Capstone Milestone Report

Using statistical and clustering approaches to study growth and check if Investment is Important for Cities Growth

1. INTRODUCTION

1.1 The study and the client

By the moment we are debating this study, all brazilians are drowned in corruption scandals from politics that develop in many outcomes that range from gossips to economic crisis. One of the main concerns is the management of public resources and the effect of what is promised and what is built with the promises.

One of the main questions that this study aims to answer is the relevance for the investment variables (public and private) in the growth of the cities of the São Paulo province, which is very related to the corruption chapters. A second, and natural, question this study would answer is where growth and its particularities differ from city to city.

The main client for this study is the Public Governor of the province which could, with the results, measure the relevance of the investment as a policy for growth and balance the distribution of this incentive. We could say that the population of the cities and the province could be a additional client, which could use the results as an argument for claiming for a more fair management.

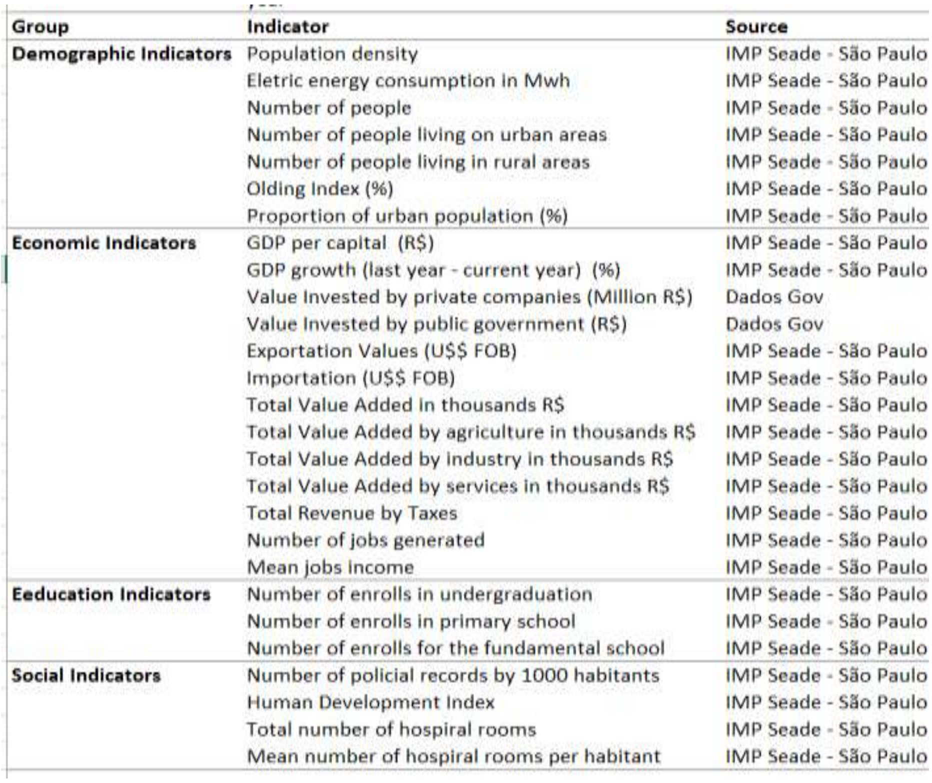
1.2 The data set

The data set used for this study was a merge of two sources:

Investment (public and private) values by year:
http://dados.gov.br/dataset

The rest of the features (demographics, social, economic, etc) :
http://www.imp.seade.gov.br/

The data ranged from the years 2000 to 2016 for all of the 645 cities of the province of São Paulo, Brazil. The features of choice were based on the main pillars of growth that are: education, economics, demographics, and social. The list of features are:



Group	Indicator	Source
Demographic Indicators	Population density	IMP Seade - São Paulo
	Eletric energy consumption in Mwh	IMP Seade - São Paulo
	Number of people	IMP Seade - São Paulo
	Number of people living on urban areas	IMP Seade - São Paulo
	Number of people living in rural areas	IMP Seade - São Paulo
	Olding Index (%)	IMP Seade - São Paulo
	Proportion of urban population (%)	IMP Seade - São Paulo
Economic Indicators	GDP per capital (R$)	IMP Seade - São Paulo
	GDP growth (last year - current year) (%)	IMP Seade - São Paulo
	Value Invested by private companies (Million R$)	Dados Gov
	Value Invested by public government (R$)	Dados Gov
	Exportation Values (U$$ FOB)	IMP Seade - São Paulo
	Importation (U$$ FOB)	IMP Seade - São Paulo
	Total Value Added in thousands R$	IMP Seade - São Paulo
	Total Value Added by agriculture in thousands R$	IMP Seade - São Paulo
	Total Value Added by industry in thousands R$	IMP Seade - São Paulo
	Total Value Added by services in thousands R$	IMP Seade - São Paulo
	Total Revenue by Taxes	IMP Seade - São Paulo
	Number of jobs generated	IMP Seade - São Paulo
	Mean jobs income	IMP Seade - São Paulo
Eeducation Indicators	Number of enrolls in undergraduation	IMP Seade - São Paulo
	Number of enrolls in primary school	IMP Seade - São Paulo
	Number of enrolls for the fundamental school	IMP Seade - São Paulo
Social Indicators	Number of policial records by 1000 habitants	IMP Seade - São Paulo
	Human Development Index	IMP Seade - São Paulo
	Total number of hospiral rooms	IMP Seade - São Paulo
	Mean number of hospiral rooms per habitant	IMP Seade - São Paulo

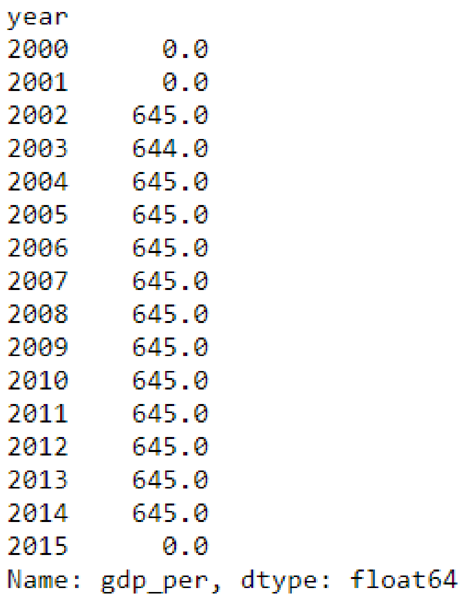
As seen above, this data set was constructed so it would cover 4 main areas on growth: demographic, economic, education and social.This would be important when considering growth in different places because it permits us to explore different aspects of development in different places, giving the possibility of inumerous comparisons.

One important question this data set would not answer is how technology measures would affect cities growth. Not enough reliable data was found at this point. This part of the analysis would be an important point for future works in the area.

1.3 Data Wrangling

The focus of this analysis was on the three main variables for the study: GDP per capita, public and private investment, which are the core for all answer to be solved. So for the data wrangling we are going to check the availability of this set of features and look for the best year to use all of them together for future analysis.

When looking to GDP per capita missing values we could see:

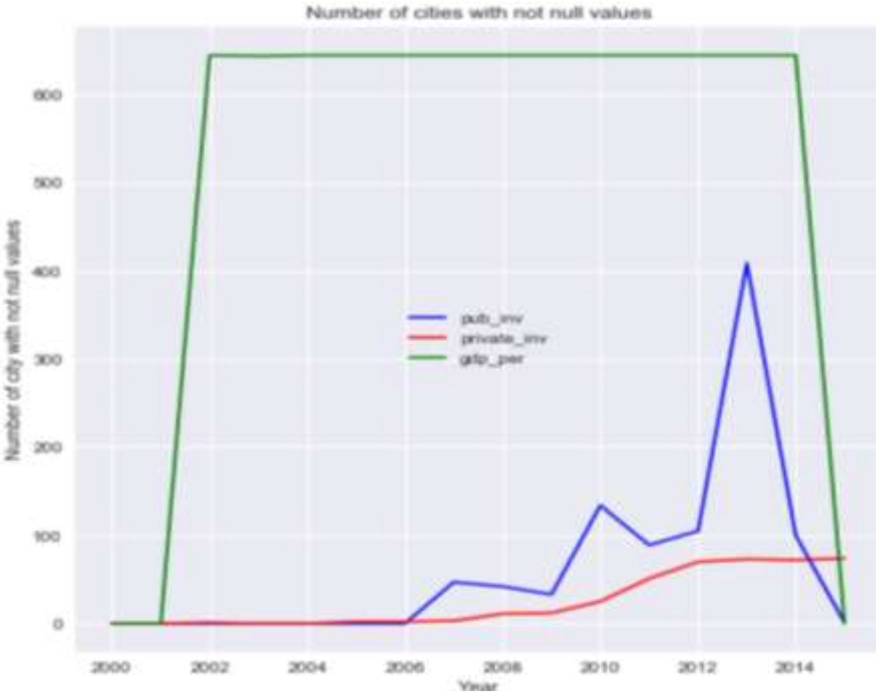


```
year
2000      0.0
2001      0.0
2002    645.0
2003    644.0
2004    645.0
2005    645.0
2006    645.0
2007    645.0
2008    645.0
2009    645.0
2010    645.0
2011    645.0
2012    645.0
2013    645.0
2014    645.0
2015      0.0
Name: gdp_per, dtype: float64
```

Number of entries for GDP per capita variable, per year

Here we could conclude that 2000, 2001 and 2015 were years that should not be used to this analysis, once there were no entries for any one of them.

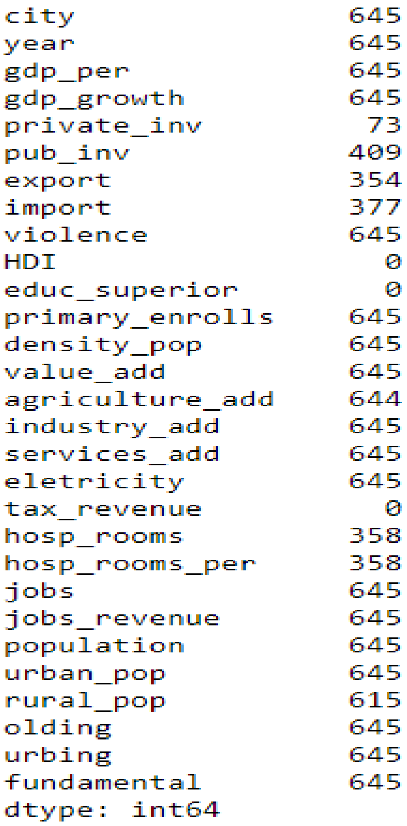
Going a little deeper and trying to see the year with the best combination for the three variables mentioned above, we could see a clear choice as the plot below shows:



Number of entries for the 3 main features, by year

Looking at this plot the scenario is set. For GDP per capita we can chose from 2001 to 2014, as we seen above. But for investment, our best chance is located in 2013, once that private investment is in its top level, and public investment reaches its peek too.

Once that the year of 2013 is a clear choice, our final check would be in the values availability for the other features in the data set.



```
city                   645
year                   645
gdp_per                645
gdp_growth             645
private_inv             73
pub_inv                409
export                 354
import                 377
violence               645
HDI                      0
educ_superior            0
primary_enrolls        645
density_pop            645
value_add              645
agriculture_add        644
industry_add           645
services_add           645
eletricity             645
tax_revenue              0
hosp_rooms             358
hosp_rooms_per         358
jobs                   645
jobs_revenue           645
population             645
urban_pop              645
rural_pop              615
olding                 645
urbing                 645
fundamental            645
dtype: int64
```

Figure 3: Data entries for all the features in the year of 2013

When looking to the Figure 3 we could see that 3 features need to be dropped: HDI, educ_superior and tax_revenue.

After this cleaning, and dropping the 3 features above, the data set was ready for further analysis.

2. PRELIMINARY RESULTS

For the start of the exploration in the data assumptions, two questions were raised:

- Are cities that received private investment in 2013 more richier (in terms of GDP per capita) than cities that did not?
- Are cities that received public investment in 2013 more richier (in terms of GDP per capita) than cities that did not?

We used the GDP per capita mean for the groups with and without investment to check for the difference and its significance.

For the first question, in spite of the fact that we have fewer data for private investment, we saw a p-value of almost 0 for the difference in means. The difference was from 16.327 R$, and proved to be significant (would not occur in the population due to chance).

On the other hand, when looking to the second question, we had a difference in means of 2.252 R$ with a p-value of 0.387. This would imply that this could occur due to chance in the population.

3. FUTURE ANALYSIS AND APPROACH

In spite of the fact that we had some answers using the means in the previous section, growth is, by many studies in the field, considered a multidimensional question. This justifies the interest of the author of this study in trying machine learning routines to solve this question with more care for this fact.

Considering the above, the approach would be the use of models that could capture in a natural way the multidimensionality of growth. Some models like the I-Distance model (the use of hyperplanes to calculate distances between objects of comparison) could capture the essence of cities with different aspects and, as consequence, different levels for the same features but with different levels of growth.

The implementation of clustering would be interesting to find some patterns within the province of São Paulo too.