# Webscraping

Rebecca Weiss

10/31/2021

---

```
#load all necessary libraries
library(XML)
library(RCurl)
library(scrapeR)
library(rvest)
library(tidyverse)
```

Clear the workspace:

```
rm(list = ls())
```

## 1. In this question, you will use rvest to parse the HTML and extract the tabular data on the "Percent ofpopulation living on less than $1.90, $3.20 and $5.50 a day" from the Wikipedia page.

### 1. (10 pts) Scrape the data from the webpage and extract the following fields: Country, < $1.90, < $3.20, < $5.50, Year and Continent. Prepare the data for analysis and ensure that the columns have meaningful names.

```
# download HTML from URL, scrape table and save into df
url <- "https://en.wikipedia.org/wiki/List_of_countries_by_percentage_of_population_living_in_poverty"
html_data <- read_html(url)
df <- html_data %>%
  html_node('.wikitable') %>%
  html_table()

# inspect table data
str(df)
```

```
## tibble [164 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Country      : chr [1:164] "Albania" "Algeria" "Angola" "Argentina" ...
##  $ < $1.90[8][5]: chr [1:164] "1.3%" "0.4%" "49.9%" "1.5%" ...
##  $ < $3.20[6]   : chr [1:164] "8.2%" "3.7%" "72.0%" "5.0%" ...
##  $ < $5.50[7]   : chr [1:164] "33.8%" "28.6%" "89.3%" "14.0%" ...
##  $ Year         : int [1:164] 2017 2011 2018 2019 2019 2014 2018 2005 2016 2019 ...
##  $ Continent    : chr [1:164] "Europe" "Africa" "Africa" "South America" ...
```

```
# rename columns
df <- rename(df, '< $1.90' = `< $1.90[8][5]`,
       '< $3.20' = `< $3.20[6]`,
       '< $5.50' = `< $5.50[7]`)


# need to remove % and divide by 100 to get numeric
df$`< $1.90` <- as.numeric(sub("%", "",df$`< $1.90`,fixed=TRUE))/100
df$`< $3.20` <- as.numeric(sub("%", "",df$`< $3.20`,fixed=TRUE))/100
df$`< $5.50` <- as.numeric(sub("%", "",df$`< $5.50`,fixed=TRUE))/100

head(df)
```

```
## # A tibble: 6 x 6
##   Country   `< $1.90` `< $3.20` `< $5.50`  Year Continent
##   <chr>         <dbl>     <dbl>     <dbl> <int> <chr>
## 1 Albania       0.013     0.082     0.338  2017 Europe
## 2 Algeria       0.004     0.037     0.286  2011 Africa
## 3 Angola        0.499     0.72      0.893  2018 Africa
## 4 Argentina     0.015     0.05      0.14   2019 South America
## 5 Armenia       0.011     0.1       0.44   2019 Asia
## 6 Australia     0.005     0.007     0.007  2014 Oceania
```

**2. (10 pts) Calculate the mean and the standard deviation of the percent of the population living under $5.50 per day for each continent. Perform a comparative analysis (i.e. explanation) of the data from each continent.**

```
# group by continent, calculate mean and sd, sort by largest
under550 <- df %>%
  group_by(Continent) %>%
  summarise('mean < $5.50' = mean(`< $5.50`),
            'stdev < $5.50' = sd(`< $5.50`)) %>%
  arrange(desc(`mean < $5.50`))

# view results
under550
```

```
## # A tibble: 7 x 3
##   Continent     `mean < $5.50` `stdev < $5.50`
##   <chr>                  <dbl>           <dbl>
## 1 Africa                0.751           0.245
## 2 Oceania               0.555           0.283
## 3 North America         0.282           0.222
## 4 South America         0.236           0.135
## 5 Asia, Europe          0.0685          0.0445
## 6 Europe                0.0546          0.0964
## 7 Asia                 NA              NA
```

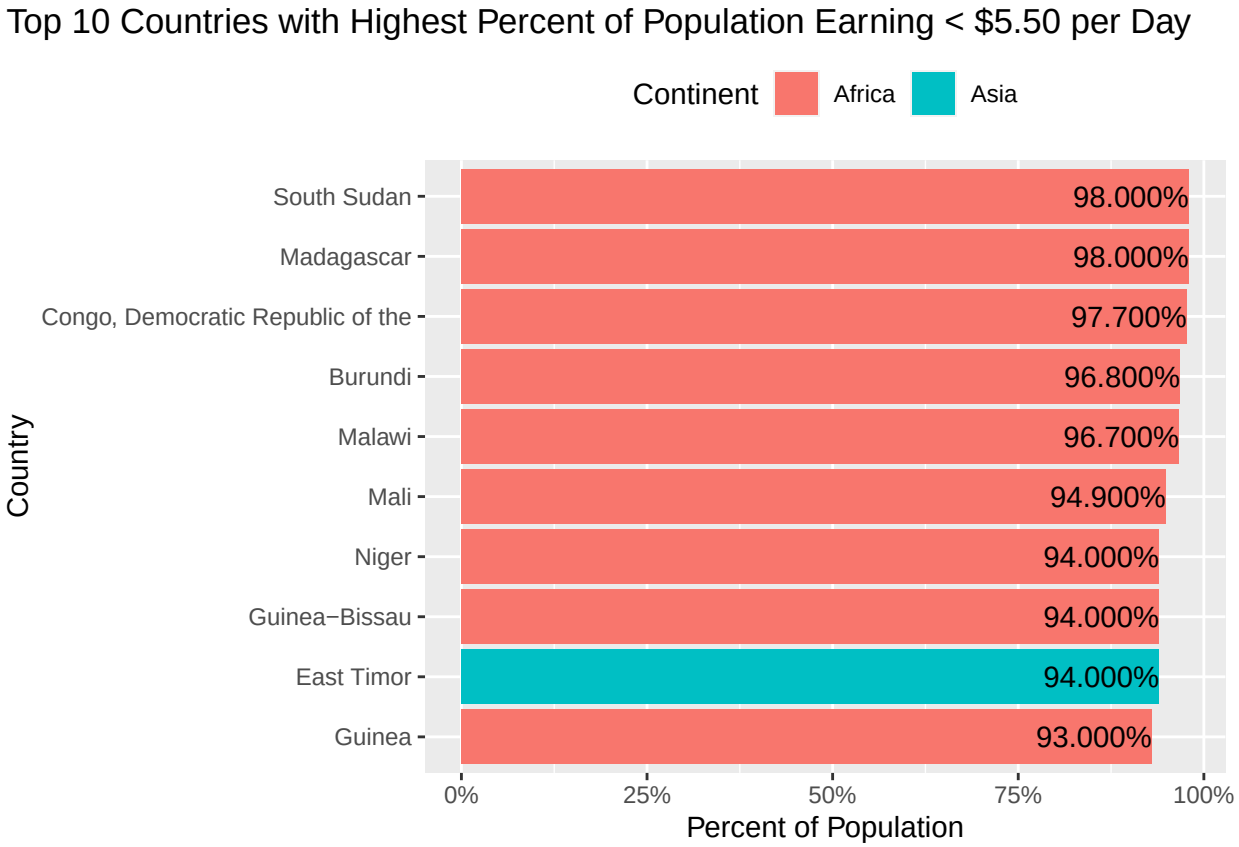
**3. (5 pts) What are the 10 countries with the highest percentage of the population having an income of less than $5.50 per day? Using a suitable chart, display the country name, the percentage and color- code by the Continent. Summarize your findings.**

```
# filter based on top 10 with < $ 5.50
top10 <- df %>%
  # select(Country, `< $5.50`, Continent) %>%
  arrange(desc(`< $5.50`)) %>%
  slice_head(n = 10)

# view results
top10
```

```
## # A tibble: 10 x 6
##    Country                          `< $1.90` `< $3.20` `< $5.50`  Year Continent
##    <chr>                                <dbl>     <dbl>     <dbl> <int> <chr>
##  1 Madagascar                           0.788     0.92      0.98   2012 Africa
##  2 South Sudan                          0.764     0.92      0.98   2016 Africa
##  3 Congo, Democratic Republic of the    0.772     0.91      0.977  2012 Africa
##  4 Burundi                              0.728     0.9       0.968  2013 Africa
##  5 Malawi                               0.692     0.894     0.967  2016 Africa
##  6 Mali                                 0.503     0.8       0.949  2009 Africa
##  7 East Timor                           0.307     0.733     0.94   2021 Asia
##  8 Guinea-Bissau                        0.684     0.845     0.94   2010 Africa
##  9 Niger                                0.454     0.769     0.94   2014 Africa
## 10 Guinea                               0.361     0.71      0.93   2012 Africa
```

```
# visualize results
ggplot(top10, aes(x=reorder(Country, `< $5.50`), y=`< $5.50`, fill = Continent,
                  label=scales::percent(`< $5.50`))) +
  geom_bar(stat='identity') +
  labs(x = "Country",
       y = "Percent of Population",
       title = "Top 10 Countries with Highest Percent of Population Earning < $5.50 per Day") +
  scale_y_continuous(labels = scales::percent) +
  geom_text(position = position_dodge(width = .9),
            hjust = 1) +
  theme(plot.title.position = "plot") +
  theme(legend.position = "top") +
  coord_flip()
```




As we can see from the output, the top 10 countries with the highest percentage of the population earning $<$ \$5.50 per day have almost *all* of their population earning that little, as they all have values greater than 93%. In the plot, you can see that 9 of the top 10 countries are in the continent of Africa, with the exception of East Timor, which are on the continent of Asia.

**4. (5 pts) Explore the countries with the lowest percentage of the population having an income of less than \$5.50 per day. What are the 5 countries with the lowest percentage, and how does the results compare to the other income groups (i.e. \$1.90 and \$3.20)?**

```
# filter based on lowest with < $ 5.50
bottom5 <- df %>%
  # select(Country, `< $5.50`, Continent) %>%
  arrange((`< $5.50`)) %>%
  slice_head(n = 5)

# view results
bottom5
```

```
## # A tibble: 5 x 6
##   Country     `< $1.90` `< $3.20` `< $5.50`  Year Continent
##   <chr>           <dbl>     <dbl>     <dbl> <int> <chr>
## 1 Switzerland     0         0         0      2018 Europe
## 2 Cyprus          0         0.001     0.001  2018 Europe
```

```
## 3 Finland          0.001     0.001     0.001  2018 Europe
## 4 Slovenia         0         0         0.001  2018 Europe
## 5 Belarus[11]      0         0         0.002  2019 Europe
```

From the output, we see the bottom 5 countries with percentage of population earning $< 5.50$ a day are Switzerland, Cyprus, Finland, Slovenia and Belarus, all with $< 0.2\%$ of their population earning $< 5.50$. These are all located on the continent of Europe, and interestingly, these 5 countries have virtually none of their population earning $< 1.90$ or 3.20 per day.

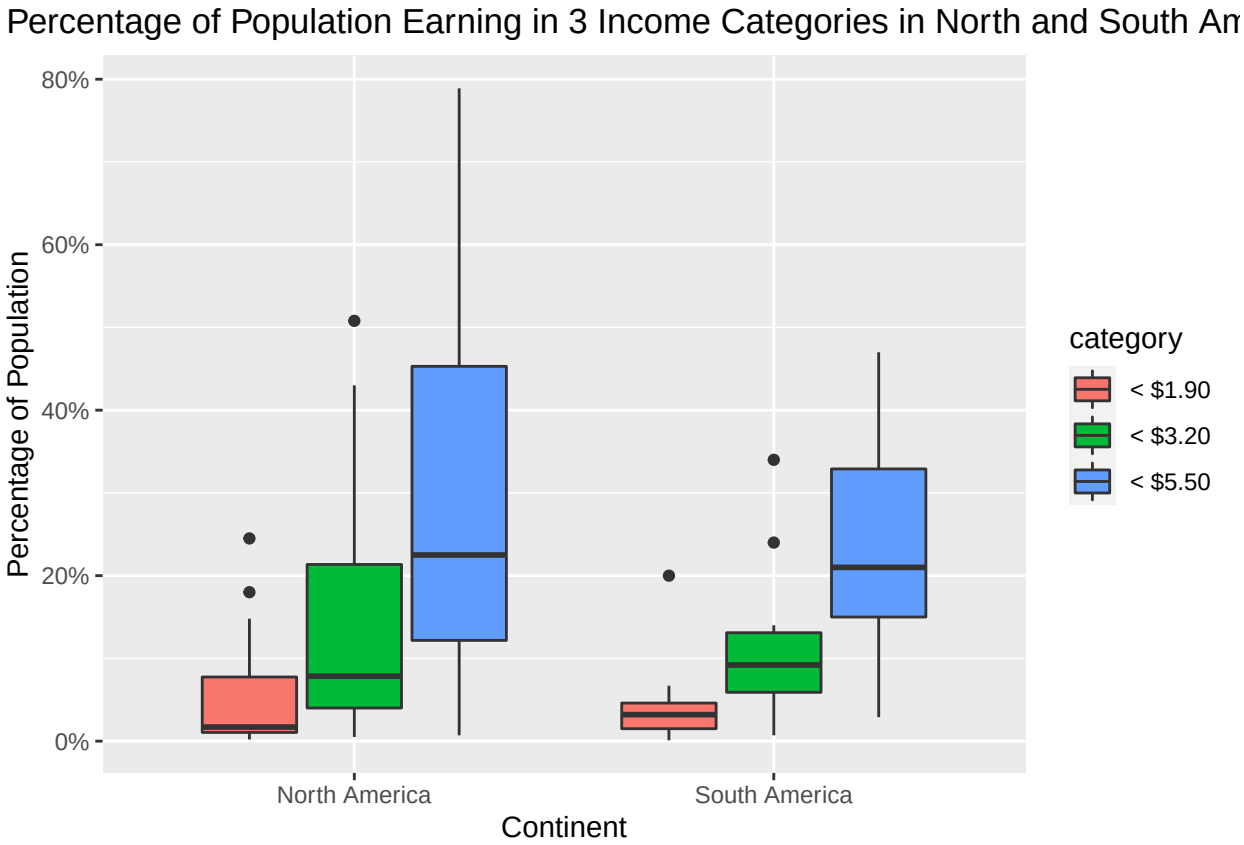
**5. (20 pts) Extract the data for any two continents of your choice. For each continent, visualize the percent of the population living on less than $1.90, $3.20 and $5.50 using box plots. Compare and contrast the results, while ensuring that you discuss the distribution, skew and any outliers that are evident.**

```
# select two continents: I will use North and South America
# pivot longer so can more easily group in box plot
conts <- df %>%
  filter(Continent == "North America" | Continent == "South America") %>%
  group_by(Continent) %>%
  pivot_longer(cols = c(`< $1.90`, `< $3.20`,`< $5.50`),
               names_to = c("category"),
               values_to = "values")

# view output
conts
```

```
## # A tibble: 81 x 5
## # Groups:   Continent [2]
##    Country    Year Continent     category values
##    <chr>     <int> <chr>         <chr>     <dbl>
##  1 Argentina  2019 South America < $1.90   0.015
##  2 Argentina  2019 South America < $3.20   0.05
##  3 Argentina  2019 South America < $5.50   0.14
##  4 Belize     2021 North America < $1.90   0.18
##  5 Belize     2021 North America < $3.20   0.43
##  6 Belize     2021 North America < $5.50   0.51
##  7 Bolivia    2019 South America < $1.90   0.032
##  8 Bolivia    2019 South America < $3.20   0.08
##  9 Bolivia    2019 South America < $5.50   0.2
## 10 Brazil     2019 South America < $1.90   0.046
## # ... with 71 more rows
```

```
# create a grouped boxplot for each continent, separated by category of income
ggplot(conts, aes(x=Continent, y=values, fill=category)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::percent) +
  labs(y="Percentage of Population",
       x="Continent",
       title = "Percentage of Population Earning in 3 Income Categories in North and South America") +
  theme(plot.title.position = "plot")
```




From the output boxplot, we see that there is a drastic range in the values for percentage of population earning $< 5.50$, 3.20, and 1.90 between North and South America. In general, North America has a wider distribution for all 3 categories than South America. In both continents, the $< 1.90$ and 3.20 both have outliers that exceed much beyond the boxplot range. In North America, the IQR depicted in the boxplots appears to show that the percentage of the population falling into any of the three categories is wider than the South America one, but the median value sits at about the same place. Additionally, it looks like both continents have the widest range for $< 5.50$, but the North America group has a skew towards higher percentage of the population than lower falling into each category, which you can see by the length of the "whiskers" on top of the boxes on the left.