A blog post about World Happiness, by Sharad Yadav

Photo credit- hubspot.net

1. **Problem Definition.** – The World Happiness Report is a landmark survey of the state of global happiness. Since 2013 the United Nations has celebrated the International Day of Happiness as it got recognition across the world that the government, society uses the happiness index to reform their policy.

   Happiness score decides the ranking of a country's happiness rank i.e. measure on social, economy (GDP per capita income) family, Health(life expectancy),Trust (government corruption),generosity, Dystopia Residual.

2. **Data Analysis.** – In this study I am going to explore data of 2017 world happiness insight with the help of sklearn's library.

```
1  # lets start with importing necessary library
2  import numpy as np
3  import pandas as pd
4  from sklearn.preprocessing import StandardScaler
5  from sklearn.linear_model import LinearRegression
6  from statsmodels.stats.outliers_influence import variance_inflation_factor
7  from sklearn.model_selection import train_test_split
8  from sklearn.model_selection import train_test_split,GridSearchCV
9  from sklearn.metrics import accuracy_score,confusion_matrix ,roc_auc_score,roc_curve,r2_score
10 import statsmodels as sm
11 import matplotlib.pyplot as plt
12 import seaborn as sns
13 import pickle
14 import warnings
15 warnings.filterwarnings("ignore")
```

```
1  #read the file and see the column  so 1. it observe top 5 country have socre more than 7
2  df=pd.read_csv(r"C:\Users\INPshy\Desktop\DATA Science\happiness_score_dataset.csv")
3  df.head()
```

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201
2	Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.49204
3	Norway	Western Europe	4	7.522	0.03880	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	2.46531

Image: - screenshot of sklearn's library

You can find the code on git hub
https://github.com/sharadyadav1988/practice-project/blob/main/World%20Happines%20project%20(1).ipynb

The Happiness Score is driven by six factors — economic production, social support, life expectancy, freedom, absence of corruption, generosity. Each of these factors contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has scores world's lowest national averages for each of the six factors.
Summary of six factors:-

1. GDP per capita:GDP per capita is a measure of a country's economic output that accounts per person of the country.

2. Social support: Social support means having friends and other people, including family, who stand beside you in times of crises helping each other.

3. Healthy life expectancy: Healthy Life Expectancy is the average Of life that any person can expect to live without any illness

4. Freedom: Freedom of choice describes an individual's ability to choose whatever he/she wants to do.

5. Generosity: the quality of being kind and generous liberal view.

6. Perceptions of corruption: The Corruption Perceptions Index (CPI) is an index published annually by Transparency International since 1995 which ranks countries "by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys.

EDA:-

As per 2017 report top 5 country is Switzerland, Iceland, Denmark, Norway, and Canada based on six parameters which we are discussing about

Out[2]:

	Country	Region	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Switzerland	Western Europe	1	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	Iceland	Western Europe	2	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201
2	Denmark	Western Europe	3	7.527	0.03328	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	2.49204
3	Norway	Western Europe	4	7.522	0.03880	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	2.46531
4	Canada	North America	5	7.427	0.03553	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	2.45176

```
3]:  1 df.describe()  #our target variable is happines score so focus should be there
```

Out[3]:

	Happiness Rank	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
count	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000
mean	79.493671	5.375734	0.047885	0.846137	0.991046	0.630259	0.428615	0.143422	0.237296	2.098977
std	45.754363	1.145010	0.017146	0.403121	0.272369	0.247078	0.150693	0.120034	0.126685	0.553550
min	1.000000	2.839000	0.018480	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.328580
25%	40.250000	4.526000	0.037268	0.545808	0.856823	0.439185	0.328330	0.061675	0.150553	1.759410
50%	79.500000	5.232500	0.043940	0.910245	1.029510	0.696705	0.435515	0.107220	0.216130	2.095415
75%	118.750000	6.243750	0.052300	1.158448	1.214405	0.811013	0.549092	0.180255	0.309883	2.462415
max	158.000000	7.587000	0.136930	1.690420	1.402230	1.025250	0.669730	0.551910	0.795880	3.602140

Image:-screenshot of dataset describe

If we talk about the whole world happiness rank region wise Australia, New Zealand, North America, Western and Eastern Europe Latin America followed by Eastern Asia region
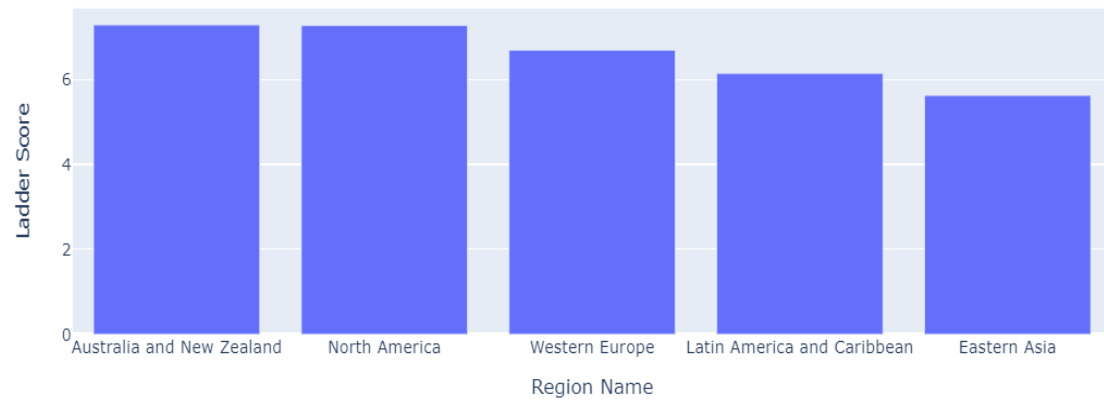
Image:-Ladder Score Region wise

here below scatter plot with respect to happiness score major factor is Economy (GDP per capita income) , Family ,health (life expectancy),Freedom .as per scatter plot these factor shows positive linear relationship means if GPD ,Freedom, Family health (life expectancy) increasing happiness score will also increase in the same proportion .

Trust(Government corruption) and Generosity slightly less relationship with happiness score.
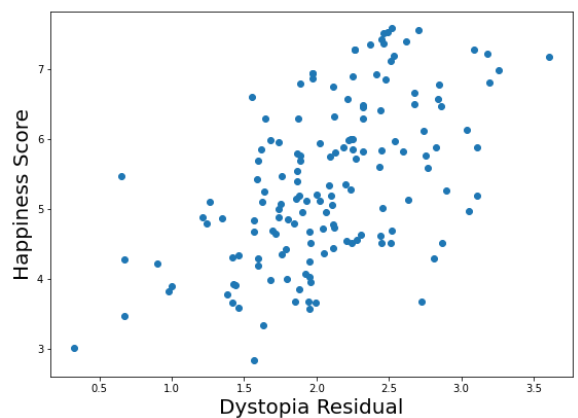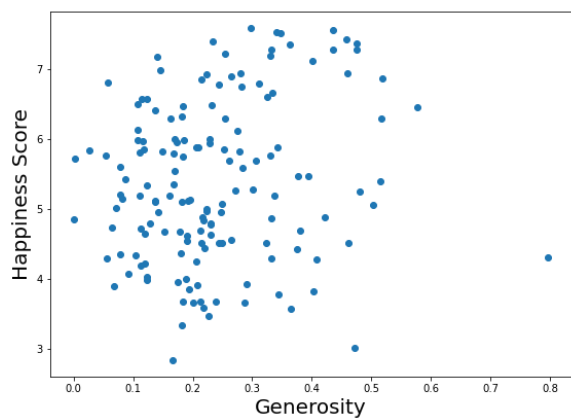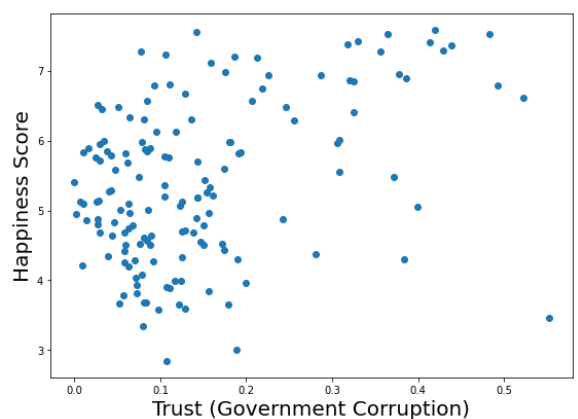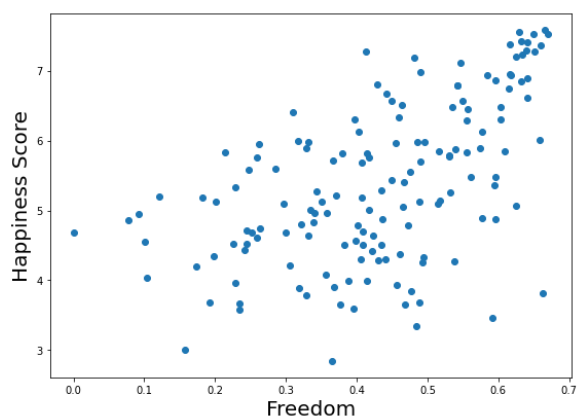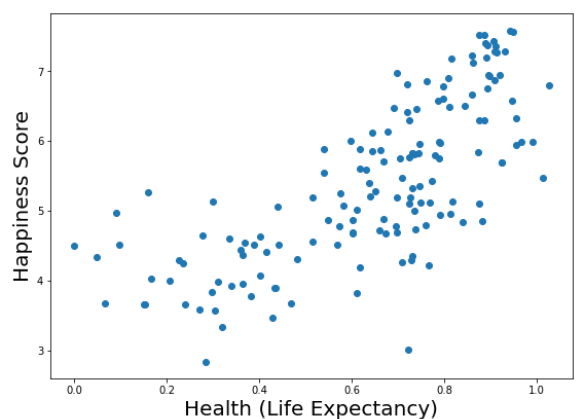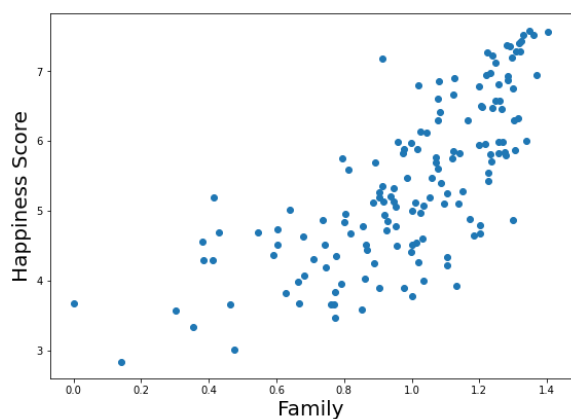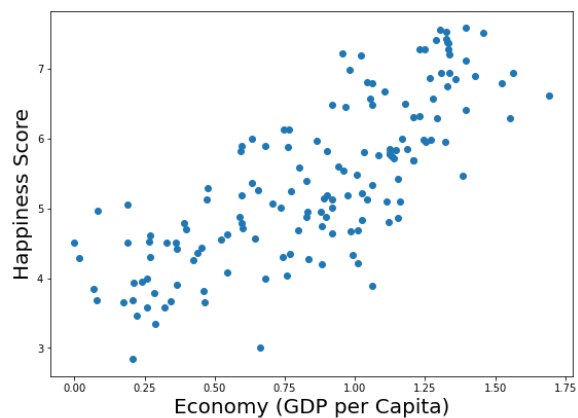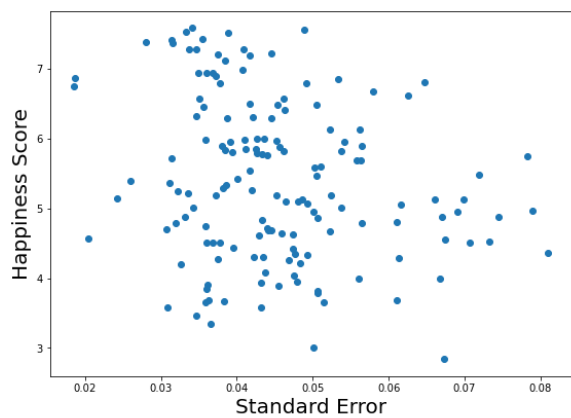
Image: - Scatter plot of features data

Moving forward if we plot heatmap to identify exact correlation with dependent variable happiness score, here in heatmap lighter orange. shed i.e.(GDP,Family ,Health,)reflect strong relationships while Freedom Generosity don't show such strong positive relationships.

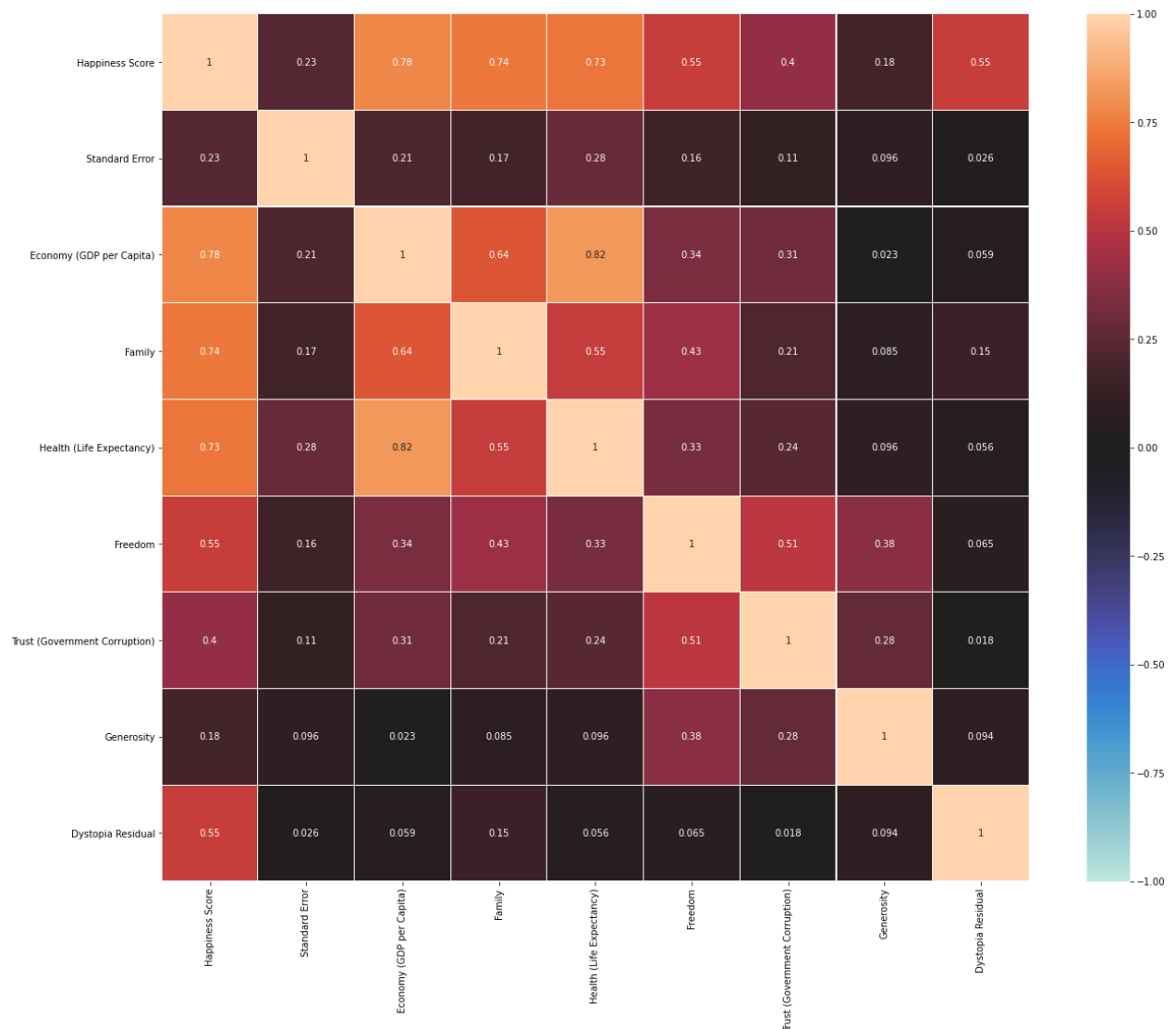|	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
Happiness Score	1	0.23	0.78	0.74	0.73	0.55	0.4	0.18	0.55
Standard Error	0.23	1	0.21	0.17	0.28	0.16	0.11	0.096	0.026
Economy (GDP per Capita)	0.78	0.21	1	0.64	0.82	0.34	0.31	0.023	0.059
Family	0.74	0.17	0.64	1	0.55	0.43	0.21	0.085	0.15
Health (Life Expectancy)	0.73	0.28	0.82	0.55	1	0.33	0.24	0.096	0.056
Freedom	0.55	0.16	0.34	0.43	0.33	1	0.51	0.38	0.065
Trust (Government Corruption)	0.4	0.11	0.31	0.21	0.24	0.51	1	0.28	0.018
Generosity	0.18	0.096	0.023	0.085	0.096	0.38	0.28	1	0.094
Dystopia Residual	0.55	0.026	0.059	0.15	0.056	0.065	0.018	0.094	1

Image: - Heatmap to check features correlation

Moving forward after understanding relationship between features and label now we separate features and label, features scaling & train test split to build model
Building Machin Learnings Models:-

```
1  #divide data into features and label
2  Y=data_cleaned['Happiness Score']
3  X=data_cleaned.drop(['Happiness Score'],axis=1)
```

Screenshot of features selection

After that we import the linear regression model from the sklearn library to train data and check the model score and how the model performs after training.

```
1  #split data in train and test,and build the model
2  x_train , x_test, y_train, y_test=train_test_split(X_scaled,Y,test_size=.25,random_state=300)
```

```
1  from sklearn.linear_model import LinearRegression
2
3  lm=LinearRegression()
4  lm.fit(x_train,y_train)  #model is ready .fit behind the scene will act like gradient descet
```

2]: LinearRegression()

```
1  Happy.head(2)
```

3]:

	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201

Screenshot of model building process

With Linear regression, I have matched predicted values with actual values. Further build different models and do hyperparameter tuning to improvise model performance but Linear Regression performing best model among all.

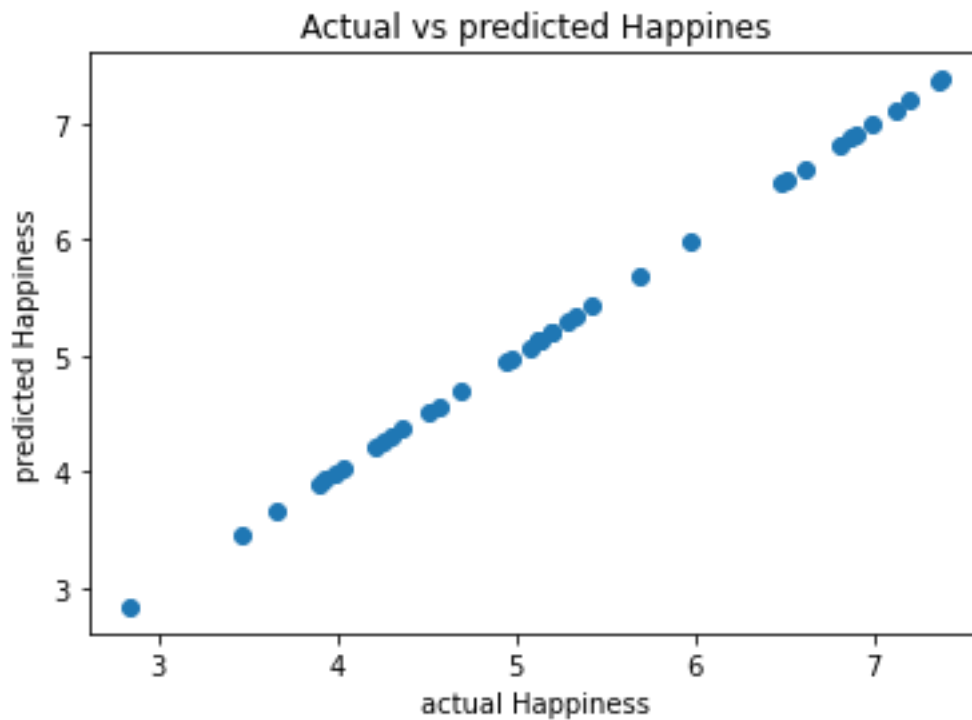Below graph elaborate how model works actually actual vs predicted

Image: -*Scatterplot*

I have also performed cross validation to make sure that my model is not over-fitted and cross validation score confirms that model is accurate, for cross checking Linear regression model I have fitted first rank features data then result what I get is really match with actual data.

```
3  lm=LinearRegression()
4  lm.fit(x_train,y_train)   #model is ready .fit behind the scene will act like gradient descet
```

Out[152]: LinearRegression()

```
In [153]:  ▶  1  Happy.head(2)
```

Out[153]:

	Happiness Score	Standard Error	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	7.587	0.03411	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	2.51738
1	7.561	0.04884	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	2.70201

```
In [154]:  ▶  1  #lets check our our model is how much accurate with actual data
              2  print('Happiness Score:' ,lm.predict(scaler.transform([[.03411,1.39651,1.34951,0.94143,0.66557,0.41978,0.29678,2.51738]]
```

Happiness Score: [7.58690282]

```
In [155]:  ▶  1  #adjusted R2 score
              2  lm.score(x_train,y_train)      #model score is to high we need to check overfitting further
```

Out[155]: 0.9999999438697404

```
In [156]:  ▶  1  #accuracy of Linear regression r2 score
              2  r2_score(y_pred,y_test)
```

Out[156]: 0.9999999382519397

```
In [157]:  ▶  1  #cross validation score
              2  cross_val_score(lm,X_scaled,Y,cv=5).mean()
```

Out[157]: 0.9999983314864249

Screenshot of model building & score

Conclusions: –

The model is calculating happiness score based on six major features economic production, social support, life expectancy, freedom, absence of corruption, generosity as if now, but it may add some few features in coming years and trained model accordingly. Of course there are different tools and approach to do analysis and trained model in better way.

Thanks for reading of my blog on 2017 world happiness index data analysis.