**Data Description:**

In order to complete this project five data sets are needed.

1. **Dataset 1:** The list of all postal codes in Glasgow

This dataset is scraped from the Wikipedia page given in [5]. It contains the postcode district, Post town, coverage and local authority area information of 56 postal codes. The dataset is processed, which reduces the postcodes to 50 valid entries. In order to find the latitude and longitude information of these 50 entries, the postcode district and the first coverage area are feed into the geolocator.geocode function that uses the FourSquare agent. This information will be concatenated with the dataset acquired from the Wikipedia webpage. The final postal code data frame is given below.

Graphical user interface, application

Description automatically generated Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated  
Graphical user interface, text, application

Description automatically generated

In order to shortlist the postal codes for clustering, the distance to the city centre is needed. There is no dataset available for the distance of a postal code to the Glasgow city centre. So, this is determined by finding the geographical coordinates of the city centre (using geolocator.geocode and FourSquare agent) and then using the haversine formula to find the distance between the latitude and longitude of the city centre to those of each of the 50 postal codes. Postal codes that are within 5 km of the city centre are considered for further analysis. Only 22 postal codes meet this requirement.

1. **Dataset 2:** The location of the Glasgow Subway station

This dataset is publicly available from the Glasgow City Council at this link [6]. The dataset is given below. Graphical user interface, application, Excel

Description automatically generated

Graphical user interface, application, Excel

Description automatically generated

This dataset lists the addresses of the 15 subway stations in Glasgow. In order to find the geographical coordinates, the addresses are fed to the geolocator.geocode function. The coordinates are then used to find the distance of each postal code to the nearest subway station. The postal codes that are within 2 km of the subway station are shortlisted. Only 10 postal codes meet this requirement.

1. **Dataset 3 and 4:** The Royal mail dataset and the Glasgow City Council dataset.

There is no dataset that maps postal codes directly to population and ethnicity data. Hence in order to extract this information two datasets are used. The first dataset is a Royal mail dataset [7]. This data set contains all the postal codes in Glasgow. It has almost 15,000 data entries. It maps the postal codes in Glasgow to the corresponding datazones. The Glasgow City Council dataset then maps datazones to population information [8]. The data set has almost 6200 entries. The data sets are processed to extract the information of the population at each postcode and their ethnicities. The following dataset is extracted. This information will then be inputted to the unsupervised machine learning algorithm.

Graphical user interface, application, table

Description automatically generated

1. **Dataset 5:** The venues data

This data is extracted using the FoursquareAPI. Venues in the ten postal codes identified above are extracted. All venues within a 1.5 km radius of the postal code are imported. The dataset is cleaned to remove any repeating venues. The venue categories are one-hot encoded. The one-hot encoded venues dataframe is merged with the population data and then fed into an unsupervised learning algorithm (K-means) for clustering into 5 categories. The clusters are then examined to find possible business opportunities at each postal code.