IBM-Coursera Capstone Project

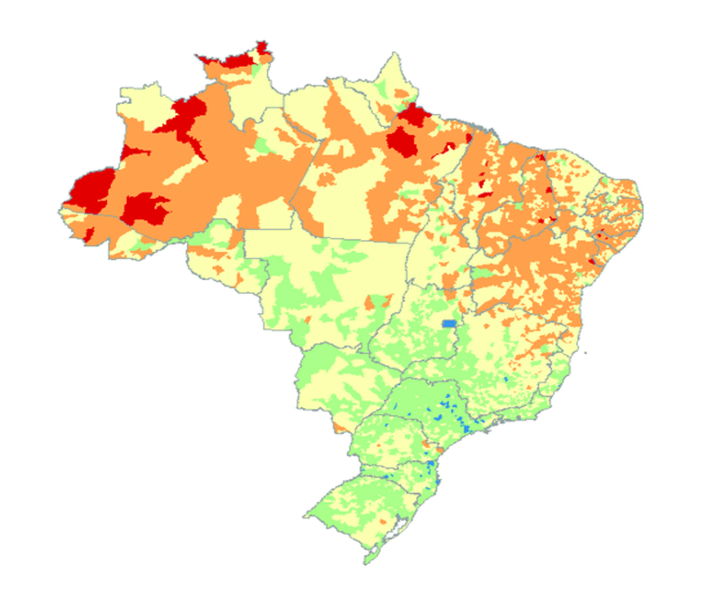
HUMAN DEVELOPMENT INDEX IN BRAZIL:

ANALYSIS OF FACTORS INFLUENCING RESULTS

(material for Blogpost publication in specialized media)

Paulo Bernardocki

2019-01-12



# Introduction/Business Problem

This report is intended to different levels of government audiences in Brazil, from municipality to country level. It intends to explore the effect of different demographic and economical data on the Human Development Index (HDI), for each municipality in the country. There are more than 5,000 municipalities in Brazil.

HCI has historically been correlated by the government with Per Capita Income (PCI). One of the objectives of this report is initially identify the actual correlation between these two indexes.

Additionally, this report will explore the effect of other variables in the HCI. It was chosen one variable related to infrastructure available in the city, and another one related to the economic activity in each municipality: The variables chosen for this task were:

* Infrastructure: number of Radio Base Stations (RBS) providing cellular service in each municipality.  
  The number of radio base station is considered a strong economic indicator of how developed a society is in general.
* Economic-activity: number of Venues listed in Foursquare per municipality.  
  As well as the number of base stations, the number of business venues is a good indication of economic activity.

This report will analyze the effect of these variables in combination with the PCI, and whether they can be used as attributes to predict the HCI of a given municipality.

Population will be used for data normalization wherever applicable.

# Area of analysis profile

Brazil is the largest country in both South America and Latin America. At 8.5 million square kilometers (3.2 million square miles), and with over 208 million people, Brazil is the world's fifth-largest country by area and the fifth most populous.

Brazil has a mixed economy with abundant natural resources. After rapid growth in preceding decades, the country entered an ongoing recession in 2014 amid a political corruption scandal and nationwide protests.

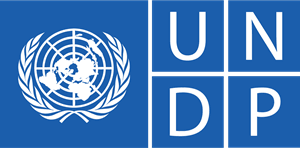
Its GDP (PPP) per capita was $15,919 in 2017[241] putting Brazil in the 77th position according to IMF data. Active in agricultural, mining, manufacturing and service sectors.

# Data

# Sources and Concerns

Brazil is composed of 5,570 municipalities. In order to collect the information related to them, collection was performed from several different sources, indicated below:

* **Human Development Index:** UNDP-United Nations Development Programme



* **Per Capita Income :** IBGE – Instituto Brasieiro de Geografia e Estatística (Brazilian Institute for Geographics and Statistics)
* **Population:** IBGE – Instituto Brasieiro de Geografia e Estatística (Brazilian Institute for Geographics and Statistics)



* **Venues:** Foursquare



* **Radio Base Stations:** ANATEL – Agência Nacional de Telecomunicações (Brazilian National Telecommunications Agency)



Population will be used directly as an attribute, as well as for feature engineering. For example, it will be used to normalize the RBS and Venues data. Therefore, the RBS and Venues data will be converted to ‘per capita’ values, as (intuitively) it is expected that one city with twice the number of inhabitants tends to have two times more venues and RBS deployed to serve its population.

The data will be filtered to identify outliers. E.g. São Paulo city is the largest city in the south hemisphere, and certainly will distort the numbers. Similarly, cities on the lower limit (e.g. under 1,000 habitants) will be reviewed if shall be included.

Foursquare limits the number of venues per request to 100. This cap will create a distortion in any linearity analysis, therefore it will be verified if the cities hitting this cap shall be removed from the analysis to avoid its implications.

By cross-checking the data against the databases, the information needed for the analysis is available for 2970 municipalities (main limitation is on PCI data), which is an amount considered satisfactory for the analysis.

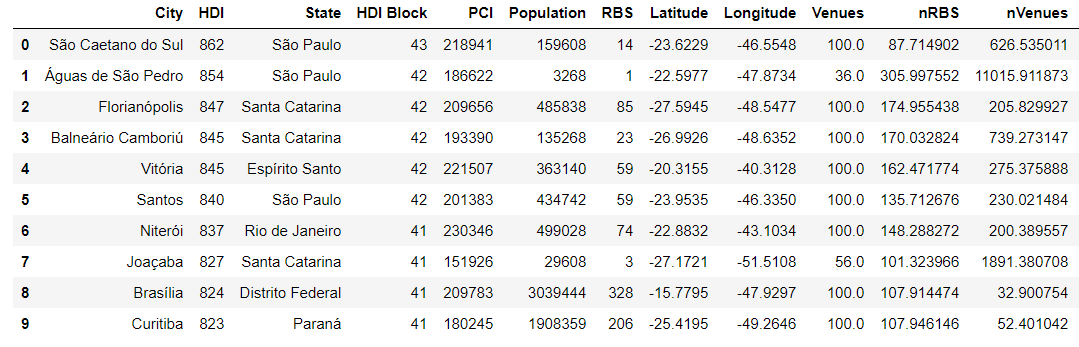
A radius of 3km was used to identify venues from the city center. Lat/long data was obtained using Python Nominatim package.

# Preparation

As mentioned before, the following normalizations were performed.

* **Venues**: was normalized per population, therefore generation the attribute **nVenues** (normalized Venues)
* **RBS**: was normalized per population, therefore generation the attribute **nRBS** (normalized RBS)
* HDI: was discretized in order to allow classiciation analysis. The HDI values are continuous in the range 0-1000. The discretization created steps 50 wide, creating 20 steps in the full range. The discretized attribute is called **HDI Block**.

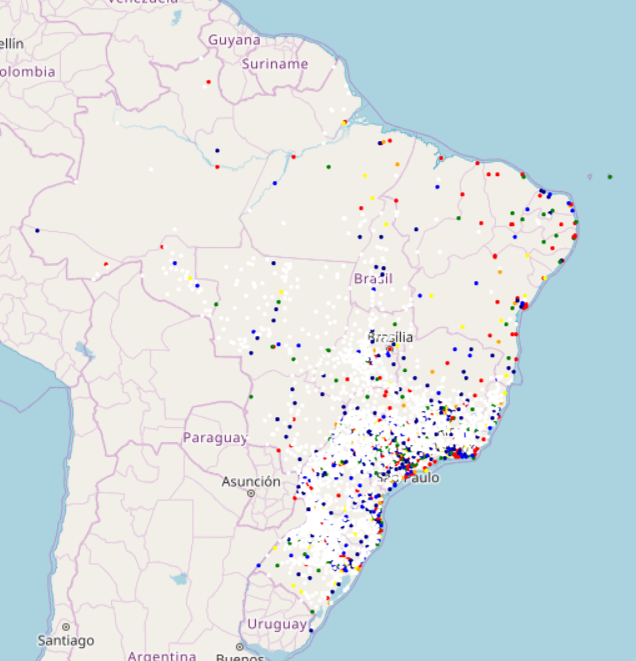
So ,the final data structure used is:

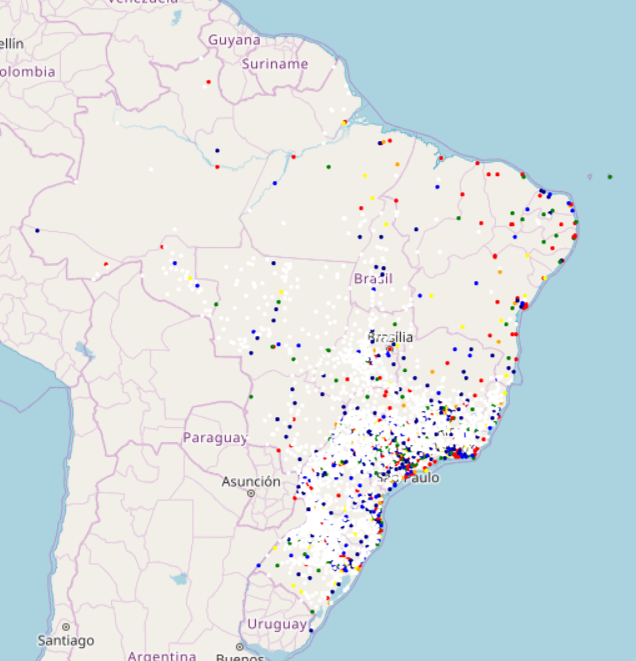


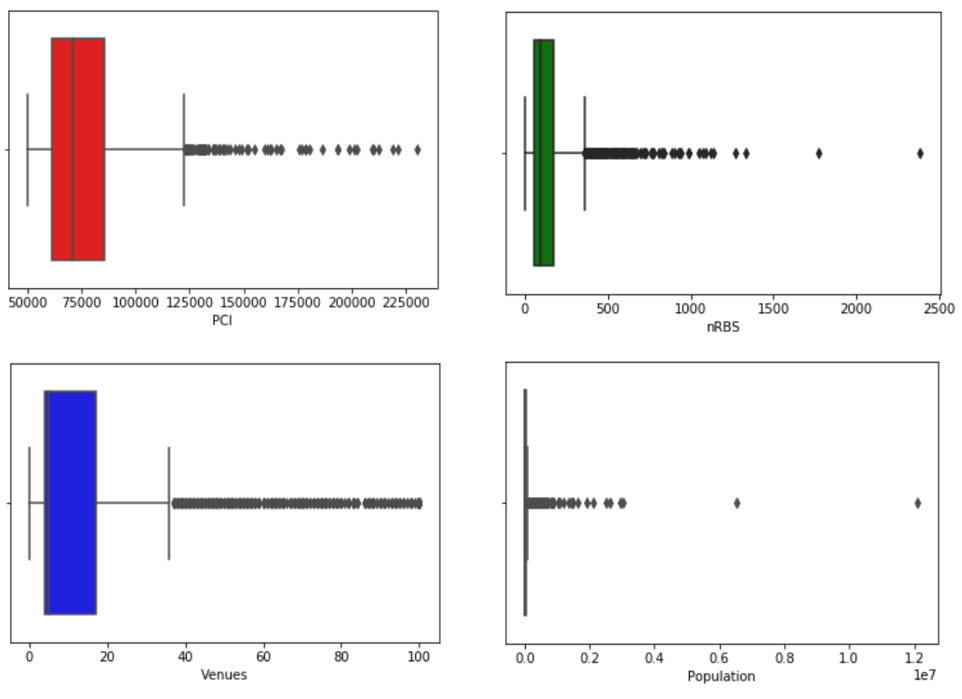
# Exploratory Data Analysis

Initially the distribution of business venues across Brazil was plotted. This allow the observation of both the distribution of municipalities (each dot), and the amount of venues in each one (white=none, blue=few, red=lots).

It can be clearly seen from the picture that there is a concentration of economic activity towards the southeast of the territory, al well as closer to the coast in comparison to the interior.



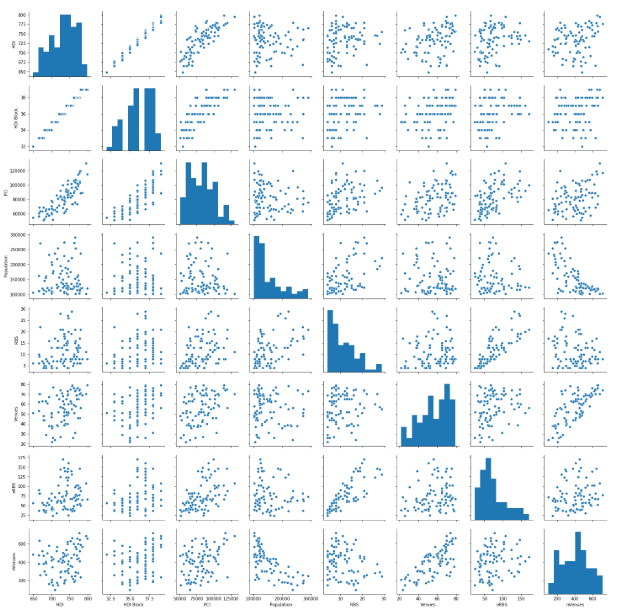


A Boxplot distribution was performed in the main variables in order to identify outliers:

For removing outliers, the following filters were applied:

* Venues: removed values above 80
* Population: removed values below 1k and above 300k inhabitants
* PCI: removed values above 150k

Next it was plotted a pairplot of all variables against each other, in order to observe correlations. It can be noted that PCI has an interesting result against HDI, with the others having a subtler influence.



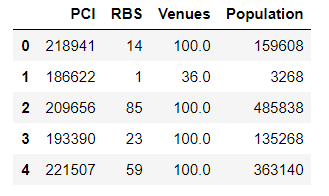
With the removal of outliers, the data become more evenly distributed and it is considered now good for initial data analysis

# Data Analysis

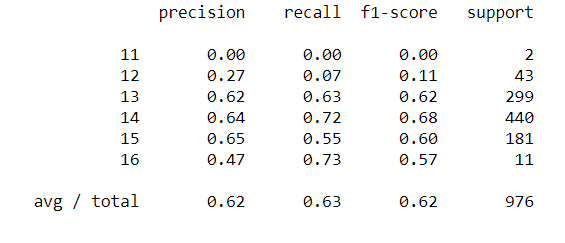
## 1 - CLASSIFICATION

The first method used for analysis was classification. Random Forest, Support Vector and kNN were used in order to evaluate which could provide better results. All have provided comparable results, with a slight advantage to kNN, which will therefore be used.

# k-Nearest Neighbors

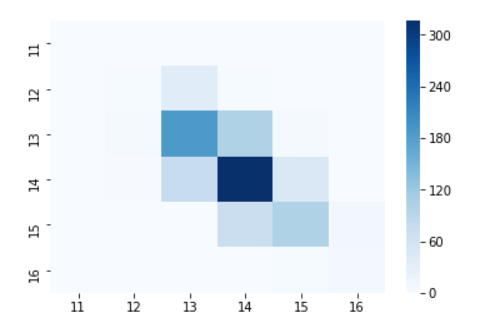
* Different values of K were tested, and the best result was found using k=9
* Train test split: 66 / 33
* Attributes were normalized using Standard Scaler
* All variables contributed positively for the result, with more importance on PCI.
* For kNN the normalized values were NOT used:
* 

CLASSIFICATION REPORT:



CONFUSION MATRIX:

HDI Blocks



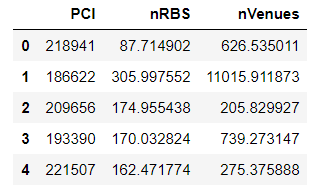
The result in the range of 60-65% was considered good, particularly because a close observation in the confusion matrix indicated that a considerable number of points were misclassified by just one adjacent step.

This leads to the next analysis, as HDI is a continuous variable, a Linear Regression will be performed.

## 2 - LINEAR REGRESSION

The parametrization for linear regression was:

* Train test split: 66 / 33
* Normalized variables were used: nRBS and nVenues



The resulted parameters for linear regression model were:

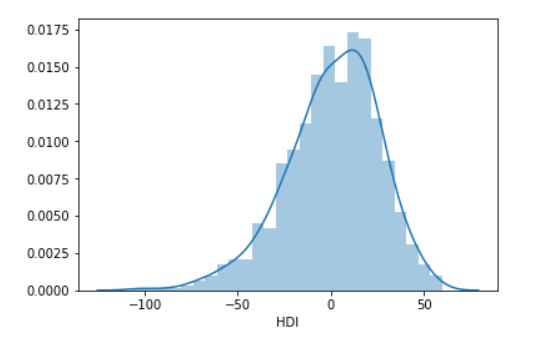


The error was low, compared to the magnitude of the target variable:



And the error distribution approximates a Normal distribution:

ERROR



## CONCLUSION

The usage of demographic, infrastructure and economic variables successfully predicted the behavior of the Human Capital Index across Brazil.

It was noted that the Per Capita Income has a relevant role in this calculation, balanced with the infrastructure proxy (number of Radio Base Stations for cellular service) as well as the economic proxy (the number of business venues in each city) has also a role in this calculation.

This information can be used by majors looking for improving the welfare of its population. Investing in infrastructure and economic growth can leverage the overall welfbeing state.