```
hdi_index <- read.csv("Human Development Index and Components.csv")
```

```
dim(hdi_index) # Gives the number of rows and columns
```

```
195 · 10
```

```
row.names(hdi_index)
```

```
'1' · '2' · '3' · '4' · '5' · '6' · '7' · '8' · '9' · '10' · '11' · '12' · '13' · '14' · '15' · '16' · '17' · '18' · '19' · '20' · '21' ·
'22' · '23' · '24' · '25' · '26' · '27' · '28' · '29' · '30' · '31' · '32' · '33' · '34' · '35' · '36' · '37' · '38' · '39' · '40' ·
'41' · '42' · '43' · '44' · '45' · '46' · '47' · '48' · '49' · '50' · '51' · '52' · '53' · '54' · '55' · '56' · '57' · '58' · '59' ·
'60' · '61' · '62' · '63' · '64' · '65' · '66' · '67' · '68' · '69' · '70' · '71' · '72' · '73' · '74' · '75' · '76' · '77' · '78' ·
'79' · '80' · '81' · '82' · '83' · '84' · '85' · '86' · '87' · '88' · '89' · '90' · '91' · '92' · '93' · '94' · '95' · '96' · '97' ·
'98' · '99' · '100' · '101' · '102' · '103' · '104' · '105' · '106' · '107' · '108' · '109' · '110' · '111' · '112' · '113' ·
'114' · '115' · '116' · '117' · '118' · '119' · '120' · '121' · '122' · '123' · '124' · '125' · '126' · '127' · '128' · '129' ·
'130' · '131' · '132' · '133' · '134' · '135' · '136' · '137' · '138' · '139' · '140' · '141' · '142' · '143' · '144' ·
```

```
str(hdi_index)
```

```
'data.frame':   195 obs. of  10 variables:
 $ HDI.rank                              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Country                               : chr  "Switzerland" "Norway" "Iceland" "Hong Kong, China (SAR)" ...
 $ HUMAN_DEVELOPMENT                     : chr  "VERY HIGH " "VERY HIGH " "VERY HIGH " "VERY HIGH " ...
 $ Human_Development_Index..HDI.         : chr  "0.962" "0.961" "0.959" "0.952" ...
 $ Life_expectancy_at_birth              : num  84 83.2 82.7 85.5 84.5 81.4 83 82 80.6 81.7 ...
 $ Expected.years.of.schooling           : chr  "16.5" "18.2" "19.2" "17.3" ...
 $ Mean.years.of.schooling               : chr  "13.9" "13" "13.8" "12.2" ...
 $ Gross.national.income..GNI..per.capita: chr  "66,933" "64,660" "55,782" "62,607" ...
 $ GNI.per.capita.rank.minus.HDI.rank    : chr  "5" "6" "11" "6" ...
 $ HDI.rank.1                            : chr  "3" "1" "2" "4" ...
```

```
names(hdi_index)
```

```
'HDI.rank' · 'Country' · 'HUMAN_DEVELOPMENT' · 'Human_Development_Index..HDI.' · 'Life_expectancy_at_birth' · 'Expected.years.of.schooling' ·
'Mean.years.of.schooling' · 'Gross.national.income..GNI..per.capita' · 'GNI.per.capita.rank.minus.HDI.rank' · 'HDI.rank.1'
```

Double-click (or enter) to edit

```
hdi_index$Expected.years.of.schooling   <- as.numeric(hdi_index$Expected.years.of.schooling)
```

```
Warning message in eval(expr, envir, enclos):
“NAs introduced by coercion”
```

```
hdi_index$Gross.national.income..GNI..per.capita <- gsub(",", "", hdi_index$Gross.national.income..GNI..per.capita)
hdi_index$Gross.national.income..GNI..per.capita   <- as.integer(hdi_index$Gross.national.income..GNI..per.capita )
```

```
Warning message in eval(expr, envir, enclos):
“NAs introduced by coercion”
```

```
hdi_index$Human_Development_Index..HDI.  <- as.numeric(hdi_index$Human_Development_Index..HDI. )
(object=3)
```

```
Warning message in eval(expr, envir, enclos):
“NAs introduced by coercion”
3
```

```
hdi_index$Mean.years.of.schooling  <- as.numeric(hdi_index$Mean.years.of.schooling)
```

```
Warning message in eval(expr, envir, enclos):
“NAs introduced by coercion”
```

```
str(hdi_index)
```

```
'data.frame':   195 obs. of  10 variables:
 $ HDI.rank                              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Country                               : chr  "Switzerland" "Norway" "Iceland" "Hong Kong, China (SAR)" ...
 $ HUMAN_DEVELOPMENT                     : chr  "VERY HIGH " "VERY HIGH " "VERY HIGH " "VERY HIGH " ...
 $ Human_Development_Index..HDI.         : num  0.962 0.961 0.959 0.952 0.951 0.948 0.947 0.945 0.942 0.941 ...
 $ Life_expectancy_at_birth              : num  84 83.2 82.7 85.5 84.5 81.4 83 82 80.6 81.7 ...
 $ Expected.years.of.schooling           : num  16.5 18.2 19.2 17.3 21.1 18.7 19.4 18.9 17 18.7 ...
 $ Mean.years.of.schooling               : num  13.9 13 13.8 12.2 12.7 13 12.6 11.6 14.1 12.6 ...
 $ Gross.national.income..GNI..per.capita: int  66933 64660 55782 62607 49238 60365 54489 76169 54534 55979 ...
 $ GNI.per.capita.rank.minus.HDI.rank    : chr  "5" "6" "11" "6" ...
 $ HDI.rank.1                            : chr  "3" "1" "2" "4" ...
```

```
table(hdi_index$HUMAN_DEVELOPMENT)
```

```
          HIGH        LOW     MEDIUM      OTHER  VERY HIGH
            49         32         44          4         66
```

```
table(hdi_index$Mean.years.of.schooling)
```

```
      2.1  2.2  2.3  2.6  2.9    3  3.1  3.2  3.6  3.8  4.1  4.3  4.4  4.5  4.6  4.7
        2    1    1    1    1    1    1    3    1    1    1    2    1    2    2    1
      4.9    5  5.1  5.2  5.4  5.6  5.7  5.9    6  6.2  6.3  6.4  6.7    7  7.1  7.2
        2    1    6    2    3    2    4    2    1    3    1    2    2    1    3    4
      7.3  7.4  7.6  7.8  7.9    8  8.1  8.3  8.4  8.5  8.6  8.7  8.8  8.9    9  9.2
        2    2    2    1    1    1    3    1    1    1    3    4    3    2    3    3
      9.3  9.4  9.6  9.8  9.9   10 10.2 10.3 10.4 10.5 10.6 10.7 10.8 10.9   11 11.1
        2    2    2    2    3    1    1    2    2    3    5    1    3    3    1    3
     11.3 11.4 11.6 11.7 11.8 11.9 12.1 12.2 12.3 12.4 12.5 12.6 12.7 12.8 12.9   13
        6    7    3    1    1    2    1    5    2    2    4    3    2    3    4    3
     13.2 13.3 13.4 13.5 13.7 13.8 13.9 14.1
        1    2    2    2    1    2    1    1
```

```
any(is.na(hdi_index))
```

```
TRUE
```

```
sum(is.na(hdi_index))
```

```
12
```

```
hdi_index <- hdi_index[complete.cases(hdi_index), ]
```

```
sum(is.na(hdi_index))
```

```
0
```

```
dim(hdi_index)
```

```
191 · 10
```

Working with data.We are going to use following libraries

1. tidyr
2. dplyr
3. ggplot

```
hdi_index%>%
 filter(HDI.rank>50)%>%
  group_by(HUMAN_DEVELOPMENT)
```

A grouped_df: 141 ×

HDI.rank	Country	HUMAN_DEVELOPMENT	Human_Development_Index..HDI.	Life_expectancy_at_birth	Expected.years.of.schooling
<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
51	Brunei Darussalam	VERY HIGH	0.829	74.6	14.0
52	Russian Federation	VERY HIGH	0.822	69.4	15.8
53	Romania	VERY HIGH	0.821	74.2	14.2
54	Oman	VERY HIGH	0.816	72.5	14.6
55	Bahamas	VERY HIGH	0.812	71.6	12.9
56	Kazakhstan	VERY HIGH	0.811	69.4	15.8
57	Trinidad and Tobago	VERY HIGH	0.810	73.0	14.5
58	Costa Rica	VERY HIGH	0.809	77.0	16.5
58	Uruguay	VERY HIGH	0.809	75.4	16.8
60	Belarus	VERY HIGH	0.808	72.4	15.2
61	Panama	VERY HIGH	0.805	76.2	13.1
62	Malaysia	VERY HIGH	0.803	74.9	13.3
68	Grenada	HIGH	0.795	74.9	16.7

```
# Exploratory Data Analysis in R with ggplot2 and dplyr:
install.packages("ggplot2")
# install.packages("dplyr")
install.packages("ISLR")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
install.packages("tidyverse")  # Install the tidyverse package
library(tidyverse)  # Load the tidyverse package
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘textshaping’, ‘conflicted’, ‘jsonlite’, ‘ragg’


Warning message in install.packages("tidyverse"):
“installation of package ‘textshaping’ had non-zero exit status”
Warning message in install.packages("tidyverse"):
“installation of package ‘ragg’ had non-zero exit status”
Warning message in install.packages("tidyverse"):
“installation of package ‘tidyverse’ had non-zero exit status”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✓ ggplot2 3.4.2     ✓ purrr   1.0.1
✓ tibble  3.2.1     ✓ dplyr   1.1.2
✓ tidyr   1.3.0     ✓ stringr 1.5.0
✓ readr   2.1.4     ✓ forcats 1.0.0

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

```
install.packages("ggpubr")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘numDeriv’, ‘SparseM’, ‘MatrixModels’, ‘minqa’, ‘nloptr’, ‘RcppEigen’, ‘carData’, ‘abind’, ‘pbkrte
```

```
library(ggplot2) #data visualization
library(dplyr) #data manipulation
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:
```

```
    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

```
library(tidyr)
library(lubridate)
library(ggpubr)
```

```
Attaching package: ‘lubridate’

The following objects are masked from ‘package:base’:

    date, intersect, setdiff, union
```

```
min_life_expectancy <- min(hdi_index$Life_expectancy_at_birth)
max_life_expectancy <- max(hdi_index$Life_expectancy_at_birth)
```

```
cat("Lowest life expectancy at birth:", min_life_expectancy, "\n")
cat("Highest life expectancy at birth:", max_life_expectancy, "\n")
```

```
Lowest life expectancy at birth: 52.5
Highest life expectancy at birth: 85.9
```

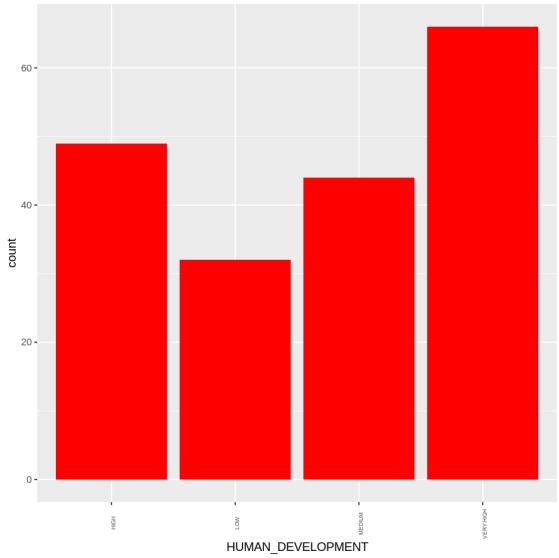
```
lowest_life_expectancy <- hdi_index[which.min(hdi_index$Life_expectancy_at_birth),"Country" ]
highest_life_expectancy <- hdi_index[which.max(hdi_index$Life_expectancy_at_birth), "Country"]
```

```
print(paste("Country with the lowest life expectancy:", lowest_life_expectancy,",that is: ",min_life_expectancy))
print(paste("Country with the highest life expectancy:", highest_life_expectancy,",that is: ",max_life_expectancy))
```

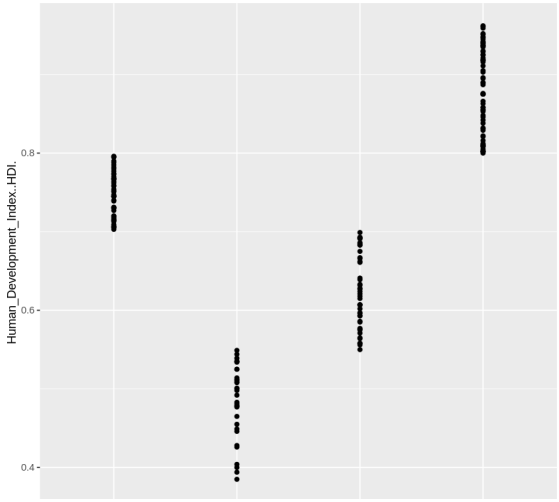
```
[1] "Country with the lowest life expectancy: Chad ,that is:  52.5"
[1] "Country with the highest life expectancy: Monaco ,that is:  85.9"
```

```
hdi_index = select(hdi_index,HDI.rank,Country,HUMAN_DEVELOPMENT, Human_Development_Index..HDI.  , Life_expectancy_at_birth, Expected.year
```

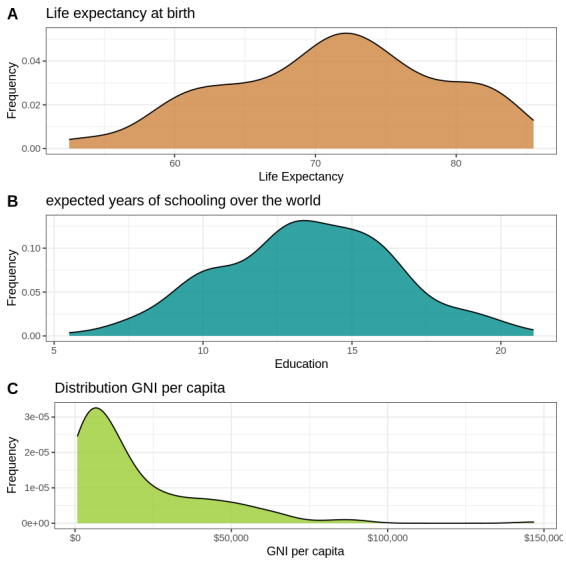
```
ggplot(hdi_index,aes(x = HUMAN_DEVELOPMENT))+
geom_bar(fill= "red" )+
theme(axis.text.x = element_text(angle = 90, size = 5))
```




```
ggplot(hdi_index,aes(x = HUMAN_DEVELOPMENT, y = Human_Development_Index..HDI.))+
geom_point()+
theme(axis.text.x = element_text(angle = 90, size = 5))
```

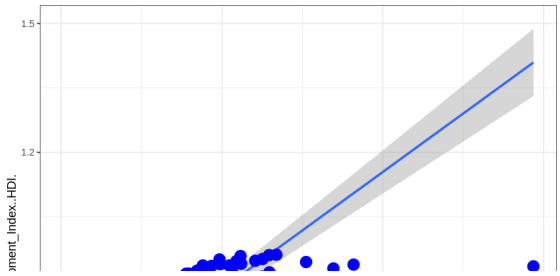



```
# Apply bw theme
theme_set(theme_bw())
#1 Distribution of quantitative variables
g1 <- ggplot(hdi_index, aes(x = Life_expectancy_at_birth)) +
    geom_density(aes(fill=Life_expectancy_at_birth), alpha=0.8,fill = "tan3") +
    labs(x = "Life Expectancy", y = "Frequency",
        title = "Life expectancy at birth")

g2 <- ggplot(hdi_index, aes(x = Expected.years.of.schooling)) +
    geom_density(aes(fill=Expected.years.of.schooling), alpha=0.8,fill = "cyan4") +
    labs(x = "Education", y = "Frequency",
        title = "expected years of schooling over the world")

g3 <- ggplot(hdi_index, aes(x = Gross.national.income..GNI..per.capita)) +
    geom_density(aes(fill=Gross.national.income..GNI..per.capita), alpha=0.8,fill = "yellowgreen") +
    labs(x = "GNI per capita", y = "Frequency",
        title = "Distribution GNI per capita ") + scale_x_continuous(label = scales::dollar)

ggarrange(g1, g2, g3, labels = c("A", "B", "C"), ncol = 1, nrow = 3)
```



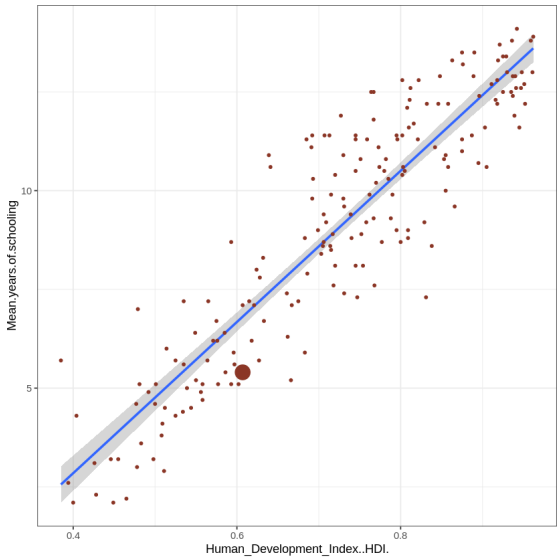

```
#2 Correlation
g <- ggplot(hdi_index, aes(x=Gross.national.income..GNI..per.capita, y=Human_Development_Index..HDI.)) +
    geom_smooth(method="lm") +geom_count(col=" blue", show.legend=F)
plot(g)
```

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



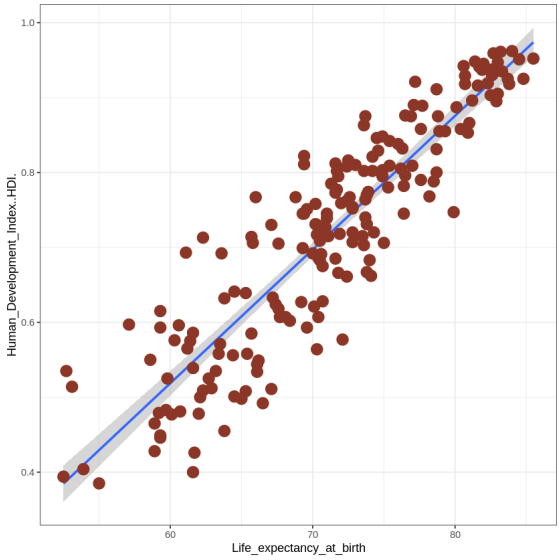
```
#2 Correlation
g <- ggplot(hdi_index, aes(x=Human_Development_Index..HDI., y=Mean.years.of.schooling)) +
    geom_smooth(method="lm") +geom_count(col="tomato4", show.legend=F)
plot(g)
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 4 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 4 rows containing non-finite values (`stat_sum()`).”
```



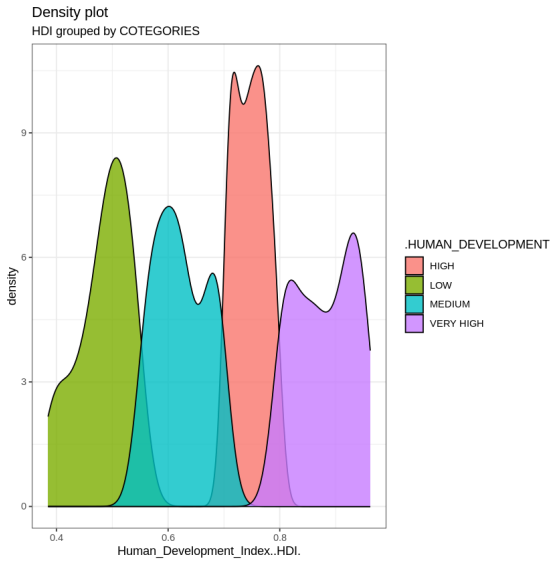

```
#2 Correlation
g <- ggplot(hdi_index, aes(x=Life_expectancy_at_birth, y=Human_Development_Index..HDI.)) +
    geom_smooth(method="lm") +geom_count(col="tomato4", show.legend=F)
plot(g)
```

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




```
#density plot
g1 <- ggplot(hdi_index, aes(Human_Development_Index..HDI.)) +
    geom_density(aes(fill=factor(HUMAN_DEVELOPMENT)), alpha=0.8) +
    labs(title="Density plot", subtitle="HDI grouped by COTEGORIES",
        x="Human_Development_Index..HDI.", fill=".HUMAN_DEVELOPMENT")
plot(g1)
```

```
Warning message:
“Removed 4 rows containing non-finite values (`stat_density()`).”
```

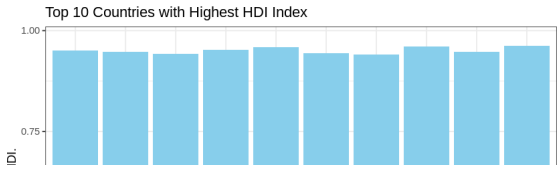



```
sorted_data <- hdi_index %>% arrange(desc(Human_Development_Index..HDI.))

top_10_countries <- head(sorted_data, 10)

top_10_countries
```

A data.frame: 10 × 8

	HDI.rank	Country	HUMAN_DEVELOPMENT	Human_Development_Index..HDI.	Life_expectancy_at_birth	Expected.years.of.schooling
	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	1	Switzerland	VERY HIGH	0.962	84.0	16.5
2	2	Norway	VERY HIGH	0.961	83.2	18.2
3	3	Iceland	VERY HIGH	0.959	82.7	19.2
4	4	Hong Kong, China (SAR)	VERY HIGH	0.952	85.5	17.3
5	5	Australia	VERY HIGH	0.951	84.5	21.1
6	6	Denmark	VERY HIGH	0.948	81.4	18.7
7	7	Sweden	VERY HIGH	0.947	83.0	19.4
8	8	Ireland	VERY HIGH	0.945	82.0	18.9
9	9	Germany	VERY HIGH	0.942	80.6	17.0
10	10	Netherlands	VERY HIGH	0.941	81.7	18.7

```
# Create the ggplot bar plot
ggplot(top_10_countries, aes(x = Country, y = Human_Development_Index..HDI.)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(x = "Country", y = "Human_Development_Index..HDI.") +
  ggtitle("Top 10 Countries with Highest HDI Index") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
head(hdi_index,2)
```

A data.frame: 2 × 10

	HDI.rank	Country	HUMAN_DEVELOPMENT	Human_Development_Index..HDI.	Life_expectancy_at_birth	Expected.years.of.schooling
	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	1	Switzerland	VERY HIGH	0.962	84.0	16.5
2	2	Norway	VERY HIGH	0.961	83.2	18.2

```
first_ranked_hdi <- hdi_index[hdi_index$HDI.rank == 1, "Human_Development_Index..HDI."]
```

```
India_hdi <- hdi_index[hdi_index$Country == "India", "Human_Development_Index..HDI."]
```

```
India_rank <- hdi_index[hdi_index$Country == "India", "HDI.rank"]
```

```
India_hdi
```

```
0.633
```

```
cat("India's Rank in HDI",India_rank)
```

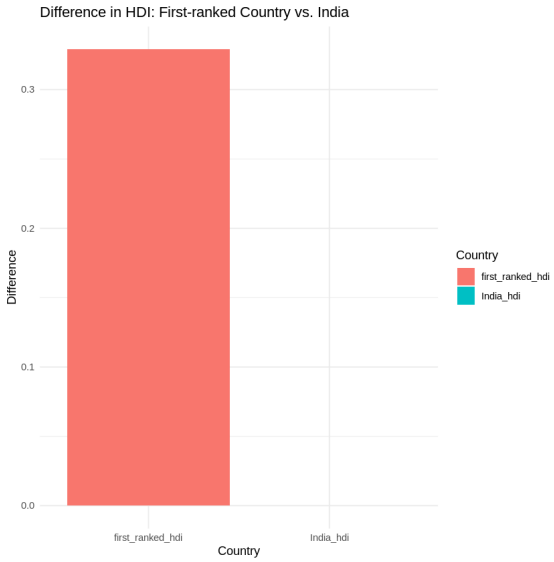
```
India's Rank in HDI 132
```

```
difference <- first_ranked_hdi - India_hdi
```

```
print(difference)
```

```
[1] 0.329
```

```
df <- data.frame(Country = c("first_ranked_hdi", "India_hdi"), Difference = c(difference, 0))
ggplot(df, aes(x = Country, y = Difference, fill = Country)) +
  geom_bar(stat = "identity") +
  labs(title = "Difference in HDI: First-ranked Country vs. India",
       x = "Country", y = "Difference") +
  theme_minimal()
```




```
countries <- c("India", "Pakistan", "Bangladesh", "Nepal", "Bhutan", "Myanmar", "China")
filtered_data <- hdi_index[hdi_index$Country %in% countries, ]
```

```
filtered_data
```

A data.frame: 7 × 1

	HDI.rank	Country	HUMAN_DEVELOPMENT	Human_Development_Index..HDI.	Life_expectancy_at_birth	Expected.years.of.schoolin
	<int>	<chr>	<chr>	<dbl>	<dbl>	<dbl
79	79	China	HIGH	0.768	78.2	14.
127	127	Bhutan	MEDIUM	0.666	71.8	13.
129	129	Bangladesh	MEDIUM	0.661	72.4	12.
132	132	India	MEDIUM	0.633	67.2	11.
143	143	Nepal	MEDIUM	0.602	68.4	12.
149	149	Myanmar	MEDIUM	0.585	65.7	10.
161	161	Pakistan	LOW	0.544	66.1	8.

```
ggplot(filtered_data, aes(x = Country, y = Human_Development_Index..HDI.)) +
  geom_bar(stat = "identity", fill = "pink") +
  geom_text(aes(label = HDI.rank    ), vjust = -0.5, size = 3, color = "black") +
  labs(title = "HDI of India and Neighboring Countries with Rank",
       x = "Country", y = "HDI")
```