Coursera Capstone Project

Title: Discovering Consumer Preferences in Different Income Group in New York

Name: Tang Ka Kit (Jay)

Contact: tangjay58@gmail.com

Project Link on GitHub:

https://github.com/jaytang0508/Coursera_Capstone/blob/master/Cousera_Capstone%20(3).ipynb

1. Introduction

This project divides the neighborhoods in New York City (NYC) into three income groups, and compares the consumer behaviors among the groups by displaying the citizens' favourite types of venues to visit in each group. This helps new entrepreneurs getting to know the market of their potential business better by finding hints about the questions: 'What kind of business should we establish?' and 'Which income group should we target?'. In addition, at the end of the project, there is also a simple program designed for advising the entrepreneurs at which exact neighborhoods they should set up the business.

2. Data Acquisition

Data in this project are divided into three parts:

a. Median Income of the neighborhoods in NYC – this dataset is mainly used to separate the neighborhoods into three different income groups. Data which are converted to an excel file and extracted only the median income column afterwards are retrieved from Renthop: https://www.renthop.com/study/assets/new-york-city-cost-of-living-2017/nyc-2br-median-rent-and-income-table.html
b. Longitude & Latitude data of the neighborhoods in NYC – this dataset is used to extract the common venues of each neighborhoods from FourSquare, as well as to display the location of neighborhoods on a map. They are retrieved from New York City, Department of City Planning: https://geo.nyu.edu/catalog/nyu_2451_34572
c. Common Venue data of the neighborhoods in NYC – this dataset is used for comparing the preferences of people in different neighborhoods. They are

retrieved from FourSquare API. The retrieving process is shown in the working process in my Jupyter Notebook on GitHub.

3. Methodology

The project is divided into four steps:

a. Import and Clean data

b. Explore the Neighborhood

c. Analyze the relationship between income group and common venues

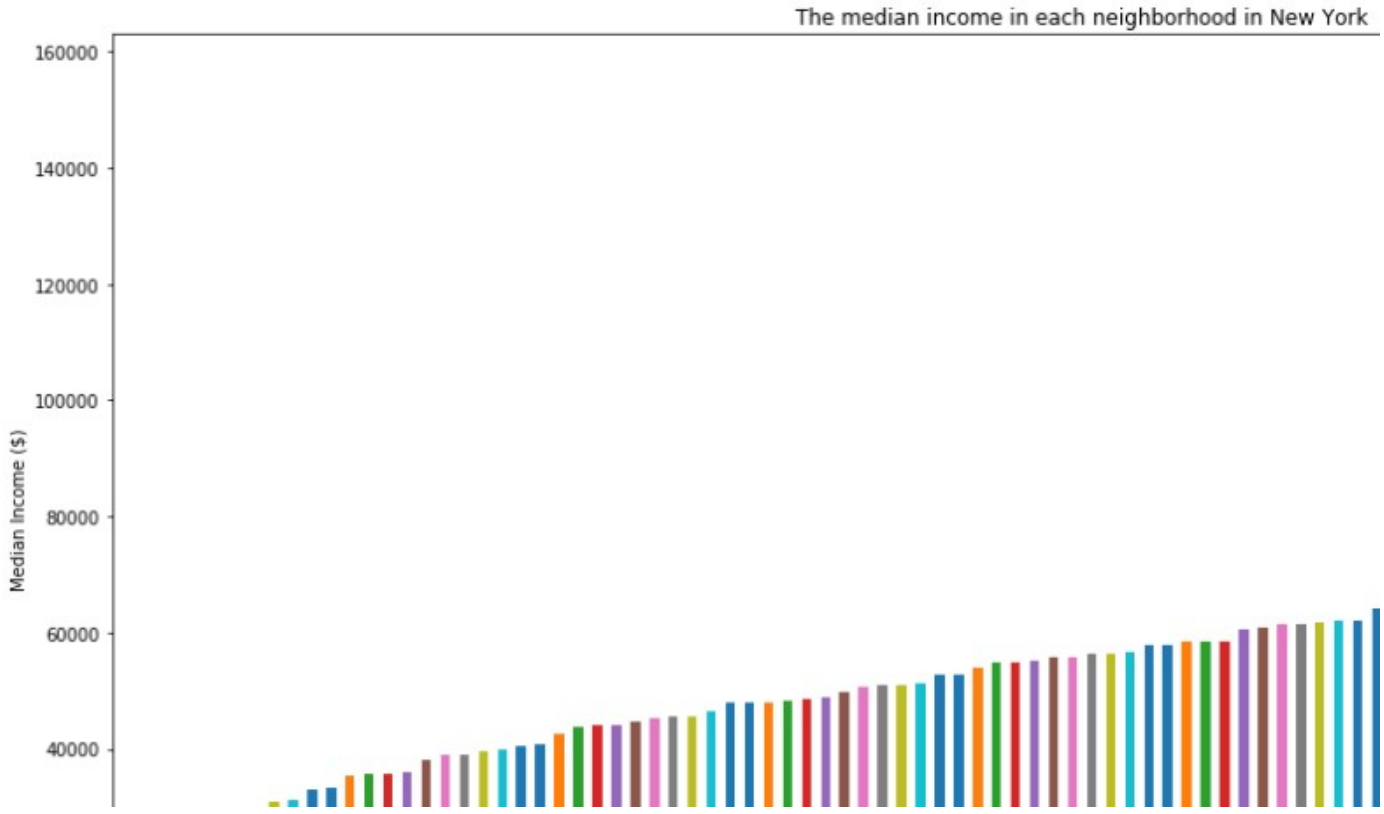
d. Create a tool to display where the venues are famous to visit

The first two parts are mainly importing all the data needed, cleaning and organising the data into a single dataframe for easier observation. The dataframe is shown here:

	Neighborhood	Borough	Median Income in dollar	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue
0	Long Island City	Queens	28,378	40.750217	-73.939202	Hotel	Coffee Shop
1	Williamsburg	Brooklyn	21,502	40.707144	-73.958115	Bagel Shop	Coffee Shop
2	Lower East Side	Manhattan	31,273	40.717807	-73.980890	Coffee Shop	Café

The whole dataframe after merging all data together has a scale of 102 rows (neighborhoods) and 10 columns (only top 5 most common venues are extracted), although first five rows are displayed only for convenient.

In part three, the analysis part, neighborhoods with different median income are shown on a graph to check whether simple classification can be done.



However, as shown on the graph, it is tough to classify them into three groups without bias,and thus, K-means-clustering on the median income column is applied in order to separate them into groups with less biasedness. The three groups (High Income Group with 12 neighborhoods, Moderate Income Group with 45 neighborhoods, Low Income Group with 45 neighborhoods) then contains various neighborhoods with their median income falling into the range discovered by the clustering method. The first and last elements in each group are then shown below:

	Income Group	Neighborhood	Borough	Median Income in dollar	Latitude	Longitude	1st Most Common Venue	2nd M Comm Ve
0	Low	Port Morris	Bronx	20334	40.801664	-73.913221	Storage Facility	Latin Amer Restau
45	Moderate	Queensboro	Queens	53836	40.744572	-73.825809	Chinese	Bus Sta

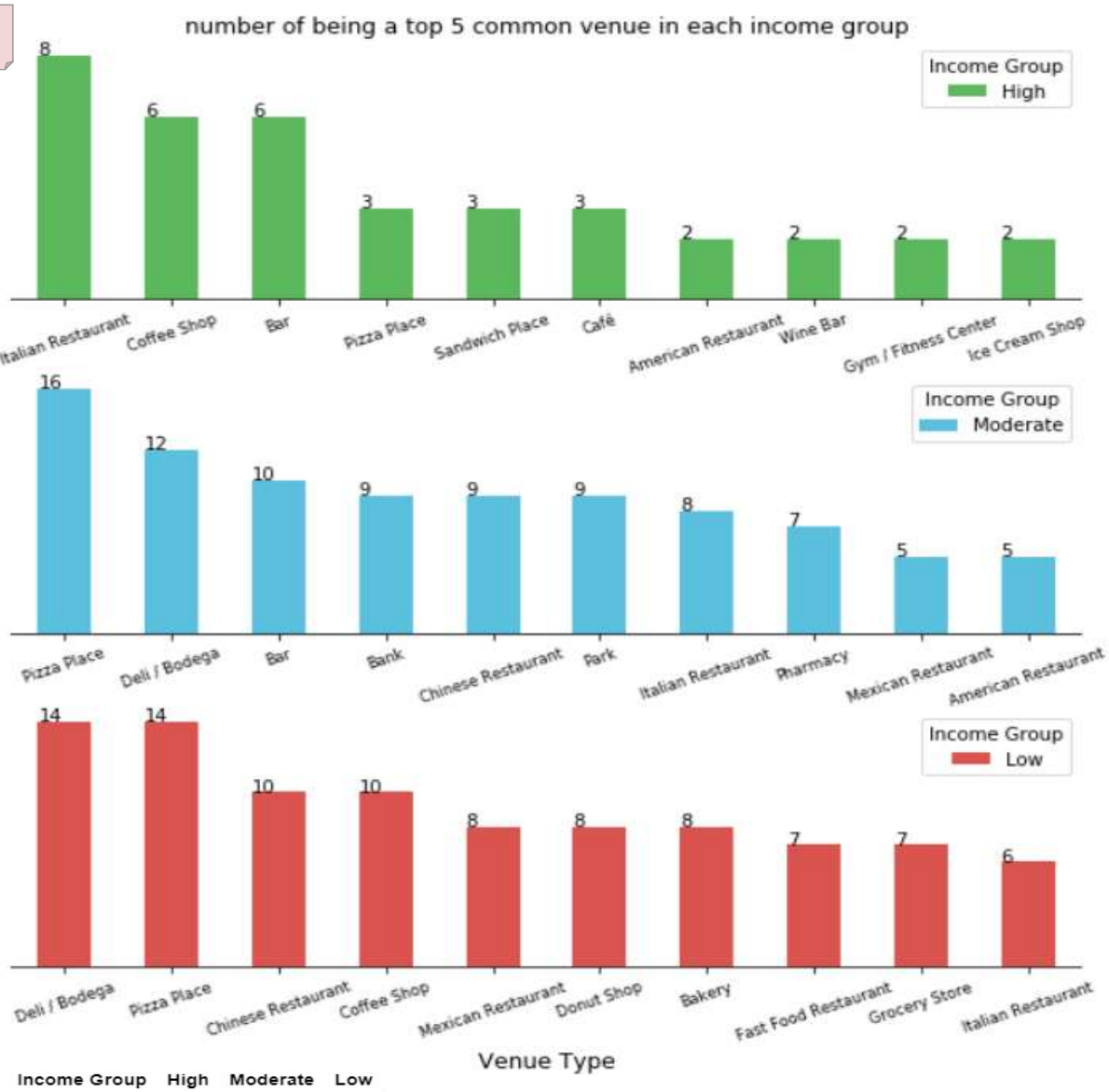
	Income Group	Neighborhood	Borough	Median Income in dollar	Latitude	Longitude	1st Most Common Venue	2nd Co
44	Low	Murray Hill	Queens	52696	40.748303	-73.978332	Korean Restaurant	Coffe
89	Moderate	Red Hook	Brooklyn	85496	40.676253	-74.012759	Seafood	Art

Since the group separation has been settled down, counting of being the top 5 common venues can be done in each of the group such that the most common venues in each group can be found based on the counting, and the assumption that the top 5 common venues in each neighborhood do not vary a lot in terms of their popularities. Here is then the counting of each venue in each income group (showing top 10 only):

Income Grou	Income Group	Income Gro
Venu	Venue	Ven
Italian Restaura	Pizza Place	Deli / Bode
Coffee Sho	Deli / Bodega	Pizza Pla
Ba	Bar	Chinese Restaura
Pizza Plac	Bank	Coffee Sh
Sandwich Plac	Chinese Restaurant	Mexican Restaura
Ca	Park	Donut Sh

A bar chart is also generated for better illustration about the in-group comparison:

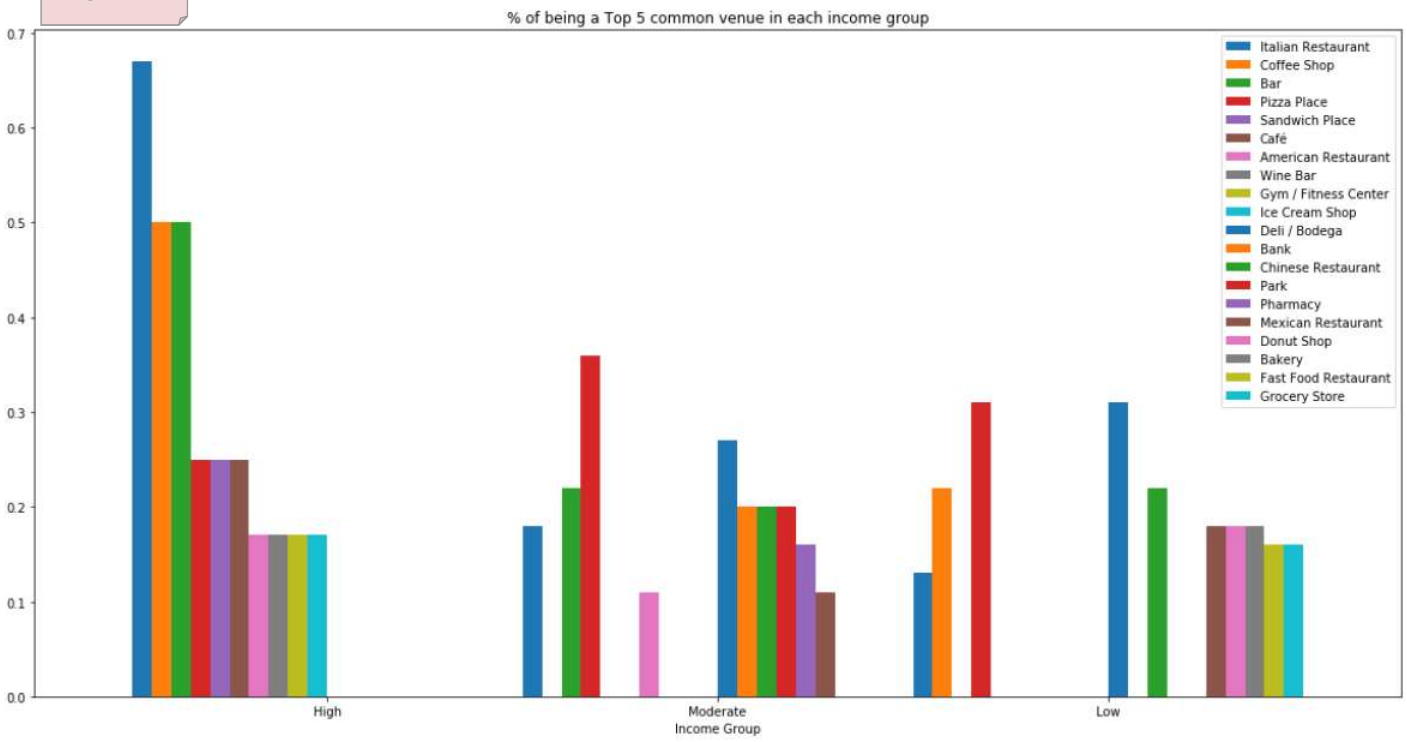
Figure 1



Income Group	High	Moderate	Low
Italian Restaurant	0.67	0.18	0.13
Coffee Shop	0.50	NaN	0.22
Bar	0.50	0.22	NaN
Pizza Place	0.25	0.36	0.31
Sandwich Place	0.25	NaN	NaN
Café	0.25	NaN	NaN
American Restaurant	0.17	0.11	NaN
Wine Bar	0.17	NaN	NaN
Gym / Fitness Center	0.17	NaN	NaN
Ice Cream Shop	0.17	NaN	NaN
Deli / Bodega	NaN	0.27	0.31
Bank	NaN	0.20	NaN
Chinese Restaurant	NaN	0.20	0.22
Park	NaN	0.20	NaN
Pharmacy	NaN	0.16	NaN
Mexican Restaurant	NaN	0.11	0.18
Donut Shop	NaN	NaN	0.18
Bakery	NaN	NaN	0.18
Fast Food Restaurant	NaN	NaN	0.16
Grocery Store	NaN	NaN	0.16

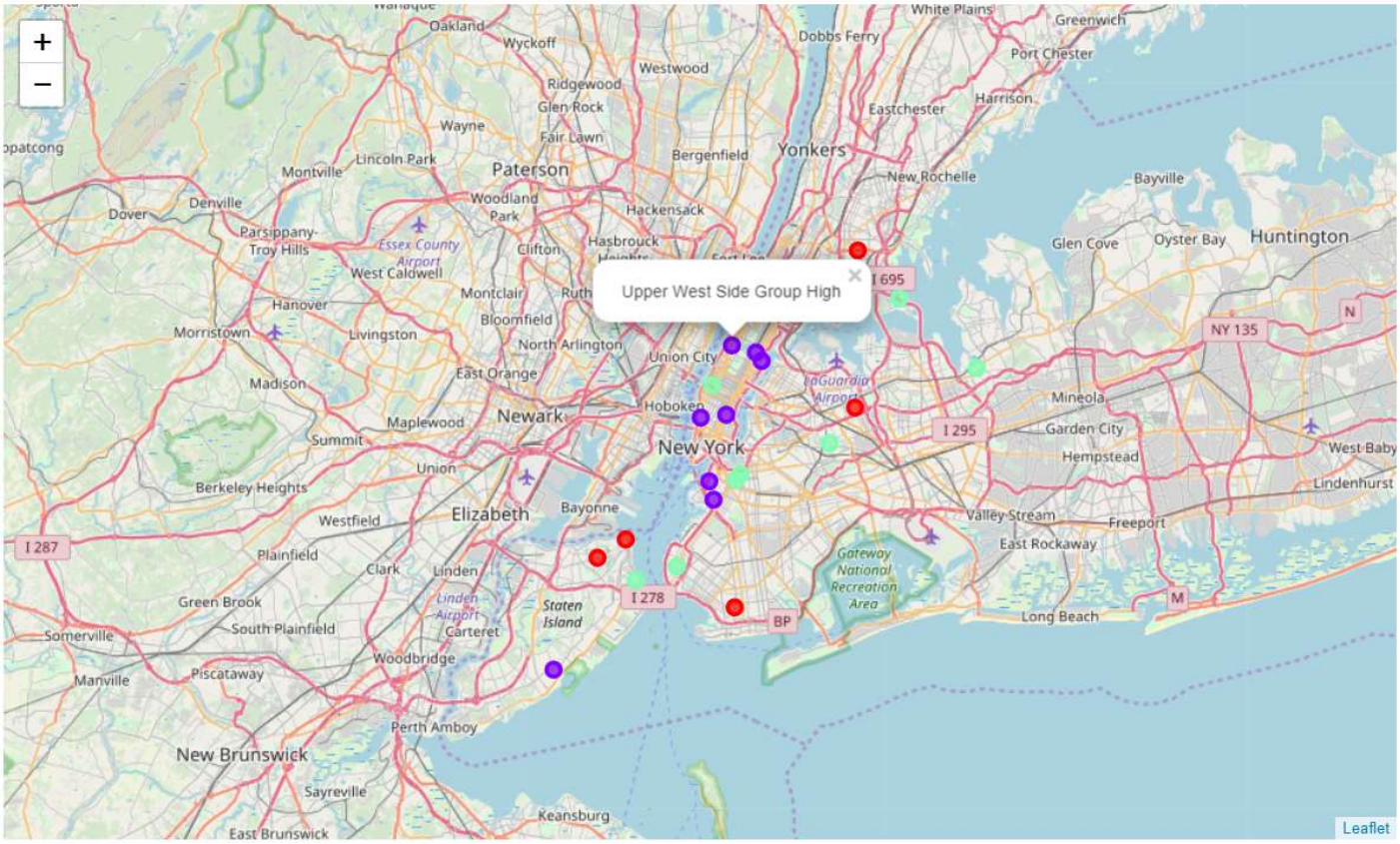
However, for inter-group comparison, as mentioned, the number of neighborhoods in each cluster varies a lot (12:45:45), which causes a non-precise comparison. Hence, the dataframe of each income group is divided by their number of neighborhoods in each group. Percentage of being a top 5 common venue in each group is then used, instead of number, to show contrast among the groups. Furthermore, for clarity, the dataframes of each group are merged into one (see figure on the left hand side), whilst the separated bar charts unify to only one (see figure on the next page).

Figure 2



At the last part of the entire project, a programme named *search_fame('venue name')* is written **to show the neighborhood and its income group of the venue being a top 5 common venue**. Italian restaurant is used as an example in the following graph.

```
search_fame('Italian Restaurant')
```



4. Result

In the result discussion session, intra-group comparison and inter-group comparison will be dicussed to find out information about the consumer preference in different area in NYC.

a. Intra-group comparison

This answers the question of ‘what kind of business should one establish in each of the income group’ if one is financially strong and wants to expand the business portfolio. At the first glance of figure 1, the outstanding venue in each group would be {High: Italian Restaurant, Moderate: Pizza Place, Low: Deli/ Bodega and Pizza Place}. Hence, these type of venues should be set up in each of the income group, whereas determining which neighborhoods inside the group can be checked through the ‘search_fame()’ attribute. Besides, base on the assumption that preference of each neighborhood among the same group holds, or rather say, do not vary a lot, setting up those venues in other neighborhoods in the same group are worthwhile to give a try.

b. Inter-group comparison

This answers the question of ‘which income group one should target’ if one has determined doing a certain kind of business, say, a young entrepreneur with idea but financially weaker. By looking at figure 2, if a color bar appears on the chart once, it certainly means that doing business in that income group is having more advantages than working it out in other groups, for instance, building a bank in the moderate income group would outperform doing it in the low or high income group. However, if a color bar appears twice or above on the chart, the group with the highest percentage has to be chosen, for example, there are three bars representing Italian restaurants, but clearly serving Italian food in the high income group tends to make more profit. Nonetheless, the graph also shows the best group for expanding the business, whereas the program advices the exact place for the expansion.

5. Recommendation for refinement

a. Enlarge the data set for the high income group (12 neighborhoods only, compared to the other two with 45 neighborhoods) by enlarging the scope to the entire United States.
b. Append the data set for the turnover of all venues to see whether being a common venue is positively correlated to a more profit-making venue.

c. Apply logistic regression for the part of inter-group comparison to get the actual probability of being a common venue by inputting the actual median income of a neighborhood.

6. Conclusion

This project focuses on answering the questions about the market situation in NYC which mainly benefits entrepreneurs. By grouping the neighborhoods into three income groups, they could earn insights about the consuming properties or where consumers would like to spend their money about each income group. The project also helps them to target the places and groups they should start or expand their businesses. However, base on the challenge on extra data acqusition, the project can still be potentially refined in order to reach a preciser result.