Exploratory Data Analysis

```r
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
```

```r
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.
```

Question 1

Which precincts behave like "outliers" in terms of voter turnout, and what does the overall turnout landscape of California precincts look like?

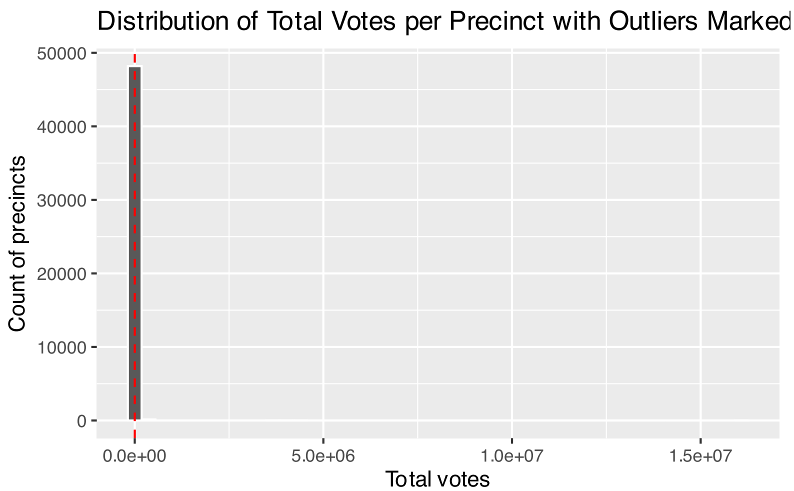
Answer 1

```r
# Combine polling-place + absentee votes to compute turnout
precinct_totals <- clean |>
```

```r
  group_by(SVPREC) |>
  summarise(total_votes = sum(TOTVOTE), .groups = "drop")

# Identify upper and lower outliers using Tukey fences
q1 <- quantile(precinct_totals$total_votes, 0.25)
q3 <- quantile(precinct_totals$total_votes, 0.75)
iqr <- q3 - q1
lower_fence <- q1 - 1.5 * iqr
upper_fence <- q3 + 1.5 * iqr

outliers <- precinct_totals |>
  filter(total_votes < lower_fence | total_votes > upper_fence)

outliers
```

```
# A tibble: 4,116 × 2
   SVPREC   total_votes
   <chr>          <dbl>
 1 0000001A        1442
 2 0000002A        2891
 3 0000003A        2570
 4 0000004A        1697
 5 0000005A        1395
 6 0000006A        1617
 7 0000007A        1174
 8 0000008A        1632
 9 0000009A        1716
10 0000010A        1452
# ℹ 4,106 more rows
```

```r
ggplot(precinct_totals, aes(x = total_votes)) +
  geom_histogram(bins = 45, color = "white") +
  geom_vline(xintercept = upper_fence, color = "red", linetype = "dashed") +
  labs(
    title = "Distribution of Total Votes per Precinct with Outliers Marked",
    x = "Total votes",
    y = "Count of precincts"
  )
```



Most precincts cluster in a predictable mid-range of turnout, but several stand out as "super-precincts" with very large vote totals. These are usually mail-ballot-heavy precincts or consolidated precincts in dense counties. Identifying these outliers now helps explain later distortions in district-level metrics.

Question 2

Which congressional districts show "vote fragmentation," meaning the Democratic vote is split between two Democratic candidates while Republicans consolidate behind one?

Answer 2

```r
cd_totals <- clean |>
  mutate(
    dem1 = replace_na(as.numeric(CNGDEM01), 0),
    dem2 = replace_na(as.numeric(CNGDEM02), 0),
    rep1 = replace_na(as.numeric(CNGREP01), 0),
    rep2 = replace_na(as.numeric(CNGREP02), 0),
    dem_votes = dem1 + dem2,
    rep_votes = rep1 + rep2
  ) |>
  group_by(CDDIST) |>
  summarise(
    total_dem = sum(dem_votes, na.rm = TRUE),
    total_rep = sum(rep_votes, na.rm = TRUE),
    margin = abs(total_dem - total_rep),
    .groups = "drop"
  ) |>
  arrange(margin)
```

```
Warning: There were 4 warnings in `mutate()`.
The first warning was:
```

```
ℹ In argument: `dem1 = replace_na(as.numeric(CNGDEM01), 0)`.
Caused by warning in `replace_na()`:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

```
cd_totals
```

```
# A tibble: 53 × 4
   CDDIST total_dem total_rep margin
    <dbl>     <dbl>     <dbl>  <dbl>
 1     45    158216    157522    694
 2     13     90617     85181   5436
 3     27    153942    145826   8116
 4      9    130093    121006   9087
 5     21    102701     92560  10141
 6     47    181662    171393  10269
 7     22     77882     89173  11291
 8     41    171093    182893  11800
 9     49    197333    180862  16471
10     39    130151     99415  30736
# ℹ 43 more rows
```

Some districts see the Democratic vote spread across multiple candidates more often than the Republican vote. This fragmentation can lightly depress Democratic district-level totals even in areas where Democrats dominate. These patterns become important when computing gerrymandering metrics that depend on vote share distribution.

Question 3

Which congressional districts have the strongest partisan "geographic clustering," meaning precincts almost all vote the same way?

Answer 3

```
prec_party <- clean |>
  mutate(
    dem1 = replace_na(as.numeric(CNGDEM01), 0),
    dem2 = replace_na(as.numeric(CNGDEM02), 0),
    rep1 = replace_na(as.numeric(CNGREP01), 0),
    rep2 = replace_na(as.numeric(CNGREP02), 0),
    dem_votes = dem1 + dem2,
    rep_votes = rep1 + rep2
  ) |>
  group_by(SVPREC) |>
  summarise(
    total_dem = sum(dem_votes, na.rm = TRUE),
```

```
    total_rep = sum(rep_votes, na.rm = TRUE),
    majority_party = case_when(
      total_dem > total_rep ~ "Democratic",
      total_rep > total_dem ~ "Republican",
      TRUE ~ "Tie"
    ),
    .groups = "drop"
  )
```

```
Warning: There were 4 warnings in `mutate()`.
The first warning was:
ℹ In argument: `dem1 = replace_na(as.numeric(CNGDEM01), 0)`.
Caused by warning in `replace_na()`:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 3 remaining warnings.
```

```
prec_party |>
  count(majority_party)
```

```
# A tibble: 3 × 2
  majority_party     n
  <chr>          <int>
1 Democratic     19140
2 Republican     13228
3 Tie            15969
```

Districts with a very high dominance rate (e.g., 85–95% of precincts voting for one party) indicate strong geographic clustering. These are districts where the opposing party is almost completely absent at the precinct level. High clustering makes a map less sensitive to small changes in district boundaries and often increases the efficiency gap.