ISM - 6423

Group - DATA VIZARDS

PROJECT REPORT

on

WORLD HAPPINESS REPORT

Shashank Ahuja
Anirudh Bhagat
Tushar Mallya
Priya Patil
Anuja Phadtare
Himanshu Vaishnav
Aaliya Yunus

INTRODUCTION

Starting from 2012, every year United Nations Sustainable Development Solutions Network releases the World Happiness Report that states which country is ranked “happiest” in the whole world. Happiness is progressively considered an important factor and must be considered while building public policies and measuring social progress. In this report, we give consideration to the variation of happiness across countries based on 6 key factors. The *Happiness Score* was constructed by taking the data from Gallup World Poll.

The variables utilized reflect what has been extensively found in research literature imperative in explaining national-level contrasts in life assessments. Some of the vital factors such as unemployment and inequality are not considered because equivalent global information are not yet available for the countries in dataset. The factors are intended to show critical lines of correlation as opposed to casual assessments.

Goals of Analysis

We wish to do the analysis to find the following:

•Countries ranking highest in happiness from the time period 2015, 2016 and 2017
•Any significant changes in the *Happiness Score* of countries
•Major key factors contributing to the *Happiness Score*

DATA ELEMENTS

For our research and analysis, we wish to predict the happiness score using 6 key factors which are the independent variables.

DataSet

Panel Data for 156 countries for the time period 2015, 2016 and 2017.

Dependent Variable: *Happiness score*

Independent Variables:

•*Economy GDP per capita:* Purchasing power of an individual in a country which is produced by World Bank.

•*Health life expectancy:* Individual's life expectancy

•*Family:* Quality of family life of an individual

•*Trust:* Government's measure of corruption

•*Freedom:* Freedom of an individual to make life decisions.

•*Generosity***:** How generous is the population. Eg. Amount of donations made in the past month/quarter.

ANALYSIS METHODS USED

Data Analysis helps us find patterns from large amount of data. These patterns are used to make business decisions.

We used both “Supervised” and “Unsupervised” data analysis methods to find the happiness score and classify countries accordingly.

1. DESCRIPTIVE ANALYSIS

Descriptive analysis is used to simply describe what is going on in the data. For our analysis, we used STATA to do the descriptive analysis using the following commands:

a) **describe:** This command is used to describe various variables in the dataset like their name, storage type, display format, description, etc.

```
. describe

Contains data
  obs:           470
 vars:            10
 size:        27,260
```

variable name	storage type	display format	value label	variable label
country	str24	%24s		Country
happinessscore	float	%9.0g		Happiness.Score
economygdpper~a	float	%9.0g		Economy..GDP.per.Capita.
family	float	%9.0g		Family
healthlifeexp~y	float	%9.0g		Health..Life.Expectancy.
freedom	float	%9.0g		Freedom
generosity	float	%9.0g		Generosity
trustgovernme~n	float	%9.0g		Trust..Government.Corruption.
dystopiaresid~l	float	%9.0g		Dystopia.Residual
year	int	%8.0g		Year

b) **summarize:** This command is used to calculate and display a variety of univariate summary statistics of the dataset. It includes the mean, standard deviation, minimum and maximum values of the dataset.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
happinesss~e	470	5.370728	1.136998	2.693	7.587
economygdp~a	470	.9278301	.415584	0	1.870766
family	470	.9903466	.3187071	0	1.610574
healthlife~y	470	.579968	.2401608	0	1.02525
freedom	470	.4028277	.1503558	0	.66973
generosity	470	.1756054	.1319093	0	.8380752
trustgover~n	470	.2014255	.1332114	0	.81971
dystopiare~l	470	2.092717	.5657718	.32858	3.83772
year	470	2015.994	.8169067	2015	2017

For our analysis, we tried to summarize the data together first and then separately for each year (2015, 2016 and 2017)

```
. sort year

. by year: summ
```

-> year = 2015

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
happinesss~e	158	5.375734	1.14501	2.839	7.587
economygdp~a	158	.8461372	.4031208	0	1.69042
family	158	.9910459	.2723691	0	1.40223
healthlife~y	158	.6302594	.2470778	0	1.02525
freedom	158	.4286149	.1506928	0	.66973
generosity	158	.1434218	.1200341	0	.55191
trustgover~n	158	.2372955	.1266849	0	.79588
dystopiare~l	158	2.098977	.5535498	.32858	3.60214
year	158	2015	0	2015	2015

-> year = 2016

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
happinesss~e	157	5.382185	1.141674	2.905	7.526
economygdp~a	157	.9538798	.4125954	0	1.82427
family	157	.7936211	.2667057	0	1.18326
healthlife~y	157	.557619	.2293492	0	.95277
freedom	157	.3709939	.1455068	0	.60848
generosity	157	.1376238	.1110379	0	.50521
trustgover~n	157	.2426349	.1337557	0	.81971
dystopiare~l	157	2.325807	.54222	.81789	3.83772
year	157	2016	0	2016	2016

-> year = 2017

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
happinesss~e	155	5.354019	1.13123	2.693	7.537
economygdp~a	155	.9847182	.4207927	0	1.870766
family	155	1.188898	.2872629	0	1.610574
healthlife~y	155	.5513408	.2370727	0	.9494924
freedom	155	.408786	.1499973	0	.6582487
generosity	155	.2468835	.1347804	0	.8380752
trustgover~n	155	.1231202	.1016606	0	.4643078
dystopiare~l	155	1.850238	.5000284	.3779137	3.117485
year	155	2017	0	2017	2017

We observed the following important things after running the command:

- The mean of *Happiness Score* has not changed significantly over the years 2015 – 2017 (~5.370)
- There is an increase in the mean *Economy GDP* across the years
- The mean of *Generosity* has remained the same during the years 2015 – 2016 (0.143, 0.137), but increased during 2017 (0.246)
- There is a decrease in the mean *Life Expectancy* across the years

c) **correlate:** This command is used to display the correlation matrix for the given variables in the dataset. In our analysis, we tried to find the correlation among our dependent and independent variables for each year (2015, 2016 and 2017) after sorting the data year wise.

```
. by year: correlate
```

```
-> year = 2015
(country ignored because string variable)
(obs=158)
```

	happin~e	econom~a	family	health~y	freedom	genero~y	trustg~n	dystop~l	year
happinesss~e	1.0000								
economygdp~a	0.7810	1.0000							
family	0.7406	0.6453	1.0000						
healthlife~y	0.7242	0.8165	0.5311	1.0000					
freedom	0.5682	0.3703	0.4415	0.3605	1.0000				
generosity	0.3952	0.3079	0.2056	0.2483	0.4935	1.0000			
trustgover~n	0.1803	-0.0105	0.0875	0.1083	0.3739	0.2761	1.0000		
dystopiare~l	0.5305	0.0401	0.1481	0.0190	0.0628	-0.0331	-0.1013	1.0000	
year	.	.	.	.	.	.	.	.	.

```
-> year = 2016
(country ignored because string variable)
(obs=157)
```

	happin~e	econom~a	family	health~y	freedom	genero~y	trustg~n	dystop~l	year
happinesss~e	1.0000								
economygdp~a	0.7903	1.0000							
family	0.7393	0.6695	1.0000						
healthlife~y	0.7654	0.8371	0.5884	1.0000					
freedom	0.5668	0.3623	0.4502	0.3412	1.0000				
generosity	0.4020	0.2942	0.2136	0.2496	0.5021	1.0000			
trustgover~n	0.1568	-0.0255	0.0896	0.0760	0.3618	0.3059	1.0000		
dystopiare~l	0.5437	0.0686	0.1197	0.1009	0.0916	-0.0029	-0.1330	1.0000	
year	.	.	.	.	.	.	.	.	.

```
-> year = 2017
(country ignored because string variable)
(obs=155)
```

	happin~e	econom~a	family	health~y	freedom	genero~y	trustg~n	dystop~l	year
happinesss~e	1.0000								
economygdp~a	0.8125	1.0000							
family	0.7527	0.6883	1.0000						
healthlife~y	0.7820	0.8431	0.6121	1.0000					
freedom	0.5701	0.3699	0.4250	0.3498	1.0000				
generosity	0.1553	-0.0190	0.0517	0.0632	0.3161	1.0000			
trustgover~n	0.4291	0.3509	0.2318	0.2798	0.4992	0.2942	1.0000		
dystopiare~l	0.4754	0.0242	0.0705	0.0550	0.0819	-0.1166	-0.0228	1.0000	
year	.	.	.	.	.	.	.	.	.

We found out that

- There is a **positive** correlation between *Economy GDP Per Capita* and *Happiness Score*. The correlation value increases over the years (0.781 in 2015, 0.7903 in 2016 and 0.8125 in 2017)
- There is a **strong positive** correlation between *Economy GDP Per Capita* and *Health Life Expectancy*. The correlation value increases over the years (0.816 in 2015, 0.8371 in 2016, 0.8431 in 2017)

2. REGRESSION ANALYSIS

In statistical modelling, regression analysis is a method used to find out the relationship between the dependent variable and one or more independent variables(predictor variables). There are many types of regression analysis methods that are useful for different kinds of datasets, like Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Stepwise Regression, etc.

However, we identified that “Multiple Linear Regression” is the best fit for our analysis. It involves one dependent variable and two or more independent variables to explain the relationship between them.

With respect to our dataset, we have one dependent variable called Happiness Score and 6 independent variables (*Economy GDP Per Capita, Health Life Expectancy, Family, Freedom, Generosity, Trust)*. We ran the regression for each year to find out any relationships among the variables.

```
. by year, sort : regress happinessscore economygdppercapita family healthlifeexpectancy freedom generosity trustgovernmentcorrup
> tion
```

-> year = 2015

Source	SS	df	MS	Number of obs	=	158
				F(6, 151)	=	87.81
Model	159.982546	6	26.6637577	Prob > F	=	0.0000
Residual	45.8520174	151	.303655744	R-squared	=	0.7772
				Adj R-squared	=	0.7684
Total	205.834563	157	1.31104817	Root MSE	=	.55105

happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	.8606572	.2203113	3.91	0.000	.4253664	1.295948
family	1.408892	.2226751	6.33	0.000	.9689305	1.848853
healthlifeexpectancy	.9753089	.316293	3.08	0.002	.3503776	1.60024
freedom	1.333433	.3850157	3.46	0.001	.5727193	2.094146
generosity	.784538	.4365328	1.80	0.074	-.077963	1.647039
trustgovernmentcorruption	.388933	.3910029	0.99	0.321	-.3836102	1.161476
_cons	1.860185	.1904824	9.77	0.000	1.48383	2.23654

-> year = 2016

Source	SS	df	MS	Number of obs	=	157
				F(6, 150)	=	92.65
Model	160.127377	6	26.6878962	Prob > F	=	0.0000
Residual	43.2059002	150	.288039335	R-squared	=	0.7875
				Adj R-squared	=	0.7790
Total	203.333278	156	1.30341845	Root MSE	=	.53669

happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	.7214128	.2171212	3.32	0.001	.2924019	1.150424
family	1.229754	.2297496	5.35	0.000	.7757909	1.683718
healthlifeexpectancy	1.436403	.348912	4.12	0.000	.7469858	2.12582
freedom	1.513935	.3879659	3.90	0.000	.747351	2.280519
generosity	.9189269	.4647615	1.98	0.050	.0006022	1.837252
trustgovernmentcorruption	.1594942	.362106	0.44	0.660	-.5559929	.8749813
_cons	2.190294	.1583237	13.83	0.000	1.877461	2.503126

-> year = 2017

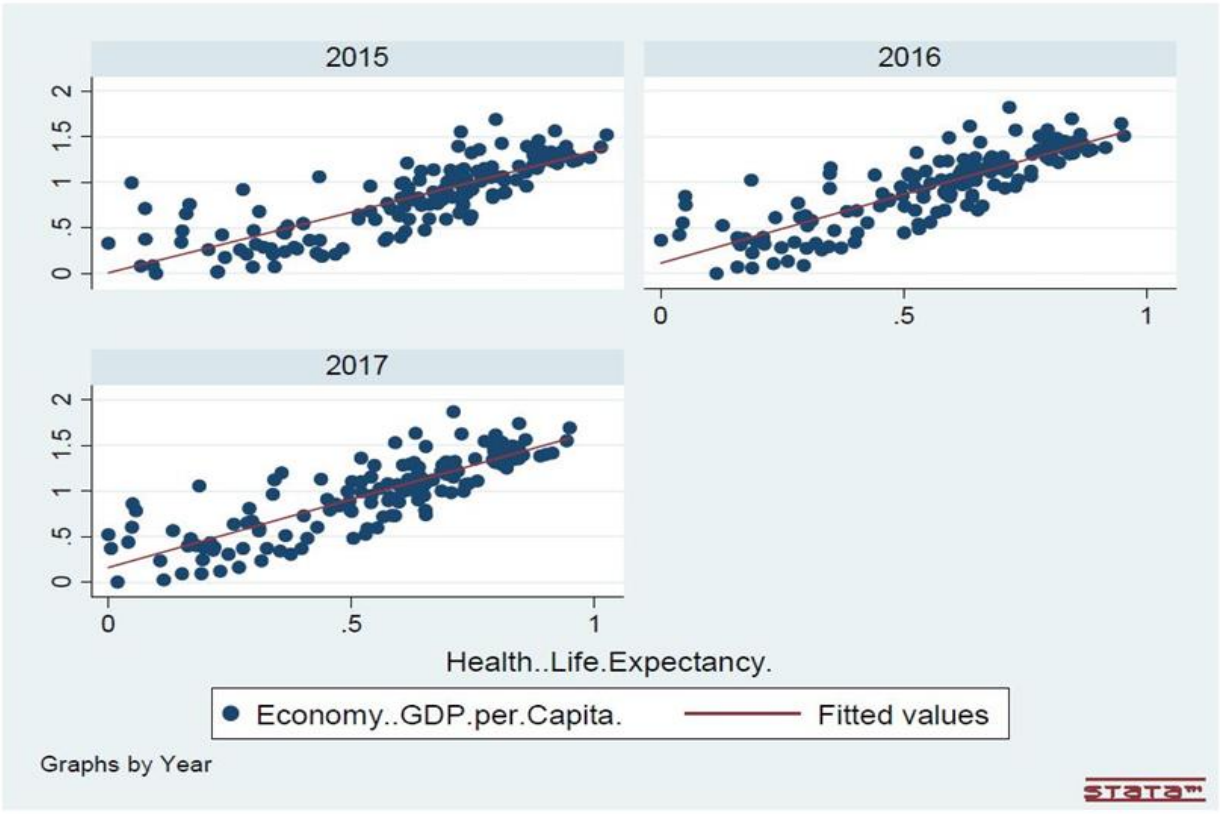
Source	SS	df	MS	Number of obs	=	155
				F(6, 148)	=	106.84
Model	160.105299	6	26.6842166	Prob > F	=	0.0000
Residual	36.9656542	148	.249767934	R-squared	=	0.8124
				Adj R-squared	=	0.8048
Total	197.070954	154	1.27968152	Root MSE	=	.49977

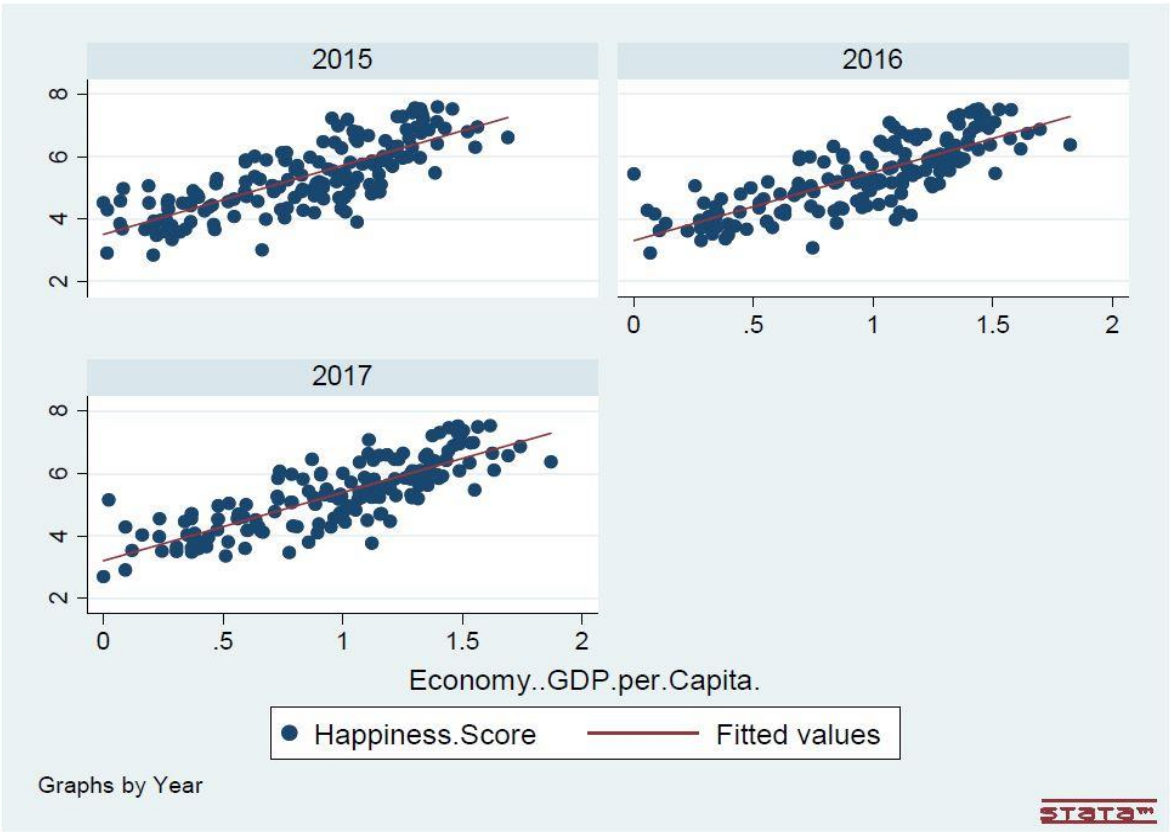
happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	.7844334	.2045131	3.84	0.000	.3802906	1.188576
family	1.117771	.2020608	5.53	0.000	.7184743	1.517068
healthlifeexpectancy	1.28888	.3215255	4.01	0.000	.6535064	1.924254
freedom	1.475715	.3425093	4.31	0.000	.798875	2.152556
generosity	.3807181	.3293271	1.16	0.250	-.2700725	1.031509
trustgovernmentcorruption	.8266072	.4843307	1.71	0.090	-.1304895	1.783704
_cons	1.743029	.1873581	9.30	0.000	1.372786	2.113271

Regression analysis for the years 2015, 2016 and 2017 indicates that

- *Economy GDP per capita*, *Life expectancy*, *Freedom* and *Family* are statistically significant as their p-value < 0.05
- *Trust* and *Generosity* are not statistically significant at 5% of significance level
- R-Squared value is increasing over the years, which indicates an improved regression model

The following scatter plot shows the relationship between *Economy GDP per capita and Health Life expectancy*.


2015
2016
2017
0
.5
1
1.5
2
Health..Life.Expectancy.
Economy..GDP.per.Capita.
Fitted values
Graphs by Year


2015
2016
2017
2
4
6
8
0
.5
1
1.5
2
Economy..GDP.per.Capita.
Happiness.Score
Fitted values
Graphs by Year

3. INTERACTION TERMS ANALYSIS

An interaction term may emerge when we are considering the relationship between three or more variables, and depicts a situation in which the synchronous impact of the two variables is not additive on the third variable.

In our project we have generated three interaction terms:

h1 = economygdppercapita * freedom

h2 = healthlifeexpectency * economygdppercapita

h3 = healthlifeexpectency * freedom

We ran regression on *Happiness Score* with these interaction terms along with other independent variables for years 2015, 2016 and 2017.

Year 2015

```
. reg happinessscore economygdppercapita healthlifeexpectancy freedom generosity trustgovernmentcorruption h1 h2 h3 if year==2015
```

Source	SS	df	MS	Number of obs	=	158
				F(8, 149)	=	53.12
Model	152.399147	8	19.0498933	Prob > F	=	0.0000
Residual	53.4354168	149	.358626959	R-squared	=	0.7404
				Adj R-squared	=	0.7265
Total	205.834563	157	1.31104817	Root MSE	=	.59885

happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	.98454	.6968732	1.41	0.160	-.3924905	2.361571
healthlifeexpectancy	-1.354282	1.12342	-1.21	0.230	-3.574174	.8656111
freedom	-1.032424	1.119214	-0.92	0.358	-3.244006	1.179157
generosity	-.130515	.5032204	-0.26	0.796	-1.124885	.8638552
trustgovernmentcorruption	.1021149	.4402238	0.23	0.817	-.7677731	.9720029
h1	-.0074452	1.511625	-0.00	0.996	-2.994437	2.979546
h2	.5649026	.6365827	0.89	0.376	-.6929931	1.822798
h3	4.994368	2.46526	2.03	0.045	.1229819	9.865754
_cons	4.07341	.4638415	8.78	0.000	3.156853	4.989967

Year 2016

```
. reg happinessscore economygdppercapita healthlifeexpectancy freedom generosity trustgovernmentcorruption h1 h2 h3 if year==2016
```

Source	SS	df	MS	Number of obs	=	157
				F(8, 148)	=	57.61
Model	153.909045	8	19.2386306	Prob > F	=	0.0000
Residual	49.4242328	148	.333947519	R-squared	=	0.7569
				Adj R-squared	=	0.7438
Total	203.333278	156	1.30341845	Root MSE	=	.57788

happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	1.292714	.6069257	2.13	0.035	.0933549	2.492074
healthlifeexpectancy	-.4940061	1.03782	-0.48	0.635	-2.544865	1.556853
freedom	1.071634	.9000735	1.19	0.236	-.7070218	2.850289
generosity	.2044947	.5489954	0.37	0.710	-.8803875	1.289377
trustgovernmentcorruption	.0738442	.4019837	0.18	0.855	-.7205249	.8682132
h1	-1.395405	1.447493	-0.96	0.337	-4.255829	1.465018
h2	.601734	.6521568	0.92	0.358	-.6870076	1.890476
h3	4.557398	2.414788	1.89	0.061	-.2145186	9.329316
_cons	3.143106	.3748775	8.38	0.000	2.402303	3.88391

Year 2017

```
. reg happinessscore economygdppercapita healthlifeexpectancy freedom generosity trustgovernmentcorruption h1 h2 h3 if year==2017
```

Source	SS	df	MS	Number of obs	=	155
				F(8, 146)	=	65.05
Model	153.894867	8	19.2368584	Prob > F	=	0.0000
Residual	43.1760867	146	.295726621	R-squared	=	0.7809
				Adj R-squared	=	0.7689
Total	197.070954	154	1.27968152	Root MSE	=	.54381

happinessscore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
economygdppercapita	1.406069	.6045923	2.33	0.021	.2111851	2.600952
healthlifeexpectancy	-.5004242	1.05872	-0.47	0.637	-2.592821	1.591973
freedom	1.337966	.8013355	1.67	0.097	-.2457501	2.921682
generosity	.2846342	.3693756	0.77	0.442	-.4453796	1.014648
trustgovernmentcorruption	.0299654	.5868814	0.05	0.959	-1.129915	1.189846
h1	-1.236014	1.349368	-0.92	0.361	-3.902831	1.430803
h2	.5724603	.5760518	0.99	0.322	-.5660172	1.710938
h3	3.703614	2.234086	1.66	0.100	-.7117117	8.11894
_cons	2.911523	.3562431	8.17	0.000	2.207463	3.615582

After running the regression, we noticed that

- the interaction term h3 (p-value<0.05) had significant impact on *Happiness Score* only in the year 2015.

- the interaction terms h1 and h2 do not have a significant impact across all the three years as the p-value is greater than 0.05.

4. K-MEANS CLUSTERING

K- Means clustering is a type of unsupervised learning. The goal of the clustering is to find similar groups which have not been explicitly labeled in the data. In K-Means clustering, number of groups is represented by K, which is defined by user before running K-Means clustering algorithm.

So, for our K-Means Clustering, we used SAS JMP for the analysis. The inputs for our K-Means clustering are:

Number of clusters: 3

Principal components: EconomyGDPperCapita, HealthLifeExpentancy, and Freedom

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	3	-2.6269	Optimal CCC

Columns Scaled Individually

Control Panel

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	187	19	0
2	141		
3	142		

Cluster Means

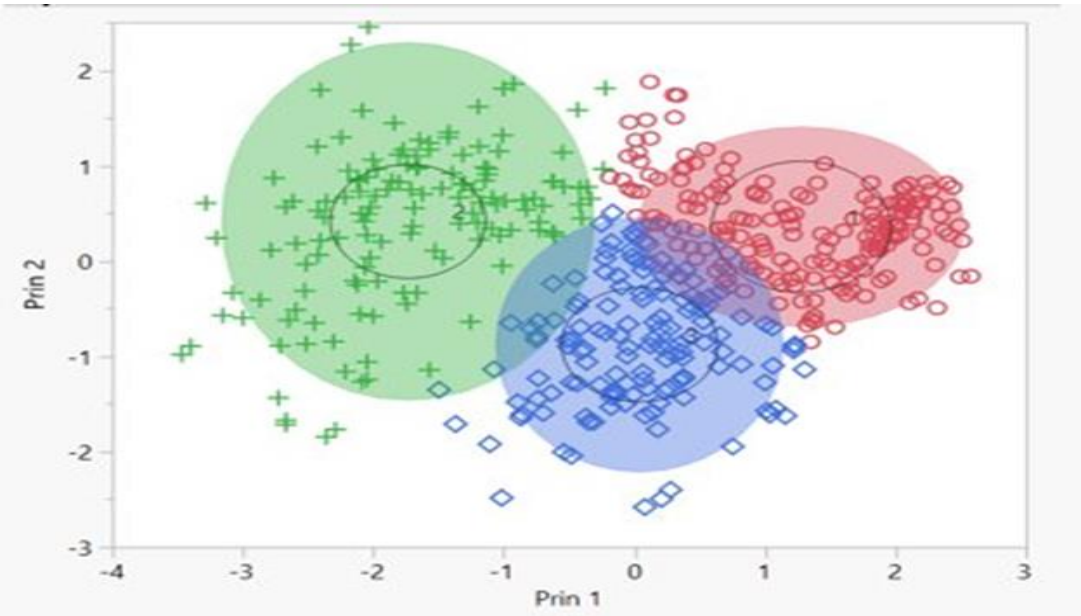
Cluster	EconomyGdpPerCapita	HealthLifeExpectancy	Freedom
1	1.22187621	0.74838465	0.5345452
2	0.42476167	0.27893588	0.34553481
3	1.04012596	0.6570922	0.2862582

In the above figure,

•**Cluster 1** is performing best in EconomyGDPperCapita, HealthLifeExpentancy, and Freedom

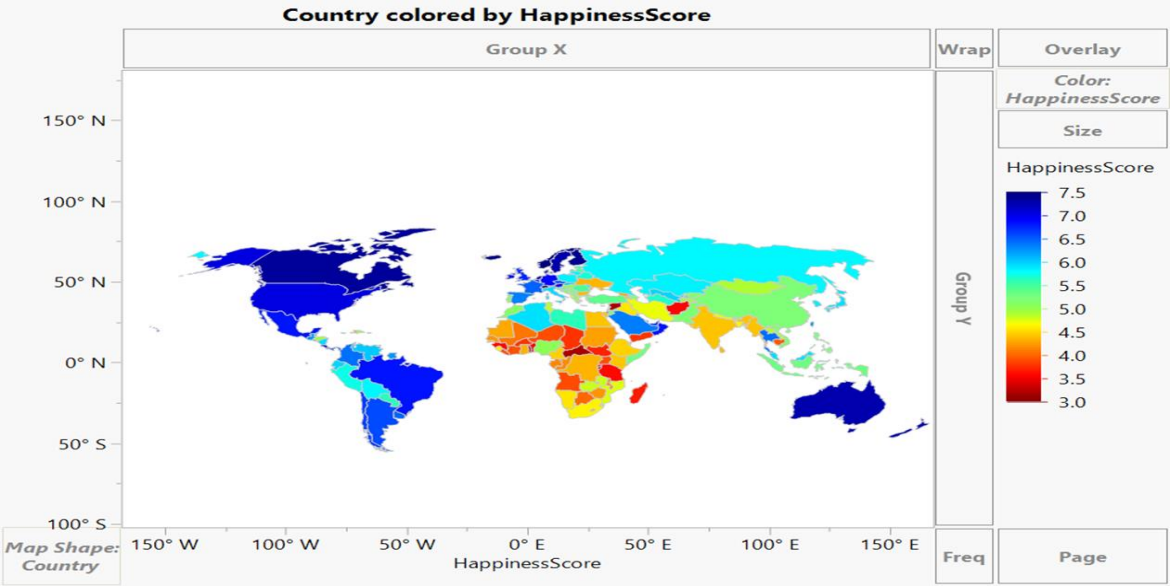
•**Cluster 3** is performing moderately in EconomyGDPperCapita and HealthLifeExpentancy but performing poor in freedom compared to all

•**Cluster 2** is performing poor in EconomyGDPperCapita and HealthLifeExpectancy but for Freedom, it is performing better than cluster 3



Above figure is the Biplot (2D) of the K-Means Clustering (K=3). Where color red group is cluster 1, color green group is cluster 2, and color blue group is cluster 3.

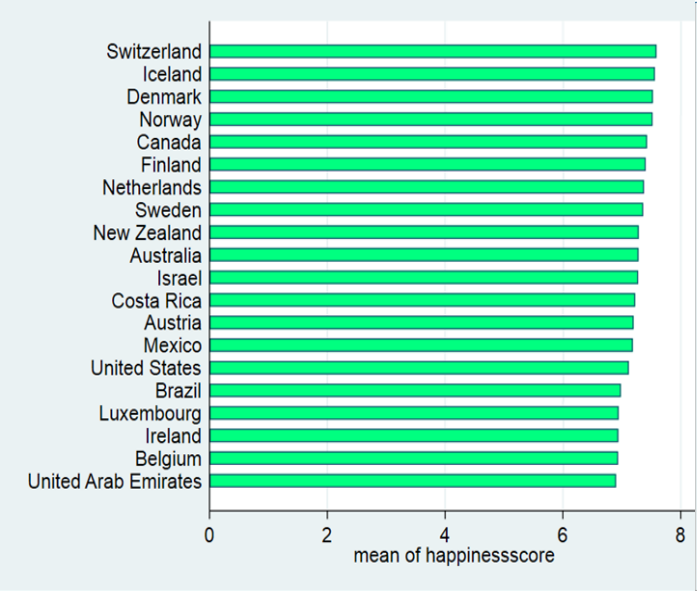
HEAT MAP OF HAPPINESS SCORE:



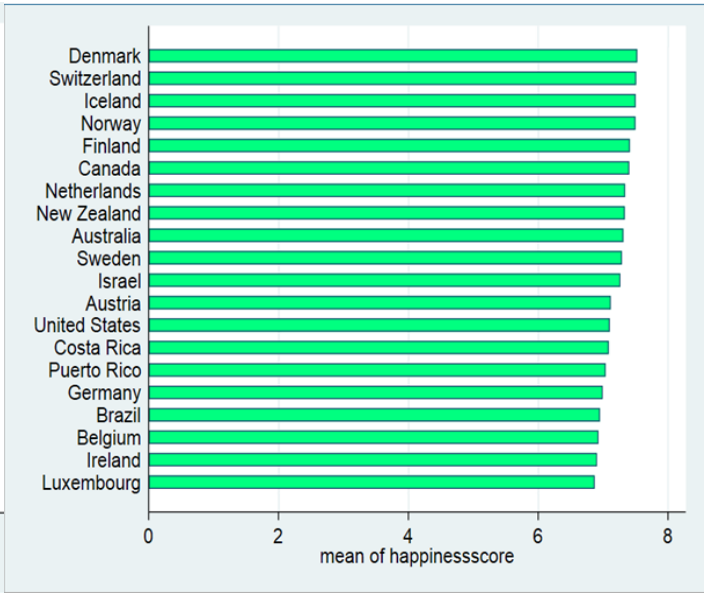
In the above figure, we used country variable to map the world and each country is colored by Happiness Score from lowest (dark red color) to highest (dark blue color) with score ranging from 3.0 to 7.5.

As we can infer from the figure that Norway, Denmark, Canada, Australia are in the top countries with highest Happiness Score and Syria, Tanzania, and Yemen are the countries with lowest Happiness Score.

Top 20 countries with highest mean of *Happiness Score* in 2015, 2016 and 2017

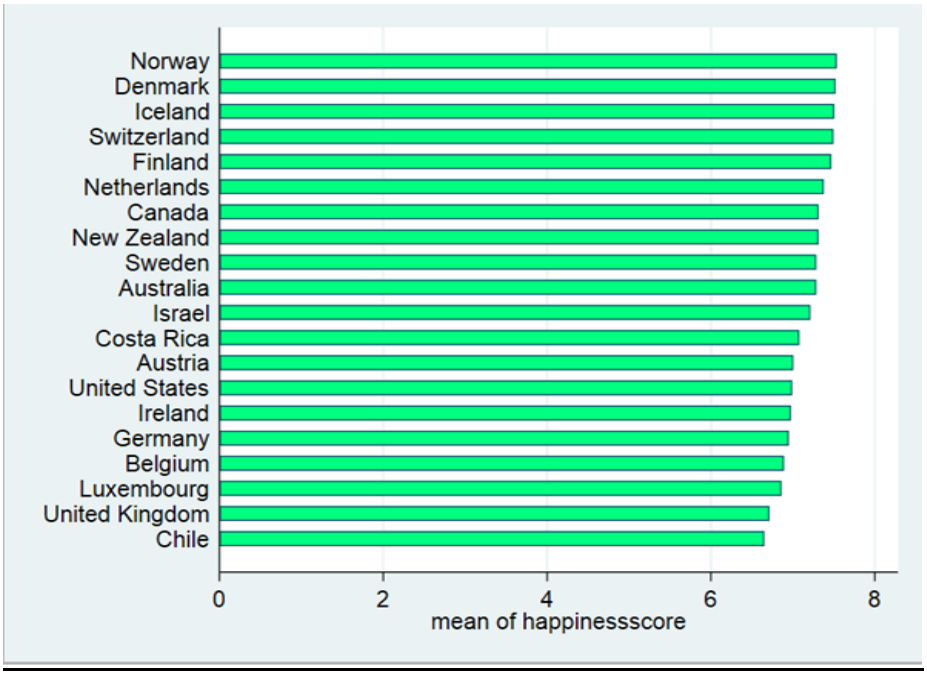


Year 2015



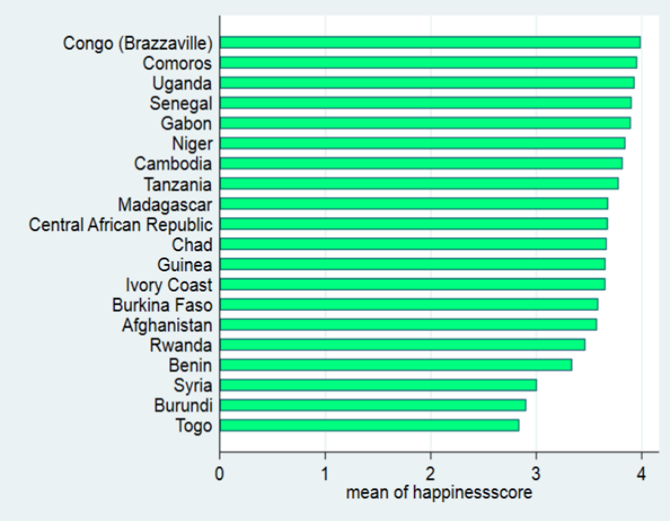
Year 2016

Year 2017

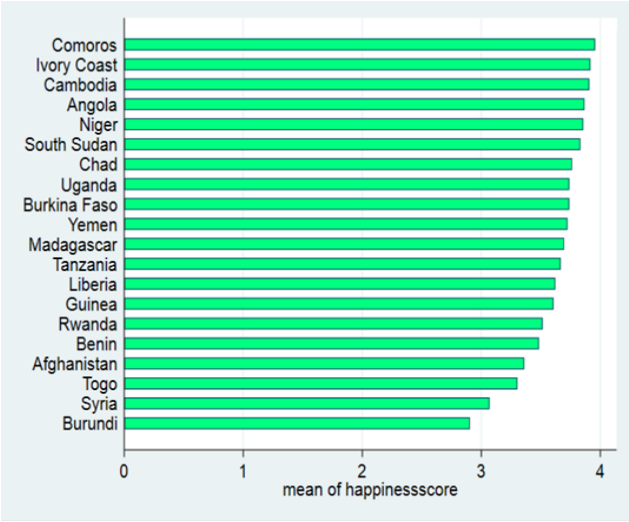


Although the top ten countries remain the same in all three years i.e. 2015, 2016 and 2017, there has been some shuffling of rankings, as is to be anticipated among countries so closely clustered in average scores.

Bottom 20 countries with lowest mean of happiness score in 2015, 2016 and 2017

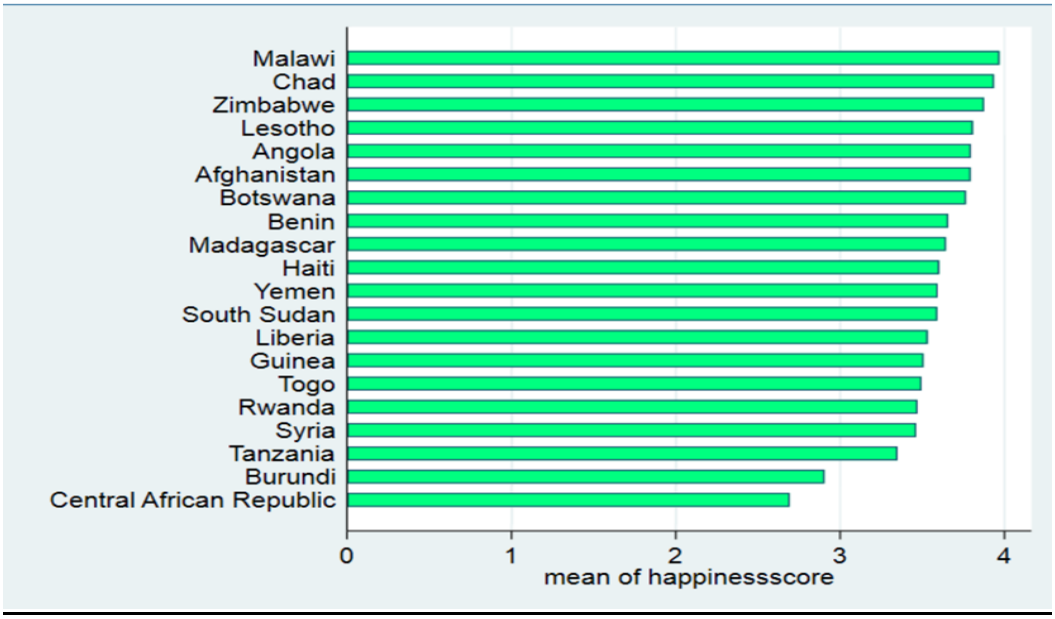


Year 2015



Year 2016

Year 2017



Unlike the top 10 countries, some of the bottom 10 countries are not consistent in holding their rankings over the years 2015, 2016 and 2017 and have showcased considerable changes in their happiness score.

CONCLUSION:

We observed that widening the concentration from Economy GDP to happiness, one can significantly increase the number of ways to improve the lives of the unhappy in different countries. The high values of following six variables significantly contributed to the happiness mean to mark the countries with highest scores: Economy GDP per capita, Health life expectancy, Family, Trust, Freedom, Generosity.

Norway bagged the first position as the happiest country in the world as per the 2017 report. It did so by investing their abundant natural resources into long-term sustainable growth as opposed to short-term gains. The specific analysis of USA ranking also enlightens that American happiness is adversely affected by social factors even after having favorable economic conditions.

Fresh possibilities to increase happiness can be created by aiming at the social sources of well-being thus reducing the dependency on the materialistic or money-oriented

resources. to pursue happiness. There is more scope of research to understand all the relationships of factors that influence the social fundamentals of happiness and consider alternatives to improve those fundamentals.