IBM Applied Data Science Capstone Report

Brian Jairam

February 7, 2020

Choosing location for new restaurants is tough. I propose a method of selecting the most advantageous U.S. cities to expand to by clustering cities with respect to the number of restaurants of similar type, and selecting targets either from cities in our current market's cluster, or in cities in clusters for which our type of food is common. This market segmentation information can also be useful for determine what products for our target markets.

1 The Problem

Suppose we are a the owner of a thriving U.S based restaurant chain with ambitions of expanding beyond the confines of our midsize city. Motivated by passion, fame, and of course, cold hard cash, we set out to raise the capital needed to open a brand new location. There is but one problem: where should we open our restaurant?

Our first instinct is to go somewhere big. More people means more exposure, right? However, our cousin (who works in the kitchen) reminds us that cities like New York or Los Angeles, while packed with potential customers, are also packed with stiff competition. However, he agrees that my mother's suggestion of Nampa, Idaho is not ideal either. Good, we do not really want to go home any time soon anyways.

On the other hand, having gained in-depth knowledge of our particular market, we are a bit hesitant to jump into cities that are vastly different from ours. If possible, we would like to set up shop in a city with similar tastes to where we are now.

So armed with a little knowledge of data analytics we picked up on Coursera, we decide to choose our target city methodically...

2 The Data

A data set of the 1000 largest US Cities by Population, hosted on OpenDataSoft was used. For each city in the data set, restaurants within 20 km of each city centre was found using the Foursquare API. Geo data for the boundaries of US States was obtained from the Folium documentation, for use in visualizing the number of restaurants found per state .

3 Methodology

1000 of the most populous cities in the US were imported as a pandas data frame, including included their populations, and coordinates. These cities were mapped using Folium for a visualization of their proximities.

Up to 200 restaurants were found for each city using the Foursquare API. For each restaurant, its name, location, and category was recorded in a pandas data frame. The number of restaurants found for each city was then calculated.

To get a better sense of which states had more food establishments with respect to their populations, the number of restaurants located in and populations of cities of the fifty states were summed. A ratio was obtained by dividing the total population of the major cities in each state by the number of restaurants found in these cities. The resulting ratios were plotted on a Folium Choropleth map.

The total number of restaurants of each type was also found, and the top twenty were visualized with a bar chart using the Seaborn library. The number of unique categories was also counted.

The cities were then clustered by the frequencies of each type of restaurant they contained. Using a one-hot vector transformation, taking a mean, and grouping by each city, these frequencies were tabulated in a data frame.

A k-means algorithm was used to cluster this data set. The elbow method was used to select the value of k.

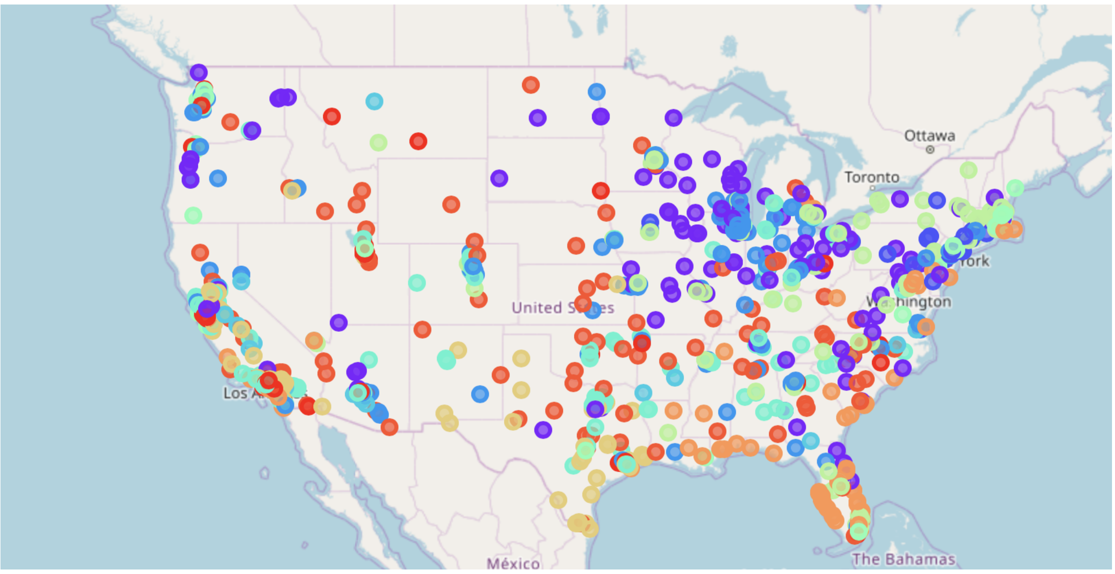
Based on the labels generated from the algorithm, cities were plotted on a map with colours corresponding to their cluster. The number of cities contained in each cluster was found as well.

The clusters were then examined visually for trends. Using a sample of a few cities from each cluster, a table was generated with the twenty most common types of restaurants in these cities. Inrtra-cluster similarities, as well as inter-cluster differences between cities was noted.

4 Results

The k-means clustering resulted in 11 clusters being generated. While some clusters looked more ambiguous than others, some clusters exhibit clear trends of certain types of restaurants. For example, in cluster 5, fast food establishments are prevalent. In addition, it was observed that cities that were geographically close were more likely to be clustered together, which is to be expected.

The following is a graph of the clusters generated for the mainland US.



We notice that orange dots tend to fall on the southeast coast. This corresponds to cluster 9, in which most cities have Mexican as its most prevalent restaurant type. This makes sense given the demographics of state of Florida in particular.

In contrast, most of the purple dots fall in smaller cities. These dots belong to cluster 1, which usually has a relatively more diverse selection of food.

These results can be used to determine where a new restaurant should be opened given a previous location's city or the type of food they sell. If the restaurant is located in a certain city, it would be advisable to target cities in the same cluster so that the results of market research can be reused. If the restaurant wants to expand to an area that is proven to be hospitable to their kind of cuisine, then the stakeholders can target clusters of cities that like that kind of cuisine. A more interesting use of this data would be to find cities with not as many restaurants of a particular type, but is clustered to cities where that type is popular. This would present unrealized potential for profit.

5 Discussion

There were a few issues with the analysis of this data set using Foursquare. Firstly, the Foursquare API is limited to only a certain number of calls per day, and a certain number of results per query. As a result, a smaller set of results was used.

Also, there were 75 cities which had no data. While this is unlikely to affect our results, this could also be indicative of certain markets that Foursquare's categorization approach or data collection methods are not suited for.

6 Conclusion

With the results of the analysis gathered, we leave the office feeling content. We have created a shortlist of a few cities our team feels good about expanding too. In fact, our cousin will be the one choosing his favourite from the list in lieu of a raise. We need to save some money right now after all!