Technical Report

Shisham Adhikari, Maggie Slein, Grayson White

Data (and data cleaning)

The GSS dataset

(Info about the GSS)

Filtering

Most of this filtering was done for the `infer` package `gss` dataset and can be attributed to authors of that package. We have included more rows and columns than that package, however, much initial tidying and subsetting can be attributed to them (Bray et al. 2020).

```r
load("gss/gss_orig.rda")
gss_subset <- gss_orig %>%
  filter(!stringr::str_detect(sample, "blk oversamp")) %>% # this is for weighting
  select(
    year,
    age,
    sex,
    college = degree,
    partyid,
    hompop,
    hours = hrs1,
    income,
    class,
    finrela,
    wrkgovt,
    marital,
    educ,
    race,
    incom16,
    weight = wtssall
  ) %>%
  mutate_if(is.factor, ~ fct_collapse(., NULL = c("IAP", "NA", "iap", "na"))) %>%
  mutate(
    age = age %>%
      fct_recode("89" = "89 or older",
                 NULL = "DK") %>%
      as.character() %>%
      as.numeric(),
    hompop = hompop %>%
      fct_collapse(NULL = c("DK")) %>%
      as.character() %>%
      as.numeric(),
    hours = hours %>%
      fct_recode("89" = "89+ hrs",
                 NULL = "DK") %>%
      as.character() %>%
      as.numeric(),
    weight = weight %>%
      as.character() %>%
      as.numeric(),
    partyid = fct_collapse(
      partyid,
      dem = c("strong democrat", "not str democrat"),
      rep = c("strong republican", "not str republican"),
      ind = c("ind,near dem", "independent", "ind,near rep"),
      other = "other party"
    ),
    income = factor(income, ordered = TRUE),
    college = fct_collapse(
      college,
      degree = c("junior college", "bachelor", "graduate"),
      "no degree"  = c("lt high school", "high school"),
      NULL = "dk"
    )
  ) %>%
  filter(year >= 2000) %>%
  filter(partyid %in% c("dem", "rep")) %>%
  drop_na()
```

Given our goal to understand which factors influence party affiliation in the US, we selected year (year of the election), age (age of voter at election), college degree (degree or no degree), partyid (democrat or republican), *hompop* (what does this represent), *hours=hrs1* (what does this represent), income (what are the bounds of the predictors–are there groups), class (?), finrela (?), wrkgovt(?), marital(?), educ (?), race (?), income16(?), weight = wtssall (?)

Why did we chose these from the dataset? Why did we exclude other variables? What are the possible implications of this?

Exploratory Data Analysis

A presentation of graphical and numerical summaries of the data (along with a discussion of their relevance to modeling assumptions and further analysis), a description of the statistical methods used to analyze your data, and diagnostics of the appropriateness of any models or inference procedures you will apply in the Results section.

Below are plots that show the distribution of political party affiliation between democrat and republican as well as the distrutbution of all the predictors included in this dataset.There appears to me more democrats than republicans represented in this dataset, which could be because democrats are more likely to participate in this survey or it could be by chance. Most of our predictors appear to normally distributed, except for income, hompop, and weight. None of the predictors appear to have a strong relationship with political party affiliation, which is not surprising given that there are roughly the same amount of democrats and republicans in each state.

1. We also need to talk about any potential collinearity but I'm not sure how to do that 2. Any statistical or numeric summaries that are missing here?

```r
#checking data structure

nrow(gss_subset)
```

```
## [1] 5800
```

```r
ncol(gss_subset)
```

```
## [1] 16
```

```r
str(gss_subset)
```

```
## tibble [5,800 x 16] (S3: tbl_df/tbl/data.frame)
##  $ year   : num [1:5800] 2002 2002 2002 2002 2002 ...
##   ..- attr(*, "label")= chr "gss year for this respondent "
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ age    : num [1:5800] 25 43 46 71 37 23 33 57 42 63 ...
##  $ sex    : Factor w/ 2 levels "male","female": 2 1 1 2 1 1 1 1 2 1 ...
##  $ college: Factor w/ 2 levels "no degree","degree": 1 2 1 1 1 1 2 2 2 2 ...
##  $ partyid: Factor w/ 5 levels "dem","ind","rep",..: 3 3 3 3 3 1 1 1 1 1 ...
##  $ hompop : num [1:5800] 1 1 2 1 1 3 4 2 1 1 ...
##  $ hours  : num [1:5800] 40 72 40 24 50 60 70 40 65 44 ...
##  $ income : Ord.factor w/ 12 levels "lt $1000"<"$1000 to 2999"<..: 12 12 12 11 12 12 12 12 12 12 ...
##  $ class  : Factor w/ 6 levels "lower class",..: 3 3 3 2 3 2 2 3 2 3 ...
##  $ finrela: Factor w/ 6 levels "far below average",..: 3 4 4 3 3 3 3 3 4 4 ...
##  $ wrkgovt: Factor w/ 3 levels "government","private",..: 2 2 2 2 2 2 2 2 2 1 ...
##  $ marital: Factor w/ 5 levels "married","widowed",..: 3 1 3 3 5 4 1 1 5 5 ...
##  $ educ   : Factor w/ 22 levels "0","1","2","3",..: 15 17 15 13 16 13 17 17 17 18 ...
##  $ race   : Factor w/ 3 levels "white","black",..: 1 1 1 1 1 2 3 1 1 1 ...
##  $ incom16: Factor w/ 7 levels "far below average",..: 3 4 4 3 2 3 3 4 2 4 ...
##  $ weight : num [1:5800] 0.558 0.558 1.116 0.558 0.558 ...
```

```r
head(gss_subset)
```

```
## # A tibble: 6 x 16
##    year   age sex   college partyid hompop hours income class finrela wrkgovt
##   <dbl> <dbl> <fct> <fct>   <fct>    <dbl> <dbl> <ord>  <fct> <fct>   <fct>
## 1  2002    25 fema~ no deg~ rep          1    40 $2500~ midd~ average private
## 2  2002    43 male  degree  rep          1    72 $2500~ midd~ above ~ private
## 3  2002    46 male  no deg~ rep          2    40 $2500~ midd~ above ~ private
## 4  2002    71 fema~ no deg~ rep          1    24 $2000~ work~ average private
## 5  2002    37 male  no deg~ rep          1    50 $2500~ midd~ average private
## 6  2002    23 male  no deg~ dem          3    60 $2500~ work~ average private
## # ... with 5 more variables: marital <fct>, educ <fct>, race <fct>,
## #   incom16 <fct>, weight <dbl>
```

```
tail(gss_subset)
```

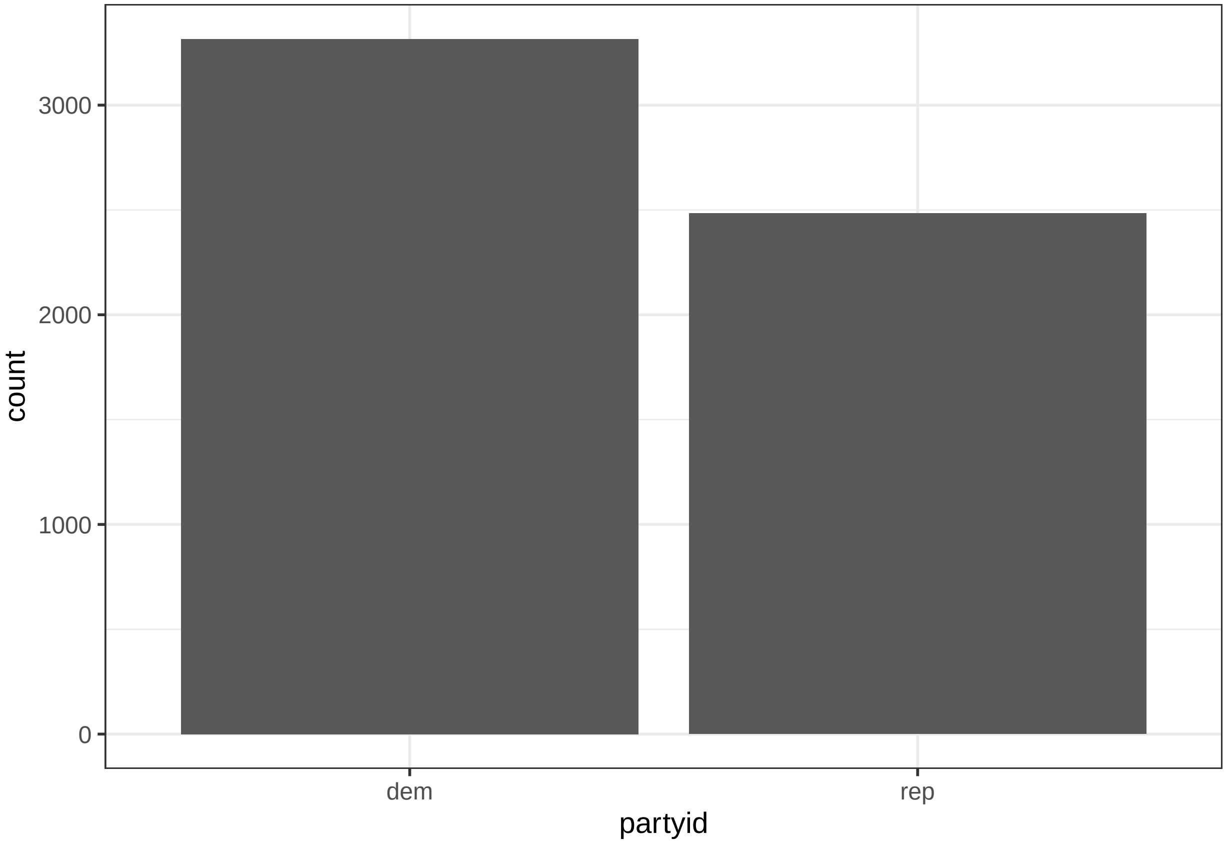
```
## # A tibble: 6 x 16
##    year   age sex   college partyid hompop hours income class finrela wrkgovt
##   <dbl> <dbl> <fct> <fct>   <fct>    <dbl> <dbl> <ord>  <fct> <fct>   <fct>
## 1  2018    21 fema~ no deg~ dem          7    42 $8000~ work~ average govern~
## 2  2018    28 fema~ no deg~ dem          2    40 $2000~ work~ average private
## 3  2018    56 male  degree  rep          2    44 $2500~ midd~ above ~ private
## 4  2018    53 male  degree  rep          2    46 $2500~ midd~ above ~ private
## 5  2018    43 fema~ degree  rep          2    40 $2500~ midd~ average private
## 6  2018    75 fema~ no deg~ rep          2    36 $2500~ work~ below ~ private
## # ... with 5 more variables: marital <fct>, educ <fct>, race <fct>,
## #   incom16 <fct>, weight <dbl>
```

```
party_afill<-gss_subset$partyid
summary(party_afill)
```

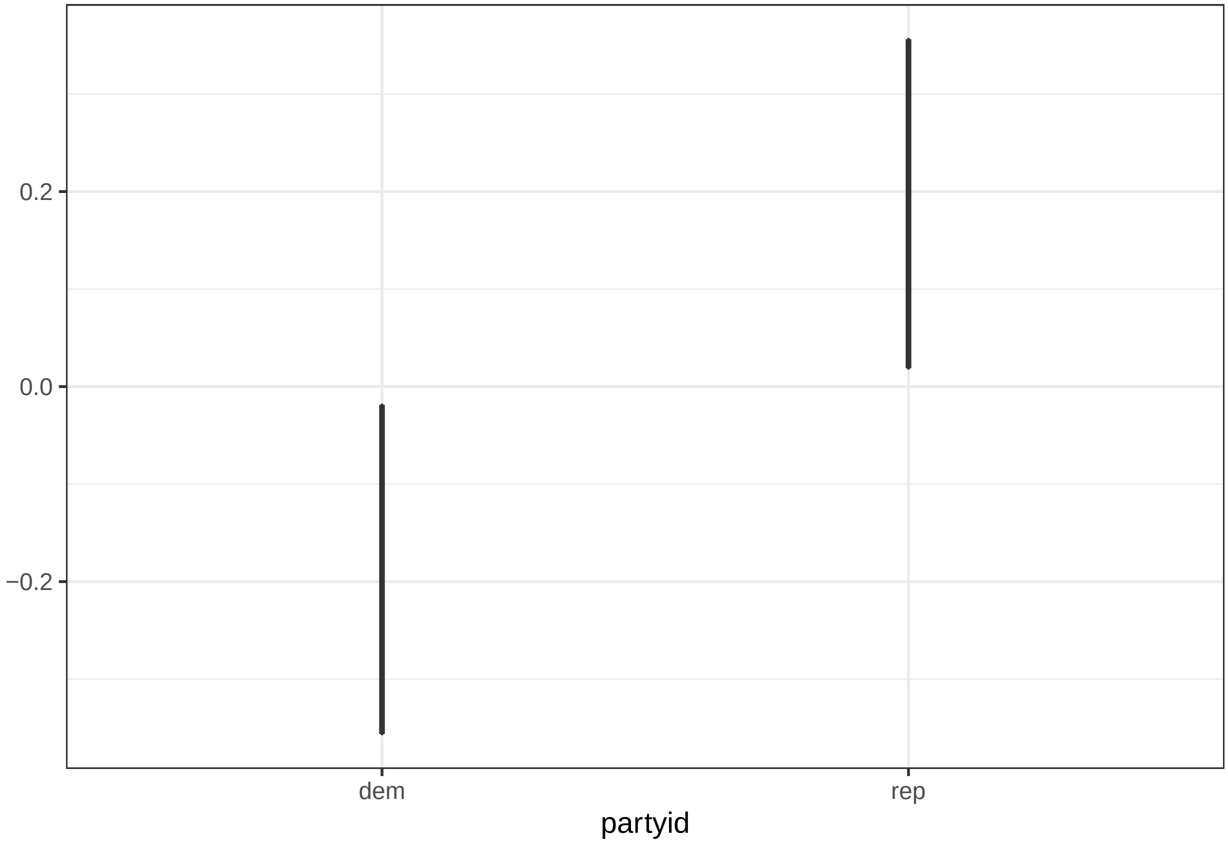
```
##   dem   ind   rep other    DK
##  3316     0  2484     0     0
```

```
#histograms
ggplot(gss_subset, aes(x = partyid)) +
  geom_histogram(stat = "count")
```

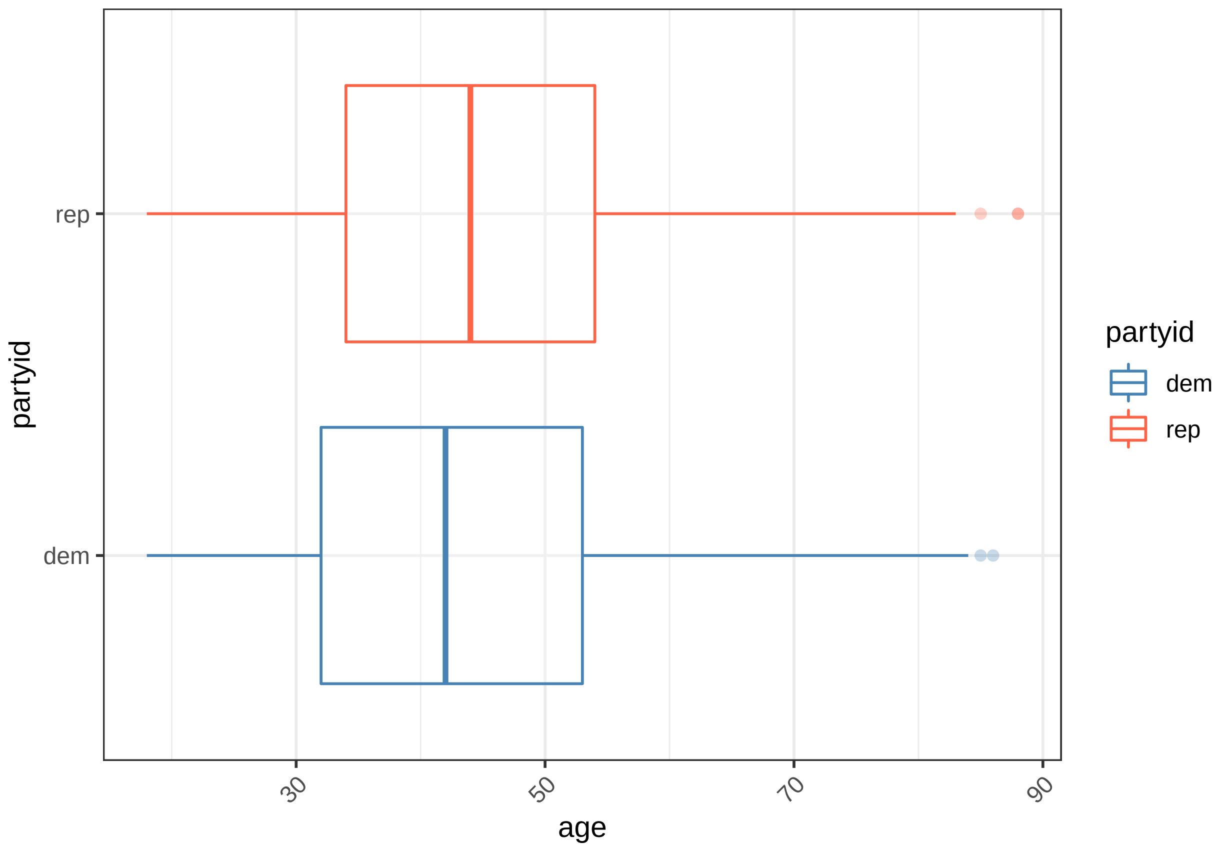
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



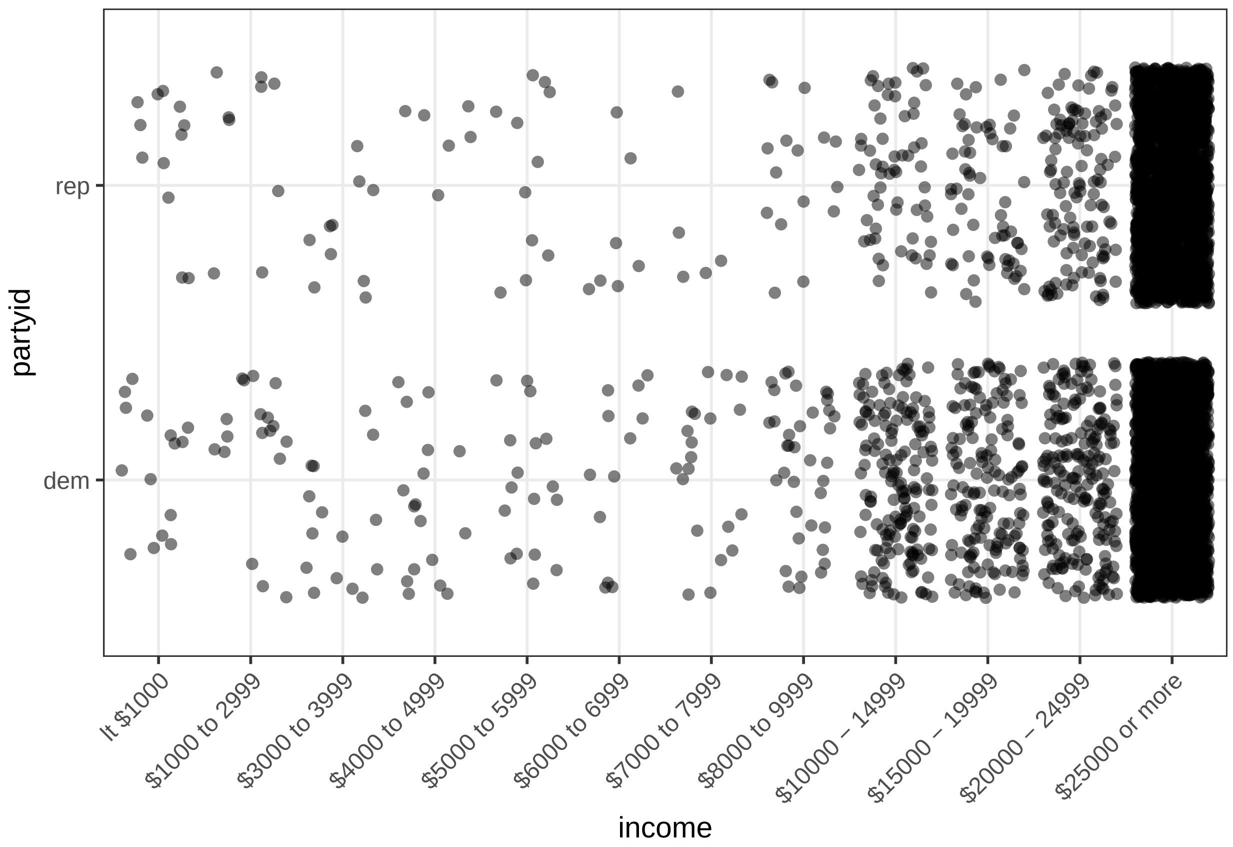
```
#boxplot
ggplot(gss_subset, aes(x = partyid)) +
  geom_boxplot()
```



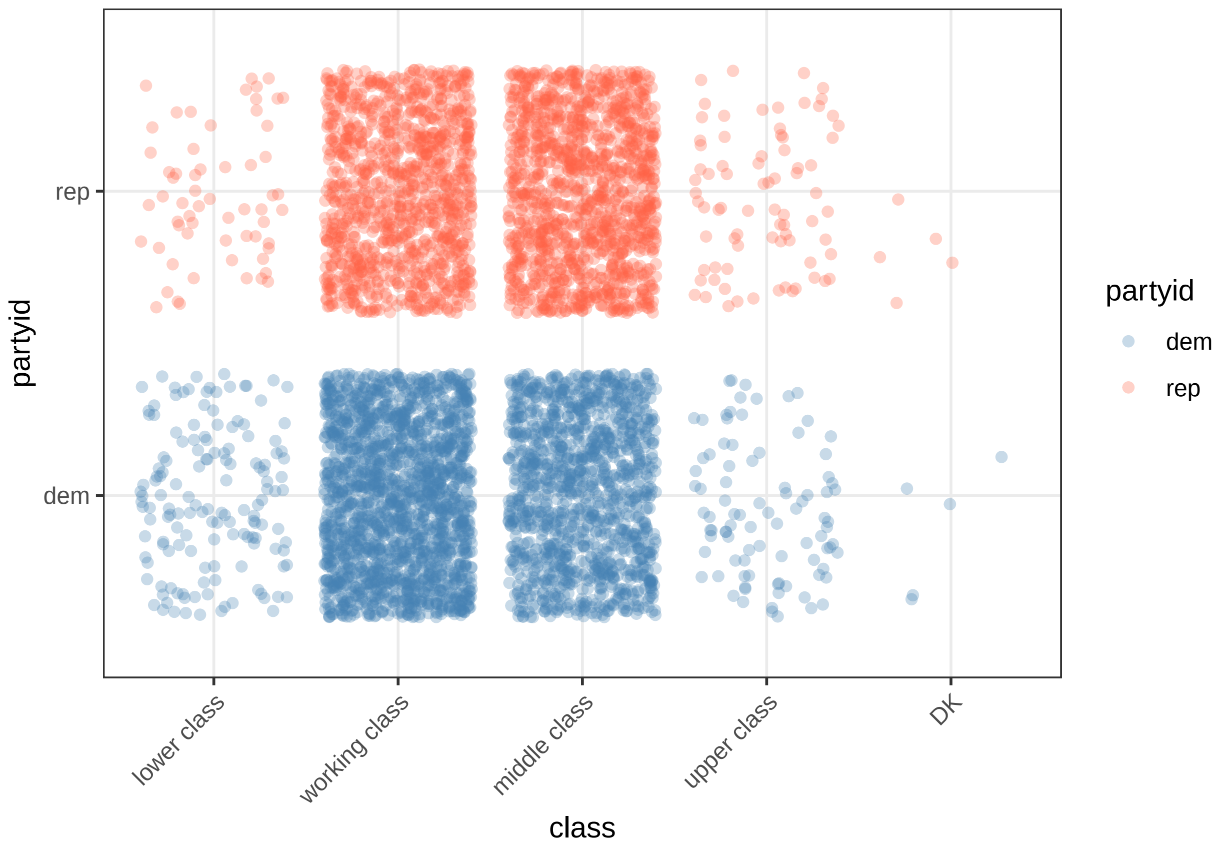
```
#pairwise scatterplots
# library(GGally)
# g2<-ggpairs(gss_subset %>% dplyr::select(-educ),
# lower = list(continuous = wrap("points", alpha = 0.3, size=0.1)),
# upper = list(combo = wrap("box_no_facet", alpha=0.25, outlier.size = .25),
# continuous = wrap("cor", size=2)))
# g2

#plotting randomly selected predictors against party affiliation
ggplot(gss_subset, aes(x = age,
                       y = partyid,
                       color = partyid)) +
  geom_boxplot(alpha = 0.3) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_color_manual(values = c("steelblue", "tomato"))
```



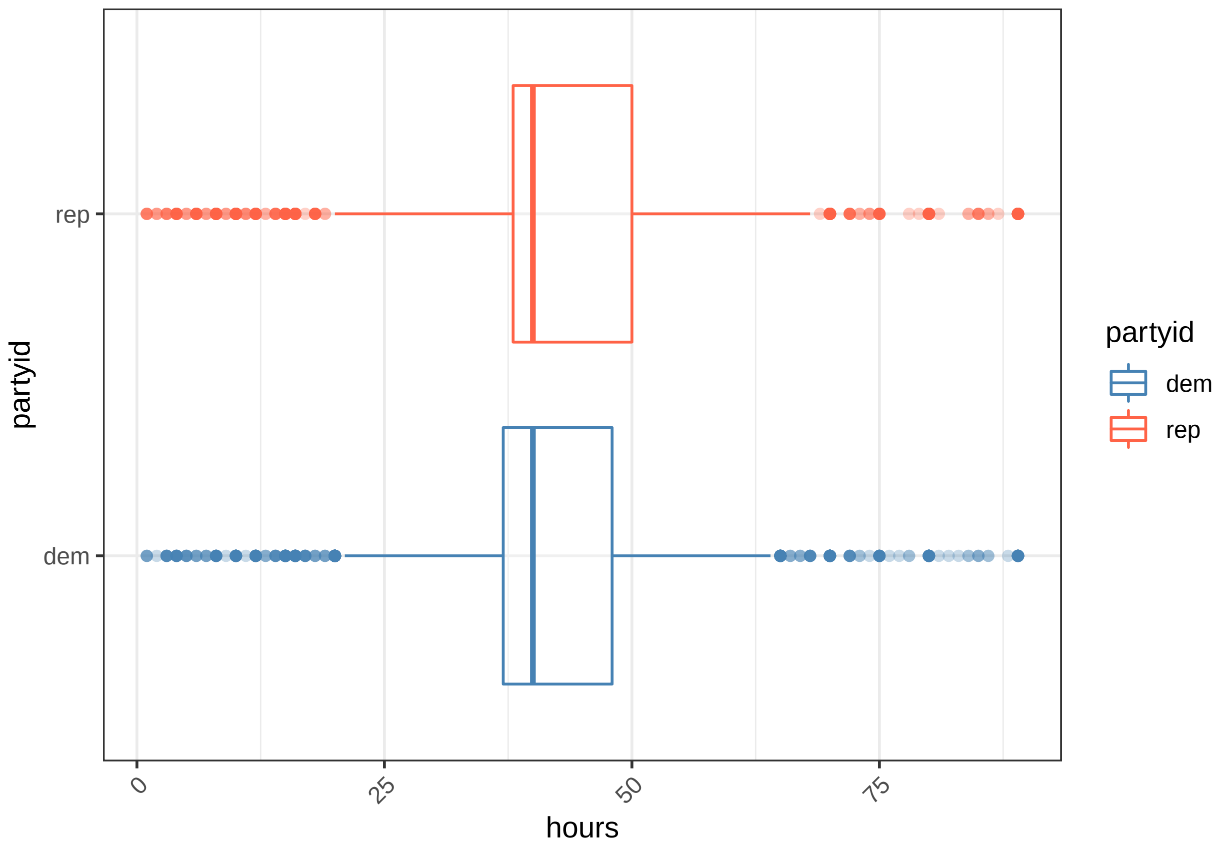
```
ggplot(gss_subset, aes(x = income,
                       y = partyid))+
  geom_jitter(alpha = 0.5)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
ggplot(gss_subset, aes(x = class,
                       y = partyid,
                       color = partyid)) +
  geom_jitter(alpha = 0.3) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_color_manual(values = c("steelblue", "tomato"))
```



```
ggplot(gss_subset, aes(x = hours,
                       y = partyid,
                       color = partyid)) +
  geom_boxplot(alpha = 0.3)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_color_manual(values = c("steelblue", "tomato"))
```



Bray, Andrew, Chester Ismay, Evgeni Chasnovski, Ben Baumer, and Mine Cetinkaya-Rundel. 2020. *Infer: Tidy Statistical Inference*. https://CRAN.R-project.org/package=infer.