Assignment1

Biplav Timalsina

January 31, 2018

STAT 757 Applied Regression Analysis

---

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.

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's 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.

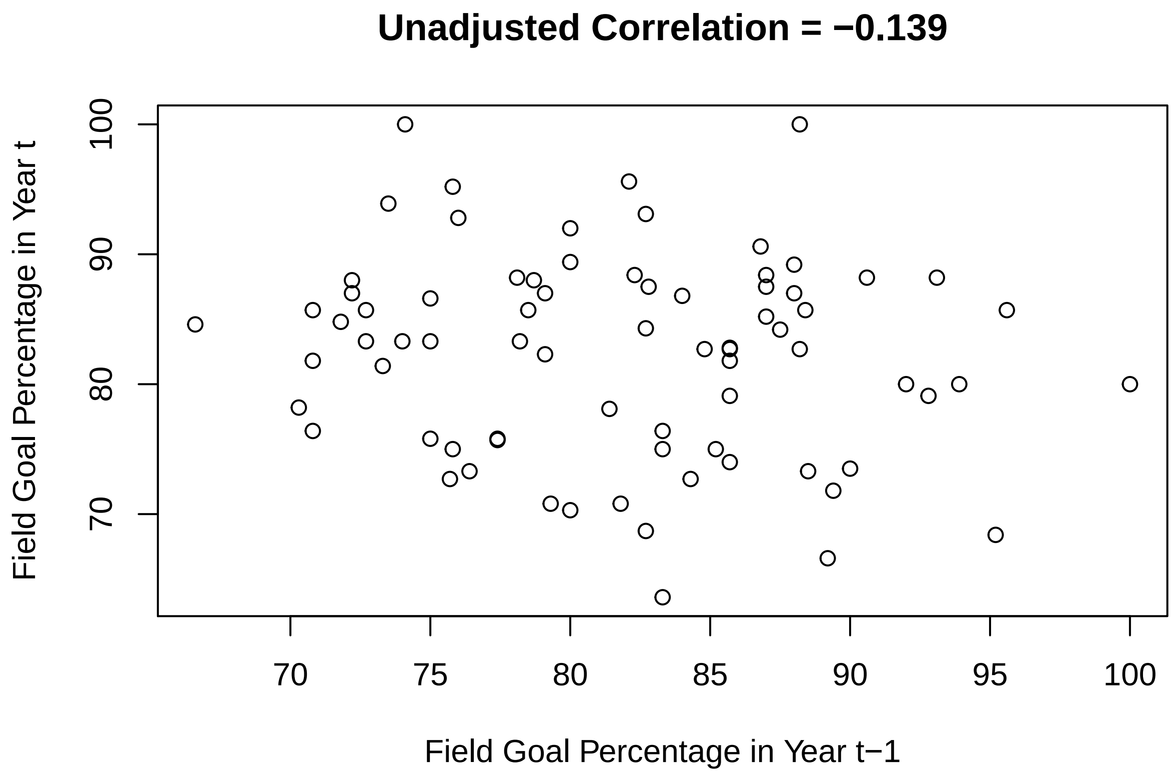
NFL Kicker Data

Part 1

```r
kicker <- read.csv("F:/unr/FieldGoals2003to2006.csv",header=T)

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")
```




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

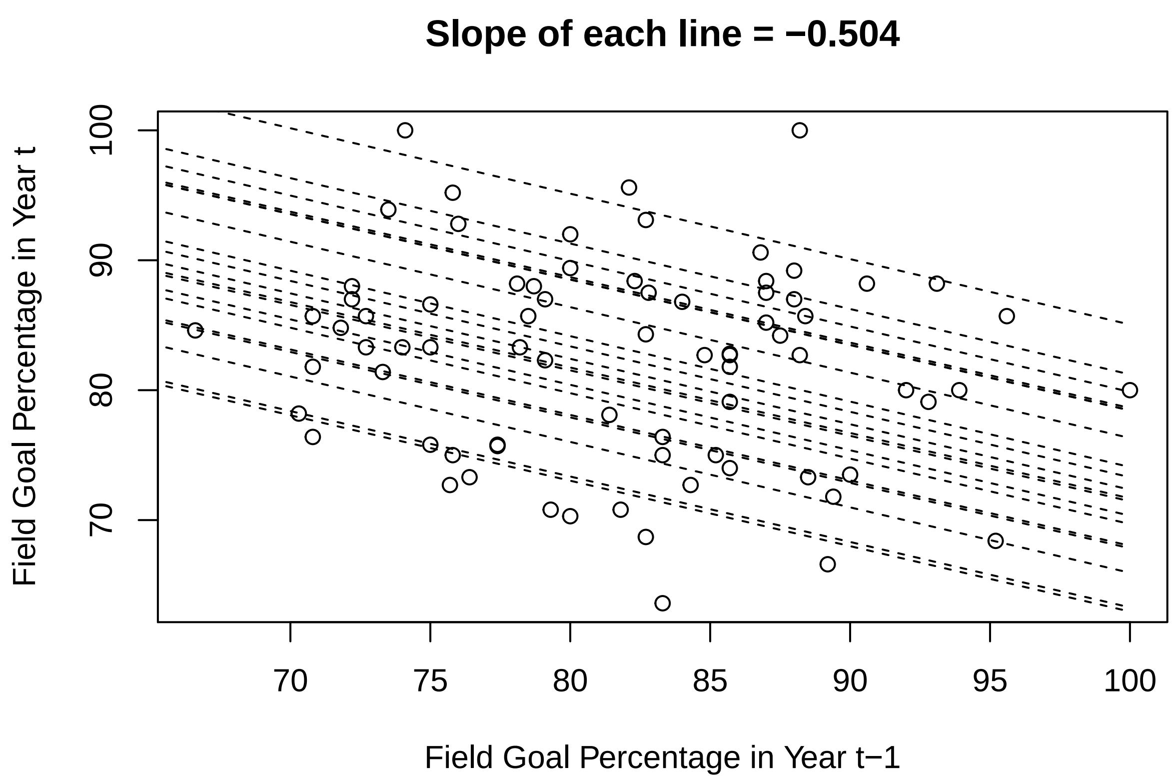
```
## Analysis of Variance Table
##
## Response: FGt
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## FGtM1        1   87.20  87.199  1.9008 0.176047
## Name        18 2252.47 125.137  2.7279 0.004565 **
## FGtM1:Name 18  417.75  23.209  0.5059 0.938592
## Residuals  38 1743.20  45.874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#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
##            126.6872              -4.6463
##      NameJason Elam     NameJason Hanson
##             -3.0167               2.1172
```

```
##            NameJay Feely             NameJeff Reed
##                -10.3737                   -8.2955
##         NameJeff Wilkins          NameJohn Carney
##                  2.3102                   -5.9774
##            NameJohn Hall            NameKris Brown
##                 -8.4865                  -13.3598
##          NameMatt Stover      NameMike Vanderjagt
##                  8.7363                    4.8955
##         NameNeil Rackers          NameOlindo Mare
##                 -6.6200                  -13.0365
##          NamePhil Dawson         NameRian Lindell
##                  3.5524                   -4.8674
##        NameRyan Longwell  NameSebastian Janikowski
##                 -2.2315                   -3.9763
##        NameShayne Graham                     FGtM1
##                  2.1350                   -0.5037
```

```
#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)}
```

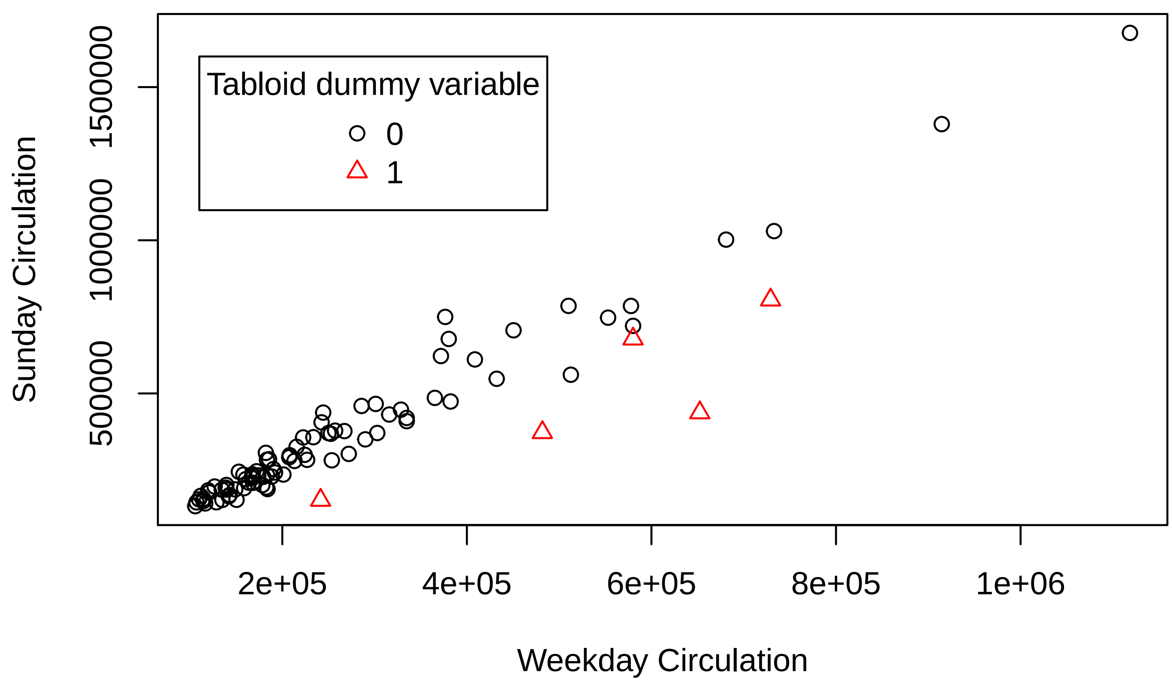


```r
detach(kicker)
```

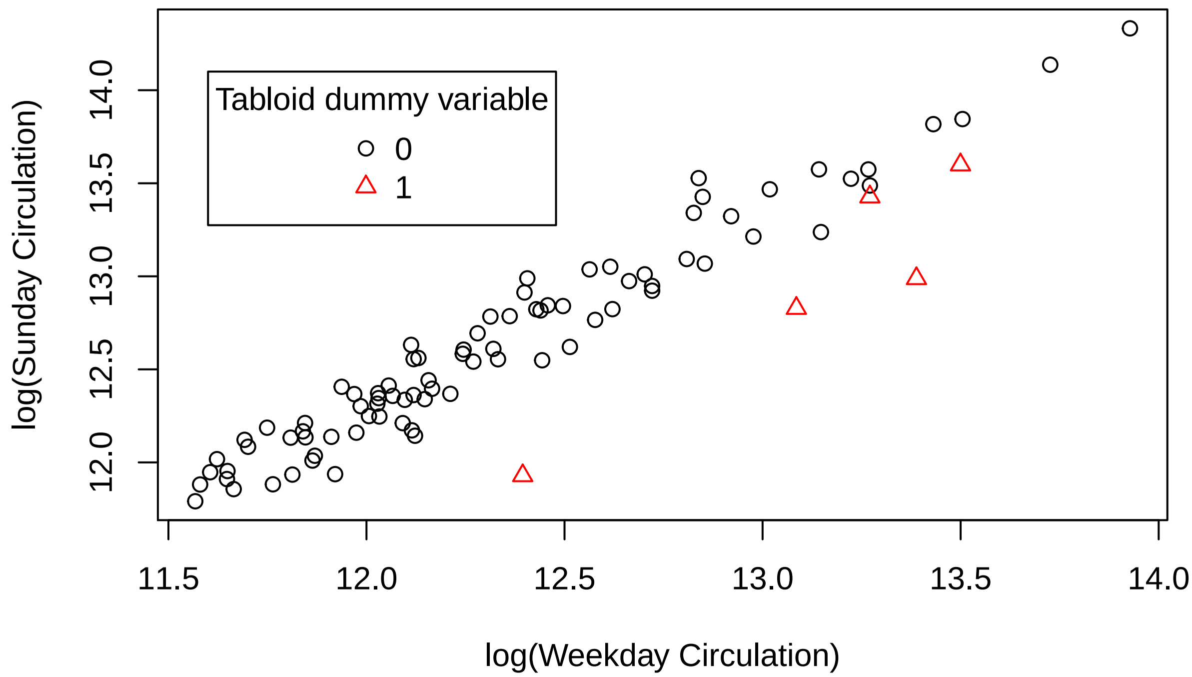
#Part 2

```r
circulation <- read.table("F:/unr/4th sem/applied regression analysis/Assignments/HW1/circulation.txt",
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")
```




```r
#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")
```



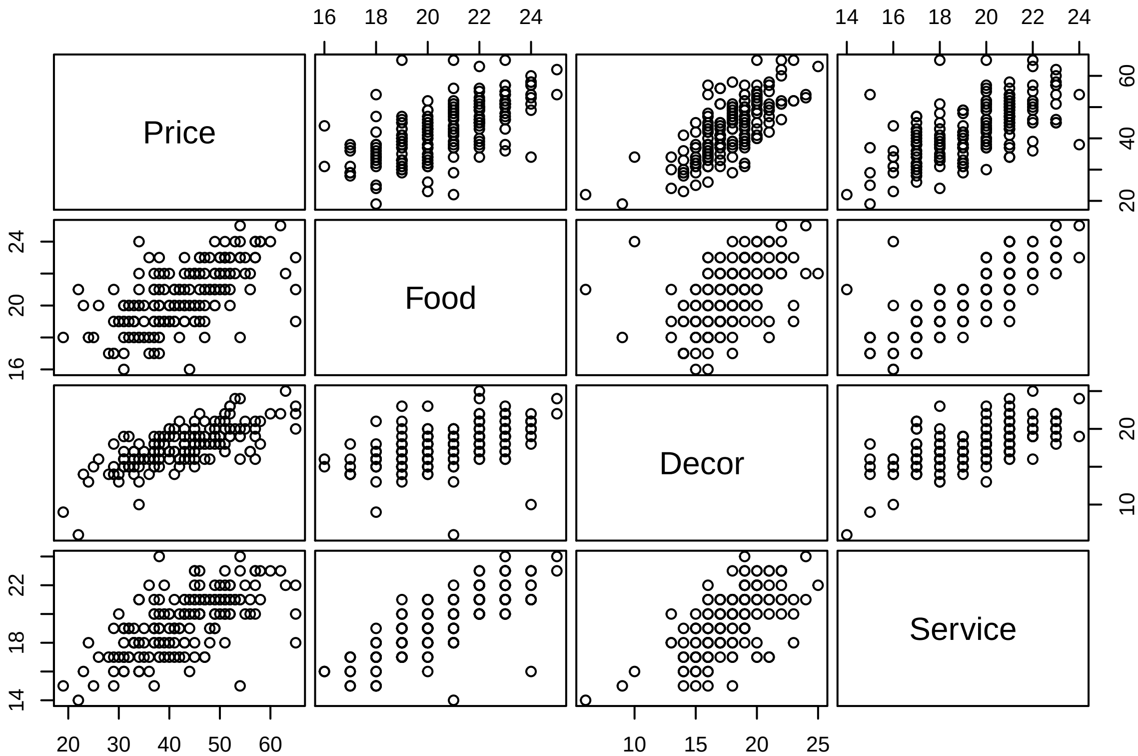

```
detach(circulation)
```

#Part 3

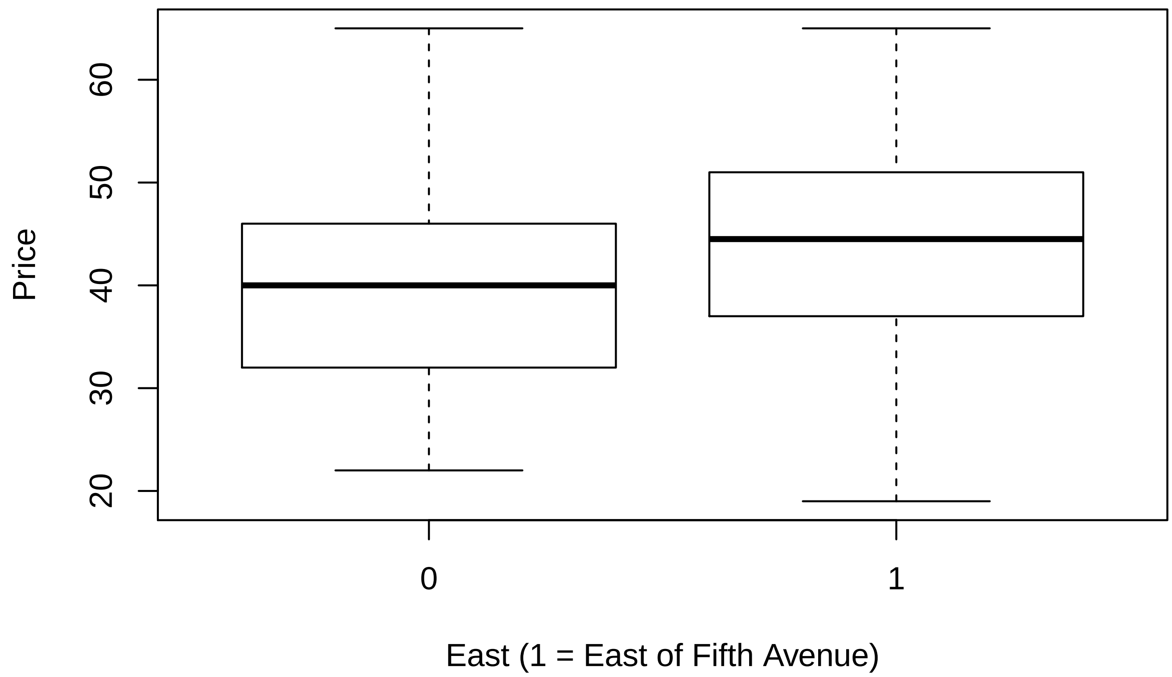
Hi All,

```
nyc <- read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW1/nyc.csv",header=TRUE)
attach(nyc)

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




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

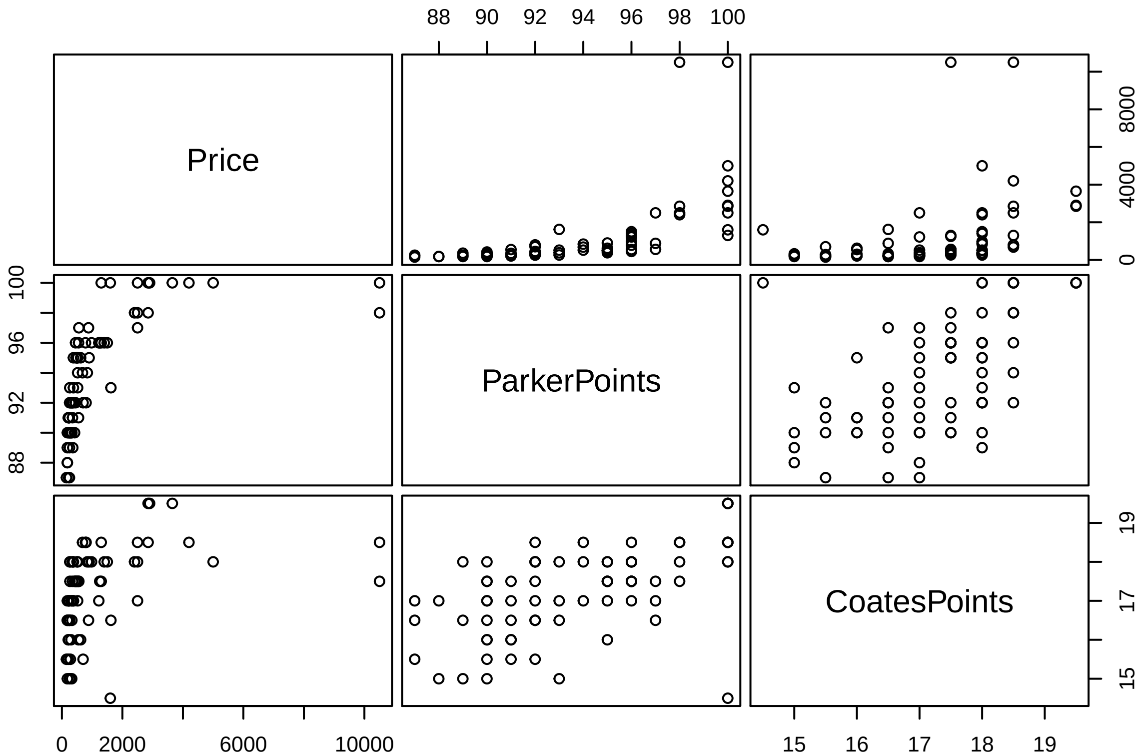



```
detach(nyc)
```

#Part 4

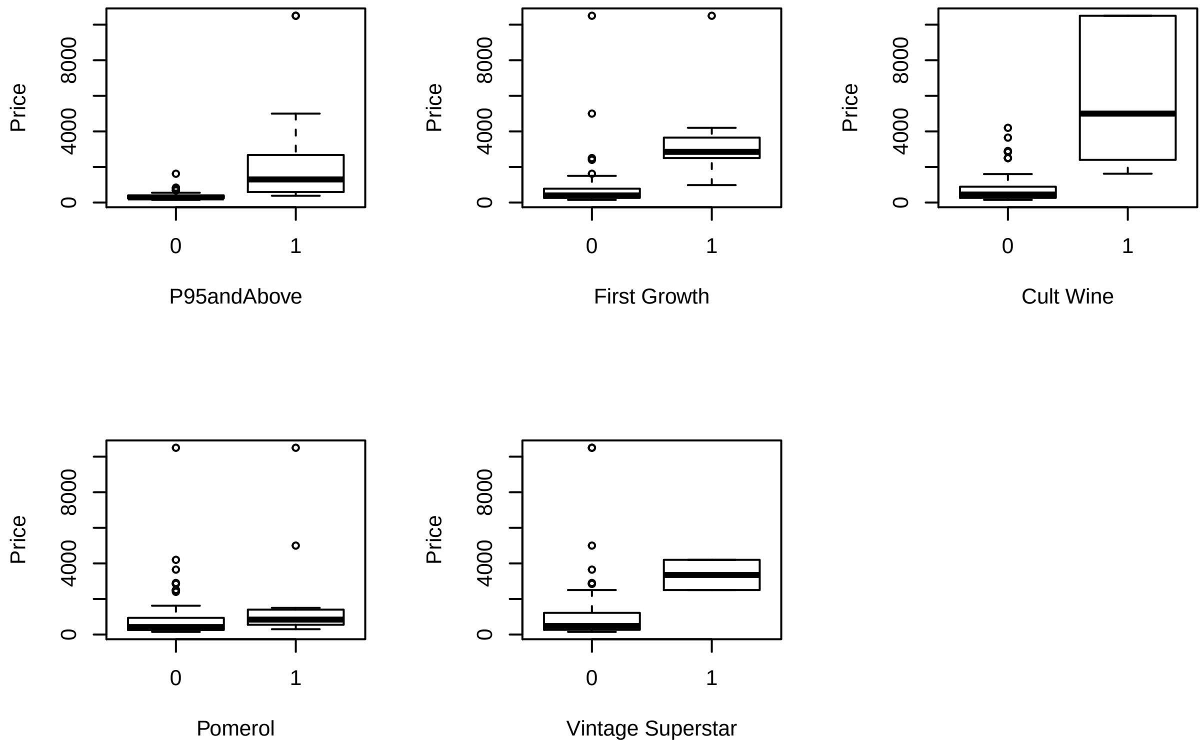
```
Bordeaux <- read.csv("F:/unr/4th sem/applied regression analysis/Assignments/HW1/Bordeaux.csv", header='
attach(Bordeaux)

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

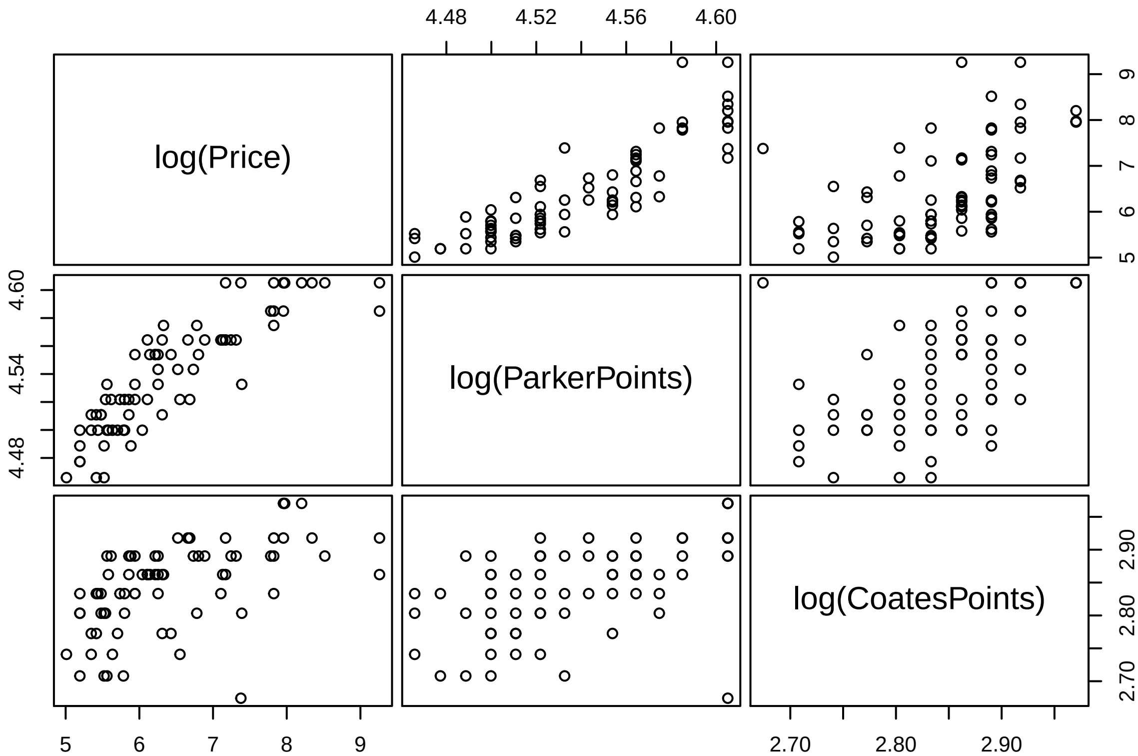



```
#Figure 1.8 on page 11
par(mfrow=c(2,3))
boxplot(Price~P95andAbove,ylab="Price",xlab="P95andAbove")
boxplot(Price~FirstGrowth,ylab="Price",xlab="First Growth")
boxplot(Price~CultWine,ylab="Price",xlab="Cult Wine")
boxplot(Price~Pomerol,ylab="Price",xlab="Pomerol")
boxplot(Price~VintageSuperstar,ylab="Price",xlab="Vintage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

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

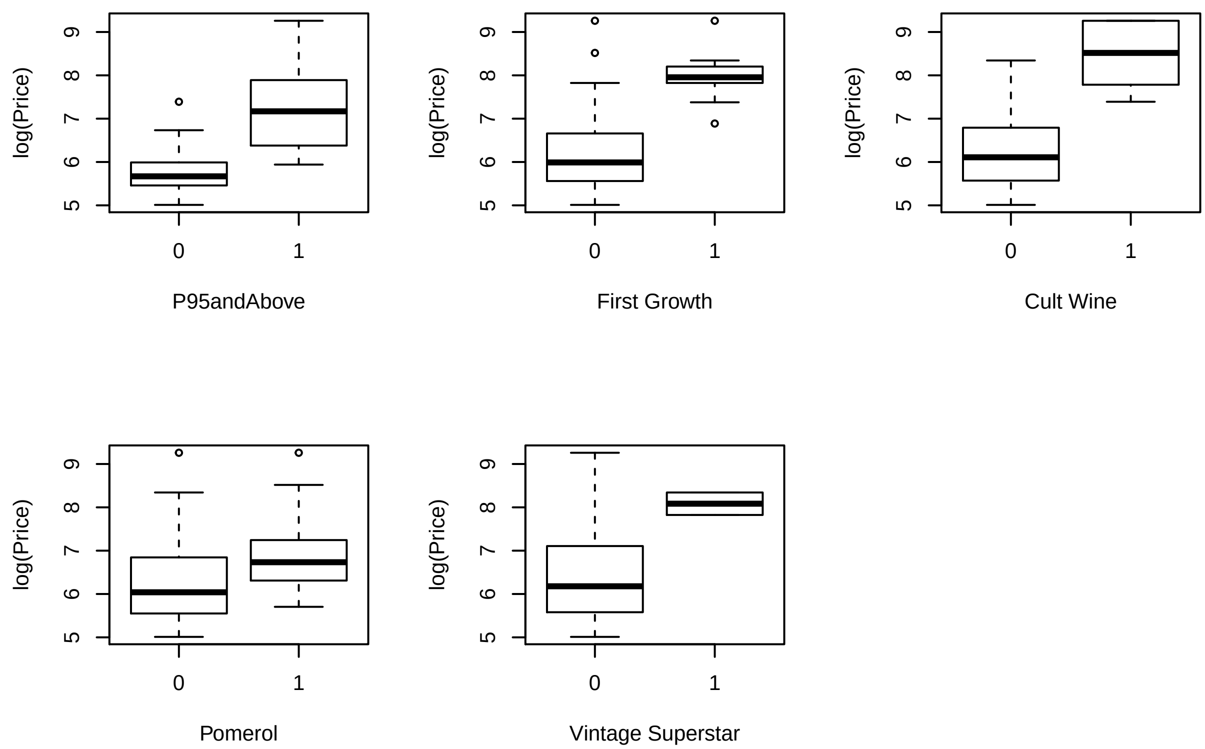



```r
pairs(log(Price)~log(ParkerPoints)+log(CoatesPoints),data=Bordeaux,gap=0.4,cex.labels=1.5)
```




```
#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)
```



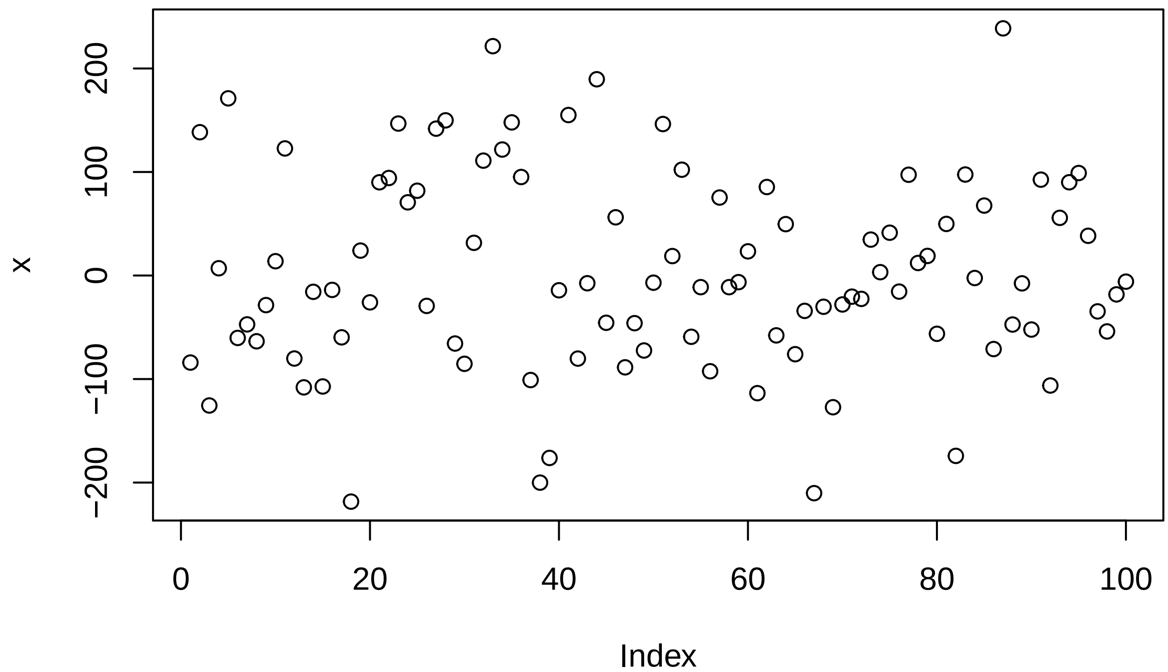
3. Generating fake data [30 points]

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(5)
x<-rnorm(100,mean=0,sd=100)
summary(x)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -218.400  -59.320   -9.399    3.164   76.970  238.700
```

```
plot(x)
```

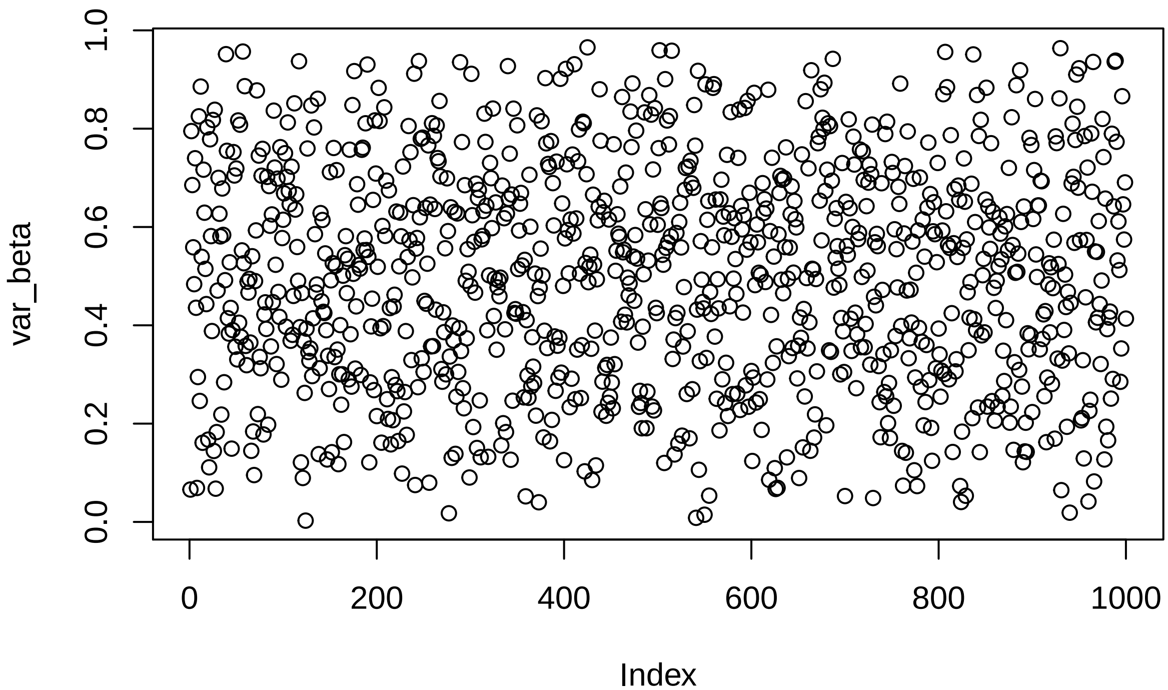


3.2 Generate 1000 random variates from a beta distribution with the parameters α and β both equal to 2. Summarize and plot the data. (Set a seed to make it reproducible).

```
set.seed=10
var_beta<-rbeta(1000,2,2)
summary(var_beta)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002704 0.327400 0.506900 0.501100 0.673200 0.965400
```

```
plot(var_beta)
```

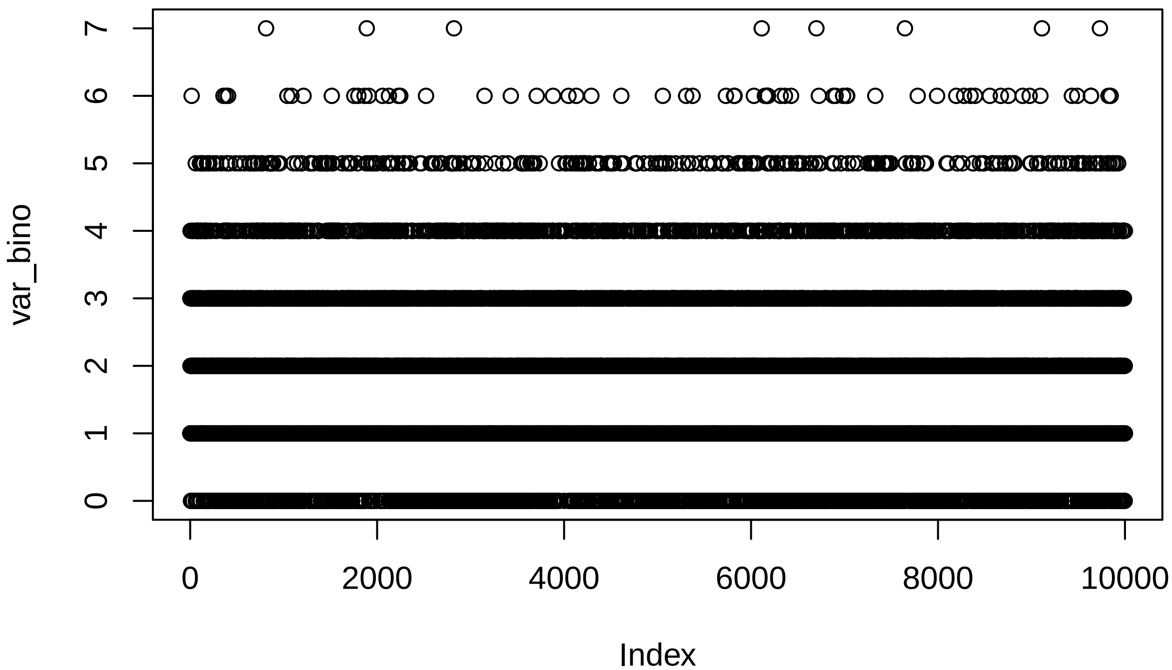


3.3 Generate 10000 random variates from a binomial distribution with the parameters $n = 10$ and $p = 0.2$. Summarize and plot the data. (Set a seed to make it reproducible).

```
set.seed=15
var_bino<-rbinom(10000,10,0.2)
summary(var_bino)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.007   3.000   7.000
```

```
plot(var_bino)
```



var_bino
Index
0
1
2
3
4
5
6
7
0
2000
4000
6000
8000
10000