Open a new restastaurant in São Paulo city

Rodrigo Mendrico

January, 2021

Introduction

The idea is to help a new investment company to discover where is the best place and the best type of restaurant to open in the city of Sao Paulo,Brazil.

I will use the foursquare to check how other restaurants are graded and how they are distibuted among the Sao Paulo city´s neighborhood among with other data from city´s oficial site. I will will check for opportunities identifying the type of restaurants and where are the best neighborhoods to apply.

Data acquisition and cleaning

2.1 Foursquare Data

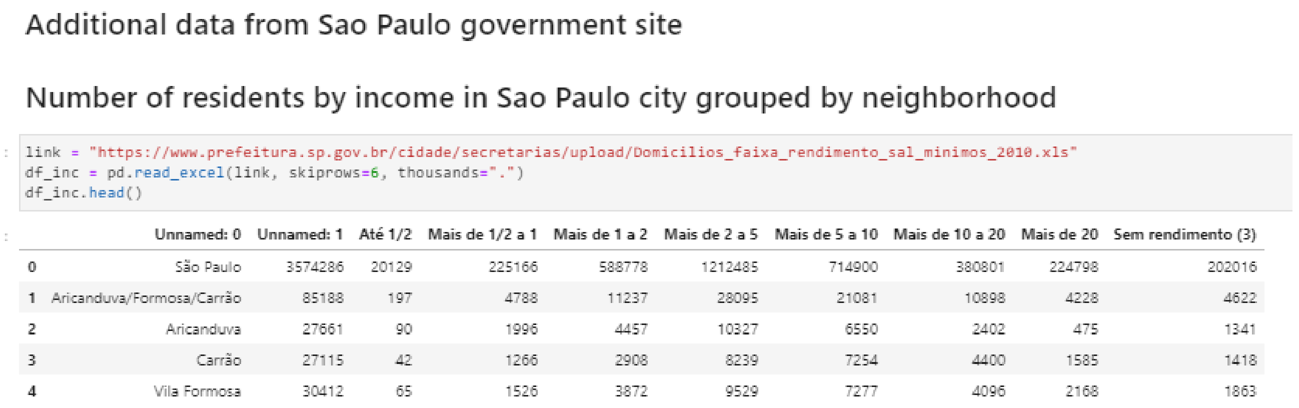
I will use foursquare api data. Foursquare is a social media website that collects information about places around the world. The documentation how to use this api is available at https://developer.foursquare.com/docs/places-api/ To use you will need to create an account on this website. Some api calls are available for free and others you need to acquire the premium category. This api will be use to explore data about Venues in the city of Sao Paulo.

2.2 Geopy

Geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources. I will Use geopy library to get the latitude and longitude values for Sao Paulo city. Geopy information is available at https://geopy.readthedocs.io/en/stable/

2.3 How I will use the data

I will use the foursquare to check how other restaurants are graded and how they are distibuted among the Sao Paulo city´s neighborhood. This will give a clue for the best type of restaurant/cousine to open checking the restaurants with minor grades and telling in which neighbors this type of cousine is not available yet. Therefore, I will get some other statistics data to include in the dataset from Sao Paulo City official site https://www.prefeitura.sp.gov.br/. I will grab the educational data by neighborhood (Level of scholarity) and financial data (#people with best income rate) to decide which would be the best neighbor to open the new restaurant. Another datasource that migh be used is the about number of new houses/apartments build by neighborhood and include on the dataset in order to get a more accurate model because when will have more new residents we have more customers. I will use clustering and cloropleth in order to visualize and base the study.

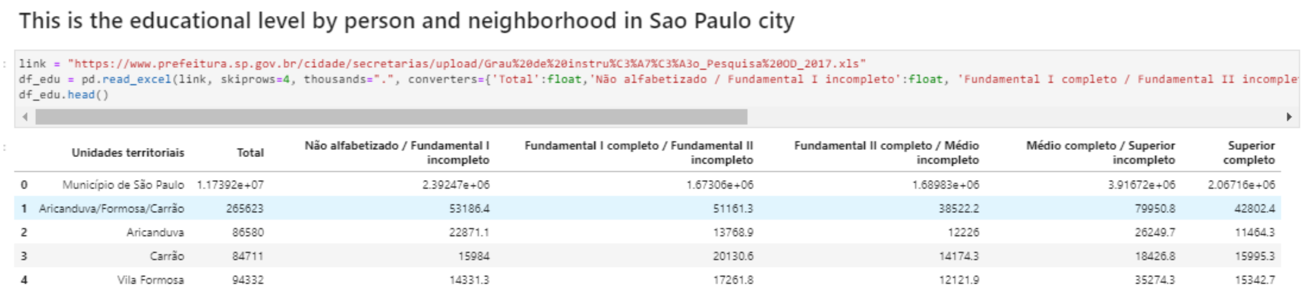


Additional data from Sao Paulo government site

Number of residents by income in Sao Paulo city grouped by neighborhood

```
link = "https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/Domicilios_faixa_rendimento_sal_minimos_2010.xls"
df_inc = pd.read_excel(link, skiprows=6, thousands=".")
df_inc.head()
```

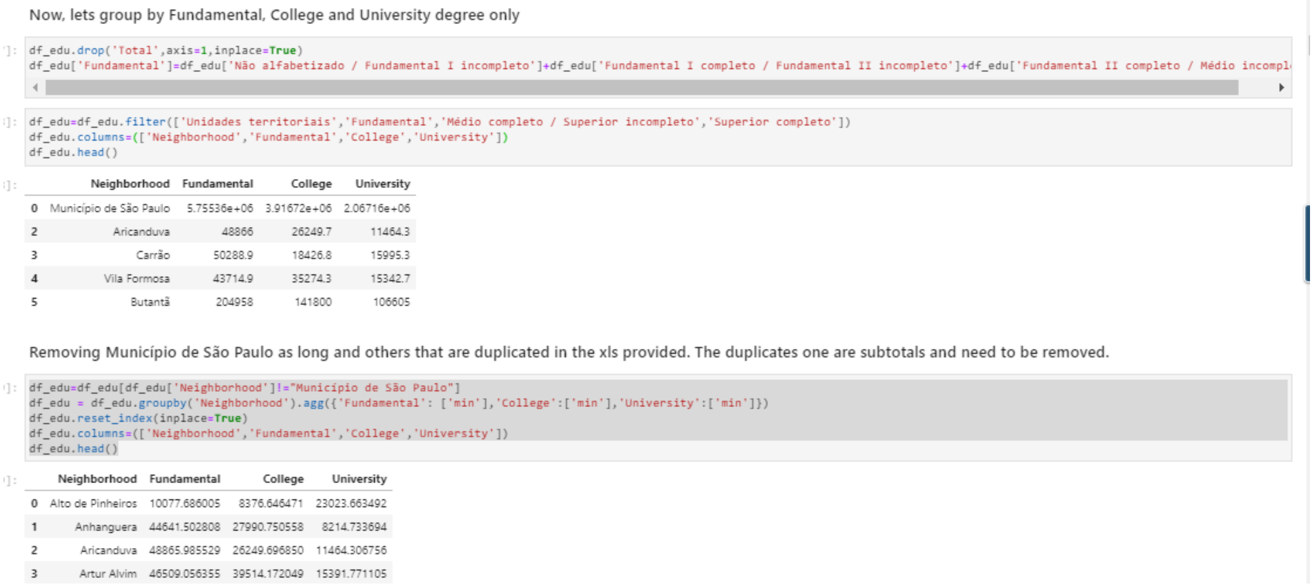
	Unnamed: 0	Unnamed: 1	Até 1/2	Mais de 1/2 a 1	Mais de 1 a 2	Mais de 2 a 5	Mais de 5 a 10	Mais de 10 a 20	Mais de 20	Sem rendimento (3)
0	São Paulo	3574286	20129	225166	588778	1212485	714900	380801	224798	202016
1	Aricanduva/Formosa/Carrão	85188	197	4788	11237	28095	21081	10898	4228	4622
2	Aricanduva	27661	90	1996	4457	10327	6550	2402	475	1341
3	Carrão	27115	42	1266	2908	8239	7254	4400	1585	1418
4	Vila Formosa	30412	65	1526	3872	9529	7277	4096	2168	1863



This is the educational level by person and neighborhood in Sao Paulo city

```
link = "https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/Grau%20de%20instru%C3%A7%C3%A3o_Pesquisa%20OD_2017.xls"
df_edu = pd.read_excel(link, skiprows=4, thousands=".", converters={'Total':float,'Não alfabetizado / Fundamental I incompleto':float, 'Fundamental I completo / Fundamental II incomple
df_edu.head()
```

	Unidades territoriais	Total	Não alfabetizado / Fundamental I incompleto	Fundamental I completo / Fundamental II incompleto	Fundamental II completo / Médio incompleto	Médio completo / Superior incompleto	Superior completo
0	Município de São Paulo	1.17392e+07	2.39247e+06	1.67306e+06	1.68983e+06	3.91672e+06	2.06716e+06
1	Aricanduva/Formosa/Carrão	265623	53186.4	51161.3	38522.2	79950.8	42802.4
2	Aricanduva	86580	22871.1	13768.9	12226	26249.7	11464.3
3	Carrão	84711	15984	20130.6	14174.3	18426.8	15995.3
4	Vila Formosa	94332	14331.3	17261.8	12121.9	35274.3	15342.7



Now, lets group by Fundamental, College and University degree only

```
df_edu.drop('Total',axis=1,inplace=True)
df_edu['Fundamental']=df_edu['Não alfabetizado / Fundamental I incompleto']+df_edu['Fundamental I completo / Fundamental II incompleto']+df_edu['Fundamental II completo / Médio incompl
```

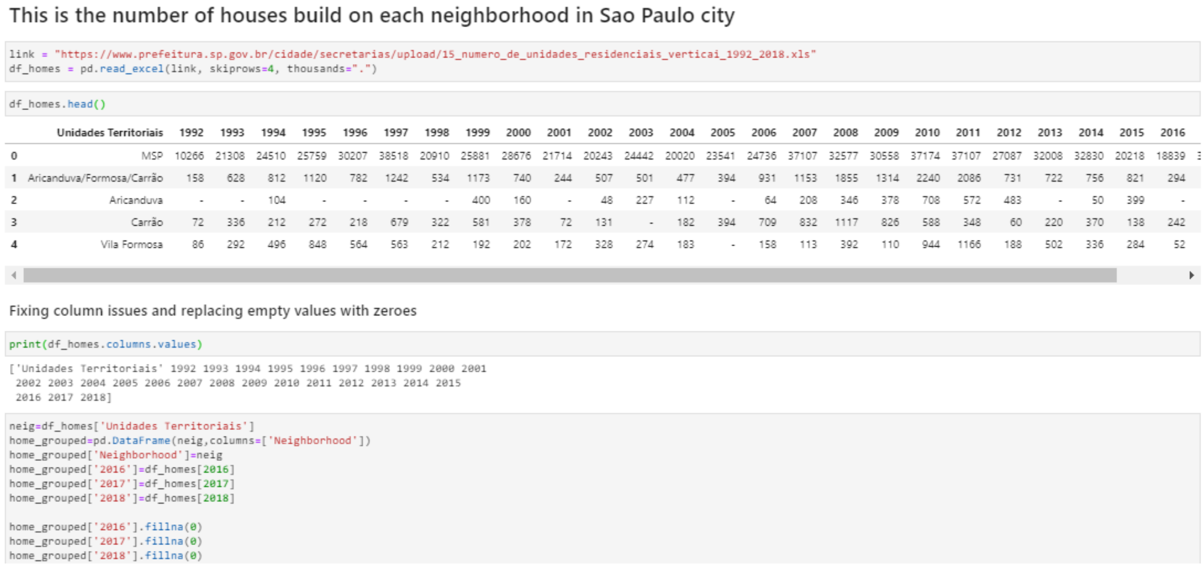
```
df_edu=df_edu.filter(['Unidades territoriais','Fundamental','Médio completo / Superior incompleto','Superior completo'])
df_edu.columns=(['Neighborhood','Fundamental','College','University'])
df_edu.head()
```

	Neighborhood	Fundamental	College	University
0	Município de São Paulo	5.75536e+06	3.91672e+06	2.06716e+06
2	Aricanduva	48866	26249.7	11464.3
3	Carrão	50288.9	18426.8	15995.3
4	Vila Formosa	43714.9	35274.3	15342.7
5	Butantã	204958	141800	106605

Removing Município de São Paulo as long and others that are duplicated in the xls provided. The duplicates one are subtotals and need to be removed.

```
df_edu=df_edu[df_edu['Neighborhood']!="Municipio de São Paulo"]
df_edu = df_edu.groupby('Neighborhood').agg({'Fundamental': ['min'],'College':['min'],'University':['min']})
df_edu.reset_index(inplace=True)
df_edu.columns=(['Neighborhood','Fundamental','College','University'])
df_edu.head()
```

	Neighborhood	Fundamental	College	University
0	Alto de Pinheiros	10077.686005	8376.646471	23023.663492
1	Anhanguera	44641.502808	27990.750558	8214.733694
2	Aricanduva	48865.985529	26249.696850	11464.306756
3	Artur Alvim	46509.056355	39514.172049	15391.771105



This is the number of houses build on each neighborhood in Sao Paulo city

```
link = "https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/15_numero_de_unidades_residenciais_verticai_1992_2018.xls"
df_homes = pd.read_excel(link, skiprows=4, thousands=".")
```

```
df_homes.head()
```

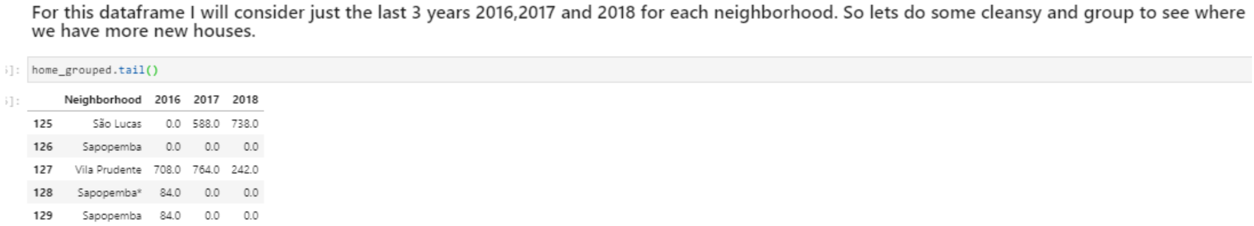
	Unidades Territoriais	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
0	MSP	10266	21308	24510	25759	30207	38518	20910	25881	28676	21714	20243	24442	20020	23541	24736	37107	32577	30558	37174	37107	27087	32008	32830	20218	18839
1	Aricanduva/Formosa/Carrão	158	628	812	1120	782	1242	534	1173	740	244	507	501	477	394	931	1153	1855	1314	2240	2086	731	722	756	821	294
2	Aricanduva	-	-	104	-	-	-	-	400	160	-	48	227	112	-	64	208	346	378	708	572	483	-	50	399	-
3	Carrão	72	336	212	272	218	679	322	581	378	72	131	-	182	394	709	832	1117	826	588	348	60	220	370	138	242
4	Vila Formosa	86	292	496	848	564	563	212	192	202	172	328	274	183	-	158	113	392	110	944	1166	188	502	336	284	52

Fixing column issues and replacing empty values with zeroes

```
print(df_homes.columns.values)
```

```
['Unidades Territoriais' 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001
 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
 2016 2017 2018]
```

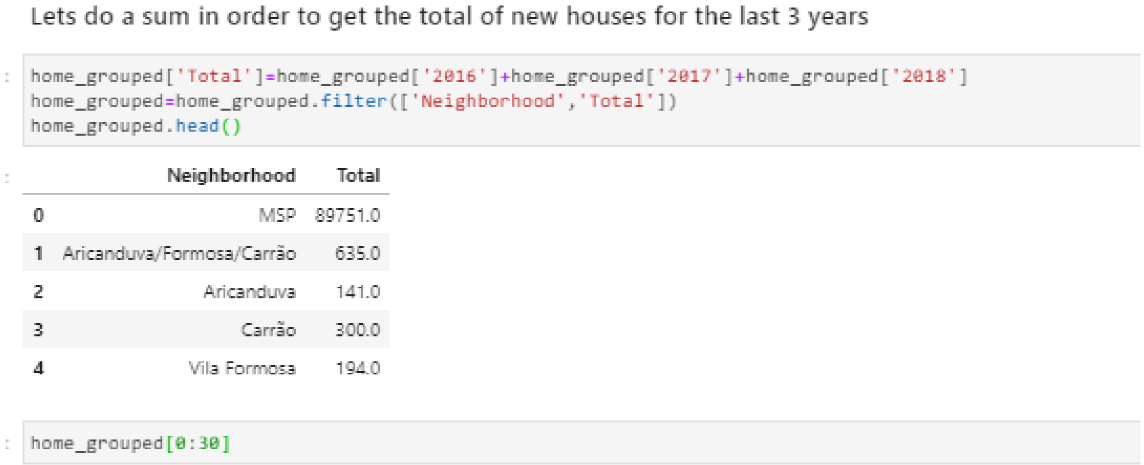
```
neig=df_homes['Unidades Territoriais']
home_grouped=pd.DataFrame(neig,columns=['Neighborhood'])
home_grouped['Neighborhood']=neig
home_grouped['2016']=df_homes[2016]
home_grouped['2017']=df_homes[2017]
home_grouped['2018']=df_homes[2018]

home_grouped['2016'].fillna(0)
home_grouped['2017'].fillna(0)
home_grouped['2018'].fillna(0)
```



For this dataframe I will consider just the last 3 years 2016,2017 and 2018 for each neighborhood. So lets do some cleansy and group to see where we have more new houses.

```
home_grouped.tail()
```

	Neighborhood	2016	2017	2018
125	São Lucas	0.0	588.0	738.0
126	Sapopemba	0.0	0.0	0.0
127	Vila Prudente	708.0	764.0	242.0
128	Sapopemba*	84.0	0.0	0.0
129	Sapopemba	84.0	0.0	0.0



Lets do a sum in order to get the total of new houses for the last 3 years

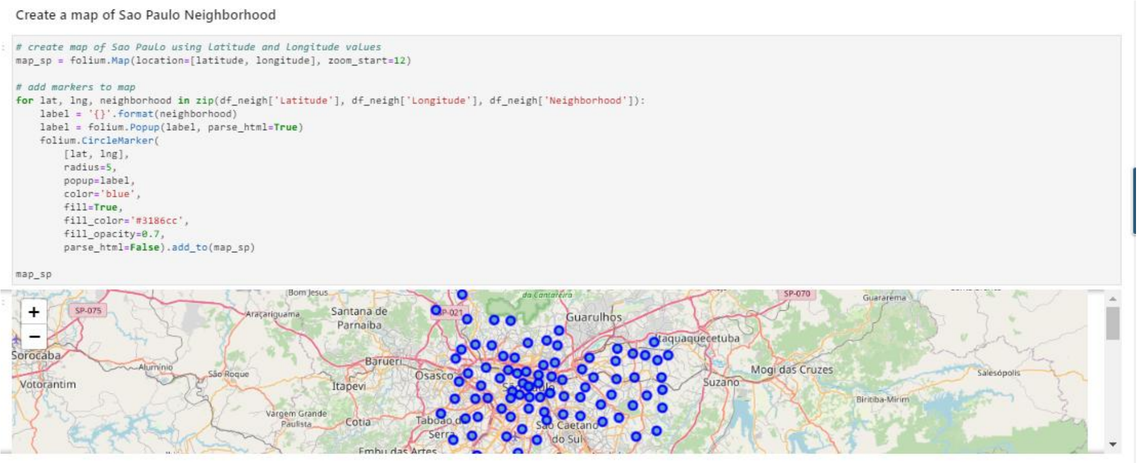
```
home_grouped['Total']=home_grouped['2016']+home_grouped['2017']+home_grouped['2018']
home_grouped=home_grouped.filter(['Neighborhood','Total'])
home_grouped.head()
```

	Neighborhood	Total
0	MSP	89751.0
1	Aricanduva/Formosa/Carrão	635.0
2	Aricanduva	141.0
3	Carrão	300.0
4	Vila Formosa	194.0

```
home_grouped[0:30]
```

Exploratory Data Analysis

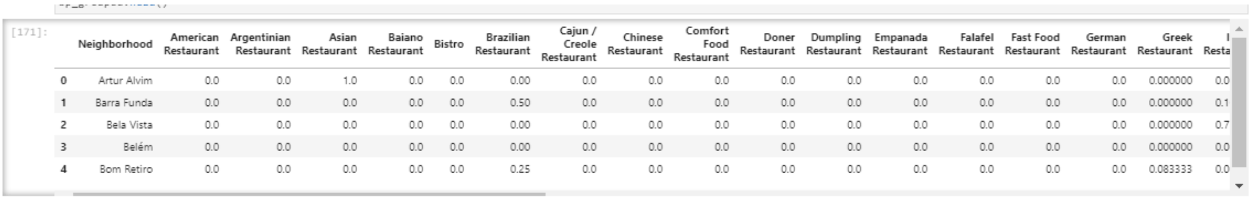
After do some data cleansy, we will treat the Income dataframe and Neighborhoods in order to get the geografical coordinates using Geopy. We also will uses Foursquare to get restaurants locations. We will looking mainly for the type of the restaurants and group them. After get the foursquare data, will be possible to make a cluster and plot a map about the categories we have found. Later we will work with education and new houses data in order to do some clustering also and compare to understand which neighborhoods has the best education and income rate. These will probably will be our preffered locations.



Create a map of Sao Paulo Neighborhood

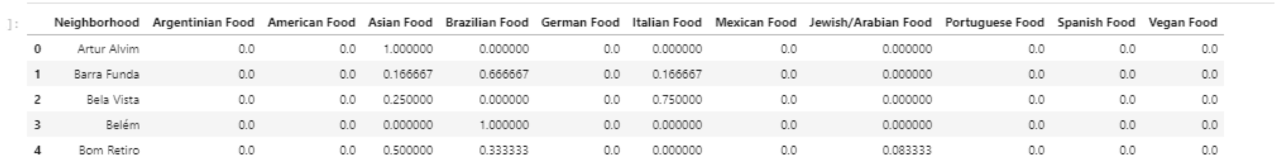
```
# create map of Sao Paulo using latitude and longitude values
map_sp = folium.Map(location=[latitude, longitude], zoom_start=12)

# add markers to map
for lat, lng, neighborhood in zip(df_neigh['Latitude'], df_neigh['Longitude'], df_neigh['Neighborhood']):
    label = '{}'.format(neighborhood)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_sp)

map_sp
```

Using Foursquare it was possible to identify the types of restaurants in each neighborhood and than group by some categories.



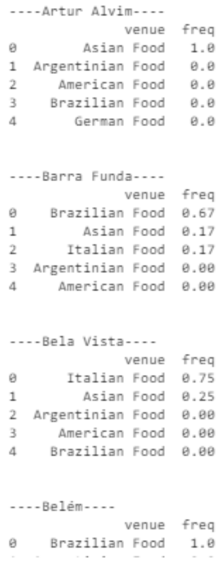
[171]:

	Neighborhood	American Restaurant	Argentinian Restaurant	Asian Restaurant	Baiano Restaurant	Bistro	Brazilian Restaurant	Cajun / Creole Restaurant	Chinese Restaurant	Comfort Food Restaurant	Doner Restaurant	Dumpling Restaurant	Empanada Restaurant	Falafel Restaurant	Fast Food Restaurant	German Restaurant	Greek Restaurant	I Resta
0	Artur Alvim	0.0	0.0	1.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	Barra Funda	0.0	0.0	0.0	0.0	0.0	0.50	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.1
2	Bela Vista	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.7
3	Belém	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
4	Bom Retiro	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.083333	0.0



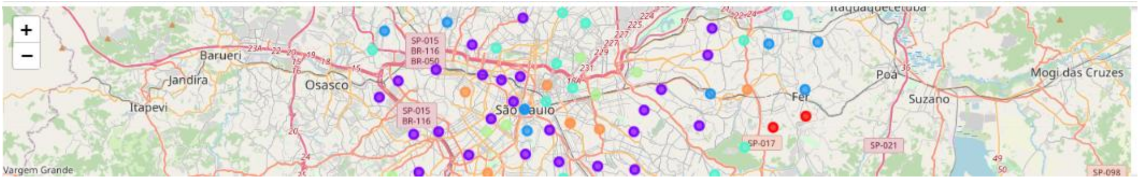
	Neighborhood	Argentinian Food	American Food	Asian Food	Brazilian Food	German Food	Italian Food	Mexican Food	Jewish/Arabian Food	Portuguese Food	Spanish Food	Vegan Food
0	Artur Alvim	0.0	0.0	1.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
1	Barra Funda	0.0	0.0	0.166667	0.666667	0.0	0.166667	0.0	0.000000	0.0	0.0	0.0
2	Bela Vista	0.0	0.0	0.250000	0.000000	0.0	0.750000	0.0	0.000000	0.0	0.0	0.0
3	Belém	0.0	0.0	0.000000	1.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0
4	Bom Retiro	0.0	0.0	0.500000	0.333333	0.0	0.000000	0.0	0.083333	0.0	0.0	0.0

We made an exploratory analysis getting the 5 top categories in each neighborhood.

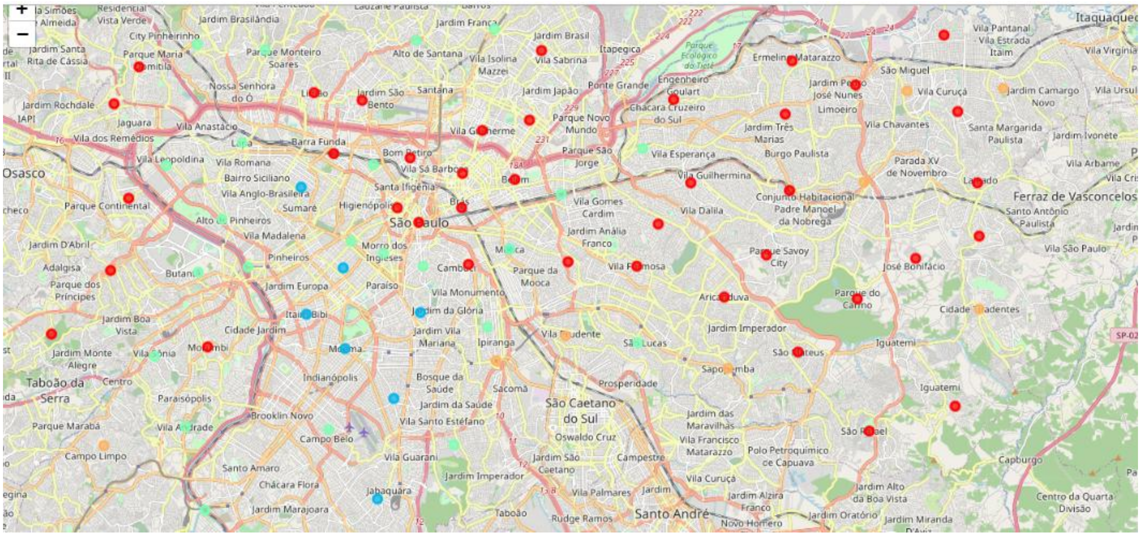


```
----Artur Alvim----
              venue  freq
0        Asian Food   1.0
1  Argentinian Food   0.0
2     American Food   0.0
3    Brazilian Food   0.0
4       German Food   0.0


----Barra Funda----
              venue  freq
0    Brazilian Food  0.67
1        Asian Food  0.17
2      Italian Food  0.17
3  Argentinian Food  0.00
4     American Food  0.00


----Bela Vista----
              venue  freq
0      Italian Food  0.75
1        Asian Food  0.25
2  Argentinian Food  0.00
3     American Food  0.00
4    Brazilian Food  0.00


----Belém----
              venue  freq
0    Brazilian Food   1.0
```

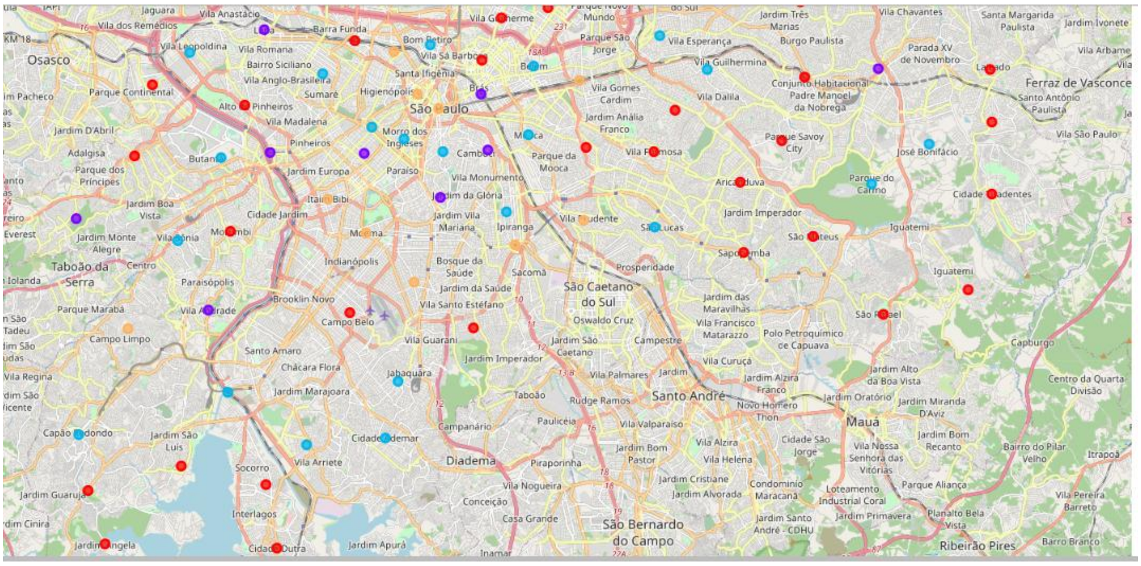
Exploratory by the type of Restaurants



Exploratory By Education Degree



Exploratory by house building



Results for Education

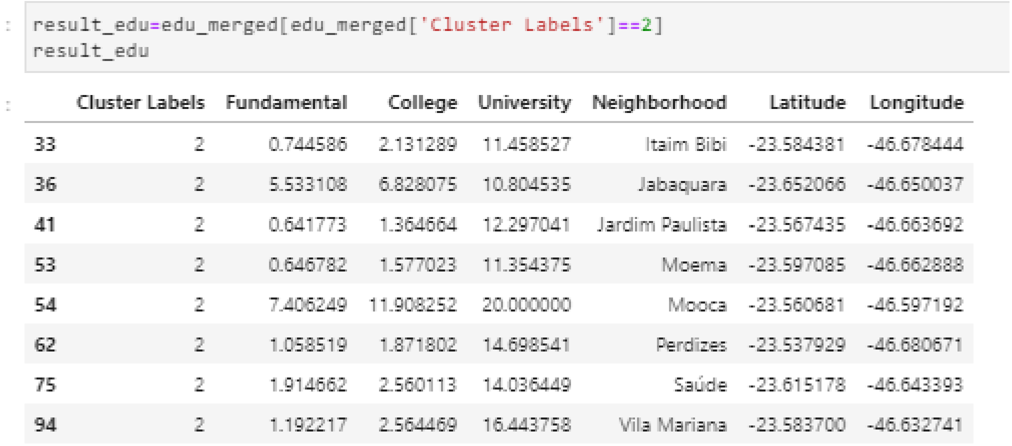
- We have more people with low education level, incomplete high school or college degree

- Low education with few people with University degree

- We have good education skills at the average with College and University degree on this group.

- We have more people with University mainly in this group.

Taking this analysis we must consider Purple and Orange neighborhoods.



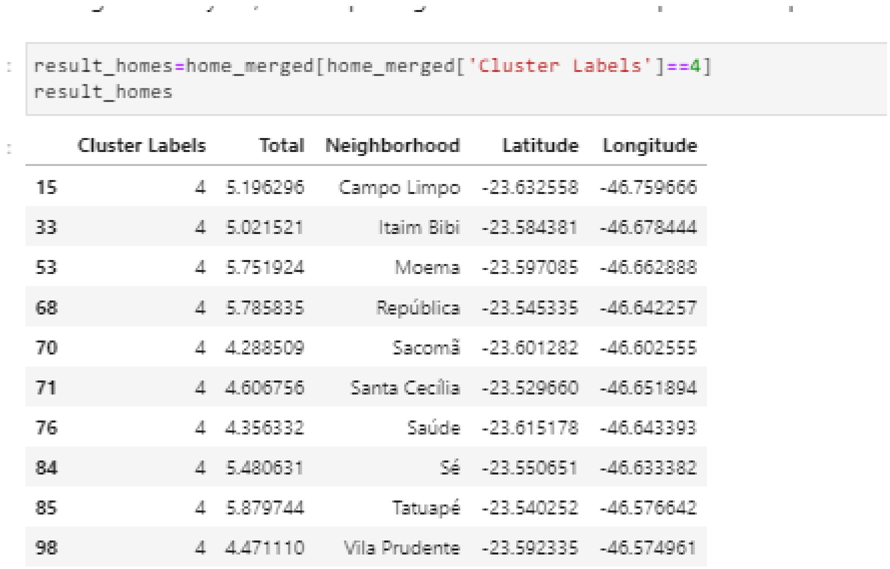
```
result_edu=edu_merged[edu_merged['Cluster Labels']==2]
result_edu
```

	Cluster Labels	Fundamental	College	University	Neighborhood	Latitude	Longitude
33	2	0.744586	2.131289	11.458527	Itaim Bibi	-23.584381	-46.678444
36	2	5.533108	6.828075	10.804535	Jabaquara	-23.652066	-46.650037
41	2	0.641773	1.364664	12.297041	Jardim Paulista	-23.567435	-46.663692
53	2	0.646782	1.577023	11.354375	Moema	-23.597085	-46.662888
54	2	7.406249	11.908252	20.000000	Mooca	-23.560681	-46.597192
62	2	1.058519	1.871802	14.698541	Perdizes	-23.537929	-46.680671
75	2	1.914662	2.560113	14.036449	Saúde	-23.615178	-46.643393
94	2	1.192217	2.564469	16.443758	Vila Mariana	-23.583700	-46.632741

Results for House Building

- Red Few houses builded
- Blue Up to 1.500 houses builded
- Orange Up to 2.000 houses builded
- Purple More than 2.000 houses builded

Taking this analysis , the Purple region has more development and possible new customers.



```
result_homes=home_merged[home_merged['Cluster Labels']==4]
result_homes
```

	Cluster Labels	Total	Neighborhood	Latitude	Longitude
15	4	5.196296	Campo Limpo	-23.632558	-46.759666
33	4	5.021521	Itaim Bibi	-23.584381	-46.678444
53	4	5.751924	Moema	-23.597085	-46.662888
68	4	5.785835	República	-23.545335	-46.642257
70	4	4.288509	Sacomã	-23.601282	-46.602555
71	4	4.606756	Santa Cecília	-23.529660	-46.651894
76	4	4.356332	Saúde	-23.615178	-46.643393
84	4	5.480631	Sé	-23.550651	-46.633382
85	4	5.879744	Tatuapé	-23.540252	-46.576642
98	4	4.471110	Vila Prudente	-23.592335	-46.574961

Discussion section

According the above results we can notice that Moema, Saúde and Itaim Bibi are the recommended Neighbors to open a new restaraunt as it has more people with high education and with more investiment for new house building. We are not taking in count the total population in each neighbour but this indicator are enough in order to suppose good neighbours/districts to open.

Let´s see the top 3 types of restaurant in each neighborhood that we discovered:

Moema

- Jewish/Arabian Food
- Asian Food
- Brazilian Food

Saúde

- Brazilian Food
- Asian Food
- Vegan Food

Itaim Bibi

- Brazilian Food
- Asian Food
- Italian Food

Conclusion section

Taking in count the type of restaurants we could consider the 3dr type of restaurant in each neighborhood for instance. We could consider this because other restaurants should have less demand and this top 3 are already stabilished and success type of restaurants on each location.

Using this methodology, we can conclude therefore that we could open a new Italian Restaurant on Itaim Bibi, a new Vegan restaurant on Saude or a new Brazilian restaurant on Moema neighborhood.