Questions 1-3
learndis.xlsx.
Overview of Learning Disabilities in Children Study

__1. (20 points total)

__a) By any means you like, import learndis.xlsx into R Studio or Stata. Tip - While in Excel, take care that the columns are saved in the appropriate formats (numeric or text).

__b) Produce a description of your imported dataset
In R the command is **`str( )`**

__c) Save your imported data to a permanent Stata or R dataset.

```
# Set working directory
setwd("~/Desktop")

# Check working directory
getwd()

# Turn off scientific notation
options(scipen=999)

# Clear the Decks
rm(list = ls())

#  Q1A:  Import learndis.xlsx
library(readxl)
learndis <- read_excel("learndis.xls")

#  Q1B:  Produce description of dataset
str(learndis)
summary(learndis)

#  Q1C:  Save to a permanent dataset
save(learndis, file="learndis.Rdata")
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     105 obs. of  6 variables:
 $ grade   : num  2 5 3 2 2 5 3 4 2 2 ...
 $ gender  : num  1 0 0 0 0 0 1 1 1 0 ...
 $ placemen: num  1 0 0 0 0 0 1 1 0 0 ...
 $ readcomp: num  72 81 67 84 76 83 65 65 82 72 ...
 $ mathcomp: num  68 82 115 98 78 88 92 73 96 84 ...
 $ iq      : num  55 71 71 89 80 80 97 67 87 105 ...
     grade          gender       placemen         readcomp         mathcomp            iq       
 Min.   :1.000   Min.   :0.0   Min.   :0.0000   Min.   : 22.00   Min.   : 61.00   Min.   : 51.0  
 1st Qu.:2.000   1st Qu.:0.0   1st Qu.:0.0000   1st Qu.: 69.75   1st Qu.: 77.00   1st Qu.: 74.0  
 Median :2.000   Median :0.0   Median :0.0000   Median : 79.00   Median : 84.00   Median : 80.0  
 Mean   :2.543   Mean   :0.4   Mean   :0.3714   Mean   : 77.63   Mean   : 86.28   Mean   : 81.5  
 3rd Qu.:3.000   3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.: 85.00   3rd Qu.: 95.00   3rd Qu.: 89.0  
 Max.   :5.000   Max.   :1.0   Max.   :1.0000   Max.   :107.00   Max.   :121.00   Max.   :105.0  
                                                NA's   :29       NA's   :11
```

__2. (20 points total)

__a) For all variables: Create variable labels

__b) For discrete variables ONLY: Create variable value labels.

__c) For all variables, as needed: Assign missing value codes.
Dear class: Huzzah. R imports blanks as missing for you!

__d) Produce again a description of your imported dataset
In Stata the command is **`str( )`**

```
#  Q2A:  For all variables, produce variable labels
library(Hmisc)
label(learndis$grade) <- "Grade Level"
label(learndis$gender) <- "Gender"
label(learndis$placemen) <- "Type of Placement"
label(learndis$readcomp) <- "Reading Comprehension"
label(learndis$mathcomp) <- "Math Comprehension"
label(learndis$iq) <- "Intellectual Ability"

#  Q2B:  For discrete variables ONLY:  Label discrete variable values
learndis$genderf <- factor(learndis$gender,
                    levels = c(0,1),
                    labels = c("male", "female"))
learndis$placemenf <- factor(learndis$placemen,
                    levels = c(0,1),
                    labels = c("Part-time", "Full-time Segregated"))


#  Q2C:  For all variables, as needed:  Assign missing value codes.
#  Preliminary:  Check for missing values (see p 78 of Unit 4 notes)
colSums(is.na(learndis))


#Q2D:   Produce, again, a description of your dataset
str(learndis)
```

```
    grade    gender  placemen  readcomp  mathcomp        iq   genderf placemenf
        0         0         0        29        11         0         0         0

Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     105 obs. of  8 variables:
 $ grade    : num  2 5 3 2 2 5 3 4 2 2 ...
  ..- attr(*, "label")= chr "Grade Level"
 $ gender   : num  1 0 0 0 0 0 1 1 1 0 ...
  ..- attr(*, "label")= chr "Gender"
 $ placemen : num  1 0 0 0 0 0 1 1 0 0 ...
  ..- attr(*, "label")= chr "Type of Placement"
 $ readcomp : num  72 81 67 84 76 83 65 65 82 72 ...
  ..- attr(*, "label")= chr "Reading Comprehension"
 $ mathcomp : num  68 82 115 98 78 88 92 73 96 84 ...
  ..- attr(*, "label")= chr "Math Comprehension"
 $ iq       : num  55 71 71 89 80 80 97 67 87 105 ...
  ..- attr(*, "label")= chr "Intellectual Ability"
 $ genderf  : Factor w/ 2 levels "male","female": 2 1 1 1 1 1 2 2 2 1 ...
 $ placemenf: Factor w/ 2 levels "Part-time","Full-time Segregated": 2 1 1 1 1 1 2 2 1 1 ...
```

__3. (20 points total)

__a) Create a grouped variable that you name **quart_iq** that has values 1, 2, 3, and 4 according to quartile of value of **iq**.

__b) Label your variable **quart_iq**

__c) Attach variable value labels to the values 1, 2, 3, and 4 of **quart_iq**

```
library(gtools)
library(Hmisc)

#Q3A:   Produce quart_iq that has value 1, 2, 3, and 4 according to quartiles of iq
#Q3C:   Following also answers question 3c (attach variable value labels)
learndis$quart_iq <- quantcut(learndis$iq, q=4, labels=c("Quartile 1","Quartile 2", "Quartile 3",
"Quartile 4"),na.rm=TRUE)


#Q3B:  Label your variable quart_iq
label(learndis$quart_iq) <- "Quartile IQ"

# Check.
table(learndis$quart_iq)
```

```
Quartile 1 Quartile 2 Quartile 3 Quartile 4
        31         22         26         26
```

Questions 4-5
gss1000.xlsx.
Overview of General Social Survey (GSS)

__4. (20 points total)

The variable **fepol** contains responses to the question "Female not suited for politics" and is coded 1=yes and 0=no. This is potentially confusing since a value of 1 is saying that the respondent believes females are not suited for politics.

__a) Create a more straightforward variable that you name **fepol_yes** and that is a reverse coding of **fepol.**

__b) Label this variable "Females suited for politics"

__c) Assign value labels of "Yes" to the value 1 and "No to the value 0.

```
# Import data
library(readxl)
gss1000 <- read_excel("gss1000.xlsx")


# Quick look at distribuiton of fepol
library(summarytools)
freq(gss1000$fepol)

#  Q4A:  Create fepol_yes (1=suited, 0=not) that is a reverse coding of fepol
gss1000$fepol_yes <- NA
gss1000$fepol_yes[gss1000$fepol==1] <- 0            # Not suited for politics is now coded = 0
gss1000$fepol_yes[gss1000$fepol==0] <- 1            # Suited for politics is now coded = 1

#  Q4B:  Label variable
label(gss1000$fepol_yes)  <- "Yes/no Suited"

# Q4C:   Assign variable value labels
gss1000$fepolf <- factor(gss1000$fepol,
                    levels = c(0,1),
                    labels = c("Suited", "Not Suited"))
gss1000$fepol_yesf <- factor(gss1000$fepol_yes,
                    levels = c(0,1),
                    labels = c("Not Suited", "Suited"))

# Check
table(gss1000$fepolf, gss1000$fepol_yesf)
```

```
Frequencies
gss1000$fepol
Type: Numeric

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          0    496     78.86          78.86     49.60          49.60
          1    133     21.14         100.00     13.30          62.90
      <NA>    371                                37.10         100.00
     Total   1000    100.00         100.00    100.00         100.00




            Not Suited Suited
  Suited             0    496
  Not Suited       133      0
```

__5. (20 points total)

The variable **socbar** contains responses to the question "Spend evening at bar" and is coded 1=never, 2 =once a year, 3=sev times a year, 4=once a month, 5=sev times a mnth, 6=sev times a week, 7=almost daily. There are missing values.

___5a) Create a new variable **socbar4** that is a grouping of the values of **socbar** as follows:

IF socbar =	THEN code socbar4 =	Assign socbar4 value label
missing	1	Unknown
1	2	Never
2 or 3 or 4	3	At most 1x/month
5 or 6 or 7	4	At least several x/month

__5b) Label your new variable.

___5c) Label your new variable values.

```
# Preliminary look at distribution of socbar
library(summarytools)
freq(gss1000$socbar)

#  Q5A:  Create scobar4
gss1000$socbar4 <- 1
gss1000$socbar4[gss1000$socbar==1] <- 2
gss1000$socbar4[(gss1000$socbar >=2) & (gss1000$socbar <= 4)] <- 3
gss1000$socbar4[(gss1000$socbar >=5) & (gss1000$socbar <= 7)] <- 4


#  Q5B:
#  Q4B:  Label variable
library(Hmisc)
label(gss1000$socbar4)  <- "Time Spend in Bar, Grouped"

# Q4C:   Assign variable value labels
gss1000$socbar4f <- factor(gss1000$socbar4,
                    levels = c(1,2,3,4),
                    labels = c("1=Unknown", "2=Never", "3=At most 1x/mos", "4=Several x/mos"))

# Check
freq(gss1000$socbar4f)
```

```
Frequencies
gss1000$socbar
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          1    321     48.86          48.86     32.10          32.10
          2     97     14.76          63.62      9.70          41.80
          3     74     11.26          74.89      7.40          49.20
          4     67     10.20          85.08      6.70          55.90
          5     58      8.83          93.91      5.80          61.70
          6     34      5.18          99.09      3.40          65.10
          7      6      0.91         100.00      0.60          65.70
       <NA>    343                                34.30         100.00
      Total   1000    100.00         100.00    100.00         100.00
Frequencies
gss1000$socbar4f
Type: Factor

                        Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------------------- ------ --------- -------------- --------- --------------
             1=Unknown    343     34.30          34.30     34.30          34.30
               2=Never    321     32.10          66.40     32.10          66.40
      3=At most 1x/mos    238     23.80          90.20     23.80          90.20
       4=Several x/mos     98      9.80         100.00      9.80         100.00
                 <NA>      0                                0.00         100.00
                Total   1000    100.00         100.00    100.00         100.00
```