Capstone project report

Contents

[Introduction 1](#_Toc21946480)

[Data used 1](#_Toc21946481)

[Data on Toronto neighbourhoods 2](#_Toc21946482)

[Analysis on Milan data 2](#_Toc21946483)

[Data acquisition and cleaning 2](#_Toc21946484)

[Clustering of Milan’s neighbourhoods 3](#_Toc21946485)

[Comparison to Toronto neighbourhoods 3](#_Toc21946486)

[Analysis of the distribution of price tier and rating variables concerning food venues in Milan 3](#_Toc21946487)

# Introduction

I will be acquiring information regarding Milan's boroughs and will perform an analysis similar to what we've done with Toronto and NY. I'll then proceed to compare it to other cities. The goal is to get insight about how cities from very different countries compare and if a clustering algorithm, depending on the criteria used, can distinguish their neighbourhoods. The analysis will provide information on how to distinguish between parts of a city, and how different kinds of neighbourhood (residential, touristic, financial...) might compare between cities.

As an example, I am going to use information about the most popular venues per neighbourhood, and use it to cluster them. Each neighbourhood popular venues blend, made of restoration, services and attraction venues in different proportions, can in fact give us insights about its geographical location and position relatively to tourism, business or services centres.

I am then going to carry out an analysis of Milan’s neighbourhood based on food venues rating and price tier. The goal is to correlate those two indicators with the neighbourhood location and nearby presence of tourism, business or services centres, possibly comparing it to the previous analysis.

This project is obviously limited in scope but it provides an interesting starting point for further work on the same data and topics.

# Data used

Milan data are not readily available as for NY or Toronto; basic information about Milan’s boroughs and neighbourhoods are found on Wikipedia (<https://en.wikipedia.org/wiki/Zones_of_Milan>), and the spatial coordinates can be obtained through services as geocoder.

An alternative and more efficient solution is to use an institutional database, which is available online (<http://dati.comune.milano.it/dataset/5c6519f6-6d26-41c9-b53b-6106e08d1b90/resource/533b4e63-3d78-4bb5-aeb4-6c5f648f7f21/download/ds634_civici_coordinategeografiche_20190902_csv.zip>), and contains spatial coordinates of each and every single address in Milan and relative neighbourhood information. By averaging the coordinates for all addresses specific to one neighbourhood it is possible to get a good estimate of each neighbourhood centre location, especially given the size of the database. Location of the coordinates on the map confirms the correctness of the operation.

Information on the neighbourhood popular venues is then obtained through Foursquare API, whereas data on food venues ratings and price tier is available through Yelp API.

Data from Toronto from the previous part of the project course.

## Data on Toronto neighbourhoods

Data on Toronto has been acquired as part of the project course, in the previous module. Information about neighbourhoods and boroughs was scraped from the Wikipedia page <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> , and the coordinates through the geocoder package, or provided with the course material, depending on the correct functioning of geocoder. I have then acquired information about popular venues in central Toronto through the Foursquare API, and clustered the neighbourhoods according to their venues composition. The result is available on <https://github.com/RaffaToSpace/Coursera_Capstone/blob/master/Toronto_clustering_analysis.ipynb> , and shows how neighbourhoods in central Toronto are pretty homogeneous in composition, with some exceptions due to the presence of services or infrastructure (such as schools, parks, etc.) in some neighbourhoods.

## Data acquisition and cleaning

Once obtained the geographical data of Milan from the institutional website I copied it into a dataframe and proceeded into extracting relevant information from it. The table contains a number of empty cells and non-valid values, requiring cleaning and sorting. I extracted a list of neighbourhoods and averaged the spatial coordinates of all the addresses for each one of them to find the central neighbourhood set of coordinates. As displayed on a map in Figure 1, the list of neighbourhoods is sufficiently representative of the areas of town and the dots are evenly spaced, except for some representing more isolated areas on the outskirts.

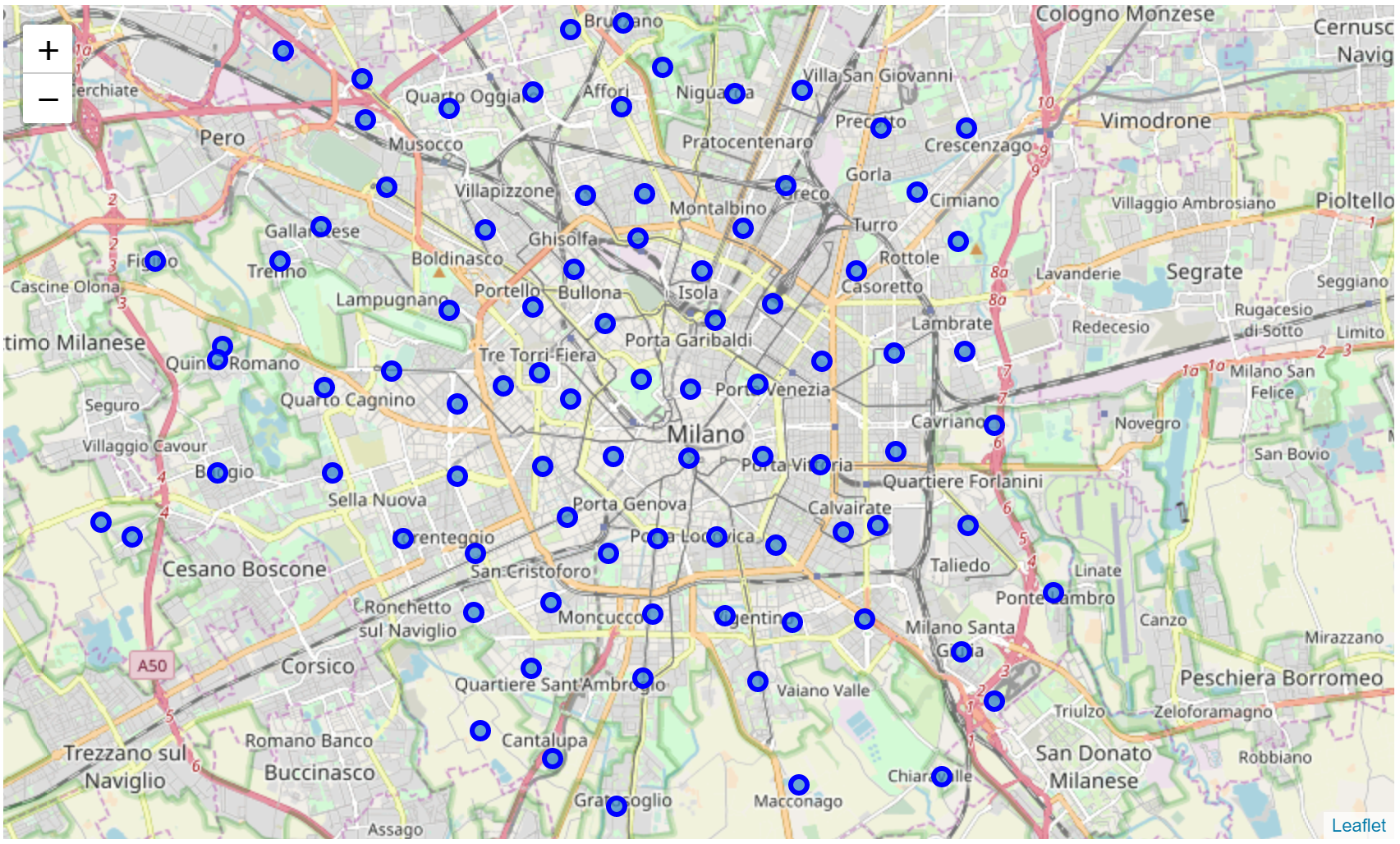


Figure 1 Map of Milan, with location of the neighbourhoods used for the analysis.

## Use of Foursquare API

I used the Foursquare API to acquire information on the venue composition of each area, similarly to what done in the previous parts of the Capstone course. A dataframe of 2370 venues was thus created (Figure 2).



Figure 2 First few lines of the popular venues dataframes obtained through Foursquare API.

I then transformed the dataframe in a “one-hot encoding” format, i.e. in which all the venue categories are extracted and set as dataframe columns, and the value in them represents the relative frequency (values 0-1) of that venue category among popular venues in the neighbourhood. As an example, if “Coffee shop” has value 0.1 for a given neighbourhood, it means that 10% of the popular venues in it are coffee shops. The analysis showed how there are 265 venue categories in the dataframe. This “one-hot encoded” dataframe will then be used for the clustering analysis.

Lastly, I made a dataframe listing the top 10 venue categories per neighbourhood, as shown in Figure 3. This will be used to interpret the results of clustering, and is very easily readable.



Figure 3 First few lines of the dataframe listing the most popular venues per area.

## Use of Yelp API

Using Yelp API is just as simple as simple as Foursquare’s. Complete instructions on its use can be found on <https://www.yelp.com/developers/documentation/v3/business_search> and <https://python.gotrained.com/yelp-fusion-api-tutorial/>. I used to obtain information about the most popular food venues around each coordinates point representing a neighbourhood, in terms of venue name, price tier and rating. With respect to Foursquare, Yelp is more customer-rating oriented and is thus more reliable for ratings-based analysis. Moreover, Yelp API has a more generous daily allowance of free calls per user, meaning that is easier to collect the required data, and the *json* text file is in my opinion easier to search for information.

The structure of the dataframe generated with information from Yelp is shown in Figure 4.



Figure 4 First few lines of the dataframe showing popular food venues per area, their price tier (column 4) and rating (column 5), both on a 1 to 5 scale.

After encoding the price tier information in *float* type numbers, I proceeded into grouping the information by borough, thus getting an average price tier and rating score per borough, plus a combined indicator named *heat*, as shown in the table in Figure 5. The *heat* parameter is in this case the simple sum of rating and price tier, but it can easily be tuned depending on the analysis needs.

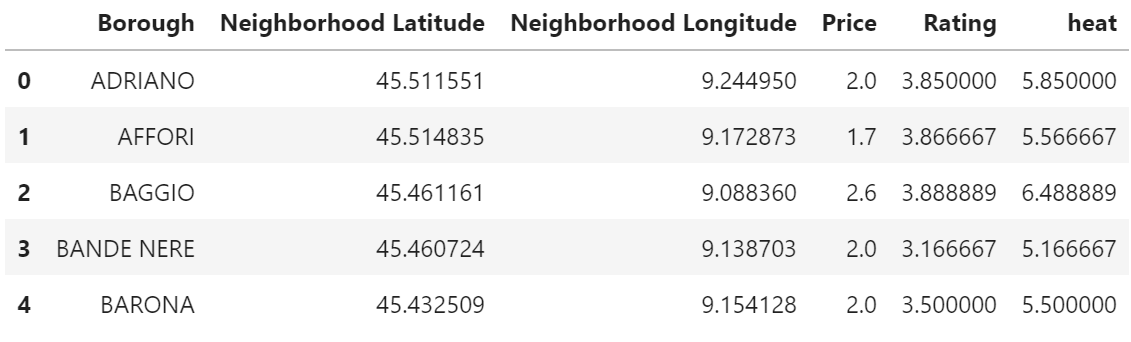


Figure 5 First few lines of the dataframe with average price tier and ratings per area, plus a combined indicator names “heat”.

# Model and clustering outcomes

## Clustering of Milan’s neighbourhoods, based on venue composition

The first analysis I did was to cluster Milan’s boroughs on the basis of the popular venues composition as acquired through Foursquare API. I used the scikit learn KMeans clustering algorithm, and tried different number of clusters to find the optimum results, i.e. a clustering with some significance and without excessive polarisation or a number of single occurrence clusters. I found that with 3 or fewer clusters basically all neighbourhoods are clustered together, whereas with more than 5 the number of single-member clusters increases. I went for 4 clusters, as it yields a neighbourhood division of some significance. The outcomes of the clustering are shown in Figure 6,.

A clear pattern emerges from the clustering. Note that using a different number of clusters doen not change general result as the main groups remain. Increasing the number of clusters generates very small, even unitary clusters which could be genuinely different from the others and might be valuable for other kinds of analysis.

* 0: mostly suburban or semi-rural location. Characterised by the presence of sport infrastructure.
* 1: contains all central and urban locations, plus others close to important transport links such as motorways and railways. Featuring primarly restaurants, cafes, pizza places and similar public venues.
* 2: outer locations, close to the outer ring motorway. Mainly hotels and restauration.
* 3: this cluster represents a very clear belt around town, mostly just outside the city inner ring road (Circonvallazione). Prevalence of pizza places, shops and services.

The most interesting point is the comparison between clusters 1 and 3, where we can see a slight different focus towards services, shops and fast food when getting out of the center towards more working class areas.

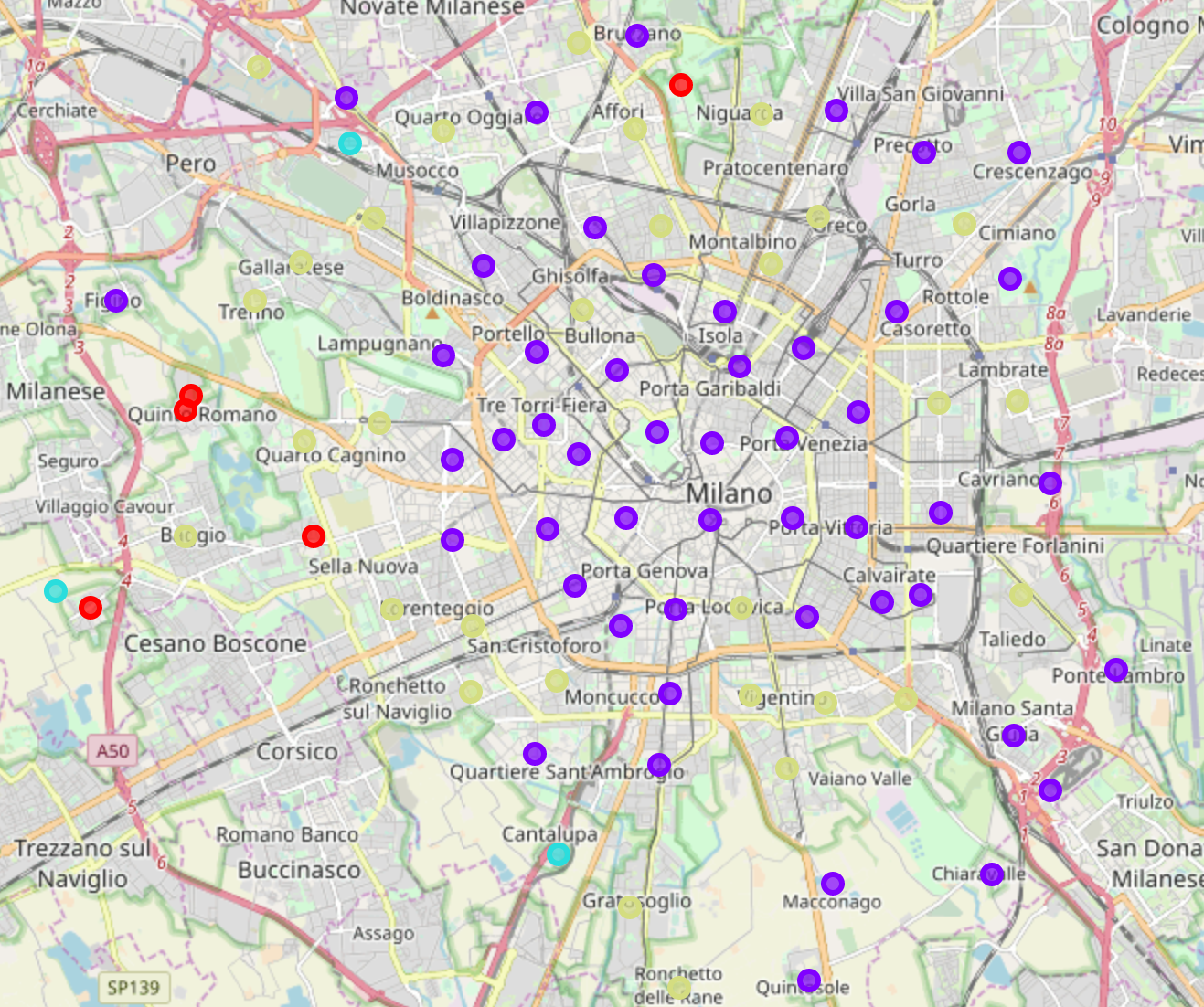


Figure 6

## Comparison to Toronto neighbourhoods

# Outcomes of the data analysis

## Analysis of the distribution of price tier and rating variables concerning food venues in Milan