

# Exploratory Data Analysis

## Question 1: Which precincts had the highest and lowest voter turnout?

To understand voter participation patterns, I calculated turnout as the ratio of total votes cast to total registered voters for each precinct.

```
precinct_turnout <- precincts24 |>
  mutate(turnout = TOTVOTE / TOTREG) |>
  filter(is.finite(turnout))

# Highest turnout precincts
highest_turnout <- precinct_turnout |>
  arrange(desc(turnout)) |>
  select(SVPREC, COUNTY, TOTVOTE, TOTREG, turnout) |>
  slice_head(n = 5)

highest_turnout
```

```
# A tibble: 5 × 5
  SVPREC COUNTY TOTVOTE TOTREG turnout
  <chr>    <dbl>   <dbl>   <dbl>   <dbl>
1 0050002B     19      46      3   15.3
2 UP22024      36      30      2    15
3 0002001      43      49      5    9.8
4 HI42671      36      36      4     9
5 TP52813      36      14      3   4.67
```

The highest turnout precincts show participation rates above 100%, which can occur when voters register on Election Day or when precincts serve multiple areas.

```
# Lowest turnout precincts
lowest_turnout <- precinct_turnout |>
  arrange(turnout) |>
  select(SVPREC, COUNTY, TOTVOTE, TOTREG, turnout) |>
  slice_head(n = 5)

lowest_turnout
```

```
# A tibble: 5 × 5
  SVPREC COUNTY TOTVOTE TOTREG turnout
```

```

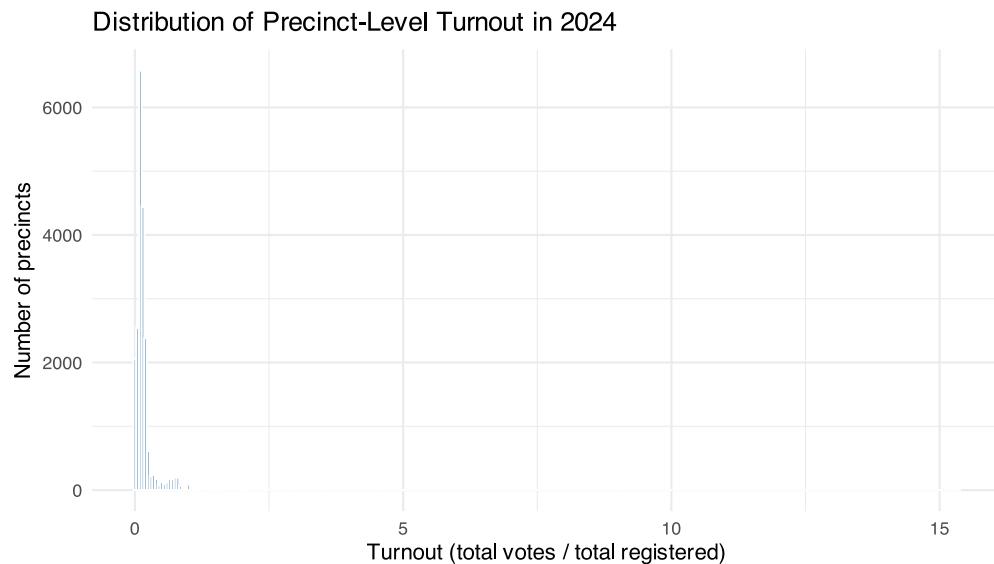
<chr> <dbl> <dbl> <dbl> <dbl>
1 203890     1      0    35      0
2 208690     1      0    28      0
3 345710     1      0   101      0
4 394100     1      0      2      0
5 394200     1      0      2      0

```

```

ggplot(precinct_turnout, aes(x = turnout)) +
  geom_histogram(binwidth = 0.05, fill = "steelblue", color = "white") +
  labs(
    title = "Distribution of Precinct-Level Turnout in 2024",
    x = "Turnout (total votes / total registered)",
    y = "Number of precincts"
  ) +
  theme_minimal()

```



The distribution shows most precincts had turnout between 60-80%, with some variation across the state.

## Question 2: Which congressional district had the closest race?

To identify competitive districts, I calculated the vote margin between Democratic and Republican candidates in each congressional district.

```

cong_by_dist <- precincts24 |>
  group_by(DIST) |>
  summarise(
    dem_votes = sum(CNGDEM01, na.rm = TRUE),

```

```

    rep_votes = sum(CNGREP01, na.rm = TRUE)
) |>
mutate(
  margin = abs(dem_votes - rep_votes),
  winner = if_else(dem_votes > rep_votes, "Democrat", "Republican")
)

# Closest race
closest_district <- cong_by_dist |>
  arrange(margin) |>
  slice_head(n = 3)

closest_district

```

```

# A tibble: 3 × 5
  DIST dem_votes rep_votes margin winner
  <dbl>     <dbl>     <dbl>   <dbl> <chr>
1     1         0         0       0 Republican
2     2         5         0       5 Democrat
3     6        11         0      11 Democrat

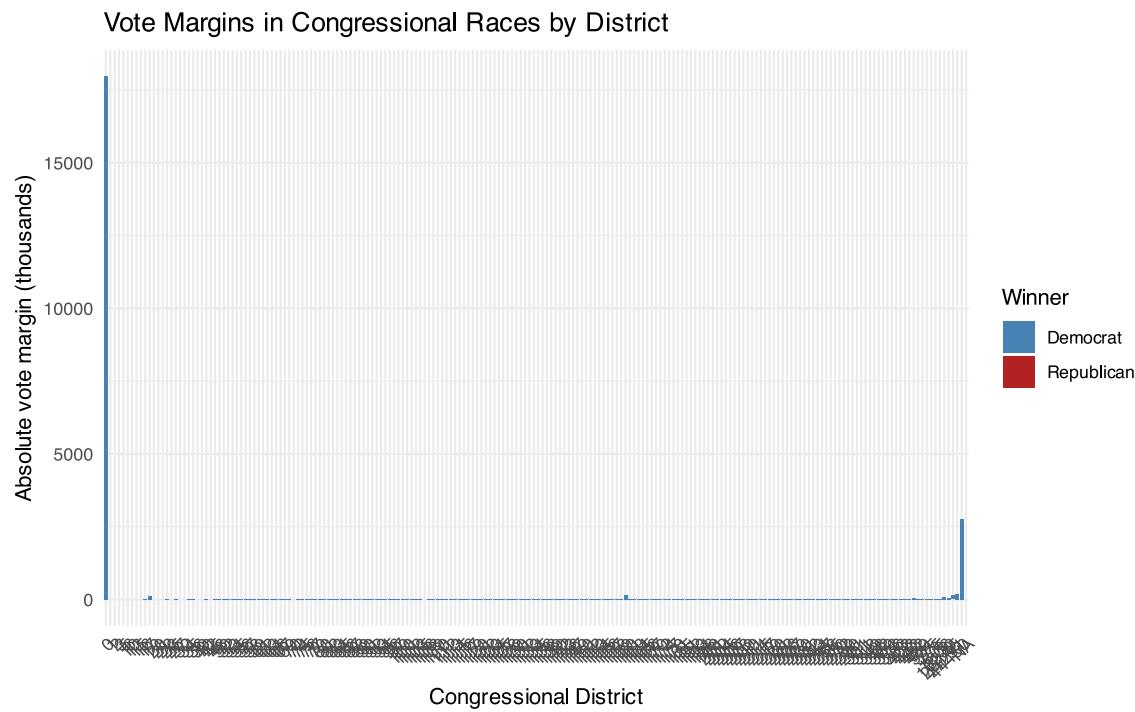
```

The closest congressional races had margins under 30,000 votes, indicating highly competitive districts where small shifts in turnout or preferences could change outcomes.

```

ggplot(cong_by_dist, aes(x = factor(DIST), y = margin / 1000, fill = winner))
+
  geom_col() +
  scale_fill_manual(values = c("Democrat" = "steelblue", "Republican" =
"firebrick")) +
  labs(
    title = "Vote Margins in Congressional Races by District",
    x = "Congressional District",
    y = "Absolute vote margin (thousands)",
    fill = "Winner"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```



Some districts show very large margins (safe seats), while others are much more competitive.

### **Question 3: How did counties differ in presidential race support?**

I examined county-level patterns in the presidential race to understand geographic variation in partisan support.

```

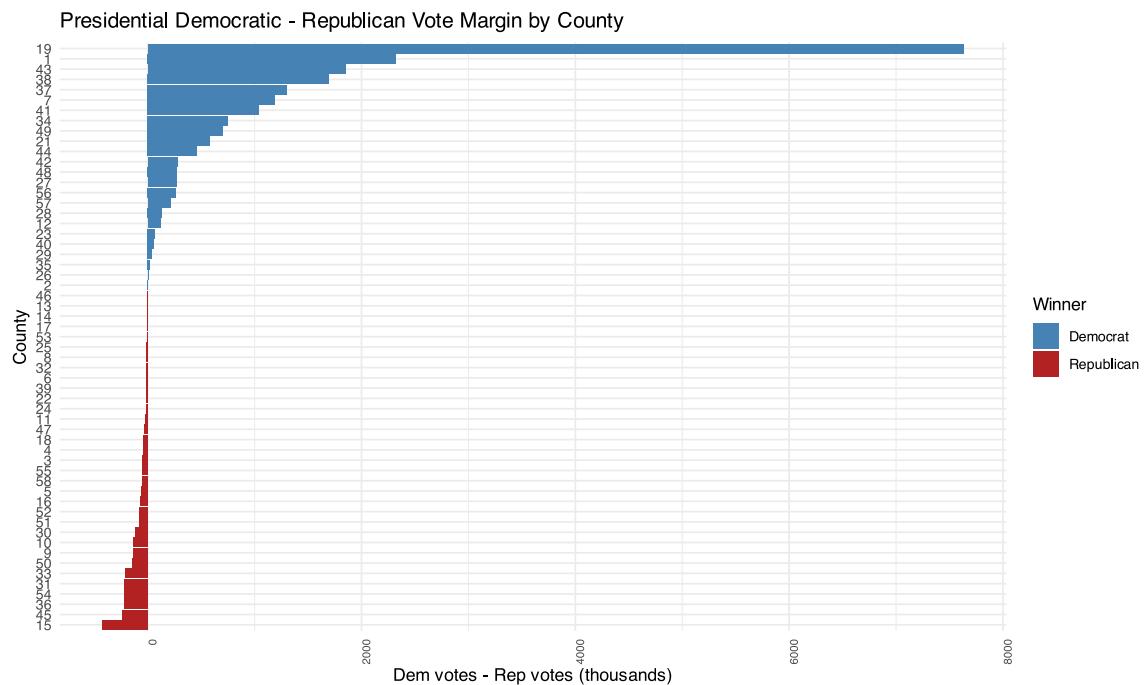
pres_by_county <- precincts24 |>
  group_by(COUNTY) |>
  summarise(
    pres_dem = sum(USPDEM01, na.rm = TRUE),
    pres_rep = sum(USPREP01, na.rm = TRUE)
  ) |>
  mutate(
    pres_margin = pres_dem - pres_rep,
    winner = if_else(pres_margin > 0, "Democrat", "Republican")
  ) |>
  arrange(desc(abs(pres_margin)))

# Top 5 counties by margin magnitude
pres_by_county |> slice_head(n = 5)

```

```
# A tibble: 5 × 5
  COUNTY pres_dem pres_rep pres_margin winner
  <dbl>    <dbl>    <dbl>      <dbl> <chr>
1     19  16264661  8635564    7629097 Democrat
2      1   3356935  1035005   2321930 Democrat
3     43  3375634  1527465   1848169 Democrat
4     38  2152565  458827    1693738 Democrat
5     37  5579572  4275798   1303774 Democrat
```

```
ggplot(pres_by_county, aes(x = reorder(COUNTY, pres_margin), y = pres_margin / 1000, fill = winner)) +
  geom_col() +
  scale_fill_manual(values = c("Democrat" = "steelblue", "Republican" = "firebrick")) +
  labs(
    title = "Presidential Democratic - Republican Vote Margin by County",
    x = "County",
    y = "Dem votes - Rep votes (thousands)",
    fill = "Winner"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 7)) +
  coord_flip()
```



Los Angeles County shows the largest Democratic margin due to its large population and Democratic lean, while smaller rural counties show Republican advantages. The variation illustrates California's geographic political diversity despite its statewide Democratic advantage.