Coursera CapStone Project: clustering stations in the Italian region 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 and understand if some stations should deserve a different status, according to the characteristics of the nearby area.

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 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.

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.

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 will be derived the latitude and longitude, 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.