A description of the problem and a discussion of the background (15marks)

> ***"****Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data to solve or execute. Remember that data science problems always target an audience and are meant to help a Group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem."*

A major coffee shop chain ("Company"), which is currently operating in Athens has been partially acquired (80% of the shares) from a Canadian Group ("Group"). The coffee shop chain has more than 100 shops operating successfully in Greece. According to information provided by the Company, key factor for the success of its retail shops, is the "ideal" location selection of its shops.

One of its most profitable coffee shop (highest sales, gross profit and net profit margins) is operating in the Neighbourhood of Kolonaki in the centre of Athens.

The Group is considering starting to expand its businesses abroad. In specific the management of the Group is considering opening coffee shops in USA & Canada. For that purpose, they need to commit and "risk" substantial capital expenditures and opex.

Before starting to commit any funds for expanding its businesses abroad, the management of the Group has decided to start with a "pilot project" (opening only one shop in one country).

The main purpose of the pilot project is to figure out if the successful "Greek coffee shop model" could be replicated in Canada and the USA, without any major adjustments on the marketing "mix" & policy or alternatively if the existing business & marketing model must be adjusted substantially in order to penetrate the foreign markets.

As a first step the Group has decided that it will be more efficient to start with Canada/Toronto since the Group is based in Canada and it has sufficient local know how & it would be more cost effective to start the "pilot project" in Canada instead of USA .

The second Step is to identify the characteristics/venue categories of the Greek Neighbourhood , where the Company's existing most successful shop is located, i.e Kolonaki, Greece.

The next step is to identify one or several Boroughs/Neighbourhoods of Toronto Centre, with similar venue categories & characteristics to Kolonaki Athens. By identifying the common venues of Kolonaki-Athens and Toronto Centre, the Company will be able to open a shop in Toronto, which has similar venues categories to Athens Centre.

For preparing the above the Company will use the services of a Data scientist.

After the Data scientist suggest to the Company the relevant Boroughs/Neighbourhoods of Toronto Centre, which are similar to Kolonaki Athens, the Company will use its own additional "selection criteria ", which are currently not disclosed to the Data scientist, in order to select the final Borough/Neighbourhood , where the new coffee shop will operate.

Identifying the Neighbourhoods of Toronto Central which share similar characteristics to Athens , is critical for the following indicative reasons :

1. Company success is strongly correlated with proper "location selection"
2. In case of failure of the "new " Toronto coffee shop to achieve sales target, such failure will not primarily be attributed to "wrong" location selection, since both Athens shop and Canadian shop will have similar venue categories. Indicatively the Company may attribute a potential failure to "price policy" of the new shop or "low acceptance" of the Toronto retail clients for the specific coffee taste or ineffective business promotion, or lack of brand awareness etc.

Depending on the success or not of the new pilot shop the Company will decide if expansion with the existing business model is making economic sense or if further adjustments of the business model are necessary.

For executing the above steps the Company is asking for the professional services of a Data Scientist

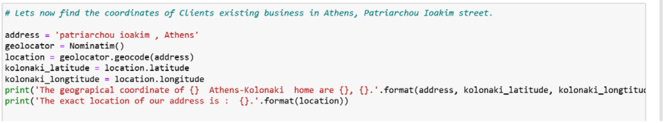
The data scientist has decided to utilize Foursquare in order to explore the venue categories near the Athens shop. Then the Data scientist will use public available information from several sources (see next section) in order to find the Boroughs and Neighbourhoods in Toronto Center.
In addition the Data scientist will utilize again Foursquare in order to find venue categories for each Borough and Neighbourhood in Toronto centre.
Finally he/she will create a cluster (using KMeans from sklearn cluster) of Boroughs and Neighbourhoods in Toronto, which share similar venue categories to Athens Greece and he/she will deliver the findings of his/her work to the management of the Company.

A description of the data and how it will be used to solve the problem. (15marks)

> *“Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.”*

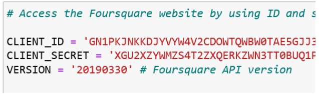
For the specific project assigned by the Group we need the following data:

- Coordinates of the existing location in Kolonaki Athens and Coordinates of Toronto Centre. In that respect we will use from “geopy.geocoders” , “Nominatim”, which converts an address into latitude and longitude values :



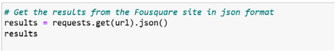
```
# Lets now find the coordinates of Clients existing business in Athens, Patriarchou Ioakim street.

address = 'patriarchou ioakim , Athens'
geolocator = Nominatim()
location = geolocator.geocode(address)
kolonaki_latitude = location.latitude
kolonaki_longtitude = location.longitude
print('The geograpical coordinate of {}  Athens-Kolonaki  home are {}, {}.'.format(address, kolonaki_latitude, kolonaki_longtitu
print('The exact location of our address is :  {}.'.format(location))
```

- Venues for Athens and Toronto city near our targeted Borough/neighbourhood. In that respect we will utilize “Foursquare”:



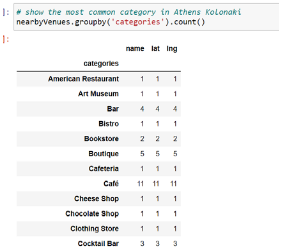
```
# Access the Foursquare website by using ID and s

CLIENT_ID = 'GN1PKJNKKDJYVYW4V2CDOWTQWBW0TAE5GJJ3
CLIENT_SECRET = 'XGU2XZYWMZS4T2ZXQERKZWN3TT0BUQ1F
VERSION = '20190330' # Foursquare API version
```

- We will need to import “requests” which is a library to handle requests in “json” format in order to get from Foursquare the requested venues info :



```
# Get the results from the Fousquare site in json format
results = requests.get(url).json()
results
```

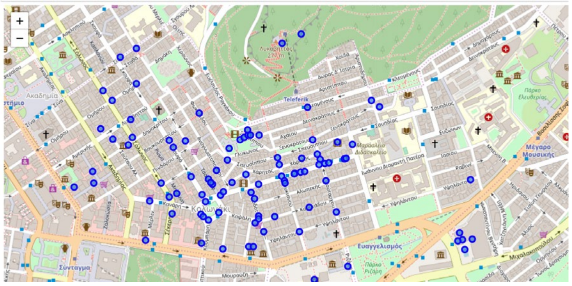
- We will then import from “pandas.io.json”, “ json_normalize” , in order to transform JSON file into a pandas data frame and then we can Group the results as following:



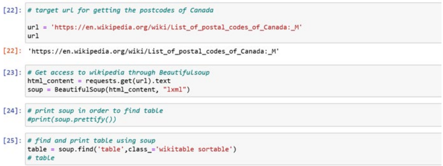
```
]: # show the most common category in Athens Kolonaki
   nearbyVenues.groupby('categories').count()
]:
```

categories	name	lat	lng
American Restaurant	1	1	1
Art Museum	1	1	1
Bar	4	4	4
Bistro	1	1	1
Bookstore	2	2	2
Boutique	5	5	5
Cafeteria	1	1	1
Café	11	11	11
Cheese Shop	1	1	1
Chocolate Shop	1	1	1
Clothing Store	1	1	1
Cocktail Bar	3	3	3

- We will also map the results of venues near our existing shop in Athens as following (using folium)

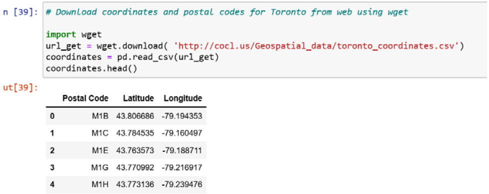


- We will use Beautiful soup in order to get from Wikipedia postal code, Neighbourhood and Borough of Toronto :



```
[22]: # target url for getting the postcodes of Canada

      url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
      url

[22]: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

[23]: # Get access to wikipedia through Beautifulsoup
      html_content = requests.get(url).text
      soup = BeautifulSoup(html_content, "lxml")

[24]: # print soup in order to find table
      #print(soup.prettify())

[25]: # find and print table using soup
      table = soup.find('table',class_='wikitable sortable')
      # table
```

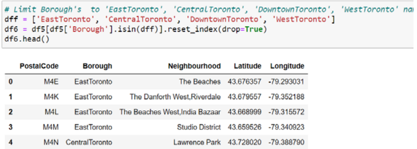
- We will use pandas in order to convert the above results into a pandas data frame
- We will use wget in order to download from cocl.us /Geospatial_data the coordinates of Toronto and the relevant postal code :



```
n [39]: # Download coordinates and postal codes for Toronto from web using wget

        import wget
        url_get = wget.download( 'http://cocl.us/Geospatial_data/toronto_coordinates.csv')
        coordinates = pd.read_csv(url_get)
        coordinates.head()
ut[39]:
```

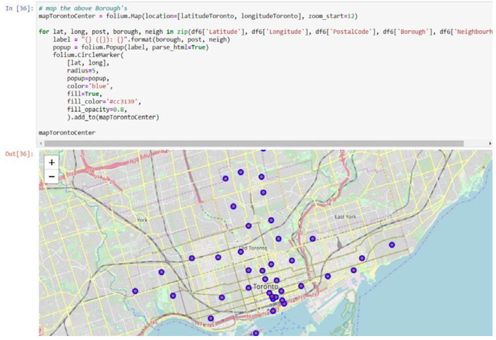
	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

- We will then merge the two tables in order to combine the info related to neighbourhood Borough Postal Code and coordinates and after doing some fine tuning in the data we get the final table and map (using folium):



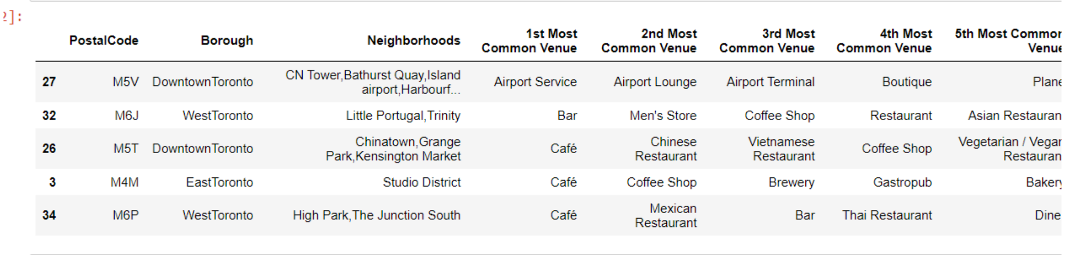
```
# Limit Borough's  to 'EastToronto', 'CentralToronto', 'DowntownToronto', 'WestToronto' na
dff = ['EastToronto', 'CentralToronto', 'DowntownToronto', 'WestToronto']
df6 = df5[df5['Borough'].isin(dff)].reset_index(drop=True)
df6.head()
```

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M4E	EastToronto	The Beaches	43.676357	-79.293031
1	M4K	EastToronto	The Danforth West,Riverdale	43.679557	-79.352188
2	M4L	EastToronto	The Beaches West,India Bazaar	43.668999	-79.315572
3	M4M	EastToronto	Studio District	43.659526	-79.340923
4	M4N	CentralToronto	Lawrence Park	43.728020	-79.388790



```
In [36]: # map the above Borough's
mapTorontoCenter = folium.Map(location=[latitudeToronto, longitudeToronto], zoom_start=12)

for lat, long, post, borough, neigh in zip(df6['Latitude'], df6['Longitude'], df6['PostalCode'], df6['Borough'], df6['Neighbourh
    label = "{} ({}): {}".format(borough, post, neigh)
    popup = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, long],
        radius=5,
        popup=popup,
        color='blue',
        fill=True,
        fill_color='#cc3139',
        fill_opacity=0.8,
        ).add_to(mapTorontoCenter)

mapTorontoCenter
```

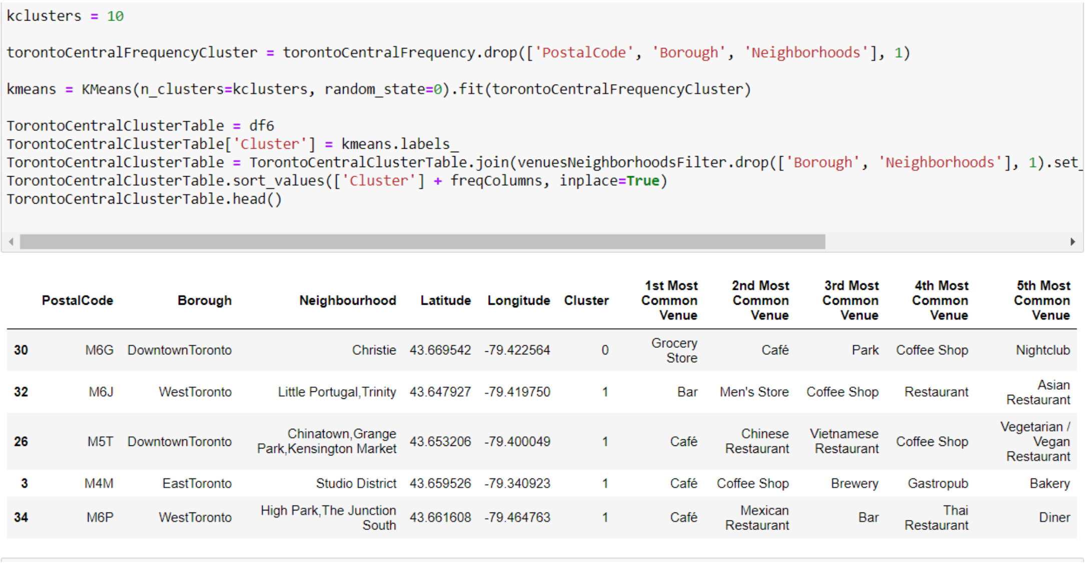
- Using Foursqure and with some data adjustments we get a list of Boroughs & neighbourhoods with the most common venues in Toronto Centre:



2]:

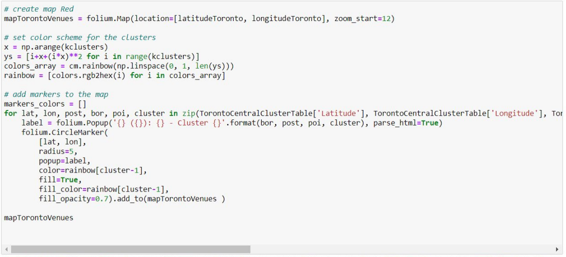
	PostalCode	Borough	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Commor Venue
27	M5V	DowntownToronto	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Service	Airport Lounge	Airport Terminal	Boutique	Plane
32	M6J	WestToronto	Little Portugal,Trinity	Bar	Men's Store	Coffee Shop	Restaurant	Asian Restauran
26	M5T	DowntownToronto	Chinatown,Grange Park,Kensington Market	Café	Chinese Restaurant	Vietnamese Restaurant	Coffee Shop	Vegetarian / Vegar Restauran
3	M4M	EastToronto	Studio District	Café	Coffee Shop	Brewery	Gastropub	Baker
34	M6P	WestToronto	High Park,The Junction South	Café	Mexican Restaurant	Bar	Thai Restaurant	Dine

- Using “sklearn.cluster” we import “KMeans” in order to create clusters of boroughs and Neighbourhoods and then we will choose the one or several “candidates” that are similar to our Athens shop Again we map our clusters using Folium library:

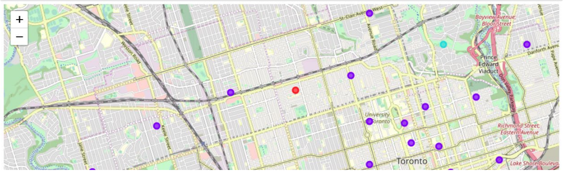


```
kclusters = 10

torontoCentralFrequencyCluster = torontoCentralFrequency.drop(['PostalCode', 'Borough', 'Neighborhoods'], 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(torontoCentralFrequencyCluster)

TorontoCentralClusterTable = df6
TorontoCentralClusterTable['Cluster'] = kmeans.labels_
TorontoCentralClusterTable = TorontoCentralClusterTable.join(venuesNeighborhoodsFilter.drop(['Borough', 'Neighborhoods'], 1).set_
TorontoCentralClusterTable.sort_values(['Cluster'] + freqColumns, inplace=True)
TorontoCentralClusterTable.head()
```

	PostalCode	Borough	Neighbourhood	Latitude	Longitude	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
30	M6G	DowntownToronto	Christie	43.669542	-79.422564	0	Grocery Store	Café	Park	Coffee Shop	Nightclub
32	M6J	WestToronto	Little Portugal,Trinity	43.647927	-79.419750	1	Bar	Men's Store	Coffee Shop	Restaurant	Asian Restaurant
26	M5T	DowntownToronto	Chinatown,Grange Park,Kensington Market	43.653206	-79.400049	1	Café	Chinese Restaurant	Vietnamese Restaurant	Coffee Shop	Vegetarian / Vegan Restaurant
3	M4M	EastToronto	Studio District	43.659526	-79.340923	1	Café	Coffee Shop	Brewery	Gastropub	Bakery
34	M6P	WestToronto	High Park,The Junction South	43.661608	-79.464763	1	Café	Mexican Restaurant	Bar	Thai Restaurant	Diner



```
# create map Red
mapTorontoVenues = folium.Map(location=[latitudeToronto, longitudeToronto], zoom_start=12)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i+x+(i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, post, bor, poi, cluster in zip(TorontoCentralClusterTable['Latitude'], TorontoCentralClusterTable['Longitude'], Tor
    label = folium.Popup('{} ({}): {} - Cluster {}'.format(bor, post, poi, cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(mapTorontoVenues )

mapTorontoVenues
```



From the above cluster of Toronto centre we will choose the one that is shares similar characteristics (venue categories) to Athens shop