Battle of the Neighbourhoods
New York vs Dublin

Coursera IBM Applied Data Science Capstone

Warren Shelley

31st of October 2020

Dublin Image (https://unsplash.com/photos/XI1yvtqKjl8)
New York Image (Photo by Quintin Gellar from Pexels)

Introduction

New York city is one of the most diverse and multicultural urban metropolises in USA and the world. It is the largest city in the USA with a population of approximately 8.3 million[1]. As a centre of multiculturalism, New York City has been a gateway for immigration to the USA throughout the years. Today it is the most linguistically diverse city in the world[2] that is home to 3.2 million immigrants[3].

Irish immigrants are historically one of the largest of these immigrant populations. This large pattern of immigration began during the Great Irish Famine of the 1840s and continued through much of the 20th century. As a result Irish culture has had a significant impact of the development of neighbourhoods within New York City.

Dublin is the largest city in Ireland and it is one of the hubs of Irish cultural. A comparison of the types of venues in these cities may reveal the similarity between neighbourhoods in the two cities.

Problem

A large company with its HQ in New York City has plans to open a European HQ in the centre of Dublin city. As one of several new employee of this company I have been assigned to a new role in the European HQ, but will have to move to New York for several months for job specific training. In order to prepare for the move I will investigate which neighbourhoods of New York are most similar to those in Dublin. Next I will investigate the average rent price for a one bedroom property in each of these neighbourhoods. This will be presented to the HR department of the company so that they can provide guidance to other employees making a transition from Dublin to New York. The main goal of this analysis is to find the most affordable New York neighbourhood which is similar to neighbourhoods in Dublin, so that those moving to New York for training purposes have a smooth as possible transition.

Data

To tackle this problem the following data sources are required:

- New York City data containing neighbourhood and borough information.
- New York venue information.
- Dublin data containing neighbourhood and borough information.
- Dublin venue data.
- New York average rent price per neighbourhood data.
- Dublin average rent price per neighbourhood data.

The data will be sourced from multiple open source sources. The New York neighbourhood information was extracted from the NYU dataset JSON file (https://geo.nyu.edu/catalog/nyu_2451_34572). The Dublin neighbourhood dataset was built by using the python BeautifulSoup library to web scrape the neighbourhood information from the list of Dublin postal districts as found on Wikipedia (https://en.wikipedia.org/wiki/List_of_Dublin_postal_districts), and reformatting into a

structured format. The geographical coordinates of neighbourhoods was found using the Python Geocoder package.

The venue information for each neighbourhood is retrieved using the Foursquare API. The Foursquare API has one of the largest databases of 105+ million places and is used by over 125,000 developers.

The rent data will be extracted from (https://streeteasy.com) which provides rental data for each neighbourhood in New York. The Dublin rental data is acquired from the Irish Central Statistics Office annual Rent Tenancy Board report.

Methodology

First the New York data was extracted from the JSON file and added to a pandas dataframe. The dataframe was filtered to only include neighbourhoods on the island of Manhattan. The Folium library was used to plot the location of each of these venues overlaid on a map of New York, as shown in Figure 1.

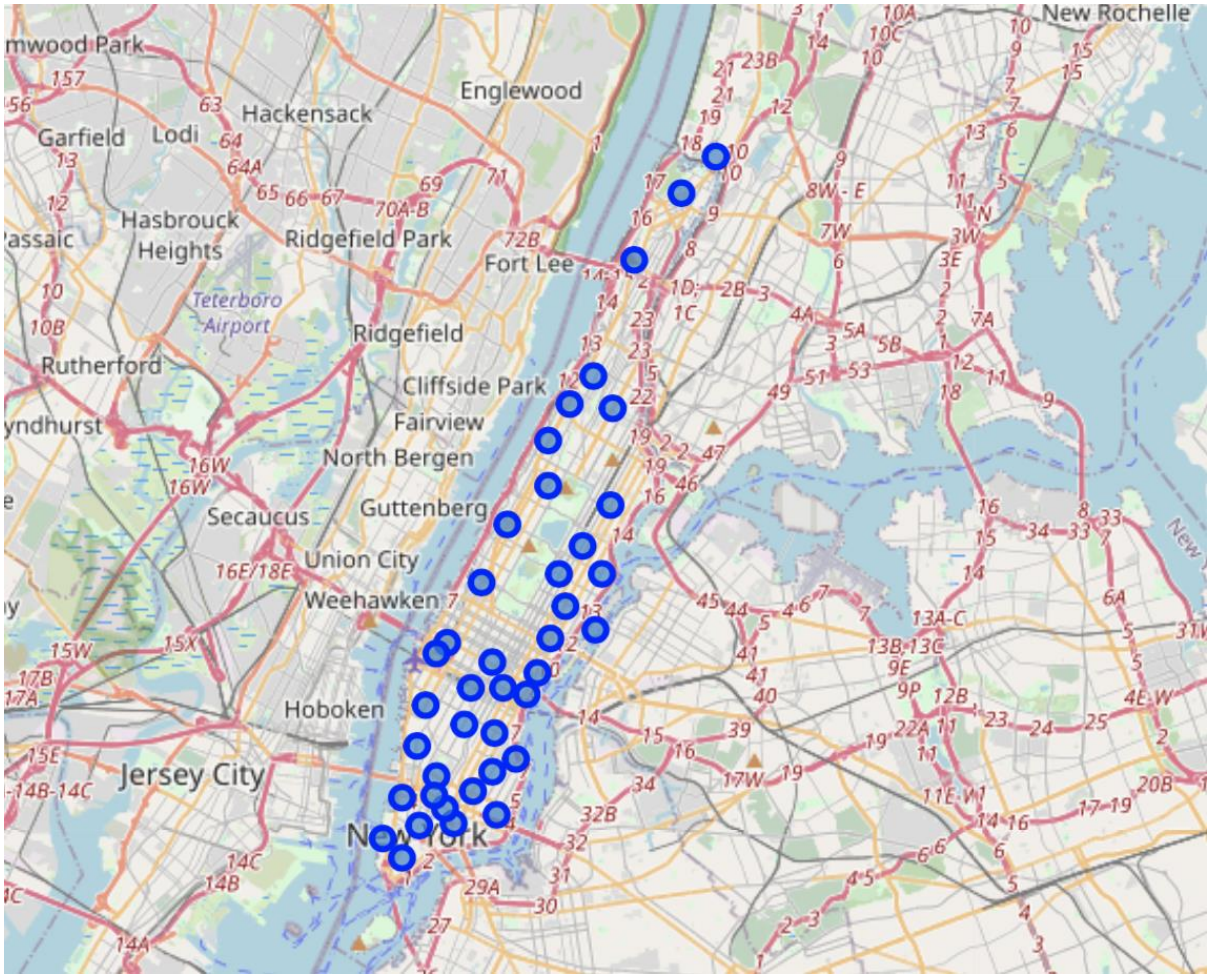

Figure 1. Folium map of NYC neighbourhoods.

The Dublin neighbourhood information was scrapped from the Wikipedia page cited above and read into a pandas dataframe. This data set required some cleaning before it was suitable for analysis. First the rows with duplicate postal codes were combined together into a single row and python geocoder was used to determine the latitude and longitude of each neighbourhood. Once again the Folium library was used to overlay the Dublin neighbourhood locations on a map of Dublin as in Figure 2.

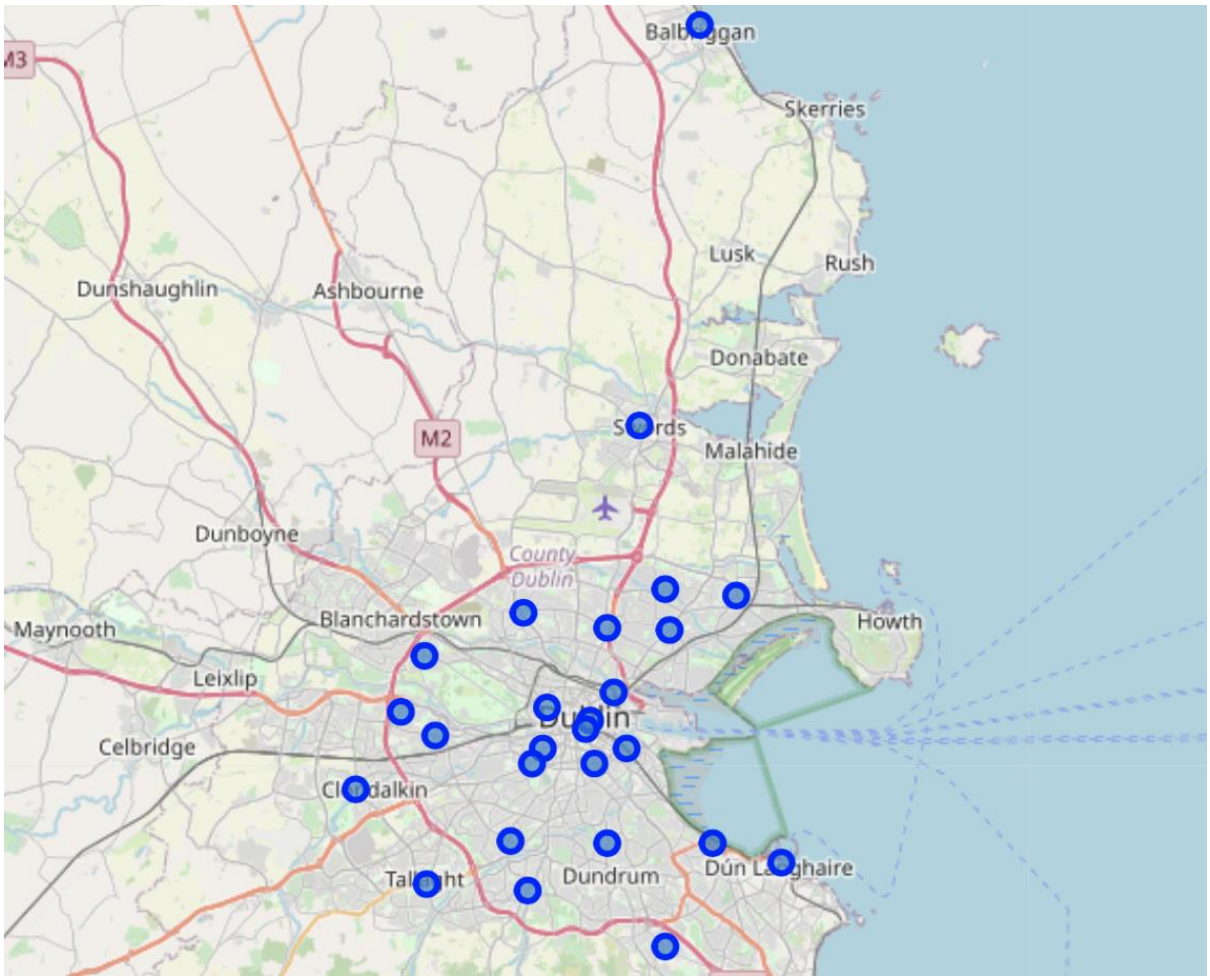

Figure 2. Folium map of Dublin Neighbourhoods

The average rent price for a one bedroom property in each neighbourhood was determined. The dataset for each city was structured in a different manner, thus some data cleaning was required. It was decided that 2018 Q3 rent prices would be used as October 2018 was the most recent accessible open source data on New York City rent prices. The monthly rent price breakdown was aggregated to determine the average quarterly rent price for each neighbourhood. An exchange rate of $1 to €0.86, as was the case in Q3 2018, was used to convert New York rent prices to euro, allowing for a direct comparison between the two cities.

For the Dublin rent data, a number of neighbourhoods which fall under the same postal code had their own separate entry in the data table. In order to make this compatible with the earlier neighbourhood data these rows were aggregated to get a mean rent price for each neighbourhood.

In order to determine the similarity between neighbourhoods, the Foursquare API was used to extract the venue information of each neighbourhood. Once the venue information was obtained, one hot encoding was used to calculate the mean frequency that each venue occurs in each neighbourhood. This will provide insight as to what venues are most common in each neighbourhood, and will provide the basis upon which the neighbourhoods are clustered.

K-means clustering was used in order to cluster the neighbourhoods based on those with similar venues. This method can quickly cluster neighbourhoods into a specified number of clusters based on the desired features, which is ideal for this analysis. The optimum number of clusters was found to be seven clusters. In order to cluster the neighbourhoods across the two cities, the two frequency tables were concatenated into one large data table and the clustering was performed across the entire set of neighbourhoods. The results of this clustering will be presented in the following section

Results and Discussion

The newly clustered neighbourhoods were colour coded and over plot on a map of New York as shown in Figure 3. There are three primary types of neighbourhood in New York. The vast majority of neighbourhoods fall into cluster 6 (orange), with seven neighbourhoods falling into cluster 0 (red) and three neighbourhoods labelled as cluster 1 (purple). The key now is to compare these neighbourhoods with the cluster labels assigned to the Dublin neighbourhoods.

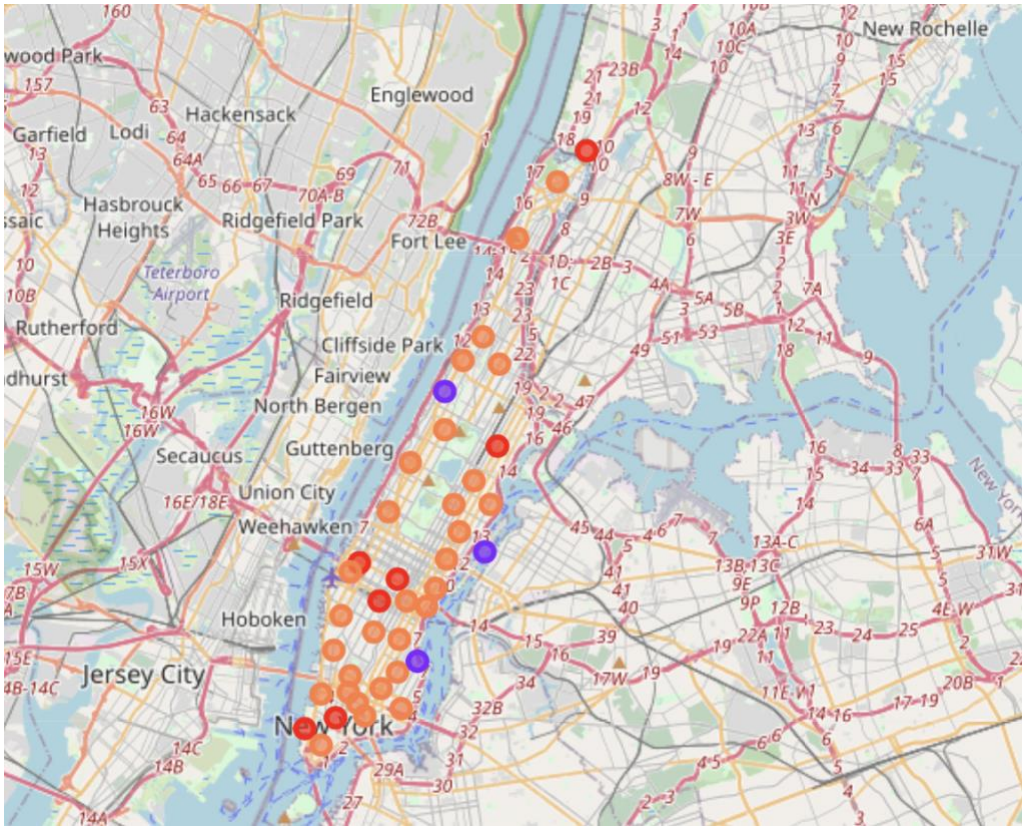

Figure 3. Cluster labelled New York neighbourhoods.

Figure 4, presents the labelled Dublin neighbourhoods. It is clear from this figure that Dublin's neighbourhoods have greater diversity than those in New York, with there being six distinct clusters in Dublin, compared to three in New York. Most neighbourhoods fall into cluster 3 (cyan) or cluster 5 (green). The key neighbourhoods of interest are those with labels of cluster 0 (red) and cluster 1 (purple) which are the neighbourhoods that share similarities with the New York neighbourhoods.

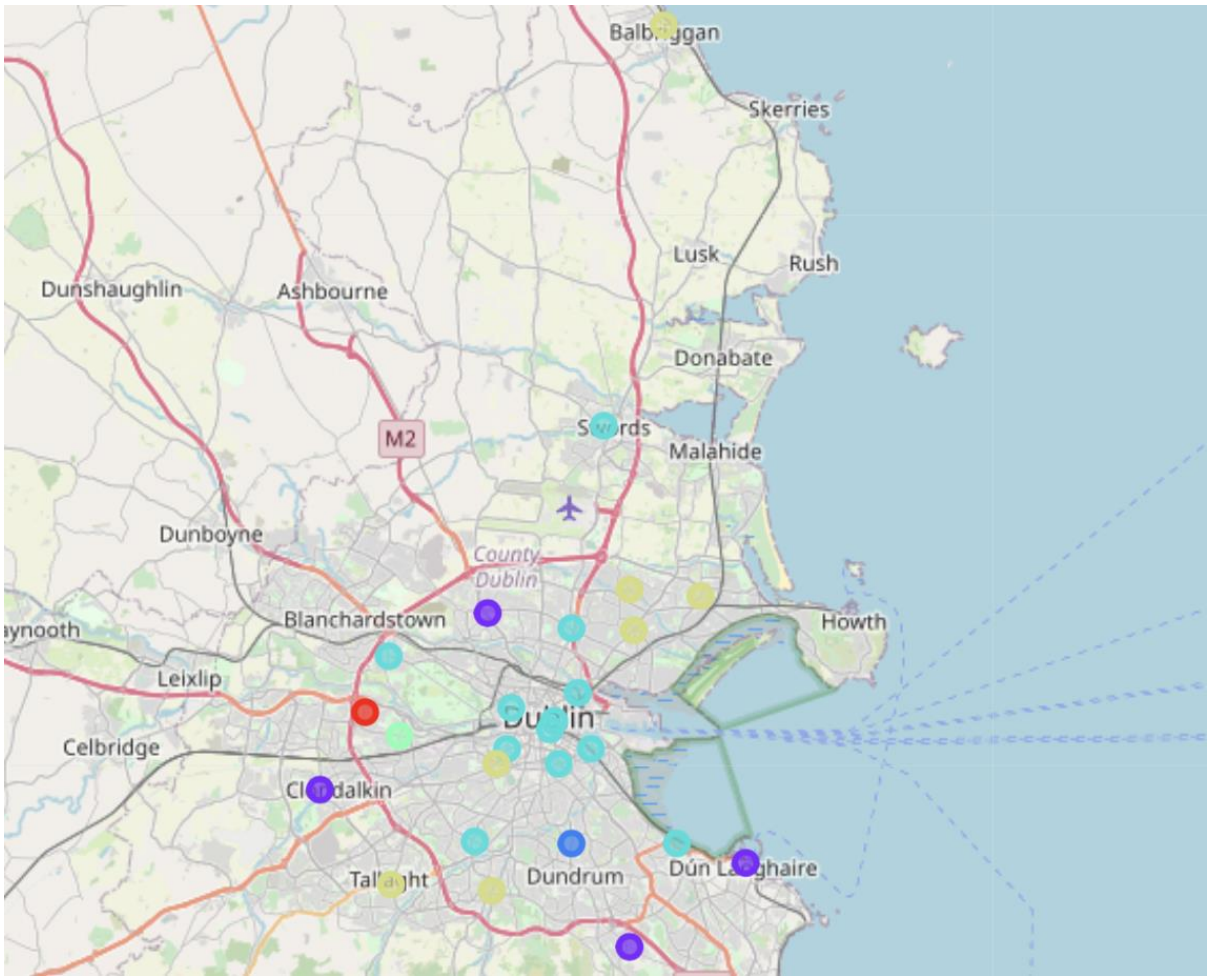

Figure 4. Cluster labelled Dublin neighbourhoods.

From this analysis, the Dublin neighbourhoods of Clondalkin, Dun Laoghaire, Sandyford and Finglas has venues most similar to those found in the New York neighbourhoods of Morningside Heights, Roosevelt Island and Stuyvesant Town. These neighbourhoods lie on the outskirts of Dublin city and at the periphery of Manhattan Island. Additionally the Dublin neighbourhood of Palmerstown has venues most similar to the New York neighbourhoods of Clinton, East Harlem, Battery Park, Civic center, Midtown south, and Midtown.

Now that a number of similar neighbourhoods have been identified, the rent prices of a one bedroom property in each of these neighbourhoods shall be determined. A small number of neighbourhoods did not have any associated rent information in the rental datasets, these neighbourhoods were dropped from further analysis. First, the rent prices from the New York neighbourhoods will be analysed as shown in Figure 5. Rent prices in Manhattan range from between €1519 in Inwood to €3953 in Tribeca. Both of these are labelled as cluster 6, which was the most common cluster type in

Manhattan. This cluster had the broadest range of prices compared to all the clusters analysed. For cluster 0, the prices ranged from €1773 in East Harlem to €3427 in Midtown South. It appears that East Harlem is the outlier in this group, as all the other neighbourhoods assigned to cluster 0 have an average rent price greater than €3200. The two remaining neighbourhoods in cluster 1 have a price of €2329 in Morningside Heights and €2506 in Roosevelt Island.

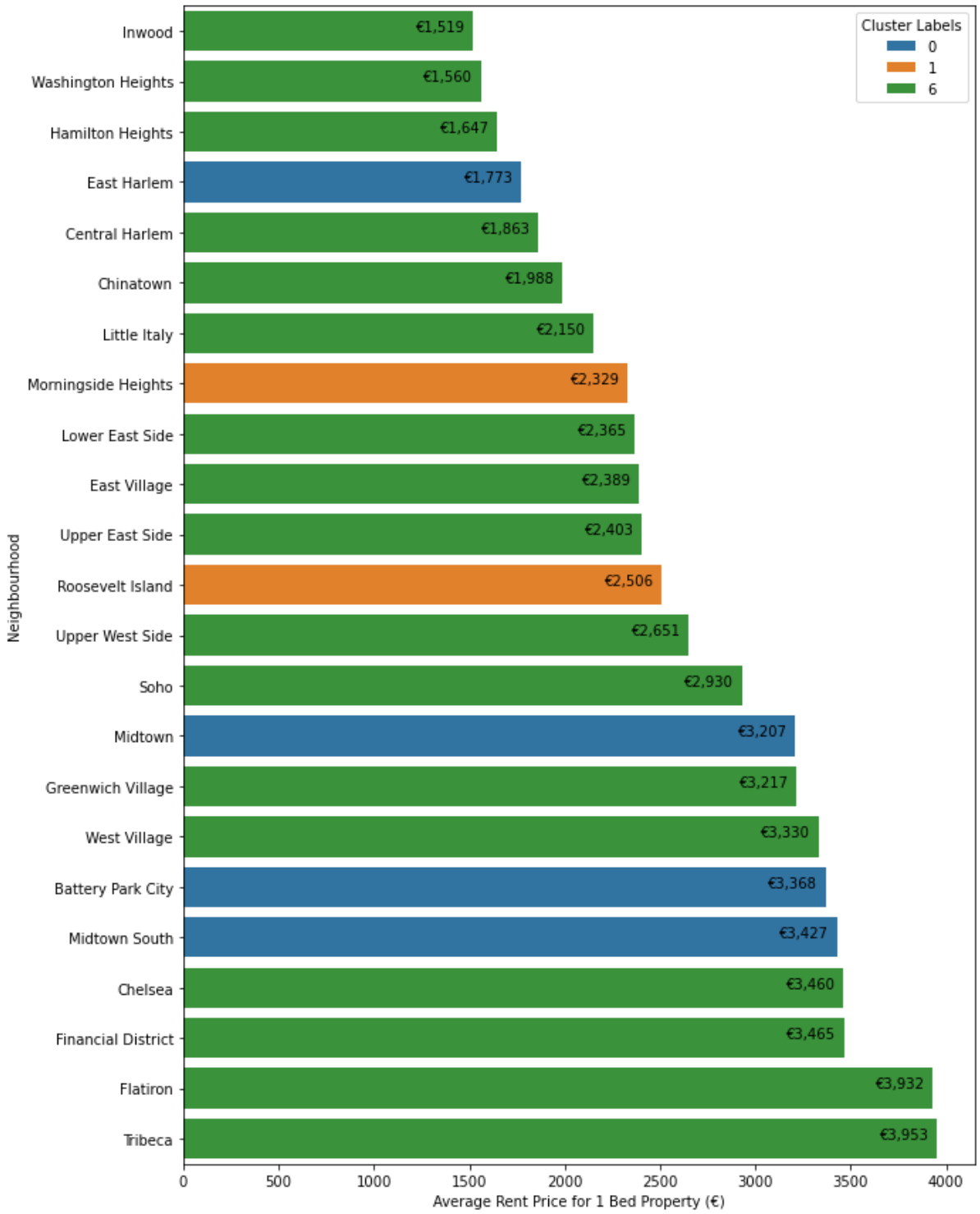

Figure 5. Manhattan average monthly rent price.

Figure 6, demonstrates the average rent prices for a one bed property in Dublin. It is clear that these rent prices fall into a much narrower range than those found in Manhattan. The lowest rent price is €904 found in Balbriggan which is the neighbourhood located furthest from Dublin City centre. The most expensive neighbourhood is Dublin 2, located in the heart of the city, with an average rent of €1428. The main thing to note is that all Dublin rent prices are less than the cheapest rent found in Manhattan.

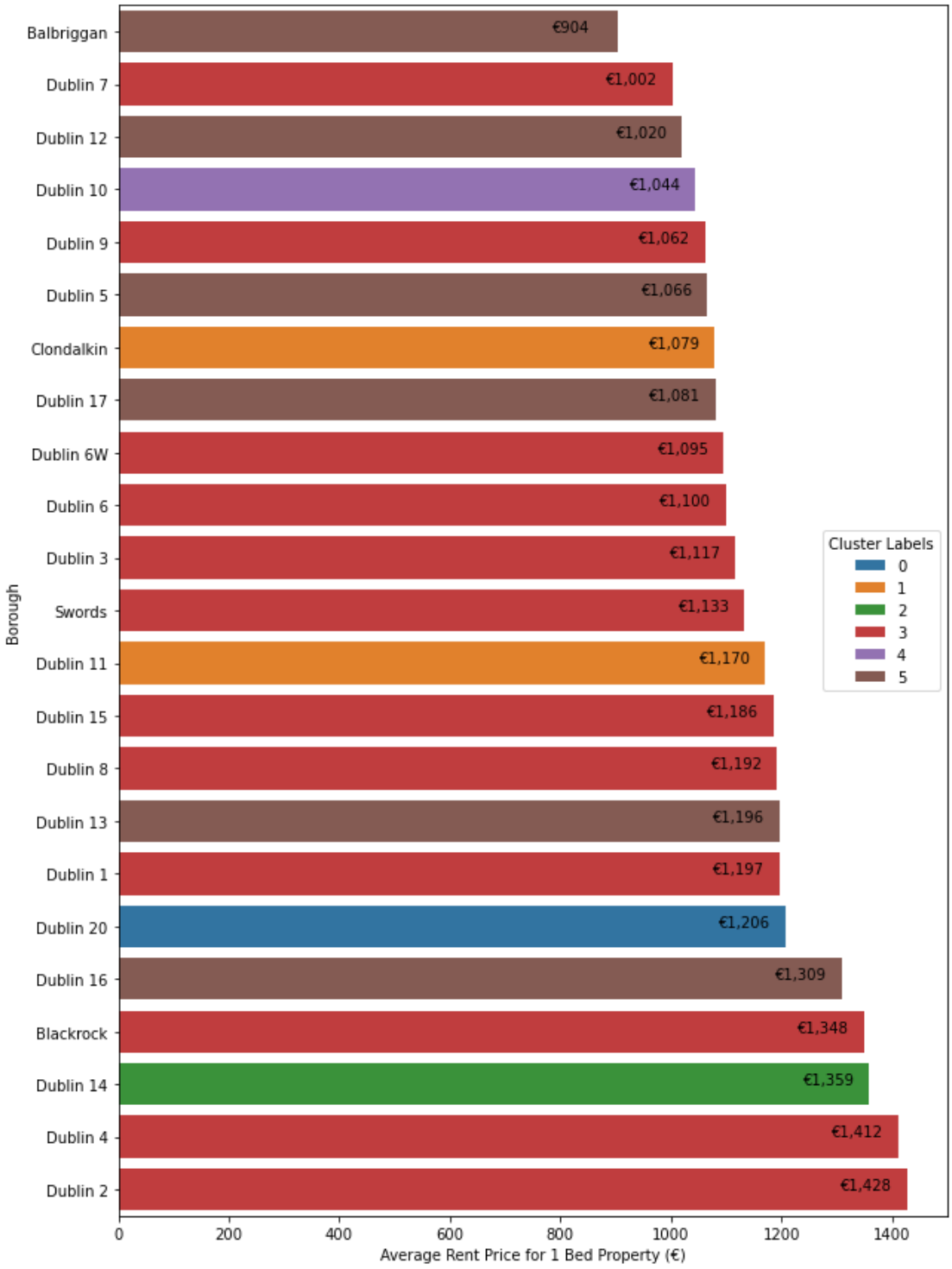

Figure 6. Dublin average monthly rent price.

Finally, the rent prices across all the similar neighbourhoods in each city are shown in Figure 7 below. This further demonstrates the stark difference in average rent price between the neighbourhoods in each city, with the rents in Manhattan being between 65-217% higher than those found in Dublin. Recalling from previously, cluster 1 had the most Dublin neighbourhoods in common with those in found in New York. Comparing these rent prices, the most cost effective option would be the neighbourhood of Morningside Heights. Alternatively, the neighbourhood of Clondalkin in Dublin assigned cluster label 0, shares similarities with a number of Manhattan neighbourhoods, with East Harlem being the cheapest possible choice of Manhattan neighbourhood which shares similarities with a Dublin neighbourhood.

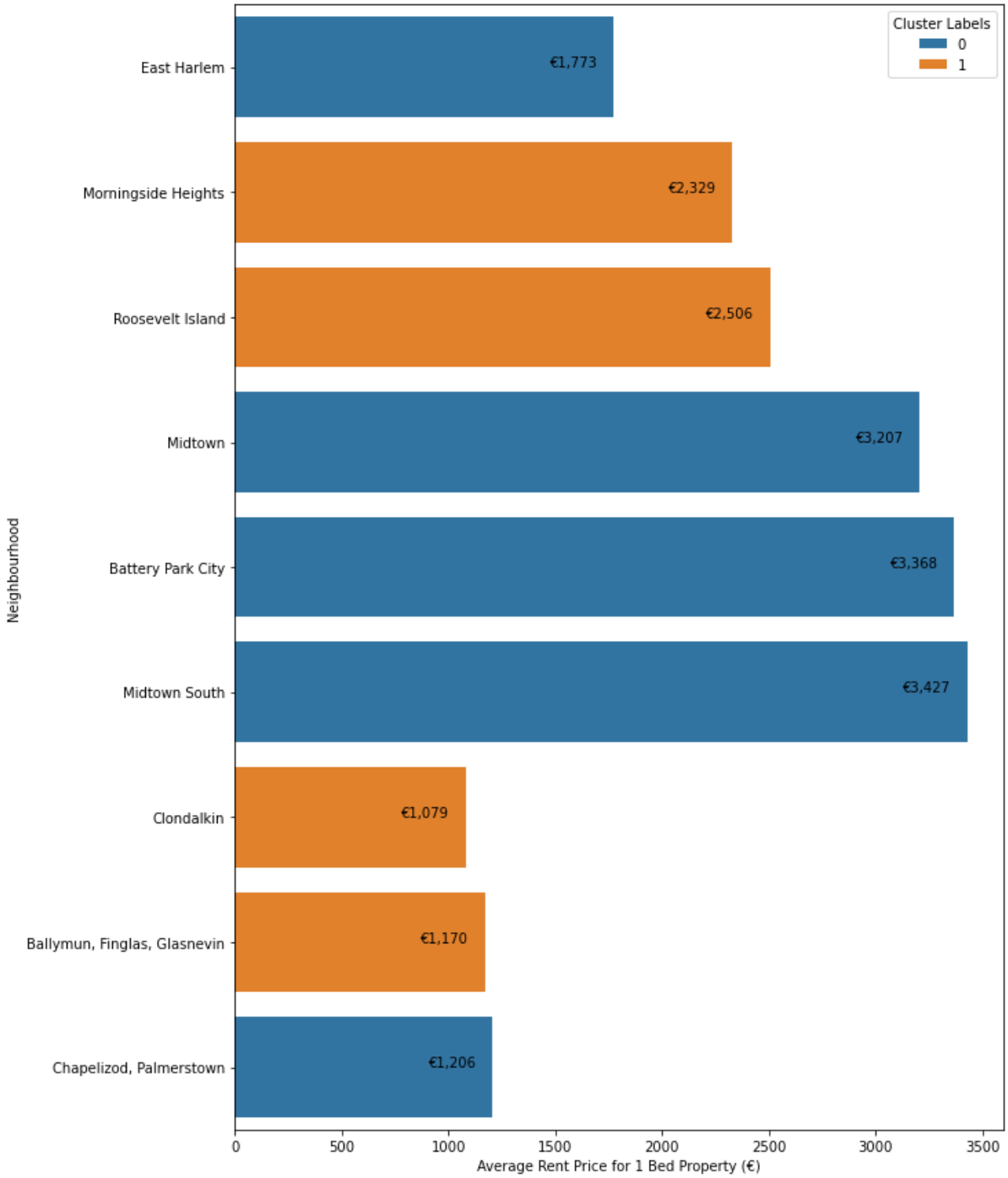

Figure 7. Average monthly rent price across similar neighbourhoods in each city.

The major caveat of this analysis is that the similarity between neighbourhoods is only defined by the venues found in that neighbourhood. This analysis does not account for any demographic, infrastructure or economic information about each neighbourhood. It also does not capture information on the differing cost of living in each city and the typical salary expectations of the inhabitants of each city. However, combined with some local knowledge, this analysis does perform a useful foundation upon which HR can base their neighbourhood recommendations on when helping new employees transition to New York.

The two Manhattan neighbourhoods that should be recommended to employees moving from Dublin to New York are East Harlem and Morningside Heights. These neighbourhoods are recommended on the basis that they share similar venues to those found in Dublin and they have the most affordable rent prices out of all the neighbourhoods that share similarities. Location wise these neighbourhoods are located on the North East and North West sides of Manhattan island, and would likely require a short commute downtown to HQ. This is very similar to the commute these employees would be used to in Dublin, as all the neighbourhoods that they shared similarities with lay on the outskirts of Dublin city.

Conclusion

The goal of this analysis was to provide information to the company HR department in order to guide their neighbourhood recommendations to employees temporarily moving to New York for training ahead of the opening of the European HQ. The goal is to find the neighbourhoods in Manhattan with the most affordable rents for a one bedroom property, which share similarities t neighbourhoods in Dublin.

A number of datasets were combined in order to perform this analysis. The Foursquare API was leveraged to the find the different types of venues found in each neighbourhood. The similar neighbourhoods in Dublin and New York were found using the K-means clustering method. This combined with an analysis of the average monthly rent price in each neighbourhood led to the two Manhattan neighbourhoods of Morningside Heights and East Harlem being selected for future recommendation. This is due to their similarity to a number of Dublin neighbourhoods and their relatively lower rent prices.

Acknowledgements

I would like to thank the hard working people at IBM for creating this incredibly informative course. I would also like to thank my fellow students who have kindly contributed to the discussion forums and provided valuable feedback during the peer reviewed assignments.

References

[1] US Census Bureau (https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045219)
[2] Business Insider, December 29 2019 (https://www.businessinsider.com/queens-languages-map-2017-2?r=US&IR=T)
[3] US Census Bureau (https://archive.vn/20200213122423/https://factfinder.census.gov/bkmk/table/1.0/en/ACS/15_1YR/B05007/1600000US3651000)