DS AI mini project

Happiness World Report

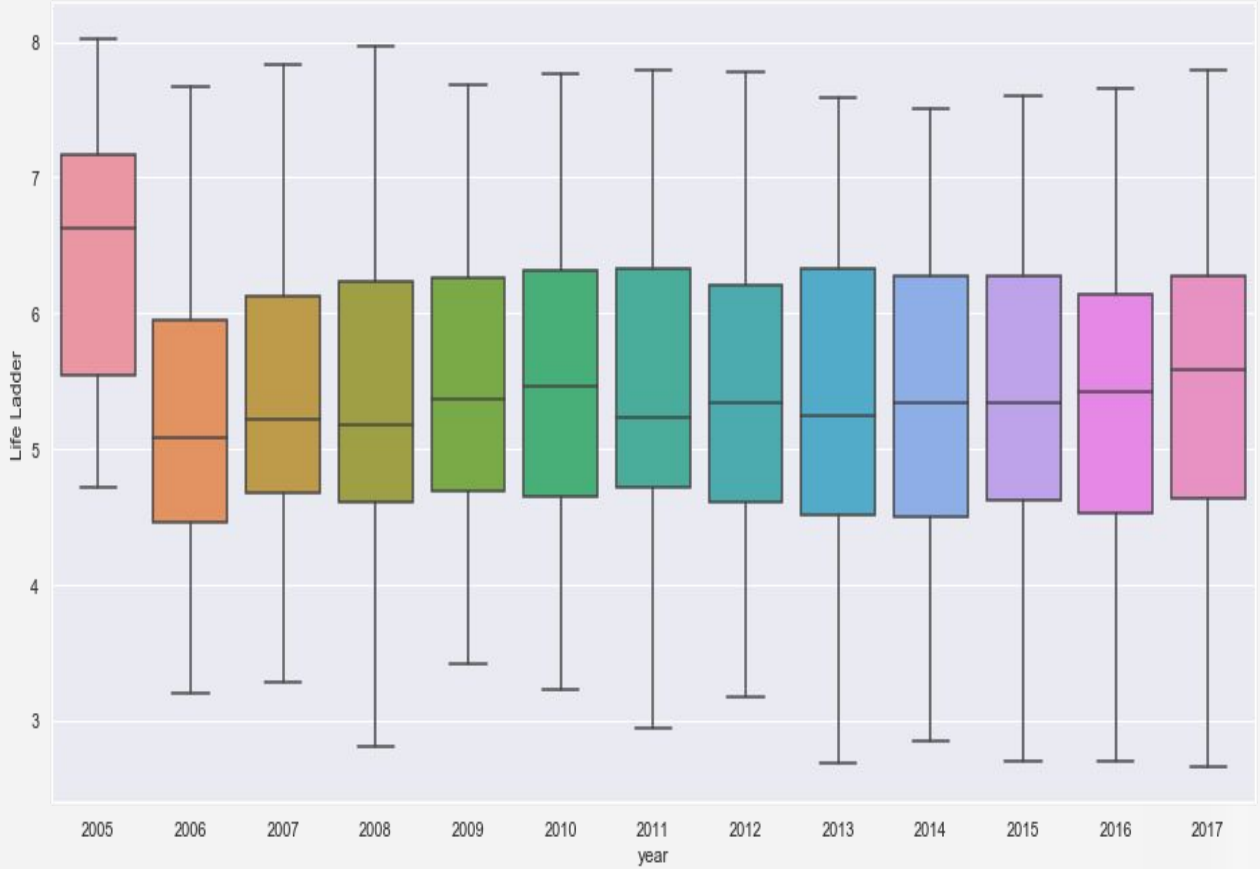
Damien Goh, Wu Wenshan, Truong Cong Cuong

Objectives

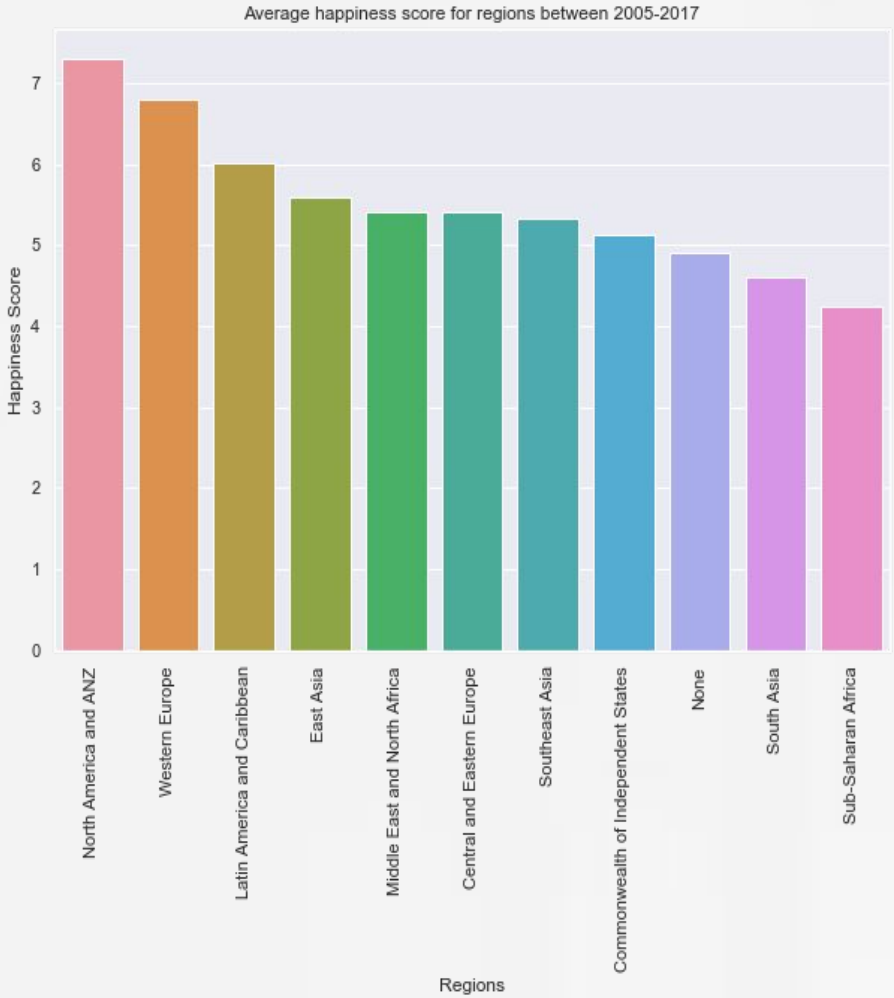
1. Choose a regression model to best learn the distribution of data & prediction of life ladder based on new data published this year
2. Clustering of data points and analysis on what the clusters mean

Data exploration

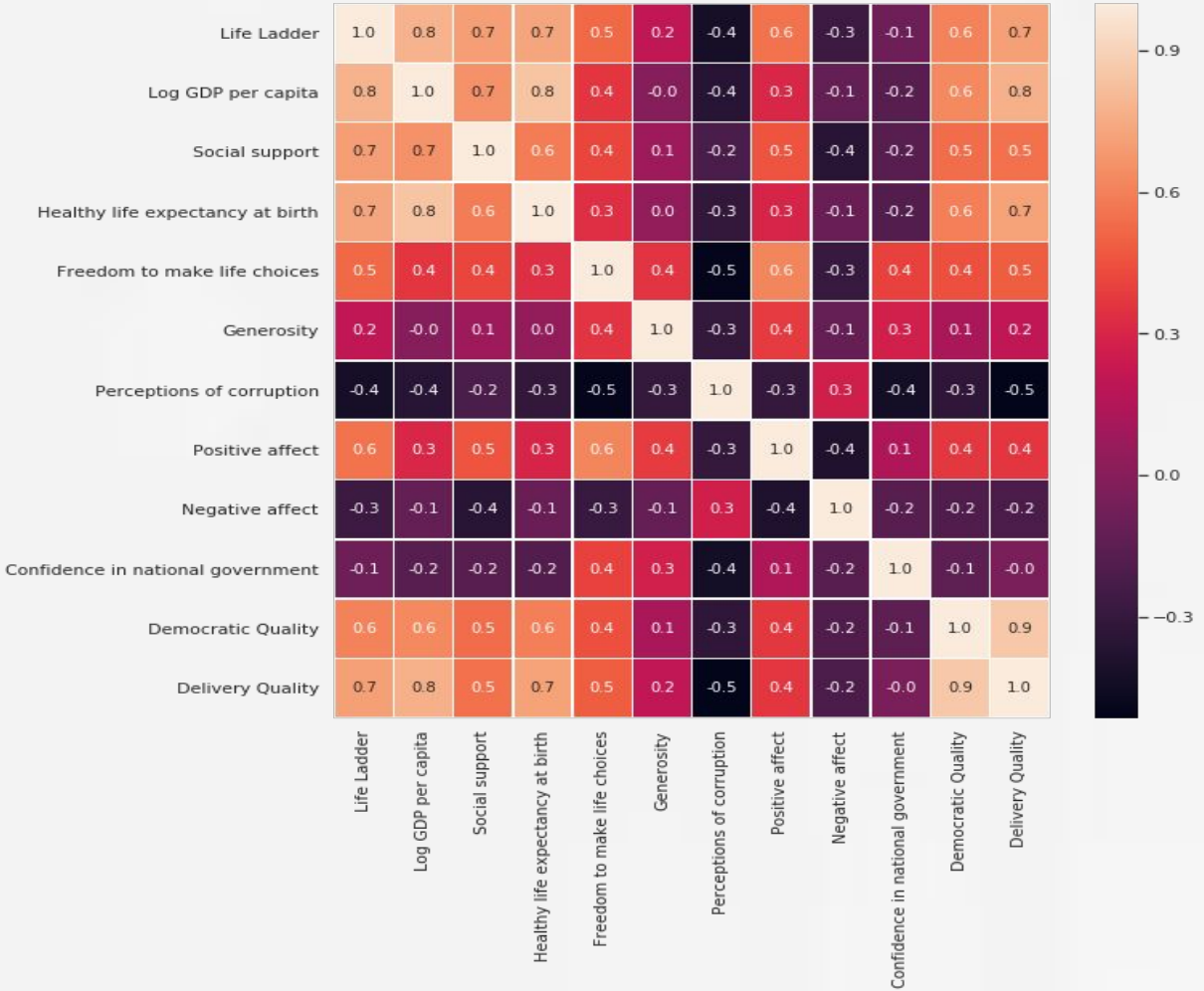
Distribution of Life Ladder for all countries



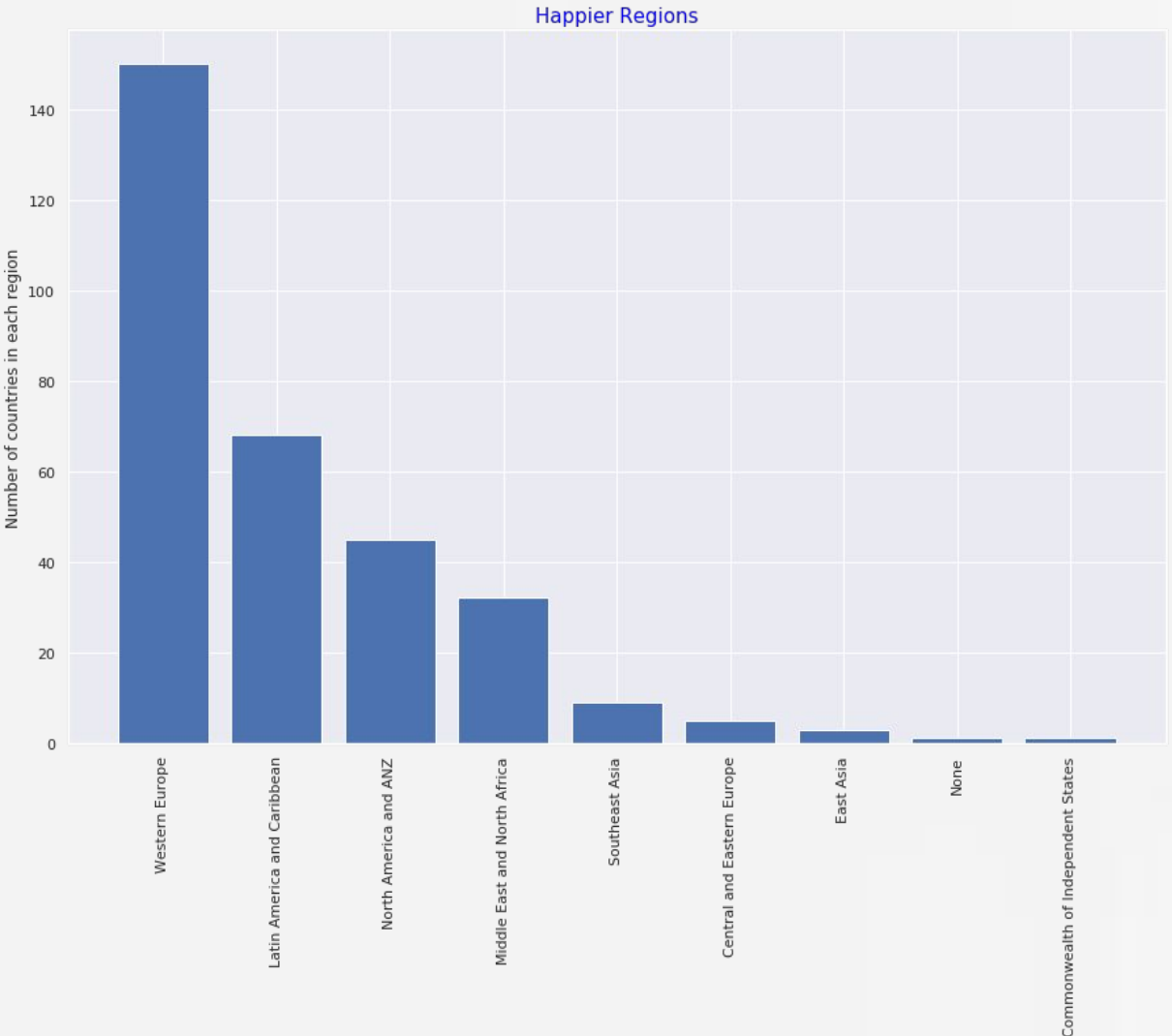
Average life ladder for different regions for the given years

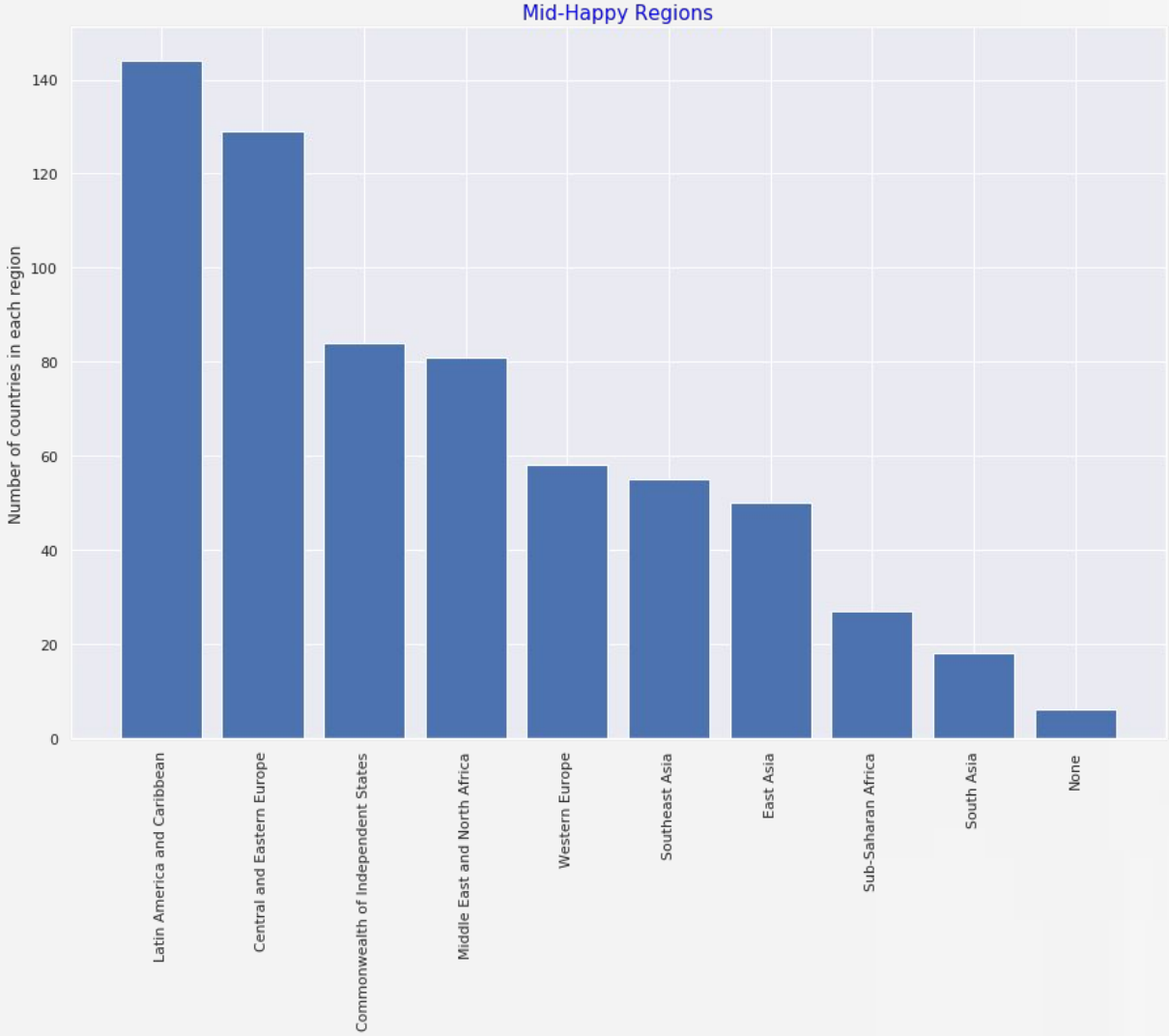


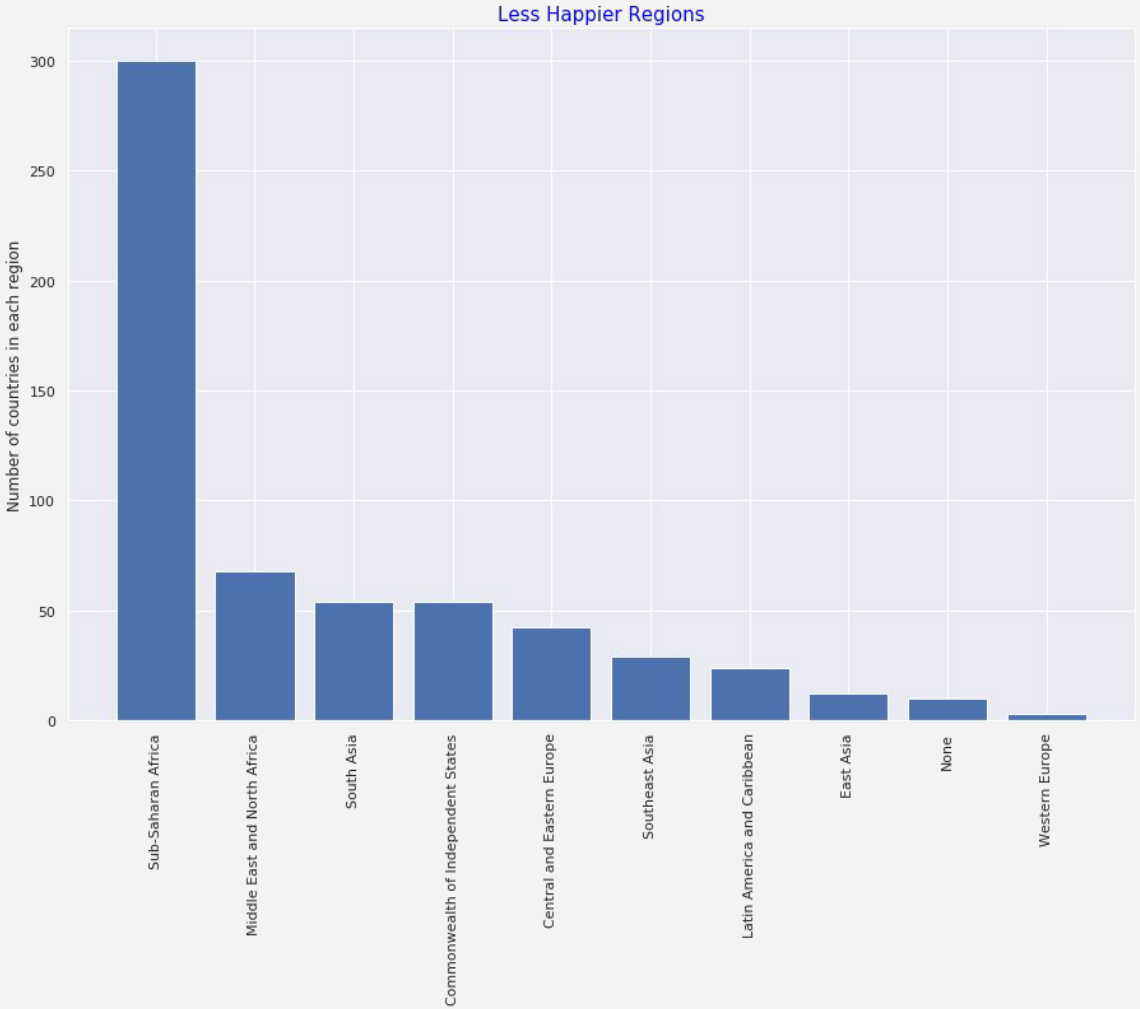
Relationship of variables compared to one another



Important variables against Life Ladder


Happier Regions
Number of countries in each region
0
20
40
60
80
100
120
140
Western Europe
Latin America and Caribbean
North America and ANZ
Middle East and North Africa
Southeast Asia
Central and Eastern Europe
East Asia
None
Commonwealth of Independent States


Mid-Happy Regions
Number of countries in each region
0
20
40
60
80
100
120
140
Latin America and Caribbean
Central and Eastern Europe
Commonwealth of Independent States
Middle East and North Africa
Western Europe
Southeast Asia
East Asia
Sub-Saharan Africa
South Asia
None


Less Happier Regions
Number of countries in each region
300
250
200
150
100
50
0
Sub-Saharan Africa
Middle East and North Africa
South Asia
Commonwealth of Independent States
Central and Eastern Europe
Southeast Asia
Latin America and Caribbean
East Asia
None
Western Europe

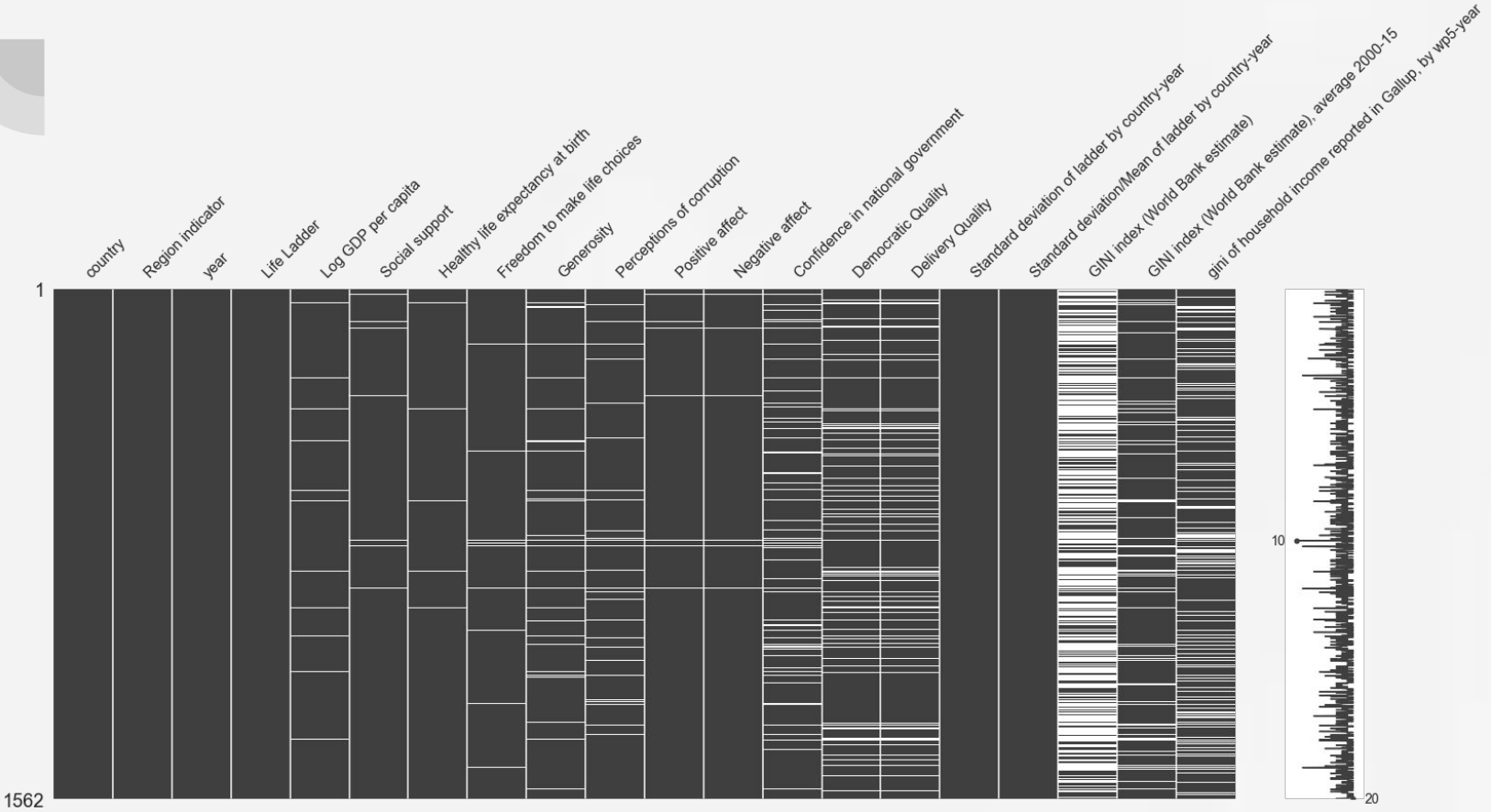
Objective 1: Regression problem

Choose a regression model to best learn the distribution of data & prediction of life ladder based on new data published this year

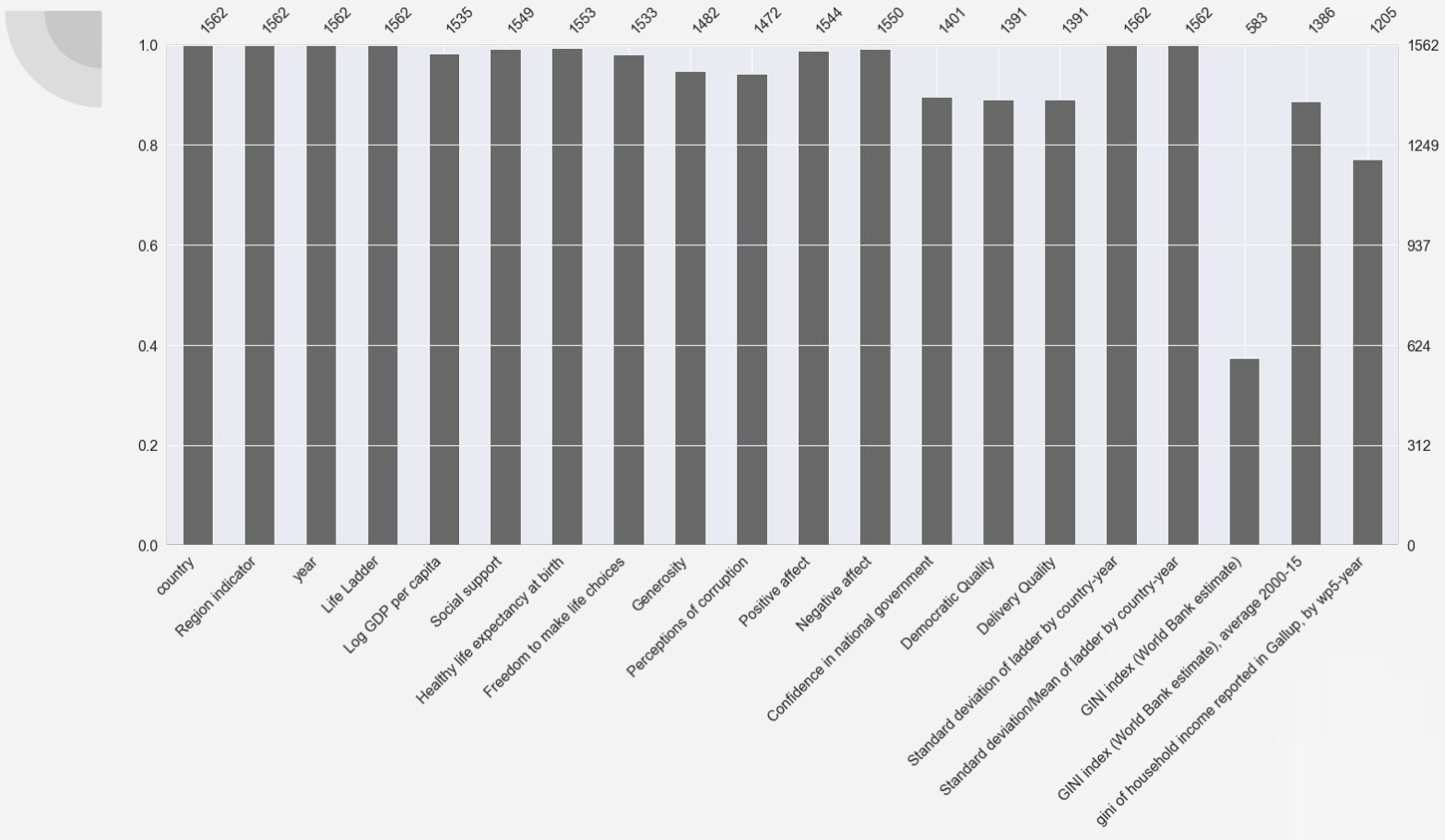
Step 1: Data preparation

- Fill in missing values in the dataset by interpolation
- Deletion of some columns

Visualise the locations where the values are missing

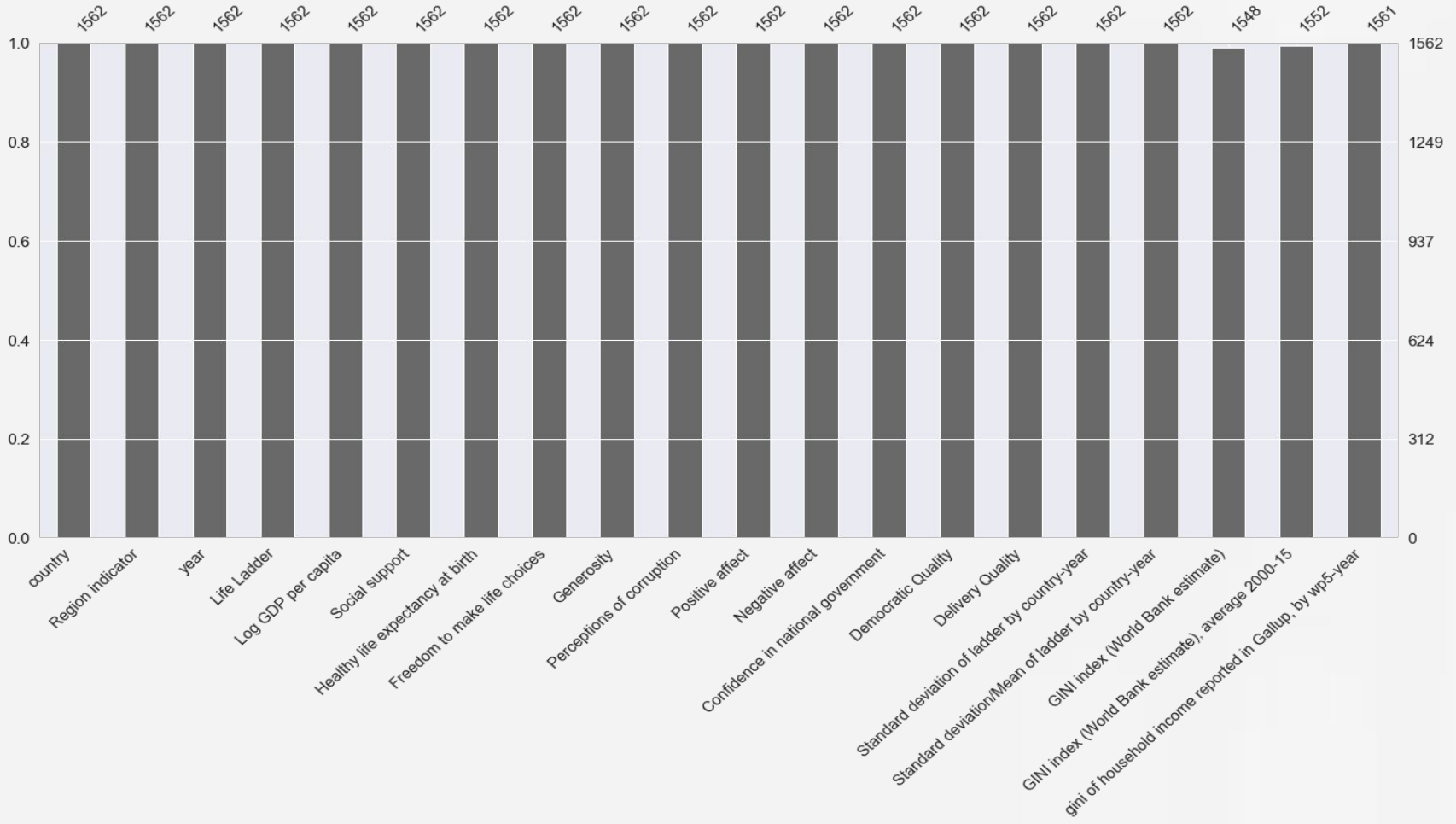


Counting the data points present for each variable in the dataset

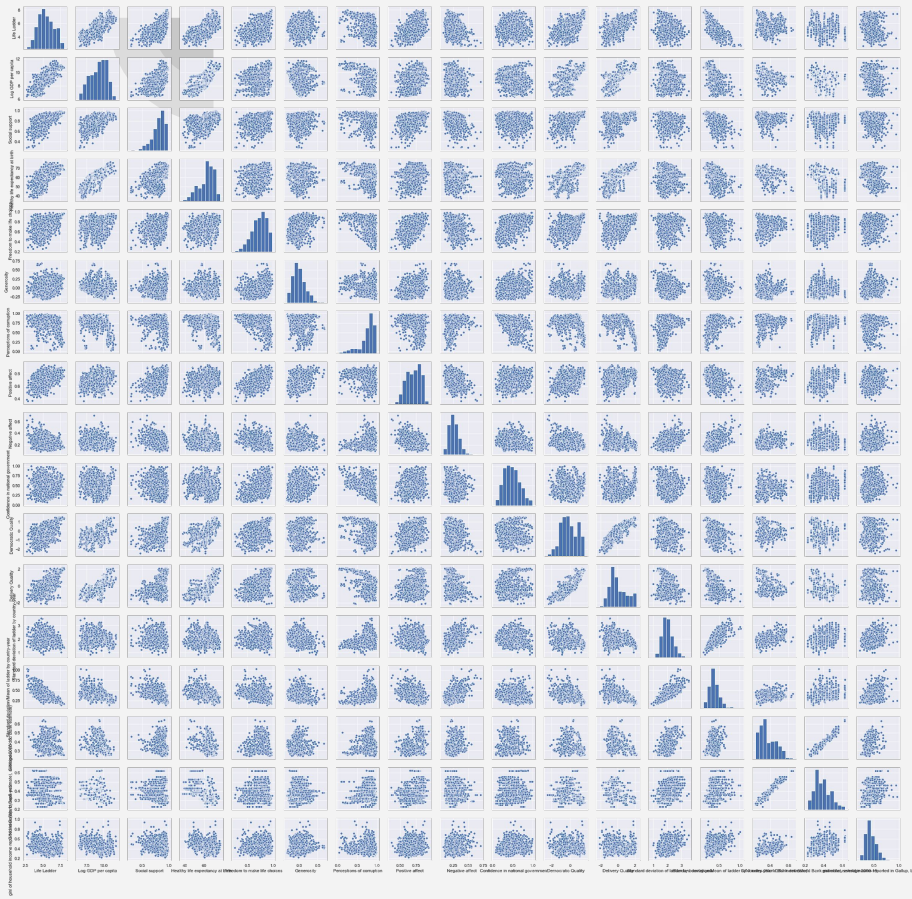


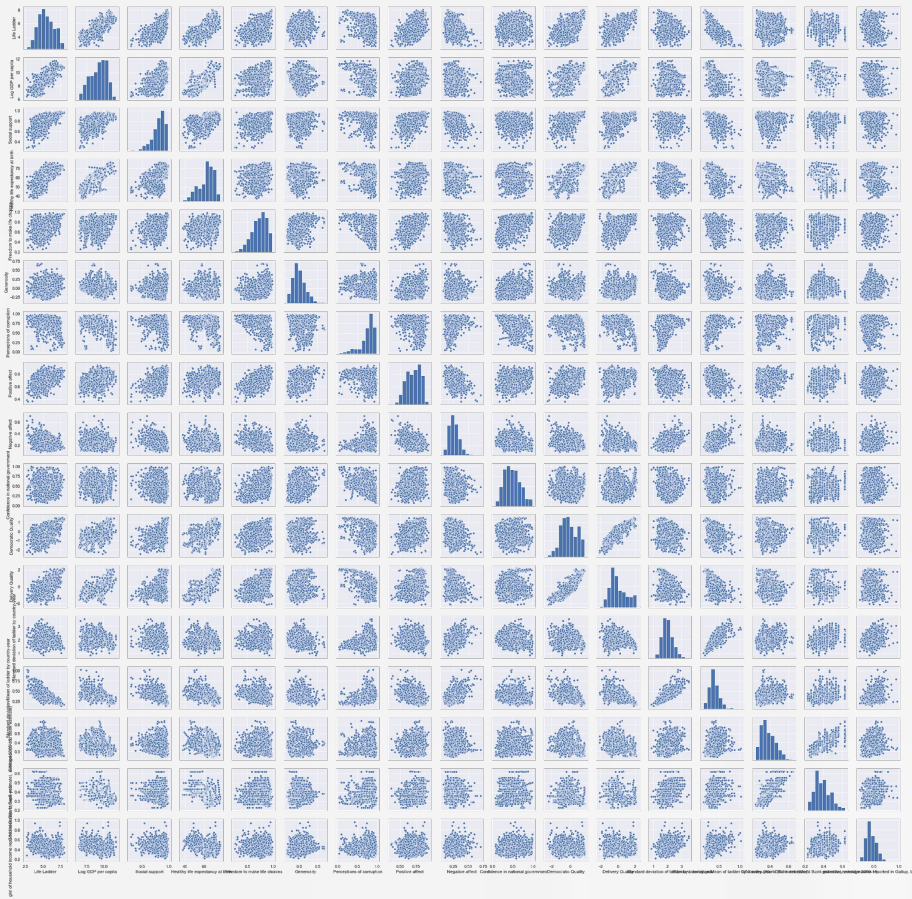
Counting the data points present for each variable in the dataset

Counting the data points present for each variable in the dataset

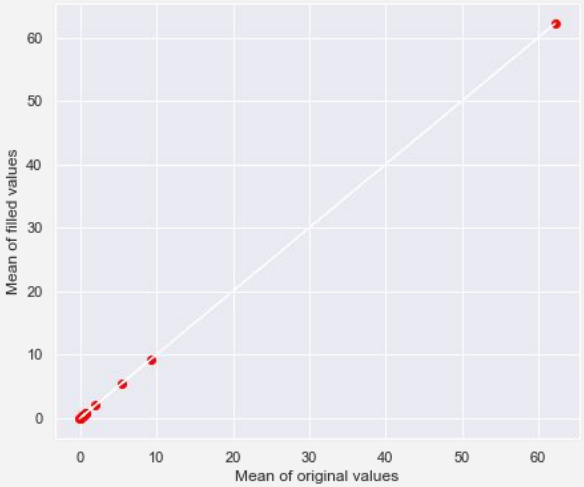


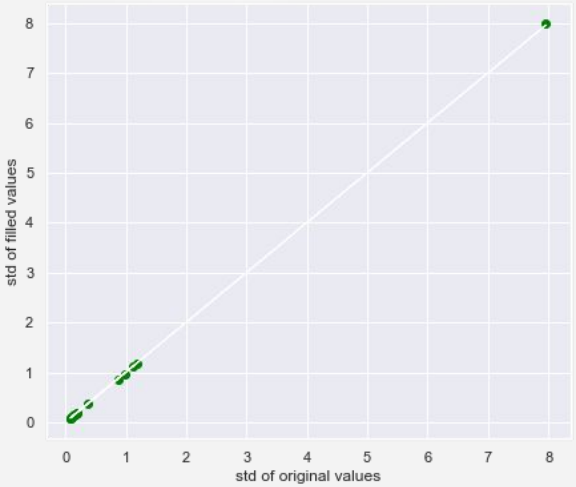
Distribution of original data VS distribution of filled in data

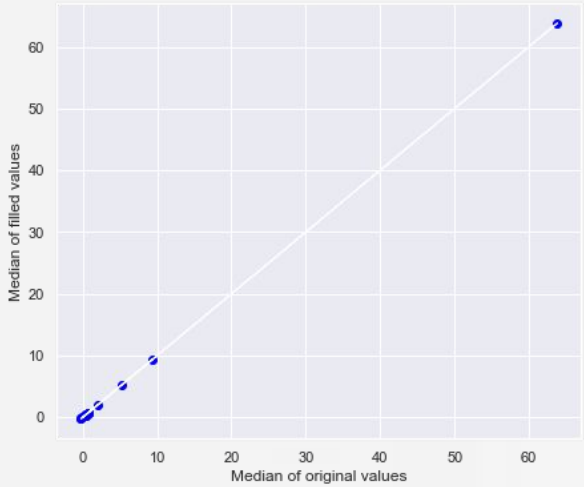




Distribution of original data VS distribution of filled in data







Step 2: Build models for prediction the life ladder

- Linear regression model
- Random forest regression model
- Multi-layer perceptron regression model

Step 3: Cross-validation is used to evaluate all three models and accuracies of prediction on test dataset is calculated using R^2

Step 4: Prediction of life ladder in 2018 based on new data

Summary of results

Cross-validation scores of models:
Linear regression model: 0.747
Random forest regression model: 0.759
Multi-perceptron regression model: 0.857

Accuracy scores of models on test data:
Linear regression model: 0.765
Random forest regression model: 0.888
Multi-perceptron regression model: 0.842

Accuracy scores on new data:
Linear regression model: 0.714
Random forest regression model: 0.774
Multi-perceptron regression model: 0.744

Conclusion:

Based on the scores obtained from cross validation and those from the prediction for the test datasets, Multi-Layer Perceptron regressor and Random Forest regressor have higher scores and accuracies than the linear regression model, and hence would be the better models for Life Ladder predictions in the future.

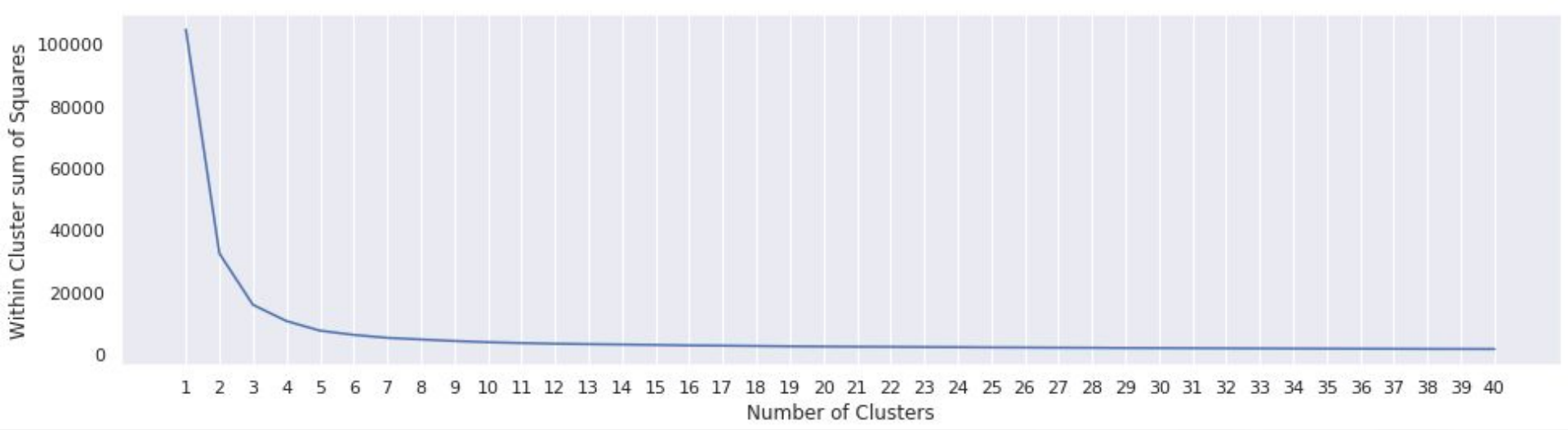
Objective 2: Clustering problem

Clustering

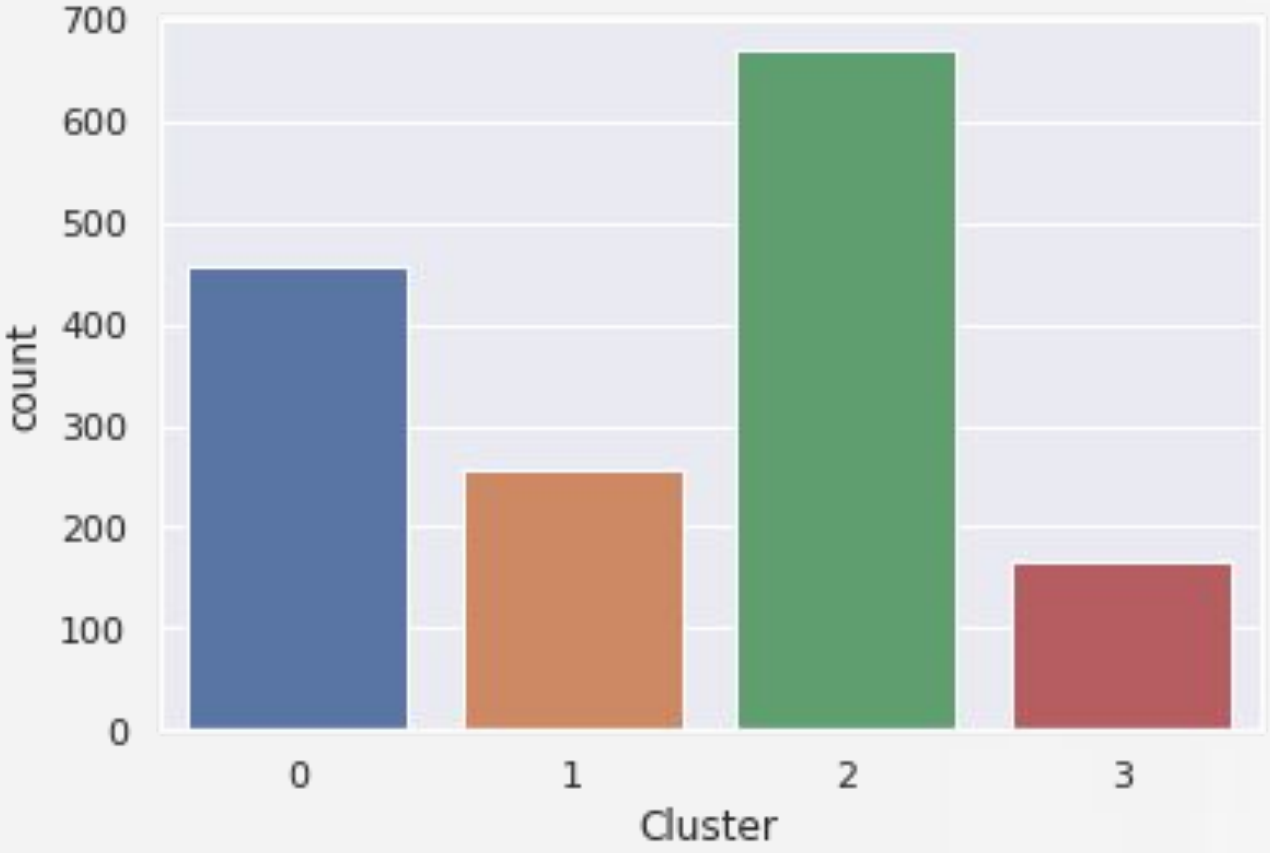
- K-Means
- BIRCH Clustering

Variables in clustering: Life Ladder, Log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, Democratic Quality, Positive affect, Generosity

K-Means Clustering



Within Cluster of Squares


700
600
500
400
300
200
100
0
count
0
1
2
3
Cluster

Clusters in K-Means

Cluster 1 and 3: Sub-Saharan Africa - the worst performer in most aspects

Cluster 0: Mostly Europe, some East Asia and American countries - the best performers in most aspects

Cluster 2: Latin America, and Caribbean, Commonwealth of Independent States, Middle East and North Africa, Central and Eastern Europe

BIRCH Clustering

Clusters in BIRCH Clustering

Cluster 0 and Cluster 3: Sub-Saharan Africa - the worst performer in most aspects

Cluster 1: Latin America, and Caribbean, Commonwealth of Independent States, Middle East and North Africa, Central and Eastern Europe

Cluster 2: Mostly Europe, some East Asia and American countries - the best performers in most aspects

Conclusion

Two methods yield similar results. While Sub-Saharan Africa countries perform almost worst in many factors, Western Europe countries perform best.

Work allocation

Damien Goh: Exploratory data analysis

Wu Wenshan: Data preparation and building of regression models

Truong Cong Cuong: Clustering

Notes

All of the work, except Prediction of 2018 Life Ladder, is based on the data extracted from the file: WHR2018Chapter2OnlineData.xls