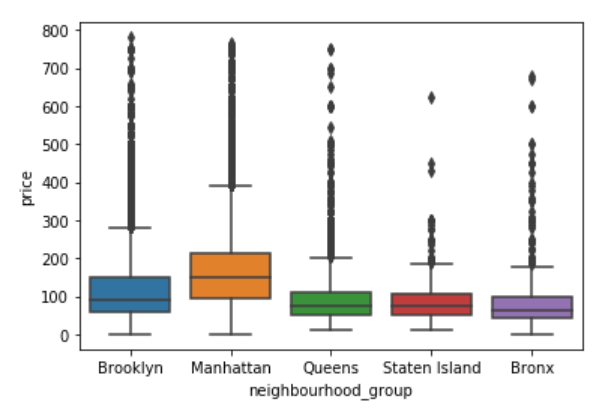
For the final analysis my dataframe is the following:



**Methodology Section:**

My reports main components are data acquisition, data cleaning, some plots to visualize the connection between the variables, merging the dataframes to create the final table, and after them the final analysis.

So in the data acquisition component with pandas library I read the .csv file from my computer. As for the Foursquare data I used a function to make the API calls, and retrieve the data from the server.

Data cleaning is in my opinion the most important part of any data analysis. „Garbage in, garbage out” as many people says. I included the variables I need from all dataframes, I detected the outliers and missing data in the dataframes and removed them. Converted datatypes such as from int to float type. Created some dummy variables where it needed. Here is a histogram to show the distribution of price variable. The most of the prices fall between 0 and 200. The next table shows the statistical numbers of the dataframe. It seems that the mean of price is 137 $. Relatively large std with 102. I visualized the data with the help of boxplots. I really like boxplots. They clearly shows the outliers, means and interquartile ranges.

Now I filtered my dataframe to cases in the neighborhood of Staten Island because the limits of api calls and memories. On Staten Island there are 356 Airbnbs. These are involved in my analysis. I imported the Folium library to make a visualization on a map:



Next I grouped the prices by neighborhoods. There comes the data of average prices of Airbnbs of neighborhoods in Staten Island.

The Foursquare API got me 5528 rows of data and I grouped them by neighborhoods to count them.

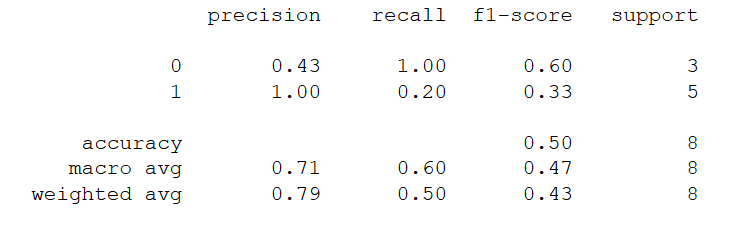
Here I made a little regression analysis. This showed me no significant connection between venue count and prices: **y hat = 104,7 - 0,03 \* Venues**

Next I included in the analysis the reviwes per month and the minimum nights columns.

Multiple linear regression gave me the coefficients: -0,03 for venue count, ,12,0 for reviwes per month and -1,77 for minimum nights.

I choose Logistic Regression Analysis as the algorithm I will use, because I wanted to get the probability of prices bigger than 90$ based ont he features I choosed. For that I created a dummy variable: good\_price. I considered a good price bigger than 90$ per night. I created a training and test set and did the model including the variables mentioned before.

**Results:**



This is the evaluation of my log.reg. model. The accuracy score is not bad at all, but with the f1 score of 0,5 there is no predictive power in the features I used.

**Conclusion:**

So as a conclusion I can say that the number of venues nearby the Airbnbs doesn't matter in the term of prices. The minimum nights neither do matter and as for the reviews it seems that investors have to buy flats, which they can list on lower prices. On lower prices more and more costumers will come. Investors will earn bigger profit.