A Statistical Analysis on World Happiness Index Data

PRESENTED BY : ROHIT TIWARI AND ANKIT GAHLAWAT

Department of Statistics, Ramanujan College, University of Delhi

Table of Contents

Abstract

Presentation Title: A Statistical Analysis on World Happiness Index Data

Institution: Ramanujan College, University of Delhi

Abstract

The pursuit of happiness is one of the most basic aspects of life. However, there are many ways to define happiness and what it means to be happy. The word happiness can refer to a variety of states, but it typically implies a feeling of pleasure and contentment.

Since Happiness is a broad concept and many can define it in their own way. Now come to think of World Happiness Report, Presented by United Nations Sustainable Development Solutions Network, took some parameters that according to them may represents the broader concept of happiness if we talk about the well-being of the countries as well.

In this report we performed a detailed analysis on those factors using happiness index data that ranges from 2005 to 2022. Our analysis is divided into four sections namely “Worldwide”, “Region wise”, “Factors” and we also made a different section that only focuses on India. In this report we checked and looked upon the trends made by individual countries and Regions in their happiness indices along following years and also did some analysis on what effects COVID – 19 brought on happiness indices. Where as in ‘Factors’ section we scrutinized on all factors and according to the results decided, what factor out of all contributed to the most and what to the least.

We also broke some famous and layman sayings such as “Money cannot buy happiness” and did some estimations like what proportion of every factor should each and every country aim to attain a good happiness score. And cross-checked if it is the regions as a whole differ in their happiness levels.

Introduction

Happiness is an emotional state characterized by feelings of joy, satisfaction, contentment, and fulfillment. While happiness has many different definitions, it is often described as involving positive emotions and life satisfaction. When most people talk about happiness, they might be talking about how they feel in the present moment, or they might be referring to a more general sense of how they feel about life overall.

Because happiness tends to be such a broadly defined term, psychologists and other social scientists typically use the term 'subjective well-being' when they talk about this emotional state. Just as it sounds, subjective well-being tends to focus on an individual's overall personal feelings about their life in the present. Two key components of happiness (or subjective well-being) are:

- **The balance of emotions:** Everyone experiences both positive and negative emotions, feelings, and moods. Happiness is generally linked to experiencing more positive feelings than negative ones.
- **Life satisfaction:** This relates to how satisfied you feel with different areas of your life including your relationships, work, achievements, and other things that you consider important.

Since 2005, most of the humans on Earth have been given a nearly annual reminder that there are entire nations of people who are measurably happier than they are. This uplifting yearly notification is known as the **World Happiness Report**.

The World Happiness Report, one of the best tools for evaluating global happiness, is based on how ecstatic people perceive themselves to be. It considers six characteristics to rank countries on overall happiness: **GDP per capita, social support, life expectancy, freedom to make choices, generosity, and perception of corruption.**

The World Happiness Report is a 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 (quality of) life factors. As of February 2022, Finland had been ranked the happiest country in the world four times in a row. The report primarily uses data from the Gallup World Poll. Each annual report is available to the public to download on the World Happiness Report website.

With the release of each report, which is published by the United Nations Sustainable Development Solutions Network, the question is not which country will appear at the top of the rankings, but rather which Northern European country will. Finland has been the world’s happiest country for five years running; Denmark and Norway hold all but one of the other titles (which went to Switzerland in 2015). Since 2002, the World Happiness Report has used statistical analysis to determine the world's happiest countries. In its 2022 update, the report concluded that Finland is the happiest country in the world. To determine the world's happiest country, researchers analyzed comprehensive Gallup polling data from 149 countries for the

past three years, specifically monitoring performance in six particular categories: gross domestic product per capita, social support, healthy life expectancy, freedom to make your own life choices, generosity of the general population, and perceptions of internal and external corruption levels.

Methodology

To conduct the project, we have collected the data used in publishing The World Happiness Report, one of the best parameters for evaluating global happiness. Data from surveys is collected from people in over 150 countries every year. It considers six characteristics to rank countries on overall happiness: GDP per capita, social support, life expectancy, freedom to make choices, generosity, and perception of corruption.

We have taken data till 2020 in our project. 2021 data did not have value of the various factors, only the happiness score was listed as it was used to be in the early years.

Graphical Representation of the data

1. Pie Chart

A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

2. Bar Graphs

Bar graphs is used to display the category of data and it compares the data using solid bars to represent the quantities

3. Line Graph

A line graph is a graphical display of information that changes continuously over time. Within a line graph, there are various data points connected together by a straight line that reveals a continuous change in the values represented by the data points.

4. Boxplots

Boxplots are a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").

5. PP Plot

A P–P plot (probability–probability plot or per cent–per cent plot or P-value plot) is a probability plot for assessing how closely two data sets agree, which plots the two cumulative distribution functions against each other.

Descriptive Statistics:

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures; which enables us to present the raw data in a more meaningful way at the initial stages of analysis. There are 4 measures:

• **Measure of central tendency** – A measure of central tendency is a summary statistic that represents the centre point or typical value of a dataset. The three most common measures of central tendency are the **mean**, **median**, and **mode**. Each of these measures calculates the location of the central point using a different method.

• **Measure of dispersion** - They are used to describe the variability in the data set and the most common measures of dispersion are range, standard deviation, variance, and quartiles.

• **Measure of skewness**- Skewness means lack of symmetry in data. There is a need to study skewness to know about the shape of the curve which is drawn with the given data. A curve is said to be skewed if

1. Mean, Median and Mode all three fall at different points i.e. they are unequal.
2. Quartiles are not equidistant from the Median

• **Measure of kurtosis**- This measure helps us to know about the flatness or the peakedness of the frequency curve. There are 3 types of curves under this:

1. Leptokurtic curve
2. Normal curve
3. Platykurtic curve

Statistical Tests

Normality assumption

Statistical errors are common in scientific literature, and about 50% of the published article has at least one error. Many of the statistical procedures including correlation, regression and analysis of variance, parametric tests, based on the assumption that the data follows a normal distribution or a Gaussian distribution (after Johann Karl Gauss, 1777-1855); that is, it is assumed that the populations from which the samples are taken are normally distributed. The assumption of normality is especially critical when constructing reference intervals for variables. Normality and other assumptions should be taken seriously, when these assumptions do not hold, it is impossible to draw accurate and reliable conclusions about reality.

With large sample sizes, the violation of the normality assumption should not cause major problems; this implies that we can use parametric procedures even when the data are not normally distributed. If we have samples consisting of hundreds of observations, we can ignore the distribution of the data. According to the central limit theorem, (a) if the sample data are approximately normal then the sampling distribution too will be normal; (b) in large samples, the sampling distribution tends to be normal, regardless of the shape of the data; and (c) means of random samples from any distribution will themselves have normal distribution. Although true normally is considered to be a myth, we can look for normality visually by using normal plots or by significance tests, which is, comparing the sample distribution to a normal one. It is important to ascertain whether data show a serious deviation from normality.

The different hypothesis tests used in the research are:

Normal test:

Normal test is the test of significance when sample size is large. For large values of n, the number of trials, almost all the distributions, for example: Poisson, Negative Binomial etc. are very closely approximated to normal distribution. Thus, in this case we apply the normal test, which is based upon the fundamental area property of the normal probability curve.

Inferential Statistics

Generally, Inferential statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn. Because the aim of inferential statistics is to draw conclusions from a sample and generalize them to a population. So, we need to have confidence that our sample accurately reflects the population.

There are two main areas of inferential statistics:

- **Estimating Parameters:**
  This means taking a statistic from a sample data (i.e., as sample mean) and use it to say something about a population parameter (i.e., population mean)
- **Hypothesis Testing:**
  This is where you can use sample data to answer research questions.

Testing of Hypothesis

Hypothesis testing was introduced by Ronald Fisher, Jerzy Neyman, Karl Pearson and Pearson's son Egon Pearson. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter.

Key Terms:

Null Hypothesis: A null hypothesis is a type of hypothesis used in statistics that proposes no statistical significance exists in a set of given observations. It is denoted by Ho.

Alternate hypothesis: A hypothesis complementary to null hypothesis. It is denoted by H1.

Types of error –

Type 1 error: Reject H0, when H0 is true. It is denoted by alpha.

Type 2 error: Accept H0, when H0 is false. It is denoted by beta.

LEVEL OF SIGNIFICANCE: It is size of type 1 error or also called maximum procedure's risk.

One Tailed Test: One-tailed test is a statistical test in which the critical area of the distribution is one sided. And if the sample being tested falls into the one side, the alternative hypothesis will be accepted. It is either right tailed or left tailed.

Two Tailed Test: A test of any statistical hypothesis where the alternate hypothesis being two tailed. This is called two-tailed test.

Paired T-Test

A t-test allows us to compare the differences (measured in means or average values) of the two data sets and determine if they came from the same population. The t-test takes a sample from each of the two data sets and establishes the problem by stating the null hypothesis that there is no significant difference between the means of the two data sets. Based on the standard formulas, the value of calculated t is calculated and is compared against the standard (tabulated) value and accordingly the decision regarding the acceptance and rejection of null hypothesis is made.

Paired t-test for Difference of Means

$t = d / S/\sqrt{n}$

where,

d is the mean difference
n is the sample size (i.e., size of d).
S is the standard deviation of d

Correlation:

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

• The coefficient of determination (denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.

Spearman's Rank Correlation Coefficient:

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, measures the strength and direction of association between two ranked variables.

Kruskal Wallis Test:

The Kruskal Wallis test is the non-parametric alternative to the One Way ANOVA. Non-parametric means that the test doesn't assume your data comes from a particular distribution. The H test is used when the assumptions for ANOVA aren't met (like the assumption of normality). It is sometimes called the one-way ANOVA on ranks, as the ranks of the data values are used in the test rather than the actual data points. The test determines whether the medians of two or more groups are different. Like most statistical tests, you calculate a test statistic and compare it to a distribution cut-off point. The test statistic used in this test is called the H statistic.

The hypotheses for the test are:

H0: The sample comes from population medians are equal.

H1: The sample comes from population medians are not equal.

The Kruskal Wallis test will tell you if there is a significant difference between groups. However, it won't tell us which groups are different.

Anova (Analysis Of Variance)

Analysis of variance is a statistical method used to test difference between two or more means. According to Prof. R.A. Fisher, Analysis of Variance (ANOVA) is the "Separation of variance ascribable to one group

of causes from the variance ascribable to other groups". By this technique the total variation the sample data is expressed as the sum of variation due to the non-negative components where each of these components is a measure of the variation due to some specific independent source or factor. Usually the hypothesis we deal with in this method is :

H0 : There's no difference in the means of different factors undertaken

H1 : There's difference in means in the factors undertaken

Multiple Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables
- The independent variables are not too highly correlated with each other
- y_i observations are selected independently and randomly from the population
- Residuals should be normally distributed with a mean of 0 and variance σ

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

$$yi = в0 + в1.xi1 + в2.xi2 + . . . + вp.xip + \epsilon$$

where, for $i=n$ **observations:**

yi = dependent variable

xi = explanatory variables

$в0$ = y-intercept (constant term)

βp = slope coefficients for each explanatory variable

ϵ=the model's error term (also known as the residuals)

Analysis

About Dataset

1	Country name	Regional indicator	Happiness score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
2	Finland	Europe	7.842	10.775	0.954	72	0.949	-0.098	0.186
3	Denmark	Europe	7.62	10.933	0.954	72.7	0.946	0.03	0.179
4	Switzerland	Europe	7.571	11.117	0.942	74.4	0.919	0.025	0.292
5	Iceland	Europe	7.554	10.878	0.983	73	0.955	0.16	0.673
6	Netherlands	Europe	7.464	10.932	0.942	72.4	0.913	0.175	0.338
7	Norway	Europe	7.392	11.053	0.954	73.3	0.96	0.093	0.27
8	Sweden	Europe	7.363	10.867	0.934	72.7	0.945	0.086	0.237
9	Luxembourg	Europe	7.324	11.647	0.908	72.6	0.907	-0.034	0.386
10	New Zealand	Oceania	7.277	10.643	0.948	73.4	0.929	0.134	0.242
11	Austria	Europe	7.268	10.906	0.934	73.3	0.908	0.042	0.481
12	Australia	Oceania	7.183	10.796	0.94	73.9	0.914	0.159	0.442
13	Israel	Asia	7.157	10.575	0.939	73.503	0.8	0.031	0.753
14	Germany	Europe	7.155	10.873	0.903	72.5	0.875	0.011	0.46
15	Canada	The Americas	7.103	10.776	0.926	73.8	0.915	0.089	0.415
16	Ireland	Europe	7.085	11.342	0.947	72.4	0.879	0.077	0.363

This is the image of data available on the World Happiness Report website. It considers six characteristics to rank countries on overall happiness

1. **Logged Gross Domestic Product (GDP) per capita**: It shows a country's GDP divided by its total population. **(Limit: 0 to 12)**
2. **Social support**: It means having friends and other people, including family, to turn to in times of need or crisis to give you a broader focus and positive self-image. **(Limit: 0 to 1)**
3. **Health Life expectancy**: It is calculated by constructing a life table. A life table incorporates data on age-specific death rates for the population in question, which requires enumeration data for the number of people, and the number of deaths at each age for that population. **(Limit: Number of years)**
4. **Freedom to make choices**: It is a human right **(Limit: 0 to 1)**
5. **Generosity**: It is the quality of being kind and generous. **(Limit: -1 to +1)**
6. **Perception of Corruption**: Corruption Perceptions Index (CPI) is an index that ranks countries by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys. **(Limit: 0 to 1)**

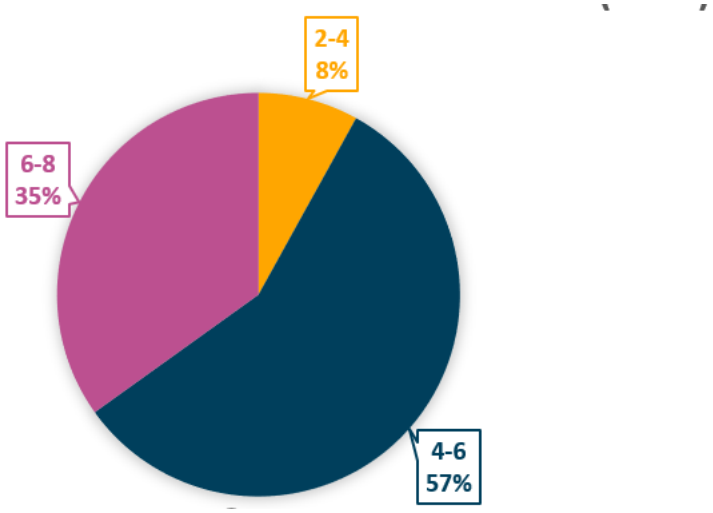
We have taken data till 2020 from the World Happiness report of 2021 in our project. 2021 data from Report of 2022 does not have all the factors included in the data, only the happiness score was listed as it was used to be in the early years.

Our analysis is divided into four sections namely “Worldwide”, “Region wise”, “Factors” and we also made a different section that only focuses on India.

Section A: Worldwide

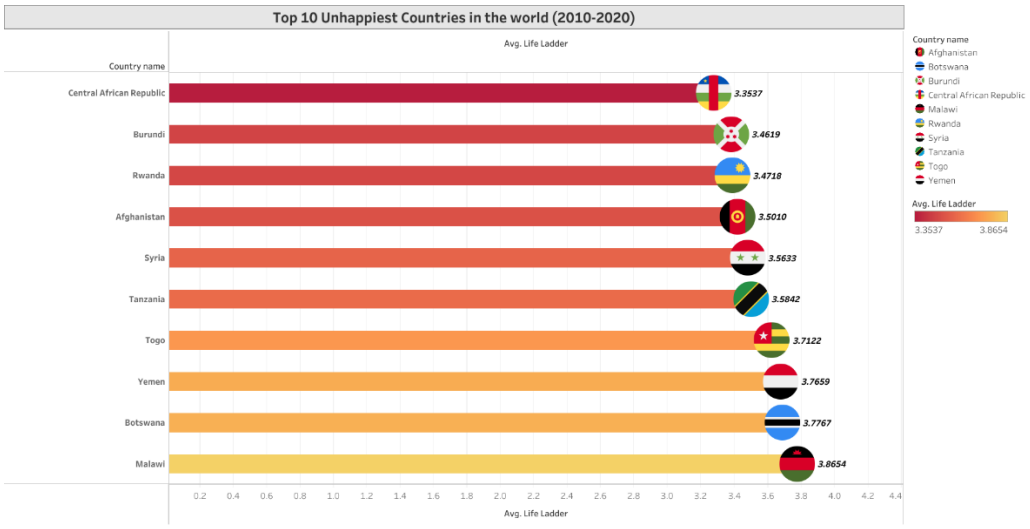
In this section we have solely focused on the worldwide scenario by explaining the difference in factors over the span of 10 years, top and bottom 10 countries of last decade, and checked upon different factors if they have changed in the span of last 10 years .

1) Proportion of Happiness Score



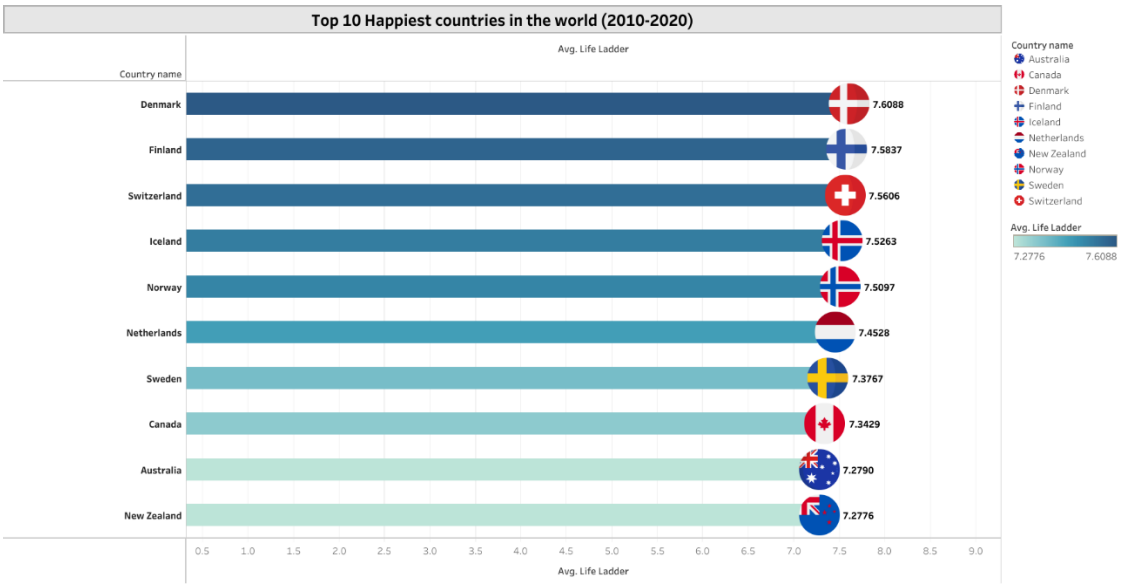
In this pie chart we can observe that about 8 % of countries have their Happiness index ranging from 2-4, 35 % of countries having happiness index ranging from 6-8 and 57 % of countries have their happiness index ranging from 4 to 6.

2) Top 10 Unhappiest Country from 2010 - 2020



This EDA represents the Bottom 10 happiest countries where on horizontal axis we have Average of Life Ladder and on Y axis we have countries out of which Malawi tops in it's ladder score naming itself the country with lowest happiness index and the unhappiest country and we can also see there's less difference between countries in their ladder score.

3) Top 10 Happiest Countries from 2010-2020



This EDA represents the Top 10 Happiest countries where on horizontal axis we have average life ladder and on y axis we have countries out of which Denmark tops in it's ladder score naming itself the country with highest happiest index making it the Happiest Country.

4) One tailed Paired T test

One tailed paired T test is used to check if there's a significance diff. in span 10 years of data among different factors (2011 – 2020) .

Ho: The means of Factors in 2011 and 2020 are significantly equal worldwide i.e. difference between means is zero

H1: The means of Factors in 2011 and 2020 are significantly unequal worldwide i.e. difference between 2020 and 2011 is greater than zero

The results are as follows

Factors	p value	result	Mean 2011	Mean 2020
Freedom to make life choices	0	Reject Null Hypothesis	0.726	0.791
Generosity	0.21	Fail to reject Null Hypothesis	-0.015	-0.024
Health Life Expectancy	0	Reject Null Hypothesis	60.69	65.205
Logged GDP	0	Reject Null Hypothesis	9.310	9.450
Perception of Corruption	0.34	Fail to reject Null Hypothesis	0.723	0.729
Social Support	0.1528	Fail to reject Null Hypothesis	0.810	0.816

t-Test: Paired Two Sample for Means		
	2011	*2021*
Mean	**9.310325926**	**9.450376991**
Variance	1.340349341	1.315737666
Observations	135	135
Pearson Correlation	0.990792816	
Hypothesized Mean Difference	0	
df	134	
t Stat	-10.3816853	
P(T<=t) one-tail	**3.42795E-19**	
t Critical one-tail	1.656304542	
P(T<=t) two-tail	6.8559E-19	
t Critical two-tail	1.977825758	

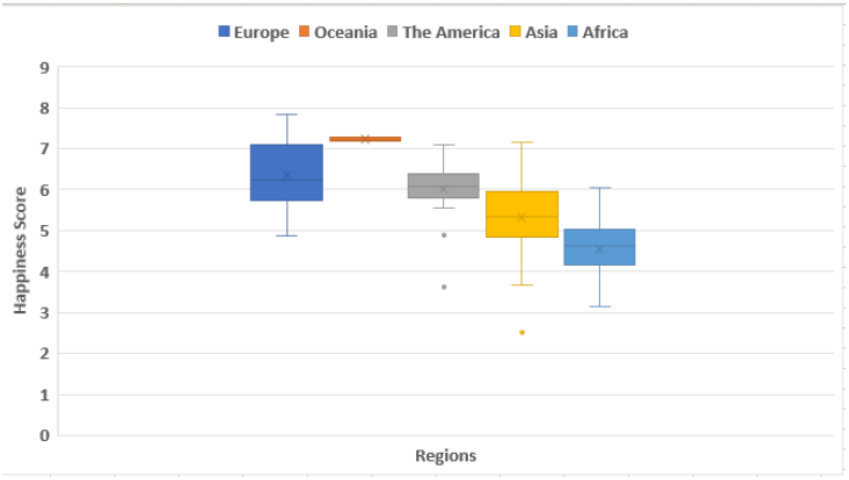
When p value is zero, null hypothesis is rejected. So we can conclude that **Freedom to make life choices, Health life expectancy at birth** and **Logged GDP** has significantly positively changed in the span of 10 years.

Section B: Regionwise

In this section we customized our data region-wise to check if there's a trend we can get if the data would have been in a region-wise format.

We have further classified our data in this section namely Asia, Africa, Oceania, Europe and The Americas where in The Americas we have 20+ countries, in Oceania we have 2 countries, in Africa we have 40+ countries, where in Asia we have 35+ and in Europe we have 45+ countries. and have analyzed and compared those regions with each other.

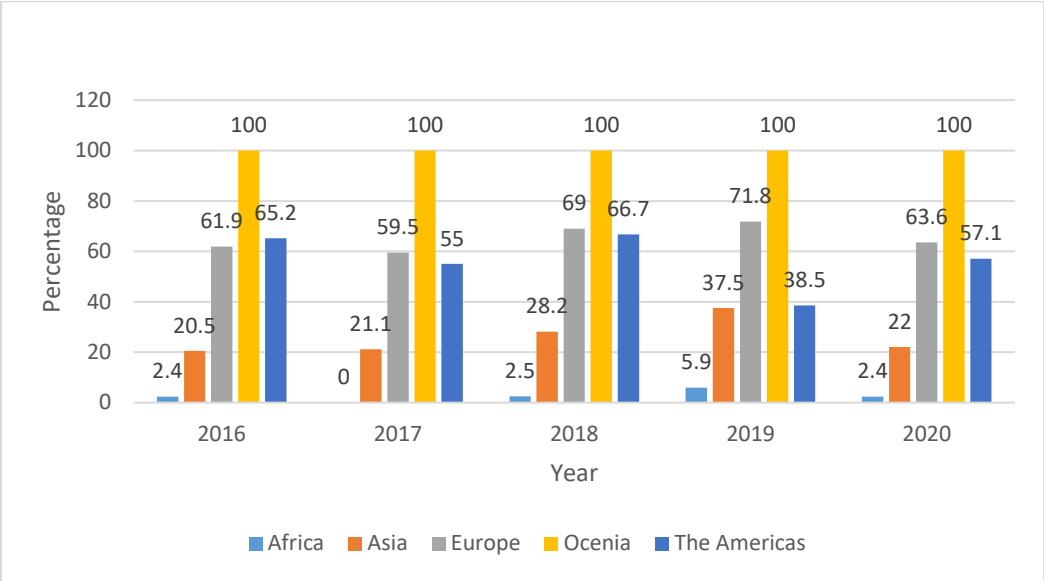
1) Boxplot of Happiness Score of different regions



From the box plot , we can clearly observe that the range of happiness of Europe is from 4.875 to 7.842 in which 50% of the observation lies between 5.73375 and 7.07975. Whereas oceania has the smallest range.

We can also observe few outliers in the case of the America and asia.

2) Percentage of countries in different regions having Happiness score greater than 6



From the above graph we can observe that the Americas faces overall reduction in the no. Of countries having ladder score greater than 6 from 2016 to 2020. Whereas Africa does not faces a significant change over the years but have the least percentage. Oceania having the highest percentage maintain the constant record of 100% countries over the years

Europe does not faces a significant difference from 2016 to 2020 but having the highest value in 2020 that is 71.8% of the countries. And for Asia also, the maximum no. of countries having ladder score more than 6 is observed in 2019 that is 37.5% of the countries.

3) Anova

Ho: There is no significant difference in means among all the regions
H1: There is a significant difference in means among all the regions

Summary Table

Year	p-value	Result	Similar Regions
2011	0	Fail to reject the null hypothesis	America and Europe
2012	0	Fail to reject the null hypothesis	America and Europe
2013	0	Fail to reject the null hypothesis	America and Europe
2014	0	Fail to reject the null hypothesis	America, Oceania and Europe
2015	0	Fail to reject the null hypothesis	America, Oceania and Europe
2016	0	Fail to reject the null hypothesis	America, Oceania and Europe
2017	0	Fail to reject the null hypothesis	America and Europe
2018	0	Fail to reject the null hypothesis	America and Europe
2019	0	Fail to reject the null hypothesis	America, Asia and Europe
2020	0	Fail to reject the null hypothesis	America and Europe

Year 2011 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	66.319	3	22.106	29.172	.000
Within Groups	103.060	136	.758		
Total	169.379	139			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.873/0.875) where as all the regions differ in their ladder score(p value= 0.000).

Multiple Comparisons

Dependent Variable: Ladder_Score

	(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Africa	Asia	-1.173668*	.192840	.000	-1.67513	-.67221
		Europe	-1.809345*	.190590	.000	-2.30495	-1.31374
		The Amer	-1.977021*	.224895	.000	-2.56183	-1.39221
	Asia	Africa	1.173668*	.192840	.000	.67221	1.67513
		Europe	-.635676*	.188072	.005	-1.12473	-.14662
		The Amer	-.803352*	.222765	.002	-1.38263	-.22408
	Europe	Africa	1.809345*	.190590	.000	1.31374	2.30495
		Asia	.635676*	.188072	.005	.14662	1.12473
		The Amer	-.167676	.220820	.873	-.74189	.40654
	The Amer	Africa	1.977021*	.224895	.000	1.39221	2.56183
		Asia	.803352*	.222765	.002	.22408	1.38263
		Europe	.167676	.220820	.873	-.40654	.74189
Games-Howell	Africa	Asia	-1.173668*	.184263	.000	-1.65800	-.68934
		Europe	-1.809345*	.184354	.000	-2.29355	-1.32514
		The Amer	-1.977021*	.207655	.000	-2.53259	-1.42146
	Asia	Africa	1.173668*	.184263	.000	.68934	1.65800
		Europe	-.635676*	.201454	.012	-1.16427	-.10708
		The Amer	-.803352*	.222975	.004	-1.39613	-.21057
	Europe	Africa	1.809345*	.184354	.000	1.32514	2.29355
		Asia	.635676*	.201454	.012	.10708	1.16427
		The Amer	-.167676	.223050	.875	-.76043	.42508
	The Amer	Africa	1.977021*	.207655	.000	1.42146	2.53259
		Asia	.803352*	.222975	.004	.21057	1.39613
		Europe	.167676	.223050	.875	-.42508	.76043

*. The mean difference is significant at the 0.05 level.

Year 2012 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	83.370	3	27.790	36.883	.000
Within Groups	98.704	131	.753		
Total	182.075	134			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.465/0.482) where as all the regions differ in their ladder score (p value= = 0.000).

Multiple Comparisons

Dependent Variable: Ladder_Score

	(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Africa	Asia	-1.161277*	.207707	.000	-1.70180	-.62076
		Europe	-1.850653*	.206598	.000	-2.38829	-1.31302
		The Amer	-2.180851*	.238883	.000	-2.80250	-1.55920
	Asia	Africa	1.161277*	.207707	.000	.62076	1.70180
		Europe	-.689376*	.192909	.003	-1.19139	-.18736
		The Amer	-1.019574*	.227148	.000	-1.61068	-.42846
	Europe	Africa	1.850653*	.206598	.000	1.31302	2.38829
		Asia	.689376*	.192909	.003	.18736	1.19139
		The Amer	-.330198	.226134	.465	-.91867	.25827
	The Amer	Africa	2.180851*	.238883	.000	1.55920	2.80250
		Asia	1.019574*	.227148	.000	.42846	1.61068
		Europe	.330198	.226134	.465	-.25827	.91867
Games-Howell	Africa	Asia	-1.161277*	.181109	.000	-1.63892	-.68364
		Europe	-1.850653*	.184069	.000	-2.33597	-1.36534
		The Amer	-2.180851*	.196820	.000	-2.71050	-1.65121
	Asia	Africa	1.161277*	.181109	.000	.68364	1.63892
		Europe	-.689376*	.216274	.011	-1.25700	-.12175
		The Amer	-1.019574*	.227225	.000	-1.62257	-.41658
	Europe	Africa	1.850653*	.184069	.000	1.36534	2.33597
		Asia	.689376*	.216274	.011	.12175	1.25700
		The Amer	-.330198	.229592	.482	-.93901	.27861
	The Amer	Africa	2.180851*	.196820	.000	1.65121	2.71050
		Asia	1.019574*	.227225	.000	.41658	1.62257
		Europe	.330198	.229592	.482	-.27861	.93901

*. The mean difference is significant at the 0.05 level.

Year 2013 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	84.158	3	28.053	38.711	.000
Within Groups	100.730	139	.725		
Total	184.887	142			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.873) where as all the regions differ in their ladder score (p value= = 0.000).

Multiple Comparisons

Dependent Variable: Ladder_Score

	(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Africa	Asia	-1.173668*	.192840	.000	-1.67513	-.67221
		Europe	-1.809345*	.190590	.000	-2.30495	-1.31374
		The Amer	-1.977021*	.224895	.000	-2.56183	-1.39221
	Asia	Africa	1.173668*	.192840	.000	.67221	1.67513
		Europe	-.635676*	.188072	.005	-1.12473	-.14662
		The Amer	-.803352*	.222765	.002	-1.38263	-.22408
	Europe	Africa	1.809345*	.190590	.000	1.31374	2.30495
		Asia	.635676*	.188072	.005	.14662	1.12473
		The Amer	-.167676	.220820	.873	-.74189	.40654
	The Amer	Africa	1.977021*	.224895	.000	1.39221	2.56183
		Asia	.803352*	.222765	.002	.22408	1.38263
		Europe	.167676	.220820	.873	-.40654	.74189
Games-Howell	Africa	Asia	-1.173668*	.184263	.000	-1.65800	-.68934
		Europe	-1.809345*	.184354	.000	-2.29355	-1.32514
		The Amer	-1.977021*	.207655	.000	-2.53259	-1.42146
	Asia	Africa	1.173668*	.184263	.000	.68934	1.65800
		Europe	-.635676*	.201454	.012	-1.16427	-.10708
		The Amer	-.803352*	.222975	.004	-1.39613	-.21057
	Europe	Africa	1.809345*	.184354	.000	1.32514	2.29355
		Asia	.635676*	.201454	.012	.10708	1.16427
		The Amer	-.167676	.223050	.875	-.76043	.42508
	The Amer	Africa	1.977021*	.207655	.000	1.42146	2.53259
		Asia	.803352*	.222975	.004	.21057	1.39613
		Europe	.167676	.223050	.875	-.42508	.76043

*. The mean difference is significant at the 0.05 level.

Year 2014 :

-Significant difference in means :

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	77.117	4	19.279	26.672	.000
Within Groups	99.750	138	.723		
Total	176.867	142			

-Exactly what regions are differing in means :

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America", "Europe" and "Oceania" are only countries when that doesn't differ in their means, p-value = 0.201(Oceania and Europe), p-value = 1.000(Europe and The America), p-value = 0.235(Oceania and The America), where as all the regions differ in their ladder score.

Multiple Comparisons

Dependent Variable: Ladder_Score

Tukey HSD

(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Asia	Africa	.956611*	.194186	.000	.41985	1.49338
	Europe	-.763116*	.185580	.001	-1.27609	-.25014
	Oceania	-2.092500*	.615664	.008	-3.79431	-.39069
	The Americas	-.783952*	.228145	.007	-1.41459	-.15332
Africa	Asia	-.956611*	.194186	.000	-1.49338	-.41985
	Europe	-1.719727*	.192064	.000	-2.25063	-1.18883
	Oceania	-3.049111*	.617650	.000	-4.75641	-1.34182
	The Americas	-1.740563*	.233450	.000	-2.38586	-1.09527
Europe	Asia	.763116*	.185580	.001	.25014	1.27609
	Africa	1.719727*	.192064	.000	1.18883	2.25063
	Oceania	-1.329384	.614998	.201	-3.02935	.37058
	The Americas	-.020836	.226341	1.000	-.64648	.60481
Oceania	Asia	2.092500*	.615664	.008	.39069	3.79431
	Africa	3.049111*	.617650	.000	1.34182	4.75641
	Europe	1.329384	.614998	.201	-.37058	3.02935
	The Americas	1.308548	.629153	.235	-.43054	3.04764
The Americas	Asia	.783952*	.228145	.007	.15332	1.41459
	Africa	1.740563*	.233450	.000	1.09527	2.38586
	Europe	.020836	.226341	1.000	-.60481	.64648
	Oceania	-1.308548	.629153	.235	-3.04764	.43054

*. The mean difference is significant at the 0.05 level.

Year 2015 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	93.544	4	23.386	33.314	.000
Within Groups	99.683	142	.702		
Total	193.227	146			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America", "Europe" and "Oceania" doesn't differ in their means, p-value = 0.349(Europe and Oceania), p-value = 0.955(Europe and The Americas), where as all the regions differ in their ladder score.

Multiple Comparisons

Dependent Variable: Ladder_Score

Tukey HSD

(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Asia	Africa	1.075066*	.187408	.000	.55723	1.59290
	Europe	-.826083*	.185271	.000	-1.33801	-.31415
	Oceania	-1.947641*	.607452	.014	-3.62612	-.26917
	The Americas	-.671596*	.223403	.026	-1.28889	-.05430
Africa	Asia	-1.075066*	.187408	.000	-1.59290	-.55723
	Europe	-1.901149*	.182886	.000	-2.40649	-1.39581
	Oceania	-3.022707*	.606728	.000	-4.69919	-1.34623
	The Americas	-1.746662*	.221429	.000	-2.35850	-1.13482
Europe	Asia	.826083*	.185271	.000	.31415	1.33801
	Africa	1.901149*	.182886	.000	1.39581	2.40649
	Oceania	-1.121558	.606072	.349	-2.79622	.55311
	The Americas	.154487	.219623	.955	-.45236	.76134
Oceania	Asia	1.947641*	.607452	.014	.26917	3.62612
	Africa	3.022707*	.606728	.000	1.34623	4.69919
	Europe	1.121558	.606072	.349	-.55311	2.79622
	The Americas	1.276045	.618794	.242	-.43377	2.98586
The Americas	Asia	.671596*	.223403	.026	.05430	1.28889
	Africa	1.746662*	.221429	.000	1.13482	2.35850
	Europe	-.154487	.219623	.955	-.76134	.45236
	Oceania	-1.276045	.618794	.242	-2.98586	.43377

*. The mean difference is significant at the 0.05 level.

Year 2016 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	86.839	4	21.710	29.880	.000
Within Groups	103.171	142	.727		
Total	190.009	146			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America", "Europe" and "Oceania" doesn't differ in their means, p-value = 0.414(Europe and Oceania), p-value = 0.986(Europe and The Americas), where as all the regions differ in their ladder score.

Multiple Comparisons

Dependent Variable: Ladder_Score

Tukey HSD

(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Asia	Africa	.853801*	.190658	.000	.32699	1.38062
	Europe	-.951505*	.189548	.000	-1.47525	-.42776
	Oceania	-2.024077*	.617986	.011	-3.73166	-.31649
	The Americas	-.838686*	.224096	.002	-1.45789	-.21948
Africa	Asia	-.853801*	.190658	.000	-1.38062	-.32699
	Europe	-1.805307*	.187136	.000	-2.32239	-1.28822
	Oceania	-2.877878*	.617251	.000	-4.58343	-1.17233
	The Americas	-1.692487*	.222059	.000	-2.30607	-1.07891
Europe	Asia	.951505*	.189548	.000	.42776	1.47525
	Africa	1.805307*	.187136	.000	1.28822	2.32239
	Oceania	-1.072571	.616909	.414	-2.77718	.63204
	The Americas	.112820	.221107	.986	-.49813	.72377
Oceania	Asia	2.024077*	.617986	.011	.31649	3.73166
	Africa	2.877878*	.617251	.000	1.17233	4.58343
	Europe	1.072571	.616909	.414	-.63204	2.77718
	The Americas	1.185391	.628384	.329	-.55092	2.92171
The Americas	Asia	.838686*	.224096	.002	.21948	1.45789
	Africa	1.692487*	.222059	.000	1.07891	2.30607
	Europe	-.112820	.221107	.986	-.72377	.49813
	Oceania	-1.185391	.628384	.329	-2.92171	.55092

*. The mean difference is significant at the 0.05 level.

Year 2017 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Life Ladder

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	70.580	3	23.527	36.135	.000
Within Groups	88.546	136	.651		
Total	159.126	139			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.698) where as all the regions differ in their ladder score (p value = 0.000).

Multiple Comparisons

Dependent Variable: Life Ladder

Tukey HSD

(I) regions	(J) regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Africa	Asia	-.635228656*	.1827845218	.004	-1.11066586	-.159791454
	Europe	-1.72546275*	.1782650457	.000	-2.18914443	-1.26178107
	The Americas	-1.48731411*	.2209757116	.000	-2.06208965	-.912538556
Asia	Africa	.635228656*	.1827845218	.004	.1597914545	1.110665857
	Europe	-1.09023410*	.1806518691	.000	-1.56012410	-.620344095
	The Americas	-.852085449*	.2229056671	.001	-1.43188097	-.272289931
Europe	Africa	1.72546275*	.1782650457	.000	1.261781070	2.189144433
	Asia	1.09023410*	.1806518691	.000	.6203440951	1.560124097
	The Americas	.2381486467	.2192149209	.698	-.332046945	.8083442381
The Americas	Africa	1.48731411*	.2209757116	.000	.9125385558	2.062089654
	Asia	.852085449*	.2229056671	.001	.2722899308	1.431880968
	Europe	-.238148647	.2192149209	.698	-.808344238	.3320469447

*. The mean difference is significant at the 0.05 level.

Year 2018 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Life Ladder

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	78.379	3	26.126	38.811	.000
Within Groups	92.898	138	.673		
Total	171.277	141			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.915) where as all the regions differ in their ladder score (p value= = 0.000).

Multiple Comparisons

Dependent Variable: Life Ladder

Tukey HSD

(I) regions	(J) regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Africa	Asia	-.844416412*	.1846356146	.000	-1.32458193	-.364250899
	Europe	-1.81597364*	.1812660685	.000	-2.28737627	-1.34457101
	The Americas	-1.67291697*	.2211003879	.000	-2.24791319	-1.09792074
Asia	Africa	.844416412*	.1846356146	.000	.3642508986	1.324581926
	Europe	-.971557229*	.1824524883	.000	-1.44604528	-.497069179
	The Americas	-.828500553*	.2220740967	.002	-1.40602902	-.250972090
Europe	Africa	1.81597364*	.1812660685	.000	1.344571009	2.287376275
	Asia	.971557229*	.1824524883	.000	.4970691792	1.446045280
	The Americas	.1430566765	.2192806007	.915	-.427206988	.7133203414
The Americas	Africa	1.67291697*	.2211003879	.000	1.097920741	2.247913190
	Asia	.828500553*	.2220740967	.002	.2509720898	1.406029016
	Europe	-.143056677	.2192806007	.915	-.713320341	.4272069884

*. The mean difference is significant at the 0.05 level.

Year 2019 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

Life Ladder

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	35.541	3	11.847	20.786	.000
Within Groups	50.725	89	.570		
Total	86.266	92			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that the pair "The America" and "Europe" and "Asia" and "The Americas" doesn't differ in their means (having p-value = 0.112 and p-value = 0.596 respectively) where as all the regions differ in their ladder score (p value= = 0.000).

Multiple Comparisons

Dependent Variable: Life Ladder

Tukey HSD

(I) regions	(J) regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Africa	Asia	-.800064522*	.2393193952	.007	-1.42666046	-.173468579
	Europe	-1.67533396*	.2194084246	.000	-2.24979817	-1.10086974
	The Americas	-1.12526475*	.2781509454	.001	-1.85353106	-.396998435
Asia	Africa	.800064522*	.2393193952	.007	.1734685789	1.426660464
	Europe	-.875269437*	.1958611951	.000	-1.38808132	-.362457552
	The Americas	-.325200228	.2599798057	.596	-1.00589011	.3554896568
Europe	Africa	1.67533396*	.2194084246	.000	1.100869743	2.249798175
	Asia	.875269437*	.1958611951	.000	.3624575523	1.388081323
	The Americas	.5500692099	.2417763083	.112	-.082959524	1.183097944
The Americas	Africa	1.12526475*	.2781509454	.001	.3969984351	1.853531063
	Asia	.3252002276	.2599798057	.596	-.355489657	1.005890112
	Europe	-.550069210	.2417763083	.112	-1.18309794	.0829595244

*. The mean difference is significant at the 0.05 level.

Year 2020 :

-Significant difference is means

We can see that there's a significant difference between different regions' mean as F statistic in the table below is less than 0.05(p = 0.000).

ANOVA

ladder_Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	77.351	3	25.784	42.140	.000
Within Groups	87.495	143	.612		
Total	164.847	146			

-Exactly what regions are differing in means ?

Since there was a significant difference in means as shown above but that doesn't show what regions are differing, in this table given below we can see that only "The America" and "Europe" doesn't differ in their means (p-value = 0.386) where as all the regions differ in their ladder score (p value= = 0.000).

Multiple Comparisons

Dependent Variable: ladder_Score

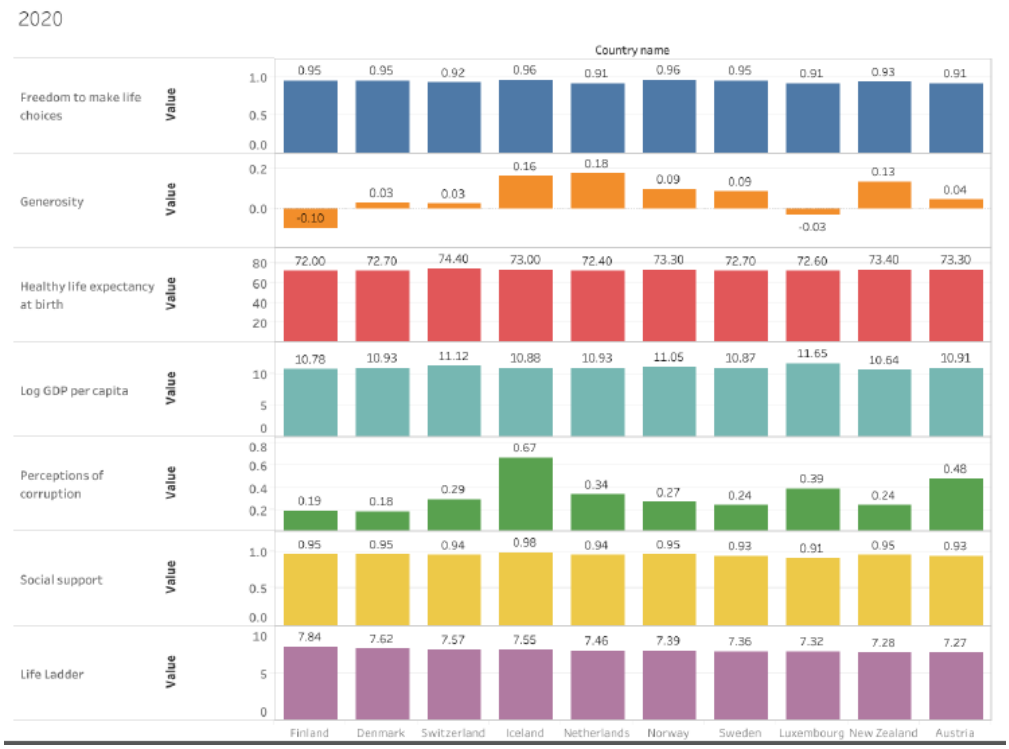
	(I) Regions	(J) Regions	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Games-Howell	Africa	Europe	-1.811726*	.156968	.000	-2.22350	-1.39995
		The Amer	-1.493476*	.185498	.000	-1.99179	-.99516
		Asia	-.771476*	.174325	.000	-1.23031	-.31264
	Europe	Africa	1.811726*	.156968	.000	1.39995	2.22350
		The Amer	.318250	.198305	.386	-.21042	.84692
		Asia	1.040250*	.187895	.000	.54704	1.53346
	The Amer	Africa	1.493476*	.185498	.000	.99516	1.99179
		Europe	-.318250	.198305	.386	-.84692	.21042
		Asia	.722000*	.212309	.007	.15831	1.28569
	Asia	Africa	.771476*	.174325	.000	.31264	1.23031
		Europe	-1.040250*	.187895	.000	-1.53346	-.54704
		The Amer	-.722000*	.212309	.007	-1.28569	-.15831

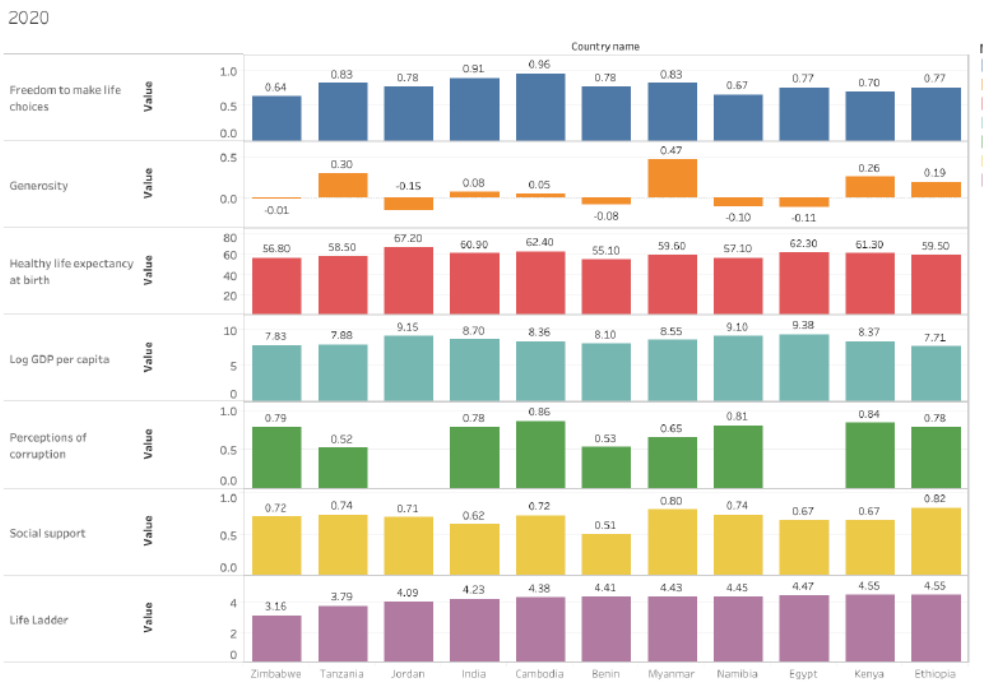
*. The mean difference is significant at the 0.05 level.

Section C: Factorwise

In this section we focused ourselves on the trend if we can get on different factors over the span of 10 years in this section we checked if factors differ or not and if they do can we see any trend in it.

1) Comparison of top 10 and bottom 10 countries





For the top 10 countries, the **Happiness Score** range lies between 0.91 and 0.96. For the bottom 10 the range lies between 0.64 and 0.91.

For the top 10 countries, the **Generosity** score range lies between -0.1 and 0.18. For the bottom 10 the range lies between -0.15 and 0.47.

For the top 10 countries, the **Healthy life expectancy** score range lies between 72 and 74.4. For the bottom 10 the range lies between 55.1 and 67.2.

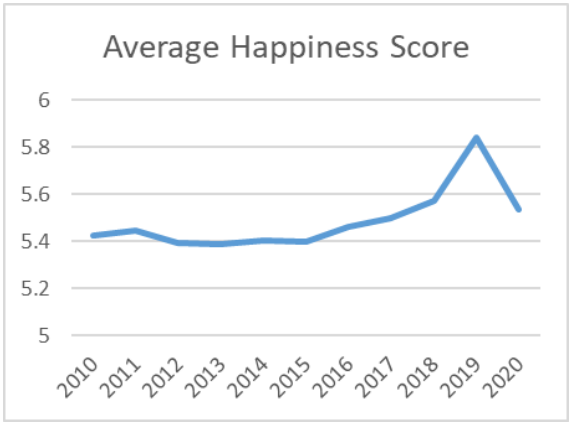
For the top 10 countries, the **log GDP** per capita score range lies between 10.64 and . For the bottom 10 the range lies between 7.71 and 9.38.

For the top 10 countries, the **Perception of corruption** score range lies between 0.18 and 0.67. For the bottom 10 the range lies between 0 and 0.86.

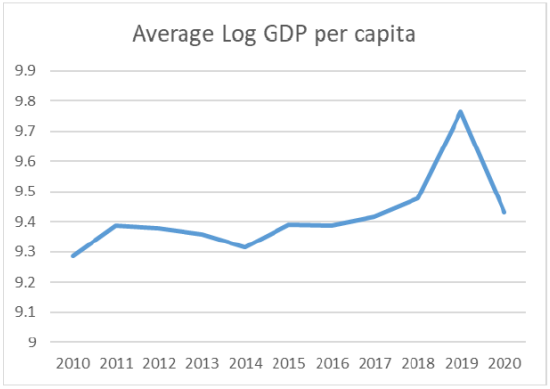
For the top 10 countries, the **Social support score** range lies between 0.91 and 0.95. For the bottom 10 the range lies between 0.51 and 0.82.

For the top 10 countries, the **Life Ladder Score** range lies between 7.27 and 7.84. For the bottom 10 the range lies between 3.16 and 4.55

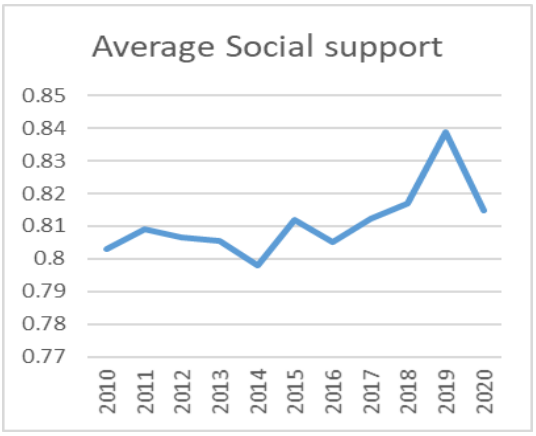
2) Analysing Different Factors Trend For The Past 10 Years



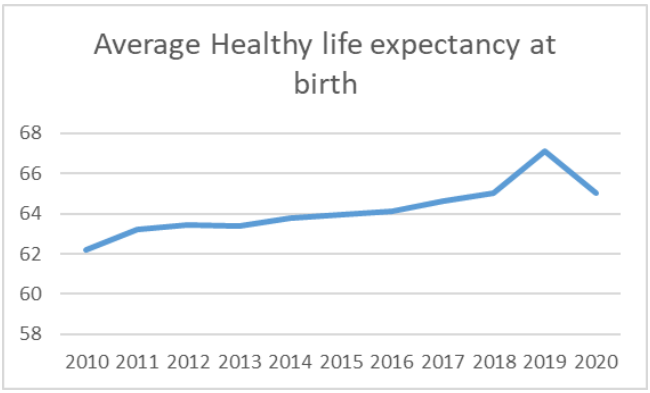
From the data, we observe that there is a significant increase in 2019 in happiness score and then again faces a sudden fall.



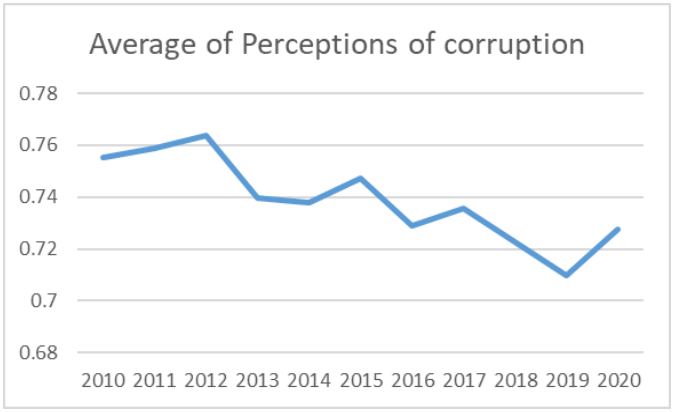
From the data, we observe that there is a significant increase in 2019 in happiness score and then again faces a sudden fall. And overall faces slight increase over the years



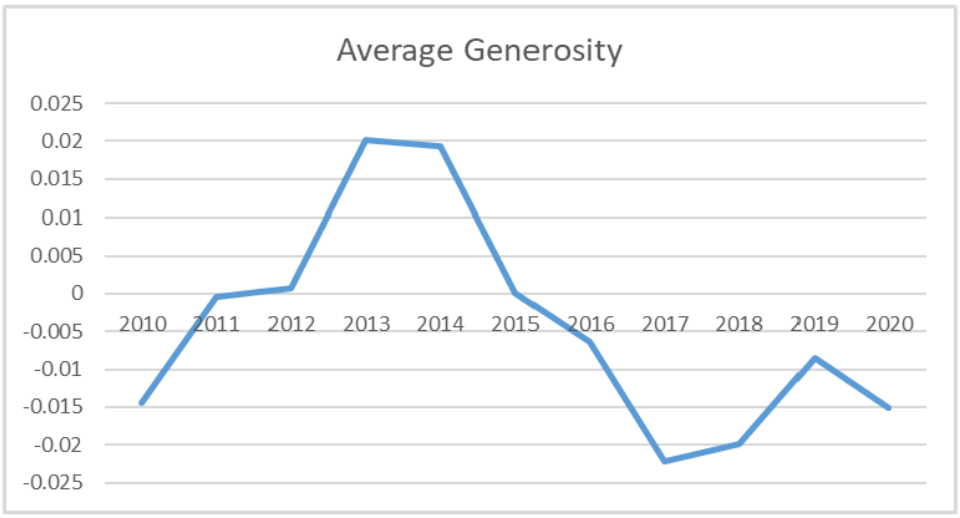
From the data, we observe that there is a significant increase in 2019 in happiness score and then again faces a sudden fall. And overall faces slight increase over the years.



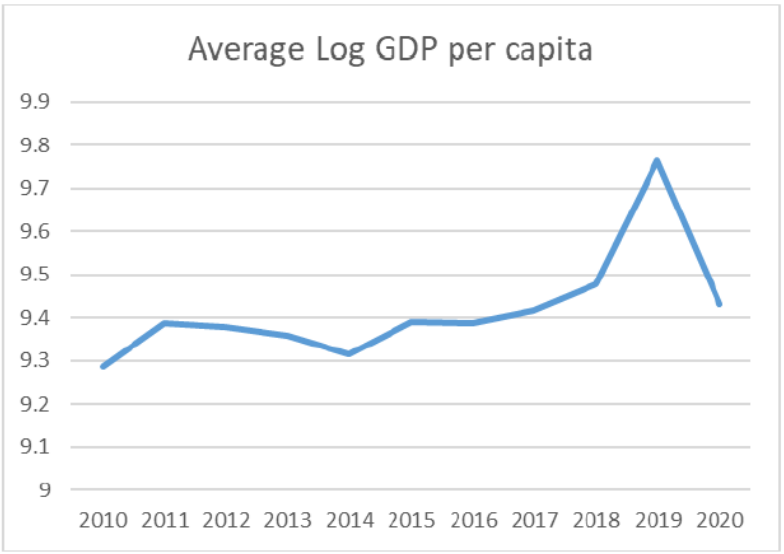
From the data, we can observe that there is a overall slight increase in healthy life expectancy over the years. There is a sudden jump in 2019 that is 67.1175.



From the data, we observe that there is overall decrease over the years. And least value is observed in 2019 and the highest in 2012.



The pattern is not fixed here but we can observe the highest value in 2013 and least value in 2017.



There is overall increase in the values over the years. The graph shown is approximately linear except 2011 and 2019 where we observe the least and highest value respectively.

3) Kruskal Wallis Test

Ho: There is no significant difference in Factors between years 2017-2021
H1: There is a significant difference in Factors between years 2017-2021

Summary Table

Factors	p value	Result
Freedom to make life choices	0.263	Retain Null Hypothesis
Generosity	0.818	Retain Null Hypothesis
Health Life Expectancy	0.009	Reject Null Hypothesis
Logged GDP	0.149	Retain Null Hypothesis
Perception of Corruption	0.882	Retain Null Hypothesis
Social Support	0.237	Retain Null Hypothesis

A Kruskal-Wallis H test showed that there was a statistically significant difference in factor Health life expectancy between the 2016-2020, $p < 0.05$

Freedom to make life choices

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Freedom to make life choices is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.263	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Generosity

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Generosity is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.818	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Health life Expectancy

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Health Life Expectancy is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.009	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Log GDP

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Log GDP per capita is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.149	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Perception of Corruption

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Perceptions of corruption is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.882	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Social Support

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Social support is the same across categories of Year.	Independent-Samples Kruskal-Wallis Test	.237	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

4) Spearman Rank Correlation

Factors	Correlation Value	Result
Logged GDP	0.809	Strong positive correlation
Social Support	0.798	Strong positive correlation
Health Life Expectancy	0.797	Strong positive correlation
Freedom to make life choices	0.607	Strong positive correlation
Generosity	-0.008	No correlation
Perception of Corruption	-0.301	Moderate negative correlation

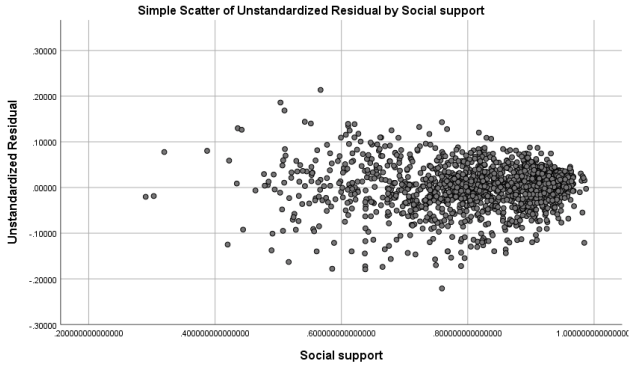
From Spearman Rank Correlation table we can see that Social Support, Health Life Expectancy, Logged GDP have correlation have strong correlation concluding that these variables can act as an important factors in providing best estimate for happiness index whereas Freedom to make life choices have comparatively low correlation and perception of corruption have moderately negative correlation where as Generosity has almost negligible correlation which we can also cross-check from Multiple Linear Regression results.

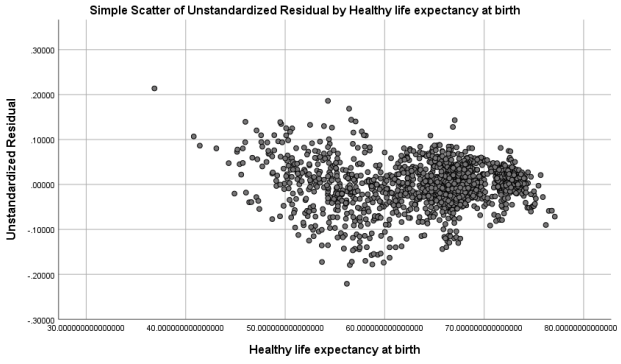
5) Multiple Linear Regression

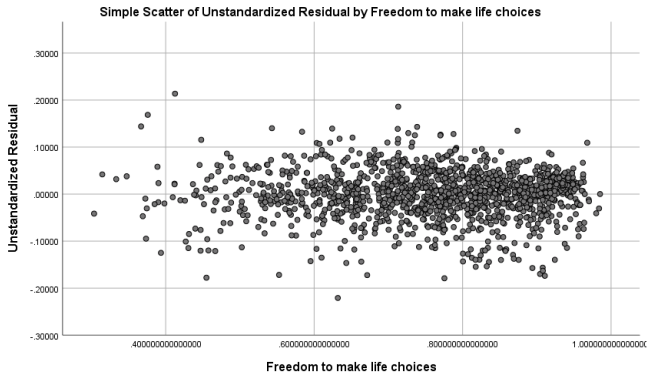
Assumptions and their Outputs :

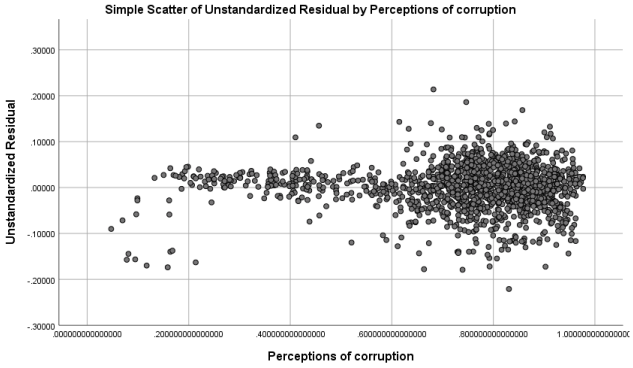
- **Residuals Vs. Predictor Plot :**

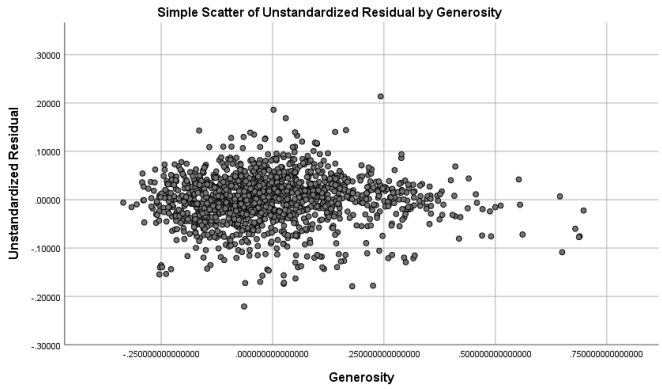
All the possible plots of Residuals Vs. Predictor Plot adhere to the conditions and assumptions of this plot i.e., a well-behaved plot will bounce randomly and form a roughly horizontal band around the residual = 0 line.

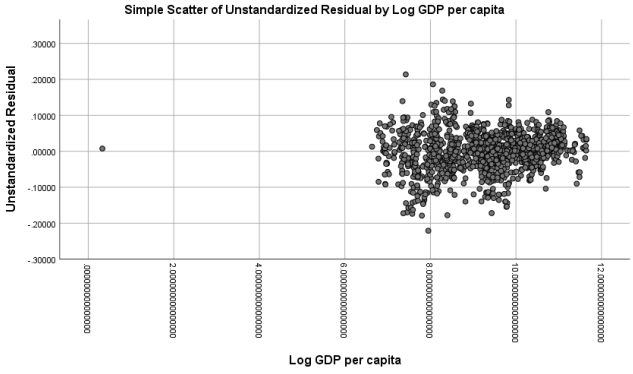


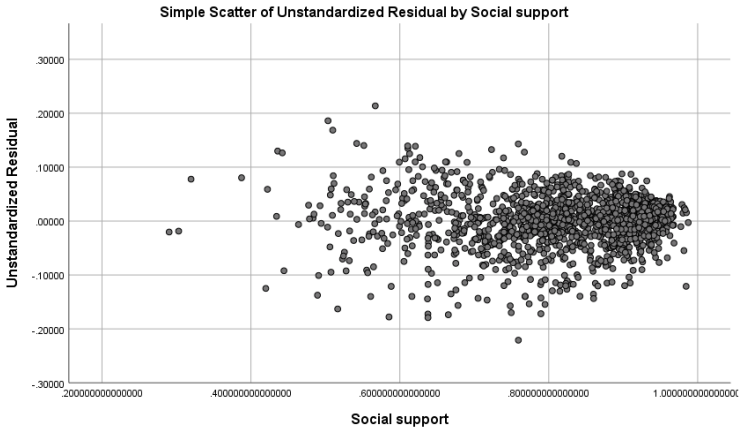




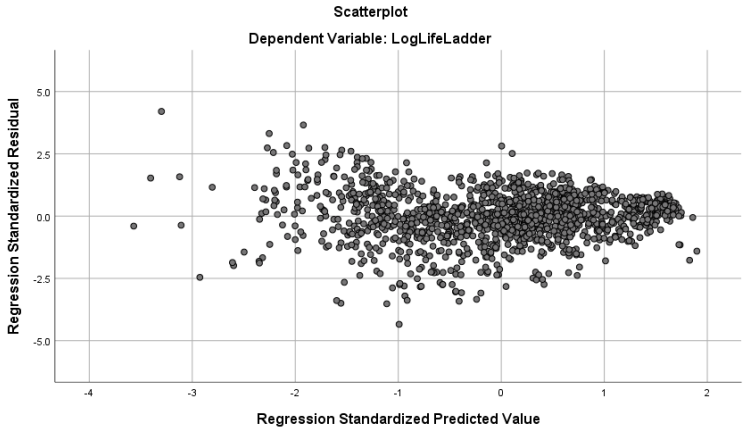






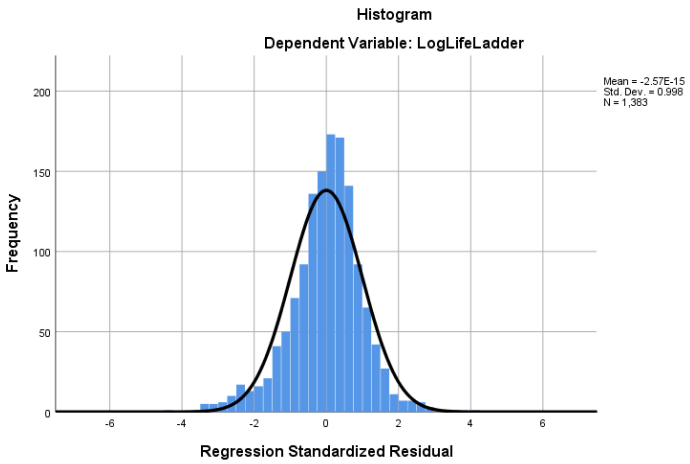


- **Residuals Vs. Fits Plot :**
  Fulfilling Residuals Vs. Fits Plot assumption required Log transformations of the Response Variable(**LifeLadder** to **LogLifeLadder**) because when we plotted the normal residual Vs. Fits plot our assumption of variance being equal of various error terms seemed violated.

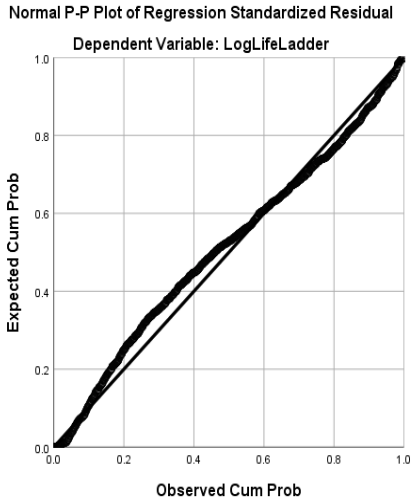


Error Terms

- **Normally Distributed**
  The third assumption is of the linear regression model is that the error terms are normally distributed.



- **P-P Plot**



Now we have the Multiple Linear Regression, to estimate the relationship between these factors(Independent variables) and the life ladder(dependent variable).

Results of Tests on Multiple Linear Regression :

Model Summary :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.856[a]	.732	.731	.11137355	.732	843.615	6	1854	.000

a. Predictors: (Constant), Perceptions_of_Corruption, Social_Support, Generosity, Freedom, Healthy_life, Log_GDP

In this table **R** represents the value the Multiple Correlation Coefficient and it can be measured as one of the quality of prediction of the dependent variable in this case it is 0.856 which is fairly a good level of prediction.

The "**R Square**" column represents the R^2 value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables is 0.732. Where as Adjusted R^2 is 0.731 and would be prefer rather than R^2 because R^2 can be manually increased by adding more and more variables whereas R^2 adjusted only considers those variables that are actually useful for the model and penalize those which are not.

Statistical Significance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	62.786	6	10.464	843.615	.000[b]
	Residual	22.997	1854	.012		
	Total	85.783	1860			

a. Dependent Variable: Log_LifeLadder

b. Predictors: (Constant), Perceptions_of_Corruption, Social_Support, Generosity, Freedom, Healthy_life, Log_GDP

The *F*-ratio in the **ANOVA** table (see below) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, $F(5, 1377) = 843.615$, $p < .05$ (i.e., the regression model is a good fit of the data).

Estimated model coefficients

Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. And here we can see we get all of our undertaken variables contribution in predicting values of ladder score that are 0.064(Log_GDP), 0.478(Social Support) ,0.006(Healthy Life Expectancy), 0.213(Freedom to make life choices) ,0.106(Generosity) ,-0.058(Perceptions of corruption).

Statistical significance of the independent variables

You can test for the statistical significance of each of the independent variables. This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population. If $p < .05$, you can conclude that the coefficients are statistically significantly

different to 0 (zero). The *t*-value and corresponding *p*-value are located in the "t" and "Sig." columns and we can see that all of our variables are Satistically Significant(p-value < 0.05).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.188	.033		5.700	.000		
	Log_GDP	.064	.005	.345	12.868	.000	.201	4.979
	Social_Support	.478	.032	.270	15.103	.000	.453	2.208
	Healthy_life	.006	.001	.216	9.054	.000	.253	3.950
	Freedom	.213	.023	.140	9.165	.000	.619	1.616
	Generosity	.106	.017	.079	6.060	.000	.841	1.189
	Perceptions_of_Corruption	-.058	.017	-.050	-3.456	.001	.689	1.451

a. Dependent Variable: Log_LifeLadder

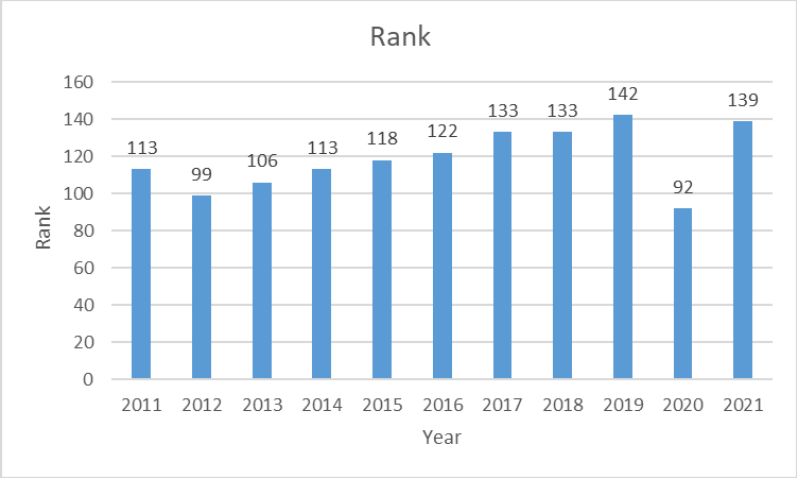
After putting it all together we get :

A multiple regression was run to predict LifeLadder from Social Support, Healthy Life Expectancy at birth, Freedom to make life choices, Generosity, Perception of corruption, Log GDP. These variables statistically significantly predicted LogLifeLadder, $F(5, 1377) = 843.615$, $p < .0005$, $R^2 = .732$.

Section D: India

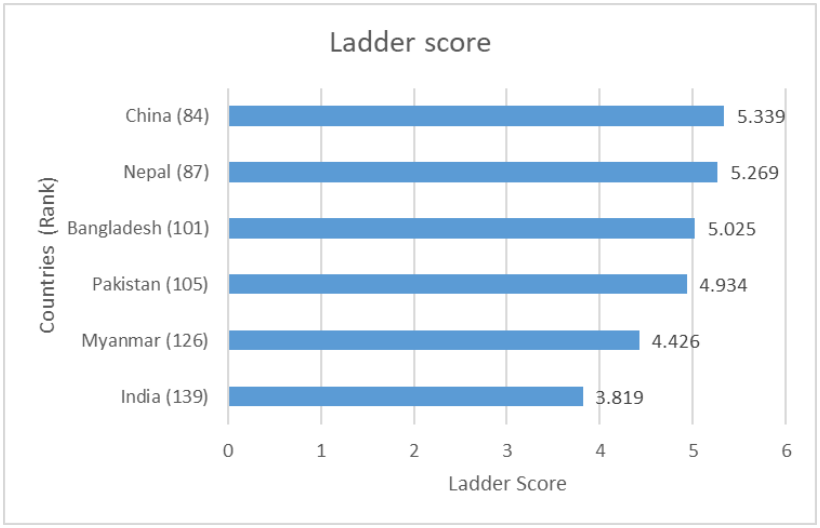
In this section, we mainly focused on India and its neighbouring countries. We have analyzed the scenario of India in accordance with different factors and compared India with other countries.

1) Rank of India in past 10 years



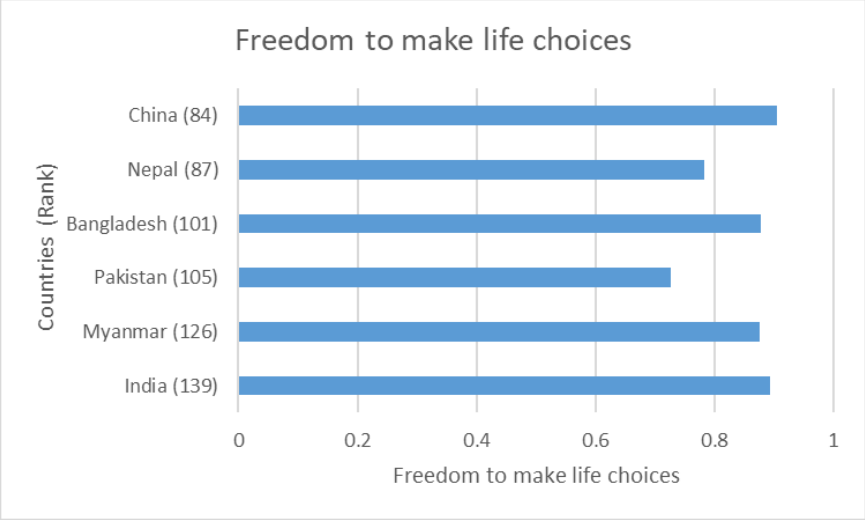
From the data, we can observe that rank of India decrease in 2012 in comparison to 2011 by 24 units and then follows an increasing pattern. There is a sudden down jump in 2020 to 92 from 142 in previous year followed by an upside jump to 139 in 2021.

2) Comparison of Happiness Score of India with it's neighbouring countries

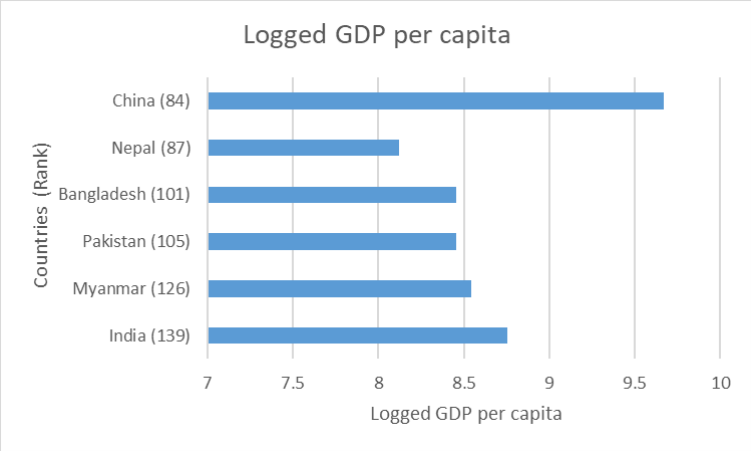


From the data, we can observe the ladder score of rank of different countries, can clearly see india has the least value of ladder score that is 3.819 whereas the china has the highest one that is 5.339

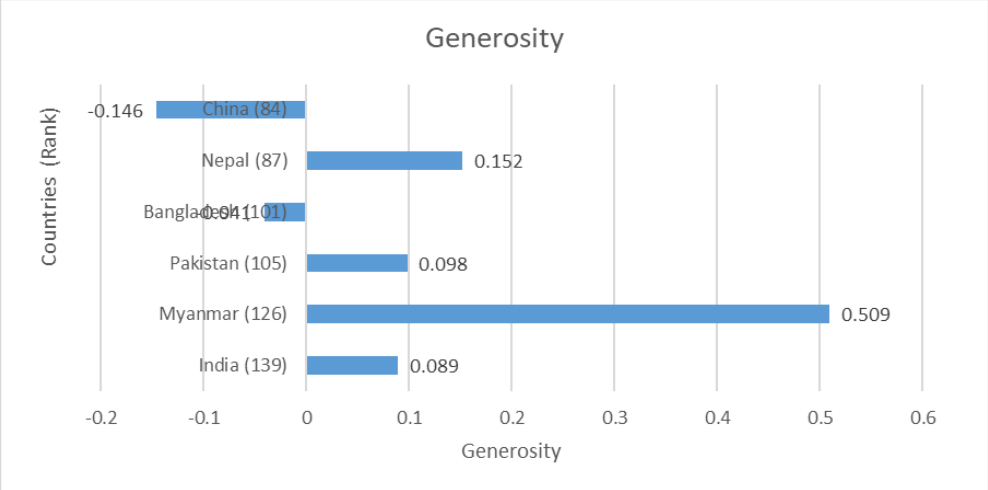
3) Comparison of different factors of India with it's neighbouring countries



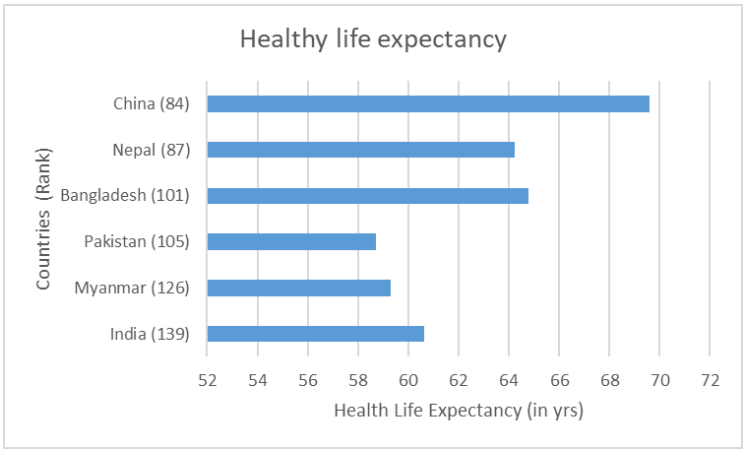
From the data, we can observe that the the rank of Pakistan is lowest in freedom to make choices score whereas the china's rank is highest one .India is having score less than the china but leading other countries like Myanmar, Nepal,etc.



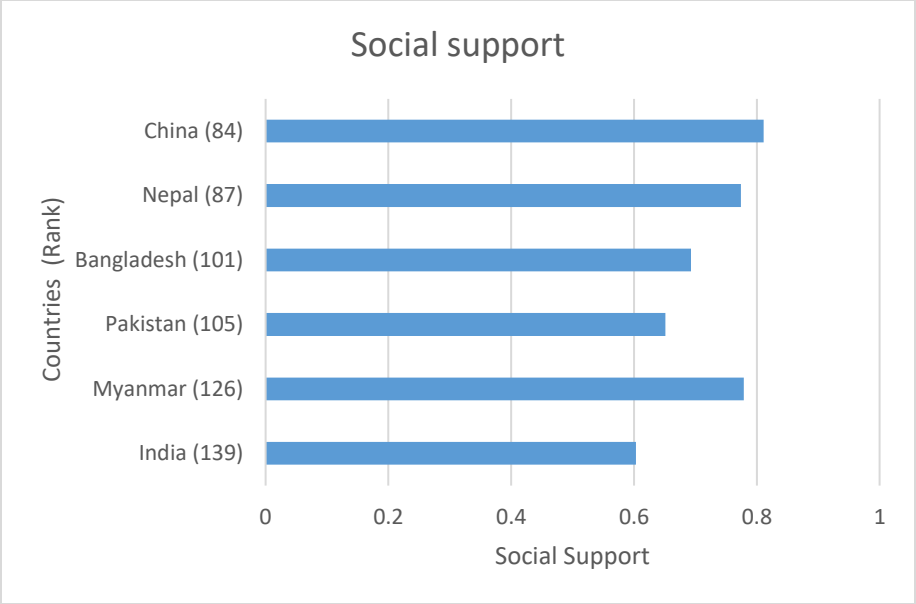
From the data, we can observe that the the rank of Nepal is lowest in logged gpd score that is having score about 8.2 whereas the china is leading among neighbouring countries in terms of log gdp having score more 9.5 out of 10 .India is having score between 8.5 and 9 that is than the china but leading all other neighbouring countries like Bangladesh, Myanmar, nepal , Pakistan.



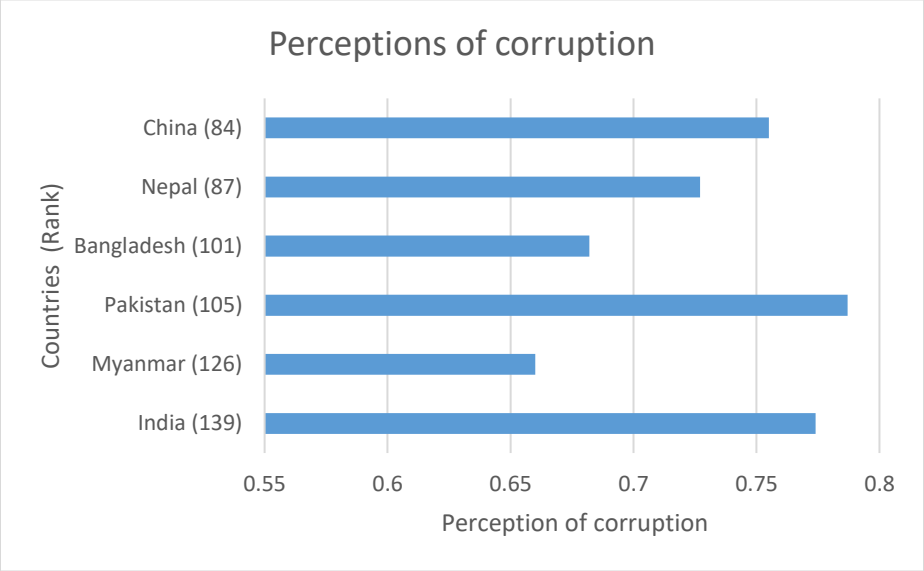
We can see generosity score of china and Bangladesh is negative that -0.146 and -0.041. Myanmar is having the highest value of generosity score that is 0 .509.



From the data, we can observe that the the rank of Pakistan is lowest in Healthy life expectancy score that is having score about 58 whereas the china is leading among neighbouring countries in terms of healthy life expectancy having score more than 65.India is having score between 60 and 65 that is less than the china, Nepal and Bangladesh but leading few neighbouring countries like Pakistan and Myanmar.



We can observe that the Social support score which is one of the important factor of happiness index, from the graph it is clearly shown Myanmar is leading in this factor over other countries and India is facing the lowest score that is 0.603.



From the data, we can observe that the the rank of Myanmar is lowest in perception of corruption score that is having score about 0.65 whereas the Pakistan is leading among neighbouring countries in terms of Perception of corruption having score more than 0.8 .India is having score between 0.75 and 0.8 that is less than the Pakistan but leading other neighbouring countries.

Conclusion

1. Money can't buy happiness. We all have heard this sentence but the data shows a different story.
   The results speak rich countries/regions are happier than the poorer ones.
2. Money is an important factor for happiness, but it is not the only factor for happiness. From correlation and multiple linear regression, social support, healthy life expectancy, and freedom to make life choices are also highly correlated.
3. Not only different countries, but different regions have different happiness levels too. Developed regions like America have higher happiness scores as compared to the least developed region like Africa.
4. To be a country with the best happiness score, top and important factors should at least have values as follows:
   Healthy life expectancy at birth should be around 70
   Logged GDP score should be around 10 t0 11
   The social support score should be above 0.9.
5. Health life expectancy, logged GDP and social support have a strong correlation, and out of that healthy life expectancy and logged GDP have significantly improved in the last 5 years. So now there is a need to have a higher value for social support for a better happiness score.
6. In the correlation paired t-test and Kruskal Wallis test, we observe that healthy life expectancy is having high correlation and it is increasing in the last few years. The health sector is developing. So, it is also leading to an increase in world happiness scores.
7. Now, we come to the scenario of India, we can observe that to improve the world happiness score of India as compared to other countries, the values should be in a manner as log GDP should be around 10 and social support is around 0.603 on an average and it should at least reach 0.9 and healthy life expectancy should be around 65-70. Freedom score is good in the case of India as compared to other countries as we have seen this earlier in the graph also. Perception of corruption is high even after some government measures, to have a good score corruption score should at least falls to 0.2. Generosity in India is less but we have seen earlier topmost countries have similar generosity scores, also it has the least correlation with happiness score so it does not affect the happiness score much.