Practice Problems

Question 1

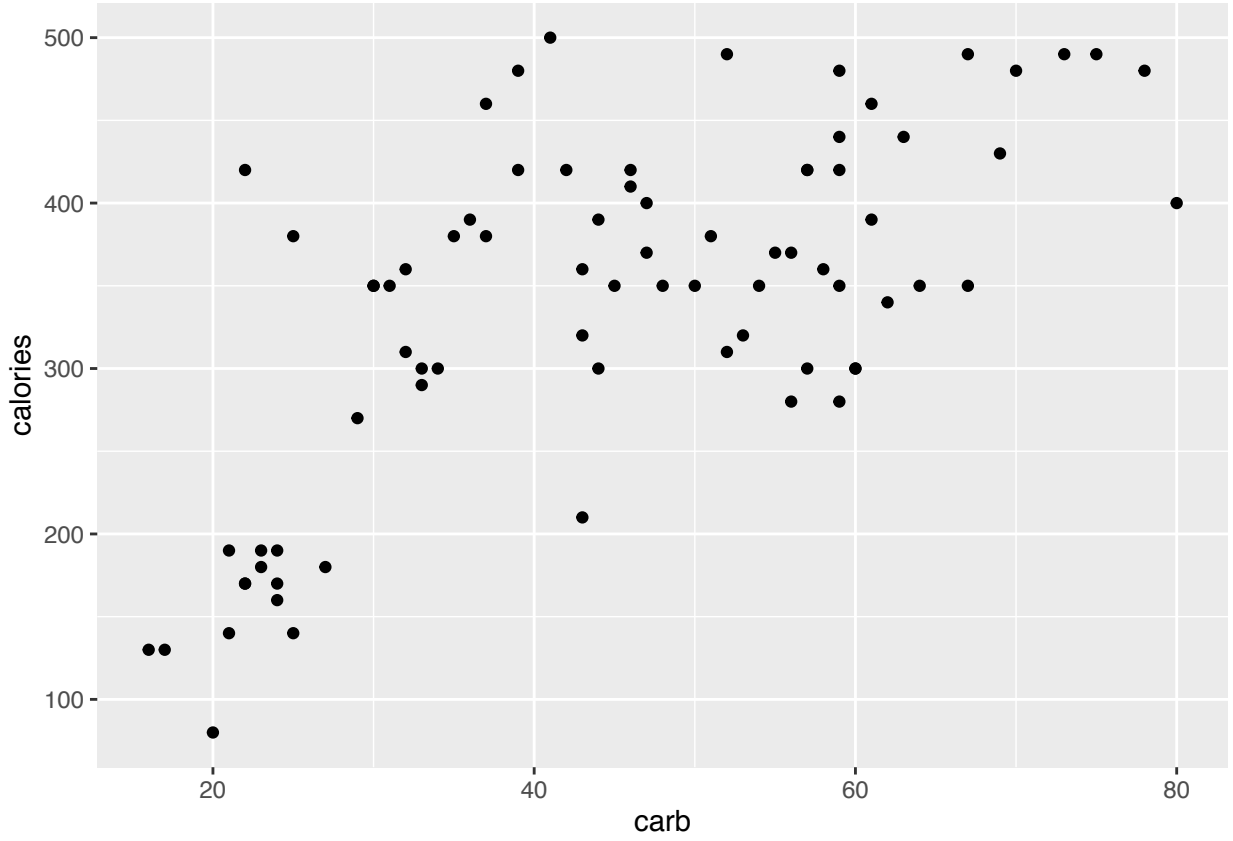
{Adapted from Exercise 7.18 in Dietz, Barr, Cetinkaya-Rundel, "OpenIntro Statistics", Second Edition} The `starbucks.csv` dataset contains data on calories and carbohydrates (in grams) in Starbucks food menu items.

```
starbucksdata<-read_csv("starbucks.csv")
glimpse(starbucksdata)
```

```
## Observations: 77
## Variables: 7
## $ item     <chr> "8-Grain Roll", "Apple Bran Muffin", "Apple Fritter", "Ban...
## $ calories <dbl> 350, 350, 420, 490, 130, 370, 460, 370, 310, 420, 380, 320...
## $ fat      <dbl> 8, 9, 20, 19, 6, 14, 22, 14, 18, 25, 17, 12, 17, 21, 5, 18...
## $ carb     <dbl> 67, 64, 59, 75, 17, 47, 61, 55, 32, 39, 51, 53, 34, 57, 52...
## $ fiber    <dbl> 5, 7, 0, 4, 0, 5, 2, 0, 0, 0, 2, 3, 2, 2, 3, 3, 2, 3, 0, 2...
## $ protein  <dbl> 10, 6, 5, 7, 0, 6, 7, 6, 5, 7, 4, 6, 5, 5, 12, 7, 8, 6, 0,...
## $ type     <chr> "bakery", "bakery", "bakery", "bakery", "bakery", "bakery"...
```

(a) Produce a plot that shows the association between carbohydrates and calories in Starbucks menu items. Describe this association.

```
starbucksdata %>%
  ggplot(aes(x=carb, y=calories)) + geom_point()
```



The association between carbohydrates and calories is a positive, moderate, and linear relationship.

(b) Before calculating anything, estimate the correlation coefficient between carbohydrates and calorie content in Starbucks menu items based on the plot you produced in (a). Justify your answer.

The correlation coefficient should be about 0.5. The correlation should be positive since there is a positive relationship between carbohydrates and calories. The absolute value for the correlation coefficient should be close to 0.5 since there is moderate relationship between carbohydrates and calories.

(c) Calculate the correlation between carbohydrate and calorie content of Starbucks menu items. How does this compare to your estimate in part (b)?

```
cor(starbucksdata$carb, starbucksdata$calories)
```

```
## [1] 0.674999
```

The correlation is slightly higher than what I have estimated, which means that the relationship between carbohydrates and calories is fairly strong.

(d) Write down a simple linear regression model to predict calories based on carbohydrate content of Starbucks menu items. Be sure to explain each term in the model.

yihat= betaohat+ beta1hatxi yihat= 120 + 6xi

yihat:estimated calories for the ith observation betaohat = 120: is the intercept parameter (the calories count corresponding to 0 carbs) beta1hat = (480-160)/80-20 = 6: is the slope parameter (the slope of the best fit line for the data points) xi:carbohydrates for the ith observation

(e) Use R to fit the regression model in (d) to these data. Report the fitted regression line and interpret the regression coefficients in the context of this study.

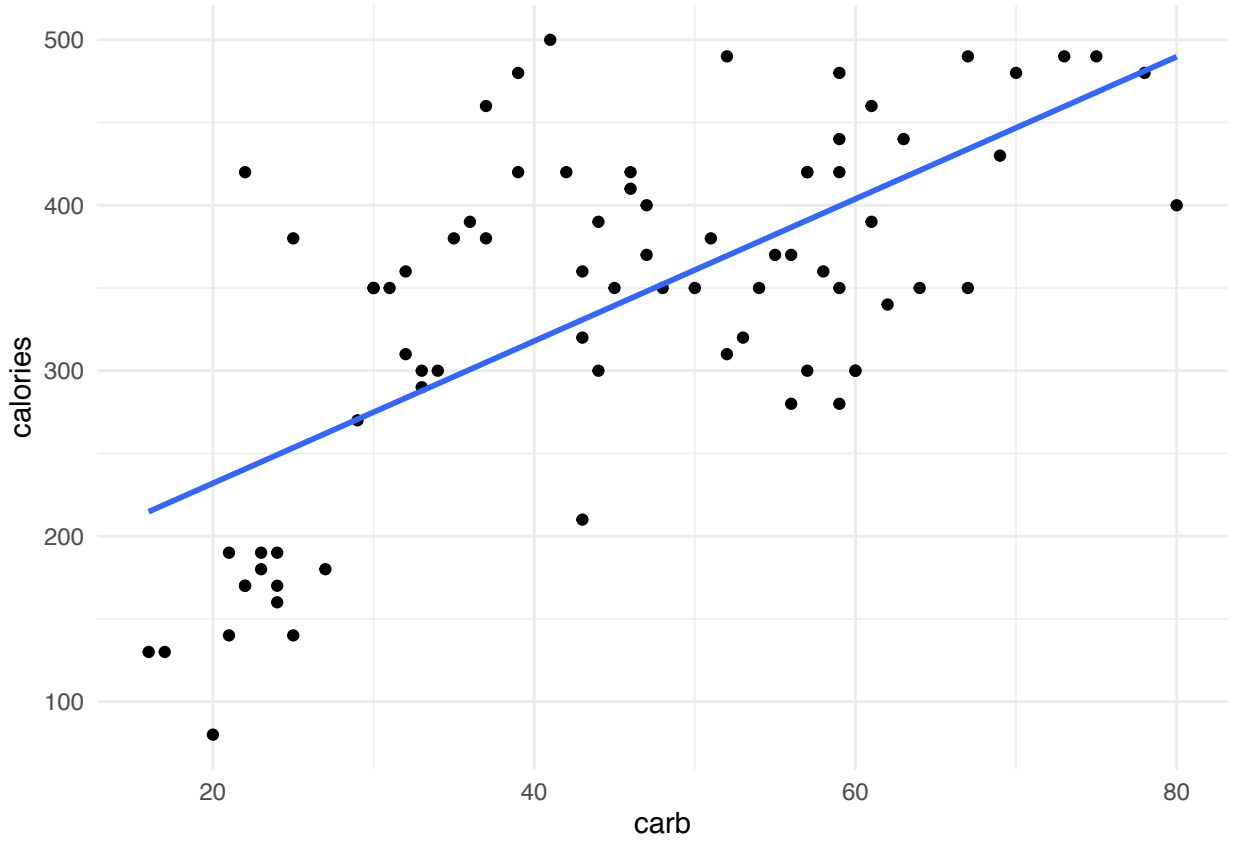
```
model <- lm(calories ~ carb , data = starbucksdata)
summary(model)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 146.020432 25.9186185 5.633805 2.927070e-07
## carb          4.297084  0.5423626 7.922900 1.672545e-11
```

yihat = 146.02 + 4.30xi The intercept parameter is about 146.02, which in this study, is the number of calories corresponding to 0 carbs. The slope parameter is about 4.30, which in this study, is the increase in calories with one unit increase of carbs.

(f) Add the estimated linear regression line that you calculated in (d) to the plot you generated in (a). Compute the coefficient of determination, R^2. How well does the linear regression line seem to capture the relationship between `carb` and `calories`? Justify your answer.

```
starbucksdata %>% ggplot(aes(x=carb, y=calories)) + geom_point() +
  geom_smooth(method="lm", se=FALSE) + theme_minimal()
```




```
summary(model)$r.squared
```

```
## [1] 0.4556237
```

Since the coefficient of determination (about 0.456) is close to 0, some of the variation in the calories are not explained by the carbs. Hence, the regression line does not seem to capture the relationship that well between carbs and calories.

Question 2

The dataset `hybrid_reg.csv` contains data on the fuel efficiencies (`mpg`), acceleration rate (`accelrate`) as well as some other variables for a sample of 153 hybrid cars. These data were used in the paper: D-J. Lim, S.R. Jahromi, T.R. Anderson, A-A. Tudorie (2014). "Comparing Technological Advancement of Hybrid Electric Vehicles (HEV) in Different Market Segments," Technological Forecasting & Social Change, http://dx.doi.org/10.1016/j.techfore.2014.05.008

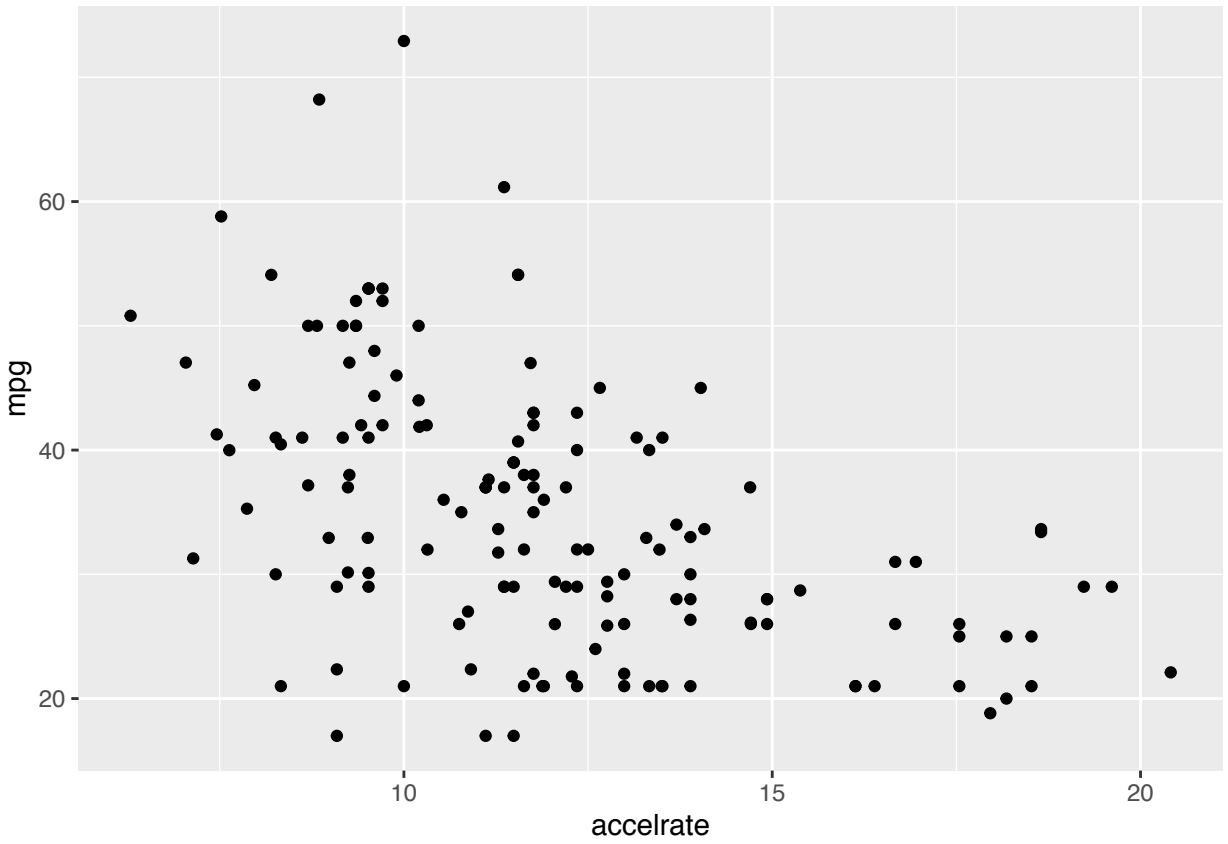
In this question you will build a linear regression model to predict the fuel efficiency (in mpg, miles per gallon) for a hybrid vehicle based on acceleration rate (i.e., the time (in seconds) it takes for a vehicle to go from 0 to 60 miles/hour). Note that higher values of `mpg` mean the vehicle is more fuel efficient.

```
cardat<-read_csv("hybrid_reg.csv")
glimpse(cardat)
```

```
## Observations: 153
## Variables: 9
## $ carid       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ vehicle     <chr> "Prius (1st Gen)", "Tino", "Prius (2nd Gen)", "Insight"...
## $ year        <dbl> 1997, 2000, 2000, 2000, 2001, 2001, 2002, 2003, 2003, 2...
## $ msrp        <dbl> 24509.74, 35354.97, 26832.25, 18936.41, 25833.38, 19036...
## $ accelrate   <dbl> 7.46, 8.20, 7.97, 9.52, 7.04, 9.52, 9.71, 8.33, 9.52, 8...
## $ mpg         <dbl> 41.26, 54.10, 45.23, 53.00, 47.04, 53.00, 53.00, 40.46,...
## $ mpgmpge     <dbl> 41.26, 54.10, 45.23, 53.00, 47.04, 53.00, 53.00, 40.46,...
## $ carclass    <chr> "C", "C", "C", "TS", "C", "TS", "TS", "MV", "TS", "C", ...
## $ carclass_id <dbl> 1, 1, 1, 7, 1, 7, 7, 4, 7, 1, 6, 7, 3, 5, 6, 1, 6, 7, 1...
```

(a) What is the association between acceleration rate and fuel efficiency (in mpg)? Produce an appropriate plot and describe the relationship.

```
cardat %>%
  ggplot(aes(x=accelrate, y=mpg)) + geom_point()
```



The association between acceleration rate and fuel efficiency is negative, moderate, and linear.

(b) Before calculating anything, estimate what you think the correlation coefficient between `accelrate` and `mpg` is based on the plot you produced in (a). Justify your answer.

The correlation coefficient should be about -0.5. The correlation should be negative since there is a negative relationship between acceleration rate and fuel efficiency. The absolute value for the correlation coefficient should be close to 0.5 since there is moderate relationship between acceleration rate and fuel efficiency.

(c) Calculate the correlation between acceleration rate and fuel efficiency of hybrid vehicles. How does this compare to your estimate in part (b)?

```
cor(cardat$accelrate, cardat$mpg)
```

```
## [1] -0.5060704
```

This is fairly close to what I estimated as the correlation coefficient in b).

(d) Write out the mathematical description of a simple linear regression model of `mpg` on `accelrate`. Describe each of the variables and parameters in the model.

yihat= betaohat+ beta1hatxi yihat= 65 - 2.25xi

yihat:estimated fuel efficiency(mpg) for the ith observation betaohat = 65: is the intercept parameter (the mpg value corresponding to 0 acceleration rate) beta1hat = (20-65)/ (20-0) = -2.25: is the slope parameter (the slope of the best fit line for the data points) xi:acceleration rate for the ith observation

(e) Use 75% of the data to select a training set, and leave 25% of the data for testing. Use the last 3 digits of your student number in `set.seed`.

```
set.seed(   )
n <- nrow(cardat)
training_indices <- sample(1:n, size=round(0.75*n))
train <- cardat[training_indices,]
test <- cardat[-training_indices,]
```

(i) Fit a linear regression model of `mpg` on `accelrate` based on the training set.

```
model2 <- lm(mpg ~ accelrate, data=train)
summary(model2)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 56.625676  3.5555938 15.925800 1.182396e-30
## accelrate   -1.754428  0.2866413 -6.120638 1.380068e-08
```

yihat = 56.63 -1.75xi

yihat:estimated fuel efficiency(mpg) for the ith observation xi:acceleration rate for the ith observation

(ii) Calculate the coefficient of determination based on the training set. Does this model describe the variation in fuel efficiencies well?

```
summary(model2)$r.squared
```

```
## [1] 0.2489809
```

Since the coefficient of determination (about 0.249) is close to 0, much of the variation in the mpg value is not explained by the accelrate value. Hence, the regression line does not seem to capture the relationship that well between mpg value and the accelrate value.

(iii) How well does this perform as a predictive model? Calculate RMSE on the training and test data using the linear regression model you fit on the training set.

```
yhat_test <- predict(model2, newdata = test)
y_test <- test$mpg;
n_test <- nrow(test)
# RMSE for predictions in testing dataset
sqrt(sum((y_test - yhat_test)^2) / n_test)
```

```
## [1] 9.708365
```

```r
yhat_train <- predict(model2, newdata = train)
y_train <- train$mpg;
n_train <- nrow(train)
# RMSE for predictions in training dataset
sqrt(sum((y_train - yhat_train)^2) / n_train)
```

```
## [1] 9.401881
```

Since there is a relatively small difference (3%) between the RMSE for the training data and that for the testing data. This suggests that our predictive model may not be overfitting the training data and is fairly useful for to make predictions for new observations.

(f)

(i) Use the model you built based on the 75% training test to predict the fuel efficiency in mpg for a hybrid vehicle that can accelerate from 0 to 60 m/h in 10 seconds. Show your calculations.

The model is: yihat = 56.63 -1.75xi. Plug in 10 for xi to get: yihat = 56.63 - 1.75*10 = 39.13 (miles per gallon). This is the fuel efficiency in mpg for a hybrid vehicle that can accelerate from 0 to 60 m/h in 10 seconds.

(ii) From the plot produced earlier in this question we can see that there was one vehicle with an acceleration time of 10 seconds in the sample. The actual fuel efficiency of this vehicle was 21 mpg. What is the residual? Show your calculations.

Residual ei = yi - yihat = 21 - 39.13 = -18.13

Question 3

The internet movie database, http://imdb.com/, is a website devoted to collecting movie data supplied by studios and fans. It claims to be the biggest movie database on the web and is run by amazon. More about information imdb.com can be found online, http://imdb.com/help/show_leaf?about, including information about the data collection process, http://imdb.com/help/show_leaf?infosource. The data we will use for this question is packaged in the R library `ggplot2movies` in the data frame `movies`. The help documentation for this data (`?ggplot2movies::movies`) describes each of the variables.

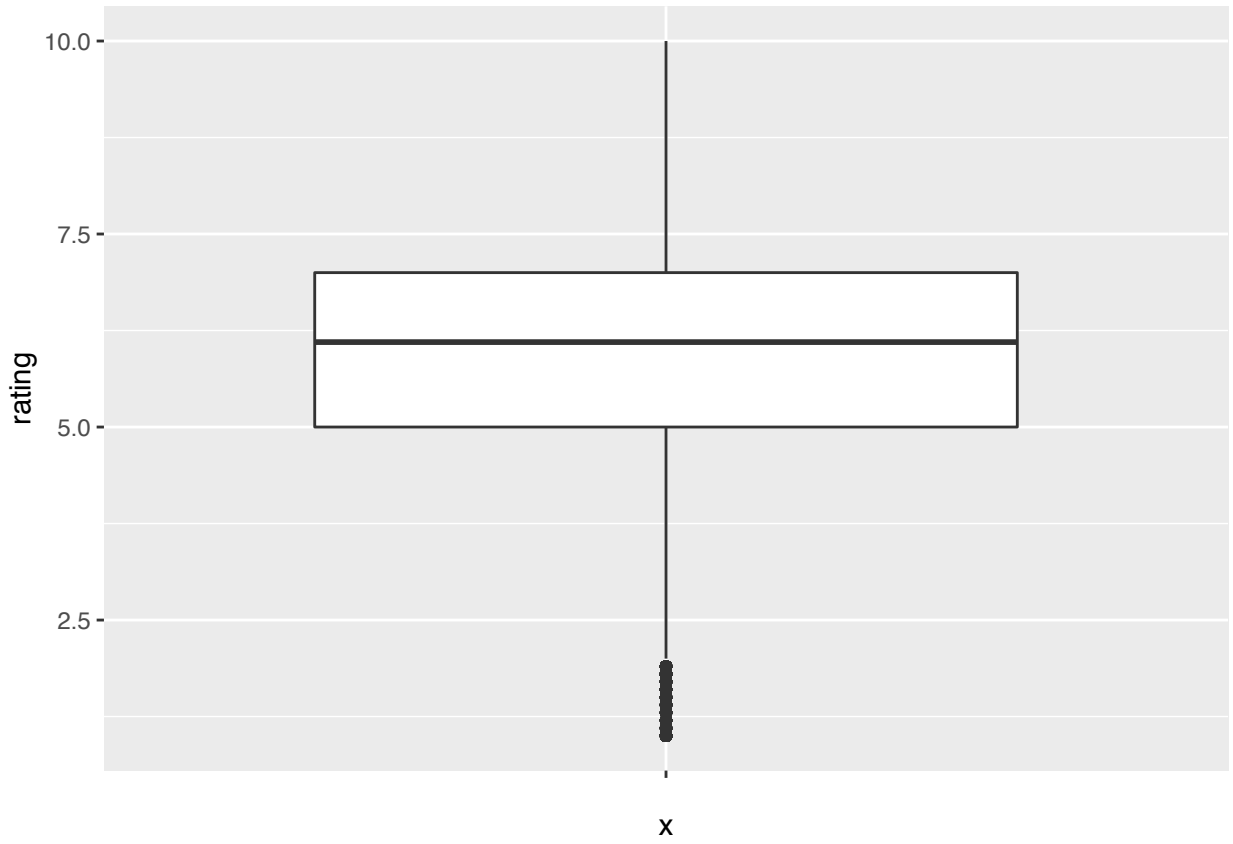
In this question you will use linear regression to build a model that predicts IMDB user ratings based on covariates in the `movies` data. More information about IMDB user ratings is available here.

```
library(ggplot2movies)
library(tidyverse)
glimpse(movies)
```

```
## Observations: 58,788
## Variables: 24
## $ title       <chr> "$", "$1000 a Touchdown", "$21 a Day Once a Month", "$4...
## $ year        <int> 1971, 1939, 1941, 1996, 1975, 2000, 2002, 2002, 1987, 1...
## $ length      <int> 121, 71, 7, 70, 71, 91, 93, 25, 97, 61, 99, 96, 10, 10,...
## $ budget      <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ rating      <dbl> 6.4, 6.0, 8.2, 8.2, 3.4, 4.3, 5.3, 6.7, 6.6, 6.0, 5.4, ...
## $ votes       <int> 348, 20, 5, 6, 17, 45, 200, 24, 18, 51, 23, 53, 44, 11,...
## $ r1          <dbl> 4.5, 0.0, 0.0, 14.5, 24.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5...
## $ r2          <dbl> 4.5, 14.5, 0.0, 0.0, 4.5, 4.5, 0.0, 4.5, 4.5, 0.0, 0.0,...
## $ r3          <dbl> 4.5, 4.5, 0.0, 0.0, 0.0, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, ...
## $ r4          <dbl> 4.5, 24.5, 0.0, 0.0, 14.5, 14.5, 4.5, 4.5, 0.0, 4.5, 14...
## $ r5          <dbl> 14.5, 14.5, 0.0, 0.0, 14.5, 14.5, 24.5, 4.5, 0.0, 4.5, ...
## $ r6          <dbl> 24.5, 14.5, 24.5, 0.0, 4.5, 14.5, 24.5, 14.5, 0.0, 44.5...
## $ r7          <dbl> 24.5, 14.5, 0.0, 0.0, 0.0, 4.5, 14.5, 14.5, 34.5, 14.5,...
## $ r8          <dbl> 14.5, 4.5, 44.5, 0.0, 0.0, 4.5, 4.5, 14.5, 14.5, 4.5, 4...
## $ r9          <dbl> 4.5, 4.5, 24.5, 34.5, 0.0, 14.5, 4.5, 4.5, 4.5, 4.5, 14...
## $ r10         <dbl> 4.5, 14.5, 24.5, 45.5, 24.5, 14.5, 14.5, 14.5, 24.5, 4....
## $ mpaa        <chr> "", "", "", "", "", "", "R", "", "", "", "", "", "", ""...
## $ Action      <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1...
## $ Animation   <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Comedy      <int> 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1...
## $ Drama       <int> 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0...
## $ Documentary <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ Romance     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Short       <int> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0...
```

(a) Create at least one appropriate graphical summary and relevant numercial summaries of the distribution of `rating`. Briefly explain why you chose these summaries and interpret the features of the distribution of IMDB ratings.

```
movies %>%
  ggplot(aes(x= " ", y=rating)) + geom_boxplot()
```



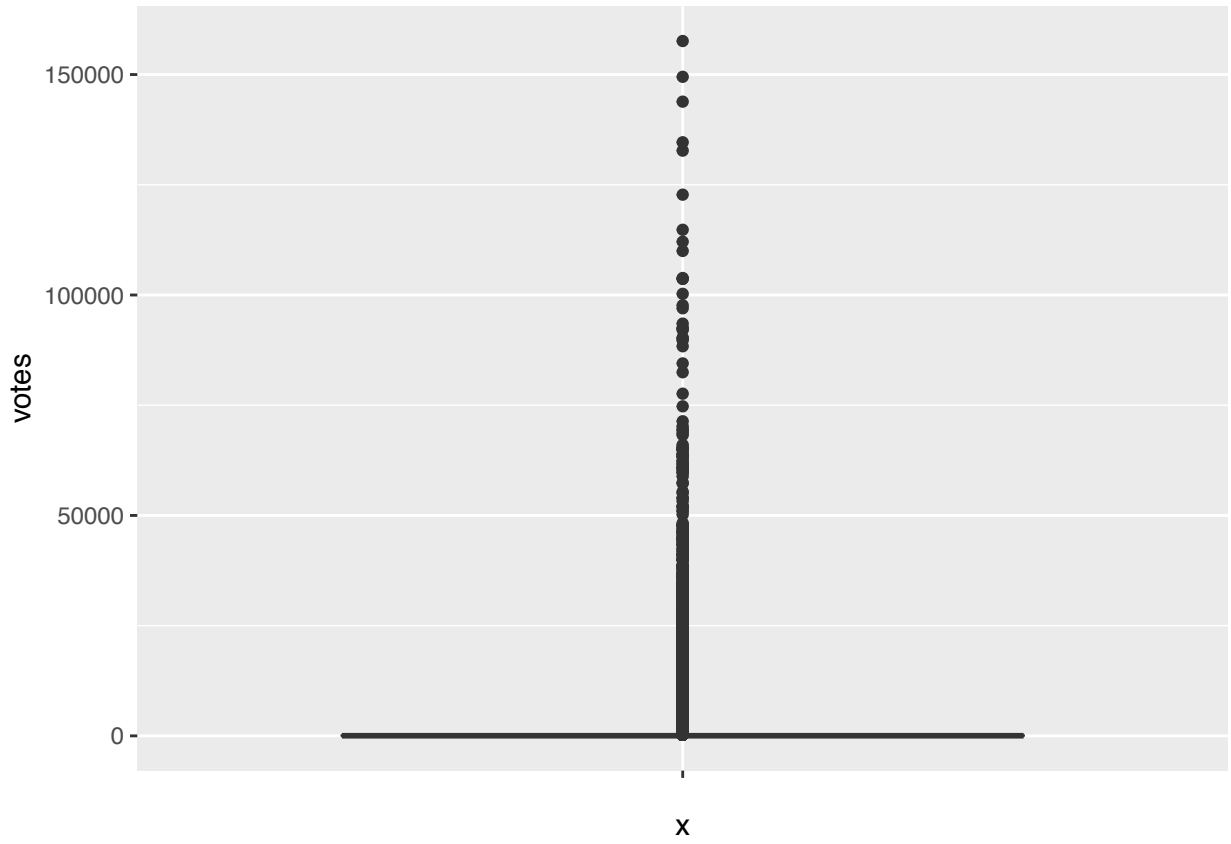

```
movies%>%
  summarise(n=n(),
            min = min(rating),
            max = max(rating),
            mean = mean(rating),
            median = median(rating))
```

```
## # A tibble: 1 x 5
##       n   min   max  mean median
##   <int> <dbl> <dbl> <dbl>  <dbl>
## 1 58788     1    10  5.93    6.1
```

The boxplot shows the distribution of the IMDB ratings and the summary table calculates the precise statistics of the ratings. We can see that the distribution is slightly left skewed, ranging from 1 to 10, with a mean of 5.93 and median of 6.1 for 58788 of the ratings.

(b) Create similar summaries that you produced in (a), but for the variable `votes` and interpret the features of its distribution.

```
movies %>%
  ggplot(aes(x= " ", y=votes)) + geom_boxplot()
```



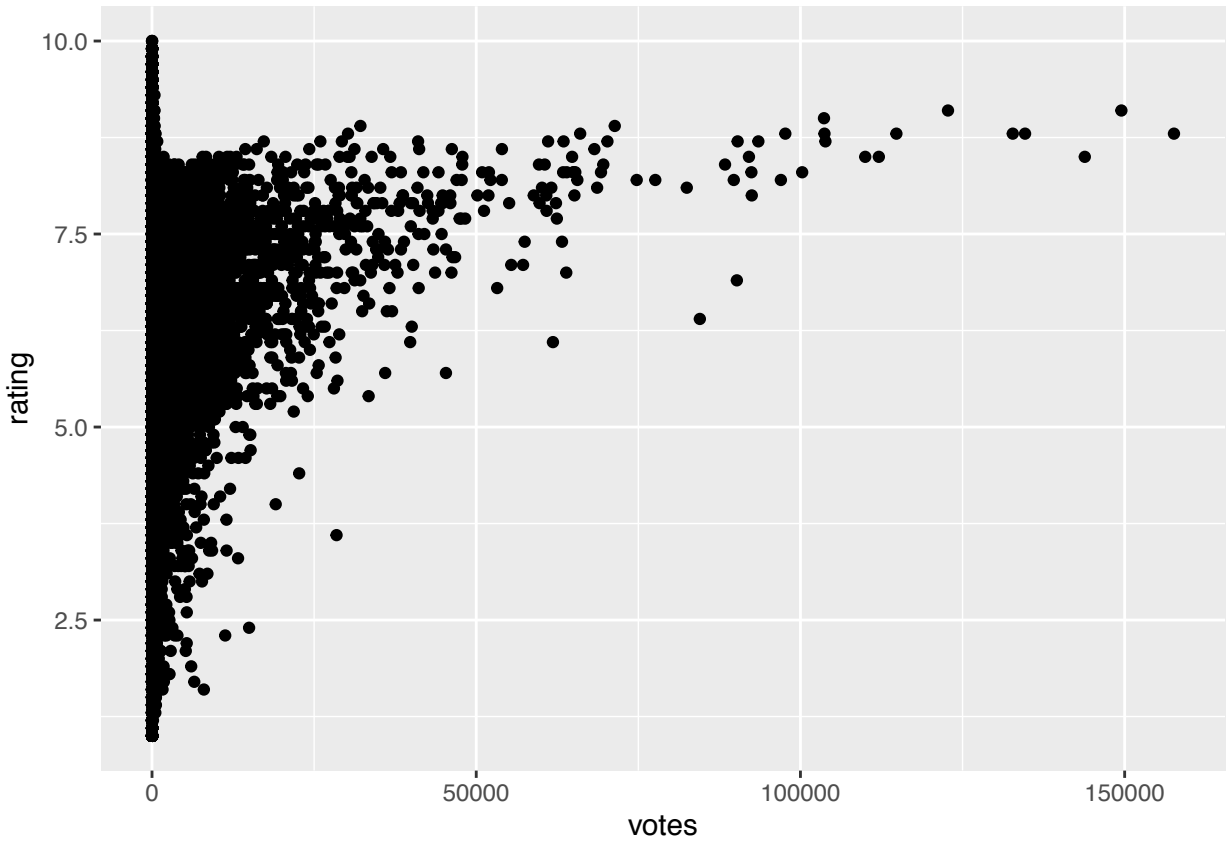

```
movies%>%
  summarise(n=n(),
            min = min(votes),
            max = max(votes),
            mean = mean(votes),
            median = median(votes))
```

```
## # A tibble: 1 x 5
##       n   min    max  mean median
##   <int> <int>  <int> <dbl>  <dbl>
## 1 58788     5 157608  632.     30
```

The boxplot shows the distribution of the votes and the summary table calculates the precise statistics of the votes. We can see that the distribution is very right skewed, with some outliers making the distribution range from 5 to 157608 , with a mean of 632.13 and median of 30 for 58788 of the ratings.

(c) Produce a plot that shows the association between `votes` and `rating` of release. Describe this association in 2-3 sentences.

```
movies %>%
  ggplot(aes(x=votes, y=rating)) + geom_point()
```



The association between acceleration votes and rating is positive, weak, and linear.

(d) Calculate the correlation between number of votes and and IMDB rating. Does it make sense to interpret the correlation in this situation? Why or why not?

```
cor(movies$votes, movies$rating)
```

```
## [1] 0.1037069
```

Having a correlation of nearly 0 means that there is little to no strength in the linear relationship between vote and rating. This correlation make sense and suggests that the linear relationship between the two variables is weak, so we probably shouldn't use a linear model to represent the relationship.

Question 4

The `Loblolly` data frame contains data on the growth of Loblolly pine trees from 14 different seed sources. In particular, it includes the ages (in years) and heights (in feet) of a sample of 14 Loblolly pines measured at 6 ages.

```
library(datasets)
glimpse(Loblolly)
```

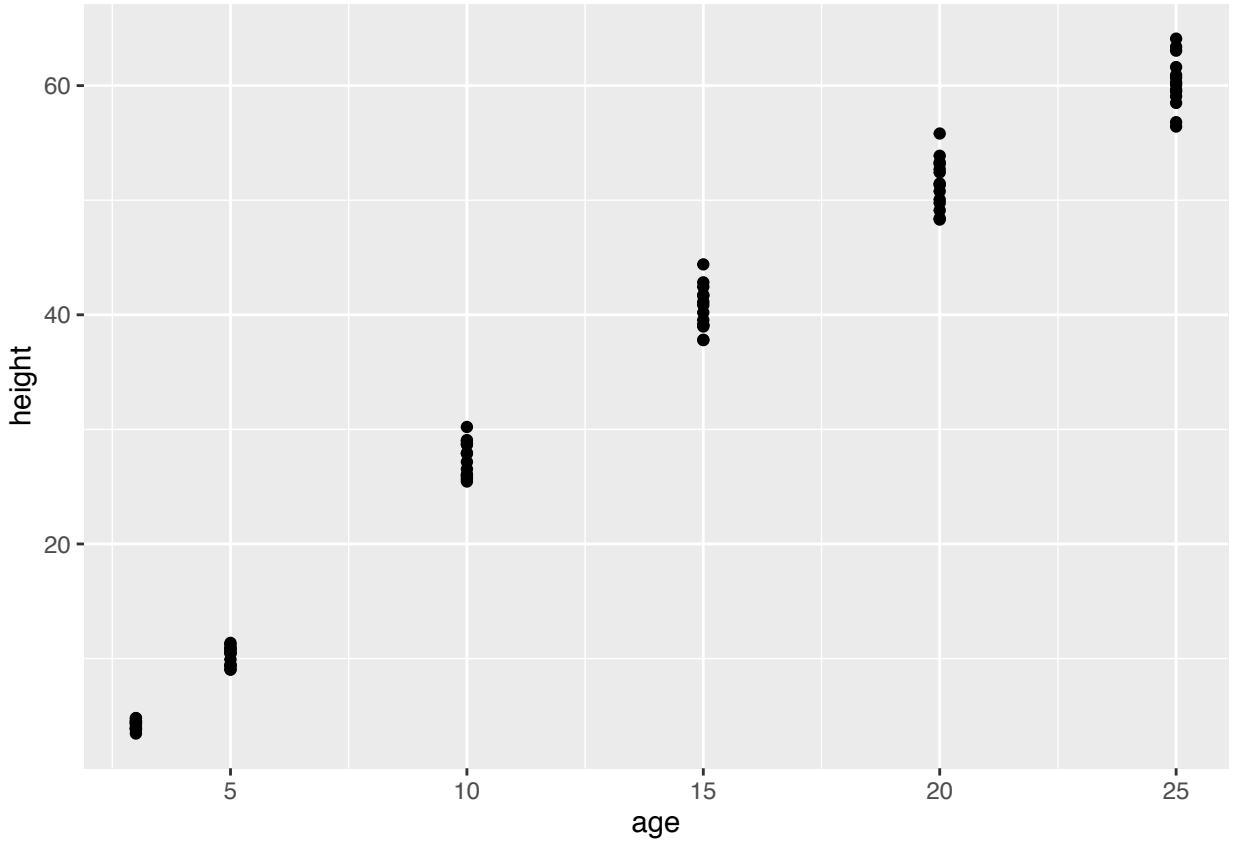
```
## Observations: 84
## Variables: 3
## $ height <dbl> 4.51, 10.89, 28.72, 41.74, 52.70, 60.92, 4.55, 10.92, 29.07,...
## $ age    <dbl> 3, 5, 10, 15, 20, 25, 3, 5, 10, 15, 20, 25, 3, 5, 10, 15, 20...
## $ Seed   <ord> 301, 301, 301, 301, 301, 301, 303, 303, 303, 303, 303, 303, ...
```

```
table(Loblolly$Seed)
```

```
##
## 329 327 325 307 331 311 315 321 319 301 323 309 303 305
##   6   6   6   6   6   6   6   6   6   6   6   6   6   6
```

(a) Create a data summary that shows the association between the age and height of Loblolly pine trees. Describe this association.

```
Loblolly %>%
  ggplot(aes(x=age, y=height)) + geom_point()
```



(b) Before calculating anything, estimate what you think the correlation coefficient between Loblolly pine age and height is based on the plot you produced in (a). Justify your answer.

I think the correlation shold be fairly close to 1 as there seems to be a fairly strong and positive relationship between age and height.

(c) Calculate the correlation between head age and height of Loblolly trees. How does this compare to your estimate in part (b)?

```
cor(Loblolly$age, Loblolly$height)
```

```
## [1] 0.9899132
```

The correlation is 0.99 which is similar to what I have estimated in b).

Question 5

The Housing data for 506 census tracts of Boston from the 1970 census. The dataframe `BostonHousing2` contains the original corrected data by Harrison and Rubinfeld (1979).

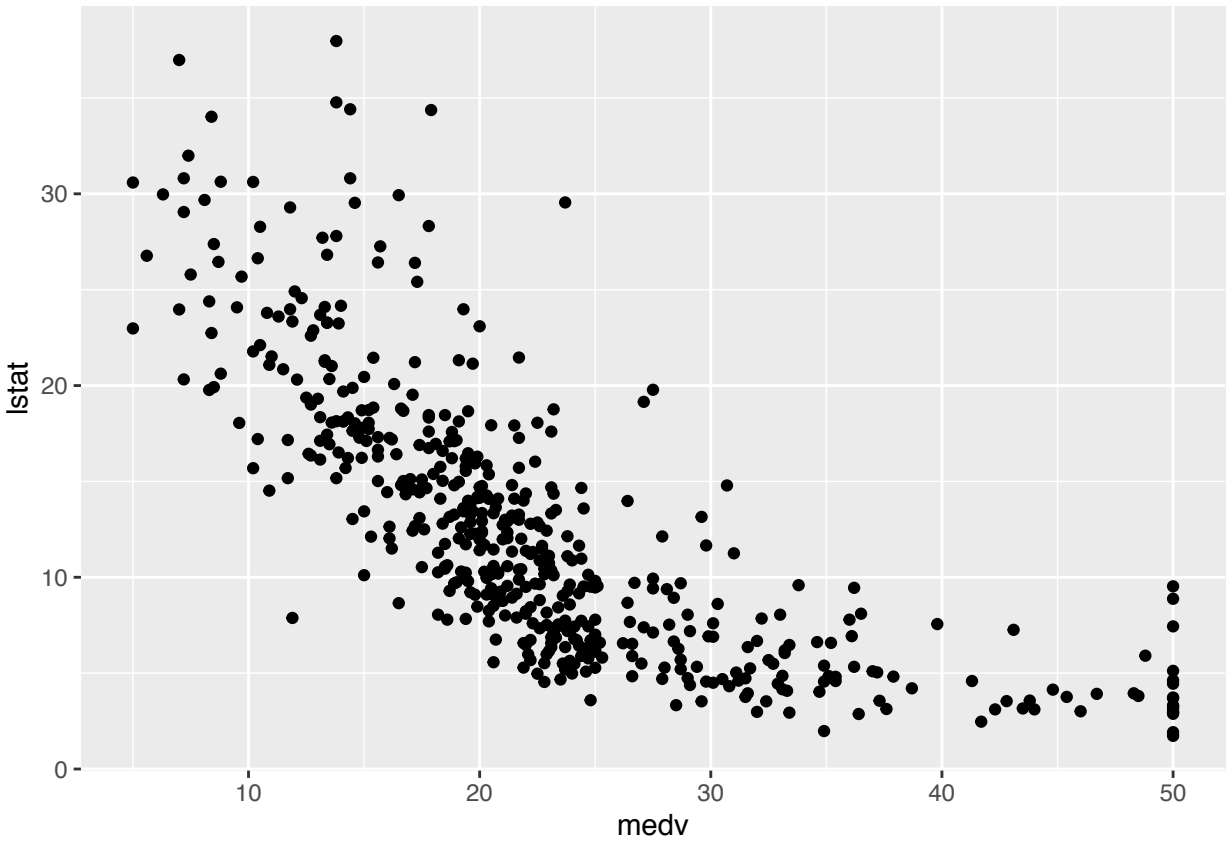
In this question you will build a linear regression model to predict the median value of owner occupied homes in USD 1000's `medv`.

```
library(mlbench)
data("BostonHousing2")
glimpse(BostonHousing2)
```

```
## Observations: 506
## Variables: 19
## $ town    <fct> Nahant, Swampscott, Swampscott, Marblehead, Marblehead, Mar...
## $ tract   <int> 2011, 2021, 2022, 2031, 2032, 2033, 2041, 2042, 2043, 2044,...
## $ lon     <dbl> -70.9550, -70.9500, -70.9360, -70.9280, -70.9220, -70.9165,...
## $ lat     <dbl> 42.2550, 42.2875, 42.2830, 42.2930, 42.2980, 42.3040, 42.29...
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9,...
## $ cmedv   <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 22.1, 16.5, 18.9,...
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.088...
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5...
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87,...
## $ chas    <fct> 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, 0.5...
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.6...
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9...
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9...
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4,...
## $ tax     <int> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311,...
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2,...
## $ b       <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396...
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17...
```

(a) Create a scatterplot of median value of homes `medv` and percentage of lower status of the population that lives in the census tract `lstat`. Describe the relationship.

```
BostonHousing2 %>%
  ggplot(aes(x=medv, y=lstat)) + geom_point()
```



There is a negative, strong, linear relationship between median value of homes and percentage of lower status of the population that lives in the census tract.

(b) Write out the mathematical description of a simple linear regression model of `medv` on `lstat`. Describe each of the variables and parameters in the model.

yihat= betaohat+ beta1hat*xi is the estimation of lstat corresponding to different medv values.

yihat:estimated lstat for the ith observation xi:medv for the ith observation betaohat: the lstat corresponding to 0 medv beta1hat: change in lstat with an increase in one unit of medv

(c) Use 80% of the data to select a training set, and leave 20% of the data for testing. Fit a linear regression model of `medv` on `lstat` on the training set.

```
set.seed(136)
n <- nrow(BostonHousing2)
training_indices <- sample(1:n, size=round(0.8*n))
train <- BostonHousing2[training_indices,]
test <- BostonHousing2[-training_indices,]

model5 <- lm(lstat ~ medv, data=train)
summary(model5)$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 25.1550805 0.63601654  39.55098 8.006305e-141
## medv        -0.5600065 0.02611815 -21.44128  1.333841e-68
```

yihat= 25.16 + -0.56*xi

yihat:estimated lstat for the ith observation xi:medv for the ith observation

(i) Calculate RMSE on the training and test data using the linear regression model you fit on the training set. Is there evidence of overfitting?

```
yhat_test <- predict(model5, newdata = test)
y_test <- test$lstat;
n_test <- nrow(test)
# RMSE for predictions in testing dataset
sqrt(sum((y_test - yhat_test)^2) / n_test)
```

```
## [1] 4.677768
```

```
yhat_train <- predict(model5, newdata = train)
y_train <- train$lstat;
n_train <- nrow(train)
# RMSE for predictions in training dataset
sqrt(sum((y_train - yhat_train)^2) / n_train)
```

```
## [1] 4.854191
```

Since there is a relatively small difference (4%) between the RMSE for the training data and that for the testing data. This suggests that our predictive model may not be overfitting the training data and is fairly useful for to make predictions for new observations.

(ii) Calculate the coefficient of determination.

```
summary(model5)$r.squared
```

```
## [1] 0.5328773
```

The coefficient of determination is 0.5328773.