Exploratory Data Analysis

```
#Set Up
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.1     ✔ stringr   1.5.2
✔ ggplot2   4.0.0     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.1.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
g24_clean <- read_csv("data/g24_sov_by_g24_svprec_clean.csv")
```

```
Rows: 51123 Columns: 76
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr (49): fips, svprec, svprec_key, election, geo_type, assaip01,
assdem01, ...
dbl (27): county, addist, cddist, sddist, bedist, totreg, demreg, repreg,
ai...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
precinct_data <- g24_clean |>
    filter(!str_detect(svprec, "TOT"))
```

Question 1

Which county had the closest race?

Answer 1

```
county_results <- precinct_data |>
    group_by(county) |>
    summarize(dem_votes = sum(demvote, na.rm = TRUE), rep_votes = sum(repvote,
na.rm = TRUE)) |>
    mutate(margin = abs(dem_votes - rep_votes), total_votes = dem_votes +
rep_votes, pct_margin = margin /total_votes) |>
    arrange(margin)
closest_county <- county_results |>
    slice(1)
closest_county
```

```
# A tibble: 1 × 6
  county dem_votes rep_votes margin total_votes pct_margin
   <dbl>     <dbl>     <dbl>  <dbl>       <dbl>      <dbl>
1      1         0         0      0           0        NaN
```

Question 2

How many SV precincts are in the data set? What is the the min, max and median votes cast?

Answer 2

```
num_precincts <- nrow(precinct_data)
precinct_data |>
    summarize(min_total_vote = min(totvote, na.rm = TRUE), median_total_vote =
median(totvote, na.rm = TRUE), max_total_vote = max(totvote, na.rm = TRUE))
```

```
# A tibble: 1 × 3
  min_total_vote median_total_vote max_total_vote
           <dbl>             <dbl>          <dbl>
1              0                87           5812
```

Question 3

How does the total ballots vary across counties?

Answer 3

```
county_data <- precinct_data |>
    group_by(county) |>
    summarize(num_precincts = n(), median_total_vote = median(totvote, na.rm =
TRUE), mean_total_vote = mean(totvote, na.rm = TRUE)) |>
    arrange(desc(median_total_vote))
county_data
```

```
# A tibble: 57 × 4
   county num_precincts median_total_vote mean_total_vote
    <dbl>         <int>              <dbl>           <dbl>
 1     19          6233               430             609.
 2     32            37               281             276.
 3     45           144               274.            626.
 4      5            59               228             456.
 5     39           554               218.            483.
 6     58            79               204             385.
 7     51            56               202             712.
 8     43          1586               190             483.
 9     38          1044               183             395.
10      1          1158               181             590.
# ℹ 47 more rows
```