Quality of Life by Country: A Clustering Analysis

Katherine M. Prioli

December 22, 2018

Abstract

BACKGROUND As economic globalization increases, it is important to understand how countries compare on key quality of life (QoL) measures when grouped by level of human development, and how these comparative results change by grouping approach. In the context of recent news reports about decreasing life expectancy in the United States, it is of particular interest to understand how the United States performs compared to other countries of a similar developmental level. METHODS Data for QoL measures of interest in calendar year 2016 was gathered and merged into a single country-level dataset. Each variable was examined via descriptive statistics and univariate visualizations, and pairwise relationships were investigated with a correlation matrix. The top- and bottom-ranking 20 countries were determined for each QoL measure. For those variables for which the United States was not among the 20 highest performing countries, a series of pairwise k-means cluster analyses were run, with a sensitivity analysis performed on the number of clusters. RESULTS The analytic dataset included 9 QoL variables (8 continuous, one categorical). On country-level analysis, the US was not among the top 20 countries for 5/8 (62.5%) continuous variables, yielding 10 pairwise comparisons for clustering. Cluster analyses showed that, for the US, life expectancy at birth was most predictive of cluster position, and was most sensitive to changes in the number of clusters. CONCLUSION Though the United States has the world's largest economy, it underperforms on several key QoL measures. K-means clustering provides a powerful tool for investigating these deficits. These findings may help inform health-related public policy aimed at decreasing deficits in life expectancy.

Background

The Centers for Disease Control and Prevention has recently issued a report indicating that life expectancy in the United States decreased in 2017 as compared to 2016, with the overwhelming majority of deaths caused by heart disease and cancer, arguably preventable illnesses (Murphy et al. (2018)). Given that the United States has the world's largest economy, this decline in life expectancy is particularly concerning, and indicates that national wealth may not be predictive of citizens' longevity (The World Bank (2018)).

As the world's economies trend toward globalism, there is increasing interest in understanding how these nations compare on key quality of life (QoL) factors, including but not limted to life expectancy. Several organizations report on QoL measures as they evolve, including among others the World Economic Forum (WEF), World Health Organization (WHO), and the United Nations Development Programme. The QoL measures reported by these bodies can be either unidimensional values or compound scores calculated from several factors of interest.

The objective of this study was to explore the relationships between key QoL indicators by country, with particular focus on how the United States ranks, through a series of visualizations and k-means clustering analyses.

Methods

This study included country-level QoL indicators as described in Table 1.

Table 1. Country-Level QoL measures.

Measure	Single or Compound	Description	Source[1]
Social Progress Index	Compound	Social progress level, scaled from 0:100 and comprising three broad categories: basic human needs (e.g., nutrition, safety), foundations of wellbeing (e.g., basic knowledge, environmental quality), and opportunity (e.g., personal rights, freedoms)	1
Gross Domestic Product (GDP)	Single	Valued in $US 2018	2
Human Development Index (HDI)	Compound	Developmental level, scale of 0:1	3
Human Development Index (HDI)	Compound	Developmental category, four levels (low, medium, high, very high)	3
World Happiness Score	Compound	Happiness score, scaled from 0:10, based on several factors including per-capita GDP, healthy life expectancy, social support, freedoms, and perception of corruption	4
Global Gender Gap Index	Single	Gender equality index, scaled from 0:1, based on measurements of gender-related gaps in such dimensions as economic participation, level of education, health and survival, and political offices held	5
Infant mortality rate	Single	Number of infant deaths per 1,000 live births	6
Life expectancy at birth	Single	Expected life at birth, both genders	7
Life expectancy at sixty	Compound	Expected remaining life years at age 60, both genders	7

[1] Source: **1** Social Progress Imperative (2018b), **2** The World Bank (2018), **3** The United Nations Development Programme (2018b), **4** Helliwell, Layard, and Sachs (2018), **5** World Economic Forum (2016b), **6** World Health Organization (2018b), **7** World Health Organization (2018a)

Data for these measures was obtained for calendar year 2016 in `.csv` or `.xls(x)` formats. Additionally, dataframe containing country identifiers (full names and three-letter codes) was generated from the `countrycode` library to facilitate merging the datafiles into one tibble.

Wrangling and Exploration

Each country-level datafile was imported, wrangled as needed, then tested against the dataframe containing country identifiers via `anti_join()` to identify mismatches. Mismatching country names were manually recoded for each datafile, then all datafiles were merged using serial joins. Countries with wholly missing data were excluded. The resulting dataframe, titled `alldata`, is presented in Table 2.

Table 2. `alldata` dataframe contents.

Source	Variable Name	Description
Social Progress Imperative (2018a)	SPI	Social Progress Index value (scale of 0:100)
The World Bank (2018)	GDP_USD_2018	Gross Domestic Product (valued in $US 2018)
The United Nations Development Programme (2018a)	HDIrank	Human Development Index ranking
The United Nations Development Programme (2018a)	HDIindex	HDI index value (scale of 0:1)
The United Nations Development Programme (2018a)	HDI_cat	HDI index category (4 levels)
Helliwell, Layard, and Sachs (2018)	happiness	World Happiness Score (scale of 0:10)
World Economic Forum (2016a)	gendereq	Gender Equality Index (scale of 0:1)
World Health Organization (2018b)	infantmort	Infant mortality rate
World Health Organization (2018a)	birth_MF	Life expectancy at birth, males & females
World Health Organization (2018a)	sixty_MF	Life expectancy at 60 years, males & females

At least one univariate visualization was generated for each variable in `alldata` via `ggplot()`, and a correlation matrix was produced with `GGally::ggpairs()` to investigate pairwise relationshps between variables. Next, a series of ordered country-as-factor bivariate visualizations were created to explore the top and bottom 20 countries by ranking within each variable, with the United States denoted in red.

K-Means Clustering

For the variables for which the United States was not among the top 20 performing countries on country-level visualization, a series of k-means cluster analyses was performed. K-means clustering is described elsewhere; briefly, given bivariate data and a desired number k of groups, this classification algorithm classifies the points in the two-dimensional plane to minimize the total within-cluster variation for all clusters (James et al. (2013)). This is an iterative process that works by establishing k centroids, classifying each point by which centroid is closest, then moving the centroids to the center of their corrresponding clusters, and repeating the process. Iteration terminates when the centroids no longer move, and the classification established in this terminal iteraton is the clustering.

The Human Development Index categorizes the world's countries into four developmental levels (low, medium, high, and very high); thus k-means clustering analysis was performed assuming 4 clusters. Missing values were excluded to ensure the clustering algorithm would run, and a function was written to subset the clustering dataset (named `clusterdata`) to the variables of interest for each k-means analysis. On each output plot, the United States was identified by an enlarged `geom_point()`. A Shiny application was written to allow sensitivity analysis of the effect of varying k, with particular attention paid to any transitions between clusters for the United States with changing k. For brevity, code for the Shiny application is not provided here, but is available on the GitHub repository for this project (Prioli (2018)). The application can be accessed at https://kmprioli.shinyapps.io/MAT_8790_kmeans/.

Example Code

For brevity, example code for wrangling, plotting, and clustering is presented here; full code is available in the GitHub repository for this project (Prioli (2018)).

The libraries required for this analysis were loaded as shown below.

```r
library(tidyverse)
library(readxl)           # For importing .xls(x) datasets
library(countrycode)      # For establishing uniform country identifiers
library(ggthemr)          # For prettifying output
library(gridExtra)        # For grid.arrange()
library(grid)             # For textGrob() to annotate grid.arrange() elements
library(kableExtra)       # For nicer output tables
library(GGally)           # For ggpairs() correlation matrix
library(wesanderson)      # For Wes Anderson palette

ggthemr("fresh")                                                  # Prettifying plot framework
wes <- wes_palette("Darjeeling1", 5, type = "discrete")      # Establishing color scheme for cluster plots
```

Code for establishing a crosswalk for country names and 3-letter codes is shown below.

```r
countries_full <- codelist_panel %>%
  select(country.name.en, year, genc3c, iso3c, wb_api3c) %>%
  group_by(country.name.en) %>%
  mutate(maxyr = max(year)) %>%
  ungroup %>%
  mutate(maxyr = case_when(
    maxyr == year ~ 1,
    TRUE ~ 0
  )) %>%
  filter(maxyr == 1) %>%
  select(-maxyr) %>%
  distinct()

countries_full <- countries_full %>%
  mutate(country3 = case_when(
    iso3c == genc3c & iso3c == wb_api3c ~ iso3c,
    is.na(iso3c) == FALSE ~ iso3c,
    is.na(iso3c) == TRUE & is.na(genc3c) == FALSE ~ genc3c,
    is.na(iso3c) == TRUE & is.na(genc3c) == TRUE & is.na(wb_api3c) == FALSE ~ wb_api3c
  )) %>%
```

```r
  rename(country = country.name.en) %>%
  arrange(country)

countries <- countries_full %>%
  select(country, country3)
```

Code for importing and wrangling each data file and standardizing country names (example shown for the Social Progress Index data) is given below.

```r
# Importing, subsetting, renaming

SPI_2016_raw <- read_xlsx("data/SPIdata.xlsx", sheet = 4)

SPIdata <- SPI_2016_raw %>%
  select(2:3) %>%
  rename(`SPI` = `Social Progress Index`,
         country3 = Code)

# Standardizing country names by using `anti_join()` to see which
# countries in `SPIdata` don't have a match in the `countries` dataframe

SPIanti <- SPIdata %>%
  anti_join(countries, by = "country3") %>%
  select(country3) %>%
  arrange(country3) %>%
  unique()
dim(SPIanti)
```

```
## [1] 5 1
```

```r
# Correcting for mismatches with `countries` using `mutate()`

SPIdata <- SPIdata %>%
  mutate(country3 = case_when(
    country3 == "CHI" ~ as.character(NA), # Nonstandard code for Chile; omitting (no data in these rows)
    country3 == "KSV" ~ "XKS",            # Nonstandard code for Kosovo
    country3 == "NCY" ~ as.character(NA), # Turk. Repub. of N. Cyprus; omitting (conflict w/Cyprus)
    country3 == "SML" ~ as.character(NA), # Unable to determine
    country3 == "WBG" ~ as.character(NA), # West Bank / Gaza Strip; omitting (conflict w/Palestine)
    TRUE ~ as.character(country3)
  )) %>%
  filter(!is.na(country3))

SPIanti <- SPIdata %>%
  anti_join(countries, by = "country3") %>%
  select(country3) %>%
  arrange(country3) %>%
  unique()
dim(SPIanti)
```

```
## [1] 0 1
```

```r
# Removing unneeded files

rm(list = c("SPI_2016_raw", "SPIanti"))
```

Shown below is the code used to combine individual data files into one dataframe and filter out countries with no data.

```r
joindata_1 <- full_join(countries, HDIdata, by = "country")
joindata_2 <- left_join(joindata_1, SPIdata, by = "country3")
joindata_3 <- left_join(joindata_2, WHRdata, by = "country")
joindata_4 <- left_join(joindata_3, genderdata, by = "country")
joindata_5 <- left_join(joindata_4, infantmortdata, by = "country")
```

```
joindata_6 <- left_join(joindata_5, lifeexpdata, by = "country")
joindata_7 <- left_join(joindata_6, GDPdata, by = "country3")

joinsub <- joindata_7 %>%
  arrange(country) %>%
  mutate(exclude_flag = case_when(
    is.na(HDIrank) == TRUE &
      is.na(HDIindex) == TRUE &
      is.na(HDI_cat) == TRUE &
      is.na(SPI) == TRUE &
      is.na(happiness) == TRUE &
      is.na(gendereq) == TRUE &
      is.na(infantmort) == TRUE &
      is.na(birth_MF) == TRUE &
      is.na(sixty_MF) == TRUE &
      is.na(GDP_USD_2018) == TRUE ~ TRUE,
    TRUE                          ~ FALSE
  )) %>%
  filter(exclude_flag == FALSE) %>%
  select(-exclude_flag)

alldata <- joinsub %>%
  mutate(country = factor(country)) %>%
  mutate(country3 = factor(country3)) %>%
  mutate(US = case_when(                              # Flag for US vs non-US
    country == "United States" ~ "US",
    TRUE                       ~ "Non US"
  )) %>%
  mutate(color = case_when(                           # Color var for use in country-level plots
    country == "United States" ~ "#FF0000",
    TRUE                       ~ "#545454"
  ))

alldata <- alldata[c(1:2, 13:14, 6, 12, 3:5, 7:11)]

len <- dim(alldata)[[1]]                              # For use in calculating histogram bin number

# write_csv(alldata, paste0("data/alldata_", lubridate::today(),".csv"))  # Uncomment to export data
```

The following example code shows the steps taken for univariate explorations, including plotting and descriptive statistics.

```
SPI_hist <- ggplot(data = alldata, aes(x = SPI)) +
  geom_histogram(bins = ceiling(sqrt(len - sum(is.na(alldata$SPI))))) +
  xlab("Social Progress Index") +
  ylab("Count") +
  ggtitle("Figure 1.  Social Progress Index Distribution")

SPIsumm <- broom::tidy(round(summary(alldata$SPI), digits = 3))
sd <- round(sd(alldata$SPI, na.rm = TRUE), digits = 3)
SPIsumm <- cbind(SPIsumm, sd)
colnames(SPIsumm) <- c("Min", "Q1", "Median", "Mean", "Q3", "Max", "NA", "SD")
SPIsumm <- SPIsumm[c(1:6,8,7)]
SPIsumm_grob <- tableGrob(t(SPIsumm), theme = ttheme_minimal())

grid.arrange(SPI_hist, SPIsumm_grob, nrow = 1, widths = c(0.8, 0.2))
```

Example code for ordered country-level plots is shown below.

```
alldata_SPI <- alldata %>%
  filter(!is.na(SPI) == TRUE) %>%
  arrange(desc(SPI)) %>%
```

```
  select(SPI, country, US, color)

alldata_SPI_top20 <- alldata_SPI %>% head(20)
alldata_SPI_bot20 <- alldata_SPI %>% tail(20)
alldata_SPI40 <- bind_rows(alldata_SPI_top20, alldata_SPI_bot20)

colors <- alldata_SPI40$color[order(alldata_SPI40$SPI)]

SPI_country_point <- ggplot(data = alldata_SPI40, aes(x = SPI,
                                                      y = fct_reorder(country, SPI), color = US)) +
  geom_point() +
  scale_color_manual(values = c("US" = "#FF0000", "Non US" = "#5BBCD6")) +
  theme(axis.text.y = element_text(color = colors)) +
  guides(color = FALSE) +
  xlim(0, 100) +
  xlab("SPI") +
  ylab("Country") +
  ggtitle("Figure 9.  Social Progress Index by Country")
```

Finally, example clustering code is presented below.

```
kmdata <- kmdf(clusterdata, "country", "happiness", "gendereq")

set.seed(19811221)
km_subset <- kmeans(kmdata[, 2:3], 4)
km_subset_cluster <- as.factor(km_subset$cluster)

clusterdata1 <- cbind(clusterdata, km_subset_cluster)

km_happiness_gendereq_plot <- ggplot(data = clusterdata1,
                                     aes(x = happiness, y = gendereq,
                                         color = km_subset_cluster,
                                         size = US,
                                         shape = HDI_cat)) +
  geom_point() +
  scale_color_manual(values = wes) +
  scale_shape_manual(values = c(18, 17, 15, 16)) +
  guides(color = guide_legend(title = "Cluster"),
         size = FALSE,
         shape = guide_legend(reverse = TRUE, title = "HDI Category")) +
  xlab("Happiness Score") +
  ylab("Gender Equality Index") +
  ggtitle("Figure 16.  Cluster Analysis, Gender Equality Index vs. Happiness Score")
km_happiness_gendereq_plot
```

Results

Data Exploration and Visualizations

The Social Progress Index data ranges from 26.01 to 89.62 and appears trimodal (Fig. 1).

Figure 1. Social Progress Index Distribution

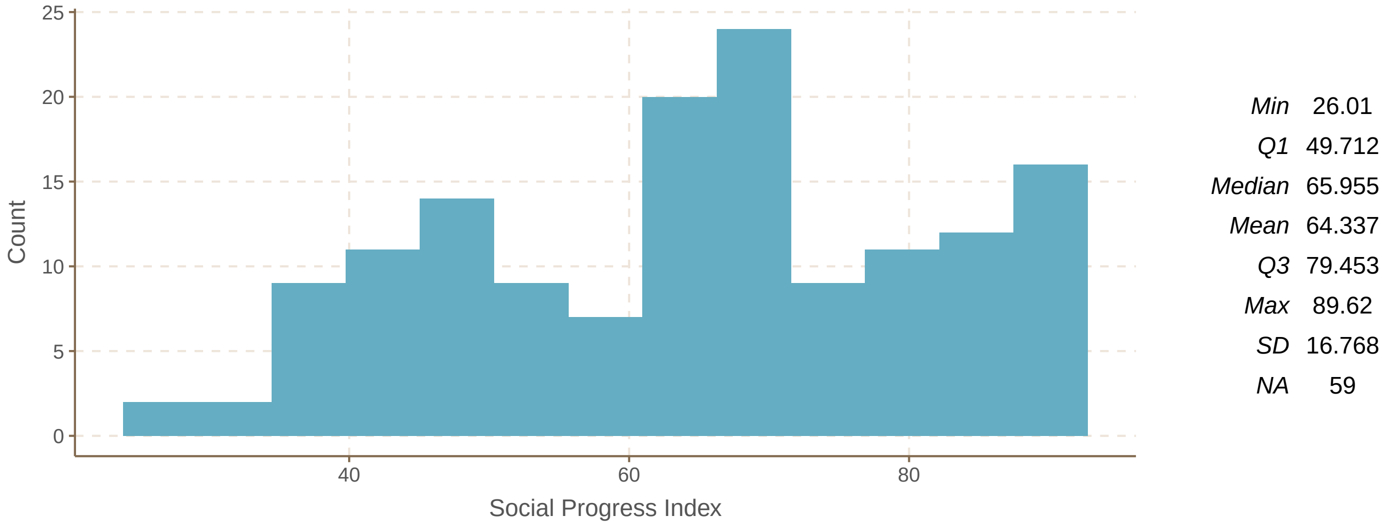
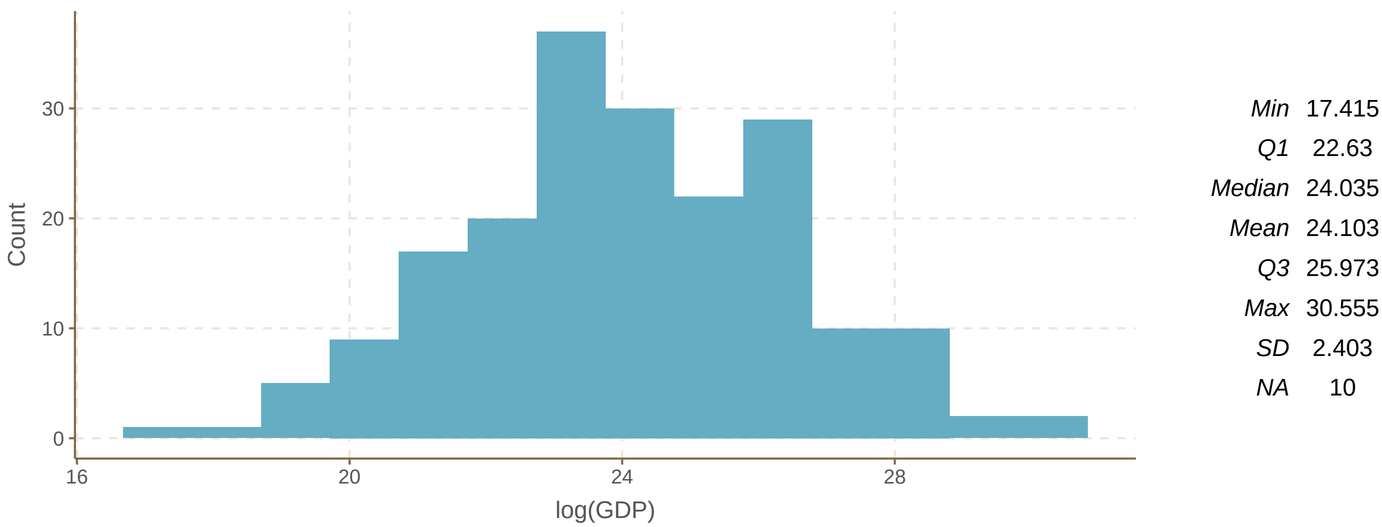


Table 3 shows summary statistics for the raw GDP values. These are unwieldy; however, after log transformation, the data is reasonably normally distributed, with mean 24.1 and standard deviation 2.4 (Fig. 2).

Table 3. Summary Statistics for `GDP_USD_2018`

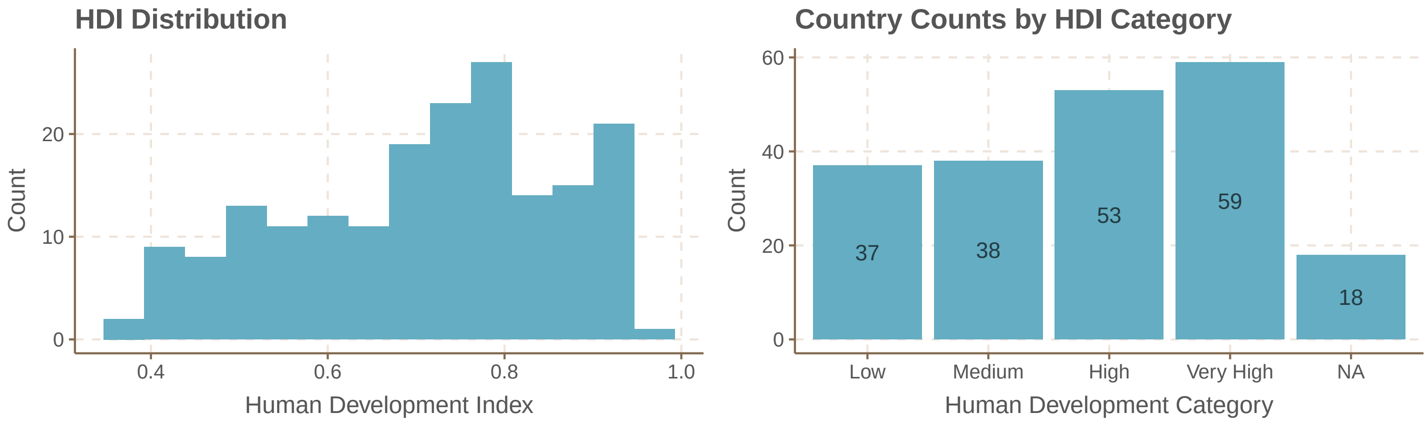
Min	Q1	Median	Mean	Q3	Max	SD	NA
36572612	6734069913	27424071373	383069641832	1.90463e+11	1.86245e+13	1.640295e+12	10

Figure 2. Gross Domestic Product Distribution, Log Transform



The Human Development Index data (Fig. 3) appears multimodal, with the "very high" developmental category the most represented in the data.

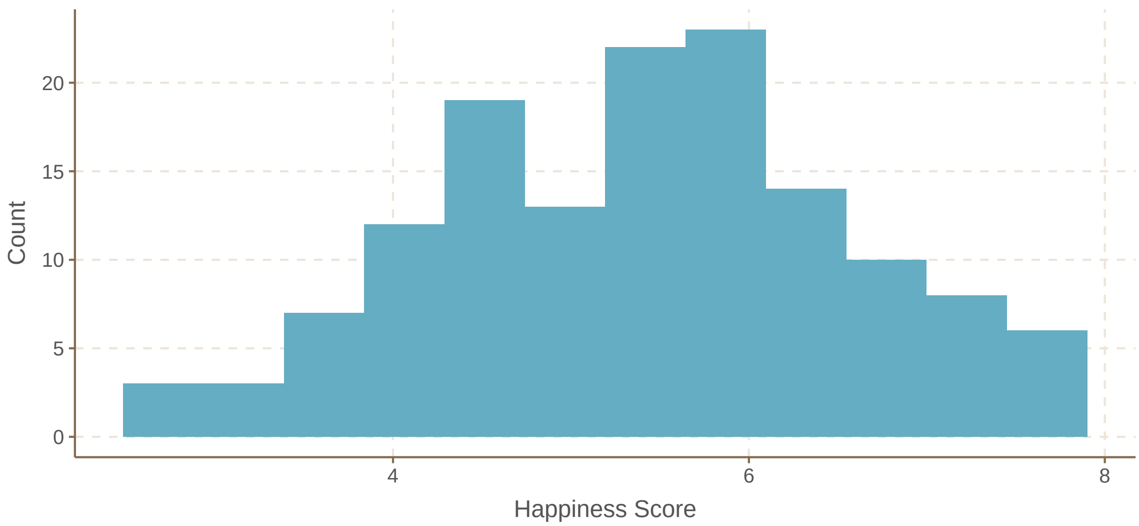
Figure 3. Human Development Index



Min	Q1	Median	Mean	Q3	Max	SD	NA
0.351	0.589	0.737	0.709	0.822	0.951	0.153	19

Figure 4 shows that the Happiness Score is reasonably normally distributed, having values ranging from 2.69 to 7.66.

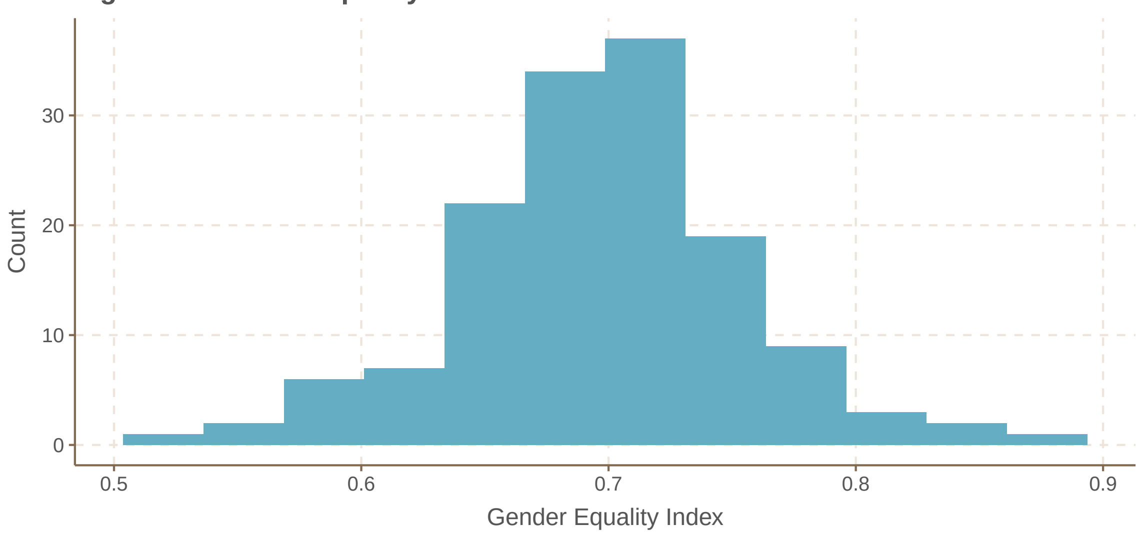
Figure 4. Happiness Score Distribution



Min	2.693
Q1	4.52
Median	5.424
Mean	5.402
Q3	6.145
Max	7.66
SD	1.146
NA	65

Figure 5 shows that the gender equality index is roughly symmetric in distribution, with mean and median quite close in value (0.6953 and 0.6983 respectively).

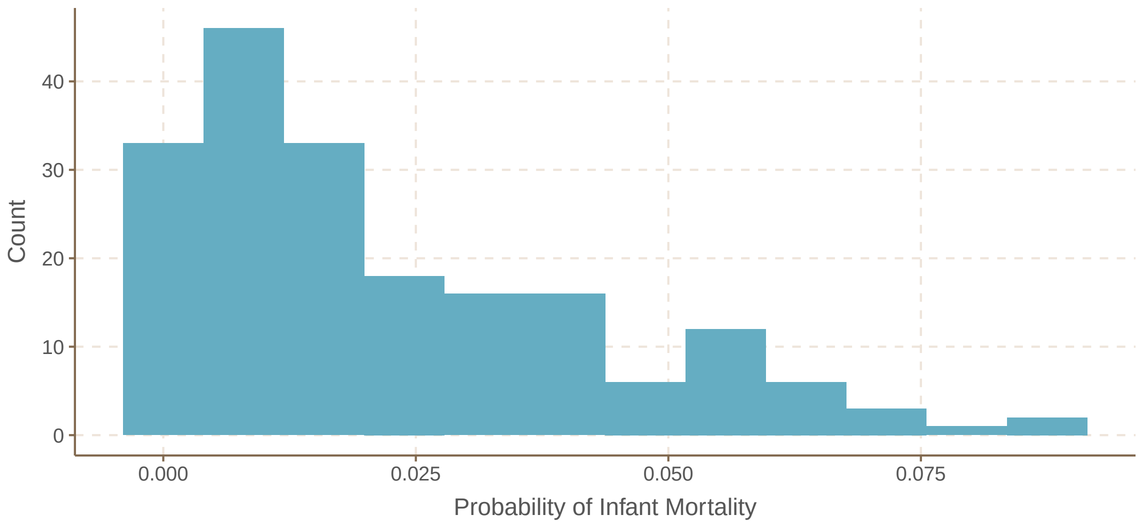
Figure 5. Gender Equality Index Distribution



Min	0.516
Q1	0.665
Median	0.698
Mean	0.695
Q3	0.729
Max	0.874
SD	0.057
NA	62

The infant mortality data (Fig. 6) is strongly right-skewed.

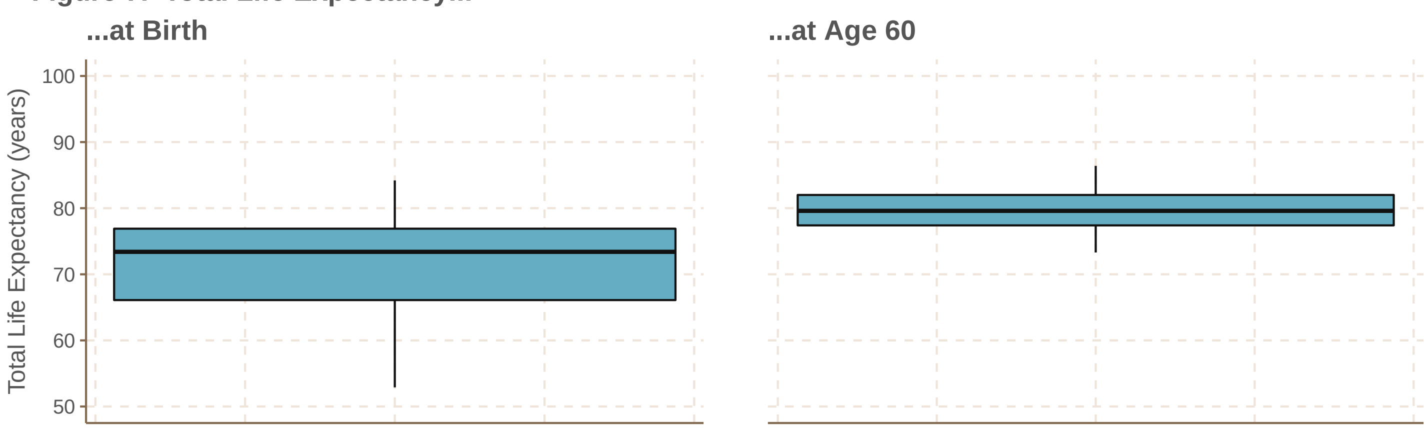
Figure 6. Infant Mortality Distribution



Min	0.002
Q1	0.007
Median	0.015
Mean	0.022
Q3	0.034
Max	0.089
SD	0.02
NA	13

The total life expectancy data (Fig. 7) shows that total life expectancy for those who reach 60 years of age is greater than that for the general population at birth.

Figure 7. Total Life Expectancy...



Min	Q1	Median	Mean	Q3	Max	SD	NA
52.9	66.1	73.4	71.9	76.9	84.2	7.564	24

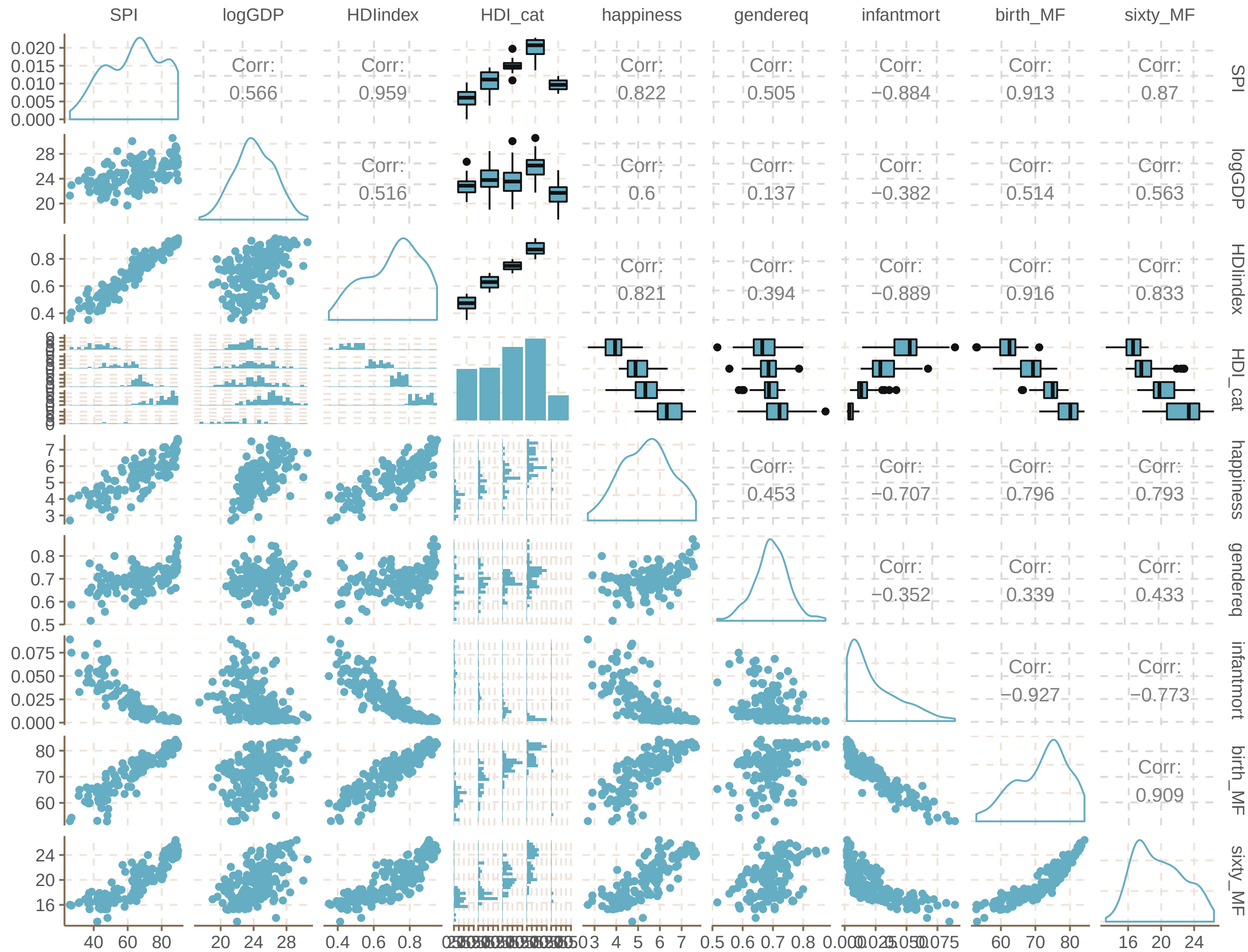
Min	Q1	Median	Mean	Q3	Max	SD	NA
73.3	77.4	79.6	79.793	82	86.4	2.981	24

When investigating pairwise relationships (Fig. 8), strong positive linear relationships were seen between `HDIindex` and `SPI`, `happiness`, and `birth_MF`; between `SPI` and `happiness`, `birth_MF`, and `sixty_MF`; and between `happiness` and `sixty_MF`. Additionally, strong positive relationships that are possibly nonlinear were seen between `HDI_index` and `sixty_MF`, and between `birth_MF` and `sixty_MF`.

Strong negative relationships were seen between `infantmort` and `birth_MF`, between `HDIindex` and `infantmort`, and between `SPI` and `infantmort`, though the latter two of these may not necessarily be linear. A strong negative nonlinear relationship was seen between `infantmort` and `sixty_MF`.

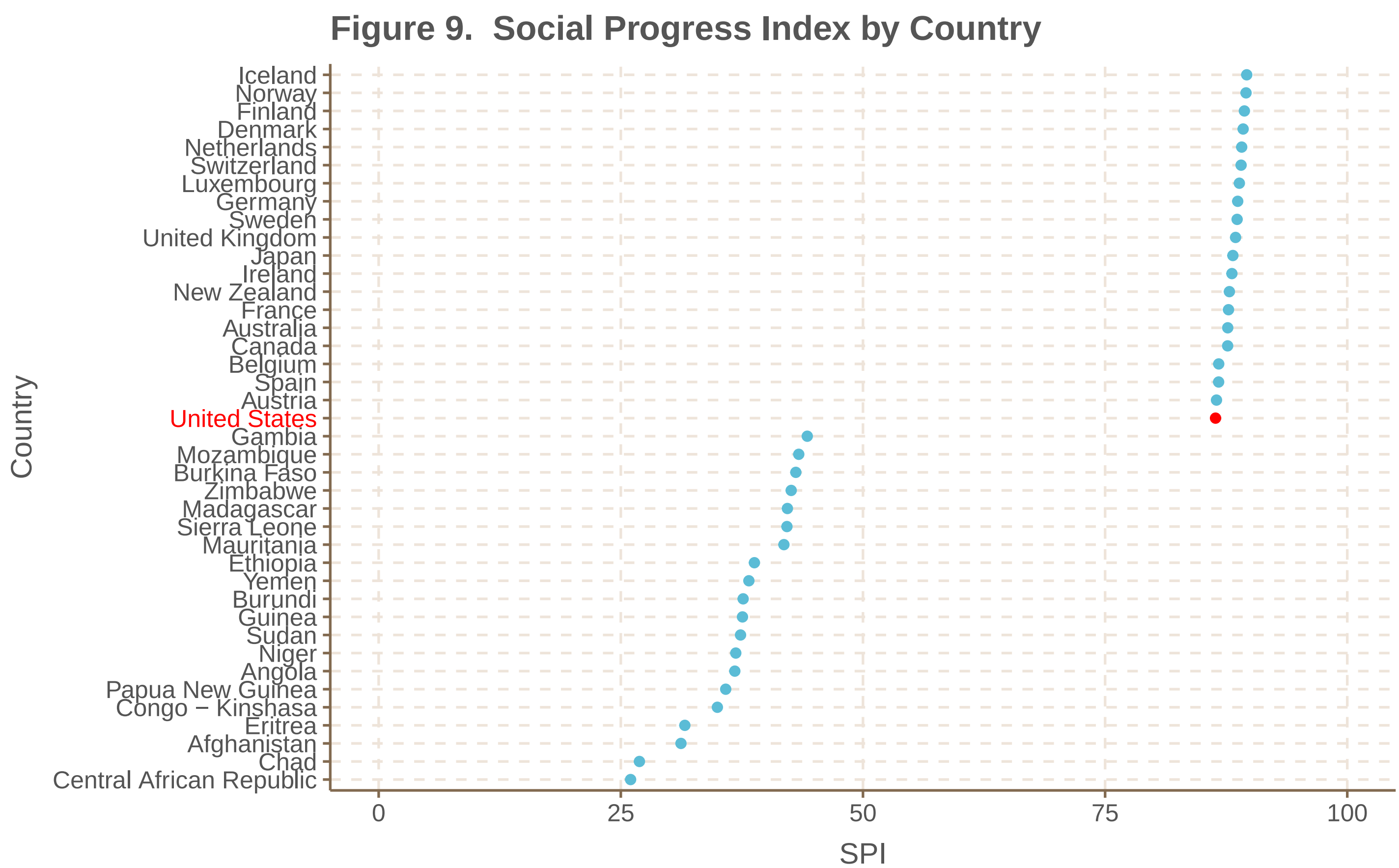
Strong differences were seen by `HDI_cat` for `infantmort`, `birth_MF`, and `sixty_MF`.

Figure 8. Correlation Matrix

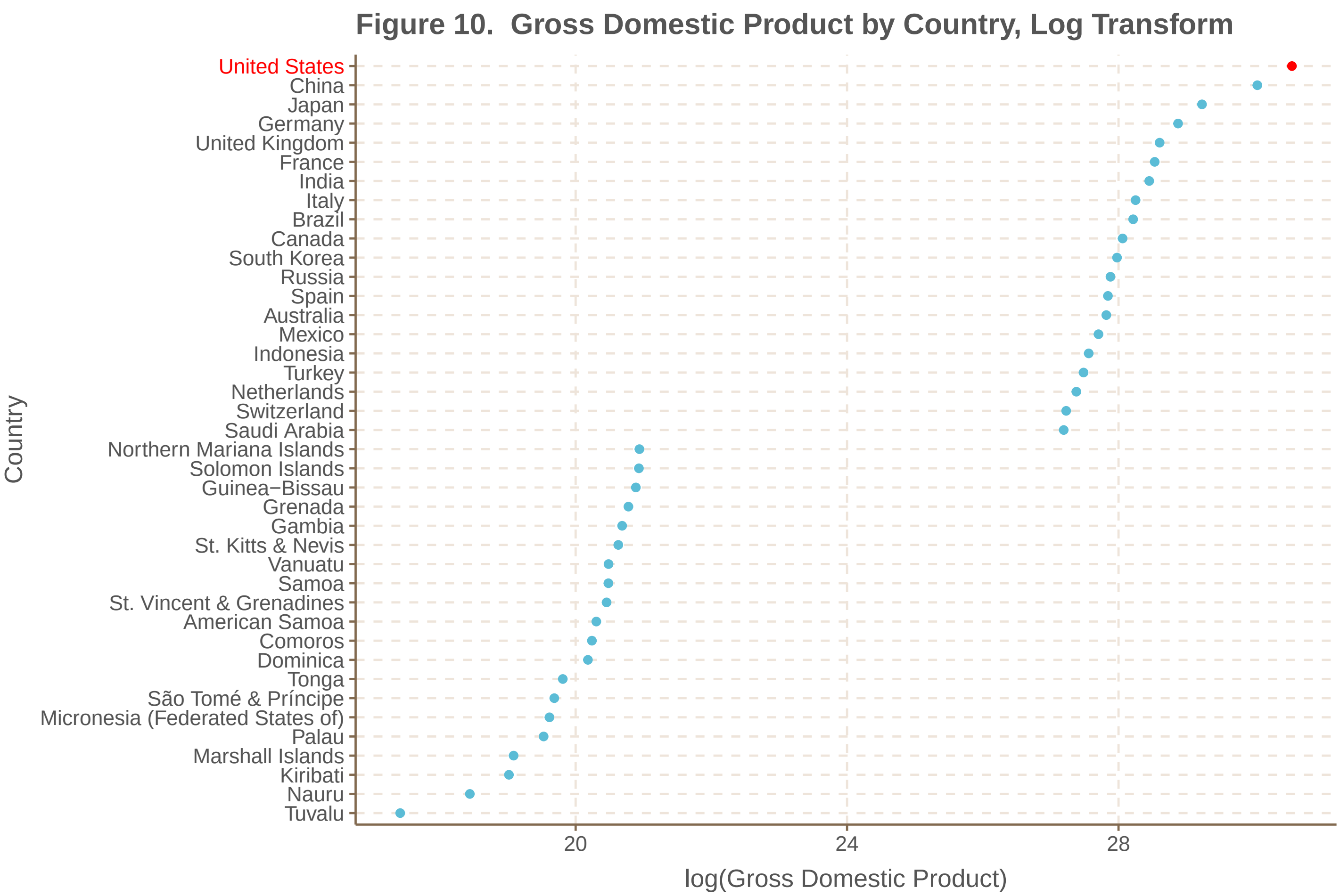


Top and Bottom 20 Countries by QoL Measure

When assessing the top and bottom 20 countries by Social Progress Index, the United States was found to rank twentieth (Fig. 9).

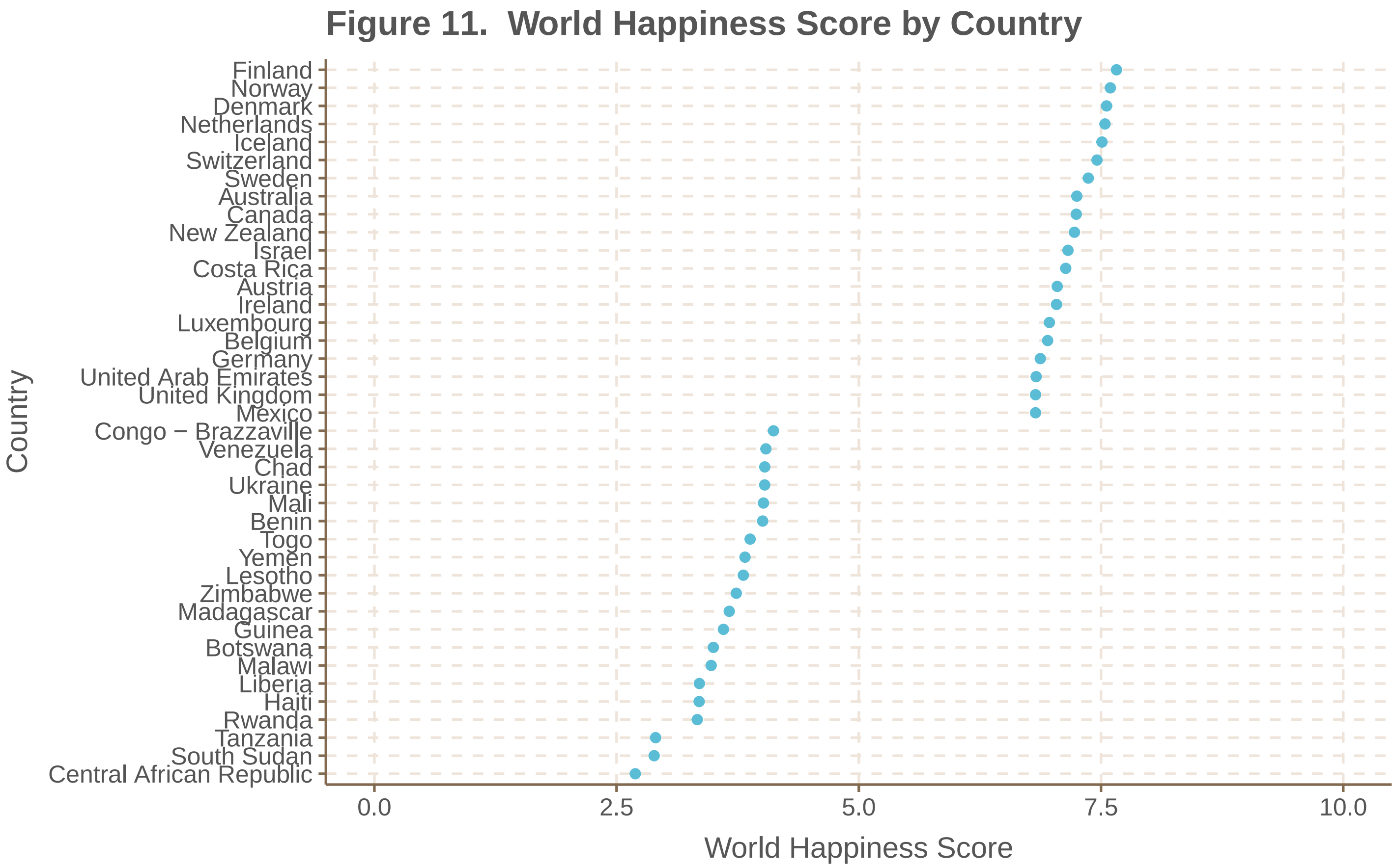


The United States has the world's largest GDP (Fig. 10).



The United States is not among the top 20 countries in terms of happiness; it ranks 21st (Fig. 11).

Figure 11. World Happiness Score by Country

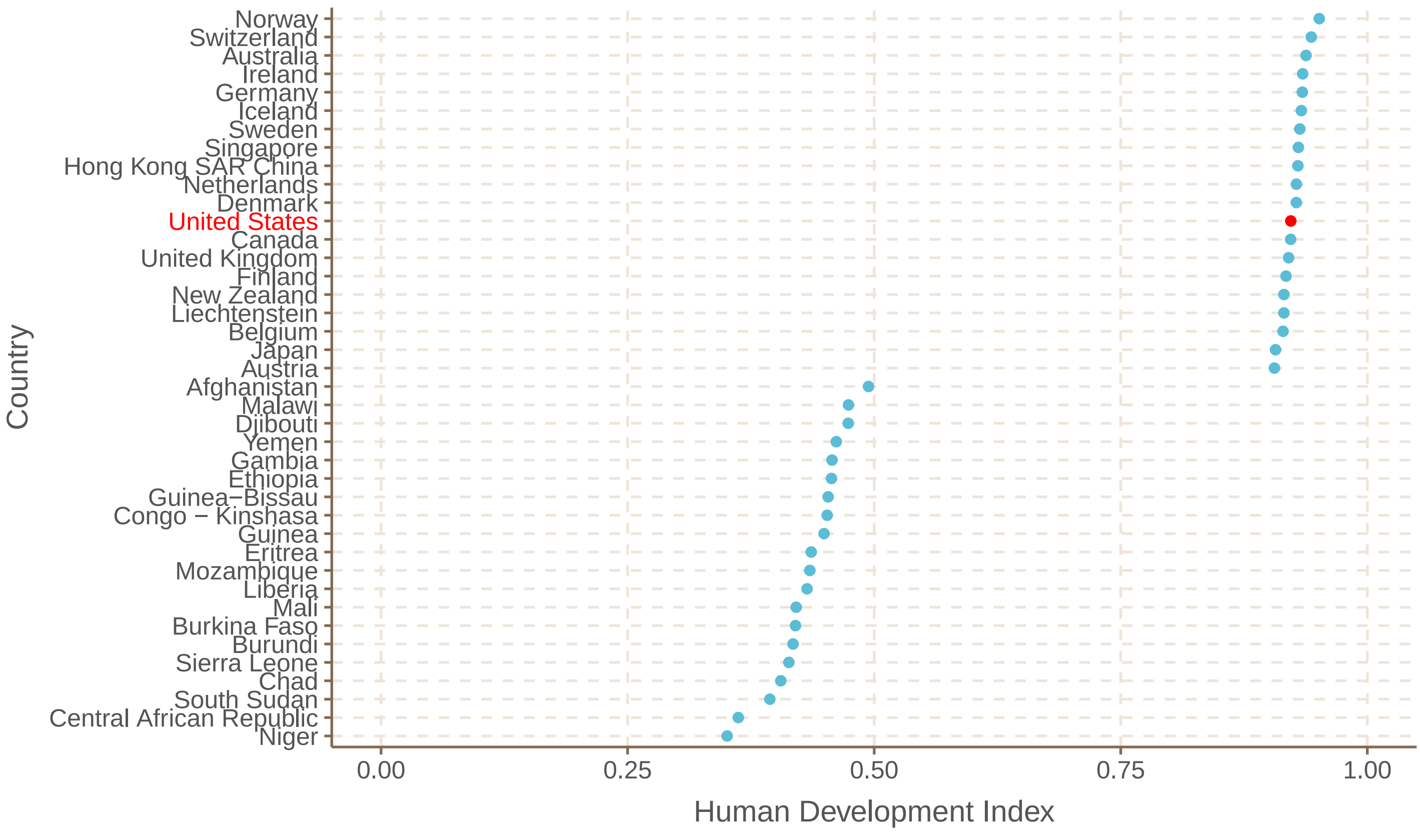



```r
which(alldata_WHR$country == "United States")
```

```
## [1] 21
```

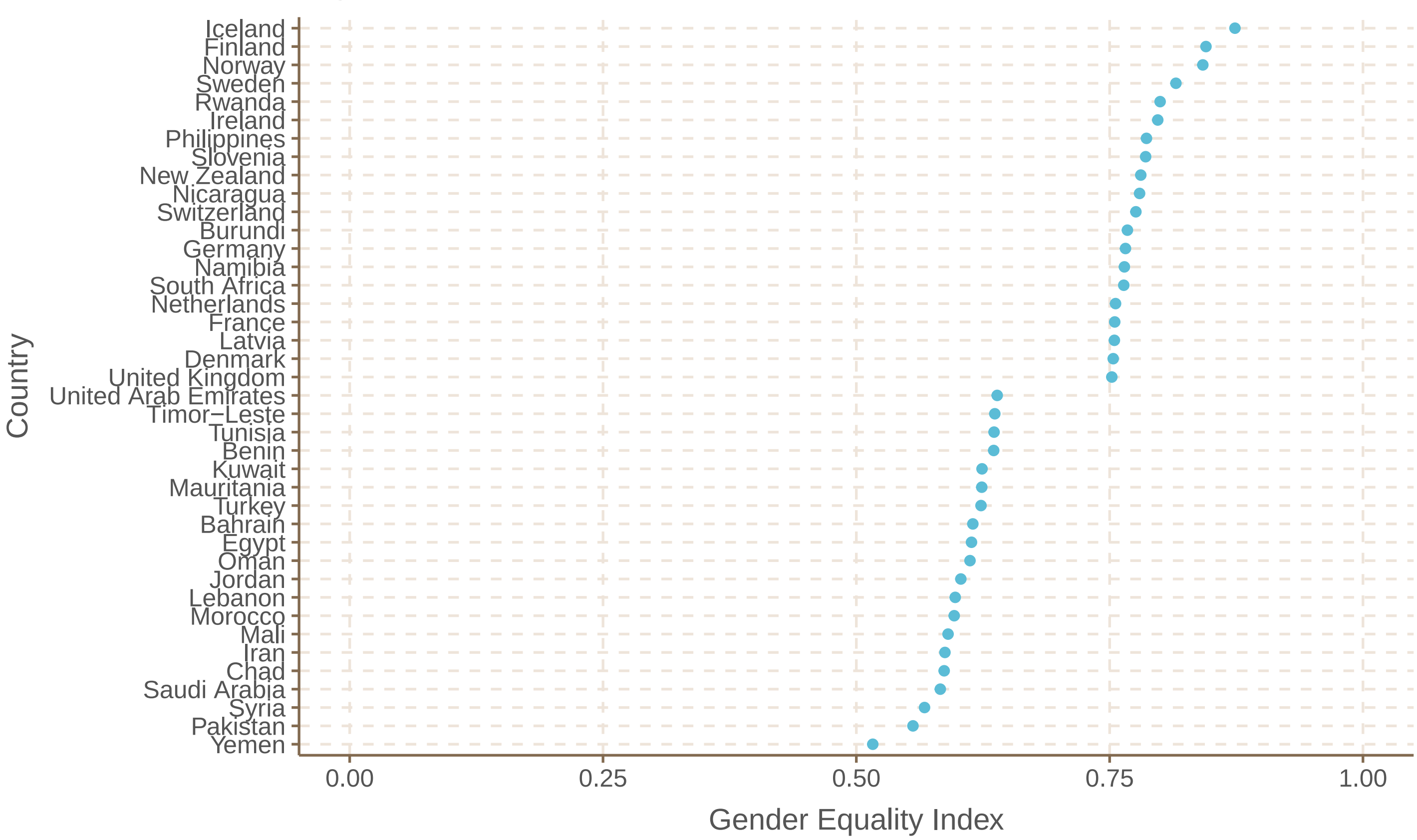
Per Fig. 12, the United States ranks twelfth by HDI.

Figure 12. Human Development Index by Country



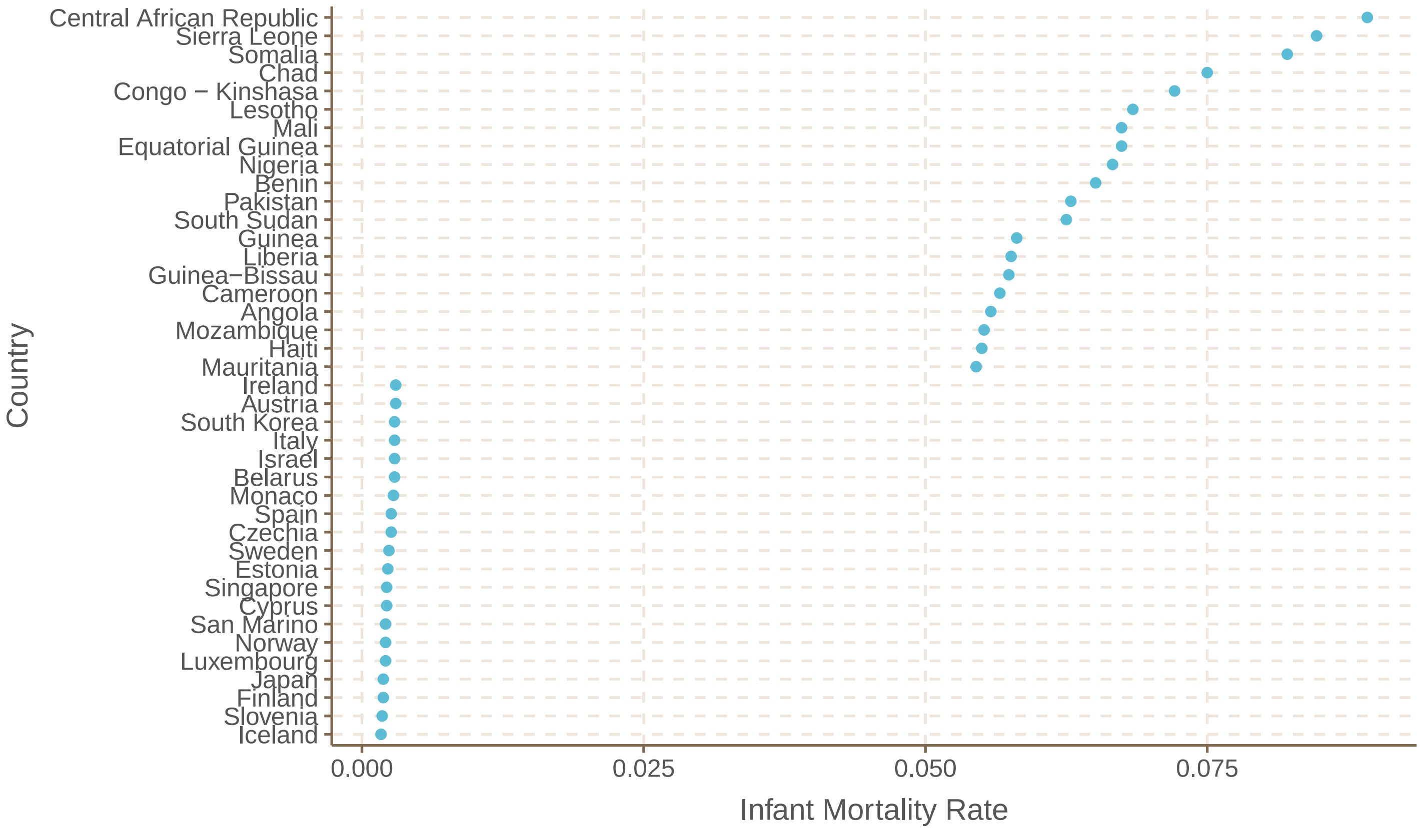
The United States is not among the top 20 countries in terms of gender equality; it ranks 45th (Fig. 13).

Figure 13. Gender Equality Index by Country



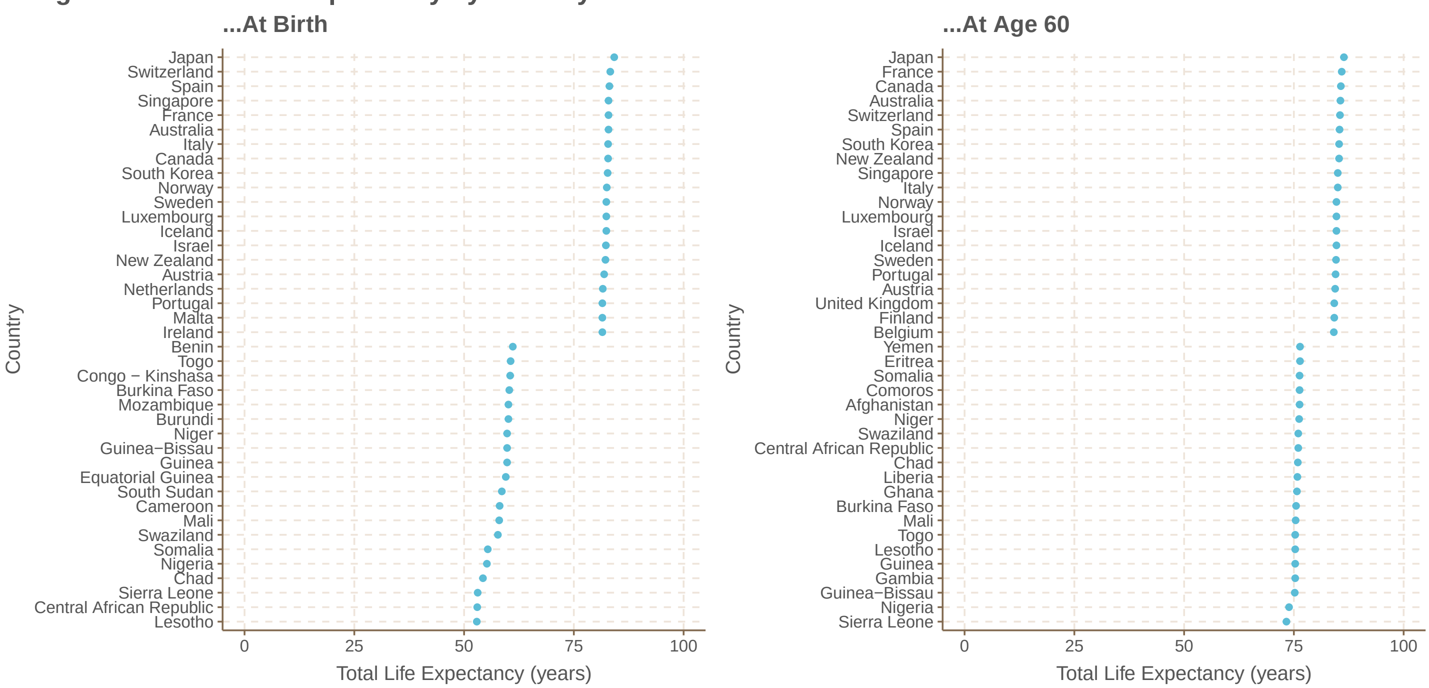
The United States has the world's 46th lowest infant mortality rate (Fig. 14).

Fig. 14. Infant Mortality Rate



Finally, per Fig. 15, the United States is not among the top 20 countries for life expectancy, ranking 34th and 31st respectively for life expectancy at birth and at 60 years of age.

Figure 15. Total Life Expectancy by Country...



Per the above country-level analyses, the United States was not among the top 20 performing countries for 5/8 (62.5%) of the variables assessed: happiness, gender equality, infant mortality, and life expectancy at birth and at 60 years. These five fields were examined via clustering.

Clustering Analysis

Ten pairwise cluster analyses (CAs) were performed assuming $k = 4$ clusters, shown in Table 4.

Table 4. Cluster Analyses Considered

Cluster Analysis	Variables
1	`happiness`, `gendereq`
2	`happiness`, `infantmort`
3	`happiness`, `birth_MF`
4	`happiness`, `sixty_MF`
5	`gendereq`, `infantmort`
6	`gendereq`, `birth_MF`
7	`gendereq`, `sixty_MF`
8	`infantmort`, `birth_MF`
9	`infantmort`, `sixty_MF`
10	`birth_MF`, `sixty_MF`

For Cluster Analysis #1 (Fig. 16) considering `happiness` and `gendereq`, the United States was clustered with other highly-developed nations; this remained stable when testing $k = 3$ and $k = 5$ clusters.

Figure 16. CA #1: Gender Equality Index vs. Happiness Score

Figure 17 depicts Cluster Analysis #2 for happiness score vs. infant mortality rate. The United States was again clustered with other highly-developed nations, and remained so when testing $k = 3$ and $k = 5$ clusters.

Figure 17. CA #2: Infant Mortality Rate vs. Happiness Score

Cluster Analyses 3 and 4 (Fig. 18) compared happiness score to both life expectancy variables. The US was clustered with other highly developed nations when considering total life expectancy at age 60, but for total life expectancy at birth, the US occupied a mixed cluster containing some countries at the "very high" development level, and some at the "high" level. For total life expectancy at age 60, the results were stable when testing $k = 3$ and $k = 5$ clusters; however, for life expectancy at birth, the US entered the cluster dominated by "very high"-level countries at $k = 3$.

Figure 18. CA #s 3 & 4: Happiness Score vs. Total Life Expectancy...

Figure 19 presents Cluster Analysis #5. The US occupied a mixed cluster including countries from all HDI categories, and this did not change with varying k.

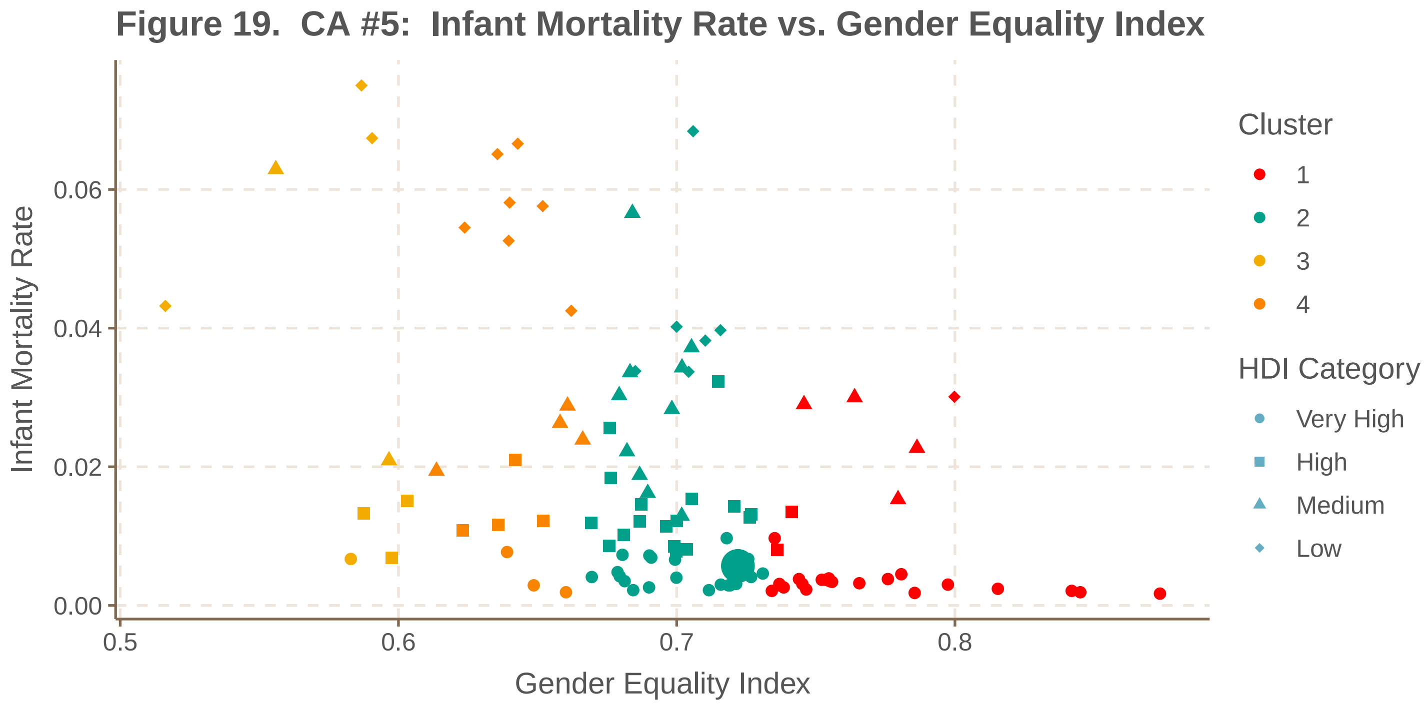


Figure 20 depicts Cluster Analyses 6 and 7, and Figure 21 shows Cluster Analyses 8 and 9. For both of these sets of analyses, the US was differentially clustered in the same manner as described for Cluster Analyses 3 and 4, and also had the same sensitivity to changing k as seen in those analyses. These results indicate that total life expectancy at birth is predictive of cluster number to some degree.

Figure 20. CA #s 6 & 7: Gender Equality Index vs. Total Life Expectancy...

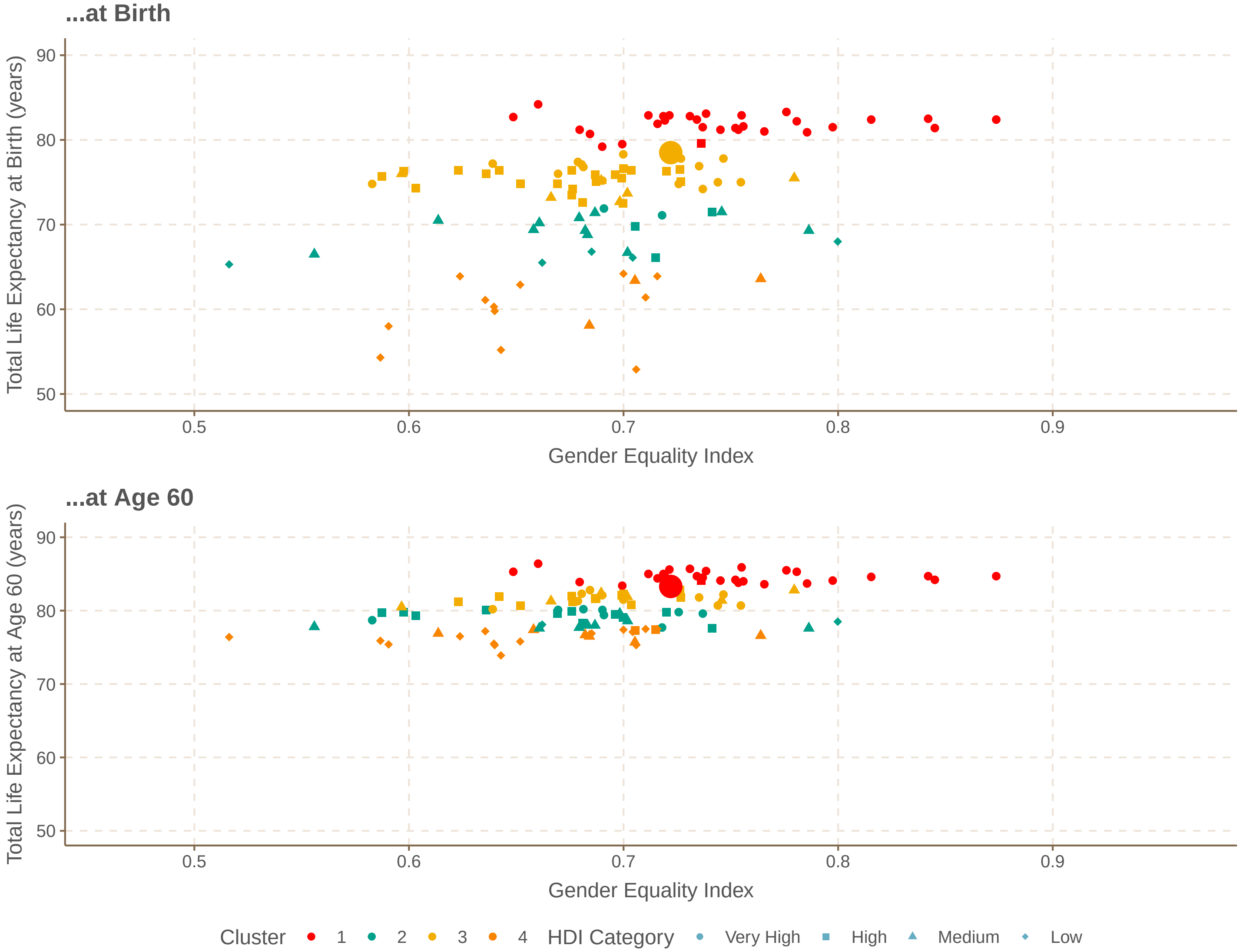
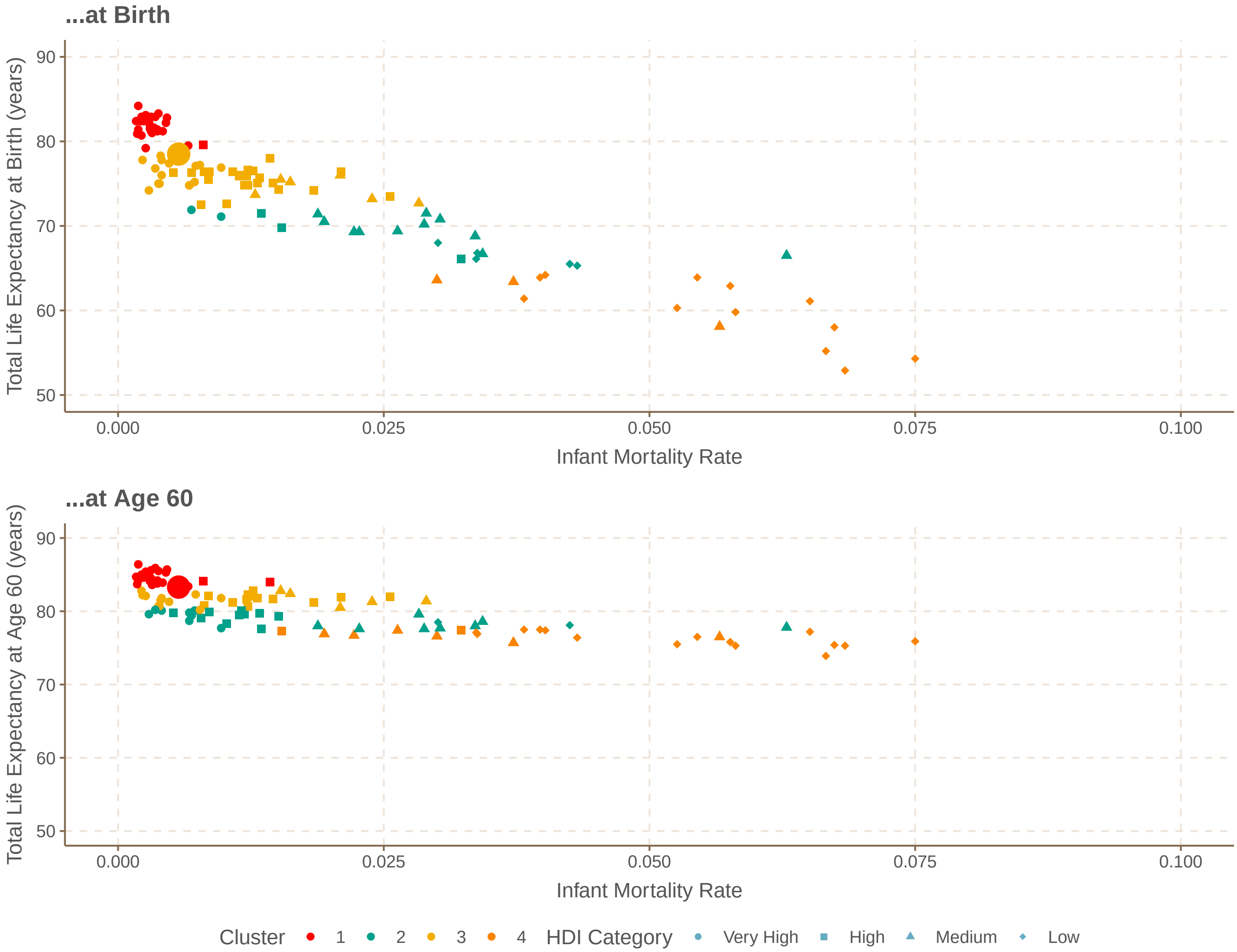


Figure 21. CA #s 8 & 9: Infant Mortality Rate vs. Total Life Expectancy...



Finally, Fig. 22 shows Cluster Analysis #10. The United States was again found in a mixed cluster, containing countries in the medium, high, and very high HDI categories. US cluster position remained unchanged for $k = 5$, but for $k = 3$, the US moved to the first cluster, containing mostly countries in the very high HDI category.

Figure 22. CA #10: Total Life Expectancy at Age 60 vs. at Birth

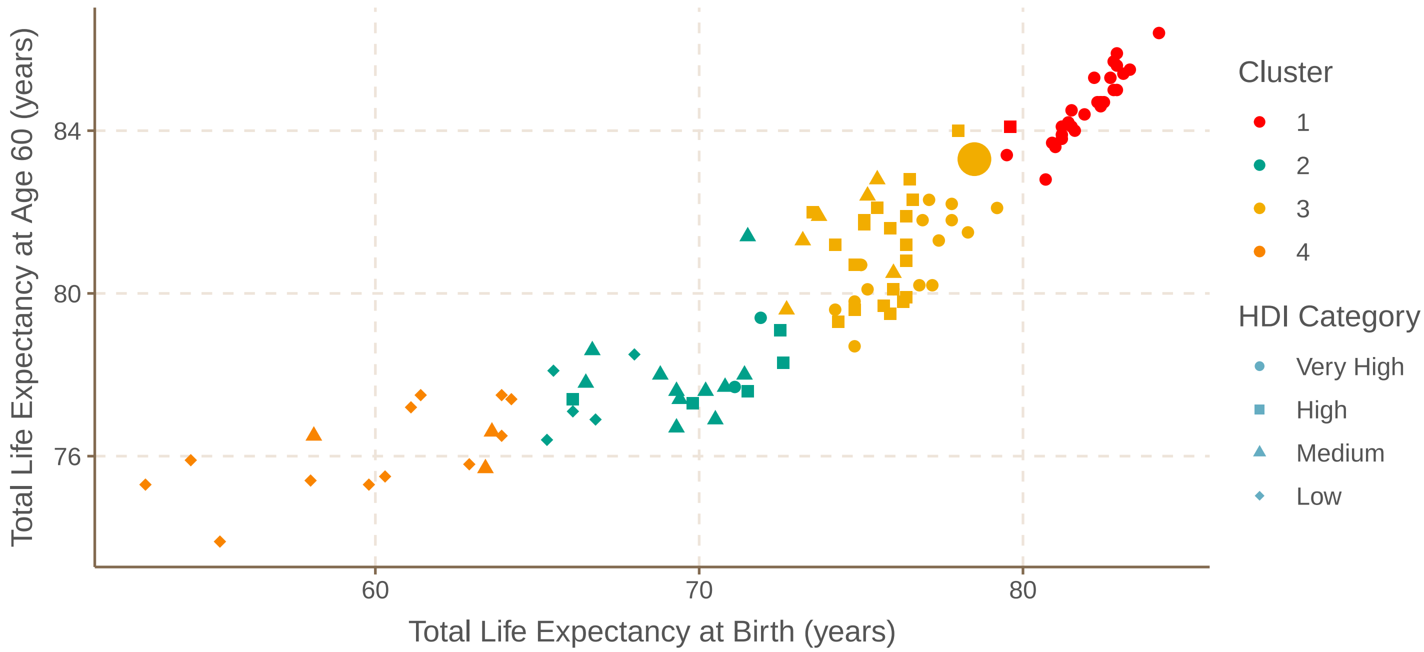


Table 5 characterizes the HDI category makeup of the cluster containing the US for all 10 cluster analyses. All CAs in which the US did not classify along with other countries in the "very high" HDI category involved the total life expectancy at birth variable, and these were sensitive to decreasing k.

Table 5. Characteristics of Clusters Containing US

Cluster Analysis	Base Case Classification	$k = 3$ Classification	$k = 5$ Classification
1	Primarily "very high"	No change	No change
2	Primarily "very high"	No change	No change
3	Mixed cluster: "high" and "very high"	Moved to primarily "very high" cluster	No change
4	Primarily "very high"	No change	No change
5	Primarily "very high"	No change	No change
6	Mixed cluster: "medium", "high", and "very high"	Moved to primarily "very high" cluster	No change
7	Primarily "very high"	No change	No change
8	Mixed cluster: "high" and "very high"	Moved to primarily "very high" cluster	No change
9	Primarily "very high"	No change	No change
10	Mixed cluster	Moved to primarily "very high" cluster	No change

Discussion

This study is useful in understanding some of the ways in which the United States lags behind other countries having a similar level of development. The US has the world's largest GDP, but this study shows that quality of life measures do not necessarily correspond to national wealth. Of the 8 continuous QoL variables considered, the United States did not appear within the top 20 performing countries for 5 (62.5%): happiness score, gender equality index, infant mortality rate, and the two total life expectancy variables. On clustering analysis, the US often clustered with countries of lower developmental level by HDI, particularly when considering total life expectancy at birth, and in some cases, these results remained stable when varying cluster number. These findings help pinpoint the ways in which the US underperforms.

Limitations

There are several limitations to this study. Missing data presented a challenge: countries with incomplete data were omitted for the clustering analysis, reducing the dataset from $n = 205$ to $n = 111$ countries. Several of the omitted countries were of low-to-medium developmental level, and omitting these may bias the results of the analysis. Second, several of these variables appear interrelated. For example, a strong correlation between `HDIindex` and `SPI` was seen in Fig. 8. These are compound measures calculated by different agencies, and there is some intersection in the subitems they each include. This correlation limits the usefulness of these two variables. Finally, limiting the data to 2016 prevents the ability to perform a longitudinal analysis. While countries in the "very high" HDI category are likely to be stable or slow-changing in QoL measures, developing countries or those with political instability may be more volatile, and these changes could be analyzed if data were present for even as short a time period as a decade.

Surprises and Challenges

Given that clustering considered the subset of variables for which the United States did not rank within the top 20 countries, the most surprising finding is that, for some of the cluster analyses, the US was classified with other countries in the "very high" HDI category, and these findings remained stable when varying k. Specific examples are gender equality index vs. happiness score (Fig. 16) and infant mortality rate vs. happiness score (Fig. 17). An additional surprise was that, for most of the clustering analyses, the clusters were largely stratified either vertically or horizontally on bivariate plotting, indicating that classification was driven by only one of the two variables considered. Cluster Analysis #7 was a weak exception to this pattern (Fig. 20).

The major technical challenge in this study was the data wrangling step: though not particularly difficult, it was time-consuming, and the manual country name recoding was tedious. In an attempt to partially automate matching on country name, fuzzy matching was attempted, but performance was poor, and I was unable to find a way to fuzzy-match by varying both string length and character distance. For example, matching the strings "Trinidad and Tobago" and "Trinidad & Tobago" would require matching with fuzzy string length (the first string is 2 characters longer than the second) whereas the strings "Cabo Verde" and "Cape Verde" would require matching with fuzzy character distance ("ca**bo**" vs. "ca**pe**"). The other challenge in this study is interpretive: the clustering visualizations contain the relevant results, but are not intuitive to interpret. The shapes given by the `shape` aesthetic are difficult to differentiate when using small shapes, but larger shapes obscure results due to overlapping, and it would not be appropriate to use jittering to avoid overlapping when visualizing cluster analyses.

Future Directions

Future work should include bringing additional years of data into the dataset and performing a longitudinal analysis. Additionally, since there is concern about both life expectancy and infant mortality rate in the US, relevant single-item QoL measures and epidemiological data should be incorporated to inform robust linear models aimed at predicting infant mortality rate and life expectancy. These analyses could be useful in informing policy decisions about public health and health education.

Conclusion

Though the United States has the world's largest economy, it underperforms on several key QoL measures. K-means clustering provides a powerful tool for investigating these deficits, and clustering results can help shape future research and public policy.

References

Helliwell, John F., Richard Layard, and Jeffrey D. Sachs. 2018. "World Happiness Report." http://worldhappiness.report/ed/2018/.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. "An Introduction to Statistical Learning." Springer. https://www-bcf.usc.edu/~gareth/ISL/.

Murphy, Sherry L., Jiaquan Xu, Kenneth D. Kochanek, and Elizabeth Arias. 2018. "Mortality in the United States, 2017.

NCHS Data Brief, No 328." National Center for Health Statistics. https://www.cdc.gov/nchs/products/databriefs/db328.htm.

Prioli, Katherine M. 2018. "MAT_8790_Final_Project." https://github.com/kmprioliPROF/MAT_8790_Final_Project.

Social Progress Imperative. 2018a. "Social Progress Index." https://www.socialprogress.org/?tab=4.

———. 2018b. "Social Progress Index - Overview." https://www.socialprogress.org/index.

The United Nations Development Programme. 2018a. "Human Development Index." http://hdr.undp.org/en/data.

———. 2018b. "Human Development Reports - Human Development Index (HDI)." http://hdr.undp.org/en/content/human-development-index-hdi.

The World Bank. 2018. "Gross Domestic Product." https://data.worldbank.org/indicator/ny.gdp.mktp.cd?view=map&year_high_desc=true.

World Economic Forum. 2016a. "Gender Equality." http://reports.weforum.org/global-gender-gap-report-2016/rankings/.

———. 2016b. "Measuring the Global Gender Gap." http://reports.weforum.org/global-gender-gap-report-2016/measuring-the-global-gender-gap/.

World Health Organization. 2018a. "Life Expectancy." http://apps.who.int/gho/data/view.main.SDG2016LEXv?lang=en.

———. 2018b. "Probability of Dying Per 1000 Live Births." http://apps.who.int/gho/data/view.main.182?lang=en.