BUSINESS INSIGHT AND ANALYSIS REPORT FOR GLASGOW

Table of Contents

Introduction

Glasgow has a population of over 600,000 people [1]. This population is projected to rise in the coming years. Migration, both national and international, is attributed as the main driver of this growth [2]. This rapid population growth brings with it a lot of new opportunities within the city of Glasgow for businesses to cater to a wider variety of clientele.

Most of the Glasgow population is clustered around the city centre which is the main hub of economic activity. Glasgow also has a very reliable public transport system. According to the SPT Subway survey, 97% of subway commuters report overall satisfaction [3]. Therefore, subways are a popular mode of transport for the local population trying to commute within the city. 13.1 million subway passenger journeys were recorded in the 2018 to 2019 period [4].

So, areas close to the city centre (~ 5 km from the city centre) and close to a subway station (~ 2 km from the nearest subway station) offer businesses a great deal of opportunity and exposure to a wide customer base. So, these areas are ideal places to start a new business. In Glasgow's highly competitive business environment, identifying business opportunities and choosing the optimum location is a pivotal decision for new small businesses. Choosing the right location can have considerable impact on future profits and prospects for a business.

Therefore, this project aims to provide data driven insight to businesses by identifying possible business opportunities within Glasgow and the ideal location (postal district) for those business opportunities. The project will use venue information acquired using the FourSquareAPI to identify the most common venues within each postal code. The total population and their ethnicities are also acquired using datasets from the Glasgow City Council and the Royal mail. Insight from this data will allow postal codes to be clustered according to the most common venues and their population characteristics. This will offer insight and analysis into the type of business competition at each postal code and possible untapped business opportunities within Glasgow.

Data Description:

In order to complete this project five data sets are used. The datasets come from three sources, web pages, publicly available datasets and FourSquareAPI.

Dataset 1: The list of all postal codes in Glasgow with their latitude and longitude information. This data set is a combination of data acquired from the Web and the FourSquareAPI.

This dataset is scraped from the Wikipedia page given in [5]. It contains the following parameter information for the 56 G category postal codes.

- Postcode district
- Post town
- Coverage
- Local authority area

The dataset is processed. The non-geographical postal codes are removed. This reduces the postcode to 50 valid entries. A new column is added to the dataset, called 'District'. The District is the first coverage area. In order to find the latitude and longitude information of these 50 entries, the postcode and the corresponding district are feed into the geolocator.geocode function that uses the FourSquare agent. The function to find the latitude and longitudes of the postal codes is given below.

```
from geopy.geocoders import Nominatim
for n in range (0,Glasgow_df.shape[0]):
    address = '{}, {}, Glasgow'.format(Glasgow_df['Postcode district'][n], Glasgow_df['District']
[n])
    geolocator = Nominatim(user_agent="foursquare_agent")
    location = geolocator.geocode(address)
    latitude=location.latitude
    longitude=location.longitude
    print('Index:',n,'{},{} has latitude='.format(Glasgow_df['Postcode district'][n], Glasgow_df['Di
strict'][n]),latitude,'and longitide=', longitude)
```

The information retrieved from the foursquare agent is saved in a csv file. This file is then imported and merged with the dataset acquired from the Wikipedia webpage. The final postal code dataframe is given below.

Table (1): The list of all post codes

	Postcode district	Post town	Coverage	Local authority area	District	Latitude	Longitude
0	G1	GLASGOW	Merchant City	Glasgow City	Merchant City	55.859126	-4.246316
1	G2	GLASGOW	Blythswood Hill, Anderston (part)	Glasgow City	Blythswood Hill	55.863319	-4.261671
2	G3	GLASGOW	Anderston, Finnieston, Garnethill, Park, Woodl...	Glasgow City	Anderston	55.859699	-4.271292
3	G4	GLASGOW	Calton (part), Cowcaddens (part), Drygate, Kel...	Glasgow City	Calton	55.853724	-4.232824
4	G5	GLASGOW	Gorbals	Glasgow City	Gorbals	55.851813	-4.253163
5	G11	GLASGOW	Broomhill, Partick, Partickhill	Glasgow City	Broomhill	55.877340	-4.321745
6	G12	GLASGOW	West End (part), Cleveden, Dowanhill, Hillhead...	Glasgow City	West End	55.879886	-4.292322
7	G13	GLASGOW	Anniesland, Knightswood, Yoker	Glasgow City	Anniesland	55.890019	-4.332739
8	G14	GLASGOW	Whiteinch, Scotstoun	Glasgow City	Whiteinch	55.875338	-4.337189
9	G15	GLASGOW	Drumchapel	Glasgow City	Drumchapel	55.912244	-4.369171
10	G20	GLASGOW	Maryhill, North Kelvinside, Ruchill	Glasgow City	Maryhill	55.890787	-4.290405

11	G21	GLASGOW	Balornock, Barmulloch, Cowlairs, Royston, Spri...	Glasgow City	Balornock	55.881657	-4.214442
12	G22	GLASGOW	Milton, Parkhouse, Possilpark	Glasgow City	Milton	55.897001	-4.246675
13	G23	GLASGOW	Lambhill, Summerston	Glasgow City	Lambhill	55.895932	-4.265246
14	G31	GLASGOW	Dennistoun, Haghill, Parkhead (part)	Glasgow City	Dennistoun	55.861736	-4.219010
15	G32	GLASGOW	Carmyle, Tollcross, Mount Vernon, Lightburn, S...	Glasgow City	Carmyle	55.831113	-4.155489
16	G33	GLASGOW	Cardowan, Carntyne, Craigend, Cranhill, Gartha...	North Lanarkshire, Glasgow City	Cardowan	55.860114	-4.172197
17	G34	GLASGOW	Easterhouse, Easthall, Provanhall	Glasgow City	Easterhouse	55.866744	-4.105808
18	G40	GLASGOW	Bridgeton, Calton, Dalmarnock	Glasgow City	Bridgeton	55.846267	-4.222231
19	G41	GLASGOW	Pollokshields, Shawlands	Glasgow City	Pollokshields	55.842266	-4.284997
20	G42	GLASGOW	Battlefield, Govanhill, Mount Florida, Strathb...	Glasgow City	Battlefield	55.824493	-4.266598
21	G43	GLASGOW	Mansewood, Newlands, Pollokshaws	Glasgow City	Mansewood	55.815297	-4.302073
22	G44	GLASGOW	Cathcart, Simshill, Croftfoot, King's Park, Mu...	Glasgow City, East Renfrewshire	Cathcart	55.816988	-4.262424
23	G45	GLASGOW	Castlemilk	Glasgow City	Castlemilk	55.805271	-4.230395
24	G46	GLASGOW	Arden, Carnwadric, Deaconsbank, Giffnock, Kenn...	East Renfrewshire, Glasgow City	Arden	55.806630	-4.330756
25	G51	GLASGOW	Cessnock, Plantation	Glasgow City	Cessnock	55.854152	-4.294398
26	G52	GLASGOW	Cardonald, Hillington, Penilee, Mosspark	Glasgow City	Cardonald	55.848291	-4.349974
27	G53	GLASGOW	Darnley, Pollok, Crookston, Parkhouse, Priesth...	Glasgow City	Darnley	55.808944	-4.340953
28	G60	GLASGOW	Bowling, Old Kilpatrick	West Dunbartonshire	Bowling	55.815327	-4.266320
29	G61	GLASGOW	Bearsden	East Dunbartonshire	Bearsden	55.920461	-4.333205
30	G62	GLASGOW	Baldernock, Milngavie, Mugdock	East Dunbartonshire, Stirling	Baldernock	55.947113	-4.280062
31	G63	GLASGOW	Balfron, Balmaha, Blanefield, Croftamie, Dryme...	Stirling	Balfron	56.071609	-4.333266
32	G64	GLASGOW	Bishopbriggs, Torrance	East Dunbartonshire	Bishopbriggs	55.904409	-4.226358
33	G65	GLASGOW	Croy, Kilsyth, Twechar	North Lanarkshire, East Dunbartonshire	Croy	55.889372	-4.204707
34	G66	GLASGOW	Clachan of Campsie, Haughhead, Kirkintilloch, ...	East Dunbartonshire	Clachan of Campsie	55.988662	-4.228600
35	G67	GLASGOW	Cumbernauld (south)	North Lanarkshire	Cumbernauld	55.944109	-3.994124
36	G68	GLASGOW	Cumbernauld (north), Dullatur	North Lanarkshire	Cumbernauld	55.874026	-4.184127
37	G69	GLASGOW	Baillieston, Bargeddie, Chryston, Garrowhill, ...	Glasgow City, North Lanarkshire	Baillieston	55.850211	-4.110833
38	G71	GLASGOW	Birkenshaw, Bothwell, Broomhouse, Tannochside,...	Glasgow City, North Lanarkshire, South Lanarks...	Birkenshaw	55.862058	-4.209337
39	G72	GLASGOW	Blantyre, Cambuslang	South Lanarkshire	Blantyre	55.793522	-4.092311
40	G73	GLASGOW	Rutherglen	South Lanarkshire	Rutherglen	55.828389	-4.211653
41	G74	GLASGOW	East Kilbride (north), Thorntonhall	South Lanarkshire	East Kilbride	55.760176	-4.179078
42	G75	GLASGOW	Auldhouse, East Kilbride (south)	South Lanarkshire	Auldhouse	55.818785	-4.295358
43	G76	GLASGOW	Busby, Carmunnock, Clarkston, Eaglesham, Water...	East Renfrewshire, Glasgow City	Busby	55.786185	-4.245825
44	G77	GLASGOW	Newton Mearns	East Renfrewshire	Newton Mearns	55.772673	-4.334237
45	G78	GLASGOW	Barrhead, Neilston, Uplawmoor	East Renfrewshire	Barrhead	55.801103	-4.389983
46	G81	CLYDEBANK	Dalmuir, Duntocher, Faifley, Hardgate	West Dunbartonshire	Dalmuir	55.924539	-4.402629
47	G82	DUMBARTON	Cardross, Milton, Renton	West Dunbartonshire, Argyll and Bute	Hardgate	55.964424	-4.657399
48	G83	ALEXANDRIA	Alexandria, Arrochar, Aldochlay, Ardlui, Ballo...	West Dunbartonshire, Argyll and Bute, Stirling	Alexandria	55.987874	-4.582209
49	G84	HELENSBURGH	Clynder, Cove, Garelochhead, Kilcreggan, Rhu, ...	Argyll and Bute	Clynder	55.855911	-4.305579

The 50 postcodes are plotted on a map using the folium library. They are shown in Figure 1. As can be seen from the map, the postal codes are dispersed over a very large area. As most of the economic activity in Glasgow is near the city centre, Postal codes that are closer are considered more suitable locations for opening businesses because of their higher exposure to a greater client base.

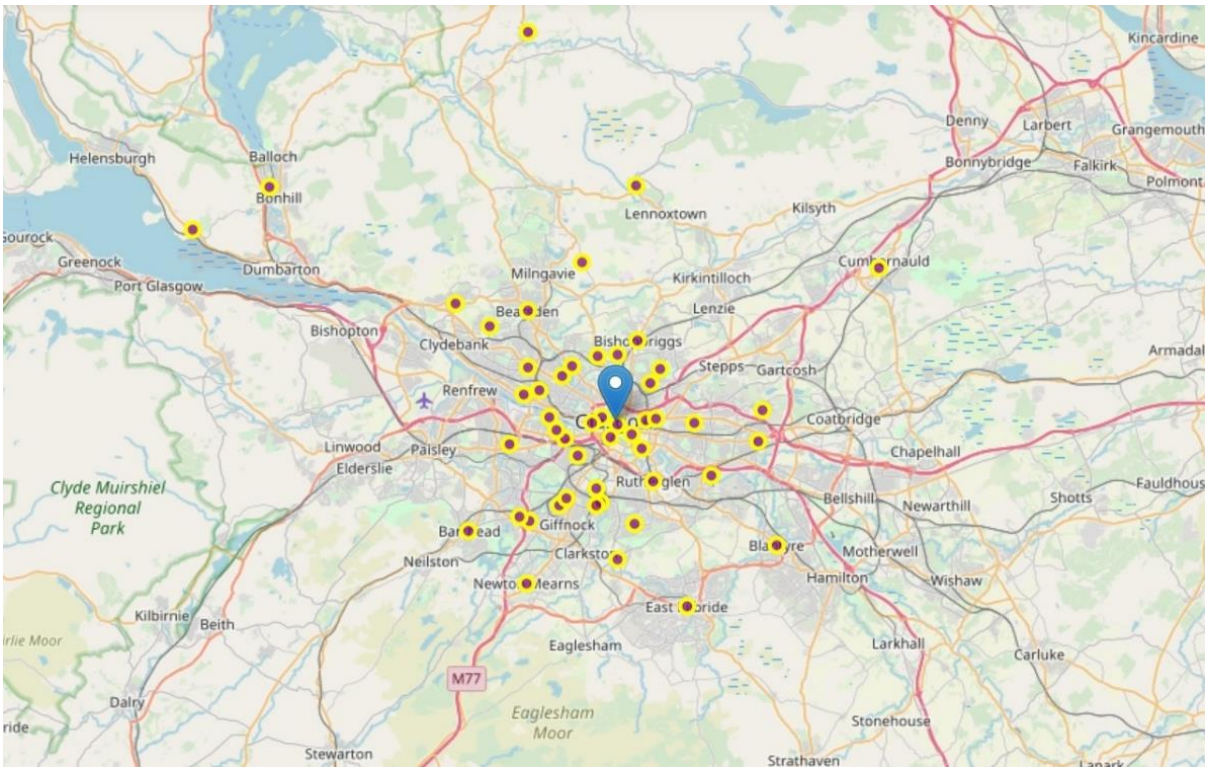

Figure (1): All the G Postal codes in Glasgow

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 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. The function used for this is given below.

```
from math import sin, cos, sqrt, atan2, radians

def calculate_distance_from_Glasgow_City_Center(latitude,longitude):
    # approximate radius of earth in km
    R = 6373.0
    lat1 = radians(55.8609825) # Latitude Coordinates for the Glasgow City Center
    lon1 = radians(-4.2488787) # Longitude Coordinates for the Glasgow City Center
    lat2 = radians(latitude)
    lon2 = radians(longitude)
    #Haversine Formula to find distance
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))

    distance = R * c
    return distance #The distance is returned in km
```

Postal codes that are within 5 km of the city centre are considered for further analysis. The rest are dropped from the dataframe because of the lower business exposure they offer. Out of the 50 postcodes, only 22 postal codes are within 5 km of the city centre. These postcodes are shown in the map in figure (2).

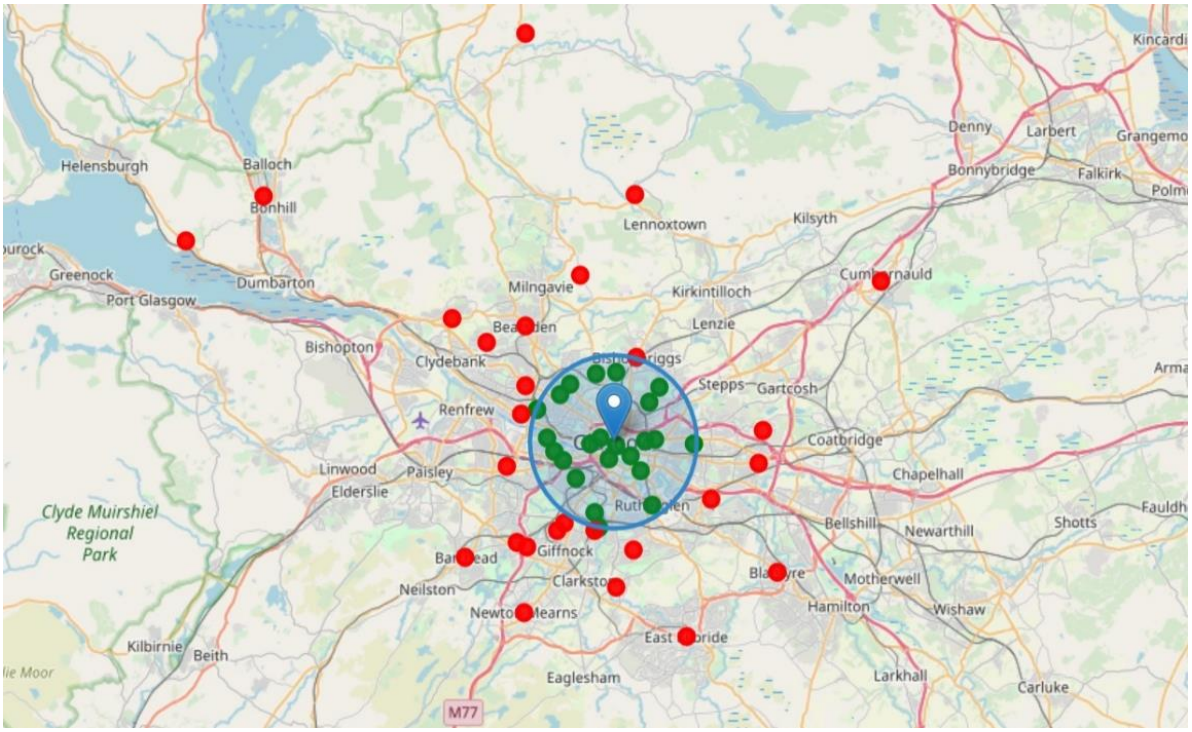

Figure (2): Postal codes that are within 5 km of the City Centre

Dataset 2: The addresses of the Glasgow Subway station and the latitude and longitude of those subway station. This data set is acquired from a publicly available dataset and the FourSquareAPI.

This dataset is publicly available from the Glasgow City Council at this link [6]. The dataset contains 14 information parameters. The dataset is given below.

Table (2): The Subway addresses dataset

	Station Name	Address Line 1	Address Line 2	Address Line 3	Postcode	Easting	Northing	Accessible	Park and Ride	Wi-Fi	Bike Shelter	Special Car Sharing Parking Spaces	Lost Property Number	Nearby
0	Bridge Street	67	Eglinton Street	Glasgow	G5 9NR	258710	664400	Yes	Yes	Yes	Yes	No	0141 335 3276.	Citizen's Theatre, 02 Academy.
1	Buchanan Street	174	Buchanan Street	Glasgow	G1 2JZ	259073	665577	Yes	No	Yes	No	No	0141 335 3276.	Queen's Street Station, Buchanan Bus Station, ...
2	Cessnock	1	Cessnock Street	Glasgow	G51 1AS	256470	664490	Yes	No	Yes	No	No	0141 335 3276.	Glasgow Science Centre, IMAX, SECC
3	Cowcaddens	11	Dundasvale Court, Unit 1	Glasgow	G4 0SN	258679	666191	Yes	No	Yes	No	No	0141 335 3276.	Glasgow School of Art, GFT, Tenement House.
4	Govan	737	Govan Road	Glasgow	G51 2YJ	255490	665643	Yes	No	Yes	No	No	0141 335 3276.	Southern General Hospital, Govan Old Parish Ch...
5	Hillhead	248	Byres Road	Glasgow	G12 8SH	256615	667052	Yes	No	Yes	No	No	0141 335 3276.	Byres Road, Ashton Lane, Botanic Gardens, Univ...
6	Ibrox	124	Copland Road	Glasgow	G51 2UB	255830	664787	Yes	No	Yes	No	No	0141 335 3276.	Bellahouston Park
7	Kelvinhall	7	Dalcross Pass	Glasgow	G11 5RA	256184	666598	Yes	No	Yes	No	No	0141 335 3276.	Kelvingrove Museum, Western Infirmary
8	Kelvinbridge	99	South Woodside Road	Glasgow	G4 9HG	257475	666922	Yes	Yes	Yes	Yes	No	0141 335 3276.	Kelvingrove Park,
9	Kinning Park	30	Cornwall Street	Glasgow	G41 1AH	256864	664281	Yes	No	Yes	No	No	0141 335 3276.	NaN
10	Partick	15	Merkland Street	Glasgow	G11 6DB	255635	666503	Yes	No	Yes	Yes	No	0141 335 3276.	Partick Rail Station, Partick Bus Station, Riv...
11	Sheilds Road	360	Scotland Street	Glasgow	G5 8QF	257660	664205	Yes	Yes	Yes	Yes	Yes	0141 335 3276.	Scotland Street School, Museum Of Education
12	St. Enoch	10	St Enoch Square	Glasgow	G1 4DB	258946	665038	Yes	No	Yes	No	No	0141 335 3276.	Glasgow Central Station, St. Enoch Centre.
13	St Georges Cross	51	Great Western Road	Glasgow	G4 9AH	258106	666560	Yes	No	Yes	No	No	0141 335 3276.	The Mitchell Library.
14	West Street	281	West Street	Glasgow	G5 8NB	258220	664150	Yes	No	Yes	No	No	0141 335 3276.	NaN

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 information acquired from the FourSquare agent is saved in an excel file which is then imported and merged with the data frame shown above. The location of the 15 subway stations are shown in figure (3). Area within 2 km of the subway station is shaded blue.

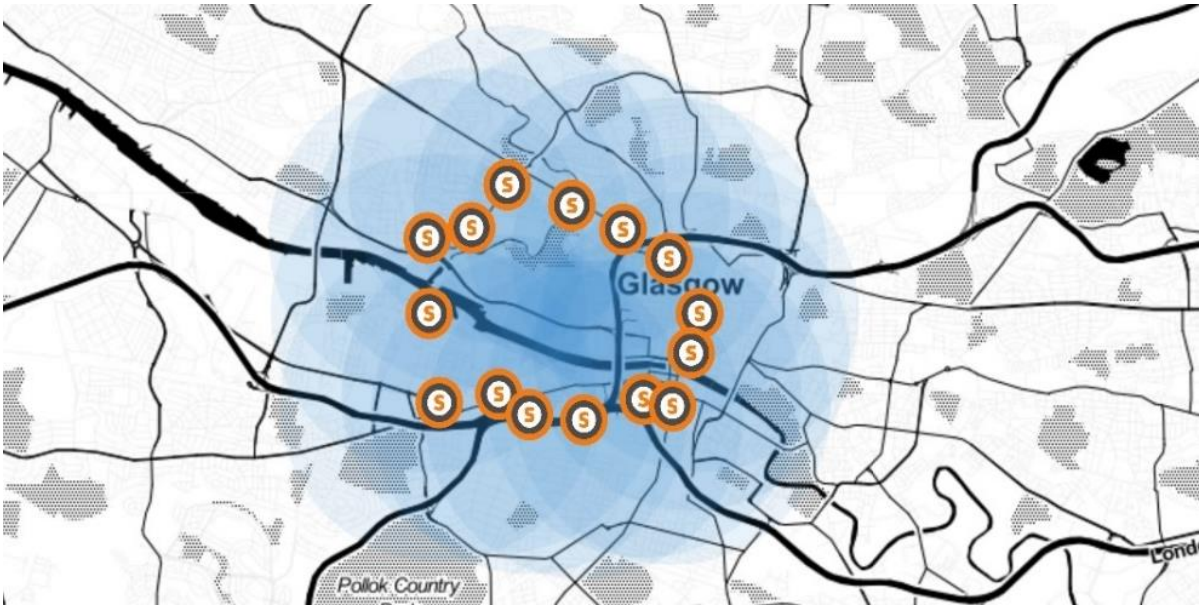

Figure (3): The location of the 15 subway stations in Glasgow

In the 2018-2019 period, 13.1 million passenger journeys were recorded by the Glasgow subway company SPT [4]. As subways are a popular mode of transport in the city of Glasgow, areas around the subway station are suitable places to start a new business within the city. In order to find the distance of a postal code to the nearest subway station the following function is used. It returns the distance to the nearest subway station in km and the name of the nearest station.

```
# Function that returns the distance of the closest subway station and the name of that station
def calculate_distance_from_Subway_station(latitude,longitude):
    Station_names=['Bridge Street',
                   'Buchanan Street',
                   'Cessnock',
                   'Cowcaddens',
                   'Govan',
                   'Hillhead',
                   'Ibrox',
                   'Kelvinhall',
                   'Kelvinbridge',
                   'Kinning Park',
                   'Partick',
                   'Sheilds Road',
                   'St. Enoch',
                   'St Georges Cross',
                   'West Street']
    distancesarray=np.zeros((len(Station_names)))
    for n in range (len(Station_names)):
        # approximate radius of earth in km
        R = 6373.0
        lat1 = radians(Glasgow_subway['Latitude'][n])
        lon1 = radians(Glasgow_subway['Longitude'][n])
        lat2 = radians(latitude)
        lon2 = radians(longitude)
        #Haversine Formula to find distance
        dlon = lon2 - lon1
        dlat = lat2 - lat1
        a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
        c = 2 * atan2(sqrt(a), sqrt(1 - a))
        distancesarray[n] = R * c

    min_distance=min(distancesarray)
    index_min=np.argmin(distancesarray)
    station=Station_names[index_min]

    return min_distance, station #The distance is returned in km
```

Out of the 22 postal codes that are shortlisted on the basis of distance from the city centre, only 10 are within 2 km of a subway station. They are shown in figure (4) as green dots. The final 10 postal codes are shown in table (3).

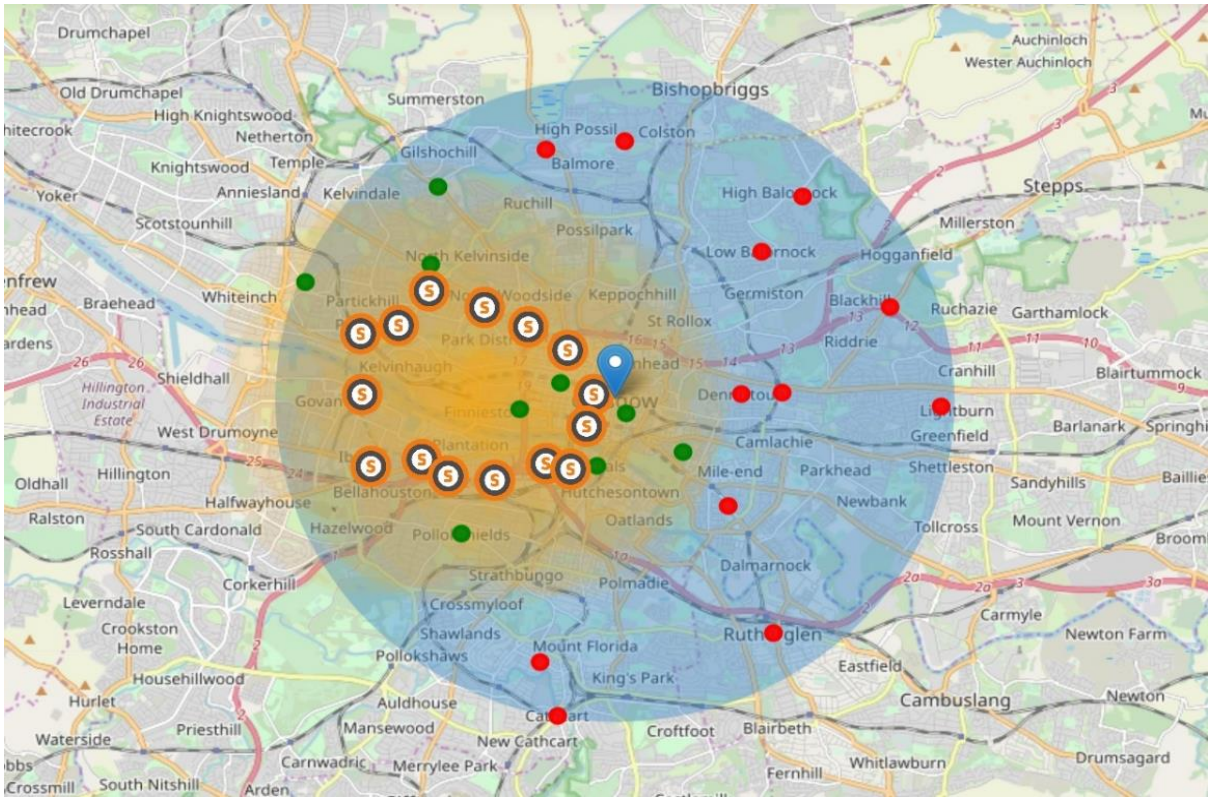

Figure (4): The postal codes that are within 5 km of the city centre and that are 2 km from the subway station.

Table (3): The list of postal codes that are within 5 km of the city centre and that are within 2 km of the postal codes

	Postcode district	Post town	Coverage	Local authority area	District	Latitude	Longitude	Distance from Center	Distance from Nearest Subway	Nearest Subway
0	G1	GLASGOW	Merchant City	Glasgow City	Merchant City	55.859126	-4.246316	0.261226	0.550108	Buchanan Street
1	G2	GLASGOW	Blythswood Hill, Anderston (part)	Glasgow City	Blythswood Hill	55.863319	-4.261671	0.839709	0.526778	Buchanan Street
2	G3	GLASGOW	Anderston, Finnieston, Garnethill, Park, Woodl...	Glasgow City	Anderston	55.859699	-4.271292	1.406410	0.912032	West Street
3	G4	GLASGOW	Calton (part), Cowcaddens (part), Drygate, Kel...	Glasgow City	Calton	55.853724	-4.232824	1.287026	1.474669	St. Enoch
4	G5	GLASGOW	Gorbals	Glasgow City	Gorbals	55.851813	-4.253163	1.054402	0.390402	Bridge Street
5	G11	GLASGOW	Broomhill, Partick, Partickhill	Glasgow City	Broomhill	55.877340	-4.321745	4.897992	1.124514	Partick
6	G12	GLASGOW	West End (part), Cleveden, Dowanhill, Hillhead...	Glasgow City	West End	55.879886	-4.292322	3.430967	0.411294	Hillhead
7	G20	GLASGOW	Maryhill, North Kelvinside, Ruchill	Glasgow City	Maryhill	55.890787	-4.290405	4.207631	1.629398	Hillhead
8	G41	GLASGOW	Pollokshields, Shawlands	Glasgow City	Pollokshields	55.842266	-4.284997	3.069140	0.927243	Kinning Park
9	G51	GLASGOW	Cessnock, Plantation	Glasgow City	Cessnock	55.854152	-4.294398	2.941483	0.159598	Cessnock

Dataset 3 and 4: The Royal mail dataset and the Glasgow City Council dataset. These are publicly available datasets

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. This dataset contains seven parameters of information. They are:

- Postcode unit
- Type of Postcode,
- Sector Community Planning Partnership
- Multi member Ward/ Local Area Partnership
- Data Zone
- Data Zone Rank (SIMD 2012)
- Is this in a bottom 15% Data Zone?

The first five rows of the data frame are shown below. The dataset is processed in order to extract the unique Data zones of each postal district.

Table (4): Royal Mail Dataset

	Postcode unit	Type of Postcode	Sector Community Planning Partnership	Multi member Ward / Local Area Partnership	Data Zone	Data Zone Rank (SIMD 2012)	Is this in a bottom 15% Data Zone?
0	G1 1BL	small user	North West	Anderston/City	S01003366	4833	no
1	G1 1BQ	small user	North West	Anderston/City	S01003358	1847	no
2	G1 1BX	Large User	North West	Anderston/City	S01003399	4029	no
3	G1 1DA	small user	North West	Anderston/City	S01003358	1847	no
4	G1 1DG	small user	North West	Anderston/City	S01003358	1847	no

Once the unique data zones of each postal code are obtained, the population and the ethnicity of the population at the postal codes is found using the Glasgow City Council data set [8]. The data set has almost 6200 entries. It is processed in order to extract and sum the population statistics of the data zones of each of the postal code. Firstly, the information regarding the total population is extracted. Next, the number and percentage of White, Asian, African and Other population living at each postcode is also extracted. The extracted dataset for the 10 postal codes is shown in table (5). This information will later be inputted to the unsupervised machine learning algorithm for clustering.

Table (5): The Extracted population information

	All people	White	White Percentage	Asian	Asian Percentage	African	African Percentage	Post Code	Other	Other Percentage
0	36994	26830	72.525274	8792	23.766016	639	1.727307	G41	733	1.981402
1	36269	31681	87.350079	2866	7.902065	1028	2.834376	G20	694	1.913480
2	31367	27572	87.901298	2814	8.971212	249	0.793828	G12	732	2.333663
3	26862	20312	75.616112	5083	18.922642	478	1.779465	G3	989	3.681781
4	24474	18245	74.548500	4353	17.786222	1008	4.118657	G4	868	3.546621
5	24405	20653	84.626101	2641	10.821553	740	3.032166	G51	371	1.520180
6	23157	20916	90.322581	1598	6.900721	169	0.729801	G11	474	2.046897
7	13291	10450	78.624633	2084	15.679783	306	2.302310	G1	451	3.393274
8	11781	9669	82.072829	1283	10.890417	611	5.186317	G5	218	1.850437
9	9736	7578	77.834840	1565	16.074363	218	2.239113	G2	375	3.851684

This information is plotted in order to better understand the population characteristics of each post code. The population at the 10 shortlisted post codes is shown in figure (5) in descending order of total population. G41, G20 and G12 each a population of over 30,000 people.

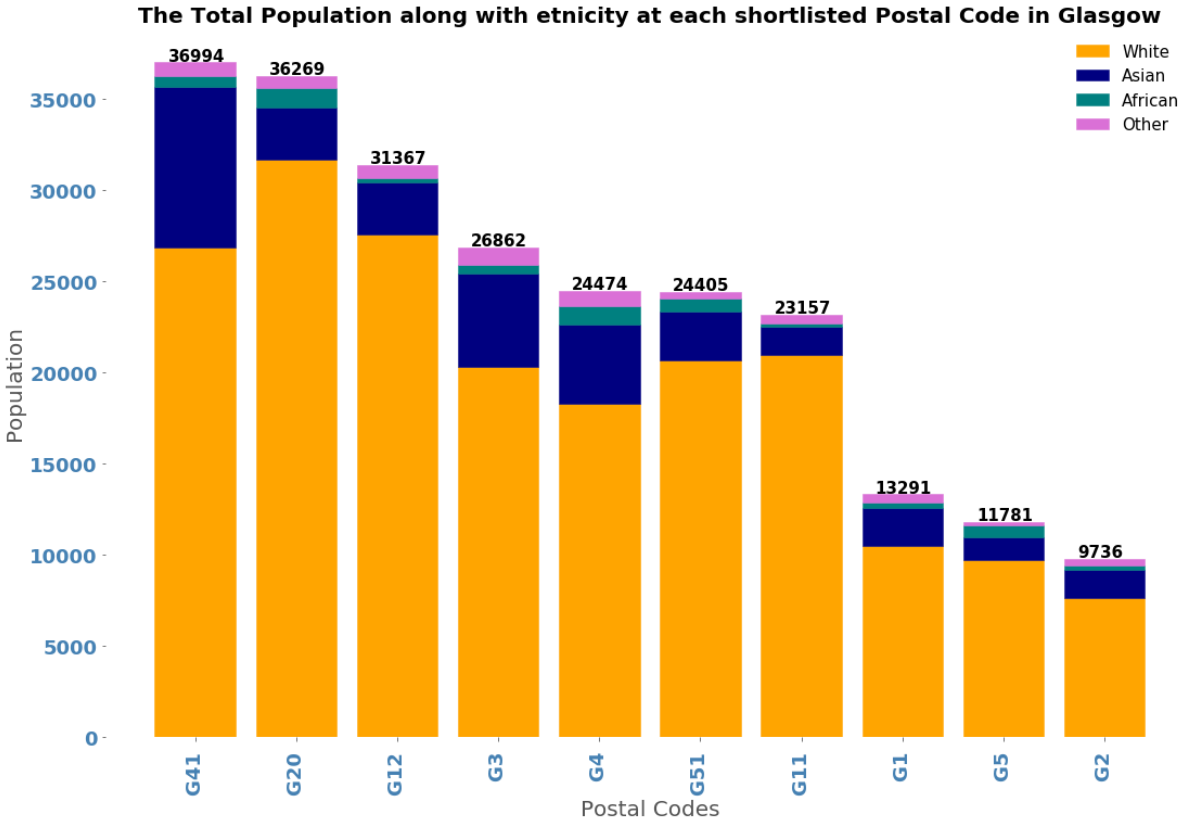

Figure (5): Total populations at each postcode

The top three ethnicity percentages of the population are plotted in figure (6) below. The figure shows that some postal codes have very similar ethnicity percentages.

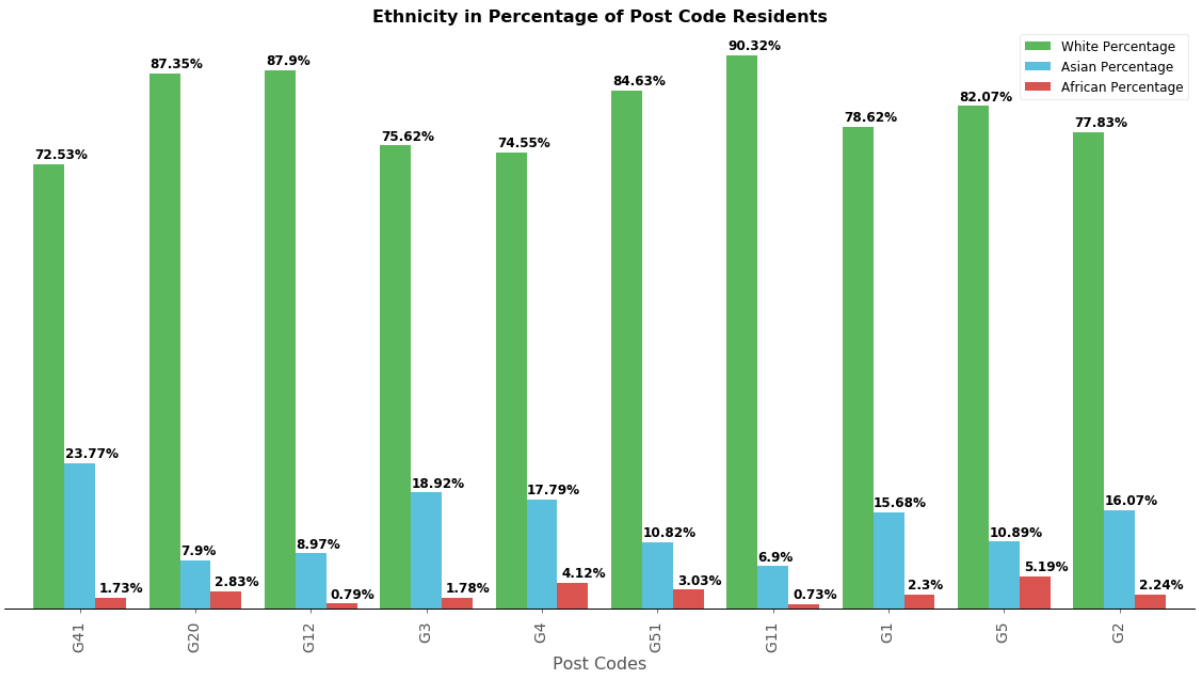

Figure (6): Ethnicity breakdown at each postal code

In order to better visualize the similarity of the ethnicities at each postal code, doughnut plots of ethnicity percentages are plotted in figure (7). As can be seen from the doughnut plots, some post codes have similar population diversity like G12 and G11. This similarity maybe due to the geographical proximity of the postal codes to each other. G4 and G3 are neighbouring postal codes and they have similar population diversity as well. In order to better cluster these postal codes, machine learning will be used.

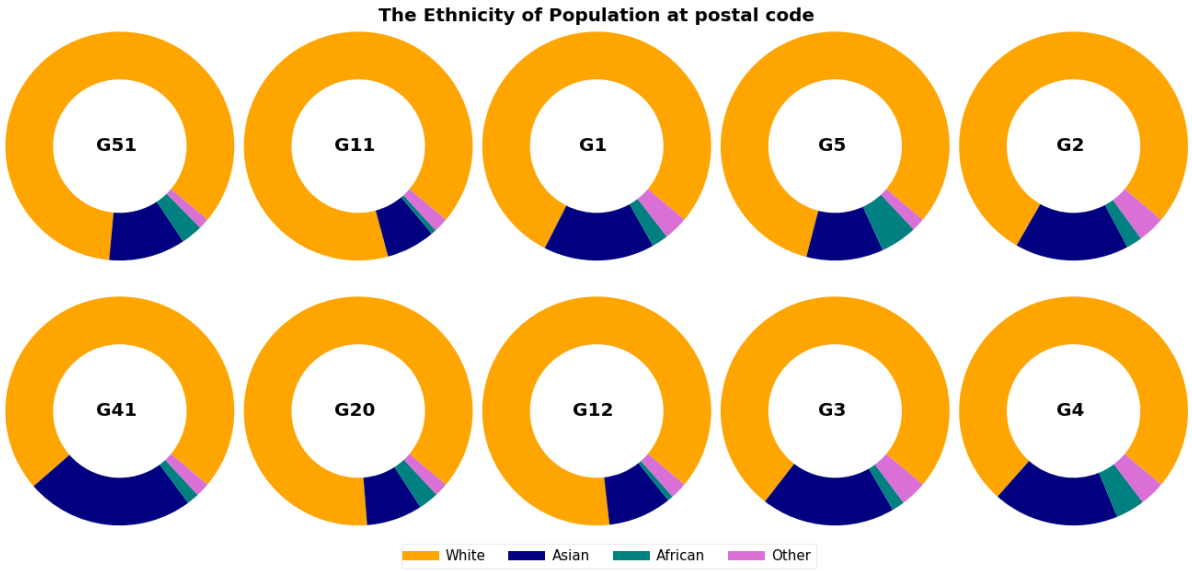

Figure (7): Ethnicity doughnut plots of each postal code

Dataset 5: The venues data. This data set is acquired from the Foursquare API

Shortlisted Post codes
G1 G2 G3 G4 G5 G11 G12 G20 G41 G51

The final shortlisted postcodes are listed in the table above. Venues in these ten postal codes are extracted. All venues within a 1.5 km radius of the postal code are imported from the API. The search area is shown as purple in the map below.

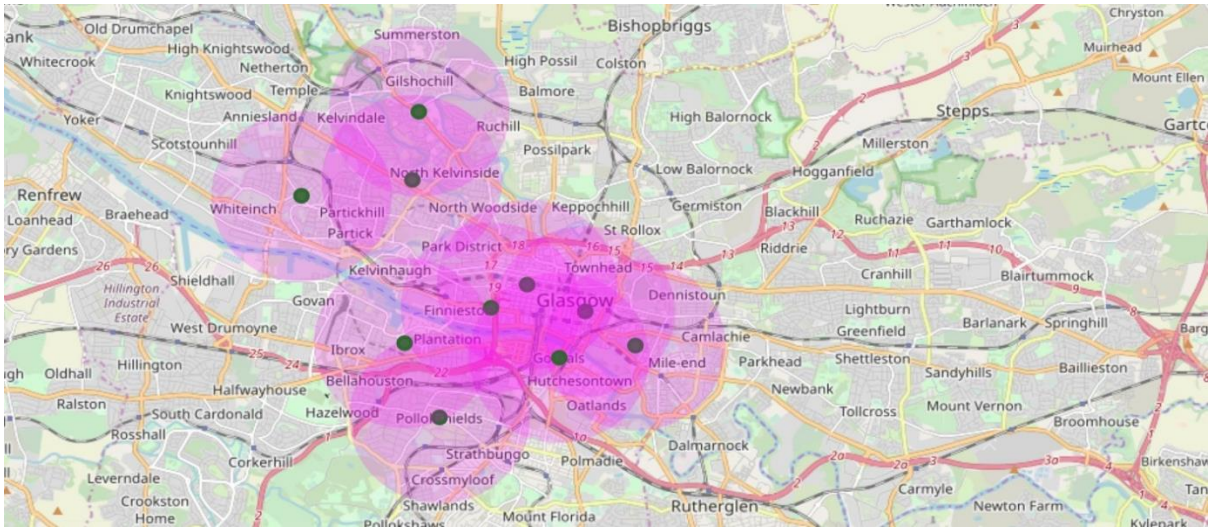

Figure (8): The search area for the venues

Other postal codes border the shortlisted postal codes. So, when 1.5 km radius is specified as a parameter of search in the Foursquare API, it might result in returning venues that belong to the neighbouring postal codes as well. These venues will be removed in the data processing stage. Also, there is a lot of overlap between the areas of search. Therefore, the FourSquareAPI will return the same venues multiple times. The repeating entries will be removed in the data processing stage. The API returns a total of 786 venues. The breakdown of the venues is given in the table below.

Postal Code	G1	G2	G3	G4	G5	G11	G12	G20	G41	G51	Total
Venues Returned	96	96	96	93	94	53	92	19	58	89	786

The 786 data entries are processed in order to remove the repeating venues. This reduces the dataset to 420 unique venues in the 10 postal code. 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.

Methodology

The project is divided into 4 stages. The stages are explained in detail below.

Stage 1-

In this stage, the postal codes are shortlisted based on whether they are 5 km from the city centre and whether they are within 2 km of a subway station. The first step of the project is to import the list of all the postal codes from the Wikipedia webpage in [5]. This dataset is imported using the BeautifulSoup library. The imported dataset is processed to remove the non-geographical postal codes. This reduces the data set to 50 valid entries. Next, the latitude and longitude of each postal code is found using the geolocator.geocode function. This information is then merged with the original data frame.

Like many big cities, most of the economic and business activity in Glasgow is focused around the city centre. Postal codes that are closer to the city centre are considered more suitable locations for opening businesses because of the myriad of opportunities that they offer business owners. In order to shortlist the postal codes, the distance to the city centre is determined by using the geographical coordinates of the city centre and the haversine formula to find the distance between the city centre and each of the 50 postal codes. Only the 22 Postal codes that are within 5 km of the city centre are considered for further analysis.

In order to further shortlist the postal codes, Dataset 2 is imported. It is publicly available at [6]. It contains the addresses of each of the 15 stations in Glasgow. The addresses are fed into the geolocator.geocode function to find the latitude and longitude of each of the subway stations. In order to find the postal codes that are within 2 km of the subway stations, the haversine formula is used to find the distance of each postcode to the nearest subway station. Out of the 22 postal codes, only 10 meet this requirement. These 10 postal codes are the ones whose population characteristics are analysed in stage 2.

Stage 2-

In this stage of the project, the population data is extracted using two datasets. The first dataset used is the Royal Mail dataset found at [7]. This dataset maps the postal codes to data zones. It is used to find the unique data zones for each of the 10 postal codes. The second dataset is the Glasgow City Council dataset which contains the population data collected in the last national census. This dataset maps data zones to population. So, in this stage of the project the unique data zones collected from the royal mail dataset is used to sum the population statistics for each of the postal code. The data extracted from the datasets are

- All people (total population at post code)
- White (total number of people having White ethnicity)
- White Percentage (Percentage of people at post code having White ethnicity)
- Asian (total number of people having Asian ethnicity)
- Asian Percentage (Percentage of people at post code having Asian ethnicity)
- African (total number of people having African ethnicity)
- African Percentage (Percentage of people at post code having African ethnicity)
- Other (total number of people having ethnicity other than White, Asian and African)
- Other Percentage (Percentage of people at post code having ethnicity other than White, Asian and African)

This information along with the top venues in each of the postal code are feed into an unsupervised machine learning algorithm for clustering into five categories.

Stage 3-

In this stage of the project, all venues information at each of the 10 postal codes is extracted. The search area of 1.5 km around the postal codes is defined as the search area. The credentials are defined to communicate with the API. The API is used to explore a location using the url below.

```
https://api.foursquare.com/v2/venues/ explore ?
client_id= CLIENT_ID &client_secret= CLIENT_SECRET &ll= LATITUDE , LONGITUDE &v= VERSION &limit= LIMIT
```

The limit is set as 100 for each postal code. The API returns the venues within 1.5 km of each postal code location. Due to the area overlap, venues are returned multiple times. Venues from the neighbouring postal codes are also returned. A total of 786 venues are returned. These are processed to remove venues from neighbouring postal codes and the repeating venues. There are 420 unique business venues in the ten shortlisted postal codes. There are 112 unique venue categories these will be one hot encoded in the next stage.

Stage 4-

The first three stages of the project are mainly focused with data extraction and transformation. In this stage, the data sets extracted in the previous stages are prepared, normalized and merged to be fed into the unsupervised learning algorithm. The K-means algorithm is used to group the postal codes based on their similarities to each other. The categories of the venues are one-hot encoded. The population dataset is normalized. The one-hot encoded venue categories dataset and the population characteristics dataset are merged. This is feed to the K-means clustering algorithm. The results of the clustering algorithm are discussed in the next section.

Results

The top five venues of each of the ten post codes are determined. The table below shows the results. Food venues are common in most of the Postal codes except for G51 and G5.

	Postcode district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	G1	Coffee Shop	Bar	Pub	Seafood Restaurant	Café
1	G11	Café	Pub	Supermarket	Restaurant	Chinese Restaurant
2	G12	Bar	Coffee Shop	Café	Pub	Italian Restaurant
3	G2	Bar	Hotel	Café	Italian Restaurant	Chinese Restaurant
4	G20	Bakery	Soccer Stadium	Gym	Sandwich Place	Pub
5	G3	Restaurant	Indian Restaurant	Bar	Coffee Shop	Hotel
6	G4	Café	Deli / Bodega	Grocery Store	Indian Restaurant	Pub
7	G41	Café	Pub	Fast Food Restaurant	Sandwich Place	Coffee Shop
8	G5	Hotel	American Restaurant	Fast Food Restaurant	Movie Theater	Seafood Restaurant
9	G51	Hotel	Sandwich Place	Lebanese Restaurant	Climbing Gym	Scenic Lookout

The venues data and the population data are fed to the K-means clustering algorithm. The 10 shortlisted postal codes shown above are clustered into 5 categories. The results of the clustering by the unsupervised machine learning algorithm are given in the table below. The five clusters are colour coded and shown in the map in figure (9).

Postal Code	G3	G4	G41	G5	G51	G11	G12	G1	G2	G20
Cluster Number	1	1	1	2	2	3	3	4	4	5
Colour on map	Red	Red	Red	Purple	Purple	Blue	Blue	Green	Green	Orange

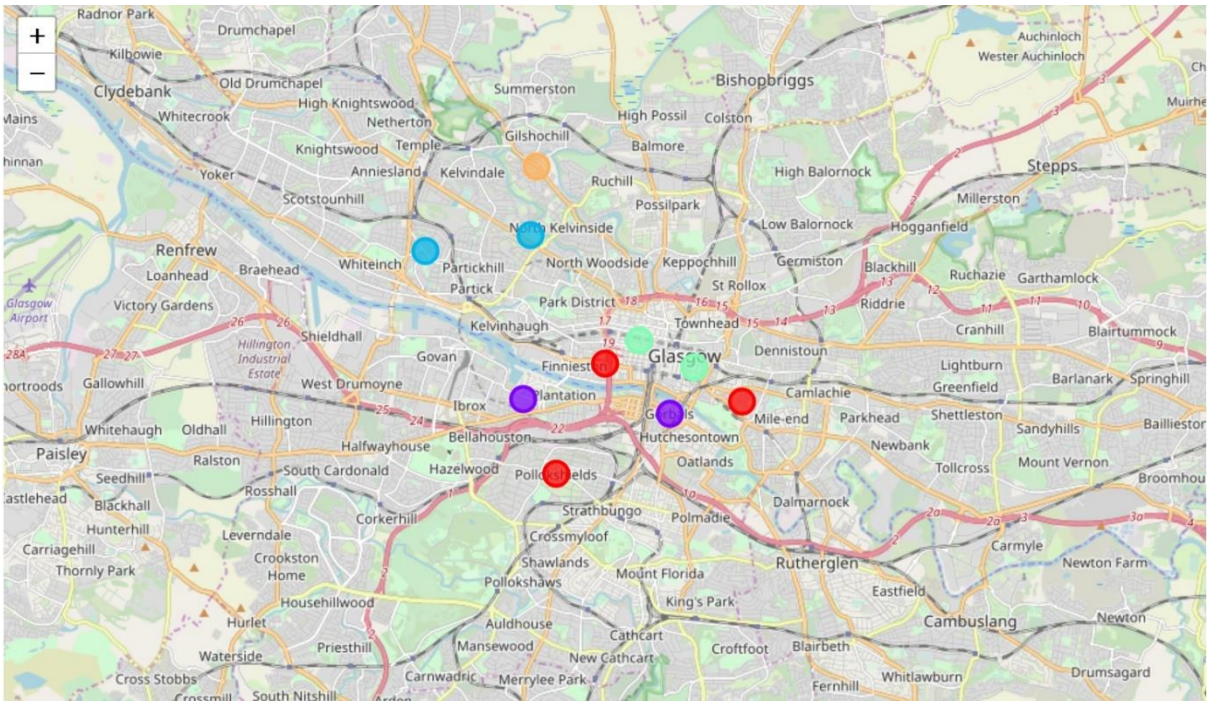

Figure (9): Clustered postal codes

Discussion

In this section, similarities between the postal codes in the cluster are identified and possible opportunities in those clusters are also identified for businesses.

Cluster 1: G3, G4, G41

This Cluster has member postal codes with the highest population diversity. It seems that food related venues are really common in these postal codes.
Possible Opportunities at these postal codes- This cluster has among the highest Asian Population density in Glasgow. Asian themed businesses are likely to succeed in these three postal codes. So, these three postal codes G3, G4 and G41 are recommended as venues for new Asian businesses.

Cluster 2: G5, G51

This cluster has a lot of hospitality and entertainment related venues.
Possible opportunities at these Post codes- At these postal codes there are not a lot of bars/pubs. So, in G5 and G51 new bars and pubs will face little competition. So, they are identified as possible business opportunities in this cluster.

Cluster 3: G11, G12

This cluster has medium total population and low diversity. Food related venues are common in this cluster.
Possible opportunities at these Post codes- These Postal codes are close to one of the largest universities in Glasgow. So, G11 and G12 offer opportunities for student service businesses like books, stationary, printing etc. These businesses have the potential to succeed in this area and have little competition at these post codes.

Cluster 4: G1, G2

These post codes have low population and medium diversity. Food related venues are common in this cluster. These postcodes also have the lowest distance from the city centre.
Possible opportunities at these Post codes- In this cluster, postal codes are near the city centre. So, at G1 and G2 there are opportunities for businesses that want to provide services to tourist like tour operators, souvenir shops etc.

Cluster 4: G20

This cluster has only one member. It is furthest away from the city centre and has a high population.
Possible opportunities at this Post code- This post code lacks entertainment venues. So new business can explore opportunities like opening a movie cinema or music venue at G20. This might be popular with residents who do not like travelling all the way to the city centre for entertainment.

Conclusion

In this Project, publicly available datasets from the Glasgow City Council and the Royal Mail are used alongside data collected from the FourSquareAPI in order to obtain data driven insight into possible business opportunities in Glasgow and the best venues for those opportunities. The summary of the results are as follows.

K-means Cluster label	0	1	2	3	4
Postal codes in the cluster	**G3, G4 and G41**	**G5 and G51**	**G11 and G12**	**G1 and G2**	**G20**
Description of the clusters	Cluster members have the highest population diversity and food related venues are the commonest in this cluster.	This cluster has a lot of hospitality and entertainment venues in the postcode	This cluster has medium population and low diversity. Food related venues are common in this cluster	These post codes have low population and medium diversity. Food related venues are common in this cluster. These postcodes also have the lowest distance from the city centre.	This cluster is furthest away from the city centre and has a high population with no entertainment venues
Possible Business Oportunities in this Cluster	These Postal codes have among the highest Asian population density in Glasgow. Non-food Asian themed businesses are likely to face little competition in these three postal codes. So, G3, G4 and G41 are recommended as venues for new Asian businesses.	In this cluster, there are a lot of hotels, bowling alleys, music venues and not a lot of food venues like restaurants, bars/pubs etc. So, new restaurants, bars and pubs will face little competition. So, they are identified as possible business opportunities in this cluster.	The Postal codes in this cluster are close to one of the largest universities in Glasgow. So, this cluster offers opportunities for student service businesses like books, stationary, printing shops etc. These businesses have the potential to succeed in this area and have little competition at these postcodes.	This cluster has postal codes that are near the city centre. So, at these postal codes there are opportunities for businesses that want to provide services to tourists like tour operators, souvenir shops, hotels etc.	This post code lacks entertainment venues. So, in G20 there are opportunities for entertainment & leisure businesses like movie cinemas, music clubs etc that want to target a new sector of the Glasgow population.

G3, G41 and G4 are suitable for Asian non-food businesses. G5 and G51 are suitable for opening bars and pubs. G11 and G12 are suitable fore businesses targeting students. G1 and G2 are suitable for businesses targeting tourists. G20 is suitable for entertainment related businesses.

References:

[1] https://www.citypopulation.de/en/uk/scotland/glasgow_city/S19000899__glasgow/
[2] https://www.glasgowlive.co.uk/news/glasgow-news/glasgows-population-continues-rise-the-14550988
[3] http://www.spt.co.uk/documents/latest/rtp220618_agenda6.pdf
[4] https://www.statista.com/statistics/305479/passenger-journeys-on-the-glasgow-underground-united-kingdom-uk/
[5] https://en.wikipedia.org/wiki/G_postcode_area
[6] https://data.glasgow.gov.uk/dataset/glasgow-subway-station-locations/resource/8c057567-55e7-4631-ae77-e37b4a0562e8
[7] http://www.glasgow.gov.uk/CHttpHandler.ashx?id=16002&p=0
[8] https://data.glasgow.gov.uk/dataset/national-identity-by-ethnic-group-2011-census