Glasgow's Music Hotspot
Coursera IBM Capstone Project

A. Introduction

The basis of this study is to help a group of investors source a suitable location for a new music hub location in Glasgow, Scotland.

Glasgow is Scotlands largest metropolis, with nearly 600,000 residents located in its central region and up to 1.8 million living in the surrounding suburbs. It's long been considered the prominent cultural hot spot in not just Scotland, but the UK. This was supported recently with the announcement that Glasgow is the UK's first UNESCO City of Music and UK's top cultural and creative city (European Commission 2019).

From gigs in small venues and bars to incredible shows in arenas and city parks, Glasgow is a hotspot for live performance. Whether bagpipes or beats, traditional or techno, Glasgow is not afraid to make some noise.

For this reason, Ive decided to investigate the ideal location to open a new music hub. A one stop venue where musicians come to be educated, rehearse and perform their own music.

Venue Location Criteria

- It's important that this location is suitable to draw in the maximum number of people, therefore amenities such as public transport and restaurants/cafes/bars would ideally be available.
- This venue will require a building with a large sqft area. This might rule out certain central locations due to size constraints.
- Lastly, information about rental and property price data will be important. An up and coming location, slightly outside the city centre may be more affordable and therefore more preferable.

Tools and Libraries for Analysis

To help choose a location for the hub, I will analyse the various districts and neighbourhoods with these tools:

- Beautiful Soup: Used to scrape website data for further processing
- Pandas: Create data frames for easy integration and manipulation of data
- Scikit-learn — KNN imputer class to fill in missing values based on nearest neighbours
- Geopy Geocoders: To convert address into latitude and longitude values using Nominatim
- Matplotlib: For plotting various graphs
- Folium Map: Visualise your data on a Leaflet map
- Choropath Map: Visualise how rental prices vary across a Glasgow neighbourhoods on a thematic map using colour variations to highlight differences

- K-means Clustering: Unsupervised learning technique. Creates clusters of similar neighbourhoods from the scraped data. This information will be then highlighted and clearly displayed on the folium map. It's this technique that will allow me to make informed decisions on best location candidates that fit the criteria set above.

B. Data

The data used to analyse neighbourhoods was sourced from several locations using web scraping or via an API.

GLASGOW POSTAL CODES *https://en.wikipedia.org/wiki/G_postcode_area*

Scraped from wikipedia using the beautiful soup library and subsequently cleaned up to label the folium map with accurate district information.

GEO JSON DATA - *https://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/scottish-postcode-directory/2021-1*

I found this data freely available from the National Records of Scotland website. It was used to create a choropleth map using geo json data of each post code area. I then superimposed a colormap of variances in rent prices per neighbourhood.

PROPERTY RENTAL DATA *www.citylets.co.uk*

Unfortunatley, I never found a data source for commercial property rental prices per postcode area. I did however find a dataset containing average rental prices of properties by bedroom number. Although not exactly what i was looking for, I believed it would act as an accurate indicator of price variation between postcodes none the less.

There were a few missing values in this dataset which I had to account for using imputation. I discuss this further in the methodology section.

FOURSQUARE LOCATION DATA

Using the API service, this will provide useful location data including — local amenities i.e popular shops, cafes and also transport links.

C. Methodology

Create Initial Data Frame

To create my initial data frame, I scraped postcode data using the beautiful soup library. I then went on to clean and remove any unnecessary characters and postcodes that were out with the city limits.

	Postcode District	Coverage
0	G1	Merchant City
1	G2	Blythswood Hill, Anderston (part)
2	G3	Anderston, Finnieston, Garnethill, Park, Woodl...
3	G4	Calton (part), Cowcaddens (part), Drygate, Kel...
4	G5	Gorbals

Initial data frame

Acquire Longitude and Latitude Data with Geo Locator

Next, I used the postcodes from the table to search for the longitude and latitude information using the geo locator library.

I had some minor issues initially when some postcodes returned incorrect locations e.g certain postcodes having coordinates in the Philippines!

After messing around with the address parameters however, I finally had all the geo data and added it to the data frame.

	Postcode District	Coverage	Lat	Long
0	G1	Merchant City	55.859773	-4.252307
1	G2	Blythswood Hill, Anderston (part)	55.863471	-4.258902
2	G3	Anderston, Finnieston, Garnethill, Park, Woodl...	55.867193	-4.273059
3	G4	Calton (part), Cowcaddens (part), Drygate, Kel...	55.871637	-4.253744
4	G5	Gorbals	55.843056	-4.242441

Data frame with latitude and longitude data

Scrape Rent Prices

As mentioned earlier, I couldn't find a working dataset with typical commercial rental prices per postcode.

Instead, I settled with a dataset containing the 2019 average prices for renting a 2 bedroom property.

I finished of the data frame with the price data attached.

	Postcode District	Coverage	Lat	Long	Price (£)
0	G1	Merchant City	55.859773	-4.252307	1011.0
1	G2	Blythswood Hill, Anderston (part)	55.863471	-4.258902	951.0
2	G3	Anderston, Finnieston, Garnethill, Park, Woodl...	55.867193	-4.273059	880.0
3	G4	Calton (part), Cowcaddens (part), Drygate, Kel...	55.871637	-4.253744	824.0
4	G5	Gorbals	55.843056	-4.242441	720.0
5	G11	Broomhill, Partick, Partickhill	55.871535	-4.302969	874.0
6	G12	West End (part), Cleveden, Dowanhill, Hillhead...	55.883231	-4.294669	961.0
7	G13	Anniesland, Knightswood, Yoker	55.883533	-4.327100	665.0
8	G14	Whiteinch, Scotstoun	55.880055	-4.352957	590.0
9	G15	Drumchapel	55.911237	-4.360579	NaN
10	G20	Maryhill, North Kelvinside, Ruchill	55.882537	-4.272827	722.0
11	G21	Balornock, Barmulloch, Cowlairs, Royston, Spri...	55.884370	-4.228781	547.0
12	G22	Milton, Parkhouse, Possilpark	55.881005	-4.255906	NaN

Data frame with rental prices included. Notice the NaN values.

Impute Missing Price Values

K-nearest neighbour is an algorithm that is useful for matching a point with its closest (k) neighbours in a multi-dimensional space.

It can be used for data that are continuous, discrete, ordinal and categorical which makes it particularly useful for dealing with all kind of missing data.

Since there was a large number of missing values, I decided not to simply remove all the districts with missing values. Instead, I landed on KNN imputation, witch out of the various other imputing techniques, including replacing values with mean, mode etc, I concluded that KNN imputation would be best method to provide good rent price estimates, for each neighbourhood, based on the the prices of its surrounding postcode areas.

Sci-kit learns KNN imputer class made this job simple using a smallish k size of 4 and setting the weights parameter to ‘distance’. This would weigh neighbourhoods closest more highly than those further away.

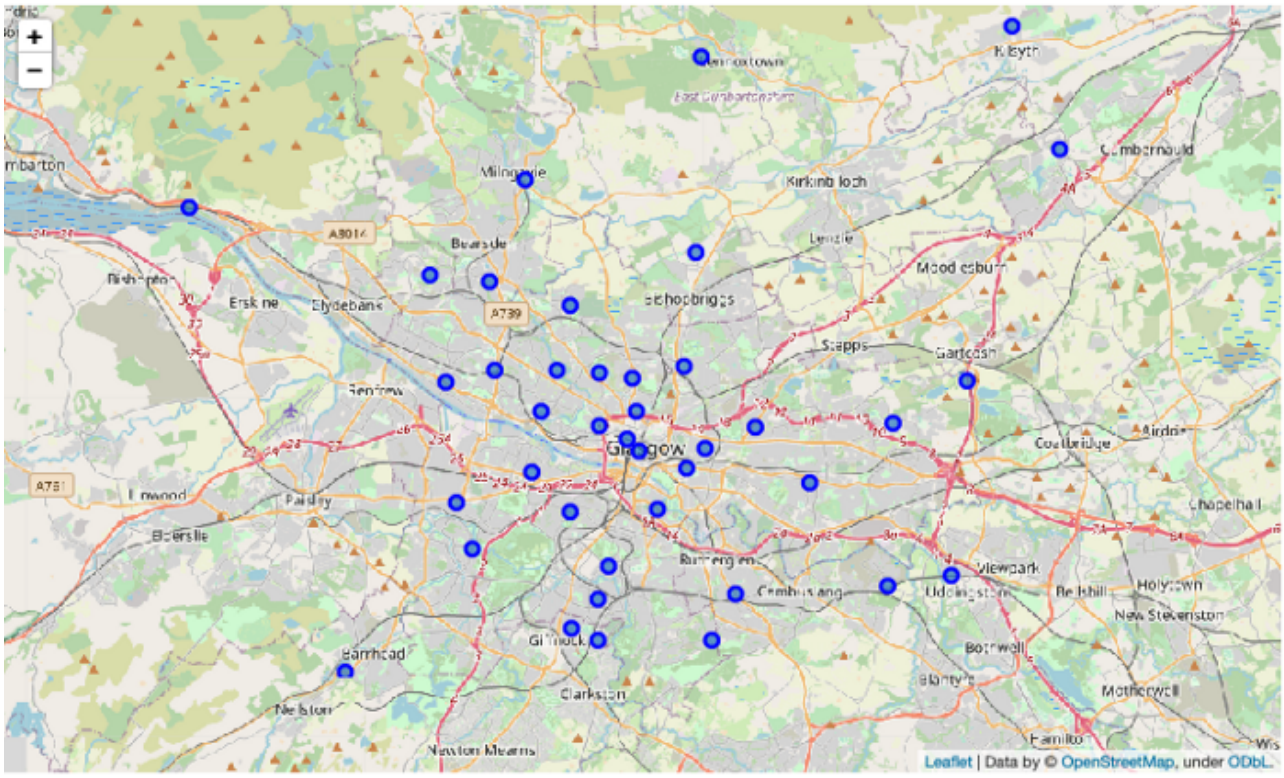
This leaves us with our complete data frame we can use for further analysis.

	Postcode District	Coverage	Lat	Long	Price (£)
0	G1	Merchant City	55.859773	-4.252307	1011
1	G2	Blythswood Hill, Anderston (part)	55.863471	-4.258902	951
2	G3	Anderston, Finnieston, Garnethill, Park, Woodl...	55.867193	-4.273059	880
3	G4	Calton (part), Cowcaddens (part), Drygate, Kel...	55.871637	-4.253744	824
4	G5	Gorbals	55.843056	-4.242441	720
5	G11	Broomhill, Partick, Partickhill	55.871535	-4.302969	874
6	G12	West End (part), Cleveden, Dowanhill, Hillhead...	55.883231	-4.294669	961
7	G13	Anniesland, Knightswood, Yoker	55.883533	-4.327100	665
8	G14	Whiteinch, Scotstoun	55.880055	-4.352957	590
9	G15	Drumchapel	55.911237	-4.360579	676
10	G20	Maryhill, North Kelvinside, Ruchill	55.882537	-4.272827	722
11	G21	Balornock, Barmulloch, Cowlairs, Royston, Spri...	55.884370	-4.228781	547
12	G22	Milton, Parkhouse, Possilpark	55.881005	-4.255906	876

Complete data frame with added imputed price data

Create Folium Map

To further understand and visualise the Glasgow neighbourhoods I utilised the folium library to build an interactive map with markers of each postcode location.



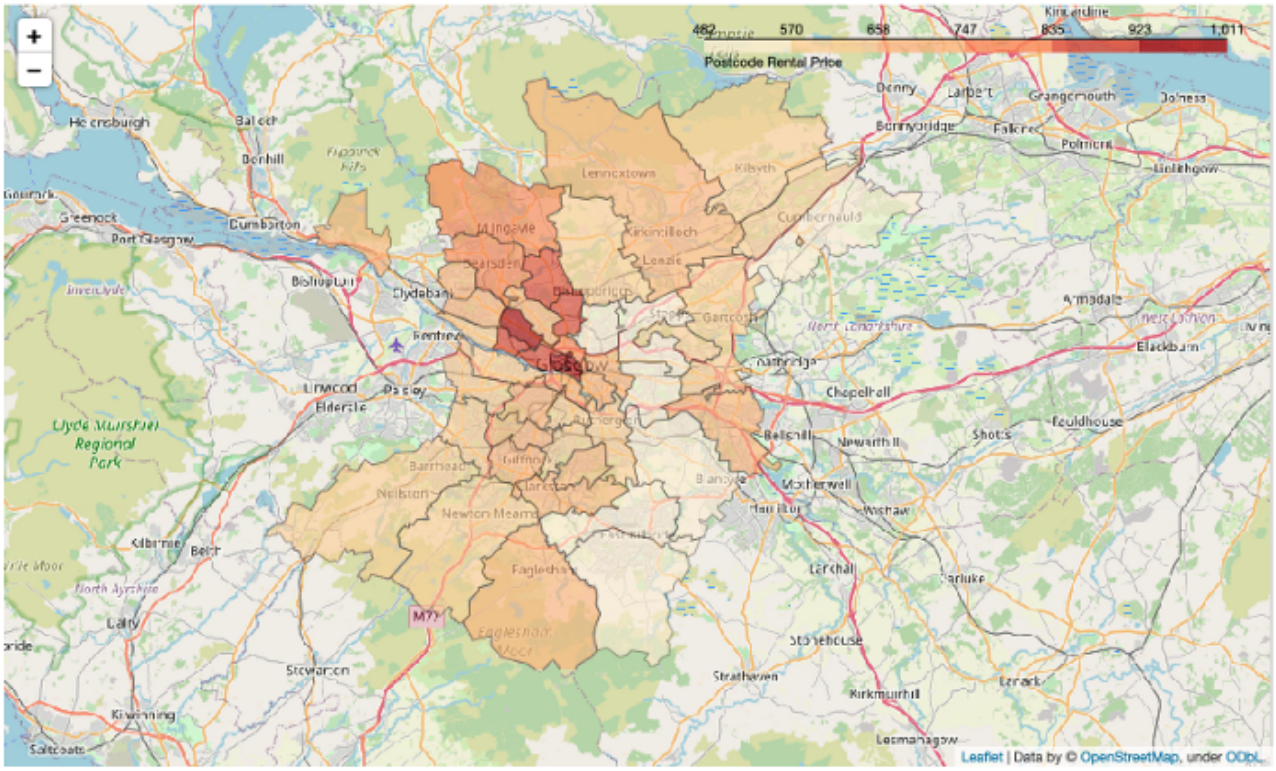
Foursquare API

Finally, using the foursquare API I searched for a list of venues in a 1000m radius of each postcode district. I then created a data frame out of all the venues in each area.

D. Data Analysis

Comparing Average Rent Prices For Each Neighbourhood

To visualise the rent price variation per neighbourhood, I created a choropleth map. This is a great tool to quickly see which area had the highest and lowest rent prices.



Choropleth map of Glasgow rental prices

For example, in dark red is the most expensive neighbourhoods to rent a property. These include the city centre (G1, G2) the west end (G3, G12) and some of the western suburbs (G61, G62).

Explore Neighbourhood Venues

To analyse the various amenities available to each postcode area, I grouped the venues by category type and counted the total categories per neighbourhood. There were 198 unique venues in the city.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Anderston, Finnieston, Garnethill, Park, Woodlands (part), Yorkhill	100	100	100	100	100	100
Anniesland, Knightswood, Yoker	32	32	32	32	32	32
Arden, Carnwadric, Deaconsbank, Giffnock, Kennishead, Thornliebank, northeast Newton Mearns	27	27	27	27	27	27
Auldhouse, East Kilbride (south)	6	6	6	6	6	6
Baillieston, Bargeddie, Chryston, Garrowhill, Gartcosh, Gartloch, Moodiesburn, Muirhead, Springhill	5	5	5	5	5	5
Baldernock, Milngavie, Mugdock	23	23	23	23	23	23
Balornock, Barmulloch, Cowlairs, Royston, Springburn, Sighthill	11	11	11	11	11	11
Barrhead, Neilston, Uplawmoor	8	8	8	8	8	8
Battlefield, Govanhill, Mount Florida, Strathbungo (part), Toryglen	43	43	43	43	43	43
Bearsden	13	13	13	13	13	13
Birkenshaw, Bothwell, Broomhouse, Tannochside, Uddingston, Viewpark	9	9	9	9	9	9
Bishopbriggs, Torrance	14	14	14	14	14	14

Total count of unique venue categories per neighbourhood

The amenity rich areas can be seen with 100+ unique venues. This is mostly in city centre areas as expected.

One Hot Encoding

One hot encoding is a technique used to encode categorical data (here the venue categories in each of the neighbourhoods) to a binary format for use in ML modelling.

It creates a sparse data frame, taking every venue category variable and allocating it a 1 if present in that specific neighbourhood and 0 if absent.

	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Asian Restaurant	Athletics & Sports	Auto Garage	Automotive Shop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Bed & Breakfast	Beer Bar	Beer Store
0	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Merchant City	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	Merchant City	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

One hot encoding table

This was then further processed by grouping the rows by neighbourhood and calculating the mean frequency of occurrence for each venue category.

Basically it allows us now to notice the most common venue types in each neighbourhood.
For any investors this is useful information to understand as it clearly illustrates the overall character of an area. For example, is it a cosmopolitan area with cafes and bars? An outskirts area with shopping centres, or an industrial area etc?

We processed this data and presented it in a data frame with the top 5 venue categories per neighbourhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Anderston, Finnieston, Garnethill, Park, Woodl...	Bar	Restaurant	Indian Restaurant	Café	Coffee Shop
1	Anniesland, Knightswood, Yoker	Café	Platform	Discount Store	Train Station	Pub
2	Arden, Carnwadric, Deaconsbank, Giffnock, Kenn...	Supermarket	Pizza Place	Train Station	Deli / Bodega	Park
3	Auldhouse, East Kilbride (south)	Department Store	Athletics & Sports	Health & Beauty Service	Supermarket	Café
4	Baillieston, Bargeddie, Chryston, Garrowhill, ...	Construction & Landscaping	Chinese Restaurant	Train Station	Gift Shop	Women's Store
5	Baldernock, Milngavie, Mugdock	Bar	Hotel	Supermarket	Plaza	Café
6	Balornock, Barmulloch, Cowlairs, Royston, Spri...	Train Station	Fast Food Restaurant	Discount Store	Soccer Field	Pawn Shop
7	Barrhead, Neilston, Uplawmoor	Grocery Store	Fast Food Restaurant	Electronics Store	Supermarket	Business Service
8	Battlefield, Govanhill, Mount Florida, Strathb...	Café	Pub	Italian Restaurant	Grocery Store	Construction & Landscaping
9	Bearsden	Platform	Grocery Store	Fast Food Restaurant	Food & Drink Shop	Pharmacy

Top venue categories per neighbourhood

K-Means Clustering

K-means is an unsupervised learning technique, that is it doesn't require any labelling of the data before use. Its goal is to group similar instances into clusters.

Scaling Price Data

I first had to scale the price data using Sci-kit learns StandardScaler method.

Standardisation is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

Once this was done, I added it into the one hot encoded data frame with all the venue category data. We now have our data set ready for K-means clustering.

Silhouette Score

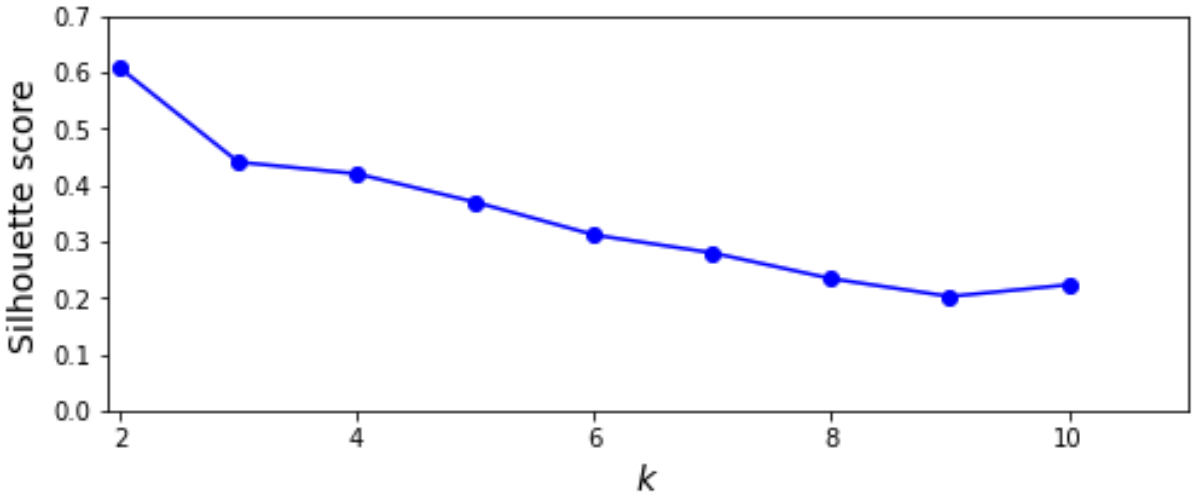
It's no easy task finding the optimum number of clusters to accurately represent your data, but there are some tools that help make a decision.

To help we use a metric called a silhouette score.

A silhouette score varies between -1 and 1, with close to 1 indicating instances are well inside their cluster and close to 0 meaning they are on the cluster boundary. Close to -1 means they are likely in the wrong cluster.

By iterating through different numbers of clusters (k) we can plot a graph to highlight the 'elbow' or drop off, where adding more clusters is unlikely to help.

```
The silhouette score for 2 is: 0.6076404431248543

The silhouette score for 3 is: 0.4405248709164939

The silhouette score for 4 is: 0.4197892700422071

The silhouette score for 5 is: 0.36938616402641006

The silhouette score for 6 is: 0.31129560129994294

The silhouette score for 7 is: 0.2793603942067682

The silhouette score for 8 is: 0.2336777554421722

The silhouette score for 9 is: 0.20208431498684515

The silhouette score for 10 is: 0.2228253542438868
```



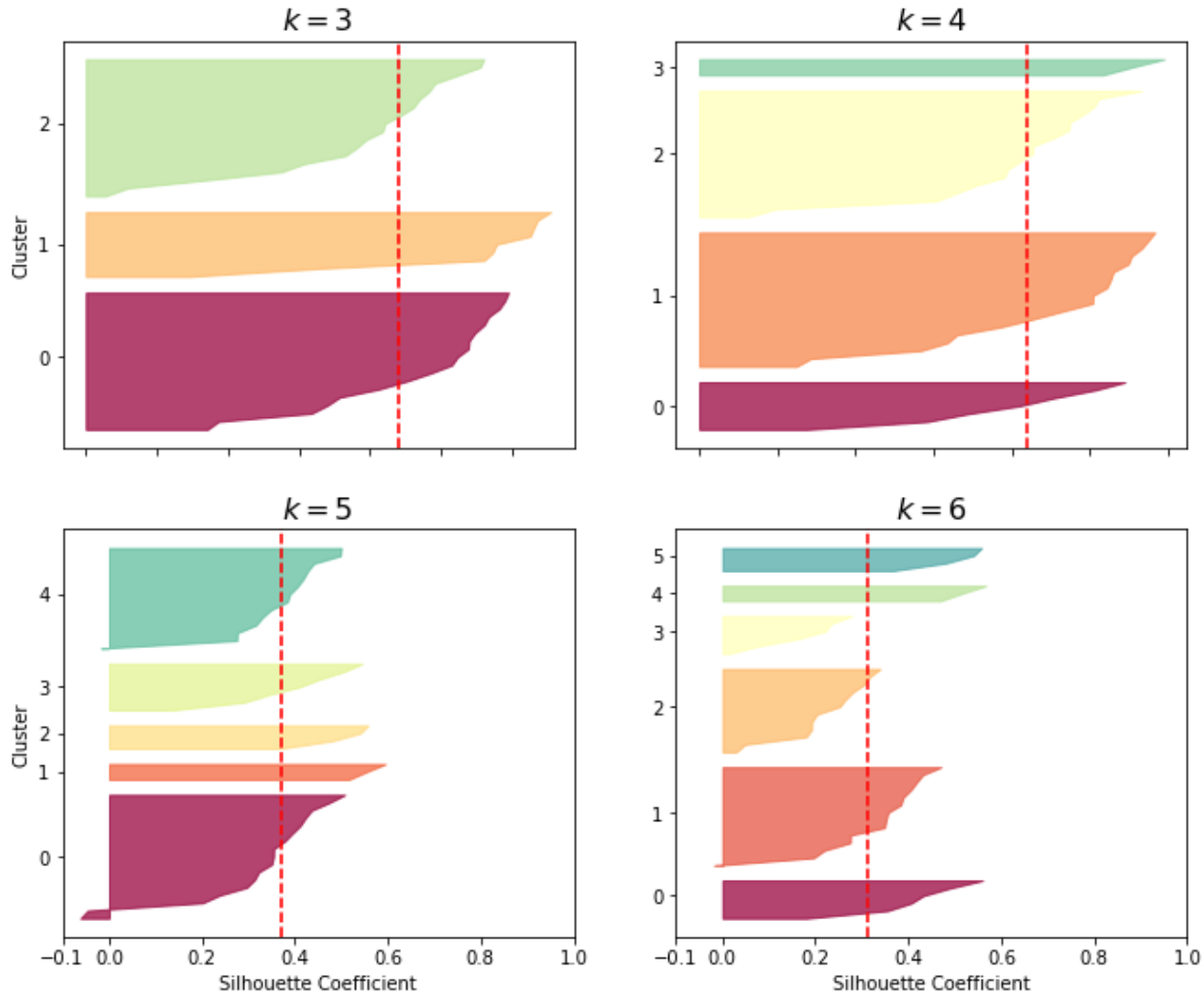
Silhouette Scores. Notice the elbow at cluster 3 where it dips and begins to level out. This is a good indicator of the ideal cluster.

Silhouette Diagram

I felt the graph above wasn't overly helpful in making a call on the cluster numbers. So i decided to plot a silhouette diagram.

From the graphs below , the height of each bar represents the number of instances in each cluster, and the width is the sorted silhouette coefficients of the instances in each cluster (wider is better).

The dashed line is the silhouette score, or the mean silhouette coefficient.
We want all clusters to be beyond this line as that means each instance is contained far from the boundary of the cluster.



We can see from the diagram clusters 3, 4, 5 look most promising.

Ive decided to settle on cluster 4, since the clusters are more evenly shaped and all reach over the dotted line.

Fitting the Data

After fitting the K means algorithm to the data. Finally i added the cluster column to our the data frame, organising it by these clusters.

This is our final data frame.

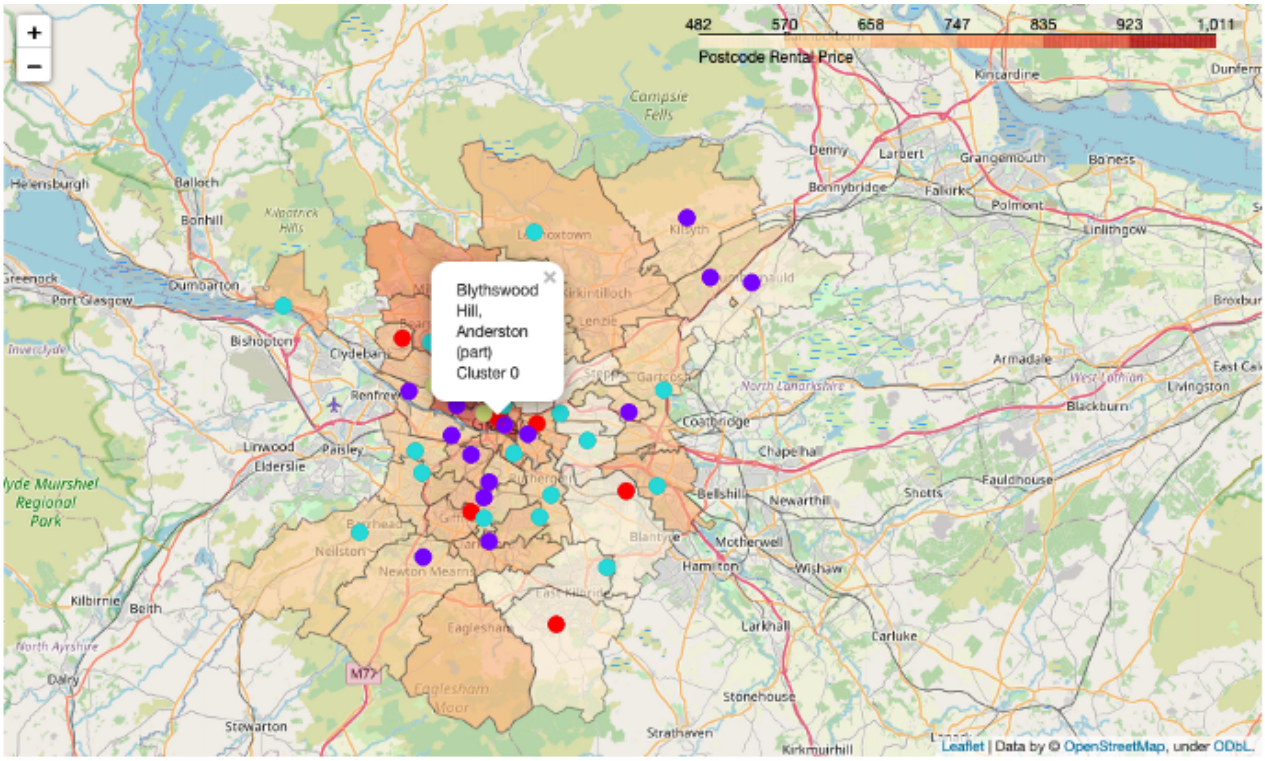
	Postcode District	Coverage	Lat	Long	Price (£)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	G1	Merchant City	55.859773	-4.252307	1011	1	Coffee Shop	Italian Restaurant	Bar	Café	Pub
1	G2	Blythswood Hill, Anderston (part)	55.863471	-4.258902	951	0	Bar	Hotel	Cocktail Bar	Coffee Shop	Italian Restaurant
2	G3	Anderston, Finnieston, Garnethill, Park, Woodl...	55.867193	-4.273059	880	3	Bar	Restaurant	Indian Restaurant	Café	Coffee Shop
3	G4	Calton (part), Cowcaddens (part), Drygate, Kel...	55.871637	-4.253744	824	2	Hotel	Bar	Italian Restaurant	Coffee Shop	Chinese Restaurant
4	G5	Gorbals	55.843056	-4.242441	720	2	Dance Studio	Gym	Flea Market	Stadium	Garden
5	G11	Broomhill, Partick, Partickhill	55.871535	-4.302969	874	1	Café	Bar	Italian Restaurant	Pub	Indian Restaurant
6	G12	West End (part), Cleveden, Dowanhill, Hillhead...	55.883231	-4.294669	961	1	Bar	Coffee Shop	Restaurant	Grocery Store	Italian Restaurant
7	G13	Anniesland, Knightswood, Yoker	55.883533	-4.327100	665	3	Café	Platform	Discount Store	Train Station	Pub
8	G14	Whiteinch, Scotstoun	55.880055	-4.352957	590	1	Clothing Store	Pharmacy	Fast Food Restaurant	Coffee Shop	Jewelry Store
9	G15	Drumchapel	55.911237	-4.360579	676	0	Construction & Landscaping	Discount Store	Pharmacy	Train Station	Supermarket
10	G20	Maryhill, North Kelvinside, Ruchill	55.882537	-4.272827	722	2	Grocery Store	Café	Pizza Place	Pub	Fast Food Restaurant
11	G21	Balornock, Barmulloch, Cowlairs, Royston, Spri...	55.884370	-4.228781	547	3	Train Station	Fast Food Restaurant	Discount Store	Soccer Field	Pawn Shop
12	G22	Milton, Parkhouse, Possilpark	55.881005	-4.255906	876	1	Supermarket	Discount Store	Harbor / Marina	Gas Station	Canal

Final data frame with clusters, prices and top venue categories

E. Results

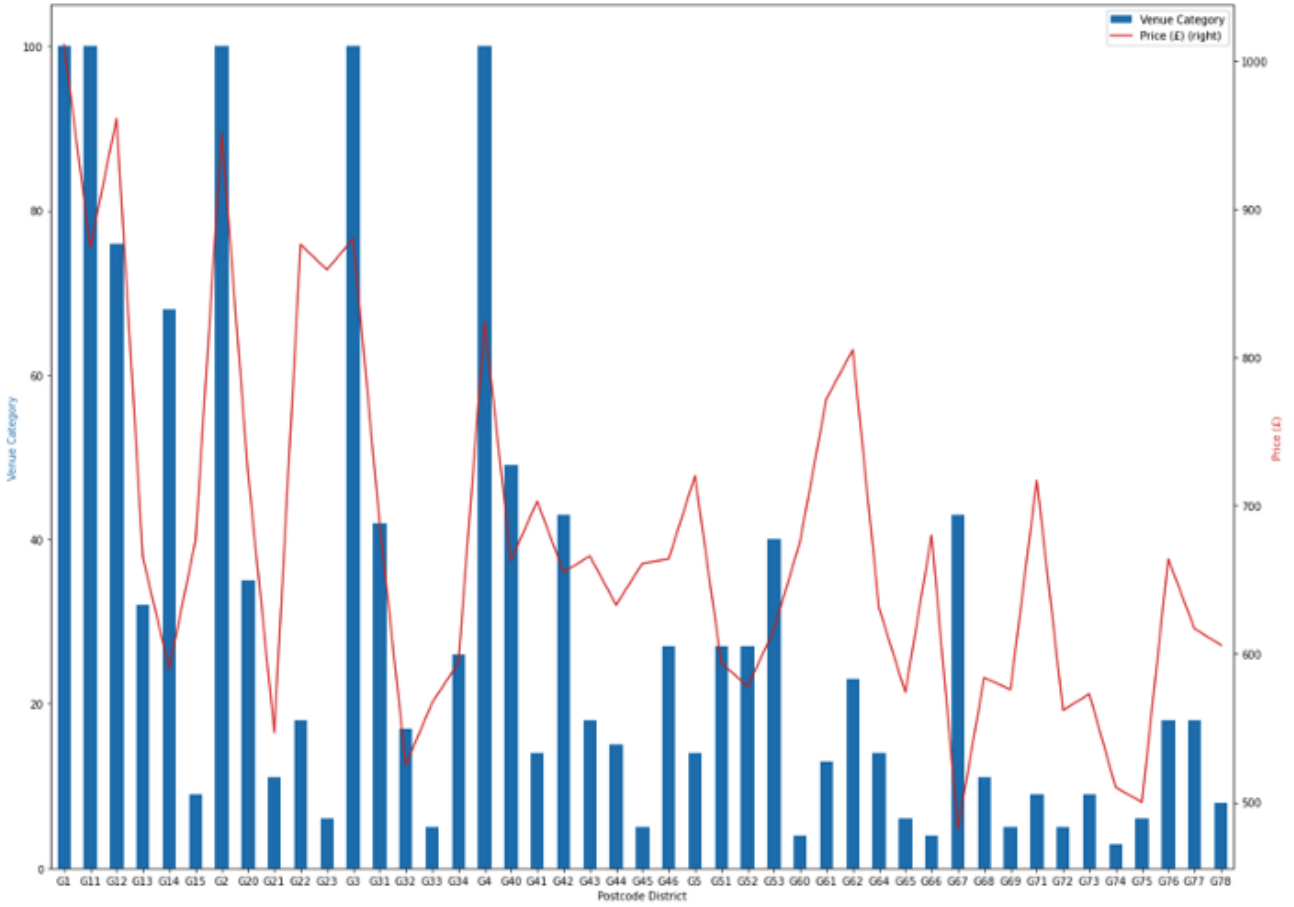
To best visualise the neighbourhood allocation to each cluster, I added them to the choropleth map.

Now together with the price colormap we can better inform investors in potential venue candidates.



Unfortunately, I struggled to find any clear differences between the clusters, with many sharing similar features. This made it hard to draw any clear conclusions from the clustering alone.

To help better choose candidates based on our initial criteria of high amenities/low rental costs, I plotted a graph to help spot the best matches.



Graph of Number of Venue Categories, Rental prices per Postcode District

From this plot we can clearly see a correlation between postcodes with high number of venue categories (local amenities) and high average rent price.

Our goal was to find locations with both the highest number of amenities for the lowest rent price. From the plot, my top choices include:

- G14 — Whiteinch, Scotstoun,
- G40 — Bridgeton, Calton, Dalmarnock
- G42 — Battlefield, Govanhill, Mount Florida,
- G67 — Cumbernauld (south)
- G53 — Darnley, Pollok, Crookston

The venue category totals together with prices are displayed in the table ‘Top picks’ below.

	Postcode District	Coverage	Lat	Long	Price (£)	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
8	G14	Whiteinch, Scotstoun	55.880055	-4.352957	590	1	Clothing Store	Pharmacy	Fast Food Restaurant	Coffee Shop	Jewelry Store
18	G40	Bridgeton, Calton, Dalmarnock	55.854865	-4.227452	663	1	Brewery	Pharmacy	Supermarket	Train Station	Bakery
20	G42	Battlefield, Govanhill, Mount Florida, Strathb...	55.826175	-4.267864	655	1	Café	Pub	Italian Restaurant	Grocery Store	Construction & Landscaping
27	G53	Darnley, Pollok, Crookston, Parkhouse, Priesth...	55.831183	-4.339166	616	2	Clothing Store	Coffee Shop	Cosmetics Shop	Burger Joint	Italian Restaurant
34	G67	Cumbernauld (south)	55.944794	-3.988847	482	1	Clothing Store	Discount Store	Fast Food Restaurant	Supermarket	Pizza Place

Top locations based on high amenities and low rent prices

F. Discussion

The aim of this investigation was to highlight suitable areas to plan a music hub in Glasgow.

I hoped the K-means clustering algorithm would have return a list of similar districts with common amenities and prices to help answer this question. But instead no strong conclusions could be drawn based on the generic clusters produced.

As with most machine learning algorithms, poor data in, leads to poor data out. Limitations including not finding proper commercial property rental prices and having to impute missing values in what price data we had, could have contributed to the poor modelling results.

Due to the lack of clear differences in the clusters, the only safe conclusions we could gather from the data was from the plot of amenities vs price and manually searching for best candidates.

I guess in this case the simplest solution is the best.

Lastly the choropleth maps proved to be a useful visualisation tool to compare rental prices for each neighbourhood at a glance. If the clustering had yielded significant similar neighbourhoods, plotted on the map would have been a great tool to quickly see possible venue suiters.

G. Conclusion

This project attempted to find a solution to a hypothetical business problem utilising some of the current data science tools and techniques.
Although in this occasion a perfect solution was not found, these techniques when used alongside good data can produce accurate results which would be highly beneficial to anyone looking to gain insights into selected city districts and neighbourhoods in future projects.