Homework 1: Data Cleaning, Merging and Aggregation in R

Emma Jay

January 24, 2025

```
#libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(writexl)

#default theme for ggplot2
ggplot2::theme_set(ggplot2::theme_minimal(base_size = 16))

#default parameters for knitr
knitr::opts_chunk$set(
  fig.width = 8,
  fig.asp = 0.618,
  fig.retina = 2,
  dpi = 150,
  out.width = "70%"
)
```

Part I: Survey Data Manipulation

1. Exploratory Data Analysis

```
#load data
gactt_data <- read_csv("data/GACTT_RESULTS_ANONYMIZED_HW1.csv")
```

```
## Rows: 3280 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (7): submission_id, zip, age, gender, cups, home_brew, party
## dbl (1): cups_num
##
## 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.
```

```
#print first 6 rows and last 4 rows
gactt_data %>%
  head(6)
```

```
## # A tibble: 6 x 8
##   submission_id zip   age             gender cups  cups_num home_brew      party
##   <chr>         <chr> <chr>           <chr>  <chr>    <dbl> <chr>          <chr>
## 1 gMR29l        <NA>  18-24 years old <NA>   <NA>        NA <NA>           <NA>
## 2 BkPN0e        <NA>  25-34 years old <NA>   <NA>        NA Pod/capsule m~ <NA>
## 3 W5G8jj        <NA>  25-34 years old <NA>   <NA>        NA Bean-to-cup m~ <NA>
## 4 4xWgGr        <NA>  35-44 years old <NA>   <NA>        NA Coffee brewin~ <NA>
## 5 QD27Q8        <NA>  25-34 years old <NA>   <NA>        NA Pour over      <NA>
## 6 V0LPeM        <NA>  55-64 years old <NA>   <NA>        NA Pod/capsule m~ <NA>
```

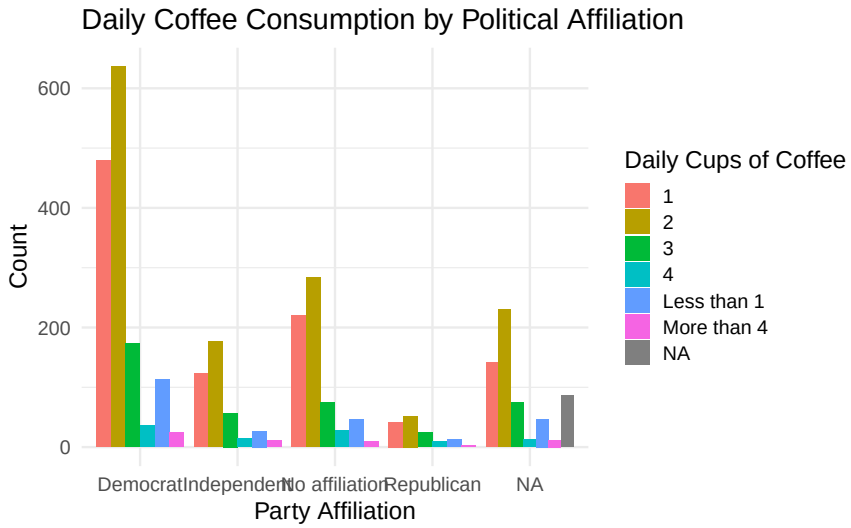
```
gactt_data %>%
  tail(4)
```

```
## # A tibble: 4 x 8
##   submission_id zip   age             gender cups       cups_num home_brew party
##   <chr>         <chr> <chr>           <chr>  <chr>         <dbl> <chr>     <chr>
## 1 42EpEY        91505 25-34 years old <NA>   More than~        5 Espresso~ <NA>
## 2 g5ggRM        60131 18-24 years old Male   1                 1 Espresso~ Demo~
## 3 rlgbDN        2351  25-34 years old Male   2                 2 Pour over Demo~
## 4 0EGYe9        32765 25-34 years old Female 1                 1 Pour ove~ Demo~
```

```
#print class of each variable
gactt_data %>%
  sapply(class)
```

```
## submission_id           zip           age        gender          cups
##   "character"   "character"   "character"   "character"   "character"
##      cups_num     home_brew         party
##     "numeric"   "character"   "character"
```

```
#plot - relationship between cups and party
gactt_data %>%
  ggplot(aes(x = party, fill = cups)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Daily Coffee Consumption by Political Affiliation",
    x = "Party Affiliation",
    y = "Count",
    fill = "Daily Cups of Coffee"
  )
```



Graph Interpretation: The graph shows that Democrats were the most identified with group in the data set, followed by Republicans. In every political affiliation group, two cups of coffee was the most reported daily coffee intake, followed by one and then three (in every category but NA where NA was the third most responded).

2. Merging Survey Data with ZIP Code Metadata

```
#load zip code metadata
zip_codes <- read_csv("data/zip_code_database.csv")
```

```
## Rows: 42735 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): zip, type, primary_city, acceptable_cities, unacceptable_cities, s...
## dbl  (4): decommissioned, latitude, longitude, irs_estimated_population
##
## 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.
```

```
#merge zip column with metadata to include state-level info in survey dataset
gactt_zip <- full_join(gactt_data, zip_codes, by = "zip")

#check match using number of NA values - should be same as unmatched zip
table(gactt_zip$zip %in% zip_codes$zip) %>%
  as.data.frame() %>%
  print()
```

```
##    Var1  Freq
## 1 FALSE   442
## 2  TRUE 43838
```

```
table(is.na(gactt_zip$state)) %>%
  as.data.frame() %>%
  print()
```

```
##    Var1  Freq
## 1 FALSE 43838
## 2  TRUE   442
```

Number of Unmatched Zip Codes: There are 442 unmatched zip codes.

3. Aggregation and Insights

```
#calculate average daily coffee consumption (cups) by state
avg_cups <- gactt_zip %>%
  mutate(cups = as.numeric(cups)) %>%
  group_by(state) %>%
  summarize(average_cups = mean(cups, na.rm = TRUE))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `cups = as.numeric(cups)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
head(avg_cups)
```

```
## # A tibble: 6 x 2
##   state average_cups
##   <chr>        <dbl>
## 1 AA             NaN
## 2 AE             NaN
## 3 AK               3
## 4 AL               1.75
## 5 AP             NaN
## 6 AR               1.75
```

```
#most preferred homebrew coffee method (home_brew) in each state
top_homebrew <- gactt_zip %>%
  filter(!is.na(home_brew)) %>%
  group_by(state, home_brew) %>%
  tally() %>%
  slice_max(order_by = n, n = 1) %>%
  ungroup() %>%
  select(state, home_brew)

head(top_homebrew)
```

```
## # A tibble: 6 x 2
##   state home_brew
##   <chr> <chr>
## 1 AK    Pour over, Other
## 2 AL    Espresso
## 3 AR    Pour over
## 4 AZ    Pour over
## 5 CA    Pour over
## 6 CO    Pour over
```

```r
#political affiliation breakdown (party) in each state (percentage of respondents identifying with demo
politic_state <- gactt_zip %>%
  group_by(state, party) %>%
  count() %>%
  group_by(state) %>%
  mutate(percentage = (n / sum(n))) %>%
  ungroup () %>%
  select(state, party, percentage) %>%
  pivot_wider(names_from = party, values_from = percentage, values_fill = list(percentage = 0))

head(politic_state)
```

```
## # A tibble: 6 x 6
##   state  `NA` Democrat Independent `No affiliation` Republican
##   <chr> <dbl>    <dbl>       <dbl>             <dbl>      <dbl>
## 1 AA    1      0           0                 0          0
## 2 AE    1      0           0                 0          0
## 3 AK    1      0           0                 0          0
## 4 AL    0.978  0.0118      0.00591           0.00355    0.00118
## 5 AP    1      0           0                 0          0
## 6 AR    0.989  0.00421     0.00140           0.00421    0.00140
```

```r
#save aggregated data frame as a variable survey_state
survey_state <- avg_cups %>%
  left_join(top_homebrew, by = "state") %>%
  left_join(politic_state, by = "state")

head(survey_state)
```

```
## # A tibble: 6 x 8
##   state average_cups home_brew        `NA` Democrat Independent `No affiliation`
##   <chr>        <dbl> <chr>           <dbl>    <dbl>       <dbl>             <dbl>
## 1 AA             NaN <NA>            1      0           0                 0
## 2 AE             NaN <NA>            1      0           0                 0
## 3 AK               3 Pour over, Oth~ 1      0           0                 0
## 4 AL               1.75 Espresso     0.978  0.0118      0.00591           0.00355
## 5 AP             NaN <NA>            1      0           0                 0
## 6 AR               1.75 Pour over    0.989  0.00421     0.00140           0.00421
## # i 1 more variable: Republican <dbl>
```

Part II: Election Data Analysis

1. Cleaning the Election Data

```r
#1. load 2024 election data
election_2024 <- read_csv("data/election_2024.csv")
```

```
## Rows: 56 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (4): state, harris_votes_share, trump_votes_share, other_votes_share
## dbl (2): harris_ev, trump_ev
## num (4): total_votes, harris_votes, trump_votes, other_votes
## lgl (1): other_ev
##
## 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.
```

```
election_2024 %>%
  sapply(class)
```

```
##              state        total_votes       harris_votes harris_votes_share
##        "character"          "numeric"          "numeric"        "character"
##          harris_ev        trump_votes  trump_votes_share           trump_ev
##          "numeric"          "numeric"        "character"          "numeric"
##        other_votes  other_votes_share           other_ev
##          "numeric"        "character"          "logical"
```

```
#2. clean data
clean_election2024 <- election_2024 %>%
  mutate(
    harris_votes_share = as.numeric(str_remove_all(harris_votes_share, "%"))/ 100,
    trump_votes_share = as.numeric(str_remove_all(trump_votes_share, "%")) / 100,
    other_votes_share = as.numeric(str_remove_all(other_votes_share, "%")) / 100,
other_ev = as.numeric(other_ev))

sapply(clean_election2024, class)
```

```
##              state        total_votes       harris_votes harris_votes_share
##        "character"          "numeric"          "numeric"          "numeric"
##          harris_ev        trump_votes  trump_votes_share           trump_ev
##          "numeric"          "numeric"          "numeric"          "numeric"
##        other_votes  other_votes_share           other_ev
##          "numeric"          "numeric"          "numeric"
```

2. Merging Survey and Election Data

```
#merge survey_state with election data using state column

#resolve state name issue across data frames
state_mapping <- data.frame(
  full_state = c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut", "
  state_abbreviation = c("AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE", "FL", "GA", "HI", "ID", "IL",

merge_state <- clean_election2024 %>%
  left_join(state_mapping, by = c("state" = "full_state")) %>%
  left_join(survey_state, by = c("state_abbreviation" = "state"))
```

```
## Warning in left_join(., survey_state, by = c(state_abbreviation = "state")): Detected an unexpected
## i Row 10 of `x` matches multiple rows in `y`.
```

```
## i Row 112 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

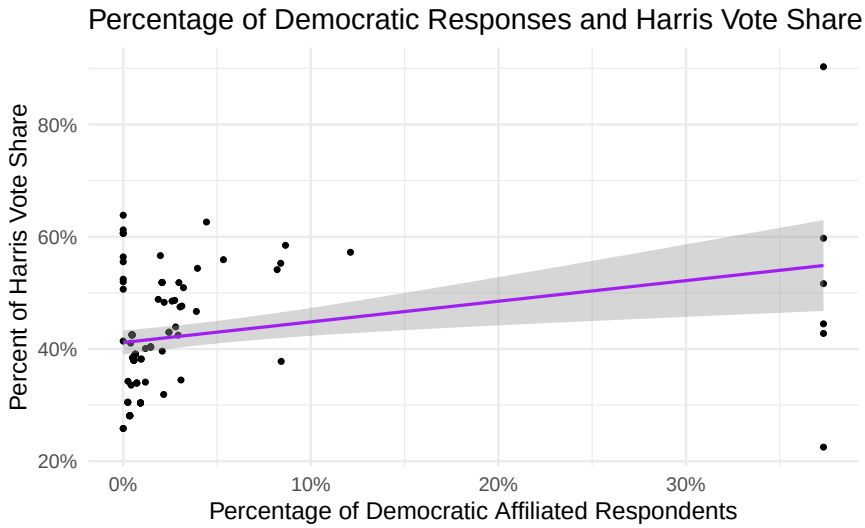
```
head(merge_state)
```

```
## # A tibble: 6 x 19
##   state      total_votes harris_votes harris_votes_share harris_ev trump_votes
##   <chr>            <dbl>        <dbl>              <dbl>     <dbl>       <dbl>
## 1 Alabama        2265090       772412              0.341        NA     1462616
## 2 Alaska          338177       140026              0.414        NA      184458
## 3 Arizona        3390161      1582860              0.467        NA     1770242
## 4 Arkansas       1182676       396905              0.336        NA      759241
## 5 California    15865475      9276179              0.585        54     6081697
## 6 Colorado       3192745      1728159              0.541        10     1377441
## # i 13 more variables: trump_votes_share <dbl>, trump_ev <dbl>,
## #   other_votes <dbl>, other_votes_share <dbl>, other_ev <dbl>,
## #   state_abbreviation <chr>, average_cups <dbl>, home_brew <chr>, `NA` <dbl>,
## #   Democrat <dbl>, Independent <dbl>, `No affiliation` <dbl>, Republican <dbl>
```

3. Comparative Analysis and Visualization

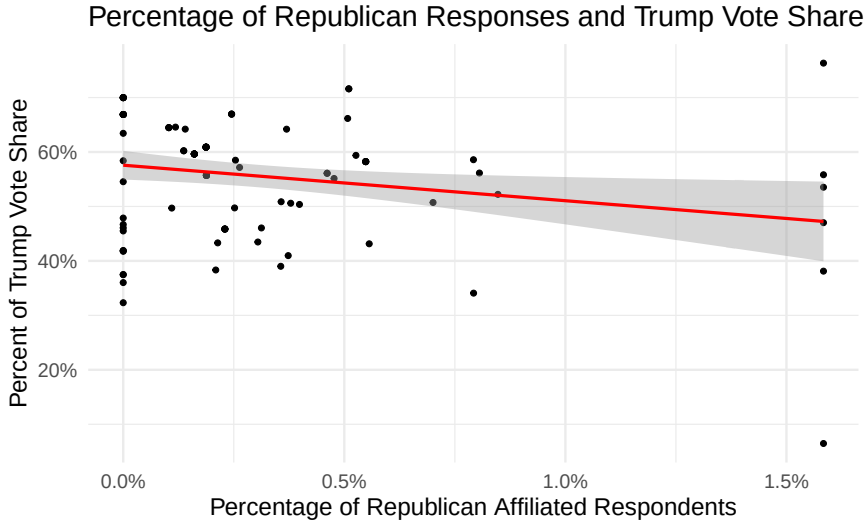
```
#1.
# relationship between Democratic respondents and Harris vote share
merge_state %>%
  ggplot(aes(x = Democrat, y = harris_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "purple") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Percentage of Democratic Responses and Harris Vote Share by State",
    x = "Percentage of Democratic Affiliated Respondents",
    y = "Percent of Harris Vote Share"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Republicans and Trump
merge_state %>%
  ggplot(aes(x = Republican, y = trump_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Percentage of Republican Responses and Trump Vote Share by State",
    x = "Percentage of Republican Affiliated Respondents",
    y = "Percent of Trump Vote Share"
  )
```

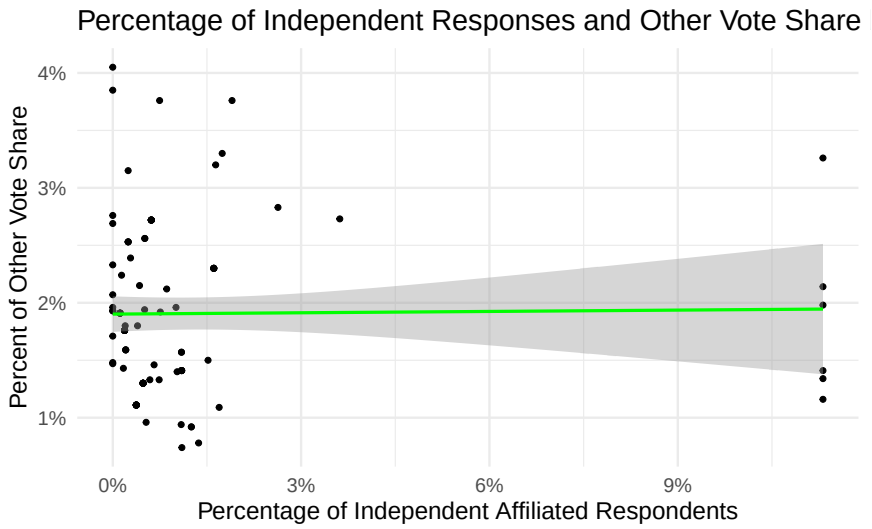
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Other and Independent
merge_state %>%
```

```
  ggplot(aes(x = Independent, y = other_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "green") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Percentage of Independent Responses and Other Vote Share by State",
    x = "Percentage of Independent Affiliated Respondents",
    y = "Percent of Other Vote Share"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



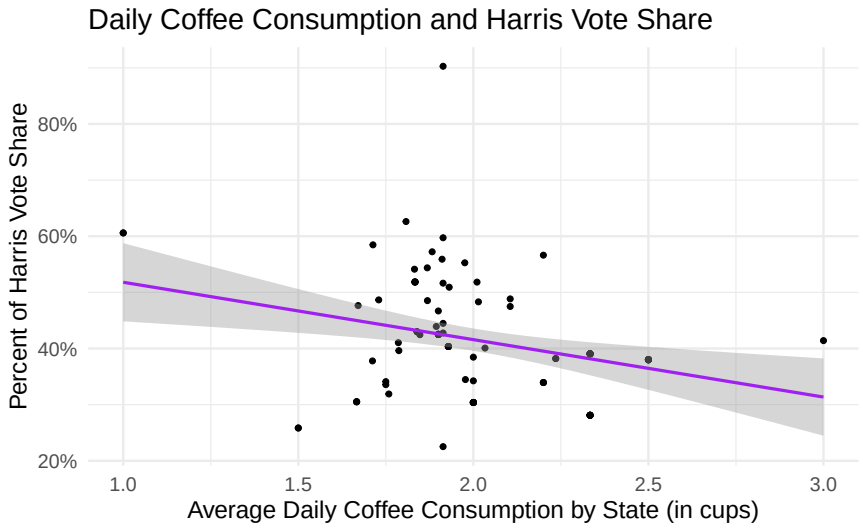
Limitations of Data: The data represented above contains states with limited numbers of respondents overall and high rates of respondents that did not complete the political affiliation question, resulting in an NA value. These limitations could lead to the data being significantly skewed or not representative of the actual state populations.

```
#relationship between daily coffee consumption (cups) and voting outcomes
#harris
merge_state %>%
  ggplot(aes(x = average_cups, y = harris_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "purple") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Daily Coffee Consumption and Harris Vote Share",
    x = "Average Daily Coffee Consumption by State (in cups)",
    y = "Percent of Harris Vote Share"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 7 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

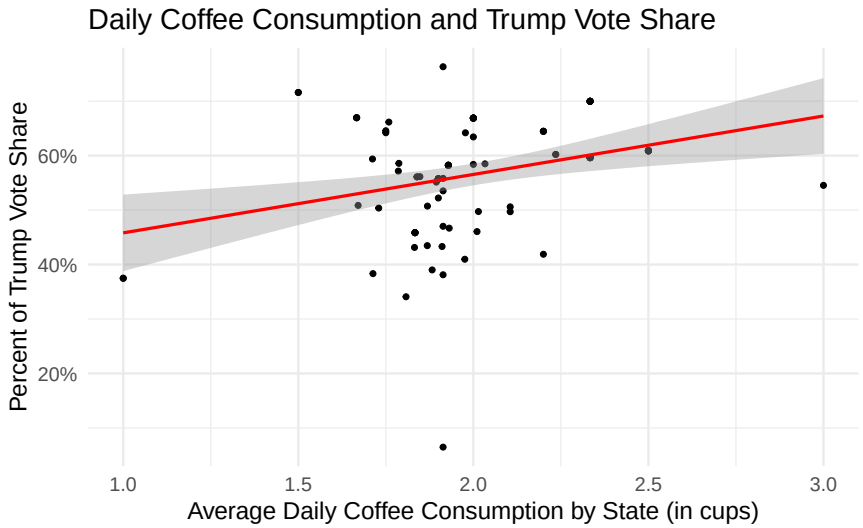
```
## Warning: Removed 7 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
#trump
merge_state %>%
  ggplot(aes(x = average_cups, y = trump_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Daily Coffee Consumption and Trump Vote Share",
    x = "Average Daily Coffee Consumption by State (in cups)",
    y = "Percent of Trump Vote Share"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

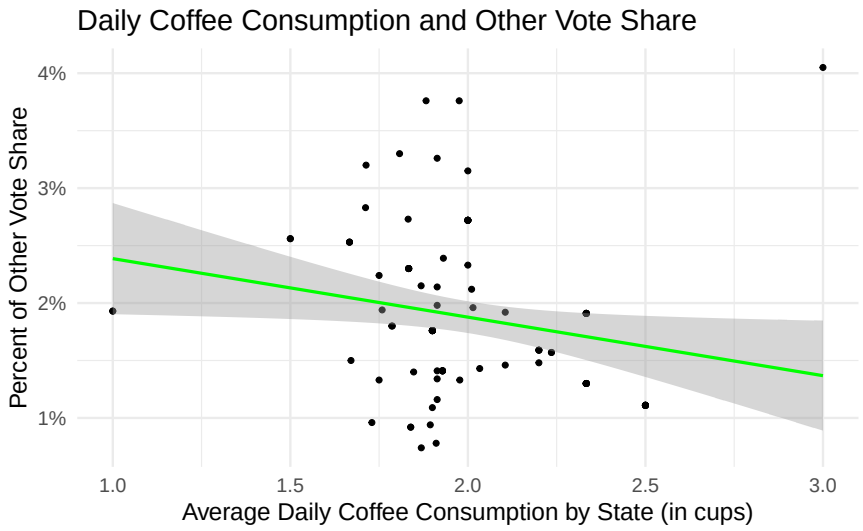
```
## Warning: Removed 7 rows containing non-finite outside the scale range (`stat_smooth()`).
## Removed 7 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
#other
merge_state %>%
  ggplot(aes(x = average_cups, y = other_votes_share)) +
  geom_point() +
  geom_smooth(method = "lm", color = "green") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Daily Coffee Consumption and Other Vote Share",
    x = "Average Daily Coffee Consumption by State (in cups)",
    y = "Percent of Other Vote Share"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 7 rows containing non-finite outside the scale range (`stat_smooth()`).
## Removed 7 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Interpretation: The scatter plots indicate that as the average daily coffee consumption of a state increases, the Harris and Other share

of the vote decreases while the Trump vote share increases. Each graph shows a clustering of responses around two cups of coffee on average per day, where the regression line has the smallest standard error compared to either. 1 or 3 cups where the standard error is significantly larger.

Save Results

```
#save final merged dataset with aggregated survey results and election outcomes

 write_xlsx(merge_state, "overview_hw1.xlsx")
```