IBM Applied Data Science – Final Project Submission

Use of Economic Indicators and Population Density to optimize franchise location

Prepared By: Stephan G. Maher

Part 1 – Problem definition and Data Sources

1. Question to be answered

Can we use economic indicators, population density, and information on competitors for each zip-code, to find an optimal location for our franchise?

1. Problem Background

My children work for a local delicatessen. Besides providing locally sourced cold- cuts, the deli also makes sandwiches and hoagies and either a soup or hot sandwich selection ranging daily from hot roast beef, pork, or meatballs. If I were to secure franchise rights, one of the first problems would be where would I locate the franchise? Before creating a financial profile, I would like to get the economic outlook and demographic data to see if I can optimize a location. The original store is in the Roxborough section of Philadelphia, specifically the “19128” zip code. I am also located in the same zip code.

2. Uses for this data

The problem to find a new location affects everyone. The real estate industry compiles data from the Census, IRS, and other sources to match home buyers with the best options. Recently Amazon accepted bids from locations all over the USA to open a second major hub. The small business association has statistics about the overall business outlook. Many factors can go into the decision for the optimal location, depending on the industry or personal preferences. While searching for data, I noted that small businesses make up over 95% of all U.S. businesses and 47% of U.S. annual GDP. Only about 40% of these businesses are profitable. By applying data science, I hope to find the optimal location for a franchise. This would be a good starting point for the business plan.

3. Description of the Data

For this evaluation, I would like to keep the commute within 20 miles, but far enough away from the current location so as not to be in competition. Some neighborhoods while close, are separate communities that would not directly compete with the target store. There are 3 suburbs of Philadelphia which fall within my initial radius. These are Montgomery, Chester, and Delaware counties. I will probably discount many Philadelphia zip codes due to the number of restaurants per zip code, but I will need to review population data before doing so. I’m also excluded New Jersey, which is within the 20-mile radius, due to the fact that some data is State specific. I would like to get information about the economic stability for each zip code, as well as rental cost information. Also, I would like information about competing stores, and population density.

I was able to find the following information by zip codes

- Zip Code with Spatial Latitude and Longitude per zip
- Population and population density by zip code
- Personal financial information by zip code.
- Residential housing and rent pricing by zip code – From this historic Data I computed a one and three year growth by zip code.
- As part of the IRS Data, I was able to find information on small business income. This is included with the Personal Financial information

4. Data and Sources

Location Data https://simplemaps.com/static/data/us-zips/1.73/basic/simplemaps_uszips_basicv1.73.zip This site contains a free table of US Zip codes in excel. The free version comes with Geo Spatial data by zip code, population and population density information. I found 180 zip codes with their spatial reference in the four-county area, and 138 of those exist within a "20 mile" radius of the initial store zip code of 19128. I also got municipal boundary data from Pennshare.

The next data set comes from the IRS. This data is made up of financial demographics by zip code for 2018. The IRS data was found at the following site SOI Tax Stats - Individual Income Tax Statistics - 2018 ZIP Code Data (SOI) | Internal Revenue Service (irs.gov) The most current year's data is 2018.

Census and Small Business Association data relies on NAICS codes. The specific code for a delicatessen comes from the following site, NAICS Code: 722513 Limited-Service Restaurants | NAICS Association. The basic 2-digit code for food services is 72. I was able to get consumer spending by zip code from the census at the following link, although that took the better part of a day to navigate the map. You can only get one piece of information at a time at the zip code level. At that point I had had enough. The reference for this site is Census Business Builder.

The last dataset that I imported was from Zillow. I have historical data on the median house price from Zillow from 1996 to the present. I computed the upward and downward trends for 1 year and 3 years prior to 2018, to match the census data. Zillow data can be found here. Housing Data - Zillow Research.

I also need to mention FourSquare as I'm using their data for free. I also used mapshaper to convert my Pennsylvania boundaries to GeoJson format.

5. Methodology

The first database that I found contained zip code information for the state of Pennsylvania. Once I had that, I could start to gather other data. I soon found that it was not that easy when i tried to build a choropleth map. In order to display much of the data that I had, I needed to find a geo Jason

File that provides the boundary markers of each zip code. I never did find one, but i researched and found a shape file from Pennshare and found a website where you can load a shape file and get back a Geo Json file. Next, I found three sources of data. IRS data up to 2018, Census data for 2018, and housing prices which provided a list of PA zip codes and the annual avg cost per zip code. The first thing I did was find the zip code for 19128, then I calculated a radius of 20 miles from that location and included the zip codes for four counties within that radius. I appended the distance column to the zip code file. I started with 173 zip codes, then I added the IRS data, I did not have lat. Lon. Information for 34 zip codes, so I dropped them. These could be postal box zip codes, and when I added the shape file, if information was missing, I could start over at that point. Then I appended each remaining file. While adding Consumer spending information, I only found 57 records with matching data. I also had the information at the county level. I adjusted the data with the mean by county. Housing prices are median home price. Once all the data was loaded, I needed to adjust my geo Json file to only include the zip codes that I would be using. I also needed to delete the existing file and recreate it each time I ran the process, then open the new file and add shapes to my map. I created a separate zip code data frame for the 19128 location of the original store. That left me with 123 zip codes to manage. To test this data I needed to apply Clustering. My first test involved Agglomerative Hierarchical Clustering. I tested using different linkages, and the data came out the same, so I used complete.

The second Clustering algorithm I used was K-means clustering. I should mention that for both clustering algorithms the data must be scaled. To find the best fit for K-means I used the bent elbow test. The theory is that as you add clusters, the sun of the squared distance decreases rapidly until it turns and the slope descends less. This usually leaves an elbow shaped line, and the point of the elbow is the maximized number of clusters. Again four seemed optimal, and I initiated this 10 times. It should be noted that k-means clustering will produce different results every time it is run, where hierarchical clustering will produce the same results run after run. The last item I used was the choropleth map, which visually shows the

6. Results

What I found was that the majority of zip codes ended up in two main groups, and 4 zip codes, two each ended up in 2 clusters. One cluster in the wealthiest suburbs of Philadelphia and the other cluster is in the business district of center city Philadelphia. The k-means algorithm grouped 4 of the center city zip codes into 1 group but was not visually remarkable. The choropleth map with folium markers does a good job of showing the concentration of wealth in the city and suburbs. And the zip code locations. If you review the data, you will see that the business income is smaller in center city as opposed to the suburbs, since there are a larger number of restaurants concentrated in 1 area. I centered my search around an area with the highest concentration of wealth. I had a problem getting the circle markers to display popup information, but what I found shows promise. Businesses just outside of the highest income zip codes have a greater density of people. This combined with higher-than-average zip codes looks like a business advantage. It looks like some people have already taken advantage of this situation, as there is a high concentration around Ardmore

7. Discussion

The main issue I had was finding data. The census bureau contains a lot of data, but small business facts are either generalized or very difficult to pull. The census bureau provides a number of maps and tables which makes it difficult to scrape information. I also learned more than what the course taught about mapping data and shape files. Basically I was trying to see if the data would show some other areas of opportunity than what I had already predicted. Basically, if I were to do this again, I would try to pull in more demographic data besides income and housing costs, and try to find more information about spending habits

8. Conclusion

If I wanted to open up a franchise location, I would look to see if there is an ideal location in Ardmore, or expand my search zone a little and look in Radnor, which is west of Villanova.

Thanks! Good luck on your project! Be kind when grading! Have a Nice Holiday!

Stay Safe!

Need small business data

References

1 https://data.pa.gov/api/views/INLINE/rows.csv?accessType=DOWNLOAD
County Latitude and longitude information
2 https://data.pa.gov/api/views/342b-rkgt/rows.csv?accessType=DOWNLOAD
Quarterly Census of Employment and Wages 2016-2018

3.https://data-phl.opendata.arcgis.com/datasets/b54ec5210cee41c3a884c9086f7af1be_0.csv
Zipcodes_poly.csv

5.https://www2.census.gov/programs- surveys/cbp/datasets/2018/zbp18totals.zip

6. https://www2.census.gov/programs-surveys/cbp/datasets/2018/cbp18st.zip?#

7. https://www2.census.gov/programs-surveys/cbp/datasets/2018/zbp18detail.zip

8. https://www2.census.gov/programs-surveys/cbp/datasets/2018/cbp18st.zip?#

9.https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/state/detail/SCPRC-EST2018-18+POP-RES.csv

10.https://www2.census.gov/programs-surveys/popest/datasets/2010- 2018/counties/totals/co-est2018-alldata.csv

11.https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/asrh/cc-est2019-agesex-42.csv

12. https://www.irs.gov/pub/irs-soi/18zp39pa.xlsx