Relationship Between Venue Categories and Income in Districts of the City of São Paulo, Brazil

Gabriel Siqueira Kakizaki

September 25, 2020

1 Introduction

1.1 Background

São Paulo, according to the Globalization and World Cities (GaWC) Research Network, is an alpha global city, along with Los Angeles and Amsterdam. It is in the ranking of the wealthiest and the most populous cities of the world, known for its cultural, social and ethnic diversity. Often holding international events, it is sought-after because of its opportunities for business.

1.2 Problem

Opening a new business in such place can be challenging by a number of reasons. Choosing a location in itself has a number of things to consider, and the income of people living in the surrounding area might be one of them. The aim of this project is to discover, if any, the relation between the type of venue and the income of people living in the districts of the city of São Paulo.

1.3 Stakeholder interest

The main audience of this project are entrepreneurs choosing where to open a business in São Paulo. The results could also be useful for people just wanting to move to the city, real estate agents, and investors because this information might give an edge on decision-making.

2 Data

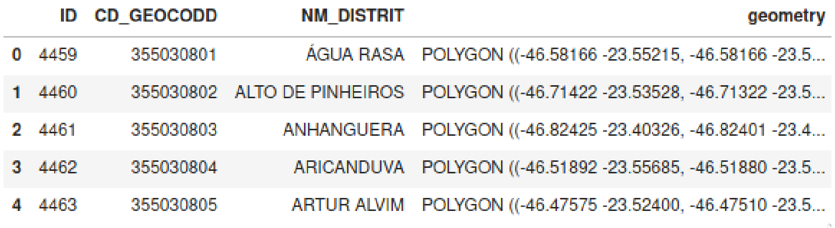
2.1 Data sources

Income and geographic data came from IBGE (Brazilian Institute of Geography and Statistics), and can be accessed here. More specifically, the income data comes from the 2010 Population Census (to know more about click here), and geographic data comes from here (portuguese) in form of shapefiles for geographic information system (GIS) software. Venue data was gathered using the Foursquare Places API with a free account created for this project.

2.2 Data description

2.2.1 Geographic data

The shapefiles came with information about the district boundaries, along with its name, code and id. This data covered the whole state of São Paulo, not just the city. A sample of it is shown in Figure 1.



	ID	CD_GEOCODD	NM_DISTRIT	geometry
0	4459	355030801	ÁGUA RASA	POLYGON ((-46.58166 -23.55215, -46.58166 -23.5...
1	4460	355030802	ALTO DE PINHEIROS	POLYGON ((-46.71422 -23.53528, -46.71322 -23.5...
2	4461	355030803	ANHANGUERA	POLYGON ((-46.82425 -23.40326, -46.82401 -23.4...
3	4462	355030804	ARICANDUVA	POLYGON ((-46.51892 -23.55685, -46.51880 -23.5...
4	4463	355030805	ARTUR ALVIM	POLYGON ((-46.47575 -23.52400, -46.47510 -23.5...

Figure 1: Geopandas dataframe

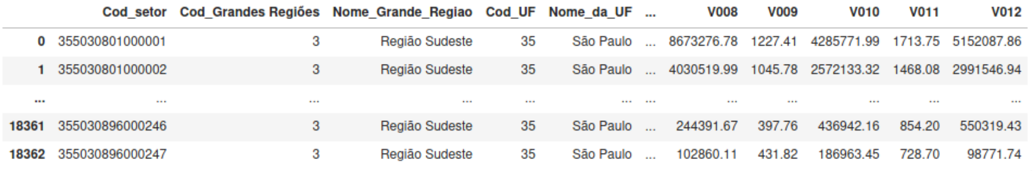
2.2.2 Census data

The census table contained various attributes, and the statistics columns were labeled with “VXXX”, where XXX is a three-digit number. The relationship between labels and their meaning can be found on the 2010 census documentation.

As we were interested only on the income, only the column “V009” was used, which translating its name to english means: “Value of the average nominal monthly income of persons aged 10 and over (with and without income)”. There weren’t any missing values for this column.

Originally, each row is associated with a unit called "census sector" (portuguese: setor censitário), which is normally a lot smaller than the actual district, composed only of one or a few city blocks.

A part of this table can be seen in Figure 2.

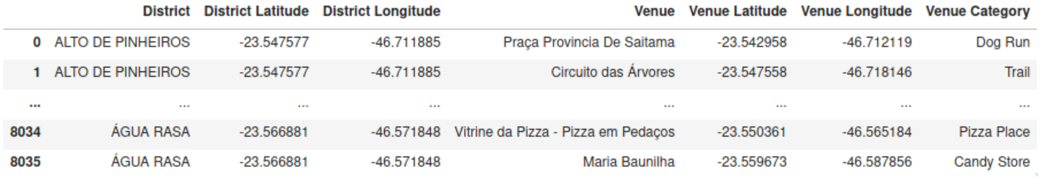


	Cod_setor	Cod_Grandes Regiões	Nome_Grande_Regiao	Cod_UF	Nome_da_UF	...	V008	V009	V010	V011	V012
0	355030801000001	3	Região Sudeste	35	São Paulo	...	8673276.78	1227.41	4285771.99	1713.75	5152087.86
1	355030801000002	3	Região Sudeste	35	São Paulo	...	4030519.99	1045.78	2572133.32	1468.08	2991546.94
...	...	...	...	...	...	...	...	...	...	...	...
18361	355030896000246	3	Região Sudeste	35	São Paulo	...	244391.67	397.76	436942.16	854.20	550319.43
18362	355030896000247	3	Região Sudeste	35	São Paulo	...	102860.11	431.82	186963.45	728.70	98771.74

Figure 2: Pandas dataframe containing the census data.

2.2.3 Venues

Foursquare API was used to collect the name, category, latitude and longitude values for a total of 8036 venues in a 2 km radius around the centroid of each of the 96 districts. The results returned from the API in JSON format were merged into a dataframe containing the district name, and coordinates as shown in Figure 3.



	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ALTO DE PINHEIROS	-23.547577	-46.711885	Praça Provincia De Saitama	-23.542958	-46.712119	Dog Run
1	ALTO DE PINHEIROS	-23.547577	-46.711885	Circuito das Árvores	-23.547558	-46.718146	Trail
...	...	...	...	...	...	...	...
8034	ÁGUA RASA	-23.566881	-46.571848	Vitrine da Pizza - Pizza em Pedaços	-23.550361	-46.565184	Pizza Place
8035	ÁGUA RASA	-23.566881	-46.571848	Maria Baunilha	-23.559673	-46.587856	Candy Store

Figure 3: Pandas dataframe containing the venue information.

3 Methods

All data processing, analysis and modeling was done using Python and various libraries. This work can be seen in a Jupyter notebook at my github page (github.com/kakig) on the repository `Coursera_Capstone`.

3.1 Exploratory data analysis (EDA)

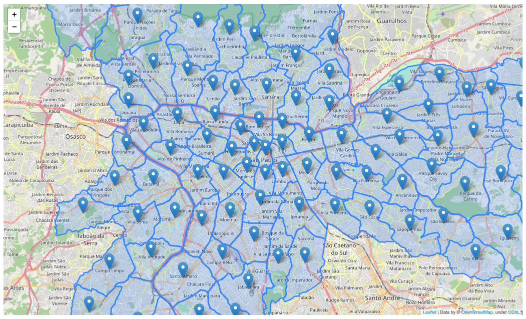


Figure 4: Map of São Paulo showing districts with markers on its centroids.

As shapefiles came with information for the whole state, only data about the city was selected. Then, centroids were calculated for each district using its geometry information, for use with the Foursquare API. Some conversion between different coordinate reference systems (CRS) was needed for this calculation and for plotting maps, which also required generating a GeoJSON object. To verify that the area and placement of the districs and the calculated centroids were correct, a map was created for visualization, as shown in Figure 4.

Proceeding to census data, to get the average income per district instead of per census sectors, data was grouped by district name and the mean was taken. The mean income for the districts was R$1576.82, and the minimum and maximum as R$396.48 and R$5402.81 respectively. There are few outliers, but they are not very large and should not disturb our analysis. The distribution of the values can be seen in Figure 5.

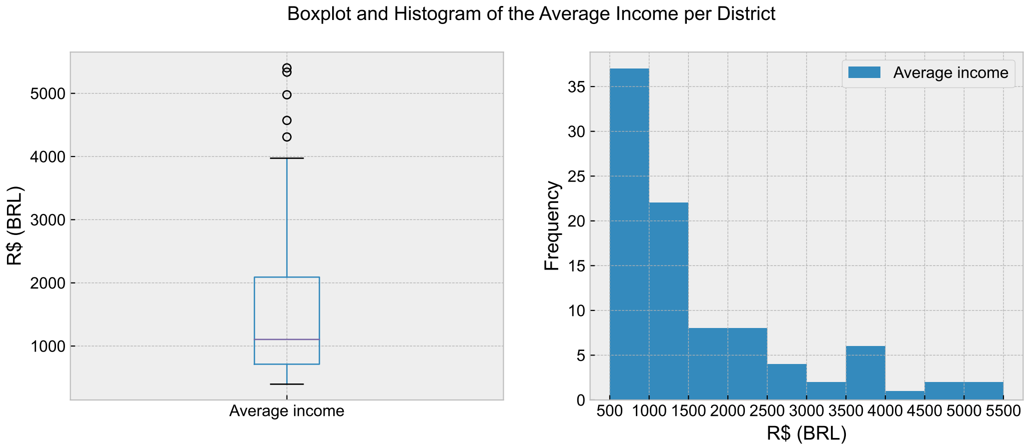


Figure 5: Boxplot (left) and histogram (right) of income. Values in Brazilian Real.

Joining the income with geographical data, it is possible to visualize where the income is higher or lower. In Figure 6 we could observe that higher income is more concentrated to the center of the city, and lowers as we go far from the center. If the income has no impact on the type of venues, then we should see the same, or at least close frequencies of venues across the districts. If it does make a difference, and that is our hypothesis, we should see different ratios of venues across different income ranges.

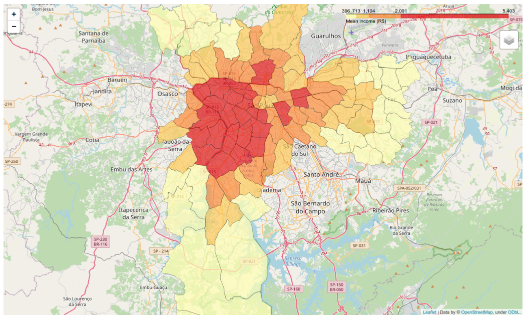


Figure 6: Average income per district of São Paulo. Red means higher income.

Switching focus to working with venue information, due to the difference in number of venues between districts, first, one-hot encoding was used to transform the table, keeping only the category value for each venue. There were 363 unique venue categories that became the columns for the table. Data in this format was used to visualize the most common venues in the whole city. Then, the values were grouped by district with and the mean was taken, so we ended up with the frequency of venue categories in percentage for each district. This may allow for a better and more fair comparison during modeling.

3.2 Modeling

Machine learning models were built using the scikit-learn library. Both unsupervised (K-Means) and supervised (Lasso, Random Forest) learning approaches were used in a complementary manner.

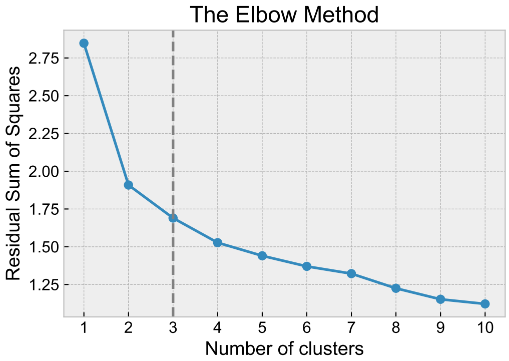


Figure 7: Elbow method for choosing the number of clusters in K-Means

K-Means algorithm was used to segment n data points (districts) into K segments, or clusters, with similar characteristics (venue frequency). Note that income data **was not** used to train this algorithm. As it cannot automatically determine the number of clusters in the data, the elbow method was used for choosing the optimal K, which was three in this case.

The target of prediciton for regression was the mean income for each district. Due to our sample size, models were evaluated using *k-fold* crossvalidation ($k = 5$) to estimate the coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE).

Lasso was used instead of ordinary least squares regression due to the large number of features (363 venue categories), as it tends to have fewer non-zero coefficients, thus being simpler.

Random forest was also used because the data might have a non-linear relationship, so a decision tree model can better fit the data and it still can be used to explain feature importances.

4 Results

4.1 Clustering

K-Means segmented the districts mainly into 2 clusters, and the third cluster contained only one district, the most southern one which contains an environmental preservation area, and because of that it was excluded from further analysis.

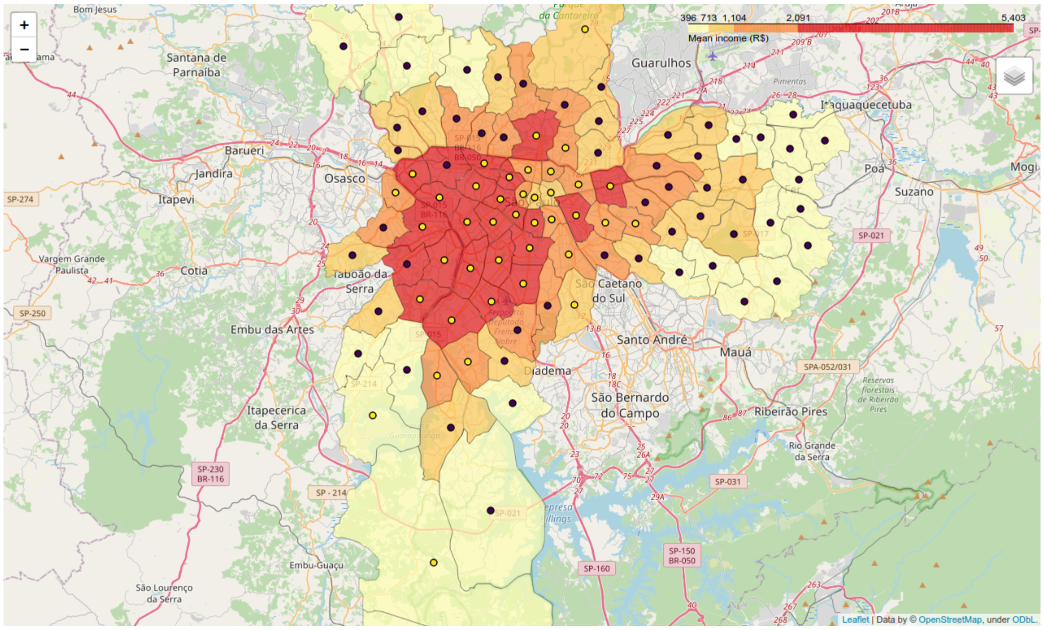


Figure 8: Choropleth map of income per district. Dark purple circles indicate the district is part of the first cluster, and yellow circles, the second.

The first cluster had 55 districts with a mean income of R$924.24 and the second had 40 districts, and a mean income of R$2503.62, more than double than the first one. The first cluster also had mostly bakeries and pizza places as most common venues, while the second had a more diverse selection of most common venues. As can be seen in Figure 8, districts from the second cluster are usually located more to the center.

4.2 Regression

Model	R^2 (mean, std)	MAE	RMSE
Lasso	0.53 (0.19)	615	810
Random Forest	0.58 (0.15)	566	752

Table 1: Coefficient of determination, mean absolute error and root mean squared error for lasso and random forest models.

Table 1 shows the performance of the two regression models. Random forests performed better lasso in all metrics, thus it was used to explain which venue categories are correlated with high and low income districts using SHAP values as shown in Figure 9. We can see that higher bakery frequency tends to lower the value of the predicted income, confirming what was shown with clustering. In general, restaurants of different (foreign) cuisines tend to increase the value of the output of the model, indicating that they are located in higher income areas.

It is also interesting to see that spa, vegetarian/vegan restaurant and art museum categories impact positively the model output, reinforcing the common sense conception that these places are frequented by people with higher economic status.

5 Discussion

5.1 Recommendations

For further research, the relationship between venue types and other kinds of data could be explored, such as with crime rates, renting prices, etc. This may lead to other insights about venue localization and where it might be best to open a new place. Also, using a smaller terrain unit than a district could be useful for pinpointing more exact locations suited for opening a new business.

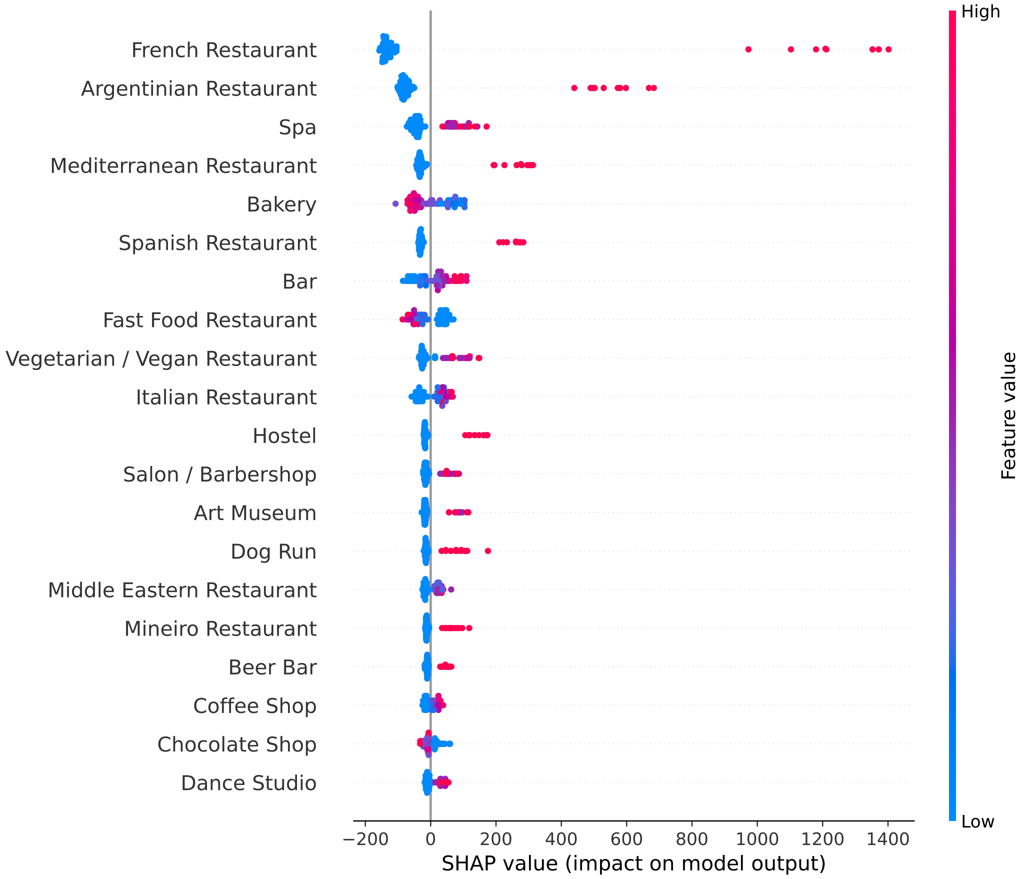


Figure 9: SHAP values for the top 20 features sorted by average impact on model output.

While working with geographical data, caution should be used to verify that the information about all places is correct. I had problems with districts, and previously with neighborhoods that had the exact same name in this and other cities. Also the names can contain abbreviations and sometimes do not match where they should.

5.2 Limitations

The census data was gathered in 2010, and venue data is from 2020. This may impact the models accuracy and the explanations of which venue types are correlated with higher or lower income areas.

Venue data was gathered in a 2 km radius from the districts centroid, because the Foursquare API is limited to a circular search around latitude and longitude coordinates. This 2 km radius is good for most of the districts, but this selection isn't perfect. The venue data for smaller districts agglomerated in the center may have included venues in other adjascent districts, and districts at the edge of the city might have venues from other cities, due to conurbation. While I could have restricted venues to be inside of the district, this would result in less data for the machine learning algorithms, and in practice, these kinds of political divisions aren't really obstacles for people going to the venues.

6 Conclusion

In this study I analyzed the relationship between the venue categories and the income for districts of city of São Paulo, and identified that few venue categories are correlated to the income, especially restaurants from foreign cuisine. It is also worth noting that bakeries and fast food restaurants appear more in lower income areas. This information could be useful to help entrepreneurs deciding the location of a new business, to know better their public based on the income.