Problem Set 1: R, R Markdown, Conceptual Foundations of ML

Candidate Number: 18195

09 February 2021

Part 1: Short Answer Questions

1. Imagine you have been hired as a data consultant. Your client has given you the task of building a classifier for a new dataset they have constructed. In each of the following 5 scenarios, would you recommend a flexible statistical learning method or an inflexible approach? Why? (2-3 sentences per scenario)

   a) There is a large sample size of $N = 5$ billion, a large number of predictors $p = 100,000$, and the client is limited in their computing resources.
   b) Large sample size of $N = 5$ billion, and small number of predictors $p = 6$.
   c) Large number of predictors, $p = 125,000$, sample size $N = 2000$ is relatively small.
   d) Based on exploratory analysis of the data, it appears that the predictors and the response have a non-linear relationship.
   e) The error term has very large variance.

a) Inflexible approach. The large number of sample size relative to predictors means that using a flexible approach should not result in overfitting, because there is enough variation per predictor that reduces the chances of the nodel memorising the dataset. However being flexible also means estimating more parameters which, given the company's lack of computing power, may not be feasible,

b) Flexible approach. If computing power is not a constraint, given the large number of observations a flexible approach should not lead to overfit even if the number of predictors is relatively small.

c) Inflexible approach. Because sample size is small but p is very large, using a flexible approach may lead to overfit where the training algorithm simply memorises the data. In this case using an inflexible approach like OLS regression may be better at predicting new data.

d) Flexible approach. If the relationship is non linear then traditional methods like OLS may not fit the model very well. In theory OLS can still work if enough polynomials are added but this will depend on how many observations there are and how flexible the function is - if n is too small then the model is likely to overfit, and if the function is extremely flexible then a large number of polynomials will be required which also tends to generate overfit.

e) If the error term has a very large variance it means that a lot of the variance is left unexplained by the current model. This may be the case when using inflexible models that are unable to fit flexible functional forms. In this case it may be wise to use a more flexible model, but not too flexible such that it overfits.

2. How is a **parametric** approach different from a **non-parametric** approach to statistical learning? How does each approach go about estimating f? Name three advantages and three disadvantages of each approach. (2-3 sentences per approach)

Parametric:

Advantage: 1. Parametric approaches like OLS will always produce unbiased betas which allows us to perform causal inference. 2. Computationally it is less costly since it does not require large n or large p to perform.

Disadvantage: 1. Because we assume a parametric form for f, the model we choose will usually not match the true unknown form of f. 2. Will perform poorly if the true functional form is very flexible.

Non-parametric:

Advantage: 1. Non-parametric methods do not make explicit assumptions about the functional form of f, which makes it much more flexible than parametric methods. 2. Produces very good predictions 3. We do not need to understand what the model does with the Ps, only that it makes good predictions.

Disadvantage: 1. Because no assumptions about parameters are made, large N is required. 2. Being too smooth can result in overfitting. 3. Can be computationally costly and time consuming since large n and large p is required

3. *ISL 2.4 Exercise 2*

a.This is a regression problem because CEO salary is a a continuous variable. We are also interested in causal inference since the focus is on the factors (x) that affects CEO salary (y). n=500 and p=3.

b. Classification problem since the outcome is a discrete choice between success and failure. We are interested in prediction since we want to predict if the new product is a success or failure. n=20, p=13.
c. Regression problem since the outcome percentage change is a continuous variable. We are interested in prediction. n=52, p=3.

4. *ISL 2.4 Exercise 3*
5. What are the two kinds of "big data" Rocio Titiunik wrote about in her paper on big data? What are some benefits and drawbacks of each kind of big data analysis for social scientific inquiry? Can either kind of big data solve the fundamental problem of causal inference? (5-10 sentences)

Big data can be large P or large N. Having a larger N is useful since there is more variation in the data and this may help us produce more accurate betas. However, no amnount of N can solve the problem of causal inference if the model is wrongly specified. For instance, if there is an omitted variable that is highly correlated with the error term. In addition, there is no N large enough that can cover all possible distributions that generated the data in the first place, and thus increasing N does not guarantee that we will converge on the true parameter.

Large P data allows us to reduce the problem of omitted variables and also estimate more flexible functional forms. In theory, if we have all the P in the world that could be correlated to a certain Y, we could circumvent the fundamental problem of causal inference. However, the problem is that this only holds if we have access to *all* variables necessary for exogeneity to hold, which is highly implausible. Even if this were possible, methods that allow for p>n require sparsity assumptions which can only be justified by strong theory,

Part 2: Coding Questions

6. In the next problem set, we will use `for` loops and `if/else` statements to implement k-fold cross-validation. To prepare you for this, we'll practice them using the fibbonacci sequence. The fibbonacci sequence is a sequence where each number is the sum of the two preceding ones: $(0,)1,1,2,3,5,\ldots$. Using `for` loops and `if/else` statements, write code that will output the sum of the first 50 terms of the fibbonacci sequence. Include zero as the first term.

```r
v<-c(0,1)
for (i in 1:48){
  v[i+2]<-v[i+1]+v[i]
}
v
```

```
##  [1]          0          1          1          2          3          5
##  [7]          8         13         21         34         55         89
## [13]        144        233        377        610        987       1597
## [19]       2584       4181       6765      10946      17711      28657
## [25]      46368      75025     121393     196418     317811     514229
## [31]     832040    1346269    2178309    3524578    5702887    9227465
## [37]   14930352   24157817   39088169   63245986  102334155  165580141
## [43]  267914296  433494437  701408733 1134903170 1836311903 2971215073
## [49] 4807526976 7778742049
```

7. *ISL 2.4 Exercise 10* (Note: 1. You will need to install the `MASS` library from CRAN. 2. Please break text out of code blocks when explaining or reporting your answers.)

```r
# Code for 10 a) goes here
library(MASS)
dim(Boston)
```

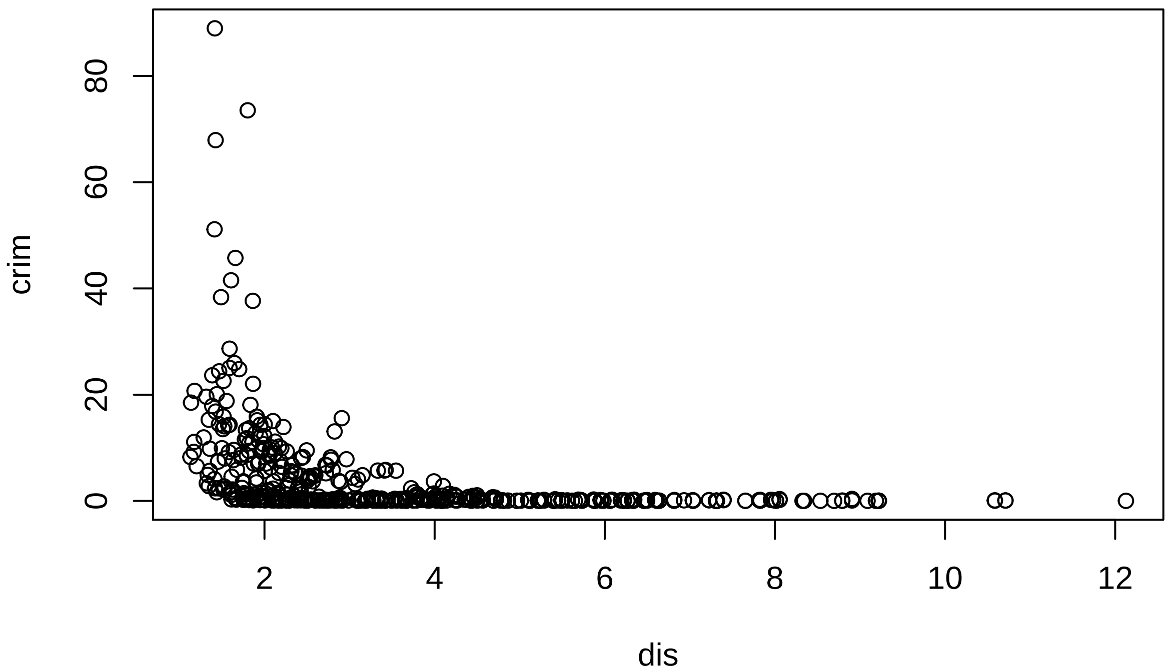
```
## [1] 506  14
```

```r
?Boston
```

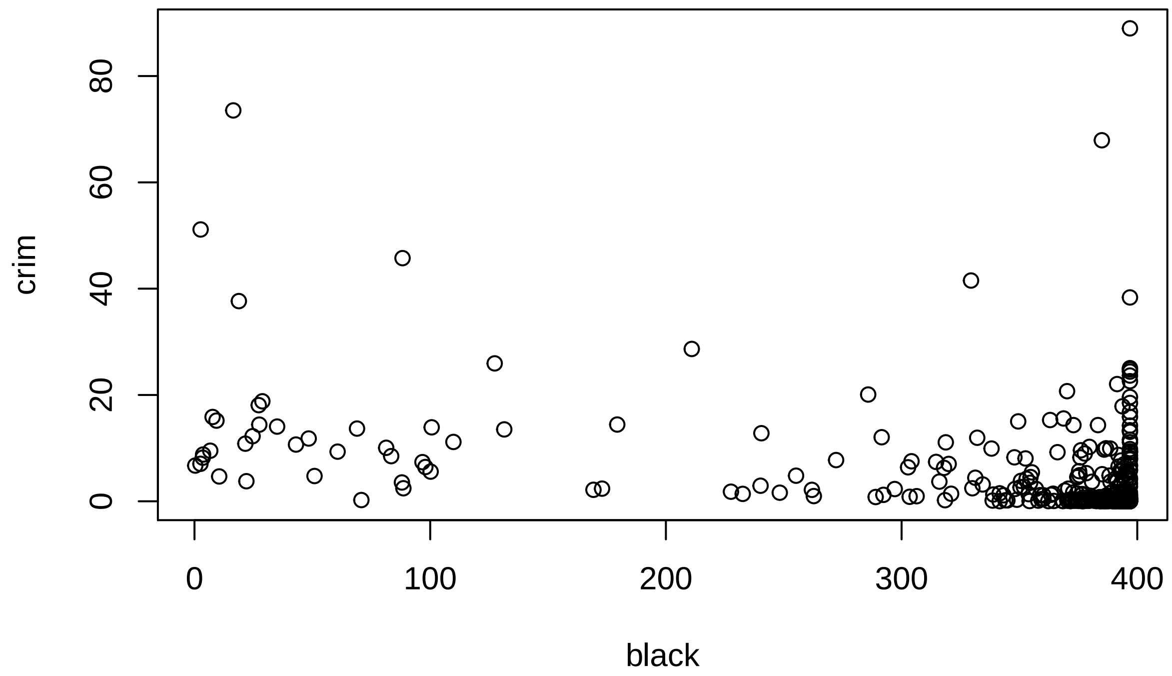
```
## starting httpd help server ... done
```

506 rows and 14 columns. Each row contains an observation at the town level, while the columns represent town level variables.

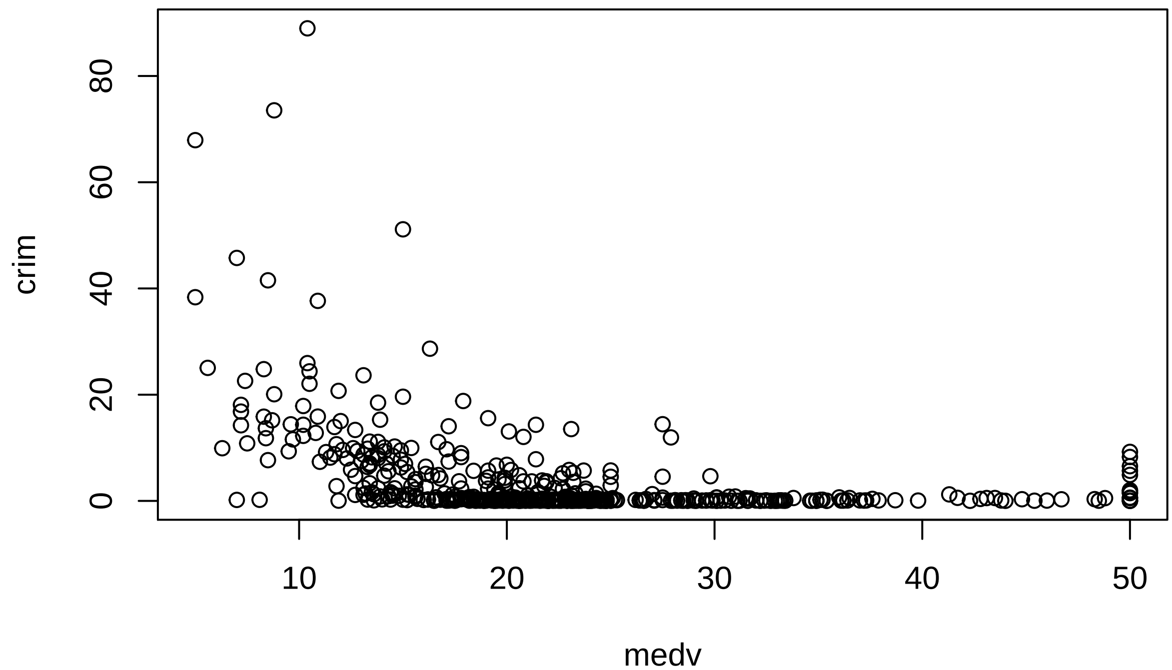
```r
# Code for 10 b) goes here
attach(Boston)
plot(dis,crim)
```




```
plot(black,crim)
```




```r
plot(medv,crim)
```



It appears that towns very close to employment centres have high crim rate, but this drops sharply once the distance increases slightly.

There also seems to be no correlation between the presence of black minority and crime rates.

Finally, as the median value of the property in the town increases, the crime rate decreases.

```r
# Code for 10 c) goes here
summary(lm(crim~.,data=Boston))
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
```

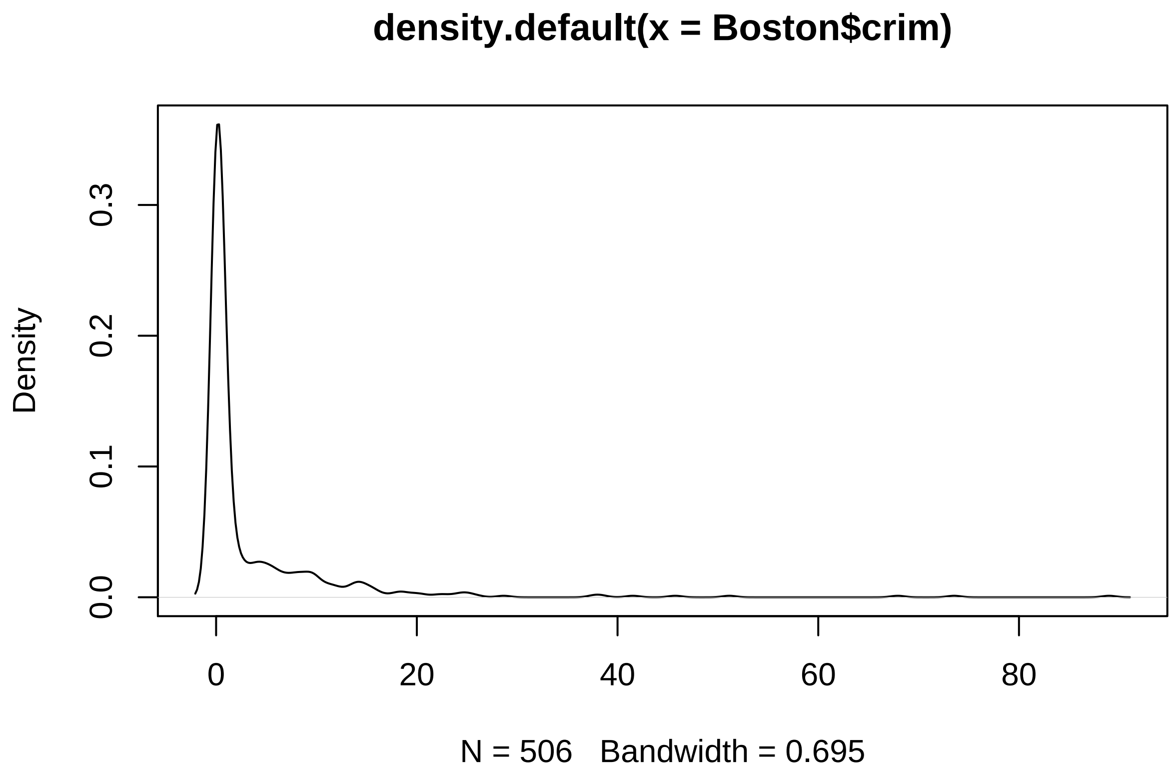
```
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Proportion of residential land zoned for lots over 25,000 sq.ft and an index of accessibility to radial highways are positively significantly associated with higher crime rates, while distance to employment centres, proportion of blacks, and the median value of owner occupied homes are negatively significantly associated with lower crime rates.

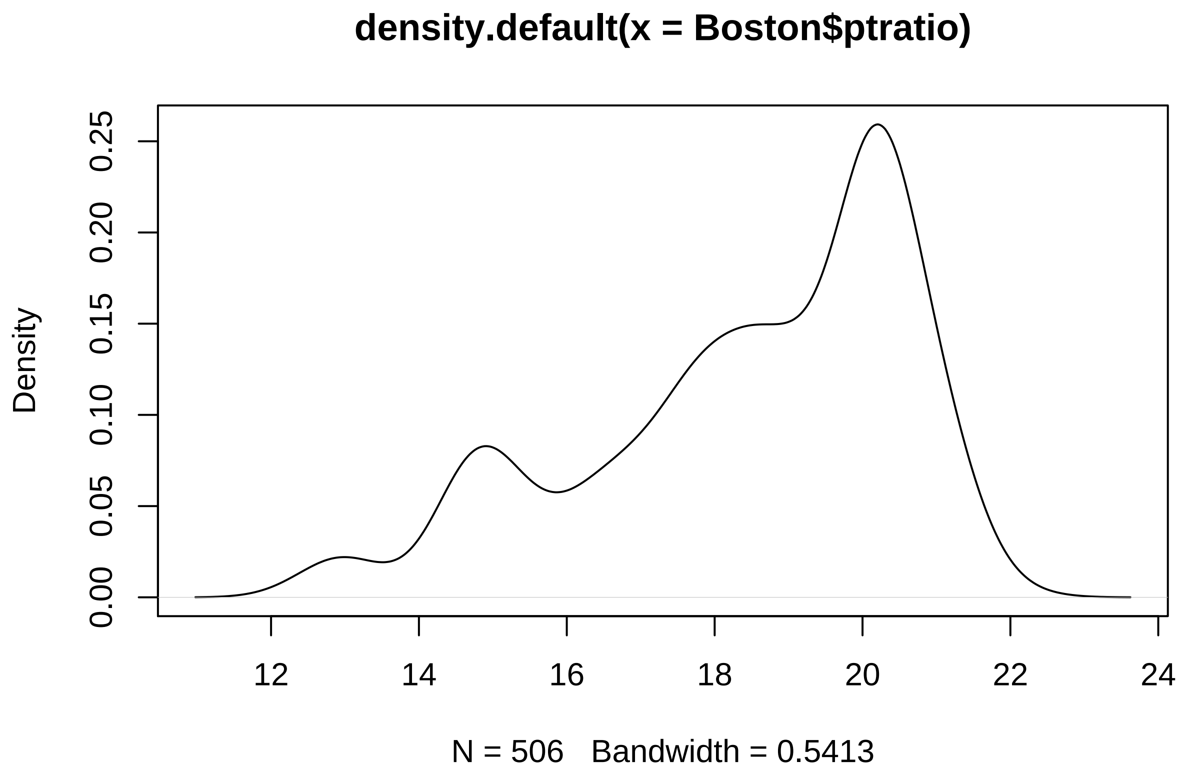
```r
# Code for 10 d) goes here
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

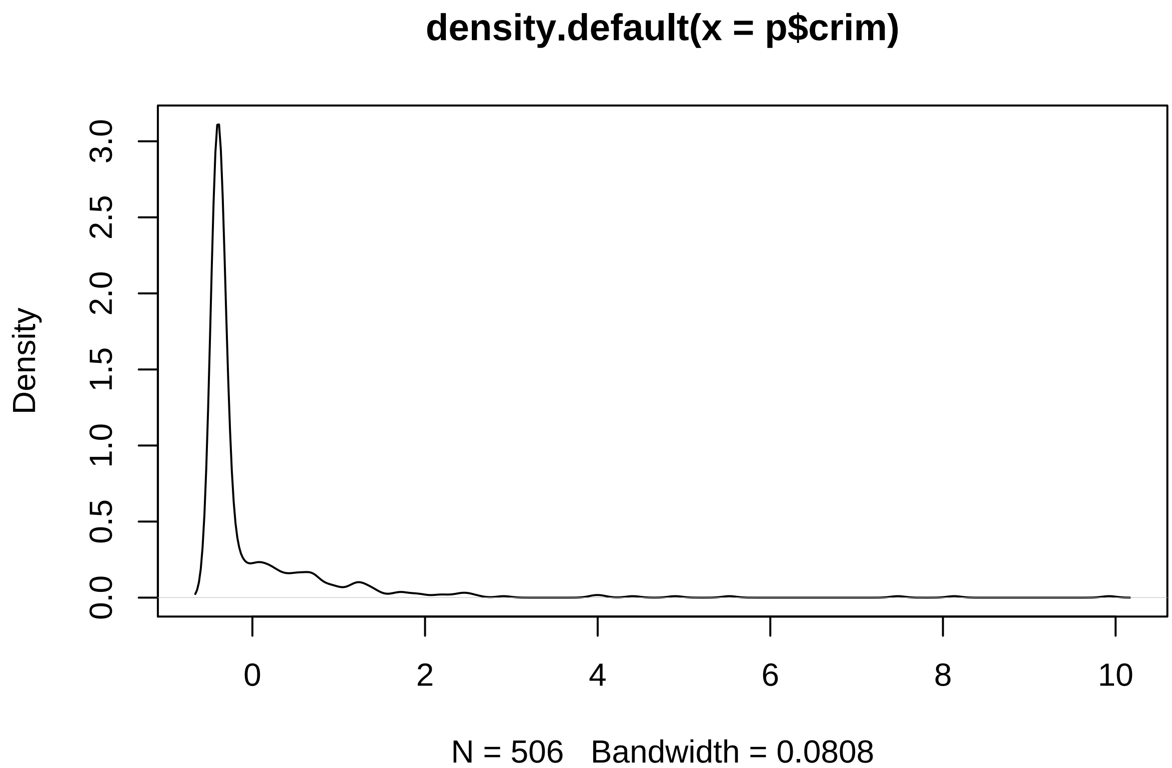
```r
plot(density(Boston$crim))
```



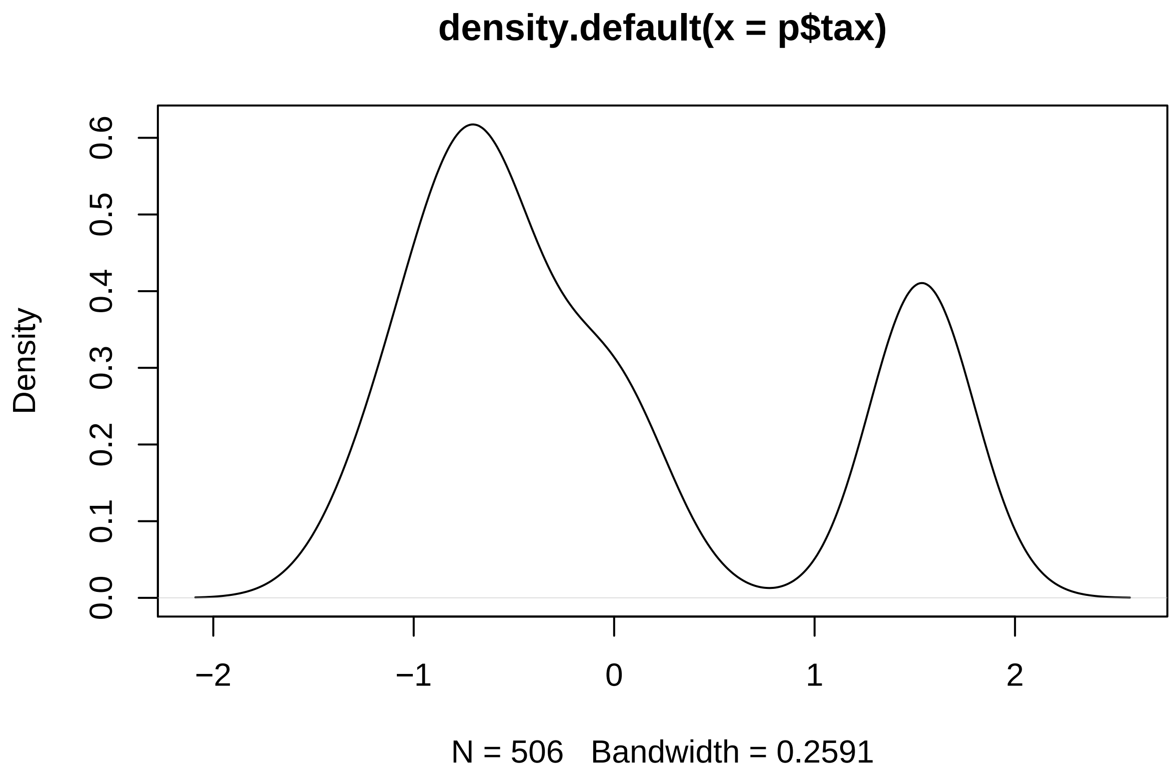

```r
plot(density(Boston$ptratio))
```



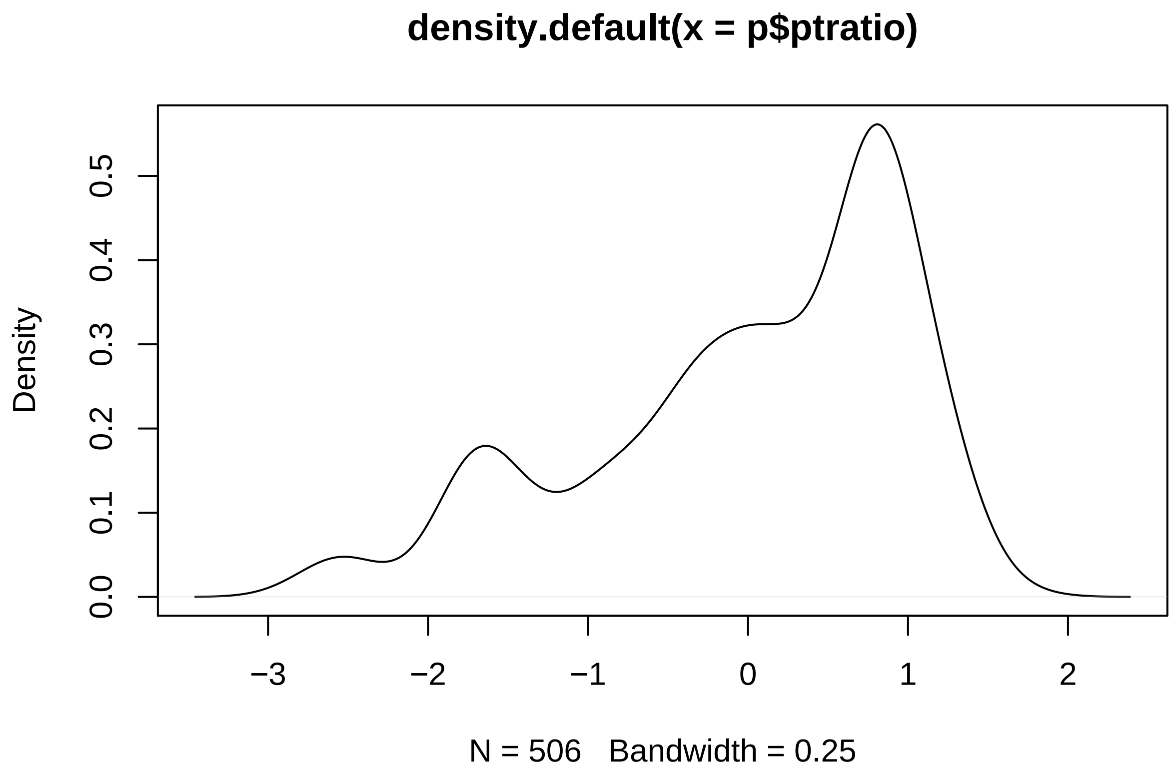

```
p<-scale(Boston) %>% as.data.frame()
plot(density(p$crim))
```




```
plot(density(p$tax))
```




```
plot(density(p$ptratio))
```



Some Boston suburbs are as far as 10 SD higher than the mean with regards to crime rate. As for tax rates, there are two significant peaks; about 60% of suburbs are between 0 to -1 SD away from the mean while another 40% are between 1-2 SD higher than the mean. Finally, the distribution of student pupil ratio is slightly left skewed with about 50% of the suburbs at 1 SD above the mean.

```r
# Code for 10 e) goes here
sum(Boston$chas)
```

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

There are 35 suburbs that bound the Charles river

```r
# Code for 10 f) goes here
summary(Boston)
```

```
##       crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   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
```

The median student teacher ratio in this dataset is 19.05.

```r
# Code for 10 g) goes here
ind<-which(Boston$medv==min(Boston$medv))
Boston[ind,]
```

```
##        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    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
```

```r
summary(Boston)
```

```
##       crim                zn             indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   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
```

There are two suburbs that have the minimum median value of owner occupied homes. Both these suburbs have 0 proportion of residential land zoned for lots over 25,000 sq.ft, do not bound the Charles river, are in the 4th quartile of proportion of non-retail business acres per town and nitrogen oxide concentration, in the 1st quartile of average number of rooms per dwelling, are completely built prior to 1940 for owner occupied homes, are very close to employment centres (1st quartile), are the highest in the dataset in terms of accessibility to radial highways, are verging on the 4th quartile when it comes to property tax, have a student teacher ratio that's verging on the 4th quartile, are in the 3rd and 4th quartile respectively in terms of proportion of blacks, are in the 4th quartile when it comes to lower status of the population.

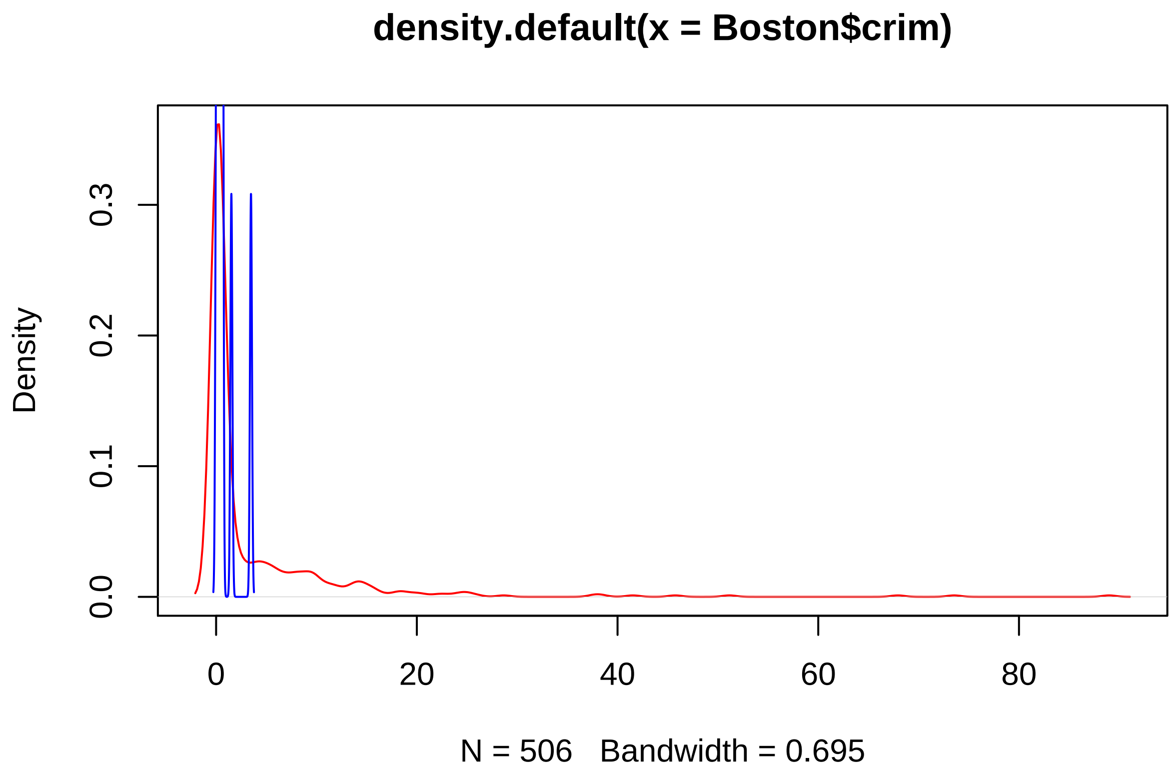
```r
# Code for 10 h) goes here
sum(Boston$rm>7)
```

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

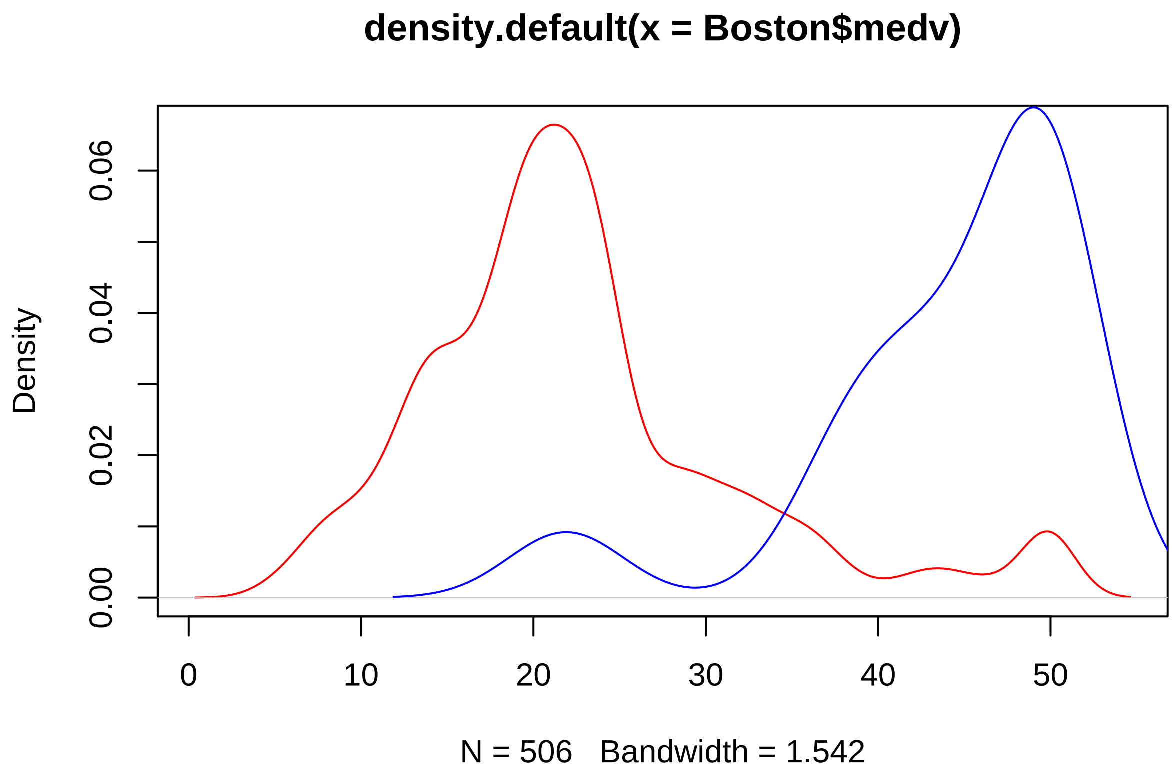
```r
sum(Boston$rm>8)
```

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

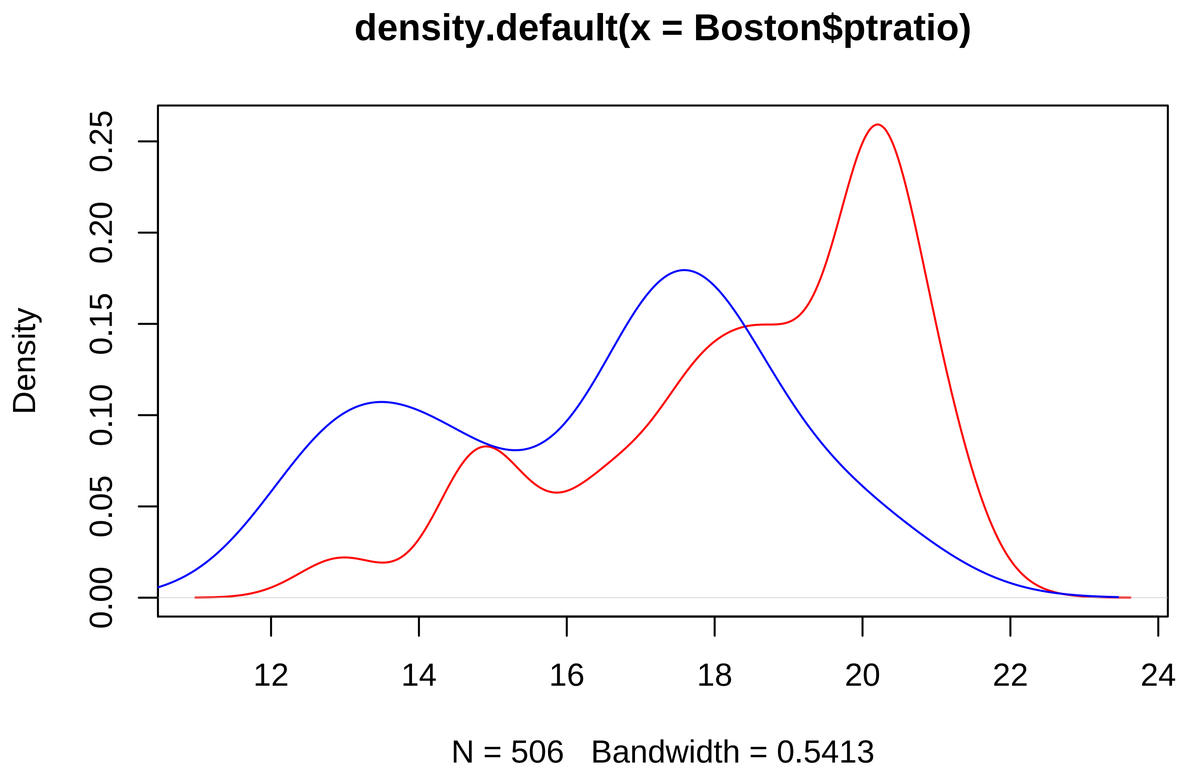
```r
plot(density(Boston$crim),col='red');lines(density(Boston%>%filter(rm>8)%>%.$crim),col='blue')
```



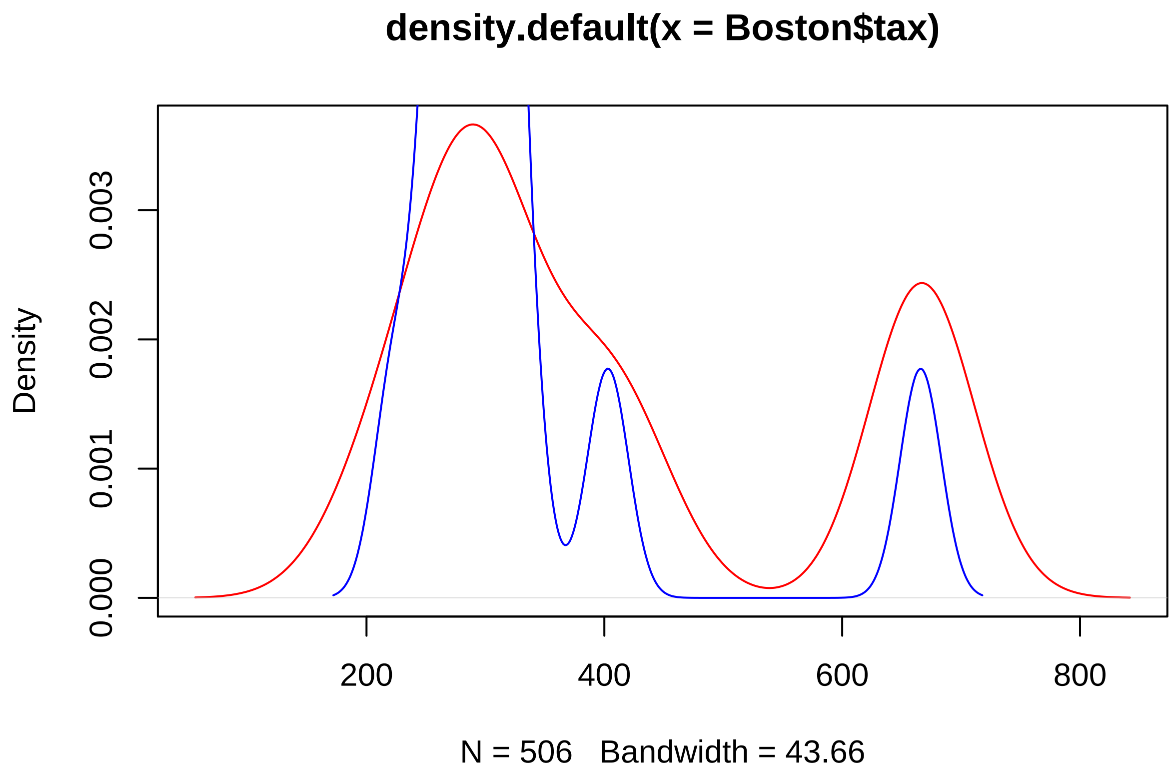

```
plot(density(Boston$medv),col='red');lines(density(Boston%>%filter(rm>8)%>%.$medv),col='blue')
```




```
plot(density(Boston$ptratio),col='red');lines(density(Boston%>%filter(rm>8)%>%.$ptratio),col='blue')
```




```r
plot(density(Boston$tax),col='red');lines(density(Boston%>%filter(rm>8)%>%.$tax),col='blue')
```



The number of suburbs with more than an average of 7 and 8 rooms per dwelling are 64 and 13 respectively. Crime statistics and taxes in suburbs with an average of more than 8 rooms per dwelling is similar to the larger dataset, but the median value of its houses are significantly higher. Its student teacher ratios are also lower,

8. Using R Markdown, write some notes on the differences between supervised and unsupervised approaches to statistical learning. Use headers of different sizes, italic and bold text, numbered lists, bullet lists, and hyperlinks. If you would like, use inline LaTeX (math notation).

Supervised vs Unsupervised Learning

Supervised learning refers to when there is a target response y and corresponding predictors. Unsupervised learning refers to when there is no target response, merely observations. In supervised learning what we are interested in is prediction, specifically prediction of y. If we have a set of y then we are able to train the data set and test it to compute the accuracy, In unsupervised learning like cluster analysis, we are not predicting anything. Instead, the algorithm looks at the data and sorts them into groups by looking for patterns.