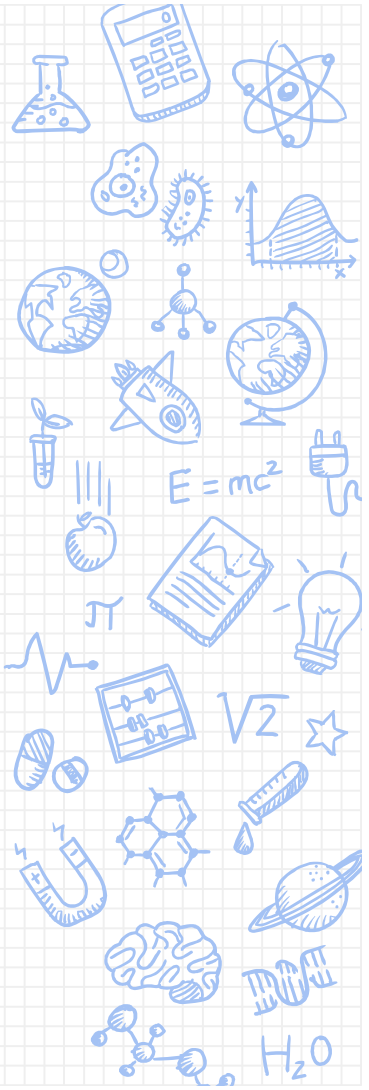
Capstone Project
The Battle of Neighborhoods

São Paulo and its contrasts

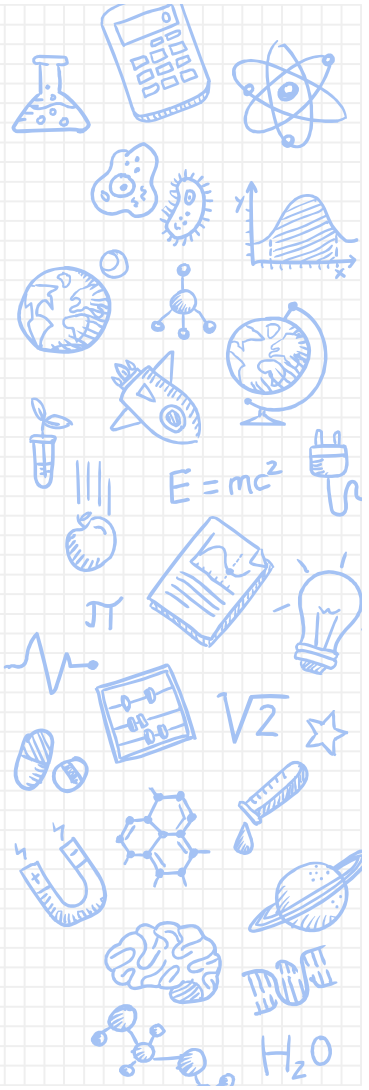
- The wealthiest city in the country and of South Hemisphere
- Representing alone 10.7% of all Brazil GDP
- The 13th most populous
- It exerts strong influence in commerce, finance, arts and entertainment in the country
- With a Human Development Index like a European country in some regions and low as a score of 0.700 in other regions.

How to find the right spot for your business in this city?



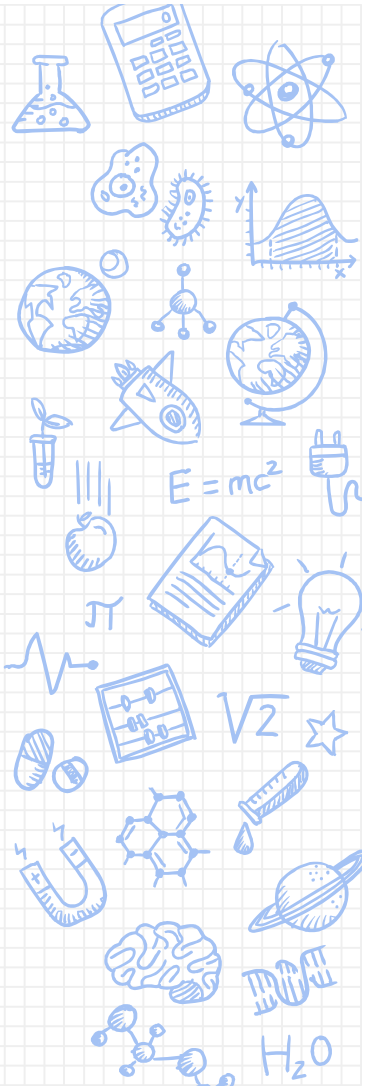
Data

- **Geographic Boundaries**: Latitude and Longitude
- **Population and Area**: Demographical Density
- **Aging Index**: Qty pop over 60 years / Qty pop below 14 years multiplied by 100
- **HDI**: The data is from 2000
- **Crime Rate**: the aggregation of all types of crimes divided by the population and multiplied by 1000. (2019)
- **Foursquare API**: Venues Categories and Locations



Methodology

1. Reduce the type of Venues Categories given by the Foursquare.
2. Transposed the base and aggregated the values by districts.
3. Normalized the values and replace any NAN value with a 0.
4. Descriptive statistics and correlation matrix.
5. Crime Rate: the aggregation of all types of crimes divided by the population and multiplied by 1000. (2019)
6. Apply the unsupervised learning algorithm, K-Means

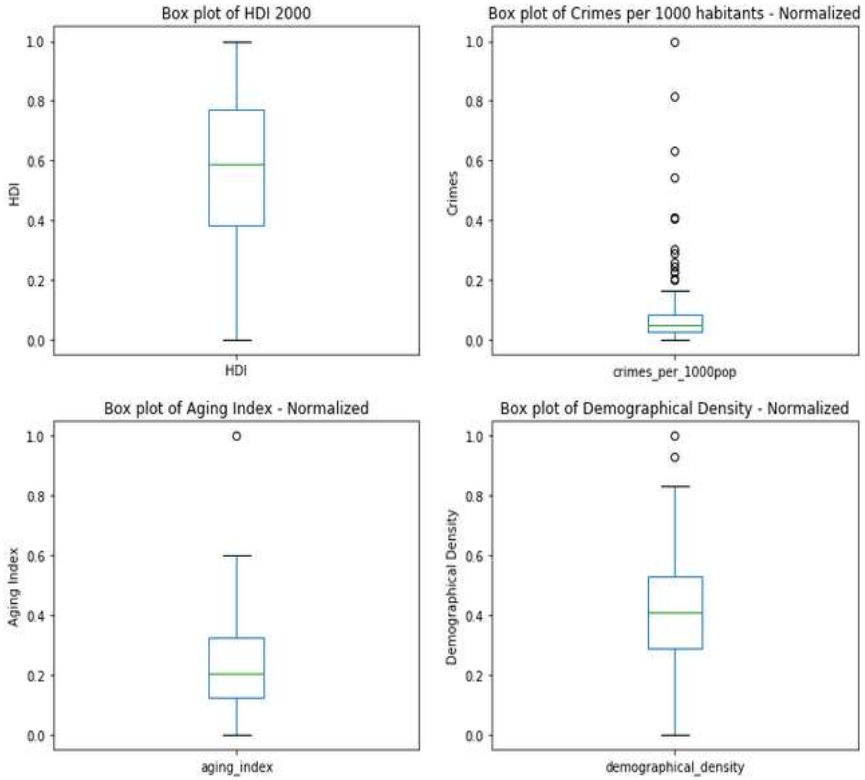


Analysis – Descriptive Statistics

	demographical_density	aging_index	HDI	crimes_per_1000pop	Bakery	Bank	Bar	Bus Station	Club	Coffee Shop	...
count	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	96.000000	...
mean	0.412344	0.234549	0.580208	0.100229	0.074632	0.000795	0.054698	0.004804	0.005764	0.026679	...
std	0.196399	0.154391	0.249326	0.162388	0.088116	0.005415	0.124247	0.021470	0.023293	0.053824	...
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	0.287501	0.121013	0.383654	0.025750	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
50%	0.408798	0.205258	0.586538	0.047621	0.053713	0.000000	0.000000	0.000000	0.000000	0.000000	...
75%	0.529667	0.322840	0.771154	0.085586	0.102778	0.000000	0.063322	0.000000	0.000000	0.036045	...
max	1.000000	1.000000	1.000000	1.000000	0.500000	0.050000	1.000000	0.142857	0.173913	0.333333	...

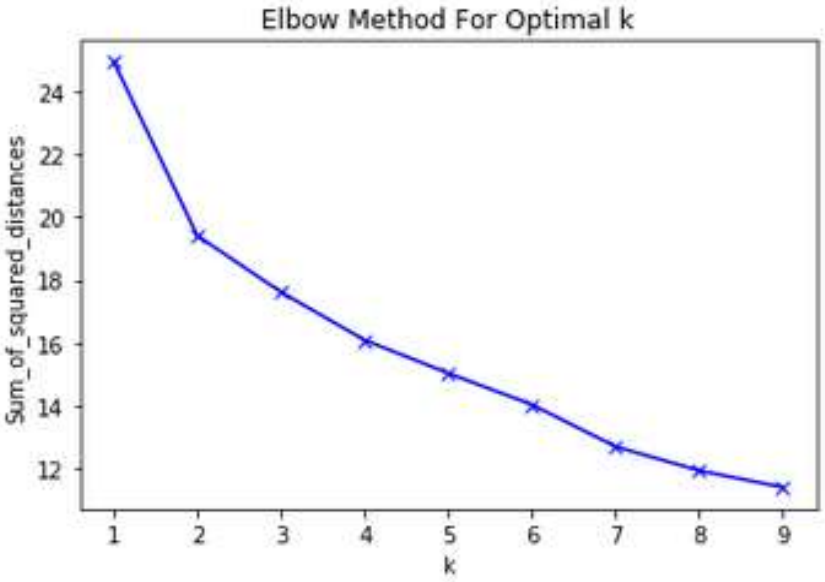
- ✓ Low mean for aging_index, crimes_per_100pop and Venues Categories in general
- ✓ 96 districts

Analysis – Box Plot



- ✓ For crimes per 1000 habitants have a large amount of outliers
- ✓ Few outliers for Aging Index and Demographical Density

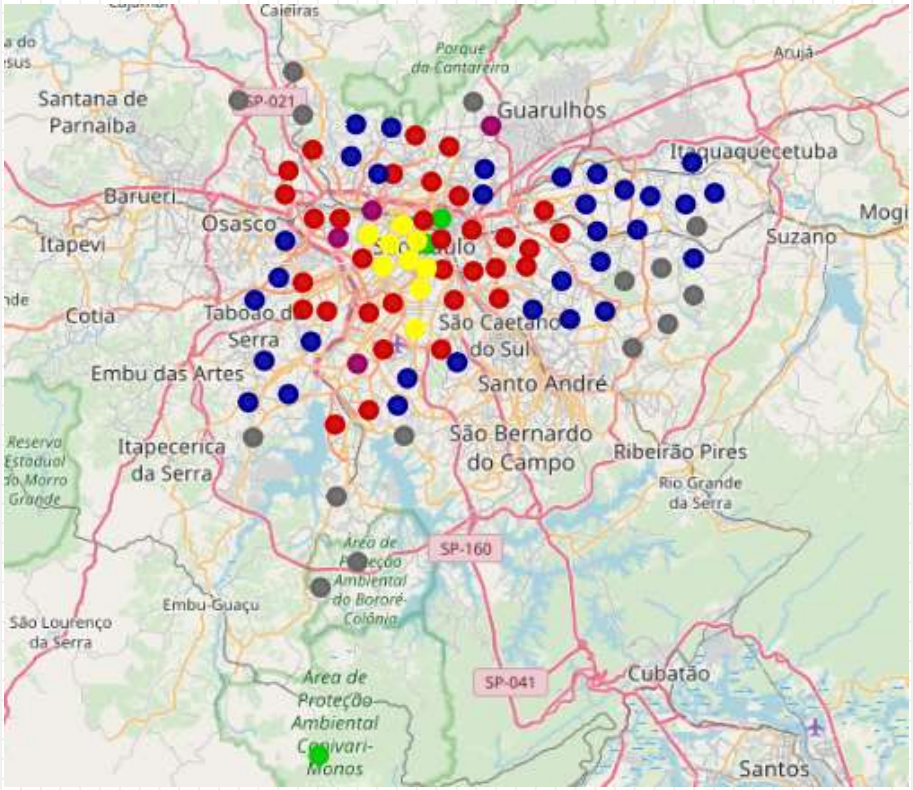
Analysis – K-Means



- ✓ Elbow around 2 Clusters but too low for the need for our analysis
- ✓ Testing various values for K I get the best value of 6.

Analysis – Clusters

It is important to note the pattern of points distribution on the map



Results and Discussion

Cluster Labels	demographical_density	aging_index	HDI	crimes_per_1000pop
0	0.245525	0.099269	0.288718	0.039738
1	0.330136	0.296171	0.757343	0.083851
2	0.705601	0.368039	0.920513	0.175710
3	0.214252	0.135558	0.408974	0.816405
4	0.533994	0.163991	0.427644	0.033064
5	0.231680	0.571810	0.795192	0.292543

- Due to the k-means algorithm trying to fit all the data in a cluster, we have data in some groups that differ from the group average.
- The Venues do not have a very well-defined pattern, becoming partly random as we could anticipate with the correlation matrix.

Conclusion

- The study gives us some interesting results, forming clusters with a certain consistency in the average and distribution of its indicators
- Due to the k-means method, some districts were allocated to groups whose value was very different from the rest
- As a next step, it would be interesting to add more metrics and filter the information from the Foursquare API for the business area that the stakeholders really want to establish