

CptS 475/575: Data Science, Fall 2018

Sukhjinder Singh

16 September 2018

Assignment 2: R basics and Exploratory Data Analysis

Exercise 1:

- (a) Use the `read.csv()` function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
college <- read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/College.csv")
```

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

```
View(college)
```

You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands

You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

(c)

- i) Use the `summary()` function to produce a numerical summary of the variables in the data set. (Respond to this question with the mean graduation rate included in the summary result).

```
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
```

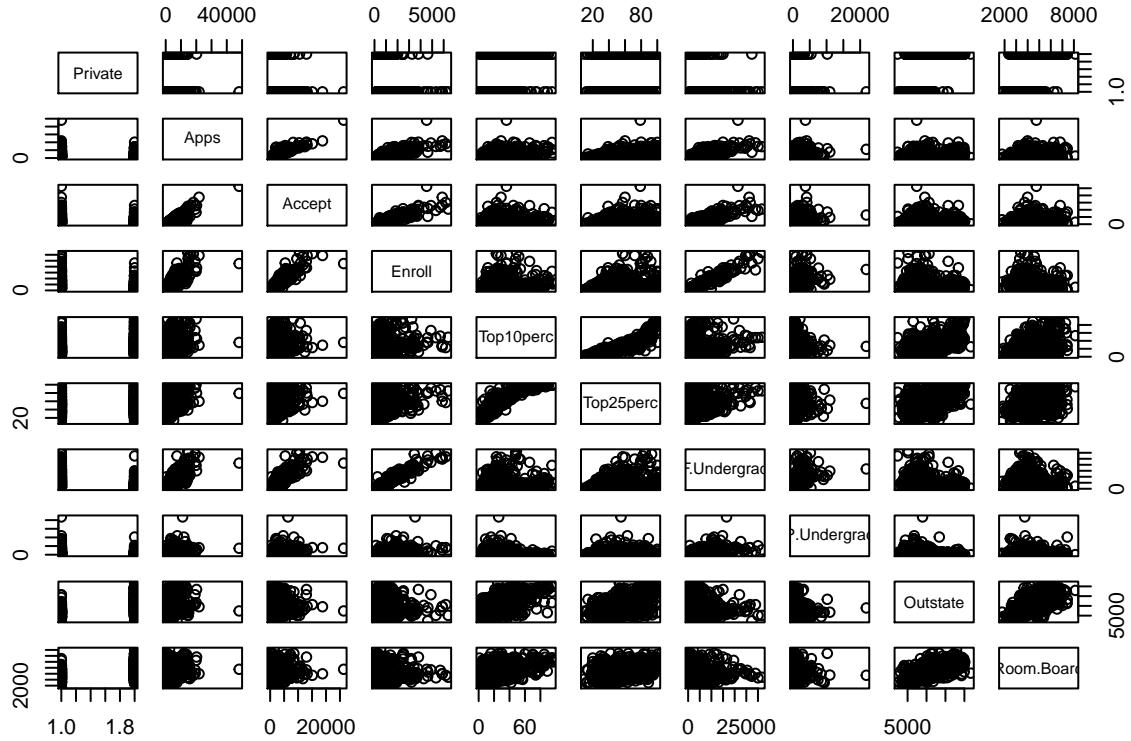
```

##   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

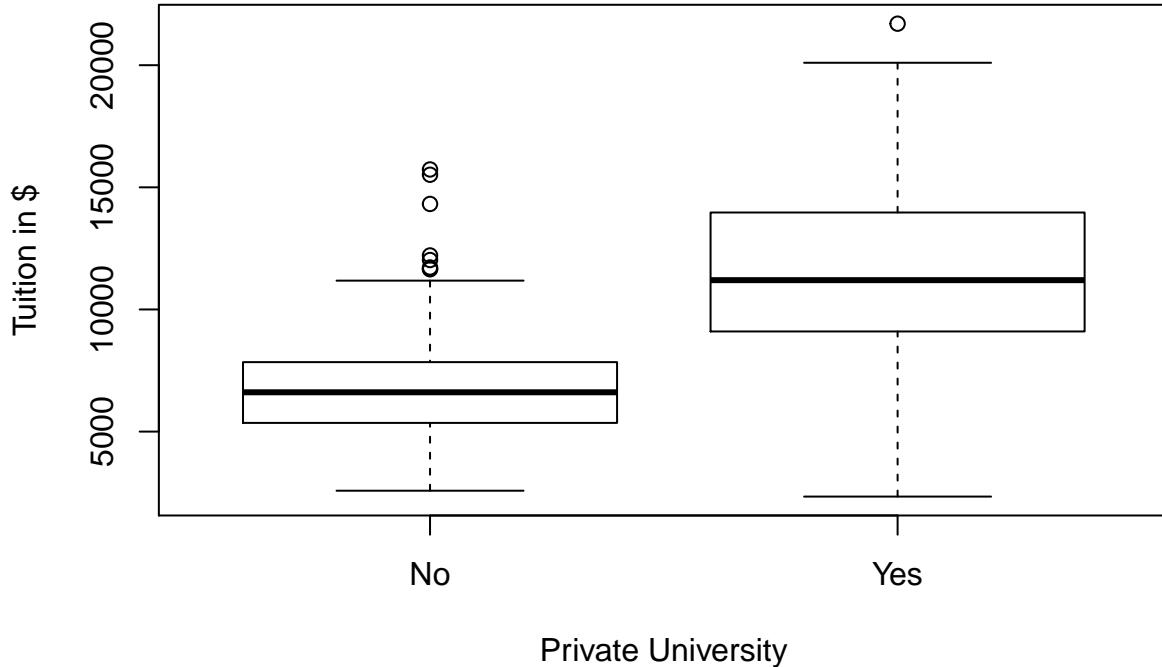
```

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

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



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



Boxplots of Outstate versus Private: Private universities have higher out of state tuition

- iv) Create a new qualitative variable, called Top, 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 25% of their high school classes exceeds 50%.

```
Top <- rep("No", nrow(college))
Top[college$Top25perc > 50] <- "Yes"
Top <- as.factor(Top)
college <- data.frame(college, Top)
```

Use the summary() function to see how many top universities there are.

```
summary(college)
```

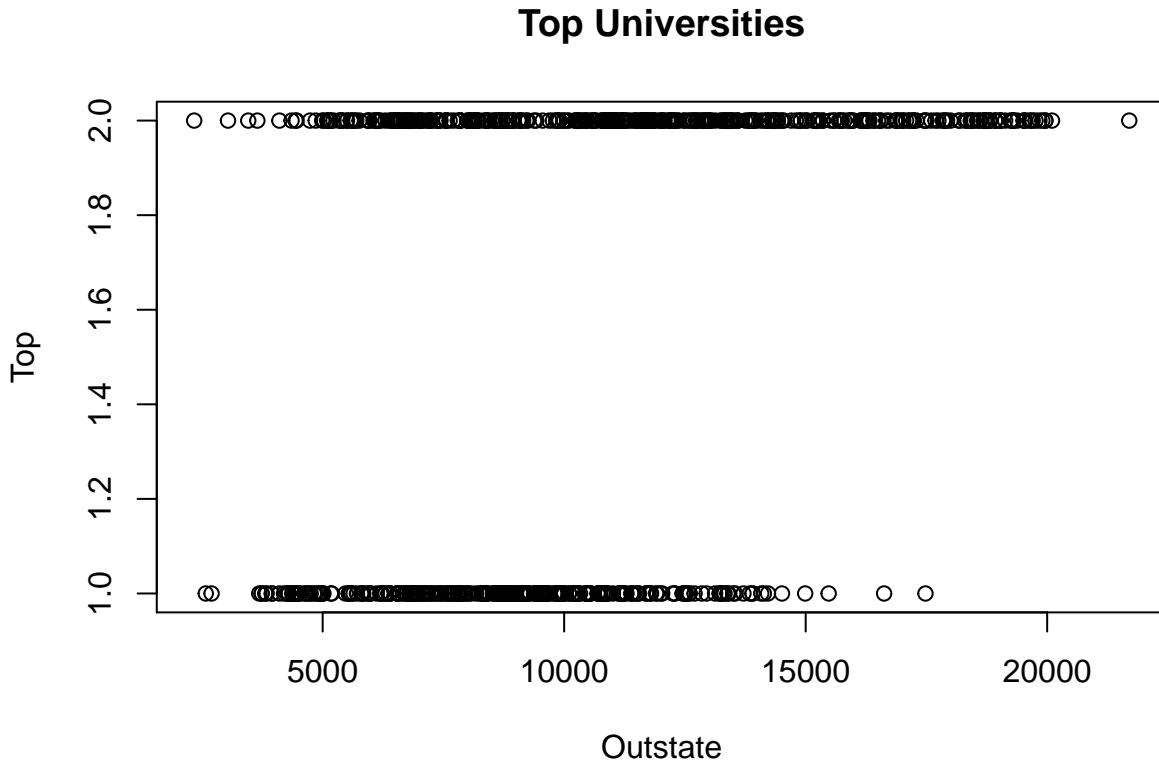
```
##  Private      Apps      Accept      Enroll      Top10perc
##  No :212    Min.   : 81    Min.   : 72    Min.   : 35    Min.   : 1.00
##  Yes:565   1st Qu.: 776   1st Qu.: 604   1st Qu.: 242   1st Qu.:15.00
##                Median :1558   Median :1110   Median :434    Median :23.00
##                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      Top
## Min.     :10.00   No :328
## 1st Qu.: 53.00  Yes:449
## Median   : 65.00
## Mean     : 65.46
## 3rd Qu.: 78.00
## Max.    :118.00

```

Now use the plot() function to produce side-by-side boxplots of Outstate versus Top. Ensure that this figure has an appropriate title and axis labels.



Boxplots of Outstate versus Top: Top universities have higher out of state tuition

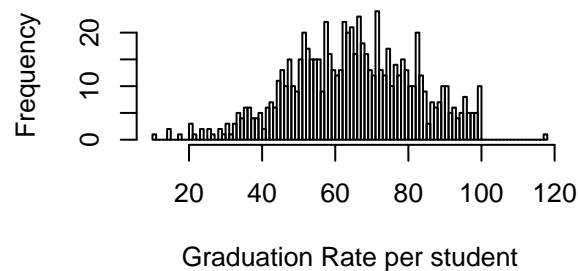
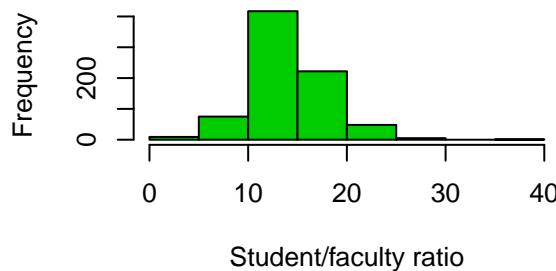
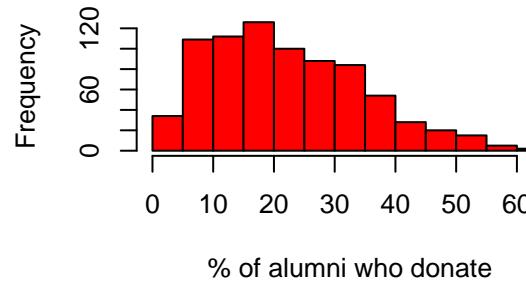
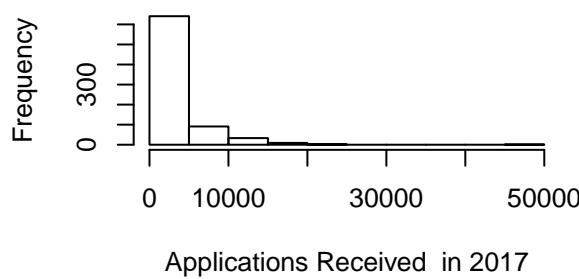
- v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways. Again, ensure that this figure has an appropriate title and axis labels.

```

par(mfrow=c(2,2))
hist(college$Apps, xlab = "Applications Received in 2017", main = "")
hist(college$perc.alumni, col=2, xlab = "% of alumni who donate", main = "")

```

```
hist(college$S.F.Ratio, col=3, breaks=10, xlab = "Student/faculty ratio", main = "")  
hist(college$Grad.Rate, breaks=100, xlab = "Graduation Rate per student", main = "")
```



vi. Continue exploring the data, and provide a brief summary of what you discover. You may use additional plots or numerical descriptors as needed. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

Followings are the observations in the dataset

1). Which university has the most liberal acceptance rate

```
acceptance_rate <- college$Accept / college$Apps  
row.names(college)[which.max(acceptance_rate)]
```

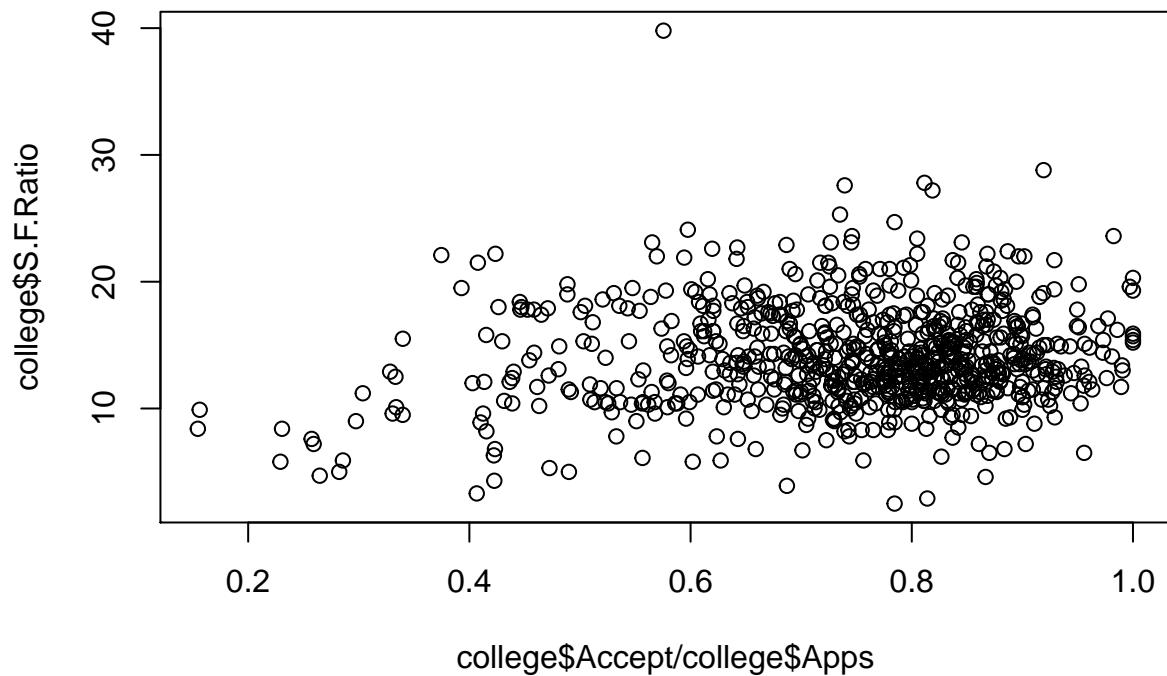
```
## [1] "Emporia State University"
```

2). The university has the highest acceptance rate

```
row.names(college)[which.max(acceptance_rate)]
```

```
## [1] "Emporia State University"
```

3). Colleges with low acceptance rate tend to have low S:F ratio.



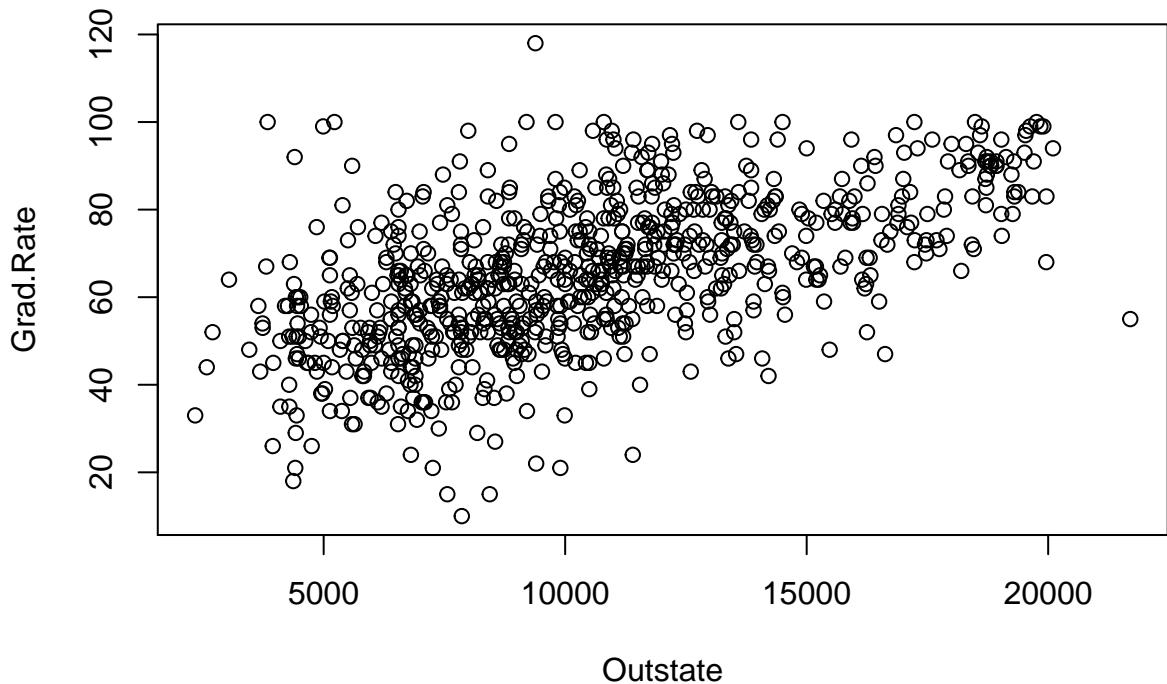
4). The university with the most students in the top 10% of class

```
row.names(college)[which.max(college$Top25perc)]
```

```
## [1] "Bowdoin College"
```

5).High tuition correlates to high graduation rate

```
plot(Grad.Rate ~ Outstate, data = college)
```



Exercise 2

This exercise involves the Auto.csv data set found on the course website. Make sure that the missing values have been removed from the data. To do this, consider the na.strings parameter of read.csv(), as well as the na.omit() function.

```
## [1] 392   9
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 :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00  3rd Qu.:8.000   3rd Qu.:275.8  3rd Qu.:126.0
##  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.:2225  1st Qu.:13.78  1st Qu.:73.00  1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615  3rd Qu.:17.02  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)           :365
```

- (a) Which of the predictors are quantitative, and which are qualitative? I have used integer value for qualitative variables such as 1,2,3.

```
Auto$originf <- factor(Auto$origin, labels = c("American", "European", "Japanese"))
with(Auto, table(originf, origin))
```

```
##      origin
## originf    1   2   3
## American 245  0   0
## European   0  68  0
## Japanese   0   0  79
```

Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year.

Qualitative: name, origin, originf

- (b) What is the range of each quantitative predictor? You can answer this using the range() function. Hint: consider using R's sapply() function to take the range of multiple features in a single function call.

```
#Qualitative predictors are followings
qualitative_columns <- which(names(Auto) %in% c("name", "origin", "originf"))
qualitative_columns
```

```
## [1] 8 9 10
# Apply the range function to the columns of Auto data that are not qualitative
sapply(Auto[, -qualitative_columns], 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) What is the mean and standard deviation of each quantitative predictor?

#Calculating mean

```
sapply(Auto[, -qualitative_columns], mean)
```

```
##          mpg cylinders displacement horsepower weight
## [1,] 23.445918      5.471939    194.411990 104.469388 2977.584184
## acceleration      year
## [2,] 15.541327    75.979592
```

#Calculating Standard Deviation

```
sapply(Auto[, -qualitative_columns], sd)
```

```
##          mpg cylinders displacement horsepower weight
## [1,] 7.805007      1.705783    104.644004 38.491160 849.402560
## acceleration      year
## [2,] 2.758864      3.683737
```

(d) Now remove the 25th through 75th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
sapply(Auto[-seq(25, 75), -qualitative_columns], mean)
```

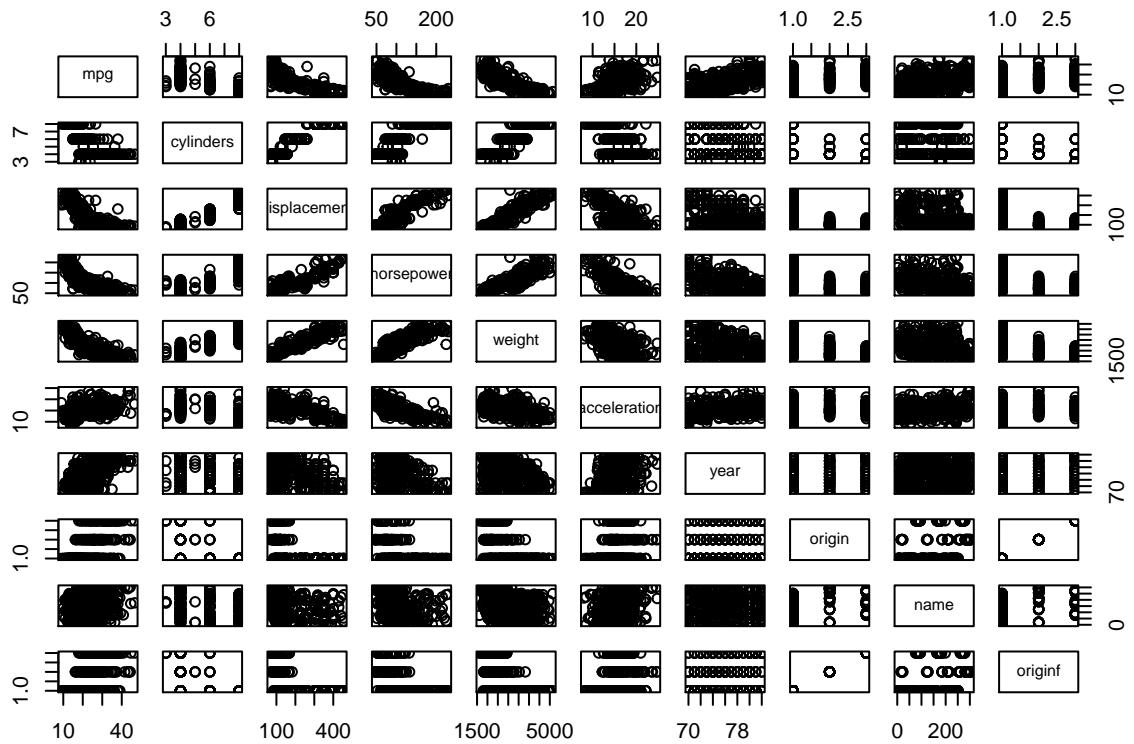
```
##          mpg cylinders displacement horsepower weight
## [1,] 24.195894      5.360704    187.167155 101.395894 2920.947214
## acceleration      year
## [2,] 15.650147    76.683284
```

```
sapply(Auto[-seq(25, 75), -qualitative_columns], sd)
```

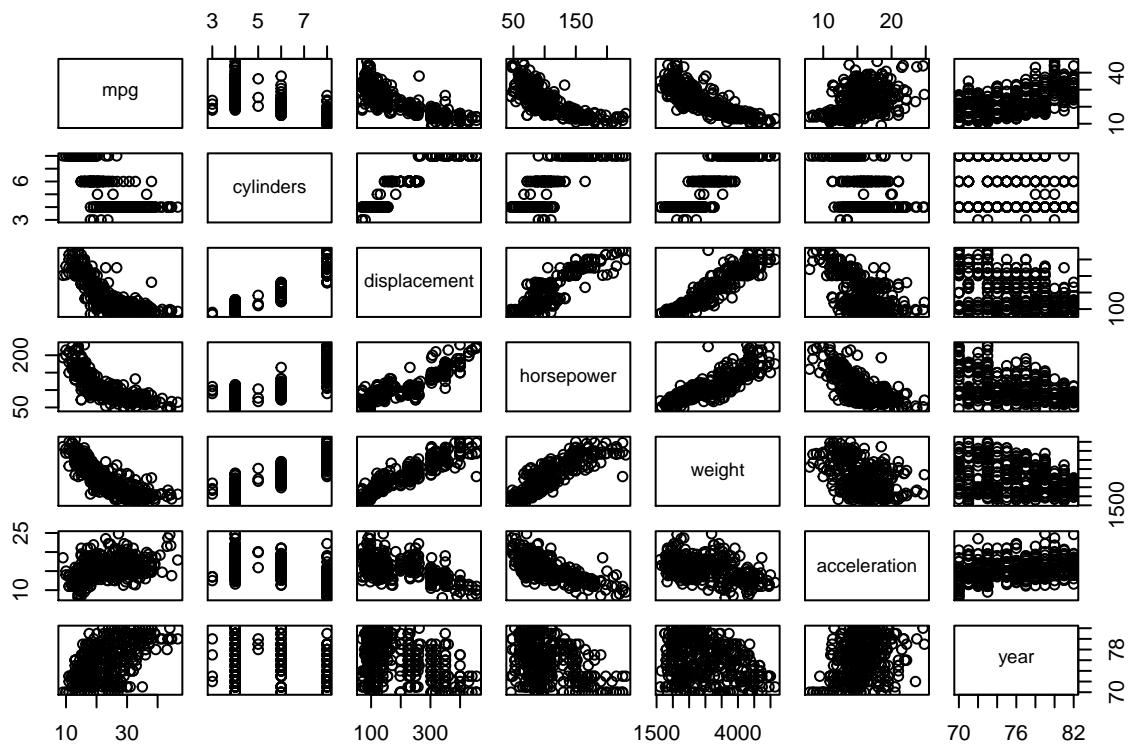
```
##          mpg cylinders displacement horsepower weight
## [1,] 7.720533      1.657987    101.119840 36.298742 799.636754
## acceleration      year
## [2,] 2.755216      3.424735
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
# Part (e):
pairs(Auto)
```

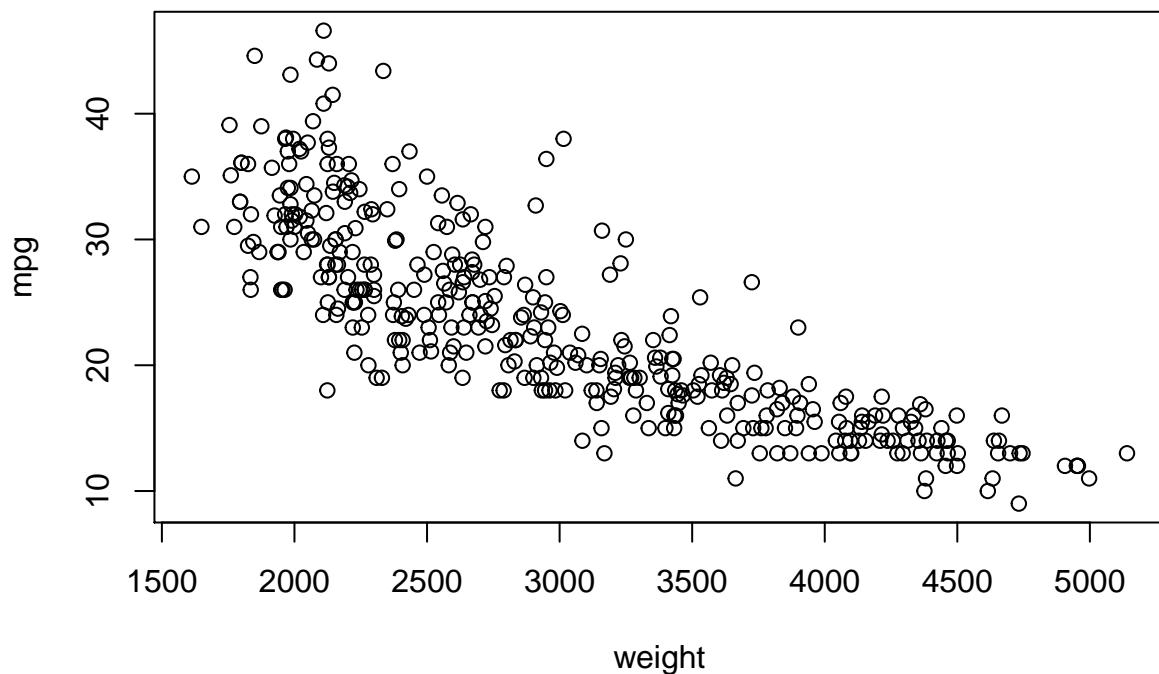


```
pairs(Auto[, -qualitative_columns])
```

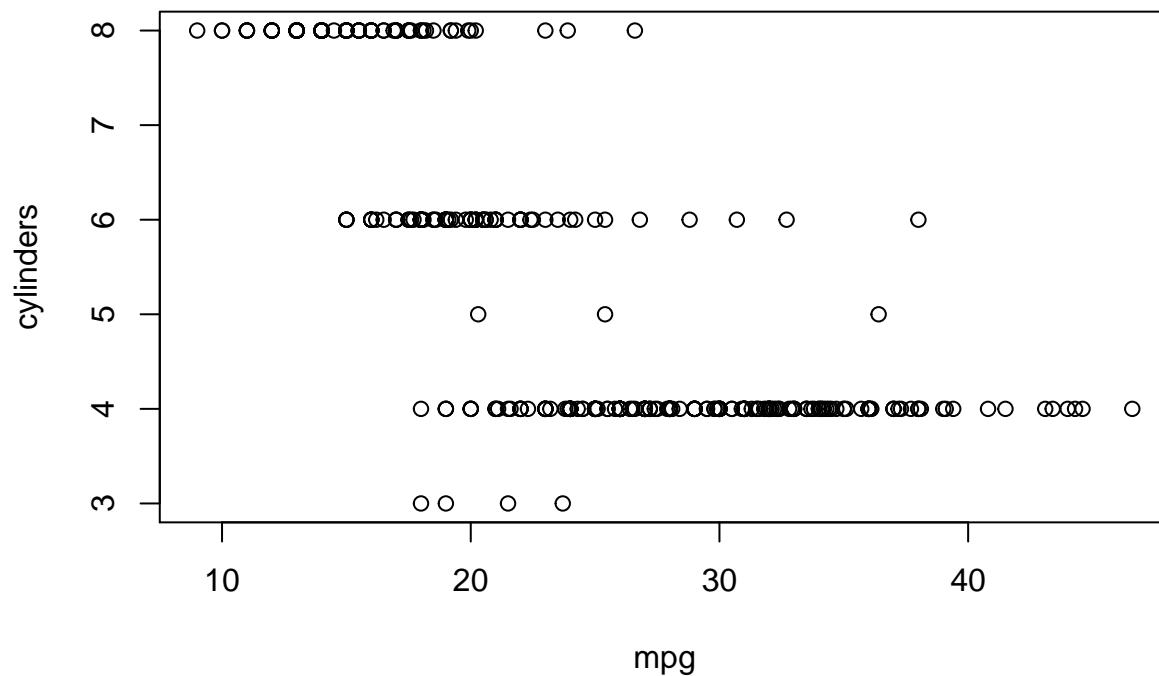


Lower mpg correlates with Heavier weight.

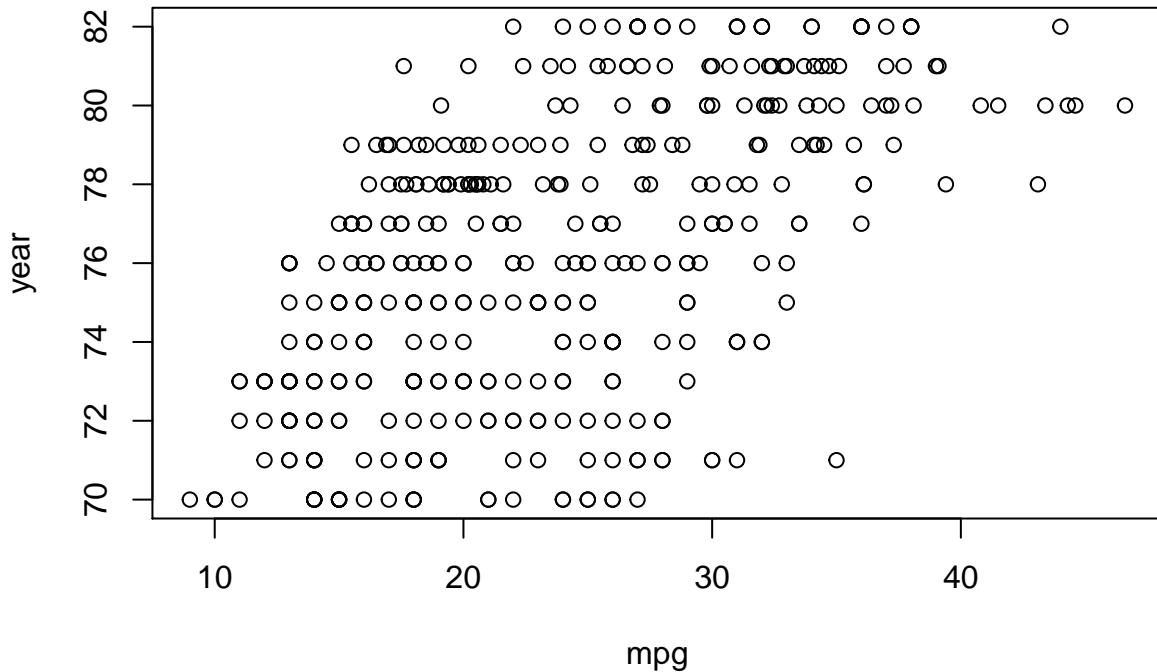
```
plot(mpg ~ weight, data=Auto)
```



```
# less mpg, more cylinders.  
with(Auto, plot(mpg, cylinders))
```



```
# Cars become more efficient over time.  
with(Auto, plot(mpg, year))
```



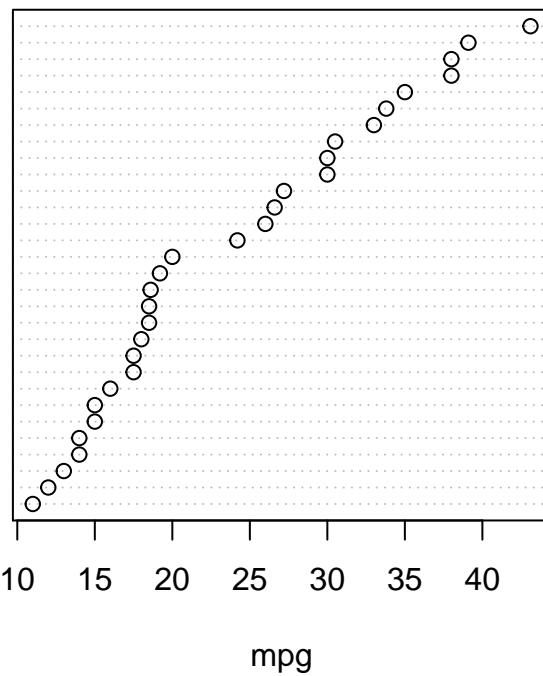
Plotting some mpg vs. some of our qualitative features for first 30 samples:

```
#observation of 30 samples
Auto.sample <- Auto[sample(1:nrow(Auto), 30), ]

# order them
Auto.sample <- Auto.sample[order(Auto.sample$mpg), ]

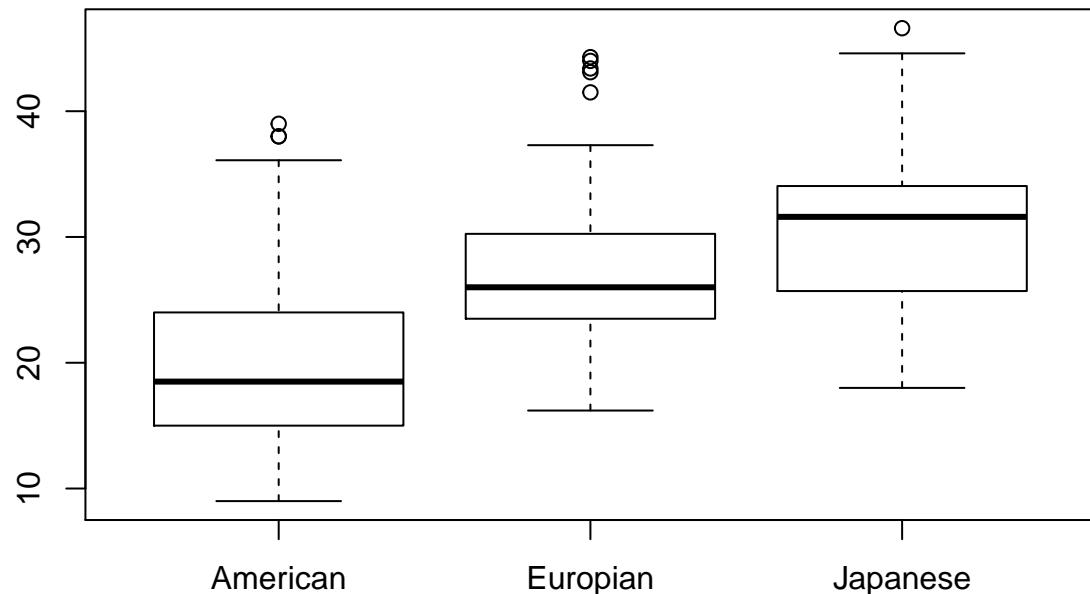
# plot them using a "dotchart"
with(Auto.sample, dotchart(mpg, name, xlab = "mpg"))
```

volkswagen rabbit custom diesel
 toyota starlet
 datsun 310 gx
 honda civic
 triumph tr7 coupe
 subaru dl
 honda civic cycc
 volkswagen dasher
 fiat 124s
 plymouth reliant
 buick skyhawk
 renault 5 (sw)
 datsun 810 maxima
 amc gremlin
 pontiac phoenix lj
 dodge aspen
 pontiac ventura si
 mercury lebaron town @ country (sw)
 plymouth fury
 chevrolet chevelle malibu classic
 amc pacer d4
 amc rebel ss
 chevrolet monte carlo
 dodge coronet custom
 plymouth cuda 340
 buick estate wagon (sw)
 plymouth volare premier v8
 buick electra 225 custom
 mercury marquis



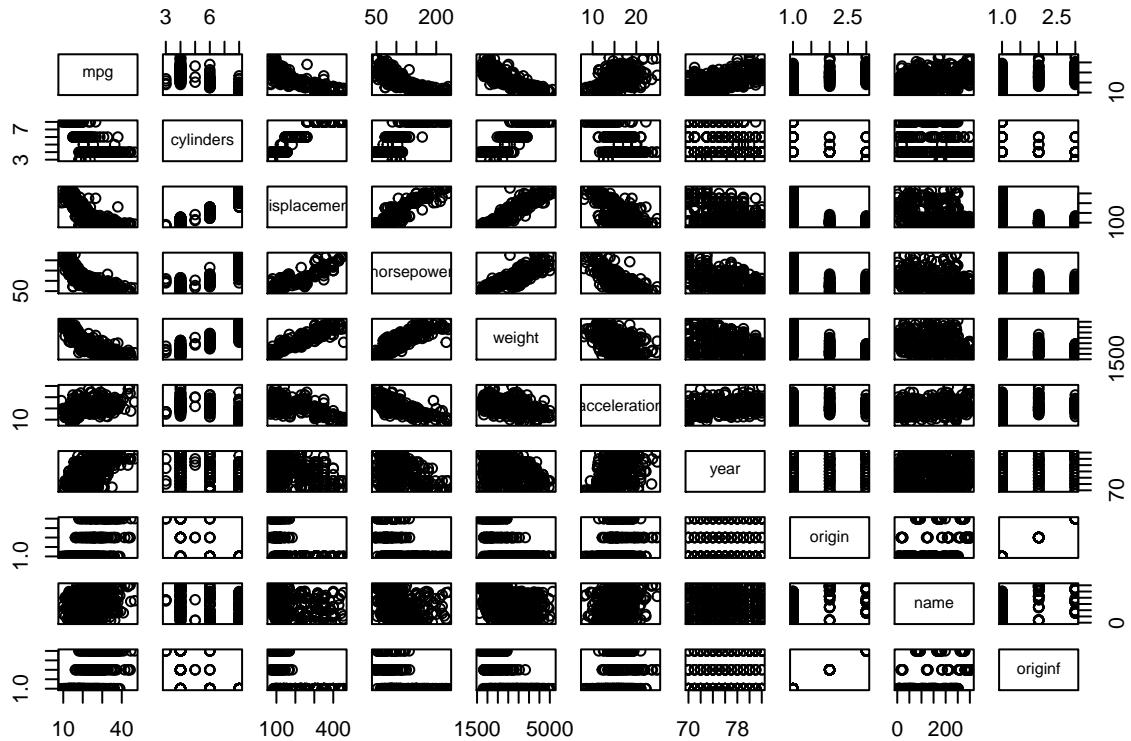
Box plot based on origin:

```
with(Auto, plot(originf, mpg), ylab = "mpg")
```



- (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.

```
pairs(Auto)
```



In the description, All the predictor mpg show some correlations with millage(mpg). The name predictor mpg has too little observations per name, so if we using this as a predictor is likely to get the results in overfitting the data and this dataset will not be able to well generalized.