Project Report

Big Data and Artificial Intelligence

-

Abhishek Mangla
Abhishek Sehgal
Anurag Sharma
Arvind Sharma
Praneeth Yaramosu

Introduction

What Is Happiness?

Most of us probably don't believe we need a formal definition of happiness; we know it when we feel it, and we often use the term to describe a range of positive emotions, including joy, pride, contentment, and gratitude.

Inspiration

What countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness? How did country ranks or scores change between the 2015 and 2016 as well as the 2016 and 2017 reports? Did any country experience a significant increase or decrease in happiness?

The World Happiness Report is an annual publication of the United Nations Sustainable Development Solutions Network. It contains articles, and rankings of national happiness based on respondent ratings of their own lives,which the report also correlates with various life factors.

Dataset Description

The dataset we are using gives us data about the Happiness Index of a Country and multiple factors that affect it.

The data consists of happiness data from 2015 - 2019, which is combined in our analysis to make predictive models.

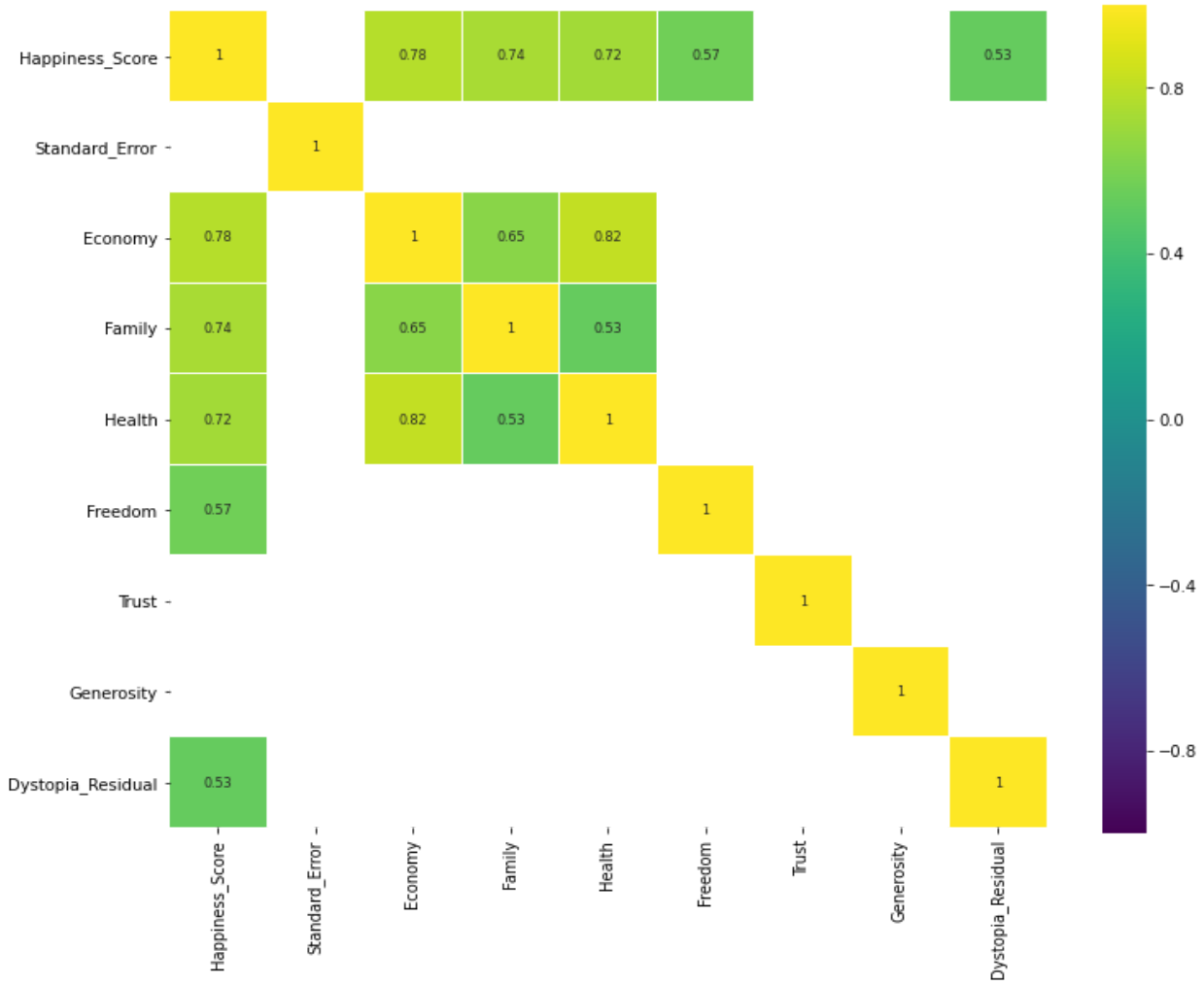
The total dataset consists of over 1000 records and 9 columns. The data has been extracted from Kaggle at "2019: State of World Happiness. What drives us?".

The different columns in our dataset are:

- **Country**: Name of the country.
- **Region**: Region the country belongs to.
- **Happiness Rank**: Rank of the country based on the Happiness Score.
- **Happiness Score**: Metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
- **Economy**: GDP Score corresponding to the country's GDP.
- **Family**: Normalized Family Score.
- **Health**: Normalized Life Expectancy Score.
- **Freedom**: Normalized Freedom Score.
- **Trust**: Normalized Trust Score
- **Generosity**: Normalized Generosity Score.

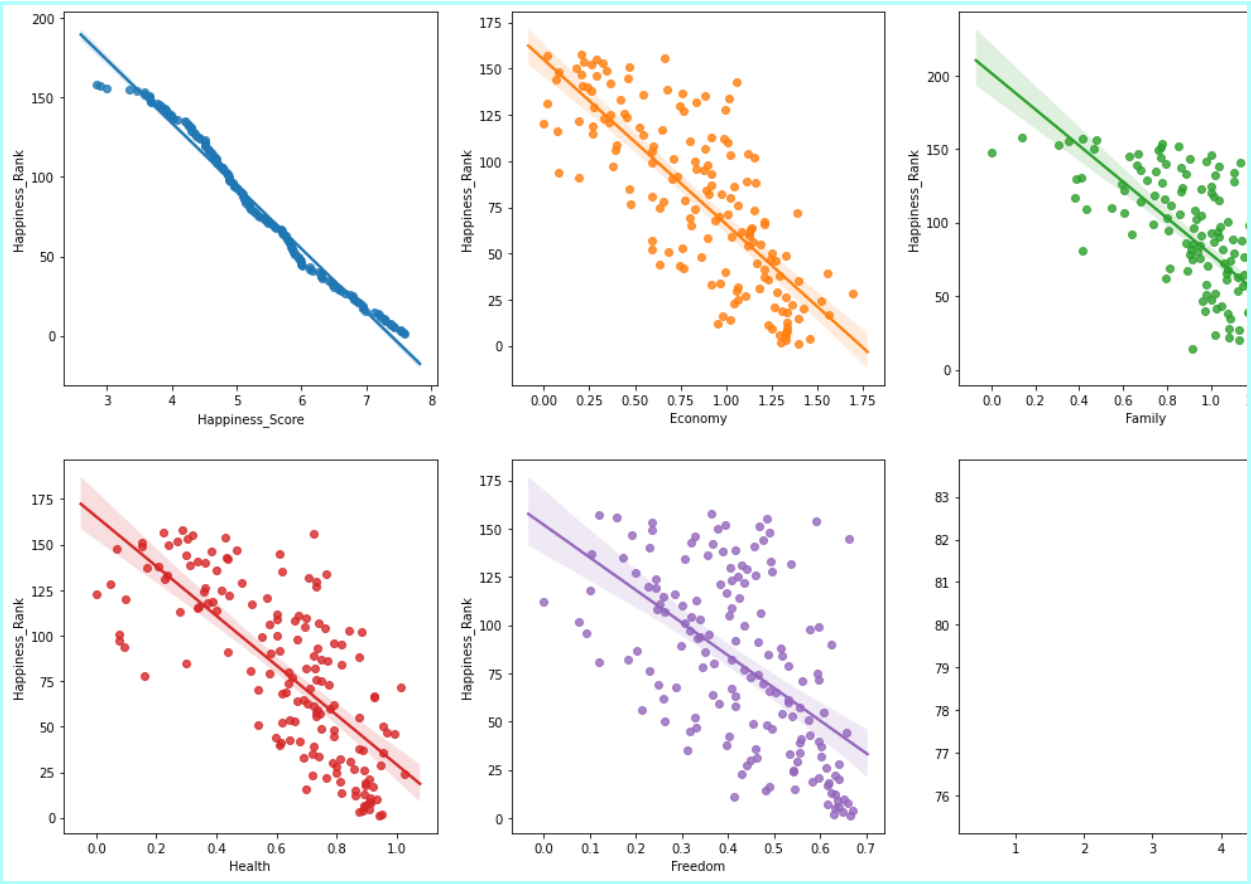
EDA

Correlation matrix - 2015



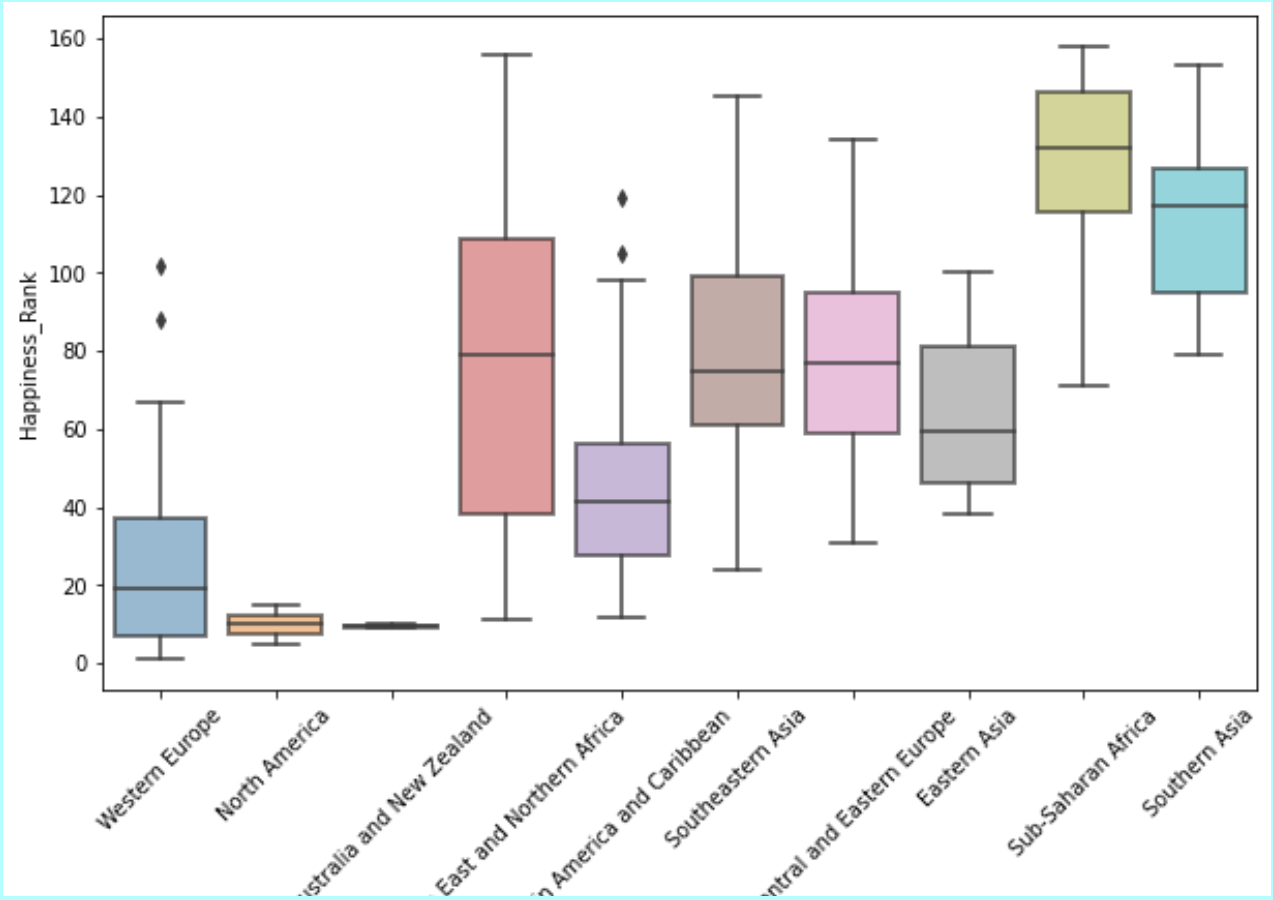
Observation - From the correlation matrix, we can clearly see that Economy and Family are the two biggest factors affecting the Happiness Score, which in turn affects the Happiness Rank.

Scatter Plot - 2015



Observation - It is also visible in the scatter plot that Economy and Family have the strongest correlation with Happiness Rank.

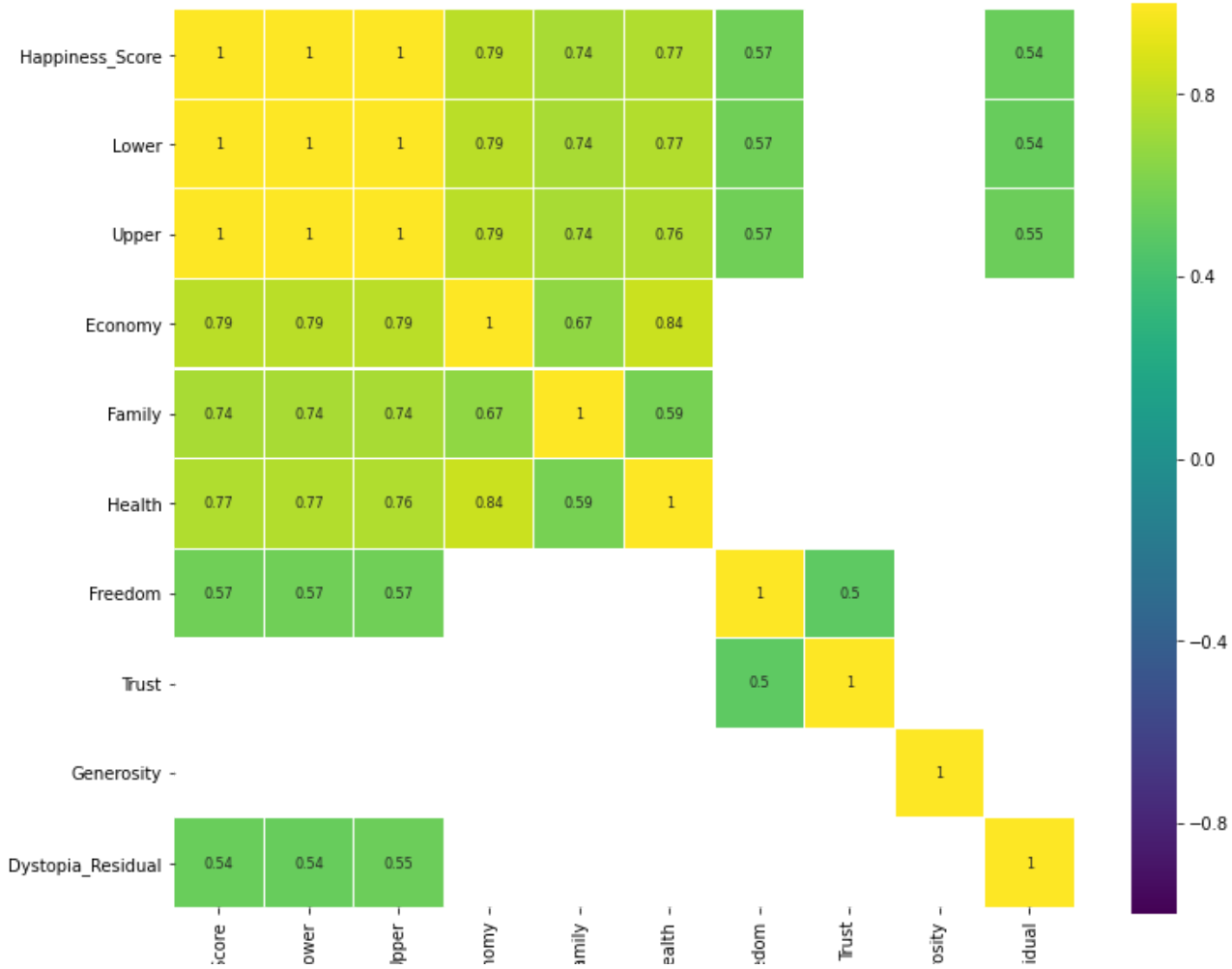
Region-wise Box Plots - 2015



Observation -

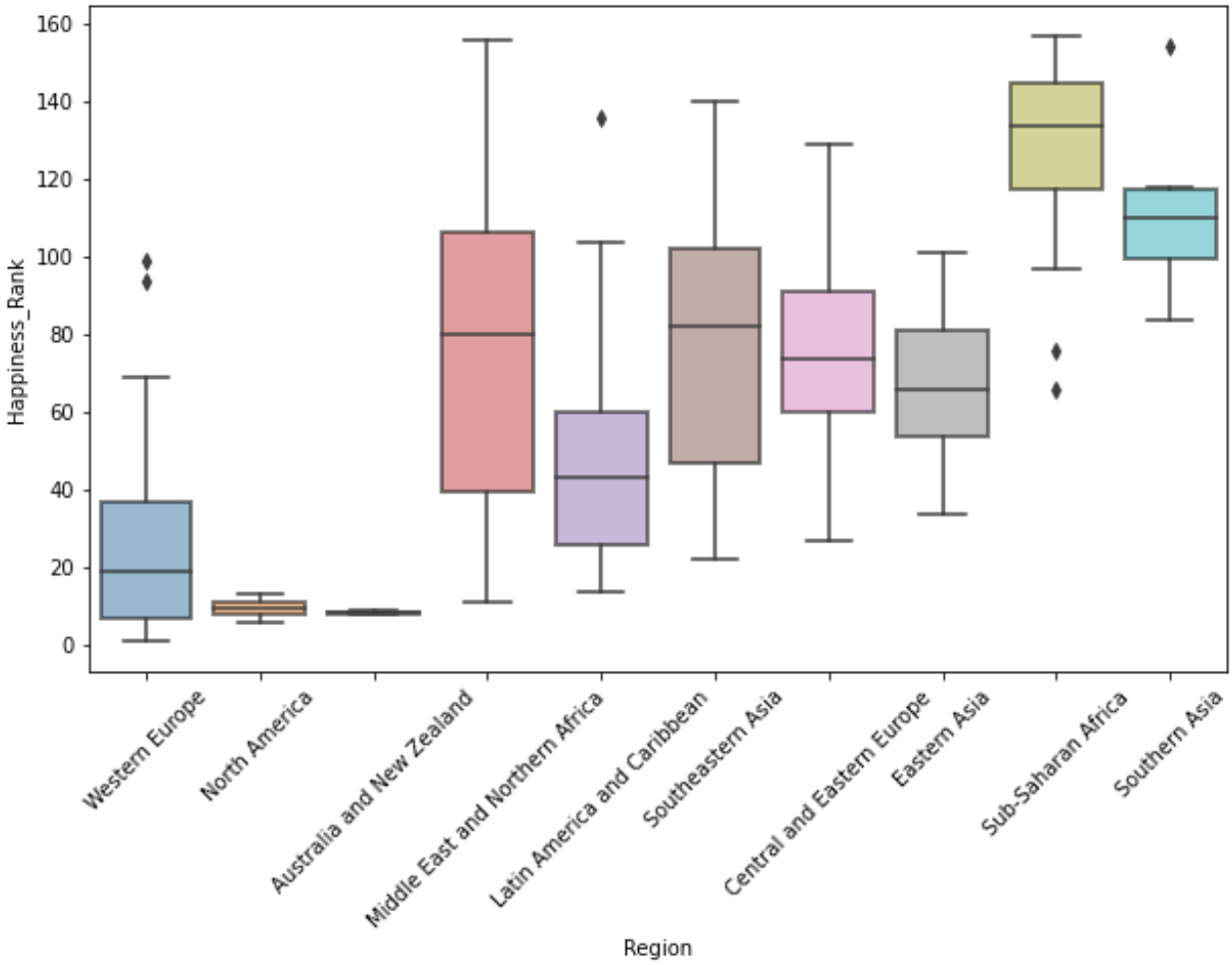
1. The Australia and New Zealand region has the highest variation in Happiness rank whereas North America has the least variation.
2. Happiness rank is highest in Sub-Saharan Africa and lowest in North America

Correlation matrix - 2016



Observation - From the correlation matrix, we can clearly see that Economy and Family are the two biggest factors affecting the Happiness Score, which in turn affects the Happiness Rank.

Region-wise Box Plots - 2016



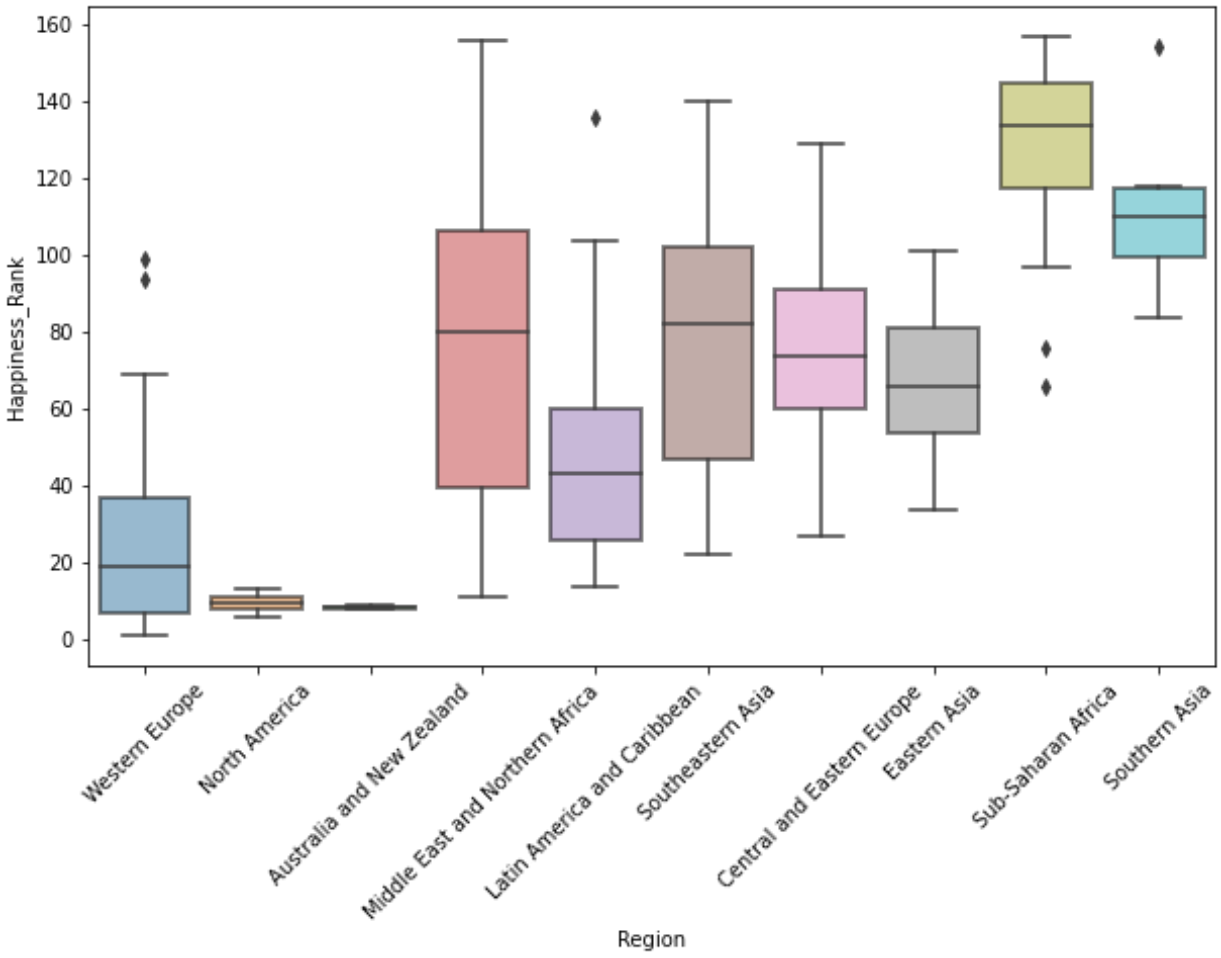
Observation -

1. The Australia and New Zealand region has the highest variation in Happiness rank whereas North America has the least variation.
2. Happiness rank is highest in Sub-Saharan Africa and lowest in North America

Correlation matrix - 2017

Observation - From the correlation matrix, we found that Economy and Health were the biggest factors impacting Happiness Rank, which is different from previous years where Family was an important factor.

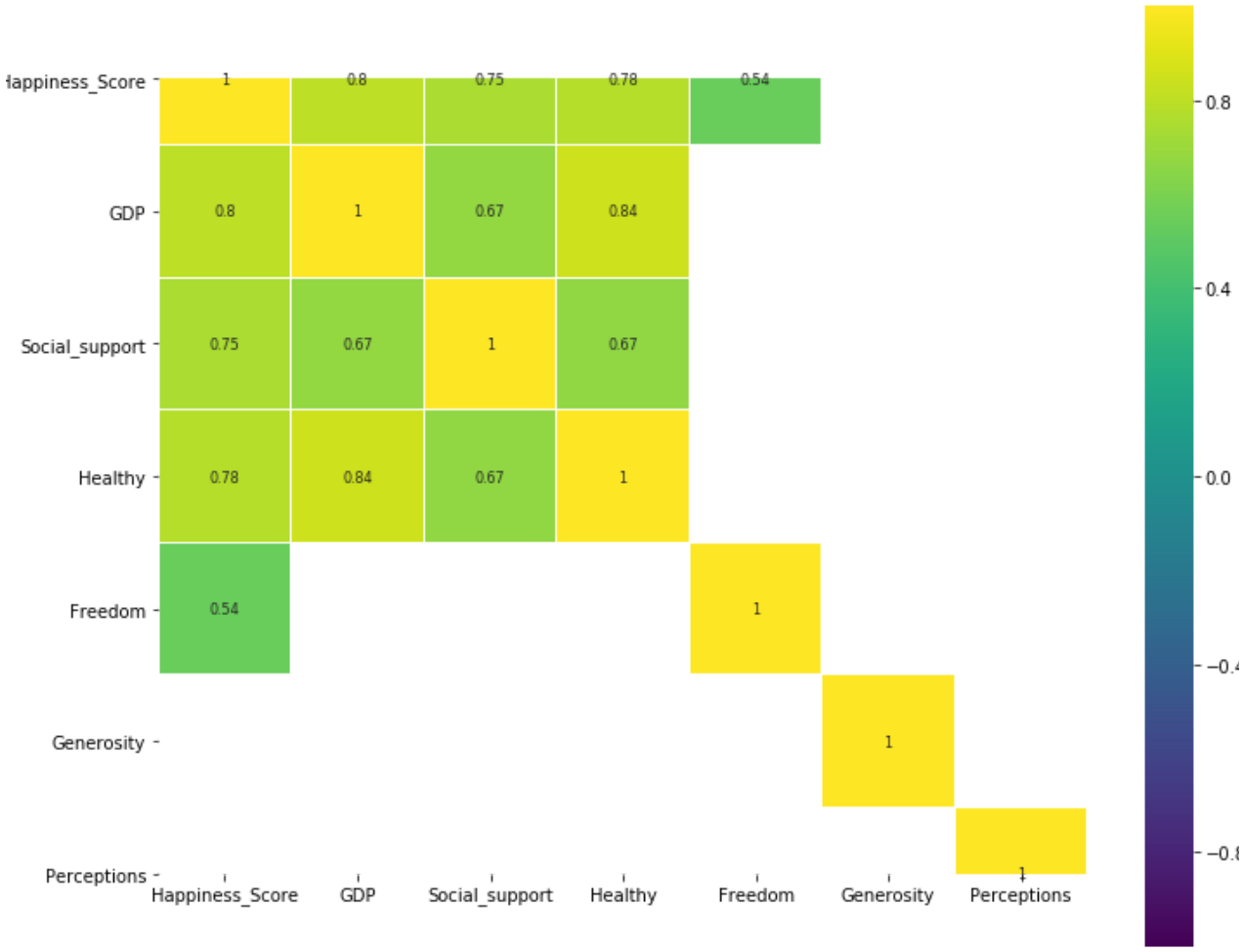
Region-wise Box Plots - 2017



Observation -

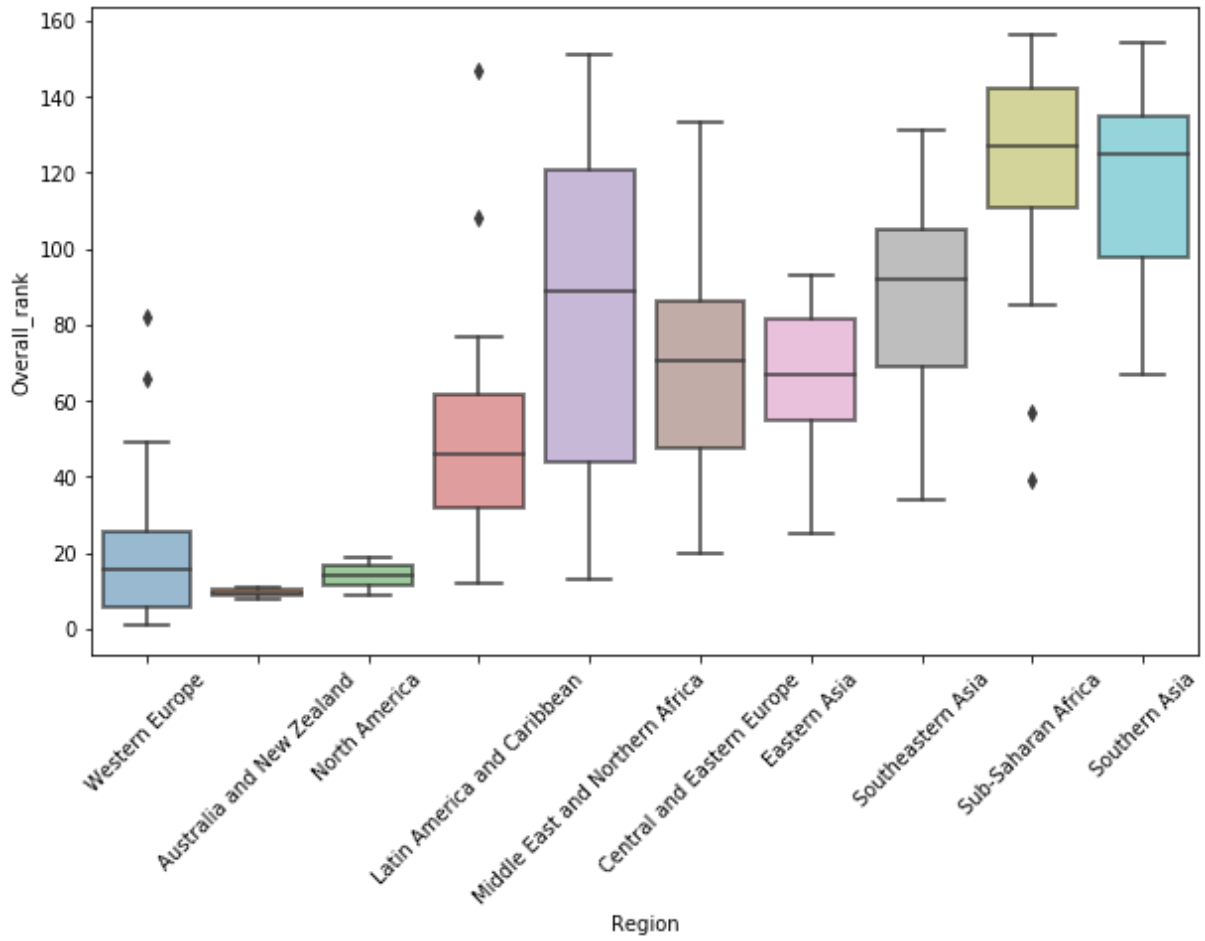
1. The Australia and New Zealand region has the highest variation in Happiness rank whereas North America has the least variation.
2. Happiness rank is highest in Sub-Saharan Africa and lowest in North America

Correlation matrix - 2018



Observation - From the correlation matrix, we found that GDP and health were the biggest factors impacting Happiness Rank, which is different from 2015 and 2016, where Family was an important factor, but like 2017

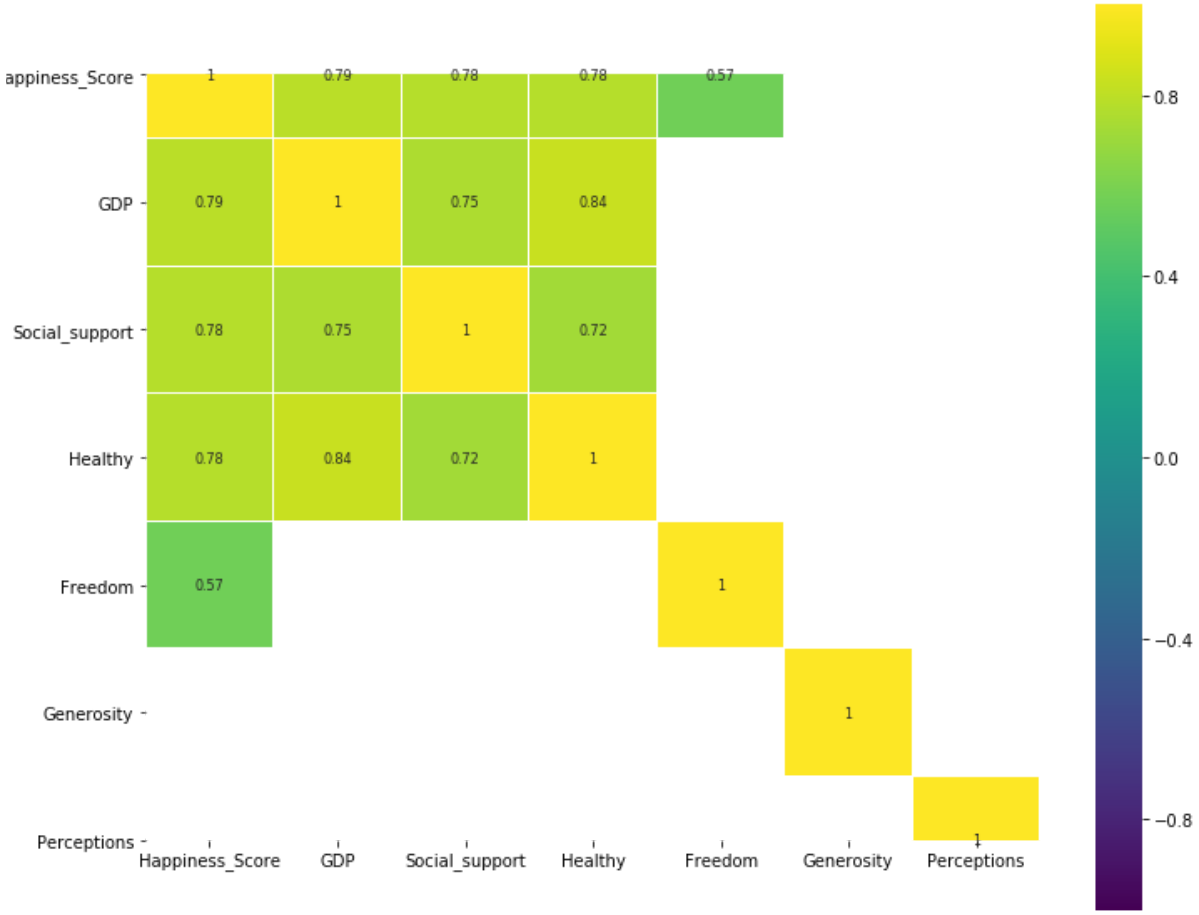
Region-wise Box Plots - 2018



Observation -

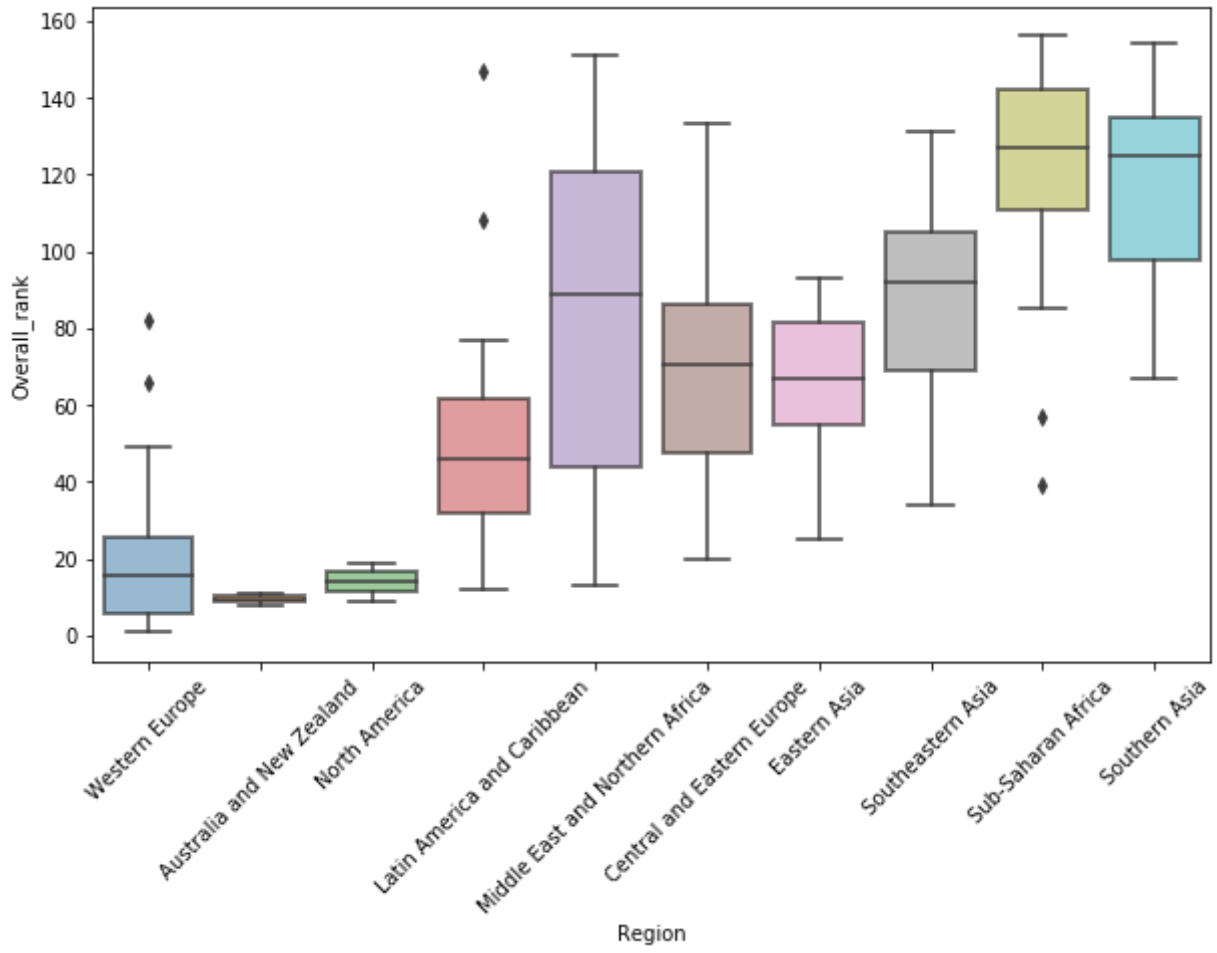
1. The Australia and New Zealand region has the highest variation in Happiness rank whereas North America has the least variation.
2. Happiness rank is highest in Sub-Saharan Africa and lowest in North America

Correlation matrix - 2019



Observation - From the correlation matrix, we found that GDP (Economy), health, and social support (Family) were the biggest factors impacting Happiness Rank.

Region-wise Box Plots - 2019



Observation -

1. Latin America and Caribbean region has the highest variation in Happiness rank whereas Australia and New Zealand has the least variation.
2. Happiness rank is highest in Sub-Saharan Africa and lowest in Australia and New Zealand Region.

Trends Discovered From EDA:

Conclusions drawn from the EDA-

1. From 2015-2019: Although Economy/GDP has been the biggest factor for happiness rank, the importance of health and social support has varied over the years.
2. The Happiness rank of Australia and New Zealand has reduced from 2015-2019 i.e. the government must have taken positive steps in the direction to improve happiness Index.
3. The Happiness rank of Latin America and Caribbean region has increased from 2015-2019 i.e. this means that either their economy is not doing well or people don't feel they are healthy or they don't have social support.

Inference Modelling with Linear Regression

Linear Regression is a great way to find inferences of the different features used in modelling.

In [9]:
```
from sklearn.linear_model import LinearRegression
X = df.iloc[:,2:].values
y=df.iloc[:,1].values
linearmodel = LinearRegression()
linearmodel.fit(X,y)
#To retrieve the intercept:
print(linearmodel.intercept_)
#For retrieving the slope:
print(linearmodel.coef_)
coefficients = list(linearmodel.coef_)
```

```
2.1761565086618972
[1.13958779 0.64450235 1.00934031 1.47859991 0.86345112 0.59165257]
```

The coefficients of the different factors obtained using linear regression are:

- Economy : 1.1395877901894251
- Family : 0.644502352531811
- Health: 1.009340313690316
- Freedom: 1.4785999099314577
- Trust: 0.8634511188181094
- Generosity: 0.5916525703015013

This shows that Freedom and Economy are the top two factors that affect the happiness index.

Predictive Modelling

Regression Trees

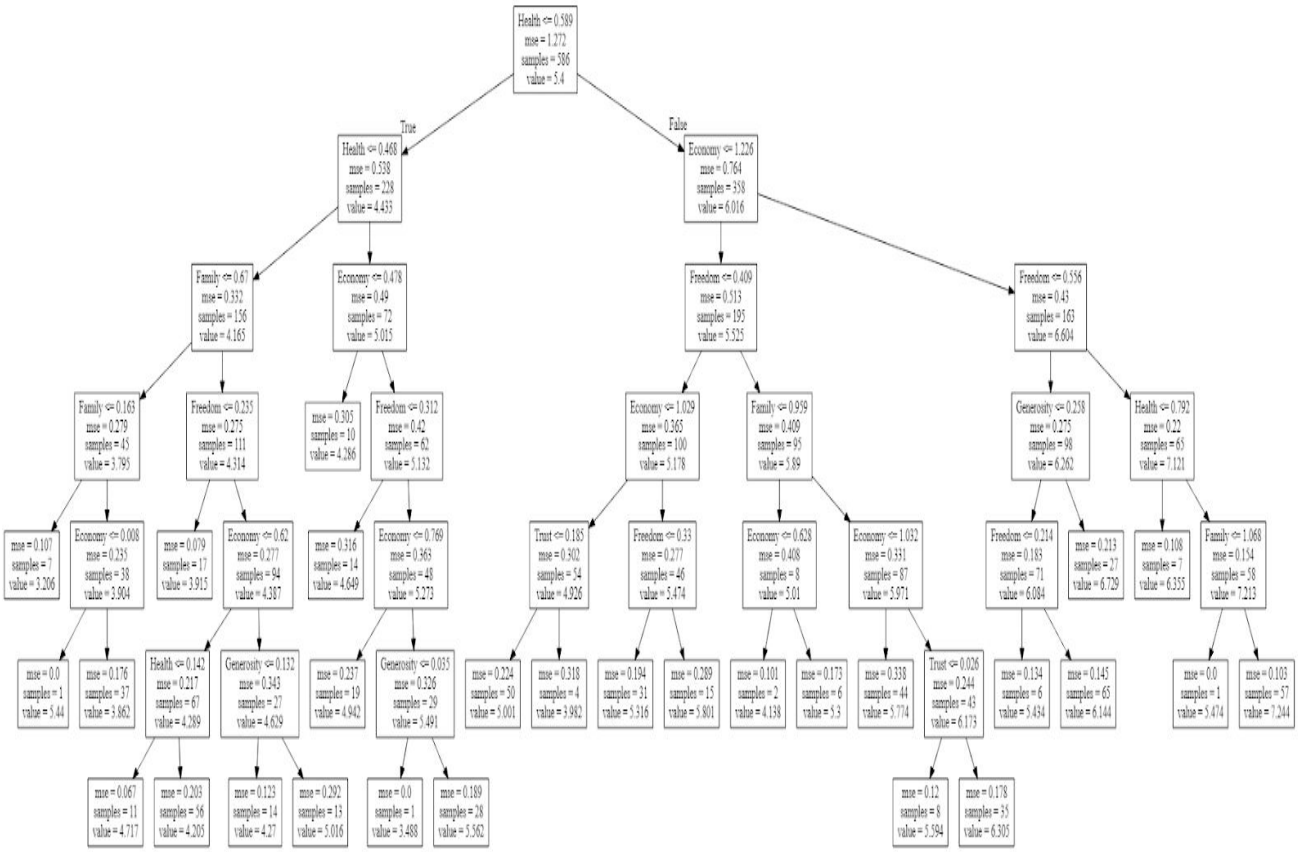
Regression trees are one of the simplest predictive modeling techniques with very good results.

We have run multiple regression tree models by pruning the maximum leaf nodes through a range of values with Mean Squared Error as the metric to determine the best model.

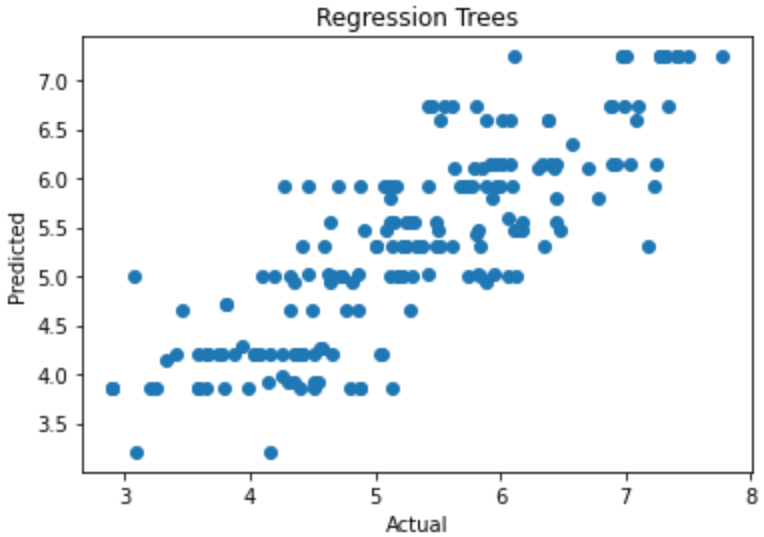
In [12]:
```
#Regression Trees
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
leaf_nodes = [5,6,7,8,9,10,12,14,16,18,20,22,24,26,28,30]
regressor = DecisionTreeRegressor()
MSE_decisiontrees = []
for n in leaf_nodes:
    regressor.set_params(max_leaf_nodes = n)
    regressor.fit(X_train, y_train)
    MSE_decisiontrees.append(mean_squared_error(y_test,regressor.predict(X_test)))

print(min(MSE_decisiontrees))
print("leaf_nodes: " + str(leaf_nodes[MSE_decisiontrees.index(min(MSE_decisiontrees))]))
```

```
0.37037858523069284
leaf_nodes: 28
```

The best model is obtained at 28 max leaf nodes and the Mean Squared Error is obtained as 0.37037858523069284.



The scatter plot for the predicted and actual values of the best tree model:



Lasso and Ridge Regression

Linear Regression is a great model for inference but the model usually overfits when the coefficients are very large i.e. for a very small change in the training data, we see a considerably big change in the regression line.

To encounter this problem some kind of penalty must be induced. This is where Lasso and Ridge regression come into play. These models make sure that the model is not overfit and thus help in making a better predictive model.

We have tried a range of different alpha values to tune the model and find the best model possible.

In [31]:
```
#Lasso
from sklearn.linear_model import Lasso

lasso = Lasso()

parameters = [1e-15, 1e-10, 1e-8, 1e-4, 1e-3,1e-2, 1, 5, 10, 20]
MSE_lasso = []

for a in parameters:
    lasso.set_params(alpha = a)
    lasso.fit(X_train, y_train)
    MSE_lasso.append(mean_squared_error(y_test,lasso.predict(X_test)))

print(min(MSE_lasso))
print("alpha: " + str(parameters[MSE_lasso.index(min(MSE_lasso))]))
```

```
0.31337499447089145
alpha: 0.01
```

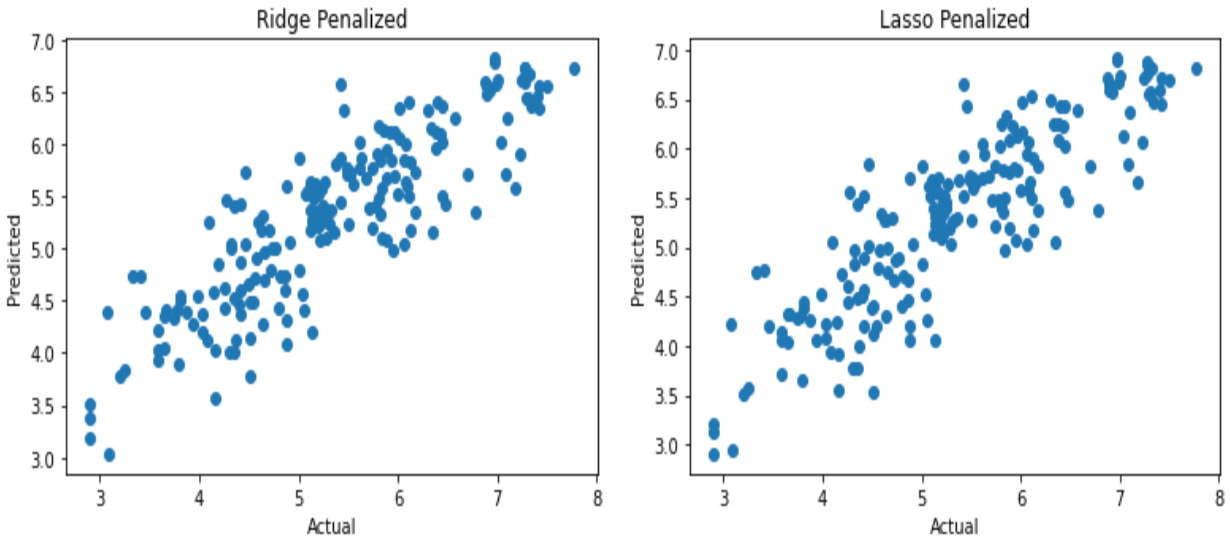
In [32]:
```
#Ridge
from sklearn.linear_model import Ridge
ridge = Ridge()

parameters = [1e-15, 1e-10, 1e-8, 1e-4, 1e-3,1e-2, 1, 5, 10, 20]
MSE_ridge = []

for a in parameters:
    ridge.set_params(alpha = a)
    ridge.fit(X_train, y_train)
    MSE_ridge.append(mean_squared_error(y_test,ridge.predict(X_test)))

print(min(MSE_ridge))
print("alpha: " + str(parameters[MSE_ridge.index(min(MSE_ridge))]))
```

```
0.3166900272196124
alpha: 5
```

The best parameters are as follows:

Parameters/Model	LASSO	RIDGE
Alpha	0.01	5
Mean Squared Error	0.31337499447089145	0.3166900272196124

The scatter plots of the predicted and actual values for the best models is:



Random Forest

Random Forest is an ensemble method which combines the results from multiple decision trees to provide the best output. Also, Random Forest is a great method to determine Feature Importance. It tells what is the impact of each predictor on the target variable.

One more important feature of the Random Forest model is that it avoids overfitting by randomly selecting features to choose from, reducing bias and increasing variance.

```
[ ] #Importing the Emsemble method
    from sklearn.ensemble import RandomForestRegressor
    regMod = RandomForestRegressor(max_depth=3, random_state=10)
    regMod.fit(X_train, y_train)
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=3, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=None, oob_score=False,
                      random_state=10, verbose=0, warm_start=False)
```

```
[ ] #Let's look at the feature importance
    for feature, importance in zip(df.columns[2:], regMod.feature_importances_):
      print('The feature importance of ' + str(feature) + ' is ' + str(importance))
```

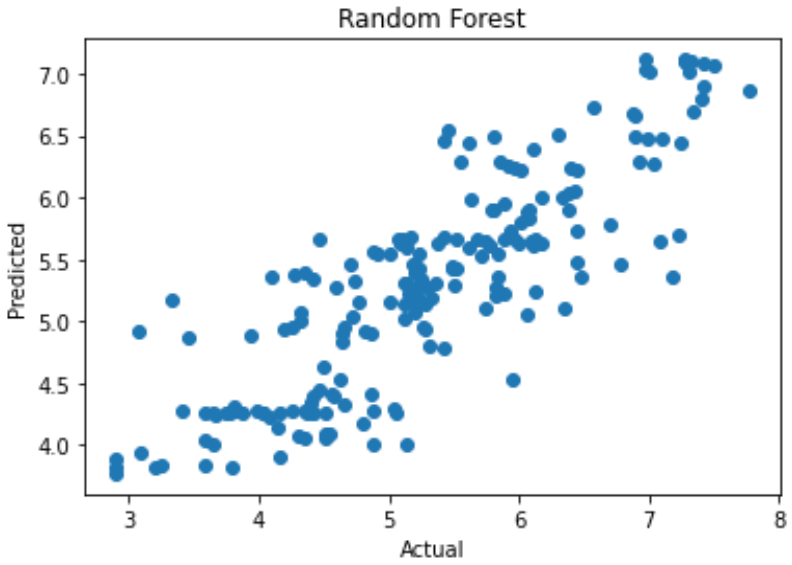
```
The feature importance of Economy is 0.4761396972273258
The feature importance of Family is 0.038917718570730286
The feature importance of Health is 0.374537397787535
The feature importance of Freedom is 0.07973262674129458
The feature importance of Trust is 0.010432875046151577
The feature importance of Generosity is 0.02023968462696275
```

```
[ ] mse = metrics.mean_squared_error(y_test, regMod.predict(X_test))
    print(mse)
```

```
0.35206705584154196
```

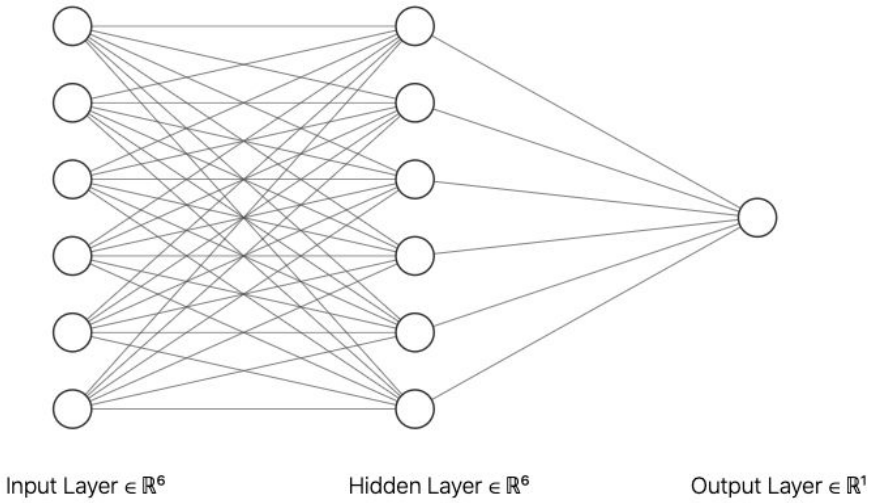
The MSE from this model is obtained as 0.35206.

The scatter plots of the predicted and actual values for the random forest model is:



Deep Neural Net

Deep Learning methods mimic actual learning with an architecture of neurons, connected with each other. Even though these models are like a black box and are really hard to interpret, they provide very accurate predictions.

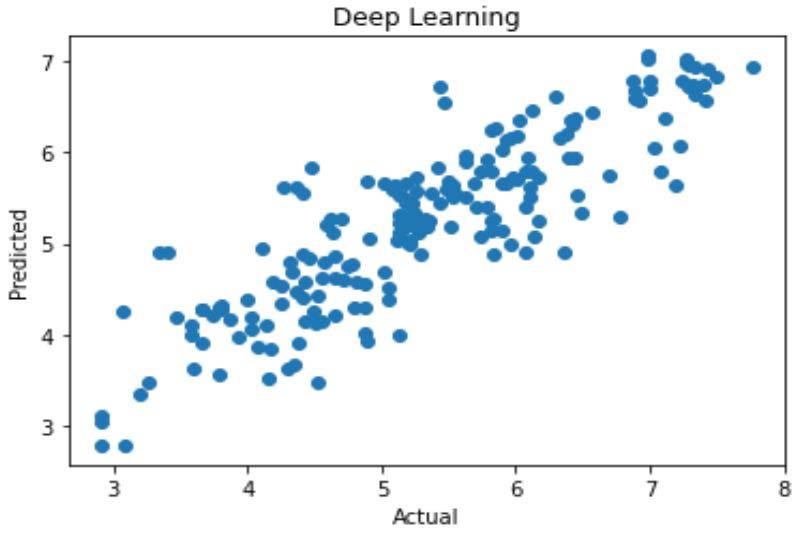


```
[ ] #Doing the Deep Neural Net
    from tensorflow import keras
```

```
[ ] #let's initiliaze a model with 10 dene layers, the output layer has to be with a linear activation
    model = tf.keras.Sequential()
    model.add(tf.keras.layers.Dense(5, input_shape = [6]))
    model.add(tf.keras.layers.Dense(1, activation='linear'))
    model.compile(optimizer='sgd', loss = 'mean_squared_error')
    model.fit(X_train, y_train, epochs=100)
```

```
Epoch 1/100
19/19 [==============================] - 0s 1ms/step - loss: 2.1959
Epoch 2/100
19/19 [==============================] - 0s 983us/step - loss: 0.4961
Epoch 3/100
19/19 [==============================] - 0s 978us/step - loss: 0.4719
Epoch 4/100
19/19 [==============================] - 0s 969us/step - loss: 0.4197
Epoch 5/100
19/19 [==============================] - 0s 970us/step - loss: 0.3943
Epoch 6/100
```

The MSE from this model is obtained as 0.2985
The scatter plots of the predicted and actual values for the random forest model is:



Model Conclusions

1. All the models provide a good fit to the data and are really good at predicting.

2. As expected the Deep Learning approach is the best for predicting.

3. One can never rely on one model, it is always a mix of various models to get the best out of each one of those.

Business Value and Relevance

The relevance of our analysis is most to recent times. COVID19 is a pandemic which has shaken the whole world. Every country is finding it difficult to contain the spread of the virus. People have to stay quarantined in their homes, maintain social distancing. Jobs are lost, businesses are not functioning, restaurants are empty.

There is a lot the governments and the organizations are doing to prevent it from spreading further and containing it soon by developing a vaccine for the same. The effort is good but there but there will be a lot which will have to be done even after this is over.

Issues

Reduced Freedom of Making Choices

People have been advised to stay inside their homes. People are not being able to go out, go to offices, colleges and schools. This has led to a decrease in the sense of freedom of the people of that country. Which eventually leads in the decrease in the happiness index.

Declining Trust in Government

No one in the world had any experience dealing with such a pandemic which is why none of the countries were able to handle this in a very good way. There has been a huge loss of lives and financial loss which is why the trust of people in the government has declined.

Plummeting GDP

As the rate of economic transactions in the world has declined, there has been a decrease in the GDP of every county which also affects the happiness index in a major way.

Fall in Life Expectancy

There has been a huge number of fatalities due to the corona virus which has caused an expected decrease in the life expectancy rate.

Things To Be Focused Upon:

Social Relationships

It will be crucial to maintain our social relationships with others during this economic downturn to not lose our identity held before COVID-19. Social relationships play a major role in maintaining happiness levels in people's lives.

Natural Happiness

Environments need to be accessible and sustainable for its people. Governments will have to promote its citizens to go out and enjoy nature and surroundings so that they can regain the natural happiness they have lost due to being advised to stay inside the houses.

Generosity

Communities will have to work together to build better lives for one another. The more social connection a country or city has, the lesser amount of inequality there will be.

Health Spending

The health spending must occur regardless of how much room in the budget a country may have. Low-income countries urgently need grants or zero-interest loans to finance the health spending they might not otherwise be able to afford.

Expanding Social Safety Nets

Governments should protect people from the economic impact of this global health crisis. A family-operated restaurant in a tourism-reliant country, or the employees of a factory shut down because of a local quarantine will need support to weather the crisis. Providing tax relief for people and businesses who can't afford to pay can be of the ways the government can help decrease the burden of the citizens.

It is also important to communicate to the public how emergency action and changes to original budgets are compatible with stability and sustainability.