Captsone 2 Milestone Report 1

Problem Statement

Location is one of the most important factors that determines whether a restaurant will succeed or fail. Restaurateurs must carefully select locations to open new restaurants in order to give their business the best chance to succeed, but how do they go about making that decision? Several factors should be considered including: How many people live in the area? What kind of income do they have? How many and what kind of other restaurants are already in the area? The plan for this project is to provide a tool that will simplify this decision making process by displaying the relevant information on an interactive map, including a single metric that will rank zip codes based on the likely opportunity for a successful new restaurant. With this tool, restaurateurs could select the type of cuisine for which they would like to open a restaurant and immediately see a map showing where the best opportunity is. Additionally they would see a list of the top 10 zip codes with more information about each zip code including the population, median income, and number of restaurants of their chosen cuisine type.

Gathering the data

The data for this project will come from three sources. First, the U.S. Census provides population, income, and age data at the zip code level. Second, the city of New York provides publicly available health inspection data that includes both the zip code and cuisine type for all restaurants in New York City that have had a health inspection. Finally, jsspina.carto.com provides a geojson file containing the coordinates needed to map the boundaries of every zip code in New York City.

The population, income, and age data were acquired from factfinder.census.gov. On this website you are able to search for and extract data in CSV format. I start by selecting the geography type of 5-digit zip code tabulation area, and then select all of New York. Then I select

topics > people > basic count/estimate > population total. I choose the 2017 5-year estimate as it is the most recently available data. I then modify the table by transposing rows and columns so that each row is a zip code and I extract the data in CSV format. I then repeat these steps to get the income and age data so that I have three separate CSV files.

For the NY health inspection data, I visited data.cityofnewyork.us and chose to export the data in CSV format. Finally the zip code coordinates were found at jsspina.carto.com and were available for download in geojson format. I saved all of these files to the data folder in my project directory.

Cleaning the data

I used Jupyter Notebook to load each of the data files and do data cleaning. For the population, income, and age data I simply dropped all unnecessary columns and then merged the three dataframes on the zip code key. For the age data I reduced the number of bins (columns) from 18 to 4. Those columns corresponded to the proportion of the population in each zip code that were between the ages of 0-14, 15-29, 30-54, and 55 and above.

In the health inspection data each row was a single inspection, which meant that I had duplicate restaurants because most restaurants are inspected more than once. I used Pandas' `drop_duplicates()` method to remove rows that had the same name, building, street, zip code, phone, etc. This reduced the number of rows from 401,000 to about 27,000. I then used a `groupby` statement to calculate the proportion of each cuisine type in each zip code. I now had a data frame where each row is a zip code and each column is a cuisine type, with the value being the proportion. This data contained 227 unique zip codes.

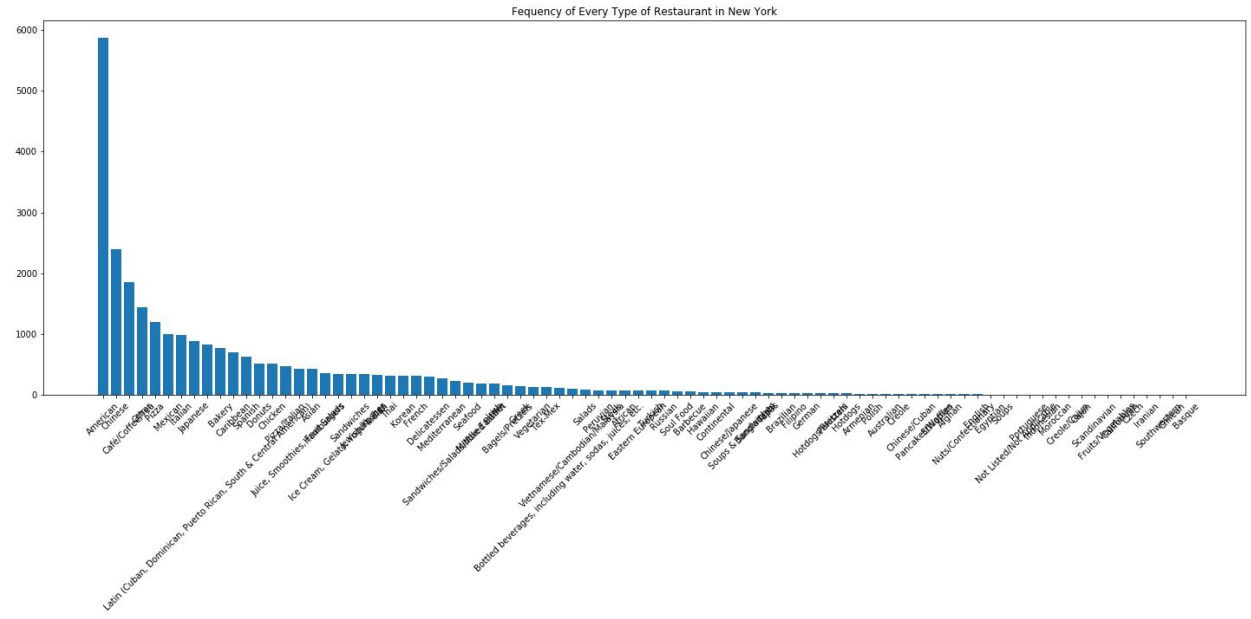
I then checked the coordinates data to see how many zip codes it had, and found that it had some duplicates which I removed using dictionary comprehension. The coordinates data

had some zip codes that were not in the health inspection data and vice versa, so I removed zip codes from each until I had 220 unique zip codes that were in both dataframes.

I then did a left merge of the census data onto the restaurant data. Because the census data contained all zip codes in the state of New York, the left merge excluded all zip codes which were not in New York City. I also included a column for the total number of restaurants in each zip code as well as number of restaurants per capita. I then found 41 rows for which the median income column was null. I examined these rows and found that they were very small zip codes with very few restaurants (many having less than 5) so I removed these rows as well, leaving a final count of 179 zip codes to be mapped.

Data Exploration

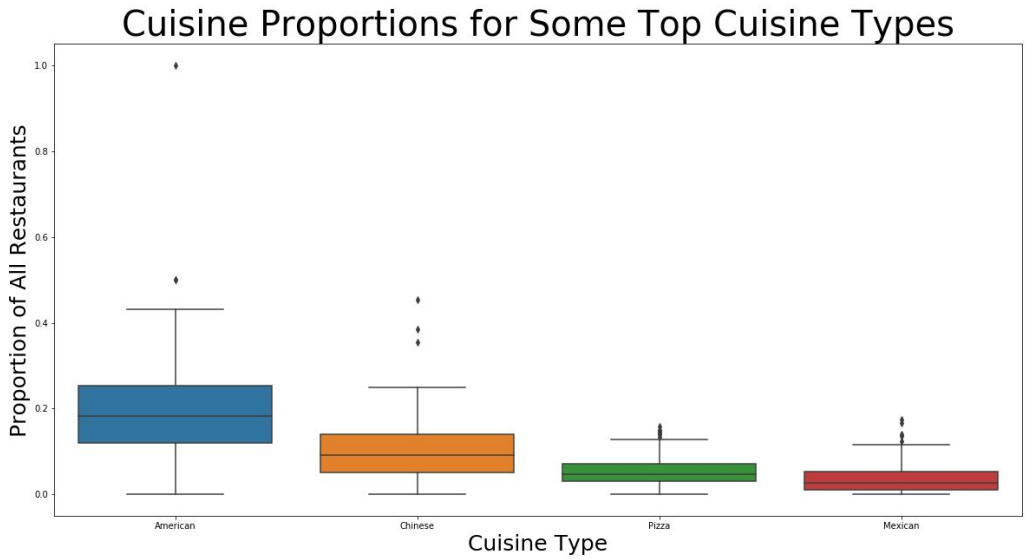
I plotted which cuisine types occurred most often and found the data to be very top heavy.



I found that American restaurants make up over 20% of all restaurants in NYC, and the top 25% of cuisine types make up over 80% of all restaurants in NYC. I believe based on this data that

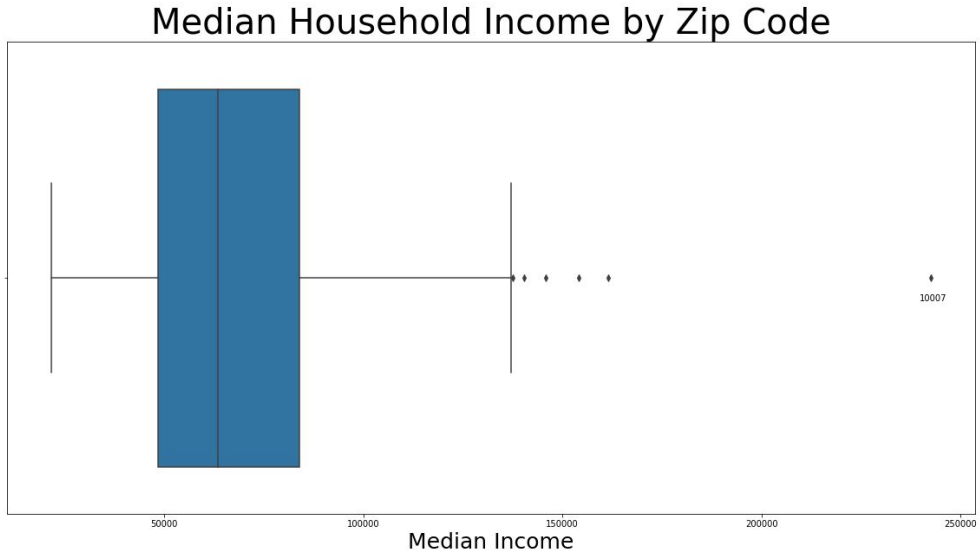
the tool I am designing will be best suited for those top cuisine types as there will be sufficient data to make proper comparisons between zip codes.

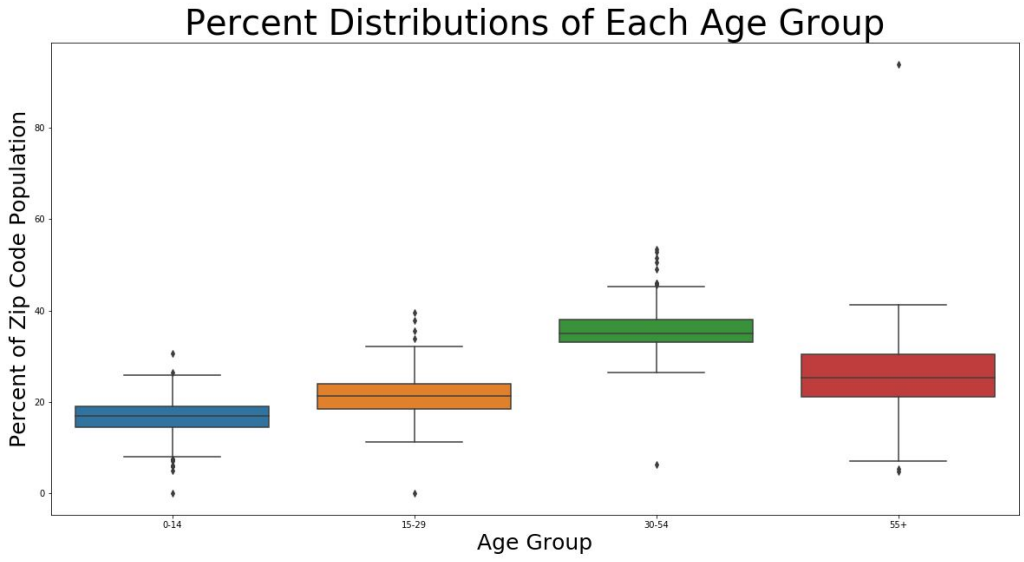
I also wanted to see how the proportions of some top cuisine types varied across zip codes, so I made a barplot.



Even for some of the top cuisine types (Pizza, Mexican) the variation in proportion from zip code to zip code is quite low. Because of this, I think it will be best to MinMax scale these values to get a better comparison between zip codes.

I also looked at the distributions of Median Income and the 4 age groups.





Based on the way the data looks, I think the best way to move forward with this tool is to create a single metric that can be used to rank the zip codes based the opportunity for success for opening a new restaurant. Because the various data are on different scales and I am most interested in ranking the zip codes, it is appropriate to scale all of the relevant columns with a

MinMax scaler so that they vary between 0 and 1. I can then simply add up all the columns of interest to get a final metric.