World Happiness Analysis

Overview

1. Dataset
2. Exploratory Data Analysis (EDA)
3. Machine Learning based prediction models
4. Clustering
5. Outcomes
6. Work Distribution

Objectives:

- The most efficient machine learning algorithm for predicting Happiness

- The most prominent factor affecting happiness in:
  - The World
  - Different sub-regions

Dataset

- World Happiness Report 2018
- Data from year 2015-2017

Exploratory Data Analysis (EDA)

Data Preparation

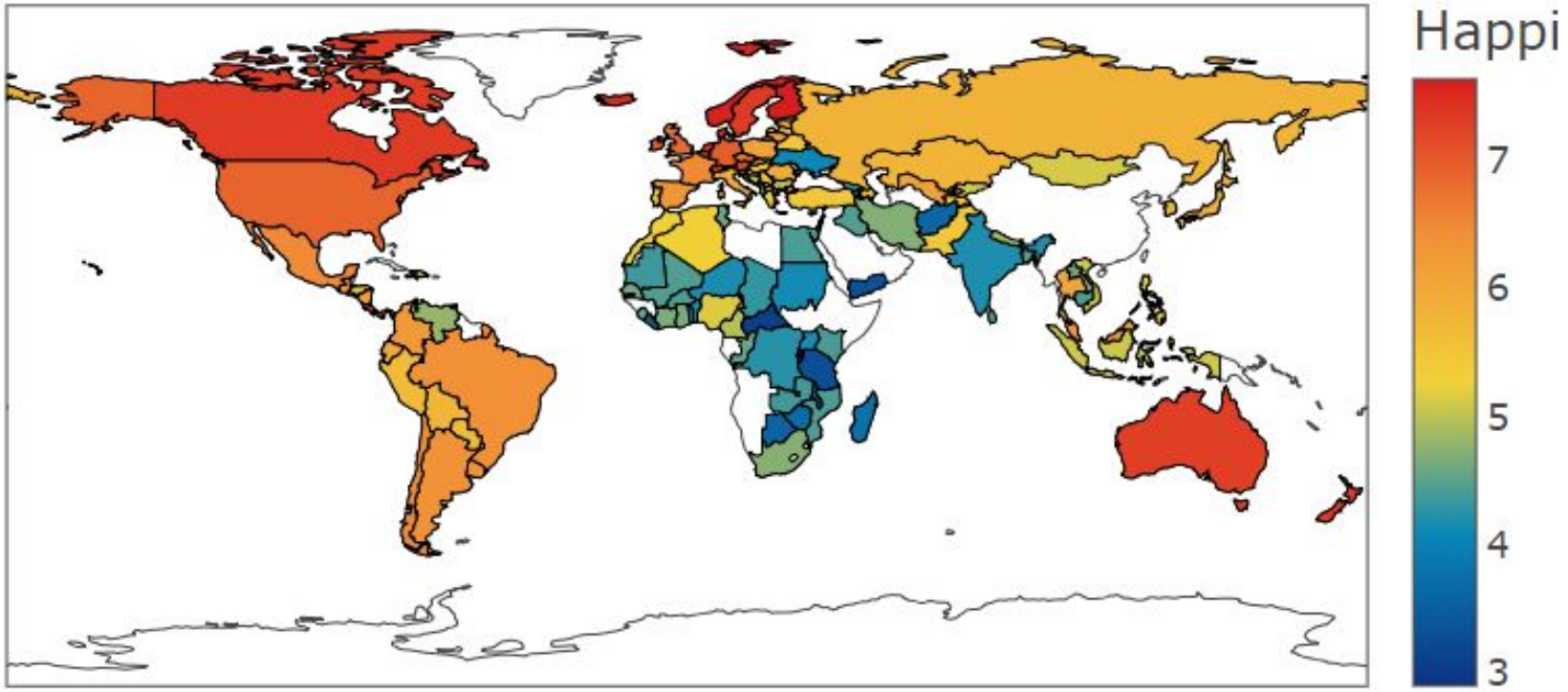
1. Remove NULL values
2. Modify the naming of the variable
3. Extract variables

Statistical Summary

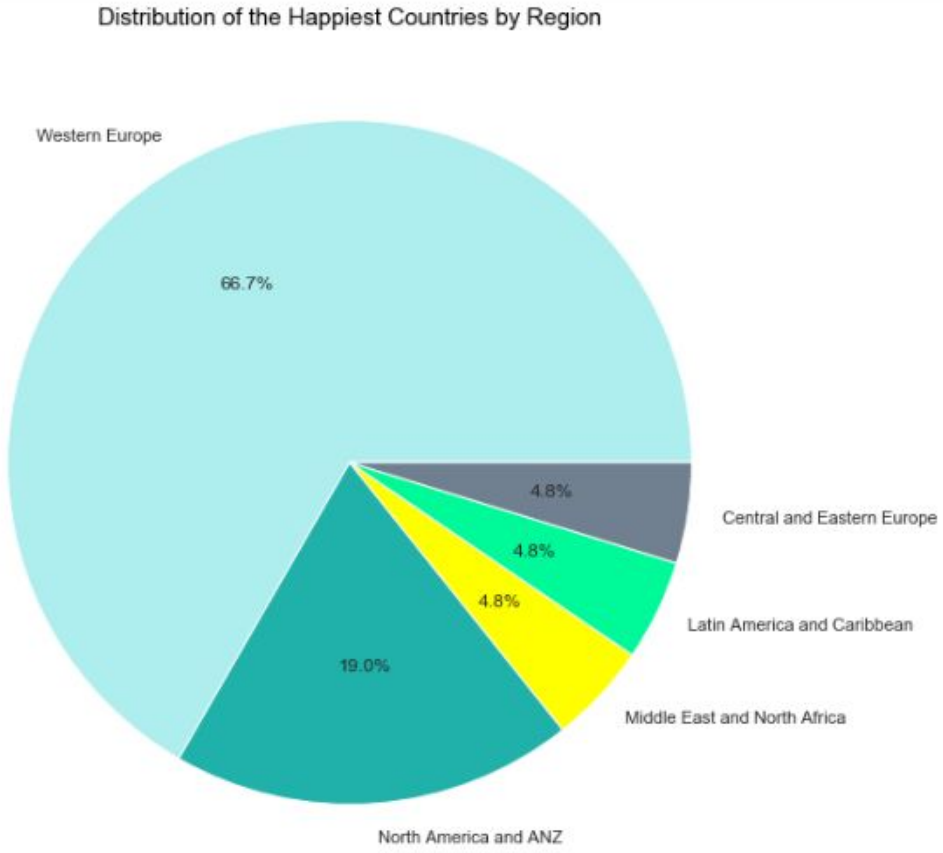
	Happiness_score	Log_of_GDP_per_person	Social_support	Healthy_life_expectancy	Freedom_to_make_life_choices	Generosity	Per
count	129.000000	129.000000	129.000000	129.000000	129.000000	129.000000	
mean	5.410946	9.222133	0.806110	63.151731	0.755963	0.296275	
std	1.145166	1.212248	0.125217	7.818432	0.136250	0.174722	
min	2.905000	6.473706	0.305565	43.994526	0.387829	0.033717	
25%	4.471000	8.297338	0.733782	57.630005	0.671541	0.167456	
50%	5.358000	9.478037	0.819712	65.049049	0.777240	0.251631	
75%	6.260000	10.210842	0.908002	68.871552	0.860594	0.381380	
max	7.632000	11.458786	0.977497	75.720833	0.983573	0.775616	

Average happiness score : **5.41**
Happiness Range: **4.21 - 6.51**

World Happiness Score Distribution



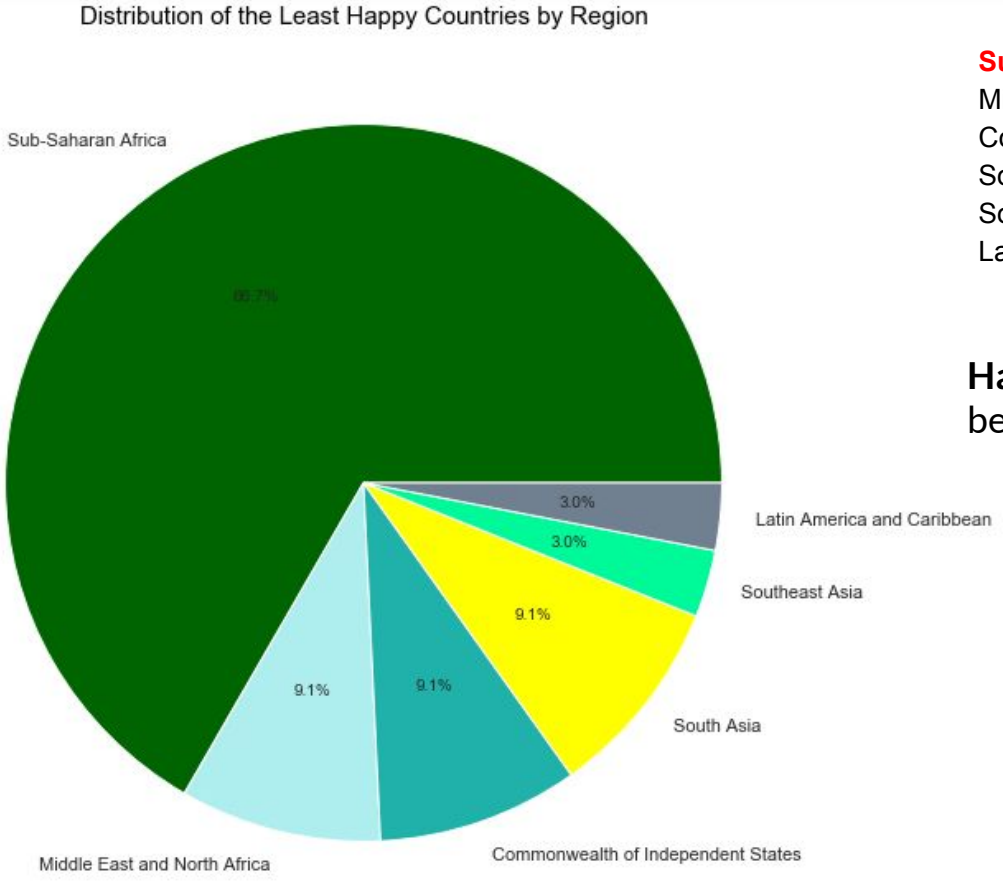
Most Happy by Regions



Western Europe	**14**
North America and ANZ	4
Middle East and North Africa	1
Latin America and Caribbean	1
Central and Eastern Europe	1

Happiness Score: 6.5 above

Least Happy by Regions



Sub-Saharan Africa	**22**
Middle East and North Africa	3
Commonwealth of Independent States	3
South Asia	3
Southeast Asia	1
Latin America and Caribbean	1

Happiness Score:
below 4.5

Happiness by countries

	country	Happiness_score
44	Finland	7.632
105	Norway	7.594
37	Denmark	7.555
57	Iceland	7.495
133	Switzerland	7.487
99	Netherlands	7.441
25	Canada	7.328
100	New Zealand	7.324
132	Sweden	7.314
6	Australia	7.272

Top 10 Most Happy

	country	Happiness_score
22	Burundi	2.905
26	Central African Republic	3.083
137	Tanzania	3.303
153	Yemen	3.355
117	Rwanda	3.408
78	Liberia	3.495
53	Haiti	3.582
84	Malawi	3.587
18	Botswana	3.590
0	Afghanistan	3.632

Top 10 Least Happy

Happiness by countries

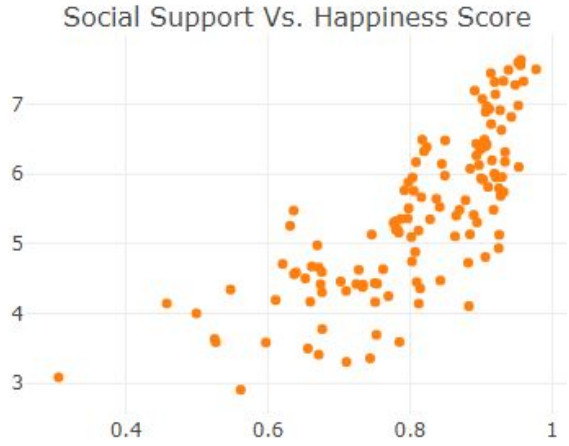
	country	Changes_in_happiness_scores
8	Latvia	1.026
2	Bulgaria	1.021
10	Macedonia	0.880
6	Hungary	0.810
9	Lithuania	0.660

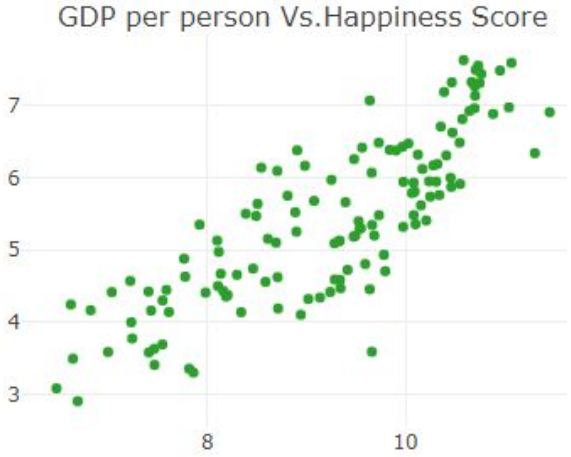
Top 5 countries getting happier

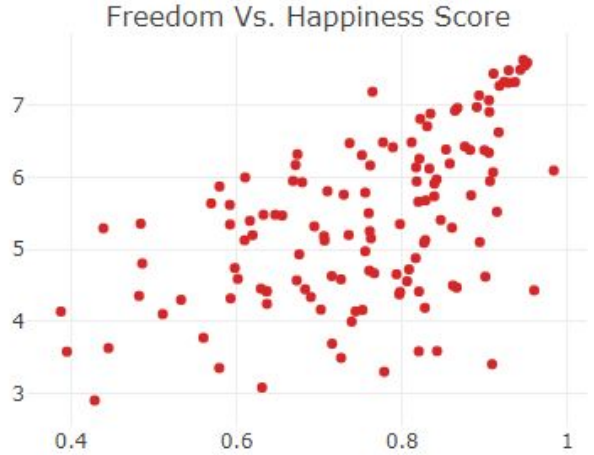
	country	Changes_in_happiness_scores
0	Albania	-0.791
3	Croatia	-0.198
11	Montenegro	0.221
12	Poland	0.275
1	Bosnia and Herzegovina	0.313

Top 5 countries getting sadder

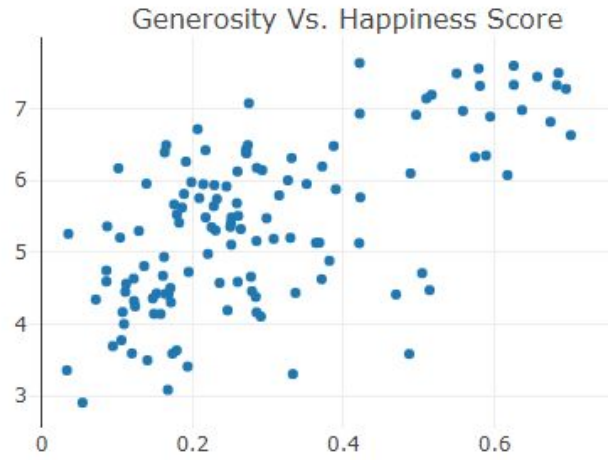
Factors Vs Happiness Score

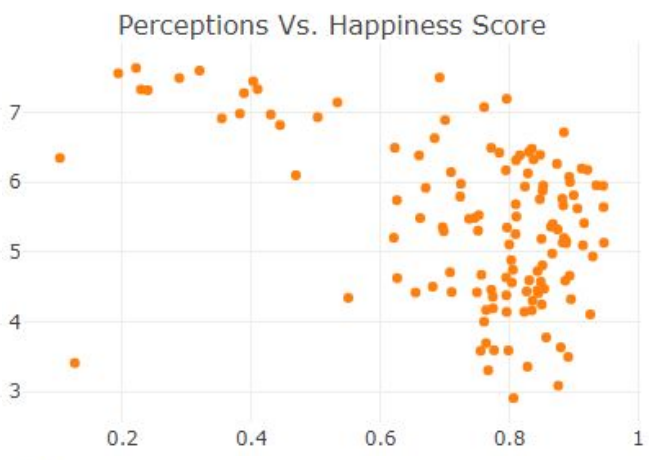


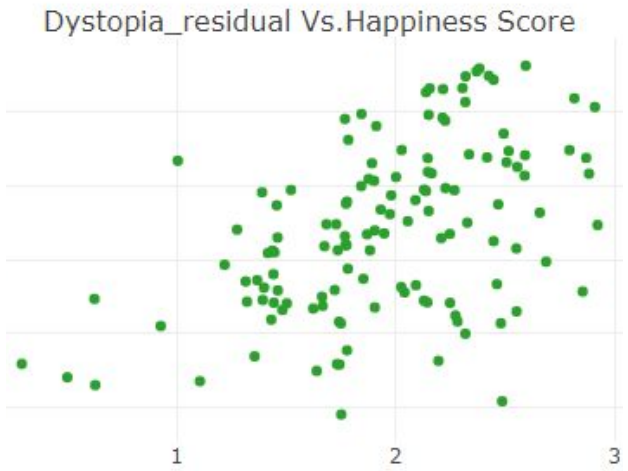




Factors Vs Happiness Score

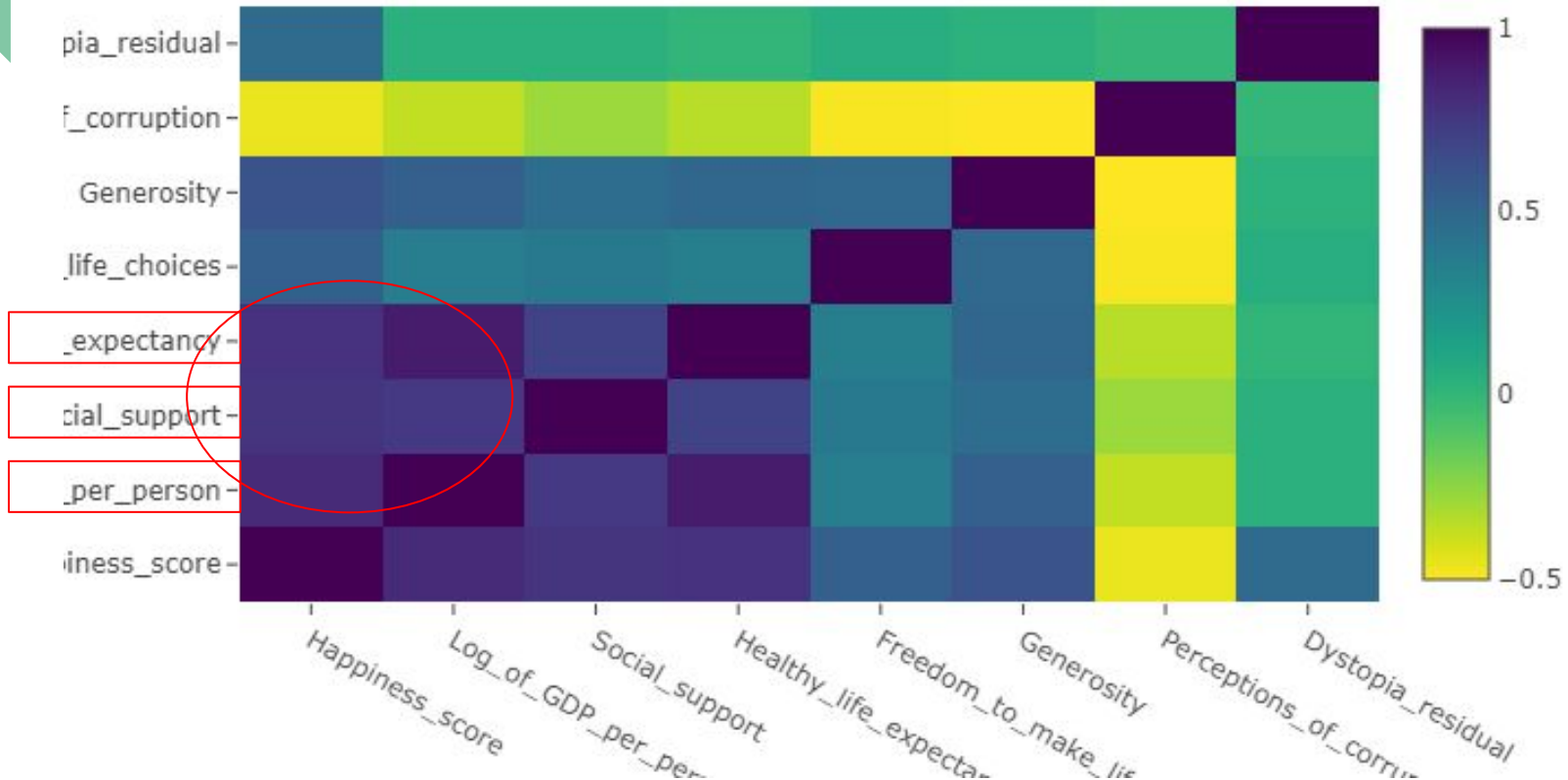




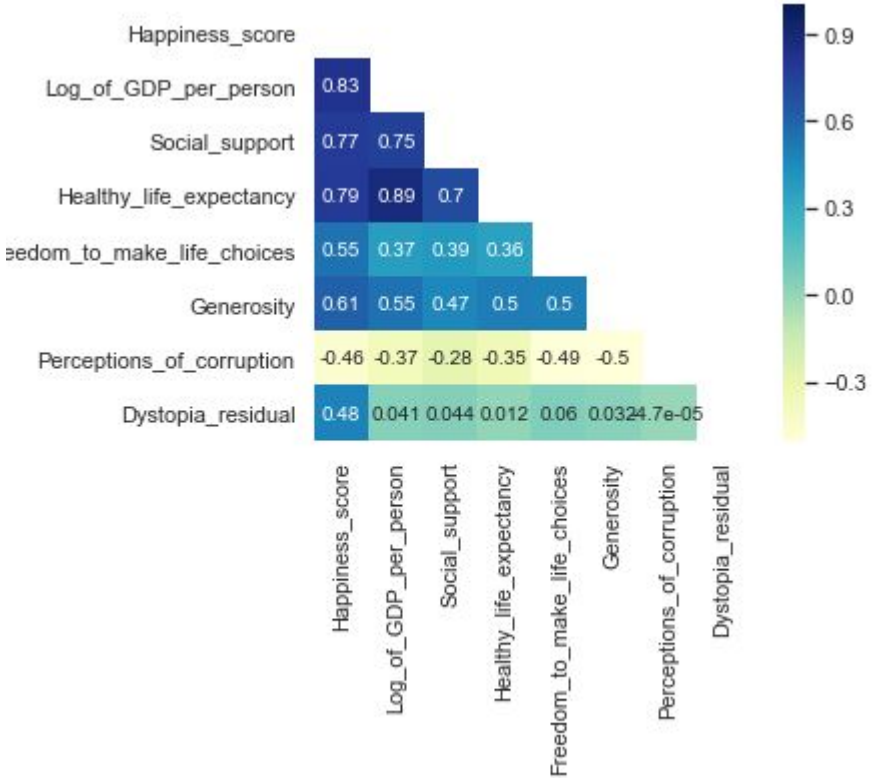


Moderate Strong Positve Relations: Social Support, Life Expectancy, GDP
Not So Strong Relations: Freedom of life choice, Generosity, Dystopia Residual
Negative Relation: Perception of corruption

Heatmap


pia_residual
_corruption
Generosity
life_choices
_expectancy
cial_support
_per_person
iness_score
Happiness_score
Log_of_GDP_per_per
Social_support
Healthy_life_expecta
Freedom_to_make_lif
Generosity
Perceptions_of_corru
Dystopia_residual
1
0.5
0
−0.5

Heatmap corr

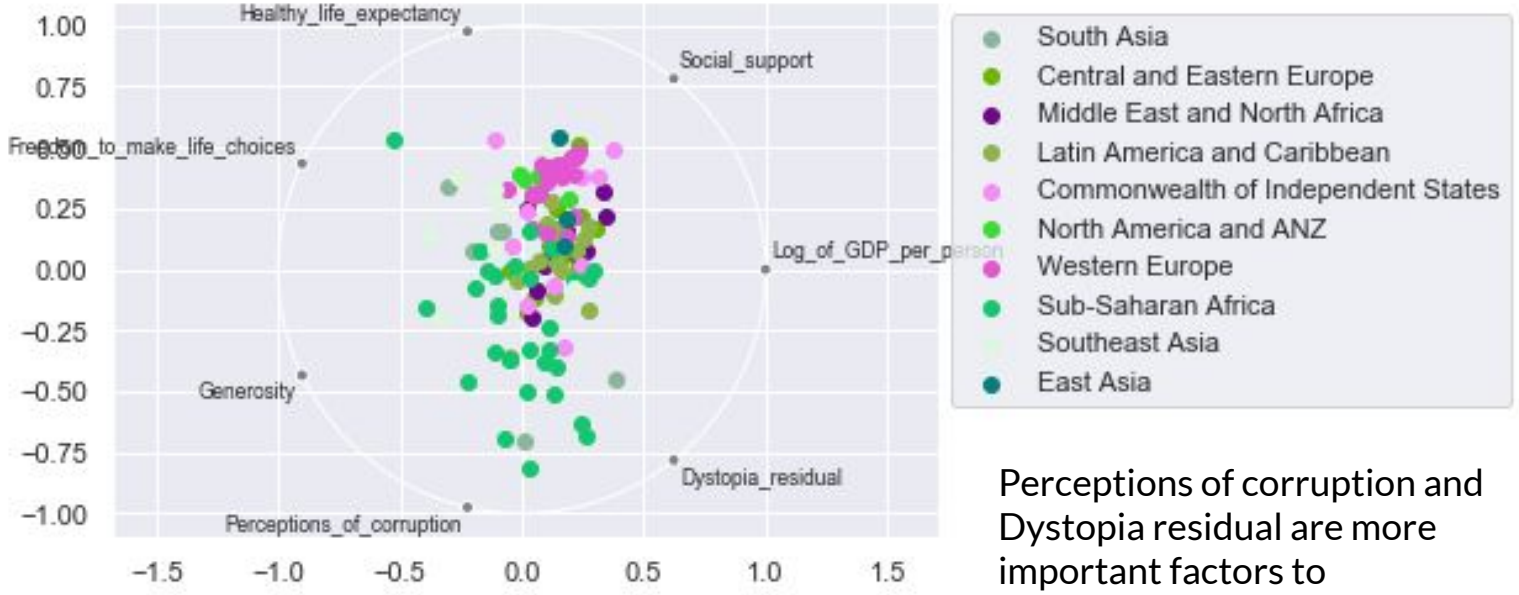


1. GDP Vs Happiness
2. Social Support Vs Happiness
3. Life Expectancy Vs Happiness
4. Life Expectancy Vs GDP
5. Life Expectancy Vs Social Support
6. Social Support Vs GDP
7. Perceptions of corruption increase lead to decrease of happiness

Top 3 Variables affect Happiness

GDP is the most important variable towards happiness,follow by Life expectancy and social support

Importance of Variables to Region



Perceptions of corruption and Dystopia residual are more important factors to sub-saharan Africa which is the number one in the list of unhappy regions.

Best Countries Analysis

	country	Happiness_score
44	Finland	7.632
105	Norway	7.594
37	Denmark	7.555
57	Iceland	7.495
133	Switzerland	7.487
99	Netherlands	7.441
25	Canada	7.328
100	New Zealand	7.324
132	Sweden	7.314
6	Australia	7.272

Top 10 Most Happy

	country	Log_of_GDP_per_person
146	Luxembourg	11.458786
93	Singapore	11.308526
150	Norway	11.071932
144	Ireland	11.046095
154	Switzerland	10.956634
79	United States	10.886991
148	Netherlands	10.767729
153	Sweden	10.750161
138	Denmark	10.736396
143	Iceland	10.711179

Top 10 GDP

	country	Social_support
143	Iceland	0.977497
78	New Zealand	0.959619
139	Finland	0.956149
138	Denmark	0.955462
28	Uzbekistan	0.953000
144	Ireland	0.952268
150	Norway	0.952147
76	Australia	0.948141
155	United Kingdom	0.942501
154	Switzerland	0.938523

Top 10 Social Support

	country	Healthy_life_expectancy
93	Singapore	75.720833
31	Japan	75.088242
152	Spain	74.363487
33	South Korea	74.041695
145	Italy	73.782936
154	Switzerland	73.173759
143	Iceland	72.755981
153	Sweden	72.745270
76	Australia	72.650299
137	Cyprus	72.608994

Top 10 Life Expectancy

Best Countries Analysis

Variables	Countries in Common
GDP & Life Expectancy	Singapore, Switzerland,Sweden,Iceland
GDP & Social Support	Norway,Ireland,Switzerland, Denmark Iceland
GDP & Happiness Score	Norway,Switzerland,Netherlands,Sweden,Denmark,Iceland
Happiness Score & Social Support	Iceland, New Zealand,finland,denmark,Norway,Australia,Switzerland
Happiness Score & Life Expectancy	Iceland, Switzerland, Sweden,Australia
Social Support and Life Expectancy	Iceland,Australia,switzerland

Top 3 best countries: **Switzerland, Iceland,Norway**

Findings from EDA

- GDP is the most important factor influence the happiness score, followed by life expectancy and social support

- Perceptions of corruption increase lead to decrease of happiness score
- Perceptions of corruption and Dystopia residual are more important factors contribute to the least happy region

- Log of GDP per person and Social Support are more important factors contribute to the most happy region

- Top 3 best countries: **Switzerland, Iceland,Norway**

Findings from EDA

- Top 3 best countries: **Switzerland, Iceland,Norway**

Switzerland / Population (2017)	Iceland / Population (2017)	Norway / Population (2017)
8.42 million (2017)	338,349 (2017)	5.258 million (2017)

- Population might be a factor influence the happiness score
- **Switzerland:** Free market economies
  Iceland: Natural resource and Tourism
  Norway: Oil Gas production and fish, forests, and minerals

Machine Learning Based Prediction Models

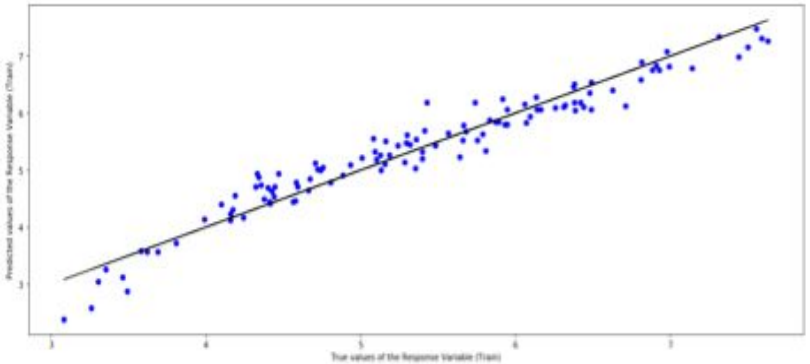
Machine Learning Algorithms Used :

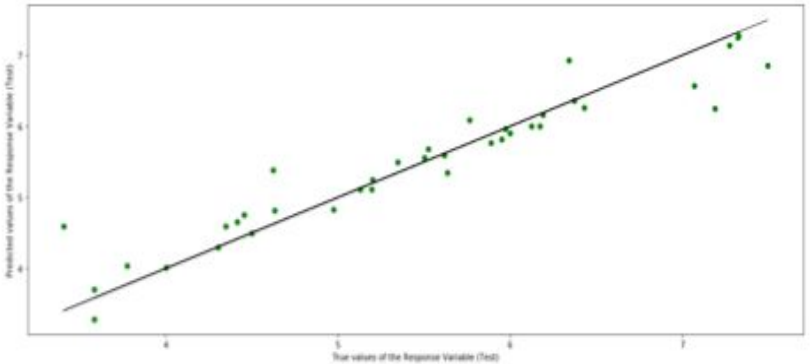
1. Linear Models
2. Support Vector Regression
3. Random Forest Regression
4. Neural Networks

Linear Models

1. Multivariate Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net
5. Bayesian Ridge Regression

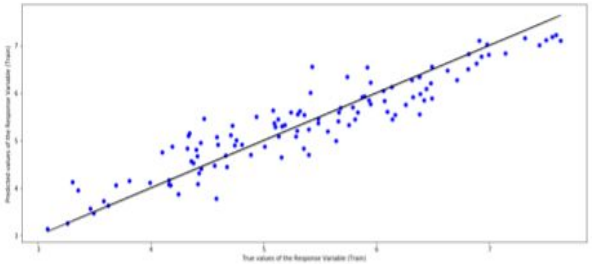
Multivariate Linear Regression

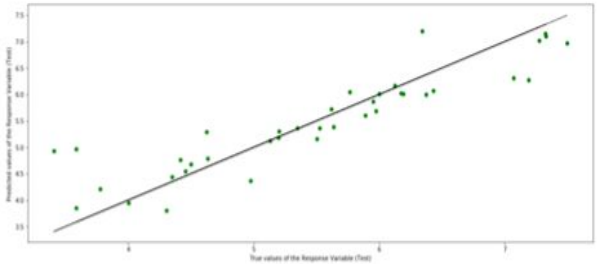


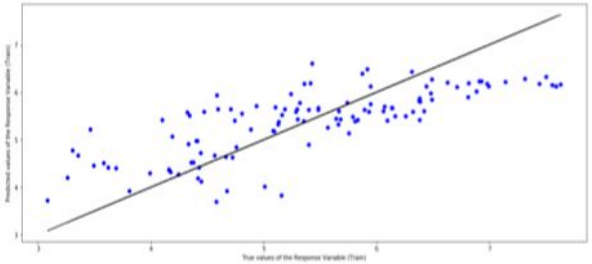


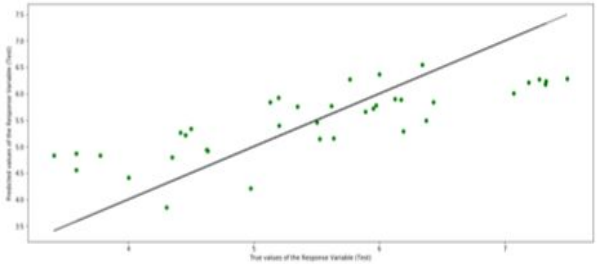
```
Goodness of Fit of Model        Train Dataset Linear
Mean Squared Error (MSE)        : 0.0714594516280009

Goodness of Fit of Model        Test Dataset Linear
Mean Squared Error (MSE)        : 0.12338277223266597
```

Ridge Regression and Lasso Regression









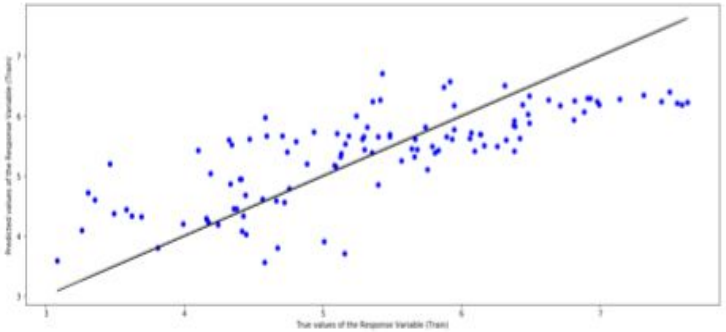
Goodness of Fit of Model Train Dataset Ridge
Mean Squared Error (MSE) : 0.1565631033856096

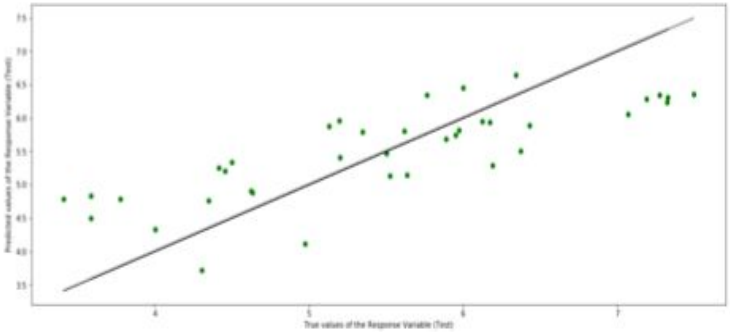
Goodness of Fit of Model Test Dataset Ridge
Mean Squared Error (MSE) : 0.23975075205859164

Goodness of Fit of Model Train Dataset Lasso
Mean Squared Error (MSE) : 0.5090139161539985

Goodness of Fit of Model Test Dataset Lasso
Mean Squared Error (MSE) : 0.53751069946733

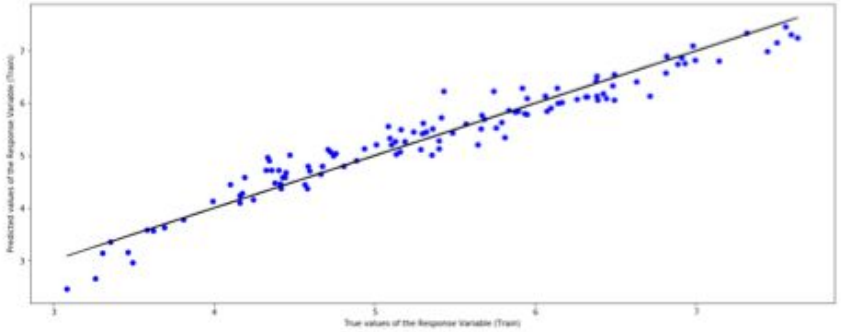
Elastic Net Regression

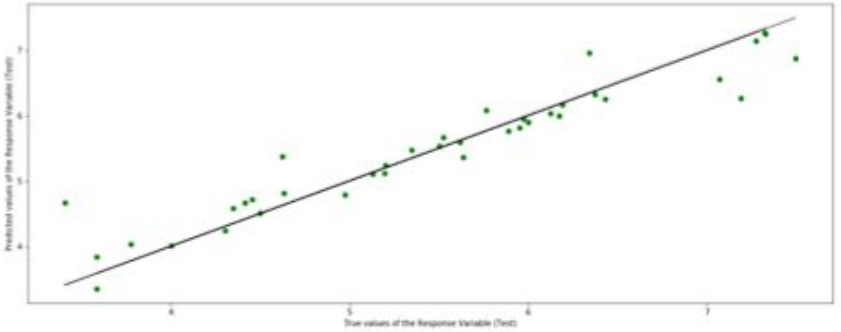




```
Goodness of Fit of Model          Train Dataset Elastic Net
Mean Squared Error (MSE)          : 0.4975326039133302

Goodness of Fit of Model          Test Dataset Elastic Net
Mean Squared Error (MSE)          : 0.5130442550733312
```

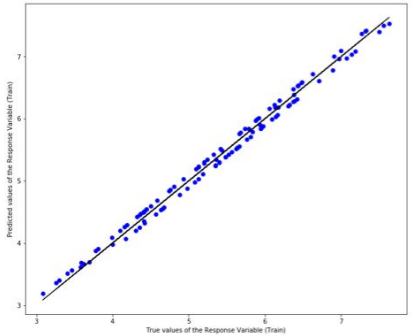
Bayesian Ridge Regression

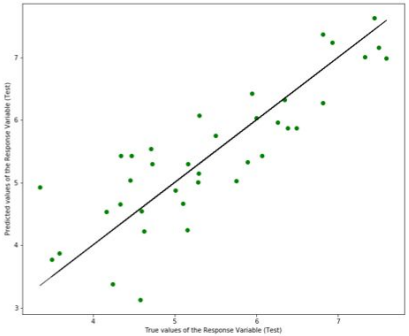


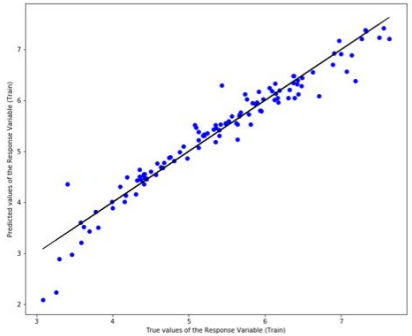


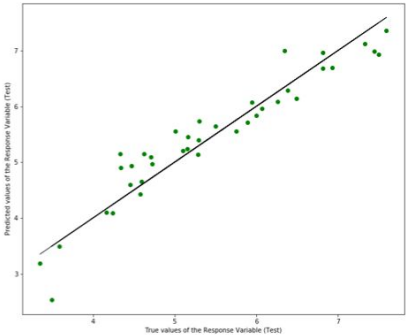
```
Goodness of Fit of Model        Train Dataset Bayesian Ridge
Mean Squared Error (MSE)        : 0.07239337075460665

Goodness of Fit of Model        Test Dataset Bayesian Ridge
Mean Squared Error (MSE)        : 0.1269202545421794
```

Support Vector Regression



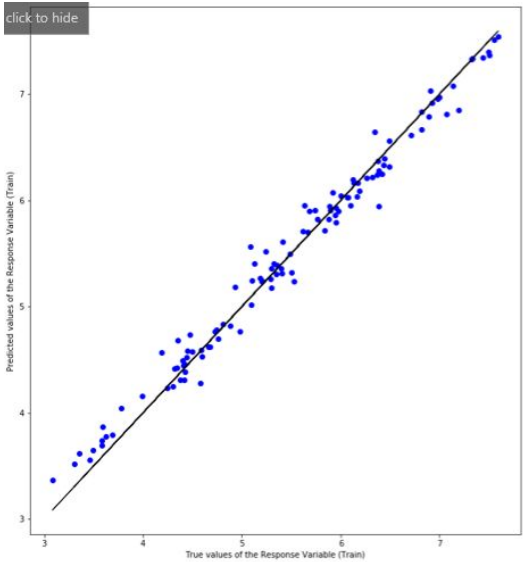


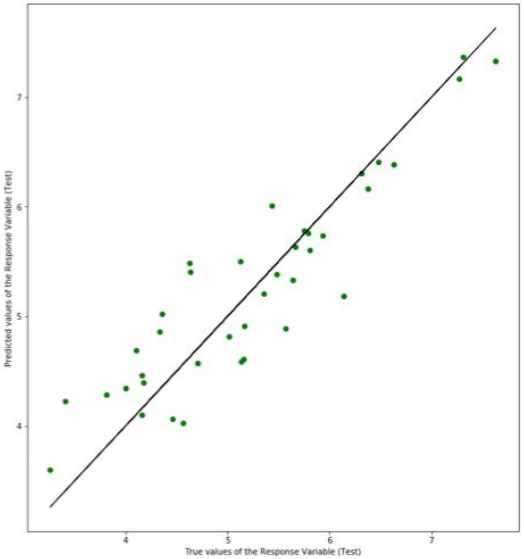





```
Goodness of Fit of Model        Train Dataset RBF
Mean Squared Error (MSE)        : 0.008789776745780341

Goodness of Fit of Model        Test Dataset RBF
Mean Squared Error (MSE)        : 0.3868240071214362

Goodness of Fit of Model        Train Dataset LIN
Mean Squared Error (MSE)        : 0.07410400159010574

Goodness of Fit of Model        Test Dataset LIN
Mean Squared Error (MSE)        : 0.1255759640370924
```

Random Forest Regression



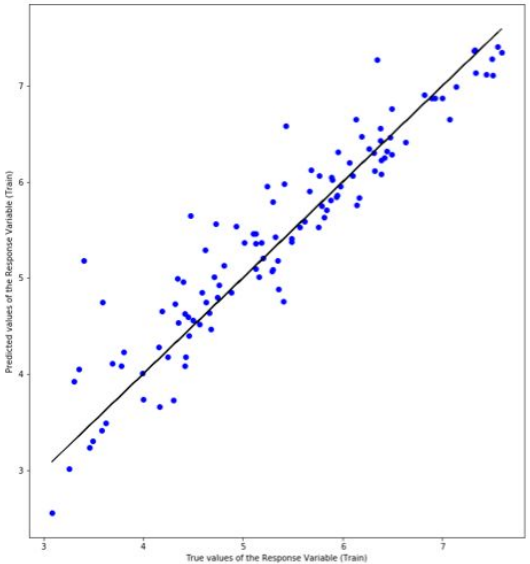


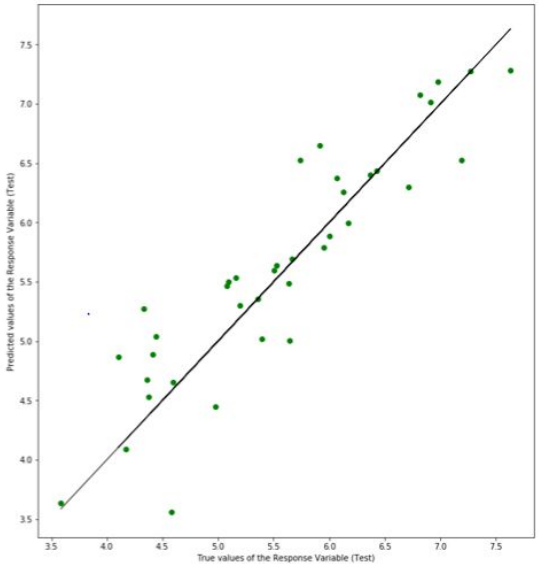

```
Goodness of Fit of Model        Train Dataset
Mean Squared Error (MSE)        : 0.023924491017398256

Goodness of Fit of Model        Test Dataset
Mean Squared Error (MSE)        : 0.18697539871304247
```

Neural Networks

1. Using MLP Regressor in Scikit
2. Using Keras with TensorFlow

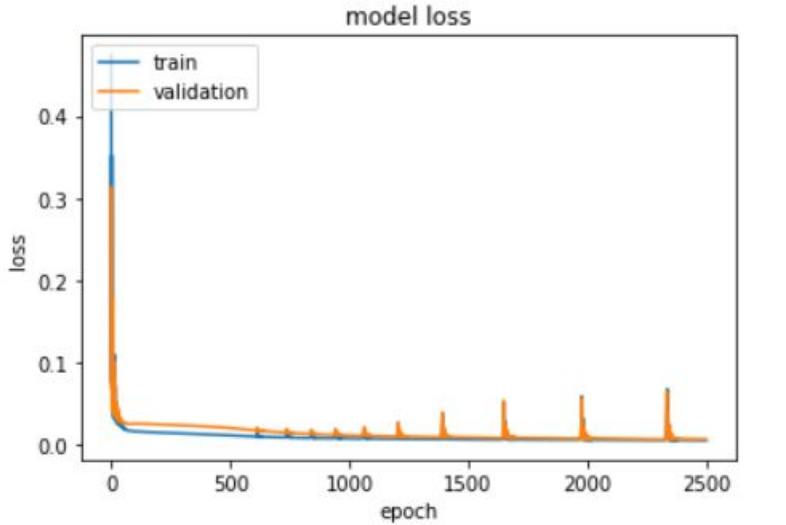
Using Scikit



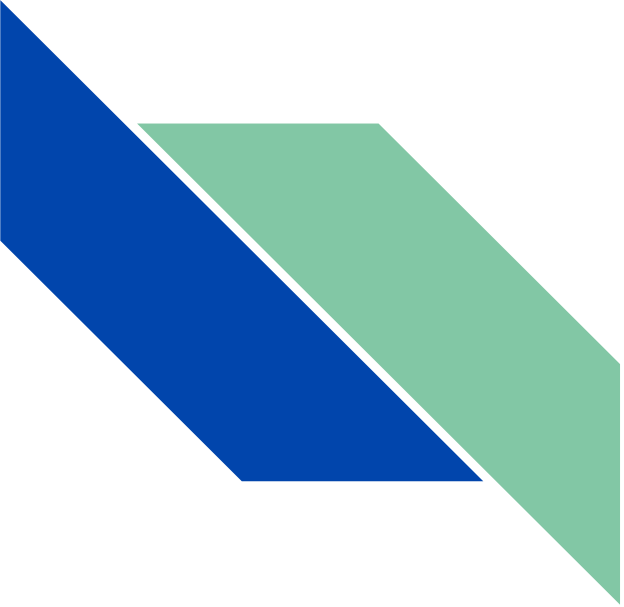



```
Goodness of Fit of Model        Train Dataset
Mean Squared Error (MSE)        : 0.16168592284911787

Goodness of Fit of Model        Test Dataset
Mean Squared Error (MSE)        : 0.1783300676335056
```

Using Keras with TensorFlow




```
Goodness of Fit of Model        Train Dataset
Mean Squared Error (MSE)        : 0.006453124621924726

Goodness of Fit of Model        Test Dataset
Mean Squared Error (MSE)        : 0.0045279696275074026
```



Clustering

- KMeans++
- DBSCAN

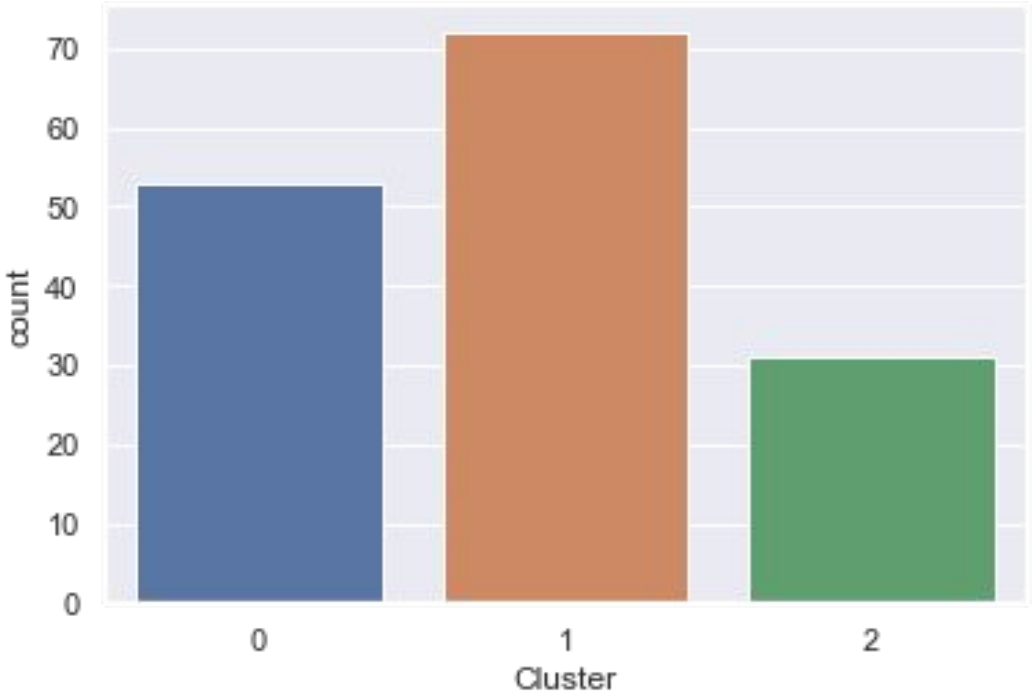
Data Preparation for Clustering

- Done on GDP, Life Expectancy & Social Support (Based on Exploratory Data Analysis)

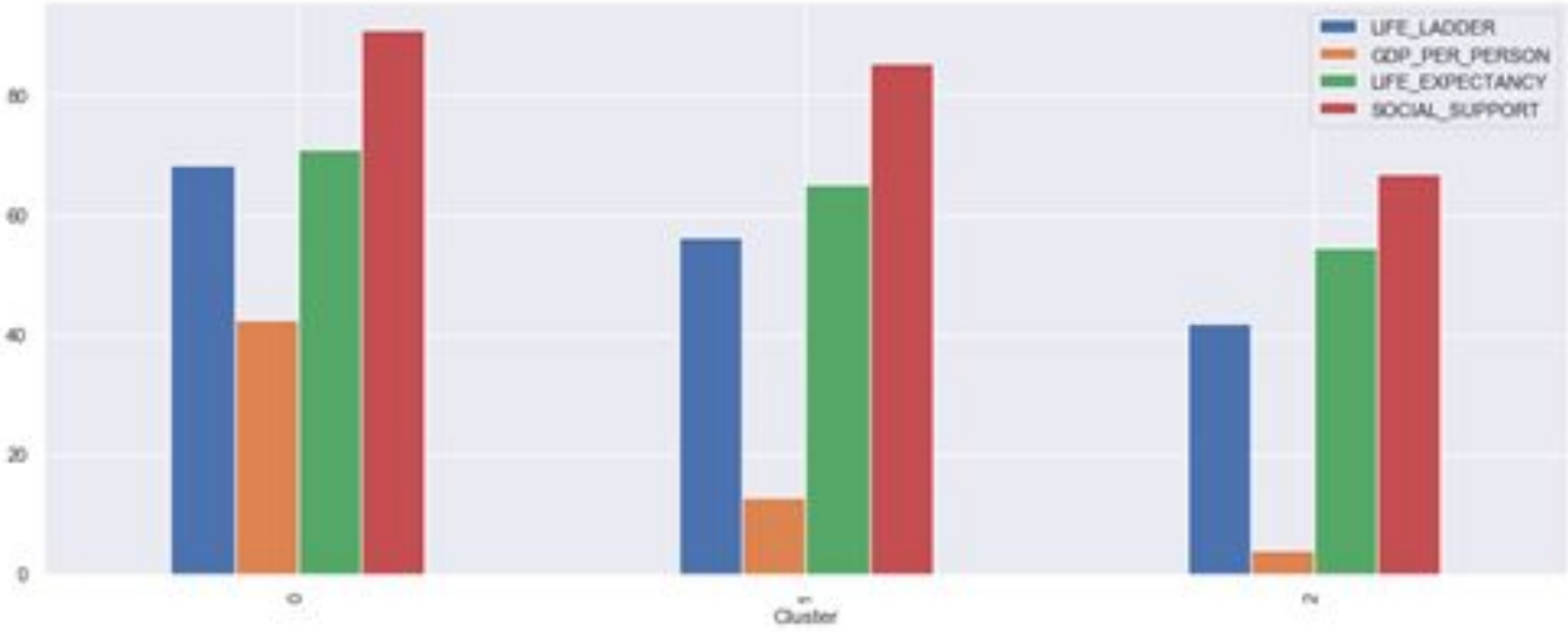
- Rescaling of all variables to, out of 100

1. Making all column Names uppercase
2. Removing 2015-2017 at the end of the columns
3. Replacing spaces with '_'
4. Setting Country Name as Index
5. Replacing NULL Values with mean of the column

K Means on World Data



Clusters using K Means

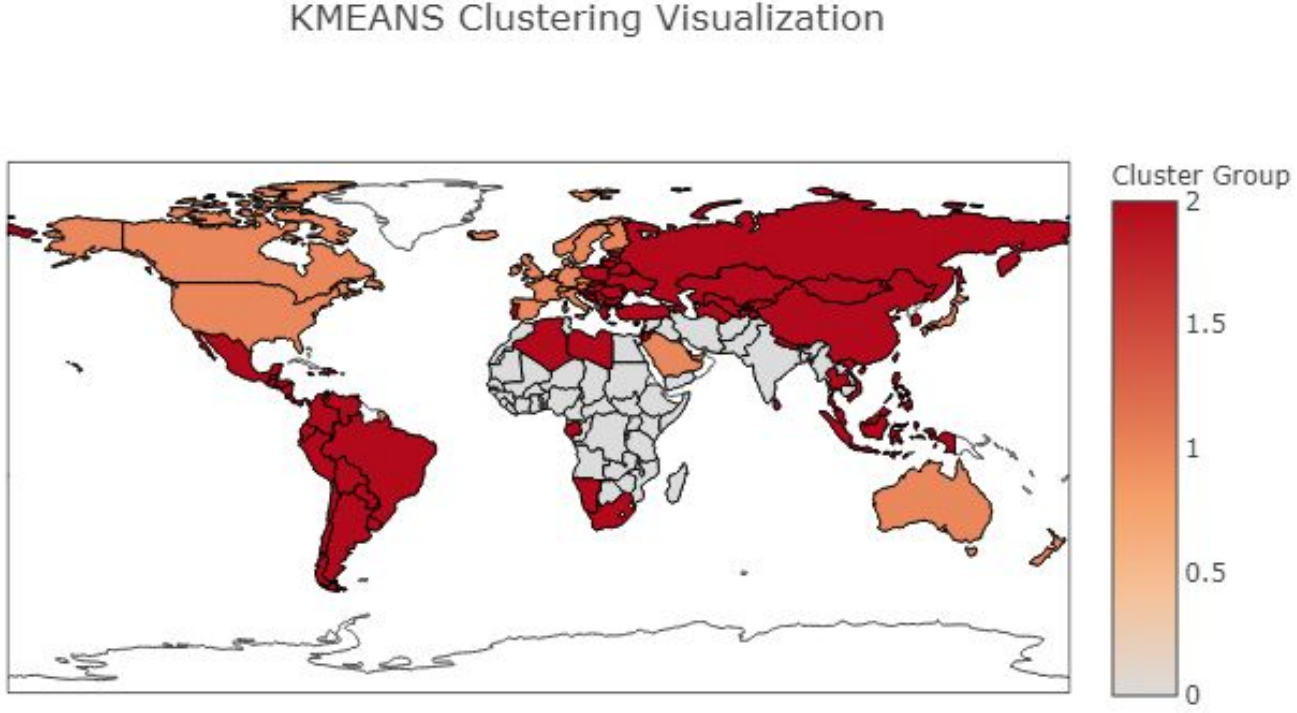


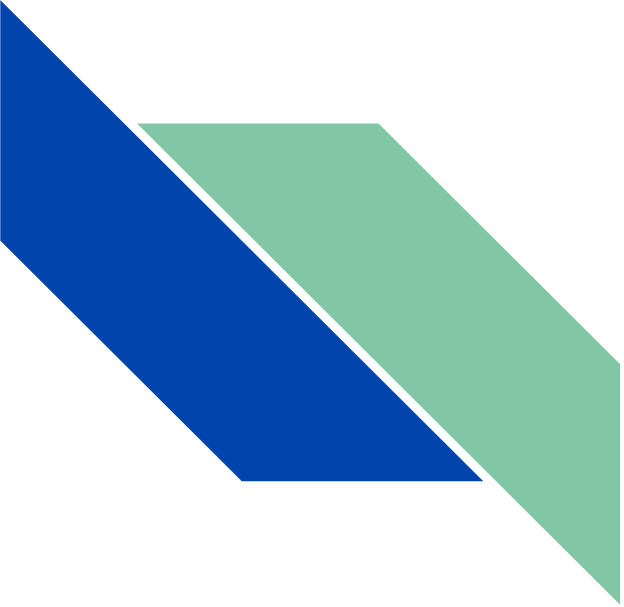
Most Happy

Moderately Happy

Least Happy

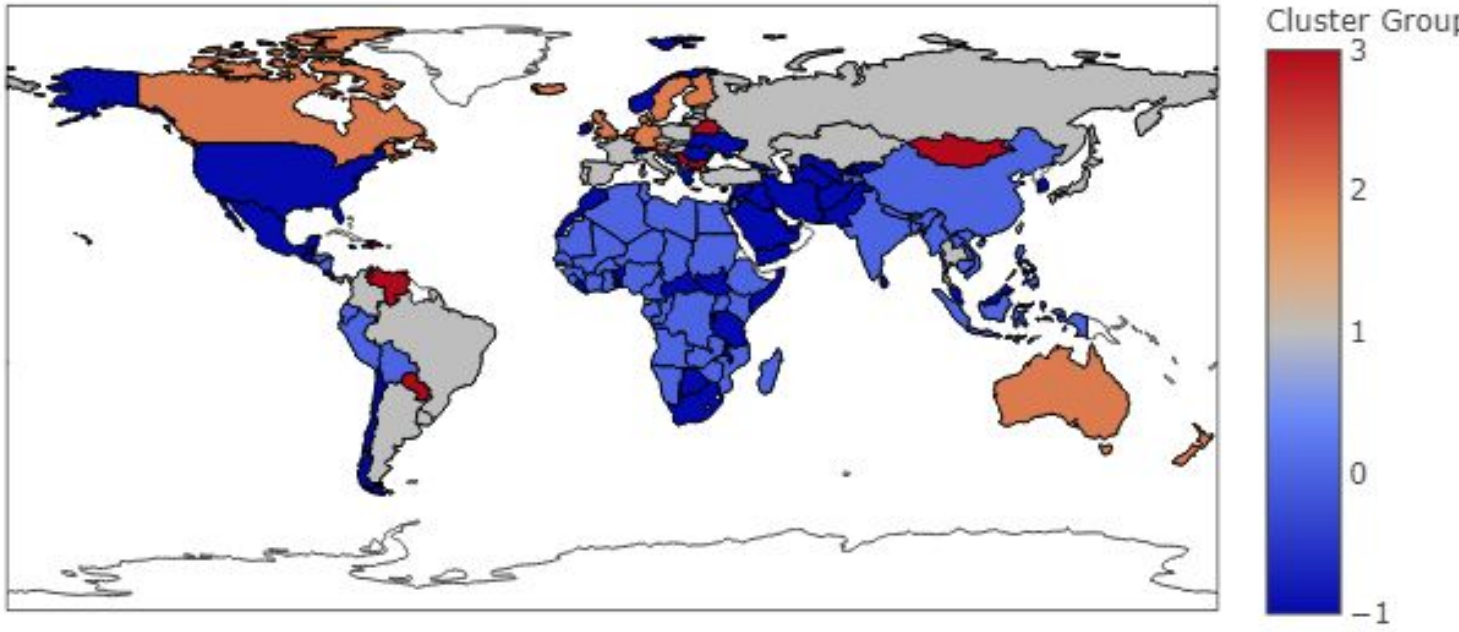
Clustering Visualisation





Clustering using DBSCAN

DBSCAN Clustering Visualization



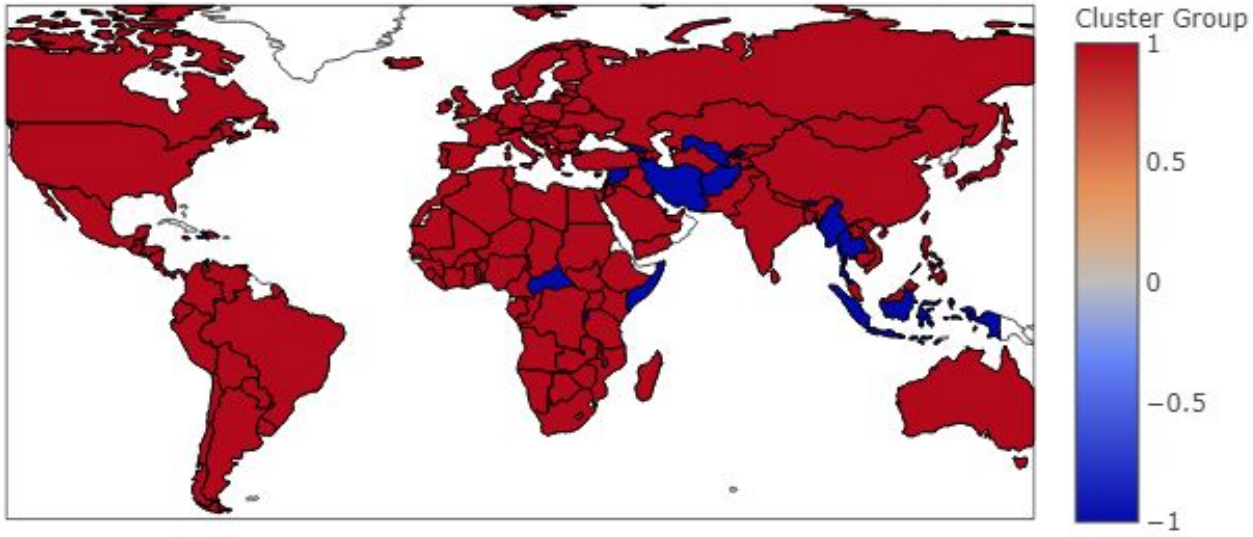
Imp Points

- ➔ Number of clusters not initially defined
  - ◆ EPS
  - ◆ Number of Samples
- ➔ Noise Points

Anomaly Detection

using Local Outlier Factor Algorithm

Anomaly Visualization

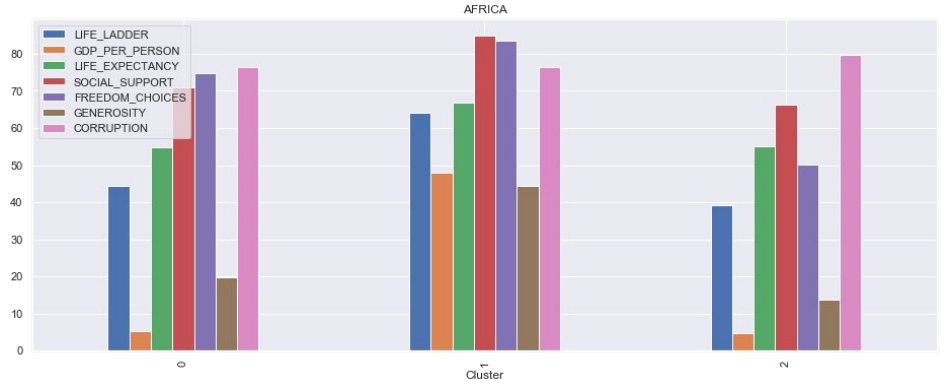


Anomalies

- Central African Republic
- Haiti
- Indonesia
- Myanmar
- Qatar
- Rwanda
- Somalia
- Uzbekistan

Clustering for Independent Regions

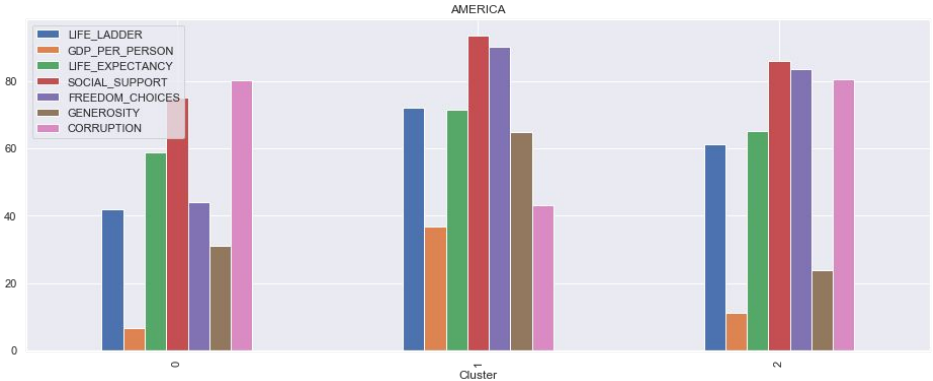
Analysing the differentiating factors among different regions



AFRICA

Within Cluster Sum of Squares : 32146.45

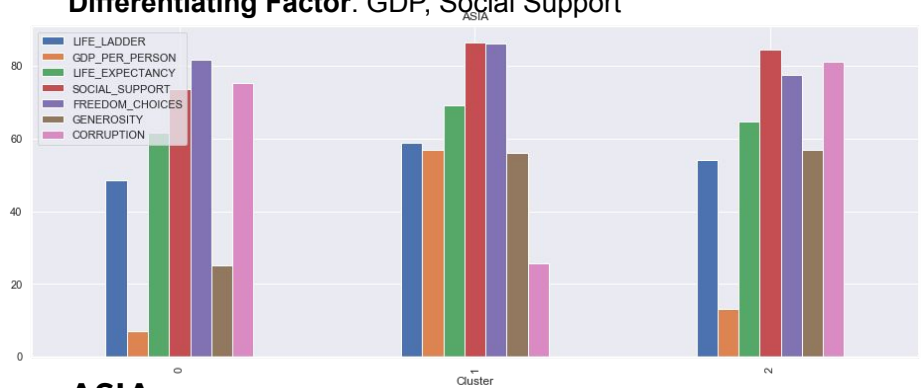
Differentiating Factor: GDP, Social Support



AMERICA

Within Cluster Sum of Squares : 6609.34

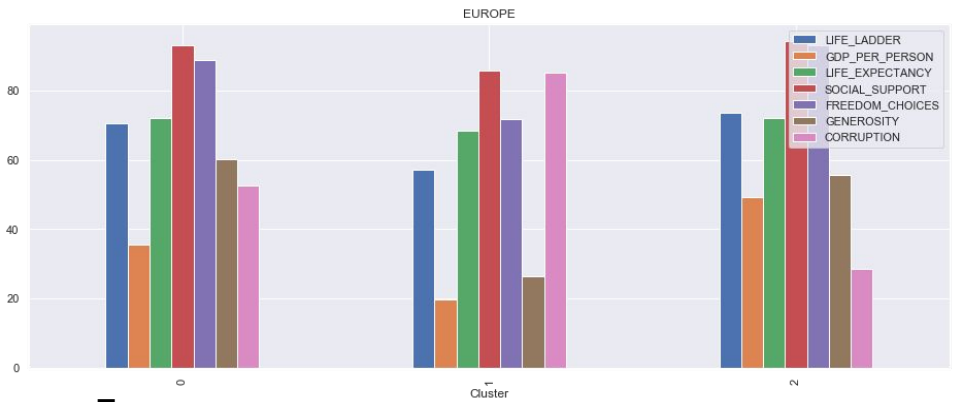
Differentiating Factor: GDP, Generosity



ASIA

Within Cluster Sum of Squares : 13263.38

Differentiating Factor: GDP, Corruption



Europe

Within Cluster Sum of Squares : 13309.47

Differentiating Factor: Freedom of Choice, Generosity

Outcomes

- The most efficient machine learning algorithm for predicting Happiness is **Neural Networks** executed with Tensor Flow.

- The most prominent factor affecting happiness in:
  - **The World:** GDP, Life Expectancy & Social Support
  - Different **sub -regions:**
    Africa : GDP, Social Support, Corruption
    America: GDP, Generosity
    Asia: GDP, Corruption
    Europe: Freedom of Choice, Generosity

Work Distribution

Work Assigned	Person In Charged	Remarks
Exploratory Data Analysis (EDA)	Wang Wen	**Tools**:Plotly,Seaborn and basic library Powerpoint Slides
Machine Learning Based Prediction Models	Manav Arora	**Tools:** Scikit, TensorFlow and basic libraries Powerpoint Slides
Clustering & Anomaly Detection	Aditya Bansal	Using K Means++,DBSCAN **Tools:** Scikit, Plotly Powerpoint Slides