Project to evaluate business opportunities in the bakery sector near subway stations located in the south zone of Rio de Janeiro

Renan Massena de Oliveira
01/05/2022

Executive Summary

The project evaluates the possibility of entrepreneurship in the bakery business near subway stations in Rio de Janeiro, taking into account geolocation data of potential customers and competitors.

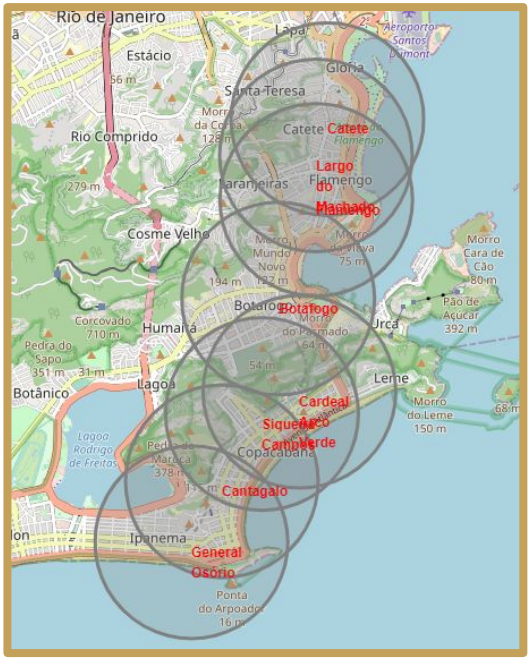
Will be used Web Scrapping techniques and data clustering with K-means, taking into account the Standard Scalar and Min Max Scalar methods. To determine the appropriate number of clusters the elbow and silhouette coefficient methods were used.

The models developed were satisfactory for analysis.

Introduction

The subway stations in Rio de Janeiro are characterized by the high flow of passers-by, and by connecting various neighborhoods and service networks these points can present opportunities for entrepreneurs in the food business.

The south zone of Rio de Janeiro was used as a study reference because it is an area that receives many tourists and citizens of the city, because it is a beach area, safe, and with wide coverage of services.



Methodology

Data Collection

The data was collected through Fousquare's API taking into consideration bakeries within a 1500 meter radius of the metro stations of Catete, Largo do Machado, Flamengo, Botafogo, Cardeal Arcoverde, Siqueira Campos, Cantagalo, and General Osório.

To better understand the market in which these bakeries are inserted, a survey was conducted of businesses within a 500-meter radius in the areas of dining and drinking, retail, commercial and professional services, landmarks and outdoors, arts and entertainment, sports and recreation, community and government, and travel and transportation

To verify the business opportunities with the public of other types of establishments and also possible competitors.

Data Handling

The work of extracting and treating the information that was consolidated in rows was carried out, this information was consolidated in another notebook and in it the distance of the bakeries from the subway stations was calculated and the duplicates with the greatest distance to other stations were excluded.

After cleaning, a new query was performed in Fousquare, searching the businesses near the bakeries, and for this it was necessary to do a new data treatment to group the subcategories referenced in the collection with the categories informed on the API website.

```
categorias = {}
for id in df_places.columns.values.tolist()[1:]:
    if (str(id)[:2]) not in categorias:
        categorias[str(id)[:2]] = []
        categorias[str(id)[:2]].append(id)
    else:
        categorias[str(id)[:2]].append(id)
```

```
df_categorias = pd.DataFrame()
for keys, values in Categorias.items():
    df_categorias[keys] = df_places[values].sum(axis=1)
```

```
df_categorias.rename(columns={'13' : 'Refeições e Bebidas',
                              '17' : 'Varejo',
                              '11' : 'Serviços comerciais e profissionais',
                              '16' : 'Pontos de referência e ao ar livre',
                              '10' : 'Arts and Entertainment',
                              '18' : 'Esportes e Recreação',
                              '12' : 'Comunidade e Governo',
                              '19' : 'Viagens e Transporte',
                              '15' : 'Saúde e medicina'},inplace=True)
```

Data Handling

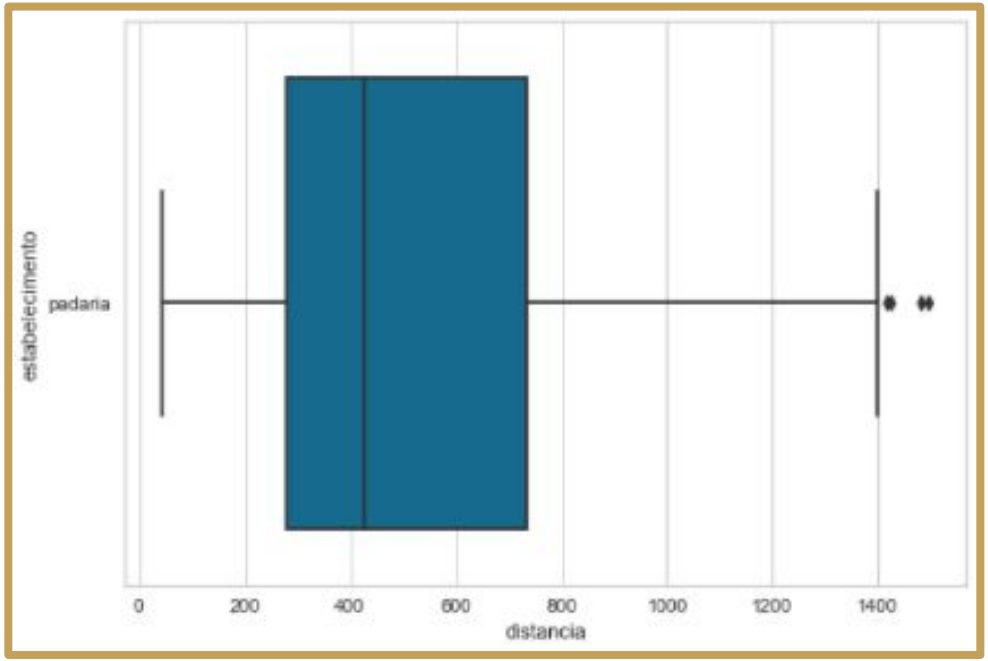
After consolidating the categories, the remaining columns were included to form the dataframe with the necessary information to begin the data analysis phase of the 199 bakeries evaluated.

```
['name', 'endereço', 'estabelecimento', 'latitude', 'longitude',
 'estacao_nome', 'estacao_lat', 'estacao_lon', 'distancia',
 'Refeições e Bebidas', 'Varejo', 'Serviços comerciais e profissionais',
 'Pontos de referência e ao ar livre', 'Arts and Entertainment',
 'Esportes e Recreação', 'Comunidade e Governo', 'Viagens e Transporte',
 'Saúde e medicina'],
```

Data Analysis

Data Analysis

I started the data analysis step using the boxplot graph to visualize the proximity of the bakeries to the subway stations, where the horizontal axis presents the distance by meters from the stations.

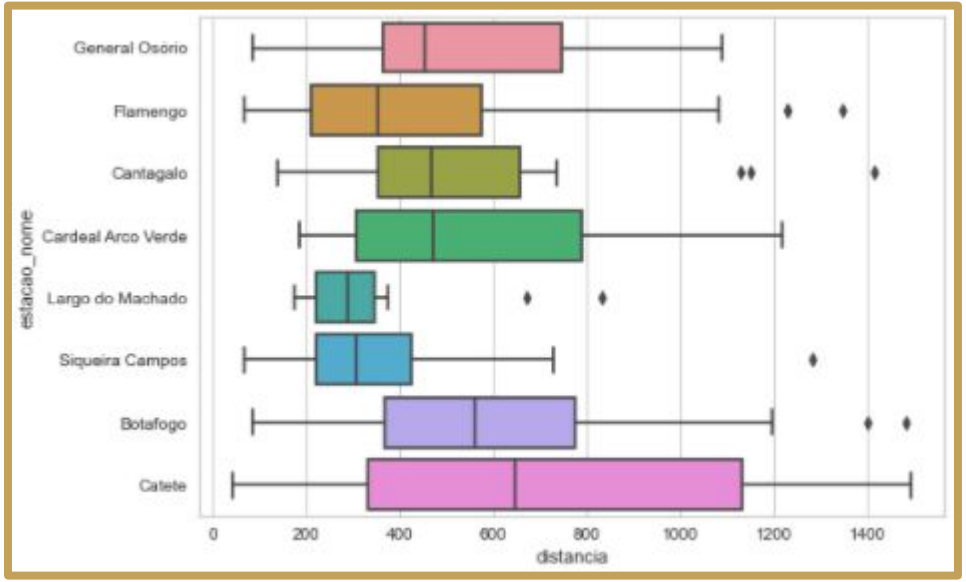


boxplot - distribution of distance between companies and nearest stations

Data Analysis

To better understand the characteristics of the neighborhoods studied, an analysis of the distribution of the distance of bakeries in relation to subway stations was carried out, using the boxplot graph.

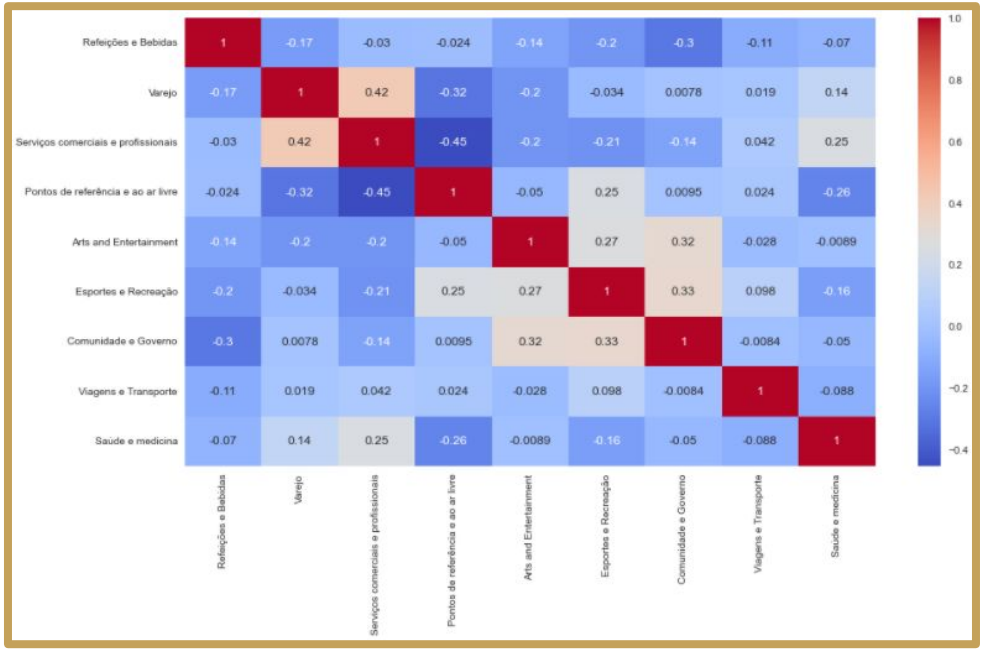
We can notice a difference between the Catete and Largo do Machado stations, which, although they are subsequent stations, have distinct profiles.



boxplot - distance of companies from nearest stations

Data Analysis - Correlation Map

To understand the possible correlation between the variables, the heat map correlation function was used.



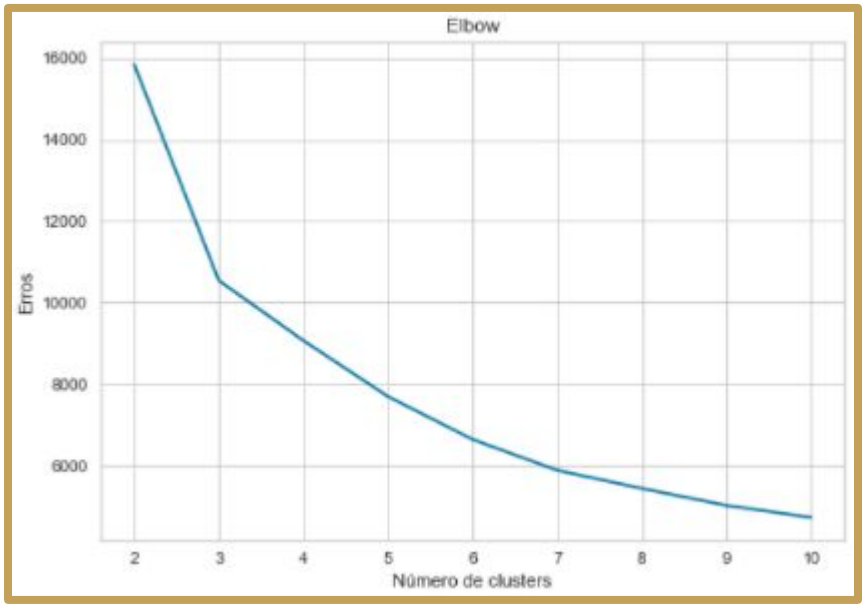
Handling data for Clustering

Handling data for Clustering - K-Means - Elbow method

Next, the simulation function of the clusters was performed using the K-means method, and then a graph was made using the elbow method.

```
errors = []
for i in range(2, 11):
  km = KMeans(n_clusters=i)
  km.fit(df_dum)
  errors.append(km.inertia_)
```

```
plt.plot(range(2,11), errors)
plt.title('Elbow')
plt.xlabel('Número de clusters')
plt.ylabel('Erros')
plt.show()
```



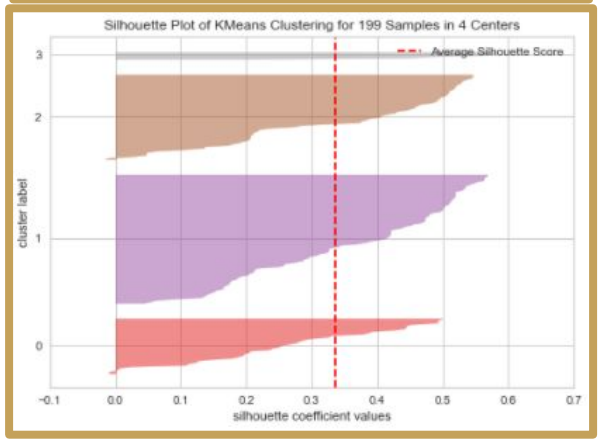
Handling data for Clustering - K-Means - Silhouette Visualizer

To facilitate the analysis the Silhouette Visualizer function was used, which presented the following graphs.

When comparing the result of the distributions and other cluster simulations this model presented a satisfactory distribution of the groups with 4 clusters.

```
for i in range(2,11):
    kmeans = KMeans(n_clusters=i)
    yellow_visualizer = SilhouetteVisualizer(kmeans)
    yellow_visualizer.fit(df_dum)
    yellow_visualizer.show()
```

4 centers

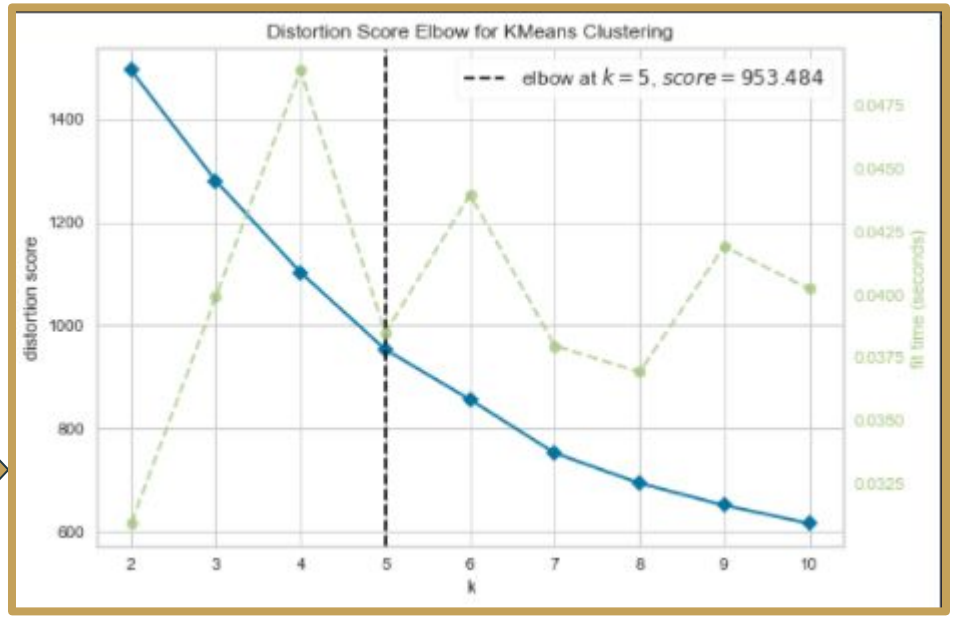


Handling data for Clustering - Standard Scaler - Elbow method

At first this method of standardization proved satisfactory, however, it will be necessary to analyze the silhouette method to understand the best number of clusters.

```
std = StandardScaler()
df_std = std.fit_transform(df_dumm)
df_std = pd.DataFrame(df_std, columns=df_dumm.columns)
```

```
kmeans = KMeans()
yellow_visualizer = KElbowVisualizer(kmeans, k=(2,11))
yellow_visualizer.fit(df_std)
yellow_visualizer.show()
```

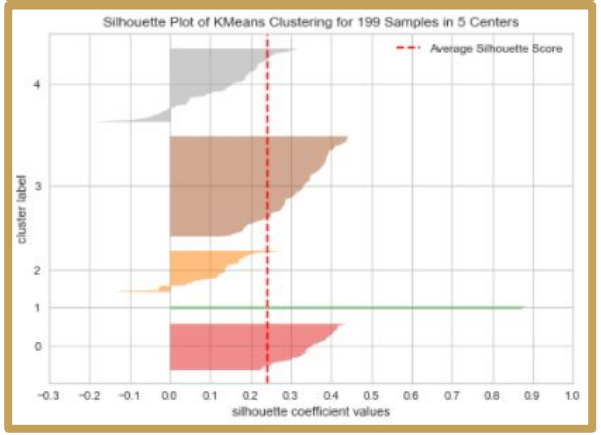


Handling data for Clustering - Standard Scaler - Silhouette Visualizer

The Silhouette Visualizer function shows that the cluster groupings are not performed similarly. And that is why this standardization method was not adopted.

```
for i in range(2,12):
  kmeans = KMeans(n_clusters=i)
  yellow_visualizer = SilhouetteVisualizer(kmeans)
  yellow_visualizer.fit(df_std)
  yellow_visualizer.show()
```

5 centers

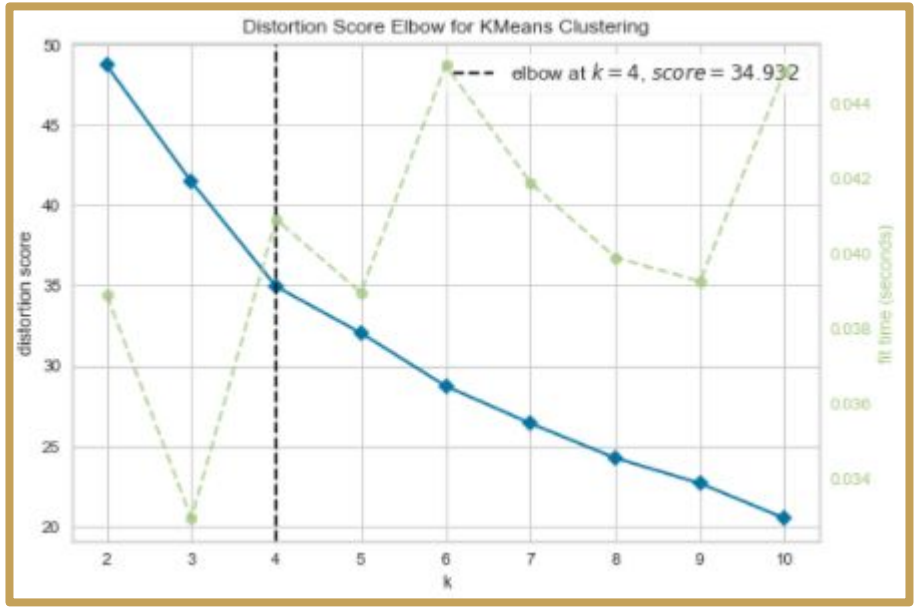


Handling data for Clustering - Min Max - Elbow method

This method showed that the best clustering would be with 6 groups, to analyze if this distribution is similar we will see the Silhouette Visualizer tool on the next slide.

```
mmax = MinMaxScaler()
df_mmax = mmax.fit_transform(df_dumm)
df_mmax = pd.DataFrame(df_mmax, columns=df_dumm.columns)
```

```
kmeans = KMeans()
yellow_visualizer = KElbowVisualizer(kmeans, k=(2,11))
yellow_visualizer.fit(df_mmax)
yellow_visualizer.show()
```

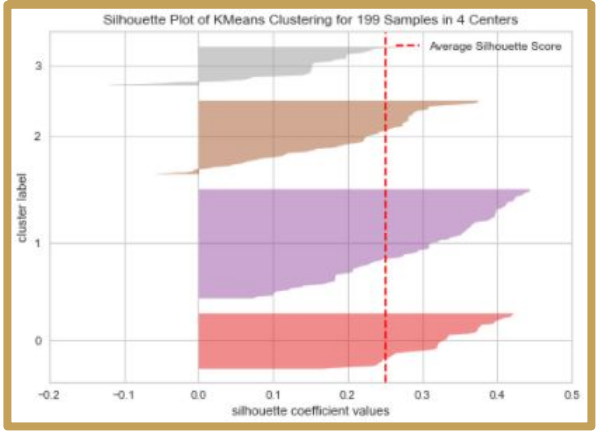


Handling data for Clustering - Min Max - Silhouette Visualizer

The Silhouette Visualizer function shows that the cluster groupings are not performed similarly. And that is why this standardization method was not adopted.

```
for i in range(2,11):
  kmeans = KMeans(n_clusters=i)
  yellow_visualizer = SilhouetteVisualizer(kmeans)
  yellow_visualizer.fit(df_mmax)
  yellow_visualizer.show()
```

4 centers

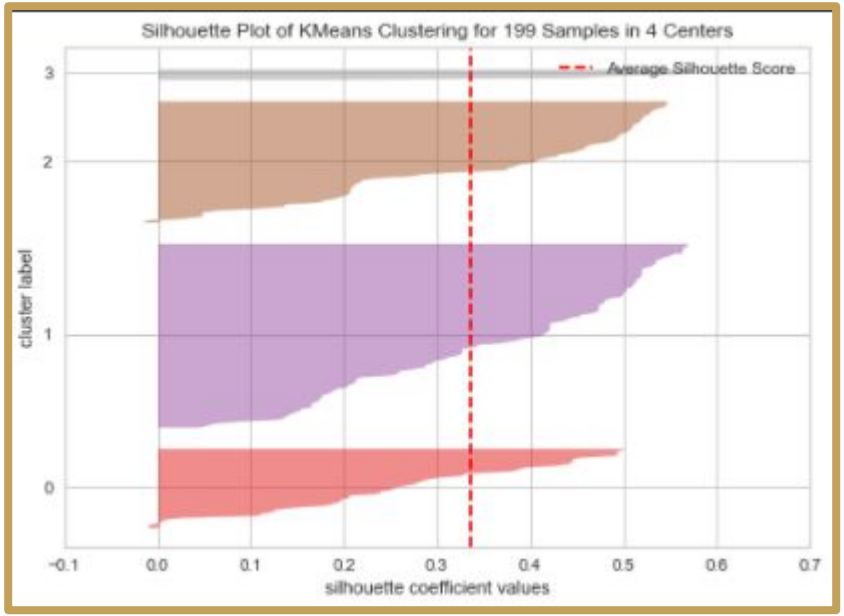


Handling data for Clustering - Selected model

After analyzing the information demonstrated above, simulations of various models were performed to better understand the distribution of establishments and their attributes by groups.

After this conference it was chosen the clustering model with **4 clusters**. Without using the Standard Scaler and the Min Max.

```
kmeans = KMeans(n_clusters=4)
clusters = kmeans.fit_predict(df_dum)
```



Qualify the Clusters

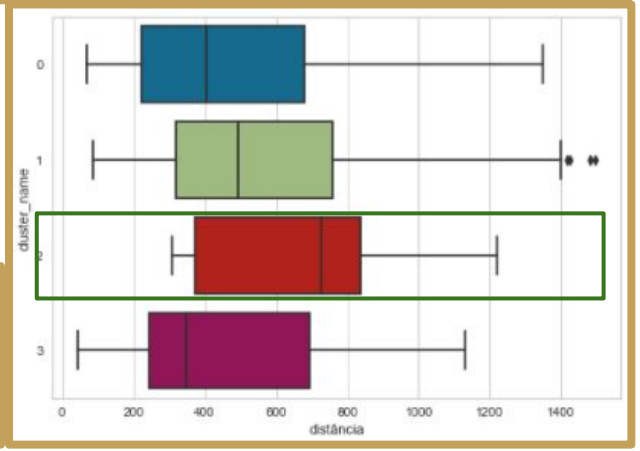
```
df.groupby('cluster')[['Refeições e Bebidas', 'Varejo', 'Serviços comerciais e profissionais',
       'Pontos de referência e ao ar livre', 'Arts and Entertainment',
       'Esportes e Recreação', 'Comunidade e Governo', 'Viagens e Transporte',
       'Saúde e medicina']].mean().T
```

cluster	estabelecimento	
0	padaria	87
1	padaria	54
2	padaria	11
3	padaria	47

cluster	0	1	2	3
Refeições e Bebidas	47.678161	58.537037	23.636364	37.340426
Varejo	9.873563	5.722222	3.818182	10.595745
Serviços comerciais e profissionais	5.367816	4.074074	2.272727	5.510638
Pontos de referência e ao ar livre	1.689655	2.685185	3.818182	1.787234
Arts and Entertainment	4.287356	3.277778	3.363636	5.297872
Esportes e Recreação	0.931034	0.888889	1.363636	1.340426
Comunidade e Governo	0.068966	0.092593	0.272727	0.382979
Viagens e Transporte	0.034483	0.018519	0.000000	0.212766
Saúde e medicina	0.206897	0.148148	0.000000	0.319149

Observing the distribution of businesses and competition, cluster number two was chosen, by comparison with the average of the other clusters it represents an area with good service coverage and little competition, however these locations are more distant from the subway stations.

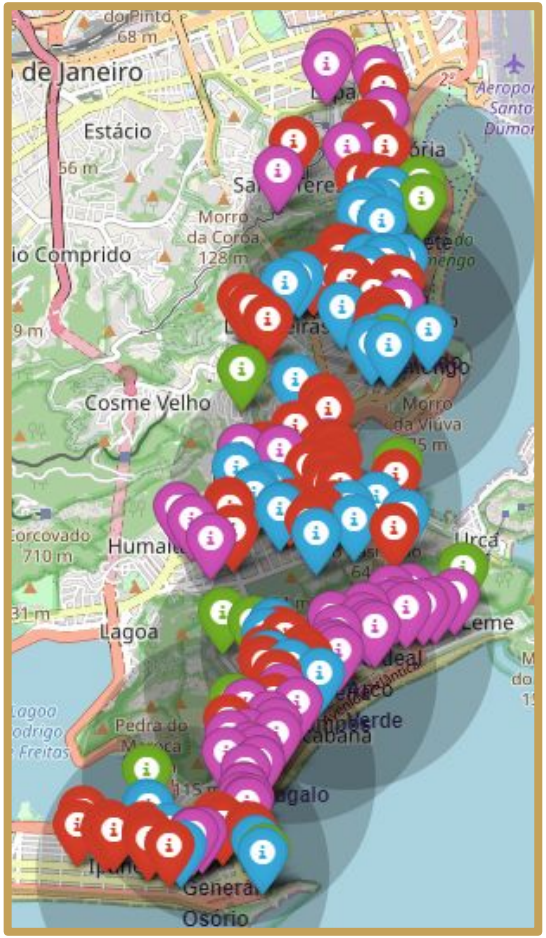
```
cluster_name = {0 : '- serviços / ++ concorrência',
                1 : '++ serviços / + concorrência',
                2 : '+ serviços / - concorrência',
                3 : '+ serviços / + concorrência'}
```



Representation of clusters by geolocation

On the map the clusters are represented by the following colors:

0 = Red (‘- services / ++ competition’);
1 = Purple (‘++ services / + competition’);
2 = Green (‘+ services / - competition’);
3 = Blue (‘+ services / + competition’);



Conclusion

The study used a geolocation API to identify possible locations for opening a bakery, and from the data collected we identified that areas a little far from the metro can offer good opportunities for entrepreneurship, because there is a good supply of services and less competition.