Chapter_01

March 19, 2022

0.1 STATISTICAL LEARNING

0.1.1 Basic Functions:

rnorm( ) function generates a vector of random normal variables, with first argument ‘n’ the sample size

[1]:
```
x = rnorm(50)
y = x + rnorm(50, mean = 50, sd = .1)
```

In the above code snippet, ‘x’ is a vector of 50 random variables ‘y’ is a vector of 50 random variables that are distributed around the point/mean 50

[2]:
```
# What's the corelation value between x and y(round off to two decimal places)??
# use R's corelation function; ?cor() - describes the corelation function
# correlation = ?

# your code here

correlation = round(cor(x,y))
correlation
```

```
1
```

[3]:
```
round(correlation)
stopifnot(round(correlation,2) == 1)
```

```
1
```

0.1.2 Matrices:

Defining Matrices: Column First

[4]:
```
mat = matrix(c(1,2,3,4,5,6), 3, 2)
mat
```

```
                            1   4
A matrix: 3 × 2 of type dbl 2   5
                            3   6
```

Row First

[5]:
```r
mat = matrix(c(1,2,3,4,5,6), nrow = 3, ncol = 2, byrow = TRUE)
mat
```

```
                                1   2
A matrix: 3 × 2 of type dbl     3   4
                                5   6
```

Indexing data in matrices

[6]:
```r
mat[1,1]
mat[c(1,2), c(1,2)]
```

1

```
A matrix: 2 × 2 of type dbl     1   2
                                3   4
```

[7]:
```r
Matrix = matrix(1:20, 4 , 5)
Matrix
```

```
                                1   5   9    13   17
A matrix: 4 × 5 of type int     2   6   10   14   18
                                3   7   11   15   19
                                4   8   12   16   20
```

[8]:
```r
# Create a submatrix containing the values of ( 3rd & 4th row, 1st & 4th column)
# sub_matrix = ?

# your code here
sub_matrix = Matrix[c(3,4), c(1,4)]
```

[9]:
```r
stopifnot(sub_matrix[1,1] == 3, sub_matrix[2,2] == 16)
```

0.1.3 Loading Data (using ISLR2)

[10]:
```r
library(ISLR2) #import library ISLR2
```

[11]:
```r
head(Auto)
```

A data.frame: 6 × 9

	mpg <dbl>	cylinders <int>	displacement <dbl>	horsepower <int>	weight <int>	acceleration <dbl>	year <int>	origin <int>
1	18	8	307	130	3504	12.0	70	1
2	15	8	350	165	3693	11.5	70	1
3	18	8	318	150	3436	11.0	70	1
4	16	8	304	150	3433	12.0	70	1
5	17	8	302	140	3449	10.5	70	1
6	15	8	429	198	4341	10.0	70	1

[12]:

```
#what are the dimensions of the Auto dataset??
#?dim() - describe the dimension function
#auto_d = ?

# your code here
auto_d = dim(Auto)
auto_d
```

1. 392 2. 9

[13]:

```
stopifnot(matrix(auto_d)[1,1] == 392, matrix(auto_d)[2,1] == 9)
```

[14]:

```
#define all the columns of the Auto Dataset
#?names() - describe the columns of the data set
#cols = ?

# your code here
cols = names(Auto)
cols
```

1. 'mpg' 2. 'cylinders' 3. 'displacement' 4. 'horsepower' 5. 'weight' 6. 'acceleration' 7. 'year' 8. 'origin' 9. 'name'

[15]:

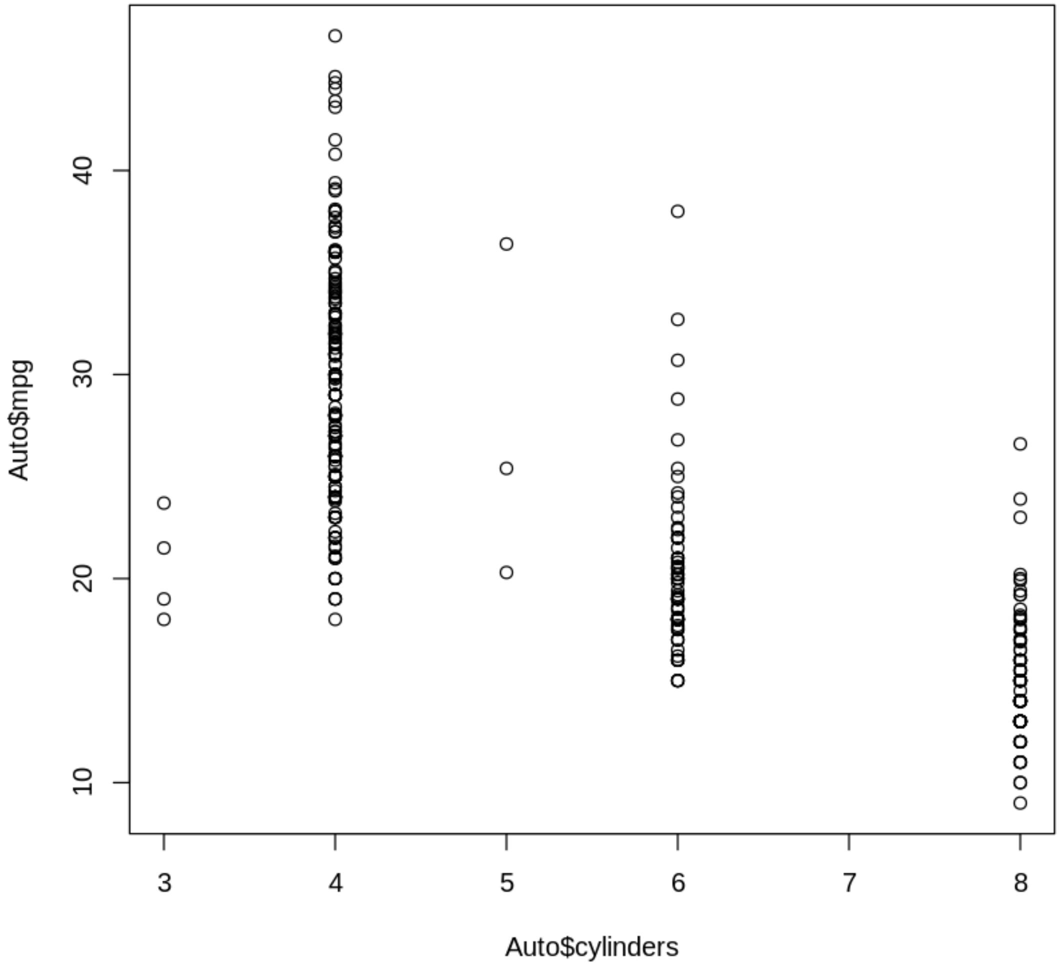
```
stopifnot(length(cols)== 9, cols[1]=='mpg')
```

0.1.4 Plot variables - Data set

Plotting variables of a data set

[16]:

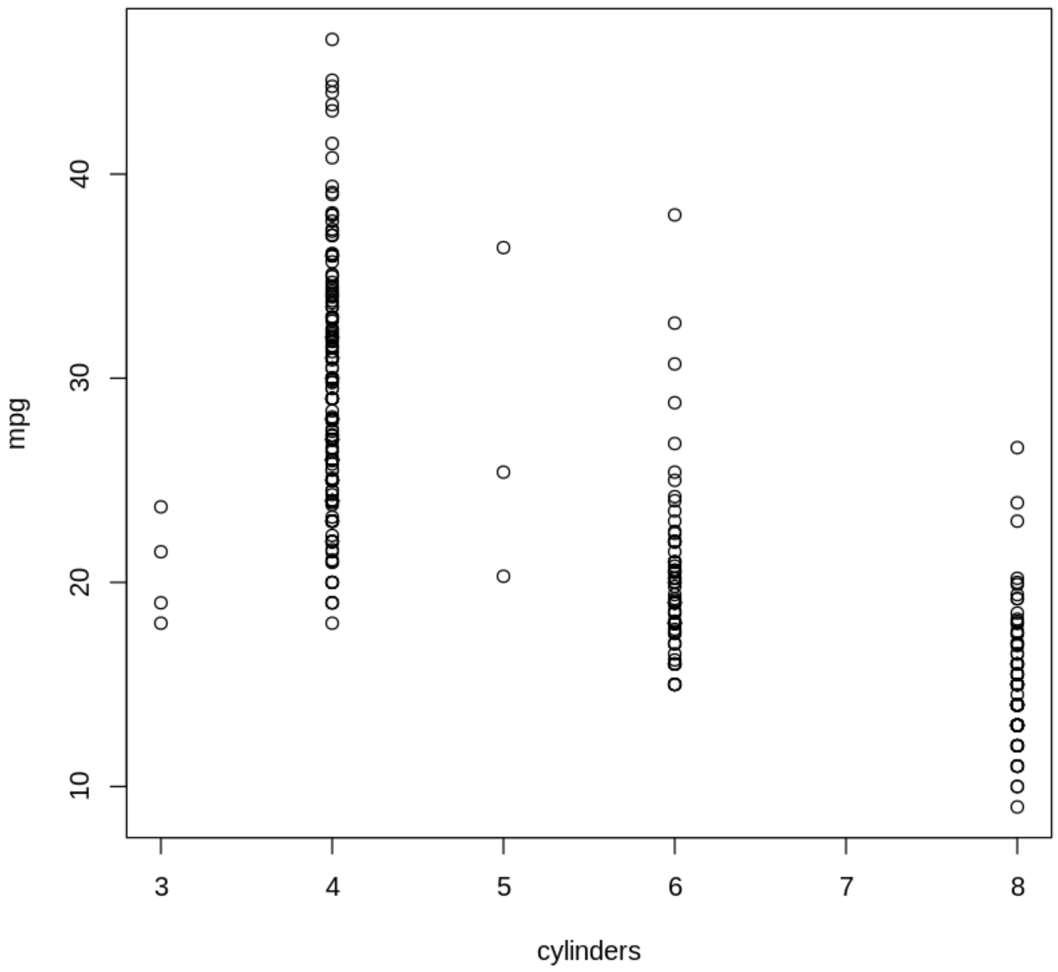
```
plot(Auto$cylinders, Auto$mpg)
```



Attach the data set to the context of R Reference the columns directly without use of the data set

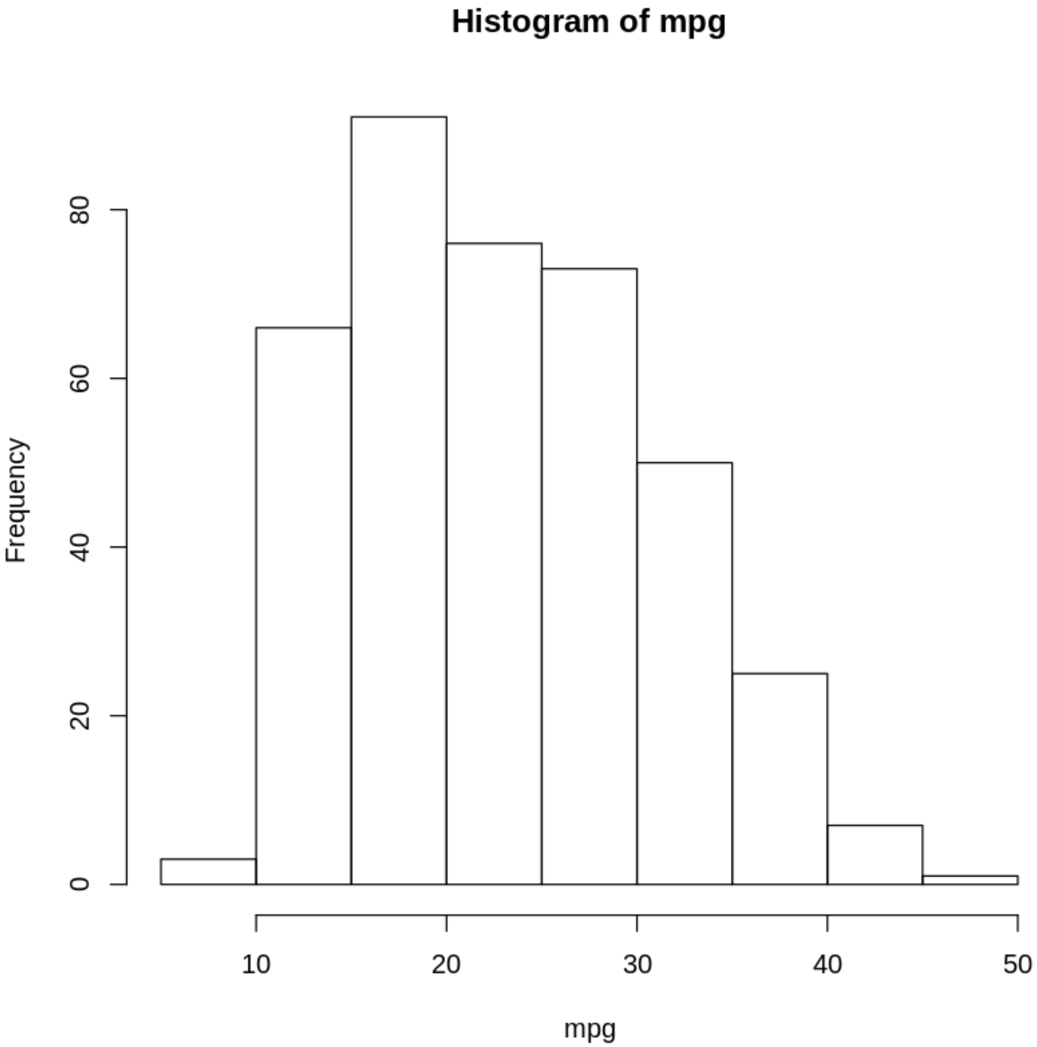
[17]:

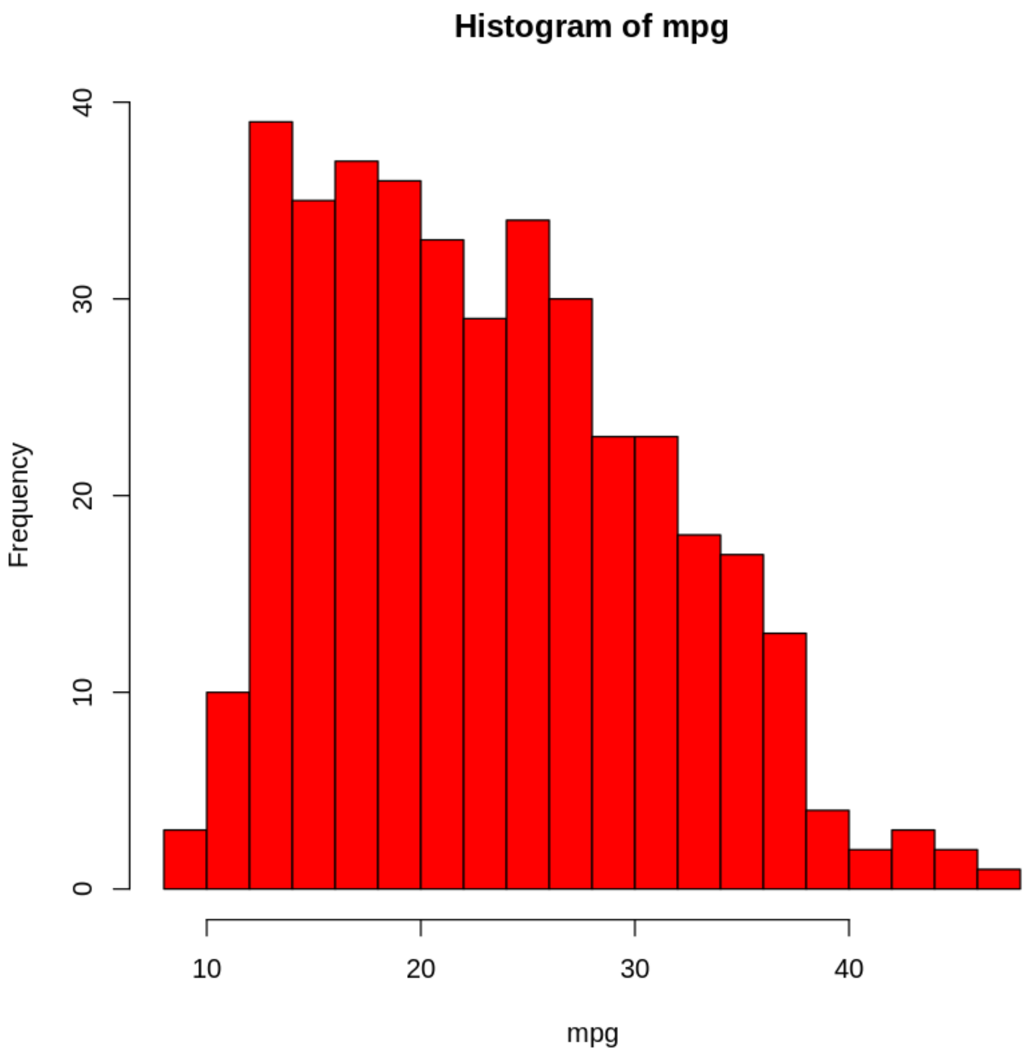
```
attach(Auto)
plot(cylinders,mpg)
```



Plotting Histograms

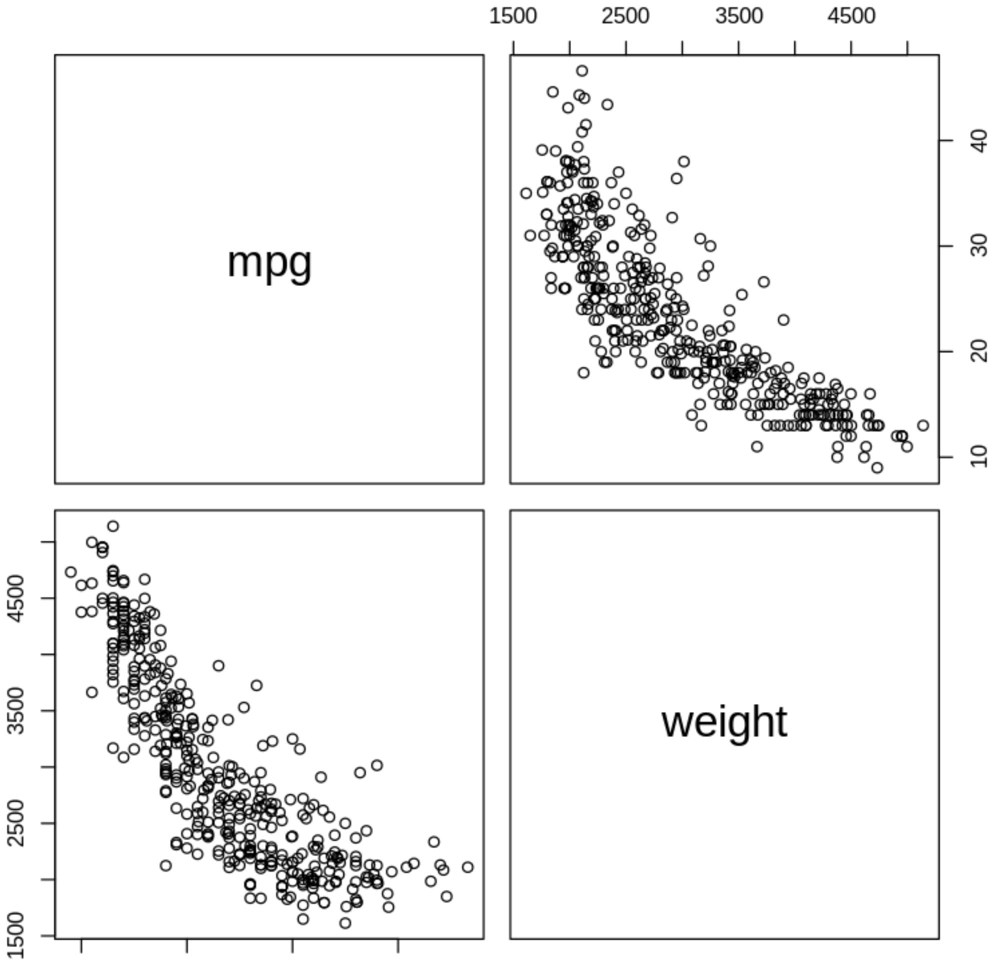
```
[18]: hist(mpg)
      hist(mpg, col=2, breaks = 15)
```


Histogram of mpg
Frequency
0
20
40
60
80
10
20
30
40
50
mpg

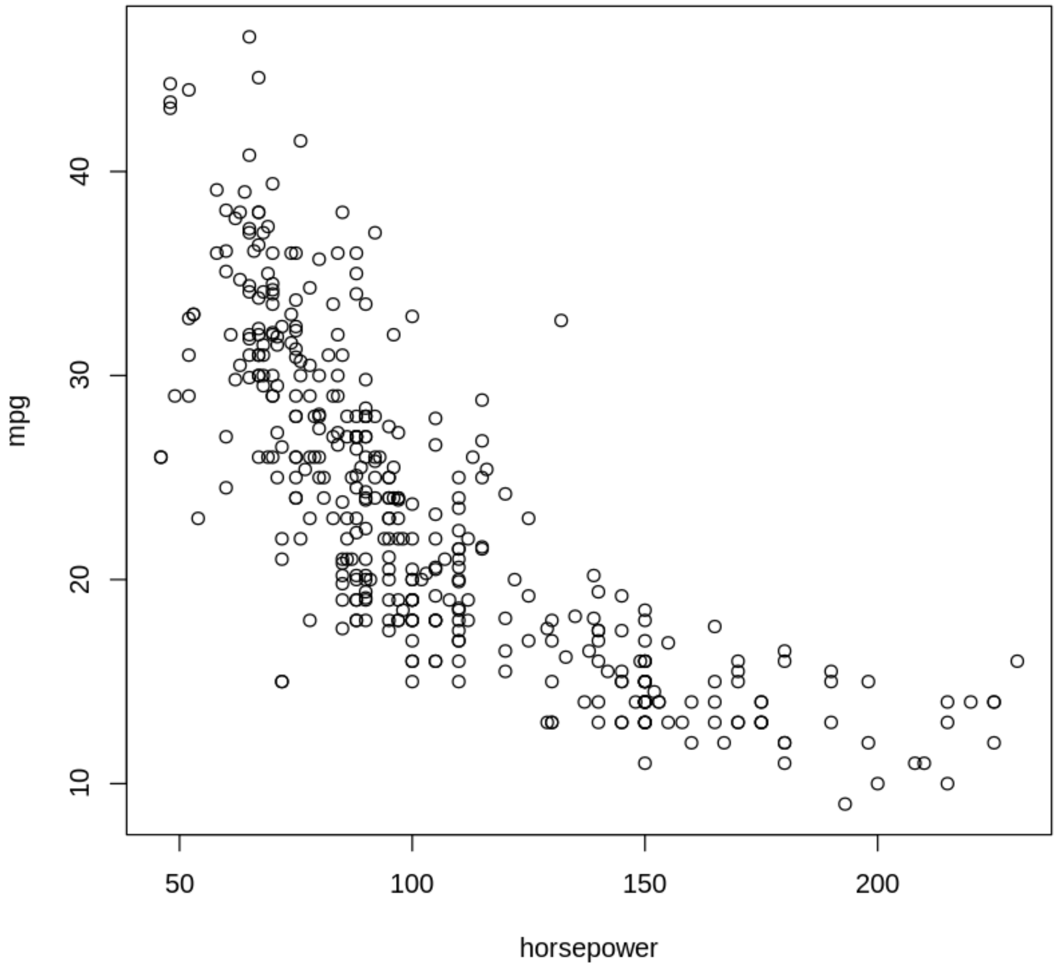


The pairs( ) function creates a scatterplot matrix, i.e. a scatterplot for every pair of variables.

```
[19]: pairs(~mpg + weight, data = Auto)
```



```
[20]: plot(horsepower, mpg)
      identify(horsepower, mpg, name)
```



Summary( ) Function - It produces a numerical summary of each variable in a particular data set

```
[21]: summary(Auto)
```

```
      mpg          cylinders      displacement     horsepower        weight    
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613  
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804  
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978  
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140  
                                                                               
  acceleration        year           origin                      name    
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5  
 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
 Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                 (Other)           :365
```

[22]:

```
#Find the summary of the mpg column
#mpg_summary = ?

# your code here
mpg_summary = summary(Auto$mpg)
```

[23]:

```
stopifnot(mpg_summary['Median'] == 22.75, mpg_summary['Max.'] == 46.60)
```