HW2 R Markdown

Santanu Mukherjee

9/11/2021

R Markdown

Chapter 2 page 52:

Q1

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method. Justify your answer.

a. The sample size n is extremely large, and the number of predictors p is small.

Answer : BETTER. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.

b. The number of predictors p is extremely large, and the number of observations n is small.

Answer : WORSE. A flexible method will cause overfitting because of the small number of observations.

c. The relationship between the predictors and response is highly non-linear.

Answer : BETTER. With more degrees of freedom, a flexible method would fit better than an inflexible one to find the non-linear effect.

d. The variance of the error terms, i.e., σ 2 = Var(ε), is extremely high.

Answer : WORSE. A flexible method would capture too much of the noise in the data because of large variance of the errors.

Q2

Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

Answer : Regression and inference with n=500 and p=3

b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Answer : Classification and prediction with n=20 and p=13

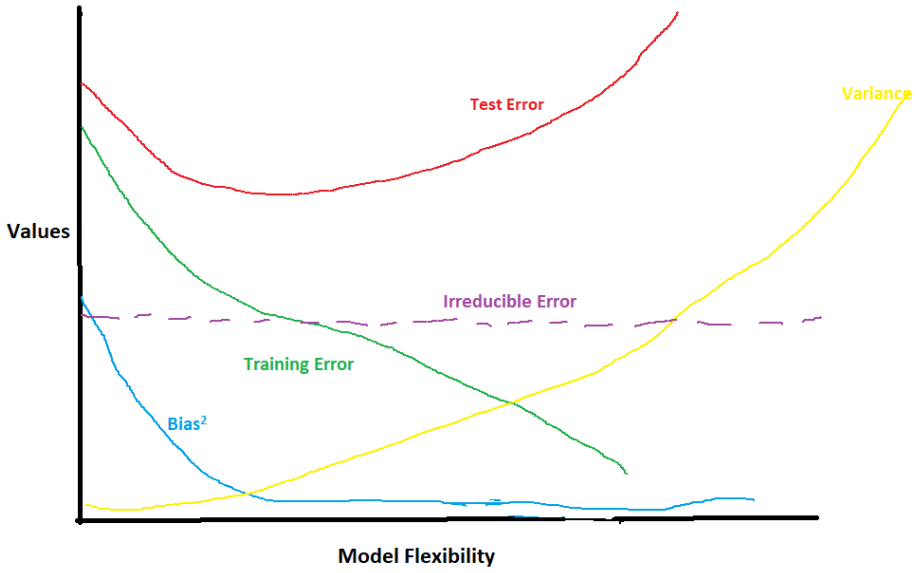
c. We are interested in predicting the % change in the USD / Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record

the % change in the USD / Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Answer : Regression and prediction with n=52 and p=3

Q3

a)Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



Sketch with the five curves

b)Explain why each of the five curves has the shape displayed in part (a).

Bias is the error introduced when the complexity of a problem is not sufficiently modeled by the simplicity of the chosen method (e.g. linear regression for non-linear relationships). As model flexibility increases (linear->trees->boosting, decreasing K in KNN, etc.), bias decreases monotonically, because less assumptions are being made about the data structure and its relationship with the response.

Variance refers to the amount by which our predictions would change if the training data were changed, and can be thought of as the error introduced when a model is overfit to the training data. As model flexibility increases, variance increases monotonically, because the method becomes more specified (and then overspecified) to the nuances of the training data, to the point where $\hat{f}$ doesn't generalize to new data.

Training Error decreases monotonically as flexibility increases. More flexible methods are generally higher variance, and can learn more complex relationships more completely, but also run the risk of overfitting, which is seen where the training error and test error diverge. Think of a decision tree, where the number of terminal nodes = the number of training observations (this model will have 0 training error and a high test error).

Test Error decreases, levels-out then increases. The minima is the point of optimal bias-variance tradeoff, where E[(Y− $\hat{Y}$)^2]=[Bias($\hat{f}$ (X))]^2+Var($\hat{f}$ (X))+Var(ϵ) is minimized. To the right of this minima, the method is overfitting ($\hat{f}$ is too high variance to make up for its lack of bias), and to the left the method is underfitting ($\hat{f}$ is too

high bias to make up for its lack of variance).

Irreducible Error refers to the error introduced by inherent uncertainty/noise in the system being approximated. It is constant and > 0 regardless of the flexibility of the model, because ϵ may contain unmeasured variables not in X that could be used to predict y, and because ϵ may contain unmeasurable variation in y that could not be accounted for in X even if we wanted to. This means that it doesn't matter how closely $\hat{f}$ models the 'true' function f, there will still be an (unknown) minimum error of Var(ϵ)>0.

Q10

Part a)

To begin, load in the Boston data set.The Boston data set is part of the MASS library in R. How many rows are in this data set? How many columns? What do the rows and columns represent?:

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

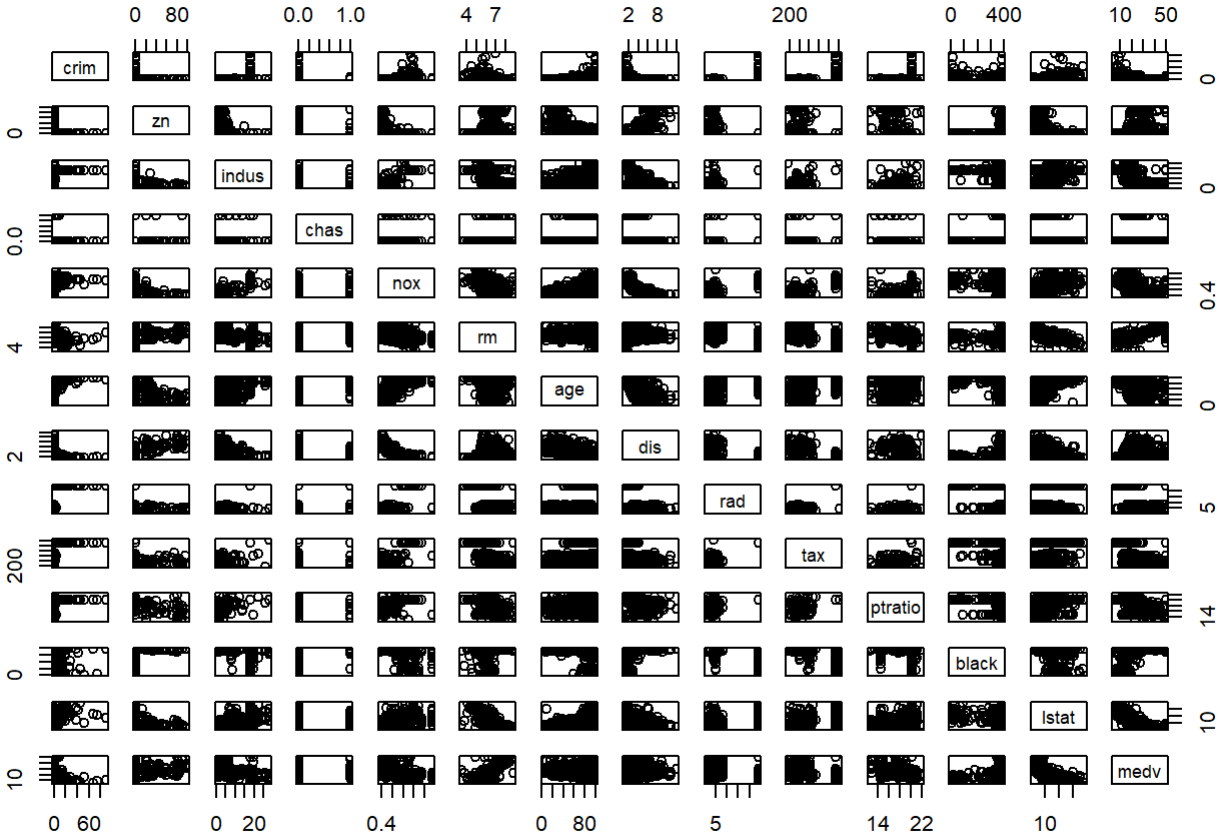
```
## [1] 506  14
```

Answer a) The Boston data frame has 506 rows and 14 columns.This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Massachusetts.Each row represent the set of predictor observations for a given Neighborhood in Boston. Each column represent each predictor variable for which an observation was made in 506 neighborhoods of Boston.

Part b)

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

```
## 'data.frame':    506 obs. of  14 variables:
##  $ 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    : num  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 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ 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 ...
```



From the diagram, it looks like certain variables appear to be correlated. A correlation matrix would be helpful to find the correlation.

Part c)

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

```
##         rad         tax       lstat         nox       indus        medv
##  0.62550515  0.58276431  0.45562148  0.42097171  0.40658341 -0.38830461
##       black         dis         age     ptratio          rm          zn
## -0.38506394 -0.37967009  0.35273425  0.28994558 -0.21924670 -0.20046922
##        chas
## -0.05589158
```

The code above provides the correlation coefficient between crime rates and other variables and is printed in order of absolutes values. So , we can see that the variables rad, tax is positively correlated and is above 0.5 and the variable Chas has a very low negative correlation.

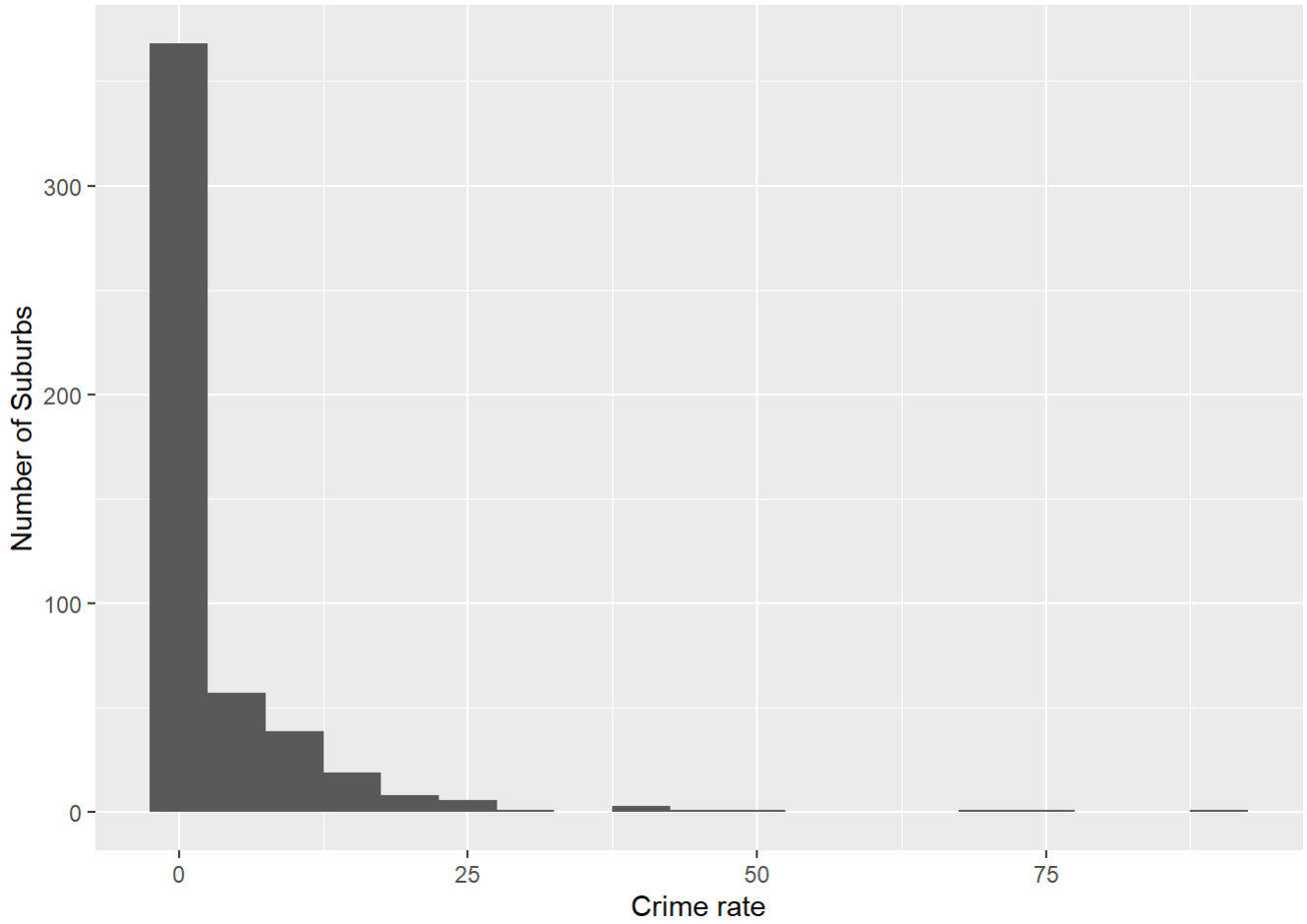
Part d)

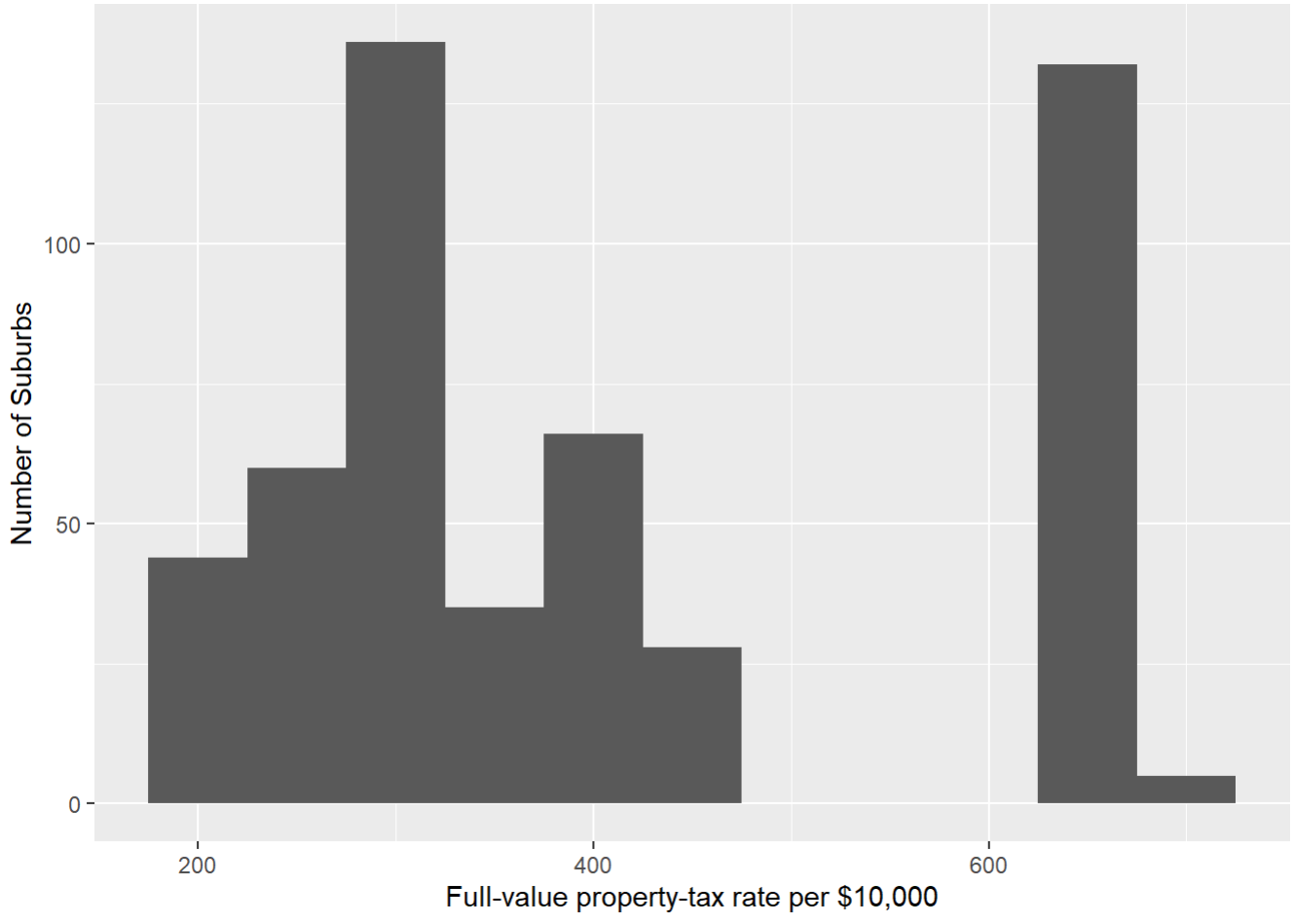
Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

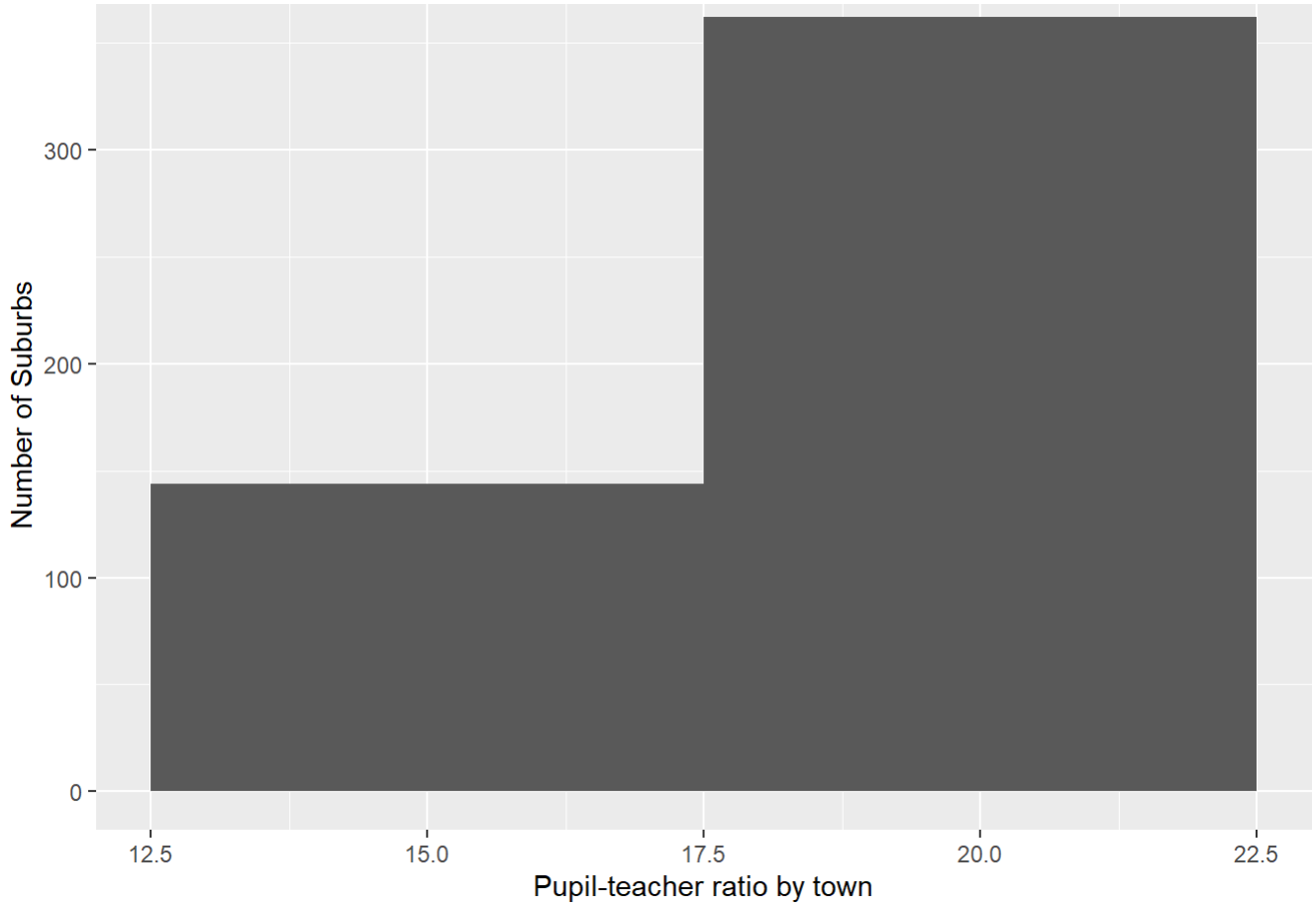
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```


Number of Suburbs
300
200
100
0
0
25
50
75
Crime rate


Number of Suburbs
100
50
0
200
400
600
Full-value property-tax rate per $10,000



We see that the median and maximum crime rate values are respectively about 0.26% and 89%. The data points to the fact that there are some neighborhoods where the crime rate is alarmingly high.

```
## [1] 0.1067194
```

11% of the neighborhoods have crime rates above 10%

```
## [1] 0.02173913
```

2% of the neighborhoods have crime rates above 25%

```
## [1] 0.007905138
```

0.8% of the neighborhoods have crime rates above 50%

Based on the histogram of the Tax rates, they are few neighborhoods where rates are relative higher. The median and average tax amount are $330 and $408.20 (per Full-value property-tax rate per $10,000) respectively.

```
## [1] 0.729249
```

73% of the neighborhood pay tax less than $600

```
## [1] 0.270751
```

27% of the neighborhood pay tax $600 or greater

Part e)

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

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

There are 35 suburbs in the Boston data set that bound the Charles river.

Part f)

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.60   17.40   19.05   18.46   20.20   22.00
```

The median pupil-teacher ratio is 19 pupils for each teacher.

Part 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?

```
##        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    20.2 396.9 30.59
##     medv
## 399    5
```

Suburb #399 with a median value of $5000.

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

Based on the summary information, here are some facts:

• Crime is very high compared to median and average rates of all Boston neighborhoods.
• No residential land zoned for lots over 25,000 sq.ft. This applies to more than half of the neighborhoods in Boston.
• Proportion of non-retail business acres per town is very high compared to most suburbs.
• This suburb is not one of the suburbs that bound the Charles river.
• Nitrogen oxides concentration (parts per 10 million) is one of the highest.
• Average number of rooms per dwelling is one of the lowest.
• Highest proportion of owner proportion of owner-occupied units built prior to 1940.
• One of the lowest weighted mean of distances to five Boston employment centers.
• Highest index of accessibility to radial highways.
• One of the highest full-value property-tax rate per $10,000.
• One of the highest pupil-teacher ratio by town.
• Highest value for 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
• One of the highest lower status of the population (percent).
• Lowest median value of owner-occupied homes in $1000s.

Based on the list above, suburb 399 can be classified as one of the least desirable places to live in Boston.

Part 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.

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

There are 64 suburbs with more than 7 rooms per dwelling.

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

There are 13 suburbs with more than 8 rooms per dwelling.

```
##       crim               zn             indus             chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      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
```

• Crime is very less compared to the overall Boston neighborhoods.
• No residential land zoned for lots over 25,000 sq.ft. This applies to more than half of the neighborhoods in Boston.
• Proportion of non-retail business acres per town is low compared to most suburbs.
• This suburb is not one of the suburbs that bound the Charles river.
• One of the lowest weighted mean of distances to five Boston employment centers.
• Not so high index of accessibility to radial highways.
• Average - full-value property-tax rate per $10,000.
• One of the highest pupil-teacher ratio by town.
• Moderate value for 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
• Not in One of the highest lower status of the population (percent).
• It has the median median value of owner-occupied homes in $1000s.

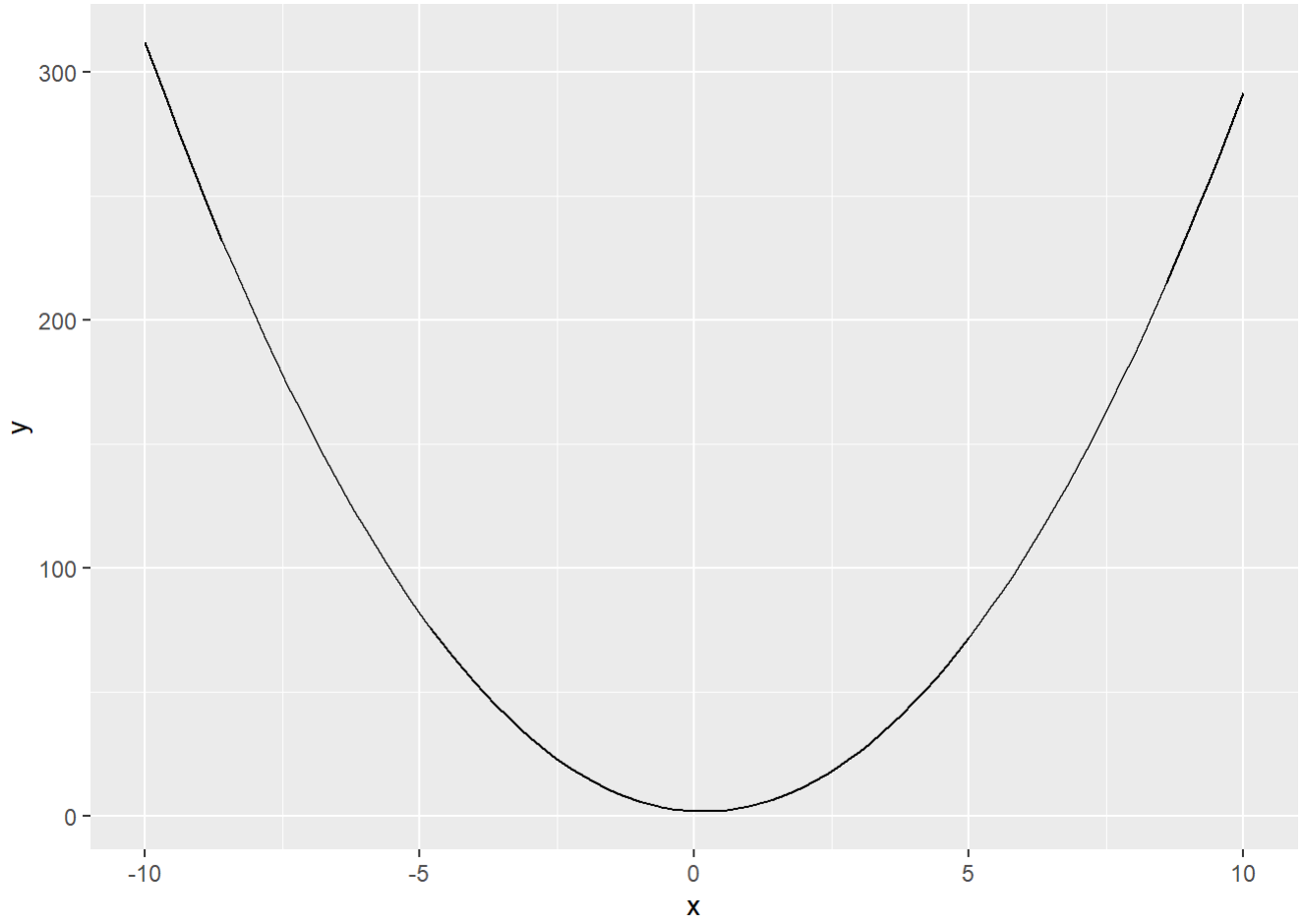
Not from the text Book:

Question 1

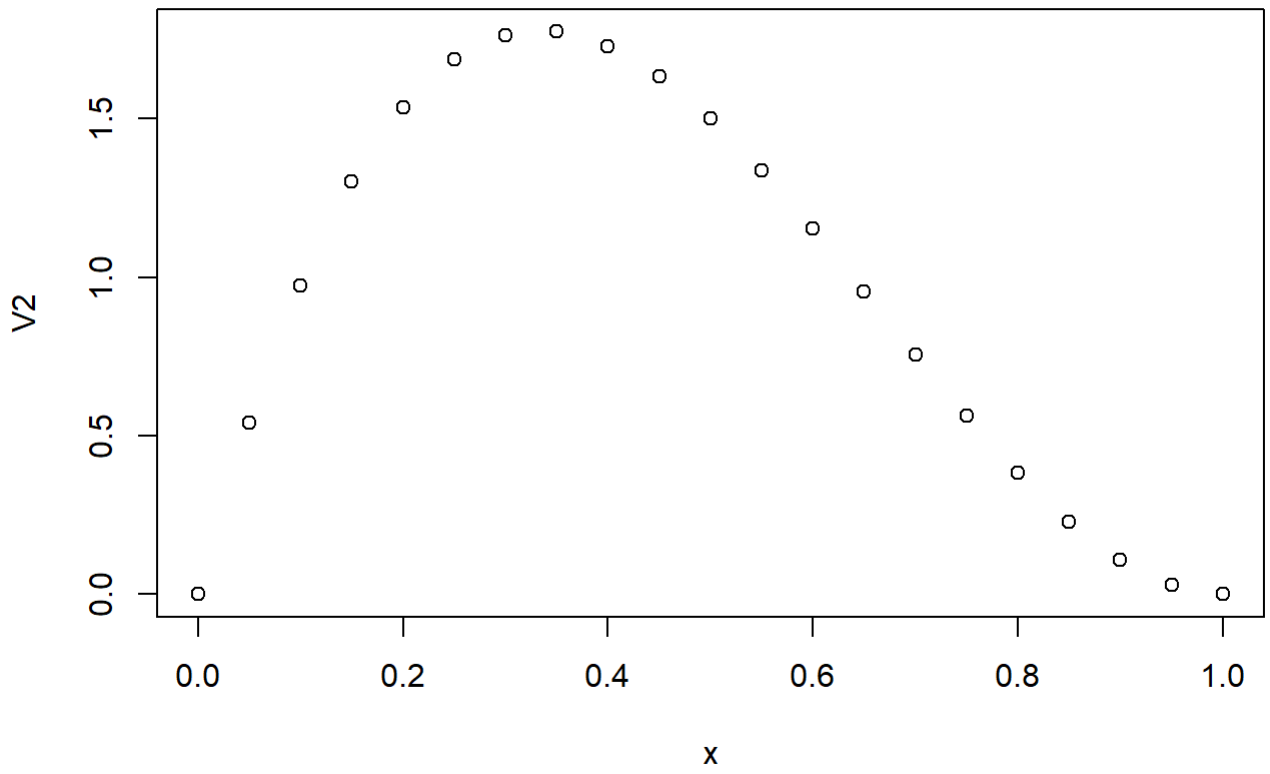
Suppose you have the following functions. Write an R function to each of them and then make a plot for each one

a. $f(x) = 2 + 3x^2 - x$, in the range of (-10,10).

```
##  [1] 312 254 202 156 116  82  54  32  16   6   2   4  12  26  46  72 104 142 186
## [20] 236 292
```



b. $f(x) = 1/B(2,3) * x(1-x)^2$, for 0 < **x** < 1, where **B(,)** is a beta function, in the range of (0,1).



Question 2

Create a dataframe with the following command

```
set.seed(123)
df = data.frame(x1 = rnorm(10), x2 = rpois(10,3), x3 = runif(10,-1,1), x4 =
rgamma(10,2,3))
```

```
##             x1 x2          x3        x4
## 1  -0.56047565  5  0.92604847 0.1590834
## 2  -0.23017749  4  0.80459809 0.3726197
## 3   1.55870831  3  0.38141056 0.2467049
## 4   0.07050839  8  0.59093484 1.0325922
## 5   0.12928774  4 -0.95077263 0.4424049
## 6   1.71506499  4 -0.04440806 0.5645746
## 7   0.46091621  3  0.51691908 0.1433038
## 8  -1.26506123  3 -0.56718413 0.3778739
## 9  -0.68685285  2 -0.36363798 0.3867932
## 10 -0.44566197  1 -0.53674843 0.4485639
```

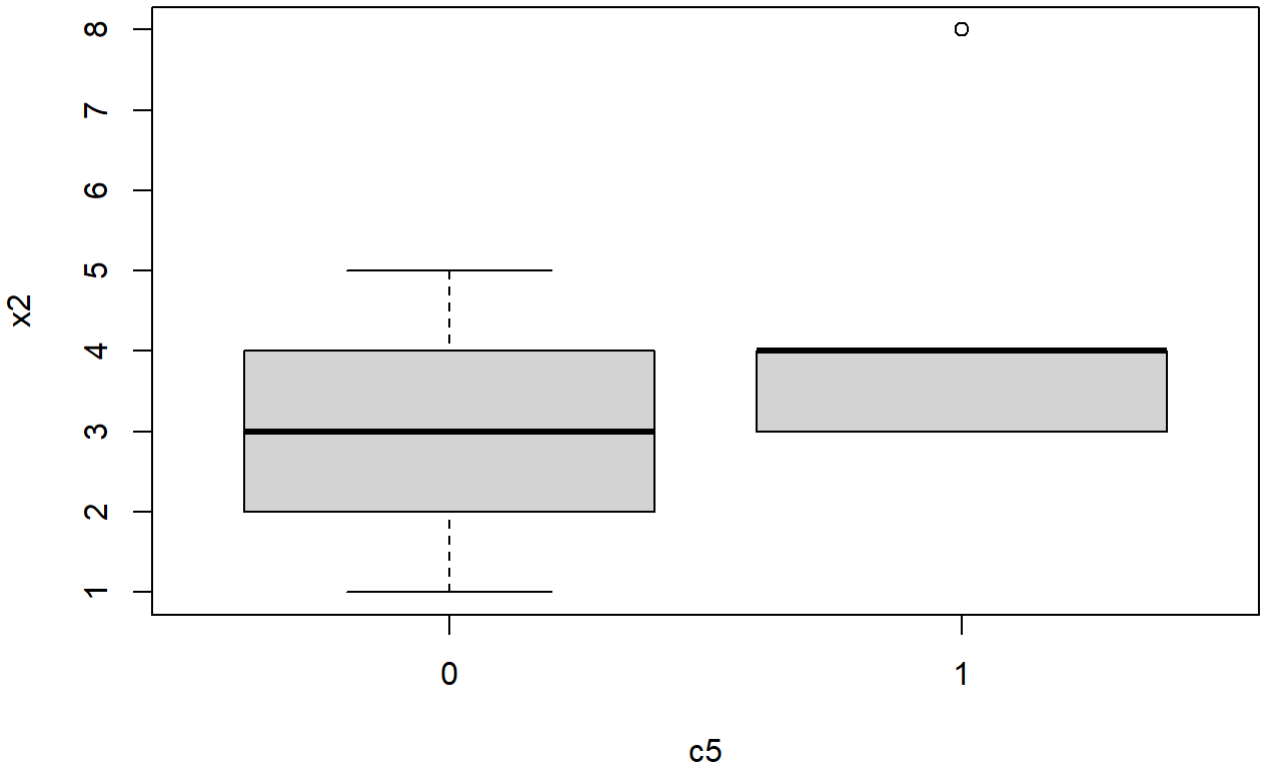
a. Obtain the means of all columns using apply().

```
##         x1         x2         x3         x4
## 0.07462564 3.70000000 0.07571598 0.41745145
```

b. Add another column, named c5, which is 1 for all x1>=0 and 0 otherwise.

```
##             x1 x2          x3        x4 c5
## 1  -0.56047565  5  0.92604847 0.1590834  0
## 2  -0.23017749  4  0.80459809 0.3726197  0
## 3   1.55870831  3  0.38141056 0.2467049  1
## 4   0.07050839  8  0.59093484 1.0325922  1
## 5   0.12928774  4 -0.95077263 0.4424049  1
## 6   1.71506499  4 -0.04440806 0.5645746  1
## 7   0.46091621  3  0.51691908 0.1433038  1
## 8  -1.26506123  3 -0.56718413 0.3778739  0
## 9  -0.68685285  2 -0.36363798 0.3867932  0
## 10 -0.44566197  1 -0.53674843 0.4485639  0
```

c. Draw box plots of x2 for different c5 values.



Question 3

In this problem, you search the internet using “auto_mpg dataset” to find an automobile mpg data. Most likely you would be able to find it in either at “kaggle”, or “UCI Machine Learning Repository.” Note that you do not use the original data.

a) Create a new project and import the data into your RStudio. Show the first 3 rows of the data by using head command.

```
## # A tibble: 6 x 9
##     mpg cylinders displacement horsepower weight acceleration `model year`
##   <dbl>     <dbl>        <dbl> <chr>       <dbl>        <dbl>        <dbl>
## 1    18         8          307 130          3504         12             70
## 2    15         8          350 165          3693         11.5           70
## 3    18         8          318 150          3436         11             70
## 4    16         8          304 150          3433         12             70
## 5    17         8          302 140          3449         10.5           70
## 6    15         8          429 198          4341         10             70
## # ... with 2 more variables: origin <dbl>, car name <chr>
```

b) Check the classes of your variables by using sapply command. What are the classes of horsepower, model_year and name?

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    "numeric"    "numeric"    "numeric"  "character"    "numeric"    "numeric"
##   model year       origin     car name
##    "numeric"    "numeric"  "character"
```

Horsepower is character, Model Year is Numeric and Car Name is character.

c) From the original data, horsepower is supposed to be numeric. Do you see any problem? In R, any missing value is labeled as "NA". Try to clean the data (actually the horsepower column) and replace any character to "NA".

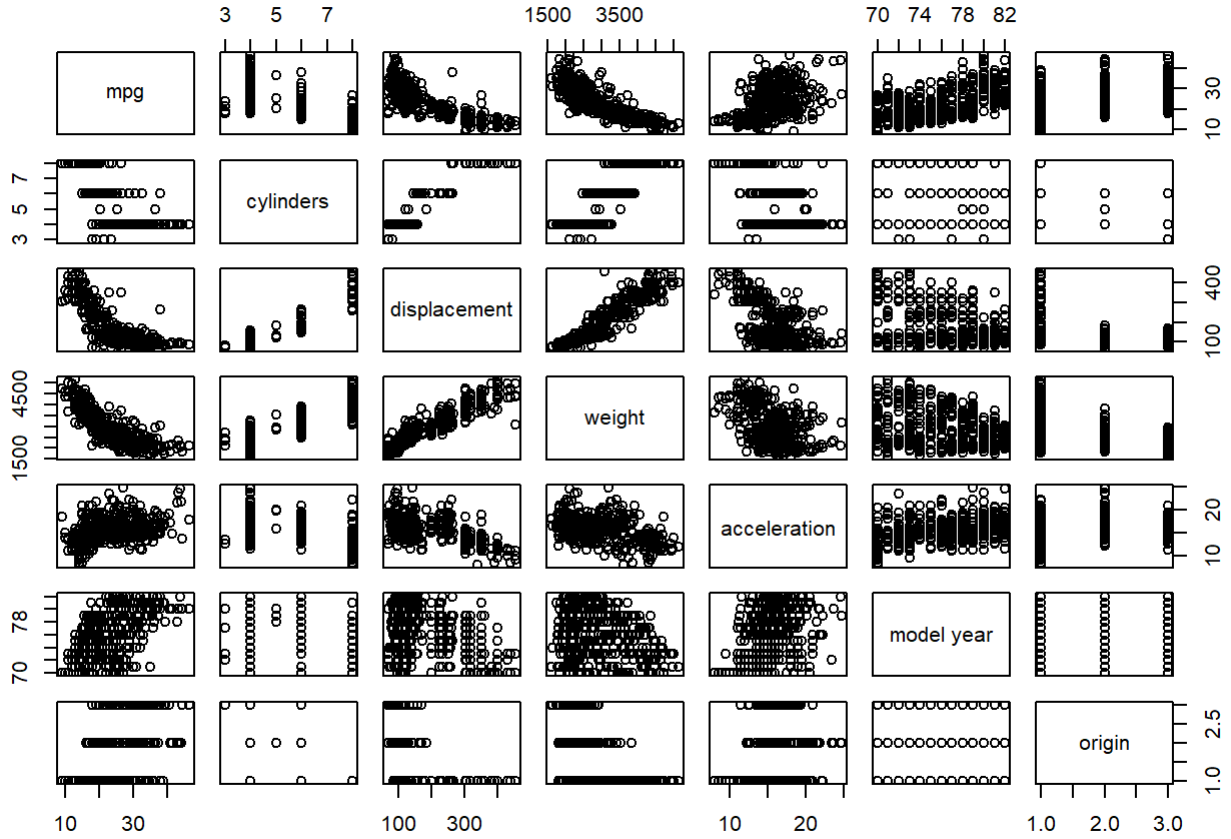
```
## # A tibble: 398 x 9
##      mpg cylinders displacement horsepower weight acceleration `model year`
##    <dbl>     <dbl>        <dbl> <chr>       <dbl>        <dbl>        <dbl>
##  1    18         8          307 130          3504         12             70
##  2    15         8          350 165          3693         11.5           70
##  3    18         8          318 150          3436         11             70
##  4    16         8          304 150          3433         12             70
##  5    17         8          302 140          3449         10.5           70
##  6    15         8          429 198          4341         10             70
##  7    14         8          454 220          4354          9             70
##  8    14         8          440 215          4312          8.5           70
##  9    14         8          455 225          4425         10             70
## 10    15         8          390 190          3850          8.5           70
## # ... with 388 more rows, and 2 more variables: origin <dbl>, car name <chr>
```

```
## # A tibble: 398 x 9
##      mpg cylinders displacement horsepower weight acceleration `model year`
##    <dbl>     <dbl>        <dbl> <chr>       <dbl>        <dbl>        <dbl>
##  1    18         8          307 130          3504         12             70
##  2    15         8          350 165          3693         11.5           70
##  3    18         8          318 150          3436         11             70
##  4    16         8          304 150          3433         12             70
##  5    17         8          302 140          3449         10.5           70
##  6    15         8          429 198          4341         10             70
##  7    14         8          454 220          4354          9             70
##  8    14         8          440 215          4312          8.5           70
##  9    14         8          455 225          4425         10             70
## 10    15         8          390 190          3850          8.5           70
## # ... with 388 more rows, and 2 more variables: origin <dbl>, car name <chr>
```

NOTE: Changed the value of “?” to NA for all values in the variable Horsepower in this dataset.

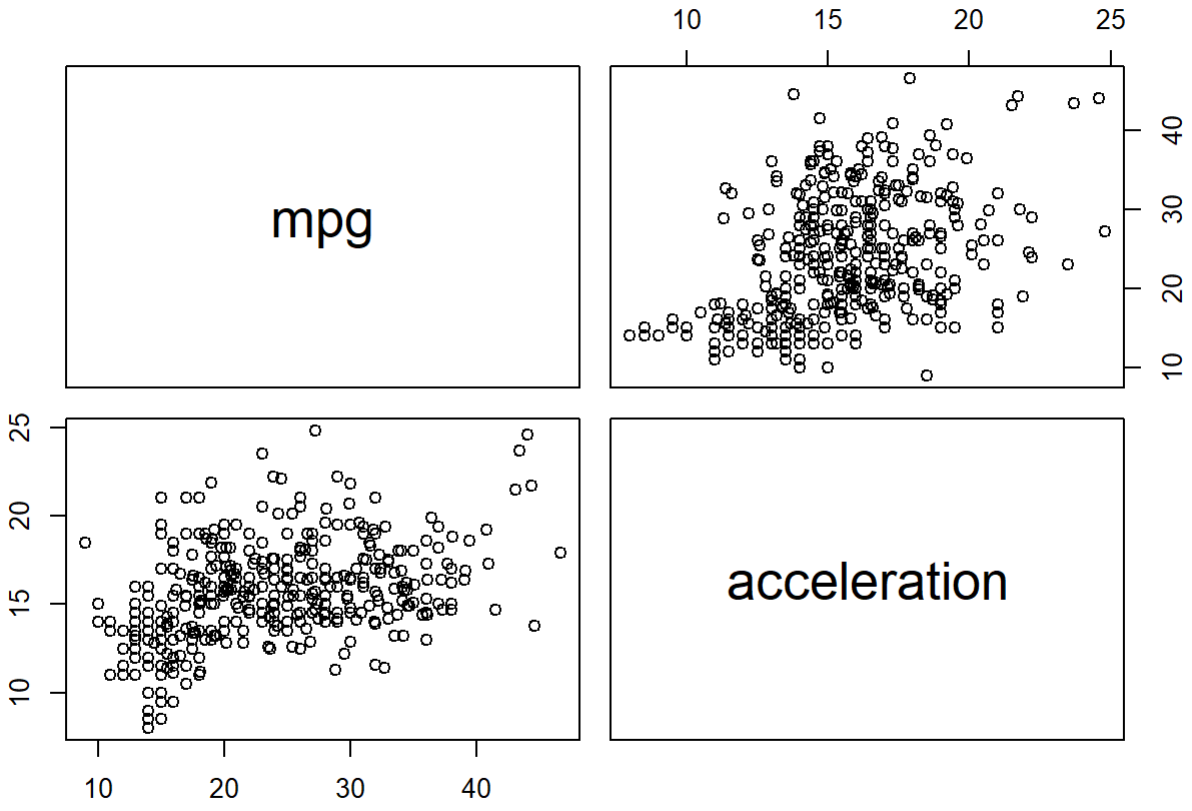
d) Do a summary analysis of the data (numeric variables) by checking each variable’s range, extreme values, mean, median, standard deviation, etc. Check the correlations among the variables and plot pairwise graph between each two variables by using command pairs.

```
##       mpg          cylinders      displacement       weight      acceleration
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   :1613   Min.   : 8.00
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2   1st Qu.:2224   1st Qu.:13.82
##  Median :23.00   Median :4.000   Median :148.5   Median :2804   Median :15.50
##  Mean   :23.51   Mean   :5.455   Mean   :193.4   Mean   :2970   Mean   :15.57
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:3608   3rd Qu.:17.18
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :5140   Max.   :24.80
##    model year        origin
##  Min.   :70.00   Min.   :1.000
##  1st Qu.:73.00   1st Qu.:1.000
##  Median :76.00   Median :1.000
##  Mean   :76.01   Mean   :1.573
##  3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :82.00   Max.   :3.000
```

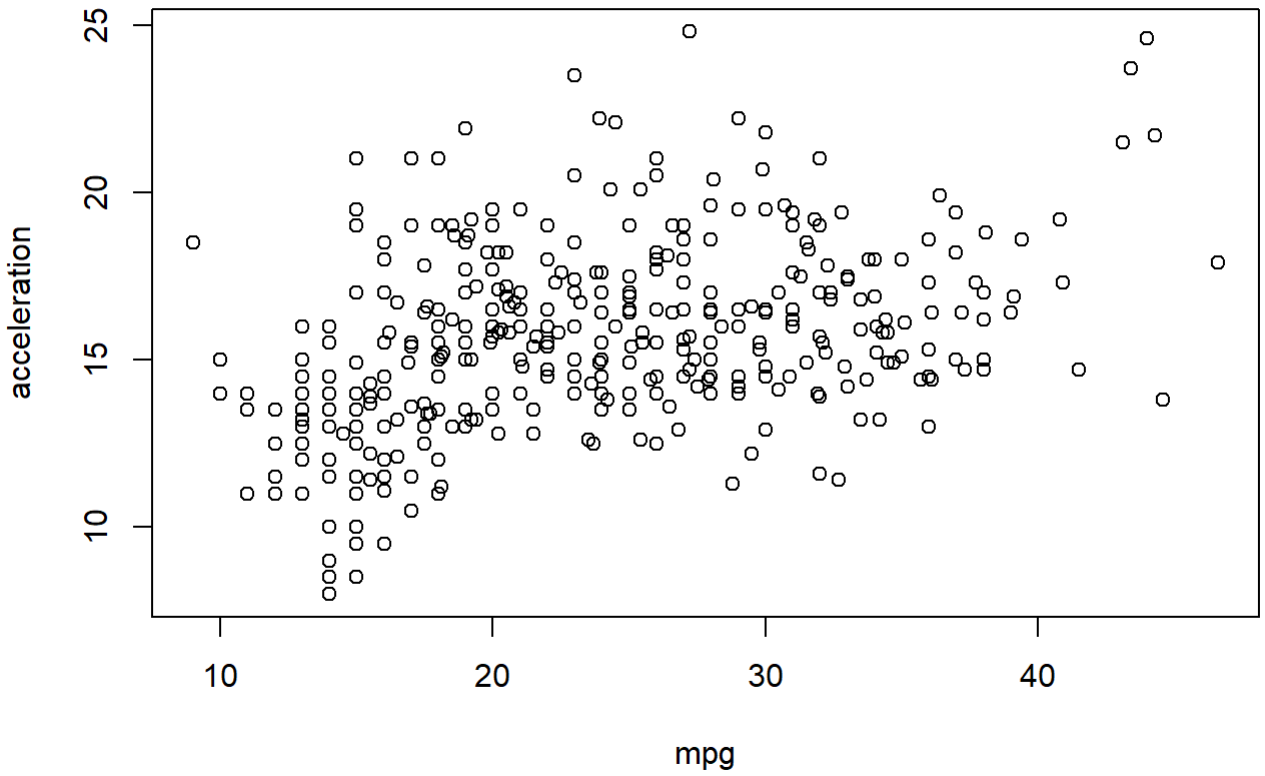


e) Create a two-variable data, with only acceleration and mpg in it. Make a scatter plot between them by using mpg as y-axis variable. Do you see strong correlation between the two variables? In addition, what is a correlation?

```
##       mpg         acceleration
##  Min.   : 9.00   Min.   : 8.00
##  1st Qu.:17.50   1st Qu.:13.82
##  Median :23.00   Median :15.50
##  Mean   :23.51   Mean   :15.57
##  3rd Qu.:29.00   3rd Qu.:17.18
##  Max.   :46.60   Max.   :24.80
```



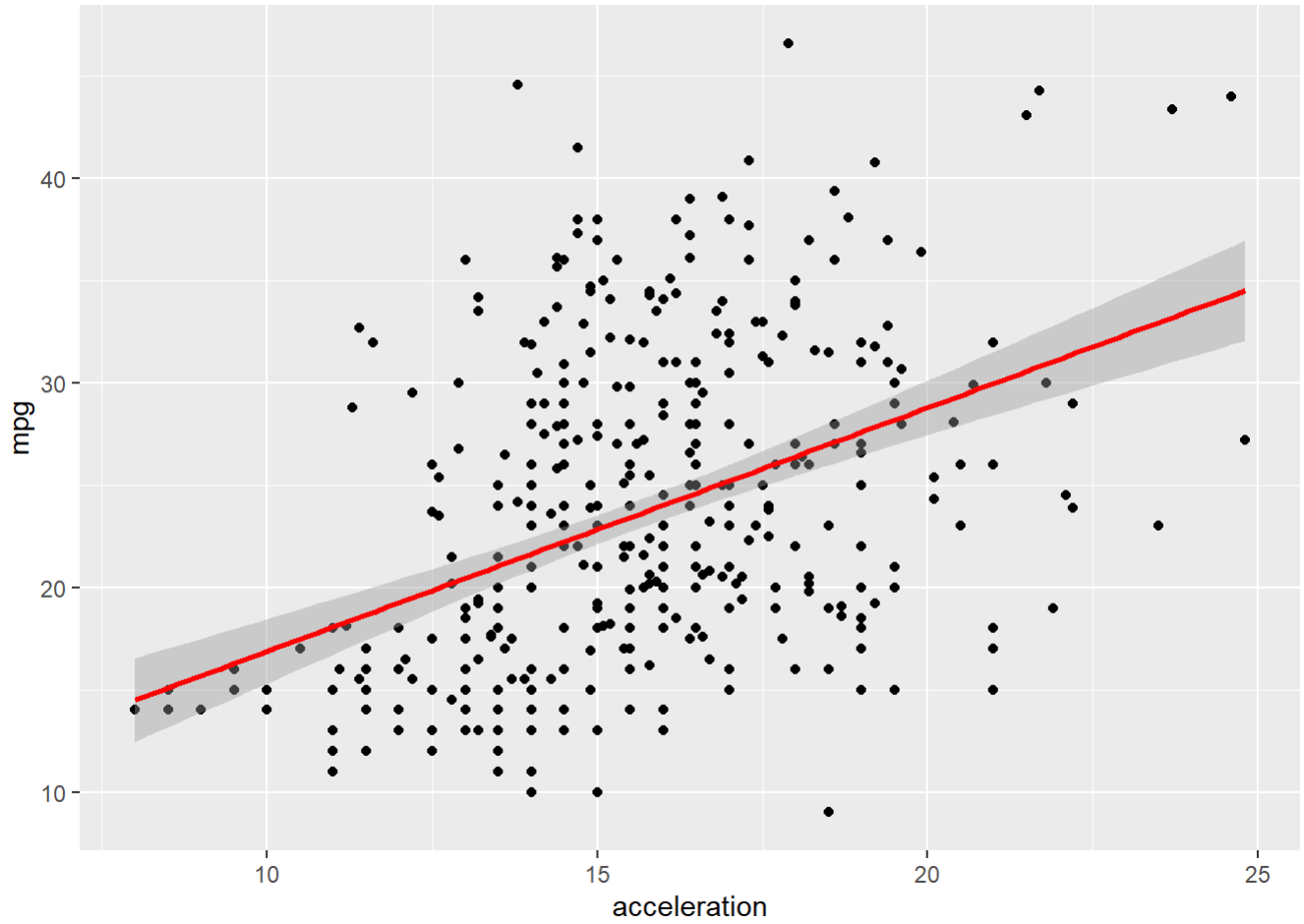

```
##                     mpg acceleration
## mpg          1.0000000    0.4202889
## acceleration 0.4202889    1.0000000
```



Correlation is positive BUT NOT really strong between “mpg” and “acceleration”.

f) Run a linear regression between the variables in part e) and use mpg as the response variable, acceleration as the predictor. What’s your conclusion for this analysis? In addition, add the regression line to the plot in part e).

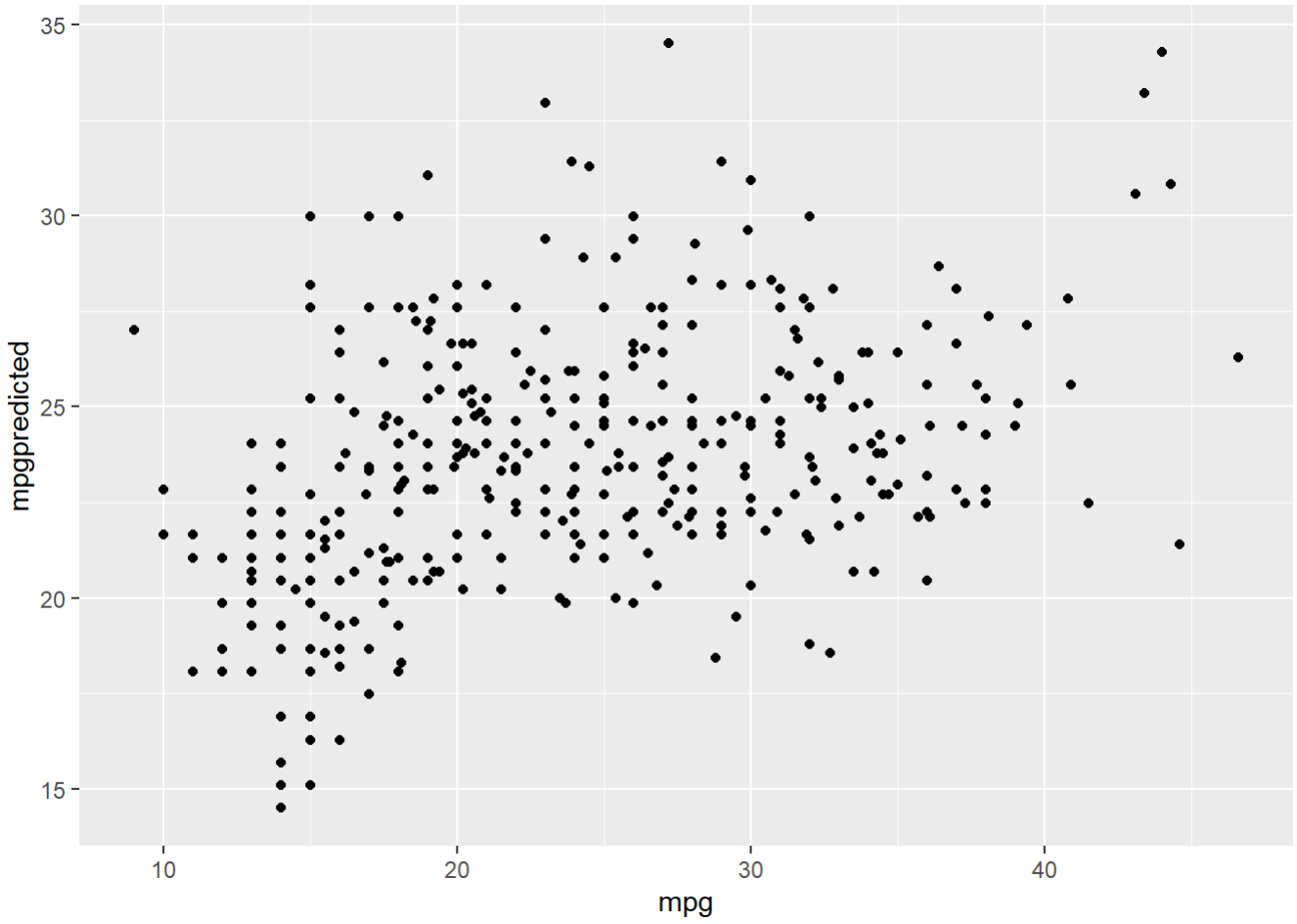
```
##
## Call:
## lm(formula = mpg ~ acceleration, data = autompgtwo)
##
## Coefficients:
##  (Intercept)  acceleration
##        4.970         1.191
```



g) Using the regression in part f), make a prediction of mpg for each acceleration values in the data.Draw a scatter plot between the original mpg and the predicted mpg. Comment.

Based on the data from f), the intercept is 4.970 and coefficient of acceleration is 1.191.
So, the equation of the line is $mpg = 4.970 + 1.191(acceleration)$

```
## # A tibble: 398 x 3
##      mpg acceleration mpgpredicted
##    <dbl>        <dbl>        <dbl>
##  1    18         12           19.3
##  2    15         11.5         18.7
##  3    18         11           18.1
##  4    16         12           19.3
##  5    17         10.5         17.5
##  6    15         10           16.9
##  7    14          9           15.7
##  8    14          8.5         15.1
##  9    14         10           16.9
## 10    15          8.5         15.1
## # ... with 388 more rows
```



The predicted mpg is close to the original mpg. The variation is not significant.

h) MSE is an abbreviation for Mean Squared Error, which is the average of the squared differences between the estimated and the truth value (or observed value). For the results in part g), treat the original mpg as true values, and predicted mpg as estimates. Find the MSE of this prediction.

```
## # A tibble: 398 x 3
##      mpg acceleration mpgpredicted
##    <dbl>        <dbl>        <dbl>
##  1    18         12           19.3
##  2    15         11.5         18.7
##  3    18         11           18.1
##  4    16         12           19.3
##  5    17         10.5         17.5
##  6    15         10           16.9
##  7    14          9           15.7
##  8    14          8.5         15.1
##  9    14         10           16.9
## 10    15          8.5         15.1
## # ... with 388 more rows
```

```
##
## Call:
## lm(formula = mpg ~ mpgpredicted, data = autompgMSE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.007  -5.636  -1.242   4.758  23.192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.00106    2.57597   0.000        1
## mpgpredicted  1.00017    0.10851   9.217   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.101 on 396 degrees of freedom
## Multiple R-squared:  0.1766, Adjusted R-squared:  0.1746
## F-statistic: 84.96 on 1 and 396 DF,  p-value: < 2.2e-16
```

```
## [1] 50.17219
```

The MSE of this prediction is 50.17219

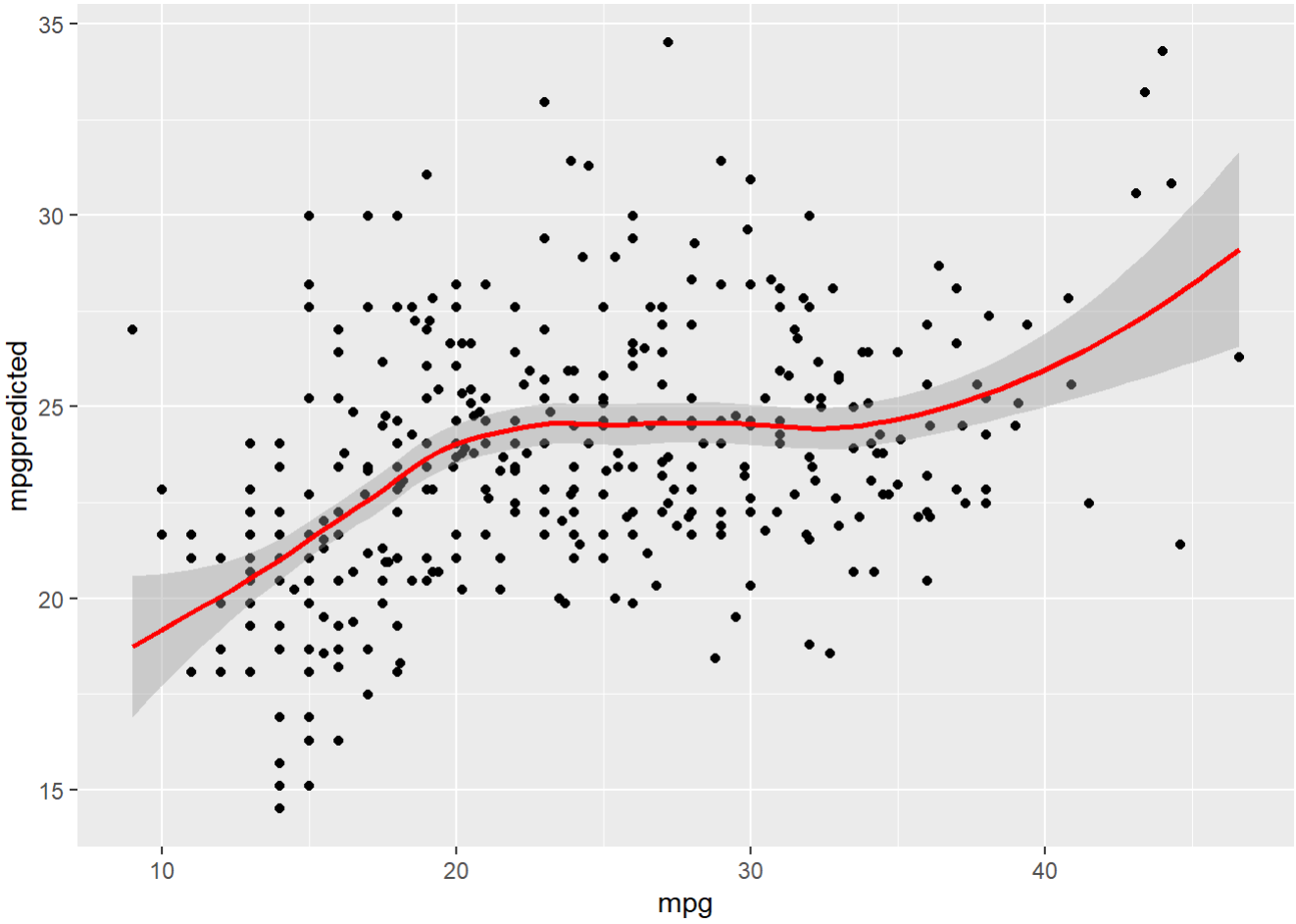
i) The Locally Estimated Scatterplot Smoothing, or LOESS, is a moving regression to fit data more smoothly. Use the loess function in R to make a LOESS regression between acceleration and mpg. What is the MSE of the prediction in this case? Comment, including the results in part h). Add the LOESS regression line into the graph you drew for part h)

```
## # A tibble: 398 x 3
##      mpg acceleration mpgpredicted
##    <dbl>        <dbl>        <dbl>
##  1    18         12           19.3
##  2    15         11.5         18.7
##  3    18         11           18.1
##  4    16         12           19.3
##  5    17         10.5         17.5
##  6    15         10           16.9
##  7    14          9           15.7
##  8    14          8.5         15.1
##  9    14         10           16.9
## 10    15          8.5         15.1
## # ... with 388 more rows
```

```
## Call:
## loess(formula = mpg ~ acceleration, data = autompgLOESS)
##
## Number of Observations: 398
## Equivalent Number of Parameters: 5.44
## Residual Standard Error: 6.961
```

```
## Call:
## loess(formula = mpg ~ acceleration, data = autompgLOESS)
##
## Number of Observations: 398
## Equivalent Number of Parameters: 5.44
## Residual Standard Error: 6.961
## Trace of smoother matrix: 5.97  (exact)
##
## Control settings:
##   span     :  0.75
##   degree   :  2
##   family   :  gaussian
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE
## drop.square:  FALSE
```

```
## [1] 47.67229
```



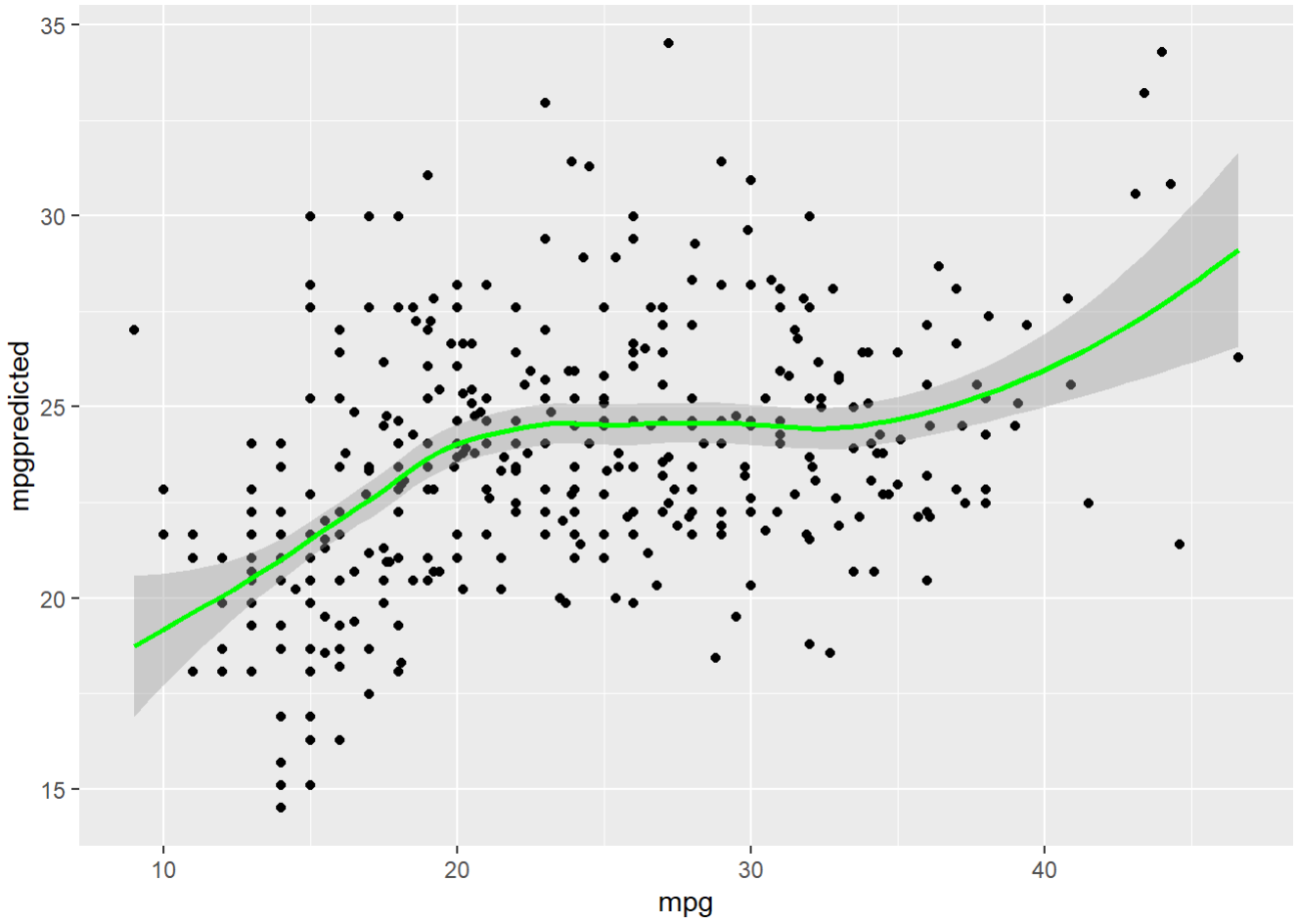
The MSE of this prediction in case of LOESS is 47.67229. The LOESS regression is local and it is a better fit.

j) Using summary to check the result of your LOESS regression in part i). The span, in the Control settings, is a smoothing parameter. Now try to run another (or more if you like) LOESS regression by adding span option in your loess command. Comment on the results.

```
## Call:
## loess(formula = mpg ~ acceleration, data = autompgLOESS, span = 10)
##
## Number of Observations: 398
## Equivalent Number of Parameters: 2.98
## Residual Standard Error: 7.041
```

```
## Call:
## loess(formula = mpg ~ acceleration, data = autompgLOESS, span = 10)
##
## Number of Observations: 398
## Equivalent Number of Parameters: 2.98
## Residual Standard Error: 7.041
## Trace of smoother matrix: 3.01  (exact)
##
## Control settings:
##   span     :  10
##   degree   :  2
##   family   :  gaussian
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE
## drop.square:  FALSE
```

```
## [1] 49.19256
```



LOESS regression is a non-parametric method of regression where least squares regression is performed in localized subsets. The Span option in LOESS controls the amount of smoothing for the default loess smoother. Smaller Span numbers produce wigglier lines, larger numbers produce smoother lines.