DAT 301: Project 1: World Happiness Report

Anand Chamoli

2025-09-26

1. Introduction

In this project, I analyze the 2019 World Happiness Report dataset, which is published by the United Nations Sustainable Development Solutions Network. The dataset measures global happiness using survey data, where people self-report their life satisfaction on a scale from 0 (worst possible life) to 10 (best possible life). It also includes factors that may explain happiness, such as GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption.

Research Questions

- Which factors are most strongly associated with happiness?
- How do happiness levels differ across countries and regions?
- Can we build a simple regression model to predict happiness scores?

2. Data Loading

We loaded the dataset from Kaggle (World Happiness Report 2019). The dataset contains country-level observations. The column names are long and not user-friendly, so we will rename them to shorter, more meaningful names.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(dplyr)

setwd("/Users/anandchamoli/RegressionExplorer")

happiness = read.csv("2019.csv")
str(happiness)
```

```
## 'data.frame':    156 obs. of  9 variables:
##  $ Overall.rank               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country.or.region          : chr  "Finland" "Denmark" "Norway" "Iceland" ...
##  $ Score                      : num  7.77 7.6 7.55 7.49 7.49 ...
##  $ GDP.per.capita             : num  1.34 1.38 1.49 1.38 1.4 ...
##  $ Social.support             : num  1.59 1.57 1.58 1.62 1.52 ...
##  $ Healthy.life.expectancy    : num  0.986 0.996 1.028 1.026 0.999 ...
##  $ Freedom.to.make.life.choices: num  0.596 0.592 0.603 0.591 0.557 0.572 0.574 0.585 0.584 0.532 ..
##  $ Generosity                 : num  0.153 0.252 0.271 0.354 0.322 0.263 0.267 0.33 0.285 0.244 ...
##  $ Perceptions.of.corruption  : num  0.393 0.41 0.341 0.118 0.298 0.343 0.373 0.38 0.308 0.226 ...
```

```r
head(happiness)
```

```
##   Overall.rank Country.or.region Score GDP.per.capita Social.support
## 1            1           Finland 7.769          1.340          1.587
## 2            2           Denmark 7.600          1.383          1.573
## 3            3            Norway 7.554          1.488          1.582
## 4            4           Iceland 7.494          1.380          1.624
## 5            5       Netherlands 7.488          1.396          1.522
## 6            6       Switzerland 7.480          1.452          1.526
##   Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 1                   0.986                        0.596      0.153
## 2                   0.996                        0.592      0.252
## 3                   1.028                        0.603      0.271
## 4                   1.026                        0.591      0.354
## 5                   0.999                        0.557      0.322
## 6                   1.052                        0.572      0.263
##   Perceptions.of.corruption
## 1                     0.393
## 2                     0.410
## 3                     0.341
## 4                     0.118
## 5                     0.298
## 6                     0.343
```

3. Data Wrangling

Here, we renamed variables for clarity and selected only the relevant columns. We then checked for missing values and used na.omit() to ensure a clean analysis.

```r
data = happiness %>%
  rename(Country = Country.or.region, GDP = GDP.per.capita, Support = Social.support, Life = Healthy.li
  select(Country, Score, GDP, Support, Life, Freedom, Generosity, Corruption)

# checking missing values
colSums(is.na(data))
```

```
##    Country      Score        GDP    Support       Life    Freedom Generosity
##          0          0          0          0          0          0          0
## Corruption
##          0
```

```
# drop rows with missing values
data = na.omit(data)
```

4. Exploratory Data Analysis (EDA)

4.1 Stats Summary: This gives us an overview of the distribution of happiness scores, GDP per capita, life expectancy, and social support.

```
summary(data$Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.853   4.545   5.380   5.407   6.184   7.769
```

```
summary(data[, c("GDP", "Life", "Support")])
```

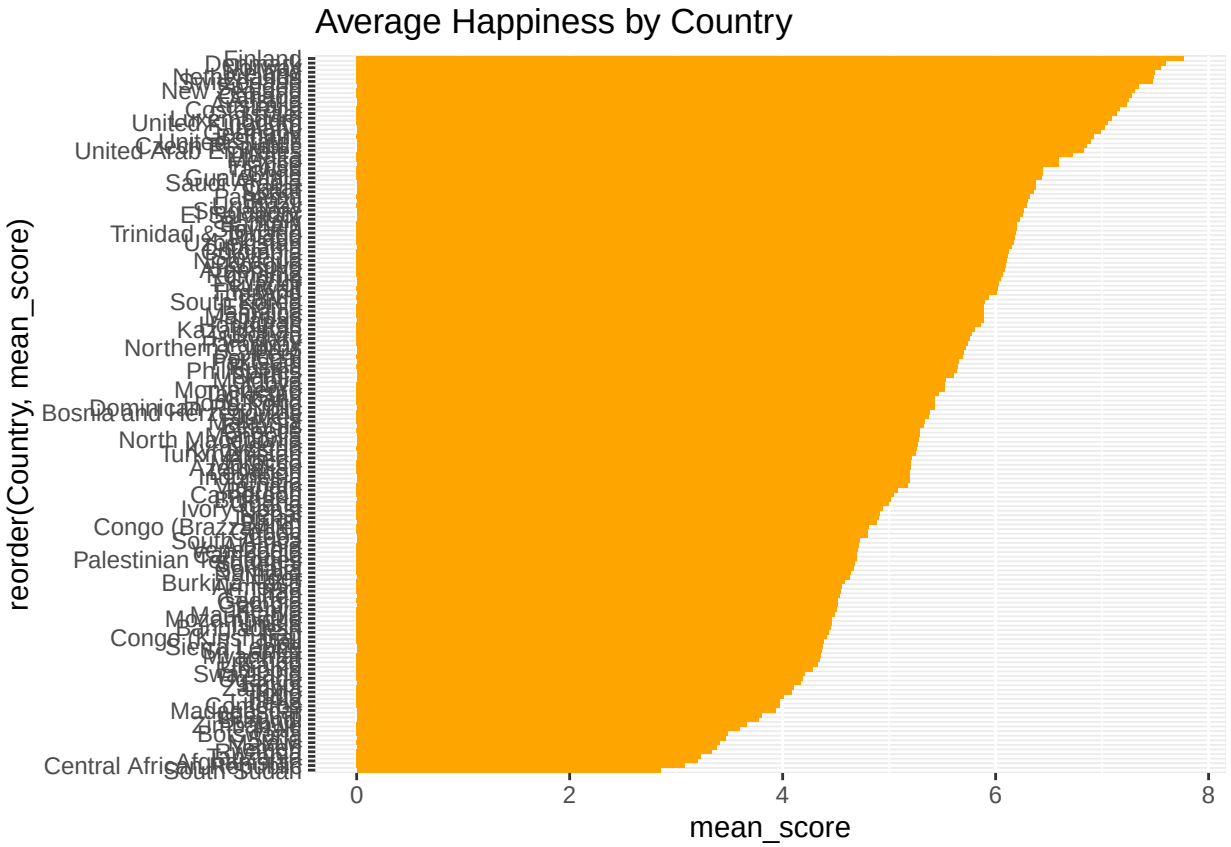
```
##       GDP              Life           Support
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.6028   1st Qu.:0.5477   1st Qu.:1.056
##  Median :0.9600   Median :0.7890   Median :1.272
##  Mean   :0.9051   Mean   :0.7252   Mean   :1.209
##  3rd Qu.:1.2325   3rd Qu.:0.8818   3rd Qu.:1.452
##  Max.   :1.6840   Max.   :1.1410   Max.   :1.624
```

4.2 Distribution of Happiness Scores: Most countries have happiness scores between 4 and 6. Very few countries score above 8 or below 3, making the distribution slightly skewed to the lower side.

```
ggplot(data, aes(x = Score)) + geom_histogram(bins = 20, fill = "turquoise", color = "white") + labs(ti
```

4.3 Average Happiness by Countries: Since there are many countries, this chart is crowded. To simplify, we also highlight the Top 10 and Bottom 10 countries.

```
data %>%
  group_by(Country) %>%
  summarise(mean_score = mean(Score, na.rm = T)) %>%
  ggplot(aes(x = reorder(Country, mean_score), y = mean_score)) + geom_col(fill = "orange") + coord_fli
```



4.4 Top 10 and Bottom 10 Countries by Happiness:

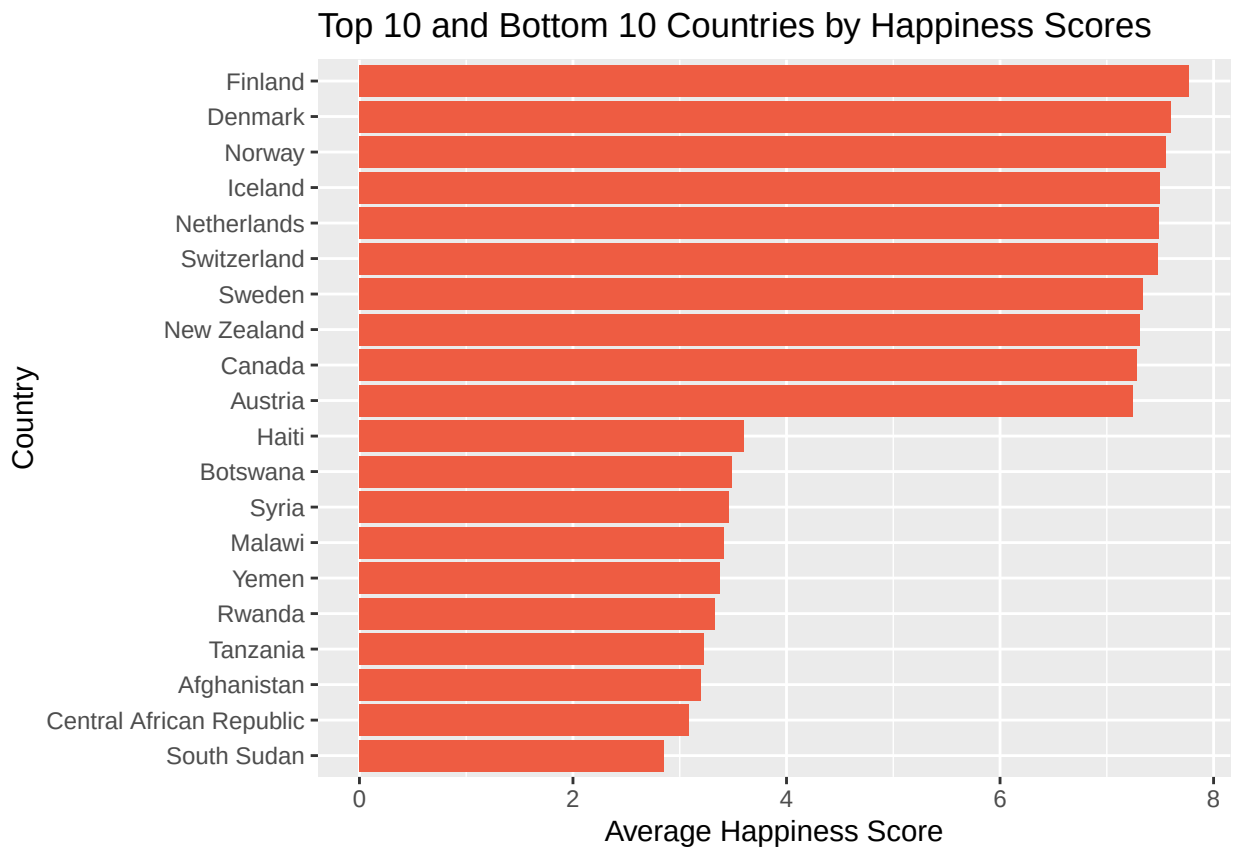
- The Top 10 countries include Finland, Denmark, and Norway, all with strong economies and high social support.
- The Bottom 10 countries include South Sudan and Afghanistan, where war, poverty, and instability lower happiness levels.

```
country_stats = data %>%
  group_by(Country) %>%
  summarise(mean_score = mean(Score, na.rm = T))

top_10 = country_stats %>% arrange(desc(mean_score)) %>% head(10)
bottom_10 = country_stats %>% arrange(mean_score) %>% head(10)

top_bottom = bind_rows(top_10,bottom_10)

ggplot(top_bottom, aes(x=reorder(Country, mean_score), y = mean_score)) + geom_col(fill = "tomato2") + c
```



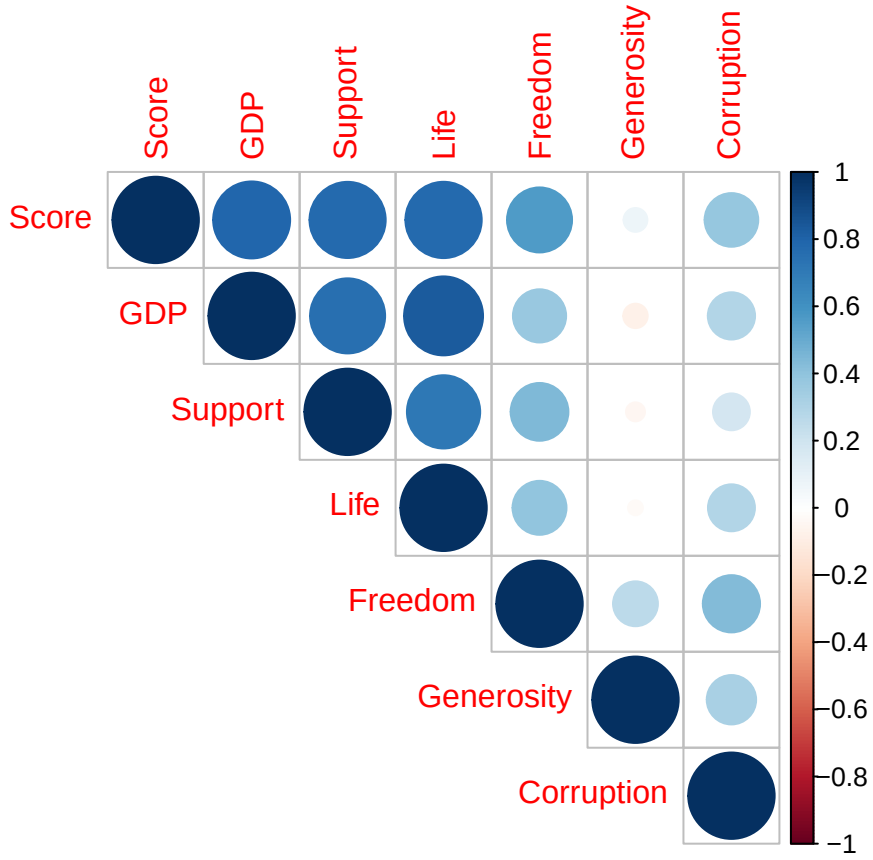
4.5 Correlation Analysis:

- Happiness Score shows strong positive correlations with GDP, Life Expectancy, and Social Support.
- Freedom also has a significant relationship.
- Generosity and Corruption show weaker connections.

```
corr_data = data %>%
  select(Score, GDP, Support, Life, Freedom, Generosity, Corruption)

corr_matrix = cor(corr_data, use = "complete.obs")

corrplot(corr_matrix, method = "circle", type = "upper")
```

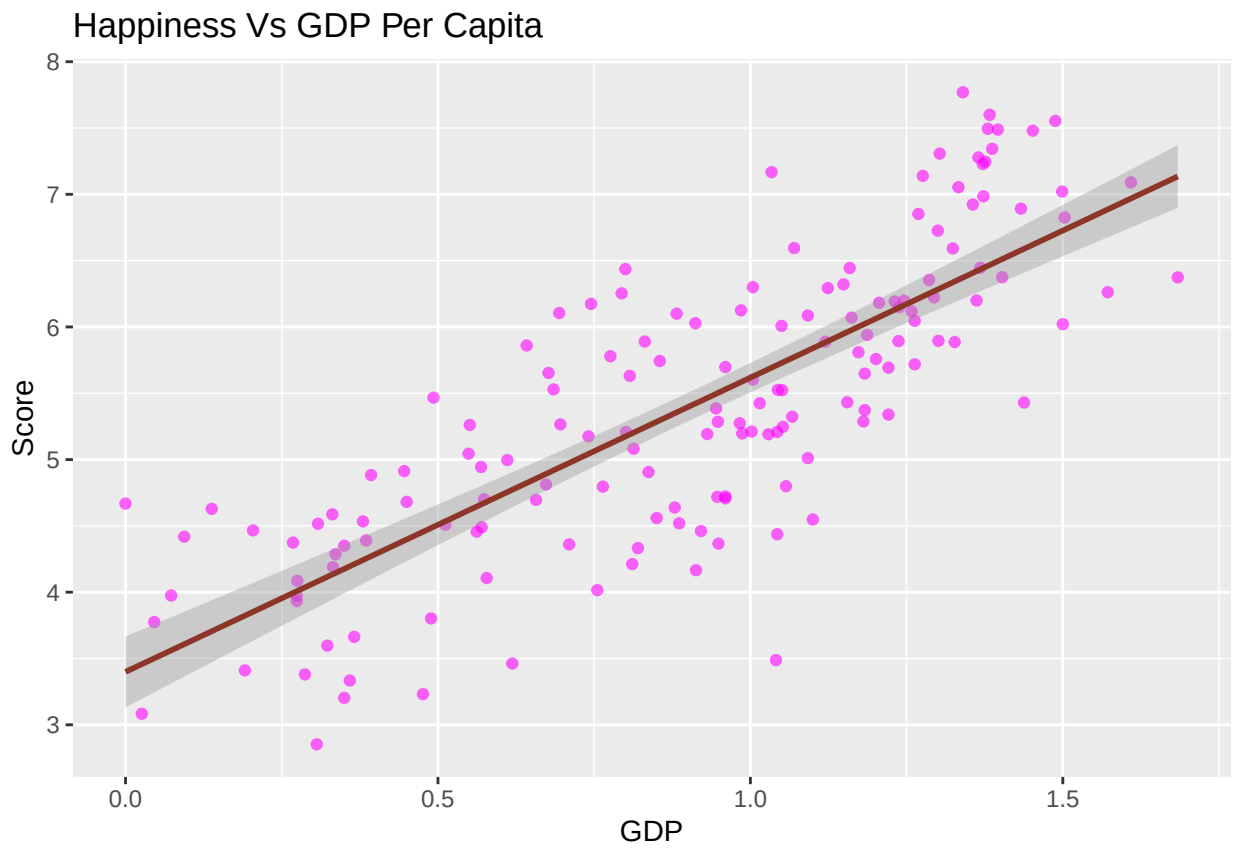


5. Data Visualization

- Countries with higher GDP per capita tend to report higher happiness.
- Countries with longer healthy life expectancy also report higher happiness.

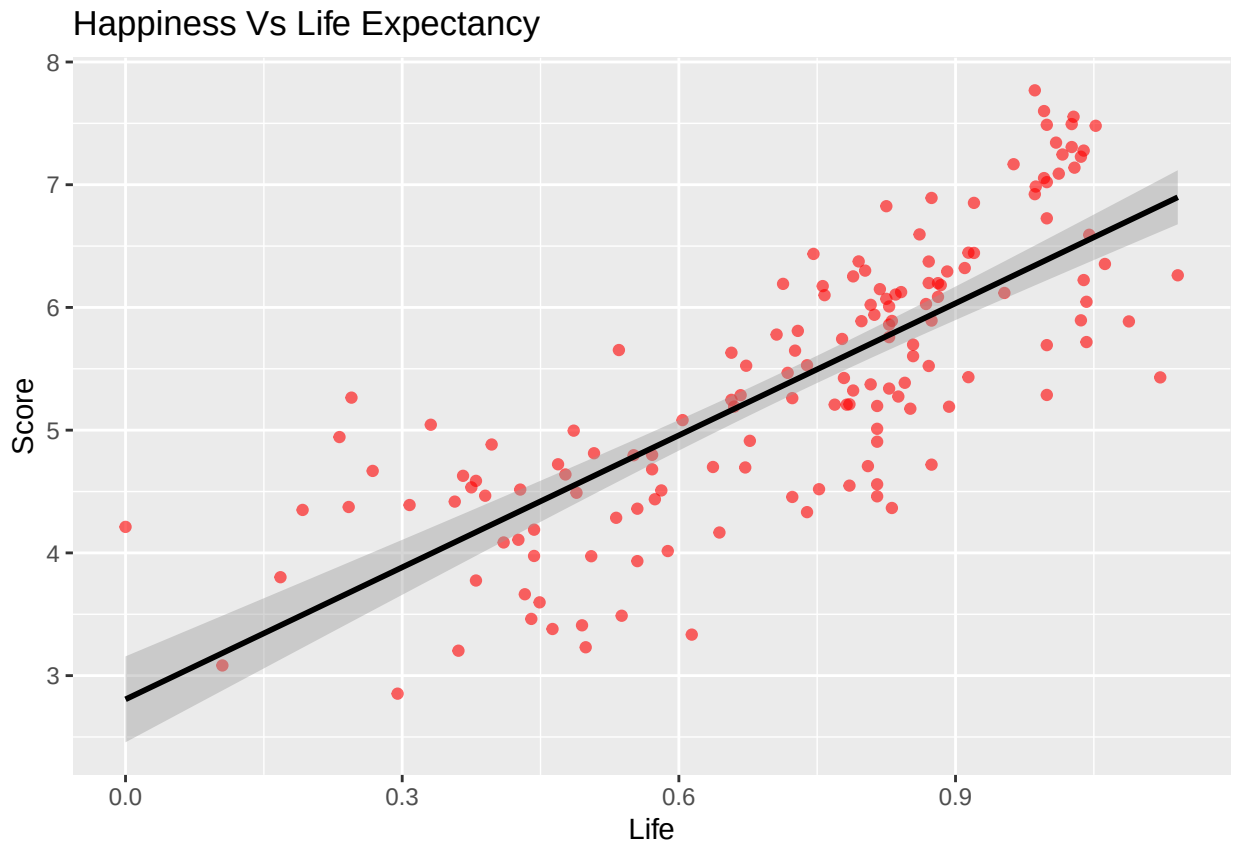
```
ggplot(data, aes(x = GDP, y = Score,)) + geom_point(color = "magenta", alpha = 0.6) + geom_smooth(metho
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data, aes(x = Life, y = Score,)) + geom_point(color = "red", alpha = 0.6) + geom_smooth(method =
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



6. Regression Analysis

- All four predictors (GDP, Life Expectancy, Support, and Freedom) are statistically significant ($p < 0.05$).
- Freedom has the largest coefficient, showing the strongest influence on happiness.
- The model explains about 77% of the variation in Happiness Scores ($R^2 = 0.77$).
- Prediction errors are small, with residuals centered around zero, suggesting a good model fit.

```
regression_model = lm(Score ~ GDP + Life + Support + Freedom, data = data)
summary(regression_model)
```

```
##
## Call:
## lm(formula = Score ~ GDP + Life + Support + Freedom, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86584 -0.34594  0.03403  0.43676  1.13076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8921     0.1994   9.491  < 2e-16 ***
## GDP           0.8105     0.2165   3.745 0.000256 ***
```

```
## Life          1.1414     0.3373   3.384 0.000910 ***
## Support       1.0166     0.2347   4.331 2.70e-05 ***
## Freedom       1.8458     0.3404   5.423 2.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5398 on 151 degrees of freedom
## Multiple R-squared:  0.7709, Adjusted R-squared:  0.7649
## F-statistic:   127 on 4 and 151 DF,  p-value: < 2.2e-16
```

7. Conclusion: From this analysis, we find that GDP, Life Expectancy, Social Support, and Freedom to make life choices are the strongest drivers of happiness across countries.

- Wealthier countries with longer life expectancy and strong social systems rank higher in happiness.
- Freedom plays the largest role, suggesting that economic prosperity alone does not guarantee happiness — personal and social freedoms are equally important.
- The top-ranked countries (Finland, Denmark, Norway) combine high GDP, life expectancy, and strong governance.
- Countries with conflict and poverty remain at the bottom, highlighting how instability impacts well-being.

8. Limitations

- Only 2019 data is used; results may change across years.
- Happiness is self-reported and may vary due to cultural biases.
- Variables like corruption are hard to quantify.

9. Future Work

- Extend analysis to 2015–2019 to see trends.
- Include cultural, political, or environmental variables for richer insights.

10. References

World Happiness Report 2019 (Kaggle)