K-Means Clustering for Travel Recommendations within Ireland

Introduction:

The country of Ireland is known for its rich history and scenic landmarks. From Trinity College and the Book of Kells to the many medieval castles, there is plenty to see. As Ireland is relatively small – roughly half the size of New York, a traveler is not limited to the city where their flight lands. They can easily travel by car across the country. However, the itinerary still needs to be prioritized based on what are the best locations to see given our areas of interest. For this study, we used Four Square to search for venues across the categories of Historic sites, Monuments, Landmarks and Memorial sites.

Background:

When planning for travel, tourists will often search Four Square. Four square provides venue recommendations based on search criteria. However, a manual search requires running multiple queries for the cities in Ireland based on each areas of interest. Given that we have selected 3 areas of interest, we would need to run 99 queries to capture the data for all cities. This project aims to simplify this process by running a Four Square search for the 33 largest cities in Ireland for our top areas of interest. We'll then cluster the results into the locations which contain the most venues of interest and provide recommendations.

The objective is for our model to recommend cities with the most venues of interest. Travelers can then plan their trips accordingly.

Interest:

Tourists want to see as much during their travels. By providing recommendations of what cities provide the most venues within their areas of interest, travelers can plan accordingly and see as much as possible during their trip.

Data:

Data Sources:

There are two sources of Data being used for this study. The first is a CSV (**IE.CSV**) containing Geocoordinate data for Ireland. This file was downloaded from **Simplemaps.com**. The raw file contains the City Name, Latitude, Longitude, County and Population for the 33 largest cities in Ireland.

The second source of data is the Four Square venues data which is accessed using the Four Squares API. The Four Square API allows for exploration of venue data by location using search or category data. It also selects venues within a given radius (in meters) and allows for the limitation of number of results.

Data Cleansing:

There was little data cleansing required for the IE.CSV file. I removed the population feature because it was not required for analysis. The City, County, Latitude and Longitude fields were kept.

The Four Square venues query returns venue data in JSON format. This data was flattened and loaded to a Dataframe. Only venues with valid categories were selected.

Feature Selection:

The following features were input into the analysis:

- City
- County
- Latitude
- Longitude
- Venue Name
- Venue Categories
- Venue Latitude
- Venue Longitude

Data Analysis

I began the analysis by plotting the cities in Ireland based on their geo-coordinates. The spread of cities across the country highlighted the challenge of planning which cities to visit:

As most flights to Ireland land in Dublin, I started analysis here. I ran a four square query four our categories of interest. I started by using a radius of 50 miles with a limit of 1000 but I found that this caused Four Square to pull in data that we were less interested in. I reduced this to 10 miles and 500 venues and the results looked much better:

Results of Four Square query for Dublin for historical sites, landmarks and monuments within 10 miles of the city. I noticed that Four Square extended my categories from three to nineteen but these all fell within our area of interest. 61 venue results were returned. I was happy to see that Four Square picked up Trinity College and Book of Kells even though it was categorized as a Library instead of Historic site.

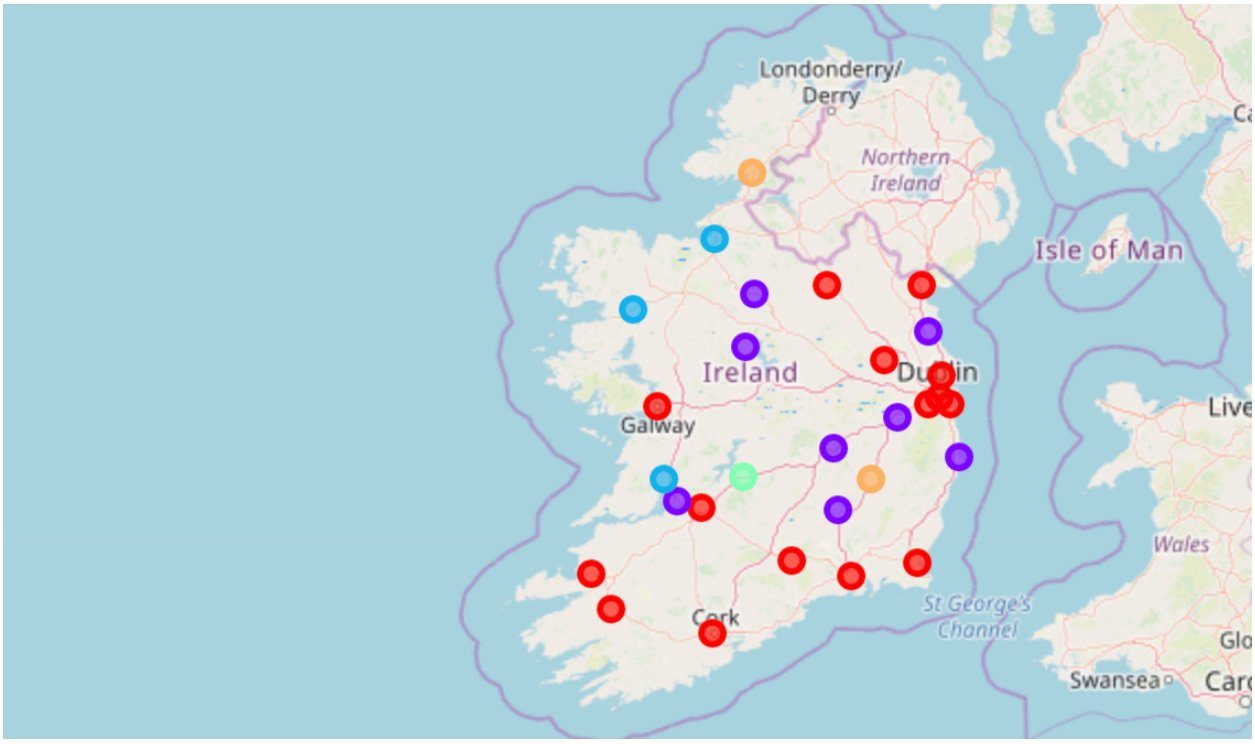
	name	categories	lat	lng
0	Old City Area	Historic Site	53.344347	-6.268818
1	Phoenix Park, Park Gate St. Entrance	Historic Site	53.350219	-6.300200
2	Trinity College Old Library & The Book of Kell...	College Library	53.343692	-6.256907
3	Marsh's Library	Historic Site	53.339037	-6.270671
4	Liffey Boardwalk	Scenic Lookout	53.346797	-6.261810
5	The Spire of Dublin / An Túr Solais (The Spire...	Monument / Landmark	53.349805	-6.260260
6	Abbey Theatre	Theater	53.348542	-6.257492
7	Dublin City Wall	Historic Site	53.344186	-6.274500
8	Kilmainham Gaol	Museum	53.341849	-6.308478
9	General Post Office (GPO)	Historic Site	53.349507	-6.260277

I then ran the same Four Square query for all 33 cities in my list. This resulted in 324 venues. The top 61 were still in Dublin but additional venues found in cities such as Cork, Limerick, Galway and Dunleavy.

Additional analysis was done to group venues by city. If a city had no venues of interest, I removed it from the list. There were 5 cities that fell into this category.

K-Means Clustering:

After one hot encoding the features, K-means clustering was run in order to see if we could cluster these results into groups of cities – i.e. which group or groups best provide the venues that we want to see. A K value of 5 was chosen. The results of the clustering are as follows:



Key:

- Cluster 1 – Red
- Cluster 2 – Purple
- Cluster 3 – Blue
- Cluster 4 – Green
- Cluster 5 – Orange

As there are still a lot of dots on this chart, I did consider reducing the K number however upon analysis of the clusters, I determined that it provided good insights into our data.

Results:

Cluster 1:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Dublin	Historic Site	Monument / Landmark	Scenic Lookout	Castle	Plaza
1	Cork	Historic Site	Monument / Landmark	Memorial Site	Capitol Building	Castle
2	Limerick	Historic Site	Monument / Landmark	Museum	Memorial Site	Capitol Building
3	Galway	Monument / Landmark	Historic Site	Memorial Site	Capitol Building	Castle
4	Waterford	Historic Site	Monument / Landmark	Park	Memorial Site	Capitol Building
6	Tralee	Monument / Landmark	Historic Site	Memorial Site	Capitol Building	Castle
9	Killarney	Monument / Landmark	Castle	Historic Site	Memorial Site	Capitol Building
17	An Cabhán	Monument / Landmark	Historic Site	Memorial Site	Capitol Building	Castle
19	Dún Dealgan	Historic Site	Monument / Landmark	Memorial Site	Capitol Building	Castle
20	Clonmel	Historic Site	Monument / Landmark	Memorial Site	Capitol Building	Castle
24	Swords	Historic Site	Monument / Landmark	Garden	Scenic Lookout	Plaza
25	Tallaght	Monument / Landmark	Historic Site	Scenic Lookout	Castle	Plaza
26	Wexford	Historic Site	Monument / Landmark	Memorial Site	Capitol Building	Castle
27	Trim	Historic Site	Monument / Landmark	Memorial Site	Capitol Building	Castle
31	Dunleary	Historic Site	Monument / Landmark	Scenic Lookout	Castle	Plaza

---

Cluster 1 contained cities with a very good representation of the venues we are looking for. The most common venue across most cities are Historic sites followed by Monuments/Landmarks. The top cities in this list are Dublin, Cork, Limerick, Galway and Waterford..

Cluster 2:

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	Drogheda	Historic Site	Trail	Memorial Site	Capitol Building	Castle
7	Kilkenny	Historic Site	Castle	Trail	Memorial Site	Capitol Building
0	Shannon	Historic Site	Trail	Memorial Site	Capitol Building	Castle
2	Ros Comáin	Historic Site	Trail	Memorial Site	Capitol Building	Castle
3	Wicklow	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Naas	Historic Site	Trail	Memorial Site	Capitol Building	Castle
5	Port Laoise	Historic Site	Memorial Site	Trail	Capitol Building	Castle
8	Carrick on Shannon	Historic Site	Trail	Memorial Site	Capitol Building	Castle

In Cluster 2 cities, Historic Sites are still the most common venue but then it moves on to Trails, Capital buildings - venues were are less interested in. A fair number of Castles on the list but as the 5th most common venue, there are not likely many in each location. Still it might be worth traveling to Cluster 2 cities if they are in route to other cities on the itinerary.

Cluster 3

Cluster 3 is a little weaker still – with historic sites coming down further in the list.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
	Drogheda	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Kilkenny	Historic Site	Castle	Trail	Memorial Site	Capitol Building
	Shannon	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Ros Comáin	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Wicklow	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Naas	Historic Site	Trail	Memorial Site	Capitol Building	Castle
	Port Laoise	Historic Site	Memorial Site	Trail	Capitol Building	Castle
	Carrick on Shannon	Historic Site	Trail	Memorial Site	Capitol Building	Castle

Cluster 4

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
30	Nenagh	Trail	Memorial Site	Capitol Building	Castle	College Library

Theres only one city in cluster 4 leading to the conclusion that it may have been worth trying different K values. On inspection however, it shows that the most common venue is a Trail - this would probably be a city to deprioritize.

Cluster 5

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
13	Donegal	Monument / Landmark	Memorial Site	Capitol Building	Castle	College Library
15	Carlow	Monument / Landmark	Memorial Site	Capitol Building	Castle	College Library

Cluster 5 has some interesting venues – no historic sites but some monuments/landmarks, memorial sites and castles. However both of these locations are far away from the others so these would likely be prioritized.

Conclusions:

Based on the data, Four Square can be used to provide good recommendations on cities with the most venues of interest. Cluster 1 is clearly the winner – with areas of interest appearing in the top 5 venues. Cluster 2 and 3 are a bit weaker, but could be worthwhile depending upon itinerary. If they are on the way – it could be worth a stop.

Of the Cluster 1 cities, Dublin, Limerick, Galway , Cork and Waterford have the most number of venues which match our desired categories.

It is worth noting, that care is needed to correctly set four square parameters. Categories are often extended to subcategories and may bring in data that we are not interested in. In addition, setting too high radius or limit parameters can result in undesired data. It took me several iterations to come up with a combination of category, radius and limit parameters to adequately meet the criteria. Failure to do so would lead to a lot of data cleansing requirements and unpredictable results.

Despite the limitations, the model provides us with a reasonable view of what cities have the most to offer in Ireland,