hw1

Question 2

a. Classification Prediction n=200 p =4
b. Regression Prediction n= 20 p=3
c. Regression Inference n= 600 p=5

Question 3

a. Political Science: During an election,whether a person casts his vote or not. We can sdetermine if he casts vote or not and calculate approximately the total number of votes to be casted in the elections.

   Response: Vote

   Predictors: Sex, Age, Income, Education Goal = Prediction

   Sports: Suppose we take a cricket game between Team A and Team B for analysis. This is an example of classification of whether Team A would win the game or Team B. The result depends on different factors such as the location, players in the team, team's history, etc.

   Response: Team A or Team B

   Predictors: Players, Location of Match, Climate on that day, History of the team's winning in that location.

   Goal: Prediction.

   Area of my choice: To decide the stadium for a particular sport during Olympics.

   Response: Stadium name

   Predictors: Type of Game, Number of Competitors, Number of sub categories, Previous year population attending the game, Previous year spending; Profit/Loss analysis for the past Olympic games etc., Stadium Size, Population size

   Goal: Inference

b. Agriculture: Relationship between water use and amount of fertilizer/type

   Predictors: year, amount of water used for a particular year,amount of fertilizer used/type for that year

   Goal = Inference

   Business= effect of currency value on company's share/stock value or company's profits.

Predictors: currency value over the years, company's share value/profit over the years, no. of employees, avg. salary range etc.)

Goal = Prediction

Area of my choice: In real estate, to predict the price of the house with its relationship to the square footage.

Response: House sale price

Predictors: Square feet

Goal: Prediction.

c. Education: Which student classification needs more attention. Depending on the research of charateristics of the students, they are clustered into groups.

Response: Groups(High scorer, Medium scorer, Low scorer)

Predictors: psychological situation, environment, aptitude, attitude

Goal= Inference.

Meteorology: Ozone level prediction in various states of the country

Response: Ozone level

Predictors: weekdays ozone level, weekends ozone level, type of industries, number of vehicles bought in that year, number of vehicles registered in that state etc.

Goal= Inference

Area of my choice: To determine whether the people in a particular community are store loyal or brand loyal.

Response: Store or Brand loyal

Predictors: Age, Sex, Income, Location, etc.

Goal= Inference

Question 4

a. sqrt [(X1test- X1)^2 + (X2test-X2)^2 + (X3test-X3)^2]

```
train <- data.frame(Obs = c(1:6), X1 = c(0,2,0,-1,-3,1), X2 = c(4,0,1,1,0,0),
X3 = c(0,1,4,2,1,1), Y = c("Green", "Red", "Red", "Green", "Green", "Red"))

test <- data.frame(train, x1test = rep(0, 6), x2test = rep(0, 6), x3test = re
p(0, 6))

test$EuclideanDistance <- sqrt((test$x1test - test$X1)^2 + (test$x2test - tes
```

```
t$X2)^2 + (test$x3test - test$X3)^2)
test
```

```
##   Obs X1 X2 X3     Y x1test x2test x3test EuclideanDistance
## 1   1  0  4  0 Green      0      0      0          4.000000
## 2   2  2  0  1   Red      0      0      0          2.236068
## 3   3  0  1  4   Red      0      0      0          4.123106
## 4   4 -1  1  2 Green      0      0      0          2.449490
## 5   5 -3  0  1 Green      0      0      0          3.162278
## 6   6  1  0  1   Red      0      0      0          1.414214
```

b. If K = 1, the closest point to it is Observation 6, hence our prediction=Red

c. If K=3, the three points tell red, green and red. Hence our prediction is red.

d. For this question, since the data is small we need a high k value because we do not want to catch any noise.

Question 5a

```
college <- read.csv("College.csv")
```

Question 5b

```
#View(college)
rownames(college) <- college[,1]
college <- college[,-1]
head(college[,1:5])
```

```
##                                Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
## Albertson College                Yes  587    479    158        38
```

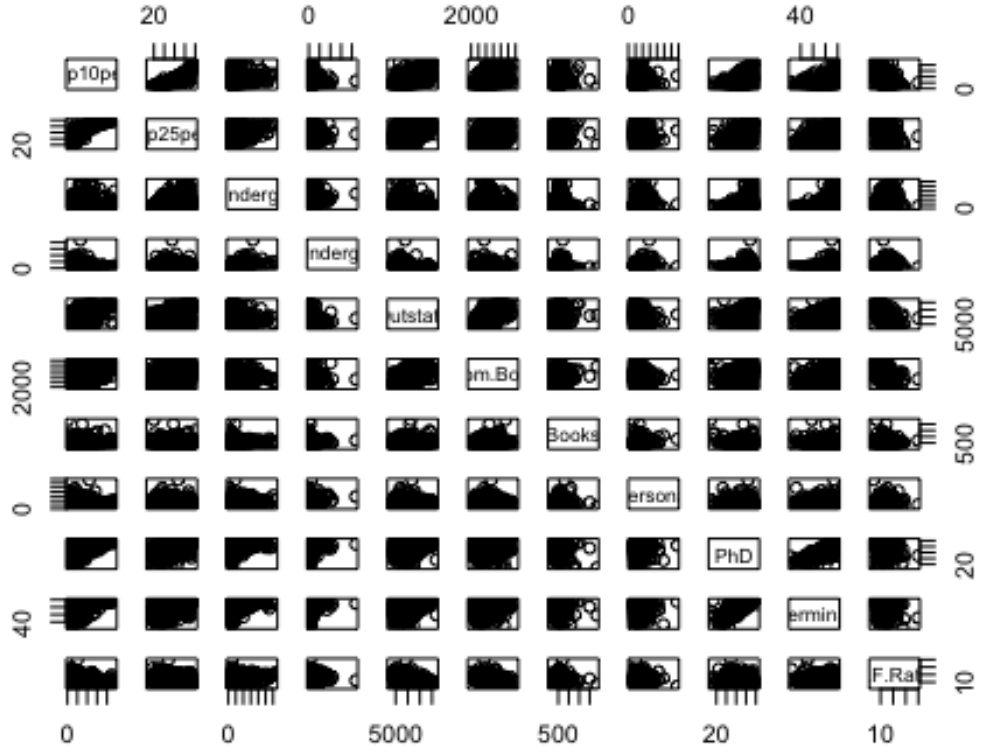
Question 5c

```
summary(college)
```

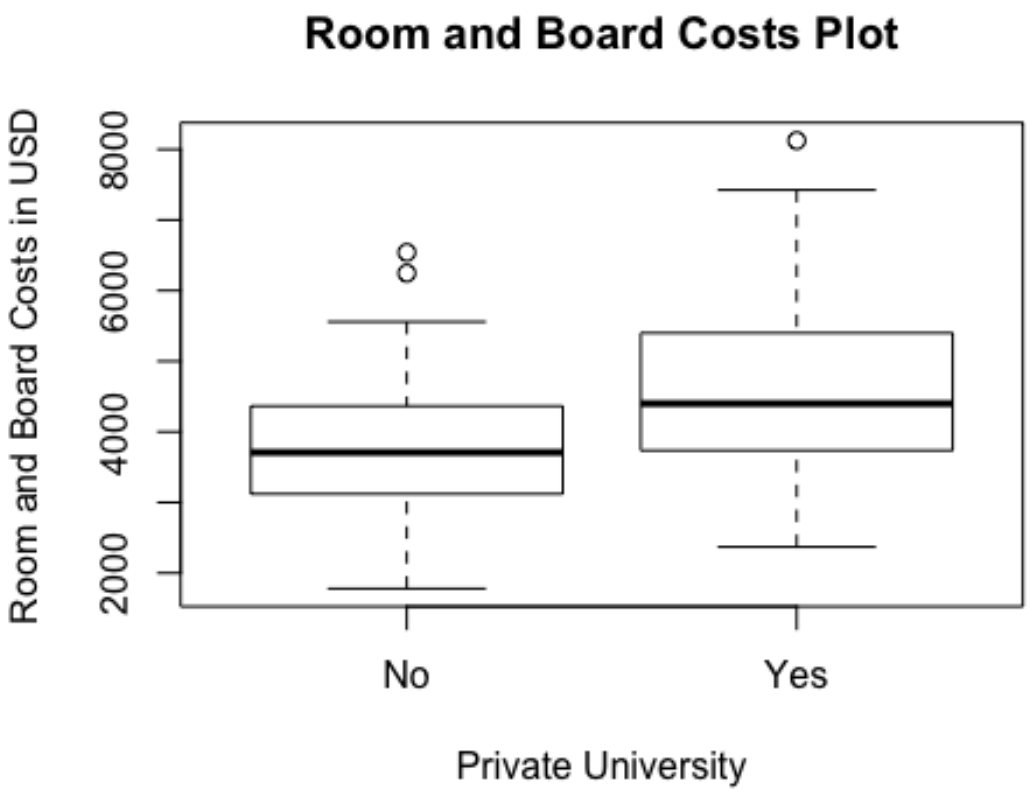
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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
```

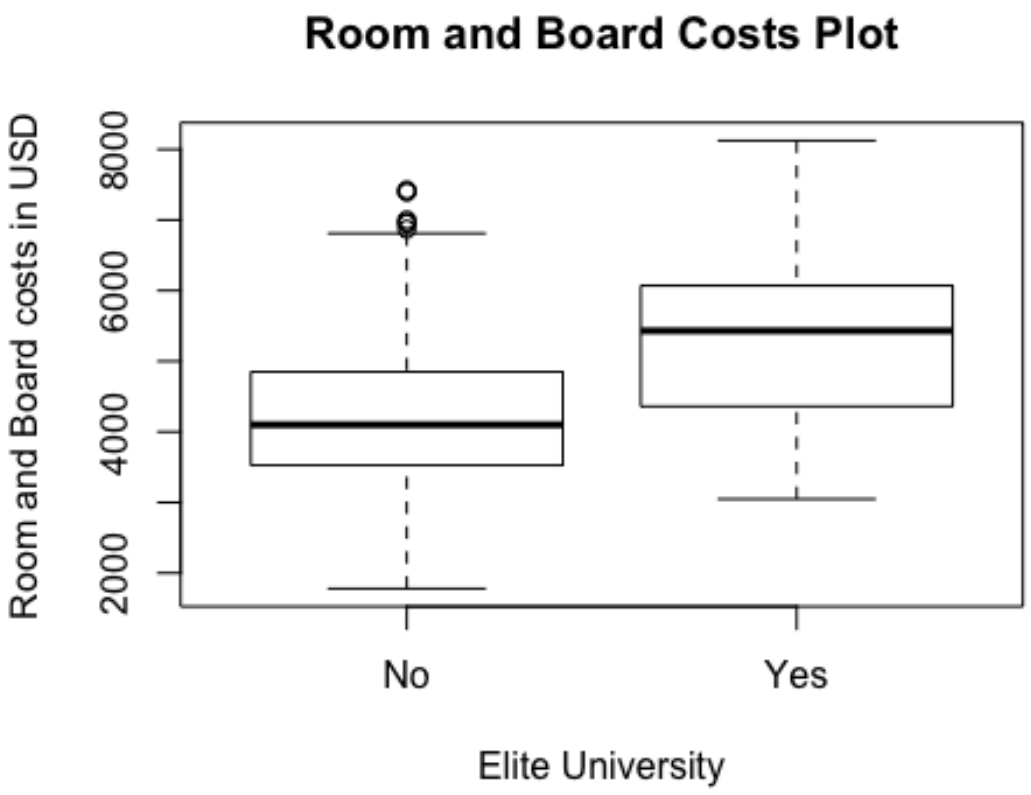
```
pairs(college[, 5:15])
```



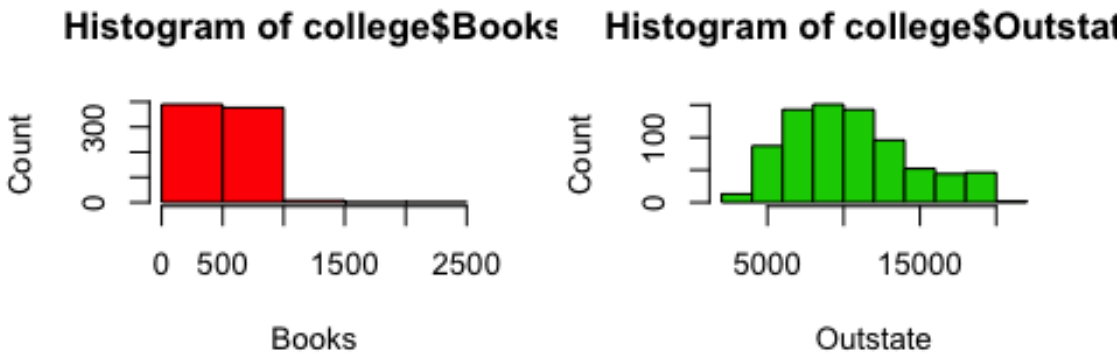

```
plot(college$Private, college$Room.Board, xlab = "Private University", ylab =
"Room and Board Costs in USD", main = "Room and Board Costs Plot")
```

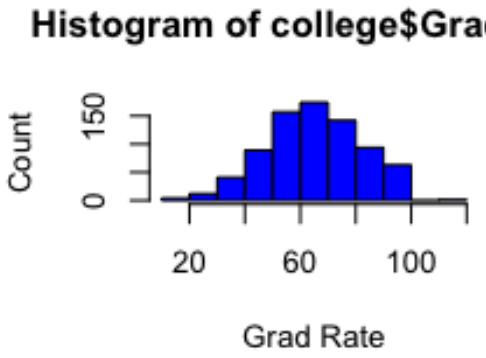


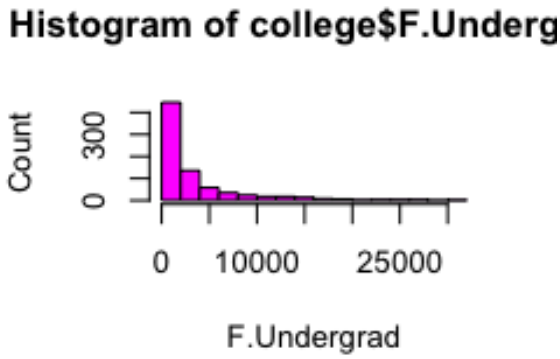

```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college=data.frame(college, Elite)
summary(college$Elite)

##  No Yes
## 699  78

plot(college$Elite, college$Room.Board, xlab = "Elite University", ylab ="Roo
m and Board costs in USD", main = "Room and Board Costs Plot")
```




```
par(mfrow = c(2,2))
hist(college$Books, col = 2, xlab = "Books", ylab = "Count", breaks=5)
hist(college$Outstate, col = 3, xlab = "Outstate", ylab = "Count", breaks=10)
hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count", breaks=1
5)
hist(college$F.Undergrad, col = 6, xlab = "F.Undergrad", ylab = "Count", brea
ks=20)
```







Question 6a

```
library(ISLR)

Auto <- na.omit(Auto)
```

"Name"" is qualitative and rest of the predictors are quantitative.

```
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
```

```
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

We observe that "origin" variable takes only values of 1,2,3 and probably should be a factor.

```
Auto$origin <- as.factor(Auto$origin)
```

To understand the numeric or quantitative variables

```
quant <- sapply(Auto, is.numeric)
quant
```

```
##          mpg    cylinders displacement   horsepower       weight
##         TRUE         TRUE         TRUE         TRUE         TRUE
## acceleration         year       origin         name
##         TRUE         TRUE        FALSE        FALSE
```

All variables except origin and name are quantitative.

Question 6b

```
sapply(Auto[, c(1,2,3,4,5,6,7)], range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
```

Question 6c

```
sapply(Auto[, quant], function(x) c(mean(x), sd(x)))
```

```
##            mpg cylinders displacement horsepower    weight acceleration
## [1,] 23.445918  5.471939      194.412  104.46939 2977.5842    15.541327
## [2,]  7.805007  1.705783      104.644   38.49116  849.4026     2.758864
##           year
## [1,] 75.979592
## [2,]  3.683737
```

Question 6d

```
sapply(Auto[1:(nrow(Auto)-50), quant], range)
```

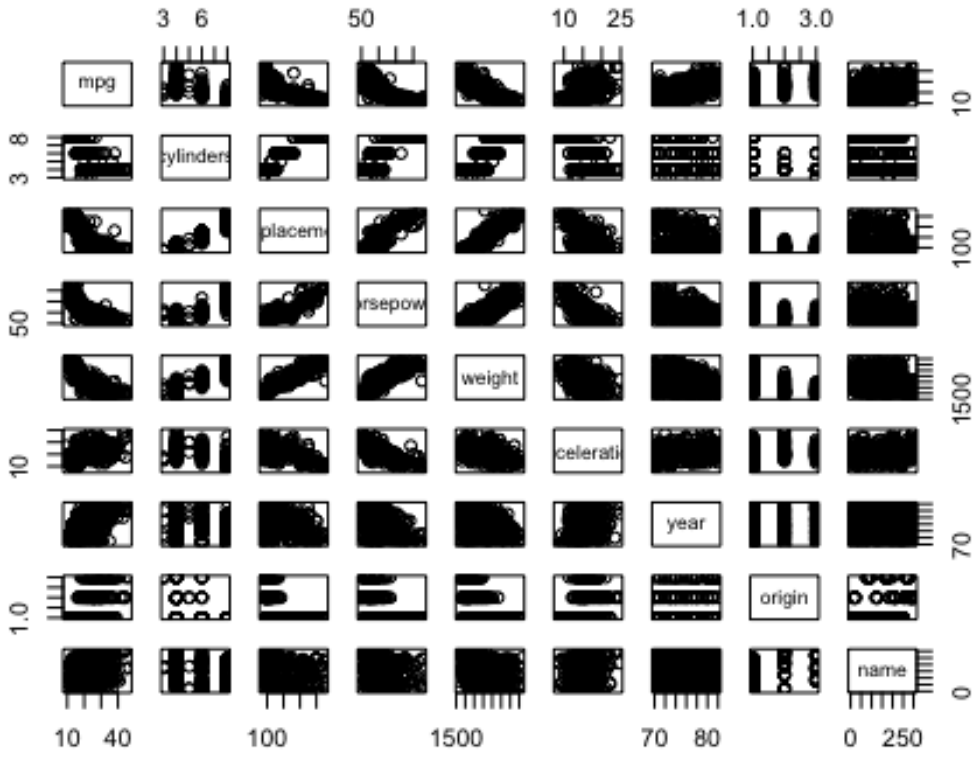
```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   81
```

```
sapply(Auto[1:(nrow(Auto)-50), quant], function(x) c( mean(x), sd(x)))
```

```
##            mpg cylinders displacement horsepower    weight acceleration
## [1,] 22.315497  5.622807     203.3260  107.82164 3045.5789     15.38363
## [2,]  7.456385  1.741626     107.4587   39.69976  872.8565      2.78170
##           year
## [1,] 75.157895
## [2,]  3.196164
```

Question 6e

```
pairs(Auto)
```



We can see that Displacement, weight and horsepower, weight are highly positively correlated. While horsepower, acceleration and mpg, weight are negatively correlated.

Question 6f

From the pairs plot, we can see that mpg is correlated to displacement, horsepower, weight,year.

```
cor(Auto$mpg, Auto$horsepower)
```

```
## [1] -0.7784268
```

```
cor(Auto$mpg, Auto$displacement)
```

```
## [1] -0.8051269
```

```
cor(Auto$mpg, Auto$weight)
```

```
## [1] -0.8322442
```

```
cor(Auto$mpg, Auto$year)
```

```
## [1] 0.580541
```

From the correlation factors, it is understood that mpg is negatively correlated to horsepower, displacement and weight, and is positively correlated to the year.