STA304 Final Project ANNI LIN

Abstract

The data for this report is on the World Happiness Score on 2017. The data is provided by the website Kaggle, a large online data sharing platform where users from across the world can share useful data with each other. The content of this set of data contains happiness score for different country and related variables such as GDP, government performance, health, freedom and etc.

Introduction

The goal of this analysis is to explore the connection behind the global happiness score, what are the causes for a high happiness score and what factors contribute to the score. This analysis will be exploring the relationship between GDP versus happiness score, government corruption versus happiness core and health versus happiness score, I will also be observing how different countries are rated in terms of happiness score. Specifically, I am interested in finding out if higher GDP and better economy can increase the happiness score. Also interested in determining if a better government, a government with less corruption can also bring its people a higher happiness score, lastly, if better health can bring higher happiness score. For all three independent variables, government corruption, GDP and health, there will be three scatterplots constructed to illustrate the relationship, to demonstrate how GDP and government corruption value corresponds to the happiness score. Logistic regression will also be used to predict future growth of the happiness core when GDP increase or government corruption decreases. By using these methods and regression model, we will be able to tell and prove the cause and factors that are closely related to the happiness score. # Data

The content of Figure 1 is on the relationship between health and happiness score, I first use r code to plot a scatterplot, then I use the abline function to create a linear regression through the data points. Specifically, I want to explore if better health can lead to a higher happiness score. The variable introduced in this figure is health..life.expectancy and Happiness.score, both variables are numeric variable. The variable Health..Life.Expectancy. contains 156 observations, it represents the life expectancy value, which are all less than 1 and are in decimal form. The other variable Happiness score is the dependent variable, it also contains 156 observations, it represents the happiness score for a specific country and the max score value is 10. From Figure 1, there is a clear positive correlation between the two variables, so I have applied linear regression model to demonstrate more specific numeric details.

In terms of Figure 2, I am exploring the relationship between economy and happiness score, I pick the variable Economy..GDP.per.Capita. and Happiness.score to see if richer country gets an overall higher happiness score. Similarliy, I have constructed a scatterplot using the plot function, the variable Economy..GDP.per.Capita. contains the same number of observations as previous variables, which is also a numeric variable, it represents a country's GDP per capita. The other variable is happiness score, which is already introduced earlier, by using the scatterplot and linear regression model, I am able to determine what is the relationship between economy and happiness score.

In Figure 3, I selected the variable Trust..Government.Corruption. and Happiness.score, both variables are still numeric variable, both variables contain 156 observations. I will start by introducing Trust..Government.Corruption, this variables represents government corruption, which also means how terrible the government is. I want to explore how does corruption contributes to the happiness score, if a corrupt

government leads to lower happiness score. Since these are both numeric variables, I will still use scatterplot to see the correlation between the two variables. A linear line is also applied to get more information from the scatterplot.

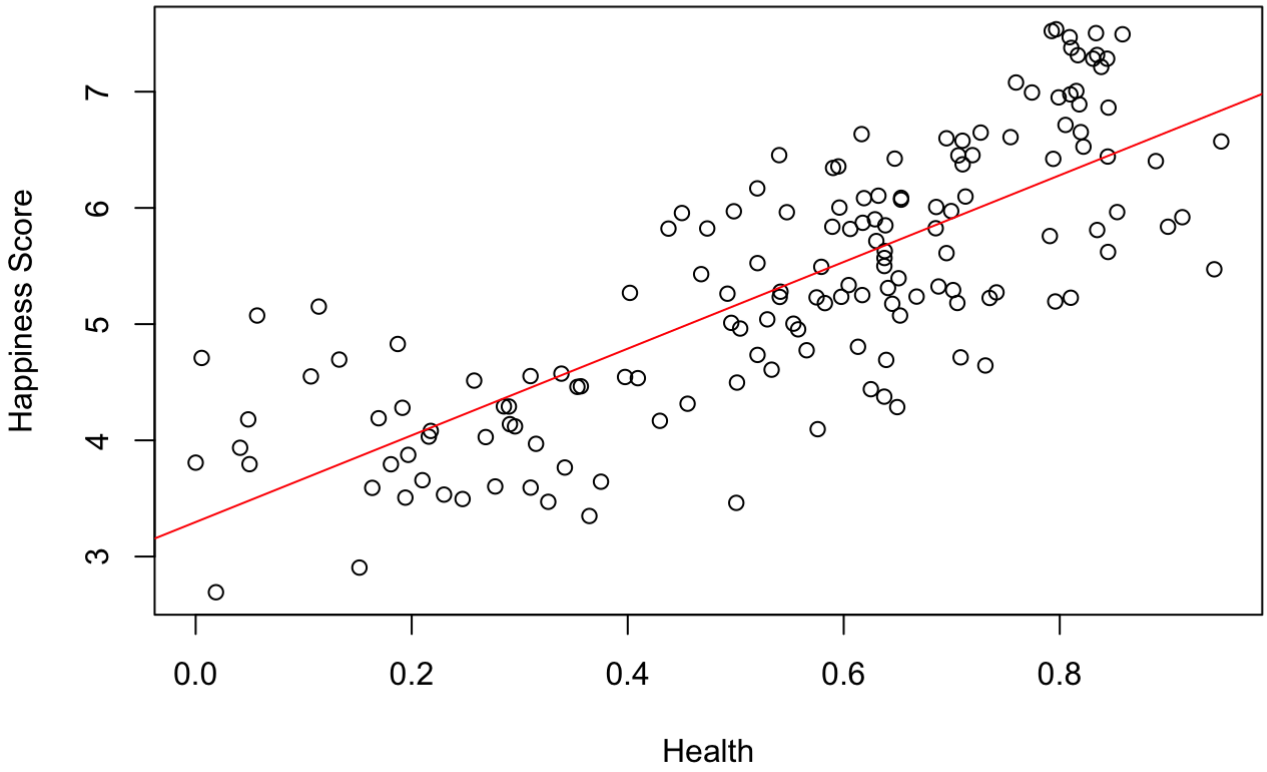
Model

The three independent variables I have chosen for my report is GDP, health and government corruption, I have constructed two different models to analyze and find the relationship that these variables have with happiness score. The first model is based on graph, for all three variables, I have constructed three scatterplots to find the correlation. Figure 1 is a scatterplot which contains data points combining health and happiness score, Figure 2 and Figure 3 are also scatterplots which reveals the relationship between GDP with happiness score and Corruption with happiness score. The reason why I chose to use scatterplot is because the variables that I am dealing with are all numeric, so it is better to fit them on a scatterplot and observe the pattern and trend.

In terms of the other method, I believe that there exists positive correlation between GDP and Happiness score, also a positive correlation on health and happiness score, and a negative correlation between corruption and happiness score. In order to demonstrate these relationships, I decided to use linear regression model, the explanatory variables in this case will be GDP, health and government corruption, whereas the response variable will be the happiness score. Base on the linear regression model, beta 1 will be the estimated slope of the regression line. I predict that beta 1 will be positive for the variable GDP and health since they should form an upward sloping line. On the other hand, the model should output beta 2 as a negative number since the correlation between corruption and happiness score should be negatively correlated and download sloping.

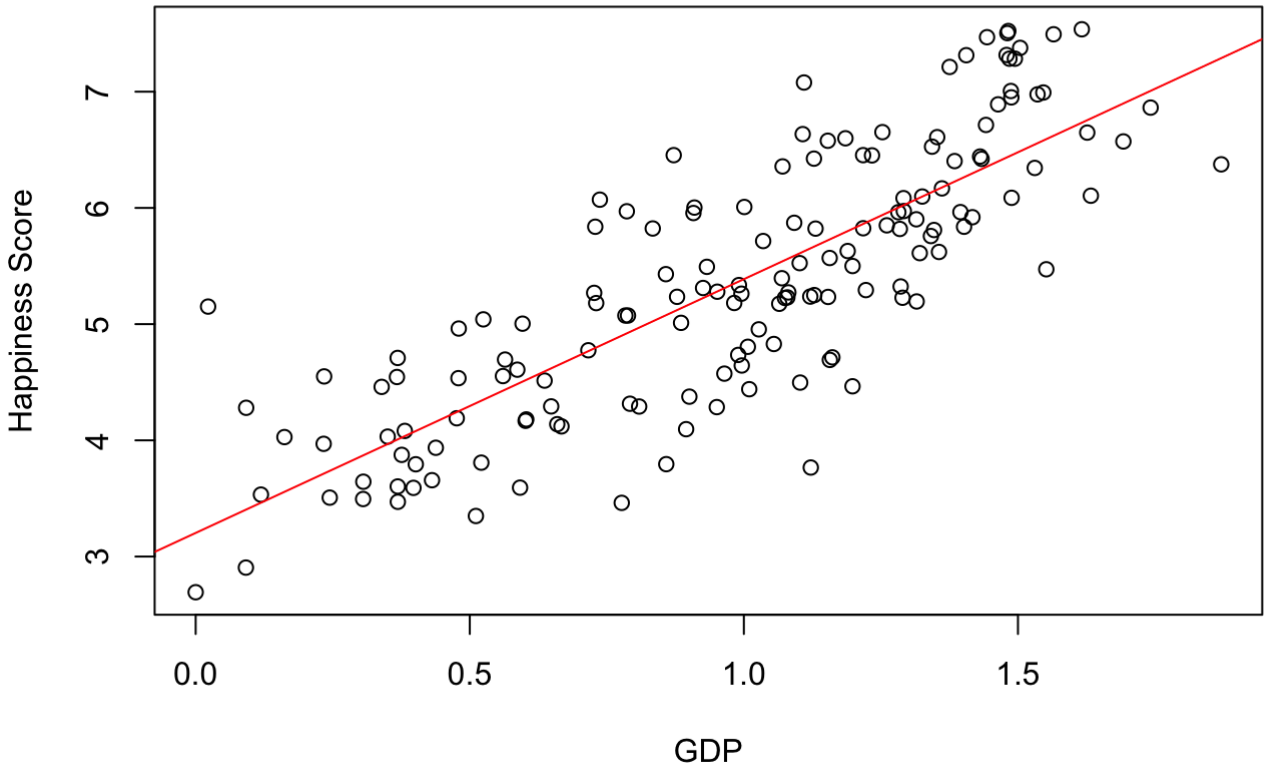
Results

Figure 1. Scatterplot of Health and Happiness Score



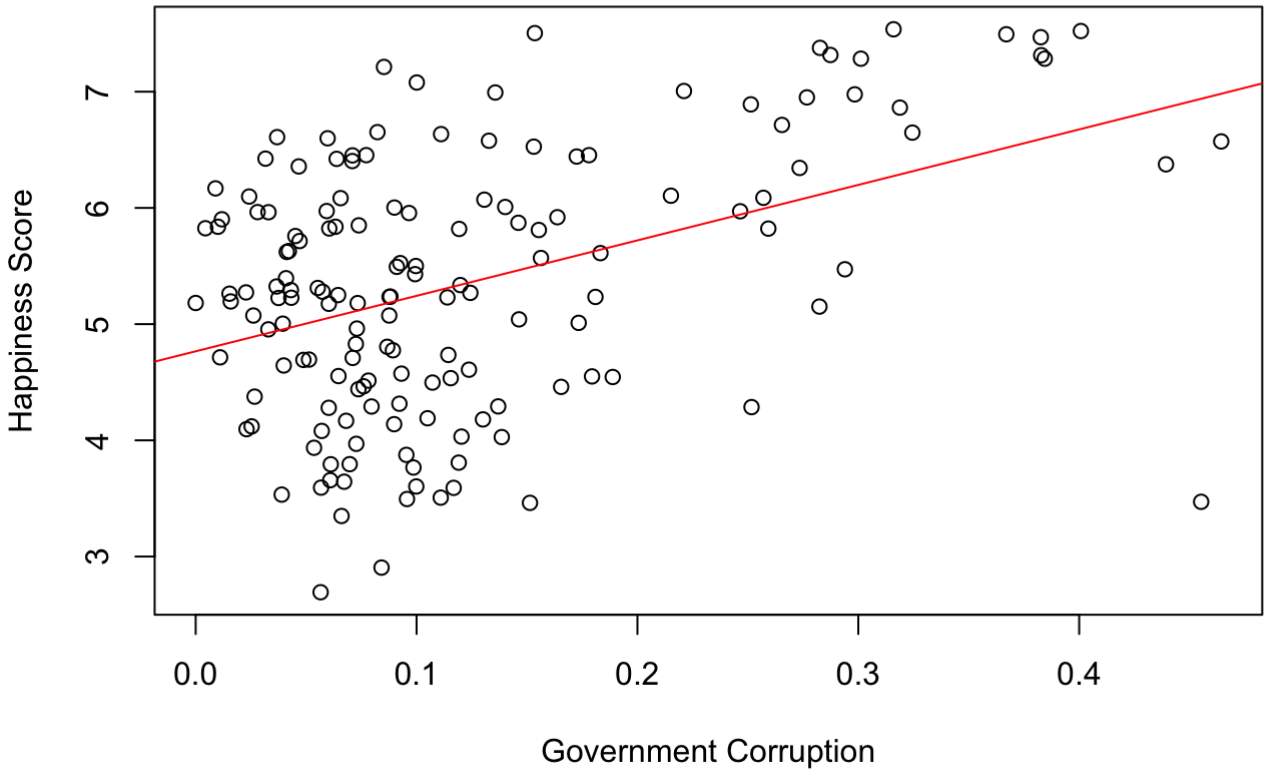

```
##
## Call:
## lm(formula = score ~ health)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70245 -0.52220 -0.03812  0.51574  1.56478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2969     0.1442   22.86   <2e-16 ***
## health        3.7312     0.2405   15.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7074 on 153 degrees of freedom
## Multiple R-squared:  0.6114, Adjusted R-squared:  0.6089
## F-statistic: 240.8 on 1 and 153 DF,  p-value: < 2.2e-16
```

Figure 2. Scatterplot of Economy and Happiness Score




```
##
## Call:
## lm(formula = score ~ gdp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88807 -0.45200 -0.05328  0.49425  1.89833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2032     0.1356   23.62   <2e-16 ***
## gdp           2.1842     0.1267   17.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6617 on 153 degrees of freedom
## Multiple R-squared:  0.6601, Adjusted R-squared:  0.6579
## F-statistic: 297.1 on 1 and 153 DF,  p-value: < 2.2e-16
```

Figure 3. Scatterplot of Government corruption and Happiness Score




```
##
## Call:
## lm(formula = score ~ gov)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4687 -0.7705  0.1210  0.7813  2.0398
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7662     0.1296  36.784  < 2e-16 ***
## gov           4.7746     0.8126   5.876 2.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 153 degrees of freedom
## Multiple R-squared:  0.1841, Adjusted R-squared:  0.1788
## F-statistic: 34.53 on 1 and 153 DF,  p-value: 2.538e-08
```

Discussion

Summary The purpose of my report and analysis is to explore the main contributors to a country's happiness score, to see what causes a country's happiness score to change. I have picked three meaningful variables to use as contributors and explore their relationship with happiness score. The variable Economy..GDP.per.Capita.,

Health..Life.Expectancy. and Trust..Government.Corruption. are all independent variables that I planned for the x axis, the dependent variable in this analysis is the variable happiness.score, which will be on the y axis. The reason why I chose GDP and health as my independent variables is because they are very essential to people's life, the third variable government corruption is also important since government contributes to lots of policies and rules that affect people's everyday life. After finalizing my three important explanatory variables, I use the data to create three graphic models to see the correlation and distribution, I have also used linear regression model to analyze the relationship between my independent and dependent variables. In the end, the results from the graphs and regression model will provide detail on the relationship each variable has with happiness score.

Figure 1 is about the relationship between Health life expectancy and happiness score. The data points from the scatterplot are mostly grouped and are moving upwards, there are no extreme values found. Base on the pattern and trend of the data points, we see that as health life expectancy value goes up, happiness score goes up as well, so a linear line can be fitted through the points and is able to well represent the trend and relationship. The red linear regression line has a upward slope, which means that as health value gets higher, happiness score value will also increase. Since most points can be fitted through by the linear regression line, it is evident that there exists a positive correlation between health and happiness score. Besides, a person's physical and mental health may be fundamentally connected (Goldberg, 2019), meaning that healthier body can leads to a healthier mindset, hence increases overall happiness. The regression model also supports the relationship, one increase in health life expectancy will cause the happiness score to go up by 3.7312.

Base on the scatterplot in Figure 2, we can see a clear relationship between GDP and happiness score. By looking at the spread of the data points, there is no clear sign of outliers, most data can be fitted on a straight line that is upward sloping. It is also clear that as GDP data gets larger and move to the right side of the graph, the corresponding happiness score also moves up in the graph. Which means that as a country's GDP per capital increases, the happiness score of that country will also increase. The article from Forbes supports this idea by saying that financial freedom is an important component of freedom in general (Zitelmann, 2020). Nowadays, wealth is an important part of everyone's life, wealth is not the only factor that brings happiness, but it is no doubt an essential factor that contributes to people's happiness. Being able to afford things that we like and not stressed by debt or financial obligation can indeed increase the sense of freedom in general. According to the red regression line and regression model from Figure 1, the slope is upward sloping and indicates a positive correlation between GDP and happiness score. Furthermore, when GDP per capita increase by 1, happiness score will go up by 2.1842.

The scatterplot from figure 6 is about the relationship between government corruption and happiness score, unlike previous scatterplot, there is no clear trend or pattern base on the spread of the data points. We can see many of the data points are scattered on the left portion of the graph, indicating that a less corrupt government does not always mean a higher happiness score. Looking at the x axis in the range of 0.0 to 0.1, this represents a very good government, however the happiness score is still spread across the y axis, meaning that even when there is a good government, happiness score is not always at its highest. It is true that there are many factors contribute to a person's sense of well-being, where one lives can be particularly important (loprespub, 2018). Government policies and rules does influence a person's life and a corrupt government can put extra problem on people's life. However, the impact of government corruption should not be as significant as the other two variables since government action only creates influences and does not directly control over a person's life. In comparison, health has a steeper linear slope than GDP, which means that health is the most direct variable that can leads to change in happiness score. Whereas the scatterplot for government corruption is not as well organized as the other two, hence can not really reflect the relationship.

Weaknesses

The weakness of my report is that the analyze of each variable could be more detailed and more in depth. Specifically, the last variable government corruption is not evident enough compare to the other two, the mathematical regression line suggests a beta value of around 4, which is confusing, and the overall spread appeared on the scatterplot is also not as clear.

Next Steps

One of the improvements that can be made is to introduce other type of graphs, throughout the analysis, the only type of graph that I am using is scatterplot, I should try to use other function to plot other type of graphs. Another improvement is that most of the data values are in decimal places and could be confusing to readers, I should manipulate the data values to a more suitable value, such as changing 0.7 to 70 percent. Last improvement that I will do next time is to find a larger data set, a data set that includes more observations and values.

References

Loprespub. “Well-Being in Public Policy: Should Governments Worry about Happiness?” HillNotes, 24 July 2018, hillnotes.ca/2018/03/20/well-being-in-public-policy-should-governments-worry-about-happiness/.

Costa, Anita. “5 Ways Being Healthy Makes You Happy.” Best Health Magazine Canada, Best Health Magazine Canada, 16 July 2019, www.besthealthmag.ca/best-you/mental-health/5-ways-being-healthy-makes-you-happy/.

Zitelmann, Rainer. “Does Money Make People Happy?” Forbes, Forbes Magazine, 2 July 2020, www.forbes.com/sites/rainerzitelmann/2020/07/06/does-money-make-people-happy/?sh=70c65f15574d.

Appendix

GitHub Link: