



Segmenting and Clustering Neighborhoods- London and Paris

Lavenya Mohanasundaram
May 2021

CONTENTS



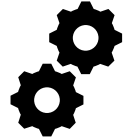
Business Case



Data Acquisition



Data Preparation



Exploratory Data Analysis



Data Visualization



Result & Discussion



Conclusion & Future Work



Segmenting and Clustering Neighborhoods for London and Paris

Objective of this project is to help people to choose their destinations depending on the experiences that the neighborhoods have to offer and what they would want to have.


So that stakeholders and globetrotters can make informed decisions and address any concerns they have including the different kinds of cuisines, provision stores and what the city has to offer.



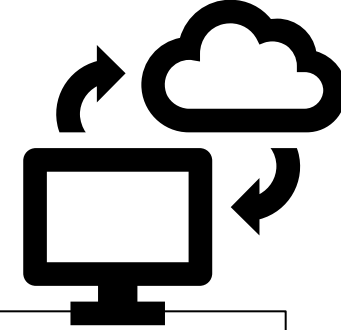
Target Audience

It will help people making smart and efficient decision on selecting great neighborhoods out number of other postal area in both the cites London and Paris.

People will get the awareness of area and neighborhood before visiting these big cities.




DATA ACQUISITION



- ✓ London : Data was obtained from web source by scrapping method (https://en.wikipedia.org/wiki/List_of_areas_of_London)
- ✓ Paris : Available data was obtained from web source by leverage JSON (<https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e>)
- ✓ Geographical location for London and Paris obtained by using
 - ArcGIS API
 - Foursquare API



DATA PREPARATION



Data Collection

collecting required data for both the cities London and Paris.

Data : Postal Codes
Neighborhoods
Borough



No
Duplicate
Values
found in the
dataset



Data Cleaning

Replacing spaces with
underscore in borough column
in London data
Break down the nested fields
and create dataframe that
needed



Feature Selection

For both the datasets,
we select Borough
Neighborhood
Postal codes
Geolocation



Feature Engineering

Process the data by
selecting only the
neighborhoods pertaining
to 'London' and 'Paris'



Data Geocoding

To get the geolocation
data we leverage
ArcGIS API and
Foursquare API



Data is ready for data analysis.
We populate the data into a
pandas dataframe.



EXPLORATORY DATA ANALYSIS

K – Mean Clustering:

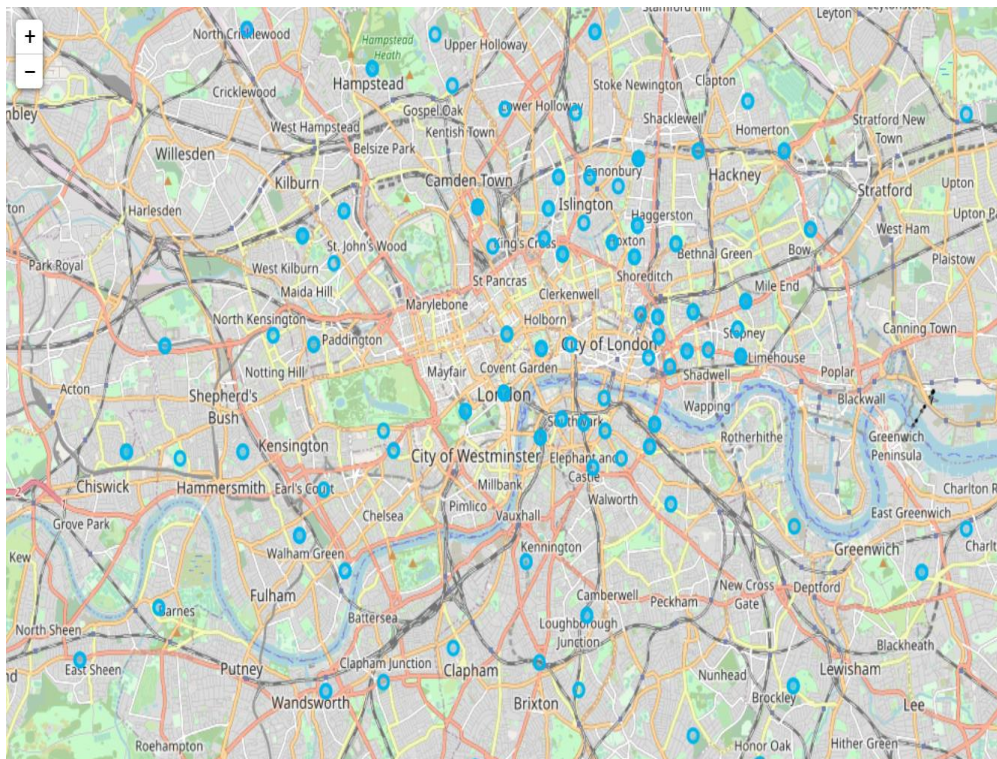
K-means clustering unsupervised Machine Learning algorithm is used to cluster the neighborhoods based on the category of venues near the neighborhoods for both cities London and Paris. We will be going with the number of clusters as 5.

WordCloud:

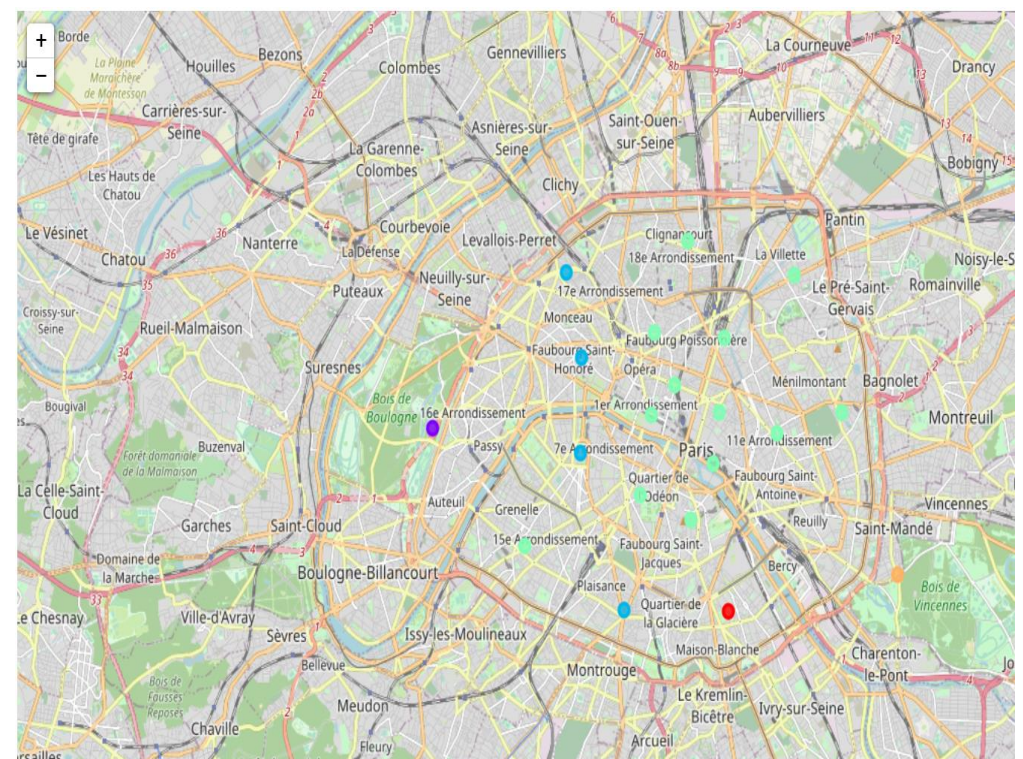
Word Cloud(Natural Language Processing) is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. We built a word cloud for each 1st Most Common Venue in London & Paris to discover dominant Venues.



Map of clustered neighborhoods of London:



Map of clustered neighborhoods of Paris:





Result and Discussion

The results from K – Mean Clustering show that we can categorize the neighborhoods into 5 clusters based on the frequency of occurrence.

Cluster Label	Clusters in London	Clusters in Paris
Cluster 1	Neighborhoods with a low number of frequencies (1 Record)	Neighborhoods with a low number of frequencies (1 Record)
Cluster 2	Neighborhoods with a high number of frequencies (294 Records)	Neighborhoods with a moderate number of frequencies (4 Records)
Cluster 3	Neighborhoods with a low number of frequencies (1 Record)	Neighborhoods with a high number of frequencies (13 Records)
Cluster 4	Neighborhoods with a low number of frequencies (2 Records)	Neighborhoods with a low number of frequencies (1 Record)
Cluster 5	Neighborhoods with a moderate number of frequencies (10 Records)	Neighborhoods with a low number of frequencies (1 Record)



Result and Discussion

The results from Wordcloud for each 1st Most Common Venue in London and Paris to discover dominant Venues and it gives:



1st Most Common Venue in London



1st Most Common Venue in Paris



Result & Discussion



By analyzing these five clusters obtained for cities London and Paris, we can see that some of the clusters are more suited for restaurants, café, plaza, art museums and hotels.

These clusters contain a higher degree of restaurants, hotels, multiplex, cafes, bars, other food joints and low degree other of venues like train station, bus station, fish market, gym, performing arts venue and smoke shop, to name a few.



Conclusion & Future Work

This project had performed the process of identifying the business problems, specifying the data required, extracting and preparing the data, visualizing the results, performing machine learning by clustering the data into 5 clusters based on their frequency similarities, tackling and reaching to a definitive solution to business problems for both the cities London and Paris.

Built useful model for the cities of London and Paris and see how attractive it is to potential tourists and migrants. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighborhoods finally concluding with clustering similar neighborhoods together. We could see that each of the neighborhoods in both the cities have a wide variety of experiences to offer which is unique.

Inclusively, it will help the stakeholders, immigrants and globetrotters may decide which city is preferable then according to their fondness and considering the factors determined in this project.



PERFECT
London
& *Paris*
ITINERARY



PERFECT
London
& *Paris*
ITINERARY



THANK YOU

Lavenya Mohanasundaram
May 2021

