HW1

Ramtin Boustani - SUID# 05999261

Problem 1

Exercise 1

(a)

flexible is better
With high observation and low number of variables we can have a good estimate to real function

(b)

inflexible is better
If use flexible model with large number of predicators we can get in trap of overfitting issue.

(c)

flexible is better
because true f is very non-linear be more flexible can estimate real function better

(d)

inflexible is better
Becasue of hight irreducible error we have more high noise and unstability in the system so it is better to be inflexible to be immune of those errors. Due to high noise and error we will have different off $\hat{f}$ as result we have higher $var(\hat{f}x_0)$

Problem 2

Exercise 2

(a)

Regression - Inference
n=500 (top firms)
p=4 (profit, #employee, industry, salary)

(b)

Classification - Prediction
n=20 (similar products)
p=13 (price charged, marketing budget, competition price, 10 other)

(c)

Regression - Prediction
n=52 (number of weeks for 2012)
p=3 (% change in US market, % change in British market, % change in German market)

Problem 3

Exercise 4

(a)

- Classification
  - Passing/Failing exam
    Output could be one of Pass/Fail and inputs are students homeworks and exams grades with final passing or failing label. Which exams and homeworks have the most result in the final exam failing (inference). Predicting final exam result (Prediction).
  - likes or dislikes Tweet
    Output could be one of like/dislike and inputs are similar tweets with label. (Prediction)
  - Text sentiment classification.
    Output could be one of positive/negative/neutral and inputs are labeled words with sentiment. (Prediction)

(b)

- Regression
  - Estimate property value similar to Zillow (Prediction)
  - Estimate stock price using related financial information (Prediction)
  - Effects of chocolate on blood pressure (inference)

(c)

- Clustering
  - Image segmentation
  - Breast cancer cell clustering
  - Handwriting character recognition

Problem 4

Exercise 5

- Advantage of Very Flexible
  - Good fit
  - Decreasing $Bias(\ \hat{f}(x_0))^2$
- Disadvantage of Very Fleible
  - Overfitting
  - Increasing $var(\hat{f}(x_0)$

Very Flexible is good for Prediction
Less Flexible is good for Inference

Problem 5

Exercise 8

(a)

```
install.packages("ISLR", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/nm/g09p009s0qb231f3_hsjz7r40011sl/T//RtmpCEfL3Q/downloaded_packages
```

```
library(ISLR)
college_df = ISLR::College
head(college_df[,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
```

(b)

```
head(rownames(college_df))
```

```
## [1] "Abilene Christian University" "Adelphi University"
## [3] "Adrian College"               "Agnes Scott College"
## [5] "Alaska Pacific University"    "Albertson College"
```

```
head(college_df[,1])
```

```
## [1] Yes Yes Yes Yes Yes Yes
## Levels: No Yes
```

(c)

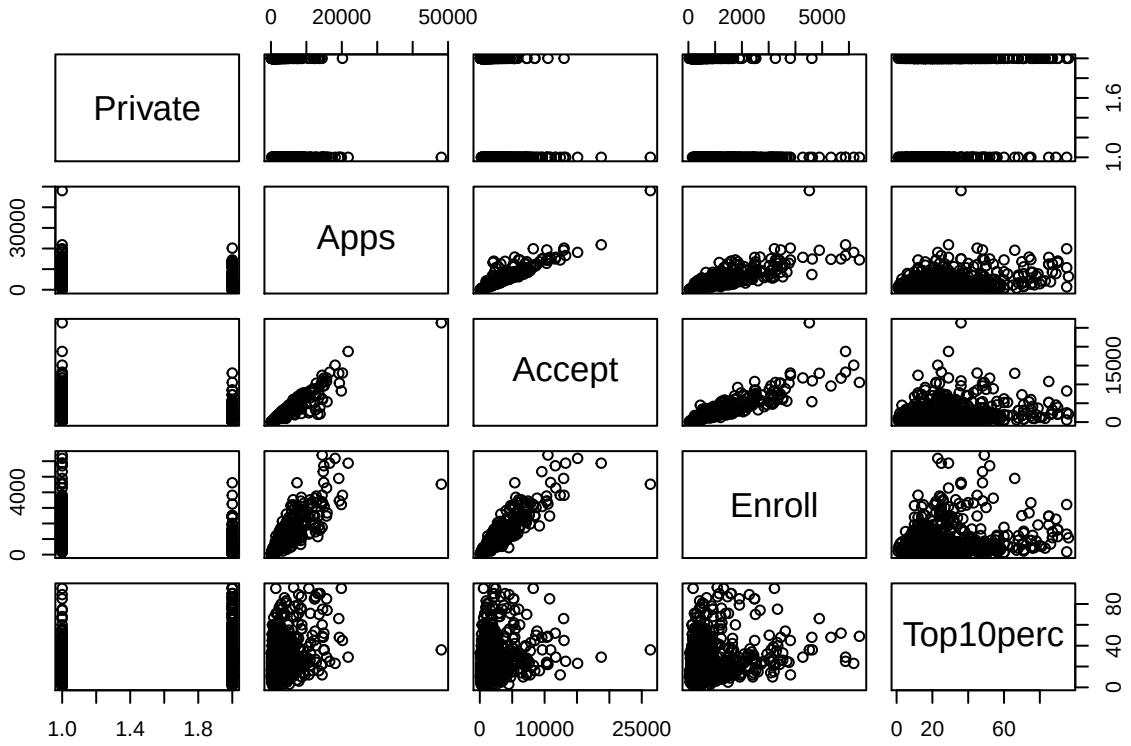
i.

```
summary(college_df[,1:5])
```

```
##  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
```

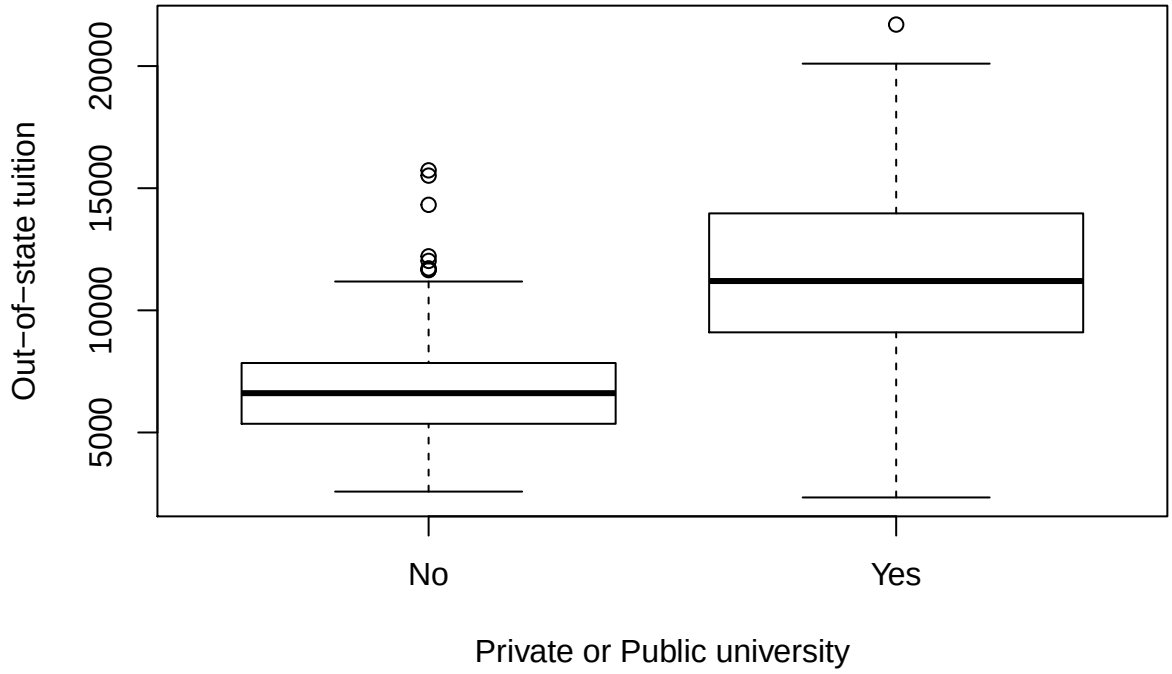
ii.

```
pairs(college_df[,1:5])
```



iii.

```
plot(college_df$Private, college_df$Outstate, xlab="Private or Public university", ylab="Out-of-state tu
```



.iv

```
Elite = rep("No", nrow(college_df))
Elite[college_df$Top10perc>50]="Yes"
```

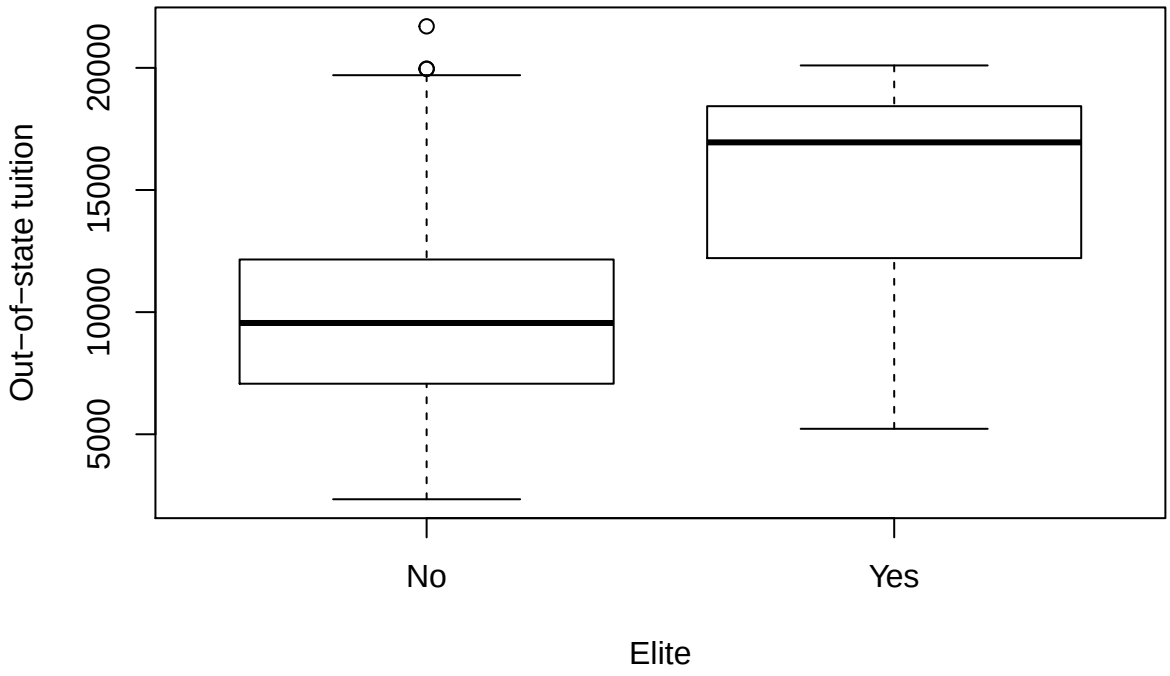
```
Elite=as.factor(Elite)
college_df = data.frame(college_df, Elite)
show(college_df[1:5,c(1,ncol(college_df))])
```

```
##                              Private Elite
## Abilene Christian University     Yes    No
## Adelphi University               Yes    No
## Adrian College                   Yes    No
## Agnes Scott College              Yes   Yes
## Alaska Pacific University        Yes    No
```

```
summary(college_df$Elite)
```

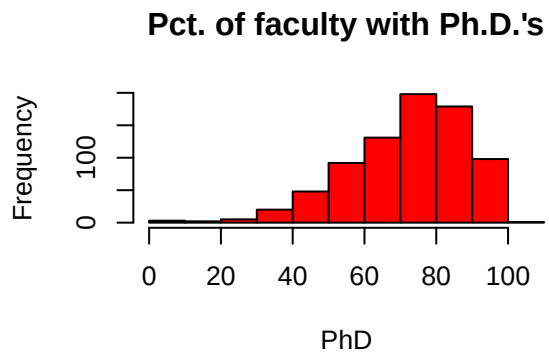
```
##  No Yes
## 699  78
```

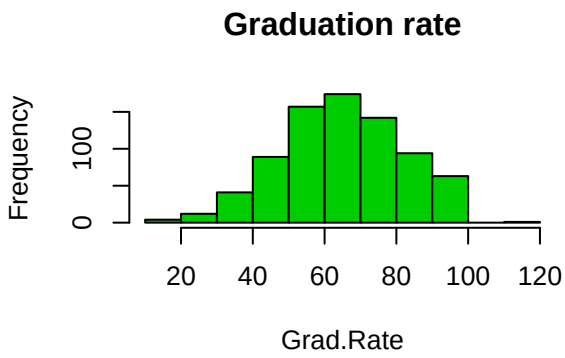
```
plot(college_df$Elite, college_df$Outstate, xlab="Elite", ylab="Out-of-state tuition")
```

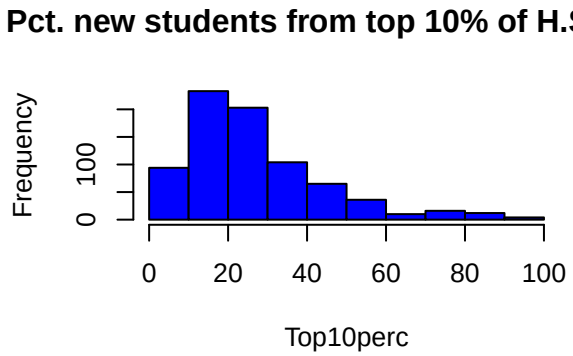


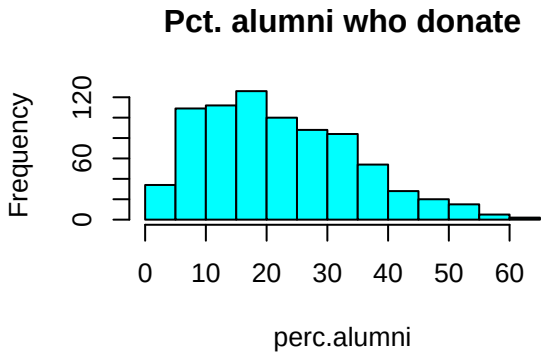
v.

```
par(mfrow=c(2,2))
hist(college_df$PhD, col = 2, xlab = "PhD", main = "Pct. of faculty with Ph.D.'s")
hist(college_df$Grad.Rate, col = 3, xlab = "Grad.Rate", main = "Graduation rate")
hist(college_df$Top10perc, col = 4, xlab = "Top10perc", main = "Pct. new students from top 10% of H.S. c
hist(college_df$perc.alumni, col = 5, xlab = "perc.alumni", main = "Pct. alumni who donate")
```



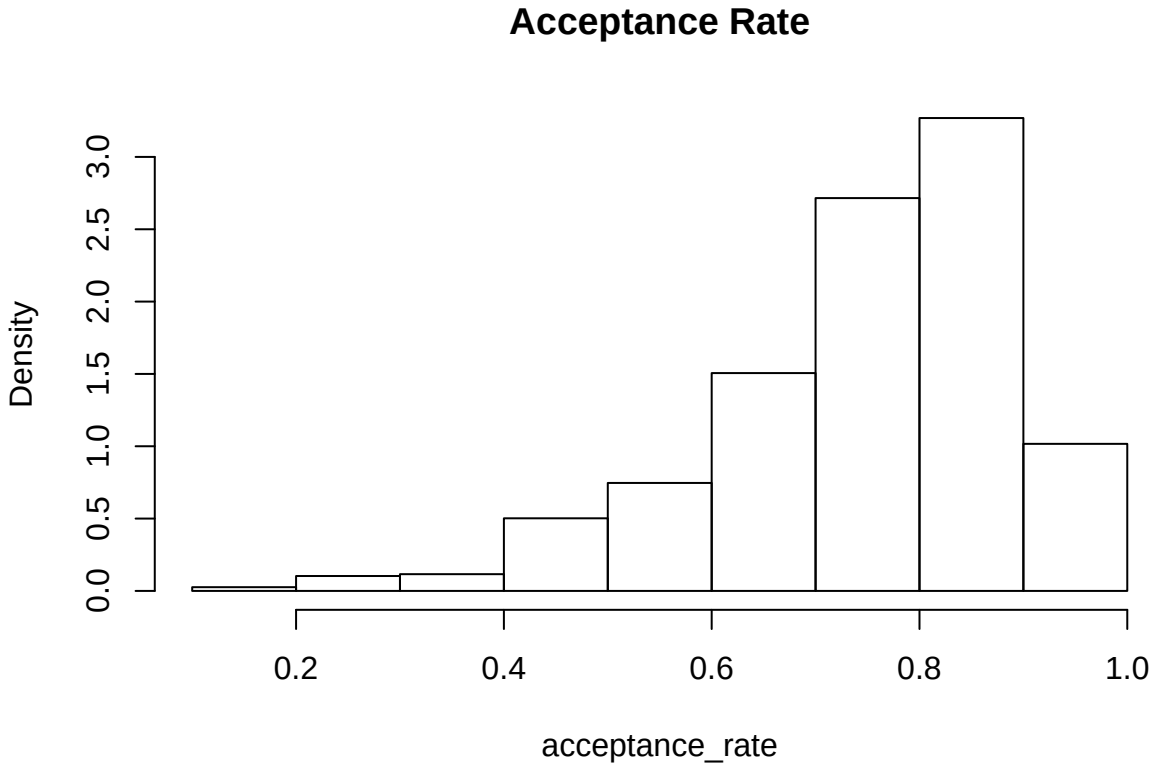






vi.

```
acceptance_rate = college_df$Accept/college_df$Apps
hist(acceptance_rate, main = "Acceptance Rate" , probability = TRUE)
```



```r
summary(acceptance_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1545  0.6756  0.7788  0.7469  0.8485  1.0000
```