IBM Applied Data Science Specialization Capstone Project

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

Arshiya Tripathy

Introduction

- A lot of people who must 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
- 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 travelers to choose the city based on their preferences, making their stay more comfortable/enjoyable.

Problem Statement

- 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.
- 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

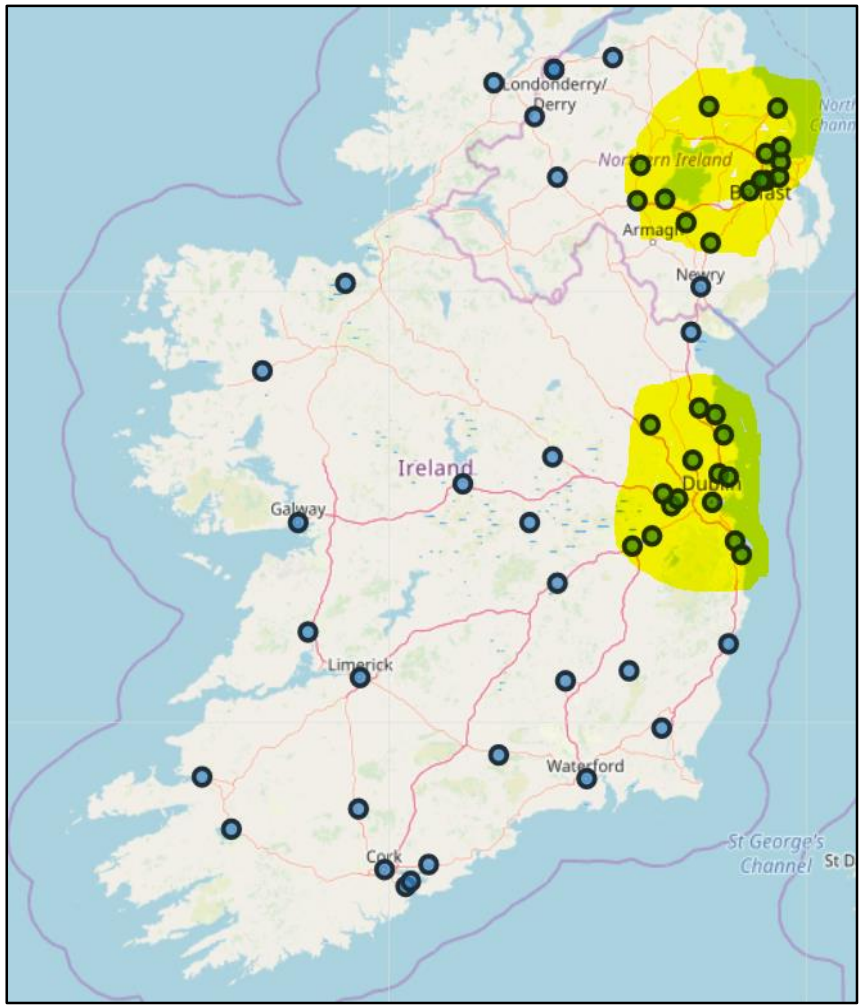
Data for analysis

- A list of 60 largest town and cities on Island of Ireland is collected by web-scrapping the Wikipedia
- Geopy python library is used to get the city centre's latitude and longitude
- Latitude and longitude are used to get Foursquare's common venues for each of the cities

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

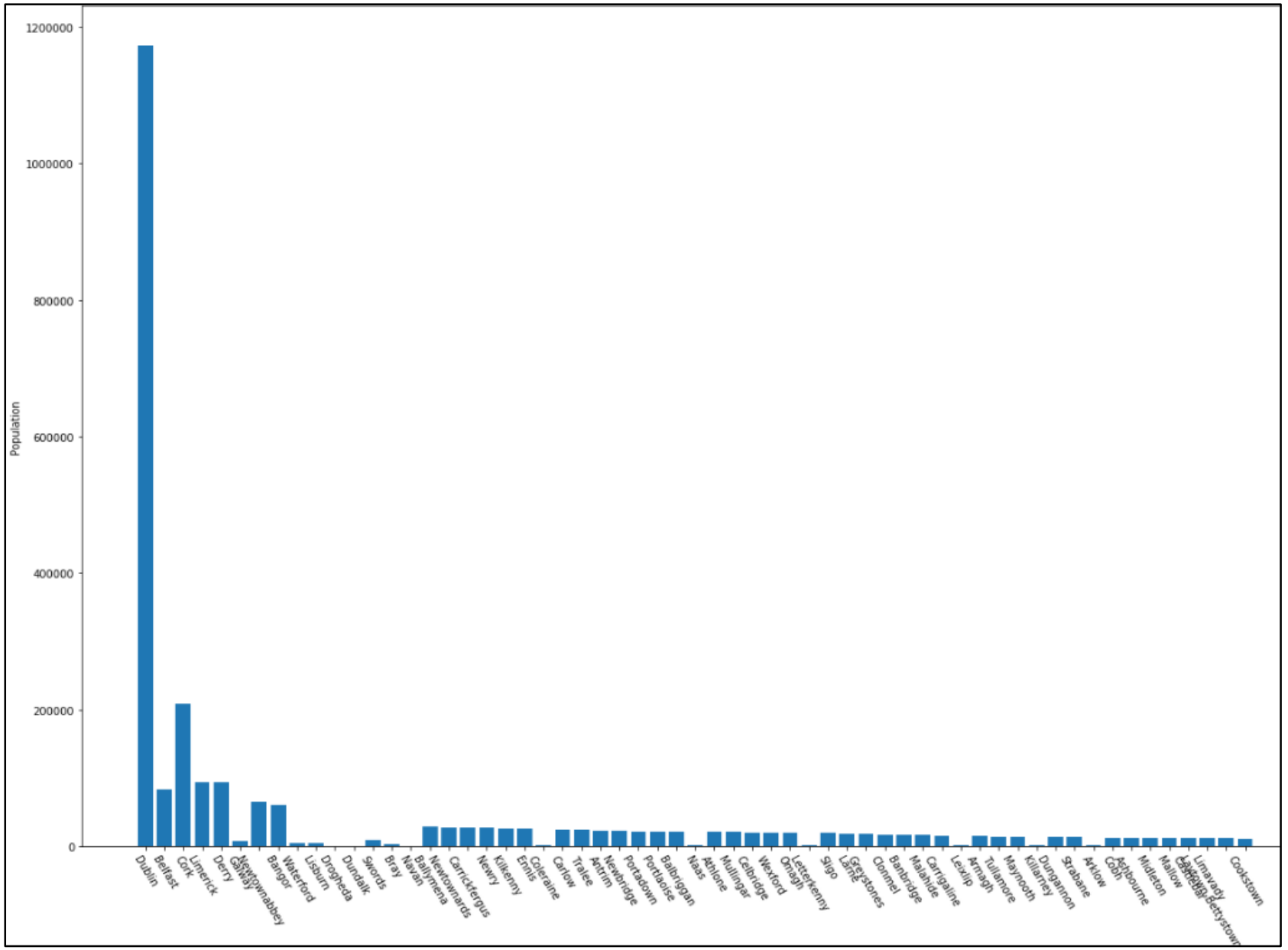
EDA

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.



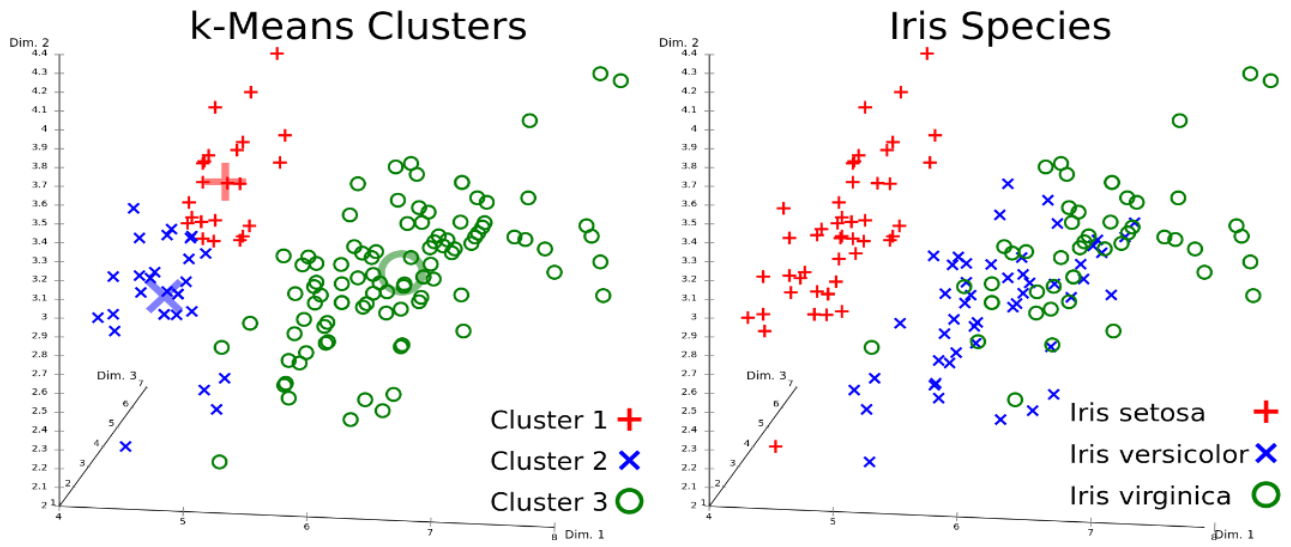
EDA

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



K-means clustering

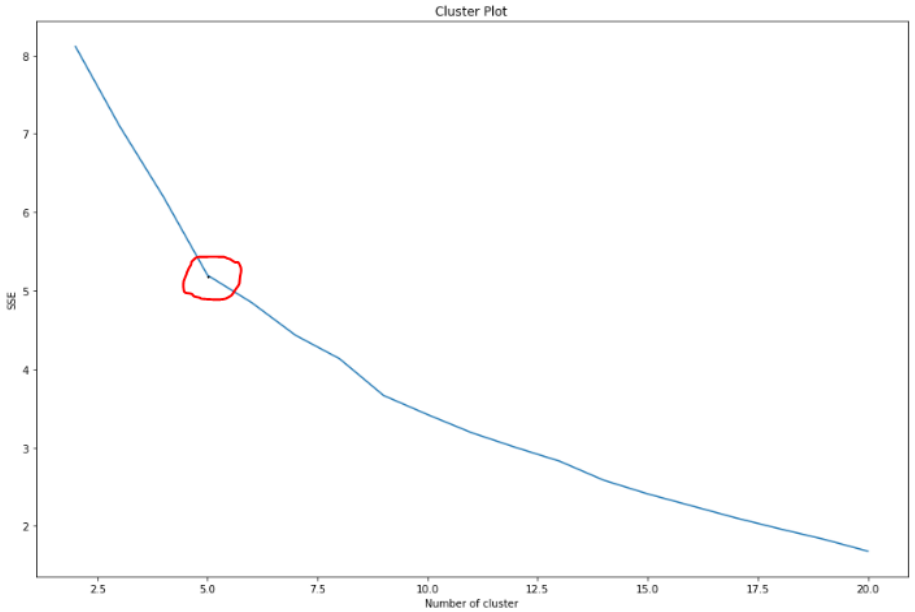
K-means clustering is an unsupervised machine learning technique which is used to mine unknown information from set of objects based on their distance (in multi-dimension space) from each other. The distance is computed based on the location of each object, which is defined by it's features. Example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



Source: https://en.wikipedia.org/wiki/K-means_clustering

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 elbow point in here is selected as the best number of cluster, i.e 5 clusters



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

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