Plots

Sara Huston

Data Manipulation

Read in files based off of path, type, and headers with specific NA values accounted for

```
w1_child <- read.csv("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Data/
                     header = TRUE, na.strings = c("9", "8", "98", "99"))
w2_child <- read.table("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Dat
                       header = TRUE, , na.strings = c("-999"))
w3_child <- read.csv("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Data/
                     header = TRUE, na.strings = c("9", "8", "98", "99"))
w4_child <- read.csv("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Data/
                     header = FALSE, na.strings = c("9", "8", "98", "99"))
names_w4_child <- read.table("~/Computational Statistics/Computational Statistics/Assignments/Assignment
educinc <- read.csv("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Data/e
                    header = TRUE, na.strings = c("."))

# Add transposed variable names to w4_child based off of the provided text file
names(w4_child) <- t(names_w4_child)

# Make all headers into lower case for easier merging
names(w1_child) <- tolower(names(w1_child))
names(w2_child) <- tolower(names(w2_child))
names(w3_child) <- tolower(names(w3_child))
names(w4_child) <- tolower(names(w4_child))
names(educinc) <- tolower(names(educinc))

# install.packages("dplyr")
# install.packages("tidyverse")
library("tidyverse")
```

```
## Warning: package 'tidyr' was built under R version 4.3.1

## Warning: package 'readr' was built under R version 4.3.1

## Warning: package 'dplyr' was built under R version 4.3.1

## Warning: package 'stringr' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.1
```

```
## -- 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   1.5.1
## v ggplot2   3.4.2     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
```

```r
library("dplyr")

# Merge files by famid and select specified variables
w1234 <- (list(w1_child, w2_child, w3_child, w4_child, educinc)
          %>% reduce(full_join, by = "famid")) %>% dplyr::select(famid, c01cohort, c01gender, c01school
                                                                 contains("atts"), contains("pcmp"), co
                                                                 contains("dscr"), contains("atod"), fa

# View variables to double check things
names(w1234)
```

```
##   [1] "famid"     "c01cohort" "c01gender" "c01school" "c01sibli"  "c01atts01"
##   [7] "c01atts03" "c01atts04" "c01atts05" "c01atts06" "c01atts07" "c01atts08"
##  [13] "c01atts09" "c01atts10" "c01atts11" "c01atts12" "c01atts13" "c01atts14"
##  [19] "c01atts15" "c01atts16" "c02atts01" "c02atts03" "c02atts04" "c02atts05"
##  [25] "c02atts06" "c02atts07" "c02atts08" "c02atts09" "c02atts10" "c02atts11"
##  [31] "c02atts12" "c02atts13" "c02atts14" "c02atts15" "c02atts16" "c03atts01"
##  [37] "c03atts03" "c03atts04" "c03atts05" "c03atts06" "c03atts07" "c03atts08"
##  [43] "c03atts09" "c03atts10" "c03atts11" "c03atts12" "c03atts13" "c03atts14"
##  [49] "c03atts15" "c03atts16" "c04atts01" "c04atts03" "c04atts04" "c04atts05"
##  [55] "c04atts06" "c04atts07" "c04atts08" "c04atts09" "c04atts10" "c04atts11"
##  [61] "c04atts12" "c04atts13" "c04atts14" "c04atts15" "c04atts16" "c01pcmp01"
##  [67] "c01pcmp02" "c01pcmp03" "c01pcmp04" "c01pcmp05" "c01pcmp06" "c01pcmp07"
##  [73] "c01pcmp08" "c01pcmp09" "c02pcmp01" "c02pcmp02" "c02pcmp03" "c02pcmp04"
##  [79] "c02pcmp05" "c02pcmp06" "c02pcmp07" "c02pcmp08" "c02pcmp09" "c03pcmp01"
##  [85] "c03pcmp02" "c03pcmp03" "c03pcmp04" "c03pcmp05" "c03pcmp06" "c03pcmp07"
##  [91] "c03pcmp08" "c03pcmp09" "c04pcmp01" "c04pcmp02" "c04pcmp03" "c04pcmp04"
##  [97] "c04pcmp05" "c04pcmp06" "c04pcmp07" "c04pcmp08" "c04pcmp09" "c01attt01"
## [103] "c01attt02" "c01attt03" "c01attt04" "c01attt05" "c01attt06" "c01attt07"
## [109] "c01attt08" "c01attt09" "c02attt01" "c02attt02" "c02attt03" "c02attt04"
## [115] "c02attt05" "c02attt06" "c02attt07" "c02attt08" "c02attt09" "c03attt01"
## [121] "c03attt02" "c03attt03" "c03attt04" "c03attt05" "c03attt06" "c03attt07"
## [127] "c03attt08" "c03attt09" "c04attt01" "c04attt02" "c04attt03" "c04attt04"
## [133] "c04attt05" "c04attt06" "c04attt07" "c04attt08" "c04attt09" "c01dscr01"
## [139] "c01dscr02" "c01dscr03" "c01dscr04" "c01dscr05" "c01dscr06" "c01dscr07"
## [145] "c01dscr08" "c01dscr09" "c01dscr10" "c01dscr11" "c01dscr12" "c01dscr13"
## [151] "c01dscr14" "c01dscr15" "c01dscr16" "c01dscr17" "c01dscr18" "c01dscr19"
## [157] "c02dscr01" "c02dscr02" "c02dscr03" "c02dscr04" "c02dscr05" "c02dscr06"
## [163] "c02dscr07" "c02dscr08" "c02dscr09" "c02dscr10" "c02dscr19" "c02dscr11"
## [169] "c02dscr12" "c02dscr13" "c02dscr14" "c02dscr15" "c02dscr16" "c02dscr17"
## [175] "c02dscr18" "c03dscr01" "c03dscr02" "c03dscr03" "c03dscr04" "c03dscr05"
## [181] "c03dscr06" "c03dscr07" "c03dscr08" "c03dscr09" "c03dscr10" "c03dscr11"
## [187] "c03dscr12" "c03dscr13" "c03dscr14" "c03dscr15" "c03dscr16" "c03dscr17"
```

```
## [193] "c03dscr18" "c03dscr19" "c04dscr01" "c04dscr02" "c04dscr03" "c04dscr04"
## [199] "c04dscr05" "c04dscr06" "c04dscr07" "c04dscr08" "c04dscr09" "c04dscr10"
## [205] "c04dscr11" "c04dscr12" "c04dscr13" "c04dscr14" "c04dscr15" "c04dscr16"
## [211] "c04dscr17" "c04dscr18" "c04dscr19" "c01atod01" "c01atod02" "c01atod03"
## [217] "c01atod04" "c01atod05" "c01atod06" "c01atod07" "c01atod08" "c01atod09"
## [223] "c01atod10" "c01atod11" "c01atod12" "c01atod13" "c01atod14" "c01atod15"
## [229] "c01atod16" "c01atod17" "c01atod18" "c01atod19" "c01atod20" "c01atod21"
## [235] "c01atod22" "c01atod23" "c01atod24" "c01atod25" "c01atod26" "c02atod01"
## [241] "c02atod02" "c02atod03" "c02atod04" "c02atod05" "c02atod06" "c02atod07"
## [247] "c02atod08" "c02atod09" "c02atod10" "c02atod11" "c02atod12" "c02atod13"
## [253] "c02atod14" "c02atod15" "c02atod16" "c02atod17" "c02atod18" "c02atod19"
## [259] "c02atod20" "c02atod21" "c02atod22" "c02atod23" "c02atod24" "c02atod25"
## [265] "c02atod26" "c03atod01" "c03atod02" "c03atod03" "c03atod04" "c03atod05"
## [271] "c03atod06" "c03atod07" "c03atod08" "c03atod09" "c03atod10" "c03atod11"
## [277] "c03atod12" "c03atod13" "c03atod14" "c03atod15" "c03atod16" "c03atod17"
## [283] "c03atod18" "c03atod19" "c03atod20" "c03atod21" "c03atod22" "c03atod23"
## [289] "c03atod24" "c03atod25" "c03atod26" "c04atod01" "c04atod02" "c04atod03"
## [295] "c04atod04" "c04atod05" "c04atod06" "c04atod07" "c04atod08" "c04atod09"
## [301] "c04atod10" "c04atod11" "c04atod12" "c04atod13" "c04atod14" "c04atod15"
## [307] "c04atod16" "c04atod17" "c04atod18" "c04atod19" "c04atod20" "c04atod21"
## [313] "c04atod22" "c04atod23" "c04atod24" "c04atod25" "c04atod26" "fameduc"
## [319] "income"    "c01edex01" "c01edex02" "c02edex01" "c02edex02" "c03edex01"
## [325] "c03edex02" "c04edex01" "c04edex02"
```

```r
# Reverse code for pcmp 1 and 2
pcmp_01_02_cols <- c(grep("pcmp01", names(w1234)), grep("pcmp02", names(w1234)))
w1234[ ,pcmp_01_02_cols] <- 5 - w1234[ ,pcmp_01_02_cols]

# Compute averages based on sets of columns for variable sets and place average into a new variable
w1234$c01attt <- rowMeans(w1234[c(grep("c01attt", names(w1234)))], na.rm = TRUE)
w1234$c04attt <- rowMeans(w1234[c(grep("c04attt", names(w1234)))], na.rm = TRUE)

w1234$c01pcmp <- rowMeans(w1234[c(grep("c01pcmp", names(w1234)))], na.rm = TRUE)
w1234$c04pcmp <- rowMeans(w1234[c(grep("c04pcmp", names(w1234)))], na.rm = TRUE)

w1234$c01dscr <- rowMeans(w1234[c("c01dscr07", "c01dscr08", "c01dscr09", "c01dscr10")], na.rm = TRUE)
w1234$c04dscr <- rowMeans(w1234[c("c04dscr07", "c04dscr08", "c04dscr09", "c04dscr10")], na.rm = TRUE)

w1234$c01atts <- rowMeans(w1234[c("c01atts03", "c01atts07", "c01atts08", "c01atts10")], na.rm = TRUE)
w1234$c02atts <- rowMeans(w1234[c("c02atts03", "c02atts07", "c02atts08", "c02atts10")], na.rm = TRUE)
w1234$c02atts <- rowMeans(w1234[c("c03atts03", "c03atts07", "c03atts08", "c03atts10")], na.rm = TRUE)
w1234$c04atts <- rowMeans(w1234[c("c01atts03", "c04atts07", "c04atts08", "c04atts10")], na.rm = TRUE)

w1234$c01atod <- rowSums(w1234[c("c01atod01", "c01atod02", "c01atod03", "c01atod04", "c01atod05", "c01a
w1234$c04atod <- rowSums(w1234[c("c04atod01", "c04atod02", "c04atod03", "c04atod04", "c04atod05", "c04a

## Final Quiz Questions

# 1) Dimension of your final data frame (row x column)
dim(w1234)
```

```
## [1] 674 338
```

```r
# 2) What is the max average education level obtained by parents ("fameduc")?
max_educ <- max(w1234$fameduc, na.rm = TRUE)
max_educ
```

```
## [1] 19
```

```r
# 3) How many children have parents with this level of education?
sum(w1234$fameduc >= max_educ, na.rm = TRUE)
```

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

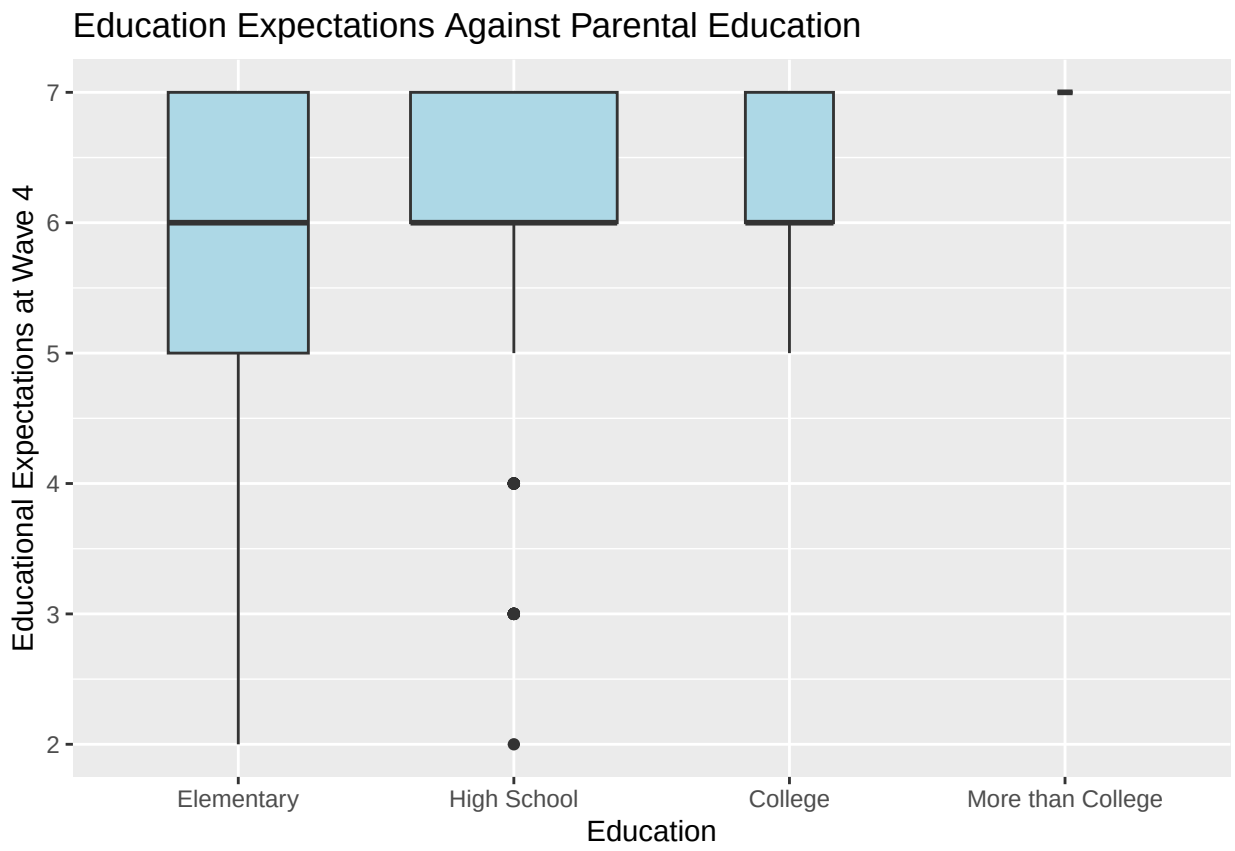
```r
# 4) What is the mean for variable "c01attt" for those who have 4 siblings?
four_sibs <- w1234[w1234$c01sibli == 4 & !is.na(w1234$c01sibli), ]
mean(four_sibs$c01attt)
```

```
## [1] 2.907292
```

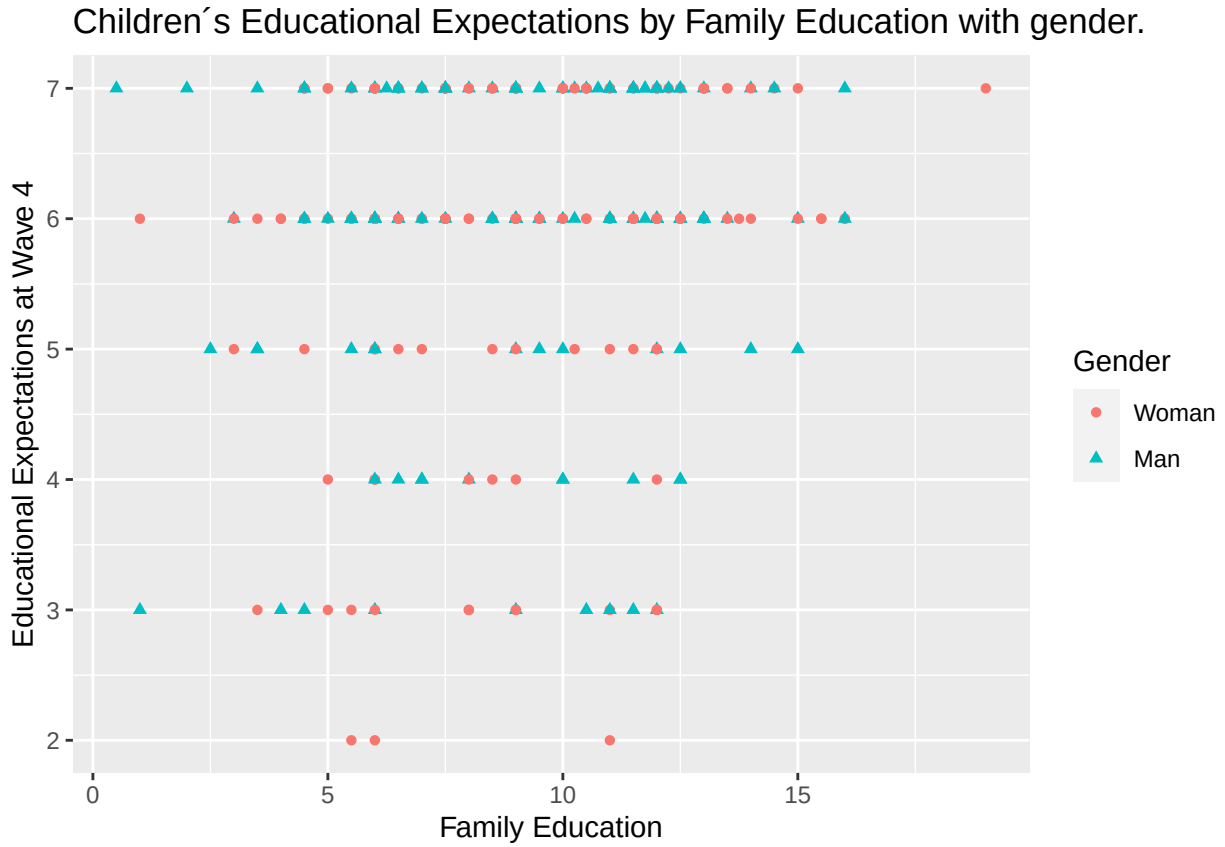
```r
# 5) Create difference scores between Waves 1 and 4 for all scales. Which scale has
# the greatest average absolute change (i.e., difference)?
score_variables <- c("atts", "pcmp", "attt", "dscr", "atod")
w1234[paste("difference_", score_variables, sep = "")] <- abs(w1234[paste("c01", score_variables, sep =

# "atod" has the greatest average absolute change
ave_abs_change <- colMeans(w1234[ , c("difference_atts", "difference_pcmp", "difference_attt", "differe
ave_abs_change[which.max(ave_abs_change)]
```

```
## difference_atod
##        9.980712
```

```r
w1234$gender_r[w1234$c01gender == 1] <- "male"
w1234$gender_r[w1234$c01gender == 2] <- "female"

# Create a new variable called "newedu" w/ various levels
w1234$new_edu[w1234$fameduc < 7 & !is.na(w1234$fameduc)] <- "Elementary"
w1234$new_edu[w1234$fameduc >= 7 & w1234$fameduc < 13 & !is.na(w1234$fameduc)] <- "High School"
w1234$new_edu[w1234$fameduc >= 13 & w1234$fameduc < 17 & !is.na(w1234$fameduc)] <- "College"
w1234$new_edu[w1234$fameduc >= 17 & !is.na(w1234$fameduc)] <- "More than College"

# Order the levels and create the factor
w1234$new_edu <- factor(c(w1234$new_edu), levels = c("Elementary", "High School", "College", "More than

# Remove na from each plot
w1234_clean <- subset(w1234, !is.na(new_edu) & !is.na(c04edex01) & !is.na(fameduc))

# Create a box plot with education levels and educational experiences at wave 4
g <- ggplot(data = w1234_clean, mapping = aes(new_edu, c04edex01))
g + geom_boxplot(data = w1234_clean, mapping = aes(new_edu, c04edex01), varwidth=T, fill = 'lightblue')
  labs(title = "Education Expectations Against Parental Education", x = "Education", y ="Educational Exp
```




```
# Create a scatter plot with educational expectations at wave 4 (c04edex01) on the Y-axis with "fameduc
ggplot(data = w1234_clean, mapping = aes(x = fameduc, y = c04edex01, color = gender_r, shape = gender_r
  labs(title = "Children's Educational Expectations by Family Education with gender.",
       x = "Family Education", y = "Educational Expectations at Wave 4") +
  scale_shape_discrete(name ="Gender", breaks = c("female", "male"), labels = c("Woman", "Man")) +
  scale_colour_discrete(name ="Gender", breaks = c("female", "male"), labels = c("Woman", "Man"))
```

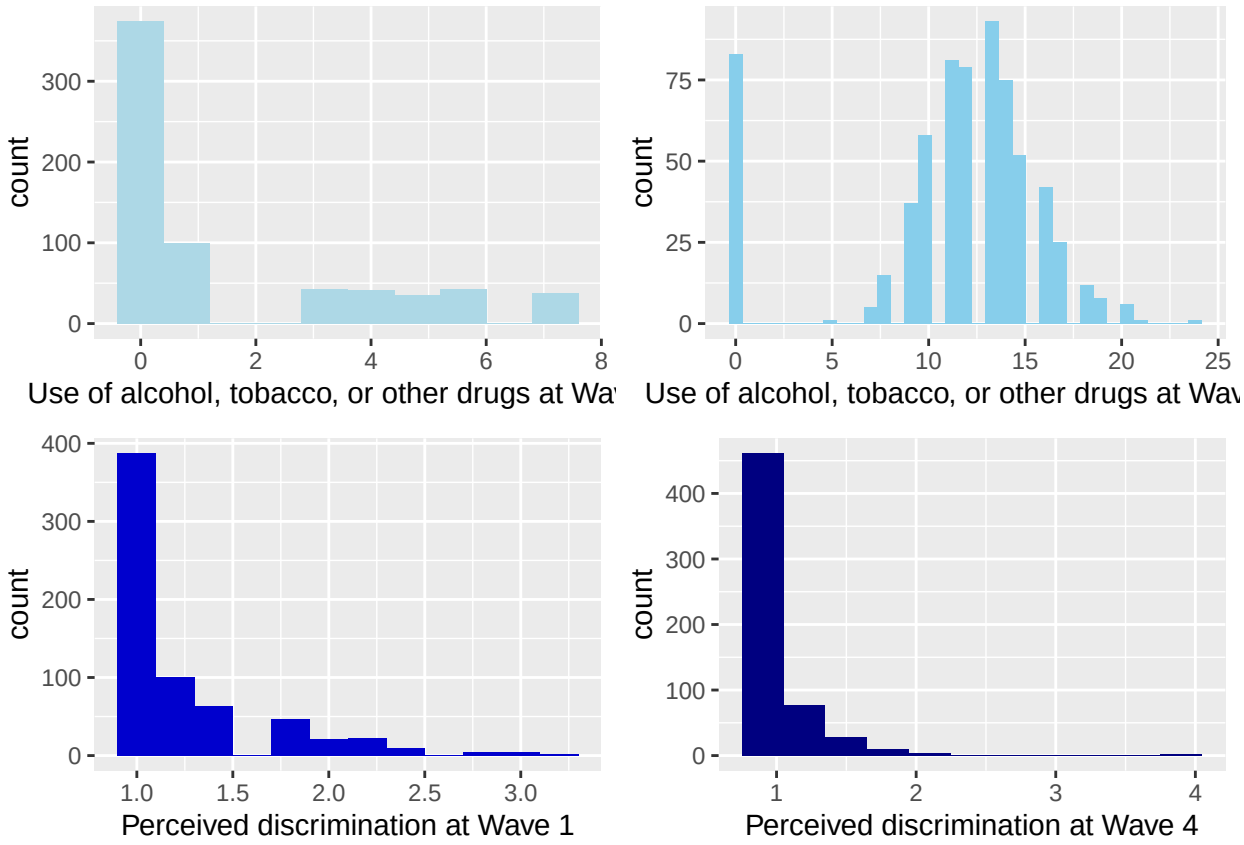



```
# Plot histograms of perceived discrimination and use of alcohol, tobacco, or other drugs at waves 1 an
t1 <- ggplot(data = subset(w1234, !is.na(c01atod)), mapping = aes(c01atod))
t2 <- ggplot(data = subset(w1234, !is.na(c04atod)), mapping = aes(c04atod))
t3 <- ggplot(data = subset(w1234, !is.na(c01dscr)), mapping = aes(c01dscr))
t4 <- ggplot(data = subset(w1234, !is.na(c04dscr)), mapping = aes(c04dscr))

g1 <- t1 + geom_histogram(fill = "lightblue", binwidth = 0.8)+ labs(x = "Use of alcohol, tobacco, or otl
g2 <- t2 + geom_histogram(fill = "skyblue", binwidth = 0.7)+ labs(x = "Use of alcohol, tobacco, or other
g3 <- t3 + geom_histogram(fill = "mediumblue", binwidth = 0.2)+ labs(x = "Perceived discrimination at Wa
g4 <- t4 + geom_histogram(fill = "navyblue", binwidth = 0.3)+ labs(x = "Perceived discrimination at Wave

# install.packages("gridExtra")
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.3
```

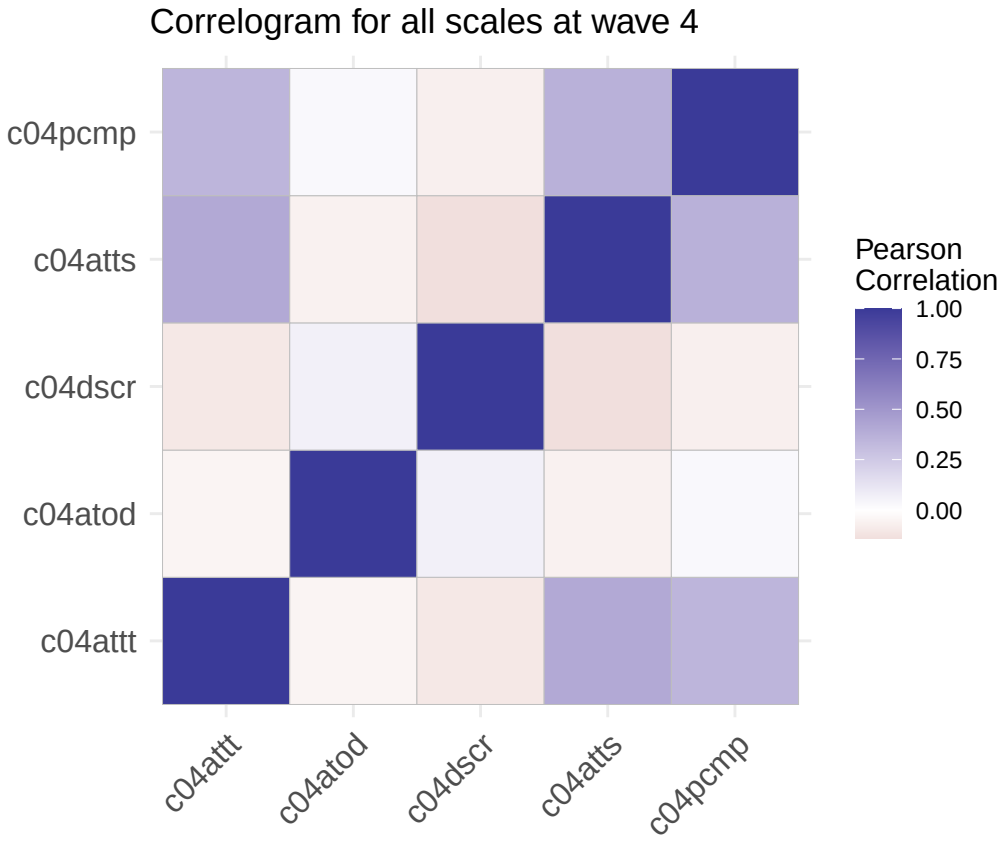
```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(g1, g2, g3, g4, ncol = 2, nrow = 2)
```




```
# Describe the distributions in two sentences
# Perceived discrimination decreases from wave 1 to wave 4, with an increase of ~100 in 1 scores.
# Alcohol, tobacco, or other drug use increased heavily from wave 1 to wave 4.

# Create a correlogram for all 5 scales at Wave 4
cor_scales <- cor(w1234[, c("c04attt", "c04atod", "c04dscr", "c04atts", "c04pcmp")], use = "complete.obs
cor_scales <- round(cor_scales, 2)
# install.packages("ggcorrplot")
library(ggcorrplot)
ggcorrplot(cor_scales) + labs(title = "Correlogram for all scales at wave 4") +
  scale_fill_gradient2(name="Pearson\nCorrelation")
```

```
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```


Correlogram for all scales at wave 4
c04pcmp
c04atts
c04dscr
c04atod
c04attt
c04attt
c04atod
c04dscr
c04atts
c04pcmp
Pearson Correlation
1.00
0.75
0.50
0.25
0.00