# COURSERA Capstone Project: Diversity of Restaurants in Singapore

Author: JL

## 1. Introduction

Singapore has the second greatest population density in the world, and has very vibrant and diverse communities. As a home to a wide range of cultures, ethnicities and religions, Singapore has a rich choice of different cuisines and restaurants. People from same background and culture tend to gather spatially and form local communities, and it is assumed that such spatial pattern of communities can be reflected by the popularity and distribution of different types of restaurants. For example, places that have many Chinese restaurants may be the places that Chinese communities stay or visit most. In addition, a map that presents the clusters of different types of cuisines in Singapore can be treated as a kind of food guide map for tourists and local citizens, which can be interesting despite commercial valuable.

## 2. Data Description

To achieve the product as described above, two major datasets are required, the spatial data of Singapore and food related Point of Interests (POI).

- Spatial data of Singapore

To facilitate urban planning, the Urban Redevelopment Authority (URA) divides Singapore into regions, planning areas and subzones. The Planning Regions are divided into smaller Planning Areas. Each Planning Area is further divided into smaller subzones which are usually centred around a focal point such as neighbourhood centre or activity node. There are over three hundred subzones of a total of 55 planning areas, organised into 5 regions. To achieve a more detailed investigation, this project will be conducted in the subzone level.

The Singapore subzone shapefile data can be downloaded from the following link on data.gov.sg (https://data.gov.sg/dataset/master-plan-2019-subzone-boundary-no-sea). There are a total of 325 Singapore subzones in the data downloaded.

The subzones are presented as polygons in the original shapefile. With the help of QGIS (a spatial analysis tool), we can extract the centroid of each subzone polygon (as shown in figure 1). The subzone information including latitude, longitude and name, can be exported as a csv file for the use of POI collection later, an example is given in figure 2.

*Figure 1 Singapore subzones*

| | Lon | Lat | SUBZONE_NAME | SUBZONE_CODE | PLN_AREA_NAME | PLN_AREA_CODE | REGION_NAME | REGION_CODE |
|---|---|---|---|---|---|---|---|---|
| **0** | 103.872352 | 1.288517 | MARINA EAST | MESZ01 | MARINA EAST | ME | CENTRAL REGION | CR |
| **1** | 103.837500 | 1.294016 | INSTITUTION HILL | RVSZ05 | RIVER VALLEY | RV | CENTRAL REGION | CR |
| **2** | 103.837064 | 1.291286 | ROBERTSON QUAY | SRSZ01 | SINGAPORE RIVER | SR | CENTRAL REGION | CR |
| **3** | 103.698639 | 1.262532 | JURONG ISLAND AND BUKOM | WISZ01 | WESTERN ISLANDS | WI | WEST REGION | WR |
| **4** | 103.846053 | 1.294046 | FORT CANNING | MUSZ02 | MUSEUM | MU | CENTRAL REGION | CR |

*Figure 2 Example of the SG_Subzone data*

- Food related POI (restaurants) data

The restaurant data can be collected from Foursquare. We can search for all the POIs under the “Food” category (Foursquare categoryID is '4d4b7105d754a06374d81259') around each subzone centroid. The searching buffer is defined as 1 km to ensure a good coverage, the limit of venues returned per request was set as 100. A total of 16294 venues were collected. A sample of the collected POIs after processing is shown in figure 3. The POIs will be joint with subzones and more exploratory analysis will be conducted in the following sessions.

| | name | categories | lat | lng |
|---|---|---|---|---|
| **0** | Yen Yakiniku | Japanese Restaurant | 1.281074 | 103.845743 |
| **1** | Bam! Tapas-Sake Bar | Tapas Restaurant | 1.278393 | 103.844426 |
| **2** | Tippling Club | Restaurant | 1.279420 | 103.843848 |
| **3** | PS.Cafe | Café | 1.280468 | 103.846264 |
| **4** | Fat Prince | Kebab Restaurant | 1.277801 | 103.845202 |
| **5** | Lolla | Spanish Restaurant | 1.281034 | 103.845708 |
| **6** | Park Bench Deli | Deli / Bodega | 1.279872 | 103.847287 |
| **7** | Pantler | Bakery | 1.280137 | 103.847256 |
| **8** | Maxwell Food Centre | Food Court | 1.280291 | 103.844742 |
| **9** | Dumpling Darlings | Dumpling Restaurant | 1.280483 | 103.846942 |
| **10** | Super Star K Korean BBQ | Korean Restaurant | 1.278003 | 103.843680 |

*Figure 3 data sample of collected POIs*

# 3. Methodology

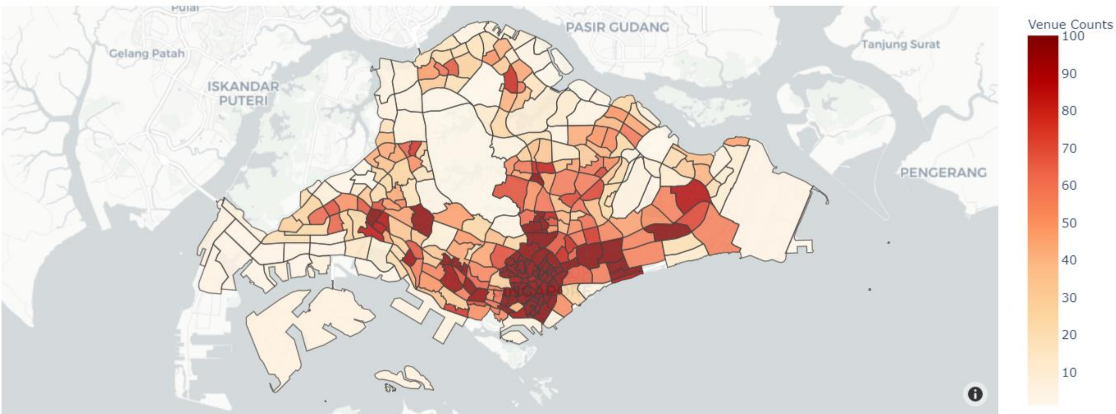
## 3.1 Exploratory Data Analysis

We can first check how many food related POIs we have collected for each subzone. We can use plotly package to create a map of venue counts. It can be observed that restaurants are mostly located in the city centre. Some subzones are missing because there is no POI within the searching radius, and these subzones are the natural areas.

In [5]:

```
# choropleth map
map_venueCounts = px.choropleth_mapbox(count_bySubzone, geojson=SG_subzone_polygon,
                locations='Subzone', featureidkey="properties.SUBZONE_N",
                color='Venue Counts',
                color_continuous_scale="OrRd",
                hover_data=['Subzone', 'Venue Counts'],
                mapbox_style="carto-positron",
                zoom=zoom_level,
                center=city_centre,
                opacity=0.8,
                title = "Venue counts by subzone"
                )
map_venueCounts.update_layout(height=500, margin={"r":0,"l":0,"b":0})
map_venueCounts.show()
```

Venue counts by subzone

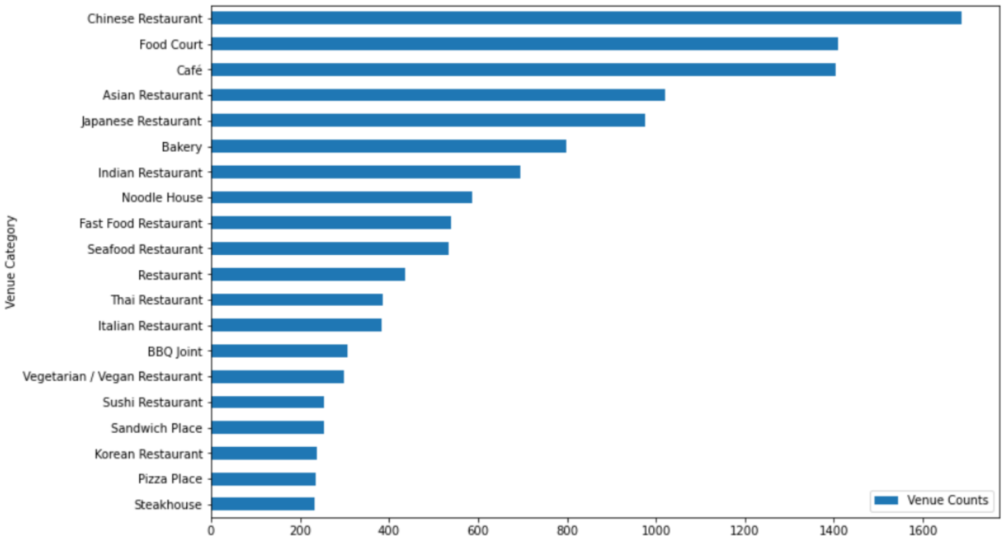


We can also check what restaurant types are most common in Singapore. As the plot shows, Chinese Restaurant is the most common type in Singapore.

```
In [6]:  # group venue by category
         count_byType = pd.DataFrame({'Venue Counts':SG_venues.groupby('Venue Category')['Venue'].count()}).reset_index()
         df_plot = count_byType.sort_values(by='Venue Counts',ascending=False).head(20)

         # create bar plot of the top 20 categories
         df_plot.sort_values(by='Venue Counts').plot.barh(x='Venue Category', y='Venue Counts',figsize=(12,8))

Out[6]: <AxesSubplot:ylabel='Venue Category'>
```




## 3.2 Data processing

Calculate distribution of different restaurant types of each subzone.

```
In [7]:  # one hot encoding
         SG_onehot = pd.get_dummies(SG_venues[['Venue Category']], prefix="", prefix_sep="")

         # move Subzone column to the first column
         Venue_categories = sorted(SG_venues['Venue Category'].unique().tolist())

         # add Subzone column back to dataframe
         SG_onehot['Subzone'] = SG_venues['Subzone']
         SG_onehot = SG_onehot[['Subzone'] + Venue_categories]

         print("SG_onehot shape: ", SG_onehot.shape)

         # group rows by neighborhood and by taking the mean of the frequency of occurrence of each category
         SG_grouped = SG_onehot.groupby('Subzone').mean().reset_index()

         print("SG_grouped shape: ", SG_grouped.shape)
         SG_grouped.head()

         SG_onehot shape:  (16294, 115)
         SG_grouped shape:  (321, 115)
Out[7]:
```

| | Subzone | African Restaurant | American Restaurant | Argentinian Restaurant | Asian Restaurant | Australian Restaurant | BBQ Joint | Bagel Shop | Bakery | Beijing Restaurant | Belgian Restaurant | Bistro | Breakfast Spot | Buffet | Burger Joint | Burmese Restaurant | Bu F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | ADMIRALTY | 0.0 | 0.000000 | 0.0 | 0.047619 | 0.0 | 0.023810 | 0.0 | 0.071429 | 0.0 | 0.0 | 0.023810 | 0.000000 | 0.023810 | 0.0 | 0.0 | |
| **1** | AIRPORT ROAD | 0.0 | 0.000000 | 0.0 | 0.047619 | 0.0 | 0.000000 | 0.0 | 0.095238 | 0.0 | 0.0 | 0.000000 | 0.142857 | 0.000000 | 0.0 | 0.0 | |
| **2** | ALEXANDRA HILL | 0.0 | 0.010309 | 0.0 | 0.061856 | 0.0 | 0.041237 | 0.0 | 0.072165 | 0.0 | 0.0 | 0.000000 | 0.020619 | 0.010309 | 0.0 | 0.0 | |
| **3** | ALEXANDRA NORTH | 0.0 | 0.000000 | 0.0 | 0.034483 | 0.0 | 0.017241 | 0.0 | 0.068966 | 0.0 | 0.0 | 0.034483 | 0.017241 | 0.000000 | 0.0 | 0.0 | |
| **4** | ALJUNIED | 0.0 | 0.000000 | 0.0 | 0.050000 | 0.0 | 0.040000 | 0.0 | 0.010000 | 0.0 | 0.0 | 0.010000 | 0.010000 | 0.000000 | 0.0 | 0.0 | |

check top 10 restaurant types of each subzone

Out[8]:

| | Subzone | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ADMIRALTY | Chinese Restaurant | Food Court | Fast Food Restaurant | Bakery | Soup Place | Italian Restaurant | Pizza Place | Asian Restaurant | Seafood Restaurant | Japanese Restaurant |
| 1 | AIRPORT ROAD | Food Court | Breakfast Spot | Cafeteria | Bakery | Noodle House | Chinese Restaurant | Restaurant | Asian Restaurant | Indian Restaurant | Fast Food Restaurant |
| 2 | ALEXANDRA HILL | Chinese Restaurant | Food Court | Noodle House | Café | Bakery | Asian Restaurant | BBQ Joint | Indian Restaurant | Sandwich Place | Fast Food Restaurant |
| 3 | ALEXANDRA NORTH | Chinese Restaurant | Café | Noodle House | Bakery | Food Court | Indian Restaurant | Fast Food Restaurant | Asian Restaurant | Hainan Restaurant | Bistro |
| 4 | ALJUNIED | Chinese Restaurant | Noodle House | Food Court | Dim Sum Restaurant | Seafood Restaurant | Asian Restaurant | Vegetarian / Vegan Restaurant | BBQ Joint | Café | Thai Restaurant |

## 3.3 Clustering

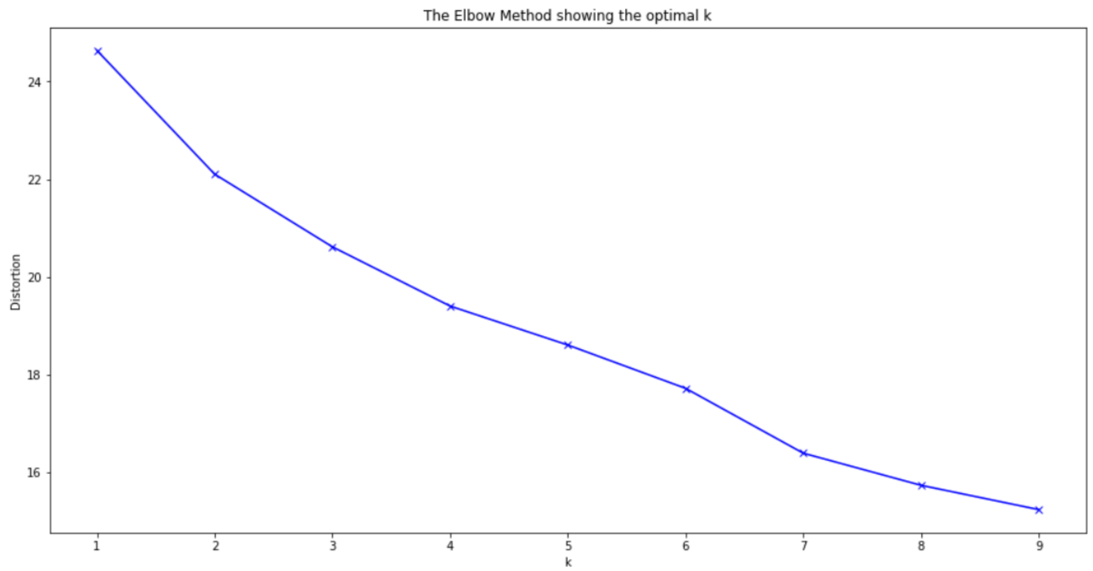
We are going to use K-means clustering method to identify subzones of similar restaurant distributions. As the basic parameter of K-means, we first need to identify the optimal number of clusters in our case. The elbow test is conducted here, the optimal K value is identified as 5.

In [9]:

```
# elbow method of determine K for K means
SG_grouped_clustering = SG_grouped.drop('Subzone', 1)

distortions = []
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(SG_grouped_clustering)
    distortions.append(kmeanModel.inertia_)

plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```




## 4. Results

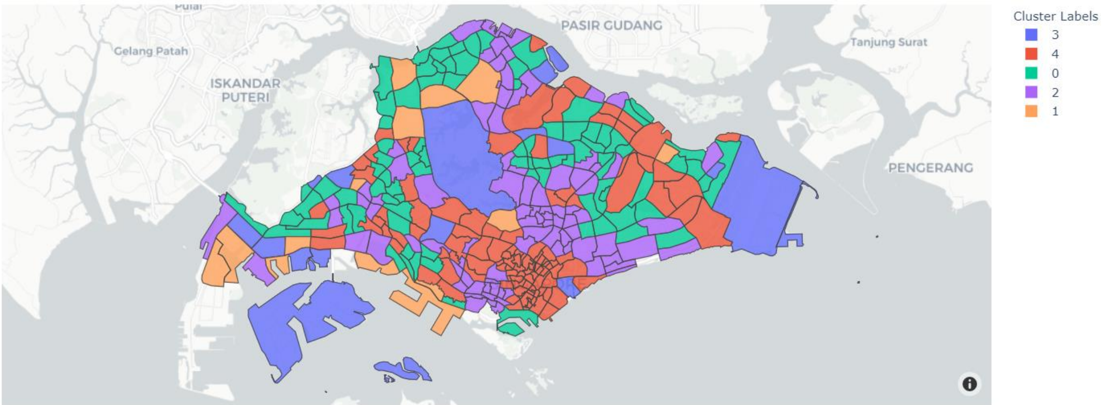
We can create a map to show the clusters of the subzones based on their restaurant distribution. You can click on each subzone to see the subzone name, total venue counts and top 3 types. As shown in the map, the city centre areas are mostly cluster_2 and cluster_4, while the outer ring areas are mostly cluster_0, and areas with smaller population density are cluster_3.

In [12]:
```
# Subzone clusters based on restaurant distribution

map_cluster = px.choropleth_mapbox(SG_merged, geojson=SG_subzone_polygon,
               locations='Subzone', featureidkey="properties.SUBZONE_N",
               color='Cluster Labels',
               hover_data=['Subzone', 'Venue Counts','1st Most Common Venue', '2nd Most Common Venue', '3rd Most Common Venue'],
               mapbox_style="carto-positron",
               zoom=zoom_level,
               center=city_centre,
               opacity=0.8,
               title = "Subzone clusters based on restaurant distributions"
               )

map_cluster.update_layout(height=500, margin={"r":0,"l":0,"b":0})
map_cluster.show()
```

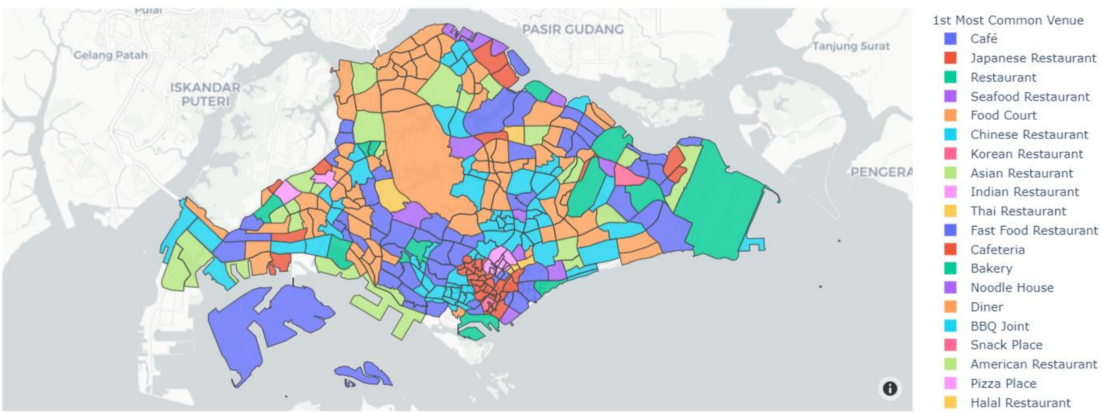
Subzone clusters based on restaurant distributions



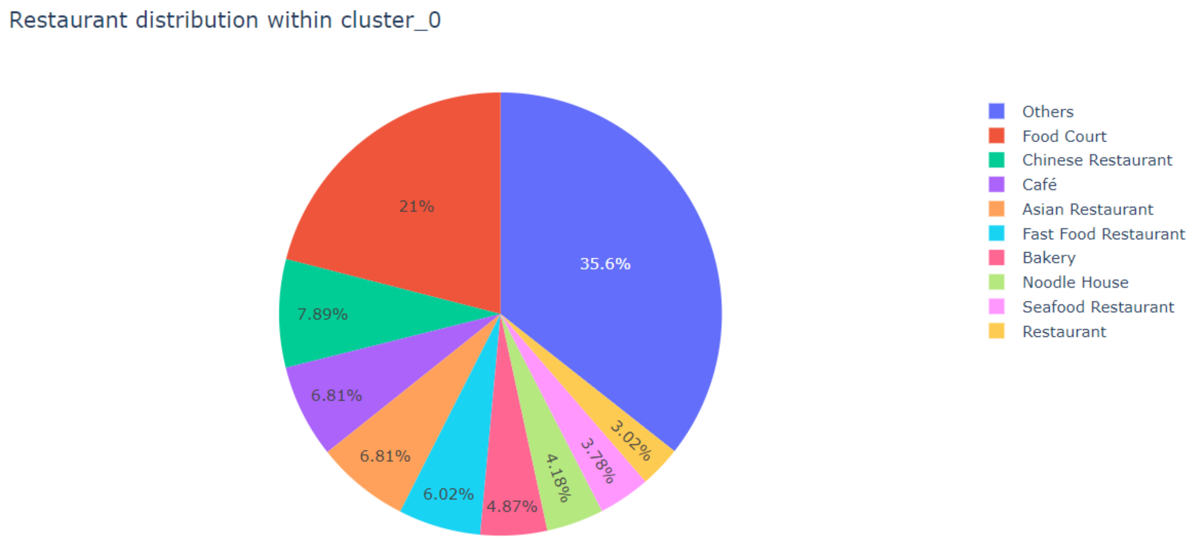

We can also identify the most common restaurant type of each subzone. It can be observed that in the south centre (the CBD area), Japanese restaurants are most common. In the northern part, food court is the most common type, which is a very typical Singaporean place where diverse food choices are provided.

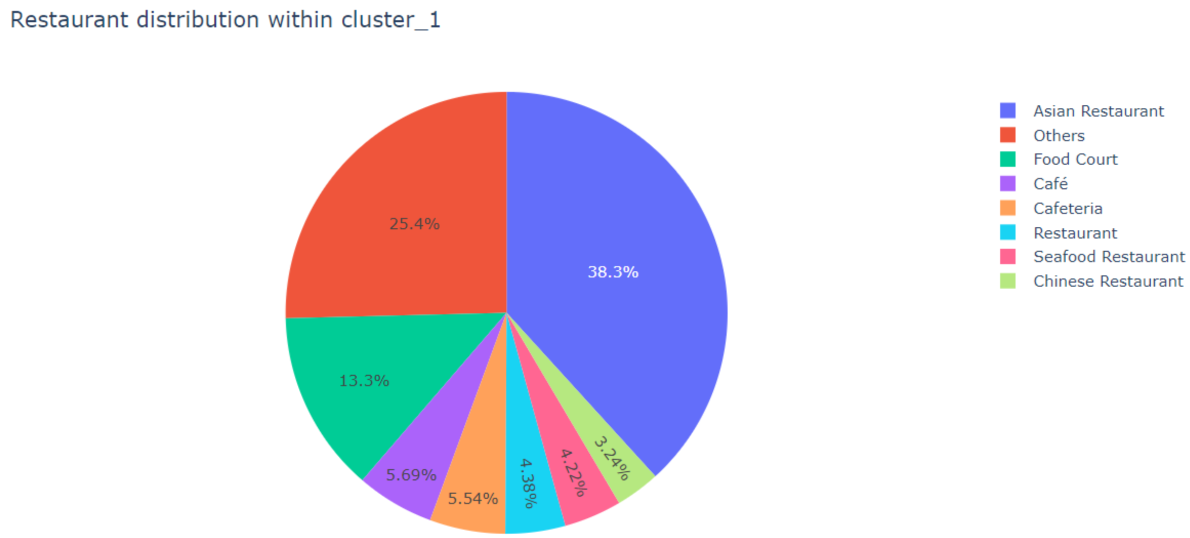
In [13]:
```
# Most common restaurant type of each subzone
map_mostCommon = px.choropleth_mapbox(SG_merged, geojson=SG_subzone_polygon,
               locations='Subzone', featureidkey="properties.SUBZONE_N",
               color='1st Most Common Venue',
               hover_data=['Subzone', '1st Most Common Venue'],
               mapbox_style="carto-positron",
               zoom=zoom_level,
               center=city_centre,
               opacity=0.8,
               title = "Most common restaurant type of each subzone"
               )
map_mostCommon.update_layout(height=500, margin={"r":0,"l":0,"b":0})
map_mostCommon.show()
```

Most common restaurant type of each subzone

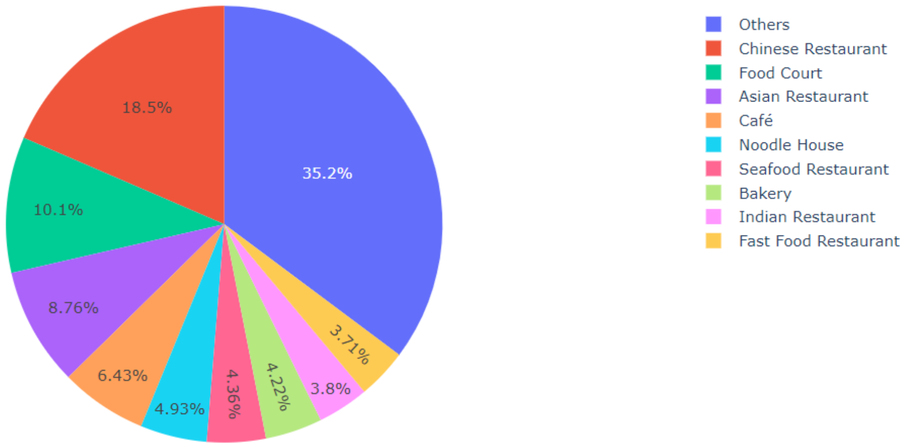


As the subzones have been grouped into 5 clusters, we can look into each of them. The distribution of different restaurant types of each cluster is displayed in the pie charts.

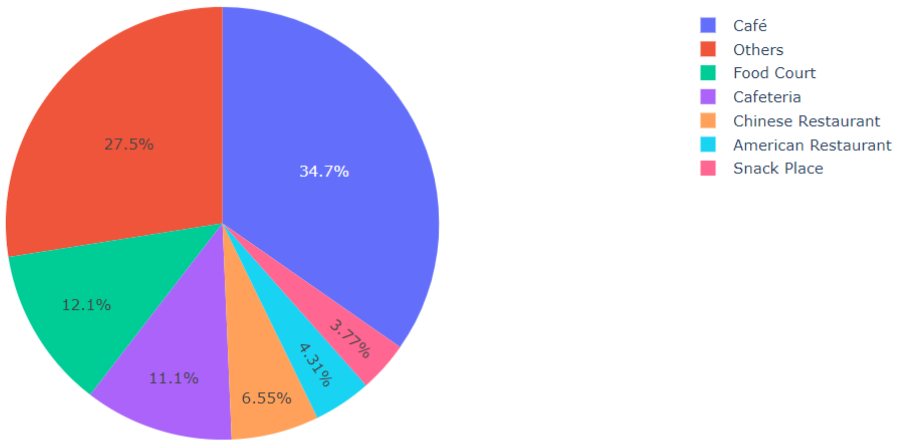


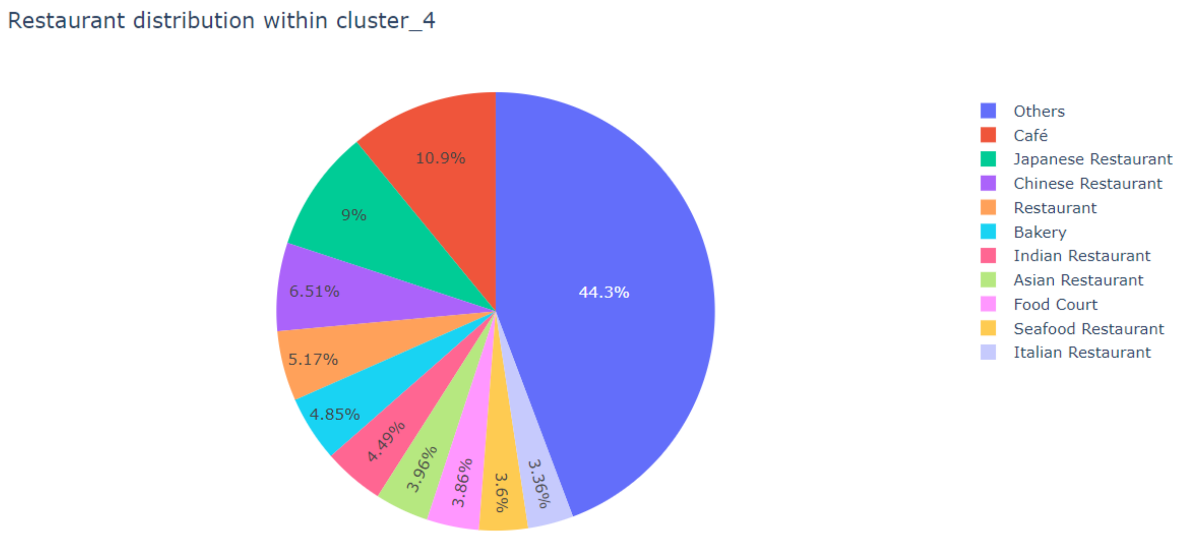



Restaurant distribution within cluster_2




Restaurant distribution within cluster_3






# 5. Conclusion

This project is trying to explore the diversity of restaurants in Singapore and provide an overview of distributions of different restaurant types in different subzones. By analysing the top 100 food-related Foursquare POIs of each subzone, we successfully identified the distribution of different cuisines in the local communities, and observed some interesting spatial patterns of the subzone clusters purely based on food preferences. If someone is interested in opening a restaurant in Singapore, they can use the outcome of this project for reference in popularity of the relevant restaurant types. This is just a practical project with very limited time. However, it can be extended and improved in many aspects.