IBM Data Science Certification - Capstone Report

Brianna Shannon

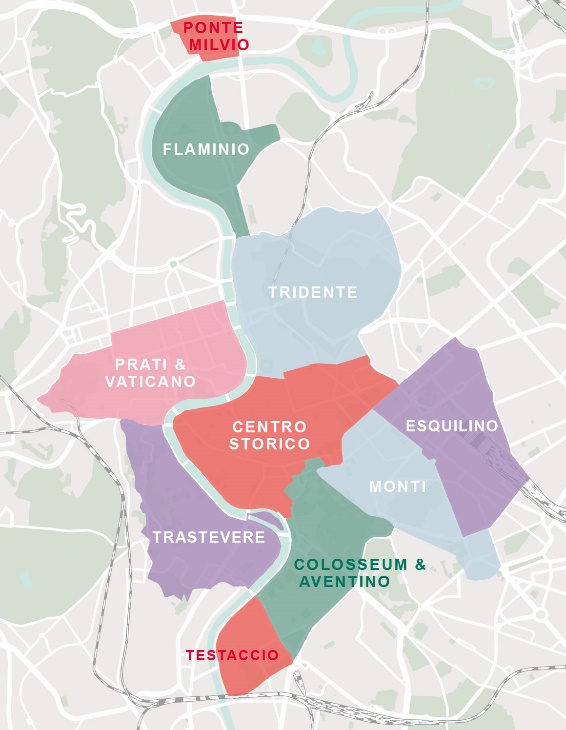
Dec. 2020

**Introduction:**

Rome is the second most visited city in Europe after Paris for tourists. Because Rome has such a deep, rich history dating back to 735 BC, there’s a lot to see and do. Ancient Rome was setup on the seven hills of Rome, which segmented much of Romans’ activities and interactions, so many tourist attractions are also segmented into distinct areas. Because of this segmentation and the number of things to see, the goal of this assignment is to cluster neighborhoods by tourist attractions, food, and nightlife to help tourists maximize their visit. By doing so, these clusters will help tourists plan out which attractions to see on the same day and where to eat on the same day.

**Data:**

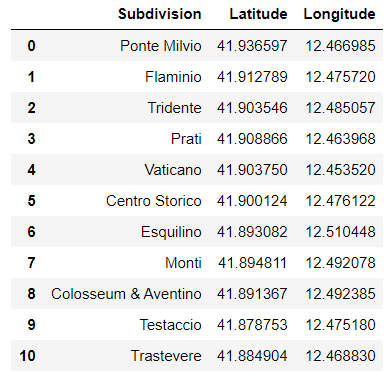
The subdivisions of Rome will be pulled from the following link: <https://www.timetomomo.com/en/destination/rome/>. The subdivisions are not in a table but rather a picture, which is shown below in Figure 1.



**Figure 1: Subdivisions of Rome**

Using the list compiled from this image, the coordinates will be pulled from google maps. The data table with the coordinates of each subdivision in seen in Table 1.

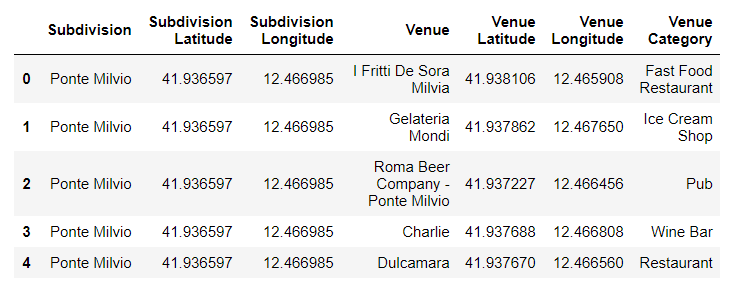
**Table 1: Subdivision Data**



**Methodology:**

Using this table, the Foursquare API will be used to pull all nearby attractions. The location data for each subdivision used to pull attractions using the Foursquare API. The API results will be combined with the subdivision to give Table 2.

**Table 2: Foursquare Data with Subdivision Data**



This data table contains nearly 700 rows. To check that the venue category only includes relevant tourist attracts, each unique venue category will be assessed. Figure 2 shows all the unique venue categories.



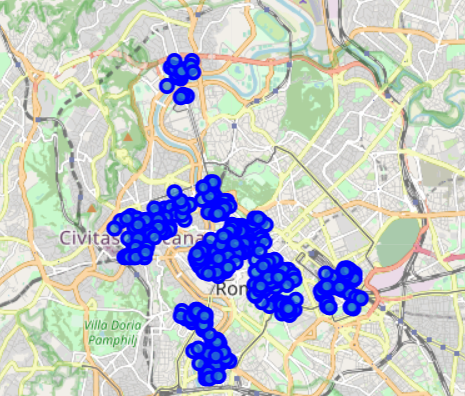
**Figure 2: Unique Venue Categories**

Some of the venue categories are not regular tourist categories, such as a cemetery. These unwanted categories will be relabeled and removed. Additionally, since the goal of this project is to consider common tourist attractions, food, and nightlife, these categories will be made and filled with the relevant venue categories in Figure 2. After doing this, the cleaned venue categories are seen in Figure 3. By getting rid of unwanted categories, the data set has decreased from nearly 700 rows to about 570 rows.



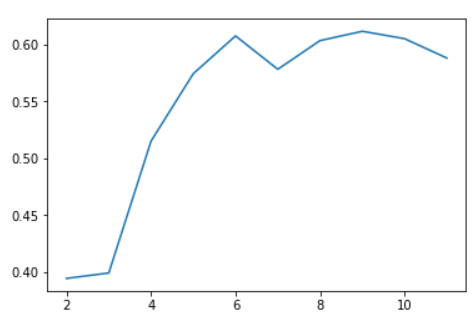
**Figure 3: Relevant Venue Categories**

Before clustering the data, the map of all the locations is seen in Figure 4. Just by visually looking at the map, its clear there are already some distinct clusters.



**Figure 4: Unclustered Map**

Now that all the data is pulled from Foursquare and is cleaned, k-means will be used to cluster the data. Since the goal of this project is to optimize tourists visits so they can visit attractions near one another, and therefore decrease overall traveling time, the k-means clustering will be done on the venue coordinates. The silhouette score will be used to find the optimal value. Since there 11 subdivisions, the maximum k value will be 11. The silhouette score was found for every k value from 2 to 11. Then, the data was plotted, and the maximum silhouette score is the optimal k value. The plot is shown in Figure 5. Using the k value of 9, the K-means model was ran and clusters were assigned.

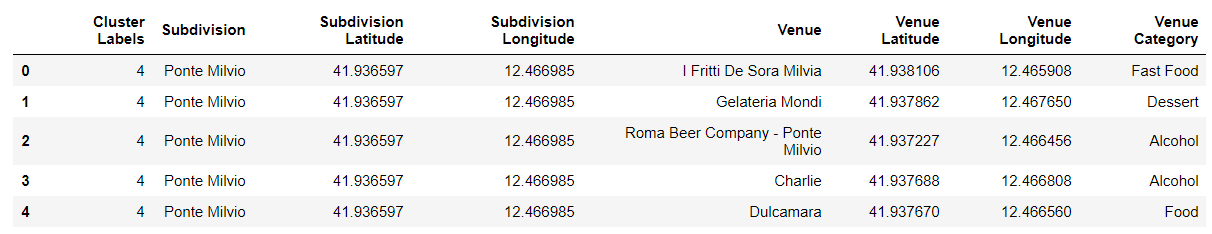


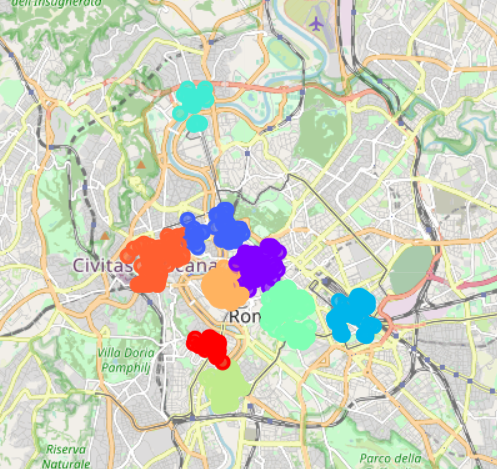
**Figure 5: Silhouette Plot**

**Results:**

As stated earlier, a k-means classification model ran with a k value of 9. A cluster label was then assigned to every row for every venue as shown in Table 3. This data was then plotted on a map as shown in Figure 6.

**Table 3: K-Means Output Table**





**Figure 6: Clustered Map**

**Discussion:**

Since there are only 9 clusters and 11 subdivisions, some of subdivisions were lumped together into the same cluster. For example, Prati was combined with either Vaticano or Flaminio. Also, some locations in Centro Storico were clustered with the Tridente subdivision. Centre Storico had the most attractions of any subdivision, so it makes sense that it had its own cluster and had locations in a second cluster. Recommendations to improve this model would be to remove certain categories. This would help make the model more customizable. For example, a family with young kids would most likely not want to visit nightclubs or bars, so the category of alcohol could be removed for them. Furthermore, the current model does not consider price. Some tourists may be on a tighter budget, so some venues may be too expensive. An added feature could be a price filter that removes anything above or below a given budget.

**Conclusion:**

This project successfully clustered tourist attractions (food, nightlife, and tourist attractions) in Rome. By having these nine distinct clusters, tourists can optimize their trip to the city by better planning their day to include venues nearby in order to reduce traveling time. Therefore, it is recommended that tourists visit one cluster each day if visiting for 9 or more days or visit adjacent clusters on the same day if visiting for less than 9 days.