IBM Applied Data Science Capstone Project Proposal

A. Chen

May 16, 2020

1 The Problem

Suppose we are the owner of a thriving restaurant in the US and are looking to expand beyond the confines of your city. As you begin to map out your expansion plans, we are struck with a tough question: Where should we go?

Perhaps we might consider large cities. After all, a larger market means more customers. However, at the same time, a larger market also means more competition.

Nor would a small city be necessarily ideal as an expansion location. There may not be enough customers to make the expansion worthwhile, even if we are the sole provider of our cuisine.

Finally, we may want to expand to a location that has similar tastes to the city in which we are located. That would allow us to put our extensive research and experience to good use.

Now, with some data analytics we recently picked up, let us shortlist some locations methodically.

2 The Data

A dataset of the top 1000 US cities by population, hosted on OpenDataSoft, will be used to determine possible cities for expansion.

Data from the Foursquare API will be to determine the most common restaurant types in each city within a radius of 20km.

3 Methodology

Population and location data for the 1000 most populous cities in the US was imported into a pandas dataframe. The data was then cleaned and saved for future use.

Next, the Foursquare API was used to obtain up to 100 recommended restaurants in each city within a radius of 20km. Each venue's name, category, and location was saved.

The total number of restaurants of each category was computed, and the top twenty categories were visualized using the Seaborn library. The number of unique categories was also counted.

Using a one-hot vector transformation, taking a mean, and grouping by city, the proportion of each type of restaurant was saved in a dataframe. The most common venues were determined for each city.

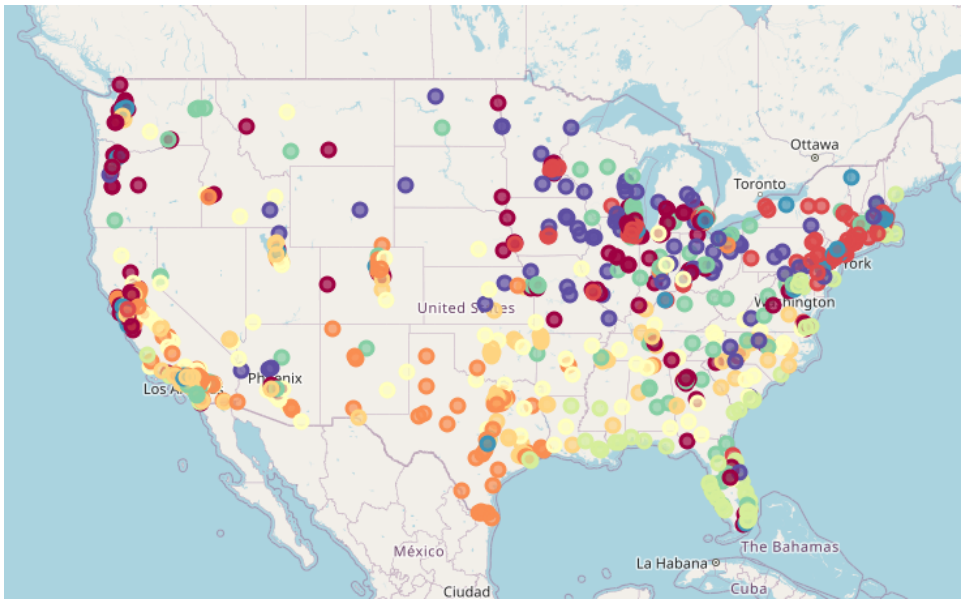
An elbow visualization was used to determine the optimal number of clusters for a k-means algorithm. The cities were then clustered, and a new dataframe was formed with the cluster labels and the most common venues.

These cities were then plotted on a map, with their colour based on the labels generated from the clustering algorithm. 5 cities from each cluster were then examined for preference trends.

In accordance with business requirements, another k-means clustering was performed on the cluster with which the home city was labeled, forming a shortlist for potential expansion targets.

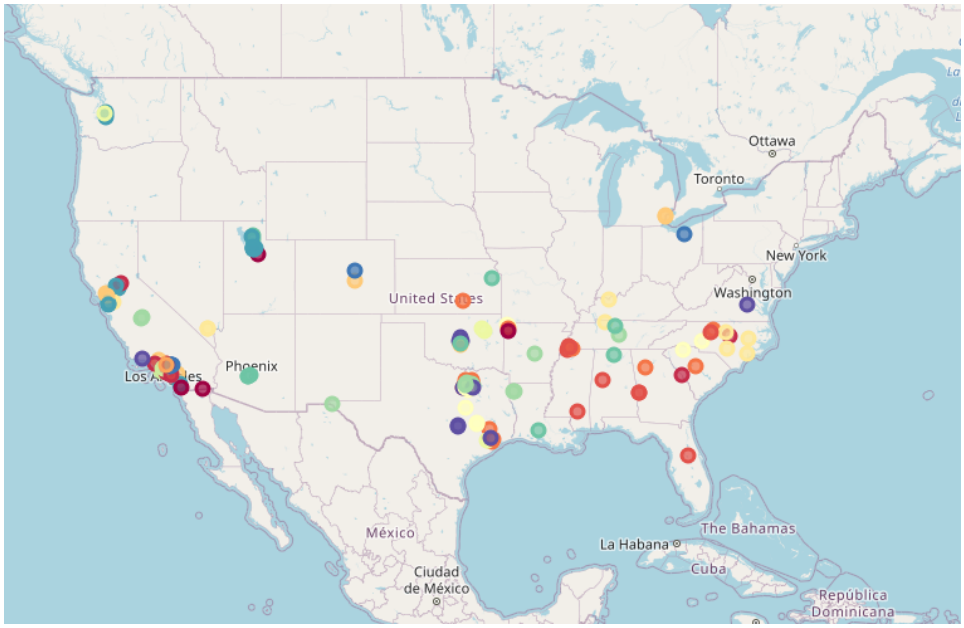
4 Results

The elbow method resulted in 9 clusters being generated. Most clusters had some differentiation: for example, cluster 3 showed a high prevalence of Mexican restaurants, and Cluster 9 showed a high prevalence of pizza and sandwich places. Geographical proximity, though not a factor for the clustering algorithm, seemed to be correlated with the cluster labels. For example, most of the Cluster 6 cities were on the southeast coast.



The majority of the cities in cluster 3 (orange) were observed to be located in Texas or California.

The example of an Italian restaurant in Mesquite, Texas was used to demonstrate an application of the results. A second k-means clustering was performed on the cluster in which Mesquite was located, yielding a list of 12 possible expansion targets.



5 Discussion

These results can be used to determine prospects for expansion given a starting city or the cuisine of the current restaurant. If the restaurant is located in a specific city, the clustering results can be used to determine cities in which previous market research would be applicable. Moreover, if the restaurant wants to expand to a city that prefers their kind of cuisine, the results can be used to determine cities where the restaurant would be welcomed.

Finally, an interesting application of the results would be to choose cities in which a specific cuisine is not common, but are clustered to cities that enjoy said cuisine. In conjunction with the population growth data, these cities could represent sources of unrealized profit, allowing the business to grow with the city.

There are some concerns with the use of the Foursquare API. First, the explore tool was used, yielding recommended venues, which are subject to user bias. Moreover, Foursquare imposes a limit of 100 results per search, so each city only had a maximum of 100 data points for analysis. Finally, some

cities did not have a full 100 restaurants in a 20km radius. As such, the computed preferences for these cities may not have been accurate reflections of their tastes.

6 Conclusion

With the results of the analysis in hand, we now have a much shorter list of target cities. Market research can be much more in-depth and specialized to each city, saving us money and yielding more useful data. Perhaps we could perform analysis on those results as well!