Impact of Economic Growth and Education in Mortality in the 21st Century

Extended Analysis  
  
Team #6 | Carlos Chávez, Yohana Sandoval, Daniel Duran, Eider Giraldo, Fernán Marsiglia, Juan Pablo Alzate, Marta Ordoñez.

# Overview

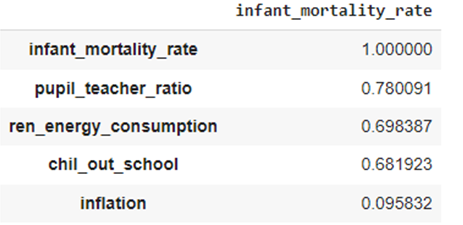
The exploratory data analysis showed the following characteristics:

* Low correlation between variables.
* There was an atypical behavior of the variables in which they are displayed in a solitary way. In the cases they are grouped they show a completely opposite trend, perhaps evidencing a simpson’s paradox.
* When models of linear and multiple regression are utilized with the variables, the results do not appear to fit with the models.

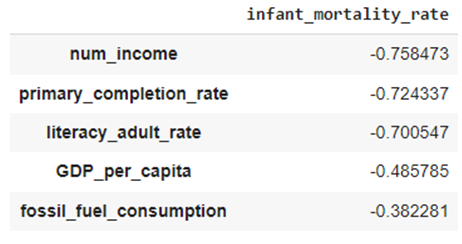
After taking into account the aforementioned characteristics evidenced in the EDA, an analysis of the data frames was carried out to find out what was generating these results. We identified one problem which had to do with the merging of the data frames and it was due to the fact that we had two different data sources: the WHO and the World Bank. For example, the WHO dataset has variables with mortality and data grouped by sex, by all genders and by age, as well as the standardized death rate. In contrast, in the World Bank datasets, the data is generated for all genders, without the previously mentioned groupings made by the WHO. Due to the differences and groupings in the dataset variables, once the data frames were joined, they showed inconsistencies and the models did not fit. For this reason, we decided to use a World Bank dataset on infant mortality. This was done with the hope to have the same data format and continue investigating the same dependent variable, only with the data limited or filtered to a specific age group. In addition, when the importance of the mortality variable for the countries is analyzed as an indicator of health and progress, it is the data on infant mortality that carry most weight. In accordance with the decision made, we went on cleaning up the infant mortality dataset and then, we merged the tables and came out with a final dataframe.

With the new data frame, we tested how the independent variables education and economy are correlated with the new dependent variable mortality in infants. The following correlations were obtained:

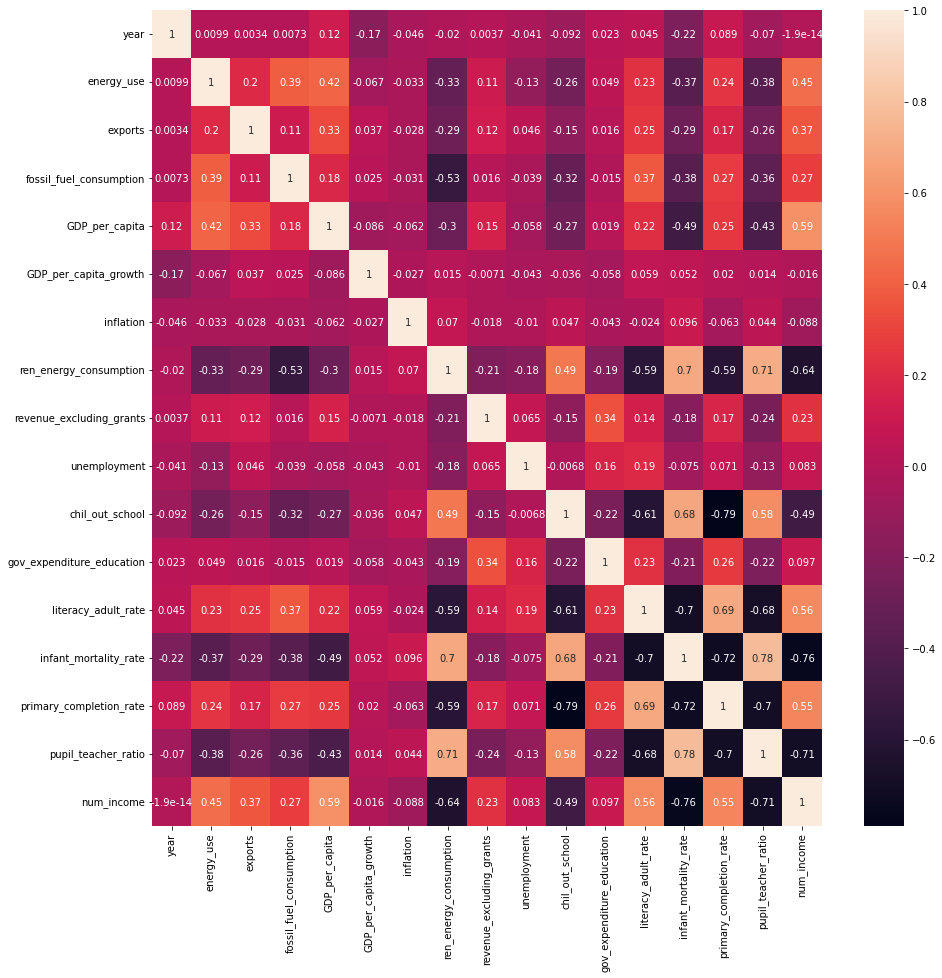
* Positive correlations:



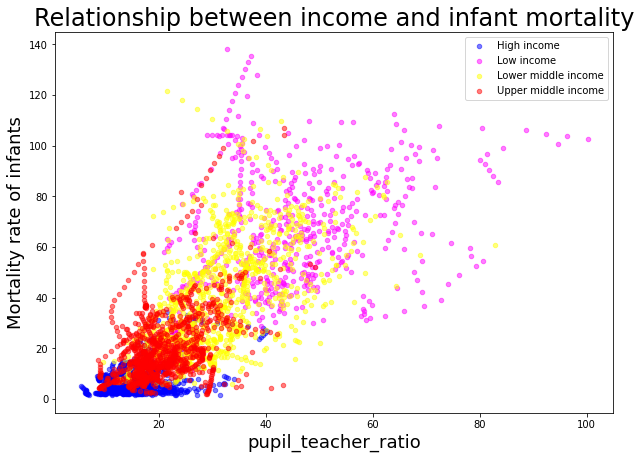
* Negative Correlations:



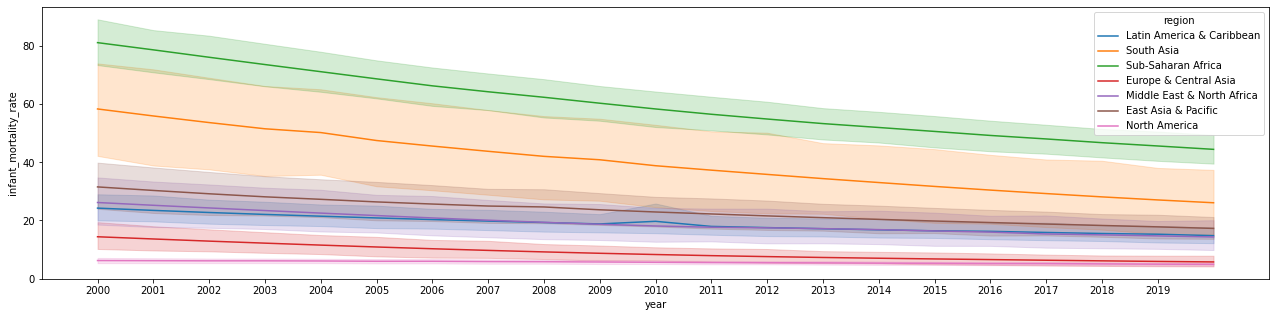
* Heat Map of the correlations found:



The heat map shows that the positive and negative correlations found are located within the independent and dependent variables.



This graphic shows the relationship between the income of the countries and the infant mortality rate. We may observe that high and upper middle income countries have a low mortality rate, while low and lower middle income countries have high mortality rates.



This linear diagram shows a general trend towards the decreasing of the mortality rate through time for all regions. Sub Saharan Africa and the south of Asia, show higher mortality rates than the other regions.

# 

# Method

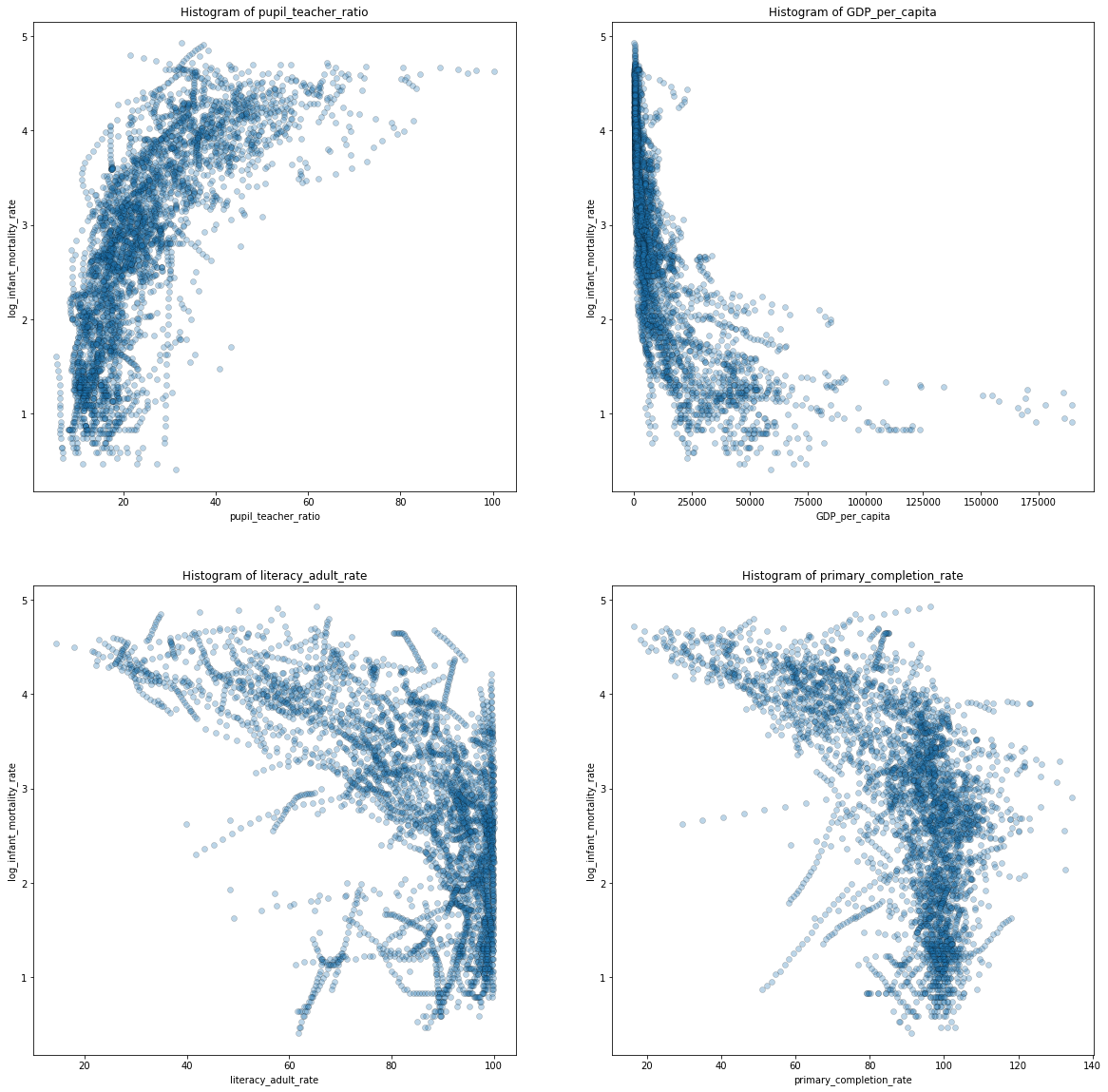
According to the previously presented date, we used the linear regression in order to model the variable that obtained the highest EDA, in this case, the pupil\_ teacher\_ratio, whose results were: R-squared:0.609

As we introduced to the model other variables and characteristics, the R2 increased. The following variables were added: num\_income, GDP\_per\_capita, infant\_mortality\_rate, literacy\_adult\_rate y primary\_completion\_rate, and the results were: R2 0.72.

The combination of the variables in the regression model: 'infant\_mortality\_rate ~ pupil\_teacher\_ratio + primary\_completion\_rate + ren\_energy\_consumption + región + income', showed a better adjustment with an R2 of 0.757.

As we may observe the linear regression model went from an R2 of 0.60 to 0.75, showing a better adjustment.

Ahora bien, el modelo de regresión lineal con las variables de tasa de mortalidad infantil vs las regiones + ingresos presentaron un R2 de 0.695 y los p-value de las regiones Latin America & North America son mayores a 0.05.



# Conclusion

The linear regression models showed an increase in the adjustment as more variables were added to the calculation. We can also notice that after the initial mistake was corrected, we started to obtain results that were closer to what we initially expected, i.e., countries with more resources had less mortality rate than those with less economical power.

Furthermore, it is shown that the linear model is a good approximation to understanding how the variables behave. It is also true that there may be other more complex models such as the supported vector machine or clustering models based on linear models that would give greater precision and accuracy, however these models tend to be poorly interpretable and not the best when performing a descriptive analysis. The latter would be more suitable for predictive rather than descriptive projects.

As we progress through the course we will implement more algorithms to better understand the relationships of our variables and see how they interact to answer our initial question.