# Moving from a smaller town to a large city in Scotland

### Exploring similarities and differences between hometown and cities and towns of interest

#### Applied Data Science Capstone by IBM

#### Part of IBM Data Science Professional Certificate

###### By Emil Fuerstenberg Haegg

## Introduction

### Background

Scotland is a country that is part of the United Kingdom and makes up the northern part of the island known as Great Britain. Scotland has a varied nature with the southern and eastern parts mostly consisting of rural lowlands, while the north-western parts are aptly known as the Highlands. The Highlands have a varied geology with mountains, countless islands, forests and bogs, but is sparsely populated. Most of the country's population is instead concentrated to the two larger cities in the south: Glasgow and Edinburgh, and the surrounding areas.

In this project we aim to find a suitable location to relocate from a city to another area, considering some of the factors which would influence the choice. A relative planning to relocate from the town of Stornoway located on the island of Lewis and Harris at the north-western edge of the country. For work reasons they are moving to the area of Glasgow and Edinburgh. Scotland has a varied nature with the southern and eastern parts mostly consisting of rural lowlands, while the north-western parts are aptly known as the Highlands. The Highlands have a varied geology with mountains, countless islands, forests and bogs, but is sparsely populated. Most of the country's population is instead concentrated to the two larger cities in the south: Glasgow and Edinburgh, and the surrounding areas.

### Business Problem

The problem consists of narrowing down the choice of data for comparing the areas, depending on what the stakeholders, in this case our relative, consider the most important. The number of different comparisons which could be made are vast, and input from our stakeholder will guide which direction to take. Our stakeholder is looking to move to town which is similar to their hometown.

As a starting point, we look to compare the cities and town on the following features: venues, population size and GDP per capita. We aim present this comparison visually. This way our stakeholder can get an initial overview on similarities and differences, as a starting point for further exploration.

## Data

### Sources

For this project the following sources of data was used.

* Wikipedia page containing table of the 51 largest cities and towns of Scotland, and their population. [1]
* Wikipedia page containing information on the hometown of Stornoway. [2]
* Wikipedia page containing table of the GDP per capita for the councils of Scotland. [3]
* Nominatim geolocator for retrieving coordinates for cities and towns of Scotland. [4]
* Foursquare places API, to request information of venues. [5]
* Github repository containing GeoJSON with boundaries for the councils of Scotland. [6]

Data could also be gathered from official registers but using Wikipedia as a source enables us to make use of this vast collection of knowledge, which is Wikipedia, while demonstrating the usage of commonly used tools for handling unstructured data.

This project consisted of two separate analysis steps on two sets of data. The first part consisted of visualising similarity of Scottish cities towns by nearby venues and population, the second part was visualising GDP per capita for areas in Scotland using a choropleth map. These two analyses required different approach and was performed separately, but the data was in similar format, so similar methods were used.

### Retrieval and wrangling of raw data

The Wikipedia pages was read as a text string using the Requests module. This was the parsed as html using the package BeautifulSoup. Using BeautifulSoup, the tables of interest were retrieved, and the columns needed read into DataFrames, for further processing.

For the DataFrame holding the list of cities and towns, and their population, the entry for Stornoway was added manually as a new row. Some cleaning was needed, unwanted characters removed, and the population numbers cast to integers, the country, UK was added to the end of the string for each city and town (hereafter known as just city)

The GeoJSON was downloaded using wget and opened for reading. By studying the file in a texteditor, the structure was understood. This way the names of all the council areas could be retrieved. The file was read into a variable for further use.

For the DataFrame having the information on councils and the GDP per capita quite some cleaning was needed. The areas did not exactly match those in the GeoJson file, sometimes the councils where further subdivided for GDP per capita and sometimes several councils were combined instead. Since the areas to be displayed on the map came from the GeoJSON file, it was necessary to match each of those areas to a GDP per capita entry. This was performed by studying overlap between representation of areas from the two sources, manually adding and removing rows to the DataFrame. For the largest council area Highlands, four different areas from the DataFrame had to be combined, with the GDP per capita added as an average of that of those four areas, a crude but straightforward method.

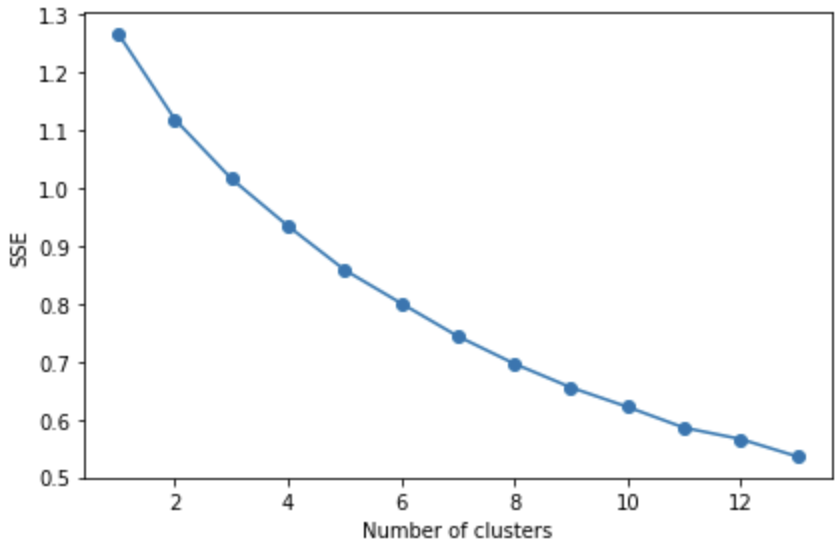
With these two dataframes, the first one containing name and population for the hometown as well as the 51 most populous cities and towns of Scotland, and the second one containing GDP per capita for all council areas of Scotland, the raw data was available for further analysis.

## Methodology

The Nominatim geolocator was used to retrieve the coordinates for each city. With this data, the Foursquare places API was called with requests for venue: -name, -longitude –latitude and –category for up to 100 venues per city, within a radius of 2500 meter. A large radius was used since we were interested in the occurrence of venues within larger areas, standing for what would be available for somebody living in that area. This way between 11 and 100 venues was retrieved for each city. This data was One Hot–encoded resulting in a DataFrame of shape (3005, 209), 3005 different venues of 208 unique categories.

These venues were grouped by city and by taking the mean of the frequency of occurrence of each category. The 10 most common venue-category could then be listed for each city.

To study similarities and differences between the cities, K-means clustering was used, grouping similar cities together based on the 10 most common venues in the within 2500 meter from the cities central coordinate. The number of clusters to use was decided by plotting the sum of squares error for 1-14 clusters. As can be seen in the graph below, no definite elbow was noticed but the decision was to use k = 5 for the clustering.



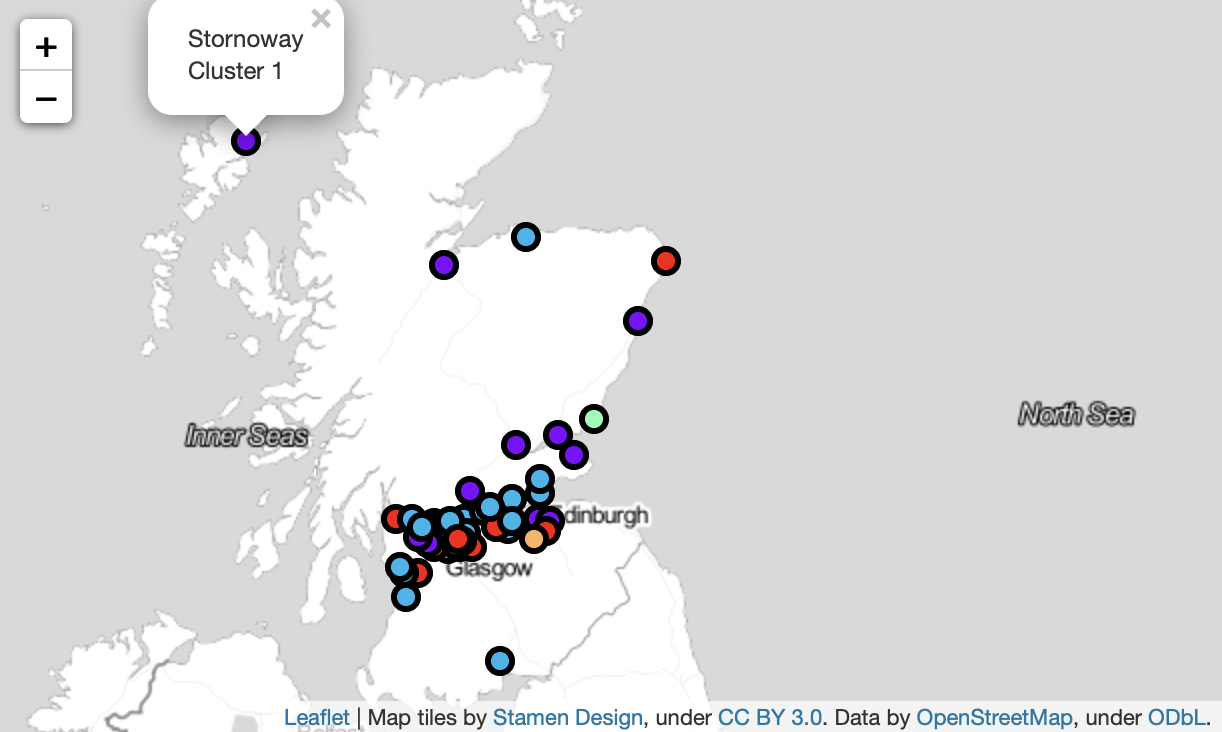
Columns containing the resulting clusters as well as the 10 most common areas was added to the DataFrame with the cities and their population through merging. A black and white map centred on Scotland was drawn using the folium library. The cites was plotted with colour representing the cluster they belong to.

To visualize the population of each city, another map was drawn with the radius of each city's marker proportional in size to the city's population.

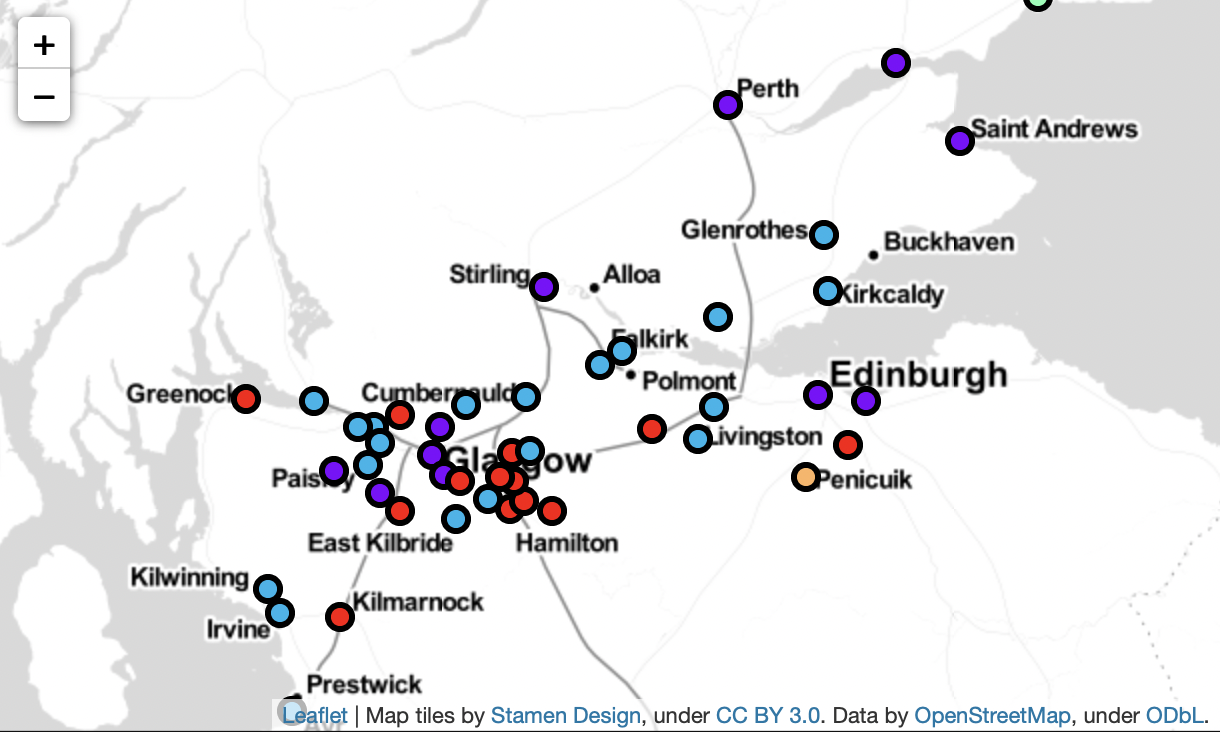
To create a choropleth map showing the GDP per capita for all council areas, the GeoJSON-file with the borders of the different areas, was used to layer on top of the map, showing the boundaries of the council areas. The areas were coloured according to GDP per capita. This was done using folium.Choropleth, with geo\_data from the GeoJSON-file, keyed on the name-entries of that file, with the DataFrame containing council names and GDP-per capita as data. 10 equally spaced bins were used to divide the GDP per capita from min to max.

The three maps above were also combined into a single map showing all the information together and coloured in the best way to avoid any ambiguity. This was done for the stakeholder to have a quick visual overview of the features studied,

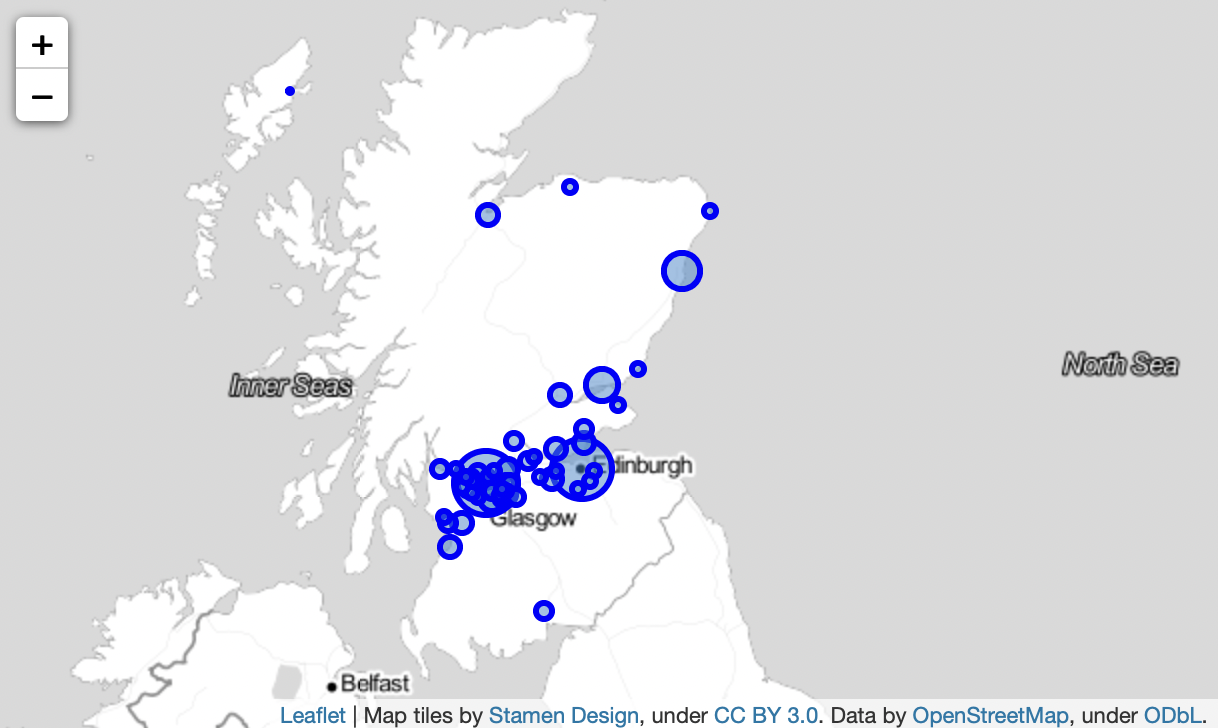
## Results



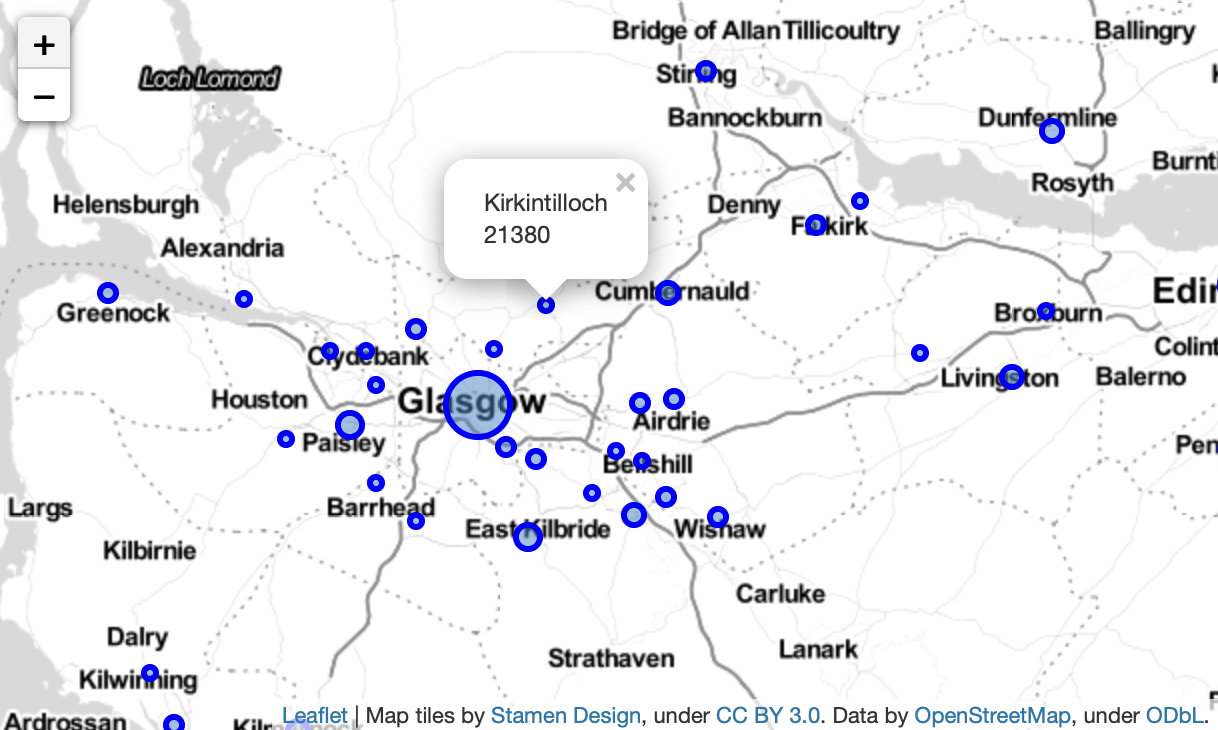
Map showing clusters based on 10 most common venue categories.



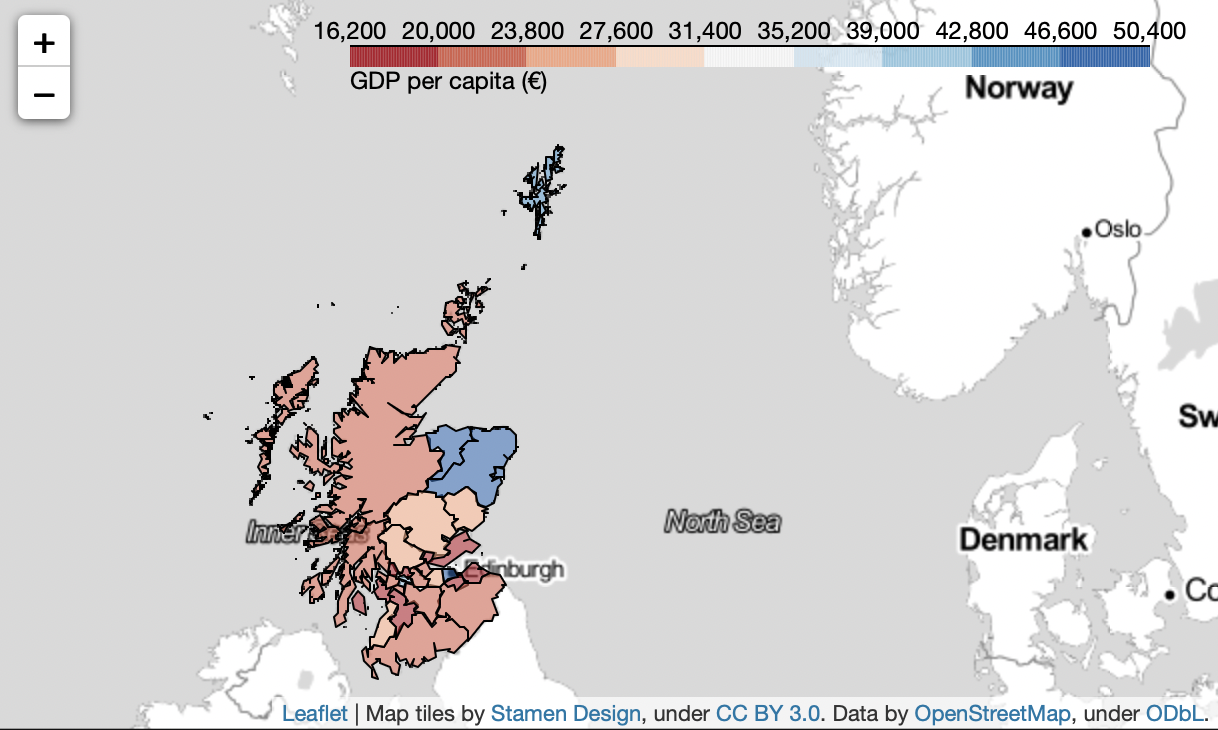
Zooming in on the more populated areas shows better detail.



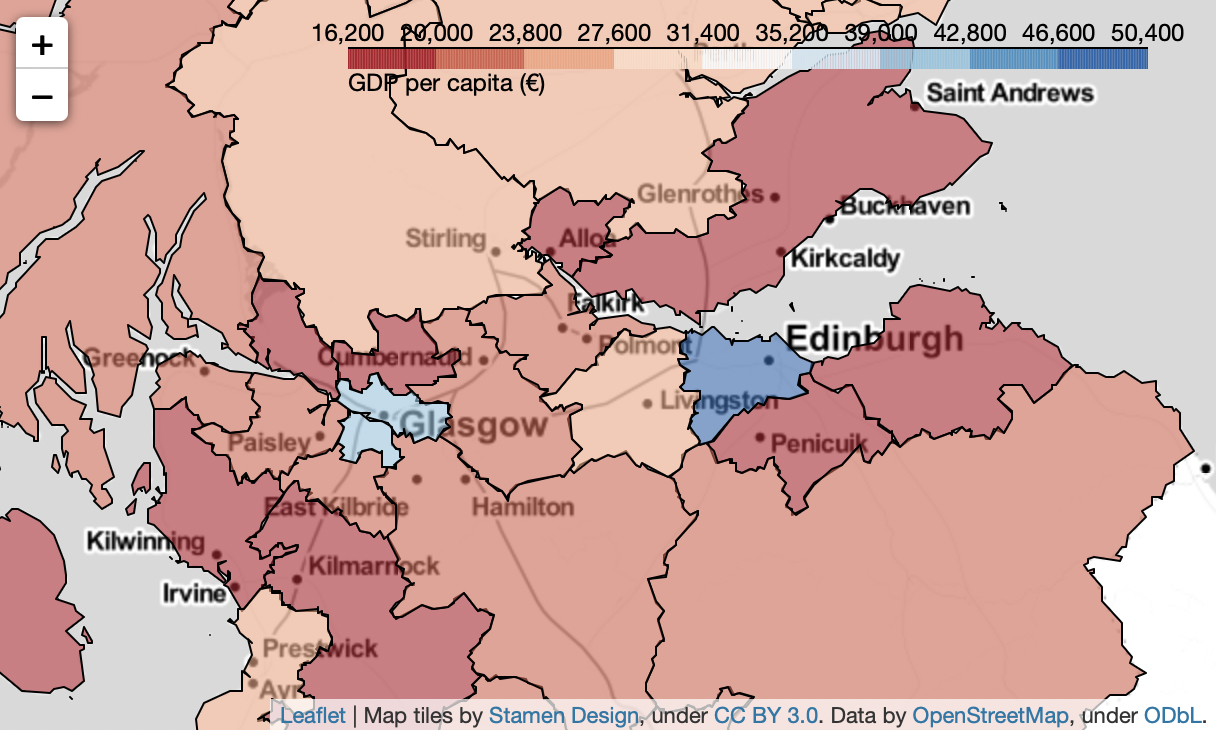
Marker proportional to the city's population size.



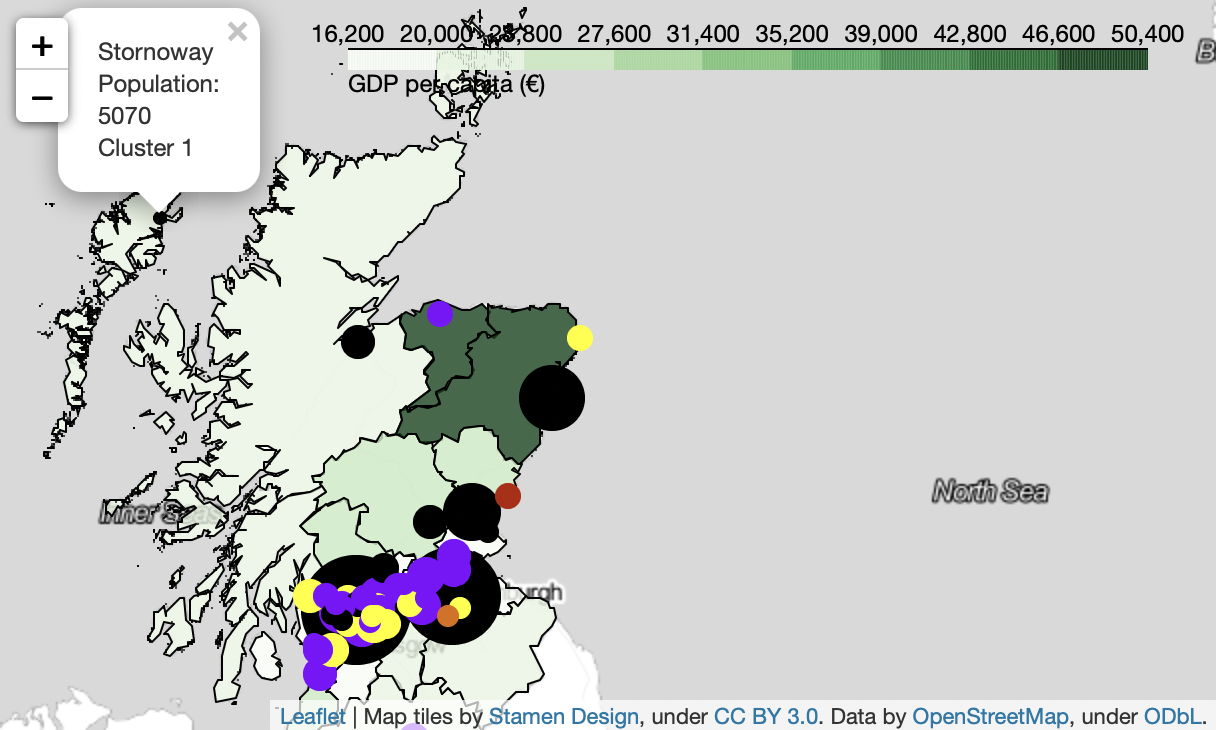
Popups shows name and population for cities.



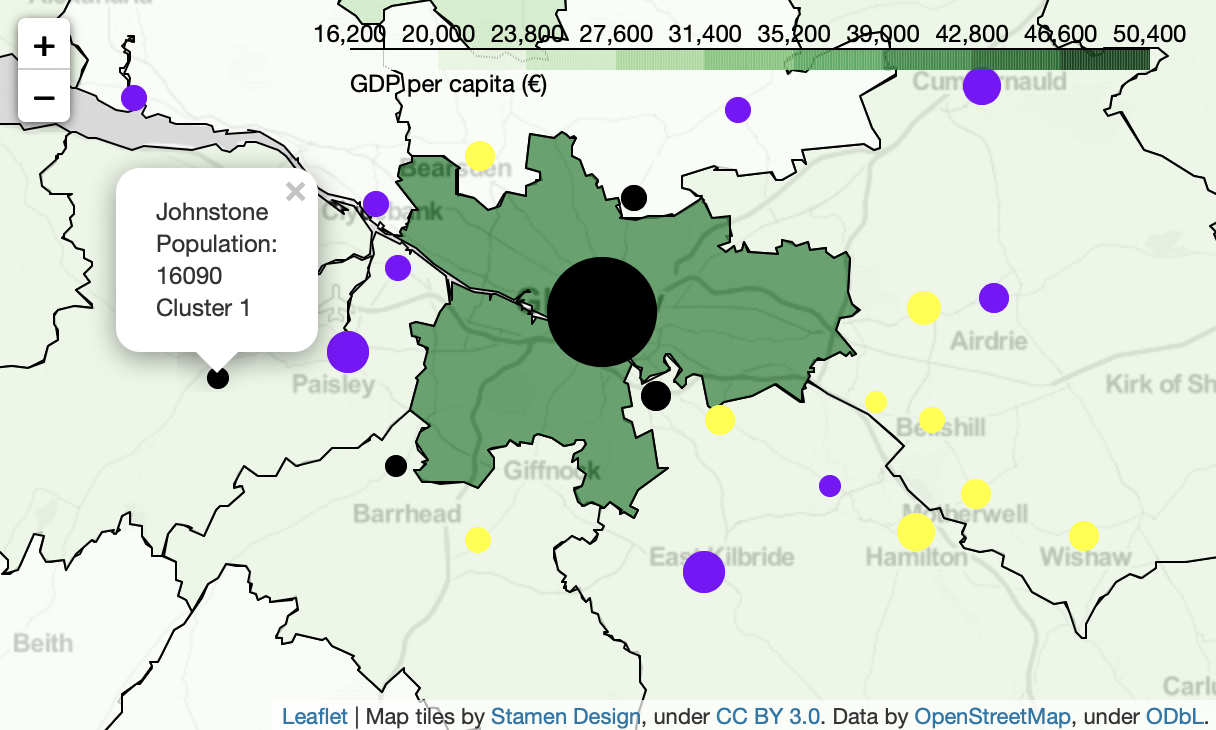
GDP per capita for the council areas.



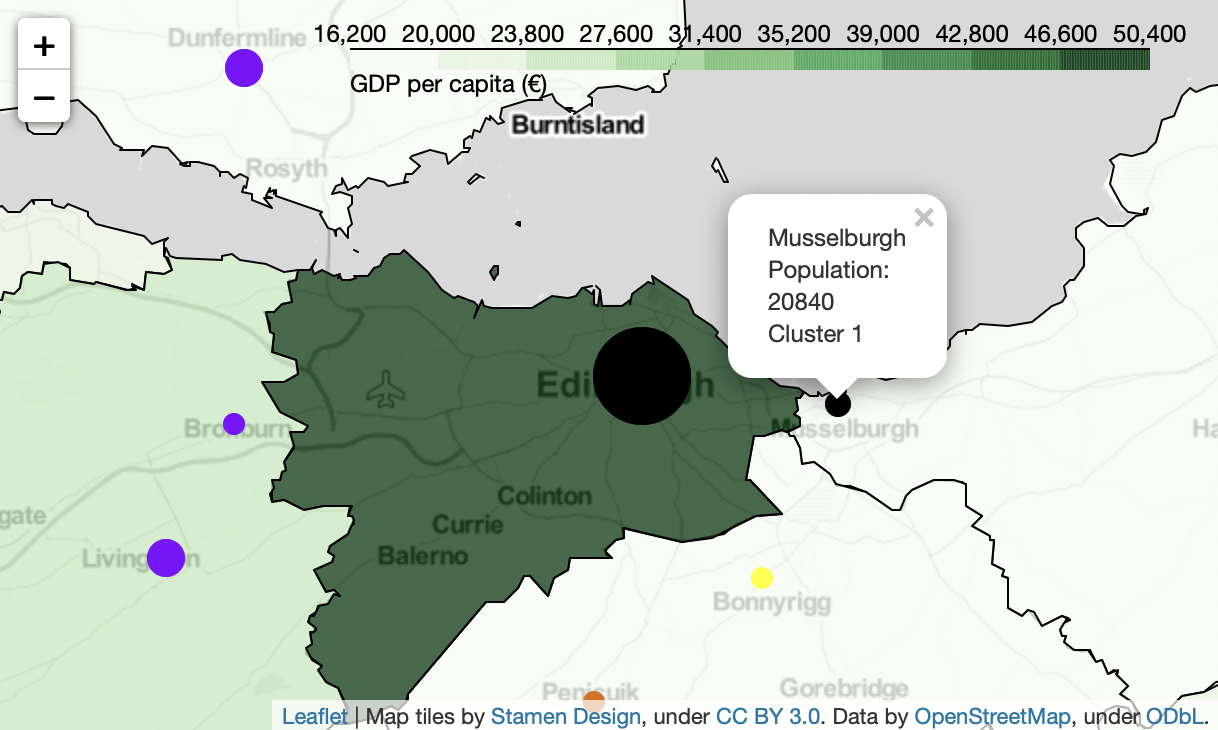
A better view over the area surrounding Glasgow and Edinburgh.



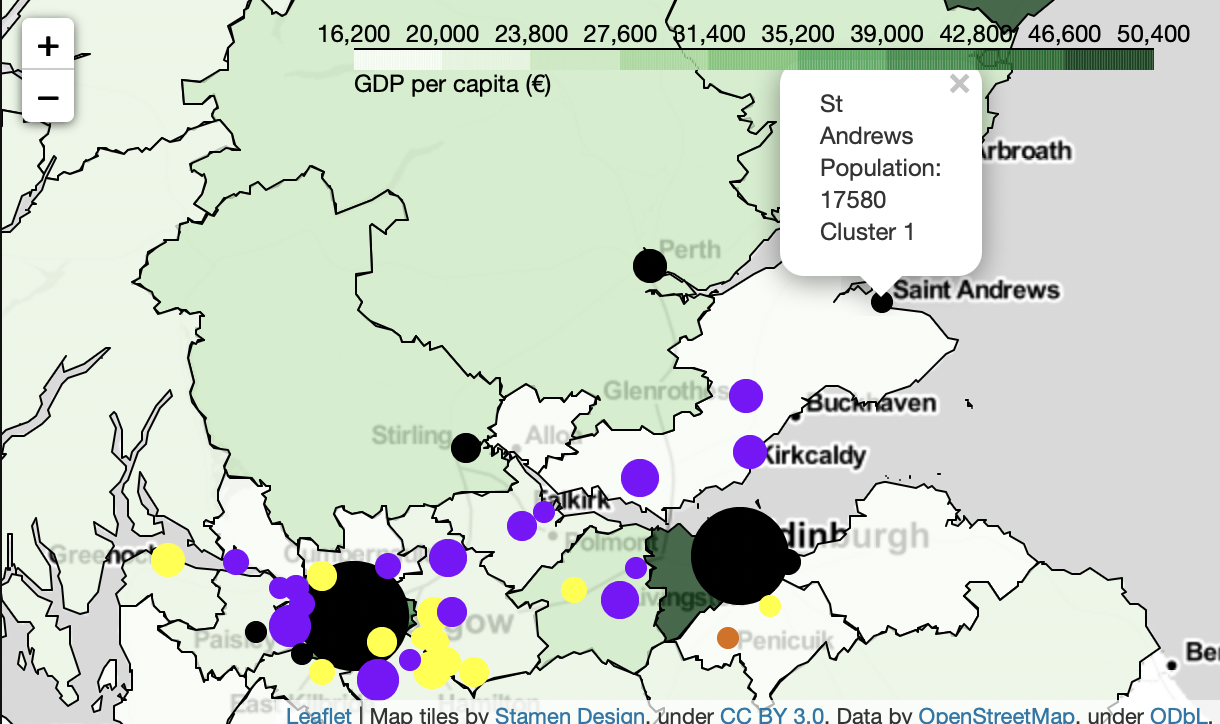
Combined map, popup showing stakeholder’s hometown.



A candidate city to investigate for our stakeholder.



Another candidate, showing similarities to the stockholder's hometown.



Yet another candidate city, a bit further out.

## Discussion

The visualisations created provides lots of information in a format which is easy to overview and would be a starting point for our stockholder for finding a suitable city to move to. With this goal of finding a city similar to their hometown there is of course a vast number of features to consider, it has been necessary to narrow down on a few interesting features, for this project.

The clustering provides information on the similarity of the venues available in the different cities. I think the clustering has worked quite well in this case, where single separate cities and their centres, for example Glasgow and Edinburgh, but also smaller but isolated towns, for example Inverness are classed together, here hotels, cafes and pubs are common for example. Suburbs with shopping centres, supermarkets and fast-food restaurants belong to another cluster, and so on.

Population size is a feature is well represented on a map like this, quickly giving information, compared to for example a table.

GDP per capita might not be the best feature to study the economy of an area, since it will naturally be higher for a city centre where much business is carried out, and low in an affluent suburban town where, people, while well off, money are not circulated. Going forward the average income for people in an area could be a better measurement of the economic situation there.

Some of the candidate cities most similar to the stakeholder's hometown can be seen marked in the three final maps in the Results-section. Since the hometown is a small town with a smaller population, these are all bigger towns in comparison. By finding data on smaller town in the south as well, better candidates could be found.

Of course, studying on the three features above, this can only be a basic measure on similarity, but a good starting-point, were other more specific needs: train access, nature, closeness to city centre or water, for example, would be the next step. Data science is an iterative process, where we are learning more about the data, by studying the data, going back and forth to solve the business problem as well as possible.

## Conclusion

In this project, I have studied the similarity between a smaller town in the northernmost Scotland to cities and towns in the more populous southern parts of the country. Using many popular data science tools, methods, libraries, and packages, with the help of unsupervised machine learning and visualisation, I have provided insight to the stakeholder and helped them with their business problem.

In addition to providing value to the stakeholder, this project has also demonstrated the usage of some of these tools taught in the courses included in the IBM Data Science Professional Certificate. I have learnt a lot, and it has sometimes been difficult to know where to stop exploring, often the initial data was not the best choice for the job, learning throughout the process, changing approach, and adjusting.

## References (data sources)

[1] https://en.wikipedia.org/wiki/List\_of\_towns\_and\_cities\_in\_Scotland\_by\_population

[2] https://en.wikipedia.org/wiki/Stornoway

[3] https://en.wikipedia.org/wiki/Economy\_of\_Scotland

[4] https://nominatim.org

[5] https://developer.foursquare.com

[6] https://github.com/martinjc/UK-GeoJSON/blob/master/json/administrative/sco/lad.json