EDUC 641 Lab: Applied Statistics in Education and Human Services I

Lab 2: 10/4 and 10/5

## Importing data, exploring categorical variables, and exporting results

**Goals:**

1. Data types in R
2. Import data
3. Identify the main features of a dataset
4. Describe categorical variables of your dataset (tables and plots)

**Note:**

1. Work in groups and discuss with your peers as you go. Write down your questions as they arise.
2. Use Pink and Yellow stickies.
3. \*\*Everything inside the single quotes (‘) symbol and gray area is code. When typing the code in R, skip the quotes sign and type whatever is inside.
4. Run code through R Script line-by-line. To run a code, go anywhere on the line where you wrote the code and either click ‘Run’ using the mouse or press Ctr/Command + Enter on your keyboard. The output will be printed on the console.

**Foundational Concept**

**Data types in R:** R has four main data types: **integer**, **character**, **numerical**/**double**, and **logical**. This is *different* from the measurement scale you discussed in class. Information we need to reuse is saved in the names object in R.

integer\_eg <- 2L  
character\_eg <- "hi"  
logical\_eg <- FALSE  
double\_eg <- 2

To obtain information about objects, we can use the ‘**str()**’ and ‘**type of ()**’ functions.

**What if we need to group numbers or characters?**

group\_of\_numbers <- c(1,3,6,9)  
group\_of\_letters <- c("a","b", "c", "d")  
arrange\_numbers <- matrix(data = c(0,10,13,0,51,5,8,3,0),  
 nrow = 3,  
 ncol = 3,  
 byrow = TRUE)  
data\_eg <- mtcars

*Note: There are several types of objects you will work with in R. In my experience, vectors are the most common (like* group\_of\_numbers and data\_eg, above).

**Today’s Lab**

1. **Import dataset in R:**
2. First, download the [dataset](https://daviddliebowitz.github.io/EDUC641_24F/data/cat.csv) from the [class website](https://daviddliebowitz.github.io/EDUC641_24F/schedule.htmlhttps://daviddliebowitz.github.io/EDUC641_24F/schedule.html) and save it inside the *data* folder of your R-project
3. Open the R-project and the R-script you created in the labs folder last time.

cat\_df <- read.csv("data/name\_of\_dataset.csv")

1. **Identify the key features of the dataset.**
2. We can click the name of the dataset object (cat\_df) in Environment to open the dataset in a new tab. We can also use several functions to obtain information about the dataset.
3. **str()**
4. **dim()**
5. **colnames()**
6. **rownames()**
7. **head()**

Based on your output above, how many rows and columns are in the dataset? What variables are there?

1. To access specific columns of the dataset, we do the following:

name\_of\_data$name\_of\_column

1. **Describing categorical variables**

R uses a special datatype called factor to work with categorical variables. Therefore, the first cleaning step is to convert all categorical variables to factors.

1. First, you need to know how data is coded (levels) that appear in the column and what they actually mean (labels) based on the information you have on your dataset from your codebook.

name\_of\_data$name\_of\_categorical\_column <- factor(name\_of\_data$name\_of\_categorical\_column,  
 levels = c(...), #write the levels as they appear in the data object  
 labels = c(...) #write the labels, i.e., what they represent as characters  
 )

1. Counts and Proportion

table(name\_of\_data$name\_of\_categorical\_column)  
prop.table(name\_of\_data$name\_of\_categorical\_column))

1. Bar chart

library(tidyverse)   
#If you don't have tidyverse installed, do install.packages("tidyverse") first  
  
ggplot(name\_of\_data) +  
 geom\_bar(aes(x = name\_of\_categorical\_column), fill = "choose\_a\_color\_of\_your\_choice")+  
 labs(  
 x = "Give a descriptive x-axis label",  
 y = "Give a descriptive y-axis label",  
 title = "Give a descriptive title"  
 ) +  
 theme\_minimal()

**BONUS**

**Exporting tables and figures:** For assignment 1, you can either take a screenshot of your output, type out the tables, or export your results. Below are some ways to save your results.

1. First, save the tables and plots into a named object, like this,

counts <- table(name\_of\_data$name\_of\_categorical\_column)

barplot <- ggplot(name\_of\_data) +  
 geom\_bar(aes(x = name\_of\_categorical\_column), fill = "choose\_a\_color\_of\_your\_choice")+  
 labs(  
 x = "Give a meaningful x-axis label",  
 y = "Give a meaningful y-axis label",  
 title = "Give a meaningfule title"  
 ) +  
 theme\_minimal()

1. To export, use the following code. The files will be saved in the path you specify.

write.csv(counts, "tables/counts.csv")

ggsave(plot = barplot, "figures/barplot.jpeg") #uses default width and height, which can be changed

**Summarize for yourself.** Write all the functions you learned today to help yourself with the assignment!