````
---
title: "Problem Set 1"
name: Clare Porter
subtitle: "Due date: Monday, 15 September"
format:
  html:
    self-contained: true
editor: visual
execute:
  echo: true
  warning: false
---

Please upload your completed assignment to the ELMs course site (under the assignments menu) and to your class Github repository. You need to upload both your Quarto document (with R code, as needed) and its rendered output (either as a PDF or HTML file).

**Total points: 25**

::: callout-note
To complete this problem set, you will need access to the `polisciols` R package. This package is not published on CRAN[^1], so you will need to install it using the following code:

```{r}
#| eval: false

install.packages("devtools")

devtools::install_github("hgoers/polisciols")
```

Remember, you only need to do this once on your computer. Run this in the console. Do not include it in your Quarto document.
:::

[^1]: The [Comprehensive R Archive Network (CRAN)](https://cran.r-project.org/) hosts many R packages that can be installed easily using the familiar `install.packages()` function. These packages have gone through a comprehensive quality assurance process. I wrote `polisciols` for this class and will update it regularly. I, therefore, will not host it through CRAN: the quality assurance process takes too long to be practical for our weekly schedule. Instead, you are downloading it directly from its Github repository.

The [American National Election Studies](https://electionstudies.org/) surveys a representative sample of US voters prior to and following each US Presidential Election. We will use their survey from the last US Presidential Election (held in 2024) to learn more about US voters.

::: callout-tip
The `nes2024` data set in the `polsciols` R package provides you with a selection of the hundreds of questions the ANES asked of all `r nrow(polisciols::nes2024) |> scales::comma()` respondents to the 2024 survey. Each row provides data for one respondent.
:::

**For this problem set, only include observations for which full responses were recorded (exclude those who, for any question included in the dataset, refused to answer it, didn't know the answer, or have missing responses).**

## Question 1

Prior to the 2024 US Presidential Election, the ANES asked respondents their highest level of school completed or highest degree received.
````

Part A

Points: 2

What was the most common highest level of education obtained by respondents? What was the least common? *Use the `education` variable from the `nes2024` data set to answer these questions.*

"Some post-high school, no bachelor's degree" is the highest level of education obtained by respondents. "Less than high school credential" was the least common level of education obtained by respondents (excluding respondents who did not answer, refused to answer, or were unsure of their education level).

```{r}
## Save dataset + edu variable
polisciols::nes2024
nes2024 <- nes2024
education <- data$education

## Preview proportion of respondents in each education level (numerically and visually)
tabyl(nes2024, education)

ggplot(nes2024, aes(y = education)) +
  geom_bar() +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"),
        plot.title.position = "plot") +
  labs(
    x = "Count of respondents",
    y = NULL,
    caption = "Source: ANES 2024 Survey"
  ) +
  scale_x_continuous(labels = scales::label_comma())

```

Part B

Points: 2

Describe the distribution of the total number of respondents that obtained each highest level of education. Include a plot of those counts.

Using the code in Part A, I find the following count of respondents which have obtained each level of education:

- Less than high school credential: 283

- High school credential: 973

- Some post-high school, no bachelor’s degree: 1738

- Bachelor’s degree: 1384

- Graduate degree: 1087

Below is the plot of highest education level achieved produced using the code in Part A:

This distribution of education level is unimodal, peaking post-high school without a bachelors degree and skewed left towards the highest level of education.

Question 2

The ANES also asked each respondent how often they pay attention to what is going on in government and politics. Let's use these responses to learn more about how much attention US voters pay to politics ahead of a Presidential election.

Part A

Points: 2

How often did individuals pay attention to what was going on in government and politics? What was the most popular level of attention? What was the least? *Use the `attention_to_politics` variable from the `nes2024` data set to answer these questions.*

About 98% of individuals pay at least some attention to what was going on in government and politics, with respondents on average paying attention "most of the time."

The majority of respondents (approx. 37%) pay attention to what is going on in government and politics "most of the time."

"Never" paying attention to what was going on in government and politics is the least popular level of attention (approx. 1.8%).

```{r}
## Preview amount of time respondents pay attention to gov.t and politics
tabyl(nes2024, attention_to_politics)

ggplot(nes2024, aes(y = attention_to_politics)) +
  geom_bar() +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"),
        plot.title.position = "plot") +
  labs(
    x = "Count of respondents",
    y = NULL,
    caption = "Source: ANES 2024 Survey"
  ) +
  scale_x_continuous(labels = scales::label_comma())

## By converting this categorical data to numeric data, we can confirm our observations by calculating the mean and median
nes2024$attention_to_politics <- as.factor(nes2024$attention_to_politics)
nes2024$attention_to_politics_num <- as.numeric(nes2024$attention_to_politics)

# A note: as.numeric organizes based on alphabetical order: About = 1, Always = 2, Most = 3, Never = 4, Some = 5
mean(attention_to_politics_num, na.rm = T)
median(attention_to_politics_num, na.rm = T)

# A note: output tells us that the mean is the more informative statistic
```

Part B

Points: 3

Describe the distribution of the total number of respondents who provided each answer to this question. Does this distribution indicate that individuals tend to pay very little, a moderate, or a lot of attention to politics prior to a US Presidential Election? Include a plot of these counts in your answer.

Using the code in Part A, we can find the following count of respondents' attention level:

- Never: 101

- Some of the time: 1109

- About half the time: 1111

- Most of the time: 2055

- Always: 1143

The distribution of attention level is unimodal, peaking where respondents pay attention "most of the time" and skewed left towards a larger attention to politics. This distribution and mean indicates that individuals pay a lot of attention to politics prior to a US Presidential Election.

Below is the plot of respondents' attention level produced using the code in Part A:

Question 3

Examine the following vector of (fake) student IQ test scores:

```{r}
iq <- c(
  145, 139, 126, 122, 125, 130, 96, 110, 118, 118, 101, 142, 134, 124, 112, 109,
  134, 113, 81, 113, 123, 94, 100, 136, 109, 131, 117, 110, 127, 124, 106, 124,
  115, 133, 116, 102, 127, 117, 109, 137, 117, 90, 103, 114, 139, 101, 122, 105,
  97, 89, 102, 108, 110, 128, 114, 112, 114, 102, 82, 101
)
```

Part A

Points: 3

Find the five-number summary, mean, and standard deviation for these data. Do you think any of these students might be outliers in this distribution? If so, what IQ test score did they get and how do you know?

Using the code below, I find the following statistics of the above vector of IQ test scores:

- Mean: 115 (or 114.98)

- SD: 14.8

- Min: 81

- P25: 104

- Median: 114

- P75: 125

- Max: 145

Following the rule of thumb for outliers, which indicates an observation is an outlier if 1.5 times the IQR above P75 or Q3, or below P25 or Q1, I do not suspect an outlier exists in this data.

```{r}
mean(iq)
```

```
skim(iq)

## Turn list into data frame in order make a boxplot
iq_df <- as.data.frame(iq)

ggplot(iq_df, aes(x = iq)) +
  geom_boxplot() +
  theme_minimal() +
  theme(
    axis.text.y = element_blank()
  )

## Check for an Outlier
# IQR = P75 - P25 = 125 - 104 = 21
# Lower bound = 104 - (1.5x21) = 72.5; this is lower than the min in the data (81)
# Upper bound = 125 + (1.5x21) = 156.5; this is higher than the max in the data (145)
```

Part B

Points: 3

In large populations, IQ test scores are standardized to have a mean of 100 and a standard deviation of 15. In what way does the distribution among these students differ from the overall population? Do not use the words and phrases "mean" or "standard deviation" in your answer.

The IQ scores in this sample is generally higher than what is typically seen in the general population, with most students in the sample scoring above what is considered average. The variation amongst students in the sample is also slightly less than what is typical in the broader population, concentrating in the more upper range.

Question 4

Points: 4

The ANES collects information about each respondent's demographics, including their race, age, household income, and party affiliation. Provide the "type" (categorical or continuous) of variable each of those demographic characteristics represents.

All four demographic characteristic listed above were represented as categorical variables:

- Race: (non-Hispanic White, non-Hispanic Black, Hispanic, non-Hispanic other)

- Age: (18-29, 30-44, 45-59, 60 years or older)

- Household income: (less than \$40,000, \$40,000-\$74,999, \$75,000-\$124,999, \$125,000 or more)

- Party Affiliation: (Inapplicable, Democratic party, Republican party, None, Independent, Another Party)

(For good measure, the following additional demographic data was also categorical)

- Education attainment: (less than high school graduate, high school graduate, some college, college degree, graduate degree)

- Sex: (male, female)

- Home tenure (home owner, other)

- Nation of birth (U.S. born, foreign born)

-   Marital status (married, widowed, divorced or separated, never married)

-   Population density (less than 471 people/square mile, 471-2,377 people/square mile, 2,377-5,505 people/square mile, more than 5,505 people/square mile)

-   Metropolitan status (non-metropolitan, metropolitan)

Please note that all the respondent's demographic information mentioned in this question does not appear on this data set for me (see code below). However, the [ANES user guide code book](https://electionstudies.org/anes_timeseries_2024_csv_20250808/) provides ample information about metadata collected, and was used to help answer this question.

```{r}
#### Question 4 ####
glimpse(nes2024)
names(nes2024)
names(nes)

#Please note that respondent's demographic information does not appear
```

Question 5

The ANES also asked respondents prior to the election to rate Kamala Harris and Donald Trump on a scale from 0 to 100.

Part A

Points: 3

Plot all respondents' answers to these two questions on a single density plot. There are several peaks in these distributions. What do these peaks have in common? What might they suggest about how people answer questions that ask them to provide a single number along a large range?

The peaks and valleys of the below graph seem to fall on round numbers (e.g., 0, 25, 50, 75, 80, 100). This observation demonstrates how people tend to choose familiar numbers and round them off, treating a continuous data as categorical. Respondents chose numbers which felt meaningful and easy to interpret, expressing strong opinions (e.g., 0 or 100), or neutrality (e.g., 50), rather than using the scale to provide a more nuanced rating.

Below is a single density plot illustrating respondents rating of presidential candidates:

```{r}
## In order to plot two columns of data on a single density plot, I will need to transform the data
nes_new <- nes2024 %>%
  pivot_longer(
    cols = c(therm_harris, therm_trump),
    names_to = "candidate",
    values_to = "therm_rating"
  )

## Plot data with distinguishing colors
ggplot(nes_new, aes(x = therm_rating, color = candidate)) +
  geom_density() +
  theme_minimal() +
  labs(
    x = "Rating of Presidential Candidates",
  )  +
  scale_color_manual(values = c("therm_harris" = "blue", "therm_trump" = "red")) +
```

```
  theme(legend.title = element_blank())
```

Part B

Points: 3

President Trump won this election. Would you predict this outcome from the way this representative sample responded to these questions? Provide a brief description of the reasoning behind your answer.

Based on the data used in this exercise, no, I would not predict this outcome. After seeing a clear difference in averages and medians between the two candidates' ratings, I wanted to confirm my hypothesis by calculating the statistical significance of the difference in means. The result is a *very* small P value. As the P value is much smaller than 0.01, we can conclude with confidence that there was a meaningful difference in how respondents felt about the two candidates.

```{r}
skim(nes2024)
## We see clear differences in the average ratings and median between Trump and Harris
## Let's conduct a t-test to see if that difference is statistically significant

tttest <- t.test(nes_clean$therm_harris, nes_clean$therm_trump, paired = TRUE)
tttest
```