



Data Science Capstone Project

Aditi Tambat

August 21, 2022



Abstract

This report discusses the use of an Api to build upon the comparison between two cities Paris and London and the correlation whether some one should start a business in either of the two cities. Various machine learning algorithms were used to compare and differentiate the two cities as well as other geolocation tools to check either of the city is good to start an Artificial Intelligence Business



1 - Introduction

The final course of the Data Science Professional Certificate consist of a capstone project where in all the skills and relevant knowledge that one has gathered from this 9 intense courses has to be applied on a final capstone project . The final problem as well as the analysis is the left for the reader to explore and decide. The idea uses location data with the help of the foursquare api that can be leveraged into coming up with a problem that the foursquare location data to solve it or just in contrast to compare cities or neighbourhoods of ones own choice



Target Audience

Potential Entrepreneur who want to start a business relating to Machine learning /AI/data science.

People who want to choose which city to live in the future


Course instructors and learners who will grade this project as well as showcase what I have learnt through this course.

Students who want learn and explore new things



2 - Business Problem

In this ever changing world of technology and reforms the use of AI will dominate and change most of the world and industries as we know so among the two busiest cities in the world which one would a person be willing to start a business in AI. Various factors would be included such as pricing, multiculturalism, language barriers and so on would influence this decision



	Place Name	State	County	City	Latitude	Longitude
0	Paris 01 Louvre	Île-de-France	Paris	Paris	48.8592	2.3417
1	Paris 02 Bourse	Île-de-France	Paris	Paris	48.8655	2.3426
2	Paris 03 Temple	Île-de-France	Paris	Paris	48.8637	2.3615
3	Paris 04 Hôtel-de-Ville	Île-de-France	Paris	Paris	48.8601	2.3507
4	Paris 05 Panthéon	Île-de-France	Paris	Paris	48.8448	2.3471

Figure 1: Paris Geolocation Dataset

- Data

Various data sets were collected, reformatted and analysed in order to get the required results. Some of them include



<http://www.cgedd.developpement-durable.gouv.fr/house-prices-in-france-property-price-index-french-a1117.html> - House Prices in France



<https://www.kaggle.com/alphaepsilon/housing-prices-dataset> - Housing Dataset



Artificial Intelligence

In regard to analysing which city would be best suited for a new AI startup a handful of datasets were extracted using web scarping tools.The final dataset was then merged and only the companies that were located in London and Paris were extracted. The code below shows how it was achieved

```
import requests
import pandas as pd

url = 'https://golden.com/list-of-artificial-intelligence-companies/'
html = requests.get(url).content
df_list = pd.read_html(html)
df = df_list[-1]
print(df)
df.to_csv('my data.csv')
```



4 - Methodology

An in-depth research of the dataset has been done and a thorough analysis of the various features and methods have been investigated to ensure the maximum accuracy of the model as possible. After reduction of the number of features in the data frame by replacing them with more useful data cluster analysis was done to find the best cluster of both Paris and London and then correlation and various other visual graphs were used to compare the two cities.

GeoLocation

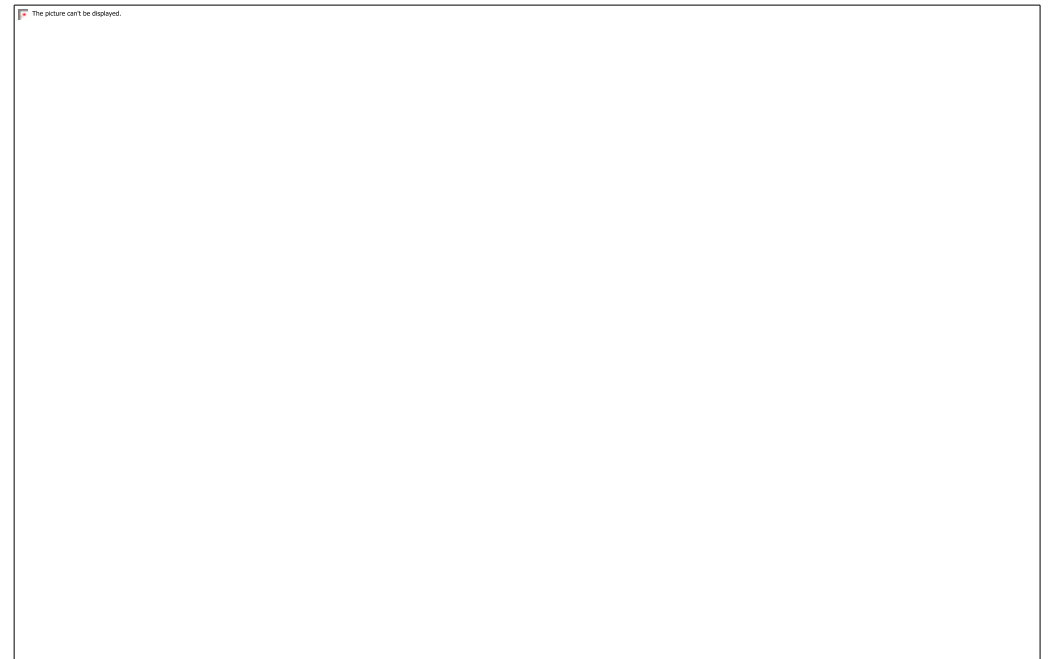
The algorithm below gets the required latitude and longitude of London(similar code has been coded for Paris) using the Google Maps Geocoder API.

```
address = "London, UK"

geolocator = Nominatim(user_agent="uk_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geographical coordinates of London are {},
      {}'.format(latitude, longitude))
```

Folium

Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. It uses the Open StreetMap technology. The code below shows only of Paris but a similar code has been coded even for London.



Python Code

create map of Paris using latitude and longitude values

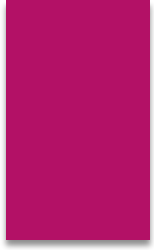
```
map_paris = folium.Map(location = [latitude, longitude], zoom_start = 11)
```

add markers to map

```
for lat, lng, county, name in zip(rparis['Latitude'],  
    rparis['Longitude'], rparis['County'], rparis['Place Name']):  
    label = '{} {}'.format(county, name)  
    label = folium.Popup(label, parse_html = True)  
    folium.CircleMarker(  
        [lat, lng],  
        radius = 5,  
        popup = label,  
        color = 'red',  
        fill = True,  
        fill_color = '#3186cc',  
        fill_opacity = 0.7,  
        parse_html = False).add_to(map_paris)
```

```
map_paris # show the map of paris with markers from the dataset
```

 The picture can't be displayed.



Foursquare API

The Foursquare API allows application developers to interact with the Foursquare platform. With the help of the Foursquare API venues and various other location and landmarks were extracted and merged into a dataframe.

```
LIMIT = 10 # limit of number of venues returned by Foursquare API

radius = 500 # define radius

url =
    'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius'
CLIENT_ID,
CLIENT_SECRET,
VERSION,
neighborhood_latitude,
neighborhood_longitude,
radius,
LIMIT)
results = requests.get(url).json()

# function that extracts the category of the venue
def get_category_type(row):
    try:
```

```
categories_list = row['categories']
    except:
categories_list = row['venue.categories']

    if len(categories_list) == 0:
        return None
    else:
        return categories_list[0]['name']

venues = results['response']['groups'][0]['items']

nearby_venues = json_normalize(venues) # flatten JSON

# filter columns
filtered_columns = ['venue.name', 'venue.categories',
                    'venue.location.lat', 'venue.location.lng']
nearby_venues = nearby_venues.loc[:, filtered_columns]

# filter the category for each row
nearby_venues['venue.categories'] =
    nearby_venues.apply(get_category_type, axis = 1)

# clean columns
nearby_venues.columns = [col.split(".")[-1] for col in
    nearby_venues.columns]

nearby_venues.head()
```

One Hot Encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

```
# one hot encoding
paris_onehot = pd.get_dummies(paris_venues[['Verne Category']], prefix =
    "", prefix_sep = "")

# add neighborhood column back to dataframe
paris_onehot['Neighborhood'] = paris_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [paris_onehot.columns[-1]] +
    list(paris_onehot.columns[:-1])
paris_onehot = paris_onehot[fixed_columns]

paris_onehot.head()
```

	Neighborhood	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
6	Bethnal	Turkish Restaurant	Asian Restaurant	Sushi Restaurant	Grocery Store	Indian Restaurant	Bakery	Out / Storage	Portuguese Restaurant	Coffee Shop	Gym / Fitness Center
5	Brent	Pub	Coffee Shop	Park	Pub	Indian Restaurant	Eastern European Restaurant	Supermarket	Fast Food	Japanese Restaurant	Out / Storage
2	Brenton	Pizza Place	Supermarket	Coffee Shop	Grocery Store	Pub	Deli/Meat Store	Italian Restaurant	Pub & Chess Shop	Pharmacy	Cafe
3	Camden	Japanese Restaurant	Pizza Place	Coffee Shop	Wine Bar	Italian Restaurant	Tea Restaurant	Italian Restaurant	Street	Hotel	Mexican Restaurant
4	City of London	Shopping Type	Hotel	Wine Place	Warehouse	Department Store	Pizza Place	Wine Store/Shop	Fast Food	French Restaurant	Botanical Garden

Figure 5: Top 10 Most visited Venues for London

Cluster Labels	Neighborhood	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
6	1	Paris 01 Louvre	Place	French Restaurant	Cocktail Bar	Church	Freemason Place	Chinese Restaurant	Park	Coffee Shop	Art Gallery
5	3	Paris 02 Marais	French Restaurant	Place	Bakery	Roman Restaurant	Restaurant	Swedish Shop	Fortune Shop	Brasserie	Farmers Market
2	2	Paris 03 Tuileries	Bakery/Place	Wine Bar	Park	Tea Room	Burger Joint	Restaurant	Cocktail Bar	Swedish Restaurant	Farmers Market
3	2	Paris 04 Hotel de Ville	Ice Cream Shop	Bakery Shop	Art Gallery	Art Museum	Cocktail Bar	Fountain	Street Shop	Lebanese Restaurant	Pub
4	3	Paris 05 Marais	Place	French Restaurant	Bar	Russian Restaurant	Museum/Landmark	Russian Museum	Ice Cream Shop	Bakery	Coffee

K- Means Clustering

After the venues were put into a dataframe, The K- Means Clustering Machine Learning Algorithm was used to train the data and get the desired clusters. The first task was finding the optimal K and as there were two different datasets to explore

- For Paris the optimal K found out to be was - 5
- For London the optimal K found out to be was - 6

After finding the Optimal K the data was trained using KMeans

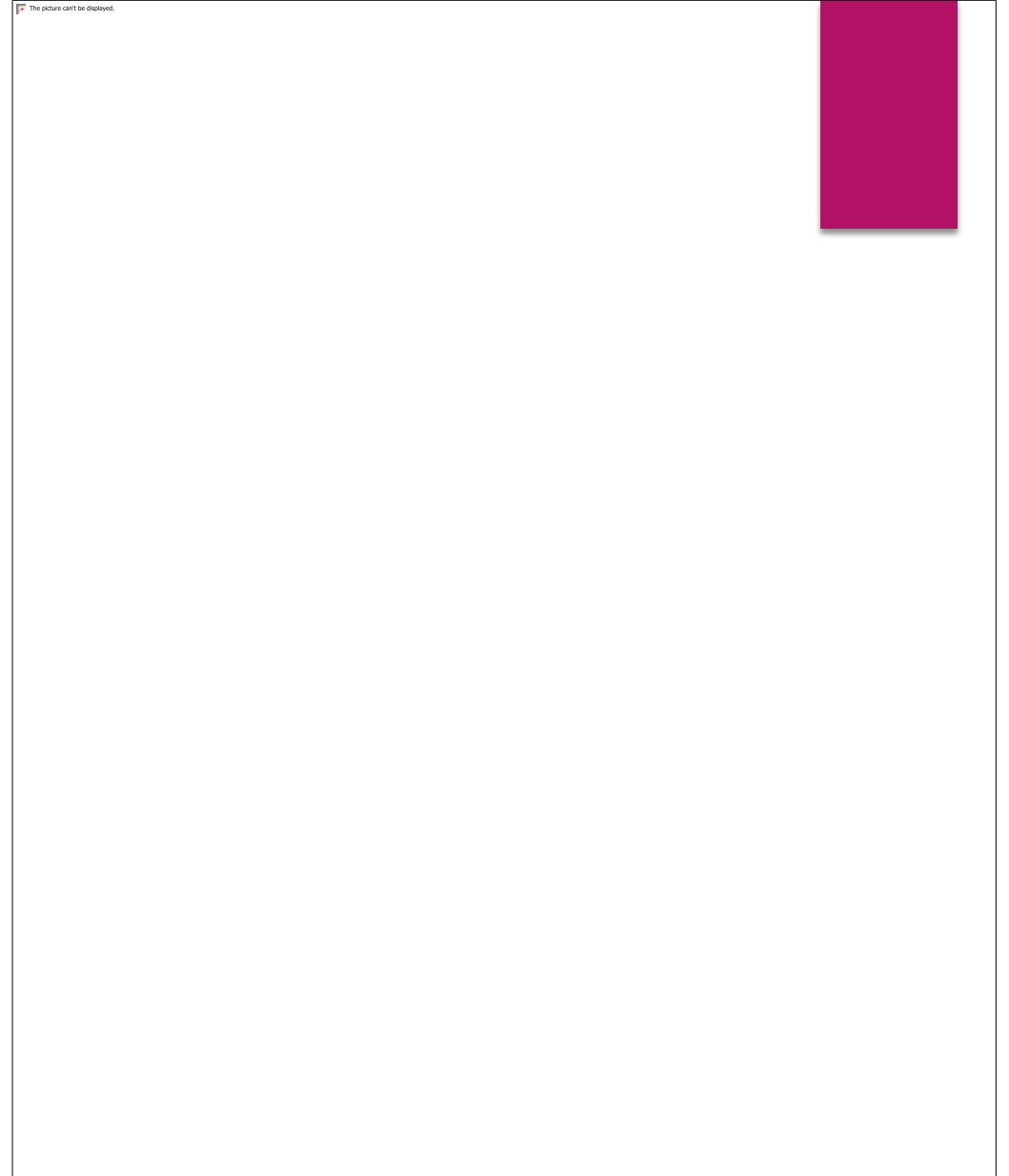
```
# set number of clusters
kclusters = int(len(rlondon["District"].unique()) / 4)
london_grouped_clustering = london_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters = kclusters, random_state =
                 1).fit(london_grouped_clustering)
```

Finally the data was then grouped into clusters as shown



Paris



London

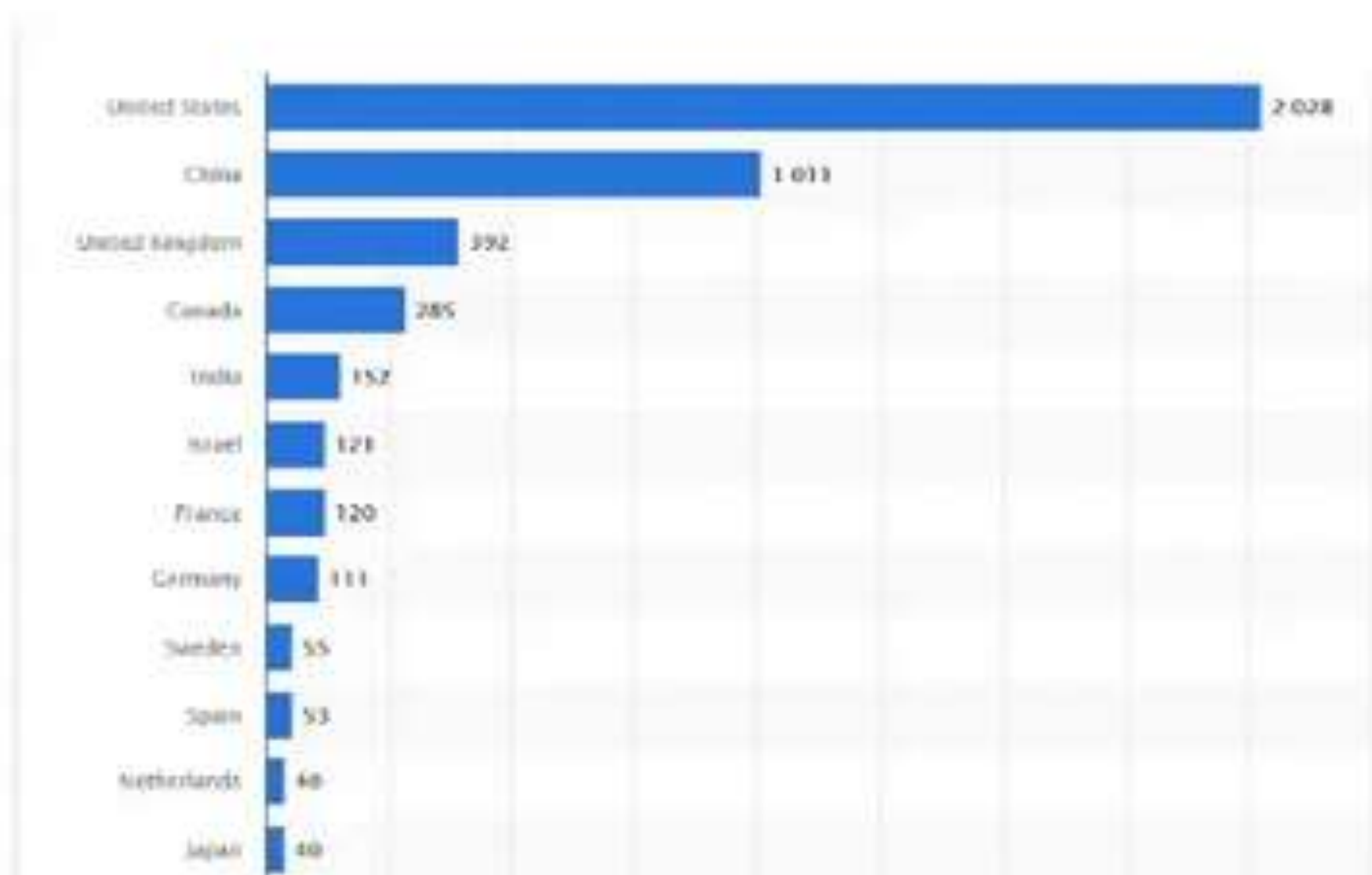


Figure 9: Countries with the maximum number of Artificial Intelligence Companies/Startups

5 - Results and Discussion

Finally we have reached the main part of our report. Let us break this down into two parts

Comparison of London and Paris

Similarities

- Both cities are multicultural and diverse in their own ways and share a rich history of their own.
- Most of the famous neighbourhoods have a restaurant as its top most Common Venue.
- Example : In Paris the Louvre is one of the most famous icons if not in the world also and its most common venue is the Plaza/French Restaurant.

- Similarly for the famous icon in London that is Westminster are Pubs and restaurants.
- The top 3 most common Venue points for London are

Coffee Shop

Hotel

Cafe

- The top 3 most common Venue points for Paris are

Coffee Shop

Pub

Cafe

- Both have an almost comparable population size of about 8-10 million.
- Both have an overwhelming power of attraction.
- When comparing the prices of the venues both of them are expensive in their own ways and offer high quality food ,concerts, exhibition etc



Differences

- While looking at the maps one can observe that Paris is more compact and one can walk around much more freely without the use of transport
- London on the other hand requires the use of transport as its much larger on the scale.
- In terms of population density Paris definitely outweighs London by a ratio of 4:1.
- By a recent comparison and taking a look of the most visited venues Paris definitely has a higher number of restaurants of a ratio of almost 3:1 and according to studies restaurants in Paris have earned higher Michelin Stars than London's.
- In terms of Leisure and entertainment London definitely has more spots than Paris. A simple example would be that London has more museums than Paris in a ratio of 8:5.
- Paris definitely hosts three of the top 10 most visited at-traction sites while London has none.
- London definitely has more people from abroad.
- London has a lower temperature than Paris on average

Discussion

There are major challenges while constructing a dataset ie:



The dataset for the Artificial Intelligence wasn't readily available and so had to be scrapped from multiple sources which often leads to inconsistency happening as well as errors.



Only a random sample of 0.05 percent was taken into consideration. A good and optimal model would take a testing data and a training data and would train it on the complete dataset multiple times.



The data obtained through the API calls would return different results each time its called. Multiple trials and error runs are required to get the desired result.



The districts have too complex geometry which would bring an error in our analysis if the venues are too close to each other. This is one of the reason why Pipelines are required .However nodoubt that if this process was to be repeated multiple times the de-sired outcome would have generated and a better comparison could have been made.



6 - Conclusion

After an in depth review of the comparison between London and Paris and which city would be a better place to start an Artificial Intelligence Company or invest multiple conclusions can be drawn. One of them being that both cities are diverse in their own ways and boast a culture unlike no other. Artificial Intelligence is a booming topic and recently more people have started investing into it as well as companies automating their processes. Both cities offer a wide range of opportunities for anyone starting to invest in Artificial Intelligence or even start a company and various factors were shown. Finally a better model could be made by various other methods and much stronger Machine Learning Algorithms like KD Tree which have a much faster run time algorithm of

$O(N \log(N))$ vs $KNN O(N^2)$

.Furthermore, clustering however did help us to highlight the most optimal venues and areas. Finally correlation does not imply causation and so any result here is subject to change on various other trends and opinions and datasets



7 - Acknowledgements

I sincerely thank all the course instructors who have taken their time and effort into making this Professional Certificate worth the effort .I also want to state that these are my opinions and are subject to change as well as I am grateful for all resources and knowledge that I have learnt throughout this course. Thank you to all the peer reviewers that have graded my projects