How many questions did you complete (a completed question means that all the sub parts were done)? Write your answer as a fraction of the total number of questions **on the very top of your assignment: Example 10/15**

Please answer all questions. Remember this assignment is worth 15/4% and is your first assignment for the course. Use R markdown to create the final document and store in a safe area till finished, all working must be shown in the assignment answers.

Be careful with your files and be organized. Keep all data, R, R Markdown etc files inside the one directory.

All statistical computing is to be done in **R** , this does not mean I want screeds of output! Only use **R** when needed and only to answer the question.

Please note that **MS**=Mendenhall and Sincich, *STATISTICS for science and engineering* 6th edition. You will need to convert the `.xls` files into `.csv` files in excel and use `read.table( ..., header=TRUE,sep=‘‘,’’)` or you can use `read.csv()` on csv files or you may wish to use the `readxl` package to read .xls files directly

Once you have made the R script file do the following:

- Make the .rmd file in RStudio.
- Use RStudio to knit the R markdown document to an html file.
- Place both the rmd and html files in the **dropbox before the due date**.

Late assignments get zero.

Please answer the following questions as found in MS as well as the additional questions placed in the text below.

All working MUST be shown and answers formatted in R markdown as shown in class

1. Summarize how I will workout your final grade for the course. Give percentages etc. Give my grading scale also e.g. What percentage is an A etc.

2. A biologist wants to make a coplot of **LENGTH Vs WEIGHT** given **RIVER*SPECIES** for fish caught in the Tennessee river and recorded in the **DDT.csv** data set, so that each point is colored according to the variable **MILE** which is treated as a factor (Qualitative variable).

```
> head(ddt)
  RIVER MILE  SPECIES LENGTH WEIGHT DDT
1   FCM    5 CCATFISH   42.5    732  10
2   FCM    5 CCATFISH   44.0    795  16
3   FCM    5 CCATFISH   41.5    547  23
4   FCM    5 CCATFISH   39.0    465  21
5   FCM    5 CCATFISH   50.5   1252  50
6   FCM    5 CCATFISH   52.0   1255 150
# The following code may help
m=with(ddt, as.numeric(factor(MILE))) # A
length(unique(m)) #B
```

   (a) Make the coplot as the biologist required **Hint:** Use `coplot()`, Lab 1, the code provided, and plotting options pch and col to differentiate the MILE variable. You should be able to produce something like what is shown below

   (b) Interpret the lower left three conditional plots.

   (c) What does line A do?

   (d) What does line B do?

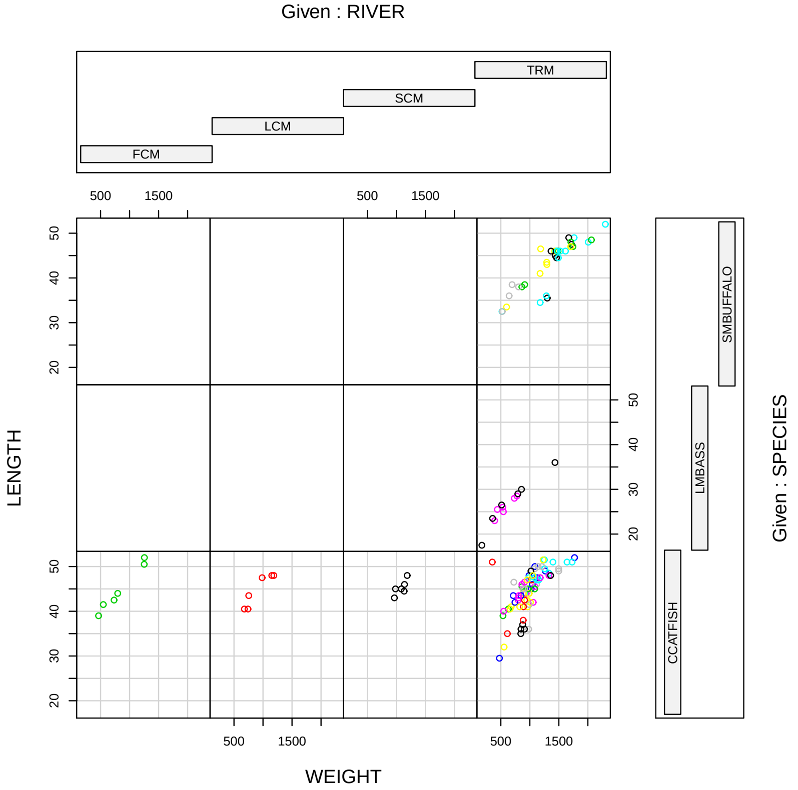
   (e) Why are the top six plots empty?

   (f) What is the mean value of DDT found in the sample of CCATFISH caught in the FCM river? **Hint:**

```
ddt=read.csv("..\\CSV\\DDT.csv")
head(ddt)
subset(ddt,RIVER=="FCM" & SPECIES=="CCATFISH",) #or
ddt[ddt$RIVER=="FCM" & ddt$SPECIES=="CCATFISH",]
```

3. **MS 1.14 - pg 8**

4. **MS page 12,13** Read pages 12 and 13 about random sampling designs and answer the following:



(a) What are the names of the four random sampling designs (1 simple and 3 more complex).

(b) Give a brief description of each.

5. **MS 1.15 - pg 15** – Use `sample(...,replace=FALSE)`, if `mtbe` is the dataframe then we need a random sample of the rows. If v is a vector containing a random sample of row indices then `mtbe[v,]` will be the random sample.

```
mtbe=read.csv("..\\CSV\\MTBE.csv", header=TRUE) # You will need to change the address
head(mtbe) # First six lines
dim(mtbe) # rows and columns
ind=sample(1:223,5,replace=FALSE) # random indices
mtbe[ind,]
```

(a) Answer the additional problems below

(i) Remove all the rows in mtbe that contain one or more NA's `mtbeo=na.omit(mtbe)`

(ii) Now calculate the standard deviation (`sd()` in R) of the depth of wells which have "Bedrock" as the Aquifier (this is using the entire `mtbeo` data frame), **Hint: You will need to alter the following code**

```
depth=mtbeo[mtbeo$Aquifier=="Unconsoli",]$Depth
  mean(depth)
```

6. **MS 1.16 - pg 15** – Use `sample(...,replace=FALSE)`, if `eq` is the dataframe then we need a

random sample of the rows. If v is a vector containing a random sample of row indices then `eq[v,]` will be the random sample.

(a) Answer the additional problems below

   (i) Make the following plot `plot(ts(eq$MAG))` and record it here:

   (ii) Using the entire `eq` data frame find the median (`median()`) of the MAGNITUDE variable.

7. MS STATISTICS IN ACTION Read the story on page 18 then answer the following:

(a) What is the data collection method?

(b) What is the population?

(c) Give the names of all the **qualitative** variables.

8. MS 2.1 - pg 26 Use `pareto()` **Hint:**

```
freq=c(15,8,63,20)
RL=c("None","Both","LegsO","WheelsO")
l=rep(RL,freq)
```

9. MS 2.4 - pg 27 - Please use the `pareto()` function I made.

10. MS 2.10 - pg 28 – Use `pie3D()` from plotrix package (may need to install it) **Hint:**

```
swd=read.csv("..//CSV//SWDEFECTS.csv", header=TRUE)
head(swd)
library(plotrix)
tab=table(swd$defect)
rtab=tab/sum(tab)
round(rtab,2)
pie3D(rtab,labels=list("OK","Defective"),main="pie plot of SWD")
```

11. MS 2.72 - pg 70 When answering this question you will need to do most of the construction by hand. Unlike other questions please follow parts a) -m) in conjunction with MS as I have given below. For constructing the histogram and table below use the left end point as 8.0 and right end point as 10.6, with 9 classes. After constructing table 1 make the graph in **R** using `barplot(...,space=0)`, use the classes as names to the vector containing the frequencies.

(a) Fill out the table when constructing the Histogram in pt a). Then plot the histogram by first creating a vector, 'v' say, of relative frequencies, then use `names(v)` and assign class names to each component, finally using `barplot(v,space=0)` make your plot.

Class	Class Interval	Data Tabulation	Frequency	Relative Frequency
1	8.0000-8.2889			
2				
3				
4				
5				
6				
7				
8				
9				
Total				

Table 1: Histogram table

(b) Use the `stem()` function in **R** for part b).

(c) Use **R** to make the histogram. Do NOT use `hist()`

Hint: You may wish to use the following functions `subset(...,subset=LOCATION=="NEW")`, `cut()`, `table()`, `barplot(...,space=0)` and `?cut` etc See in class instruction concerning this and ..,

```
new<-subset(voltage.df,subset=LOCATION=="NEW")
new$VOLTAGE->vtn
vtn
max(vtn)
min(vtn)
lept<-min(vtn)-0.05
rept<-max(vtn)+0.05
rnge<-rept-lept
inc<-rnge/9
inc
seq(lept, rept,by=inc)->cl
cl
cvtn<-cut(vtn,breaks=cl)
new.tab=table(cvtn)
barplot(new.tab,space=0,main="Frequency Histogram(NEW)",las=2)
hist(vtn,nclass=10)
```

(d) Now complete d)-m) – You can use any of the built in R functions

12. MS 2.73 - pg 70

13. MS 2.80 - pg 72

14. MS 2.84 - pg 74

15. Using the ddt data set re-create the plot below using `ggplot`.
Make sure your plot is titled with your name. NB – You MUST use `ggplot()`

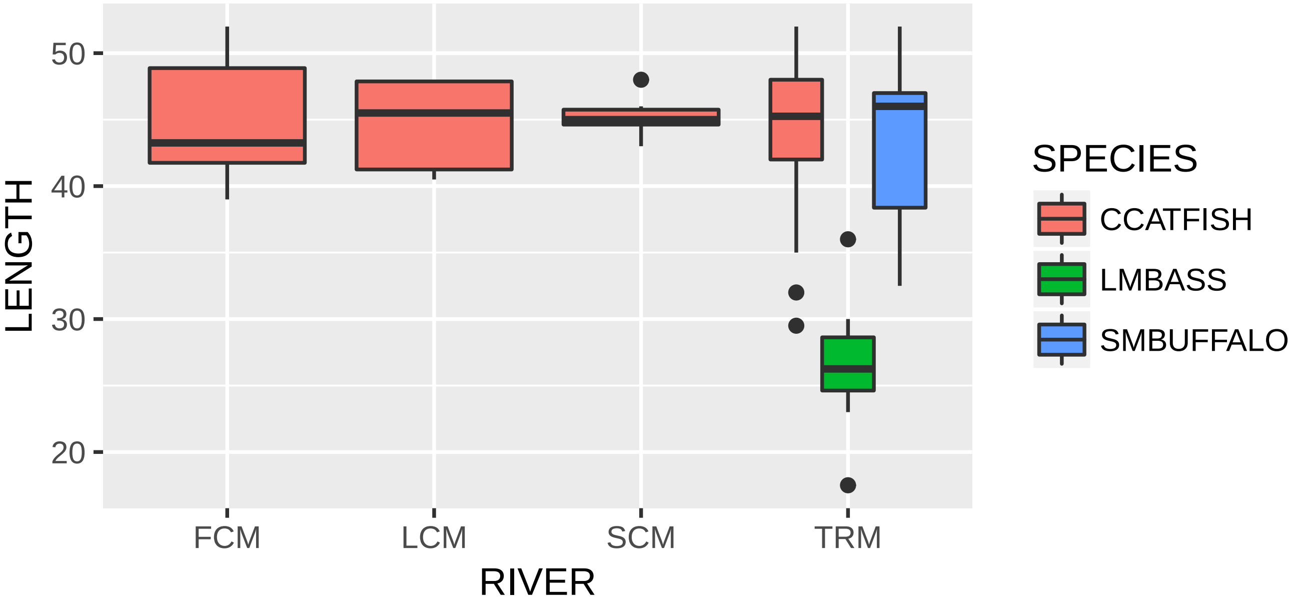


Figure 1: GGPLOT used to make this image