Capstone Project
IBM Data Science Professional Certification
Jim Kaczkowski
June, 2020

Examining clusters of Chicago restaurants by Zip Code, population and income

Introduction/ Business Problem

With the advent of the Covid-19 pandemic, restaurants face increasing challenges with business closures, employee lay-offs and, upon re-opening, maintaining profitability while conforming to public health, social distancing requirements. To mitigate the loss of revenue, some establishments are trying to promote out-of-home sales, enabled by online delivery companies such as GrubHub, Uber Eats, etc. The challenge for restaurants however is maintaining profitability when the online delivery companies can charge up to 30% commission on each order. See LA Times article and Forbes article.

There are many factors that are critical to a restaurant's success, such as proximity to customers, competition, labor costs, commercial rents, street location, etc. One hypothesis for successful restaurants in a post-Covid-19 world is that they will need to find more customers with more income and locate closer to them to enable more direct order pickup without reliance on 3rd party delivery companies.

This analysis will examine whether restaurant cluster data from Fousquare, combined with income and population data by zip code can define new opportunities for opening restaurants by existing restaurateurs and new restaurant investors.

Data

This analysis will use the Foursquare Places API as well as publicly available data to rank Chicago zip codes by population density, average income, and the number of competing restaurants. The following data sets will be used:

1. US Zip Codes with Latitude and Longitude This national data set supplies zip codes with geospatial coordinates. Latitude and Longitude from Chicago zip codes, beginning with "606", will be extracted to form the foundation of later API calls, using Foursquare.
2. Population by US Zip Code This is a Kaggle data set that shows population by zip code, as of 2010.
3. Income by US Zip Code This is a large data set from the IRS, from which median income by zip code can be extracted.
4. The Foursquare Places API Foursquare data will be used to specify the number of restaurant venues by zip code.

Methodology

The approach will be to develop stratified maps showing Chicago zip codes, separated into tiers, based on population density and median income. Foursquare venue data will then be overlaid to show clustering of venues by zip code. This analysis will then aim to enhance any clustering by examining primarily food venues while excluding businesses such as facial salons and yoga studios. Lastly the analysis will incorporate income and population by zip code to test whether more well defined clusters emerge.

Population Data

Population data was drawn from a Kaggle data set for the zip codes that begin with "606", representing the city of Chicago. A .csv file was uploaded into a SQL database, from which queries

could be performed. Once the relevant data was converted into a dataframe, histogram and choropleth maps were developed, using MatPlotLib and Folium respectively.

Income Data

Income data was drawn from IRS data set for the zip codes that begin with "606", representing the city of Chicago. A .csv file was uploaded into a SQL database, from which queries could be performed. Total individual income for each zip code was divided by the total number of individual tax returns to derive an "Average Income" for each zip code. Once the relevant data was converted into a dataframe, histogram and choropleth maps were developed, using MatPlotLib and Folium respectively.

Chicago Zip Code Coordinate Data

Chicago Zip Code Coordinate data was drawn from public data from opendatasoft.com. A geojson file was uploaded to the Github repository for this project, from which choropleth maps could be developed.

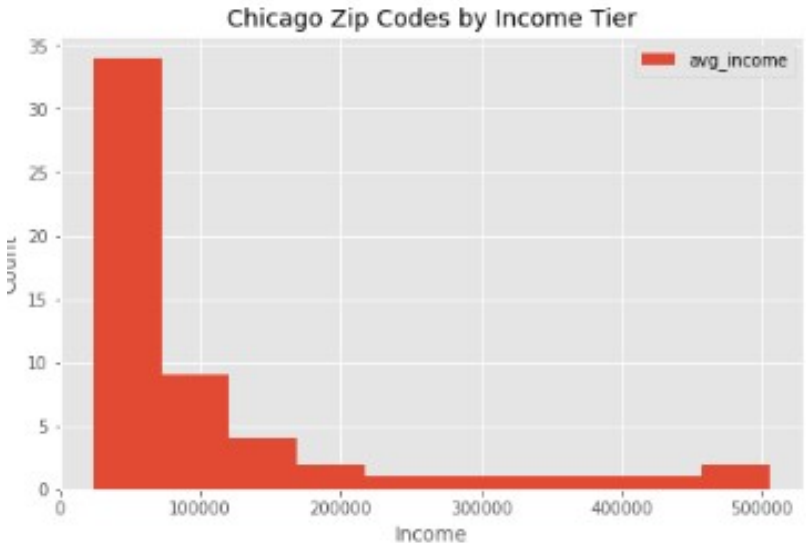
Foursquare Data

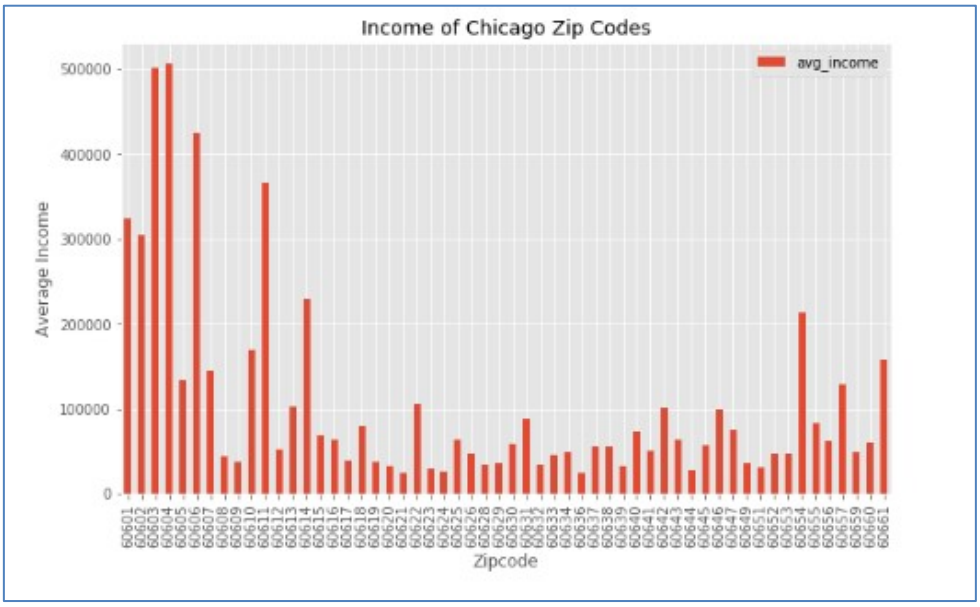
Venue data for Chicago zip codes was drawn from Foursquare Places API for the zip codes that begin with "606", representing the city of Chicago. A json file was downloaded and converted to a dataframe. The first level of clustering downloaded all venues for a given zip code, including restaurants, yoga studios, nail salons, etc. A second level of clustering was performed that eliminated the non-food establishments to be more relevant to the restaurant industry. A third level of clustering added income and population by zip code to the food concentrated list of venues. Due the large range of both population and income by zip code, income and population were normalized, using MinMax scaling from the SKLearn python library.

Results

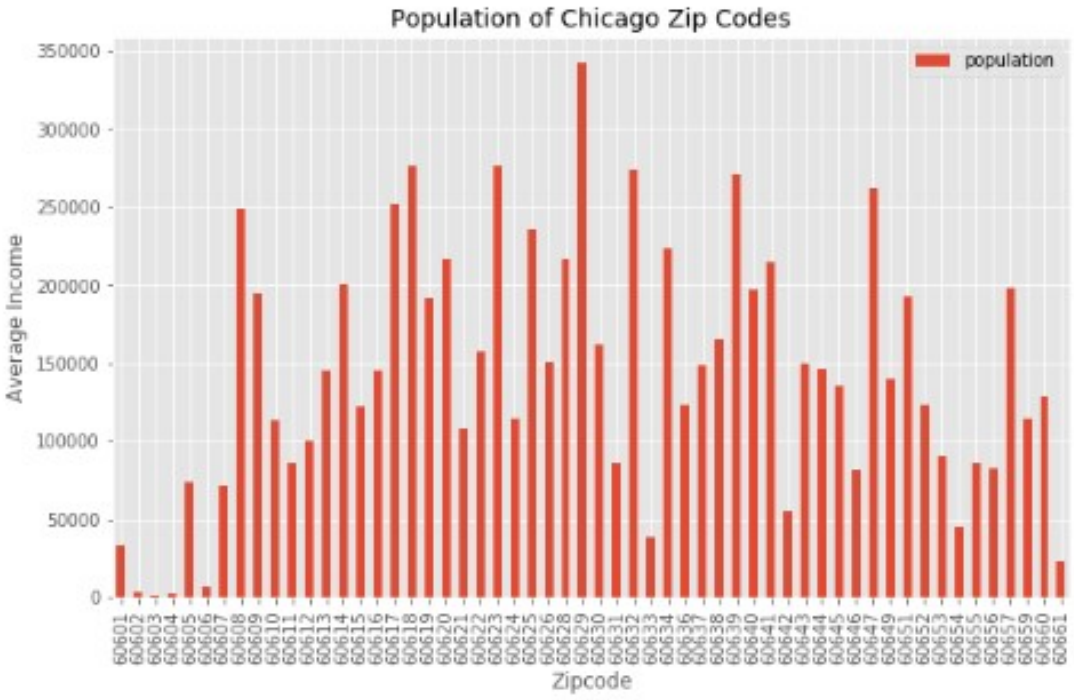
For the three variables: income, population and venue information, the results are as follows:

- Average Individual Income is highly skewed toward the central downtown area of Chicago (two zip codes). Income in 2 zip codes is roughly five times that of 34 other zip codes in the city.

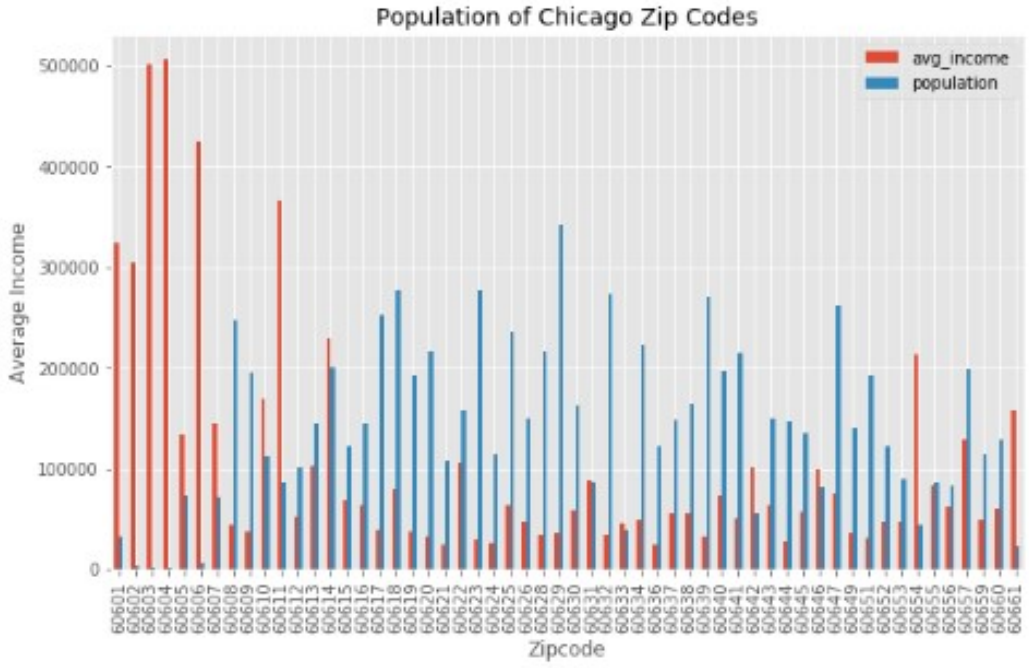




- Population, on the other hand, is highly skewed away from the city center of Chicago (the three zip codes on the left of the histogram).



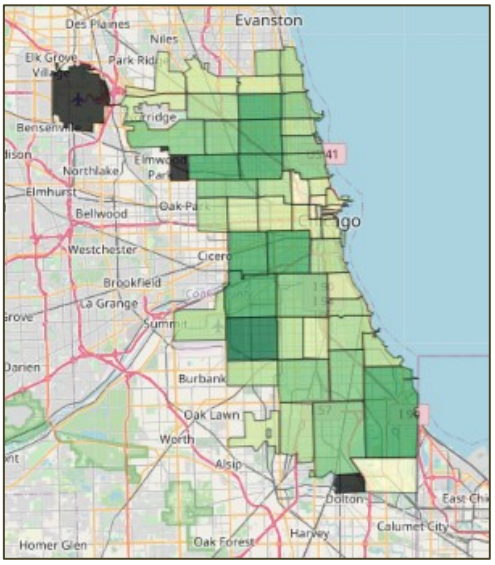
- A histogram of both income and population shows the least populated zip codes in the city center have significantly higher income.



- Choropleth maps of population vs income demonstrate the inverse relationship of population and income

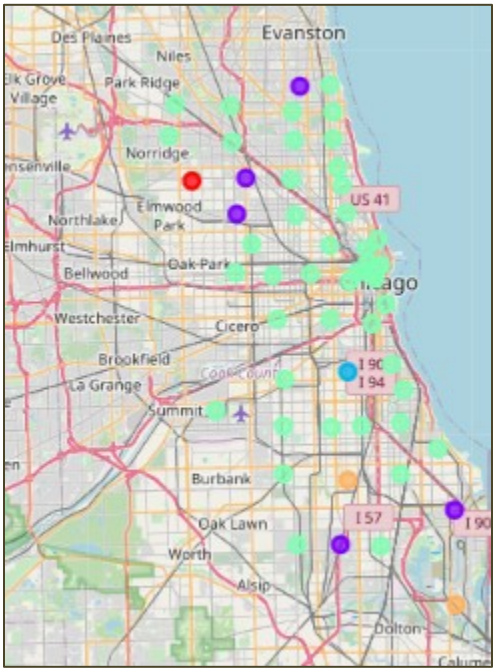
Income by Zip Code

Population by Zip Code



- Foursquare data was downloaded for each and clustered using K-Means clustering. The clusters were visualized via a folium map.

Venue Clusters in Chicago



- Specific cluster profiles of venues were not necessarily definitive as the figure above indicates. Specific cluster profiles can be viewed as part of the Github repository for this project.
- One of the challenges of the venue clustering may have been the inclusion of numerous non-food venues in the Foursquare results.

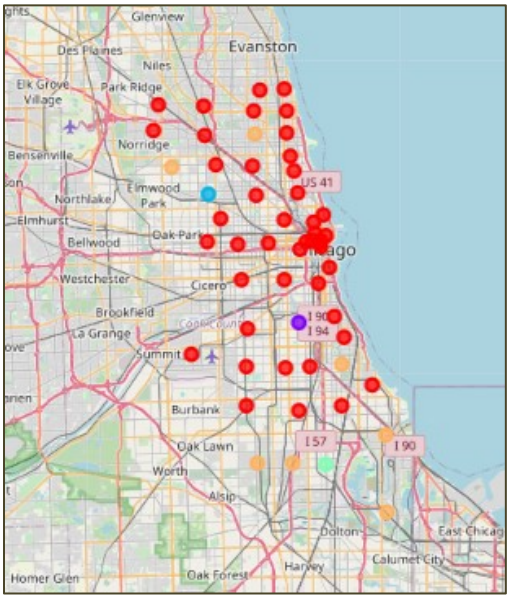
In [344]:
```
chicago_merged.loc[chicago_merged['Cluster Labels'] == 2, chicago_merged.columns[[0] + list(range(4, chicago_merged.shape[1]))]]
```

Out[344]:

	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	60656	Football Stadium	Historic Site	Park	Athletics & Sports	Sporting Goods Shop	Burger Joint	Bistro	Sushi Restaurant	Museum	English Restaurant
11	60632	Gas Station	Park	Fast Food Restaurant	Sandwich Place	Fried Chicken Joint	Donut Shop	Bar	Bank	Chinese Restaurant	Pizza Place
16	60617	Food	Bar	Wine Bar	Park	Eye Doctor	Empanada Restaurant	English Restaurant	Event Service	Exhibit	Falafel Restaurant
22	60623	Discount Store	Pizza Place	Gym / Fitness Center	Food Truck	Bike Rental / Bike Share	Café	Train Station	Park	Yoga Studio	Empanada Restaurant
24	60625	Park	Bank	Mexican Restaurant	Track	Gym	Soccer Field	Ice Cream Shop	Empanada Restaurant	English Restaurant	Event Service
29	60631	Food	Donut Shop	Bank	Event Service	Coffee Shop	Falafel Restaurant	English Restaurant	Exhibit	Eye Doctor	Yoga Studio
31	60633	Food	Greek Restaurant	Park	Discount Store	Lounge	Empanada Restaurant	English Restaurant	Event Service	Exhibit	Dumpling Restaurant
32	60634	Bar	Gaming Cafe	Gym Pool	Pet Store	Park	Yoga Studio	Exhibit	Empanada Restaurant	English Restaurant	Event Service
34	60637	Cosmetics Shop	Lake	Caribbean Restaurant	Salon / Barbershop	Discount Store	Park	Diner	Fast Food Restaurant	Falafel Restaurant	Eye Doctor

- In order to refine the analysis, non-food venues were removed from the Foursquare results to determine if clustering was more defined. K-Means clustering was then used. The results do not appear to have improved the definition of the clusters.

"Mostly Food" Venue Clusters in Chicago

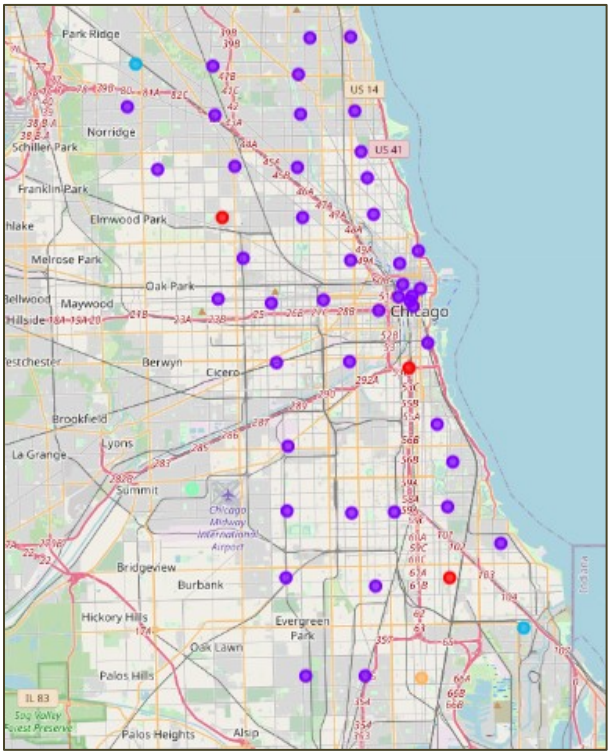


- Specific cluster profiles of venues from "mostly food venues" were not necessarily definitive as the figure above indicates. Specific cluster profiles can be viewed as part of the Github repository for this project.

A third component was then the addition of population and income data to the Foursquare venue data by zip code. Due to the large ranges of both income and population, those variables were normalized, using MinMax Scaling from the SkLearn library, so as not to skew the clustering.

Unfortunately, the addition of income and population data has not improved the segmentation across zip codes in Chicago neighborhoods. This may be explained by the relatively flat levels of income and population across across 34 out of the 56 Chicago Zip Codes.

"Mostly Food" Venues with Population and Income Data



Discussion

The objective for this analysis was to determine if population and income data provided more granularity into clustering of food venues in Chicago. Based on available data, the answer is that population and income don't improve the clustering of food venues in Chicago.

- The general uniformity of population in the zip codes outside the city center provides little illumination on specific opportunities for new restaurant location.
- The extreme high income of the two downtown zip codes appears to be offset by their sparse population. Elsewhere, income is generally uniform and thus provides little illumination on specific opportunities for new restaurant location.

Additional data may help to improve this analysis, such as the following:

- Data on commercial real estate costs by zip code to calculate potential return on investment for a new restaurant.
- Data on automobile and pedestrian traffic by zip code to assess viability of drive-through and "take-out" opportunities for new establishments.

- More detailed data on “restaurant type” and cuisine to assess more local opportunities for new establishments. (The Foursquare type information appears to consist of user supplied verbiage on restaurant type and thus leads to inconsistency of classification.)

It also must be considered if Zip Code is too large a designation to provide meaningful segmentation in a city like Chicago.

Conclusion

This analysis is a capstone project to demonstrate my mastery and use of skills taught in 8 previous courses in the IBM Data Science Professional Certification. In this analysis I have employed the following tools and techniques:

- Jupyter Notebooks
- Watson Data Studio
- Downloading and scraping web data from various sources such as government data portals and Kaggle.
- Data Hygiene and Clean up
- Creating SQL databases and performing SQL queries
- Manipulating JSON and GeoJSON files for both data analysis and geo-spatial data visualization
- Downloading venue and location data via API calls from the Foursquare database
- Use of Folium and MatPlotLib for data visualization, namely:
    - Histograms
    - Choropleth maps
    - Geo-location maps
- Use of K-Means Clustering and data normalization
- Sharing of work via the Github community

This project and this course sequence have provided me with the basic understanding of the tools of Data Science so that I can develop my skills further with more complex data. The skills and opportunities that I have acquired in this course sequence have ignited a curiosity for me to explore more complex analytical techniques and more complex data sets. I truly appreciate this first step in my journey to a career in Data Science.