CS 750 Final Project: Predicting Happiness by Country

Devin Bouchard, Travis Calley, Ryan Reynolds

- **Motivation/Introduction**: Which addresses these issues:

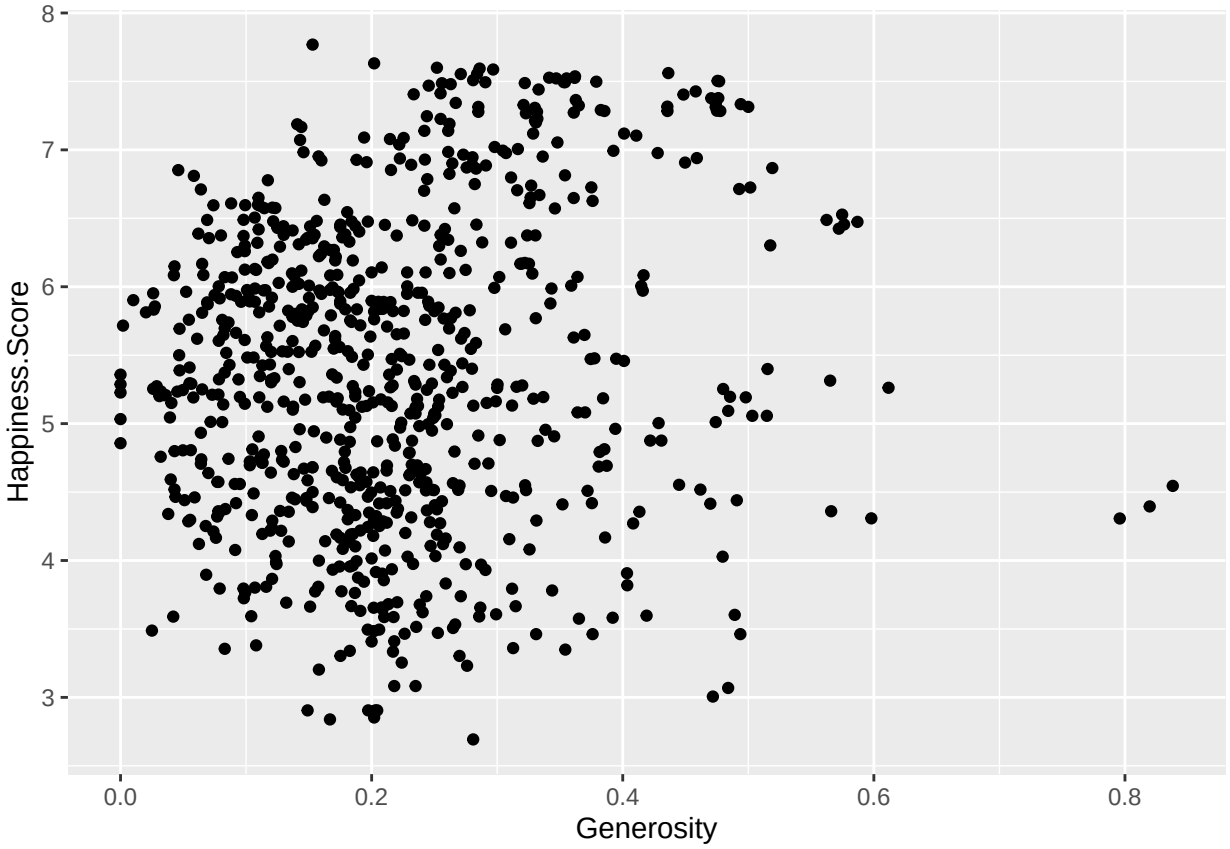
  1. What is the problem? The problem in this project is to determine and generate models to predict happiness of a country based on the following predictors: economy, family, health (life expectancy), trust (government corruption), freedom, and generosity.
  2. Is it prediction or inference? This is an inference problem because we are trying to learn about the data generation process. We are trying to generate a model that can accurately predict a given countries happiness score/rank.
  3. Is it classification or regression? This is a regression problem because we are aiming to predict a hapiness rank which is a continuous value. There happiness score is a value that is scaled accordingly and does not categorize the data set. We are using a few factors to predict happiness score/rank.
  4. Why is the problem important? This problem is important because happiness is important. Hapiness is vital in a society to achieve goals and innovation to advance humanity. Without hapiness motivation is low and nobody would be willing to do anything for the greater good. In this project we will understand the things that make a population happy in the hope of understanding what would have to changed to increase the happiness of a large group.
  5. What does success look like? A successful project would have a training data set that has all of the features included but the predictor for happiness rank. This predictor would then be predicted based on the hapiness rank of other countries with the same features. We should also be able to look at the coorelation of these features the predictor to determine with of the feawtures has the largest impact in making people happy.
  6. What are the data sources that will be used. Is it likely that they will suffice to achieve the goals? We will be using data from the years 2015-2019 for happiness rank by country. The data was downloaded from https://www.kaggle.com/unsdsn/world-happiness.

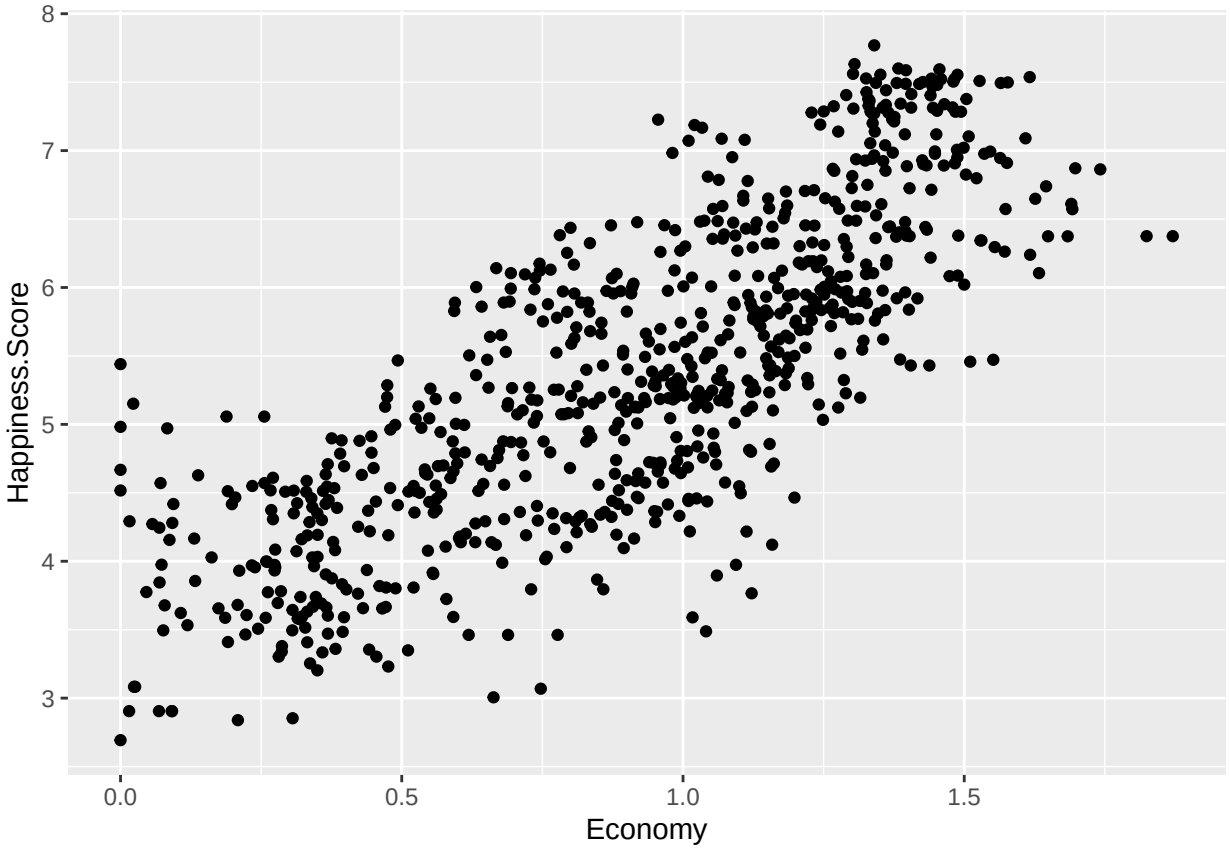
- **Evaluation methodology**: You should answer questions like:

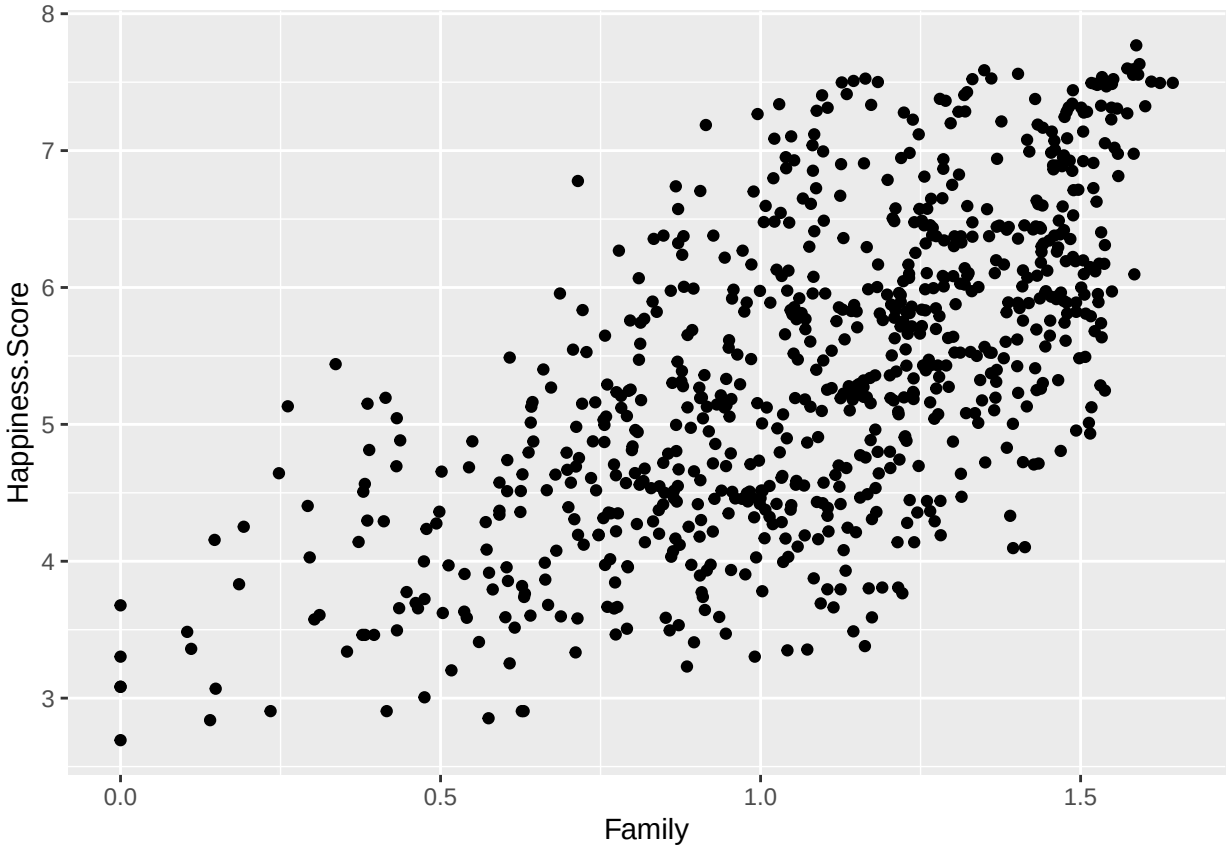
  1. What is the right metric for success? A good metric would be the Mean square error of our predicted hapininess rankings for a country compared to the actual hapiness rankings given in the original data set. This number should be as small.
  2. How good does it need to be for the project to succeed? For example, does the prediction error needs to be at most 5%? What about the area under the curve. Argue why. Determining which countries are the happiest based on several factors is important for individuals as well as countries as a whole. If an individual is looking to live somewhre else, this data would be useful to see what factors matter most to people in the area as well as the overall happiness of that country. For countries, it would be useful to look at this data and see where their country is lacking and try to improve on those areas. Countries would be able to use the models we create to see where they would rank based on changes in each of the feature categories. Therefore, the data should have at most a 10% prediction error. It is important to ensure a low prediction error on data that could have a strong influence on decisions people make which could have a strong influence on their lives or the lives of others.

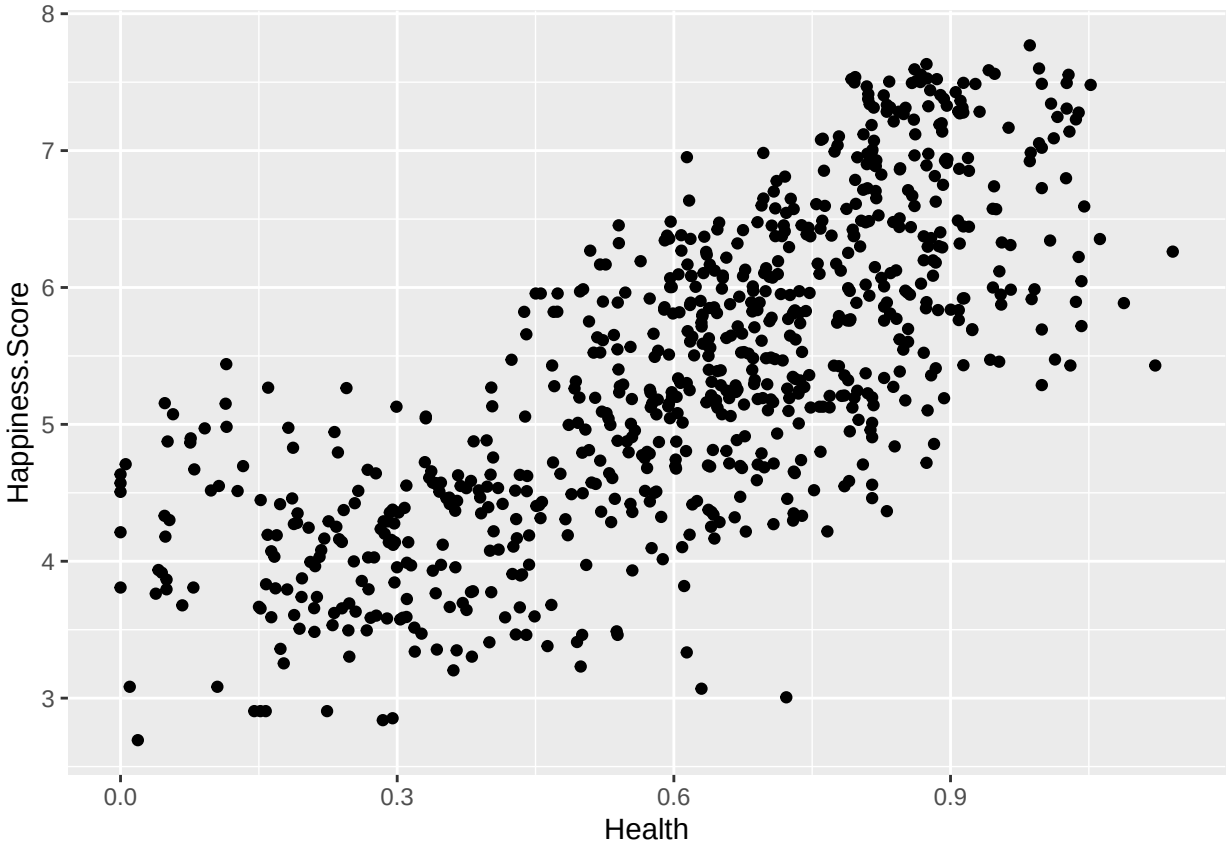
3. Use a test set? Bootstrapping to understand parameter variability? We used a test set to validate the methods we chose to analyze the data. Since there were over 700 data points in the data set, we felt that splitting the data into training and test sets was enough to verify the performance. We split the data into 70% training and 30% test.
4. How to make sure that the results are valid? Comparing the models to a subset of the data and calculating the MSE is a useful way to determine the validity of the moel. It is also wise to use bootstrapping here to be sure one can have a strong confidence in the results that are obtained.

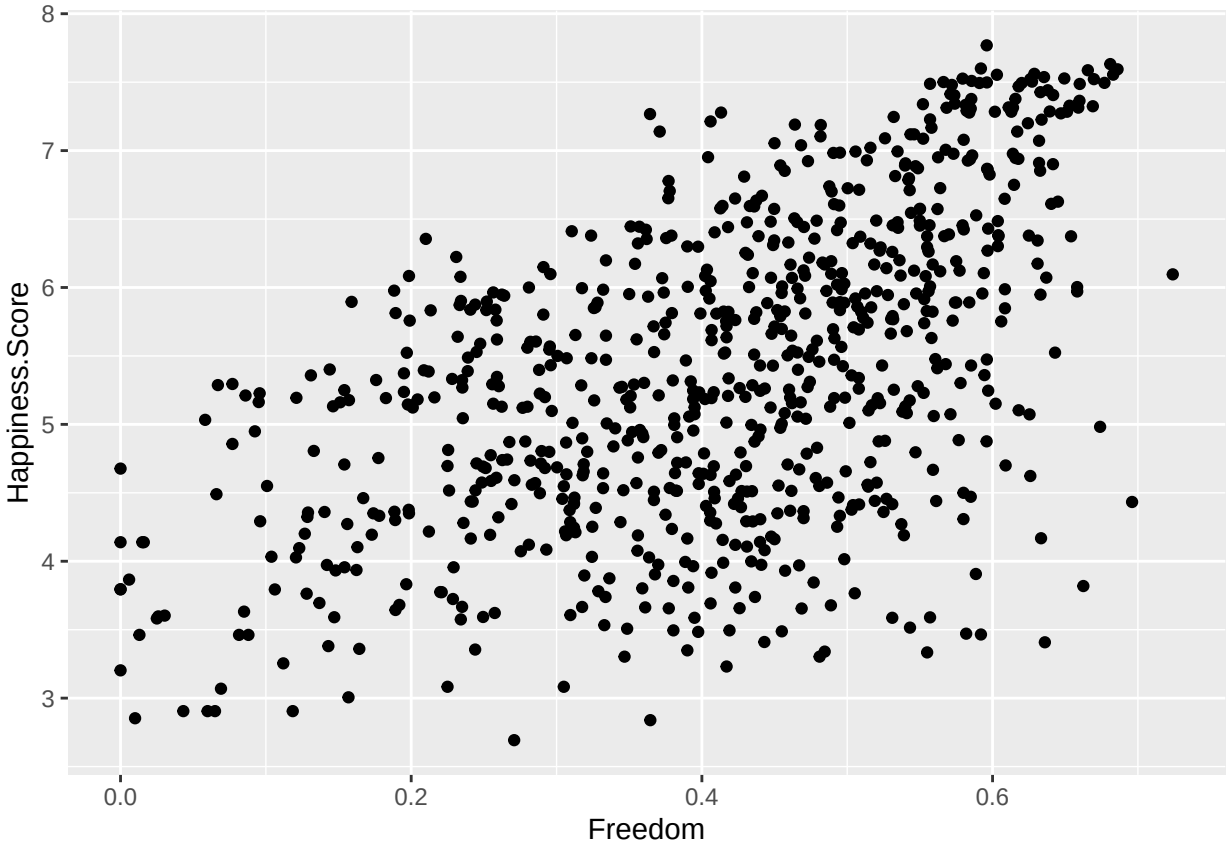
- **Implementation**

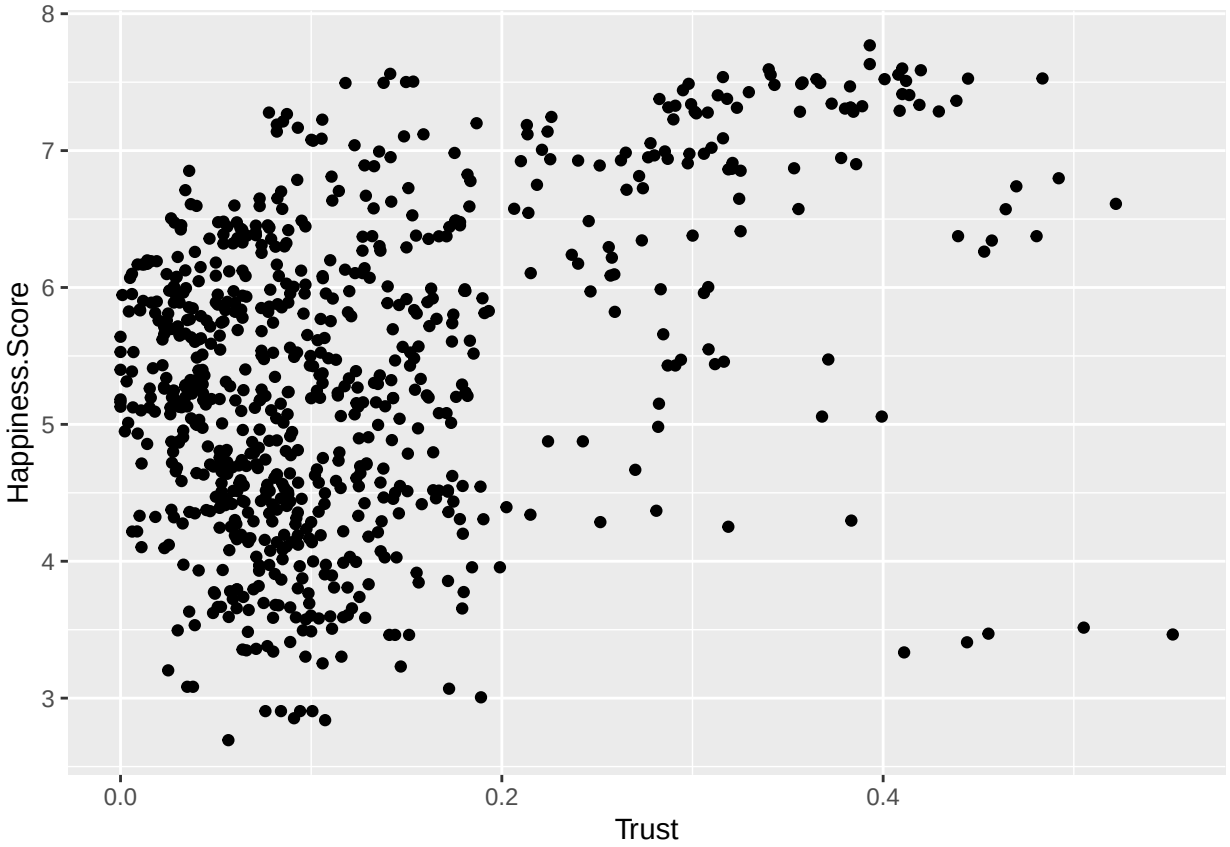



Happiness.Score
8
7
6
5
4
3
0.0
0.5
1.0
1.5
Economy


Happiness.Score
8
7
6
5
4
3
0.0
0.5
1.0
1.5
Family


Happiness.Score
8
7
6
5
4
3
0.0
0.3
0.6
0.9
Health


Happiness.Score
8
7
6
5
4
3
0.0
0.2
0.4
0.6
Freedom

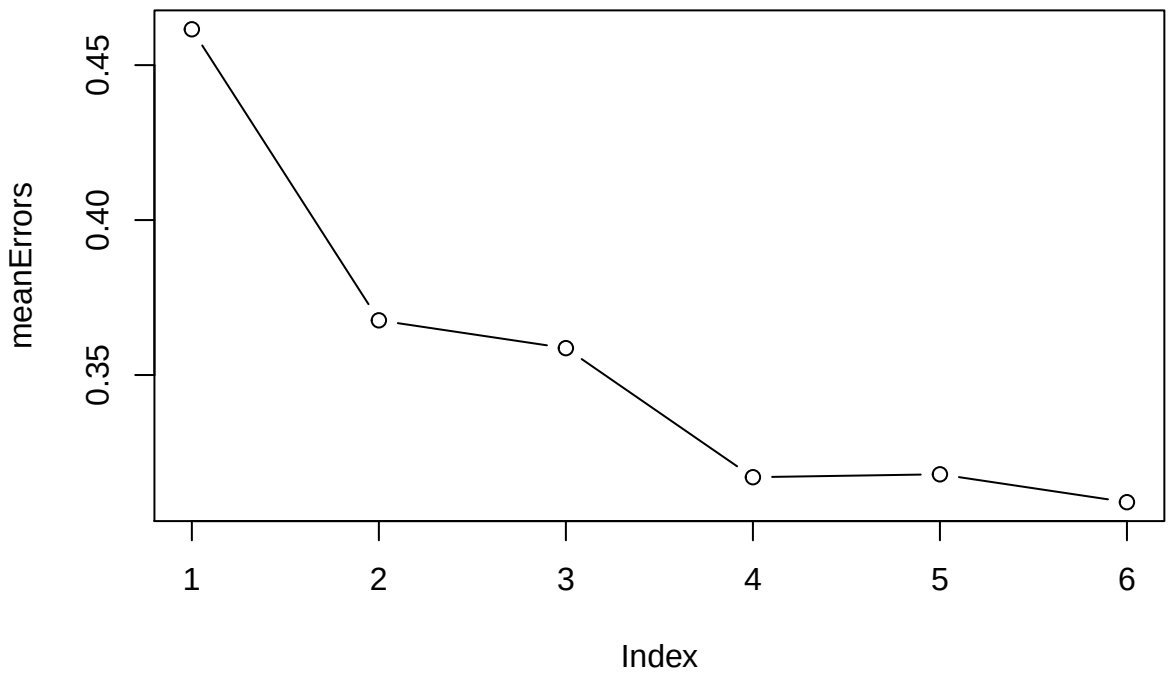


Using the plots of the above features we can determine the features that have the largest affect on the hapiness scores of the countries in our dataset. If the data on the plot is in a linear shape with a high slope, then the feature has high coorelation to hapiness.

Based on the plots, the features with the highest coorelation are Economy, Family, and Health. This tells us that for a population to be happiest, it is most important fot them to have high values in these areas of life.

```
## Subset selection object
## Call: regsubsets.formula(Happiness.Score ~ Economy + Family + Health +
##     Freedom + Trust + Generosity, data = train[folds != k, ],
##     nvmax = 6)
## 6 Variables  (and intercept)
##            Forced in Forced out
## Economy        FALSE      FALSE
## Family         FALSE      FALSE
## Health         FALSE      FALSE
## Freedom        FALSE      FALSE
## Trust          FALSE      FALSE
## Generosity     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          Economy Family Health Freedom Trust Generosity
## 1  ( 1 ) "*"     " "    " "    " "     " "   " "
## 2  ( 1 ) "*"     " "    " "    "*"     " "   " "
## 3  ( 1 ) "*"     " "    "*"    "*"     " "   " "
## 4  ( 1 ) "*"     "*"    "*"    "*"     " "   " "
## 5  ( 1 ) "*"     "*"    "*"    "*"     "*"   " "
```

```
## 6  ( 1 ) "*"     "*"    "*"    "*"     "*"   "*"
```



If using few features was desired, this best subset selection algorithm would provide the best set of features to use based on the total number of desired features. The printout above gives the best subset of features to use for every number of features from 1 to 6 in this case. For example, if it was desired to only use one feature, the printout above says using the "Economy" feature will yield the most accurate result.

```
## Linear

## Test MSE: 0.3011421

##
##
## Lasso

##
## Test MSE: 0.3014939

##
##
## SVM radial

##
## Test MSE: 0.2597847
```

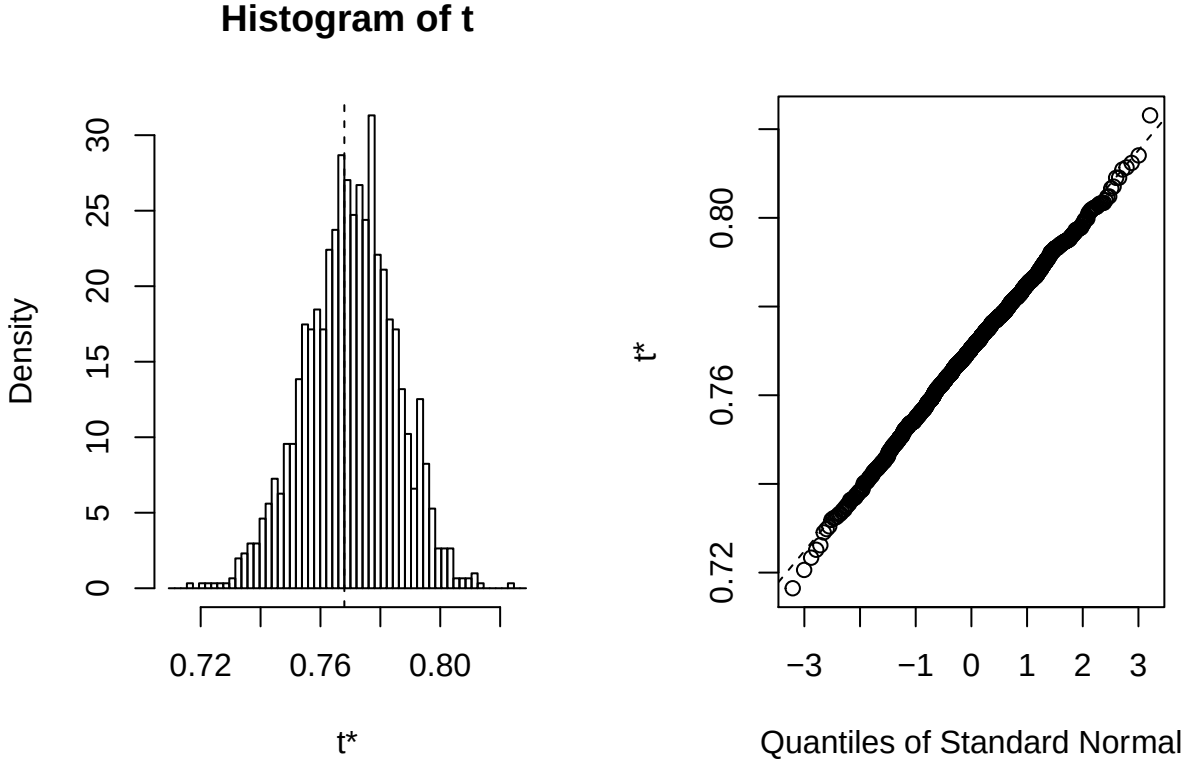
```
##
##
## SVM linear

##
## Test MSE: 0.2968839

##
##
## Boosted Tree

##
## Test MSE: 0.2839667

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d.all, statistic = bs, R = 1500, formula = Happiness.Score ~
##     Economy + Family + Family:Health + Health + Freedom + Trust +
##         Generosity)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.7679815 0.002069636  0.01506598
```



```
## Test RMSE:

## 32.04757

##            Country Happiness.Rank Happiness.Score Happiness.Rank.Pred
## 1     2019-Finland              1           7.769                   1
## 2 2018-Switzerland             10           7.487                   2
## 3      2018-Sweden             14           7.314                   3
## 4      2018-Norway              2           7.594                   4
## 5     2017-Finland             12           7.469                   5
## 6      2015-Norway              4           7.522                   6
##   Happiness.Score.Pred  Economy   Family    Health   Freedom     Trust
## 1             7.634298 1.340000 1.587000 0.9860000 0.5960000 0.3930000
## 2             7.599391 1.420000 1.549000 0.9270000 0.6600000 0.3570000
## 3             7.581896 1.355000 1.501000 0.9130000 0.6590000 0.3830000
## 4             7.578847 1.456000 1.582000 0.8610000 0.6860000 0.3400000
## 5             7.479691 1.443572 1.540247 0.8091577 0.6179509 0.3826115
## 6             7.472911 1.459000 1.330950 0.8852100 0.6697300 0.3650300
##   Generosity
## 1  0.1530000
## 2  0.2560000
## 3  0.2850000
## 4  0.2860000
## 5  0.2454828
## 6  0.3469900
```

- **Results**: Describe the results of the method. Describe how well the method did in the evaluation and compare with prior work (if applicable). Discuss what the results mean in the context of the problem definition. Is there anything that can be done to improve the results, or are they good enough? What about confidence in the results?

Knowing that this was a regression and inference problem, we chose methods that would best fit this classification. We started by attempting a best subset selection. The data we had did not have many features but we wanted to see if any of the features had a negative effect on model prediction. It turned out that using a subset of 4 and 5 features was very close to using all six features. Knowing this, we decided to fit models with all six features since the Test MSEs for six features was always better than four or five. The first model we tried was a linear regression model which attempted to predict the happiness score based on the six relevant features: Economy, Family, Health, Freedom, Trust, and Generosity. The MSE obtained from this method was about 0.301. We tried using bootstrapping with multiple regression methods; however, we were only able to get it working using linear regression. The linear regression results from boosting showed a quantile plot that strongly follows a normal distribution with high confidence. The histogram shows that most of the data falls within 2 standard deviations of the mean, these are both good results that give us strong confidence for linear regression.

To create a model that more correctly predicted happiness rank and score we added an interaction effect between the features Family and Health (Family:Health). Adding this interaction effect lowered MSE across the board for all models that were created. Also, when comparing the ranks in the above table it made the results more closely reflect what we were expecting. Finland was accurately predicted as the most happy country with the other results not being too far off from what they were supposed to be.

The next method we tried was Lasso. Even though we had more data points than features, we found that Lasso provided a good model on the miniproject datasets which were similar to this happiness dataset. The Lasso method also used all six features and produced almost the same MSE as the linear model at just over 0.301. In addition to these methods, we also tried SVMs and Boosted Trees. We found that a radial SVM

performed the best out of any method. THe MSE for a radial SVM using six features was 0.259. We also compared the ranks themselves instead of just the scores that were predicted. The ranks appear closer than the predicted score makes them seem. While they are not perfectly ordered, each rank is within about 5-10 places of where it actually should have been.