Happiness around the World

Josh Stremmel, Jooan Choi, Namrata Singh

Table of contents

Introduction

For this project, we set out to explore what drives happiness in different countries and how it changes over time or varies by region. We used data from the World Happiness Report, a well-known dataset created by the United Nations Sustainable Development Solutions Network. It's packed with information on happiness scores for about 155 countries, along with factors like income, social support, and personal freedoms. The dataset is freely available on Kaggle, and we worked with data from 2017 to 2019, last updated in 2020. The happiness score is a number that sums up how happy people are in each country, and we looked at things like GDP per capita, social support, and freedom to figure out what affects it. We also added region labels for a few countries to see how happiness differs across places like Europe or Asia.

Methodology

The World Happiness Report data is super detailed, but it needed some cleanup before we could dive in. Each year (2017, 2018, and 2019) had different column names, like “Happiness.Score” in 2017 but just “Score” in 2018 and 2019. That made things tricky, so we used R packages like dplyr and tidyr to standardize the names. We also noticed some countries were listed differently, like “Trinidad & Tobago” in one year and “Trinidad and Tobago” in another, so we made those consistent. Plus, there were missing values, especially in 2018’s “Perceptions of corruption” column, where some entries were “N/A.” We fixed that by turning “N/A” into proper NA values and removing rows with missing data.

One big decision was how to handle regions. We didn’t have region info for every country, so we created a list of 17 countries like Norway, India, and Brazil and assigned them regions such as Western Europe or South Asia. Countries without region info were labeled “Other.” We decided to use all three years of data (2017-2019) to get a sense of how happiness trends over time, thinking it would give us a fuller picture.

Exploratory Data Analysis

First, we wrangled the dataset to make it easier to analyze. This meant fixing column names, standardizing country names, and getting rid of missing values, leaving us with 466 rows of data across 2017 to 2019. We also added a region column for 17 countries to compare places like Europe and Asia. To dig deeper, we created smaller datasets, like one for the happiest countries in 2019, to study what factors boost happiness, how regions differ, and whether happiness scores changed over the years.

Quantitative

In our number-crunching phase, we looked at how the happiness score relates to factors like GDP per capita, social support, life expectancy, freedom, generosity, and perceptions of corruption. We made graphs, like scatter plots and a correlation heatmap, to see how these factors connect. We also ran a regression analysis to figure out which factors have the biggest impact on happiness scores. Basically, we tested different combinations to get a clear idea of what’s driving happiness.

Average Happiness Score by Region

Table 1: Average Happiness Score by Region

Region	Count	Average Score	SD Score
East Asia	9	5.671000	0.3272751
Latin America	6	6.404500	0.2018938
North America	3	7.307333	0.0261024
Oceania	6	7.288167	0.0352274
Other	415	5.219046	1.0189858
South Asia	3	4.173333	0.1506929
Sub-Saharan Africa	3	4.758333	0.0612073
Western Europe	21	7.496810	0.1101134

The table represents the average happiness scores of different regions of the world, based on the countries included in the dataset from 2017 to 2019. Western Europe has the highest average score (7.50), followed by North America and Oceania. On the other hand, regions such as sub-Saharan Africa and South Asia have significantly lower levels of happiness. The standard deviation column helps highlight how consistent or varied the scores are within each region.

Correlation Matrix of Happiness Factors

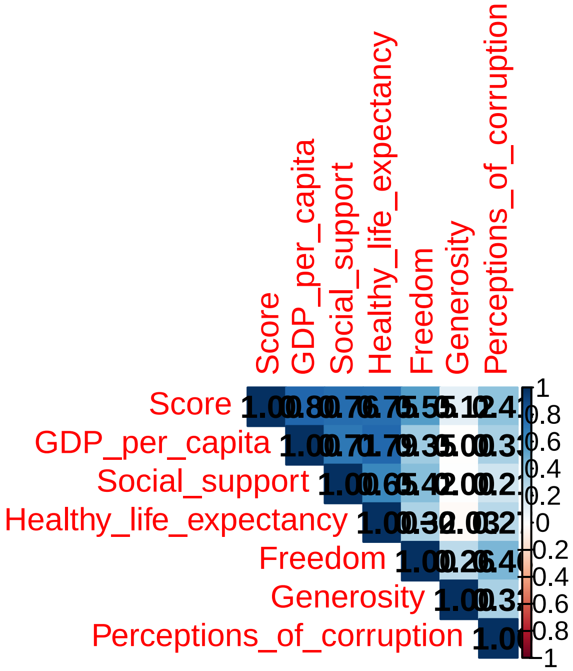


Figure 1: Correlation Matrix of Happiness Factors

This correlation heatmap shows the relationship between the happiness score and the variables that contribute to it. A strong positive correlation was observed between happiness and GDP per capita (0.80), social support (0.76), and healthy life expectancy (0.75), indicating that they are the main drivers of happiness. Other factors such as freedom and generosity show moderate to weak correlations, while perception of corruption shows the weakest positive correlation (0.12).

Regression Analysis of Happiness Score

This table presents the results of multiple linear regression analysis using the happiness score as a dependent variable. All six predictors, including GDP, social support, healthy life expectancy, freedom, generosity, and perception of corruption, are statistically significant, with GDP, social support, and freedom having the greatest impact. A high R^2 value (0.794) indicates that this model accounts for a significant portion of the variation in the happiness score. This supports the idea that economic and social conditions are key drivers of people's happiness.

Happiness Score vs. GDP per Capita

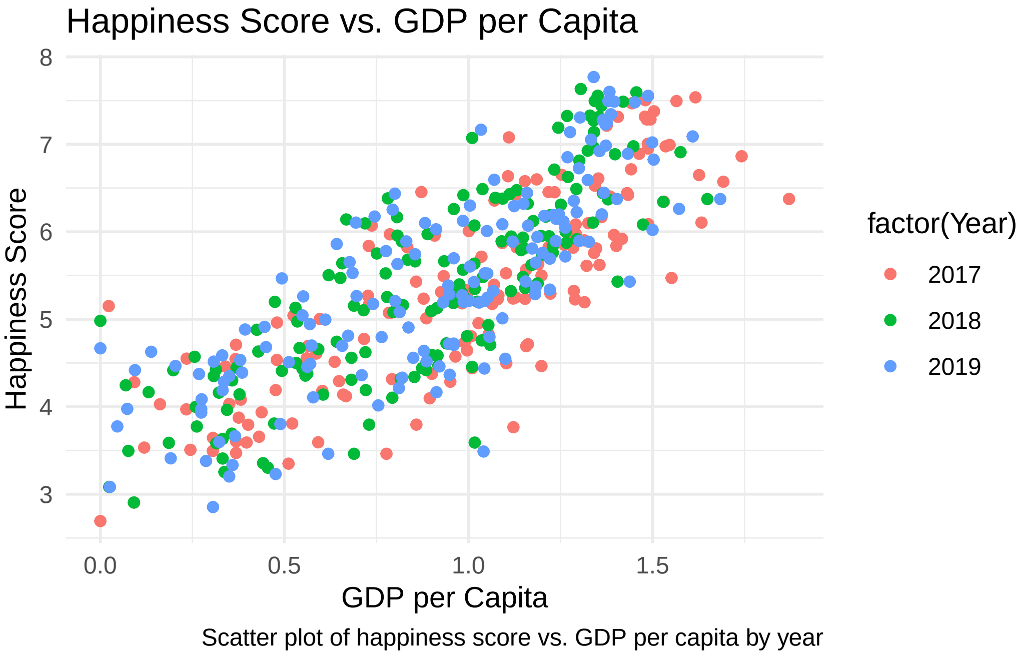


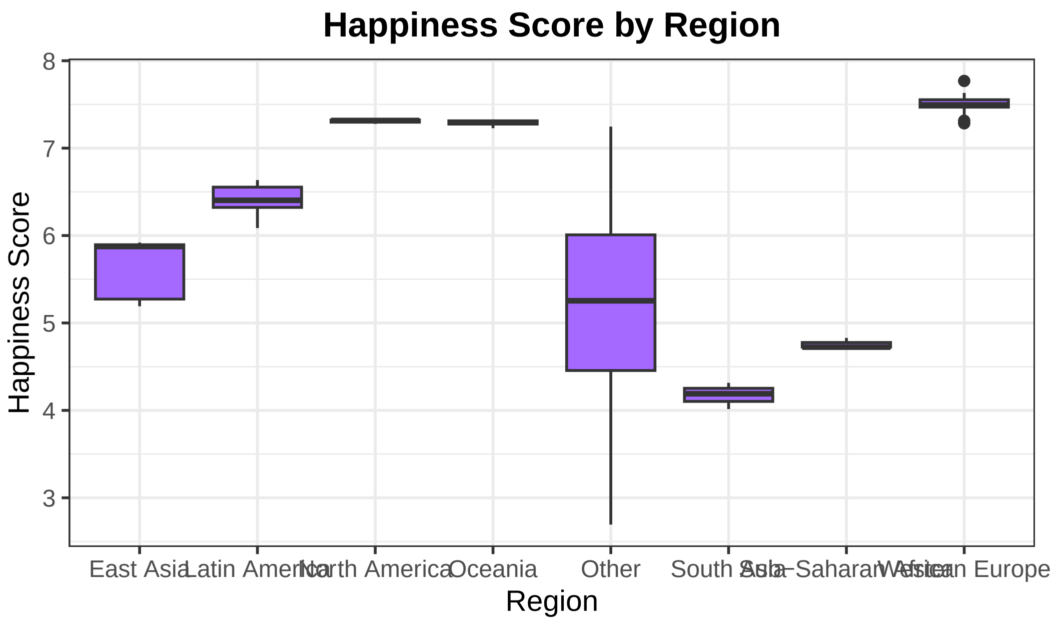
Figure 2: Happiness Score vs. GDP per Capita

Table 2: Regression Analysis of Happiness Score

```
=====================================================
                              Dependent variable:
                          ---------------------------
                                     Score
-----------------------------------------------------
GDP_per_capita                     0.854***
                                    (0.109)

Social_support                     1.107***
                                    (0.122)

Healthy_life_expectancy            1.057***
                                    (0.155)

Freedom                            1.445***
                                    (0.190)

Generosity                          0.481**
                                    (0.224)

Perceptions_of_corruption          0.830***
                                    (0.293)

Constant                           1.793***
                                    (0.112)

-----------------------------------------------------
Observations                          466
R2                                   0.794
Adjusted R2                          0.791
Residual Std. Error            0.511 (df = 459)
F Statistic                294.145*** (df = 6; 459)
=====================================================
Note:                     *p<0.1; **p<0.05; ***p<0.01
```

This scatterplot shows the relationship between GDP and happiness scores for all countries and years (2017-2019). Each dot represents the country for a particular year and is separated by the year. Positive trends are evident: countries with higher GDP per capita tend to report higher happiness scores. The distribution of dots also suggests some variation, indicating that wealth is important but not the only factor influencing happiness.

Happiness Score by Region



Boxplot of happiness scores across regions

Figure 3: Happiness Score by Region

This boxplot compares happiness scores for regions around the world. Western Europe, North America, and Oceania show the highest median happiness levels with relatively low variability. On the other hand, regions such as sub-Saharan Africa and South Asia report the lowest scores. The widespread distribution of values in the ‘Other’ category reflects the various countries included in it. This visualization clearly shows how the regional context affects overall well-being.

Happiness Score by Year

This box plot shows the distribution of happiness scores for each year from 2017 to 2019. The medians remained fairly stable over time, suggesting that global happiness levels did not

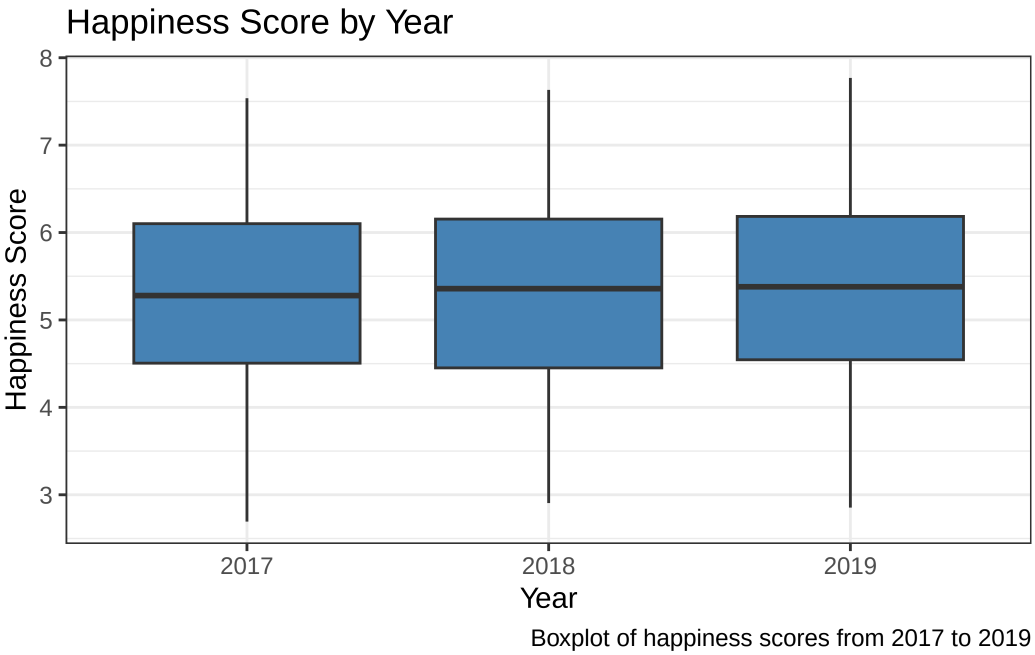


Figure 4: Happiness Score by Year

undergo significant changes during this period. Consistent spread and extent by year indicate that individual countries may fluctuate, but overall global happiness patterns remained stable.

Top 10 Happiest Countries in 2019

The bar chart shows the 10 countries with the highest happiness scores in 2019. Finland, Denmark, and Norway top the list, continuing the trend of Nordic countries ranking highest in world happiness. Their scores were all closely grouped with 7.5 points or higher, showing high performance in factors such as income, social support, and quality of life.

Results and Conclusion

Starting with a clean and standardized dataset from the World Happiness Report (2017-2019), we explored the factors that make countries happy and how happiness varies by year and region. We first made sure the data was usable—fixing inconsistent country names, aligning column names across years, and removing missing values. In addition, we added local labels for selected countries to allow regional comparisons.

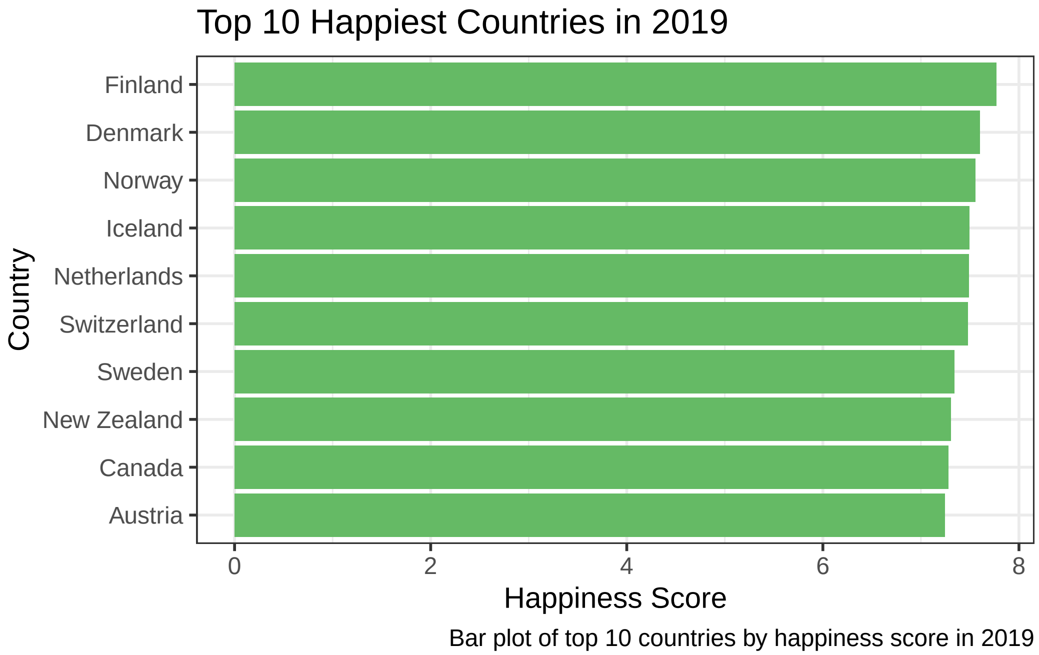


Figure 5: Top 10 Happiest Countries in 2019

Our results showed that GDP, social support, and healthy life expectancy were the most important factors associated with increased happiness scores. This was supported by both correlation analysis and regression modeling. Western European countries such as Finland and Denmark consistently ranked highest, probably because they performed well in these key areas.

Boxplots by region showed distinct differences in happiness levels in regions such as sub-Saharan Africa and South Asia, where on average they were low. Over the course of three years, the happiness score remained fairly stable, suggesting that there was no significant change worldwide during this period.

In conclusion, our analysis confirms that economic power alone cannot explain happiness. Countries with strong social systems, good health outcomes, and greater freedom tend to be happier overall. Therefore, improving happiness means focusing on the welfare, trust, and support of society as well as wealth.

References

Sustainable Development Solutions Network (2020). World Happiness Report [Dataset]. https://www.kaggle.com/datasets/unsdsn/world-happiness

Code Appendix

```
library(dplyr)
library(readr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(stargazer)
library(kableExtra)
# Importing the data from github
data_2017 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo
data_2018 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo
data_2019 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo

# Fixing column names
data_2017 <- data_2017 %>%
  rename(
    Country = "Country",
    Score = "Happiness.Score",
    GDP_per_capita = "Economy..GDP.per.Capita.",
    Social_support = "Family",
    Healthy_life_expectancy = "Health..Life.Expectancy.",
    Freedom = "Freedom",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Trust..Government.Corruption."
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

data_2018 <- data_2018 %>%
  rename(
    Country = "Country or region",
    Score = "Score",
    GDP_per_capita = "GDP per capita",
    Social_support = "Social support",
    Healthy_life_expectancy = "Healthy life expectancy",
    Freedom = "Freedom to make life choices",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Perceptions of corruption"
  ) %>%
  mutate(
    Perceptions_of_corruption = na_if(Perceptions_of_corruption, "N/A"),
    Perceptions_of_corruption = as.numeric(Perceptions_of_corruption)
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

data_2019 <- data_2019 %>%
  rename(
    Country = "Country or region",
    Score = "Score",
    GDP_per_capita = "GDP per capita",
    Social_support = "Social support",
    Healthy_life_expectancy = "Healthy life expectancy",
    Freedom = "Freedom to make life choices",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Perceptions of corruption"
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

# Add Year column
data_2017$Year <- 2017
data_2018$Year <- 2018
data_2019$Year <- 2019
combined_data <- bind_rows(data_2017, data_2018, data_2019)

# Fixing country name
combined_data <- combined_data %>%
  mutate(Country = recode(Country,
                          "Trinidad & Tobago" = "Trinidad and Tobago",
                          "Hong Kong S.A.R., China" = "Hong Kong",
                          "Taiwan Province of China" = "Taiwan",
                          "Macedonia" = "North Macedonia"
  ))

# Remove N/A
combined_data <- combined_data %>%
  drop_na()

# Check Data
#str(combined_data)
#summary(combined_data)

# Mapping regional data
region_data <- tibble(
```

```r
  Country = c("Norway", "Denmark", "Iceland", "Switzerland", "Finland", "Netherlands",
              "Canada", "New Zealand", "Sweden", "Australia", "India", "China",
              "Japan", "South Korea", "Brazil", "Argentina", "South Africa"),
  Region = c("Western Europe", "Western Europe", "Western Europe", "Western Europe",
             "Western Europe", "Western Europe", "North America", "Oceania",
             "Western Europe", "Oceania", "South Asia", "East Asia", "East Asia",
             "East Asia", "Latin America", "Latin America", "Sub-Saharan Africa")
)

# Merging regional data
combined_data <- combined_data %>%
  left_join(region_data, by = "Country") %>%
  mutate(Region = replace_na(Region, "Other"))

# Summarize stat
summary_stats <- combined_data %>%
  summarise(across(c(Score, GDP_per_capita, Social_support, Healthy_life_expectancy,
                     Freedom, Generosity, Perceptions_of_corruption),
                   list(mean = ~mean(., na.rm = TRUE),
                        sd = ~sd(., na.rm = TRUE),
                        min = ~min(., na.rm = TRUE),
                        max = ~max(., na.rm = TRUE))))

# Correlation
cor_matrix <- cor(combined_data %>% select(-Country, -Year, -Region))

# Regression model
model <- lm(Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
            Freedom + Generosity + Perceptions_of_corruption, data = combined_data)

# Summarize by region
region_summary <- combined_data %>%
  group_by(Region) %>%
  summarise(
    Count = n(),
    Avg_Score = mean(Score, na.rm = TRUE),
    SD_Score = sd(Score, na.rm = TRUE)
  )
region_summary %>%
  kable(col.names = c("Region", "Count", "Average Score", "SD Score"),
        caption = "Average Happiness Score by Region") %>%
  kable_classic()
```

```r
corrplot(cor_matrix, method = "color", type = "upper", addCoef.col = "black")
stargazer(model, type = "text")
ggplot(combined_data, aes(x = GDP_per_capita, y = Score, color = factor(Year))) +
  geom_point() +
  labs(title = "Happiness Score vs. GDP per Capita", x = "GDP per Capita", y = "Happiness Sco
       caption = "Scatter plot of happiness score vs. GDP per capita by year") +
  theme_minimal()
ggplot(combined_data, aes(x = Region, y = Score)) +
  geom_boxplot(fill = "#A569FF") +
  labs(title = "Happiness Score by Region", x = "Region", y = "Happiness Score",
       caption = "Boxplot of happiness scores across regions") +
  theme_bw() +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
ggplot(combined_data, aes(x = factor(Year), y = Score, group = Year)) +
  geom_boxplot(fill = "#4682B4") +
  labs(title = "Happiness Score by Year", x = "Year", y = "Happiness Score",
       caption = "Boxplot of happiness scores from 2017 to 2019") +
  theme_bw()
top_10 <- combined_data %>% filter(Year == 2019) %>% slice_max(Score, n = 10)
ggplot(top_10, aes(x = reorder(Country, Score), y = Score)) +
  geom_bar(stat = "identity", fill = "#65BA65") +
  coord_flip() +
  labs(title = "Top 10 Happiest Countries in 2019", x = "Country", y = "Happiness Score",
       caption = "Bar plot of top 10 countries by happiness score in 2019") +
  theme_bw()
# Load packages
library(dplyr)
library(readr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(stargazer)
library(kableExtra)

# Importing the data from github
data_2017 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo
data_2018 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo
data_2019 <- read_csv("https://raw.githubusercontent.com/Stat184-Spring2025/Course_Project_Jo

# Fixing column names
data_2017 <- data_2017 %>%
  rename(
    Country = "Country",
    Score = "Happiness.Score",
    GDP_per_capita = "Economy..GDP.per.Capita.",
    Social_support = "Family",
    Healthy_life_expectancy = "Health..Life.Expectancy.",
    Freedom = "Freedom",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Trust..Government.Corruption."
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

data_2018 <- data_2018 %>%
  rename(
    Country = "Country or region",
    Score = "Score",
    GDP_per_capita = "GDP per capita",
    Social_support = "Social support",
    Healthy_life_expectancy = "Healthy life expectancy",
    Freedom = "Freedom to make life choices",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Perceptions of corruption"
  ) %>%
  mutate(
    Perceptions_of_corruption = na_if(Perceptions_of_corruption, "N/A"),
    Perceptions_of_corruption = as.numeric(Perceptions_of_corruption)
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

data_2019 <- data_2019 %>%
  rename(
    Country = "Country or region",
    Score = "Score",
    GDP_per_capita = "GDP per capita",
    Social_support = "Social support",
    Healthy_life_expectancy = "Healthy life expectancy",
    Freedom = "Freedom to make life choices",
    Generosity = "Generosity",
    Perceptions_of_corruption = "Perceptions of corruption"
  ) %>%
  select(Country, Score, GDP_per_capita, Social_support, Healthy_life_expectancy, Freedom, Ge

# Add Year column
```

```r
data_2017$Year <- 2017
data_2018$Year <- 2018
data_2019$Year <- 2019
combined_data <- bind_rows(data_2017, data_2018, data_2019)

# Fixing country name
combined_data <- combined_data %>%
  mutate(Country = recode(Country,
                          "Trinidad & Tobago" = "Trinidad and Tobago",
                          "Hong Kong S.A.R., China" = "Hong Kong",
                          "Taiwan Province of China" = "Taiwan",
                          "Macedonia" = "North Macedonia"
  ))

# Remove N/A
combined_data <- combined_data %>%
  drop_na()

# Check Data
str(combined_data)
summary(combined_data)

# Mapping regional data
region_data <- tibble(
  Country = c("Norway", "Denmark", "Iceland", "Switzerland", "Finland", "Netherlands",
              "Canada", "New Zealand", "Sweden", "Australia", "India", "China",
              "Japan", "South Korea", "Brazil", "Argentina", "South Africa"),
  Region = c("Western Europe", "Western Europe", "Western Europe", "Western Europe",
             "Western Europe", "Western Europe", "North America", "Oceania",
             "Western Europe", "Oceania", "South Asia", "East Asia", "East Asia",
             "East Asia", "Latin America", "Latin America", "Sub-Saharan Africa")
)

# Merging regional data
combined_data <- combined_data %>%
  left_join(region_data, by = "Country") %>%
  mutate(Region = replace_na(Region, "Other"))

# Summarize stat
summary_stats <- combined_data %>%
  summarise(across(c(Score, GDP_per_capita, Social_support, Healthy_life_expectancy,
                     Freedom, Generosity, Perceptions_of_corruption),
                   list(mean = ~mean(., na.rm = TRUE),
                        sd = ~sd(., na.rm = TRUE),
                        min = ~min(., na.rm = TRUE),
                        max = ~max(., na.rm = TRUE))))

# Correlation
cor_matrix <- cor(combined_data %>% select(-Country, -Year, -Region))

# Regression model
model <- lm(Score ~ GDP_per_capita + Social_support + Healthy_life_expectancy +
            Freedom + Generosity + Perceptions_of_corruption, data = combined_data)

# Summarize by region
region_summary <- combined_data %>%
  group_by(Region) %>%
  summarise(
    Count = n(),
    Avg_Score = mean(Score, na.rm = TRUE),
    SD_Score = sd(Score, na.rm = TRUE)
  )

# Happiness score table by region
region_summary %>%
  kable(col.names = c("Region", "Count", "Average Score", "SD Score"),
        caption = "Average Happiness Score by Region") %>%
  kable_classic()

# Correlation matrix
corrplot(cor_matrix, method = "color", type = "upper", addCoef.col = "black")

# Regression analysis
stargazer(model, type = "text")

# Score vs gdp
ggplot(combined_data, aes(x = GDP_per_capita, y = Score, color = factor(Year))) +
  geom_point() +
  labs(title = "Happiness Score vs. GDP per Capita", x = "GDP per Capita", y = "Happiness Sc
       caption = "Scatter plot of happiness score vs. GDP per capita by year") +
  theme_minimal()

# Score vs region
ggplot(combined_data, aes(x = Region, y = Score)) +
  geom_boxplot(fill = "#A569FF") +
  labs(title = "Happiness Score by Region", x = "Region", y = "Happiness Score",
       caption = "Boxplot of happiness scores across regions") +
  theme_bw() +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))

# Year
ggplot(combined_data, aes(x = factor(Year), y = Score, group = Year)) +
  geom_boxplot(fill = "#4682B4") +
  labs(title = "Happiness Score by Year", x = "Year", y = "Happiness Score",
       caption = "Boxplot of happiness scores from 2017 to 2019") +
  theme_bw()

# Top 10 heppy country
top_10 <- combined_data %>% filter(Year == 2019) %>% slice_max(Score, n = 10)
ggplot(top_10, aes(x = reorder(Country, Score), y = Score)) +
  geom_bar(stat = "identity", fill = "#65BA65") +
  coord_flip() +
  labs(title = "Top 10 Happiest Countries in 2019", x = "Country", y = "Happiness Score",
       caption = "Bar plot of top 10 countries by happiness score in 2019") +
  theme_bw()
```