Battle of the Dublin Districts

Yash Karle, March 2021

1. Introduction

1.1 Background

- Housing crisis in Dublin

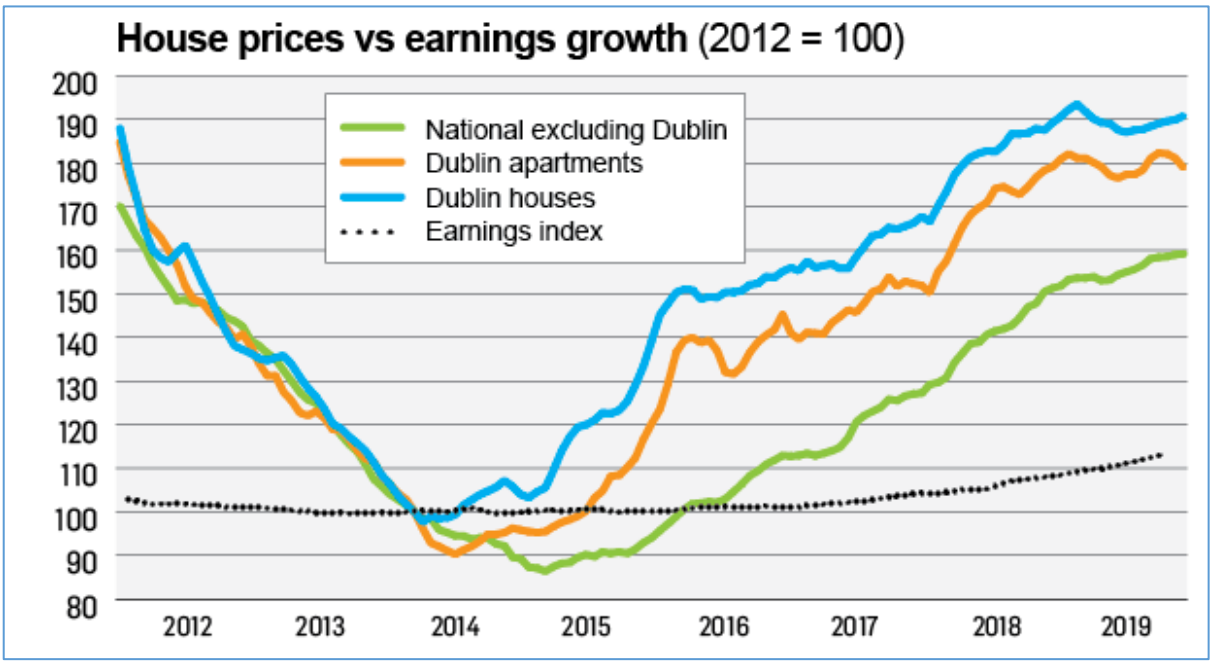

Figure 1: Dublin House Prices [1]

As you can see in the figure house prices in Dublin have risen by almost 90 percent, while wages have increased by only 18 percent since 2012. This is an ongoing social and economic problem in Ireland that the Irish government is trying to solve. Ireland needs an estimated 30,000 new units built annually and it's been lagging on that front by quite some margin in the last few years. This has led to multiple problems forefront of which are high levels of homelessness, longer commute times and greater wage pressures to meet the costs of living in Dublin. Families, renters, first time buyers and the elderly all have different housing needs as well as drastically different demands from the surrounding neighbourhood. Thereby, mixed housing is the need of the hour to support these varied range of demographically distinct population clusters.

- 15-minute city – impact of Covid-19

Given this context in terms of a national housing crisis (effect magnified in the capital which has majority of the urban population) arrives a global pandemic – Covid-19. Dublin City Council have accelerated their urban development plans along with the "Dublin Chamber" [2] initiative which tries to mimic 15-minute city initiative pilots in cities across the globe including Paris, Barcelona, Melbourne and London. As part of this and as the name suggests emphasis will be given on the needs of the ever-increasing urban density, enhanced public transport and investment in public realm.

They are essentially proposing better livability and walkability of urban neighbourhoods to ensure more sustainable communities. Especially in the post-Covid world with the changing ways of working it is a great time to kickstart reimagining of neighbourhoods.

1.2 Problem

- The perfect neighbourhood - 5K lockdowns

The problem lies in the background and is very much relevant to the current time. Level-5 restrictions in Ireland meant people had to restrict their movements within 5 km radius from their house. This is both a concern and an opportunity from the urban development point of view. Beyond this, a perfect neighbourhood in such scenarios would be able to satisfy all the "local" needs of the surrounding population in that neighbourhood. Different neighbourhoods would have diverse demographic distribution amongst its populations. As a local business it should be focusing on these demands of their most immediate potential customers and in turn it involves a mindset shift for these customers where they choose local businesses over bigger name brands.

- Skewed distribution of amenities and opportunities

Some neighbourhoods lack essential facilities and amenities let aside recreational spaces. Especially given the current circumstances, we are forced to focus on smaller parts of a neighbourhood (2km, 5km radius) to see if the businesses' and retailer's setup as part of these local towns are enough to meet the majority needs of the immediate surrounding population. The problem here is that some neighbourhoods have ample of amenities, shops and businesses whereas others do not. To add to the problem and what is prevalent especially in Dublin (housing crisis) the construction of houses is skewed on top of this which magnifies the gap between the demand (more houses built at the right places suited for a specific population) and the supply (a smaller number of houses built that too at places where the neighbourhood isn't necessarily suited for the needs of the population).

1.3 Interest

- Find the right house for me!

Apart from the urban development and housing perspective there is a personal interest in this project where I am exploring neighbourhoods (clustering) and then identifying attributes of houses in those similar neighbourhoods that directly/indirectly affect (promote/demote) the price of the houses given that my aim is to find the perfect house for me in Dublin. This should provide guidance when shortlisting properties to view before buying a house where we identify suitable neighbourhoods and try to see if the data can justify the price tags depending on the house attributes and neighbourhood characteristics. In terms of the arguments provided in the earlier section, the aim is to identify neighbourhoods for people sharing similar needs (e.g. travel time to work, parks in the vicinity, restaurants, shops in the neighbourhood) as myself.

- Hyper-personalisation

  Lastly, we look at potential value that this project might provide for local small businesses when they try to target their nearby customers with relevant and personalized ad campaigns. With hyper-personalisation in mind, zip-codes or neighbourhood indexing can help local businesses to target audiences nearby with the aim to upsell relevant produces/services which should benefit the customers given the current circumstances. This should largely benefit local stores without any digital presence where advertisers/local council authorities can promote products/services based on the location of the viewer.

2. Data acquisition and cleaning

2.1 Data sources

- Daft-scraper API

  As seen below, this is a very useful API [3] which is simple to use and get up to speed.

```
options = [
   PropertyTypesOption([PropertyType.ALL]),
   LocationsOption([Location.DUBLIN_COUNTY]),
   AdStateOption(AdState.AGREED)
]
```

  The sample example above shows a search using the API filter options to get all listings in "Dublin county where a property sale has been agreed for any type of property" as listed on the popular Irish Property website (Daft.ie)
  We fetch all such listings and build a dataframe containing all the useful features for each property which as seen below would consist of <price', 'facilities', 'address', 'num_bedrooms', 'num_bathrooms', 'latitude', 'longitude'>
  This data would help us recommend properties to the targeted end-user as well as the geographical coordinates would help us visually analyse the data in question.

- Foursquare API

  This involves a similar approach taken during the previous weeks in this course where we had analysed different neighbourhoods in Toronto, Canada.
  The challenge here is to obtain different districts comprising within Dublin City and obtain their respective geographical coordinates using Nominatim geolocator.
  The sample code given in the notebook shows how we plan to construct the final dataframe where each row would be an individual venue along-with the attributes of each of the venues including their geolocation coordinates.
  OneHotEncoding can be used to get a feature representing distribution of different types of venues as well as the most popular and dominating venue type in each of the districts within Dublin city.

2.2 Data cleaning

- Feature Extraction
  - Neighbourhood – We are fetching this feature from the “title” obtained from the Daft-scraper API response. Essentially, we are splitting the string using ‘,’ and trying to pick the last part in it which in most cases is the neighbourhood. The

values for this represent each of the Dublin districts (example Dublin1, Dublin 2 up-to Dublin 24). As you can see below Dublin 15 has the greatest number of properties in our working dataset.

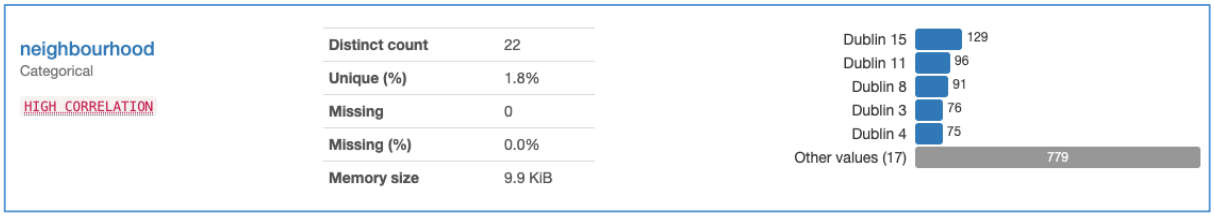

Figure 2: 'Neighbourhood' feature inspection

- SellerId – This represents the unique identifier for the seller associated with the completed sale of each property advertised on Daft. This is directly available for each record as fetched from the field "seller".

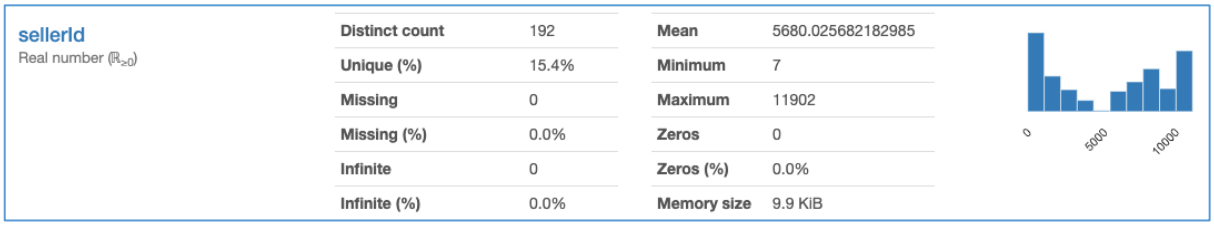

Figure 3: 'SellerId' feature inspection

- FloorArea – This was available in a nested field fetched from "floorArea"; the only changes were needed for records where the unit of area was acres and not squared metres. For consistent comparison we have converted all records to squared metres.
- Longitude – This was another nested field fetched from the field "point". This represents the longitude geolocation value for the precise location of the respective property.

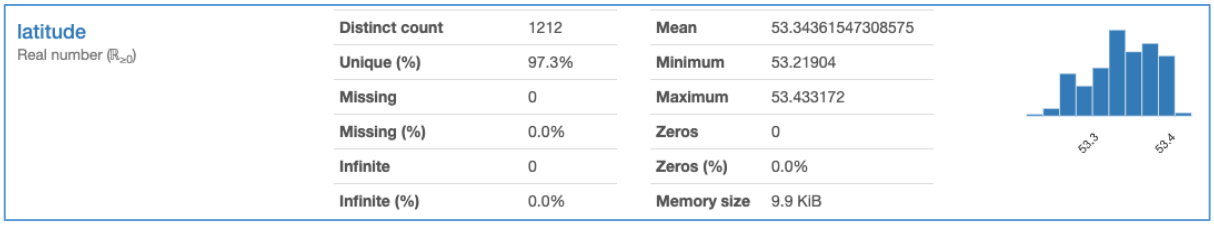

Figure 4: 'Longitude' feature inspection

- Latitude – This was another nested field fetched from the field "point". This represents the latitude geolocation value for the precise location of the respective property.

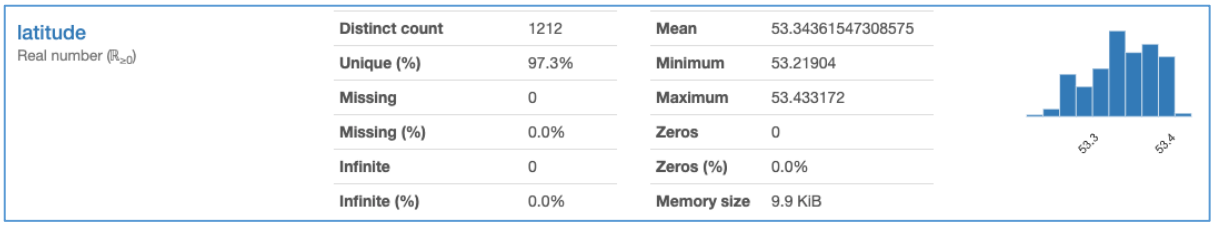

Figure 5: 'Latitude' feature inspection

- BerRating – Lastly, this feature was directly available from the "rating" field received in the API response. This represents the electricity/energy consumption rating of the respective property. As you can see below, values range from A1 (best) to F/G (worst) depending on the energy consumption indicators.

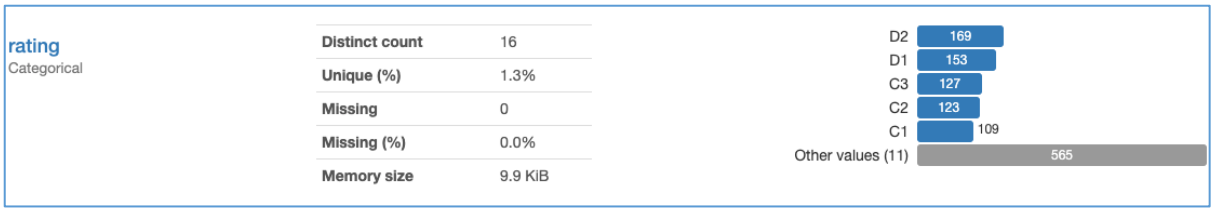

Figure 6: 'BerRating' feature inspection

- Missing data
  - Price – For records where "price" field had a 0 value, we fetched the same from the "abbreviatedPrice" field for the same record (if available). Values that where represented using 'k'(thousands) and 'M' (millions) were taken care of and converted to unit prices for consistent comparison.
  - <NumBedrooms, NumBathrooms, BerRating> – For certain property types (like Site/Townhouse) these fields are not applicable or irrelevant so we simply replace the null values with drastically different values from the non-null records to ensure they are classified separately in the final analysis.

2.3 Feature selection

- Feature Synthesis
  - PricePerBedroom – This feature (Price / NumBedrooms) is quite self-explanatory where we are trying to find out the price per each bedroom for a given property. This feature should help guide any observations for the relation between number of bedrooms and price for a given neighbourhood.
  - <DeltaAvgPrice, DeltaMedianPrice> – These two features are constructed to assess the deviation of prices (around the mean and median respectively) for properties within each neighbourhood just to gauge the fluctuation of prices within that neighbourhood.
  - NorthSouthFlag – This is another simple feature created based on knowledge of how districts are named in Dublin. All the odd districts are classified as 'North' and all the even districts are classified as 'South'. This should help us compare clusters of properties on the opposite side of the Dublin city center.
  - DistanceToCity – This is an additional feature apart from the <latitude, longitude> pair where we are interested to see how the distance from the city

center has an effect on the prices of the properties. As you can see from the notebook, we have used the haversine distance formula between two geo-coordinates to calculate this feature value for each property.
    - DaysSincePublished – Straightforward feature created from the 'AdPublishDate' field that the API mentions which is essentially counting days since the advert was published and the date we are running the analysis to give us a rough estimate as to how much time the advert has been active.
    - <NumFood, NumRecreation, NumShop> – Lastly this set of features are curated using a custom method written to process the results received from the Foursquare API. Drawing inspiration from the local lockdowns imposed during the ongoing pandemic we are trying to see how much of an impact the immediate neighbourhood (i.e. within a 5km radius from the concerned property location) shops, restaurants and recreational amenities have on the property price.

- Aggressive filtering
    - Zero Price – We have removed all records that had a zero valued price (including checking abbreviated price value). Any property record without a price value doesn't add any value to this project's hypothesis so we simply remove all such rows.
    - Null values for newly added features – Any custom features that we created as discussed in the previous section and thereby that have null values are also removed from the analysis dataframe. We have taken this route of aggressive filtering to make sure that whatever data records we have are of good quality and highly informative to draw reliable insights from. There might have been better techniques to deal with such instances but they are out of scope of this project.

3. Exploratory Data Analysis

As we can see below, pandas-profiling is the tool of choice for doing some very high-level Exploratory Data Analysis (EDA). We see in Figure 7 below the typical (n x n) matrix where n is the number of features. The intersecting blocks/cells in the matrix are represented as heat map which gives us an indication about the correlation between each pair of features. Here are some of the interesting observations by merely looking at the matrix below where we assess the values of the correlation coefficient (as r approaches +1 the pair of features are positively correlated and likewise as r approaches -1 the pair are negatively correlated):

1. (Longitude, Price) – They show a positive moderate correlation suggesting as the Longitude value increases (as we move across the city from West to East) the Price value increases proportionately.
2. (DistanceToCity, NumBedrooms) – Again this pair shows a moderate positive correlation and it indicates that as DistanceToCity value increases (as we move away from the city center) NumBedrooms increases proportionately which makes sense as we can expect more larger houses/properties as we move towards the suburbs / outskirts of the city.
3. (NumFood, DistanceToCity) – This pair shows a very strong negative correlation suggesting almost in all cases as the NumFood value increases DistanceToCity increases proportionately too. An expected observation which suggests there are more Food places (restaurants and bars) in a 5km radius from the houses as we move closer to the city center.

4. (NumRecreation, DistanceToCity) – This pair too shows a moderate negative correlation which again suggests there are more Recreation places in a 5km radius from the houses as we move closer to the city center.
5. (NumShop, DistanceToCity) – Finally this pair shows a very strong positive correlation which indicates there are more shops away from the city center compared to closer to it. Perhaps an indication of bigger stores, warehouses and shopping outlets in suburbs outside town owing to the more open spaces available.

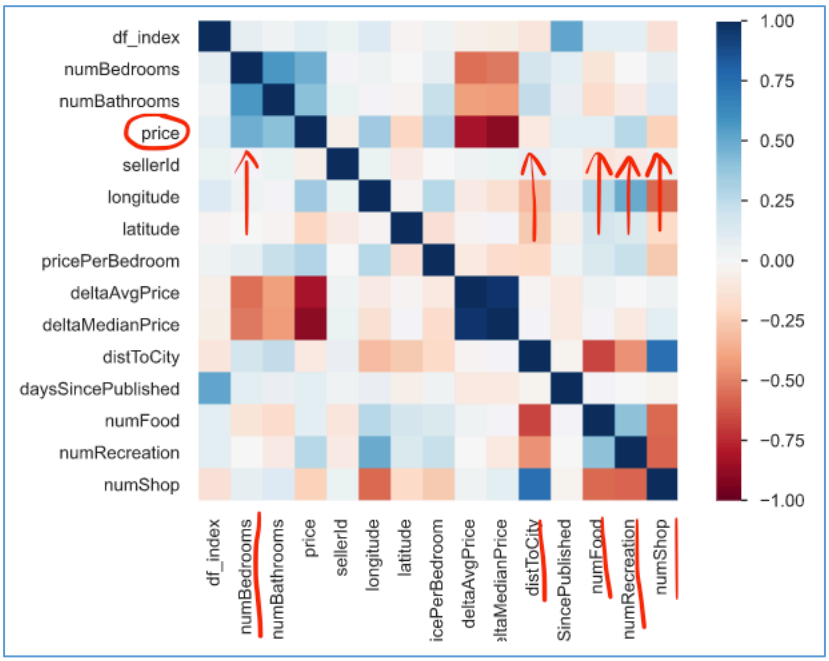

Figure 7: Pearson's 'r' (correlation coefficient) matrix

3.1 Relationship between Price and NumBedrooms

Next, we look at more such pairs and their direct interactions focusing purely on the price of houses. First such pair of interest is <Price, NumBedrooms>. As quite obviously seen from Figure 8 below, NumBedrooms has discrete values from 1 to 10 and the center of distribution being around 3 bedrooms. We can see a clear spike in terms of the price as we move from left to right with a peak around 4-bedroom houses thereby suggesting a direct correlation between the two features. This is completely expected where the prices of a house increase with increase in number of bedrooms within that house. What's slightly interesting to note here is that 3/4-bedroom houses seem to have the highest price indicating that neighbourhood of the house might influence its price and not just mere number of bedrooms. We will explore the influence of neighbourhood and other features on price in the further sections below.

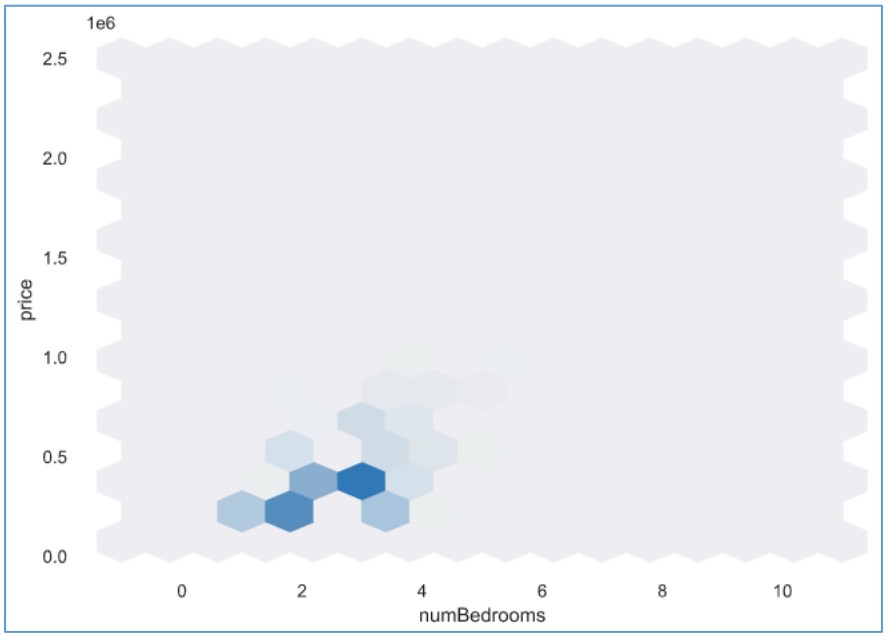

Figure 8: Price v/s NumBedrooms

3.2 Relationship between Price and DistanceToCity

Below in Figure 9 we can see the interaction between <Price, DistanceToCity>. Noticeably there are 2 peaks in terms of value of price. We see that the price peaks at about 3kms from the city in terms of distance and then another smaller spike in price at around 10kms. Apart from that there is a hint of negative correlation where we see prices falling as the distance from the city increases.

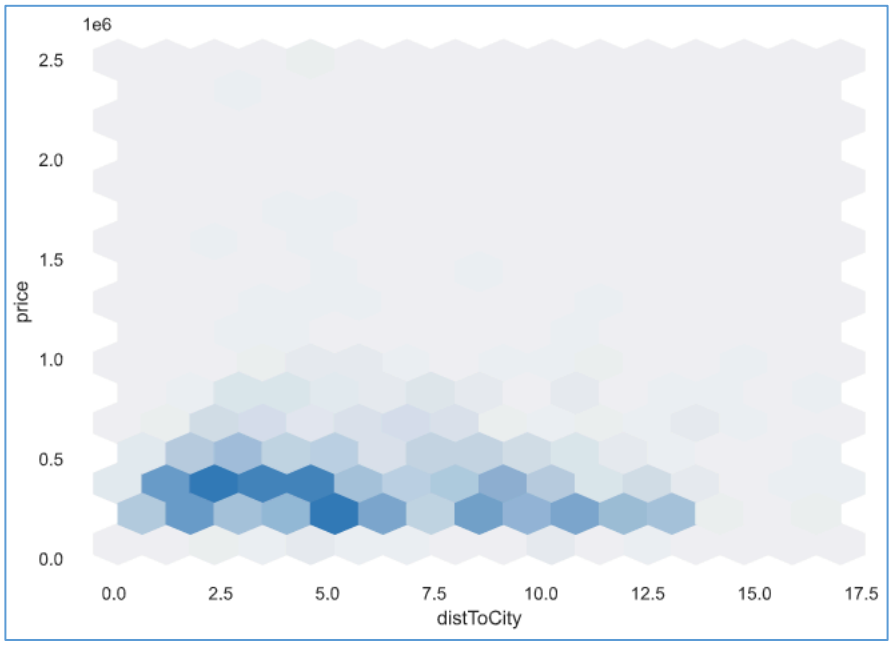

Figure 9: Price v/s DistanceToCity

3.3 Relationship between Price and NumFood

Next, we observe the interaction of features <Price, NumFood>. Here the relation is much more straightforward as seen in Figure 10 below. The price peaks gradually as the NumFood approaches 50. There is also a prevalent direct correlation as discussed before where the price increases as number of food places increase in the immediate neighbourhood of a property. This shows how much of a key influential feature NumFood is as far as the price is concerned.

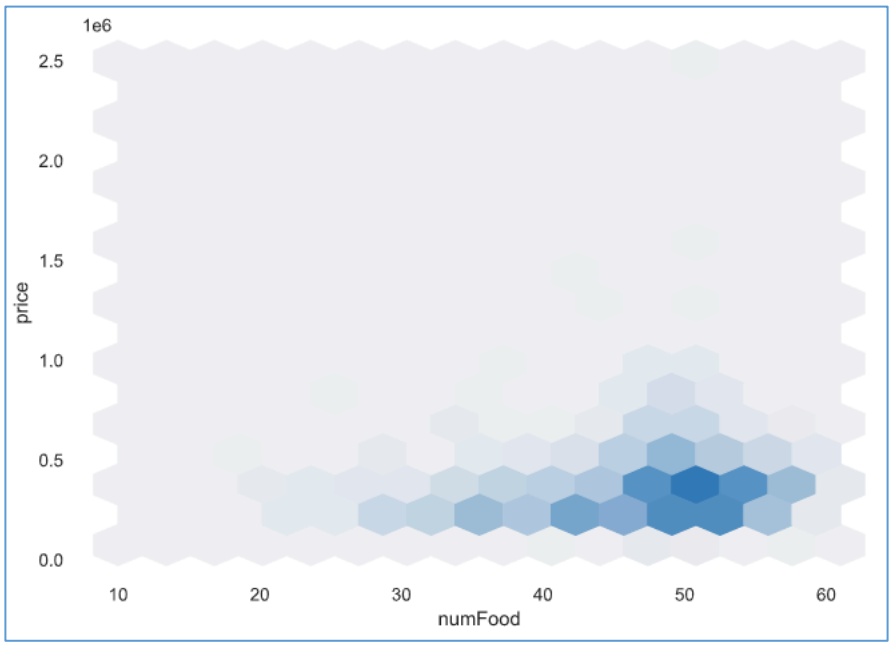

Figure 10: Price v/s NumFood

3.4 Relationship between Price and NumRecreation

Another similar such interaction is <Price, NumRecreation>. As seen in Figure 11 below, the price is quite evenly distributed from in between 10-15 number of recreational places. This region is also the densest part of the plot meaning most houses, irrespective of their prices, do have these many recreational places in the immediate neighbourhood. That also suggests that in terms of its interaction with Price, NumRecreation won't be that much of a helpful predictor or regressor. Lastly, as also discussed previously the plot below seem to suggest that NumRecreation is ever so slightly positively correlated with Price.

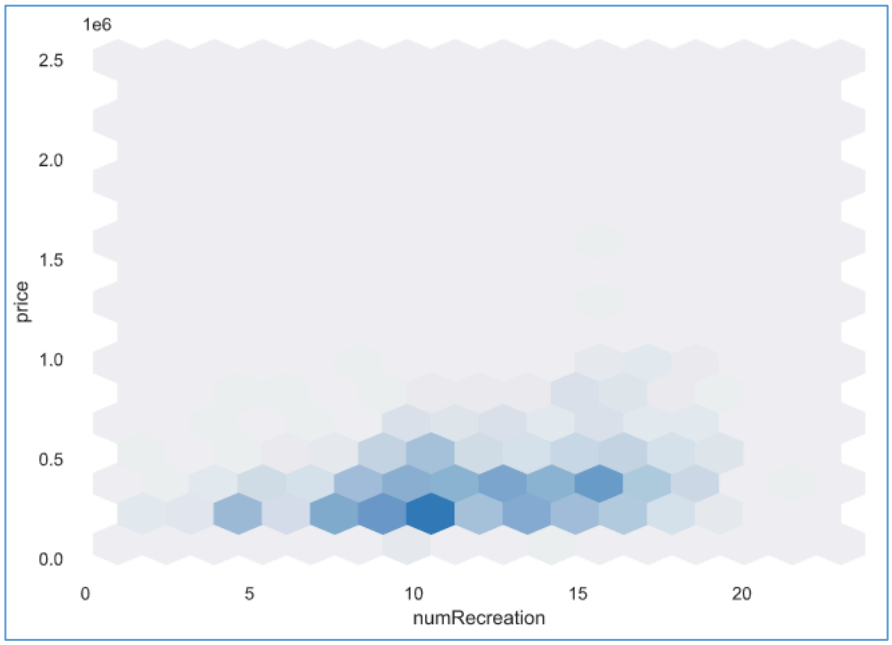

Figure 11: Price v/s NumRecreation

3.5 Relationship between Price and NumShop

Lastly, we look at the interaction between <Price, NumShop>. As seen in Figure 12 below, values for number of shops are spread across a wide range right from 5 to 30 shops in within 5kms of the property. In terms of relation with price, there is a small peak at 5 shops but it would be very strange to draw any conclusions from this but the only plausible explanation to this might be that the peak indicates residential area. The peak seems to taper as the number of shops increase suggesting a negative correlation with price.

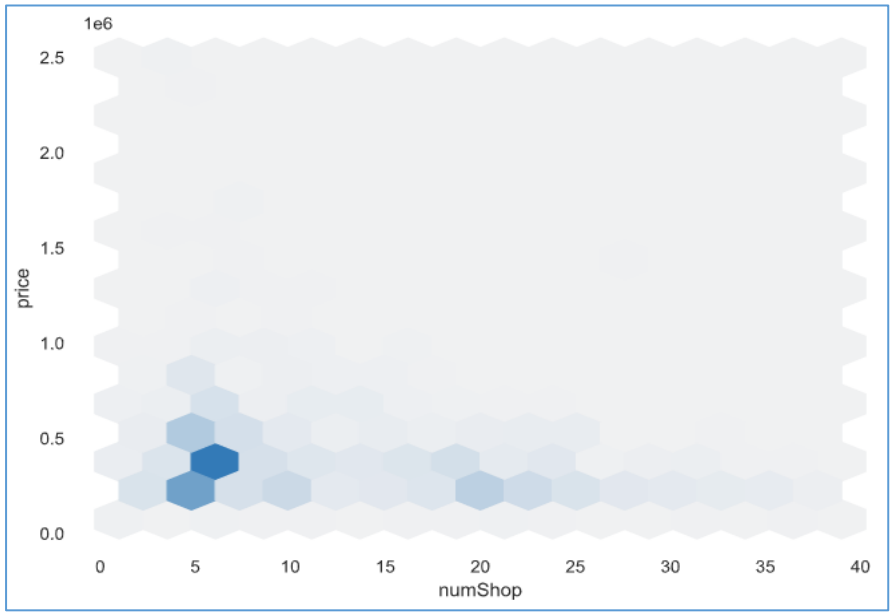

Figure 12: Price v/s NumRecreation

4 Clustering

4.1 DBSCAN clustering

- Property characteristics: <NumBedrooms, NumBathrooms, FloorArea>

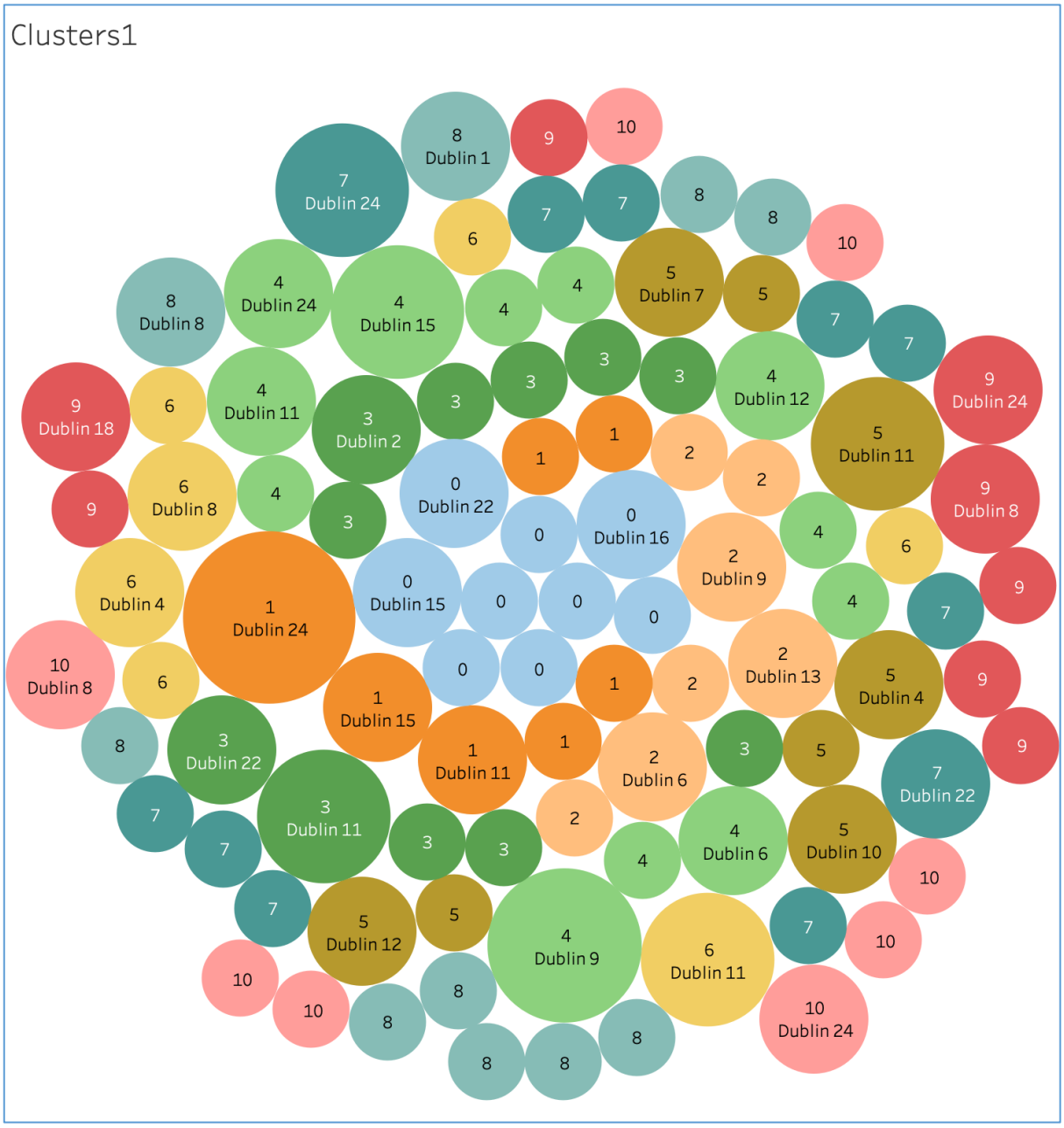

Figure 13: Clusters based on Property characteristics

The attempt here is to use DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm to create clusters within the property records we have. DBSCAN algorithm fits our purpose as it is density and spatial based so it should take into consideration the relative positions of the houses as seen on a map and more importantly it inherently deals with noise in the data. Firstly, we look at clustering the properties depending on their own characteristics (area, bedrooms, bathrooms). As seen in the Figure 13 above, we then look at each of the clusters and number of properties

that each neighbourhood has within that cluster. This should tell us similar/dissimilar neighbourhoods purely from the property characteristics.
Here are some of the observations after we visualize the clusters:

- D15, D16, D22 share same sized representation within Cluster 0
- D9, D13 are again very similar neighbourhoods looking at Cluster 2
- D6, D11, D12, D24 represent similar sizes in Cluster 4
- D4, D7, D10, D12 are all similar within Cluster 5
- D1, D8 have similar property characteristics within Cluster 8
- D8, D24 appear as similar sized in Clusters 9,10

The observations tell us that there are pairs of neighbourhoods where the residential setting or at least the part of it share some characteristics in terms of the property characteristics. For example, if we were to make sense of the fact that D9 and D13 are in the same cluster it would suggest that most houses in the area are apartments with a median of 2 bedrooms per apartment and average floor area of 70 sq. ft. (looking in more depth at the data within the records for each cluster)

- Location characteristics: <Longitude, Latitude>

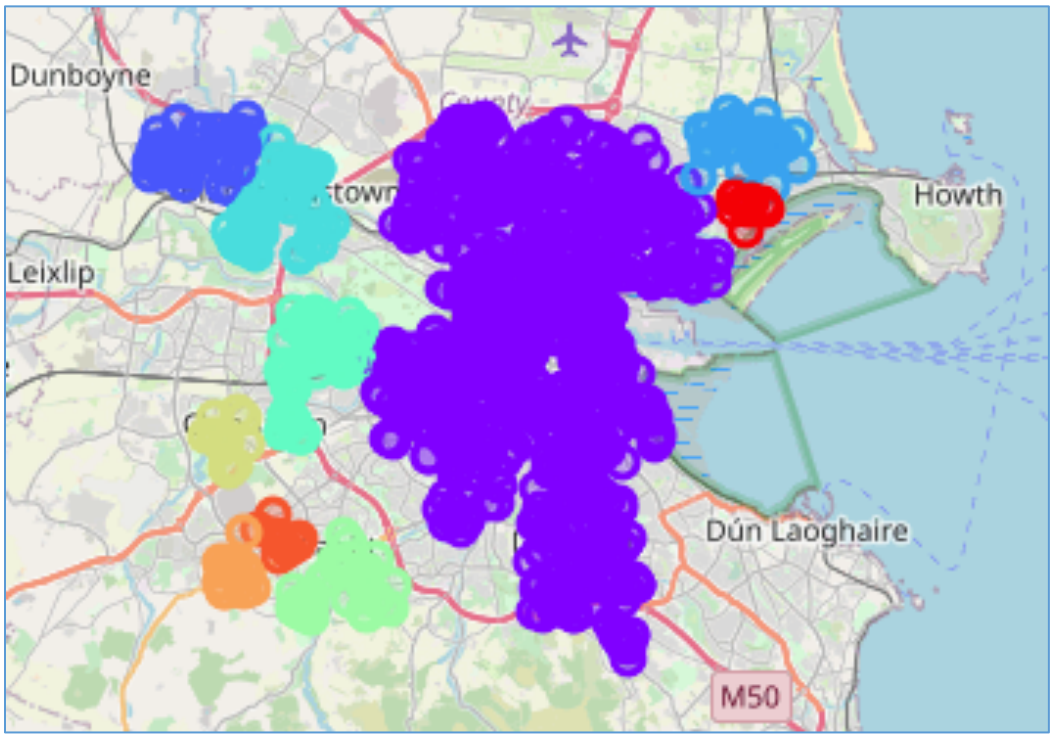

Figure 14: Clusters based on Location characteristics

Next, we again apply the DBSCAN clustering algorithm but this time it's merely based on the geolocation characteristics (longitude, latitude) of the property in terms of where it is placed on the Dublin city map. As we can see in Figure 14 above, after trying with few different parameter values for the DBSCAN algorithm we get this interesting distribution of properties within clusters which are dense yet distinct enough at the same time. In terms of observing the plotted clusters above, it's very obvious to see that the largest cluster in purple covers the inner Dublin city (within the M50 highway which is essentially a ring road along the circumference of Dublin city). What's more interesting

is the fact that the algorithm was able to distinguish between spatially farther but less dense areas as we can see from the smaller clusters on the map. All the clusters are on or outside the M50 boundary and are basically suburbs or towns within Dublin county but outside Dublin city.

- Neighbourhood characteristics: <NumFood, NumRecreation, NumShop>

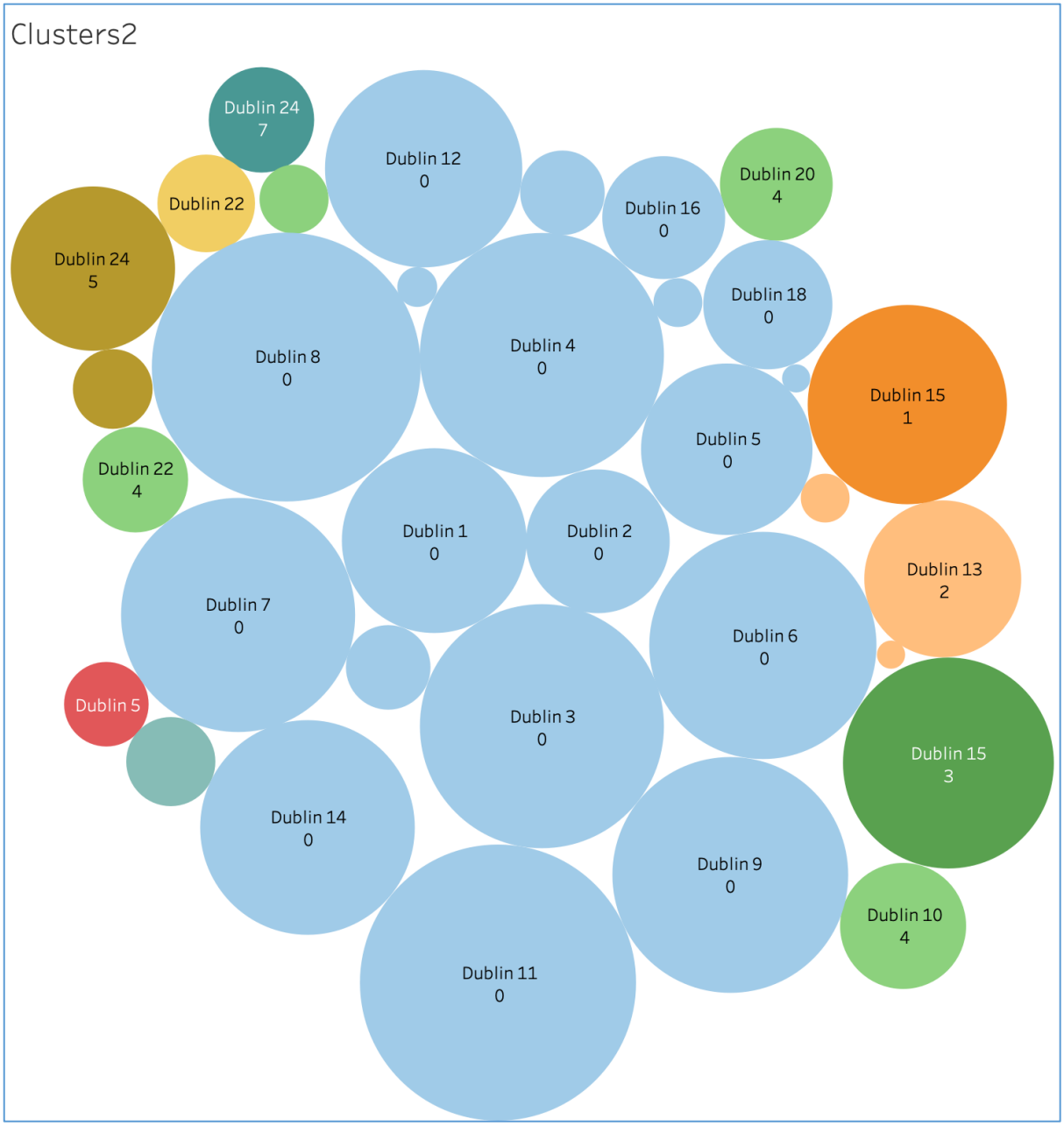

Figure 15: Clusters based on Neighbourhood characteristics

Lastly, we look at the neighbourhood characteristics (shops, restaurants and recreational places) when applying the DBSCAN clustering algorithm as we did above with the property characteristics. Again, we follow the same approach of trying to group the clusters by their neighbourhood and try to see if any neighbourhoods are similar or dissimilar. It will be interesting to see why these clusters were formed the way they

were and explore more in terms of granular factors of neighbourhood planning but this is something beyond the scope of this project. As we can see from the Figure 15 above, here are some of the important high-level observations:

- D2, D16, D18 share the same neighbourhood characteristics as seen from Cluster 0 all of which are south Dublin neighbourhoods. Similarly, D1, D5 can be seen as similar neighbourhoods which in turn are north Dublin neighbourhoods.
- As far as the remainder of Cluster 0 is concerned, D3, D4, D6, D7, D9, D12, D14 are all fairly equal sized and stacked together. That leaves D8, D11 as the largest sized neighbourhoods within Cluster 0.
- D10, D20, D22 again all of which are southside neighbourhoods hold very similar neighbourhood characteristics and it is represented from Cluster 4
- Apart from we see singleton neighbourhood clusters which suggests that these clusters are quite unique in terms of the distribution of shops, restaurants and recreational places – for example D15 in Cluster 1 and Cluster 3, D13 in Cluster 2, D24 in Cluster 5 and Cluster 7

4.2 Solution to the problem

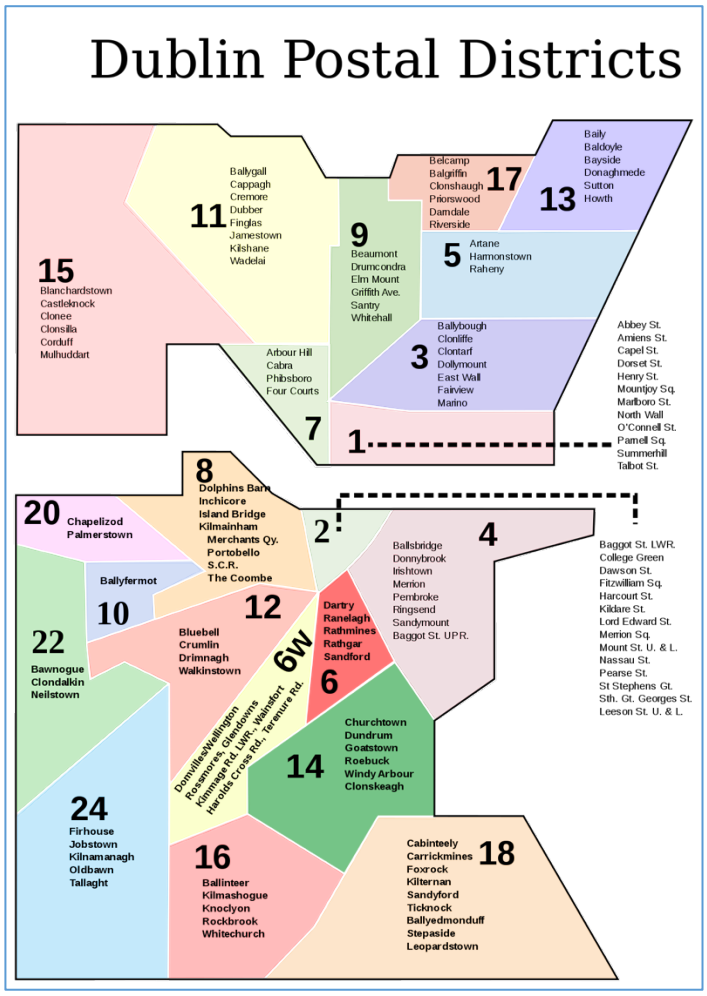

Figure 16: Dublin Postal Districts [4]

Now in this section we aim to link all the observations and analyses above to our original identified problem. First of which moving to a newly defined concept about localism and "15-minute cities" in the post pandemic era. Here is a brief flavour of recommendations that can be made putting the problem at hand in context and then purely revisiting the observations above:

1. New localism – Urban planning and development have been rapidly accelerated ever since the onset of the pandemic and local governments are now focusing on their plans ahead as countries and neighbourhoods look towards leaving from the lockdown world. The clusters and the observations around how amenities are distributed within a city or city suburbs for that matter can provide some important leads for the decision-making involved around the whole "15-minute city" ideology. These decisions can be based on focusing on smaller neighbourhoods and what's lacking to make it self-dependent and accessible for the majority mass living/working there.
2. Pricing of houses – This observational analysis directly affects three different types of personas namely:
    a. Developers involved in constructing housing plans in a profitable and sustainable manner
    b. Financial Services/Banks in terms of deciding optimal mortgage rates for lending
    c. Lastly, the actual buyers or investors who are trying to find the perfect match that satisfies all their criteria.
3. Hyper-personalisation – To boost digital presence of local small businesses where the retailers/restaurateurs can target more relevant and immediately local audience. This should increase their relevant customer reach and boost engagement meanwhile saving their shipping costs. From the other perspective, thinking of the present time and context it would suit the residents of a particular neighbourhood to work, shop, eat and enjoy responsibly all locally with minimum travelling needs.

5. Conclusion

To summarize the project and the implementation, we have looked at fetching the housing data using the Daft-scraper API and joined the same with the Foursquare API to curate some useful features which allows us to understand the interactions with price in terms of price indicators (promoters and demoters) as well as get to know more about the different districts within Dublin city looking from multiple different perspectives – urban development, post-pandemic town planning, uplifting small businesses and even help new buyers understand the market triggers and plan their move into a potentially new neighbourhood.

6. Future directions

Some additional explorations possible beyond the current scope of this project:

- Fetch seller names from seller id and see if the data makes sense
- Most common & least common venues for each neighbourhood
- Removing outliers for price and then redo price deciles
- Popular transport routes, commute time to city center as an influencer on price
- Schools in the neighbourhood influencing house prices
- Crime rates in a neighbourhood and correlation with some of the other price indicators
- Pricing per sq. ft of area

References:

1. Irish Time Article: https://www.irishtimes.com/life-and-style/homes-and-property/ireland-s-housing-crisis-in-five-revealing-graphs-1.4150332
2. Dublin Chamber of Commerce Blueprint Document: https://www.dublinchamber.ie/DublinChamberofCommerce/media/banners/Dublin_The-15-Minute-City.pdf
3. Daft-scraper API: https://github.com/TheJokersThief/daft-scraper
4. Dublin Postal Districts Wiki: https://en.wikipedia.org/wiki/List_of_Dublin_postal_districts