```r
library(ggplot2)
library(GGally)
```

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

```r
library(NbClust)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(ggrepel)
library(MASS)
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.2
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## 
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
## 
##     margin
```

Problem 1

(a)

```
# Load data
salary <-  read.csv("data/salary.txt", sep = ",", header = TRUE)

# Create a factor variable for district size
salary <- within(salary, {
  size <- factor(districtSize, levels = c(1, 2, 3))
})

# View the structure of the data
str(salary)
```

```
## ’data.frame’:    325 obs. of  5 variables:
##  $ District    : chr  "Dubuque" "West Des Moines" "Ankeny" "Muscatine" ...
##  $ districtSize: int  3 3 3 3 3 3 3 3 3 3 ...
##  $ salary      : num  37730 39110 39501 38559 39079 ...
##  $ experience  : num  14.5 10.7 11.7 14.5 13 ...
##  $ size        : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
```

(b)

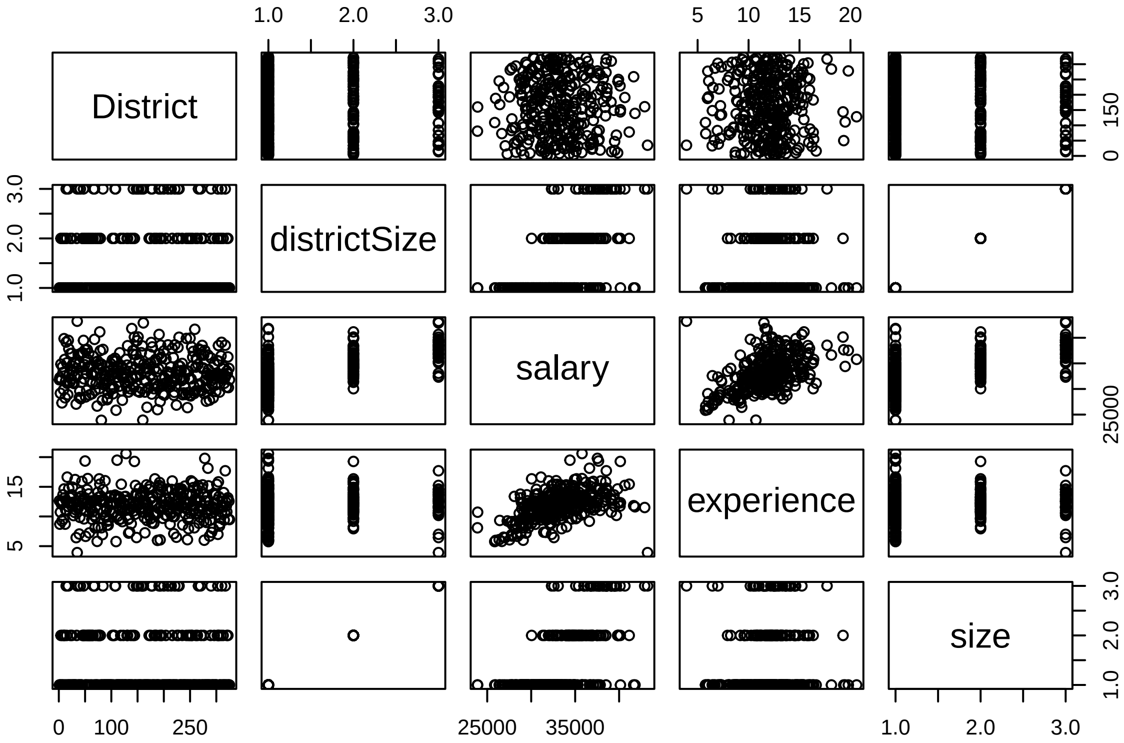
```
# Numerical summeries
summary(salary)
```

```
##    District          districtSize       salary        experience     size   
##  Length:325         Min.   :1.000   Min.   :23890   Min.   : 3.91   1:223  
##  Class :character   1st Qu.:1.000   1st Qu.:30848   1st Qu.:10.44   2: 68  
##  Mode  :character   Median :1.000   Median :32868   Median :11.97   3: 34  
##                     Mean   :1.418   Mean   :33168   Mean   :11.86          
##                     3rd Qu.:2.000   3rd Qu.:35297   3rd Qu.:13.33          
##                     Max.   :3.000   Max.   :43233   Max.   :20.60          
```

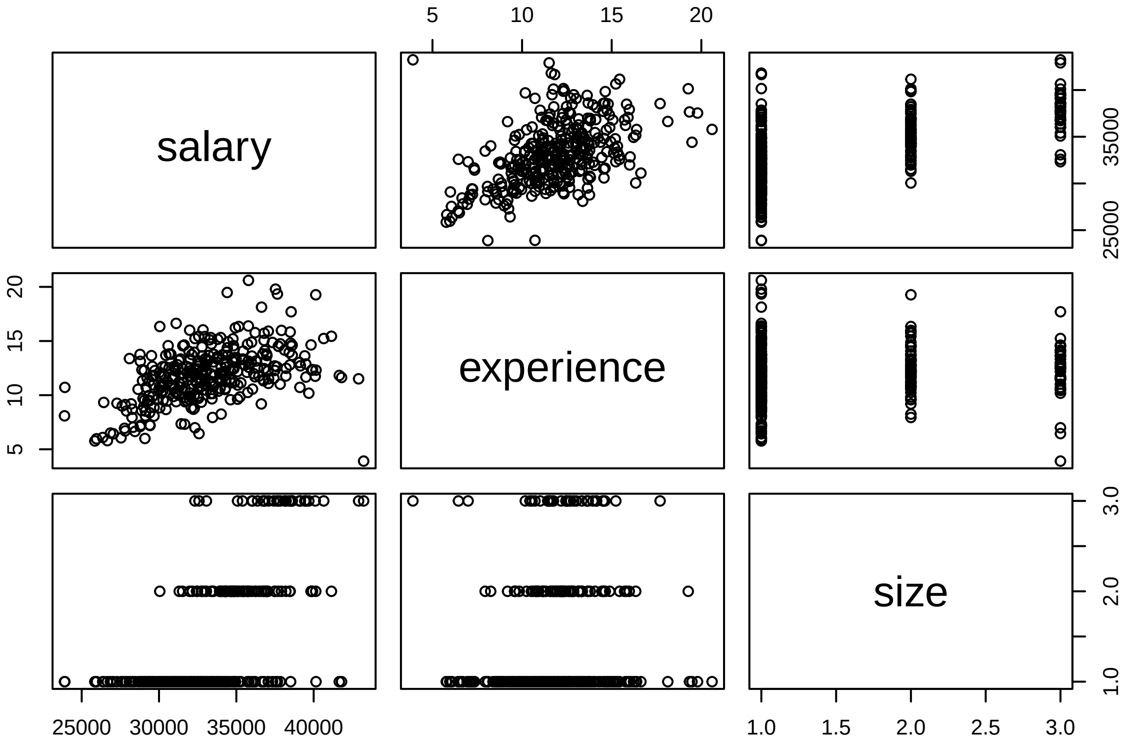
```
avg_salary <- mean(salary$salary)
aggregate(salary ~ size, data = salary, mean)
```

```
##   size   salary
## 1    1 31798.97
## 2    2 35326.11
## 3    3 37834.15
```

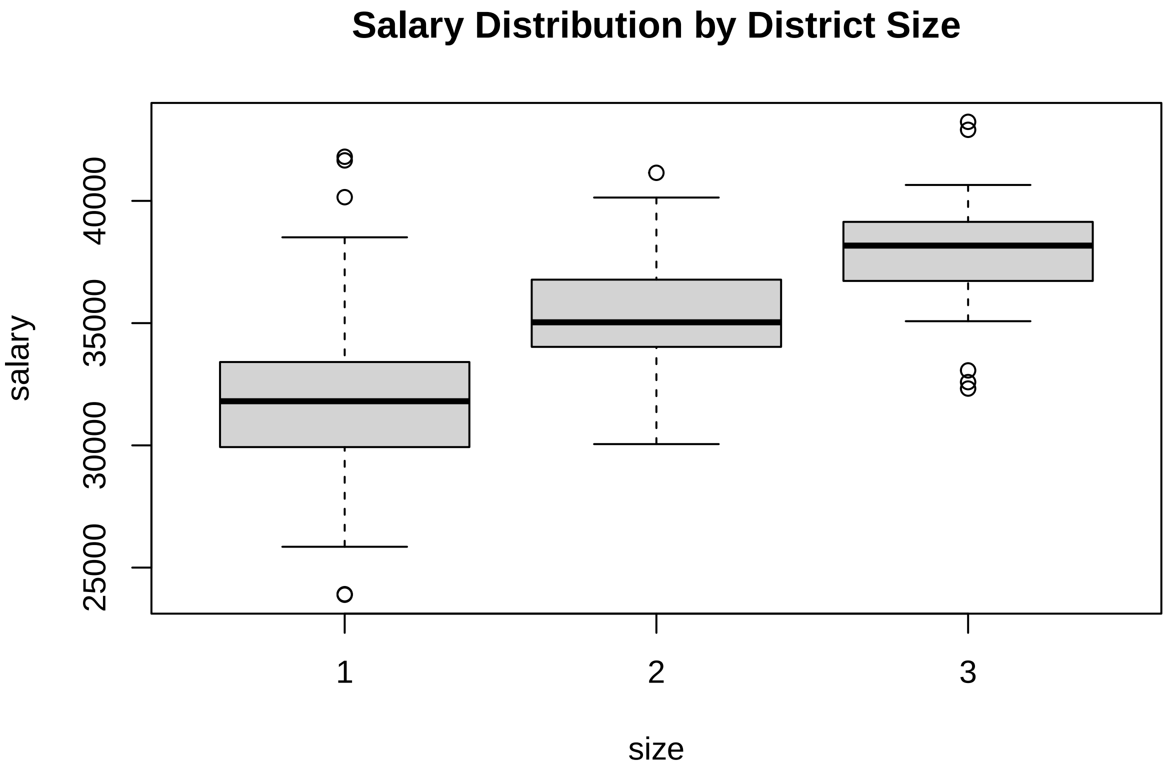
```
# Graphical summeries
plot(salary)
```



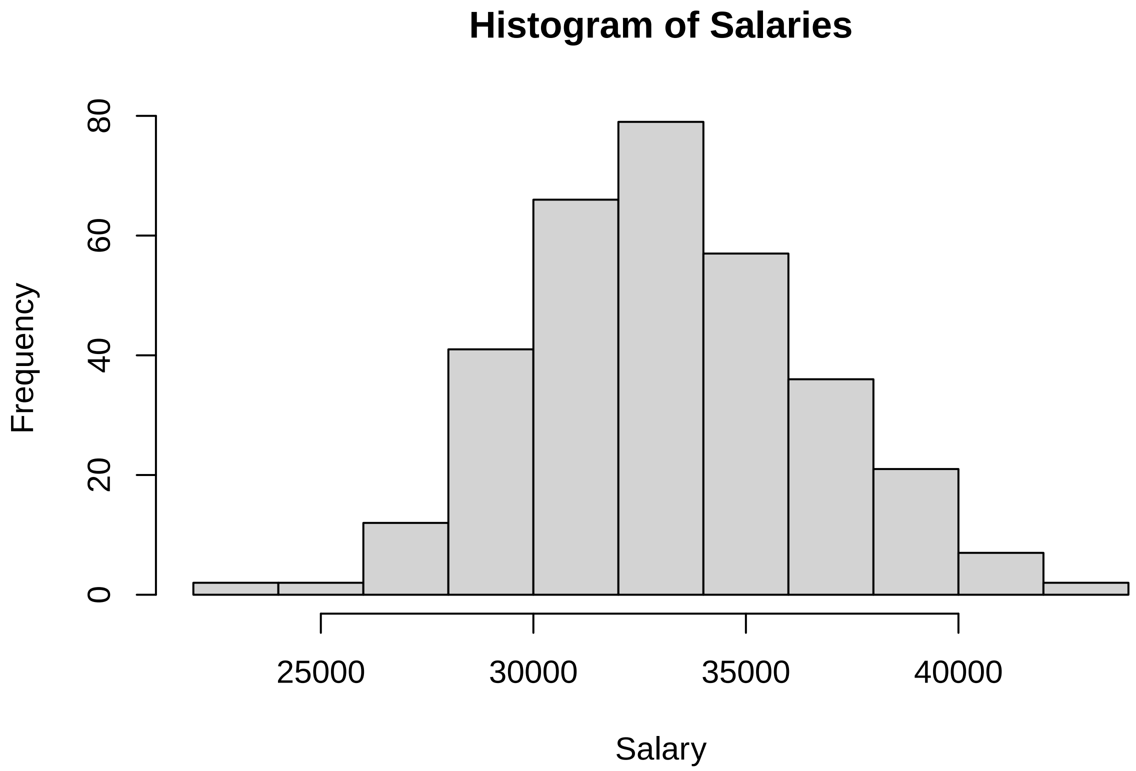
```
pairs(salary[c("salary", "experience", "size")])
```



```r
boxplot(salary ~ size, data = salary, main="Salary Distribution by District Size")
```



```r
hist(salary$salary, main = "Histogram of Salaries", xlab = "Salary")
```



The average salary is: 3.3168327×10^4

(c)

```
model_A <- lm(salary ~ experience + districtSize, data = salary)
model_B <- lm(salary ~ experience + size, data = salary)

summary(model_A)
```

```
##
## Call:
## lm(formula = salary ~ experience + districtSize, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7446.0 -1307.9  -180.7  1142.7 10099.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22065.55     622.14   35.47   <2e-16 ***
## experience     586.42      48.79   12.02   <2e-16 ***
## districtSize  2924.90     184.47   15.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2225 on 322 degrees of freedom
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5748
## F-statistic:   220 on 2 and 322 DF,  p-value: < 2.2e-16
```

```
summary(model_B)
```

```
##
## Call:
## lm(formula = salary ~ experience + size, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7577.1 -1283.9  -108.9  1141.3 10220.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24995.49     589.35  42.412   <2e-16 ***
## experience    584.15      48.96  11.932   <2e-16 ***
## size2        3088.00     310.73   9.938   <2e-16 ***
## size3        5732.28     410.84  13.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 321 degrees of freedom
## Multiple R-squared:  0.578,  Adjusted R-squared:  0.5741
## F-statistic: 146.6 on 3 and 321 DF,  p-value: < 2.2e-16
```

```
AIC(model_A, model_B)
```

```
##         df      AIC
## model_A  4 5937.259
## model_B  5 5938.828
```

```
BIC(model_A, model_B)
```

```
##         df      BIC
## model_A  4 5952.394
## model_B  5 5957.747
```

Looking at the AIC and BIC, model A slightly outperforms model B, which is surprising given that variables districtSize and size in theory represent exactly the same information. Considering the summary of both models, we can see that the linear model takes each factor of variable size in consideration with their own estimate (Intercept, size2, size3), while model A with the variable districtSize only uses one estimate which is multiplied with the districtSize. Therefore, we may conclude that model B slightly overfits the values because it has more variables to estimate on, and thats why we see a slight difference in performance between the models.

(d)

```
summary(model_B)
```

```
##
## Call:
## lm(formula = salary ~ experience + size, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7577.1 -1283.9  -108.9  1141.3 10220.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24995.49     589.35  42.412   <2e-16 ***
## experience    584.15      48.96  11.932   <2e-16 ***
## size2        3088.00     310.73   9.938   <2e-16 ***
## size3        5732.28     410.84  13.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 321 degrees of freedom
## Multiple R-squared:  0.578,  Adjusted R-squared:  0.5741
## F-statistic: 146.6 on 3 and 321 DF,  p-value: < 2.2e-16
```

Looking at the p-values we can see that all variables are significant (<2e-16). Therefore, each variable explains part of the variance and should not be dropped. We can interpret the Estimate as follows: The intercept is the base salary for every teacher. With each year of experience, we can add its coefficient to the salary, and depending on the districtSize, we can do the same.

(e)

```
model_C <- lm(salary ~ I(experience - 13) + size, data = salary)
summary(model_C)
```

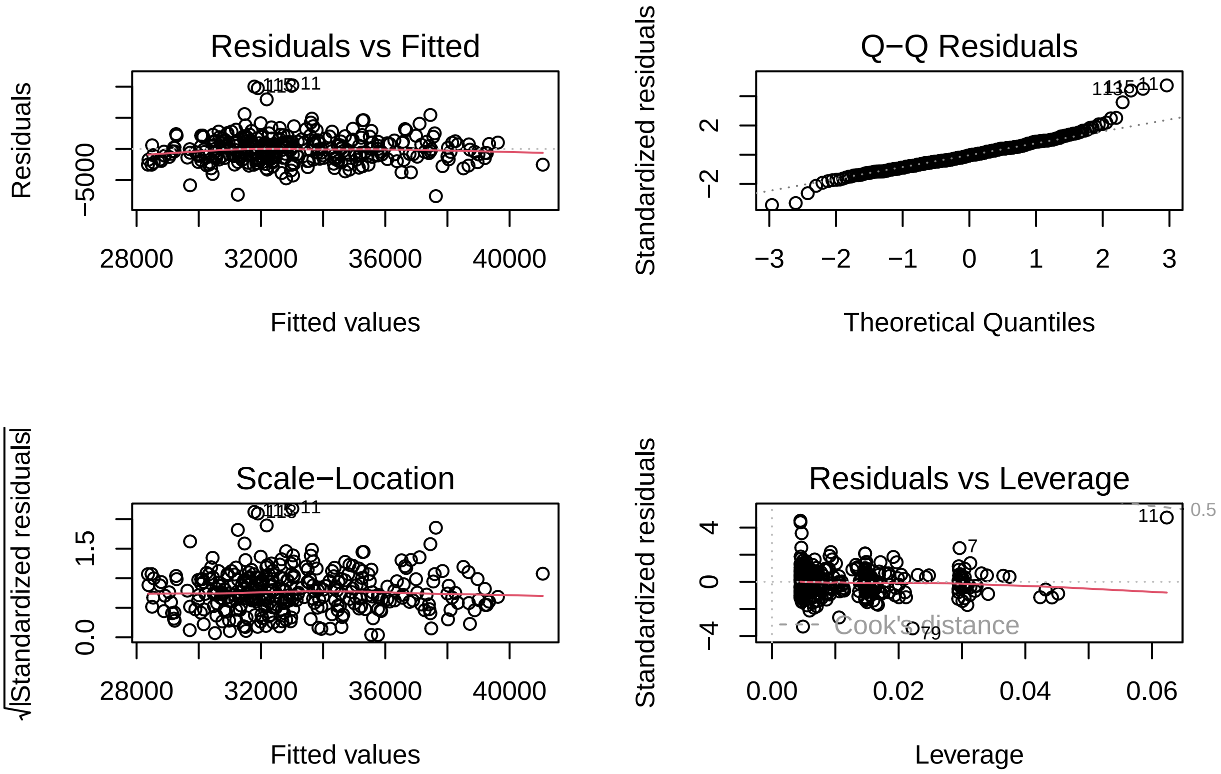
```
##
## Call:
## lm(formula = salary ~ I(experience - 13) + size, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7577.1 -1283.9  -108.9  1141.3 10220.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32589.46     163.20 199.691   <2e-16 ***
## I(experience - 13)   584.15      48.96  11.932   <2e-16 ***
## size2               3088.00     310.73   9.938   <2e-16 ***
## size3               5732.28     410.84  13.952   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2227 on 321 degrees of freedom
## Multiple R-squared:  0.578,  Adjusted R-squared:  0.5741
## F-statistic: 146.6 on 3 and 321 DF,  p-value: < 2.2e-16
```

The Intercept changes, but surprisingly by more than +13. All other coefficients stay the same, and all stay significant.

(f)

```r
par(mfrow = c(2,2))
plot(model_B)
```



Residuals vs Fitted: We can observe that the residuals are evenly spread and that there is no distinct pattern, therefore we can conclude that the data is linear. Q-Q Residuals: The Residuals follow the normal distribution for the most part, though both ends do deviate slightly. Still, we would argue that the error terms are normally distributed. Scale-Location: We observe a horizontal line which is spread equally, therefore homoscedasticity holds. Residuals vs Leverage: There are no points that have a particularly high leverage on the model.

(g)

```r
new_data_A <- data.frame(experience = 10, districtSize = 3)
prediction_A <- predict(model_A, newdata = new_data_A)

new_data_B <- data.frame(experience = 10, size = as.factor(3))
prediction_B <- predict(model_B, newdata = new_data_B)
```

Prediction for Model A: 3.6704423×10^4 Prediction for Model B: 3.6569287×10^4

Problem 2

(a)

The standard error of Beta_hat_2, we can use the following formula: SE(Beta_hat_2) = sqrt(s^2 * (X_t * X)^-1[3,3]) = sqrt(2 * 2) = 2

(b)

We can perform a t-test to test the Hypothesis that Beta2 = 0:

```r
# t-statistic
t_stat_beta2 <- 15 / 2

# p-value
p_value_beta2 <- 2 * pt(-abs(t_stat_beta2), df = 22)
p_value_beta2
```

```
## [1] 1.694205e-07
```

We get a significant value as the p-value, therefore we can reject the hypthesis that the true value = 0. Note: We have 22 degrees of freedom from the number of points - number of predictors: 25 - 3 = 22

(c)

```r
# TODO: Cant solve atm
```

(d)

We can perform a t-test to test the Hypothesis that beta1 - beta2 = 0

```r
# t-statistic for beta1 = beta2
t_stat_diff <- (12 - 15) / sqrt(6)

# p-value
p_value_diff <- 2 * pt(t_stat_diff, df = 22)
p_value_diff
```

```
## [1] 0.233624
```

We get a high p-value (0.23), therefore we don't reject the hypothesis that beta1 = beta2.

(e)

```r
# Total sum of squares
SST <- 120

# Residual sum of squares (SSE)
SSE <- 2 * (25 - 3)

# Regression sum of squares (SSR)
SSR <- SST - SSE

# F-statistic
F_stat <- (SSR / 2) / (SSE / 22)
p_value_F <- pf(F_stat, df1 = 2, df2 = 22, lower.tail = FALSE)

# R-squared
R_squared <- SSR / SST

list(F_stat = F_stat, p_value_F = p_value_F, R_squared = R_squared)
```

```
## $F_stat
## [1] 19
##
## $p_value_F
## [1] 1.610593e-05
##
## $R_squared
## [1] 0.6333333
```

From R_squared, we can see that 63.3% of the variance in y has been explained by the model.