A

MINOR PROJECT REPORT

on

UNDERSTANDING THE REGIONAL DIFFERENCES IN WORLD HAPPINESS INDEX

BE (AI & DS) VI Semester

By

Mayush Timmapuram (160120771039)

Rithvik Ramdas (160120771043)

Under the guidance of

Dr. T Prathima

Assistant Professor

IT Department

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY (A)

(Affiliated to Osmania University; Accredited by NBA(AICTE) and NAAC(UGC), ISO Certified 9001:2015)

KOKAPET(V), GANDIPET(M), RR District HYDERABAD - 75

Website: www.cbit.ac.in

2022-2023

CERTIFICATE

This is to certify that the project work entitled **"Understanding the Regional Differences in World Happiness Index"** submitted to CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY, in partial fulfilment of the requirements for the completion of Minor Project – II of VI Semester B.E. in Artificial Intelligence and Data Science, during the Academic Year 2022-2023, is a record of original work done by **Mayush Timmapuram (160120771039) and Rithvik Ramdas (160120771043)** during the period of study in the Department of AI & DS, CBIT, HYDERABAD, under our guidance.

Project Guide
Dr. T Prathima
Assistant Professor, Dept. of IT,
CBIT, Hyderabad.

Head of the Department
Dr K. Ramana
Associate Professor, Dept. of AI&DS,
CBIT, Hyderabad.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to **Dr. T Prathima,** our project guide, for her invaluable guidance and constant support, along with her capable instruction and persistent encouragement.

We are grateful to our Head of Department, **Dr K. Ramana**, for his steady support and for the provision of every resource required for the completion of this project.

We would like to take this opportunity to thank our Principal, **Dr P. Ravinder Reddy**, as well as the Management of the Institute, for having designed an excellent learning atmosphere.

Our thanks are due to all members of the staff and our lab assistants for providing us with the help required to carry out the groundwork of this project.

ABSTRACT

The World Happiness Report is a publication of the Sustainable Development Solutions Network, powered by the Gallup World Poll data. The World Happiness Report reflects a worldwide demand for more attention to happiness and well-being as criteria for government policy. It reviews the state of happiness in the world today and shows how the science of happiness explains personal and national variations in happiness.

The project involves collecting and cleaning the data, followed by exploratory data analysis to understand the relationships between variables. We will then use feature selection techniques to identify the most significant factors affecting a country's happiness level. Next, we performed exploratory data analysis to understand the relationships between variables and identified the most crucial features that affect a country's happiness level. Using this information, we then build a predictive model using machine learning algorithms such as regression and decision trees to predict a country's happiness score based on its quality-of-life factors such as - economic, social, and health indicators.

The project aims to provide insights into the factors that contribute to a country's happiness and provide a useful tool for policymakers and governments to develop strategies to improve the well-being of their citizens. Overall, this project demonstrates the potential of data science in providing insights into complex societal issues such as happiness and well-being.

CONTENTS

S. No	*Topic*	*Page No.*
	ACKNOWLEDGEMENT	iii
	ABSTRACT	iv
	CONTENTS	v
	LIST OF FIGURES	vi
	LIST OF TABLES	vii
1	**INTRODUCTION**	1
2	**LITERATURE SURVEY**	2
3	**DATASET DESCRIPTION**	7
4	**METHODOLOGY**	
	4.1 Preprocessing Tasks	10
5	**RESULTS AND COMPARISON**	
	5.1 Top 10 Countries: Mean Happiness Score Trends	17
	5.2 Top 10 Most Progressive Countries	18
	5.3 Regional Analysis on Happiness Scores Of Countries	21
	5.4 Finding Correlation Using Correlation Matrix Heatmap	22
	5.5 Multiple Linear Regression	23
	5.6 Support Vector Regression	24
	5.7 Decision Tree Regression for Predicting Happiness Scores	26
	5.8 KNN Classifier for Predicting Happiness Levels	28
	5.9 Naïve Bayes Classification Model	29
	5.10 Feature Importance	31
6	**CONCLUSION AND FUTURE SCOPE**	
	6.1 Conclusion	33
	6.2 Future Scope	33
	BIBLIOGRAPHY	35

LIST OF FIGURES

Figure No.	*Name of The Figure*	*Page No.*
5.1	Mean Happiness Score Trends	17
5.2	Scatter plot of Top 10 most progressive countries from 2015-2022	19
5.3	Trends of Top 10 most Progressive Countries from 2015-2022	20
5.4	Boxplot for regional analysis of happiness scores	21
5.5	Correlation Heatmap	22
5.6	Multiple Linear Regression	24
5.7	Support Vector Regression (SVR)	26
5.8	Scatter plot of Decision Tree Regression	27
5.9	Decision Tree	28
5.10	KNN Model	29
5.11	Naïve Bayes Classification Model	30
5.12	Feature Importance (KNN)	32
5.13	Feature Importance (Naïve Bayes)	32

LIST OF TABLES

Table No.	*Name of The Table*	*Page No.*
2.1	Literature Survey	2
3.1	Dataset Description	7
5.1	Multiple Linear Regression Metrics	24
5.2	Support Vector Regression Metrics	26
5.3	Decision Tree Regression Metrics	28

1. INTRODUCTION

The government of Bhutan follows a philosophy called Gross Domestic Happiness (GDH), sometimes also referred to as Gross National Happiness (GNH), in the 1970s. It is used to measure the state of well-being and the collective happiness of the population of Bhutan. Bhutan's approach to measuring happiness highlights the importance of looking beyond traditional economic indicators like GDP (Gross Domestic Product) when assessing the well-being of a population [1]. This philosophy has inspired similar efforts in other countries and regions and has contributed to a growing interest in understanding the factors that contribute to overall happiness and life satisfaction.

In this report, we examine the data behind the Global Happiness Index to better understand regional differences in happiness. Specifically, we look at the 10 most progressive countries based on their rate of increase in happiness scores from 2015 to 2022, and the last few countries with the lowest happiness scores. By examining the factors that contributed to the impressive progress of the top 10 countries, we aim to identify policy recommendations that can help other countries emulate their success.

Similarly, by analyzing the challenges faced by the few lowest ranked countries, it seeks to suggest policy solutions that can help improve the well-being of its citizens and overall well-being. Our analysis provides valuable insight into the factors that contribute to well-being and policy interventions that may promote well-being, contributing to the broader debate on measuring progress beyond GDP.

2. LITERATURE SURVEY

Table 2.1: Literature Survey

Reference No.	Paper Title	Results	Gaps
1.	Analyzing happiness index as a measure along with its parameters and strategies for improving India's rank in world happiness report. [2]	Performed inference on India's happiness index.	1. Limited scope 2. Lack of clarity in methodology 3. Limited generalizability: The study focuses on India's ranking in the World Happiness Report, and the findings may not be generalizable to other countries or regions. The paper could benefit from a comparative analysis of happiness across different countries or regions.
2.	Predicting Quality of Life using Machine Learning: case of World Happiness Index [3]	Regression tasks by using various loss functions. Purpose of this paper is to explore the performance of machine learning algorithms used in the literature for Quality-of-Life index prediction.	1. Classification models have not been used to predict data. 2. Limited explanation on the data preprocessing steps: While the authors mention that they unified the column names and dropped features that were not common to all years, they do not provide a clear explanation of how they handled outliers, or any other preprocessing steps

			that may have been applied to the data. 3. Limited comparison with other studies: A comprehensive comparison of the paper's results with other studies in the literature is not provided. 4. Feature/variable importance is not taken into account. 5. Decision tree regression using Gini index is not used as a metric to analyze the data.
3.	S. D. N. S. Meghna Chaudhary, "NETWORK LEARNING APPROACHES TO STUDY WORLD," arXiv, p. 13, 2020.	Created a knowledge graph that summarizes past literature on how the descriptors considered in the study influence happiness.	1. Limited scope and representativeness of the data. 2. Lack of comparative analysis with other statistical methods. 3. Insufficient details on the methodologies and models used. 4. Inadequate interpretation of the findings and their implications. 5. Failure to identify specific research gaps or future directions.

4.	How Happy Are We Actually? A Posetic Analysis of the World Happiness Index 2016 2019 Denmark as an Exemplary Case [4]	The present study applies an alternative ap-proach based on partial ordering methodology where all seven indicators are included. simultaneously in the analyses without any pretreatment nor weighting. It comprises of a posetic analysis of the index indicators for the years 2016–2019.	1. Inadequate treatment of the seven indicators and lack of weighting in the analysis. 2. Insufficient discussion on the implications of the partial ordering approach. 3. Limited focus on the Top 10 countries, with little exploration of the remaining countries. 4. Lack of consideration for uncertainty and potential effects on the results. 5. Incomplete examination of the reasons for incomparability and mathematical contradictions in indicator values. 6. Insufficient discussion on the practical implications and usefulness of the World Happiness Index as a tool for governments. 7. Need for further research to address uncertainty and compare the results with the original rankings.
5.	Analysis on the Relationships on the Global Distribution of	The study found that the clustering of countries based	1. Limited discussion on the specific relationships between the World

	the World Happiness Index and Selected Economic Development Indicators [5]	on economic development indicators yielded valuable insights. However, no single cluster emerged as the best performing in terms of the economic variables studied. The analysis highlighted the persistent poor performance of certain countries and emphasized the importance of addressing corruption. The findings underscored the need for policy frameworks that prioritize the well-being, education, income, and reduced corruption to enhance the quality of life for citizens.	Happiness Index (WHI) and economic development indicators. 2. Insufficient explanation of the cluster analysis methodology and its implications for the grouping of countries. 3. Lack of detailed analysis on the specific characteristics and performance of each cluster. 4. Incomplete exploration of the impact of corruption and potential strategies for addressing it.
6.	A Data Analysis of the World Happiness Index and its Relation	The analysis of the World Happiness Report reveals that	1. Global North had higher average happiness scores,

	to the North-South Divide	the Global North exhibits a significantly higher average happiness score compared to the Global South. The top 10 happiest nations belong to the Global North, while the 10 least happy nations are predominantly from the Global South. Future research should focus on identifying the key factors contributing to this North-South divide in happiness levels	while Global South had lower scores. 2. The top 10 happiest countries belonged to Global North, while bottom 10 least happy countries were from Global South. 3. Statistical hypothesis test confirmed significant difference in happiness between regions. 4. Further research needed to explore contributing factors and address potential biases in World Happiness Index computation

3. DATASET DESCRIPTION

The dataset used for the World Happiness Index is an annual report that measures the happiness and well-being of people in different countries around the world. The report is published by the United Nations Sustainable Development Solutions Network and includes a variety of factors that contribute to overall happiness and well-being, such as economic prosperity, social support, life expectancy, freedom, and absence of corruption [6].

The dataset is widely used by researchers, policymakers, and the public to understand the state of happiness and well-being around the world, to identify trends and patterns over time, and to inform policies and interventions that promote human flourishing.

The dataset used for the World Happiness Index includes the following columns:

Table 3.1: Dataset Description

Feature Name	Description	Domain
Country	This column contains the names of the countries that are included in the report.	Categorical
Region	This column contains the region to which the country belongs.	Categorical
Happiness Rank	This column contains the rank of the country based on its level of happiness.	Numerical

Happiness Score	This column contains the score assigned to each country based on several factors that contribute to happiness, such as economic prosperity, social support, life expectancy, freedom, and absence of corruption.	Numerical
Standard Error	This column contains the standard error of the happiness score for each country, which measures the variability of the score.	Numerical
Economy (GDP per capita)	This column contains the economic prosperity of each country measured by its gross domestic product (GDP) per capita.	Numerical
Family	This column contains the social support that citizens receive from their families, friends, and communities.	Numerical
Health (Life Expectancy)	This column contains the life expectancy of citizens in each country, which is an indicator of overall health and well-being.	Numerical
Freedom	This column contains the level of freedom that citizens must make choices and pursue their goals.	Numerical

Trust (Government Corruption)	This column contains the level of trust that citizens have in their government and institutions, as well as the level of corruption that exists in these institutions.	Numerical
Region	This column contains the region to which the country belongs.	Numerical

Overall, the World Happiness Report dataset provides a comprehensive picture of the factors that contribute to happiness and well-being in different countries and regions of the world.

4. METHODOLOGY

We began by analyzing the mean happiness scores for countries from 2015 to 2019. We then delved deeper into the data by examining the trends in the top 10 countries with the highest mean happiness scores over the years. Additionally, we identified the top 10 most progressive countries from 2015 to 2019, using the change in happiness score over the period as a measure of progress.

To gain a more comprehensive understanding of the factors that contribute to happiness, we analyzed the mean values of the individual factors that make up the World Happiness Index. We also examined the average values of happiness variables for different regions to identify any regional differences in happiness.

Furthermore, we explored the relationship between the happiness score and each of the individual factors using scatter plots, along with regression lines to determine the strength of the relationship. To understand the correlations between these variables, we created a correlation matrix heat map. We also categorized the happiness scores into three levels and applied classification algorithms, namely K-Nearest Neighbors and Naive Bayes, to classify countries based on their happiness scores. Finally, we identified the most prominent features that contribute to happiness using an important analysis.

The study utilized a multivariate descriptive - exploratory data analysis using the cluster analysis technique. The data for Life Expectancy, Expected Years in Schooling, and Gross National Income Per Capita were retrieved from the United Nations Development Program (UNDP) Human Development Report 2014. The data for the World Happiness Report 2014 and the Corruption Perception Index were downloaded from the United Nations Sustainable Development Solutions Network (SDSN) and Transparency International online resources [7] [8].

4. 1 PREPROCESSING TASKS

The datasets from 2015-2022 were taken, and the 2017 dataset was chosen as the primary dataset. The preprocessing tasks were then performed, converting the remaining datasets to match the columns of the 2017 dataset.

4.1.1 Task 1 - Changing Column Names

The column names were modified in accordance with the primary dataset, which led to the renaming of columns in the other datasets. This task has been done so that all the datasets from 2015 to 2022 can be combined at the end.

Code:

```
df_15=plyr::rename(df_15, replace = c("Happiness Rank" = "Happiness.Rank",
                    "Happiness Score" = "Happiness.Score",
                    "Economy (GDP per Capita)" = "Economy..GDP.per.Capita.",
                    "Health (Life Expectancy)" = "Health..Life.Expectancy.",
                    "Trust (Government Corruption)" =
"Trust..Government.Corruption.",
                    "Dystopia Residual"="Dystopia.Residual"
                    ))
df_16=plyr::rename(df_16, replace = c( "Happiness Rank" = "Happiness.Rank",
                    "Happiness Score" = "Happiness.Score",
                    "Economy (GDP per Capita)" = "Economy..GDP.per.Capita.",
                    "Health (Life Expectancy)" = "Health..Life.Expectancy.",
                    "Trust (Government Corruption)"  =
"Trust..Government.Corruption.",
                    "Dystopia Residual"="Dystopia.Residual"
                    ))
df_18=plyr::rename(df_18, replace = c( "Country.or.region"="Country",
                    "Overall.rank"="Happiness.Rank" ,
                    "GDP.per.capita"="Economy..GDP.per.Capita.",
                    "Healthy.life.expectancy"="Health..Life.Expectancy.",
                    "Freedom.to.make.life.choices"="Freedom",
                    "Perceptions.of.corruption"="Trust..Government.Corruption.",
                    "Social.support"="Family",
                    "Score"="Happiness.Score"))
str(df_18)
```

```
df_19=plyr::rename(df_19, replace = c( "Country.or.region"="Country",
                        "Overall.rank"="Happiness.Rank" ,
                        "GDP.per.capita"="Economy..GDP.per.Capita.",
                        "Healthy.life.expectancy"="Health..Life.Expectancy.",
                        "Freedom.to.make.life.choices"="Freedom",
                        "Perceptions.of.corruption"="Trust..Government.Corruption.",
                        "Social.support"="Family",
                        "Score"="Happiness.Score"))
str(df_19)
# Adding a happiness rank column as it is not present in the dataset
df_20$Happiness.Rank <- rank(-df_20$Ladder.score)

# Arranging columns
df_20 <- df_20 %>%
  select(Country.name, Happiness.Rank, everything())

# Renaming the necessary column names
df_20 = plyr::rename(df_20, replace = c(
  "Country.name"="Country",
  "Logged.GDP.per.capita"="Economy..GDP.per.Capita.",
  "Healthy.life.expectancy"="Health..Life.Expectancy.",
  "Freedom.to.make.life.choices"="Freedom",
  "Perceptions.of.corruption"="Trust..Government.Corruption.",
  "Social.support"="Family",
  "Ladder.score"="Happiness.Score"
))

# Convert to numeric
df_20$Happiness.Score <- as.numeric(df_20$Happiness.Score)
df_20$Economy..GDP.per.Capita. <- as.numeric(df_20$Economy..GDP.per.Capita.)/10
df_20$Health..Life.Expectancy. <- as.numeric(df_20$Health..Life.Expectancy.)/100
str(df_20)
```

```
# Adding a happiness rank column as it is not present in the dataset
df_21$Happiness.Rank <- rank(-df_21$Ladder.score)

df_21 = plyr::rename(df_21, replace = c(
  "Country.name"="Country",
  "Logged.GDP.per.capita"="Economy..GDP.per.Capita.",
  "Healthy.life.expectancy"="Health..Life.Expectancy.",
  "Freedom.to.make.life.choices"="Freedom",
  "Perceptions.of.corruption"="Trust..Government.Corruption.",
  "Social.support"="Family",
  "Ladder.score"="Happiness.Score"
))

# Arranging columns
df_21 <- df_21 %>%
  select(Country, Happiness.Rank, everything())

df_21$Economy..GDP.per.Capita. <- as.numeric(df_21$Economy..GDP.per.Capita.)/10
df_21$Health..Life.Expectancy. <- as.numeric(df_21$Health..Life.Expectancy.)/100

str(df_21)
df_22 = plyr::rename(df_22, replace = c(
  "RANK"="Happiness.Rank" ,
  "Explained.by..GDP.per.capita"="Economy..GDP.per.Capita.",
  "Explained.by..Healthy.life.expectancy"="Health..Life.Expectancy.",
  "Explained.by..Freedom.to.make.life.choices"="Freedom",
  "Explained.by..Perceptions.of.corruption"="Trust..Government.Corruption.",
  "Explained.by..Social.support"="Family",
  "Score"="Happiness.Score"
))
```

4.1.2 Task 2 - Inserting New Columns “Year” And “Region”

To ensure proper record-keeping of the happiness scores and regional information, we introduced two new columns, "Year" and "Region," while merging all the datasets from 2015-2022. The "Year" column tracks the year of each happiness score recording, while the "Region" column keeps track of the various regions within countries.

Code:

```
df_15<-cbind(Year=2015,df_15)

df_16<-cbind(Year=2016,df_16)

df_17<-cbind(Year=2017,df_17)

df_18<-cbind(Year=2018,df_18)

df_19<-cbind(Year=2019,df_19)

df_20<-cbind(Year=2020,df_20)

df_21<-cbind(Year=2021,df_21)

df_22<-cbind(Year=2022,df_22)

common_region <- unique(subset(df1, Region!="NA", c(Country, Region)))

head(common_country)
assign_region <- function(x){
Region <- common_region$Region[common_region$Country == x]
}

for(country in common_country)
df1$Region[df1$Country == country] <- assign_region(country)
```

4.1.3 Task 3 - Merging The Datasets from 2015-2022

The datasets from 2015-2022 were merged to create a consolidated dataset.

Code:

```
df15_16<-dplyr::bind_rows(df_15,df_16)

df15_16_17<-dplyr::bind_rows(df15_16,df_17)

df18_19<-dplyr::bind_rows(df_18,df_19)

df_20_21 <- dplyr::bind_rows(df_20,df_21)

str(df_20_21)
df_20_21_22 <- dplyr::bind_rows(df_20_21,df_22)

df_15_16_17_18_19<-dplyr::bind_rows(df18_19,df15_16_17)

df<-dplyr::bind_rows(df_15_16_17_18_19, df_20_21_22)

head(df)
```

4.1.4 Task 4 - Removing Unnecessary Columns

Columns that were not related to the "Quality of life" factors were excluded from the consolidated dataset.

Code:

```
# df = subset(df, select = -
cower.Confidence.Interval,Upper.Confidence.Interval,Dystopia.Residual,Standard.Error
,Whisker.high,Wh        isker.low))

colSums(is.na(df))
```

4.1.5 Task 5 - Filtering Uncommon Data In Country Column

Countries that were not present in all the datasets were removed from the consolidated dataset.

Code:

```
aggregate(df$Country, by=list(df$Year), FUN=length)
Country_2015 = subset(df, Year == 2015)$Country
Country_2016 = subset(df, Year == 2016)$Country
Country_2017 = subset(df, Year == 2017)$Country
```

```
Country_2018 = subset(df, Year == 2018)$Country
Country_2019 = subset(df, Year == 2019)$Country
Country_2020 = subset(df, Year == 2020)$Country
Country_2021 = subset(df, Year == 2021)$Country
Country_2022 = subset(df, Year == 2022)$Country
common_country
=intersect(intersect(intersect(intersect(intersect(intersect(Country_2015,
Country_2016),Country_2017),Country_2018),Country_2019),Country_2020),Country_2021),Country_
2022)
length(common_country)
```

4.1.6 Task 6 - Filling Na Values with Mean Value Of The Respective Columns

The missing values (NA values) in the dataset were filled with the mean value of the respective columns. This approach helps to ensure that the dataset remains complete and that the missing values are replaced with representative values based on the average of each column.

Code:

```
df$Trust..Government.Corruption.[is.na(df$Trust..Government.Corruption.)] <-
median(df$Trust..Government.Corruption., na.rm = T)

df$Happiness.Score[is.na(df$Happiness.Score)] <- median(df$Happiness.Score, na.rm = T)

df$Economy..GDP.per.Capita.[is.na(df$Economy..GDP.per.Capita.)] <-
median(df$Economy..GDP.per.Capita., na.rm = T)
df$Family[is.na(df$Family)] <- median(df$Family, na.rm = T)
df$Health..Life.Expectancy.[is.na(df$Health..Life.Expectancy.)] <-
median(df$Health..Life.Expectancy., na.rm = T)

df$Freedom[is.na(df$Freedom)] <- median(df$Freedom, na.rm = T)
df$Generosity[is.na(df$Generosity)] <- median(df$Generosity, na.rm = T)

colSums(is.na(df))
```

5. RESULTS AND COMPARISON

5.1 TOP 10 COUNTRIES: MEAN HAPPINESS SCORE TRENDS

The graph provides an overview of how the mean happiness scores have changed over the years for the top 10 countries.

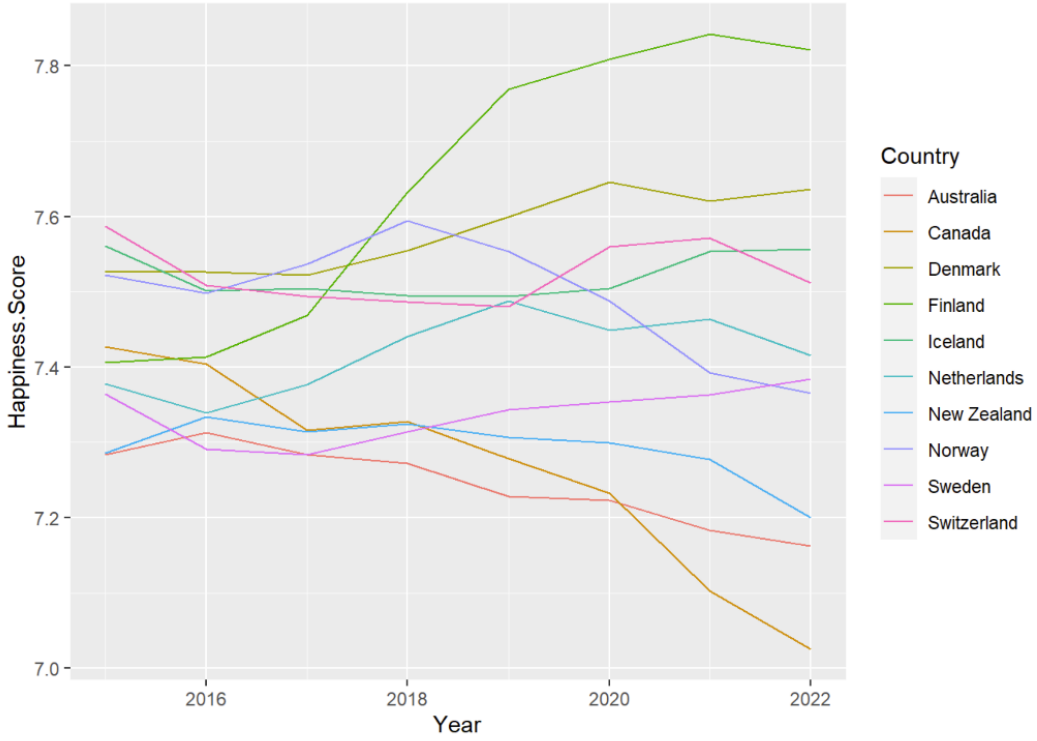

Figure 5.1: Mean Happiness Score Trends

Code:

```
Top10_happy_country_DF = df1 %>%
 group_by(Country) %>%
 summarise(mscore = mean(Happiness.Score)) %>%
 arrange(-mscore) %>%
 slice_head(n=10)

Top10_happy_country_DF_list = c(Top10_happy_country_DF$Country)

df1_Top10_happy_country = subset(df1,Country %in% Top10_happy_country_DF_list)

ggplot(df1_Top10_happy_country,  aes(x = Year,y = Happiness.Score,color = Country))+  geom_line()
```

Key inferences from the graph in Figure 5.1:

Finland: While Finland consistently maintains the highest happiness score, its progression over the years may not be as pronounced since it already starts with a high score.

Denmark: Denmark has a relatively high starting score and shows a consistent upward trend, indicating progression over time.

Switzerland: Switzerland also starts with a high score and demonstrates a relatively stable trend with slight fluctuations.

Iceland: Iceland's happiness score remains relatively stable with minor variations over the years.

Norway: Norway starts with a high score and shows a slightly increasing trend.

Netherlands: The Netherlands starts with a high score and maintains a relatively stable trend with slight variations.

Sweden: Sweden's happiness score remains relatively stable with minor fluctuations.

New Zealand: New Zealand starts with a high score and demonstrates an upward trend, indicating progression over time.

Canada: Canada starts with a high score and shows a slightly increasing trend.

Australia: Australia's happiness score remains relatively stable with minor variations. Based on these observations, it appears that Denmark, New Zealand, and Canada have shown noticeable progression in happiness scores over the years among the top 10 happiest countries.

5.2 TOP 10 MOST PROGRESSIVE COUNTRIES

The scatter plot and line plot were utilized to identify the countries that have made the most progress in terms of their happiness score from 2015 to 2022.

Code:

```
df1 %>%
  mutate(y = as.character(Year)) %>%
  select(y, Country, Region, Happiness.Score) %>%
  pivot_wider(names_from = y, values_from = Happiness.Score,
              names_prefix = "y_") %>%
  mutate(p = (y_2022 - y_2015)/y_2015 * 100) %>%
  arrange(-p) %>%
  slice_head(n = 10) %>%
  ggplot(aes(reorder(Country, p), p)) +
  geom_point() +
  theme_bw() +
  coord_flip() +
  labs(title = "The 10 most progressive countries from 2015 - 2022",
       y = "Percentage Increase of Happiness Score", x = "")
```

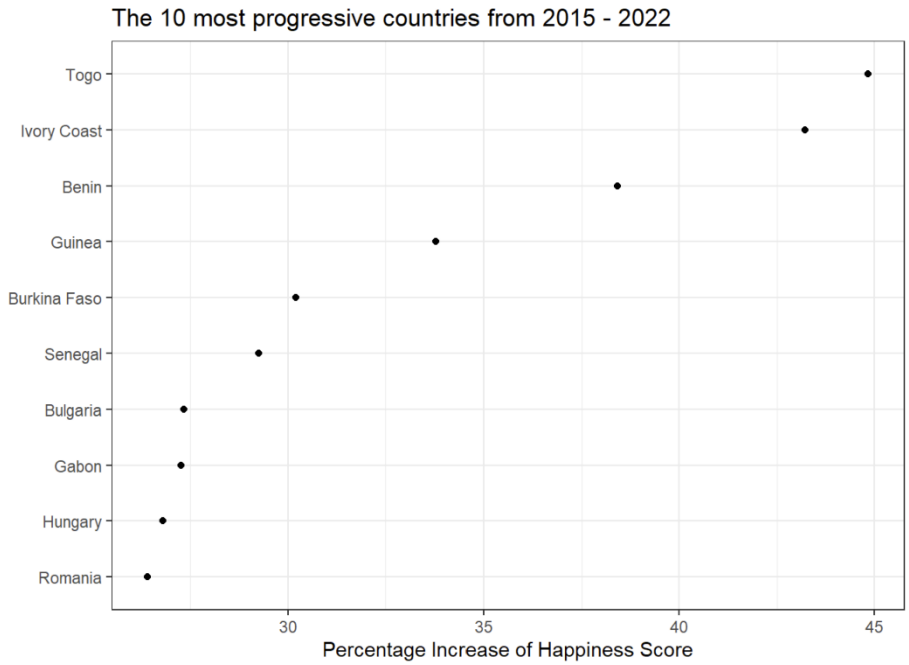

Figure 5.2: Scatter plot of Top 10 most progressive countries from 2015-2022

From the scatter plot in Figure 5.2, we can infer that the top 10 most progressive countries in terms of happiness score have all experienced significant increases in their happiness scores from 2015 to 2022. The countries that have progressed the most are Togo, Ivory Coast, and Benin, all from the Sub-Saharan Africa region. They have experienced a percentage increase of 44.84%, 43.23%, and 38.41%, respectively.

```
Top10_Progress_country_df = df1 %>%
 mutate(y = as.character(Year)) %>%
 select(y, Country, Region, Happiness.Score) %>%
 pivot_wider(names_from = y, values_from = Happiness.Score,
         names_prefix = "y_") %>%
 mutate(p = (y_2022 - y_2015)/y_2015 * 100) %>%
 arrange(-p) %>%
 slice_head(n = 10)

Top10_Progress_country_df_list = c(Top10_Progress_country_df$Country)

df1_Top10_Progress_country = subset(df1,Country %in% Top10_Progress_country_df_list)
ggplot(df1_Top10_Progress_country,  aes(x = Year,y = Happiness.Score,color = Country))+  geom_line()
```

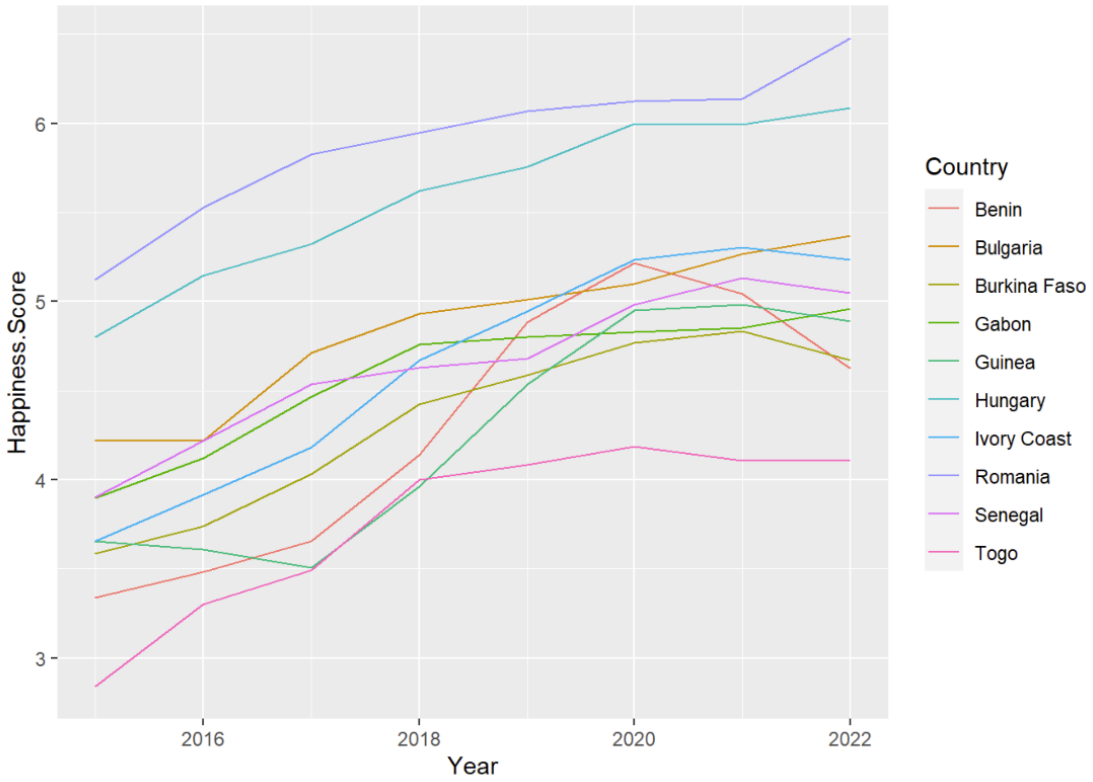

Figure 5.3: Trends of Top 10 most Progressive Countries from 2015-2022

From the line plot in Figure 5.3, we can infer that all countries have experienced an upward trend in happiness scores over the years, indicating progress. Based on the plot, it seems that Togo and Ivory Coast have shown the most significant improvement in happiness scores, followed by Benin and Guinea. Hungary and Romania, which are both from the Central and Eastern Europe region, also show a significant improvement in happiness scores. Overall, the line plot suggests that the top 10 most progressive countries have made significant progress in improving their citizens' happiness and well-being over the years.

5.3 REGIONAL ANALYSIS ON HAPPINESS SCORES OF COUNTRIES

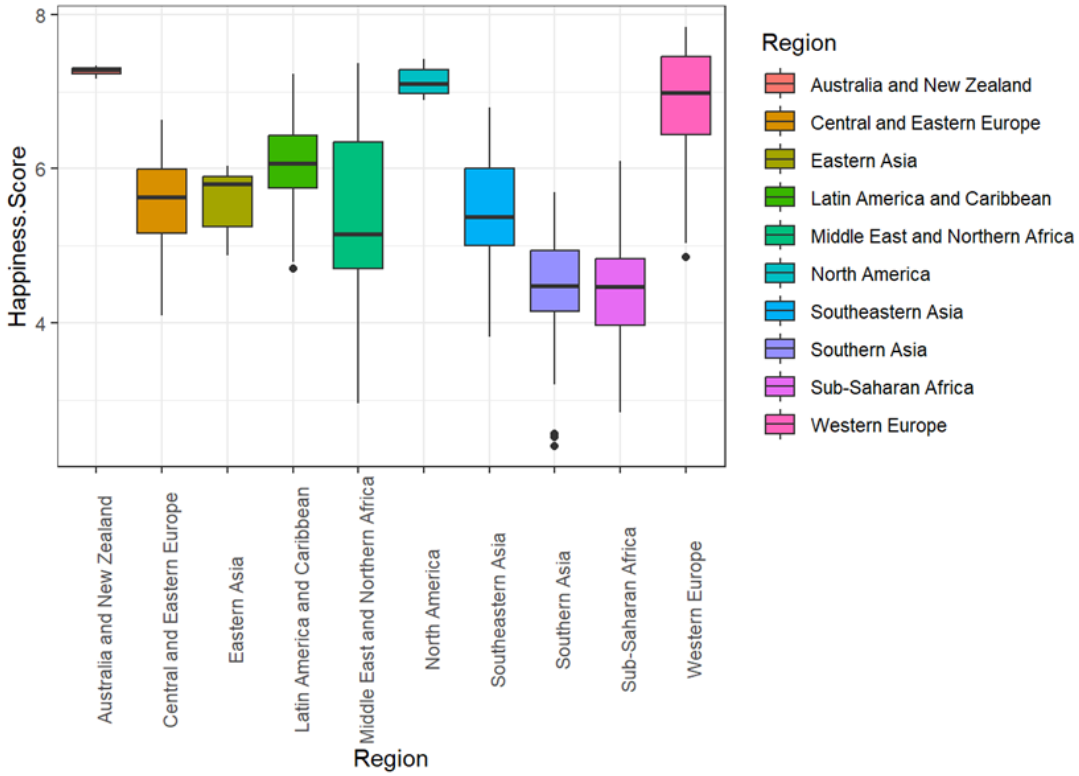

Figure 5.4: Boxplot for regional analysis of happiness scores

The boxplot in Figure 5.4 gives you a visual representation of the distribution of happiness scores across different regions. You can compare the median, the range, and the spread of happiness scores across the different regions.

Australia and New Zealand, North America and Western Europe have the highest median happiness scores and the narrowest range of scores, indicating that people in these regions are generally happier and there is less variability in their happiness levels.

Sub-Saharan Africa has the lowest median happiness score and the widest range of scores, indicating that people in this region are generally less happy, and there is a lot of variability in their happiness levels.

Latin America and the Caribbean and Southeast Asia have relatively high median happiness scores but wider ranges of scores compared to Western Europe and North America, indicating that people in these regions are generally happy but there is more variability in their happiness levels.

Overall, the boxplot can give you a good sense of the differences in happiness levels across regions and how the variability in happiness scores compares across regions.

5.4 FINDING CORRELATION USING CORRELATION MATRIX HEATMAP

The correlation heatmap shows the correlation coefficients between all pairs of variables in the dataset. The correlation coefficient ranges from -1 to 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation.

Code:

```
dataset = select(df1,-c("Year","Country","Happiness.Rank","Region"))
head(dataset)
library(GGally)
ggcorr(dataset, label = TRUE, label_round = 2, label_size = 3.5, size = 2, hjust = .85) +
 ggtitle("Correlation Heatmap") +
 theme(plot.title = element_text(hjust = 0.5))
```

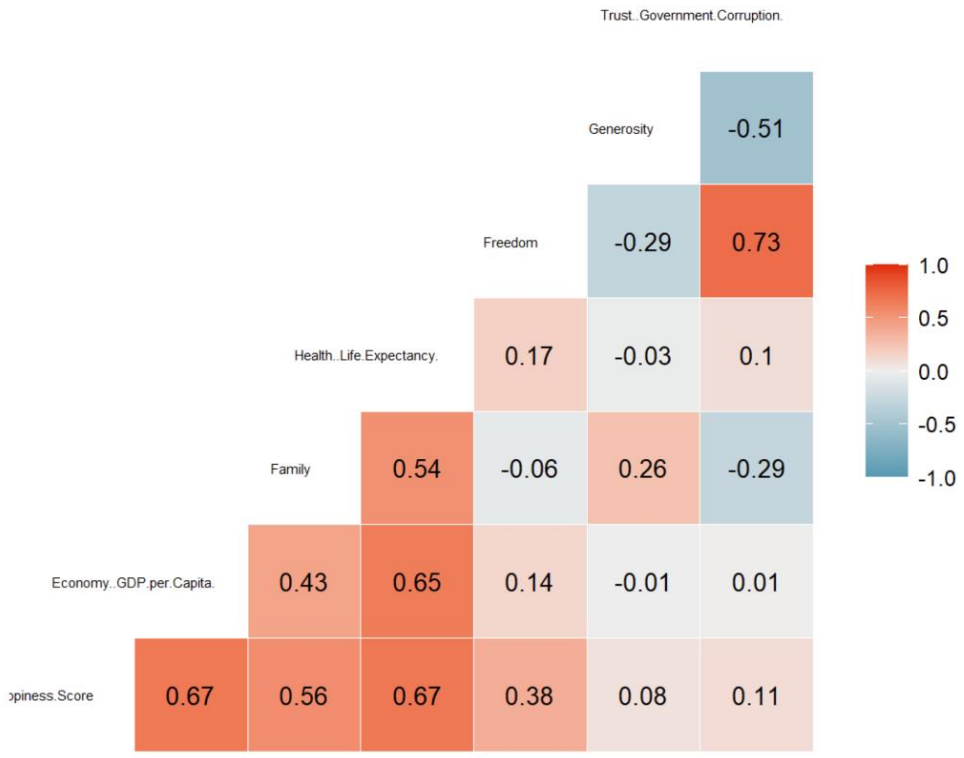

Figure 5.5: Correlation Heatmap

From the heatmap in Figure 5.5, we can infer that there is a strong positive correlation between "Happiness.Score" and GDP per capita, Healthy life expectancy, and Family, which suggests that countries with higher GDP per capita, longer healthy life expectancy, and stronger social support tend to have higher levels of happiness. There is also a moderate positive correlation between "Happiness.Score" and Freedom to make life choices and Generosity.

On the other hand, there is a negative correlation between "Happiness.Score" and Corruption, which indicates that countries with higher levels of corruption tend to have lower levels of happiness. There is also a negative correlation between Freedom to make life choices and Corruption, which suggests that countries with higher levels of corruption tend to have less freedom to make life choices

5.5 MULTIPLE LINEAR REGRESSION

Multiple linear regression was employed to predict the happiness score values, leveraging a statistical technique that models the relationship between a dependent variable and two or more independent variables. This technique expands upon the foundational concept of simple linear regression, which focuses solely on a single independent variable.

Code:

```
lm_model = lm(formula = Happiness.Score ~ .,
              data = data_train)
summary(lm_model)
y_pred_lm = predict(lm_model, newdata = data_test)
Actual_lm = data_test$Happiness.Score

Pred_Actual_lm <- as.data.frame(cbind(Prediction = y_pred_lm, Actual = Actual_lm))

gg.lm <- ggplot(Pred_Actual_lm, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Multiple Linear Regression", x = "Actual happiness score",
       y = "Predicted happiness score") +
  theme(plot.title = element_text(face="bold", size = (15)),
        axis.title = element_text(size = (10)))
gg.lm
```

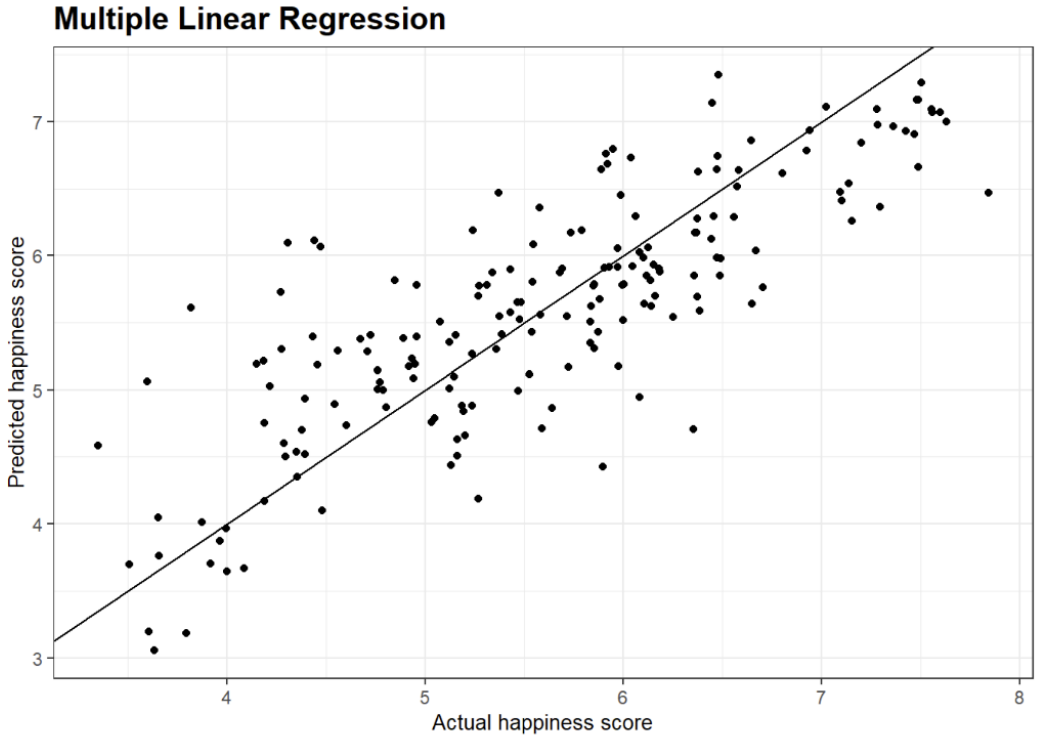

Figure 5.6: Multiple Linear Regression

The scatter plot in Figure 5.6 helps visualize the model's performance in predicting happiness scores. As the points are closely clustered around the regression line and follow a relatively linear pattern, it indicates that the model's predictions are accurate and consistent with the actual values.

Performance Metrics:

Table 5.1: Multiple Linear Regression Metrics

R2	RMSE	MAE
0.7867092	0.4797292	0.3779097

5.6 SUPPORT VECTOR REGRESSION

Support Vector Regression (SVR) was employed to predict happiness score values, utilizing a machine learning algorithm that extends the principles of Support Vector Machines (SVM) to regression tasks. SVR specializes in modeling and predicting continuous numeric variables by establishing a relationship between the dependent variable and a set of independent variables. Its objective is to discover an optimal hyperplane that effectively fits the data while maximizing the margin around the hyperplane. This hyperplane is determined based on support vectors, which are the data points closest to the hyperplane. By leveraging these support vectors, SVR

can accurately predict happiness scores and handle non-linear relationships between the variables.

Code:

```
library(e1071)

regressor_svr = svm(formula = Happiness.Score ~ .,
            data = data_train,
            type = 'eps-regression',
            kernel = 'radial')
# Predicting happiness score with SVR model
y_pred_svr = predict(regressor_svr,  newdata = data_test)

Pred_Actual_svr <- as.data.frame(cbind(Prediction = y_pred_svr, Actual = data_test$Happiness.Score))


Pred_Actual_lm.versus.svr <- cbind(Prediction.lm = y_pred_lm, Prediction.svr = y_pred_svr, Actual =
data_test$Happiness.Score)


gg.svr <- ggplot(Pred_Actual_svr, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "SVR", x = "Actual happiness score",
       y = "Predicted happiness score") +
  theme(plot.title = element_text(face = "bold", size = (15)),
        axis.title = element_text(size = (10)))
gg.svr
```

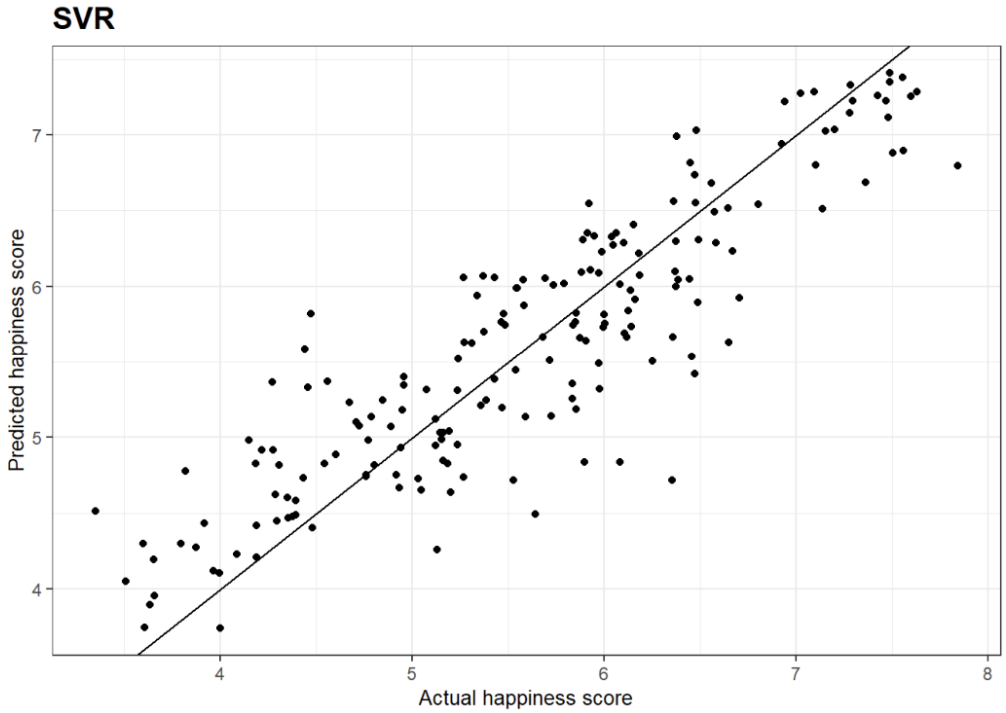

Figure 5.7: Support Vector Regression (SVR)

The plot in Figure 5.7 helps to visualize the performance of the SVR model in predicting happiness scores. As the points closely align with the regression line and exhibit a relatively linear pattern, it suggests that the SVR model accurately captures the relationship between predictors and happiness scores.

Performance Metrics:

Table 5.2: Support Vector Regression Metrics

R2	RMSE	MAE
0.7867092	0.4797292	0.3779097

5.7 DECISION TREE REGRESSION FOR PREDICTING HAPPINESS SCORES

A decision tree regression model was employed to predict the values of happiness scores in the project. A decision tree regression model is a machine learning algorithm that utilizes a tree-like structure to model the relationships between input features and output values. It works by recursively splitting the data into subsets based on the input feature that provides the best split, with the goal of minimizing the residual error between the predicted and actual values.

Code:

```
library(rpart)
regressor_dt = rpart(formula = Happiness.Score ~ .,
```

```
            data = data_train,
            control = rpart.control(minsplit = 10))
# Predicting happiness score with Decision Tree Regression
y_pred_dt = predict(regressor_dt, newdata = data_test)

Pred_Actual_dt <- as.data.frame(cbind(Prediction = y_pred_dt, Actual =
data_test$Happiness.Score))


gg.dt <- ggplot(Pred_Actual_dt, aes(Actual, Prediction )) +
  geom_point() + theme_bw() + geom_abline() +
  labs(title = "Decision Tree Regression", x = "Actual happiness score",
       y = "Predicted happiness score") +
  theme(plot.title = element_text(face = "bold", size = (15)),
        axis.title = element_text(size = (10)))
gg.dt
library(rpart.plot)
prp(regressor_dt)
```

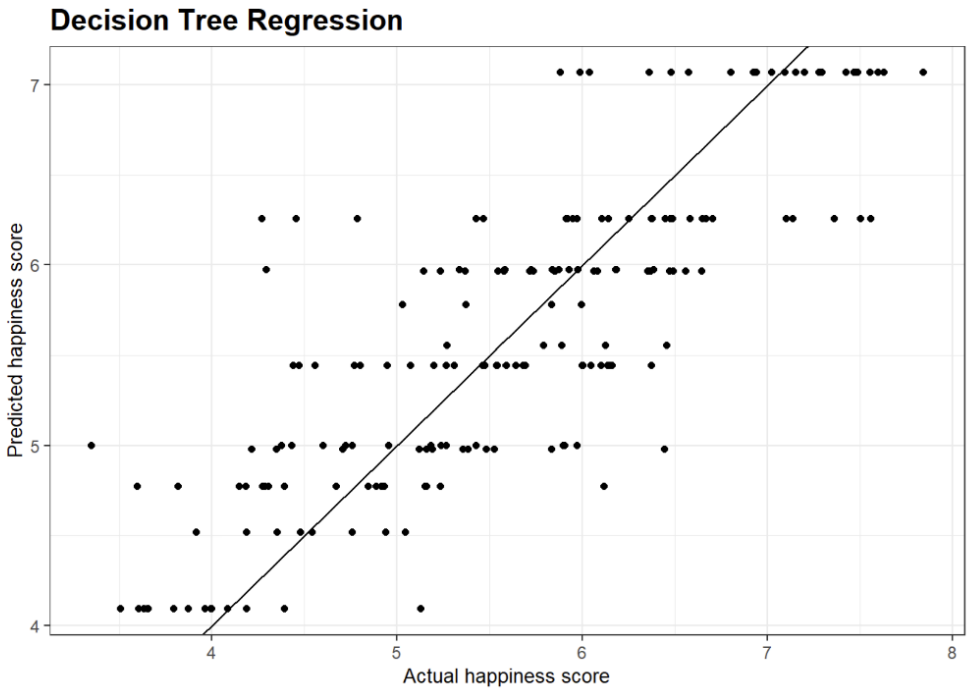

Figure 5.8: Scatter plot of Decision Tree Regression

The plot in Figure 5.8 displays a scatter plot of the actual versus predicted happiness score for the test data set, along with a diagonal line that represents perfect prediction. We can see that

the model generally predicts the happiness score accurately, with a few outliers where the prediction deviates significantly from the actual value.

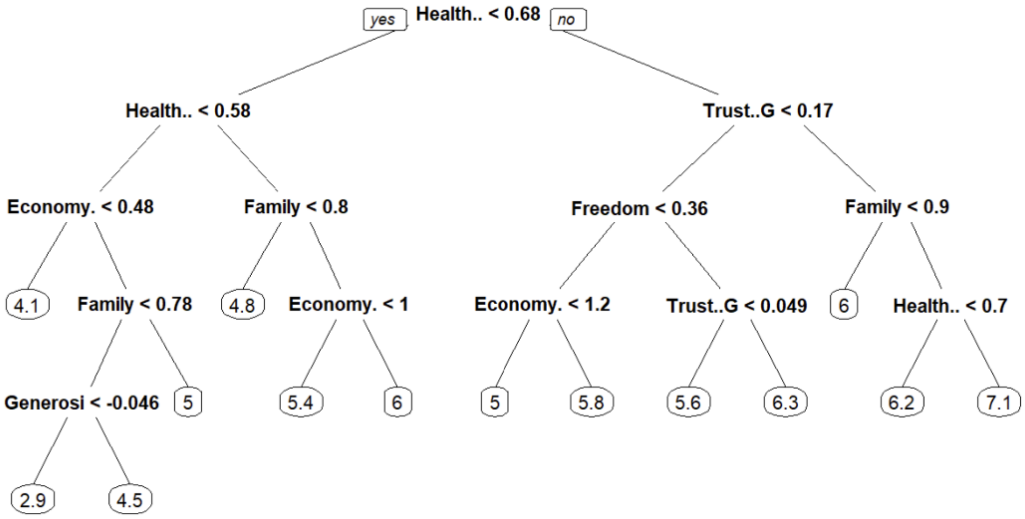

Figure 5.9: Decision Tree

From the decision tree in Figure 5.9, the first decision point is based on the "GDP per capita" feature, with a split at a value of 0.691. The tree then splits further based on other features such as "Social support", "Healthy life expectancy", and "Freedom to make life choices". The split points and feature importance are determined by the algorithm used to build the tree.

Performance Metrics:

Table 5.3: Decision Tree Regression Metrics

R2	RMSE	MAE
0.6752957	0.5908649	0.4587602

5.8 KNN CLASSIFIER FOR PREDICTING HAPPINESS LEVELS

The KNN model provides insights into the relationship between predictor variables and happiness scores, allowing us to make predictions and identify the relative importance of several factors influencing happiness. This can be illustrated in Figure 5.10

Code:

```
set.seed(123)
library("MLmetrics")
```

```
model_knn <- train(
 Happy.Level~.,
 data=data_train,
 trControl=tc,
 preProcess = c("center","scale"),
 method="knn",
 metric='Accuracy',
 tuneLength=20
 )

model_knn
plot(model_knn)
pred_knn <- predict(model_knn, data_test)

cm_knn<-confusionMatrix(pred_knn, data_test$Happy.Level)

cm_knn
```

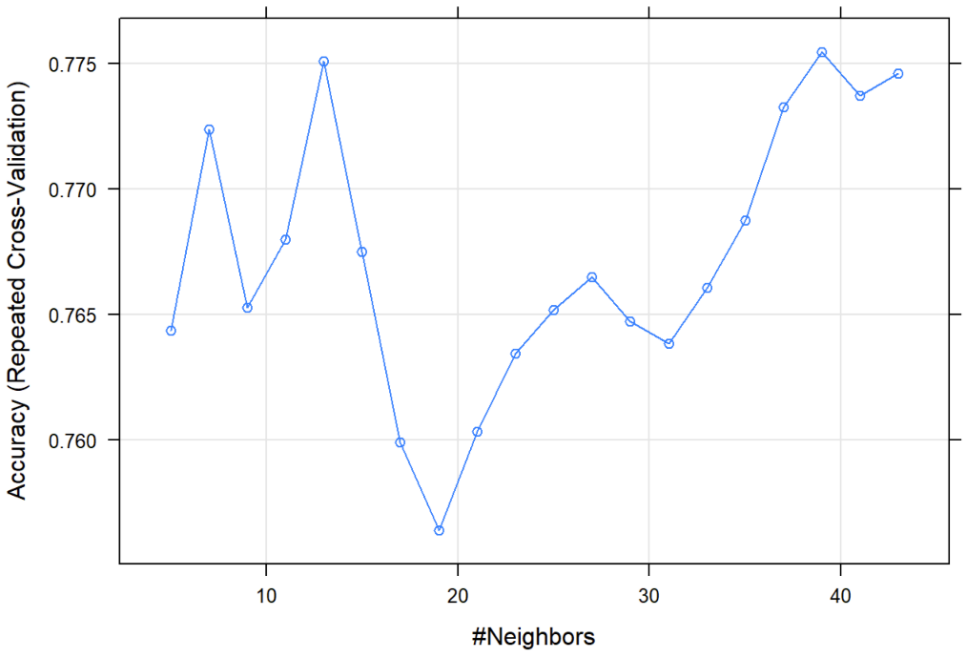

Figure 5.10: KNN Model

5.9 NAÏVE BAYES CLASSIFICATION MODEL

Naive Bayes is a probabilistic algorithm that uses the Bayes theorem with the "naive" assumption of conditional independence between features. This model was applied to predict the happiness level. It is trained on the training data with the outcome variable "Happy. Level"

and all the other variables as predictors. The model is pre-processed by centering and scaling the predictors.

Code:

```
model_nb <- train(Happy.Level~.,
                  data_train,
                  method="naive_bayes",
                  preProcess = c("center","scale"),
                  metric='Accuracy',
                  trControl=tc)

model_nb
plot(model_nb)
pred_nb <- predict(model_nb, data_test)

cm_nb<-confusionMatrix(pred_nb, data_test$Happy.Level)

cm_nb
```

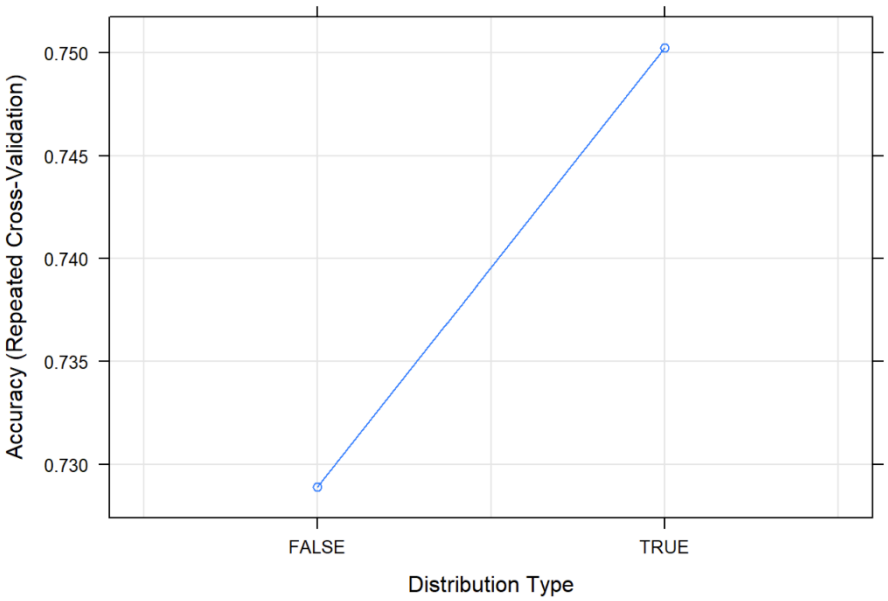

Figure 5.11: Naïve Bayes Classification Model

The plot in Figure 5.11 shows a density plot for each feature in the dataset, separated by 'Happy. Level' classes. From the plot, we can see the distribution of each feature for the 'Happy' and 'Unhappy' classes. This can help us identify which features are more important for

distinguishing between happy and unhappy countries. Overall, the plot can provide insights into the underlying data and help us understand the behavior of the model.

5.10 FEATURE IMPORTANCE

Feature importance refers to the measure of the relative importance or contribution of each feature (also known as predictor variable or independent variable) in a machine learning model. It helps in understanding which features have the most noteworthy influence on the target variable (also known as the dependent variable) and can provide insights into the underlying relationships and patterns in the data. It helps in identifying the variables that have a strong influence on the model's predictions and those that have a lesser impact. This is illustrated in Figure 5.12.

Code:

```
# Create object of importance of our variables
knn_importance <- varImp(model_knn)

# Create box plot of importance of variables
ggplot(data = knn_importance, mapping = aes(x = knn_importance[,1])) + # Data & mapping
  geom_boxplot() + # Create box plot
  labs(title = "Variable importance: K-Nearest Neighbours ") + # Title
  theme_light() # Theme

# Second Plot
# Create object of importance of our variables
nb_importance <- varImp(model_nb)

# Create box plot of importance of variables
ggplot(data = nb_importance, mapping = aes(x = nb_importance[,1])) + # Data & mapping
  geom_boxplot() + # Create box plot
  labs(title = "Variable importance: Naive Bayes model") + # Title
  theme_light() # Theme
```

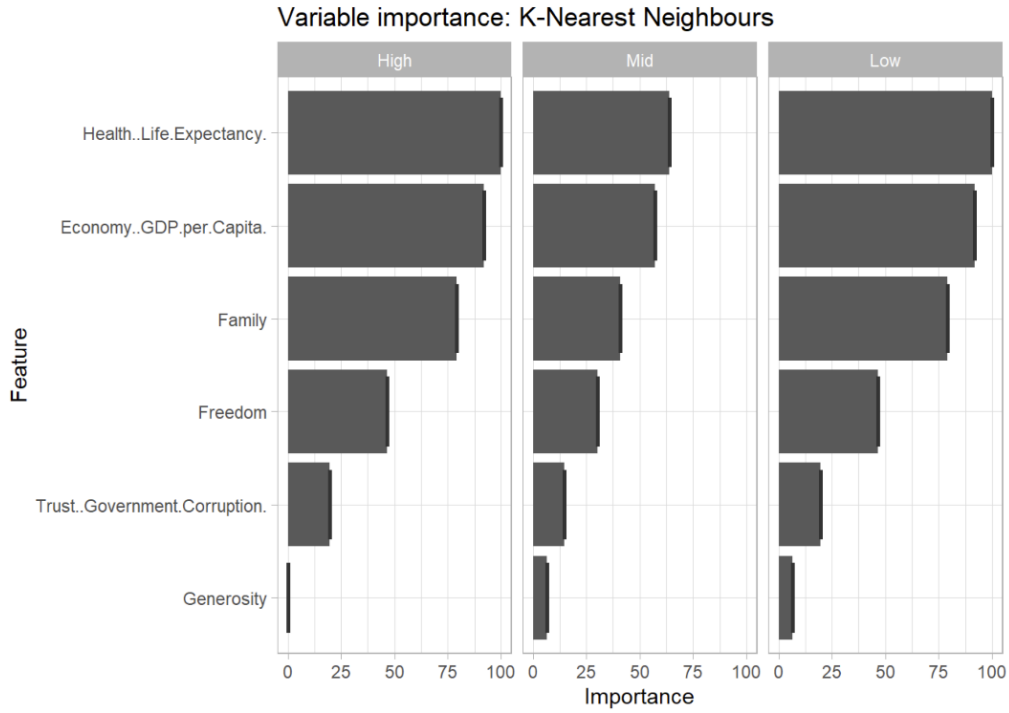

Figure 5.12: Feature Importance (KNN)

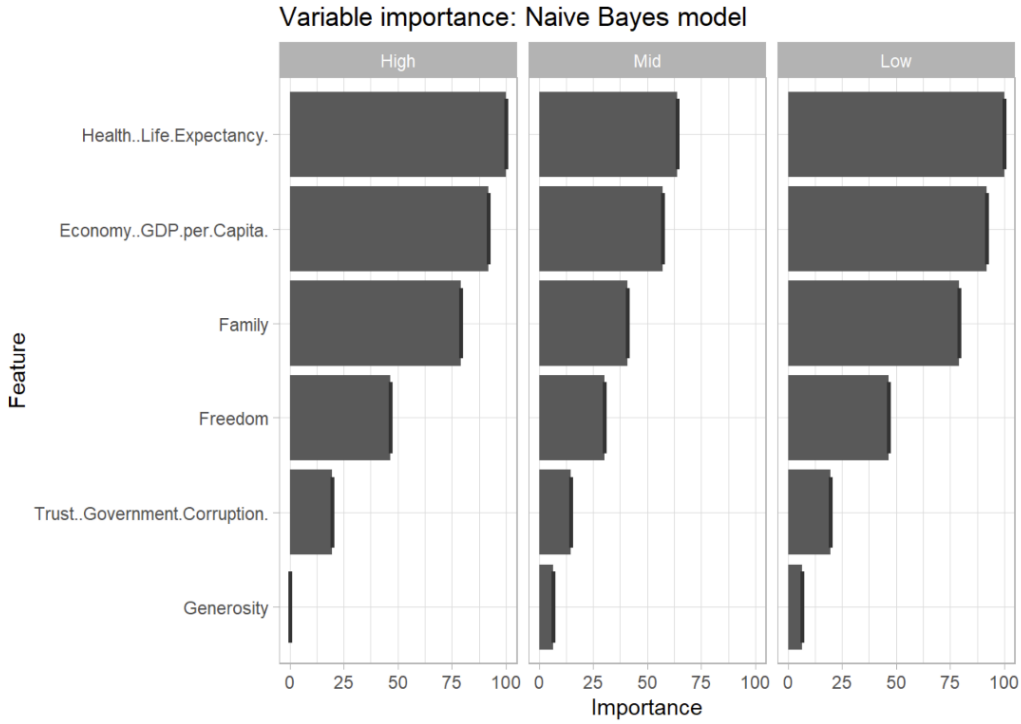

Figure 5.13: Feature Importance (Naïve Bayes)

The feature-importance analysis as seen in Figure 5,13, indicated that "Health life Expectancy" had a higher impact on predicting happiness scores compared to other quality of life factors. This implies that the health aspect plays a significant role in determining happiness levels.

6. CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

In conclusion, this study has shown that happiness depends on an enormous range of influences. By analyzing data from the World Happiness Report from 2015 to 2022, we have been able to identify some of the key factors that contribute to happiness, such as economic prosperity, social support, and a sense of freedom and autonomy. Thus, regular collection of happiness data on a large scale can inform policymaking and help us identify what "deliverables" should be created to foster well-being.

In other words, moving to a happier country could plausibly make you happier. For the same reason, moving to a less happy country could reduce your level of happiness. It is also important to note that happiness is not only an individual phenomenon, but also a collective one. Emotions are contagious, even at a national level.

In addition to our regional analysis, we further examined the policies implemented by the top 3 progressive countries to understand the factors that contributed to their remarkable increase in happiness scores. Our findings reveal that these countries, Brenin, Togo, and Ivory Coast, have implemented a range of strategies aimed at improving the well-being and happiness of their citizens. These strategies include investments in social support systems, healthcare infrastructure, and initiatives promoting transparency and trust in governance. The success of these policies in driving significant improvements in happiness scores underscores the importance of targeted interventions and evidence-based policymaking. By studying and adopting these best practices, other countries can potentially replicate the success achieved by these progressive nations and enhance the well-being of their own populations.

6.2 FUTURE SCOPE

Therefore, a happy population can contribute to the happiness of others in the same community or country, while an unhappy population can have a negative impact on others. This highlights the importance of not only individual happiness, but also the well-being of society as a whole. Policymakers can use this information to design policies that promote happiness and well-being, ultimately contributing to a happier and healthier world.

Deep learning techniques can be employed to analyze unstructured data sources such as social media posts, news articles, and online forums. By applying sentiment analysis, natural language processing, and deep learning algorithms, you can gain insights into public sentiment and its relationship to happiness levels. This can provide real-time or near-real-time information about happiness trends and public well-being.

Building on the project's findings, personalized happiness assessment models can be developed. These models can consider individual-level data, such as demographics, personal preferences, and lifestyle factors, to provide tailored recommendations for individuals to enhance their well-being and happiness.

Exploring the impact of cultural and contextual factors on happiness can provide valuable insights. This may involve considering cultural values, social norms, and specific contextual circumstances to understand how they influence happiness levels in different regions or countries.

BIBLIOGRAPGY

[1] "Bhutan’s Gross National Happiness Index | OPHI," 2 12 2020. [Online]. Available: https://ophi.org.uk/policy/gross-national-happiness-index/.

[2] S. Ahtesham, "Analyzing happiness index as a measure along with its parameters and strategies for improving India’s rank in world happiness report," *ICTACT Journals,* vol. 6, no. 1, p. 4, 2020.

[3] N. S. Ayoub Jannani, "Predicting Quality of Life using Machine Learning:," *IEEE,* 2021.

[4] L. Carlsen, "How Happy Are we Actually? A Posetic Analysis," *International Journal of Community Well-Being,* p. 12, 2019.

[5] C. A. L. R. E. T. Eden J. Garces, "Analysis on the Relationships on the Global Distribution of the World Happiness Index and Selected Economic Development Indicators," *Open Access Library Journal,* vol. 6, p. 16, 2019.

[6] "About | The World Happiness Report," 2023. [Online]. Available: https://worldhappiness.report/about/.

[7] " UN Sustainable Development Solutions Network (SDSN)," UN, [Online]. Available: https://www.unsdsn.org/.

[8] "Transparency International," [Online]. Available: https://www.transparency.org/en/library.

[9] "Togo country profile," 28 April 2022. [Online]. Available: https://www.bbc.com/news/world-africa-14106781.

[10] T. W. Bank, "The World Bank in Togo," 31 March 2023. [Online]. Available: https://www.worldbank.org/en/country/togo/overview#3.

[11] M. C. N. S. Siddharth Dixit, "NETWORK LEARNING APPROACHES TO STUDY WORLD," *arxiv,* p. 13, 2020.