STA205 Final Project - Exploring the GSS Part I

Introduction

In this study, we use the 2016 General Social Survey (GSS), a dataset on trends in attitudes, behaviours, and attributes of people from the United States, to investigate various social issues of interest. We first discuss and analyze work harassment before delving into time spent on emails. Lastly, we attempt to find correlation between political views and views on scientific research.

Section I: Harassment at Work

1. What are the possible responses to this question and how many respondents chose each of these answers?

```
gss16 %>%
  filter(!is.na(harass5)) %>%
  group_by(harass5) %>%
  summarize(count=n())
```

```
## # A tibble: 3 x 2
##   harass5                                                 count
##   <chr>                                                   <int>
## 1 Does not apply (i do not have a job/superior/co-worker)    96
## 2 No                                                       1136
## 3 Yes                                                       237
```

A: The possible responses to this question are "Does not apply (i do not have a job/superior/co-worker)" (96 people), "No" (1136 people), and "Yes" (237 people).

2. What percent of the respondents for whom this question is applicable (i.e. excluding `NA`s and `Does not applys`) have been harassed by their superiors or co-workers at their job.

```
gss16 %>%
  filter(!is.na(harass5), harass5 != "Does not apply (i do not have a job/superior/co-worker)") %>%
  group_by(harass5) %>%
  summarize(count=n()) %>%
  mutate(percent = count/sum(count))
```

```
## # A tibble: 2 x 3
##   harass5 count percent
##   <chr>   <int>   <dbl>
## 1 No       1136   0.827
## 2 Yes       237   0.173
```

A: 17.26% of respondents that are applicable for this question have been harassed by their superiors or coworkers at their jobs.

Section II: Time Spent On Email

The 2016 GSS also asked respondents how many hours and minutes they spend on email weekly. The responses to these questions are recorded in the `emailhr` and `emailmin` variables. For example, if the response is 2.5 hrs, this would be recorded as `emailhr = 2` and `emailmin = 30`.

3. Create a new variable called `email` that combines these two variables to reports the number of minutes the respondents spend on email weekly.

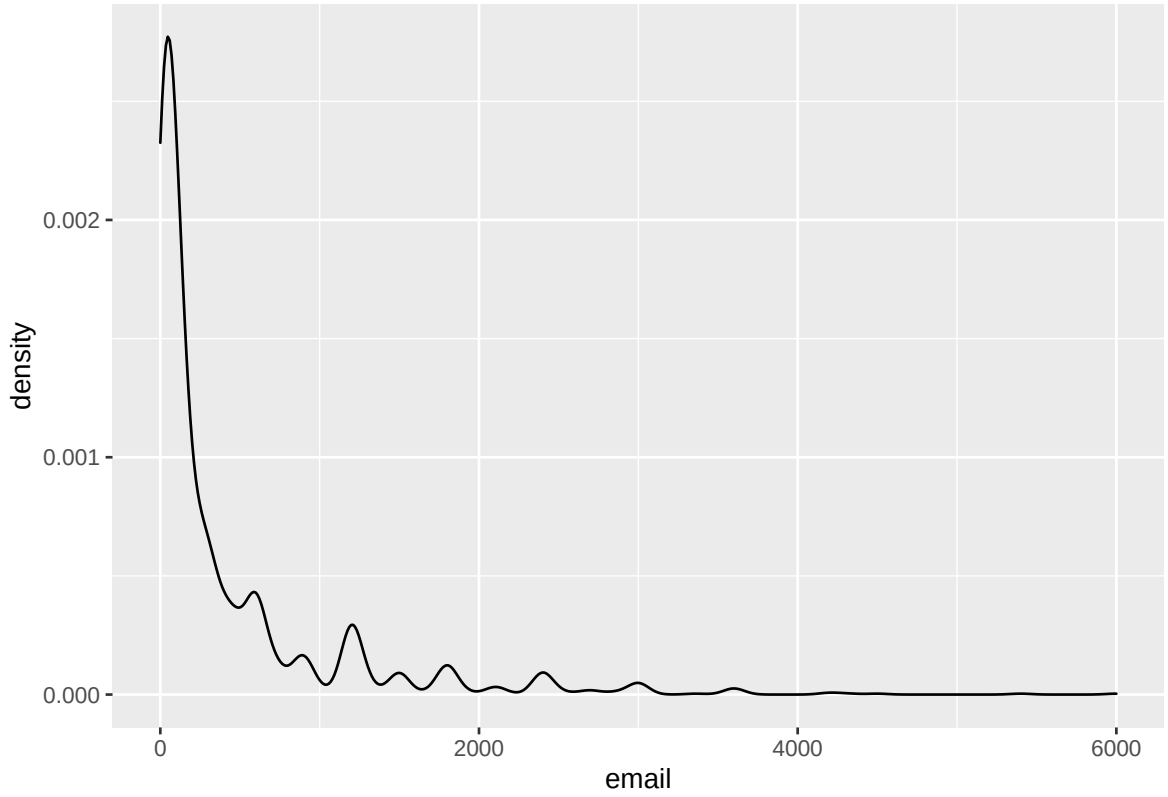
```
(gss16 <- gss16 %>%
  mutate(email = 60 * emailhr + emailmin))
```

```
## # A tibble: 2,867 x 10
##    harass5 emailmin emailhr  educ polviews    advfront snapchat instagrm wrkstat
##    <chr>      <dbl>   <dbl> <dbl> <chr>       <chr>    <chr>    <chr>    <chr>
##  1 <NA>           0      12    16 Moderate    Strongl~ <NA>     <NA>     Workin~
##  2 <NA>          30       0    12 Liberal     Disagree No       No       Workin~
##  3 No            NA      NA    16 Conservati~ <NA>     No       No       Retired
##  4 <NA>          10       0    12 Moderate    Disagree <NA>     <NA>     Workin~
##  5 No            NA      NA    18 Slightly l~ <NA>     Yes      Yes      Workin~
##  6 <NA>           0       2    14 Slightly l~ Strongl~ No       Yes      Keepin~
##  7 <NA>           0      40    14 Slightly l~ Strongl~ <NA>     <NA>     Workin~
##  8 No            NA      NA    11 Slghtly co~ <NA>     Yes      Yes      Workin~
##  9 <NA>           0       0    12 <NA>        Agree    <NA>     <NA>     Workin~
## 10 No            NA      NA    14 Conservati~ <NA>     No       No       Retired
## # ... with 2,857 more rows, and 1 more variable: email <dbl>
```

4. Visualize the distribution of this new variable. Find the mean and the median number of minutes respondents spend on email weekly. Is the mean or the median a better measure of the typical among of time Americans spend on email weekly? Why?

```
ggplot(data=gss16) +
geom_density(mapping=aes(x=email))
```

```
## Warning: Removed 1218 rows containing non-finite values (stat_density).
```




```
summary(gss16)
```

```
##    harass5             emailmin         emailhr             educ
##  Length:2867        Min.   : 0.000   Min.   :  0.000   Min.   : 0.00
##  Class :character   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:12.00
##  Mode  :character   Median : 0.000   Median :  2.000   Median :13.00
##                     Mean   : 3.319   Mean   :  6.892   Mean   :13.74
##                     3rd Qu.: 0.000   3rd Qu.:  8.000   3rd Qu.:16.00
##                     Max.   :59.000   Max.   :100.000   Max.   :20.00
##                     NA's   :1218     NA's   :1218      NA's   :9
##    polviews           advfront           snapchat           instagrm
##  Length:2867        Length:2867        Length:2867        Length:2867
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    wrkstat              email
##  Length:2867        Min.   :   0.0
##  Class :character   1st Qu.:  50.0
##  Mode  :character   Median : 120.0
##                     Mean   : 416.8
##                     3rd Qu.: 480.0
##                     Max.   :6000.0
##                     NA's   :1218
```

A: The distribution of the number of minutes spend on email weekly is a right-skewed distribution.

The median is 120 minutes, and the mean is 416.8 minutes. In this case our median seems to be a better measure because it is closer to the highest point of the curve (which is where the majority of the population lies).

5. Create another new variable, `snap_insta` that is coded as "Yes" if the respondent reported using any of Snapchat (`snapchat`) or Instagram (`instagrm`), and "No" if not. If the recorded value was `NA` for both of these questions, the value in your new variable should also be `NA`.

```
gss16 <- gss16 %>%
mutate(snap_NA =
           ifelse(is.na(snapchat) & is.na(instagrm), NA, "No")
) %>%
mutate(snap_insta =
           ifelse((snapchat == "Yes" | instagrm == "Yes"), "Yes", snap_NA))
gss16 <- subset(gss16, select=-c(snap_NA))
gss16
```

```
## # A tibble: 2,867 x 11
##    harass5 emailmin emailhr  educ polviews    advfront snapchat instagrm wrkstat
##    <chr>      <dbl>   <dbl> <dbl> <chr>       <chr>    <chr>    <chr>    <chr>
##  1 <NA>           0      12    16 Moderate    Strongl~ <NA>     <NA>     Workin~
##  2 <NA>          30       0    12 Liberal     Disagree No       No       Workin~
##  3 No            NA      NA    16 Conservati~ <NA>     No       No       Retired
##  4 <NA>          10       0    12 Moderate    Disagree <NA>     <NA>     Workin~
##  5 No            NA      NA    18 Slightly l~ <NA>     Yes      Yes      Workin~
##  6 <NA>           0       2    14 Slightly l~ Strongl~ No       Yes      Keepin~
##  7 <NA>           0      40    14 Slightly l~ Strongl~ <NA>     <NA>     Workin~
##  8 No            NA      NA    11 Slghtly co~ <NA>     Yes      Yes      Workin~
##  9 <NA>           0       0    12 <NA>        Agree    <NA>     <NA>     Workin~
## 10 No            NA      NA    14 Conservati~ <NA>     No       No       Retired
## # ... with 2,857 more rows, and 2 more variables: email <dbl>, snap_insta <chr>
```

6. Calculate the percentage of Yes's for `snap_insta` among those who answered the question, i.e. excluding `NA`s.

```
gss16 %>%
  filter(!is.na(snap_insta)) %>%
  group_by(snap_insta) %>%
  summarize(n=n()) %>%
  transmute(snap_insta, n, perc=n/sum(n))
```

```
## # A tibble: 2 x 3
##   snap_insta     n  perc
##   <chr>      <int> <dbl>
## 1 No           858 0.625
## 2 Yes          514 0.375
```

A: The percentage of people who have either Instagram or Snapchat is (surprisingly) only 37.46 percent.

7. What are the possible responses to the question *Last week were you working full time, part time, going to school, keeping house, or what?* and how many respondents chose each of these answers? Note that this information is stored in the `wrkstat` variable.

```
gss16 %>%
  group_by(wrkstat) %>%
  summarize(count=n()) %>%
  arrange(desc(count))
```

```
## # A tibble: 9 x 2
##   wrkstat          count
##   <chr>            <int>
## 1 Working fulltime  1321
## 2 Retired            574
## 3 Working parttime   345
## 4 Keeping house      284
## 5 Unempl, laid off   118
## 6 Other               89
## 7 School              76
## 8 Temp not working    57
## 9 <NA>                 3
```

A: The possible responses to the question are working fulltime (1321), retired (574), working part time (345), keeping house (284), unemployed or laid off (118), other (89), school (76), and temporarily not working (3).

8. Fit a model predicting `email` (number of minutes per week spent on email) from `educ` (number of years of education), `wrkstat`, and `snap_insta`. Interpret the slopes for each of these variables.

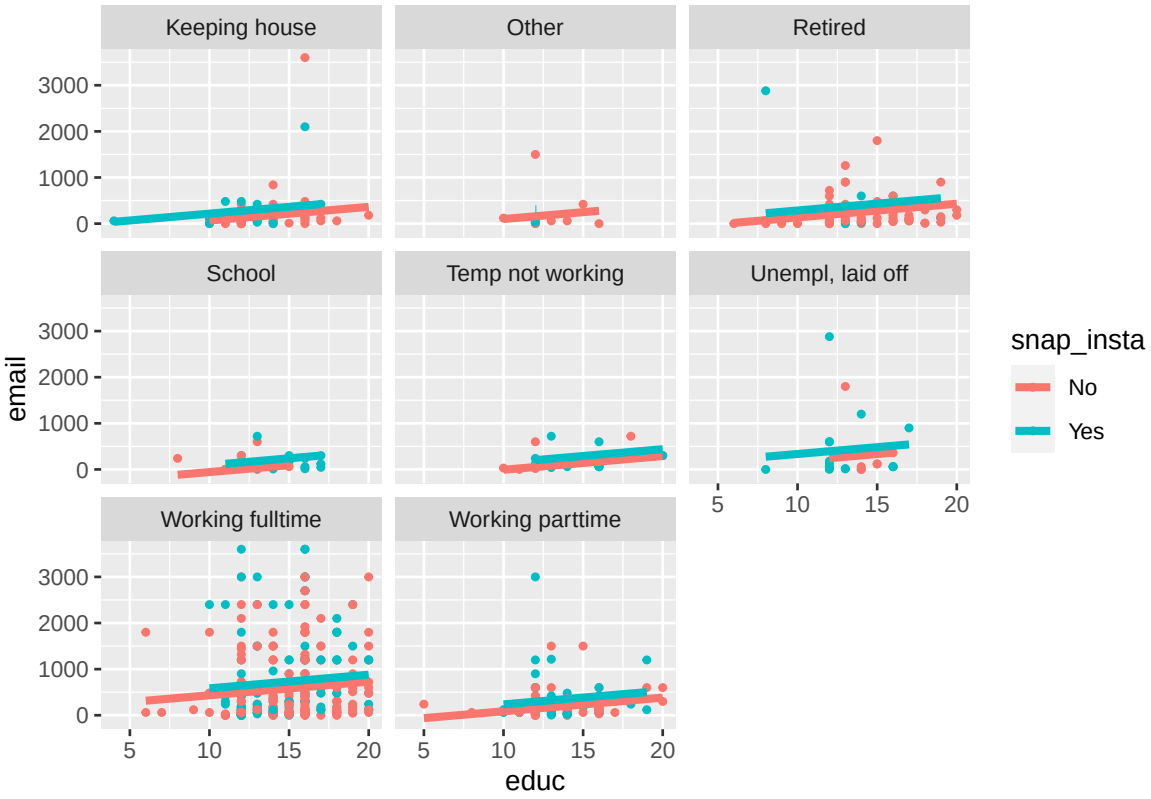
```
filteredgss <- gss16 %>%
  filter(!is.na(educ) & !is.na(email))
email_main_fit <- linear_reg() %>%
  set_engine("lm") %>%
  # fit(email ~ educ, data = filteredgss)
  fit(email ~ educ + wrkstat + snap_insta, data = gss16)

email_main_fit %>% tidy()
```

```
## # A tibble: 10 x 5
##    term                     estimate std.error statistic   p.value
##    <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)               -230.     150.      -1.53  0.126
##  2 educ                        29.6      9.60     3.09  0.00211
##  3 wrkstatOther                33.1    209.       0.158 0.875
##  4 wrkstatRetired              68.3    111.       0.615 0.539
##  5 wrkstatSchool             -124.     144.      -0.860 0.390
##  6 wrkstatTemp not working    -73.7    154.      -0.479 0.632
##  7 wrkstatUnempl, laid off    118.     151.       0.783 0.434
##  8 wrkstatWorking fulltime    367.      87.7      4.18  0.0000326
##  9 wrkstatWorking parttime     18.9    102.       0.186 0.853
## 10 snap_instaYes              150.      52.7      2.84  0.00460
```

```
email_main_fit_aug <- augment(email_main_fit$fit)
```

```
email_main <- ggplot(email_main_fit_aug) +
geom_point(mapping = aes(x=educ, y= email, color=snap_insta), size=1) +
```

```
geom_line(mapping = aes(x=educ, y=.fitted, color = snap_insta), size=1.5) +
facet_wrap(~wrkstat)
email_main
```



A: For both Instagram and Snapchat users, the relationship between amount of time spent on emails and one's education seems to show only minor positive correlation. Even across different occupations, the difference was minor. Perhaps we can say that one's education has slight correlation with how much time one spends on their email.

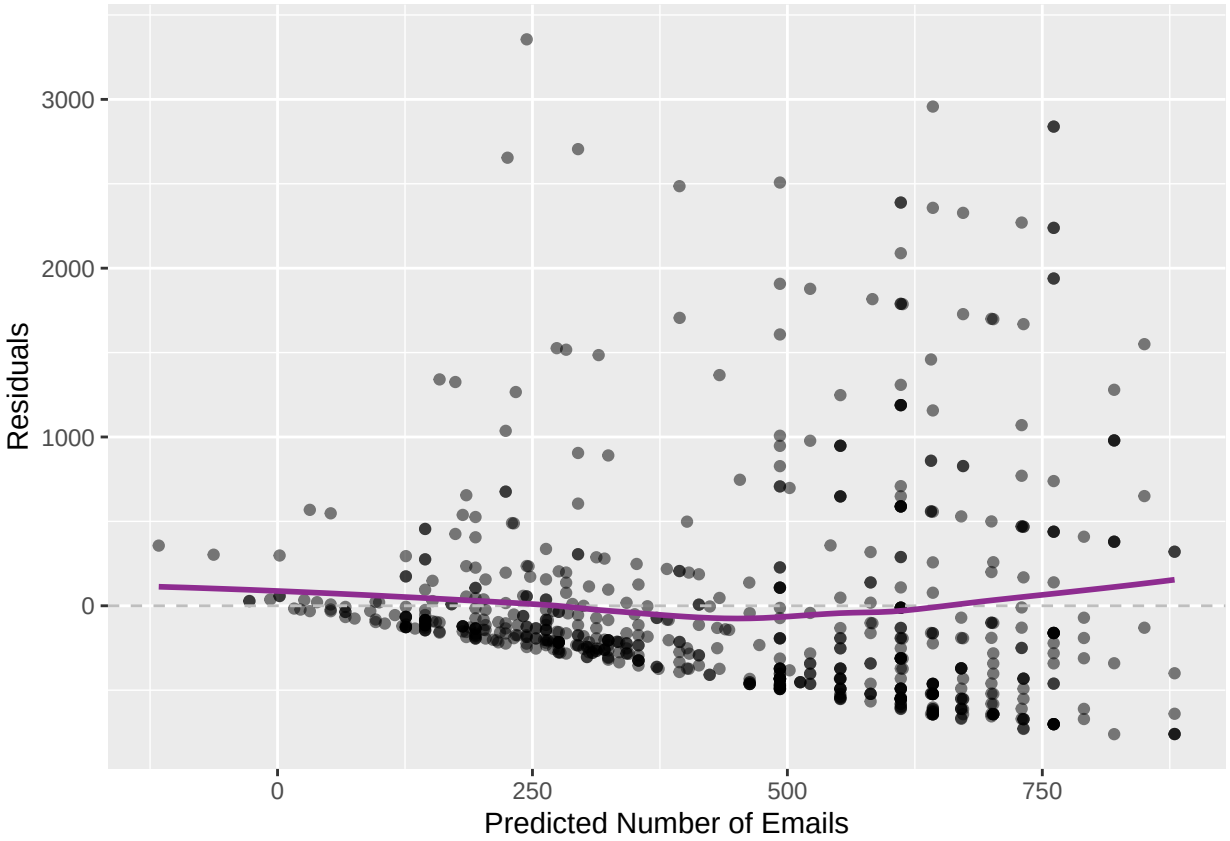
Another trend we see through the graphs is that those who use either Instagram or Snapchat tend to check their email more. This might, of course be connected with more device usage (those who use social media tend to use their devices more).

Slope interpretation: With all else held constant, for each additional year on education, the number of minutes spend on email weekly is expected to be higher, on average, by a factor of 29.63.

9. Create a predicted values vs. residuals plot for this model. Are there any issues with the model? If yes, describe them.

```
ggplot(email_main_fit_aug, mapping = aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "gray", lty = "dashed") +
    geom_smooth(color = "#8E2C90", se = FALSE) +
  labs(x = "Predicted Number of Emails", y = "Residuals")
```

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



A: When constructing a residuals plot, we look for points that are randomly scattered around the x-axis, with no specific pattern. If this is not the case, then a linear model may not be the best model to for prediction. In this example, the points are not randomly scattered, as the majority of them are close to the x-axis and sometimes clusters form. There is almost no correlation between the residuals and predicted values.

Section III: Political Views and Science Research

10. In a new variable, recode `advfront` such that Strongly Agree and Agree are mapped to `"Yes"`, and Disagree and Strongly disagree are mapped to `"No"`. The remaining levels can be left as is. Don't overwrite the existing `advfront`, instead pick a different, informative name for your new variable.

```
gss16 <- gss16 %>%
  mutate(advYesNo = advfront)
gss16$advYesNo[gss16$advfront == "Strongly agree" | gss16$advfront == "Agree" ] <- "Yes"
gss16$advYesNo[gss16$advfront == "Strongly disagree" | gss16$advfront == "Disgree" ] <- "No"
select(gss16, advfront, advYesNo)
```

```
## # A tibble: 2,867 x 2
##    advfront       advYesNo
##    <chr>          <chr>
##  1 Strongly agree Yes
##  2 Disagree       Disagree
##  3 <NA>           <NA>
##  4 Disagree       Disagree
```

```
##  5 <NA>           <NA>
##  6 Strongly agree Yes
##  7 Strongly agree Yes
##  8 <NA>           <NA>
##  9 Agree          Yes
## 10 <NA>           <NA>
## # ... with 2,857 more rows
```

11. In a new variable, recode polviews such that Extremely liberal, Liberal, and Slightly liberal, are mapped to "Liberal", and Slghtly conservative, Conservative, and Extrmly conservative disagree are mapped to "Conservative". The remaining levels can be left as is. Make sure that the levels are in a reasonable order. Don't overwrite the existing polviews, instead pick a different, informative name for your new variable.

```
gss16 <- gss16 %>%
  mutate(ConsOrLib = polviews)
gss16$ConsOrLib[gss16$polviews == "Extremely liberal" | gss16$polviews == "Slightly liberal"] <- "L
gss16$ConsOrLib[gss16$polviews == "Extrmly conservative" | gss16$polviews == "Slghtly conservative"
gss16 %>%
  select(polviews, ConsOrLib)
```

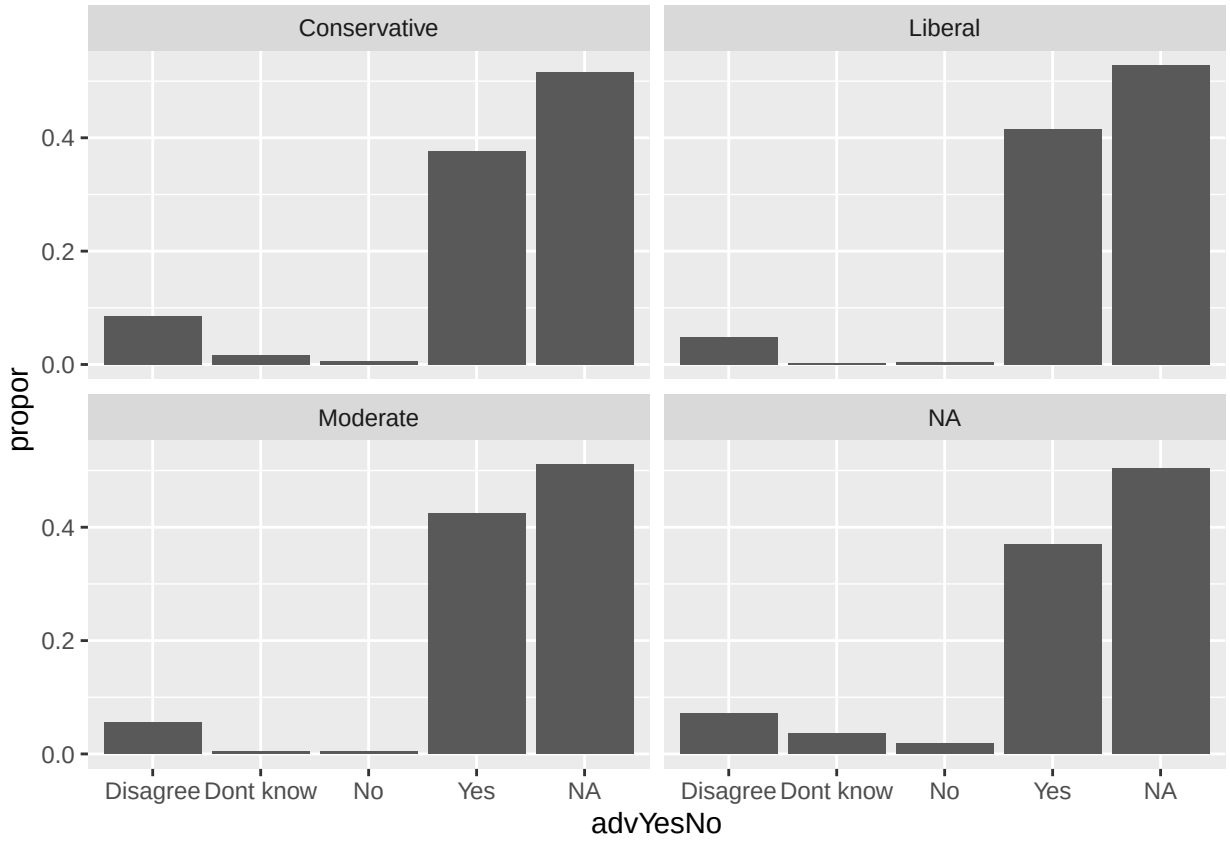
```
## # A tibble: 2,867 x 2
##    polviews             ConsOrLib
##    <chr>                <chr>
##  1 Moderate             Moderate
##  2 Liberal              Liberal
##  3 Conservative         Conservative
##  4 Moderate             Moderate
##  5 Slightly liberal     Liberal
##  6 Slightly liberal     Liberal
##  7 Slightly liberal     Liberal
##  8 Slghtly conservative Conservative
##  9 <NA>                 <NA>
## 10 Conservative         Conservative
## # ... with 2,857 more rows
```

12. Create a visualization that displays the relationship between these two new variables and interpret it.

```
vis_data = gss16 %>%
  group_by(ConsOrLib, advYesNo) %>%
  summarize(count = n()) %>%
  mutate(propor = count/sum(count))
```

```
## `summarise()` has grouped output by 'ConsOrLib'. You can override using the
## `.groups` argument.
```

```
ggplot(data=vis_data) +
  geom_col(aes(y=propor,x=advYesNo)) +
  facet_wrap(~ConsOrLib)
```

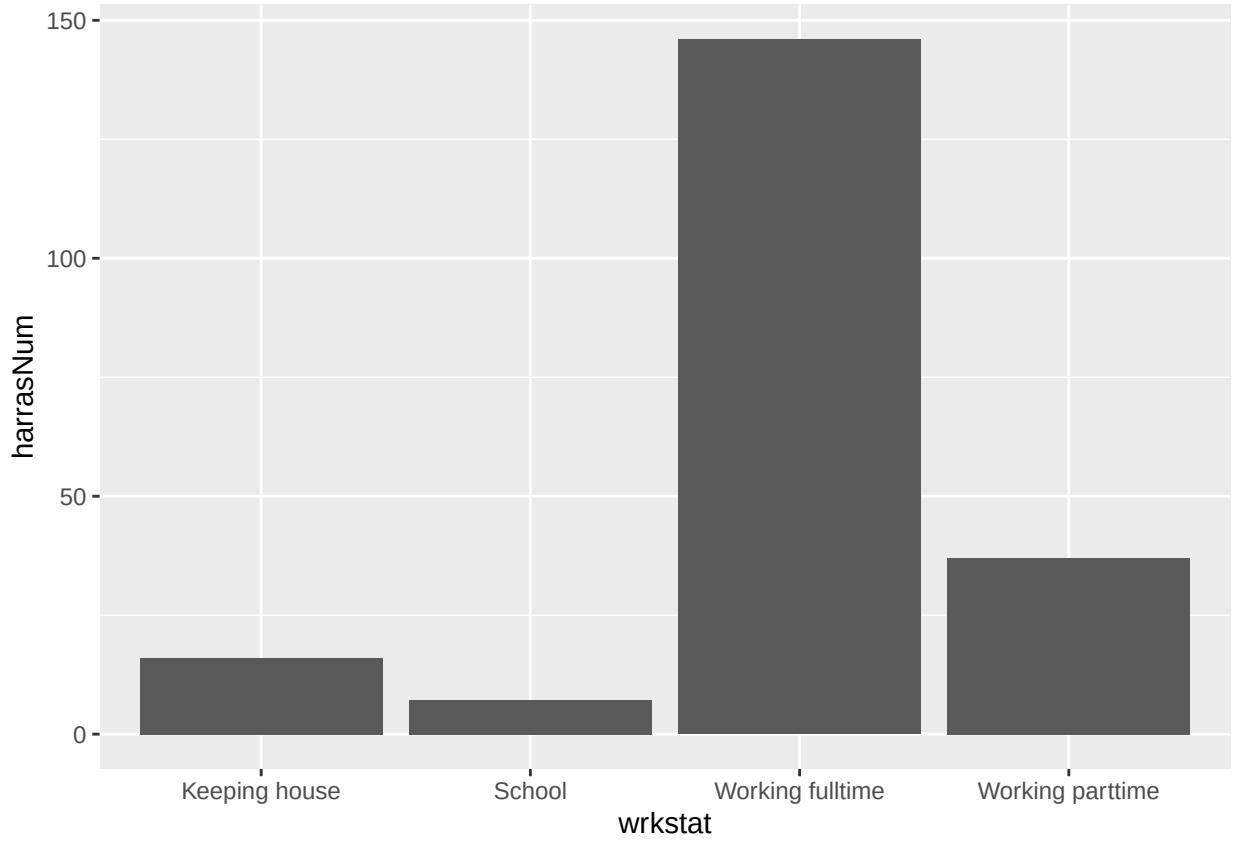


A: People who are more liberal or moderate tend to agree more with the federal government supporting scientific research and its necessity. Conservatives have the most percentage of people who disagree, followed by moderates and liberals.Perhaps the more liberal one is, the more likely that he or she supports scientific research.

Section IV: Personal Questions

13. For those that are still working, how does work status relate to harrassment at work? Are full time workers less likely or more likely to be harrassed?

```
gss16 <- gss16 %>%
  mutate(harrasNum = ifelse(harass5 == "Yes", 1, 0))
HarrassVsWrkstat <- gss16 %>%
  filter(!is.na(harass5) & !is.na(wrkstat)
         & wrkstat %in%
           c("Working fulltime", "Working parttime", "School", "Keeping house"))

ggplot(HarrassVsWrkstat) + geom_col(aes(x=wrkstat, y=harrasNum))
```



A: From the graph, we see that those who work fulltime are more likely to be harassed. This makes sense when we consider that those who work fulltime work more and thus have a higher chance of exposure to work harassment.

14. Do one's political views relate to harrassment at work?

```
vis_data <- gss16 %>%
  filter(!is.na(harass5) & !is.na(ConsOrLib)) %>%
  filter(harass5 == "Yes" | harass5 == "No") %>%
  group_by(ConsOrLib, harrasNum) %>%
  summarize(n=n()) %>%
  mutate(propor = n/sum(n),
         harass = ifelse(harrasNum == 0, "No", "Yes"))
```

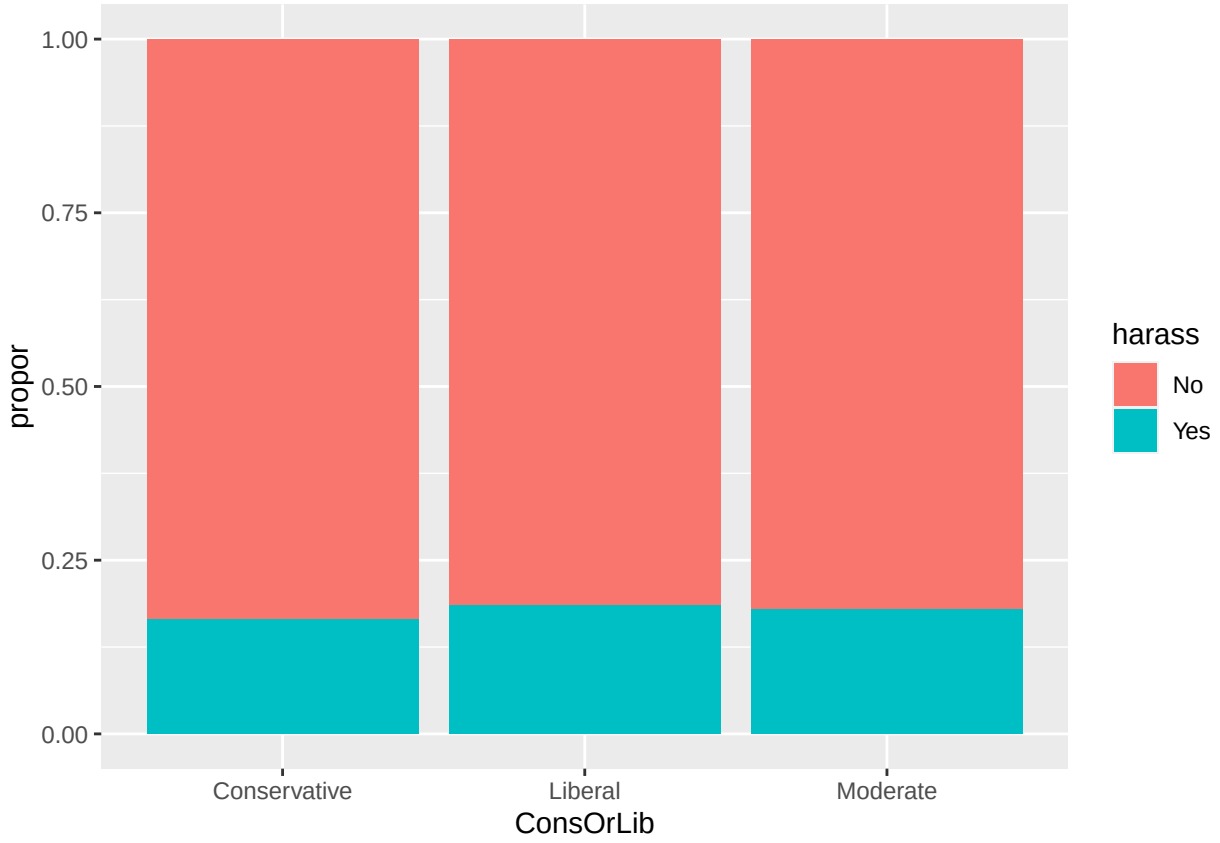
```
## `summarise()` has grouped output by 'ConsOrLib'. You can override using the
## `.groups` argument.
```

```
vis_data
```

```
## # A tibble: 6 x 5
## # Groups:   ConsOrLib [3]
##   ConsOrLib    harrasNum     n propor harass
##   <chr>            <dbl> <int>  <dbl> <chr>
## 1 Conservative         0   366  0.836 No
## 2 Conservative         1    72  0.164 Yes
```

```
## 3 Liberal              0   327  0.815 No
## 4 Liberal              1    74  0.185 Yes
## 5 Moderate             0   400  0.821 No
## 6 Moderate             1    87  0.179 Yes
```

```
ggplot(vis_data) +
  geom_col(aes(x=ConsOrLib, y=propor, fill=harass))
```



A: The proportion of those who have been harassed at work and those who haven't seem to be identical across political views. We might conclude that political views barely affects work harassment.

Conclusion

Through this analysis, we've realized a few things.

1. Work harassment is still a problem that we need to recognize, especially for fulltime workers.
2. Those who are more educated and use social media more may be slightly more likely to spend more time on his or her email.
3. The more conservative one's political views are, the more likely he/she disagrees with federal support of scientific research and its necessity.

   However, the scope and validity of these questions is limited because of the data. We have no way of knowing from which population group the samples were taken from, and if there's any sampling bias involved. Although the data was from 2016, a noticeable portion of those interviewed were retired, so

the age distribution may be left-skewed. All of these unknown factors will affect our interpretation of the data and the understanding of our analysis.