Capstone Project
The Battle of Neighborhoods

Prospects of a Coffee Shop close to business areas in Milan, Italy.

1 Part 1

This project is the subject of the IBM Capstone Project which is the last assignment of the Coursera's course Applied Data Science Capstone.

In this document define a business problem, look for data in the web and, use Foursquare location data to compare different districts within municipalities of Milan to figure out which neighborhood is suitable for starting a coffee shop business. As prepared for the assignment, I go through the problem designing, data preparation and final analysis section step by step.

Detailed codes and images are given in Github and link can be found at the end of the document.

1.1 Discussion and Background of the Business Problem

Milan is a metropolis in Italy's northern Lombardy region. It is a global capital of fashion and design. Home to the national stock exchange, it's a financial hub also known for its high-end restaurants and shops. Milan is definitely one of the best places to start up a new business.

During the daytime, especially in the morning and lunch hours, office areas provide huge opportunities for coffee shops. Reasonably priced (one lunch meal 5€) shops are usually always full during the lunch hours (11 am — 2 pm) and, given this scenario, we will go through the benefits and pitfalls of opening a breakfast cum lunch coffee shop in highly dense office places.

The core of Milano is made of 9 municipalities but, we will later concentrate on 4 most busiest business boroughs of Milan: Centro Storico, Stazione Centrale, Città Studi e Porta Garibaldi to target daily office workers.

We will go through each step of this project and address them separately. I first outline the initial data preparation and describe future steps to start the battle of neighborhoods in Milan.

1.2 Target Audience

This project is mainly aimed at group of people:

- Entrepreneurs who wants to invest or open a coffee shop. This analysis will be a comprehensive guide to start or expand coffee shops targeting the large pool of office workers in Milan during lunch hours.
- People with passion for data that want see an example of applied data science

1.3 Data Section

1.3.1 Milan Municipalities Table from Wikipedia

I first make use of page Zones of Milan from Wikipedia to scrap the table to create a data-frame. For this, I've used requests and Beautifulsoup4 libraries to create a data-frame containing name of the 9 municipalities of Milan.

We start as below:

```
response = requests.get('https://en.wikipedia.org/wiki/Zones_of_Milan').text
soup = BeautifulSoup(response,'lxml')
table = soup.find('table', {'class':'wikitable sortable'})
```

```
table_rows = table.find_all('tr')

res = []
for tr in table_rows:
    td = tr.find_all('td')
    row = [tr.text.strip() for tr in td if tr.text.strip()]
    if row:
        res.append(row)


df = pd.DataFrame(res, columns=["Num", "LongName", "Area", "Population","Density", "Districts"])
#remove column Districts
df = df.drop(columns=['Districts'])
df
```

After little manipulation, the data-frame is obtained as below:

	Num	LongName	Area	Population	Density
0	1	Centro storico	9.67	96,315	11,074
1	2	Stazione Centrale, Gorla, Turro, Greco, Cresce...	12.58	153.109	13,031
2	3	Città Studi, Lambrate, Porta Venezia	14.23	141,229	10,785
3	4	Porta Vittoria, Forlanini	20.95	156.369	8,069
4	5	Vigentino, Chiaravalle, Gratosoglio	29.87	123,779	4,487
5	6	Barona, Lorenteggio	18.28	149,000	8,998
6	7	Baggio, De Angeli, San Siro	31.34	170,814	6,093
7	8	Fiera, Gallaratese, Quarto Oggiaro	23.72	181,669	8,326
8	9	Porta Garibaldi, Niguarda	21.12	181,598	9,204

Data-frame from Wikipedia Table.

1.3.2 Dataset Coordinates of Milan boroughs: Geopy Client

Next objective is to get the coordinates of these 4 major districts using geocoder class of Geopy client. Using the code snippet as below:

```
from geopy.geocoders import Nominatim
geolocator = Nominatim()
location = geolocator.geocode("Milano, MI, Lom, Italia")
address = []
coord = []
address = df['Shortname']+", Milano, MI, Lom, Italy"
coord = address.apply(geolocator.geocode).apply(lambda x: (x.latitude, x.longitude))
df['Coordinates'] = coord
df
```

]:

	Num	LongName	Area	Population	Density	Shortname	Coordinates
0	1	Centro storico	9.67	96,315	11,074	Centro storico	(45.41921235, 9.07080197950279)
1	2	Stazione Centrale, Gorla, Turro, Greco, Cresce...	12.58	153.109	13,031	Stazione centrale	(45.4866591, 9.2072566)
2	3	Città Studi, Lambrate, Porta Venezia	14.23	141,229	10,785	Città studi	(45.4770557, 9.2265746)
3	4	Porta Vittoria, Forlanini	20.95	156.369	8,069	Porta vittoria	(45.4622607, 9.2095796)
4	5	Vigentino, Chiaravalle, Gratosoglio	29.87	123,779	4,487	Vigentino	(45.4399296, 9.2004923)
5	6	Barona, Lorenteggio	18.28	149,000	8,998	Barona	(45.4388451, 9.1546701)
6	7	Baggio, De Angeli, San Siro	31.34	170,814	6,093	Baggio	(45.4614328, 9.0910822)
7	8	Fiera, Gallaratese, Quarto Oggiaro	23.72	181,669	8,326	Fiera	(45.5202499, 9.0789880116852)
8	9	Porta Garibaldi, Niguarda	21.12	181,598	9,204	Porta garibaldi	(45.4806652, 9.1868884)

1.3.3 Average Property Price in Major Municipalities of Milan: Web Scrapping

Another factor that can guide us later for deciding which district would be best to open a coffee shop is, the average price of a property in the 9 boroughs. I get this information from scrapping 'Mercato Immobiliare a Milano' web-page, similarly to the Wiki page before. As I want to consider the 4 busiest business municipalities of Milan as mentioned in section 1. They are: "Centro Storico", "Stazione Centrale", "Città Studi", "Porta Garibaldi".

The data-frame looks as below:

	Shortname	Price
0	Centro storico	€ 6.500 /m²
1	Città studi	€ 3.950 /m²
2	Porta garibaldi	€ 5.500 /m²
3	Stazione centrale	€ 4.550 /m²

1.3.4 Joining results

Now we merge price property dataset with boroughs dataset and we obtain the following result.

```
df3 = pd.merge(df, df2, on='Shortname', how='inner')
```

```
df3
```

	Num	LongName	Area	Population	Density	Shortname	Latitude	Longitude	Price
0	1	Centro storico	9.67	96,315	11,074	Centro storico	45.419212	9.070802	€ 6.500 /m²
1	2	Stazione Centrale, Gorla, Turro, Greco, Cresce...	12.58	153.109	13,031	Stazione centrale	45.486659	9.207257	€ 4.550 /m²
2	3	Città Studi, Lambrate, Porta Venezia	14.23	141,229	10,785	Città studi	45.477056	9.226575	€ 3.950 /m²
3	9	Porta Garibaldi, Niguarda	21.12	181,598	9,204	Porta garibaldi	45.480665	9.186888	€ 5.500 /m²

1.4 Conclusion Part1

We get the initial dataframe with Names of Major Municipalities, and corresponding coordinates of those major districts and average property price. Before comparing all the municipalities, since we want to concentrate only on lunch coffee shops targeting the office workers, we need to get the idea about the best business areas in Milan. Here we want to concentrate on the best four boroughs:

- Centro storico
- Stazione Centrale
- Città Studi
- Porta Garibaldi

So as the next step we will use Foursquare data and obtain information on coffee shops. With these, we can start with our battle of neighborhoods for opening a coffee shop in Milan.

2 Part 2 Explore neighborhoods using Foursquare data

2.1 Using Foursquare Location Data

Foursquare is a very comprehensive data provider. For this business problem we have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these 4 Major Districts of Milan.

The popular spots returned depends on the highest foot traffic and thus it depends on the time when the call is made. So we may get different popular venues depending upon different time of the day.

The call returns a JSON file and we need to turn that into a data-frame. Here I've chosen 100 popular spots for each major districts within a radius of 1 km.

Below is the data-frame obtained from the JSON file that was returned by Foursquare:

	District	Dist_Latitude	Dist_Longitude	Venue	Venue_Lat	Venue_Long	Venue_Category
300	Porta garibaldi	45.507704	9.17941	Virgin Active	45.502018	9.182590	Gym / Fitness Center
301	Porta garibaldi	45.507704	9.17941	Al Paradiso Della Pizza	45.511351	9.175416	Pizza Place
302	Porta garibaldi	45.507704	9.17941	Birrificio La Ribalta	45.507038	9.173219	Brewery
303	Porta garibaldi	45.507704	9.17941	Istanbul Kebab	45.510421	9.176128	Kebab Restaurant
304	Porta garibaldi	45.507704	9.17941	Esselunga	45.512380	9.173461	Supermarket
305	Porta garibaldi	45.507704	9.17941	Total Natural Training	45.509142	9.194221	Gym
306	Porta garibaldi	45.507704	9.17941	Spirit de Milan	45.506678	9.159744	Ballroom
307	Porta garibaldi	45.507704	9.17941	Il Bucatino con Giardino	45.502088	9.165959	Italian Restaurant
308	Porta garibaldi	45.507704	9.17941	Sushi Ran	45.500348	9.170942	Japanese Restaurant
309	Porta garibaldi	45.507704	9.17941	Hotel La Residenza	45.512413	9.178770	Hotel
310	Porta garibaldi	45.507704	9.17941	Piazza Dergano	45.504034	9.175839	Plaza
311	Porta garibaldi	45.507704	9.17941	Palestra McFIT	45.504708	9.198828	Gym
312	Porta garibaldi	45.507704	9.17941	Biologic Bar & Restaurant	45.512302	9.178709	Hotel Bar
313	Porta garibaldi	45.507704	9.17941	MiSciolgo	45.499214	9.165308	Ice Cream Shop
314	Porta garibaldi	45.507704	9.17941	Teatro della Cooperativa	45.515042	9.190504	Theater

---

2.2 Visualization and Data Exploration

2.2.1 Folium Library and Leaflet Map

Folium is a python library that can create interactive leaflet map using coordinate data. Since I am interested in restaurants as popular spots first I create a data-frame where the 'Venue_Category' column in previous data-frame contains the words "Cof", "Caf", "Bar".

I used the following snippet of code:

```
# Create a Data-Frame out of it to Concentrate Only on Coffee Shops
Milan_Cafe = Milan_Venues[Milan_Venues['Venue_Category'].str.contains('Caf|Cof|Bar')].reset_index(drop=True)
Milan_Cafe.index = np.arange(1, len(Milan_Cafe)+1)
print ('Shape of the Data-Frame with Venue Category Cafe: ', Milan_Cafe.shape)
Milan_Cafe.head(100)
```

```
Shape of the Data-Frame with Venue Category Cafe:  (45, 7)
```

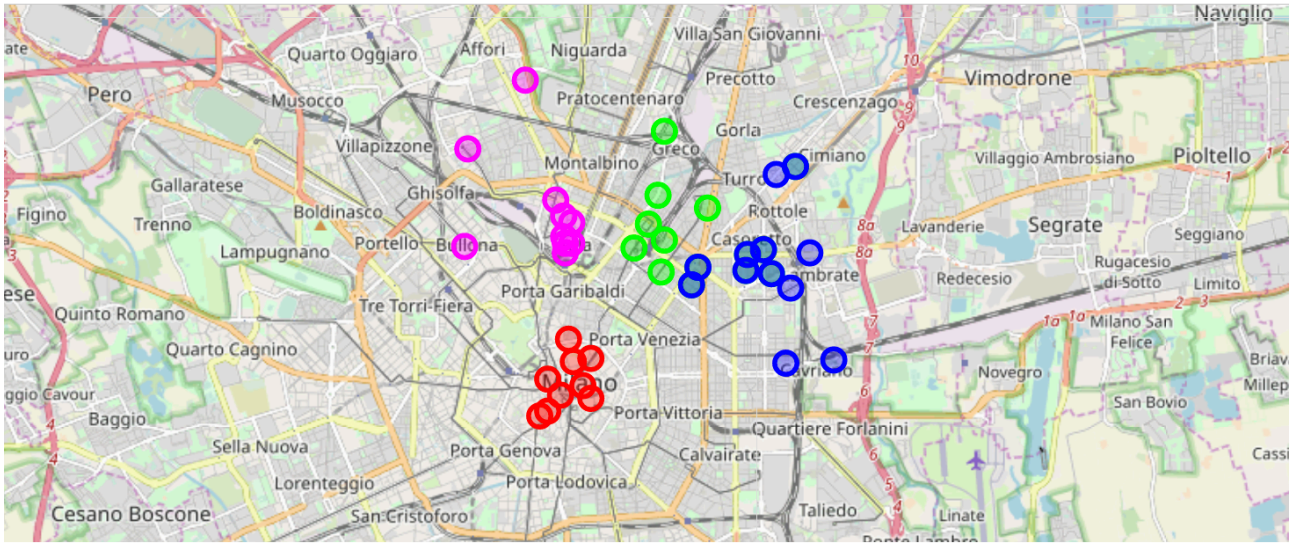
7]:

	District	Dist_Latitude	Dist_Longitude	Venue	Venue_Lat	Venue_Long	Venue_Category
1	Centro storico	45.467281	9.185962	Starbucks Reserve Roastery	45.464920	9.186153	Coffee Shop
2	Centro storico	45.467281	9.185962	Signorvino	45.467243	9.183567	Wine Bar
3	Centro storico	45.467281	9.185962	Bulgari Lounge Bar	45.470014	9.188943	Cocktail Bar
4	Centro storico	45.467281	9.185962	Lavazza Coffee Design	45.466274	9.190975	Coffee Shop
5	Centro storico	45.467281	9.185962	B Café	45.462640	9.183381	Café
6	Centro storico	45.467281	9.185962	Signorvino	45.464552	9.192885	Wine Bar

Next step is to use this dataframe to create a leaflet map with Folium to see the distribution of the venues in the 4 major districts.

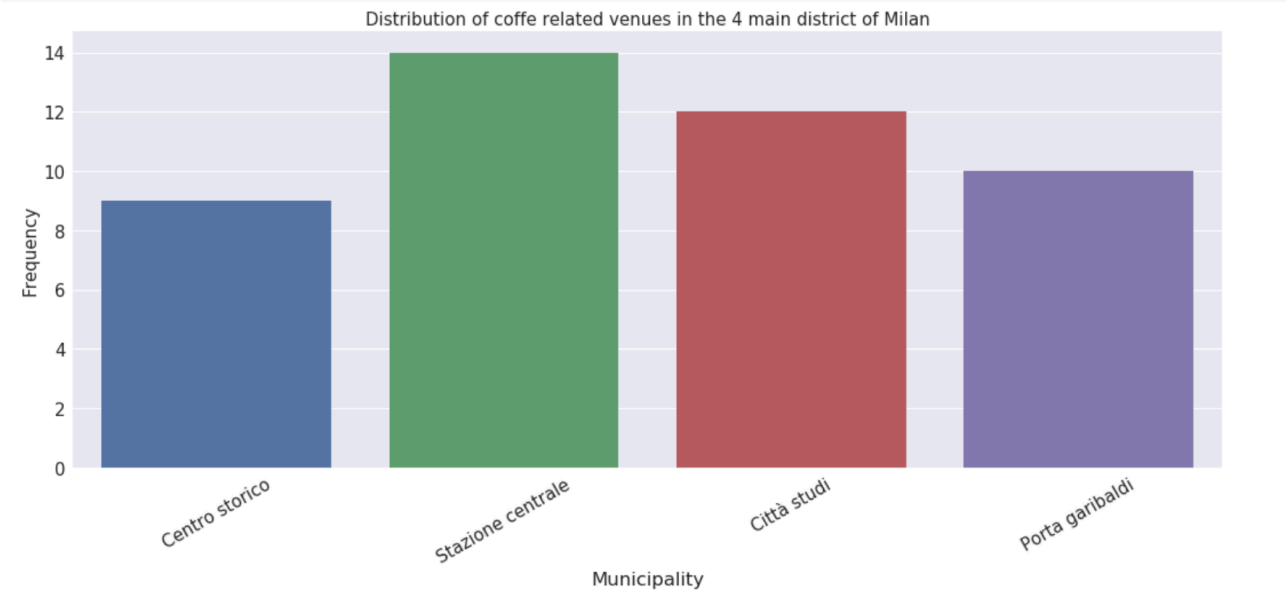
```
map_cafes = folium.Map(location=[latitude, longitude], zoom_start=11, tiles="openstreetmap",
                       attr="<a href=https://github.com/python-visualization/folium/>Folium</a>")

# set color scheme for the Venues based on the Major Municipalities
Municips = ['Centro storico', 'Stazione centrale', 'Porta garibaldi', 'Città studi']

x = np.arange(len(Municips))

rainbow = ['#00ff00', '#ff00ff','#0000ff','#ffa500' ,'#ff0000']

# add markers to the map
# markers_colors = []
for lat, lon, poi, distr in zip(Milan_Cafe['Venue_Lat'],
                                Milan_Cafe['Venue_Long'],
                                Milan_Cafe['Venue_Category'],
                                Milan_Cafe['District']):
    label = folium.Popup(str(poi) + ' ' + str(distr), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=7,
        popup=label,
        color=rainbow[Municips.index(distr)-1],
        fill=True,
        fill_color=rainbow[Municips.index(distr)-1],
        fill_opacity=0.3).add_to(map_cafes)
```

With the code snippet above the leaflet map looks as below

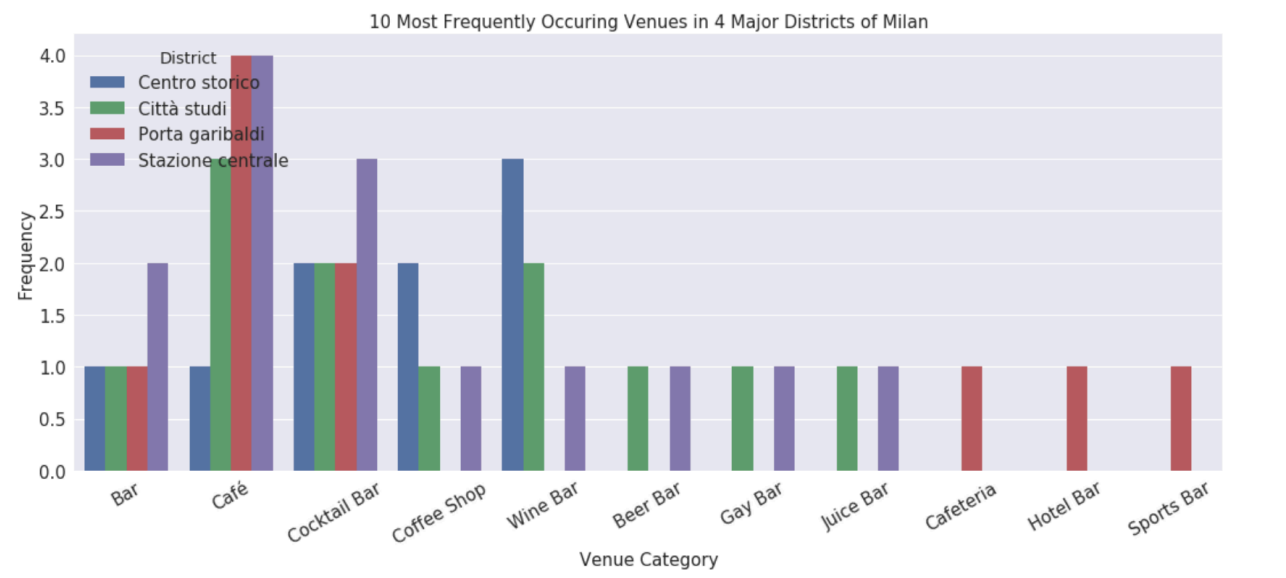


2.2.2 Exploratory Data Analysis

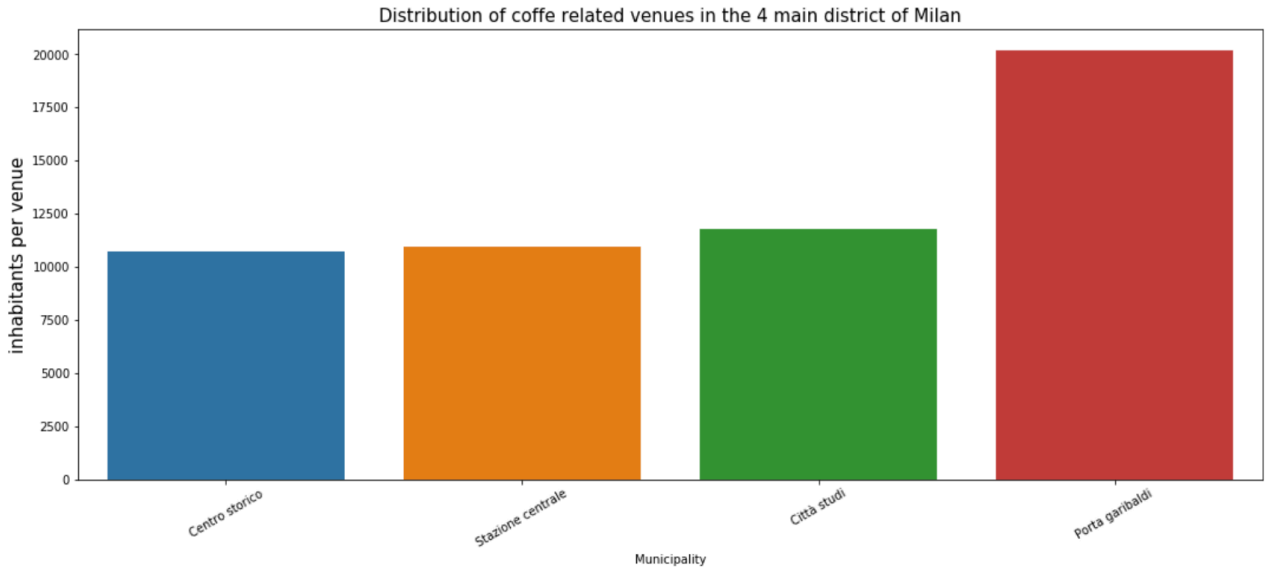
There are 45 unique venues distributed in the main 4 districts of Milan as shown in the following picture:



The following picture shows the frequency of the single coffee related categories in each of the main 4 districts:



By calculating the population by the number of venues we obtain the following picture:



2.3 Results and Discussion

Purpose of this project was to identify Milan areas with low number of coffee shops in order to aid stakeholders in narrowing down the search for optimal location for a new venue.

Foursquare offered us the opportunity to explore Milan's neighborhoods, by extracting all the venues related to Coffee shop, Café, Bar, and other similar venues located in the area.

We limited our analysis to the 4 main district (municipalities) of Milan: Centro storico, Stazione centrale, Porta Garibaldi, Città Studi. The results shows that they are the most districts with the highest average of property price.

Foursquare data resulting from our analysis shows that the number of Cafe, Bar and Coffee shop is quite low considering the population and the size of the area.

The most common category is Cafè and they are concentrated in Stazione centrale and Porta Garibaldi.

Porta Garibaldi is the district with the highest ratio inhabitants per venue (over than 20 thousands). As this district also is one of the most expanding business areas in Milan it looks as best location for opening a new Café or Coffee shop.

Anyway the accuracy of data purely depends on the data provided by FourSquare, as there could be venues that are not registered there.

Also recommended district should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

Find the code in Github.