DATA 606 Data Project Proposal

Andrew Bowen

Setup

```r
library(dplyr)
```

Data Preparation

```r
# load data
data_url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/fifa/fifa_countries_audience

fifa <- read.csv(data_url)
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for. Research Question: Is there a stignificant difference in the mean viewership of soccer in Europe vs another continent (confederation)?

Cases

What are the cases, and how many are there? There are viewership data for 191 countries. There are 6 confederations in total. We are interested in comparing the European confederation (UEFA) to the other ones.

Data collection

Describe the method of data collection. This data was collected from FIFA TV viewership during a world cup. It was cleaned and put into csv format by FiveThirtyEight and lives in their fifa data GitHub repo

Type of study

What type of study is this (observational/experiment)? This is an observational study as the researchers are not actively changing conditions to influence behaviors/outcomes.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link. This data comes from a FiveThirtyeight data set on FIFA viewership. We'll focus on the GDP-weighted TV viewership column `gdp_weighted_share`, as that accounts for population differences.

Dependent Variable

What is the response variable? Is it quantitative or qualitative? The dependent/response variable is the `gdp_weighted_share`, which is a country's GDP-weighted audience share (as a percentage) of all viewers of a world cup. This is qualitative data

Independent Variable(s)

Our independent variable is the Confederation (UEFA - Europe, CONCACAF - North America, etc.) to which a country belongs. This is categorical.

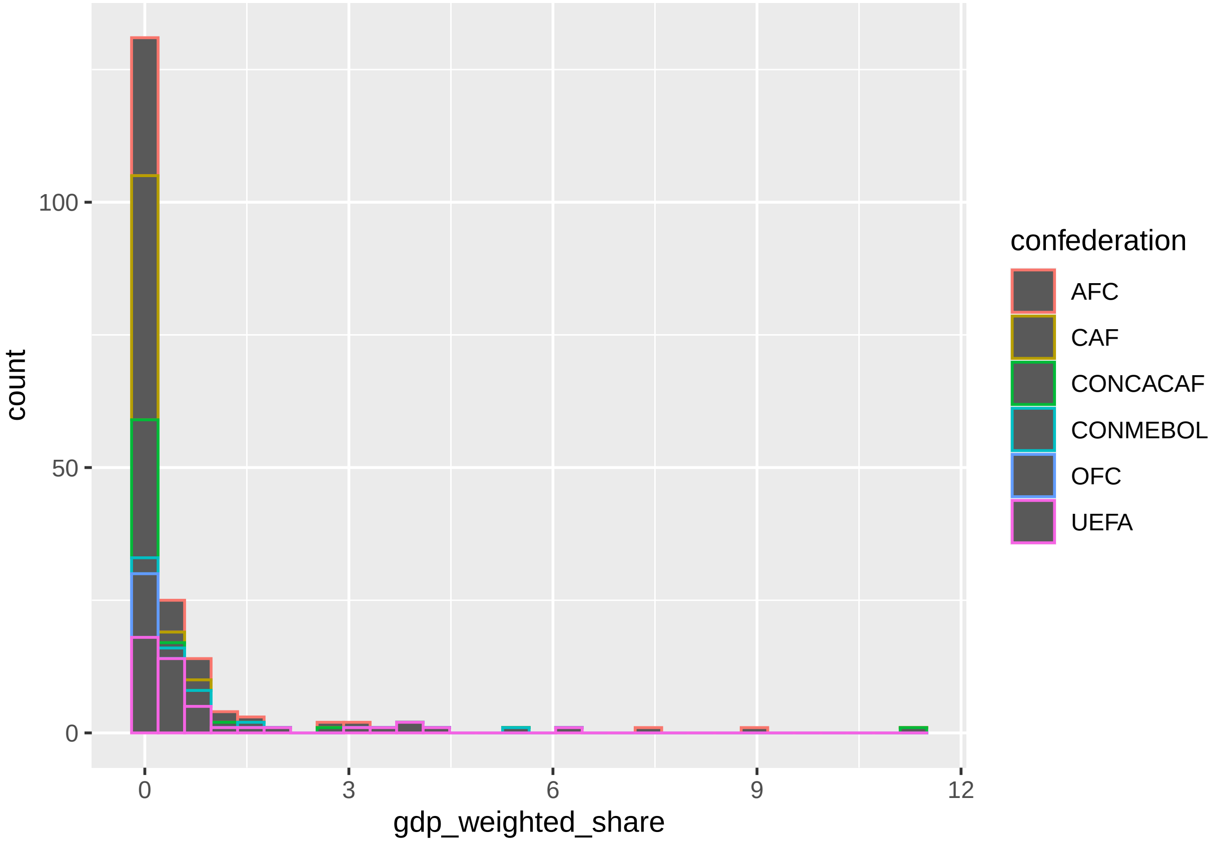
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
library(ggplot2)

ggplot(fifa, aes(x=gdp_weighted_share, group=confederation, color=confederation)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These distributions do not appear normal, so we'll use the median as our summary stat.

```
fifa %>%
    group_by(confederation) %>%
    summarise(median_gdp_weighted_share = median(gdp_weighted_share))
```

```
## # A tibble: 6 x 2
##   confederation median_gdp_weighted_share
##   <chr>                             <dbl>
## 1 AFC                                 0.1
## 2 CAF                                 0
## 3 CONCACAF                            0
## 4 CONMEBOL                            0.5
## 5 OFC                                 0
## 6 UEFA                                0.3
```

Hypothesis Testing

- H_0: The mean GDP-weighted viewership share of the world cup in Europe is *not higher* than that of other confederations
- H_a: The mean GDP-weighter viewership share of the world cup is *higher* than that of other confederations

Running a t-test between these two groups with a significance level $\alpha = 0.05$

```r
# Filtering into our two groups: Europe vs not Europe.
europe <- fifa %>% filter(confederation == "UEFA")
other_countries <- fifa %>% filter(confederation != "UEFA")

# Running one-tailed t-test using R built-in
t.test(europe$gdp_weighted_share, other_countries$gdp_weighted_share, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  europe$gdp_weighted_share and other_countries$gdp_weighted_share
## t = 1.7859, df = 77.677, p-value = 0.03901
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.02926207        Inf
## sample estimates:
## mean of x mean of y
## 0.8478261 0.4165517
```

Since our p-value is less than our significance level ($\alpha = 0.05$), we can reject the null hypothesis and state that the average weighted viewership of the world cup is higher in Europe than in other confederations.