WORLD HAPPINESS REPORT

Cheong Jin Hui, Darren Choo, Darren Wong
DFF2, Group 3

What is the world happiness report?

Landmark survey of the state of global happiness.

Contains articles and rankings of national happiness, based on respondent ratings of their own lives (correlates to life factors).

Our Motivation

We chanced upon an article about Singapore's happiness score faring at 6.4 points
(highest amongst its immediate regional neighbour countries)

We wanted to figure out what factors affected the happiness score of a country → sieve out most important

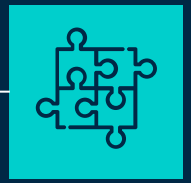
Use existing data to build a model to **predict** future data.

Singapore	6.4
Thailand	6
Philippines	5.9
Malaysia	5.4
Indonesia	5.3

Problem Statement

To build a model that best predicts the life ladder of different countries considering various regression models.

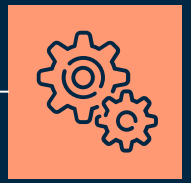
Data Preparation



01

Removal of variables

The variables are removed due to irrelevance or missing data



02

Filling up of missing data

We filled up the missing data using the median

03

Removal of additional variables

The variables are removed because of their low correlation

Exploratory Data Analysis

Analysis of Missing and Irrelevant Data.

Finding the relationship between the variables and life ladder and choosing to drop certain variables

Analysis of Outliers of the different variables.

Exploratory Analysis & Data Preparation

Irrelevant Data to be dropped

- Country name
- Standard deviation of ladder by country-year
- Standard deviation/Mean of ladder by country-year

Identifying Missing Data

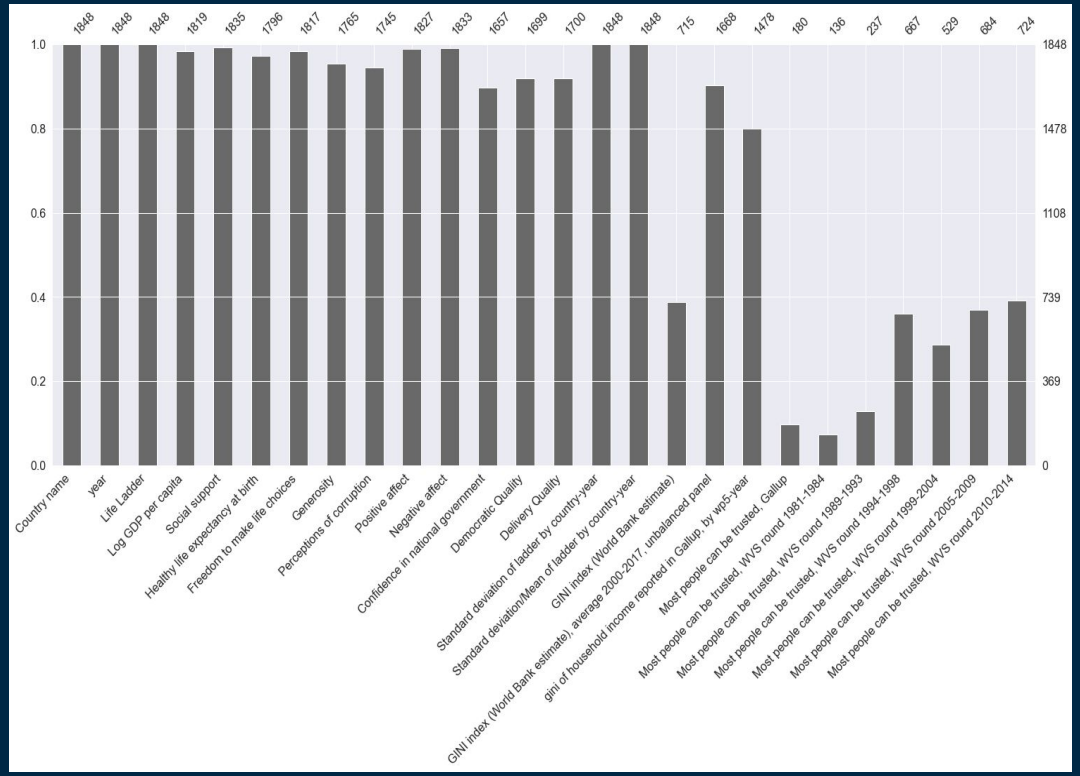
```
In [8]: data.isna().sum()

Out[8]: Country name                                                      0
        year                                                              0
        Life Ladder                                                       0
        Log GDP per capita                                               29
        Social support                                                   13
        Healthy life expectancy at birth                                 52
        Freedom to make life choices                                     31
        Generosity                                                       83
        Perceptions of corruption                                       103
        Positive affect                                                  21
        Negative affect                                                  15
        Confidence in national government                               191
        Democratic Quality                                              149
        Delivery Quality                                                148
        Standard deviation of ladder by country-year                      0
        Standard deviation/Mean of ladder by country-year                 0
        GINI index (World Bank estimate)                               1133
        GINI index (World Bank estimate), average 2000-2017, unbalanced panel  180
        gini of household income reported in Gallup, by wp5-year        370
        Most people can be trusted, Gallup                             1668
        Most people can be trusted, WVS round 1981-1984                1712
        Most people can be trusted, WVS round 1989-1993                1611
        Most people can be trusted, WVS round 1994-1998                1181
        Most people can be trusted, WVS round 1999-2004                1319
        Most people can be trusted, WVS round 2005-2009                1164
        Most people can be trusted, WVS round 2010-2014                1124
        dtype: int64
```

Methods:

Utilised data exploratory tools to find out the exact number of missing data

Visualisation of missing data



Methods:

Utilised missingno library to visualise the missing data on a bar graph

Actions:

Dropped Variables with more than 60% of data missing.

Variables with missing data to be dropped

- GINI index (World Bank estimate)
- GINI index (World Bank estimate), average 2000-2017, unbalanced panel'
- gini of household income reported in Gallup, by wp5-year
- Most people can be trusted, Gallup
- Most people can be trusted, WVS round 1981-1984
- Most people can be trusted, WVS round 1989-1993
- Most people can be trusted, WVS round 1994-1998
- Most people can be trusted, WVS round 1999-2004
- Most people can be trusted, WVS round 2005-2009
- Most people can be trusted, WVS round 2010-2014

Filling up remaining data

Methods:

Filled up rest of the data with **median**.

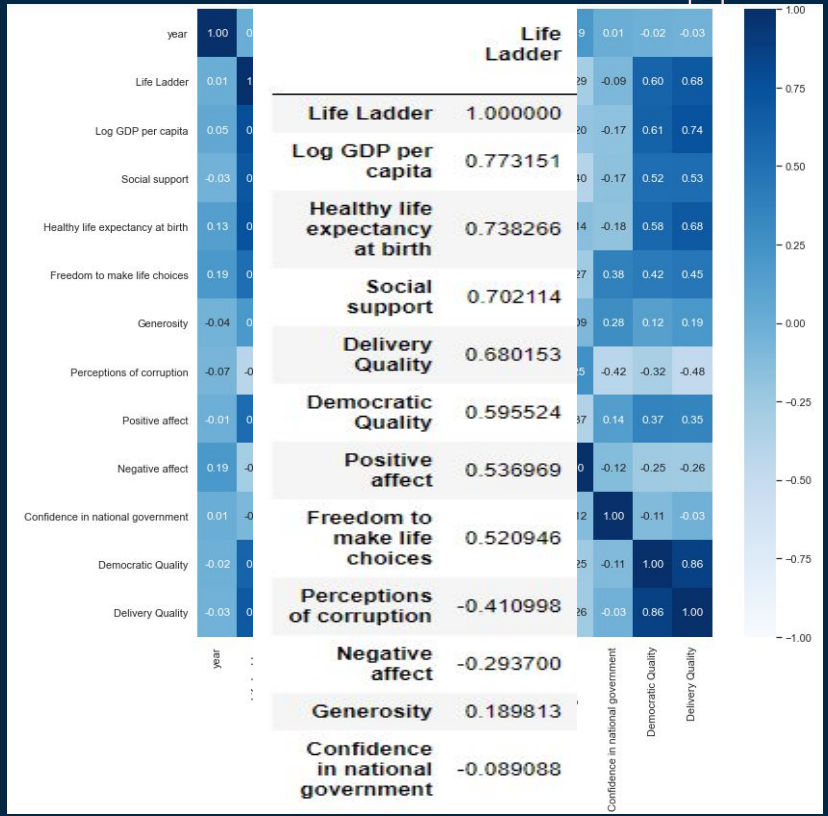
```
In [14]: data_dropped_1.isna().sum()

Out[14]: year                                      0
         Life Ladder                               0
         Log GDP per capita                        0
         Social support                            0
         Healthy life expectancy at birth          0
         Freedom to make life choices              0
         Generosity                                0
         Perceptions of corruption                 0
         Positive affect                           0
         Negative affect                           0
         Confidence in national government         0
         Democratic Quality                        0
         Delivery Quality                          0
         dtype: int64
```

Analysis of Variables

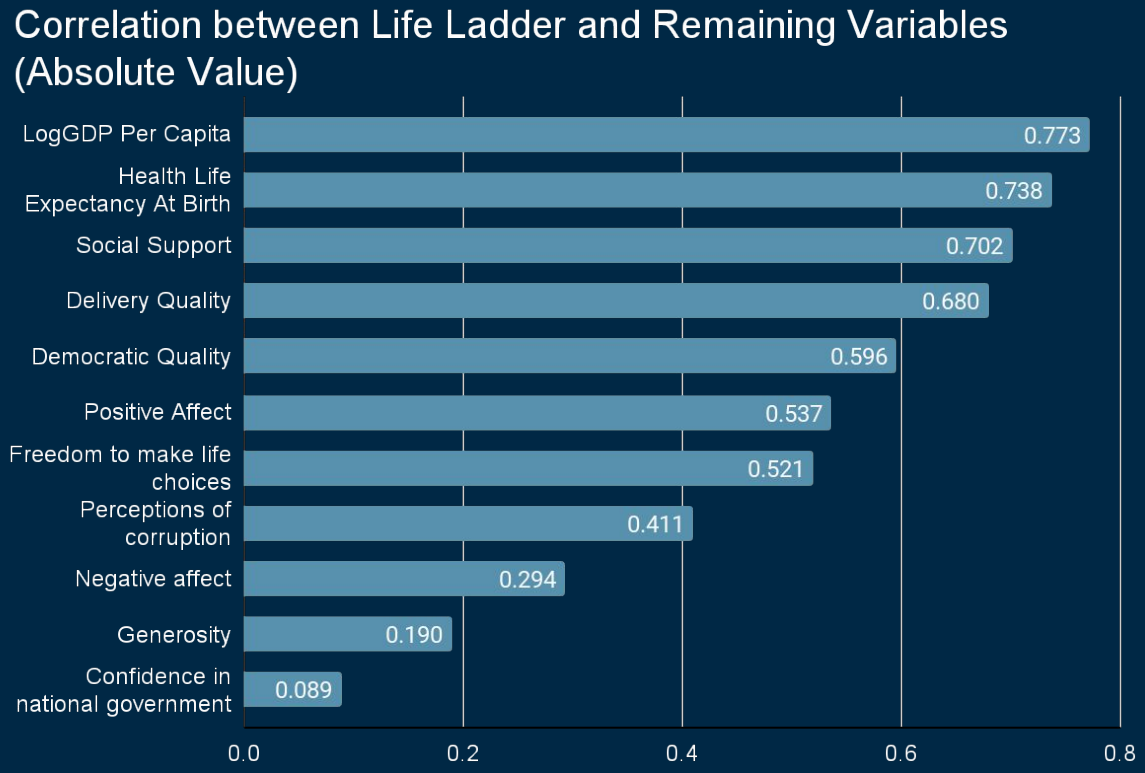
Methods:

Plotted the correlation matrix of the remaining variables as well as a table showing the correlation between Life Ladder and the remaining variables.



	Life Ladder
Life Ladder	1.000000
Log GDP per capita	0.773151
Healthy life expectancy at birth	0.738266
Social support	0.702114
Delivery Quality	0.680153
Democratic Quality	0.595524
Positive affect	0.536969
Freedom to make life choices	0.520946
Perceptions of corruption	-0.410998
Negative affect	-0.293700
Generosity	0.189813
Confidence in national government	-0.089088

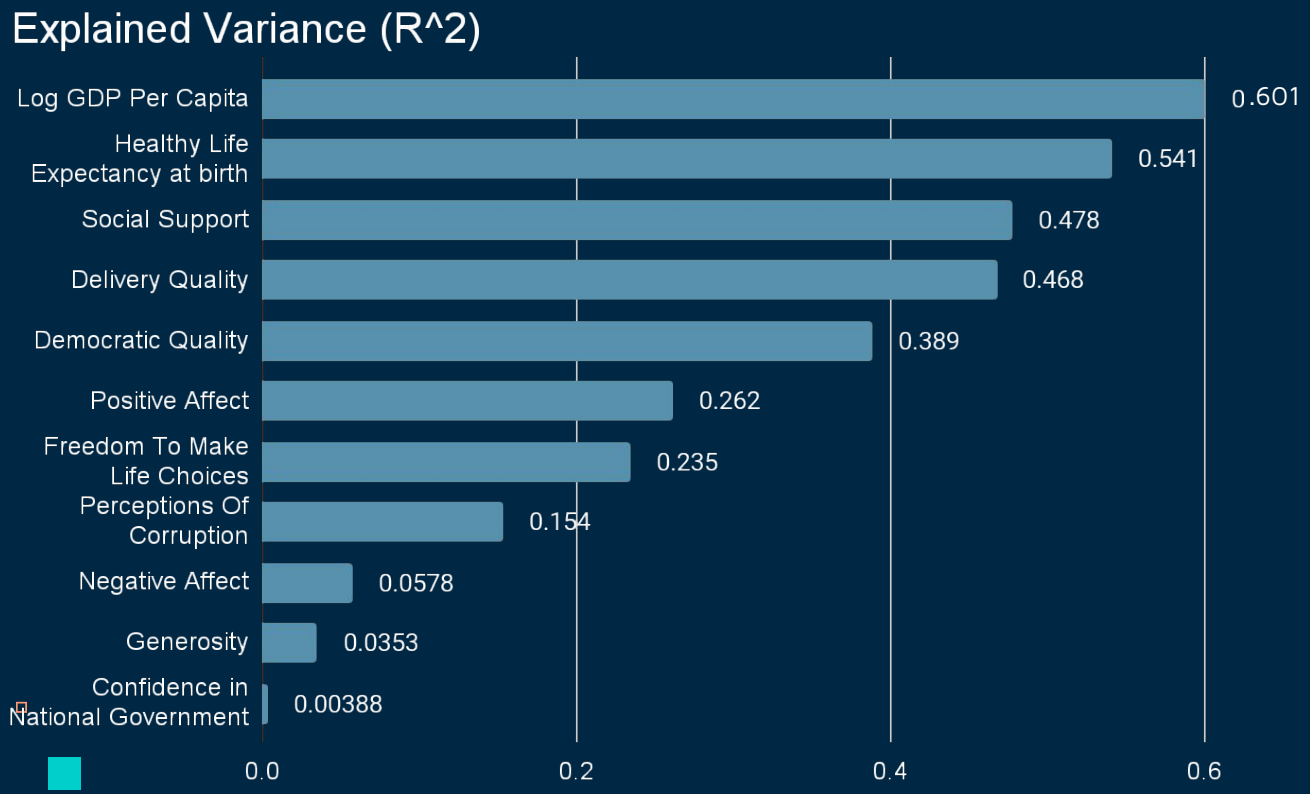
Analysis of Variables



Analysis:

- LogGDP Per capita had the strongest correlation with Life Ladder.
- Generosity, Confidence in national government, and negative affect had a weak correlation with Life Ladder and thus were least important in predicting "Life Ladder".

Uni-Variate Regression



Methods:

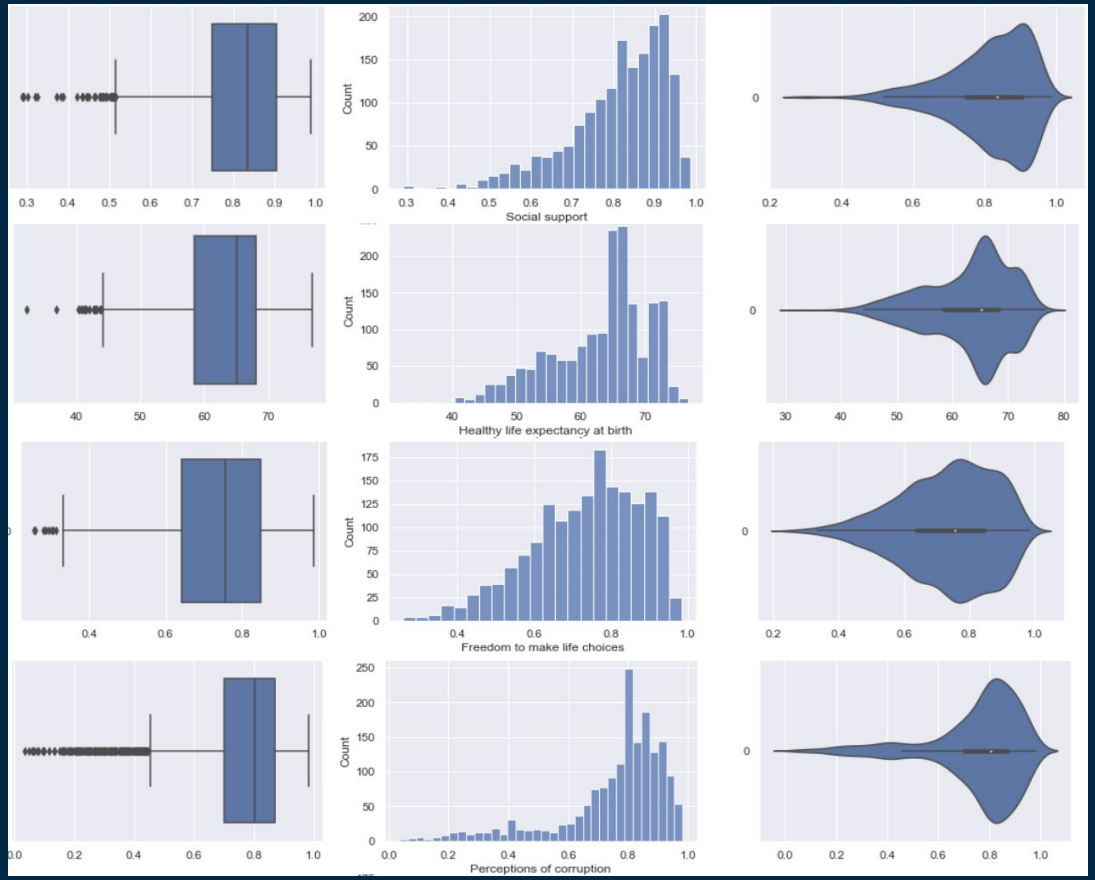
Utilise Uni-Variate regression to further evaluate the relationship between our predictor variables and response variable Life Ladder.

Analysis:

Confidence in national government, generosity, and negative affect had extremely low R^2 values.

Actions:

Decided to drop these 3 variables based on correlation and R^2 values.



Data Visualisation and Analysis

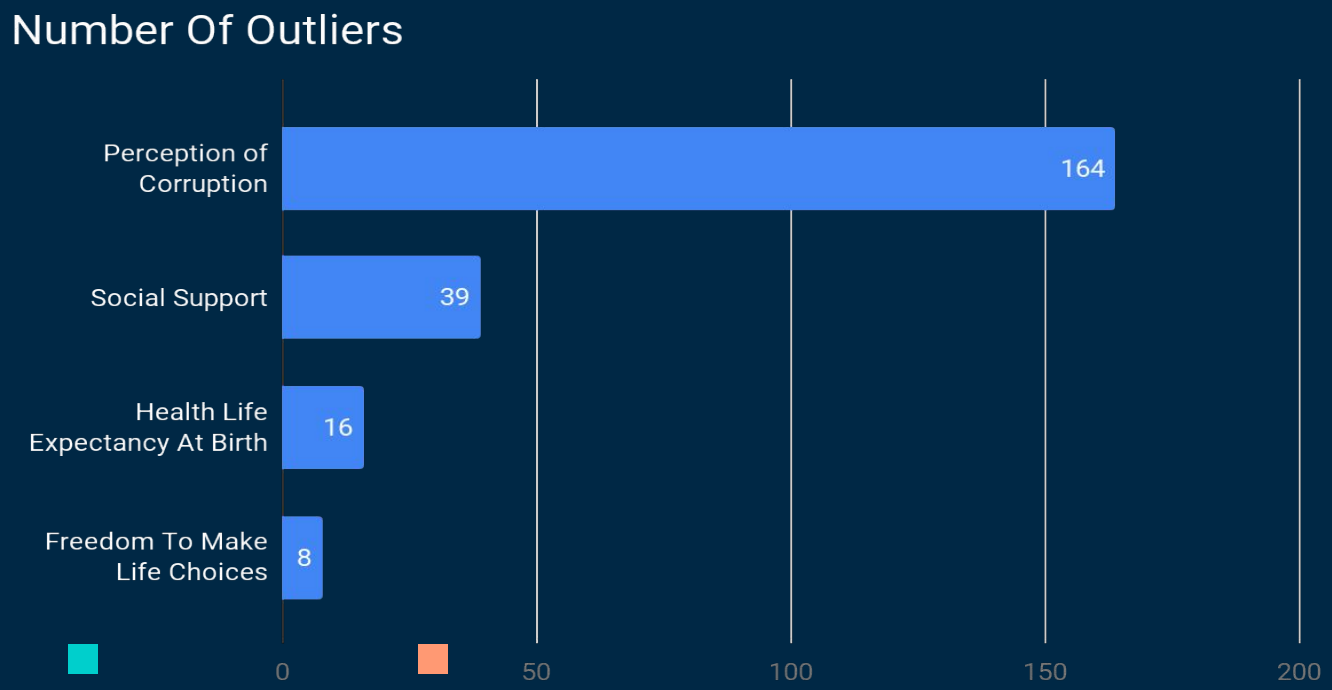
Methods:

Plotted box, bar, and violin plots to visualise the data.

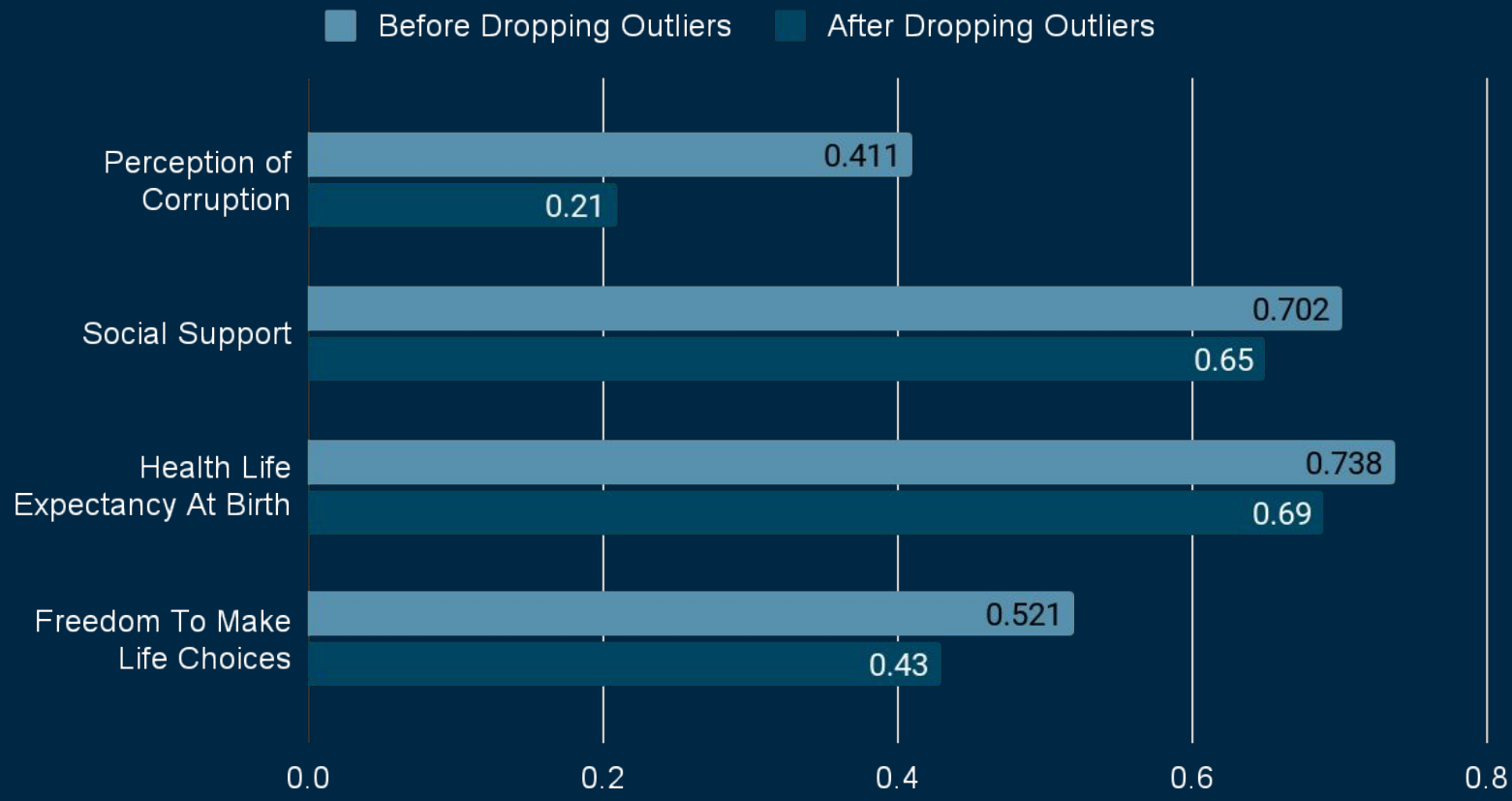
Variables with outliers :

- Social support
- Health life expectancy at birth
- Freedom to make life choices
- Perceptions of corruption

Analysis Of Outliers



Correlation (Absolute Value)



Variables Kept

- Log GDP Per Capita
- Health Life Expectancy At Birth
- Social Support
- Delivery Quality
- Democratic Quality
- Positive Affect
- Freedom to make life choices
- Perceptions of Corruption

To Keep	To Remove
-Log GDP Per Capita -Social Support -Health Life Expectancy At Birth -Freedom To Make Life Choices -Perceptions of Corruptions -Positive affect -Negative affect. -Democratic Quality -Delivery Quality	-GINI index (World Bank estimate) -GINI index (World Bank estimate), average 2000-2017, unbalanced panel' -gini of household income reported in Gallup, by wp5-year -Most people can be trusted, Gallup -Most people can be trusted, WVS round 1981-1984 -Most people can be trusted, WVS round 1989-1993 -Most people can be trusted, WVS round 1994-1998 -Most people can be trusted, WVS round 1999-2004 -Most people can be trusted, WVS round 2005-2009 -Most people can be trusted, WVS round 2010-2014

Machine Learning

Regression models

- Multivariate Linear regression
- Random Forest regression
- eXtreme Gradient boosting regression

Reasons for choosing our models

Regression models	How the model works
Multivariate Linear regression	Uses the least-squares approach
Random Forest regression	Uses bagging algorithm which is the process of creating and merging a collection of independent, parallel decision trees using different subsets of the training data
eXtreme Gradient boosting regression	Uses an iterative approach to combine a number of weak, sequential models to create one strong model by focusing on the mistakes in the prior iterations

Machine Learning

Cross-Validation Methods

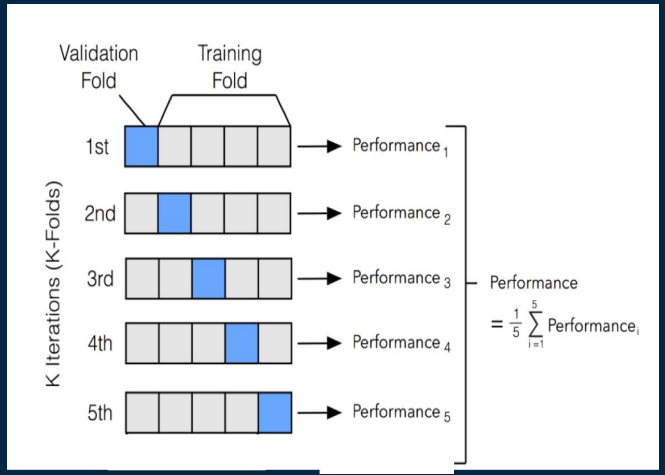
- Hold-out CV
- K-fold CV

Reason for using k-fold CV

1. K-fold CV results in a less biased model compared to holdout CV since the model has the opportunity to train on multiple train-test splits. Every observation has the chance of appearing in both train and test sets.

2. Hold out CV ,on the other hand, is dependent on just one random train-test split. That makes the hold-out method score dependent on how the data is split into train and test sets.

Example of how k-fold CV works



Reason for using k-fold CV

Using holdout CV method for linear regression

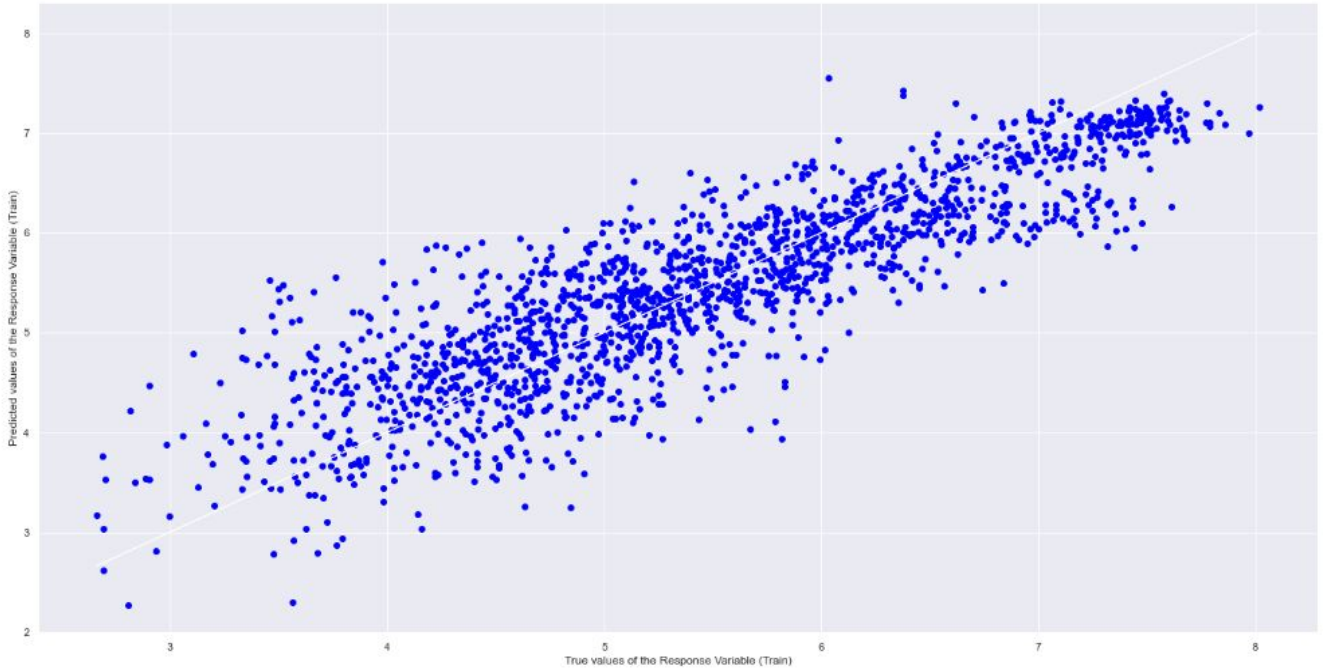
```
Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.721488766669981
Mean Squared Error (MSE)        : 0.3255203183216921
```

Using k-fold CV method for linear regression

```
Goodness of Fit of Model
Scores of the model is  : [0.7794031  0.75312542 0.77348087 0.72775014 0.7324435 ]
Explained Variance (R^2)        : 0.7532406048969661
Mean Squared Error (MSE)        : 0.30841598431115264
```

By using k-fold CV, it gives us more accurate results as compared to a normal holdout CV method. Hence we decide to use k-fold CV to measure the accuracy of our all our models.

Multivariate Linear Regression

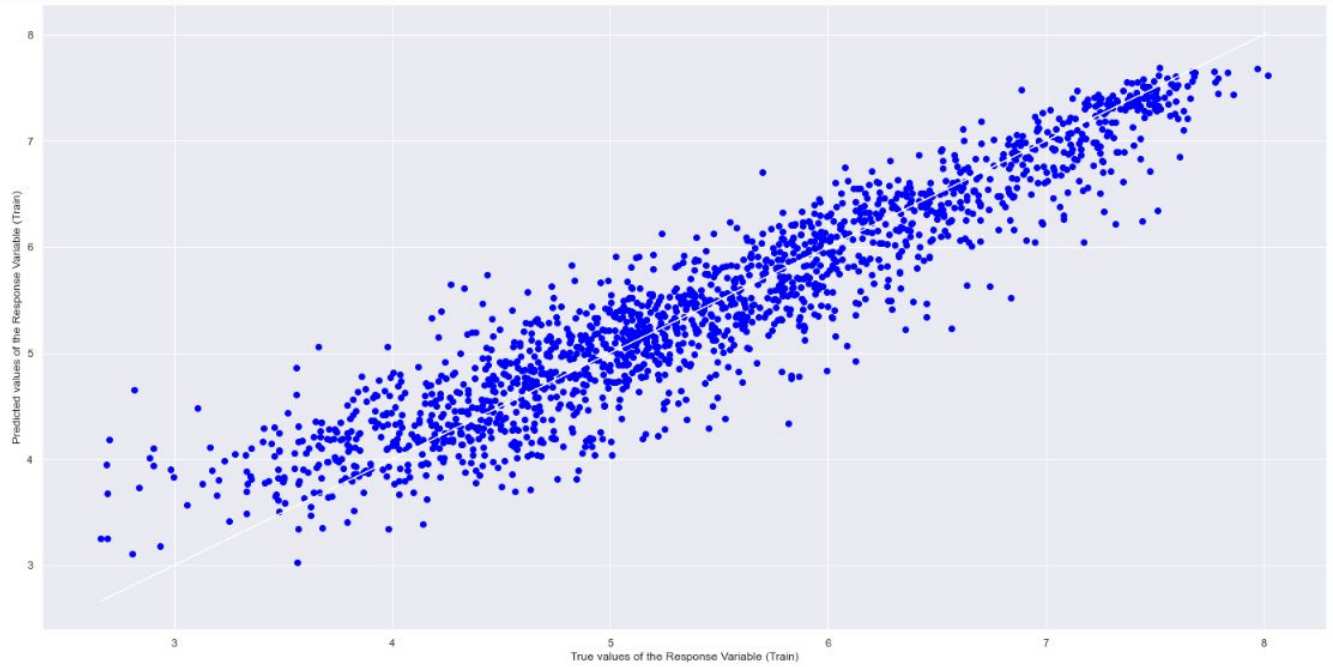



```
Goodness of Fit of Model
Scores of the model is  : [0.77928503 0.75284544 0.77323218 0.72792989 0.73263695]
Explained Variance (R^2)        : 0.7531858970298412
Mean Squared Error (MSE)        : 0.3084783865914527
```

Explained Variance:

0.753

Random Forest Regression

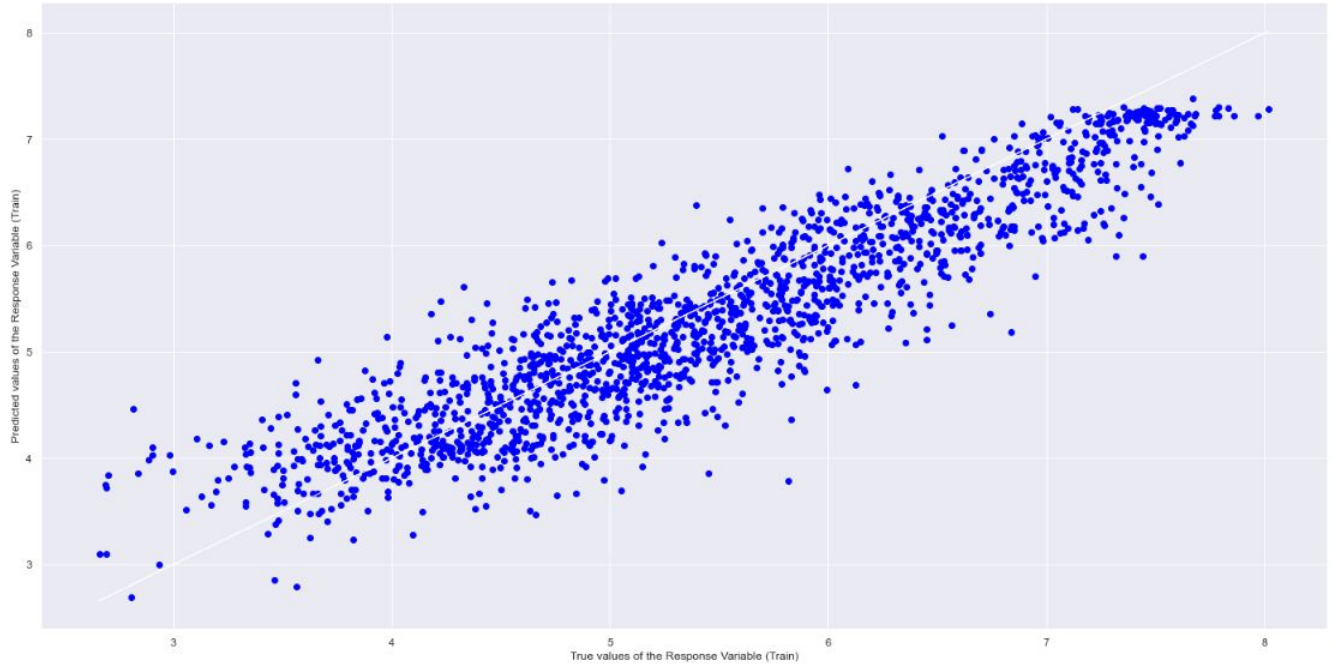



```
Goodness of Fit of Model
Scores of the model is  : [0.88659449 0.89074927 0.88040048 0.84298767 0.8803495 ]
Explained Variance (R^2)        : 0.8762162820214539
Mean Squared Error (MSE)        : 0.15596938105256694
```

Explained Variance:

0.876

eXtreme Gradient Boosting Regression

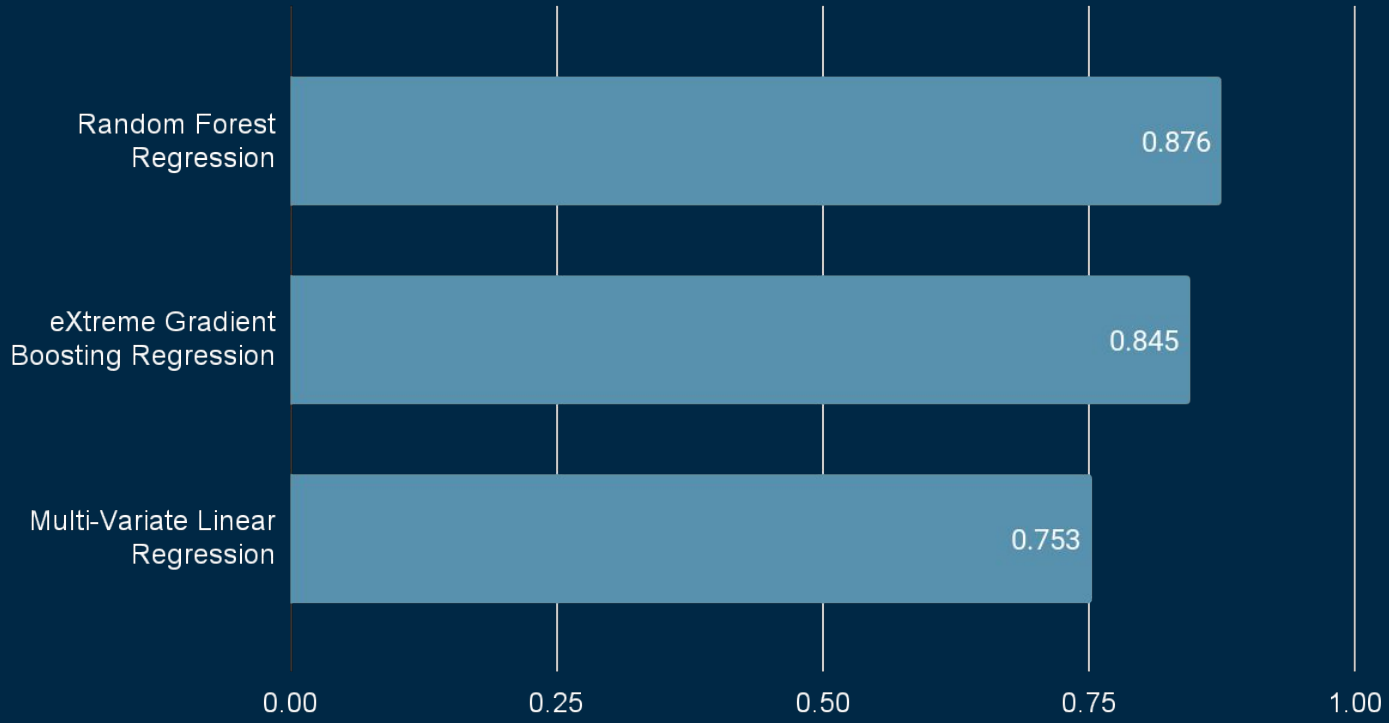



```
Goodness of Fit of Model
Scores of the model is  : [0.85803256 0.85442201 0.85630996 0.80772294 0.84900812]
Explained Variance (R^2)        : 0.8450991187715999
Mean Squared Error (MSE)        : 0.2162859744754668
```

Explained Variance:

0.845

Explained Variance



Comparing the explained variance values obtained by the different models

Random Forest Regression is the best model

Using RFR to predict 2019 data

```
count =0
error_total =0
for i in data_predict_list:
    predicted = rand_forest.predict([i])[0]
    actual =list_2019_life_ladder[count]
    error = (abs(predicted - actual))/actual*100
    count += 1
    error_total += error

print(error_total/count)
accuracy = 100-error_total/count
print(accuracy)
```

```
9.511980083832997
90.488019916167
```

Using our model to predict 2019 data, we are able to get an accuracy of **90.5%**

Conclusion and Recommendations

The most important variables in predicting life ladder

Looking at the correlation and explained variance of the variables in our dataset, these are the three most important variables in predicting life ladder as their correlation and explained variance values were the highest.

1. Log gdp per capita
2. Social Support
3. Health life Expectancy at birth

Therefore, Singapore should focus on improving these 3 components to bring out the best increase in health ladder.

Conclusion and Recommendations

The best model in predicting life ladder

We concluded that random forest regressor is the best model to predict life ladder across the three models that we used.

Countries that wish to predict their life ladder should use random forest regressor.

What we learnt

1. Application of new machine learning models such as Random Forest Regression and eXtreme Gradient boosting regression
2. New visualisation techniques using the missingno library
3. Application of k-fold CV

THANK YOU