Spatial Analysis of racial and income inequality in São Paulo.

André Leite Rodrigues

Jun/2020

1. Introduction.

Brazil has the second highest concentration of income in the world, says UN report [1]. In Brazil, the richest 1% concentrates 28.3% of the country's total income. That is, almost a third of the income is in the hands of the wealthiest. The richest 10% in Brazil, on the other hand, account for 41.9% of the total income.

This inequality of income reflects in the occupation of urban spaces. In the city of São Paulo it's possible see big mansions in the best places of the city (in terms of infrastructure) in contrast with big favelas in the poorest places.

But in addition to income inequality, Brazil still faces a problem of structural racism. Dimensions of our history and culture that have allowed privileges associated with “whiteness” and disadvantages associated with “color” to endure and adapt over time.[2] Brazil was the last country in the Americas to abolish slavery. Without compensation or preparation to insert these ex-slaves into society and without treating racism as a problem, more than 130 years after the abolition the Brazilian inequality is colored: the majority of the poor population is black in contrast to the rich population where the majority is white .

1.1 Objective

Use the techniques learned during the 9 Data Science course modules offered by IBM on the Coursera platform to analyze data from the São Paulo census, to understand if race is correlated with income inequality and whether this inequality is reflected in the occupation of space in the city of São Paulo.

2. Data acquisition and source

First, the census data from the Brazilian Institute of Geography and Statistics (IBGE) will be analyzed [3]. The IBGE promotes the national census every 10 years, but due to COVID-19 this year's census is delayed. Therefore, in this study I will use the data from the 2010 census. The census data are available for everyone to access on the Internet at www.ibge.gov.br in the downloads section> Census> 2010 population census> Universe results> Aggregated by census sectors . You can download all research documentation as well.

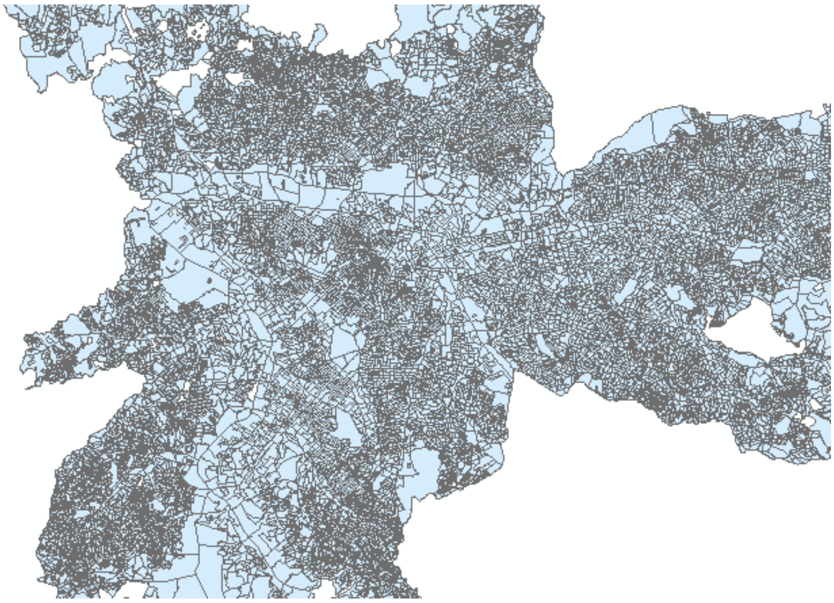
The files used for this analysis will be:

- DomicílioRenda_SP1.csv: file containing the average nominal income of the census sector. From this table, the following variables will be used:
    - Cod_setor: census sector id;
    - V002: Total nominal monthly income of private households.

- Pessoa03_SP1.csv: file containing the number of resident people and the declared race. From this table the following variables will be used:

- Cod_setor: census sector id;
- V001: Total resident people;
- V002: Resident people and color or race – white;
- V003: Resident people and color or race - black;
- V004: Resident people and color or race – yellow;
- V005: Resident people and color or race - brown;
- V006: Resident people and color or race – native;
- V087 to V116: Residents man;
- V167 to V246: Residents woman;

The geographical unit worked on in this study is the census sector. The census sector is the territorial unit established for the purpose of cadastral control, formed by a continuous area, located in a single urban or rural framework, with the size and number of households that allow for the survey by a census taker. IBGE provides the geographical limits of the census sector in .kmz and .shp formats. They are spatial vector formats in the shape of polygons. For this study I will use only the centroid of these areas as a geographical reference, transforming the polygons into points and converting them to the.json format. The figure below illustrates the division of São Paulo into census sectors.



The city of São Paulo has 18363 census tracts, each sector has an average of 300 households.

To understand the distribution of Health, Education and Culture equipment, forecast data from São Paulo will be used, available at http://geosampa.prefeitura.sp.gov.br/.

To access the data just enter the website and go to "downloads"> "equipment" and choose as bases. All variables already have a mark to which census sector they belong to.Finally, the last database will be built by Foursquare's API venues. It is mandatory and I will extract information from two diferente districts for compare.

3. Methodology

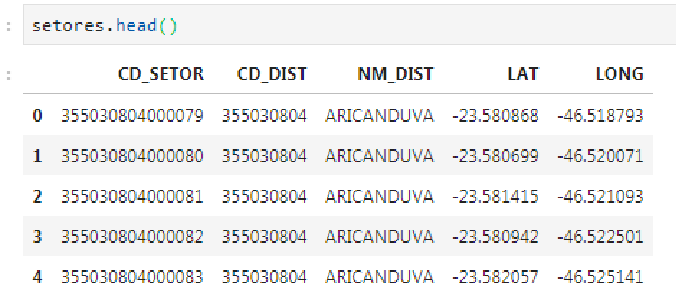
3.1. Space census sector data cleaning and structuring

The first data treated is the census sectors. Firstly, this data is made available in the shapefile (.shp) format, so I started its treatment in a specific tool for spatial data, ArcGIS. There are other tools available for download that can provide the treatment of this data, but as I have the ArcGIS license I preferred to work with this tool using the python language and its arcpy library.

After downloading from the IBGE website, some treatments are required to use the data. IBGE uses the SIRGAS 2000 cartographic reference system, but most open maps work with WGS84. Cartographic projections are systems of geographic coordinates, consisting of parallel meridians (imaginary semicircle drawn from one pole of the Earth to another) (imaginary lines parallel to the Equator), on which the spherical surface of the Earth can be represented.

In this script I selected only the city of São Paulo (3550308), area of interest for the study. Then I transformed the shape from polygons to points and exported it to the json format.

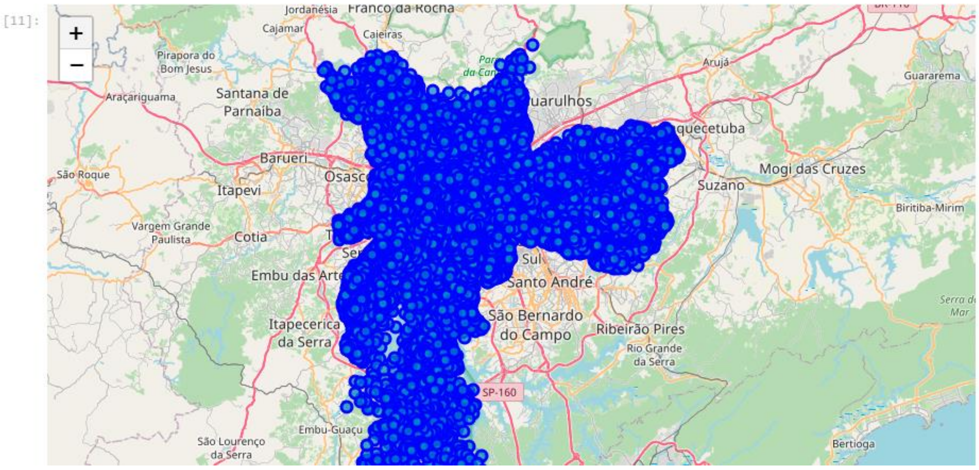
Next step is upload json file para o notebook and transforming this data of nested Python dictionaries into a pandas dataframe. So let's start by creating an empty dataframe. Then loop through the data and fill the dataframe one row at a time. The result is the follw table



```
setores.head()
```

	CD_SETOR	CD_DIST	NM_DIST	LAT	LONG
0	355030804000079	355030804	ARICANDUVA	-23.580868	-46.518793
1	355030804000080	355030804	ARICANDUVA	-23.580699	-46.520071
2	355030804000081	355030804	ARICANDUVA	-23.581415	-46.521093
3	355030804000082	355030804	ARICANDUVA	-23.580942	-46.522501
4	355030804000083	355030804	ARICANDUVA	-23.582057	-46.525141

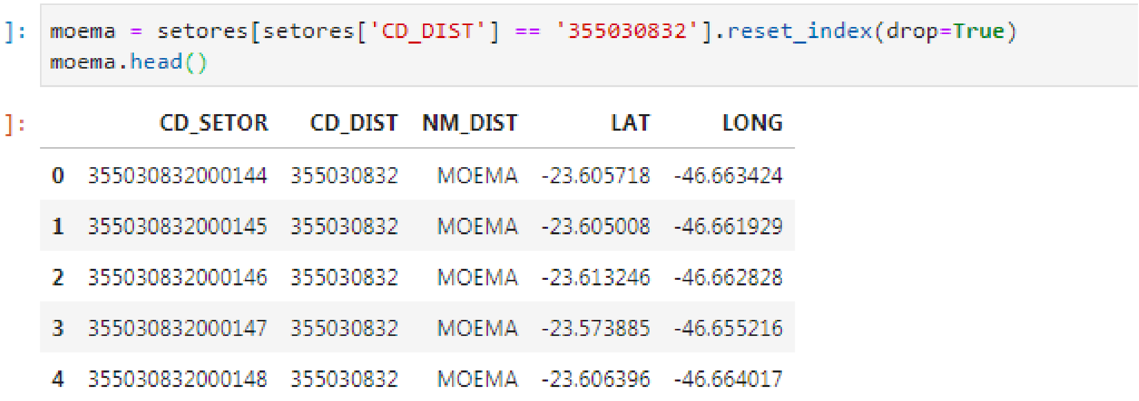
And to finish this step, a map was created to see if the data are consistent.



3.2. Extracting and preparing FourSquare data – Points of Interest (POI).

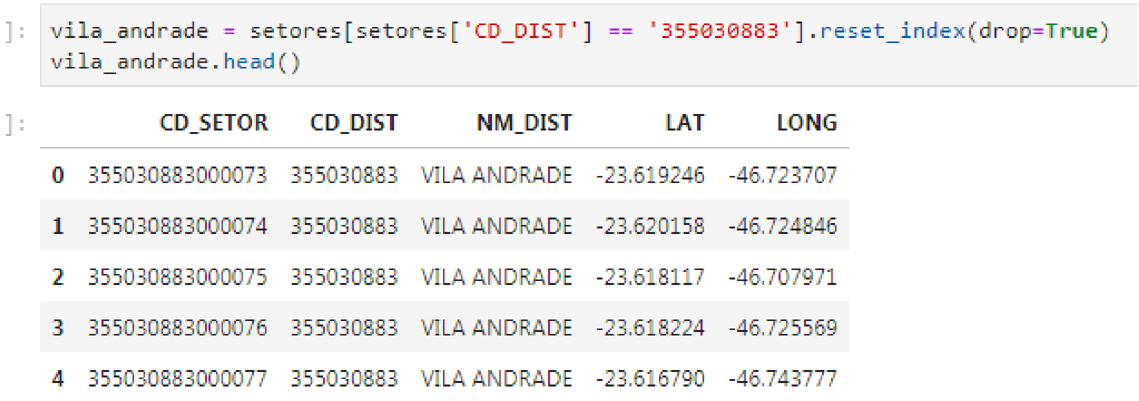
As previously described, São Paulo has more than 18,000 census sectors, which would be too much to get all information via the FourSquare API. For specific analysis of points of interest using this API, only 2 districts of the city will be analyzed: Moema and Vila Andrade.

The first step is to register on the FourSquare developer portal, create a CLIENT ID and CLIENT SECRET for our application. With that we can start the data extraction process. Then, select the census sectors belonging to the districts of Moema and Vila Andrade were selected.



```python
moema = setores[setores['CD_DIST'] == '355030832'].reset_index(drop=True)
moema.head()
```

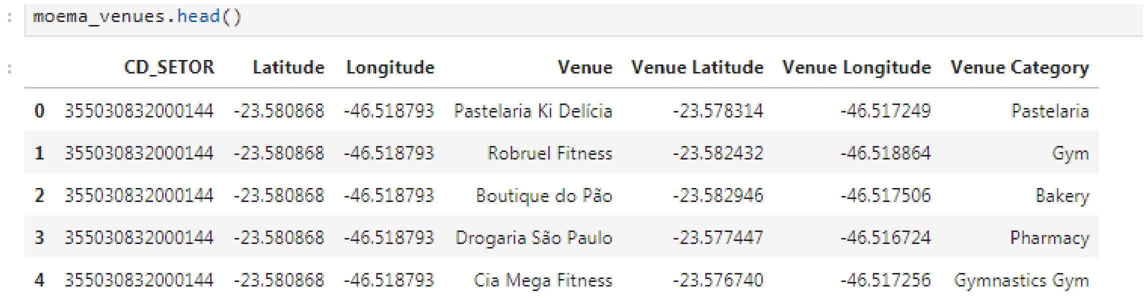
	CD_SETOR	CD_DIST	NM_DIST	LAT	LONG
0	355030832000144	355030832	MOEMA	-23.605718	-46.663424
1	355030832000145	355030832	MOEMA	-23.605008	-46.661929
2	355030832000146	355030832	MOEMA	-23.613246	-46.662828
3	355030832000147	355030832	MOEMA	-23.573885	-46.655216
4	355030832000148	355030832	MOEMA	-23.606396	-46.664017



```python
vila_andrade = setores[setores['CD_DIST'] == '355030883'].reset_index(drop=True)
vila_andrade.head()
```

	CD_SETOR	CD_DIST	NM_DIST	LAT	LONG
0	355030883000073	355030883	VILA ANDRADE	-23.619246	-46.723707
1	355030883000074	355030883	VILA ANDRADE	-23.620158	-46.724846
2	355030883000075	355030883	VILA ANDRADE	-23.618117	-46.707971
3	355030883000076	355030883	VILA ANDRADE	-23.618224	-46.725569
4	355030883000077	355030883	VILA ANDRADE	-23.616790	-46.743777

After that, a process was made to consult each of the id codes of the census sectors and search for the first 10 venues within a radius of 500 meters, creating a table containing with the CD_SETOR, the Venue and its category (in addition to the respective latitudes and longitudes).



```python
moema_venues.head()
```

	CD_SETOR	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	355030832000144	-23.580868	-46.518793	Pastelaria Ki Delícia	-23.578314	-46.517249	Pastelaria
1	355030832000144	-23.580868	-46.518793	Robruel Fitness	-23.582432	-46.518864	Gym
2	355030832000144	-23.580868	-46.518793	Boutique do Pão	-23.582946	-46.517506	Bakery
3	355030832000144	-23.580868	-46.518793	Drogaria São Paulo	-23.577447	-46.516724	Pharmacy
4	355030832000144	-23.580868	-46.518793	Cia Mega Fitness	-23.576740	-46.517256	Gymnastics Gym

The same procedure was carried out for the Vila Andrade district. Finally, the data was exported to csv files.

3.3. Creating SQL DataBase in IBM Db2 on Cloud

A cloud database is a database service built and accessed through a cloud platform. It serves many of the same functions as traditional databases with the added flexibility of cloud computing.

To create this database I navigate to IBM Cloud and select the Db2 on Cloud service. Type a service instance name, choose the region to deploy to, as well as an org and space for the service, then click create. You should be able to see details related to connection configuration when you open the web console for the Db2 on Cloud service. The connection details include the following: a host name, which is a unique name or label assigned to any device that is connected to a specific computer network. A port number, which is the database port. The database name, which is the database name. A user ID, which is the username you'll use to connect. Password, is the password you'll use to connect.

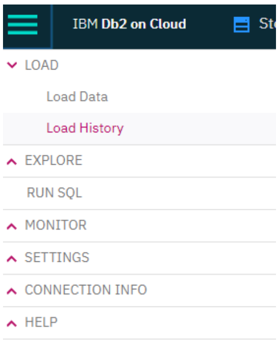
3.3.1. Loading Tables

The follow tables was load in Db2:

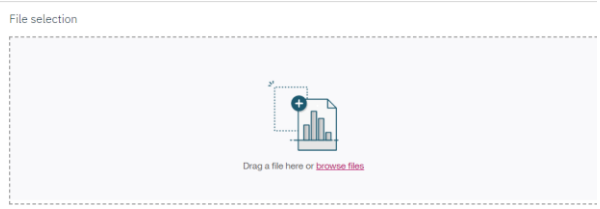
- Pessoa03_SP1.csv as POPULATION: this table contain the data about color, race and gender from IBGE CENSUS;
- DomicílioRenda_SP1.csv as INCOME: this table contain the data about income from IBGE CENSUS;
- GEOINFO_CT_CULTURA_2018_Variáveis.csv as CULTURA: Information from São Paulo city about culture equipaments like librarys, theaters;

- GEOINFO_CT_EDUCACAO_2018_Variáveis.csv as EDUCATION: Information from São Paulo city about schools and students;
- GEOINFO_CT_SAUDE_2018_Variáveis.csv as HEALTH: Information from São Paulo city about clinics, First Aid centers, Hospitals and beds;
- Setores_censo_2010 as SETORES: Information about sector ID and district id.

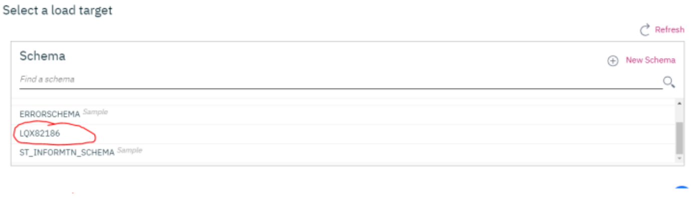
To upload tables, it’s just open IBM Db2 console and click in Load Data:



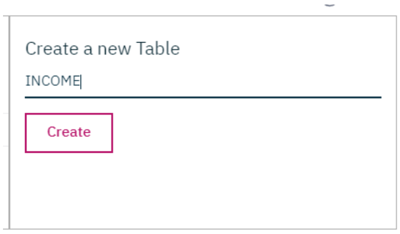
The next step is choose the file



Select a load target



Create New Table

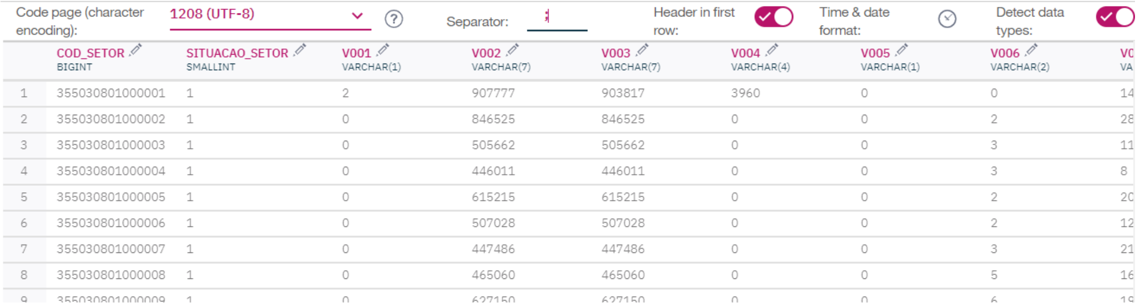


Create a new Table

INCOME

Create

Verify the fields



Code page (character encoding): 1208 (UTF-8)

Separator: ;

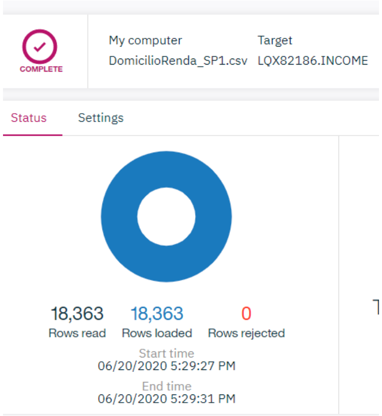
Header in first row: ✓

Time & date format:

Detect data types: ✓

	COD_SETOR BIGINT	SITUACAO_SETOR SMALLINT	V001 VARCHAR(1)	V002 VARCHAR(7)	V003 VARCHAR(7)	V004 VARCHAR(4)	V005 VARCHAR(1)	V006 VARCHAR(2)	V0 VA
1	355030801000001	1	2	907777	903817	3960	0	0	14
2	355030801000002	1	0	846525	846525	0	0	2	28
3	355030801000003	1	0	505662	505662	0	0	3	11
4	355030801000004	1	0	446011	446011	0	0	3	8
5	355030801000005	1	0	615215	615215	0	0	2	20
6	355030801000006	1	0	507028	507028	0	0	2	12
7	355030801000007	1	0	447486	447486	0	0	3	21
8	355030801000008	1	0	465060	465060	0	0	5	16
9	355030801000009	1	0	627150	627150	0	0	6	19

And upload



COMPLETE

My computer: DomicilioRenda_SP1.csv

Target: LQX82186.INCOME

Status | Settings

18,363 Rows read

18,363 Rows loaded

0 Rows rejected

Start time
06/20/2020 5:29:27 PM

End time
06/20/2020 5:29:31 PM

3.3.2. Prepare table Summary sector

Back to the notebook, I imported 3 libraries: ibm_db, pandas and ibm_db_dbi. And loaded the SQL extension. The follow steps was create a database connection with my credentials and conect with my database.

Using SQL query, I group by sector ID, count the number of equipaments per sector and sum the number of students in EDUCATION table. I Do the same metodology for HEALTH but sum the number of hospital beds e for CULTURE I just count the number of equipaments.

This results in 3 new tables, who I export to csv and upload in Db2 as SETOR_CULT, SETOR_EDUC and SETOR_HEALTH.

The sector id is may key between this tables. Using the follow script I created the summary_sector table:

SELECT Setores.CD_SETOR, Setores.CD_DIST, Setores.NM_DIST, POPULATION.V001 AS Pop_TT, POPULATION.V002 AS Pop_white, DOUBLE(POPULATION.V002)/DOUBLE(POPULATION.V001) AS PERC_WHITE, POPULATION.V003+POPULATION.V005 AS Pop_black, DOUBLE(POPULATION.V003+POPULATION.V005)/DOUBLE(POPULATION.V001) AS PERC_BLACK, POPULATION.V004+POPULATION.V006 AS POP_other, DOUBLE(POPULATION.V004+POPULATION.V006)/DOUBLE(POPULATION.V001) AS PERC_OTHER, INCOME.V002 AS TT_INCOME, DOUBLE(INCOME.V002)/DOUBLE(POPULATION.V001) AS INCOME_PER_CAPTA, POPULATION.VMAN AS MAN, DOUBLE(POPULATION.VMAN)/DOUBLE(POPULATION.V001) AS PERC_MAN, POPULATION.VWOMAN AS WOMAN, DOUBLE(POPULATION.VWOMAN)/DOUBLE(POPULATION.V001) AS PERC_WOMAN, SETOR_EDUC.TT_SCHOOLS, SETOR_EDUC.TT_STUDENTS, SETOR_CULT.TT_CULTURE_EQUP, SETOR_HEALTH.TT_HEALTH_EQUIP, SETOR_HEALTH.TT_LEITOS

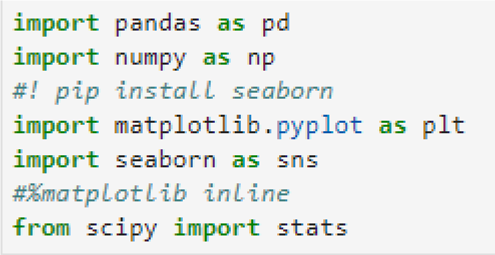
FROM ((((Setores LEFT JOIN POPULATION ON Setores.CD_SETOR = POPULATION.Cod_setor) LEFT JOIN INCOME ON Setores.CD_SETOR = INCOME.Cod_setor) LEFT JOIN SETOR_EDUC ON Setores.CD_SETOR = SETOR_EDUC.SETCENS) LEFT JOIN SETOR_CULT ON Setores.CD_SETOR = SETOR_CULT.SETCENS) LEFT JOIN SETOR_HEALTH ON Setores.CD_SETOR = SETOR_HEALTH.SETCENS

WHERE (((POPULATION.V001) Is Not Null));

Finally I use pandas.read_sql to create a pandas table.

3.4. Exploratory data analysis.

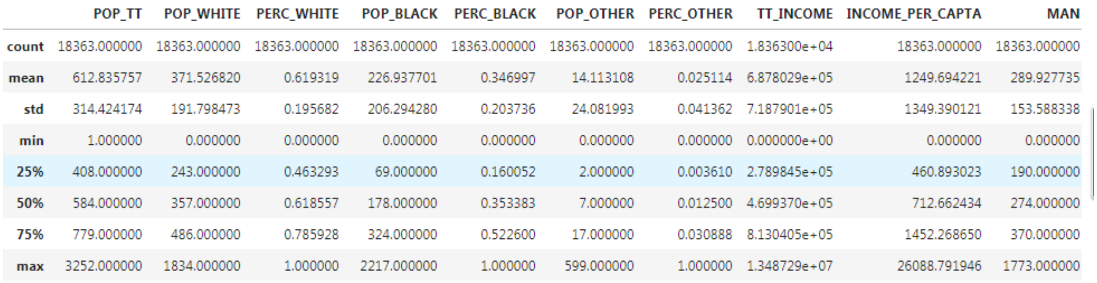
First import libraries:



```
import pandas as pd
import numpy as np
#! pip install seaborn
import matplotlib.pyplot as plt
import seaborn as sns
#%matplotlib inline
from scipy import stats
```

3.4.1. Analysis by census sectors

For a first view of the Summary_sector table I use the .describe () function so I can evaluate the count, average, standard deviation, minimum, maximum and quartiles for each variable.

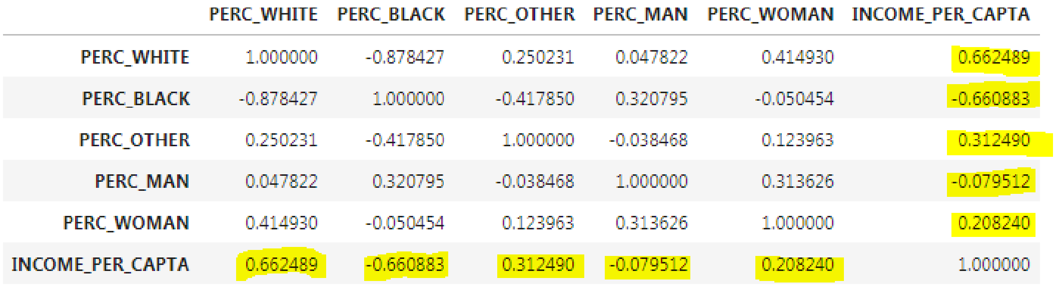


	POP_TT	POP_WHITE	PERC_WHITE	POP_BLACK	PERC_BLACK	POP_OTHER	PERC_OTHER	TT_INCOME	INCOME_PER_CAPTA	MAN
count	18363.000000	18363.000000	18363.000000	18363.000000	18363.000000	18363.000000	18363.000000	1.836300e+04	18363.000000	18363.000000
mean	612.835757	371.526820	0.619319	226.937701	0.346997	14.113108	0.025114	6.878029e+05	1249.694221	289.927735
std	314.424174	191.798473	0.195682	206.294280	0.203736	24.081993	0.041362	7.187901e+05	1349.390121	153.588338
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000
25%	408.000000	243.000000	0.463293	69.000000	0.160052	2.000000	0.003610	2.789845e+05	460.893023	190.000000
50%	584.000000	357.000000	0.618557	178.000000	0.353383	7.000000	0.012500	4.699370e+05	712.662434	274.000000
75%	779.000000	486.000000	0.785928	324.000000	0.522600	17.000000	0.030888	8.130405e+05	1452.268650	370.000000
max	3252.000000	1834.000000	1.000000	2217.000000	1.000000	599.000000	1.000000	1.348729e+07	26088.791946	1773.000000

As expected, the variables related to population and income are filled for all 18,363 census sectors.

Only 4,727 sectors have schools, 721 cultural facilities and 855 health facilities. Therefore, I will exclude the use of this information for analysis at the scale of the census sector. This information will be better used when the level of scale is the district.

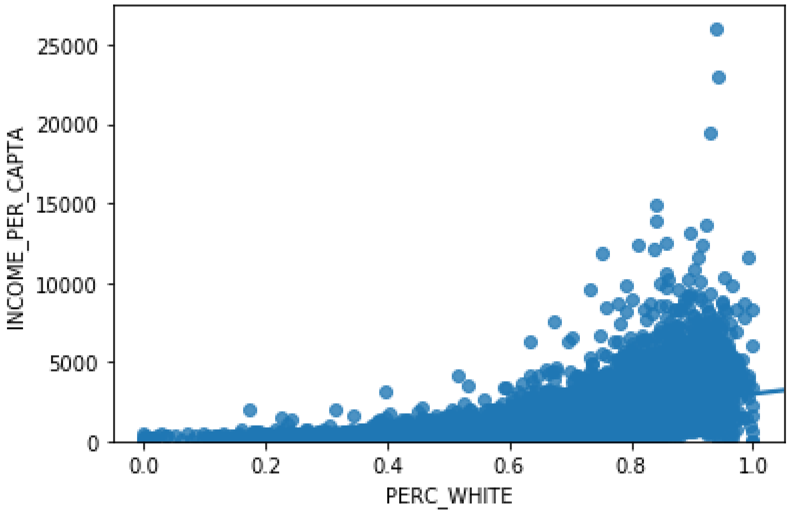
Given this, I establish the correlation between the proportional variables of the population in each district and the per-capita income.



	PERC_WHITE	PERC_BLACK	PERC_OTHER	PERC_MAN	PERC_WOMAN	INCOME_PER_CAPTA
PERC_WHITE	1.000000	-0.878427	0.250231	0.047822	0.414930	0.662489
PERC_BLACK	-0.878427	1.000000	-0.417850	0.320795	-0.050454	-0.660883
PERC_OTHER	0.250231	-0.417850	1.000000	-0.038468	0.123963	0.312490
PERC_MAN	0.047822	0.320795	-0.038468	1.000000	0.313626	-0.079512
PERC_WOMAN	0.414930	-0.050454	0.123963	0.313626	1.000000	0.208240
INCOME_PER_CAPTA	0.662489	-0.660883	0.312490	-0.079512	0.208240	1.000000

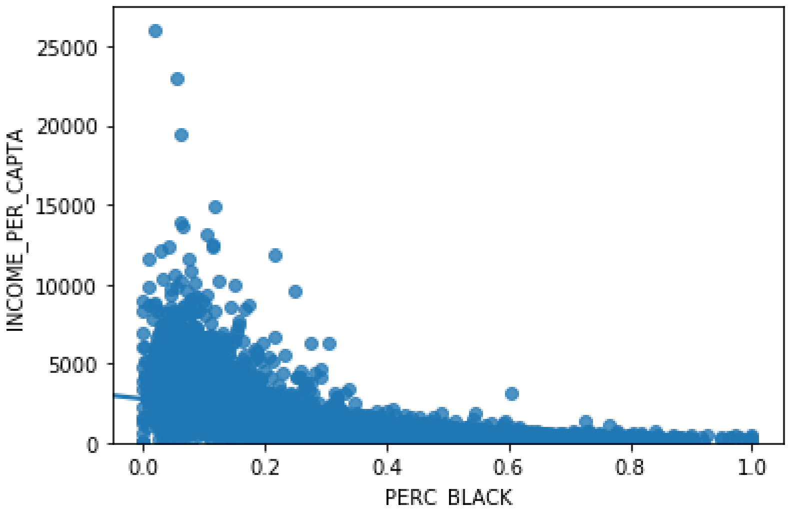
By the Pearson correlation index, a moderate / strong correlation was identified between income and percentages of white and black people, however the correlation at the level of the census sector was weak in relation to gender. There is also a strong correlation between the percentages of whites and blacks, which may indicate that there may be spatial segregation.

The white population of the city represents 60.6% of the total declared to the IBGE. When constructing the dispersion graph of the census areas, we can observe that there is a growth correlation between the percentage of white people in the sector and the income per capita.

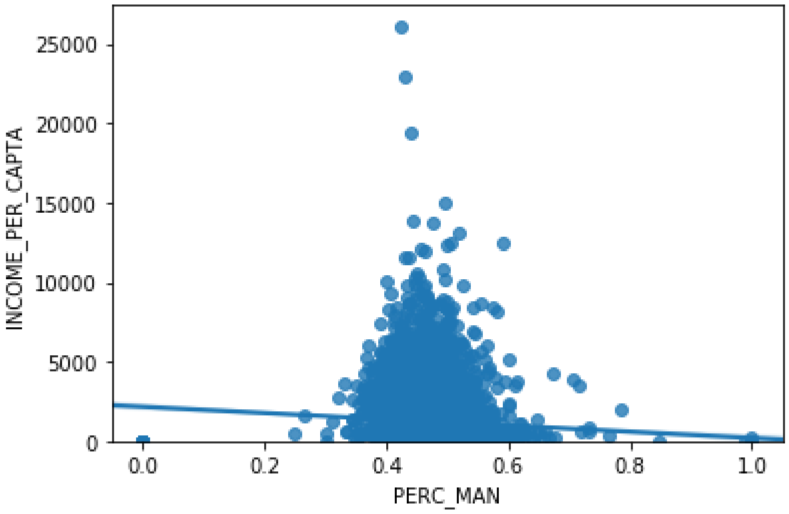


The Pearson Correlation Coefficient is 0.6624886961070122 with a P-value of P = 0.0

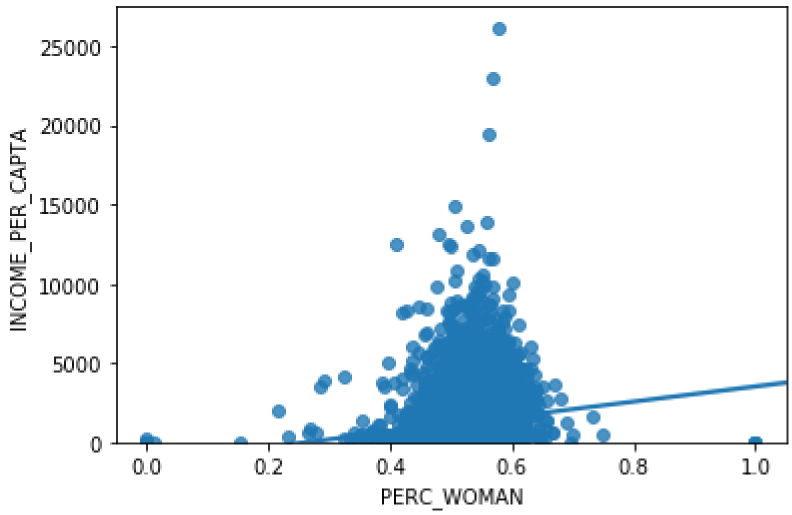
The city's black population represents 37% of the total declared to the IBGE. When constructing the dispersion graph of the census areas, we can observe that there is a decreasing correlation between the percentage of black people in the sector and the income per capita.



The population of men in the city of São Paulo is 5,323,943 and represents 47.3% of the total population. The Pearson Correlation Coefficient is -0.07951240563479849 with a P-value of P = 3.7925656303285465e-27



The population of women in the city of São Paulo is 5,925,231 and represents 52.7% of the total population. The Pearson Correlation Coefficient is 0.20824026521970707 with a P-value of P = 4.9674315895493 82e-179

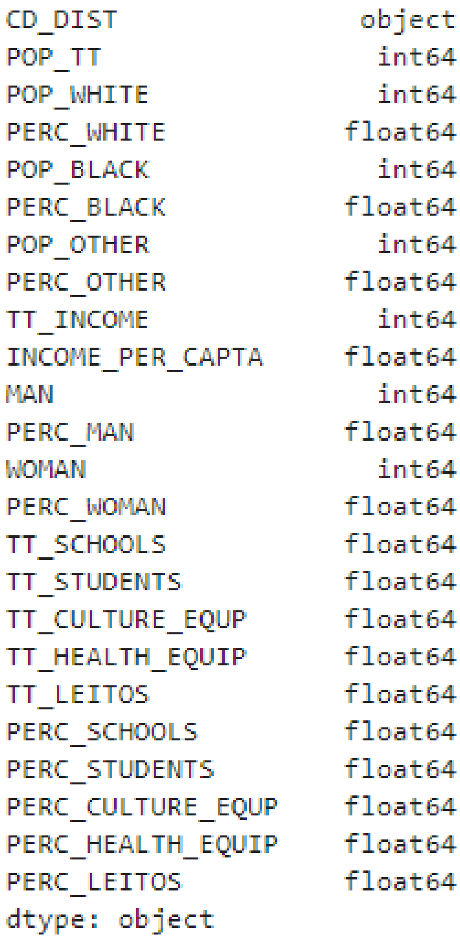


3.4.2. Analysis by districts.

Districts are territories in which municipalities are subdivided into which administrative, judicial, fiscal, police or health authority is exercised. Each census sector is spatially contained in a district of the city. All 18,363 census sector areas can be grouped into 96 city districts.

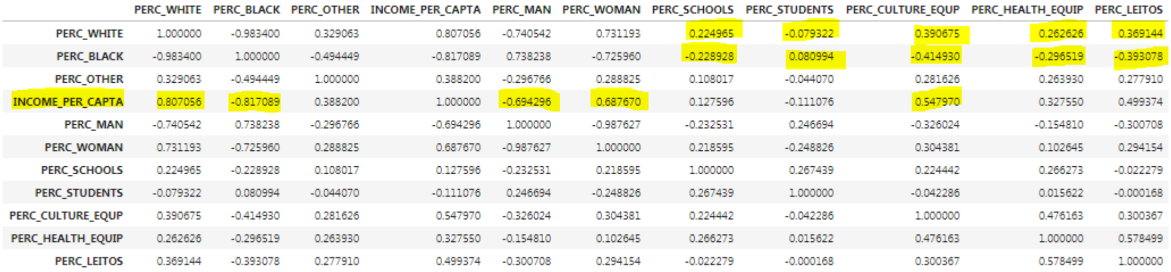
The first 9 digits of the sector ID form the district ID.

The first step was to group the CD_DIST column using the groupby function and sum the variables with absolute values. Then, all variables were proportional to the total number of the population for us to compare, including information on health equipment, culture and schools.



```
CD_DIST                 object
POP_TT                   int64
POP_WHITE                int64
PERC_WHITE             float64
POP_BLACK                int64
PERC_BLACK             float64
POP_OTHER                int64
PERC_OTHER             float64
TT_INCOME                int64
INCOME_PER_CAPTA       float64
MAN                      int64
PERC_MAN               float64
WOMAN                    int64
PERC_WOMAN             float64
TT_SCHOOLS             float64
TT_STUDENTS            float64
TT_CULTURE_EQUP        float64
TT_HEALTH_EQUIP        float64
TT_LEITOS              float64
PERC_SCHOOLS           float64
PERC_STUDENTS          float64
PERC_CULTURE_EQUP      float64
PERC_HEALTH_EQUIP      float64
PERC_LEITOS            float64
dtype: object
```

The first action was to create an overview of the correlation between the proportional variables in the table.

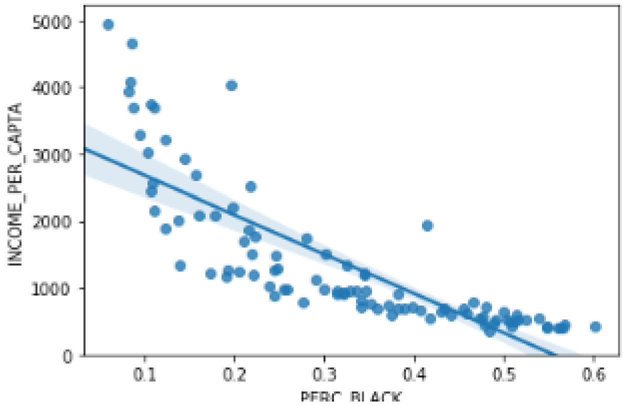


	PERC_WHITE	PERC_BLACK	PERC_OTHER	INCOME_PER_CAPTA	PERC_MAN	PERC_WOMAN	PERC_SCHOOLS	PERC_STUDENTS	PERC_CULTURE_EQUP	PERC_HEALTH_EQUIP	PERC_LEITOS
PERC_WHITE	1.000000	-0.983400	0.329063	0.807056	-0.740542	0.731193	0.224965	-0.079322	0.390675	0.262626	0.369144
PERC_BLACK	-0.983400	1.000000	-0.494449	-0.817089	0.738238	-0.725960	-0.228928	0.080994	-0.414930	-0.296519	-0.393078
PERC_OTHER	0.329063	-0.494449	1.000000	0.388200	-0.296766	0.288825	0.108017	-0.044070	0.281626	0.263930	0.277910
INCOME_PER_CAPTA	0.807056	-0.817089	0.388200	1.000000	-0.694296	0.687670	0.127596	-0.111076	0.547970	0.327550	0.499374
PERC_MAN	-0.740542	0.738238	-0.296766	-0.694296	1.000000	-0.987627	-0.232531	0.246694	-0.326024	-0.154810	-0.300708
PERC_WOMAN	0.731193	-0.725960	0.288825	0.687670	-0.987627	1.000000	0.218595	-0.248826	0.304381	0.102645	0.294154
PERC_SCHOOLS	0.224965	-0.228928	0.108017	0.127596	-0.232531	0.218595	1.000000	0.267439	0.224442	0.266273	-0.022279
PERC_STUDENTS	-0.079322	0.080994	-0.044070	-0.111076	0.246694	-0.248826	0.267439	1.000000	-0.042286	0.015622	-0.000168
PERC_CULTURE_EQUP	0.390675	-0.414930	0.281626	0.547970	-0.326024	0.304381	0.224442	-0.042286	1.000000	0.476163	0.300367
PERC_HEALTH_EQUIP	0.262626	-0.296519	0.263930	0.327550	-0.154810	0.102645	0.266273	0.015622	0.476163	1.000000	0.578499
PERC_LEITOS	0.369144	-0.393078	0.277910	0.499374	-0.300708	0.294154	-0.022279	-0.000168	0.300367	0.578499	1.000000

When placed at district level, the correlation between income and percentage of whites and blacks increases to 0.8 and the variables of structures such as schools, libraries, theaters, hospitals and beds show a growth correlation, being when the district has more whites and is decreasing when the district has more black people.

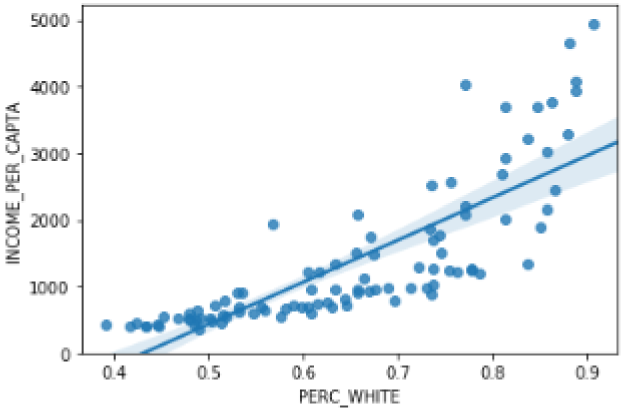
When comparing income at the district level we have:

Per capita income compared to the percentage of black population:



The Pearson Correlation Coefficient is -0.8170894312959299 with a P-value of P = 3.278212047348941e-24

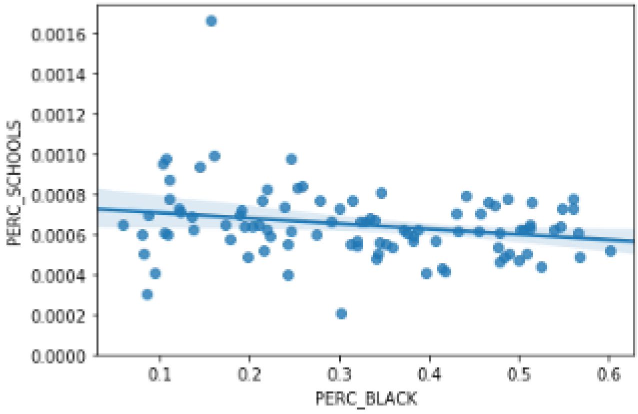
Per capita income compared to the percentage of white population:



The Pearson Correlation Coefficient is 0.8070564754629121 with a P-value of P = 3.1464332612695534e-23

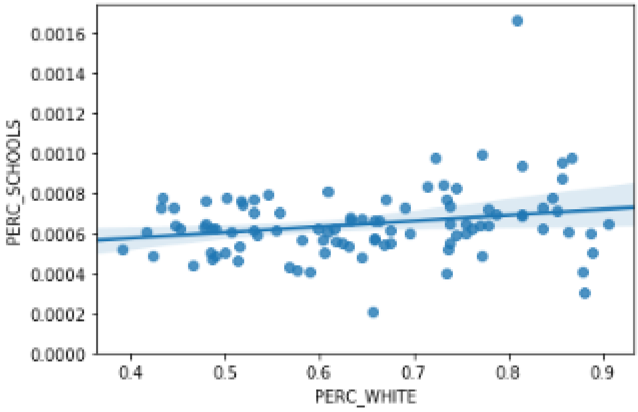
The same trend is observed for the availability of cultural facilities, schools and hospitals.

Proportion of schools per inhabitant compared to proportion of black population:



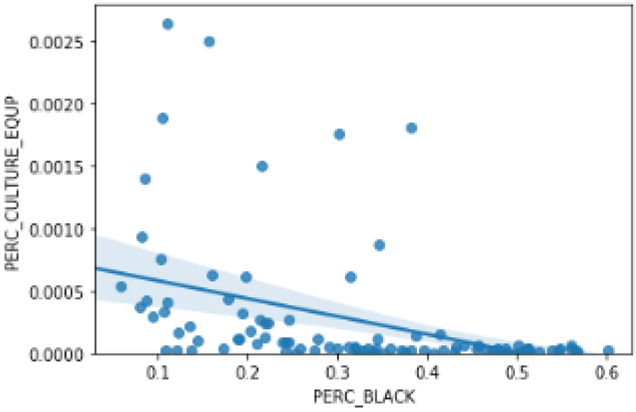
The Pearson Correlation Coefficient is -0.22892761686205754 with a P-value of P = 0.024863042345446253

Proportion of schools per inhabitant compared to proportion of white population:



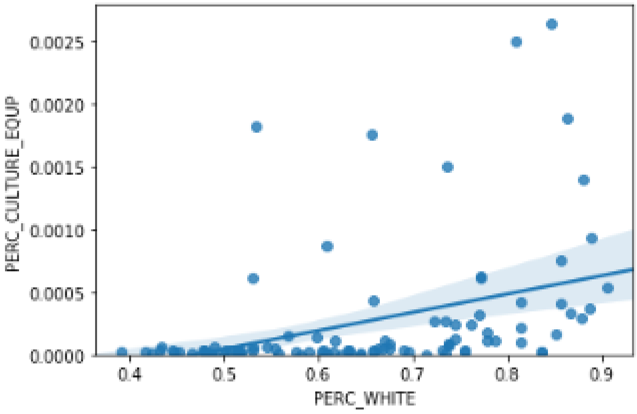
The Pearson Correlation Coefficient is 0.22496464720879644 with a P-value of P = 0.027548995953717785

Proportion of cultural spaces per inhabitant with proportion of black population:



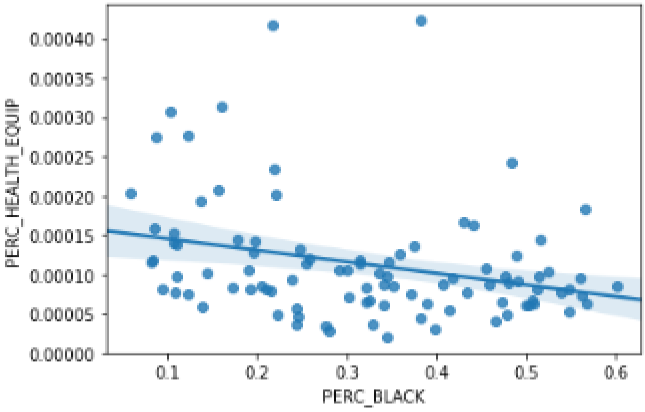
The Pearson Correlation Coefficient is -0.4149303403542528 with a P-value of P = 2.6301309434569084e-05

Proportion of cultural spaces per inhabitant with proportion of white population:



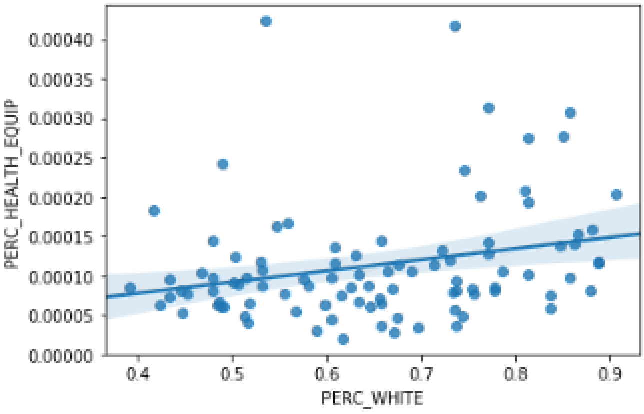
The Pearson Correlation Coefficient is 0.39067480552196565 with a P-value of P = 8.310631495373544e-05

Proportion of hospitals and healthcare facilities per capita with proportion of black population:



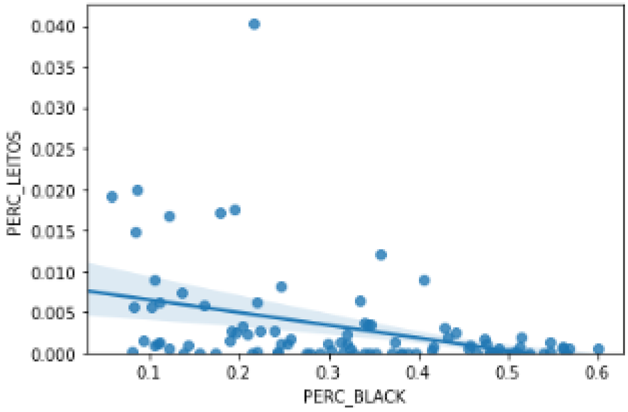
The Pearson Correlation Coefficient is -0.2965190544759442 with a P-value of P = 0.0033516445854855567

Proportion of hospitals and healthcare facilities per capita with proportion of white population:



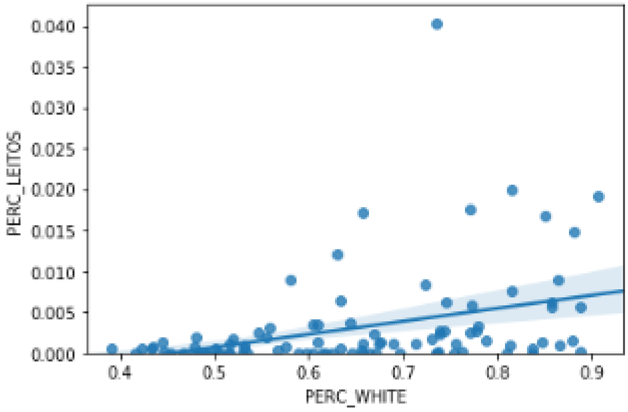
The Pearson Correlation Coefficient is 0.262626233881216 with a P-value of P = 0.009736798524530946

Proportion of hospital beds per inhabitant with proportion of black population:



The Pearson Correlation Coefficient is -0.39307811546210625 with a P-value of P = 7.444491517151547e-05

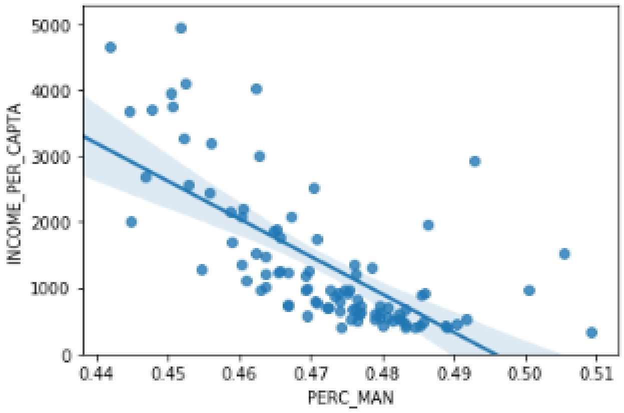
Proportion of hospital beds per inhabitant with proportion of white population:



The Pearson Correlation Coefficient is 0.36914449022611084 with a P-value of P = 0.00021466883451653064

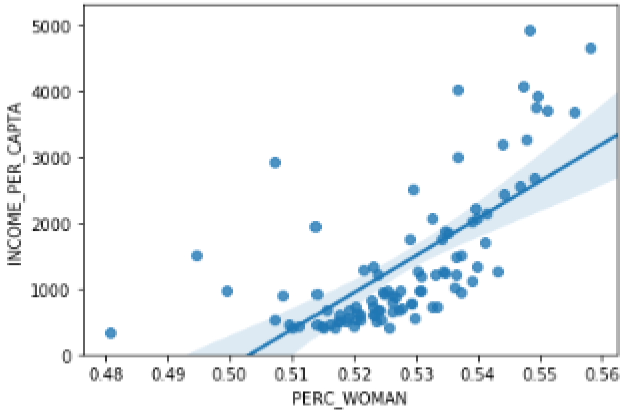
Income at the district level showed a moderate correlation depending on gender.

Per capita income and percentage of men in the district:



The Pearson Correlation Coefficient is -0.6942955688734946 with a P-value of P = 4.361287499926767e-15

Per capita income and percentage of women in the district:



The Pearson Correlation Coefficient is 0.6876696602137802 with a P-value of P = 1.0030213089943034e-14

3.5. Clustering using k-means

K-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data. In this study, I used k-Means for segmentation.

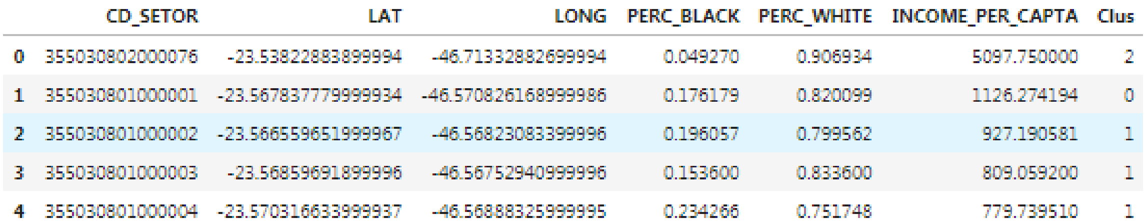
3.5.1. Census Sector

Based on the exploratory analysis of the data, the percentages of white population, percentages of black population and per capita income of the sector were used to compose this model.

The data were normalized on the standard deviation due to the different magnitude of the per capita income data with the population proportions. The method used to normalize was StandardScaler ().

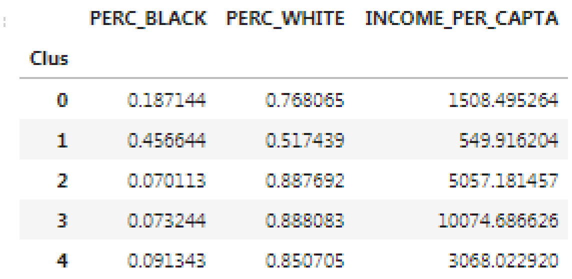
The centroid initialization method chosen was “k-means ++”, which intelligently selects the initial centers for the cluster to accelerate convergence. The number of defined clusters was 5. The number of times the algorithm was run with different centroid seeds was 12.

The model created the 5 clusters and classified based on these variables each of the census sectors in the city.



	CD_SETOR	LAT	LONG	PERC_BLACK	PERC_WHITE	INCOME_PER_CAPTA	Clus
0	355030802000076	-23.53822883899994	-46.71332882699994	0.049270	0.906934	5097.750000	2
1	355030801000001	-23.567837779999934	-46.570826168999986	0.176179	0.820099	1126.274194	0
2	355030801000002	-23.566559651999967	-46.56823083399996	0.196057	0.799562	927.190581	1
3	355030801000003	-23.56859691899996	-46.56752940999996	0.153600	0.833600	809.059200	1
4	355030801000004	-23.570316633999937	-46.56888325999995	0.234266	0.751748	779.739510	1

The following table shows the average of the groups



Clus	PERC_BLACK	PERC_WHITE	INCOME_PER_CAPTA
0	0.187144	0.768065	1508.495264
1	0.456644	0.517439	549.916204
2	0.070113	0.887692	5057.181457
3	0.073244	0.888083	10074.686626
4	0.091343	0.850705	3068.022920

The number of sectors per group was:

- 0: 3,504
- 1: 11,988
- 2: 901
- 3: 67
- 4: 1,903

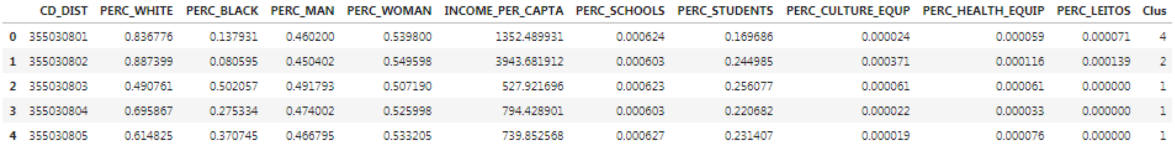
3.5.2. Districts

Based on the exploratory analysis of the data, the following were used to compose this model: the percentages of the white population, the percentages of the black population, the per capita income of the sector, the percentage of men and women, the number of schools per inhabitant, the number of students per inhabitant, the number of cultural facilities per inhabitant, the amount of health facilities per inhabitant and the number of hospital beds per district.

The data were normalized on the standard deviation using the StandardScaler () method.

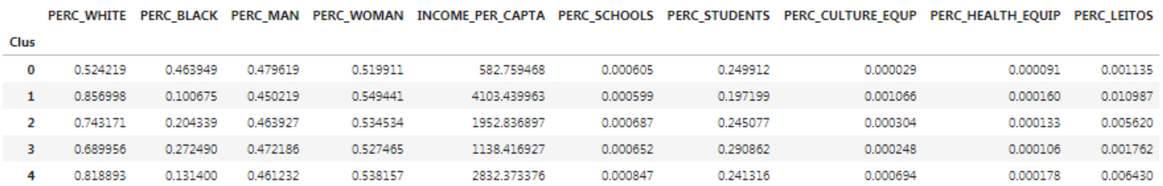
The centroid initialization method chosen was “k-means ++”, which intelligently selects the initial centers for the cluster to accelerate convergence. The number of defined clusters was 5. The number of times the algorithm was run with different centroid seeds was 12.

The model created the 5 clusters and classified based on these variables each of the city's districts.



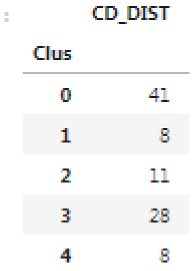
	CD_DIST	PERC_WHITE	PERC_BLACK	PERC_MAN	PERC_WOMAN	INCOME_PER_CAPTA	PERC_SCHOOLS	PERC_STUDENTS	PERC_CULTURE_EQUP	PERC_HEALTH_EQUIP	PERC_LEITOS	Clus
0	355030801	0.836776	0.137931	0.460200	0.539800	1352.489931	0.000624	0.169686	0.000024	0.000059	0.000071	4
1	355030802	0.887399	0.080595	0.450402	0.549598	3943.681912	0.000603	0.244985	0.000371	0.000116	0.000139	2
2	355030803	0.490761	0.502057	0.491793	0.507190	527.921696	0.000623	0.256077	0.000061	0.000061	0.000000	1
3	355030804	0.695867	0.275334	0.474002	0.525998	794.428901	0.000603	0.220682	0.000022	0.000033	0.000000	1
4	355030805	0.614825	0.370745	0.466795	0.533205	739.852568	0.000627	0.231407	0.000019	0.000076	0.000000	1

The following table shows the average of the groups



	PERC_WHITE	PERC_BLACK	PERC_MAN	PERC_WOMAN	INCOME_PER_CAPTA	PERC_SCHOOLS	PERC_STUDENTS	PERC_CULTURE_EQUP	PERC_HEALTH_EQUIP	PERC_LEITOS
Clus										
0	0.524219	0.463949	0.479619	0.519911	582.759468	0.000605	0.249912	0.000029	0.000091	0.001135
1	0.856998	0.100675	0.450219	0.549441	4103.439963	0.000599	0.197199	0.001066	0.000160	0.010987
2	0.743171	0.204339	0.463927	0.534534	1952.836897	0.000687	0.245077	0.000304	0.000133	0.005620
3	0.689956	0.272490	0.472186	0.527465	1138.416927	0.000652	0.290862	0.000248	0.000106	0.001762
4	0.818893	0.131400	0.461232	0.538157	2832.373376	0.000847	0.241316	0.000694	0.000178	0.006430

The number of districts per group was:



	CD_DIST
Clus	
0	41
1	8
2	11
3	28
4	8

3.6. Predictive model SVM.

SVM works by mapping data to a large-sized resource space, so that data points can be categorized, even when data is not linearly separable. A separator between the categories is found and the data is transformed so that the separator can be drawn as a hyperplane. After that, the characteristics of the new data can be used to predict the group to which a new record should belong.

3.6.1. SVM CENSUS sector.

Given the 5 groups created based on data from the city of São Paulo, this model aims to classify a sector in another city based on information on the percentage of whites, percentage of blacks and per capita income.

The data were normalized on the standard deviation due to the different magnitude of the per capita income data with the population proportions. The method used to normalize was StandardScaler (). Then they are divided into training and test data in the proportion of 80% for training and 20% for testing. The training sample had 14,690 sectors and the test sample had 3,673.

Then SVM functionality was imported from the sklearn library. The SVM algorithm offers a choice of kernel functions to perform its processing. Basically, mapping data in a higher dimensional space is called kernelling. The mathematical function used for the transformation is known as the kernel function and can be of different types. In this case, the linear function was chosen.

To evaluate this model, the F1 score was used, which considers the precision by the recall r of the test to calculate the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier and r is the number of correct positive results. divided by the number of all relevant samples. A Jaccard index was also used, which measures the similarity between sets of finite samples and is defined as the size of the intersection divided by the size of the union of the sample sets.

The F1 Score was equal to 1 and the Jaccard index is also the maximum score.

3.6.1. SVM District.

Given the 5 groups created based on data from the city of São Paulo, this model aims to classify a district in another city based on information on the percentages of white population, the percentages of black population, the per capita income of the sector, the percentage of men and women, the number of schools per capita, the number of students per capita, the amount of cultural facilities per capita, the amount of health facilities per capita and the number of hospital beds per district.

The data were normalized on the standard deviation due. The method used to normalize was StandardScaler (). Then they are divided into training and test data in the proportion of 80% for training and 20% for testing. The training sample had 76 sectors and the test sample had 20.

Then SVM functionality was imported from the sklearn library. The SVM algorithm offers a choice of kernel functions to perform its processing. Basically, mapping data in a higher dimensional space is called kernelling. The mathematical function used for the transformation is known as the kernel function and can be of different types. In this case, the linear function was chosen.

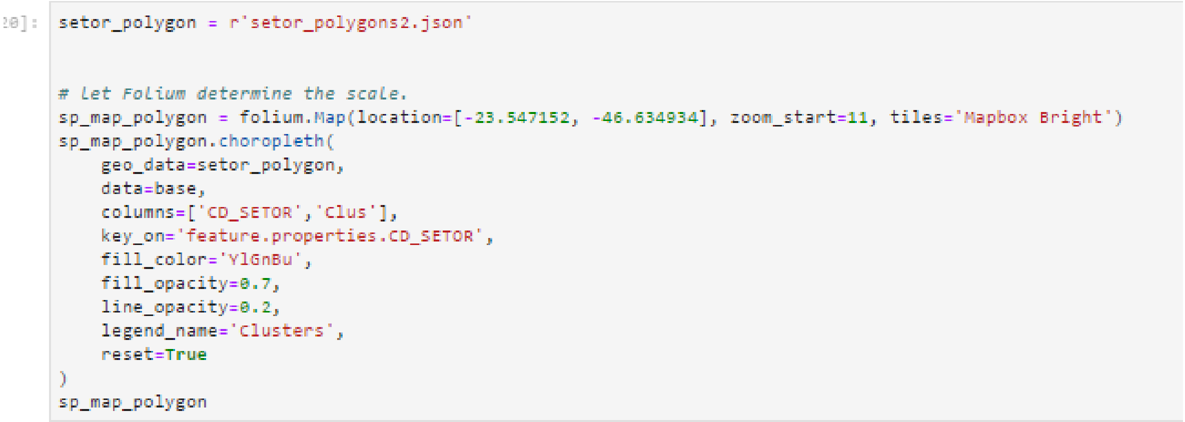
To evaluate this model, the F1 score was used, which considers the precision by the recall r of the test to calculate the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier and r is the number of correct positive results divided by the number of all relevant samples. A Jaccard index was also used, which measures the similarity between sets of finite samples and is defined as the size of the intersection divided by the size of the union of the sample sets.

The F1 Score was equal to 0.93 and the Jaccard index was 0.93. Indicating a good rating.

4. Results

With the sectors clustered and with the prediction model created, the first step is to place the information on the map to visualize the spatial distribution and understand if there is a spatial continuity of the clusters.

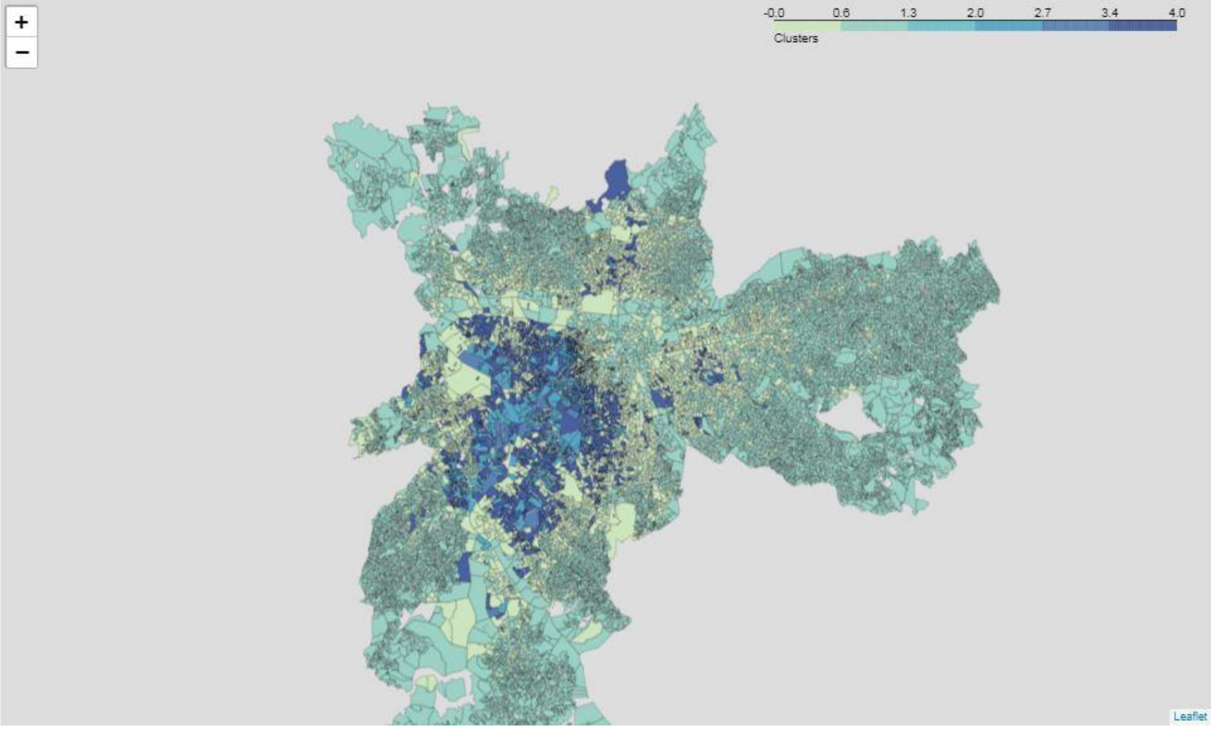
To do this was used the folium library and the following configuration:



```
20]: setor_polygon = r'setor_polygons2.json'


     # Let Folium determine the scale.
     sp_map_polygon = folium.Map(location=[-23.547152, -46.634934], zoom_start=11, tiles='Mapbox Bright')
     sp_map_polygon.choropleth(
         geo_data=setor_polygon,
         data=base,
         columns=['CD_SETOR','Clus'],
         key_on='feature.properties.CD_SETOR',
         fill_color='YlGnBu',
         fill_opacity=0.7,
         line_opacity=0.2,
         legend_name='Clusters',
         reset=True
     )
     sp_map_polygon
```

As a result we can see the map below:



Groups 2, 3 and 4 where the average percentage of the black population is below 10% (far below 37% of the total city), with an average income above R $ 3,068 are concentrated in the most central area of the city with minor exceptions of sectors.

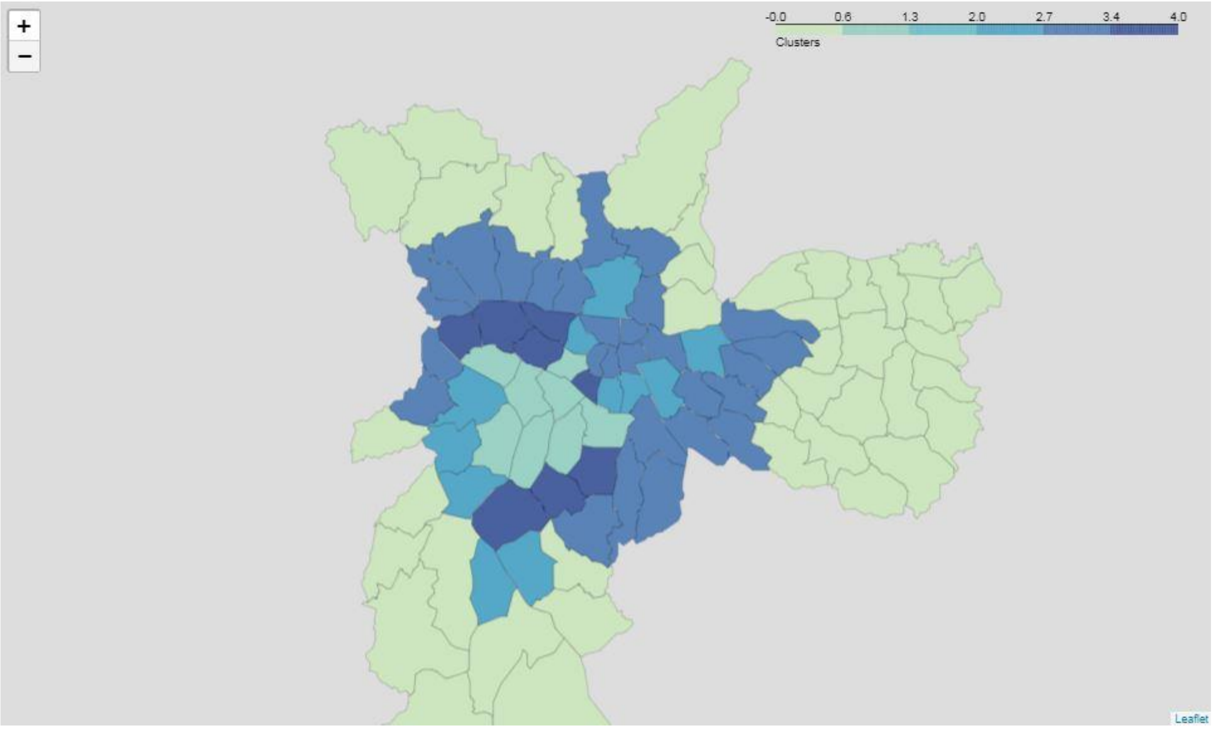
Group 0, which has an average of 18% blacks, has a lower average income around R $ 1,508 and is spatially on the edge of the central areas.

Group 1, which represents the majority of the city's sectors, has an average of 45.6% of the black population and has the lowest average per capita income in the city. These sectors are distributed in the peripheral parts of the city of São Paulo.

The SVM predictive model showed a high capacity to classify this same condition in other Brazilian cities to understand whether this spatial phenomenon is repeated.

The following model showed the classification of the districts using information other than the percentages of the white population, the percentages of the black population, the per capita income of the sector, the percentage of men and women, the number of schools per inhabitant, the number of students per inhabitant, the amount of cultural equipment per inhabitant, the amount of health equipment per inhabitant and the number of hospital beds per district.

As a result we can see the map below:



Groups 4 and 1 represent 16 districts of the city and are located in the most central part. The average white population is over 80% and black below 15% (in contrast to the representativeness of 37% of the black population in the city). These clusters have the best average per capita income and group 4 has the best average of schools per inhabitant and the best average of health equipment. Group 1, on the other hand, had the best average of cultural equipment and hospital beds. These groups together show that there is a concentration of structures in areas predominantly occupied by white people.

Groups 2 and 3 that are on the edges of groups 1 and 4. They function as a kind of geographical boundary between group 0 and the others. All averages have no specific prominence, they are in the middle both geographically and by population and structure. However, the average black population is still below the city, 20.4% for 2 and 27.2% for 3.

Finally, group 0 with the highest average percentage of blacks in the city and with the worst indicators of structure forming the peripheral areas of the city.

The SVM predictive model showed a high capacity to classify this same condition in other Brazilian cities to understand whether this spatial phenomenon is repeated.

5. Discussions

Income inequality in the city of São Paulo and access to urban equipment showed a strong correlation with color. Places with a higher percentage of whites have higher per capita income and better distribution of cultural, education and health structures.

To remedy this inequality, it is necessary to look at the problem in a social way. Brazilians often deny the existence of racism in our society, but it is no coincidence that poverty is concentrated in places where the majority is black.

It is necessary for governments to face this problem and adopt policies for the inclusion and distribution of income aimed at the black population. Creating a quota system at universities and encouraging companies to hire more black people through tax incentives may be an option.

Another suggestion is to bring more structures to the poorest parts of the city (where the black population is concentrated). Improving transportation, building cultural centers and schools helps to stimulate people in the search for knowledge and capacity. Increasing education increases access to income.

6. **Conclusion**

Racism is related to inequality in the distribution of Brazilian income and is reflected in the occupation of spaces in the city of São Paulo.

The data shown in this study of the correlation between the percentage of black residents, lower income, distance from central and more structured parts of the city, show that São Paulo needs inclusion policies that take into account the color / race factor.

It is necessary to consider that racism is a structural problem in Brazilian society and to initiate projects that ensure that the black population has more access to education and opportunities to improve their income. Only then can we reduce social inequality in Brazil.

References

[1]http://hdr.undp.org/sites/default/files/hdr2019.pdf

[2] https://assets.aspeninstitute.org/content/uploads/files/content/docs/rcc/RCC-Structural-Racism-Glossary.pdf

[3]https://www.bbc.com/portuguese/brasil-44034767