Data Science Capstone

Final Project Paper

Introduction

In this scenario, we are looking to use Python coding, Foursquare data, and K-means clustering to determine the best city to open an upscale cocktail bar in Scotland. There are many ways to determine the ideal location for a new business, from a large potential customer base, to a large tourism sector, to finding a gap in a particular market. In this instance, we will be looking to the best city based on the types of venues that are already common in the given cities. Our investors are interested in opening within the country of Scotland, but in a city where competition for an upscale cocktail bar will be low. Additionally, it should complement other venue types, such as entertainment, sport, and theater type venues. Pubs would be considered similar establishments, whereas full-scale restaurants and coffee shops would not be considered similar. Other retail outlets are also not competitors.

To carry this out, we will use an analytical approach and consider at least the 50 largest cities in the country. We will be using a machine learning technique to break all these cities in Scotland up into distinct groups based on the types of businesses already present in those cities, and then use those to determine how unique a new upscale cocktail bar would be.

The stakeholders are our investors in such an establishment and, eventually, those who have to permit this new establishment. In showing that machine learning algorithms have been used to determine what city to target, we can hopefully persuade investors and convince city officials that the business will thrive and help further establish the entertainment centers of said city. Both of these audiences are very aware of the different market possibilities, but would like to use data to determine the optimum location. They are not interested in overtaking competitors, and rather are focused on finding market gaps.

Data

We will be using data from several sources and combining them to carry out our analysis.

First, the largest cities list comes from Wikipedia -- it will be scraped from the page. Although there are other dimensions to this dataset, we will mainly just use the Locality name. These data were accessed in early 2018, so it is up to date enough to use for this analysis.

Second, I will pull venue data from Foursquare using the developer API. This is also data updated to 2018, so it will be up to date enough to use for this project. Specifically, we will be using coordinates of city centers scraped from a previous project to access a certain radius around cities, and to collect the number of venue types in that area.

Methodology

The first step was to scrape the list of cities from the Wikipedia page listing the top cities in Scotland by population. This was brought in and was formatted into a Pandas dataframe.

Then, a csv file of pre-collected coordinates for all the cities in Scotland was uploaded and merged with the largest cities list.

For quality control purposes, a map of Scotland was created to make sure there were not errors in importing or with the coordinates used. The map showed no outliers, and that all the points were where they were expected to be.

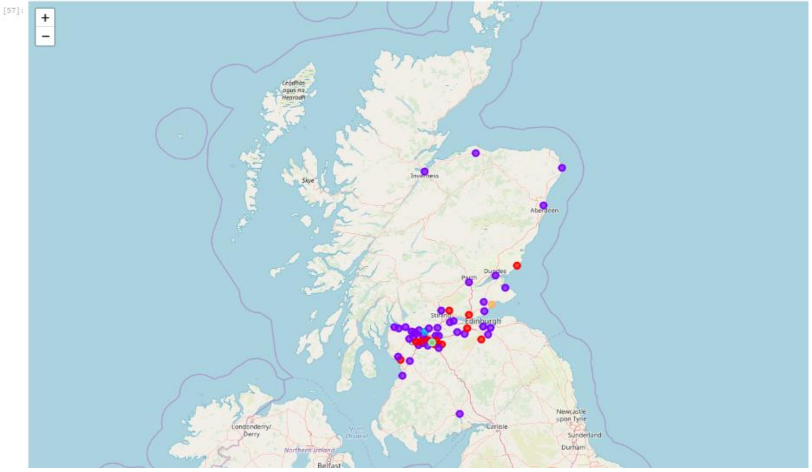
Next, i connected to the Foursquare developer API using unique login credentials generated earlier in this course. This was used as a pathway to get 100 venues in each of the cities, which we could strip of many of the details to just use venue type. Each of these venue types were converted to one-hot (binary) coding and then grouped and sorted to find the most common venue types in each city.

Using this output, we then clustered the set of cities (filtering out those which returned a n/a value) using k-means clustering based on the top 10 most common venue types in each city. The resulting clusters grouped together cities with similar economic makeup. Using this, we can investigate which cluster has the venue makeup we're looking for to open our new unique establishment.

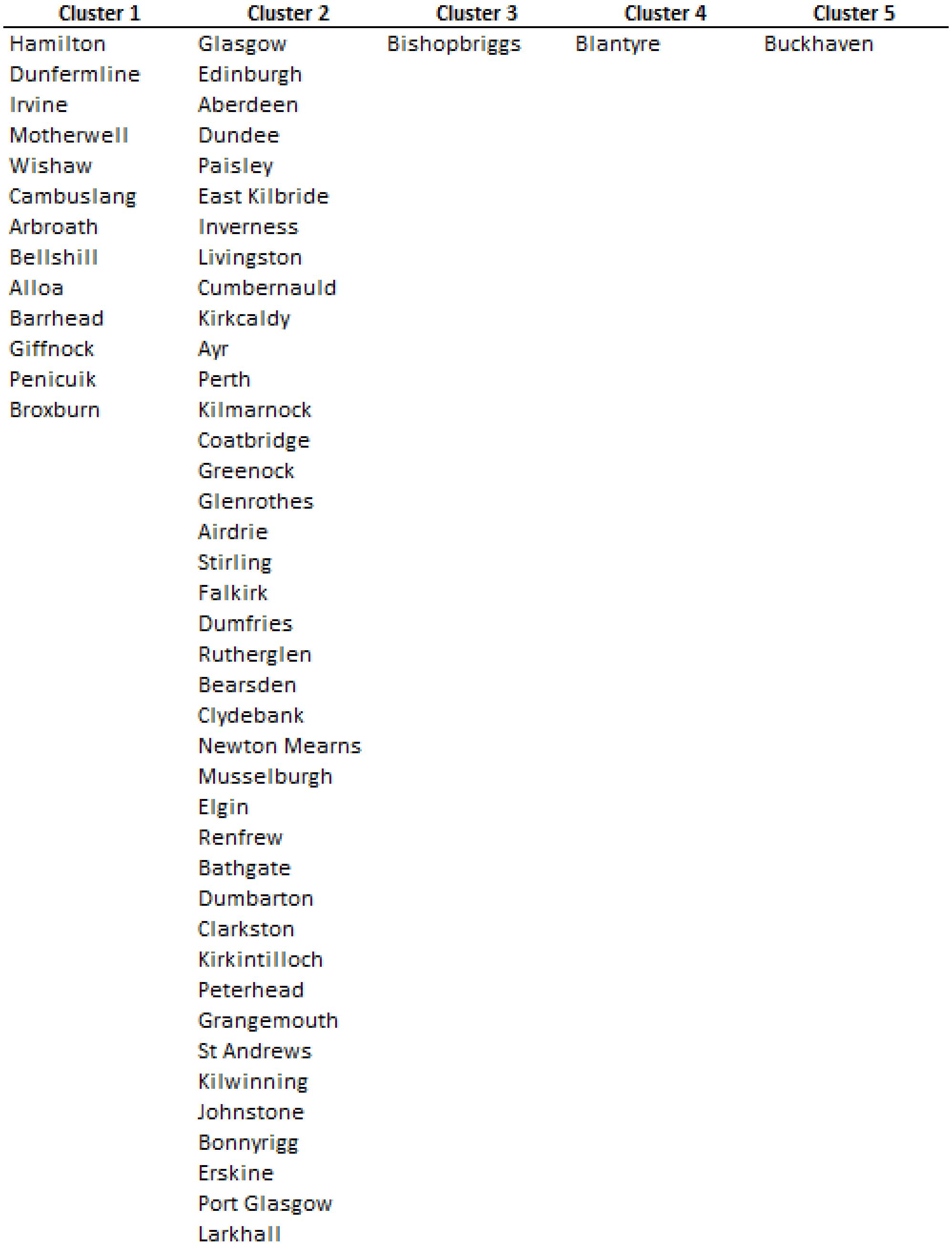
Results

In the end, we have five unique clusters based on venue types. These ranged from the largest cluster having a size of 40 cities, with the three smallest clusters having a single city, Bishopbriggs, Blantyre, and Buckhaven. Based on the venue makeup of these five clusters, we can see that they are indeed quite different, and can say that our clusters are a successful representation.

Below is a map of the five clusters generated using matplotlib.



The clusters are shown below:



Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Hamilton	Glasgow	Bishopbriggs	Blantyre	Buckhaven
Dunfermline	Edinburgh			
Irvine	Aberdeen			
Motherwell	Dundee			
Wishaw	Paisley			
Cambuslang	East Kilbride			
Arbroath	Inverness			
Bellshill	Livingston			
Alloa	Cumbernauld			
Barrhead	Kirkcaldy			
Giffnock	Ayr			
Penicuik	Perth			
Broxburn	Kilmarnock			
	Coatbridge			
	Greenock			
	Glenrothes			
	Airdrie			
	Stirling			
	Falkirk			
	Dumfries			
	Rutherglen			
	Bearsden			
	Clydebank			
	Newton Mearns			
	Musselburgh			
	Elgin			
	Renfrew			
	Bathgate			
	Dumbarton			
	Clarkston			
	Kirkintilloch			
	Peterhead			
	Grangemouth			
	St Andrews			
	Kilwinning			
	Johnstone			
	Bonnyrigg			
	Erskine			
	Port Glasgow			
	Larkhall			

Discussion

Looking further into the types of venues within each of these, we determine that our final cities will be one of the smaller clusters. The largest two clusters were put together based on similarities that include the presence of bars and pubs, which are often the second, third, or fourth most common venue type. Additionally, while there are some “complementary” types of establishments in these clusters (such as cafes and restaurants), it isn’t as common that there are event and entertainment venues.

One of our clusters, Bishopbriggs, seems to be the ideal location for this venture. Bars or pubs are not listed as a top ten venue type. Additionally, many of the most common venue types seem complementary such as a park and an event space. Given that Bishopbriggs is a small suburb on the outskirts of Glasgow, a new upscale cocktail bar could be a unique offering in this neighborhood, being a venue type traditionally found in urban downtown areas. We think it would be a good move by our investors to jump on this market opportunity.

Conclusion

In this report and corresponding materials, we have used an analytical approach using multiple sources of third-party data to determine the optimum city in which to open a new upscale cocktail bar. Using stakeholder feedback and parameters, we have determined that the Glasgow-area suburb Bishopbriggs seems to fit many of the criteria we are looking for, having many of the complementary types of venues that would help funnel potential customers, as well as a market opportunity as a drinking establishment.