Capstone Project

Robert Joseph

July 18, 2020

Abstract

This report discusses the use of Foursquare Api to build upon the comparison between Paris and London and the correlation whether someone should start a buisness 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.

London is the capital and largest city of England and the United Kingdom.Standing on the River Thames in the south-east of England, at the head of its 50-mile (80 km) estuary leading to the North Sea, London has been a major settlement for two millennia.

Paris is the capital and most populous city of France, with an estimated population of 2,150,271 residents as of 2020, in an area of 105 square kilometres (41 square miles).Since the 17th century, Paris has been one of Europe's major centres of finance, diplomacy, commerce, fashion, science and arts.

The main Goal of this project that I have chosen would be to evaluate the comparison between Paris and London as well as point out the differences. Another factor to be included is which city would be more ideal to start an Artificial Intelligence company and the various factors correlating to it as both cities are major cities and global hotspots in the world for tech companies.

Target Audience

- Potential Entreprenuers who want to start a business relating to Machine learning /AI/data science.
- People who wanna 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.

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, multiculturism, 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

3 - 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
- https://data.world/datasets/real-estate - Numerous Datasets for different categories
- https://data.london.gov.uk/dataset?tag=start-ups - Data sets for london
- https://www.kaggle.com/tags/companies - Various companies and their datasets

More datasets were included and merged to get the final dataset relating to the idea.The use of even foursquare datasets were used and important features such as Housing Prices, Locality, Famous icons, Restaurant Prices, transportation facilities, technological hotspots as well as access to a high wifi speed and so on were all assessed.

London

Almost half a million lines of records were present for the London dataset and had to sampled such that only 160 rows were extracted

	Postcode	Country	County	District	Latitude	Longitude
0	BR1 1AA	England	Greater London	Bromley	51.401546	0.015415
1	BR1 1AB	England	Greater London	Bromley	51.406333	0.015208
2	BR1 1AD	England	Greater London	Bromley	51.400057	0.016715
3	BR1 1AE	England	Greater London	Bromley	51.404543	0.014195
4	BR1 1AF	England	Greater London	Bromley	51.401392	0.014948

Figure 2: London Geolocation Dataset

as it had too many records to handle.
This was achievable with the help of

```
# Python Code
rlondon = london.sample(frac = 0.0005) # only 0.0005 % of the data was
    randomly selected
rlondon.head()
```

Various columns which served no purpose to the analysis were dropped such as 'Police Station Code', Lower layer super output area', 'Rural/urban','Region', 'Altitude', 'London zone' etc.
Finally the resulting dataset is shown in Figure 2

Paris

Similar to the London Dataset there were about 30,000 rows out of which only a sample was taken by using the same python code as state above.
The paris dataset had no columns which needed to be dropped and so was retained in its original state.

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 scarpping 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')
```

These are some of the links

- http://analytics.dkv.global/data/pdf/AI-in-UK/AI-in-UK-1000-UK-AI-Companies-Profiles.pdf
- https://craft.co/artificial-intelligence-companies-in-paris-fr?page=1
- https://clutch.co/fr/developers/artificial-intelligence

Finally to compare them various visualisation tools were used and articles referenced in order to reach the final conclusion.

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.

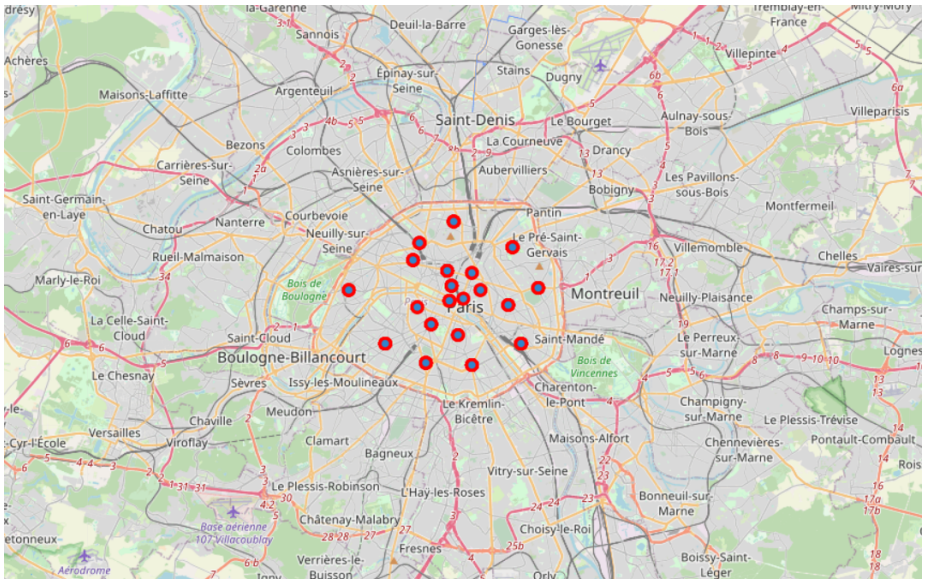

Figure 3: Map of Paris with Markers

```
# 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
```

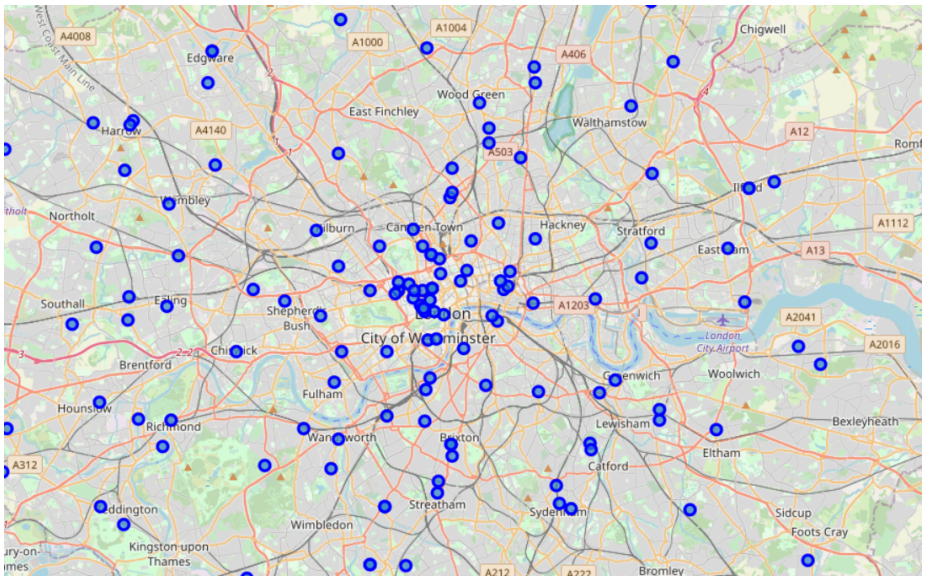

Figure 4: Map of London with Markers

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[['Venue 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
0	Barnet	Turkish Restaurant	Italian Restaurant	Sushi Restaurant	Grocery Store	Indian Restaurant	Bakery	Deli / Bodega	Portuguese Restaurant	Coffee Shop	Gym / Fitness Center
1	Brent	Pub	Coffee Shop	Park	Platform	Indian Restaurant	Eastern European Restaurant	Supermarket	Food Truck	Japanese Restaurant	Deli / Bodega
2	Bromley	Pizza Place	Supermarket	Coffee Shop	Grocery Store	Pub	Stationery Store	Indian Restaurant	Fish & Chips Shop	Pharmacy	Café
3	Camden	Japanese Restaurant	Pizza Place	Coffee Shop	Beer Bar	Italian Restaurant	Tapas Restaurant	Malay Restaurant	Market	Hotel	Mexican Restaurant
4	City of London	Boxing Gym	Hotel	Burrito Place	Steakhouse	Department Store	Pizza Place	Indie Movie Theater	Event Space	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
0	3	Paris 01 Louvre	Plaza	French Restaurant	Cocktail Bar	Church	Pedestrian Plaza	Chinese Restaurant	Park	Coffee Shop	Art Gallery	Garden
1	0	Paris 02 Bourse	French Restaurant	Plaza	Bakery	Ramen Restaurant	Restaurant	Souvlaki Shop	Perfume Shop	Bookstore	Farmers Market	Coffee Shop
2	2	Paris 03 Temple	Sandwich Place	Wine Bar	Park	Tea Room	Burger Joint	Restaurant	Cocktail Bar	Seafood Restaurant	Farmers Market	Wine Shop
3	2	Paris 04 Hôtel-de-Ville	Ice Cream Shop	Souvenir Shop	Art Gallery	Art Museum	Cocktail Bar	Fountain	Gourmet Shop	Lebanese Restaurant	Pub	Alsatian Restaurant
4	3	Paris 05 Panthéon	Plaza	French Restaurant	Bar	Korean Restaurant	Monument / Landmark	Science Museum	Ice Cream Shop	Bakery	Creperie	Grocery Store

Figure 6: Top 10 Most visited Venues for Paris

Most Visited Venues

The top 10 most visited venues were extracted from each neighbourhood and then merged together to form another dataset.

```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending = False)
    return row_categories_sorted.index.values[0:num_top_venues]

num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1,
            indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
```

```
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] =
    paris_grouped['Neighborhood']

for ind in np.arange(paris_grouped.shape[0]):
   neighborhoods_venues_sorted.iloc[ind, 1:] =
       return_most_common_venues(paris_grouped.iloc[ind, :],
       num_top_venues)

neighborhoods_venues_sorted.head()
```

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

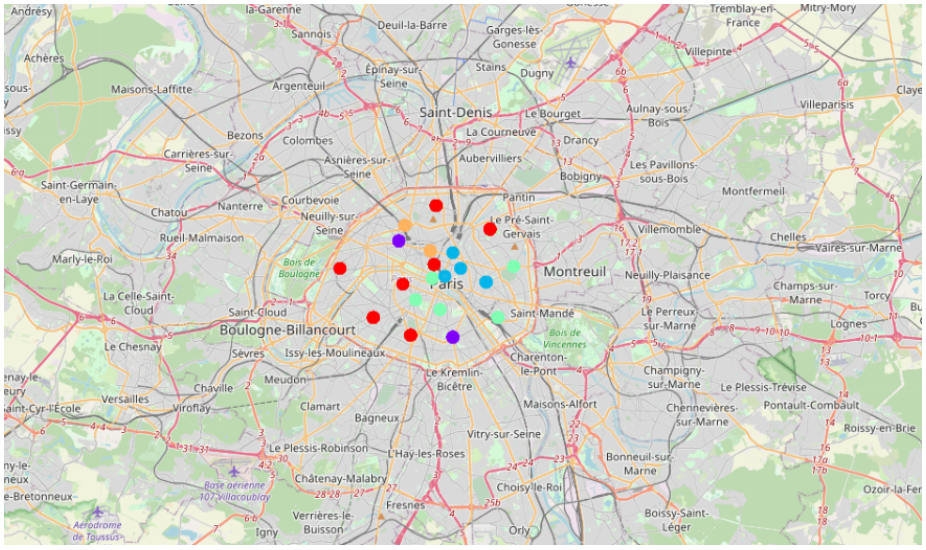

Figure 7: Map of Paris Clustered

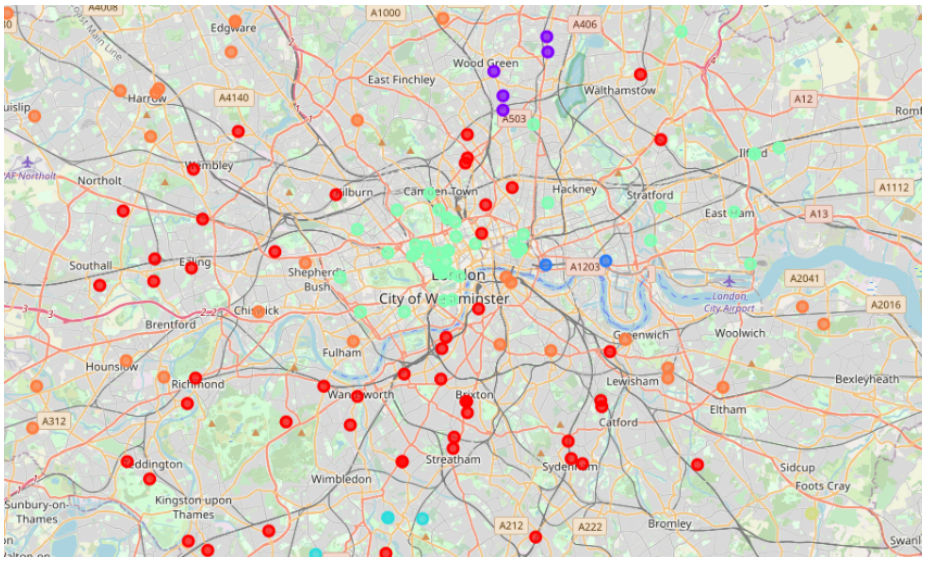

Figure 8: Map of London Clustered

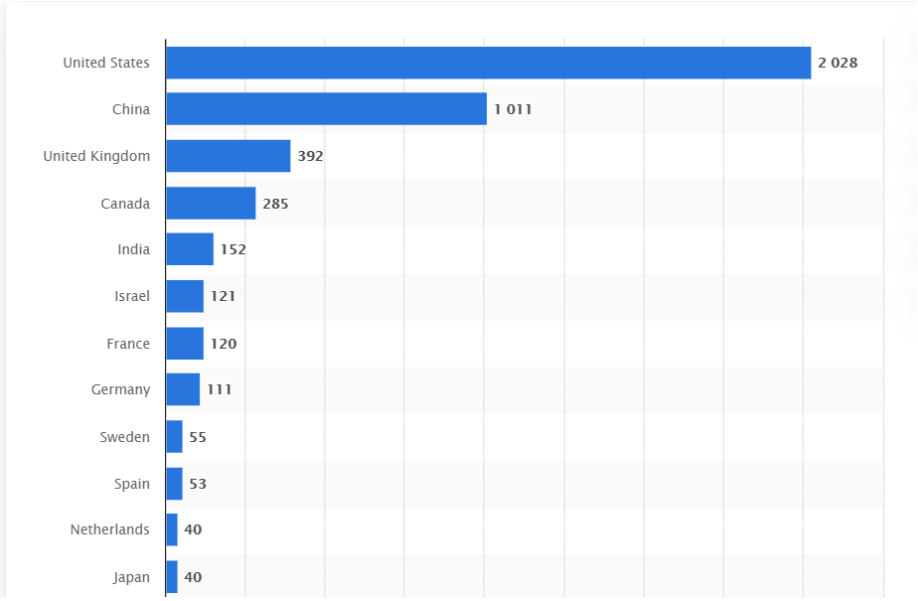

Figure 9: Countries with the maximum number of Artificial Intelligence Companies/Startups

Artificial Intelligence

For the artificial intelligence dataset after cleaning and webscrapping the number of startups were plotted and number of AI companies in each country was counted. Our main focus however is on the United Kingdom specifically London and France specifically Paris.

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 commont venue is the Plazza/French Restaurant.
  - Similarly for the famous icon in London that is Westminister 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 attraction sites while London has none.
  - London definitely has more people from abroad.
  - London has a lower temperature than Paris on average.

Artificial Intelligence

Now while comparing which city would be better to start an Artificial Intelligence company London definitely has more AI companies.
London has almost 758 AI companies out of which 645 are headquatered in the capital.
According to the Mayor of London Sadiq Khan " "There are few areas of innovation that have the power to define our future economy and society more than artificial intelligence".
According to a recent study london has 868 Number of AI Jobs per 1 Million in the city compared to Paris 660.
One of the major regions this is true is also because of the Worldclass Universities which are present in London such as Cambrdige/Kings/Imperial and so on. London's status as a global financial

Tech hub							
	2013	**2014**	**2015**	**2016**	**2017**	**2018**	**Total**
San Francisco	£418.08m	£1.83bn	£2.07bn	£4.46bn	£806.03m	£1.84bn	**£11.44bn**
Beijing	£11.66m	£53.75m	£197.32m	£599.56m	£1.63bn	£1.07bn	**£3.57bn**
New York	£79.43m	£165.62m	£318.28m	£667.51m	£593.85m	£1.2bn	**£3.05bn**
Shanghai	–	£1.28m	£400.93m	£16.10m	£1.6bn	£453.61	**£2.47bn**
London	£9.85m	£41.16m	£67.04m	£166.04m	£228.97m	£326.90m	**£839.96m**
Paris	£1.92m	£2.83m	£23.49m	£61.49m	£99.45m	£132.40m	**£321.48m**
Singapore	£13.76m	£13.89m	£70.92m	£55.59m	£106.52m	£30.81m	**£291.49m**
Tel Aviv	£14.80m	£17.12m	£5.49m	£39.04m	£112.25m	£89.01m	**£277.71m**
Berlin	£7.09m	£0.79m	£23.60m	£17.41m	£17.67m	£21.06m	**£87.62m**
Bangalore	£1.31m	£32.29m	£45.75m	£1.96m	£36.71m	£18.65m	**£136.67m**

Figure 10: Table showing venture capital funding into AI companies across major global cities from Jan 2013- August 2018

services centre and leader in the development of financial technologies has helped the city's AI finance ecosystem to flourish While in the figure the dataset hasnt been completed , similar trends have been found in regard to London being one of the top most hub spots for Artificial Intelligence.

"We would like France to be one of the leaders of AI but we want Europe to be the champion as well, we have the means and now we have to create the conditions that will enable us to get there " - Emmanuel Macron

Paris however does not fall behind in its capabilities of being an AI hotspot. Companies such as Facebook, Google, Samsung all intend to open a Paris AI center and help Paris in its global push into the AI market.

The goverment has invested almost 1.8Billion Euros for Artificial Intelligence and promotes that people come and invest in Paris.

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 no doubt that if this process was to be repeated multiple times the desired outcome would have generated and a better comparison could have been made.

6 - Conclusion

After an indepth 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 $O(Nlog(N))$ 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 Proffesional 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. I also want to thank God, my Family as well as friends who have made this a reality and for supporting me throughout.Thank you to all the peer reviewers that have graded my projects.