Homework One

Yael Beshaw

2025-01-02

Load Necessary Packages

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## -- 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   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.1
## -- 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(tidycensus)
```

Use the FiveThirtyEight presidential elections data to answer the following questions about the 2020 general election results.

```
url<-"https://raw.githubusercontent.com/fivethirtyeight/election-results/main/election_results_president

presidential_elections<-read_csv(url)
```

```
## Rows: 7423 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): state_abbrev, state, office_name, stage, party, candidate_name, bal...
## dbl (8): id, race_id, office_id, cycle, politician_id, candidate_id, votes, ...
## lgl (4): office_seat_name, special, unopposed, winner
##
## 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 One:

Create a data frame with the two party vote share and the winning candidate for each state (plus D.C.) in the 2020 presidential election.

```
election_2020 <- presidential_elections|>
  select(cycle, stage, state, candidate_name, votes, winner)|>
  #select columns of interest
  filter(cycle==2020)|> #2020
```

```
  filter(stage== "general")|> #presidential election
  filter(!is.na(state)) #each state + D.C.
```

```
pres_2020 <- election_2020|>
  select(state, candidate_name, votes)|> #filter the df columns
  group_by(state) |> #organize by state
  filter(candidate_name %in% c("Joe Biden", "Donald Trump"))|> #two party
  filter(str_detect(state, "CD-[0-9]") ==FALSE) #remove CD's for Maine/Nebraska
```

```
#troubleshoot
#issue where "New York" was duplicated, assuming it did not filter out
#state vs city

pres_2020 <- pres_2020|>
  filter(!(state == "New York" & candidate_name == "Joe Biden" &
             votes == 386010))|>
  filter(!(state == "New York" & candidate_name == "Donald Trump" &
             votes ==   295657))

#utilize pivot wider to show vote shares between Trump and Biden
pres_wide <- pres_2020|>
  pivot_wider(names_from = candidate_name,
              values_from = votes)|>
  arrange(state) #alphabetical order

pres_wide
```

```
## # A tibble: 51 x 3
## # Groups:   state [51]
##    state                `Donald Trump` `Joe Biden`
##    <chr>                         <dbl>       <dbl>
##  1 Alabama                     1441170      849624
##  2 Alaska                       189951      153778
##  3 Arizona                     1661686     1672143
##  4 Arkansas                     760647      423932
##  5 California                  6006429    11110250
##  6 Colorado                    1364607     1804352
##  7 Connecticut                  714717     1080831
##  8 Delaware                     200603      296268
##  9 District of Columbia          18586      317323
## 10 Florida                     5668731     5297045
## # i 41 more rows
```

Question Two: Use the data frame you created in the prior step to calculate Biden's share of the two-party vote in each state (i.e. Biden votes / (Biden votes + Trump votes)

```
new_pres2020 <- pres_wide|>
  mutate(`Biden's Share`= `Joe Biden`/(`Joe Biden` + `Donald Trump`))

new_pres2020
```

```
## # A tibble: 51 x 4
## # Groups:   state [51]
##    state                `Donald Trump` `Joe Biden` `Biden's Share`
##    <chr>                         <dbl>       <dbl>           <dbl>
##  1 Alabama                     1441170      849624           0.371
```

```
##  2 Alaska                       189951      153778          0.447
##  3 Arizona                     1661686     1672143          0.502
##  4 Arkansas                     760647      423932          0.358
##  5 California                  6006429    11110250          0.649
##  6 Colorado                    1364607     1804352          0.569
##  7 Connecticut                  714717     1080831          0.602
##  8 Delaware                     200603      296268          0.596
##  9 District of Columbia          18586      317323          0.945
## 10 Florida                     5668731     5297045          0.483
## # i 41 more rows
```

Question Three:

Use the following code to download the ACS estimated median household income for each state and then use a join to add this column to your data.

```
median_income <- get_acs(geography = "state",
                         variables = c(median_income = "B19013_001"),
                         year = 2020)
```

```
## Getting data from the 2016-2020 5-year ACS
```

```
votes_and_median <-new_pres2020|>
  left_join(median_income, by=join_by(state == NAME))|>
  mutate(`Median Household Income`= estimate)|>
  select(state, `Donald Trump`, `Joe Biden`, `Biden's Share`,
         `Median Household Income`)
```

Question Four:

Run a linear regression to calculate the effect of median income on Biden's statewide two party vote share. Produce a formatted table to display your results and briefly discuss your findings.

```
library(flextable)
```

```
##
## Attaching package: 'flextable'
```

```
## The following object is masked from 'package:purrr':
##
##     compose
```

```
model<-lm(`Biden's Share` ~ `Median Household Income`  , data= votes_and_median)
summary(model)
```

```
##
## Call:
## lm(formula = `Biden's Share` ~ `Median Household Income`, data = votes_and_median)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.223946 -0.034994 -0.002569  0.043908  0.241640
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.223e-02  7.333e-02  -0.303    0.763
## `Median Household Income`  7.984e-06  1.112e-06   7.181 3.48e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.08688 on 49 degrees of freedom
## Multiple R-squared:  0.5127, Adjusted R-squared:  0.5028
## F-statistic: 51.56 on 1 and 49 DF,  p-value: 3.482e-09
```

As seen in the table below, for every one unit increase in the median household income of a state (including D.C.) in the United States. The expected mean for Biden's Share of the vote goes up by 7.98e-06 units. This is statistically significant as we observe the p-value to less than 0.05 and we therefore reject the null hypothesis that there is is no effect of median household income. While the effect of the median household income seems small, it is rather much larger when we consider that the unit of measure is in dollars. Thus, a state that has a median household value around $25,000 greater than another (i.e., Alabama vs Alaska), we can estimate that Biden's Vote Share increases by 0.1996 or almost 20%.

```
as_flextable(model)
```

```
## Warning: fonts used in `flextable` are ignored because the `pdflatex` engine is
## used and not `xelatex` or `lualatex`. You can avoid this warning by using the
## `set_flextable_defaults(fonts_ignore=TRUE)` command or use a compatible engine
## by defining `latex_engine: xelatex` in the YAML header of the R Markdown
## document.
```

	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	-0.022	0.073	-0.303	0.7631
'Median Household Income'	0.000	0.000	7.181	0.0000***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 0.08688 on 49 degrees of freedom

Multiple R-squared: 0.5127, Adjusted R-squared: 0.5028

F-statistic: 51.56 on 49 and 1 DF, p-value: 0.0000