Out[2]:

Coursera - Applied Data Science Capstone

Introduction

London is the capital of the United Kingdom. With a Metropolitan area spanning 3.2 thousand square miles & a population of 14 million people London exerts significant influence in areas such as finance, politics, the arts, culture & sport. Amongst this the London Underground is central to enabling London to function on a daily basis.

The London Underground first opened in 1863 & now operates 270 stations. With an estimates 5 million people travelling each day it is central to ensuring London functions smoothly. With such high importance proximity to a station can have a major impact on any decision.

My project will focus on reviewing the area around tube stations & clustering these together. This would allow those looking to expand a business to identify areas most suitable for opening

Data

Station Data

To get the latitude and longitude of tube stations within London I've pulled data from Open Street Map (https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations). This data contains 302 location points though some of these are for separate parts of the same location. Where possible data used was taken from the platform to remove duplication

As we can see the station data contains several useful fields. For this analysis we'll just be retaining name, latitude & longitude

Out[15]:

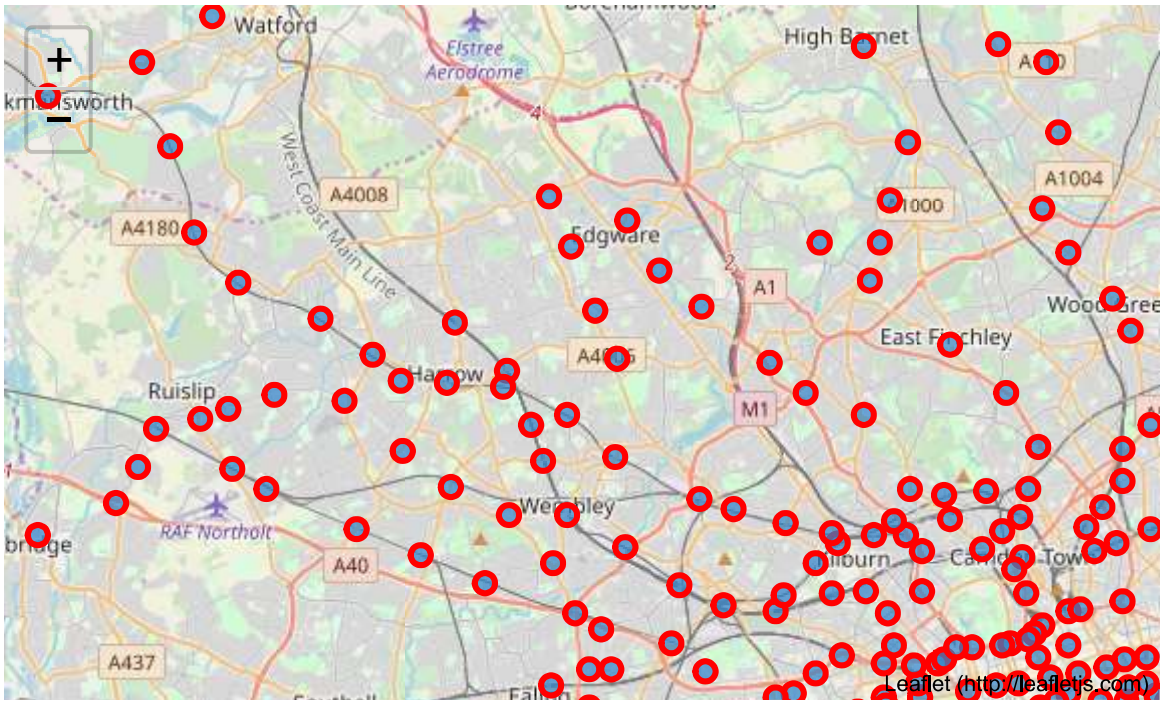
	name	latitude	longitude	platform / entrance	collected by	collected on	
0	Acton Town	51.502500	-0.278126	Platform	User:Gagravarr	24/11/06	Distric Piccac
1	Acton Central	51.50883531	-0.263033174	Entrance	User:Firefishy	08/05/2007	Londo Overg
2	Acton Central	51.50856013	-0.262879534	Platform	User:Firefishy	08/05/2007	Londo Overg
3	Aldgate	51.51394	-0.07537	Aldgate High Street entrance	User:Morwen	28/4/2007	Metrop
4	Aldgate East	51.51514	-0.07178	Entrance	User:Parsingphase	(2006)	Distric Hamm & City

Out[16]:

	name	rows
276	West Hampstead	4
0	Acton Central	2
285	Willesden Junction	2
240	Stratford	2
206	Richmond	2

Using Folium we're able to overlay a map of London with the data points that we've collected. We can see quite clearly that central London is quite densely covered by stations. While there are some isloated stations that exist on commuter lines

Out[22]:



Venue Data

Data for venues surrounding tube stations has been collected using the Foursquare API. This provides important information such as Venue name, category and latitude & Londitude

For this analysis the data has been restricted to just 100 venues within a 500m radius of the stations location

Out[68]:

	station	station latitude	station longitude	venue	venue latitude	venue longitude	venue category
0	Acton Central	51.508835	-0.263033	The Station House	51.508877	-0.263076	Pub
1	Acton Central	51.508835	-0.263033	Acton Park	51.508595	-0.261573	Park
2	Acton Central	51.508835	-0.263033	The Rocket	51.508772	-0.263787	Pub
3	Acton Central	51.508835	-0.263033	Everyone Active	51.506608	-0.266878	Gym / Fitness Center
4	Acton Central	51.508835	-0.263033	Putt in the Park	51.508583	-0.260178	Mini Golf

Exploratory Data Analysis

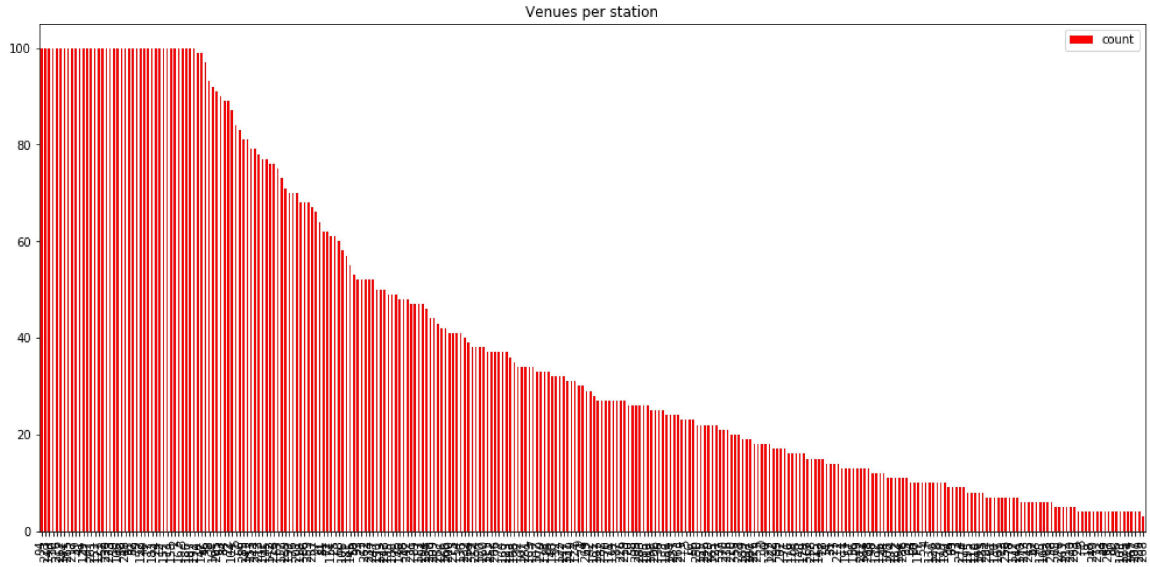
Prior to clustering any of the data we want to get a feel for venues data by performing some simple data analysis

As we'd expect stations that are in more central areas of London such as Waterloo & Camden are surrounded by a large number of venues. This is primarily to cater to tourists and the large number of businesses in the area

Stations outside of central London typically have less surrounding venues, this is likely down to being residential areas rather than commuter venues

Out[26]:

	station	count
94	Goodge Street	100
223	South Kensington	100
149	Leicester Square	100
30	Brixton	100
216	Shepherd's Bush	100



Coffee shops are the most frequent venue category within 500m of 66 stations, followed by cafes at 50 stations.

Out[30]:

	venue category	station
20	Coffee Shop	66
14	Café	50
40	Pub	35
28	Grocery Store	21
31	Hotel	18
32	Indian Restaurant	10
3	Bakery	8
19	Clothing Store	8
13	Bus Stop	6
22	Convenience Store	5

Stations that are situated in large tourist areas tend to display a greater variety of venue categories within the immediate radius

Out[31]:

	station	venue category
10	Baker Street	66
248	Tottenham Court Road	64
121	Holborn	63
40	Canary Wharf	63
192	Piccadilly Circus	62
30	Brixton	60
265	Warren Street	60
149	Leicester Square	60
153	London Bridge	60
216	Shepherd's Bush	60

Clustering stations

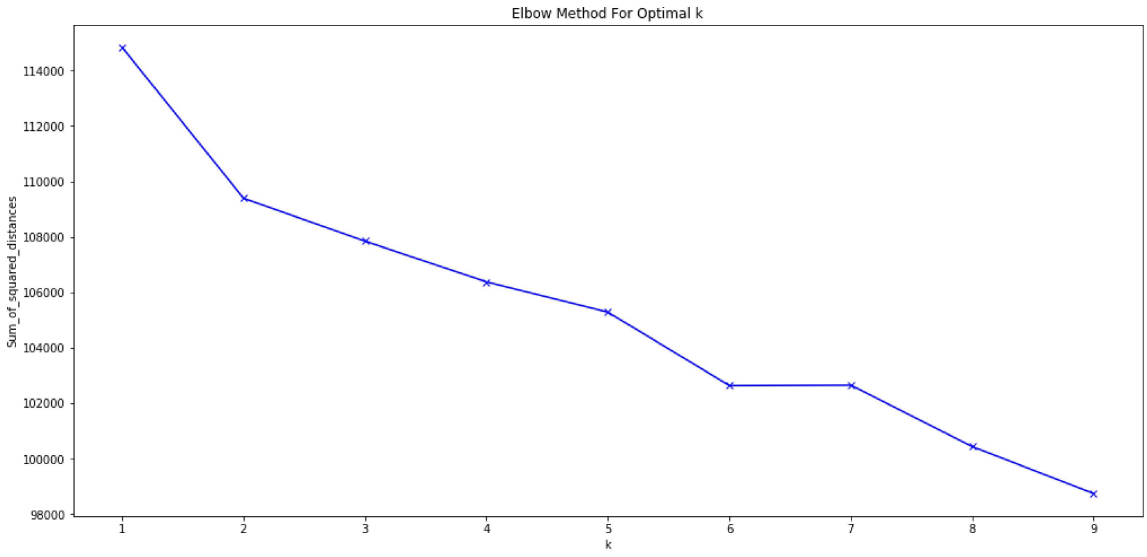
To cluster each station we'll create a vector representation using venue categories. Stations will be clustered together using KMeans. As we need to specify the number of clusters we'll identify this using the Elbow method.

Out[33]:

venue category	Accessories Store	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Antique Shop
station							
Acton Central	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Acton Town	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aldgate	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aldgate East	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Alperton	0.0	0.0	0.0	0.0	0.0	0.0	0.0

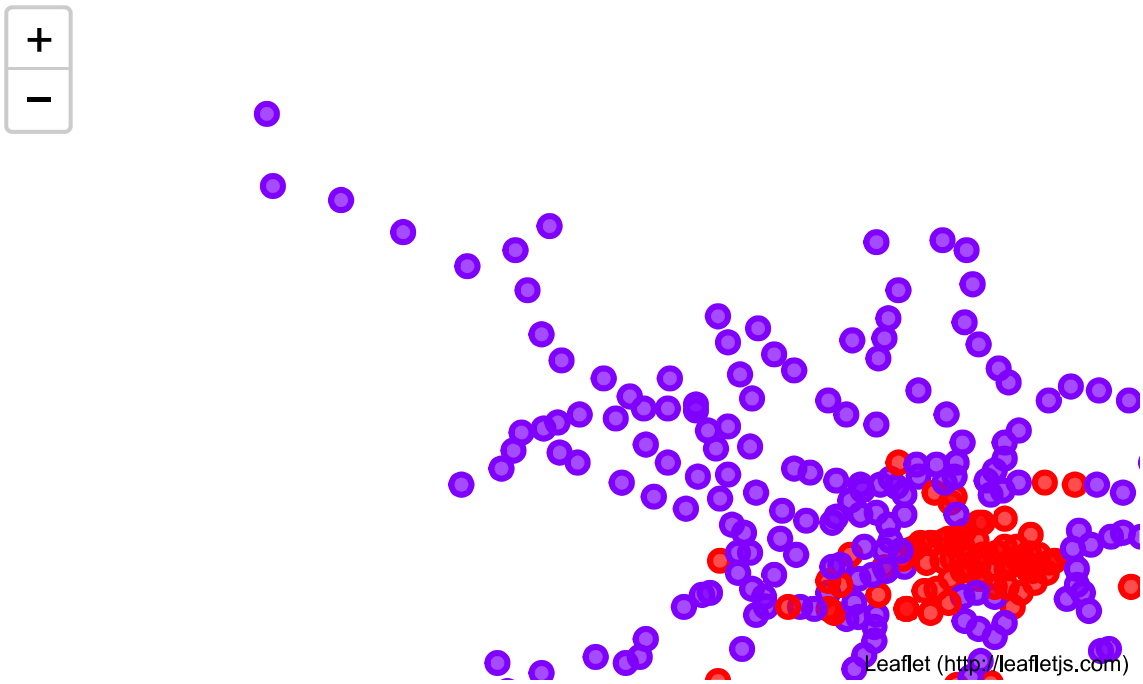
5 rows × 396 columns

Based on the Elbow plot I selected 2 different clusters for my analysis



After plotting the clusters we can see that one cluster seems to be predominatly covering zone 1 of London. This is generally the densest area and contains numerous shopping, entertainment and business districts

Out[41]:



Conclusions

London tube stations appear to be effectively clustered into two groups

This does appear to have sound rationale as the center cluster mostly covers zone 1 which is more dense and contains shopping and business districts

The second clusters mostly appears to be concentrated in more residential areas

Potential future analysis

Extend the Foursquare area to use more data rather than just 100 merchants

Using house price data provided by the UK government could further enrich the clustering and provide more insight

Utilising the users API from Foursquare one could start to look into building connections amongst different tube stations. This would open the way for some interesting applications of Network and graphical models