World Happiness Report

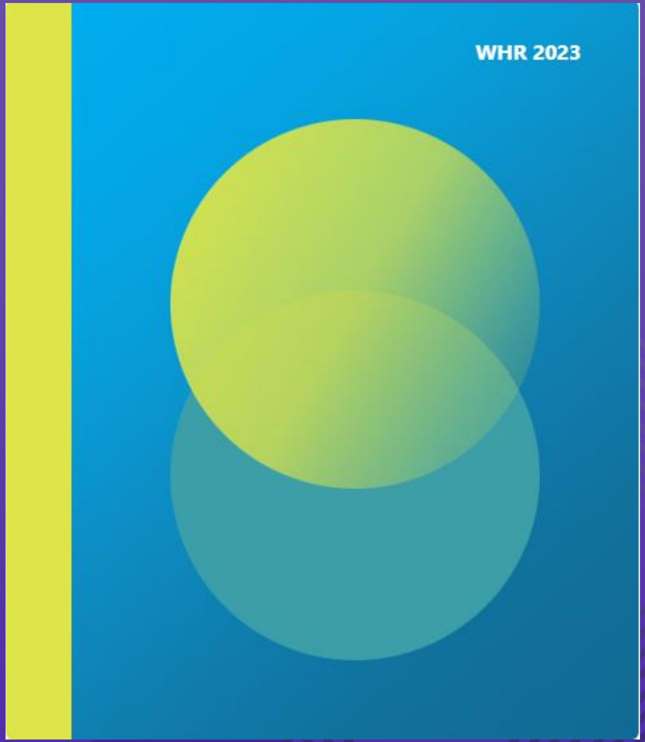
SC1015 B128 Team 2

Lim Sin Le Dion, Marvin Beh Chee Wen, Merwyn Masagca

Table of contents

Context

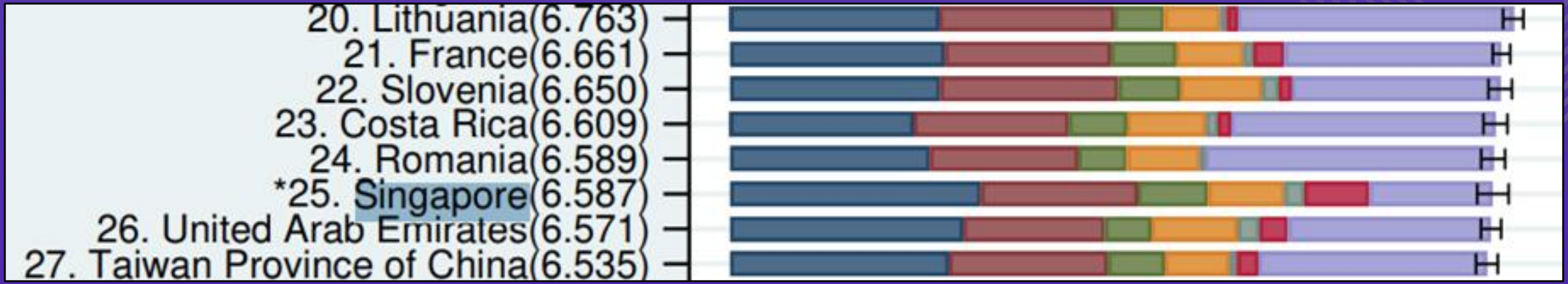
- Annual report published by United Nations Sustainable Development Solutions Network (SDSN)
- Based on a variety of factors
- Widely recognised as a key resource for understanding global happiness and well-being



Our Motivation

Singapore was crowned the happiest country in Asia and ranked 25th in the UN's World Happiness Report 2023.

Our project aims to understand which factors are the most important in predicting happiness in a country and hence build a model that best predicts happiness in a country.



01 Problem Statement

Problem Statement

To build a model that best predicts the factor life ladder, and compare them using various regression models

02

Exploratory Data Analysis

Data Cleaning & Analysis of Data

Data Preparation

Filling in missing data

Filling up missing data within the dataset using the median

Removal of irrelevant variables

Variables we deem as irrelevant to our problem statement

Removal of low correlation variables

Dropping variables due to their low correlation to life ladder

Fill missing data with median & dropping irrelevant variables

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2199 entries, 0 to 2198
Data columns (total 11 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Country name                      2199 non-null   object
 1   year                              2199 non-null   int64
 2   Life Ladder                       2199 non-null   float64
 3   Log GDP per capita                2179 non-null   float64
 4   Social support                    2186 non-null   float64
 5   Healthy life expectancy at birth  2145 non-null   float64
 6   Freedom to make life choices      2166 non-null   float64
 7   Generosity                        2126 non-null   float64
 8   Perceptions of corruption         2083 non-null   float64
 9   Positive affect                   2175 non-null   float64
 10  Negative affect                   2183 non-null   float64
dtypes: float64(9), int64(1), object(1)
memory usage: 189.1+ KB
```

```
df.fillna(value = df.median(), inplace = True)
dfnew = df.drop(columns = ['year', 'Country name'])
dfnew.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2199 entries, 0 to 2198
Data columns (total 9 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Life Ladder                       2199 non-null   float64
 1   Log GDP per capita                2199 non-null   float64
 2   Social support                    2199 non-null   float64
 3   Healthy life expectancy at birth  2199 non-null   float64
 4   Freedom to make life choices      2199 non-null   float64
 5   Generosity                        2199 non-null   float64
 6   Perceptions of corruption         2199 non-null   float64
 7   Positive affect                   2199 non-null   float64
 8   Negative affect                   2199 non-null   float64
dtypes: float64(9)
memory usage: 154.7 KB
```

Variables dropped

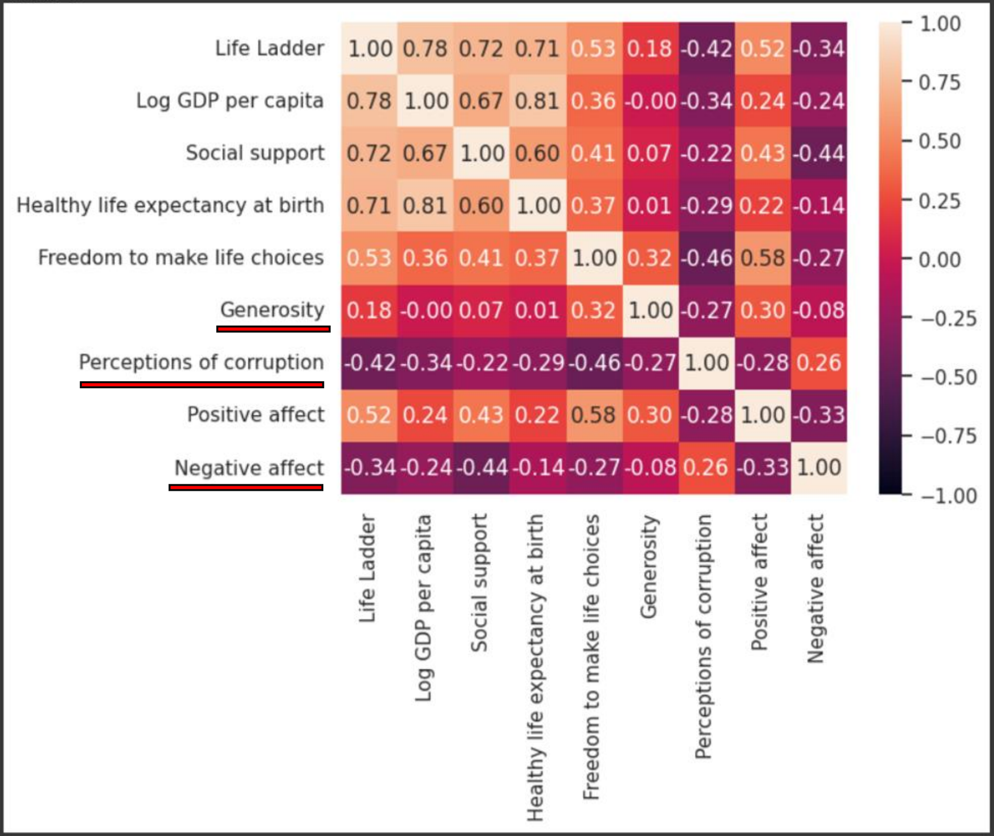
1. Year
2. Country name

Heatmap to visualise correlation

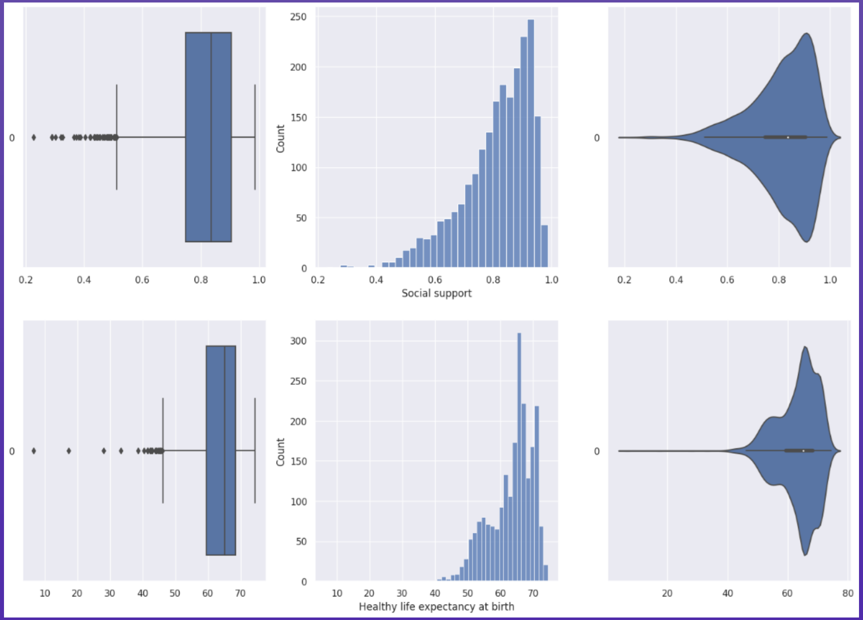
Find out how strong the relationships of variables are to one another

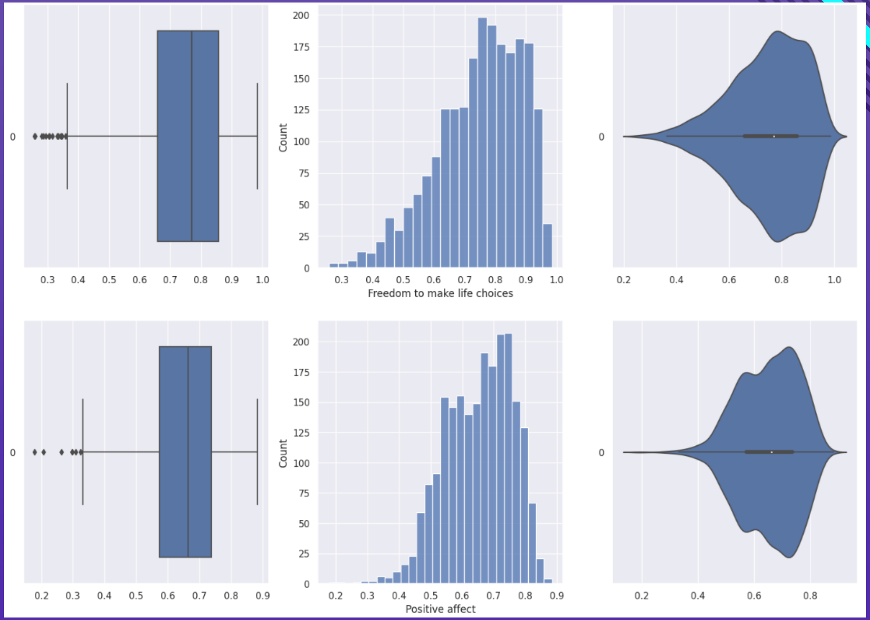
```
sb.heatmap(dfnew.corr(), vmin = -1, vmax = 1, annot = True, fmt = ".2f")
```

<Axes: >



Variables with outliers





Variables in order of most outliers to least

1. Social Support
2. Healthy life expectancy at birth
3. Freedom to make life choices
4. Positive affect
5. Log GDP per capita

```
Life Ladder                          1
Log GDP per capita                   1
Social support                      52
Healthy life expectancy at birth    26
Freedom to make life choices        15
Positive affect                      7
dtype: int64
```

Dropping outliers

```
[ ] # Find the rows where ANY column is True
    outliers = rule.any(axis = 1)   # axis 0 is row, 1 is column
    outliers.value_counts()
```

```
False    2111
True       88
dtype: int64
```

```
[ ] # Which row indices correspond to outliers in the dataframe?
    outliertrue = outliers.index[outliers == True]
    outliertrue
```

```
Int64Index([   0,    5,    9,   10,   11,   12,   13,  134,  147,  148,  181,
             182,  183,  186,  187,  188,  189,  190,  191,  192,  193,  194,
             215,  216,  218,  288,  289,  290,  292,  344,  345,  346,  347,
             348,  349,  350,  420,  475,  610,  676,  766,  767,  768,  769,
             770,  771,  774,  837,  883,  884, 1114, 1116, 1117, 1118, 1119,
            1120, 1121, 1168, 1179, 1186, 1192, 1333, 1335, 1336, 1337, 1487,
            1488, 1491, 1637, 1647, 1709, 1759, 1830, 1884, 1950, 1951, 1952,
            1953, 1954, 1955, 1956, 1995, 2134, 2182, 2183, 2184, 2185, 2186],
           dtype='int64')
```

```
# Remove the outliers based on the row indices obtained above
dfclean.drop(axis = 0,             # 0 drops row 1 drops column
            index = outliertrue, # this takes a list as input
            inplace = True)         # not overwritten by default

dfclean
```

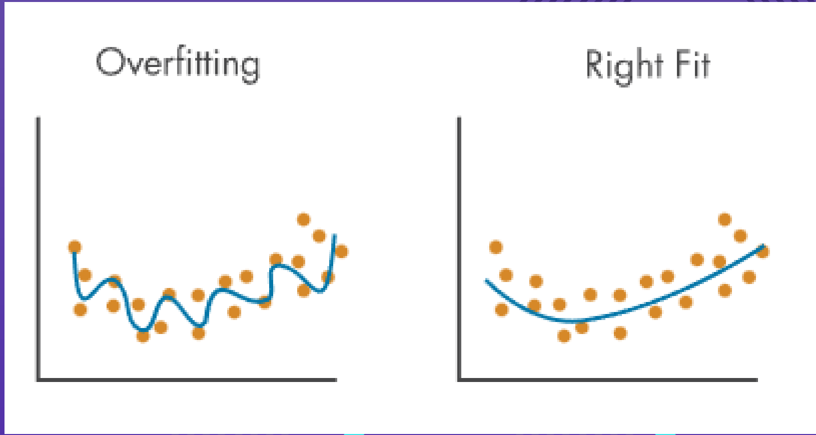
Overfitting

What is overfitting?

- A model thats fits the training set well but testing set poorly

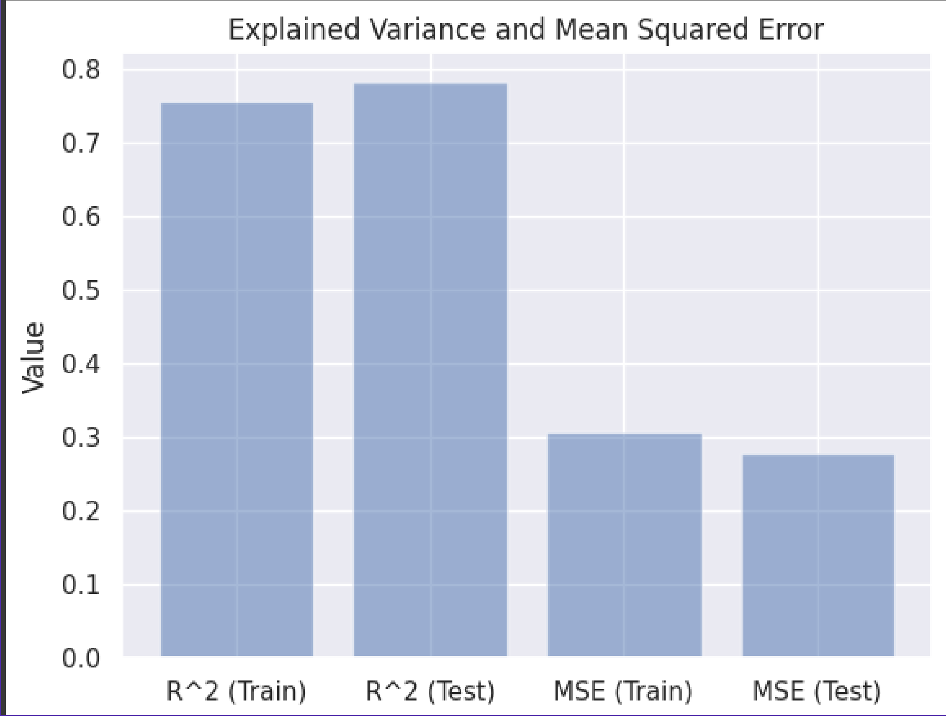
Overfitting can occur due to a model being trained ”too well”, in other words,

- The model is too custom-tailored to the train data such that it would perform poorly when placed in a different sets of data.
- Such a model will not fit future observations and affect accuracy of prediction of future data.



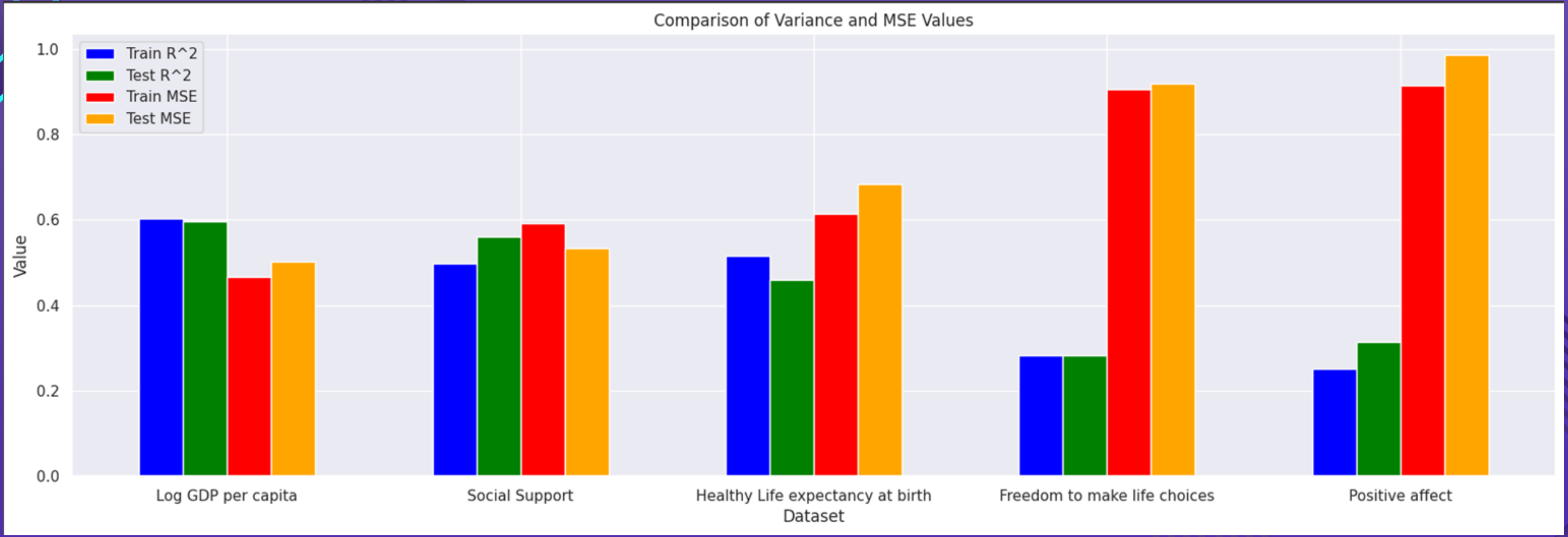
Checking for overfitting

```
Explained Variance (R^2) on Train Set   : 0.7565295551398689
Explained Variance (R^2) on Test Set    : 0.7830400641183496
Mean Squared Error (MSE) on Train Set   : 0.30730671948178584
Mean Squared Error (MSE) on Test Set    : 0.278114018425742
```



Our model does not show signs of overfitting

Uni-variate Linear Regression



Factors most likely to affect happiness in a country

Log GDP per capita

Social Support

Healthy Life Expectancy at Birth

We aim to build a regression model solely using these predictors and see whether there are improvements compared to using the model that includes all predictors.

03

Machine Learning

Building regression model

Machine Learning

Metric for comparison

1. Explained Variance (R^2)

1. Mean Squared Error (MSE)

Machine Learning

Regression models

1. Multi-Variate Linear Regression
2. K-fold Multi-Variate Linear Regression
3. Random Forest Regression
4. Ridge Regression

How the models work

Multi-variate Linear regression	Fits a linear model that uses least squares approach to minimise residual sum of squares between observed data in the dataset and the predicted data by linear approximation
K-Fold Multivariate Linear Regression	K-fold linear regression is a variant of the linear regression that uses k-fold cross-validation to evaluate the performance of the model and improve its accuracy

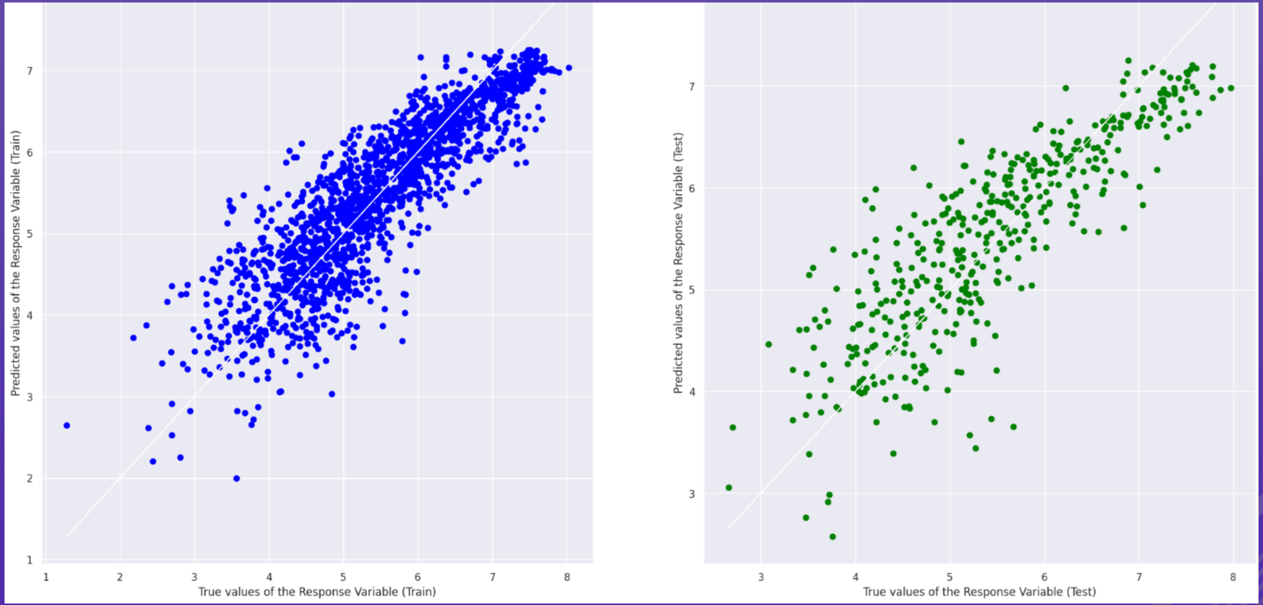
How the models work

Random Forest Regression	Random forest regression uses an ensemble of decision trees to perform regression tasks. In a random forest, multiple decision trees are trained independently on different subsets of the data, and their predictions are combined to make a final prediction
Ridge Regression	Ridge Regression uses L2 regularization to prevent overfitting, By adding a penalty term to the loss function that penalizes large values of the model's coefficients

Multi-Variate Linear Regression

Explained Variance $(R^2) = 0.71$

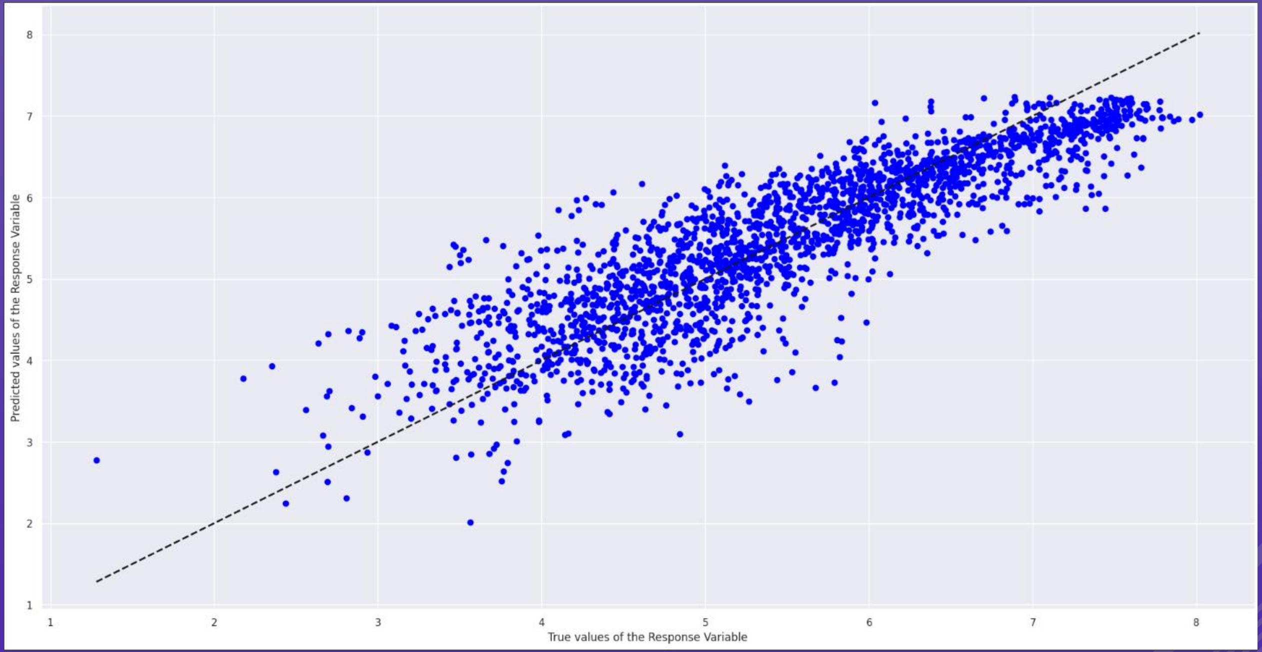
Mean Squared Error (MSE) = 0.35



K-fold Multi-Variate Linear Regression

Explained Variance (R^2) = 0.76

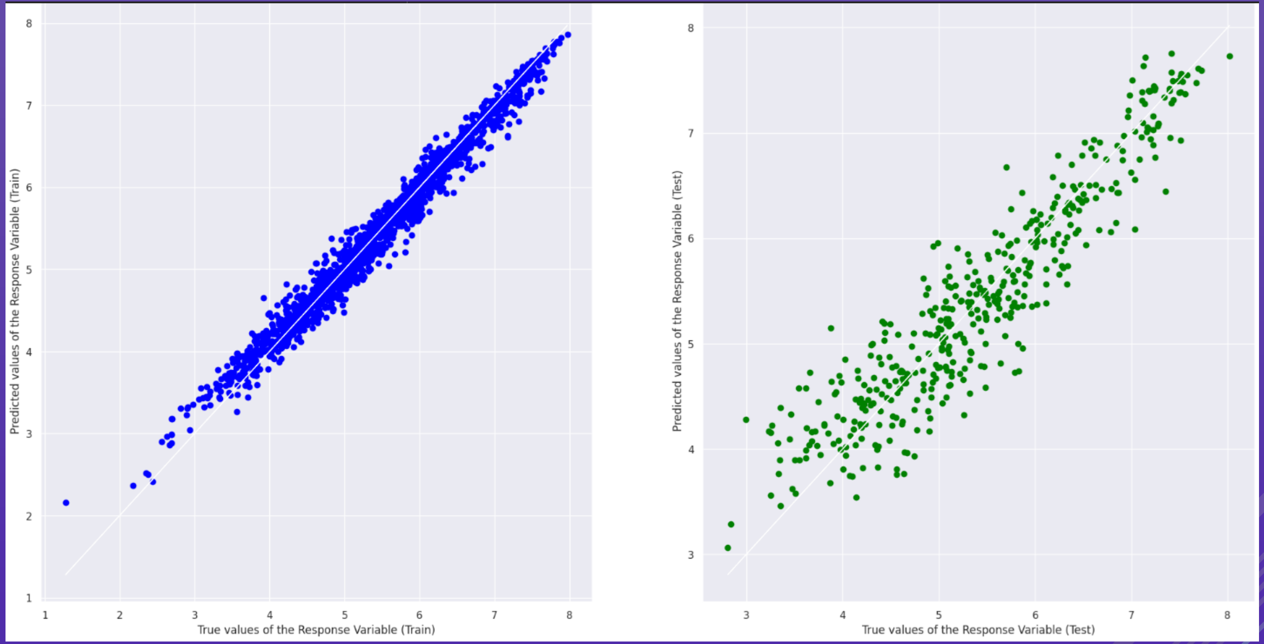
Mean Squared Error (MSE) = 0.30



Random Forest Regression

Explained Variance (R^2) = 0.87

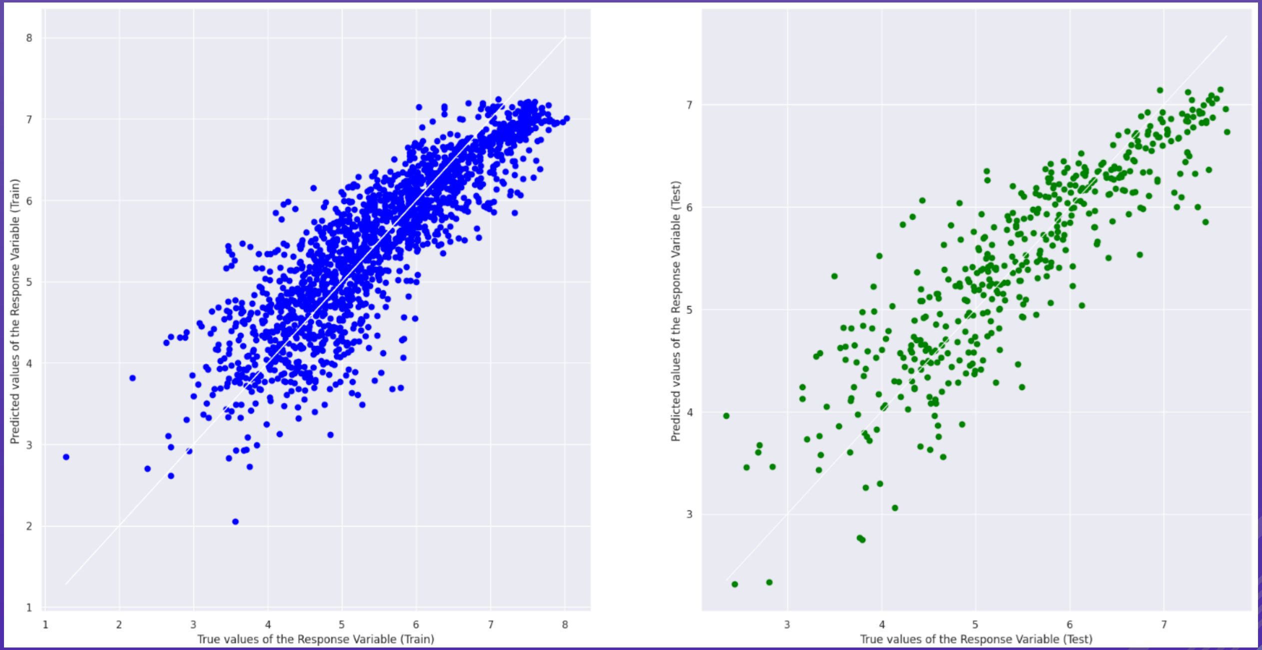
Mean Squared Error (MSE) = 0.16



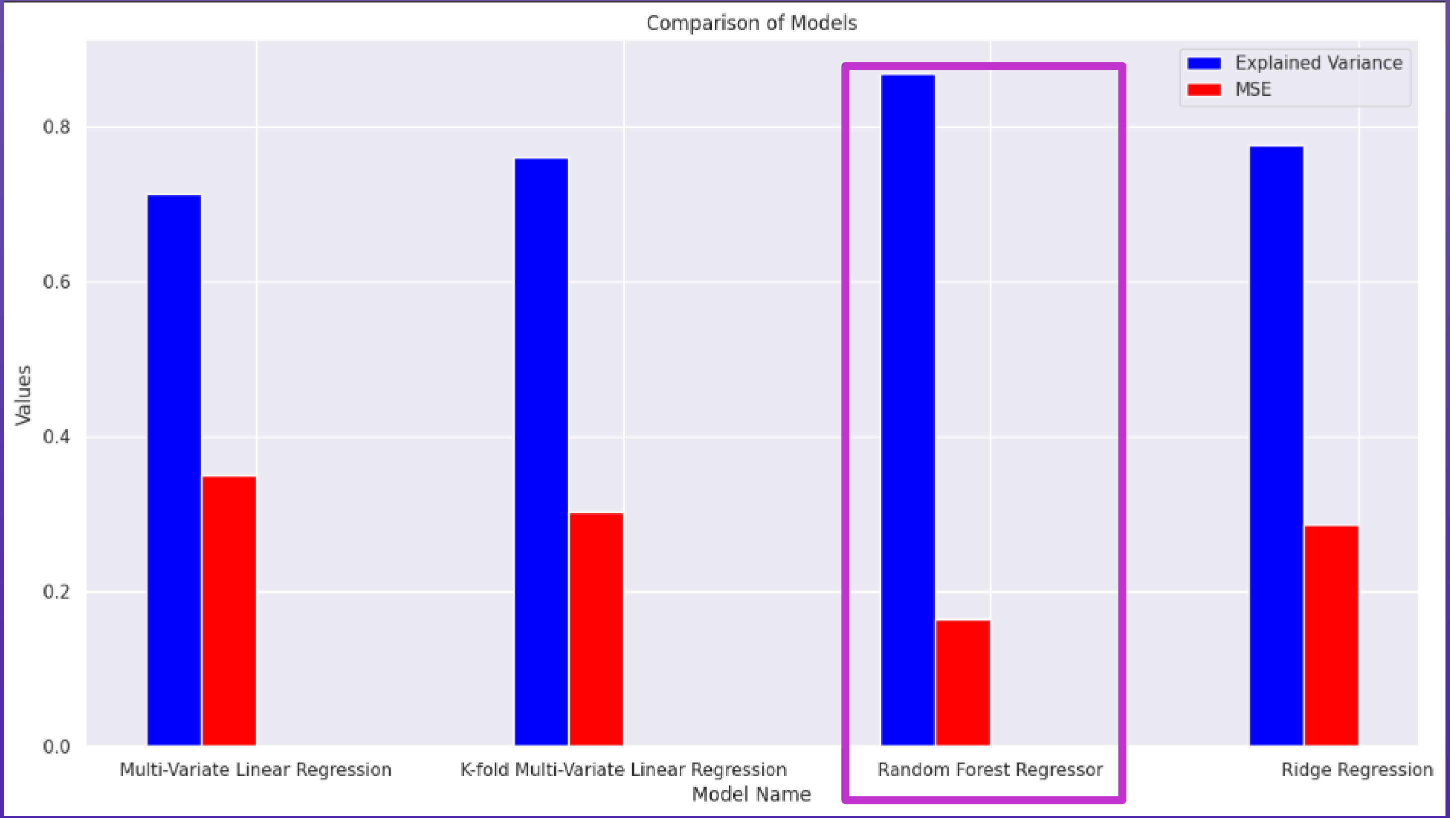
Ridge Regression

Explained Variance (R^2) = 0.78

Mean Squared Error (MSE) = 0.29



Comparison of Models



Random Forest Regressor is the best model

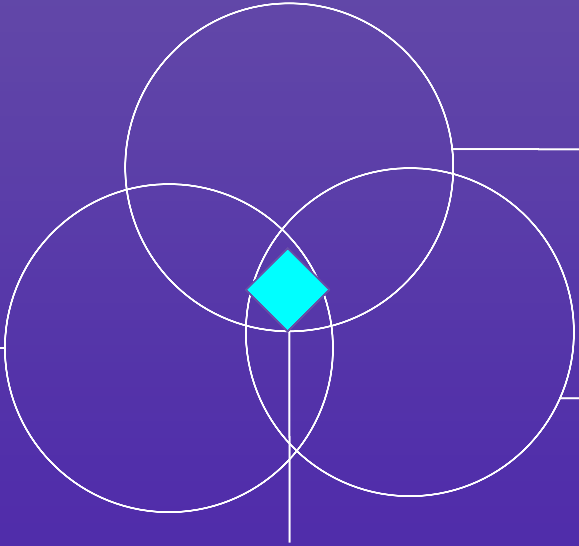
04

Conclusion

Most Important Factors, Which Regression Model to Use, Recommendations

Most Important Factors to Happiness in a country

3. Healthy Life expectancy at birth



1. GDP per capita

2. Social Support

Recommendation:
Maximise these factors

Thank You