

Problem Set 3: Data Visualisation

Qi Liu

February 18, 2026

1 Data Preparation and Sample Restrictions

This analysis uses the Canadian Election Study (CES) 2015. Following the assignment requirements, I first restricted the sample to “Good quality” respondents using the `discard` indicator to reduce noise from low-quality interviews. Turnout was then recoded from the voting question (`p_voted`) into a binary variable where “Yes” equals 1 and “No” equals 0. Don’t know and Refused answers were treated as missing. Age was computed from the birth year (2015 – birth year) and grouped into four standard categories.

```
1  # Good quality filter (supports numeric or text codings)
2  discard_txt <- tolower(trimws(as.character(ces_raw$discard)))
3  discard_num <- suppressWarnings(as.numeric(as.character(ces_raw$
4  discard)))
5
6  good_quality <- dplyr::case_when(
7    !is.na(discard_num) ~ discard_num == 0,
8    discard_txt %in% c("good quality", "good_quality", "good") ~ TRUE,
9    TRUE ~ FALSE
10 )
11
12 ces0 <- ces_raw %>% filter(good_quality)
13
14 # Turnout recode (supports numeric or text codings)
15 voted_txt <- tolower(trimws(as.character(ces0$p_voted)))
16 voted_num <- suppressWarnings(as.numeric(as.character(ces0$p_voted)))
17
18 ces1 <- ces0 %>%
19   mutate(turnout = case_when(
20     !is.na(voted_num) & voted_num == 1 ~ 1,
21     !is.na(voted_num) & voted_num == 5 ~ 0,
22     !is.na(voted_num) & voted_num %in% c(8,9) ~ NA_real_,
23     voted_txt %in% c("yes", "y") ~ 1,
24     voted_txt %in% c("no", "n") ~ 0,
25     TRUE ~ NA_real_
26   )) %>%
27   filter(!is.na(turnout))
28
29 # Age groups (age column treated as birth year)
```

```

29 ces2015 <- ces1 %>%
30 mutate(birth_year = suppressWarnings(as.numeric(as.character(age))),
31 birth_year = if_else(birth_year >= 1900 & birth_year <= 2005, birth_
year, NA_real_),
32 age_years = 2015 - birth_year,
33 age_group = case_when(
34 age_years < 30 ~ "<30",
35 age_years <= 44 ~ "30-44",
36 age_years <= 64 ~ "45-64",
37 age_years >= 65 ~ "65+",
38 TRUE ~ NA_character_
39 ),
40 age_group = factor(age_group, levels = c("<30", "30-44", "45-64", "65+"))
41 ) %>%
42 filter(!is.na(age_group))

```

Listing 1: Data Cleaning and Recoding (matching PS03.R)

2 Data Visualization

2.1 Turnout Rate by Age Group

Figure 1 displays the mean turnout rate for each age category. The results reveal a strong monotonic relationship: electoral participation increases with age. Substantively, this aligns with political behavior theories suggesting that older citizens possess higher community stakes and more established voting habits.

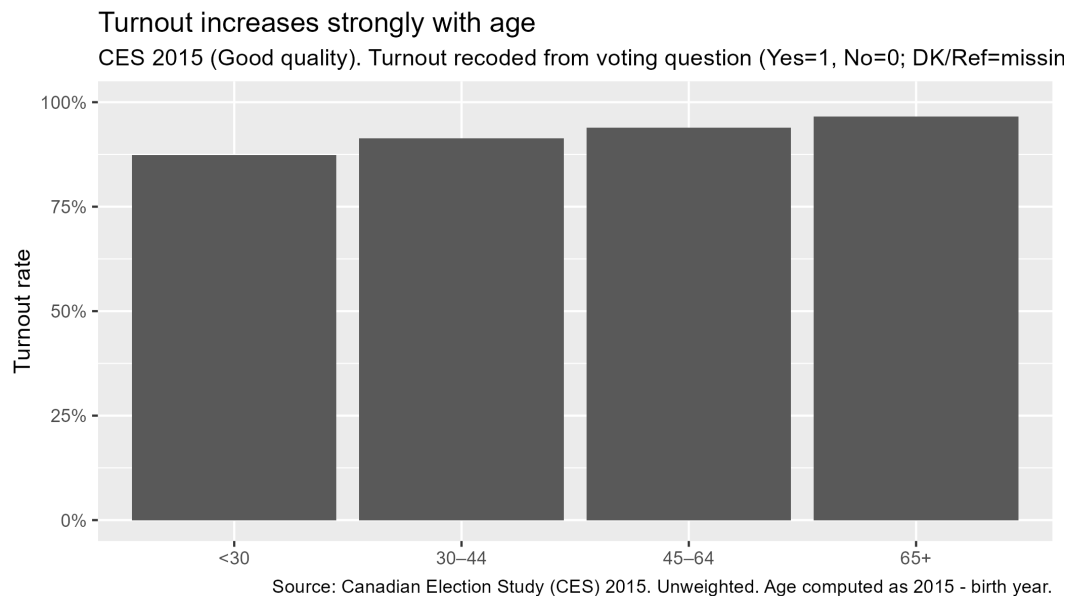


Figure 1: Turnout rate by age group (unweighted).

2.2 Ideological Self-Placement by Party

Figure 2 compares the distribution of left-right ideological self-placement (0-10) across major parties. The density curves show that Conservative supporters are concentrated on the right, NDP supporters on the left, and Liberal supporters cluster around the center-right.

```
1 party_num <- suppressWarnings(as.numeric(as.character(ces$vote_for)))
2 party_txt <- tolower(trimws(as.character(ces$vote_for)))
3
4 ces_ideo <- ces %>%
5 mutate(
6   selfplace = suppressWarnings(as.numeric(as.character(p_selfplace))),
7   party = case_when(
8     !is.na(party_num) & party_num == 1 ~ "Liberal",
9     !is.na(party_num) & party_num == 2 ~ "Conservative",
10    !is.na(party_num) & party_num == 3 ~ "NDP",
11    !is.na(party_num) & party_num == 4 ~ "Bloc Qu b cois",
12    !is.na(party_num) & party_num == 5 ~ "Green",
13    party_txt %in% c("liberal","liberals") ~ "Liberal",
14    party_txt %in% c("conservative","conservatives") ~ "Conservative",
15    party_txt %in% c("ndp","new democratic party") ~ "NDP",
16    party_txt %in% c("bloc","bloc quebecois","bloc qu b cois") ~ "Bloc
17    Qu b cois",
18    party_txt %in% c("green","greens") ~ "Green",
19    TRUE ~ NA_character_
20  ) %>%
21  filter(!is.na(selfplace), selfplace >= 0, selfplace <= 10) %>%
22  filter(!is.na(party))
23
```

Listing 2: Sample restrictions for Figure 2 (major parties)

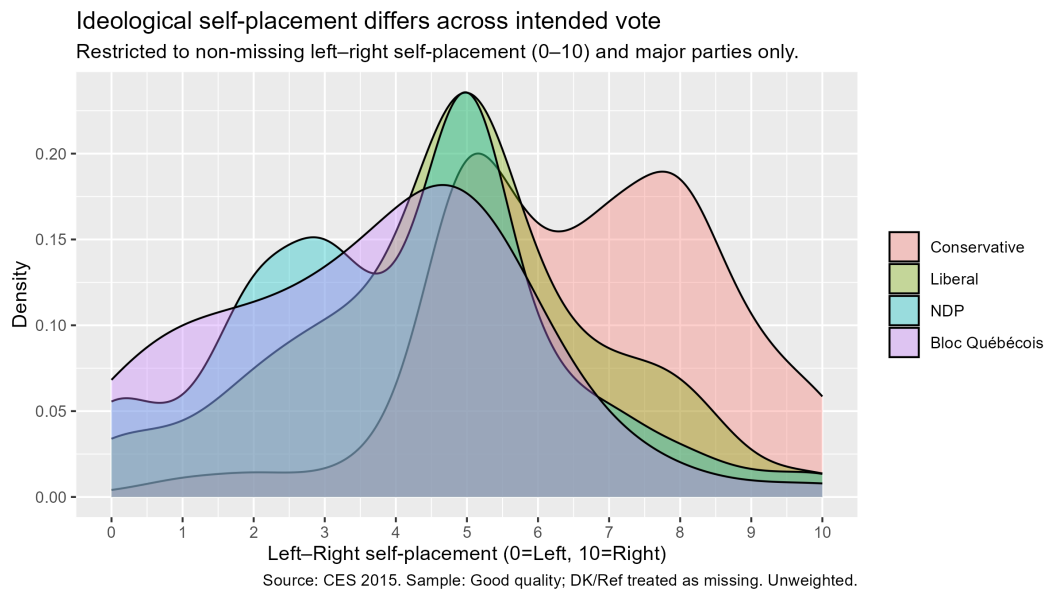


Figure 2: Density plot of ideological self-placement by intended vote.

2.3 Turnout by Income and Province

Figure 3 presents counts of turnout by household income, faceted by province. The plot illustrates that while turnout generally remains high across income groups in this "Good quality" sample, the raw counts are significantly influenced by the varying sample sizes across Canadian provinces (e.g., Ontario vs. PEI).

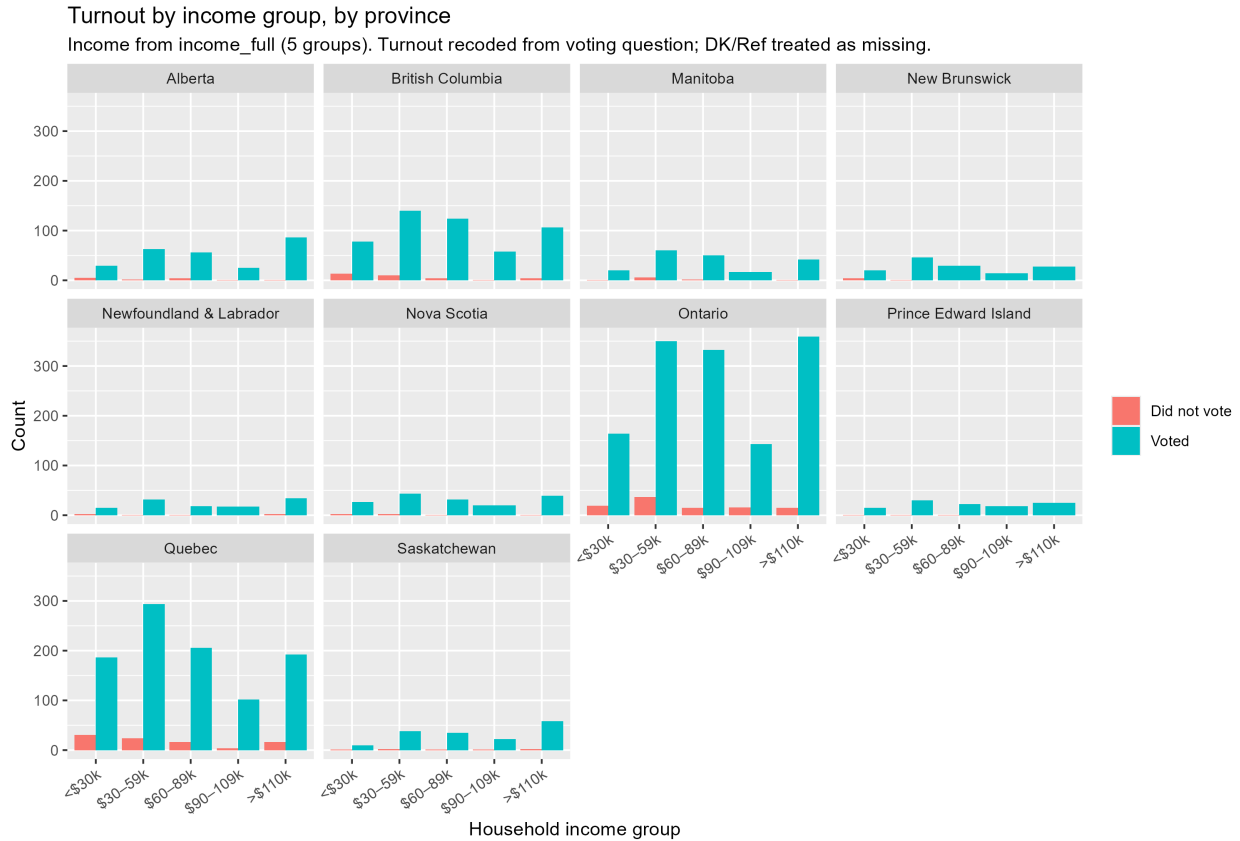


Figure 3: Histogram counts of turnout by income group, faceted by province.

2.4 Polished Visualization

For the final requirement, I applied a custom theme `theme_ps3()` to the age-turnout plot. This version includes direct annotations using `ggrepel` to highlight the participation gap between the youngest and oldest cohorts, making the substantive takeaway immediately clear to the reader.

```
1 theme_ps3 <- function() {
2   theme_minimal(base_size = 12) +
3   theme(
4     plot.title.position = "plot",
5     plot.title = element_text(face = "bold", size = 15),
6     plot.subtitle = element_text(size = 11),
7     plot.caption = element_text(size = 9, color = "grey30"),
8     panel.grid.minor = element_blank(),
9     legend.position = "top"
10  )
11 }
12
13 p1_final <- p1 + theme_ps3() +
14 ggrepel::geom_label_repel(
```

```

15   data = turnout_by_age %>% filter(age_group %in% c(gap_df$min_group,
16   gap_df$max_group)),
17   aes(label = paste0(as.character(age_group), ": ", scales::percent(
18   turnout_rate, accuracy = 1))),
19   nudge_y = 0.08,
20   show.legend = FALSE
  )

```

Listing 3: Reusable custom theme + direct annotation (ggrepel)

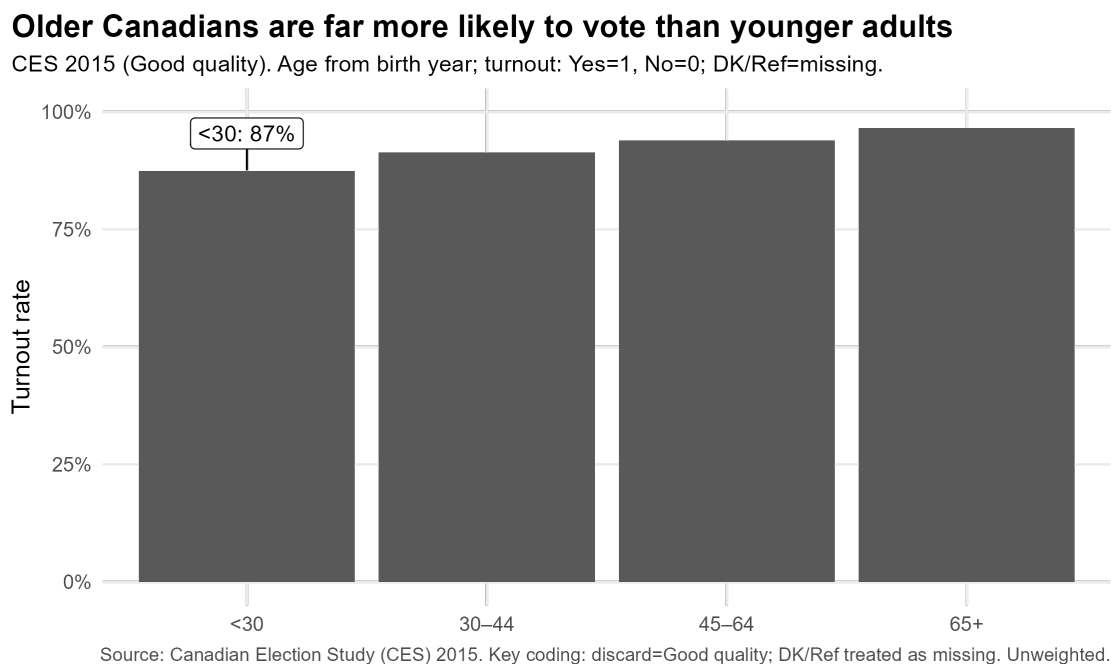


Figure 4: Final Polished Plot: Age and Turnout Engagement.

3 Reflection and Limitations

These descriptive visualizations suggest strong associations, particularly between age and turnout. However, they do not imply causality. Factors such as education and political interest likely confound these relationships. Furthermore, using counts in Figure 3 makes provincial comparisons difficult; using within-province proportions would be a more robust approach for future research.