Data Science Capstone Project

# Introduction/Business problem

Metro Manila, simply Manila, is the National Capital Region and the Philippines' prime tourist destination. Manila comprises 17 cities and municipalities, including the capital city, Manila City. Though it is the smallest region in the country, Metro Manila is the most populous of the twelve defined metropolitan areas in the Philippines and the 19th most populous in the world (Koop, 2021). Being the capital, Manila is considered to be the center of commerce, education, and entertainment of the country.

In this Capstone Project, I want to know the most common places available to Manila's people. After I determine the Top 3 Most Common Places, I will cluster the type of business and determine what kind of establishment is best to set up in a particular city. Lastly, I also want to know the most common cuisines people prefer to eat here in the Metro.

# Description of the data

The different districts present in Metro Manila were obtained from the Philippines Statistics Authority records (PSA). Every city has been divided into separate smaller neighborhoods or "barangays," as termed in the Philippines. To make the data processing easier, Microsoft Excel is used to clean the PSA data while also retaining the different Metro Manila cities.

Each city's central location's latitude and longitude is requested from Google Cloud's Geocoding API. These coordinates will then be used as input in the Foursquare API, which is used to obtain the different venues present at each city within a 3000-meter radius relative to its coordinates.

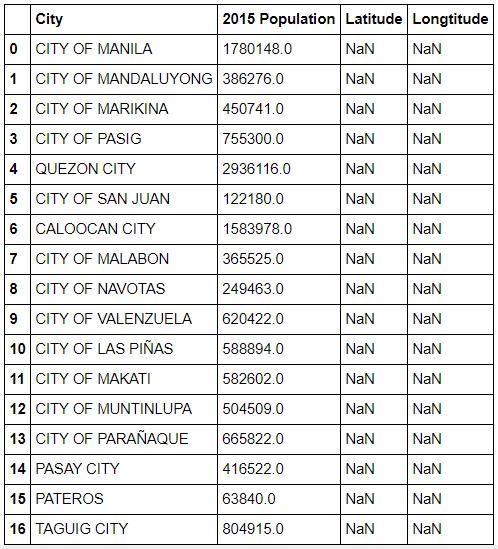
Lastly, from the Foursquare API's response, the result will be parsed to obtain necessary values such as each venue's coordinates, name, venue sub-category, and main category. The primary type is how Foursquare classifies the venue. To name a few, this can be food, arts, and education. Venue sub-category is the specific category a venue is classified as, i.e., Japanese Restaurant for Food.

# Methodology

This section will describe the data analysis and how I used the data to yield the results.

I cleaned the PSA data and loaded it to the Jupyter Notebook as a Pandas data frame[.](https://en.wikipedia.org/wiki/Districts_of_Cologne) For this, I used the pandas read function. Once I loaded the data, I had to clean it further by renaming two (2) columns for Latitude and Longitude while removing an extra ghost column imported from the file. The table below shows the processed data frame named "NCR\_data."

**Table 1 NCR\_data data frame**



As discussed, each city's central location's latitude and longitude is requested from Google Cloud's Geocoding API. The data has provisions for the Latitude and Longitude but had NaN values. Using the Geocoding API documentation, I looked up each city and directly added the values to the NCR\_data table.



Figure 1 Code used to collect coordinates and directly append to the main data frame

I used Folium to plot the data points, embedded on an interactive Map, to verify the collected coordinates.

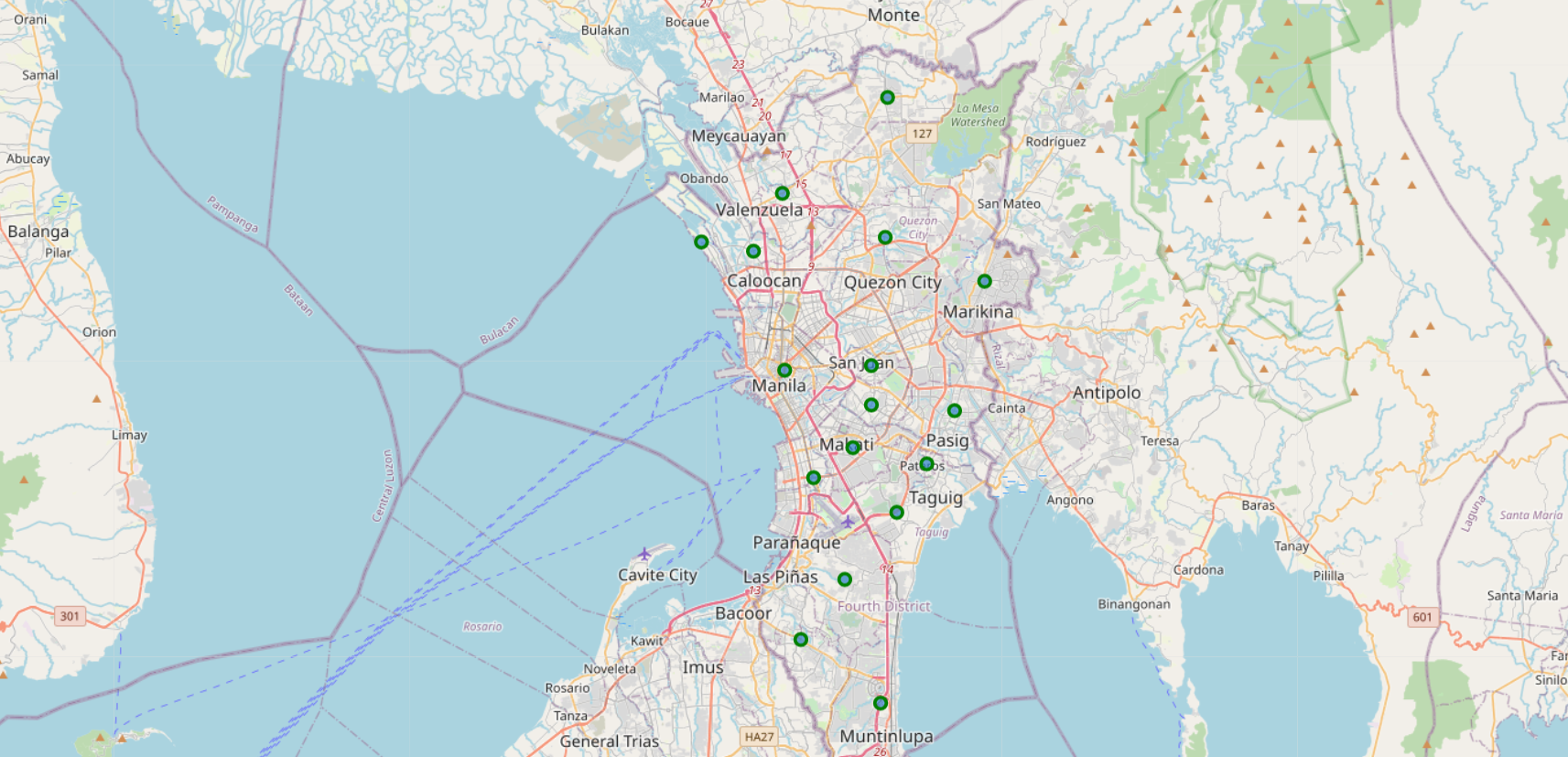
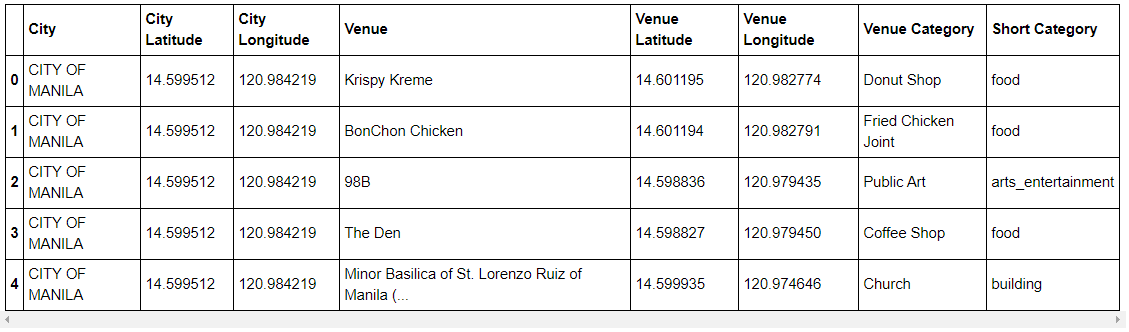


Figure 2 Plotted points in Folium map

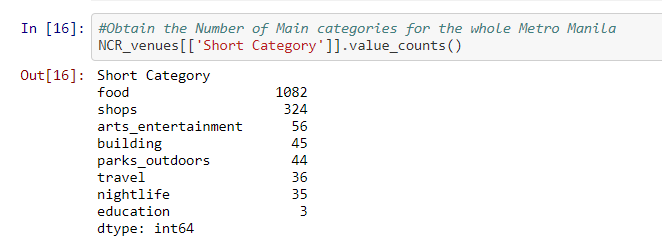
Since the coordinates have been collected, I was able to proceed to use the Foursquare API. A search radius of 3000 meters was used to return 1625 data points from each city's center. It is important to note that Foursquare does not directly return the main category (named 'Short Category' in the table) a venue is classified under.

Table 2 NCR\_venue Data Frame with data from Foursquare



There are two (2) main reasons as to why I added the Short Category column. The first reason is I will be able to determine the top places available in Metro Manila. Coincidentally, this method is more efficient in clustering cities with the same kinds of places/businesses. Secondly, adding the Short Category will also allow me to select only the data I need for a specific type; in this case, I needed the 'food' category.

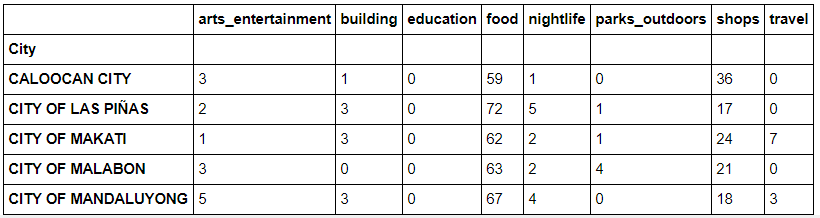
Using Pandas to manipulate the data, each venue type's sum is outputted in descending order along with its total.



It will be a good idea to visualize the data. However, with the values returned by the system, it is clear what the top 3 types of venues are.

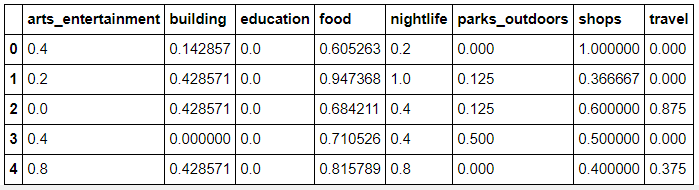
In order to obtain a data frame that contains the totaled summary of available places/businesses in each city, the data frame is first transformed by one-hot encoding (0/1) the venue types and then adding up the values per city.

Table 3 Total Location Types per City



The summarized data frame is further processed using the MinMax Scaler to obtain values between 0 and 1. It is essential to note the arrangement of the index due to the merging of tables later on.

Table 4 Scaled Table of Location Type per City



With the data transformed to 1s and 0s, it is now possible to cluster the dataset. However, there are methods, like the Elbow method and the Silhouette score, to determine the optimal number of clusters (k) to analyze the data. For this analysis, the Silhouette score was used because the sample size is less than 18, which is the minimum for the Elbow method.

An advantage of using the Silhouette score in determining the optimal number of clusters is that this test readily shows the optimum number. In previous tests, I observed that the optimal number of groups is 4. However, due to further data processing, the code has gotten mixed up and returned two (2) as the optimum number of clusters.

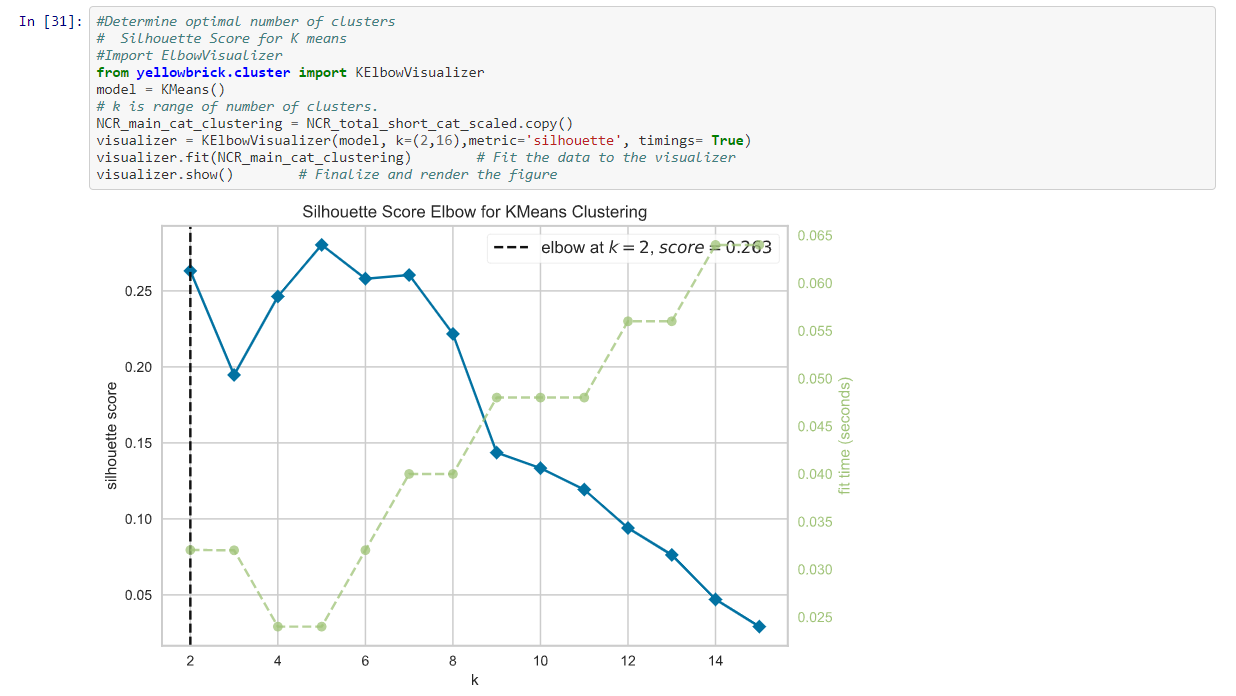


Figure 3 Silhouette Score Method to Determine Optimum Number of Groups

I ran an unsupervised machine learning algorithm with all this data, specifically a k-means clustering algorithm from the Scikit-learn package. One could use the elbow method to define the k value systematically. Still, I chose k to be four since most of my tests returned four as the optimum number.

The same steps were recreated to determine the clusters of restaurants for each city. However, in selecting the datapoints, only those with the 'food' location type were selected. After saving this to a data frame, the data was visualized through a bar graph and histogram to determine the distribution of the data set.

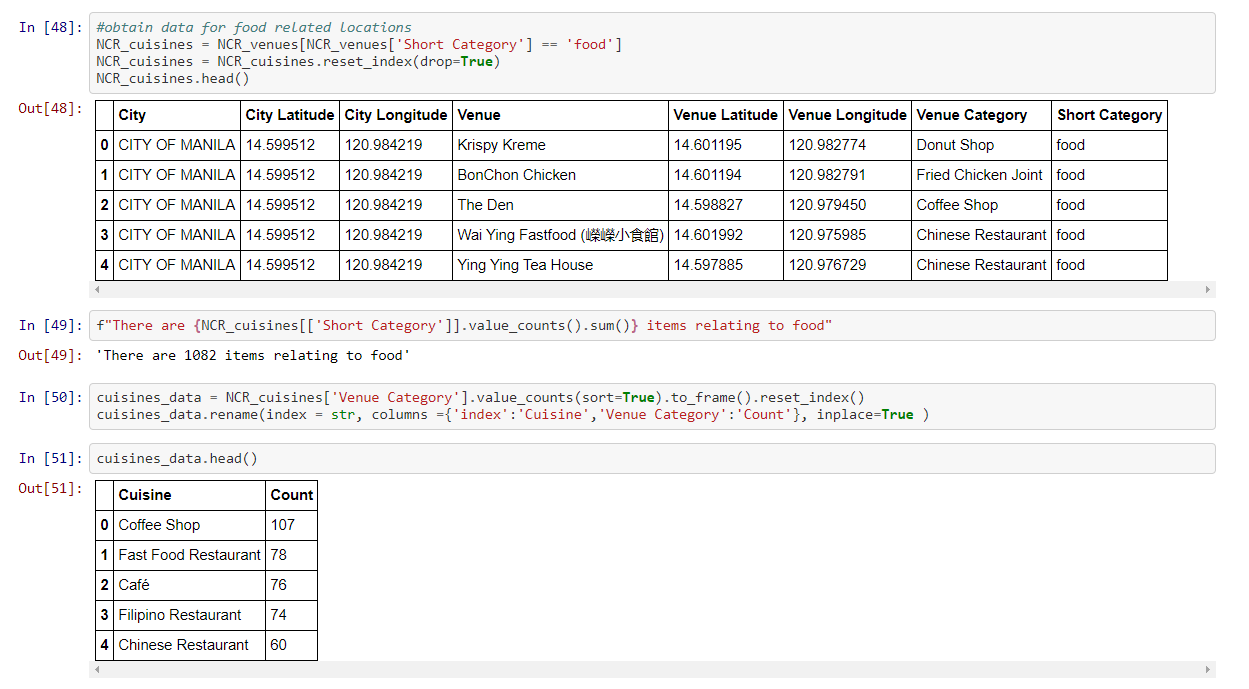


Figure 4 Code to Filter Food Location Type Data

I had determined that the majority of the food-venue category has a value of 1 to 11.6. As much as possible, the data should be flat because I am concerned only with the top cuisines available in Manila.

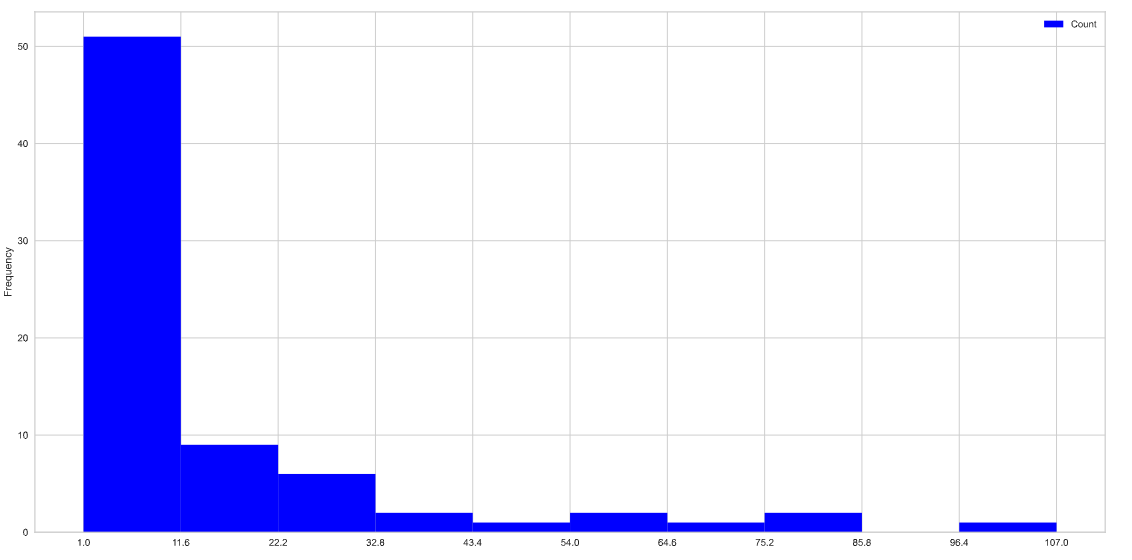


Figure 5 Histogram of Cuisines available in Metro Manila

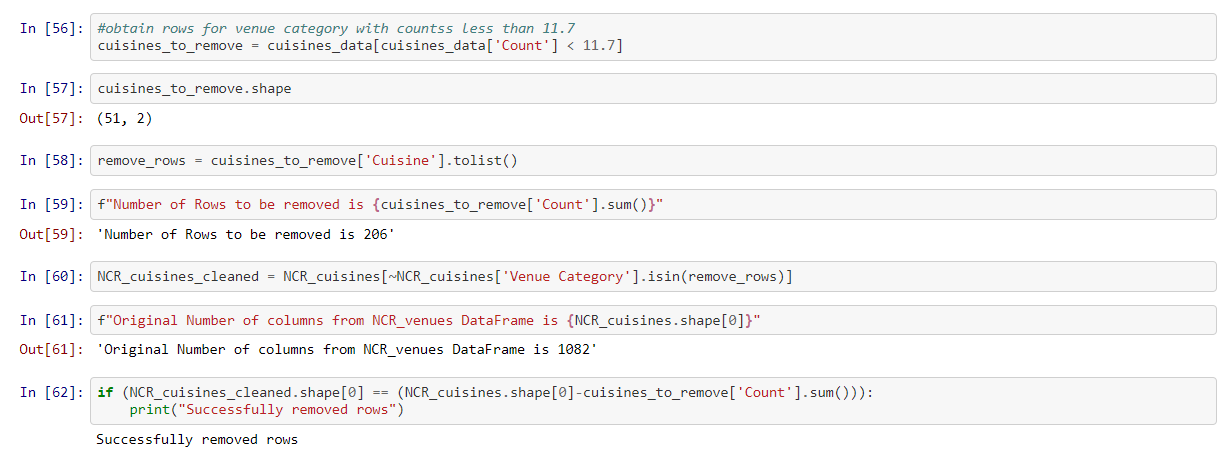
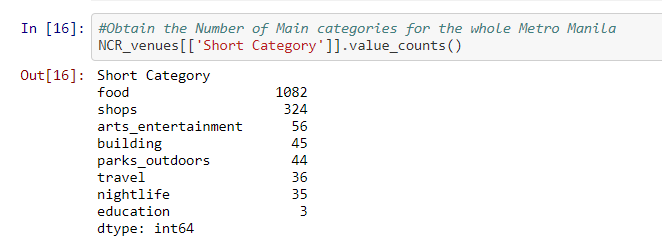
In order to easily remove the unwanted cuisines in the dataset, I used the code in the figure below. The logic behind this is that I collect the unwanted cuisines and save this as a list. From the filtered food data frame, I only selected the rows, *not in* the list (seen in line 60).  


Figure 6 Code to Clean Unwanted Data

The data is the one-hot encoded using the mean and was processed using the k-means algorithm with the optimum number of clusters obtained from the Silhouette score.

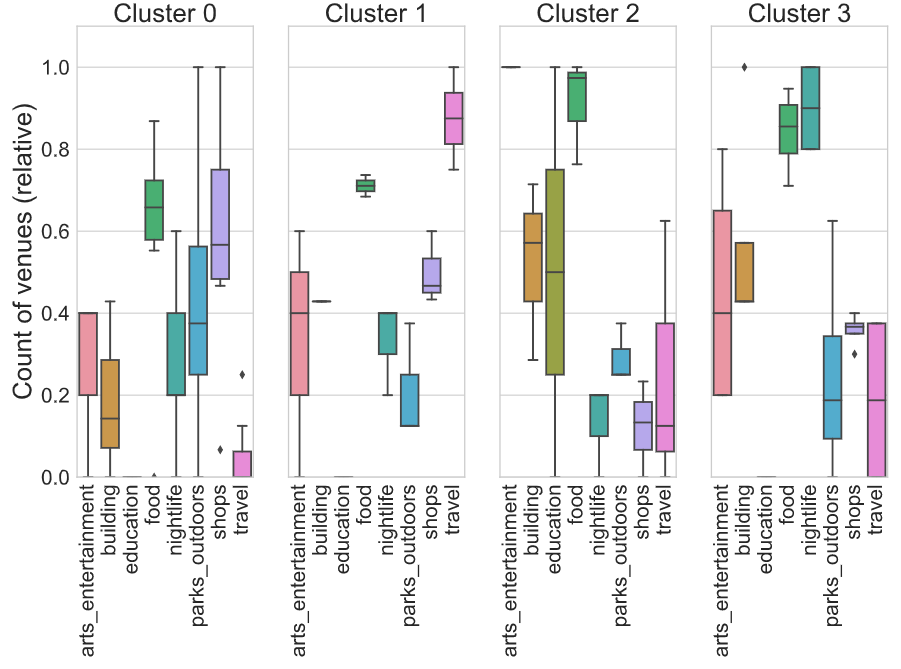
# Results



The picture above shows that the top most-common venue types in Metro Manila are Food, Shops, and Art and Entertainment.

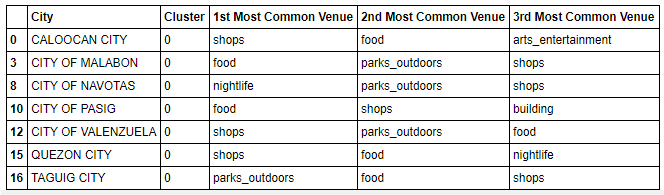
## Venue Type Cluster

Four (4) clusters of location type concentrations for Metro Manila are shown in the box plot below. The different clusters assigned from 0 to 3 show the most common venue type available in each city. However, it is still difficult to discern the depicted data, so I provided more context on the next page.

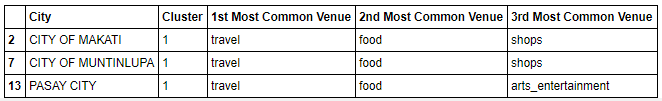


For simplicity's sake, Clusters 0 to 3 have been renamed to Clusters 1 to 4. The analysis has outputted groups that are divided into shops, travel, arts and entertainment, and buildings. The clusters have similarities, but some refinement can still be done to represent more accurate data.

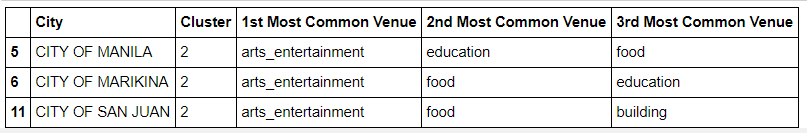
### Cluster 1: Shops



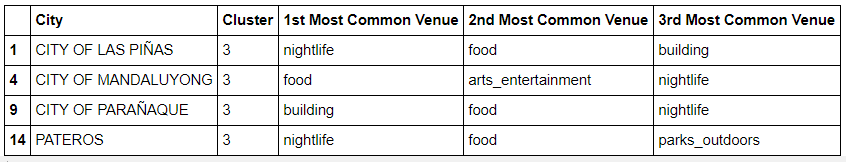
### Cluster 2: Travel



### Cluster 3: Arts and Entertainment

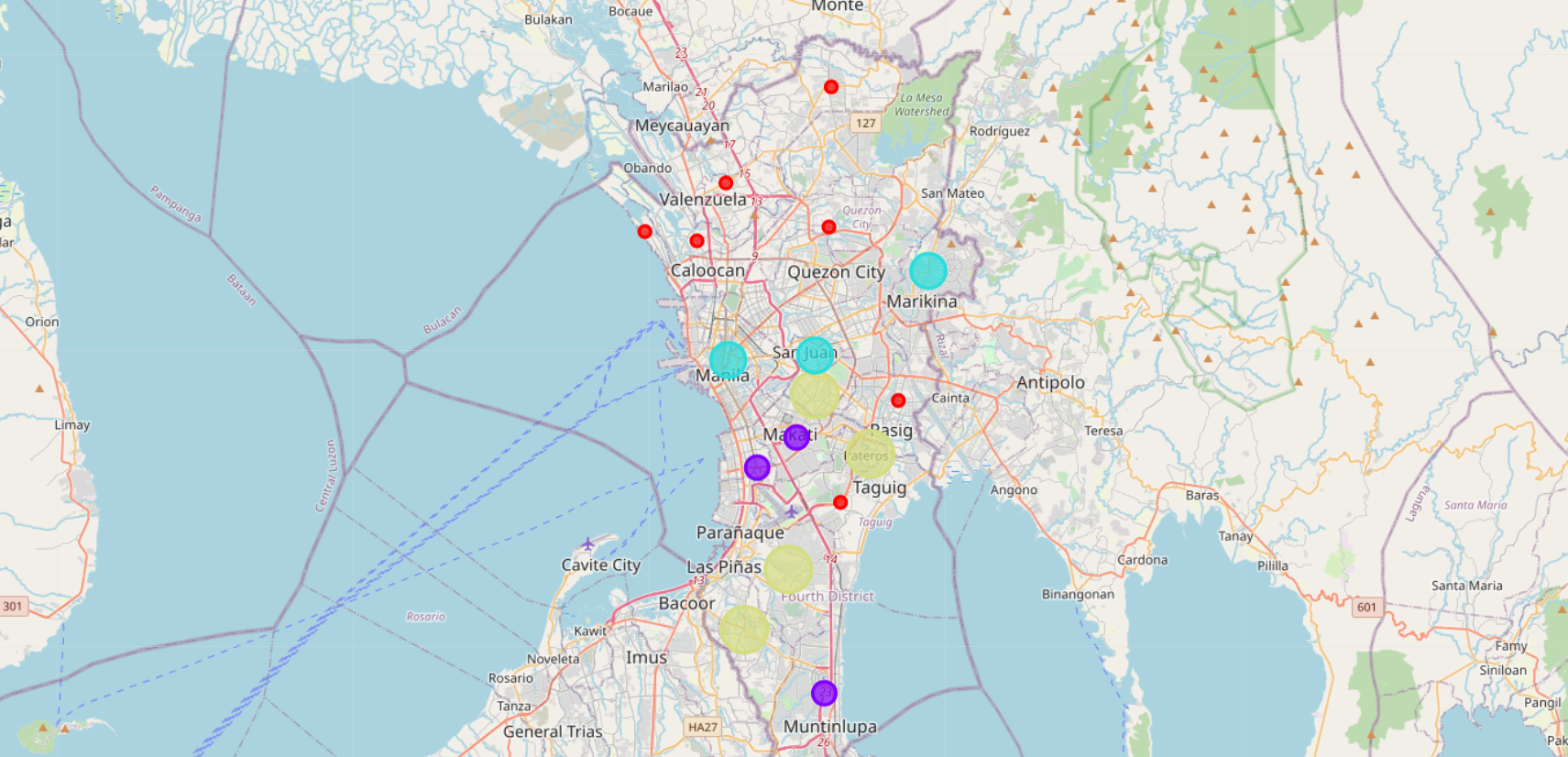


### Cluster 4: Building and Nightlife



We can now use the cluster labels to show the city districts marked with a cluster-specific color on a map, again using Folium.

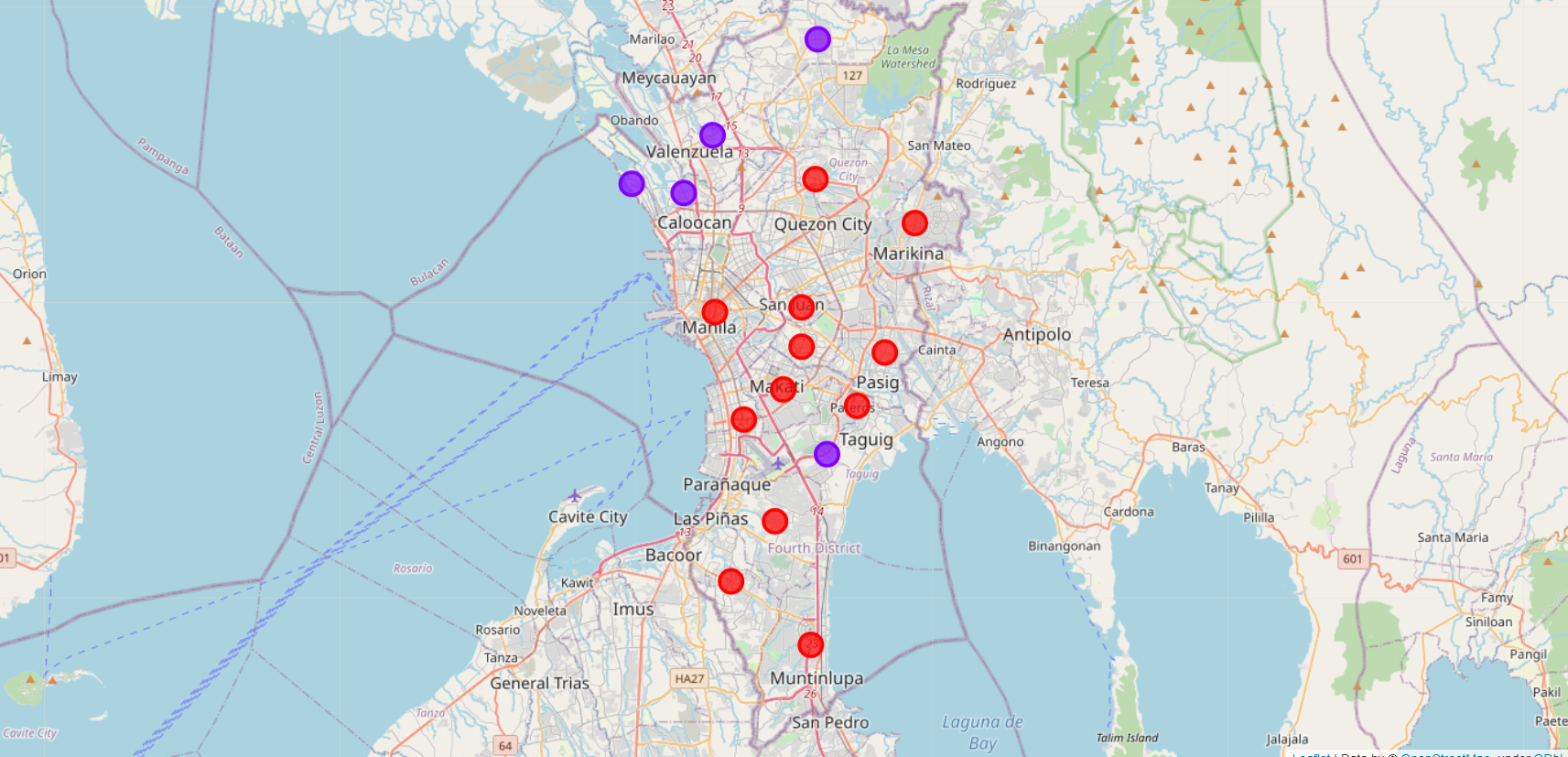
The different clusters are shown on the map. For easier visualization, I have used different sizes for each cluster – this, however, does not accurately depict the size of the cluster – to differentiate each of them easily. There are seventeen (17) bubbles for the seventeen (17) cities, with four (4) different colors for the four (4) clusters.



With the analysis complete, we can now define where a person can set up an establishment or business with the different clusters in Metro Manila.

## Cuisine Cluster

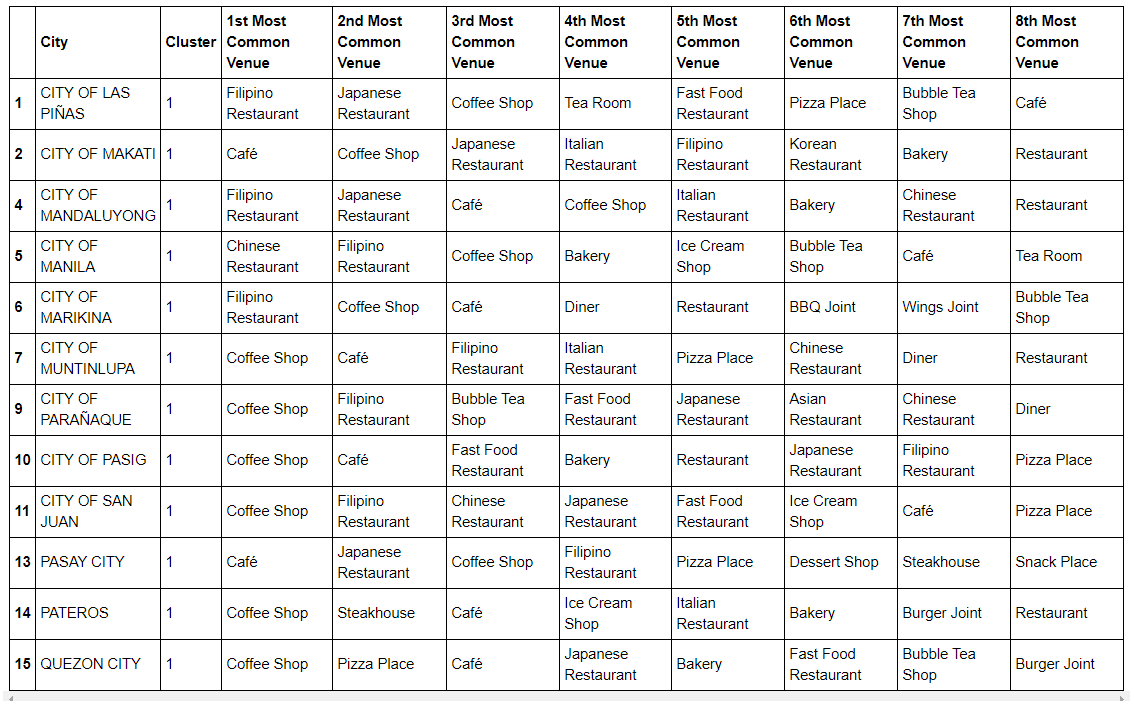
All of the significant cuisines are represented thanks to the K Means clustering method. Cluster 0, where people have more choices for fast food. Cluster 1 is where Restaurants and Cafés are the 1st choices for people.



### Cluster 0: Fast Food



### Cluster 1: Coffee Shop and Restaurants



# Conclusion

From the results, we can see the readily available venue type for people residing in Metro Manila. Subsequently, we can also know what type of business is best established per city while even knowing where to eat. Although the analysis is not perfect and further refinement is needed, it is a good stepping stone in learning data science.

# ****Appendix****

The publications of Dr. Johannes Wagner and Kristian Jackson have helped me in various stages of this project. Additionally, the IBM Labs resources that were taken up to this point have helped in reviewing codes needed to analyze and visualize the data.

Link to Dr. Johannes Wagner's blog post: <https://www.linkedin.com/pulse/applied-data-science-capstone-project-restaurant-wagner-mba/>

Link to Kristian Jackson's Git Hub Repository:

<https://github.com/kristianjackson/Coursera_Capstone>

Link to Koop, 2021: <https://www.visualcapitalist.com/most-populous-cities-in-the-world/>

Link to the PSA Data:

<https://psa.gov.ph/sites/default/files/attachments/hsd/pressrelease/2015_Table%201_Legislative%20Districts.xlsx>