IBM Applied Data Science Specialization

Capstone Project

Arshiya Tripathy

Estimating similarities between cities using Foursquare data and K-means clustering

12/03/2020

Thursday

Contents

1. Introduction

A lot of people who have to travel to different cities for work or vacation for a longer duration would want to understand how similar or different is that place in comparison to their current residential city or a city they have lived in past or are very familiar with. Someone would like to visit/stay in a city very similar to their hometown and someone would want the complete opposite. Adare, County Limerick in Ireland is a very quite town full of heritage whereas Dublin is modern metropolitan city filled with top end restaurants, theatres, shopping centres etc. A city can be defined by the venues and places it has to offer, some can be rich in culture with lots of museums , chic cafes and other can be fast and modern with high-end bars, fast-food chains etc. Thus, two cities can be compared based on similarities or differences between what venues each has to offer to it's residents. This profiling can be used by travellers to choose the city based on their preferences, making their stay more comfortable/enjoyable.

Ireland

2. Problem Statement

Cities and towns in the island of Ireland (Norther Ireland and Republic of Ireland) are selected for this analysis. These cities are then clustered to determine which set of cities are similar or different than each other based on the venues present in them. This clustering will help decide the traveller, based on his/her preference which city should he/she visit next. The traveller can compare their current city or any other city they are familiar with to decide the next city they want to visit by looking at the cities present in the same cluster of their current or other know city.

3. Data for analysis

A list of 60 largest town and cities on Island of Ireland is collected by web-scrapping the Wikipedia page https://en.wikipedia.org/wiki/List_of_settlements_on_the_island_of_Ireland_by_population. This gives the name of the city, province, county and country's name along with a small description and population.

Geopy python library is used to get the city centre's lat. and long. For each of the 60 cities. The city name is used to query foursquare database to get a list of popular venues around the latitude and longitude of the city's centre. All this information will be used to create clusters of cities based on the venues they have to offer.

Sample of the dataset, which is finally produced by combining the **Wikipedia**, **Geopy**'s latitude and longitude and **Foursquare's** common venues for each of the cities is provided below, settlement being city/town's name:

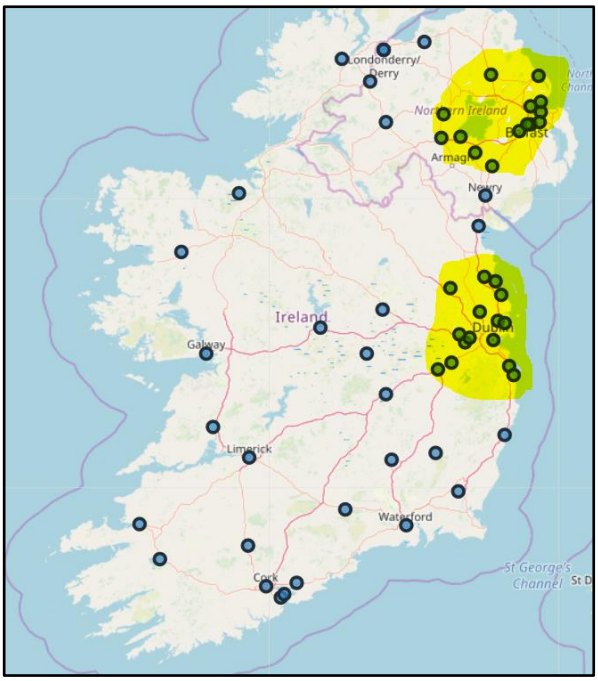
Settlement	Population	Province	County	Jurisdiction	Description	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
Dublin	1173179	Leinster	Dublin	Republic	Capital city of the Republic of Ireland and ha...	53.349764	-6.260273	Pub	Coffee Shop	Café	Clothing Store	Hotel	Restaurant
Cork	208669	Munster	Cork	Republic	Largest city in the province of Munster in the...	51.897928	-8.470581	Pub	Café	Coffee Shop	Burger Joint	Restaurant	Bar
Limerick	94192	Munster	Limerick, Clare	Republic	Principal city of Ireland's Mid-West Region an...	52.661252	-8.630124	Café	Pub	Hotel	Italian Restaurant	Coffee Shop	Restaurant

4. Methodology

4.1 EDA

Location

By plotting these cities on the Ireland's map it can be seen that most of them are present on the eastern side of the island and most of them are concentrated near two locations, Belfast and Dublin.

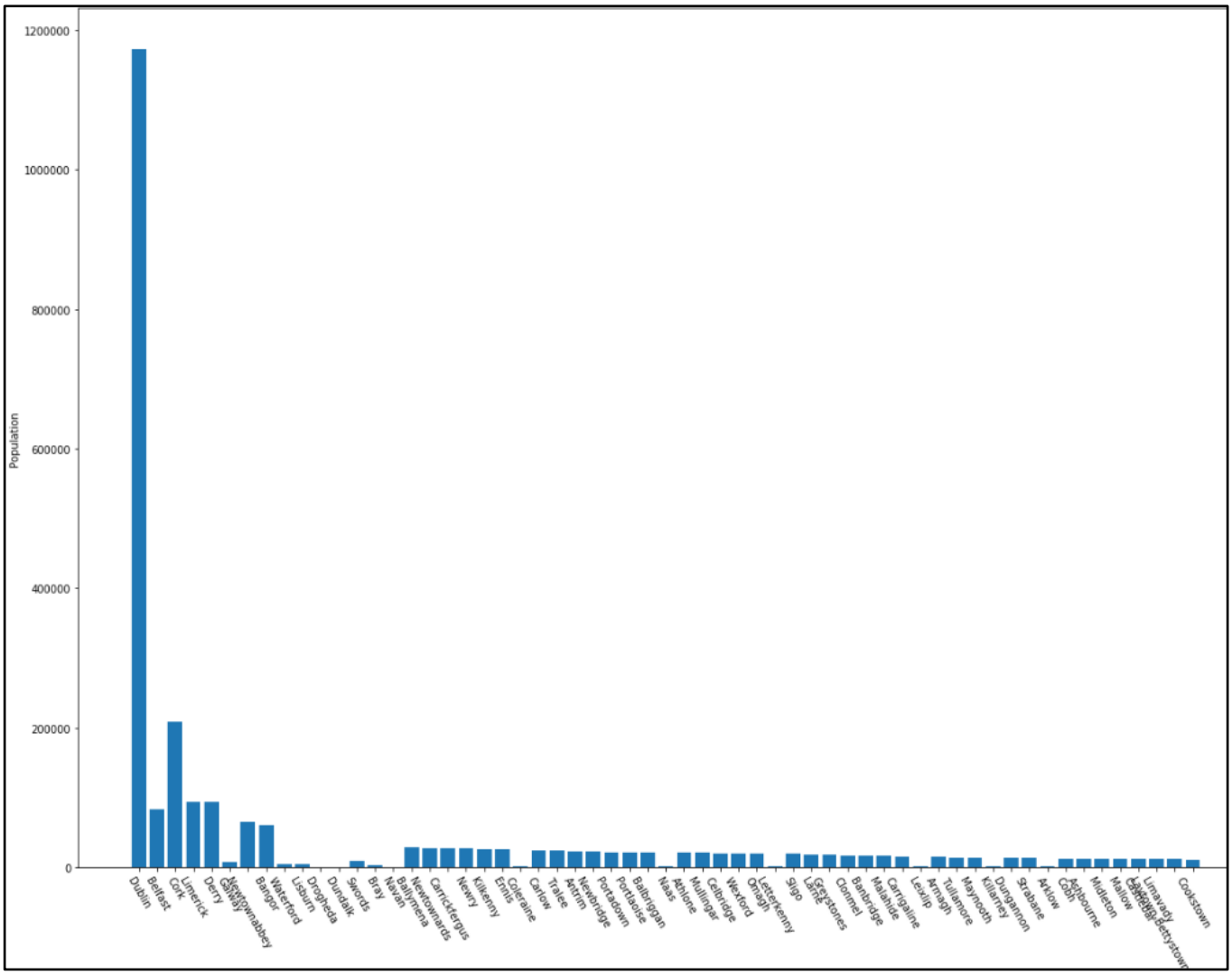


Looking at the location, most of them are present in the Republic of Ireland (40) than Northern Ireland (20). Leinster province also has most of the cities, 24 of them.

Jurisdiction Province	Northern	Republic
Connacht	0	3
Leinster	0	24
Munster	0	11
Ulster	20	1

Population

The top 3 biggest cities based on the population are Dublin, Belfast and Cork, Dublin being the biggest one having over a million residents.



4.2 Clustering

Analytical Data Set

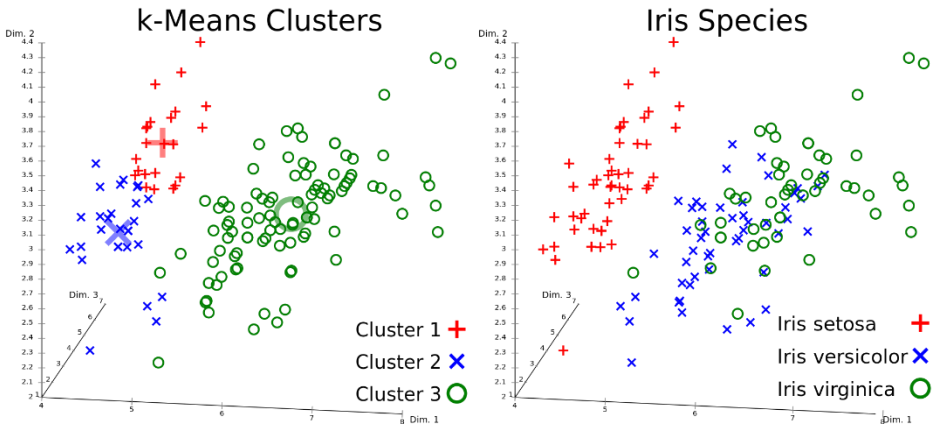
Since this project is around helping the traveller find similar cities based on the popular locations/venues offered by each city, these Irish cities are clustered based upon their venues only. The top 10 common venues of each city are captured and then transformed using one-hot-encoding. Below is the final table after implementing one-hot-encoding which will be fed to the clustering algorithm.

Castle	Cheese Shop	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop	Comedy Club	Comic Shop	Concert Hall	Construction & Landscaping	Convenience Store	Cosmetics Shop	Deli / Bodega	Department Store	Dessert Shop
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

K-Means Clustering

According to **wikipedia**, ***k*-means clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*-Means minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can for example be found using k-medians and k-medoids.

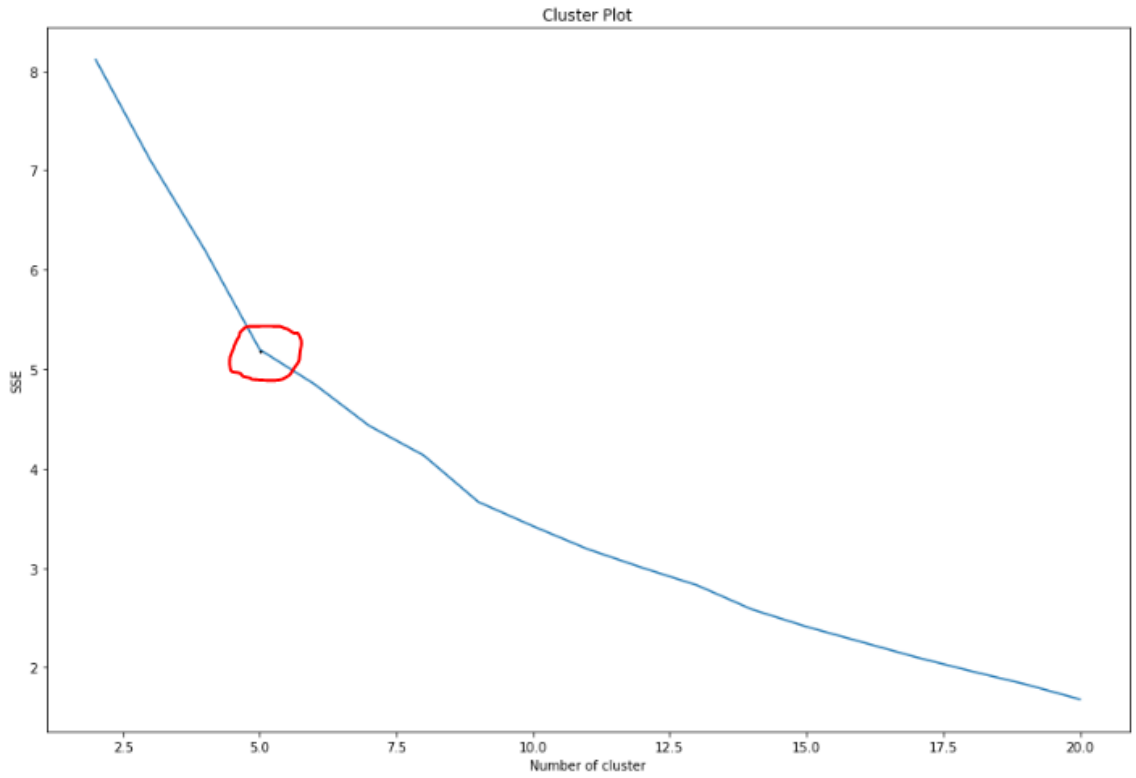
In short, K-means clustering is an unsupervised machine learning technique which is used to mine unknown information from set of objects based on their distance (in multi-dimension space) from each other. The distance is computed based on the location of each object, which is defined by it's features. Example of clusters produced by it.



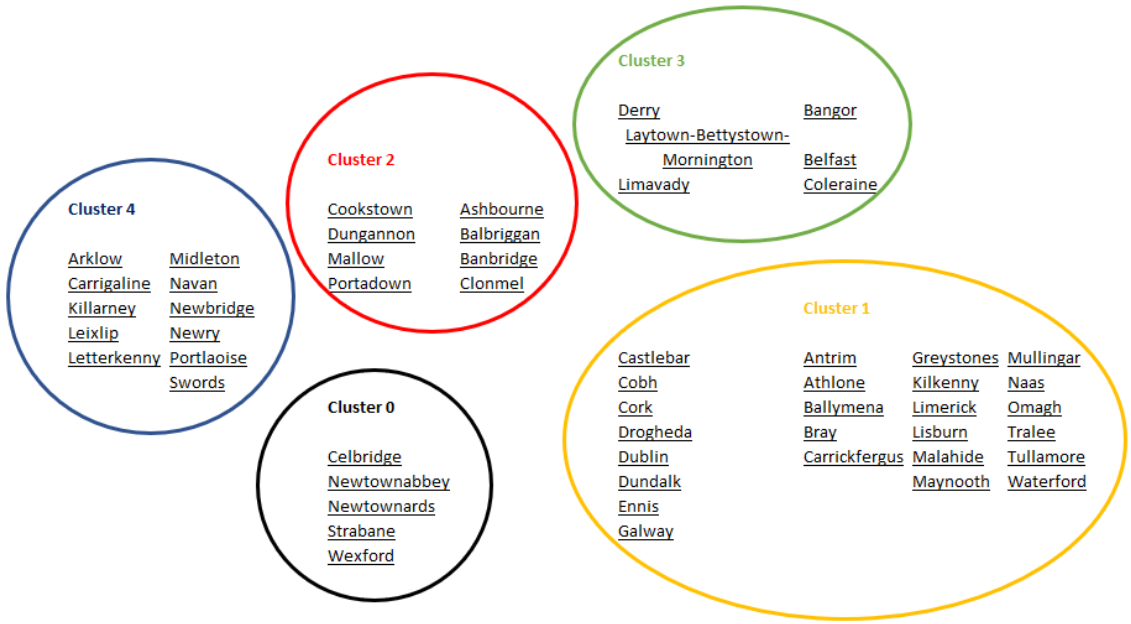
.

5. Results

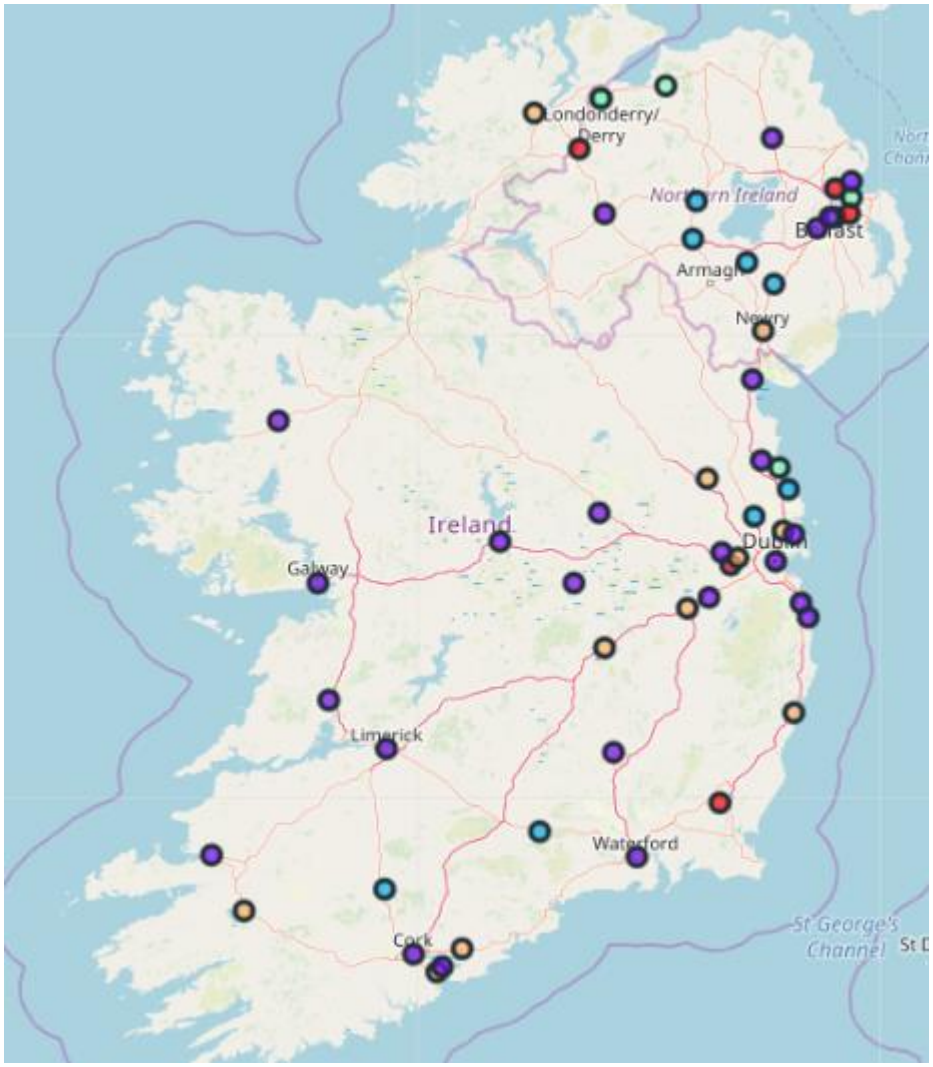
The K-means clustering is done to group the 60 Irish cities according to their popular places/venues associated with their city centre. The dataset, with, hot-encodes venues, is used to run k-means algorithm on. The number of clusters is decided by re-running the algorithm with different k values and determining the SSE for each of them. The elbow point in here is selected as the best number of cluster, i.e 5.



Cluster 1 has most number of cities, 25; cluster 0 has the least number of cities, i.e 5.



Also, These clusters are all over the Ireland.



The most common venues/places over Ireland are Pubs. This was no surprise as Irish people are know all over the world for their pub culture and wiskeys.

Venue Type	Freq. of 1st most popular in a city
Pub	14
Hotel	7
Coffee Shop	6
Pizza Place	5
Café	4
Supermarket	3
Diner	2
Restaurant	2
Italian Restaurant	2
Clothing Store	2
Plaza	1
Department Store	1
Spa	1
Grocery Store	1
Gift Shop	1

```
Chinese Restaurant     1
Fish & Chips Shop      1
Food                   1
```

6. Discussion

From this clustering exercise, it can be shown that the cities can be clustered based on the type of popular places present in them. Any traveller can find a city which suits his/her taste of the shops, eating places etc. They can also compare their current city with the clusters and see which one is most similar to their current neighbourhood. This will help them decide how comfortable they will be if they move there. If this methodology of clustering is scaled up, by bringing in other macro features like population, cultural diversity, living expenses etc. These clusters can be modelled better and new insights can be discovered.

7. Conclusion

60 biggest cities/towns across island of Ireland were clustered based on their popular venue types into 5 clusters. K-means unsupervised machine learning algorithm was used for this exercise, to discover similarities and differences between them. Tools like **foursquare** API, **geopy** library and **web-scrapping** were used to collect the data. This proof of concept shows potential and can be scaled up both horizontally and vertically.