Project 2, Dataset 3

Alice Ding

2023-02-27

Overview

For this dataset, I'll be using the first one posted by Farhana and it holds vote counts for two states. This data has one row per political candidate and includes the following columns:

- Candidate
- CA
- FL

The last 2 columns are vote counts for those states.

```
vote_data <- read.csv("https://raw.githubusercontent.com/addsding/data607/main/project2/vote-counts.csv
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
## incomplete final line found by readTableHeader on
## 'https://raw.githubusercontent.com/addsding/data607/main/project2/vote-counts.csv'
```

```
head(vote_data)
```

```
##         Candidate      CA      FL
## 1 Hillary Clinton 5931283 4485745
## 2    Donald Trump 3184721 4605515
## 3    Gary Johnson  308392  206007
## 4      Jill Stein  166311   64019
```

To get this data tidy, we'll be pivoting the last two columns to have one column per candidate and state combination.

Tidying the Data

Pivot time!

```
vote_data_pivot <- pivot_longer(vote_data, cols=2:3, names_to="state", values_to="votes")
head(vote_data_pivot)
```

```
## # A tibble: 6 x 3
##   Candidate       state votes
##   <chr>           <chr> <int>
## 1 Hillary Clinton CA    5931283
## 2 Hillary Clinton FL    4485745
## 3 Donald Trump    CA    3184721
## 4 Donald Trump    FL    4605515
## 5 Gary Johnson    CA     308392
## 6 Gary Johnson    FL     206007
```

Data looks good, no other cleaning necessary!

Analysis

Overall, who has the most votes? And what percentage of votes did they get?

```
votes <- vote_data_pivot |>
  group_by(Candidate) |>
  summarise(sum_votes = sum(votes),
            .groups = 'drop') |>
  mutate(freq = formattable::percent(sum_votes / sum(sum_votes)))

votes
```

```
## # A tibble: 4 x 3
##   Candidate       sum_votes freq
##   <chr>               <int> <formttbl>
## 1 Donald Trump      7790236 41.11%
## 2 Gary Johnson       514399 2.71%
## 3 Hillary Clinton  10417028 54.97%
## 4 Jill Stein         230330 1.22%
```

From the sum, it looks like Hilary Clinton had the most votes at 54.97% for CA and FL with Donald Trump in 2nd place at 41.11%. Does this change if we group by state as well?

```
cali_votes <- vote_data_pivot[vote_data_pivot$state=="CA",] |>
  group_by(Candidate) |>
  summarise(sum_votes = sum(votes),
            .groups = 'drop') |>
  mutate(freq = formattable::percent(sum_votes / sum(sum_votes)))

cali_votes
```

```
## # A tibble: 4 x 3
##   Candidate       sum_votes freq
##   <chr>               <int> <formttbl>
## 1 Donald Trump      3184721 33.21%
## 2 Gary Johnson       308392 3.22%
## 3 Hillary Clinton   5931283 61.84%
## 4 Jill Stein         166311 1.73%
```

```
fl_votes <- vote_data_pivot[vote_data_pivot$state=="FL",] |>
  group_by(Candidate) |>
  summarise(sum_votes = sum(votes),
            .groups = 'drop') |>
  mutate(freq = formattable::percent(sum_votes / sum(sum_votes)))

fl_votes
```

```
## # A tibble: 4 x 3
##   Candidate       sum_votes freq
##   <chr>               <int> <formttbl>
## 1 Donald Trump      4605515 49.20%
## 2 Gary Johnson       206007 2.20%
## 3 Hillary Clinton   4485745 47.92%
## 4 Jill Stein          64019 0.68%
```

By state, Clinton is very much the majority in CA. in FL however, it's much closer with Trump actually taking the lead. This makes sense as FL tends to lean more Republican while CA is very much a Democratic state.

Conclusion

With the help of pivoting, this data was easy to tidy up and then analyze with `dplyr`. It would be more interesting to see other states' voting data as well as perhaps what percentage of the state is in which political party to see if the votes lined up with affiliation. It'd also be interesting to see the percentage of folks who didn't vote if possible and see how that swung the outcome of this election in another analysis.