# Insurance-PPA4.R

Souvik

2020-11-11

```
setwd("C:/Users/Souvik/Downloads/PPA")

library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Loading required package: carData
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.3
```

```
## corrplot 0.84 loaded
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.0.3
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.0.3
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
## The following object is masked from 'package:car':
## 
##     logit
```

```
insurance <- read.csv("insurance_LR1.csv", stringsAsFactors = TRUE)
dim(insurance)
```

```
## [1] 1338    7
```

```
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
summary(insurance)
```

```
##       age            sex           bmi           children     smoker    
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064  
##  1st Qu.:27.00   male  :676   1st Qu.:26.27   1st Qu.:0.000   yes: 274  
##  Median :39.00                Median :30.40   Median :1.000             
##  Mean   :39.21                Mean   :30.66   Mean   :1.095             
##  3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000             
##  Max.   :64.00                Max.   :53.13   Max.   :5.000             
##                               NA's   :2                                 
##        region       charges     
##  northeast:324   Min.   : 1122  
##  northwest:325   1st Qu.: 4740  
##  southeast:364   Median : 9382  
##  southwest:325   Mean   :13270  
##                  3rd Qu.:16640  
##                  Max.   :63770  
## 
```

```
#to convert variable into factor variable
insurance$sex <- as.factor(insurance$sex)
summary(insurance)
```

```
##       age            sex           bmi           children     smoker    
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064  
##  1st Qu.:27.00   male  :676   1st Qu.:26.27   1st Qu.:0.000   yes: 274  
##  Median :39.00                Median :30.40   Median :1.000             
##  Mean   :39.21                Mean   :30.66   Mean   :1.095             
##  3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000             
##  Max.   :64.00                Max.   :53.13   Max.   :5.000             
##                               NA's   :2                                 
##        region       charges     
##  northeast:324   Min.   : 1122  
##  northwest:325   1st Qu.: 4740  
##  southeast:364   Median : 9382  
##  southwest:325   Mean   :13270  
##                  3rd Qu.:16640  
##                  Max.   :63770  
## 
```

```
insurance$children <- as.factor(insurance$children)
summary(insurance)
```

```
##       age            sex           bmi        children smoker    
##  Min.   :18.00   female:662   Min.   :15.96   0:574    no :1064  
##  1st Qu.:27.00   male  :676   1st Qu.:26.27   1:324    yes: 274  
##  Median :39.00                Median :30.40   2:240              
##  Mean   :39.21                Mean   :30.66   3:157              
##  3rd Qu.:51.00                3rd Qu.:34.70   4: 25              
##  Max.   :64.00                Max.   :53.13   5: 18              
##                               NA's   :2                          
##        region       charges     
##  northeast:324   Min.   : 1122  
##  northwest:325   1st Qu.: 4740  
##  southeast:364   Median : 9382  
##  southwest:325   Mean   :13270  
##                  3rd Qu.:16640  
##                  Max.   :63770  
## 
```

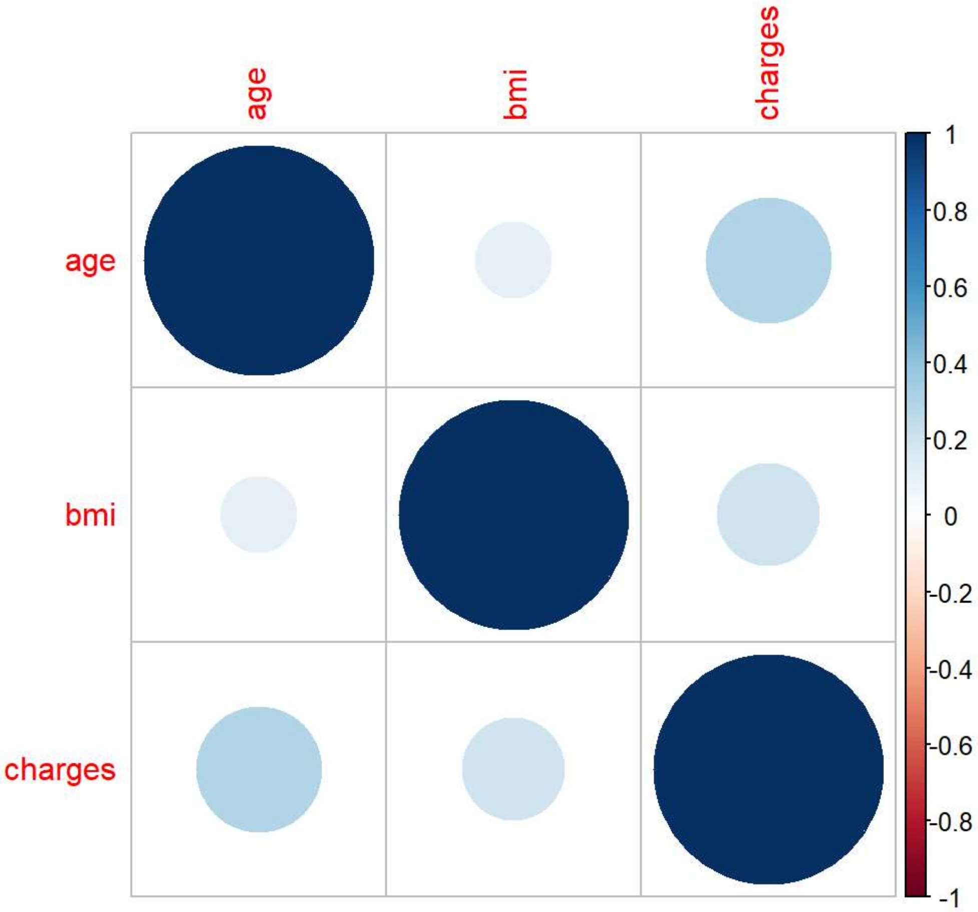
```
#treat Missing values
#is.na(insurance$bmi)

#insurance$bmi[is.na(insurance$bmi)] <- mean(insurance$bmi, na.rm = TRUE)
#insurance$bmi[is.na(insurance$bmi)] <- 0
insurance <- na.omit(insurance)
summary(insurance)
```

```
##       age            sex           bmi           children smoker
##  Min.   :18.00   female:662   Min.   :15.96   0:574    no :1062
##  1st Qu.:26.75   male  :674   1st Qu.:26.27   1:323    yes: 274
##  Median :39.00                Median :30.40   2:240
##  Mean   :39.21                Mean   :30.66   3:157
##  3rd Qu.:51.00                3rd Qu.:34.70   4: 24
##  Max.   :64.00                Max.   :53.13   5: 18
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:324   1st Qu.: 4744
##  southeast:363   Median : 9389
##  southwest:325   Mean   :13281
##                  3rd Qu.:16687
##                  Max.   :63770
```

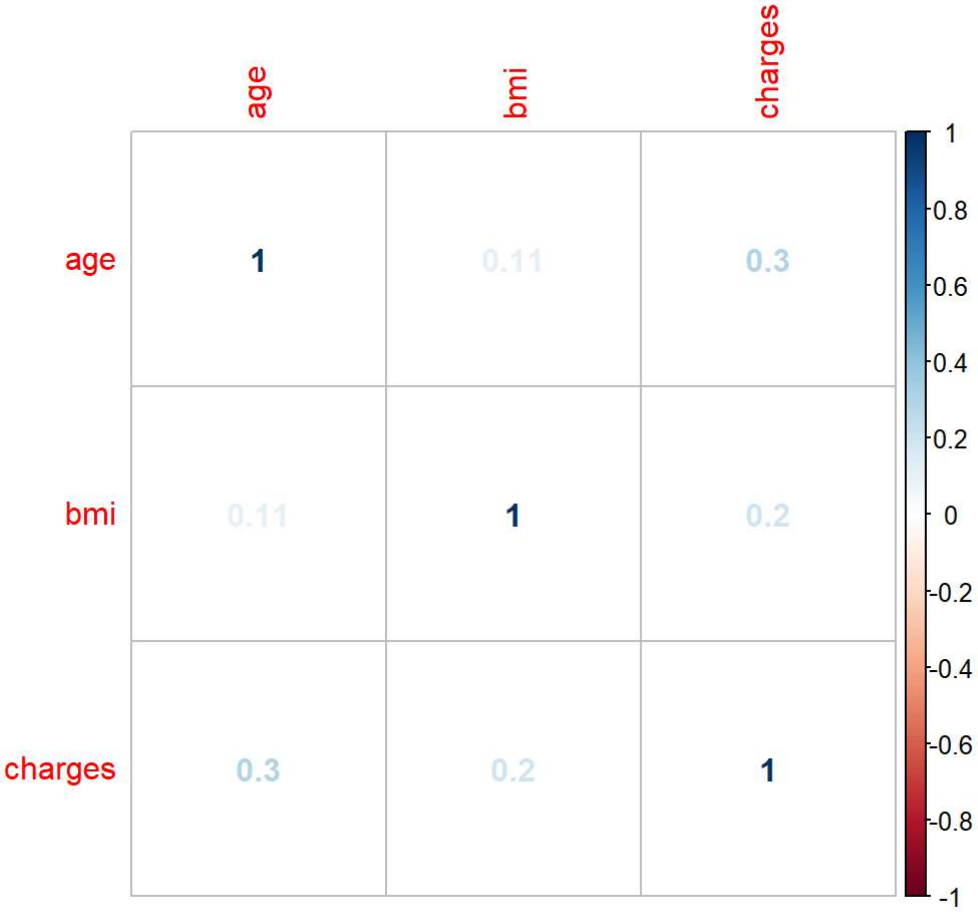
```
cr <- cor(insurance[c("age","bmi","charges")])
cr
```

```
##               age       bmi   charges
## age     1.0000000 0.1092419 0.2988486
## bmi     0.1092419 1.0000000 0.1983498
## charges 0.2988486 0.1983498 1.0000000
```

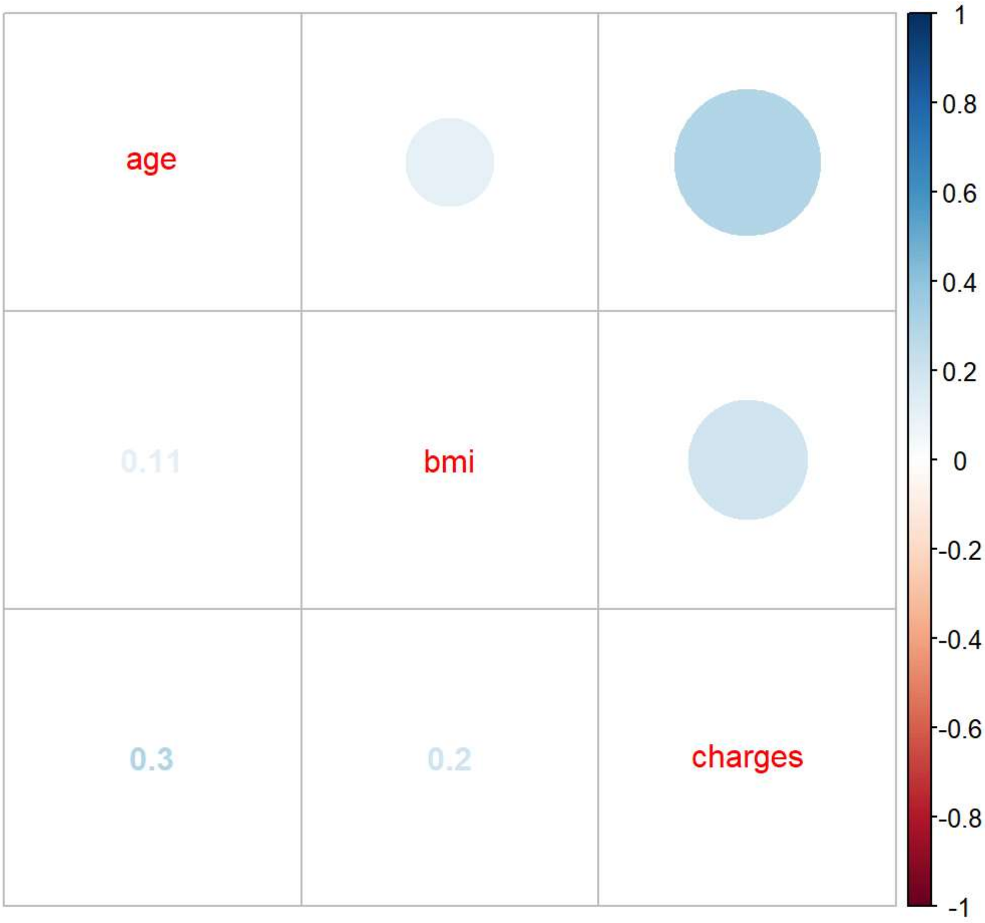
```
corrplot(cr, type = "full")
```



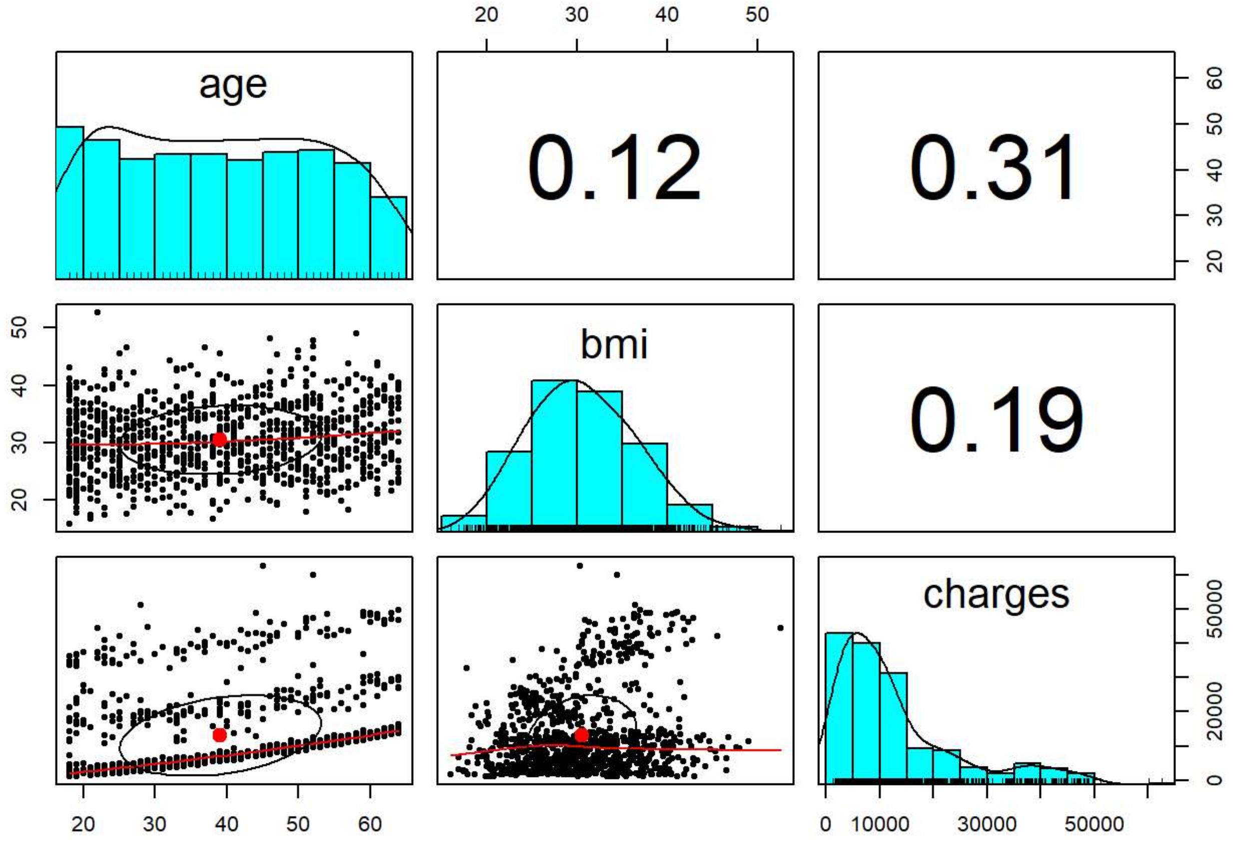
```
corrplot(cr,method = "number")
```



```
corrplot.mixed(cr)
```



```
#To Dummy variables

insurance$smoker_y <- ifelse(insurance$smoker == "yes", 1,0)
insurance$smoker_n <- ifelse(insurance$smoker == "no", 1,0)
insurance$region_se <- ifelse(insurance$region == "southeast", 1,0)
insurance$region_ne <- ifelse(insurance$region == "northeast", 1,0)
insurance$region_sw <- ifelse(insurance$region == "southwest", 1,0)
insurance$region_nw <- ifelse(insurance$region == "northwest", 1,0)

#Splitting of dataset into training and testing

split <- sample.split(insurance$charges, SplitRatio = 0.7)
training_data <- subset(insurance, split == "TRUE")
testing_data <- subset(insurance, split == "FALSE")

#pair.panels

pairs.panels(training_data[c("age","bmi","charges")])
```



```
#Linear Regression
model1 <- lm(charges ~ age, data = training_data)
summary(model1)
```

```
##
## Call:
## lm(formula = charges ~ age, data = training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7897  -6488  -5807   5664  47953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2905.76    1086.59   2.674  0.00762 **
## age           260.77      26.18   9.960  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11230 on 933 degrees of freedom
## Multiple R-squared:  0.0961, Adjusted R-squared:  0.09513
## F-statistic:  99.2 on 1 and 933 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(charges ~ age+bmi, data = training_data)
summary(model2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi, data = training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12066  -6885  -4970   6660  48095
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5884.97    2039.49  -2.886    0.004 **
## age           244.97      26.03   9.411  < 2e-16 ***
## bmi           308.28      60.83   5.068 4.86e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11090 on 932 degrees of freedom
## Multiple R-squared:  0.1203, Adjusted R-squared:  0.1185
## F-statistic: 63.75 on 2 and 932 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(charges ~ age+bmi+smoker_y, data = training_data)
summary(model3)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker_y, data = training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12116  -3177  -1106   1294  29554
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10325.14    1123.22  -9.192   <2e-16 ***
## age            239.32      14.28  16.754   <2e-16 ***
## bmi            309.87      33.38   9.282   <2e-16 ***
## smoker_y     23187.02     498.43  46.520   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6084 on 931 degrees of freedom
## Multiple R-squared:  0.7354, Adjusted R-squared:  0.7345
## F-statistic: 862.5 on 3 and 931 DF,  p-value: < 2.2e-16
```

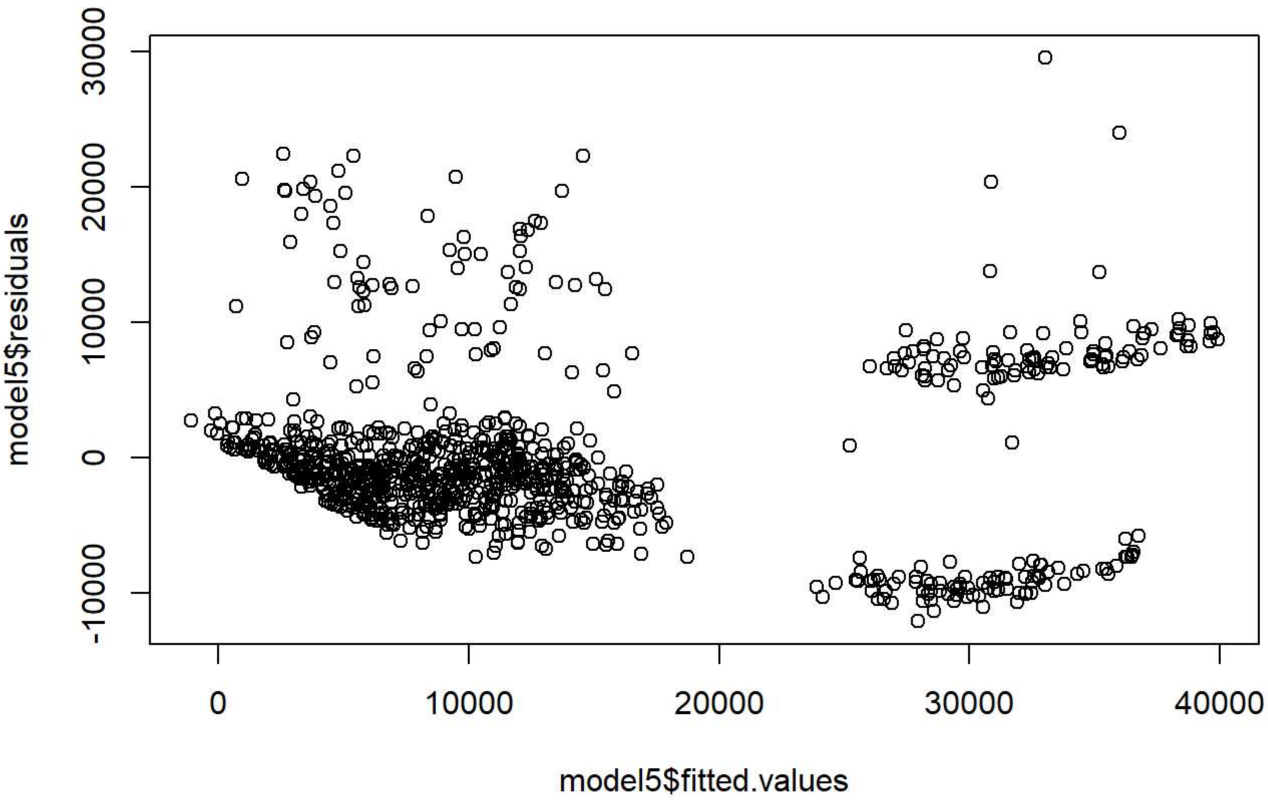
```
model4 <- lm(charges ~ age+bmi+smoker_y+region_nw, data = training_data)
summary(model4)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker_y + region_nw, data = training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12004  -3177  -1130   1310  29631
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10501.49    1150.13  -9.131   <2e-16 ***
## age            239.19      14.29  16.739   <2e-16 ***
## bmi            313.06      33.69   9.293   <2e-16 ***
## smoker_y     23195.49     498.70  46.512   <2e-16 ***
## region_nw      334.35     466.32   0.717    0.474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6085 on 930 degrees of freedom
## Multiple R-squared:  0.7355, Adjusted R-squared:  0.7344
## F-statistic: 646.7 on 4 and 930 DF,  p-value: < 2.2e-16
```

```
model5 <- lm(charges ~ age+bmi+smoker_y, data = training_data)
summary(model5)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker_y, data = training_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12116  -3177  -1106   1294  29554
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10325.14    1123.22  -9.192   <2e-16 ***
## age            239.32      14.28  16.754   <2e-16 ***
## bmi            309.87      33.38   9.282   <2e-16 ***
## smoker_y     23187.02     498.43  46.520   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6084 on 931 degrees of freedom
## Multiple R-squared:  0.7354, Adjusted R-squared:  0.7345
## F-statistic: 862.5 on 3 and 931 DF,  p-value: < 2.2e-16
```

```
plot(model5$fitted.values,model5$residuals)
```



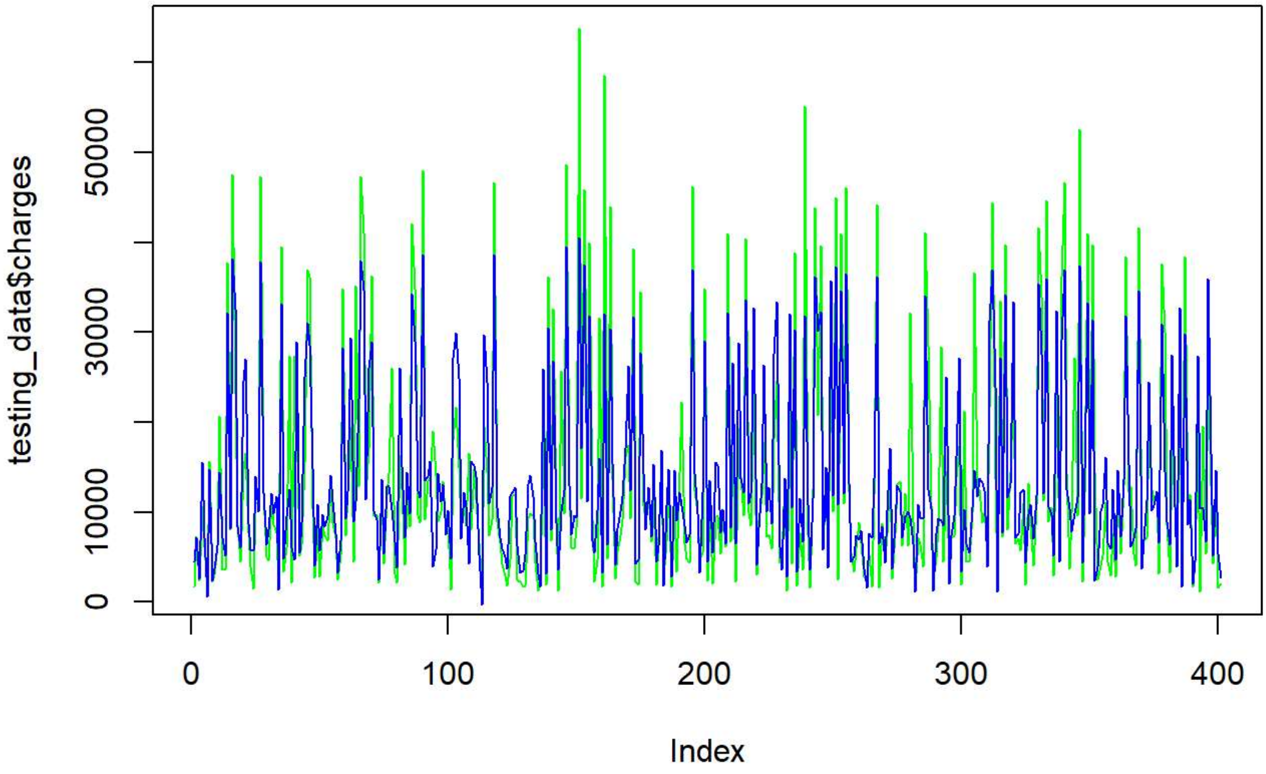
```
#prediction
prediction <- predict(model5, testing_data)
head(prediction)
```

```
##          2          8         18         19         25         36
##  4446.7164  7125.2096  2567.8693 15564.1165  7213.5215   550.8663
```

```
head(testing_data$charges)
```

```
## [1]  1725.552  7281.506  2395.172 10602.385  6203.902  1625.434
```

```
plot(testing_data$charges,type="l",col="green")
lines(prediction,type="l",col="blue")
```



```
#marketing dataset
#amount spend dependent variable


#set WD on top
#Write all the libraries on the top
#There should be no views in your code
#No error in model
```