Improving the quality of life and development in a large city
- Ishan Kumar

1.INTRODUCTION

1.1.Problem

São Paulo is the most populous city in the american continent and the richest city in Brazil, but suffers from the inequality of development between the central region and the periphery.

The economic and social development of the peripheries can improve the lives of all, including the inhabitants of the central region.

São Paulo is the capital of the state of São Paulo (r.1)

São Paulo is a state of Brazil (r.1)

1.1.Audience

This report is intended for those responsible for urban development in the city of São Paulo

1.2. Purpose

The city of São Paulo is divided into districts (or neighborhoods). Let's compare the central and peripheral districts with regard to nearby venues, the human development index, average monthly salaries and population. Which districts deserve greater attention from the authorities?

2.DATA DESCRIPTION

To compare São Paulo's central region and peripheral region, we will study where the largest populations, the best salaries, the stores and the services are concentrated.

Let us scrap some Wikipedia pages "List of districts of São Paulo by population"(r2)

[1]:

	Position	District	Population2010	Population2000
0	1	Grajaú	444593	333436
1	2	Sapopemba	296042	282239
2	3	Jardim Ângela	291798	245805
3	4	Brasilândia	280069	247328
4	5	Capão Redondo	275230	240793

And scrap "List of districts of São Paulo by Human Development Index" (r3)

	District	Human Development Index	Classification
0	Moema	0.981	very high
1	Pinheiros	0.980	very high
2	Perdizes	0.977	very high
3	Jardim Paulista	0.975	very high
4	Alto de Pinheiros	0.972	very high

"Human Development Index (HDI) was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. The HDI can also be used to question national policy choices" (r4). And it is based on long and healthy life, knowledge and descent standard of living.

Let us merge these dataframes and look for districts' coordinates in Google Maps (r5):

[4]:

	Unnamed: 0	Position	District	Population2010	Population2000	Human Development Index	Classification	Latitude	Longitude
0	0	1	Grajaú	444593	333436	0.754	medium	-23.771911	-46.669070
1	1	2	Sapopemba	296042	282239	0.786	medium	-23.573739	-46.524185
2	2	3	Jardim Ângela	291798	245805	0.750	medium	-23.700988	-46.769102
3	3	4	Brasilândia	280069	247328	0.769	medium	-23.470400	-46.689832
4	4	5	Capão Redondo	275230	240793	0.782	medium	-23.668870	-46.769991

We will explore districts venues with Foursquare API (r6):

[16]:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Sapopemba	-23.573739	-46.524185	Pizzaria Família K	-23.571293	-46.524938	Pizza Place
1	Sapopemba	-23.573739	-46.524185	Nova Colonial Pizzaria e Choperia	-23.569667	-46.524516	Pizza Place
2	Sapopemba	-23.573739	-46.524185	Padaria Monsenhor	-23.571328	-46.527176	Bakery
3	Sapopemba	-23.573739	-46.524185	Subway	-23.571205	-46.523084	Sandwich Place
4	Sapopemba	-23.573739	-46.524185	Hortifrut Jd. Vila Formosa	-23.571360	-46.525167	Grocery Store

Let us scrap wages by district (r7), convert the values to dollars and merge to other dataframes:

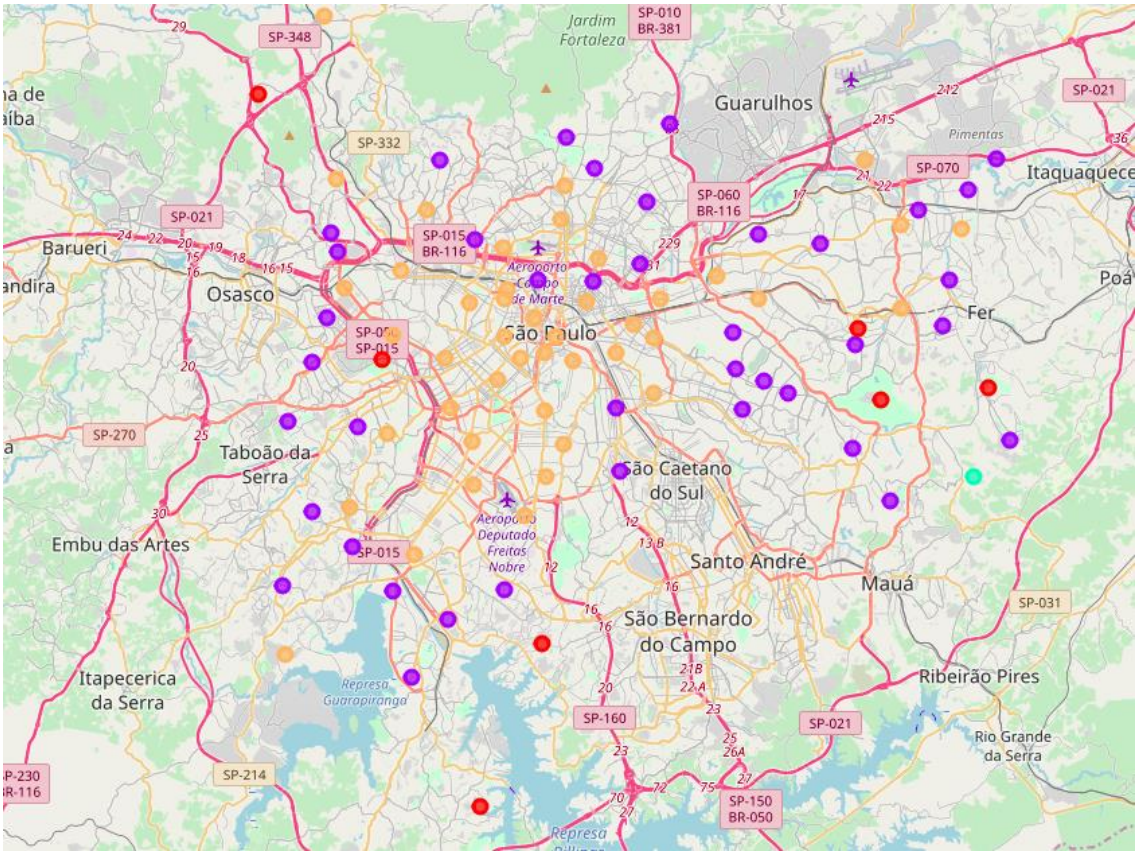
[27]:

	Unnamed: 0	Position	District	Population2010	Population2000	Human Development Index	Classification	Latitude	Longitude	Average Monthly Wage R$	Average Monthly Wage U$
0	0	1	Grajaú	444593	333436	0.754	medium	-23.771911	-46.669070	1852.28	561.296970
1	1	2	Sapopemba	296042	282239	0.786	medium	-23.573739	-46.524185	2500.50	757.727273
2	2	3	Jardim Ângela	291798	245805	0.750	medium	-23.700988	-46.769102	1889.36	572.533333
3	3	4	Brasilândia	280069	247328	0.769	medium	-23.470400	-46.689832	1680.36	509.200000
4	4	5	Capão Redondo	275230	240793	0.782	medium	-23.668870	-46.769991	2018.27	611.596970

3.METHODOLOGY

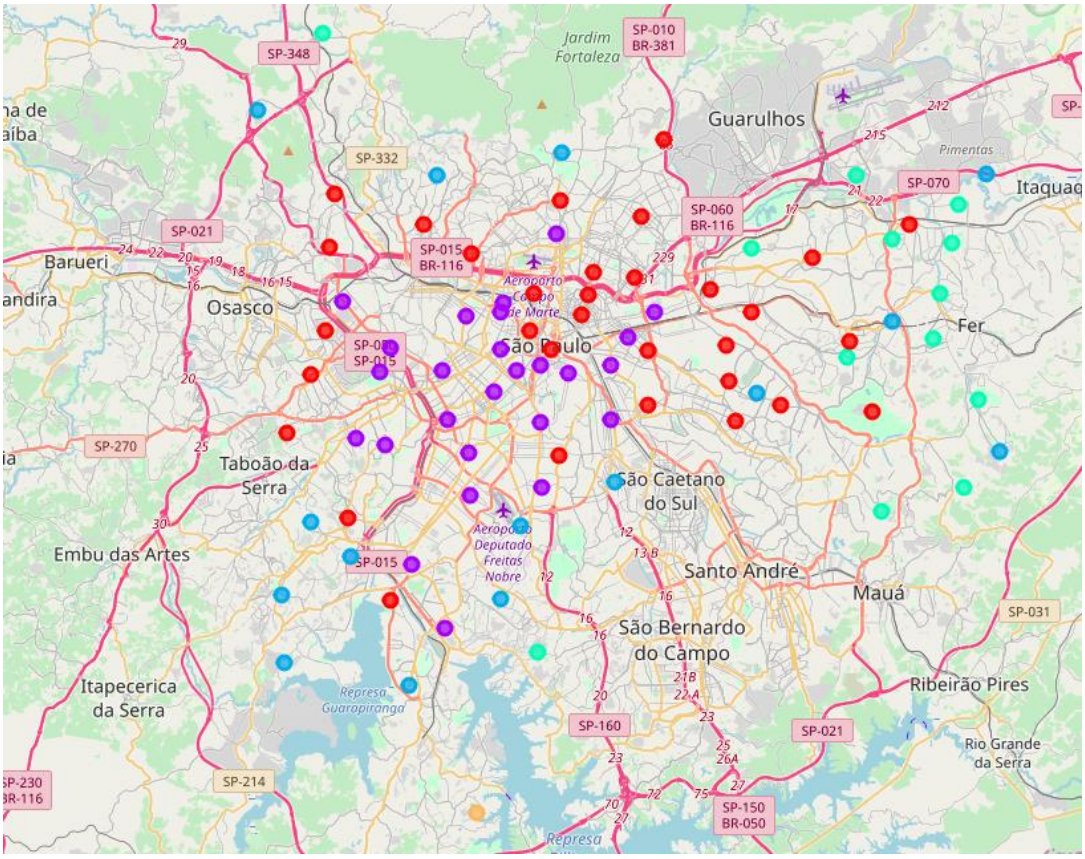
As in week 3 laboratory, we have grouped districts considering nearby venues (from Foursquare). We have used K-Means algorithm in Scikit-learn because we want to compare central districts and peripheral districts.

Here we have the Folium map with the clusters:



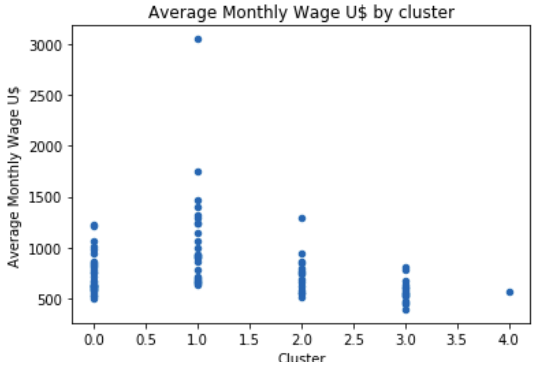
We can see the yellow group in the center of the map and the purple group more distant. And other groups (red and green) even more distant.

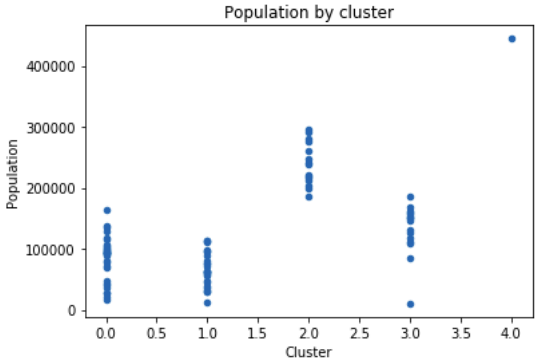
We have tried clustering considering other districts' attributes like population, HDI, Average Monthly Wage:



We can see a red cluster in the center-north of the city, and we can see a purple cluster in the center-south. And we have blue, yellow and green clusters more distant from the center.

And we plot some scatter graphs with these groups from the second clustering:





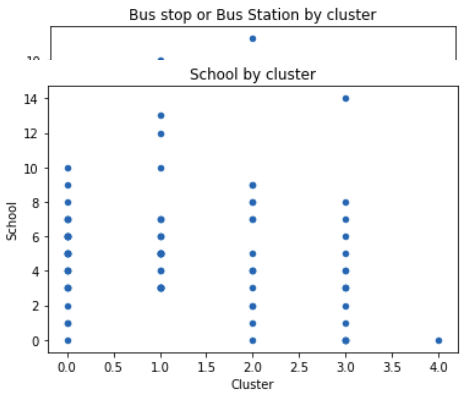
To consider venues and HDI, and wage and population, we have selected some venues categories and we have grouped and renamed the categories returned by Foursquare into a set of new “categories”:

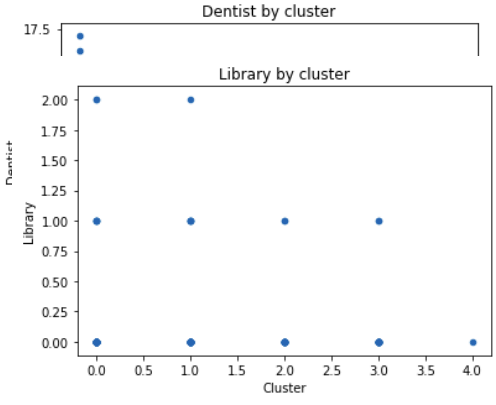
```
array(['Bus', "Dentist's Office", 'School', 'Entertainment', 'Sports',
       'Health', 'Library', 'Daycare', 'Metro Station', 'Train Station',
       'Bookstore'], dtype=object)
```

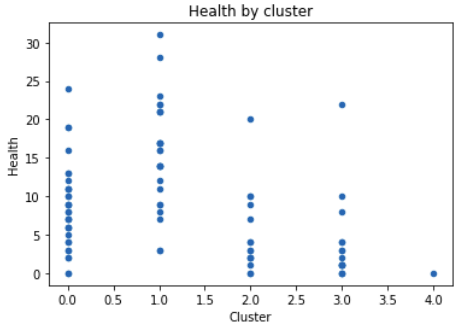
[33]:

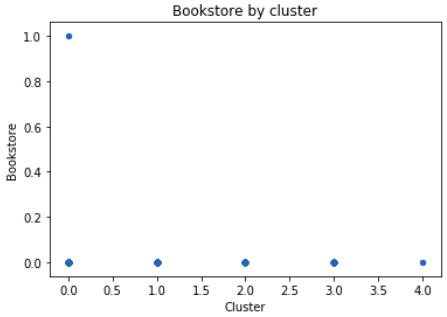
	Cluster Labels	District	Population2010	Human Development Index	Classification	Latitude	Longitude	Average Monthly Wage R$	Average Monthly Wage U$	Bus	Dentist	School	Entertainment	Health	Sports	Library	Daycare	Metro Station	Train Station	Bookstore	Total
0	4	Grajaú	444593	0.754	medium	-23.771911	-46.669070	1852.28	561.296970	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
1	2	Sapopemba	296042	0.786	medium	-23.573739	-46.524185	2500.50	757.727273	3.0	2.0	4.0	1.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	15.0
2	2	Jardim Ângela	291798	0.750	medium	-23.700988	-46.769102	1889.36	572.533333	0.0	1.0	2.0	0.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	8.0
3	2	Brasilândia	280069	0.769	medium	-23.470400	-46.689832	1680.36	509.200000	2.0	10.0	3.0	1.0	3.0	5.0	1.0	0.0	0.0	0.0	0.0	25.0
4	2	Capão Redondo	275230	0.782	medium	-23.668870	-46.769991	2018.27	611.596970	1.0	5.0	8.0	1.0	4.0	7.0	0.0	0.0	0.0	0.0	0.0	26.0

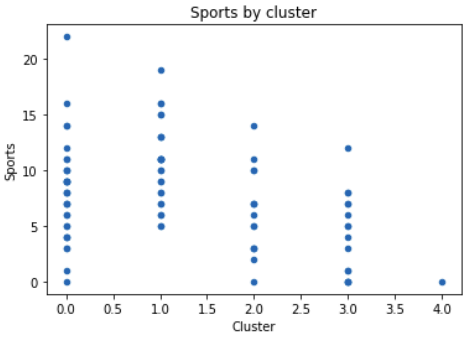
We have compared the existing clusters with the quantity of each venue.

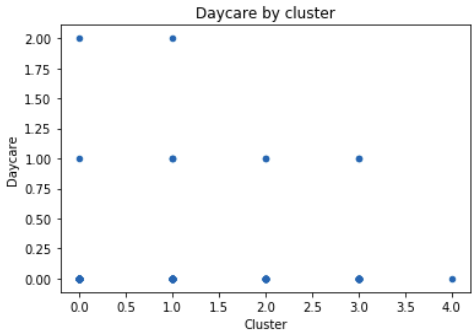


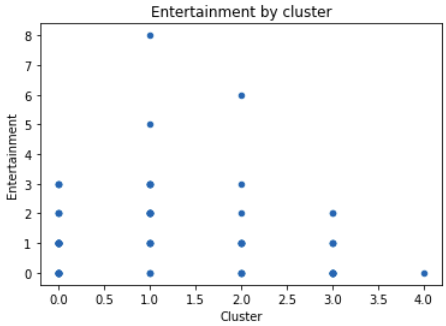


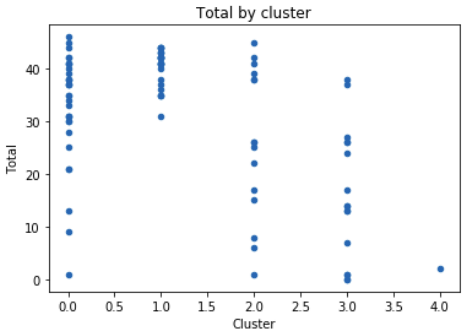






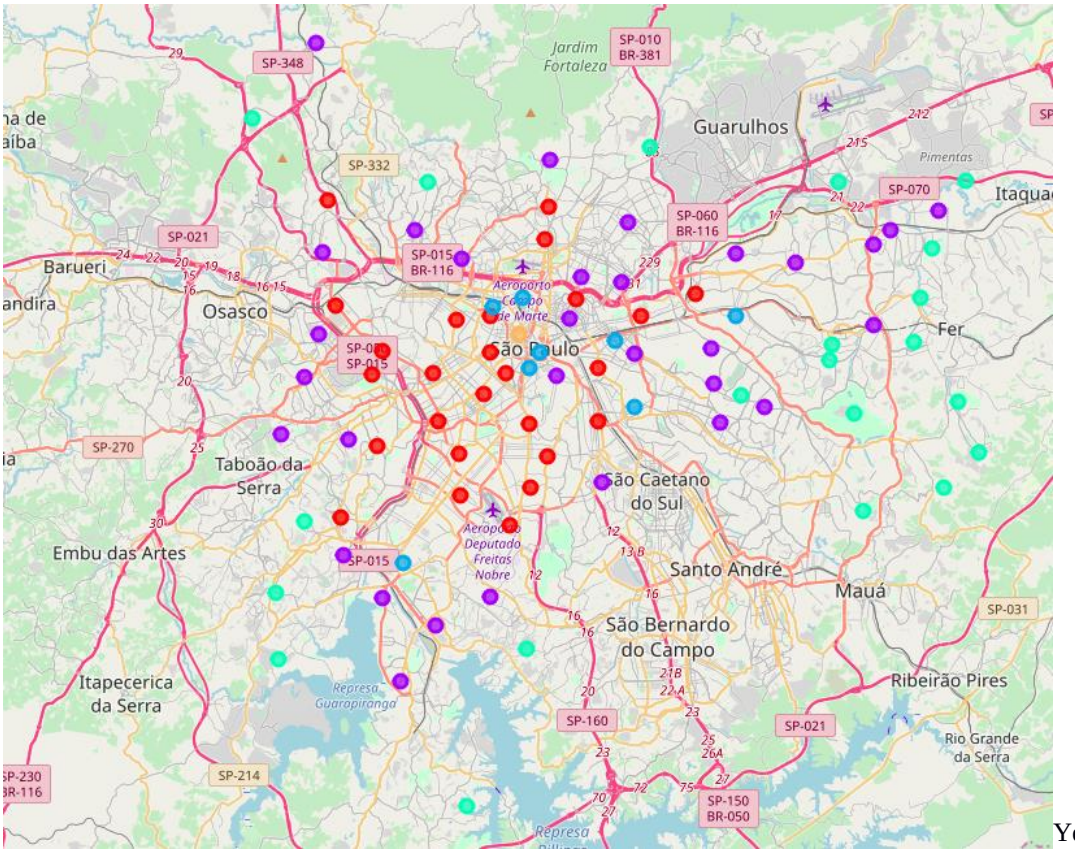






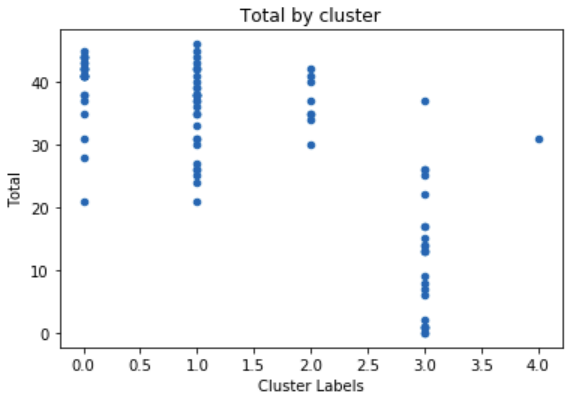
We can see the cluster 4 is very poor. It consists of one district, Grajaú.
And we see the whole city needs more libraries, bookstores and daycares.
Clusters 0, 1, 2 and 3 have districts with different numbers of venues.

And we have applied k-means again considering HDI, wages, population and venues:

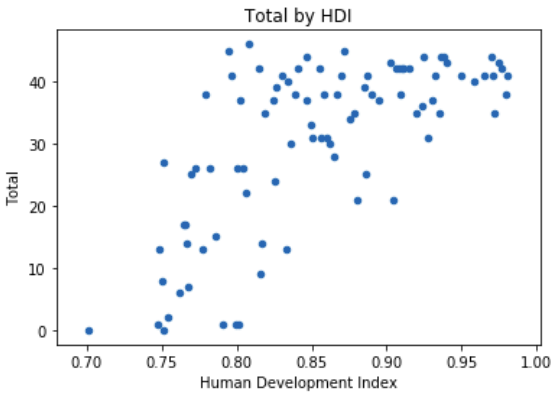


Yellow group is in the center the city (district República) and really has a nice infrastructure. Green group 3 really covers some poor districts.

Let us plot some charts (considering the new clusters). Here we have total venues by cluster:



And here we compare total of venues by human development index:



Cluster 3:

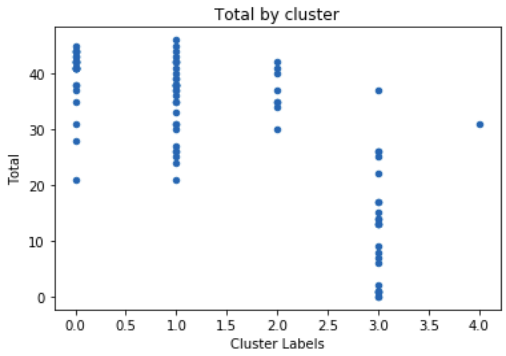
Launcher | Coursera-Capstone-week-5 | Code | Python 3

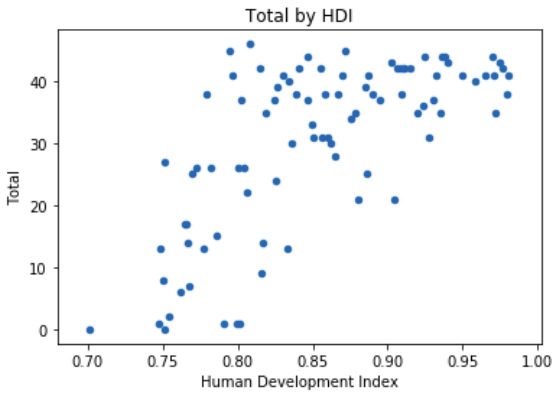
[50]:

ster bels 2	Cluster Labels	District	Population2010	Human Development Index	Classification	Latitude	Longitude	Average Monthly Wage R$	Average Monthly Wage U$	Bus	Dentist	School	Entertaiment	Health	Sports	Library	Daycare	Metro Station	Train Station	Bookstore	Total
3	4	Grajaú	444593	0.754	medium	-23.771911	-46.669070	1852.28	561.296970	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
3	2	Sapopemba	296042	0.786	medium	-23.573739	-46.524185	2500.50	757.727273	3.0	2.0	4.0	1.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	15.0
3	2	Jardim Ângela	291798	0.750	medium	-23.700988	-46.769102	1889.36	572.533333	0.0	1.0	2.0	0.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	8.0
3	2	Brasilândia	280069	0.769	medium	-23.470400	-46.689832	1680.36	509.200000	2.0	10.0	3.0	1.0	3.0	5.0	1.0	0.0	0.0	0.0	0.0	25.0
3	2	Capão Redondo	275230	0.782	medium	-23.668870	-46.769991	2018.27	611.596970	1.0	5.0	8.0	1.0	4.0	7.0	0.0	0.0	0.0	0.0	0.0	26.0
3	2	Itaim Paulista	241026	0.762	medium	-23.469455	-46.405854	2520.70	763.848485	2.0	1.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	6.0
3	2	Jaraguá	220292	0.791	medium	-23.439354	-46.782427	2783.40	843.454545	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	2	Cidade Tiradentes	219868	0.766	medium	-23.601429	-46.398844	4252.35	1288.590909	5.0	3.0	2.0	0.0	3.0	3.0	0.0	1.0	0.0	0.0	0.0	17.0
3	2	Campo Limpo	216098	0.806	high	-23.634547	-46.754958	1801.48	545.903030	0.0	3.0	4.0	3.0	2.0	10.0	0.0	0.0	0.0	0.0	0.0	22.0
3	3	Lajeado	185184	0.748	medium	-23.526313	-46.429817	1470.31	445.548485	2.0	3.0	4.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	13.0
3	3	Vila Curuçá	162486	0.765	medium	-23.502311	-46.423077	2562.64	776.557576	1.0	0.0	3.0	1.0	4.0	8.0	0.0	0.0	0.0	0.0	0.0	17.0
3	3	Pedreira	158656	0.777	medium	-23.696262	-46.637714	2672.96	809.987879	3.0	0.0	4.0	1.0	1.0	4.0	0.0	0.0	0.0	0.0	0.0	13.0
3	3	Cachoeirinha	157408	0.802	high	-29.946917	-51.103680	1892.59	573.512121	0.0	4.0	3.0	0.0	22.0	7.0	1.0	0.0	0.0	0.0	0.0	37.0
3	3	São Rafael	151017	0.767	medium	-23.629185	-46.459661	1750.59	530.481818	2.0	5.0	2.0	0.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	14.0
3	3	Parelheiros	146212	0.747	medium	-23.820943	-46.704386	1734.37	525.566667	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
3	3	Cidade Líder	130255	0.817	high	-23.556407	-46.477807	1799.51	545.306061	0.0	2.0	5.0	0.0	1.0	5.0	0.0	1.0	0.0	0.0	0.0	14.0
3	3	Iguatemi	126645	0.751	medium	-23.618039	-46.417409	2002.58	606.842424	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	3	Ermelino Matarazzo	116632	0.801	high	-23.470018	-46.473195	1831.38	554.963636	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	3	Guaianases	111325	0.768	medium	-23.576556	-46.409522	1502.85	455.409091	1.0	0.0	3.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	7.0
3	3	José Bonifácio	108366	0.804	high	-23.547519	-46.433093	1984.27	601.293939	1.0	6.0	1.0	2.0	8.0	8.0	0.0	0.0	0.0	0.0	0.0	26.0
3	0	Artur Alvim	104864	0.833	high	-23.548794	-46.476466	1654.12	501.248485	2.0	1.0	4.0	0.0	0.0	5.0	0.0	0.0	1.0	0.0	0.0	13.0
3	0	Jaçanã	92836	0.816	high	-23.453125	-46.572076	2043.27	619.172727	2.0	3.0	1.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	9.0
3	0	Parque do Carmo	69630	0.799	medium	-23.582371	-46.464393	1913.98	579.993939	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	3	Marsilac	10180	0.701	medium	-23.894952	-46.707810	1287.32	390.096970	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5.RESULTS & DISCUSSION

The last image shows some districts we should pay attention to. The following chart show that clustering 3 covers districts with a low number of selected venues (related to transport, health, education).





And the following chart shows that human development index accompanies the total of the selected venues:

We have limited the number of venues and distance when applying the Foursquare api, and there are libraries in schools and universities, even só, it is surprising the low amount of libraries and bookstores available throughout the city.

6.CONCLUSION

To improve human development index, people needs nice health, transport, education, entertainment. Some districts of São Paulo need improvement in these areas, mainly some districts in the extreme South and many districts of East.

The whole city needs more libraries, bookstores and daycares.

7. REFERENCES

r1.Wikipedia - São Paulo - https://pt.wikipedia.org/wiki/S%C3%A3o_Paulo
r2.Wikipedia - Lista dos distritos de São Paulo por população - https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_popula%C3%A7%C3%A3º
r3.Wikipedia - Lista dos distritos de São Paulo por Índice de Desenvolvimento Humano - https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_%C3%8Dndice_de_Desenvolvimento_Humano
r4.Human Development Index (HDI) - http://hdr.undp.org/en/content/human-development-index-hdi
r5.Google Maps
r6.Foursquare
r7.Média de salário em SP vai de R$ 1,2 mil em Marsilac a R$ 10 mil no Campo Belo - https://g1.globo.com/sao-paulo/noticia/media-de-salario-em-sp-vai-de-r-12-mil-em-marsilac-a-r-10-mil-no-campo-belo.ghtml