**Project Report:**

**PREDICTIVE MODELING ON WORLD HAPPINESS REPORTS 2015-2020**

Data Science 2: Statistics for Data Science

University of Waterloo

**Contents**

[Summary: 2](#_heading=h.gjdgxs)

[Introduction: 2](#_heading=h.30j0zll)

[Objectives: 2](#_heading=h.1fob9te)

[Data Set: 3](#_heading=h.3rdcrjn)

[Data Preparation & Challenges: 3](#_heading=h.2et92p0)

[Addition of columns for concatenated dataset: 3](#_heading=h.tyjcwt)

[Standardized naming of variables: 4](#_heading=h.3dy6vkm)

[Checking for missing values: 4](#_heading=h.1t3h5sf)

[Challenges in Data Preparation: 4](#_heading=h.4d34og8)

[Data exploration: 4](#_heading=h.26in1rg)

[Regression Analysis: 4](#_heading=h.lnxbz9)

[Transformation of Data: 8](#_heading=h.35nkun2)

[Fitting Linear Regression (OLS): 9](#_heading=h.1ksv4uv)

[Predictive Modeling: 10](#_heading=h.44sinio)

[Forward Stepwise Regression: 12](#_heading=h.2jxsxqh)

[Interpreting the coefficients: 12](#_heading=h.z337ya)

[Hypothesis Testing - Interpretation of p-value: 13](#_heading=h.3j2qqm3)

[Conclusion: 13](#_heading=h.1y810tw)

[References: 14](#_heading=h.4i7ojhp)

# Summary:

Happiness is an arbitrary concept and humans have always tried to find its source. Our group worked together to explore the relations of GDP per capita, healthy life expectancy (HLE), social support, freedom to make life choices, generosity, and perception about governments with the ultimate Target Variable- the “Happiness Score”. We download the datasets from Kaggle. Upon visually inspecting the datasets, we had to find consistency in names of variables, add variables Rank and Region, and finally drop null values from our concatenated dataset. Explanatory analysis revealed a general trend of dips and highs in variables Family, Freedom, and Generosity in years 2015-2020. Variables Happiness, Economy, Trust remained more or less constant across these years. Analysis shows the happiest nations are from regions Australia and New Zealand, and North America. Least happy nations represent Sub-Saharan Africa. R2 value indicated Economy, Family, Health, Freedom has strong and positive association with target variable but Trust and Generosity do not have strong association for some value with target variable.

# Introduction:

“Happiness is not by chance but by choice” ~Jim Rohn

For people, happiness either comes from within oneself or by helping others. But one notion is common: a human has to select happiness. In pursuit of happiness and its factors, we worked on data collected by Gallop World Poll to identify factors that contribute to the happiness of societies.

The Sustainable Development Solutions Network of UN releases the annual World Happiness Report (WHR) which is a survey of the state of general happiness in world countries. This report publishes happiness scores and different variables to correlate global bliss with life factors. The first WHR was published in 2012 as a result of 2011 adopted resolution ‘65/309 Happiness: Towards a Holistic Definition of Development’.

World Happiness Report takes data from Gallop World Report (GWP), World Health Organization (WHO), and World Development Indicators (WDI) of the World Bank. This report uses multiple factors but six are main variables to correlate with happiness (WHR n.d.). These are GDP per capita, healthy life expectancy (HLE), social support, freedom to make life choices, generosity, and perception about governments. All these six variables are correlated with Happiness Score or subjective well-being which is measured through Cantril Ladder. The Happiness Score is an average value for each country from responses collected through 2005-2019 (Helliwell *et al.*, 2015-2020).

The datasets from 2015-2020 were included for analysis in this report. We used multiple linear regression to model the relation of our predictors with our target variable Happiness Score. We also analyzed the correlation of each variable with our target variable.

# Objectives:

Our hypothesis is: GDP per capita, healthy life expectancy (HLE), social support, and freedom to make life choices, are associated with the happiness of citizens of a country.

We aimed to understand whether the chosen variables could define happiness of a nation. We also wanted to infer the individual relationship of the target with each explanatory variable.

# Data Set:

The datasets included in the study come from WH reports of 2015-2020. Six downloaded datasets had six main variables which were included for modeling. Information on each variable is given below (Helliwell *et al.*, 2015-2020).

1. **Happiness score:** Data for this variable is collected from years 2005 and onwards by Gallop World Poll. This variable is measured through Cantril Ladder, by using a simple question: “Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" Happiness Score is an average of national responses collected as answers to the above mentioned question. The citizens of each country answered according to their self-perceived happiness.
2. **GDP per capita:** These values are from the World Development Indicators (WDI) and the Penn World.
3. **Healthy life expectancy (HLE):** Taken from the World Health Organization's (WHO) Global Health Observatory data repository.
4. **Social support:** It is the national average of binary response (YES/NO : 0/1) to availability of friends or family in times of need.
5. **Freedom to make life choices:** This variable is the national average of responses collected by GWP.
6. **Generosity:** This variable recorded responses at country level whether citizens had donated to charity in the last two months.
7. **Perception about government:** This variable recorded the average of binary responses on the question of trust of governments of their countries.

# Data Preparation & Challenges:

Six datasets were downloaded from <https://www.kaggle.com/mathurinache/world-happiness-report>, these datasets had multiple variables with different names corresponding to the same variable. We ran a visual exploration of data in each dataframe, mainly to understand what features were present in each dataset and if they could be concatenated. These are the key findings from the initial exploration of data sets:

## Addition of columns for concatenated dataset:

1. A separate column for “year” had to be included before concatenating the dataframes.
2. The variable "Happiness Ranking" or "Ranking" which indicates the ranking of countries in each year was missing from data sets of 2017-2020. This variable was extracted from Year and Score columns.
3. Other key variables are Country, and Region. The variable Region was also missing in the dataset of 2017, 2018, and 2019.
4. A column of variable Region was added before concatenating the datasets of years 2017-2019. We have empty values in the Region column. Data for the years 2017, 2018, and 2019 did not have Region, so we populated the data for those years. The existing values in the column had inconsistencies (i.e. South Asia and also Southern Asia, which should have the same value). To address this issue, we used 2015 data as a dictionary and used lambda function to populate consistent region names for the countries across the dataset.

## Standardized naming of variables:

1. Our target variable is "Happiness Score" which was also referred to "Ladder Score", or "Score" in different years' data. As described above, there are 6 features (independent variables): Economy, Family, Health, Freedom, Trust, and Generosity. These features are named and organized differently in our data sets. Names of all the variables were organized and renamed before concatenating.
2. Columns in addition to the above mentioned columns vary across dataframes, most of which are being supplementary statistical measures. We dropped these columns for our analyses as we had only focused on six main features.

## Checking for missing values:

Missing values were checked in the concatenated dataset. There was only one missing value in variable Trust which was dropped using dropna().

## Challenges in Data Preparation:

Two major challenges were faced while preparing data which were corrected as described in the above section. Those challenges were:

1. The data from different years were not consistent and 6 years under study had 3 - 4 different approaches in structuring the data, naming the columns and other attributes. In order to resolve this issue, we did additional research to understand different approaches and to make sure we are interpreting the data correctly.
2. Some steps of data exploration and quality check involved visual/manual reading of data and identifying inconsistencies (e.g. in Country names and Region data). While this was somewhat time-consuming and prone to errors, we double-checked the data before and after cleaning to ensure consistency and accuracy.

# Data exploration:

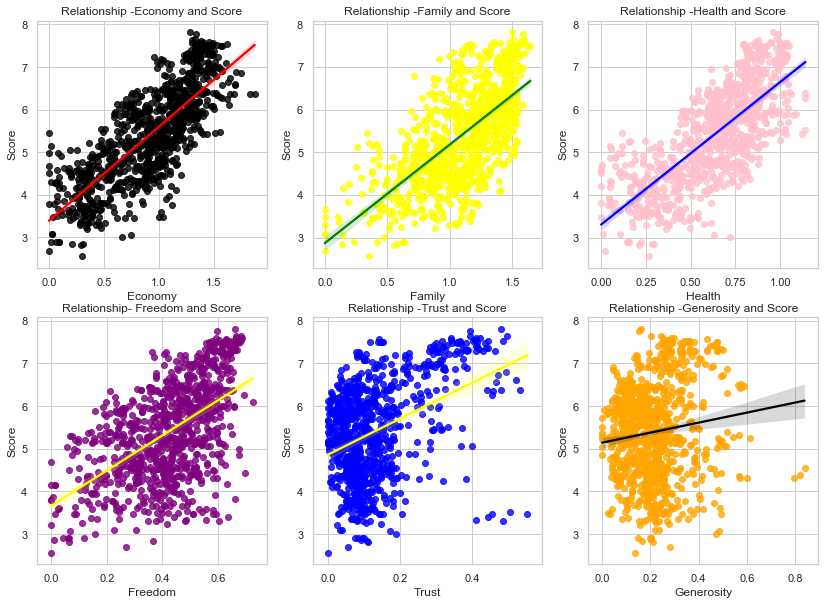
The analysis of the relationship between the 6 main variables in the dataset and a country’s happiness score proved to provide multiple interesting findings. By utilizing techniques in data science our group was able to determine the correlation between these variables and happiness score and whether any positive linear relationship exists.

# Regression Analysis:

Our group employed a linear regression model as a predictive modelling technique to determine whether there happened to be any positive relationship between the potential 6 variable indicators and happiness score (target variable). We chose Linear Regression Model as predictive Modelling Technique since our response variable was continuous in nature. But for applying Linear Regression, it must meet all the assumptions of Linear Regression.

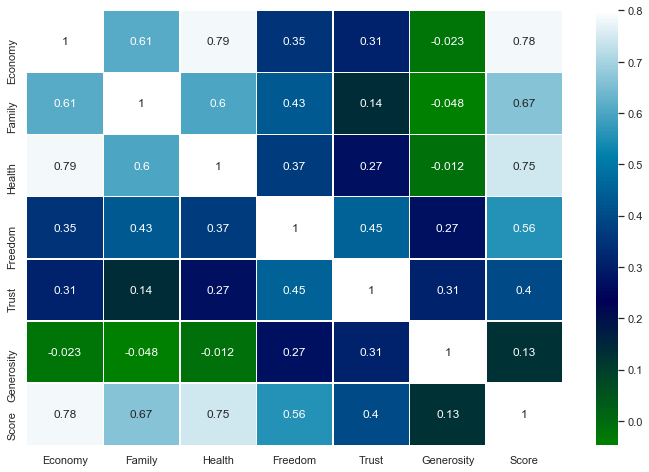
**Assumptions:**

1. Linearity (each Variable is linearly related to outcome
2. Residuals are Normal
3. No or Little Multicollinearity
4. Residuals have Constant Variability
5. **Linearity:** To determine whether linearity was present between the 6 main variables and happiness score, we formulated scatter plots to outline the relationship.

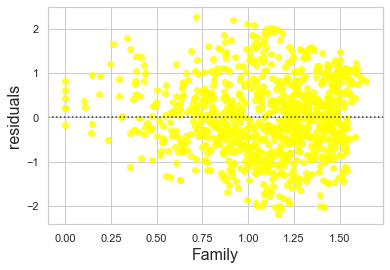
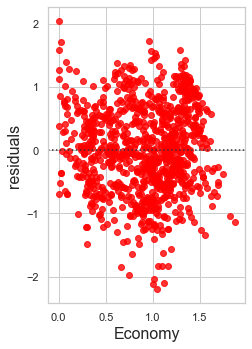


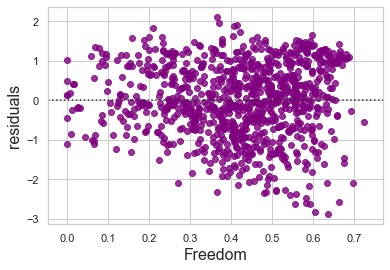
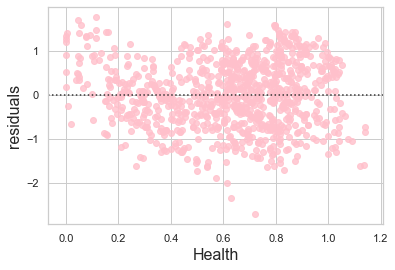
1. **Economy:** (GDP per capita) it shows strong linear positive relation with response variable Score. The regression line also seemed to suggest that the happiness score is dependent on the economy, as we noticed a larger economy tended to have a country with a higher happiness score. Unlike the other variables, this seems to pose a more consistent relationship with less present outliers. Nevertheless, only further analysis will determine if this correlation equals to causation.
2. **Family:** This variable shows strong positive Linear relationship with response variable Score. Though through visual analysis there does seem more outliers present that the economy, we relationship does seem apparent with a consistent pattern between the analysis of the explanatory X variable (family) and target Y variable (happiness score). A causation would not be incredibly surprising as one would think that a strong family dynamic would mean that citizens within that country are happier.
3. **Health:** This variable also seems to present a strong positive linear relationship with happiness score. This is understandable, as one would assume that if citizens within a country are healthier, they are in less discomfort than those which are in poor health which would subsequently increase their happiness. However, further hypothesis testing would be required to determine causation.
4. **Freedom:** This variable shows strong positive Linear relationship with response variable Score. A country with more freedom seems to indicate a country that is also happier based on our linear regression scatterplot.
5. **Trust:** This shows a strong Linear relationship with smaller values of x but for larger values of x the relationship is not Linear. Nevertheless, most of the data points within the graph seem to be skewed. Through data transformation it may indeed begin to present a positive relationship between our target variable (happiness score) and trust. Some of the values are away from the cloud.
6. **Generosity:** Analyzing our linear regression graph seems to present a similar situation when we were analyzing the trust variable, with the majority of X values appearing skewed on the scatterplot graph.

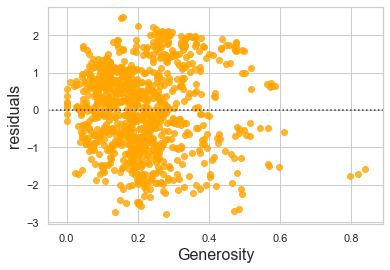
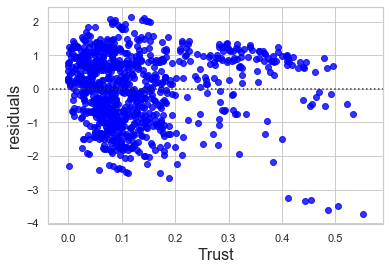
Our observation was further validated by plotting a heatmap which analyzed the relationship between all variables present within our analysis. Our heatmap (ranged from -1 to 1 indicating the strength of a linear relationship between variables) indicates a strong relationship between these 4 variables and happiness score (Freedom = 0.56, Health = 0.75, Economy = 0.78, Family = 0.67). Conversely, generosity and trust (respectively 0.13 and 0.4 on the heatmap) do not seem to indicate any association with happiness score. Nonetheless, we will transform this data to remove the noticeable skew that was apparent within our linear regression plots to see whether this may then present a positive linear relationship.



1. **Residuals are Normal:**







1. **Economy:** Residuals are nearly normal
2. **Family:** Residuals are nearly normal
3. **Health:** Residuals are nearly normal
4. **Freedom:** Residuals are nearly normal
5. **Trust:** Residuals are not nearly normal
6. **Generosity:** Residuals are not nearly normal
7. **No Multicollinearity Check**

If any 2 or more than independent variables are highly correlated with each other, they will influence the entire model and lead to wrong predictions. Multicollinearity may impact slope while also affecting the standard error which will also be higher in the case of multicollinearity. Multicollinearity can be checked using heatmap and VIF.

The best way to identify multicollinearity is to calculate the Variance Inflation Factor (VIF) corresponding to every independent Variable in the Dataset. Multicollinearity is not always a problem.

**Predictors VIF**

Economy 18.084174

Family 18.578173

Health 20.016245

Trust 3.403368

Generosity 4.218613

Freedom 13.283738

When we care more about how much each individual feature rather than a group of features affects the response variable Score, then removing multicollinearity may be a good option. In our analysis we care about how a group of features affect the target variable Score so there is no need to fix this.

## Transformation of Data:

By creating graphs outlining the constant residuals, our initial assumptions about the data points for trust and generosity are not validated as both these variables display skewness. Since this can degrade the performance of our predictive model to determine happiness score while also violating the assumption of linearity, transformation for trust and generosity’s dataset must occur in order to determine whether a linear relationship truly exists.

Though **trust and generosity** did not initially seem to indicate any association, using techniques in data transformation it is possible that altering the data we may then prove to show a relationship between variables which previously did not seem to be related in any aspect. **Square root transformation** for these datasets seemed most appropriate as the majority of X values were skewed to the right. Log transformation was not possible as some of the data points were valued at 0 which prevented us from employing this method. Upon transforming the datasets using a square root transformation, we determined that both trust and generosity presented a linear relationship with happiness score, though this was not apparent when we first created our scatterplots. This was further validated through the creation of new scatterplots for the trust and generosity variables using square root transformation.

This would also require us to create an adjusted heatmap utilizing our original results for variables such as **freedom, health, family and economy** while employing the new results from our transformed data for **trust and generosity**. Our new results for trust and generosity do indicate an increase in the association relative to happiness score (Trust = 0.36, Generosity = 0.12). Nevertheless, this would still indicate that generosity does not present a significant enough association with happiness score and therefore should be dropped from our predictive model.

## Fitting Linear Regression (OLS):

**Freedom:** The LARGEST regression coefficient is for **Freedom of 1.15859** means that on average, every 1 unit increase in Freedom (to make choices) is associated with an increase of 1.158 in Happiness score. This was further apparent in our earlier scatterplot.

**Economy:** Another significant feature, indicated by a high regression coefficient, is **Economy of 1.1133.** This means that on average, each point increase in Economy is associated with an increase of 1.11 in happiness. This means that a country with the means to provide for its citizens through a high GDP per capita would also increase the happiness of that country.

**Health:** Our third most significant feature is **Health of .9914.** This means that on average, each unit increase in Healthy life expectancy is associated with an increase of happiness score

**Trust:** Our Fourth most significant feature is **Trust of 0.7832**. This means that on average, each unit increase in Government Trust (the perception of citizens having an uncorrupt government) is associated with an increase in 78.32 in happiness.

An interesting aspect of this is how the heatmap showed a lower correlation between trust and happiness score. However, after employing an OLS stats model using all variables, it is apparent that trust is a decent indicator of happiness, and that if citizens do not trust a country’s government, they are more likely to be unhappy. Additionally, this coefficient of determination or r-squared, tells us that 75% percentage of the total variance in the Score can be explained by the linear regression model. This is an important statistic that measures how 'good' our model is at predicting Score.

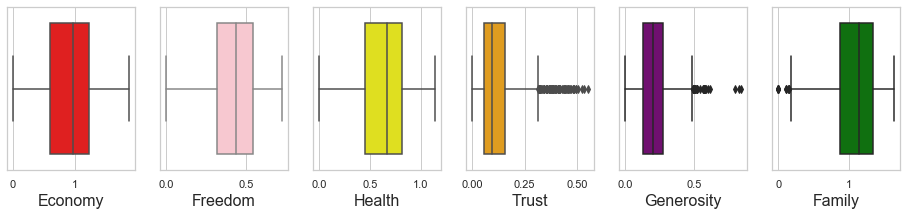
# Predictive Modeling:

We performed our predictive modeling as follows:

1. Descriptive analysis on the Data
2. Data Treatment (outlier Fixing)
3. Data Modeling
4. Evaluate the performance of Model

Following are the details.

1. Descriptive analysis is performed using boxplot and describes method to see the distribution of each predictor.



1. Data Treatment (Outlier Fixing)

Trust and Generosity both show outliers and would not be ideal for use in our predictive model. Additionally, the low level of association to the target variable further validates this notion.

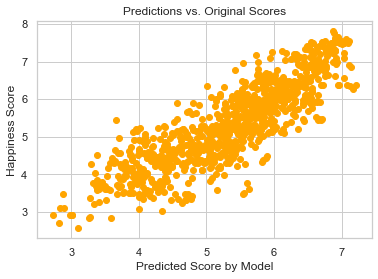
1. Data Modeling: Regression analysis

Regression analysis is a form of predictive modeling technique which finds the relationship between a target and (predictor). This technique is used for finding the causal effect relationship between variables.

Train and Test Split: For training and testing our linear regression model we split the data into training and testing part using sklearn library 70% for training part and 30% for testing part.

Feature Selection: Rank Feature is strongly negatively correlated with our Target Feature, It should be dropped. Keeping both rank and Score does not make sense. Year, Region , Country and Rank all are not required for building a model since year , region and country cannot contribute to happiness so better to remove them.

Residual Plot: We create a scatterplot between the predicted Scores, available in m.fittedvalues (where m is the fitted model) and the original Scores. We evaluate this plot to see how well our regression model predicts the score of given data in our data set. A perfect model would get us a scatterplot where all the data lies on the 45 degree line. That would mean that x = y, and every predicted Score would have equaled the actual score. Conclusion of Pattern The data does not show any pattern between fitted values and residuals, no assumptions have been violated.

****

This coefficient of determination or r-squared, tells us that 75% percentage of the total variance in the Score can be explained by the linear regression model. This is an important statistics that measures how 'good' our model is at predicting Score.

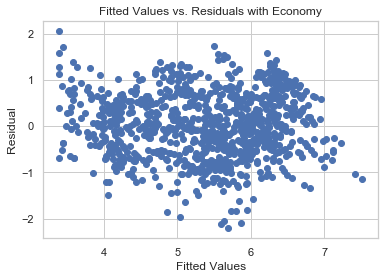
Evaluating the fit of model with Single Strong predictor Economy: We took a reduced model using only our strongest predictor, Economy, to predict Score. We evaluated this model using:

\* Residual plot

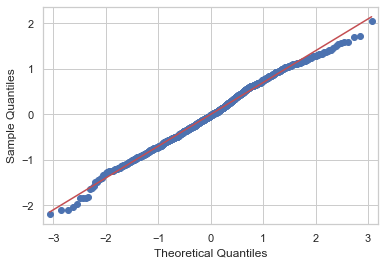
\* Q-Q plot

\* predicted Score by model vs Actual Score

*Residual plot:* This plot is testing that the error in the model is normally distributed. It does so by ensuring that there is no pattern between the fitted values and the residuals. No clear pattern in the plot means no assumptions have been violated.

****

*Q-Q plot:*The Residual Q-Q plot tests for normality of the distribution and does a better job at showing outliers and distribution of the data than a residual plot. It is easier to see patterns in the single line of the Q-Q plot than a scatter plot. On the x-axis we have the cumulative sum of standard normal distribution. On y-axis cumulative we have the sum of sample distribution. In our graph, some points at the lower end and upper end are not normal; otherwise the rest of the part is normal.

****

## Forward Stepwise Regression:

This method is also used for finding and selecting features. In this method RMSE decreases every time with the addition of a new feature. This was apparent as every time we added an additional variable (freedom, health, family and economy) the RMSE fell which indicated that each of these variables could be used as an effective predictor of happiness score.

## Interpreting the coefficients:

Based on our new OLS utilizing our trained test split, we can determine how inserting specific values for coefficients can alter the results for happiness score. Our model is useful as it could be used to predict the future happiness score of a country based on historical trends. By employing the 4 main variables (health, freedom, family, economy) that impact happiness score, data science can be utilized to determine and predict the happiness score of any country if we have tabulated the necessary statistical data.

Based on our findings:

1. **Health:** An increase in *1* unit of Health would cause an increase of *0.9901* in happiness score (Score – Target Variable)
2. **Freedom:** An increase in *1* unit of Freedom would cause an increase of *1.9173* in happiness score (Score – Target Variable)
3. **Family:** An increase in *1* unit of Health would cause an increase of *0.5687* in happiness score (Score – Target Variable)
4. **Economy:** An increase in *1* unit of Health would cause an increase of *1.1825* in happiness score (Score – Target Variable)

OLS Regression Results

==============================================================================

Dep. Variable: Score R-squared: 0.749

Model: OLS Adj. R-squared: 0.748

Method: Least Squares F-statistic: 693.8

Date: Sun, 13 Dec 2020 Prob (F-statistic): 3.33e-277

Time: 21:02:03 Log-Likelihood: -788.56

No. Observations: 934 AIC: 1587.

Df Residuals: 929 BIC: 1611.

Df Model: 4

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

Intercept 2.2754 0.071 32.235 0.000 2.137 2.414

Economy 1.1825 0.078 15.218 0.000 1.030 1.335

Freedom 1.9173 0.136 14.070 0.000 1.650 2.185

Health 0.9901 0.123 8.034 0.000 0.748 1.232

Family 0.5687 0.077 7.363 0.000 0.417 0.720

==============================================================================

Omnibus: 15.481 Durbin-Watson: 1.414

Prob(Omnibus): 0.000 Jarque-Bera (JB): 15.986

Skew: -0.290 Prob(JB): 0.000338

Kurtosis: 3.271 Cond. No. 15.0

==============================================================================

## Hypothesis Testing - Interpretation of p-value:

P-value for each of the predictors (Economy, Family, Freedom and health) is less than significance level (0.5) so we can say that they are predictors for our response variable. This would mean we would reject the null hypothesis (H₀) and that there would indeed be a causal relationship between our explanatory variables and response variable (happiness score).

**Conclusion:**

Freedom would have the most significant impact on a country’s happiness score. Both Economy and Health indicate that it almost has a 1:1 increase between the explanatory X variables and target Y variable (happiness score). Family seems to have the lowest impact on happiness score. This would seem to indicate that a country’s citizens value liberty above all other societal wellness indicators. A strong economy and GDP per capita is also indicative of a happier society, which would make sense as if a country does not have to worry about providing the bare necessities to their citizens, they are more likely to be happy (though looking at inequality in the distribution of resources could be a separate topic of interest). Surprisingly, health is of less importance than a country’s economy, with citizens being ready to sacrifice a healthy populace with more ailments rather than be economically poor. Finally, family seems to be the variable with the lowest impact on happiness. Though the family dynamic is important, it would seem that citizens prefer to live in a wealthier, healthier and more prosperous society.

# References:

1. World Happiness Report WHR (n.d.). FAQs, retrieved from: <https://worldhappiness.report/faq/> retrieved on: 8th December 2020.
2. Helliwell JF, Layard R, Sachs J (2015). Statistical Appendix for Chapter 2, <https://s3.amazonaws.com/happiness-report/2015/StatisticalAppendixWHR3-April-16-2015.pdf>
3. Helliwell JF, Huang H, Wang S (2016). Statistical Appendix for “The Distribution of World Happiness”, Chapter 2, World Happiness Report Update 2016, <https://s3.amazonaws.com/happiness-report/2016/StatisticalAppendixWHR2016.pdf>
4. Helliwell JF, Huang H, Wang S (2017). Statistical Appendix for “The social foundations of world happiness”, Chapter 2, World Happiness Report 2017, <https://s3.amazonaws.com/happiness-report/2017/StatisticalAppendixWHR2017.pdf>
5. Helliwell JF, Huang H, Wang S (2018). Statistical Appendix 1 for Chapter 2 of World Happiness Report 2018, <https://s3.amazonaws.com/happiness-report/2018/Appendix1ofChapter2.pdf>
6. Helliwell JF, Huang H, Wang S (2019). Statistical Appendix 1 for Chapter 2 of World Happiness Report 2019, <https://s3.amazonaws.com/happiness-report/2019/WHR19_Ch2A_Appendix1.pdf>
7. Helliwell JF, Huang H, Wang S, Norton M (2020). Statistical Appendix for Chapter 2 of World Happiness Report 2020 https://happiness-report.s3.amazonaws.com/2020/WHR20\_Ch2\_Statistical\_Appendix.pdf