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.

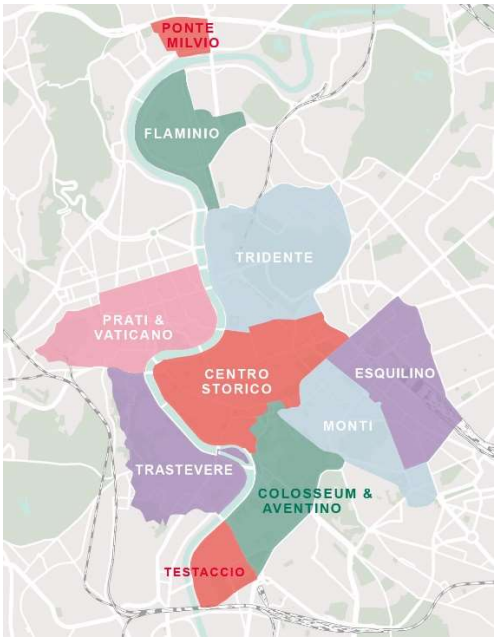

Figure 1: Subdivisions of Rome

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

	Subdivision	Latitude	Longitude
0	Ponte Milvio	41.936597	12.466985
1	Flaminio	41.912789	12.475720
2	Tridente	41.903546	12.485057
3	Prati	41.908866	12.463968
4	Vaticano	41.903750	12.453520
5	Centro Storico	41.900124	12.476122
6	Esquilino	41.893082	12.510448
7	Monti	41.894811	12.492078
8	Colosseum & Aventino	41.891367	12.492385
9	Testaccio	41.878753	12.475180
10	Trastevere	41.884904	12.468830

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

	Subdivision	Subdivision Latitude	Subdivision Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ponte Milvio	41.936597	12.466985	I Fritti De Sora Milvia	41.938106	12.465908	Fast Food Restaurant
1	Ponte Milvio	41.936597	12.466985	Gelateria Mondi	41.937862	12.467650	Ice Cream Shop
2	Ponte Milvio	41.936597	12.466985	Roma Beer Company - Ponte Milvio	41.937227	12.466456	Pub
3	Ponte Milvio	41.936597	12.466985	Charlie	41.937688	12.466808	Wine Bar
4	Ponte Milvio	41.936597	12.466985	Dulcamara	41.937670	12.466560	Restaurant

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.

```
['Abruzzo Restaurant', 'Accessories Store', 'American Restaurant', 'Argentinian Restaurant', 'Art Gallery', 'Art Museum', 'A
rts & Crafts Store', 'Athletics & Sports', 'Bakery', 'Bar', 'Bed & Breakfast', 'Beer Bar', 'Beer Garden', 'Beer Store', 'Bis
tro', 'Bookstore', 'Boutique', 'Brazilian Restaurant', 'Breakfast Spot', 'Brewery', 'Bubble Tea Shop', 'Burger Joint', 'Caf
é', 'Cemetery', 'Chinese Restaurant', 'Church', 'Clothing Store', 'Cocktail Bar', 'Coffee Shop', 'Comfort Food Restaurant',
'Comic Shop', 'Concert Hall', 'Convenience Store', 'Cosmetics Shop', 'Cupcake Shop', 'Deli / Bodega', 'Department Store', 'D
essert Shop', 'Diner', 'Event Space', 'Farmers Market', 'Fast Food Restaurant', 'Filipino Restaurant', 'Flea Market', 'Foo
d', 'Food & Drink Shop', 'Fountain', 'French Restaurant', 'Furniture / Home Store', 'Garden', 'Gastropub', 'German Restauran
t', 'Gift Shop', 'Gluten-free Restaurant', 'Gourmet Shop', 'Greek Restaurant', 'Gym', 'Historic Site', 'History Museum', 'Ho
stel', 'Hotel', 'Hotel Bar', 'Ice Cream Shop', 'Indian Restaurant', 'Indie Movie Theater', 'Irish Pub', 'Italian Restauran
t', 'Japanese Restaurant', 'Jazz Club', 'Jewelry Store', 'Juice Bar', 'Lingerie Store', 'Lounge', 'Market', 'Mediterranean R
estaurant', 'Mexican Restaurant', 'Monument / Landmark', 'Multiplex', 'Museum', 'Music Venue', 'Nightclub', 'Noodle House',
'Other Nightlife', 'Outdoors & Recreation', 'Park', 'Pastry Shop', 'Performing Arts Venue', 'Pet Store', 'Pie Shop', 'Pizza
Place', 'Plaza', 'Pub', 'Public Art', 'Racetrack', 'Restaurant', 'Road', 'Rock Club', 'Roman Restaurant', 'Roof Deck', 'Sala
d Place', 'Sandwich Place', 'Scenic Lookout', 'Sculpture Garden', 'Seafood Restaurant', 'Shopping Mall', 'Spanish Restauran
t', 'Sporting Goods Shop', 'Steakhouse', 'Supermarket', 'Sushi Restaurant', 'Taco Place', 'Temple', 'Theater', 'Toy / Game S
tore', 'Train Station', 'Trattoria/Osteria', 'Vegetarian / Vegan Restaurant', 'Wine Bar', 'Wine Shop', "Women's Store"]
```

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.

```
['Fast Food' 'Dessert' 'Alcohol' 'Food' 'Outdoors' 'Museum' 'Store'
 'Religious Site' 'Art']
```

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.

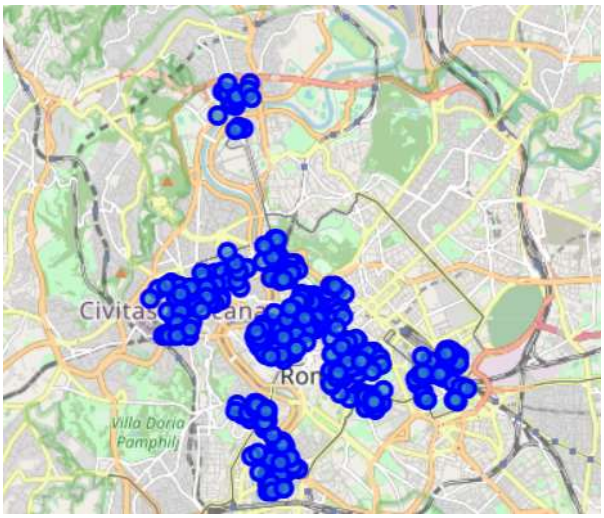

Figure 4: Unclustered Map

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.

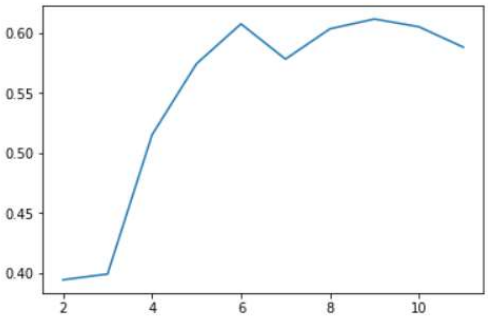

Figure 5: Silhouette Plot

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

	Cluster Labels	Subdivision	Subdivision Latitude	Subdivision Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	4	Ponte Milvio	41.936597	12.466985	I Fritti De Sora Milvia	41.938106	12.465908	Fast Food
1	4	Ponte Milvio	41.936597	12.466985	Gelateria Mondi	41.937862	12.467650	Dessert
2	4	Ponte Milvio	41.936597	12.466985	Roma Beer Company - Ponte Milvio	41.937227	12.466456	Alcohol
3	4	Ponte Milvio	41.936597	12.466985	Charlie	41.937688	12.466808	Alcohol
4	4	Ponte Milvio	41.936597	12.466985	Dulcamara	41.937670	12.466560	Food

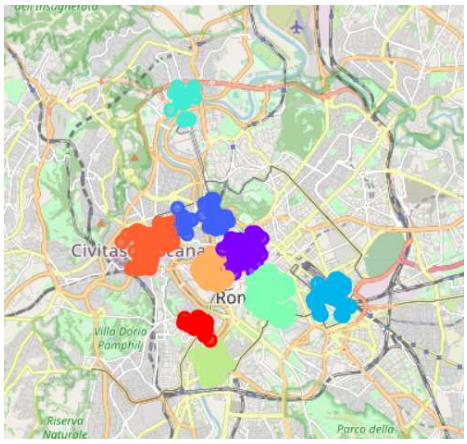

Figure 6: Clustered Map

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.