Clustering Neighborhoods in Mexico City's inner city

By: Luis Botello

1 Introduction

Mexico City is one of the most populous cities in the world, home to one of the largest metro systems in the world. The system served 1.655 billion passengers in 2019. However, Mexico City is also ranked as one of the most congested cities in the world. This project aims to better understand neighborhoods surrounding Mexico City's metro stations. The results from the analysis can serve city planners identify general movement of people, identify stations needing expansion, or identify areas for new stations.

2 Data Acquisition and Data Cleaning

2.1 Data Sources

Two data sources were utilized for this project. Data regarding metro stations (name, location, year-built) were scraped from Wikipedia. The Foursquare Place API was utilized to obtain data of the venues surrounding each metro station. Any venue within a 300 m. radius of the station was considered a venue of such station. The data points obtained for each venue are the following: Venue Name, Venue Category, Latitude, Longitude, Distance (m.) from Station.

2.2 Data Cleaning

The only data cleaning involved formatting the data into a useful data frame. Data from both sources were imported in json format, therefore, useful information had to be extracted utilizing a couple lines of code.

3 Exploratory Data Analysis

3.1 Visualizing Mexico City's Metro Stations

The first step was visualizing all metro stations on a map to get a better picture of how the metro system is laid out around Mexico City.

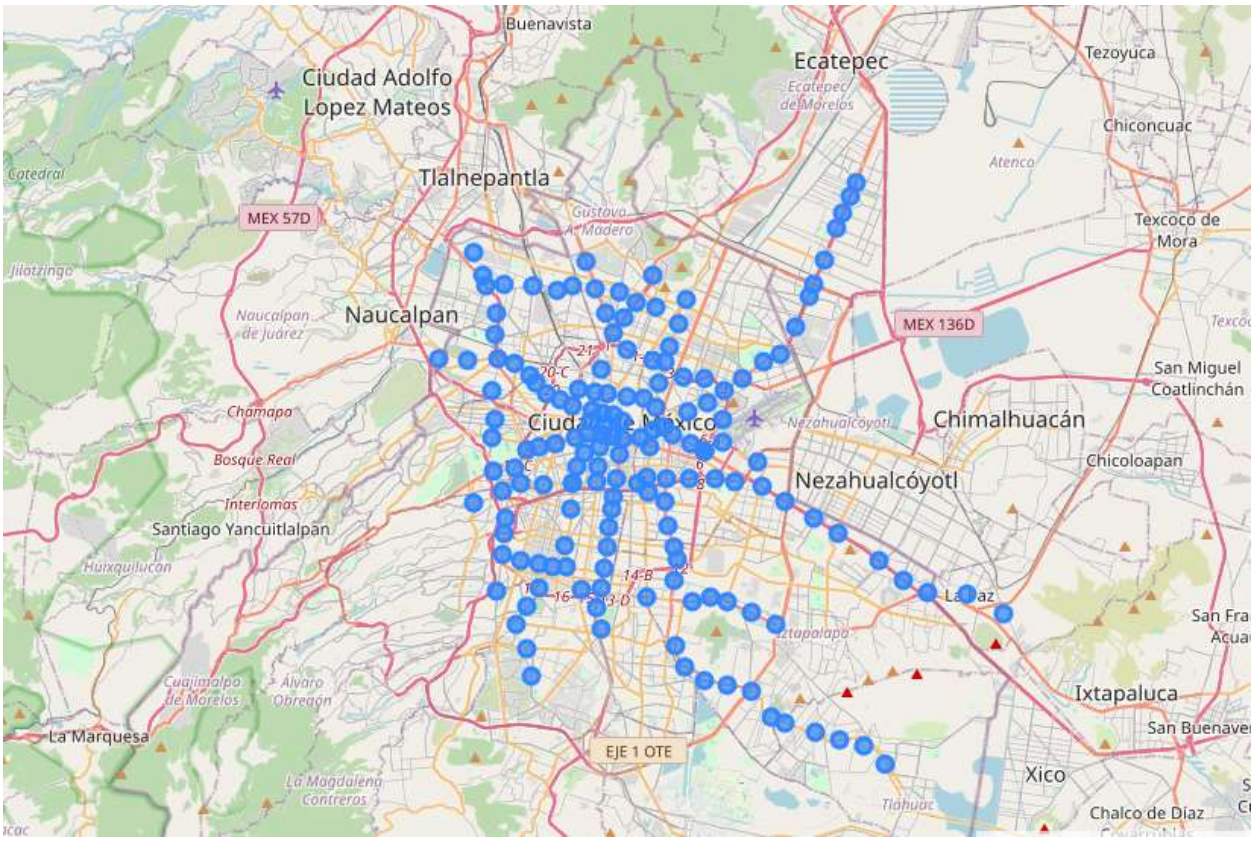

Figure 1

As it is expected, metro stations are more heavily concentrated on the city center. Furthermore, one can observe that the metro system does not cover the entirety of the city. From this observation and a quick google search, I discovered that Mexico City also possesses an extensive bus system to complement the metro system. A more in-depth analysis would require including the bus stations as well as to cover the entirety of the city. However, due to the scope of this project, only metro stations will be considered.

3.2 Most Common Venue Categories

To get a better picture of the city, I wanted to know which venue categories were the most common in the city. The results can be observed in the figure below.

	Category	No. of Venues
0	Mexican Restaurant	466
1	Taco Place	341
2	Bar	106
3	Bakery	98
4	Café	97
5	Coffee Shop	88
6	Convenience Store	84
7	Ice Cream Shop	77
8	Restaurant	74
9	Pizza Place	74

c

Figure 2

Not surprisingly, the most and second most common venue categories were Mexican Restaurant and Taco Place, respectively. One interesting thing to note, 8 out the 10 most common categories were food venues.

3.3 Stations with a Low Number of Venues Surrounding Them

Furthermore, I wanted to identify stations with a low number of venues surrounding them. The goal was to identify these stations to remove them from the analysis. Stations with such few venues is the equivalent of a record with few data, which would prevent our model from getting reliable results. Stations with less than five venues surrounding them can be observed below.

	Station	No. of Venues
0	Impulsora	1
1	Plaza Aragón	1
2	Martín Carrera	2
3	General Anaya	2
4	UAM-I	2
5	La Paz	3
6	Talismán	3
7	Viveros / Derechos Humanos	4
8	Politécnico	4
9	Nezahualcóyotl	4
10	Deportivo 18 de Marzo	4
11	Tacubaya	4

Figure 3

A total of 34 venues were removed from the dataset that was used for our model. The initial data set contained a total of 3,327 venues, meaning that we removed around 1% of the total venues initially identified.

4 Predictive Modeling

I applied K-Means clustering with 3 and 4 clusters. The clustering was performed based on the number of venues in each category each metro station had surrounding them. The dataset contained a total of 176 different venue categories, from which the clustering was performed. Note, that stations with less than five venues were not included in the analysis. However, they were later added to the visualization (in black) to keep track of them. The figure below depicts the results utilizing 3 clusters.

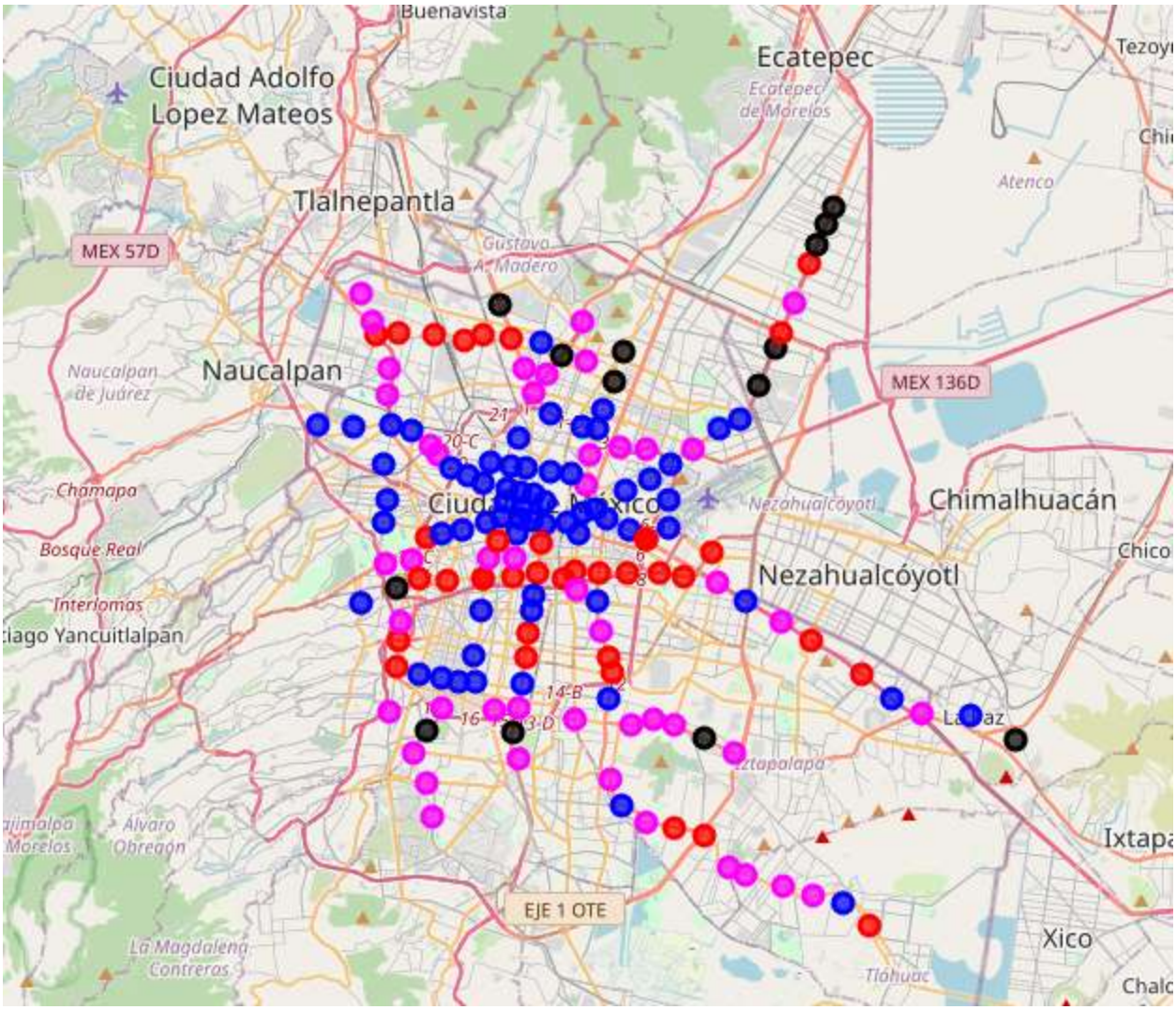

Figure 4

From the figure above, we can observe that there is a concentration of blue stations in the city center. Red and magenta stations seem to be more randomly distributed, however, there seems to be two corridors of red stations running in the east-west direction. One corridor is in the north and the other one south of the city center. Black stations, stations with less than five venues, seem to be in the outskirts of the city for the most part, which would suggest that those are residential or less urbanized areas. Results using 4 clusters can be observed in the figure below.

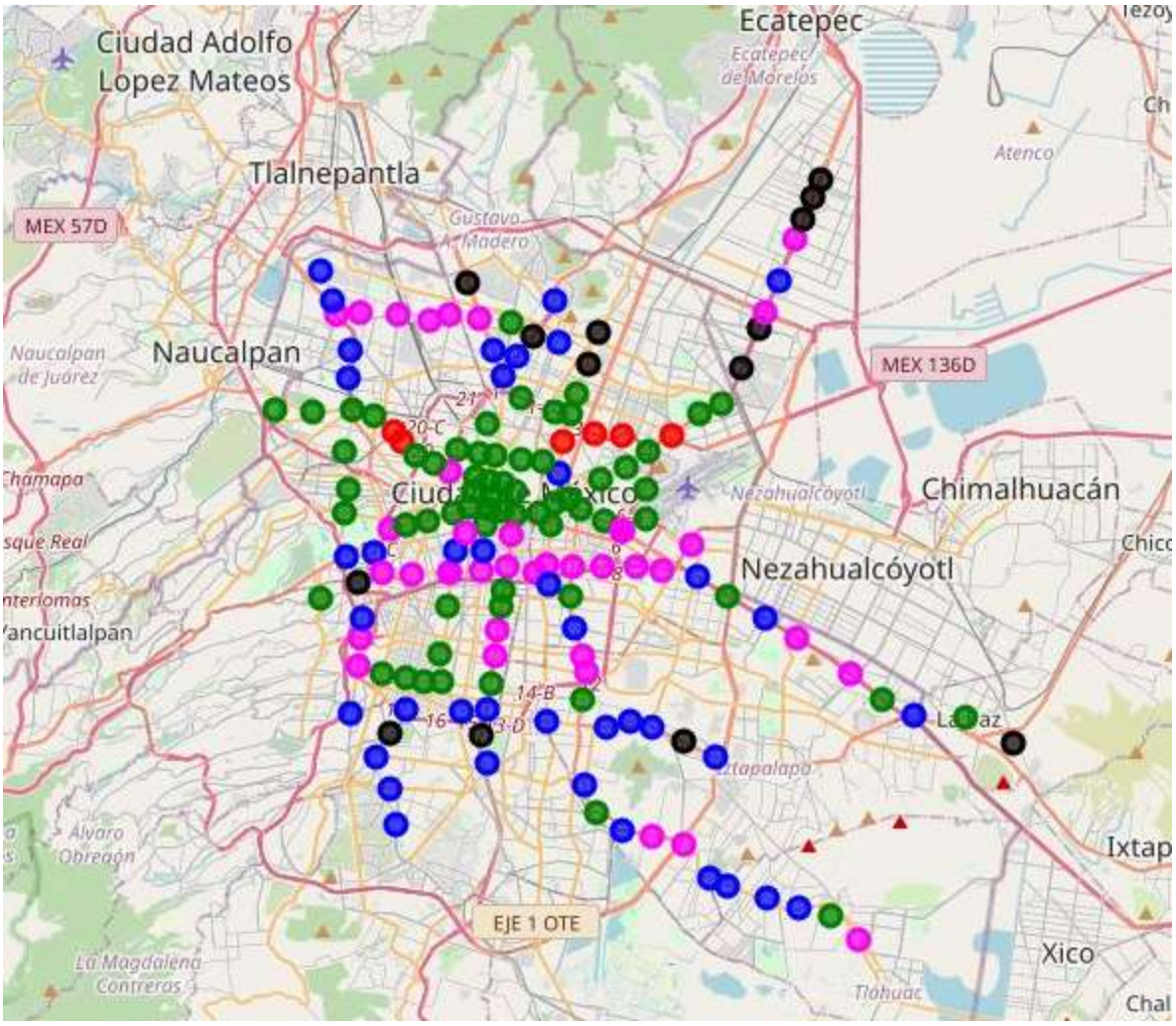

Figure 5

K-Means with four clusters does not seem to provide further insights. When compared to the 3 clusters model, they look awfully similar. The only difference being the new cluster (4) in red, however, there are only 5 stations classified as red.

I wanted to dig deeper into the 3-clusters model (Figure 4) so I compiled a table for each cluster containing the ten most common venue categories, which can be observed below.

Cluster 0	
Category	No. of Venues
Park	41
Convenience Store	38
Bakery	37
Mexican Restaurant	35
Taco Place	31
Café	25
Coffee Shop	20
Gym	19
Public Art	16
Seafood Restaurant	15

Cluster 1	
Category	No. of Venues
Taco Place	232
Mexican Restaurant	122
Office Supplies Store	54
Brewery	33
Clothing Store	30
Print Shop	29
Bakery	28
Bar	24
Pizza Place	24
Restaurant	23

Cluster 2	
Category	No. of Venues
Mexican Restaurant	308
Taco Place	78
Bar	70
Coffee Shop	58
Café	54
Ice Cream Shop	51
Pizza Place	37
Restaurant	36
Bakery	33
Hotel	32

Figure 6

An initial observation is that all three clusters are dominated by food venues and all three contain the Taco Place category in the top 5. However, looking at the tables above and the 3-cluster model map (Figure 4) we can draw some insights. Blue stations tend to be concentrated in the city center, where it seems that bars are prevalent. Note the other two clusters do not contain the bar category in the top ten. Furthermore, the hotel category places in the top ten. All these observations suggest that blue stations are tourist areas as the have plenty of hotels and bars.

Regarding magenta stations, which seem to be located throughout the city, have parks as their most common category. Also note that the gym category places at number eight. This suggest that neighborhoods surrounding magenta stations are residential areas.

Stations labeled under cluster 1 (red) seem to be neighborhoods that will have a lot of office space. This conclusion is derived from the fact that office supply stores and print shops are both in the top ten venues categories.

5 Conclusions

In this analysis, I tried to cluster neighborhoods surrounding each metro station of Mexico City's metro system. The 3-cluster model showed the city center to be more tourist oriented and vibrant. On the other hand, the outskirts of the city seem to be of mixed use, varying from office oriented neighborhoods to what seems to be heavily residential neighborhoods. This data could be useful to city planners to identify the flow of people from residential to office space areas. Furthermore, it could help identify area needing expansion. The next steps for this analysis would be to aggregate the bus stations that complement the metro system. This with the objective of obtaining a more accurate and more encompassing representation of Mexico City.