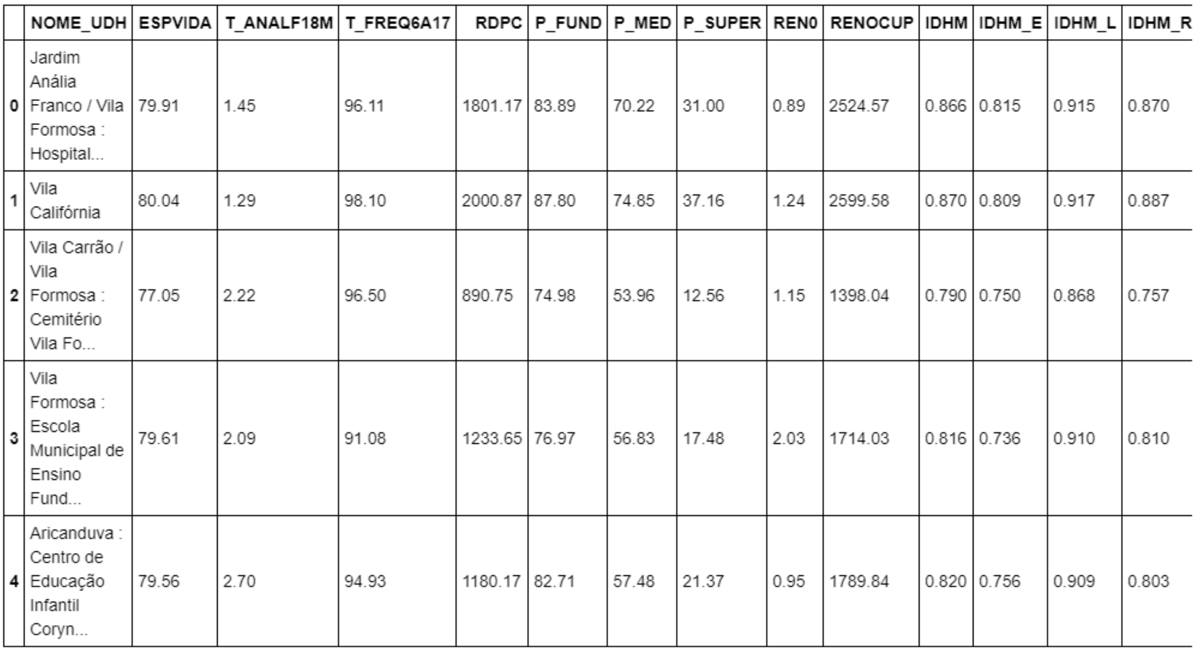
# Sao Paulo Education Inequalities and its impact

## 2. Data Collection and Preparation

**Data set 1:** Socio-Demografic 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



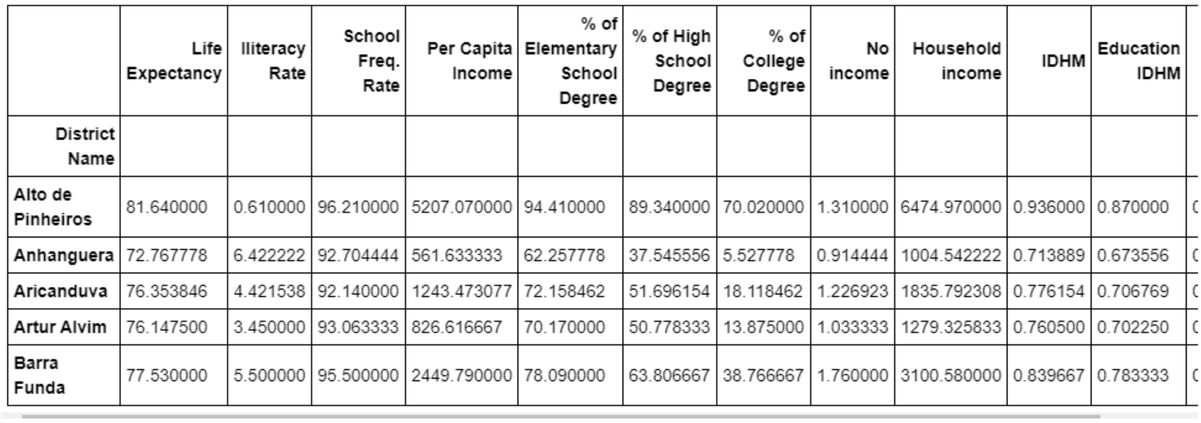
| | NOME_UDH | ESPVIDA | T_ANALF18M | T_FREQ6A17 | RDPC | P_FUND | P_MED | P_SUPER | REN0 | RENOCUP | IDHM | IDHM_E | IDHM_L | IDHM_R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Jardim Anália Franco / Vila Formosa : Hospital... | 79.91 | 1.45 | 96.11 | 1801.17 | 83.89 | 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 |

Data Preparation:

1)Select the relevant indexes for this study

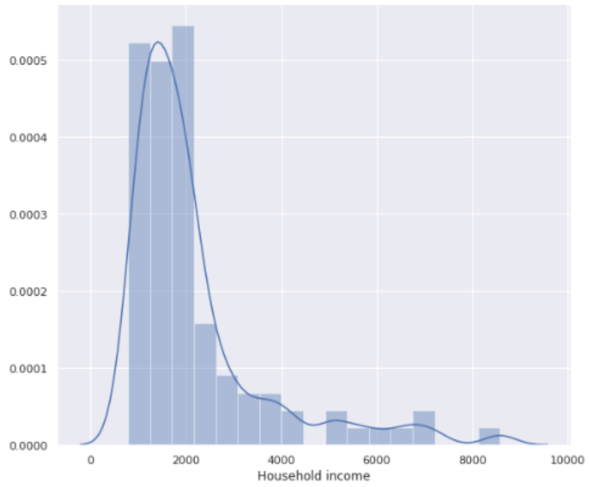
Life Expectancy, Illiteracy Rate, School Freq. Rate, Per Capita Income, % of Elementary School Degree, % of High School Degree, % of College Degree, No income, Household income, IDHM, Education IDHM, Health IDHM , Income IDHM, District Name

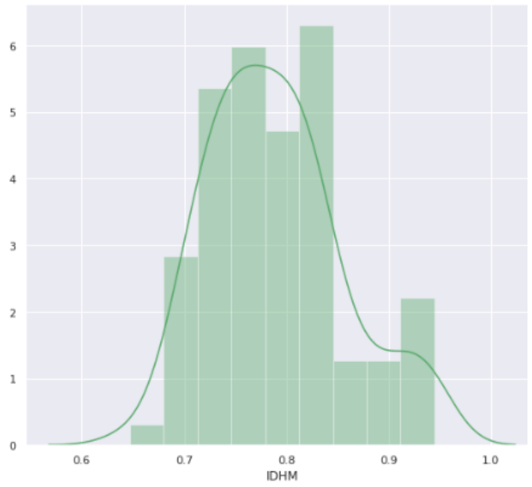
2) Group by each districts



| | Life Expectancy | Iliteracy Rate | School Freq. Rate | Per Capita Income | % of Elementary School Degree | % of High School Degree | % of College Degree | No income | Household income | IDHM | Education IDHM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **District Name** | | | | | | | | | | | |
| **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 |
| **Anhanguera** | 72.767778 | 6.422222 | 92.704444 | 561.633333 | 62.257778 | 37.545556 | 5.527778 | 0.914444 | 1004.542222 | 0.713889 | 0.673556 |
| **Aricanduva** | 76.353846 | 4.421538 | 92.140000 | 1243.473077 | 72.158462 | 51.696154 | 18.118462 | 1.226923 | 1835.792308 | 0.776154 | 0.706769 |
| **Artur Alvim** | 76.147500 | 3.450000 | 93.063333 | 826.616667 | 70.170000 | 50.778333 | 13.875000 | 1.033333 | 1279.325833 | 0.760500 | 0.702250 |
| **Barra Funda** | 77.530000 | 5.500000 | 95.500000 | 2449.790000 | 78.090000 | 63.806667 | 38.766667 | 1.760000 | 3100.580000 | 0.839667 | 0.783333 |

3) Check the data consistency and basic statistics. Ex.: Histograms

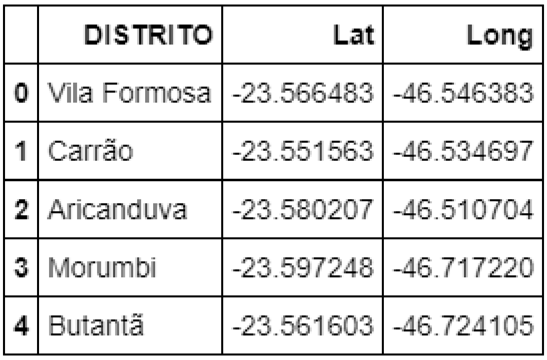




**Data set 2:** Geografical Data for district location

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

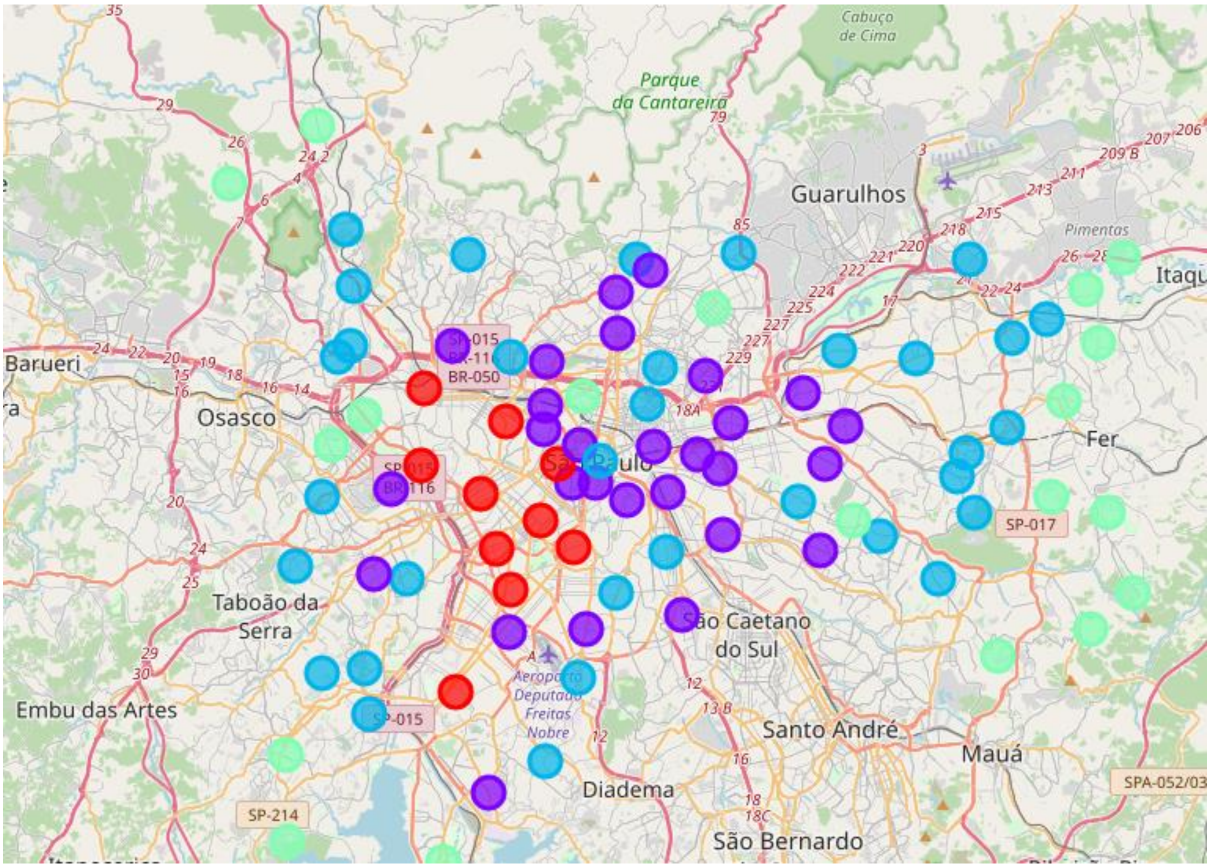


|   | DISTRITO | Lat | Long |
|---|---|---|---|
| 0 | Vila Formosa | -23.566483 | -46.546383 |
| 1 | Carrão | -23.551563 | -46.534697 |
| 2 | Aricanduva | -23.580207 | -46.510704 |
| 3 | Morumbi | -23.597248 | -46.717220 |
| 4 | Butantã | -23.561603 | -46.724105 |

Data Preparation:

1) Join that information with the Social Index DataFrame

2) Plot in a map the location of each district



**Data set 3:** Availability of Schools for each districts

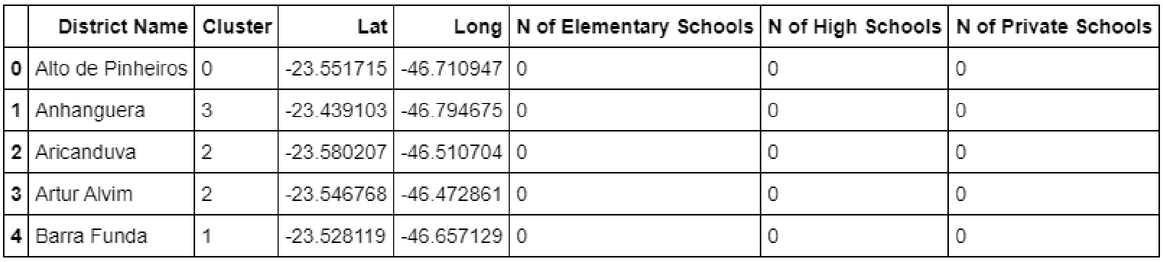
<u>Source:</u> Foursquare API

<u>Data Content:</u> Information of a venue in a given radius of a location

<u>Data Preparation</u>:

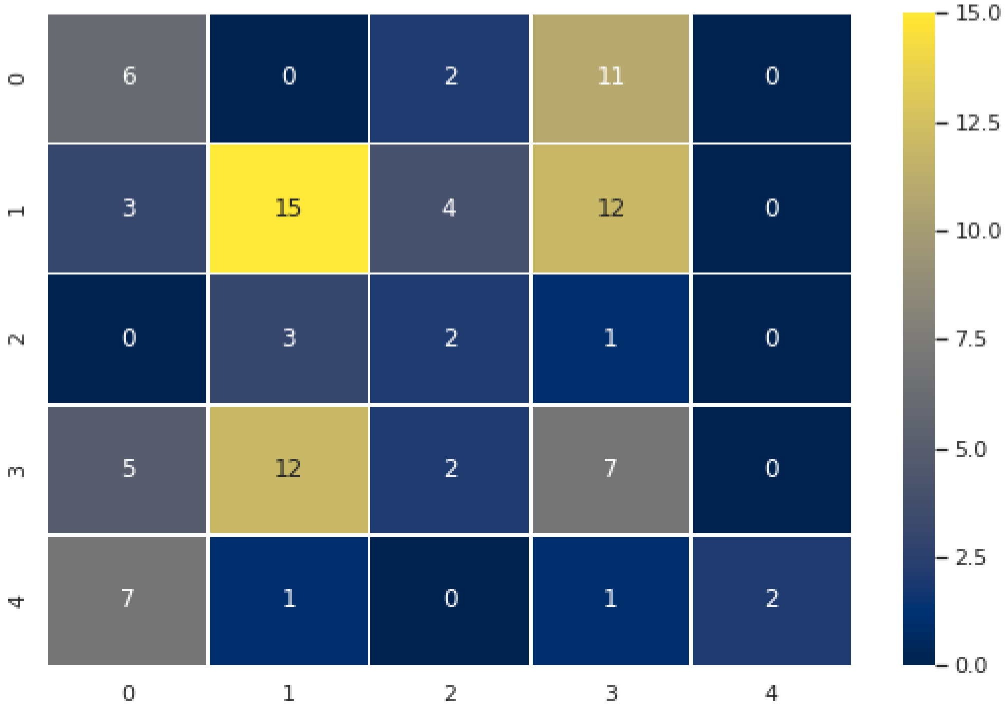
1) Fetch the number of venues within a 3km radius of districs center for the categories bellow:

```
Elementary_SchoolID= '4f4533804b9074f6e4fb0105'
Middle_SchoolID='4f4533814b9074f6e4fb0106'
Private_SchoolID='52e81612bcbc57f1066b7a46'
```



| | District Name | Cluster | Lat | Long | N of Elementary Schools | N of High Schools | N of Private Schools |
|---|---|---|---|---|---|---|---|
| 0 | Alto de Pinheiros | 0 | -23.551715 | -46.710947 | 0 | 0 | 0 |
| 1 | Anhanguera | 3 | -23.439103 | -46.794675 | 0 | 0 | 0 |
| 2 | Aricanduva | 2 | -23.580207 | -46.510704 | 0 | 0 | 0 |
| 3 | Artur Alvim | 2 | -23.546768 | -46.472861 | 0 | 0 | 0 |
| 4 | Barra Funda | 1 | -23.528119 | -46.657129 | 0 | 0 | 0 |

2) After using Scikit-learn to cluster the districts by K-mean method separately by data set 1 and data set 3, prepare a correlation matrix and check for a tendency and an eventual outlier



| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 6 | 0 | 2 | 11 | 0 |
| 1 | 3 | 15 | 4 | 12 | 0 |
| 2 | 0 | 3 | 2 | 1 | 0 |
| 3 | 5 | 12 | 2 | 7 | 0 |
| 4 | 7 | 1 | 0 | 1 | 2 |