Group 7
Dystopia

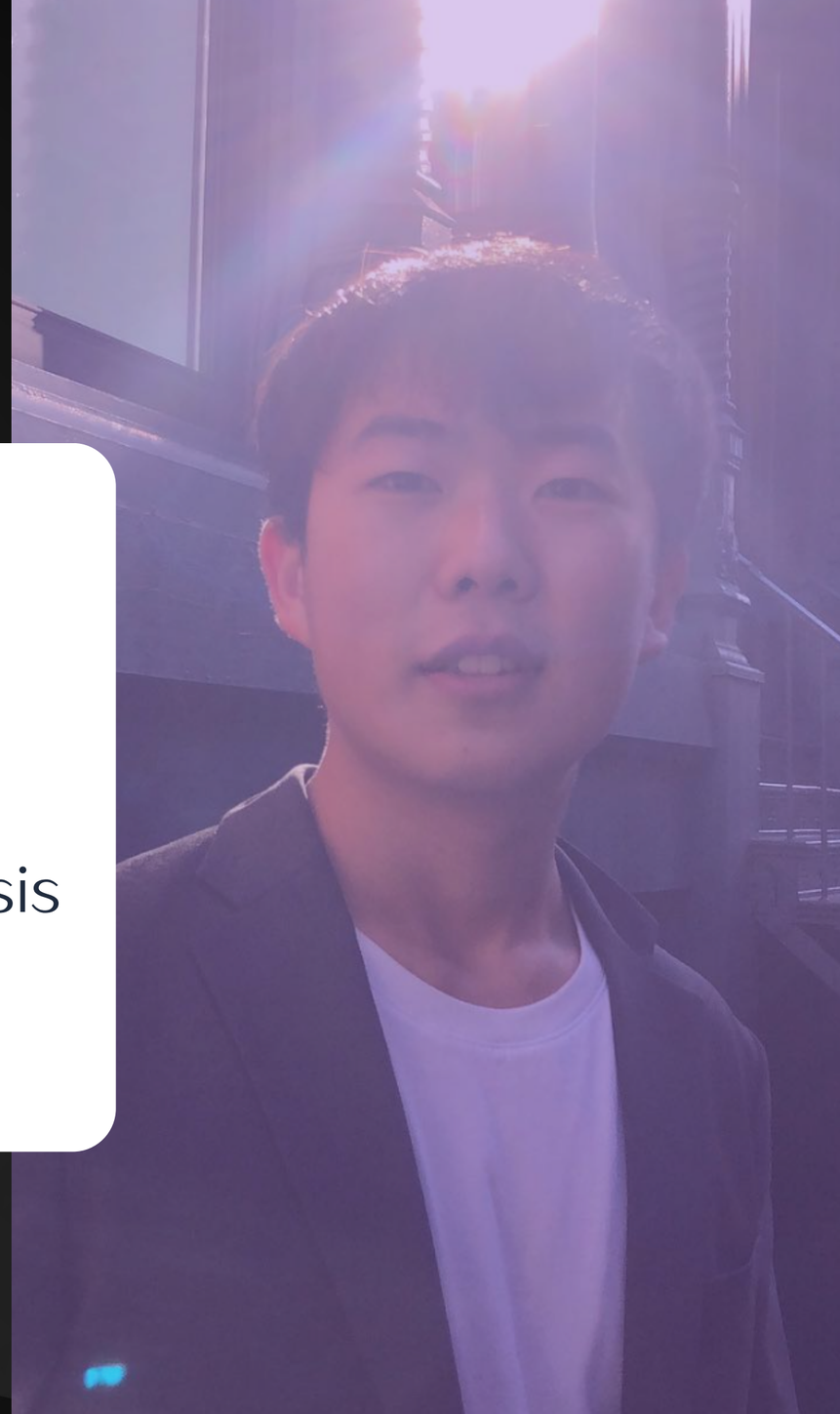
Wharton Data Science Academy 2023

Assessing Happiness

Ritwij Ghosh – Jeesung Lee – Felix Liu – Dmitri Zhao – Steven Kwon

Overview

1. The Problem
2. The Goal
3. Happiness Data analysis
4. Wharton GYP Data
5. World country Data Analysis & Modeling
6. Conclusion & Limitation

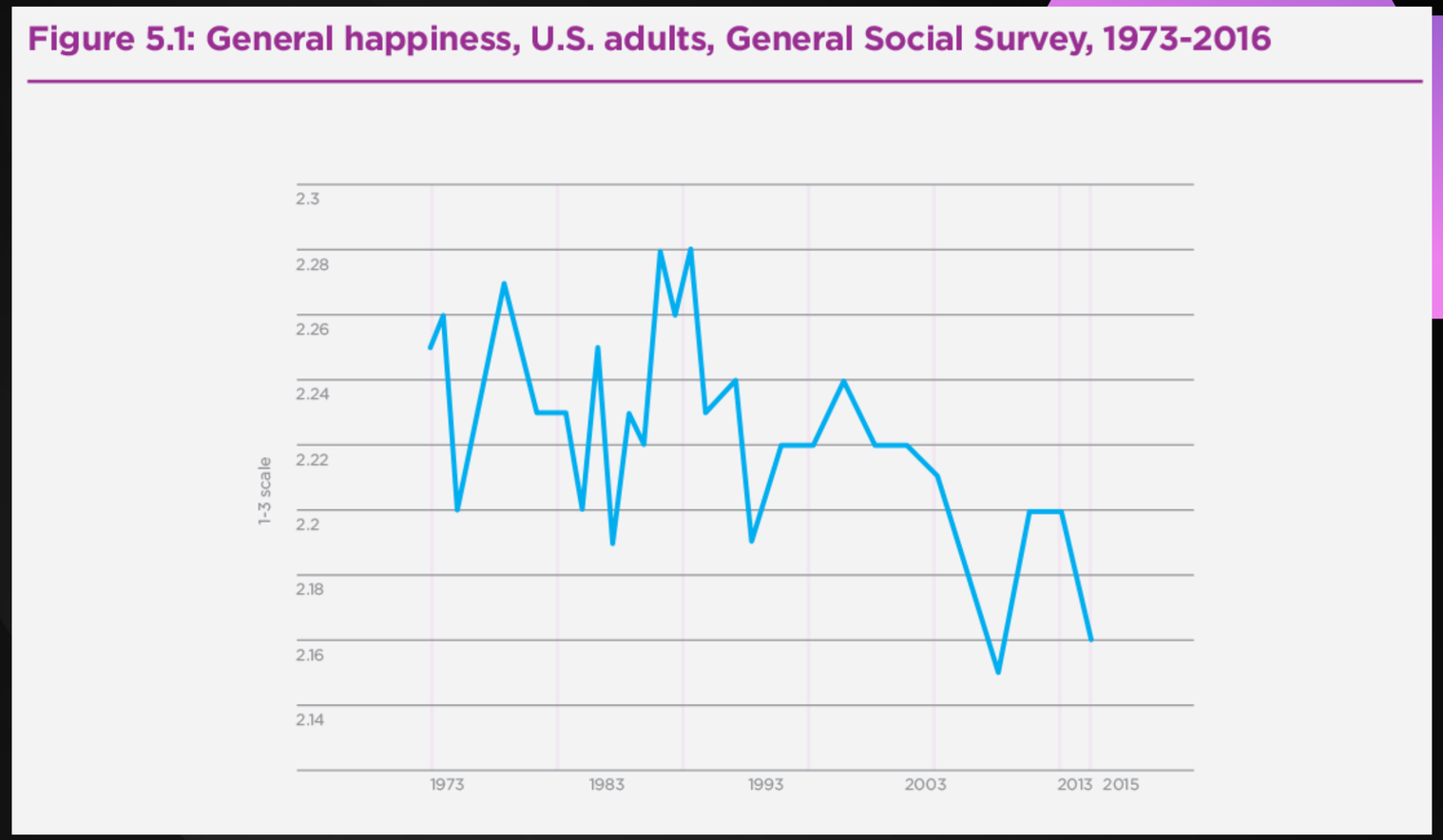


The Problem

01 Happiness rates have been decreasing in the U.S and around the globe.

02 Even after the Great Recession in 2009, the happiness rate never reached its former peak rate in the 1990s.

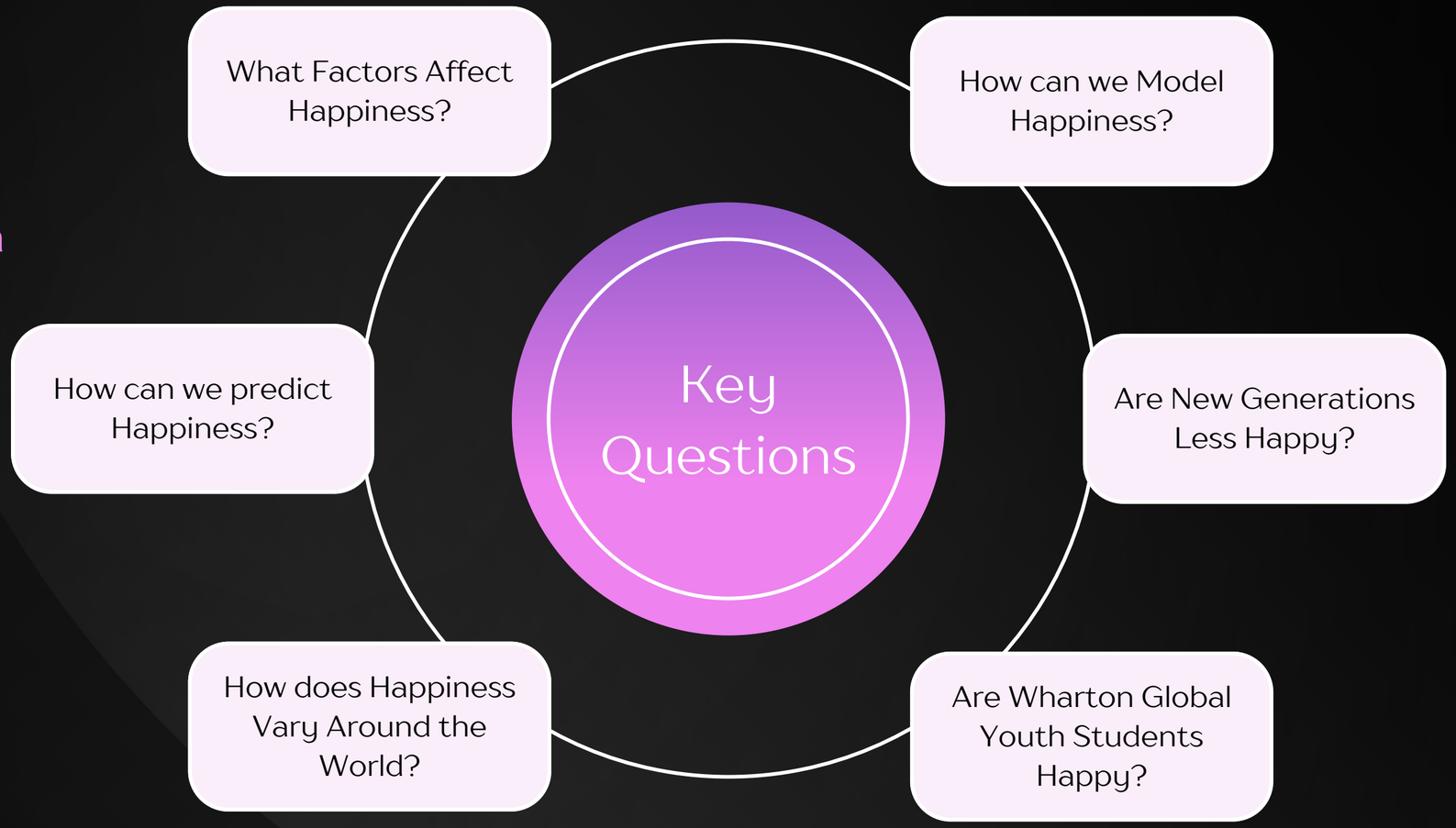
03 By determining the factors which correlate with the happiness rate, we are aiming to analyze the dataset.



https://worldhappiness.report/ed/2019/the-sad-state-of-happiness-in-the-united-states-and-the-role-of-digital-media/

Our Goals

What can we, as Data Scientists, study to make a difference?



Data set 1

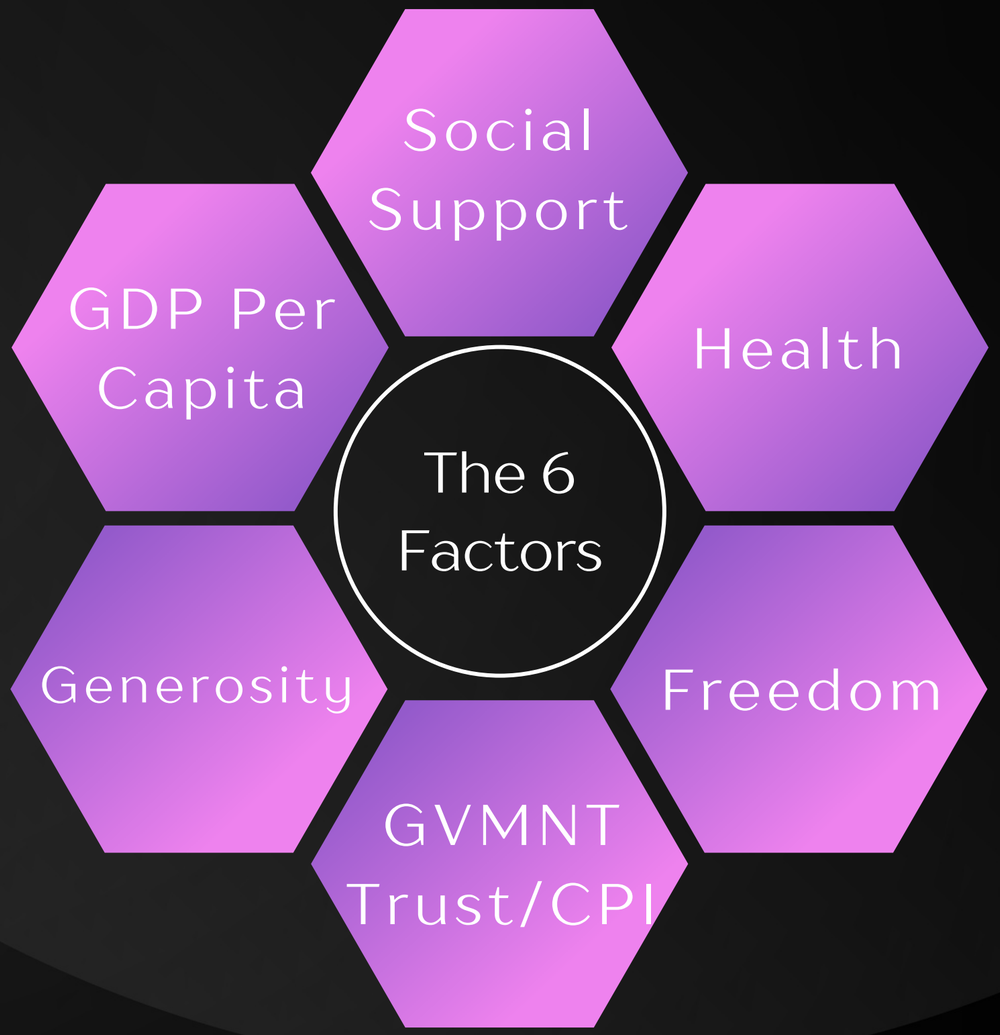
- World Happiness Score Report
- Calculates a Happiness Score for each Country using 6 Factors
- Uses Data from Gallup World Poll and the Cantril Ladder Visual

Home | The World Happiness Report

The World Happiness Report is a publication of the Sustainable Development Solutions Network, powered b...

HappinessRpt

World Happiness Report



Data set 2

- Published by the UN Sustainable Development Solutions Network
- World Data 2023
- 133 countries
- Country specific factors

https://unstats.un.org/unsd/undataforum/index.html

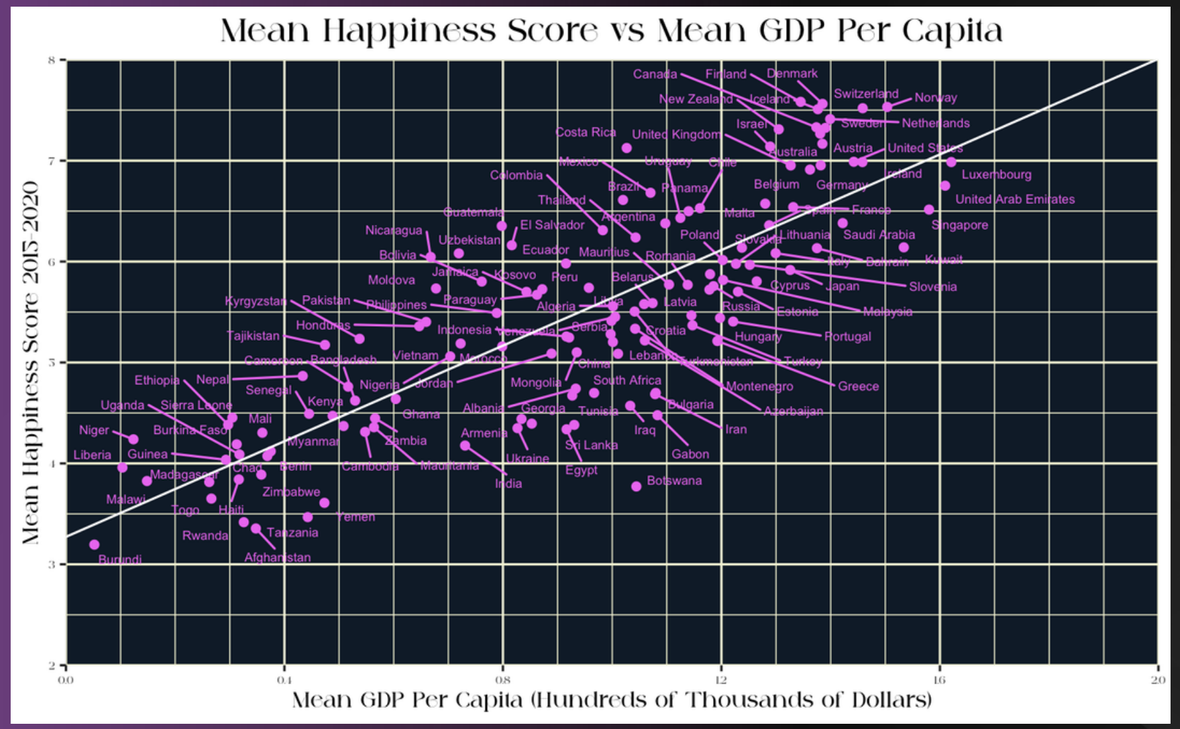
World Data 2023

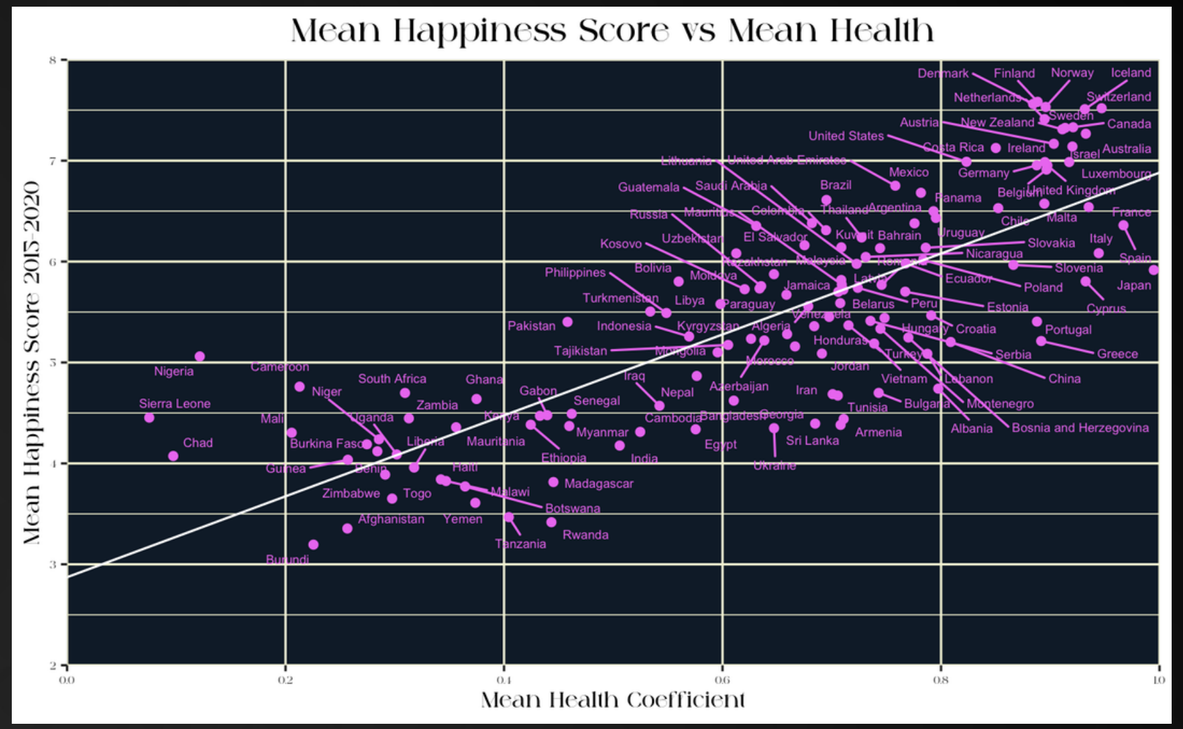
- Country
- Density (P/Km2)
- Abbreviation
- Agricultural Land (%)
- Land Area (Km2)
- Armed Forces Size
- Birth Rate
- Calling Code
- Capital/Major City
- CO2 Emissions
- CPI
- CPI Change (%)
- Currency_Code
- Fertility Rate
- Forested Area (%)
- Gasoline_Price
- GDP
- Gross Primary Education Enrollment (%)
- Gross Tertiary Education Enrollment (%)
- Infant Mortality
- Largest City
- Life Expectancy
- Maternal Mortality Ratio
- Minimum Wage
- Official Language
- Out of Pocket Health Expenditure (%)
- Physicians per Thousand
- Population
- Population: Labor Force Participation (%)
- Tax Revenue (%)
- Total Tax Rate
- Unemployment Rate
- Urban Population
- Latitude
- Longitude

36 Features -> 28 Features

Linear Regression & Multiple Regression

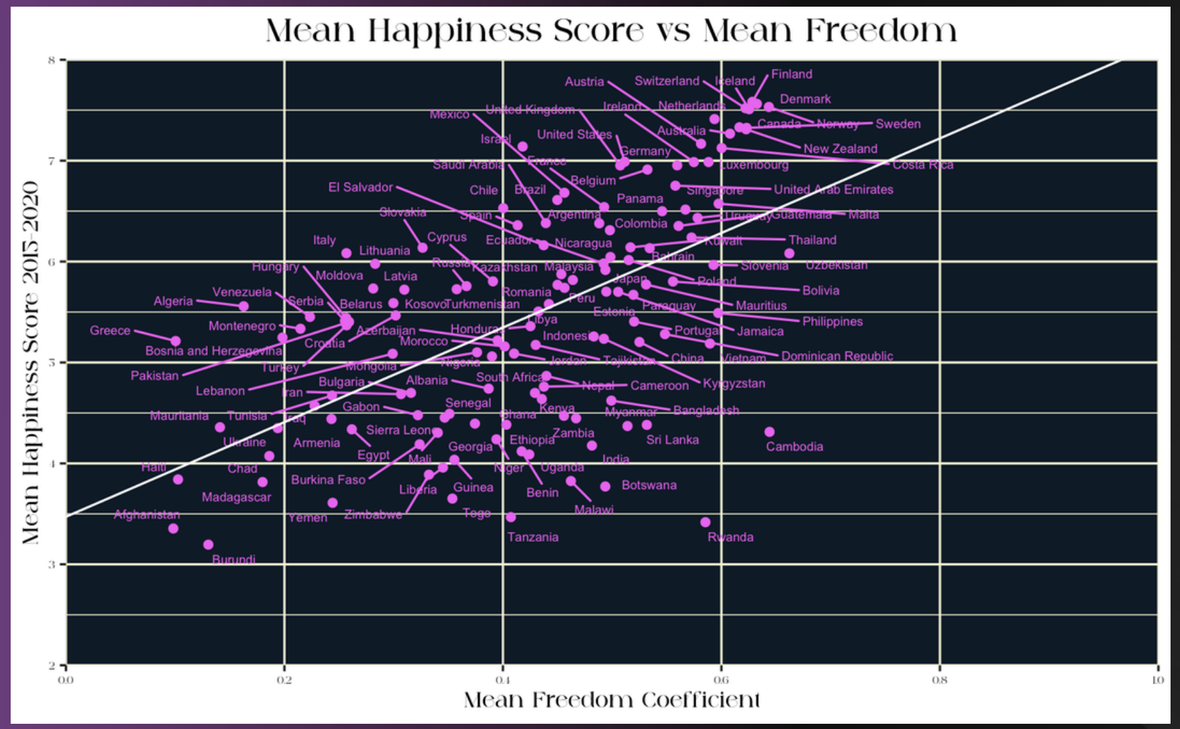
World Happiness Report

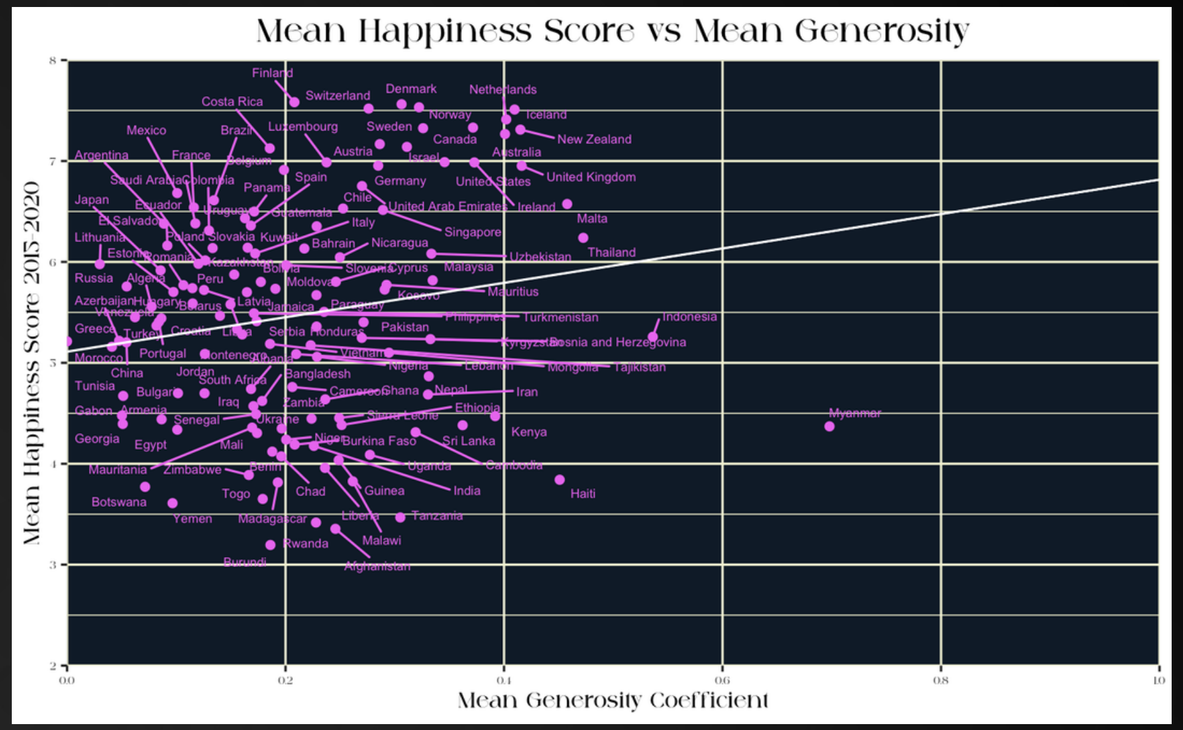




Linear Regression & Multiple Regression

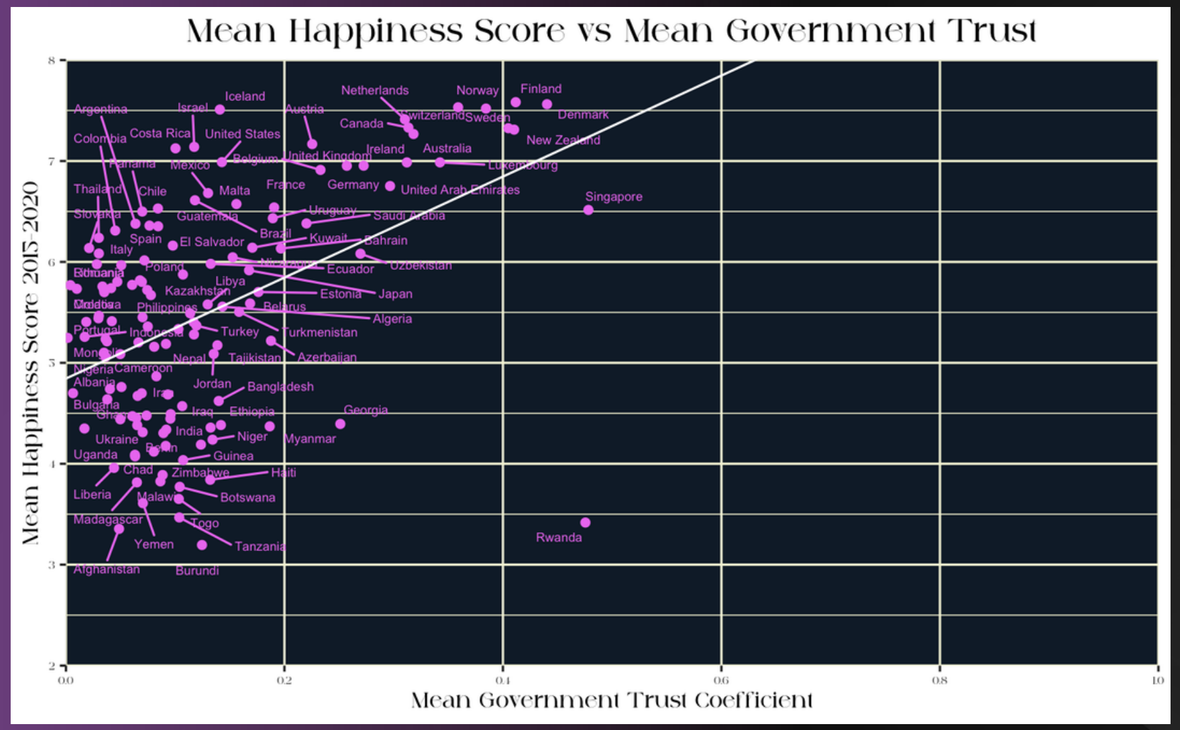
World Happiness Report

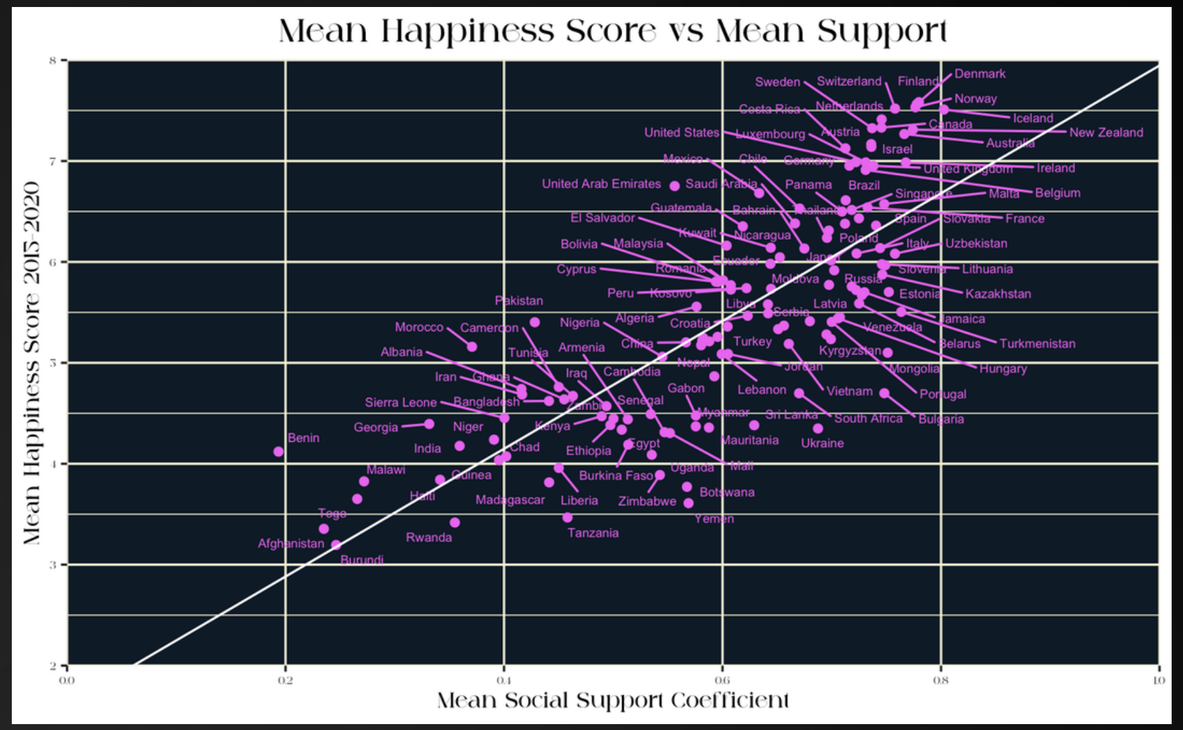




Linear Regression & Multiple Regression

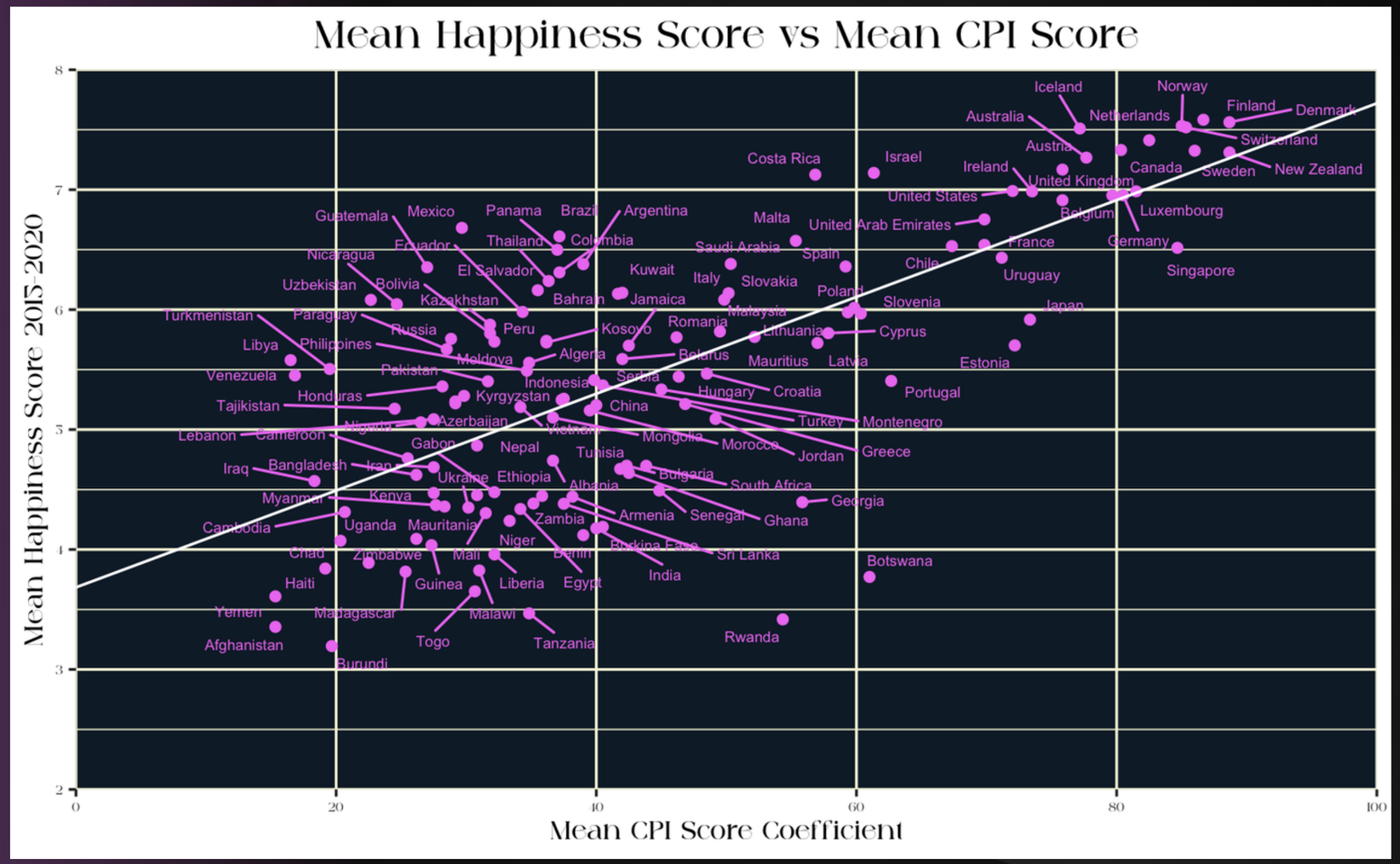
World Happiness Report





Linear Regression & Multiple Regression

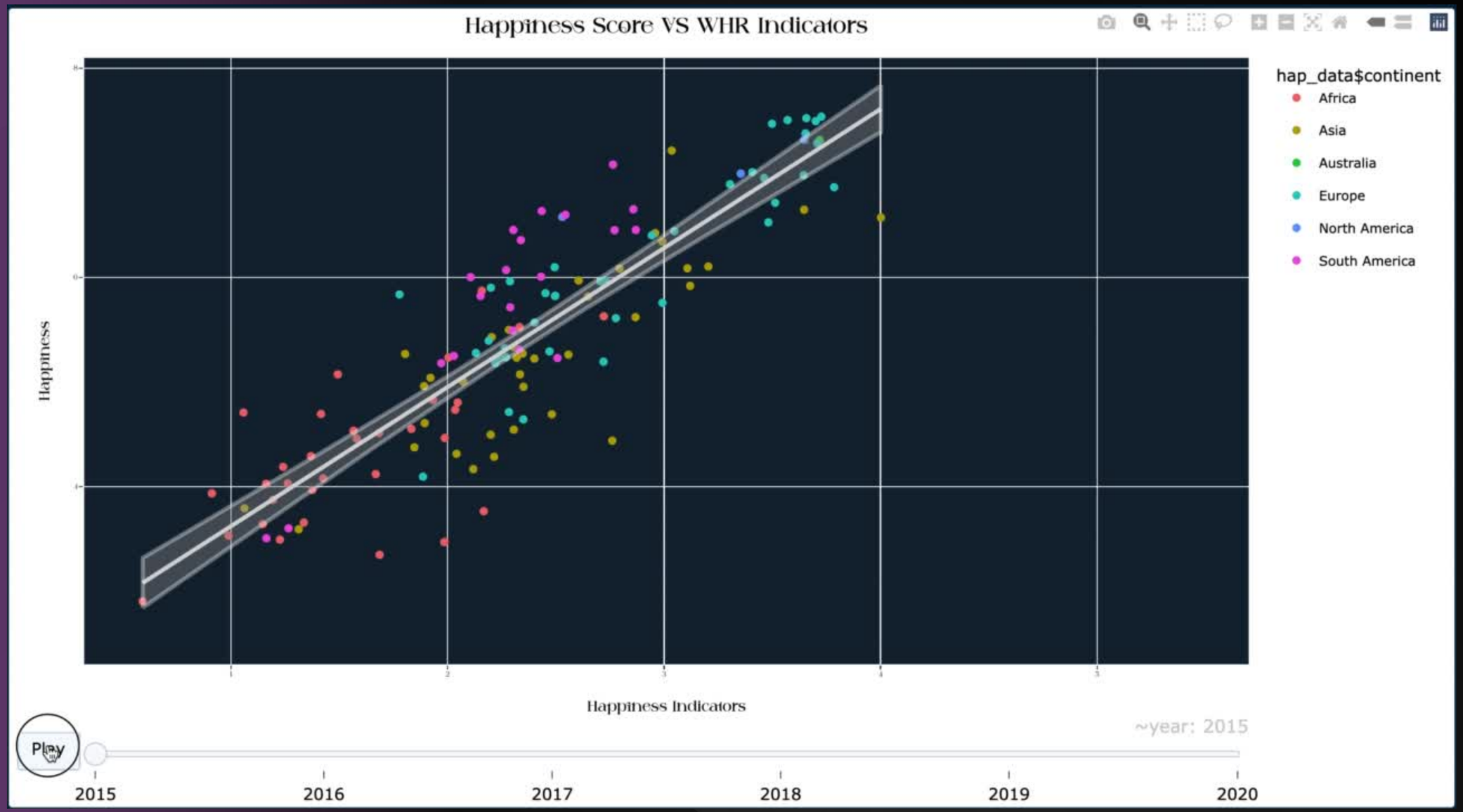
World Happiness Report



Equation for World Happiness

A multiple regression line constructed using the Betas of the 6 factors that define Happiness Score

$$y_{pred} = 2.3323 + 1.4215x_{gdp} + 1.3008x_h + 0.1.5899x_f + 0.8579x_g + 0.7382x_{gov} + 0.0382x_s$$



How can we predictively apply this Model?

We apply it to you.

To Us.

To Wharton.

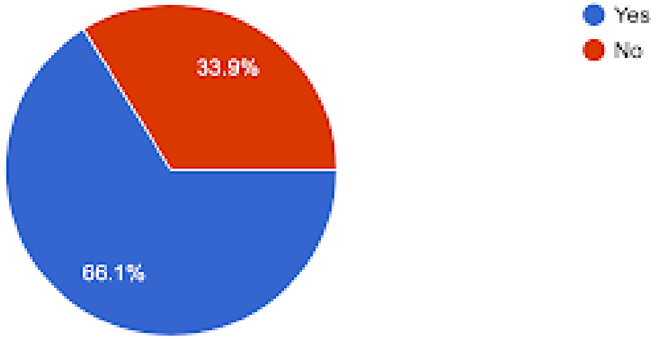
Wharton GYP Data

What is your current Program with Wharton Global Youth?

127 responses

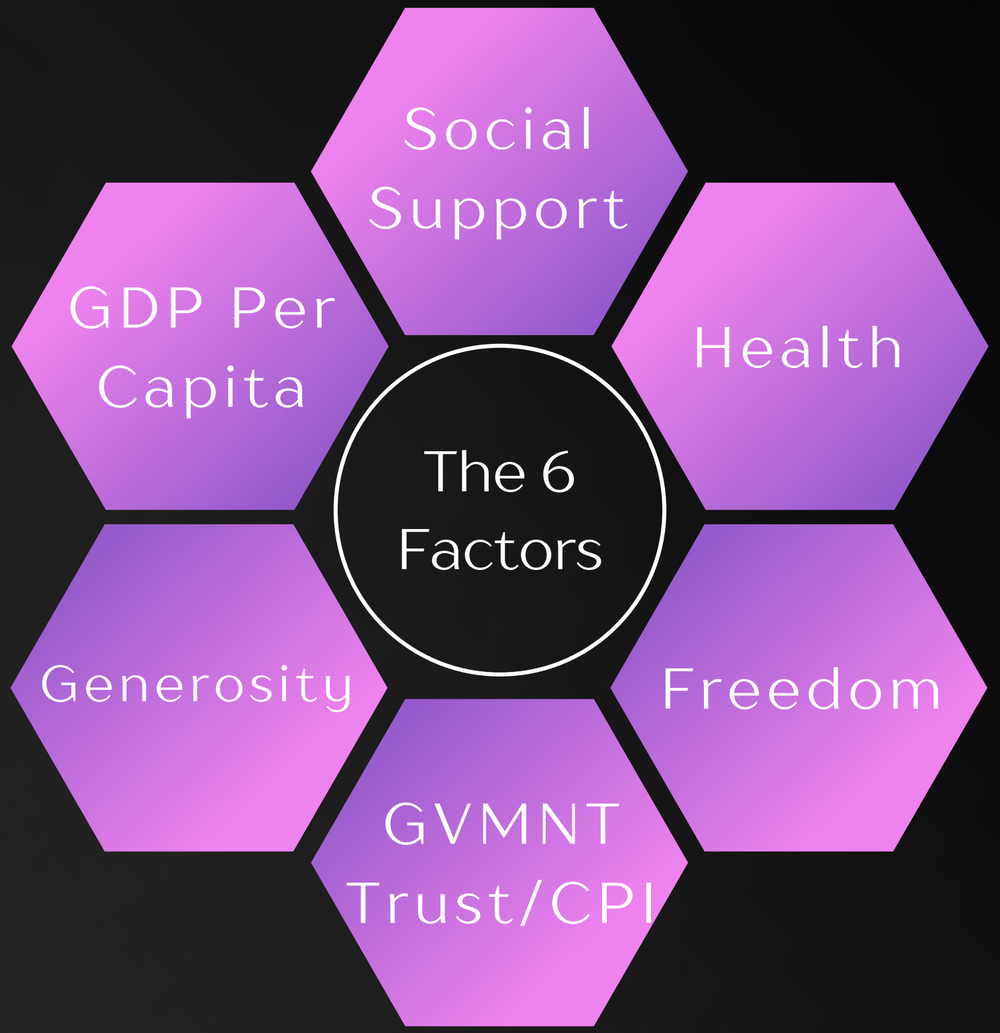
Do you live inside the United States of America (All Territories Included)?

127 responses



Wharton GYP Data

- Overall Comparison
- Happiness in Data Science VS LBW
- Happiness in USA Residents VS International Students



Wharton GYP Data

	Mean	Max	Min
Overall	4.08	6.18	0.83
DSA	4.28	5.72	0.83
LBW	3.96	6.18	2.33
USA Residents	4.06	6.18	0.83
International Residents	4.14	5.86	2.48

World Data 2023 Information

- Multiple Linear Regression Fit
- LASSO and Backward Elimination
- Decision Trees

Backward Elimination

```
Call:
lm(formula = happiness_score ~ . - country, data = no_na_country_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62473 -0.30738  0.00703  0.36056  1.71364 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               9.258e+00  3.768e+00   2.457 0.016342 *  
density                  -3.115e-04  4.091e-04  -0.761 0.448863    
agricultural_land        -3.685e-03  5.414e-03  -0.681 0.498268    
land_area                 4.813e-09  4.672e-08   0.103 0.918232    
armed_forces             -7.458e-07  5.554e-07  -1.343 0.183415    
birth_rate               -3.929e-02  6.183e-02  -0.635 0.527085    
carbon_emissions         -1.487e-07  3.455e-07  -0.431 0.668039    
cpi                      -9.776e-04  1.905e-03  -0.513 0.609406    
cpi_change               -2.537e-03  1.665e-02  -0.152 0.879285    
fertility_rate            3.149e-01  4.372e-01   0.720 0.473572    
forested_area             6.988e-03  5.250e-03   1.331 0.187235    
gasoline_price           -9.077e-02  3.422e-01  -0.265 0.791576    
gdp                       1.480e-14  8.083e-14   0.183 0.855217    
primary_edu_enrollment    5.678e-03  7.611e-03   0.746 0.457973    
tertiary_edu_enrollment   8.231e-03  5.155e-03   1.597 0.114581    
infant_mortality         -2.726e-04  1.476e-02  -0.018 0.985316    
life_expectancy          -4.810e-02  4.380e-02  -1.098 0.275652    
maternal_mortality       -2.809e-03  8.683e-04  -3.235 0.001818 ** 
minimum_wage              1.320e-01  4.226e-02   3.124 0.002550 ** 
pocket_health_expenditure 4.559e-03  5.648e-03   0.807 0.422090    
physicians_per_thousand   1.242e-01  8.934e-02   1.390 0.168766    
population               -8.350e-10  2.267e-09  -0.368 0.713731    
labor_force              -9.626e-03  1.087e-02  -0.886 0.378702    
tax_revenue               1.230e-03  1.620e-02   0.076 0.939711    
total_tax                -1.308e-03  6.422e-03  -0.204 0.839146    
unemployment_rate        -7.546e-02  2.184e-02  -3.456 0.000912 ***
urban_population          6.065e-09  5.313e-09   1.141 0.257351    
latitude                 -2.638e-03  3.906e-03  -0.675 0.501493    
longitude                -4.446e-03  1.435e-03  -3.099 0.002743 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6719 on 74 degrees of freedom
Multiple R-squared:  0.7293,	Adjusted R-squared:  0.6268 
F-statistic: 7.119 on 28 and 74 DF,  p-value: 6.751e-12
```

```
Call:
lm(formula = happiness_score ~ armed_forces + forested_area + 
    tertiary_edu_enrollment + life_expectancy + maternal_mortality + 
    minimum_wage + unemployment_rate + urban_population + longitude, 
    data = no_na_country_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.71467 -0.32501  0.00207  0.33956  1.71659 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              8.982e+00  1.829e+00   4.912 3.86e-06 ***
armed_forces            -8.101e-07  3.181e-07  -2.546 0.012529 *  
forested_area            8.220e-03  3.113e-03   2.640 0.009715 ** 
tertiary_edu_enrollment  1.349e-02  3.428e-03   3.934 0.000161 ***
life_expectancy         -5.050e-02  2.500e-02  -2.020 0.046297 *  
maternal_mortality      -2.894e-03  6.693e-04  -4.324 3.85e-05 ***
minimum_wage             1.298e-01  2.902e-02   4.473 2.18e-05 ***
unemployment_rate       -5.785e-02  1.543e-02  -3.750 0.000307 ***
urban_population         3.381e-09  1.409e-09   2.399 0.018414 *  
longitude               -4.086e-03  1.140e-03  -3.584 0.000541 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6327 on 93 degrees of freedom
Multiple R-squared:  0.6983,	Adjusted R-squared:  0.6691 
F-statistic: 23.92 on 9 and 93 DF,  p-value: < 2.2e-16
```

Relaxed LASSO

$$\min_{\beta_0, \beta_1, \beta_2, \ldots, \beta_p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2 + \lambda(|\beta_1| + |\beta_2| + \cdots + |\beta_p|) \right\}$$

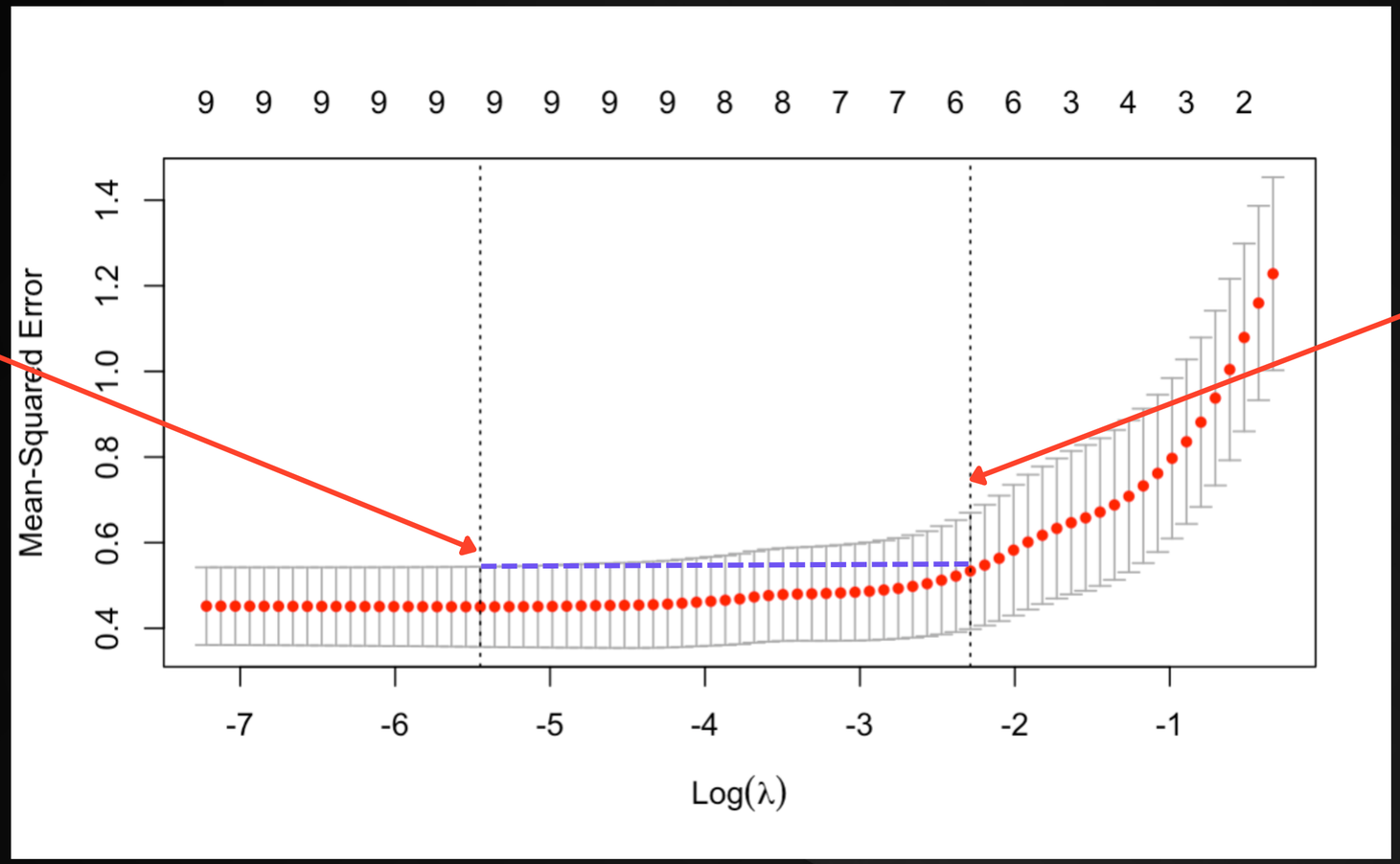
(Intercept)	5.5125049
armed_forces	-0.3283747
forested_area	0.1612251
tertiary_edu_enrollment	0.3822414
life_expectancy	-0.2911267
maternal_mortality	-0.5650687
minimum_wage	0.4066356
unemployment_rate	-0.2437893
urban_population	0.2977736
longitude	-0.2342157

Find the minimum beta values using the L1 penalization. Then, construct a multiple linear regression model for prediction.

$$y_{pred} = 5.513 - 0.328x_{af} + 0.161x_{fa} + 0.382x_{tee} - 0.291x_{le} - 0.565x_{mm} + 0.407x_{mw} - 0.244x_{ur} + 0.298x_{up} - 0.234x_{l}$$

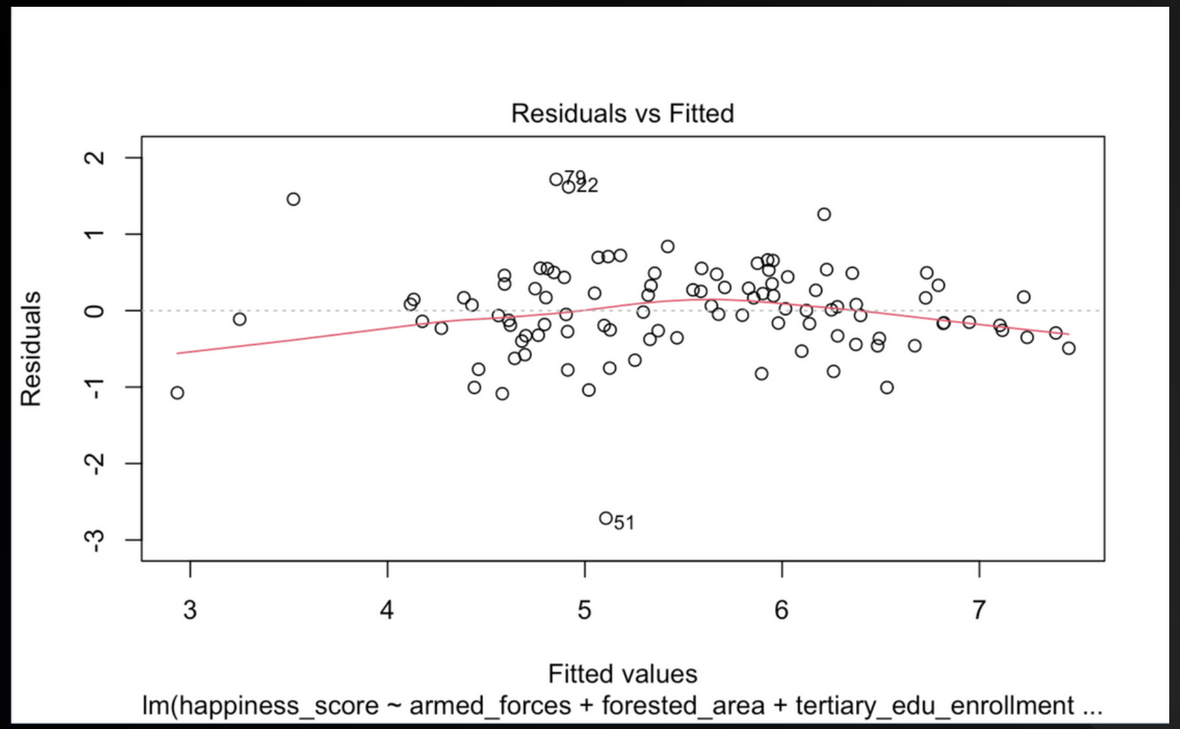
Lambda Min & 1se

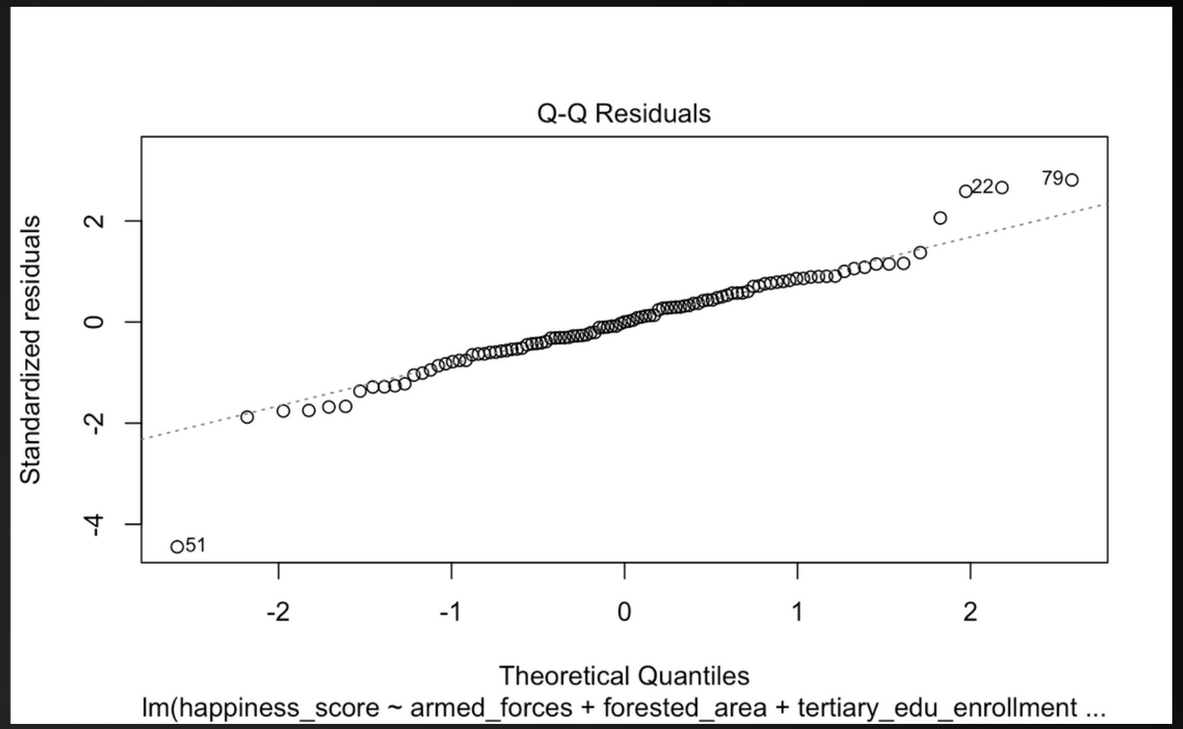
0.003912633



0.1114332

Assumptions





Random Test Prediction

Regarding that the data of Ecuador was not used in the model training, we could predict the happiness score of Ecuador using the other columns.

```
fit.lm <- lm(happiness_score ~ armed_forces + forested_area + tertiary_edu_enrollment + life_expectancy + maternal_mortality + minimum_wage + unemployment_rate +
urban_population + longitude, data = country_data)
```

Actual

5.56

Predict

6.26

Testing Error & Training Error

Testing Error

-0.187

Training Error

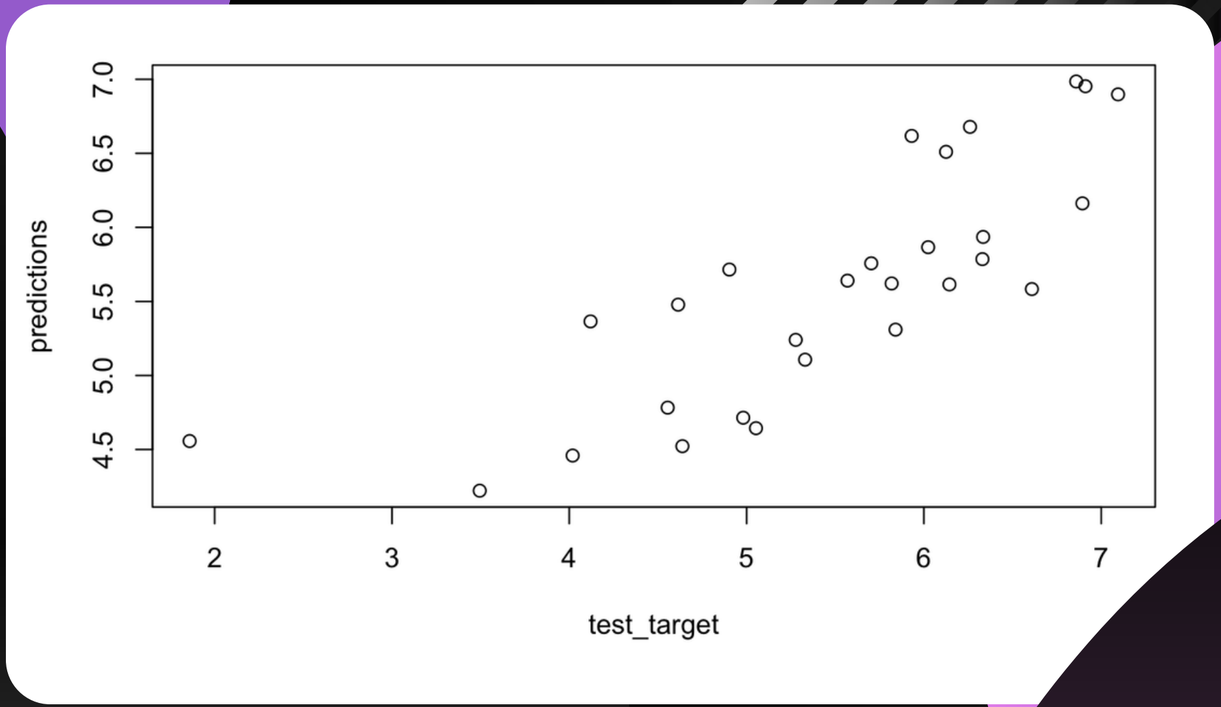
0.00000000000000213

Testing Error is minimal, meaning that the model is accurate

Obviously, the Training Error is miniscule which proves that the model is not overfitted.

Random Forest

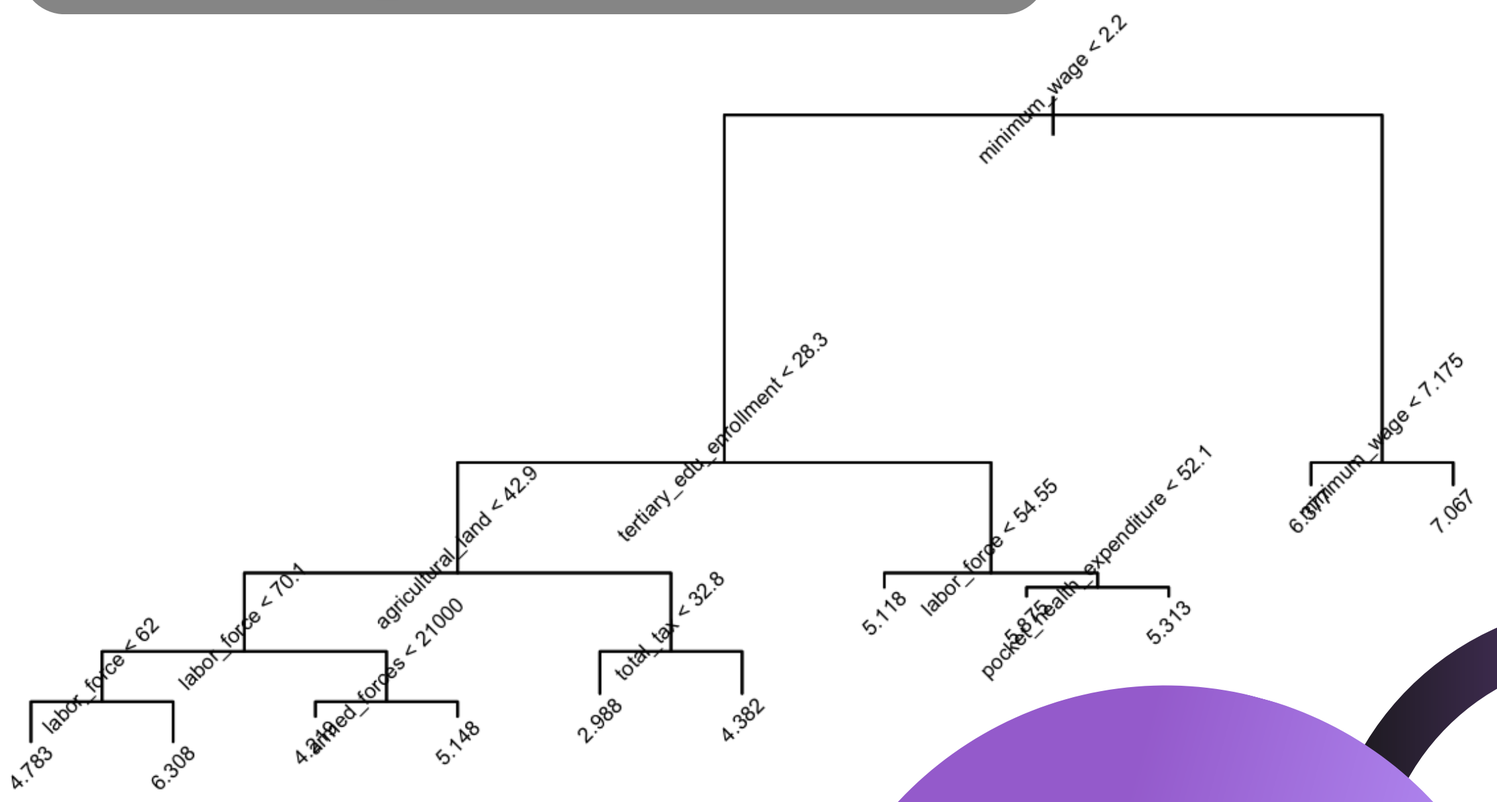
Although a Random Forest is not the best fit for our low sample size, we wanted an additional model that, able to handle complex relationships.



With a 61% R squared, our random forest model is well equipped to provide

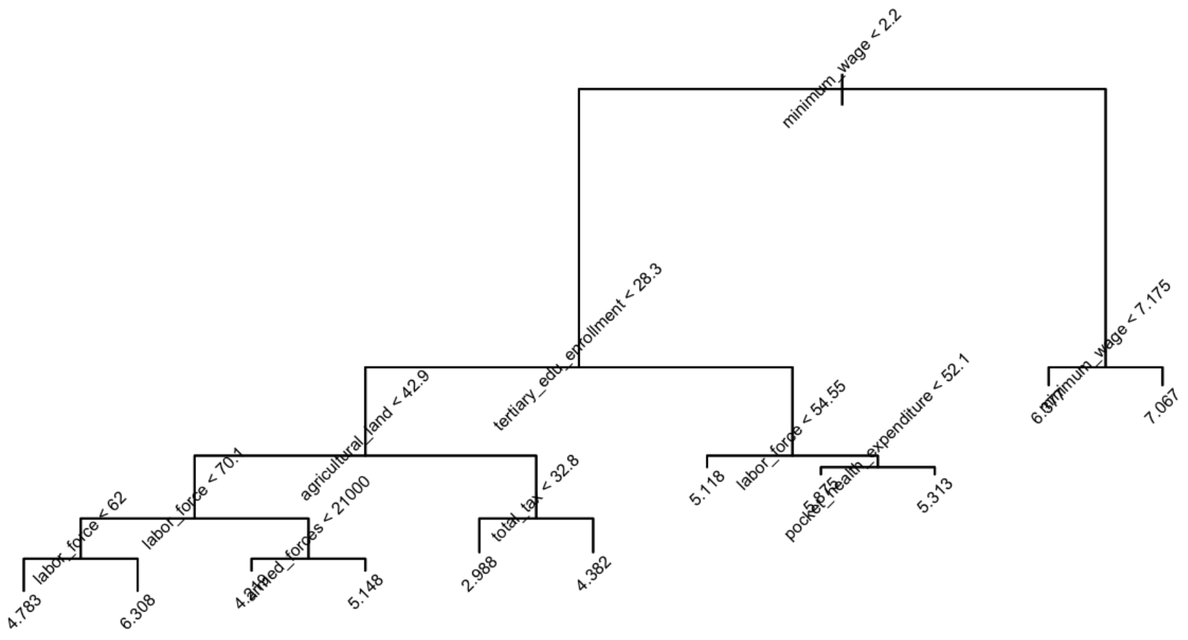
Affected by our sample size, the RMSE sits at about 0.7, which should influence any insights from this model.

Single Tree



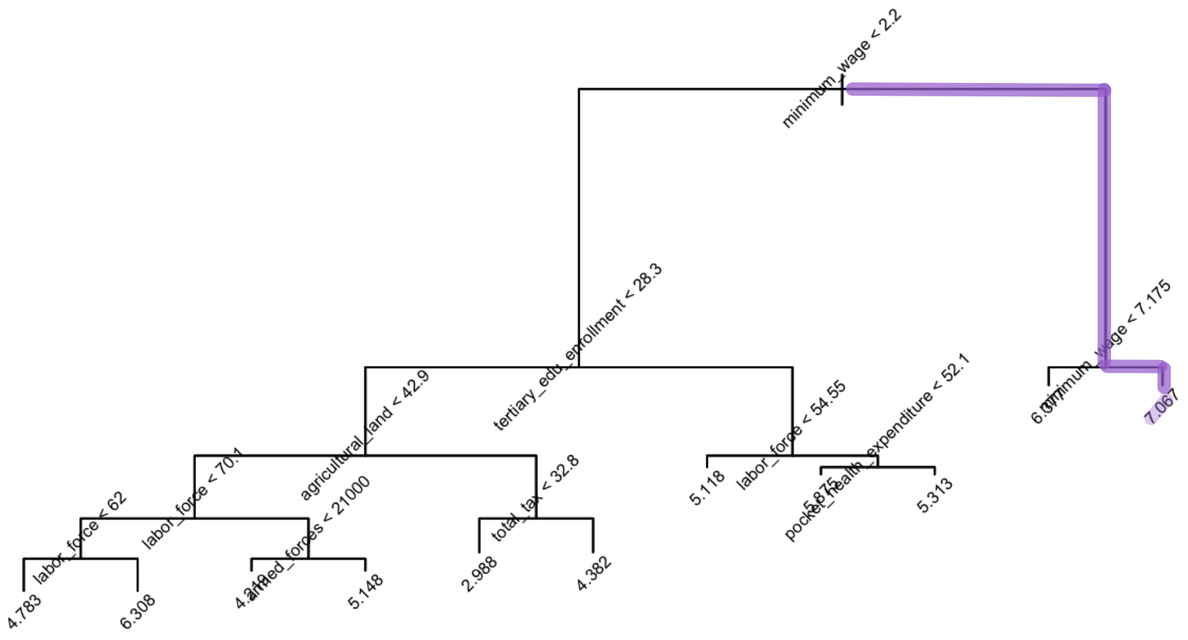
Taking a tree from our forest, we have a very interpretable model, which also offers insight on the effects of our variables on happiness

Decision Tree Testing: Canada



Minimum Wage	**9.51**
Tertiary Education Enrollment	68.9
Agricultural Land	6.9
Labour Force	65.1

Decision Tree Testing: Canada



7.06

6.96

Conclusion

This data case study offered a meticulous and comprehensive analysis of the World Data 2023 data set, shining light on the relationship between country-specific parameters and happiness. The model is empirically tested through regression analysis, allowing us to effectively anticipate happiness levels in countries through quantifying the factors that affect a country.

Armed Forces	Forested Area	Life Expectancy
Tertiary Education	Maternal Mortality	Minimum Wage
Unemployment Rate	Urban Population	Longitude
Population	Agricultural Land	

Limitations

Sources of error for our multiple regression

1 Wharton Data
a. Actual Happiness Score
b. Assumption that variables can correspond

2 World 2023
Small dataset