WORLD HAPPINESS ANALYSIS

CSE 351: Introduction to Data Science

Edward Ng
Sean Xia
Tracy Ho

TABLE OF CONTENTS

01

DATA CLEANING

Name Unification, Replacing Outliers, and Filling in Missing Values.

02

DATA ANALYSIS

Some noticeable trends and correlations in the data.

03

MODELING

Value Predictions for the Happiness Scores.

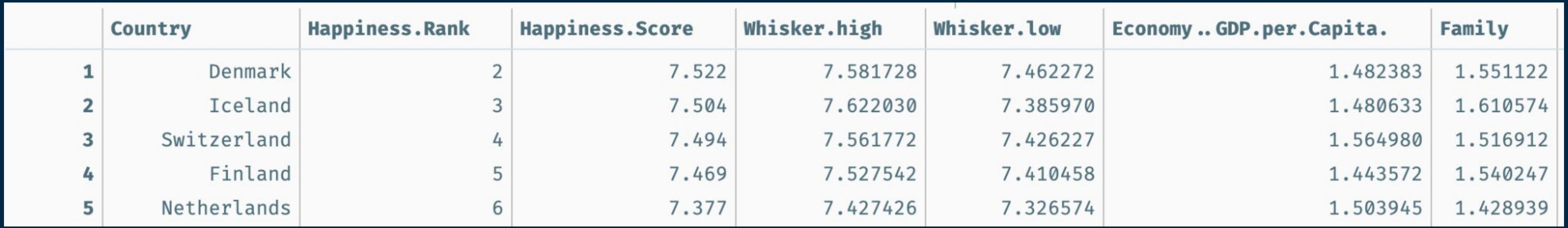
DATA CLEANING

01

Concatenating the Data

The data provided was separated by years. Thus, we wanted to combine all of the data together so that it was easier to manage and work with. However, doing so created multiple issues.

Column Name Unification



	Country	Happiness.Rank	Happiness.Score	Whisker.high	Whisker.low	Economy..GDP.per.Capita.	Family
1	Denmark	2	7.522	7.581728	7.462272	1.482383	1.551122
2	Iceland	3	7.504	7.622030	7.385970	1.480633	1.610574
3	Switzerland	4	7.494	7.561772	7.426227	1.564980	1.516912
4	Finland	5	7.469	7.527542	7.410458	1.443572	1.540247
5	Netherlands	6	7.377	7.427426	7.326574	1.503945	1.428939

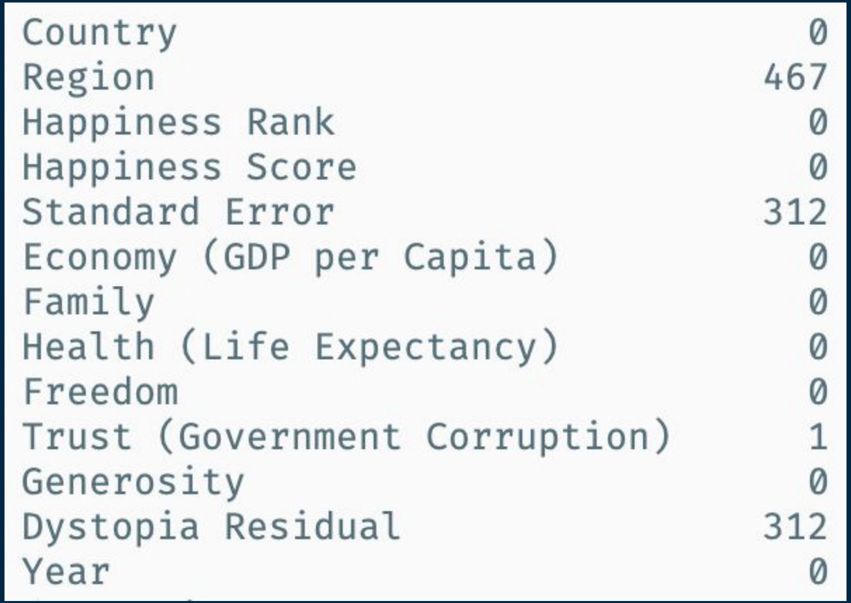
The first apparent issue is the naming of each column, leading to columns to be separated in concatenation even though they represent the same data.

Confidence Intervals v. Standard Error

For some years, the standard error was reported for the happiness score. However, for other years, it was reported as upper and lower confidence levels.

To concatenate these together, we simply dropped both upper/lower confidence levels and calculated the standard error.

Missing Values



```
Country                              0
Region                             467
Happiness Rank                       0
Happiness Score                      0
Standard Error                     312
Economy (GDP per Capita)             0
Family                               0
Health (Life Expectancy)             0
Freedom                              0
Trust (Government Corruption)        1
Generosity                           0
Dystopia Residual                  312
Year                                 0
```

A count of the missing values in our data based on column parameter.

Missing Regions

Regional data was included in the 2015/2016 dataset. However, it was dropped in the others.

To alleviate the simple issue, we created a dictionary mapping of countries to regions from the 2015/2016 dataset.

We used this dictionary to generate the regions that were missing from the other datasets.

Furthermore, it helped detect inconsistent country names.

Missing Standard Errors/Trust Ranks/Corruption Residuals

The missing standard error was dropped in 2019, but we still wanted to keep this information so we filled the missing values with 0.

The 2019 set also dropped corruption residuals. However, we did not see any valuable insight so we dropped the variable.

There was one missing trust rank, indicating some data error. To fix this, we imputed with the mean for that country.

Outlying Values

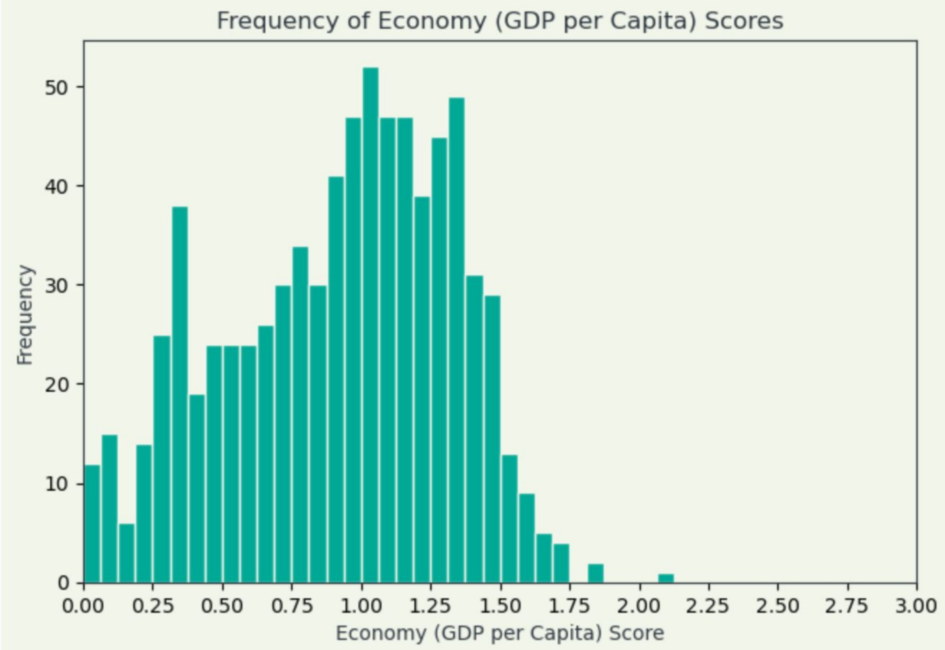
Outlying Values

Outlier Detection

To detect outliers, we simply visualized and binned the frequency of each variable. Because they resembled normal curves, we labeled points that fell 3 standard deviations away from the mean as outliers.

Outlier Imputation

For each outlying value that was identified, we simply replaced the value with the mean of the values in the entire dataset.



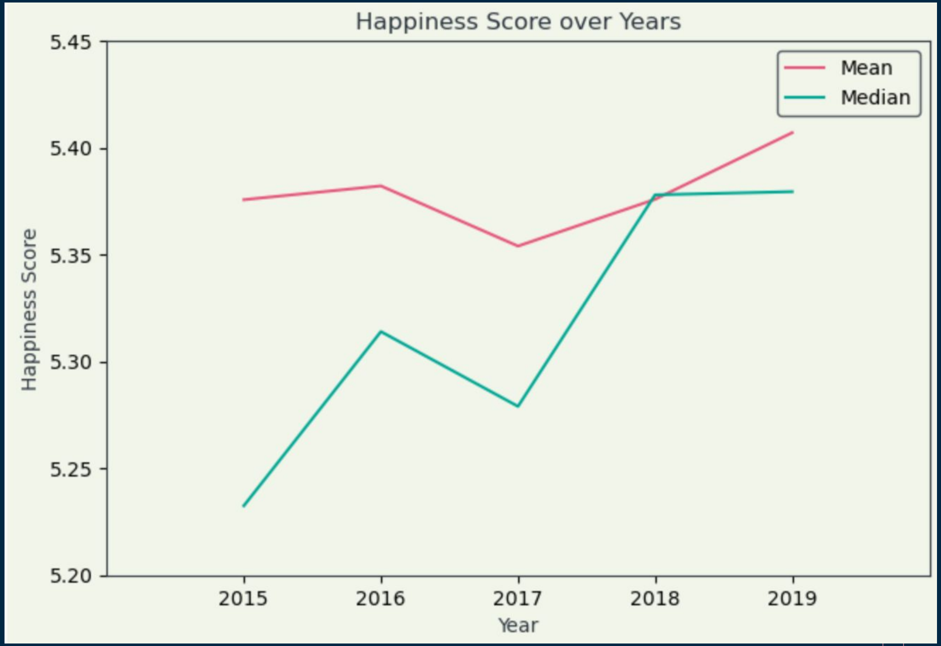
Frequency Distribution of Economy Scores

DATA ANALYSIS

02

Happiness Score Central Tendencies

We calculated the mean/median happiness scores for each year and then plotted the results to analyze the trends.



Stable Country Happiness Ranks

A stable happiness rank means that the country has gone through the least changes in ranking.

Top 5 Stable Countries

- Iceland (4)
- Switzerland (6)
- Denmark (8)
- Canada (9)
- Norway (9)

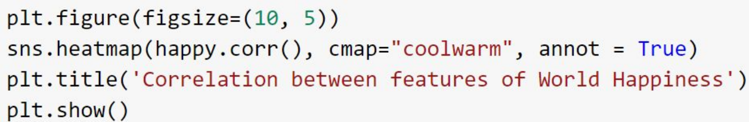
Increasing Country Happiness Ranks

For a country’s happiness rank to be improving, it must increase throughout all five years.

Top 5 Increasing Countries

- Benin (+208)
- Ivory Coast (+203)
- Burkina Faso (+189)
- Cambodia (+181)
- Togo (+177)

VISUALIZE HAPPINESS SCORE’S RELATIONSHIP

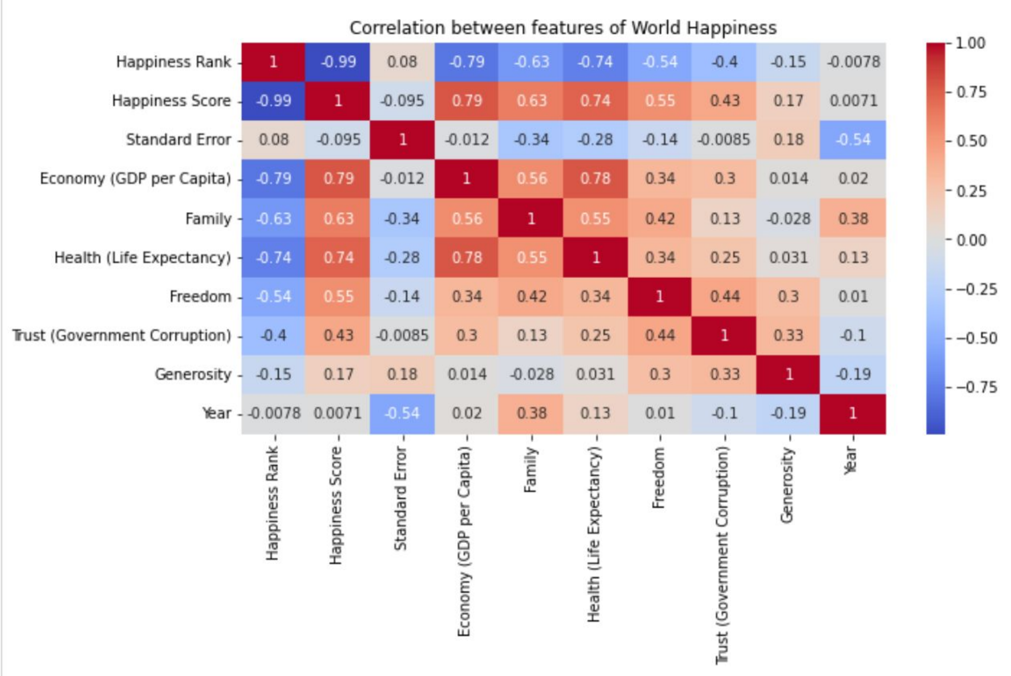


```
plt.figure(figsize=(10, 5))
sns.heatmap(happy.corr(), cmap="coolwarm", annot = True)
plt.title('Correlation between features of World Happiness')
plt.show()
```

In order to visualize the relationship between happiness score and other features, I decided to use a Heatmap.

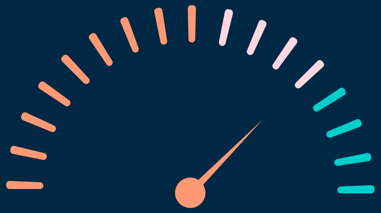
The heatmap generates a correlation matrix that shows that correlation value between the variables.

From this heatmap, we can see that the Happiness score has a high correlation with Economy, Health, and Family.



HOW TO INCREASE HAPPINESS SCORE?

From the heatmap, we can see that the Happiness score has a high correlation with Economy, Health, and Family.



IMPROVE QUALITY OF EDUCATION

INCREASE JOB SKILLS

Therefore if we were the president of a country, we would first increase the GDP per capita of the country as there is a strong correlation between the GDP and Happiness score. To increase the GDP, we would improve the quality of education and increase job skills within the country. This will also increase the life expectancy of the country as it is highly correlated with GDP.

03

MODELING

LINEAR REGRESSION

LINEAR REGRESSION

WHAT FEATURES DID WE TRAIN ON?

- Economy (GDP per Capita)
- Freedom

Happiness Score Mean Squared Error: 0.6058271239843792

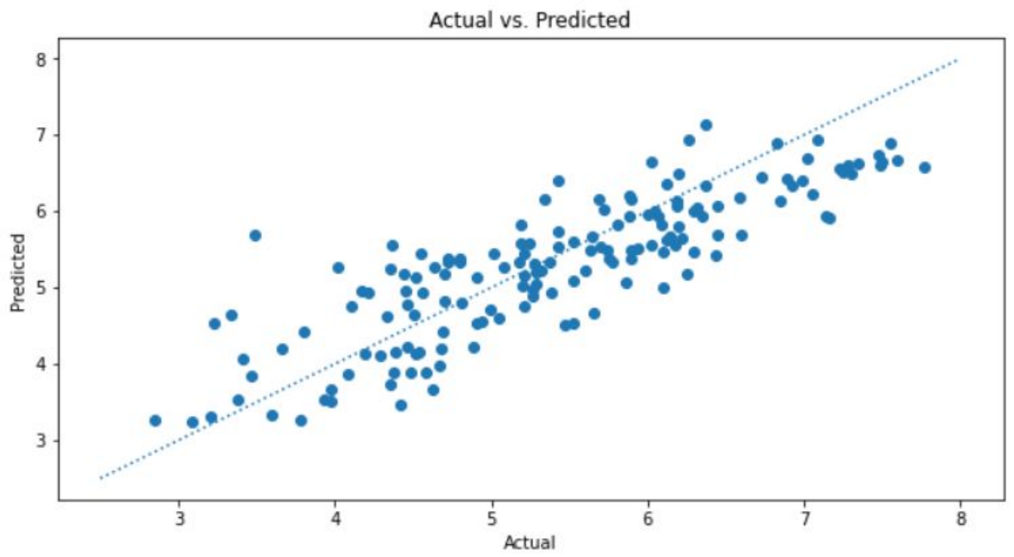
The mean squared error is the standard deviation of the difference between the predicted and actual values

WHY?

We trained our linear regression model with only Economy and Freedom variables as they were highly correlated with Happiness Score. Even though Health and Family was also highly correlated with the Happiness Score, it is also highly correlated with Economy. Putting highly correlated features provide no additional information for modeling, therefore it was not included

ACTUAL VS. PREDICTED VALUES

When plotting the actual vs predicted values, you can see that the data follows a linear trend and the points are relatively close to the actual points.



RIDGE REGRESSION

RIDGE REGRESSION

WHY RIDGE REGRESSION?

Advantages of ridge regression:

- Ridge regression is well-suited for models being trained on data with heavy multicollinearity, or a large correlation between different features, as one would expect between features like the economy, family, health, and generosity.
- Another feature one would expect to be correlated in the data we have is freedom and trust in government.
- Avoids overfitting a model

Disadvantages of ridge regression:

- As it is unable to perform feature selection, all predictors are in the final model meaning if there are useless features, it will not be removed.

RIDGE REGRESSION

WHAT FEATURES DID WE TRAIN ON?

- Economy (GDP per Capita)
- Family
- Health (Life Expectancy)
- Freedom
- Trust (Government Corruption)
- Generosity

HOW DID WE TRAIN?

We split the model into training (2015, 2016, 2017, 2018) and testing (2019) sets, then used ridge regression in order to build a model to predict the Happiness Scores.

Mean Squared Error: 0.5758837127385211

The mean squared error is the standard deviation of the difference between the predicted and actual values

WHY?

The features that we excluded were ['Country', 'Region', 'Happiness Rank', 'Year', 'Standard Error'] because happiness score is not dependent on these values. We are allowed to train the model with all these other intercorrelated features by the properties of ridge regression.

ROBUST REGRESSION

ROBUST REGRESSION

WHY ROBUST REGRESSION (RANdom SAmple Consensus (RANSAC))?

Advantages of robust regression:

- Robust regression is less susceptible to outliers compared to regression methods that are based on least square estimation. Least square estimation are very susceptible to outliers because the variances are evaluated quadratically.
- Probably the best known Robust Regression algorithm is the Random Sample Consensus (RANSAC) algorithm

Disadvantages of robust regression:

- A disadvantage of RANSAC is that there is no upper bound on the time it takes to compute these parameters.
- Sometimes too many iterations are required and we can likely do better than brute force sampling.

ROBUST REGRESSION

WHAT FEATURES DID WE TRAIN ON?

- Economy (GDP per Capita)
- Family
- Health (Life Expectancy)
- Freedom
- Trust (Government Corruption)
- Generosity

HOW DID WE TRAIN?

We split the model into training (2015, 2016, 2017, 2018) and testing (2019) sets, then used the RANdom SAmple Consensus (RANSAC) algorithm in order to build a model to predict the Happiness Scores.

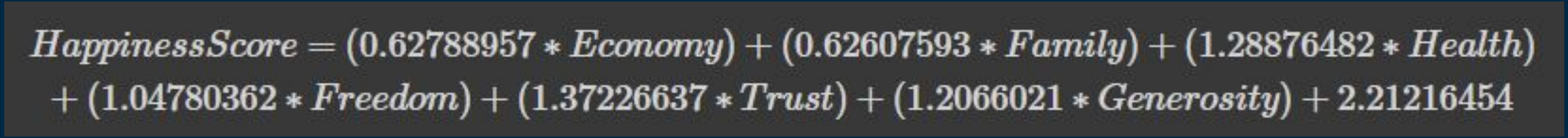
Mean Squared Error: 0.5811039449579765

The mean squared error is the standard deviation of the difference between the predicted and actual values

WHY?

We trained on the same features as we did on ridge regression because all of these features would logically contribute to a person’s happiness. If any country had a feature which is largely different from other countries, but had similar values for all the other features, that one still feature would still have a big effect on happiness alone.

HAPPINESS FORMULA



$$HappinessScore = (0.62788957 * Economy) + (0.62607593 * Family) + (1.28876482 * Health) + (1.04780362 * Freedom) + (1.37226637 * Trust) + (1.2066021 * Generosity) + 2.21216454$$

Out of our models, Ridge Regression performed the best in predicting the values of the Happiness Score as it’s mean squared error was the lowest. Due to that, we invented our formula based on that model.

THANKS

From,

Sean Xia, Tracy Ho, Edward Ng

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik
Please keep this slide for attribution