Exploratory Data Analysis

Question 1

Which precincts had the closest congressional race ### Answer 1

```
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
```

```
sov_2024 <- read_csv("data/g24_sov_by_g24_svprec.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_votes <- sov_2024 |>
    mutate(
        CNGDEM01 = as.numeric(CNGDEM01),
        CNGREP01 = as.numeric(CNGREP01)
    ) |>
```

```
    select(
        COUNTY,
        SVPREC_KEY,
        CDDIST,
        dem_votes = CNGDEM01,
        rep_votes = CNGREP01
    ) |>
    mutate(
        total_votes = dem_votes + rep_votes,
        dem_share = dem_votes / total_votes
    )
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
ℹ In argument: `CNGDEM01 = as.numeric(CNGDEM01)`.
Caused by warning:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
# Here we compute how close each Congressional race was by taking the

# absolute difference between Democratic and Republican votes. Smaller

# values mean closer races.

closest_precincts <- precinct_votes |>
mutate(
diff_congress = abs(dem_votes - rep_votes)
) |>
filter(total_votes > 0) |>
arrange(diff_congress) |>
select(SVPREC_KEY, COUNTY, CDDIST, dem_votes, rep_votes, diff_congress)

closest_precincts
```

```
# A tibble: 34,028 × 6
   SVPREC_KEY    COUNTY CDDIST dem_votes rep_votes diff_congress
   <chr>          <dbl>  <dbl>     <dbl>     <dbl>         <dbl>
 1 06001453900        1     14        17        17             0
 2 06001542300        1     14        12        12             0
 3 06001835830        1     14        67        67             0
 4 06005CP38A         3      5        37        37             0
 5 060074302          4      1        21        21             0
 6 060074807          4      1         6         6             0
 7 06013ANTI136       7     10         9         9             0
```

```
 8 06013BREN124       7     10         5         5             0
 9 06013BRHL101       7     10         5         5             0
10 06013BRHL106A      7     10        21        21             0
# ℹ 34,018 more rows
```

```
# The top rows show the precincts where the Congressional vote totals were the
closest.
```

Question 2

What is the range of total votes per precinct in the 2024 election?

Answer 2

```
# Here we summarize total precinct size using min, median, mean, and max.

# This tells us how large or small precincts were in terms of turnout.

precinct_size_stats <- precinct_votes |>
summarize(
min_votes = min(total_votes, na.rm = TRUE),
median_votes = median(total_votes, na.rm = TRUE),
mean_votes = mean(total_votes, na.rm = TRUE),
max_votes = max(total_votes, na.rm = TRUE)
)

precinct_size_stats
```

```
# A tibble: 1 × 4
  min_votes median_votes mean_votes max_votes
      <dbl>        <dbl>      <dbl>     <dbl>
1         0          108      2188.   3324096
```

This tells us the range of precinct sizes, smallest -> largest

Question 3

How many Congressional districts were won by Democrats vs Republicans?

Answer 3

```
# First we aggregate votes to the district level by summing all precincts

# within the same Congressional district. Then we identify which party won.

district_votes <- precinct_votes |>
group_by(CDDIST) |>
```

```r
summarize(
dem_votes = sum(dem_votes, na.rm = TRUE),
rep_votes = sum(rep_votes, na.rm = TRUE),
total_votes = sum(total_votes, na.rm = TRUE)
) |>
mutate(
winner = if_else(dem_votes > rep_votes, "DEM", "REP")
)

district_votes
```

```
# A tibble: 53 × 5
   CDDIST dem_votes rep_votes total_votes winner
    <dbl>     <dbl>     <dbl>       <dbl> <chr>
 1      0  52843210  34247705    87090915 DEM
 2      1    110472    208150      318622 REP
 3      2    272384    106407      378791 DEM
 4      3    187960    233895      421855 REP
 5      4    227321    114644      341965 DEM
 6      5    134467    214223      348690 REP
 7      6    165386    121625      287011 DEM
 8      7    197361     98273      295634 DEM
 9      8    201756     70932      272688 DEM
10      9    130093    121006      251099 DEM
# ℹ 43 more rows
```