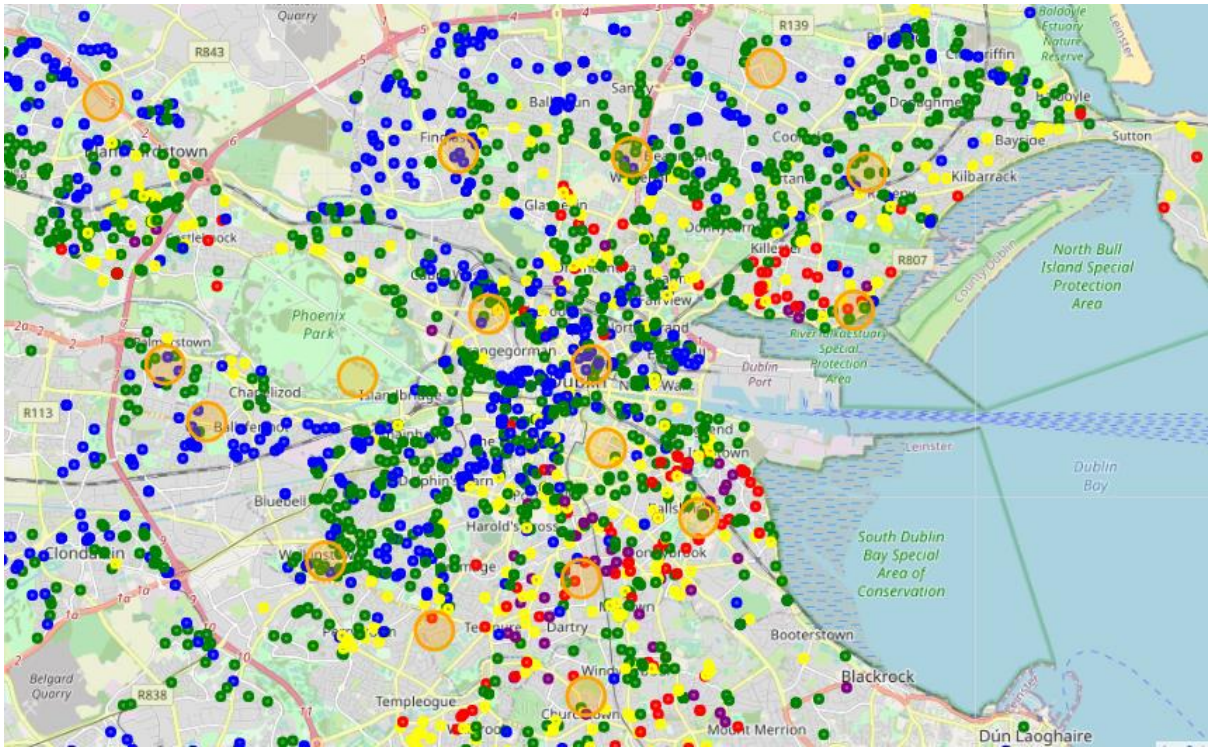
DECODING DUBLIN THROUGH DATA

Utilising k-Means Clustering and additional data analysis of geospatial data, house prices and pub ratings to determine the best place to emigrate to in Dublin, Ireland from Morningside, Johannesburg, South Africa

IBM Data Science Certification Capstone Project



Author: Keegan Moore
Management Consultant
BscEng (Ind)

Date: 17 May 2020

Contents

Table of Figures

DECODING DUBLIN THROUGH DATA

Utilising k-Means Clustering and additional data analysis of geospatial data, house prices and pub ratings to determine the best place to emigrate to in Dublin, Ireland from Morningside, Johannesburg, South Africa

IBM Data Science Certification Capstone Project

1. Introduction

1.1 Background

Dublin is the capital and largest city of the Republic of Ireland and is situated on Ireland's East coast, within the province of Leinster. Dublin is known as the technology hub of Europe, giving home to many top firms such as Google, Airbnb, Facebook and Accenture.

Being a Consultant within the Technology and Applied Intelligence field I have recently decided to emigrate from my current home in Johannesburg, South Africa to Dublin, Ireland - but much like Johannesburg, Dublin is incredibly varied amongst it's different areas (or Postal Codes) and a number of factors may change depending on where one chooses to live within Dublin including:

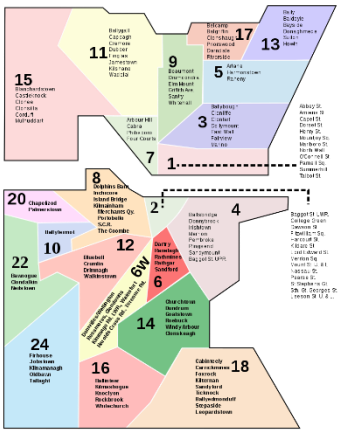
- Venues and general surrounds - the area may be more nature focussed with many parks, or perhaps have a more arts or dining focus
- Housing - the price of housing may vary wildly depending on area
- The vibe – What is Dublin without Pubs and music? An area may have lots of pubs or be a quiet residential area
- There may be many other factors that may come into play such as schools, proximity to work, malls, golf courses etc, depending on one's priorities

Dublin is made up of 18 distinct Postal Codes which make up the various areas of Dublin. These include [1]:

Table 1: List of Postal Codes in Dublin

Northside	Southside
Dublin 1 (D1)	Dublin 2 (D2)
Dublin 3 (D3)	Dublin 4 (D4), Dun Laoghaire Rathdown
Dublin 5 (D5)	Dublin 6 (D6), Dun Laoghaire Rathdown
Dublin 7 (D7)	Dublin 6W (D6W), South Dublin
Dublin 9 (D9)	Dublin 8 (D8)
Dublin 11 (D11), Fingal	Dublin 10 (D10)
Dublin 13 (D13), Fingal	Dublin 12 (D12)
Dublin 15 (D15), Fingal	Dublin 14 (D14), Dun Laoghaire Rathdown
Dublin 17 (D17), Fingal	Dublin 16 (D16), Dun Laoghaire Rathdown
	Dublin 18 (D18), Dun Laoghaire Rathdown
	Dublin 20 (D20)
	Dublin 22 (D22)
	Dublin 24 (D24)

Dublin Postal Districts



1.2 Problem

Given the fact that I don't know Dublin's various areas at all and have never even visited the country before, choosing exactly where to live is a challenging task. I do however know Johannesburg well - hence I would like my new home in Dublin to of be similar in style, and have a similar feel and surroundings to Morningside, Johannesburg, but also be within 10Km of Dublin City centre where I'll be working. Given that I am coming from South Africa, with a current average exchange rate of around R20 to the Euro, it should also be within an area that has house prices in the region of no more than €400 000 – €500 000 but cheaper if possible and preferably be an area with a great Irish vibe (ie. Have great pubs and music) so that I can experience Ireland at its best. This project will therefore aim to utilise data, and various analytical and Data Science techniques to aid in making that decision.

1.3 Interest and Value

Beyond merely giving value to myself in a personal capacity, there are approximately 23,000 emigrants [1] from South African annually, and as many as 258 million people living outside of their country of birth as of 2018 [2]. Very often these people will leave to their new country knowing very little about it, and simply 'wing it' along the way. As such, a similar approach to the one taken in this report can be taken for any location in the world.

2. Data Acquisition and Literature Review

In order to solve the problem at hand a large amount of reference, geospatial and pricing data will be required. For this I will make use of several tools, libraries and APIs, as follows:

1. Geopy Geocode Nominatim [4]: Geopy is a geospatial client for Python and other web services. Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, and landmarks across the globe and includes classes for the OpenStreetMap Nominatim, Google Geocoding API (V3), and many other geocoding services

2. Foursquare Places API [5]: Foursquare is the most trusted, independent location data platform for understanding how people move through the real world. Foursquare currently has over 60 million registered users sharing locations, venues, pictures, ratings etc around the globe. The Foursquare Places API offers real-time access to Foursquare's global database of rich venue data and user content to power location-based experiences. With a Personal developer account one can commit 99500 standard calls and 50 complex calls to the Foursquare API per day, with endpoints ranging from Venue names to user reviews.

3. PSRA - Dublin Residential Property Price Register [6]: This is openly available data published by the Property Services Regulatory Authority in Ireland. It includes Date of Sale, Price and Address of all residential properties purchased in Ireland since the 1st January 2010. In its present form it does not possess geographical coordinates, so Geopy will need to be used to enrich the data further. Note that Geopy is notoriously unstable and using Geocode on thousands of addresses (most of them with peculiar Irish names and areas) takes an incredibly long time and will also produce many null values. As such, I have included the code in my Juptyer Notebook, but for simplicity have saved the geographical coordinates to a csv file obtained from The Irish Property Price Register – Geocoded to Small Areas [7].

4. BeautifulSoup [8]: BeautifulSoup is a Python library that makes it easy to scrape information from web pages. It allows an HTML script to be parsed, iterated, searched, and modified.

3. Methodology & Exploratory Data Analysis

Given that there are three specific criteria to fulfil in order to answer the problem statement, the following methodology and analysis was used for each one:

3.1 Overview of Methodology

3.1.1 Area Clustering (finding similar areas in Dublin to Morningside, Johannesburg):

K-Means clustering of all areas in Dublin and Morningside, Johannesburg will be done in order to visually identify places in Dublin that are similar to Morningside. The K-Means clustering machine learning algorithm was chosen for this as it allows the greatest flexibility in selecting and viewing various values of K (number of clusters), as well as the fact that it is fast, robust, and gives reliable results for data sets that are distinct and well separated from each other. This suits the dataset well. The overview of the methodology to achieve this is:

- The geographical coordinates of the various Dublin postal codes will be obtained using Geopy Nominatim Geocoder. I will then manually add my desired South African address in Morningside, Johannesburg to the list of addresses so that its geographical locations are added as well.
- Next, I will calculate the distance from the centre of Dublin City (near to where I will be working) to all the Postal Code Areas using a Great Circle (or Hoversine) formula that I will define as a function and remove all areas that are further than 10km's away (except for Morningside, Johannesburg as I want to compare this later).
- Once complete, the Foursquare Places API will be used to scour the surrounding area within a 3km radius of the geographic locations of each Postal Code obtained from Geopy, and the various venues will be collated.
- The OneHotEncoding technique will then be used to create dummy variables of all venues and the frequency of venues within the area will then be calculated and sorted from most to least frequent.
- The K-Means Clustering Machine Learning algorithm will then be utilised to compare all areas to each other and create clusters of areas that are most alike. These results will be displayed on a Folium map. This will indicate which areas of the 22 areas in Dublin are most like Morningside, Johannesburg and are less than or equal to 10km from the City centre.

3.1.2 Examining Dublin's Housing Prices by Area

The K-Means clustering algorithm above, if done correctly, will produce various clusters showing the different areas of Dublin which are similar to Morningside, Johannesburg. However since I am looking for one definitive result I will need to dig deeper into other factors, such as housing prices, to narrow it down. My limit for buying a house is €500,000, and cheaper is certainly preferable. This section will attempt to a) understand the layout of Dublin and where the various low-cost and expensive areas are, and then b) rank the various areas by their average house prices to focus the decision:

- The Dublin Residential Property Price Register data will be read into Python and cleaned to remove blanks, certain unused columns and excessively high or low prices (outliers).
- The data will the be enriched with geographical coordinates so that it can be mapped (this can be done with the Geopy Geocoder but as per section 2 part 3 the coordinates have been saved to a csv file to save time so that it can be read in directly).

- Data analysis of the price will be performed in several ways including placing each house into a price 'bucket' and colour coding it so that, when mapped onto a Folium map, it is possible to see the prices of different areas and how they change.
- Finally, the average price of the houses within each area will be obtained by grouping the 3000 houses by area and then aggregating them. Once the average price is obtained the areas will be ranked from cheapest to most expensive.

3.1.3 Finding the best Vibe – or in this case, the top 10 Irish Pubs

They say that the best thing about Ireland, other than the Galway Girls and the Leprechauns, are the Irish Pubs. They are a large part of Irish culture and entertainment. As such, although this is a lower priority than sections 3.1 and 3.2, the aim is to find out a) where the top 5 Pub areas are in Dublin by looking at the average number of 'Likes' for various pubs scraped using the Foursquare API and b) find what the names are of the top 10 pubs by number of 'Likes' (as well as where to find them. Using 'Likes' is not necessarily the most accurate measure as many people don't leave 'Likes' for a venue, but it will be used as a fun exploration of the city.

- Once again Foursquare Places API will be utilised to search the areas suggested as being most like Morningside, Johannesburg in section 3.1 for venues with the keyword "Pub" within a 3km radius.
- The data will be further cleaned to ensure that irrelevant venues (such as "Public library", or 'Garden Bar') are removed by filtering on category. There may also be overlaps between areas, so duplicates will be removed, keeping the area name that is closest to the pub.
- The number of 'Likes' for each Pub will be determined. These will then be averaged and grouped by area and then ranked from most to least 'Likes'.
- A list of all pubs will be created, sorted from highest to lowest, in order to obtain a top 10 list. Their locations on a Folim map will be plotted in order to better understand where in the city they are.
-

3.4 Deciding Where to Live

The initial investigation starts with 24 potential areas in Dublin in which to live. Utilising the 10Km rule, and the clustering information from section 3.1 as a primary guide, the goal is to first narrow this number down. Second, using the housing data ranking and a budget of €500 000 at maximum (preferably cheaper) the choices will hopefully further be narrowed. Lastly, the pub rankings will hopefully reveal the areas with a great vibe to allow for the selection a specific area (and of course find the best pub in town!)

3.2 Data Analysis

3.2.1 Area Clustering

Using the names of each area within Dublin and Geopy Nominatim, the latitudes and longitudes of each area are obtained and then mapped using the Folium library in Python to get an idea of where each area is:

	Neighbourhood	Main Area	Latitude	Longitude
0	Dublin 1	Northside	53.3525	-6.25665
1	Dublin 3	Northside	53.3612	-6.18547
2	Dublin 5	Northside	53.3835	-6.18192
3	Dublin 7	Northside	53.3606	-6.28447
4	Dublin 9	Northside	53.386	-6.24558
5	Dublin 11	Northside	53.3866	-6.29263
6	Dublin 13	Northside	53.3857	-6.07538
7	Dublin 15	Northside	53.3949	-6.38921
8	Dublin 17	Northside	53.4004	-6.20949
9	Dublin 2	Southside	53.3389	-6.25271
10	Dublin 4	Southside	53.3275	-6.22749
11	Dublin 6	Southside	53.3177	-6.25953
12	Dublin 6W	Southside	53.3093	-6.29943
13	Dublin 8	Southside	53.3503	-6.32021
14	Dublin 10	Southside	53.3432	-6.36096
15	Dublin 12	Southside	53.3205	-6.32882
16	Dublin 14	Southside	53.2986	-6.2582
17	Dublin 16	Southside	53.2557	-6.29634
18	Dublin 18	Southside	53.2409	-6.18335
19	Dublin 20	Southside	53.3524	-6.37233
20	Dublin 22	Southside	53.317	-6.43852
21	Dublin 24	Southside	53.248	-6.42435
22	Panther Road, Boskruin, GA	Northside	-26.0886	27.9575

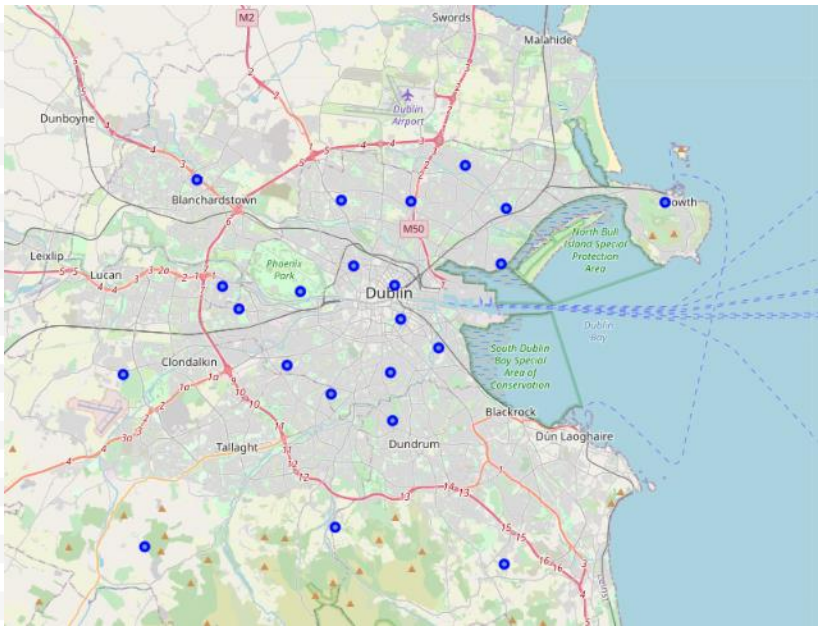


Figure 1: All Postal Codes in Dublin represented by their epicentre

As can be seen, many of the areas seem to be quite far from the city centre. Therefore, a calculation of distance to each area is required, and anywhere further than 10Km must be removed. The distance was calculated using a Hoversine formula. As such Dublin 16, Dublin 18, Dublin 22 and Dublin 24 are removed leaving the following:

```
from math import radians, cos, sin, asin, sqrt

def checkradius(lon1, lat1, lon2, lat2):
    """
    Calculate the great circle distance between two points
    on the earth (specified in decimal degrees)
    """
    # convert decimal degrees to radians
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 6371 # Radius of earth in kilometers. Use 3956 for miles
    return c * r
```

Figure 2: The hoversine function and the resulting dataset

Neighbourhood	Main Area	Latitude	Longitude	Distance
Dublin 1	Northside	53.352488	-6.256646	1.18707
Dublin 3	Northside	53.361223	-6.185467	4.03279
Dublin 5	Northside	53.383454	-6.181923	5.65706
Dublin 7	Northside	53.360551	-6.284470	3.23664
Dublin 9	Northside	53.386050	-6.245577	4.35489
Dublin 11	Northside	53.386614	-6.292627	5.56907
Dublin 17	Northside	53.400361	-6.209491	6.30182
Dublin 2	Southside	53.338940	-6.252713	1.16908
Dublin 4	Southside	53.327507	-6.227486	2.35119
Dublin 6	Southside	53.317698	-6.259525	3.47142
Dublin 6W	Southside	53.309282	-6.299435	5.69598
Dublin 8	Southside	53.350263	-6.320213	5.24824
Dublin 10	Southside	53.343217	-6.360964	7.95195
Dublin 12	Southside	53.320529	-6.328824	6.51055
Dublin 14	Southside	53.298647	-6.258201	5.48839
Dublin 20	Southside	53.352370	-6.372325	8.71501
Morningside, GA	Northside	-26.080900	28.062100	9432.02

Now that the list of potential areas is finalised, the Foursquare API is used to obtain the JSON files of all venues within a 3km radius of each of the above locations:

```
def getNearbyVenues(names, latitudes, longitudes, radius=3000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']
```

Figure 3: Part of the function used to call the Foursquare Places API

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dublin 1	53.352488	-6.256646	147 Deli	53.353410	-6.259807	Deli / Bodega
1	Dublin 1	53.352488	-6.256646	Gate Theatre	53.353113	-6.261997	Theater
2	Dublin 1	53.352488	-6.256646	Dealz	53.350623	-6.263183	Discount Store
3	Dublin 1	53.352488	-6.256646	El Grito Mexican Taqueria	53.357390	-6.256618	Mexican Restaurant
4	Dublin 1	53.352488	-6.256646	La Pausa Caffe	53.356173	-6.265484	Coffee Shop

Figure 4: The resulting dataset after pull the venue information from the Foursquare Places API

In the image above the Venue Category of each venue can be seen. 125 unique categories were found in the data. This is now turned into 'Dummy variables' using OneHotEncoding. This means that if a particular venue exists in the data it given a value of 1, otherwise it is given a value of 0. In the image below only zeros can be seen, indicating that none of those categories exist in Dublin 1.

Beach	Beer Bar	Bistro	...	Turkish Restaurant	vegetarian / Vegan Restaurant	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Yoga Studio	Zoo	Zoo Exhibit	Neighbourhood
0	0	0	...	0	0	0	0	0	0	0	0	0	Dublin 1
0	0	0	...	0	0	0	0	0	0	0	0	0	Dublin 1
0	0	0	...	0	0	0	0	0	0	0	0	0	Dublin 1
0	0	0	...	0	0	0	0	0	0	0	0	0	Dublin 1
0	0	0	...	0	0	0	0	0	0	0	0	0	Dublin 1

Figure 5: The resulting dataset after using the OneHotEncoding technique to obtain dummy variables for all categories

Now that everything is 1's and 0's, each category is grouped by Neighbourhood and the average value for each category is calculated, giving a measure of the frequency of its presence within each Neighbourhood. Now it is possible to investigate the top categories within various areas:

```
----Dublin 1----
                venue  freq
0                Café  0.10
1         Coffee Shop  0.08
2                 Pub  0.06
3  Italian Restaurant  0.04
4           Bookstore  0.04
```

```
----Morningside, GA----
                venue  freq
0               Hotel  0.12
1  Italian Restaurant  0.08
2                Café  0.08
3         Coffee Shop  0.06
4          Restaurant  0.04
```

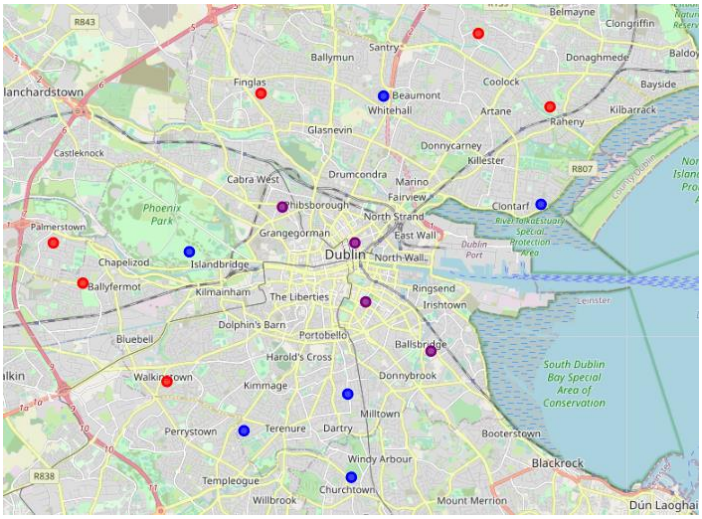
Figure 6: The top 5 venues by frequency for Dublin 10 and Morningside, Johannesburg

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Dublin 1	Café	Coffee Shop	Pub	Plaza	Bookstore	Cocktail Bar	Ice Cream Shop	Italian Restaurant	Park	Deli / Bodega
Dublin 10	Supermarket	Bar	Fast Food Restaurant	Park	Pub	Coffee Shop	Bistro	Motorcycle Shop	Burger Joint	Diner
Dublin 11	Supermarket	Park	Restaurant	Pub	Coffee Shop	Asian Restaurant	Hotel	Gastropub	Café	Fast Food Restaurant
Dublin 12	Supermarket	Pub	Park	Coffee Shop	Bar	Grocery Store	Tram Station	Shopping Mall	Motorcycle Shop	Clothing Store
Dublin 14	Park	Café	Pub	Restaurant	Clothing Store	Burger Joint	Coffee Shop	Movie Theater	Bakery	Japanese Restaurant

Figure 7: The resulting dataframe after sorting the top 10 venues for each area based on their frequency

Finally, now that we understand what kind of venues exist within each area, K-Means clustering can be used to understand which areas are most like each other. During this process the K-Means model is fit with the data as shown above (after removing the Neighbourhood name). The number of clusters has been set to 3 for this investigation, as more clusters than that produced random outliers which didn't add value to the analysis. The results of the model are shown below on Folium maps:

The results of the 3 clusters shown for the Areas in Dublin



The results of the 3 clusters shown for Morningside, Johannesburg

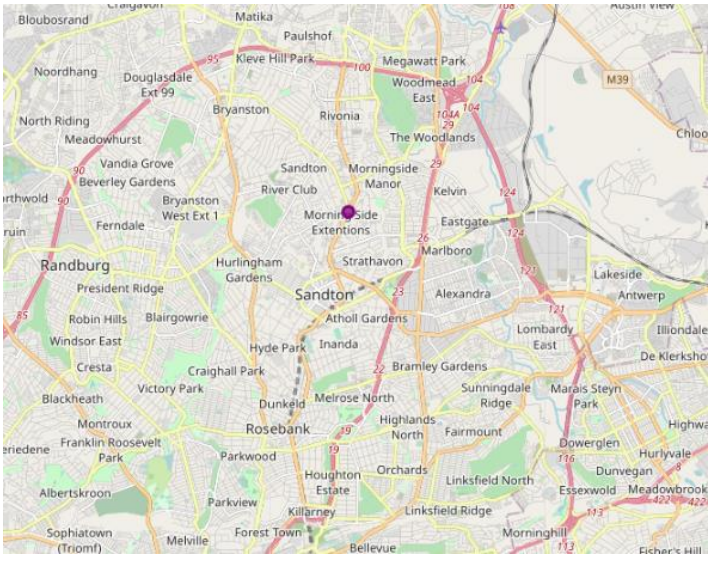


Figure 8: The result of the K-Means clustering algorithm in Dublin, represented on a Folium map

From the results of the K-Means clustering analysis Morningside, Johannesburg is most like Dublin 1, Dublin 2, Dublin 4 and Dublin 7 as they are all represented by purple dots (indicating the same cluster). Further analysis of house pricing and Pubs will hopefully narrow down this list to enable the best decision to be taken.

3.2.2 Housing Prices Analysis

During this stage of the analysis an understanding of Dublin's house prices will be gained. More than 3000 houses worth of publicly available data will be used for this analysis (PSRA - Dublin Residential Property Price Register 2017).

Similarly to section 3.2.1, this data will be read into a Pandas dataframe and Geopy Nominatim used to Geocode the locations:

address	postal_code	price	latitude	longitude
Calmount Park, Calmount Ave, Ballymount, Dubli...	Dublin 15	445475.0	53.312959	-6.345802
Alexander Court, Pembroke Street Upper, Dublin...	Dublin 2	7750000.0	53.334872	-6.254379
1 Belmont Park, Dublin 4, D04 P8E2, Ireland	Dublin 4	660000.0	53.322400	-6.238416
1 Bolbrook Ave, Tymon South, Dublin 24, D24 Y9...	Dublin 24	239000.0	53.286741	-6.341993
3 Boyne House, Custom House Square, Mayor Stre...	Dublin 1	280000.0	53.349353	-6.242606

Figure 9: Dataframe containing housing data obtained from the PSRA, enriched with geographical coordinates

To ensure that only valid data is used, all houses priced more than €2 million and less than €50,000 have been removed. To gain a better understanding of the spread of housing prices within various areas, as well as the averages, the below boxplot has been generated. From this it can be seen that Dublin 4, 14 and 16 seem to have particularly high average prices:

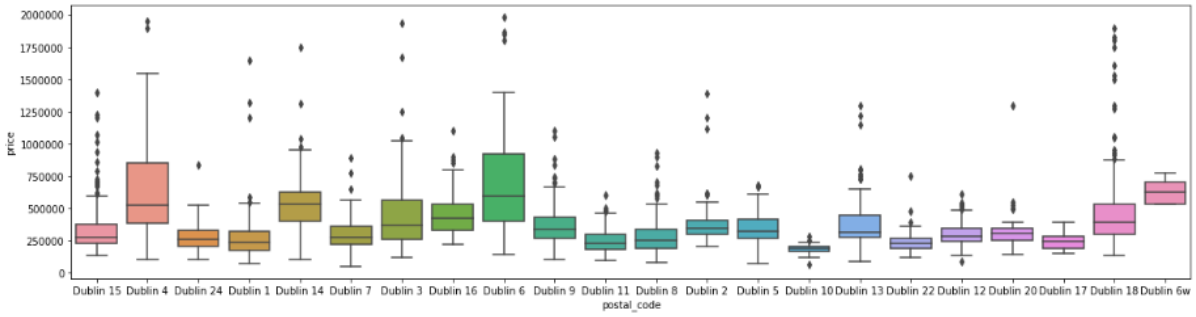


Figure 10: A boxplot showing the average, inter-quartile ranges and spread of the house prices by Area

Although data is available for houses of all kinds, it is unrealistic at present to spend more than €1 million for a house (even that is rather high, as €500,000 or less is preferable), so the data will be narrowed to show only houses that are less than €1 million in price. Now that this is complete, each house will be placed into a 'bucket' representing the price range in which it fits. The various buckets and their averages are shown in the boxplot below:

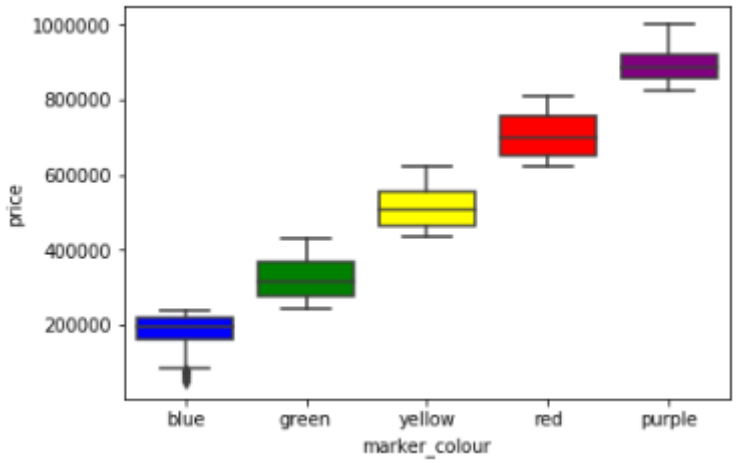


Figure 11: Boxplot showing the average price and price pread of the assigned 'Buckets'

Once each element of data is assigned a particular bucket based on its price, we are able to plot each house's location with a colour representation of the bucket in which it fits on a Folium map. We will overlay this data with pale orange circles showing the epicentres of each one of Dublin's 22 areas. This representation is displayed below. From this it can be seen that the more expensive houses seem to be on the South side of Dublin, with the exceptionally expensive houses in areas like Dublin 4 and Dublin 6, whilst the more 'middle class' houses exists in Dublin 1, Dublin 2, Dublin 7 and other areas further away from the city. Note that there are very few houses in the dataset forDublin 16 and 24. This is not a concern, however, as these areas were removed from consideration for being too far from the city centre.

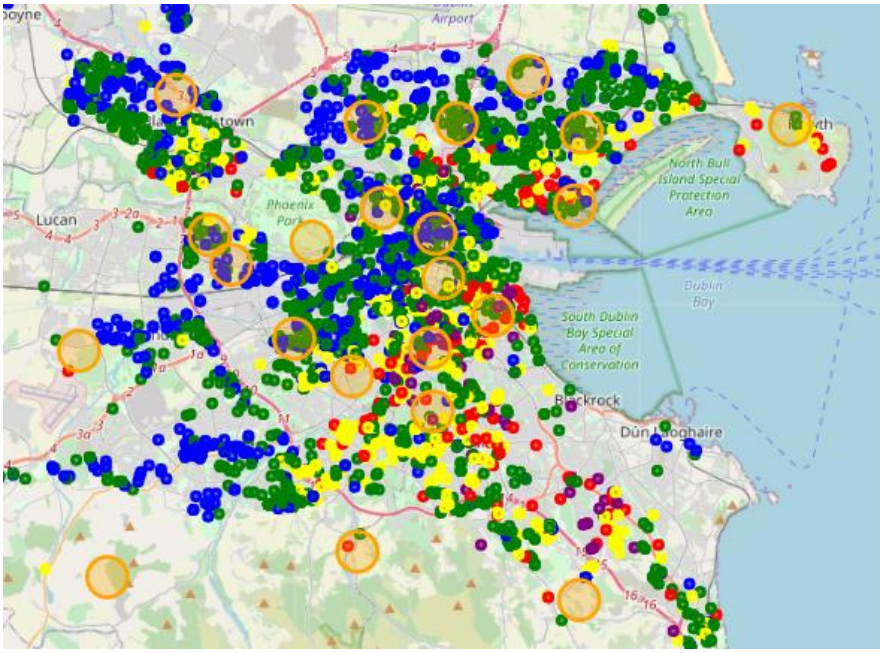


Figure 12: A folium map showing all houses under €1 million, coloured by price 'Bucket' The orange circle represent the epicentres of each one of Dublin's postal code areas

As a final method of ranking the areas by price, the below graph shows the average price of all 3000 houses sorted by Area. As can be seen, Dublin 6, 4, 6W and 14 have prices over €500,000 on average. For simplification Dublin 1,2,4 and 7 (the four areas most similar to Morningside, Gauteng) have been filtered out in the subsequent graph:

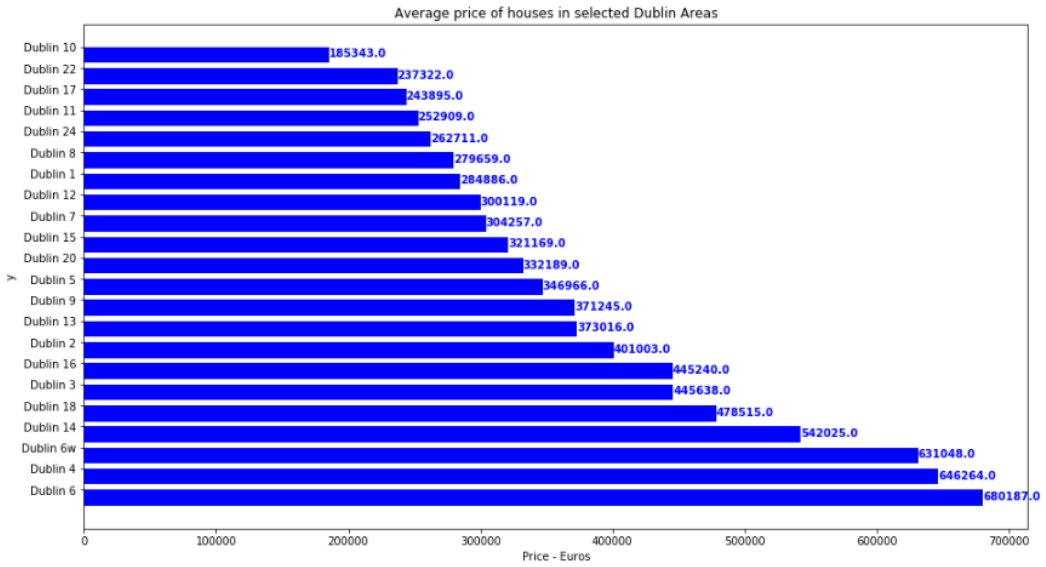


Figure 13: Bar chart showing the average price of all Dubin areas in descending order

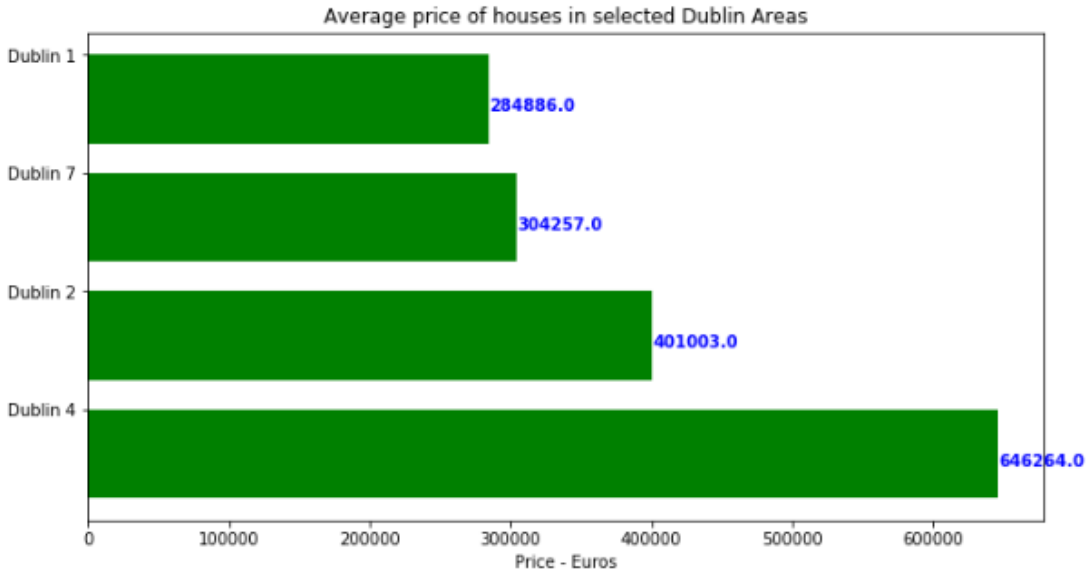


Figure 14: Bar chart showing the average price of houses in Dublin 1,2,3 and 7 in descending order

From the above graphs it can be seen that Dublin 4 will not be an optimal fit due to the fact that the average house price is far beyond €500,000 and so this option can be removed – leaving Dublin 1, 2 and 7 still in play.

3.2.3 Finding the best vibe (the top pub areas and top pubs in Dublin)

For some additional information to narrow the potential list further – and mostly just for fun – an analysis of Dublins pubs will now be done. In order to do this the location data from section 4.1 is read into a Pandas Dataframe.

Next, each location will be looped through and then searched for all pubs within a radius of 3kms using the Foursquare Places API. This will collect the venue's name, latitude, longitude, category, as well as other data. The data is then cleaned to ensure that only pubs are included in the data and not other random venues that happen to have the words 'pub' in them.

```
#check for pubs
Check = 0
for index, row in df_coords.iterrows():
    x_df = GetPubData(latitude = row['Latitude'],longitude = row['Longitude'], search_query = "pub", radius = 3000)
    x_df['neighborhood'] = df_coords.at[index , 'Neighbourhood']
    if Check != 0:
        df_PubData = df_PubData.append(x_df, sort = 'False')
    else:
        df_PubData = x_df
        df_PubData['neighborhood'] = df_coords.at[index , 'Neighbourhood']
        Check = 1
df_Pubdata = df_PubData.reset_index(drop = True)
df_PubData.head()
```

```
https://api.foursquare.com/v2/venues/search?client_id=BJAKOBU2AS1ZQPMVLIE3JQRBOFQWPZONVEQLVS5RF4TXC4KG&client_secret=2MUXVFOAOUIBESPYVU5S2N1SX
XYKTZQDQQ431YLA1WICHO4K&ll=53.3524881,-6.256645689721826&v=20180604&query=pub&radius=3000&limit=100
```

country	crossStreet	distance	formattedAddress	id	labeledLatLngs	lat	lng	name	neighborhood
Ireland	NaN	362	[72-74 Parnell St, Dublin, Dublin City, Ireland]	4f3d6c3ee4b0c30dffa53d07	[{'label': 'display', 'lat': 53.35228833905803...	53.352288	-6.262096	The Parnell Heritage Pub & Grill	Dublin 1
Ireland	NaN	718	[Temple St, Dublin, Dublin City, Ireland]	4d7a008df87b236a99ff381f	[{'label': 'display', 'lat': 53.35746058494907...	53.357461	-6.263539	The Temple Kavanaghs Pub	Dublin 1
Ireland	NaN	754	[9 College St, Dublin, Dublin City, Ireland]	4bb7b2f9b35776b00308c801	[{'label': 'display', 'lat': 53.34572589214289...	53.345726	-6.257483	Doyles Pub	Dublin 1
Ireland	NaN	1005	[41 Blessington St, Ireland]	523db48e498e8f61a1cc1ddf	[{'label': 'display', 'lat': 53.35705694381323...	53.357057	-6.269704	Blessington House PUB	Dublin 1
Ireland	NaN	267	[Ireland]	4d701010b73bb1f769c5b372	[{'label': 'display', 'lat': 53.35298688288603...	53.352987	-6.260578	Shakespeare Pub	Dublin 1

Figure 15: A snippet of code used to find all pubs pn the Foursquare API and the resulting Dataframe

The number of pubs found within the set parameters are then counted and grouped by Area and sorted in descending order, as below, to get an idea of where the most pubs are in Dublin. It was somewhat surprising that some of the areas were so low, and there may well be other venues that are not specifically called a 'Pub' or are perhaps listed as a restaurant rather than a pub or bar. Nonetheless, I shall continue with a focus on venues specifically designated as 'Pubs' – after all, what is a true Irish Pub if it is not actually called a 'Pub'?

```
Dublin 7           35
Dublin 1           34
Dublin 2           32
Dublin 4           30
Dublin 6           29
Dublin 8           23
Dublin 14          16
Dublin 11          12
Dublin 6W          12
Dublin 9           11
Dublin 12           8
Dublin 10           7
Dublin 5            6
Dublin 17           6
Dublin 3            6
Dublin 20           5
Morningside, GA     2
```

Figure 16: The number of pubs found within a 3Km radius in each area, sorted

Although we have the number of pubs within the various areas, we do not yet know where the best pubs are. For this the number of 'Likes' for each venue will be extracted from the Foursquare Places API (if the venue has any 'Likes'). After data cleaning, the 'Likes' are then sorted in descending order, as below, to find the top 10 rated pubs. These have been displayed on a Folium map for better visualisation of where they are:

```
for index, rows in df_Likes.iterrows():
    venue_id = rows['id']
    url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(venue_id, CLIENT_ID, CLIENT_SECRET, VERSION)
    result = requests.get(url).json()

    try:
        df_Likes.at[index , 'Likes'] = (result['response']['venue']['likes']['summary'])
    except:
        df_Likes.at[index , 'Likes'] = 0
```

name	neighborhood	categories	distance	lat	lng	id	Likes
The Brazen Head	Dublin 8	Irish Pub	2974	53.344982	-6.276335	4ade0eedf964a520747021e3	787
The Bernard Shaw	Dublin 14	Pub	3702	53.331712	-6.264264	4ade0f0cf964a520f07021e3	450
O'Donoghue's	Dublin 2	Pub	118	53.338314	-6.254162	4ade0eeff964a5207e7021e3	318
Kehoe's	Dublin 4	Pub	2610	53.341143	-6.259437	4bd42c04046076b027707771	236
The Barge	Dublin 14	Pub	3548	53.330493	-6.260525	4b044d87f964a5201b5322e3	176
Toners Pub	Dublin 2	Pub	124	53.337838	-6.252410	4b05d950f964a52056e422e3	151
Arthur's Pub	Dublin 1	Pub	1926	53.343260	-6.281174	4d81270ddbc5f04dc25405b7	145
Doyles Pub	Dublin 1	Pub	754	53.345726	-6.257483	4bb7b2f9b35776b00308c801	95
Slattery's	Dublin 4	Pub	1174	53.337183	-6.234549	4b52519df964a520f37627e3	86
John Kavanagh's The Gravediggers	Dublin 11	Pub	2345	53.369486	-6.272062	4ade0ef5f964a520947021e3	63

Figure 17: Snippet of code to get the numbe of 'Likes' per Pub from the Foursquare API (if any) and the resulting Dataframe

The top 10 pubs, by Likes, are shown below with their geographic locations:

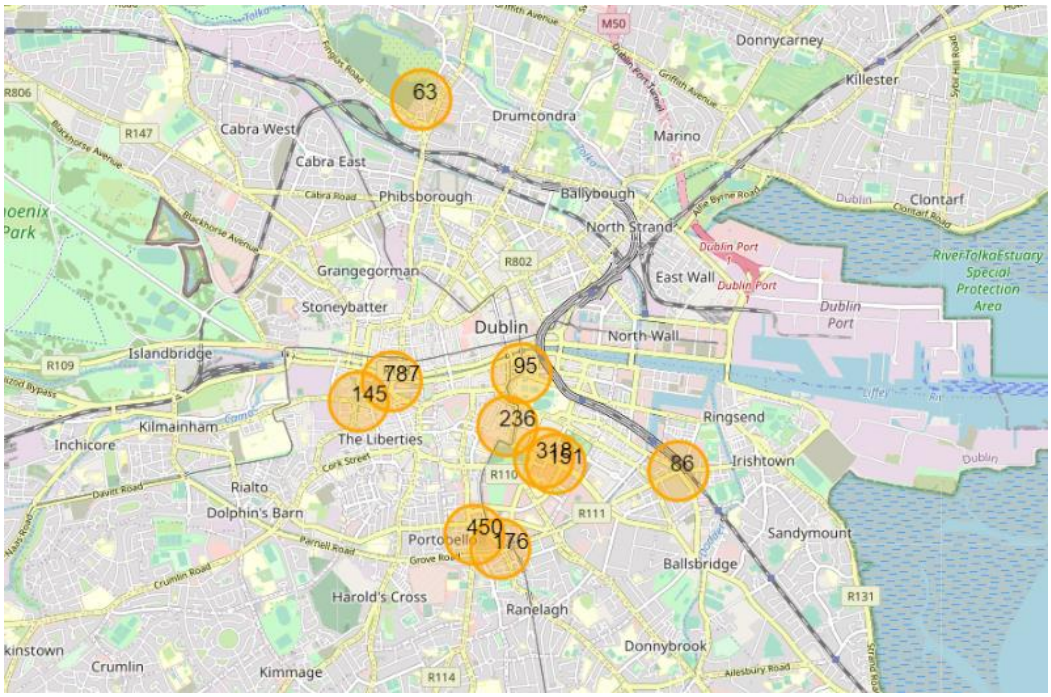


Figure 18: A Folium map showing the top 10 pubs by number of 'Likes'

Finally, to get a better understanding of where the top 5 pub areas are, the number of ‘Likes’ for each area will be aggregated and sorted from highest to lowest, as displayed below:

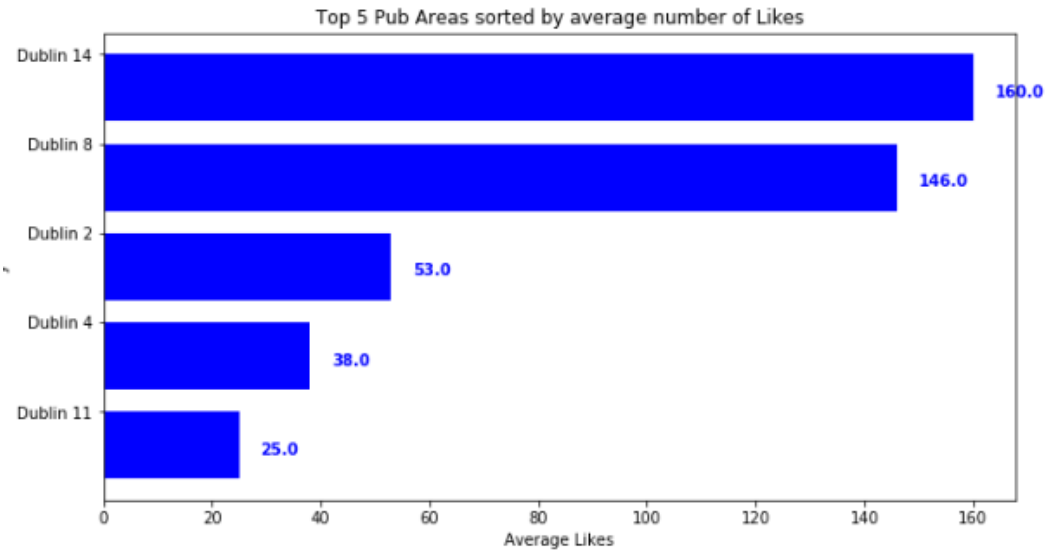


Figure 19: Bar graph showing the top 5 areas by average number of 'Likes' sorted in descending order

4. Results

- All areas considered in Dublin are within 10kms of the city centre, except for Dublin 16, 18, 20 and 22.
- After compiling the data and fitting it to the K-Means clustering machine learning algorithm it was revealed that Dublin 1,2,4 and 7 are most like Morningside, Johannesburg, giving a far more targeted list of areas for further consideration. From the analysis it seems that the style of the city (venues and surrounds) begin to change approximately 4Kms from the city centre, and this changes even more so approximately 6.5Kms outside of the city.
- The analysis of the Dublin Housing data explained a lot about what Dublin's financial structure is as a region, where the expensive areas are versus the middle-class areas. It revealed that Dublin 6,4, 6W and 14 had the highest average prices for houses. Dublin 4, being one of the potential regions for consideration, had an average house price of €646,000 which is substantial, whilst Dublin 1,2 and 7 were well below €500,000 at €284 000, €304 000 and €400 000 respectively. Dublin 13 was of particular interest as well, being far from the city, in a cove surrounded by ocean and yet not too pricey – this may prove to be a potential spot for a small holiday home or simply somewhere to find a nice Airbnb for the weekend.
- The Pub exploration (ie. Searching for an area with the best vibe) revealed a number of interesting things. First, that top areas for Irish Pubs according to the average number of 'Likes' is Dublin 14, 8, 2 ,4 and 11 respectively. Dublin 2 and Dublin 4 are the only areas in the top 5 that are also suggested in the clustering analysis.
- Of the suggested areas in the clustering analysis, none of them are far from the city centre with the nearest being Dublin 1 and 2 (1.2 Kms away) and the furthest being Dublin 7 (3.23 Km away)

5. Discussion

The investigation, utilising various Data Science tools and techniques, was certainly able to answer the problem statement sufficiently well. It was able to accurately determine areas that are similar to Morningside (this was verified after the fact with a few Google searches of Dublin 2, it's areas, descriptions and surroundings). Furthermore, it was simply and elegantly able to visualise data that showed Dublin's housing prices and best pubs in a manner that is easy to understand and follow for someone who has never been to Ireland before. With that being said, there are certainly limitation to this analysis and areas for further improvement. Some of these limitations and future considerations include:

- Clustering: The number of clusters (K) that are chosen will affect the sensitivity of the analysis and thus selecting various values for K will change the results.
- Radius and Limit: When using the Foursquare API a radius and limit must be set. By increasing or decreasing the radius and the limit on the number of venues that are accepted, the results could vary significantly as new venues are compared and taken into account. During this investigation a radius of 3km was chosen and the maximum limit used was 100 so as not to max out the number of calls available to the API per day and also to be time efficient. By increasing the limit, we may be able to get a better representation of an area in future, although I would not recommend increasing the radius as an areas' feel, style and layout can change significantly over even 1 or 2 kilometers.

- Overlapping: Because each area is a different size there may be some overlapping in the radius' during analysis. Whilst this isn't perfect, the nearest venues will be selected first by the Foursquare API, so this doesn't have too much of an effect on the overall result given a relatively low limit.
- Data cleansing: Much of the data, particularly during the Housing investigation, is not perfect and as such many rows had to be removed or data replaced. This may affect the results somewhat.
- Foursquare limitations: Depending on the licence that one has, the number of calls to the Foursquare API available per day may be limiting and may dictate the parameters of one's investigation. For most investigations however, with a Personal account, this shouldn't matter.
- Searching for Venues and their 'Likes': When searching for venue types on Foursquare a keyword is used, such as 'Pub'. This will return many venues that are correct but may miss some that are designated differently by category, such as 'restaurant', even if they do technically have a Pub. Furthermore, when scraping the 'Likes' for each venue, this assumes that people have indeed taken the time to 'Like' the venue on the Foursquare app, and not on some other platform instead.

6. Conclusion

- After removing all areas further than 10Kms (Dublin 16,18,20 and 22) left 18 areas for consideration.
- The clustering analysis revealed that Dublin 1,2,4 and 7 were most like Morningside, Johannesburg. With this being a key requirement, it thus narrowed the search considerably.
- The house pricing analysis revealed Dublin 4 average house prices to be substantially higher than the required maximum of €500,000, leaving Dublin 1, 2 and 7 for consideration.
- Finally, the Pub exploration showed that, over the remaining contenders, Dublin 2 was in the top 5 pub areas (at number 3). Dublin 2 further sports the 3rd and 6th best overall Pubs by 'Likes' – O'Donoghues and Toner's Pub.
- With Dublin 2 also only approximately 1.2Kms away from the Dublin city centre it would mean less transportation costs. This, along with the vibe of the area, justify the higher average house prices than Dublin 1 and 7, as the house would likely maintain its value in future.
- Given all the above, Dublin 2 seems the most logic place to make my move – although I will definitely be visiting the top pub, The Brazen Head in Dublin 8, shortly after arriving.

References

[1] Wikipedia. (2020, March 3). *List of Dublin postal districts*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_Dublin_postal_districts

[2] Buckham, D. (2019, November 25). *Emigration – what are the facts?* Retrieved from Business News: https://www.cbn.co.za/featured/emigration-what-are-the-facts/

[3] Hill, A. (2018, September 10). *Migration: how many people are on the move around the world?* Retrieved from The Guardian: https://www.theguardian.com/news/2018/sep/10/migration-how-many-people-are-on-the-move-around-the-world

[4] Esmukov, K. (2018). *Geopy*. Retrieved from Geopy: https://geopy.readthedocs.io/en/stable/

[5] Foursquare. (2020). Retrieved from Foursquare: https://foursquare.com/

[6] Dept of Public Expenditure and Reform, I. (2017). *PSRA - Dublin Residential Property Price Register*. Retrieved from Data.Gov.IE: https://data.gov.ie/dataset/property-price-register

[7] Lynn, S. (2017). *The Irish Property Price Register – Geocoded to Small Areas*. Retrieved from https://www.shanelynn.ie/the-irish-property-price-register-geocoded-to-small-areas/

[8] Foundation, P. S. (2020). *beautifulsoup4 4.9.0*. Retrieved from PyPi: https://pypi.org/project/beautifulsoup4/