TU Dortmund
Department of Statistics

**Case Studies I**

# Project 1: Descriptive data analysis

## Goodbye Germany - A selected analysis of the World Happiness Report

Lecturers:
Prof. Dr. Paul Bürkner
Dr. Daniel Habermann
Lars Kühmichel

Author: Rahul Ramesh Vishwkarma
Matriculation Number.: 236862

Group Number.: 17

Group members: Md Jubirul Alam, Rafay Maqsood,
Syed Hassan Saqlain Tayyab

October 11, 2024

# Contents

# 1 Introduction

The happiness of the citizen is a fundamental goal of every nation. Since this notably helps to understand their well-being and also assists governments in deciding on future policies in favour of the quality life of an individual. Considering the development of a society, the study of individuals' happiness has received a lot of importance in the last two decades. This study particularly uses descriptive statistical analysis tools to analyse the variables that measure subjective well-being (life satisfaction), a country's GDP (gross domestic product), and the freedom to make their own choices in life.

The data set is a small part of the original data set from the World Happiness Report (WHR) from 2005 to 2018, covering 165 countries. First, we consider the variable for life satisfaction. This variable provides valuable information and helps to determine the country with the most or least satisfied citizens. Additionally, the variables i.e. a country's GDP, Freedom to make own choices, that are responsible for an individual's happiness are discussed. Secondly, the measure of dispersion of income of a country (Gini index.) is discussed to know the variability of income among the countries.

Later, we consider the potential problems that are responsible for the missing values in the data set and discuss how to manage these values. Then we also talk about the variables that are crucial for Asian food quality in a country. Here, three countries, namely Germany, China, and Hong Kong S.A.R. of China(for short, Hong Kong), are compared for food quality. Lastly, the trends of a few variables in different countries over the period are compared. In the second section, the data set is described and the variables are explained. The third section is dedicated to statistical methods that we use as a toolbox for the fourth section. In the fourth section, we use descriptive statistical analysis along with scatter plots, box plots, line plots, and some discrete measures of statistics to explain variables' descriptive properties and relations between them. In the fifth section, the results from the analysis are explained, summarized, and given a vision for possible further analysis.

# 2 Problem statement

## 2.1 Data set and its reliability

The data set is a small extraction from the World Happiness Report (2019)(Helliwell (2019)), considering the years from 2005 to 2018. The data was collected with various methods, such as surveys from the GULLUP World Poll (GWP), estimations from the World Bank's Global Economic Prospects, government statistical agencies, and World Bank indicators (WDI). Originally, the data set consisted of 1704 observations for 165 countries from 2005 to 2018, with 26 variables. However, we consider the data set available only for the last year of these countries then removing missing values from it, which counts only 157 countries. Then we consider the variables `Life.Ladder`, `Log.GDP.per.capita`, and `Life.Freedom` for our first analysis, which are all numeric apart from `Country.name` and `Year`. Since almost 2% of given countries are missing in the complete data set, we can still get good estimations for our expected result. We also extracted the variable `avg.GINI` for the last year of each country, removing the missing values, and we left with 139 countries for the second analysis. Since this data is an estimate from the World Bank Country Department and government statistical agencies, we still rely on this available data for the second analysis.

## 2.2 Explanation of variables

For the variable `Life.Ladder`, citizens were asked to evaluate their lives on a scale of numbers from 0 to 10 (corresponding from the worst to the best) at the current moment. Then the average of these survey results is taken as a single response for the country. The `Log.GDP` is the log of GDP per capitta. GDP per capita is a monetary quantification of a country, calculated by summing up the incomes from all the reliable sources without considering any subsidies, divided by the population. The data is not available for 2018, but it is estimated from available time series data from OECD Economics and the World Bank's Global Economic Prospects. Similar to the `Life.Ladder`, the variable `Life.Freedom` is based on the national average of the question in regard to the current satisfaction level of the freedom they have to make their own decisions about their lives. Citizens were surveyed for household income by government statistical agencies and

World Bank country departments in local currency. The income was later converted into international dollars as a common measure among all countries. Then the Gini index is calculated for each country, and this quantifies the inequality of income among the countries. However, we use the average of the Gini index between the years 2000 and 2016 as an estimation for the Gini index for 2017–18. This variable in our data set is named `avg.gini`.

## 2.3 Aim of this Project

The aim of the project is to analyse the data with descriptive statistical tools. In the first part of the analysis, we figure out the best or worst country in 2018 with respect to the variable `Life.Ladder`. We also examine the pair-wise dependencies (linear dependencies) among the variables. Then we look at the countries with the lowest and highest `avg.gini` (Gini index). In the fourth section of the analysis, we compare the variables' values for the assessment of the quality of Asian food. In the final analysis, we discuss the improvements and deteriorations over the years in the few countries' well-being and in economy.

# 3 Statistical methods

## 3.1 Mean

Let $X$ be a real valued random variable and $x_1, x_2, \ldots, x_n$ are its realizations (observations) in the sample (data) of size $n$ (Akinkunmi (2019)), then the arithmetic mean (mean, $\bar{x}$) is defined as,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

i.e., taking the sum of realizations divided by the sample size. The mean is affected by the extreme values, and it is a single value that describes the centredness of the observations (realizations).

## 3.2 Median and q-quantiles

For ordinal data, q-quantiles measure the dispersion in the data by dividing the data into $q$ (nearly)equal-sized groups. Here, we consider 4-quartiles, which are $Q_1$ (i.e., $\frac{(n+1)}{4}th$ value in the data set), $Q_2$ (also known as $Median$), and $Q_3$ (i.e., $\frac{3(n+1)}{4}th$ value in the data set), which divide the data into four nearly equal parts. $Q_1$ and $Q_3$ form ($IQR$) (inter-quartile range) by the formula $IQR = (Q_3 - Q_1)$, which consists of 50% of the data that resides in the mid(Paul Newbold (2019)). The interquartile range is also used as a dispersion measure. Similarly, $Median$ divides the data into two equal parts, i.e., 50%-50%. As mentioned before, for the mean, the median is not affected by the extreme values. For the sorted data $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ from the data $x_1, x_2, \ldots, x_n$, the $Median$ is given by(Hay-Jahans (2019)),

$$Median = \begin{cases} \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{if } n \text{ is even} \\ x_{(\frac{n+1}{2})}, & \text{if } n \text{ is odd} \end{cases} \tag{2}$$

## 3.3 Gini index

For a categorical random variable $X$ with categories, $1, 2, ..., g$ and if $n_k$ be the number of observations from the category $k$; $k$ =$1, 2, ..., g$ from a total number of observations $n$ (James G. (2023)). Then the Gini Index ($G$) is defined as,

$$G = \sum_{k=1}^{g} \frac{n_k}{n}(1 - \frac{n_k}{n}) \tag{3}$$

Note that here, $n = \sum_{k=1}^{g} n_k$. $G$ accounts for the total variability present in the data over all $g$ categories.

## 3.4 Correlation coefficient

Let $X$ and $Y$ be the random variables with respective realizations $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ then the coefficient of correlation is given by(Akinkunmi (2019)),

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

Where, $\bar{x}$ and $\bar{y}$ are the mean of $X$ and $Y$, respectively.

The $r_{XY}$ is known as Pearson's correlation coefficient and quantifies the level of linear relationship among two variables. Also $r_{XY}$ lies between $-1$ and 1, when $r_{XY} = 0$, it implies no linear relationship between the variables/ However, it is possible that they can be non-linearly related. If $-1 \leq r_{XY} < 0$ implies negative relationship, while $0 < r_{XY} \leq 1$ implies positive. The $r_{XY}$ is independent of the unit used for measuring $X$ and $Y$.

## 3.5 Types of plots

### 3.5.1 Box plots and Whiskers

The box plot is a graphical representation of data to understand dispersion via quartiles and extreme values(Akinkunmi (2019)). In the box plot diagram (see Figure 2), the box is the inter-quartile range ($IQR$) with median being a line inside the box. The first whisker line is drawn from $x_{min}$(Minimum), being the smallest value within the distance of 1.5 $IQR$, to $Q1$. Similarly, the second line drawn from $x_{max}$(Maximum), being the largest value within the distance of 1.5 $IQR$, to $Q3$. The points lie outside these whiskers range are called outliers. The x-axis is usually used for showing measurement units, whereas the y-axis is the range of values in the data. Typically, box plots are used to compare different sets of data.

### 3.5.2 Scatter plots

For a numeric vector observation $(x, y)$ of a bivariate variable $(X, Y)$ (Hay-Jahans (2019)), the scatter plot is a two-dimensional graph representing the data points in the graph (see Figure 1(a)). The graph is useful to observe the relationship between two variables. Each point in the plot is a tuple $(x, y)$, where $x$ and $y$ correspond to $X$ and $Y$. If points appear to be co-linear with a positive slope, then the relationship is positive, i.e., $r_{XY} > 0$. In contrast to this, if the slope is negative, it implies a negative relationship ($r_{XY} < 0$). The best thing about the scatter plots is that the non-linear relationship among the variables can also be seen; for example, the relationship can also be circular and given by $X^2 + Y^2 = r^2$, where $r > 0$.

### 3.5.3 Line plot

The line charts are used to plot data points $(x, y)$, where the variable $x$ is sorted and then the $(x, y)$'s are plotted and are connected by a straight line, consequently from left to right(Akinkunmi (2019)). The variable $x$ is often a time, for example, in days, months, years, etc. This also called as time plot or time series. To know the changes in variables over time and trend, line charts are best. For an example, see Figures 4(a) and 4(b).

## 3.6 Software

For our analysis, we used the statistical programming language R version 4.2.3 (2023-03-15 ucrt)(R Development Core Team (2020)). Apart from the base R package, several other packages are also used, such as *dplyr* [version 1.1.1](Wickham *et al.* (2021)) for data manipulation, and *ggplot* [version 3.4.4](Hadley Wickham ORCID (2020)) for visualization.

# 4 Statistical analysis

Now, all the tools explained in the former section are used to derive valuable information from the data set. Firstly, from the summary table, properties of the distribution of a variable are derived. Then a linear relationship is explained from scatter plots and quantified by a measure of correlation. Later, the dispersion of income is described via a box plot, and the existence and manipulation of missing data values are justified. Box plots are then used for Asian food quality comparisons among countries. Additionally, the direction of variables over time is explained with the help of line plots.

## 4.1 Analysis of life satisfaction score

Firstly, from Table 1, on average, people in the world have 5.461 `Life.Ladder` (life satisfaction score), which is quite similar to the median value, which does not differ by 0.011. However, only 25% of the world's population has a `Life.Ladder` value within the range of 6.235 to 7.858. In the extreme cases, Finland is at the apex of a satisfactory lifestyle, followed by Denmark and Switzerland in 2018. Additionally, the trend of `Life.Ladder` has shown consistently positive over the course of the period for Finland (see Figure 6). In contrast to this result, Afghanistan has declined drastically in terms of quality of life in the meantime, and its score was the lowest in 2018 and was led by the countries, South Sudan and Yemen.

Table 1: Descriptive Summary of `Life.Ladder`

| Variable | Mean | Min. | 1st Qu. | Median | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| `Life.Ladder` | 5.461 | 2.694 | 4.659 | 5.472 | 6.235 | 7.858 |

### Linear dependencies between variables

From Figure 1, the variable `Life.Ladder` is strongly correlated with the variables `Log.GDP` and `Life.Freedom`. Also, there is a strong indication of the linear dependence of `Log.GDP` with a correlation of 0.75. This also depicts that the higher the GDP of a country, the more likely its citizens are to be happy. In the case of `Life.Freedom`, the correlation is 0.5 but significant for the linear relationship. It is quite convincing that individuals who have the freedom to make their own decisions regarding their lives are the ones who are more satisfied with their lives.

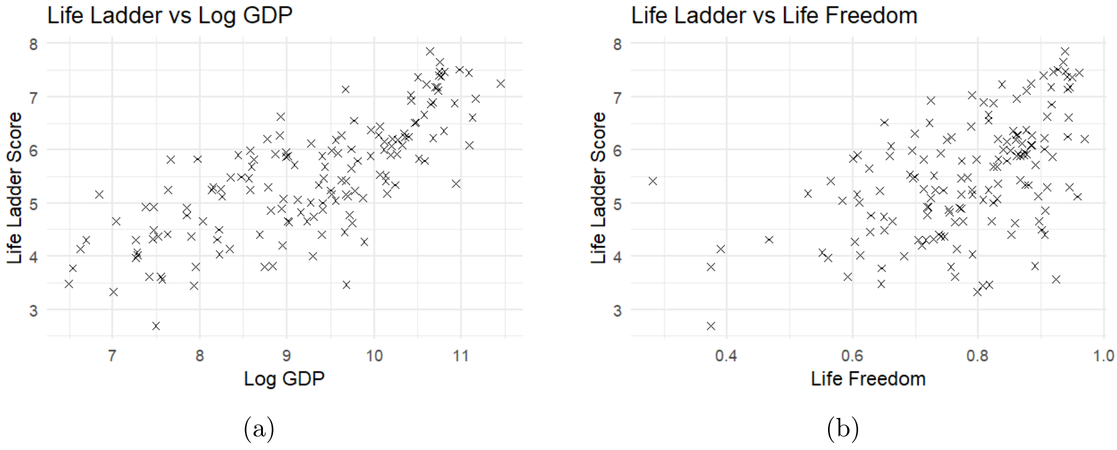


Figure 1: Scatter plot of `Life.Ladder` with respect to `Log.GDP` and `Life.Freedom`

## 4.2 Gini Index of Households Income

Gini indices for the last year of all countries are estimated from 2000 to 2016. From Figure 2, about 50% of countries (average Gini index) fall under the range of 0.211 and 0.434 for `avg.gini`. The country Azerbaijan in 2018 had the lowest `avg.gini` (0.211), which implies the incomes of the citizens are distributed uniformly. However, among the ten countries with the lowest `avg.gini`, eight are from Europe. In contrast to that, the three South African countries, namely Botswana, South Africa, and Namibia, have the highest `avg.gini`. This also means that the variation among the household incomes of these countries is too high.

## 4.3 Handling the missing values

For a few countries, there is an unavailability of data for `Life.Ladder` in the years 2016–18. For these missing data, the values are substituted from the last year's available

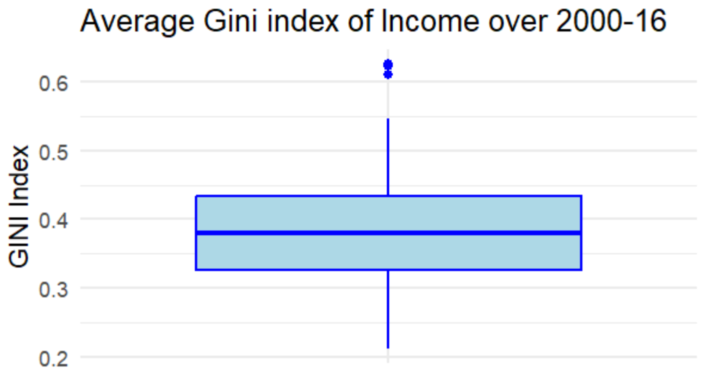


Figure 2: Box plot of `avg.gini`

data, but only up to 3 years. This imputation also makes sense, as the happiness of a country is a long term smooth process. Since, for longer years back we have many missing values and this might be due to the unavailability of equipment back in the time or people's unawareness about the confidentiality of the survey. For `Log.GDP` values, Somalia and Taiwan are missing for the years 2017–18, and these values are imputed by The World Factbook via the PPP (Purchasing Power Parities) method. The PPP is a measure of a country's monetory value by comparing the relative price of the same product in two different countries with their own currency. Also, in most of the countries, relevant questions were not asked by GWP in the survey.

In our analysis, As the given data is sensitive to country's economy and culture, we tried to not use the imputed data values in this analysis. At the same time, we aim to use the latest available data for each country in the data set. Most countries do have data for either 2018 or 2017 or both. However, for `Log.GDP`, we estimated it from the last 16 years of data by taking the average over this period. This we use as an estimated `Log.GDP` for the year 2018.

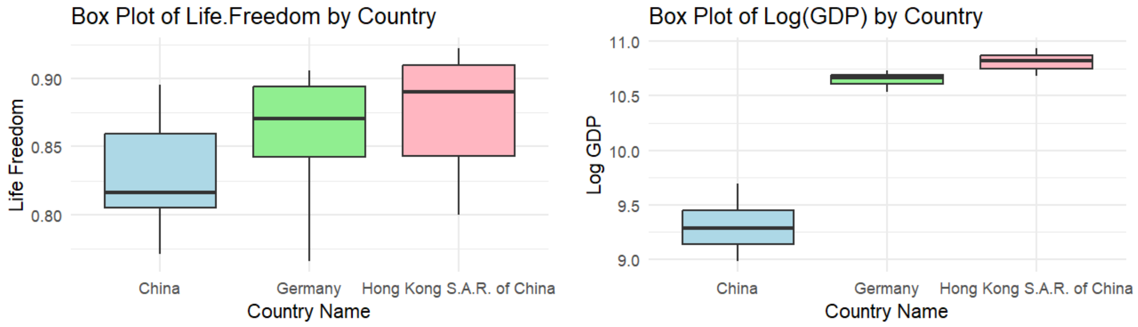
After data filtration for the latest available data for the countries, we lost only 14 out of 165 countries for the analysis of the variables `Life.Freedom` and `Log.GDP`.

## 4.4 Asian Food comparison among countries

Since the country's economy helps to improve the infrastructure, this also leads to tourism, hotels, and restaurants. Hence, the variable `Log.GDP` is considered the influencing factor for food quality here. In addition to that, the freedom to make one's own choices also influences one's ambition to become a chef or a cook, which helps to keep enthusiasm in a person's profession for the work. Hence, `Life.Freedom` is likely to affect a country's food taste and quality. In this section, we used two variables, `Life.Freedom` and `Log.GDP` to quantify the Asian food quality in Germany, China, and Hong Kong.

Over the course of the period, It can easily be seen from Figure 3(a), that not only has Hong Kong shown the highest variation in `Life.Freedom` values, but also half of the time Hong Kong has the higher value than 0.8904, which is astonishing. Similarly, Germany is not far behind. Contrary to this, China is far behind both countries, and not more than one-fourth of the time it has a higher value than this limit.

For the variable `Log.GDP`, Germany and Hong Kong seem to be quite stable over these years from Figure 3(b), but Hong Kong is much more appealing; it has gone beyond the limit of 10.75 for three-fourths of the time. On the other hand, Germany has not gone more than one-fourth of the time beyond this limit. In the meantime, China has never reached this limit, not even once.



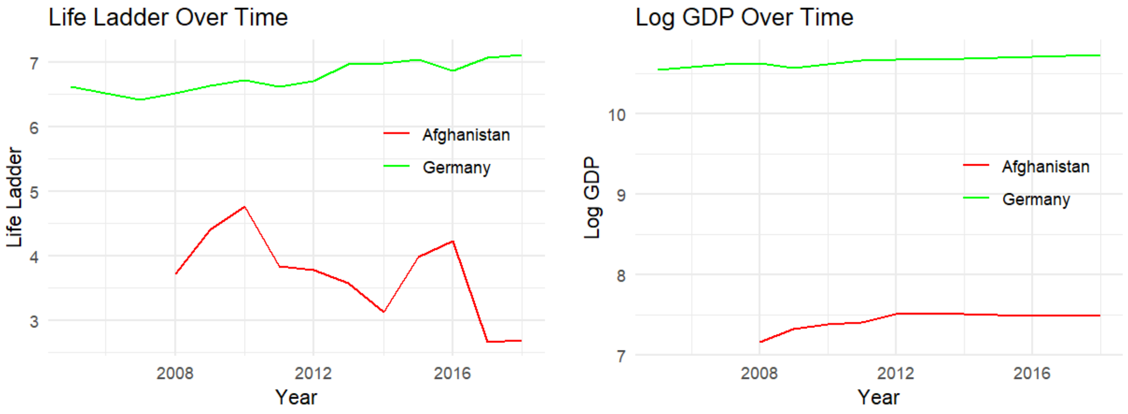
(a) (b)

Figure 3: Distribution of `Life.Freedom` and `Log.GDP` for China, Germany and Hong Kong

## 4.5 Trend of variables for Germany, Afghanistan and India

Over a long period of time, Figure 4(a) explains that Germany has been improving in terms of the life satisfaction scale, i.e., `Life.Ladder`. The growth of `Life.Ladder` seems to be pretty constant. While the evaluation of life satisfaction has deteriorated over the same period, with some fluctuation in Afghanistan. Regardless of these fluctuations, `Life.Ladder` has a negative trend. In similar fashion from Figure 4(b), the economic growth, i.e., `Log.GDP`, is increasing consistently during the whole period for Germany. On the other side, in Afghanistan, it appears to be growing tremendously. Nevertheless, the growth became stable after the year 2012.

From the first task, the variables `Life.Ladder` and `Log.GDP` are positively correlated, with a correlation coefficient of 0.75. But after considering these two line plots of India in Figure 5 (a) and (b), the relation seems to be counterintuitive. For India, these variables are appeared to be negative correlated. Since, for a high value of `Log.GDP` and the value of `Life.Ladder` is low. This can be easily seen from Figure 6.



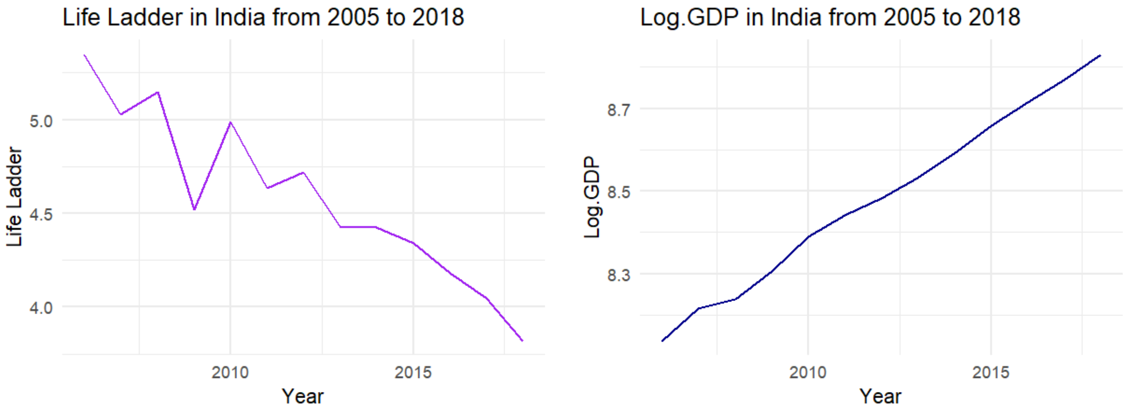
(a)

(b)

Figure 4: Trends of `Life.Ladder` and `Log.GDP` of Germany and Afghanistan

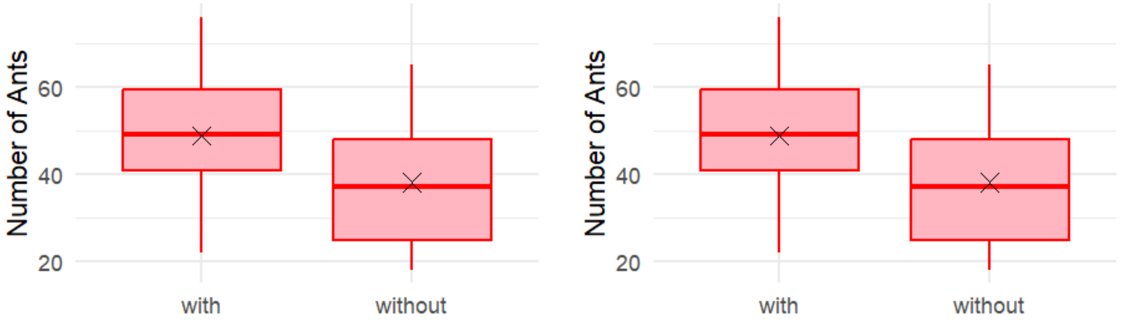
# 5 Summary

The descriptive analysis is conducted on an extract of a data set from the World Happiness Report (WHR) 2019. The goal of the project is to understand the well-being of



(a) (b)

Figure 5: Trend of `Life.Ladder` and `Log.GDP` of India



(a) (b)

Figure 6: Distribution of `Life.Ladder` and `Log.GDP` of India

citizens via other variables that affect citizens' happiness. Also, a comparison of Asian food quality among countries are discussed. Further, the long term change in the lifestyle and economy among few countries are considered.

Firstly from `Life.Ladder`, Finland is the happiest country in the world and tends to be on top in the future. However, in Afghanistan, quality of life is at the worst stage. The correlation coefficient suggests, a strong linear dependency between life satisfaction and economic growth. Similarly, life satisfaction and freedom to make own decisions, are quite significantly dependent.

Secondly, the Gini index and its box plot are used to show inequalities in wages among the countries. The household incomes are very similar for the first eight European countries. While, Botswana, South Africa, and Namibia, income distribution is scattered. More than half of the countries do not have an average Gini index, implies a need for an improvement. Then, the possible sources of missing values along with the imputation methods are considered. Then, the Asian food quality among Germany, China, and Hong Kong is compared with box plots. The results were more favourable to Hong Kong, where China is far behind.

Finally, life satisfaction and economy have been improved consistently over the time in Germany. While, Afghanistan's economy has not grown since 2012 with happiness score is also getting worse. In India, counterintuitive relationship among these two variable has seen, which turns out to be negatively correlated.

It concludes, the low happiness score suggest a need for improvements in the economy and citizens' freedom. It is reasonable to say, Germany, with constant growth in economy and happiness, will continue in the future. For Afghanistan, the government can make better policies by considering economy, people's happiness and freedom. In addition, other variables such as generosity among the people, perceptions of corruption and social support from friends and family etc., to determine the influence on happiness can be used. One can consider the Gini index and happiness score, for further analysis.

# Bibliography

Akinkunmi, Mustapha. 2019. *Introduction to Statistics Using R.* Morgan and Claypool Publishers.

Hadley Wickham ORCID, Winston Chang. 2020. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* R package version 3.5.1.

Hay-Jahans, Christopher. 2019. *R Companion to Elementary Applied Statistics.* Chapman  Hall.

Helliwell, J., Huang H. R.  Wang S. 2019. World Happiness Report 2019.

James G., Witten D., Hastie T. Tibshirani R. 2023. *An Introduction to Statistical Learning: With Applications in R.* Springer.

Paul Newbold, William Carlson, Betty Thorne. 2019. *Statistics for Business and Economics.* Chapman  Hall.

R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Wickham, Hadley, François, Romain, Henry, Lionel, & Müller, Kirill. 2021. *dplyr: A Grammar of Data Manipulation.* R package version 1.0.6.

# Appendix

## A Additional figures

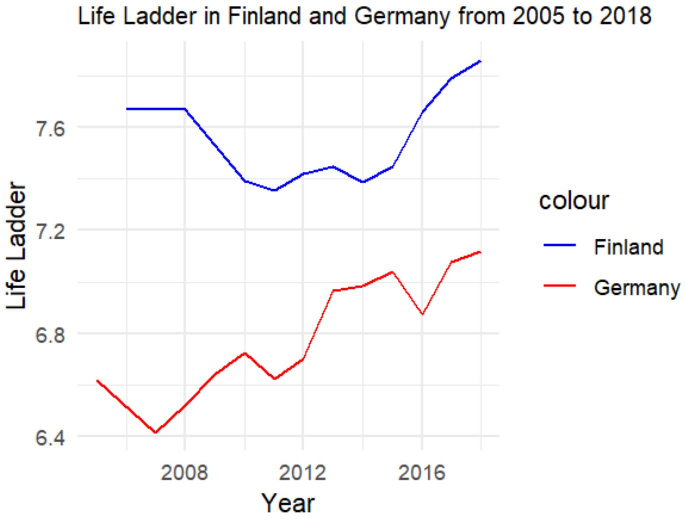


Figure 7: Line plot of `Life.Ladder` of India

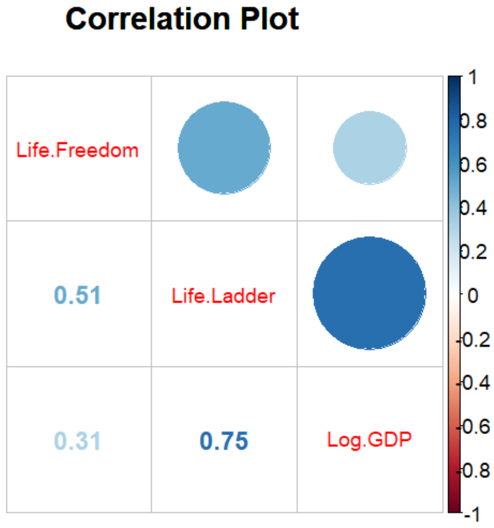


Figure 8: Correlation Plot of `Life.Ladder`, `Log.GDP` and `Life.Freedom`