IBM Data Science Professional Certificate

Capstone Project

Finding the Best Area to Start a Restaurant Business in San Francisco

Tianfang Chen

Oct. 2019

I. Introduction

Restaurant is one of the most common small business entrepreneurs start. It is also one of the business with least technical barrier of entry. At the same time, the restaurant business has very low profit margins, with the average number between three to five percent. Also, opening a new restaurant requires quite an amount of initial capital investment if renting or buying a venue. Therefore, it is very important to start off on the right foot by choosing the best area to open the restaurant.

This project offers an infographic view of the demographics and restaurant competitions in each neighborhood of San Francisco, alongside with some examples of where different types of restaurant should best be started. Anyone who wants to start a new restaurant in San Francisco can use the report or the interactive tool on the Jupyter Notebook as a guide to find the optimal place to start a restaurant based on the two elements.

II. Data Description

Data used in the project are demographics and restaurant competitions, and each comes from a different source.

Demographics data come from the San Francisco Planning Department (https://default.sfplanning.org/publications_reports/SF_NGBD_SocioEconomic_Profiles/2012-2016_ACS_Profile_Neighborhoods_Final.pdf). Demographics data are found for each neighborhood, and combined to form the overall dataset. The dataset includes features like total population, race percentages, and median household income.

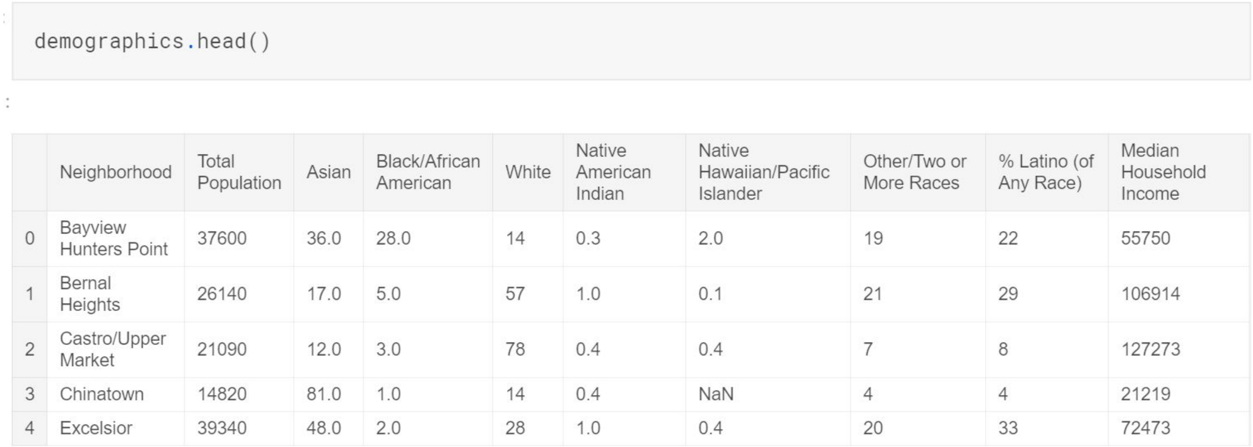
Restaurant competition data come from Foursquare. Restaurant venue data are collected via Foursquare location API, and then taken the total amount of each type of restaurants within each neighborhood to form the overall restaurant competition dataset.

III. Methodology

Data Acquisition and Cleaning

There are two main datasets used in the project. One is the demographics; another is restaurant competitions.

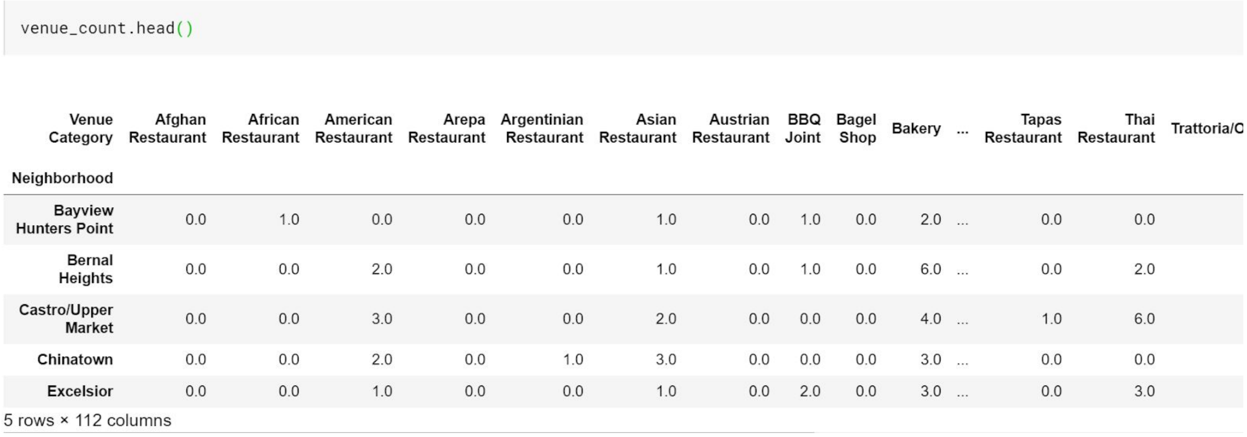
The demographics information is readily available in PDF report from San Francisco Planning Department at https://default.sfplanning.org/publications_reports/SF_NGBD_SocioEconomic_Profiles/2012-2016_ACS_Profile_Neighborhoods_Final.pdf. Relevant data are picked out and put into CSV file. Dataset is cleaned by changing data type for race and filling n/a with 0.



```
demographics.head()
```

	Neighborhood	Total Population	Asian	Black/African American	White	Native American Indian	Native Hawaiian/Pacific Islander	Other/Two or More Races	% Latino (of Any Race)	Median Household Income
0	Bayview Hunters Point	37600	36.0	28.0	14	0.3	2.0	19	22	55750
1	Bernal Heights	26140	17.0	5.0	57	1.0	0.1	21	29	106914
2	Castro/Upper Market	21090	12.0	3.0	78	0.4	0.4	7	8	127273
3	Chinatown	14820	81.0	1.0	14	0.4	NaN	4	4	21219
4	Excelsior	39340	48.0	2.0	28	1.0	0.4	20	33	72473

The restaurant competition data are pulled from Foursquare API in the Jupyter notebook. First, the latitude and longitude of each neighborhood is acquired with geocoder API and put into a Pandas dataframe. The location information is then used to find all restaurants within one mile radius of the location. The list of restaurants is then grouped by the type of restaurant to make up the final restaurant competition dataset.



```
venue_count.head()
```

Venue Category	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Austrian Restaurant	BBQ Joint	Bagel Shop	Bakery	...	Tapas Restaurant	Thai Restaurant	Trattoria/C
Neighborhood														
Bayview Hunters Point	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	...	0.0	0.0	
Bernal Heights	0.0	0.0	2.0	0.0	0.0	1.0	0.0	1.0	0.0	6.0	...	0.0	2.0	
Castro/Upper Market	0.0	0.0	3.0	0.0	0.0	2.0	0.0	0.0	0.0	4.0	...	1.0	6.0	
Chinatown	0.0	0.0	2.0	0.0	1.0	3.0	0.0	0.0	0.0	3.0	...	0.0	0.0	
Excelsior	0.0	0.0	1.0	0.0	0.0	1.0	0.0	2.0	0.0	3.0	...	0.0	3.0	

5 rows × 112 columns

Exploratory Data Analysis

As the project is focused on creating a visualization for demographics and restaurant competition information for each neighborhood, there is less need to draw insights from the datasets by its own. But it is still quite interesting to examine them especially the demographics dataset.

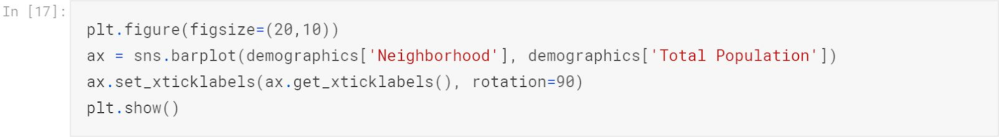
Couple things of interest are the distribution of population, distribution of median household income, and whether there's a correlation between median household income and percentage of each race. For population and median household income, bar charts and box plots are produced. And for the correlation, a heatmap is produced.

Mapping

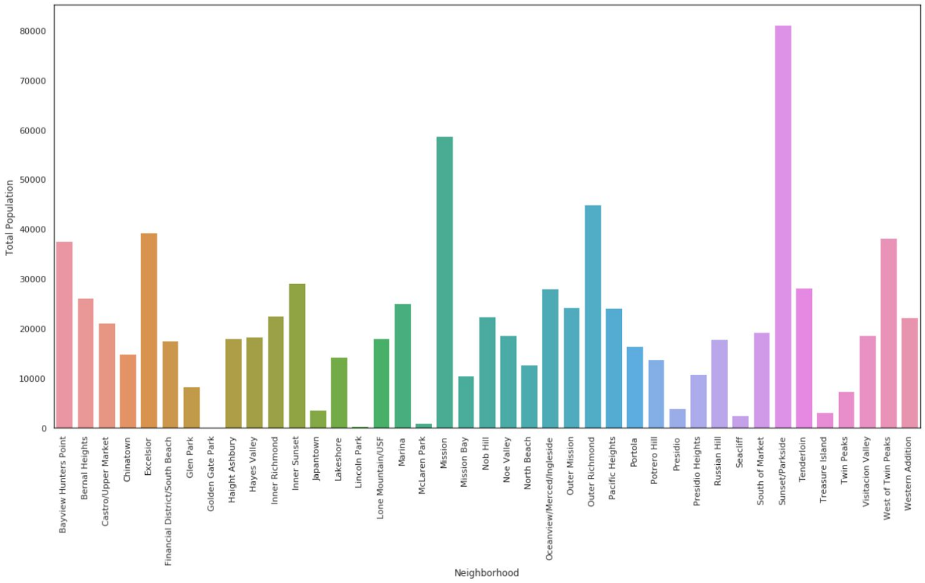
A visualization of demographics and restaurant competition information is produced with the Folium library. When clicking on each neighborhood, one can see the relevant information in a pop-up.

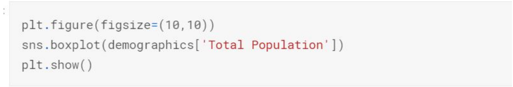
IV. Results

Population

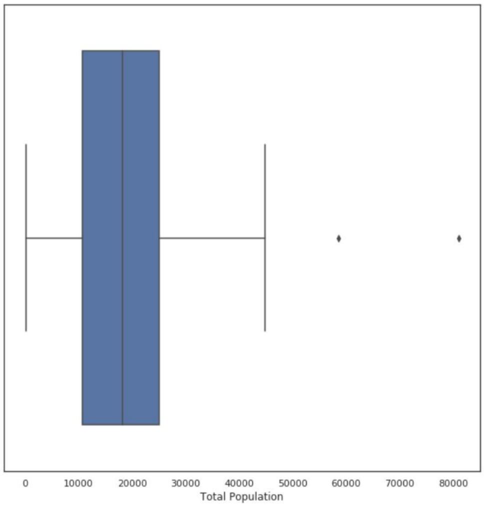


```
In [17]:
plt.figure(figsize=(20,10))
ax = sns.barplot(demographics['Neighborhood'], demographics['Total Population'])
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
plt.show()
```



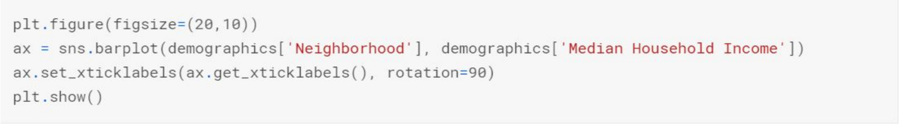


```
plt.figure(figsize=(10,10))
sns.boxplot(demographics['Total Population'])
plt.show()
```

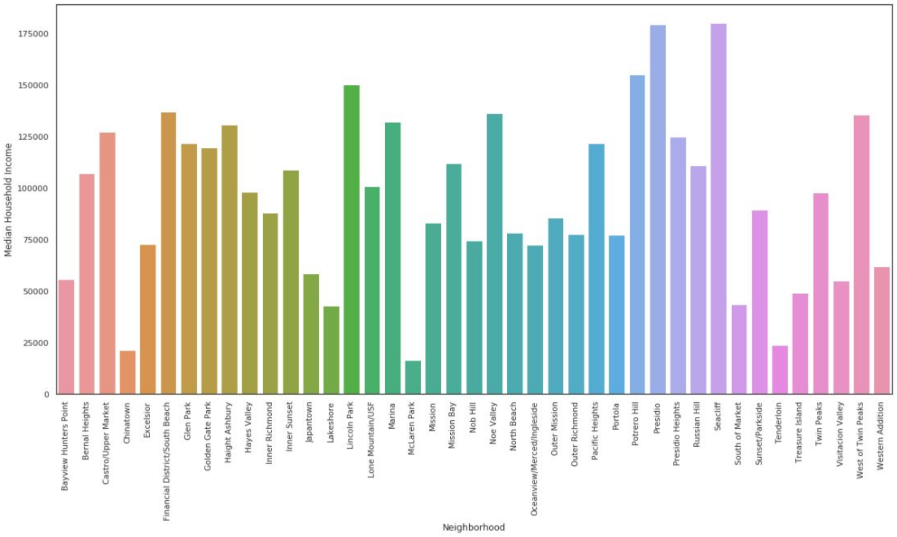


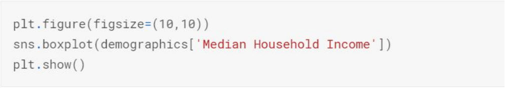
From the above graphs, we can see that there's a widespread distribution of population ranging from a low of lower than 100 to a high of higher than 80000. There are also two outliers on the higher side.

Median Household Income

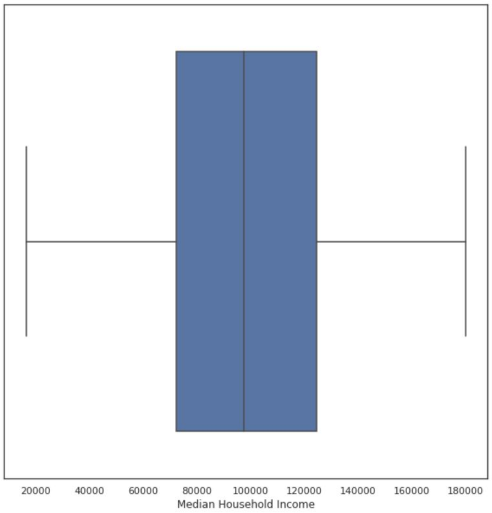


```
plt.figure(figsize=(20,10))
ax = sns.barplot(demographics['Neighborhood'], demographics['Median Household Income'])
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
plt.show()
```



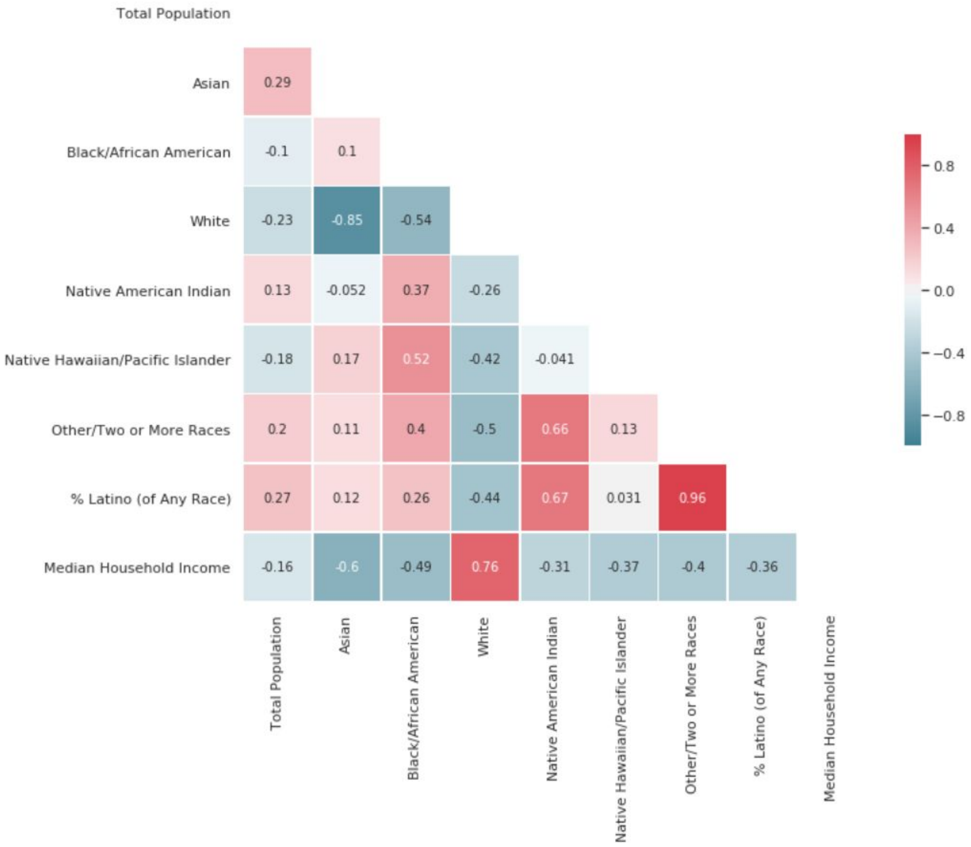


```
plt.figure(figsize=(10,10))
sns.boxplot(demographics['Median Household Income'])
plt.show()
```



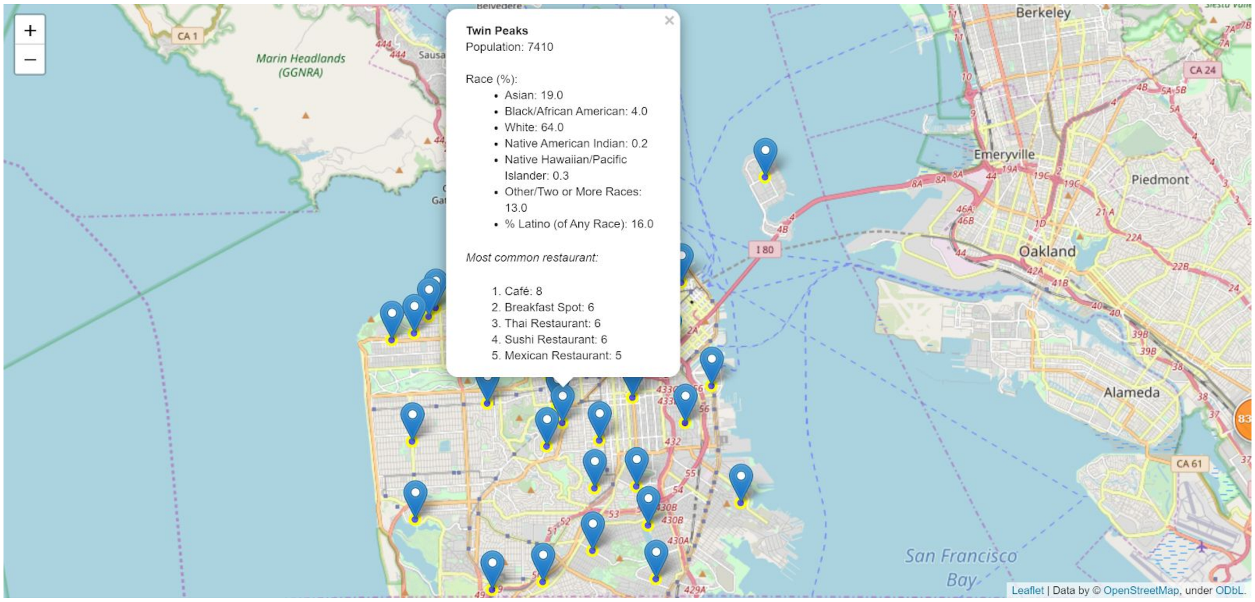
From the above graphs, we can see that there's a widespread distribution of median household income ranging from a low of around 20000 to a high of around 180000. There are no outliers.

Correlation Heatmap



The most interesting part of the heatmap is the last line Median Household Income's correlation with the race percentages. There is a strong positive correlation (r=0.76) between the percentage of White residence and Median Household Income, and moderate or weak negative correlation between the percentage of other races and Median Household Income.

Mapping



Twin Peaks
Population: 7410

Race (%):

- Asian: 19.0
- Black/African American: 4.0
- White: 64.0
- Native American Indian: 0.2
- Native Hawaiian/Pacific Islander: 0.3
- Other/Two or More Races: 13.0
- % Latino (of Any Race): 16.0

Most common restaurant:

1. Café: 8
2. Breakfast Spot: 6
3. Thai Restaurant: 6
4. Sushi Restaurant: 6
5. Mexican Restaurant: 5

Above is an example of how clicking on a neighborhood would look like on the visualization. Population, race percentage, and the most common types of restaurants can be found on the pop-up.

V. Discussion

Use Case

A prospective restaurant owner can use the interactive visualization map to see relevant information--population, race percentage, and count of current restaurants--to better choose a location for his/her future restaurant.

For example, let's say that Diego is looking to start an authentic Mexican restaurant. Then he can find the neighborhood with high population, high percentage of Latino residents, and lower number of Hispanic food restaurants. And he can easily do that with the interactive map without having to look for different data himself.

Limitation

Some limitations of the current codes include:

- The use of one mile radius for finding restaurants in the neighborhood. The latitude and longitude are not at the center of each neighborhood, and the neighborhoods' boundaries are not one mile circles centered at the latitudes and longitudes.
- Foursquare API has a limit of 100 for each request. The maximum amount of restaurants retrieved for each neighborhood is 100 while there might be more than that in the area.

Future Improvement

The above limitations should be improved upon in the future. Circumventing the two limitations can be done by picking enough amount of distanced latitude and longitude pairs on the map, and finding restaurants near each location pairs. The method would generate a list of all the restaurants in San Francisco. And then the locations of each restaurant would be compared with the boundary of each neighborhood to create accurate lists of restaurants in each neighborhood.

To make the tool easier in the future, a search algorithm can be developed--when a prospective restaurant owner searches for a type of restaurant, the tool will give the recommendation automatically.

VI. Conclusion

In this project, we have acquired and cleaned data from online sources, performed exploratory data analysis, and created the interactive map in Jupyter notebook. Prospective restaurant owners can visualize relevant information of each neighborhood with the interactive map, and therefore choose the best place to start the restaurant easier and better.