Introduction:

It is always desirable to open a business at a place where the supply is low and the demand is high but it is hard to figure out where it is. Many things have to be taken into consideration, such as the demographic and economic characteristics of the area, competition in the area and other business that might complement your own. One can’t easily go to every corner of the city to collect all the information. Fortunately, with the help from the city’s geographical data and the venues data from Foursquare, we are able to solve the problem without going on the field. This project will use the city of Cleveland as an example to demonstrate how data science can help people who are looking to open a common business like a pizza place or a sandwich place find the ideal position where the demand is high and the supply is low.

This process can be broken down to two major steps. First, one can acquire data from Foursquare, find out the popular business in all neighborhoods and group the neighborhoods into clusters where the neighborhoods in the same cluster have similar venues. It provides a pattern of venues that are widely successful in many neighborhoods. Then we are able to compare the venues between neighborhoods inside the same cluster and find out whether a common business is missing in a neighborhood. If a common business is missing it means the supply is low and we would know about the demand by checking the other venues in the neighborhood.

Data

The geojson data of the city of Cleveland is used. It consists of the names and the GPS coordinates of 36 neighborhoods in Cleveland. Based on the coordinates, venues within the radius of 1km of each neighborhood are extracted from Foursquare API. A total of 1064 venues are returned under 224 unique categories. The top 10 venues for each neighborhood are sorted out and then converted from categorical variables to dummy variables for the convenience of clustering. These venues will help us determine whether there is an opportunity for a certain business in a neighborhood.

Methodology

Based on the venues, the Cleveland neighborhoods are first grouped into 3 clusters using K-Means algorithm. The number of clusters is finalized after testing various numbers and 3 clusters seems to be the best fit for Cleveland neighborhoods. The neighborhoods in the same cluster would have a lot of common venues and the neighborhoods in different clusters are less similar. The reason for clustering is that if a lot of neighborhoods are in the same cluster, it suggests a pattern of venues is widely used and successful. The largest cluster is analyzed. The top 10 most common venues in this cluster are sorted out with their appearance percentage. Neighborhoods that don’t have the same top 10 venues as the cluster may have low supplies and high demands for the missing venues. That’s how business opportunities can be found. In the article, I examine the neighborhoods that have neither of the top 2 venues of the cluster, both of which happen to be food related.

Results

Top 10 venues of each neighborhood is acquired and shown in the Jupyter notebook. 36 Cleveland neighborhoods are partitioned into 3 clusters. The largest cluster has 29 neighborhoods. The top 10 most common venues in this cluster are pizza place, sandwich place, grocery store, bar, discounted store, pharmacy, fast food restaurant, convenience store, deli/bodega and Bank. 65.4% of the neighborhoods in this cluster have a pizza place in their own top 10 venue and 57.7% of the neighborhoods in this cluster have a sandwich place in their own top 10 venue. 6 neighborhoods in this cluster have neither a pizza place nor a sandwich place in their top 10 venues. 5 of these neighborhoods have 4 or more food-related venues in their top 10 despite not having a pizza and a sandwich place. The one left is ‘Central’ whose top 10 venues include grocery store, fast food restaurant, college arts building, restaurant, discount store, and light rail station. It has 2 food-related venues in its top 10, one of which is fast food.

Discussion and conclusion

The biggest cluster of neighborhoods based on venues suggests a successful business model, especially for food-related venues where they benefit from the high demand and are well complemented by other businesses in the neighborhood. 5 of the top 10 venues in this cluster are food related and the top 2 are pizza places and sandwich places. In the 6 neighborhoods that don’t have a pizza place and a sandwich place, 5 of them have other restaurants nearby to fill the vacancy.

However, there are not as many food places in Central. There are only 2 food-related venues in its top 10 and only one of them is fast food. Considering there is a college arts building and a light rail station nearby in addition to other stores, the demand for food venues like pizza or sandwich should be high.

Central has a low supply and a high demand for food industry. Opening a pizza place or a sandwich place in this neighborhood might be a good idea.