Ch.2 Exercises: Statistical Learning

Conceptual

1.

(a)

- **Better** : A large sample size means a flexible model will be able to better fit the data.

(b)

- **Worse** : A flexible model would likely overfit. Flexible methods generally do better when large datasets are available.

(c)

- **Better** : Flexible methods perform better on non-linear datasets as they have more degrees of freedom to approximate a non-linear

(d)

- **Worse** : A flexible model would likely overfit, due to more closely fitting the noise in the error terms than an inflexible method. In other words, the data points will be far from f (ideal function to describe the data) if the variance of the error terms is very high. This hints that f is linear and so a simpler model would better be able to estimate f.

2.

(a)

- **Regression**;The response in this case is quantitative and so this is a regression problem.
- **Inference**; We want to understand how the predictors impact the salary of a CEO, and not actually predict the salary of a CEO. Therefore, inference is our aim.
- n=500. p=profit, number of employees, industry.

(b)

- **Classification**; response(success or failure) is a binary value.
- **Prediction**; we want to know predicted value of the target.
- n=20. p=price, marketing budget, competition price and 10 other variables.

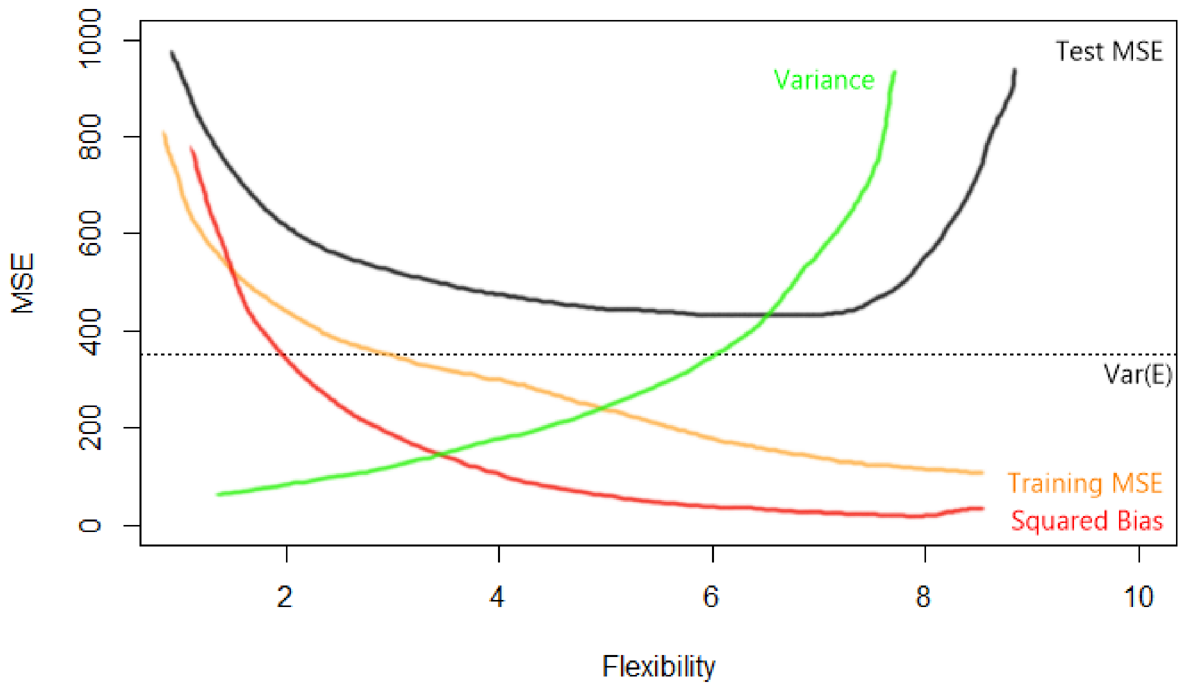
(c)

- **Regression**; response(% change in the USD/Euro) is a quantitative value.
- **Prediction**, we want to predict the response.

- n=52. p=% change in US market, % change in UK market, % change in German market.

3.

(a)



(b)

- Var(E) is the irreducible error and the test predictions cannot be better than this, therefore it is a straight line. Test MSE reduces to an optimum point as increased flexibility means a better fit, with further increases leading to overfitting. Training MSE continues to reduce as more flexibility means the method can very closely fit the training data. Variance increases as the method tends to overfit as flexibility increases (fitting training data too well and not generalizing to test data). Generally, bias is reduced as the flexibility increases due to the method being better able to fit the data.

4.

(a)

Classification methods would be useful for applications where the outcomes are to be classified into a category, this can be a binary classification or a multi-class classification. Some areas where classification could be useful:

- **Breast cancer prediction:** Given a set of predictors such as a mammogram scan, age, family history, lifestyle and other variables, and a response of *Yes*(has cancer) and *No*(does not have cancer) – we can then train a model to predict whether a patient has breast cancer.
- **Classifying species of plants:** Given a set of images of a plant, a model can be trained that will classify that plant into one of the trained species. This is a multi-class classification problem. The response would be the species name and the predictors would be images of that species.

- **Fraud detection:** Classify whether a transaction is fraudulent, given data like the transaction amount, location, purchased item or service, previous customer transactions etc. The response would be “Yes” or “No”, and our aim is to make a prediction.
- **Stock price:** Classify whether a stock will go up or down in price the next day given a set of financial data and news from the preceding week. The aim is to make a prediction.

(b)

Regression methods are useful when we have a quantitative response; that is where we need to predict a numerical value for our response. Some areas where regression could be useful are:

- **House price factors:** Given a set of predictors such as location, house features, median income for the area and so on and the house price as the response/target, we can train a model to infer the impact of those variables on house prices.
- **Salary:** Predict the salary of an individual given their education, work history, skillsets and other relevant data (age, sex, etc.). The response is the salary amount.
- **Sales:** Predict unit sales of a product given marketing data such as TV, Radio or Internet advert expenditure, and use it to infer the importance of each advertising method. The response is the unit sales of the product.
- **Driving Insurance premium:** Given a set of variables such as the drivers history, age, type of vehicle, expected yearly mileage and the premium as the response, we can train a model to predict the insurance premium for new customers.

(c)

Cluster analysis is useful in cases where we do not have a target response available – i.e. unsupervised learning. We can aim to ascertain whether observations can be classed into distinct groups or understand if there are any underlying relationship between variables. Some areas where this can be useful are:

- **Tissue classification:** : Clustering can be used to separate different types of tissue in medical images. This can be useful in identifying groups of tissue that are not normal and need further study.
- **Market research:** Differentiate a group of people within a city into distinct market segments to increase marketing effectiveness or identify new opportunities. Given data such as incomes, location, age, sex, opinion polls and so on for a city, we can segment the city into different consumer areas.
- **Image segmentation:** Separate an image into different regions to make object recognition easier. For example, segmenting image frames from a video camera in a car into ‘other vehicles’, ‘humans’, ‘road signs’ and so on can help ADAS (Advanced driver-assistance systems) in vehicles make the correct decision.
- **Gaming market segmentation:** Given a set observations with variables such as age, location, income, sex, hours spent gaming, gaming devices used and so on. We could use cluster analysis to see if these observations fall into distinct groups. If there are distinct groupings, then it could be helpful with further study – say for example one grouping could represent casual gamers and the other hardcore gamers, and another one could be newer gamers (say people over the age 60).

5. & **6.**

- Flexible methods work well when the underlying function is non-linear. The predictions in general have a lower bias but can have a higher variance, as these models are more likely to overfit the data.

- Less flexible methods do not tend to overfit the data but can have a high bias when the underlying function is non-linear. They can also use fewer observations and parameters, particularly when it is assumed that the underlying function is linear. Flexible methods tend to require a larger number of observations and parameters, and can lead to overfitting (higher variance).
- Flexible methods (non-parametric methods) are preferable when we make no assumptions about the function to be estimated.Most real-life relationships are non-linear and so a non-parametric approach is better suited to modelling them. Flexible models by their nature are more complex and less interpretable than their linear counterparts, so even though their predictions might be more accurate, we may not be able to explain why it has made those predictions (a black box model).
- Less flexible methods (parametric) are useful if we assume or know that the underlying function is linear. As a linear relationship is assumed, the model needs to predict fewer parameters than a non-parametric method. Additionally, these models are more interpretable, and so will be preferred when we are interested in making inferences or the interpretability of the results.

7. (a) The Euclidean distance is the straight line distance between two points. This can be calculated using the Pythagorean theorem.

For 3D space we have:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

Using the above formula, we get the following distances:

$$\begin{aligned}
d(1, test) &= 3 \\
d(2, test) &= 2 \\
d(3, test) &= 3.16 \\
d(4, test) &= 2.24 \\
d(5, test) &= 1.41 \\
d(6, test) &= 1.73
\end{aligned}$$

(b)

- **Green**; as nearest single observation is green.

(c)

- **Red**; as nearest three observations are green, red and red.The probability of the test point belonging to red is 2/3 and green is 1/3. Therefore, the prediction is red.

(d)

- For highly non-linear boundaries, we would expect the best value of K to be small. Smaller values of K result in a more flexible KNN model, and this will produce a decision boundary that is non-linear. A larger K would mean more data points are considered by the KNN model and this means its decision boundary is closer to a linear shape.

Applied

8.

```r
library(ISLR)
#library(ggplot2)
```

(a) (b)

```r
college.rownames = rownames(College)
```

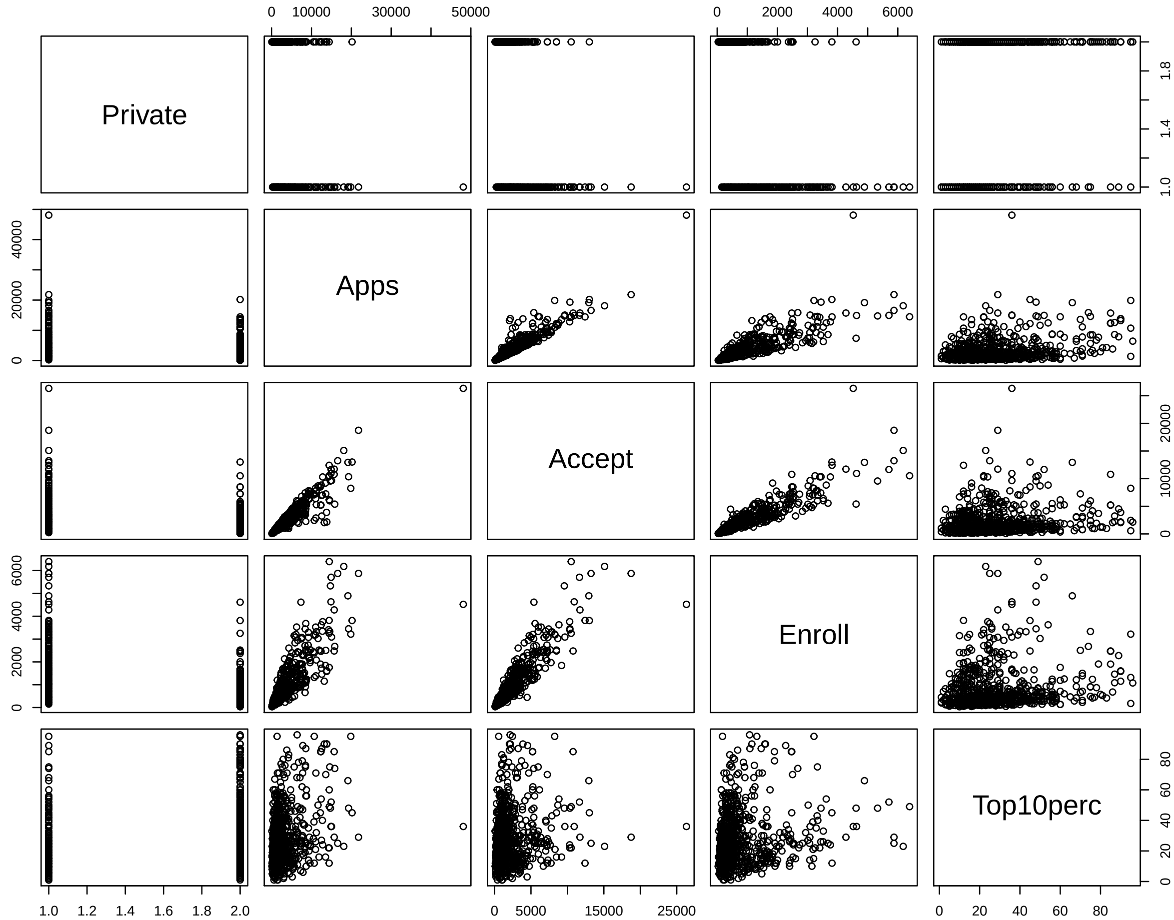
(c) i.

```r
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
```

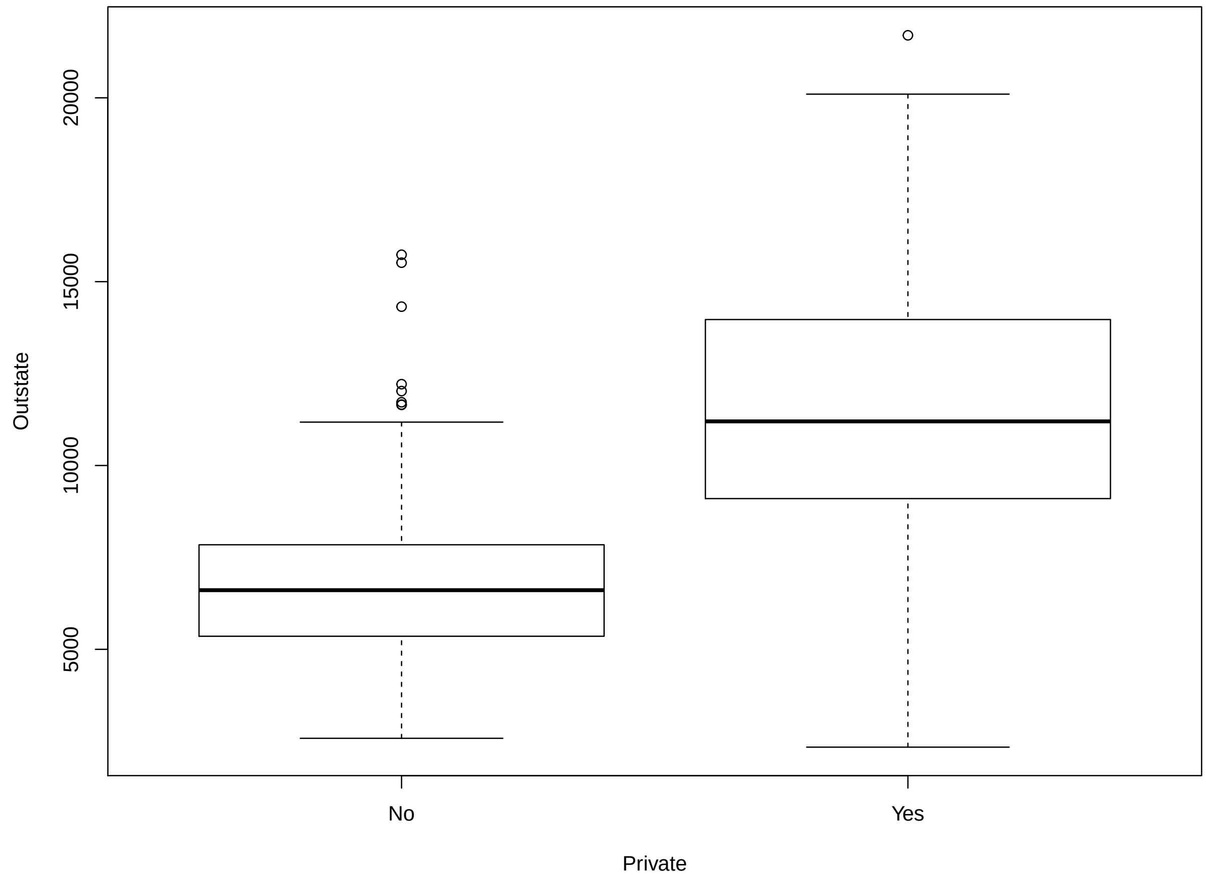
ii.

```r
# Pair plot of first 5 variables
pairs(College[,1:5])
```



iii.

```
# Boxplots of outstate vs Private.
boxplot(Outstate~Private, data = College)
```



iv.

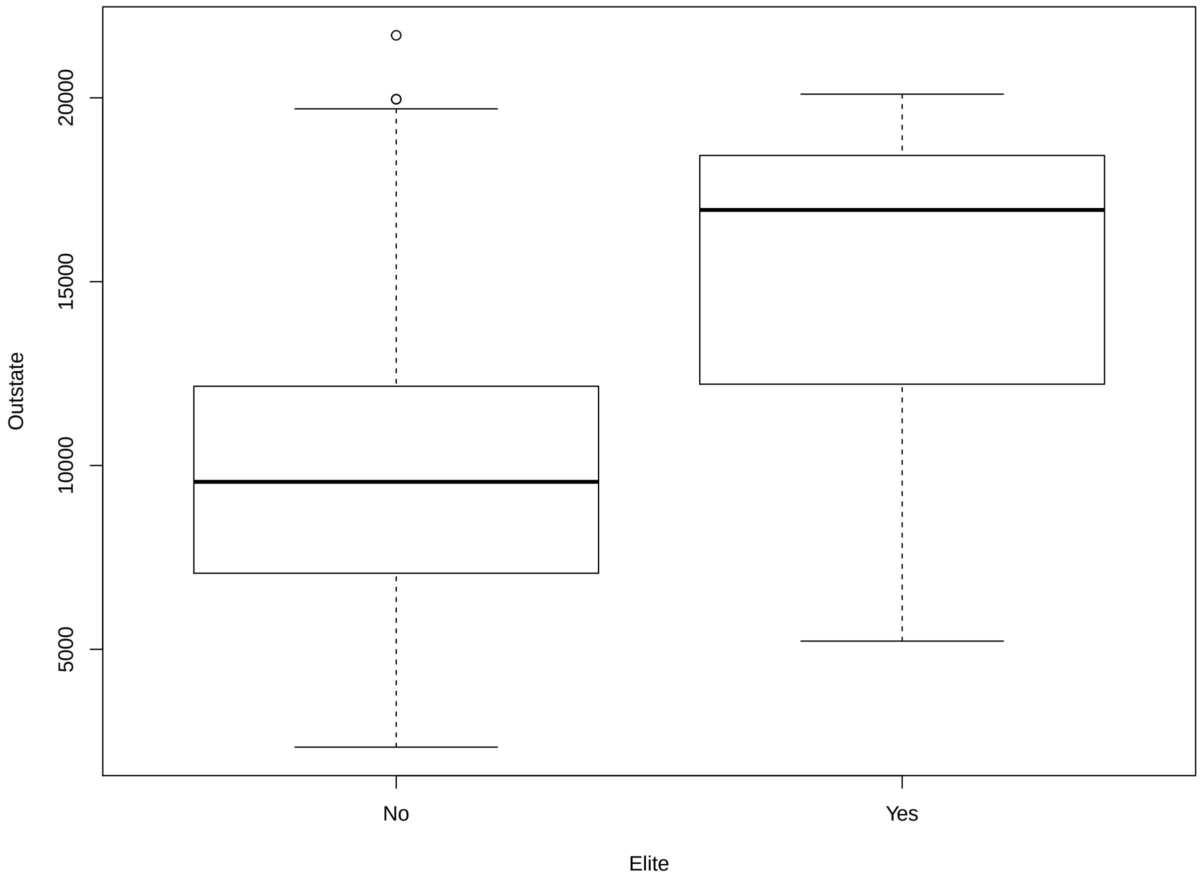
```r
# New variable 'Elite'
Elite = rep("No",nrow(College))
Elite[College$Top10perc>50] = "Yes"
college.df = data.frame(College, Elite)

summary(college.df)
```

```
##  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            
```

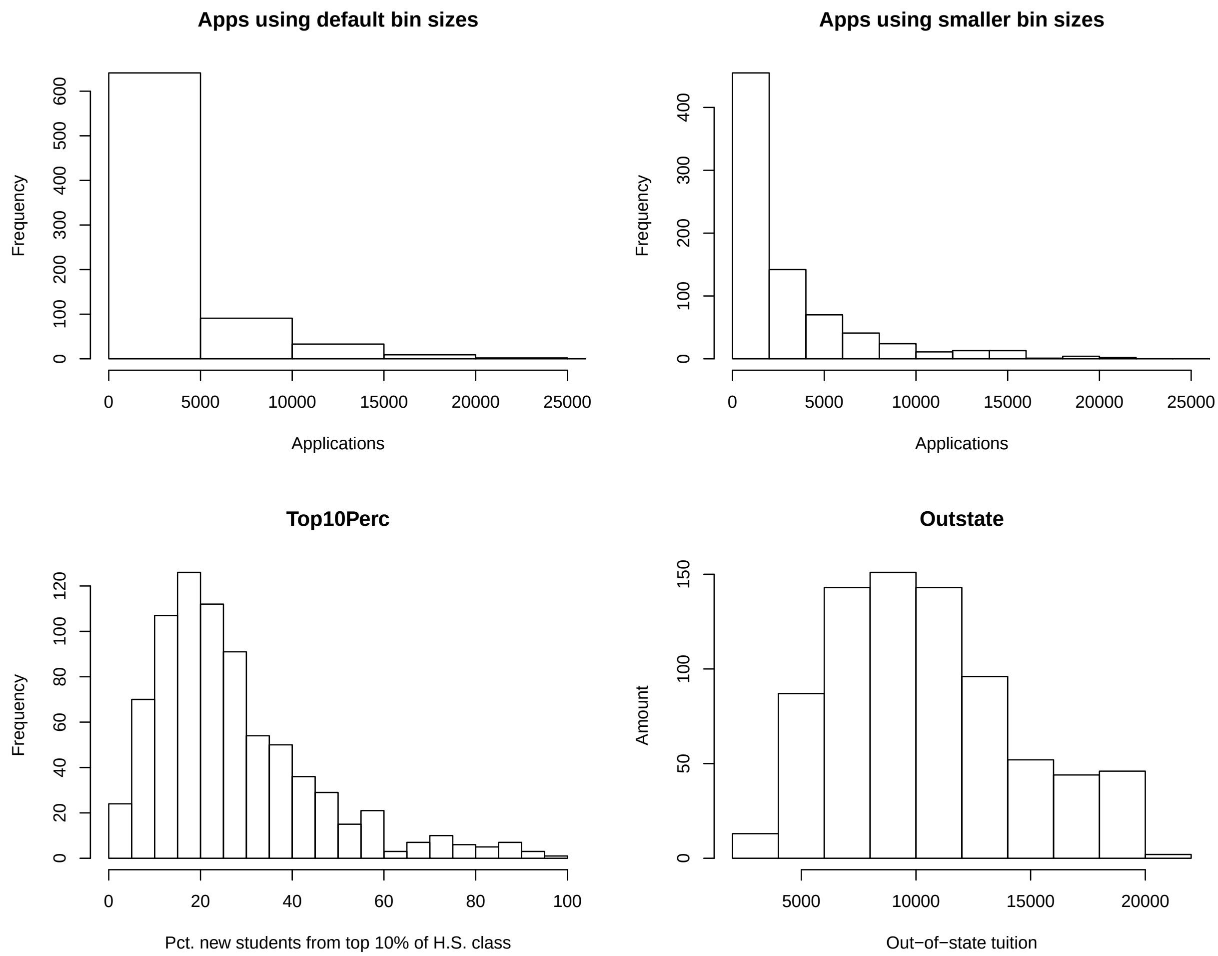
- There are 78 elite universities.

```r
boxplot(Outstate~Elite, data = college.df)
```



v.

```
par(mfrow=c(2,2))
hist(College$Apps, xlim=c(0,25000), xlab = "Applications", main = "Apps using default bin sizes")
hist(College$Apps, xlim=c(0,25000), breaks=25, xlab = "Applications",
     main = "Apps using smaller bin sizes")
hist(College$Top10perc, breaks=25, xlab = "Pct. new students from top 10% of H.S. class",
     main="Top10Perc")
hist(College$Outstate, xlab="Out-of-state tuition",ylab="Amount",main="Outstate")
```



- Histogram of Apps(Number of applications received) is highly right skewed. This shows that most universities received 5000 or fewer applications. The mean number of applications received will also be heavily skewed.
- Histogram for Top10Perc(Number of new students who are the top 10% of their class) is also right skewed; this shows that only a few universities get a majority of their new students from this cohort.

```
mean(college.df$Apps)
```

```
## [1] 3001.638
```

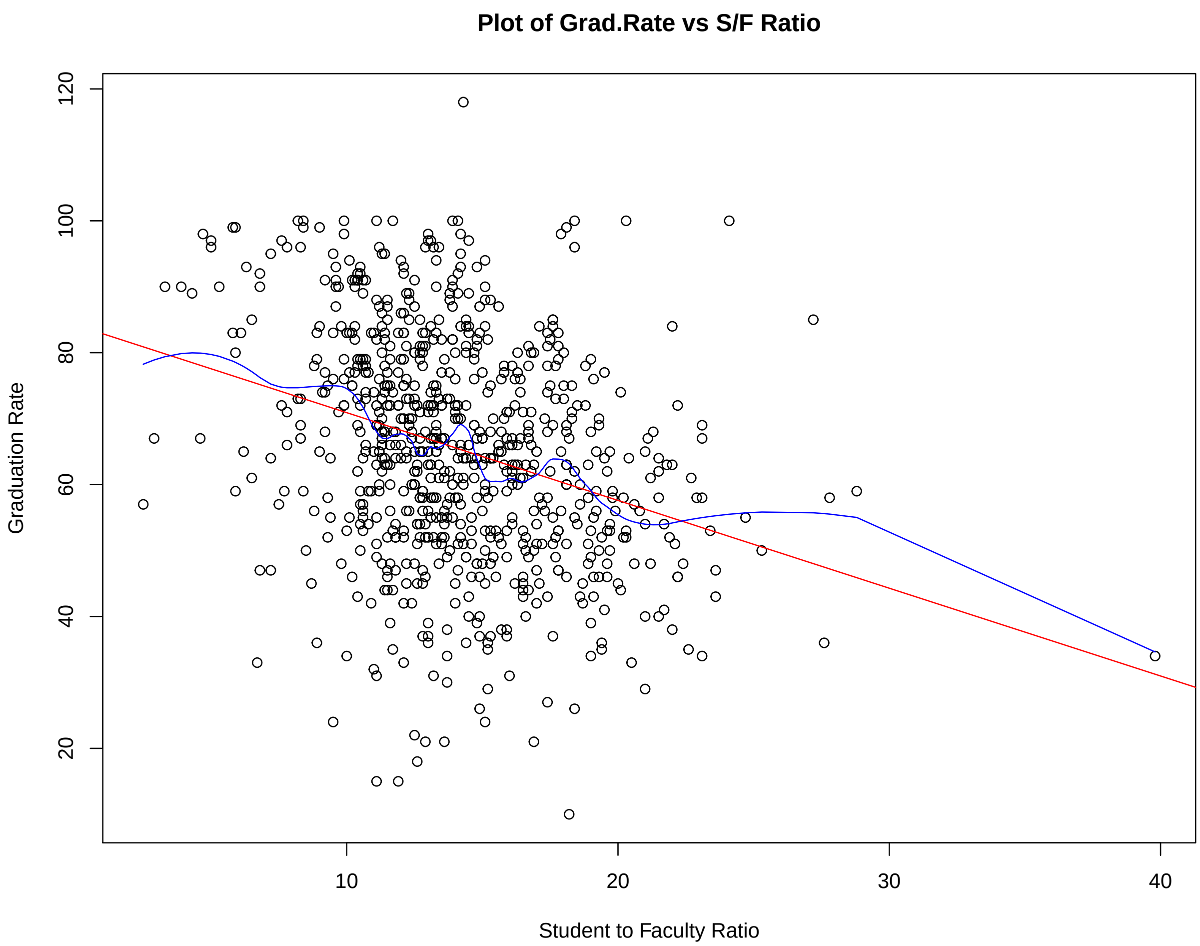
```
median(college.df$Apps)
```

```
## [1] 1558
```

vi.

```
# Exploring the relationship between Grad.Rate(Graduation Rate) and S.F.Ratio(Student/faculty ratio).
plot(college.df$S.F.Ratio, college.df$Grad.Rate,
```

```r
     xlab = "Student to Faculty Ratio", ylab = "Graduation Rate",
     main = "Plot of Grad.Rate vs S/F Ratio")

# Linear regression line.
abline(lm(college.df$Grad.Rate~college.df$S.F.Ratio), col="red")

# Local regression line with smoothing of 25%.
loessMod = loess(Grad.Rate ~ S.F.Ratio, data=college.df, span=0.25)
j = order(college.df$S.F.Ratio)
lines(college.df$S.F.Ratio[j],loessMod$fitted[j], col="blue")
```



- The results suggest a negative linear relationship between the graduation rate of students and the student to faculty ratio at universities.
- As the student to faculty ratio increases, we can expect students to have a lower graduation rate.

9.

(a)

- **Quantitative**: mpg,cylinders,displacement,horsepower, weight, acceleration, year.
- **Qualitative**: name, origin.

(b)

```r
sapply(Auto[,1:7], 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
```

(c)

```r
# Mean and standard deviation.
sapply(Auto[,1:7], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```r
sapply(Auto[,1:7], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

(d)

```r
# Remove 10th to 85th rows from Auto.
Auto.reduced = Auto[-c(10:84),]
sapply(Auto.reduced[,1:7], range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
```

```r
sapply(Auto.reduced[,1:7], mean)
```

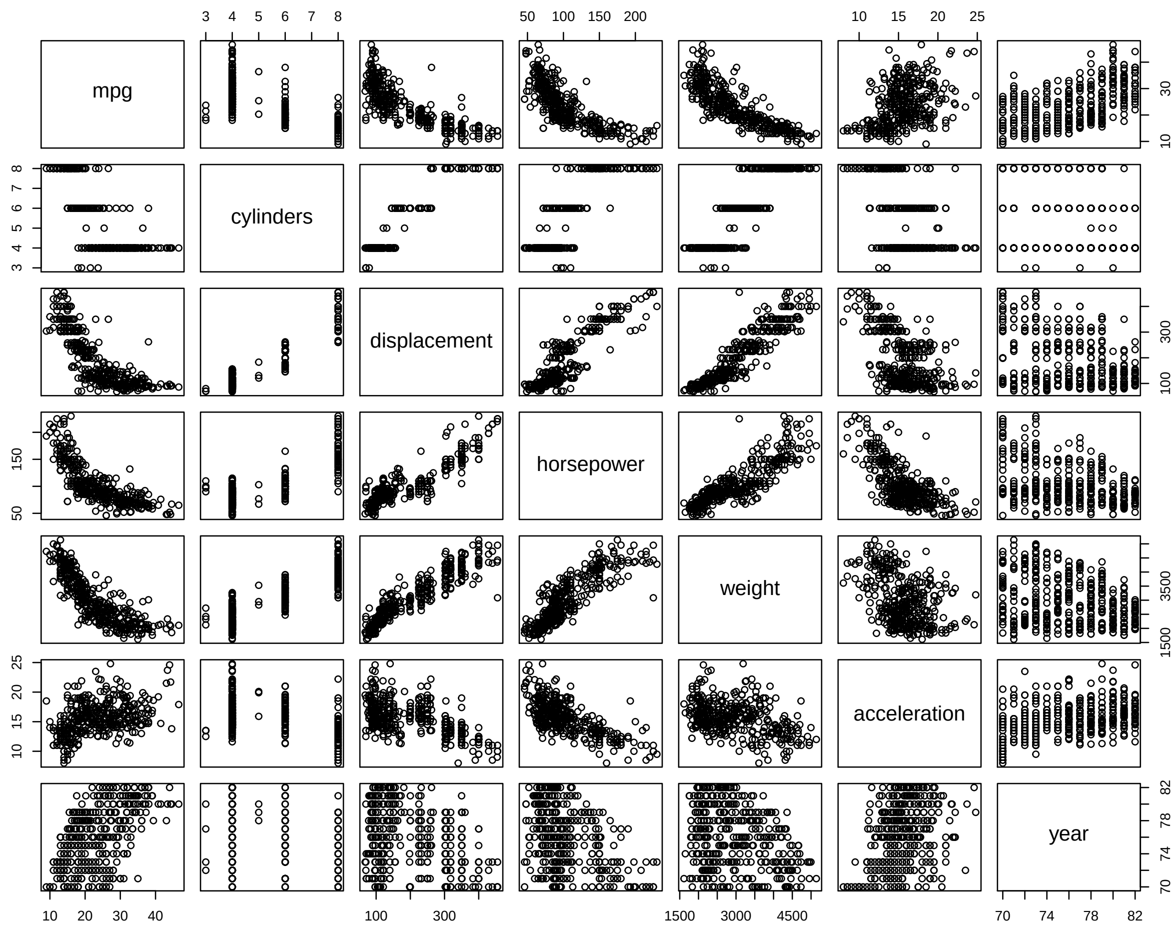
```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.368454     5.381703   187.753943   100.955836  2939.643533    15.718297
##         year
##    77.132492
```

```r
sapply(Auto.reduced[,1:7], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.880898     1.658135    99.939488    35.895567   812.649629     2.693813
##         year
##     3.110026
```

(e)

```r
pairs(Auto[,1:7])
```




```r
cor(Auto[,1:7])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
##              acceleration       year
## mpg             0.4233285  0.5805410
## cylinders      -0.5046834 -0.3456474
## displacement   -0.5438005 -0.3698552
## horsepower     -0.6891955 -0.4163615
## weight         -0.4168392 -0.3091199
## acceleration    1.0000000  0.2903161
## year            0.2903161  1.0000000
```

- From the pair plot and the correlation data, we can state there exists linear relationships between some of the variables.
- For example, `mpg` has strong negative linear relationships with `displacement`, `cylinders` and `weight`. That is we can expect the `mpg` of the car to decrease as their `displacement` and `cylinders` increase.
- `mpg` has a positive correlation with `year`, and this suggests that newer models tend to have higher `mpg` than older ones.

(f)

- Both the plots and the correlation data suggests we can predict mpg.
- An increase in the variables `displacement`, `cylinders` and `weight` will lead to a reduced `mpg`.
- Newer models `year` tend to have higher `mpg`.

10.

(a)

```r
library(MASS)
```

?Boston

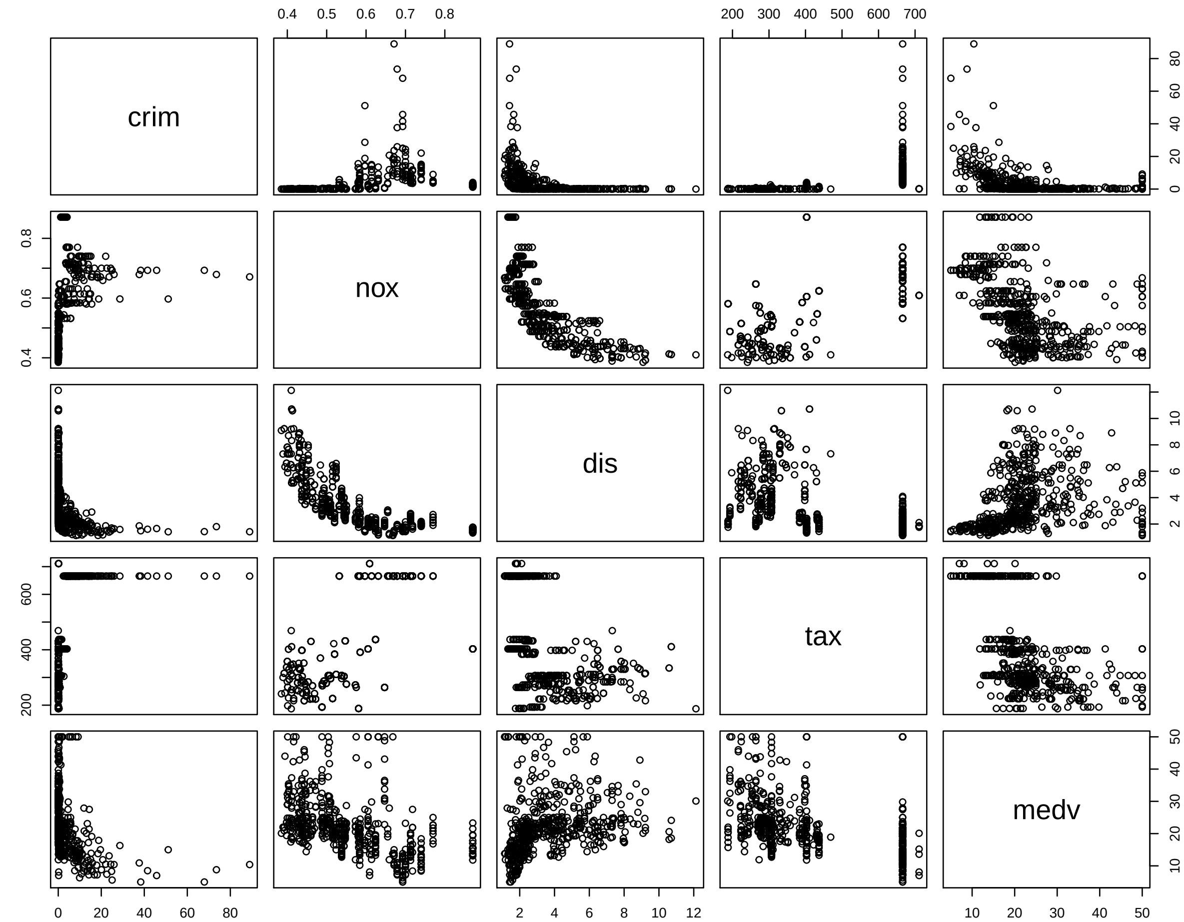
```r
dim(Boston)
```

```
## [1] 506  14
```

- 506 rows of suburbs or towns and 14 columns of predictors.

(b)

```r
# Pair plots of some variables
pairs(~crim+nox+dis+tax+medv, data = Boston)
```



- `crim` seems to have a negative linear relationship with `medv` and `dis`.
- `nox` has a negative linear relationship with `dis`.
- `dis` has a positive linear relationship with `medv`.

(c)

```
# Correlation coefficients between CRIM and all other variables.
cor(Boston[-1],Boston$crim)
```

```
##                [,1]
## zn      -0.20046922
## indus    0.40658341
## chas    -0.05589158
## nox      0.42097171
## rm      -0.21924670
## age      0.35273425
## dis     -0.37967009
## rad      0.62550515
## tax      0.58276431
## ptratio  0.28994558
## black   -0.38506394
```

```
## lstat    0.45562148
## medv    -0.38830461
```

- There are some correlations between `crim` and other variables, but they are not as strong as some of the relationships we observed in the `Auto` dataset.
- `crim` has a negative linear relationship with `medv`, `dis` and `black`.
- `crim` has a positive linear relationship with `indus`, `nox`, `rad` and `tax`.

(d)

```
# Suburbs with crime rate higher than 2 s.d from the mean(higher than 95% of suburbs).
High.Crime = Boston[which(Boston$crim > mean(Boston$crim) + 2*sd(Boston$crim)),]
range(Boston$crim) ; mean(Boston$crim) ; sd(Boston$crim)
```

```
## [1]  0.00632 88.97620
```

```
## [1] 3.613524
```

```
## [1] 8.601545
```

- There are 16 suburbs with a crime rate higher than 95% of the other suburbs.
- Some suburbs have extremely high rates of crime (5-8 s.d from the mean).
- The range is very wide, it goes from a rate of near zero to 89.

```
# Suburbs with tax rates higher than 2 s.d from the mean.
High.Tax = Boston[which(Boston$tax > mean(Boston$tax) + 2*sd(Boston$tax)),]
range(Boston$tax)
```

```
## [1] 187 711
```

- There are no suburbs with a tax rate higher than 2 s.d. from the mean. This seems reasonable as property tax rates are designed not to be extremely drastic.
- The range is narrower than the crime rate.
- Some suburbs do have tax rates higher than 1 s.d.(higher than 65% of suburbs) from the mean.

?Boston

```
# Suburbs with pupil teacher ratio higher than 2 s.d from the mean.
High.PT = Boston[which(Boston$ptratio > mean(Boston$ptratio) + 2*sd(Boston$ptratio)),]
range(Boston$ptratio)
```

```
## [1] 12.6 22.0
```

- There are no suburbs with a high pupil to teacher ratio, and this a reasonable outcome as educational laws limit the numbers of teacher or students per class/school.
- The range in quite narrow, and and all pupil teacher ratios lie within 2 s.d. of the mean.
- Some pupil teacher ratios are higher than 1 s.d.

(e)

```r
sum(Boston$chas==1)
```

```
## [1] 35
```

- 35 suburbs/towns bound the Charles river.

(f)

```r
median(Boston$ptratio)
```

```
## [1] 19.05
```

(g)

```r
which(Boston$medv == min(Boston$medv))
```

```
## [1] 399 406
```

- There are two suburbs (399 & 406) that have the lowest median property values.

```r
# Values of other predictors for suburb 399
Boston[399,]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.9 30.59
##     medv
## 399    5
```

```r
range(Boston$lstat)
```

```
## [1]  1.73 37.97
```

```r
range(Boston$ptratio)
```

```
## [1] 12.6 22.0
```

- `crim` is more than 2 s.d. above the mean - very high crime rates in this suburb. Both `ptratio`and `lstat` are close to their maximum values.

(h)

```r
# More than 7 rooms
sum(Boston$rm > 7)
```

```
## [1] 64
```

```r
# More than 8 rooms
sum(Boston$rm > 8)
```

```
## [1] 13
```

```r
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00  
```

```r
summary(subset(Boston, rm > 8))
```

```
##       crim               zn            indus             chas       
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age             dis       
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
```

```
##       rad              tax           ptratio          black      
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv     
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

- Relatively low `crim`, `lstat` and much higher `medv` when comparing the IQR range.