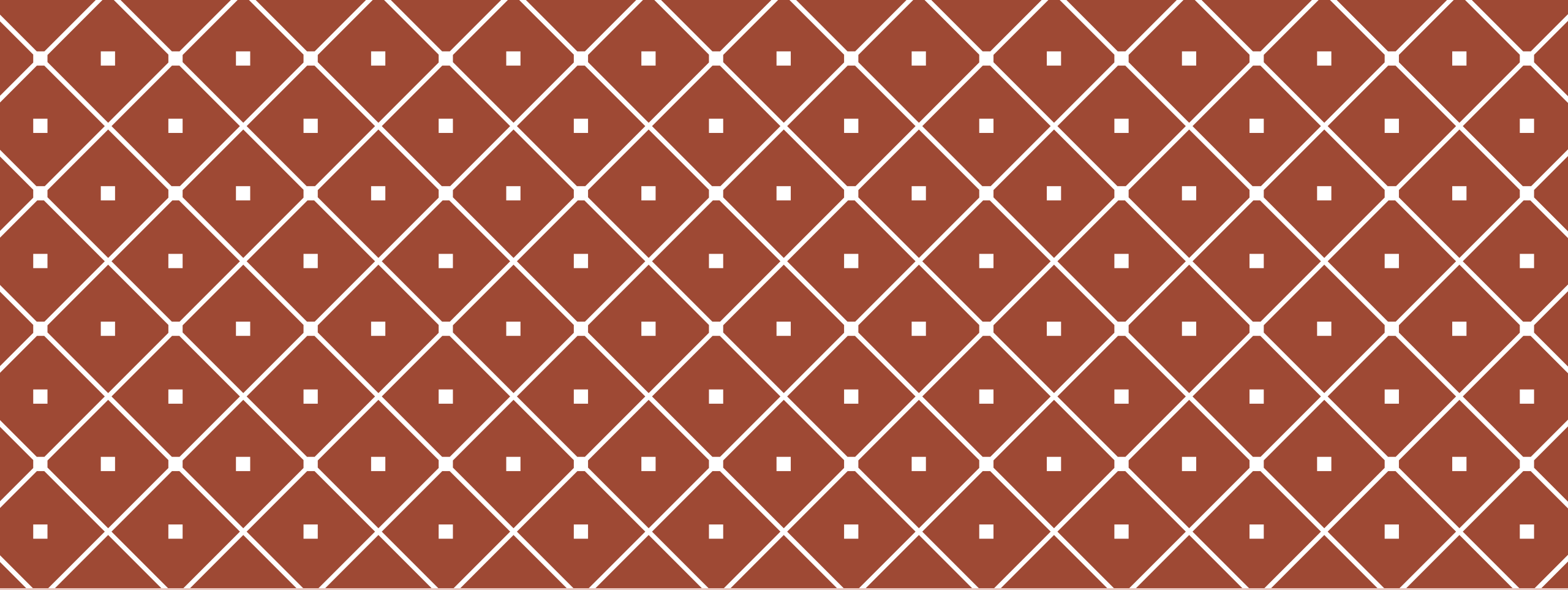


ANALYZING VENUE AVAILABILITY IN SÃO PAULO NEIGHBORHOODS

Lucas Falcão Monteiro

THE CITY OF SÃO PAULO

Largest city in Brazil

Largest metropolitan area in the southern hemisphere and in the Americas

São Paulo state accounts for 1/5 of
Brazil's population and
1/3 of its GDP.

SÃO PAULO DISTRICTS

The city is divided into 96 districts, or neighborhoods

These districts reflects Brazil's rampant inequality: the central ones have higher income and access to infrastructure, leisure and diverse establishments, while the outskirts are generally poor.

PROBLEM

Grouping neighborhoods together in terms of access to different types of venues should be able to indicate what areas are in need of investment and infrastructure, and both the private and public sector could benefit from this.

DATA

Data on districts and their localization was retrieved from the São Paulo city website

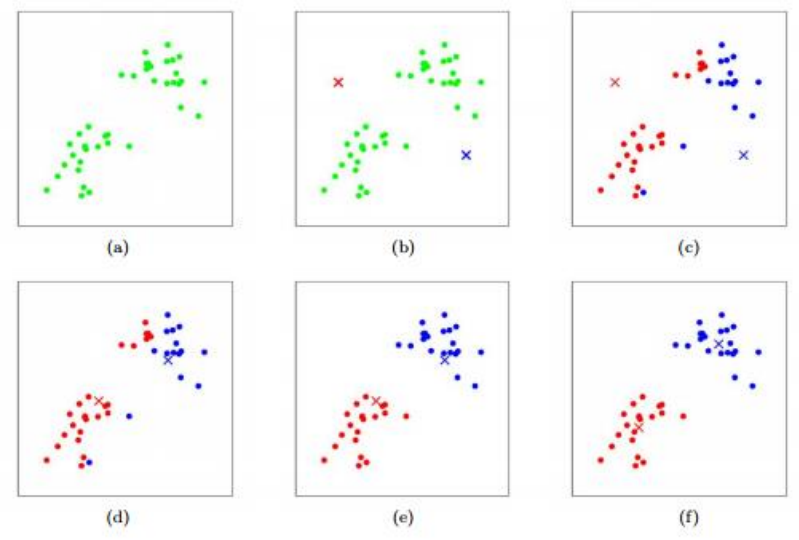
- Data was in the shapefile format. This was read and the coordinates were converted from SAD69 to the usual UTM format

Data on available venues within a radius of 1.6km was obtained from the Foursquare API

MODELING

The k-means algorithm from the sklearn package was used to cluster neighborhoods together based on number of venues, different types of venues (e.g. 60 types of venues), and also the types of venues itself (e.g. Italian Restaurant).

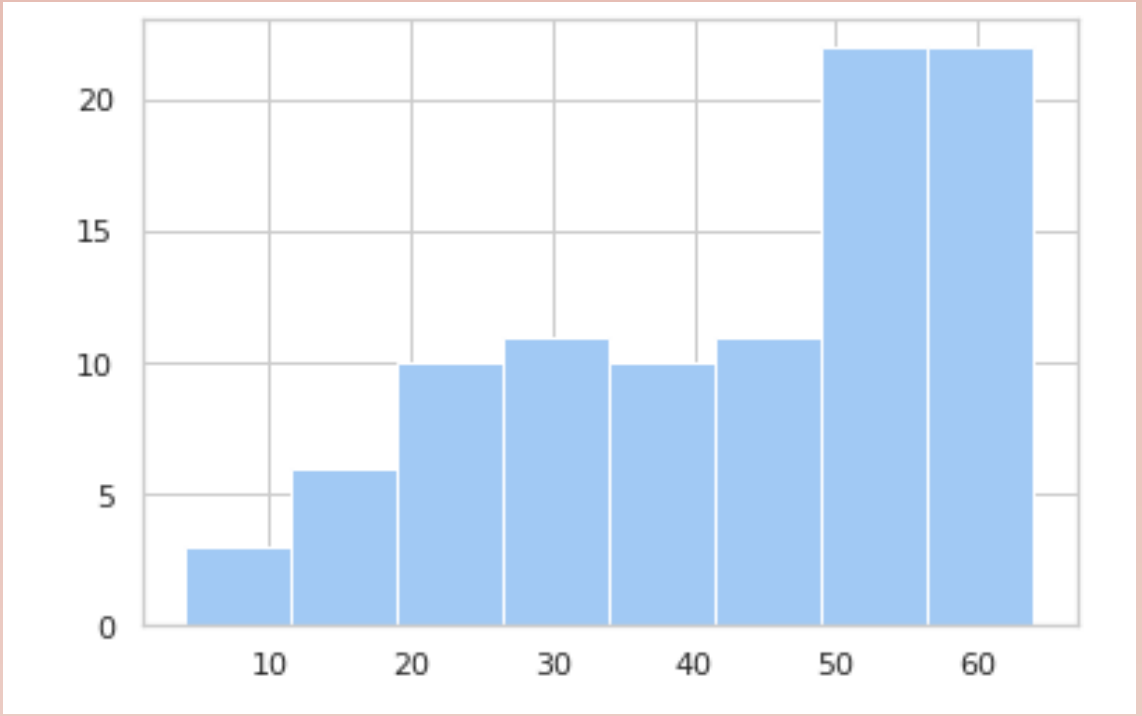
This algorithm uses the distance between each group (e.g. Euclidean) to perform successive iterations assigning them to the closest cluster center.



Source : stanford.edu/~cpiech/cs221/handouts/kmeans.html

RESULTS – HISTOGRAM OF VENUE TYPES

About 50% of the neighborhoods had a diverse number of types of venues nearby. About 40% of them had an intermediate number of venues and some 10% had very few types.

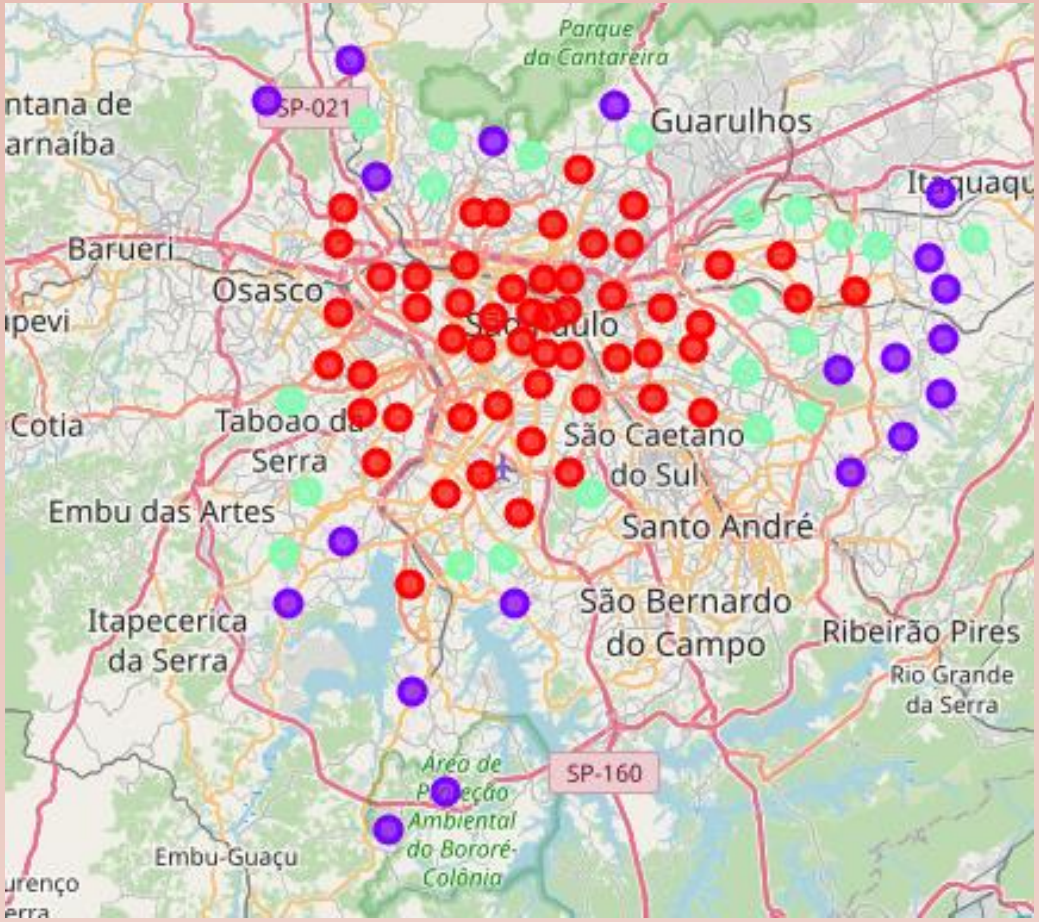


The x axis represents number of different types of venues and the y axis represents the frequency.

RESULTS – CLUSTERS ACCORDING TO NUMBER AND DIFFERENT TYPES OF VENUES

As expected, central neighborhoods had more diverse types of venues around (red circles) while neighborhoods on the outskirts lack options for venues.

3 clusters were used for k-means.

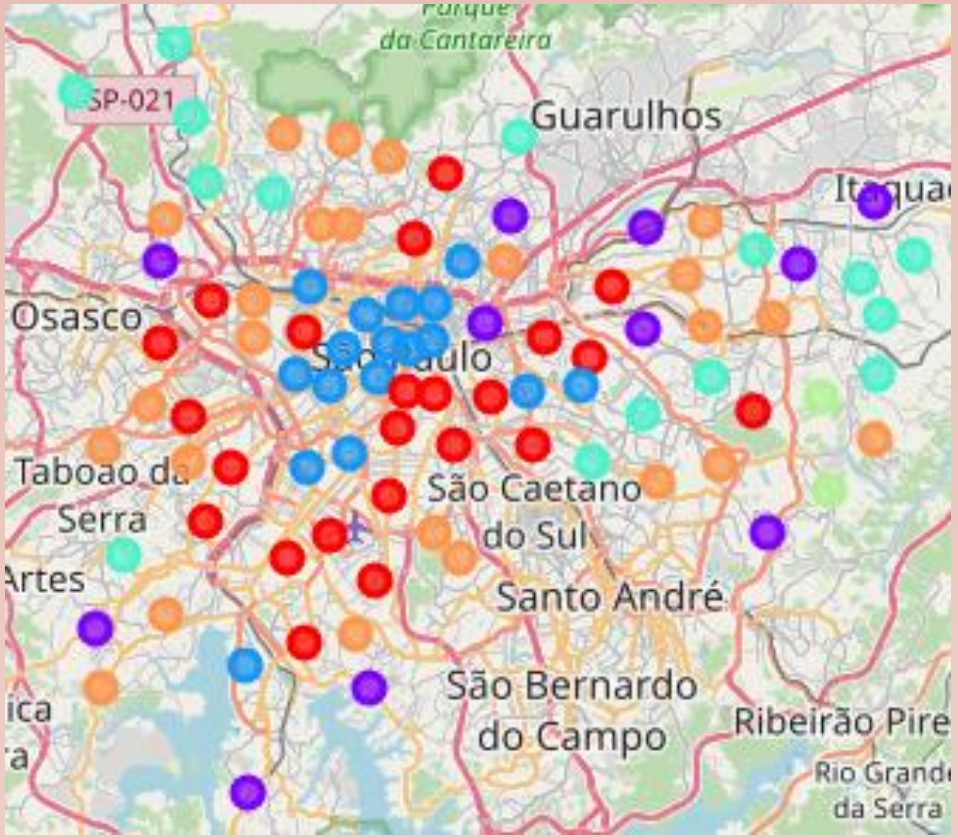


RESULTS - CLUSTERS ACCORDING TO VENUE CATEGORY

4 Neighborhoods were not included due to low number of nearby venues: Marsilac, Tremebé, Parelheiros and Grajaú

6 clusters were used in k-means

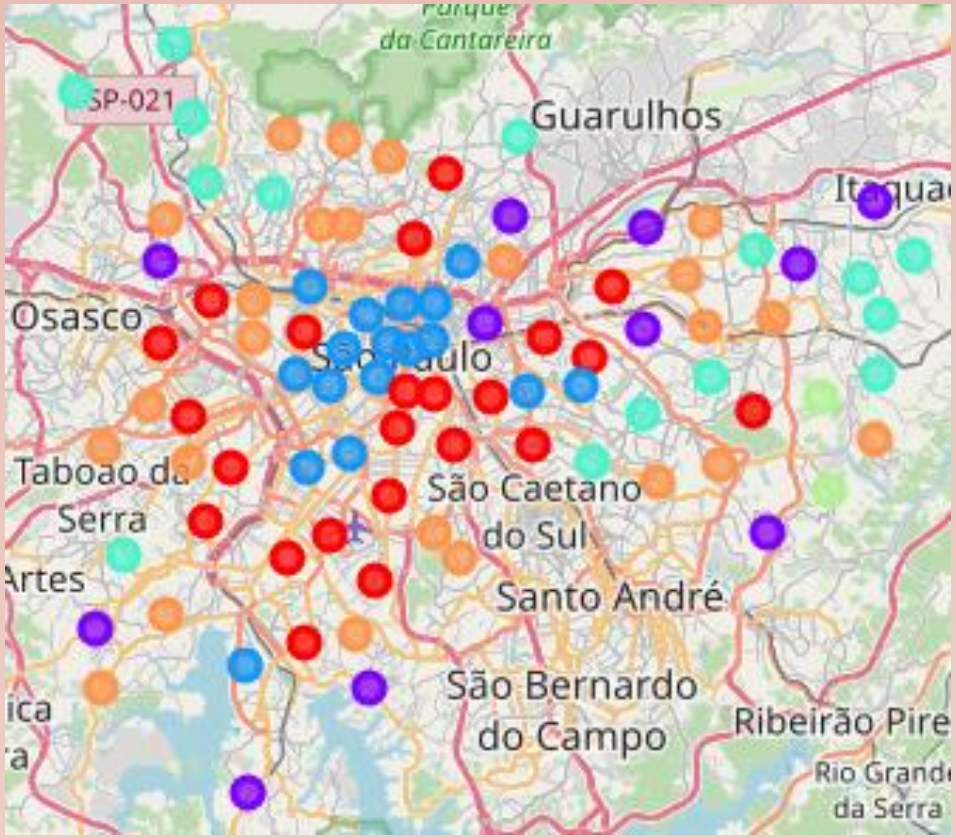
Clustering shows a general trend with spatial connection as well, but there are some interesting exceptions



RESULTS - CLUSTERS ACCORDING TO VENUE CATEGORY

Some neighborhoods in the center lack diversity (e.g. Belém, a purple spot in the middle) and others on the outskirts show higher diversity (e.g. Saúde, a blue point in the south, and Parque do Carmo, a red point in the East)

Purple, Light Green and Light Blue points indicate areas that do not show much diversity of venues.



DISCUSSION

Some shortcomings of the analysis are:

- The lack of demographic information to center the districts based on habitation and not on their borders
- Limitations on the number of venues retrieved from each neighborhood (100)
- Possibly incomplete annotation in Foursquare

Even if some neighborhoods are incompletely annotated, this could still indicate areas with lower access to technology and thus point out areas for further development

CONCLUSION / FUTURE DIRECTIONS

This project has identified neighborhoods in São Paulo with poor access to diverse venues and establishments, and the public and private sector could benefit from this information

These analyses could be improved by enhancing the locations of each neighborhood based on population density, and from proper annotation of nearby venues (such as using other platforms such as Google Maps).