DA5020.A7.Nithya.Sarabudla

nithyasarabudla

2023-10-25

loading the packages

```
library(rvest)
library(dplyr)
```

```
##
## 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(ggplot2)
library (tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()         masks stats::filter()
## x readr::guess_encoding() masks rvest::guess_encoding()
## x dplyr::lag()            masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

extract the tabular data on the “Percent of

population living on less than $1.15, $2.60 and $3.85 a day” from the Wikipedia page.

Question 1

PART 1

Scrape the data from the webpage and extract the following fields: Country, < $1.15, <

$2.60, < $3.85, Year and Continent. Prepare the data for analysis and ensure that the columns have

meaningful names.

```
# Read the HTML page from the given URL
page = read_html("https://en.wikipedia.org/w/index.php?title=List_of_sovereign_states_by_percentage_of_

# Extract the third table from the HTML page
table <- html_table(page)[3]
table
```

```
## [[1]]
## # A tibble: 166 x 6
##    Country    `$1.15` `$2.60` `$3.85`  Year Continent
##    <chr>      <chr>   <chr>   <chr>   <int> <chr>
##  1 Albania    13.35   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
##  7 Austria    0.60%   0.70%   0.80%    2019 Europe
##  8 Azerbaijan 0.00%   0.00%   0.00%    2019 Asia
##  9 Bangladesh 6.62%   37.44%  76.01%   2019 Asia
## 10 Belarus    0.00%   0.00%   0.10%    2020 Europe
## # i 156 more rows
```

```
# Convert the table to a data frame
df <- data.frame(table)

# Rename the columns
colnames(df) <- c("Country", "less_than_1.15", "less_than_2.60", "less_than_3.85", "Year", "Continent")

# Clean and convert the "less_than_1.15" column by removing any trailing characters and converting to n
df$less_than_1.15 <- as.array(df$less_than_1.15)
df$less_than_1.15 <- substring(df$less_than_1.15,1,nchar(df$less_than_1.15)-1)
```

```
df$less_than_1.15 <- as.numeric(df$less_than_1.15)

# Clean and convert the "less_than_2.60" column by removing any trailing characters and converting to n
df$less_than_2.60<- as.array(df$less_than_2.60)
df$less_than_2.60 <- substring(df$less_than_2.60,1,nchar(df$less_than_2.60)-1)
df$less_than_2.60 <- as.numeric(df$less_than_2.60)

# Clean and convert the "less_than_3.85" column by removing any trailing characters and converting to n
df$less_than_3.85 <- as.array(df$less_than_3.85)
df$less_than_3.85 <- substring(df$less_than_3.85,1,nchar(df$less_than_3.85)-1)
df$less_than_3.85<- as.numeric(df$less_than_3.85)

# Display the first few rows of the data frame
head(df)
```

```
##     Country less_than_1.15 less_than_2.60 less_than_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
```

PART 2

Calculate the mean and the standard deviation of the percent of the population living under

$3.85 per day for each continent. Perform a comparative analysis (i.e. explanation) of the data from

each continent

```
# Calculate the mean and standard deviation of "less_than_3.85" percentages for each continent
less_than_3.85_df <- df %>%
  select(Continent, less_than_3.85) %>%
  group_by(Continent) %>%
  summarise(Mean = mean(less_than_3.85, na.rm = TRUE), std_dev = sd(less_than_3.85, na.rm = TRUE))

less_than_3.85_df
```

```
## # A tibble: 7 x 3
##   Continent      Mean std_dev
##   <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
```

```
# Find and display the continent with the lowest mean poverty level
min_mean_continent <- less_than_3.85_df %>%
  filter(Mean == min(Mean)) %>%
  select(Continent) %>%
  pull()
cat("Continent with the lowest mean poverty level:", min_mean_continent, "\n")
```

```
## Continent with the lowest mean poverty level: Europe
```

```
# Find and display the continent with the lowest standard deviation
min_std_dev_continent <- less_than_3.85_df %>%
  filter(std_dev == min(std_dev)) %>%
  select(Continent) %>%
  pull()
cat("Continent with the lowest standard deviation:", min_std_dev_continent, "\n")
```

```
## Continent with the lowest standard deviation: Asia, Europe
```

```
# Find and display the continent with the highest mean poverty level
max_mean_continent <- less_than_3.85_df %>%
  filter(Mean == max(Mean)) %>%
  select(Continent) %>%
  pull()
cat("Continent with the highest mean poverty level:", max_mean_continent, "\n")
```

```
## Continent with the highest mean poverty level: Africa
```

```
# Find and display the continent with the highest standard deviation
max_std_dev_continent <- less_than_3.85_df %>%
  filter(std_dev == max(std_dev)) %>%
  select(Continent) %>%
  pull()
cat("Continent with the highest standard deviation:", max_std_dev_continent, "\n")
```

```
## Continent with the highest standard deviation: Asia
```

PART 3

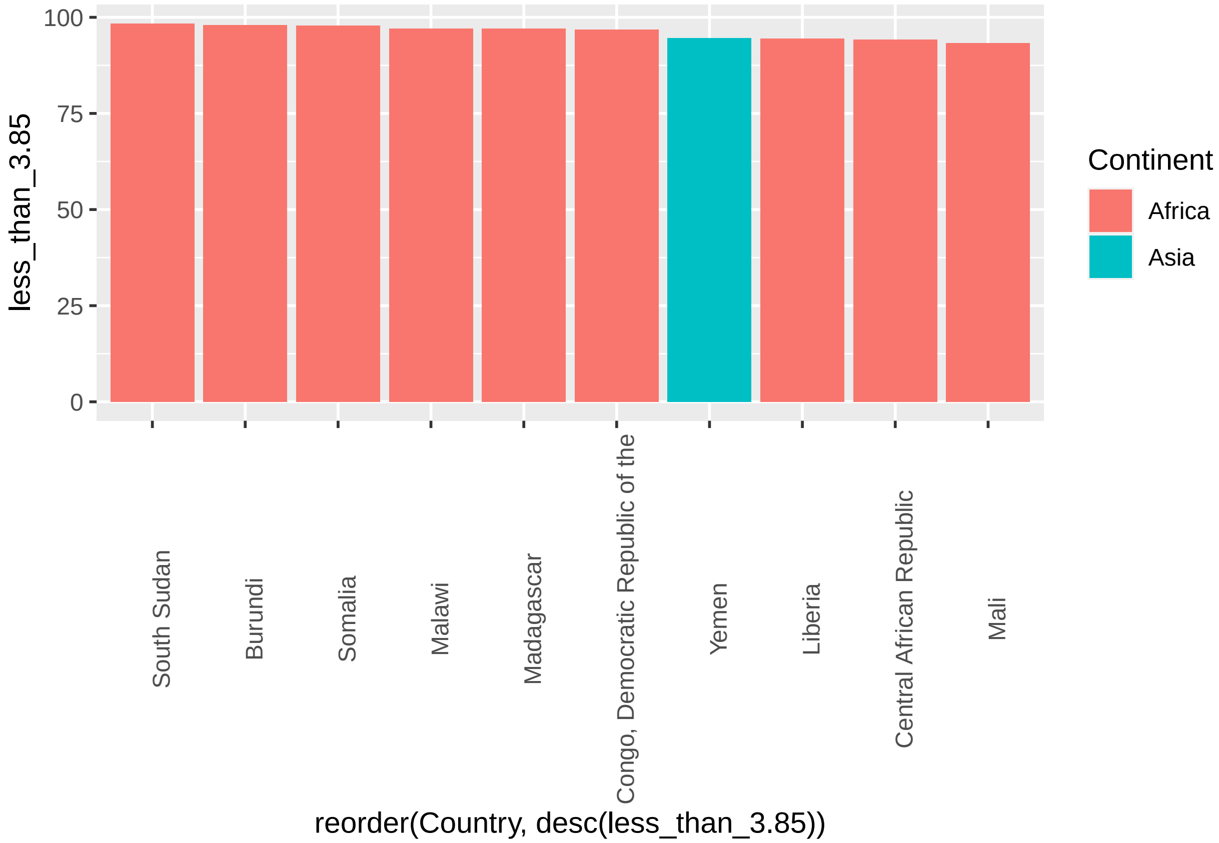
What are the 10 countries with the highest percentage of the population having an income

of less than $3.85 per day? Using a suitable chart, display the country name, the percentage and

color- code by the Continent. Summarize your findings

```r
# Extract the top 10 countries with the highest percentage of income less than $3.85
top_10_countries <- df %>%
  arrange(desc(less_than_3.85)) %>%
  head(10)

# Create a bar plot to visualize the top 10 countries' income percentages
df %>%
  arrange(desc(less_than_3.85)) %>%
  select(Country, less_than_3.85, Continent) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Country, desc(less_than_3.85)), y = less_than_3.85, fill = Continent)) +
  geom_bar(stat = 'identity') +
  theme(axis.text.x = element_text(angle = 90))
```



```r
# Summarize your findings
summary(top_10_countries)
```

```
##    Country          less_than_1.15  less_than_2.60  less_than_3.85
##  Length:10          Min.   :42.26   Min.   :74.64   Min.   :93.29
##  Class :character   1st Qu.:58.18   1st Qu.:82.98   1st Qu.:94.48
##  Mode  :character   Median :70.88   Median :88.64   Median :96.94
##                     Mean   :66.58   Mean   :86.52   Mean   :96.19
##                     3rd Qu.:75.79   3rd Qu.:90.39   3rd Qu.:97.68
##                     Max.   :80.71   Max.   :93.27   Max.   :98.44
##       Year       Continent
##  Min.   :2019   Length:10
##  1st Qu.:2019   Class :character
##  Median :2019   Mode  :character
##  Mean   :2019
##  3rd Qu.:2019
##  Max.   :2019
```

From the above plot, it's evident that the majority of the countries with the highest percentage of the population living on less than $3.85 per day are from Africa. Specifically, out of the top 10 countries in this category, 9 are from Africa, while only 1 country is from Asia. South Sudan ranks as the top country with the highest percentage of its population earning less than $3.85 per day. The 10th country on the list with the highest population living on less than $3.85 per day is Mali, also from Africa.

PART 4

Explore the countries with the lowest percentage of the population having an income of

less than $3.85 per day. What are the 5 countries with the lowest percentage, and how does the

results compare to the other income groups (i.e. $1.15 and $2.60)?

```r
# Arrange the data frame in ascending order of the "less_than_3.85" column
# and select the top 5 rows with the lowest income percentages
df %>% arrange(less_than_3.85) %>% head(5)
```

```
##                Country less_than_1.15 less_than_2.60 less_than_3.85 Year
## 1           Azerbaijan              0            0.0           0.00 2019
## 2 United Arab Emirates              0            0.0           0.00 2019
## 3              Iceland              0            0.0           0.04 2019
## 4              Belarus              0            0.0           0.10 2020
## 5              Finland              0            0.1           0.10 2019
##   Continent
## 1      Asia
## 2      Asia
## 3    Europe
## 4    Europe
## 5    Europe
```

Among the 5 countries with the lowest population percentage living on less than $3.85 per day, 2 are in Asia (Azerbaijan the United Arab Emirates) with a 0% population in this category, while the remaining 3 are from Europe (Iceland, Belarus, and Finland) with percentages ranging from 0.04% to 0.1%. For the less than $2.60 per day threshold, only Finland has 0.1%, with the rest at 0%. In all these countries, when considering less than $1.15 per day, the population percentage is 0%.

PART 5

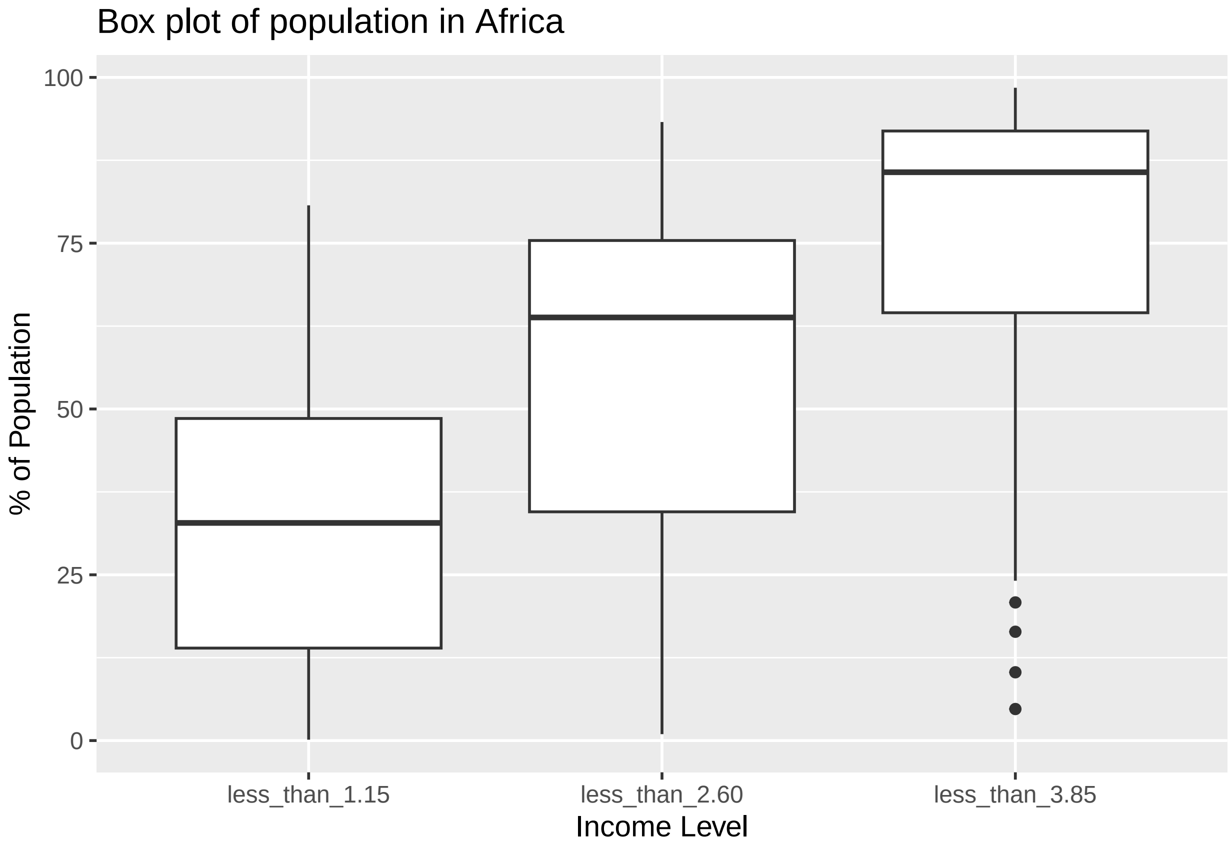
Extract the data for any two continents of your choice. For each continent, visualize the

percent of the population living on less than $1.90, $3.20 and $5.50 using box plots. Compare and

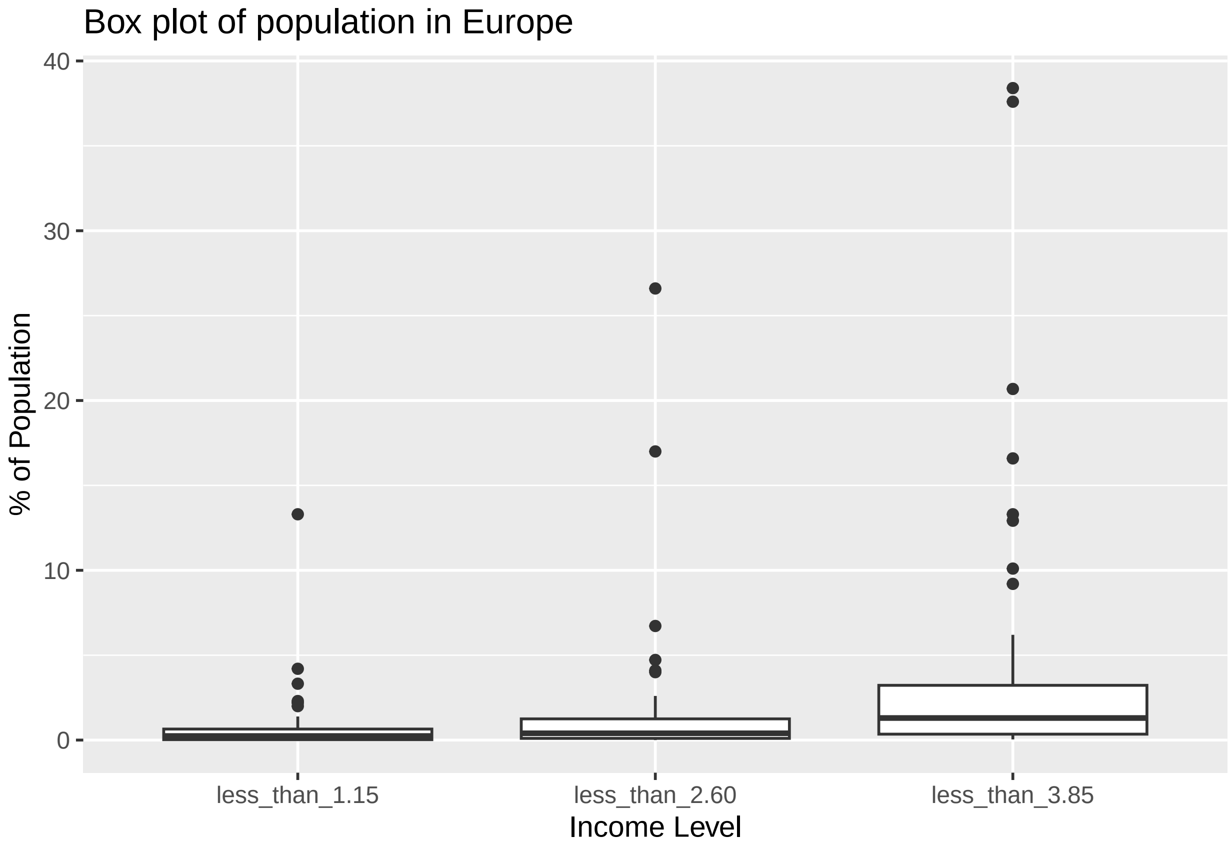
contrast the results, while ensuring that you discuss the distribution, skew and any outliers that are

evident.

```r
# Create a box plot to visualize the income distribution in Africa
df %>%
  filter(Continent == "Africa") %>%
  pivot_longer(cols = starts_with("less"), names_to = "Income Level", values_to = "Percentage") %>%
  ggplot(aes(x = `Income Level`, y = Percentage)) +
  geom_boxplot() +
  ggtitle("Box plot of population in Africa") +
  xlab("Income Level") +
  ylab("% of Population")
```



```
# Create a box plot to visualize the income distribution in Europe
df %>%
  filter(Continent == "Europe") %>%
  pivot_longer(cols = starts_with("less"), names_to = "Income Level", values_to = "Percentage") %>%
  ggplot(aes(x = `Income Level`, y = Percentage)) +
  geom_boxplot() +
  ggtitle("Box plot of population in Europe") +
  xlab("Income Level") +
  ylab("% of Population")
```



Box Plot for Asia:

Distribution: The box plot for “Asia” shows that the distribution of the population living on less than $1.15 is right skewed, while the distribution for less than $2.60 is slightly left-skewed, and the distribution for less than $3.85 is strongly left-skewed.

Skew: Right-skewed for less than $1.15, with most people having lower income. Slightly left-skewed for less than $2.60, where the majority have higher income. Strongly left-skewed for less than $3.85, indicating that the majority have even higher income levels.

Outliers: There are 4 outliers in less_than_3.85 means there are a few places or individuals in Asia with much lower income compared to most people.

Box Plot for Europe :

In Europe, the income distribution for less than $1.15, less than $2.60, and less than $3.85 is strongly right-skewed. This indicates that the majority of the population in Europe has higher income levels, with a long tail of individuals who have lower income. There are a few outliers in the three income levels, which represent exceptional cases of individuals or regions with significantly lower income than the majority of the population.