Applied Data Science Capstone Report (June 2019)

Brandon W Galloway

INTRODUCTION

In today's economic climate, a common struggle seen across America is a slowdown in the foundation of new small businesses. A lack of capital due to outstanding debts and lack of business knowledge are some of the strongest of the many key hang-ups of small business ownership. In this project I have aimed to assist in eliminating the ladder of these two problems, business knowledge. When starting a small business it can be overwhelming knowing where to start, let alone where to start your business. My analysis of the Chicago business environment strives to bring insight to aspiring small business owners as to what areas of Chicago would be the best to help them grow their business.

I. DATA

Several key data-sets where leveraged to obtain a strong underlying knowledge base for the predictive logic. The first and most key dataset utilized was the foursquare developer API [1]. This dataset allowed the analysis of the most visited businesses in a specific area along with insights into the businesses that make up a specific neighborhood. The second dataset used was the Chicago Business Data Census from the United States Census Bureau [2]. This dataset contained data about each neighborhood of Chicago's businesses, including the amount of businesses of any particular business category, their locations, and their respective size in employees. This dataset provided core insights into business saturations for every neighborhood analyzed. The third dataset leveraged for this project was the Chicago Neighborhood Data stored at Mongabay.com [3]. This website lists each zip code in the core city of Chicago along with each neighborhood the zip codes contain. This dataset served as a sort of primary-key, linking the other gathered datasets together so that conclusions could be drawn across many different data sources. The final main dataset used was the Chicago Geojson and Illinois Geojson Data publicly available from the state of Illinois and GitHub [4]. This dataset was used to format Folium maps to better allow the end users to experience data outcomes in a functional and easy-to-understand format. It contains the boundaries for the neighborhoods analyzed. All together these datasets form the core knowledge-base for this project's predictive outcomes and recommendations and I believe they serve as well tested and well formatted data sources to be utilized in this report and project.

II. METHODOLOGY

Introductory analysis of Chicago Zip-codes

To begin the project, each dataset was examined and a goal was produced. It was then determined that to start the project It would be useful to first obtain the mongabay data. The Python library BeautifulSoup was used to retrieve the data from the website and parse it into a Pandas dataframe. That dataframe was then grouped by neighborhood to give a base dataset for further examination. Taking that dataset, the next few days were spent gaining insight into the neighborhoods selected for analysis.

Illinois/Chicago Geojson Data

Next, now that the neighborhoods were parsed and the zip-codes stored, the geodata was then pulled from the before mentioned data sources and passed according to what was needed for analysis. The initial dataset contained every zip code in Illinois so it was sorted and trimmed according to the scope of the project. During this time it was noted the GeoJson would need some formatting to be used in Folium maps and that the ZCTA5CE10 property contained the zip-code for the Poly-area. From this data we

were able to easily extract the Latitude and Longitude for each zip-code which was then placed into the dataframe df_geoZips. A beta Folium visual was produced at this time for both the Lat/Long and GeoJson data as can be seen in the figures below.

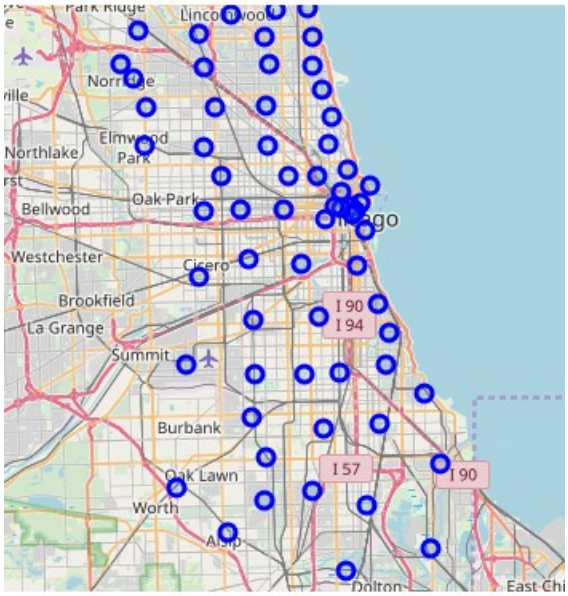

Figure 1: Latitude/Longitude Beta-Visual

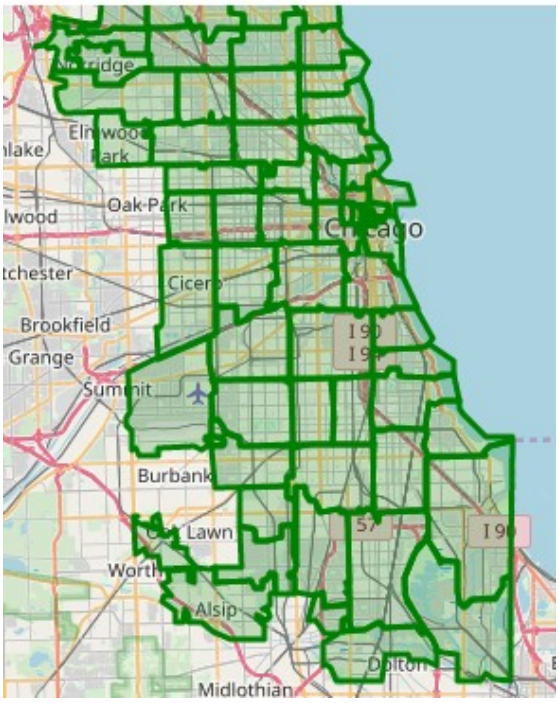

Figure 2: GeoJson Beta-Visual

Chicago Business Data

After the beta-visualization was finalized the first large data step was made through the import and placement of the Chicago Business Census Data into the df_ccd Dataframe. This dataset was sorted by our distinct zip-codes as decided in the Illinois GeoJson section and grouped by zip-code. In total over 40 thousand lines of data were collected.

Foursquare Data

To wrap up data management, foursquare data was obtained. Using the explore endpoint on each Latitude and Longitude collected during the Illinois GeoJson section, data was collected about each business in the area. That data was then onehot encoded, grouped and normalized to better fit our classification algorithms down the line. At this time a list of venue categories was also pulled down from the API and sorted according category, subcategory, and sub-subcategory for later use in a sudo-UI.

Clustering Foursquare Data

The foursquare data was clustered using K-Means clustering to help turn the unstructured data into meaningful insights into the status of each Chicago neighborhood. The result can be seen below. Clear clusters formed which was a highly desirable outcome and allowed for high confidence and ease of calculation going into cost-benefit analysis.

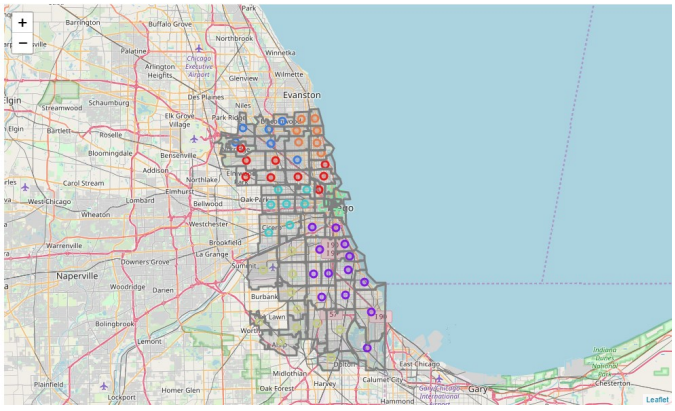

Figure 3: Clustered Foursquare Visual

Collection of Dataframes and Cost Logic

To finalize computations a cost function was derived to drive a recommendation system. We seek to minimize this function in our recommendations. This function was decided to be the sum of all 1-9 employee businesses of the selected market type in the area to a negative combined with a negative modifier for each non-related search item found in the foursquare data rankings. Scaling

small at first as to not encourage heavy competition but still leaving an impact on the later ranks to encourage a developing market.

III. Results/Discussion

Outcomes for Different Business Types?

The system was tested for many of the foursquare business types but samples from An American Restaurant, Chinese Restaurant, and public Art Sculpture are shown below.

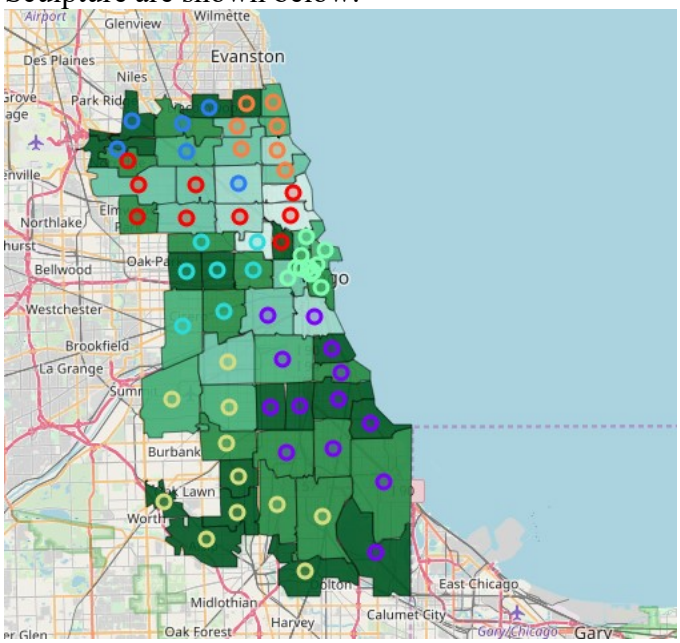

Figure 4: American Restaurant Choropleth

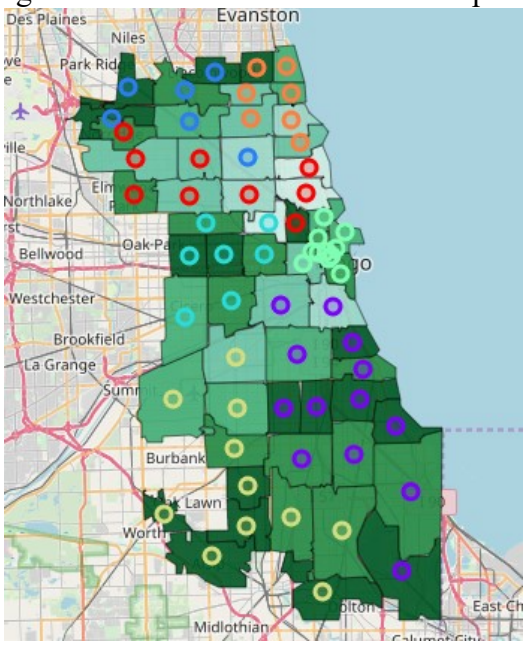

Figure 5: Chinese Restaurant Choropleth

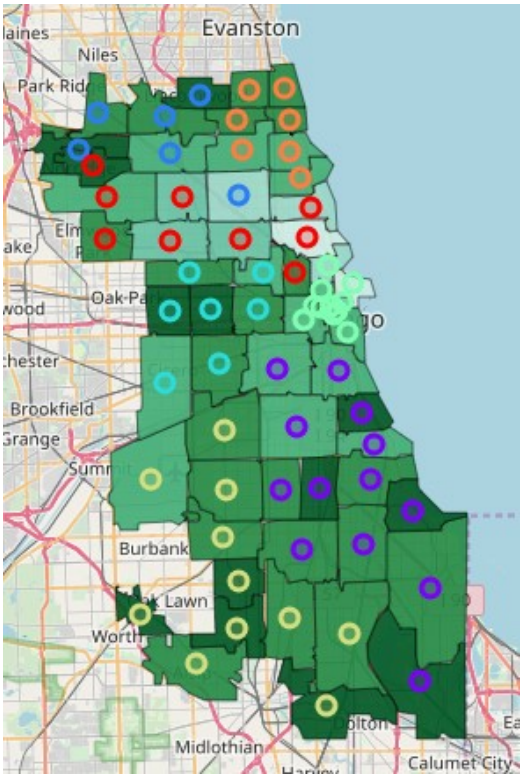

Figure 6: Public Art Choropleth

Analysis of Outcomes

From these three recommendation maps we can surmise that the south west coastline of Chicago is a prime location for developing small businesses and we can see this mirror in real life knowledge as the Oak Lawn area and lower south loop have experienced rapid growth in the past 10years and are an abundant market for trendy small businesses in the food sphere. In our Public Art outcomes it is noted the main Chicago hub is overcrowded but that the Calumet and Park Ridge areas are ripe for expansion. I believe this is due to their residential status, being quite packed with houses and lacking in parks, and their proximity to airports. Overall, I believe these outcomes do indeed hold weight in the real world and coincide nicely with my own experience of the Chicagoland area. Thus, I believe these outcomes could indeed serve small business owners and entrepreneurs well in the forming or continued expansion of their businesses.

IV. Conclusion

To conclude, it is clear that this project did produce some meaningful data in the realm of business risk assessment and could be a valuable tool to current and aspiring business owners. While currently accurate, I believe the analysis could be further expanded to include population/demographic analysis, consumer spending habits, and related synergistic and antagonistic business considerations. I found this project to be a very enjoyable experience and do intended to implement these proposed improvements in the future. Nevertheless, in its current form, the system provided ample useful information to help those looking to set foot in the Chicago business sphere.

References

[1] https://developer.foursquare.com
[2] https://www.census.gov/data/datasets/2016/econ/cbp/2016-cbp.html
[3] https://data.mongabay.com/igapo/zip_codes/metropolitan-areas/metro-zip/Chicago%20(IL)1.html
[4] https://github.com/blackmad/neighborhoods/blob/master/chicago.geojson