Applied Data Science Capstone

Markos Lolos

08 Nov 2019

Problem Definition

- Currently Brexit is affecting the City of London and many financial institutions are moving to other cities in Europe (a popular one is Paris) in order to continue servicing the European customers without interruptions and amendments. As a result many financial institutions are asking their employees to move to these cities. Let's explore which areas of these cities are similar to the areas of London, so the employers can give an informed advice to the employees.
- The goal of these project is to examine which areas of the other two cities are similar to areas of London. Based on this clustering the employees will be able either to choose the most similar area to the one they live now, or have an informed comparison while choosing the area of their new home.

Data Requirements

- The data we are going to need are the several districts of London and Paris. We are going to acquire these form the following links:
- London: https://www.milesfaster.co.uk/london-postcodes-list.htm
- Paris: https://www.annuaire-administration.com/code-postal/region/ile-de-france.html
- Additionally, we will need the data related with the venues in its city. We are going to acquire these by using the Foursquare API as required by the instructions of this project.

Data Wrangling

- We downloaded the data and created dataframes which included the information/variables required to complete the assignment
- We printed and plotted the data to see that we get the right inputs
- We found the common variables which are present in both cities based on which we would be the variables used to classify the location using k-Means.
- We transformed the categorical variables to numerical ones in order to facilitate the training of the machine learning algorithm

Methodology

- We use the mutually existing most common locations per neighborhood.
- We train the algorithm using the location data of London
- Then we predict the cluster that each of the districts of Paris should fall under

Methodology -Model Training

Let's train our model with *k*-means to cluster the neighborhoods of London using only the common location categories.

In [51]:
```
# set number of clusters
kclusters = 5

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0,n_init=20).fit(ldn_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Out[51]:
```
array([1, 4, 0, 0, 1, 0, 0, 0, 0, 2], dtype=int32)
```

In [52]:
```
# add clustering labels
neighborhoods_venues_sorted['Cluster Labels']=kmeans.labels_

ldn_merged = df_ldn2

#merging the two tables
ldn_merged = ldn_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Areas')

ldn_merged.head() # check the last columns!
```

Methodology -Model Prediction

Let's predict now the cluster of each area in Paris for the same parameters.

In [79]:
```
# run a test for k-means prediction
kmeans.predict(prs_grouped_clustering)[0:20]
```

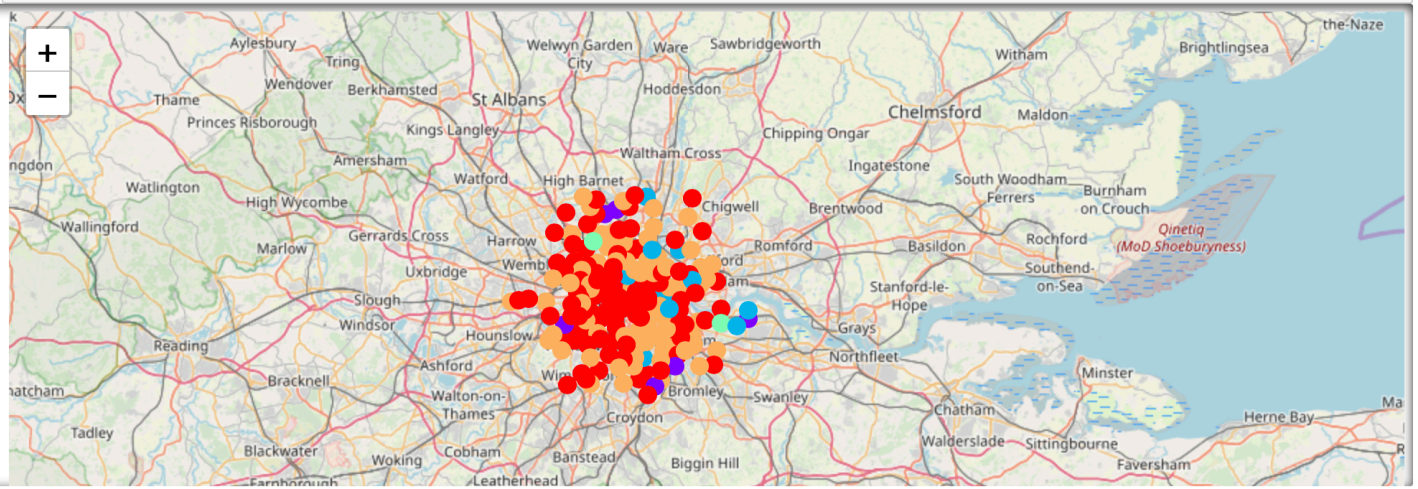
Out[79]:
```
array([0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0],
      dtype=int32)
```

In [80]:
```
# add clustering labels
prs_neighborhoods_venues_sorted['Cluster Labels']=kmeans.predict(prs_grouped_clustering)
prs_merged = df_prs
prs_neighborhoods_venues_sorted.head()
```

Methodology -London classification

```
map_clusters
```

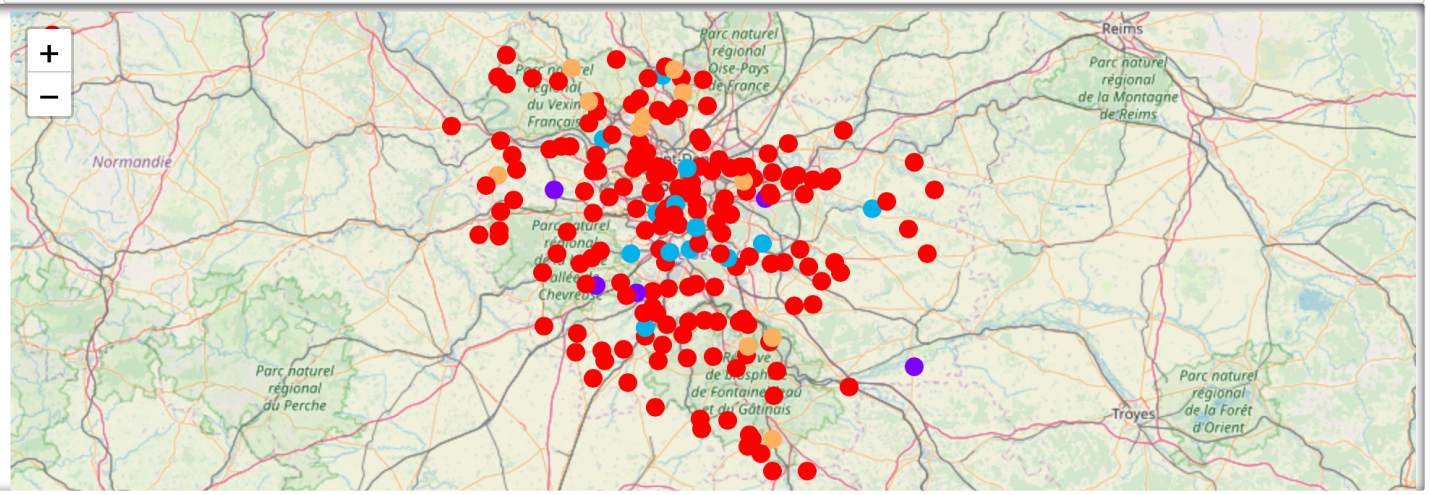
Out[53]:



Methodology -Paris prediction

```
map_clusters
```

Out[59]:



Results - London

We notice that the there are three clusters that gather the majotiry of the locations in London. Let's examine the characteristics of each cluster to learn more about these locations.

In [60]:
```
ldn_merged.groupby('Cluster Labels').count()
```

Out[60]:

Cluster Labels	Postcodes	Areas	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	111	111	111	111	111	111	111	111	111	111	111	111	111	111
1	6	6	6	6	6	6	6	6	6	6	6	6	6	6
2	11	11	11	11	11	11	11	11	11	11	11	11	11	11
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2
4	76	76	76	76	76	76	76	76	76	76	76	76	76	76

Results -Clusters

- Cluster 0 includes rural areas that have great access to shopping stores, grocery stores, activities, and restaurants. We can say that these areas are best suited to people who like to be near active day life for convenience. These places also look extremely central. with great access to gym, theatres and in some occassions hotels.
- Cluster 1 includes areas which look to be more decentralised with popular locations being, grocery stores, train stations and restaurants.

Results -Clusters

- Cluster 2 most common area are the parks of London, however they are locations around the city rather than parks in central London. Other popular areas are pubs, restaurants, pizza places and dry cleaners.
- Cluster 3 includes areas similar to the ones in Cluster 2. Probably these two clusters could have been classified as one.
- Cluster 4 has areas which most popular locations are pubs. Other popular areas are convinience stores, coffee shops, gyms and restaurants. We can conclude that these areas are central, but residential friendly for young professionals as they are not as crowded by restaurants as cluster 3 and have many ammenities near them.

Results -Paris

Prediction review for Paris based on London classification

In [66]:
```
prs_merged.groupby('Cluster Labels').count()
```

Out[66]:

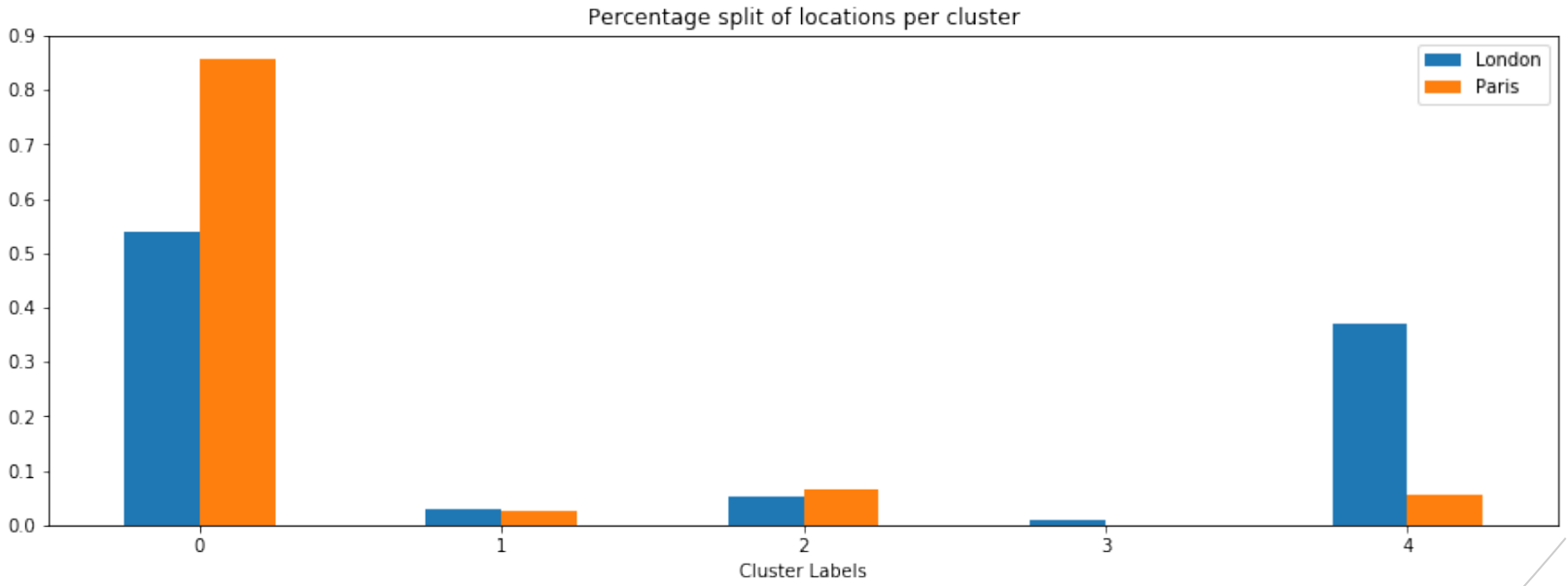
Cluster Labels	Postcodes	Districts	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	174	174	174	174	174	174	174	174	174	174	174	174	174	174
1	5	5	5	5	5	5	5	5	5	5	5	5	5	5
2	13	13	13	13	13	13	13	13	13	13	13	13	13	13
4	11	11	11	11	11	11	11	11	11	11	11	11	11	11

A problem that we faced which we will discuss later on in this assignment is that the two locations had only 282 type of locations similar, which made it impossible to get a result for some of the areas in Paris, because they had completely different popular locations.

Results


Number of locations per cluster
London
Paris
175
150
125
100
75
50
25
0
0
1
2
3
4
Cluster Labels

Results


Percentage split of locations per cluster
London
Paris
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
Cluster Labels

Discussion

- We can conclude that the first category is the dominant category in areas of London and Paris too. This category represents central shopping areas good for people who would like to be in the centre of their neiborhoods.
- The biggest difference is that for London these location account for approximately half the areas of London. On the other hand for Paris the percentage is much higher close to 85%.
- This is probably because the prediction algorithm took out most of the areas of Paris as the common location with London were not satisfactory to make a prediction. But we will explain this at the end of this section.

Discussion

- From the sample reviewed, the second category (Cluster 1) being the decentralised areas are almost equally weighted for both cities at a level of 2.5-3.0%.
- The third category (Cluster 2) being the areas near the parks are also almost equally weighted for the two cities.
- As mentioned earlier at the algorithm training paragraph, Cluster 2 and 3 should probably be the same category. This is also reflected in the fact that there were 0 predicted areas in Paris for Cluster 3.
- Given this facts and assuming that these two clusters converge on the type of areas we can say that these clusters are similar in percentage content for the two cities. More specifically by adding the percentages they approximately 6.2-6.4%.

Discussion

- Lastly, the last category (Cluster 4) is what looks to significantly differentiate the two cities. Cluster 4 is heavy in Pubs inclusion. This is a cultural element of the UK (and respectively London), which makes London so more different than the other cities of Europe. The content of pub doinated areas in London is 35-40% . On the other hand Paris only has approximately 5% of its areas being pub dominated.

Limitations

- In the process of completing this assignment we met a challenge; in some occassions the most popular places in London were completely different than popular areas in Paris and vice versa. This resulted to only use a small sample of mutually existing locations ("common_cols") to train the algorithm and predict the areas of Paris. This smaple obviously was not representative enough as in the process many areas of Paris were dropped. This is because there were not satisfactory mutually existing areas to use for the prediction. This limitation is something that we can look into exploring in the future to make the algorithm more resilient in predict areas which might have similarities but are not strictlyb identical.

Disclaimer

- This assignment was completed as part of the module "Applied Data Science Capstone". This module was completed as part of the "IBM Data Science Professional Certificate" provided on Coursera. The rights of this essay remain with its author and no copy or reproduction should be attepted without his written consent.

Thank you for reading