Final Report

Liz Kim

1/21/2025

## Table of Contents

* [1. Introduction](#introduction)
* [2. Research Question](#research-question)
* [3. Data Scraping & Tidying the Data](#Data-Scaping-and-Tidying-the-Data)
* [4. Descriptive Statistics and Plots](#descriptive-statistics-and-plots)
* [5. Data Interpretation](#data-interpretation)

# 1. Introduction

This project investigates the factors associated with national happiness, utilizing happiness scores from the World Population Review and key factors from Wikipedia tables. Key variables include GDP per capita, environmental quality, abd literacy rates offering insights into the economic, environmental, and educational factors that influence happiness levels across countries.

The reason why I’m working on this project is because it is important to understand how these factors contribute to national happiness. Happiness is a critical measure of well-being, and identifying its determinants is vital.

# 2. Research Question

Main Question: How do economic,education, and environmental variables influence national happiness levels as reported by the World Population Review?

RQ1: How does GDP per capita associate with a country’s happiness score? RQ2: How does environmental quality associate with a country’s happiness score? RQ3: How does a country’s literacy rate associate with its happiness score?

# 3. Data Scraping & Tidying the Data

Below is the code I used to scrap the data.

**Happiness Score**

# Read the CSV file  
happiest\_countries <- read\_csv("happiest-countries-in-the-world-2024.csv")

## Rows: 145 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): country  
## dbl (7): HappiestCountriesWorldHappinessReportRankings2024, HappiestCountrie...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Preview the data  
head(happiest\_countries)

## # A tibble: 6 × 8  
## country HappiestCountriesWorl…¹ HappiestCountriesWor…² LatestScoreChange  
## <chr> <dbl> <dbl> <dbl>  
## 1 India 126 4.05 0.0140  
## 2 China 60 5.97 0.152   
## 3 United States 23 6.73 -0.164   
## 4 Indonesia 80 5.57 0.293   
## 5 Pakistan 108 4.66 0.105   
## 6 Nigeria 102 4.88 -0.101   
## # ℹ abbreviated names: ¹​HappiestCountriesWorldHappinessReportRankings2024,  
## # ²​HappiestCountriesWorldHappinessReportScore2024  
## # ℹ 4 more variables: HappiestCountriesWorldHappiessReportRankings2023 <dbl>,  
## # HappiestCountriesWorldHappiessReportScore2023 <dbl>,  
## # HappiestCountriesWorldHappiessReportRankings2022 <dbl>,  
## # HappiestCountriesWorldHappiessReportScore2022 <dbl>

I requested this data table from WorldPopulationReview.com. I received an excel file through my email.

**GDP per Capita**

# Save the URL and scrape the webpage  
url <- "https://en.wikipedia.org/wiki/List\_of\_countries\_by\_GDP\_(nominal)\_per\_capita"  
gdp\_page <- read\_html(x = url)  
  
# Extract all tables with the class "wikitable"  
tables <- html\_elements(gdp\_page, css = "table.wikitable")  
  
# Convert the first table to a data frame  
gdp\_per\_capita\_table <- html\_table(tables[[1]], fill = TRUE)  
  
# Save the table as a CSV file  
write\_csv(gdp\_per\_capita\_table, "gdp\_per\_capita.csv")  
  
# Preview the table  
head(gdp\_per\_capita\_table)

## # A tibble: 6 × 7  
## `Country/Territory` `IMF[4][5]` `IMF[4][5]` `World Bank[6]` `World Bank[6]`  
## <chr> <chr> <chr> <chr> <chr>   
## 1 Country/Territory Estimate Year Estimate Year   
## 2 Monaco — — 240,862 2022   
## 3 Liechtenstein — — 187,267 2022   
## 4 Luxembourg 135,321 2024 128,259 2023   
## 5 Bermuda — — 123,091 2022   
## 6 Switzerland 106,098 2024 99,995 2023   
## # ℹ 2 more variables: `United Nations[7]` <chr>, `United Nations[7]` <chr>

**Education (Literacy Rate)**

# Save the URL and scrape the webpage  
url <- "https://en.wikipedia.org/wiki/List\_of\_countries\_by\_literacy\_rate"  
literacy\_page <- read\_html(x = url)  
  
# Extract all tables with the class "wikitable"  
tables <- html\_elements(literacy\_page, css = "table.wikitable")  
  
# Convert all table nodes into a list of data frames  
all\_tables <- html\_table(tables, fill = TRUE)  
  
# Inspect tables to find the one I want to extract  
for (i in seq\_along(all\_tables)) {  
 print(paste("Table", i))  
 print(head(all\_tables[[i]]))  
}

## [1] "Table 1"  
## # A tibble: 6 × 9  
## Country `Youth(15 to 24)` `Youth(15 to 24)` `Adult(25+)` `Adult(25+)`  
## <chr> <chr> <chr> <chr> <chr>   
## 1 Country "Rate" "Year" "Rate" "Year"   
## 2 Afghanistan \* "65.0" "2020[4]" "31.7" "2011"   
## 3 Albania \* "99.2" "2012" "97.2" "2012"   
## 4 Algeria \* "93.8" "2008" "75.1" "2008"   
## 5 American Samoa \* "97.7" "1980" "97.3" "1980"   
## 6 Andorra \* "" "" "" ""   
## # ℹ 4 more variables: `Elderly(65+)` <chr>, `Elderly(65+)` <chr>,  
## # `Youth GenderParity Index` <chr>, `Youth GenderParity Index` <chr>  
## [1] "Table 2"  
## # A tibble: 6 × 6  
## Country Literacy rate[12][13…¹ Literacy rate[12][13…² Literacy rate[12][13…³  
## <chr> <chr> <chr> <chr>   
## 1 Country Total Male Female   
## 2 Afghanis… 37.3% 52.1% 22.6%   
## 3 Albania \* 98.1% 98.5% 97.8%   
## 4 Algeria \* 81.4% 87.4% 75.3%   
## 5 Andorra \* 100.0% 100.0% 100.0%   
## 6 Angola \* 71.1% 82.0% 60.7%   
## # ℹ abbreviated names: ¹​`Literacy rate[12][13][text–source integrity?]`,  
## # ²​`Literacy rate[12][13][text–source integrity?]`,  
## # ³​`Literacy rate[12][13][text–source integrity?]`  
## # ℹ 2 more variables: `Literacy rate[12][13][text–source integrity?]` <chr>,  
## # `Literacy rate[12][13][text–source integrity?]` <chr>  
## [1] "Table 3"  
## # A tibble: 6 × 7  
## Territory `Literacy rate (all)` `Male literacy` `Female literacy`  
## <chr> <chr> <chr> <chr>   
## 1 Aruba 97.5% 97.5% 97.5%   
## 2 Cayman Islands 98.9% 98.7% 99.0%   
## 3 Guadeloupe 96.5% 96.4% 96.6%   
## 4 Guam 99.8% 99.8% 99.8%   
## 5 Kosovo 91.9% 96.6% 87.5%   
## 6 Macau 96.2% 98.0% 94.6%   
## # ℹ 3 more variables: `Gender difference[a]` <chr>, Year <chr>, Note <chr>

# Extract the second table   
literacy\_rate\_table <- all\_tables[[2]]  
write\_csv(literacy\_rate\_table, "literacy\_rate.csv")  
  
# Preview the table  
head(literacy\_rate\_table)

## # A tibble: 6 × 6  
## Country Literacy rate[12][13…¹ Literacy rate[12][13…² Literacy rate[12][13…³  
## <chr> <chr> <chr> <chr>   
## 1 Country Total Male Female   
## 2 Afghanis… 37.3% 52.1% 22.6%   
## 3 Albania \* 98.1% 98.5% 97.8%   
## 4 Algeria \* 81.4% 87.4% 75.3%   
## 5 Andorra \* 100.0% 100.0% 100.0%   
## 6 Angola \* 71.1% 82.0% 60.7%   
## # ℹ abbreviated names: ¹​`Literacy rate[12][13][text–source integrity?]`,  
## # ²​`Literacy rate[12][13][text–source integrity?]`,  
## # ³​`Literacy rate[12][13][text–source integrity?]`  
## # ℹ 2 more variables: `Literacy rate[12][13][text–source integrity?]` <chr>,  
## # `Literacy rate[12][13][text–source integrity?]` <chr>

**Environmental Quality**

# Save the URL and scrape the webpage  
url <- "https://en.wikipedia.org/wiki/Environmental\_Performance\_Index"  
environment\_page <- read\_html(x = url)  
  
# Extract all tables with the class "wikitable"  
tables <- html\_elements(environment\_page, css = "table.wikitable")  
  
# Convert all table nodes into a list of data frames  
all\_tables <- html\_table(tables, fill = TRUE)  
  
# Inspect tables to find the one I want to extract  
for (i in seq\_along(all\_tables)) {  
 print(paste("Table", i))  
 print(head(all\_tables[[i]]))  
}

## [1] "Table 1"  
## # A tibble: 6 × 6  
## `Policy Objective` `Wt. (%)` `Issue Category` `Wt. (%)` Indicator `Wt. (%)`  
## <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 Ecosystem Vitality 45% Biodiversity & Hab… 25 Marine K… 12.0   
## 2 Ecosystem Vitality 45% Biodiversity & Hab… 25 Marine H… 12.0   
## 3 Ecosystem Vitality 45% Biodiversity & Hab… 25 Marine P… 2.0   
## 4 Ecosystem Vitality 45% Biodiversity & Hab… 25 Protecte… 12.0   
## 5 Ecosystem Vitality 45% Biodiversity & Hab… 25 Species … 16.0   
## 6 Ecosystem Vitality 45% Biodiversity & Hab… 25 Terrestr… 10.0   
## [1] "Table 2"  
## # A tibble: 6 × 5  
## Country Region Value Trend `Rank 2024`  
## <chr> <chr> <dbl> <chr> <int>  
## 1 Afghanistan South Asia 31 12.8 144  
## 2 Angola Sub-Saharan Africa 40.1 8.2 106  
## 3 Albania Europe & Central Asia 52.2 6.1 47  
## 4 United Arab Emirates Middle East & North Africa 51.6 9.1 48  
## 5 Argentina Latin America & Caribbean 47 1.1 70  
## 6 Armenia Europe & Central Asia 44.9 2.0 80

# Extract the second table (Environmental Performance Index)  
environmental\_quality\_table <- all\_tables[[2]]  
write\_csv(environmental\_quality\_table, "environmental\_quality.csv")  
  
# Preview the table  
head(environmental\_quality\_table)

## # A tibble: 6 × 5  
## Country Region Value Trend `Rank 2024`  
## <chr> <chr> <dbl> <chr> <int>  
## 1 Afghanistan South Asia 31 12.8 144  
## 2 Angola Sub-Saharan Africa 40.1 8.2 106  
## 3 Albania Europe & Central Asia 52.2 6.1 47  
## 4 United Arab Emirates Middle East & North Africa 51.6 9.1 48  
## 5 Argentina Latin America & Caribbean 47 1.1 70  
## 6 Armenia Europe & Central Asia 44.9 2.0 80

Here, I included the code to tidy the tables. The datasets comprises four tables, each offering insights into countries’ happiness scores, GDP, literacy rates, and environmental performance. Tidy data follows a standardized structure where each variable forms a column, each observation forms a row, and each cell contains a single measurement.

The advantages of tidy data are numerous. It simplifies data analysis by standardizing the structure, enabling easier manipulation, visualization, and modeling. Additionally, tidy data ensures compatibility with a wide range of tools and software, making workflows more streamlined. However, achieving a tidy format often requires significant transformation of raw data, which can be time-consuming. In some cases, restructuring data into a tidy format increases the number of rows, leading to larger file sizes and potential performance challenges with very large datasets.

Despite these challenges, the benefits of tidy data outweigh the disadvantages in my dataset. I dropped NA variables, renamed variables, replaced missing or incorrectly formated values through functions, and etc. This process allowed me to easily read, interpret, and manipulate the data in order to create various appropriate plots for each dataset. Additionally, it made merging data more efficient and easy.

## Happiness Score Table

# Read the previously saved CSV file  
happiest\_countries <- read\_csv("happiest-countries-in-the-world-2024.csv")

## Rows: 145 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): country  
## dbl (7): HappiestCountriesWorldHappinessReportRankings2024, HappiestCountrie...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Clean and rename the happiest\_countries dataset  
happiest\_countries\_cleaned <- happiest\_countries %>%  
 select(country, HappiestCountriesWorldHappinessReportScore2024) %>%   
 filter(!is.na(HappiestCountriesWorldHappinessReportScore2024)) %>%   
 rename(HappinessScore = HappiestCountriesWorldHappinessReportScore2024)   
  
# Preview and save the cleaned table  
head(happiest\_countries\_cleaned)

## # A tibble: 6 × 2  
## country HappinessScore  
## <chr> <dbl>  
## 1 India 4.05  
## 2 China 5.97  
## 3 United States 6.73  
## 4 Indonesia 5.57  
## 5 Pakistan 4.66  
## 6 Nigeria 4.88

write\_csv(happiest\_countries\_cleaned, "happiest\_countries\_cleaned.csv")

## GDP Table

# Read the previously saved CSV file  
gdp\_per\_capita\_table <- read\_csv("gdp\_per\_capita.csv")

## New names:  
## Rows: 224 Columns: 7  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (7): Country/Territory, IMF[4][5]...2, IMF[4][5]...3, World Bank[6]...4,...  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `IMF[4][5]` -> `IMF[4][5]...2`  
## • `IMF[4][5]` -> `IMF[4][5]...3`  
## • `World Bank[6]` -> `World Bank[6]...4`  
## • `World Bank[6]` -> `World Bank[6]...5`  
## • `United Nations[7]` -> `United Nations[7]...6`  
## • `United Nations[7]` -> `United Nations[7]...7`

# Rename columns  
colnames(gdp\_per\_capita\_table) <- c("country", "IMF\_estimate", "IMF\_year",   
 "World\_Bank\_estimate", "World\_Bank\_year",   
 "UN\_estimate", "UN\_year")  
  
# Tidy the table  
fill\_missing\_values <- function(data) {  
 data %>%  
 select(country, IMF\_estimate) %>%   
 mutate(  
 IMF\_estimate = ifelse(IMF\_estimate == "-", NA, IMF\_estimate),   
 IMF\_estimate = gsub(",", "", IMF\_estimate),   
 IMF\_estimate = as.numeric(IMF\_estimate)   
 ) %>%  
 fill(IMF\_estimate, .direction = "down") %>%   
 filter(!is.na(IMF\_estimate)) %>%   
 rename(estimate = IMF\_estimate)   
}  
  
# Apply the cleaning function  
gdp\_per\_capita\_table\_cleaned <- fill\_missing\_values(gdp\_per\_capita\_table)

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `IMF\_estimate = as.numeric(IMF\_estimate)`.  
## Caused by warning:  
## ! NAs introduced by coercion

# Preview and save the cleaned table  
head(gdp\_per\_capita\_table\_cleaned)

## # A tibble: 6 × 2  
## country estimate  
## <chr> <dbl>  
## 1 Luxembourg 135321  
## 2 Bermuda 135321  
## 3 Switzerland 106098  
## 4 Ireland 103500  
## 5 Cayman Islands 103500  
## 6 Isle of Man 103500

write\_csv(gdp\_per\_capita\_table\_cleaned, "gdp\_per\_capita\_table\_cleaned.csv")

## Education (Literacy Rate Table)

# Read the previously saved CSV file  
literacy\_rate\_table <- read\_csv("literacy\_rate.csv")

## New names:  
## Rows: 197 Columns: 6  
## ── Column specification  
## ──────────────────────────────────────────────────────── Delimiter: "," chr  
## (6): Country, Literacy rate[12][13][text–source integrity?]...2, Literac...  
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
## Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
## • `Literacy rate[12][13][text–source integrity?]` -> `Literacy  
## rate[12][13][text–source integrity?]...2`  
## • `Literacy rate[12][13][text–source integrity?]` -> `Literacy  
## rate[12][13][text–source integrity?]...3`  
## • `Literacy rate[12][13][text–source integrity?]` -> `Literacy  
## rate[12][13][text–source integrity?]...4`  
## • `Literacy rate[12][13][text–source integrity?]` -> `Literacy  
## rate[12][13][text–source integrity?]...5`  
## • `Literacy rate[12][13][text–source integrity?]` -> `Literacy  
## rate[12][13][text–source integrity?]...6`

head(literacy\_rate\_table)

## # A tibble: 6 × 6  
## Country Literacy rate[12][13…¹ Literacy rate[12][13…² Literacy rate[12][13…³  
## <chr> <chr> <chr> <chr>   
## 1 Country Total Male Female   
## 2 Afghanis… 37.3% 52.1% 22.6%   
## 3 Albania \* 98.1% 98.5% 97.8%   
## 4 Algeria \* 81.4% 87.4% 75.3%   
## 5 Andorra \* 100.0% 100.0% 100.0%   
## 6 Angola \* 71.1% 82.0% 60.7%   
## # ℹ abbreviated names: ¹​`Literacy rate[12][13][text–source integrity?]...2`,  
## # ²​`Literacy rate[12][13][text–source integrity?]...3`,  
## # ³​`Literacy rate[12][13][text–source integrity?]...4`  
## # ℹ 2 more variables:  
## # `Literacy rate[12][13][text–source integrity?]...5` <chr>,  
## # `Literacy rate[12][13][text–source integrity?]...6` <chr>

# Rename columns  
colnames(literacy\_rate\_table) <- c("country", "total", "male", "female", "gap", "year")  
  
# Tidy the data  
literacy\_rate\_table\_cleaned <- literacy\_rate\_table %>%  
 select(country, total, year) %>%  
 mutate(  
 country = gsub("\\\*", "", country),   
 total = as.numeric(gsub("%", "", total)),   
 year = as.numeric(year)   
 )

## Warning: There were 2 warnings in `mutate()`.  
## The first warning was:  
## ℹ In argument: `total = as.numeric(gsub("%", "", total))`.  
## Caused by warning:  
## ! NAs introduced by coercion  
## ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 1 remaining warning.

# Filter out rows where country == "Country" or "World"  
literacy\_rate\_table\_cleaned <- literacy\_rate\_table\_cleaned %>%  
 filter(country != "Country",   
 country != "World")   
  
# Preview and save the cleaned table  
head(literacy\_rate\_table\_cleaned)

## # A tibble: 6 × 3  
## country total year  
## <chr> <dbl> <dbl>  
## 1 Afghanistan  37.3 2021  
## 2 Albania  98.1 2018  
## 3 Algeria  81.4 2018  
## 4 Andorra  100 2016  
## 5 Angola  71.1 2015  
## 6 Antigua and Barbuda  99 2015

write\_csv(literacy\_rate\_table\_cleaned, "literacy\_rate\_table\_cleaned.csv")

## Environmental Quality Table

# Read the previously saved CSV file  
environmental\_quality\_table <- read\_csv("environmental\_quality.csv")

## Rows: 179 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): Country, Region, Trend  
## dbl (2): Value, Rank 2024  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Tidy the data  
environmental\_quality\_table\_cleaned <- environmental\_quality\_table %>%  
 select(country = `Country`, value = `Value`)  
  
# Preview and save the cleaned table  
head(environmental\_quality\_table\_cleaned)

## # A tibble: 6 × 2  
## country value  
## <chr> <dbl>  
## 1 Afghanistan 31   
## 2 Angola 40.1  
## 3 Albania 52.2  
## 4 United Arab Emirates 51.6  
## 5 Argentina 47   
## 6 Armenia 44.9

write\_csv(environmental\_quality\_table\_cleaned, "environmental\_quality\_table\_cleaned.csv")

# 4. Descriptive Statistics and Plots

# Clean country names function  
clean\_country\_names <- function(data, country\_column = "country") {  
 data %>%  
 mutate(!!sym(country\_column) := gsub("[^[:alnum:] ]", "", !!sym(country\_column))) %>%  
 mutate(!!sym(country\_column) := trimws(!!sym(country\_column))) %>%  
 mutate(!!sym(country\_column) := tolower(!!sym(country\_column)))  
}  
  
# Clean the country names in all datasets  
happiest\_countries\_cleaned <- clean\_country\_names(happiest\_countries\_cleaned)  
gdp\_per\_capita\_table\_cleaned <- clean\_country\_names(gdp\_per\_capita\_table\_cleaned)  
literacy\_rate\_table\_cleaned <- clean\_country\_names(literacy\_rate\_table\_cleaned)  
environmental\_quality\_table\_cleaned <- clean\_country\_names(environmental\_quality\_table\_cleaned)  
  
# Rename columns to prepare for merging  
gdp\_data <- gdp\_per\_capita\_table\_cleaned %>%  
 rename(Value = estimate) %>%  
 mutate(Variable = "GDP")  
  
education\_data <- literacy\_rate\_table\_cleaned %>%  
 rename(Value = total) %>%  
 mutate(Variable = "Education")  
  
env\_quality\_data <- environmental\_quality\_table\_cleaned %>%  
 rename(Value = value) %>%  
 mutate(Variable = "EnvironmentalQuality")  
  
happiness\_data <- happiest\_countries\_cleaned %>%  
 rename(Value = HappinessScore) %>%  
 mutate(Variable = "HappinessScore")  
  
# Combine all datasets into a single long format  
long\_data <- bind\_rows(  
 gdp\_data,  
 education\_data,  
 env\_quality\_data,  
 happiness\_data  
)  
  
# Pivot to wide format  
wide\_data <- long\_data %>%  
 pivot\_wider(  
 names\_from = Variable,  
 values\_from = Value  
 )  
  
# Remove year and sexgap columns  
wide\_data\_cleaned <- wide\_data %>%  
 select(-year)   
  
# Display the Merged Data Overview  
wide\_data\_cleaned %>%  
 head(10) %>%   
 kable(  
 col.names = c("Country", "GDP", "Education", "Environmental Quality", "Happiness Score"),  
 caption = "Merged Data Overview",  
 format = "markdown"  
 )

Merged Data Overview

| Country | GDP | Education | Environmental Quality | Happiness Score |
| --- | --- | --- | --- | --- |
| luxembourg | 135321 | 100 | 75.1 | 7.12 |
| bermuda | 135321 | NA | NA | NA |
| switzerland | 106098 | 99 | 67.8 | 7.06 |
| ireland | 103500 | 99 | 65.8 | 6.84 |
| cayman islands | 103500 | NA | NA | NA |
| isle of man | 103500 | NA | NA | NA |
| norway | 90434 | 100 | 69.9 | 7.30 |
| singapore | 89370 | NA | 53.0 | 6.52 |
| united states | 86601 | 86 | 57.2 | 6.73 |
| iceland | 85787 | 99 | 64.3 | 7.53 |

I created the Merged Data Overview table to provide a comprehensive view of all the important variables in a single table. However, due to differences in country names and variations in the countries included in each dataset, many NA values are present. Therefore, for the visualizations, I will analyze two variables at a time separately. Despite the NA values, this table is still useful for quickly skimming through the data.

## RQ1: How does GDP per capita associate with a country’s happiness score?

# Clean the country names in gdp\_per\_capita\_table\_cleaned  
gdp\_per\_capita\_table\_cleaned <- gdp\_per\_capita\_table\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Clean the country names in happiest\_countries\_cleaned  
happiest\_countries\_cleaned <- happiest\_countries\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Merge GDP and HappinessScore data  
merged\_data <- gdp\_per\_capita\_table\_cleaned %>%  
 left\_join(happiest\_countries\_cleaned, by = "country")  
print(merged\_data)

## # A tibble: 221 × 3  
## country estimate HappinessScore  
## <chr> <dbl> <dbl>  
## 1 luxembourg 135321 7.12  
## 2 bermuda 135321 NA   
## 3 switzerland 106098 7.06  
## 4 ireland 103500 6.84  
## 5 cayman islands 103500 NA   
## 6 isle of man 103500 NA   
## 7 norway 90434 7.3   
## 8 singapore 89370 6.52  
## 9 united states 86601 6.73  
## 10 iceland 85787 7.53  
## # ℹ 211 more rows

# Remove missing values  
cleaned\_merged\_data <- merged\_data %>%  
 filter(!is.na(estimate), !is.na(HappinessScore))  
print(cleaned\_merged\_data)

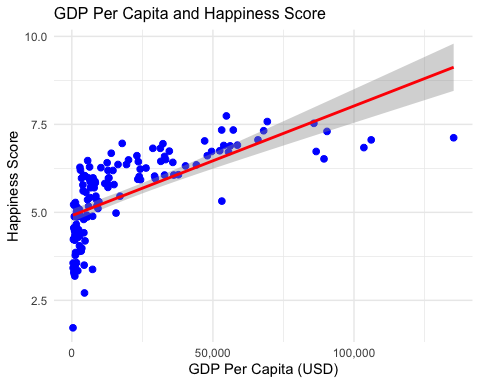
## # A tibble: 140 × 3  
## country estimate HappinessScore  
## <chr> <dbl> <dbl>  
## 1 luxembourg 135321 7.12  
## 2 switzerland 106098 7.06  
## 3 ireland 103500 6.84  
## 4 norway 90434 7.3   
## 5 singapore 89370 6.52  
## 6 united states 86601 6.73  
## 7 iceland 85787 7.53  
## 8 denmark 69273 7.58  
## 9 netherlands 67984 7.32  
## 10 australia 65966 7.06  
## # ℹ 130 more rows

# View summary statistics   
summary(cleaned\_merged\_data[, c("estimate", "HappinessScore")])

## estimate HappinessScore   
## Min. : 411 Min. :1.720   
## 1st Qu.: 2680 1st Qu.:4.635   
## Median : 7507 Median :5.805   
## Mean : 19722 Mean :5.522   
## 3rd Qu.: 29399 3rd Qu.:6.412   
## Max. :135321 Max. :7.740

# Create the scatter plot  
ggplot(data = cleaned\_merged\_data, aes(x = estimate, y = HappinessScore)) +  
 geom\_point(color = "blue", size = 2) +   
 geom\_smooth(method = "lm", color = "red", se = TRUE) +   
 scale\_x\_continuous(labels = scales::comma) +   
 labs(  
 x = "GDP Per Capita (USD)",  
 y = "Happiness Score",  
 title = "GDP Per Capita and Happiness Score"  
 ) +  
 theme\_minimal() +   
 theme(  
 plot.title = element\_text(size = 12)   
 )

## `geom\_smooth()` using formula = 'y ~ x'



## RQ2: How does environmental quality associate with a country’s happiness score?

# Clean the country names in environmental\_quality\_table\_cleaned  
environmental\_quality\_table\_cleaned <- environmental\_quality\_table\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Clean the country names in happiest\_countries\_cleaned  
happiest\_countries\_cleaned <- happiest\_countries\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Merge environmental quality data with happiness data  
merged\_data\_env <- environmental\_quality\_table\_cleaned %>%  
 left\_join(happiest\_countries\_cleaned, by = "country")  
print(merged\_data\_env)

## # A tibble: 179 × 3  
## country value HappinessScore  
## <chr> <dbl> <dbl>  
## 1 afghanistan 31 1.72  
## 2 angola 40.1 NA   
## 3 albania 52.2 5.3   
## 4 united arab emirates 51.6 6.73  
## 5 argentina 47 6.19  
## 6 armenia 44.9 5.46  
## 7 antigua and barbuda 55.6 NA   
## 8 australia 63.1 7.06  
## 9 austria 68.9 6.91  
## 10 azerbaijan 40.5 4.89  
## # ℹ 169 more rows

# Remove rows with missing values in value or HappinessScore  
cleaned\_merged\_data\_env <- merged\_data\_env %>%  
 filter(!is.na(value), !is.na(HappinessScore))  
print(cleaned\_merged\_data\_env)

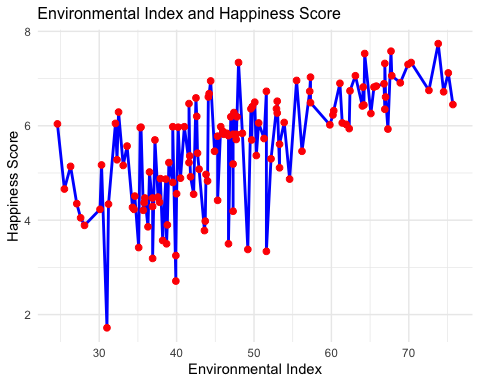
## # A tibble: 135 × 3  
## country value HappinessScore  
## <chr> <dbl> <dbl>  
## 1 afghanistan 31 1.72  
## 2 albania 52.2 5.3   
## 3 united arab emirates 51.6 6.73  
## 4 argentina 47 6.19  
## 5 armenia 44.9 5.46  
## 6 australia 63.1 7.06  
## 7 austria 68.9 6.91  
## 8 azerbaijan 40.5 4.89  
## 9 belgium 66.8 6.89  
## 10 benin 37.8 4.38  
## # ℹ 125 more rows

# View summary statistics   
summary(cleaned\_merged\_data\_env[, c("value")])

## value   
## Min. :24.60   
## 1st Qu.:38.40   
## Median :46.10   
## Mean :47.83   
## 3rd Qu.:56.70   
## Max. :75.70

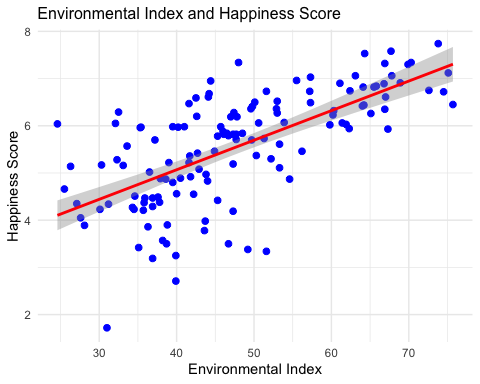
# Create the line graph  
ggplot(cleaned\_merged\_data\_env, aes(x = value, y = HappinessScore)) +  
 geom\_line(color = "blue", size = 1) +   
 geom\_point(color = "red", size = 2) +   
 labs(  
 title = "Environmental Index and Happiness Score",  
 x = "Environmental Index",  
 y = "Happiness Score"  
 ) +  
 theme\_minimal() +   
 theme(  
 plot.title = element\_text(size = 12)   
 )

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



# Create the scatter plot  
ggplot(cleaned\_merged\_data\_env, aes(x = value, y = HappinessScore)) +  
 geom\_point(color = "blue", size = 2) +   
 geom\_smooth(method = "lm", color = "red", se = TRUE) +   
 labs(  
 title = "Environmental Index and Happiness Score",  
 x = "Environmental Index",  
 y = "Happiness Score"  
 ) +  
 theme\_minimal() +   
 theme(  
 plot.title = element\_text(size = 12)   
 )

## `geom\_smooth()` using formula = 'y ~ x'



I created two graphs to explore and to find out the best graph. The line graph still helps demonstrate a general positive trend between environmental quality and happiness scores; however, I think the scatter plot graph is better.

## RQ3: How does a country’s literacy rate associate with its happiness score?

# Clean the country names in literacy\_rate\_table\_cleaned  
literacy\_rate\_table\_cleaned <- literacy\_rate\_table\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Clean the country names in happiest\_countries\_cleaned  
happiest\_countries\_cleaned <- happiest\_countries\_cleaned %>%  
 mutate(country = gsub("[^[:alnum:] ]", "", country)) %>%   
 mutate(country = trimws(country)) %>%   
 mutate(country = tolower(country))   
  
# Merge Literacy Rate and HappinessScore data  
merged\_data\_lit <- literacy\_rate\_table\_cleaned %>%  
 left\_join(happiest\_countries\_cleaned, by = "country")  
print(merged\_data\_lit)

## # A tibble: 195 × 4  
## country total year HappinessScore  
## <chr> <dbl> <dbl> <dbl>  
## 1 afghanistan 37.3 2021 1.72  
## 2 albania 98.1 2018 5.3   
## 3 algeria 81.4 2018 5.36  
## 4 andorra 100 2016 NA   
## 5 angola 71.1 2015 NA   
## 6 antigua and barbuda 99 2015 NA   
## 7 argentina 99 2018 6.19  
## 8 armenia 99.8 2020 5.46  
## 9 australia 99 NA 7.06  
## 10 austria 98 NA 6.91  
## # ℹ 185 more rows

# Remove missing values  
cleaned\_merged\_data\_lit <- merged\_data\_lit %>%  
 filter(!is.na(total), !is.na(year), !is.na(HappinessScore))  
print(cleaned\_merged\_data\_lit)

## # A tibble: 116 × 4  
## country total year HappinessScore  
## <chr> <dbl> <dbl> <dbl>  
## 1 afghanistan 37.3 2021 1.72  
## 2 albania 98.1 2018 5.3   
## 3 algeria 81.4 2018 5.36  
## 4 argentina 99 2018 6.19  
## 5 armenia 99.8 2020 5.46  
## 6 azerbaijan 99.8 2019 4.89  
## 7 bahrain 97.5 2018 5.96  
## 8 benin 42.4 2018 4.38  
## 9 bolivia 92.5 2015 5.78  
## 10 bosnia and herzegovina 98.5 2015 5.88  
## # ℹ 106 more rows

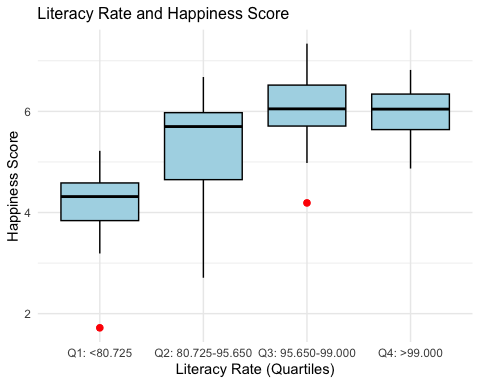
# View summary statistics   
summary(cleaned\_merged\_data\_lit[, c("total")])

## total   
## Min. : 22.30   
## 1st Qu.: 78.72   
## Median : 93.80   
## Mean : 85.00   
## 3rd Qu.: 98.40   
## Max. :100.00

# Calculate percentiles for the "total" column in the literacy rate table  
percentiles <- quantile(literacy\_rate\_table\_cleaned$total, probs = c(0.25, 0.5, 0.75, 1), na.rm = TRUE)  
print(percentiles)

## 25% 50% 75% 100%   
## 80.725 95.650 99.000 100.000

# Add bins to the data using the specified quartiles  
cleaned\_merged\_data\_lit <- cleaned\_merged\_data\_lit %>%  
 mutate(  
 literacy\_bin = cut(  
 total,  
 breaks = c(0, 80.725, 95.650, 99.000, 100),   
 labels = c("Q1: <80.725",   
 "Q2: 80.725-95.650",   
 "Q3: 95.650-99.000",   
 "Q4: >99.000"),  
 include.lowest = TRUE  
 )  
 )  
# Create the box plot  
ggplot(cleaned\_merged\_data\_lit, aes(x = literacy\_bin, y = HappinessScore)) +  
 geom\_boxplot(fill = "lightblue", color = "black", outlier.color = "red", outlier.size = 2) +  
 labs(  
 title = "Literacy Rate and Happiness Score",  
 x = "Literacy Rate (Quartiles)",  
 y = "Happiness Score"  
 ) +  
 theme\_minimal() +   
 theme(  
 plot.title = element\_text(size = 12)   
 )

 # 5. Data Interpretation {#data-interpretation}

## Data Analysis

In addressing the main question, “How do economic, environmental, and educational variables influence national happiness levels as reported by the World Population Review?”, this analysis provides evidence of significant associations between these factors and happiness scores.

*H1: I hypothesize that higher GDP per capita, measured in USD based on IMF estimates, is associated with higher national happiness levels.*

The scatter plot examining GDP per capita and happiness scores reveals a clear positive trend. Countries with higher GDP per capita tend to report higher happiness scores. This supports the hypothesis that wealthier countries have more resources to invest in public services such as healthcare, education, and infrastructure, which improve the overall quality of life.

*H2: I hypothesize that better environmental quality, as measured by the Environmental Performance Index (EPI), is associated with higher happiness.*

The line graph analyzing the association between environmental quality and happiness scores suggests a positive but variable association. Countries with higher environmental index scores generally report higher happiness levels, supporting the hypothesis that a clean and sustainable environment promotes well-being. However, the noisy pattern in the graph indicates that some countries with similar environmental scores differ significantly in happiness. The scatter plot graph shows a positive trend as well.

*H3: I hypothesize that higher literacy rates, as a measure of education, are associated with higher happiness scores.*

The box plot demonstrates a generally positive association between literacy rates and happiness scores across the first three quartiles (Q1 to Q3). However, Q3 (95.650–99.000) appears to have a slightly higher happiness scores compared to Q4 (> 99.000), despite Q4 representing the countries with the highest literacy rates. This pattern could reflect diminishing returns of literacy on happiness or challenges unique to highly literate societies, such as increased societal pressures, disparities in resource allocation, or inequality.

## Limitations and Future Directions

One limitation of this study lies in the quality and consistency of the data used. Data sources may vary in their methods of collection, measurements, and reporting standards across countries. Future research should prioritize using standardized global datasets to minimize inconsistencies and align data collection periods across variables to ensure accuracy. Conducting longitudinal studies could further enhance the understanding of how changes in variables influence happiness over time, providing a more comprehensive and precise perspective on these relationships.

As an extension of this study, we could perform statistical correlation tests, such as Pearson or Spearman correlation, to quantitatively assess the strength and direction of relationships between each variable and national happiness scores. This approach would allow us to identify which factor is most strongly correlated with happiness and could potentially be a primary determinant. Additionally, conducting regression analysis could help control for confounding variables and provide deeper insights into the relative importance of each factor in explaining variations in happiness.

## Implications

The positive link between GDP per capita and happiness emphasizes the importance of investments in public services like healthcare, education, and infrastructure, with a focus on reducing income inequality to maximize benefits. Similarly, the association between environmental quality and happiness underscores the need for sustainable policies that prioritize pollution reduction and green initiatives. The relationship between literacy rates and happiness suggests that education policies should focus not only on improving literacy but also on ensuring equitable access and opportunities, as the trend indicates that higher literacy rates do not always directly translate to higher happiness scores. Overall, these findings demonstrate the need for balanced strategies addressing economic, educational, and environmental factors to achieve sustainable well-being globally.