Final Project: Predicting Happiness

Dustin Robinson, Princess Gonzalodo, Dhara Khubchandani, Pramod Mantya Raju

Profesor Guillaume Faddoul ISYS 812-01

May 26, 2023

Table of Contents

Introduction

The team utilized three datasets to explore the potential factors influencing the happiness score of a country. Happiness is a multifaceted concept encompassing various aspects to identify the factors contributing to a higher quality of life. Understanding happiness allows researchers and policymakers to implement measures that improve individual happiness. The newfound policies may include higher minimum income thresholds, equitable disbursement of income, freedom of choice, and other variables that could increase happiness, efficiency, and productivity within a country's borders. The team referenced various World Bank and World Happiness Report organization datasets. However, despite having the validated datasets available, the team opted to utilize datasets provided by Kaggle, drastically limiting the project's significance. Nevertheless, the validated data is available from reputable sources, and further research can utilize the accurate data, among other variables, alleviating some of the mentioned limitations.

World Happiness

The first dataset the team used was World Happiness Report up to 2022 (see Appendix, Figure 1), which has 157 rows and columns titled "Overall rank," "Country or region," "Score," "GDP per capita," "Social support," "Healthy life expectancy," "Freedom to make life choices," "Generosity," and "Perceptions of corruption ." Overall rank is the ranking of the country based on happiness score. The "Country or region" describes the name of the country or region in the evaluation. "Score" represents the happiness score designated to the country or region. "GDP per capita" divides the gross domestic product by the population. "Social support" measures individuals' emotional perception towards their availability for assistance from external sources such as family, friends, and community. "Health life expectancy" has two explanations. From 2015-2019 the variable explains how a person's life expectancy contributes to happiness. In 2020 and 2021, the World Happiness Report shifted the metric to measure the average years individuals are expected to live without limitations like illness or disabilities. Furthermore, "Freedom to make life choices" is the extent to which individuals have the personal freedom to make decisions about their lives without constraints or restrictions. Moreover, "Generosity" is the willingness of individuals to participate in acts of kindness or compassion - generally measuring the society's willingness to share intangible resources. Lastly, "Perceptions of Corruption" measures how individuals perceive societal corruption.

GINI

Another dataset we used was the Gini Index per country (see Appendix, Figure 2). The GINI index measures income inequality within a country on a scale from zero to one.

The closer a country is to zero, the less income inequality it has, and vice versa. The dataset has only four columns: country_code, country_name, year, and value, representing the country's abbreviated name, the actual country name, the year the data was collected and analyzed, and the GINI index value, respectively. The dataset provides information about where each country falls on this scale, which allows us to examine income disparities. We play around with the concept of "can money buy happiness," as inequality can influence economic disparities and less social mobility.

GDP

The GDP by Country dataset contains the Gross Domestic Product per country (see Appendix, Figure 3). GDP is a measure of a country's economic performance based on the total value of goods and services produced within the borders of that specific country. The dataset was selected based on the belief that it could help explore potential relationships between GDP and happiness score, given that increased GDP might contribute to "happiness" such as access to resources. Initially, the dataset consisted of 181 rows, alphabetically representing the country's name. The columns are the years for which GDP values were derived, ranging from 1999 to 2022.

Understanding the factors that contribute to a country's happiness score can enrich the lives of people within that country. Moreover, by gaining a deeper understanding of these factors, governments can shape policies and allocate resources more effectively by aligning their goals, wages, and laws with crucial determinants of happiness. Increasing a country's happiness score ultimately means enhanced quality of life and increased satisfaction.

The Driving Question

What factors affect the happiness score of a country?

Data Cleaning

The team initially thought to merge all the data sets but rapidly realized it would require more intricacy and attention to detail. The most significant issue began in the dataset provided by Kaggle for the World Happiness Report. The column titles were primarily the same for the first three years of the World Happiness Report (2015 - 2017). The second year introduced the confidence interval for the happiness score, but in 2018, the Dystopia Residual, a measurement of a country's distance from dystopia, was removed as a value. In the two years following the first three years (2018 - 2019), the columns matched that grouping but not the first three years. The final two years of the World Happiness Report (2020 - 2021) also differed from all prior years. For example, the life expectancy score changed from measuring how life expectancy explains happiness to an average life expectancy in years. Despite many containing the same information, the mismatched columns created an insurmountable amount of null values. For example, the column titled 'Trust (Government corruption)' became 'Perceptions of corruption' but represented the same information for the category, despite the name change.

Furthermore, the World Happiness data was downloaded as a csv file for each reporting year, a total of seven csv files. Instead of merging or concatenating the mismatched columns, the team opted to subset the years of the other datasets to keep merging and the risk of data loss at adequate levels to conduct data analysis in the latter portions of the project. For example, the GINI Index and GDP datasets were separated to match the years of the World Happiness Report, meaning for the years 2015 - 2021, we had seven subsets for the GINI and GDP datasets.

The most accessible column for people to recognize was the country name column. Therefore, the team selected the country name as the index column for reindexing at the final data frame. Next, the team pressed forward with the merging process using the identifying country name column. The first datasets to start the merging process were the World Happiness Report and the GDP datasets. The datasets were merged for each subsetted year (2015 - 2021). Despite the column naming issues identified previously, the team kept the column names of the World Happiness Report intact for this merging section.

The team experienced no significant data loss with the first merge and opted to take the World Happiness and GDP combined data through another merging interval. The final merge required blending the previously combined datasets with the GINI Index. Because the combined and GINI datasets were sectioned into years, the team followed the project's standard protocol of merging the subsetted years, requiring seven more merges. The team was left with seven subsets, each containing the World Happiness Report, the GDP, and the GINI Index for the respective years. Still, the team needed to

address the column names within the World Happiness Report to create a final data frame satisfactory for data analysis.

The complete data subset represents the year (between 2015 and 2021), the World Happiness Report for the year, the GDP for the year, and the GINI score for the year, identified and indexed by the country name. Through collaboration, the team identified the clearest titles for each column and named them accordingly. The team utilized the matching titles method to preserve the most data. Each complete data subset required the World Happiness columns to be identified, renamed, and sometimes dropped.

Finally, with the column names matched, the final data frame underwent its last transformation process. The complete data subsets for each year were finally concatenated together. Still, the team needed to address the GDP issue; GDP was represented in string format—the values needed to be stripped of the commas, currency symbols, and periods. The team applied a replace function on the string symbols causing issues, converted the GDP values to integers, and finalized the data frame by setting the index to country (see Appendix, Figure 4). However, the 2020 and 2021 subsets left the data minimized to nineteen total values. The limited values would not allow appropriate data analysis, leaving the team to conclude the years could not be used for the final dataset. The team speculates the COVID pandemic left data severely underreported for those years. Nevertheless, the years could be held for other analysis, under the assumption that the data is not statistically significant and would require further research.

Data Models & Analyses

What factors affect the happiness score of a country?

The Ordinary Least Squares (OLS) regression model allowed the team to examine the relationship among the dataset variables and analyze the indication of their coefficients. First, the team isolated a subset of columns from the “final_df” data frame to use as the independent variables (features): “GDP_per_Capita, Generosity, GINI Coefficient, GDP, Family, Life Expectancy, and Freedom.” Next, the model isolated our dependent variable (target), “Happiness_Score,” and added a constant for estimating an intercept within the model. Afterward, the team specified the OLS model using the parameters “sm.OLS(target, features),” fitting the model to the dataset while storing the results. Finally, a summary showing the statistical measures, regression analysis, parameters with names, and p and t values are displayed.

The team leveraged our combined understanding of statistical analysis to analyze the model. The uncentered R-squared shows the proportion of the “Happiness Score” can be explained by the independent variables, with the best model resulting in GINI Coefficient and GDP per capita as the independent variables. The R-square shows that the independent variables explain 98.4% of happiness, suggesting a strong correlation. Other measures like the F-statistic of 1.015e+04 suggests that the model is highly significant. Furthermore, the team looked at the p-values of each independent variable. The p-values suggest the significance of each variable; the dependent variables were significant within the model with 95% confidence, meaning the variables accurately predicted the happiness score (See Appendix, Figure 5).

Our regression analysis suggests that GDP per capita and the GINI coefficient significantly predict a country’s Happiness Score. For example, an increase in GDP per capita will increase the happiness score in the respective country, indicated by its positive coefficient (3.35). Similarly, an increase in the GINI coefficient will lead to an increase in a country’s Happiness Score because of the variable’s positive coefficient (.06). The team speculates a higher GDP per capita is associated with greater levels of income inequality (indicated by GINI Coefficient) because the GDP in the respective country is higher, suggesting the country may be economically developed leading to higher income inequality. Moreover, GDP per Capita may not accurately assess individual income, and further research could include Gross National Income per Capita (GNI per Capita). Instead of measuring the disbursement of a country’s revenue over its population, the GNI per capita measures the income generated by a country and divides it evenly over its respective population. Accurately assessing income can signal policymakers to adjust laws regarding minimum incomes, leading to an increase in happiness and standard of living for the global population.

Within the model, which of the X values has the strongest correlation with happiness score?

The Pandas data manipulation library calculated the correlation between the variables in our data frame. Furthermore, the team created a heatmap for the correlation to visualize the information (see Appendix, Figures 6 and 7). On a heatmap, a lighter color indicates a lower correlation, whereas a darker one suggests a higher correlation. From the heatmap, GDP per capita shows the strongest positive correlation with a country's Happiness Score. Surprisingly, GDP has the lowest correlation with a country's happiness. The team infers that an individual's happiness is appropriately analyzed by a disbursement of the country's revenue over its population rather than the earnings of the entire nation.

Do countries with higher GDPs have higher happiness scores?

The team created subplots and bar charts to analyze the GDP and Happiness data for different countries in 2017, 2018, and 2019 (see Appendix, Figure 8). First, the team executed a Python function call for ten subplots with a grid of 2 X 3. Then, the team added a loop for Python to run through each year and select the years in the list so that the top 5 years with the highest GDP are chosen for the specific year using the condition final_df['Year']== year. The data is then sorted and stored in descending order based on the top 5 selected from the GDP column using head(5).

Inside the loop, the ax[0, i].bar() function is used to create a bar chart on the first row of the subplot grid. The x-axis of the bar chart represents the countries' index from the top_gdp data, and the y-axis represents their corresponding GDP values. Finally, the ax[0, i].set_title() function sets the title of the subplot with the year appended to 'Top GDP countries."

The second loop follows a similar structure but focuses on the Happiness Score, selecting the top 5 countries with the highest happiness score each year. Finally, the team created a bar chart for each year in the subplot grid's second row (ax[1, i]). The ax[1, i].set_ylim([6,8]) function sets the y-axis limit to a specific range of 6 to 8 to ensure consistent scaling across the subplots.

Next, the plt.subplots_adjust(hspace = 0.5, wspace = 0.2) function adjusts the spacing between the subplots to make them visually appealing and easier to interpret. Finally, the plt.show() function displays the resulting plot. From the subplot, the team concluded that the top 5 GDP countries are not the ones who are happy, and vice versa; the happiest countries do not have the highest GDPs.

What predicts a country’s distance from dystopia?

The team also sought to predict a country's distance from dystopia, a hypothetical country with a happiness score of 1.85. As the dependent variable, "Dystopia Residual" measures how far a nation is from becoming dystopian. The method used to create this model was a forward stepwise regression, a statistical method that isolates the significant predictors from all potential independent variables (see Appendix, Figure 9). Unfortunately, the complete subsets for 2020 and 2021 endured substantial data loss, and the countries left to analyze were below the minimum threshold of twenty observations. Further research requires a comprehensive dataset for conducting an in-depth analysis. Nevertheless, the team continued seeking valuable insights or identifying patterns and trends within the data.

The subset used for the model includes numerous socioeconomic and well-being metrics for nations worldwide. A carefully chosen set of attributes from this dataset were picked as potential indicators of a nation's separation from a dystopian society, including indicators of generosity, happiness, the GINI coefficient (a measure of income distribution within a country), GDP (Gross Domestic Product), GDP per capita, family support, life expectancy, freedom, and perceptions of corruption. The features were isolated and kept in a variables section during selection. In addition, the goal variable, "Dystopia Residual," was isolated from the feature variables.

Next, a stepwise regression was used to assess various feature combinations. After starting with an initial group of features, the algorithm added or eliminated characteristics based on how much they improved the model's predictive ability. The models were assessed using the adjusted R-squared score, which considered the quality of fit and the number of features within the model. The model's summary contained statistical data such as coefficients, standard errors, t-statistics, p-values, and other pertinent measurements, making it efficient and effective to evaluate each feature's importance and effect on a country's distance from dystopia. The goal was to pinpoint the subset of characteristics that most accurately predicted a nation's distance from a dystopian future. The best model offers insights into the association between the chosen features and the distance from "Dystopia."

According to the model, the chosen characteristics—Life Expectancy, Perceptions of Corruption, and GINI Coefficient—considerably influence forecasting the distance from "Dystopia." The model appears to account for roughly 99% of the variation in the predictive model based on these features, according to the R-squared value of 0.99.

Each variable's coefficients show the direction and magnitude of each feature's influence. For example, as life expectancy for a country increases, the country's distance from "Dystopia" also increases, as shown by the positive life expectancy coefficient of 0.0214. In retrospect, when a country's perceptions of corruption increase,

the distance from "Dystopia" decreases, according to the negative Perceptions of Corruption coefficient (-1.3118). The GINI Index coefficient is 0.0422, indicating that when income inequality increases, the distance from "Dystopia" trends in the same direction.

The p-values show the statistical significance of the coefficients. For example, the p-values for Life Expectancy, Perceptions of Corruption, and GINI Coefficient in this instance are, respectively, 0.0148, 0.0421, and 0.0000, indicating the three variables are statistically significant in predicting the distance from "Dystopia" with 95% confidence.

The regression model shows that a country's proximity to a dystopian state is influenced by better life expectancy, lower views of corruption, and greater economic inequality. The team speculates a longer life expectancy is associated with economically developed countries, suggesting the country may have a higher happiness score. Furthermore, income inequality is considered more significant in some developed nations, meaning the standard of living is higher, regardless of the income disparity. Moreover, the team hypothesizes that fears of corruption may add unnecessary stress to everyday lives causing greater unhappiness or less distance from the dystopian level of happiness. Still, further research can identify factors such as healthcare or other public resources that provide insights into why a country trends further from unhappiness than others. Identifying factors that may predict proximity to unhappiness can lead to research that develops methods to mitigate the issues causing a country to be closer to "Dystopia."

Has happiness increased overtime?

The team utilized a series of four graphs to scatter the happiness scores by country for each year (2015, 2016, 2017, 2018)(see Appendix, Figure 10). The graphs show the happiness trend by country for each year. If happiness increased over the years, the graph would trend higher than the previous year, and the opposite holds.

The four graphs appeared nearly identical, indicating the happiness score for a specific country by year was relatively stagnant. Still, an in-depth analysis indicates a slight decrease in overall happiness. Nevertheless, the team speculates that the World Happiness Survey respondents could have previously answered prior surveys, leading to biases or unchanged happiness measures. The team also suggests that happiness may only grow with drastic changes to a country's government, policies, or economy. In retrospect, countries with higher happiness scores may experience happiness immobility because the standard of living is considered maximized within the specific nation.

Conclusion

Despite the limitations of the project, the insights provided encourage the possibility of further research. The hypothetical idea of creating policies to generate greater happiness grounded in accurate and validated data may foster an equitable future for the global population. Policymakers can consider income and happiness when setting minimum wage laws. With further development of the ideas behind this project, governments can contract unbiased researchers to collect and mine verified and accurate data representing the entire population. Furthermore, further research can create insights and information to help legislators implement safer and more equitable decisions for the country's citizens.

The theoretical ideas within the regression models allow researchers to predict happiness and find models or variables that increase a country's distance from Dystopia (the fictitious unhappiest place on earth). Increasing distance from unhappiness does not mean happiness will increase, but it provides a pivotal perspective that cannot be overlooked. For example, life expectancy did not increase a country's happiness score but increased its distance from unhappiness. Governments may use that information to increase their budget for healthcare to promote longer lives and mitigate stress for their population. Further development of the two models can provide a valuable tool that helps promote opportunity, equity, and happiness globally.

References

Ache, M. (2022, March 19). *World Happiness Report up to 2022*. Kaggle. https://www.kaggle.com/datasets/mathurinache/world-happiness-report

Pedersen, U. T. (2023a, February 28). *Gini Index*. Kaggle. https://www.kaggle.com/datasets/ulrikthygepedersen/gini-index-per-country

Pedersen, U. T. (2023b, February 28). *Gross Domestic Product*. Kaggle. https://www.kaggle.com/datasets/ulrikthygepedersen/gdp-per-country

The World Bank. (n.d.). *GDP (Current US$)* | *Data - World Bank Data*. The World Bank. https://data.worldbank.org/indicator/SI.POV.GINI?most_recent_value_desc=true

The World Bank. (n.d.). *Gini Index* | *Data - World Bank Data*. The World Bank. https://data.worldbank.org/indicator/SI.POV.GINI?most_recent_value_desc=true

World Happiness Report. (n.d.). *The World Happiness Report*. Home | The World Happiness Report. https://worldhappiness.report/

Appendix

```
wh_2015 = pd.read_csv('2015.csv')
wh_2015.head()
```

10]:

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family
0	Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951
1	Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223
2	Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058
3	Norway	Western Europe	4	7.522	0.03880	1.45900	1.33095
4	Canada	North America	5	7.427	0.03553	1.32629	1.32261

```
wh_2016 = pd.read_csv('2016.csv')
wh_2016.head()
```

11]:

	Country	Region	Happiness Rank	Happiness Score	Lower Confidence Interval	Upper Confidence Interval	Economy (GDP per Capita)
0	Denmark	Western Europe	1	7.526	7.460	7.592	1.44178
1	Switzerland	Western Europe	2	7.509	7.428	7.590	1.52733
2	Iceland	Western Europe	3	7.501	7.333	7.669	1.42666
3	Norway	Western Europe	4	7.498	7.421	7.575	1.57744
4	Finland	Western Europe	5	7.413	7.351	7.475	1.40598

Figure 1: *Screenshot of World Happiness Index loaded into python.*

```
gini_df = pd.read_csv('gini_by_country.csv')
gini_df.head()
```

17]:

	country_code	country_name	year	value
0	AGO	Angola	2000	52.0
1	AGO	Angola	2008	42.7
2	AGO	Angola	2018	51.3
3	ALB	Albania	1996	27.0
4	ALB	Albania	2002	31.7

Figure 2: *Screenshot of Gini Coefficient loaded into Python.*

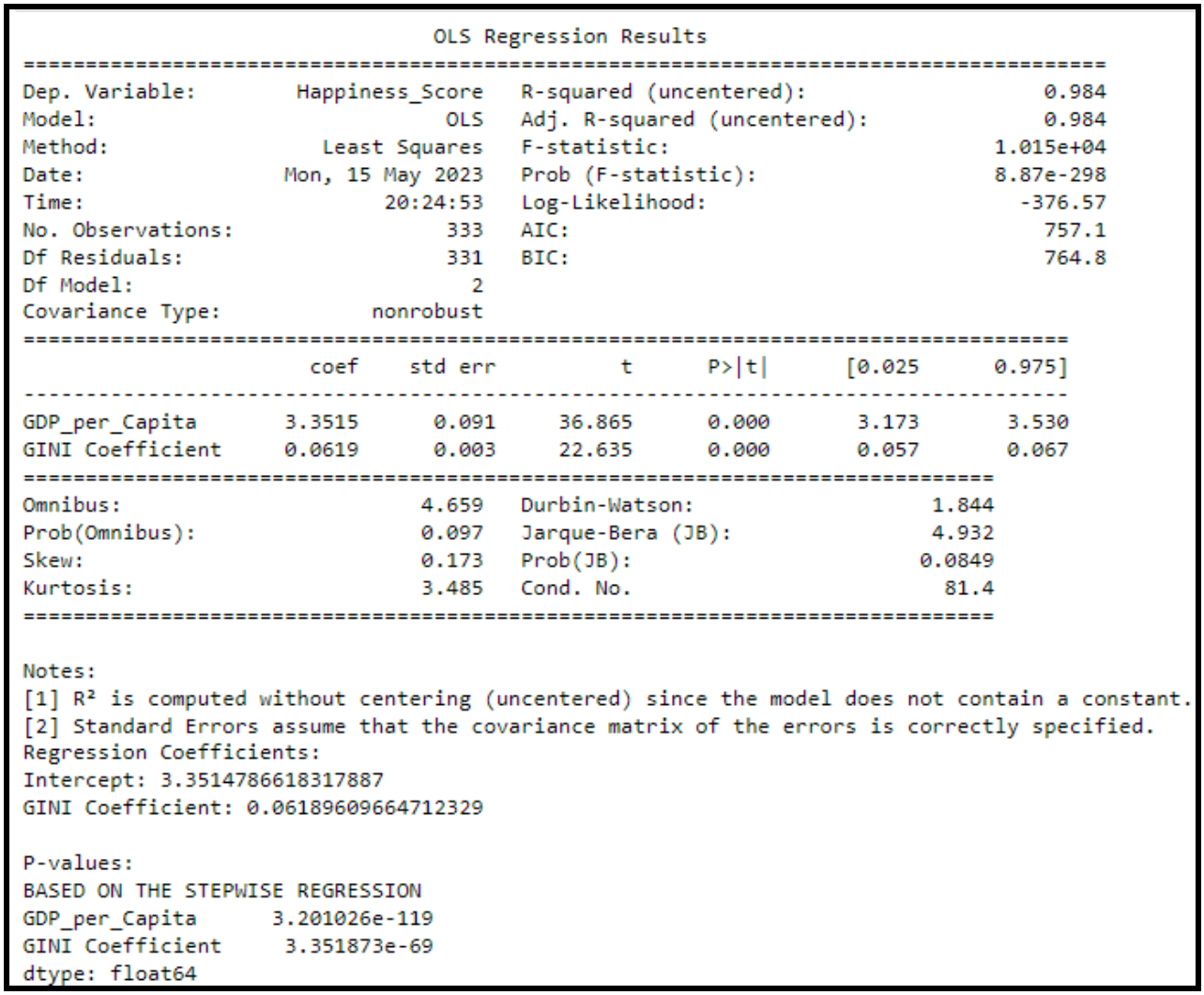
```
gdp_df = pd.read_csv('gdp_by_country.csv')
gdp_df.head()
```

	country_code	country_name	year	value
0	ABW	Aruba	1986	$1,676,310,877.06
1	ABW	Aruba	1987	$1,945,835,371.02
2	ABW	Aruba	1988	$2,308,707,372.64
3	ABW	Aruba	1989	$2,588,749,895.63
4	ABW	Aruba	1990	$2,691,300,678.70

Figure 3: *Screenshot of GDP dataset loaded into Python.*

Country	Year	GDP	Happiness_Score	GDP_per_Capita	Family	Life Expectancy	Freedom	Perceptions of Corruption	Generosity	GINI Coefficient
Albania	2015	1354642246300	4.959	0.87867	0.80434	0.81325	0.35733	0.06413	0.14272	32.8
Armenia	2015	4712024104100	4.350	0.76821	0.77711	0.72990	0.19847	0.03900	0.07855	32.4
Austria	2015	344269230000	7.200	1.33723	1.29704	0.89042	0.62433	0.18676	0.33088	30.5
Belgium	2015	416701400000	6.937	1.30782	1.28566	0.89667	0.58450	0.22540	0.22250	27.7
Benin	2015	6732814000000	3.340	0.28665	0.35386	0.31910	0.48450	0.08010	0.18260	47.8

Figure 4: *Screenshot of cleaned dataset to begin analysis.*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Happiness_Score   R-squared (uncentered):                   0.984
Model:                            OLS   Adj. R-squared (uncentered):              0.984
Method:                 Least Squares   F-statistic:                          1.015e+04
Date:                Mon, 15 May 2023   Prob (F-statistic):                    8.87e-298
Time:                        20:24:53   Log-Likelihood:                         -376.57
No. Observations:                 333   AIC:                                      757.1
Df Residuals:                     331   BIC:                                      764.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
GDP_per_Capita     3.3515      0.091     36.865      0.000       3.173       3.530
GINI Coefficient   0.0619      0.003     22.635      0.000       0.057       0.067
==============================================================================
Omnibus:                        4.659   Durbin-Watson:                   1.844
Prob(Omnibus):                  0.097   Jarque-Bera (JB):                4.932
Skew:                           0.173   Prob(JB):                       0.0849
Kurtosis:                       3.485   Cond. No.                         81.4
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Regression Coefficients:
Intercept: 3.3514786618317887
GINI Coefficient: 0.06189609664712329

P-values:
BASED ON THE STEPWISE REGRESSION
GDP_per_Capita      3.201026e-119
GINI Coefficient     3.351873e-69
dtype: float64
```

Figure 5: *Screenshot of Happiness model regression summary.*

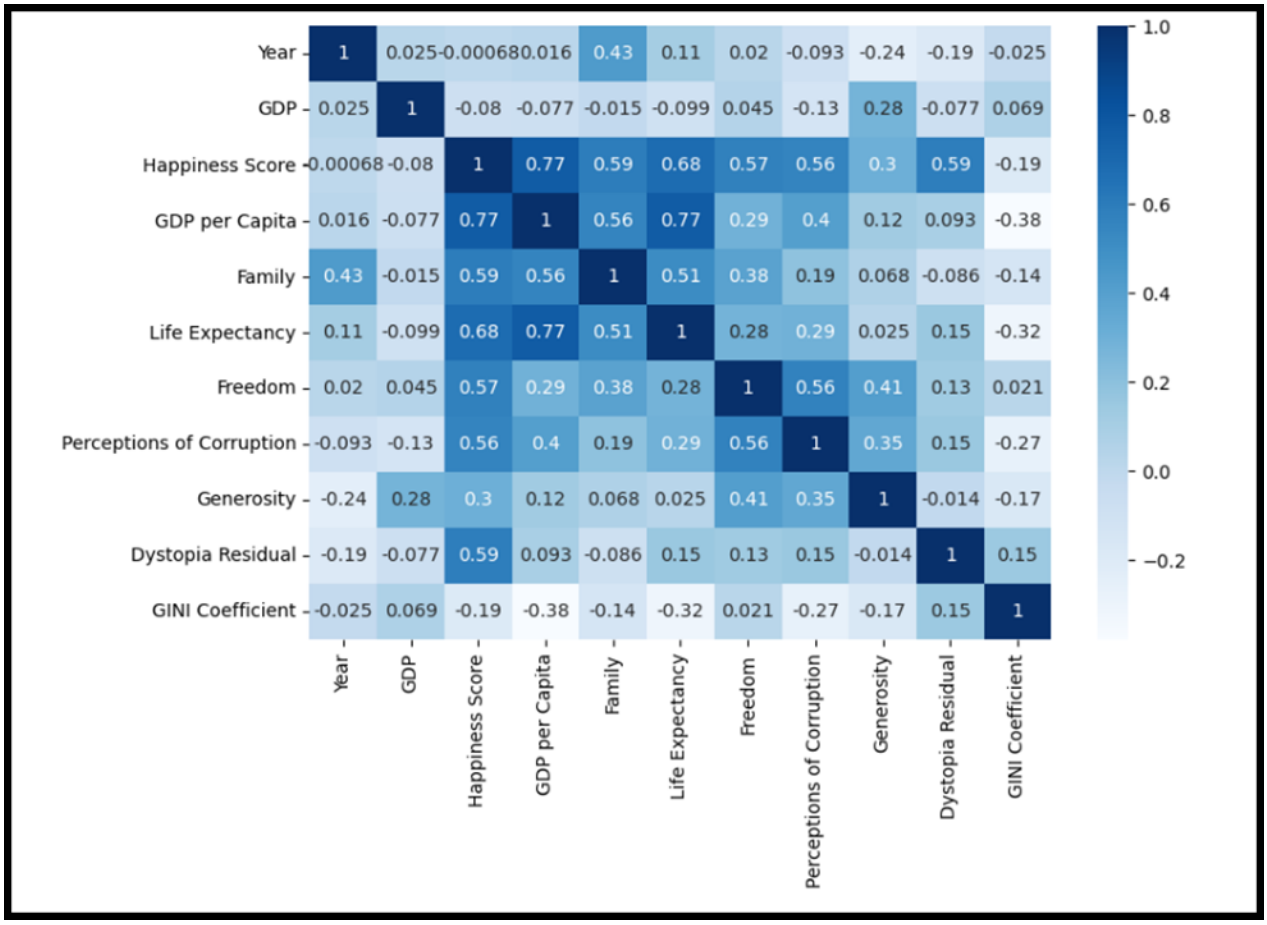


Figure 6: *Screenshot of the correlation matrix including all variables.*

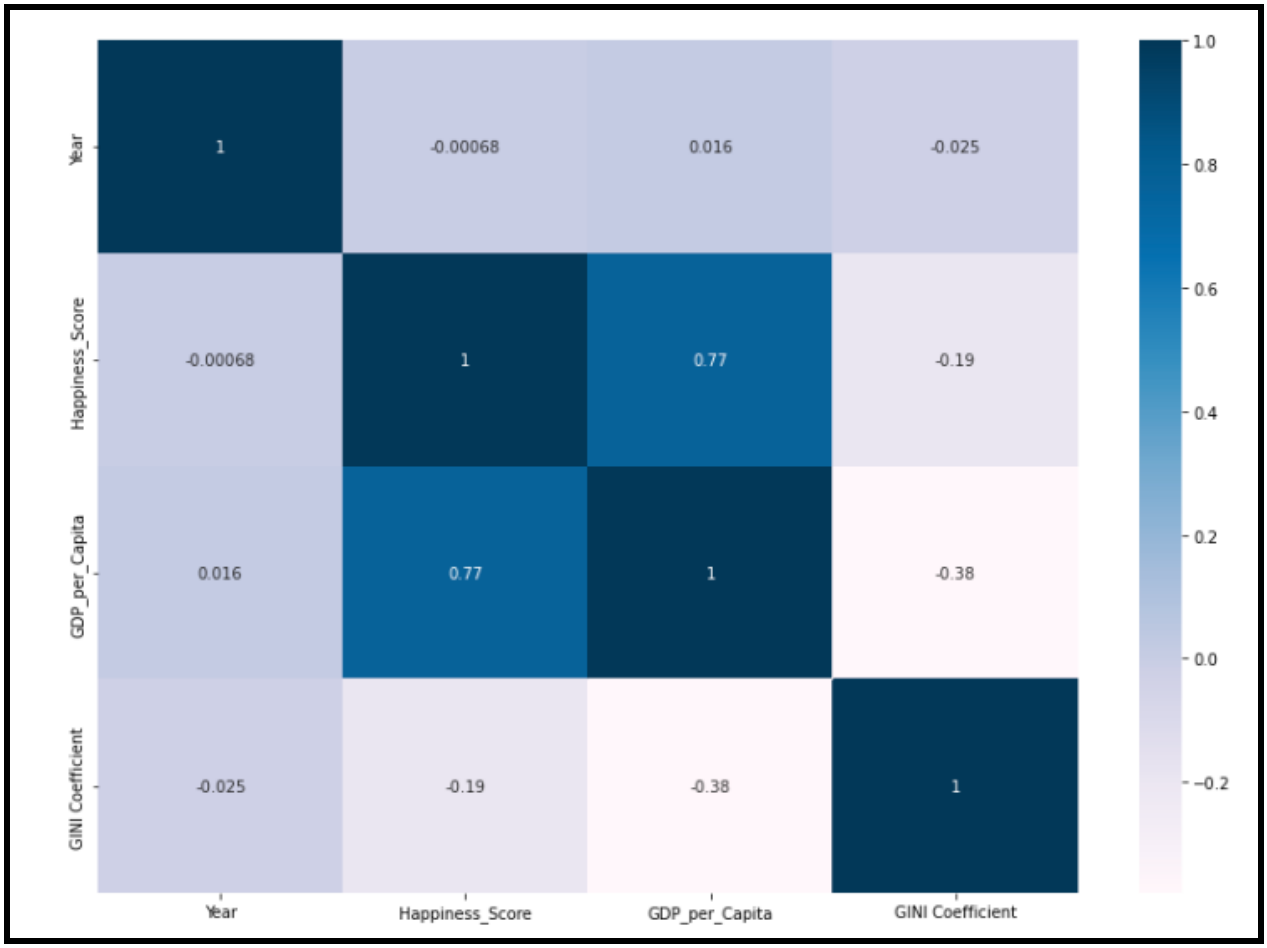


Figure 7: *Screenshot of the correlation matrix with the variables from the Happiness model.*

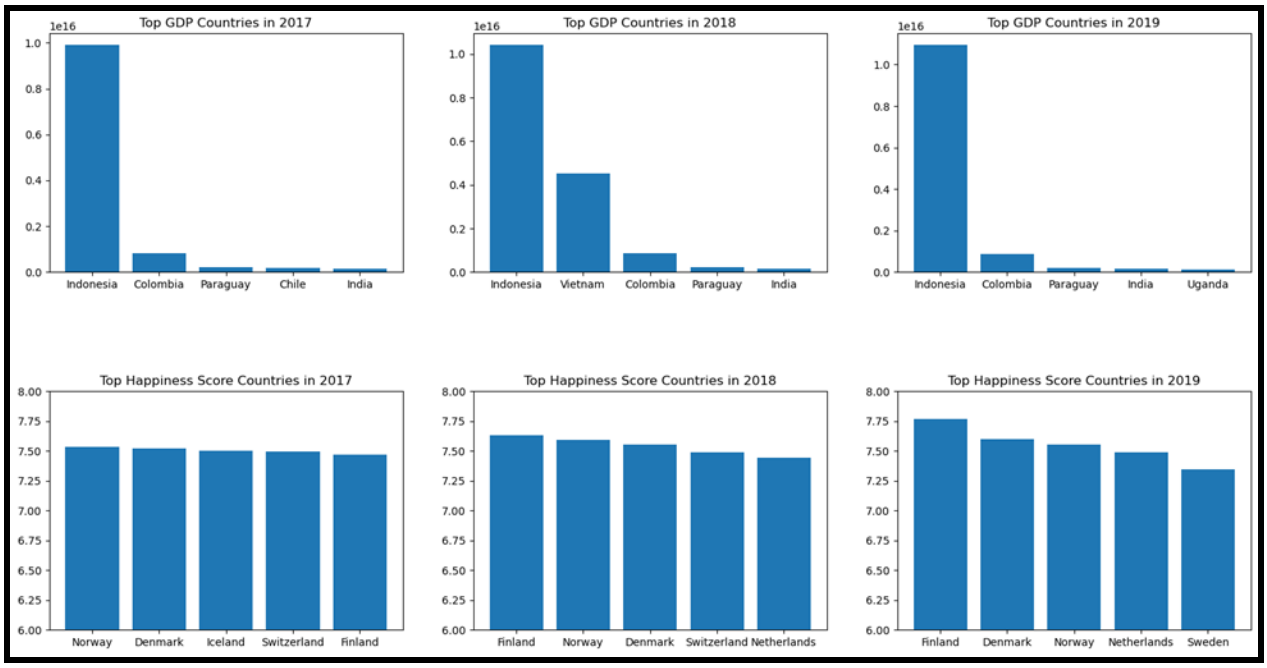
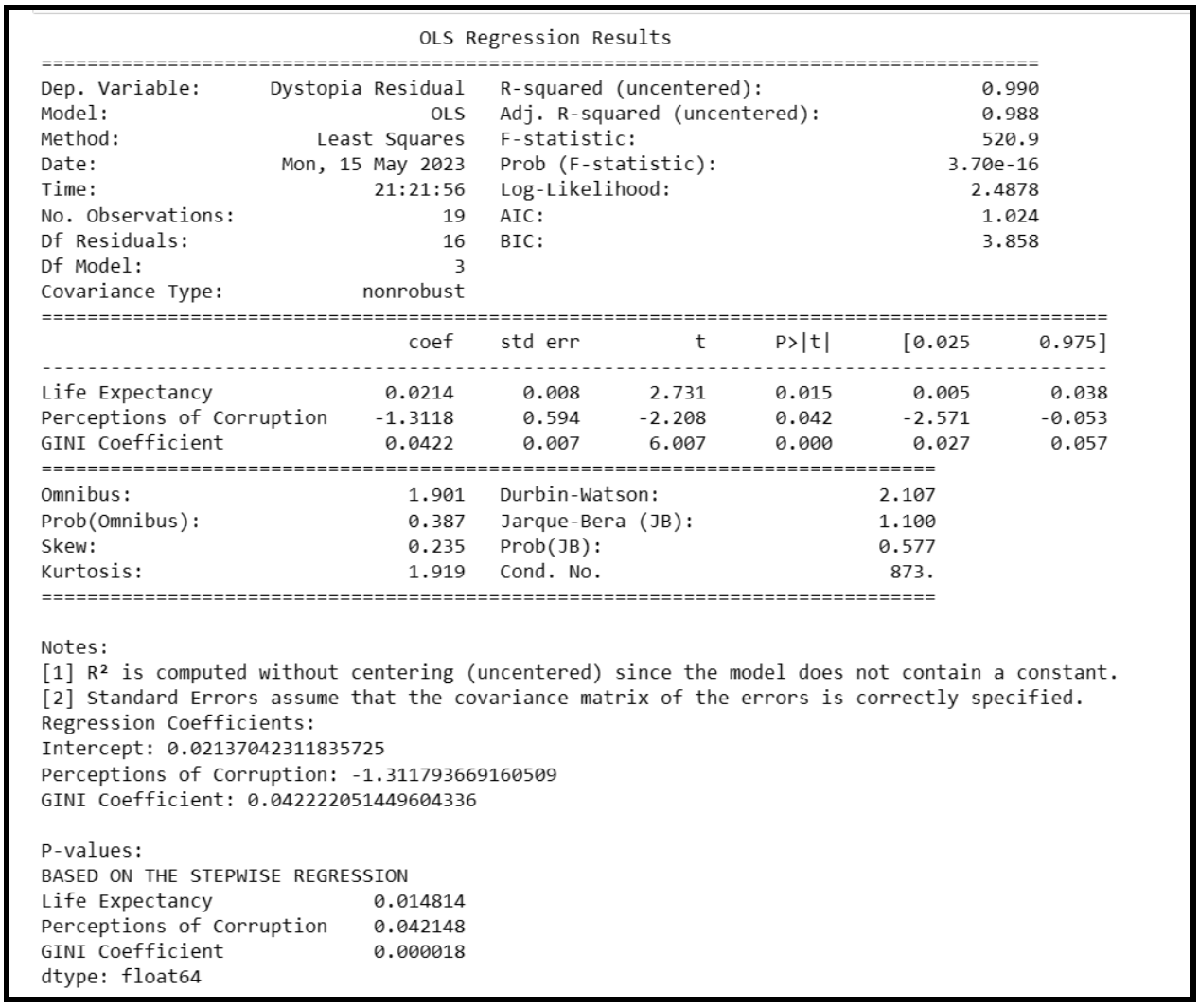


Figure 8: *Screenshot of the Highest GDPs compared to the Highest Happiness Scores for the Top 5 values in each category.*



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Dystopia Residual   R-squared (uncentered):                   0.990
Model:                            OLS   Adj. R-squared (uncentered):              0.988
Method:                 Least Squares   F-statistic:                              520.9
Date:                Mon, 15 May 2023   Prob (F-statistic):                    3.70e-16
Time:                        21:21:56   Log-Likelihood:                          2.4878
No. Observations:                  19   AIC:                                     1.024
Df Residuals:                      16   BIC:                                     3.858
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Life Expectancy                0.0214      0.008      2.731      0.015       0.005       0.038
Perceptions of Corruption     -1.3118      0.594     -2.208      0.042      -2.571      -0.053
GINI Coefficient               0.0422      0.007      6.007      0.000       0.027       0.057
==============================================================================
Omnibus:                        1.901   Durbin-Watson:                   2.107
Prob(Omnibus):                  0.387   Jarque-Bera (JB):                1.100
Skew:                           0.235   Prob(JB):                        0.577
Kurtosis:                       1.919   Cond. No.                         873.
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Regression Coefficients:
Intercept: 0.02137042311835725
Perceptions of Corruption: -1.311793669160509
GINI Coefficient: 0.042222051449604336

P-values:
BASED ON THE STEPWISE REGRESSION
Life Expectancy              0.014814
Perceptions of Corruption    0.042148
GINI Coefficient             0.000018
dtype: float64
```

Figure 9: *Screenshot of the Dystopia model regression summary.*

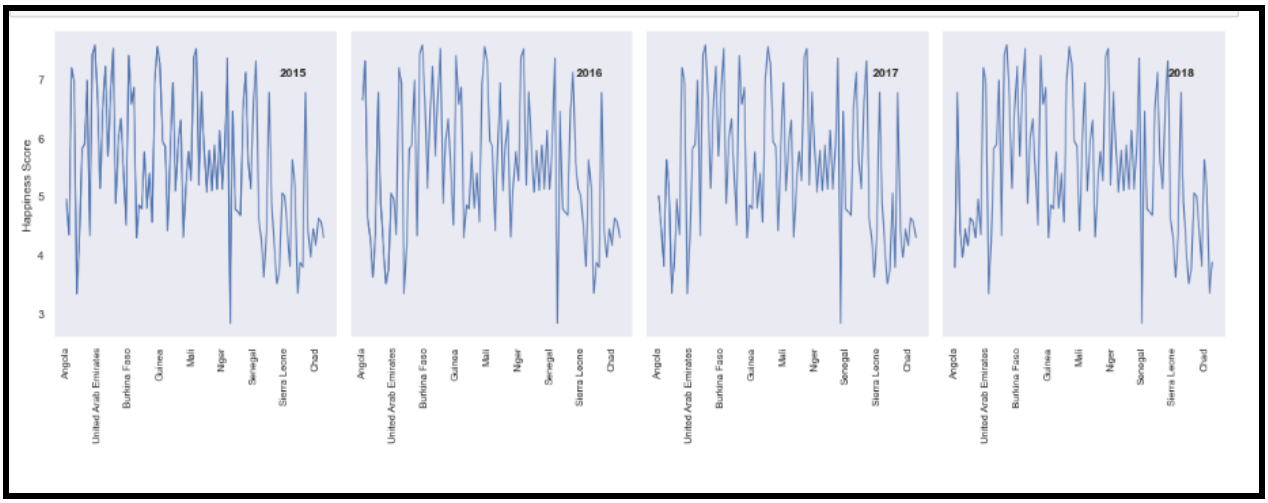


Figure 10: *Screenshot of Happiness Score by Country for the years 2015 - 2018.*