Data Science - Analysis of Metro Stations Areas in Panama City. Clustering and Classification

Ing. Diego Garcia **2020**

1. Introduction

1.1 Background

For the past 20 years, Panama has been at the top of the list of countries with the highest economic growth in the Latin American region. Panama, better known as the Global Trade Bridge has forged a logistics, commercial, and financial network promoted by the 'Panama Canal'. This has definitely been reflected in major public infrastructure projects such as the expansion of the Canal, the new terminal of Tocumen International Airport, and the construction of the first Metro line in Panama City. Indeed, the explosive growth of the Panama City, have pushed this need for a transportation system looking for an effective way to prompt the connection and mobility of Panamanians and tourist around the region. In 2014, the Metro became operational, helping then the supply of the public travel demand characteristic. The first line consists of 14 stations with 16 kilometers along. However, the second line went successfully into operation in the second quarter of 2019, making the rail system much more robust and allowing better reach and connectivity to many more districts. Sixteen more stations were added, covering 21 kilometers in the direction towards the east of the city.

1.2 Problem

Undoubtedly, the development of this Metro system is a long-term process, and it carries along with it ramifications of positive impacts, that could be analyzed in order to maximize the opportunities. One approach that I will expand here, is the clustering and segmentation of each metro station, to analyze a ratio of venues that characterize the transit of people around it. In this way, I could expect as a result, a classification of the main commercial activities in its surroundings and primary usage of each station. The information could be interest and useful in this way:

- Find places for new business development,
- Identify the urban structured to raise more strategies for the urban planification’s institutes,
- Relate the travel demand characteristic in each place,
- Boost the tourism activity by recognizing the principal venues in each station

2. Data Acquisition and Preparation

To better address this problem, I used the following data and sources:

- **List of stations and their coordinates**. For this intend used Google Maps to get the locations and the official website of the Metro of Panama https://www.elmetrodepanama.com/linea-1/, https://www.elmetrodepanama.com/linea-2/ to get the names of each station.
- **The Foursquare API**, to get the data of the most common venues surrounding each station.
- **Foursquare Main Categories data**, https://developer.foursquare.com/docs/build-with-foursquare/categories/. I scraped the link described previously, to get the main categories in which Foursquare classify their data of venues.

2.1.Description

As I mentioned above, the data of stations and coordinates were taken from the internet thanks to Google Map and the official website of the Metro of Panama. Because the information is not shown in a clear order, I arranged the data manually in a CSV file and placed it on my GitHub.

Once I set the data frame of the metro stations, I used the coordinates as input on the Foursquare API to get the data of the venues surrounding the metro station. Because Foursquare classifies its venues data in a series of categories and subcategories of venues, I preferred to rely firstly on the main categories of venues and for that, I needed to extract the data. It could be done by two methods, either by making a call to the API and making a request using the URL or by scraping their website. The second method indeed is a good way to practice scraping using the library BeautifulSoup and is not that complex. Nowadays scraping is a very demanding skill, and that's the reason that I chose it.

3. Methodology

Finally, I got the data and combined them into one frame. I proceeded to the cleaning and preparation process. Foursquare API offers a series of data of places according to the information registered by the users. That means that it might not take all of the actual venues existing around. If a venue is not registered it would be very difficult to consider it. One of the things that I noted, is that despite the input were the coordinates of the station, in some of them, foursquare recognized the station itself as a place surrounding the area. This data is redundant so I applied a filter to drop all the rows containing the stations as venues. Figure 1 shown below, is a preview of the data frame which I used initially, containing the important features. It shows only the first 5 rows.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Main Category
0	San Isidro	9.064168	-79.511222	CONWAY	9.065526	-79.514498	Department Store	Shop & Service
1	Los Andés	9.049284	-79.507917	Wendy's	9.049943	-79.507147	Fast Food Restaurant	Food
2	Los Andés	9.049284	-79.507917	El Campeón - Los Andes	9.050430	-79.507121	Clothing Store	Shop & Service
3	Los Andés	9.049284	-79.507917	McDonald's	9.050706	-79.509033	Fast Food Restaurant	Food
4	Los Andés	9.049284	-79.507917	Cinepolis	9.051966	-79.508367	Movie Theater	Arts & Entertainment

Figure 1. Main Data Frame

Figure 2 is a map of Panama City. The points represent the position of each station according to its coordinates. The route marked in red, represents the Line 1 of the metro, while the green one represents the route of stations in Line 2. This map was displayed by using the library ‘folium’.

Figure 2. Metro Line in Panama City

3.1. Exploratory Data Analysis

My first intention was to graph a chart in which we can compare the main categories that characterize the venues surrounding the Stations. For this, I needed to group all the venues using as the principal reference, the column ‘main category’. Also, I needed to add a column which I named ‘count’, that could help me to do the summary of the total venues of each category. The resulting data frame is shown in figure 3.

Main Category	count
Arts & Entertainment	12
Athletics & Sports	5
Food	210
Nightlife Spot	14
Outdoors & Recreation	17
Shop & Service	135
Travel & Transport	42

Figure 3. Data frame Venues classified by Main Category

For example, if we want to open a business, close to the station, we can have a general overview of what type of commercial locals are the most common. In figure 4, we can see the that top 3 type of venues corresponds to Food, Service & Shop and Travel & Transport. That means that from all venues registered in Foursquare and surrounding all the metro stations area, 48.28% are locals offering Food, 31.03% are in Shop and Service business, and 9.66% include mostly hotels venues, categorized as Travel & Transport.

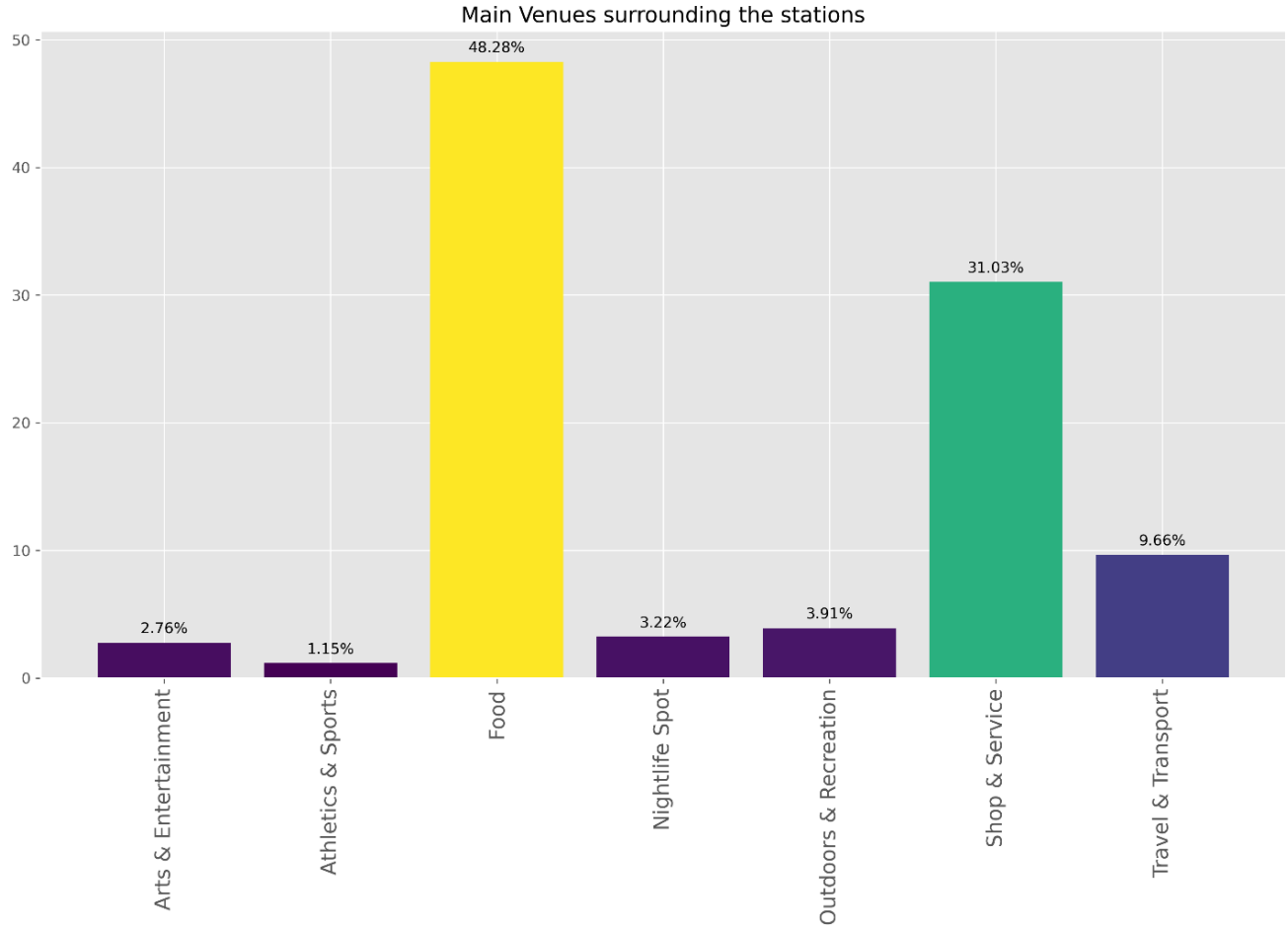

Figure 4. Main Venues surrounding the stations

Good, now let's say we want to know what are the most common type of places of food around the stations. For that I needed to filter the column 'main category' of the main data frame shown in figure 1, just to display those in the food category. After that, my main indicator was the column 'venue categories', which represent the subcategories of food places. In figure 5, it is shown that the most common type of local offering food, are the fast food restaurants, with around 25 venues in the proximity of the metro stations. While in the second place are restaurants offering a variety of menu with table service. They differ between some of them offering special dishes as an Italian restaurant and a Latin American Restaurant.

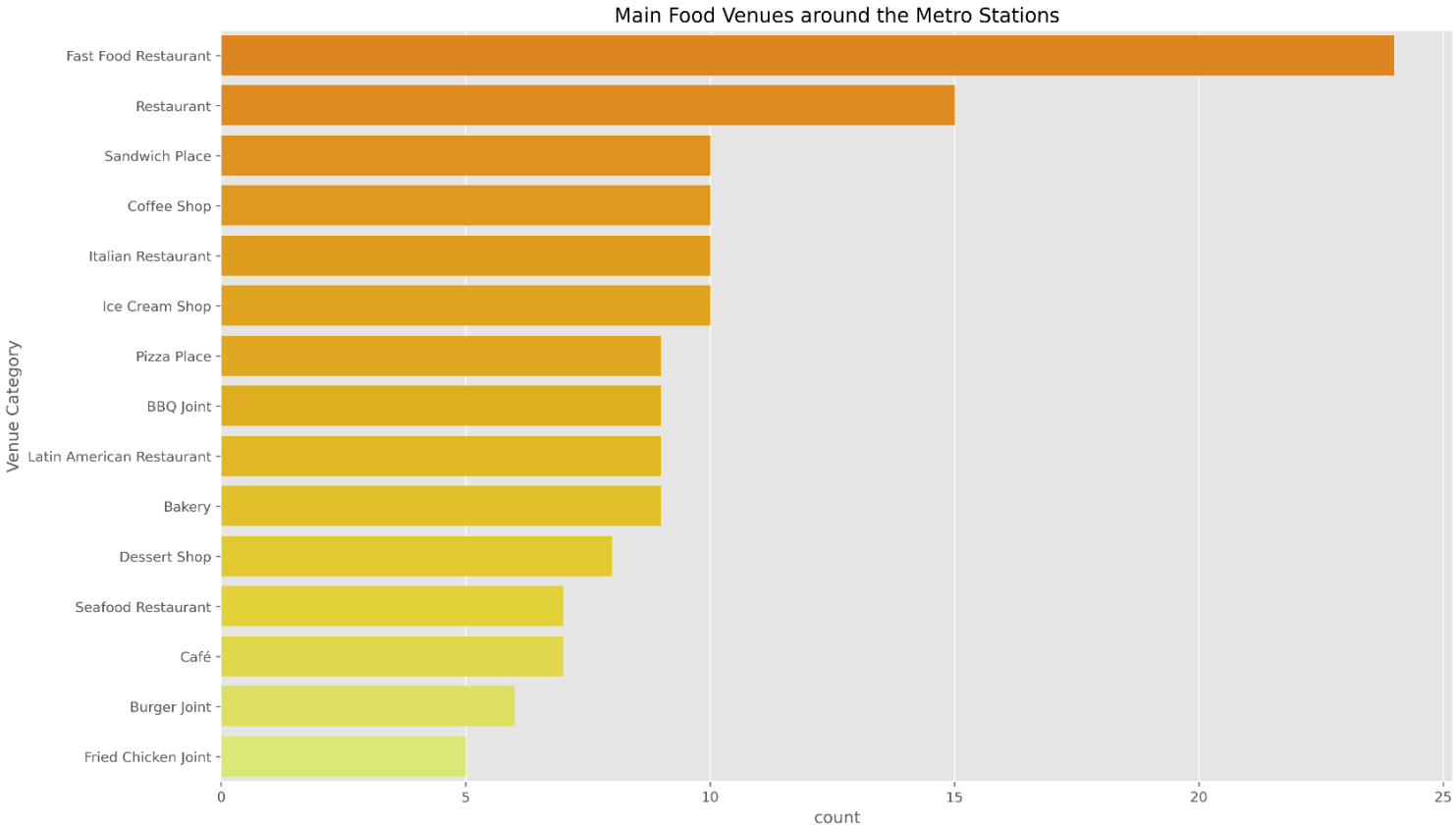

Figure 5. Main Food Venues around the Metro Stations.

I applied the same to the Shop and Service category, and it is shown in figure 6, that department stores and shopping malls are the most common type of businesses around the metro station areas. If we compare by quantity, we can conclude also that there are even more department stores than fast-food restaurants.

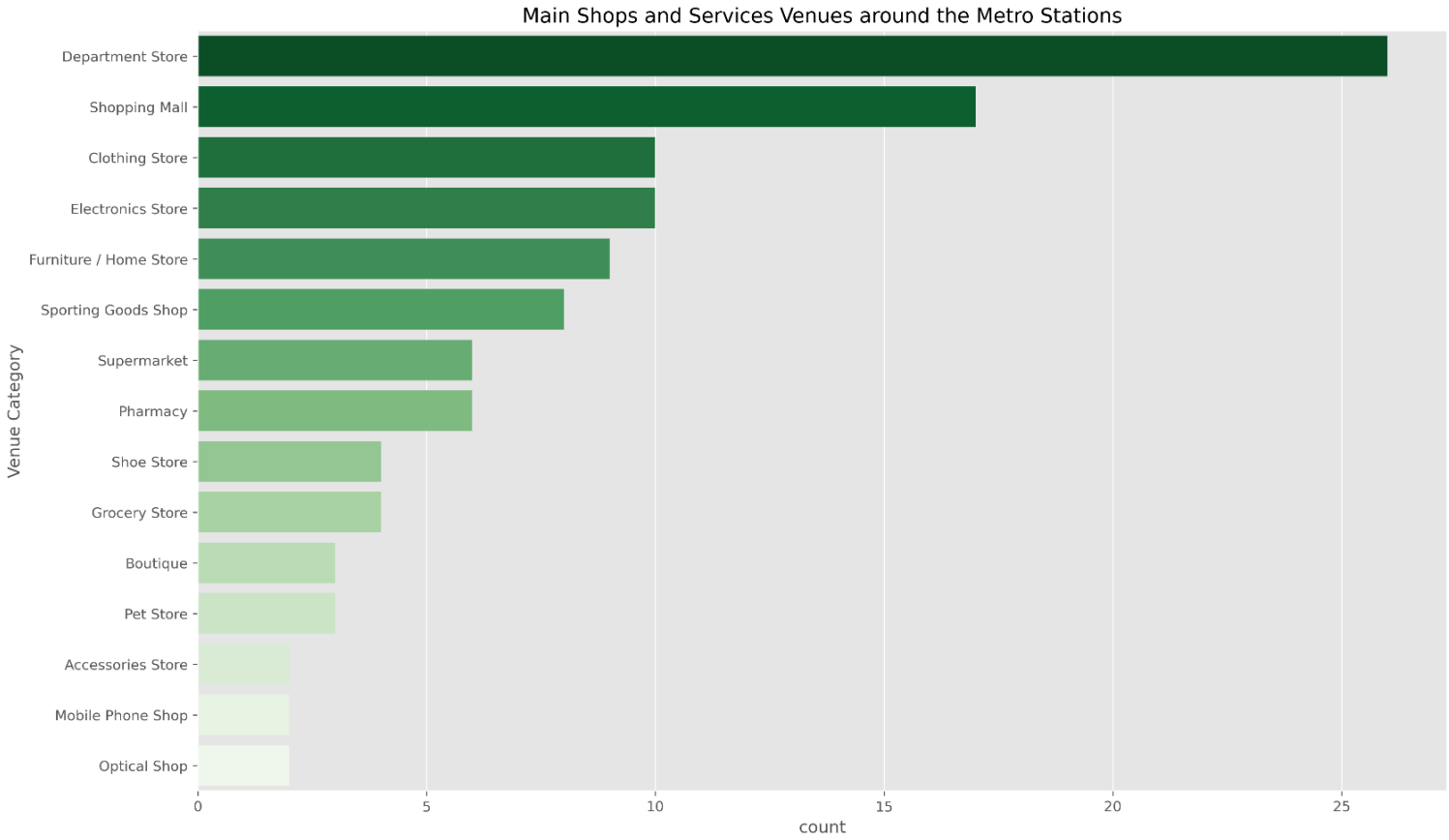

Figure 6. Main Shops and Service Venues around the Metro Stations

Ok, we saw the main types of venues surrounding the stations. Now we can make an analysis of each station, to know the top 10 most common venues around each one, and in this way, we can cluster them into groups. Clustering would help me divide the data into several groups, segregating the ones with similar traits. To get the top 10 most common venues in each station, I needed to rearrange the main data frame, by grouping the rows of the same station and by taking the mean of the frequency of occurrence of each category. Therefore this time the data frame is going to contain 28 rows, which corresponds to the number of stations, and 12 columns, 2 describing the name of station and neighborhood and, the other 10 describing the top most common venues. For example, in figure 7 we can see that in the San Isidro Station, the 1st most common venue are department stores and in the Via Argentina Station the 1st most common venues are hotels.

	Station Name	Neighborhood	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	Estación San Isidro	San Isidro	Department Store	Wings Joint	Diner	Disc Golf	Donut Shop	Electronics Store	Fast Food Restaurant	Fish & Chips Shop	Fish Market	Flea Market
1	Estación Los Andés	Los Andés	Department Store	Shoe Store	Clothing Store	Electronics Store	Fast Food Restaurant	Burger Joint	Fried Chicken Joint	Shopping Mall	Movie Theater	Pet Store
2	Estación Pan de Azúcar	Pan de Azúcar	Plaza	Park	Grocery Store	Beach	Hotel Pool	Food Truck	Disc Golf	Donut Shop	Electronics Store	Diner
3	Estación San Miguelito	Vía Simón Bolívar	Department Store	Sporting Goods Shop	Furniture / Home Store	Supermarket	Coffee Shop	Fast Food Restaurant	Wings Joint	Flea Market	Food Court	Flower Shop
4	Estación Pueblo Nuevo	Pueblo Nuevo	Dessert Shop	Department Store	Diner	Disc Golf	Donut Shop	Electronics Store	Fast Food Restaurant	Fish & Chips Shop	Fish Market	Flea Market
5	Estación 12 de Octubre	12 de Octubre	Food Truck	Restaurant	BBQ Joint	Martial Arts School	Fish & Chips Shop	Wings Joint	Fish Market	Food Court	Flower Shop	Flea Market
6	Estación El Ingenio	El Ingenio	Fast Food Restaurant	Paper / Office Supplies Store	Music Store	Rental Service	Latin American Restaurant	Bar	Furniture / Home Store	Seafood Restaurant	Flea Market	Fish Market
7	Estación Fernández de Córdoba	Vía Fernández de Córdoba	Furniture / Home Store	Ice Cream Shop	Latin American Restaurant	Sandwich Place	Chinese Restaurant	Gym	Breakfast Spot	BBQ Joint	Hostel	Food Court
8	Estación Vía Argentina	Vía Argentina	Hotel	Electronics Store	Italian Restaurant	Restaurant	Seafood Restaurant	Burger Joint	Pizza Place	Spa	Cuban Restaurant	Speakeasy
9	Estación Iglesia del Carmen	Vía España	Hotel	Italian Restaurant	Restaurant	Café	Coffee Shop	Spanish Restaurant	Vegetarian / Vegan Restaurant	Hotel Bar	French Restaurant	Food Truck
10	Estación Santo Tomas	Ave Justo Arosemena	Hotel	Bakery	Italian Restaurant	American Restaurant	Indian Restaurant	Hostel	Korean Restaurant	Bed & Breakfast	Scenic Lookout	Sandwich Place

Figure 7. Top 10 most common venue in each station

K-Means can help us group data based on the similarity of the stations to each other. In this way, we could better classify the stations according to the venues that are common between them. First, we have to determine the number of clusters or K in K-Means, for that, I used the elbow method. This graph is useful to compare the metric of accuracy and to select the best K that could suit the model. Let's take a look of figure 8.

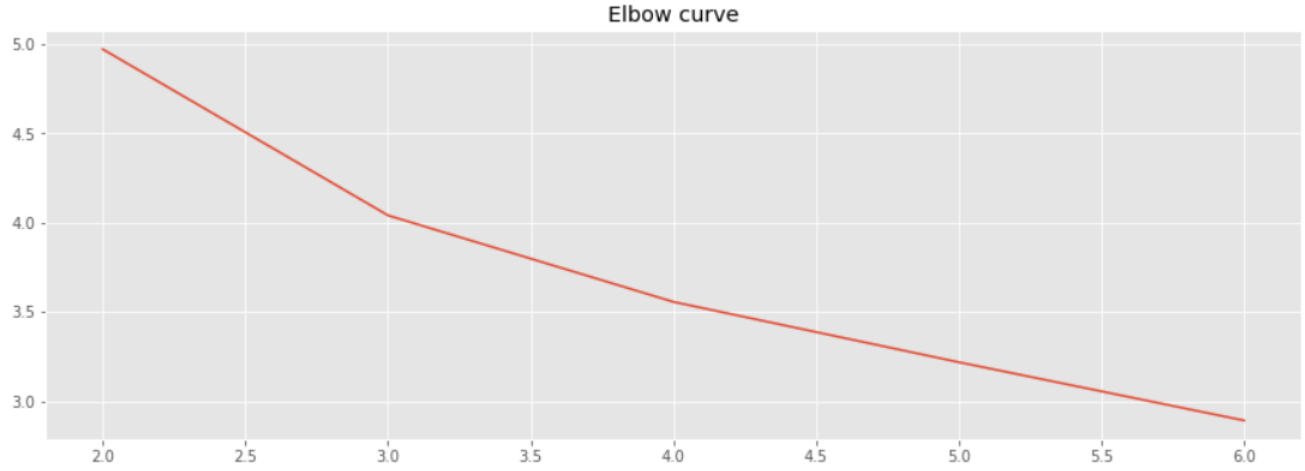

Figure 8. Elbow curve for K-Means

The inertia attribute of K-means identifies the sum of squared distances of samples to the nearest cluster center. For the samples, I used the mean of the frequency of occurrence of each category. That grade of distortion helps us identify that k=3 could best to option to run the model.

4. Results

Finally, I ran the analysis choosing k=3 as the number of clusters. The output needed from the algorithm was the corresponding cluster label of each station. For better insight, I displayed the map shown in figure 9. We can see that the stations were segregated into three groups, corresponding each one to different colors. We are going to call them Red Cluster, Purple Cluster, and Green Cluster. Also, it is notable that the cluster in purple color contains the major number of stations.

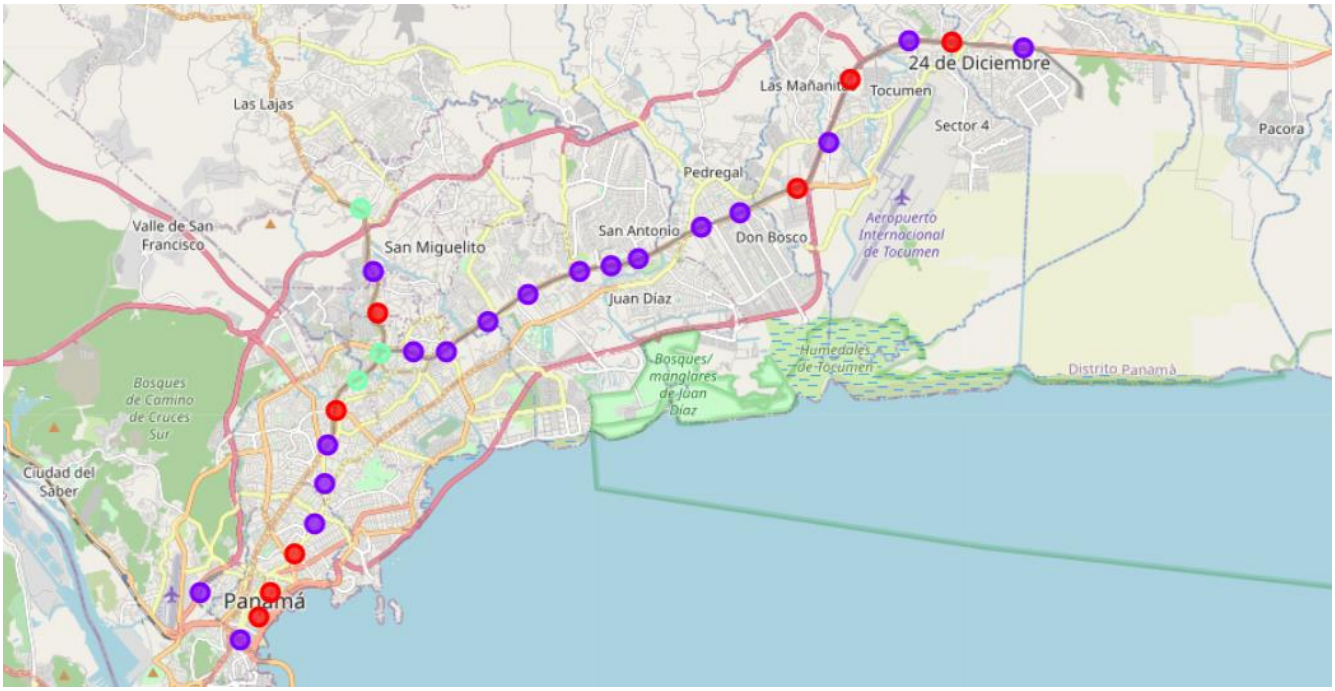

Figure 9. Clustering of the Metro Stations.

Perfect, now if we can ask ourselves, what are the common venues that characterize each cluster. Let's say if I'm looking for a station located close to hotels. I can get a frame of each cluster to compare the composition of the top 5 venues that are most common in each cluster. As I show below in figure 10, we can see the places that characterize the 3 clusters. The Red Cluster is composed mostly of hotel areas. Meaning that if I want to find a hotel in the city close to metro stations, is most probably that I'm going to find them in the area of Santo Tomas Station or Iglesia del Carmen Station, as is indicated in the map on the first 2 red labels in the city. Then we can see that the Purple Cluster is characterized mostly by fast food restaurants and shopping malls. The green one is composed mostly of department stores.

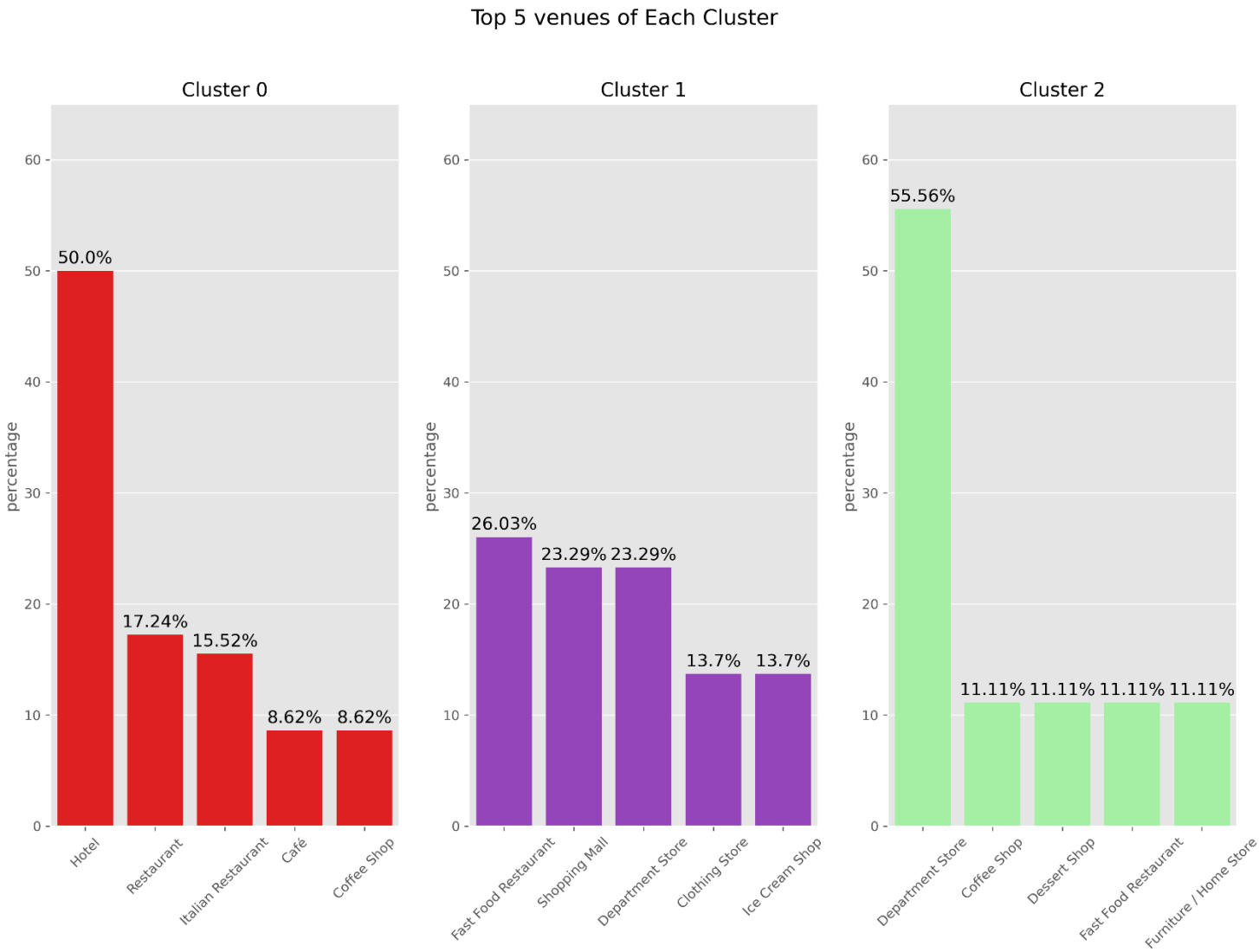

Figure 10. Top 5 venues of Each Cluster

5. Discussion

I can say that it has been a considerable approach and as I mentioned before, the model could have a certain scope due to the limitation of the data that we have. While Foursquare API might not take all of the actual venues existing around because it offers data of places limit by the information registered by the users, also Panama, as a growing country, should also consider working much more on its databases in general. For business investors for sure localization is one important variable to consider at the moment of opening a business, but certainly, it will be interesting to have more variable to study, like the average renting prices of each area, the crime rate of each area, and others more that could be considered in the future with a strong database.

6. Conclusion

As I mentioned initially, Panama is characterized for being a country strategically run by service. We saw through the exploratory analysis that there is a big number of commercial locals around these important areas, as it is the metro stations. Being the most common ones food, services, and shops. These important features to consider can be especially helpful in business and tourism purposes. Although this clustering method could be helpful also for tourism institutions to explore more the use of the station for foreigners, by identifying which area is more commercial for shopping, or in which area could be suitable to rent a place. Definitely, Data science would have a potential niche by integrating it into the future construction and planning of this long term mega project such as the metro in Panama. It has been a really interesting analysis and I know that Panama as a developing country is among the most compatible to enter the digital and big data competitiveness but there is hard work ahead.