Chapter 2 Statistical Learning

Jishen Yin

2020/4/28

```
knitr::opts_chunk$set(echo = TRUE)
library(MASS)
library(ISLR)
library(tidyverse)
library(GGally)
```

Problem 8

This exercise relates it to the `College` data set, which can be found in `ISLR` package. It contains a number of variables for 777 different universities and colleges in the US. The variables are

`Private`: Public/private indicator

`Apps`: Number of applications received

`Accept`: Number of applicants accepted

`Enroll`: Number of new students enrolled

`Top10perc`: New students from top 10% of high school class

`Top25perc`: New students from top 25% of high school class

`F.Undergrad`: Number of full-time undergraduates

`P.Undergrad`: Number of part-time undergraduates

`Outstate`: Out-of-state tuition

`Room.Board`: Room and board costs

`Books`: Estimated book costs

`Personal`: Estimated personal spending

`PhD`: Percent of faculty with Ph.D.'s

`Terminal`: Percent of faculty with terminal degree

`S.F.Ratio`: Student/faculty ratio

`perc.alumni`: Percent of alumni who donate

`Expend`: Instructional expenditure per student

`Grad.Rate`: Graduation rate

(a) Read the data into R. Call the loaded data college.

```
data(College)
```

(b) Look at the data using the fix() function.

(c)

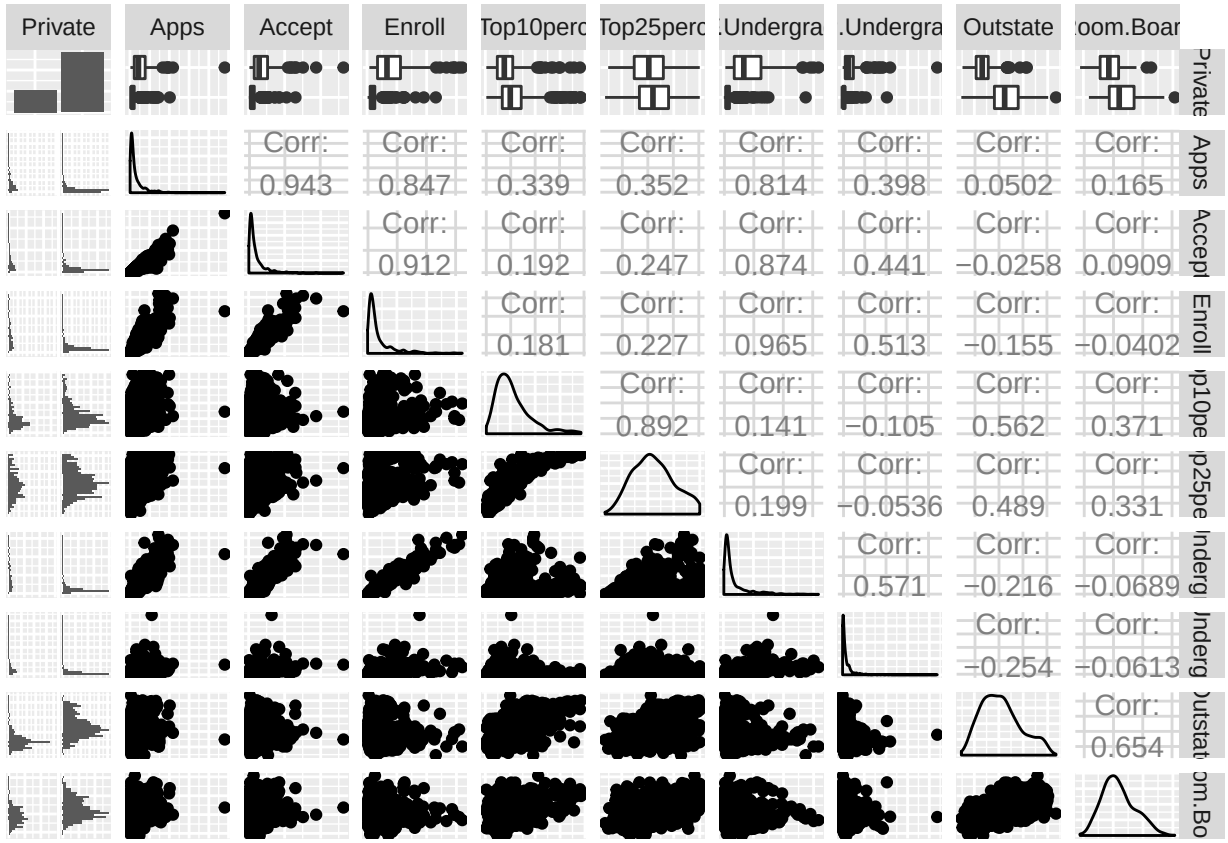
i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
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
```

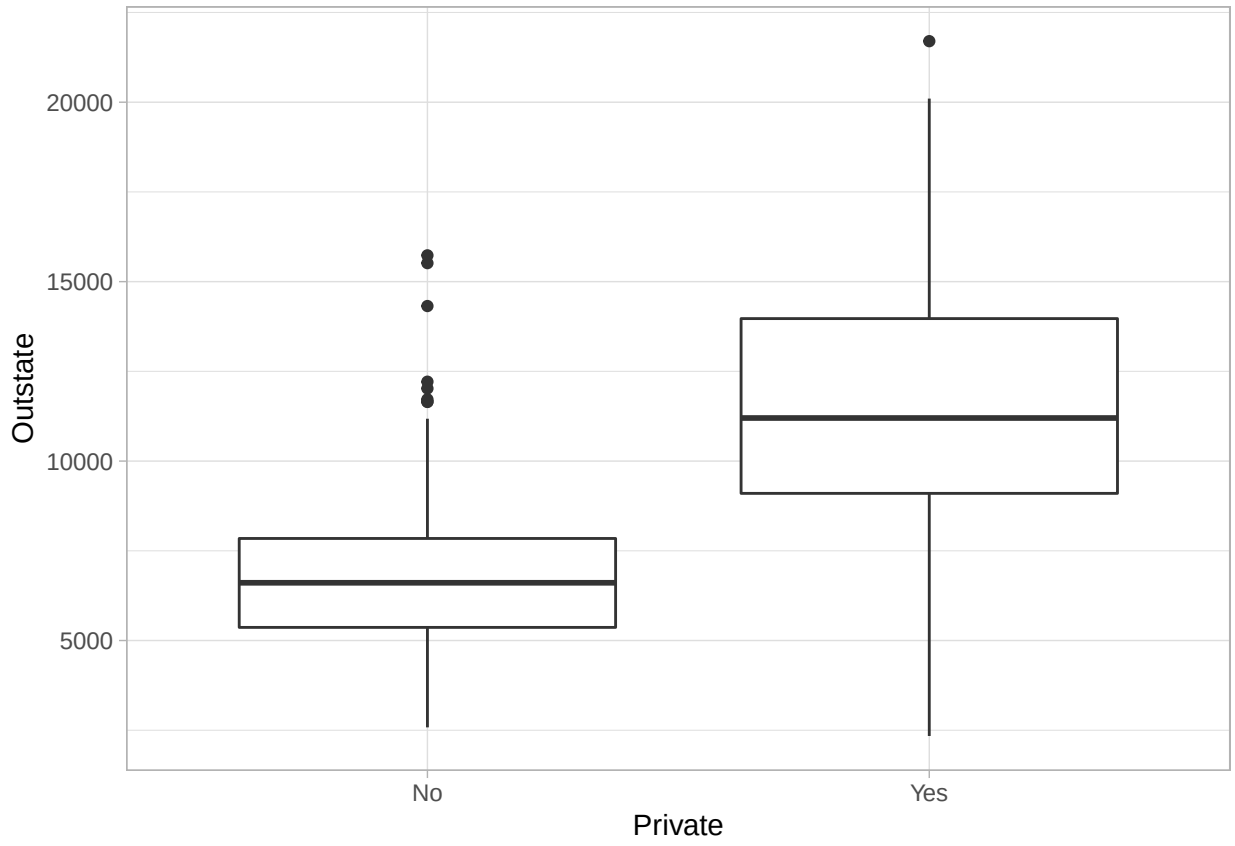
ii. Use the `pairs` function to produce a scatterplot matrix of the first columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

```
ggpairs(College[,1:10], axisLabels = "none")
```



iii. Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

```
ggplot() +
  geom_boxplot(aes(x = Private, y = Outstate), data = College) +
  theme_light()
```



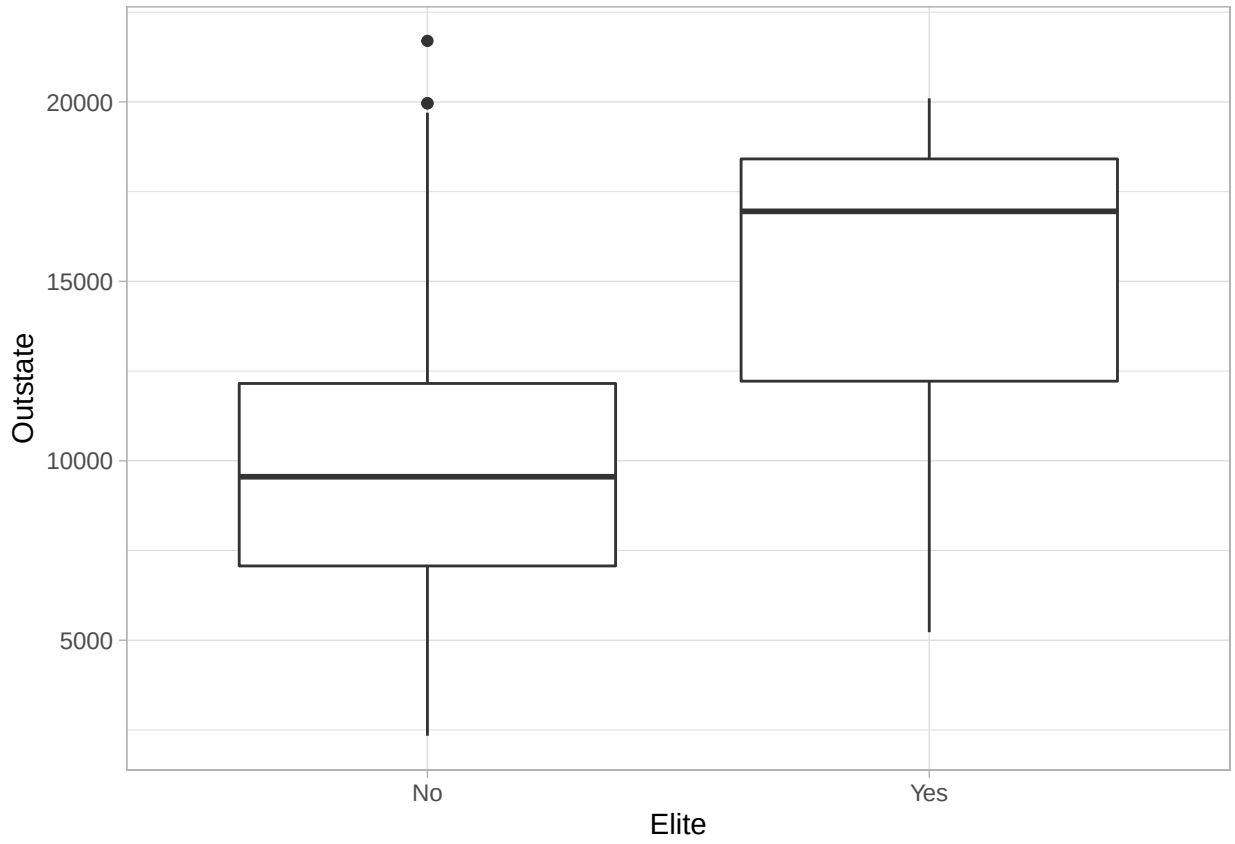
iv. Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Use th `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

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

```
summary(College$Elite)
```

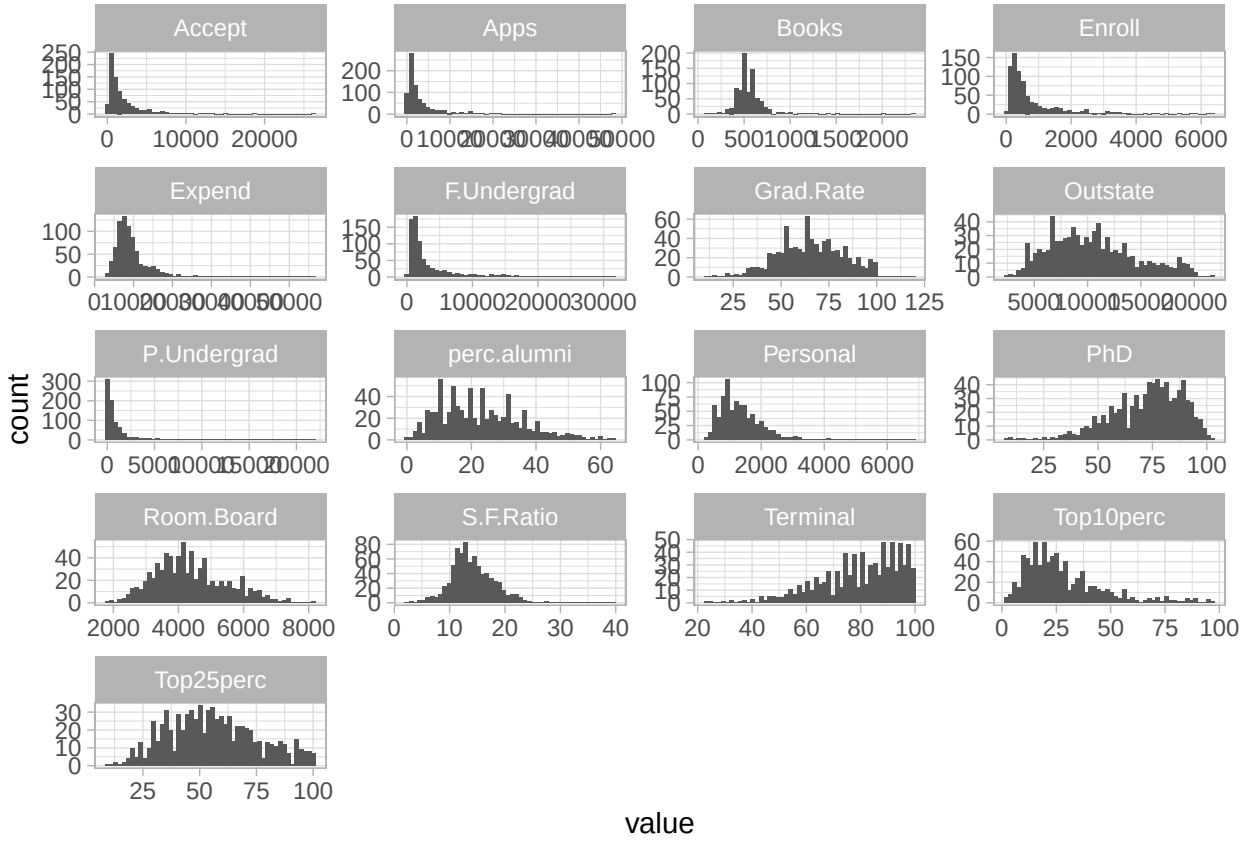
```
##  No Yes
## 699  78
```

```
ggplot() +
  geom_boxplot(aes(x = Elite, y = Outstate), data = College) +
  theme_light()
```



v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow=c(2, 2))` useful.

```
College.gathered <- College %>%
  select(-Private, -Elite) %>%
  gather(key = "variable", value = "value")

ggplot(data = College.gathered) +
  geom_histogram(aes(x = value), bins = 50) +
  facet_wrap(~variable, ncol = 4, scales = "free") +
  theme_light()
```



Problem 9

This exercise involves the `Auto` data set studied in the lab. Make sure that the missing values have been removed from the data.

```
data(Auto)
```

(a) Which of the predictors are quantitative, and which are qualitative?

With `help(Auto)`, we get the description of the data set.

`mpg`: miles per gallon

`cylinders`: Number of cylinders between 4 and 8

`displacement`: Engine displacement (cu. inches)

`horsepower`: Engine horsepower

`weight`: Vehicle weight (lbs.)

`acceleration`: Time to accelerate from 0 to 60 mph (sec.)

`year`: Model year (modulo 100)

`origin`: Origin of car (1. American, 2. European, 3. Japanese)

`name`:Vehicle name

The quantitative variables are `mpg`, `displacement`, `horsepower`, `weight` and `acceleration`. The qualitative variables are `cylinders`, `year`, `origin` and `name`.

```
Auto <- Auto %>% na.omit() %>%
  mutate(cylinders = as.factor(cylinders),
         year = as.factor(year),
         origin = as.factor(origin)) %>%
  select(-name)
```

(b) What is the range of each quantitative predictor?

```
Auto.continuous <- Auto %>%
  select(mpg, displacement, horsepower, weight, acceleration) %>%
  gather(key = "variable", value = "value")

Auto.continuous %>%
  group_by(variable) %>%
  summarise(range = max(value) - min(value))
```

```
## # A tibble: 5 x 2
##   variable      range
##   <chr>         <dbl>
## 1 acceleration   16.8
## 2 displacement  387
## 3 horsepower    184
## 4 mpg            37.6
## 5 weight       3527
```

(c) What is the mean and standard deviation of each quantitative predictor?

```
Auto.continuous %>%
  group_by(variable) %>%
  summarise(mean = mean(value),
            sd = sd(value))
```

```
## # A tibble: 5 x 3
##   variable       mean     sd
##   <chr>         <dbl>  <dbl>
## 1 acceleration   15.5   2.76
## 2 displacement  194.  105.
## 3 horsepower    104.   38.5
## 4 mpg            23.4   7.81
## 5 weight       2978.  849.
```

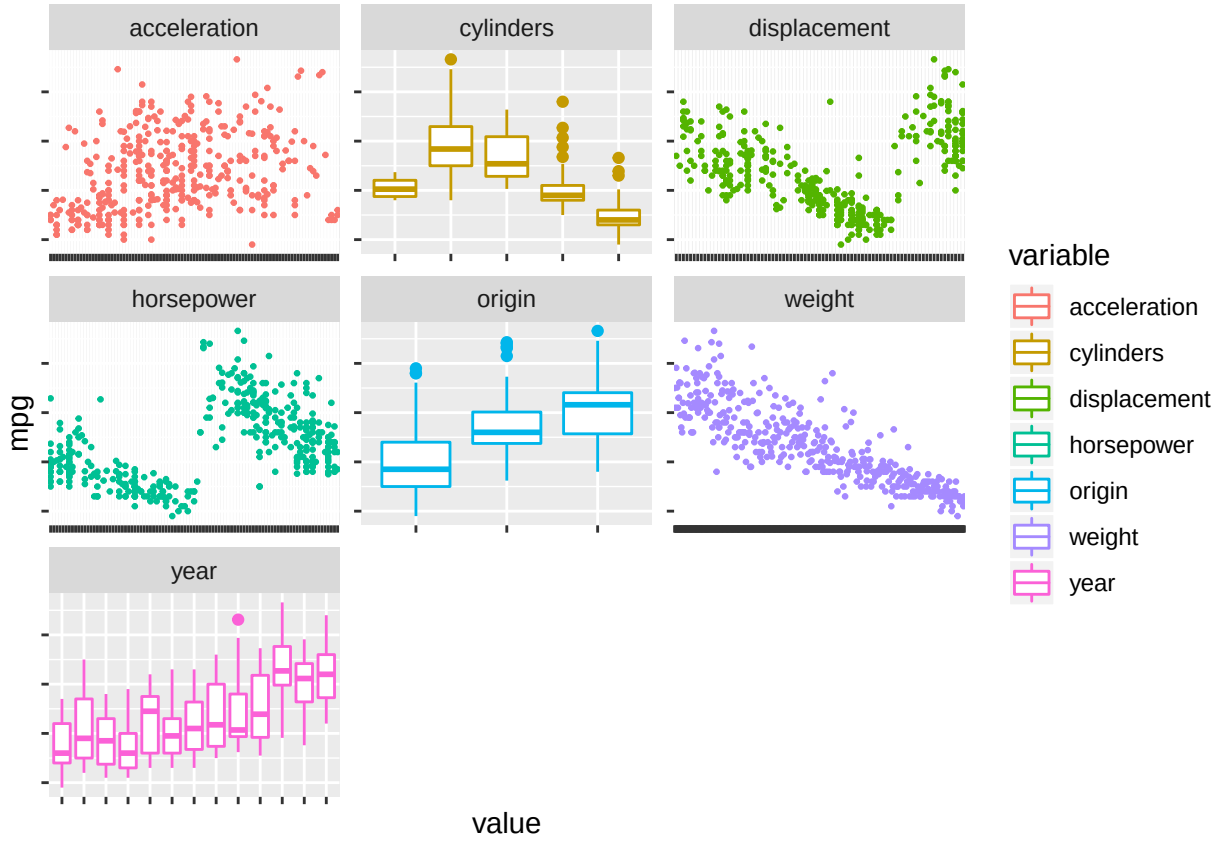
(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
Auto.subset <- Auto[c(1:9, 86:392),]
Auto.subset %>%
  select(mpg, displacement, horsepower, weight, acceleration) %>%
  gather(key = "variable", value = "value") %>%
  group_by(variable) %>%
  summarise(range = max(value) - min(value),
            mean = mean(value),
            sd = sd(value))
```

```
## # A tibble: 5 x 4
##   variable      range   mean     sd
##   <chr>         <dbl>  <dbl>  <dbl>
## 1 acceleration   16.3   15.7   2.69
## 2 displacement  387    187.   99.7
## 3 horsepower    184    101.   35.7
## 4 mpg            35.6   24.4   7.87
## 5 weight       3348   2936.  811.
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice.

```
Auto.gathered <- Auto %>%
  gather(key = "variable", value = "value", -mpg)

Auto.gathered.continuous <- Auto.gathered %>%
  filter(variable %in% c("acceleration", "displacement", "horsepower", "weight"))

Auto.gathered.discrete <- Auto.gathered %>%
  filter(!(variable %in% c("acceleration", "displacement", "horsepower", "weight")))

ggplot(Auto.gathered, mapping = aes(x = value, y = mpg, color = variable, group = value)) +
  facet_wrap(~variable, scales = "free") +
  geom_point(data = Auto.gathered.continuous, size = 0.5) +
  geom_boxplot(data = Auto.gathered.discrete) +
  theme(axis.text = element_blank())
```



(f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

`Acceleration` seems useless in predicting `mpg`. Three qualitative variables are potentially useful since the boxplots between different groups show a significant difference. `Weight` is negatively correlated with `mpg` while `displacement` and `horsepower` do not have strictly linear relationship with `mpg`. But we could use hierarchical model with these two predictors.

```
model <- lm(mpg ~ .-acceleration, data = Auto)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ . - acceleration, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.931 -1.671 -0.049  1.448 11.612
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.9706782  1.9703396  15.718  < 2e-16 ***
## cylinders4    6.9489835  1.5189655   4.575 6.51e-06 ***
## cylinders5    6.6467365  2.3240427   2.860 0.004477 **
## cylinders6    4.3050676  1.6932440   2.542 0.011413 *
## cylinders8    6.3723261  1.9615936   3.249 0.001266 **
## displacement  0.0117929  0.0067234   1.754 0.080256 .
## horsepower   -0.0395543  0.0104627  -3.781 0.000182 ***
## weight       -0.0051675  0.0005439  -9.501  < 2e-16 ***
## year71        0.9057500  0.8066633   1.123 0.262236
## year72       -0.4921367  0.8015260  -0.614 0.539593
## year73       -0.5550656  0.7185757  -0.772 0.440340
## year74        1.2376123  0.8470486   1.461 0.144840
## year75        0.8654150  0.8276285   1.046 0.296402
## year76        1.4923994  0.7942429   1.879 0.061027 .
## year77        2.9948793  0.8136242   3.681 0.000267 ***
## year78        2.9703034  0.7736654   3.839 0.000145 ***
## year79        4.8922614  0.8182998   5.979 5.30e-09 ***
## year80        9.0552685  0.8695618  10.414  < 2e-16 ***
## year81        6.4527050  0.8525066   7.569 3.02e-13 ***
## year82        7.8336547  0.8429257   9.293  < 2e-16 ***
## origin2       1.6931856  0.5155093   3.284 0.001119 **
## origin3       2.2936695  0.4957722   4.626 5.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.844 on 370 degrees of freedom
## Multiple R-squared:  0.8744, Adjusted R-squared:  0.8673
## F-statistic: 122.6 on 21 and 370 DF,  p-value: < 2.2e-16
```

Problem 10

This exercise involves the `Boston` housing data set.

(a) Load and read about the data set `Boston`.

```
data(Boston)
```

With `help(Boston)`, we can get the description of the data set:

`crim`: per capita crime rate by town.

`zn`: proportion of residential land zoned for lots over 25,000 sq.ft.

`indus`: proportion of non-retail business acres per town.

`chas`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

`nox`: nitrogen oxides concentration (parts per 10 million).

`rm`: average number of rooms per dwelling.

`age`: proportion of owner-occupied units built prior to 1940.

`dis`: weighted mean of distances to five Boston employment centres.

`rad`: index of accessibility to radial highways.

`tax`: full-value property-tax rate per $10,000.

`ptratio`: pupil-teacher ratio by town.

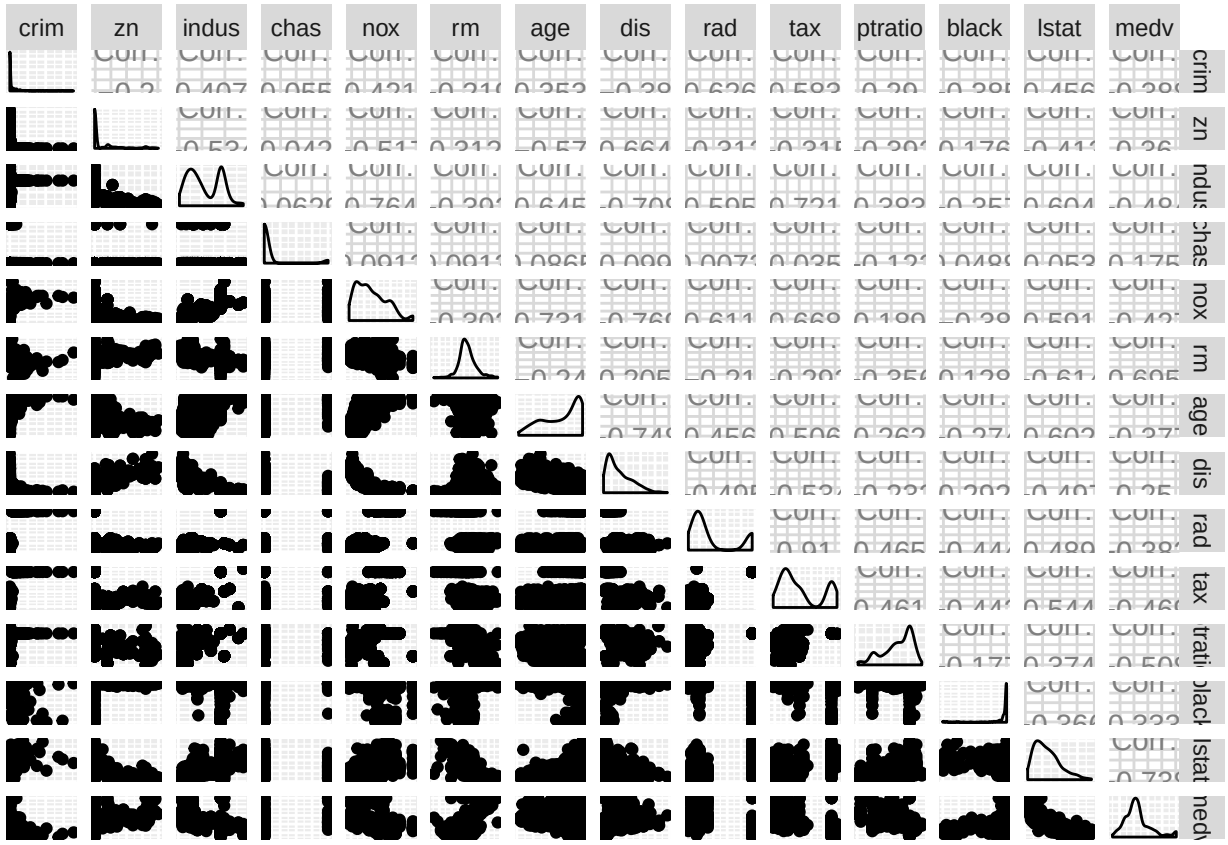
`black`: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

`lstat`: lower status of the population (percent).

`medv`: median value of owner-occupied homes in $1000s.

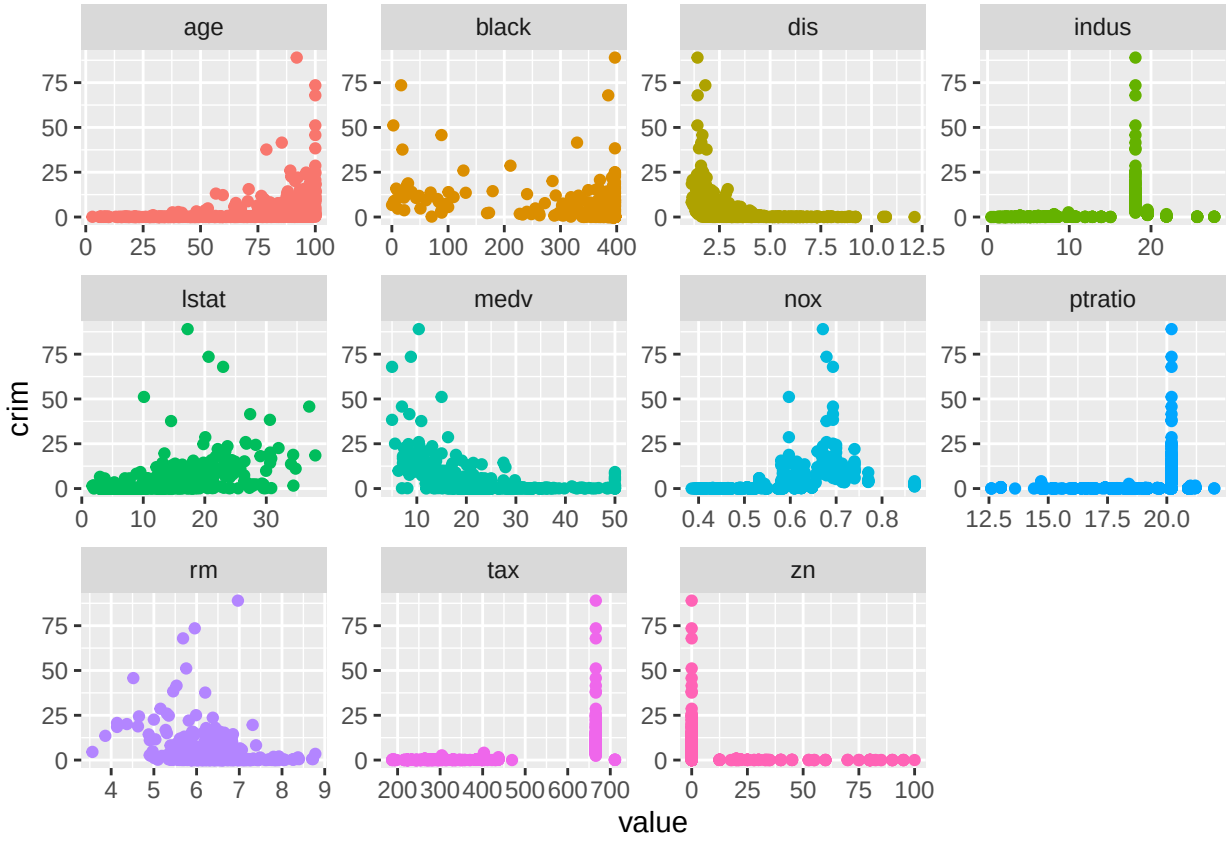
(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

```
ggpairs(Boston, axisLabels = "none")
```

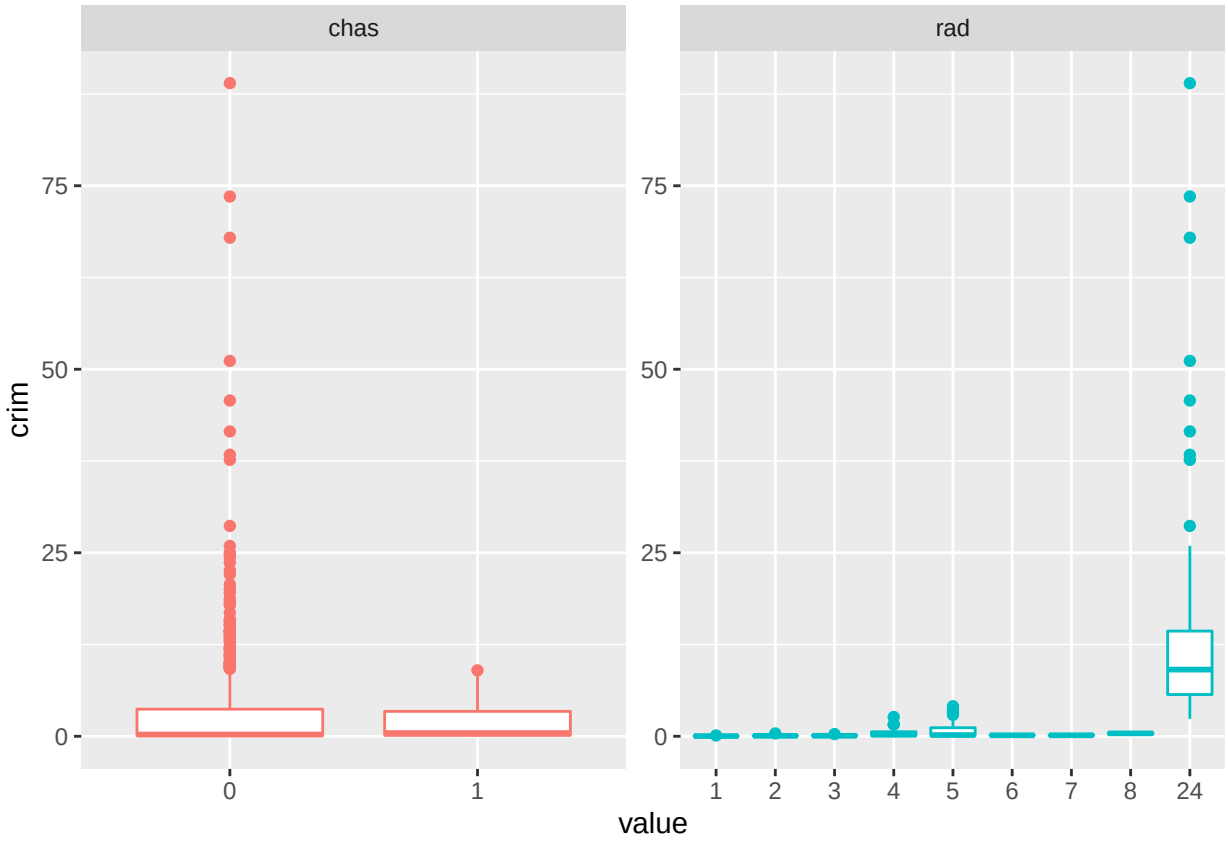


(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```
Boston.gathered <- Boston %>%
  gather(key = "variable", value = "value", -crim)

Boston.gathered.continuous <- Boston.gathered %>%
  filter(!(variable %in% c("rad", "chas")))

Boston.gathered.discrete <- Boston.gathered %>%
  filter(variable %in% c("rad", "chas")) %>%
  mutate(value = as.factor(value))

ggplot(Boston.gathered.continuous,
       aes(x = value, y = crim, color = variable)) +
  geom_point() +
  facet_wrap(~variable, scales = "free") +
  guides(color = "none")
```




```
ggplot(Boston.gathered.discrete,
       aes(x = value, y = crim, color = variable)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free") +
  guides(color = "none")
```



(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
Boston %>%
  gather(key = "variable", value = "value") %>%
  filter(variable %in% c("crim", "tax", "ptratio")) %>%
  group_by(variable) %>%
  summarise(max = max(value),
            range = max(value) - min(value))
```

```
## # A tibble: 3 x 3
##   variable   max range
##   <chr>    <dbl> <dbl>
## 1 crim      89.0  89.0
## 2 ptratio   22     9.4
## 3 tax      711   524
```

(e) How many of the suburbs in this data set bound the Charles river?

```
summary(as.factor(Boston$chas))
```

```
##   0   1
## 471  35
```

(f) What is the median pupil-teacher ratio among the towns in this data set?

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

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

(g) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
Boston[Boston["medv"] == min(Boston["medv"]),]
```

```
##        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
```

(h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

```
nrow(Boston[Boston["rm"] > 7,])
```

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

```
nrow(Boston[Boston["rm"] > 8,])
```

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