Note:

- There are three parts to this assignment.
- For Part A and Part B, please install the `tidyverse` package using `install.packages` so that you can use the `dplyr` library. Once you have done that use the `read_csv()` function to read the csv files and not the `read.csv()` function. Please send me an email if you have problems with installing. I will not be able to help or provide extensions if you contact me late.
- You will need to submit your assignment as a markdown document that the graders will run on their machine. As a precaution, please submit your corresponding html files also.

Part A (40 marks)

The file `household.csv` contains (fictional) data from a survey of 500 randomly selected households.

a. Indicate the type of data for each of the variables included in the survey.
b. For each of the categorical variables in the survey, indicate whether the variable is nominal or ordinal. Explain your reasoning in each case.
c. Create a histogram for each of the numerical variables in this data set. Indicate whether each of these distributions is approximately symmetric or skewed. Which, if any, of these distributions are skewed to the right? Which, if any, are skewed to the left?
d. Find the maximum and minimum debt levels for the households in this sample.
e. Find the indebtedness levels at each of the 25th, 50th, and 75th percentiles.
f. Find and interpret the interquartile range for the indebtedness levels of these selected households
g. Write a report that is less than 250 words that summarizes your analysis.

Name your file as A03b_Gwid.rmd. So if your GWID is G19860011 then you should name your submission file as A03b_G19860011.rmd. Please make sure that you comment your R code.

Part B (40 marks)

The file `SupermarketTransactions.csv` contains data on over 14.000 transactions. There are two numerical variables, Units Sold and Revenue. The first of these is discrete and the second is continuous. For each of the following, **do whatever it takes** to create a bar chart of counts for Units Sold and a histogram of Revenue for the given subpopulation of purchases.

a. All purchases made during January and February of 2008[1].
b. All purchase made by married female homeowners.
c. All purchases made in the state of California.
d. All purchases made in the Produce product department.

Write a report that is less than 250 words that summarizes your analysis.

Name your file as A03a_Gwid.md. So if your GWID is G19860011 then you should name your submission file as A03a_G19860011.rmd. Please make sure that you comment your R code.

[1] Use the date conversion facility in R to convert the dates which are strings to the date format by repurposing the following example: `dates <- as.Date(strDates, "%m/%d/%Y")` from this link. To compare dates you can refer to this link.

Part C

All of you must have heard about the central limit theorem (CLT). If not, have a look at this video.

a. Then run the following R commands. Please spend some time trying understand the code well.

```
rnorm2 <- function(n,mean,sd) { mean+sd*scale(rnorm(n)) }
set.seed(1239)
r1 <- rnorm2(100,25,4)
r2 <- rnorm2(50,10,3)
samplingframe <- c(r1,r2)
hist(samplingframe, breaks=20,col = "pink")
```

   Please describe the distribution that you obtain.

b. Draw 50 samples of size 15, and plot the sampling distribution of means as a histogram.
c. Draw 50 samples of size 45, and plot the sampling distribution of means as a histogram.
d. Make sure that the distributions in parts b and c are on the same plot. Explain the three histograms in terms of their differences and similarities.
e. Explain CLT in your own words. Restrict your response to less than 50 words.
f. Does this exercise help you understand CLT? If so why? If not, why not? Restrict your response to less than 50 words.

Name your file as A03c_Gwid.rmd. So if your GWID is G19860011 then you should name your submission file as A03c_G19860011.rmd. Please make sure that you comment your R code.