Midterm

Zai Rutter

11/26/2021

```
setwd("~/Documents/Upenn/Data 210/Week 5/Midterm")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fastLink)
library(dplyr)
library(haven)
genforward_sept_2017 <- read_sav("genforward sept 2017.sav")

Gs17<- genforward_sept_2017
```

Question 1

A This question makes an assumption that the survey respondent agrees that there is gridlock in Washington. This questions is also vague, what do they mean by dysfunction and gridlock? These are terms familiar to political scienctists but to people without a college degree, it can be confusing.

B Q: "Do you agree that there is bipartisan cooperation in Congress and the Senate?" R: "Strongly disagree, somewhat disagree, somewhat agree, Strongly agree, No answer"

Q: "Do you believe that"Donald Trump" is negatively contributing to bipartisan cooperation in Congress and the Senate?"

R:"Strongly disagree, somewhat disagree, Neutral, somewhat agree, Strongly agree"

Any one of the names could be used instead of Donald Trump.

Question 2

A What is the relationship between Trump voters and thier support for Green New Deal and for Medicare for all. Is there a Trump 'sect' (race, age, education ect) that generally runs counter to dominent trends.

What is the relationship between non voters and thier support for Green New Deal and for Medicare for all.

Is there a relationship between those who did not vote and thier Trump approval

Generally, who did not vote?

B Q1. ‘Do you support the decision for U.S. troops to leave Afghanistan?’ R:“Strongly support, somewhat support, Neutral, somewhat support, Strongly support, Do not know” Q2. ‘How often do you practice your religion?’ R. ‘Very often, somewhat often, Never, do not have follow a religion’

Question 3

A

```
### A
#
# What percent of the sample strongly approved or somewhat approved of the way
#  that President Trump is handling his job as president (using question Q1)?

# strong approved is 1 , somewhat is 2
attributes(Gs17$Q1)
```

```
## $label
## [1] "Overall, do you approve, disapprove, or neither approve nor disapprove of the wa"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
##               Strongly approve               Somewhat approve
##                              1                              2
## Neither approve nor disapprove            Somewhat disapprove
##                              3                              4
##            Strongly disapprove                     DON’T KNOW
##                              5                             77
##                 SKIPPED ON WEB                        refused
##                             98                             99
```

```
# number of people sampeled # 1741
sum(length(unique(Gs17$GenF_ID)))
```

```
## [1] 1741
```

```
Gs17$approve.trump <- ifelse( Gs17$Q1 <=2, T, F)
table(Gs17$approve.trump == T)
```

```
##
## FALSE  TRUE
##  1485   256
```

```
# Percent
256/1741
```

```
## [1] 0.1470419
```

B

```
###B
# What percentage of Republican men "strongly approve" or "somewhat approve"
# of the way Trump is handling his job as president?

# Male = 1, Female = 2
attributes(Gs17$gender)
```

```
## $label
## [1] "Respondent gender"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
## Unknown    Male  Female
##       0       1       2
```

```
# Republican = 6,5,7
attributes(Gs17$PartyID7)
```

```
## $label
## [1] "DATA-ONLY: Computed 7-level Party ID"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
##                     Unknown             Strong Democrat
##                          -1                           1
##           Moderate Democrat               Lean Democrat
##                           2                           3
## Don't Lean/Independent/None             Lean Republican
##                           4                           5
##         Moderate Republican           Strong Republican
##                           6                           7
```

```
Gs17$republican <- ifelse(Gs17$PartyID7 >=5,T,F)

(sum(Gs17$approve.trump[Gs17$gender == 1 & Gs17$republican == T]))/(sum(Gs17$republican == T & Gs17$gen
## [1] 0.5244444
```

```
# What is this percentage for Republican women?
(sum(Gs17$approve.trump[Gs17$gender == 2 & Gs17$republican == T]))/(sum(Gs17$republican == T & Gs17$gen
```

```
## [1] 0.4545455
```

```
# What percentage of Republican men and Republican women
# (separately) "somewhat disapprove" or "strongly disapprove" of Trump?

# Somewaht disapprove / strongly disapprove = 5 and 4
attributes(Gs17$Q1)
```

```
## $label
## [1] "Overall, do you approve, disapprove, or neither approve nor disapprove of the wa"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
##               Strongly approve               Somewhat approve
##                              1                              2
## Neither approve nor disapprove            Somewhat disapprove
##                              3                              4
##            Strongly disapprove                     DON'T KNOW
##                              5                             77
##                SKIPPED ON WEB                        refused
##                             98                             99
```

```
## rep men = 0
(sum(Gs17$approve.trump[Gs17$Q1 == 4 & Gs17$Q1 == 5 & Gs17$republican == T &
                          Gs17$gender == 1]))/(sum(length(unique(Gs17$GenF_ID))))
```

```
## [1] 0
```

```
## rep women = 0
(sum(Gs17$approve.trump[Gs17$Q1 == 4 & Gs17$Q1 == 5 & Gs17$republican == T &
                          Gs17$gender == 2]))/(sum(length(unique(Gs17$GenF_ID))))
```

```
## [1] 0
```

C Question 21 and Question 6 are the most important 21) ~ 18% 6) ~10.6%

```
# Which two issues did 2016 Trump voters indicate were the most
# important problems facing the country? What percentage of Trump voters
# listed each of these two issues as the top issue?

#Q13
#Trump Voters == 2
attributes(Gs17$Q0)
```

```
## $label
## [1] "Did you vote for Hillary Clinton, Donald Trump, someone else, or not vote in the"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
##                                 Hillary Clinton
##                                               1
##                                    Donald Trump
##                                               2
##                                    Someone else
##                                               3
## Did not vote in the 2016 presidential election
##                                               4
##                                      DON’T KNOW
##                                              77
##                                  SKIPPED ON WEB
##                                              98
##                                         refused
##                                              99
```

```r
Trumpvoters<- Gs17[Gs17$Q0 == 2,]

# most important = 1
attributes(Gs17$Q13_1)
```

```
## $label
## [1] "[Abortion] What is the most important problem facing this country today?"
##
## $format.spss
## [1] "F8.0"
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
##
## $labels
##     Most important Not most important
##                  1                  0
```

```r
Q13 <- Trumpvoters[ , grepl( "Q13" , names( Gs17 ) ) ]

Q13totals <- vector("double", ncol(Q13))


for(i in seq_along(Q13)){
  Q13totals[[i]]<-sum(Q13[[i]])
}

# Question 21 and Question 6 are the most important
# 21) 43 6) 25
#
43/236
```

```
## [1] 0.1822034
```

```
25/236
```

```
## [1] 0.1059322
```

D

```
# What percentage of 2016 Clinton voters listed these two issues are the most
#  important problem facing the country?

ClintonVoters<- Gs17[Gs17$Q0 == 1,]

ClintonQ13 <- ClintonVoters[ , grepl( "Q13" , names( Gs17 ) ) ]

C13 <- vector("double", ncol(ClintonQ13))

for(i in seq_along(ClintonQ13)){
  C13[[i]]<-sum(ClintonQ13[[i]])
}

38/853
```

```
## [1] 0.04454865
```

```
116/853
```

```
## [1] 0.1359906
```

E Wowmen under 30 and over 30, top issues are 14,6 and 3

```
####### E
# What are the top three issues that women over 30 years old care about?
# Are these top issues the same for women aged 30 and under?

## top 3 over 30 = 14,6,3
## top 3 under 30 = 14,6,3

Wunder30 <- Gs17[Gs17$gender == 2 & Gs17$AGE4 == 1,]

# top 3 over 30
Wover30<- Gs17[Gs17$gender == 2 & Gs17$AGE4 > 1,]
Wover30Q13 <- Gs17[ , grepl( "Q13" , names( Gs17 ) ) ]

Wover30Total <- vector("double", ncol(Wover30Q13))
```

```
for(i in seq_along(Wover30Q13)){
  Wover30Total[[i]]<-sum(Wover30Q13[[i]])
}

# top 3 under 30
Wunder30 <- Gs17[Gs17$gender == 2 & Gs17$AGE4 == 1,]
Wunder30Q13 <- Wunder30[ , grepl( "Q13" , names( Wunder30 ) ) ]

Wunder30Total <- vector("double", ncol(Wunder30))

for(i in seq_along(Wunder30Q13)){
  Wunder30Total[[i]]<-sum(Wunder30Q13[[i]])
}
```

Question 4

A

```
# A
library(readr)
nyc_central_park_temps <- read_csv("nyc-central-park-temps.csv")
```

```
## Rows: 54779 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): STATION, NAME
## dbl  (2): TMAX, TMIN
## date (1): DATE
##
## 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.
```

```
nyc<- separate(nyc_central_park_temps,
         col = DATE,
         into = c('Year', 'Month', 'Day'),
           sep="-")

## Which years are missing at least one day of temperature data, and
# how many days are missing?
## none??

sum(is.na(nyc$TMAX))
```

```
## [1] 0
```

```
sum(is.na(nyc$TMIN))
```

```
## [1] 0
```

```
nyc_central_park_temps %>%
  filter(is.na(TMAX))
```

```
## # A tibble: 0 x 5
## # ... with 5 variables: STATION <chr>, NAME <chr>, DATE <date>, TMAX <dbl>,
## #   TMIN <dbl>
```

B

```
### B
# Create a variable that tells us the
# difference between the highest and lowest temperature for each day.

nyc <- nyc %>%
  group_by(Year, Month, Day) %>%
  mutate(DailyDifference = lag(TMAX,0)-lag(TMIN,0)) %>%
  filter(DailyDifference ==max(DailyDifference))

# Across the full dataset, what the average of this difference?

mean(nyc$DailyDifference)
```

```
## [1] 14.73218
```

```
# Which day during this 150 year window had the biggest difference between the
# highest and lowest temperature?

# 1921, march 28 had highest difference of 48 degrees
nyc %>%
  filter(DailyDifference == 48)
```

```
## # A tibble: 1 x 8
## # Groups:   Year, Month, Day [1]
##   STATION     NAME                 Year  Month Day    TMAX  TMIN DailyDifference
##   <chr>       <chr>                <chr> <chr> <chr> <dbl> <dbl>           <dbl>
## 1 USW00094728 NY CITY CENTRAL PAR~ 1921  03    28       82    34              48
```

```
#Averaging across years, which month tends to
# have the highest average difference in daily high and low temperatures?

# May

nycsummary <- nyc %>%
  group_by(Year, Month) %>%
  mutate(average.monthly = mean(DailyDifference))

aggregate(DailyDifference ~ Month,
          nycsummary,
          mean)
```

```
##    Month DailyDifference
## 1     01        12.60602
## 2     02        13.49717
## 3     03        14.59742
## 4     04        16.31600
## 5     05        17.18156
## 6     06        16.68089
## 7     07        15.98731
## 8     08        15.40796
## 9     09        15.29867
## 10    10        14.72645
## 11    11        12.63667
## 12    12        11.80839
```

C

```
####### C
#  Load and merge in the precipitation data.
# What type of merge does it mark
# sense to perform? Which variable(s) will you merge on? Perform the merge,
# then use the results to figure out how many days in the past 150 years had a
# high temperature of at least 50 degrees and received at least 1 inch of snowfall.

nyc_central_park_precipitation <- read_csv("nyc-central-park-precipitation.csv")
```

```
## Rows: 54786 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): STATION, NAME
## dbl  (2): PRCP, SNOW
## date (1): DATE
##
## 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.
```

```
nyc_central_park_precipitation<- separate(nyc_central_park_precipitation,
               col = DATE,
               into = c('Year', 'Month', 'Day'),
               sep="-")

weather <- merge(x = nyc_central_park_precipitation,
              y = nyc,
              by = c("STATION","Year","Month","Day"),
              all = T)

# how many days in the past 150 years had a high temperature of at least 50
# degrees and received at least 1 inch of snowfall

weather %>%
  filter(TMAX >= 50) %>%
  filter ( SNOW >= 1 ) %>%
  summarise(n())
```

```
##   n()
## 1  25
```

D

Which month tends to have the most rainy days in New York City? March

What percentage of days does it usually rain in this month? 36.43011%

And which month tends to be the dryest (i.e. fewest days with precipitation)? What percentage?

October, 27.20430%

```r
###### D
# Aggregate the data by month to figure out what percentage of days have had
# preciptation since 1869.
# Your resulting dataset should have 12 rows (one per month).
# You should use the PRCP variable (and ignore the SNOW variable).

weather <- weather %>%
  mutate(Dummy = ifelse(PRCP >0,T,F))

aggregate(Dummy ~ Month,
          weather,
          mean)
```

```
##    Month     Dummy
## 1     01 0.3552688
## 2     02 0.3427762
## 3     03 0.3643011
## 4     04 0.3582222
## 5     05 0.3578495
## 6     06 0.3455556
## 7     07 0.3374194
## 8     08 0.3161290
## 9     09 0.2795556
## 10    10 0.2720430
## 11    11 0.3075556
## 12    12 0.3384946
```

```r
# Which month tends to have the most rainy days in New York City?
# March



# What percentage of days does it usually rain in this month?

# 36.43011%

# And which month tends to be the dryest (i.e. fewest days with precipitation)?
# What percentage?

# October, 27.20430%
```

E

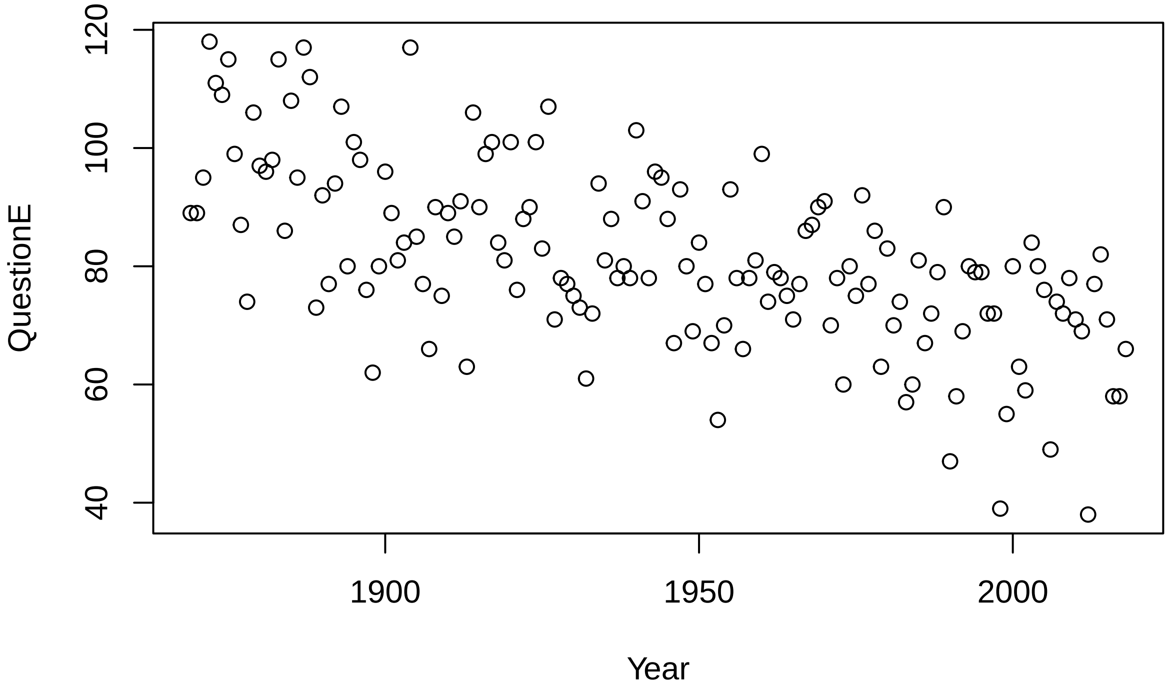
There is a steady decline of days below 32 degrees since 1869

```
####### E
# Use aggregation to figure out how many days in each year since 1869
# had a low temperature of 32 degree or below.

weather <- weather %>%
  mutate(QuestionE = ifelse(TMIN <= 32, T, F))

QE <-aggregate(QuestionE ~ Year,
          weather,
          sum)

# Use the plot() function or ggplot2 to make a simple graph of the relationship
# between the year (on the x-axis) and the number of cold days in
# Central Park (on the y-axis).
QE %>%
  plot("Year", "QuestionE" )
```



```
# What pattern do you notice in this graph

#There is a steady decline of days below 32 degrees since 1869
```