Problem Set 3

Data Visualisation for Social Scientists

Due: February 18, 2026

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Wednesday February 18, 2026. No late assignments will be accepted.

Canadian Election Study

The data for this problem set come from the Canadian Election Study (CES) in 2015. The main purpose of the study is to give a comprehensive picture of the Canadian election: why people vote as they do, what changes during campaigns and across elections, and how Canadian voting compares with that in other democracies.

Data Manipulation

1. Load the CES .csv file from GitHub into your global environment. Filter respondents to only include "high quality" participants:

   ```
   ces2015 <- ces2015 |> filter(discard == "Good quality")
   ```

2. Filter the dataset to those participants that answered the question about voting for the past election using `p_voted`. Consider respondents who gave a "Yes" answer as having voted, while "No" as not having voted. Treat "Don't know" and "Refused" as missing.

3. Create an age variable and group into categories (e.g., <30, 30-44, 45-64, 65+). Year of birth is in age (four-digit year).

```r

#Question1

# Load required libraries
library(tidyverse)
# Load CES 2015 data from local file
ces2015 <- read_csv("/Users/whiz/Desktop/whiz1/DataViz_2026/problemSets/PS03/
    my_answers/CES2015.csv")
# Keep only high-quality respondents
ces2015 <- ces2015 |>
  filter(discard == "Good quality")
# Check the dataset
glimpse(ces2015)

#Question2
# Filter and recode voting variable
ces2015 <- ces2015 |>
  mutate(
    voted = case_when(
      p_voted == "Yes" ~ 1,
      p_voted == "No" ~ 0,
      p_voted %in% c("Don't know", "Refused") ~ NA_real_,
      TRUE ~ NA_real_
    )
  ) |>
  filter(!is.na(voted))
# Check results
table(ces2015$voted)

#Question3
ces2015 <- ces2015 |>
  mutate(
    age = as.numeric(age),          # convert birth year to numeric
    age_years = 2015 - age,

    age_group = case_when(
      age_years < 30 ~ "<30",
      age_years >= 30 & age_years <= 44 ~ "30-44",
      age_years >= 45 & age_years <= 64 ~ "45-64",
      age_years >= 65 ~ "65+",
      TRUE ~ NA_character_
    )
  )

table(ces2015$age_group)
```

Table 1: Number of respondents by age group

Age_Group	Count
<30	389
30-44	893
45-64	1986
65+	1394

Data Visualization

1. Plot turnout rate by age group.

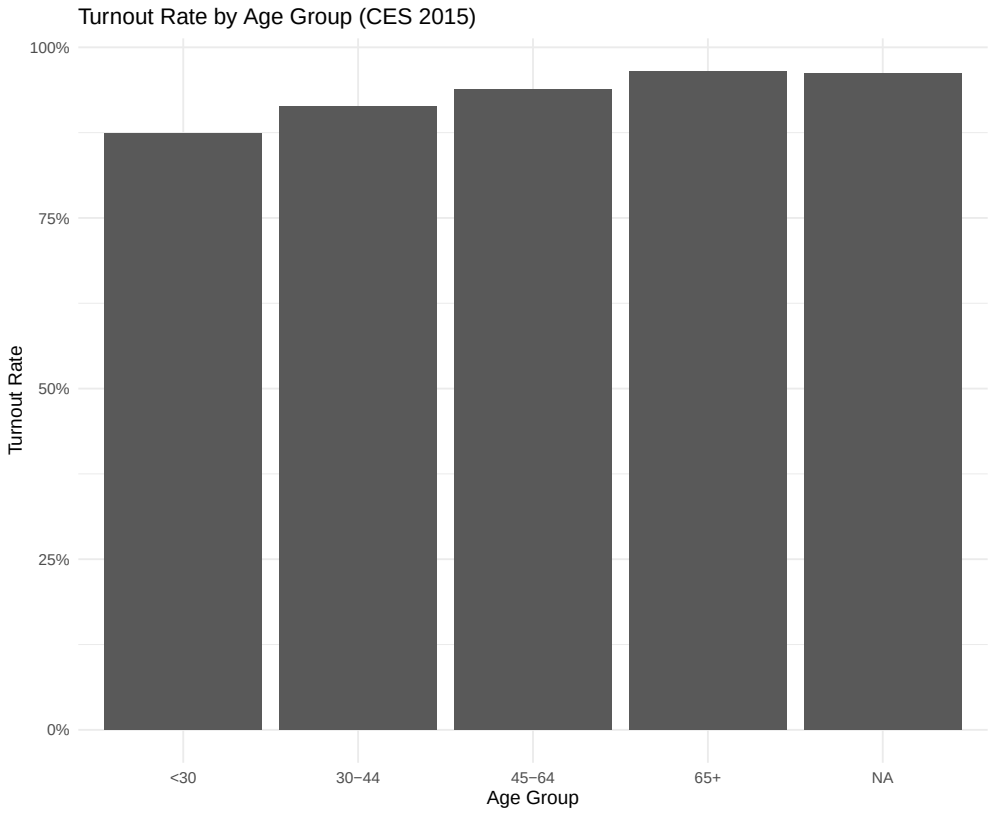

Figure 1: Turnout by age group

2. Create a density plot of ideology by party, restricting your sample to respondents with non-missing left–right self-placement (0–10 scale) and those that intended to vote for a main party (e.g., Liberal, Conservative, NDP, Bloc in Quebec, and Green).

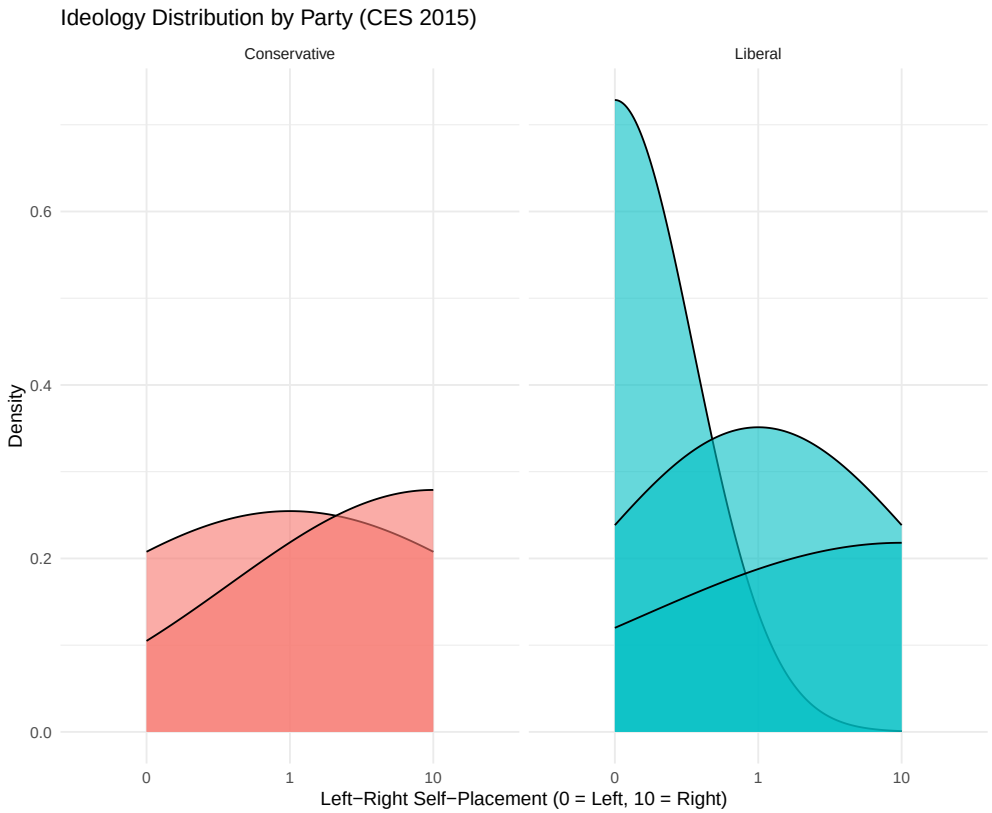

Figure 2: Turnout by age group

3. Produce histogram counts of turnout by income (`income_full`), faceted by province.

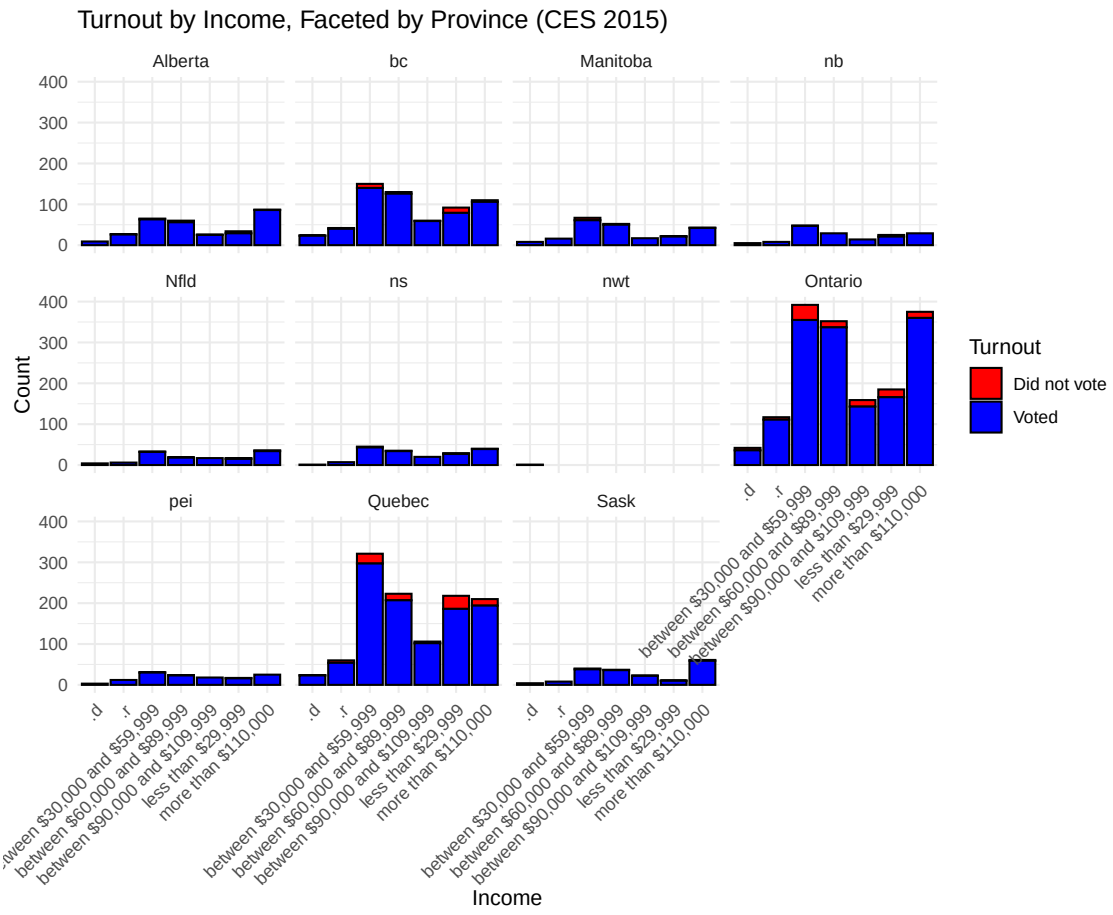

Figure 3: Turnout by age group

4. Create your own reusable custom theme. Apply your theme to one of the previous plots and add:

   (a) An improved title summarizing the main substantive takeaway.

   (b) A more informative subtitle describing the sample and variables.

   (c) A caption noting data source, weighting, and key coding decisions.

   (d) At least one direct annotation using `ggrepel` that calls out a key pattern.

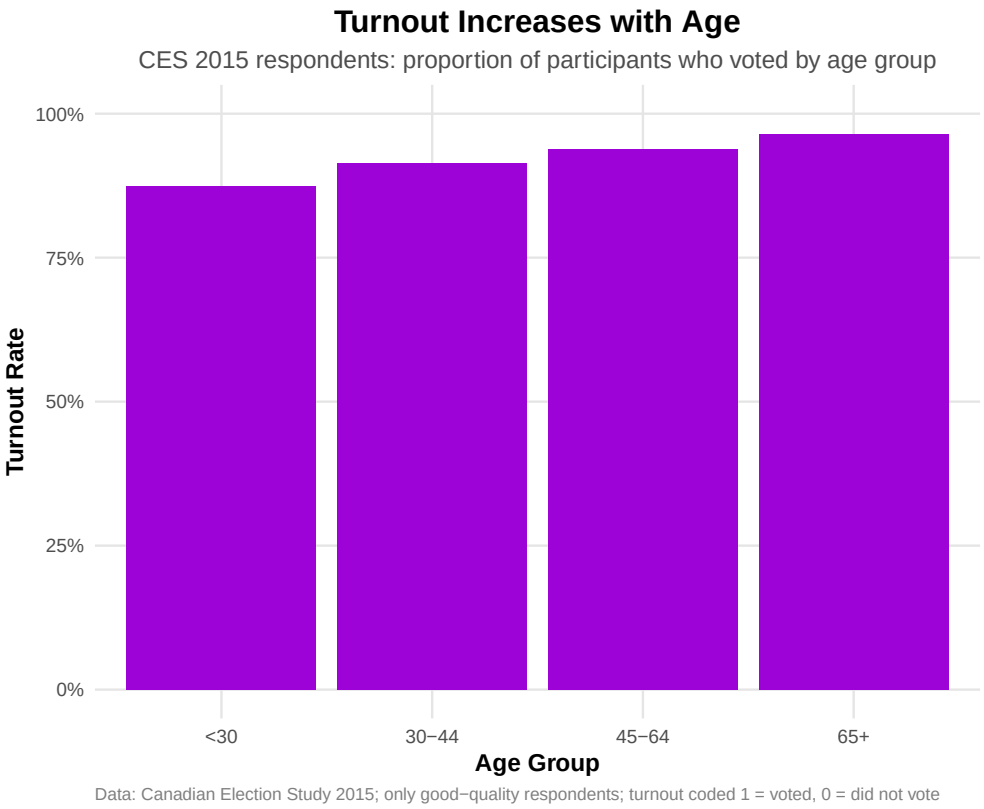

Figure 4: Turnout by age group