**Introduction**

In this Capstone project I will be identifying planning permission applications in Cork City from 2008 onwards. This data is in csv format and has been obtained from the City Council’s open data website [https://data.corkcity.ie](https://data.corkcity.ie/).

I will be looking at what neighbourhoods these applications take place in. Here the Foursquare data will help us define these neighbourhoods by the amenities available. I will also be using the supplemental data in this source to provide further useful comparisons between the segmented neighbourhoods once the clustering has been mapped.

The utility of this project lies in a number of features. On the most basic level this is will let us identify the basic neighbourhood types that attract the highest density of planning permission. Additionally the data contains a number of extra pieces of information which can be used to supplement the analysis after clustering. Primarily these applications record whether an application was successful, whether it was for one-off housing and the square footage of the application. We will be able to find the average values for each neighbourhood (or in the case of application success we can see which neighbourhoods are easiest and hardest to obtain permission for).

This information would be of significant interest to many parties. From building contractors and developers to county planners. It would help local businesses identify changing land use patterns and help banks identify which areas are going to experience significant growth.

**Data**

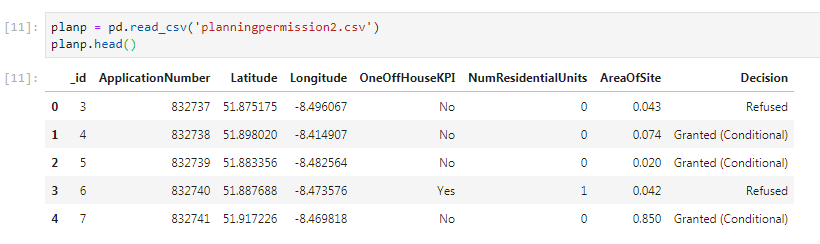
The data I am using is a csv file I obtained from <https://data.corkcity.ie>.

It is reasonably large measuring 4,975 rows by 34 columns. However of these the majority are unimportant to our investigation. As part of data preparation the following columns were kept and the rest dropped:

1. Application Number
   1. As Primary Key
2. Latitude
3. Longitude
4. OneOffHouseFlag
   1. This signals if the permission was for a one off house
5. NumberOfResidentialUnits
   1. This signifies how many homes are being built
6. AreaOfSite
   1. The total area of the property in km2
7. Decision
   1. Whether the application was accepted for rejected

The new data frame looks like this:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ApplicationNumber | Latitude | Longitude | OneOffHouseKPI | NumResidentialUnits | AreaOfSite | Decision |
| 832737 | 51.87518 | -8.49607 | No | 0 | 0.043 | Refused |
| 832738 | 51.89802 | -8.41491 | No | 0 | 0.074 | Granted (Conditional) |
| 832739 | 51.88336 | -8.48256 | No | 0 | 0.02 | Granted (Conditional) |
| 832740 | 51.88769 | -8.47358 | Yes | 1 | 0.042 | Refused |
| 832741 | 51.91723 | -8.46982 | No | 0 | 0.85 | Granted (Conditional) |

In my Jupyter notebook it looks like this

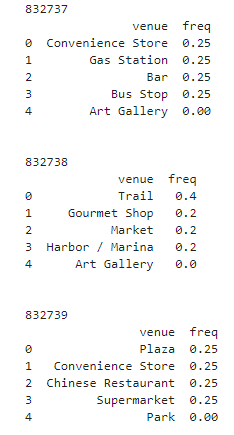
Additionally, mindful of the limits on API calls that could be made to Foursquare we adjusted the dataset to come within the call limit taking a random sample of 870 applications (roughly the calls we had left on the day).

**Methodology**

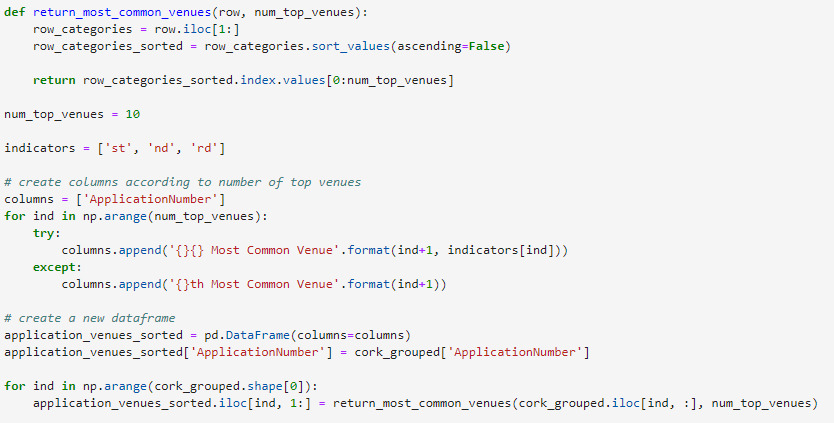
The first step was to run the various applications through the Foursquare API to assemble a list of most common nearby venues. The radius was set to 500m as this is appropriately sized for a city of Cork’s size. The limit was reduced to 20 for ease of computation.

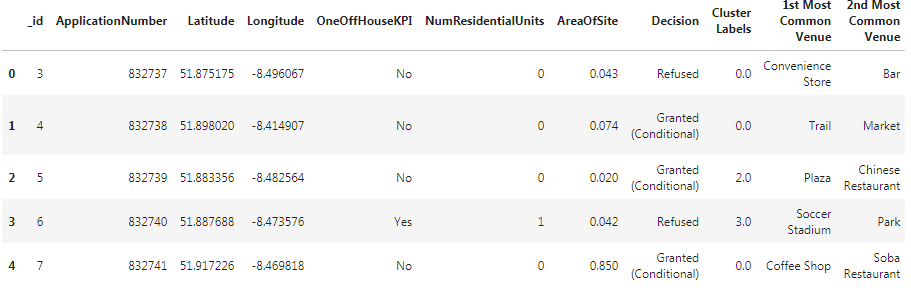
The initial API results (with one return per venue type) were first aggregated to application level with the most common venue for each application type being listed. Then these individual results were transposed onto a larger table with the application numbers serving as primary key:

This was done first through the use of one hot encoding on the venue types returned and the frequency of these dummies in each application then recorded:



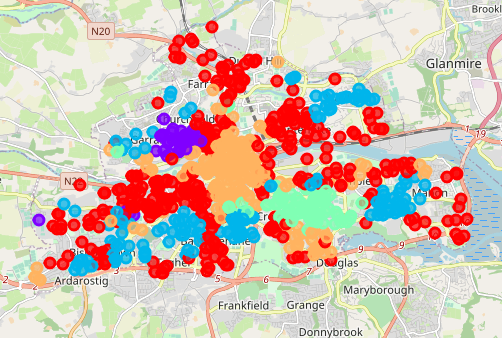
Finally these sub-categories were aggregated through a custom function:

From here we could use k means clustering to identify which neighbourhoods were the most similar to each other. We aimed for four clusters in this method and it returned the following table:



Finally we made to map the clusters using Folium and a simple line of code:

**Results**

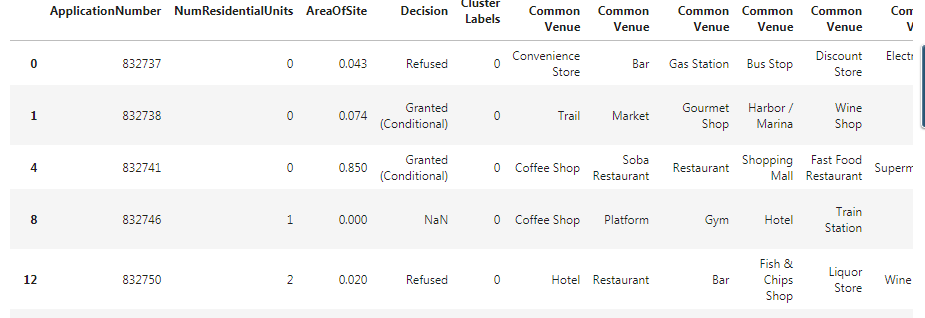
The map produced was striking:

**Discussion**

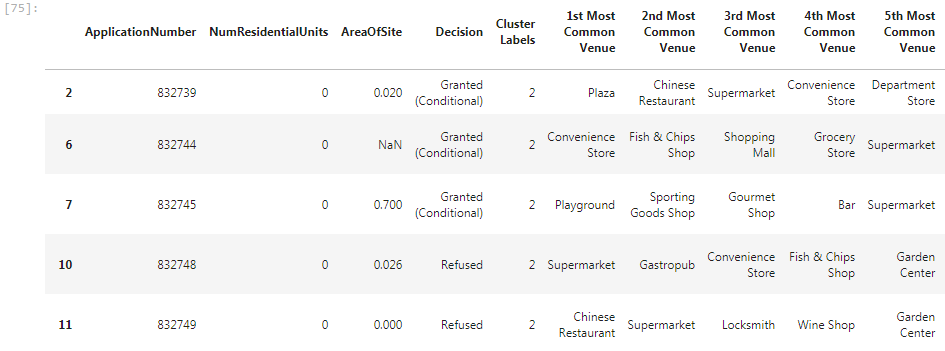
The most striking thing about the results was emergence of a clear city centre area (orange), city periphery area (red), north-west city centre (purple), south-east suburbia (green). Investigating these areas more specifically we can see that the popular venues confirm these land uses.

The city-centre area is dominated by pubs and restaurants:

The suburban area meanwhile saw more parks and sports areas:

The city periphery area has a wider array of services such as coffee shops, walking trails and even a marina:

Finally the north inner city area contains a number of fast food places and some bars:

The least apt grouping was the blue group which contains a disparate collection of resources:

**Conclusion**

This demonstrates that planning applications should distinguish strongly between suburban, city centre, inner-city areas but also that localisation can be taken too far and that many amenities and characteristics are common across urban areas (as seen in the blue areas of the map).

There is further advantage to this data with further investigation. We can see for example that suburban areas seem to see the highest rate of rejection while city centre areas see the smallest properties with applications. These simple insights can be easily expanded upon.