HW 1

Austin Pesina

2/2/2021

#2

a) This is a regression problem because all the variables have numerical values. This is also an inference problem because we are looking at the relationship between the variables and the CEO's salary. n = top 500 U.S. firms; p = 3
b) classification; prediction; n = 20; p = 3
c) regression; prediction; n = 52 p=3

#5

A flexible approach can give a better fit for non-linear models and reduces the bias. On the other hand, it requires a greater number of parameters, it overfits the model, and there is an increase in variance.

A more flexible approach is preferred when we want a prediction while a less flexible approach is preferred when we want an inference and interpretability.

#6

A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a particular form of f and requires a large sample to accurately estimate f.

The advantages of a parametric approach over a non-parametric one is that it simplifies the modeling of f down to fewer parameters so not as many observations are needed. The disadvantages are that there is a potentially inaccurate estimate of f if the form of f is wrong; it could also overfit the observations if a more flexible model is used.

#8

```
library(ISLR)
data(College)
college <- read.csv("~/Data_Mining/Data/College.csv", header=T)

head(college[, 1:5])
```

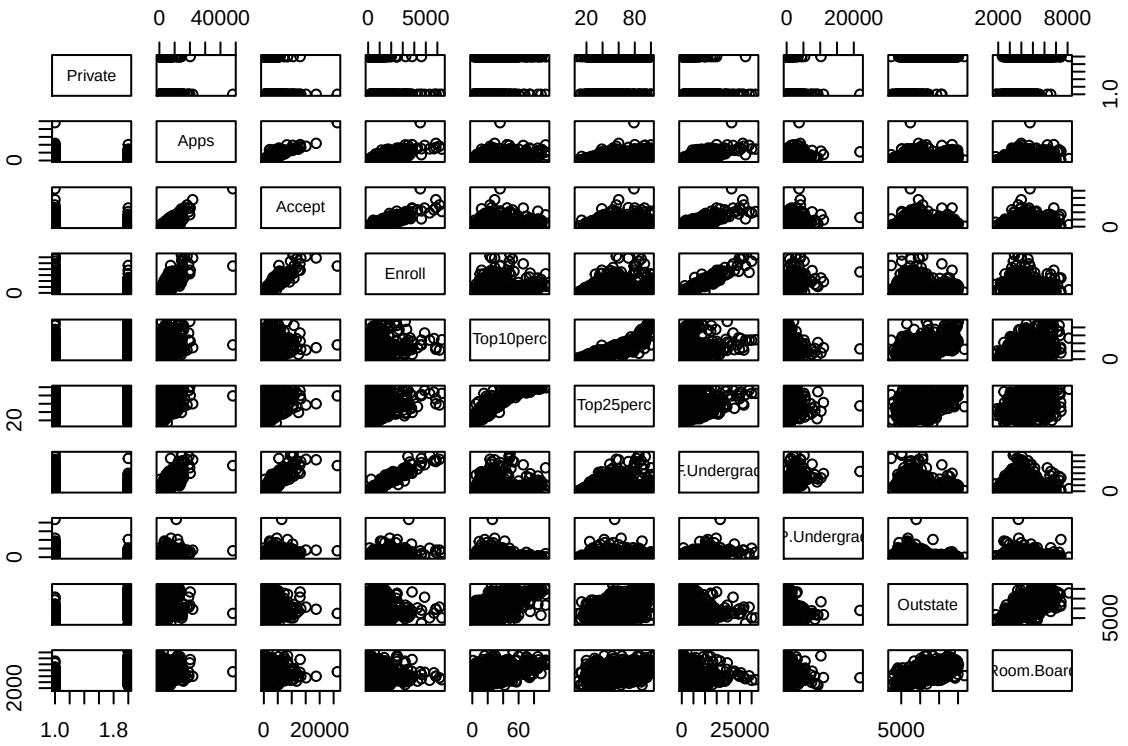
```
##                              X Private Apps Accept Enroll
## 1 Abilene Christian University     Yes 1660   1232    721
```

```
## 2           Adelphi University     Yes 2186   1924    512
## 3              Adrian College     Yes 1428   1097    336
## 4          Agnes Scott College     Yes  417    349    137
## 5    Alaska Pacific University     Yes  193    146     55
## 6            Albertson College     Yes  587    479    158
```

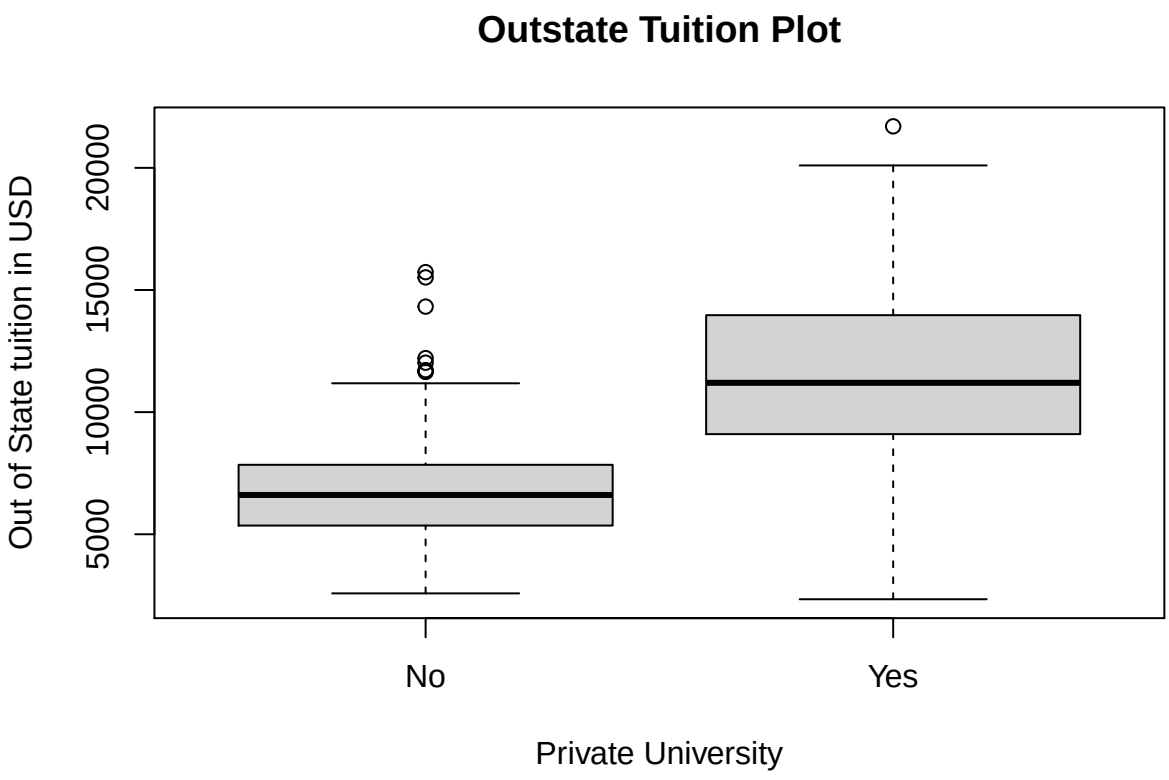
```
rownames <- college[,1]
fix(college)
college <- college[,-1]
fix(college)

college$Private<-as.factor(college$Private)

summary(college)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```
pairs(college[, 1:10])
```



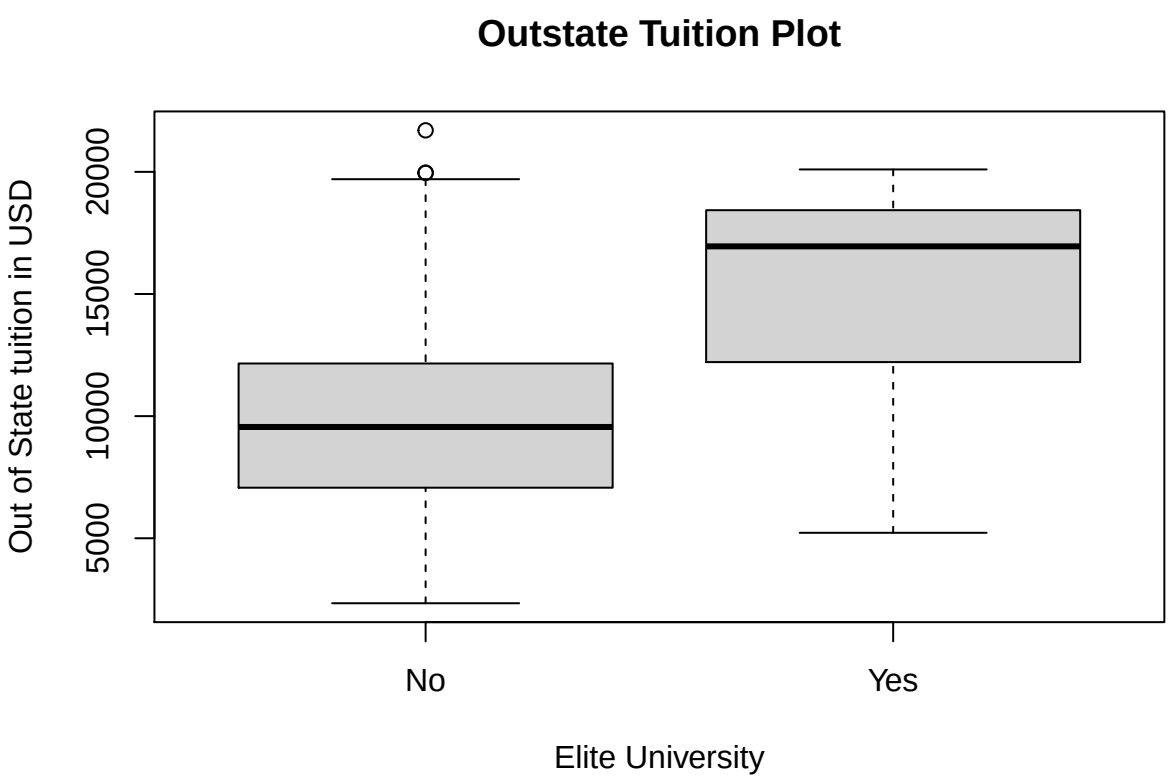

```
plot(college$Private, college$Outstate, xlab = "Private University",
     ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")
```



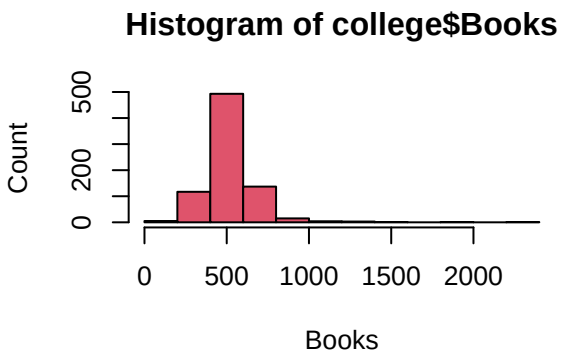

```
Elite <-rep("No", nrow(college))
Elite [college$Top10perc >50] = "Yes"
Elite <-as.factor(Elite)
college <- data.frame(college, Elite)
fix(college)
summary(college$Elite)
```

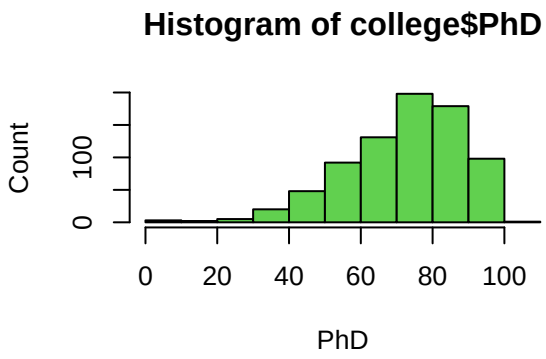
```
##  No Yes
## 699  78
```

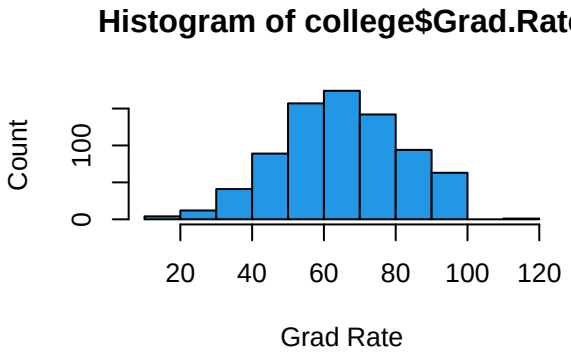
```
plot(college$Elite, college$Outstate, xlab = "Elite University",
     ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")
```

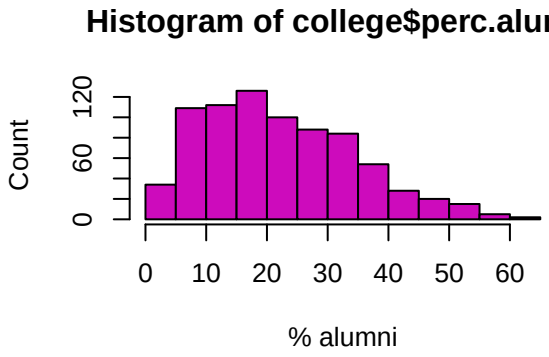



```
par(mfrow = c(2,2))
hist(college$Books, col = 2, xlab = "Books", ylab = "Count")
hist(college$PhD, col = 3, xlab = "PhD", ylab = "Count")
hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count")
hist(college$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count")
```










```
summary(college$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.