Individual project

Jianing Shi

2020/3/26

Medical Cost Personal Datasets

Introduction

Goal of the project

This is an analysis of the data set “Medical Cost Personal Datasets”. The goal of this project is to find which of the predictors have the most impact on the prediction of the personal medical cost. I am interested in how does the different factor influence on the individual medical cost billed by health insurance, such as gender, age and so on.

Describe the dataset

```
#install R package "psych"
library(psych)

## Warning: package 'psych' was built under R version 3.6.3

cost <- read.csv("C:/Users/jenny/Documents/insurance.csv", sep=",")
str(cost)

## '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(cost)

##       age            sex           bmi           children     smoker

##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :10
64
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 2
74
```

```
##  Median :39.00                Median :30.40   Median :1.000

##  Mean   :39.21                Mean   :30.66   Mean   :1.095

##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000

##  Max.   :64.00                Max.   :53.13   Max.   :5.000

##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

```
describe(cost)
```

```
##          vars    n     mean       sd  median  trimmed     mad     mi
n
## age         1 1338    39.21    14.05   39.00    39.01   17.79   18.0
0
## sex*        2 1338     1.51     0.50    2.00     1.51    0.00    1.0
0
## bmi         3 1338    30.66     6.10   30.40    30.50    6.20   15.9
6
## children    4 1338     1.09     1.21    1.00     0.94    1.48    0.0
0
## smoker*     5 1338     1.20     0.40    1.00     1.13    0.00    1.0
0
## region*     6 1338     2.52     1.10    3.00     2.52    1.48    1.0
0
## charges     7 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.8
7
##               max    range  skew kurtosis     se
## age         64.00    46.00  0.06    -1.25   0.38
## sex*         2.00     1.00 -0.02    -2.00   0.01
## bmi         53.13    37.17  0.28    -0.06   0.17
## children     5.00     5.00  0.94     0.19   0.03
## smoker*      2.00     1.00  1.46     0.14   0.01
## region*      4.00     3.00 -0.04    -1.33   0.03
## charges  63770.43 62648.55  1.51     1.59 331.07
```

In the code above, I used the r package “psych”. There is 1338 observations of 7 variables.

Inputs: 1. age: age of primary beneficiary 2. sex: insurance contractor gender, female, male 3. bmi: body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9 4. children: number of

children covered by health insurance/Number of dependents 5. smoker: smoking 6. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest

Output: 1. charges: individual medical costs billed by health insurance.

Note that the variables: sex, smoker and region are categorical variables. So I creat a table below to summrize the dataset. The dataset is simulated on the basis of demographic statistics from the US Census Bureau, according to the book from which it is from.

```
name <- c("age", "sex", "bmi", "children", "smoker", "region", "charges
")
type <- c("num", "factor", "num", "num", "factor", "factor", "num")
missingvalue <- c(rep(0,7))
mytable <- matrix(c(name, type, missingvalue), nrow = 7, ncol =3, dimna
mes = list(c(1:7), c("name", "type", "missing value")))
mytable
```

```
##   name       type     missing value
## 1 "age"      "num"    "0"
## 2 "sex"      "factor" "0"
## 3 "bmi"      "num"    "0"
## 4 "children" "num"    "0"
## 5 "smoker"   "factor" "0"
## 6 "region"   "factor" "0"
## 7 "charges"  "num"    "0"
```

What other people have done

On the website of kaggle, I found that there was a person using that dataset, he was interested in "Can you accurately predict insurance costs?" The goal of his analysis is to predict the variable charges by comparing the significance of input variables in Python. There were not much of variable selection, and there were not much explanation of the variables. For their analysis, they used a single method to predict the result, by adding or dropping the predictors, they try to achieve a higher accuracy.

Main difference

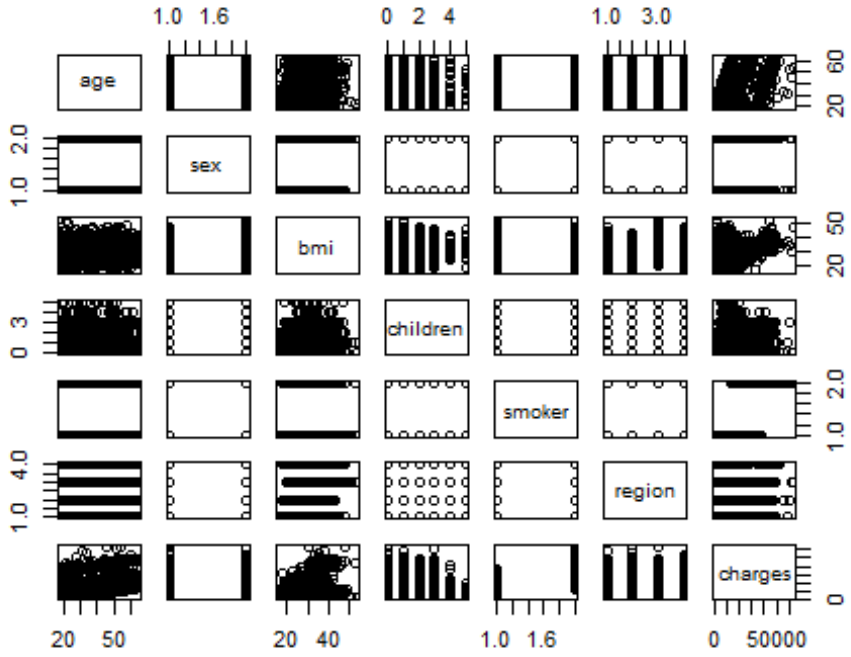
I will provide clear data visulization of the data set, and show the variable selection to avoid overfitting. Others wanted to know that how to predict the cost and reach a high accuracy. For this dataset, the output(charges) is non-categorical, so the method of predicting this data is limited. But I can still choose some different method of variable selections. By comparing the result of different variable selection method, and sub the result into the linear regression model, I will get the result of which method of variable selection is best fit for this data set.

Data visulization

I will create some graphs to provide a data visualization. For those graphs below, I want to see the distribution of each variables, the correlation between each variables, the outliers and leverage of each variables.

Graph of all variables

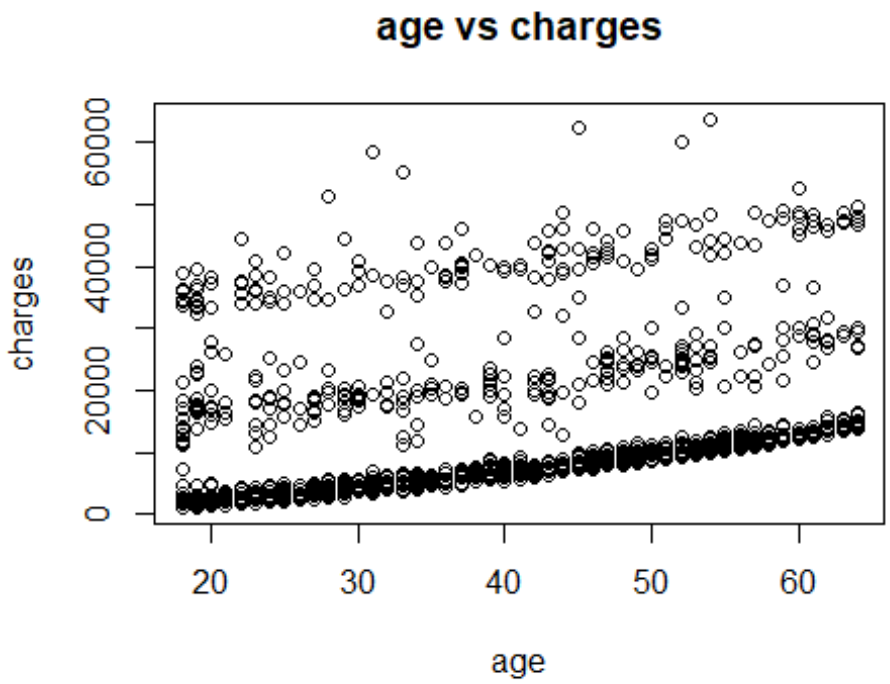
```
plot(cost)
```



The figure of all variables are not that clear to see, we can only see that there are some outliers for sex and region, and those outliers are above the maximum. I want to know more details about the variables, so I creat some graphs below.

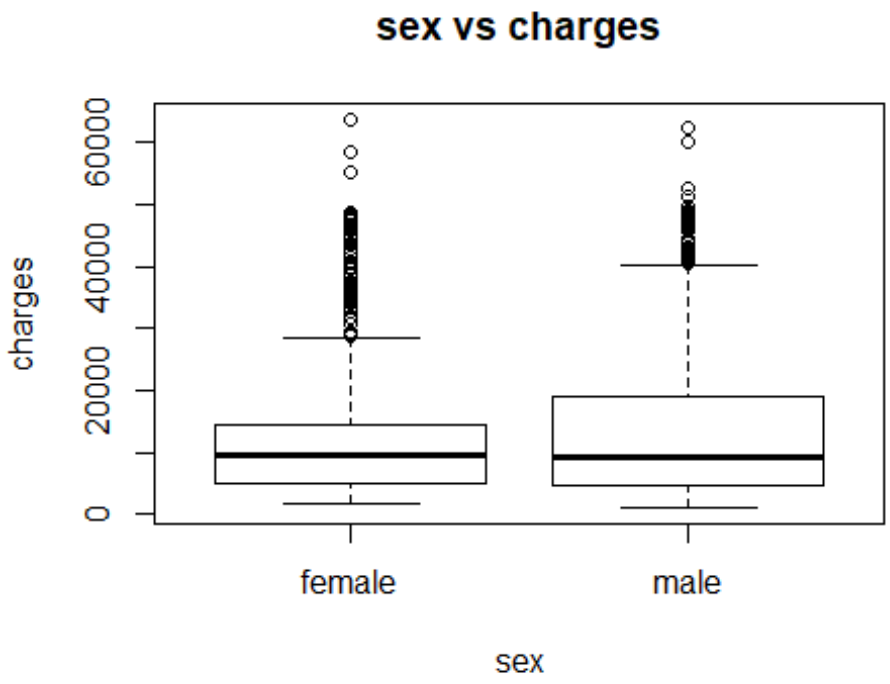
Graph of the relationship between each input and output(charges)

```
# age vs charges
plot(charges~age, data = cost, main = "age vs charges")
```



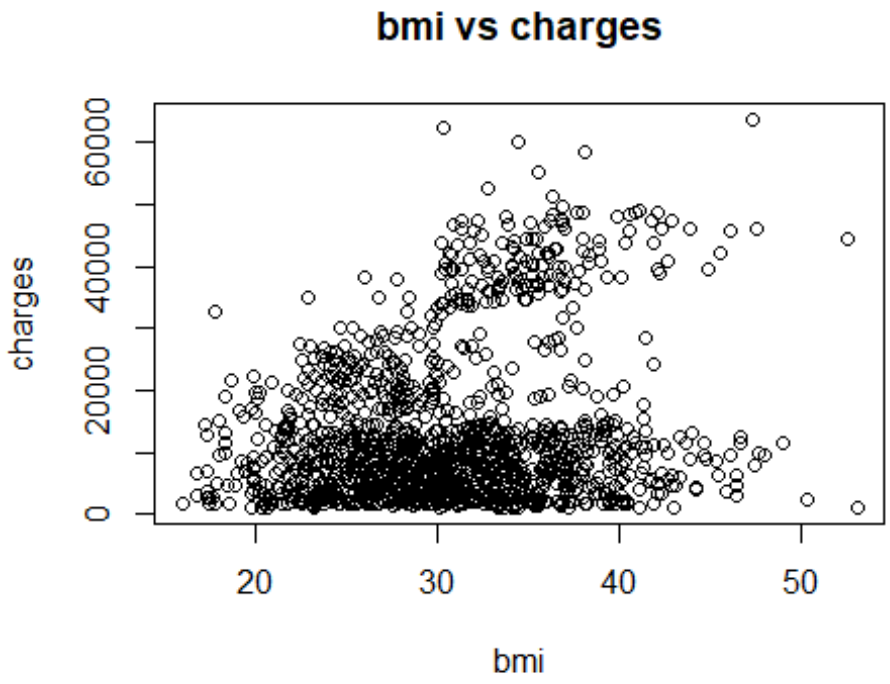
From the plot above, clearly we can see that there are three groups of charges. The lowest one is from 0 to around 10000 dollars of the charges, and this group contains the most amount of people. The middle group is about 1000 to 30000 dollars. The highest group is above 30000 dollars. Also, for each group, as the age is increasing, the charge is also increasing.

```
#sex vs charges
plot(charges~sex, data = cost, main = "sex vs charges")
```



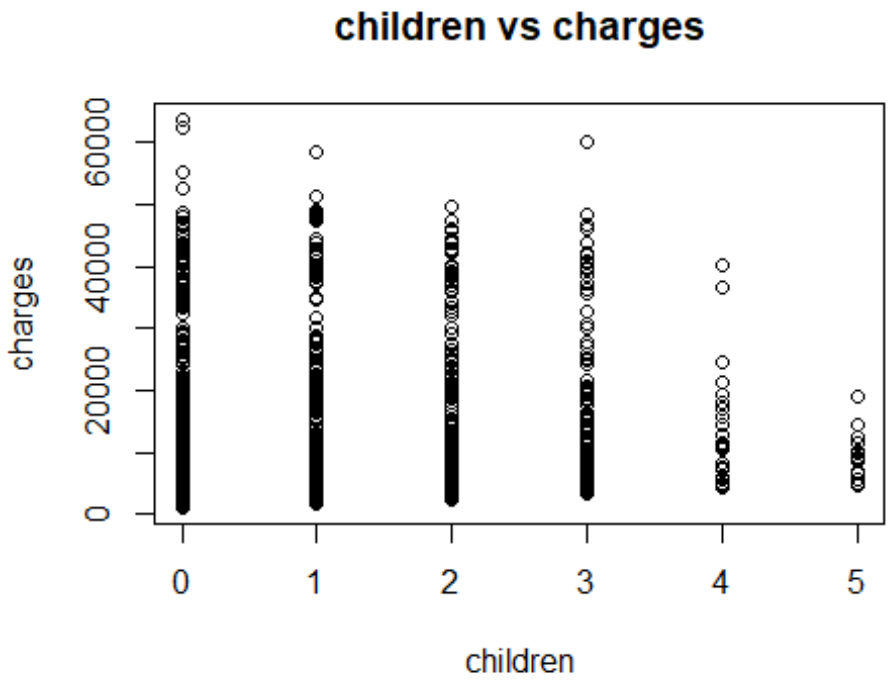
From the boxplot above, we can see that the median charge of both of female and male is around 10000 dollars, but the interquartile range for male is significantly higher than female. The minimum charge for both female and male is around 0 dollars, but the maximum charge for male is about 10000 dollars higher than female. Both of female and male have outliers.

```
#bmi vs charges
plot(charges~bmi, data = cost, main = "bmi vs charges")
```



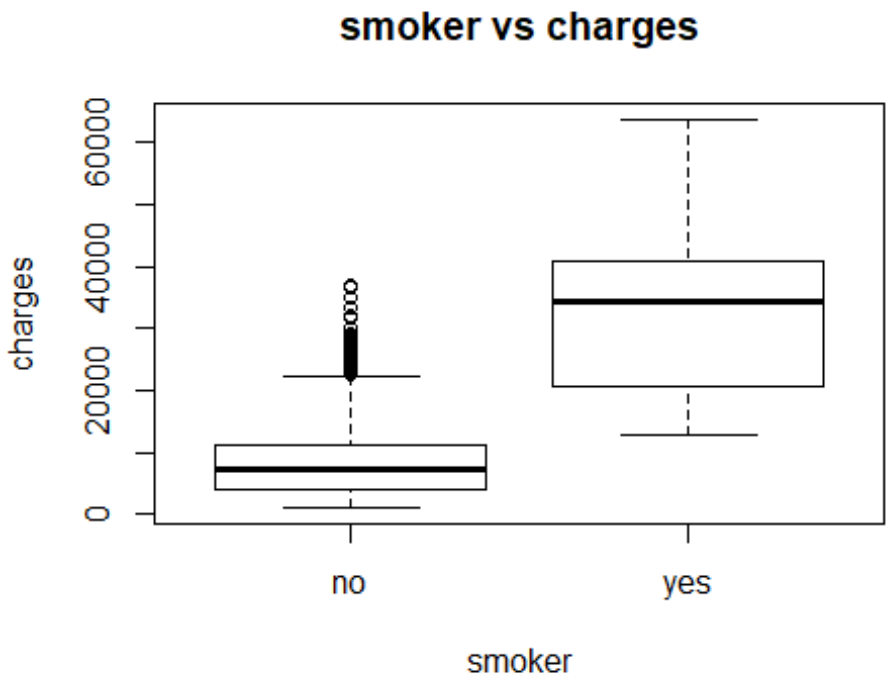
The plot above shows that the higher charges(above 30000 dollars) always happens on people who has a bmi higher than 30.

```
#children vs charges
plot(charges~children, data = cost, main = "children vs charges")
```



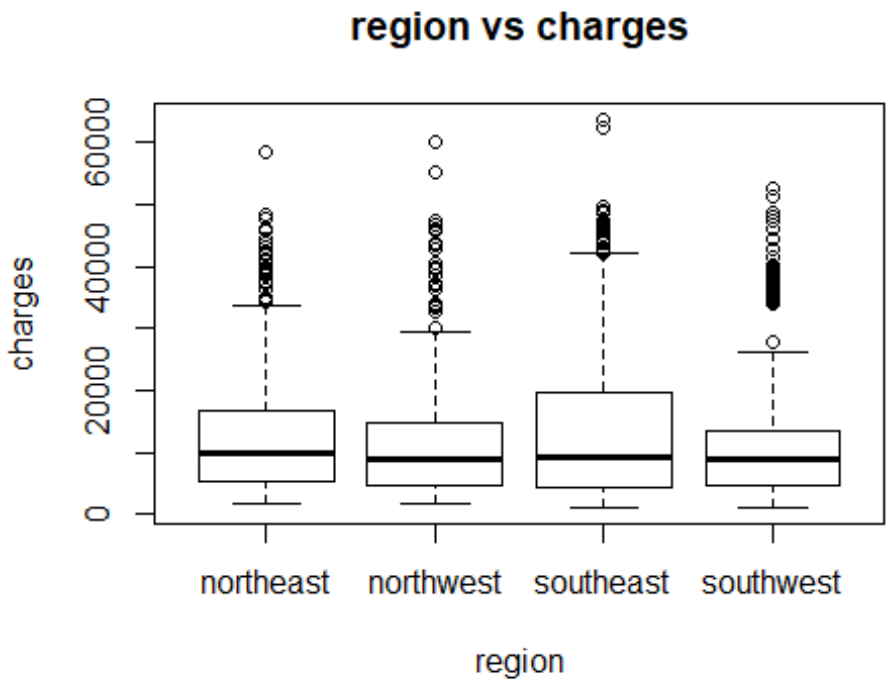
The figure above shows that people that has 0 child has the highest medical cost. Also, as the number of children increases from 0 to 5, the cost decreases.

```
#smoker vs charges
plot(charges~smoker, data = cost, main = "smoker vs charges")
```



The boxplot shows a significant difference of the charges between non-smoker and smoker. As we can see, the median charges of non-smoker is below 10000 dollars with a maximum charge below 30000 dollars(with some outliers). For smoker, the minimum charge is about 15000 dollars, which is even higher than the maximum charge of non-smoker. The median charge of smoker is around 40000 dollars and the maximum charge is above 60000 dollars.

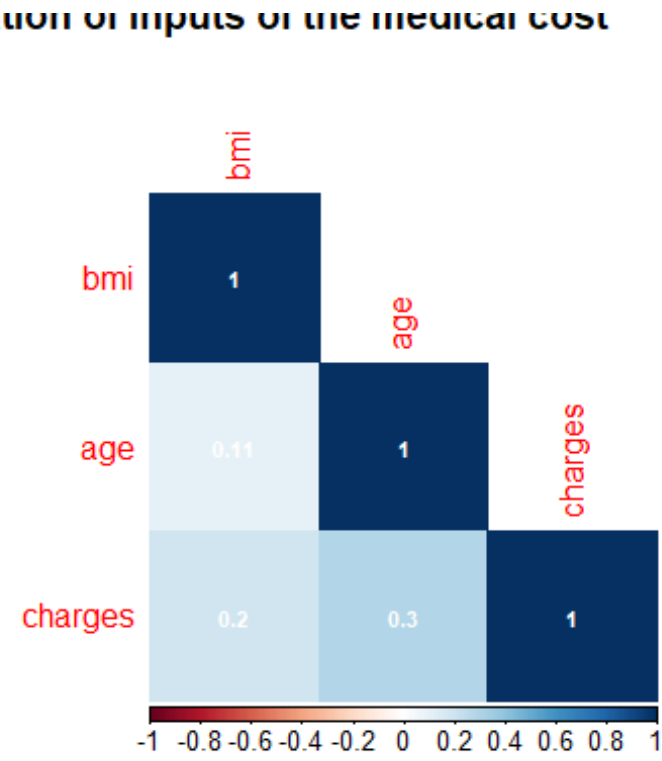
```
#region vs charges
plot(charges~region, data = cost, main = "region vs charges")
```



The median charge of all the four region are about the same, which is around 10000 dollars. The southeast region has the highest maximum charges(about 45000 dollars). The southwest region has the lowest maximum charges(below 30000 dollars). All of the four regions have outliers.

Graph of correlation

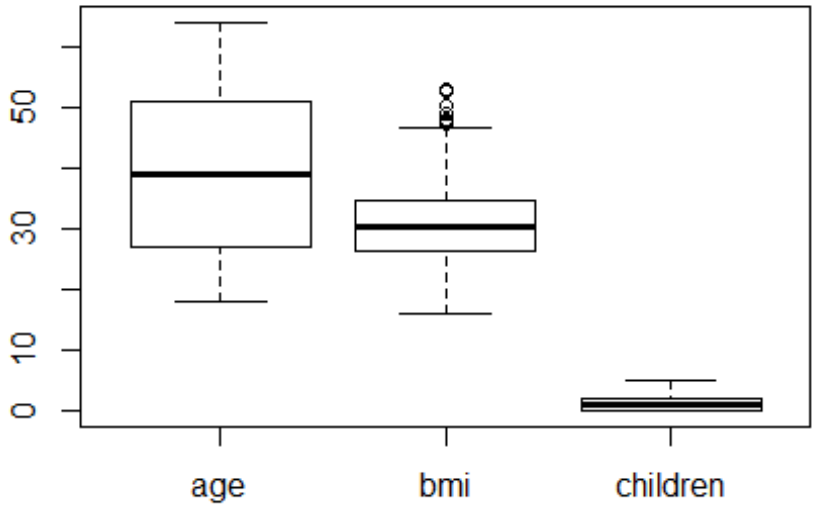
```
#graph of correlation of variables
#install r package "corrplot"
library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.3

## corrplot 0.84 loaded

cost_cor <- subset(cost, select = -c(2,5,6))
corrplot(cor(cost_cor[, -3]), method = "color", type = "lower", number.
cex = 0.7, order = "hclust",addCoef.col = "white", title = "correlation
 of inputs of the medical cost")
```



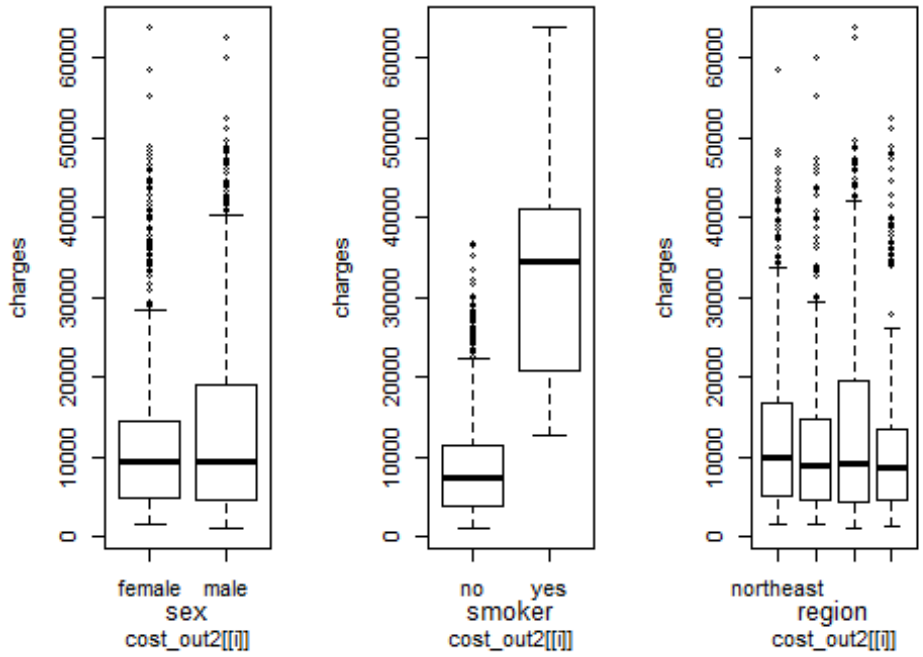
To plot the correlation graph, I use the r package "corrplot". First, I drop the categorical variables(sex, smoker and region), then I plot the correlation between the variables bmi, age and charges. The correlation is between -1 to 1, so the darker the color, the stronger the correlation. We can see that there are not such a strong correlation between those variables.

Graph of outliers

```
#boxplot of the non-categorical variables
cost_out1 <- subset(cost, select = -c(2,5,6,7))
boxplot(cost_out1)
```




```
#boxplot of the categorical variables
cost_out2 <- subset(cost, select = -c(1,3,4,7))
outlier1 <- par(mfrow = c(1,3))
for (i in 1:3){
  plot(charges~cost_out2[[i]], data = cost)
  mtext(names(cost_out2)[i], cex = 0.8, side = 1, line = 2)
}
```




```
par(outlier1)
```

For the non-categorical variables, bmi shows some outliers above maximum. For the categorical variables, I compare those variables to the output "charges", and the boxplot shows that for both genders, there are some outliers above averages; for the non-smokers, there are some outliers above averages, and for all of the four regions, there are some outliers above averages.

Histogram of each variables

```
#histogram of age
library(RColorBrewer)
col = brewer.pal(6,"Reds")
col
```

```
## [1] "#FEE5D9" "#FCBBA1" "#FC9272" "#FB6A4A" "#DE2D26" "#A50F15"
```

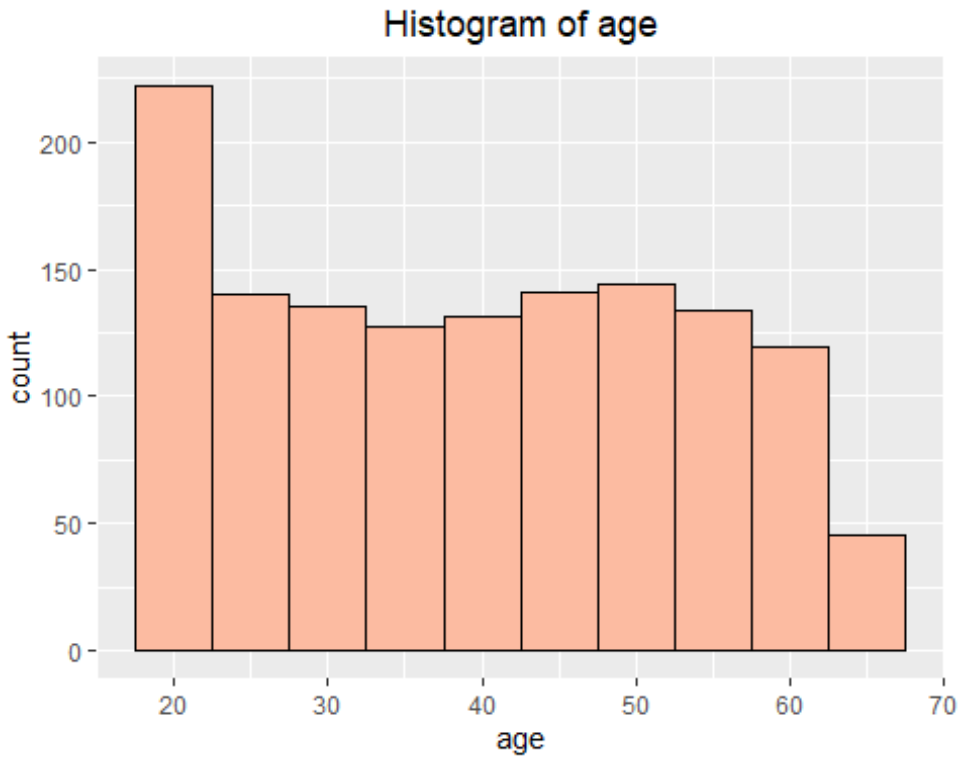
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
##
## Attaching package: 'ggplot2'
```

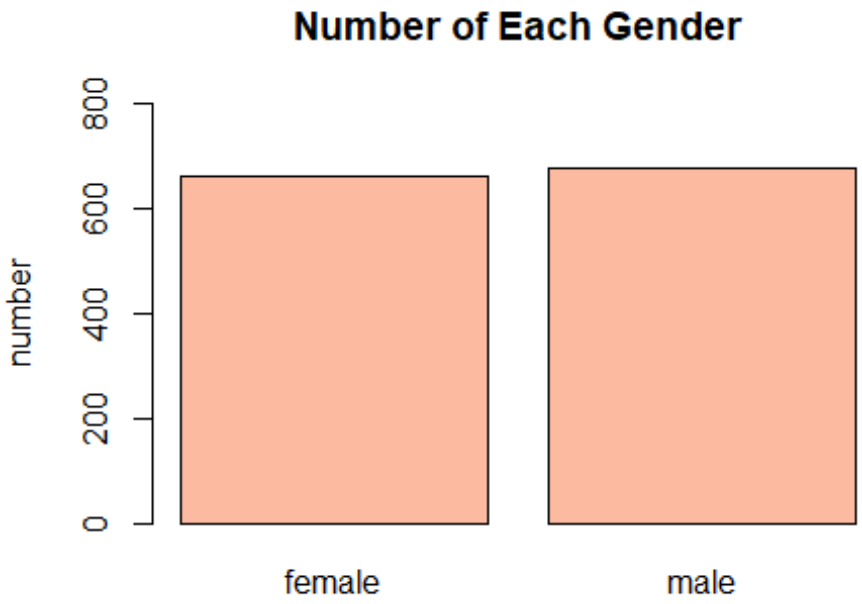
```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(data = cost, aes(x = age))+geom_histogram(aes(color = I("black"),
 fill = I("#FCBBA1")), binwidth = 5)+ggtitle("Histogram of age")+theme
(plot.title = element_text(hjust = 0.5))
```



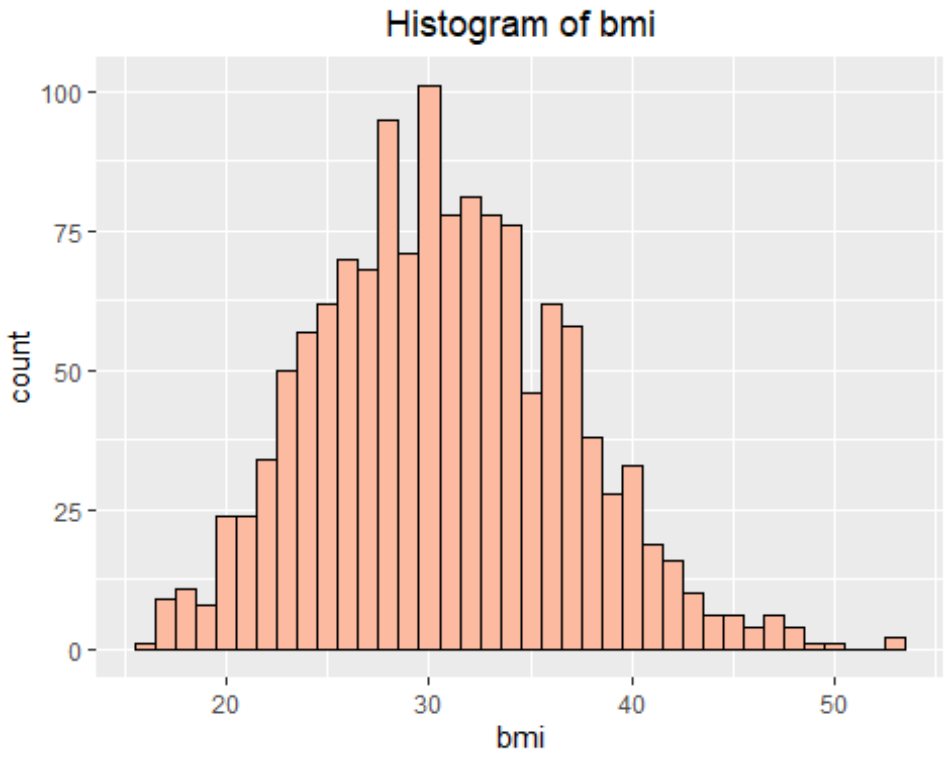
From the histogram above, we can see that the data set contains a greater number of people who is under 20, and the number of people for each other ages are about the same.

```
#histogram of sex
plot(cost$sex, main = "Number of Each Gender", ylab = "number", ylim =
c(0,800), col = "#FCBBA1")
```



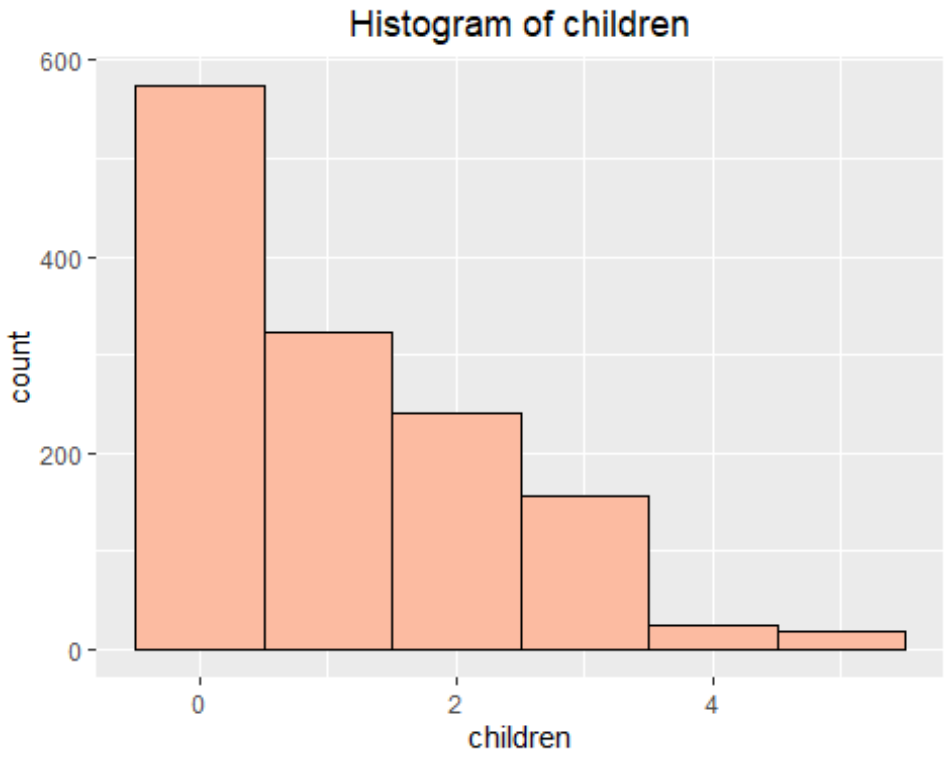
As the figure shows above, both female and male are about 600 people in the data set.

```
#histogram of bmi
ggplot(data = cost, aes(x = bmi))+geom_histogram(aes(color = I("black"),
 fill = I("#FCBBA1")), binwidth = 1)+ggtitle("Histogram of bmi")+theme
(plot.title = element_text(hjust = 0.5))
```



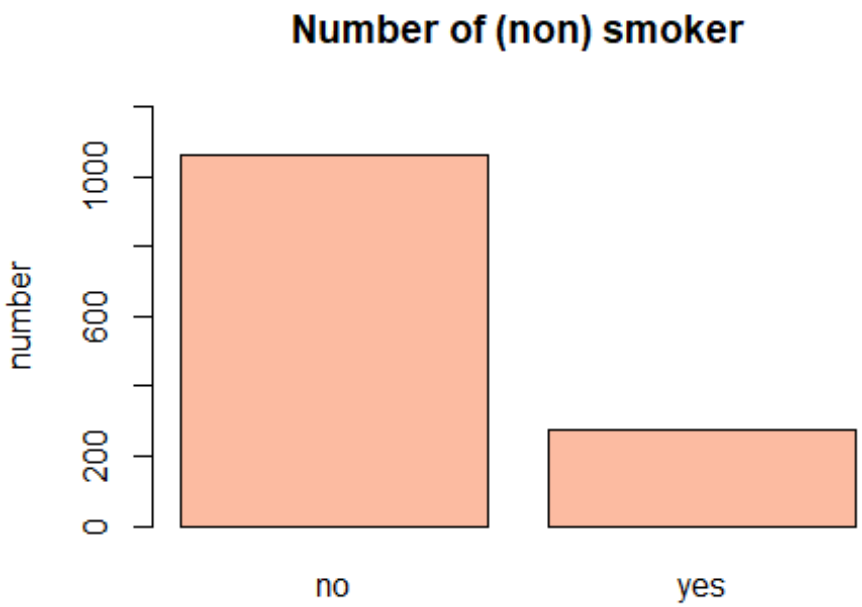
From the histogram above, we can see that the bmi of those people is normally distributed.

```
#histogram of children
ggplot(data = cost, aes(x = children))+geom_histogram(aes(color = I("bl
ack"), fill = I("#FCBBA1")), binwidth = 1, border = "black")+ggtitle("H
istogram of children")+theme(plot.title = element_text(hjust = 0.5))

## Warning: Ignoring unknown parameters: border
```



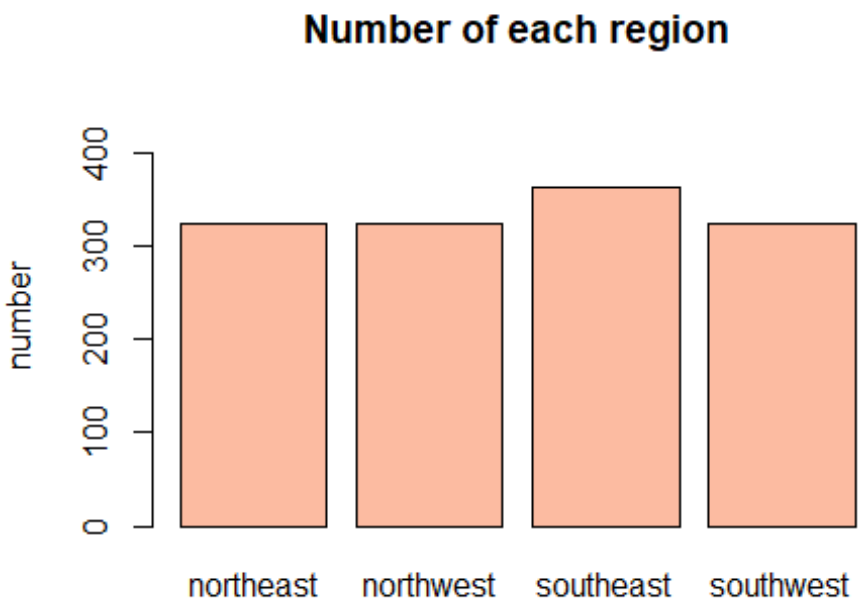
We can see that as the number of children increases, the number of people decreases. The most of people does not have a child (about 500).

```
#histogram of smoker
plot(cost$smoker, main = "Number of (non) smoker", ylab = "number", yli
m = c(0,1200), col = "#FCBBA1")
```



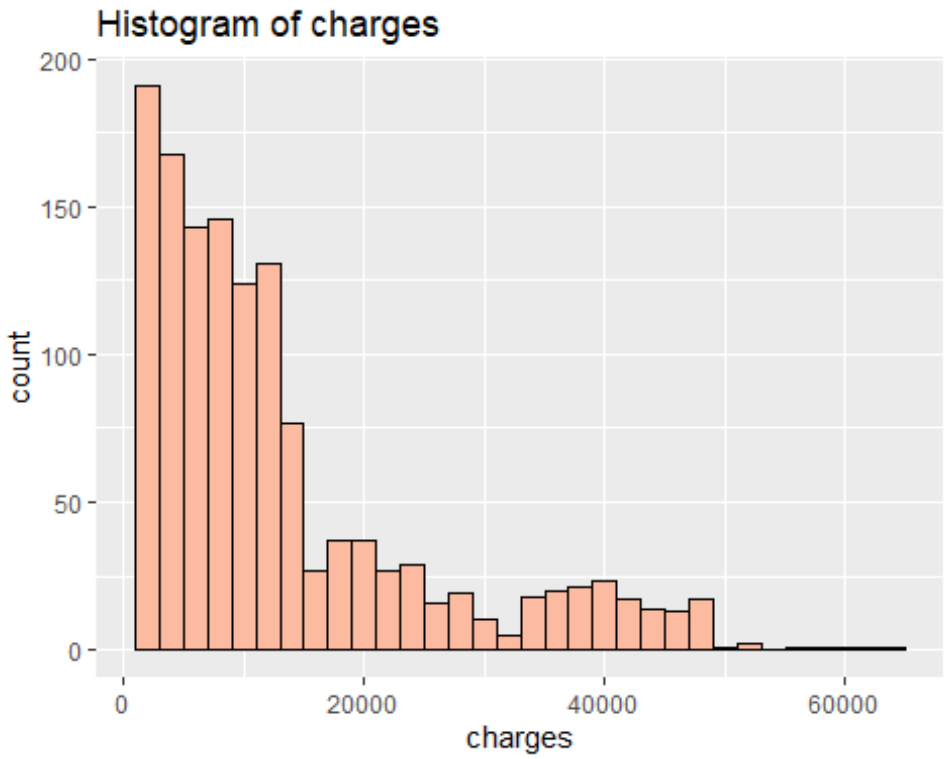
There are about 1100 non-smokers and 300 smokers are contained in the dataset.

```
#histogram of regions
plot(cost$region, main = "Number of each region", ylab = "number", ylim
 = c(0,450), col = "#FCBBA1")
```



There are about 380 people in the region of southeast, which is the highest in the dataset. The patients in other three regions are about the same, which is about 320 patients.

```
#histogram of charges
ggplot(data = cost, aes(x = charges))+geom_histogram(aes(color = I("bla
ck"), fill = I("#FCBBA1")), binwidth = 2000)+ggtitle("Histogram of char
ges")
```



From the figure above, as the charges are increasing, the number of patients are decreasing, there are about 180 patients who has a medical cost lower than 2000 dollars.

Analysis

Linear Regression Model

```
lm1 <- lm(charges~.,data = cost)
summary(lm1)

##
## Call:
## lm(formula = charges ~ ., data = cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
```

```
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
mse <- mean(lm1$residuals^2)
mse
```

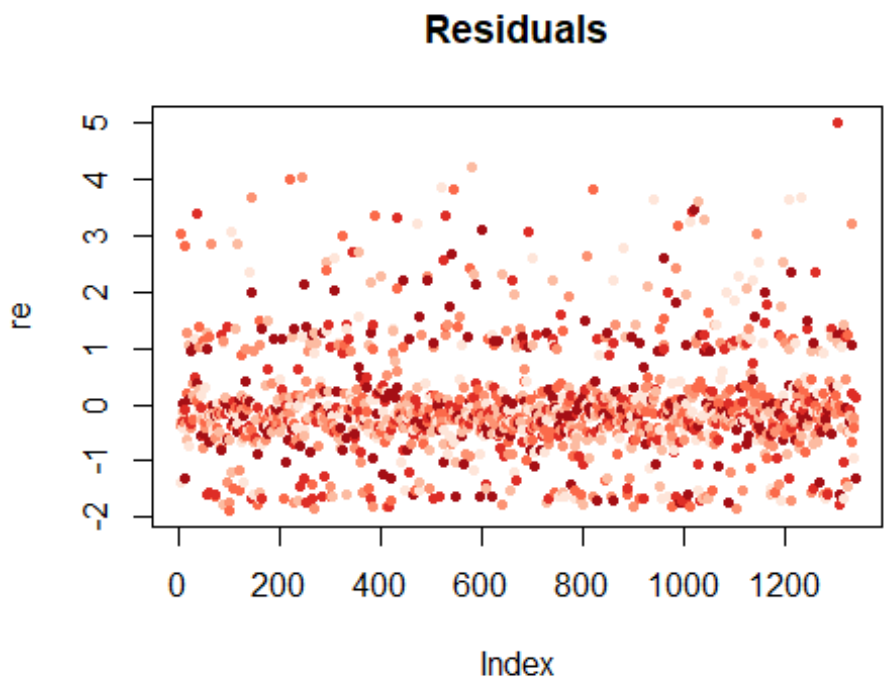
```
## [1] 36501893
```

That is the full model of linear regression, in the summary above, we can see that this model can explain 75.09% of the observations. In the summary, age, bmi, children, smoker(yes), and region(southeast and southwest) are more important than other predictors. This model is a good model, but the MSE of the model is very large, so we still want a better model.

In order to increase the accuracy of this model, we want to drop the influencial points.

Data Cleaning

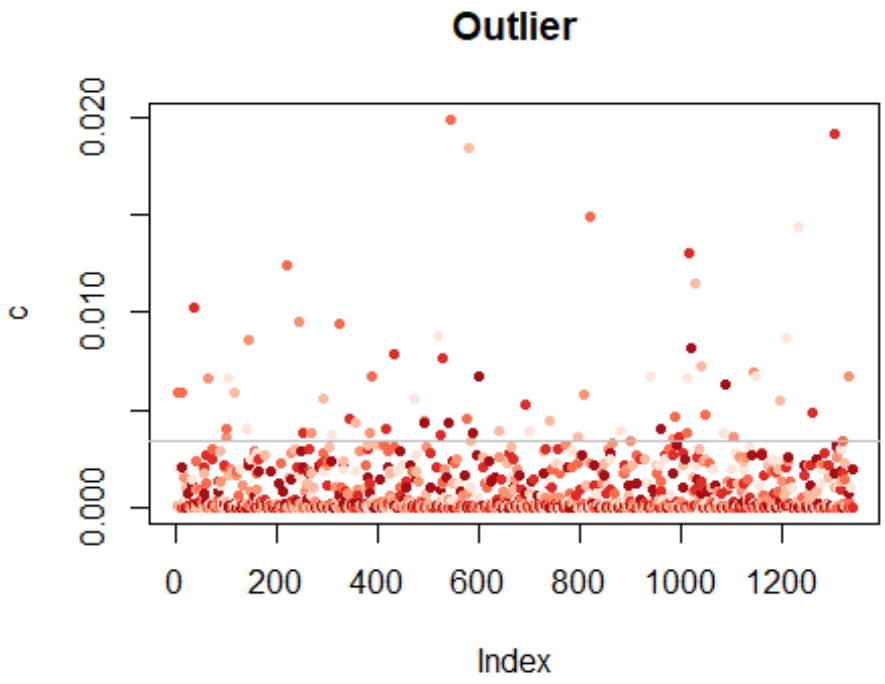
```
re = rstudent(lm1)
plot(re, pch =20, cex = 1, main= "Residuals", col = brewer.pal(6,"Reds
"))
```




```
c = cooks.distance(lm1)
h = head(cost[c > 4 * mean(c, na.rm=T), ])
h
```

```
##     age    sex    bmi children smoker    region  charges
## 4    33   male 22.705        0     no northwest 21984.47
## 10   60 female 25.840        0     no northwest 28923.14
## 35   28   male 36.400        1    yes southwest 51194.56
## 63   64   male 24.700        1     no northwest 30166.62
## 99   56   male 19.950        0    yes northeast 22412.65
## 100  38   male 19.300        0    yes southwest 15820.70
```

```
re = rstudent(lm1)
plot(c, pch = 20, cex = 1, main = "Outlier", col = brewer.pal(6,"Reds"))
abline(h = 4*mean(c, na.rm = T), col = "Grey")
```




```
cost <- cost[c <= 4 * mean(c, na.rm=T), ]
str(cost)

## 'data.frame':    1270 obs. of  7 variables:
##  $ age     : int  19 18 28 32 31 46 37 37 25 62 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 1 1 1 2 2 1
 ...
##  $ bmi     : num  27.9 33.8 33 28.9 25.7 ...
##  $ children: int  0 1 3 0 0 1 3 2 0 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 2 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2
3 3 2 1 1 3 ...
##  $ charges : num  16885 1726 4449 3867 3757 ...
```

From the plot above, we can see that there is a significantly decreasing of the influencial points. So from now on, I will use the clean data to do the following prediction.

New Linear Regression Model

```
lm2 <- lm(charges~.,data = cost)
summary(lm2)

##
## Call:
## lm(formula = charges ~ ., data = cost)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -10932.9  -1970.9   -342.2   1626.0  14959.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11787.191    792.622 -14.871  < 2e-16 ***
## age                255.145      9.424  27.073  < 2e-16 ***
## sexmale            102.992    263.505   0.391   0.6960
## bmi                308.697     23.055  13.389  < 2e-16 ***
## children           453.292    109.016   4.158 3.43e-05 ***
## smokeryes        24289.036    331.047  73.370  < 2e-16 ***
## regionnorthwest   -701.909    378.260  -1.856   0.0637 .
## regionsoutheast   -923.823    379.797  -2.432   0.0151 *
## regionsouthwest   -842.736    377.998  -2.229   0.0260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4671 on 1261 degrees of freedom
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.8354
## F-statistic:   806 on 8 and 1261 DF,  p-value: < 2.2e-16
```

```
mse <- mean(lm2$residuals^2)
mse
```

```
## [1] 21664191
```

Comparing the two linear regression model above, the value of R-squared increased from 0.7509 to 0.8381, which means there are 83.81% of the variables can be well explained by the linear model. Also, the MSE decreased significantly. The predictor region(wouthwest) is no longer a significant predictor.

Then, we can use the significant predictor to predict the model.

```
lm3 <- lm(charges~age+bmi+children+smoker+region,data = cost)
summary(lm3)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = cost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10889.0  -1956.4   -331.4   1650.1  15010.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11747.130    785.703 -14.951  < 2e-16 ***
## age                255.037      9.417  27.082  < 2e-16 ***
## bmi                309.172     23.015  13.433  < 2e-16 ***
## children           454.448    108.939   4.172 3.23e-05 ***
```

```
## smokeryes         24298.743    330.003  73.632  < 2e-16 ***
## regionnorthwest    -703.270    378.117  -1.860   0.0631 .
## regionsoutheast    -925.767    379.637  -2.439   0.0149 *
## regionsouthwest    -844.258    377.851  -2.234   0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4669 on 1262 degrees of freedom
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.8355
## F-statistic: 921.7 on 7 and 1262 DF,  p-value: < 2.2e-16

mse <- mean(lm3$residuals^2)
mse

## [1] 21666815
```

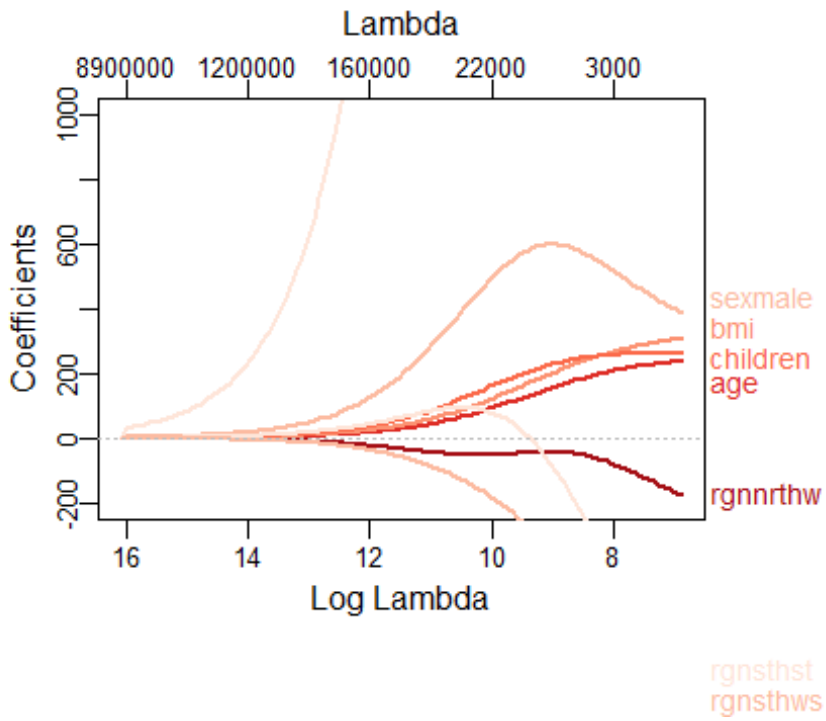
Split Data

```
dim(cost)

## [1] 1270    7

set.seed(1)
cost$charges <- as.numeric(cost$charges)
trainindex <- sample(1:635)
train <- cost[trainindex,]
test <- cost[-trainindex,]
```

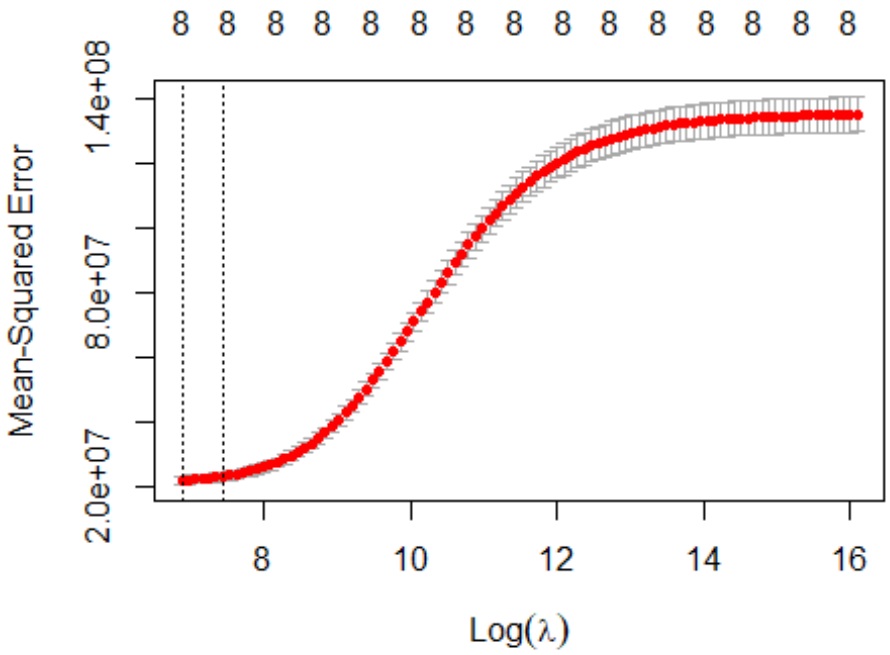
Ridge Regression Model

```
library(plotrix)

##
## Attaching package: 'plotrix'

## The following object is masked from 'package:psych':
##
##     rescale

library(glmnet)

## Warning: package 'glmnet' was built under R version 3.6.3

## Loading required package: Matrix

## Loaded glmnet 3.0-2

library(Matrix)
library(plotmo)

## Warning: package 'plotmo' was built under R version 3.6.3

## Loading required package: Formula

## Loading required package: TeachingDemos
```

```
trainm <- model.matrix(charges~.,data = train)[,-1]
testm <-  model.matrix(charges~.,data = test)

rm <- glmnet(trainm, train$charges, alpha = 0)

plot_glmnet(rm, col = brewer.pal(6,"Reds"), ylim = c(-200,1000), lwd =
2)
```




```
cvrm <- cv.glmnet(trainm, train$charges, alpha = 0)
bestlam <- cvrm$lambda.min
bestlam

## [1] 973.8359

plot(cvrm, col = brewer.pal(6,"Reds"))
```




```
predict(rm, type = "coefficients", s = bestlam)

## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)     -10996.1753
## age                238.7658
## sexmale            389.4627
## bmi                308.6549
## children           264.1117
## smokeryes        22672.8437
## regionnorthwest    -174.9563
## regionsoutheast    -735.9510
## regionsouthwest    -791.5467

ridge.pred <- predict(rm, s = bestlam, newx = testm, type = "coefficien
ts")
```

We can see that the value of lambda that results in the smallest cross-validation error is 983.349. But none of the coefficients are zero, because redge regression does not perform variable selection.

Data Sselection

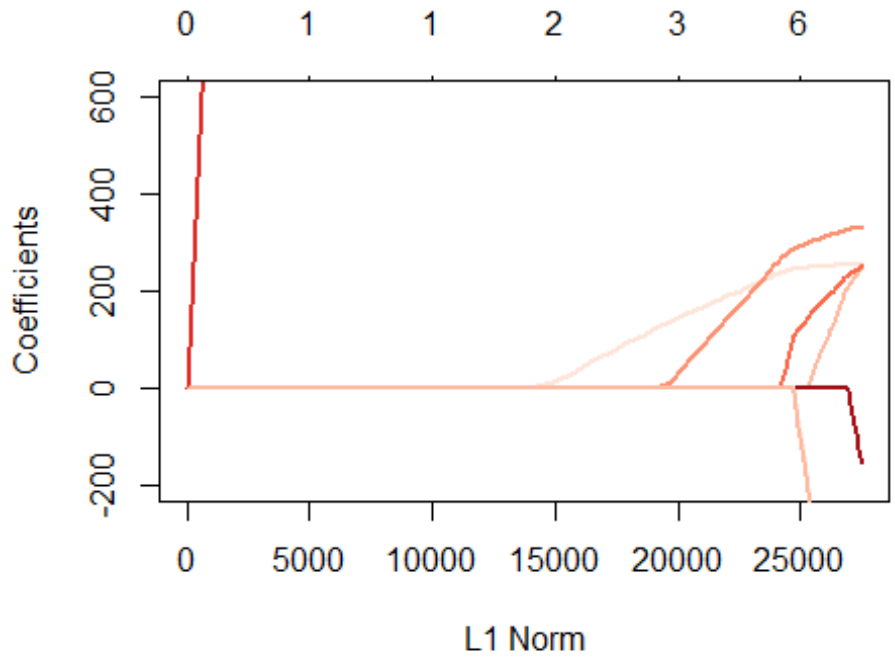
Lasso Model

```
lasso1 <- glmnet(trainm, train$charges, alpha = 1)
lasso1
```

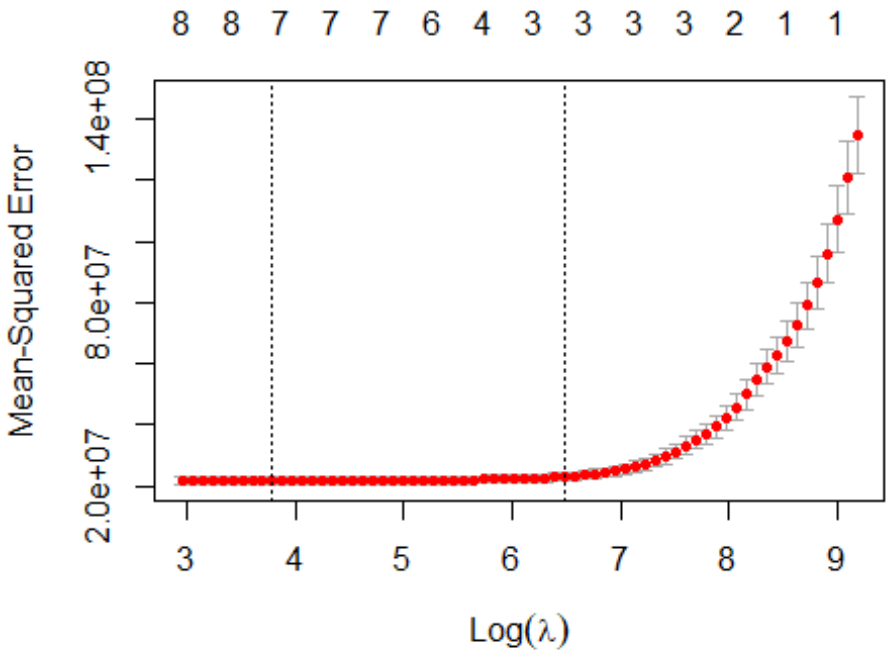
```
##
## Call:  glmnet(x = trainm, y = train$charges, alpha = 1)
##
##    Df   %Dev Lambda
## 1   0 0.0000 9738.0
## 2   1 0.1189 8873.0
## 3   1 0.2177 8085.0
## 4   1 0.2997 7367.0
## 5   1 0.3677 6712.0
## 6   1 0.4243 6116.0
## 7   1 0.4712 5573.0
## 8   1 0.5101 5078.0
## 9   1 0.5425 4627.0
## 10  1 0.5693 4216.0
## 11  2 0.6012 3841.0
## 12  2 0.6378 3500.0
## 13  2 0.6683 3189.0
## 14  2 0.6935 2906.0
## 15  2 0.7145 2647.0
## 16  2 0.7319 2412.0
## 17  2 0.7464 2198.0
## 18  3 0.7605 2003.0
## 19  3 0.7745 1825.0
## 20  3 0.7861 1663.0
## 21  3 0.7957 1515.0
## 22  3 0.8037 1380.0
## 23  3 0.8103 1258.0
## 24  3 0.8158 1146.0
## 25  3 0.8204 1044.0
## 26  3 0.8242  951.4
## 27  3 0.8274  866.9
## 28  3 0.8300  789.9
## 29  3 0.8322  719.7
## 30  3 0.8340  655.8
## 31  3 0.8355  597.5
## 32  3 0.8367  544.5
## 33  3 0.8377  496.1
## 34  3 0.8386  452.0
## 35  3 0.8393  411.9
## 36  3 0.8399  375.3
## 37  4 0.8405  341.9
## 38  4 0.8410  311.6
## 39  4 0.8414  283.9
## 40  4 0.8418  258.7
## 41  4 0.8421  235.7
## 42  5 0.8423  214.7
## 43  6 0.8427  195.7
## 44  6 0.8431  178.3
## 45  7 0.8434  162.4
## 46  7 0.8437  148.0
```

```
## 47   7 0.8439  134.9
## 48   7 0.8441  122.9
## 49   7 0.8443  112.0
## 50   7 0.8444  102.0
## 51   7 0.8445   93.0
## 52   7 0.8446   84.7
## 53   7 0.8447   77.2
## 54   7 0.8447   70.3
## 55   7 0.8448   64.1
## 56   7 0.8448   58.4
## 57   7 0.8449   53.2
## 58   7 0.8449   48.5
## 59   7 0.8449   44.2
## 60   8 0.8449   40.2
## 61   8 0.8450   36.7
## 62   8 0.8450   33.4
## 63   8 0.8450   30.4
## 64   8 0.8450   27.7
## 65   8 0.8450   25.3
## 66   8 0.8451   23.0
## 67   8 0.8451   21.0
## 68   8 0.8451   19.1
```

```
plot(lasso1, col = brewer.pal(6,"Reds"), ylim = c(-200, 600), lwd = 2)
```




```
cv.lasso  <- cv.glmnet(trainm, train$charges, alpha=1)
plot(cv.lasso)
```




```
bestlambda <- cv.lasso$lambda.min
bestlambda

## [1] 44.16227

preco <- predict(lasso1, newx = testm, s = bestlambda, type = "coeffici
ents")
preco

## 9 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)     -12430.2368
## age                254.2563
## sexmale            203.0041
## bmi                325.3836
## children           230.4965
## smokeryes        24480.0434
## regionnorthwest       .
## regionsoutheast   -695.1860
## regionsouthwest   -680.2642
```

We can see that the value of lambda that results in the smallest cross-validation error is 44.59368. In the Lasso model, only the region northwest is not significant.

Decision Tree

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.3
```

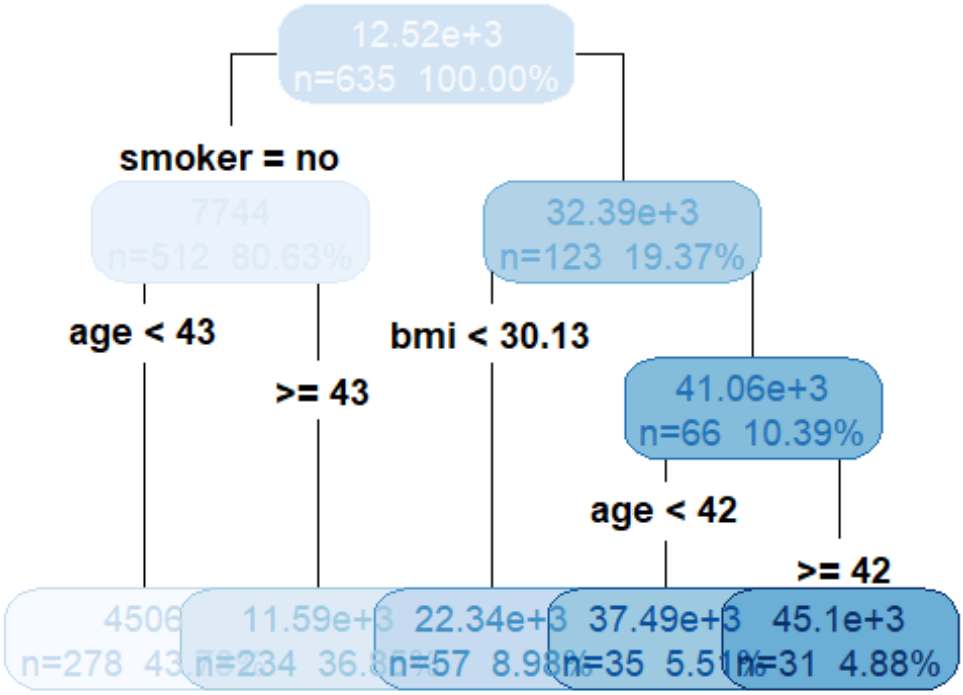
```
set.seed(1)

dt <- rpart(charges~., data = train)
summary(dt)
```

```
## Call:
## rpart(formula = charges ~ ., data = train)
##   n= 635
##
##           CP nsplit  rel error     xerror       xstd
## 1 0.70057850      0 1.00000000 1.00286859 0.07656396
## 2 0.12479769      1 0.29942150 0.30123990 0.01898833
## 3 0.07421936      2 0.17462381 0.17656046 0.01186190
## 4 0.01109755      3 0.10040446 0.10516570 0.01033465
## 5 0.01000000      4 0.08930691 0.09831858 0.01045629
##
## Variable importance
##   smoker      bmi      age      sex   region children
##       73       14       10        2        1        1
##
## Node number 1: 635 observations,    complexity param=0.7005785
##   mean=12517.51, MSE=1.353676e+08
##   left son=2 (512 obs) right son=3 (123 obs)
##   Primary splits:
##       smoker   splits as  LR,          improve=0.700578500, (0 missi
ng)
##       age      < 41.5    to the left,  improve=0.095237100, (0 missi
ng)
##       bmi      < 31.145  to the left,  improve=0.048928760, (0 missi
ng)
##       sex      splits as  LR,          improve=0.007606829, (0 missi
ng)
##       children < 1.5     to the left,  improve=0.007280629, (0 missi
ng)
##
## Node number 2: 512 observations,    complexity param=0.07421936
##   mean=7744.376, MSE=2.344717e+07
##   left son=4 (278 obs) right son=5 (234 obs)
##   Primary splits:
##       age      < 42.5    to the left,  improve=0.531429000, (0 missi
ng)
##       bmi      < 35.88   to the left,  improve=0.033858100, (0 missi
ng)
##       children < 1.5     to the left,  improve=0.017201710, (0 missi
ng)
##       sex      splits as  RL,          improve=0.008615297, (0 missi
ng)
##       region   splits as  RLLL,        improve=0.003588683, (0 missi
```

```
ng)
##   Surrogate splits:
##       bmi      < 35.6325 to the left,  agree=0.574, adj=0.068, (0 sp
lit)
##       children < 3.5     to the left,  agree=0.547, adj=0.009, (0 sp
lit)
##
## Node number 3: 123 observations,    complexity param=0.1247977
##   mean=32386.15, MSE=1.116492e+08
##   left son=6 (57 obs) right son=7 (66 obs)
##   Primary splits:
##       bmi      < 30.125  to the left,  improve=0.781149800, (0 missi
ng)
##       age      < 53.5    to the left,  improve=0.155141300, (0 missi
ng)
##       sex      splits as  LR,          improve=0.030990690, (0 missi
ng)
##       children < 1.5     to the left,  improve=0.011799300, (0 missi
ng)
##       region   splits as  RLRR,        improve=0.005199331, (0 missi
ng)
##   Surrogate splits:
##       sex      splits as  LR,          agree=0.602, adj=0.140, (0 sp
lit)
##       age      < 21.5    to the left,  agree=0.569, adj=0.070, (0 sp
lit)
##       region   splits as  RLRR,        agree=0.569, adj=0.070, (0 sp
lit)
##       children < 2.5     to the right, agree=0.553, adj=0.035, (0 sp
lit)
##
## Node number 4: 278 observations
##   mean=4505.805, MSE=9699913
##
## Node number 5: 234 observations
##   mean=11591.91, MSE=1.251537e+07
##
## Node number 6: 57 observations
##   mean=22337, MSE=2.845734e+07
##
## Node number 7: 66 observations,    complexity param=0.01109755
##   mean=41064.96, MSE=2.096014e+07
##   left son=14 (35 obs) right son=15 (31 obs)
##   Primary splits:
##       age      < 41.5    to the left,  improve=0.68956860, (0 missin
g)
##       bmi      < 37.66   to the left,  improve=0.13518720, (0 missin
g)
##       children < 0.5     to the left,  improve=0.07371368, (0 missin
g)
```

```
##       sex      splits as  RL,          improve=0.01594999, (0 missin
g)
##       region   splits as  RRLL,        improve=0.01358952, (0 missin
g)
##   Surrogate splits:
##       bmi      < 35.73   to the left,  agree=0.591, adj=0.129, (0 sp
lit)
##       region   splits as  RRLL,        agree=0.591, adj=0.129, (0 sp
lit)
##       children < 0.5     to the left,  agree=0.561, adj=0.065, (0 sp
lit)
##
## Node number 14: 35 observations
##   mean=37487.03, MSE=5821110
##
## Node number 15: 31 observations
##   mean=45104.57, MSE=7280724
```

```
rpart.plot(dt, digits = 4, fallen.leaves = TRUE, type = 4, extra = 101,
tweak = 1.6, col = brewer.pal(9,"Blues"))
```




```
dt.pre <- predict(dt, data = test)
summary(dt.pre)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4506    4506   11592   12518   11592   45105
```

```
summary(test$charges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4568    8689   12338   14233   49578
```

```
head(dt.pre, n=10)
```

```
##       137       538       497       315       284       197       32
4
##  4505.805 11591.908  4505.805 37487.026 11591.908  4505.805 11591.90
8
##       633       292       522
##  4505.805  4505.805  4505.805
```

```
MSE <- mean((train$charges-dt.pre)^2)
MSE
```

```
## [1] 12089262
```

The MSE is large.

Random Forest

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.3
```

```
## randomForest 4.6-14
```

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

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

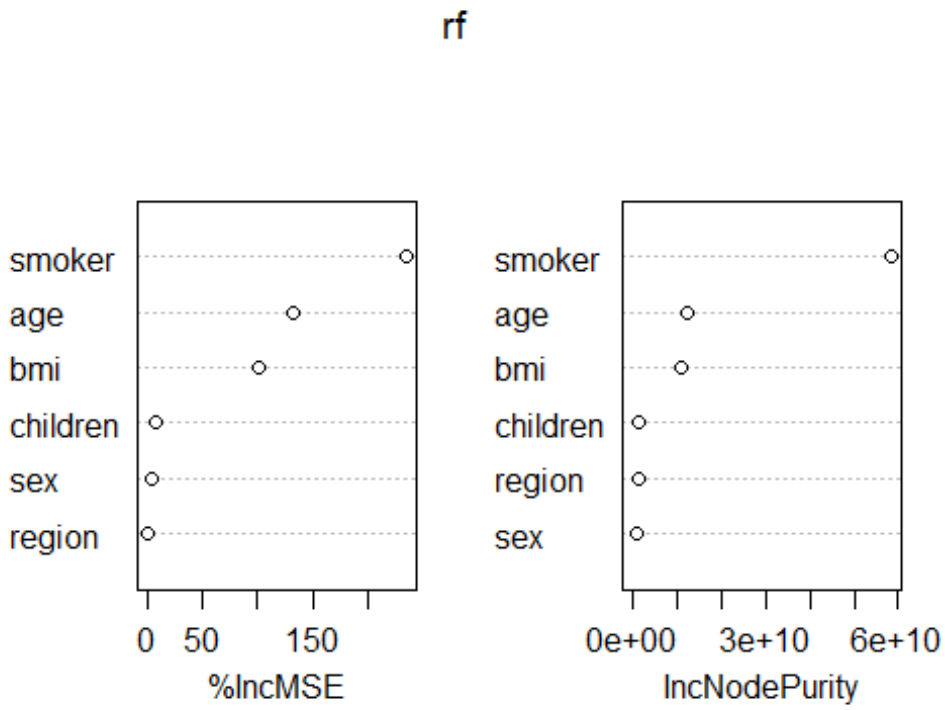
```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:psych':
##
##     outlier
```

```
rf <- randomForest(charges~., data = train, ntree = 1000, mtry = sqrt(1
1), replace = TRUE, importance = TRUE)
rf
```

```
##
## Call:
##  randomForest(formula = charges ~ ., data = train, ntree = 1000,
  mtry = sqrt(11), replace = TRUE, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
```

```
##           Mean of squared residuals: 8801756
##                     % Var explained: 93.5

rf.pre = predict(rf, test)


importance(rf)

##             %IncMSE IncNodePurity
## age      132.886863   12276996580
## sex        3.455148     588800939
## bmi      101.515763   10986463363
## children   8.384882     987668307
## smoker   234.917443   58516364324
## region     1.128408     969674723

varImpPlot(rf)
```



From the figure above, we can see that smoker, age, bmi are more important than children, sex and region.

Forward Selection

```
forward <- step(glm(charges~., data = cost), direction = "forward", tes
t = "F")

## Start:  AIC=25075.89
## charges ~ age + sex + bmi + children + smoker + region
```

```
summary(forward)

##
## Call:
## glm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = cost)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10932.9   -1970.9    -342.2    1626.0   14959.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11787.191    792.622 -14.871  < 2e-16 ***
## age                255.145      9.424  27.073  < 2e-16 ***
## sexmale            102.992    263.505   0.391   0.6960
## bmi                308.697     23.055  13.389  < 2e-16 ***
## children           453.292    109.016   4.158 3.43e-05 ***
## smokeryes        24289.036    331.047  73.370  < 2e-16 ***
## regionnorthwest   -701.909    378.260  -1.856   0.0637 .
## regionsoutheast   -923.823    379.797  -2.432   0.0151 *
## regionsouthwest   -842.736    377.998  -2.229   0.0260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 21818812)
##
##     Null deviance: 1.6819e+11  on 1269  degrees of freedom
## Residual deviance: 2.7514e+10  on 1261  degrees of freedom
## AIC: 25076
##
## Number of Fisher Scoring iterations: 2
```

The AIC of forward selection is 22537, and the predictor age, bmi, children, smoker are more significant than others.

Backward Selection

```
backward <- step(glm(charges~., data = cost), direction = "backward", t
est = "F")

## Start:  AIC=25075.89
## charges ~ age + sex + bmi + children + smoker + region
##
##            Df   Deviance   AIC    F value    Pr(>F)
## - sex       1 2.7517e+10 25074    0.1528   0.69597
## <none>        2.7514e+10 25076
## - region    3 2.7674e+10 25077    2.4504   0.06202 .
## - children  1 2.7891e+10 25091   17.2892 3.427e-05 ***
## - bmi       1 3.1425e+10 25243  179.2777 < 2.2e-16 ***
## - age       1 4.3506e+10 25656  732.9591 < 2.2e-16 ***
```

```
## - smoker    1 1.4497e+11 27184 5383.2150 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=25074.05
## charges ~ age + bmi + children + smoker + region
##
##            Df   Deviance   AIC   F value    Pr(>F)
## <none>         2.7517e+10 25074
## - region    3 2.7678e+10 25076    2.4622   0.06105 .
## - children  1 2.7896e+10 25089   17.4020 3.232e-05 ***
## - bmi       1 3.1451e+10 25242  180.4515 < 2.2e-16 ***
## - age       1 4.3509e+10 25654  733.4601 < 2.2e-16 ***
## - smoker    1 1.4573e+11 27189 5421.6522 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(backward)
```

```
##
## Call:
## glm(formula = charges ~ age + bmi + children + smoker + region,
##     data = cost)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10889.0   -1956.4    -331.4    1650.1   15010.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11747.130    785.703 -14.951  < 2e-16 ***
## age                255.037      9.417  27.082  < 2e-16 ***
## bmi                309.172     23.015  13.433  < 2e-16 ***
## children           454.448    108.939   4.172 3.23e-05 ***
## smokeryes        24298.743    330.003  73.632  < 2e-16 ***
## regionnorthwest   -703.270    378.117  -1.860   0.0631 .
## regionsoutheast   -925.767    379.637  -2.439   0.0149 *
## regionsouthwest   -844.258    377.851  -2.234   0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 21804164)
##
##     Null deviance: 1.6819e+11  on 1269  degrees of freedom
## Residual deviance: 2.7517e+10  on 1262  degrees of freedom
## AIC: 25074
##
## Number of Fisher Scoring iterations: 2
```

THe AIC is 22535, which is a little bit smaller than forward selection. The predictor age, bmi, children, smoker are significant.

SOme Comparision

Above all, we can see the variable smoker, age, bmi are alwaye significant in all of the methods. The variable “children” is significant in some of the methods, but not significant in others. SO we want to test it in the regression model.

```
# with variable children

lm4 <- lm(charges~age+bmi+children+smoker, data = train)
summary(lm4)

##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11068.6  -2010.9   -320.2   1656.5  15772.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12696.56    1073.66 -11.826   <2e-16 ***
## age            257.50      12.84  20.048   <2e-16 ***
## bmi            319.39      31.67  10.086   <2e-16 ***
## children       267.26     152.64   1.751   0.0804 .
## smokeryes    24595.93     463.99  53.009   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4620 on 630 degrees of freedom
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8426
## F-statistic: 849.2 on 4 and 630 DF,  p-value: < 2.2e-16

# without variable children

lm5 <- lm(charges~age+bmi+smoker, data = train)
summary(lm5)

##
## Call:
## lm(formula = charges ~ age + bmi + smoker, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11345.1  -2077.4   -230.9   1643.9  15773.3
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12544.25    1071.88  -11.70   <2e-16 ***
## age            258.96      12.84   20.17   <2e-16 ***
## bmi            321.82      31.69   10.16   <2e-16 ***
## smokeryes    24598.71     464.75   52.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4628 on 631 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.842
## F-statistic:  1128 on 3 and 631 DF,  p-value: < 2.2e-16
```

By comparing the two summary above, we can see that with the variable children, the adjusted R-squared is 0.8491 which is a little bit higher than the adjusted R-squared of the prediction without children (0.8481). So, as we want a higher accuracy, we choose the variable children as a predictor.

Conclusion

In conclustion, I choose the linear regression model with predictor age, bmi, smoker and children to predict the medical cost for a patient.

The final regression model is:

charges = -12561.2 + 259.12 * age + 310.88 * bmi + 348.41 * children + 25073.53 * smoker

In this project, I used some r package as following:

psych corrplor RColorBrewer ggplot2 plotrix glmnet Matrix plotmo rpart rpart.plot randomForest

From this project, I learned how to describe a dataset, how to use the color chart to optimazies my figure, how to develop my code to optimazies my figure. I also learned that for different method, there are different results, we need to test and get our final answers.

Reference

https://www.kaggle.com/janiobachmann/patient-charges-clustering-and-regression https://www.kaggle.com/datasets
https://rdrr.io/cran/lmridge/man/summary.lmridge.html
http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram https://www.kaggle.com/mirichoi0218/insurance