Coursera CapStone Project:
Clustering stations in the Italian region of Lazio

Figure 1 - Railroad system of Lazio

Index

Introduction

This project is a "Capstone Project" of Coursera - "IBM Certified Data Scientist" program.

The project is focused on the railway stations located in the Italian region of Lazio, a territory populated by 5.9 million people, visited by more than 12 million tourists each year and served by 165 train stations (41 of them inside the city of Rome).

Rete Ferroviaria Italiana is the owner of the stations, and classifies them as “bronze”, “silver”, “gold” and “platinum”, depending on the services offered at the venue, to both travelers and non-travelers.

Goal of the project is to use data gathered from Foursquare and other sources to classify the stations independently from the RFI labels, and understand if some stations should deserve a different status, according to the characteristics of the nearby area.

Figure 2 - Tiburtina Station, one of the three "Platinum" level stations in Lazio

Business Problem

RFI train stations underwent a significant transformation, from travel hubs to meeting centers and places of aggregation, going beyond their original role of serving travelers by offering diverse services to travelers and non-travelers as well, such as catering services, parking places, retail stores, coworking places and more.

The service offering of a station depends on its class, that goes from “bronze” to “silver”, “gold” and “platinum”.

Goal of this study is to use open data about the 165 train stations in the Italian region of Lazio to categorize them depending on a series of factors, such as the number of check-ins and reviews, and characteristics of the nearby area such as traffic and population, and the number of significant venues in a radius of 1.5 km from the station, such as number of restaurants, museums, universities, cafes, professional buildings, hotels, shops, gyms and more.

This categorization will be useful to compare the current classification made by RFI with the one deriving from the above parameters, to understand if some station should deserve a different status, such as a promotion or a demotion.

The main target audience will be, therefore, RFI management.

Figure 3 - One of the large restaurant and meeting venues in Termini Station, Rome

Data

The table below reports the type of data needed and their sources.

Data	Source	Last updated
List of RFI train stations in the Lazio region, with location and classification	http://www.rfi.it/rfi/LINEE-STAZIONI-TERRITORIO/Nelle-regioni/Lazio	2019
Number of check-ins and reviews of each station	FourSquare API	Daily updates
Top 100 venues in a 1000 meters range of each station, categorized by high-level groups	FourSquare API	Daily updates
Population and density of the neighborhoods hosting the stations	https://it.wikipedia.org/wiki/Municipi_di_Roma https://www.comune.roma.it/web-resources/cms/documents/Territorio2017DEF.pdf https://www.tuttitalia.it/lazio/27-comuni/popolazione/	2017 2017 2019
Pollution levels (as a rough indicator of traffic)	http://www.arpalazio.net/main/aria/sci/qa/misure/PM10.php	2019

Methodology

For each station, from the address the latitude and longitude will be derived, and those in turn will be used to retrieve the FourSquare data and venues. The match with population, density and pollution data will be done on the basis of the neighborhood hosting the station.

The above data will be collected in a single dataframe, with a row for each station and columns reporting population, density, pollution and a set of FourSquare categories with the number of top venues occurring for each category.

The features will be standardized for better manipulation of the clustering algorithm, and then a clustering will be performed with 4 clusters, the same number of the classes used by RFI.

The original classes and the new clusters will then be compared, using both tables and maps, to check for correlations between the two.

Loading Data

Loading the stations from the web page

In [ ]:
```
#loading stations from the web page

# source url
url = "http://www.rfi.it/rfi/LINEE-STAZIONI-TERRITORIO/Nelle-regioni/Lazio"

# performing the request
file = requests.get(url).text
```

Parsing the text with BeautifulSoup to retrieve the table

In [ ]:
```
# parsing data with Beautiful Soup
parsable_file = BS(file, 'lxml')

# retrieving the table
data_table_list = parsable_file.find_all('table')
data_table = data_table_list[1]
```

Converting the table into a list

In [ ]:
```
# converting the table into a list
list = pd.read_html(str(data_table), header=0)
list
```

Figure 4 - Use of Beautiful Soup to websource data

Results and Discussion

Goal of the project was to use data gathered from Foursquare and other sources to classify the stations independently and understand if some stations should deserve a different status from the one given by the network manager (RFI = Rete Ferroviaria Italiana), according to the characteristics of the nearby area.

RFI, in fact, arranged the stations into four categories, depending on the level of services offered in each station.

The list of the 165 train stations of the Italian region named Lazio was successfully websourced, as well as their addresses.

Two more datasets were websourced:

1. a list of pollution measuring stations, reporting average yearly emissions (PM10) and number of days when the maximum threshold has been trespassed (overcomings); these data were used mainly as indicators of local traffic;
2. a list of all main neighbourhoods in Lazio, with the overall population and density.

Using Nominatim, the addresses of the three POIs above were converted into geographical coordinates (longitude and latitude).

In [ ]:
```
# retrieving longitude and latitude of each train station

# instantiating a geolocator
geolocator = Nominatim(user_agent="stat_explorer")
count = 0
# retrieving data for each station
columns_list = []

for address, city, name in zip(df_stations['Indirizzo'], df_stations['Comune/Località'],
                               df_stations['Nome Stazione/fermata']):
    complete_address = address + ',' + city
    # passing the location to the geolocator
    location = geolocator.geocode(complete_address)
    if location is None:
        location = geolocator.geocode(str(name))
        print('nome '+name)
    else: print('complete_address '+complete_address)
    # retrieving latitude and longitude from the geolocator


    print(location)
    print(count)
    count = count + 1
    latitude = location.latitude
    longitude = location.longitude
    display_name = location.address
    columns_list.append([latitude, longitude, display_name])

columns_list
```

Figure 5 - Using Nominatim to retrieve latitude and longitude from address

After that, a function based on the Vincenty geodesic distance was employed, to associate the closest pollution station and the closest neighbourhood, as well as their data, to each train station.

```
# this function returns population total and density for given coordinates
def find_population(latitude, longitude):

    # station coordinates
    stat_coord = (latitude, longitude)

    temp_distance = max_distance

    population = []
    for lat, long, pop, dens in zip(df_pop_coord['latitude'], df_pop_coord['longitude'],
                                    df_pop_coord['Popolazione(ab)'], df_pop_coord['Densità(ab/km²)']):
        # neighbourhood coordinates
        pop_coord = (lat, long)
        # distance between points according to the Vincenty's formula
        distance = geopy.distance.vincenty(stat_coord, pop_coord).km
        if distance < temp_distance:
            temp_distance = distance
            temp_population = pop
            temp_density = dens

    temp_population
    temp_density

    return temp_population, temp_density

columns_list = []
count = 0

# finding the pollution and population values for each station
for latitude, longitude in zip(df_stations_coord['latitude'], df_stations_coord['longitude']):

    # finding the pollution values corresponding to the station
    pm10, overcoming = find_pollution(latitude, longitude)

    # finding the population values corresponding to the station
    population, density = find_population(latitude, longitude)
```

Figure 6 - Using Vincenty distance to associate POIs through coordinates

The final step of the data collection used FourSquare to gather, for each station, the number of venues in a range of 1500 meters, arranged in the FourSquare top 10 categories, that are:

- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport

After scaling the data, the K-Means clustering was applied, using 4 clusters, the same number of the RFI categories.

The visual analysis of the clusters showed a first cluster with an excellent presence in all categories, with the second significantly below. The third and fourth ones had smaller quantities in all venue indicators and were differentiated mostly by distribution of venues and population/pollution differences.

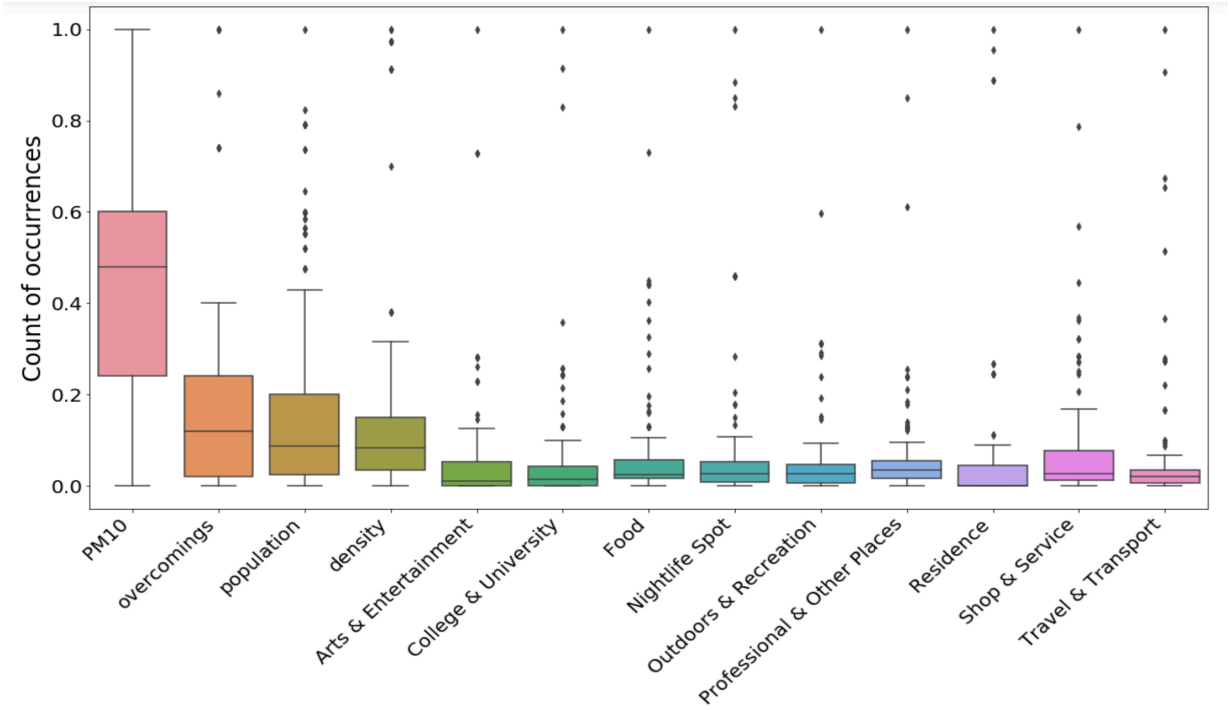

Figure 7 - The most represented venues are Food, Professional and Shops, and all venues present significant upper outliers. PM10 and overcomings are more evenly distributed.

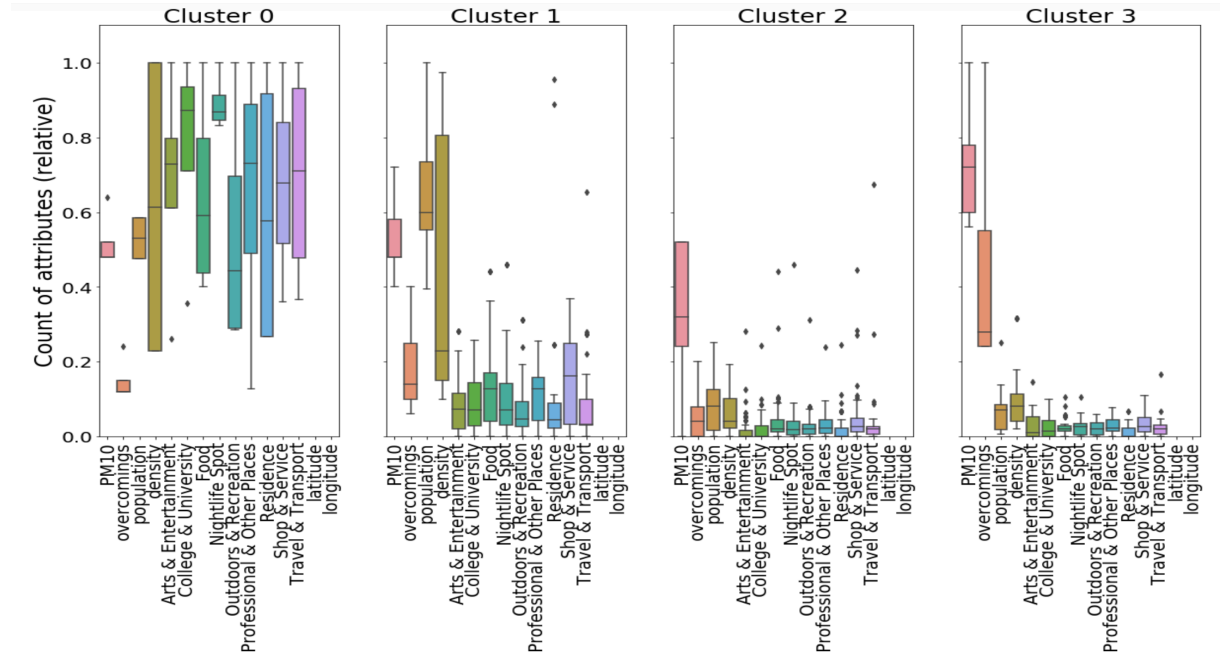

Figure 8 - The first cluster shows an excellent presence in all categories, while the second is significantly below. The third and fourth ones have smaller indicators and are differentiated mostly by distribution of venues and population/pollution differences.

84% of stations matched the original RFI classification. 16% of stations fell into the category immediately below or above. This overall agreement between original categories and computed clusters was shown by visual maps as well.

	Nome Stazione/fermata	Categoria	latitude	longitude	PM10	overcomings	population	density	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	R
0	ACQUA ACETOSA	BRONZE	41.792689	12.620391	27	13	38963	2998	3	6	0	6	6	8	3	
1	ALBANO LAZIALE	SILVER	41.726437	12.657916	27	13	18574	999	7	5	0	26	6	5	5	
2	ANAGNI-FIUGGI	SILVER	41.710359	13.096031	20	3	21249	188	0	0	0	2	1	0	0	
3	ANGUILLARA	SILVER	42.065953	12.293394	18	1	19459	259	0	1	0	6	3	1	8	
4	ANTRODOCO CENTRO	BRONZE	42.410244	13.071053	18	1	45800	335	0	0	0	7	4	0	0	
5	ANTRODOCO-BORGO VELINO	BRONZE	42.410244	13.071053	18	1	45800	335	0	0	0	7	4	0	0	
6	ANZIO	SILVER	41.452150	12.629295	25	8	55101	1262	4	6	0	25	4	11	5	
7	ANZIO COLONIA	BRONZE	41.460169	12.617093	25	8	55101	1262	1	4	0	7	7	3	1	

Figure 9 - Example of aggregation of different data on each single station

	Nome Stazione/fermata	Categoria	Cluster
8	APPIANO PROBA PETRONIA	2	1
13	BAGNI DI TIVOLI	2	3
19	CAPANNELLE	2	1
22	CASSINO	1	2
27	CECCANO	2	3
31	CIAMPINO	1	2
35	CIVITAVECCHIA	1	2
47	FIUMICINO AEROPORTO	1	2
52	FORMIA-GAETA	1	2
57	GEMELLI	2	1
73	LATINA	1	2
82	MARINO LAZIALE	2	3
85	MONTE MARIO	2	1
92	NUOVO SALARIO	2	1
95	ORTE	1	2
101	PAVONA	2	3
110	QUATTRO VENTI	2	1
116	ROMA AURELIA	2	1
117	ROMA BALDUINA	2	1
120	ROMA PRENESTINA	2	1
121	ROMA S.PIETRO	1	0
123	ROMA TIBURTINA	0	1
124	ROMA TRASTEVERE	1	0
146	TIVOLI	2	3
148	TOR VERGATA	2	1
154	VALMONTONE	2	3
164	ZAGAROLO	2	3

Figure 10 - List of the stations that received a different classification, the original category is the second-last on the right

The stations that were differently classified fell mostly into 2 categories:

1. stations in the centre of Rome, with a venue rich neighbourhood, promoted to the above category (such as Trastevere, S. Pietro, Aurelia, Quattro Venti, Balduina, Monte Mario)
2. stations playing a crucial role as a node in the transportation network, but with poorer neighbourhood, that were demoted to the category immediately below (such as Tiburtina, Fiumicino, Ciampino, Civitavecchia)

Therefore, notwithstanding the fact that FourSquare Data is un-balanced, with some categories, such as food, over-represented, the clustering made quite sense, with the stations distribution matching broadly the investments made by the network manager on each station, as reflected by the category level.

The non-matching labels made sense as well, showing that some stations correctly offer higher level of service than the one suggested by the nearby area, given the strategic position (a close-by airport or harbor), while others should deserve a promotion considering the amount of venues and activities in the nearby area.

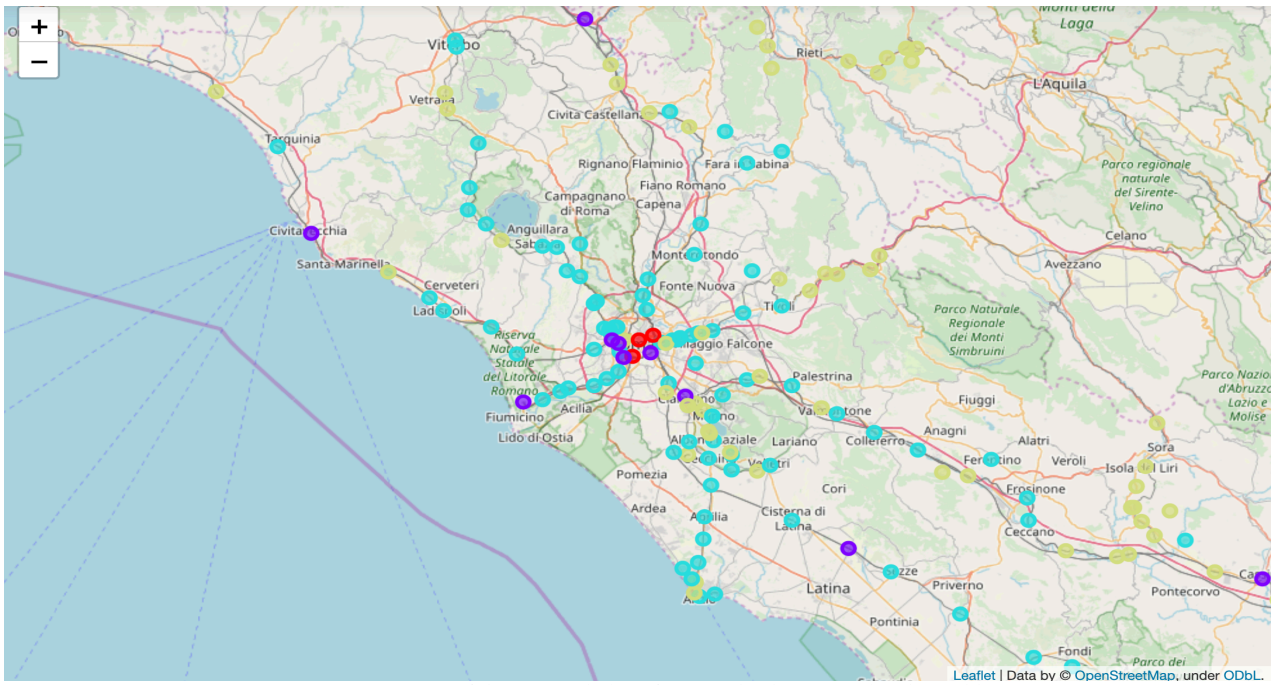

Figure 11 - Map of the original classification, harbors are clearly visible

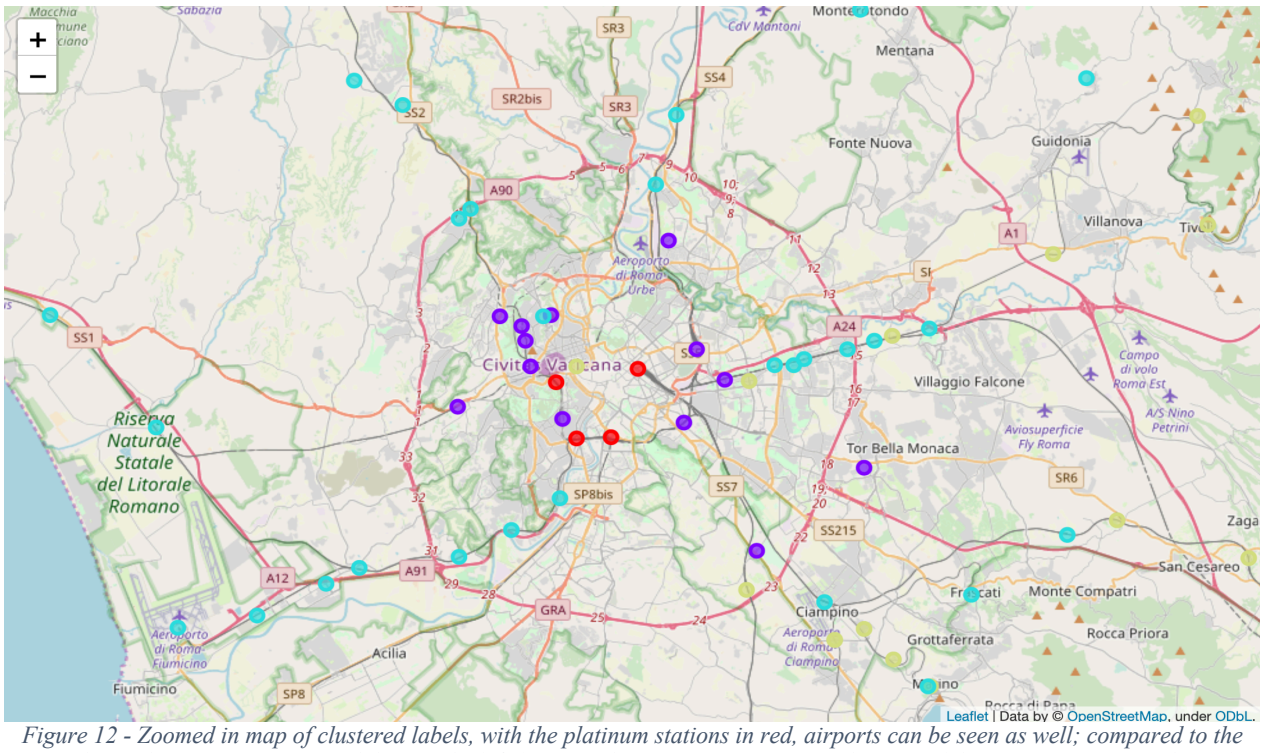

Figure 12 - Zoomed in map of clustered labels, with the platinum stations in red, airports can be seen as well; compared to the map above, there are two more red dots on the left (promoted stations: Trastevere and S. Pietro) but a missing one on the right (Tiburtina was demoted to Gold by the clustering algorithm)

Conclusions

The project made use of websourced data and a non-commercial FourSquare account.

Notwithstanding these limited resources, it proved strong insights into the areas surrounding each station, supporting the choices made by the rail network manager and giving some valuable suggestion.

Using a commercial FourSquare account could give a more granular description of venues and increase the understanding of stations' neighbourhoods.

Further applications could be:

1. exploring the predicted impact of a station promotion/demotion on its reviews;
2. exploring the correlation of check-ins and reviews to other parameters, like availability of certain venues, such as parking or bus stops or restaurants, or the lack of others, such as coworking places or professional structures;
3. the creation of a recommendation system able to advice citizens and tourists about the nearby venues of each station, depending on their needs and past choices of similar users;
4. forecasting the profitability of parking places offered by the station analyzing the demand and availability in the nearby area.

The project could be extended to the whole Italian network as well.

Figure 13 - The national RFI network