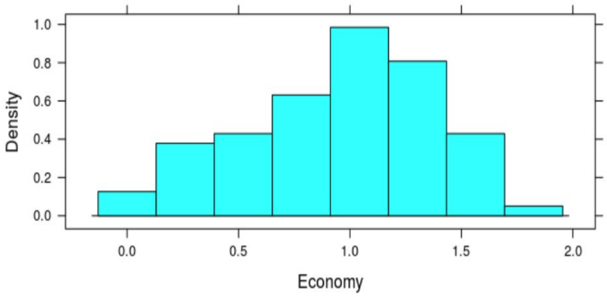
DSC 205 Project #2 Report

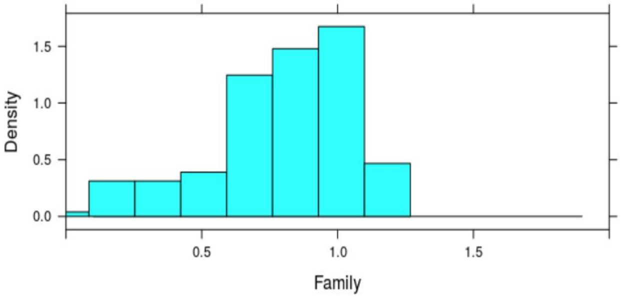
Introduction and Background:

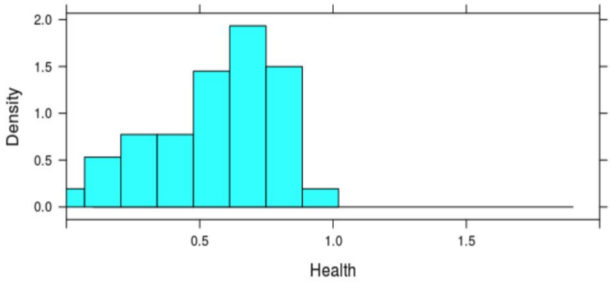
In a world where anxiety and depression rates seem to be increasing, more and more people are trying to figure out where happiness can come from. While some people believe that happiness is derived from health, others find that it can come from a sense of freedom or a strong economy. In order to determine where happiness comes from, the economic strength, families, life expectancy, perception of freedom, trust in one's government, and a country's generosity were looked at in association with various countries happiness scores. In addition to these variables, the happiness score was looked at in comparison to the suicide rates to see whether or not that would have any influence in predicting a country's happiness. Determining the variables that best predict a country's happiness might help people ascertain where they should live and what they should strive to have in life in order to be happy.

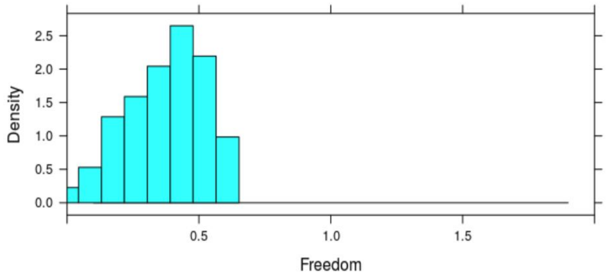
Variables:

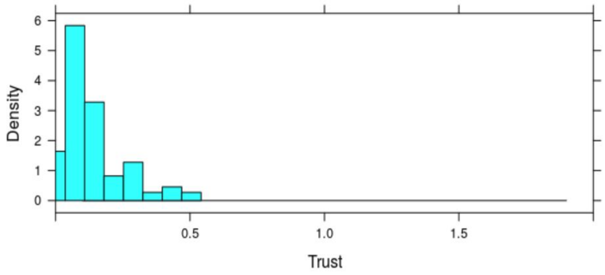
1. Economy - extent to which GDP contributes to happiness score
2. Family - extent to which family contributes to happiness score
3. Health - extent to which health contributes to happiness score
4. Freedom - extent to which freedom contributes to happiness score
5. Trust - extent to which perception of government corruption contributes to happiness score
6. Generosity - extent to which generosity of a country contributes to happiness score
7. Suicide Rate - number of suicides per 100,000 people
8. Happiness Score - based on a question that asked people to rank their happiness

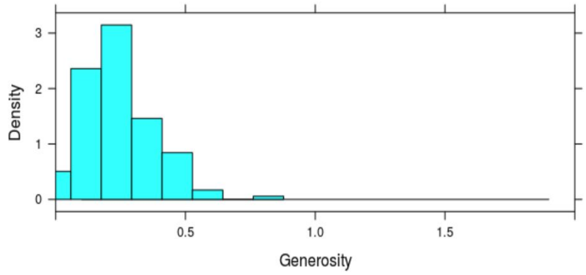




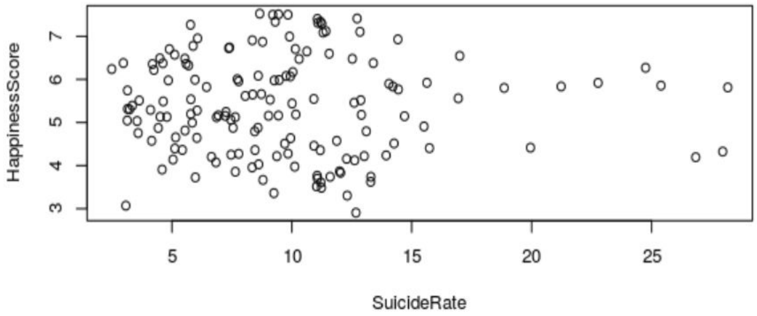








Above are the distributions of the first 6 variables, which shows the extent to which they contribute to happiness score. Based on the distributions, we predict that Economy, Family, and Health will be in the prediction model for Happiness Score because of their high extent of contribution.

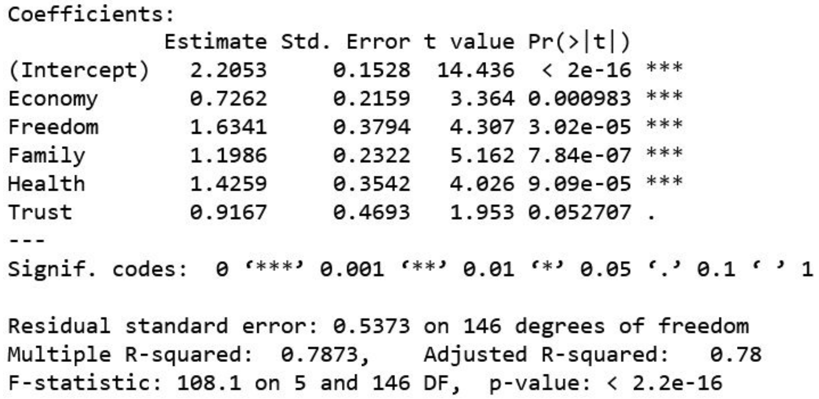


The scatter plot above shows the relationship between Suicide Rate and Hapiness Score. It appears that there is no trend, we predict that Suicide Rate will not be a predictor variable for Happiness Score.

1st-order model:

In order to find a model that can predict Happiness Score from the first 7 variables: Economy, Family, Health, Freedom, Trust, Generosity, and Suicide Rate, we ran a stepwise selection to sieve out the first order terms of the model. The 5 variables that passed the stepwise screening were Economy (x_1), Freedom (x_2), Family (x_3), Health (x_4), and Trust (x_5). Following that, we tested the model's utility. Our null hypothesis is that all β_1 through $\beta_5 = 0$, and the alternative hypothesis is that at least one $\beta_i \neq 0$, with $i \geq 0$. Our F-test resulted in a p-value of approximately 0, which is significant at a 5% significance level. Therefore, we rejected the null hypothesis and concluded that the model is useful.

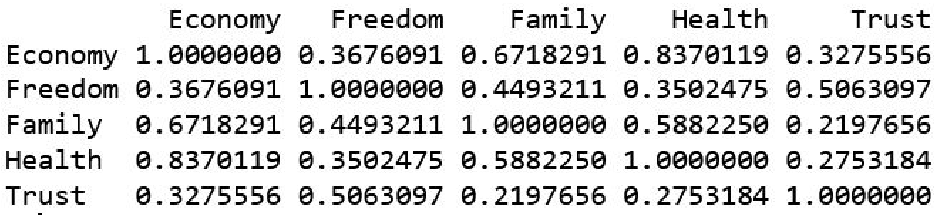
As we can see in Picture 1, 4 out of 5 terms are highly significant, which might be a sign of multicollinearity. Therefore, our next step is testing for correlation between the independent variables. In Table 1, the correlation between Health and Economy is quite high at 0.84. However, it is not high enough to conclude that there is multicollinearity between these variables. We proceeded to search for the best model using these 5 variables.
Model: $E(y) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2053     0.1528  14.436  < 2e-16 ***
Economy       0.7262     0.2159   3.364 0.000983 ***
Freedom       1.6341     0.3794   4.307 3.02e-05 ***
Family        1.1986     0.2322   5.162 7.84e-07 ***
Health        1.4259     0.3542   4.026 9.09e-05 ***
Trust         0.9167     0.4693   1.953 0.052707 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5373 on 146 degrees of freedom
Multiple R-squared:  0.7873,	Adjusted R-squared:   0.78 
F-statistic: 108.1 on 5 and 146 DF,  p-value: < 2.2e-16
```

Picture 1: Result model of stepwise regression



```
          Economy   Freedom    Family    Health     Trust
Economy 1.0000000 0.3676091 0.6718291 0.8370119 0.3275556
Freedom 0.3676091 1.0000000 0.4493211 0.3502475 0.5063097
Family  0.6718291 0.4493211 1.0000000 0.5882250 0.2197656
Health  0.8370119 0.3502475 0.5882250 1.0000000 0.2753184
Trust   0.3275556 0.5063097 0.2197656 0.2753184 1.0000000
```

Table 1: Result of correlation testing between independent variables

Interaction models:

In our next step, we try to find if there are any useful interaction terms between two of the variables we have from our first step. For each of the interaction terms, we conduct a nested F-test to see if adding the interaction term improves our original model. Our null hypotheses are the same for each test, that is the interaction coefficient $\beta_6 = 0$, and our alternative hypotheses are that $\beta_6 \neq 0$. The following table recorded the p-value of 10 nested F-test for 10 interactions at a 5% significance level, the adjusted R-squared of the model in testing and our conclusion.

Model(s): $E(y) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_i\ x_j$

Interaction (β_6)	$R_a{}^2$	p-value	Conclusion
Economy * Freedom	0.7801	0.3079	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Economy * Family	0.789	0.0082	Reject the null hypothesis. Model is an improvement.
Economy * Health	0.783	0.0859	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Economy * Trust	0.7788	0.6462	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Freedom * Family	0.7799	0.3384	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Freedom * Health	0.7821	0.1219	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Freedom * Trust	0.7787	0.7182	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Family * Health	0.7938	0.0013	Reject the null hypothesis. Model is an improvement.
Family * Trust	0.7823	0.1115	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Health * Trust	0.7791	0.5286	Fail to reject the null hypothesis. Cannot conclude that model is an improvement

After finding 2 significant interaction terms: Economy * Family (β_6) and Family * Health (β_7), we wanted to know if a model with both of the interaction terms would be an improvement to the first-order model or one of the single interaction models. We conducted three more nested F-tests with the null hypothesis that the added interaction coefficient β_6 or β_7= 0, and our alternative hypotheses are that at least one β_6 or $\beta_7 \neq 0$. The following table is our p-value and conclusions at a 5% significance level.

Model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_3 + \beta_7 x_3 x_4$

Compared to	p-value	Conclusion
First-order model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$	0.0054	Reject the null hypothesis. Model is an improvement.
Only interacting Economy * Family $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_3$	0.0630	Fail to reject the null hypothesis. Cannot conclude that model is an improvement
Only interacting Family * Health $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_3 x_4$	0.7346	Fail to reject the null hypothesis. Cannot conclude that model is an improvement

We can see that, although the newest model is an improvement from the first-order model, it is not an improvement compared to the models with only one interaction term. Therefore, we discard this new model. Between the two models with only one interaction terms, one between Economy and Family and one between Family and Health, we compared their R_a^2 value to choose the better model. The model interacting Family and Health has a higher R_a^2 of 0.7938, meaning 79.38% of the happiness score can be predicted using this model. Therefore, we choose this model to continue with our project.

2nd Order Models:

In order to determine whether or not any of the 2nd order terms of the predictor variables would improve the model, we looked at each of the 2nd order terms individually and ran a nested F-test to ascertain whether or not the 2nd order term was significant. Our null hypothesis was that $\beta_6 = 0$ and our alternative hypothesis was that $\beta_6 \neq 0$. If our nested F-test resulted in a p-value significant at the 5% significance level, we reject the null hypothesis and conclude that the model is useful. If not, we fail to reject the null hypothesis and conclude that the model is not an improvement.

Model(s): $E(y) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_i^2$

2nd Order Term $(\beta_6)^2$	p-value	Conclusion
poly(Family, 2)	0.1294	Fail to reject the null hypothesis. Cannot conclude that model is an improvement.

poly(Trust, 2)	0.0465	Reject the null hypothesis. Model is an improvement.
poly(Health, 2)	0.1717	Fail to reject the null hypothesis. Cannot conclude that model is an improvement.
poly(Economy, 2)	0.3461	Fail to reject the null hypothesis. Cannot conclude that model is an improvement.
poly(Freedom, 2)	0.8777	Fail to reject the null hypothesis. Cannot conclude that model is an improvement.

Nested F-tests with Best Interaction & 2nd Order Model :

After determining the significance of the family and health interaction and the significance of the second order Trust term, we combined both of them into a model with a 2nd-order Trust term (β_7) and an interaction term of Family and Health (β_6). Then we performed nested F-tests and compared this new model to the first order model, the interaction only model, and the 2nd-order only model. Our null hypothesis was that β_6 and $\beta_7 = 0$ while the alternative hypothesis was that either β_6 or $\beta_7 \neq 0$. If our nested F-test results in a p-value significant at the 5% significance level, we reject the null hypothesis and conclude that the model is useful. If not, we fail to reject the null hypothesis and conclude that the model is not an improvement.

Model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_3 x_4 + \beta_7 x_5^2$

Compared to	p-value	Conclusion
First-order model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$	0.00067	Reject the null hypothesis. Model is an improvement.
Interaction only model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_3 x_4$	0.0383	Reject the null hypothesis. Model is an improvement.
2nd-order only model: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_5^2$	0.0011	Reject the null hypothesis. Model is an improvement.

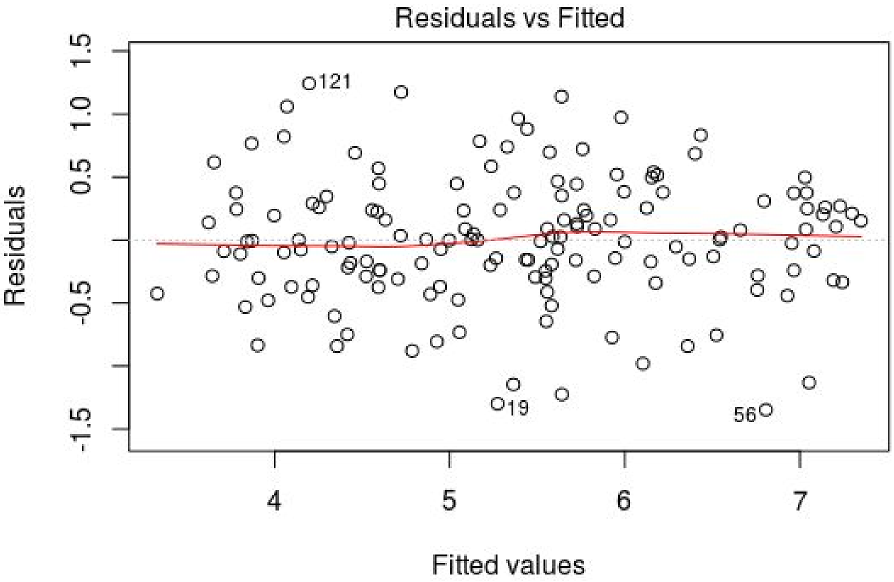
While we were able to conclude that all of the models were significant, we determined that the model with the interaction term between Family and Health and with the second order term for Trust proved to be an improvement over the first-order model, interaction only model, and the 2nd-order only model.

<u>Fitted model:</u> $\hat{y} = 3.14 + 0.76x_1 + 1.42x_2 - 0.063x_3 - 0.54x_4 + 0.588x_5 + 2.62x_3x_4 - 1.09x_5^2$

Our fitted model has an adjusted R-squared of 0.7985, indicating that 79.85% of the happiness score of a country can be predicted with this model. The model is highly significant with a p-value of approximately 0.

Residual plot for final model:

Looking at the residuals versus the fitted values of our best model, we do not see any visual trends or patterns to the plot. Additionally, the points are scattered around 0, suggesting that the x values are correctly specified in the given model.



lm(HappinessScore ~ Economy + Freedom + Family * Health + poly(Trust,2))

Limitations/Assumptions:

This data is representative of samples that were collected in 2016, so the validity of this information could have well changed since then. All of these variables are subject to change from year to year, and the data may not be as applicable to the current state of the world now as it was in 2016. Specifically, given an extreme situation such as the ongoing pandemic, people's levels of happiness may not match what we have analyzed. Additionally, the answer to whether or not people are happy is subjective and some people in certain cultures may be more willing to share that they are unhappy over others.

Conclusion:

Based on our findings and our best predicted model, if people are interested in what will most increase their happiness, they should specifically focus on family and health, live in a place where the perception of government corruption is low, the economy is strong and they have a sense of freedom.

Citations:

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

https://www.kaggle.com/unsdsn/world-happiness

https://www.kaggle.com/srinesh/passport-power-2018?select=Passport_index.csv