Analysis of Variance

Import, clean, explore data

Jenine Harris
Brown School

Exploring the data using graphics and descriptive statistics

- Data saved in an R data file with the file extension .rda can be imported using the `load()` function.

- One limitation of this is that the name of the data object resulting from `load()` is included in the .rda file, so assigning the data to a new object with a new name using `<-` does not work.

```
# load GSS rda file
load(file = "/Users/harrisj/Box/teaching/Teaching/Fall2020/data/gss2018.
```

- The data frame was automatically named `GSS`.

- Rename the data `gss.2018` and remove the `GSS` data for easier use.

```
# assign GSS to gss.2018
gss.2018 <- GSS

# remove GSS
rm(GSS)
```

Data management

- Examine five variables: `USETECH`, `HAPPY`, `SEX`, `AGE`, `DEGREE`.
- Start with `summary(object = gss.2018)` to take a first look at the data frame.

```
# examine the variables
summary(object = gss.2018)
```

```
##       YEAR          BALLOT         USETECH           HAPPY
##  Min.   :2018   Min.   :1.000   Min.   : -1.00   Min.   :1.000
##  1st Qu.:2018   1st Qu.:1.000   1st Qu.: -1.00   1st Qu.:1.000
##  Median :2018   Median :2.000   Median : 10.00   Median :2.000
##  Mean   :2018   Mean   :2.002   Mean   : 48.09   Mean   :1.855
##  3rd Qu.:2018   3rd Qu.:3.000   3rd Qu.: 80.00   3rd Qu.:2.000
##  Max.   :2018   Max.   :3.000   Max.   :999.00   Max.   :8.000
##     PARTYID         RINCOME              RACE             SEX
##  Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.: 0.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median : 9.000   Median :1.000   Median :2.000
##  Mean   :2.968   Mean   : 7.509   Mean   :1.394   Mean   :1.552
##  3rd Qu.:5.000   3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :98.000   Max.   :3.000   Max.   :2.000
##      DEGREE           EDUC            AGE           MARITAL
##  Min.   :0.000   Min.   : 0.00   Min.   :18.00   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:12.00   1st Qu.:34.00   1st Qu.:1.00
##  Median :1.000   Median :14.00   Median :48.00   Median :2.00
##  Mean   :1.684   Mean   :13.84   Mean   :49.13   Mean   :2.67
```

Using the GSS codebook to clean the data

- The `USETECH` variable had a minimum value of -1 and a maximum of 999; open the GSS Data Explorer to look for the question used to get this variable:
    - During a typical week, about what percentage of your total time at work would you normally spend using different types of electronic technologies (such as computers, tablets, smart phones, cash registers, scanners, GPS devices, robotic devices, and so on)?
- The responses should be between zero and 100 percent of the time, recode values outside this range to be NA.
- In the GSS Data Explorer there were three values outside the logical range of zero to 100: -1, 998, and 999.

Recode missing values

- Use the `mutate()` command with `na_if()` from the tidyverse package to recode the values that do not make sense.
- The `na_if()` function recodes specific values of a variable to `NA` for missing.
- In this case, make the `USETECH` variable `NA` if it has the value of -1, 998, or 999. The `na_if()`.

```
# recode USETECH to valid range
library(package = "tidyverse")
gss.2018.cleaned <- gss.2018 %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999))

# check recoding
summary(object = gss.2018.cleaned$USETECH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   15.00   60.00   55.15   90.00  100.00     936
```

Cleaning the remaining variables

- The range was now 0.00 for the minimum and 100.00 for the maximum and there are a lot of `NA` values for the USETECH variable.
- The other variables of interest are: `AGE`, `DEGREE`, `SEX`, and `HAPPY`. L
- Find the variables in the GSS Data Explorer, starting with age:
  - 89 = 89 or older
  - 98 = "Don't know"
  - 99 = "No answer"
- It seems to make the most sense to leave the 89 code for 89 or older and recode the 98 and 99 responses to be `NA`.
- Add on to the existing code and select the five variables of interest to make the data frame size more manageable.

```
# recode USETECH and AGE to valid ranges
gss.2018.cleaned <- gss.2018 %>%
  select(HAPPY, SEX, DEGREE, USETECH, AGE) %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999)) %>%
  mutate(AGE = na_if(x = AGE, y = 98)) %>%
  mutate(AGE = na_if(x = AGE, y = 99))
```

Check the recoding

```r
# check recoding
summary(object = gss.2018.cleaned)
```

```
##      HAPPY            SEX            DEGREE          USETECH
##  Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :  0.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 15.00
##  Median :2.000   Median :2.000   Median :1.000   Median : 60.00
##  Mean   :1.855   Mean   :1.552   Mean   :1.684   Mean   : 55.15
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 90.00
##  Max.   :8.000   Max.   :2.000   Max.   :4.000   Max.   :100.00
##                                                  NA's   :936
##       AGE
##  Min.   :18.00
##  1st Qu.:34.00
##  Median :48.00
##  Mean   :48.98
##  3rd Qu.:63.00
##  Max.   :89.00
##  NA's   :7
```

Complete the NA recoding

- The three other variables, `SEX`, `DEGREE`, and `HAPPY` are categorical variables; the codebook shows some categories that might be better coded as `NA`:

- DEGREE
  - 8 = "Don't know"
  - 9 = "No answer"
- HAPPY
  - 8 = "Don't know"
  - 9 = "No answer"
  - 0 = "Not applicable"

```
# recode variables of interest to valid ranges
gss.2018.cleaned <- gss.2018 %>%
  select(HAPPY, SEX, DEGREE, USETECH, AGE) %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999)) %>%
  mutate(AGE = na_if(x = AGE, y = 98)) %>%
  mutate(AGE = na_if(x = AGE, y = 99)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 8)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 8)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 0))
```

Check the NA recoding

```
# check recoding
summary(object = gss.2018.cleaned)
```

```
##      HAPPY             SEX             DEGREE           USETECH
##  Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :  0.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 15.00
##  Median :2.000   Median :2.000   Median :1.000   Median : 60.00
##  Mean   :1.844   Mean   :1.552   Mean   :1.684   Mean   : 55.15
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 90.00
##  Max.   :3.000   Max.   :2.000   Max.   :4.000   Max.   :100.00
##  NA's   :4                                       NA's   :936
##       AGE
##  Min.   :18.00
##  1st Qu.:34.00
##  Median :48.00
##  Mean   :48.98
##  3rd Qu.:63.00
##  Max.   :89.00
##  NA's   :7
```

Adding labels to categories

- Instead of using `recode_factor()`, try the `factor()` function, which has the `x =` argument for the name of the variable to change to a factor and the `labels =` argument to list the labels for each of the categories in the factor variable.

- Make sure to list the categories in the appropriate order for both the variables.

```
# recode variables of interest to valid ranges
gss.2018.cleaned <- gss.2018 %>%
  select(HAPPY, SEX, DEGREE, USETECH, AGE) %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(AGE = na_if(x = AGE, y = 98)) %>%
  mutate(AGE = na_if(x = AGE, y = 99)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 8)) %>%
  mutate(DEGREE = na_if(x = DEGREE, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 8)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 9)) %>%
  mutate(HAPPY = na_if(x = HAPPY, y = 0)) %>%
  mutate(SEX = factor(x = SEX, labels = c("male","female"))) %>%
  mutate(DEGREE = factor(x = DEGREE, labels = c("< high school",
                                                  "high school", "junior c
                                                  "college", "grad school"
  mutate(HAPPY = factor(x = HAPPY, labels = c("very happy",
                                                "pretty happy",
                                                "not too happy")))
```

Check the labels

```
# check recoding
summary(object = gss.2018.cleaned)
```

```
##              HAPPY           SEX                    DEGREE        USETECH
##  very happy    : 701   male  :1051   < high school : 262   Min.   :  0.00
##  pretty happy :1304   female:1294   high school    :1175   1st Qu.: 15.00
##  not too happy: 336                 junior college: 196   Median : 60.00
##  NA's          :   4                 college        : 465   Mean   : 55.15
##                                      grad school    : 247   3rd Qu.: 90.00
##                                                             Max.   :100.00
##                                                             NA's   :936
##       AGE
##  Min.   :18.00
##  1st Qu.:34.00
##  Median :48.00
##  Mean   :48.98
##  3rd Qu.:63.00
##  Max.   :89.00
##  NA's   :7
```

Exploratory data analysis

- Start by answering the question: do people with higher educational degrees use technology more?
- Start with EDA and check group means and standard deviations.
- To get the mean and standard deviation for each degree category, use `group_by()` with `DEGREE` and then `summarize()` with the mean and standard deviation listed.
- Use `drop_na()` so that `mean()` and `sd()` would work without using `na.rm = TRUE` for each one.

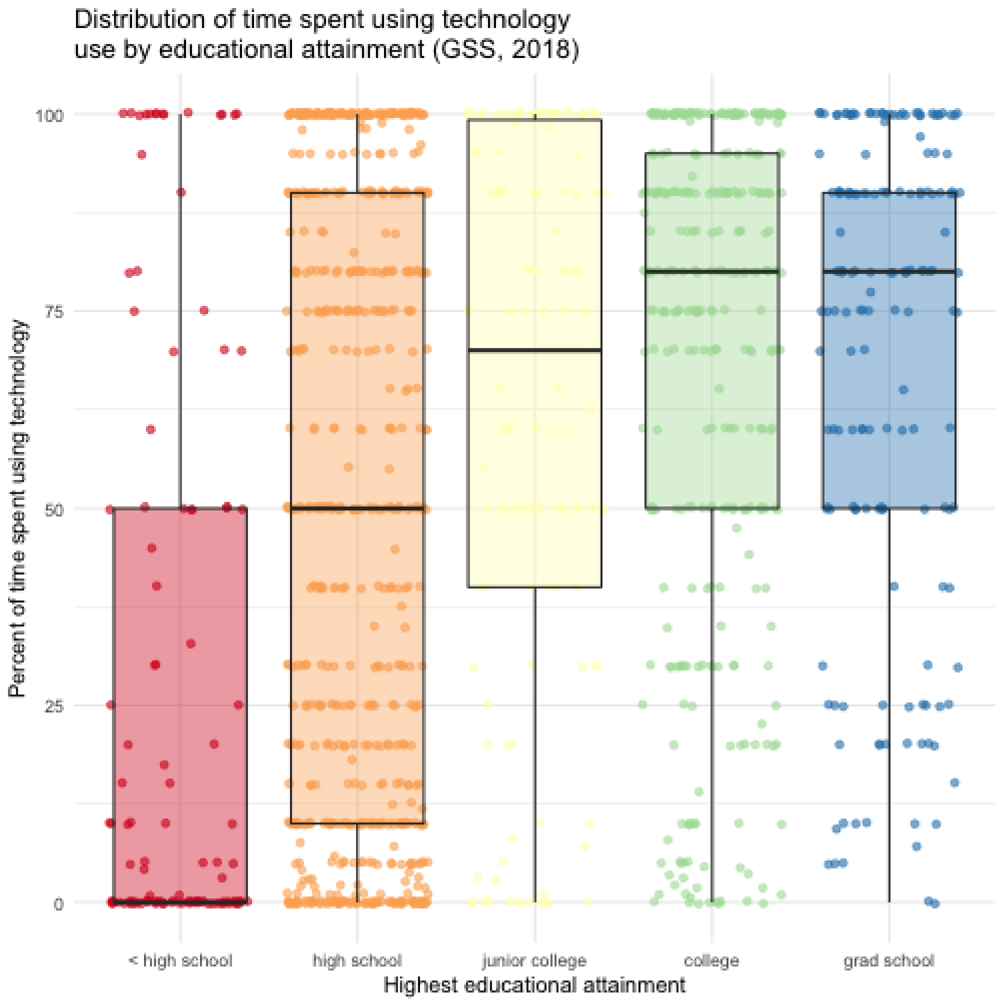
```
# mean and sd of age by group
use.stats <- gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  group_by(DEGREE) %>%
  summarize(m.techuse = mean(USETECH),
            sd.techuse = sd(USETECH))
use.stats
```

```
## # A tibble: 5 x 3
##   DEGREE         m.techuse sd.techuse
##   <fct>              <dbl>      <dbl>
## 1 < high school       24.8       36.2
## 2 high school         49.6       38.6
## 3 junior college      62.4       35.2
## 4 college             67.9       32.1
## 5 grad school         68.7       30.2
```

Visualize technology use by education group

```
# graph usetech
gss.2018.cleaned %>%
  drop_na(USETECH) %>%
  ggplot(aes(y = USETECH, x = DEGREE)) +
  geom_jitter(aes(color = DEGREE), alpha = .6) +
  geom_boxplot(aes(fill = DEGREE), alpha = .4) +
  scale_fill_brewer(palette = "Spectral", guide = FALSE) +
  scale_color_brewer(palette = "Spectral", guide = FALSE) +
  theme_minimal() +
  labs(x = "Highest educational attainment",
       y = "Percent of time spent using technology",
       title = "Distribution of time spent using technology\nuse by educ
```

Interpreting the boxplots



Floor and ceiling effects with ANOVA

- ANOVA can still be used when there are floor and ceiling effects, but with some caution.
- When there are floor or ceiling effects, this means that the variation in a measure is limited by its range.
- Since ANOVA is an analysis of *variance* which examines central tendency and variation together, the limitations of floor and ceiling effects can result in not finding differences when there are differences.
- One common reason for ceiling and floor effects is when the underlying measure has a wider range than what is measured.
    - In the case of technology use, the range of zero to 100 percent of the time is as wide as it can be, so the observations at the ceiling and floor of this measure are just reflecting very low and very high levels of technology use among many of the people in the sample.