Capstone Project - The Battle of The Neighborhoods

A. Introduction:

A.1. Background

According to Wikipedia, “Italy is the fifth most visited country in the world, with a total of 52.3 million international arrivals in 2016. Italy is well known for its cultural and environmental tourist routes and is home to 55 UNESCO World Heritage Sites, the most in the world. Rome […], Milan […], Venice, and Florence are also among the world's top 100 destinations.”

A.2. Problem Description:

The context of this Capstone project will be explained through the following scenario.

You just started to work as a Junior Data Analyst at a travel agency.

Since Italy will be a very sought travel destination after COVID-19, your manager asks you to propose a weekend travel itinerary for busy people.

Ideally, the itinerary should pack different experiences or cities for a weekend trip. Therefore, these cities should be quite different (from a venue's point of view) and quite close from a geographical point of view.

A.3. Objective

This report aims at proposing three cities for a weekend trip according to the criteria mentioned above. Therefore, we will proceed to study and analyze some provinces of the Veneto region, group them into similar clusters and analyze those clusters to gather meaningful information. That information can be used to find out three cities that are enough different (from a venue's point of view) and quite close from a geographical point of view.

A.4. Target Audience

This information provided by this report would be useful to your manager who will be able to create a weekend trip proposal in Italy.

B. Data Description:

To consider the objective stated above, we will use the following data sources:

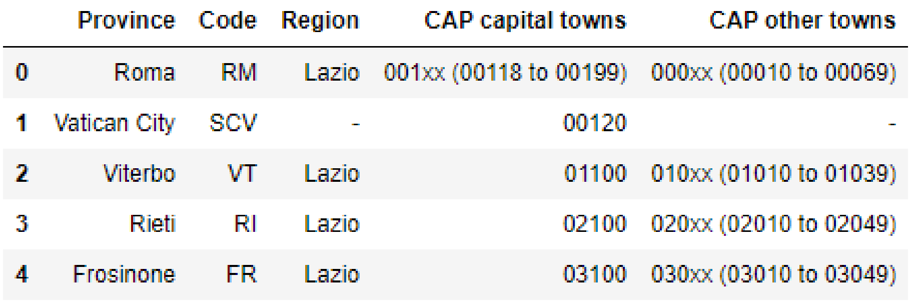
a) List of Italian provinces and Regions. The following Wikipedia page was scraped to pull out the necessary information: https://en.wikipedia.org/wiki/List_of_postal_codes_in_Italy. The information obtained i.e. the table of postal codes was transformed into a Pandas DataFrame for further analysis.

b) Coordinates data for each capital city in each region. The following csv lists the geographical coordinates of each city: https://simplemaps.com/static/data/country-cities/it/it.csv

C. Methodology:

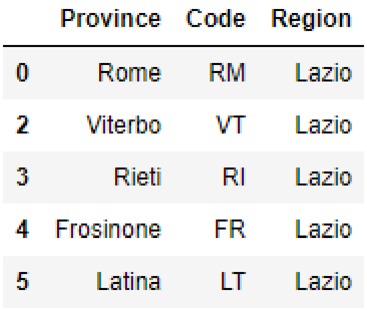
C.1. Importing data and creating Pandas DataFrames

To start with our analysis, we downloaded the Province data from the Wikipedia table into the Pandas DataFrame on the right. It shows the first five rows.



	Province	Code	Region	CAP capital towns	CAP other towns
0	Roma	RM	Lazio	001xx (00118 to 00199)	000xx (00010 to 00069)
1	Vatican City	SCV	-	00120	-
2	Viterbo	VT	Lazio	01100	010xx (01010 to 01039)
3	Rieti	RI	Lazio	02100	020xx (02010 to 02049)
4	Frosinone	FR	Lazio	03100	030xx (03010 to 03049)

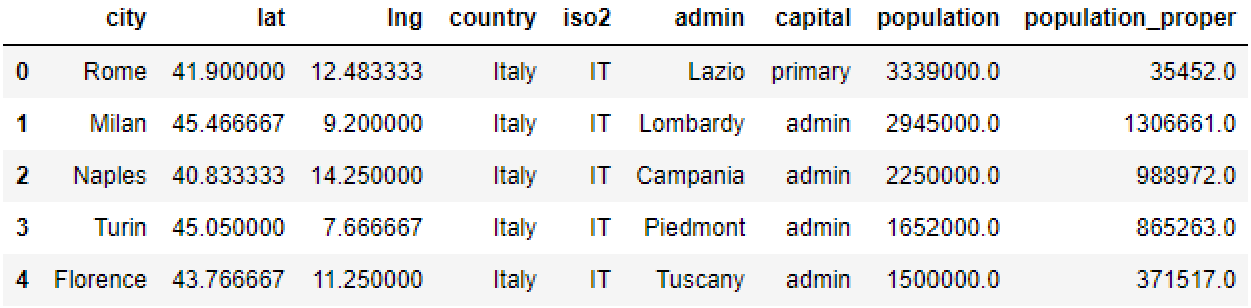
As we can see, the last two columns aren’t relevant to us. Also, some Region values are empty (Shown with - character), therefore we need to clean up the new Pandas DataFrame. On the right side, we report the first five rows of the DataFrame after the cleanup.



	Province	Code	Region
0	Rome	RM	Lazio
2	Viterbo	VT	Lazio
3	Rieti	RI	Lazio
4	Frosinone	FR	Lazio
5	Latina	LT	Lazio

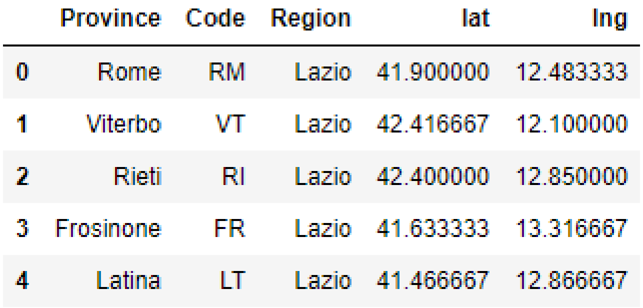
Since this data doesn’t show coordinates, we have to get them separately. The website Simplemaps.com offers coordinates information by city. This is fine because every Province takes its name from the capital city of that Province.

We downloaded a csv file from Simplemaps.com and imported it into a Pandas DataFrame. The first five rows are reported below.



	city	lat	lng	country	iso2	admin	capital	population	population_proper
0	Rome	41.900000	12.483333	Italy	IT	Lazio	primary	3339000.0	35452.0
1	Milan	45.466667	9.200000	Italy	IT	Lombardy	admin	2945000.0	1306661.0
2	Naples	40.833333	14.250000	Italy	IT	Campania	admin	2250000.0	988972.0
3	Turin	45.050000	7.666667	Italy	IT	Piedmont	admin	1652000.0	865263.0
4	Florence	43.766667	11.250000	Italy	IT	Tuscany	admin	1500000.0	371517.0

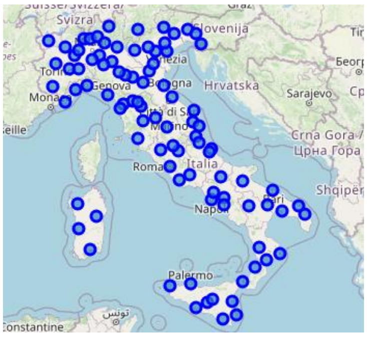
We are interested in the following columns: city, lat, lng. We clean up the remaining columns and merge the two DataFrames into the following one.



	Province	Code	Region	lat	lng
0	Rome	RM	Lazio	41.900000	12.483333
1	Viterbo	VT	Lazio	42.416667	12.100000
2	Rieti	RI	Lazio	42.400000	12.850000
3	Frosinone	FR	Lazio	41.633333	13.316667
4	Latina	LT	Lazio	41.466667	12.866667

This DataFrame shows most of the Italian Provinces (Province), the code of each Province (Code), the Region to which the Province belongs, the latitude and longitude of each value in the Province column.

C.2. Generating a map of Italy and moving towards Veneto



Initially, we generated a map of Italy to have an overview of where the Provinces are denser.

To do so, we used the python folium library. On the left side, we used the latitude and longitude values of each city/Province to superimpose a mark on the map.

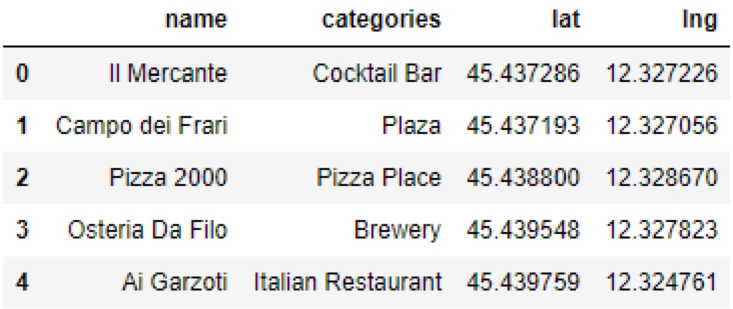
At first sight, it seems that the northern part shows a higher density. For this reason, we will focus on the Veneto region.

C.3. Utilizing Foursquare API to explore the Provinces in Veneto

As reported by Google, Veneto is a northeastern Italian region stretching from the Dolomite Mountains to the Adriatic Sea. Venice, its regional capital, is famed for its canals, Gothic architecture, and Carnival celebrations.

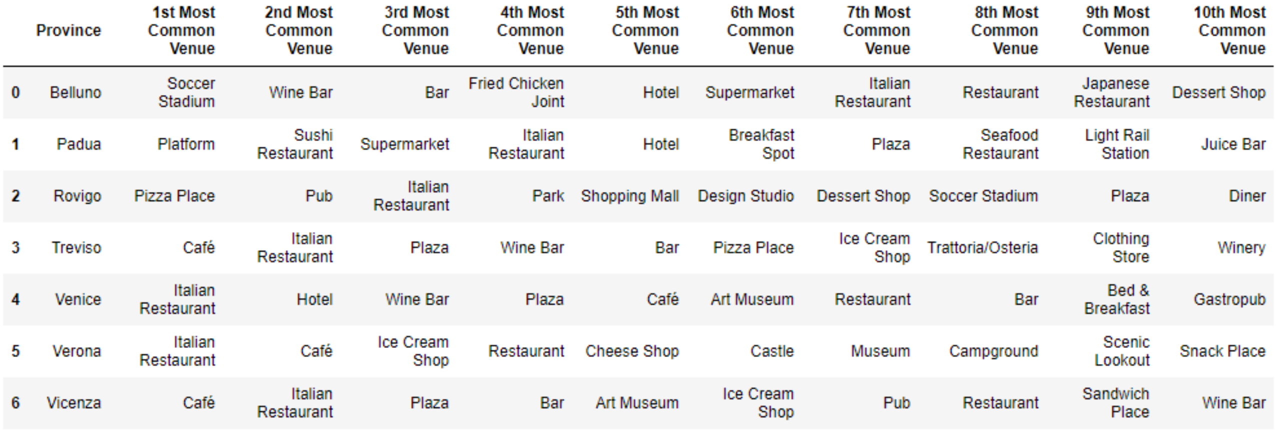
On the side, we show the location of the Veneto region.

We then used the Foursquare API to explore the Veneto Provinces and segment them. We set the LIMIT parameter to 100, which would limit the number of venues returned by the Foursquare API and the radius of 1000 meters. Here is the head of the list of Venues for the city of Venice.



	name	categories	lat	lng
0	Il Mercante	Cocktail Bar	45.437286	12.327226
1	Campo dei Frari	Plaza	45.437193	12.327056
2	Pizza 2000	Pizza Place	45.438800	12.328670
3	Osteria Da Filo	Brewery	45.439548	12.327823
4	Ai Garzoti	Italian Restaurant	45.439759	12.324761

We create a new function that will repeat the process above for all the City/Provinces in Veneto. This function will give us a list of the top 100 venues in the seven City/Provinces. Here is the outcome of this DataFrame.



	Province	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Belluno	Soccer Stadium	Wine Bar	Bar	Fried Chicken Joint	Hotel	Supermarket	Italian Restaurant	Restaurant	Japanese Restaurant	Dessert Shop
1	Padua	Platform	Sushi Restaurant	Supermarket	Italian Restaurant	Hotel	Breakfast Spot	Plaza	Seafood Restaurant	Light Rail Station	Juice Bar
2	Rovigo	Pizza Place	Pub	Italian Restaurant	Park	Shopping Mall	Design Studio	Dessert Shop	Soccer Stadium	Plaza	Diner
3	Treviso	Café	Italian Restaurant	Plaza	Wine Bar	Bar	Pizza Place	Ice Cream Shop	Trattoria/Osteria	Clothing Store	Winery
4	Venice	Italian Restaurant	Hotel	Wine Bar	Plaza	Café	Art Museum	Restaurant	Bar	Bed & Breakfast	Gastropub
5	Verona	Italian Restaurant	Café	Ice Cream Shop	Restaurant	Cheese Shop	Castle	Museum	Campground	Scenic Lookout	Snack Place
6	Vicenza	Café	Italian Restaurant	Plaza	Bar	Art Museum	Ice Cream Shop	Pub	Restaurant	Sandwich Place	Wine Bar

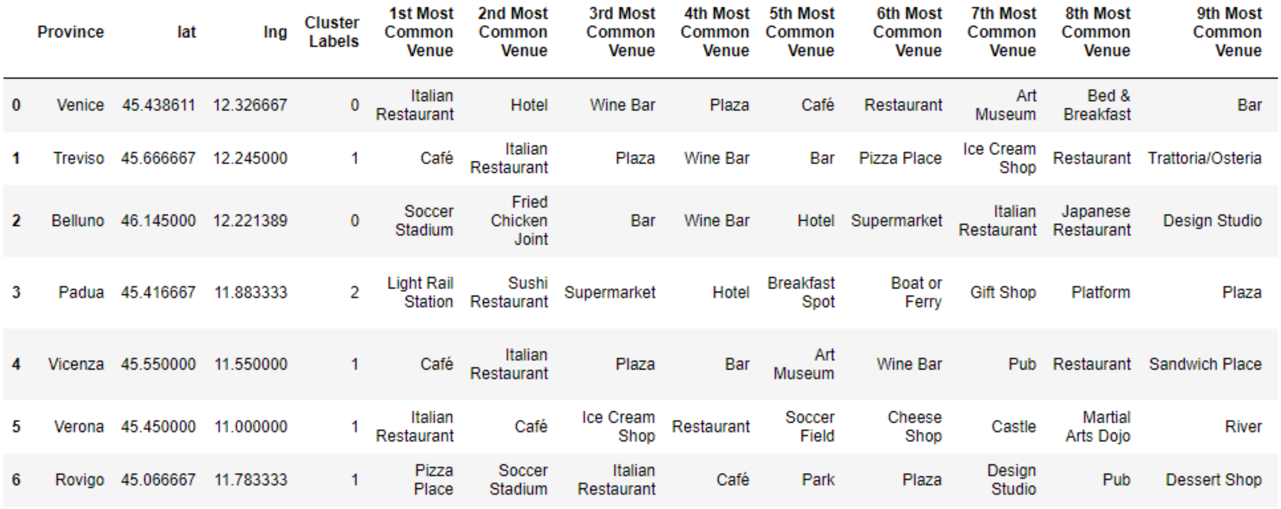
We have some common venue categories in the Provinces. We use the unsupervised learning K-means algorithm to cluster the Provinces. K-Means algorithm is one of the most common methods for clustering in unsupervised learning.

We use a k_cluster value of 3 to split the City/Provinces into 3 different clusters based on the similarity among their venues.

D. Results:

D.1. Adding the Cluster Labels to the Venue Data

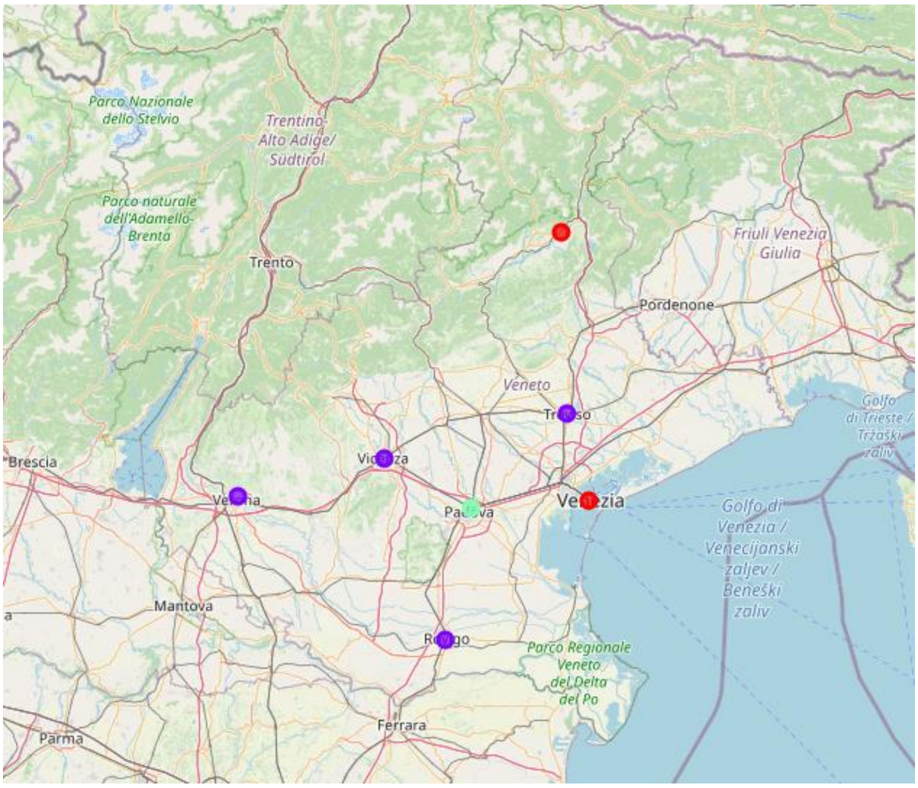
The below table depicts the clustered data along with the top 10 most common venues in that cluster. The 10th column is not visible in the table.



	Province	lat	lng	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Venice	45.438611	12.326667	0	Italian Restaurant	Hotel	Wine Bar	Plaza	Café	Restaurant	Art Museum	Bed & Breakfast	Bar
1	Treviso	45.666667	12.245000	1	Café	Italian Restaurant	Plaza	Wine Bar	Bar	Pizza Place	Ice Cream Shop	Restaurant	Trattoria/Osteria
2	Belluno	46.145000	12.221389	0	Soccer Stadium	Fried Chicken Joint	Bar	Wine Bar	Hotel	Supermarket	Italian Restaurant	Japanese Restaurant	Design Studio
3	Padua	45.416667	11.883333	2	Light Rail Station	Sushi Restaurant	Supermarket	Hotel	Breakfast Spot	Boat or Ferry	Gift Shop	Platform	Plaza
4	Vicenza	45.550000	11.550000	1	Café	Italian Restaurant	Plaza	Bar	Art Museum	Wine Bar	Pub	Restaurant	Sandwich Place
5	Verona	45.450000	11.000000	1	Italian Restaurant	Café	Ice Cream Shop	Restaurant	Soccer Field	Cheese Shop	Castle	Martial Arts Dojo	River
6	Rovigo	45.066667	11.783333	1	Pizza Place	Soccer Stadium	Italian Restaurant	Café	Park	Plaza	Design Studio	Pub	Dessert Shop

D.2. Visualizing the resulting Clusters

We use the matplotlib and folium packages to visualize the clusters on the Veneto map.


Parco Nazionale dello Stelvio
Trentino-Alto Adige/Südtirol
Parco naturale dell'Adamello-Brenta
Trento
Friuli Venezia Giulia
Pordenone
Veneto
Treviso
Golfo di Trieste / Tržaški zaliv
Brescia
Vicenza
Verona
Padova
Venezia
Golfo di Venezia / Venecijanski zaljev / Beneški zaliv
Mantova
Rovigo
Parco Regionale Veneto del Delta del Po
Ferrara
Parma

E. Discussion:

We carried out this analysis with the intent to show different cities that could be visited during a weekend trip. The initial request was to find out cities that should be quite different (from a venue's point of view) and quite close from a geographical point of view.

Looking at the table and map presented above we can see how cities/Provinces are clustered:

- Cluster 0: Venice, Belluno
- Cluster 1: Verona, Vicenza, Treviso, Rovigo
- Cluster 2: Padua

We want to visit a city for each cluster. Since Cluster 2 has only one city, Padua, this could be a city we want to visit.

In cluster 0 there are two cities, Venice and Belluno. However, Venice is much closer to Padua, therefore Venice becomes the second city on our weekend trip.

In cluster 1, we have four cities, but Treviso is the nearest to both Padua and Venice, therefore Treviso is going to be the third city on our weekend trip.

F. Conclusion:

Despite the outcome, using a different k_cluster value can show a slightly different result. As seen in the example above, data was used to cluster cities in Veneto based on the most common venues in those cities. Similarly, it could be interesting to compare clusters in different regions or countries.

G. References:

- Venice Grand Canal image: https://pixabay.com/photos/grand-canal-venice-italy-canal-918699/
- Wikipedia content: https://en.wikipedia.org/wiki/List_of_postal_codes_in_Italy
- Simplemaps.com city coordinates**:** https://simplemaps.com/static/data/country-cities/it/it.csv
- Location of Veneto region image: https://en.wikipedia.org/wiki/File:Veneto_in_Italy.svg
- Foursquare API