EPID 674: Data Management in R

Homework 4 Answer Key

Kelly Bakulski

Last compiled on February 18, 2022

# Load packages

##### Load packages for this assignment, do this every time  
library(tidyverse)  
library(here)  
library(ggpubr)  
library(viridis)  
library(RColorBrewer)

# Load data

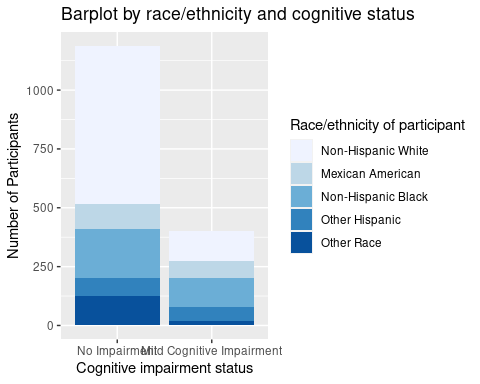
# Load saved NHANES homework dataset from the project files. This is the dataset we built in homework 2  
load(here(("nhanes\_homework\_dataset.rda")), verbose = TRUE)

## Loading objects:  
## nhanes\_homework

# 1. Create a bar plot

### Create a bar plot where cognitive impairment status (variable: MCI) is on the x-axis and the bar fill is by race/ethnicity (variable: race\_eth). Try updating the bar colors.

# Optional, drop the participants with missing cognition values  
nhanes\_cognition <- nhanes\_homework %>%  
 drop\_na(MCI)  
  
# Make the bar plot  
ggplot(nhanes\_cognition,  
 aes(x = MCI,  
 fill = race\_eth)) +  
 geom\_bar() +  
 labs(title = "Barplot by race/ethnicity and cognitive status",  
 x = "Cognitive impairment status",  
 y = "Number of Participants",  
 fill = "Race/ethnicity of participant") +   
 scale\_fill\_brewer()



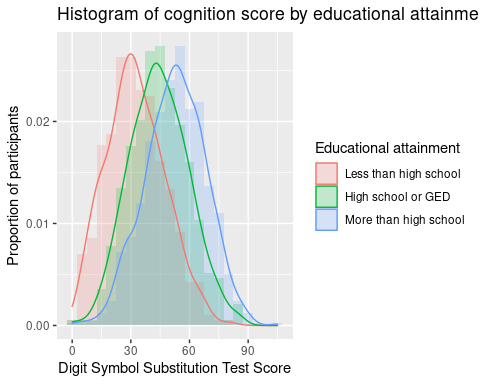
Interpret the figure: There appears to be a higher proportion of Non-Hispanic White participants in the no impairment category relative to the mild cognitive impairment category.

# 2. Create a histogram or density plot

### Make a histogram or density plot with the digit symbol substitution test score (variable: CFDDS) on the x-axis. Have the fill and/or color of the plot by category of educational attainment.

# Optional, drop participants missing education  
nhanes\_education <- nhanes\_homework %>%  
 drop\_na(education)  
  
# Overlapping histograms  
ggplot(nhanes\_education,  
 aes(x = CFDDS,  
 fill = education,  
 y = ..density..)) +   
 geom\_histogram(binwidth = 5,  
 position = "identity", #default position is stack   
 alpha = 0.2) +  
 geom\_density(aes(color = education),  
 alpha = 0) +  
 labs(x = "Digit Symbol Substitution Test Score",  
 y = "Proportion of participants",  
 title = "Histogram of cognition score by educational attainment",  
 fill = "Educational attainment",  
 color = "Educational attainment")

## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.



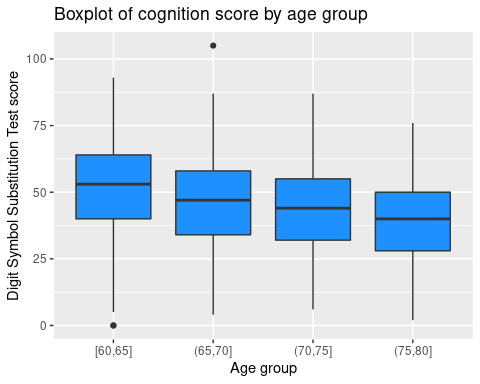
Interpret the figure: Within each educational attainment group, the distribution of digit symbol substitution score is relatively normally distributed.

# 3. Create a boxplot or violin plot

### Make a boxplot or violin plot with the digit symbol substitution test score (variable: CFDDS) on the y-axis. Split the boxes or violins by the age group (variable: age\_group) on the x-axis. Try filling the boxes with the color of your choice

# Boxplot split by age groups  
ggplot(nhanes\_homework,  
 aes(x = age\_groups,  
 y = CFDDS)) +  
 geom\_boxplot(fill = "dodgerblue") +  
 labs(x = "Age group",  
 y= "Digit Symbol Substitution Test score",  
 title = "Boxplot of cognition score by age group")

## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.



Interpret the figure: Yes, there does appear to be a relationship between age group and digit symbol substitution test. Higher age gropus have lower digit symbol substitution test scores.

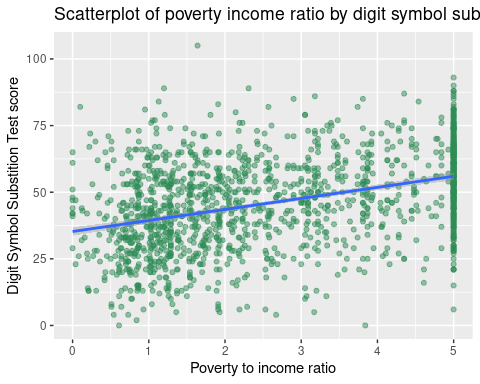
# 4. Create a scatterplot with a line of best fit

### Create a scatter plot with poverty-income ratio (variable: INDFMPIR) on the x-axis and digit symbol substitution test (variable: CFDDS) on the y-axis. Add a line of best fit for this relationship.

# Add a linear fit  
ggplot(nhanes\_homework,  
 aes(x = INDFMPIR,  
 y = CFDDS)) +  
 geom\_point(alpha = 0.5,  
 color = "seagreen") +  
 geom\_smooth(method = lm) +  
 labs(x = "Poverty to income ratio",  
 y = "Digit Symbol Substition Test score",  
 title = "Scatterplot of poverty income ratio by digit symbol substitution test")

## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.

## `geom\_smooth()` using formula 'y ~ x'



Interpret the figure: Yes, there does appear to be a relationship between poverty to income ratio and digit symbol substitution test score. Higher values of poverty to income ratio have higher digit symbol substitution test scores.

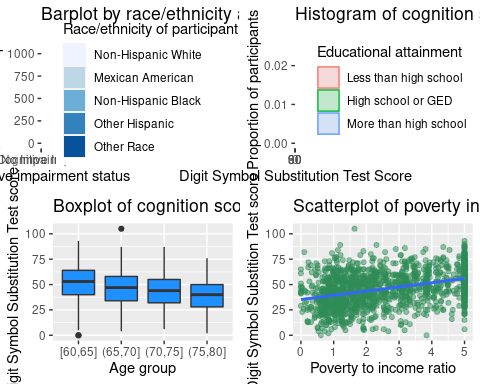
# 5. Create a multipaneled plot

### Create a multipaneled plot containing each of the previous plots from questions 1-4 (a barplot, a histogram, a boxplot or violin plot, and a scatterplot)

bar <- ggplot(nhanes\_cognition,  
 aes(x = MCI,  
 fill = race\_eth)) +  
 geom\_bar() +  
 labs(title = "Barplot by race/ethnicity and cognitive status",  
 x = "Cognitive impairment status",  
 y = "Number of Participants",  
 fill = "Race/ethnicity of participant") +   
 scale\_fill\_brewer()   
  
hist <- ggplot(nhanes\_education,  
 aes(x = CFDDS,  
 fill = education,  
 y = ..density..)) +   
 geom\_histogram(binwidth = 5,  
 position = "identity", #default position is stack   
 alpha = 0.2) +  
 geom\_density(aes(color = education),  
 alpha = 0) +  
 labs(x = "Digit Symbol Substitution Test Score",  
 y = "Proportion of participants",  
 title = "Histogram of cognition score by educational attainment",  
 fill = "Educational attainment",  
 color = "Educational attainment")  
  
box <- ggplot(nhanes\_homework,  
 aes(x = age\_groups,  
 y = CFDDS)) +  
 geom\_boxplot(fill = "dodgerblue") +  
 labs(x = "Age group",  
 y= "Digit Symbol Substitution Test score",  
 title = "Boxplot of cognition score by age group")  
  
  
scatter <- ggplot(nhanes\_homework,  
 aes(x = INDFMPIR,  
 y = CFDDS)) +  
 geom\_point(alpha = 0.5,  
 color = "seagreen") +  
 geom\_smooth(method = lm) +  
 labs(x = "Poverty to income ratio",  
 y = "Digit Symbol Substition Test score",  
 title = "Scatterplot of poverty income ratio by digit symbol substitution test")  
  
  
ggarrange(bar, hist, box, scatter,   
 nrow = 2,   
 ncol = 2)

## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.  
## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.  
## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.

## `geom\_smooth()` using formula 'y ~ x'



Interpret the figure: The histogram of cognition score by educational attainment is in the upper right hand corner.