Capstone Project- The Battle of Neighborhoods

Project Description

XPTO It is a real estate investment company, and in 2021 it intends to invest in the construction of several housing projects.

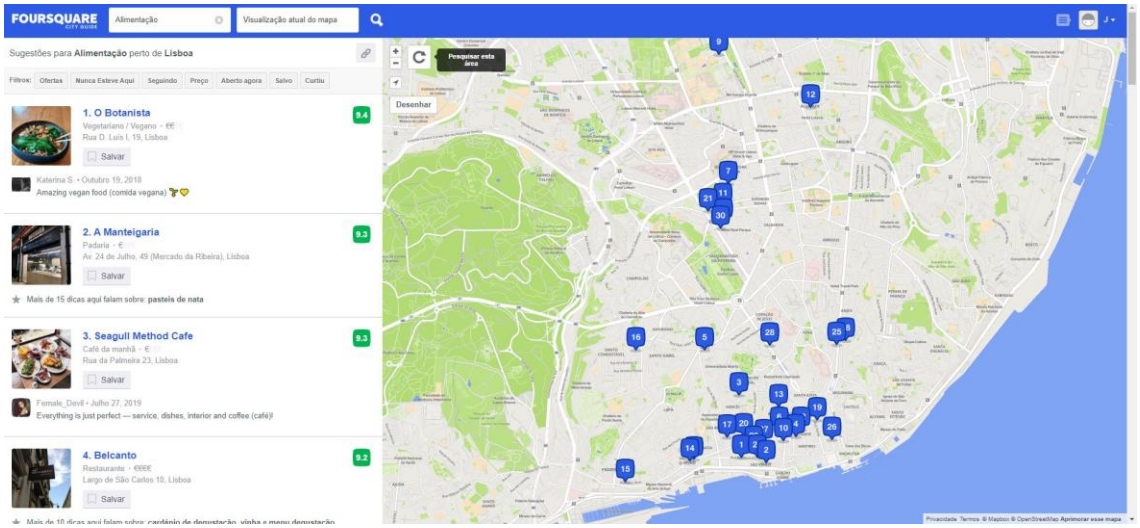
With this project we intend to help the XPTO company to choose which neighborhood has the most potential based on the distribution of various facilities available around the neighborhood. This project will use K-mean clustering unsupervised machine learning algorithm to cluster the venues based on the place category such as restaurants, parks, gyms, theaters clubs etc... In order to give a better understanding of the similarities and dissimilarities between the chosen neighborhoods to retrieve more insights and to conclude with ease which neighborhood wins over other

Data Sources

• Lisbon City Neighbourhoods:
Using Lisbon city neighborhoods zip codes (by a CSV) to pull the Latitude and Longitude information.

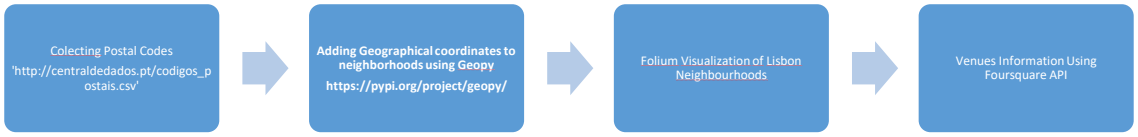
• Foursquare API:
Foursquare API has a databe with more that 115 million places, and provides the ability to

perform location search, location sharing and details about a business. This feature enabled to collect the nearby places of the neighborhoods. Due to API request limitations for the free plan, the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.



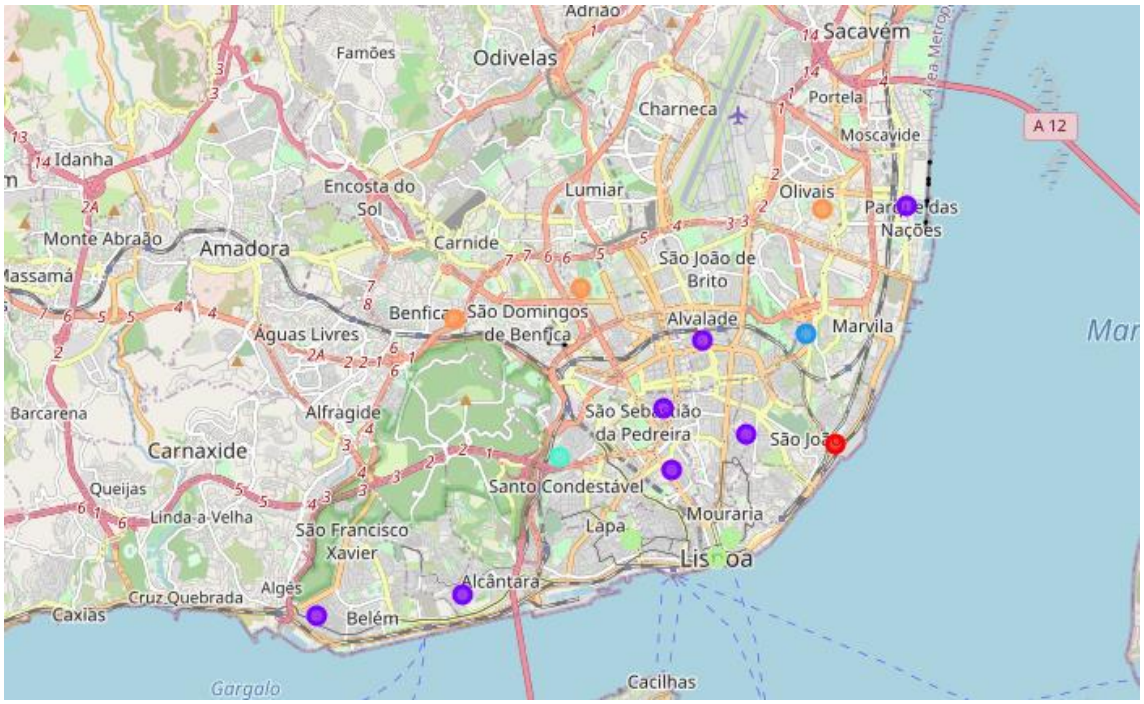
Data acquisition

The zipcodes were acquired through a CSV, and then properly treated. Then the coordinates information is added to the neighborhoods using the geopy (Python Geocoding toolbox)
The Foursquare API would enable to collect the nearby neighborhoods places information (type of venue, Latitude, Longitude, etc…). Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 700.



Exploratory Data Analysis

With Folium (Python visualization library) would be used to visualize the neighborhoods cluster distribution of Lisbon over an interactive leaflet map.

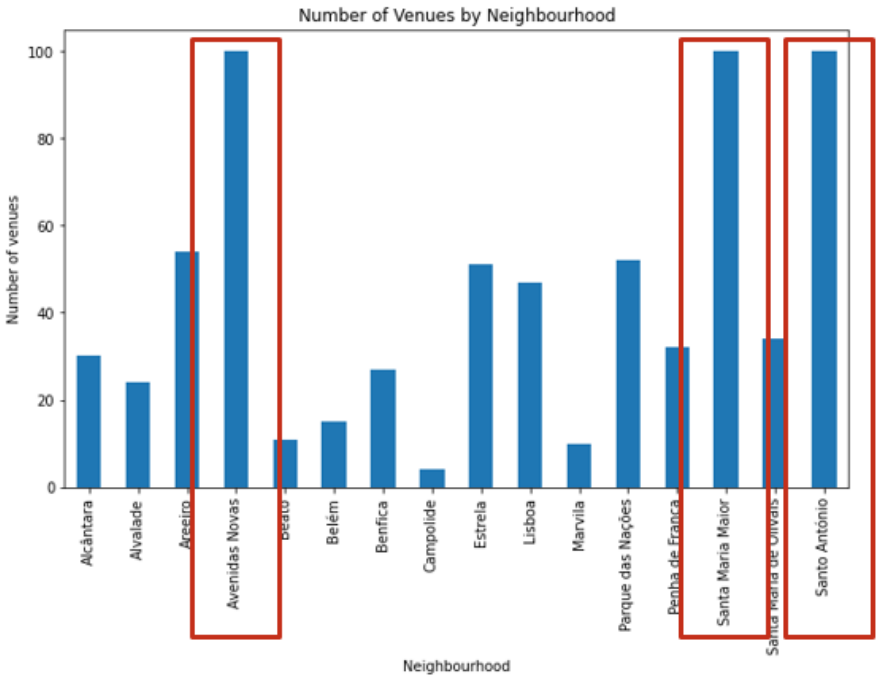


Extensive comparative neighborhoods analysis world be carried out to drive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.

Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions.

Results

Of all the neighborhoods analyzed, there are 3 that stand out for the quantity and variety of venues. These Neighborhoods are: Avenidas Novas, Santa Maria Maior and Santo Antonio

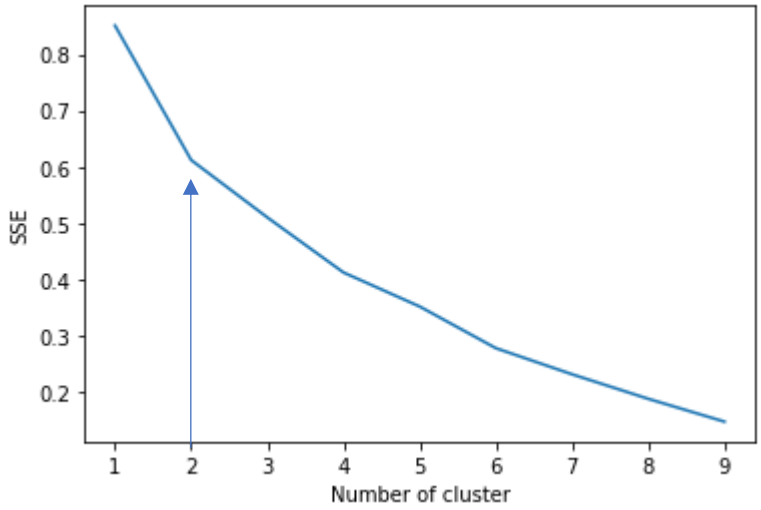


The top 5 of each neighborhood has a great similarity of venues: Restaurants, Cafes and Hotels, which means that there is no major disruption between the clusters

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alcântara	Restaurant	Snack Place	Soccer Field	Portuguese Restaurant	Café
1	Alvalade	Café	Sushi Restaurant	Portuguese Restaurant	College Cafeteria	Gym
2	Areeiro	Portuguese Restaurant	Café	Bar	BBQ Joint	Restaurant
3	Avenidas Novas	Restaurant	Hotel	Café	Bakery	Hostel
4	Beato	Restaurant	Dance Studio	Supermarket	Cafeteria	Museum
5	Belém	Portuguese Restaurant	Pizza Place	Japanese Restaurant	Bakery	Restaurant
6	Benfica	Café	Portuguese Restaurant	Bakery	Pizza Place	Gym
7	Campolide	Ice Cream Shop	Soccer Field	Hockey Arena	Bakery	Farm
8	Estrela	Portuguese Restaurant	Restaurant	Italian Restaurant	Coffee Shop	BBQ Joint
9	Lisboa	Portuguese Restaurant	Bar	Hostel	Plaza	Wine Bar
10	Marvila	Chinese Restaurant	Café	Electronics Store	Shopping Mall	Italian Restaurant
11	Parque das Nações	Ice Cream Shop	Electronics Store	Burger Joint	Gym / Fitness Center	Clothing Store
12	Penha de França	Hotel	Café	Hostel	Portuguese Restaurant	Chinese Restaurant
13	Santa Maria Maior	Portuguese Restaurant	Hotel	Hostel	Wine Bar	Café
14	Santa Maria de Olivais	Café	Portuguese Restaurant	Furniture / Home Store	Pizza Place	Grocery Store
15	Santo António	Hotel	Portuguese Restaurant	Restaurant	Breakfast Spot	Café

As a result of this low venue type diversification, applying the k-means algorithm, the ideal K value is 2, as can be seen by the elbow method result.

This method runs the k-means clustering on the dataset for a range of values for k, and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center



Also the Silhouette Coefficien, that is calculated using the mean intra-cluster distance, present similar result (he best value is 1 and the worst value is -1.)

```
For n_clusters = 2 The average silhouette_score is : 0.4413253147581664
For n_clusters = 3 The average silhouette_score is : 0.24153453722233748
For n_clusters = 4 The average silhouette_score is : 0.16749450907175745
For n_clusters = 5 The average silhouette_score is : 0.10387744379057998
For n_clusters = 6 The average silhouette_score is : 0.04933683030561553
```

Conclusion

This project concludes that the three best neighborhoods to due to the proximity of a large number and variety of venues are: Avenidas Novas, Santa Maria Maior and Santo Antonio

However, the Models can be improved by adding new datasets. Some examples of datasets that can bring a lot of valuable information :

- Public Transports
- Health Care Facilities
- Schools
- Noise measurements
- Propriety price (€/m2)