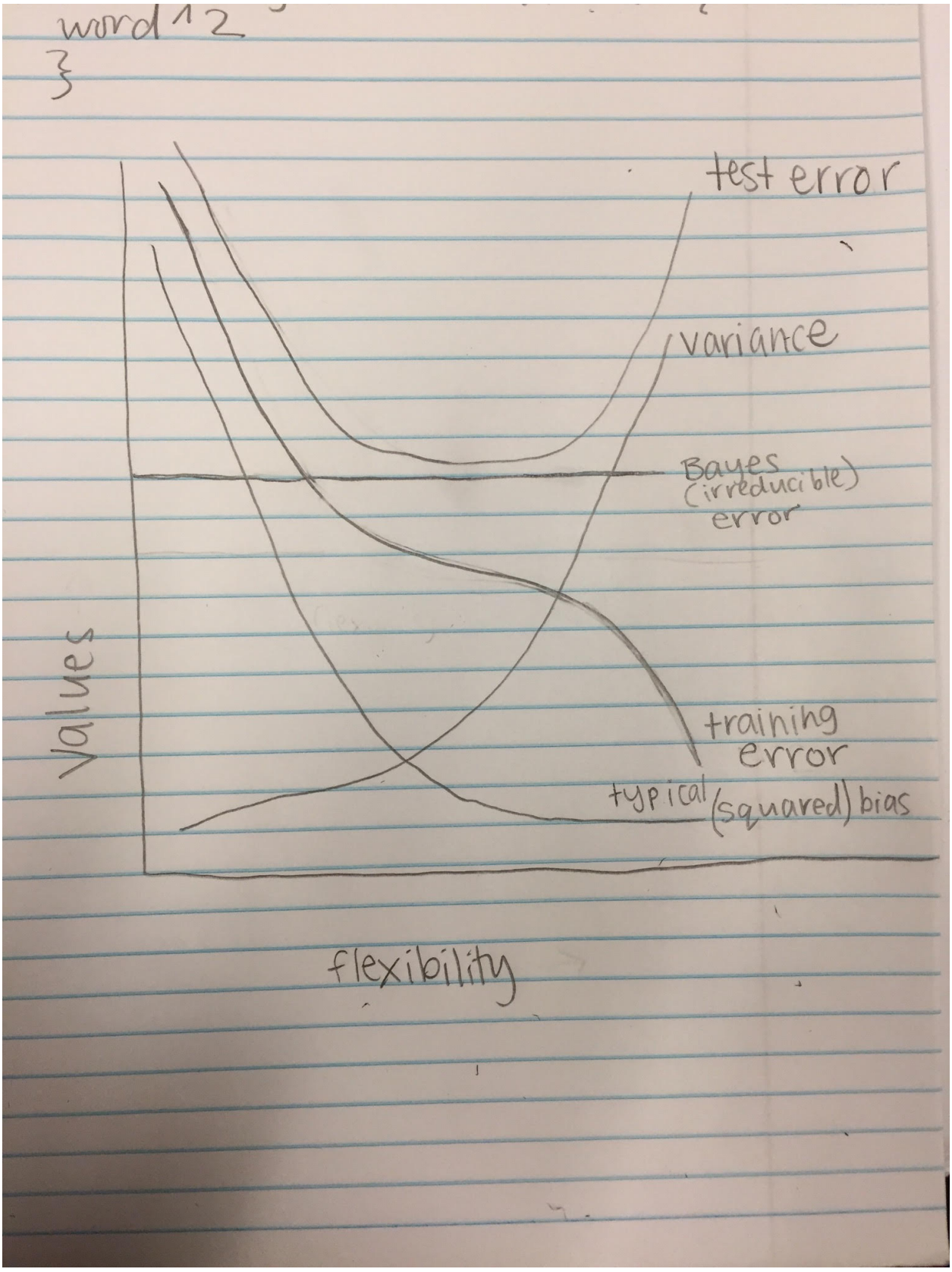


Figure 1: Here's my curves

Topic 1 Exercises

Shelby Witherby

DISCUSSION QUESTIONS:

ISL 2.4.1

a) With more data, you can train the model and support more flexibility. Therefore, the performance of the flexible model would be best.

b) The performance of the inflexible model would be best. We don't have enough observations to train the model.

c) The performance of the flexible model would be best. Since it is highly non-linear, the linear model wouldn't provide an accurate estimation for the data.

d) The performance of the inflexible model would be best. The flexible model would try to fit to the noise.

ISL 2.4.3

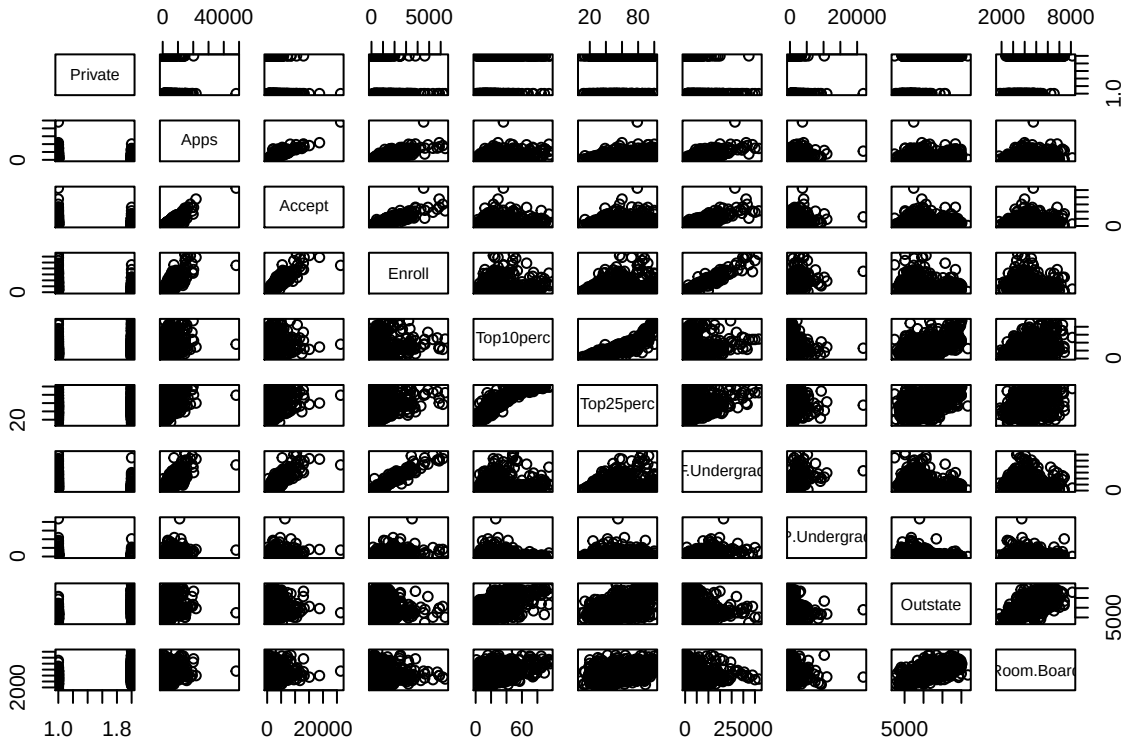
a) below you can see my sketch of curves for bias-variance decomposition

b) The variance "refers to the amount f hat would change if we estimated it with a different training data set," and "more flexible statistical methods have higher variance" according to the textbook. Thus, variance increases as flexibility increases. The training data declines as flexibility increases because as flexibility increases, the f curve fits the observed data more closely. The test error declines at first as flexibility increases, but then levels off and starts to increase again, because eventually we would be overfitting the data (provided small training error). The irreducible error is the minimum lower bound for the test error, and is a constant, so it is a straight line parallel to the x axis. The squared bias starts really high, and decreases with increased flexibility. This is because a very simple (not flexible) model, like linear regression, probably won't represent most real-life situations.

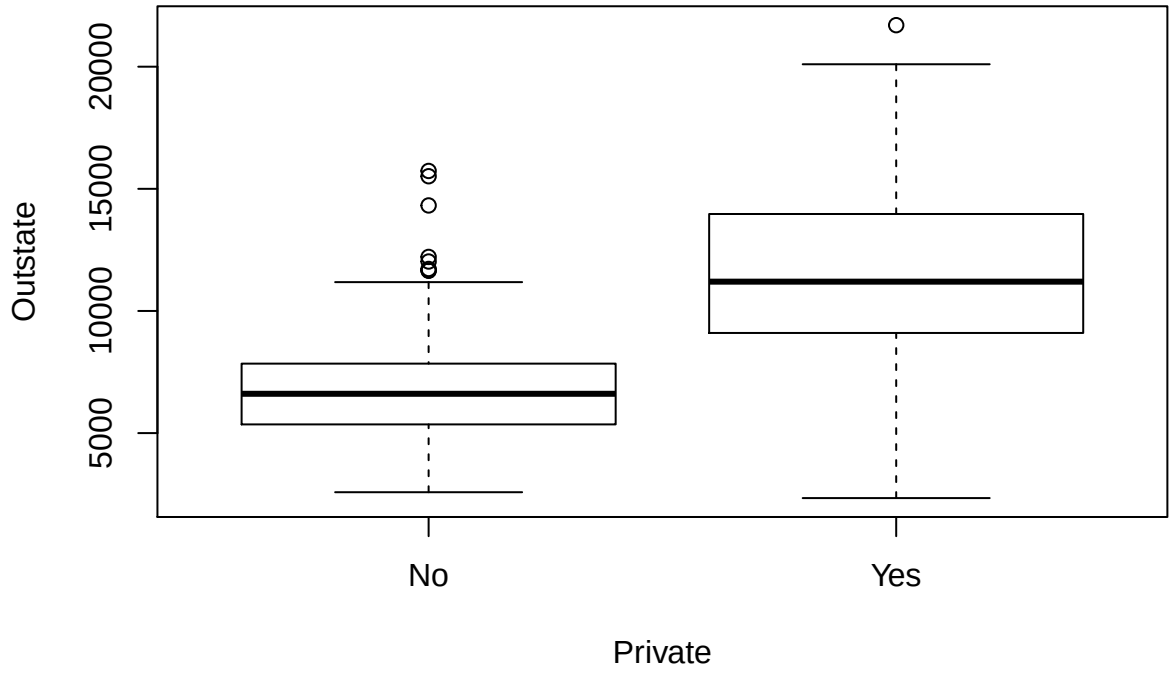
```r
library(ISLR)
data(College)
college <- read.csv("http://www-bcf.usc.edu/~gareth/ISL/College.csv", header = TRUE)
college <- College
#rownames(college)=college[,1]
#View(college)
#college = college[,-1]
#fix(college)
summary(college)
```

```
##  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
```

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




```
plot(Outstate ~ Private, data = college)
```

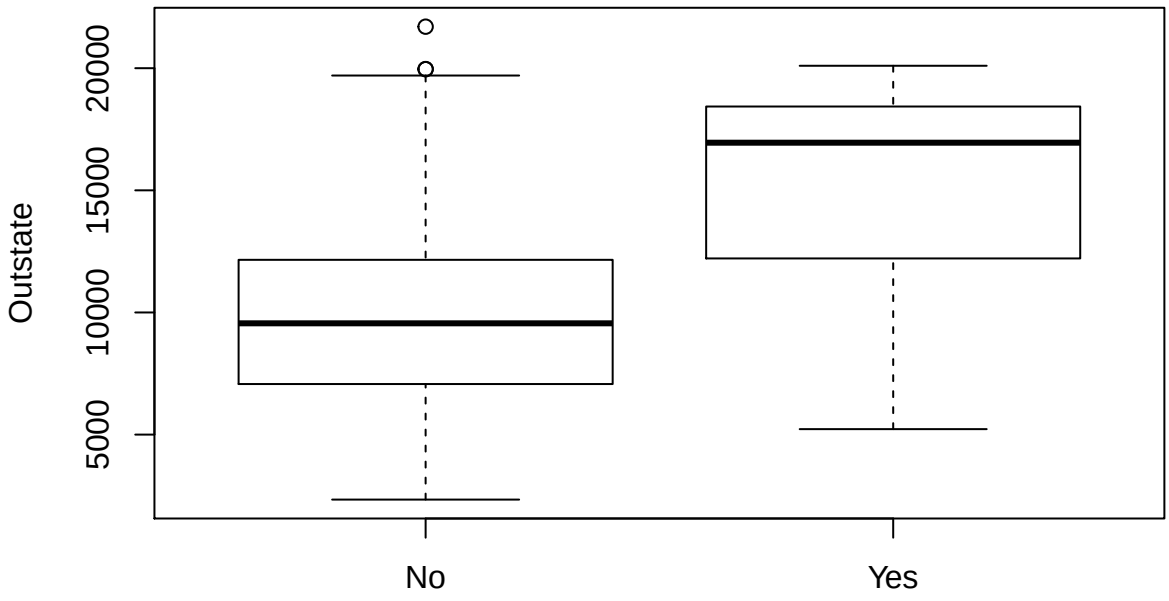



```
Elite = rep("No", nrow(college))
Elite[college$Top10perc >50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)
summary(college)
```

```
##  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      Elite
##  Min.   : 10.00   No :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```
plot(Outstate ~ Elite, data = college)
```



At elite colleges, there is a higher proportion of out of state students. (Hey Mac)

ISL 2.4.9

```
library(ISLR)
data(Auto)
Auto <- Auto[ - complete.cases(Auto),]
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 :23.00   Median :4.000   Median :151.0   Median : 93.0
##  Mean   :23.46   Mean   :5.465   Mean   :194.1   Mean   :104.4
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:264.5   3rd Qu.:125.0
##  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.:2224   1st Qu.:13.80   1st Qu.:73.00   1st Qu.:1.000
##  Median :2800   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2976   Mean   :15.55   Mean   :75.99   Mean   :1.578
##  3rd Qu.:3616   3rd Qu.:17.05   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)           :364
```

a) mpg, cylinders, displacement, horsepower, weight, acceleration, and year are quantitative. Origin and name are qualitative.

```
with(Auto, range(mpg))
```

```
## [1]  9.0 46.6
```

```
with(Auto, range(cylinders))
```

```
## [1] 3 8
```

```
with(Auto, range(displacement))
```

```
## [1]  68 455
```

```
with(Auto, range(horsepower))
```

```
## [1]  46 230
```

```
with(Auto, range(weight))
```

```
## [1] 1613 5140
```

```
with(Auto, range(acceleration))
```

```
## [1]  8.0 24.8
```

```
with(Auto, range(year))
```

```
## [1] 70 82
```

b) The range of mpg is 9 to 46.6. The range of cylinders is 3 to 8. The range of displacement is 68 to 455. The range of horsepower is 46 to 230. The range of weight is 1613 to 5140. The range of acceleration is 8 to 24.8. The range of year is 70 to 82.

```
with(Auto, c(mean(mpg), sd(mpg)))
```

```
## [1] 23.459847  7.810128
```

```
with(Auto, c(mean(cylinders), sd(cylinders)))
```

```
## [1] 5.465473 1.703152
```

```
with(Auto, c(mean(displacement), sd(displacement)))
```

```
## [1] 194.1240 104.6225
```

```
with(Auto, c(mean(horsepower), sd(horsepower)))
```

```
## [1] 104.40409  38.51873
```

```r
with(Auto, c(mean(weight), sd(weight)))
```

```
## [1] 2976.2379  850.0719
```

```r
with(Auto, c(mean(acceleration), sd(acceleration)))
```

```
## [1] 15.550384  2.756557
```

```r
with(Auto, c(mean(year), sd(year)))
```

```
## [1] 75.994885  3.675975
```

c) The mean of mpg is 23.5 and the standard deviation is 7.8. The mean of cylinders is 5.5 and the standard deviation is 1.7. The mean of displacement is 194.1 and the standard deviation is 104.6. The mean of horsepower is 104.4 and the standard deviation is 38.5. The mean of weight is 2976.2 and the standard deviation is 850.1. The mean of acceleration is 15.6 and the standard deviation is 2.8. The mean of year is 76 and the standard deviation is 3.7.

```r
Auto2 <- Auto[ - (10:85), ]

# The ranges
with(Auto2, range(mpg))
```

```
## [1] 11.0 46.6
```

```r
with(Auto2, range(cylinders))
```

```
## [1] 3 8
```

```r
with(Auto2, range(displacement))
```

```
## [1]  68 455
```

```r
with(Auto2, range(horsepower))
```

```
## [1]  46 230
```

```r
with(Auto2, range(weight))
```

```
## [1] 1649 4997
```

```r
with(Auto2, range(acceleration))
```

```
## [1]  8.5 24.8
```

```r
with(Auto2, range(year))
```

```
## [1] 70 82
```

```r
# The mean and standard deviations
with(Auto2, c(mean(mpg), sd(mpg)))
```

```
## [1] 24.427937  7.867464
```

```
with(Auto2, c(mean(cylinders), sd(cylinders)))
```

```
## [1] 5.365079 1.650146
```

```
with(Auto2, c(mean(displacement), sd(displacement)))
```

```
## [1] 187.1333 100.0462
```

```
with(Auto2, c(mean(horsepower), sd(horsepower)))
```

```
## [1] 100.75556  35.97364
```

```
with(Auto2, c(mean(weight), sd(weight)))
```

```
## [1] 2934.7333  812.5334
```

```
with(Auto2, c(mean(acceleration), sd(acceleration)))
```

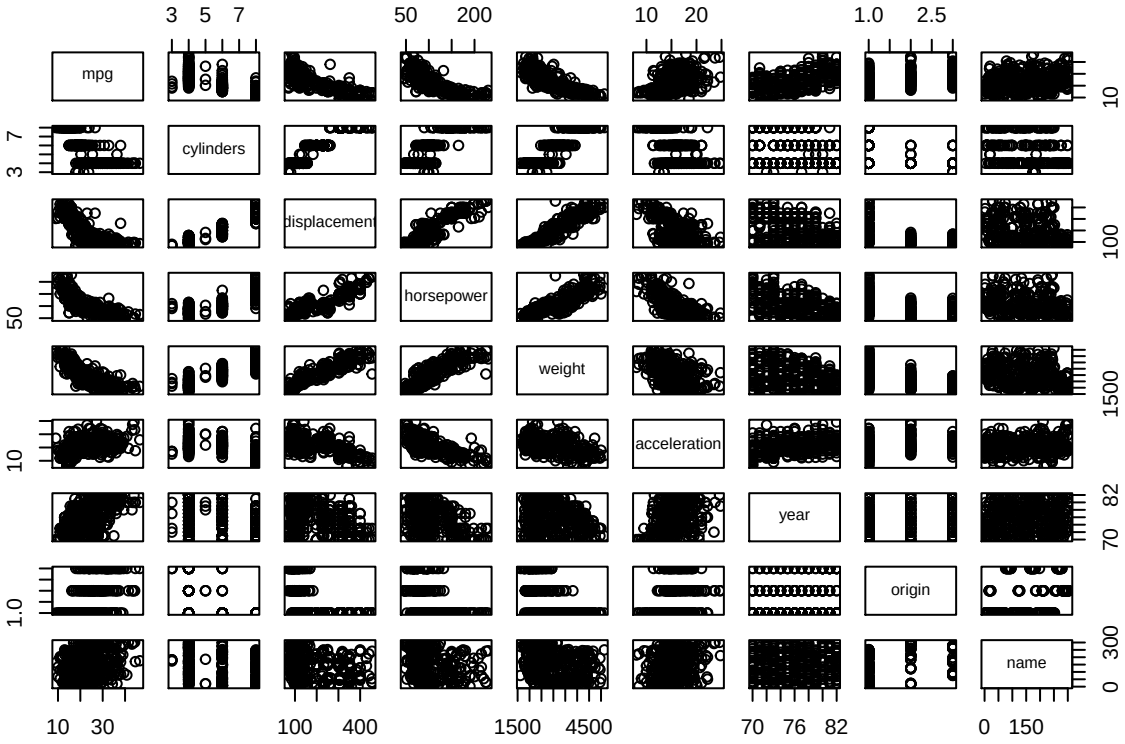
```
## [1] 15.729206  2.710061
```

```
with(Auto2, c(mean(year), sd(year)))
```

```
## [1] 77.158730  3.102323
```

d) The range of mpg is 11 to 46. The range of cylinders is 3 to 8. The range of displacement is 68 to 455. The range of horsepower is 46 to 230. The range of weight is 1649 to 4997. The range of acceleration is 8.5 to 24.8. The range of year is 70 to 82. The mean of mpg is 24 with a standard deviation of 7.9. The The mean of cylinders is 5 with a standard deviation of 1.7. The mean of displacement is 187 with a standard deviation of 100. The mean of horsepower is 100.7 with a standard deviation of 35.97. The mean of weight is 2934.7 with a standard deviation of 812. The mean of acceleration is 15.7 with a standard deviation of 2.7. The mean of year is 77 with a standard deviation of 3.

```
pairs(Auto)
```


mpg
cylinders
displacement
horsepower
weight
acceleration
year
origin
name

e) #HA! I made scatterplots of relationships with EVERY predictor!! weight has a positive linear relationship with horsepower. weight has a positive linear relationship with displacement. weight has a negative linear relationship with mpg. horsepower has a negative relationship with mpg. displacement has a negative relationship with mpg. Though the variance is high, year is positively associated with mpg - newer cars get better mileage.

f) Yes, many predictors might be useful to predict mpg. For example, variables displacement, horsepower and weight appear to have a strong correlation with mpg. My plots show that more weight results in lower mpg, more horsepower results in lower mpg, more displacement results in lower mpg. and a larger year (newer) results in higher mpg.

THEORY ASSIGNMENT:

ISL 2.4.2

a) regression, inference, n=500, p=3

b) classification, prediction, n=20, p=13

c) regression, prediction, n = 52, p = 3

ISL 2.4.7

```
sqrt((0-0)^2 + (3-0)^2 + (0-0)^2)
```

```
## [1] 3
```

```
sqrt((2-0)^2 + (0-0)^2 + (0-0)^2)
```

```
## [1] 2
```

```
sqrt((0-0)^2 + (1-0)^2 + (3-0)^2)
```

```
## [1] 3.162278
```

```
sqrt((0-0)^2 + (1-0)^2 + (2-0)^2)
```

```
## [1] 2.236068
```

```
sqrt((-1-0)^2 + (0-0)^2 + (1-0)^2)
```

```
## [1] 1.414214
```

```
sqrt((1-0)^2 + (1-0)^2 + (1-0)^2)
```

```
## [1] 1.732051
```

a)

Obs. 1: 3

Obs. 2: 2

Obs. 3: sqrt{10}

Obs. 4: sqrt{5}

Obs. 5: sqrt{2}

Obs. 6: sqrt{3}

b) Obsesrvation 5 has the smallest value for the Euclidian distance and the test point. As such, it will be the value associated with obs 5, which is green.

c) With k = 3, the three nearest are observations 5 (green), 6 (red), and 2 (red). Therefore, it will be red.

d) If the Bayes decision boundary in this problem is highly non-linear, we would expect the best value for K to be small. If it is highly nonlinear, that would mean that the model values change quickly. As such, models with small k values are able to change quickly.