WORLD HAPPINESS REPORT

STATISTICAL ANALYSIS

(MSE PROJECT)

Submitted By,

Radha Rathore

C21008

PGPDS 2021

INTRODUCTION

The happiness scores and rankings use data from the Gallup World Poll. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy at birth, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country.

While a happiness report or rank might seem trivial, it is this type of data that points to our development policies and the people's perception of how their governments perform.

The report also helps in pointing out the importance of qualitative development rather than quantitative. It takes into consideration what people really think about topics such as women's rights, corruption rates, fundamental rights and more. In a way, it is one of the most important reports for a developing economy like India.

DATASET DESCRIPTION

1. **Ladder score**: It is the score for happiness and livability. The high ladder score represents the best possible life for you and low ladder score represents the worst possible life for you.

2. **Logged GDP per capita**: this translates to a measure of national wealth since GDP market value per person also readily serves as a prosperity measure.

3. **Social support**: It is defined in terms of social network characteristics such as assistance from family, friends, neighbors and other community members.

4. **Healthy life expectancy at birth**: How long, on average, a newborn can expect to live in a healthy state, if current death rates do not change

5. **Freedom to make life choices**: Freedom of choice describes an individual's opportunity and autonomy to perform an action selected from at least two available options, unconstrained by external parties.

6. **Generosity**: Generosity is the virtue of being liberal in giving, often as gifts.

7. **Perceptions of corruption:** Corruption is a form of dishonesty or criminal offense undertaken by a person or organization entrusted with a position of authority

8. **Regional Indicator**

9. **Country**

10. **Year**

HOW THE DATA LOOKS LIKE

	Country name	year	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption	Regional indicator
0	Afghanistan	2008	3.724	7.370	0.451	50.8	0.718	0.168	0.882	South Asia
1	Afghanistan	2009	4.402	7.540	0.552	51.2	0.679	0.190	0.850	South Asia
2	Afghanistan	2010	4.758	7.647	0.539	51.6	0.600	0.121	0.707	South Asia

EDA

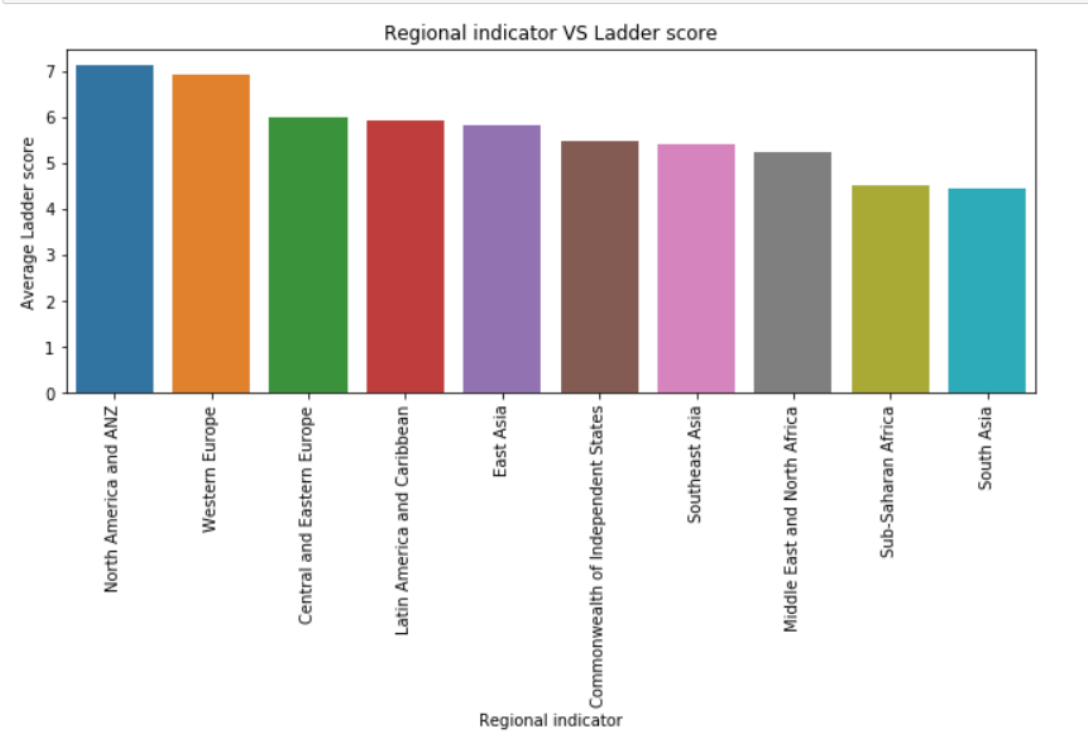
Non Graphical:

General idea of various scores for all the countries for 2021:

	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy at birth	Freedom to make life choices	Generosity	Perceptions of corruption
count	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000
mean	5.532839	9.432208	0.814745	64.992799	0.791597	-0.015134	0.727450
std	1.073924	1.158601	0.114889	6.762043	0.113332	0.150657	0.179226
min	2.523000	6.635000	0.463000	48.478000	0.382000	-0.288000	0.082000
25%	4.852000	8.541000	0.750000	59.802000	0.718000	-0.126000	0.667000
50%	5.534000	9.569000	0.832000	66.603000	0.804000	-0.036000	0.781000
75%	6.255000	10.421000	0.905000	69.600000	0.877000	0.079000	0.845000
max	7.842000	11.647000	0.983000	76.953000	0.970000	0.542000	0.939000

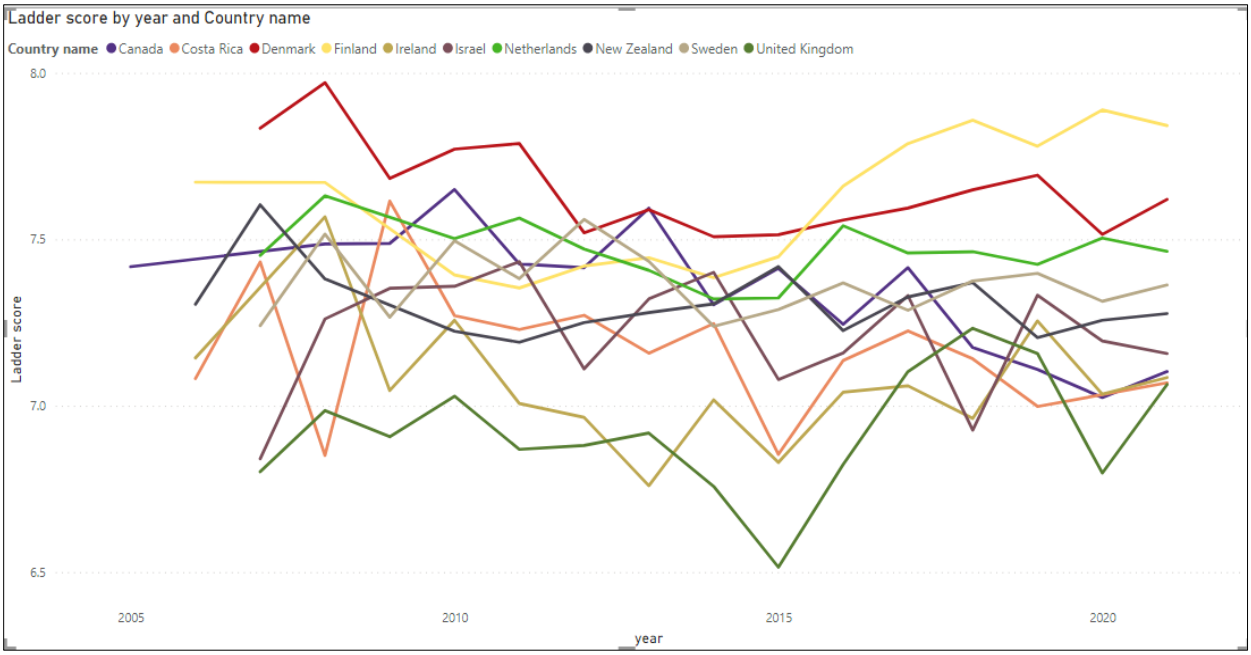
GRAPHICAL :

1. **Average Ladder score for 2021 given region wise**



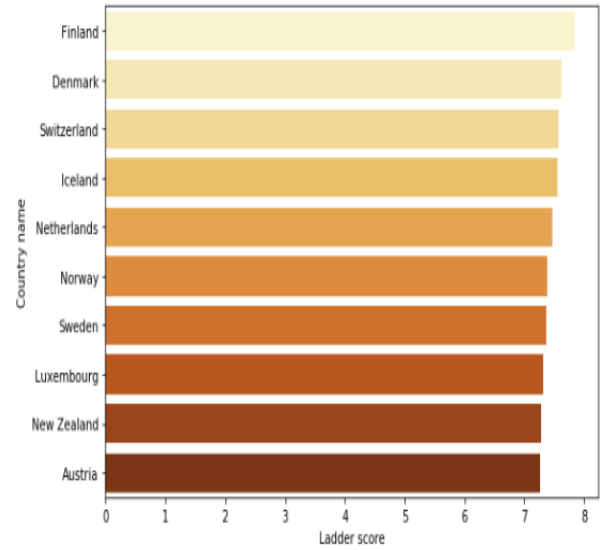
South Asia and Sub-Saharan Africa have the lower mean ladder score of happiness while North America and ANZ region have the highest score.

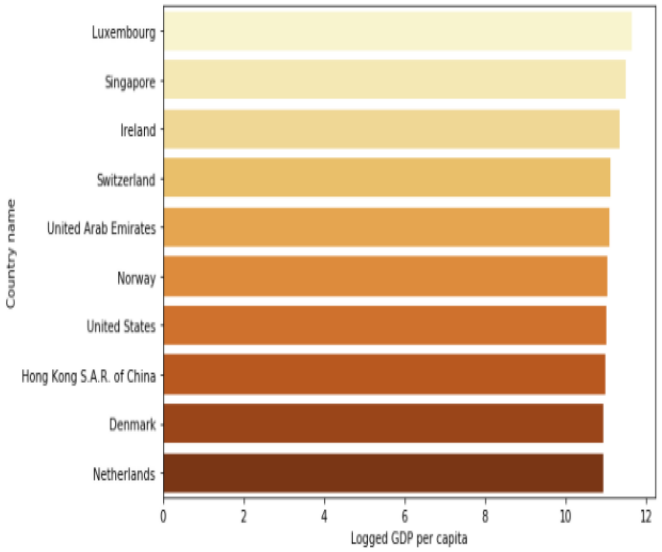
2. Ladder score of top 10 countries over the years

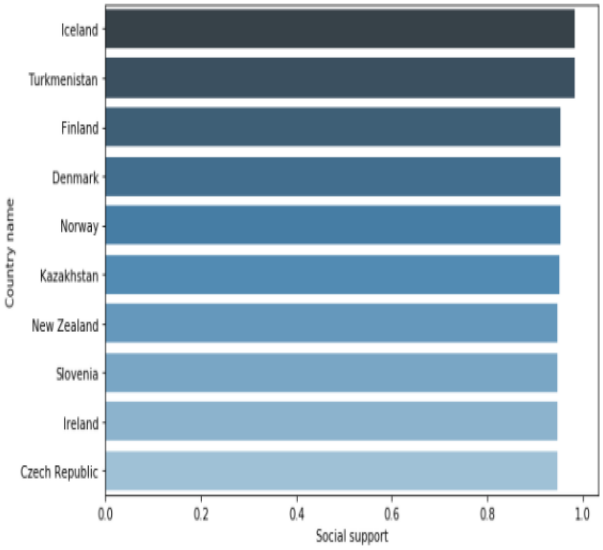


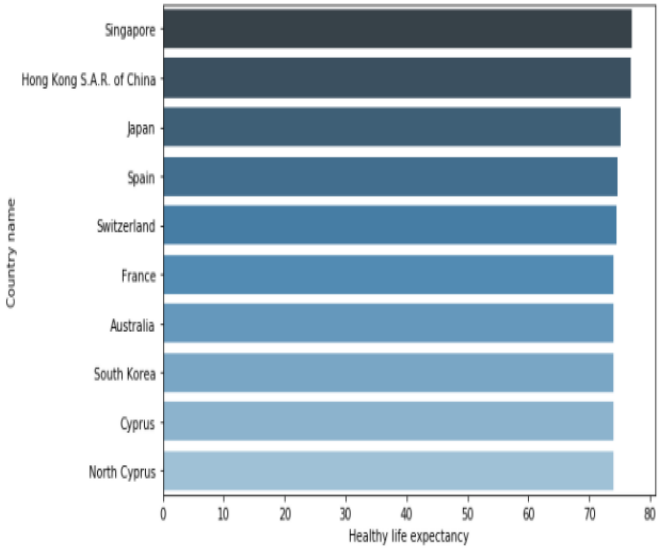
Trend of Top 10 countries by ladder score through the years

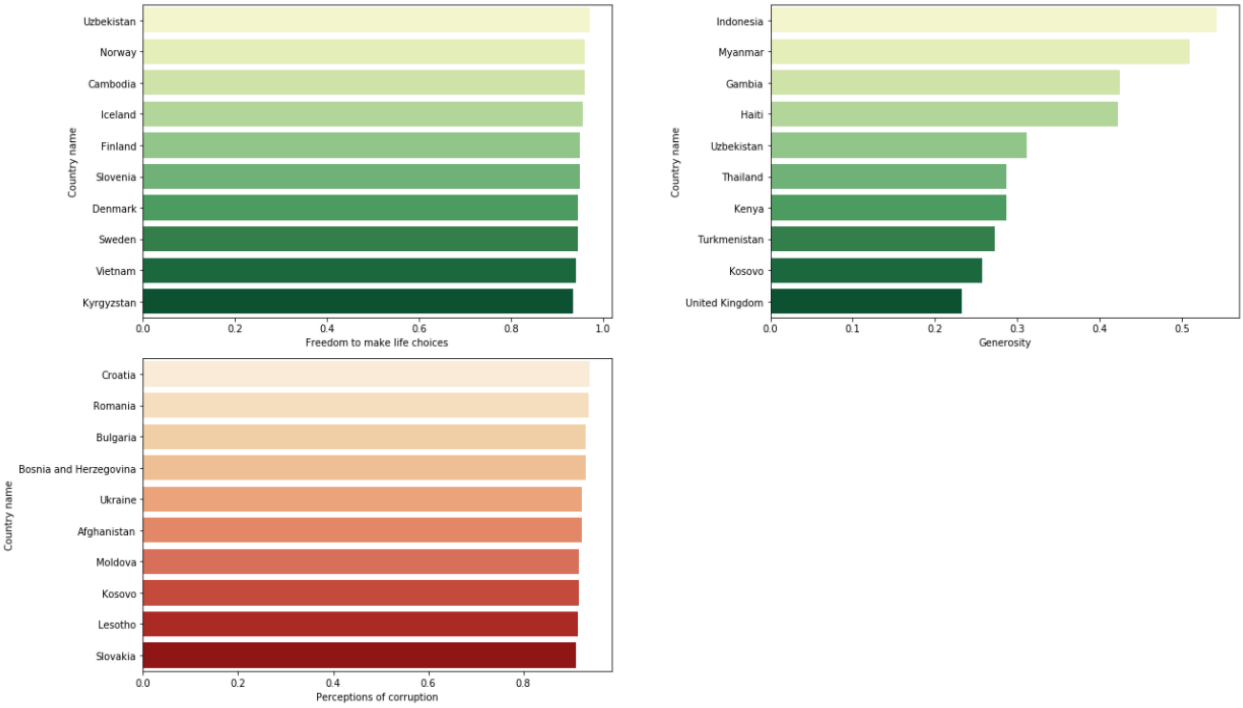
3. Top 10 countries in 2021 by various parameters



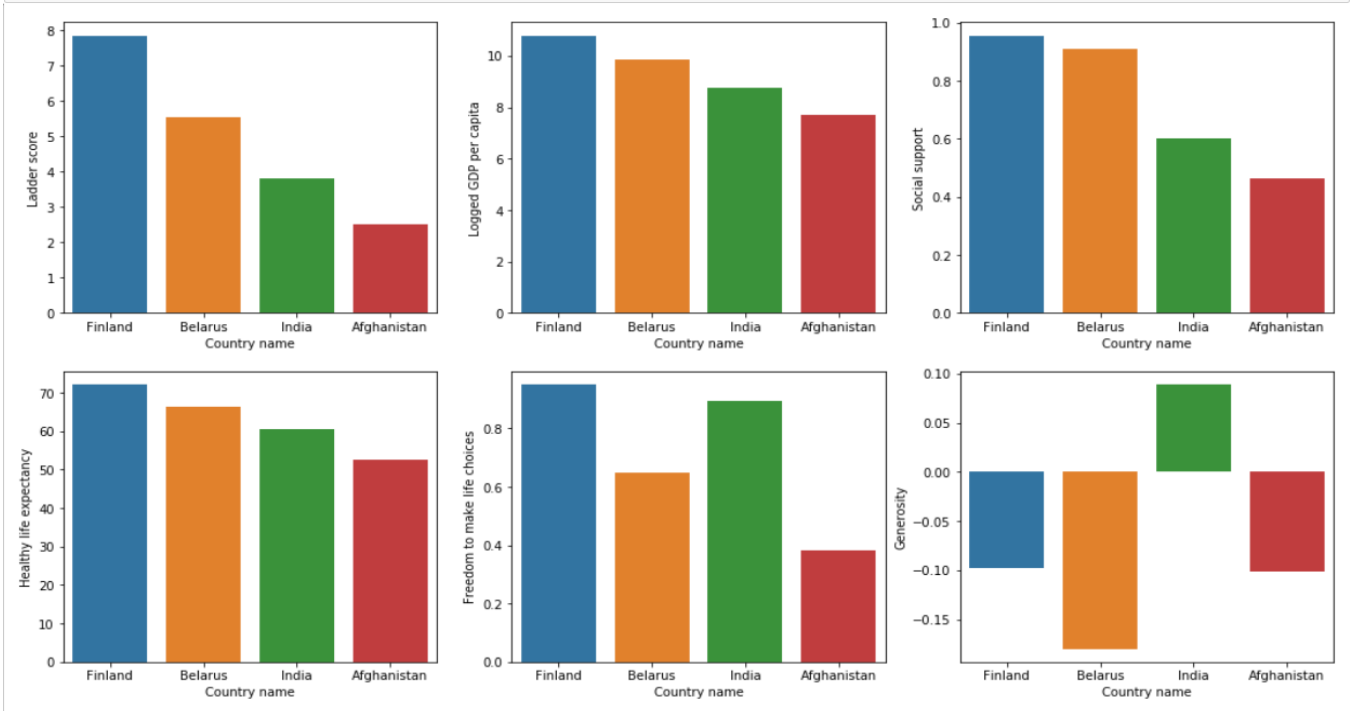








4. **Comparison of India in 2021 for different factors with country having highest, lowest and mean ladder score**



5. **Heatmap to show correlation between different variables**

HYPOTHESIS TESTING

Dividing GDP into 2 categories based on median value: High (>9.46000) and Low (<9.46000) and then conducting hypothesis test for freedom to make life choices and life expectancy at birth by dividing these data into 2 samples filtered on each GDP category

1. Life expectancy at birth when compared for samples with GDP <9.46 and >9.46 rejects null hypothesis. It means Life expectancy at birth varies significantly for countries with lower GDP than higher.

Python Result:

```
3.113993871580549e-271
reject null hypothesis
```

2. Freedom to make life choices when compared for samples with GDP <9.46 and >9.46 rejects null hypothesis. It means Freedom to make life choices vary significantly for countries with lower GDP than higher.

Python Result:

```
5.301739118068363e-34
reject null hypothesis
```

Dividing Life Expectancy at birth into 2 categories based on median value: High (>65.20) and Low (<65.20) and then conducting hypothesis test for social support

1. Social support when compared for samples with Healthy life expectancy at birth < 65.2 and >65.2 rejects null hypothesis. It means Social support vary significantly for countries with lower Healthy life expectancy at birth than higher values.

Python Result:

```
0.00031300883180986034
reject null hypothesis
```

ANOVA

Conducted ANOVA test by diving ladder score into 3 groups : <=3, 3-6, >6

1. **Social Support**

   Social support scores for countries falling in different ladder score groups vary significantly

```
p-value for significance is:  2.8664989236501846e-100
reject null hypothesis
```

2. **Perceptions of corruption**

Perceptions of corruption for countries falling in different ladder score groups vary significantly.

```
p-value for significance is:  3.1073465341901075e-39
reject null hypothesis
```

CHI - SQUARE

- Comparison of 3 groups of ladder score with the different regions that countries fall in.
- null hypothesis is that both variables are independent and alternate hypothesis is that both variables are not independent

Result:

p-value: 0.0439137502423842
Reject H0, there is a relation between the 2 categories.

REGRESSION ANALYSIS

SIMPLE LINEAR REGRESSION:

1. **Simple linear regression taking GDP as response and social support as predictor to check the relation**

```
                                 OLS Regression Results
================================================================================================
Dep. Variable:     Logged GDP per capita   R-squared (uncentered):                   0.989
Model:                               OLS   Adj. R-squared (uncentered):              0.989
Method:                    Least Squares   F-statistic:                          1.663e+05
Date:                   Fri, 23 Jul 2021   Prob (F-statistic):                        0.00
Time:                           01:47:54   Log-Likelihood:                         -2619.6
No. Observations:                   1861   AIC:                                      5241.
Df Residuals:                       1860   BIC:                                      5247.
Df Model:                              1
Covariance Type:               nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Social support    11.4071      0.028    407.751      0.000      11.352      11.462
==============================================================================
Omnibus:                      100.076   Durbin-Watson:                   0.573
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              150.540
Skew:                           0.459   Prob(JB):                     2.05e-33
Kurtosis:                       4.049   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

p value of F- statistic is significant. Social support explains 98.9% of the GDP score for the countries (Adj R sq = 0.989). P-value of social support is also significant.

2. **Simple linear regression taking Freedom to make life choices as response and social support as predictor to check the relation**

OLS Regression Results

Dep. Variable:	Healthy life expectancy at birth	R-squared (uncentered):	0.986
Model:	OLS	Adj. R-squared (uncentered):	0.986
Method:	Least Squares	F-statistic:	1.304e+05
Date:	Fri, 23 Jul 2021	Prob (F-statistic):	0.00
Time:	01:49:14	Log-Likelihood:	-6407.3
No. Observations:	1861	AIC:	1.282e+04
Df Residuals:	1860	BIC:	1.282e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Social support	77.3218	0.214	361.087	0.000	76.902	77.742

Omnibus:	85.595	Durbin-Watson:	0.536
Prob(Omnibus):	0.000	Jarque-Bera (JB):	126.900
Skew:	0.411	Prob(JB):	2.78e-28
Kurtosis:	3.979	Cond. No.	1.00

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

p value of F- statistic is significant. Social support explains 98.6% variation of the Freedom to make life choices score for the countries (Adj R sq = 0.986). P-value of social support is also significant.

MULTIPLE LINEAR REGRESSION :

1. **Multiple linear regression taking Ladder score (happiness index) as response and social support, GDP per capita, Life expectancy, Freedom to make life choices, generosity and Perceptions of corruption as predictors to check the relation and make prediction. I am using the data till 2020 to predict the ladder scores of countries for 2021.**

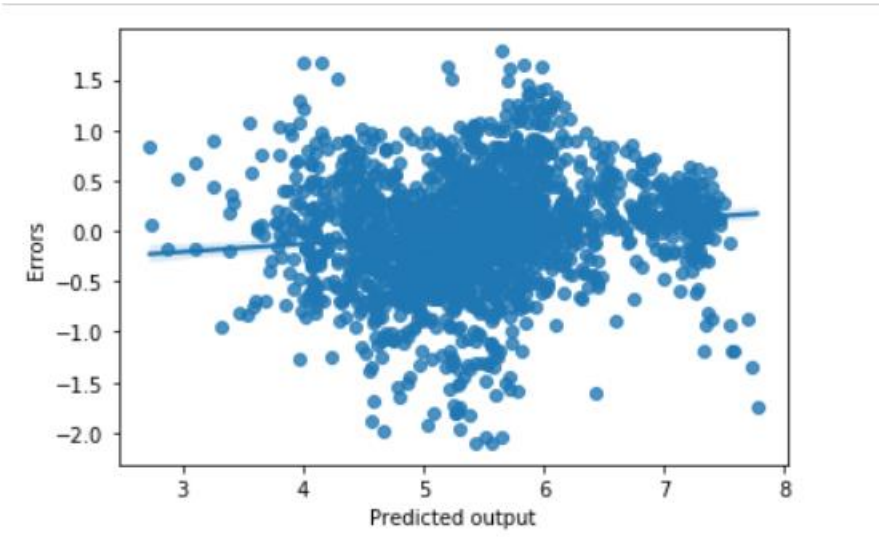
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Ladder score   R-squared (uncentered):                   0.989
Model:                            OLS   Adj. R-squared (uncentered):              0.989
Method:                 Least Squares   F-statistic:                          2.543e+04
Date:                Fri, 23 Jul 2021   Prob (F-statistic):                        0.00
Time:                        02:13:33   Log-Likelihood:                         -1511.0
No. Observations:                1712   AIC:                                      3034.
Df Residuals:                    1706   BIC:                                      3067.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Logged GDP per capita           0.2967      0.027     11.138      0.000       0.244       0.349
Social support                  2.3865      0.171     13.928      0.000       2.050       2.723
Healthy life expectancy at birth  0.0222      0.004      6.145      0.000       0.015       0.029
Freedom to make life choices    0.5060      0.116      4.380      0.000       0.279       0.733
Generosity                      0.6183      0.095      6.487      0.000       0.431       0.805
Perceptions of corruption      -1.3573      0.064    -21.227      0.000      -1.483      -1.232
==============================================================================
Omnibus:                       48.663   Durbin-Watson:                   0.552
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.998
Skew:                          -0.323   Prob(JB):                     2.09e-14
Kurtosis:                       3.682   Cond. No.                         803.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

p-value of F-stat is significant and based upon adjusted R^2 value, predictors are explaining 98.9% of the variation in ladder score, and each predictor is significant as well, given the p-value of t-stat for each predictor.

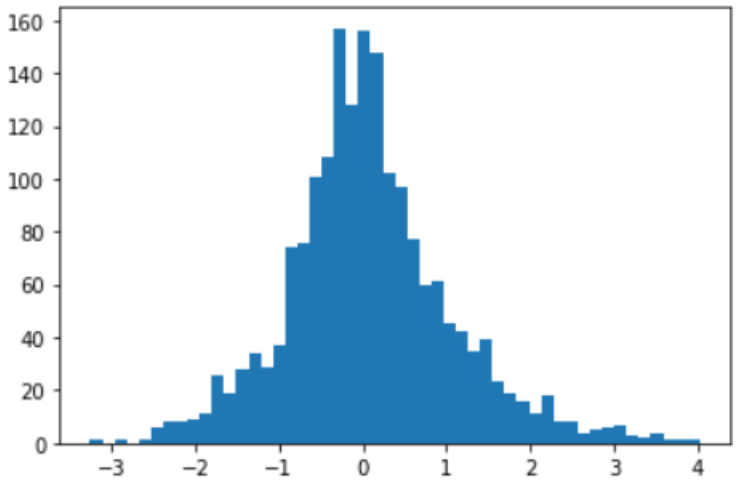
R sq and Adj R sq, both are very high, but as the data has multicollinearity, let's check.

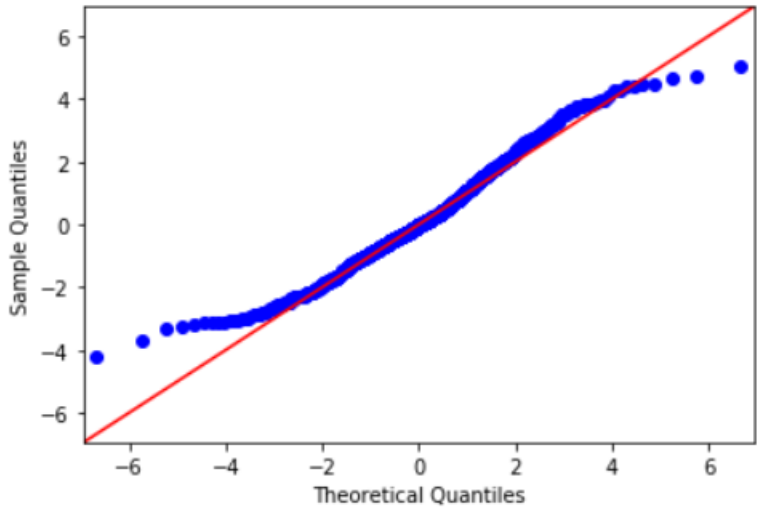
Homoscedasticity Test:



The residuals don’t have a perfect constant variance. This could be due to the presence of outliers and skewness in the data

Normality Test:



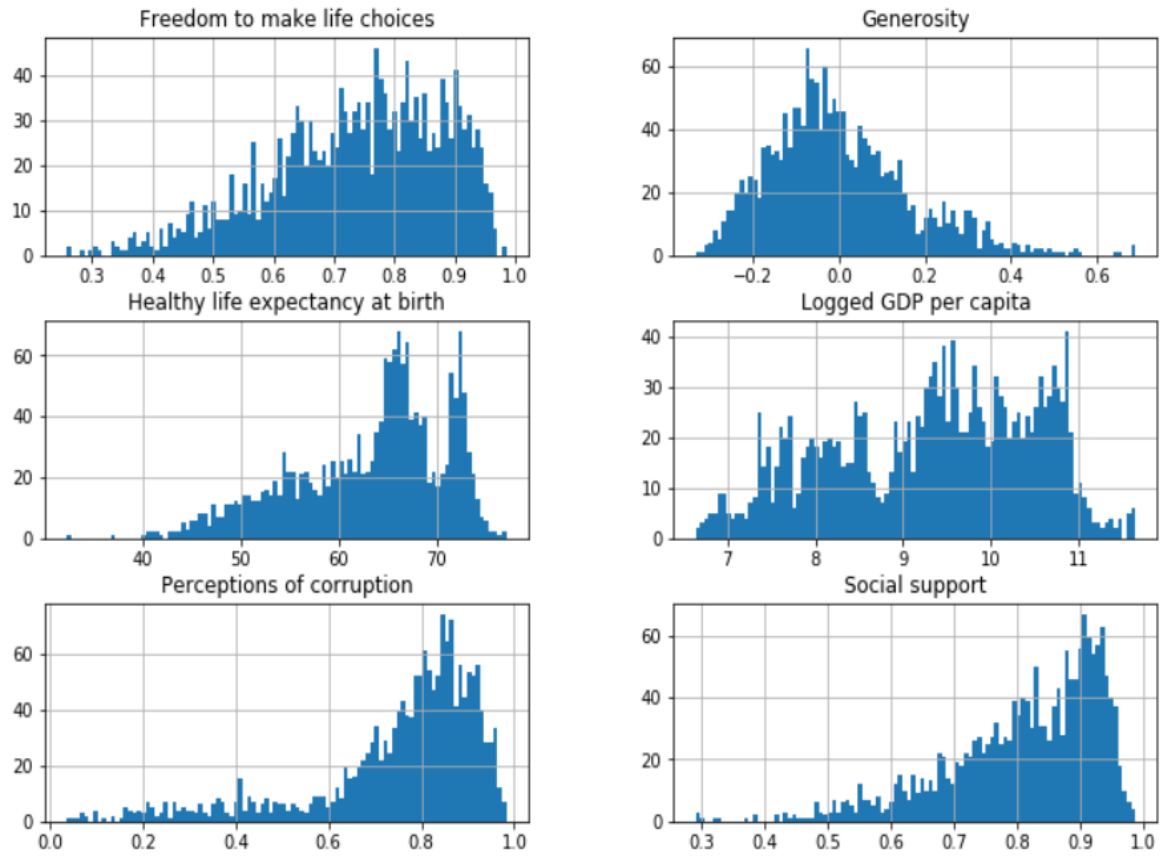


The residuals are not normally distributed as evident from the both the above plots.

Model Prediction result:

```
Baseline model result (sum(y-ybar)^2)/n : 1.1455714774109274
mean sq err = 0.3183507480968029 and rmse = 0.5642257953131911
```

Plotting histogram plot of all predictors to check the distribution:



2. Multiple Regression taking all predictors, but this time, data is scaled using Standard scalar and then regression is done on PCs

- Applied standard scalar on the data
- 3 PCS formed to handle multicollinearity and then regression done
- Explained variance ratio = 0.83 by these 3 PC

OLS Regression Results

Dep. Variable:	Ladder score	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	1703.
Date:	Fri, 23 Jul 2021	Prob (F-statistic):	0.00
Time:	02:15:25	Log-Likelihood:	-1462.1
No. Observations:	1712	AIC:	2932.
Df Residuals:	1708	BIC:	2954.
Df Model:	3		
Covariance Type:	nonrobust		

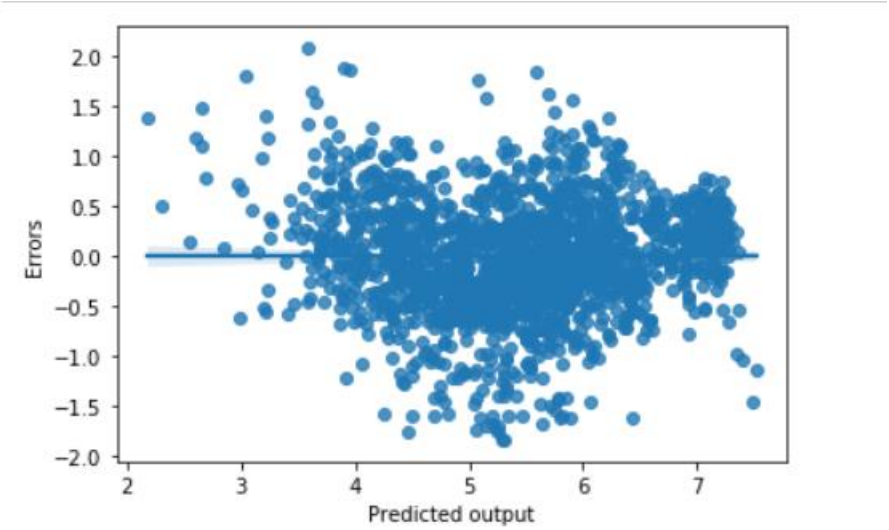
	coef	std err	t	P>|t|	[0.025	0.975]
const	5.4450	0.014	395.899	0.000	5.418	5.472
PC0	-0.5670	0.008	-71.120	0.000	-0.583	-0.551
PC1	-0.0564	0.012	-4.763	0.000	-0.080	-0.033
PC2	-0.0922	0.017	-5.450	0.000	-0.125	-0.059

Omnibus:	29.396	Durbin-Watson:	0.574
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.729
Skew:	-0.181	Prob(JB):	5.27e-10
Kurtosis:	3.684	Cond. No.	2.12

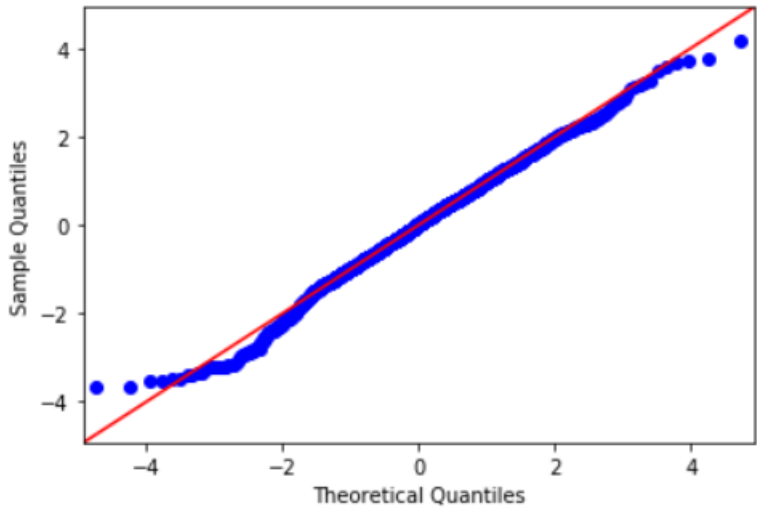
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

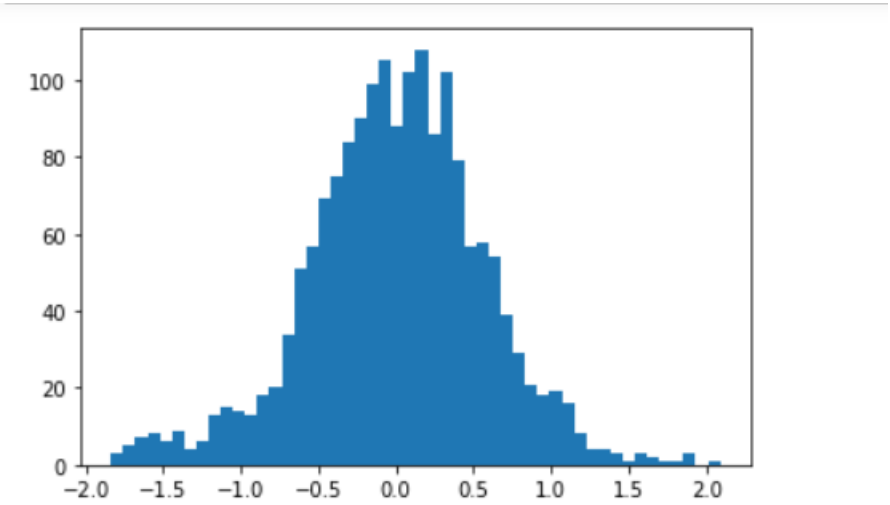
p-value of F-stat is significant and based upon adjusted R^2 value, predictors are explaining 75.0% of the variation in ladder score, and each predictor is significant as well, given the p-value of t-stat for each predictor (Principal components in this case)

Homoscedasticity Test:



Normality Test:





Model Prediction result:

```
Baseline model result (sum(y-ybar)^2)/n : 1.1455714774109274
mean sq err = 0.29393172841904014 and rmse = 0.5421547089337508
```

3. Regression taking all predictors, but this time, data is scaled using Robust scalar and then regression is done on PCs

- Taken 3 PCs
- Explained variance ratio = 0.85

About Robust Scaling:

- Since some of the features have outliers, so I will also try with robust scaling and check the results.
- Robust standardization approach is to standardizing input variables in the presence of outliers to ignore the outliers from the calculation of the mean and standard deviation, and then use the calculated values to scale the variable.
- Scaled value = (value – median) / (p75 – p25)
- The resulting variable has a zero mean and median and a standard deviation of 1, although the outliers are still present with the same relative relationships to other values.

OLS Regression Results

Dep. Variable:	Ladder score	R-squared:	0.747
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	1680.
Date:	Fri, 23 Jul 2021	Prob (F-statistic):	0.00
Time:	02:16:38	Log-Likelihood:	-1470.8
No. Observations:	1712	AIC:	2950.
Df Residuals:	1708	BIC:	2971.
Df Model:	3		
Covariance Type:	nonrobust		

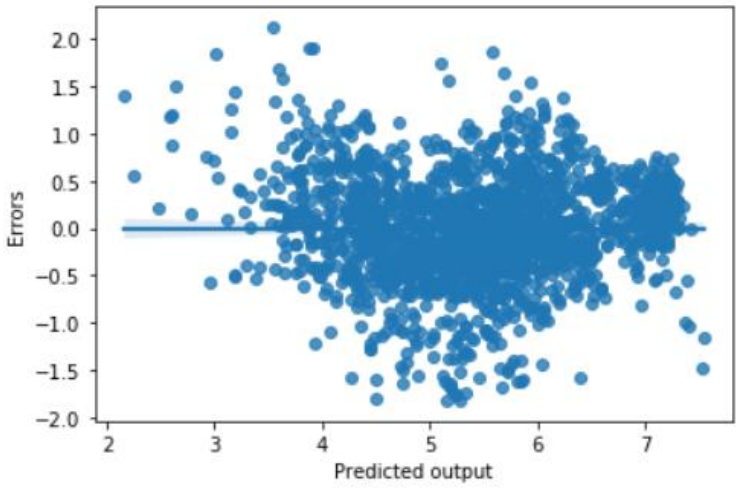
	coef	std err	t	P>|t|	[0.025	0.975]
const	5.4450	0.014	393.883	0.000	5.418	5.472
PC0	0.6771	0.010	64.668	0.000	0.657	0.698
PC1	-0.3958	0.015	-26.722	0.000	-0.425	-0.367
PC2	0.2292	0.019	12.047	0.000	0.192	0.267

Omnibus:	30.234	Durbin-Watson:	0.576
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44.940
Skew:	-0.178	Prob(JB):	1.74e-10
Kurtosis:	3.709	Cond. No.	1.82

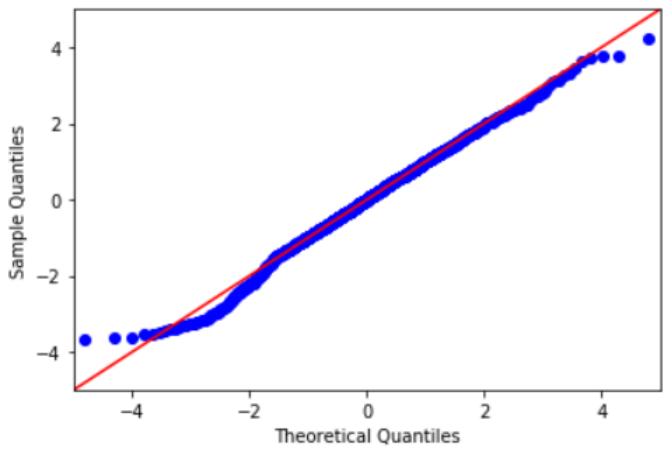
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

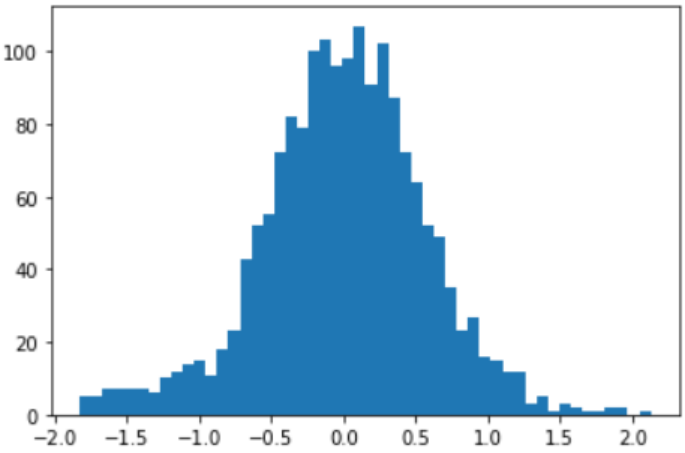
p-value of F-stat is significant and based upon adjusted R^2 value, predictors are explaining 74.6% of the variation in ladder score, and each predictor is significant as well, given the p-value of t-stat for each predictor (Principal components in this case)

Homoscedasticity Test:



Normality Test:





Model Prediction result :

```
Baseline model result (sum(y-ybar)^2)/n : 1.1455714774109274
mean sq err = 0.29581979238887274 and rmse = 0.5438931810464925
```

Comparing the mse and rmse results with that of standard scalar method, they are not that different.

CONCLUSION

- Since there is multicollinearity in the data, so I have created independent features using PCA.
- Using PCA, MSE of the model improves a little bit.
- I have attempted using 2 feature scaling methods while doing PCA: Standard scaling and Robust Scaling. Adjusted R square and MSE vary only a bit using the two methods.
- Based on the regression and statistical tests, we see that each of these features are significant in determining the happiness score of a country and are also correlated with each other. It helps to gain insights about the social and developmental aspects of the countries and how happy citizens of the countries really are and how each of these features is important in determining it.
- The correlated factors could be studied together for better understanding and creating a positive impact in the future.

References :

- Dataset and information taken from Kaggle - https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv