[DIGITARE IL NOME DELLA SOCIETÀ]

Comparative analysis of the wards in the city of Derby (United Kingdom) based on venue information, price house and criminality rate

Course: Applied Data Science Capstone Project

Author: MdN Calvo Mateo

06/07/2020

1. Introduction / Business Problem

The city of Derby, situated in the East Midlands region of United Kingdom, had a population of 256,906 habitants in 2019. The city of Derby is divided into 17 wards, or neighborhoods, as it is displayed in Figure 1.

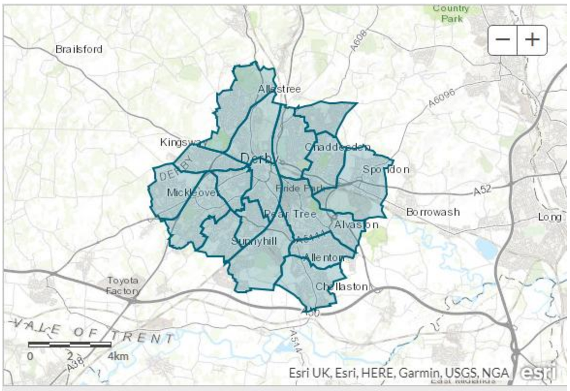


Figure 1. Map of Derby and its wards. Source: https://info4derby.derby.gov.uk/

Derby is one of the cities in United Kingdom with the highest proportion of advance manufacturing jobs across its population. Rolls-Royce plc, Toyota and Bombardier, which are 3 important multinationals dedicated to the manufacturing of aerospace engines, cars and trains, respectively, have manufacturing plants located in or around Derby.

These 3 companies employ a significant amount of employees which relocate to Derby from another countries or another places in United Kingdom for job reasons. One of the key decisions faced by people when relocating it is to decide where to rent or purchase a property to live in. Choosing the right neighborhood or ward according to personal preference is not an easy and straightforward task to accomplish in an unknown city.

This is the reason why this report will focus of comparing the 17 different wards in which the city of Derby is divided, and compare them based on the following criteria:

- Existing venues and services in each ward.
- Criminality rate.
- Median house price.

This report will be useful for those people moving to Derby or prospective property buyers in Derby since it will enable to have a better understanding of the various wards, the similarity amongst them and their main characteristics. This information will enable the potential property renters/buyers in Derby to make a more informed decision regarding where to choose to live.

2. Data Sources

With the prospects of understanding the selected characteristics of the 17 wards in which the city of Derby is divided, the following data and information has been gathered.

- The venues in each of the 17 Derby wards, based on the available information in the Foursquare API Database.
    - The information regarding the existing venues in each ward has been accessed via the Foursquare API application.
      Source: https://developer.foursquare.com/
    - The coordinates for the estimated central location for each ward have been obtained from Google Maps and manually inputted into an Excel Document.
      Source: https://www.google.com/maps

- The criminality rate in each of the wards in the period May 2019 – May 2020, measured as rate of crimes per 1000 population.
    - The information regarding criminality rates per ward in Derby, during the period May 2019 – May 2020 has been downloaded from the "Info4Derby Portal", and this information has been added to the existing Excel document already containing the coordinates for each Derby ward. **Source:** https://info4derby.derby.gov.uk/crime-and-community-safety/reports/

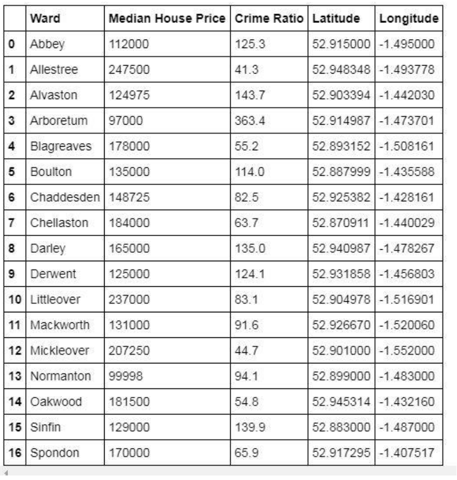
- The median price paid for all property types in each of the Derby wards during Q1 2019, expressed in GBP. All property types include Bungalows, Flats/Maisonettes, terraced houses, semi-detached houses and detached houses.
    - The information regarding median property prices per ward in Derby for the period Q1 2019 has been downloaded from the "Info4Derby Portal". This information has been added to the existing Excel document already containing the coordinates for each Derby ward and the associated criminality rates.
      Source:
      https://info4derby.derby.gov.uk/housing/report/view/d867ba1e909244c8ac75f97294cb7b93/E05001767

Once the required data for the analysis has been gathered, this project will analyse this information using a Jupyter Notebook with the programming language Python. The previously described data will be imported as a Pandas Dataframe and once prepared, this project will apply unsupervised Machine Learning methodologies in order to cluster the 17 wards in Derby in various clusters, with the objective of identifying similar clusters based on the criteria describe in the lines above. In this case, the methodology that will be used to analyze the dataset will be K-means Clustering. A detailed explanation of the methodology used in this project will be presented on the next sections of this document.

3. Methodology

This section will describe the methodology followed to analyse the 17 Derby wards based on the venue information, the mean house price and criminality rates in each of them.

After gathering the data regarding the 17 wards in Derby described in the "Data Sources" section, and store it into a Excel File, this information was imported into a Pandas Dataframe in a Jupyter notebook, while using the Python kernel (Figure 2)



	Ward	Median House Price	Crime Ratio	Latitude	Longitude
0	Abbey	112000	125.3	52.915000	-1.495000
1	Allestree	247500	41.3	52.948348	-1.493778
2	Alvaston	124975	143.7	52.903394	-1.442030
3	Arboretum	97000	363.4	52.914987	-1.473701
4	Blagreaves	178000	55.2	52.893152	-1.508161
5	Boulton	135000	114.0	52.887999	-1.435588
6	Chaddesden	148725	82.5	52.925382	-1.428161
7	Chellaston	184000	63.7	52.870911	-1.440029
8	Darley	165000	135.0	52.940987	-1.478267
9	Derwent	125000	124.1	52.931858	-1.456803
10	Littleover	237000	83.1	52.904978	-1.516901
11	Mackworth	131000	91.6	52.926670	-1.520060
12	Mickleover	207250	44.7	52.901000	-1.552000
13	Normanton	99998	94.1	52.899000	-1.483000
14	Oakwood	181500	54.8	52.945314	-1.432160
15	Sinfin	129000	139.9	52.883000	-1.487000
16	Spondon	170000	65.9	52.917295	-1.407517

Figure 2. Dataframe containing information regarding Derby Wards

A map of the city of Derby indicating the central points (Latitude and Longitude) for each of the 17 Derby wards was generated using the Folium package, as displayed in Figure 2.

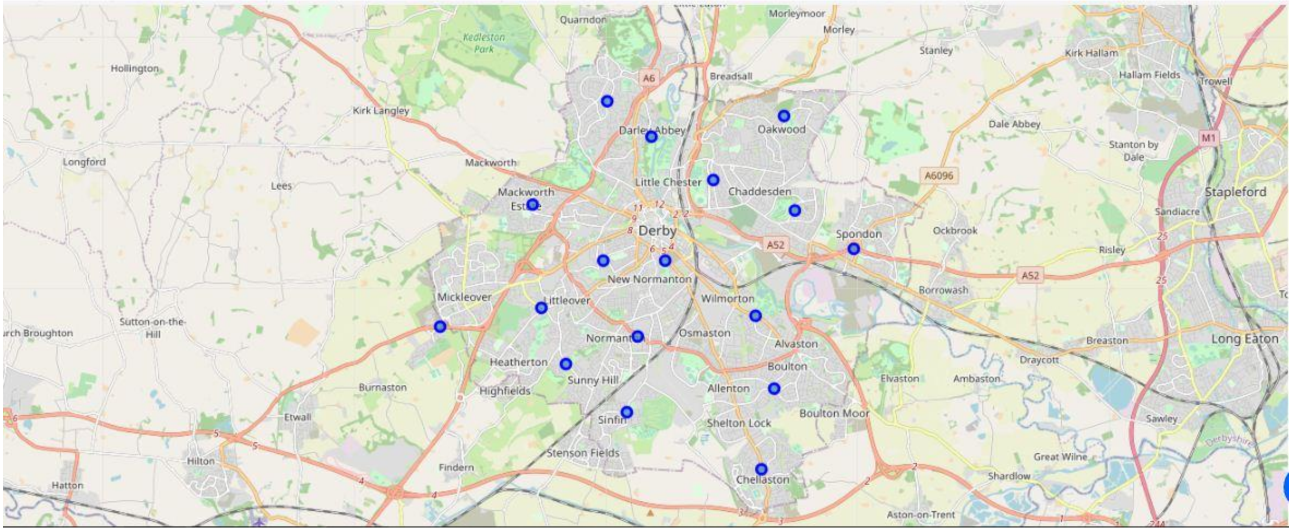


Figure 3. Folium Map indicating the 17 Derby Wards center points.

In order to gain understanding of the gathered information for the 17 wards, a scatter plot was generated comparing the "Median House Price" in each ward with its associated "Crime Ratio for 1000 habitants", using the Matplotlib Pyplot package, as displayed in Figure 3. Figure 3 indicates that the "Arboretum" ward present the highest criminality rates in the city of Derby with a significant difference from the other wards, with a crime ratio with 363.4 crimes for 1000 habitants during May 2019-May 2020. All the remaining 16 wards in Derby presented crime ratios inferior to 143 crimes for 1000 habitants during the same period of time, over 50% less than in "Arboretum".

With regards to Median House prices during Q1 2019, it can be observed a significant variability in the values: "Arboretum" presents the cheapest properties with Median values of £97,000, whereas "Allestree" presents the most expensive properties, with Median values of £247,500, 2.5 times higher than the cheapest ward in Derby.

Also, observing Figure 4, it may indicate a negative correlation between house price and criminality rate.

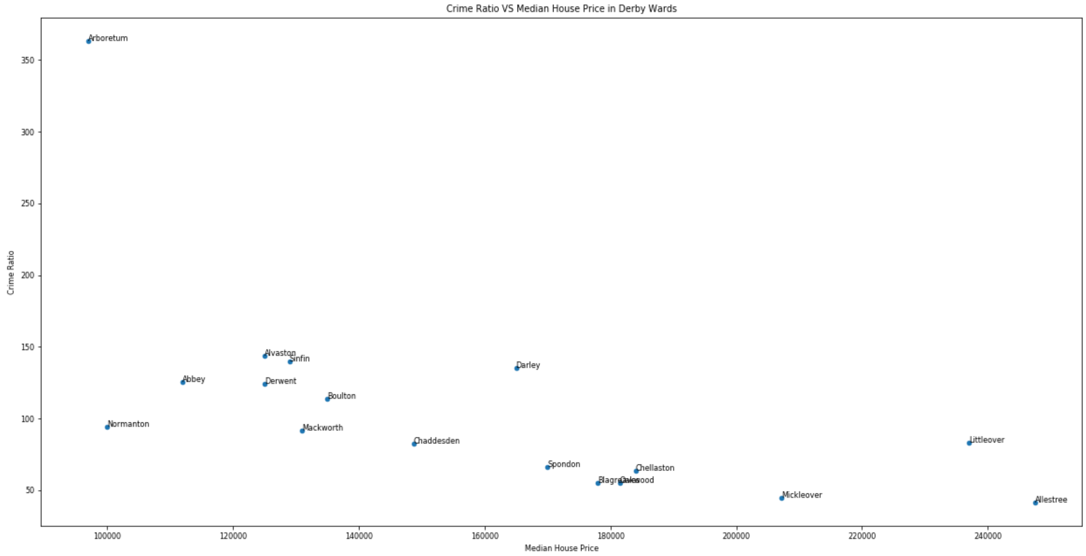


Figure 4. Scatter Plot - Median House Price VS Crime Ratio in Derby Wards

The obtained results using the Scypy package indicate that the relation between "Median House price" and "Crime ratio" presents a Pearson Coefficient (R) of -0.61066 and a p-value of 0.0092 (< 0.05). According to these results, it can be indicated that there is a strong negative correlation between the Median House price and the

Crime Ratio in the various Derby wards, since R is > -0.5, and as the p-value is less than 0.05, the certainty in these results are significant.
In Figure 5, it can be observed a Regression plot of the "Median House price" versus "Crime ratio" for the 17 Derby wards, presenting a scatter plot of the points and the best fit regression line for those variables. This chart has been obtained using the Seaborn package.

Observing the graph below, most of the Derby wards, except "Arboretum", would appear to fit well with the regression line, however the "Arboretum" ward could be considered as an outlier point in terms of crime ratio.

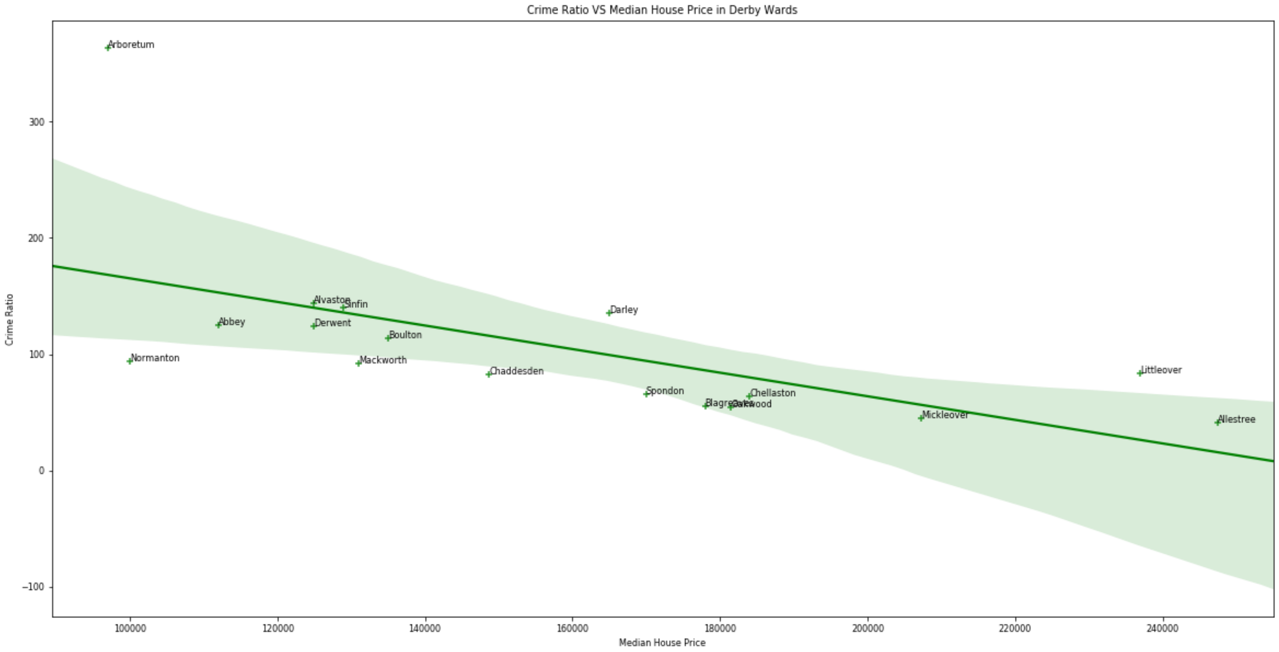


Figure 5. Regression Plot - Median House Price VS Crime Ratio in Derby Wards

After having completed a preliminary data analysis of the "Median House Price" vs "Crime Ratio" in the various Derby wards, Foursquare API was used in order to obtain information of the existing venues in each of the 17 Derby Wards. The settings used in order to obtain this information was:

- Displaying a maximum of 100 venues per ward
- Display venues within a 650m limit from the center coordinates location of each ward.

The outcome of the Foursquare database analysis showed 134 venues across the 17 Derby Wards, grouped into 61 unique venue types. All the venue information obtained from Foursquare was stored into a Pandas Dataframe to allow further analysis.

As it can be observed in Figure 6, the Top 5 most frequent types of venue in Derby across all wards, which are displayed in a Bar Chart using Matplotlib Pyplot are:

- Grocery stores, Pubs, Parks, Clothing stores and Indian restaurants: representing the 29% of all the venues in Derby.

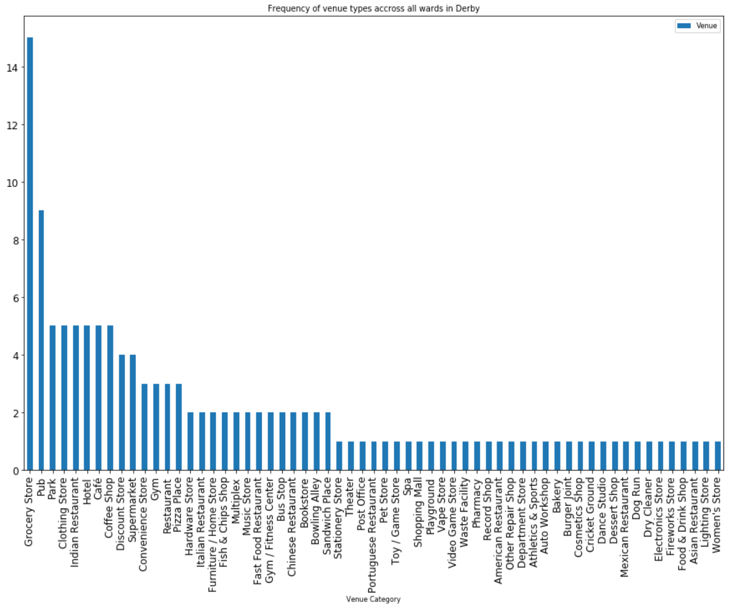


Figure 6. Bar Chart of venue types in Derby

Comparing the number of venues in each of the Derby wards, it is observed that the “Arboretum” ward concentrates most of the city venues: 54 out of 134, representing a 40% of the total. The remaining 16 wards concentrate the remaining 60% of the city venues, and each of those wards presents less than 10 venues. This information is displayed in a Bar Chart in Figure 7.

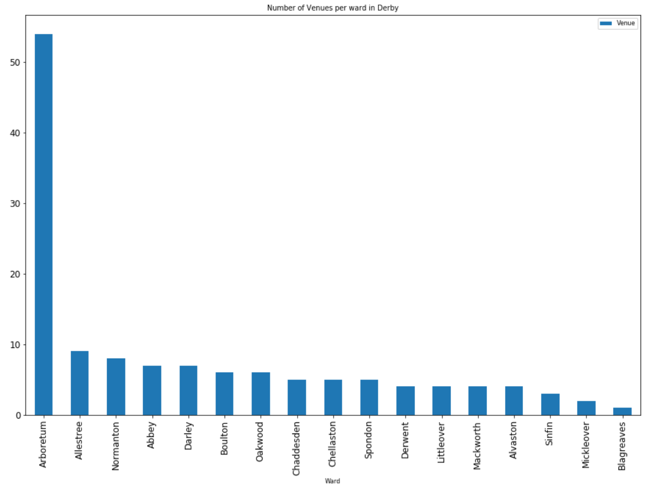
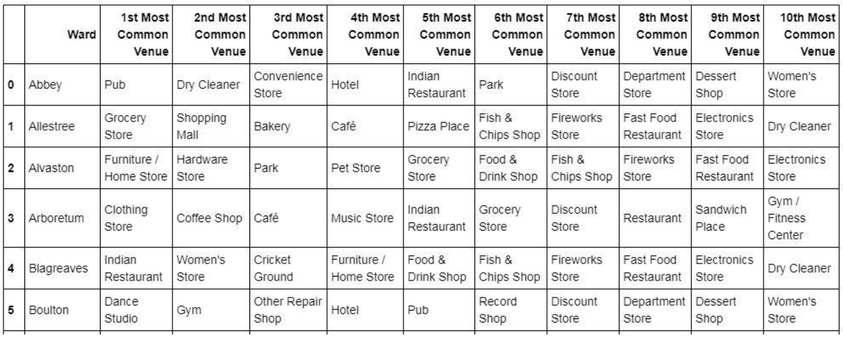


Figure 7. Frequency of venues in Derby by ward

The top 10 most common venues for each ward based on the Foursquare API data for each of the Derby wards was also calculated and stored in a Dataframe, as displayed below in Figure 8.



	Ward	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
0	Abbey	Pub	Dry Cleaner	Convenience Store	Hotel	Indian Restaurant	Park	Discount Store	Department Store	Dessert Shop	Women's Store
1	Allestree	Grocery Store	Shopping Mall	Bakery	Café	Pizza Place	Fish & Chips Shop	Fireworks Store	Fast Food Restaurant	Electronics Store	Dry Cleaner
2	Alvaston	Furniture / Home Store	Hardware Store	Park	Pet Store	Grocery Store	Food & Drink Shop	Fish & Chips Shop	Fireworks Store	Fast Food Restaurant	Electronics Store
3	Arboretum	Clothing Store	Coffee Shop	Café	Music Store	Indian Restaurant	Grocery Store	Discount Store	Restaurant	Sandwich Place	Gym / Fitness Center
4	Blagreaves	Indian Restaurant	Women's Store	Cricket Ground	Furniture / Home Store	Food & Drink Shop	Fish & Chips Shop	Fireworks Store	Fast Food Restaurant	Electronics Store	Dry Cleaner
5	Boulton	Dance Studio	Gym	Other Repair Shop	Hotel	Pub	Record Shop	Discount Store	Department Store	Dessert Shop	Women's Store

Figura 8. Extract of Dataframe indicating top 10 most common venues in each Derby Ward

After, the "One Hot" encoding technique was performed with the Venue types for each ward in order to set each venue type as a column in a Dataframe, and combine this information with the "Median House Price" and "Crime Ratio" in each ward. As each of the variables in the dataframe are expressed in different numerical ranges, the Dataframe was normalized prior to perform the Machine Learning analysis. In this case, the Standard Scaler package from Sklearn was used for normalizing the dataset.

Once the dataset including information regarding the 17 wards in Derby was normalized, an unsupervised Machine Learning technique called K-means was used in order to cluster the wards into different clusters according to:

- The available venues in each area
- The Median House prices
- The Crime Rates.

The k-means algorithm works clustering the dataset into various "k" number of clusters. In order to determine the optimum amount of clusters, or value of "k", both the Elbow method and the Silhouette Score method were applied. In both cases, the methods indicated that the optimum k value was 2, meaning that the 17 wards in Derby would be best divided into 2 clusters – See Figures 9 and 10.

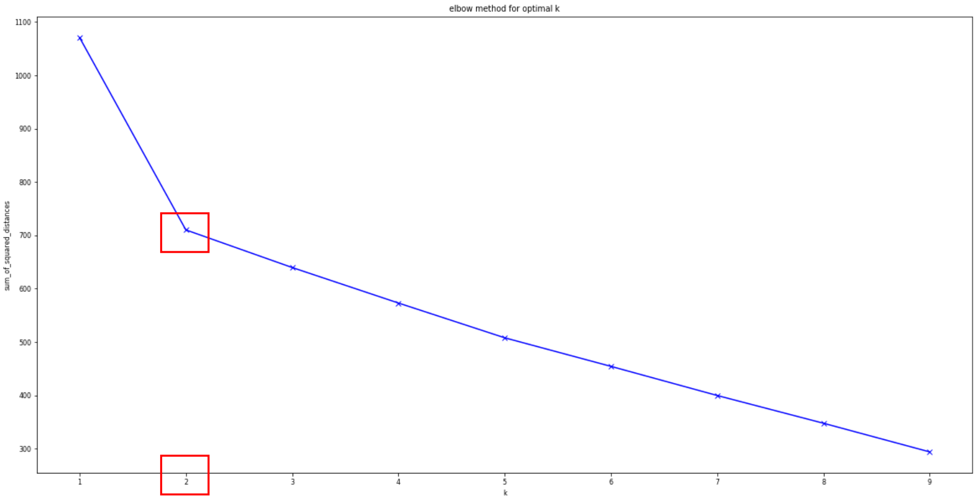


Figure 9. Elbow method to determine optimum k for K-means

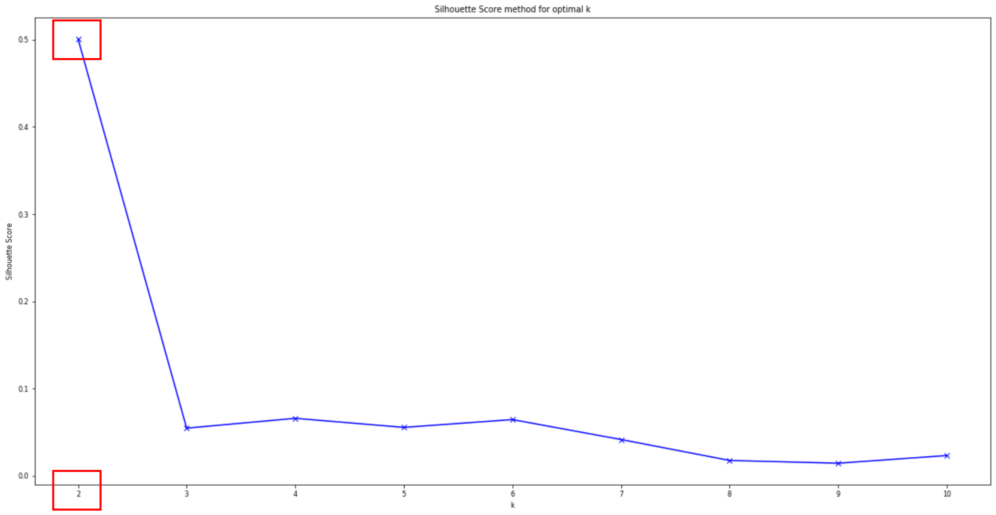
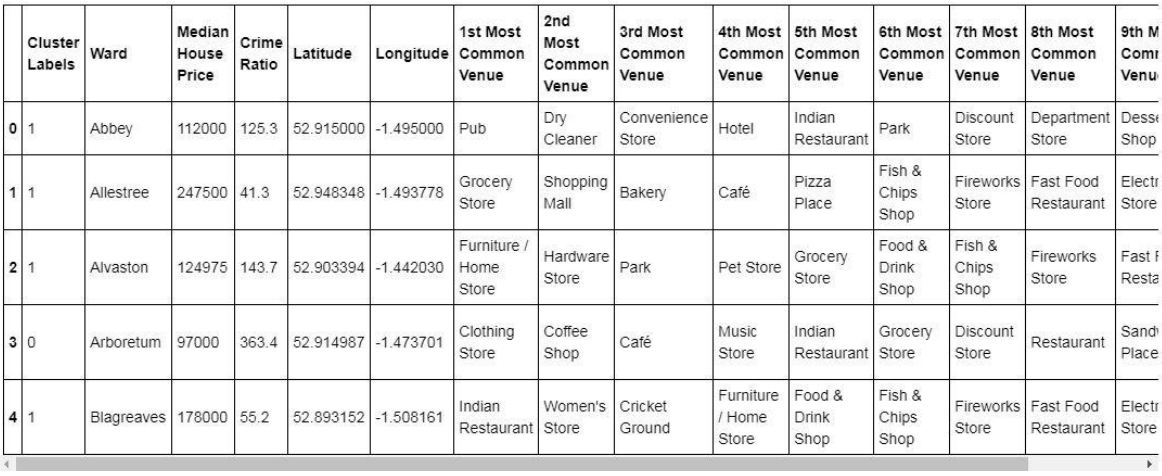


Figure 10. Silhouette Score method to determine optimum k for K-means

After having run the K-means clustering algorithm in the normalized dataset and obtained the associated cluster for each of the Derby wards, this information was added to the information Dataframe for the various wards in the column "Cluster Labels" (See Figure 11).



	Cluster Labels	Ward	Median House Price	Crime Ratio	Latitude	Longitude	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 M Comı Venu
0	1	Abbey	112000	125.3	52.915000	-1.495000	Pub	Dry Cleaner	Convenience Store	Hotel	Indian Restaurant	Park	Discount Store	Department Store	Desse Shop
1	1	Allestree	247500	41.3	52.948348	-1.493778	Grocery Store	Shopping Mall	Bakery	Café	Pizza Place	Fish & Chips Shop	Fireworks Store	Fast Food Restaurant	Electr Store
2	1	Alvaston	124975	143.7	52.903394	-1.442030	Furniture / Home Store	Hardware Store	Park	Pet Store	Grocery Store	Food & Drink Shop	Fish & Chips Shop	Fireworks Store	Fast F Resta
3	0	Arboretum	97000	363.4	52.914987	-1.473701	Clothing Store	Coffee Shop	Café	Music Store	Indian Restaurant	Grocery Store	Discount Store	Restaurant	Sandı Place
4	1	Blagreaves	178000	55.2	52.893152	-1.508161	Indian Restaurant	Women's Store	Cricket Ground	Furniture / Home Store	Food & Drink Shop	Fish & Chips Shop	Fireworks Store	Fast Food Restaurant	Electr Store

Figure 11. Dataframe including information and Cluster labels for each Derby ward.

An updated map of Derby, with its 17 wards categorized into 2 clusters was printed using the Folium package and is displayed in Figure 12. As it can be observed, the "Arboretum" ward represents itself one of the clusters – Cluster 0, whereas the 16 remaining wards in Derby are all included into the Cluster 1. The results section provides a deeper analysis of these results.

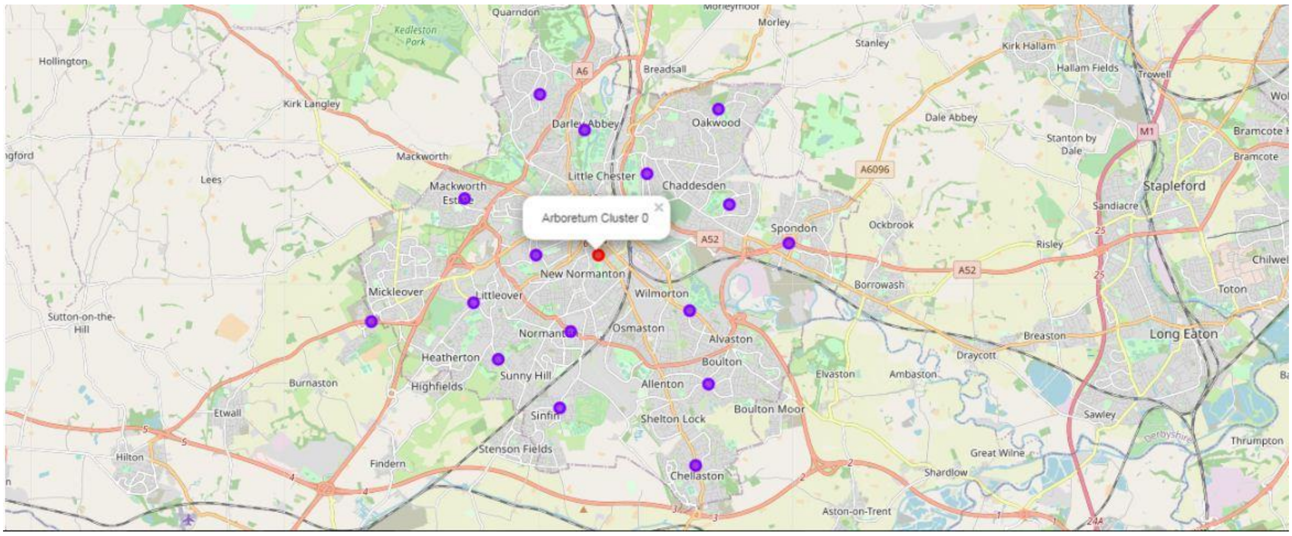


Figure 12. Clusters of Derby wards according to K-means clustering algorithm.

4. Results

After having completed the analysis presented in the Methodology section, it was concluded that according to the K-means clustering algorithm the 17 wards in Derby can be clustered into 2 clusters:

Cluster 0 - including only the "Arboretum" ward. This cluster presents and "Median House Price" of £97,000 and an average crime rate of 363.4.

This cluster gathers 54 out of the 134 venues in Derby according to Foursquare API, and the 5 most frequent types of venues are Clothing Stores, Coffee Shops, Cafes, Gyms/Fitness Centers and Indian Restaurants. All information regarding types of venues on Cluster 0 is displayed in a bar chart in Figure 13.

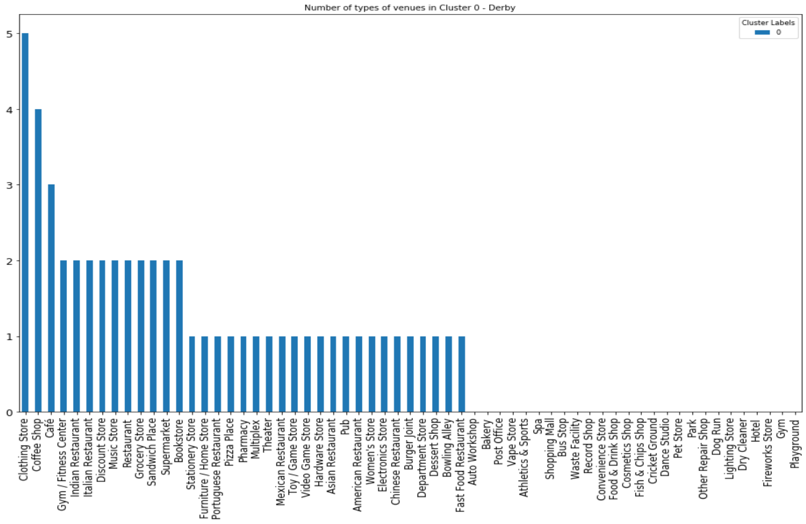


Figure 13. Bar Char - Frequency of venues per type in Cluster 0

Cluster 1 – including the remaining 16 wards in Derby. The properties in each cluster present an average "Median House Price" of £160,996.8 and an average crime rate of 91.18.

The 16 wards in cluster 1 gathers 80 out of the 134 venues in Derby according to Foursquare API, and the 5 most frequent types of venues are Grocery Stores, Pubs, Parks, Hotels and Convenience Stores. All information regarding types of venues on Cluster 1 is displayed in a bar chart in Figure 14.

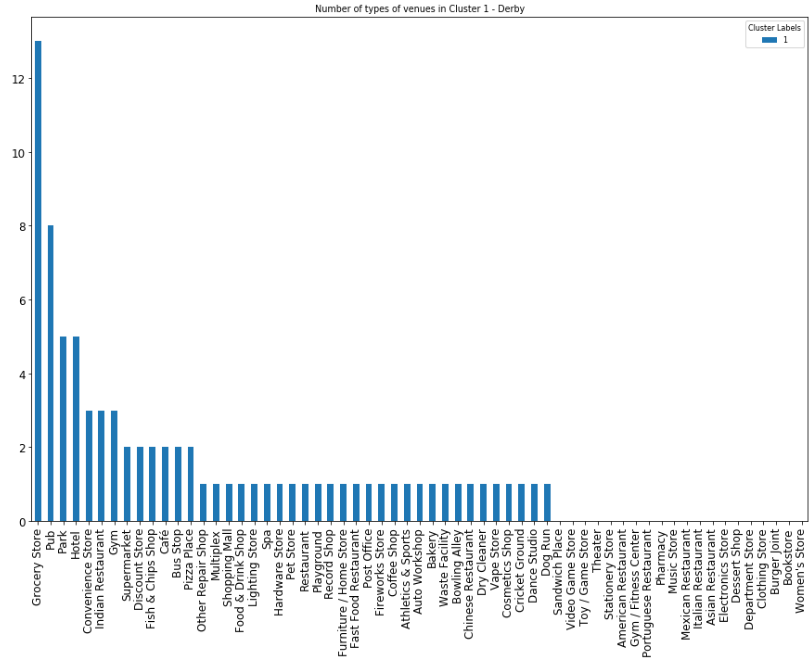


Figure 14. Frequency of venues per type in Cluster 1

Based on the previous lines, it can be stated that the "Cluster 0", which includes the Arboretum ward of the city of Derby presents the lowest price properties in the city as well as the highest criminality rate, with significant difference from that in "Cluster 1". "Cluster 0" concentrates 40% of all the venues in Derby, and seems to host a great variety of leisure related venues – from clothes stores, cafes, gyms and restaurants. Based on this, "cluster 0" could be categorized as the main shopping and hospitality district in the city.

"Cluster 1", which includes the remaining 16 wards of the city of Derby, includes the remaining 60% of the venues. Observing the type of venues: mostly supermarkets, parks and pubs, it could be said that these wards could represent the quieter and more residential areas of the city, with lower criminality rates, and higher house prices than if compared with "Cluster 0".

5. Discussion & recommendations

As indicated in the previous sections, the 17 wards in which the city of Derby is divided can be classified into 2 main clusters:

- **Cluster 0**, which covers the "Arboretum" ward which includes most of the leisure offers in the city, and presents the cheapest properties and the highest criminality rate.
- **Cluster 1**, which covers the remaining 16 wards, and represent the more residential and quieter areas of the city, with lower criminality rates and more expensive properties.

Those people considering to rent or buy a property in the city of Derby can use this information to decide where to find a suitable property. In the case of a potential buyer looking for a safe and quiet area to settle down, the information in this report would suggest to avoid the area included in "Cluster 0", and focusing his/her search in the areas included in "Cluster 1".

On contrary, for a potential buyer or renter interested in having a great variety of leisure venues in the near proximity, and is not particularly bothered about criminality rates, the information in this report would suggest him/her to focus the search in the area included in “Cluster 0”.

A potential interesting analysis to perform next, would be to further understand the differences amongst the 16 wards included in the “Cluster 1”, and categorized as Residential areas. In order to do this, it would be suggested to repeat the methodology included in Sections 3-4 to cluster the 16 wards classed as residential areas to further understand whether those can be clustered in smaller groups.

6. Conclusion

This report is focused on analyzing and comparing the 17 wards in which the city of Derby according to the median house prices in each ward during Q1 2019, the criminality rate during May 2019-May 2020 in each ward and the existing venues in each area.

Initially, it was possible to statistically estimate that there is a negative linear relationship between the crime rates and the median house prices in the various Derby wards, and a linear regression line can be fitted to explain this relationship. Those wards in which the median house prices are the lowest present the highest criminality rates.

In addition, this report analyzed and gathered the existing venues in the 17 Derby wards according to the data presented in the Foursquare API. A k-Means clustering analysis was performed across the 17 Derby wards considering the normalized information regarding available venues, Crime Rate and Median house prices in each of them.

An optimum value of 2 clusters allowed to divide the Derby wards into 2 main groups: one cluster containing most of the leisure venues in the city, with the lowest house prices and the highest criminality rates, and a second cluster containing the residential areas of the city, which present a reduced number of venues per ward, higher house prices and lower criminality rates.

The information of this report can be considered as useful and informative for potential home buyers or renters, or people moving to Derby in order to make informed decisions about where to search on a property based on their personal preferences and the main characteristics of the wards in Derby.