Data Science Capstone Project

Table of Contents

[Introduction/Business problem 1](#_Toc67005930)

[Description of the data 1](#_Toc67005931)

[Methodology 2](#_Toc67005932)

[Results 8](#_Toc67005933)

[Venue Type Cluster 8](#_Toc67005934)

[Discussion 10](#_Toc67005935)

[Conclusion 10](#_Toc67005936)

[Acknowledgement & sources 10](#_Toc67005937)

# 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 place 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.

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

# 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 different 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 main 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.

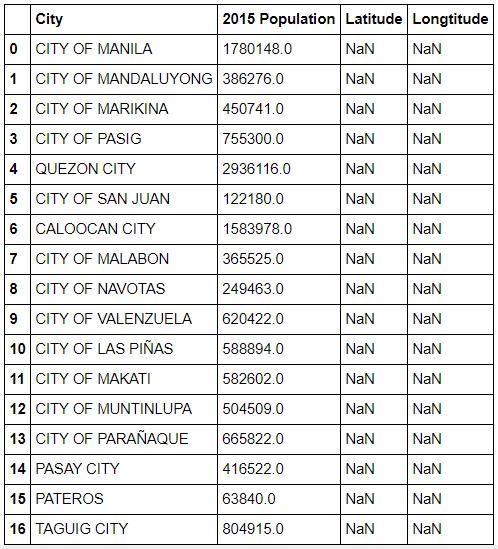
Link to the PSA Data: <https://psa.gov.ph/sites/default/files/attachments/hsd/pressrelease/2015_Table%201_Legislative%20Districts.xlsx>

# Methodology

In this section, I 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 that was imported from the file. The table below shows the processed data frame named “NCR\_data.”

**Table 1 NCR\_data Data Frame**

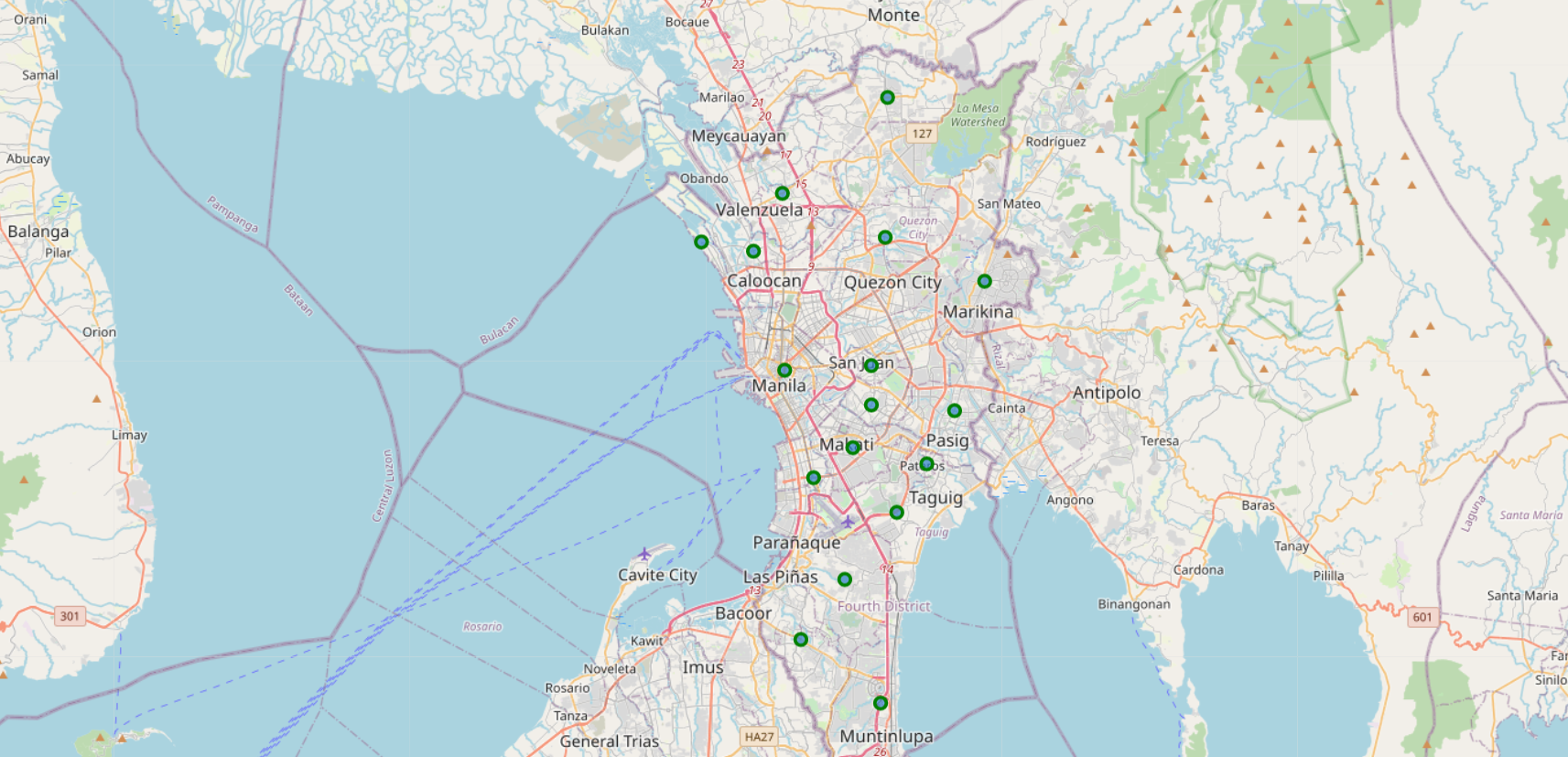


As discussed, each city's central location's latitude and longitude is requested from Google Cloud's Geocoding API. The data has provisions for the Latitude and Longitude, but had NaN values. Using the documentation of the Geocoding API, I looked up each city and directly added the values to the NCR\_data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

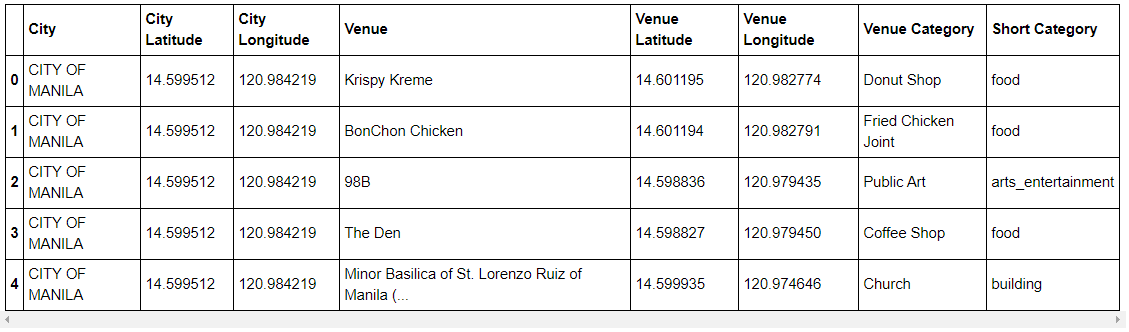


Figure 1 Code used to collect coordinates and directly append to the main data frame

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

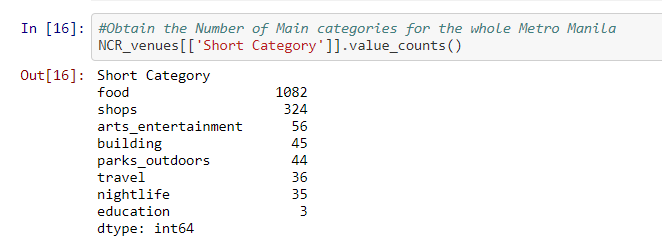


Since the coordinates have been collected, I was able to proceed to using the Foursquare API. From each city’s center, a search radius of 3000 meters was used to return 1625 data points. It is important to note that Foursquare does not directly return the main category (named ‘Short Category’ in the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



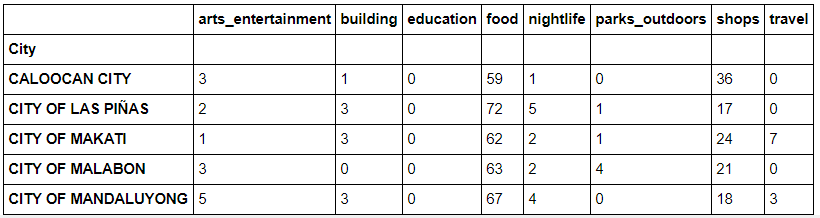
There are two (2) main reasons as to why I added the Short Category column. The first reason is I will be able to determine the top places available in Metro Manila. Coincidentally, this method is more efficient in clustering cities with the same kinds of places/businesses. Secondly, adding the Short 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, the sum of each venue type is outputted in descending order along with its total.

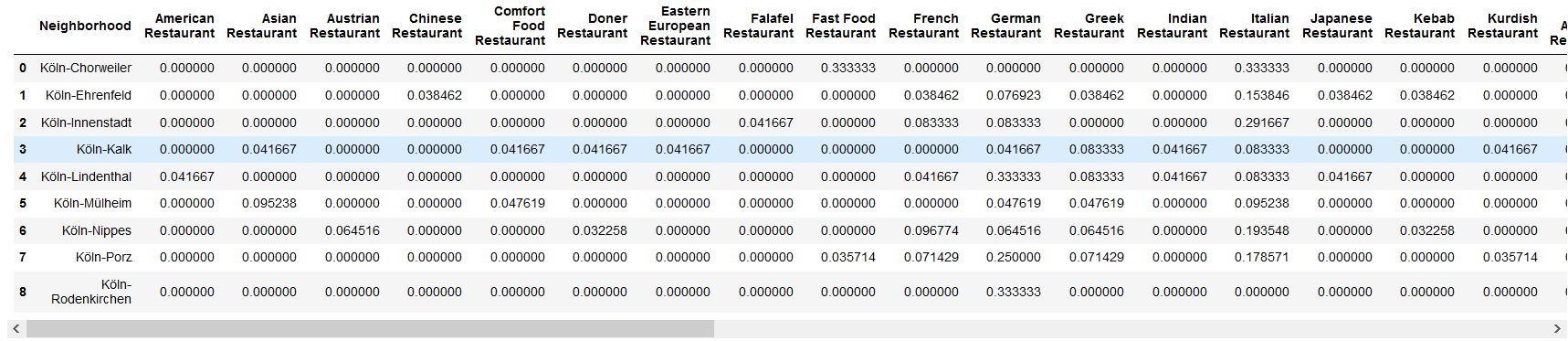


It will be a good idea to visualize the data. However, with the values returned by the system, it is clear what the top 3 types of venues are.

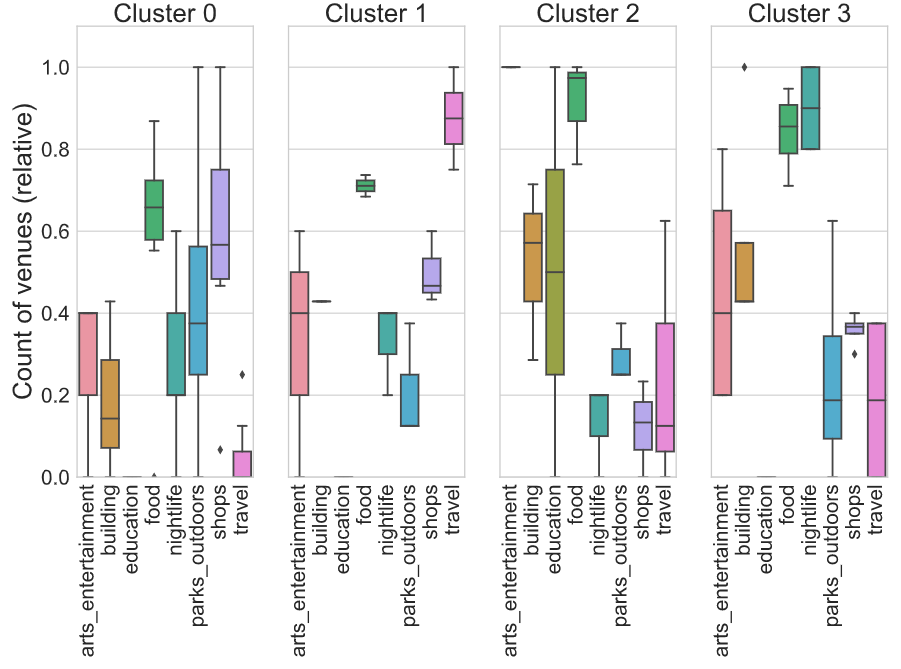
To find the cluster the different venue types in each city, the data frame is transformed the by one-hot encoding (0/1) the venue types and then adding it all up to obtain a data frame which contains the summary of available places/businesses in each city.

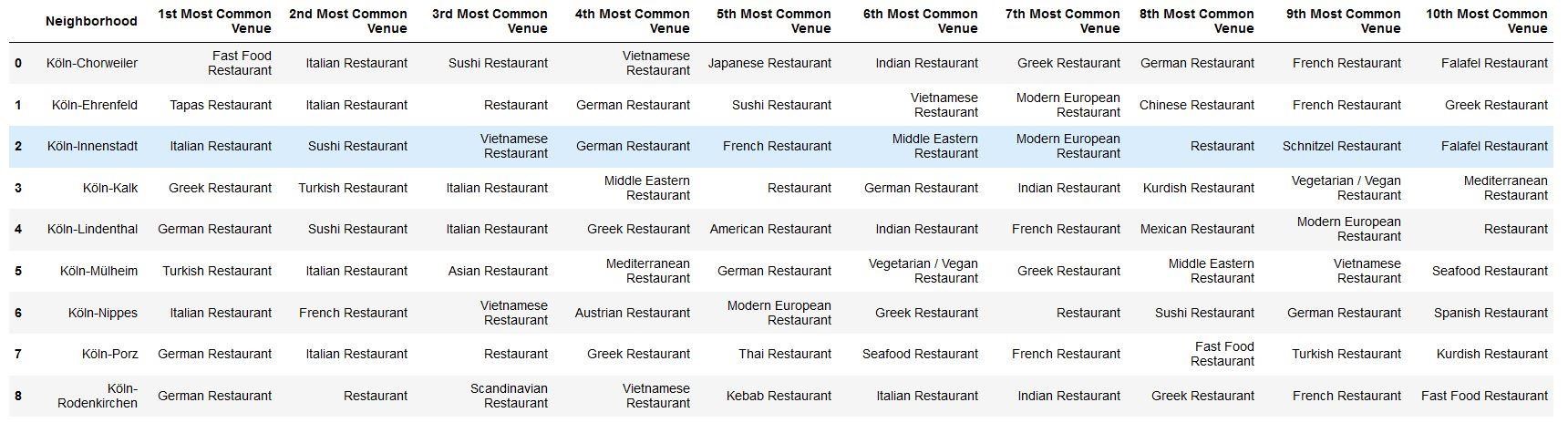


Next, I used grouping to show the frequency of each category of restaurants in each city district.



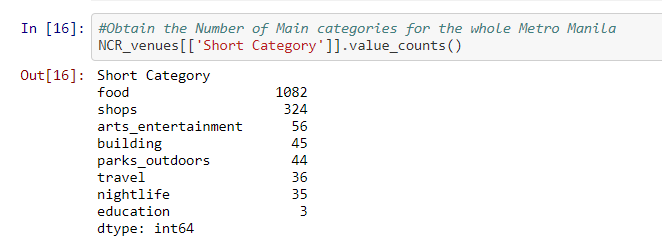
I used this information to create a data frame in which you can see the most common restaurant venue types for each city district.





Now, with all this data, I could finally run an unsupervised machine learning algorithm, more specifically, a k-means clustering algorithm from the scikit-learn package. One could use the ellbow method to systematically define the k value, but I simply chose k to be 5, having been inspired by one of the coursera courses to do so.

# Results

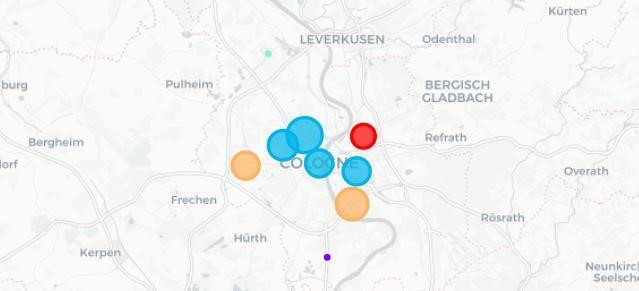


From the picture above, it is clear that the top most common venue types in Metro Manila are Food, Shops, and Art and Entertainment.

## Venue Type Cluster

What we see in the table are the city districts and their most common venues, and they now have been assigned five different cluster labels from 0 to 4.

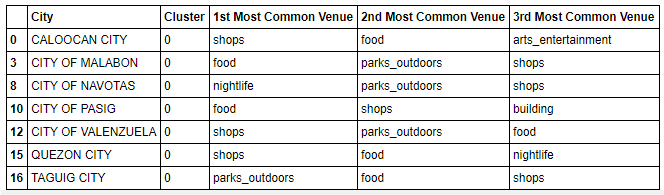
We can now use the cluster labels to show the city districts marked with a cluster-specific color on a map, again using folium:



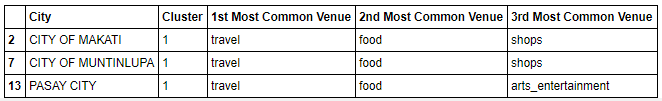
You will see nine bubbles for the nine city districts, with five different colors for the five different clusters. If you have trouble counting to five here, look for a small green dot on the upper part of the picture and a small purple dot on the lower part of the picture.

Now, what is the final result of this exercise? We now can show five clusters of restaurant type concentrations for the city of Cologne, which I named according to the restaurant concentration the data shows.

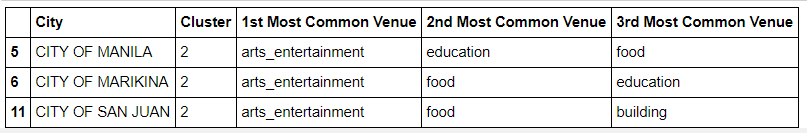
### Cluster 1



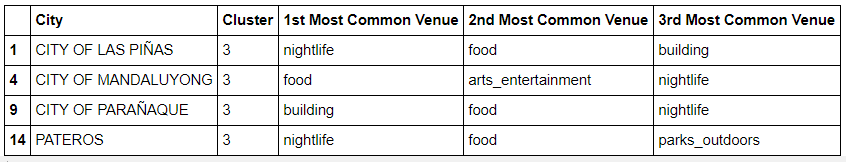
### Cluster 2



### Cluster 3



### Cluster 4



Interestingly, it is really possible to define clusters of certain cuisines in Cologne city. People living in Cologne will probably agree that these clusters sound pretty reasonable and are not too far away from what you would have expected.

# Discussion

If I reflect the work necessary to create these results, what comes to my mind is that for typical ways of scraping, cleaning, handling, transforming and visualizing data, all the tools are simply there. We just have to get to know the available open source packages and learn how to use them. What I find fantastic is that nearly all of them are free of charge. Also, a simple notebook computer is enough: in my case, I used a ThinkPad L470, more than three years old. All the rest is concentrated, creative, interesting, sometimes hard work and searching for hints, tips, examples, explanations etc. in the web. With these tools, many exciting data science use cases can be created, for all kinds of useful purposes.

# Conclusion

We achieved the goal presented at the outset of this blogpost: tourists can see in the results which city districts best match their food desires. This is just one example of fantastic data science uses cases one can realize applying technology which is available for free today! What a time to be alive.