STAT 757 - Assignment 1

Jane Doe

February 2, 2018

i. DataCamp: Introduction to R [30 points]

Please complete the course an Introduction to R. You should have received an email with an invitation link. Please email me if you did not. If you already know R, please talk to me in class and follow up with an email to opt out.

Completed. Records are in DataCamp.

ii. Instructions for the rest of this assignment

The purpose of this portion of the assignment is to get a little experience making R Markdown documents as a way of nicely formatting output from R code while exploring the datasets from Sheather Ch.1 and learning to generate realizations of random variables (aka "fake data"). Modify this RMarkdown file (STAT_757_Assignment1.Rmd) and compile your document as a PDF (or Word document if you're having LaTeX issuse) and naming it according to the format SURNAME-FIRSTNAME-Assignment1.pdf, and emailing that PDF to the instructor by the due date listed above.

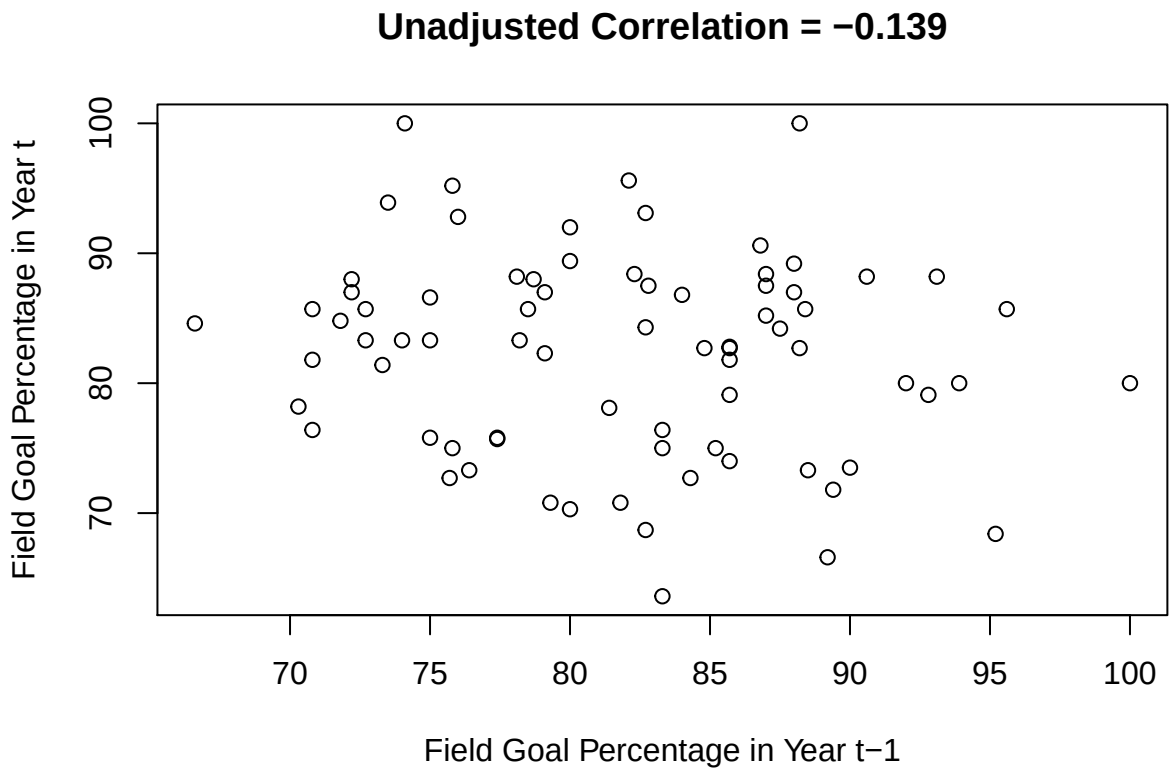
2. Reproduce the plots from Sheather Ch.1 [40 points]

Modify this file so that it reproduces all the output from the R script located at http://www.stat.tamu.edu/~sheather/book/docs/rcode/Chapter1.R. I've done the plots for the first dataset for you below. Remember that you will need to download each of the four data sets from http://www.stat.tamu.edu/~sheather/book/data_sets.php, and set your working directory (under the "Session" menu in Rstudio) appropriately. (And yes, this really is as easy as copying the blocks of R code for each dataset into this document into the appropriate places!) Need help? First, see http://rmarkdown.rstudio.com. Especially the resources under Learning More (http://rmarkdown.rstudio.com/#learning-more).

Below are the plots that appear in Chapter 1 of the textbook. They were created from the R script http://www.stat.tamu.edu/~sheather/book/docs/rcode/Chapter1.R and the data files at http://www.stat.tamu.edu/~sheather/book/data_sets.php.

Assessing the ability of NFL Kickers

```
kicker <- read.csv("~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/FieldG
## Sorry this line is too long, the data are labeled 'FieldGoals2003to2006.csv'

attach(kicker) ## THIS IS NOT USUALLY RECOMMENDED, ASK ME IN CLASS WHY NOT.

#Figure 1.1 on page 2
plot(kicker$FGtM1,kicker$FGt,
main="Unadjusted Correlation = -0.139",
xlab="Field Goal Percentage in Year t-1",ylab="Field Goal Percentage in Year t")
```




```
#p-values on page 3
fit.1 <- lm(FGt~FGtM1 +Name +FGtM1:Name,data=kicker)
anova(fit.1)
```

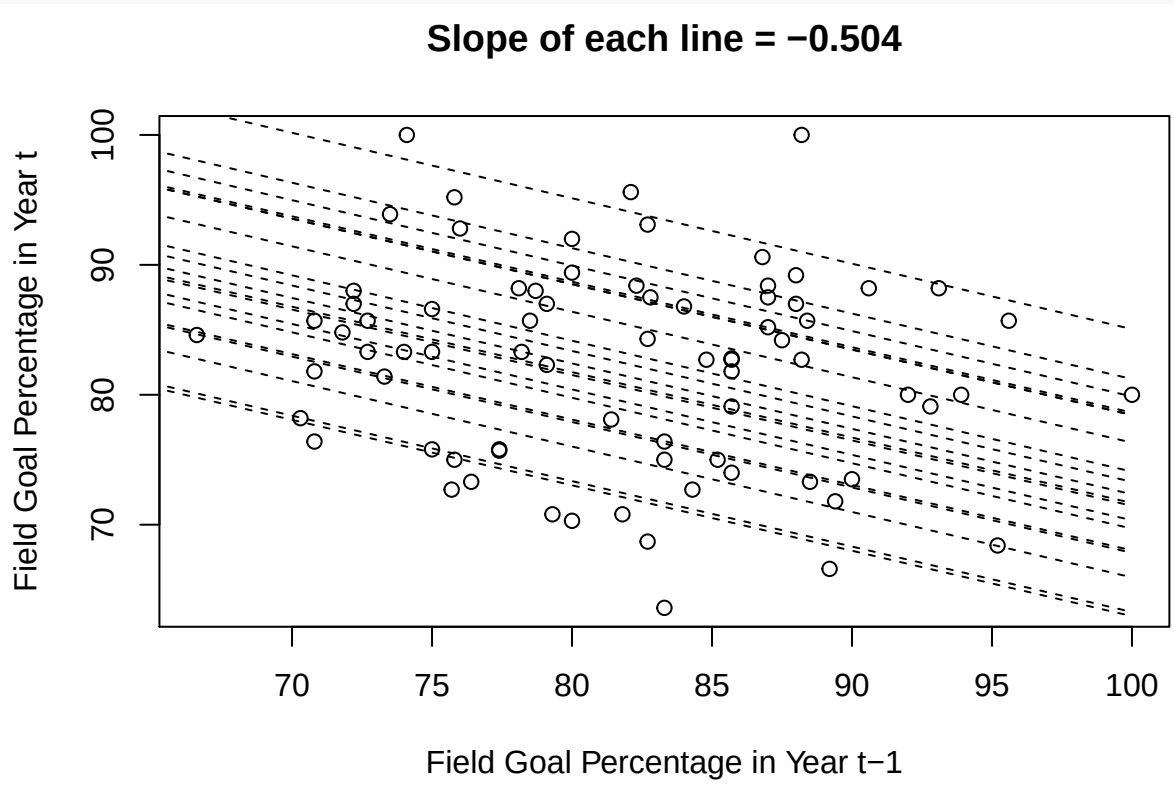
```
## Analysis of Variance Table
##
## Response: FGt
##             Df Sum Sq Mean Sq F value Pr(>F)
## FGtM1        1     87    87.2    1.90 0.1760
## Name        18   2252   125.1    2.73 0.0046
## FGtM1:Name 18    418    23.2    0.51 0.9386
## Residuals  38   1743    45.9
```

```
#slope and interecepts of lines in Figure 1.2 on page 3
fit.2 <- lm(FGt ~ Name + FGtM1,data=kicker)
fit.2
```

```
##
## Call:
## lm(formula = FGt ~ Name + FGtM1, data = kicker)
##
## Coefficients:
##               (Intercept)           NameDavid Akers
##                   126.687                   -4.646
##           NameJason Elam          NameJason Hanson
##                    -3.017                    2.117
##            NameJay Feely             NameJeff Reed
##                   -10.374                   -8.296
##         NameJeff Wilkins           NameJohn Carney
##                     2.310                   -5.977
##            NameJohn Hall            NameKris Brown
```

```
##                   -8.486                  -13.360
##          NameMatt Stover       NameMike Vanderjagt
##                   8.736                    4.896
##         NameNeil Rackers           NameOlindo Mare
##                  -6.620                  -13.036
##          NamePhil Dawson          NameRian Lindell
##                   3.552                   -4.867
##        NameRyan Longwell  NameSebastian Janikowski
##                  -2.231                  -3.976
##        NameShayne Graham                    FGtM1
##                   2.135                   -0.504
```

```
#Figure 1.2 on page 3
plot(kicker$FGtM1,kicker$FGt,
main="Slope of each line = -0.504",
xlab="Field Goal Percentage in Year t-1",
ylab="Field Goal Percentage in Year t")
tt <- seq(60,100,length=1001)
slope.piece <- summary(fit.2)$coef[20]*tt
lines(tt,summary(fit.2)$coef[1]+slope.piece,lty=2)
for (i in 2:19)
{lines(tt,summary(fit.2)$coef[1]+summary(fit.2)$coef[i]+slope.piece,lty=2)}
```

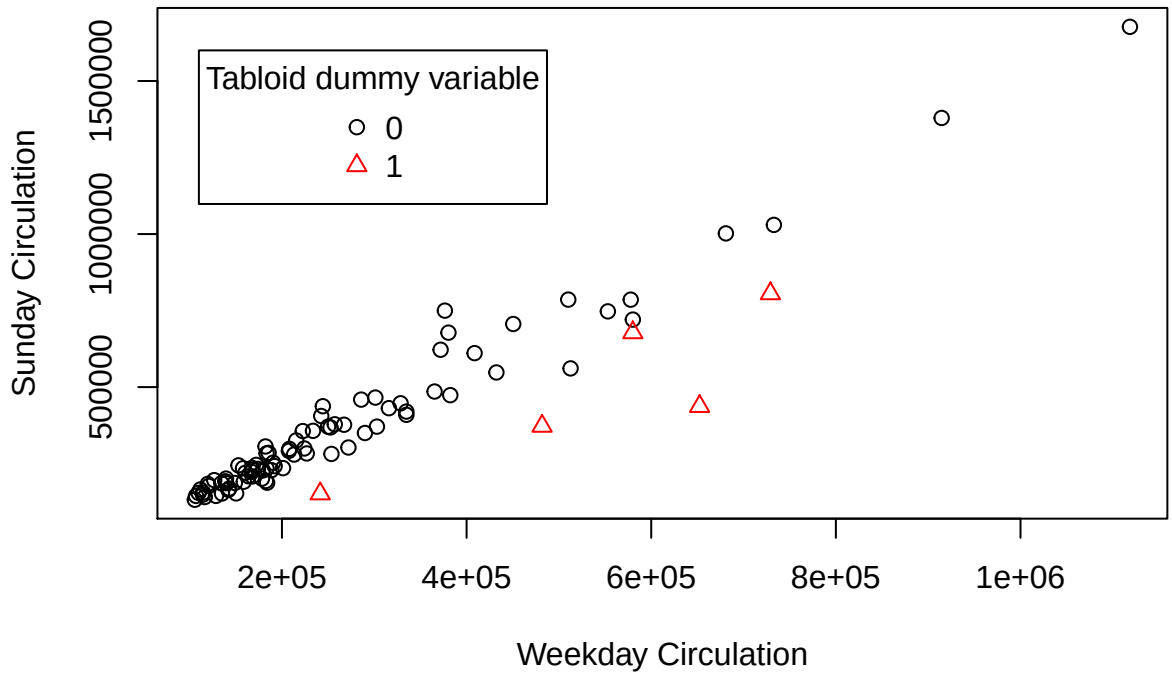



```
detach(kicker)
```

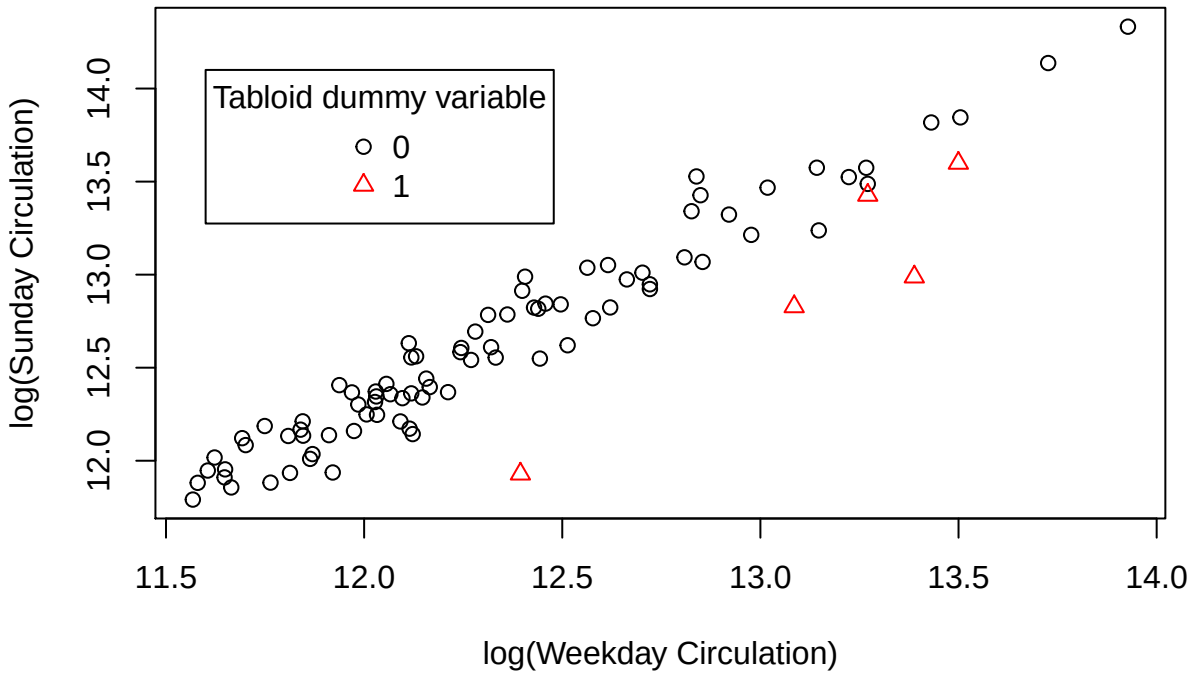
###Newspaper circulation

```
circulation <- read.table("~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/
attach(circulation)

#Figure 1.3 on page 5
```

```
plot(Weekday,Sunday,xlab="Weekday Circulation",ylab="Sunday Circulation",
pch=Tabloid.with.a.Serious.Competitor+1,col=Tabloid.with.a.Serious.Competitor+1)
legend(110000, 1600000,legend=c("0","1"),
pch=1:2,col=1:2,title="Tabloid dummy variable")
```



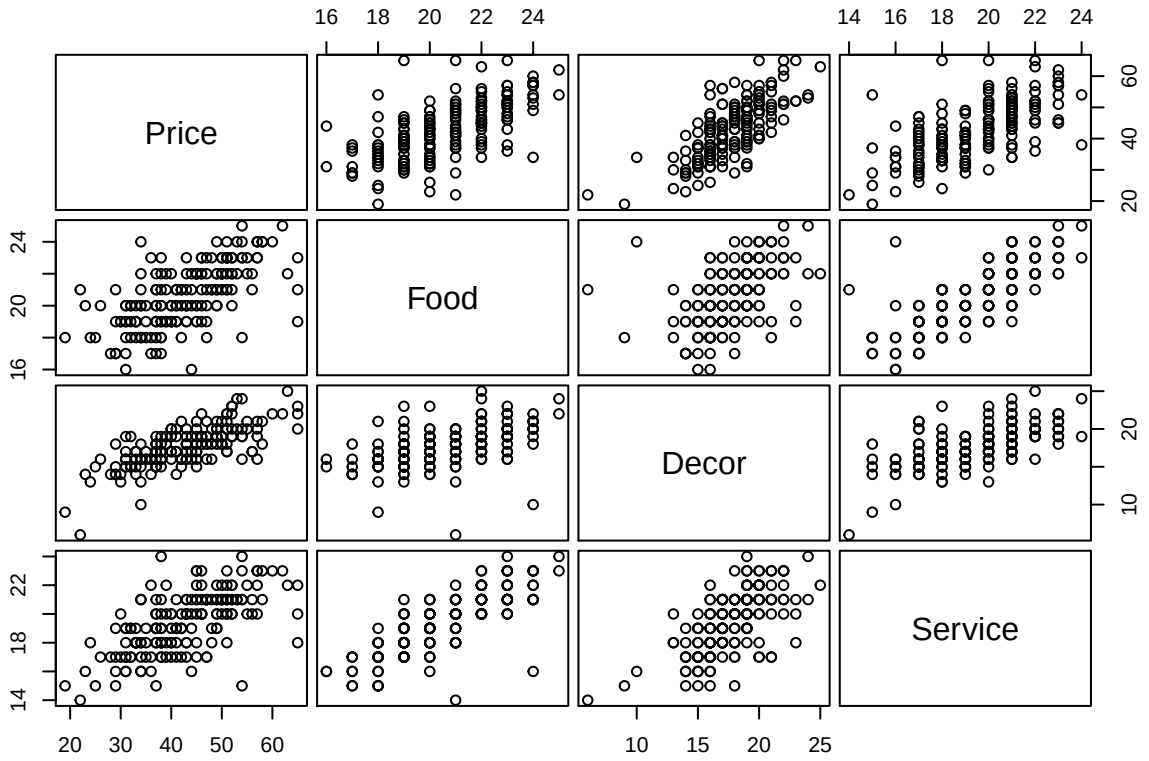

```
#Figure 1.4 on page 5
plot(log(Weekday),log(Sunday),xlab="log(Weekday Circulation)",ylab="log(Sunday Circulation)",
pch=Tabloid.with.a.Serious.Competitor+1,
col=Tabloid.with.a.Serious.Competitor+1)
legend(11.6, 14.1,legend=c("0","1"),pch=1:2,col=1:2,
title="Tabloid dummy variable")
```



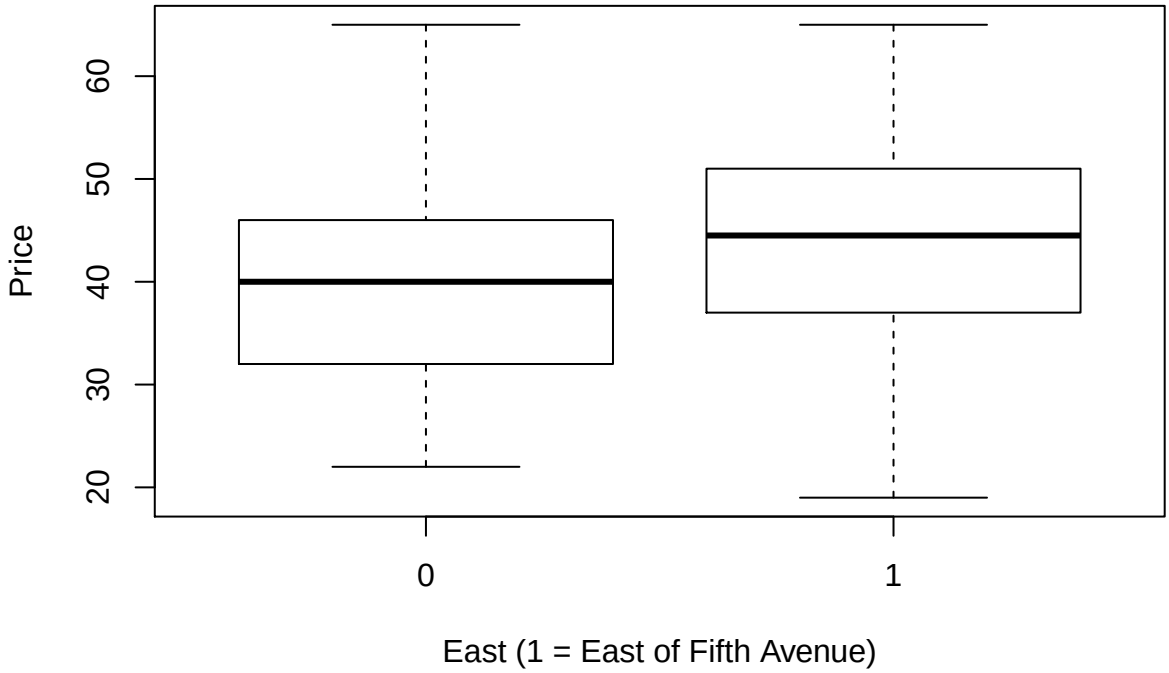

```
detach(circulation)
```

Menu pricing in a new Italian restaurant in NYC

```
nyc <- read.csv("~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/nyc.csv",
attach(nyc)

#Figure 1.5 on page 7
pairs(Price~Food+Decor+Service,data=nyc,gap=0.4,
cex.labels=1.5)
```



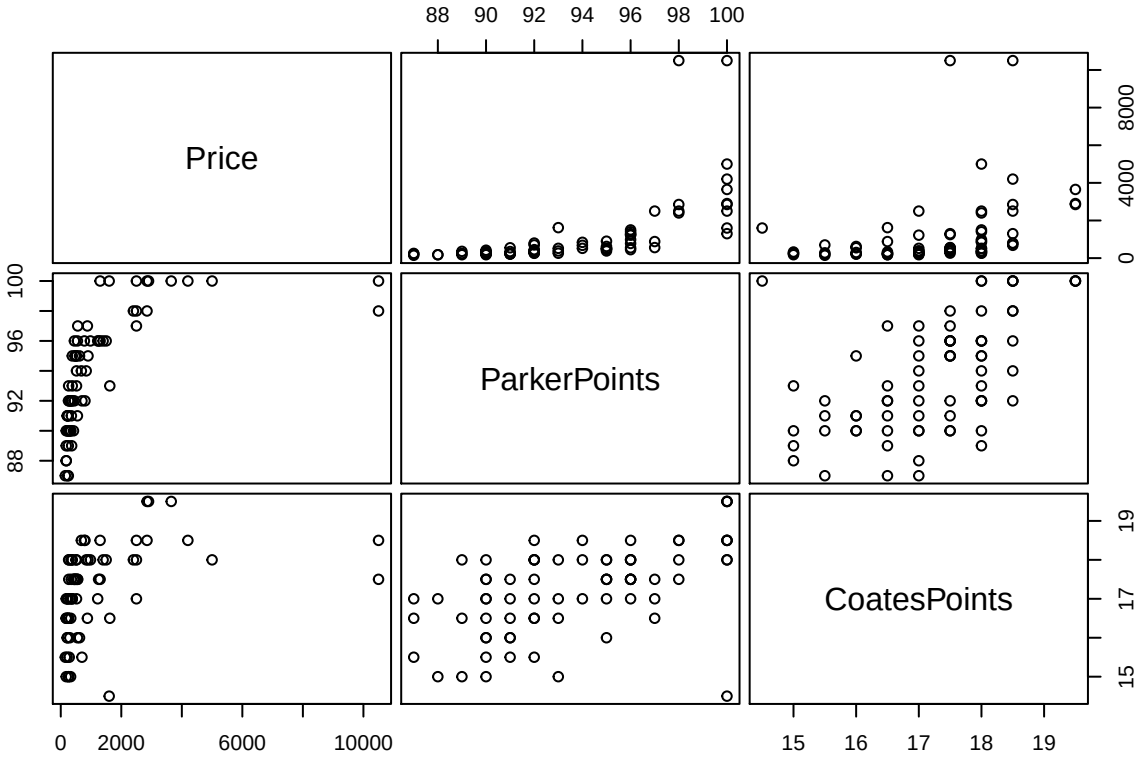

```
#Figure 1.6 on page 10
boxplot(Price~East,ylab="Price",
xlab="East (1 = East of Fifth Avenue)")
```



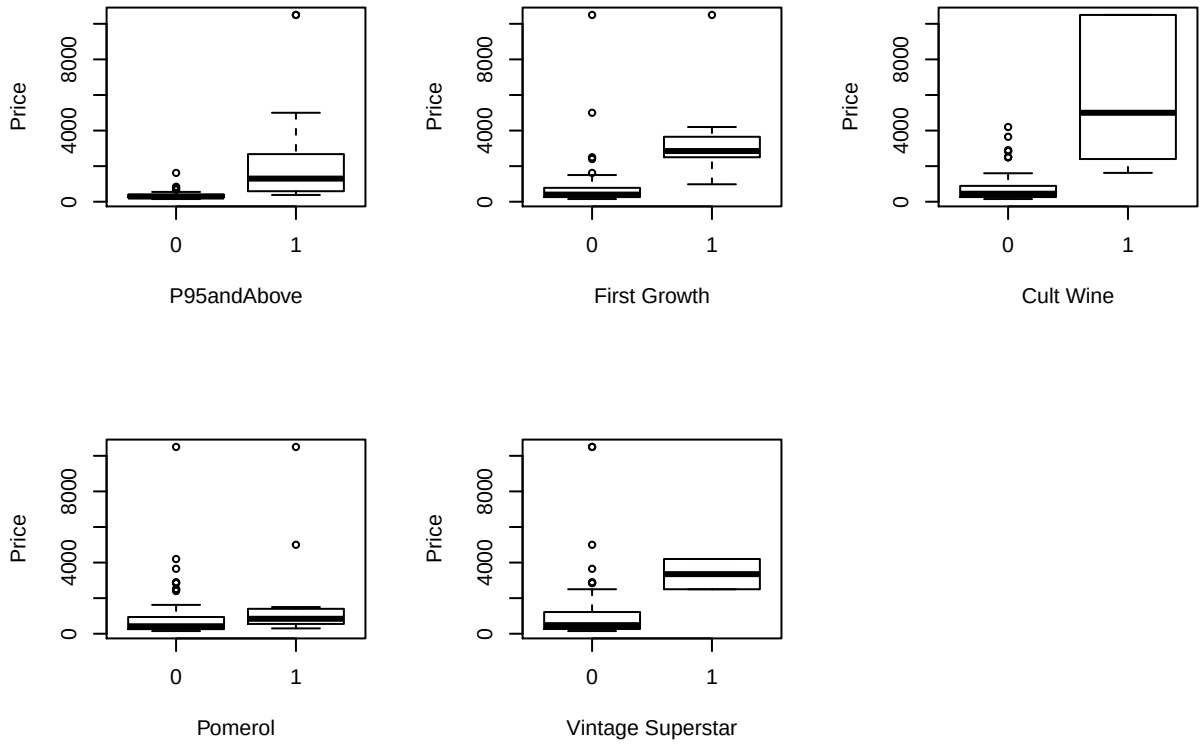

```
detach(nyc)
```

Effect of wine critics' ratings on prices of Bourdeax wines

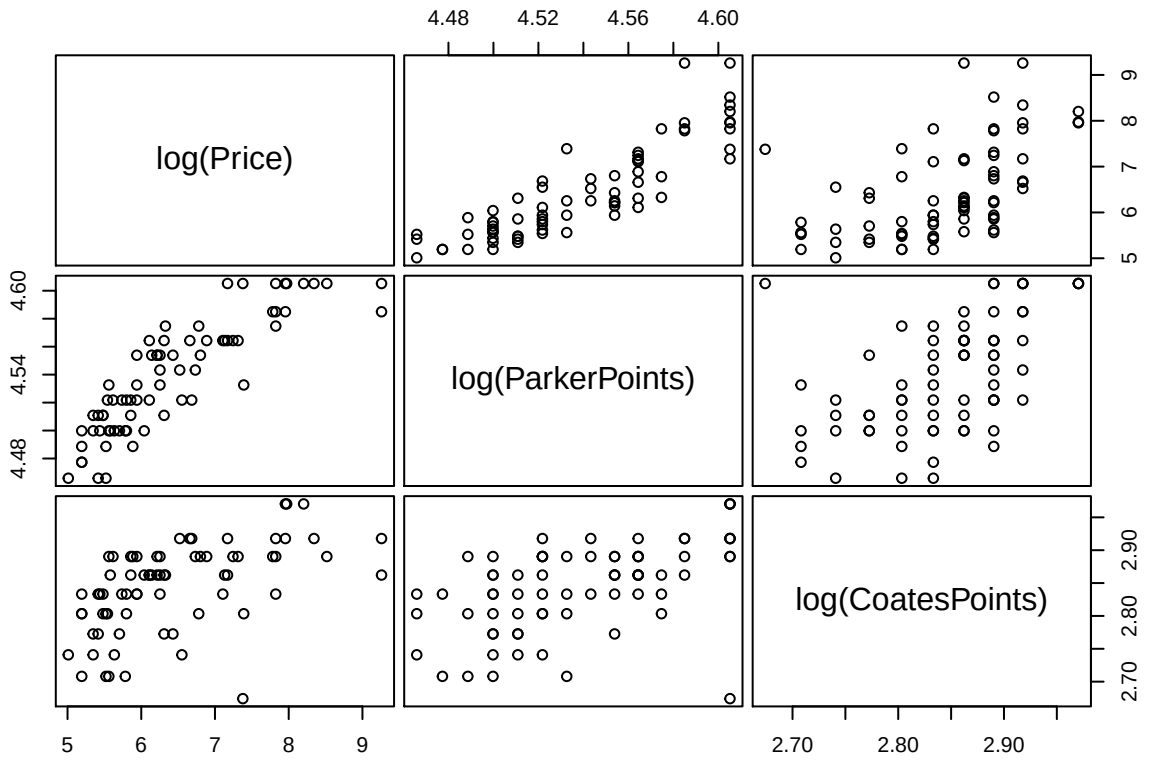
```
Bordeaux <- read.csv("~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/Bordeaux.csv")
attach(Bordeaux)

#Figure 1.7 on page 10
pairs(Price~ParkerPoints+CoatesPoints,data=Bordeaux,gap=0.4,cex.labels=1.5)
```




```
#Figure 1.8 on page 11
par(mfrow=c(2,3))
boxplot(Price~P95andAbove,ylab="Price",xlab="P95andAbove")
boxplot(Price~FirstGrowth,ylab="Price",xlab="First Growth")
boxplot(Price~CultWine,ylab="Price",xlab="Cult Wine")
boxplot(Price~Pomerol,ylab="Price",xlab="Pomerol")
boxplot(Price~VintageSuperstar,ylab="Price",xlab="Vintage Superstar")

#Figure 1.9 on page 12
par(mfrow=c(1,1))
```

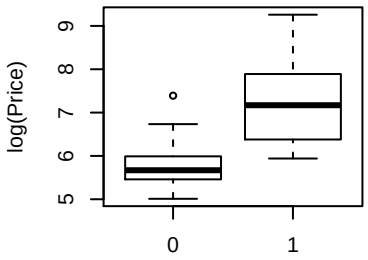


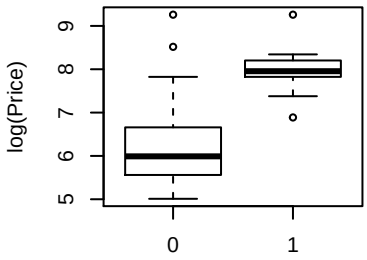

```
pairs(log(Price)~log(ParkerPoints)+log(CoatesPoints),data=Bordeaux,gap=0.4,cex.labels=1.5)
```

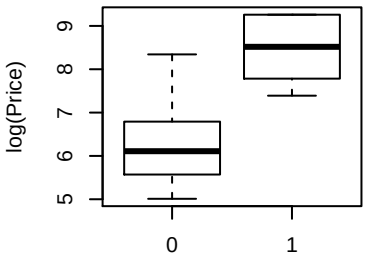


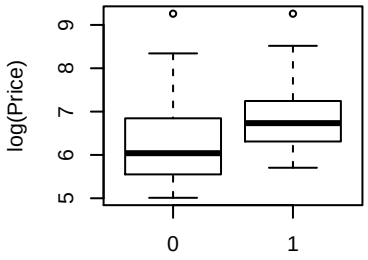

```
#Figure 1.10 on page 13
par(mfrow=c(2,3))
boxplot(log(Price)~P95andAbove,ylab="log(Price)",
xlab="P95andAbove")
```

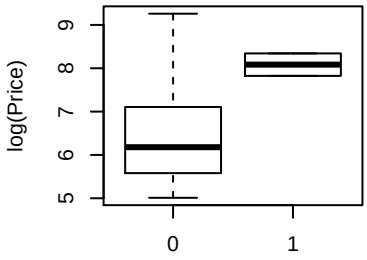
```
boxplot(log(Price)~FirstGrowth,ylab="log(Price)",
xlab="First Growth")
boxplot(log(Price)~CultWine,ylab="log(Price)",
xlab="Cult Wine")
boxplot(log(Price)~Pomerol,ylab="log(Price)",
xlab="Pomerol")
boxplot(log(Price)~VintageSuperstar,ylab="log(Price)",
xlab="Vintage Superstar")

detach(Bordeaux)
```







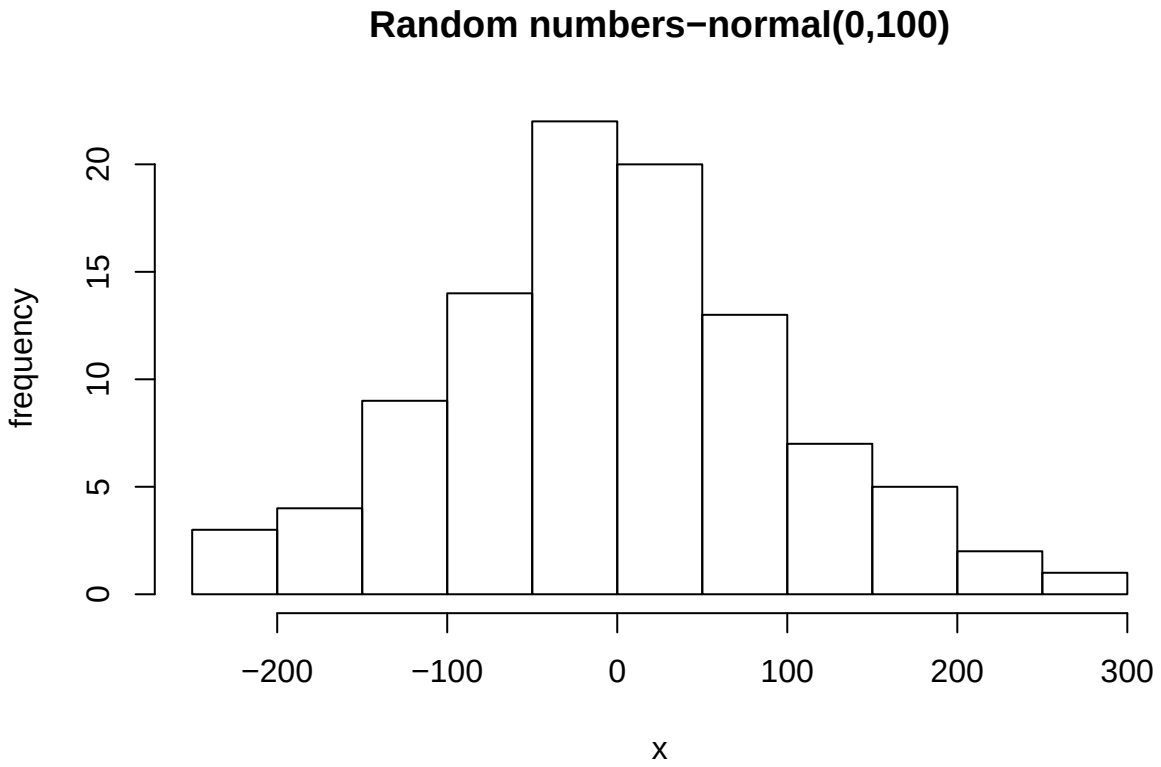




3. Generating fake data [30 points]

3.1 Generate 100 random variates from a normal distribution with mean 0 and standard deviation of 100. Summarize and plot the data. (Set a seed to make it reproducible).

```
set.seed(100)
sample1<-rnorm(100,0,100)
hist(sample1,xlab="x",ylab="frequency",main="Random numbers-normal(0,100)")
```



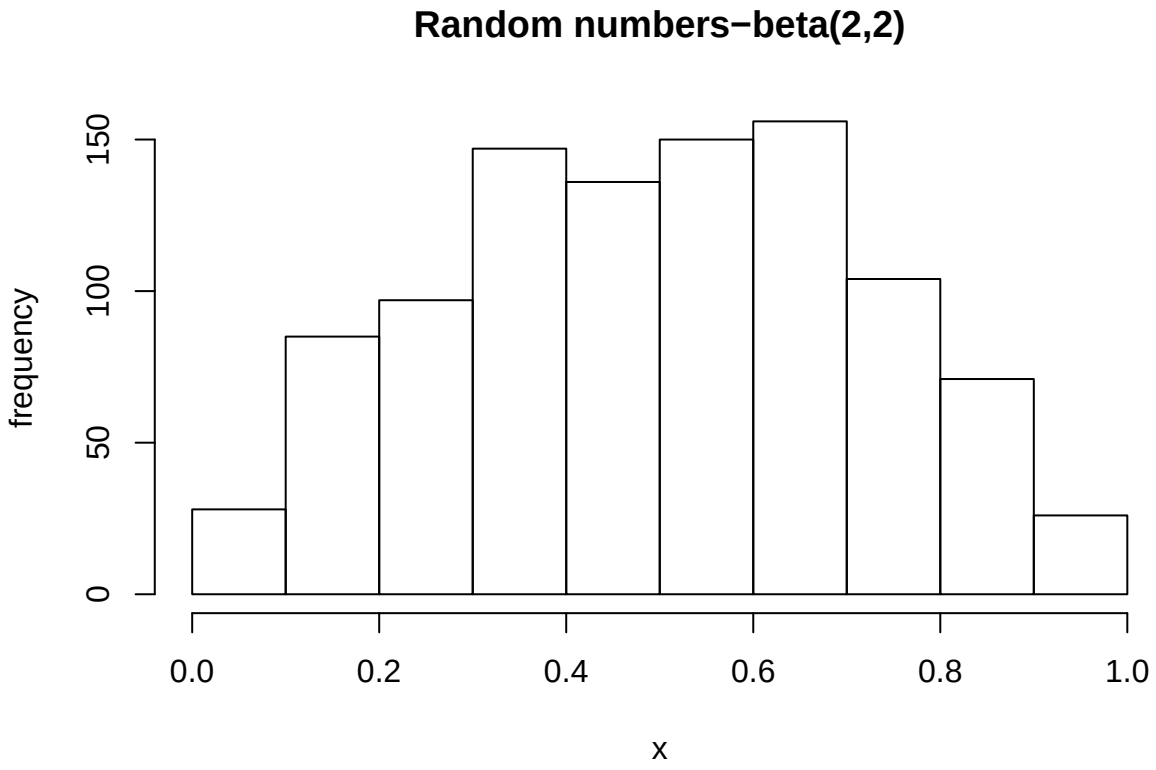

```
summary(sample1)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -227.193  -60.885   -5.942    0.291   65.589  258.196
```

At 0.29, the sample mean is fairly close to the distribution mean of zero; the min/max values are inline with the standard deviation; and the shape of the histogram looks like the bell-shaped normal distribution.

3.2 Generate 1000 random variates from a beta distribution with the parameters α and β both equal to 2. Summarize and plot the data. (Set a seed to make it reproducible).

```
set.seed(1000)
sample2<-rbeta(1000,2,2)
hist(sample2,xlab="x",ylab="frequency",main="Random numbers-beta(2,2)")
```



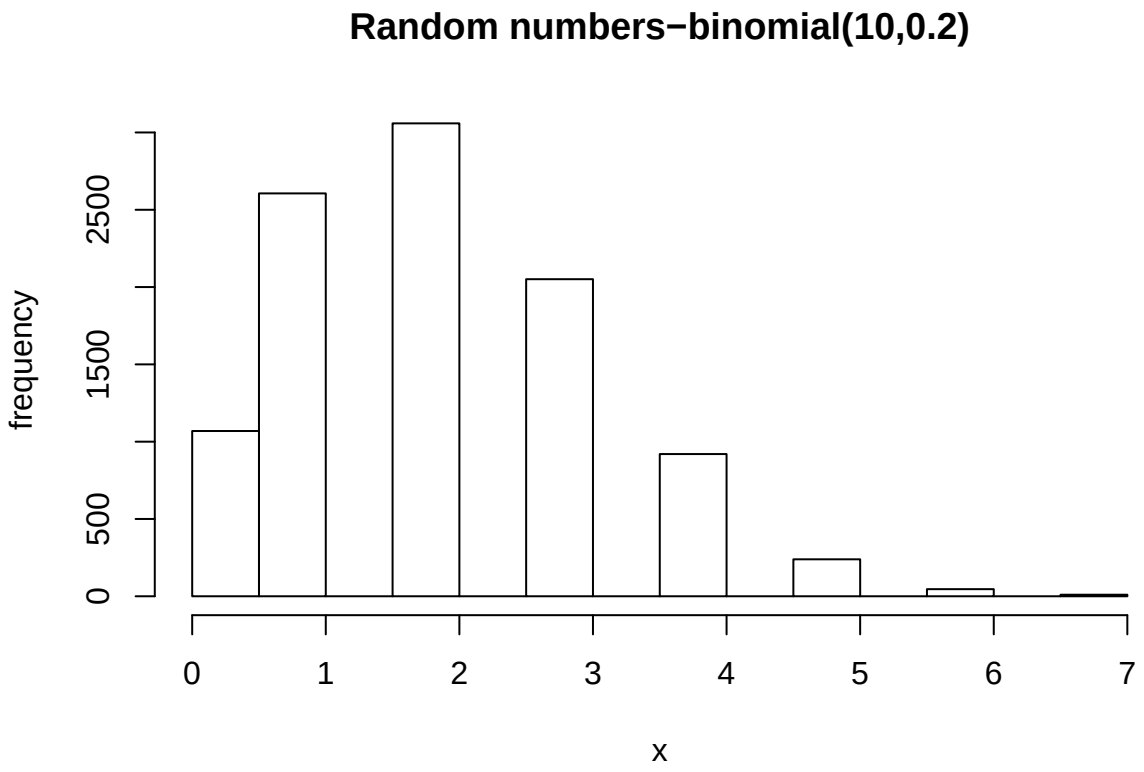

```
summary(sample2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0211  0.3285  0.5063  0.5003  0.6671  0.9789
```

With equal parameter values, the distribution should look like a bell-shaped normal distribution but with a 'fatter' bell. This is what the histogram looks like, though a little skewed to the right. More samples should smooth this out toward the center. Also, the x values are between 0 and 1, which is expected for the beta distribution.

3.3 Generate 10000 random variates from a binomial distribution with the parameters $n = 10$ and $p = 0.2$. Summarize and plot the data. (Set a seed to make it reproducible).

```
set.seed(10000)
sample3<-rbinom(10000,10,0.2)
hist(sample3,xlab="x",ylab="frequency",main="Random numbers-binomial(10,0.2)")
```




```
summary(sample3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    2.00    2.01    3.00    7.00
```

The sample mean of 2.01 is very close to the expected n and p value - 10 * 0.2 = 2, due to the high sample number (compared to the sample mean in 3.1). And as expected, the distribution is skewed to the lower numbers (because of the 0.2 parameter) and then tails off sharply at the higher values.