DA5020.A7.Hsiao-Yu.Peng

Hsiao-Yu Peng

2023-10-27

Q1

```
url <- "https://en.wikipedia.org/w/index.php?title=List_of_sovereign_states_by_percentage_of_population_
website <- read_html(url)
data_table <- html_table(website)[3]
df <- data.frame(data_table)
colnames(df) = c("Country", "less_1.15", "less_2.6", "less_3.85", "Year", "Continent")

df$less_1.15 <- as.array(df$less_1.15)
df$less_1.15 <- substring(df$less_1.15, 1, nchar(df$less_1.15)-1)
df$less_1.15 <- as.numeric(df$less_1.15)

df$less_2.6 <- as.array(df$less_2.6)
df$less_2.6 <- substring(df$less_2.6, 1, nchar(df$less_2.6)-1)
df$less_2.6 <- as.numeric(df$less_2.6)

df$less_3.85 <- as.array(df$less_3.85)
df$less_3.85 <- substring(df$less_3.85, 1, nchar(df$less_3.85)-1)
df$less_3.85 <- as.numeric(df$less_3.85)

head(df)
```

```
##     Country less_1.15 less_2.6 less_3.85 Year     Continent
## 1   Albania     13.30    26.60     38.40 2023        Europe
## 2   Algeria      0.32     2.23     20.83 2019        Africa
## 3    Angola     51.40    72.79     89.13 2019        Africa
## 4 Argentina      1.60     5.80     18.20 2020 South America
## 5   Armenia      0.40     6.90     44.70 2020          Asia
## 6 Australia      0.50     0.74      0.74 2019       Oceania
```

```
dim(df)
```

```
## [1] 166   6
```

```
sum(is.na(df))
```

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

The data frame has dimension 166 x6. We've convert the percentage of population have an income of less than 1.15, 2.6, and 3.85 into numerical type. And there is no missing value in this data frame.

Q2

```
less_than_385 <- df %>%
  select(Continent, less_3.85) %>%
  group_by(Continent) %>%
  summarise(mean(less_3.85), sd(less_3.85))

less_than_385
```

```
## # A tibble: 7 x 3
##   Continent     ‘mean(less_3.85)‘ ‘sd(less_3.85)‘
##   <chr>                     <dbl>           <dbl>
## 1 Africa                    74.3            25.4
## 2 Asia                      33.8            30.2
## 3 Asia, Europe               6.74            4.90
## 4 Europe                     5.03            9.22
## 5 North America             28.5            20.7
## 6 Oceania                   49.2            27.4
## 7 South America             21.3            12.8
```

In Africa, an average of 74.28% of the population lives on less than $3.85 per day, with significant variability (standard deviation of 25.36%). This indicates diverse living conditions across the continent, with some regions having a higher percentage in this income category. Asia has a similar wide range, with an average of 74.28% living on less than $3.85 per day, also with a standard deviation of 30.16%. Oceania has an average of 49.20% in this income category, with significant regional variations (standard deviation of 27.43%).

A combination of Asia and Europe results in an average of 6.74% of the population living on less than $3.85 per day, with relatively less variation (standard deviation of 4.90%). Europe alone has an average of 5.03% in this category, but a higher standard deviation (9.22%) indicates significant regional disparities.

South America averages approximately 21.27% in the under $3.85 category, with a lower standard deviation (12.76%), suggesting less regional variation. In North America, the average is around 28.55% in the under $3.85 category, with moderate variation (standard deviation of 20.73%) across different regions.

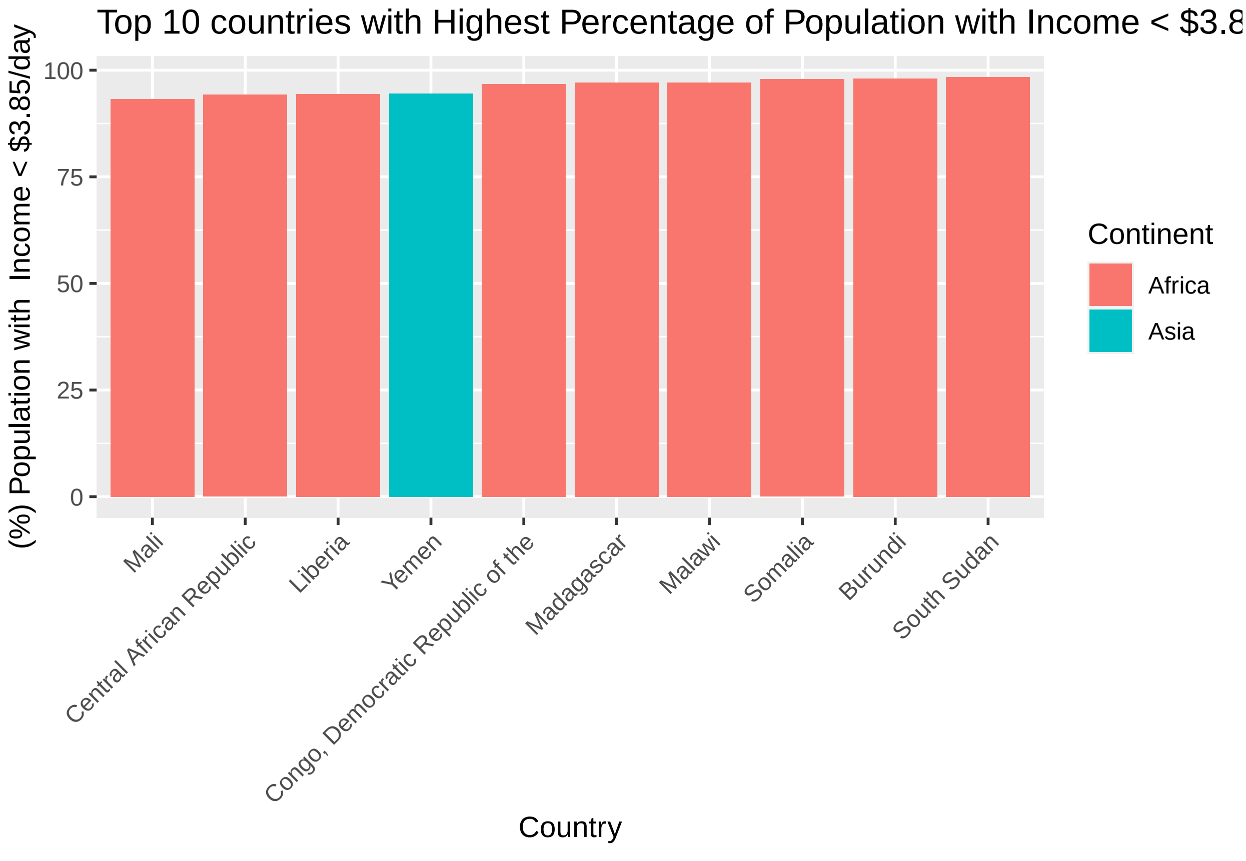
In summary, Africa and Asia have high average percentages in the income category below $3.85, with significant variability. Oceania also exhibits regional disparities. Europe and Asia/Europe have lower averages with varying regional conditions, while North America and South America fall in between. These statistics provide insights into economic disparities and living conditions across continents.

Q3

```
# Select top 10 countries
top10_countries <- df %>%
  select(Country, less_3.85, Continent) %>%
  arrange(desc(less_3.85)) %>%
  head(10)

top10_countries
```

```
##                                Country less_3.85 Continent
## 1                        South Sudan     98.44    Africa
## 2                            Burundi     98.02    Africa
## 3                            Somalia     97.87    Africa
## 4                             Malawi     97.10    Africa
## 5                         Madagascar     97.09    Africa
## 6  Congo, Democratic Republic of the     96.78    Africa
## 7                              Yemen     94.55      Asia
## 8                            Liberia     94.46    Africa
## 9           Central African Republic     94.26    Africa
## 10                              Mali     93.29    Africa
```

```
# Create the plot
ggplot(top10_countries, aes(x = reorder(Country, less_3.85), y = less_3.85, fill = Continent)) +
  geom_bar(stat='identity') +
  labs(title= 'Top 10 countries with Highest Percentage of Population with Income < $3.85 per day', x =
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The majority of the top 10 countries with the highest percentage of their population living on less than `$3.85` per day are in Africa, with one country located in Asia. In these countries, over 90% of the population has an income of less than `$3.85` per day.

Q4.

```r
# Select top 10 countries less than $3.85 per day
top5_countries_385 <- df %>%
  select(Country, less_3.85, Continent) %>%
  arrange(less_3.85) %>%
  head(5)

top5_countries_385
```

```
##                Country less_3.85 Continent
## 1           Azerbaijan      0.00      Asia
## 2 United Arab Emirates      0.00      Asia
## 3              Iceland      0.04    Europe
## 4              Belarus      0.10    Europe
## 5              Finland      0.10    Europe
```

```r
# Select top 10 countries less than $2.60 per day
top5_countries_260 <- df %>%
  select(Country, less_2.6, Continent) %>%
  arrange(less_2.6) %>%
  head(5)

top5_countries_260
```

```
##          Country less_2.6 Continent
## 1     Azerbaijan        0      Asia
## 2        Belarus        0    Europe
## 3 Czech Republic        0    Europe
## 4        Iceland        0    Europe
## 5       Maldives        0      Asia
```

```r
# Select top 10 countries less than $1.15 per day
top5_countries_115 <- df %>%
  select(Country, less_1.15, Continent) %>%
  arrange(less_1.15) %>%
  head(5)

top5_countries_115
```

```
##          Country less_1.15 Continent
## 1     Azerbaijan         0      Asia
## 2        Belarus         0    Europe
## 3 Czech Republic         0    Europe
## 4        Finland         0    Europe
## 5        Germany         0    Europe
```

```r
merged_data <- merge(merge(top5_countries_385, top5_countries_260, by = c("Country", "Continent"), all =
merged_data
```

```
##                Country Continent less_3.85 less_2.6 less_1.15
## 1           Azerbaijan      Asia      0.00        0         0
## 2               Belarus    Europe      0.10        0          0
## 3        Czech Republic    Europe        NA        0          0
## 4               Finland    Europe      0.10       NA          0
## 5               Germany    Europe        NA       NA          0
## 6               Iceland    Europe      0.04        0         NA
## 7              Maldives      Asia        NA        0         NA
## 8 United Arab Emirates      Asia      0.00       NA         NA
```

In general, the merged table reveals that these countries, primarily in Europe and Asia, report very low percentages of their populations living within these income categories. For all three income thresholds, Azerbaijan (Asia) and Belarus (Europe) have reported the lowest percentages (0.00% or 0.10%). Czech Republic reports 0% for both the `$2.60` and `$1.15` thresholds. Finland reported 0.10% for the `$3.85` income category and 0% for the `$1.15` category. Germany reports 0% for the `$1.15` threshold. Iceland has reported low percentages, 0.04% and 0.00%, respectively, for the `$3.85` and `$2.60` categories. Maldives reports a low percentage (0.00%) for the `$2.60` threshold. The United Arab Emirates reports the lowest percentage (0.00%) for the $3.85 threshold

Q5

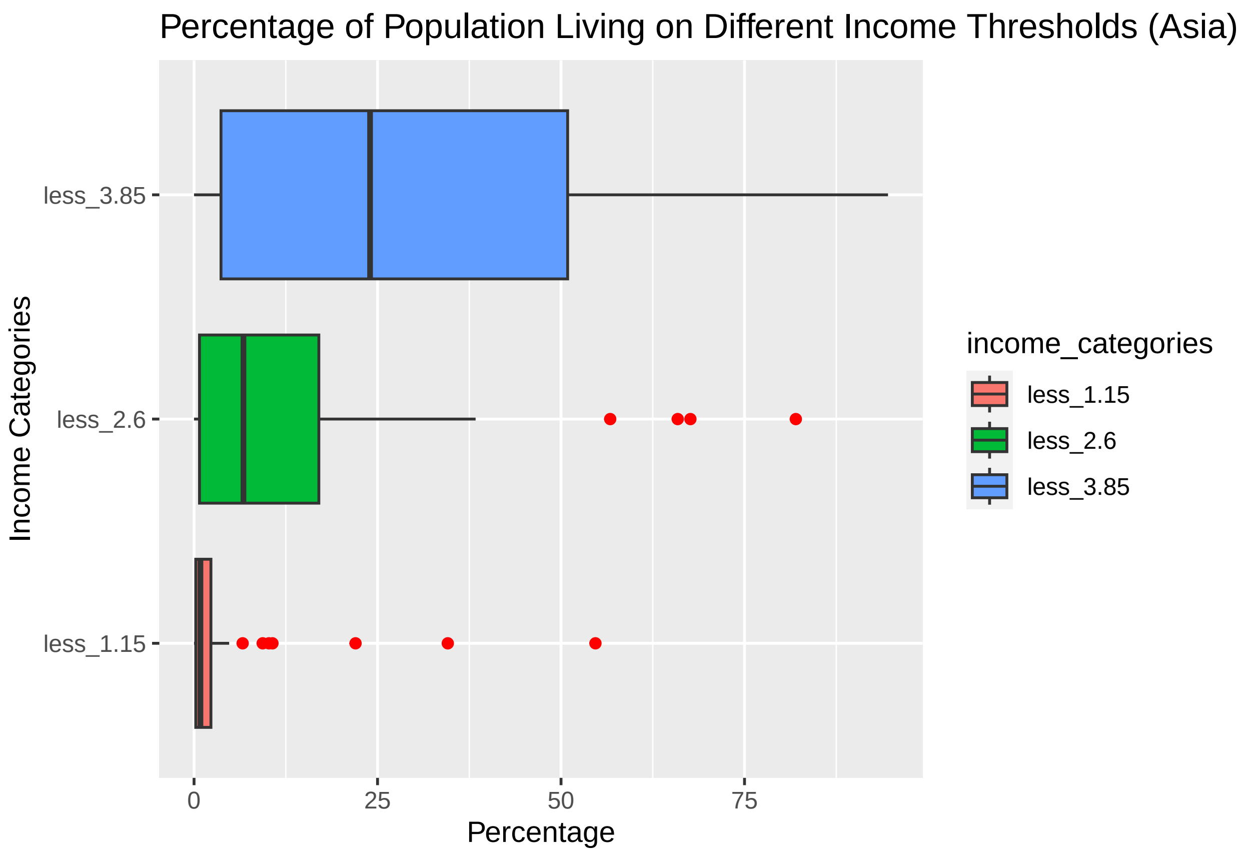
```
# Select Asia data
asia_data <- df %>%
  dplyr::filter(Continent == "Asia") %>%
  select(less_1.15, less_2.6, less_3.85)

# melting asia_data, convert column name becomes x, its values becomes y
require(reshape2)
```

```
## Loading required package: reshape2

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
asia_data <- melt(asia_data, value.name = "percentage", variable.name = "income_categories")
```

```
## No id variables; using all as measure variables
```

```
# Create Asia box plot
ggplot(data = asia_data, aes(x=income_categories, y=percentage)) +
         geom_boxplot(aes(fill=income_categories), outlier.colour="red") +
        coord_flip() +
         labs(title = "Percentage of Population Living on Different Income Thresholds (Asia)",
       x = "Income Categories", y = "Percentage")
```



In the Asia box plot, the median and the percentage of the population living on less than $1.15 are relatively low, ranging from approximately 0% to 5%. However, the presence of outlier points extends the range to between 25% and 50%.

In the less_2.6 threshold category, the median is around 10%, and there is a slight right skew in the distribution. The spread of values ranges from 0% to 20%, but it's notable that some outlier data points exceed even 75%.

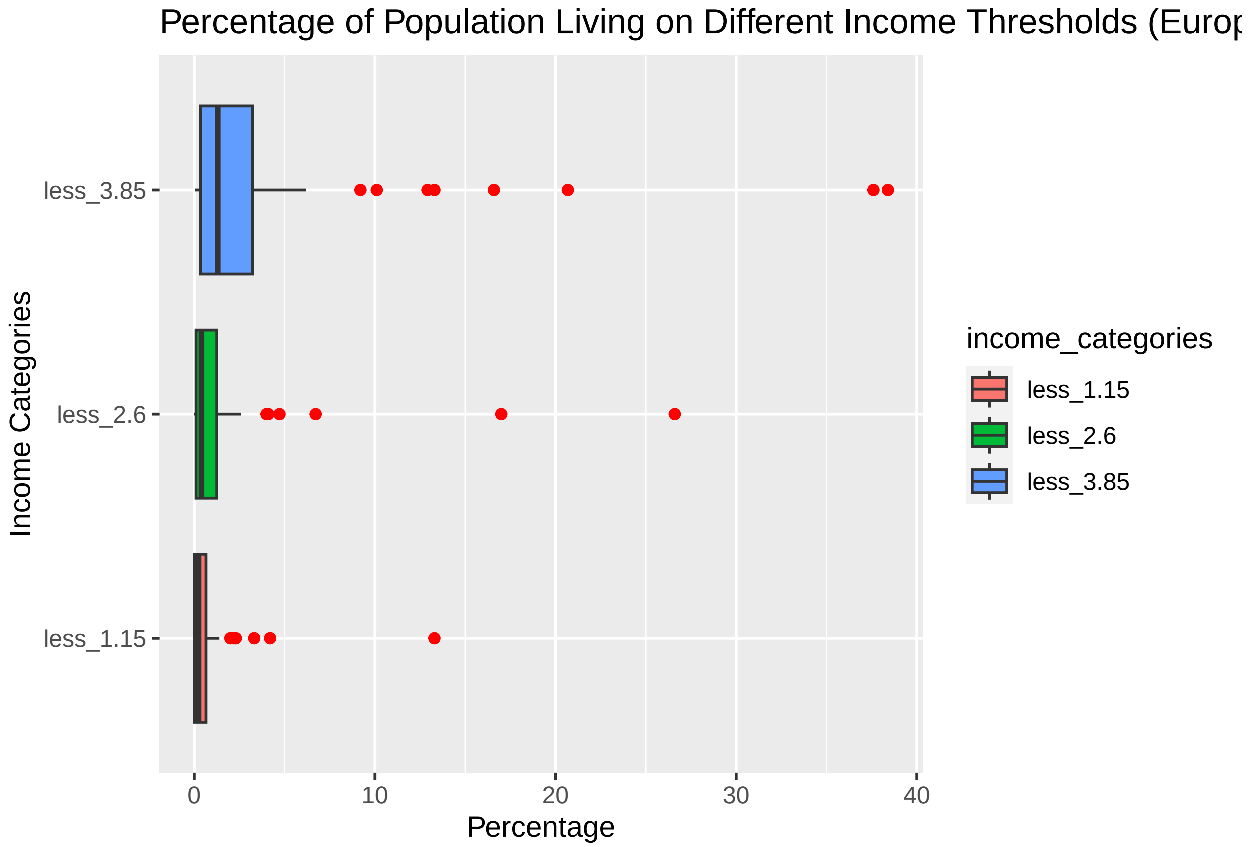
For the less_3.85 category, the data spreads from about 5% to 50%, with a median around 24%. The distribution exhibits a right skew, and there is a notable high standard deviation.

The box plots for Asia reveal that the presence of outliers significantly widens the distribution, emphasizing the income disparities within the region.

```
# Select Europe data
europe_data <- df %>%
  dplyr::filter(Continent == "Europe") %>%
  select(less_1.15, less_2.6, less_3.85)

# melting europe_data, convert column name becomes x, its values becomes y
europe_data <- melt(europe_data, value.name = "percentage", variable.name = "income_categories")
```

```
## No id variables; using all as measure variables
```

```
# Create Europe boxplot
ggplot(data = europe_data, aes(x=income_categories, y=percentage)) +
         geom_boxplot(aes(fill=income_categories), outlier.colour="red") +
         coord_flip() +
         labs(title = "Percentage of Population Living on Different Income Thresholds (Europe)",
       x = "Income Categories", y = "Percentage")
```



In the Europe box plot, all three income categories have distributions below 5%, but their outliers extend broadly.

In the less_1.15 category, both the median and the distribution are close to 0%, with a few outliers below 5%. For the less_2.6 category, the median is approximately 0%, and there is a slight right skew. The distribution ranges from 0% to 3%, with outliers ranging from 5% to over 25%. In the less_3.85 category, the median is around 2%, and the distribution spans from 0% to 4%. It exhibits a slight right skew with outlier points ranging from 10% to 38%.

The Europe box plot indicates that less than 5% of the population falls within these income thresholds, but there is a wide range of outliers.