Exploratory Data Analysis

```
── 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
Linking to GEOS 3.13.0, GDAL 3.8.5, PROJ 9.5.1; sf_use_s2() is TRUE

udunits database from /Library/Frameworks/R.framework/Versions/4.5-arm64/
Resources/library/units/share/udunits/udunits2.xml

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.
```

```
Warning: There were 44 warnings in `mutate()`.
The first warning was:
ℹ In argument: `across(...)`.
Caused by warning:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 43 remaining warnings.
```

Question 1

What are the top 3 counties in terms of proportion of votes for third party (non Democrat or Republican) canidates in the presidential election?

Answer 1

```
most_third_party <- precinct_election |>
  mutate(
    TOTAL_PRES = PRSDEM01 + PRSREP01 + PRSGRN01 + PRSLIB01 + PRSPAF01 +
PRSAIP01,
    third_party = TOTAL_PRES - (PRSDEM01 + PRSREP01)) |>
  group_by(COUNTY) |>
  summarize(
    third_party_prop = sum(third_party, na.rm = TRUE) /
                       sum(TOTAL_PRES, na.rm = TRUE)) |>
  arrange(desc(third_party_prop)) |>
  slice_head(n = 3)
most_third_party
```

```
# A tibble: 3 × 2
  COUNTY third_party_prop
   <dbl>            <dbl>
1     12           0.0446
2      1           0.0440
3     23           0.0417
```

Counties 12, 1, and 23 had the highest proportion of votes cast for third party candidates at 4.46%, 4.40%, and 4.17%, respectively.

Question 2

What is the distribution of Republican presidential vote share across precincts?

Answer 2

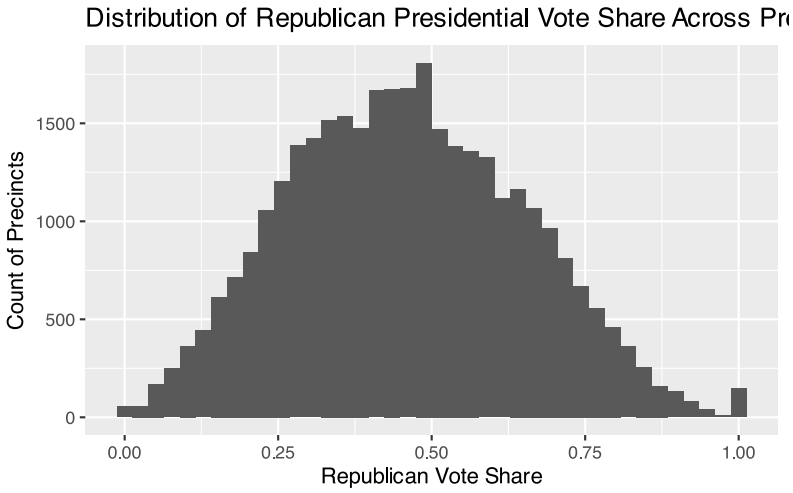
```
# Summary Statistics
precinct_election |>
  mutate(
    TOTAL_PRES = PRSDEM01 + PRSREP01 + PRSAIP01 + PRSGRN01 + PRSLIB01 +
PRSPAF01,
    rep_share = PRSREP01 / TOTAL_PRES) |>
  summarize(
    mean_rep_share = mean(rep_share, na.rm = TRUE),
    median_rep_share = median(rep_share, na.rm = TRUE),
    sd_rep_share = sd(rep_share, na.rm = TRUE),
    min_rep_share = min(rep_share, na.rm = TRUE),
    max_rep_share = max(rep_share, na.rm = TRUE),
    n_precincts = n())
```

```
# A tibble: 1 × 6
  mean_rep_share median_rep_share sd_rep_share min_rep_share max_rep_share
```

```
           <dbl>            <dbl>        <dbl>         <dbl>         <dbl>
1          0.459            0.454        0.191             0             1
# ℹ 1 more variable: n_precincts <int>
```

```
# Plot
precinct_election |>
  mutate(
    TOTAL_PRES = PRSDEM01 + PRSREP01 + PRSAIP01 + PRSGRN01 + PRSLIB01 +
PRSPAF01,
    rep_share = PRSREP01 / TOTAL_PRES) |>
  ggplot(aes(x = rep_share)) +
  geom_histogram(bins = 40) +
  labs(
    title = "Distribution of Republican Presidential Vote Share Across
Precincts",
    x = "Republican Vote Share",
    y = "Count of Precincts")
```

```
Warning: Removed 12467 rows containing non-finite outside the scale range
(`stat_bin()`).
```



The distribution of republican presidential votes was very slightly right skewed, and the mean proportion of votes for republicans in precincts was 0.459, suggesting that Trump was a minority in the majority of California precincts in the 2024 presidential election.

Question 3

What is the distribution of precinct sizes?

Answer 3

```
precinct_election |>
  mutate(
    precinct_size = PRSDEM01 + PRSREP01 + PRSAIP01 + PRSGRN01 + PRSLIB01 +
PRSPAF01) |>
  summarize(
    mean_precinct_size = mean(precinct_size, na.rm = TRUE),
    median_precinct_size = median(precinct_size, na.rm = TRUE),
    sd_precinct_size = sd(precinct_size, na.rm = TRUE),
    min_precinct_size = min(precinct_size, na.rm = TRUE),
    max_precinct_size = max(precinct_size, na.rm = TRUE),
    n_precincts = n())
```

```
# A tibble: 1 × 6
  mean_precinct_size median_precinct_size sd_precinct_size min_precinct_size
               <dbl>                <dbl>            <dbl>             <dbl>
1               352.                  116             570.                 0
# ℹ 2 more variables: max_precinct_size <dbl>, n_precincts <int>
```

The average precinct size is 352 and the standard deviation is 570 once filtered out precincts over 10,000 in size.