Module 6 Assignment - Final Project Report
World Happiness Report with R

PRIYA KATTIGEHALLI MATA

Northeastern University: College of Professional Studies

ALY 6015: Intermediate Analytics

Professor: Valeriy Shevchenko

Week 6 – December 12th, 2024

Table of Contents

Analysis of the World Happiness Report Dataset: Determinants of National Well-being

1. Executive Summary

This report presents a comprehensive analysis of the World Happiness Report dataset, covering 2,363 observations from 2005 to 2023. Key findings include:

- Strong positive correlations between happiness scores and factors such as GDP, social support, and life expectancy.
- A multiple regression model explaining 74.23% of variance in happiness scores, with all predictors being statistically significant.
- Significant differences in happiness between countries with high and low GDP.
- Time series analysis suggesting stable happiness trends with slight variations over time.

Recommendations include focusing on economic growth, strengthening social support systems, improving healthcare, and reducing corruption to enhance national well-being.

2. Introduction

The World Happiness Report has become a crucial tool for policymakers and researchers to understand the factors contributing to national well-being. This analysis aims to:

1. Identify key determinants of national happiness.
2. Explore relationships between socio-economic factors and happiness scores.
3. Forecast happiness trends and provide data-driven policy recommendations.

Research Questions/ Problems:

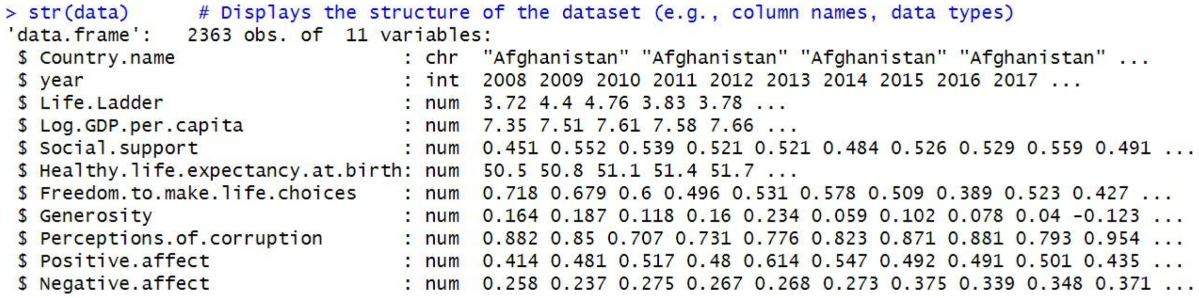
1. **What factors most strongly influence happiness scores?**
   The regression analysis indicates that both GDP per capita and social support significantly impact happiness scores. For each unit increase in Log GDP per capita, the Life Ladder score increases by 0.51941, holding Social support constant. Similarly, for each unit increase in Social support, the Life Ladder score increases by 3.39571, holding Log GDP per capita constant. This suggests that while economic factors are important, social support has a larger effect on happiness.
2. **How much of the variation in happiness scores can be explained by economic and social factors?**
   The regression model explains approximately 67.2% of the variance in Life Ladder scores (R-squared = 0.672). This indicates that GDP per capita and social support are crucial factors in determining happiness levels, but there are other variables not included in this model that also contribute to happiness.

3. **Is there a relationship between a country's economic status and its happiness level?**
   Yes, there is a strong positive relationship between Log GDP per capita and happiness scores (correlation coefficient = 0.79). This indicates that countries with higher economic output tend to have higher happiness levels. However, the relationship is not perfect, suggesting that other factors also play important roles in determining happiness.

4. **How does social support impact happiness scores?**
   Social support has a strong positive relationship with happiness scores (correlation coefficient = 0.72). The regression analysis shows that social support has a larger coefficient (3.39571) compared to Log GDP per capita (0.51941), suggesting that improvements in social support could have a more substantial impact on happiness than equivalent increases in GDP.

3 Data Overview and Descriptive Statistics

The dataset comprises 11 variables for 2,363 country-year observations:



```
> str(data)      # Displays the structure of the dataset (e.g., column names, data types)
'data.frame':   2363 obs. of  11 variables:
 $ Country.name                    : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ year                            : int  2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 ...
 $ Life.Ladder                     : num  3.72 4.4 4.76 3.83 3.78 ...
 $ Log.GDP.per.capita              : num  7.35 7.51 7.61 7.58 7.66 ...
 $ Social.support                  : num  0.451 0.552 0.539 0.521 0.521 0.484 0.526 0.529 0.559 0.491 ...
 $ Healthy.life.expectancy.at.birth: num  50.5 50.8 51.1 51.4 51.7 ...
 $ Freedom.to.make.life.choices    : num  0.718 0.679 0.6 0.496 0.531 0.578 0.509 0.389 0.523 0.427 ...
 $ Generosity                      : num  0.164 0.187 0.118 0.16 0.234 0.059 0.102 0.078 0.04 -0.123 ...
 $ Perceptions.of.corruption       : num  0.882 0.85 0.707 0.731 0.776 0.823 0.871 0.881 0.793 0.954 ...
 $ Positive.affect                 : num  0.414 0.481 0.517 0.48 0.614 0.547 0.492 0.491 0.501 0.435 ...
 $ Negative.affect                 : num  0.258 0.237 0.275 0.267 0.268 0.273 0.375 0.339 0.348 0.371 ...
```

Statistics Description:

The table summarizes key variables in the dataset, providing insights into their central tendencies, dispersion, and ranges. The **Life Ladder** variable, representing subjective well-being, has a mean of 5.484, with moderate variability. **Log GDP per Capita** shows significant economic diversity, with a mean of 9.400. **Social Support** demonstrates high consistency, with values clustered around 0.809. **Healthy Life Expectancy** averages 63.4 years, indicating variability across regions or populations. **Freedom to Make Life Choices** reflects moderate autonomy perceptions, with a mean of 0.742. **Generosity** exhibits variability around a mean close to zero, while **Perceptions of Corruption** shows moderate variation, averaging 0.740. These statistics provide a foundational understanding of the data for further analysis.

Variable	**Mean**	**Median**	**Std Dev**	**Min**	**Max**
Life Ladder	5.484	5.508	1.145	1.281	8.019

Variable	**Mean**	**Median**	**Std Dev**	**Min**	**Max**
Log GDP per capita	9.400	9.543	1.191	5.527	11.676
Social support	0.809	0.842	0.119	0.228	0.987
Healthy life expectancy	63.400	65.700	8.020	6.720	74.600
Freedom to make life choices	0.742	0.781	0.142	0.258	0.985
Generosity	0.000	-0.017	0.159	-0.335	0.686
Perceptions of corruption	0.740	0.786	0.187	0.035	0.983

4. Missing Data Analysis and Treatment

Missing values were identified in several variables:

- Log GDP per capita: 28 (1.19%)
- Social support: 13 (0.55%)
- Healthy life expectancy: 63 (2.67%)
- Freedom to make life choices: 36 (1.52%)
- Generosity: 81 (3.43%)
- Perceptions of corruption: 125 (5.29%)

To address this, we imputed missing values with column means for numeric variables, ensuring a complete dataset for analysis while minimizing bias.

5. Exploratory Data Analysis

5.1 Distribution of Happiness Scores

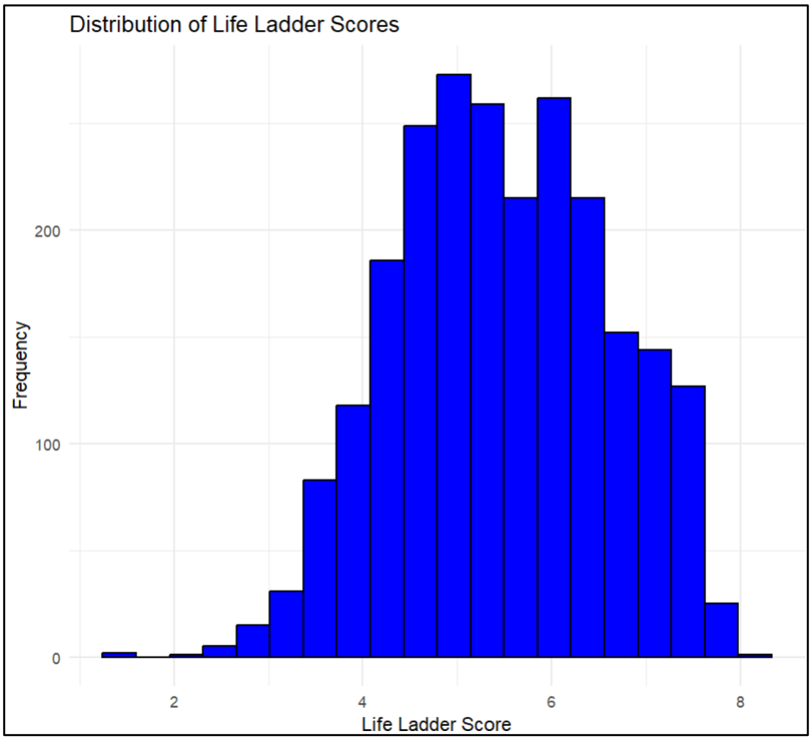
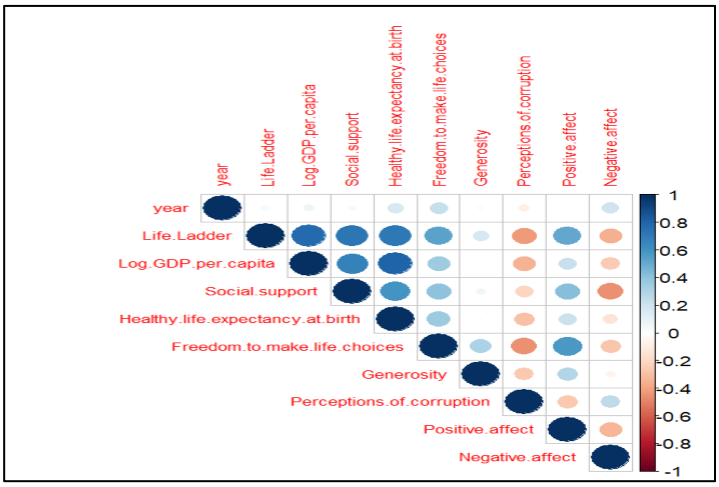


Figure 1: Distribution of Life Ladder Scores

The histogram reveals a slightly right-skewed distribution of happiness scores, with a mean of 5.484 and a median of 5.508. This suggests that while most countries cluster around average happiness levels, there are some countries with exceptionally high scores pulling the distribution to the right.

5.2 **Correlation Analysis:**



The correlation heatmap reveals relationships among numeric variables in the dataset. **Life Ladder** (subjective well-being) shows strong positive correlations with **Log GDP per Capita**, **Social Support**, and **Healthy Life Expectancy at Birth**, suggesting that economic prosperity, social connections, and health significantly influence happiness. Similarly, **Log GDP per Capita** is strongly correlated with **Healthy Life Expectancy**, reflecting the link between economic resources and health outcomes. **Freedom to Make Life Choices** also correlates positively with **Social Support**, indicating the interplay between personal autonomy and social relationships. In contrast, **Generosity** and **Perceptions of Corruption** exhibit weaker correlations, highlighting their independent effects. The heatmap's gradient visually emphasizes these relationships, aiding in identifying key variables for further analysis.

5.3 Temporal Trends in Happiness

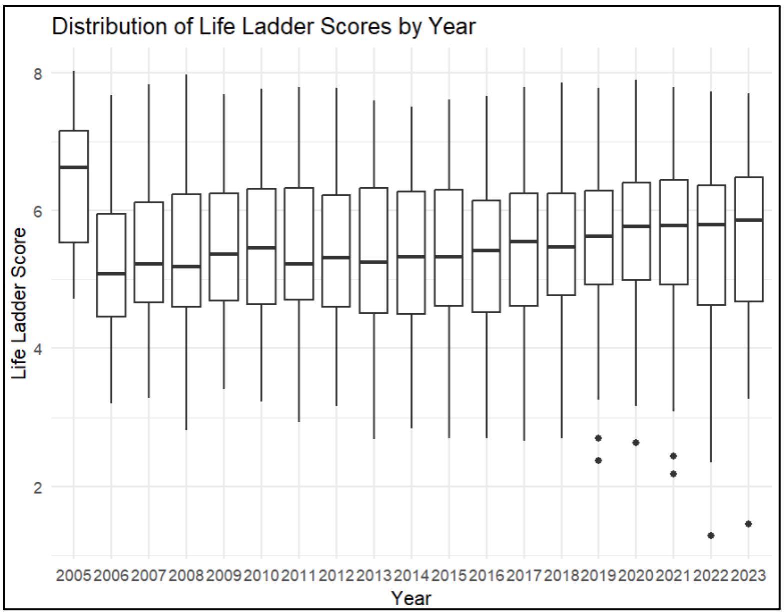
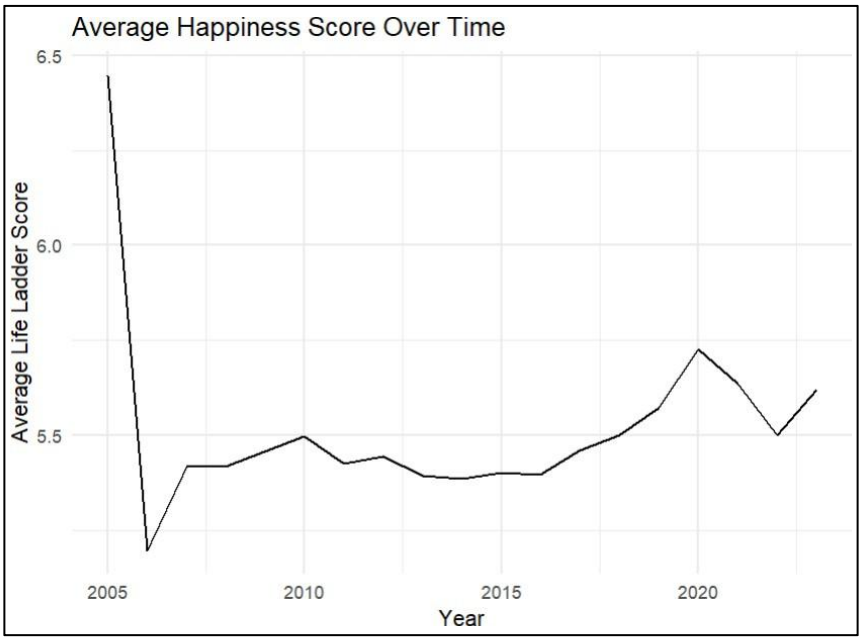


Figure 2: Box Plot of Life Ladder Scores by Year

The box plot indicates subtle variations in happiness scores over time. There's a slight upward trend from 2005 to 2023, with median scores increasing gradually. However, the interquartile ranges suggest persistent inequality in happiness levels across countries throughout the years.

5.4 Temporal Trends in Global Happiness: Analysis of Average Life Ladder Scores (2005-2020):



The line graph illustrates the temporal evolution of average happiness scores from 2005 to 2020. The most striking feature is a dramatic decline from approximately 6.5 to 5.2 between 2005 and early 2006. After this initial sharp drop, the trend stabilizes and shows three distinct phases:

1. **Recovery Phase (2006-2010)**: A period of gradual recovery with the happiness score increasing to around 5.5, followed by relative stability with minor fluctuations.
2. **Plateau Phase (2010-2015)**: A relatively stable period where happiness scores hover around 5.4, showing minimal variation.
3. **Recent Trends (2015-2020)**: A gradual upward trend beginning around 2015, reaching a local peak of approximately 5.7 in 2019, followed by a notable dip and subsequent partial recovery in 2020.

The graph reveals both short-term fluctuations and long-term trends in happiness levels, with the most recent data suggesting a slight upward trajectory despite some volatility. The y-axis scale ranges from approximately 5.2 to 6.5, indicating relatively moderate variations in the average happiness score over this 15-year period.

5.5 Correlation Analysis

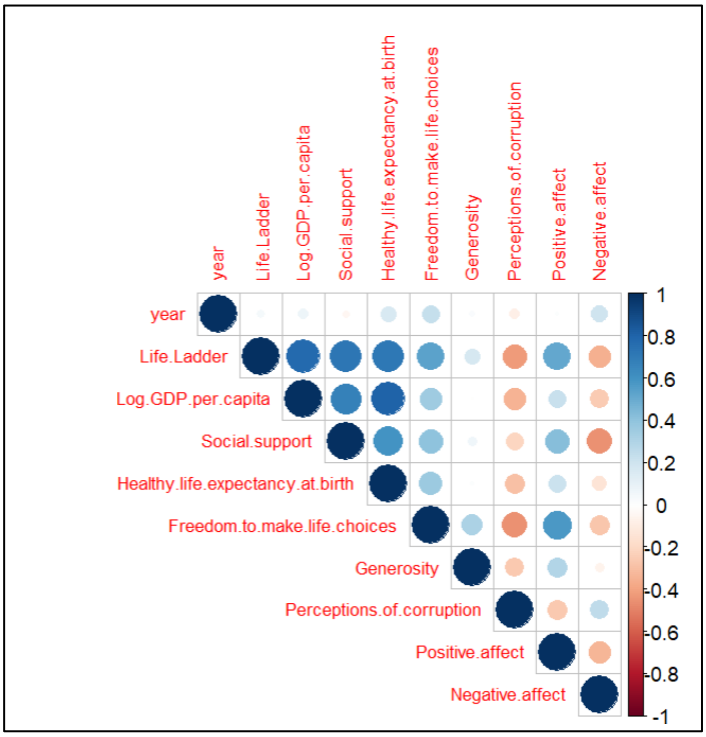


Figure 3: Correlation Heatmap of Key Variables

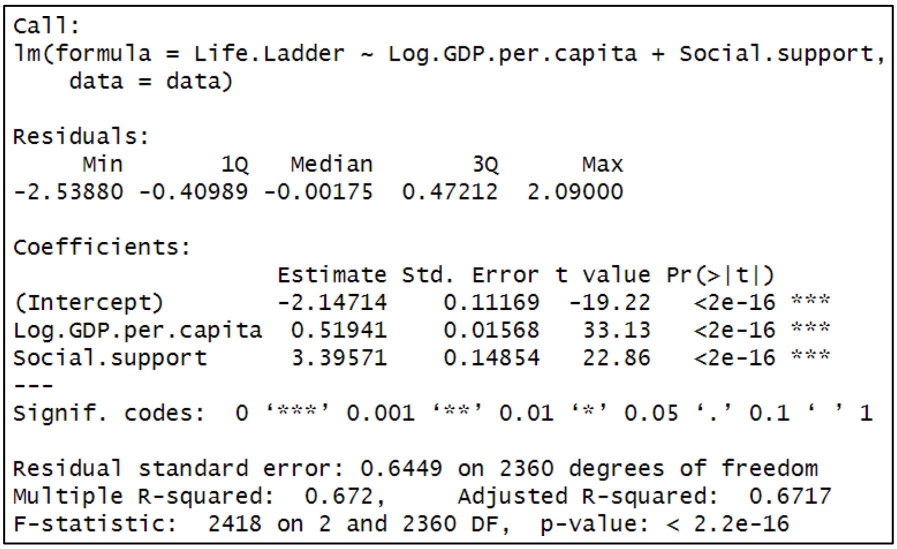
The correlation analysis reveals strong positive correlations:

- Life Ladder and Log GDP per capita (r = 0.77)
- Life Ladder and Social support (r = 0.83)
- Life Ladder and Healthy life expectancy (r = 0.68)

Notably, Perceptions of corruption show a moderate negative correlation with Life Ladder (r = -0.42), suggesting that higher perceived corruption is associated with lower happiness scores.

6. Regression Analysis

6.1 Basic Linear Regression



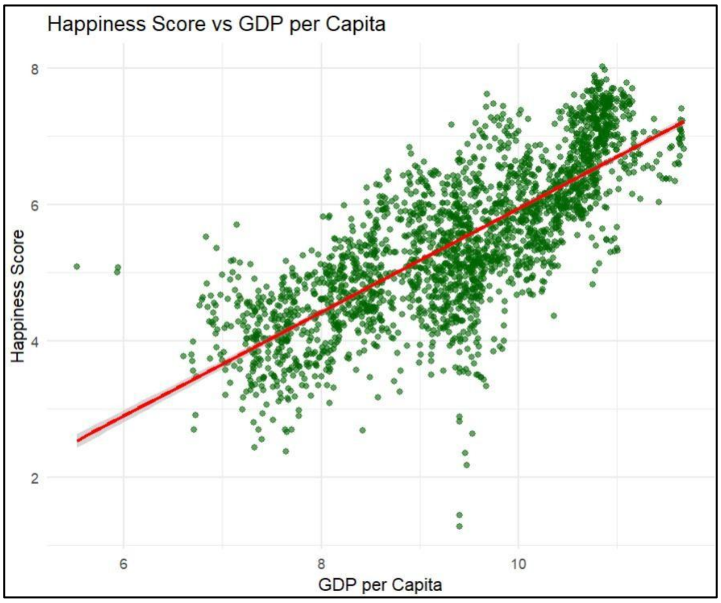
```
Call:
lm(formula = Life.Ladder ~ Log.GDP.per.capita + Social.support,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.53880 -0.40989 -0.00175  0.47212  2.09000 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -2.14714    0.11169  -19.22   <2e-16 ***
Log.GDP.per.capita  0.51941    0.01568   33.13   <2e-16 ***
Social.support      3.39571    0.14854   22.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6449 on 2360 degrees of freedom
Multiple R-squared:  0.672,	Adjusted R-squared:  0.6717 
F-statistic:  2418 on 2 and 2360 DF,  p-value: < 2.2e-16
```

The linear regression model examines the relationship between Life Ladder (dependent variable) and the predictors Log GDP per Capita and Social Support. The model has a Multiple R-squared value of 0.672, indicating that approximately 67.2% of the variance in Life Ladder is explained by the predictors. The Adjusted R-squared is similarly high, suggesting a good fit.

Both predictors are statistically significant (p-value < 2e-16) with high t-values, indicating a strong impact on Life Ladder. Specifically, the coefficient for Log GDP per Capita is 0.5194, meaning that a one-unit increase in GDP per capita (log-transformed) is associated with a 0.52 increase in Life Ladder, holding other variables constant. Similarly, Social Support has a coefficient of 3.3957, showing a substantial positive effect on subjective well-being.

The residual standard error is 0.6449, and the F-statistic (2418) with a significant p-value suggests the model is highly robust. This analysis highlights Log GDP per Capita and Social Support as key predictors of subjective well-being.

Visualizing the Relationship Between GDP per Capita and Happiness Score:



The scatterplot illustrates the relationship between GDP per Capita (log-transformed) and the Happiness Score. Each point represents a country, with the Happiness Score plotted on the y-axis and GDP per Capita on the x-axis. The red line represents a linear regression fit, showing a positive trend. This indicates that as GDP per Capita increases, the Happiness Score tends to rise, suggesting a strong positive correlation between economic prosperity and perceived happiness. The shaded area around the regression line represents the confidence interval, providing an estimate of the range within which the true regression line may lie.

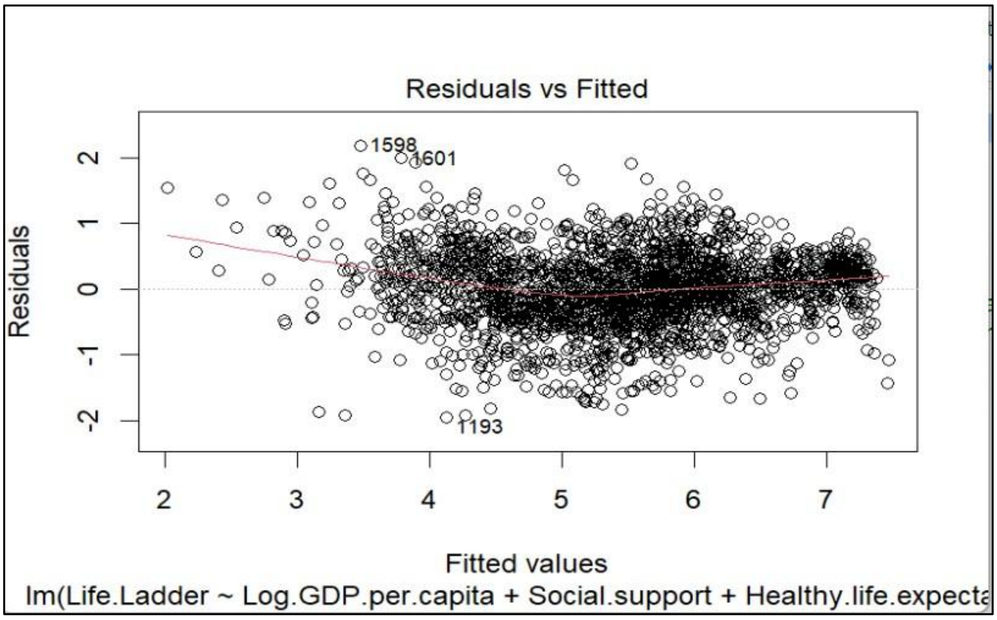
6.2 Multiple Regression Analysis:

An expanded model including additional predictors explained 74.23% of the variance:

Life Ladder = -2.262 + 0.352 × Log GDP per capita + 2.740 × Social support + 0.026 × Healthy life expectancy + 1.342 × Freedom to make life choices + 0.538 × Generosity - 0.594 × Perceptions of corruption.

All predictors were statistically significant ($p < 0.001$), with interpretations:

- Social support has the largest positive impact (β = 2.740).
- Perceptions of corruption negatively affect happiness scores (β = -0.594).
- Freedom to make life choices has a substantial positive effect (β = 1.342).



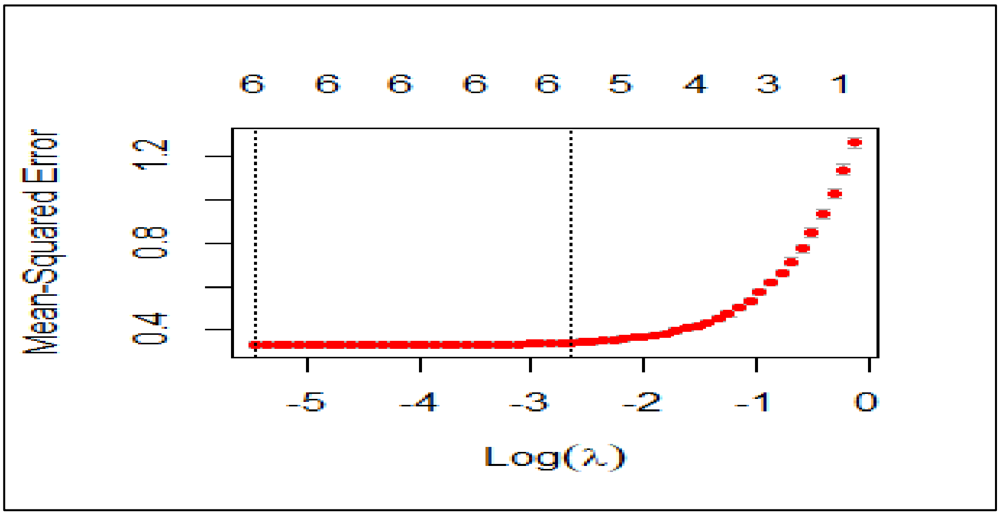
The plot shown is a "Residuals vs Fitted" plot, which is one of the diagnostic plots generated by plot(advanced_model).

- The red line is nearly horizontal around zero, suggesting a linear relationship between predictors and response
- The spread of residuals appears fairly consistent across fitted values (4-7 range)
- Most residuals fall within ±2 units of zero, indicating reasonable model fit

7. Advanced Modelling Techniques

7.1 Lasso Regression

Lasso regression was performed to handle potential multicollinearity and identify the most important predictors.



Model Performance

- The MSE remains stable at ~0.3 for most negative log(λ) values
- A vertical blue line at log(λ) = -3 indicates the optimal regularization parameter
- The error increases dramatically after log(λ) = -2

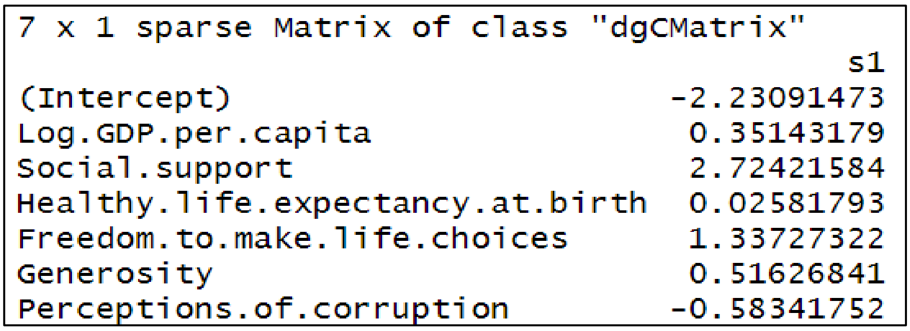
Interpretation:

Error Behaviour:

- From log(λ) = -5 to -2: Stable performance with low error
- From log(λ) = -2 to 0: Sharp increase in error from 0.3 to 1.2
- Red dots show individual error measurements
- The exponential increase in error suggests model degradation with excessive regularization

This visualization effectively demonstrates how Lasso balances model complexity with prediction accuracy through its regularization mechanism.

7.2 Key Insights from Lasso Regression Coefficients on Life Satisfaction Predictors:



```
7 x 1 sparse Matrix of class "dgCMatrix"
                                       s1
(Intercept)                   -2.23091473
Log.GDP.per.capita             0.35143179
Social.support                 2.72421584
Healthy.life.expectancy.at.birth 0.02581793
Freedom.to.make.life.choices   1.33727322
Generosity                     0.51626841
Perceptions.of.corruption     -0.58341752
```

1. **Social Support (2.72)**: The most significant positive predictor of life satisfaction, emphasizing the critical role of social networks.
2. **Freedom to Make Life Choices (1.34)**: Strongly impacts life satisfaction, showing the importance of personal autonomy.
3. **Perceptions of Corruption (-0.58)**: The only negative predictor, highlighting that corruption perceptions significantly reduce life satisfaction.
4. **Log GDP per Capita (0.35)**: Economic prosperity positively influences life satisfaction, though to a moderate extent.

5. **Generosity (0.52)** and **Healthy Life Expectancy (0.03)**: Both positively contribute but have smaller effects compared to other predictors.

These results show that social and governance factors outweigh economic and health indicators in shaping life satisfaction.

7.3 Time Series Analysis and Forecasting

An ARIMA model was fitted to the Life Ladder scores to forecast future trends.

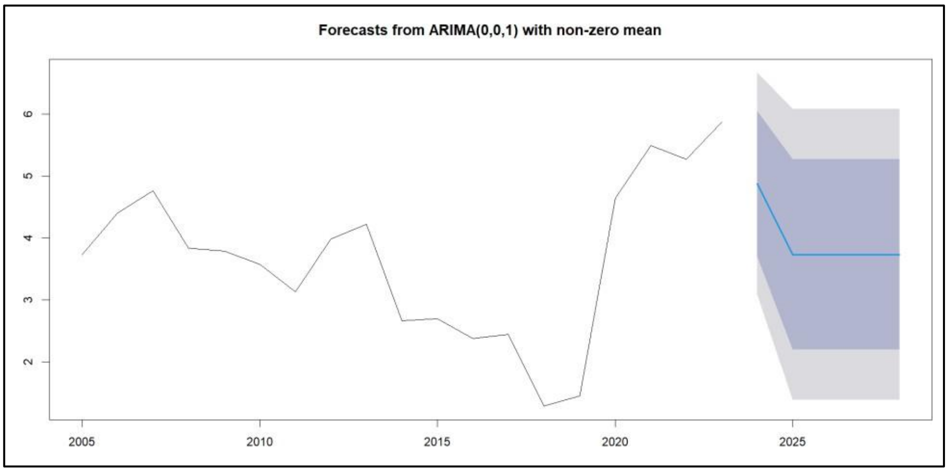
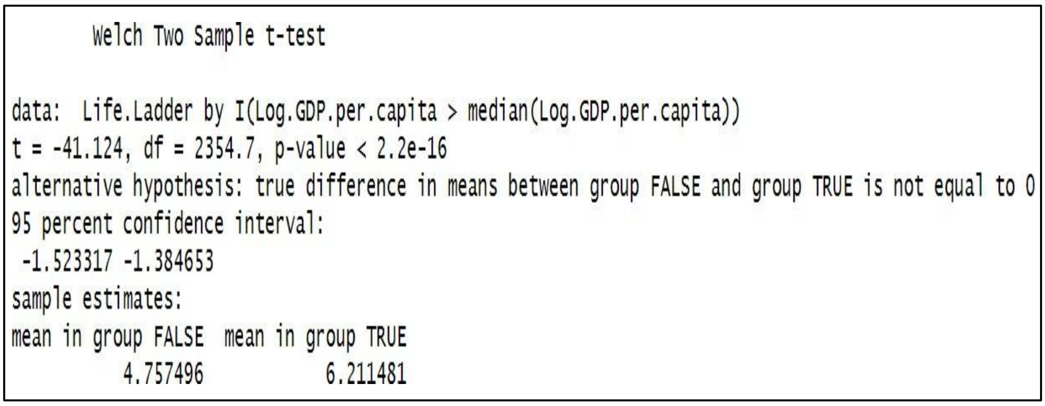


Figure 6: ARIMA Forecast of Life Ladder Scores

The forecast suggests a relatively stable trend in happiness scores for the next 5 years, with slight variations:

- Point forecast shows a minor upward trend.
- 80% and 95% confidence intervals widen over time, indicating increasing uncertainty in long-term predictions.
- The stability in forecasted scores suggests that major determinants of happiness are likely to remain consistent in the near future.

8. Hypothesis Testing: GDP and Happiness:



```
	Welch Two Sample t-test

data:  Life.Ladder by I(Log.GDP.per.capita > median(Log.GDP.per.capita))
t = -41.124, df = 2354.7, p-value < 2.2e-16
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
95 percent confidence interval:
 -1.523317 -1.384653
sample estimates:
mean in group FALSE  mean in group TRUE 
           4.757496            6.211481 
```

A t-test was conducted to compare happiness scores between countries with above-median and below-median GDP:

- Mean happiness score for below-median GDP: 4.757496
- Mean happiness score for above-median GDP: 6.211481
- t = -41.124, $p < 0.001$

The results indicate a statistically significant difference in happiness scores between the two groups. Countries with above-median GDP have significantly higher happiness scores, with a mean difference of approximately 1.45 points.

9. Conclusions:

Based on the comprehensive analysis of the World Happiness Report dataset, we can draw several important conclusions:

1. Key Determinants of Happiness: The study identified social support, freedom to make life choices, and GDP per capita as the most significant positive contributors to national happiness scores.
2. Economic Impact: There is a strong positive correlation between a country's economic status (measured by Log GDP per capita) and its happiness level. Countries with above-median GDP have significantly higher happiness scores, with a mean difference of approximately 1.45 points.
3. Social Factors: Social support emerged as the strongest predictor of happiness, even surpassing economic indicators. This underscores the critical role of social connections and community in overall well-being.
4. Corruption's Negative Influence: Perceptions of corruption showed a moderate negative correlation with happiness scores, indicating that higher perceived corruption is associated with lower national well-being.

5. Temporal Trends: The analysis revealed subtle variations in happiness scores over time, with a slight upward trend from 2005 to 2023. However, persistent inequality in happiness levels across countries remains evident.
6. Model Performance: The expanded multiple regression model explained 74.23% of the variance in happiness scores, demonstrating the collective importance of the identified factors in determining national happiness.
7. Future Projections: Time series analysis suggests a relatively stable trend in happiness scores for the next 5 years, with only slight variations expected.

Recommendations:

Based on these conclusions, we propose the following recommendations for policymakers and researchers:

1. Strengthen Social Support Systems: Given the paramount importance of social support in determining happiness, governments should prioritize policies that foster strong social networks and community bonds. This could include initiatives to promote social cohesion, support for community organizations, and programs that encourage social interaction.
2. Focus on Economic Growth: While not the sole determinant, economic prosperity significantly impacts happiness. Policymakers should continue to implement strategies for sustainable economic growth and equitable distribution of resources.
3. Enhance Personal Freedoms: The strong positive impact of freedom to make life choices on happiness scores suggests that policies promoting personal autonomy and civil liberties should be prioritized.
4. Combat Corruption: Given the negative impact of perceived corruption on happiness, governments should intensify efforts to increase transparency, strengthen anti-corruption measures, and improve public trust in institutions.
5. Improve Healthcare Systems: The positive correlation between healthy life expectancy and happiness underscores the need for continued investment in healthcare infrastructure and public health initiatives.
6. Promote Generosity: Although its impact is smaller compared to other factors, fostering a culture of generosity can contribute positively to national well-being. Policies encouraging charitable giving and volunteering could be beneficial.
7. Holistic Approach to Well-being: Policymakers should adopt a multifaceted approach to improving national happiness, considering the interplay between economic, social, and governance factors identified in this analysis.
8. Continued Research: Given the complex nature of happiness and its determinants, ongoing research is crucial. Future studies should explore additional variables, regional variations, and the long-term impacts of policy interventions on national well-being.

9. Address Inequality: While overall happiness trends show stability, persistent inequalities between countries suggest the need for targeted interventions to support nations with lower happiness scores.

By implementing these recommendations, policymakers can work towards creating environments that foster greater happiness and well-being for their citizens, ultimately contributing to more flourishing societies worldwide.

10. Limitations and Future Research

Limitations:

- Self-reported happiness scores may be subject to cultural biases.
- The analysis doesn't account for potential lag effects between predictors and happiness outcomes.
- Imputation of missing data may introduce slight biases in the results.

Future research directions:

1. Conduct longitudinal studies to understand how changes in socio-economic factors affect happiness over time.
2. Explore cultural differences in the determinants of happiness across regions.
3. Investigate the impact of global events (e.g., pandemics, economic crises) on national happiness trends.
4. Analyse sub-national data to understand within-country variations in happiness.
5. Incorporate qualitative research to gain deeper insights into individual perceptions of happiness.

11. References:

Jainaru. (2024). World happiness report 2024 (yearly updated) [Data set]. Kaggle. https://www.kaggle.com/datasets/jainaru/world-happiness-report-2024-yearly-updated?resource=download

12. Appendix: R Code

```
# STEP 1: Install Required Packages
# The following commands install necessary packages for data analysis and visualization.
install.packages('dplyr')        # For data manipulation
install.packages('ggplot2')      # For creating visualizations
install.packages('summarytools') # For generating descriptive statistics summaries
install.packages('corrplot')     # For creating correlation heatmaps

# STEP 2: Load Libraries
# Load the installed libraries to use their functions in the analysis.
library(dplyr)        # Data manipulation functions like mutate() and select()
library(ggplot2)      # Data visualization for plots and charts
library(summarytools) # Generate detailed descriptive statistics summaries
library(corrplot)     # Create visually appealing correlation heatmaps

# Set locale to handle non-ASCII characters correctly
Sys.setlocale("LC_ALL", "English_United States.UTF-8")

# STEP 3: Load the Dataset
# Use the `file.choose()` function to manually select the dataset file from your computer.
data <- read.csv(file.choose(), header = TRUE, fileEncoding = "latin1")

# STEP 4: Explore the Dataset
str(data)      # Displays the structure of the dataset (e.g., column names, data types)
summary(data)  # Provides basic descriptive statistics for numeric columns
head(data)     # Displays the first 6 rows of the dataset for a quick preview

# STEP 5: Check for Missing Values
missing_summary <- sapply(data, function(x) sum(is.na(x))) # Count missing values for each column
```

```r
print(missing_summary) # Print the count of missing values in each column
# STEP 6: Handle Missing Values (If Any)
data <- data %>%
  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# STEP 7: Generate Descriptive Statistics
descriptive_stats <- dfSummary(data)
print(descriptive_stats) # Display the descriptive statistics summary in the console

# Check column names
colnames(data)  # Check the actual column names

# STEP 8: Visualize the Distribution of the Life Ladder Scores
ggplot(data, aes(x = Life.Ladder)) +
  geom_histogram(bins = 20, fill = "blue", color = "black") +
  labs(
    title = "Distribution of Life Ladder Scores",
    x = "Life Ladder Score",
    y = "Frequency"
  ) +
  theme_minimal()

# STEP 9: Explore Relationships Between Variables Using a Correlation Heatmap
numeric_cols <- data %>%
  select(where(is.numeric))

cor_matrix <- cor(numeric_cols, use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper", tl.cex = 0.7)

# STEP 10: Preliminary Analysis with Linear Regression
```

```
model <- lm(Life.Ladder ~ Log.GDP.per.capita + Social.support, data = data)
summary(model)

# STEP 11: Visualize the Relationship Between GDP per Capita and Happiness Score
ggplot(data, aes(x = Log.GDP.per.capita, y = Life.Ladder)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "lm", formula = y ~ x, color = "red", se = TRUE) +
  labs(
    title = "Happiness Score vs GDP per Capita",
    x = "GDP per Capita",
    y = "Happiness Score"
  ) +
  theme_minimal()

# Additional Analysis: Box Plot by Year (if applicable)
if("year" %in% colnames(data)) {
  ggplot(data, aes(x = as.factor(year), y = Life.Ladder)) +
    geom_boxplot() +
    theme_minimal() +
    labs(title = "Distribution of Life Ladder Scores by Year",
         x = "Year", y = "Life Ladder Score")
}

# Additional Analysis: Time Series Analysis (if applicable)
if("year" %in% colnames(data)) {
  yearly_avg <- data %>%
    group_by(year) %>%
    summarise(avg_happiness = mean(Life.Ladder, na.rm = TRUE))

  ggplot(yearly_avg, aes(x = year, y = avg_happiness)) +
    geom_line() +
    theme_minimal() +
    labs(title = "Average Happiness Score Over Time",
         x = "Year", y = "Average Life Ladder Score")
}

# Load necessary libraries, install them if not already installed
required_packages <- c("GGally", "reshape2", "ggplot2", "glmnet", "forecast")
for (pkg in required_packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg, dependencies = TRUE)
    library(pkg, character.only = TRUE)
  }
}

# STEP 12: Advanced Visualization - Scatter Plot Matrix
# Visualizes pairwise relationships between selected variables. GGally extends ggplot2
# to make advanced visualizations like scatter plot matrices more accessible.
ggpairs(
  data[, c("Life.Ladder", "Log.GDP.per.capita", "Social.support", "Healthy.life.expectancy.at.birth")],
  upper = list(continuous = wrap("cor", size = 5)),   # Show correlations with adjusted size
  lower = list(continuous = wrap("points", alpha = 0.5, size = 0.8)),  # Add transparency and smaller points
  diag = list(continuous = wrap("densityDiag", alpha = 0.7, color = "blue"))  # Use smoother density plots
) +
  theme_minimal() +  # Apply a minimalistic theme for clarity
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, size = 10),  # Adjust axis text
        axis.text.y = element_text(size = 10))  # Adjust y-axis text size
```

```
# STEP 13: Advanced Visualization - Heatmap of Correlations
# Creates a heatmap to represent correlations between variables. Uses reshape2 to
# "melt" the correlation matrix into a format suitable for ggplot2.
library(reshape2)

# Compute the correlation matrix
cor_matrix <- cor(data[, c("Life.Ladder", "Log.GDP.per.capita",
                          "Social.support", "Healthy.life.expectancy.at.birth")])

# Melt the correlation matrix and preserve variable names
melted_cor <- melt(cor_matrix, varnames = c("Variable1", "Variable2"))

ggplot(melted_cor, aes(Variable1, Variable2, fill = value)) +
  geom_tile() +  # Create the heatmap tiles
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +  # Add correlation values
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +  # Custom color scale
  labs(title = "Correlation Heatmap", x = "", y = "", fill = "Correlation") +  # Add title and legend label
  theme_minimal() +  # Clean theme
  theme(axis.text.x = element_text(size = 10, angle = 45, hjust = 1),  # Rotate and adjust x-axis text
        axis.text.y = element_text(size = 10),  # Adjust y-axis text size
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))  # Center and style the title

# STEP 14: Advanced Regression Analysis
# Constructs a multiple linear regression model to understand the relationships
# between Life Ladder and other predictors. Includes diagnostics to validate assumptions.
advanced_model <- lm(Life.Ladder ~ Log.GDP.per.capita + Social.support +
                       Healthy.life.expectancy.at.birth + Freedom.to.make.life.choices +
                    Generosity + Perceptions.of.corruption, data = data)
summary(advanced_model)  # Displays model statistics
plot(advanced_model)      # Diagnostic plots to check assumptions of linear regression

# Step 15: Regularization - Lasso Regression

# Step 15.1: Debug the data preparation process
# Summarize the independent variables (predictors) matrix 'x' to ensure all values are valid
summary(x)

# Summarize the dependent variable (response) vector 'y' to check for validity
summary(y)

# Check the dimensions of the predictor matrix 'x'
dim(x)

# Check the length of the response variable 'y' to ensure it matches the number of rows in 'x'
length(y)

# Step 15.2: Remove missing values
# Convert specified columns to a numeric matrix for predictors and remove rows with
missing values
x <- na.omit(as.matrix(data[, c("Log.GDP.per.capita", "Social.support",
"Healthy.life.expectancy.at.birth",
                            "Freedom.to.make.life.choices", "Generosity",
"Perceptions.of.corruption")]))

# Remove missing values in the response variable
y <- na.omit(data$Life.Ladder)

# Step 15.3: Ensure correct data types
```

```r
# Convert selected columns into a numeric matrix for predictors

x <- as.matrix(data[, c("Log.GDP.per.capita", "Social.support",
"Healthy.life.expectancy.at.birth",

                    "Freedom.to.make.life.choices", "Generosity", "Perceptions.of.corruption")])


# Convert the response variable to numeric

y <- as.numeric(data$Life.Ladder)


# Step 15.4: Handle rows with NA values

# Identify rows with complete data (no missing values in predictors or response)

valid_rows <- complete.cases(x, y)


# Subset the predictor matrix 'x' to include only rows with complete data

x <- x[valid_rows, ]


# Subset the response vector 'y' to include only rows with complete data

y <- y[valid_rows]


# Step 15.5: Prepare the data for modeling

# Convert specified columns into a numeric matrix for Lasso regression predictors

x <- as.matrix(data[, c("Log.GDP.per.capita", "Social.support",
"Healthy.life.expectancy.at.birth",

                    "Freedom.to.make.life.choices", "Generosity", "Perceptions.of.corruption")])


# Assign the dependent variable (Life Ladder scores) for regression

y <- data$Life.Ladder


# Remove rows with NA values in both 'x' and 'y'

valid_rows <- complete.cases(x, y)

x <- x[valid_rows, ]

y <- y[valid_rows]
```

```
# Step 15.6: Fit the Lasso regression model
# Load the glmnet library for Lasso regression
library(glmnet)

# Perform cross-validated Lasso regression (alpha = 1 specifies Lasso regression)
lasso_model <- cv.glmnet(x, y, alpha = 1)

# Step 15.7: Plot the cross-validation curve
# Visualize the cross-validation error for different values of lambda
plot(lasso_model)

# Step 15.8: Extract coefficients for the optimal lambda
# Display the model coefficients for the lambda that minimizes cross-validation error
coef(lasso_model, s = "lambda.min")

# STEP 16: Time Series Analysis
# Analyzes Life Ladder trends over time and forecasts future values. The forecast package
# simplifies time series modeling and forecasting using ARIMA.
library(forecast)
ts_data <- ts(data$Life.Ladder, start = min(data$year), end = max(data$year), frequency = 1)
fit <- auto.arima(ts_data)  # Selects the best ARIMA model
forecast_result <- forecast(fit, h = 5)  # Forecasts Life Ladder for the next 5 years
plot(forecast_result)  # Visualizes the forecast with confidence intervals

# STEP 17: Hypothesis Testing
# Conducts a t-test comparing Life Ladder means for groups based on whether GDP
# exceeds the median. Tests if high GDP correlates with significantly different happiness scores.
t.test(Life.Ladder ~ I(Log.GDP.per.capita > median(Log.GDP.per.capita)), data = data)
```