Project Proposal

Due November 16 at 11:59pm

Aesha Gandhi, Leo Yao, Junhao Xu, Sebine Scaria

Load Packages

```
library(tidyverse)
library(readxl)
library(dplyr)
library(ggplot2)
```

Dataset 1 (top choice)

Data source: Our World in Data – Global Energy Data (OWID)

Brief description:

The dataset provides annual country-level data on energy production, consumption, and electricity generation across more than 190 countries from 1965 to 2023. Each row represents a unique country–year observation. The data is compiled from multiple international sources, including BP Statistical Review and Ember, and harmonized by Our World in Data.

The dataset includes over 100 indicators related to total and per capita energy use, electricity generation by fuel type, carbon intensity, and the shares of renewable and fossil energy. These measures allow cross-country comparisons of energy structures, economic development, and progress toward decarbonization. Key variables used in this study include GDP, population, carbon intensity of electricity, fossil fuel share, and renewable energy share.

Research question 1:

Does a country's GDP per capita significantly affect the carbon intensity of its electricity generation, and does this relationship differ between high-income and low-income countries?

- Outcome variable (include the name/description and type of variable):

- carbon_intensity_elec **(continuous)** measures the amount of carbon dioxide emitted per unit of electricity generated (grams of CO per kilowatt-hour). It captures how carbon-efficient a country's electricity production is — lower values indicate cleaner, more renewable-based power systems.

Load the data and provide a `glimpse()`:

```
#Read the dataset, select the columns to be used and drop the missing data
energy_rq1 <- read_excel("owid-energy-data.xlsx") %>%
  select(country, year, gdp, population, carbon_intensity_elec,fossil_share_elec,electricity_
  mutate(gdp_per_capita = gdp / population) %>%
  filter(!is.na(carbon_intensity_elec),
         !is.na(gdp_per_capita))

glimpse(energy_rq1)
```
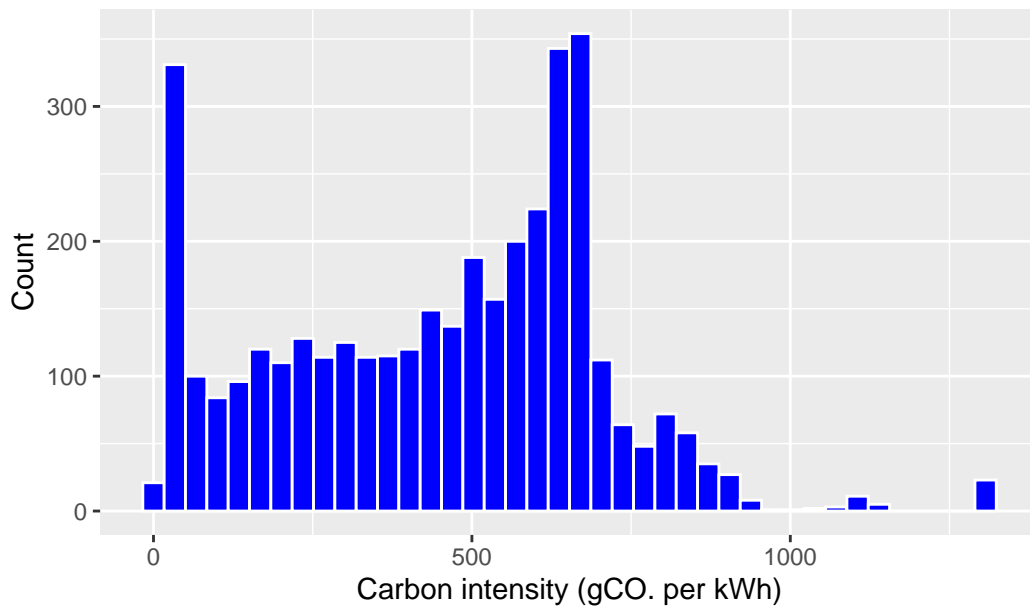
```
Rows: 3,800
Columns: 8
$ country               <chr> "Afghanistan", "Afghanistan", "Afghanistan", "A~
$ year                  <dbl> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007,~
$ gdp                   <dbl> 11283793920, 11021273088, 18804871168, 21074343~
$ population            <dbl> 20130279, 20284252, 21378081, 22733007, 2356059~
$ carbon_intensity_elec <dbl> 250.000, 217.391, 169.014, 241.758, 227.848, 21~
$ fossil_share_elec     <dbl> 35.417, 27.536, 21.127, 30.769, 29.114, 28.049,~
$ electricity_generation <dbl> 0.48, 0.69, 0.71, 0.91, 0.79, 0.82, 0.90, 1.01,~
$ gdp_per_capita        <dbl> 560.5384, 543.3414, 879.6333, 927.0372, 947.877~
```
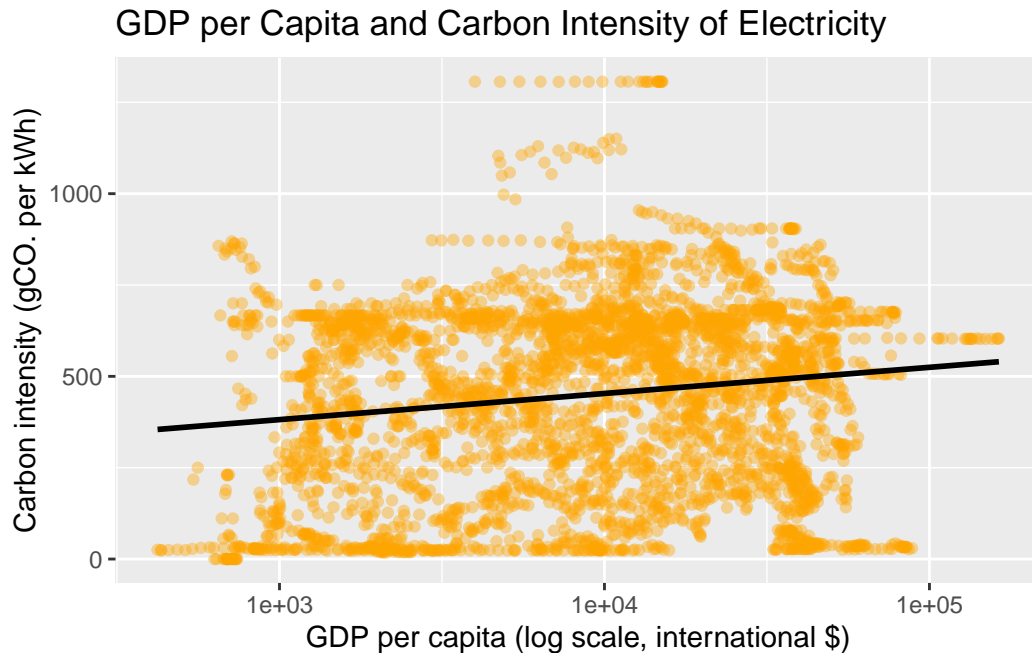
Exploratory Plots:

```
ggplot(energy_rq1, aes(x = carbon_intensity_elec)) +
  geom_histogram(bins = 40, fill = "blue", color = "white") +
  labs(
    title = "Distribution of Carbon Intensity of Electricity",
    x = "Carbon intensity (gCO per kWh)",
    y = "Count"
  )
```

Distribution of Carbon Intensity of Electricity


```
ggplot(energy_rq1, aes(x = gdp_per_capita, y = carbon_intensity_elec)) +
  geom_point(alpha = 0.4, color = "orange") +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  scale_x_log10() +
  labs(
    title = "GDP per Capita and Carbon Intensity of Electricity",
    x = "GDP per capita (log scale, international $)",
    y = "Carbon intensity (gCO  per kWh)"
  )
```

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

GDP per Capita and Carbon Intensity of Electricity

Interpretation:

The distribution of carbon intensity is bimodal and right-skewed, suggesting that countries cluster into two groups — those with predominantly renewable-based electricity (very low emissions) and those still dependent on fossil fuels (very high emissions).

Contrary to expectations, the plot shows no clear negative link between GDP per capita and carbon intensity. This likely reflects differences among high-income countries and the influence of unobserved factors such as fossil fuel share and electricity generation, which will be controlled for in the regression analysis.

Research question 2:

What factors are associated with a country having >50% renewable electricity share?

- Outcome variable (include the name/description and type of variable):

    - `renewable_majority` **(Binary)**: indicates whether a country generates more than half of its electricity from renewable energy sources (1 = yes, 0 = no).
        * Constructed as 1 if `renewables_share_elec` > 50, otherwise 0

Load the data and provide a `glimpse()`:

```
energy_rq2 <- read_excel("owid-energy-data.xlsx") %>%
  select(country, year, renewables_share_elec,
         fossil_share_elec, electricity_generation,
         gdp, population) %>%
  mutate(
    renewable_majority = if_else(renewables_share_elec > 50, 1, 0),
    gdp_per_capita = gdp / population
  ) %>%
  filter(!is.na(renewable_majority),
         !is.na(fossil_share_elec),
         !is.na(gdp_per_capita))

glimpse(energy_rq2)
```
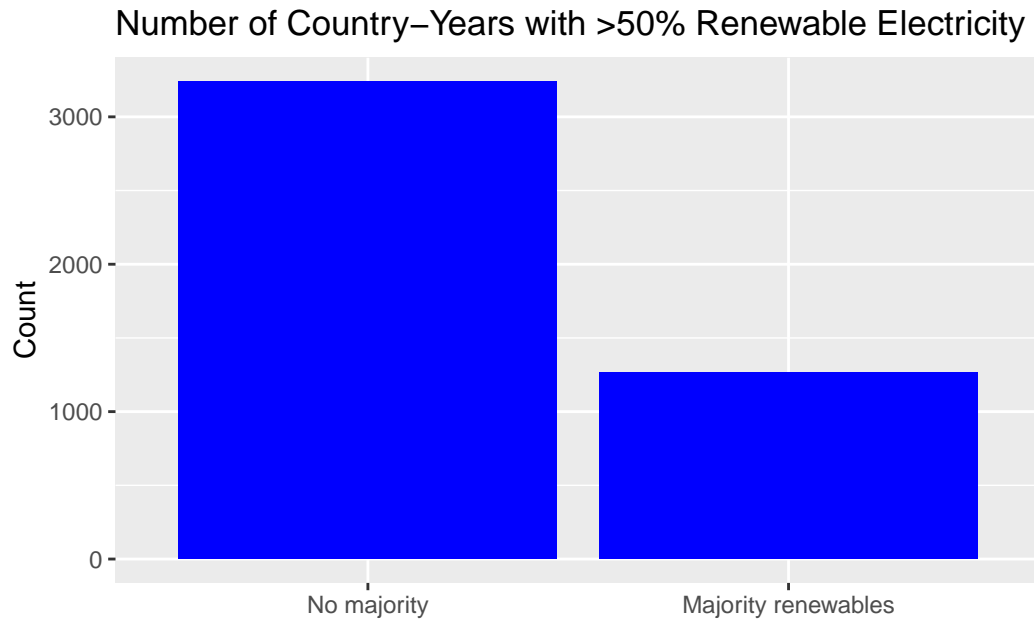
```
Rows: 4,504
Columns: 9
$ country               <chr> "Afghanistan", "Afghanistan", "Afghanistan", "A~
$ year                  <dbl> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007,~
$ renewables_share_elec <dbl> 64.583, 72.464, 78.873, 69.231, 70.886, 71.951,~
$ fossil_share_elec     <dbl> 35.417, 27.536, 21.127, 30.769, 29.114, 28.049,~
$ electricity_generation <dbl> 0.48, 0.69, 0.71, 0.91, 0.79, 0.82, 0.90, 1.01,~
$ gdp                   <dbl> 11283793920, 11021273088, 18804871168, 21074343~
$ population            <dbl> 20130279, 20284252, 21378081, 22733007, 2356059~
$ renewable_majority    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ gdp_per_capita        <dbl> 560.5384, 543.3414, 879.6333, 927.0372, 947.877~
```

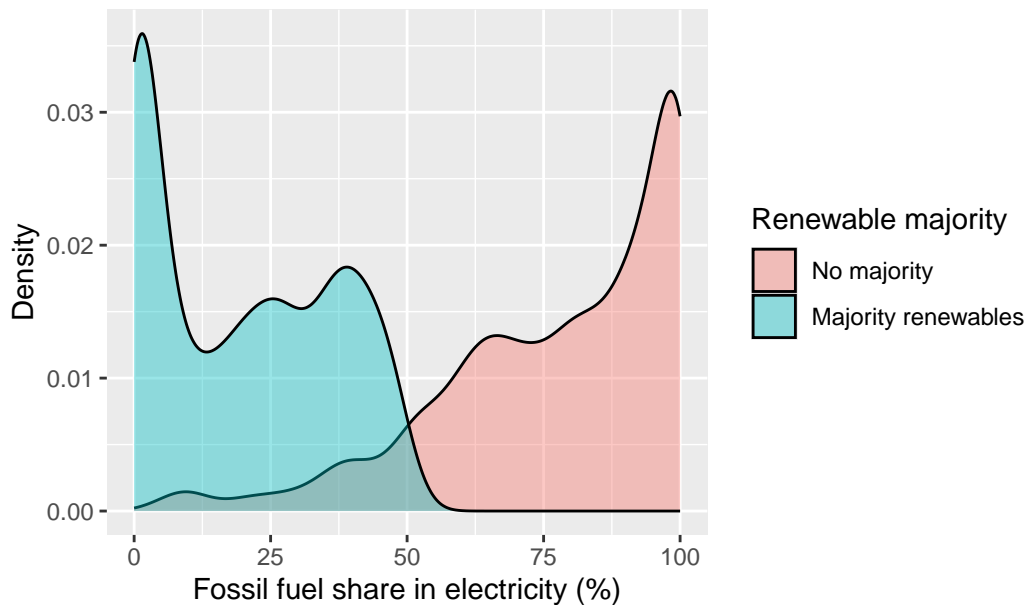**Exploratory Plots:**

```
ggplot(energy_rq2, aes(x = factor(renewable_majority,
                                  labels = c("No majority", "Majority renewables")))) +
  geom_bar(fill = "blue") +
  labs(
    title = "Number of Country-Years with >50% Renewable Electricity",
    x = "",
    y = "Count"
  )
```

Number of Country–Years with >50% Renewable Electricity


```
ggplot(energy_rq2,
       aes(x = fossil_share_elec,
           fill = factor(renewable_majority,
                         labels = c("No majority", "Majority renewables")))) +
  geom_density(alpha = 0.4) +
  labs(
    title = "Distribution of Fossil Fuel Share by Renewable-Majority Status",
    x = "Fossil fuel share in electricity (%)",
    y = "Density",
    fill = "Renewable majority"
  )
```

Distribution of Fossil Fuel Share by Renewable–Majority Statu

Interpretation:

The bar chart reveals that achieving a renewable-majority electricity mix is still relatively uncommon globally. Most countries in most years continue to rely primarily on fossil fuel–based electricity generation. This suggests that while renewable adoption has grown, the global energy transition remains incomplete.

The density plot clearly shows that countries with a renewable-majority electricity mix have very low fossil fuel shares, while those without such a majority rely heavily on fossil-based generation. The relationship appears strongly negative and nonlinear.

Dataset 2: General Social Survey (GSS)

Data source: General Social Survey (GSS), 2018 Wave

Brief description: The GSS is a nationally representative survey of U.S. adults conducted biennially since 1972. Each observation represents one respondent and includes demographic characteristics (e.g., age, race, education), socioeconomic indicators (e.g., income, occupational prestige, perceived class), and attitudinal measures (e.g., political views, confidence in institutions, happiness).

The 2018 wave contains over 2,300 respondents and several hundred variables of mixed types (continuous, ordinal, and categorical) making it suitable for regression-based inference.

Research question 1: *Is perceived social class associated with annual income after controlling for race and education?*

- **Outcome:** `realinc` – respondent's annual income in inflation-adjusted U.S. dollars (continuous)

  - **Predictors:** `class` (perceived social class), `race`, `educ`

  - **Model type:** Multiple linear regression

  - **Rationale:** This question examines whether self-perceived class accurately reflects objective income once demographic and educational differences are considered, providing inference on the social meaning of class identity.

```
# run the below lines before loading data
install.packages("remotes")
```

```
Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.4'
(as 'lib' is unspecified)
```

```
remotes::install_github("kjhealy/gssr")
```

```
Skipping install of 'gssr' from a github remote, the SHA1 (827d7f0b) has not changed sin
  Use `force = TRUE` to force installation
```

```
library(gssr)
```

```
Package loaded. To attach the GSS data, type data(gss_all) at the console.
For the panel data and documentation, type e.g. data(gss_panel08_long) and data(gss_pane
For help on a specific GSS variable, type ?varname at the console.
```

```
data <- gss_get_yr(2018)
```

```
Fetching: https://gss.norc.org/documents/stata/2018_stata.zip
```

```
head(data, 10)
```

```
# A tibble: 10 x 1,136
   year        id wrkstat   hrs1       hrs2       evwork      wrkslf  wrkgovt
   <dbl+lbl> <dbl> <dbl+lbl> <dbl+lbl>  <dbl+lbl>  <dbl+lbl>   <dbl+l> <dbl+l>
 1 2018         1 3 [with ~ NA(i) [iap]    41      NA(i) [iap] 2 [som~ 2 [pri~
 2 2018         2 5 [retir~ NA(i) [iap] NA(i) [iap]    1 [yes] 2 [som~ 2 [pri~
 3 2018         3 1 [worki~    40      NA(i) [iap] NA(i) [iap] 2 [som~ 2 [pri~
 4 2018         4 1 [worki~    40      NA(i) [iap] NA(i) [iap] 2 [som~ 2 [pri~
```

```
 5 2018          5 5 [retir~ NA(i) [iap] NA(i) [iap]      1 [yes] 2 [som~ 2 [pri~
 6 2018          6 5 [retir~ NA(i) [iap] NA(i) [iap]      1 [yes] 2 [som~ 2 [pri~
 7 2018          7 1 [worki~    35       NA(i) [iap] NA(i) [iap] 2 [som~ 1 [gov~
 8 2018          8 1 [worki~    89 [89+~ NA(i) [iap] NA(i) [iap] 2 [som~ 2 [pri~
 9 2018          9 1 [worki~    40       NA(i) [iap] NA(i) [iap] 1 [sel~ 2 [pri~
10 2018         10 1 [worki~    40       NA(i) [iap] NA(i) [iap] 2 [som~ 2 [pri~
# i 1,128 more variables: occ10 <dbl+lbl>, prestg10 <dbl+lbl>,
#   prestg105plus <dbl+lbl>, indus10 <dbl+lbl>, marital <dbl+lbl>,
#   martype <dbl+lbl>, divorce <dbl+lbl>, widowed <dbl+lbl>,
#   spwrksta <dbl+lbl>, sphrs1 <dbl+lbl>, sphrs2 <dbl+lbl>, spevwork <dbl+lbl>,
#   cowrksta <dbl+lbl>, cowrkslf <dbl+lbl>, coevwork <dbl+lbl>,
#   cohrs1 <dbl+lbl>, cohrs2 <dbl+lbl>, spwrkslf <dbl+lbl>, spocc10 <dbl+lbl>,
#   sppres10 <dbl+lbl>, sppres105plus <dbl+lbl>, spind10 <dbl+lbl>, ...
```

```r
data_clean <- data %>%
  select(realinc, class, race, educ, confed) %>%
  filter(!is.na(realinc),                        # remove missing income
         !is.na(class),
         !is.na(race),
         !is.na(educ),
         !is.na(confed),
         realinc > 0) %>%                         # drop invalid income values
  mutate(
    class = factor(class),                        # convert categorical vars to fac
    race = factor(race),
    confed = factor(confed,
                    levels = c(1, 2, 3),
                    labels = c("Hardly any", "Some", "A great deal"))
  )

glimpse(data_clean)
```

```
Rows: 1,410
Columns: 5
$ realinc <dbl+lbl>  14755.0,  72640.0, 119879.4, 119879.4,   8512.5,  24970.0~
$ class    <fct> 2, 3, 3, 4, 2, 2, 1, 3, 2, 2, 1, 2, 3, 3, 3, 3, 3, 3, 2, 1,~
$ race     <fct> 1, 1, 1, 2, 2, 1, 1, 1, 3, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1,~
$ educ     <dbl+lbl> 10, 16, 16, 18, 13, 12,  8, 12, 14, 16, 12, 16, 20, 14,  8~
$ confed   <fct> Hardly any, Some, A great deal, Some, Some, A great deal, Some~
```

Load the data and provide a `glimpse()`:

9

```
# same as above
data_clean <- data %>%
  select(realinc, class, race, educ, confed) %>%
  filter(!is.na(realinc),                          # remove missing income
         !is.na(class),
         !is.na(race),
         !is.na(educ),
         !is.na(confed),
         realinc > 0) %>%                           # drop invalid income values
  mutate(
    class = factor(class),                          # convert categorical vars to factors
    race = factor(race),
    confed = factor(confed,
                  levels = c(1, 2, 3),
                  labels = c("Hardly any", "Some", "A great deal"))
  )

glimpse(data_clean)
```

```
Rows: 1,410
Columns: 5
$ realinc <dbl+lbl>  14755.0,  72640.0, 119879.4, 119879.4,    8512.5,   24970.0~
$ class    <fct> 2, 3, 3, 4, 2, 2, 1, 3, 2, 2, 1, 2, 3, 3, 3, 3, 3, 3, 2, 1,~
$ race     <fct> 1, 1, 1, 2, 2, 1, 1, 1, 3, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1,~
$ educ     <dbl+lbl> 10, 16, 16, 18, 13, 12,  8, 12, 14, 16, 12, 16, 20, 14,  8~
$ confed   <fct> Hardly any, Some, A great deal, Some, Some, A great deal, Some~
```
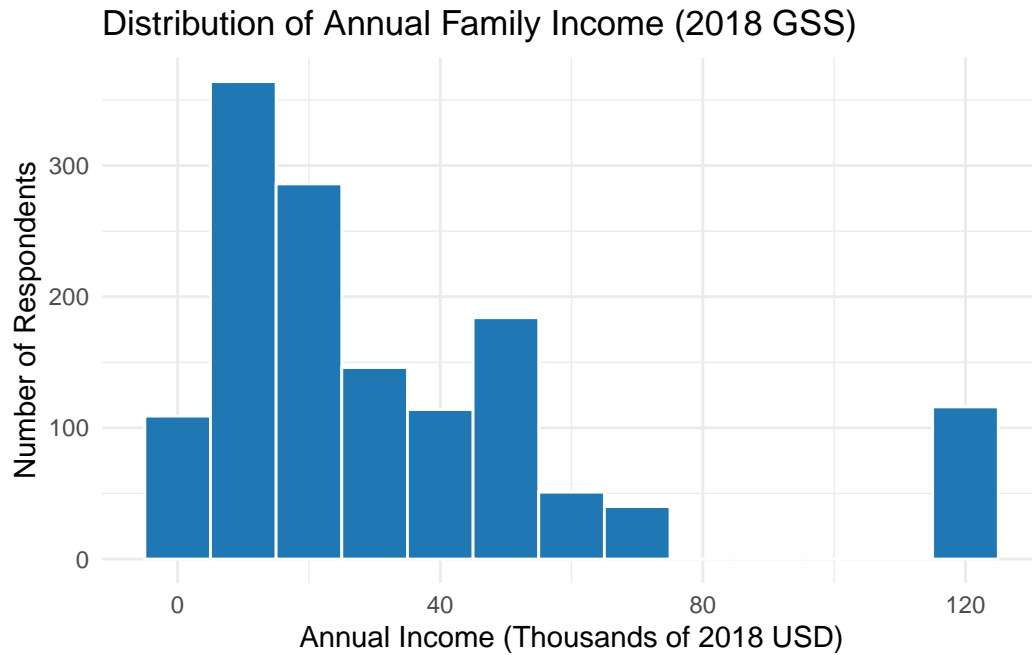
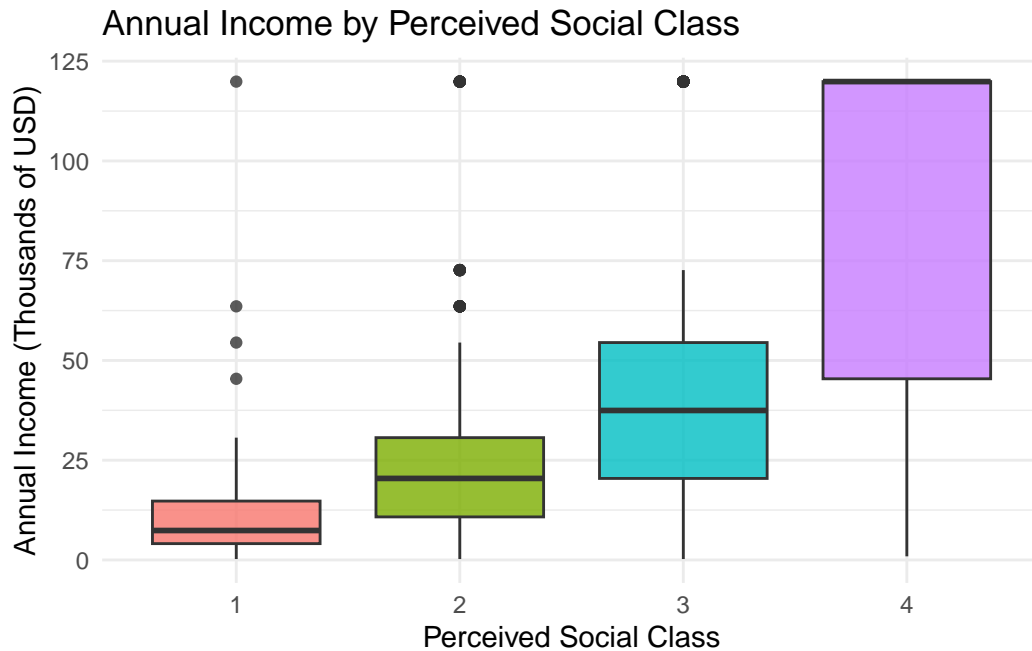**Exploratory Plots:**

a) Distribution of Income

```
ggplot(data_clean, aes(x = realinc / 1000)) +
  geom_histogram(binwidth = 10, fill = "#1f77b4", color = "white") +
  labs(
    title = "Distribution of Annual Family Income (2018 GSS)",
    x = "Annual Income (Thousands of 2018 USD)",
    y = "Number of Respondents"
  ) +
  theme_minimal()
```

Distribution of Annual Family Income (2018 GSS)

b) Income by Perceived Social Class

```r
# factor as "Lower," "Working," "Middle," "Upper"
ggplot(data_clean, aes(x = factor(class), y = realinc / 1000, fill = factor(class))) +
  geom_boxplot(alpha = 0.8) +
  labs(
    title = "Annual Income by Perceived Social Class",
    x = "Perceived Social Class",
    y = "Annual Income (Thousands of USD)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

Annual Income by Perceived Social Class

Interpretation:

The histogram shows that annual family income (`realinc`) in the 2018 GSS is strongly right-skewed, with most respondents earning under $100,000 and a small number with much higher incomes. The boxplot indicates a clear upward trend in income across perceived social classes—median income rises steadily from lower to upper class, suggesting that individuals who see themselves as higher class generally report higher actual earnings.

Research question 2: *Does confidence in institutions (e.g., the federal government) differ by race and social class?*
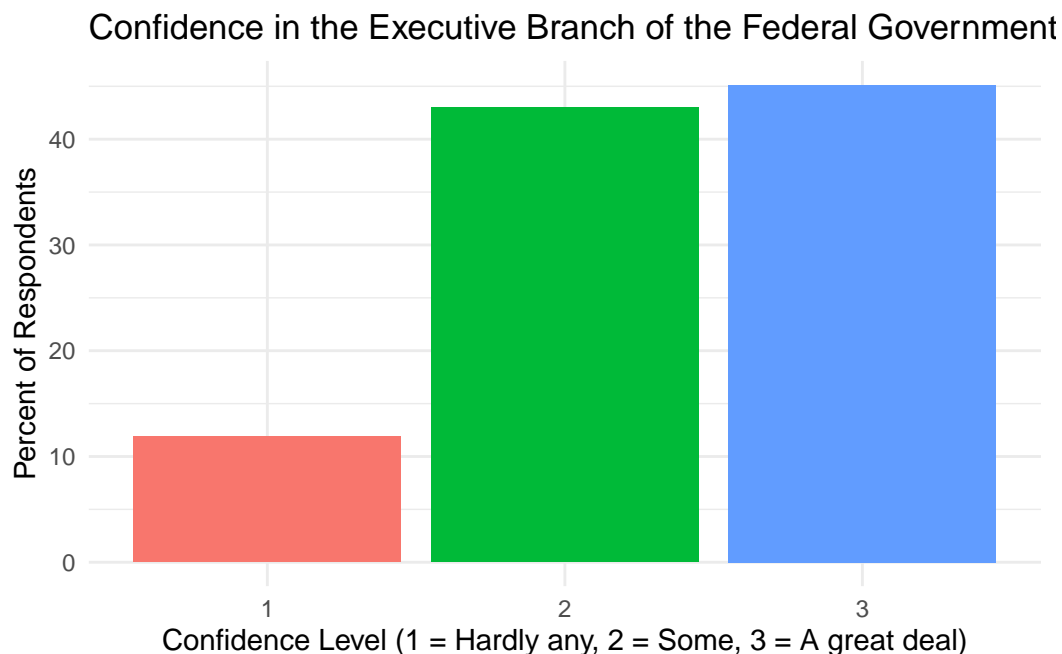
- Outcome variable (include the name/description and type of variable):
- Ordinal outcome with interaction
- **Outcome:** `confed` – confidence in government or major institutions (ordinal: 1 = hardly any, 2 = some, 3 = a great deal)
- **Predictors:** `race`, `class`, and the interaction `race × class`
- **Model type:** Ordinal logistic regression

- **Rationale:** This explores how structural and perceived social positions jointly influence institutional trust, testing whether the association between class and confidence varies across racial groups.

Exploratory Plots:

a) Distribution of confidence in executive branch of federal government
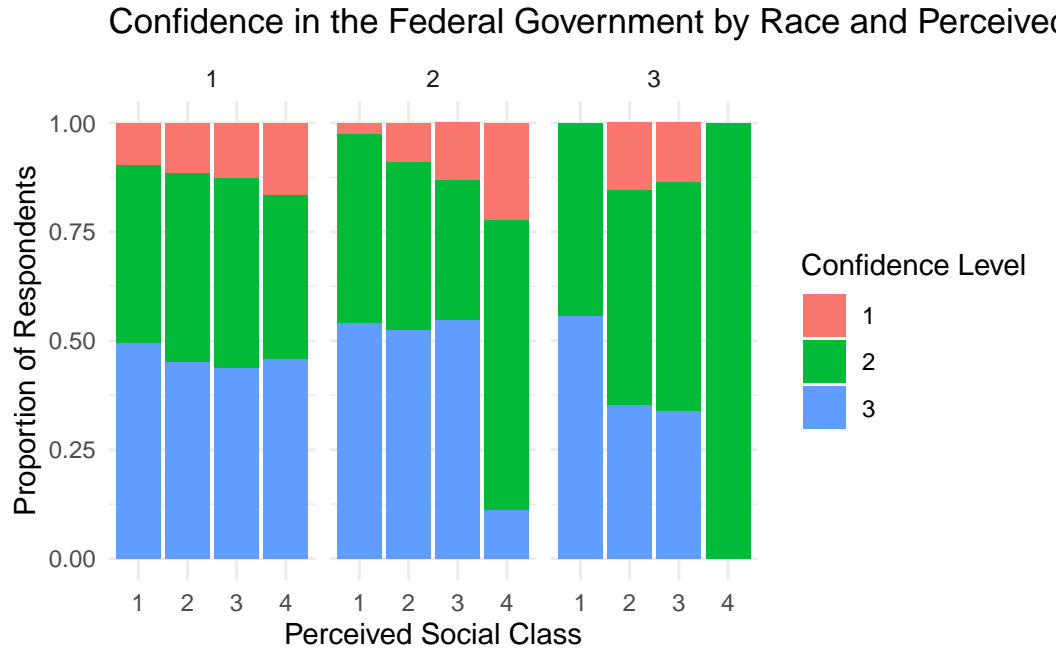
```
data_conf <- data %>%
  filter(!is.na(confed))

data_conf %>%
  count(confed) %>%
  mutate(pct = n / sum(n) * 100) %>%
  ggplot(aes(x = factor(confed), y = pct, fill = factor(confed))) +
  geom_col() +
  labs(
    title = "Confidence in the Executive Branch of the Federal Government (2018 GSS)",
    x = "Confidence Level (1 = Hardly any, 2 = Some, 3 = A great deal)",
    y = "Percent of Respondents"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```


Confidence in the Executive Branch of the Federal Government

b) Confidence by race and social class

```
data_conf <- data %>%
  filter(!is.na(confed), !is.na(race), !is.na(class))

ggplot(data_conf, aes(x = factor(class), fill = factor(confed))) +
  geom_bar(position = "fill") +
  facet_wrap(~ race) +
  labs(
    title = "Confidence in the Federal Government by Race and Perceived Social Class",
    x = "Perceived Social Class",
    y = "Proportion of Respondents",
    fill = "Confidence Level"
  ) +
  theme_minimal()
```


Confidence in the Federal Government by Perceived

Interpretation:

The bar chart of `confed` reveals that most respondents report only *some* confidence in the federal government, with fewer expressing *a great deal* of trust. When examined by race and perceived social class, the stacked bar plot shows variation in trust patterns: higher social

classes tend to display greater confidence overall, but this relationship differs by race, implying that the link between class and institutional trust may depend on racial group membership.

Team Charter

When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?

- Our team will meet weekly to work on the project, setting up a time block every week and canceling if needed. We will meet both in person and virtually, likely in person for the first few weeks and then virtually after Thanksgiving Break.

What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?

- At least a few hours notice should be given if missing meetings and the person missing the meeting should reach out proactively and catch themselves up, still assigning tasks to themselves.

How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?

- Our team communicates through a group message chat, expecting responses within the day. We will all be responsive and communicate with each other effectively. In time periods where there are deadlines, our team will make sure to be proactive in the chat.