Schools availability correlation with socio-demographics in Sao Paulo city

By Guilherme Kleine

1. Problem Introduction

Sao Paulo, Brazil is a large city with a huge disparity in standard of living and opportunities between each of its 95 districts. Although government is working to provide quality education to all population, some districts is lacking behind.

This study will try correlating the availability of schools with the quality of life of its population. Quality of life will be considered the performance of Social Indexes, such as per capital income, life expectancy, literacy rate and HDI.

The ultimate goal for this study is to check if it is possible to correlate both properties and look for outliers that can as benchmark or object of future study for public policies.

2. Analytic Approach

A Descriptive model was selected to understand effectively how the Social-Demographic characteristic of Sao Paulo districts correlates with the educational availability. Each district will be divided on separate clusters based on those two attributes. A correlation matrix will then be created to extract outlier districts, where socioeconomics and educational opportunities no not correlate.

3. Data Collection

Data was collected from official Brazilian government agencies and using Foursquare API

Data set 1: Socio- Economic Data for São Paulo city

Source: 2010 Census – available at: http://dados.prefeitura.sp.gov.br/pt_PT/dataset/censo-demografico-2000-e-2010

Data Content: More than Socio-Demografic 200 indexes for the 1595 microregions inside Sao Paulo

	NOME_UDH	ESPVIDA	T_ANALF18M	T_FREQ6A17	RDPC	P_FUND	P_MED	P_SUPER	REN0	RENOCUP	IDHM	IDHM_E	IDHM_L	IDHM_R
0	Jardim Anália Franco / Vila Formosa : Hospital...	79.91	1.45	96.11	1801.17	83.89	70.22	31.00	0.89	2524.57	0.866	0.815	0.915	0.870
1	Vila Califórnia	80.04	1.29	98.10	2000.87	87.80	74.85	37.16	1.24	2599.58	0.870	0.809	0.917	0.887
2	Vila Carrão / Vila Formosa : Cemitério Vila Fo...	77.05	2.22	96.50	890.75	74.98	53.96	12.56	1.15	1398.04	0.790	0.750	0.868	0.757
3	Vila Formosa : Escola Municipal de Ensino Fund...	79.61	2.09	91.08	1233.65	76.97	56.83	17.48	2.03	1714.03	0.816	0.736	0.910	0.810
4	Aricanduva : Centro de Educação Infantil Coryn...	79.56	2.70	94.93	1180.17	82.71	57.48	21.37	0.95	1789.84	0.820	0.756	0.909	0.803

Figure 1: Census Raw Data for the city of Sao Paulo

Data set 2: Availability of Schools for each districts

Source: Foursquare API

Data Content: Information of a venue in a 3km radius of a location for the following categories:

-Elementary School-

-Middle School

-Private School

3Km was set standard as a reasonable distance for educational daily commute

A function was created to fetch the information about venues on each categories

Auxiliary Data Set: Geographical Data for district location

Source: Official Administrative limits - available at: https://mapas.ibge.gov.br/bases-e-referenciais/bases-cartograficas/malhas-digitais.html

Data Content: Coordinates for Sao Paulo districts

	Lat	Long
DISTRITO		
Cidade Tiradentes	-23.603240	-46.399874
Itaim Paulista	-23.468716	-46.403903
Guaianases	-23.571128	-46.411152
Vila Curuçá	-23.503012	-46.415271
Iguatemi	-23.617538	-46.418492

Figure 2: Example of districts coordinates data

4. Data Preparation

For Data set 1

The data was grouped by district name. The following 13 indexes were considered relevant for this study and were extracted for the data:

Life Expectancy , Illiteracy Rate , School Freq. Rate , Per Capita Income , % of Elementary School Degree , % of High School Degree , % of College Degree , % of no income , Household income , HDI , Education HDI , Health HDI , Income HDI

These indexes reflect the main social and economic situation of each of the 93 districts in São Paulo city

District Name	Life Expectancy	Illiteracy Rate	School Freq. Rate	Per Capita Income	% of Elementary School Degree	% of High School Degree	% of College Degree	% of no income	Household income	HDI	Education HDI	He
Alto de Pinheiros	81.640000	0.610000	96.210000	5207.070000	94.410000	89.340000	70.020000	1.310000	6474.970000	0.936000	0.870000	0.94
Anhanguera	72.767778	6.422222	92.704444	561.633333	62.257778	37.545556	5.527778	0.914444	1004.542222	0.713889	0.673556	0.79
Aricanduva	76.353846	4.421538	92.140000	1243.473077	72.158462	51.696154	18.118462	1.226923	1835.792308	0.776154	0.706769	0.85
Artur Alvim	76.147500	3.450000	93.063333	826.616667	70.170000	50.778333	13.875000	1.033333	1279.325833	0.760500	0.702250	0.85
Barra Funda	77.530000	5.500000	95.500000	2449.790000	78.090000	63.806667	38.766667	1.760000	3100.580000	0.839667	0.783333	0.87

Figure 3: Census data after data preparation

In order to better understand the quality of the data and format, histograms of some of those indexes were prepared. It showed that some index follow a bell curve and other, like income distribution is a positive skewed curve, as expected.

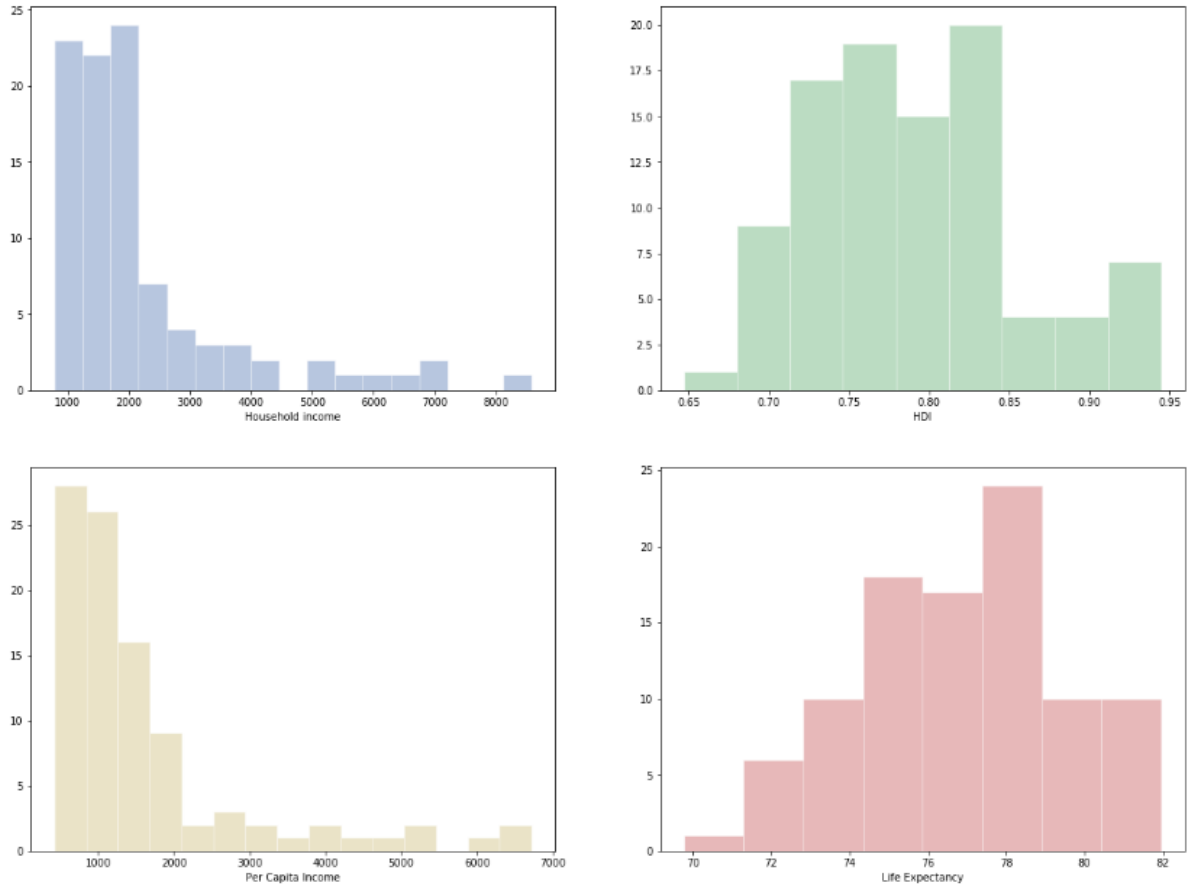

Figure 4: Social-economical indexes histograms

Population data for Data Set1 was also used to calculate the school/children ratio later on this study

For Data set 2

The data fetched by Foursquare API was consolidated and grouped by district considering the number of venues found

	District Name	N of Elementary Schools	N of High Schools	N of Private Schools
0	Alto de Pinheiros	18	14	6
1	Anhanguera	3	4	0
2	Aricanduva	24	8	2
3	Artur Alvim	28	18	4
4	Barra Funda	23	17	11

Figure 5: Example of the School availability per district

5. Modeling

First step done was to cluster each district in five clusters based on the 13 socio-economic indexes mentioned on section 4. The indexes were normalized and the algorism was run 12 times to find the optional centroid. The clusters were later renamed in order of socioeconomic development level from 0 as higher development to 4 as lower development. Figure 6 shows the average index of each cluster and Figure 7 is a map of color-coded cluster of each district.

	Life Expectancy	Illiteracy Rate	School Freq. Rate	Per Capita Income	% of Elementary School Degree	% of High School Degree	% of College Degree	% of no income	Household income	HDI	Education HDI	
Cluster												
0	81.194502	1.265746	95.985430	4896.846924	92.120030	85.102079	61.854627	1.599074	5852.654137	0.923464	0.867709	0.
1	78.623706	2.568670	93.551259	1808.139900	78.592198	62.559134	28.490517	1.375352	2367.906337	0.827069	0.762700	0.
2	76.165431	4.067583	92.790112	1123.414519	69.137109	49.590856	15.902956	1.408515	1605.294420	0.766216	0.700565	0.
3	73.627060	5.699621	92.415389	711.729763	61.584370	39.135626	7.831509	1.214390	1128.619754	0.716822	0.653118	0.
4	71.464265	7.732941	93.220882	511.386029	54.188676	30.571471	4.567059	4.078235	891.204118	0.676176	0.603118	0.

Figure 6: Average socio-economic index for each cluster

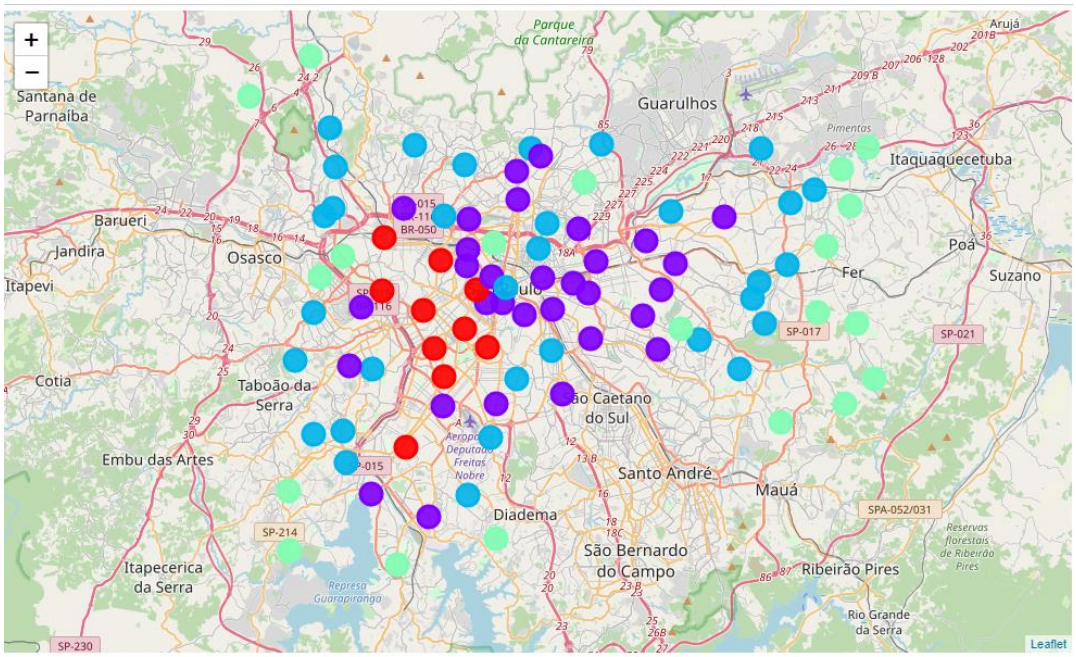

Figure 7: Location of each socio-economic cluster

The second step was to combine the data set 2 (from Foursquare API) and divide by the population retrieved from Data set 1

N of Elementary Schools	N of High Schools	N of Private Schools	Population	Elementary Schools Conc	High Schools Conc	Private Schools Conc
5	5	3	4794.0	1.042970	1.042970	0.625782
3	3	0	14629.0	0.205072	0.205072	0.000000
13	3	0	14100.0	0.921986	0.212766	0.000000
18	7	2	16982.0	1.059946	0.412201	0.117772
16	9	4	2759.0	5.799203	3.262051	1.449801

Figure 8: Example of number of schools and schools/ children ration analysis

As in the first step, a K-Mean model was used to cluster the districts in to five clusters based on the availability of Elementary, High School and Private Schools per children. The clusters was also relabeled from 0 as higher availability to 4 as lower availability. Figure 9 shows the average school availability of each cluster and Figure 10 is a map of color-coded cluster of each district.

:

School Cluster	Elementary Schools Conc	High Schools Conc	Private Schools Conc
0	5.066078	2.771277	2.435277
1	1.837831	1.095041	2.323203
2	1.526207	1.665857	0.712059
3	0.972319	0.609205	0.353792
4	0.286388	0.219133	0.072290

Figure 9: Average school per thousand child ration for each cluster

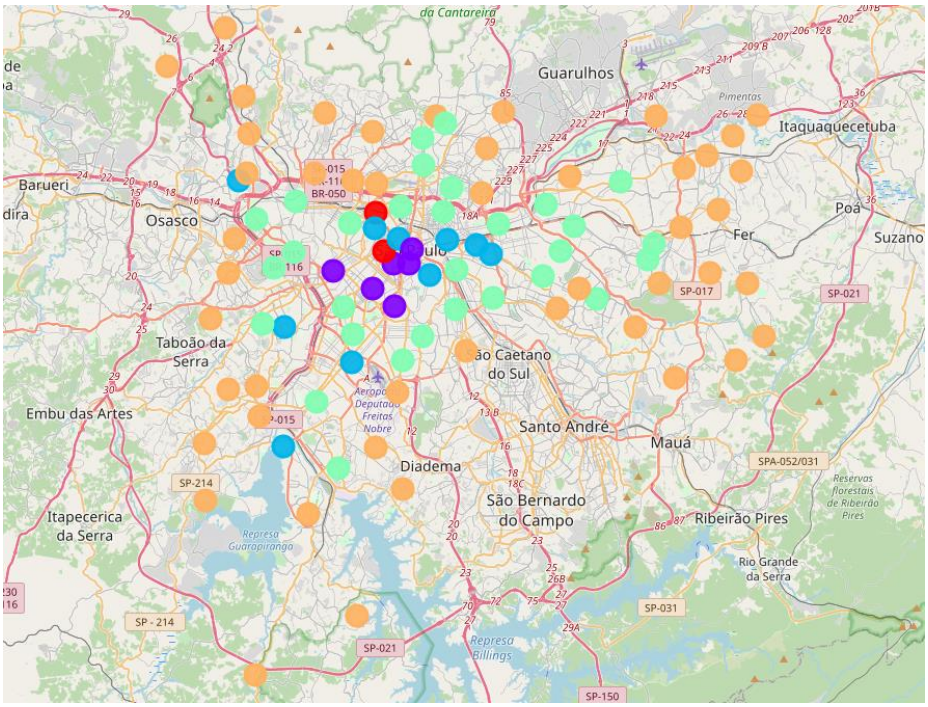

Figure 10: Location of each school avaiability cluster

6. Analysis

In order to understand how schools availability is distributed between the districts, a histogram was prepared.

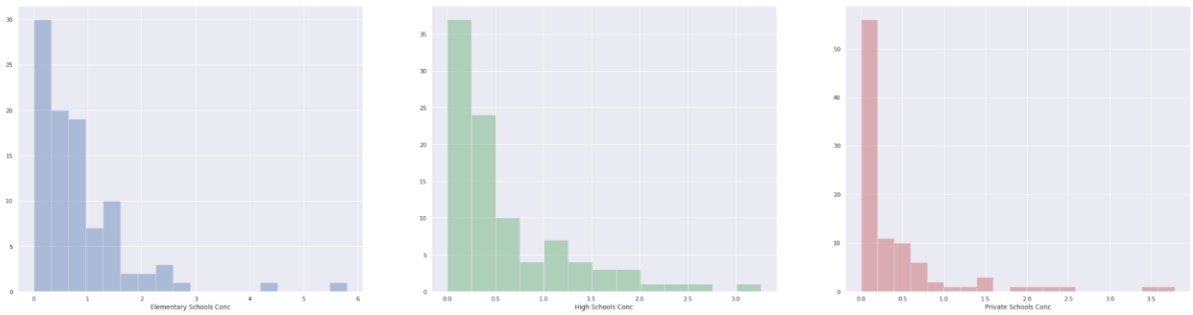

Figure 11: Histogram of school availability

It showed a very positive skewed graph, with many districts with a low school/child ratio

A correlation matrix was created to investigate the relation between school availability and socio-economic development. Figure 12 shows this matrix as a heat map. Figure 13 is a legend of the characteristic of each cluster.

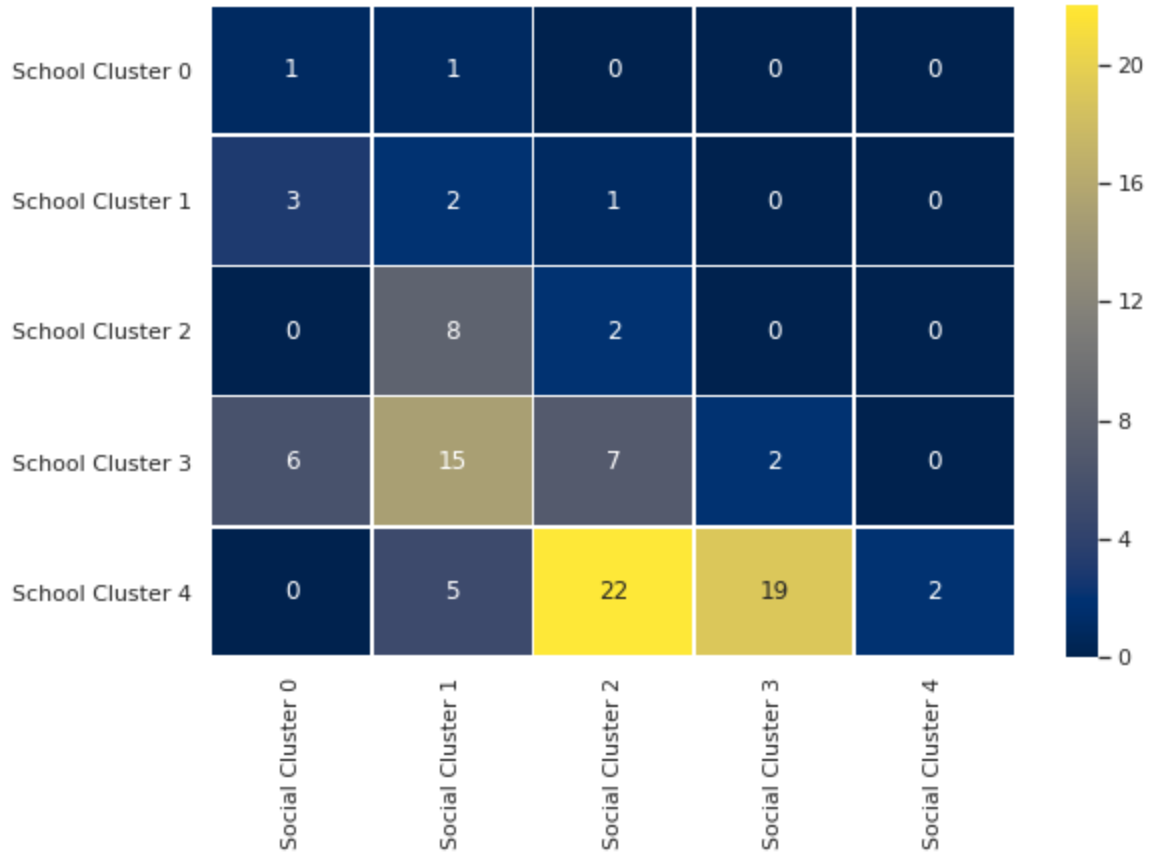

Figure 12: Correlation matrix between both clustering analyses

Number	Social Cluster	School Cluster
0	Highest Education, Highest Income, Highest Health	High number of private Schools and public schools
1	Medium Education, High Income	High number of private Schools
2	Medium Education, Medium Income	Medium number of public schools
3	Low Education, Low Income	Low number of schools
3	Lowest Education, Lowest Income, Lowest Health	Lowest number of schools

Figure 13: Clusters main characteristics

The correlation matrix showed that most districts are on the lower spectrum of schools availability. 82% (78 out of 95) are on either cluster 4 or 3. The districts on Social-economic cluster 0, 1 and 2 follow the diagonal line, where there is a correlation between both clusters, with the exceptions show on Figure 14.

	District Name	Development Cluster	School Cluster
0	Alto de Pinheiros	0	3
32	Itaim Bibi	0	3
46	Lapa	0	3
51	Moema	0	3
59	Perdizes	0	3
70	Santo Amaro	0	3

Figure 14: Affluent districts with disparity on school availability

7. Conclusion

The distribution map showed that both affluent socio-economical clusters and high availability of schools are located near the financial center of the city. The districts on the outskirts of the city showed worse socio economical indexes as well as lower school per children ratio.

This study showed a great disparity in school availability between districts, where most districts situated on the lower 2 tiers.

While it is not the intent of this study to infer causality of socio development with availability of schools, it is expected that a more affluent district would have more schools available, either public or private. This showed to be true for most districts, with a few exceptions. Those affluent districts with low availability of schools can be a target for further study or a potential location for a private schools installation.