Global HDI and Happiness

Ahalya Kudalugodaarachchi, Ian Wang, Ankita Bhattacharyya

2025-12-17

Exploratory Data Analysis

Satisfying FAIR Principles - Human Development Index (HDI) Data Set

Findable - The data is collected from the United Nations. Because of this, the data is very reliable. Also, the original table contains all of the descriptions for certain terminology. Especially what's considered "High HDI", "Medium HDI", and "Low HDI", all have their definitions listed at the bottom.

Accessible - It's freely accessible and available in the Excel format. The report from the year is also posted in a link that leads to the pdf in the same webpage.

Interoperable - At the end of the provided report, it contains all of the survey questions that were used in the interviews that took place for the countries. However, the country names themselves are not the "standardized" ones and therefore, it would be difficult to use other data tables with that of the HDI table.

Reusable - All of the other columns in the data table are used to calculate the HDI and from there the HDI rank. Therefore, the attributes listed are all important to the argument.

Conclusion - The HDI data set from the UN, overall does meet the requirements of the FAIR standards. However, the "I" or "Interoperable" is not sufficiently met compared to the others.

Satisfying CARE Principles - HDI Dataset

Collective Benefit - The UN aims to use this data to uniformly and properly track the development of each country in the world. Some territories are also included. The UN uses this data to compare countries data from the HDI standpoint but also from other attributes like mean schooling years or gross national income (GNI). This helps the UN to best figure out what each country or region needs or is affected by the most and the UN can help with suggesting policy changes or even providing aid.

Authority to Control - The data is collected by the United Nations. They had people who were sent out to all or nearly all 195 countries during a specific time period and they collected answers to the survey questions. For countries that had little internet or wifi access, surveys were conducted in-person. While the data and subsequent graphs are available to the public. The countries themselves are unable to determine how their data is used within the report.

Responsibility - The UN analyzed the data they collected themselves and released their findings in the UN report which is a booklet written in English. There is no way to have it translated to a different language. There are no other language options which greatly minimizes who is reached especially given that nearly all 195 countries have data that has been recorded.

Ethics - Data is only summarized by country and therefore nothing confidential has been released in either the report or the raw table. It ethically obtained its data with the consent of its participants.

Conclusion - Because the HDI dataset is sufficient in most of the criteria it mostly passes the CARE principles. It's only weak on the "Responsibility" aspect of how the information is released to the public.

Graphs and Analysis

Figure 1: A scatter plot of Countries' Human Development Index and Happiness Score for the year 2022.

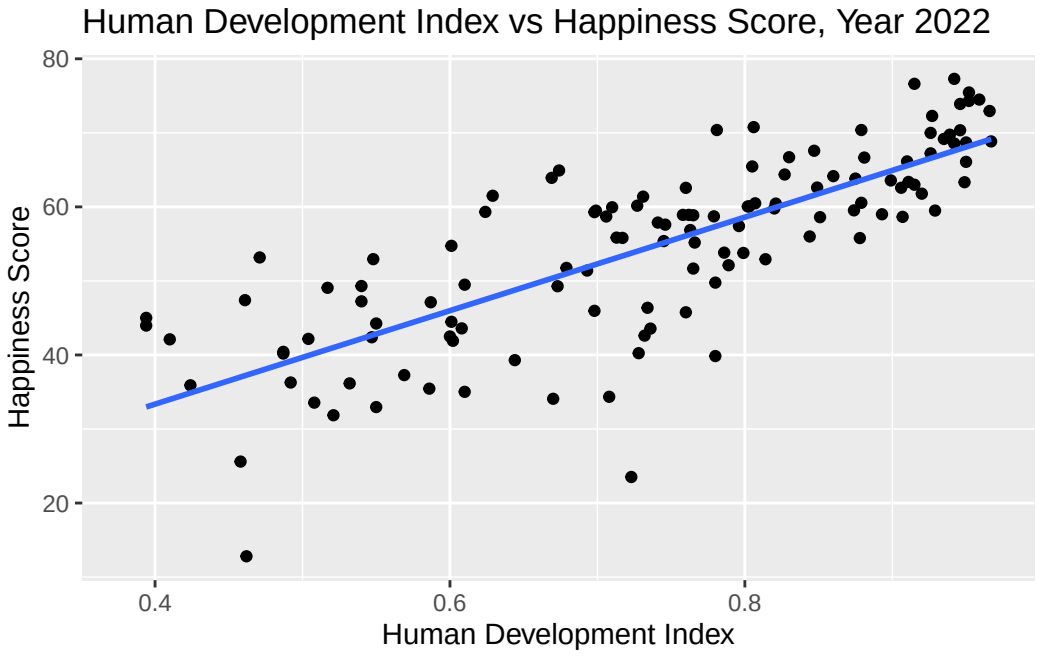
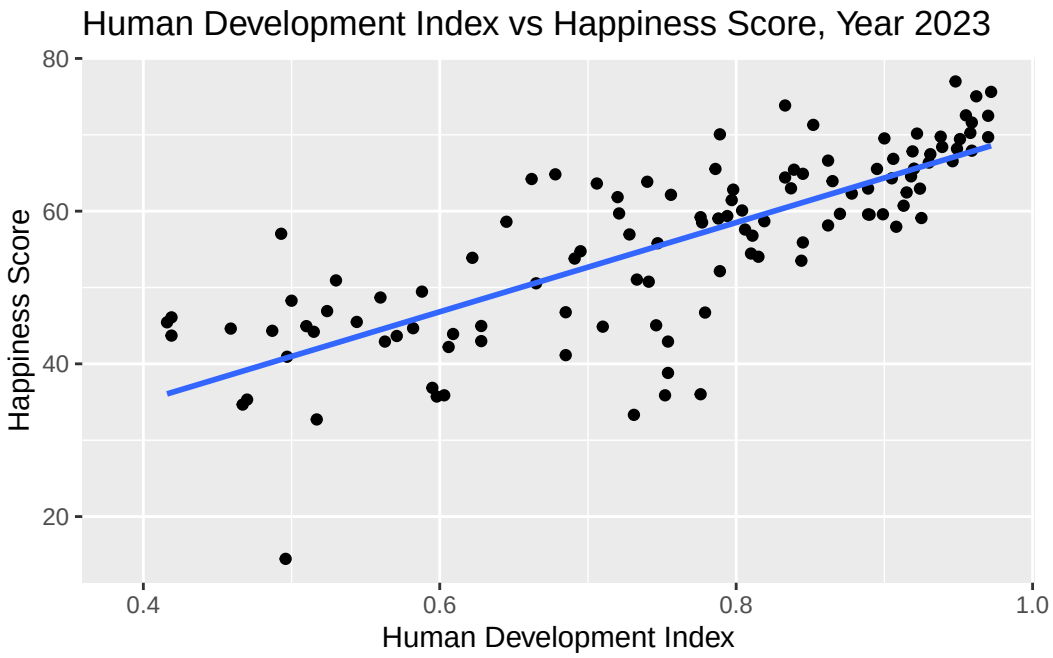


Figure 2: A scatter plot of Countries' Human Development Index and Happiness Score for the year 2023.



In regards to the graphs separately, both Figure 1 and Figure 2 have strong positive correlations between Human Development Index and Happiness Score. For the both the plots the each black dot represents a country. The blue trend line represents shows that the countries with higher human development index, which also means better education, income and life expectancy, tend to report higher happiness levels. This shows that humen development could be a significant factor in shaping the well being of nations' populations. Although overall, both graphs show

Table 1: Five Number Summary for Global Human Development Index

Countries	Minimum	Quartile 1	Median	Quartile 3	Maximum	Mean	Standard Deviation	Year
207	0.380	0.602	0.740	0.847	0.967	0.7237409	0.1551773	2022
193	0.388	0.622	0.762	0.862	0.972	0.7407979	0.1516078	2023

Table 2: Five Number Summary for Global Happiness Score

Countries	Minimum	Quartile 1	Median	Quartile 3	Maximum	Mean	Standard Deviation	Year
207	12.81	45.9600	57.880	63.5700	77.29	54.87562	12.35355	2022
193	14.46	45.4775	58.645	64.8375	76.99	55.94750	11.80071	2023

words i guess

Author Contribution

Ahalya worked on the wrangling of the Human Development Index data set, and created the Quarto document, as well as wrote the Exploratory Analysis. Ian wrote the entire README file, as well as the plan document.

Code Appendix

```
#Tidyverse style was followed throughout.
# Wrangling/Scraping and Cleaning of HDI and Happiness Score data 2022 -------
library(tidyverse)
library(rvest)

#scraping raw HDI 2022 Data
HDI_2022_raw <- read_html("https://countryeconomy.com/hdi") %>%
  html_elements(css = "table") %>%
  html_table()

HDI_2022 <- (HDI_2022_raw[[1]])

HDI_2022_clean <- HDI_2022 %>%
  select(-c(,3:5)) %>%
  separate_wider_delim(
    cols = Countries,
    delim = " [",
    names = c("Country", "junk")
    ) %>%
  select(-c(,2)) %>%
  rename()

happiRaw <- read.csv("hapiscore_whr.csv") # read in csv file

happiClean_2022 <- happiRaw %>%
  select(-c(geo, X2005:X2021, X2023)) %>% # removing unneeded columns
  rename(
    Happiness_Score_2022 = X2022, #rename to a more appropriate name
    Country = name
  )

# HDI and Happiness Score 2022 Plot
# creating the combined data
HDI_HappiScore_2022 <- full_join(
  x = HDI_2022_clean,
  y = happiClean_2022,
  by = join_by(Country == Country)
)
```

```r
#creating visualization HDI vs HappiScore 2022
ggplot(
  data = HDI_HappiScore_2022, # data set
  mapping = aes(
    x = HDI, # x variable
    y = Happiness_Score_2022 # y variable
  )
) + geom_point() + labs(
  title = "Human Development Index vs Happiness Score, Year 2022", # title
  x = xlab("Human Development Index"), # x-axis label
  y = ylab("Happiness Score") # y-axis label
) + geom_smooth( # line of best fit
  method = "lm", # linear model
  se = FALSE
)
# Wrangling and Cleaning of HDI and Happiness Score data 2023 -------

HDI_raw <- read_csv("HDR25_Statistical_Annex_HDI_Table(Table 1.csv")

HDI_clean <- HDI_raw %>%
  select(-c(1, 4:16)) %>%
  slice(-c(1:7, 82, 133, 177, 204:278)) %>%
  rename("Country" = `Table 1. Human Development Index and its components`,
         "HDI" = `...3`) %>%
  mutate(
    HDI = as.numeric(HDI)
  )

happiRaw <- read.csv("hapiscore_whr.csv") # read in csv file

# cleaning the data
happiClean <- happiRaw %>%
  select(-c(geo, X2005:X2022)) %>% # removing unneeded columns
  rename(
    Happiness_Score_2023 = X2023, #rename to a more appropriate name
    Country = name
  )

happiClean_2022 <- happiRaw %>%
  select(-c(geo, X2005:X2021, X2023)) %>% # removing unneeded columns
  rename(
    Happiness_Score_2022 = X2022, #rename to a more appropriate name
    Country = name
  )

#  HDI and Happiness Score 2023 Plot
library(tidyverse)
```

```r
library(ggplot2)


# creates full data set
HDI_Happiscore_clean <- left_join(
  x = HDI_clean,
  y = happiClean,
  by = join_by(Country == Country)
)

# HDI vs Happiness Score (2023)
ggplot(
  data = HDI_Happiscore_clean,
  mapping = aes(
    x = HDI, # x variable
    y = Happiness_Score_2023 # y variable
  )
) + geom_point() + labs(
  title = "Human Development Index vs Happiness Score, Year 2023", # title
  y = ylab("Happiness Score"), # x axis
  x = xlab("Human Development Index") # y axis
) + geom_smooth(
  method = "lm", # creates best fit line
  se = FALSE # removes shaded part
)

# Creating Summary Statistics for each attribute-----------
library(tidyverse)
library(rvest)
library(kableExtra)

HDI2022_stats <- HDI_HappiScore_2022 %>%
  summarize(
    Countries = n(),
    Minimum = min(HDI, na.rm = TRUE),
    `Quartile 1` = quantile(HDI, probs = 0.25, na.rm = TRUE),
    Median = median(HDI, na.rm = TRUE),
    `Quartile 3` = quantile(HDI, probs = 0.75, na.rm = TRUE),
    Maximum = max(HDI, na.rm = TRUE),
    Mean = mean(HDI, na.rm = TRUE),
    `Standard Deviation` = sd(HDI, na.rm = TRUE),
  ) %>%
  mutate(
    Year = "2022"
  )

Happi2022_stats <- HDI_HappiScore_2022 %>%
```

```
  summarize(
    Countries = n(),
    Minimum = min(Happiness_Score_2022, na.rm = TRUE),
    `Quartile 1` = quantile(Happiness_Score_2022, probs = 0.25, na.rm = TRUE),
    Median = median(Happiness_Score_2022, na.rm = TRUE),
    `Quartile 3` = quantile(Happiness_Score_2022, probs = 0.75, na.rm = TRUE),
    Maximum = max(Happiness_Score_2022, na.rm = TRUE),
    Mean = mean(Happiness_Score_2022, na.rm = TRUE),
    `Standard Deviation` = sd(Happiness_Score_2022, na.rm = TRUE)
  ) %>%
  mutate(
    Year = "2022"
  )


HDI2023_stats <- HDI_Happiscore_clean %>%
  summarize(
    Countries = n(),
    Minimum = min(HDI, na.rm = TRUE),
    `Quartile 1` = quantile(HDI, probs = 0.25, na.rm = TRUE),
    Median = median(HDI, na.rm = TRUE),
    `Quartile 3` = quantile(HDI, probs = 0.75, na.rm = TRUE),
    Maximum = max(HDI, na.rm = TRUE),
    Mean = mean(HDI, na.rm = TRUE),
    `Standard Deviation` = sd(HDI, na.rm = TRUE),
  ) %>%
  mutate(
    Year = "2023"
  )

Happi2023_stats <- HDI_Happiscore_clean %>%
  summarize(
    Countries = n(),
    Minimum = min(Happiness_Score_2023, na.rm = TRUE),
    `Quartile 1` = quantile(Happiness_Score_2023, probs = 0.25, na.rm = TRUE),
    Median = median(Happiness_Score_2023, na.rm = TRUE),
    `Quartile 3` = quantile(Happiness_Score_2023, probs = 0.75, na.rm = TRUE),
    Maximum = max(Happiness_Score_2023, na.rm = TRUE),
    Mean = mean(Happiness_Score_2023, na.rm = TRUE),
    `Standard Deviation` = sd(Happiness_Score_2023, na.rm = TRUE)
  ) %>%
  mutate(
    Year = "2023"
  )

HDI_stats <- bind_rows(HDI2022_stats, HDI2023_stats)
Happi_stats <- bind_rows(Happi2022_stats, Happi2023_stats)
```

```r
# Summary Statistics for Global Human Development Table ---------------

HDI_stats %>%
  kable(
    caption = "Five Number Summary for Global Human Development Index", # title of table
    booktabs = TRUE,
    align = c("l", rep("c",6)),
    format.args = list(big.mark = ',') # adds separators to the number totals
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped","condensed"),
    font_size = 10,
    stripe_color = "gray!10" # adds back stripes
  )

# Summary Statistics for Global Happiness Score Table ---------------
Happi_stats %>%
  kable(
    caption = "Five Number Summary for Global Happiness Score", # title of table
    booktabs = TRUE,
    align = c("l", rep("c",6)),
    format.args = list(big.mark = ',') # adds separators to the number totals
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped","condensed"),
    font_size = 10,
    stripe_color = "gray!10" # adds back stripes
  )
```