**Session 1:**

**Starting in R:**

getwd() # returns current working directory

setwd("/Users/CaitlynLim/Documents/RScripts/Session 1")

**Reading data:**

**Stata:**

library(rio) # load required package

library(haven)

soep <- data.frame(as\_factor(read\_dta("data1.dta"))) # read data set

soep <- spread\_attrs(droplevels(gather\_attrs(soep))) # cleaning up

# x <-read.csv("NameOfFile.csv")

# x <- read.table("NameOfFile.txt")

**Overview of data:**

str(soep) #compile in short

summary(soep) #summary

head(soep) #show first and last part

**Frequencies as proportions:**

prop.table(table(soep$emp, useNA = "always"))

prop.table(table(soep$lsat, useNA = "always"))

# How would we get percentages instead of proportions?

# employment status by gender

table(soep$sex, soep$emp, useNA = "always")

mean(soep$income, **na.rm = TRUE**)

**Analyzing data graphically:**

# univariate analysis

# distribution of income (for non-zero incomes below 200,000 EUR)

hist(soep$income[soep$income > 0 & soep$income < 2e+05],

col = "lightblue",

breaks = 100

)

# bivariate analysis

# income by occupation (for non-zero incomes below 200,000 EUR)

boxplot(soep$income ~ soep$emp,

subset = soep$income > 0 & soep$income < 2e+05

)

**Regression analysis:**

# regression of income on gender and employment status

lm(income ~ sex + emp, data = soep)

# more detailed results

summary(lm(income ~ sex + emp, data = soep))

# restrict regression analysis to respondents with positive income

summary(lm(income ~ sex + emp, data = soep, subset = income > 0))

Session 2:

y[length(y)] # last element

x[x < 16] #Highlight x < 16 you can find out the true/false

# if index is TRUE, the corresponding element of x is selected

# changing content

countries[3] <- "Italy"

y[c(1, 3, 5)] <- c(100, 110, 120) # replaces 1st, 3rd, and 5th element

y[y > 10] <- 1 # set values greater than 10 to one