Analysis of the World Happiness Report

https://github.com/yatusabe123/World-Happiness-Report-Analysis

Ameya Gokhale, Rui Yu, Yuhang Diao, Priyanka Balaji Ramanathan, Jayesh Lalwani

Summary:

The United Nations releases its highly influential World Happiness Report on an annual basis. It ranks 155 countries according to their happiness score. The citizens of each country ae asked to respond to the "Cantril Ladder," a survey that analyzes how happy an individual is in addition to how they view their country's GDP per capita, Family, Life Expectancy, Freedom, Generosity and Trust (in government). The study considers a fictional country "dystopia" to be a baseline observation with the worst rating for all six features. Subsequently, each country is compared to dystopia in terms of these features. The impact of each of these features on a country's happiness score is then determined. This project is centered around the World Happiness Report dataset from Kaggle that includes data for the years 2015 through 2019.

In the initial part of this project, various exploratory data analysis (EDA) visualizations were constructed to analyze for trends in the data. An example of one of the visualizations in this project, the relationship between happiness score and each of the six individual predictors was graphed. The EDA visualizations were conducted in R as it provides the best environment for such tasks. First however, each of the datasets from the five years needed to be aggregated and subsequently tidied. The main benefit of tidy data for this project proved to be the consistency that it offered that made operating on it simpler. Some data was incorporated from the "Gapminder" dataset, which is generated by a not-for profit startup in Sweden with the aim of promoting global development through statistical analysis. Additional predictors such as the labor force participation of a country were added from this Gapminder dataset to the aggregated data from Kaggle. Although these additional predictors benefited the entirety of the project, for the purpose of the EDA section, it helped to establish more trends between happiness score and additional variables, allowing for the project to have a more holistic visual understanding of happiness.

The latter part of this project centered around modelling. First, backward stepwise model selection in R was used to determine which combination of the predictors (including those incorporated from the Gapminder dataset) were most effective in predicting the happiness score of a country. In backward elimination, the program eliminates unnecessary variables, one step at a time to arrive at an optimal model with the most effective indicators. Secondly using python, simple regression was used on the data from 2015-2018 to predict the happiness values for 2019. Additionally, in a similar fashion, the regression techniques of LASSO and Ridge regression were used to see if they would provide more optimal models. All the regression models were compared using the mean squared error as a metric to determine which was the most optimal in modelling this data.

Methods:

The technical description of the data management and processing methods are divided into several sections below.

Tidying data (for EDA):

After the data was read into R, each column name had to be standardized as they were slightly different across in each year's data frame. The data was then combined into one singular data frame. The files in the latter years had the region and country combined in one column (although the value was the country name), hence filling up the "Region" with the "Region" value of the same country from previous years. This was corrected in the tidied data as well. The "NA" values were replaced with the average values from the column. It would have been counterproductive to delete observations (countries) with some "NA" data as this would have deleted many observations that had values for other indicators. This would have had a detrimental impact on the visualizations for these other indicators.

Exploratory Data Visualizations:

Using ggplot2, a variety of different visualizations were conducted. Ggplot2 is a graph-focused package in R that is common amongst data scientists to visualize different relationships in data. These visualizations are listed below:

1. Relationship between each indicator and happiness score
2. Regions and countries with the highest happiness score
3. Relationship between additional indicators from the Gapminder data set and happiness score
4. Heat maps of the data to show correlation between each indicator and happiness scores

Stepwise Model Selection and Gapminder data integration:

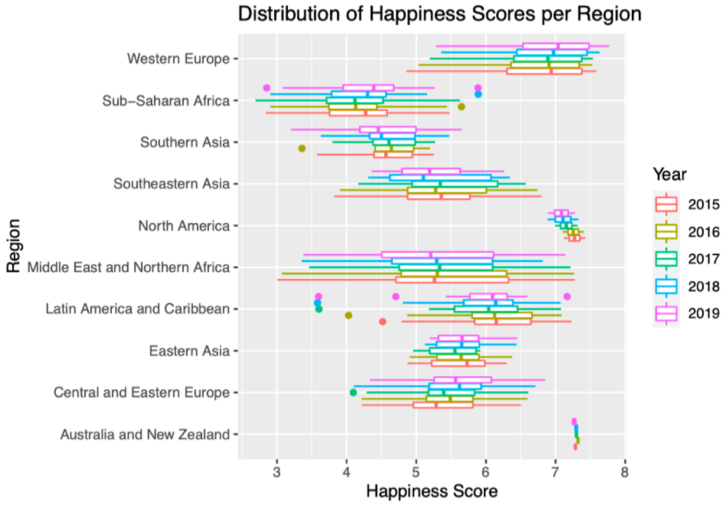
Using SQL-join techniques in R, the Gapminder data was integrated into the original dataset from Kaggle. Initially, the statistical significance for the model with each variable was determined. Then a backward elimination stepwise model selection was conducted with the entire model to remove unnecessary variables. Lastly, model diagnostics were performed on this final model to verify that it did not violate any model assumptions.

Regression Techniques:

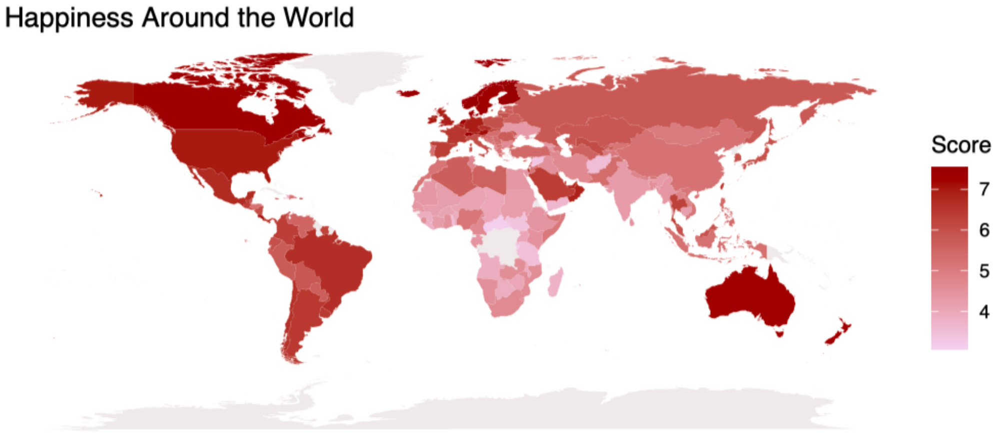
The regression techniques were carried out in python. The data had to be tidied separately in python as the goals of this portion of the project were slightly different. For the years 2015 and 2016, the last two columns were not aligned with their respective columns in the other three datasets, using the pandas package, these columns were swapped to maintain uniformity in positionality amongst all the indicators. The data from 2015 to 2018 was combined into one pandas data frame, as this is considered the "training data" or the data on which the regressions are performed. The data from 2019 is the "test data," or the dataset upon which the learned regression (from the training set) is applied to determine its effectiveness. The "Sklearn" package in python facilitates many such regression techniques. Three regression techniques were used from this Sklearn package- standard OLS, ridge and lasso regression. Ridge and Lasso regression are regularization regression techniques that have potential to solve for overfitting that may be caused by a small amount of available data. These regressions heavily penalize outliers as well, so they are a good way of accounting for them. The regressions were run on the combined dataset consisting of data from 2015 to 2018, and subsequently used to predict the happiness scores of countries in the year 2019. The mean squared error for all three of these regression techniques was calculated to determine which one of them was the most optimal regression.

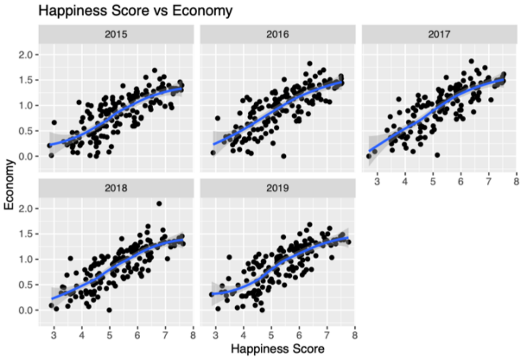
Results:

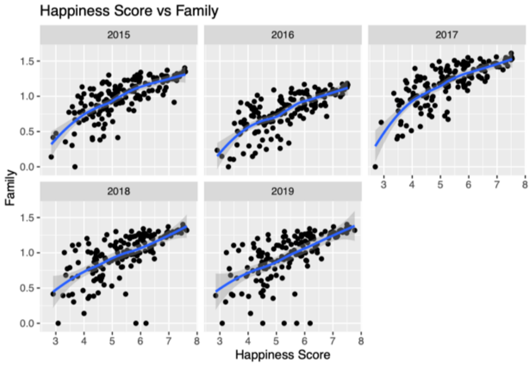
Visualizations:

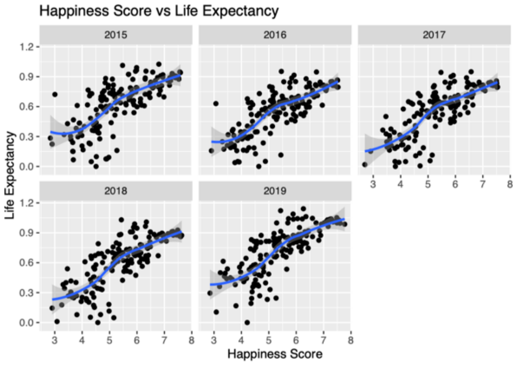


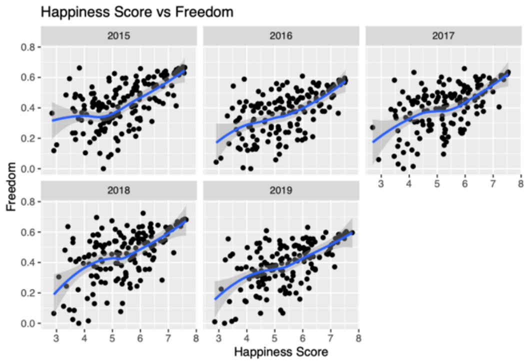
From both these visualizations, we can see that countries in Western Europe, Oceania and North America have consistently high happiness scores. Similarly, countries in Africa, Southern Asia have lower happiness scores. This ties into the logical assumption that countries that are more developed have higher happiness scores. Middle East and Northern Africa have countries that range between both ends of the happiness score spectrum which is intuitive as some countries in the Middle East are very developed.

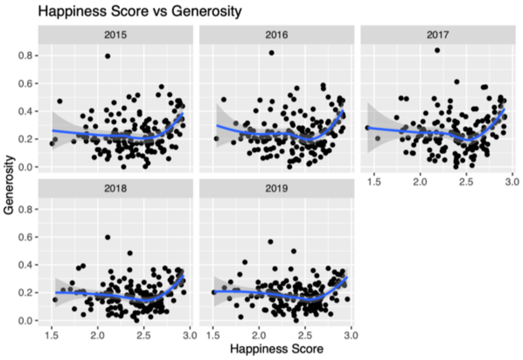


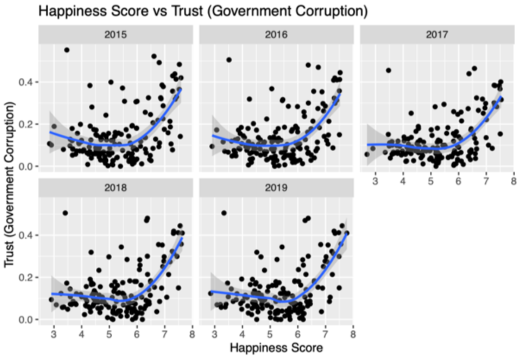

Happiness Score vs Economy
2015
2016
2017
2018
2019
Economy
2.0
1.5
1.0
0.5
0.0
3
4
5
6
7
8
Happiness Score

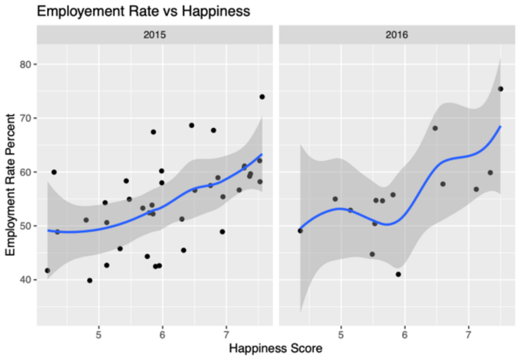

Happiness Score vs Family
2015
2016
2017
2018
2019
Family
1.5
1.0
0.5
0.0
3
4
5
6
7
8
Happiness Score

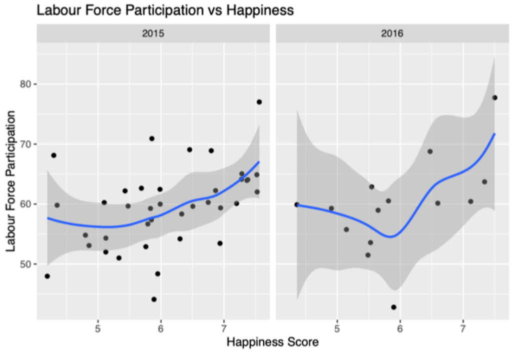

Happiness Score vs Life Expectancy
2015
2016
2017
2018
2019
Life Expectancy
1.2
0.9
0.6
0.3
0.0
3
4
5
6
7
8
Happiness Score


Happiness Score vs Freedom
2015
2016
2017
2018
2019
Freedom
0.8
0.6
0.4
0.2
0.0
3
4
5
6
7
8
Happiness Score


Happiness Score vs Generosity
2015
2016
2017
2018
2019
Generosity
0.8
0.6
0.4
0.2
0.0
1.5
2.0
2.5
3.0
Happiness Score

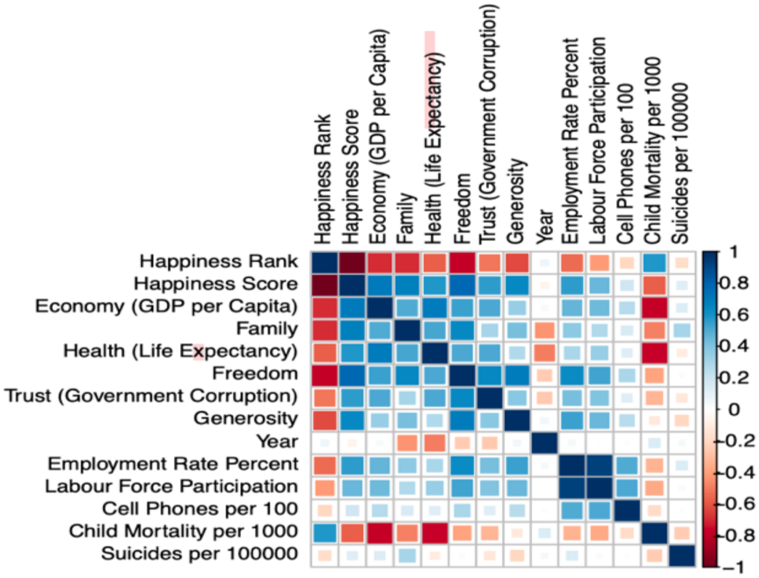

Happiness Score vs Trust (Government Corruption)
2015
2016
2017
2018
2019
Trust (Government Corruption)
0.4
0.2
0.0
3
4
5
6
7
8
Happiness Score


Employement Rate vs Happiness
2015
2016
Employment Rate Percent
80
70
60
50
40
5
6
7
Happiness Score


Labour Force Participation vs Happiness
2015
2016
Labour Force Participation
80
70
60
50
5
6
7
Happiness Score

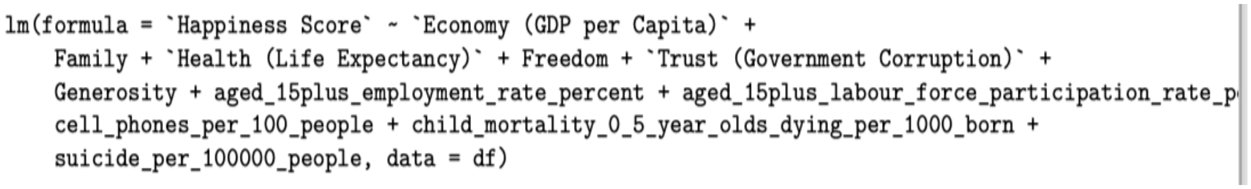
Observations regarding the relationship between each indicator and happiness score

1. There is a positive linear relationship between the economy and the happiness score for a country over each year.
2. There is a positive linear relationship between the family and the happiness score for a country over each year.
3. There is a positive linear relationship between the life expectancy and the happiness score for a country over each year.
4. There is a positive almost-linear relationship between the life expectancy and the happiness score for a country over each year. In some years (2015, 2018) the relationship is slightly curved.
5. There is no linear relationship between generosity and the happiness score for a country over each year.
6. There is a relationship between the trust in government over the years. However, the relationship is not linear.
7. There is a weak positive relationship between employment rate and happiness.
8. There is a weak positive relationship between labor force participation and happiness.

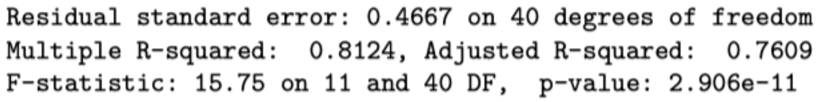


Above is a heat map that displays the correlation between different variables. One can see that there is little to no correlation between generosity and the happiness score which is consistent with previous graphs. Freedom to make life choices did not have a significant correlation with the happiness score in comparison with other factors such as the economy. There appears to be a negative correlation between child mortality and happiness score i.e., the lower the child mortality rate, the higher the happiness score. An interesting fact is that the number of suicides does not have much correlation with the happiness score (in fact it is almost positive). The employment rate percent and the labor force participation have positive correlation with each other and hence for all intents and purposes of statistical inference can be taken as one attribute. Economy appears to have a very high negative correlation with child mortality. This heat map is only for one year, but every year seems to show the same correlations.

Model Selection and Regression:

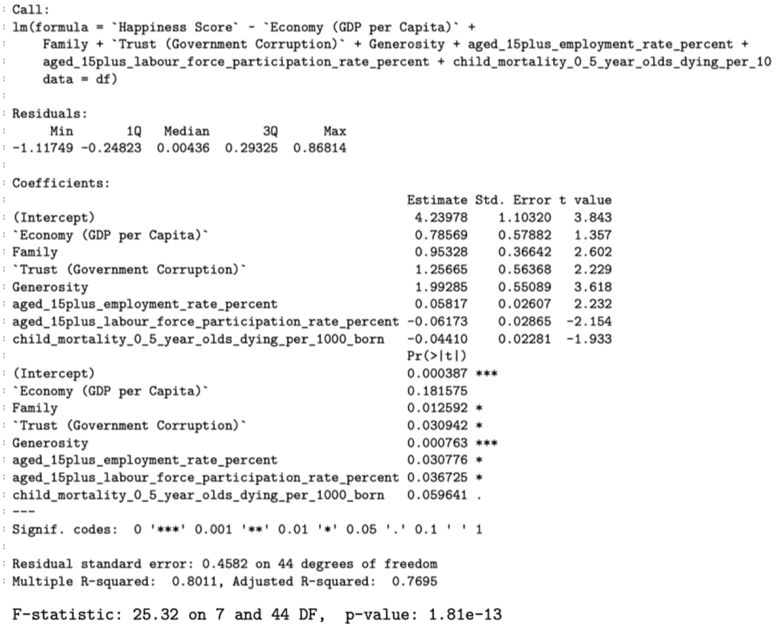


```
lm(formula = `Happiness Score` ~ `Economy (GDP per Capita)` +
    Family + `Health (Life Expectancy)` + Freedom + `Trust (Government Corruption)` +
    Generosity + aged_15plus_employment_rate_percent + aged_15plus_labour_force_participation_rate_p
    cell_phones_per_100_people + child_mortality_0_5_year_olds_dying_per_1000_born +
    suicide_per_100000_people, data = df)
```



```
Residual standard error: 0.4667 on 40 degrees of freedom
Multiple R-squared:  0.8124, Adjusted R-squared:  0.7609
F-statistic: 15.75 on 11 and 40 DF,  p-value: 2.906e-11
```

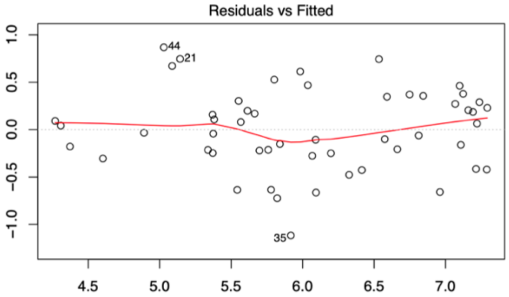
Above is the naïve model with every feature from the dataset included. The p-value of 2.906e-11 appears to be significant, indicating that the model is useful. The R-squared is 0.8124, which means that 82.24% of the variance is explained by this model. The adjusted R-squared is 0.7609 indicating that after penalizing for several parameters, 76.09% of variance is explained by this model.

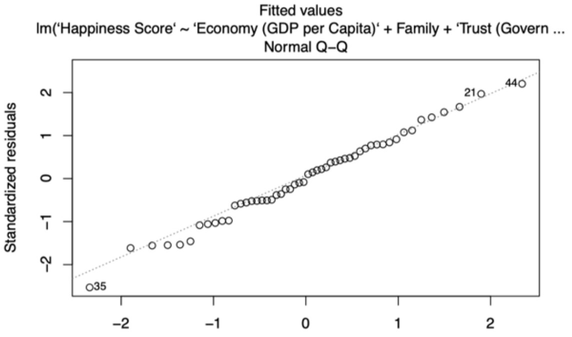


```
: Call:
: lm(formula = `Happiness Score` ~ `Economy (GDP per Capita)` +
:     Family + `Trust (Government Corruption)` + Generosity + aged_15plus_employment_rate_percent +
:     aged_15plus_labour_force_participation_rate_percent + child_mortality_0_5_year_olds_dying_per_10
:     data = df)
:
: Residuals:
:      Min       1Q   Median       3Q      Max
: -1.11749 -0.24823  0.00436  0.29325  0.86814
:
: Coefficients:
:                                                      Estimate Std. Error t value
: (Intercept)                                           4.23978    1.10320   3.843
: `Economy (GDP per Capita)`                            0.78569    0.57882   1.357
: Family                                                0.95328    0.36642   2.602
: `Trust (Government Corruption)`                       1.25665    0.56368   2.229
: Generosity                                            1.99285    0.55089   3.618
: aged_15plus_employment_rate_percent                   0.05817    0.02607   2.232
: aged_15plus_labour_force_participation_rate_percent -0.06173    0.02865  -2.154
: child_mortality_0_5_year_olds_dying_per_1000_born   -0.04410    0.02281  -1.933
:                                                     Pr(>|t|)
: (Intercept)                                         0.000387 ***
: `Economy (GDP per Capita)`                          0.181575
: Family                                              0.012592 *
: `Trust (Government Corruption)`                     0.030942 *
: Generosity                                          0.000763 ***
: aged_15plus_employment_rate_percent                 0.030776 *
: aged_15plus_labour_force_participation_rate_percent 0.036725 *
: child_mortality_0_5_year_olds_dying_per_1000_born   0.059641 .
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 0.4582 on 44 degrees of freedom
: Multiple R-squared:  0.8011, Adjusted R-squared:  0.7695

 F-statistic: 25.32 on 7 and 44 DF,  p-value: 1.81e-13
```

After doing the stepwise selection, the model with the lowest AIC values included the variables : Economy (GDP per Capita), Family, Trust (Government Corruption), Generosity,aged_15plus_employment_rate_percent,aged_15plus_labour_force_participation_rate_percent, and child_mortality_0_5_year_olds_dying_per_1000_born. After doing variable selection on the naive model, the new model is significant with a p-value of 1.81e-13. T, the R-squared is 0.8011, which means that 80.11% of variance is explained by this model. The adjusted R-squared is 0.7695 indicating that after penalizing for several parameters, 76.95% of variance explained by this model.





The Q-Q plot for residuals appears to be linear, and the Residuals vs Fitted plot does not appear to show any pattern indicating that model assumptions have not been violated.

For the regression part of the project, the 2015-2018 data was used as the training and the 2019 data was the test set.

	Training Error (MSE)	Test Error (MSE)
Simple linear regression	0.30087982427050036	0.3024847268723013
Ridge regression	0.3008827702023948	0.3024921406931356
Lasso regression	0.5183784111663383	0.49254191751994514

The errors for both the training and test sets appear to be in the same range, indicating that there is no overfitting. As there is no overfitting on the simple linear regression, it appears that regularization is unnecessary. This would explain the fact that the ridge regression seems to be slightly worse in terms of being an effective model in comparison to simple linear regression. Lasso regression appears to be the worst of the three models with the highest MSE.

Ridge regression and Lasso regression are both regularization models that work to rectify issues caused with overfitting. Overfitting occurs when a model works very well on the training data, but poorly on the test data (the data that matters as it determines if the prediction capabilities of the model are effective). This is one of the reasons that these two regressions are ineffective here. There seems to be enough data to avoid overfitting. A general rule of thumb is that if there are more predictors than observations, the lack of data for the regression models will result in overfitting. Here, there are over 600 observations in the combined dataset, and only six indicators, so regularization techniques are unnecessary. Ridge regression is done with calculating L-2 norms and works better if there are many parameters that have similar values (like this dataset). Lasso regression on the other hand is done by incorporating L-1 norms and works well when many predictors are insignificant (Lasso can essentially zero out the coefficients). Therefore, this explains why Lasso regression performs the poorest on this dataset.

Note that the data frame that included information from Gapminder was not used for this as there were very few values (only for the year 2015). So, there would have been no set to use as a test set, instead there would have been a need to divide what few values were present into a train and test set. After experimenting with this and deciding that the models were insignificant and not worth including in the report, they were excluded from the report.

Discussion:

The main goal of this project was to highlight how machine learning and statistical analysis techniques can be used congruently in data science with EDA techniques. For example, in this project, one can interpolate the EDA data and make an accurate guess on what the results of the model selection will be. It is crucial to both visualize trends in datasets so that one can effectively model them and vice-versa. It makes the data analytics process more efficient by allowing one to have prior knowledge about a dataset and thus approach it with a more concise and targeted approach. For example, if the graph visualizations show that there are a great many outliers in a dataset, one can assume that regularization techniques will be necessary to avoid overfitting of the data when modelling it. This allows one to have a more effective and targeted approach when considering regression techniques. Another example is related to the concept of feature selection. Feature selection in a model is the process of removing undesired features to better optimize the model and make it computationally more efficient. Visualizations such as the ones that determine the relationships between individual predictors and the response variable have potential to help with this. Relationships that show a higher visual correlation may be considered as optimal features in the feature selection process.

Although the data science concepts that are highlighted in this project can be used in a variety of capacities, such matrimony of regression and visualization techniques has applicability in the social sciences. The World Happiness Report is highly influential and is used by policymakers in the United Nations and individual countries. The Gapminder startup aims to improve world development through statistical analysis. Policymakers can take the concepts used in this project, apply them at a more advanced level and use them to promote development across the globe. For example, one can see with the visualizations what areas of the world lack in happiness. As intuition would suggest, they are countries that are not as developed. Policy makers can target these areas. They have visualizations and regression data to know/determine what factors are common amongst countries with higher happiness scores. Their policy decisions can thus be data driven and place more emphasis on certain areas. For example, the results from the project show that the economy is the variable with the highest correlation towards a country's happiness score. Policy makers can place more emphasis on improving the economies of such countries to maximize the happiness of their citizens. The regression models, once effectively optimized, can be used to forecast the effects of the policy makers' decisions and set goals on a year or decade basis. This forecasting has a wide variety of uses; perhaps one of the most important being the ability to see if certain set goals have been met after an amount of time has passed.

The dataset is lacking in quantity, which hinders the regression process. Perhaps more data can be incorporated into work that is done in this area. More data is not restricted to more observations, but more features or indicators as well. In the real world however, this is already being done. Gapminder, which does work in this field operates on large amounts of data. Perhaps more complicated supervised machine learning techniques can be used as well to more effectively model the data.

Statement of Contributions:

Ameya Gokhale: Wrote the report, Python regression
Rui Yu: Combining Gapminder data and perform model selection
Yuhang Diao: Combining Gapminder data and perform model selection
Priyanka Balaji Ramanathan: Tidying data and EDA visualizations
Jayesh Lalwani: EDA visualizations

References:

1. https://www.kaggle.com/unsdsn/world-happiness
2. https://www.gapminder.org/data/

Appendix:

Note the relevant code is not added here; however, the GitHub link to the repository for the project is included in the title section on the first page. There are three files, one for the visualizations, one for the model selection and one for the regression. There are some redundancies in the code (the sections that involve the tidying/merging of data) as the work was divided and different parts of the project called for slightly different data frames.