Identifying Venue types for London Borough

IBM Capstone Project

By: Daniel Simpson

Introduction

A property developer has just finished construction on locations in each of the 33 London boroughs. The developer is interested in knowing what the best businesses are to open at these new locations. They would like to identify which boroughs are similar to each other, identify the most frequently occurring businesses within each similar boroughs group, and identify businesses to open in each borough with this information. The developer would like to utilize demographic data, political data, median salary data, and religious make-up for each borough to identify similarity between boroughs. With those similar boroughs identified, the developer would then like to use venue data on businesses within each borough.

The goal of this project is to identify similar boroughs within the London area using clustering techniques. Once boroughs have been clustered into groups, individual boroughs will be looked at to identify which venue type is lacking in respect to the cluster group, to give insight into what type of venue should be opened up within each borough.

Data:

For demographic, political, salary, and religious data, publicly open available data online will be gathered.

- Borough names, political, Area, Population, and Co-ordinates data
- Ethnicity data
- Age data
- Salary data

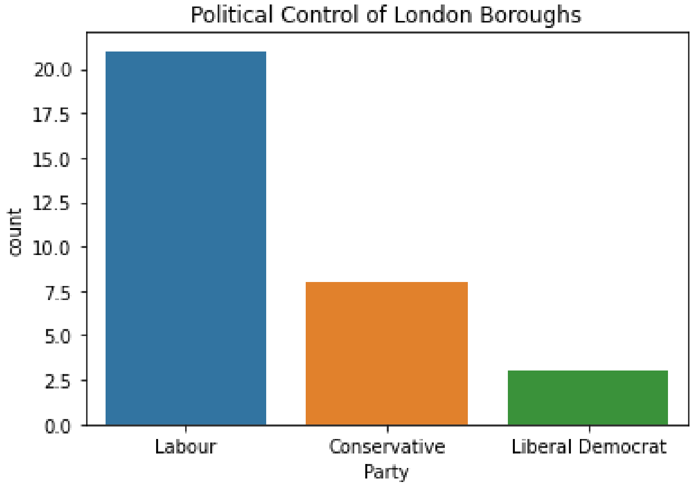
Geo data will be pulled using the Foursquare API. Web scraping will be used to gather population and demographic information for each borough within the greater London area.

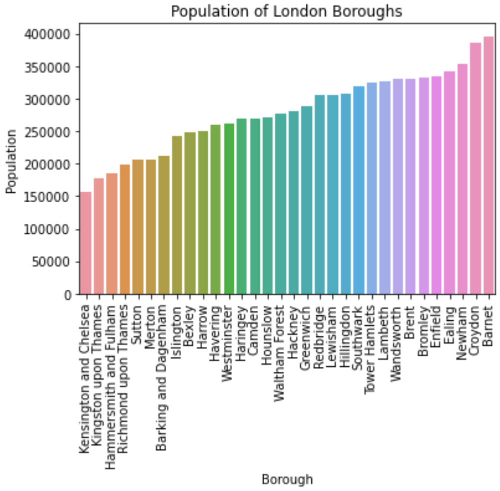
- Venue data

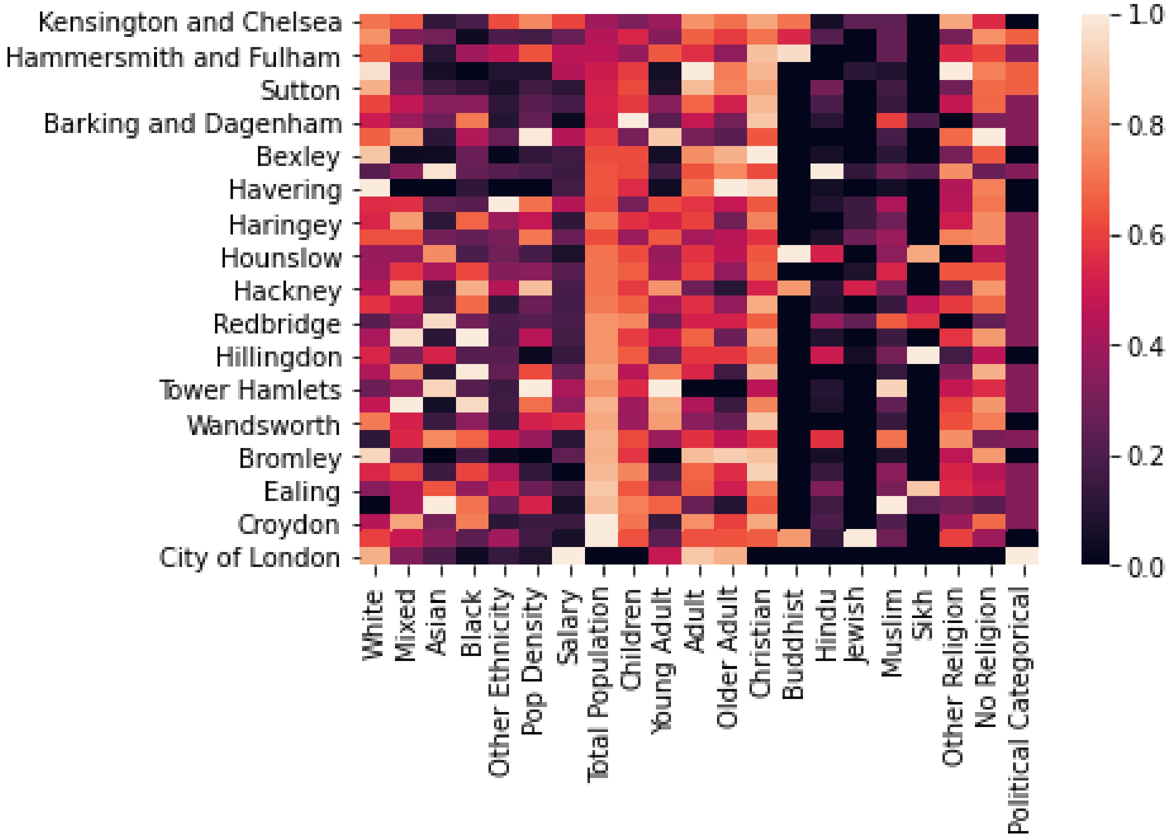
Methodology

The exploratory data analysis for each data set can be found below. Once the data was collected it was all cleaned and stored into a single Data Frame. Some of the data came in different formats so this all had to be accounted for. Every time data was collected it was observed for missing values using the missingno library. Once all the data was collected and scaled, **K-means** algorithm was used to generate similar boroughs within London based upon the demographic, political, income, and age data collected.

Once the clusters were established, then the Foursquare API was used to pull venue data within each neighborhood. The top frequently occurring venue types within each cluster were then compared to the top frequently occurring venues within each borough that fell within the respective cluster. Comparison was made to identify the most desirable venues to build within each borough.

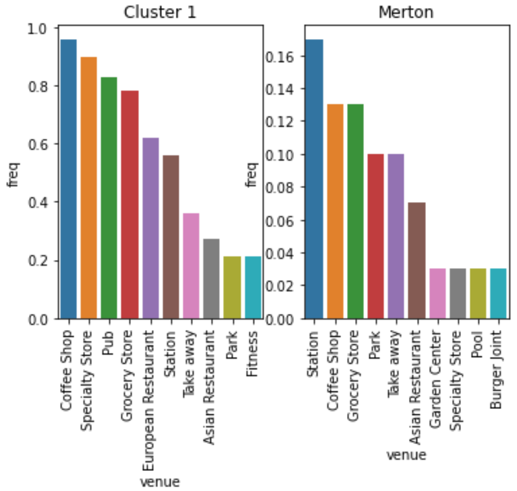


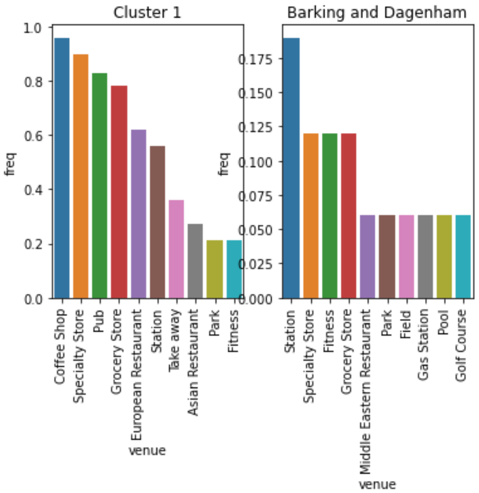


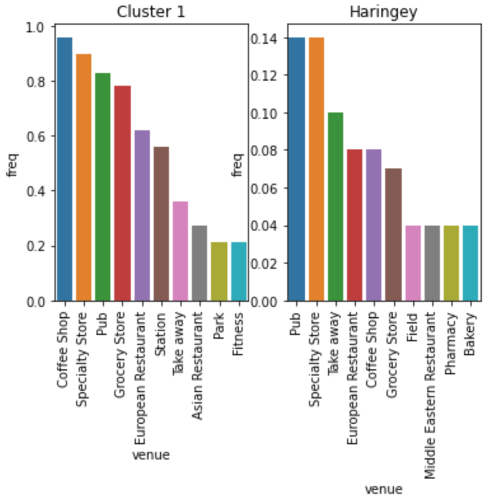


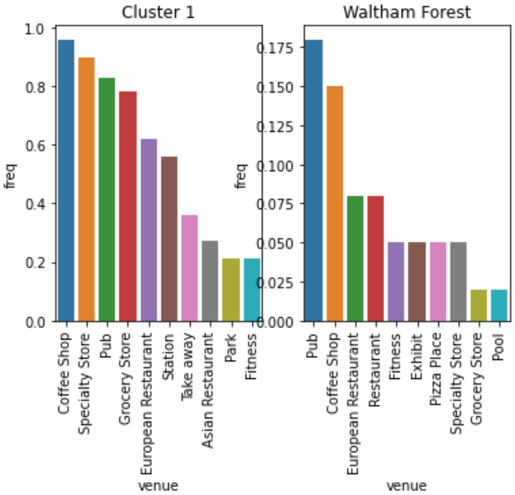
Results:

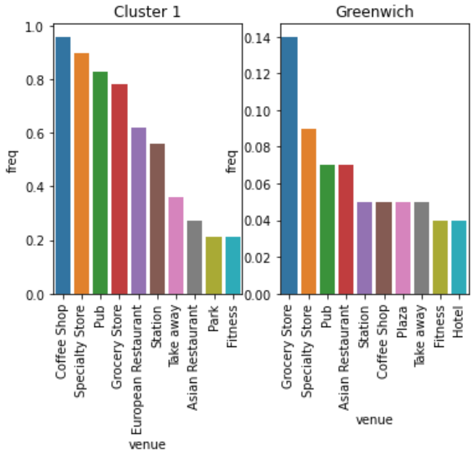
For a quick glance at all the borough's in Cluster 1 we can compare venue types:

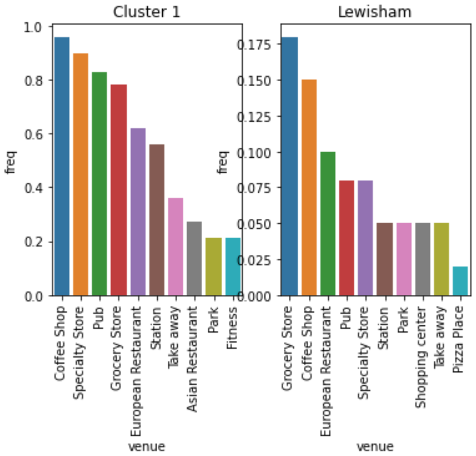


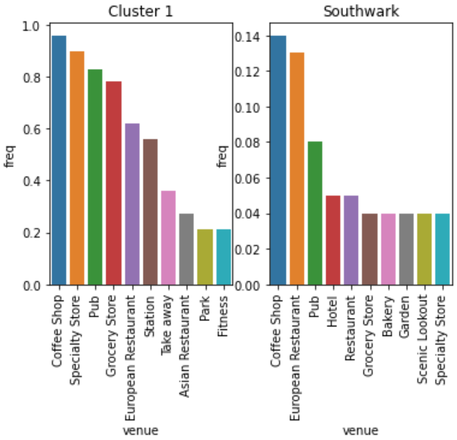


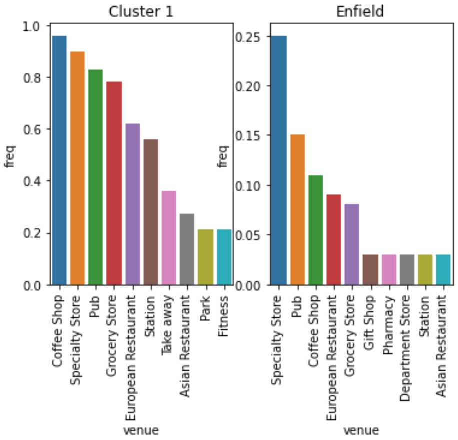


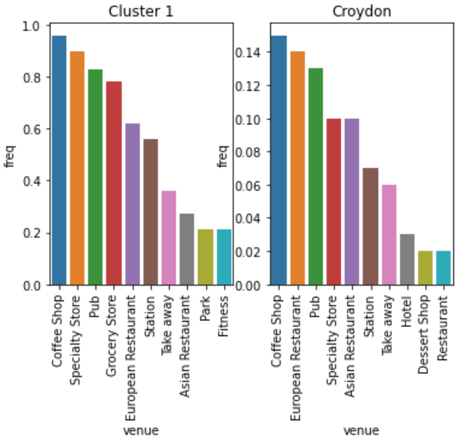







Cluster 1
freq
1.0
0.8
0.6
0.4
0.2
0.0
Coffee Shop
Specialty Store
Pub
Grocery Store
European Restaurant
Station
Take away
Asian Restaurant
Park
Fitness
venue
Southwark
freq
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Coffee Shop
European Restaurant
Pub
Hotel
Restaurant
Grocery Store
Bakery
Garden
Scenic Lookout
Specialty Store
venue


Cluster 1
freq
1.0
0.8
0.6
0.4
0.2
0.0
Coffee Shop
Specialty Store
Pub
Grocery Store
European Restaurant
Station
Take away
Asian Restaurant
Park
Fitness
venue
Enfield
freq
0.25
0.20
0.15
0.10
0.05
0.00
Specialty Store
Pub
Coffee Shop
European Restaurant
Grocery Store
Gift Shop
Pharmacy
Department Store
Station
Asian Restaurant
venue


Cluster 1
freq
1.0
0.8
0.6
0.4
0.2
0.0
Coffee Shop
Specialty Store
Pub
Grocery Store
European Restaurant
Station
Take away
Asian Restaurant
Park
Fitness
venue
Croydon
freq
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Coffee Shop
European Restaurant
Pub
Specialty Store
Asian Restaurant
Station
Take away
Hotel
Dessert Shop
Restaurant
venue

Discussion:

With the bar charts for comparison we can identify within each Borough, the best option for each of the developer's properties.

For instance, if we look at Cluster 1:

- Merton does not have a high frequency of Pubs compared to the cluster
- Barking and Dagenham does not have a high frequency of Pubs and Take away compared to the cluster
- Haringey does not have a high frequency of coffee shops and fitness centers compared to the cluster
- Waltham Forest does not have a high frequency of take away compared to the cluster
- Greenwich does not have a high frequency of European Restaurants and Grocery stores compared to the cluster
- Lewisham does not have a high frequency of Asian Restaurants and Fitness centers compared to the cluster
- Southwark does not have a high frequency of Take away and Asian Restaurants compared to the cluster
- Enfield does not have a high frequency of Take away and Asian Restaurants compared to the cluster
- Croydon does not have a high frequency of Fitness centers compared to the cluster

This comparison can be made by easily comparing the bar charts of the borough and the Cluster that borough falls within. We could then make a recommendation for all of the 33 boroughs in London.

Conclusion:

The results show us the best venue types to open at each of the developers properties around London. Additional data could be collected on the residents of each Borough, along with building information on venues not listed (such as private office spaces), to help better improve the model. Overall, the results are easy to identify and this report would be a good first step in identifying what business venues we should place at the developer's buildings.