Opening a Chicago Area Bakery

David Griffin

March 2021

Introduction

The goal of this project is to find ideal locations within DuPage Country, Illinois to open a bakery. This would be of interest to anyone looking to open a bakery within the Greater Chicago area and surrounding cities. This person/group would be asking for tangible reasons with concrete data to guide their decision making. The business use-case lies in not just making a well informed decision on potential locations to decrease the likelihood of a business failing, but to maximize the chance for the business to have great initial success, growth potential, along with background information on where and who possible competitors are located.

Data

The data will be comprise four compenents from three different sources. The first two data sets were obtained from the United States Census Burueau website (census.gov). Through their main search interface (https://data.census.gov/cedsci/), information from the 2019 American Community Survey 5-Year Estimates. This table was filtered to show all places in Illinois and can be downloaded separately and ultimately uploaded to github. The particular data being analyzed from this set is the mean household income for locations in Illinois.

The next data source from the US Census Bureau is the 2017 Economic Census for Accommodation and Food Services (https://www.census.gov/data/tables/2017/econ/economic-census/naics-sector-72.html). This was downloaded, filtered to only Illinois and uploaded to github. The data points of interest in this set are the annual sales and annual payroll by locations in Illinois for the umbrella of 'Accommodation and Food Services' of which the proposed bakery would be included.

The Wikipedia entry for DuPage County, Illinois will be webscraped to obtain a list of places (cities, town, villages). Lastly, the Foursquare API will be utilized in finding venues in each of these places along with their associated data of which primary interest is the category of each venue.

Methodology

The bulk of the work involved data cleaning. The common field between the different data sources were the locations, of which the locations obtained from the Wikipedia page referenced above were used as the primary source and the American Community Survey and Economic Census entries for their locations had to be re-formatted to match the locations as shown from the Wikipedia listings. As an example, the Economic Census listed Elmhurst as 'Elmhurst city, Illinois'. Because of this coding was needed to go through each row in the dataframe, find where terms such as 'city, Illinois', or 'village, Illinois' were used and the remove these, to leave only 'Elmhurst', as an example from above. The American Community Survey required the same data cleaning regarding the locations.

Once that had been completed, data that was not to be included in the analysis was removed. In the case of the Economic Census, only the locations, annual sales and annual payroll remaind. For the American Community Survey, this had been done prior to uploading the data file for use to github. This file only included the locations and mean household income.

A moderate amount of data cleaning was necessary for the webscraping via the Beautifulsoup library. First the tags used in the html to identify the separate locations was used to obtain the text itself. Next, the text was split by the newline character into a list variable of places. Similar to the datasets above, the locations of interest for this project contained additional text such as '(mostly)' or '(partly)'. These pieces of text were also removed in the same fashion as the processes above. Lastly, the scraped text contained mayn different types of places and were not limited to the cities/towns. To remove these extra items, the index of the first city of Aurora was found and the list was set to start at this location.

To assist with map visualizations of the locations, the Nominatim function from the Geopy library was used. This looped through each of the places in DuPage to obtain the latitude and longitude coordinates of each and then assigned to a dataframe.

With the coordinates obtained, with the Foursquare API and the functions reviewed in the IBM Applied Data Science Capstone clustering lab, a dataframe consisting of the most common venues in each place was developed.

One hot encoding was used on this dataframe to convert the categorical data of the most common venue types into numerical data as required for use in K-Means clustering. Once these conversions are complete, this data is then combined with with the dataframe created earlier containing the household income, annual sales, and annual payroll into it's own dataframe stored as 'learning_df'.

Before performing K-Means clustering, the data in 'learning_df' is standardized using StandardScaler. Standardization is recommended for machine learning algorithms to have all units of data on a standard scale to avoid bias and have the data processed more efficiently.

Finally, K-Means clustering is ran with a K of 6. Any number of clusters could be asked for, however for this use case I found 6 to be an adequate number of clusters as too little or too many clusters would make meaningful differences between locations more difficult to discern.

Results and Discussion

When reviewing the clusters, even when asking for different numbers of clusters, the city of Chicago was always placed in a cluster by itself. Intuitively, this would make sense due the size, number of venues, income levels and the economic census figures. Additionally, when reviewing its annual sales and annual payroll of Chicago compared to the other clusters its figures are so significantly greater that it to compare the remaining clusters it was most helpful to remove it from the analysis altogether. If there is a desire to open a bakery within the city of Chicago, further analysis and clustering of the city itself would be needed.

The next easiest difference to notice is the size of the clusters. Cluster 0 contains 31 places, and the remaining four clusters contain just one location. From reviewing the sales and payroll, it is easy to remove a location from further analysis and make a strong recommendation for another. Burr Ridge appears to have significantly less business in terms of sales in the food and accommodation sector than any other location in the review and I would remove it from consideration. On the opposite end, cluster 1, consisting only of the city of Naperville, the annual sales in the food sector reveal a large gap between it and the remaining clusters. Payroll is also larger for this cluster, but no to the same degree as the difference in sales. In light of this, Naperville would be the ideal recommendation for the client.

Conclusion

After this review, it is clear that with a few pieces of key information an individual is able to make narrow down locations and make better informed decisions with the assistance of machine learning, particularly K-Means clustering. Three different cities were presented to the user in their own cluster that contained meaningful differences. The cluster containing the city of Chicago, along with Cluster 0 would need further analysis and most likely additional clustering within those groups would be of benefit if the user decides against clusters 2, 4, or 5.

Further review of economic trends in these clusters could provide additional guidance to the user along with any data able to be obtained on the venues themselves found within the clusters. The latter would probably prove difficult as a majority of this information would be private.

Even without this extra information, the user can be confident in the location they ultimately choose and have a good foundation of data to base their success on.