Predicting Human Development Index with Social Metrics

Leiny Adriano

Economics and Data Science Course

Department of Economics, Carnegie Mellon University

December 19, 2020

TABLE OF CONTENTS

Introduction

The Human Development Index (HDI) is a broad measure of a nation's developmental stage in comparison with all other nations. As such, it depends on a number of macrosocial and macroeconomic metrics that reflect the development of each country. In this project, we are focusing on the social aspect of the HDI. For this reason, we have chosen the Gini Index, the World Happiness Report, and the Corruption Perception Index to compare and contrast them with the HDI. Based on our analysis, we intend to address the question of whether it is possible to predict the Human Development Index by using social metrics.

Motivation

It is often believed that a nation's development is only measure by their economic output. For example, using their Gross Domestic Product (GDP) or Gross National Income (GNI). However, to measure development we also need to take into account the social component within each country. This is why the Human Development Index (HDI) was proposed by the economist Mahbub ul Haq[1] and later adopted by the United Nations Development Program. The HDI offers a comprehensive view of not only the economic aspect but also takes into consideration social metrics. As the UN expresses it, "the HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone"[2]. The index accomplishes this by incorporating the life expectancy index, education index, and GNI per capita for each country into a single metric. However, each of the individual indexes that compose the HDI have other underlying factors that influence them, including other social factors that can be summarized with metrics. For example, the Gini Coefficient measures the statistical dispersion of income inequality by using income levels[3]. Also, the World Happiness Report is a global survey that measures the perceived happiness of a country's citizens[4]. Lastly, the Corruption Perception Index scores countries on how corrupt a country's public sector is perceived to be[5]. We intend to explain HDI through a social lens by taking into account income

inequality, happiness, and governmental corruption. All of which are variables that may explain or influence the capabilities of a country's citizens. Thus, through our statistical analysis and models we will find if there is a quantitative basis that supports our hypothesis and motivation.

Methodology

Methodological Approach

For this research project, we used statistical methods to explore the relationship between HDI and other social metrics (i.e. Gini Index, Happiness Index, Corruption Perception Index). We obtained the data by accessing public repositories and applying data cleaning methods to obtain a dataset that was relevant to our research. Then, using R, we evaluated each individual variable and assessed the statistical relationships between all the quantitative variables. Finally, we used our findings to design a multiple linear regression model that predicted HDI by using a combination of the Gini Index, Happiness Index, and Corruption Perception Index.

Data Collection

As mentioned before, we obtain the data for this research project by accessing public data repositories available through websites such as GitHub and Kaggle. In this case, we used four datasets each corresponding to an index. We used the `covid-19-data`[6] dataset from OurWorldInData because it includes the HDI for most countries. For the Gini Coefficient we used the `gini`[7] dataset from GapMinder which uses the index as measured in 2016. For the Happiness Index we used the `happiness`[8] dataset from the World Happiness Report which uses the index as measured in 2019. For the Corruption Perception Index we used the `cpi`[9] dataset from Transparency International which uses the index as measured in 2019. For reproducibility purposes, we upload all the datasets into a single public repository on GitHub which hosts the code and data used for this research project[10]. Then, we used R to merge all datasets and

create a data frame with the variables that were relevant for our research project. Lastly, we performed a thorough data cleaning process for our data to be suitable for analysis and model building.

Analysis Method

For the analysis, we first performed an EDA on the quantitative variables (i.e. HDI, Gini Index, Happiness Index, and Corruption Perception Index). We looked at their distributions and explored their relationship with HDI. Also, we measured the correlations for all the quantitative variables to select those that were relevant for a linear regression model. Then, given that all variables were viable for a model, we created linear regression models for each individual variable and all of their combinations. Finally, we selected the model that performed the best based on the statistics it produced (e.g. coefficients, p-values) and diagnostic plots.

Exploratory Data Analysis

Our dataset has six columns and 189 rows. The variables it includes are continent, country, hmn_dev_index, gini_index, happiness_index, and corruption_index. The only two categorical variables are continent and country, while the rest are continuous.

Human Development Index

As it may be observed below on Figure 1, the hmn_dev_index variable has mean 70.86 with minimum at 35.40 and maximum at 95.30. The 1st quantile is at 59.00 and the 3rd quantile is at 81.60.

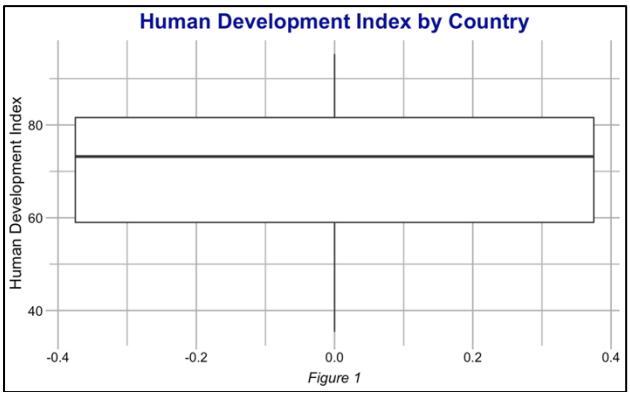


Figure 1

Below on Figure 2, we see the distribution for the hmn_dev_index variable. We observe that it follows a bimodal distribution, there are no outliers, and most countries have scores around 70 and 80. In this case, higher scores are better.

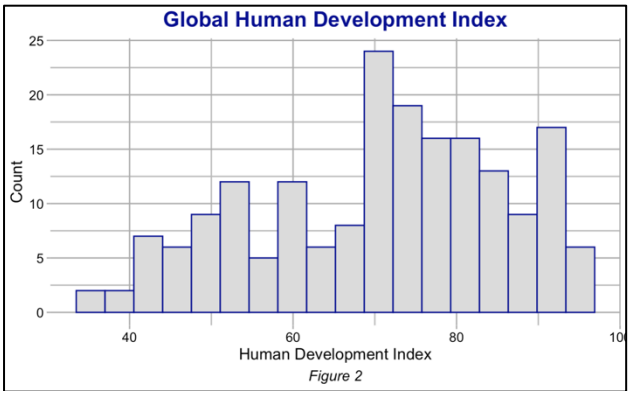


Figure 2

Gini Index

As it may be observed below on Figure 3, the gini_index variable has mean 38.91 with minimum at 25.00 and maximum at 63.00. The 1st quantile is at 33.70 and the 3rd quantile is at 42.60. The plot shows that there are some outliers above 55.00.

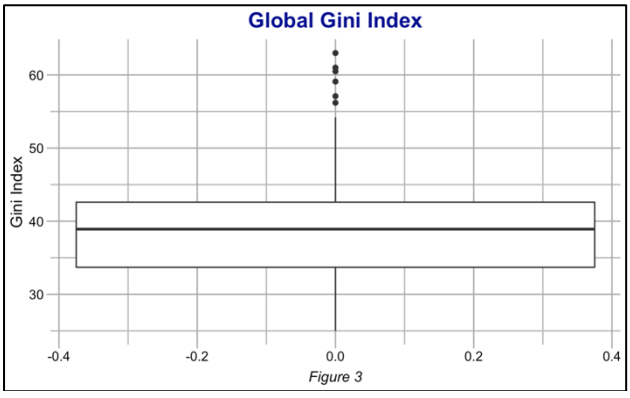


Figure 3

Below on Figure 4, we see the distribution for gini_index variable. We observe that it follows a unimodal distribution and it is skewed to the right. Because the distribution was smoothed there are no outliers. Also, most countries have scores around 36 and 42. In this case, lower scores are better.

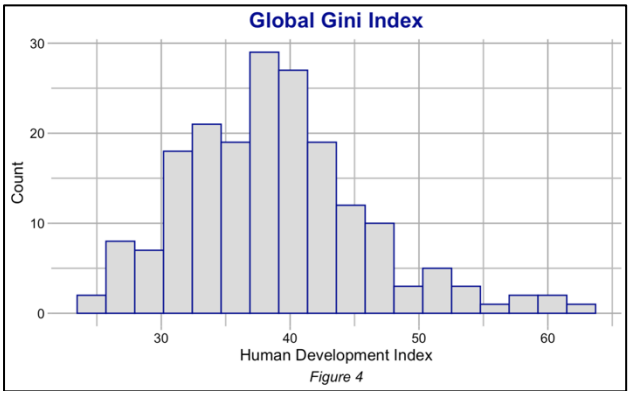


Figure 4

In Figure 5, we have a scatterplot that shows HDI on the x-axis and Gini Index on the y-axis. The observations are spread across the plot but follow a weak negative linear trend. This is highlighted by a line that depicts a linear regression for the two variables. In addition, we see that the countries are organized in clusters for their respective continents. Moreover, African countries have higher Gini coefficients and lower HDI scores, while European countries have lower Gini coefficients and higher HDI scores.

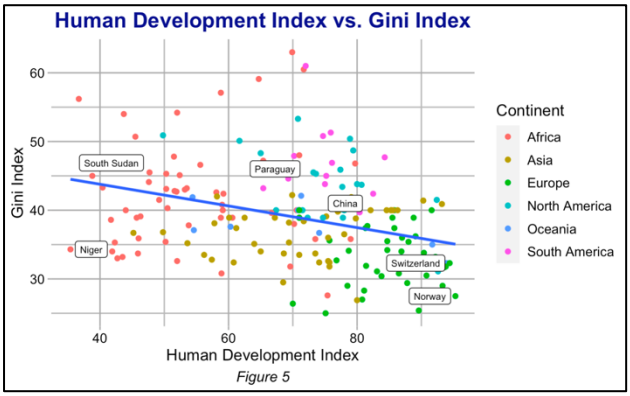


Figure 5

Happiness Index

As it may be observed below on Figure 6, the happiness_index variable has mean 5.418 with minimum at 2.853 and maximum at 7.769. The 1st quantile is at 4.722 and the 3rd quantile is at 6.028. The plot shows that there are no outliers.

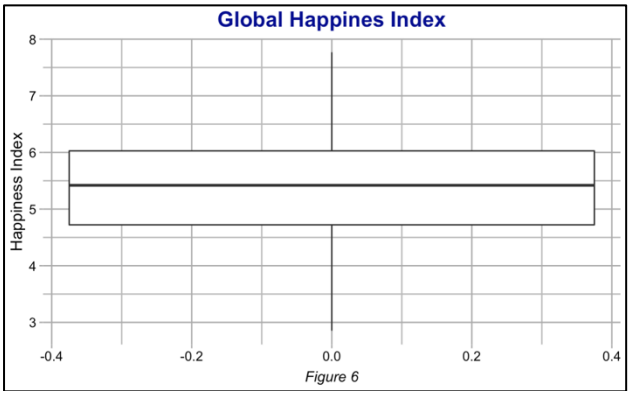


Figure 6

Below on Figure 7, we see the distribution for gini_index variable. We observe that it follows a unimodal distribution with a very high mode at around 5.5. Other than the mode, the distribution is relatively smooth with extreme score being less common. In this case, higher scores are better.

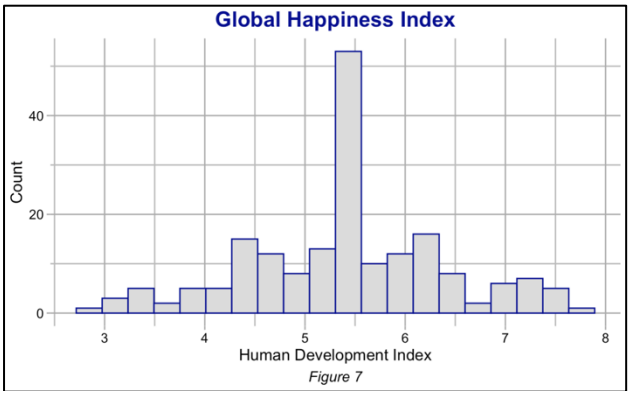


Figure 7

In Figure 8, we have a scatterplot that shows HDI on the x-axis and Happiness Index on the y-axis. The observations follow a strong positive linear trend. This is highlighted by a line that depicts a linear regression for the two variables. In this plot, countries do not form clear clusters based on the continents they belong to. However, most European countries have the highest combinations of HDI scores and Happiness Index scores.

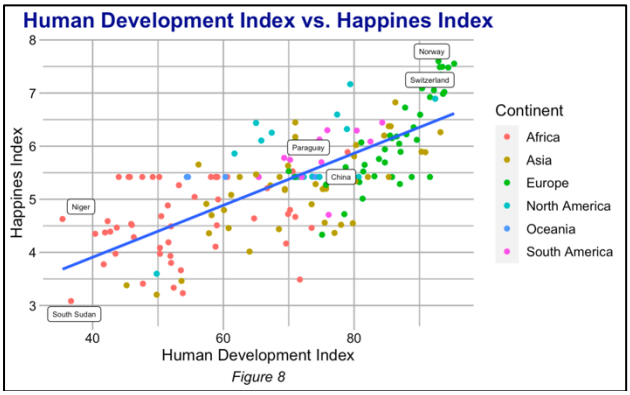


Figure 8

Corruption Perception Index

As it may be observed below on Figure 9, the happiness_index variable has mean 43.04 with minimum at 10.00 and maximum at 90.00. The 1st quantile is at 30.00 and the 3rd quantile is at 52.00. The plot shows that there are some outliers above 85.

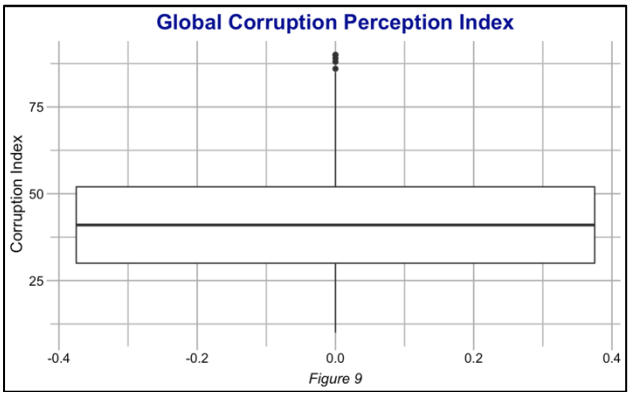


Figure 9

Below on Figure 10, we see the distribution for gini_index variable. We observe that it follows a bimodal distribution with modes at 40 and 55. In this case, higher scores are better. Because the distribution was smoothed there are no outliers. Finally, most countries have scores between 25 and 40. In this case, higher scores are better

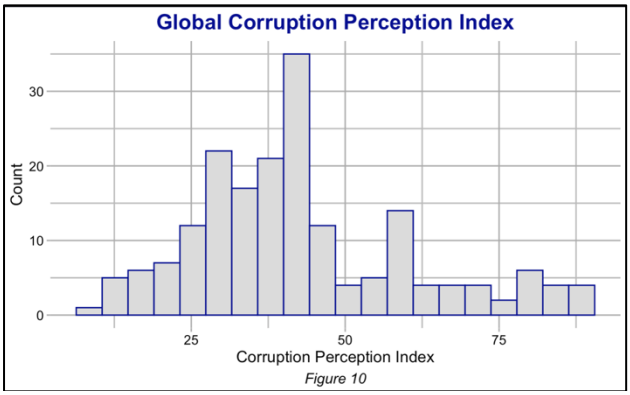


Figure 10

In Figure 11, we have a scatterplot that shows HDI on the x-axis and Corruption Perception Index on the y-axis. The observations follow a positive linear trend. This is highlighted by a line that depicts a linear regression for the two variables. In this plot, countries do not form clear clusters based on the continents they belong to. However, most European countries have the highest combinations of HDI scores and Happiness Index scores.

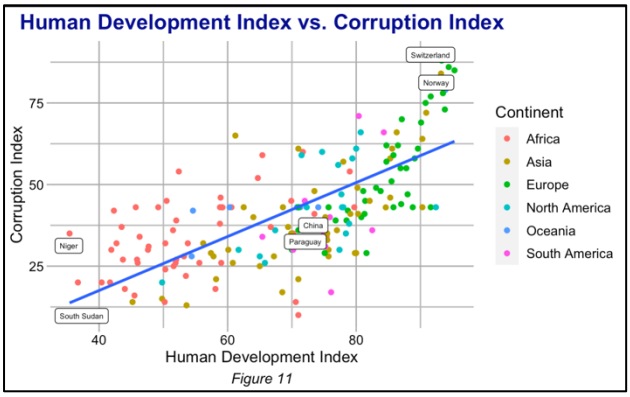


Figure 11

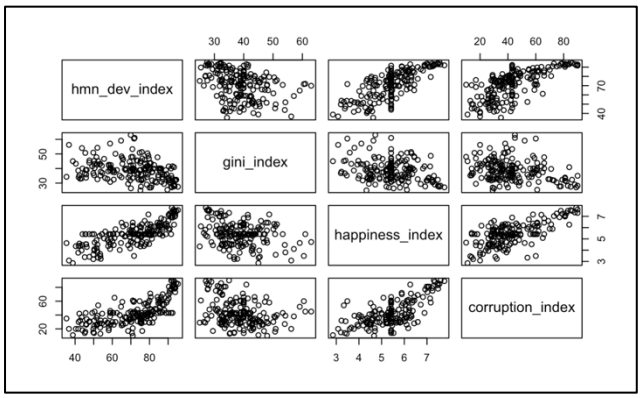
Correlation

Here, we see a correlation matrix for all the quantitative variables in the dataset. The correlation coefficients support the observations we made for all the scatterplots. HDI is weakly negatively correlated with the Gini Index, while it is strongly positively correlated with the Happiness Index and the Corruption Perception Index. However, we see that there is also a strong correlation or interaction between the Happiness Index and the Corruption Perception Index. For this reason, it might be best to not include

them both on the same model. However, we will further explore this with the diagnostic plots for the linear regression models we create.

```
##                  hmn_dev_index gini_index happiness_index corruption_index
## hmn_dev_index             1.00      -0.33            0.75             0.70
## gini_index               -0.33       1.00           -0.30            -0.29
## happiness_index           0.75      -0.30            1.00             0.69
## corruption_index          0.70      -0.29            0.69             1.00
```

Furthermore, on Figure 12 we display the plot pairs for all the quantitative variables in our dataset. This visualizations supports the observations we made with the correlation matrix.



Linear Regression Model

Human Development Index = 22.4339 + (-0.2613) * *Gini Index* + 10.8132 * *Happiness Index*

Results

Based on the summary statistics and diagnostic plots, we chose a multiple linear regression model that used the Gini Index and the Happiness Index to predict the HDI. The table below offers a summary of the coefficients and statistics associated with the linear model. The model's intercept is 22.4339 and is the expected HDI score when using the mean value of all Gini Index coefficients and Happiness Index scores. The coefficient for the Gini Index is -0.261 which means that we would expect a country's HDI score to

increase by 1 for each |-0.261| change in its Gini Index coefficient holding everything else constant. This coefficient is significant for our linear model because it has a p-value of 0.00102. The coefficient for the Happiness Index is 10.813 which means that we would expect a country's HDI score to increase by 1 for each 10.813 change in its Happiness Index score holding everything else constant. This coefficient is significant for our linear model because it has a p-value of $< 2^{-16}$. In addition, the model's residual standard error is 9.982 on 186 degrees of freedom which is the amount by which the HDI score can deviate from the true regression line, on average. Hence, the linear model's percentage error is 44.50%. The model also has an adjusted R-squared of 0.5659. Lastly, the linear produces an F-statistic with coefficient 123.5 and p-value $< 2^{-16}$.

```
##
## Descriptive statistics
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             hmn_dev_index
## ------------------------------------------------
## gini_index                    -0.261**
##                                (0.107)
##
## happiness_index               10.813***
##                                (0.767)
##
## Constant                      22.434***
##                                (6.723)
##
## ------------------------------------------------
## Observations                     189
## R2                              0.570
## Adjusted R2                     0.566
## Residual Std. Error       9.982 (df = 186)
## F Statistic           123.518*** (df = 2; 186)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Moreover, below we have the diagnostic plots for our linear model. First, on Figure 13, we have plot that depicts the residuals versus the fitted values. In the plot we observe that the observations are randomly distributed across the space and the highlighted line is roughly center at 0. However, it slightly deviate from 0 at 60 and 80 on the x-axis. Next, on Figure 14, we have the Normal Q-Q plot that visualizes the theoretical quantiles versus the standard residuals. Here we observe that the residual points roughly follow the dashed line. Although, they deviate from the dashed line at the lower extreme.

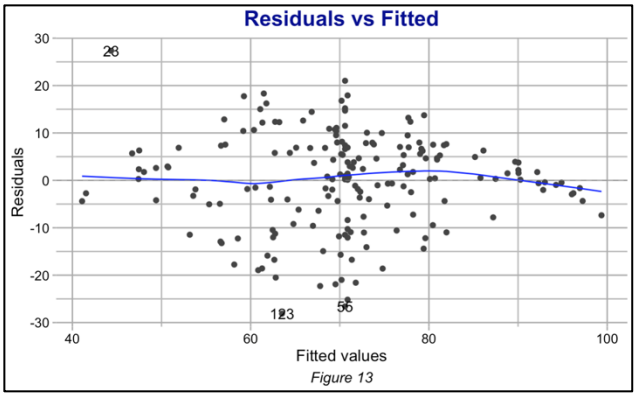


Figure 13

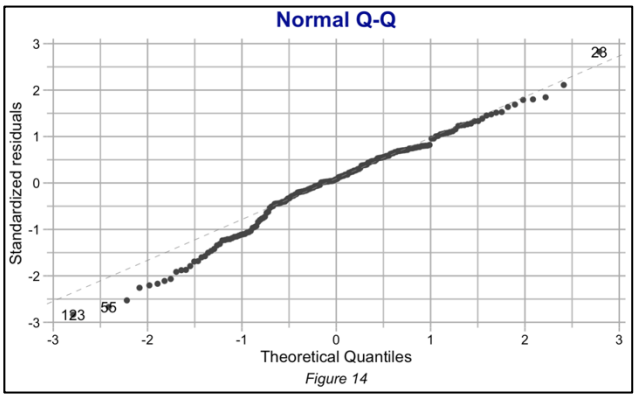


Figure 14

On Figure 15, we have the Scale-Location plot which visualizes the fitted values versus the square root of the standard residuals. In this case, the highlighted line is not horizontal and curves downwards at the extremes. This suggests a heteroscedasticity problem (i.e. non-constant variance in residual errors) with our model. Lastly, Figure 16 is the Residuals vs Leverage plot for our model. It plots the leverage versus the standard residuals. The plot shows that there outliers or high leverage points (i.e.) that may be influencing our linear model.

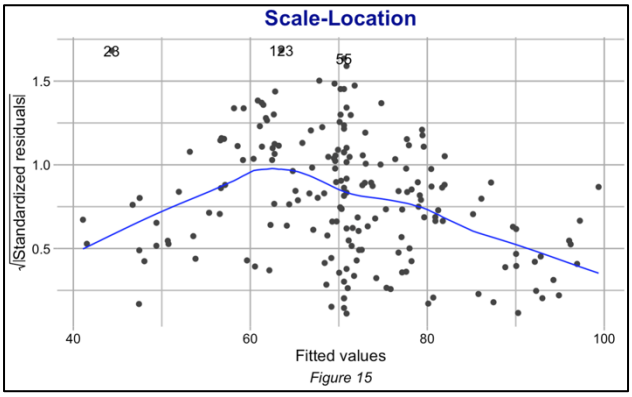


Figure 15

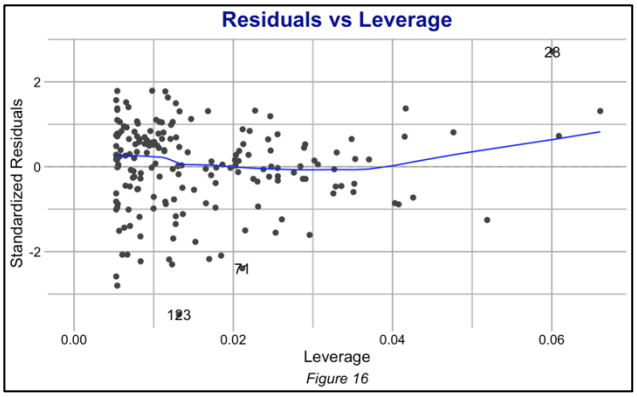


Figure 16

Conclusion

The purpose of our research project was to explore the statistical relationships between a country's Human Development Index (HDI) and a series of social metrics (i.e. Gini Index, Happiness Index, and Corruption Perception Index). We observed that all three of our predictor variables had linear relationships with HDI. Although, the ones with the strongest correlations were the Happiness Index and Corruption Perception Index, which were also mutually correlated. Based on our statistical analysis we were able to create a multiple linear regression model to predict a country's Human Development Index by using its Gini Coefficient and Happiness Index score. The diagnostic we ran on the model suggest that there are underlying issues with our predictor variables that affect the accuracy of the linear model. Mainly, the issues are related to heteroscedasticity and outliers or high leverage points. Of course, we must take into consideration that our predictor variables are social metrics that summarize particular socioeconomic phenomena for each country. As such, there are a wealth of factors that the model does not control for but that influence it. In future iterations of this research project we could gather data of more social metrics to have a broad sample and be able to narrow or amplify their statistical effects on our model. Perhaps, this social metrics could have a narrower scope such as Press Freedom Index to avoid introducing too much variability. Nonetheless, the main takeaway of our project is that it is possible to predict to some degree the Human Development Index by using social metrics. Therefore, based on our analysis, we conclude that it is possible to focus on social aspects to predict a country's HDI

Bibliography

[1]India Times. (n.d.). *The Economic Times*. Retrieved December 2020, from Definition of 'Human Development Index: https://economictimes.indiatimes.com/definition/human-development-index#:~:text=Description%3A%20Pakistani%20economist%20Mahbub%20ul,Nations%20Deve Develop%20Program%20(UNDP).

[2]United Nations Development Programme. (2020). *Human Development Reports*. Retrieved December 2020, from Human Development Index: http://hdr.undp.org/en/content/human-development-index-hdi

[3]Wikipedia. (n.d.). *Wikipedia, the free encyclopedia*. Retrieved December 2020, from Gini coefficient: https://en.wikipedia.org/wiki/Gini_coefficient

[4]World Happiness Report. (2020). Retrieved 2020 December, from World Happiness Report 2020: https://worldhappiness.report/ed/2020/

[6]Our World in Data. (2020, February). *GitHub*. Retrieved 2020 December, from covid-19-data: https://github.com/owid/covid-19-data

[7]GapMinder. (2020, April 11). *Kaggle*. Retrieved 2020 December, from Income Inequality: https://www.kaggle.com/psterk/income-inequality/metadata

[9]Transparency Internation. (2017, January 26). *Kaggle*. Retrieved December 2020, from Corruption Perceptions Index: https://www.kaggle.com/transparencyint/corruption-index/metadata

[10]Adriano, L. (2020, December 18). *GitHub*. Retrieved from Human Development by Social Metrics: https://github.com/ladrian0/human-development-by-social-metrics