Capstone Project - The Battle of Neighborhoods
Coursera IBM Applied Data Science

Exploring London Tube Neighborhood using web and Foursquare location data

Introduction/Business Problem

- London Underground, better known as the Tube, has 11 lines covering 402km and serving 270 stations. The Tube handles up to 5 million passenger journeys a day. At peak times, there are more than 543 trains whizzing around the Capital
- For this project, we want to look at the neighborhood surrounding the Tube stations and classify them based on the Restaurants and Bars closest to a station. By analyzing this data, we can classify stations and explore the opportunities to start up a new Restaurant/Bar, close to Tube Stations in London, United Kingdom

Data

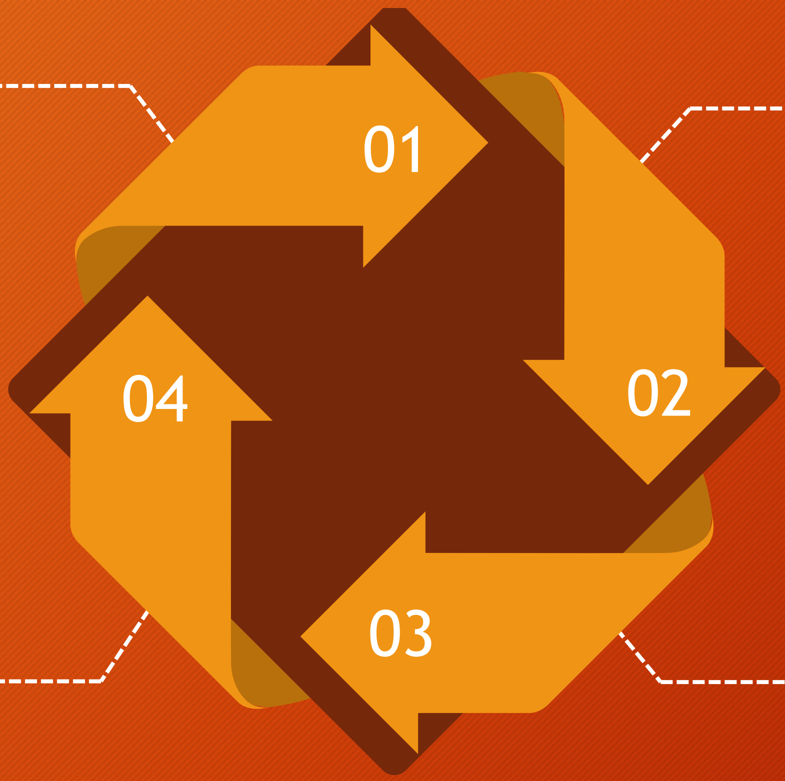
Data Required

- List of London Underground Stations
- Latitude and Longitude coordinates of the Underground Stations
- Food Venue Category data, particularly Restaurants and Bars

Data Sources

- Wikipedia pages for Neighborhood
  - https://en.wikipedia.org/wiki/List_of_London_Underground_stations
- Wikipedia pages for coordinates of the Neighborhood
  - https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations
- Foursquare API to explore food venues surrounding each station using Food sub-category id (4d4b7105d754a06374d81259)

Methodology



- Web scrapping Wiki pages for Station(Neighborhood) list & geo coordinates
- Foursquare API to explore food venues
- Filter Venue Category for Restaurants & Bars
- Data cleanup

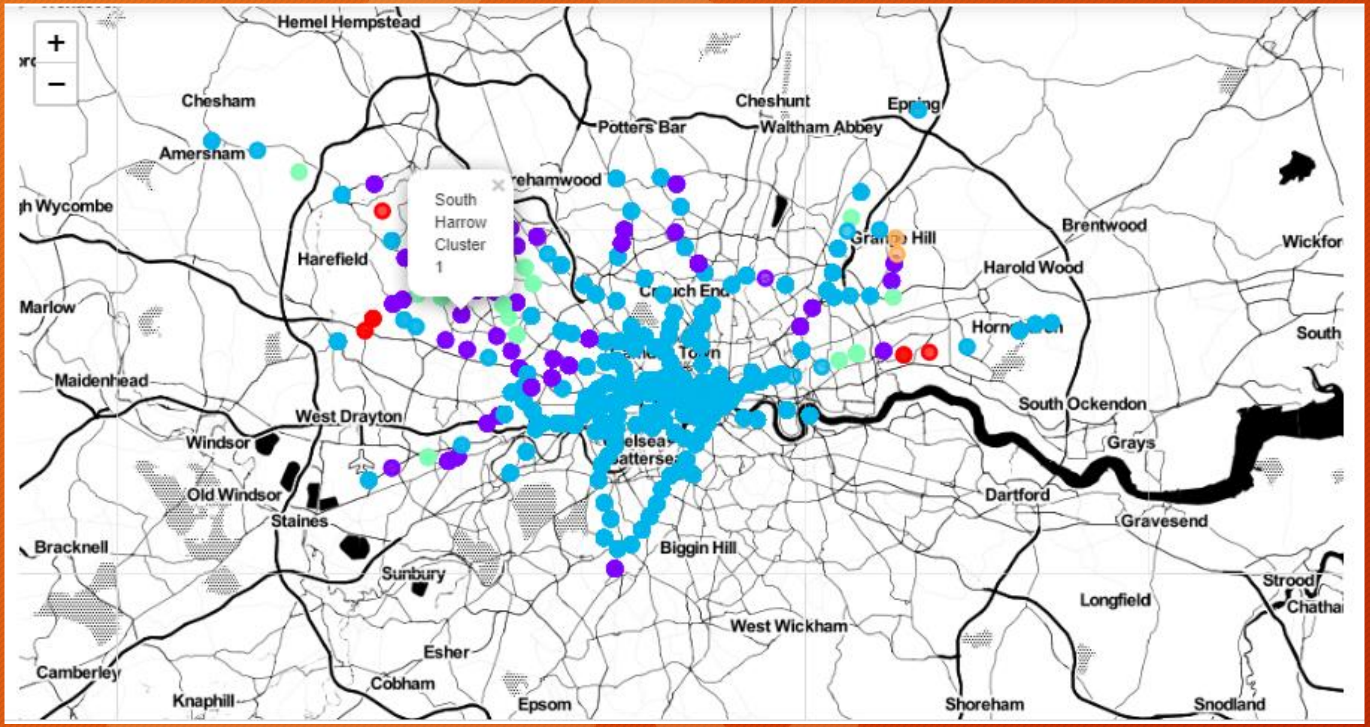
Data Exploration and Visualization using matplotlib and seaborn python library

- Feature Engineering and Clustering using K-Means
- Visualization in a map using Folium

- Examine Clusters
- Recommendations
- Conclusion

Result

Categorized the data into 5 Clusters



- ***Cluster 0 (Red):*** Chinese and Vietnamese Restaurants are predominant with no Bars
- ***Cluster 1 (Violet):*** Fast Food Restaurants are most common on this cluster
- ***Cluster 2 (Blue):*** All Restaurants & Bars are concentrated in this cluster. French, Japanese Restaurants & Bars top the list
- ***Cluster 3 (Light Green):*** Most of the Indian Restaurants with few Bars are clustered here
- ***Cluster 4 (Orange):*** Only couple of Restaurants in this cluster

Discussion

- Restaurants are more common than the Bars in each of the Station Neighborhood
- Neighborhood of London Bridge and Tooting Broadway has the highest number of Restaurants
- Liver Pool Street, Wimbledon, Tootenham Court Road Tube Stations neighborhood is dominated by Bars
- Roding Valley, Grange Hill, Buckhurst Hill to name a few are the least used Tube stations of London as a result has the least number of restaurants as per the analysis
- High Street Kensington is one of the Stations where we have the most number of Restaurants & Bars. This station is served by the Circle (yellow) line and the District (green) line, both of which are very easy to use and well-connected to major attractions in the city, justifies the reason

Recommendations

- Avoid neighborhoods in **Cluster 2**, already high concentration of Restaurants and Bars, will have intense competition

- Stations under **Cluster 4** are least used as a result has only couple of Restaurants so better to avoid

- Open new Bars in the neighborhood of **Cluster 0** with little to no competition

- **Cluster 1 & 3** are also an option for new Restaurants or Bars with moderate competition if have unique selling propositions to stand out for competition. But avoid Fast Food Restaurants and Indian Restaurants respectively

Conclusion

All of the analysis is dependent on the adequacy and accuracy of Foursquare data. Foursquare data is limited but can provide insights into a city’s development. This data could be combined with other sources to provide more accurate results. A more comprehensive analysis and future work would need to incorporate data from other external databases

Some drawbacks of this analysis are

- clustering is completely based on the most common venues obtained from Foursquare data
- results could potentially vary if we use some other clustering techniques like DBSCAN