Evaluation and Analysis of World Happiness Score (April 2020)

Akhila Saineni, Nihar Garlapati, Pranitha Chandra

Abstract—the world happiness score report is base done a wide source of data, however the key dataset used is from the Gallup World poll. In the following analysis and study, we would focus on the happiness score recorded by people of different countries through the Gallup world poll. We also explore other economical and social data points related to the same country such as GDP, Life expectancy, Freedom to make life choices, Perception of corruption & social support. We were able to segregate the countries into their respective contients and region as well. In the following report, we will continue to explore the correlation of various economic and social metrics of countries with their happiness score. A multi variate linear regression model will be dicussed to predict the happiness score of each country. At the end of the report we will conclude with major correlations identified and key economical and social metrics and their impact on the happiness score at different continents. The following report would help governments of different countries to identify various metrics that would improve the overall happiness score of their respective countries.

1 Introduction

THe world happiness score is a landmark survery where individuals from 156 countries have participated and provided statistics about the happiness. In the report, we will evaluate the happiness score results from various countries and their correlation with various economic and financial metrics of the respective countries.

The independent variables or metrics explored for this analysis are as follows. GDP Per Capita, Social Support, Freedom to make life choices, Perceptions of corruption, Health life Expectancy, Region & Continent. The response variable which we are interested is the world happiness score.

2 Exploratory data analysis

2.1 General Statistics

The statistics of the variables used in the analysis are as follows.
The Score of a country ranged from 2.8 to 7.7(M=5.42, SD=1.10), Score was normally distributed with skewness of 0.00 and kurtosis of 2.3
The GDP of a country ranged from 0.0 to 1.6(M=0.90 SD=0.39), GDP was normally distributed with skewness of -0.4 and kurtosis of 2.23
The Social support of a country ranged from 0.0 to 1.6(M=1.21 SD=0.29), Social support was normally distributed, left skewed with skewness of -0.00 and kurtosis of 1.14
The Freedom to make life choices of a country ranged from 0.0 to 0.63(M=0.39 SD=0.14), Freedom was normally distributed with skewness of -0.68 and kurtosis of 2.89
The Generosity of a country ranged from 0.0 to 0.56(M=0.18 SD=0.09), Generosity was normally distributed with skewness of 0.73 and kurtosis of 4.09
The Corruption of a country ranged from 0.0 to 0.45 (M=0.10 SD=0.00), Corruption was normally distributed, right skewed with skewness of 1.63 and kurtosis of 5.39
The Life Expectancy of a country ranged from 0.10 to 1.14 (M=0.73 SD=0.23), Corruption was normally distributed, left skewed with skewness of -0.54 and kurtosis of 2.44

After conducting data screening, the dataset checked out to be accurate with no missing data and no outliers were eliminated. We have used mahalanobis score to determine the outliers that needed exclusion.

2.2 Key Findings

Australia seems to be the continent with highest average happiness score followed by Europe. Africa seems to be the contient with the lowest average happiness score next to Asia. Refer Figure 3.3.4 for more information
Finland has the highest happiness score of 7.7 followed by Denmark with a happiness score of 7.6. Central African Republic and South Sudan have the lowest happiness scores 3.0 & 2.8 respectively.

3 Correlation

3.1 Correlation Matrix

A correlation matrix has been created using the corrplot function of the corrplot package. The correlation matrix provided us with correlation coefficients of various variables in the dataset. Happiness Score seems to have high positive relationship with GDP, Life Expectancy & Social support, medium positive relationship with freedom to make life choices and perception of corruption. Happiness score seems to have week positive correlation with

- Akhila Saineni is with Harrisburg Institute of Science and Technology, Harrisburg, PA 17101. E-mail: asaineni@my.harrisburgu.edu
- Nihar Garlapati is with Harrisburg Institute of Science and Technology, Harrisburg, PA 17101. E-mail: ngarlapati@my.harrisburgu.edu
- Pranitha Chandra is with Harrisburg Institute of Science and Technology, Harrisburg, PA 17101. E-mail: pchandra2@my.harrisburgu.edu

Genorisity. Refer figure 3.3.1 for more information

3.2 Correlation by Continents

When looked at correlation of the above mentioned economic and social metrics with happiness score by each continent separately, a similar positive correlation has been detected for GDP across different contients. However, Life expectancy seems to have strong positive relationship only in developed continents. In other words, GDP seems to be an important aspect of happiness all across the world, but life expecticeny seems to be important only for continent with mostly developed nations. Figure 3.3.2 & 3.3.3 can be reviewed for additional information

4 ANOVA

As we have fabricated the countries information into regions and continents for better overview on the data as a whole and in segments. We wanted to see if there is a difference in behavior of Happiness score among different continents and regions. Like stated before, continents and regions variable are categorical variables; having more than 2 categories hence we have decided to use ANOVA (Analysis of Variance) to see how the means of the samples are compared.

This type of statistical test compares the variance in the group means within a sample whilst considering only one independent variable or factor. The assumptions of one-way ANOVA are conducted before the test to confirm if the data meets the assumptions of Sample Independence, Normality, Variance Equality and if our dependent variable i.e. Happiness Score is continuous.

Hypothesis of One-Way ANOVA (between Continents)
- Null Hypothesis - There is no difference in the average happiness score between different continents.
- Alternative Hypothesis - There is a difference in the average happiness score between different continents

Hypothesis of One-Way ANOVA (between Regions)
- Null Hypothesis - There is no difference in the average happiness score between different regions
- Alternative Hypothesis - There is a difference in the average happiness score between different regions

Based on the One-Way ANOVA test statistics summary, it is considered that the ANOVA test is significant, therefore we reject the null hypothesis and conclude that there is a difference in the average happiness score between different continents and regions. Additionally, from the Levene's test, it is observed that the variances are equal among both the continents and regions. The above conducted ANOVA test only demonstrates if there is or there is no difference in the average happiness score between different continents and regions, but it does not provide any further information on which continents or regions the average happiness score is different. As our ANOVA test proved statistically significant, for good insight into the dataset we went ahead to perform the Post Hoc Method to identify which continents are different or are similar in terms of happiness score. In a manner, there is a significant difference in the happiness score between the continents with higher number of developing nations such as Asia and Africa when compared to continents with fewer developing nations such as Europe and North America as shown by Bonferroni's correction in Post Hoc Method.

5 REGRESSION MODEL

5.1 Introduction

To conduct the final analysis, a model is built to predict happiness score based on predictor variables. In order to achieve this, the entire dataset is divided into two datasets namely train dataset and test dataset using random sampling. The train dataset contained 75% of the original dataset and the test dataset contained remaining 25% of the original dataset. The train dataset is used to fit the regression model and test dataset is used to predict our happiness score using the model fitted on train dataset.

Linear Regression modelling is used to build our prediction model. Our data is nearly normally distributed and passes through other assumptions of linear regression. The function lm() is used in R to fit linear regression models for Happiness score variable using different sets of independent variables in each model. Furthermore, backward stepwise approach is used to analyze the models being built. As a matter of fact, we are not using regions and continents variable in our models as they were fabricated manually and added to the original dataset by us for a broader picture of the dataset besides that the two categorical variables couldn't be any significance to the modelling (as shown by the results of the model_all that includes all the variables in the dataset as predictor variables). The decision parameters Adjusted R-square and p-values of the variables are used as our base for model selection and improvement along with other factors/scores like RMSE, MSE, AIC and BIC. The model with higher value of Adjusted R-square and lower values of AIC, BIC, MSE and RMSE is considered a better fit in our modelling analysis. The final model that is considered to be our prediction model amongst all the other models, has the highest Adjusted R-square value at 0.7938 which means that the model explains approximately 79% of the variation in the happiness score variable (Figure 5.1.1) using the selected independent variables whereas the rest of the 21% of the variation in happiness score variable cannot be explained by the selected independent variables.

5.2 Residual Plots

From the figure 5.2.1, residual plots test the assumptions of whether the relationship between the variables is linear (i.e. linearity) and whether there is equal variance along the regression line (i.e. homoscedasticity). From the residual plots, the spread of the residuals in the residual’s vs fitted plot suggests there is a linear relationship between the Residuals and the fitted values, thereby indicating the points vary around the regression line in a random and constant manner. The residuals fan-out from left to right having a consistent spread around the residual line (=0). The data points do not vary in a complex fashion representing equal variance and the residual vs fitted plot shows that the error variances are equal. The corresponding Residuals vs fitted plot looks like the assumption of equal variances is not violated. Normal Q-Q plot represents that most of the data lies on the line and all or most of the values are centered on zero lie on the line whereas the points 3, 78, and 111 at the left tail deviate away from the tail. Hence, meeting the assumption of linearity and shows normal distribution of the data. The third plot (Scale-Location) does look random and does not form any identifiable shapes and the regression line is approximately horizontal with x-axis. The Residual vs Leverage Cook's plot here shows the points that have the greatest influence on the regression i.e. leverage points and points 3, 78, and 111 have greater influence on the model.

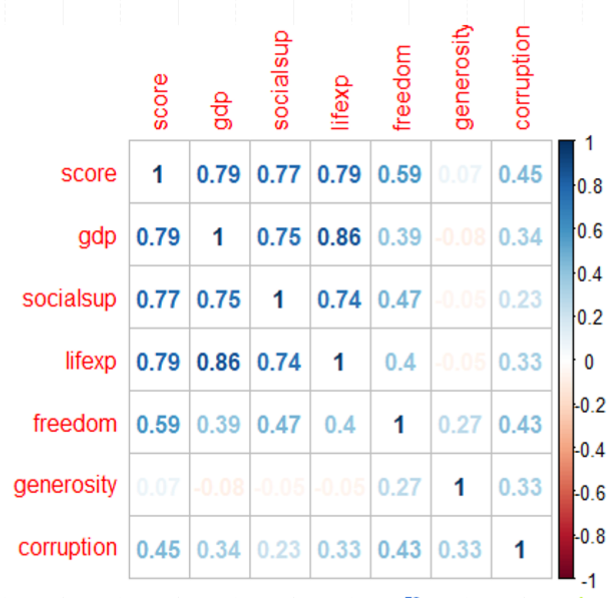
5.3 Interpretation of the Model

Interpretation of the analysis performed, we built the multiple linear regressions by considering only the significant variables against the response variable i.e. happiness score and discared insignificant predictor variables from earlier models. The best fit model achieved 0.79 adjusted R-square value, which indicates a good fit. Also, the p-value for all the predictor variables are below 0.05 which indicates the variables are statistically significant to the prediction of the happiness score. Also, checking the residual plots and QQ plot, we can see that the residuals have no pattern and are normally distributed, which means the model fit the data well. All of this is assuming that our assumptions of linear regression are fulfilled, if not the model results will be irrelevant if the assumptions were not met.

5.4 Model Prediction

When we plot the prediction of the best fit model (model_5) on the test dataset against the true values, the graph shows that the spread of the response variable is similar to the linear model. Nevertheless, we cannot depend on this output as the experiment was built on a very small dataset and this dataset does not contain other variables such as weather or education level which might help in more accurate prediction.

6 FIGURES



	score	gdp	socialsup	lifexp	freedom	generosity	corruption
score	1	0.79	0.77	0.79	0.59	0.07	0.45
gdp	0.79	1	0.75	0.86	0.39	-0.08	0.34
socialsup	0.77	0.75	1	0.74	0.47	-0.05	0.23
lifexp	0.79	0.86	0.74	1	0.4	-0.05	0.33
freedom	0.59	0.39	0.47	0.4	1	0.27	0.43
generosity	0.07	-0.08	-0.06	-0.05	0.27	1	0.33
corruption	0.45	0.34	0.23	0.33	0.43	0.33	1

Figure 3.3.1 Correlation Matrix

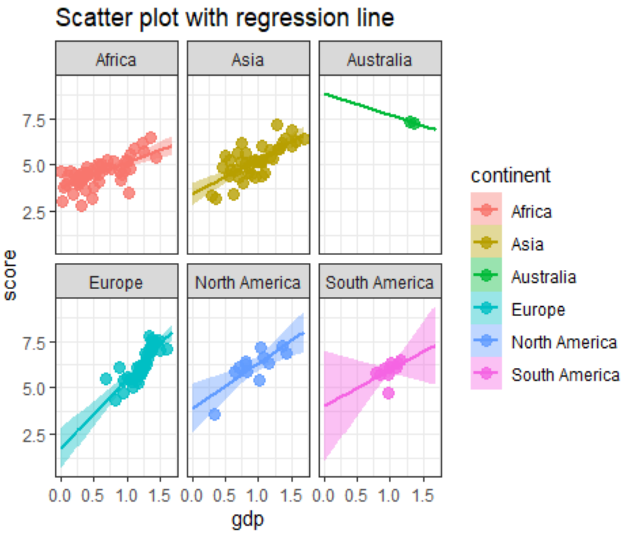


Figure 3.3.2 Scatter Plot with Regression Line of GDP

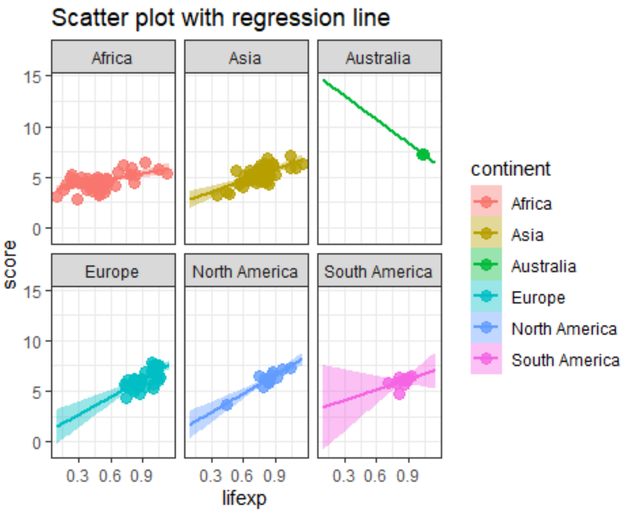


Figure 3.3.3 Scatter Plot with Regression Line of Life Expectancy

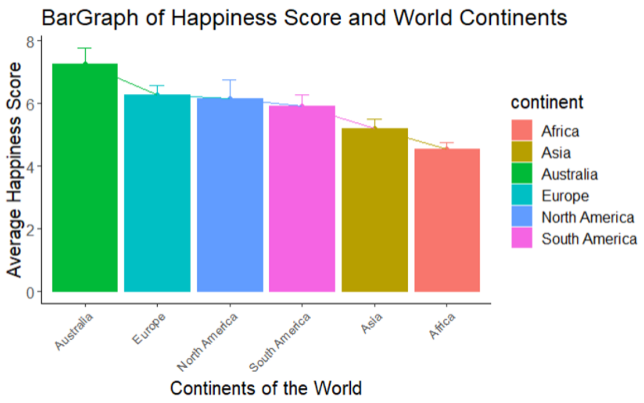


Figure 3.3.4 Average Happiness Score by Continent

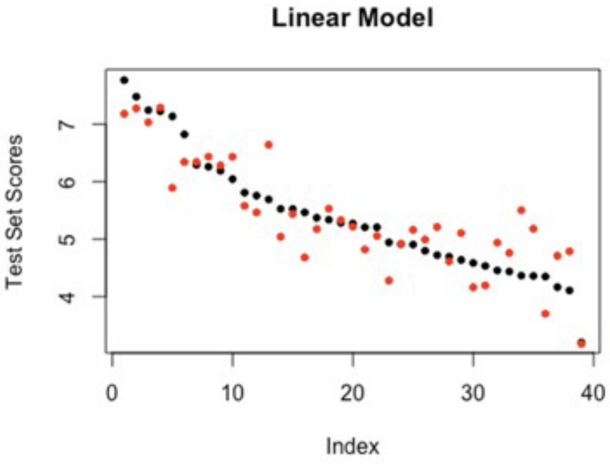


Figure 5.1.1 Linear Regression Model

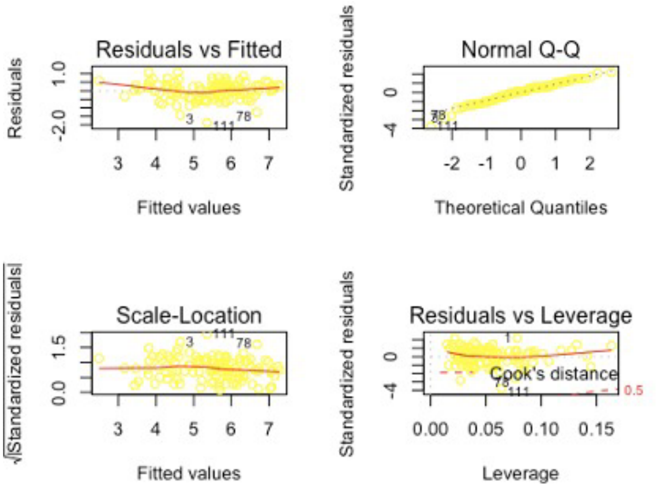


Figure 5.2.1 Residual Plots

7. Conclusion

In this study, we explore all the variables which can affect a country's happiness score. Through the exploratory analysis, we discovered that GDP, Social Support, Life Expectancy, Freedom and Corruption are the most important factors that drives the happiness score. Using multiple linear models, we see that the duration prediction has a good fit of training data and small generalization error, but prediction of happiness score seems to vary periodically, which is something that can be addressed in future work.

We can conclude that GDP, Social Support, Life Expectancy and Freedom play a major role in determining a countries happiness score. If the objective of a country to keep their citizens happy, the following analysis proves that GDP itself doesn't play a key role but metrics such as life expectancy, freedom to make life choices and perception of corruption also play a major role in improving the happiness score. In developing continents such as Africa, Life expectancy is not heavily correlated with happiness. However, GDP seems to be heavily correlated. In the north, American continent both GDP and Life expectancy play a key role in contributing to the happiness score. Most North American countries seem to have higher freedom score when compared to Asia & Africa, which also tells us why North American countries higher happiness score have compared to Asia and Africa

In the future, we will analyze the effects in specific continents and demographic covariates on happiness score. It would also be interesting to model happiness between regions and build models to predict the score given population density and education level of citizens. The model performance can be improved by applying advanced techniques such as Decision tree and Random Forest. Also, Time Series and Forecasting methods can be to understand our data trend better.

Acknowledgment

The authors wish to thank Harrisburg Institue of Technology; Professor Pavana Ganesh, Harrisburg Institute of technology. This work was supported as a part of course curriculum of Master of Science Analytics in Computational Mathematics at Harrisburg Institue of Science and Technology

References

[1] John F. Helliwell, Richard Layard, Jeffrey D. Sachs, and Jan-Emmanuel De Neve "World Happiness Report". 2020

[2] "List of countires by contient" http://statisticstimes.com/geography/countries-by-continents.php. 2019

[3] Helliwell, J., Layard, R., & Sachs, J. "World Happiness Report" New York: Sustainable Development Solutions Network. 2019