Data Analysis on the Auto, College and Boston Datsets

2024-07-01

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(skimr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(patchwork)
library(moderndive)
```

A. Read the data

```
college <- read.csv("/Users/admin/Downloads/College.csv")
```

B. Use the View() function

```
View(college)
```

C.i. A numerical summary of quantitative attributes

```
p = summary(college)
print(p)
```

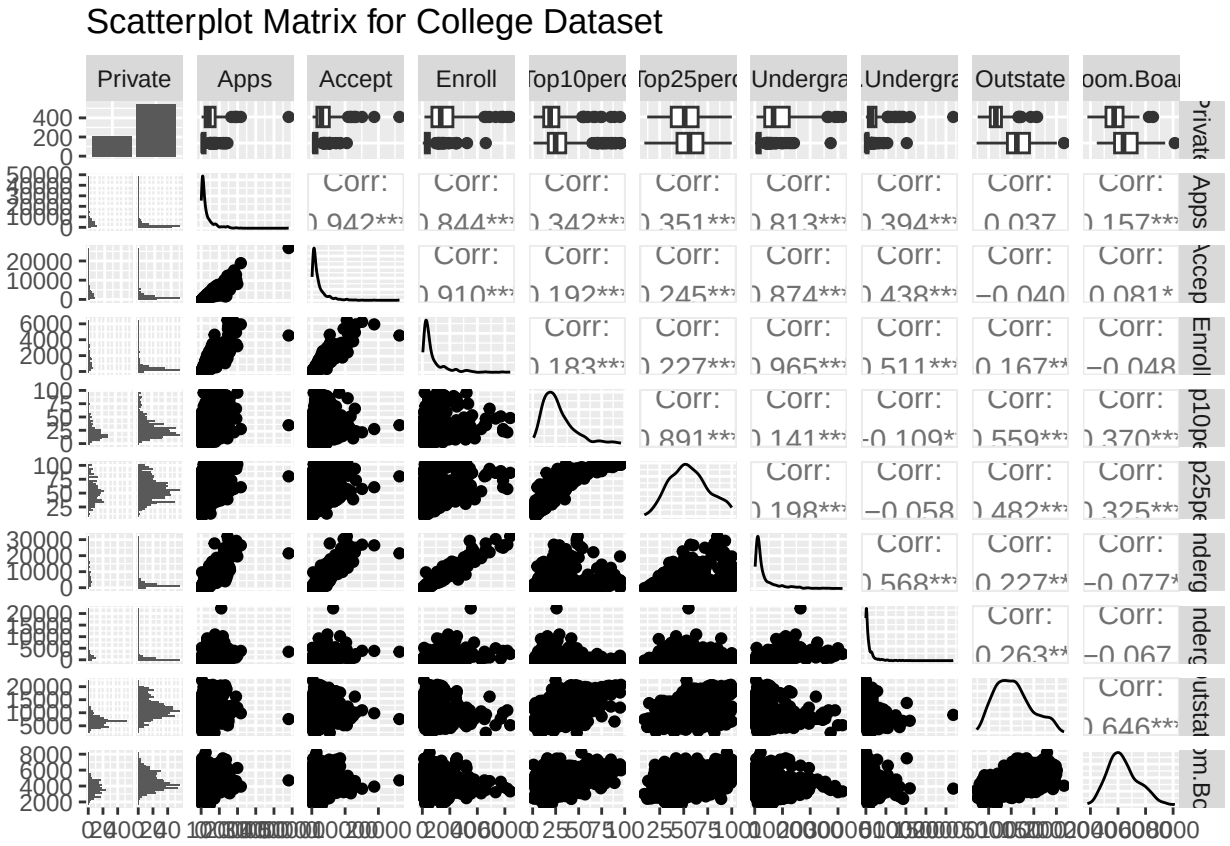
```
##       X                Private               Apps            Accept
##  Length:756         Length:756         Min.   :   81.0   Min.   :   72.0
##  Class :character   Class :character   1st Qu.:  786.5   1st Qu.:  615.5
##  Mode  :character   Mode  :character   Median : 1562.0   Median : 1111.5
##                                        Mean   : 3003.1   Mean   : 2020.3
##                                        3rd Qu.: 3629.5   3rd Qu.: 2428.0
##                                        Max.   :48094.0   Max.   :26330.0
```

```
##      Enroll         Top10perc       Top25perc      F.Undergrad
##  Min.   :  35.0   Min.   : 1.00   Min.   :  9.00   Min.   :  139
##  1st Qu.: 244.0   1st Qu.:15.00   1st Qu.: 41.00   1st Qu.: 1003
##  Median : 437.0   Median :23.50   Median : 54.00   Median : 1718
##  Mean   : 781.8   Mean   :27.62   Mean   : 55.88   Mean   : 3724
##  3rd Qu.: 902.2   3rd Qu.:35.00   3rd Qu.: 69.00   3rd Qu.: 4110
##  Max.   :6392.0   Max.   :96.00   Max.   :100.00   Max.   :31643
##   P.Undergrad         Outstate       Room.Board       Books
##  Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
##  1st Qu.:   99.0   1st Qu.: 7305   1st Qu.:3600   1st Qu.: 469.5
##  Median :  363.0   Median : 9990   Median :4200   Median : 502.0
##  Mean   :  865.6   Mean   :10443   Mean   :4358   Mean   : 547.2
##  3rd Qu.:  973.0   3rd Qu.:12931   3rd Qu.:5050   3rd Qu.: 600.0
##  Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
##     Personal         PhD            Terminal       S.F.Ratio
##  Min.   : 250   Min.   :  8.00   Min.   : 24.00   Min.   : 2.5
##  1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.00   1st Qu.:11.5
##  Median :1200   Median : 75.00   Median : 82.50   Median :13.6
##  Mean   :1341   Mean   : 72.73   Mean   : 79.81   Mean   :14.1
##  3rd Qu.:1700   3rd Qu.: 86.00   3rd Qu.: 92.00   3rd Qu.:16.5
##  Max.   :6800   Max.   :103.00   Max.   :100.00   Max.   :39.8
##   perc.alumni        Expend        Grad.Rate
##  Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:13.00   1st Qu.: 6742   1st Qu.: 53.00
##  Median :21.00   Median : 8408   Median : 65.00
##  Mean   :22.76   Mean   : 9659   Mean   : 65.55
##  3rd Qu.:31.00   3rd Qu.:10838   3rd Qu.: 78.00
##  Max.   :64.00   Max.   :56233   Max.   :118.00
```

C.ii. Scatterplot Matrix of the first ten columns of the quantitative data.

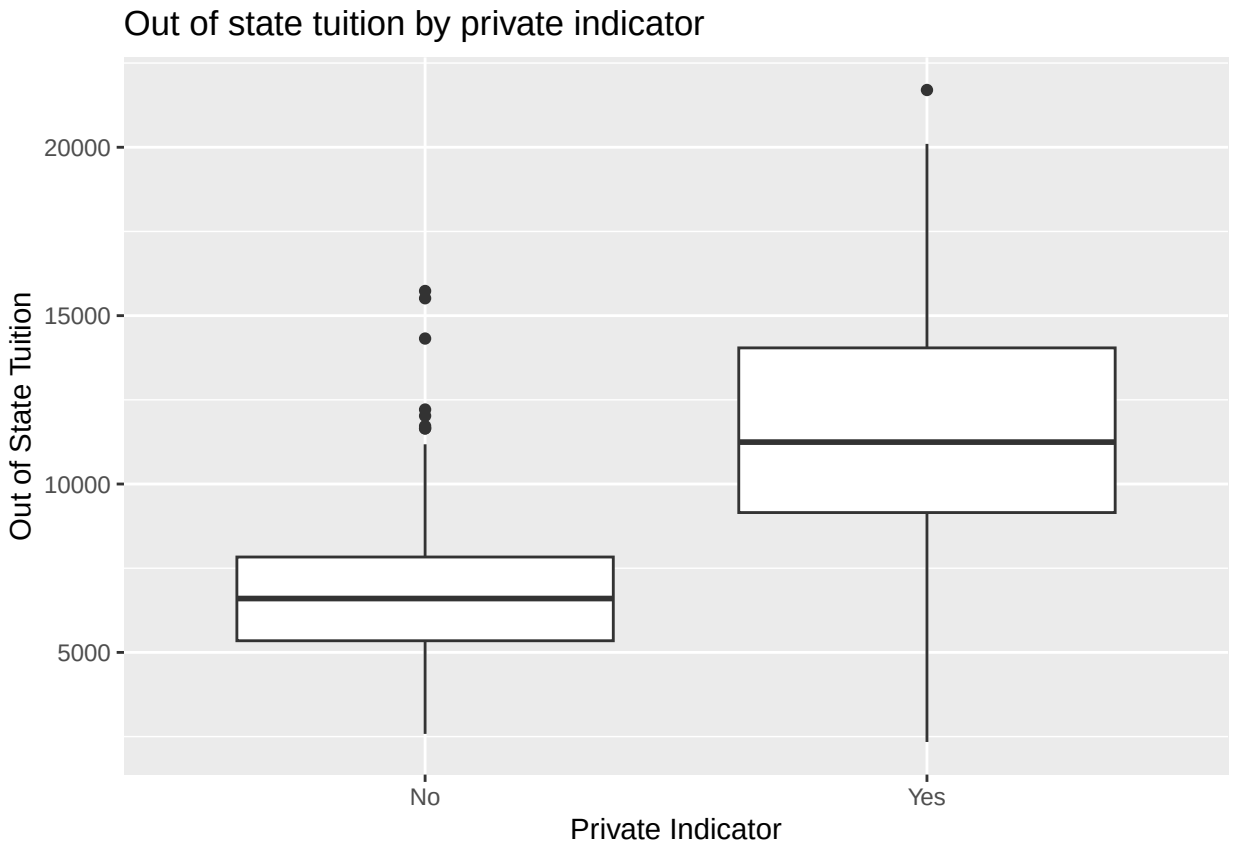
```
Scatter_matrix <- ggpairs(college, columns = c(2:11),
                          title = "Scatterplot Matrix for College Dataset")
Scatter_matrix
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



C.iii. Boxplots of Outstate versus Private

```
ggplot(college, aes(x= Private, y = Outstate)) +
  geom_boxplot() +
  labs(x = "Private Indicator", y = "Out of State Tuition",
       title = "Out of state tuition by private indicator")
```



C.iv.Creating a new qualitative variable, Elite, by binning the Top10perc attribute.

```
college$Elite <- cut(college$Top10perc, breaks=c(0, 50, 100),
                     include.lowest = T,
                     right = F)
```

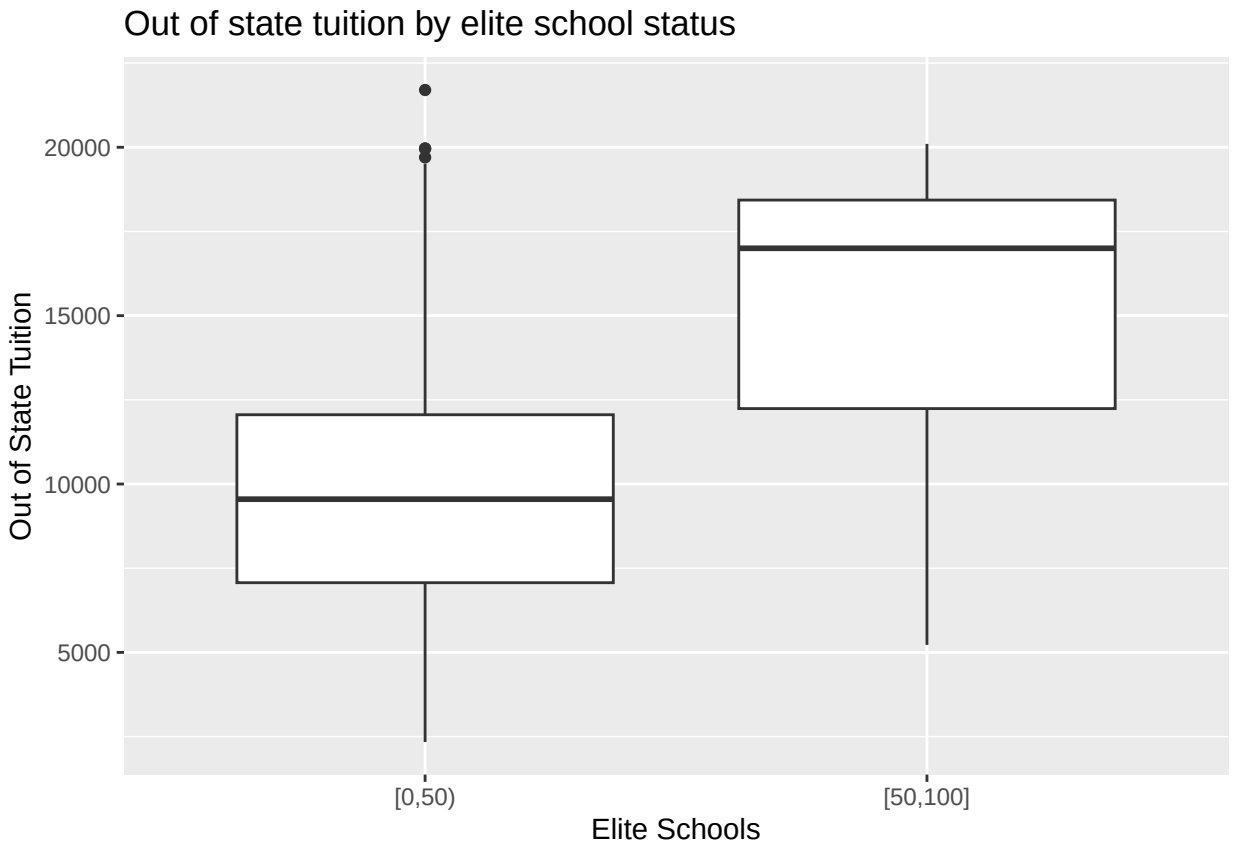
C.v. Using the summary() function to identify the number of elite universities.

```
summary(college$Elite)
```

```
##   [0,50) [50,100]
##      675       81
```

Producing boxplots of Outstate versus Elite

```
ggplot(college, aes(x= Elite, y = Outstate)) +
  geom_boxplot() +
  labs(x = "Elite Schools", y ="Out of State Tuition",
       title = "Out of state tuition by elite school status")
```



C.vi.Various histograms for quantitative variables

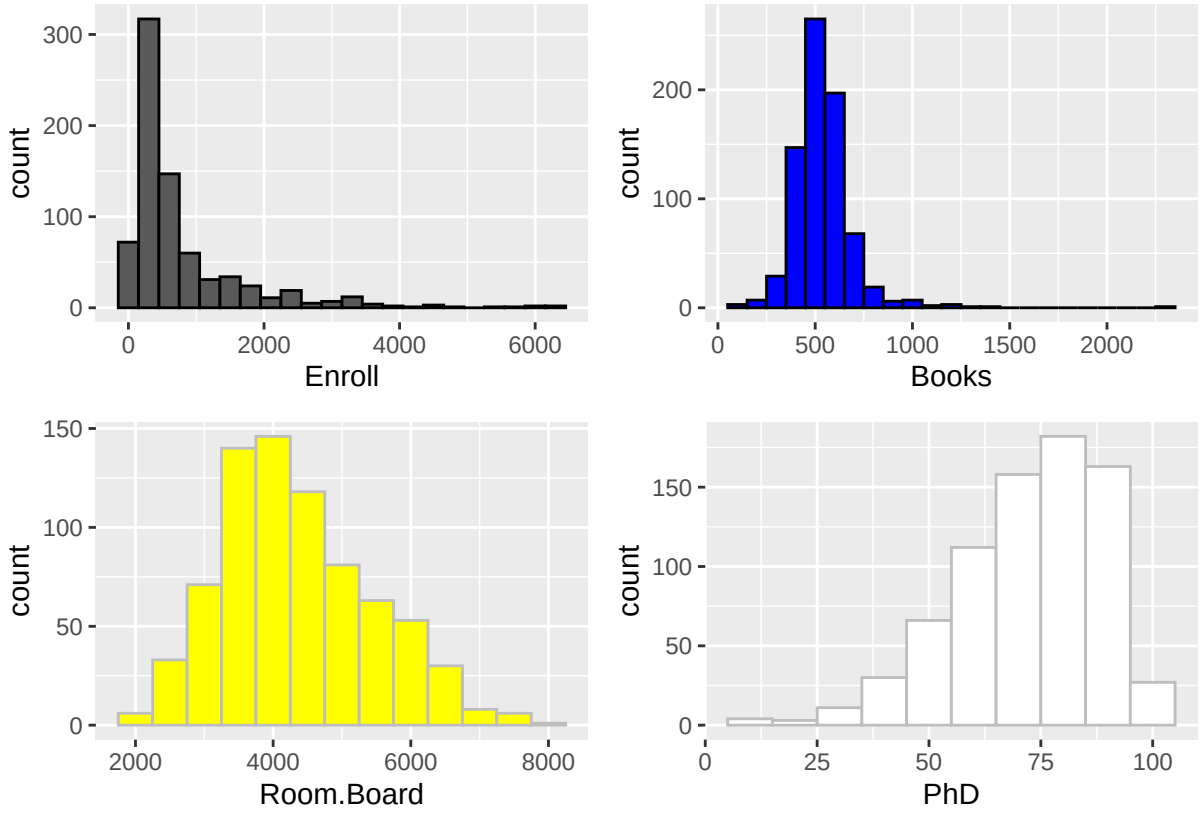
```
p1 <- ggplot(data= college, mapping= aes(x= Enroll))+
  geom_histogram(binwidth = 300, color = "black")

p2 <- ggplot(data = college, mapping=aes(x= Books)) +
  geom_histogram(binwidth = 100, fill= "blue", color= "black")

p3 <- ggplot(data= college, mapping=aes(x= Room.Board)) +
  geom_histogram(binwidth = 500, fill= "yellow", color= "grey")

p4 <-ggplot(data= college, mapping=aes(x= PhD)) +
  geom_histogram(binwidth = 10, fill= "white", color= "grey")

p1 + p2 + p3 + p4
```



C.vii. Exploratory Data Analysis

Reviewing the raw data

```
glimpse(college)
```

```
## Rows: 756
## Columns: 20
## $ X           <chr> "Abilene Christian University", "Adelphi University", "Adr~
## $ Private     <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Y~
## $ Apps        <int> 1660, 2186, 1428, 417, 193, 587, 353, 1899, 1038, 582, 173~
## $ Accept      <int> 1232, 1924, 1097, 349, 146, 479, 340, 1720, 839, 498, 1425~
## $ Enroll      <int> 721, 512, 336, 137, 55, 158, 103, 489, 227, 172, 472, 484,~
## $ Top10perc   <int> 23, 16, 22, 60, 16, 38, 17, 37, 30, 21, 37, 44, 38, 44, 23~
## $ Top25perc   <int> 52, 29, 50, 89, 44, 62, 45, 68, 63, 44, 75, 77, 64, 73, 46~
## $ F.Undergrad <int> 2885, 2683, 1036, 510, 249, 678, 416, 1594, 973, 799, 1830~
## $ P.Undergrad <int> 537, 1227, 99, 63, 869, 41, 230, 32, 306, 78, 110, 44, 638~
## $ Outstate    <int> 7440, 12280, 11250, 12960, 7560, 13500, 13290, 13868, 1559~
## $ Room.Board  <int> 3300, 6450, 3750, 5450, 4120, 3335, 5720, 4826, 4400, 3380~
## $ Books       <int> 450, 750, 400, 450, 800, 500, 500, 450, 300, 660, 500, 400~
## $ Personal    <int> 2200, 1500, 1165, 875, 1500, 675, 1500, 850, 500, 1800, 60~
## $ PhD         <int> 70, 29, 53, 92, 76, 67, 90, 89, 79, 40, 82, 73, 60, 79, 36~
## $ Terminal    <int> 78, 30, 66, 97, 72, 73, 93, 100, 84, 41, 88, 91, 84, 87, 6~
## $ S.F.Ratio   <dbl> 18.1, 12.2, 12.9, 7.7, 11.9, 9.4, 11.5, 13.7, 11.3, 11.5, ~
## $ perc.alumni <int> 12, 16, 30, 37, 2, 11, 26, 37, 23, 15, 31, 41, 21, 32, 26,~
## $ Expend      <int> 7041, 10527, 8735, 19016, 10922, 9727, 8861, 11487, 11644,~
## $ Grad.Rate   <int> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74, 68, 55~
## $ Elite       <fct> "[0,50)", "[0,50)", "[0,50)", "[50,100]", "[0,50)", "[0,50~
```

Display a random sample of 5 rows

```
college %>% sample_n(size = 5)
```

```
##                          X Private  Apps Accept Enroll Top10perc Top25perc
## 1     Syracuse University     Yes 10477   7260   2442        28        67
## 2          Wilson College     Yes   167    130     46        16        50
## 3       Butler University     Yes  2362   2037    700        40        68
## 4       Wilkes University     Yes  1631   1431    434        15        36
## 5 University of Rochester     Yes  8766   5498   1243        56        75
##   F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1       10142         117    15150       6870   635      960  73       84
## 2         199         676    11428       5084   450      475  67       76
## 3        2607         148    13130       4650   500     1600  77       81
## 4        1803         603    11150       5130   550     1260  78       92
## 5        5071         438    17840       6582   500      882  93       99
##   S.F.Ratio perc.alumni Expend Grad.Rate    Elite
## 1      11.3          13  14231        67   [0,50)
## 2       8.3          43  10291        67   [0,50)
## 3      10.9          29   9511        83   [0,50)
## 4      13.3          24   8543        67   [0,50)
## 5       5.9          23  26037        80 [50,100]
```

Computing summary statistics

The summary statistics show that there are 0 missing values. PhD is skewed left with values peaking on the right side, while S.F. Ratio and Top10per are both skewed right meaning values peak on the left side.

The statistics demonstrate that the mean percentage of faculty having a PhD is 72.7%, with a moderate spread of 16.4 (the standard deviation).

They show that the Student/Faculty ratio has a mean of 14.1 students per faculty member, with a standard deviation of 3.98.

The data also shows that the mean percentage of new students from the top 10% of high school class is 27.6 with a larger spread standard deviation of 17.6.

```
college %>%
  select(PhD,S.F.Ratio, Top10perc) %>%
  skim()
```

Table 1: Data summary

Name	Piped data
Number of rows	756
Number of columns	3
Column type frequency:	
numeric	3
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
PhD	0	1	72.73	16.40	8.0	62.0	75.0	86.0	103.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
S.F.Ratio	0	1	14.10	3.98	2.5	11.5	13.6	16.5	39.8	
Top10perc	0	1	27.62	17.61	1.0	15.0	23.5	35.0	96.0	

Correlation Matrix

There is a degree of collinearity between Expend and PhD explanatory variables as seen by the 0.4313352 correlation coefficient.

```
college %>%
  select(Top10perc, PhD, Expend) %>%
  cor()
```

```
##           Top10perc       PhD    Expend
## Top10perc 1.0000000 0.5298115 0.6621175
## PhD       0.5298115 1.0000000 0.4313352
## Expend    0.6621175 0.4313352 1.0000000
```

Correlation Coefficient

```
college %>%
  get_correlation(formula = PhD ~ Top10perc)
```

```
##         cor
## 1 0.5298115
```

There is a positive correlation between percentage of faculty with Ph.D's and new students from the top 10%.

```
p1 <- ggplot(data= college, mapping=aes(x= PhD, y = Top10perc)) +
  geom_point() +
  labs(x= "Percentage of faculty with Ph.D.'s", y = "New students from the top 10%",
       title = "Relationship between top 10% students and percentage of faculty with Ph.D.'s") +
  geom_smooth(method = "lm", se = FALSE)
```

Correlation Coefficient

```
college %>%
  get_correlation(formula = S.F.Ratio ~ Top10perc)
```

```
##          cor
## 1 -0.3899961
```

There is a slightly negative correlation between student/faculty ratio and new students from the top 10%.

```
p2 <- ggplot(data= college, mapping=aes(x= S.F.Ratio, y = Top10perc)) +
  geom_point() +
  labs(x= "Student/Faculty Ratio", y = "New students from the top 10%",
       title = "Relationship between top 10% students and and student/faculty ratio") +
  geom_smooth(method = "lm", se = FALSE)
```

Correlation Coefficient

```
college %>%
  get_correlation(formula = S.F.Ratio ~ Top10perc)
```

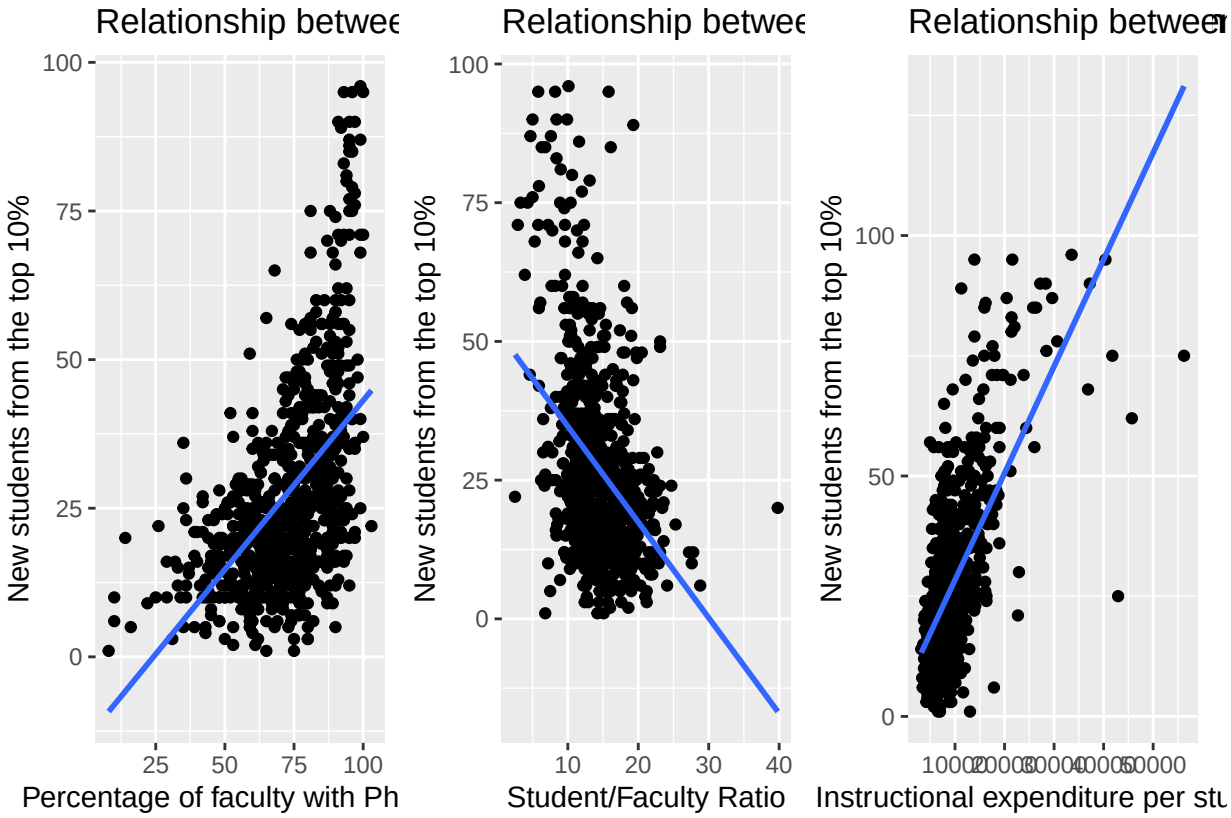
```
##          cor
## 1 -0.3899961
```

There is a positive correlation between instructional expenditure per student and new students from the top 10%.

```
p3 <- ggplot(data= college, mapping=aes(x= Expend, y = Top10perc)) +
  geom_point() +
  labs(x= "Instructional expenditure per student", y = "New students from the top 10%",
       title = "Relationship between top 10% students and instructional expenditure per student") +
  geom_smooth(method = "lm", se = FALSE)

p1 + p2 + p3
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Fit regression model

```
college_model <- lm(Top10perc ~ Expend + Private, data = college)
```

Get regression table

```
get_regression_table(college_model)
```

```
## # A tibble: 3 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept        6.30       1.15      5.45    0        4.03     8.56
## 2 Expend           0.002      0        23.5     0        0.002    0.002
## 3 Private: Yes    -0.262      1.11     -0.236   0.814   -2.44     1.92
```

Multiple Regression Model

Some trends from this model are that only private schools have funding greater than $20,000 per student and that there is a mix of new students from the top 10% in both private and public schools, althought it appears to be a greater number in private schools.

```
ggplot(college, mapping = aes(x = Expend, y = Top10perc, color = Private)) +
  geom_point() +
  labs(x = "Instructional expenditure per student", y = "New students from the top 10%",
       color = "Private School", title = "Multiple Regression Model")+
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

2a.

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

```
auto <- read.csv("/Users/admin/Downloads/Auto.csv")
auto <- na.omit(auto)
glimpse(auto)
```

```
## Rows: 371
## Columns: 9
## $ mpg          <dbl> 18, 15, 18, 16, 17, 15, 14, 14, 14, 15, 15, 14, 15, 14, 2~
## $ cylinders    <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 4, 6, 6, 6, 4, ~
## $ displacement <dbl> 307, 350, 318, 304, 302, 429, 454, 440, 455, 390, 383, 34~
## $ horsepower   <int> 130, 165, 150, 150, 140, 198, 220, 215, 225, 190, 170, 16~
## $ weight       <int> 3504, 3693, 3436, 3433, 3449, 4341, 4354, 4312, 4425, 385~
## $ acceleration <dbl> 12.0, 11.5, 11.0, 12.0, 10.5, 10.0, 9.0, 8.5, 10.0, 8.5, ~
```

```
## $ year         <int> 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 70, 7~
## $ origin       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, ~
## $ name         <chr> "chevrolet chevelle malibu", "buick skylark 320", "plymou~
```

2b.Ranges for each quantitative predictor

Converting the qualitative variables to factors.

```
auto$year <- as.factor(auto$year)
auto$origin <- as.factor(auto$origin)
auto$name <- as.factor(auto$name)
```

Converting the int and double variables to numeric

```
auto$mpg <- as.numeric(auto$mpg)
auto$cylinders <- as.numeric(auto$cylinders)
auto$displacement <- as.numeric(auto$displacement)
auto$horsepower <- as.numeric(auto$horsepower)
auto$weight <- as.numeric(auto$weight)
auto$acceleration <- as.numeric(auto$acceleration)
```

Finding the ranges for each quantitative predictor

```
range(auto$mpg)
```

```
## [1]  9.0 46.6
```

```
range(auto$cylinders)
```

```
## [1] 3 8
```

```
range(auto$displacement)
```

```
## [1]  68 455
```

```
range(auto$horsepower)
```

```
## [1]  46 230
```

```
range(auto$weight)
```

```
## [1] 1613 5140
```

```
range(auto$acceleration)
```

```
## [1]  8.0 24.8
```

2c. Mean and standard deviation for each quantitative predictor

```
auto %>%
  select(mpg, cylinders, displacement, horsepower, weight, acceleration) %>%
  skim()
```

Table 3: Data summary

Name	Piped data
Number of rows	371
Number of columns	6
Column type frequency:	

numeric	6
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
mpg	0	1	23.69	7.87	9	17.5	23.0	29.5	46.6	
cylinders	0	1	5.44	1.71	3	4.0	4.0	8.0	8.0	
displacement	0	1	192.08	105.25	68	99.5	144.0	284.5	455.0	
horsepower	0	1	104.39	39.00	46	75.0	92.0	127.0	230.0	
weight	0	1	2954.67	847.86	1613	2219.5	2755.0	3589.5	5140.0	
acceleration	0	1	15.50	2.76	8	13.7	15.5	17.0	24.8	

mpg: mean = 23.7, sd = 7.87 cylinders: mean = 5.44, sd = 1.71 displacement: mean = 192, sd= 105 horsepower: mean = 104, sd = 39 weight: mean = 2955, sd = 848 acceleration: mean = 15.5, sd = 2.76

2d.Calculations with 10 records removed

The removal of 10 records did not cause the calculations to change substantialy. In fact, the first time I completed the removal I did not see any changes in any of the variables. When I changed the records that were removed a second time, I finally saw one range change, but it only changed the minimum by 1.5.

```
removed <- 5:15
auto_less10 <- auto[-removed, ]

range(auto_less10$mpg)
```

```
## [1]  9.0 46.6
```

```
range(auto_less10$cylinders)
```

```
## [1] 3 8
```

```
range(auto_less10$displacement)
```

```
## [1]  68 455
```

```
range(auto_less10$horsepower)
```

```
## [1]  46 230
```

```
range(auto_less10$weight)
```

```
## [1] 1613 5140
```

```
range(auto_less10$acceleration)
```

```
## [1]  9.5 24.8
```

2e.Predicting gas mileage on the basis of other variables

Correlation

The correlation between cylinders and mpg is a strong correlation at -0.776896.

```r
cor(auto$cylinders, auto$mpg)
```

```
## [1] -0.776896
```

Setting the seed for reproducibility

```r
set.seed(1)
```

Sampling the dataset

```r
row.number <- sample(1:nrow(auto), 0.8 * nrow(auto))
```

Splitting the dataset into train and test sets

```r
train <- auto[row.number, ]
test <- auto[-row.number, ]

dim_train <- dim(train)
dim_test <- dim(test)
dim_train
```

```
## [1] 296   9
```

```r
dim_test
```

```
## [1] 75  9
```

Fit the linear model

```r
auto_model <- lm(mpg ~ cylinders, data = train)
summary(auto_model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.671  -3.098  -0.726   2.944  17.552
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.5379     0.9812   44.37   <2e-16 ***
## cylinders    -3.6224     0.1725  -20.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.09 on 294 degrees of freedom
## Multiple R-squared:  0.5999, Adjusted R-squared:  0.5985
## F-statistic: 440.8 on 1 and 294 DF,  p-value: < 2.2e-16
```

Make predictions on the test data

```r
predictions <- predict(auto_model, newdata = test)
```

Calculating Mean Squared Error and R-squared

```r
mse <- mean((test$mpg - predictions)^2)
mse
```

```
## [1] 19.64836
```

```
rss <- sum((test$mpg - predictions)^2)
tss <- sum((test$mpg - mean(test$mpg)) ^2)
rsquared <- 1 - (rss/tss)
rsquared
```

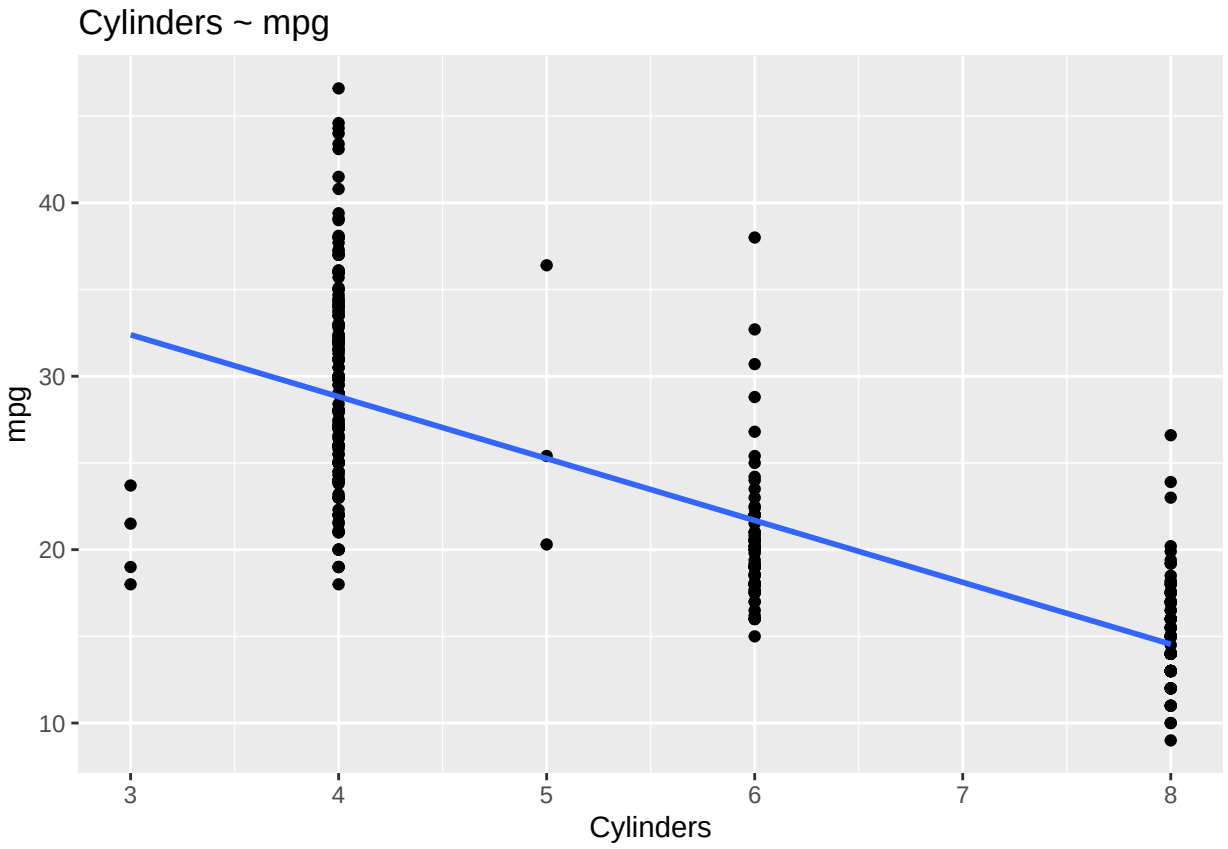
```
## [1] 0.6128595
```

The rsquared value is a little over halfway close to 1, so it isn't the strongest fit.

Scatterplot of cylinders and mpg

```
ggplot(auto, mapping = aes(x = cylinders, y = mpg)) +
  geom_point() +
  labs(x = "Cylinders", y = "mpg", title = "Cylinders ~ mpg") +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Correlation

The correlation between weight and mpg is a strong correlation at -0.8324161.

```
cor(auto$weight, auto$mpg)
```

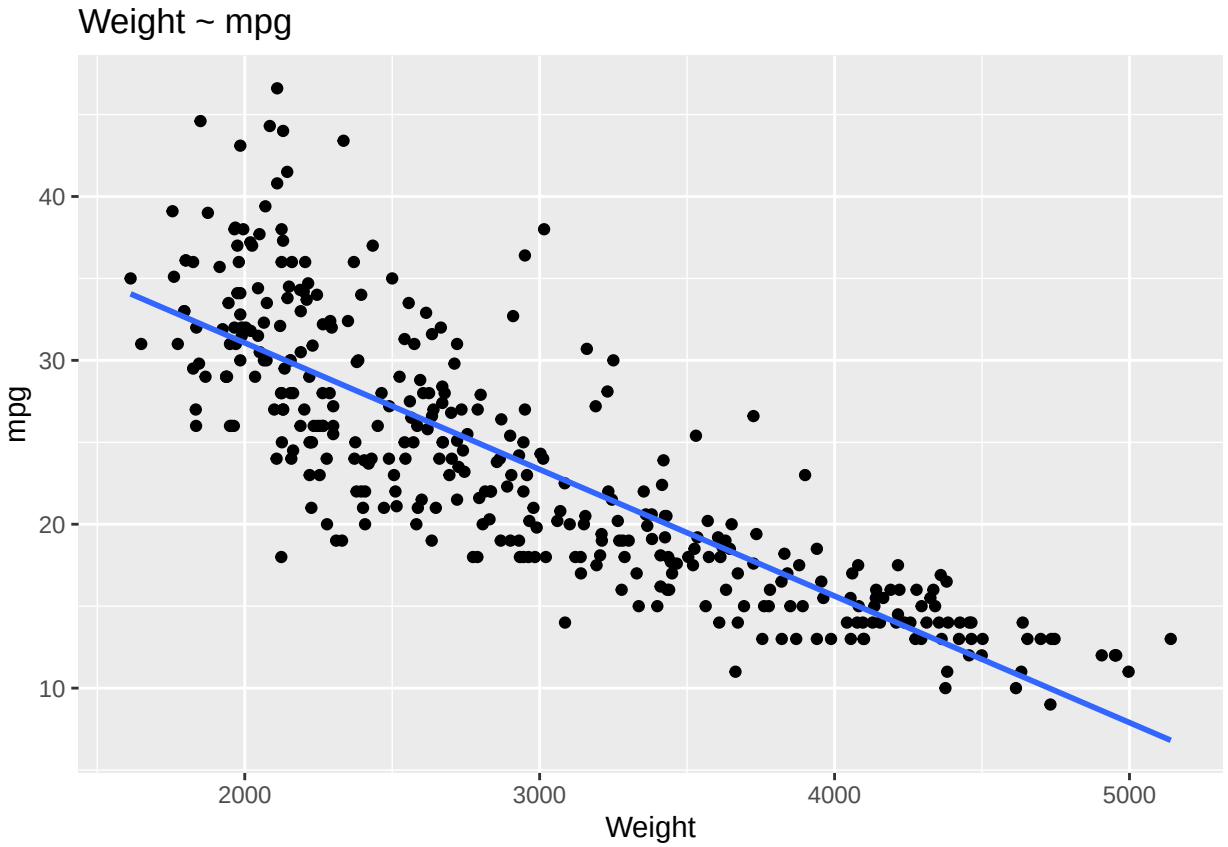
```
## [1] -0.8324161
```

Scatterplot of weight and mpg

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

```
  labs(x = "Weight", y = "mpg", title = "Weight ~ mpg") +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The analysis suggests that there is a correlation between more cylinders and a smaller mile per gallon and a higher weight and a smaller mile per gallon.

3a Loading the Boston dataset

The dataset contains 485 rows and 14 columns. The rows represent a single record in the datset whereas the columns represent the variables of the datset.

```
boston <- read.csv("/Users/admin/Downloads/boston.csv")
glimpse(boston)
```

```
## Rows: 485
## Columns: 14
## $ X       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829,~
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 1~
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87, 7.~
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524,~
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.631,~
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9, 9~
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9505~
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4,~
```

```
## $ tax     <int> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 31~
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15~
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10~
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9, 15~
```

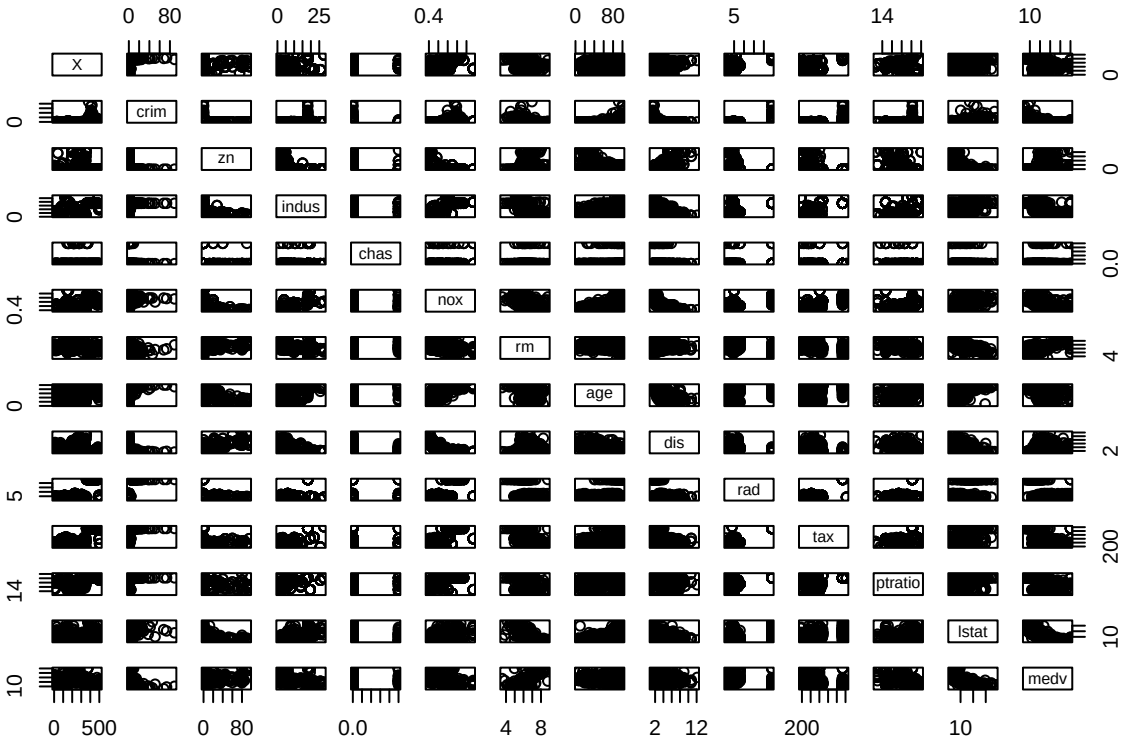
3b. Pairwise scatterpolots of the predictor columns

```
str(boston)
```

```
## 'data.frame':    485 obs. of  14 variables:
##  $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : int  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Converting ints to numeric values

```
boston$chas <- as.numeric(boston$chas)
boston$rad <- as.numeric(boston$rad)
boston$tax <- as.numeric(boston$tax)
pairs(boston)
```



The find-

ings indicate that there are some strong and weak correlations between some of the variables.

3c. Associations with per capita crime rate

```
correlations <- cor(boston)
correlations_with_crim <- correlations[, "crim"]
correlations_with_crim
```

```
##           X        crim          zn       indus        chas         nox
##  0.40088083  1.00000000 -0.19384321  0.39956287 -0.06105289  0.41409408
##          rm         age         dis         rad         tax     ptratio
## -0.22521094  0.34903165 -0.37594209  0.62284159  0.58005637  0.29166966
##       lstat        medv
##  0.45870243 -0.39228714
```

Some positive correlation between crime rate and taxes with a 0.5800564 correlation.

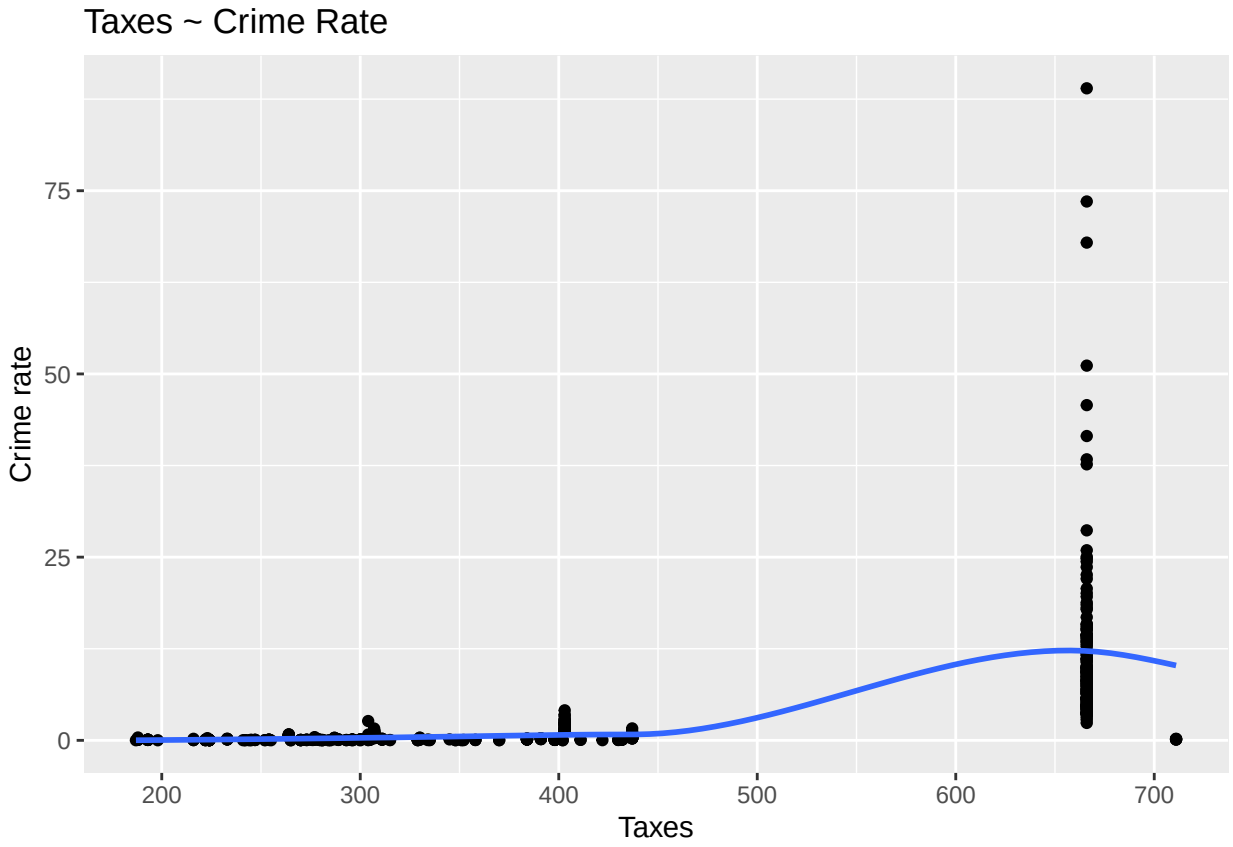
```
cor(boston$crim, boston$tax)
```

```
## [1] 0.5800564
```

```
ggplot(boston, mapping=aes(x= tax, y = crim)) +
  geom_point()+
  labs(x = "Taxes", y = "Crime rate", title = "Taxes ~ Crime Rate") +
  geom_smooth(model = "lm", se = FALSE)
```

```
## Warning in geom_smooth(model = "lm", se = FALSE): Ignoring unknown parameters:
## `model`
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Some positive correlation between crime rate and accessibility to radial highways with a 0.6228416 correlation.

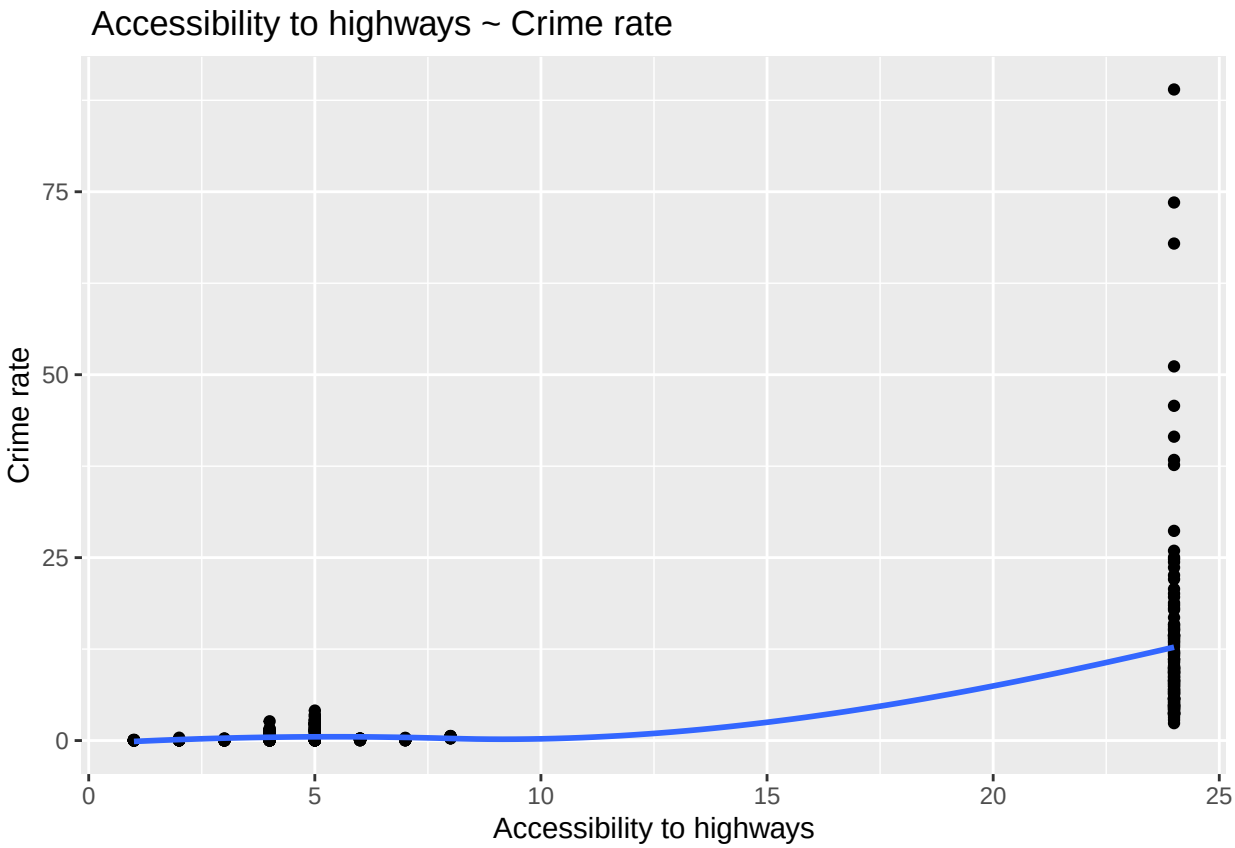
```
cor(boston$crim, boston$rad)
```

```
## [1] 0.6228416
```

```
ggplot(boston, mapping=aes(x= rad, y = crim)) +
  geom_point()+
  labs(x = "Accessibility to highways", y = "Crime rate", title = " Accessibility to highways ~ Crime ra
  geom_smooth(model = "lm", se = FALSE)
```

```
## Warning in geom_smooth(model = "lm", se = FALSE): Ignoring unknown parameters:
## `model`
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Very weak negative correlation between crime rate and average number of rooms with a -0.2252109 correlation.

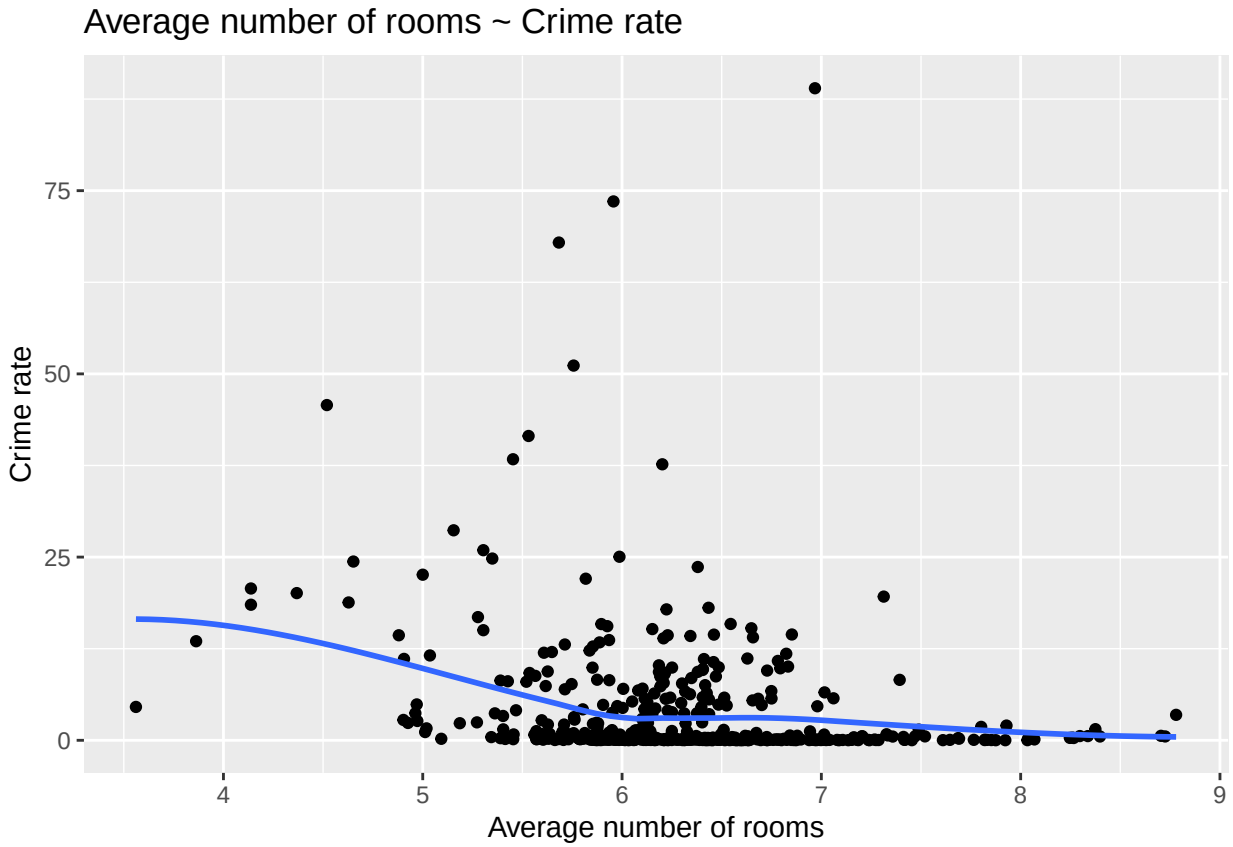
```
cor(boston$crim, boston$rm)
```

```
## [1] -0.2252109
```

```
ggplot(boston, mapping=aes(x= rm, y = crim)) +
  geom_point()+
  labs(x = "Average number of rooms", y = "Crime rate", title = "Average number of rooms ~ Crime rate")
  geom_smooth(model = "lm", se = FALSE)
```

```
## Warning in geom_smooth(model = "lm", se = FALSE): Ignoring unknown parameters:
## `model`
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



3d. Crime rates and tax rates of Boston suburbs

```
summary(boston$crim)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00632   0.08387   0.29819   3.76611   3.84970 88.97620
```

The average crime rate is 3.76, but some suburbs have much higher crime rates as evident by the maximum crime rate of 88.97620.

```
summary(boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   281.0   335.0   413.6   666.0   711.0
```

The tax rates vary greatly, with a minimum tax rate of $187 per $10,000 but some suburbs have up to $711 per $10,000.

```
summary(boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.10   18.47   20.20   22.00
```

The pupil teacher ratios have some variation with a minimum of 12.60 pupils to teachers up to 22 pupils per teachers in some suburbs.

3e. There are 35 suburbs in Boston that bound the Charles River.

```
boston %>%
  filter(chas == 1) %>%
  summarize(count = n())
```

```
##   count
## 1    35
```

3f. The median pupil-teacher ratio among the towns in the dataset is 19.1.

```
median(boston$ptratio)
```

```
## [1] 19.1
```

3g. The lowest value of median owner occupied homes is $5,000 and it belongs to the suburbs 399 and 406.

The value of the other predictors compare to overall range: Suburb 399: crime rate, percent lower status of the population, proportion of non-retail business acres per town, tax rate are over average, proportion of residential land zone for lots over 25,000 sq. ft. is at minimum and nixtric oxides concentration are below average.

Suburb 406: crime rate, percent lower status of the population,proportion of non-retail business acres per town, tax rate are over average, proportion of residential land zone for lots over 25,000 sq. ft. is at minimum and nixtric oxides concentration are below average. # Both suburbs have nearly identical numbers for the predictors.

```
summary(boston$medv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   16.70   21.20   22.53   25.00   50.00
```

```
boston %>%
  group_by(X) %>%
  filter(medv == 5)
```

```
## # A tibble: 2 x 14
## # Groups:   X [2]
##       X  crim    zn indus  chas   nox    rm   age   dis   rad   tax ptratio
##   <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1   399  38.4     0  18.1     0 0.693  5.45   100  1.49    24   666    20.2
## 2   406  67.9     0  18.1     0 0.693  5.68   100  1.43    24   666    20.2
## # i 2 more variables: lstat <dbl>, medv <dbl>
```

```
boston %>%
  filter(X == 399 | X == 406)
```

```
##     X    crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
## 1 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
## 2 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 22.98    5
```

```
summary(boston)
```

```
##        X              crim                zn             indus
##  Min.   :  1.0   Min.   : 0.00632   Min.   : 0.00   Min.   : 0.46
##  1st Qu.:143.0   1st Qu.: 0.08387   1st Qu.: 0.00   1st Qu.: 5.32
##  Median :264.0   Median : 0.29819   Median : 0.00   Median : 9.90
```

```
##  Mean   :261.9   Mean   : 3.76611   Mean   :10.23   Mean   :11.42
##  3rd Qu.:385.0   3rd Qu.: 3.84970   3rd Qu.: 0.00   3rd Qu.:18.10
##  Max.   :506.0   Max.   :88.97620   Max.   :95.00   Max.   :27.74
##       chas              nox               rm             age
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90
##  1st Qu.:0.00000   1st Qu.:0.4640   1st Qu.:5.887   1st Qu.: 46.30
##  Median :0.00000   Median :0.5380   Median :6.211   Median : 79.20
##  Mean   :0.07216   Mean   :0.5601   Mean   :6.290   Mean   : 69.55
##  3rd Qu.:0.00000   3rd Qu.:0.6310   3rd Qu.:6.629   3rd Qu.: 94.30
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00
##       dis              rad              tax           ptratio
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 2.065   1st Qu.: 4.000   1st Qu.:281.0   1st Qu.:17.40
##  Median : 3.067   Median : 5.000   Median :335.0   Median :19.10
##  Mean   : 3.646   Mean   : 9.748   Mean   :413.6   Mean   :18.47
##  3rd Qu.: 4.779   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 7.12   1st Qu.:16.70
##  Median :11.48   Median :21.20
##  Mean   :12.75   Mean   :22.53
##  3rd Qu.:17.12   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

3h. There are 64 suburbs that average more than 7 rooms per dwelling and there are 13 suburbs that average more than 8 rooms per dwelling.

The suburbs that average more than 8 rooms have an average crime rate of 0.71879, a range of taxes from $224 - $666 per $10,000 and an average distance of 3.430 to five Boston employment centers.

```
boston %>%
  filter(rm > 7) %>%
  summarize(count = n())
```

```
##   count
## 1    62
```

```
boston %>%
  filter(rm > 8) %>%
  summarize(count = n())
```

```
##   count
## 1    13
```

```
eight_rooms <- boston %>%
  filter(rm > 8)

summary(eight_rooms)
```

```
##        X              crim               zn            indus
##  Min.   : 98.0   Min.   :0.02009   Min.   : 0.00   Min.   : 2.680
##  1st Qu.:225.0   1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970
##  Median :233.0   Median :0.52014   Median : 0.00   Median : 6.200
##  Mean   :232.3   Mean   :0.71879   Mean   :13.62   Mean   : 7.078
##  3rd Qu.:258.0   3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200
##  Max.   :365.0   Max.   :3.47428   Max.   :95.00   Max.   :19.580
```

```
##       chas              nox               rm             age        
##  Min.   :0.0000   Min.   :0.4161   Min.   :8.034   Min.   : 8.40  
##  1st Qu.:0.0000   1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40  
##  Median :0.0000   Median :0.5070   Median :8.297   Median :78.30  
##  Mean   :0.1538   Mean   :0.5392   Mean   :8.349   Mean   :71.54  
##  3rd Qu.:0.0000   3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50  
##  Max.   :1.0000   Max.   :0.7180   Max.   :8.780   Max.   :93.90  
##       dis              rad              tax           ptratio     
##  Min.   :1.801   Min.   : 2.000   Min.   :224.0   Min.   :13.00  
##  1st Qu.:2.288   1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70  
##  Median :2.894   Median : 7.000   Median :307.0   Median :17.40  
##  Mean   :3.430   Mean   : 7.462   Mean   :325.1   Mean   :16.36  
##  3rd Qu.:3.652   3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40  
##  Max.   :8.907   Max.   :24.000   Max.   :666.0   Max.   :20.20  
##      lstat            medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```