null

Introduction:

Read the data

```
college<- read.csv("C:/Users/glane/Downloads/College.csv",stringsAsFactors=FALSE)
```

b

```
college$Private <- factor(college$Private)
rownames(college) = college[,1]
fix(college)
college = college[,-1]
fix(college)
```

C.1

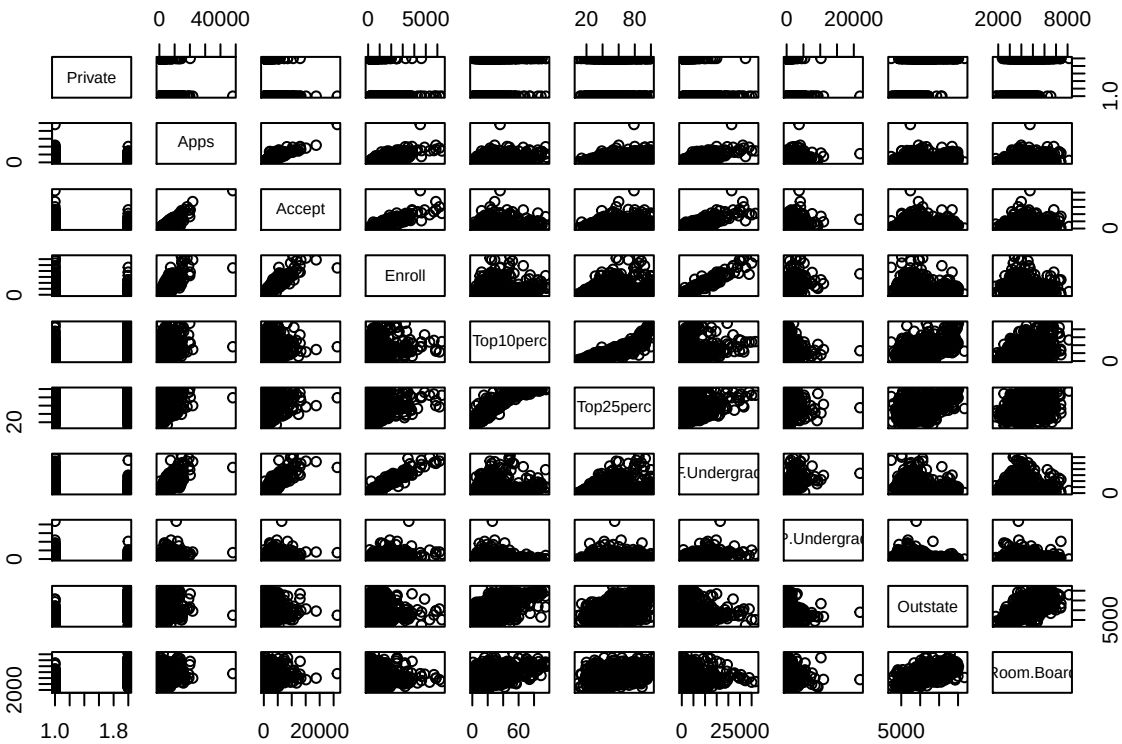
```
summary(college)
```

```
##  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
```

2.

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

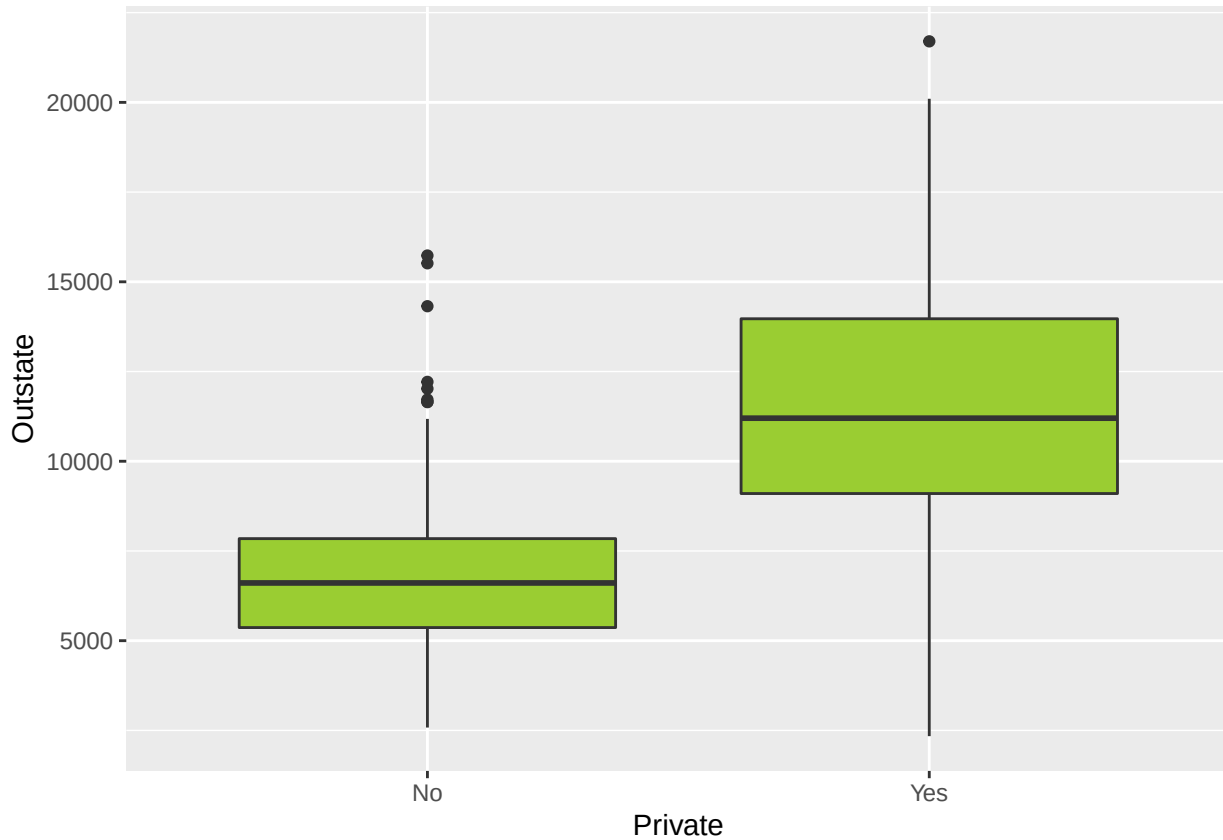


3.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```r
ggplot(data = college,aes(y=Outstate,x=Private))+geom_boxplot(fill="yellowgreen")
```



4.

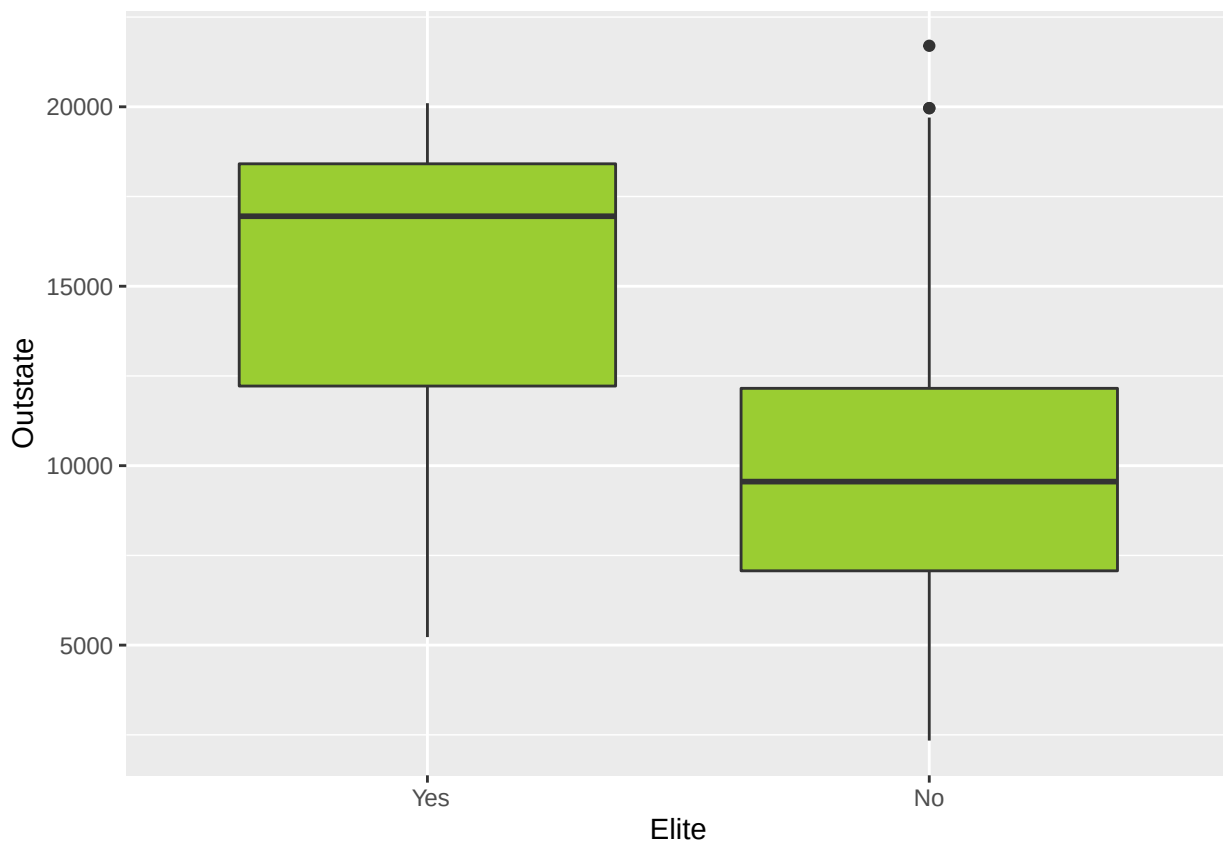
```
Elite=rep("No",nrow(college ))
Elite[college$Top10perc >50] = " Yes"
Elite=as.factor(Elite)
college=data.frame(college , Elite)
summary(college)
```

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

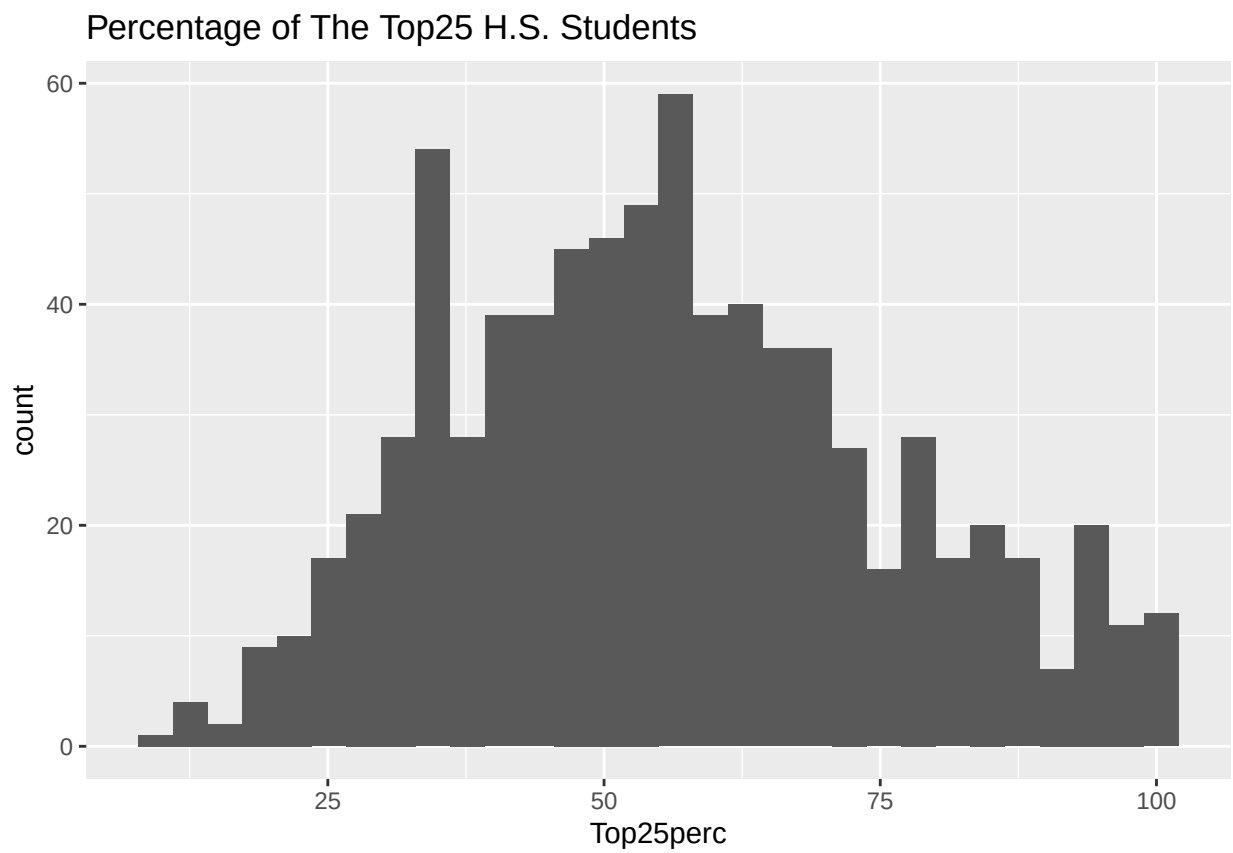
Elite uNiversities = 78

```
ggplot(data = college,aes(y=Outstate,x=Elite))+geom_boxplot(fill="yellowgreen")
```

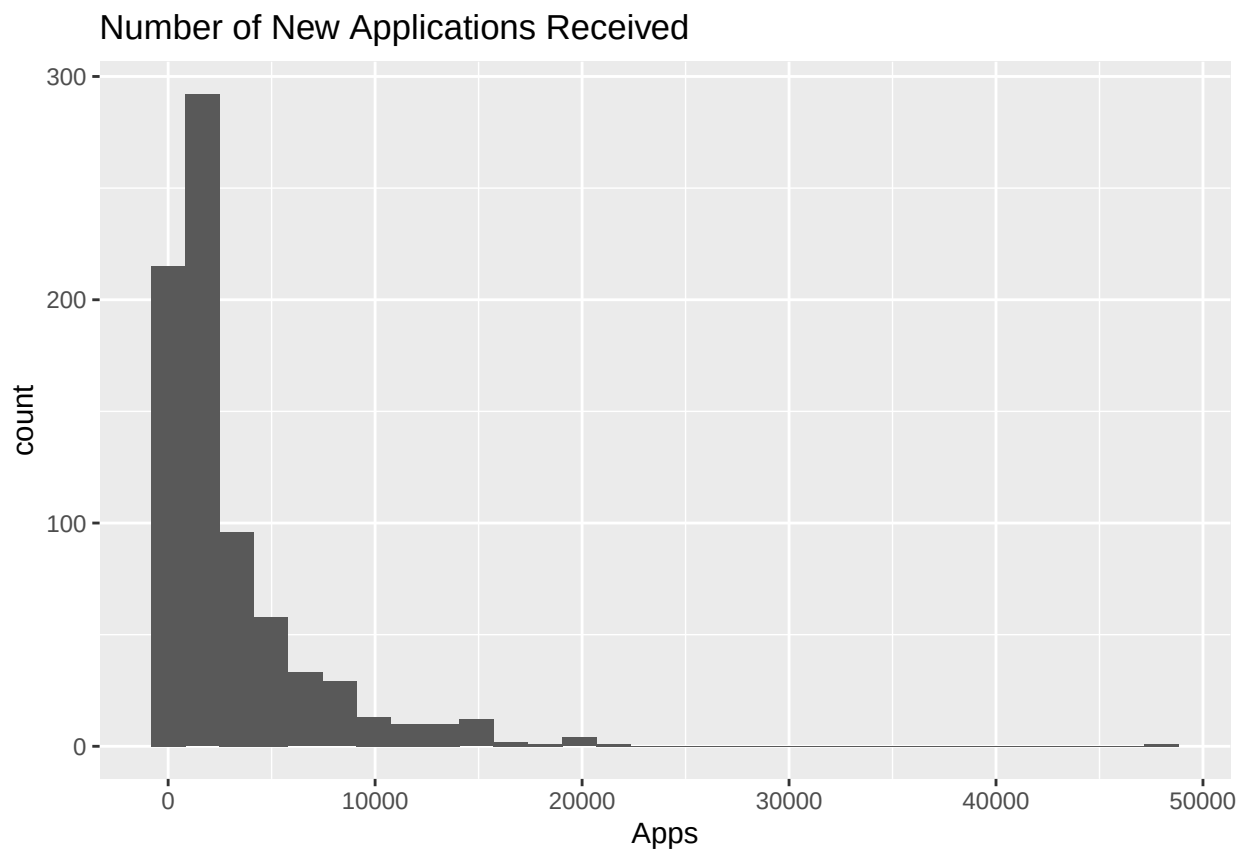


5.

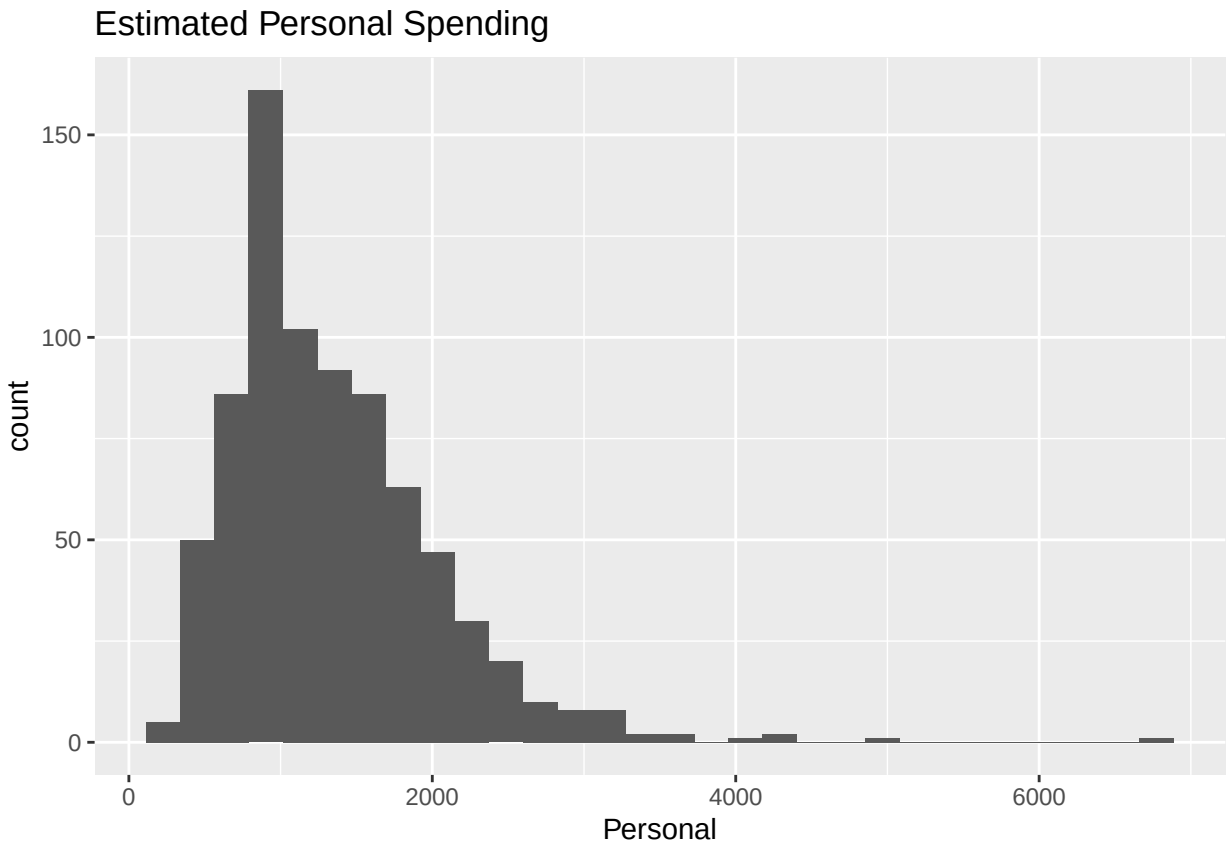
```
ggplot(data = college, aes(x=Top25perc))+geom_histogram(bins = 30)+labs(title = "Percentage of The Top25
```



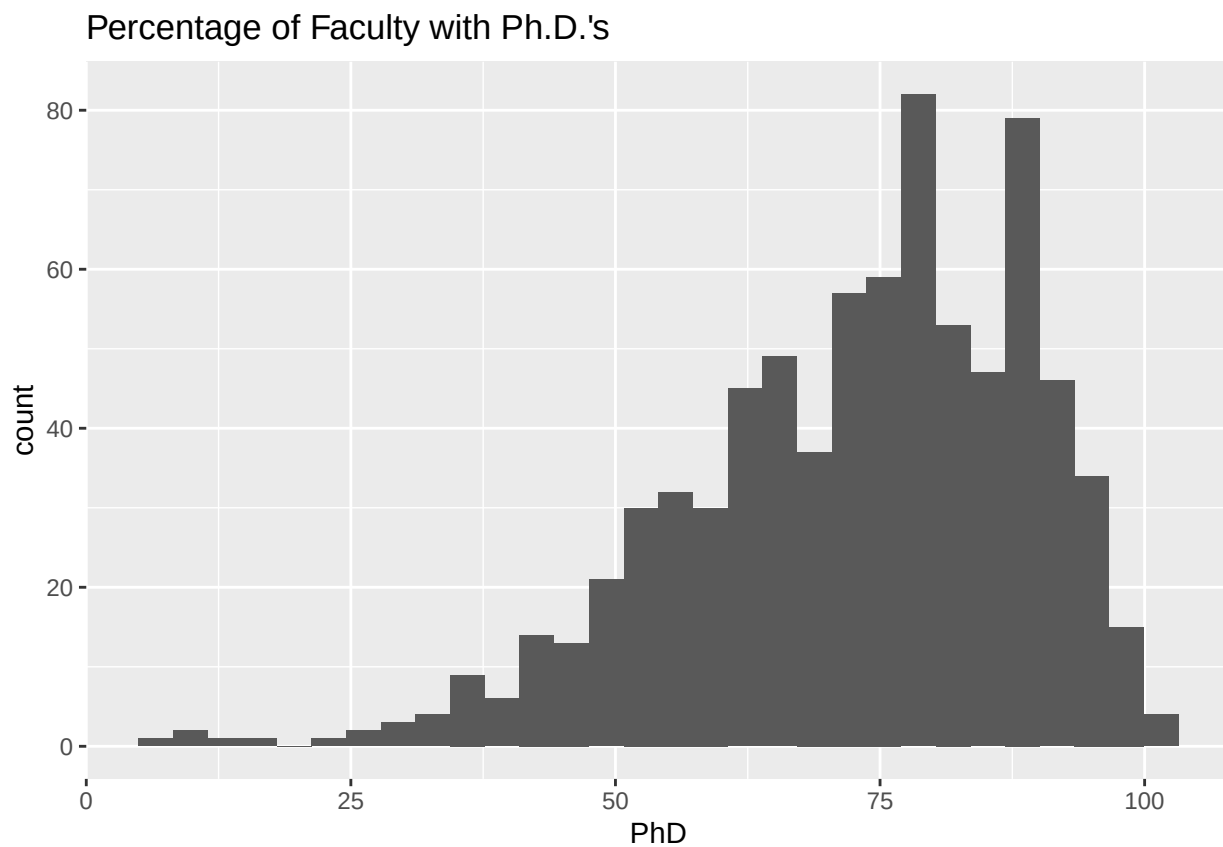

```
ggplot(data = college, aes(x=Apps))+geom_histogram(bins = 30)+labs(title = "Number of New Applications
```




```
ggplot(data = college, aes(x=Personal))+geom_histogram(bins = 30)+labs(title = "Estimated Personal Spen
```




```
ggplot(data = college, aes(x=PhD))+geom_histogram(bins = 30)+labs(title = "Percentage of Faculty with Ph
```



6.

```
summary(college$Personal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250     850    1200    1341    1700    6800
```

```
summary(college$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

There is a college with more than 100% of percentage

```
row.names(college)[which(college$PhD>100)]
```

```
## [1] "Texas A&M University at Galveston"
```

Auto data

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.3
```

```
data("Auto")
auto <- na.omit(Auto)
```

a.

```
lapply(auto, class)
```

```
## $mpg
```

```
## [1] "numeric"
##
## $cylinders
## [1] "numeric"
##
## $displacement
## [1] "numeric"
##
## $horsepower
## [1] "numeric"
##
## $weight
## [1] "numeric"
##
## $acceleration
## [1] "numeric"
##
## $year
## [1] "numeric"
##
## $origin
## [1] "numeric"
##
## $name
## [1] "factor"
```

The column name is the only not numeric, therefore it is a qualitative, it is seen that the origin column is qualitative, factors described as numbers. The other columns are all quantitatives.

```r
auto$origin <- as.factor(auto$origin)
```

b.

```r
quant = names(auto) %in% c("name", "origin")
lapply(auto[, !quant], range)
```

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

c

```r
lapply(auto[, !quant], function(x){ c('mean'=mean(x), 'sd'=sd(x))})
```

```
## $mpg
##      mean        sd
## 23.445918  7.805007
##
## $cylinders
##     mean       sd
## 5.471939 1.705783
##
## $displacement
##    mean      sd
## 194.412 104.644
##
## $horsepower
##      mean        sd
## 104.46939  38.49116
##
## $weight
##      mean        sd
## 2977.5842  849.4026
##
## $acceleration
##      mean        sd
## 15.541327  2.758864
##
## $year
##      mean        sd
## 75.979592  3.683737
```

d

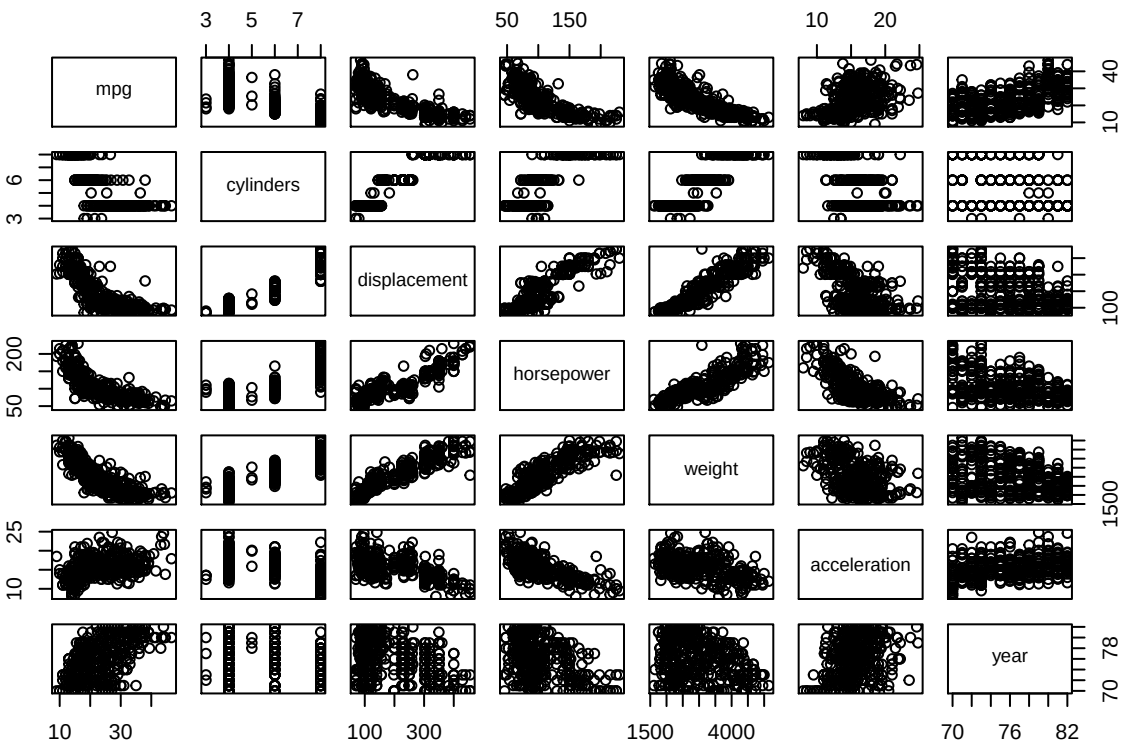
```r
lapply(auto[-(10:85),!quant], function(x){ c('mean'=mean(x), 'sd'=sd(x))})
```

```
## $mpg
##      mean        sd
## 24.404430  7.867283
##
## $cylinders
##     mean       sd
## 5.373418 1.654179
##
## $displacement
##      mean        sd
## 187.24051  99.67837
##
## $horsepower
##      mean        sd
## 100.72152  35.70885
##
## $weight
##      mean        sd
## 2935.9715  811.3002
##
```

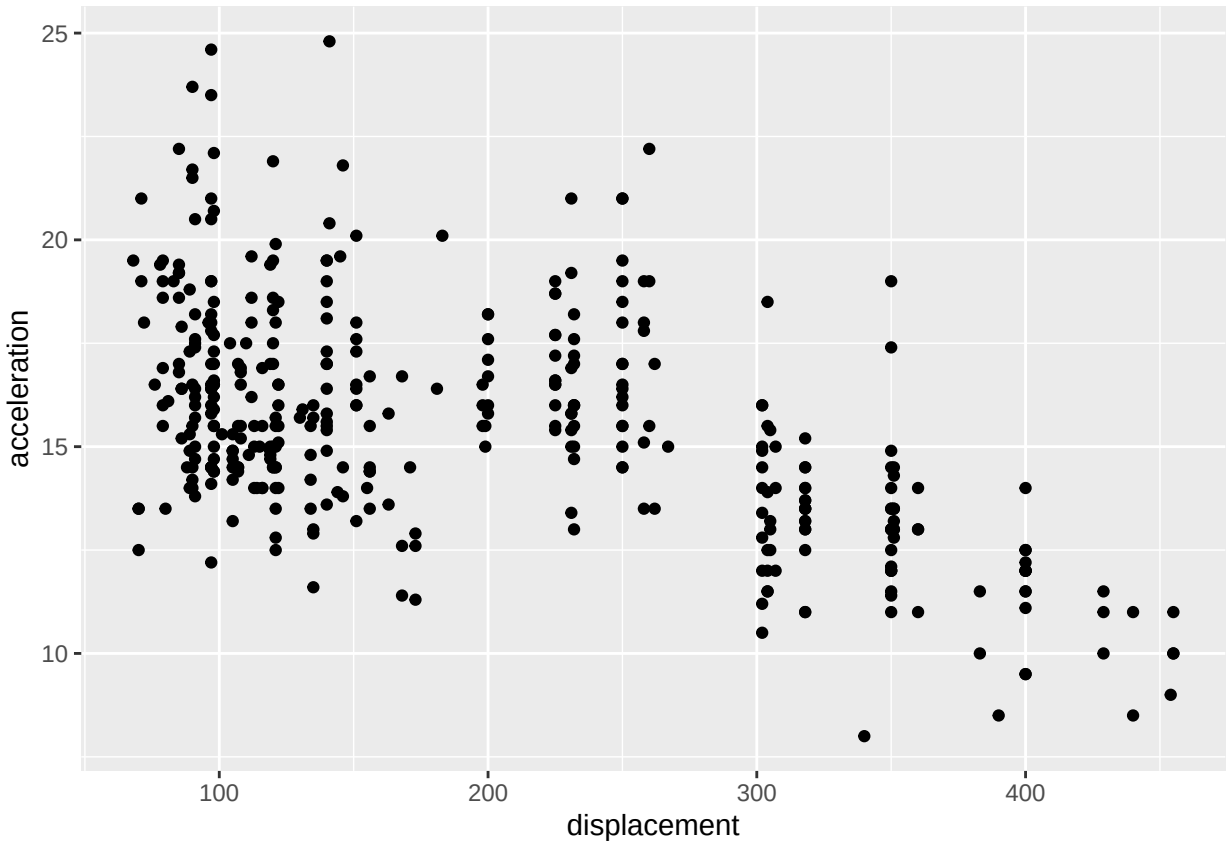
```
## $acceleration
##      mean        sd
## 15.726899  2.693721
##
## $year
##      mean        sd
## 77.145570  3.106217
```

e

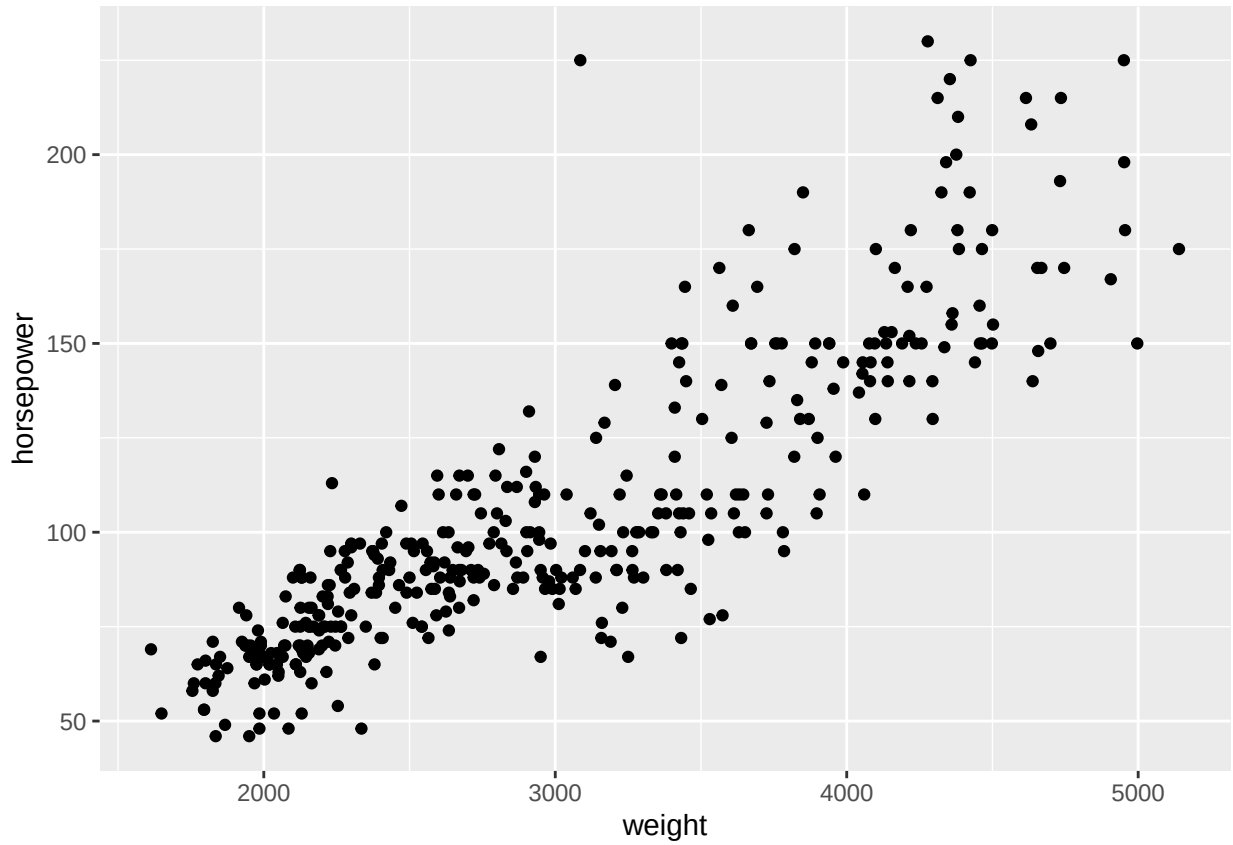
```r
pairs(auto[, !quant])
```



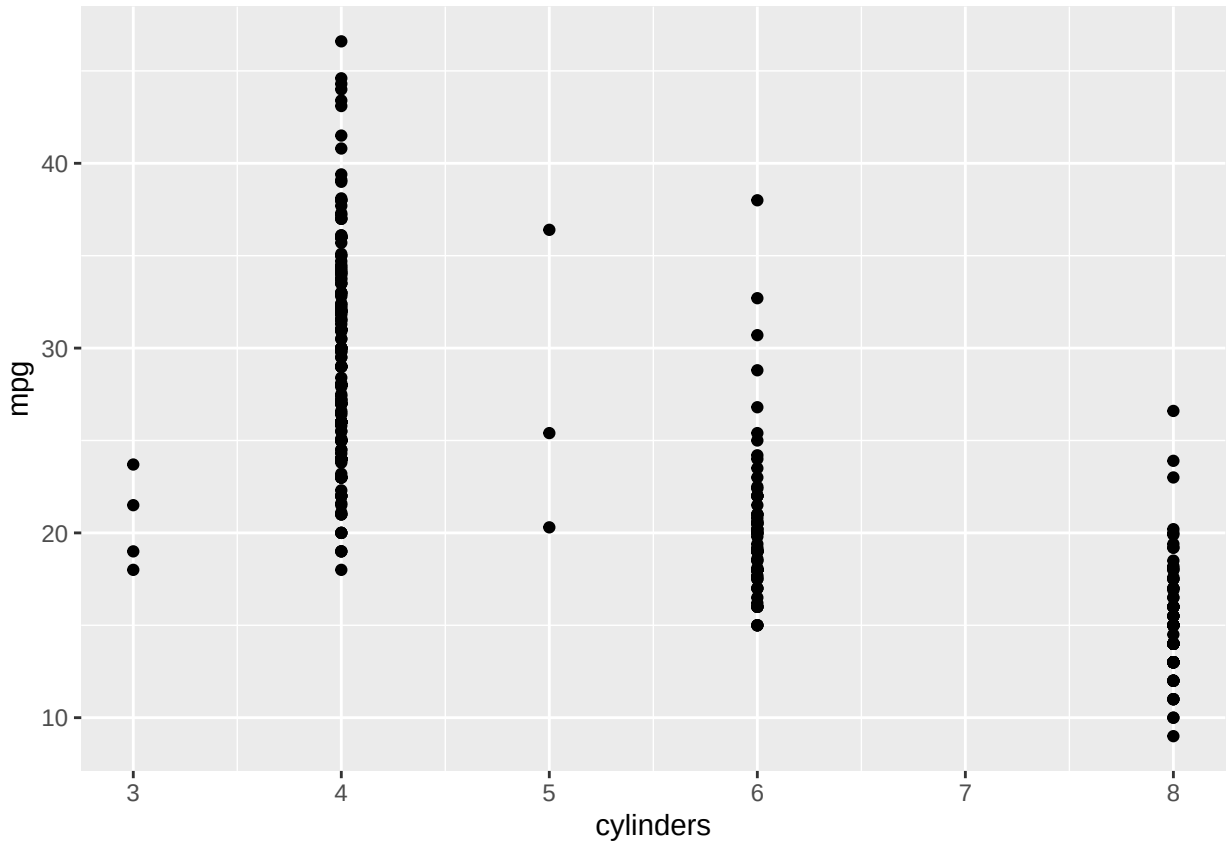

```r
ggplot(data = auto,aes(y=acceleration,x=displacement))+geom_point()
```



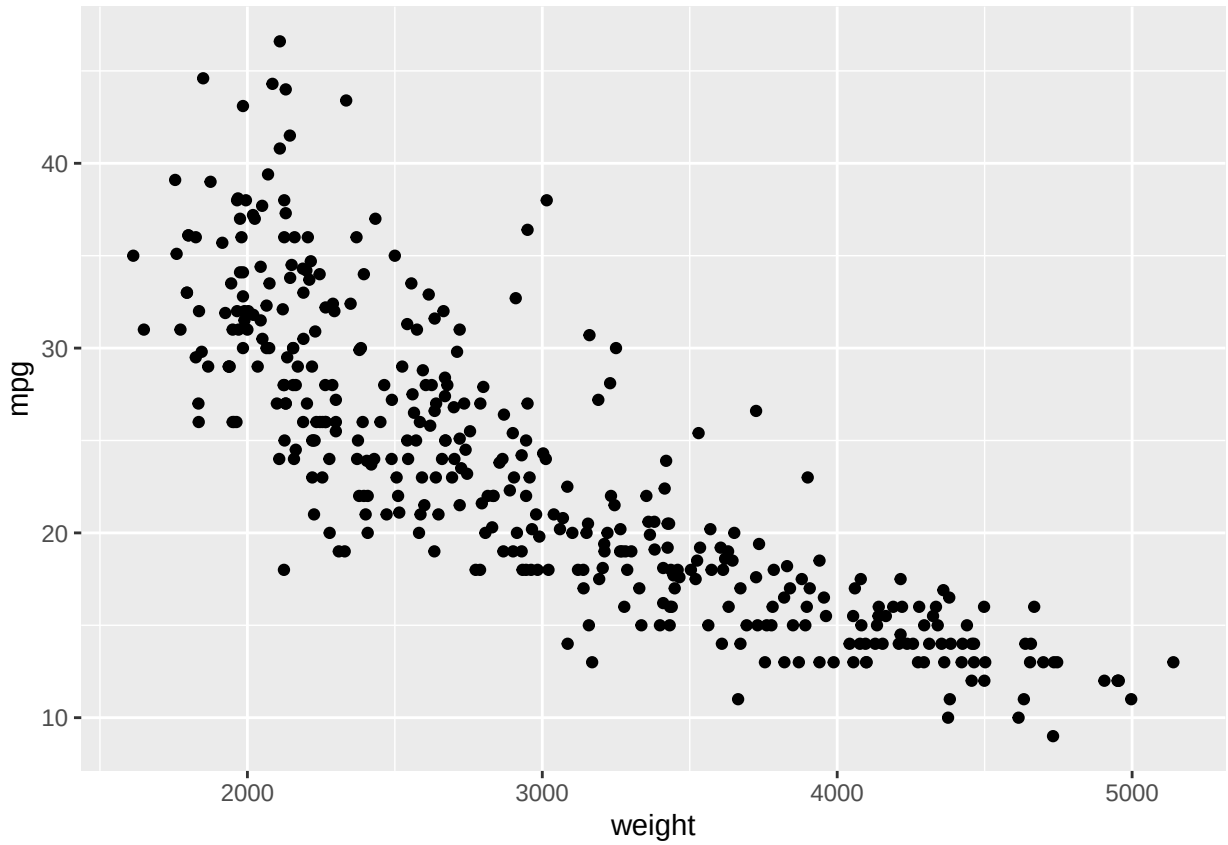

```
ggplot(data = auto,aes(y=horsepower,x=weight))+geom_point()
```




```
ggplot(data = auto,aes(y=mpg,x=cylinders))+geom_point()
```

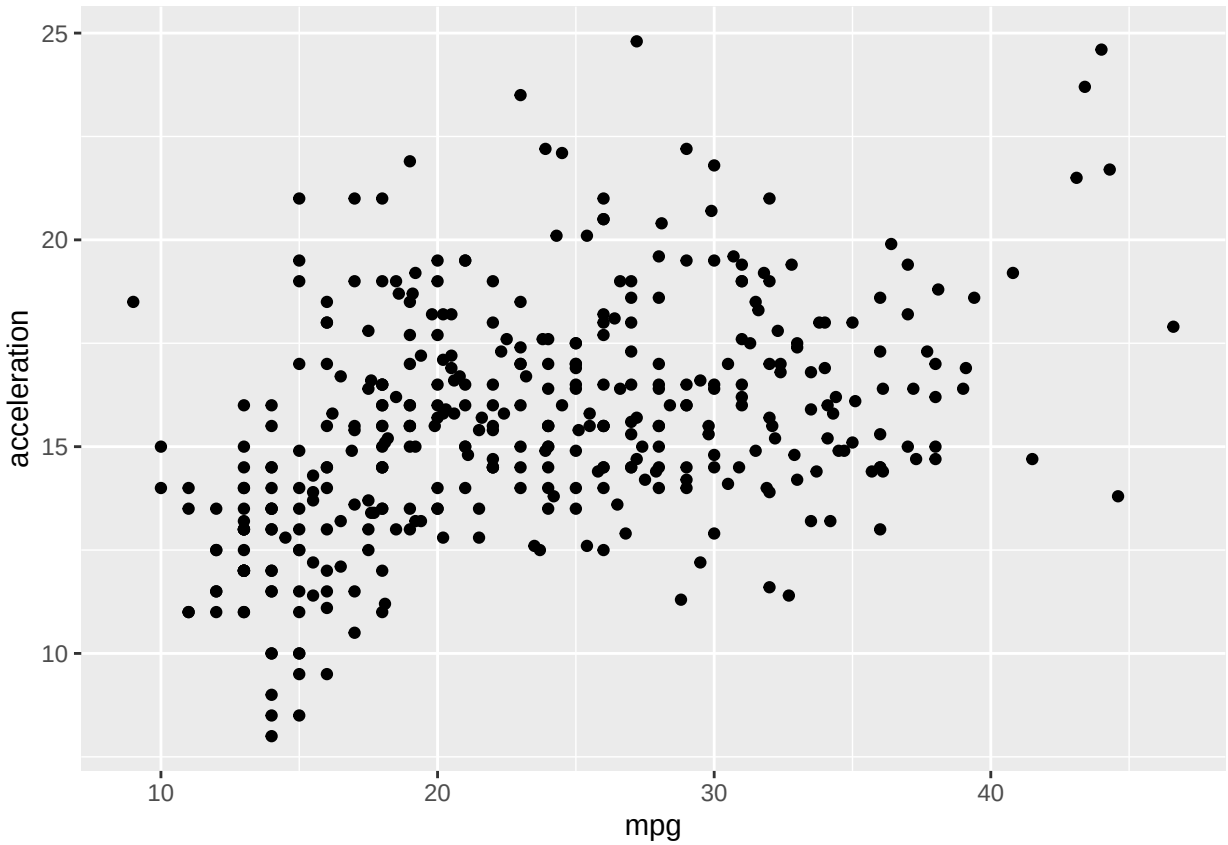



```
ggplot(data = auto,aes(y=mpg,x=weight))+geom_point()
```

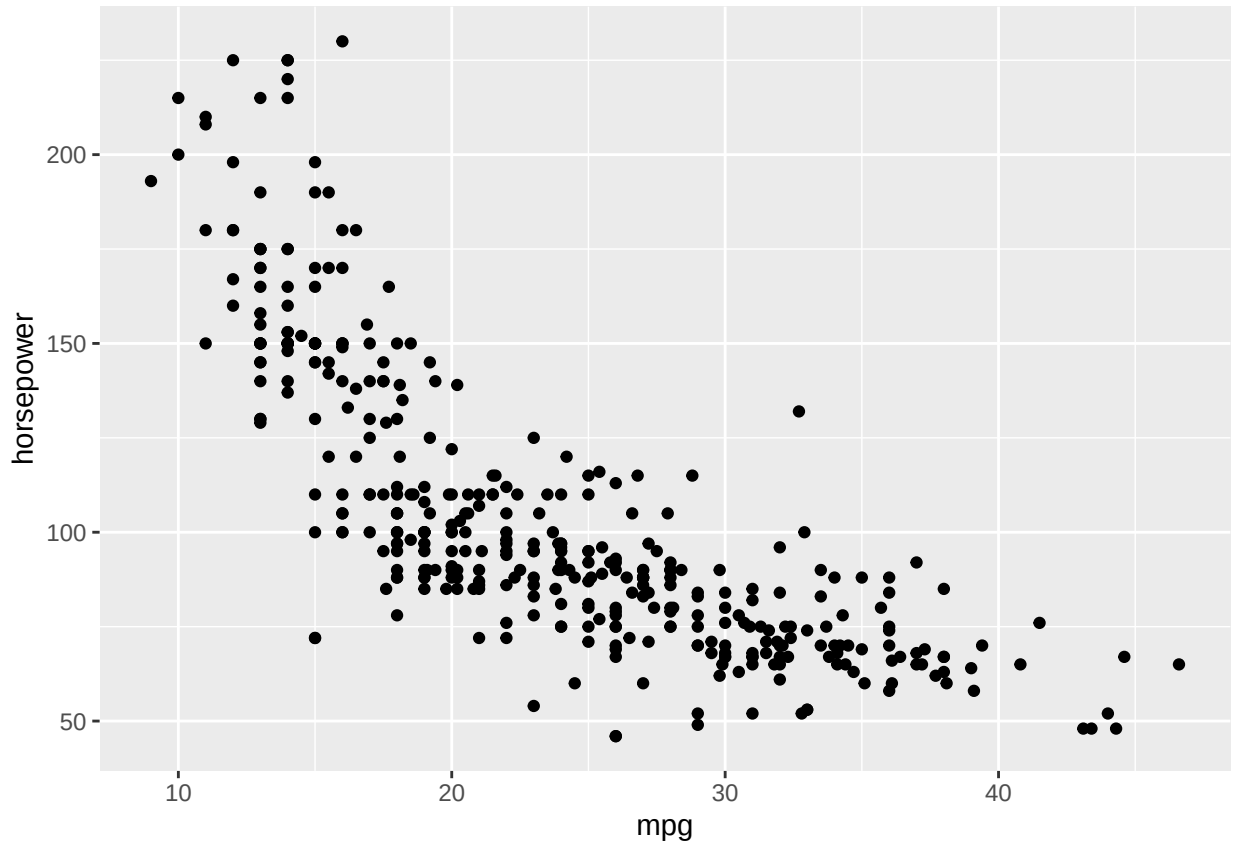


mpg tends to decrease as cylinders or weight increases, whereas horsepower increases as weight increases, and acceleration and displacement have a negative correlation.

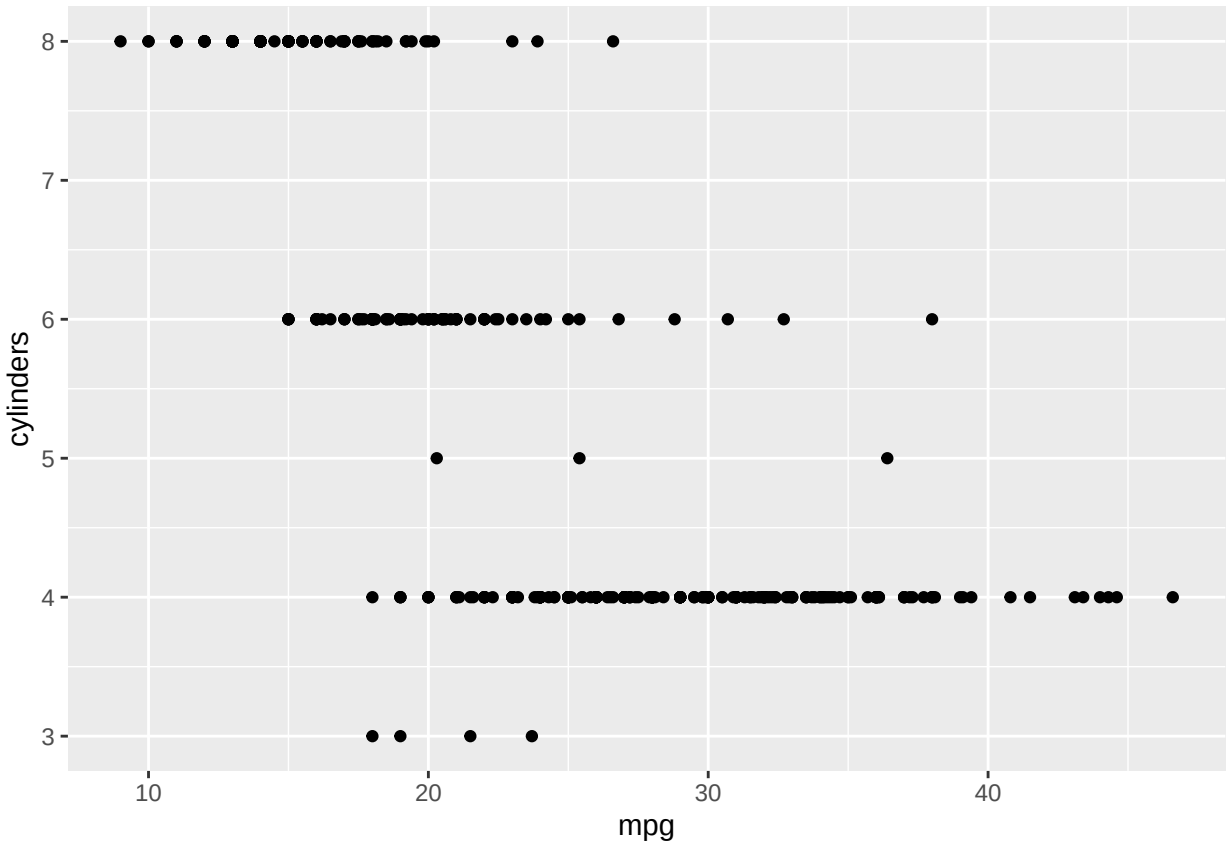
```
ggplot(data = auto,aes(y=acceleration,x=mpg))+geom_point()
```



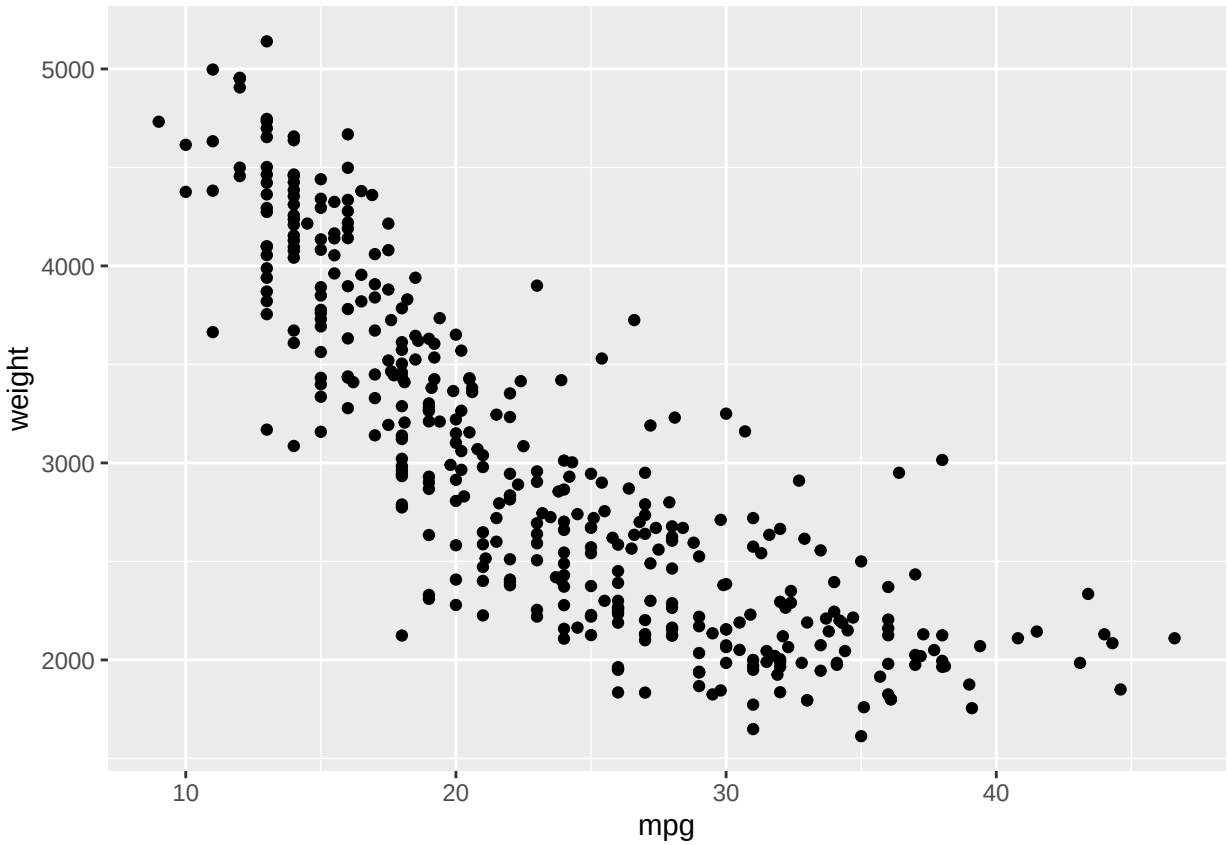

```
ggplot(data = auto,aes(y=horsepower,x=mpg))+geom_point()
```



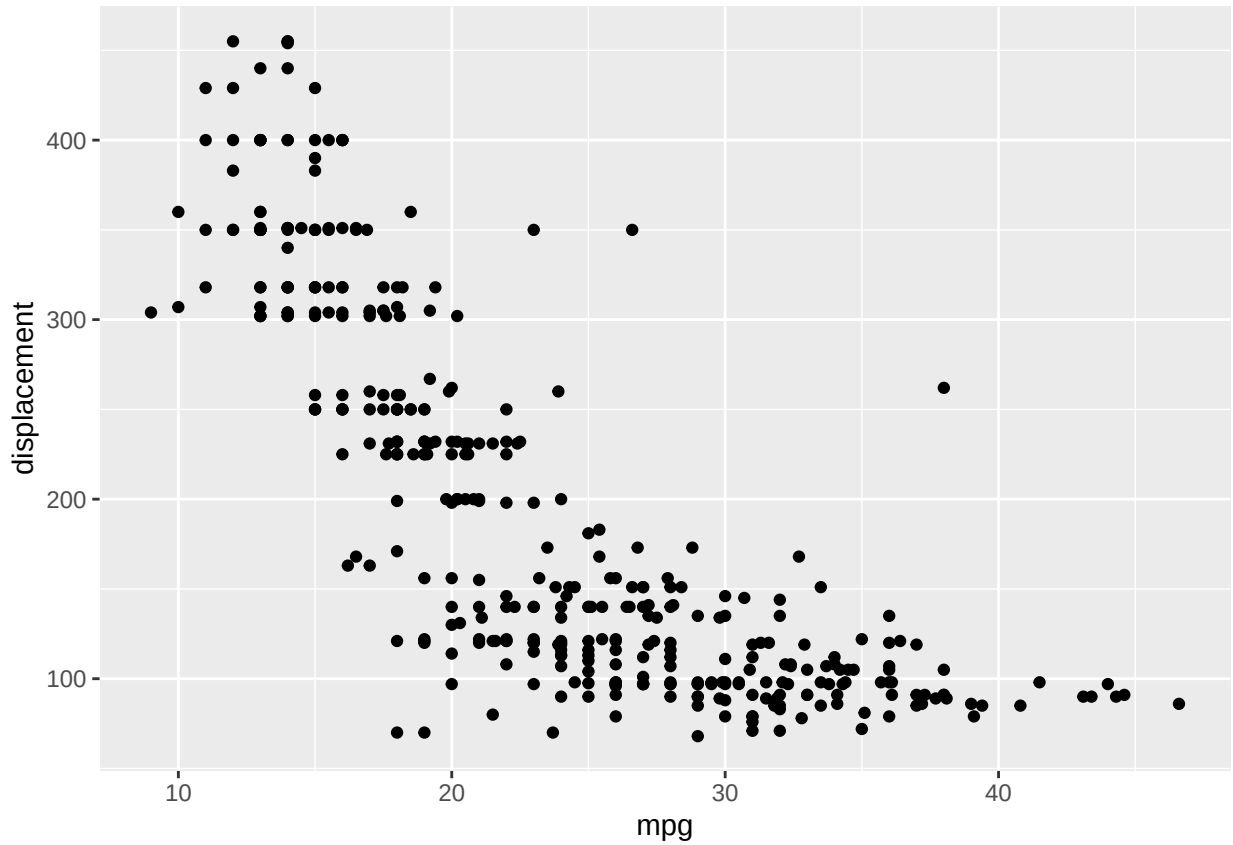

```
ggplot(data = auto,aes(y=cylinders,x=mpg))+geom_point()
```




```
ggplot(data = auto,aes(y=weight,x=mpg))+geom_point()
```




```
ggplot(data = auto,aes(y=displacement,x=mpg))+geom_point()
```



The weights, displacement and horsepower decreases with increase in mpg.

Boston housing data set a

```
library(MASS)
data("Boston")
?Boston
```

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

```
nrow(Boston)
```

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

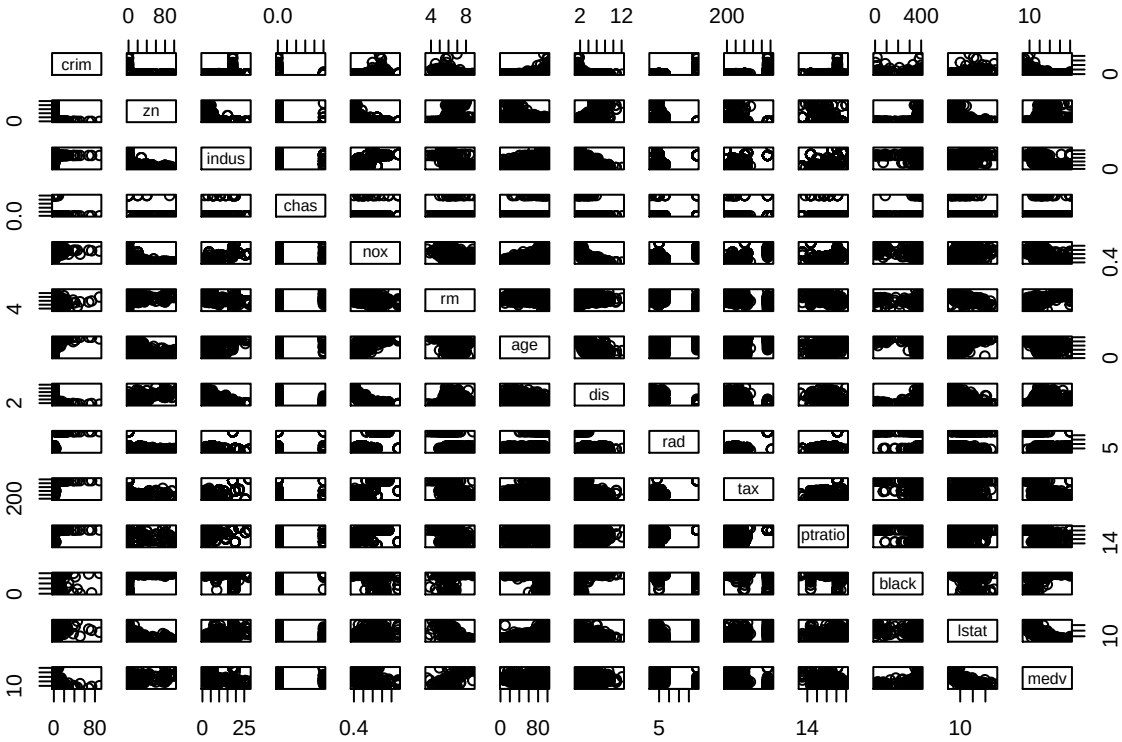
```
ncol(Boston)
```

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

The rows represent observations of the U.S. Census Tracts in the Boston Area. The columns presents the measures of the Census Variables.

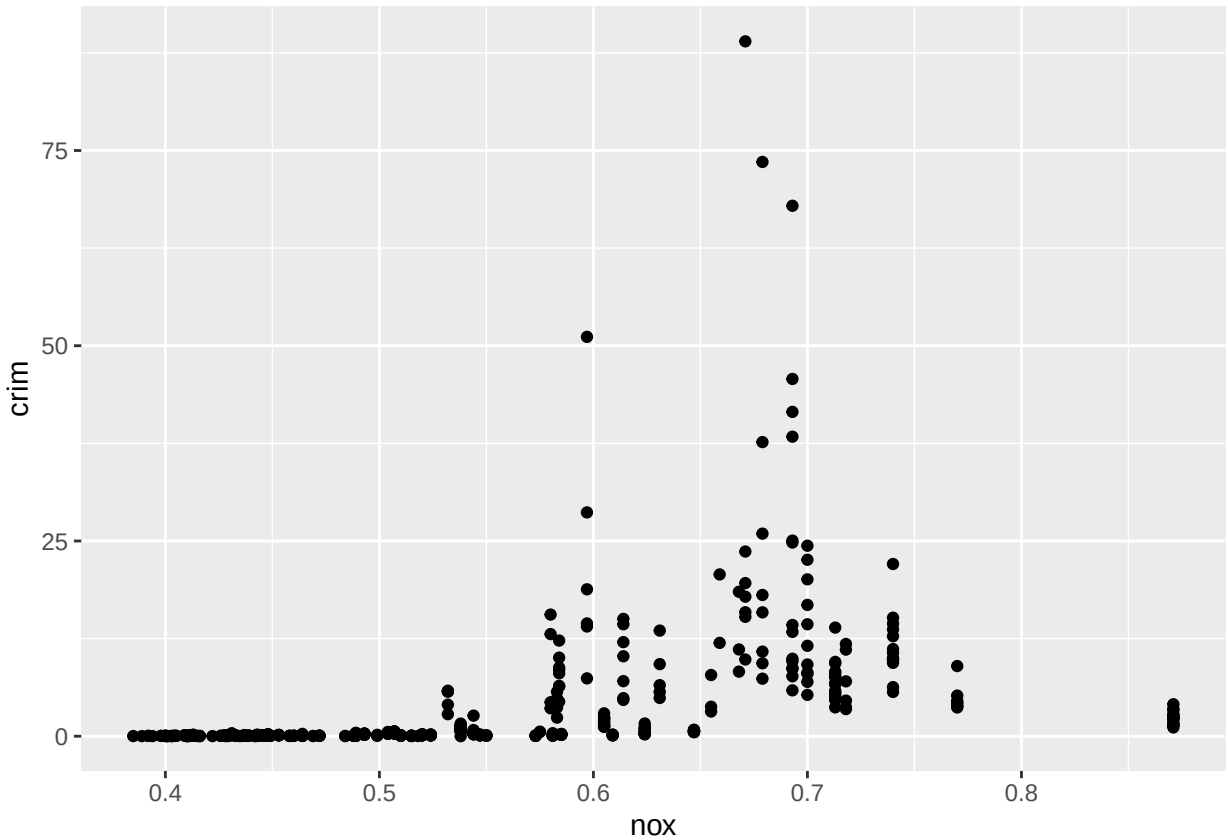
b

```
pairs(Boston)
```

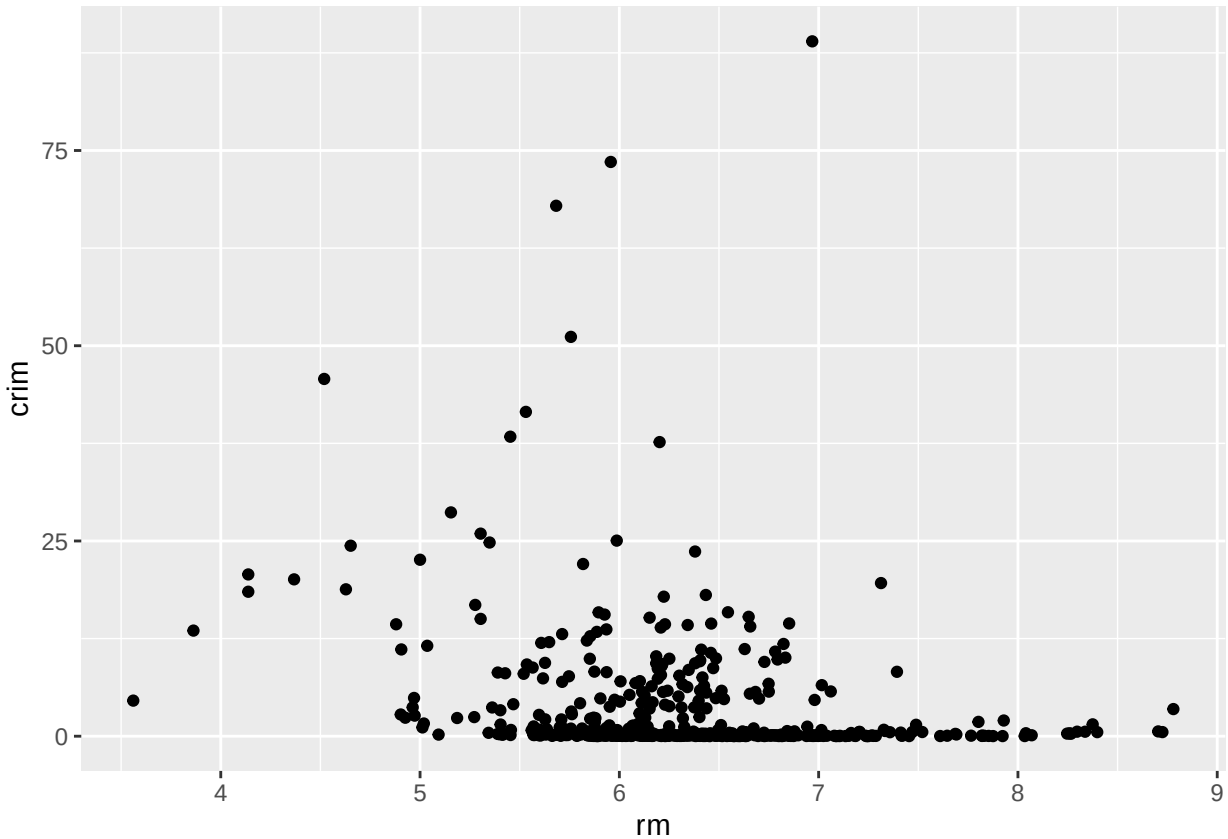


zn and distance have a negative relationship with crime. Age seems to have a positive relationship.

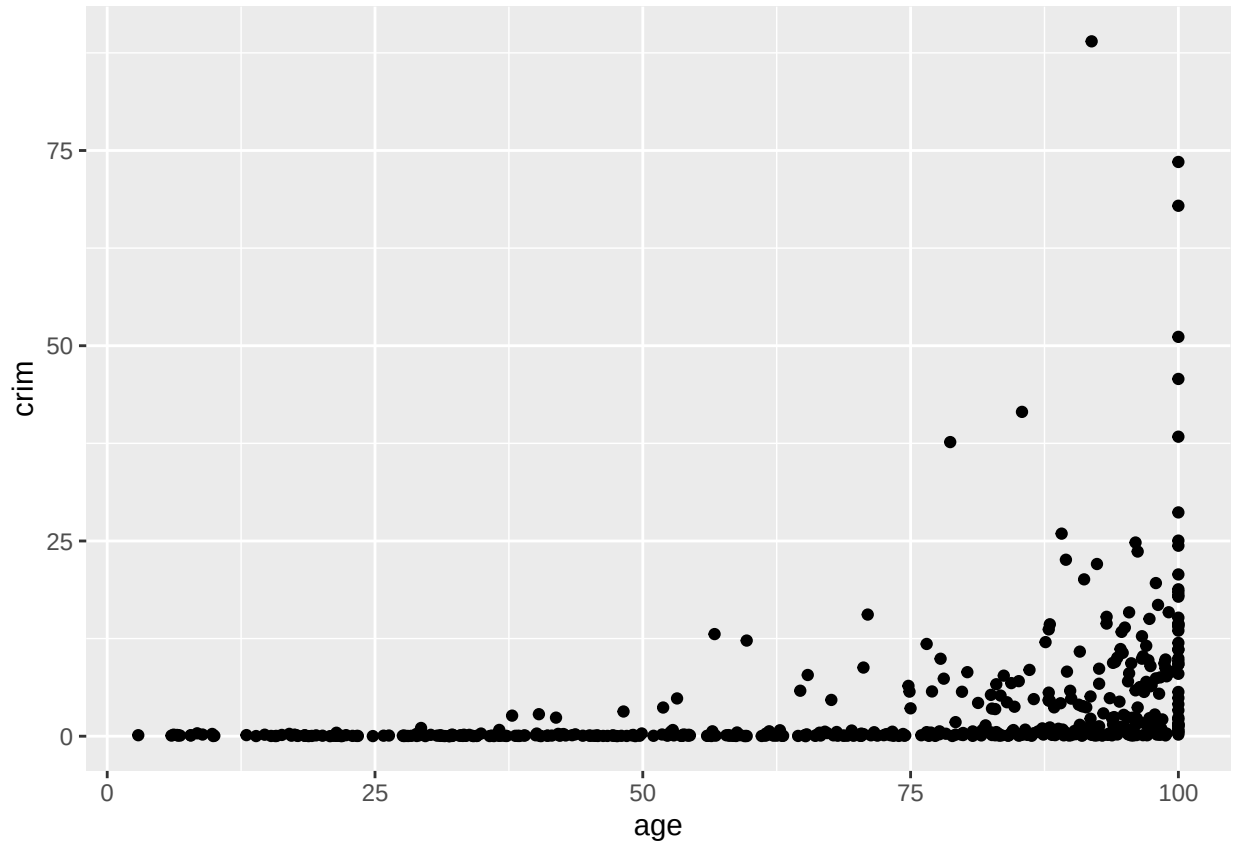
```
ggplot(data = Boston,aes(y=crim,x=nox))+geom_point()
```



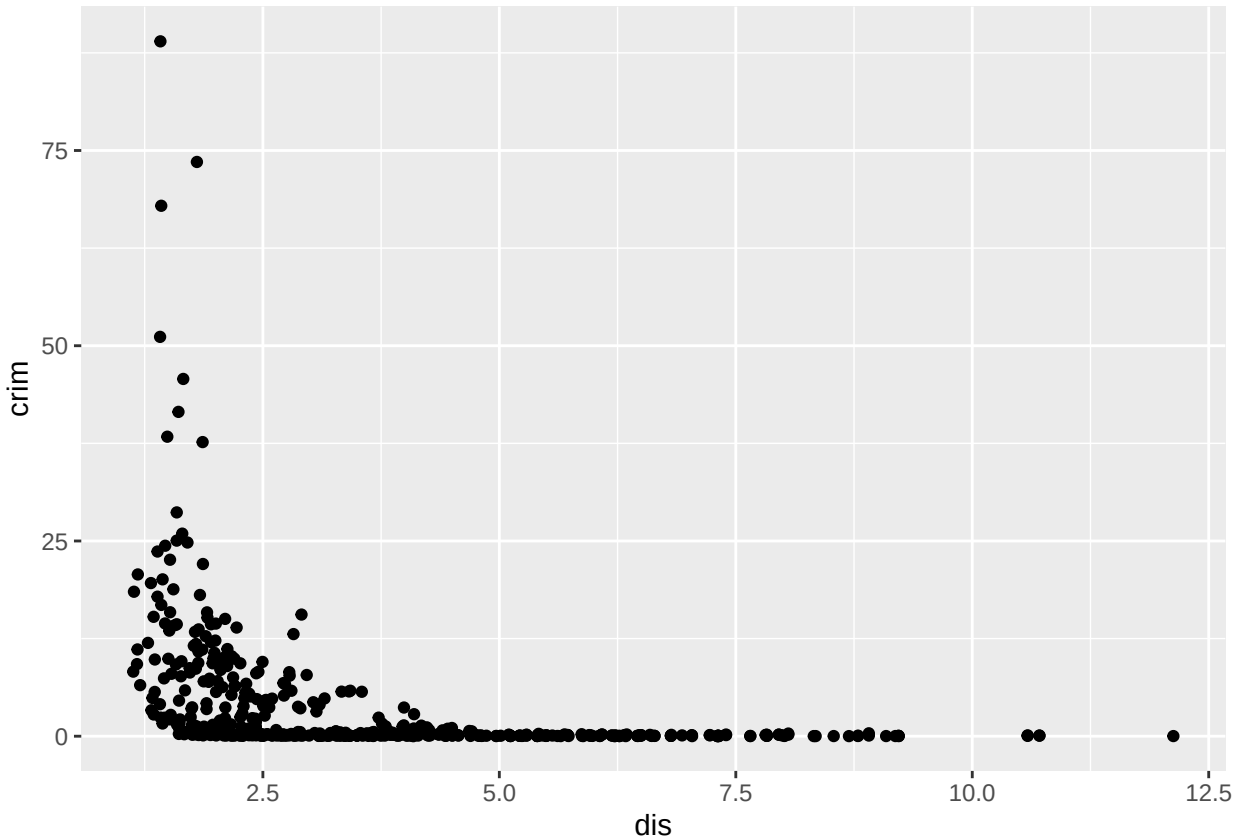

```
ggplot(data = Boston,aes(y=crim,x=rm))+geom_point()
```



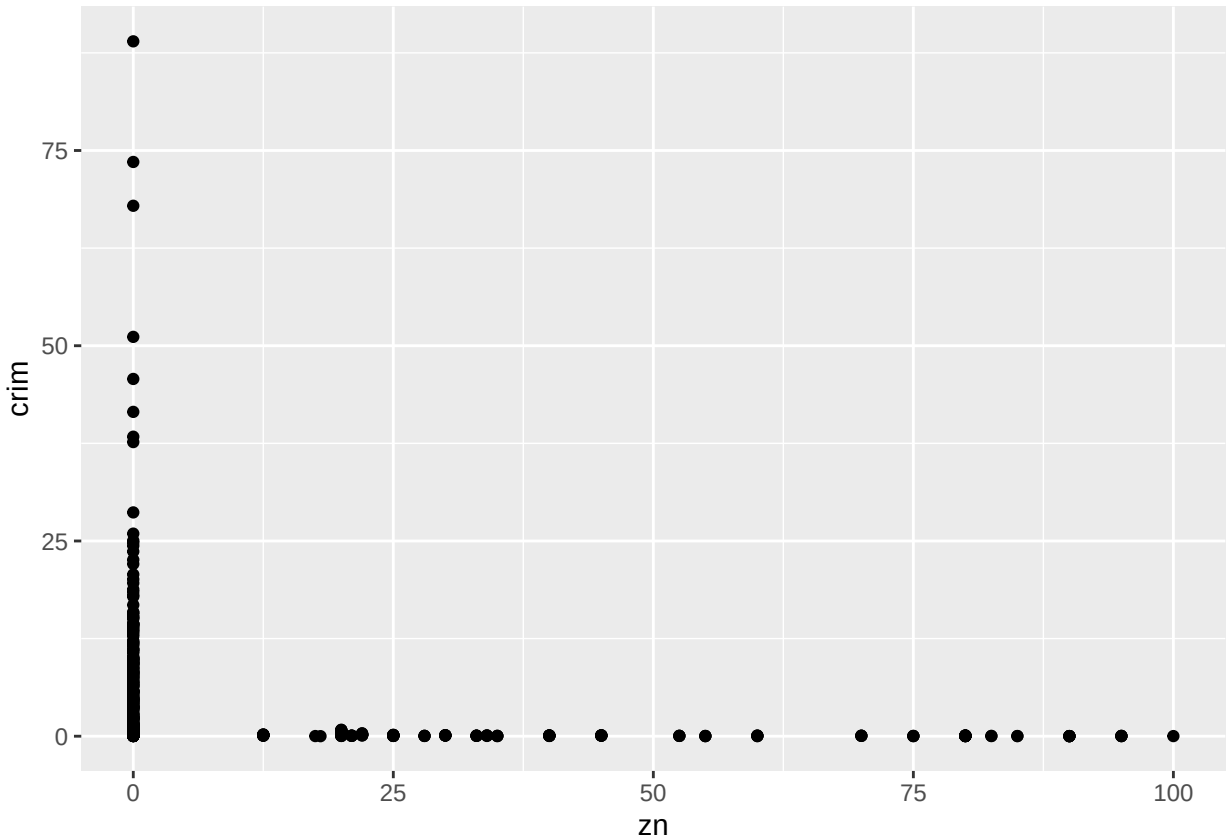

```
ggplot(data = Boston,aes(y=crim,x=age))+geom_point()
```



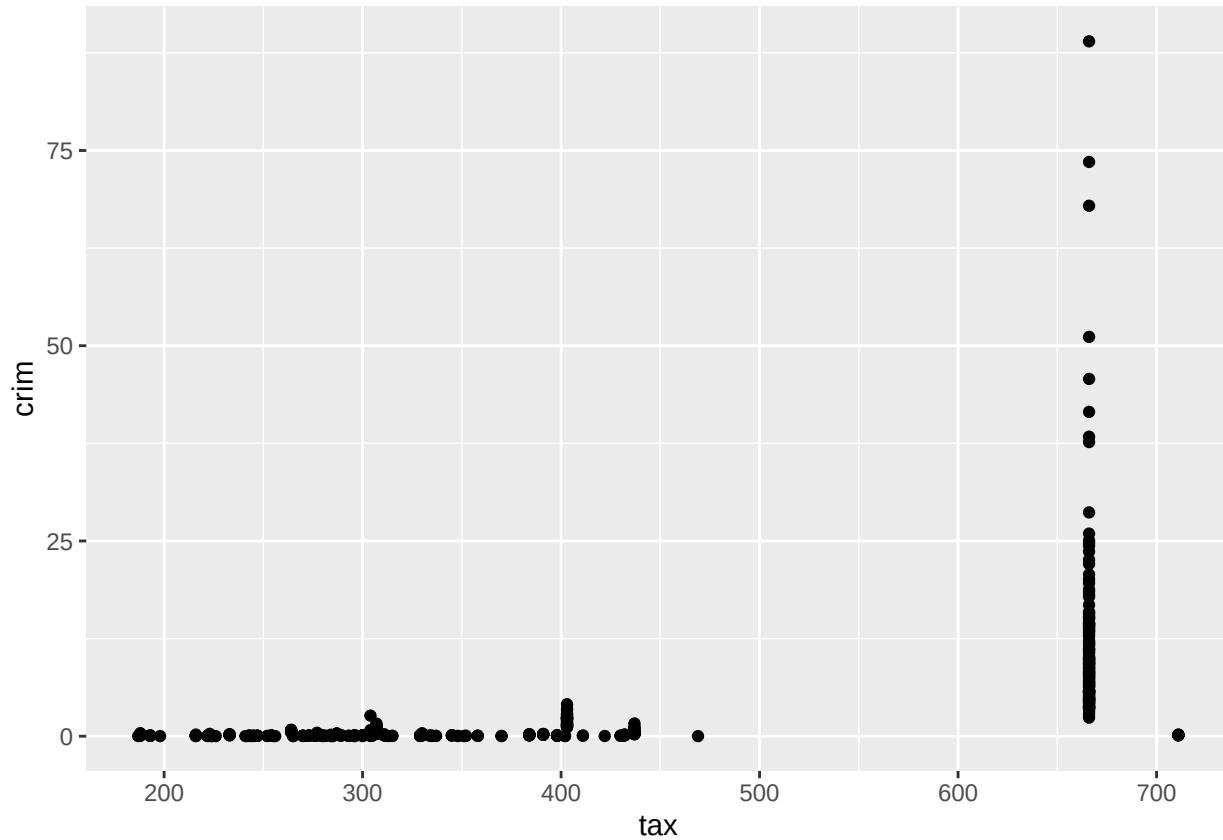

```
ggplot(data = Boston,aes(y=crim,x=dis))+geom_point()
```




```
ggplot(data = Boston,aes(y=crim,x=zn))+geom_point()
```




```r
ggplot(data = Boston,aes(y=crim,x=tax))+geom_point()
```



It seems that high crime increases in areas close to employment centers, older homes, and zones with residential lots less than 25,000 sqft. In other words, more urban or populated areas.

d Crime Rates

```
summary(Boston$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

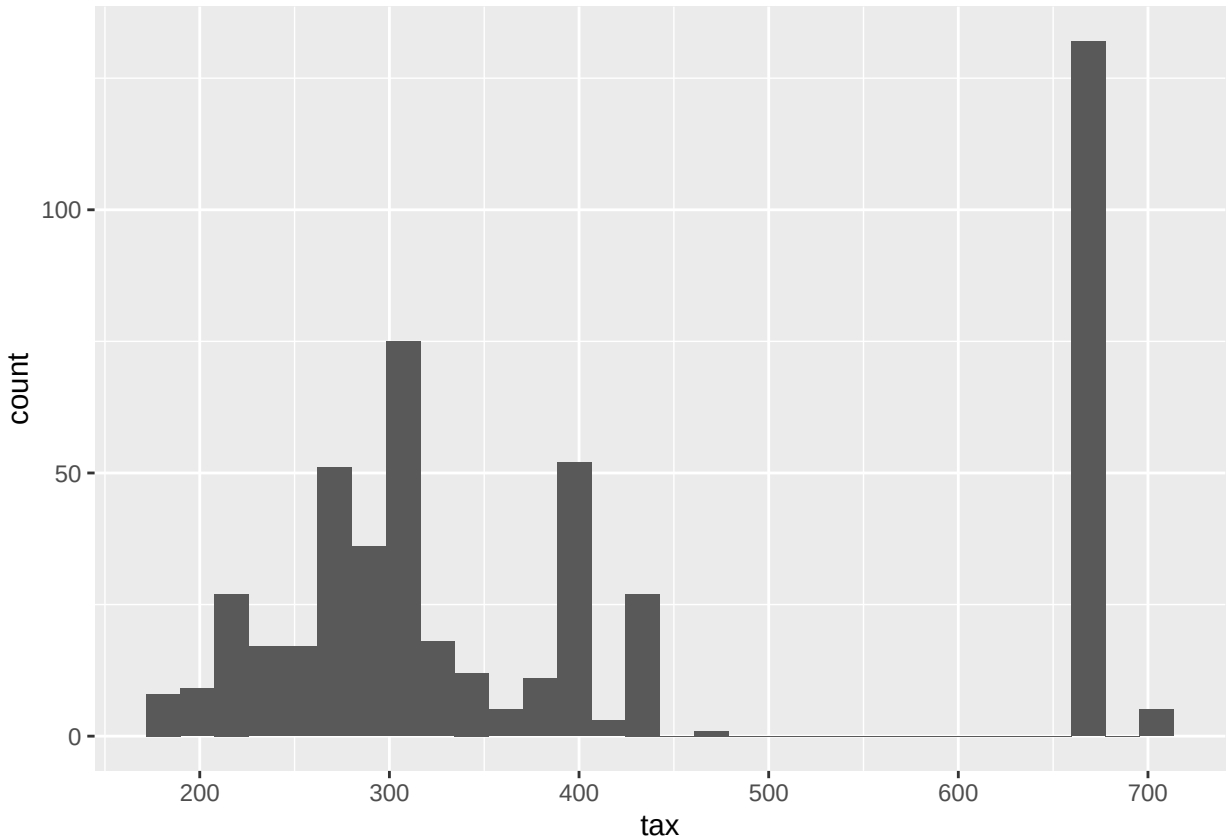
The maximum value is much higher than the 3th quartile. Counting crime rates above 30

```
length(Boston$crim[Boston$crim>30])
```

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

Tax Rates

```
ggplot(data = Boston,aes(tax))+geom_histogram(bins = 30)
```



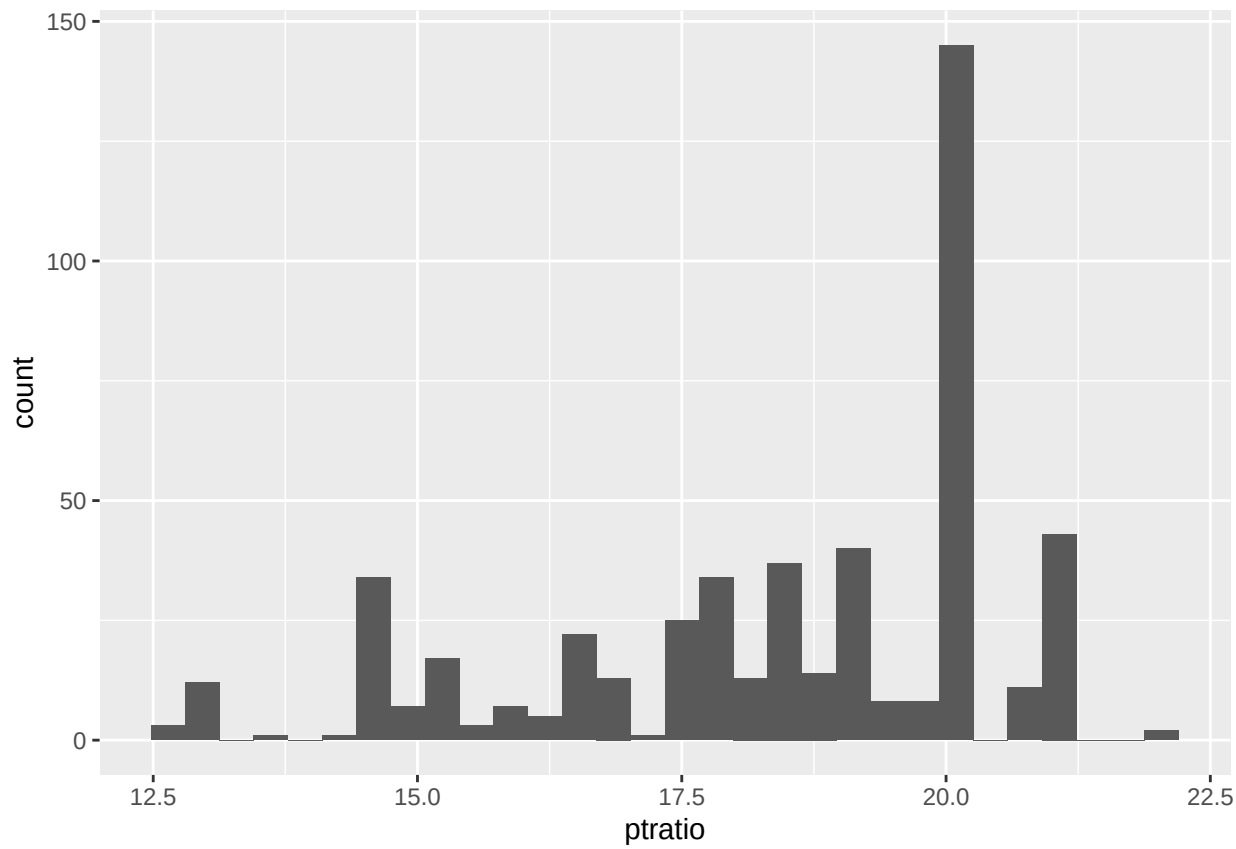
There are particulary suburbs in a higher level, counting values above 500.

```
length(Boston$tax[Boston$tax>500])
```

```
## [1] 137
```

Pupil-Teacher Ratio

```
ggplot(data = Boston,aes(ptratio))+geom_histogram(bins = 30)
```




```
length(Boston$ptratio[Boston$ptratio>19])
```

```
## [1] 253
```

There are only 18 suburbs with a crime rate greater than 20. There are 137 suburbs with a tax rate greater than 650. There are 253 suburbs with a pupil teacher ratio greater than 20. No real high crime rates, but high tax rates and high pupil teacher rates.

e.

```
nrow(Boston[Boston$chas==1,])
```

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

There are 35 suburbs that are bound by the Charles river.

f.

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

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

g.

```
min(Boston$medv)
```

```
## [1] 5
```

The 5th suburb has the lowest median value of owner occupied homes.

```
range(Boston$tax)
```

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

The range for taxes is 187 to 711.

```r
Boston[min(Boston$medv),]$tax
```

```
## [1] 222
```

The taxes for suburb 5 is 222, more on the lower end of the range.

h.

```r
nrow(Boston[Boston$rm>7,])
```

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

64 suburbs average more than 7 rooms per dwelling.

```r
nrow(Boston[Boston$rm>8,])
```

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

13 suburbs average more than 8 rooms per dwelling.

```r
Boston[Boston$rm>8,]
```

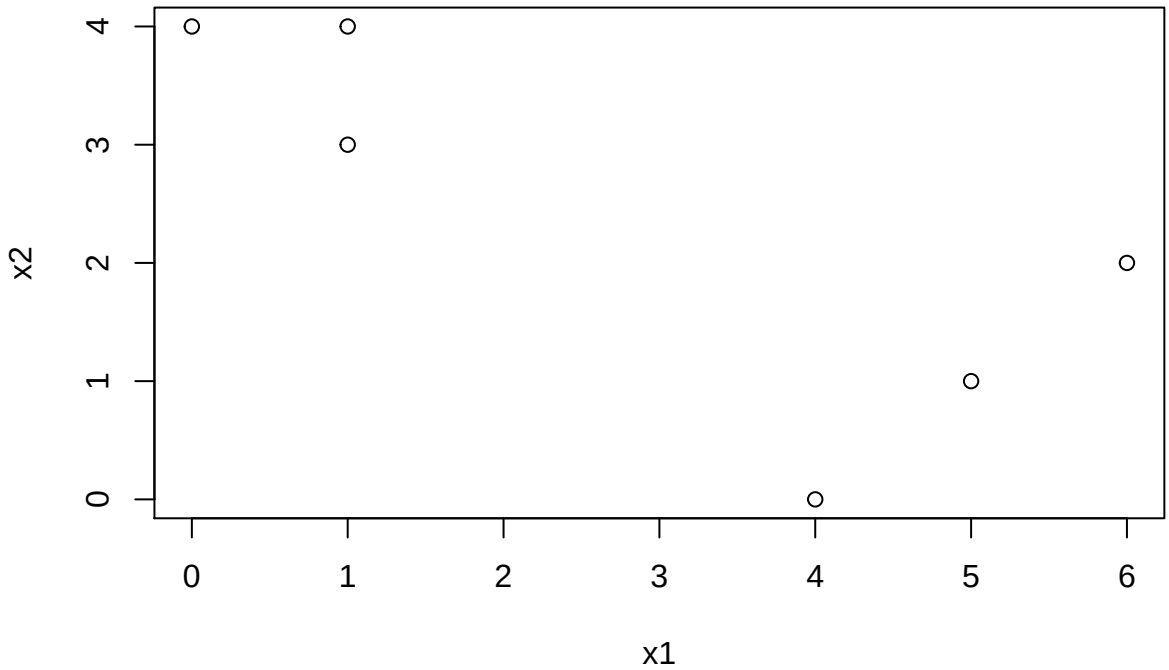
```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio  black lstat
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90  4.21
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45  3.32
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55  2.88
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05  4.14
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00  4.63
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38  3.13
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91  2.47
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95  3.95
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90  3.54
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70  5.12
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86  5.91
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54  7.44
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55  5.29
##     medv
## 98  38.7
## 164 50.0
## 205 50.0
## 225 44.8
## 226 50.0
## 227 37.6
## 233 41.7
## 234 48.3
## 254 42.8
## 258 50.0
## 263 48.8
## 268 50.0
## 365 21.9
```

There are only 2 suburbs with more than 8 rooms per dwelling that lie on the Charles river. rows : 164 and 365

K-means clustering

a.

```r
x1 <- c(1, 1, 0, 5, 6, 4)
x2 <- c(4, 3, 4, 1, 2, 0)
plot(x1,x2)
```



b.

```r
x <- cbind(x1,x2)
l<-sample(3, nrow(x), replace = T)
l
```

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

c

```r
centroid1 <- c(mean(x[l == 1, 1]), mean(x[l == 1, 2]))
centroid2 <- c(mean(x[l == 2, 1]), mean(x[l == 2, 2]))
centroid1
```

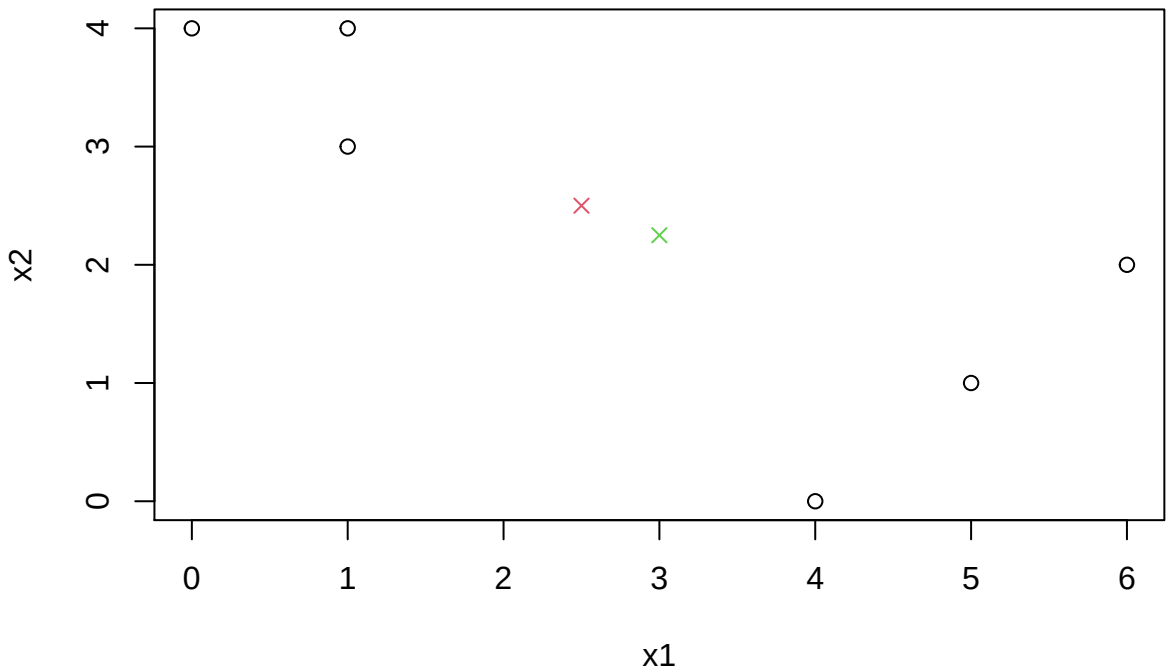
```
## [1] 2.5 2.5
```

```r
centroid2
```

```
## [1] 3.00 2.25
```

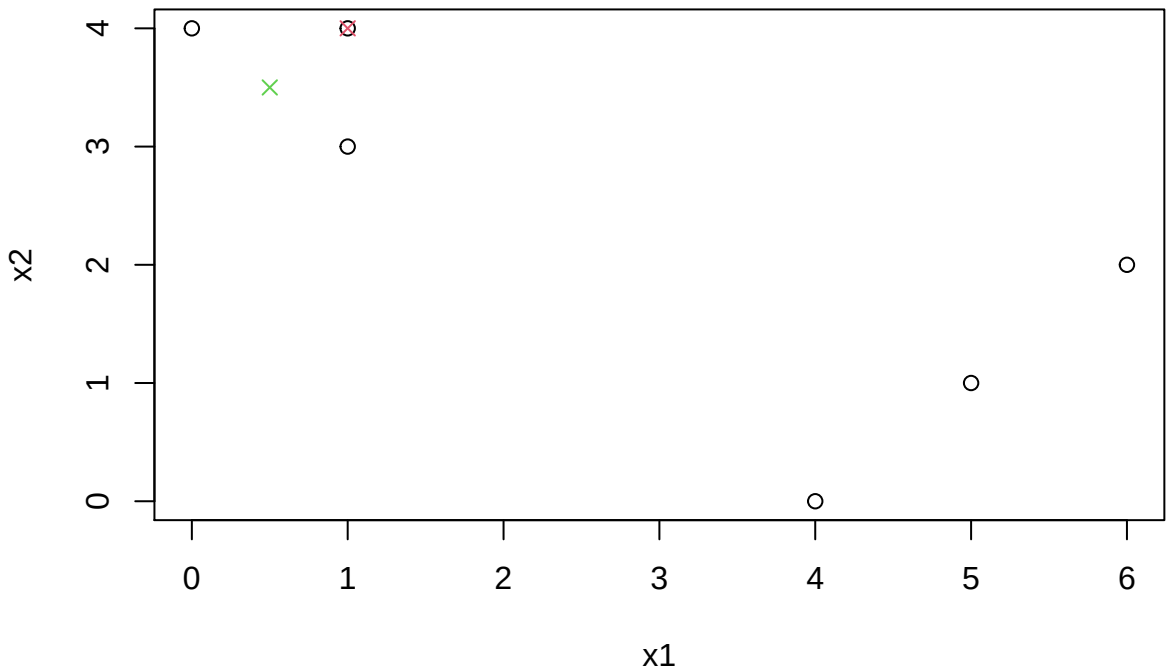
d

```r
l <- c(1, 2, 2, 3, 3, 3)
plot(x1, x2)
points(centroid1[1], centroid1[2], col = 2, pch = 4)
points(centroid2[1], centroid2[2], col = 3, pch = 4)
```



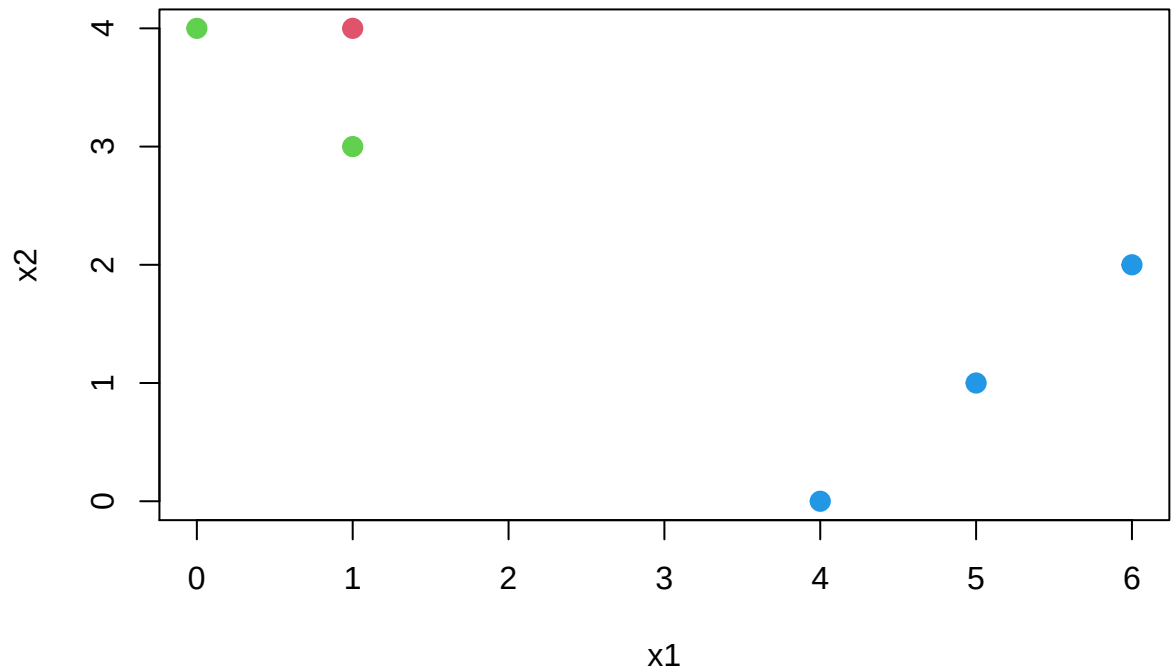
e

```
centroid1 <- c(mean(x[l == 1, 1]), mean(x[l == 1, 2]))
centroid2 <- c(mean(x[l == 2, 1]), mean(x[l == 2, 2]))
plot(x1, x2)
points(centroid1[1], centroid1[2], col = 2, pch = 4)
points(centroid2[1], centroid2[2], col = 3, pch = 4)
```



f

```
plot(x1, x2, col=(l + 1), pch = 20, cex = 2)
```



Q.6

```
set.seed(123)
n=20;p=50;s=10;mu1=c(rep(1,s),rep(0,p-s));
mu2=c(rep(0,s),rep(1,s),rep(0,p-2*s));
mu3=c(rep(0,s),rep(0,s),rep(1,s),rep(0,p-3*s));
x1=matrix(rnorm(n*p),n,p)+mu1;
x2=matrix(rnorm(n*p),n,p)+mu2
x3=matrix(rnorm(n*p),n,p)+mu3
features=rbind(x1,x2,x3)
cat=c(rep("A",n),rep("B",n),rep("C",n))
sim.data=data.frame(Class=cat,x=features)
```

c

```
km <- kmeans(features, centers=3)
table(km$cluster,cat)
```

```
##    cat
##      A  B  C
##   1 11  0  2
##   2  0 10 17
##   3  9 10  1
```

All are perfectly clustered

d

```r
km <- kmeans(features, centers=2)
table(km$cluster,cat)
```

```
##    cat
##      A  B  C
##   1 18 10  0
##   2  2 10 20
```

The middle class is forced to a wrong class. The extreme classes are classified correctly

e

```r
km <- kmeans(features, centers=4, nstart = 20)
table(km$cluster,cat)
```

```
##    cat
##      A  B  C
##   1  0  9  9
##   2  0  1 10
##   3 11  0  1
##   4  9 10  0
```

One of the classes is split into 2 classes

```r
#principal component
pc <- prcomp(features)$x
km <- kmeans(pc[, 1:2], centers=3)
table(km$cluster,cat)
```

```
##    cat
##      A  B  C
##   1 19  9  1
##   2  0  9  8
##   3  1  2 11
```

The result gives the almost identical splitting when compared with centers = 3 for actual data