# **Introduction / Business Problem**

Paris, France and Milan, Italy are the two well know and renowned destinations of Europe known as shopping and Fashion paradise. Both the city houses most world-famous apparel and accessories brand in the heart of the city. There are different studies available in literature regarding the effectiveness of both the cities as shopping destination but some of them may be lacking in data driven exercise during the decision-making process.

As part of this current study my intention is to help a small boutique start-up firm Smart Fashion inc. to take a holistic decision of opening a shop in either Paris or Milan city neighbourhood area. It is assumed that Smart Fashion inc. entirely focuses on women’s apparel and cosmetic products. While making the data driven decision, both the city is compared with respect to populations leaving in the city limits of two cities and GDP per capita. While considering GDP per capita data, the data for France and Italy are compared and it is considered that Paris and Milan GDP per capital is in-line with France and Italy’s GDP per capita value.

The objective of this capstone project is described as follows:

* Do a comparison study between Paris and Milan and chose one city for consideration.
* It is evident that city have more population within the city limit and the countries having more GDP per capita may have more spending power towards fashion or luxury items. In order to take the shop opening decision, both city’s population data and corresponding countries GDP (Gross Domestic Product) per capita data will be extracted from external data sources.
* Ger the neighbourhood data of the selected city and plot visually.
* Compare the neighbourhood details of the selected city for analysis.
* Combining all the data mentioned above a concise decision will be suggested to Smart Fashion Inc.
* Carry our clustering process to group the neighbourhood into 3 clusters and show visually.

# **Data Source & Collection**

The data used for this study are secondary data sources related to Paris and Milan’s location and neighbourhood data, data related to the two countries GDP per capita and city’s population data. During the process three different data sources are being considered and they are explained below:

* **Foursquare**: This data source will be access to get the shopping venue for the neighbourhood location of Paris, France. Foursquare provides REST (Restful Webservice) based API (Application Programming Interfaces) to get the data in JSON format. My own user account with Foursquare will be used for providing access credentials to get the data.
* **Wikipedia Page for City’s Population Data**: The Wikipedia page (<https://en.wikipedia.org/wiki/List_of_cities_in_the_European_Union_by_population_within_city_limits> ) with European city’s population data will be used to get an idea of population living within the city limits of Paris and Milan. Pandas library will be used for web scrapping of the page to het the tabular data during analysis.
* **Wikipedia Page for Country’s GDP per Capita:** The Wikipedia page (<https://en.wikipedia.org/wiki/Economy_of_the_European_Union> ) will be web scrapped to fetch the GDP per capita data of France and Italy. Pandas library will be used for web scraping the page and retrieving tabular data.
* **Neighbourhood Data of Paris**: The following API URL (<https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e> ) returns the neighbourhood data of Paris.

Both the Wikipedia pages are publicly available and accessible and there are no licencing agreements in place for accessing those data.

# **Methodology**

The methodology used in the current analysis are as follows:

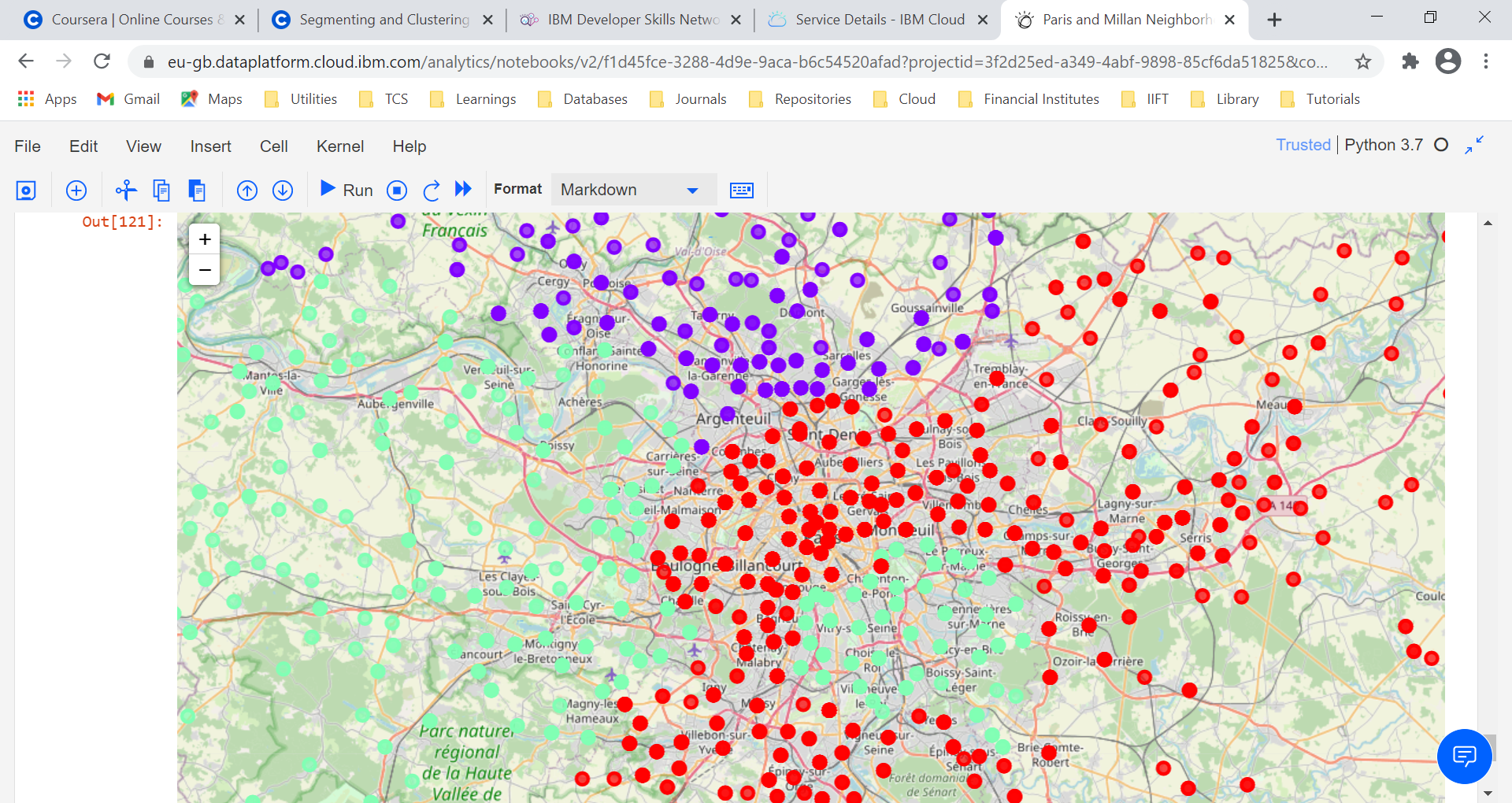
* GDP per capita data for France and Italy are compared
* Population within City limit data of Paris and Milan are considered.
* Python 3.7 is used
* REST based API endpoint is executed to get Paris neighbourhood data
* For data visualization purpose, folium library of Python is used.
* Python’s geopy library is used for retrieving any geographical location data
* Foursquare API is used to get different venues of the neighbouring locations
* Descriptive statistics (mean value) of venues corresponding to unique neighbouring locations are calculated.
* K-means clustering is carried out to divide the neighbouring locations into 3 different clusters.

# **Results**

The result of the analysis clearly shows the followings:

* Compared to Milan, Italy we can consider Paris, France as more desirable location for opening new show for Smart Fashion Inc.
* Different neighbourhood locations of Paris vary while considering Women’s store and cosmetic store locations.
* From the results it is evident that VAL-DE-MARNE has highest value for the Cosmetic Shops and this location as good value (second best) for Women's store. Though ESSONNE has highest value for Women's store but VAL-DE-MARNE is close to ESSONNE.
* If we compare cosmetic shops then VAL-DE-MARNE is way above ESSONNE.
* The K-means cluster of the neighbourhood effectively segmented the neighbourhoods into three distinct clusters as shown below.

The results obtained through cluster analysis is plotted visually as follows:



# **Discussions**

Based on the results we have received, Smart Fashion Inc. is advised to open their shop in the neighbourhood of Paris, France.

The suggested neighbourhood location for open store in VAL-DE-MARNE.

# **Conclusions**

During preparation of the report, I have used different tools and technologies like Python 3.7, Foursquare API, K-means clustering algorithm and descriptive statistics in order to suggest store opening location for Smart Fashion Inc.

While making the report, I have assumed that investment or cost of opening a shop is more or less same in the neighbouring locations of Paris which may not be true in the idea scenario.

Neighbourhood specific cost aspects may also be considered to make the prediction even better and which can be considered as way forward item.

Thank you very much.