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

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 push 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 station 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 ramification of positive impacts, that could be analyze in order to maximize the opportunities. One approach that I will expand here, is the clustering and segmentation of each metro station, for the purpose of analyze a ratio of venues that characterize the transit of people around it. In this way I could expect as a result, 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

To get a better approach to this problem, I will use the following data and sources:

- List of stations and its coordinates. For this intend I will use 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, which going to help us to get the data of the most common venues surrounding each station.
- Foursquare Categories, https://developer.foursquare.com/docs/build-with-foursquare/categories/. I will get the data by scraping the website using BS4 library

2.1. Description

As I mentioned above, the list of stations and coordinate will be taken from the internet thanks to Google Map and the official website of the Metro of Panama. Because the information isn't show in a clear order, I arranged the data manually in a csv file and placed it on my github.

Once I set a dataframe of the metro stations, I'm going to use it as input on the Foursquare API to get the data of the venues surrounding the metro station. Because Foursquare classify its venues data in a series of categories and subcategories of venues, we are going to rely firstly on the main categories of venues. For that we will need to extract the data. It could be done by two method, 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 demand skill, therefore I will choose that way.

The next step will be cleaning and preparation of data. The first intention is to graph a chart in which we can compare the main categories the characterize the venues surrounding the Stations.

For example, we want to know what is the most common type of local that offers food around the stations. After the we want to find the top 10 venues surrounding the station. For that we are going to rearrange the dataframe and in this case we are going to use the subcategories of the venues to be more specific. The goal is to make clusters of the station that share same similarities or characteristics. After the analysis we are going to be able to see the results and hopefully get some clear insights.