

Plots

Sara Huston

Assignment 1

Data Manipulation

Install and load all required packages (Don't install in rmd)

```
# install.packages("dplyr")
# install.packages("tidyverse")
# install.packages("gridExtra")
# install.packages("ggcorrplot")
# install.packages("formatR")
```

```
library("ggcorrplot")
```

```
## Loading required package: ggplot2
```

```
library("gridExtra")
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v stringr    1.5.1
## v forcats    1.0.0      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::combine() masks gridExtra::combine()
## 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
```

```
library("dplyr")
library("formatR")
```

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/Da
                    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/Assignmen
educinc <- read.csv("~/Computational Statistics/Computational Statistics/Assignments/Assignment 1/Data/
                    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))
```

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, c01sibli, contains("atts"), contains("pcmp"),
  contains("attd"), contains("dscr"), contains("atod"), fameduc, income,
  c01sibli, contains("edex"))
```

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$c01attd <- rowMeans(w1234[c(grep("c01attd", names(w1234)))], na.rm = TRUE)
w1234$c04attd <- rowMeans(w1234[c(grep("c04attd", 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$c03atts <- rowMeans(w1234[c("c03atts03", "c03atts07", "c03atts08", "c03atts10")], na.rm = TRUE)
w1234$c04atts <- rowMeans(w1234[c("c04atts03", "c04atts07", "c04atts08", "c04atts10")], na.rm = TRUE)

w1234$c01atod <- rowSums(w1234[c("c01atod01", "c01atod02", "c01atod03", "c01atod04", "c01atod05", "c01atod06"), na.rm = TRUE])
w1234$c04atod <- rowSums(w1234[c("c04atod01", "c04atod02", "c04atod03", "c04atod04", "c04atod05", "c04atod06"), na.rm = TRUE])

```

Quiz Questions

1) Dimension of your final data frame (row x column)

```
dim(w1234)
```

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

2) What is the max average education level obtained by parents (“fameduc”)?

```

max_educ <- max(w1234$fameduc, na.rm = TRUE)
max_educ

```

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

3) How many children have parents with this level of education?

```
sum(w1234$fameduc >= max_educ, na.rm = TRUE)
```

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

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] 2.9073
```

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 = "")])
ave_abs_change <- colMeans(w1234[, c("difference_atts", "difference_pcmp", "difference_attt", "difference_dscr", "difference_atod")])
ave_abs_change[which.max(ave_abs_change)]

## difference_atod
##          9.9807
```

Assignment 2

Plots

Plot 1

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

Recode gender variables

```
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"
```

Create a new variable called “newedu” w/ various levels of education

```
w1234$new_edu <- factor(c(w1234$new_edu), levels = c("Elementary", "High School", "College", "More than College"))
```

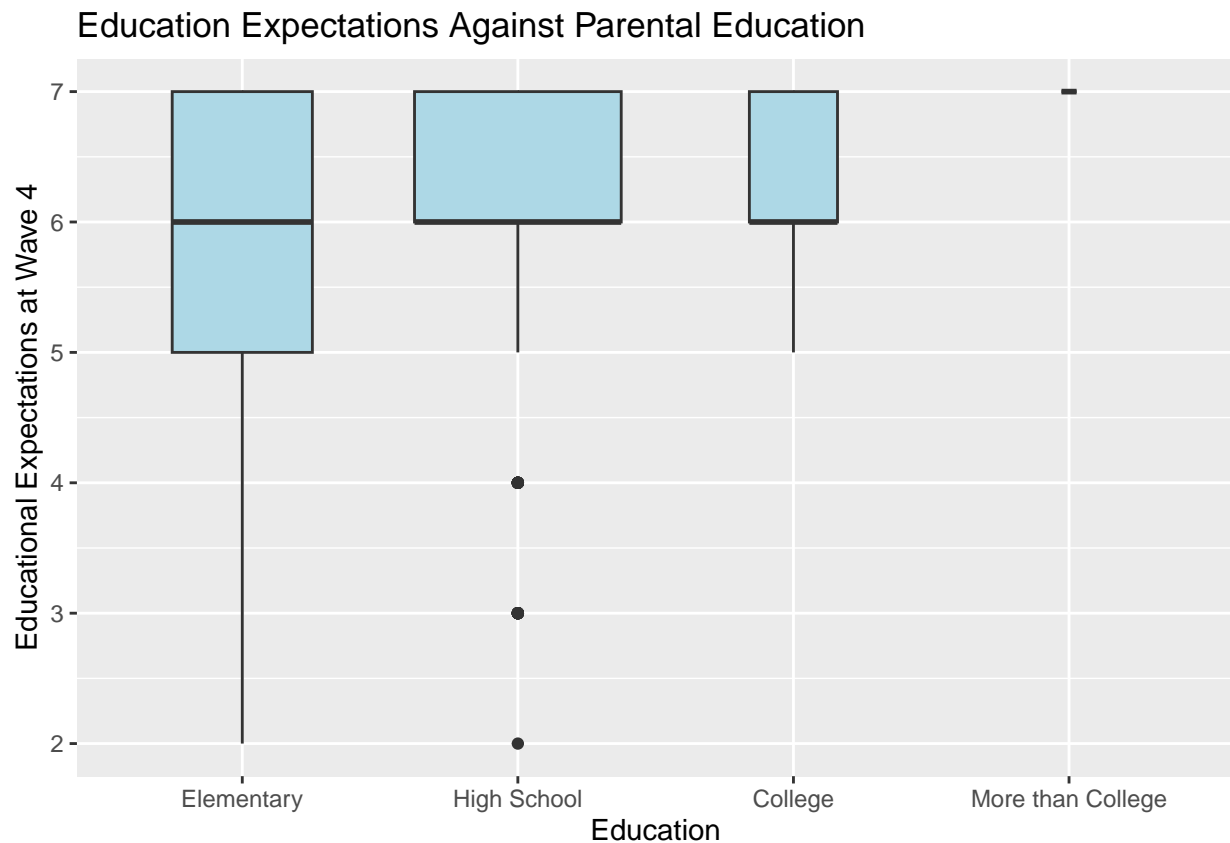
Order the education levels and factor

```
w1234_clean <- subset(w1234, !is.na(new_edu) & !is.na(c04edex01) & !is.na(fameduc))
```

Remove na from each plot

```
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 Expectations at Wave 4")
```

Create a box plot with education levels and educational experiences at wave 4

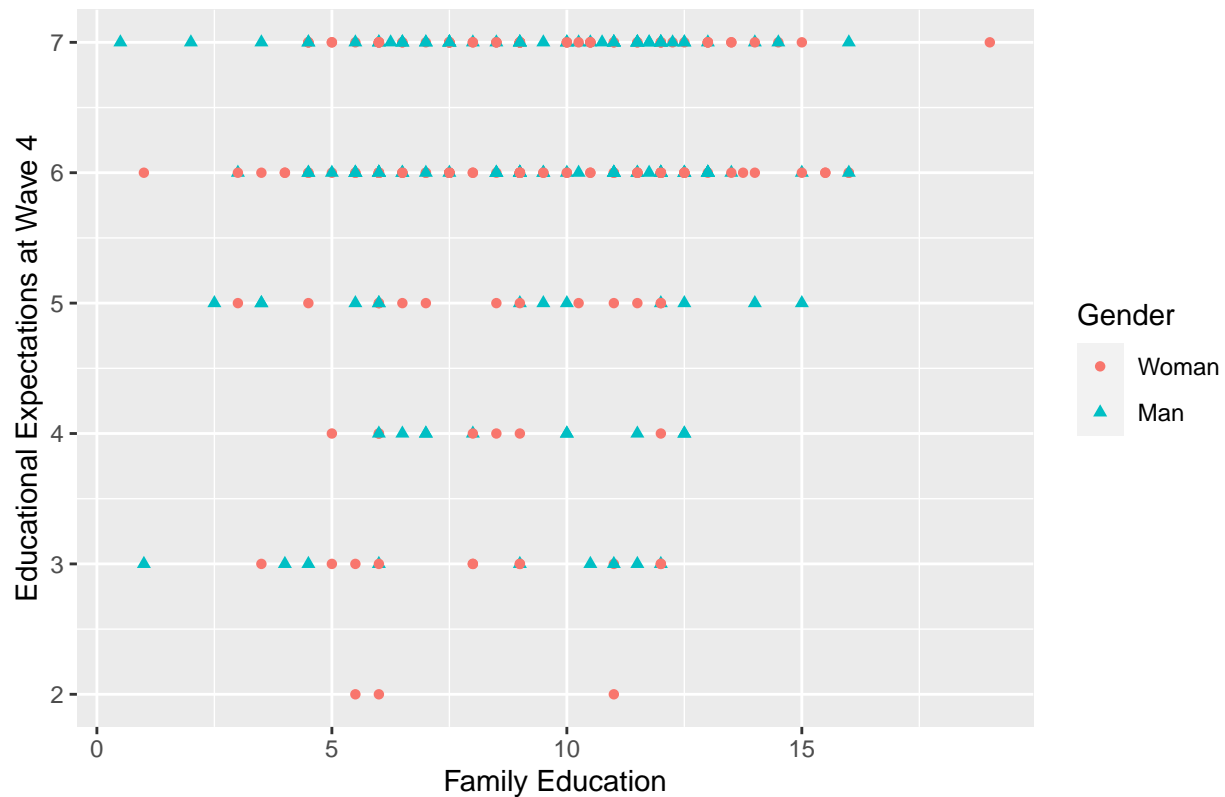


Plot 2

```
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_color_discrete(name = "Gender", breaks = c("female", "male"), labels = c("Woman", "Man"))
```

Create a scatter plot with educational expectations at wave 4 (c04edex01) on the Y-axis with “fameduc” on the X axis

Children's Educational Expectations by Family Education with gender.



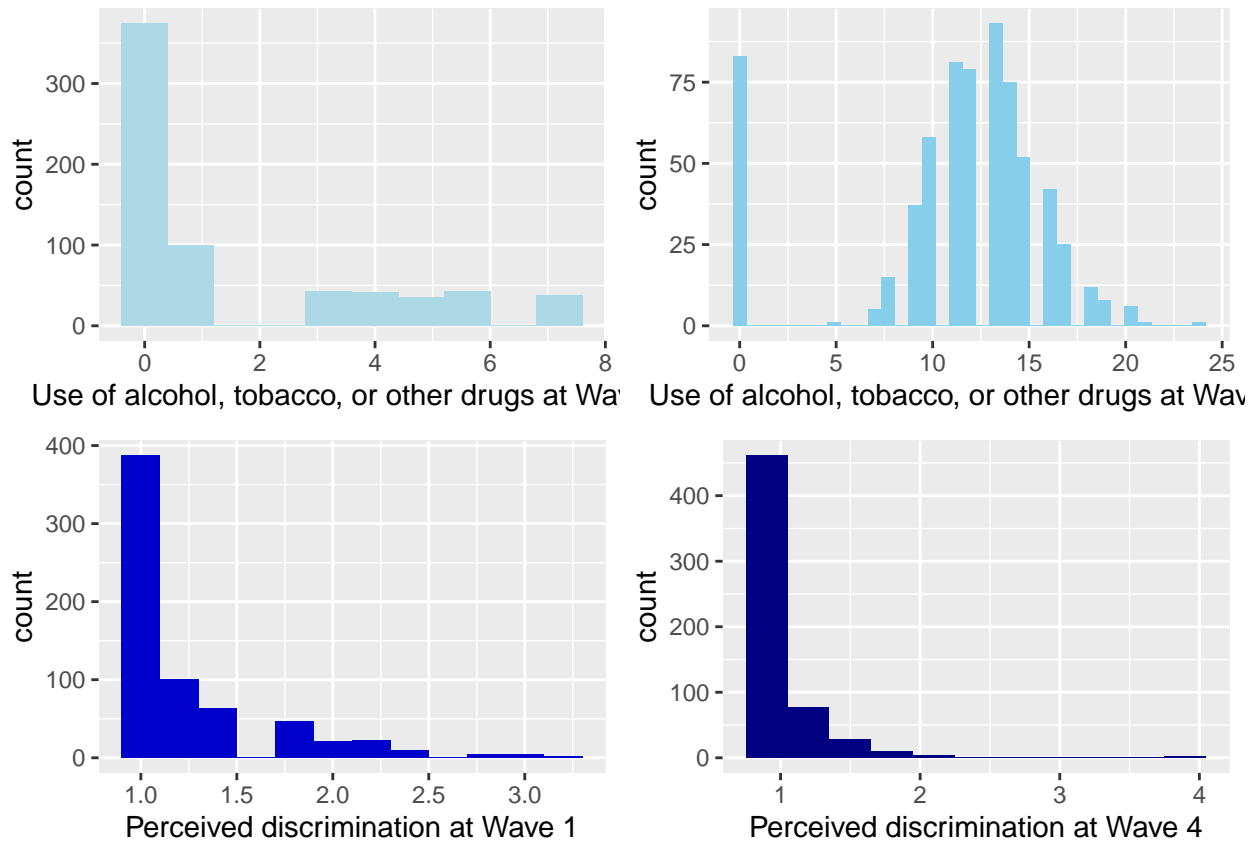
Plot 3

```
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 other drugs at Wave 1")
g2 <- t2 + geom_histogram(fill = "skyblue", binwidth = 0.7) + labs(x = "Use of alcohol, tobacco, or other drugs at Wave 4")
g3 <- t3 + geom_histogram(fill = "mediumblue", binwidth = 0.2) + labs(x = "Perceived discrimination at Wave 1")
g4 <- t4 + geom_histogram(fill = "navyblue", binwidth = 0.3) + labs(x = "Perceived discrimination at Wave 4")

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

Plot histograms of perceived discrimination and use of alcohol, tobacco, or other drugs at waves 1 and 4



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.

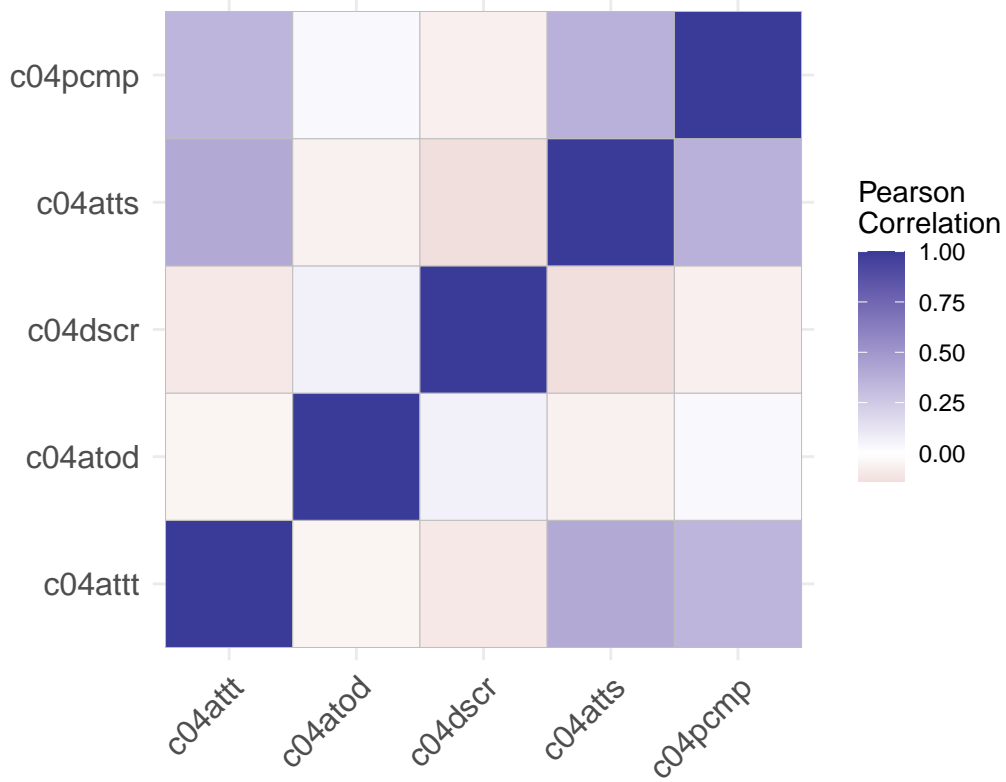
Plot 4

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

Create a correlogram for all 5 scales at Wave 4

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

Correlogram for all scales at wave 4



Quiz Questions

““