# Professional Certificate – Data Science

Coursera – IBM

## Capstone Project

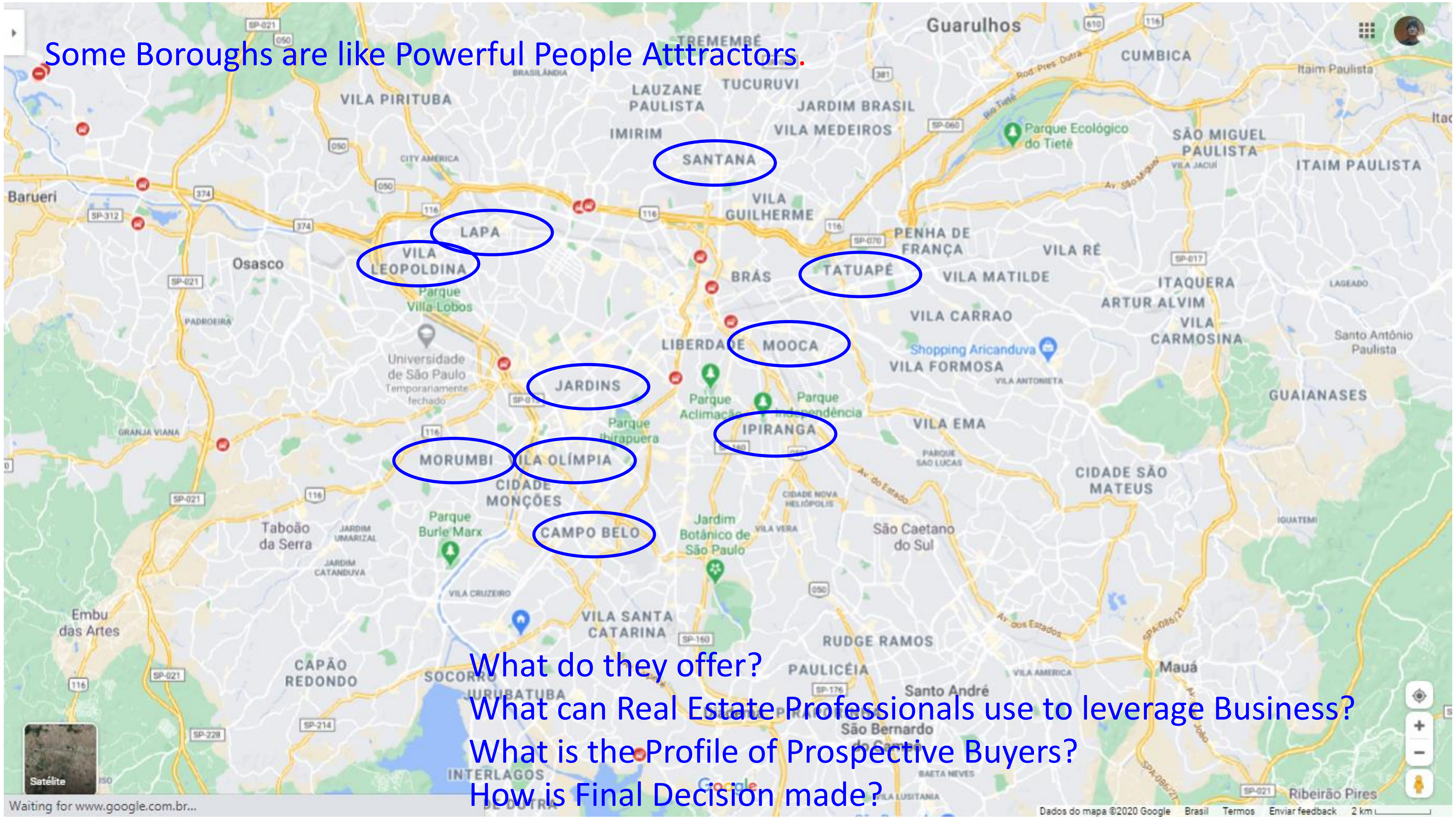
Residential Real Estate Market in São Paulo – Brazil

Alexandre Neri

# City of São Paulo - Overview

- São Paulo: Population 15 million inhabitants
- Metropolitan Area: Population 6 million inhabitants
- Average Daily Commuting Time from Satellite Cities in Metropolitan Area to São Paulo (Home-Work-Home) is 3 hours
- What do these 6 million people want? They want to move to São Paulo
- What kind of Housing do they want? The one that is affordable and suits their “Life Style”, that is, their professional and personal moment in life.

# City of São Paulo - Overview




Some Boroughs are like Powerful People Atttractors.

What do they offer?

What can Real Estate Professionals use to leverage Business?

What is the Profile of Prospective Buyers?

How is Final Decision made?

# What is available for Prospective Buyers

- **Studio**: It looks like a hotel room with some. Area ranges from 25 to 33 square meters;


- **Regular**: It is the "regular" apartment everybody is used to => 2, 3 or 4 bedrooms (sometimes with exclusive bathrooms), living room, kitchen, dining room, etc. Area ranges from 50 to 180 square meters.

CONDE DE IRAJÁ - 4 DORMS (2 SUÍTES) COM COZINHA ABERTA

# What is the Profile of Prospective Buyers

- **Studio Buyer**: usually a single young person starting a career and who has different values from their parents. A car is no longer desirable and Uber or Public Transportation are preferred. Technology-addicted with a taste for cosmopolitan cuisines;

- **Regular Buyer**: for them being married is the most common marital status and they want a place to live for several years and build a family. They are people that **could** once have been Studio buyers but also people from neighbor towns that want to save precious time by quitting commuting. Car (or cars) is a statement of social position and so are expensive Health plans;

- **Investor Buyer**: this buyer is the most senior of all types. It is common in Brazil people who retire but still need to have income. Some of these people acquire apartments (and houses) to rent. They might acquire any type of housing, Studios os Regulars, in any combination they think may maximize their income.

**A little bit into the Model.**

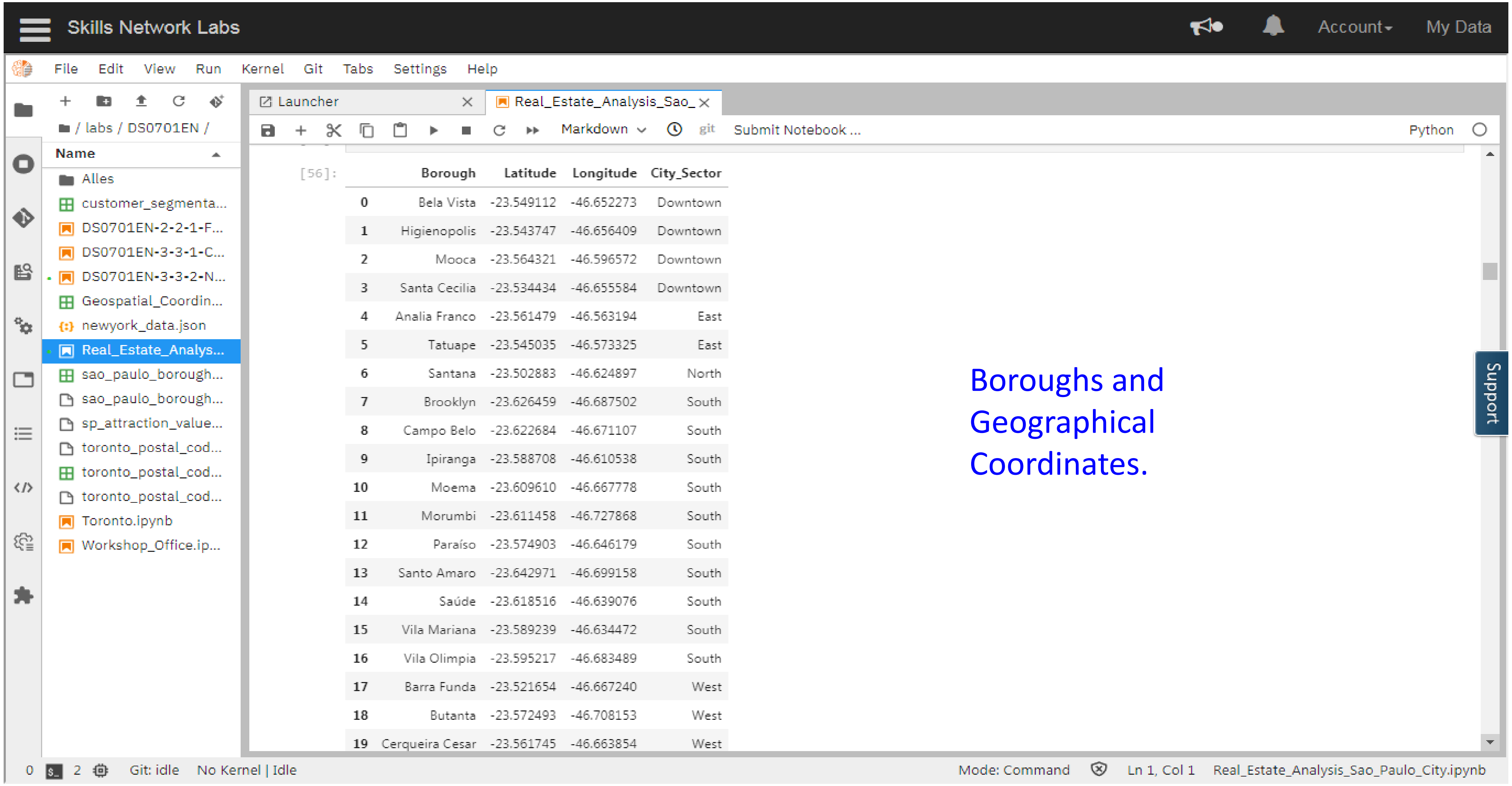
✓ ***Professionals targeted***:

Taking into account everything that was briefly explained above, the underlying idea for this Notebook lies on providing **Potential Buyers and Residential Real Estate Professionals** an index-based tool to help them identifying their targets and making together the best of business opportunities.

✓ ***Methodology:***

- Selection of 25 São Paulo boroughs, known for decades as the best locations in the city for real estate investment;
- Extraction by Foursquare API of all interesting venues around each of these boroughs;
- Further Classification of borough venues. It means concatenation of similar ones into a wider one (Bar and Pub tags into one single tag “Bar”);
- Development of "Buyer-Drivers-and-Preferences" dataframe which will lock to the Classification labels built on previous item;
- **"Attractiveness Index"** calculation which will be the multiplication of the number(frequency) of venue occurrence by the "power of attraction" it displays for each type of buyer. It is a Weighted Average calculation;
- Index-based Decisions, Insights and Statistical Inferences

# São Paulo - Boroughs in Focus:




| | Borough | Latitude | Longitude | City_Sector |
|---|---|---|---|---|
| 0 | Bela Vista | -23.549112 | -46.652273 | Downtown |
| 1 | Higienopolis | -23.543747 | -46.656409 | Downtown |
| 2 | Mooca | -23.564321 | -46.596572 | Downtown |
| 3 | Santa Cecilia | -23.534434 | -46.655584 | Downtown |
| 4 | Analia Franco | -23.561479 | -46.563194 | East |
| 5 | Tatuape | -23.545035 | -46.573325 | East |
| 6 | Santana | -23.502883 | -46.624897 | North |
| 7 | Brooklyn | -23.626459 | -46.687502 | South |
| 8 | Campo Belo | -23.622684 | -46.671107 | South |
| 9 | Ipiranga | -23.588708 | -46.610538 | South |
| 10 | Moema | -23.609610 | -46.667778 | South |
| 11 | Morumbi | -23.611458 | -46.727868 | South |
| 12 | Paraíso | -23.574903 | -46.646179 | South |
| 13 | Santo Amaro | -23.642971 | -46.699158 | South |
| 14 | Saúde | -23.618516 | -46.639076 | South |
| 15 | Vila Mariana | -23.589239 | -46.634472 | South |
| 16 | Vila Olimpia | -23.595217 | -46.683489 | South |
| 17 | Barra Funda | -23.521654 | -46.667240 | West |
| 18 | Butanta | -23.572493 | -46.708153 | West |
| 19 | Cerqueira Cesar | -23.561745 | -46.663854 | West |

Boroughs and Geographical Coordinates.

# After some data cleansing.....



Skills Network Labs    Account▾    My Data

File  Edit  View  Run  Kernel  Git  Tabs  Settings  Help

/ labs / DS0701EN /

Name
- Alles
- customer_segmenta...
- DS0701EN-2-2-1-F...
- DS0701EN-3-3-1-C...
- DS0701EN-3-3-2-N...
- Geospatial_Coordin...
- newyork_data.json
- Real_Estate_Analys...
- sao_paulo_borough...
- sao_paulo_borough...
- sp_attraction_value...
- toronto_postal_cod...
- toronto_postal_cod...
- toronto_postal_cod...
- Toronto.ipynb
- Workshop_Office.ip...

Launcher    Real_Estate_Analysis_Sao_

Code    Submit Notebook ...    Python

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Bela Vista | Edifício Consolata | Residential Building (Apartment / Condo) | -23.549069 | -46.652204 | Yes |
| 4 | Bela Vista | Estação Luz (CPTM) | Train Station | -23.536232 | -46.634527 | Yes |
| 5 | Bela Vista | Ministério das Relações Exteriores | Office | -23.549116 | -46.652432 | Yes |
| 6 | Bela Vista | Volanty | Automotive Services (Sales & Maintenance) | -23.548967 | -46.651949 | Yes |
| 7 | Bela Vista | Charme De Higienopolis | Restaurant | -23.549001 | -46.651970 | Yes |

[72]:

```
# drops columns no longer useful
new_sp_df = sp_df.drop('Decision', 1)
new_sp_df.head()
```

What the Boroughs have to offer or
What people want .....

[72]:

| | Borough | Venue_Name | Business_Category | Venue_Latitude | Venue_Longitude |
|---|---|---|---|---|---|
| 0 | Bela Vista | Estação Higienópolis - Mackenzie (Metrô) | Metro Station | -23.548846 | -46.652371 |
| 1 | Bela Vista | Palácio Paralelo | Snack Place | -23.549162 | -46.652235 |
| 2 | Bela Vista | Santa Gula | Restaurant | -23.547085 | -46.649708 |
| 3 | Bela Vista | Edifício Consolata | Residential Building (Apartment / Condo) | -23.549069 | -46.652204 |
| 4 | Bela Vista | Estação Luz (CPTM) | Train Station | -23.536232 | -46.634527 |

**Step 05:**

Frequency of Venue Type Occurence Calculation

[73]:

```
sampa_onehot = pd.get_dummies(new_sp_df[['Business_Category']], prefix="", prefix_sep="")
sampa_onehot['Borough'] = new_sp_df['Borough']

fixed_columns = [sampa_onehot.columns[-1]] + list(sampa_onehot.columns[:-1]) #establishes column sequence(order)
sampa_onehot = sampa_onehot[fixed_columns]

sampa_grouped = sampa_onehot.groupby('Borough').mean().reset_index()
```

Support

0  2  Git: idle  No Kernel | Idle    Mode: Command    Ln 1, Col 1    Real_Estate_Analysis_Sao_Paulo_City.ipynb

# What do Boroughs offer? Occurrence Frequencies

| | Borough | Services (Sales & Maintenance) | Bank | Bar | Bus Line | Bus Stop | Church & Temple | Clothing & Shoes | Coffee Shop | Coworking Space | Dentist's Office | Doctor's Office | Educational Services | Electronics & Mobile | St |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Analia Franco | 0.028571 | 0.071429 | 0.000000 | 0.014286 | 0.014286 | 0.014286 | 0.242857 | 0.042857 | 0.000000 | 0.014286 | 0.028571 | 0.100000 | 0.057143 | 0.00 |
| 1 | Barra Funda | 0.027778 | 0.055556 | 0.013889 | 0.055556 | 0.013889 | 0.000000 | 0.013889 | 0.027778 | 0.083333 | 0.000000 | 0.013889 | 0.000000 | 0.027778 | 0.01 |
| 2 | Bela Vista | 0.076923 | 0.000000 | 0.076923 | 0.000000 | 0.000000 | 0.000000 | 0.019231 | 0.019231 | 0.019231 | 0.019231 | 0.038462 | 0.076923 | 0.000000 | 0.00 |
| 3 | Brooklyn | 0.023256 | 0.034884 | 0.011628 | 0.023256 | 0.081395 | 0.000000 | 0.011628 | 0.011628 | 0.034884 | 0.011628 | 0.023256 | 0.011628 | 0.000000 | 0.02 |
| 4 | Butanta | 0.034884 | 0.069767 | 0.011628 | 0.151163 | 0.104651 | 0.000000 | 0.011628 | 0.081395 | 0.011628 | 0.046512 | 0.023256 | 0.011628 | 0.023256 | 0.00 |
| 5 | Campo Belo | 0.051948 | 0.012987 | 0.025974 | 0.000000 | 0.000000 | 0.000000 | 0.038961 | 0.000000 | 0.025974 | 0.025974 | 0.103896 | 0.025974 | 0.012987 | 0.01 |
| 6 | Cerqueira Cesar | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.012987 | 0.012987 | 0.142857 | 0.051948 | 0.051948 | 0.025974 | 0.038961 | 0.038961 | 0.012987 | 0.00 |
| 7 | Higienopolis | 0.000000 | 0.085106 | 0.021277 | 0.021277 | 0.042553 | 0.000000 | 0.042553 | 0.106383 | 0.021277 | 0.042553 | 0.063830 | 0.063830 | 0.000000 | 0.02 |
| 8 | Ipiranga | 0.169492 | 0.033898 | 0.033898 | 0.000000 | 0.000000 | 0.033898 | 0.000000 | 0.016949 | 0.067797 | 0.000000 | 0.000000 | 0.152542 | 0.000000 | 0.03 |
| 9 | Lapa | 0.047059 | 0.023529 | 0.000000 | 0.000000 | 0.011765 | 0.023529 | 0.035294 | 0.023529 | 0.035294 | 0.070588 | 0.058824 | 0.023529 | 0.023529 | 0.02 |
| 10 | Moema | 0.036364 | 0.090909 | 0.000000 | 0.000000 | 0.018182 | 0.018182 | 0.236364 | 0.072727 | 0.000000 | 0.072727 | 0.054545 | 0.000000 | 0.054545 | 0.03 |
| 11 | Mooca | 0.061728 | 0.037037 | 0.061728 | 0.012346 | 0.012346 | 0.012346 | 0.024691 | 0.024691 | 0.049383 | 0.012346 | 0.049383 | 0.037037 | 0.012346 | 0.03 |
| 12 | Morumbi | 0.101695 | 0.016949 | 0.050847 | 0.050847 | 0.016949 | 0.016949 | 0.000000 | 0.016949 | 0.000000 | 0.000000 | 0.016949 | 0.033898 | 0.000000 | 0.03 |
| 13 | Paraíso | 0.010638 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.010638 | 0.010638 | 0.031915 | 0.053191 | 0.085106 | 0.031915 | 0.021277 | 0.00 |
| 14 | Paulista | 0.000000 | 0.034884 | 0.000000 | 0.011628 | 0.023256 | 0.000000 | 0.000000 | 0.023256 | 0.069767 | 0.011628 | 0.034884 | 0.046512 | 0.011628 | 0.00 |
| 15 | Perdizes | 0.159420 | 0.000000 | 0.000000 | 0.014493 | 0.014493 | 0.000000 | 0.043478 | 0.028986 | 0.043478 | 0.000000 | 0.028986 | 0.014493 | 0.014493 | 0.04 |
| 16 | Pinheiros | 0.030303 | 0.075758 | 0.000000 | 0.000000 | 0.015152 | 0.000000 | 0.075758 | 0.045455 | 0.030303 | 0.015152 | 0.060606 | 0.015152 | 0.000000 | 0.04 |
| 17 | Santa Cecilia | 0.000000 | 0.000000 | 0.011765 | 0.000000 | 0.011765 | 0.000000 | 0.023529 | 0.023529 | 0.035294 | 0.082353 | 0.058824 | 0.000000 | 0.000000 | 0.00 |

# What do Buyers want? Preferences Matrix

```
[75]: weights = pd.read_excel("sp_attraction_values.xlsx")
weights.head()
```

[75]:

| | Buyer Type | Automotive Services (Sales & Maintenance) | Bank | Bar | Bus Line | Bus Stop | Church & Temple | Clothing & Shoes | Coffee Shop | Coworking Space | Dentist's Office | Doctor's Office | Educational Services | Electronics & Mobile | Gas Station | Gym | Laundry Service | (H Lab |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Studio Buyer | 0.00 | 0.00 | 0.02 | 0.05 | 0.05 | 0.02 | 0.040 | 0.05 | 0.05 | 0.02 | 0.02 | 0.05 | 0.05 | 0.00 | 0.05 | 0.05 | |
| 1 | Regular Buyer | 0.10 | 0.04 | 0.00 | 0.00 | 0.00 | 0.08 | 0.007 | 0.03 | 0.03 | 0.03 | 0.04 | 0.02 | 0.06 | 0.10 | 0.00 | 0.01 | |
| 2 | Investor Buyer | 0.03 | 0.01 | 0.01 | 0.04 | 0.04 | 0.05 | 0.020 | 0.03 | 0.04 | 0.04 | 0.03 | 0.05 | 0.05 | 0.03 | 0.03 | 0.03 | |

```
[76]: weights.shape[0]
```

[76]: 3

**Step 07:**

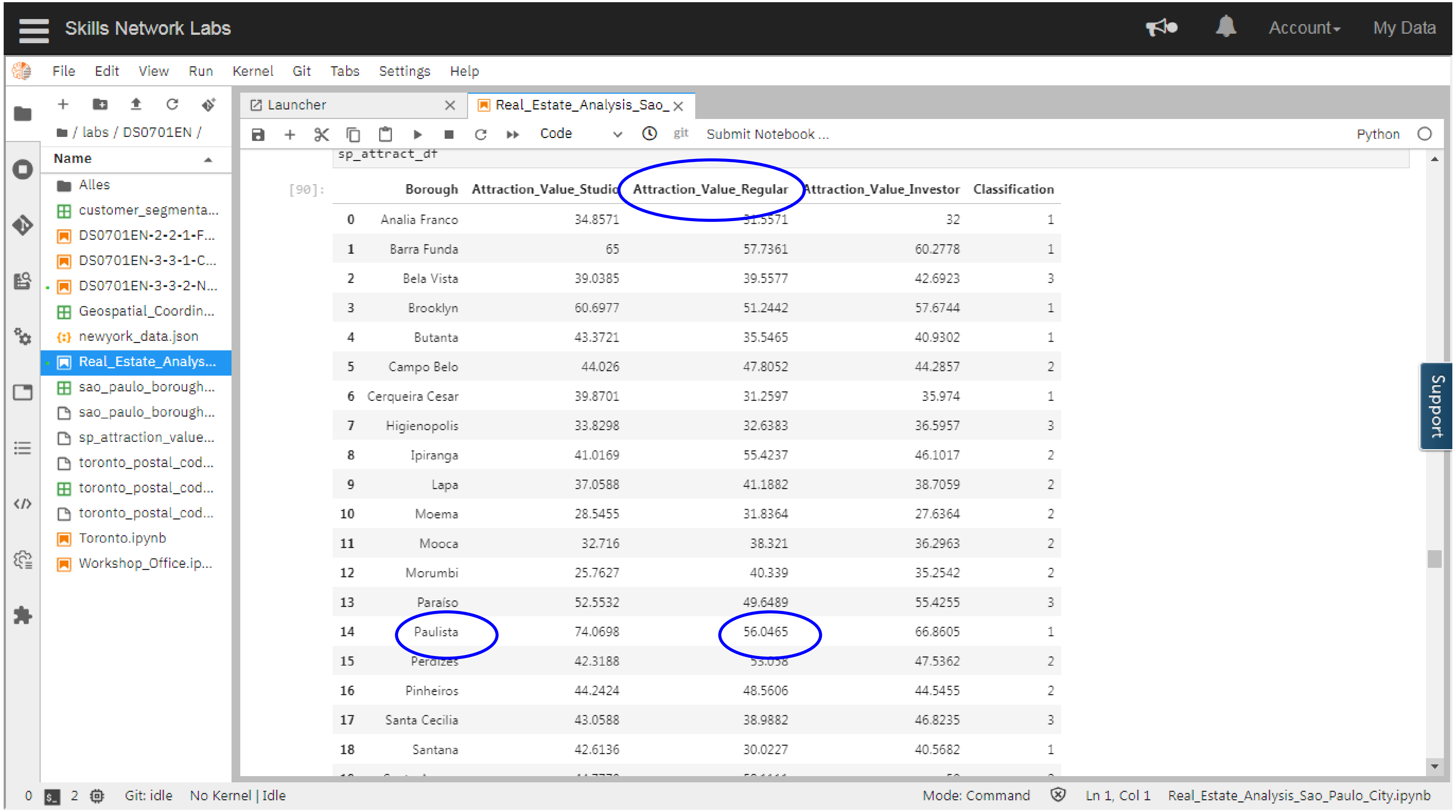
Final Dataframe Definition **(Attractiveness Indexes)** and Software-Oriented Decision

```
[77]: # define the dataframe columns
column_names = ['Borough','Attraction_Value_Studio', 'Attraction_Value_Regular','Attraction_Value_Investor', 'Classification']

# instantiate the dataframe
sp_attract_df = pd.DataFrame(columns=column_names)
```

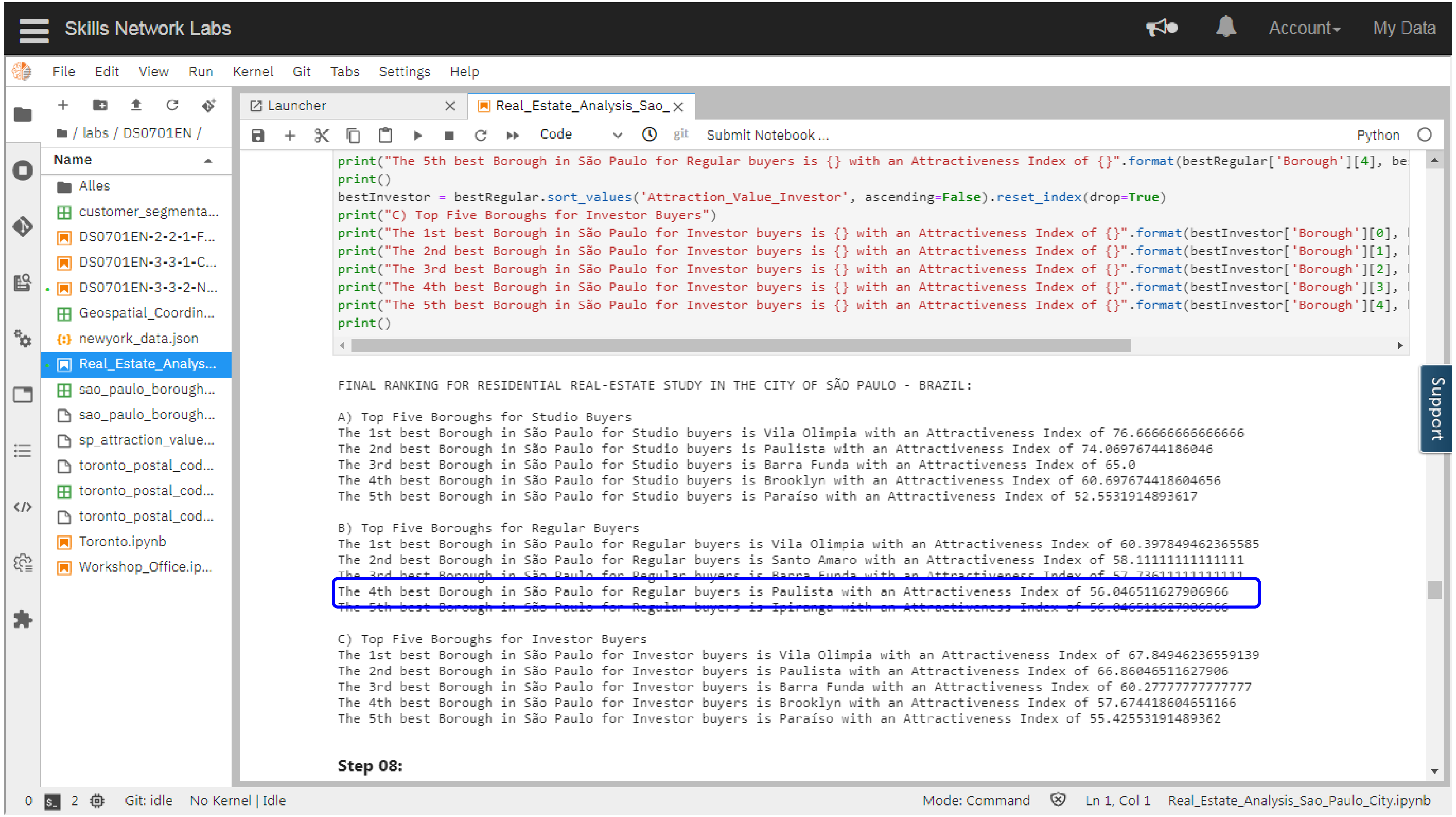
```
[78]: sp_attract_df.head()
```

# The Attractiveness Index for each Borough

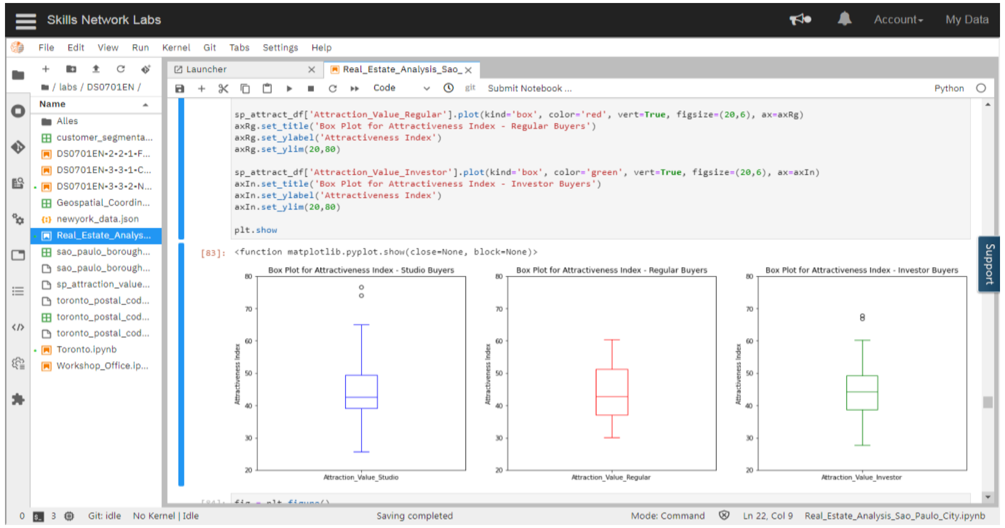


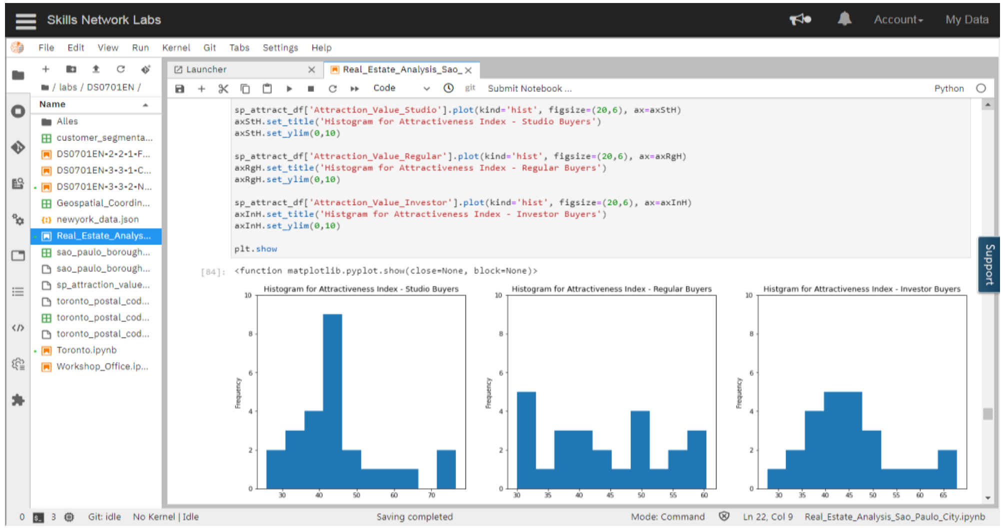

| | Borough | Attraction_Value_Studio | Attraction_Value_Regular | Attraction_Value_Investor | Classification |
|---|---|---|---|---|---|
| 0 | Analia Franco | 34.8571 | 31.5571 | 32 | 1 |
| 1 | Barra Funda | 65 | 57.7361 | 60.2778 | 1 |
| 2 | Bela Vista | 39.0385 | 39.5577 | 42.6923 | 3 |
| 3 | Brooklyn | 60.6977 | 51.2442 | 57.6744 | 1 |
| 4 | Butanta | 43.3721 | 35.5465 | 40.9302 | 1 |
| 5 | Campo Belo | 44.026 | 47.8052 | 44.2857 | 2 |
| 6 | Cerqueira Cesar | 39.8701 | 31.2597 | 35.974 | 1 |
| 7 | Higienopolis | 33.8298 | 32.6383 | 36.5957 | 3 |
| 8 | Ipiranga | 41.0169 | 55.4237 | 46.1017 | 2 |
| 9 | Lapa | 37.0588 | 41.1882 | 38.7059 | 2 |
| 10 | Moema | 28.5455 | 31.8364 | 27.6364 | 2 |
| 11 | Mooca | 32.716 | 38.321 | 36.2963 | 2 |
| 12 | Morumbi | 25.7627 | 40.339 | 35.2542 | 2 |
| 13 | Paraíso | 52.5532 | 49.6489 | 55.4255 | 3 |
| 14 | Paulista | 74.0698 | 56.0465 | 66.8605 | 1 |
| 15 | Perdizes | 42.3188 | 53.038 | 47.5362 | 2 |
| 16 | Pinheiros | 44.2424 | 48.5606 | 44.5455 | 2 |
| 17 | Santa Cecilia | 43.0588 | 38.9882 | 46.8235 | 3 |
| 18 | Santana | 42.6136 | 30.0227 | 40.5682 | 1 |

# Model Recommendations



# Statistical Analysis




**Box Plot:**

Dispersion for Studio **Attr_Idx** is the highest of all three and it probably indicates that Real Estate Enterprises are doing their homework as they have been placing their "Studio Launch Sites" in areas already perceived by prospective buyers as very promising to their life style.

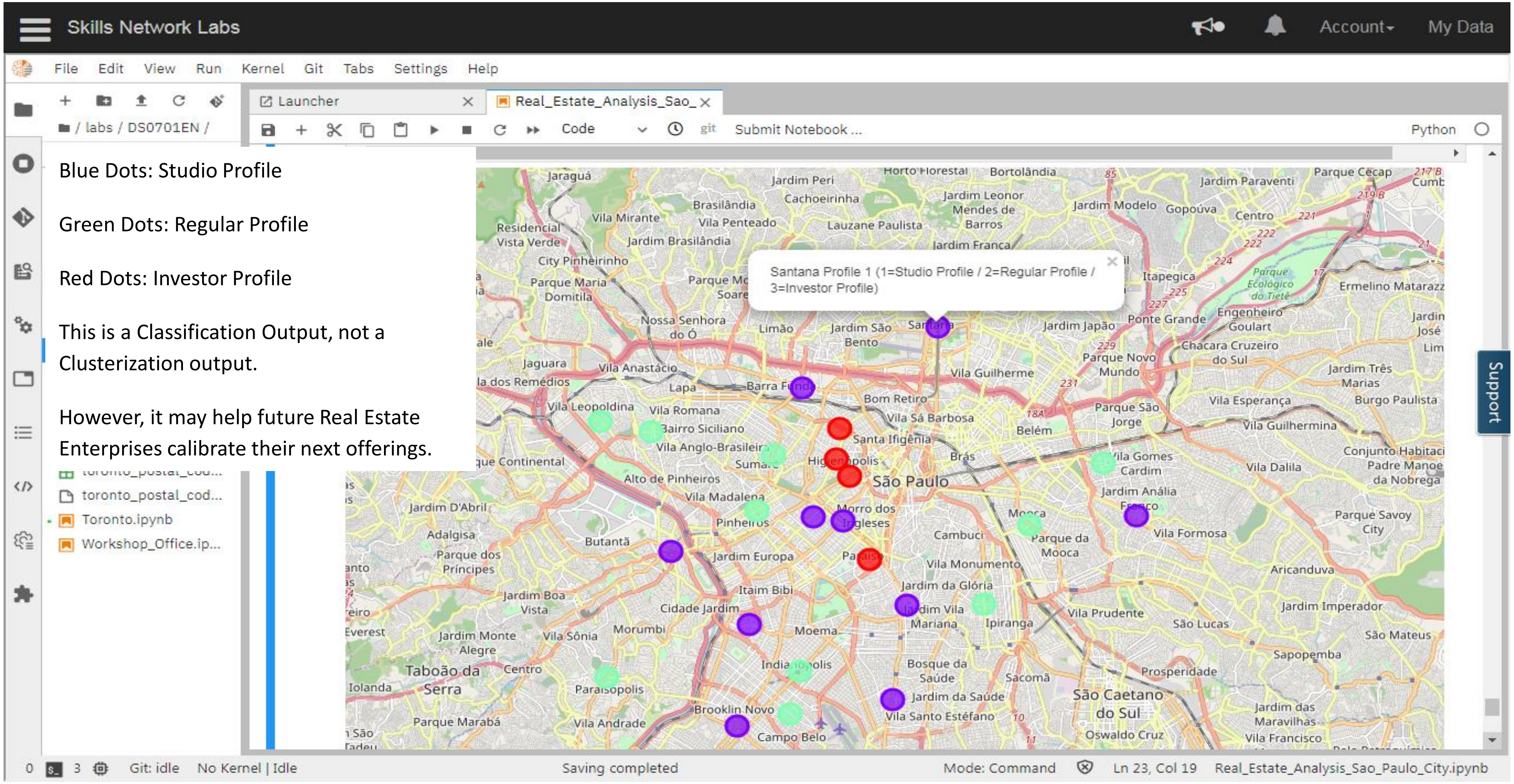
Now, putting Interquartile Range (IQR) under the microscope for all three Buyers' profile shows Regular **Attr_Idx** displays the broadest dispersion which is clearly visible by histogram visual inspection. Regular Buyer is the most frequent buyer in Showrooms and this availability in numbers translates into a more diverse palette of needs and preferences. It explains this 'fatter' histogram without any clear unique peak.

**Histogram:**

Still about IQR and histograms, **Attr_Idx** for Studio and Investor profiles show a slight tendency to symmentry, albeit being asymmentrical. The former displays a peak that indicates more focus from buyers and sellers and the latter displays a hedge-like behaviour as the result of a balanced-game between Studio and Regular profiles.

Finally, the Median (Q2) for the three **Attr_Idx** are very similar (as the mean average) which could indicate that the whole Residential Real Estate Market for Studios is still in its infancy in São Paulo and perhaps there are still room for information delivery and training for Sellers in order to improve their delivery and minimize gray areas for buyers´ clear understanding.

# Map Visualization




Blue Dots: Studio Profile

Green Dots: Regular Profile

Red Dots: Investor Profile

This is a Classification Output, not a Clusterization output.

However, it may help future Real Estate Enterprises calibrate their next offerings.

# Conclusions

- Regardless of Buyer Profile, Vila Olimpia Borough topped the bill three times which comes at no surprise at all as lots of Tech Companies and Start-Ups are establishing themselves in the area, nicknamed by the midia as the SP Sillicon Valley;

- Paulista borough also appears on the 3 rankings but it is less regarded by Regular buyers than by others. It is a borough characterized by lots of financial companies, services and goverment buildings making it not a good decision to raise kids, but a good one for young people and investors;

- Talking about Buyer Profile, choices for Studio and Investor Buyers were the same for the Top Five positions. Although it might look confusing, there is an underlying logic as Investors want the highest possible ROI, the faster the better. It takes less time to build Studio aparments and tennant turnoround is expected to be higher thus generating quick profit;

- Not a single 'Downtown' borough made the Top Five. Unfortunately, this was expected as these boroughs are still linked to a bad reputation São Paulo downtown got in the 70's and 80's;

- Another borough absent from Top Five is Morumbi, a former powerful attractor area in past decades but still lacking improvements like subway lines that are yet to reach the borough and houses are still the most commom housing. Being at the 'wrong side of the tracks', that is, at the west bank of Pinheiros river does not help either as it is widely accepted that such locations devaluate any Real Estate investment;

- In general terms, the Model works fine as it translates what is known as "common sense" into data-based recommendations for both sides of the equation (buyers and sellers). **Further enhancements could be added to it in next versions like financial data to reckon ROI versus land value (BRL/sqmt) or it could provide interaction with users via website to make the decision process faster**.