# Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Ivan Oliveira

## 1. Introduction

### 1.1 Background

The brazilian city São Paulo received a high number of immigrants after the first world war, among them are the japanese. Nowadays Brazil has the largest japanese community outside Japan[1], and there are lots of venues exclusively for japanese people.

### 1.2 Problem

In this project we will try to find an optimal location for a restaurant. This report will be targeted to stakeholders interested in opening an Japanese restaurant in São Paulo, Brazil.Since there are lots of restaurants in São Paulo, we will assume the hypothesis that crowded areas are the best locations for a restaurant, so we will focus in areas near of subway stations.With the aid of our data and data science skills, we will generate a few most promising neighborhoods based on this criteria. We will make clear the advantages of each neighborhood so that the best possible location can be chosen by stakeholders.

## 2. Data

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighborhood (any type of restaurant)
- number of japanese restaurants in the neighborhood

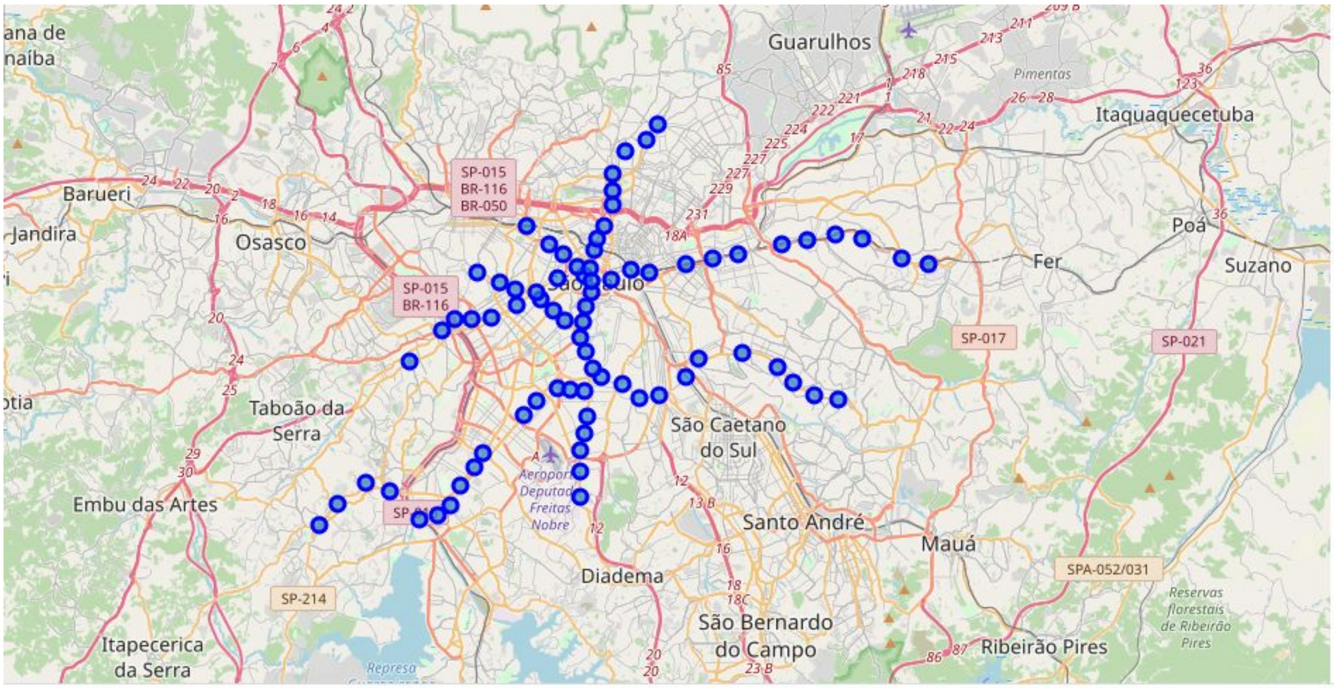
Following data sources will be used to extract/generate the required information:

- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**
- coordinate of São Paulo center will be obtained using **Google Maps API geocoding**
- csv file with coordinates of metro stations obtained on **Kaggle.com**

## 3. Methodology

### 3.1 Data Cleaning

As the first step we will build a DataFrame of all São Paulo's subway stations, then plot in a folium map.




Then using the foursquare API we will build a DataFrame of venues within 500 meters of each subway station, along with their latitude, longitude and venue's type information.

| | Stations | Station Latitude | Station Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Aacd Servidor | -23.597825 | -46.652374 | Quitanda da Cerveja | -23.597729 | -46.650111 | Brewery |
| 1 | Aacd Servidor | -23.597825 | -46.652374 | Lilló Restaurante e Pizzaria | -23.596710 | -46.649572 | Brazilian Restaurant |
| 2 | Aacd Servidor | -23.597825 | -46.652374 | Ginásio Mané Garrincha | -23.598493 | -46.652815 | Athletics & Sports |
| 3 | Aacd Servidor | -23.597825 | -46.652374 | COTP - Centro Olímpico de Treinamento e Pesquisa | -23.598675 | -46.655799 | Athletics & Sports |
| 4 | Aacd Servidor | -23.597825 | -46.652374 | Grill Hall Prazeres da Carne | -23.598110 | -46.650314 | Churrascaria |
| 5 | Aacd Servidor | -23.597825 | -46.652374 | Gola Solta | -23.599186 | -46.647956 | Bar |
| 6 | Aacd Servidor | -23.597825 | -46.652374 | Trivial do Chef | -23.596172 | -46.649123 | Restaurant |

To apply the KMeans algorithm we need a numeric DataFrame, so we build it using the function get_dummies on the above DataFrame, specifying the 'Venue Category' column, then we group by ‘Stations’ and take the mean, resulting on the below DataFrame;

```
In [18]: sp_venues_grouped = sp_venues_onehot.groupby('Stations').mean().reset_index()
         sp_venues_grouped
```

| | Stations | Acai House | Accessories Store | American Restaurant | Arcade | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Arts & Entertainment | ... | Video Game Store | Vietnamese Restaurant | Warehouse Store | Wine Bar | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Aacd Servidor | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |
| 1 | Adolfo Pinheiro | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.058824 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |
| 2 | Alto Da Boa Vista | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.024390 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |
| 3 | Alto Do Ipiranga | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |
| 4 | Ana Rosa | 0.000000 | 0.018868 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |
| 5 | Anhangabau | 0.000000 | 0.000000 | 0.010000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.010000 | 0. |
| 6 | Armenia | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0. |

```
In [19]: sp_venues_grouped.shape
```

After some work, we will gather those informations and build a DataFrame of stations alongside with the most common venues.

| | Stations | 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 | Aacd Servidor | Brazilian Restaurant | Café | Restaurant | Coffee Shop | Bakery | Athletics & Sports | Churrascaria | Sports Club | Sandwich Place | Brewery |
| 1 | Adolfo Pinheiro | Farmers Market | Restaurant | Clothing Store | Arts & Crafts Store | Theater | Rental Car Location | Beer Bar | Gym / Fitness Center | Snack Place | Concert Hall |
| 2 | Alto Da Boa Vista | Plaza | Pharmacy | Pizza Place | Coffee Shop | Gym Pool | Gym / Fitness Center | Restaurant | Farmers Market | Bar | Pet Store |
| 3 | Alto Do Ipiranga | Bar | Pet Store | Brazilian Restaurant | Gym | Pizza Place | Seafood Restaurant | Dessert Shop | Pharmacy | Convenience Store | Burger Joint |
| 4 | Ana Rosa | Bakery | Ice Cream Shop | Bar | Dessert Shop | Brazilian Restaurant | Café | Japanese Restaurant | Vegetarian / Vegan Restaurant | Cosmetics Shop | Pet Store |
| 5 | Anhangabau | Brazilian Restaurant | Café | Coffee Shop | Bar | Vegetarian / Vegan | Restaurant | Theater | Cultural Center | Bookstore | Burger Joint |

### 3.2 The KMeans algorithm and the Silhouette Score to choose the best ‘K’

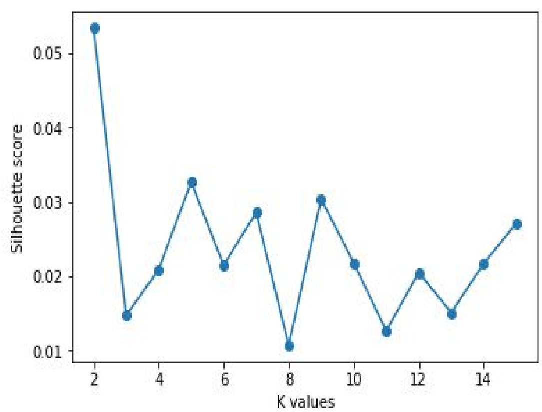
First we will scale the sp_venues_grouped panda’s dataframe using the function MinMaxScaler(). Our first job is choose the best number of clusters (K), so we use the Silhouette Score method described below;

```python
silhouette = []
K = np.arange(2, 16)

for k in K:
    kmean = KMeans(n_clusters=k, random_state=1)
    labels_pred = kmean.fit_predict(sp_venues_grouped_scaled)
    silhouette.append(silhouette_score(sp_venues_grouped_scaled, labels_pred))
```

```python
#As shown on the plot, the optimal value is 6 for K
plt.plot(K, silhouette, 'o-')
plt.xlabel('K values')
plt.ylabel('Silhouette score')
```

```
Text(0, 0.5, 'Silhouette score')
```




The best choice of clusters is 6, so running a KMeans algorithm we will get cluster labels that we use to build the new DataFrame below;

```python
k = 6
kmeans = KMeans(n_clusters=k, random_state=1)
labels = kmeans.fit_predict(sp_venues_grouped_scaled)
```

```python
stations_venues_sorted.insert(0, 'Clusters labels', labels)
```

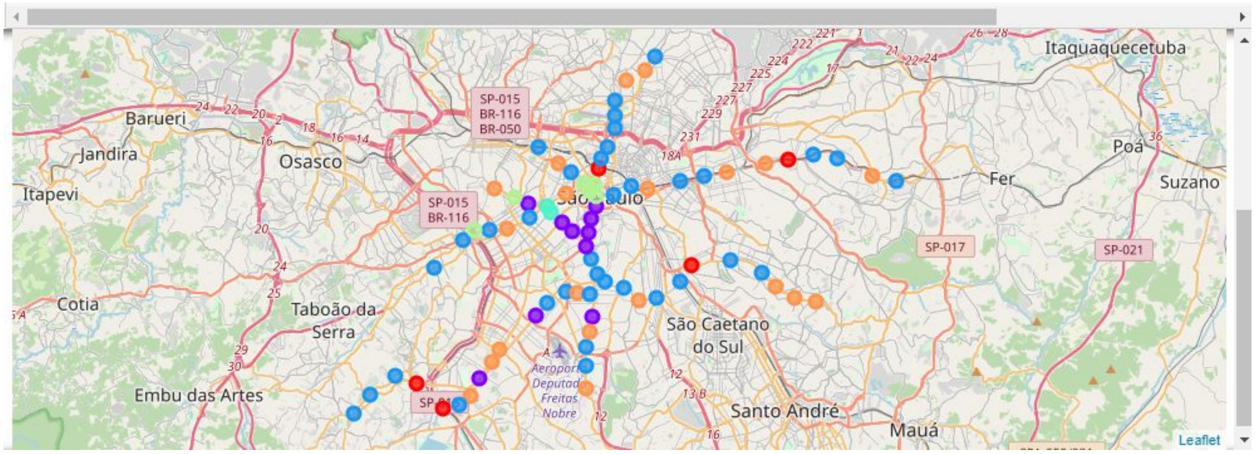
```python
metros.columns = ['Stations', 'Latitude', 'Longitude', 'Line', 'Neighborhood']
```

```python
clusters_merged = metros.join(stations_venues_sorted.set_index('Stations'), on='Stations')
```

```python
clusters_merged
```

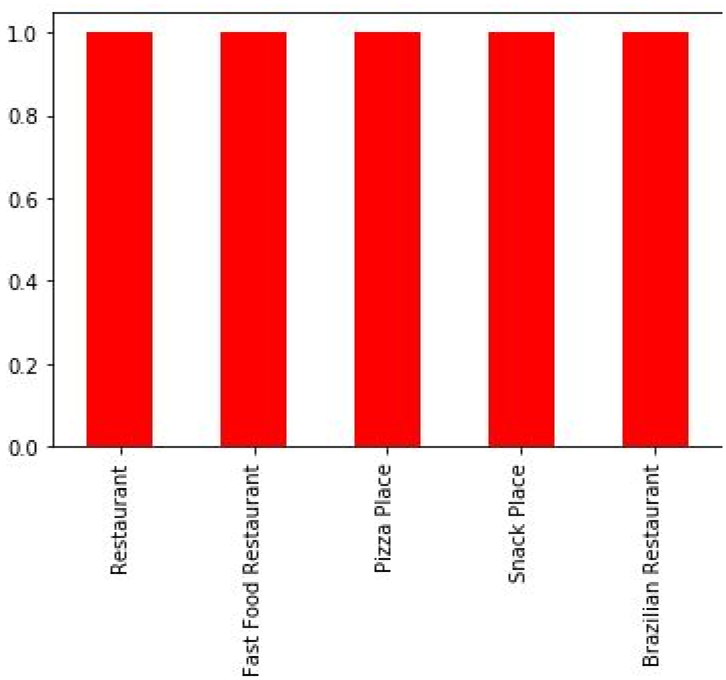
| | Stations | Latitude | Longitude | Line | Neighborhood | Clusters labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Aacd Servidor | -23.597825 | -46.652374 | ['lilas'] | ['moema', 'hospital-sao-paulo'] | 2 | Brazilian Restaurant | Café | Restaurant | Coffee Shop | Bakery | Athletics & Sports | Chu |
| 1 | Adolfo Pinheiro | -23.650073 | -46.704206 | ['lilas'] | ['largo-treze', 'alto-da-boa-vista'] | 5 | Farmers Market | Restaurant | Clothing Store | Arts & Crafts Store | Theater | Rental Car Location | |
| 2 | Alto Da Boa Vista | -23.641625 | -46.699434 | ['lilas'] | ['adolfo-pinheiro', 'borba-gato'] | 1 | Plaza | Pharmacy | Pizza Place | Coffee Shop | Gym Pool | Gym / Fitness Center | R |
| 3 | Alto Do Ipiranga | -23.602237 | -46.612486 | ['verde'] | ['santos-imigrantes', 'sacoma'] | 5 | Bar | Pet Store | Brazilian Restaurant | Gym | Pizza Place | Seafood Restaurant | |

Now we can plot the clusters using a folium map to visualize;

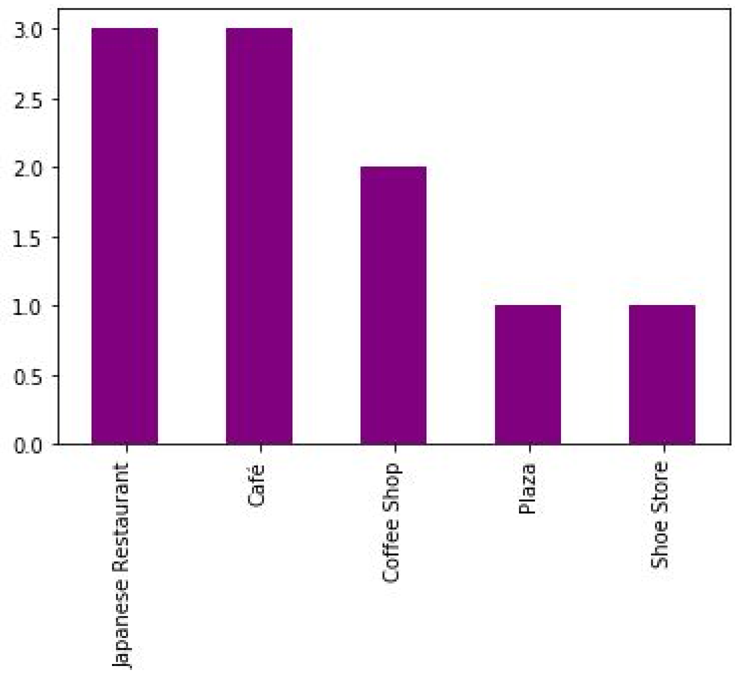



Plotting the top venues in each cluster we can get some insights;

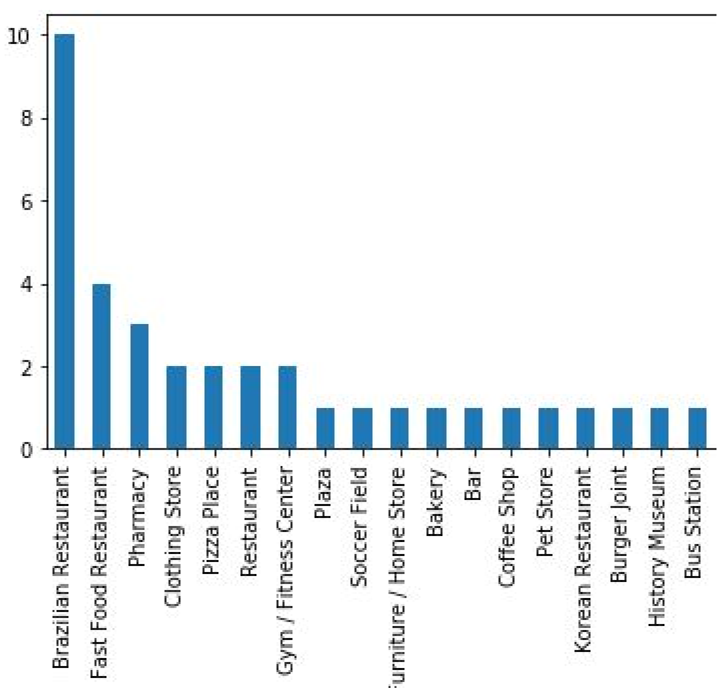
**Cluster 0**



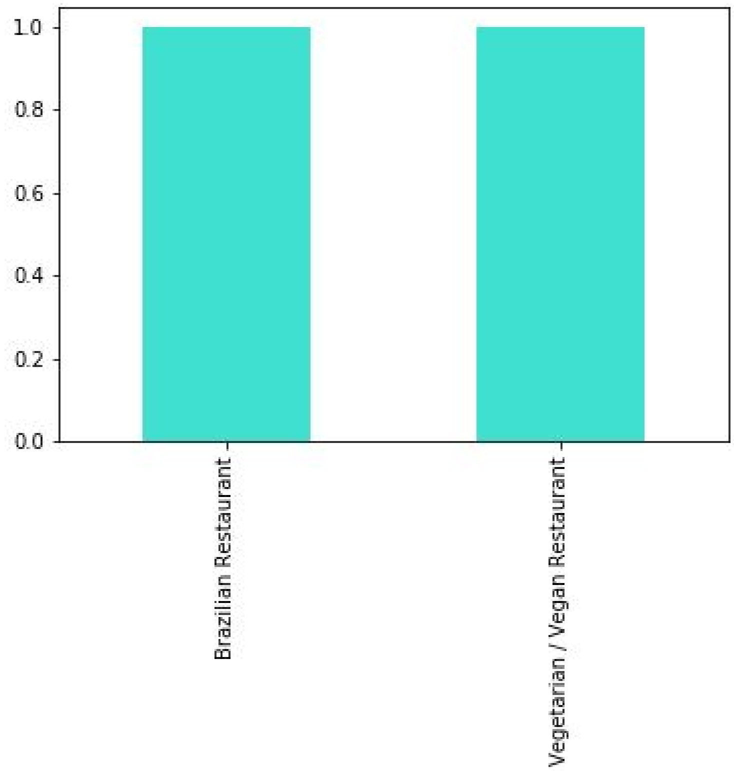
## Cluster 1



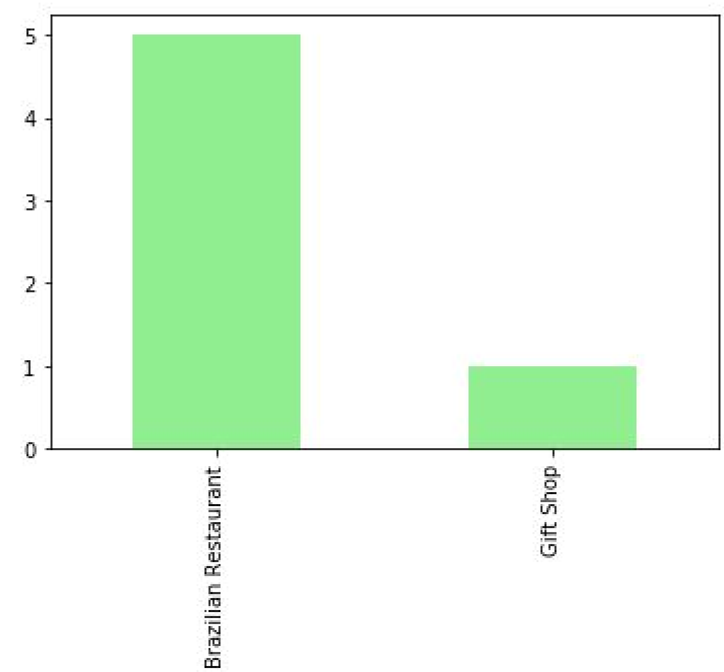

## Cluster 2



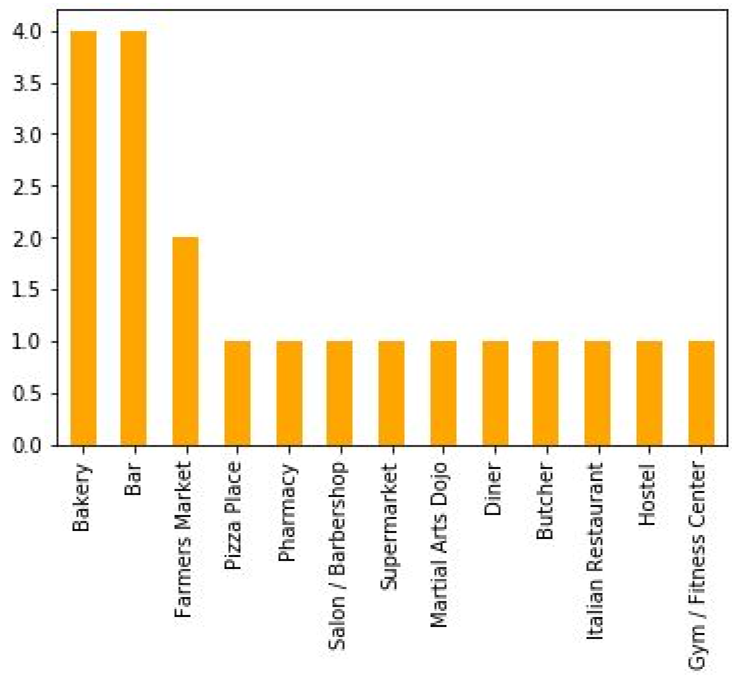
**Cluster 3**




**Cluster 4**



**Cluster 5**




## 4. Results

As we can see from the bar plots there are lots of food venues near São Paulo's subway stations, we can see that cluster 1 has the high concentration of japanese restaurants, so at first this shows that those subway stations have a high competition to our venue. With only the data provided by Foursquare API we can not make a decision, but we can argue that cluster 1 has 3 stations near of the largest japanese neighborhood in São Paulo (Liberdade, Praça da Árvore and São Joaquim), so we would have high number of clients. Using more data about those places we can refine our choice.

## 5. Discussion

The cluster 1 has a high number of japanese restaurants, so at first is advisable work with the data about cluster's 1 stations, inspec the restaurants location's is a good start point.

## 6. Conclusion

As a result we saw that there are lots of food venues near subway stations in São Paulo, but we found a cluster with a high number of Japanese food venues, so a further inspection of subway stations of this cluster can lead to a better insight.