INTRODUCTION

The factors that contribute to happiness have long been a controversial topic. Does having more money mean more happiness, and if so, to what extent? What about health; does better health lead to a greater level of happiness? What is more important, money or health? And what about having support from friends and family? We wanted to embark on a quest to understand the factors that lead to happiness. To this end, this was our research question: **What factors contribute meaningfully to one's happiness, and to what degree?**

We decided to use data from the 2019 World Happiness Report to perform our analysis. The World Happiness Report is an annual publication from the United Nations Sustainable Development Solutions Network that ranks global happiness in countries around the world. The report measures happiness based on several key variables, including GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. These factors are used to assess the overall happiness and well-being of the populations of over 150 countries, with survey data primarily sourced from the Gallup World Poll.

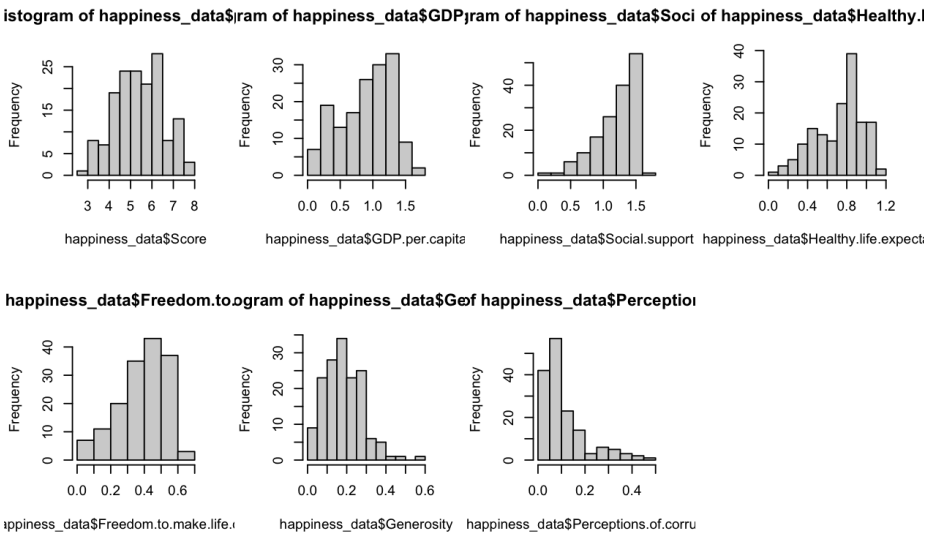
We obtained the data from Kaggle. The data can be found at this link: https://www.kaggle.com/datasets/sougatapramanick/happiness-index-2018-2019?select=2019.csv

As we wanted to explore if there were any relationships, particularly linear relationships, between our predictor variables with Happiness Score, we decided that a multiple linear regression model would work best.

This paper will begin with exploratory data analysis, followed by a review of the linear models we fit to describe our data, culminating in choosing the "best" predictive model. Finally, this paper will discuss the real-world implications and limitations of our final model.

DATA DESCRIPTION

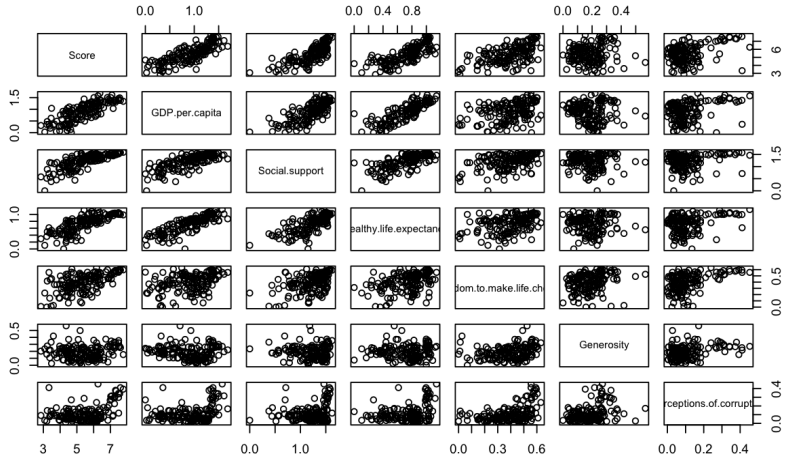
The distribution of each of the variables in our dataset was visualized using the histograms below:



Most of the variables showed deviation from the normal distribution based on the skewing of the data.

The mean and standard deviation of each variable in the data is stated in the table below:

	Score (response variable)	**GDP Per Capita (predictor)**	**Social Support (predictor)**	**Healthy Life Expectancy (predictor)**	**Freedom to Make Life Choices (predictor)**	**Generosity (predictor)**	**Perceptions of Corruption (predictor)**
Mean	5.407	0.905	1.209	0.725	0.393	0.185	0.111
SD	1.113	0.398	0.299	0.242	0.143	0.095	0.095

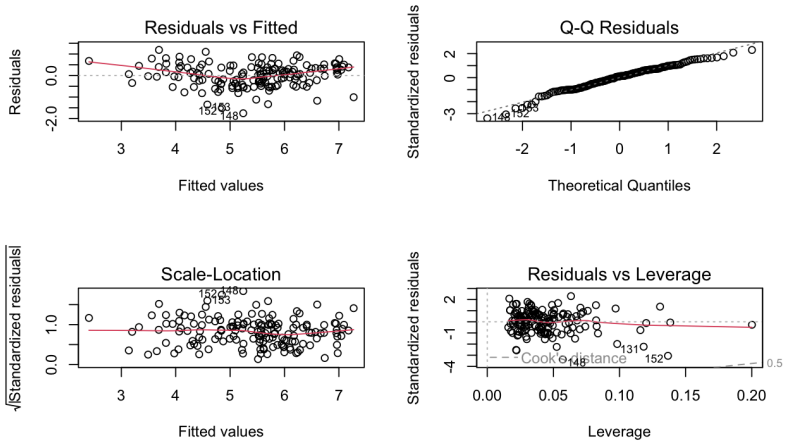


Based on our scatterplot matrix, we can see that there is some correlation between the Score response variable and most of the predictor variables. The predictor variables Generosity and Perceptions of Corruption do not show a strong correlation with Score, which is something we further explored when evaluating our potential models.

RESULTS AND INTERPRETATION

We began by generating the full linear regression model for our data before any transformations. Our original model is represented by the equation:

Happiness Score = 1.7952 + 0.7754 (GDP per Capita) + 1.1242 (Social Support)
+ 1.0781 (Healthy Life Expectancy) + 1.4548 (Freedom to Make Life Choices) + 0.4898 (Generosity)
+ 0.9723 (Perceptions of Corruption)

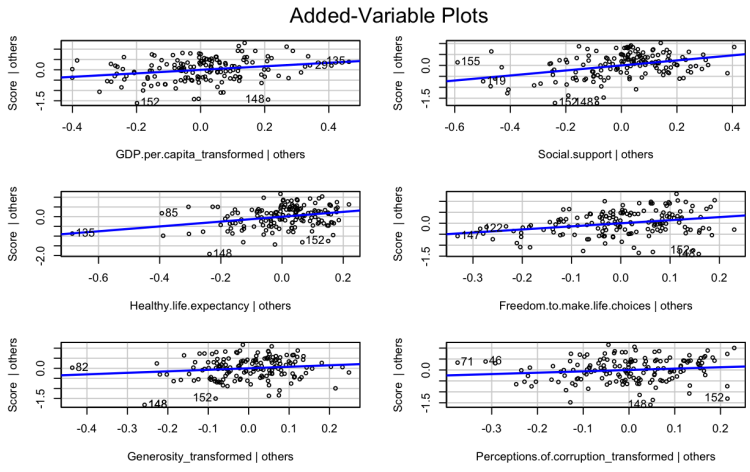


Overall, it seemed like the diagnostic plots performed well, except for the fact that there seems to be a slight fan shape in the residuals vs fitted plot, and the residuals seem to have a slight left skew from the Q-Q plot. We attempted to resolve these concerns by using the power transformation which resulted in a transformed model:

$$\text{Happiness Score} = 1.0832 + 0.8546\,(\text{GDP per Capita})^{0.73} + 1.1518\,(\text{Social Support}) + 1.2529\,(\text{Healthy Life Expectancy}) + 1.4083\,(\text{Freedom to Make Life Choices}) + 0.7509\,(\text{Generosity})^{0.33} + 0.6244\,(\text{Perceptions of Corruption})^{0.33}$$

In both our original and transformed models, however, two predictor variables, Generosity and Perceptions of Corruption, were determined to be statistically insignificant because their p-values were greater than 0.05. We were then interested in assessing if multicollinearity exists; we found that none of the VIFs were greater than 5, and the only correlation above 0.8 was between GDP.per.capita_transformed and Healthy.life.expectancy. From this, we concluded that multicollinearity was not an issue with our model.

We were then interested in visualizing the pure effect of each predictor on our response variable, happiness score:



The added variable plots of the transformed model confirmed that Generosity and Perceptions of Corruption did not have a large influence on the Score based on the more gradual slope of the lines. We

then performed variable selection to decide on our final model. We began with the method of considering all possible subsets.

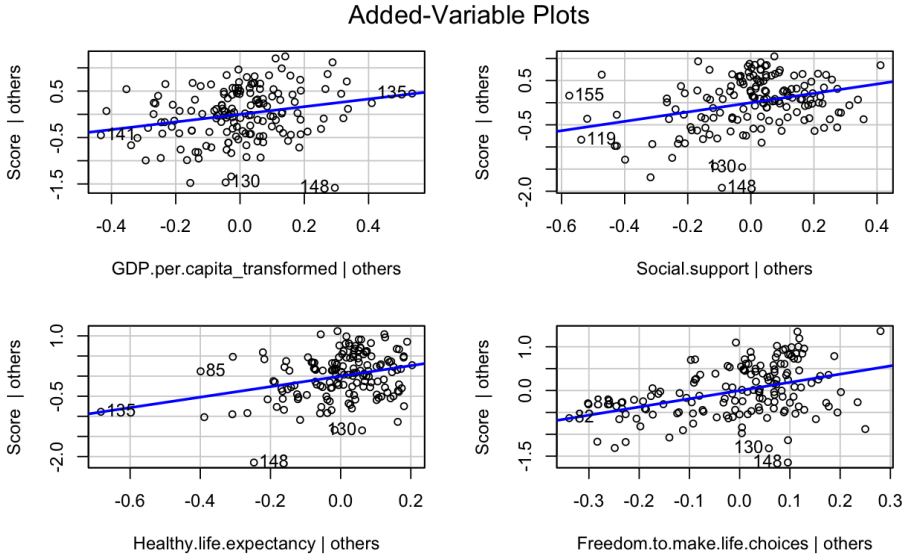
	GDP.per.capita_transformed	Social.support	Healthy.life.expectancy	Freedom.to.make.life.choices	Generosity_transformed	Perceptions.of.corruption_transformed
1	(1) " "	" "	"*"	" "	" "	" "
2	(1) " "	"*"	"*"	" "	" "	" "
3	(1) " "	"*"	"*"	"*"	" "	" "
4	(1) "*"	"*"	"*"	"*"	" "	" "
5	(1) "*"	"*"	"*"	"*"	"*"	" "
6	(1) "*"	"*"	"*"	"*"	"*"	"*"

After constructing the new subsetted models, we ran metrics on all 6 models to determine the most fit models based on their adjusted R^2 values, AICs, AICcs, and BICs. The model with all 6 predictor variables performed the best for 3 out of the 4 metrics; however, the model with 4 predictors performed best for BIC. In the subsetted 4 variable model, all variables were statistically significant, so we decided to further explore this model to determine if it would be the best fit for the data.

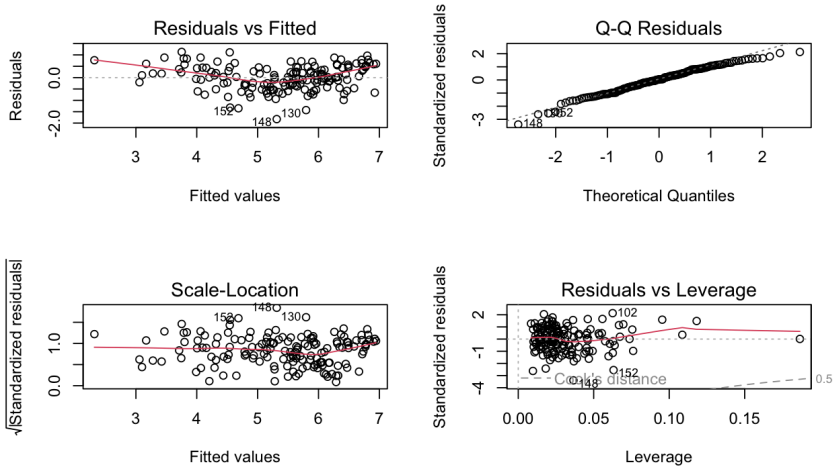
After running forward and backward stepwise variable selection with BIC and AIC as the metrics, we found that the same 4 variable model from our subsetting was also the model determined from our forward and backward BIC variable selections. Based on its performance, we decided that the 4 variable model was the best model for the data.

Happiness Score = 1.7004 + 0.8241(GDP per Capita)$^{0.73}$ + 1.0539 (Social Support) + 1.3075 (Healthy Life Expectancy) + 1.8777(Freedom to Make Life Choices)

Here are the added variable plots of our final model:



Comparing the added variable plots of the 4 variable model vs those of the full model, we saw that each variable in the 4 variable model has a significant impact on Score. In the full model, however, the slopes for the Score vs Generosity and Score vs Perceptions of Corruption were significantly less steep.



The diagnostic plots from our final four variable model showed similar results, but a slight improvement from our original model, supporting the assumption that this is the best-fit model.

DISCUSSION

In this project, we investigated how happiness would be measured based on several key variables, including GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption.

Our original full model showed potential for non-constant variance and a slightly left-skewed error term, and also statistical insignificance of the Generosity and Perceptions of Corruption predictors. Therefore, we first decided to transform the data using the Box-Cox method; the diagnostic plots of this model were not much better than our original model, and we then decided to focus on variable selection. We employed various methods for variable selection, including considering all possible subsets, using forward stepwise regression using AIC and BIC, and using backward stepwise regression using AIC and BIC. Our results pointed to either the 6 variable model or the 4 variable model. We then found that Generosity and Perceptions of Corruption were statistically insignificant in the 6 variable model, whereas all variables were significant in the 4 variable model. This was confirmed by the added variable plots of the 4 variable model, where each plot clearly seemed to have a positive slope.

In summary, through exploring the data, conducting transformations, and selecting variables, we found the factors most influential to happiness: GDP per capita, Social Support, Healthy Life Expectancy, and Freedom to Make Life Choices, concluding on a 4-variable model being the best predictive model of happiness. It was interesting that perceptions of corruption and generosity were deemed statistically insignificant; we unfortunately were unable to find any articles online about why this might be the case.

This final model has real-world implications, as it coincides with the strategies policymakers have implemented to improve happiness. Governments often invest in people's education to increase GDP per capita, enhance social welfare programs to increase social support, encourage preventative healthcare measures to increase healthy life expectancy, and enforce human rights legislation to protect civil liberties and ensure freedom to make life choices. While there is no official, objective path to a country's happiness, governments and policymakers leverage happiness research as a guide to increasing overall happiness in their country.

However, our model and dataset, and thus analysis, have limitations. As mentioned, happiness is subjective, which makes it difficult to measure without bias and probes the question as to whether any research is truly "accurate". One factor contributing to this subjectivity is cultural differences. Traditions in culture may cause differences in how each country views the indicators of happiness (e.g. freedom to make life choices). As a result, how happiness is defined for one culture may not apply to another (e.g., individualistic vs collective countries). Thus, calculations and measurements for such subjective indicators may lead to bias in the model, emphasizing the countries whose cultures are more in line with GDP per Capita, Social Support, Healthy Life Expectancy, and Freedom to make life choices. This means that we may not be truly predicting the happiness of all countries, limiting the universality of our research and analysis. Additionally, the time of data may also result in limitations to our model. The data we modeled was from 2018 - 2019. From that time, there may have been changes to happiness data. Moreover, the pandemic, in addition to how countries may have implemented policies within the past 6 years, may require a time-lag consideration to assess their impact on happiness. More recent data should be collected to understand how relevant the studied indicators, and other external factors, are to happiness today.

Ultimately, despite the limitations and restricted generality of happiness research, it still lays the foundation for understanding the key to such a crucial aspect of the world.