# 2 The Data Science Workflow

**Data Requirements**

The main districts in Paris are divided into 20 Arrondissements Municipaux (administrative districts), shortened to arrondissements.

The data regarding the districts in Paris needs to be researched and a suitable useable source identified. If it is found but is not in a useable form, data wrangling and cleaning will have to be performed.

The cleansed data will then be used alongside Foursquare data, which is readily available. Foursquare location data will be leveraged to explore or compare districts around Paris, identifying the high traffic areas where consumers go for shopping, dining and entertainment - the areas where the fashion brand are most interested in opening new stores.

**The Data Science Workflow for Part 1 & 2 includes the following:**

* **Outline the initial data that is required:**
  + District data for Paris including names, location data if available, and any other details required.
* **Obtain the Data:**
  + Research and find suitable sources for the district data for Paris.
  + Access and explore the data to determine if it can be manipulated for our purposes.
* **Initial Data Wrangling and Cleaning:**
  + Clean the data and convert to a useable form as a dataframe.

**The Data Science Workflow for parts 3 & 4 includes:**

* **Data Analysis and Location Data:**
  + Foursquare location data will be leveraged to explore or compare districts around Paris.
  + Data manipulation and analysis to derive subsets of the initial data.
  + Identifying the high traffic areas using data visualisation and tatistical nalysis.
* **Visualization:**
  + Analysis and plotting visualizations.
  + Data visualization using various mapping libraries.
* **Discussion and Conclusions:**
  + Recomendations and results based on the data analysis.
  + Discussion of any limitations and how the results can be used, and any conclusions that can be drawn.

**Import the Required Libraries**

*# Import libraries*

**import** **numpy** **as** **np** *# library to handle data in a vectorized manner*

**import** **json** *# library to handle JSON files*

**import** **pandas** **as** **pd**

!conda install -c conda-forge geopy --yes

**from** **geopy.geocoders** **import** Nominatim *# convert an address into latitude and longitude values*

**import** **requests** *# library to handle requests*

**from** **pandas.io.json** **import** json\_normalize *# tranform JSON file into a pandas dataframe*

*# Matplotlib and associated plotting modules*

**import** **matplotlib.pyplot** **as** **plt**

**import** **matplotlib.cm** **as** **cm**

**import** **matplotlib.colors** **as** **colors**

**from** **bs4** **import** BeautifulSoup

*# Import k-means from clustering stage*

**from** **sklearn.cluster** **import** KMeans

!conda install -c conda-forge folium=0.5.0 --yes

**import** **folium** *# map rendering library*

print('Libraries imported.')

# Data Research and Preparation

## Import the Paris District Data

#### Arrondissements Municipaux for Paris CSV (administrative districts)

Paris is divided into 20 Arrondissements Municipaux (or administrative districts), shortened to just arrondissements. They and normally referenced by the arrondissement number rather than a name.

Data for the arrondissements is necessary to select the most suitable of these areas for new stores.

Initially looking to get this data by scraping the relevent Wikipedia page (<https://en.wikipedia.org/wiki/Arrondissements_of_Paris>), fortunately, after much research, this data is available on the web and can be manipulated and cleansed to provide a meaningful dataset to use.

Data from Open|DATA France: <https://opendata.paris.fr/explore/dataset/arrondissements/table/?dataChart>

Also available from Opendatasoft: <https://data.opendatasoft.com/explore/dataset/arrondissements%40parisdata/export/>

*# Download the dataset and read it into a pandas dataframe.*

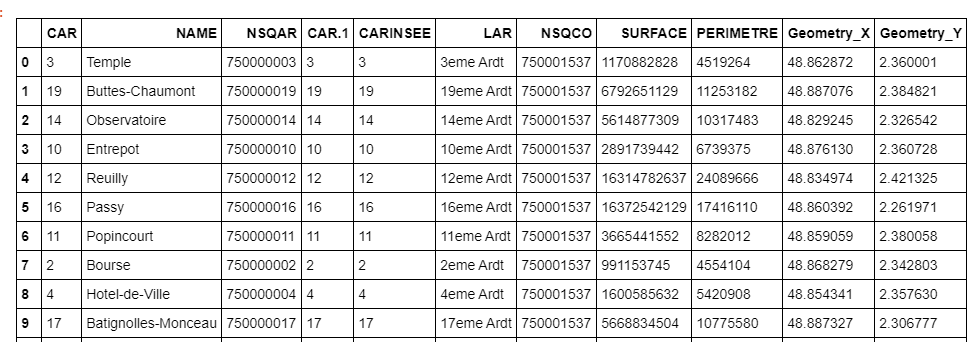
*# The Arrondissements dataset was downloaded from Paris|DATA: https://opendata.paris.fr/explore/dataset/arrondissements/table/?dataChart*

*# Then placed on the GitHub repo for the project.*

*# https://raw.githubusercontent.com/AR-data-science/Coursera\_Capstone/master/Arrondissements\_.csv*

paris = pd.read\_csv('https://raw.githubusercontent.com/AR-data-science/Coursera\_Capstone/master/Arrondissements\_.csv')

paris



## Exploring, Wrangling and Cleaning the Data

### Rename the columns 'Geometry\_X' and 'Geometry\_Y', "CAR' to 'Arrondissement' etc...

*# Rename the necessary columns 'Geometry\_X and Geometry\_Y' etc...*

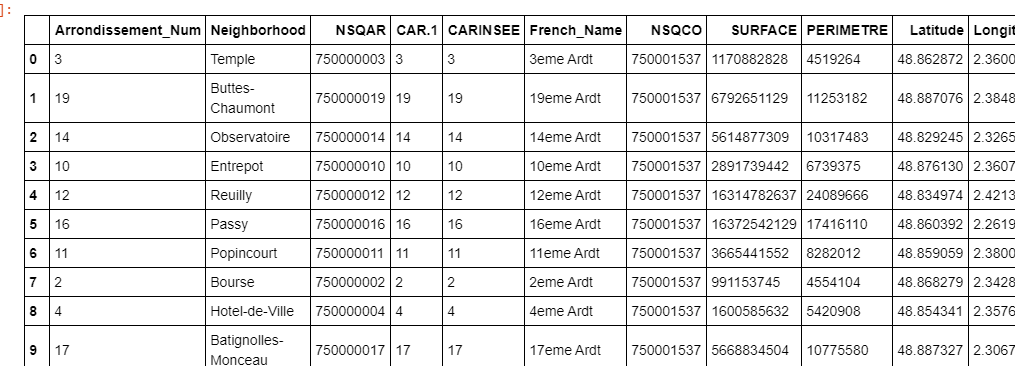
*# District : name of the central District for the Arrondissement*

*# Arrondissement : the Arrondissement or district number which is used to identify it*

*# Arrondissement\_Fr : the descriptive French label for each Arrondissement*

paris.rename(columns={'NAME': 'Neighborhood ', 'CAR': 'Arrondissement\_Num', 'Geometry\_X': 'Latitude', 'Geometry\_Y': 'Longitude', 'LAR': 'French\_Name'}, inplace=**True**)

paris



### **Clean up the dataset to remove unnecessary columns**[**¶**](https://jp-tok.dataplatform.cloud.ibm.com/data/notebooks/converter/assets/8a834e45-b7cd-41bb-97d2-a2a747fd0d2c?access_token=6bf2f41d17fb979d3b9d9ee28bab69541d04afbfd14e6e7d088b2c8f96106a7f&project=cae6b3ac-8685-44d5-b31f-45b4ed8e8214#Clean-up-the-dataset-to-remove-unnecessary-columns)

*# Clean up the dataset to remove unnecessary columns.*

*# Some of the columns are for mapping software - not required here.*

paris.drop(['NSQAR','CAR.1','CARINSEE','NSQCO','SURFACE', 'PERIMETRE' ], axis=1, inplace=**True**)

paris

