LA Metro Lines Characterization

Introduction

Driving in Los Angeles (LA) traffic could be a daunting task for tourists or people who are staying in LA for only a short while. Not only does LA have the worst traffic in the US (according to CBS news), but also confusing traffic layouts and road surfaces in dire disrepair. With LA being one of the host cities of World Cup 2026 and the host city of 2028 Olympics, it is clear that LA metro will be playing a major role in transporting tourists.
Therefore, the characterization of the metro lines would be of great interest to tourists and local businesses alike. In this study, we are looking to characterize the metro lines by the venues surrounding the metro stations. For tourists, this study provides an idea of what venues to expect along the metro lines. For businesses, this study presents an opportunities to find a niche along the lines and avoid areas where competition is overly saturated.
For this study, we will focus our efforts on studying the metro lines that pass through the city of LA, namely, the Red, the Gold and the Expo lines.

Data

In order to characterize the metro lines described previously, we will be using Four Square to collect list of venues and their categories for all metro stations along the Red, Gold and Expo lines. To do so, we'll first set up a function to loop through all metro stations to collect venue data from Four Square, then proceed to import data containing all coordinates data of all metro stations so that we can use the loop function defined previously to loop through all metro stations and collect data from Four Square.

With the quarries completed, let's inspect the data frame to see the data inside.

```
In [8]: metro_stn_venues.head()
```

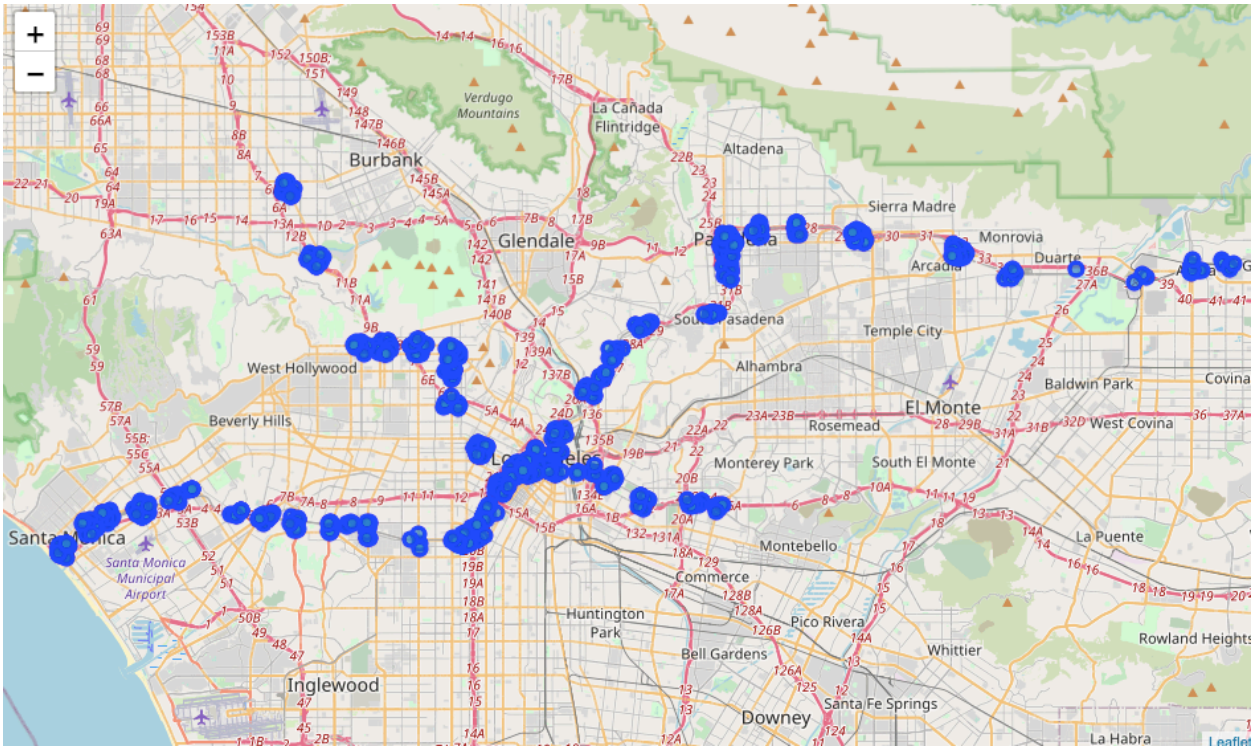
Out[8]:

	Station	Station Latitude	Station Longitude	Venue ID	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Atlantic Station	34.0334	-118.154	5397b842498ea56da1541a95	Tacos Ensenada	34.033560	-118.153599	Mexican Restaurant
1	Atlantic Station	34.0334	-118.154	4c37f8a83849c92844cebeb1	Bob's Freeze	34.032557	-118.154414	Ice Cream Shop
2	Atlantic Station	34.0334	-118.154	4ab55445f964a520fc7320e3	Los Molcajetes	34.032970	-118.155352	Latin American Restaurant
3	Atlantic Station	34.0334	-118.154	4b64c062f964a52048cd2ae3	Fish Taco Express	34.032529	-118.154530	Taco Place
4	Atlantic Station	34.0334	-118.154	4b83365ff964a520a1fd30e3	SUBWAY	34.032530	-118.153702	Sandwich Place

As shown above, the data set includes the ID, name, coordinates and categories of each vanue found as well as the station and its coordinates to which the venue is in close proximity to. With the geological data and labels, this data set allows us to see what is surrounding each metro station, and therefore able to characterize each station and metro line.

Methodology

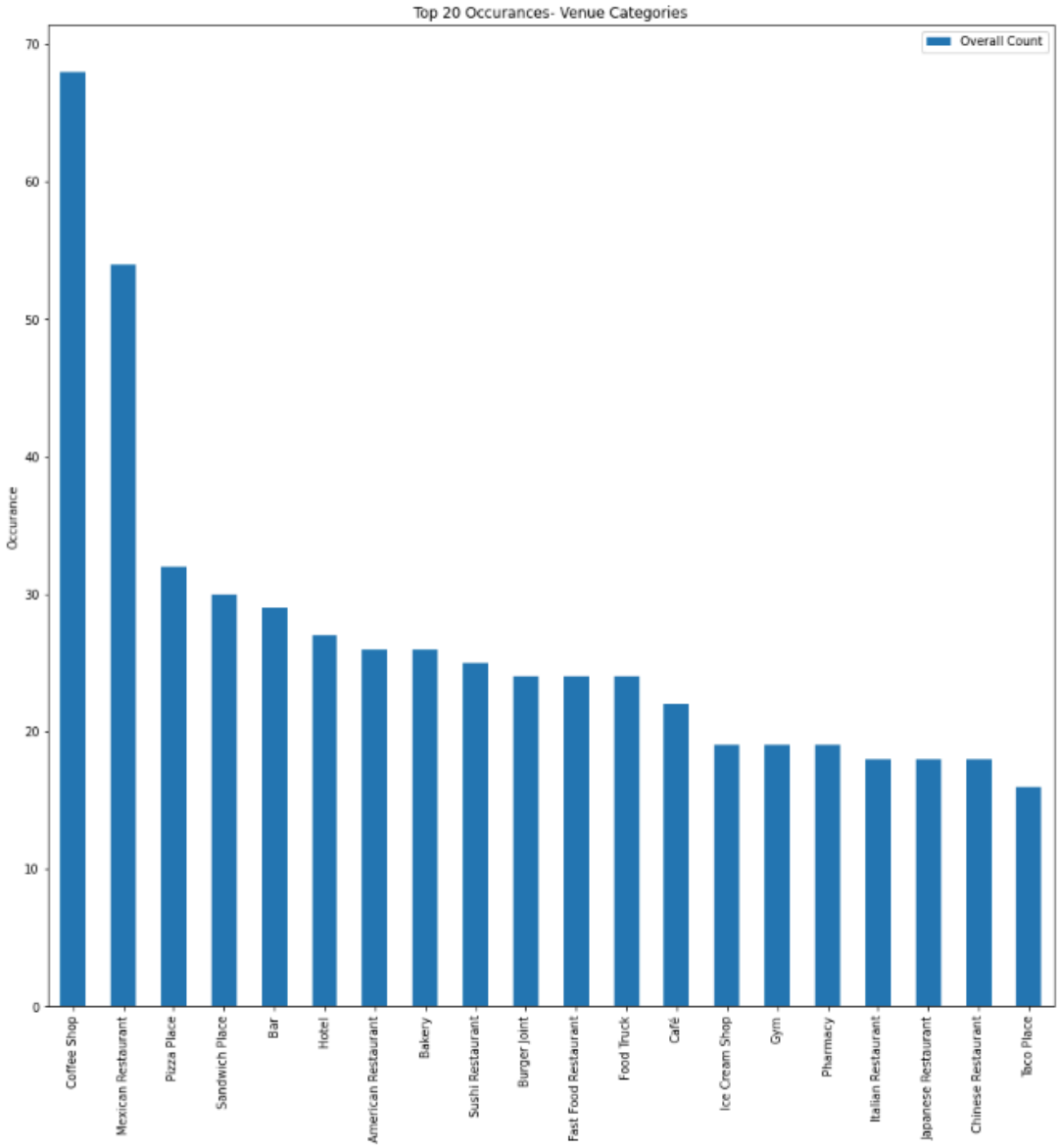
To characterize the metro lines, the first thing to do is to visualize the overall geolocation data of all venues we have gathered in the previous section. To do so, we'll create a map centered in LA with all venues plotted as dots on the map to see the distribution of venues.



The map above shows venues around the metro line stations. Not surprisingly, it appears that stations located within the city of LA have venues more densely surrounding the metro stations.

Venue Categories

Let's shift our focus to the types of venues we find around the metro stations. To do so, we'll one-hot encode the category of venues. Then, we can obtain the top categories by summing the occurrences of the encoded table vertically and sort through the list. We can then find out the top 20 occurrences along the metro lines.



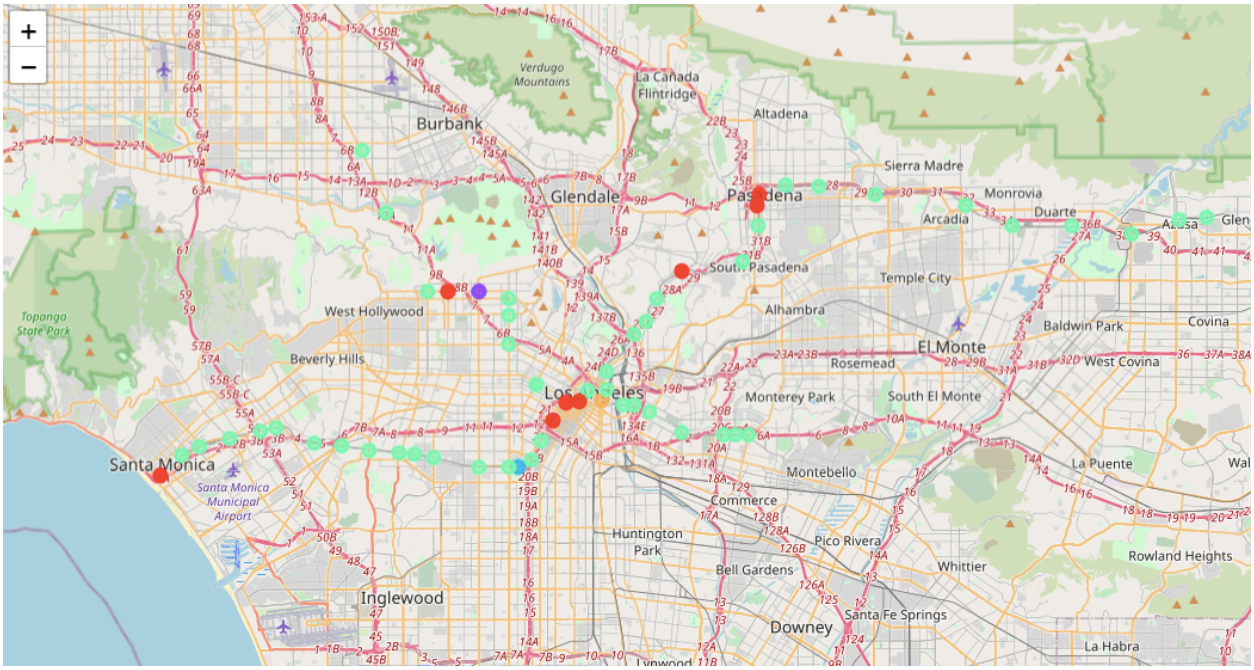
As shown in the graph above, you'll find plenty of coffee shops along the metro lines. The same applies to Mexican restaurants, pizza place and other top occurrences.

K-Means Clustering

With the data thoroughly examined, we turn our attention to seeking additional insight with machine learning. Namely, we will utilize k-means clustering to divide the metro stations into different groups.
To carry out k-means clustering on the metro stations, we first import the `Kmeans` library from scikit-learn, we then set clusters to 5 and proceed to fit the model. As a result, the `Kmeans` model produces an array of labels for all metro stations.

To help differentiate the difference between labels, we'll create a data frame of top 10 most common values for each station. Then add the kmeans labels to the data frame to complete the picture.

Next, we'll visualize the distribution of metro stations with different labels on the map. To do so, we'll add the geophysical information of each metro station into the data frame just created. A map can then be created with stations of different labels plotted in different colors.



Results and Discussion

Recommended Station to Stop By

A typical tourist would likely want to stop by the stations with most venues close by. With our data, we can generate a list of top 20 stations with most venues close by. We can accomplish this by counting the venues for each station and sort by respective venue counts. As follows.

	Station	Overall Count	Line
0	Little Tokyo/Arts District Station	95	Gold
1	Pershing Square Station	89	Red
2	Memorial Park Station	67	Gold
3	Hollywood/Western Station	60	Red
4	Highland Park Station	49	Gold
5	Expo Park/USC Station	49	Expo
6	Del Mar Station	48	Gold
7	7th Street/Metro Center Station	48	Red/ Expo
8	Downtown Santa Monica Station	47	Expo
9	Hollywood/Vine Station	46	Red
10	Vermont/Sunset Station	44	Red
11	Fillmore Station	36	Gold
12	Expo/Bundy Station	34	Expo
13	Pico Station	34	Expo
14	North Hollywood Station	31	Red
15	Arcadia Station	30	Gold
16	Universal City/Studio City Station	30	Red
17	Hollywood/Highland Station	28	Red
18	26th Street/Bergamot Station	25	Expo
19	Culver City Station	25	Expo

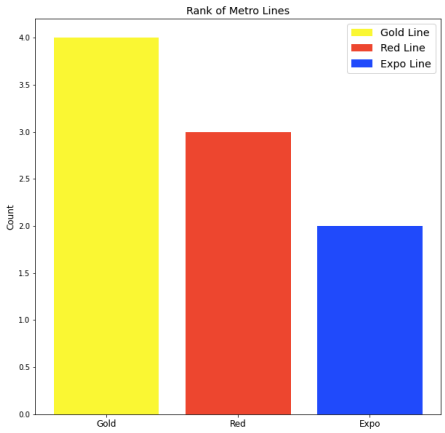
As suggested by the data, the top 3 stations to stop by are:

1. Little Tokyo/Arts District Station
2. Pershing Square Station
3. Memorial Park Station

Indeed, a little research shows that Little Tokyo/Art District and Pershing Square, both close to downtown LA, are popular tourist spots with lots of attractions and food venues. The Memorial Park Station is located in the city of Pasadena. With Old Town Pasadena close to the station, it's not hard to see why there are many venues densely surround the station.

Battle of the Lines

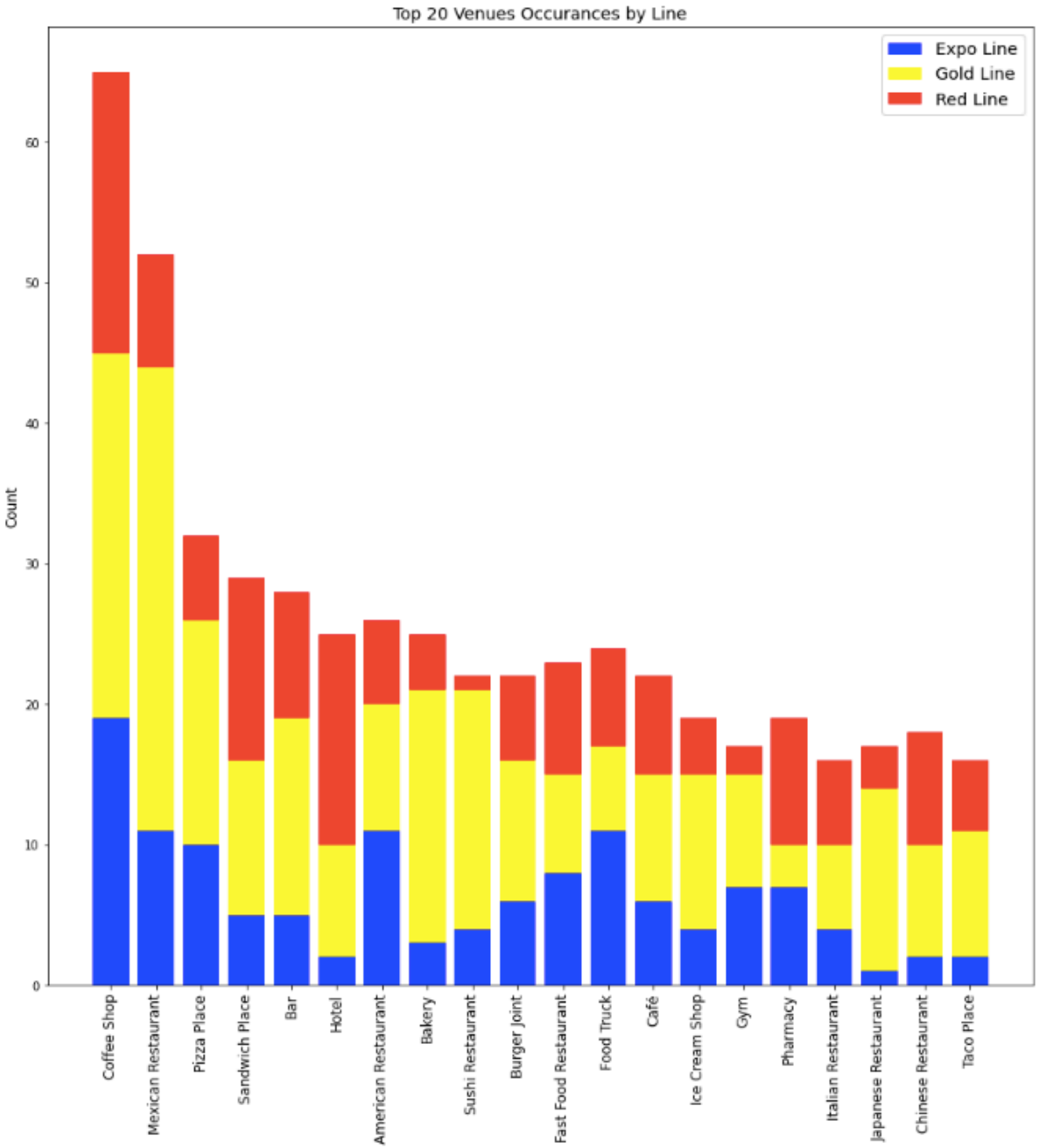
So, whcih line is the most exciting?
To answer this question, let's rank the lines by how many stations they have in the top 10 list.



As shown above, the Gold line apears to be the winner with 4 of it's stations in the top 10 list. Followed by the Red line with 3 stations in the top 10 and followed by the Expo line with 2 stations in the top 10 list.

Characterizing the Lines

Now that we know the most frequently encountered venues along the metro lines. We could then seek to categorize the lines by examining the distribution of each of the top 20 categories amongst the different lines. To do so, we group the venue occurrences by line and then create a stacked bar graph with yellow, red and blue bars representing the Gold, Red and Expo lines, respectively.



A few points are observed based on the graph shown above:

1. Large portion of Mexican restaurants are concentrated along the Gold line stations. For tourists seeking Mexican food, the gold line would offer great selection. On the contrary, stations along the Gold line are where competition is most fierce for restaurant owners looking to open a Mexican restaurant.
2. For tourists looking to find accommodation, accommodation is most likely to be found along the red line as a large portion of all hotels are located along the red line. Given the lack of sushi restaurants on the red line, there's an opportunity for restaurateurs to open a sushi restaurant along the red line.

3. For tourists looking for great American restaurants and food trucks, the Expo line offers great selections. On the other hand, if you're looking for hotel to stay, it would be better to look elsewhere than the Expo line.

Results from K-Means Cluttering

Based on the map created using k-means cluttering, we can conclude as such:

1. The majority of the metro stations falls under the group plotted as teal dots on the map, and has no distingushing features from one another. These are not particularly interesting places for tourists to visit.
2. Stations plotted in red are among city centers and have an abundance of venues nearby. These are recommended stations for tourists to stop by.
3. Lastly, we have 3 metro stations being singled out by the k-means algorithm. There are our top recommended stations to tourists each with strong distinguishing characteristics. Namely, the Hollywood/Western Station (purple), the Expo Park/USC Station (cyan) and Little Tokyo/Arts District station (orange).

Conclusion

With data collected from Four Square, we have been able characterize the metro stations and the lines they belong to as described in the results and discussion section. Through data visualization techniques, we can see the types of venues that characterize each metro line and station. Furthermore, by utilizing the k-means cluttering technique, 3 particular stations had been singled out by the algorithm as top recommended stations to visit for tourists.