Final Project : Medical Insurance Cost Prediction

Bhavana Penmatsa, Kiran Varma Kolukuluri, Roshini Padmanabha,
Sharon Sowmya Tolety

Group : BA01-6 (with Roshini, Bhavana from MIS01)

Management Information Systems Department, San Diego State University

BA 649-01: Business Analytics

Prof. Huiyu Qian

05/06/2024

Contributions of Each Member:

Bhavana Penmatsa - Regression : Three Subset Selections , Ridge regression, Classification - KNN and Report Writing

Kiran Varma Kolukuluri - Data Preparation and EDA - Univariate and Bivariate Analysis, Classification - Linear Discriminant Analysis, Report Writing

Roshini Padmanabha - Introduction, Linear regression Models, Residual Diagnosis, Lasso Regression, Conclusion and Recommendations, References, Report Formatting

Sharon Sowmya Tolety - Bivariate Analysis, Linear regression fits, Regression - Cross Validation and modifications, Classification - Logistic Regression, Conclusion

Part I Introduction

Background of the business/research problem

The business/research problem revolves around understanding the factors influencing medical expenses in the context of health insurance. With rising healthcare costs, insurance companies need to accurately predict medical expenses to set appropriate premiums and assess risks effectively. By analyzing factors such as age, BMI, smoking status, and region, the goal is to identify the primary drivers of medical expenses and improve decision-making processes in pricing and risk assessment.

Regression problem

What is the Business/Research Question?

The business/research question aims to predict medical expenses for policyholders based on various factors such as age, sex, BMI, smoking status, number of children, and region.

How accurately can machine learning models forecast medical expenses using demographic and lifestyle variables?

What is the DV?

The dependent variable (DV) in the regression problem is the "Charges" column, representing the medical expenses incurred by policyholders.

What are the Potential IVs?

Potential independent variables (IVs) include age, sex, BMI, smoking status, number of children, and region.

Why is it Important in Business/Research Practice?

Accurate prediction of medical expenses helps insurance companies set premiums that reflect the expected costs associated with covering policyholders.

Understanding the factors influencing medical expenses allows insurers to tailor insurance plans, manage risks effectively, and optimize profitability.

Classification problem

What is the Business/Research Question?

The business/research question aims to classify policyholders into high and low medical expense groups based on demographic and lifestyle factors.

Can machine learning models accurately classify policyholders as high or low medical expense individuals using available data?

What is the DV?

The dependent variable (DV) in the classification problem is the "high_charge" dummy variable, indicating whether a policyholder belongs to the high medical expense group.

What are the Potential IVs?

Potential independent variables (IVs) include age, sex, BMI, smoking status, number of children, and region.

Why is it Important in Business/Research Practice?

Classifying policyholders into high and low medical expense groups helps insurers identify individuals who may require more comprehensive coverage or interventions to manage their health.

Effective classification enables insurers to customize insurance plans, target interventions, and allocate resources efficiently, ultimately improving customer satisfaction and reducing overall healthcare costs.

Based on our dataset and research questions, the **hypotheses** may include:

Older individuals may have higher medical expenses compared to younger ones.

Smokers are likely to incur higher medical expenses than non-smokers.

Individuals with higher BMIs may face increased medical costs due to associated health risks.

Region may influence medical expenses, with certain areas experiencing higher healthcare costs than others.

Individuals with more children may have higher medical expenses due to increased healthcare needs for their family.

Part II Data Preparation and EDA

Let us now look at the Data source and summary of the data set we used:

Link of the data source:
https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction?resource=download

Number of Observations: The dataset comprises 2772 observations. Each observation in the dataset, represents an individual policyholder or insurance applicant details.

Variables of the dataset are as follows:

Sex: Gender of the policyholder.

Type: Categorical (Qualitative)

Age: Age of the policyholder.

Type: Numerical (Quantitative - continuous)

BMI (Body Mass Index): Body mass index of the policyholder, indicating body fat based on height and weight.

Type: Numerical (Quantitative - continuous)

Children: Number of children or dependents covered by the insurance policy.

Type: Numerical (Whole number)

Smoker: Smoking status of the policyholder (Yes/No).

Type: Categorical

Region: Geographic region where the policyholder resides.

Type: Categorical

Charges : Medical expenses incurred by the policyholder.

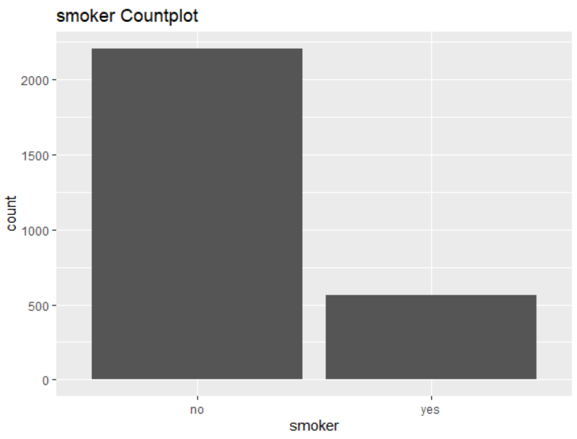
Type: Numerical (Continuous decimal)

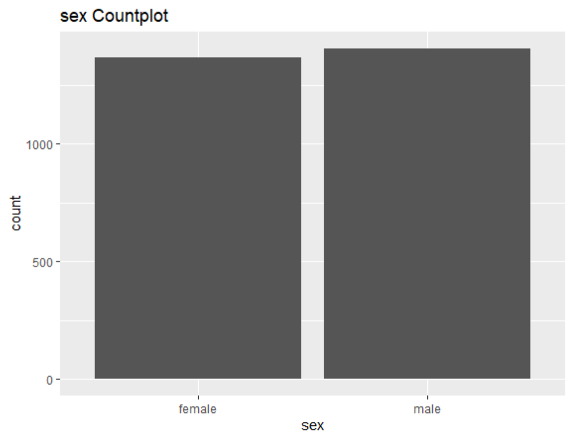
Exploratory Data Analysis

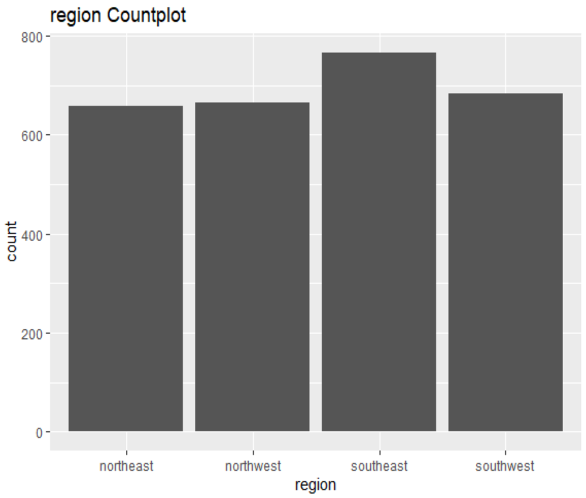
We have observed that there are no missing values in our dataset

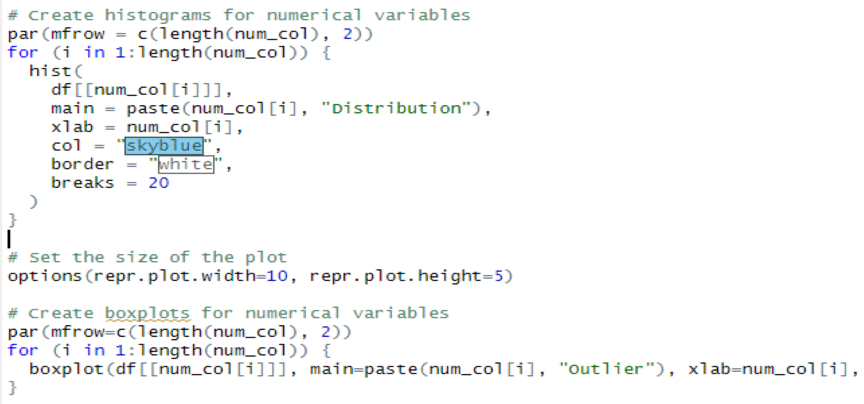
Univariate Analysis

Univariate analysis is important in regression because it allows for the individual examination of each variable in the dataset. By analyzing each variable separately, researchers can gain insights into its distribution, central tendency, dispersion, and potential outliers. This helps in understanding the characteristics of the variables before they are included in the regression model. Univariate analysis also aids in identifying any data preprocessing steps that may be necessary, such as normalization or transformation, to meet the assumptions of regression analysis. Overall, univariate analysis provides a foundational understanding of the data, which is essential for building reliable and accurate regression models.

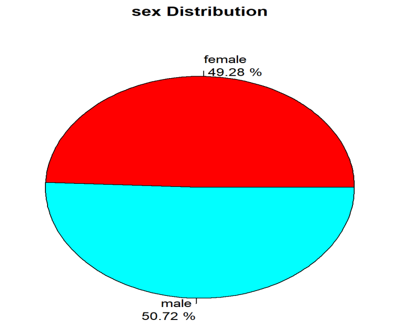
	The image is a bar graph titled "smoker countplot". The x-axis shows if someone smokes or not (yes or no), and the y-axis labeled "count" shows the number of people. There are more people who don't smoke than those who do.

	The image is a countplot, titled "sex countplot". The x-axis categorizes people by sex (female or male), and the y-axis labeled "count" shows the number of people. Here as we can see there is not much difference in the count of males and females, but males are a bit more compared to female population.

	The image is a bar graph titled "Region Countplot". The horizontal axis lists four regions in the United States: northeast, northwest, southeast, and southwest. The vertical axis labeled "count" shows the number of people in each region. It appears that the southeast has the most people, followed by the southwest, then the northwest, and least is the northeast.



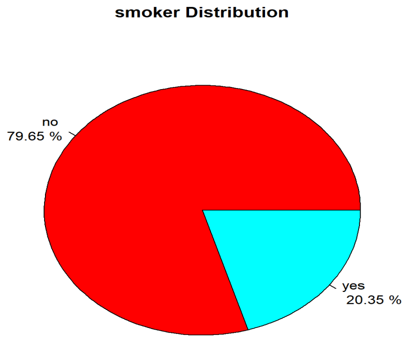
```
# Create histograms for numerical variables
par(mfrow = c(length(num_col), 2))
for (i in 1:length(num_col)) {
  hist(
    df[[num_col[i]]],
    main = paste(num_col[i], "Distribution"),
    xlab = num_col[i],
    col = "skyblue",
    border = "white",
    breaks = 20
  )
}

# Set the size of the plot
options(repr.plot.width=10, repr.plot.height=5)

# Create boxplots for numerical variables
par(mfrow=c(length(num_col), 2))
for (i in 1:length(num_col)) {
  boxplot(df[[num_col[i]]], main=paste(num_col[i], "Outlier"), xlab=num_col[i],
}
```



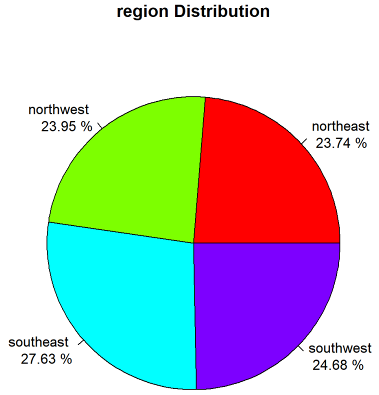
In order to deep dive into the statistics of the sex, region and smoker variables we have also plotted pie charts for the above stated variables.

We can get a clear idea and perception regarding how much percentage each category occupies.

For example with the histogram plotted for sex we can only see that Male category computes more data than the females' but with the the pie chart we can easily see that male computes 50.72% and female computes 49.28%

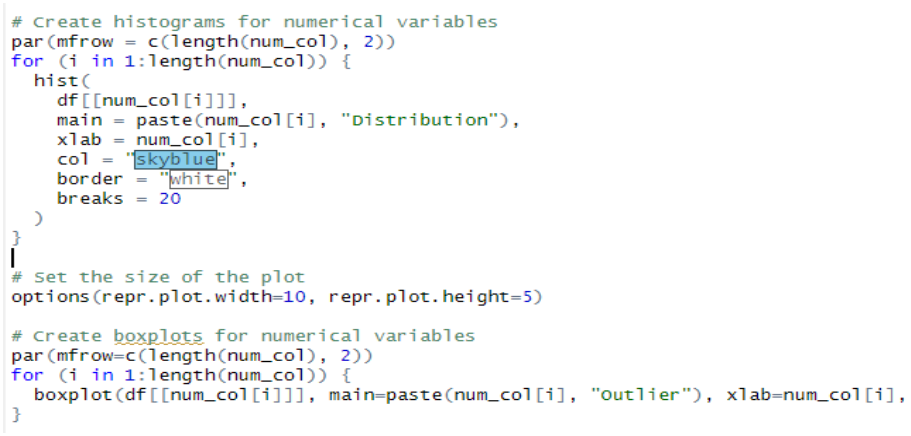


Similarly the pie chart for Smoker distribution , we can easily see that No computes to 79.65% and Yes computes to 20.35%.

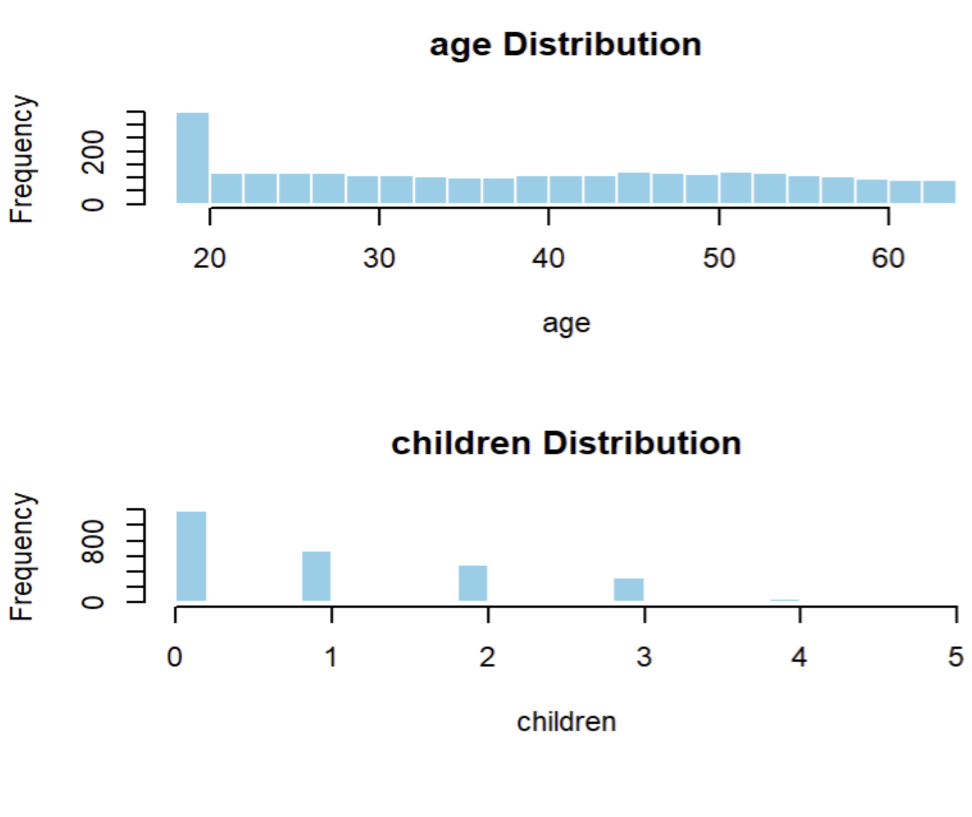


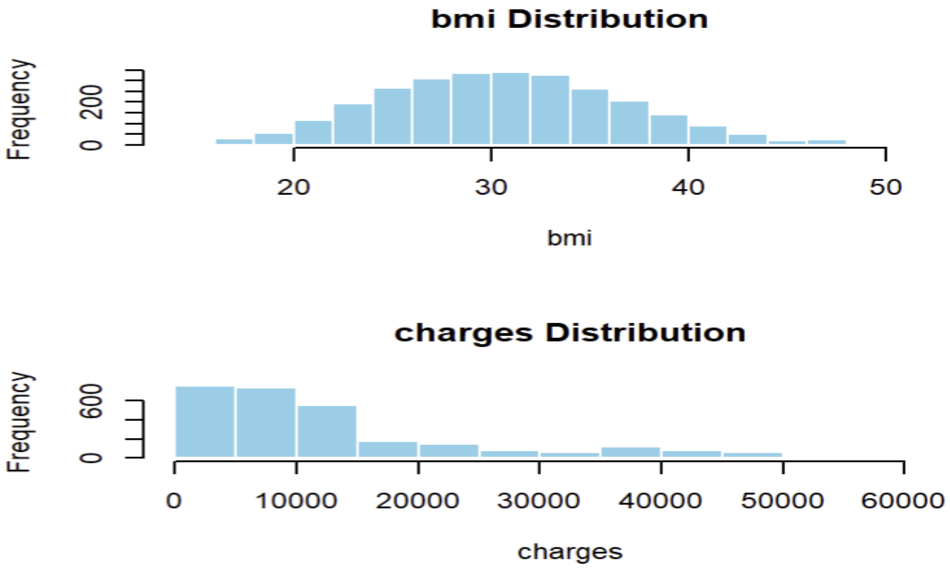
The pie chart for Region distribution ,shows that Northeast computes to 23.74%, Southwest is 24.68%, Northwest is 23.95% and Southeast computes to 27.63%.

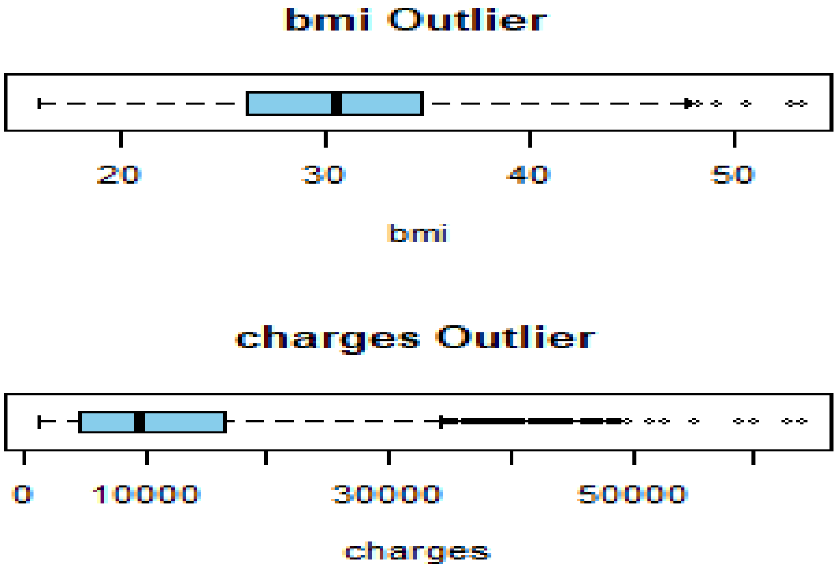
Creating Histograms for Numerical Variables



```
# Create histograms for numerical variables
par(mfrow = c(length(num_col), 2))
for (i in 1:length(num_col)) {
  hist(
    df[[num_col[i]]],
    main = paste(num_col[i], "Distribution"),
    xlab = num_col[i],
    col = "skyblue",
    border = "white",
    breaks = 20
  )
}

# Set the size of the plot
options(repr.plot.width=10, repr.plot.height=5)

# Create boxplots for numerical variables
par(mfrow=c(length(num_col), 2))
for (i in 1:length(num_col)) {
  boxplot(df[[num_col[i]]], main=paste(num_col[i], "Outlier"), xlab=num_col[i],
}
```





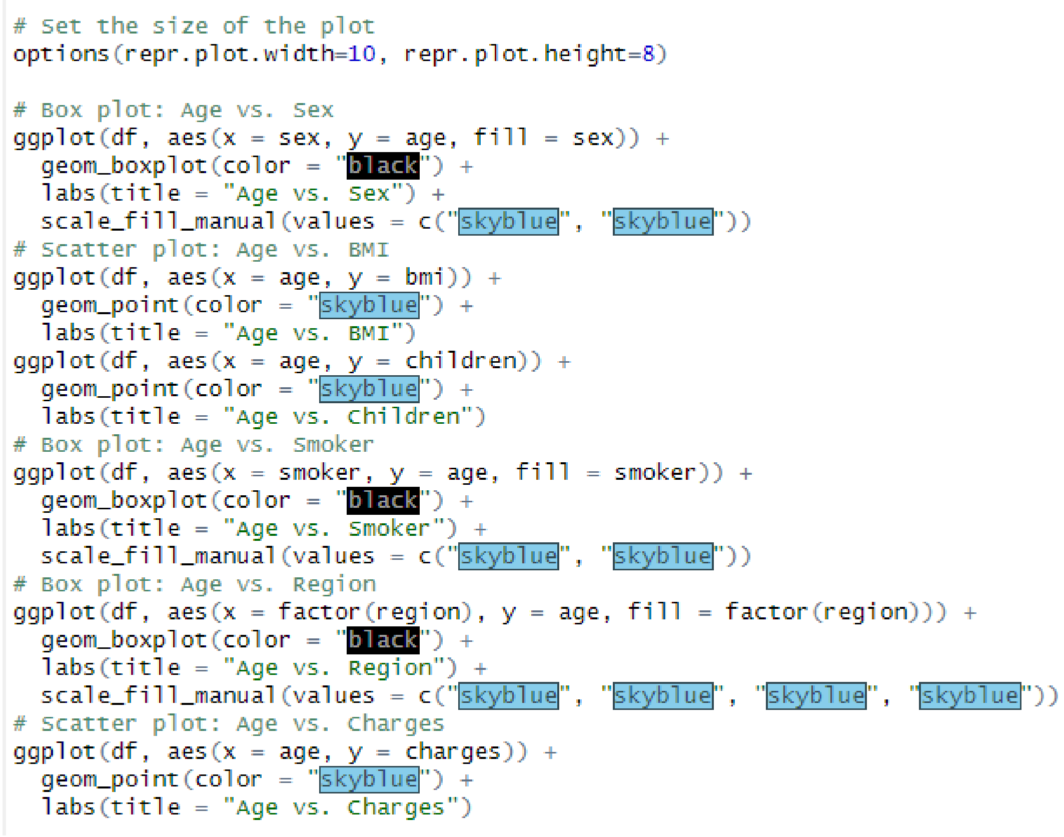


The following boxplots and histograms give us a clear idea about the data distribution and let us find the outliers in our dataset. As you can see we have plotted histograms for the quantitative variables such as age, bmi, children and charges. But, we also plotted the boxplot for the sake of scrutinizing the dataset by looking out for outliers and from the below box plots we can observe that charge variables have some outliers.

Bivariate Analysis

As the name suggests bivariate analysis concentrates on analyzing the relationship between two variables. So we plotted the necessary scatter plots and boxplots based on the variable type to have a better understanding of them. Bivariate analysis helps in identifying and understanding the relationship between the dependent variable and each independent variable separately before incorporating them into the regression model. This step is vital in determining if there's a linear or non-linear relationship between variables.

Age Vs All



```r
# Set the size of the plot
options(repr.plot.width=10, repr.plot.height=8)

# Box plot: Age vs. Sex
ggplot(df, aes(x = sex, y = age, fill = sex)) +
  geom_boxplot(color = "black") +
  labs(title = "Age vs. Sex") +
  scale_fill_manual(values = c("skyblue", "skyblue"))
# Scatter plot: Age vs. BMI
ggplot(df, aes(x = age, y = bmi)) +
  geom_point(color = "skyblue") +
  labs(title = "Age vs. BMI")
ggplot(df, aes(x = age, y = children)) +
  geom_point(color = "skyblue") +
  labs(title = "Age vs. Children")
# Box plot: Age vs. Smoker
ggplot(df, aes(x = smoker, y = age, fill = smoker)) +
  geom_boxplot(color = "black") +
  labs(title = "Age vs. Smoker") +
  scale_fill_manual(values = c("skyblue", "skyblue"))
# Box plot: Age vs. Region
ggplot(df, aes(x = factor(region), y = age, fill = factor(region))) +
  geom_boxplot(color = "black") +
  labs(title = "Age vs. Region") +
  scale_fill_manual(values = c("skyblue", "skyblue", "skyblue", "skyblue"))
# Scatter plot: Age vs. Charges
ggplot(df, aes(x = age, y = charges)) +
  geom_point(color = "skyblue") +
  labs(title = "Age vs. Charges")
```

We can see here we have plotted five graphs which have both boxplots and scatterplots.

As we know age is a quantitative variable and based on the other variables if they are a quantitative variable we plotted a scatterplot and if they are a qualitative variable we plotted a boxplot.

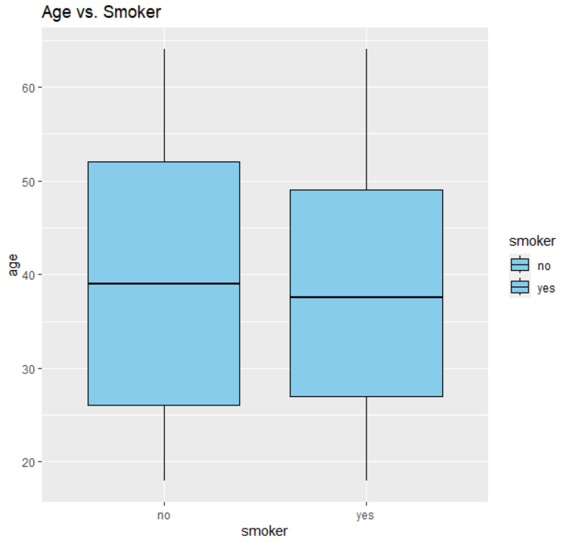
Age vs Sex - (quantitative vs qualitative) - boxplot

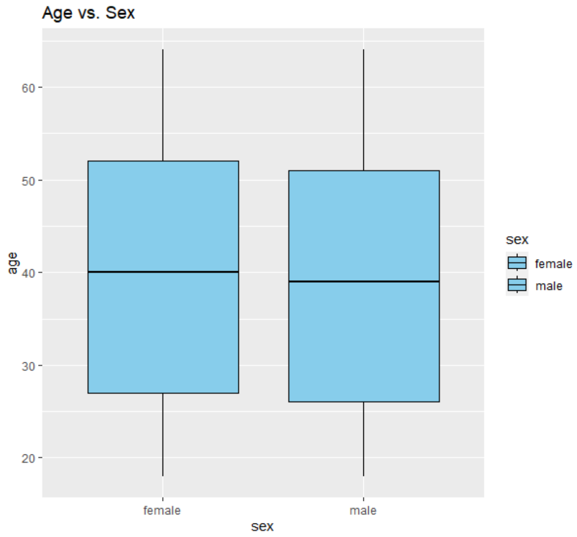
Age vs BMI - (quantitative vs quantitative) - scatterplot

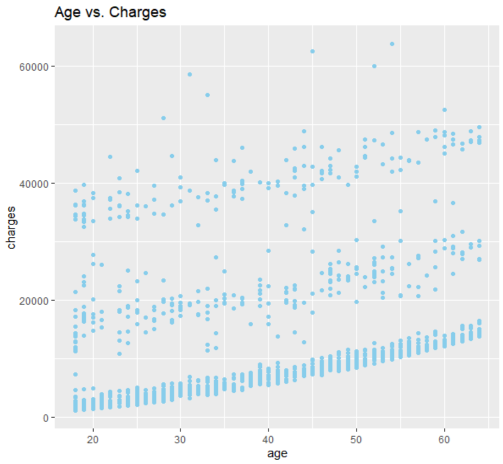
Age vs Smoker (quantitative vs qualitative) - boxplot

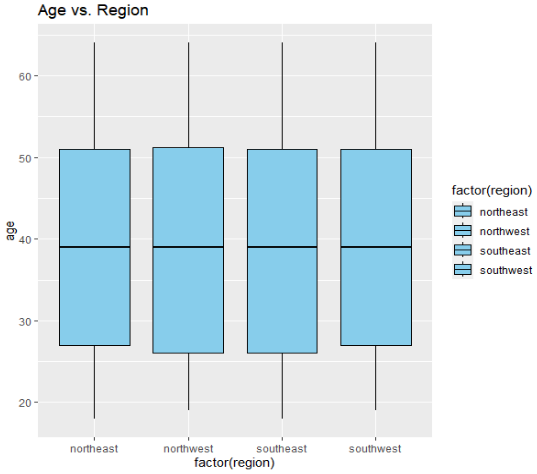
Age vs Region(quantitative vs qualitative) - boxplot

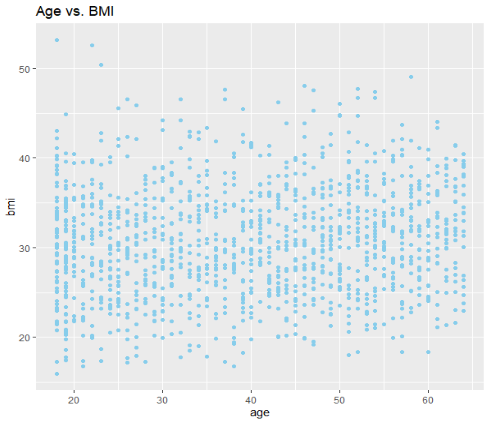
Age vs Charges - (quantitative vs quantitative) - scatterplot



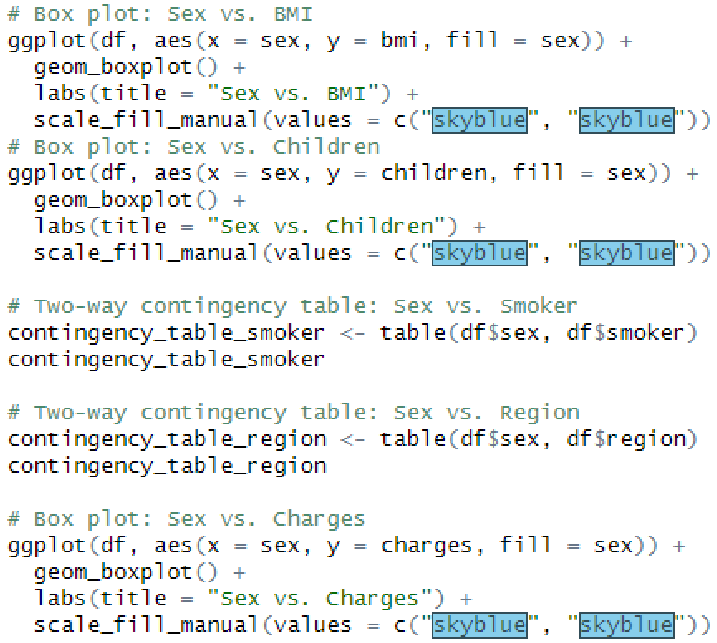








Sex Vs The Remaining Variables: Children, Smoker, Region and Charges



```
# Box plot: Sex vs. BMI
ggplot(df, aes(x = sex, y = bmi, fill = sex)) +
  geom_boxplot() +
  labs(title = "Sex vs. BMI") +
  scale_fill_manual(values = c("skyblue", "skyblue"))
# Box plot: Sex vs. Children
ggplot(df, aes(x = sex, y = children, fill = sex)) +
  geom_boxplot() +
  labs(title = "Sex vs. Children") +
  scale_fill_manual(values = c("skyblue", "skyblue"))

# Two-way contingency table: Sex vs. Smoker
contingency_table_smoker <- table(df$sex, df$smoker)
contingency_table_smoker

# Two-way contingency table: Sex vs. Region
contingency_table_region <- table(df$sex, df$region)
contingency_table_region

# Box plot: Sex vs. Charges
ggplot(df, aes(x = sex, y = charges, fill = sex)) +
  geom_boxplot() +
  labs(title = "Sex vs. Charges") +
  scale_fill_manual(values = c("skyblue", "skyblue"))
```

Since we already plotted a graph against age vs sex we have only five graphs in total with respect to sex.As we know sex is a qualitative variable and based on the the other variables if they are a quantitative variable we plotted a scatterplot and if they are a qualitative variable we plotted a boxplot.

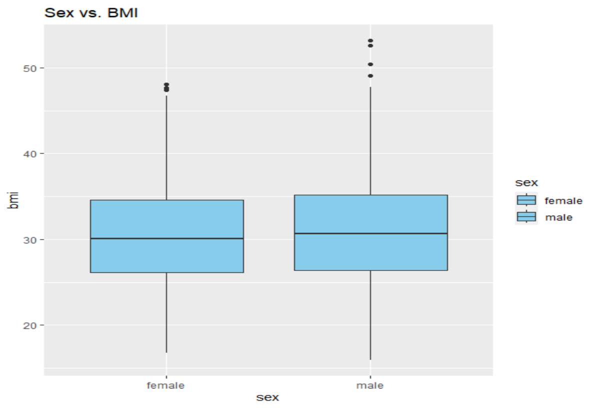
Sex vs BMI - (quantitative vs qualitative) - boxplot

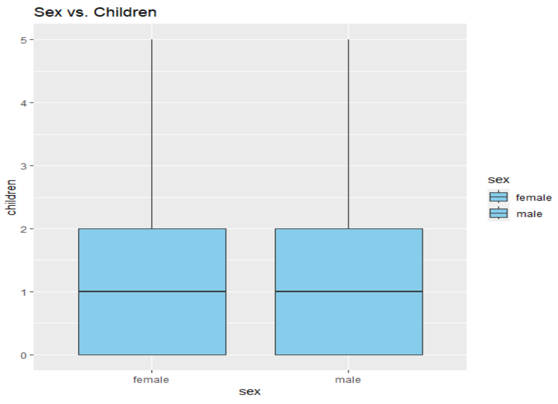
Sex vs Children - (quantitative vs qualitative) - boxplot

Sex vs Smoker (qualitative vs qualitative) - two way contingency table

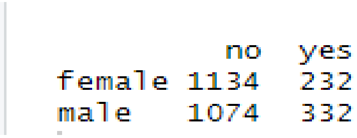
Sex vs Region(qualitative vs qualitative) - two way contingency table

Sex vs Charges - (quantitative vs qualitative) - boxplot



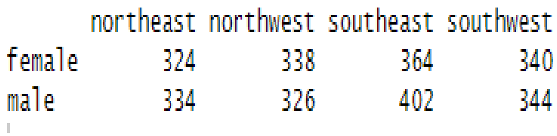


Two-way contingency table: Sex vs. Smoker



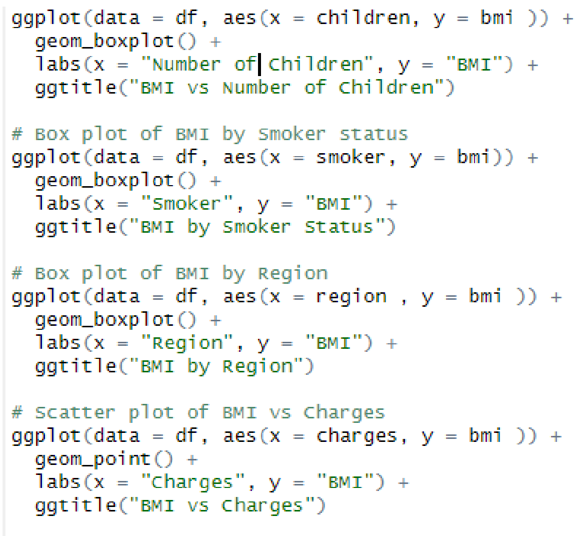
```
         no   yes
female 1134   232
male   1074   332
```

Two-way contingency table: Sex vs. Region



```
       northeast northwest southeast southwest
female       324       338       364       340
male         334       326       402       344
```

BMI vs The Remaining Variables : Children, Smoker, Region, Charges



```
ggplot(data = df, aes(x = children, y = bmi )) +
  geom_boxplot() +
  labs(x = "Number of Children", y = "BMI") +
  ggtitle("BMI vs Number of Children")

# Box plot of BMI by Smoker status
ggplot(data = df, aes(x = smoker, y = bmi)) +
  geom_boxplot() +
  labs(x = "Smoker", y = "BMI") +
  ggtitle("BMI by Smoker Status")

# Box plot of BMI by Region
ggplot(data = df, aes(x = region , y = bmi )) +
  geom_boxplot() +
  labs(x = "Region", y = "BMI") +
  ggtitle("BMI by Region")

# Scatter plot of BMI vs Charges
ggplot(data = df, aes(x = charges, y = bmi )) +
  geom_point() +
  labs(x = "Charges", y = "BMI") +
  ggtitle("BMI vs Charges")
```

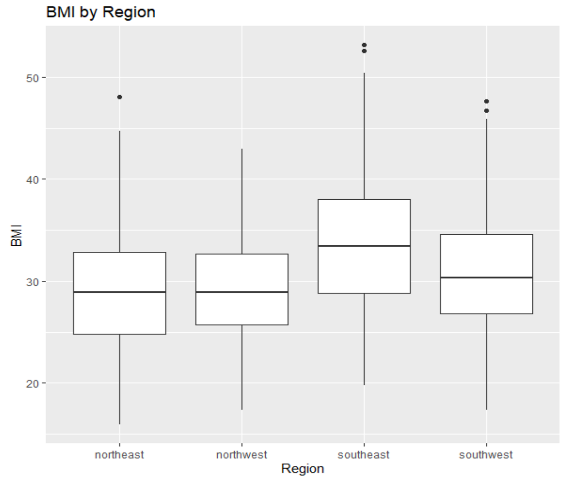
Since we already plotted a graph against bmi vs sex and age we have only four graphs in total with respect to bmi.As we know bmi is a quantitative variable and based on the the other variables if they are a quantitative variable we plotted a scatterplot and if they are a qualitative variable we plotted a boxplot.

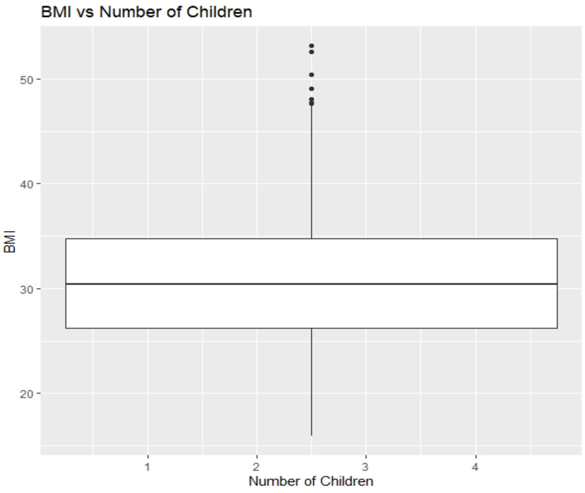
BMI vs Children - (quantitative vs qualitative) - boxplot

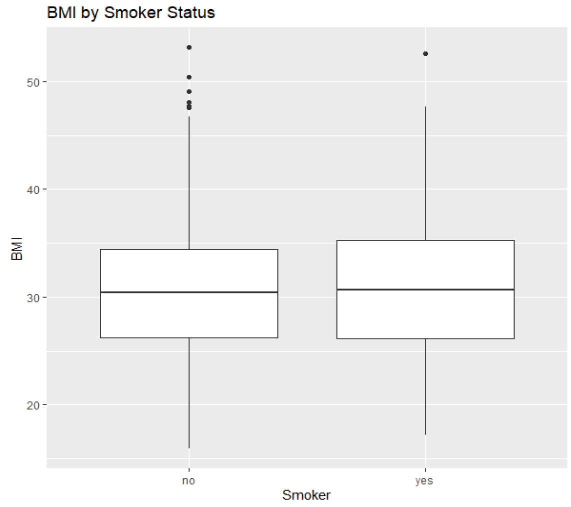
BMI vs Smoker -(quantitative vs qualitative) - boxplot

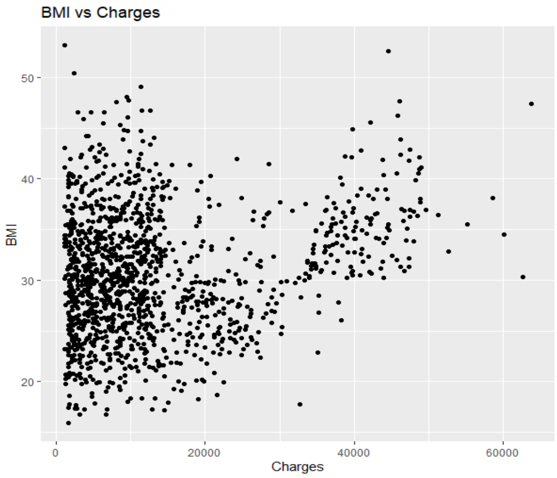
BMI vs Region - (quantitative vs qualitative) - boxplot

BMI vs Charges - (quantitative vs quantitative) - scatterplot


BMI by Region


BMI vs Number of Children


BMI by Smoker Status

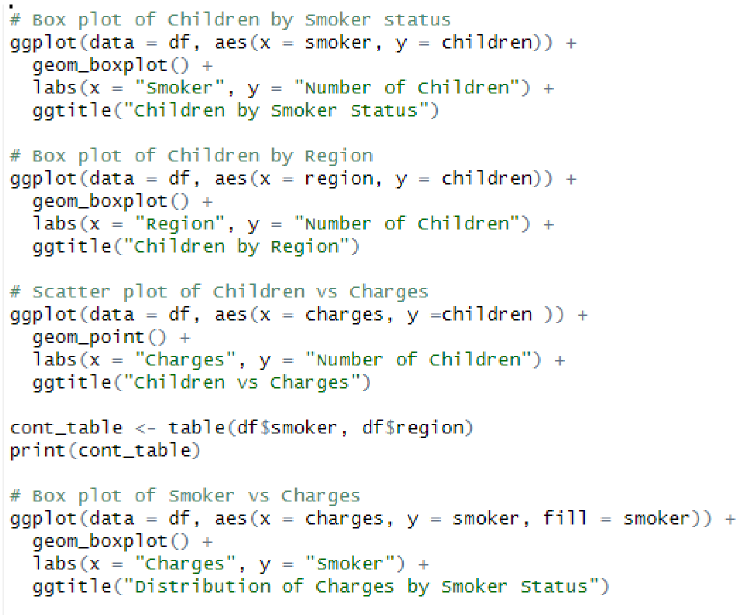

BMI vs Charges

Children vs The Remaining Variables: Smoker, Region, Charges

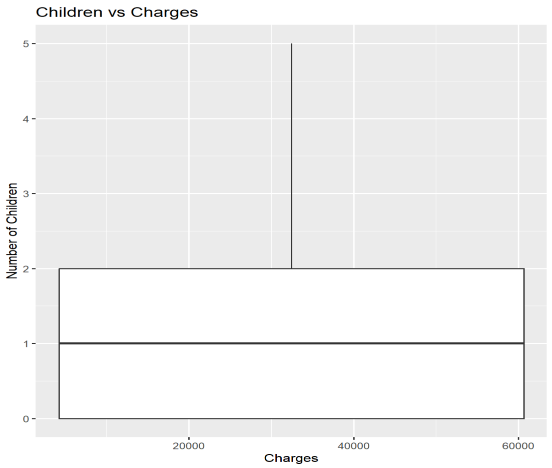
Since we already plotted a graph against children vs sex, age and bmi we have only three graphs in total with respect to children.As we know children is a qualitative variable and based on the the other variables if they are a quantitative variable we plotted a scatterplot and if they are a qualitative variable we plotted a boxplot.

Children vs Smoker -(quantitative vs qualitative) - boxplot
Children vs Region - (quantitative vs qualitative) - boxplot

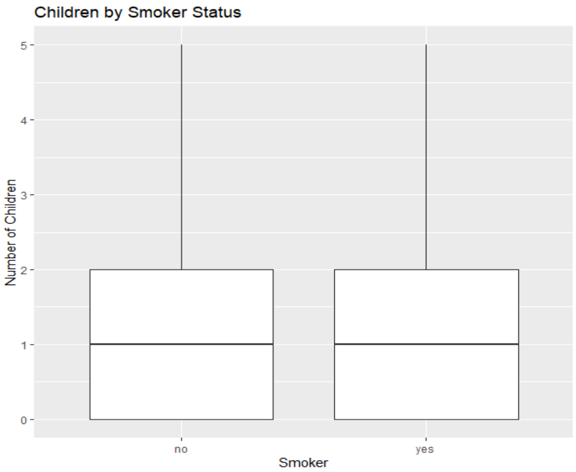
Children vs Charges - (quantitative vs qualitative) - boxplot



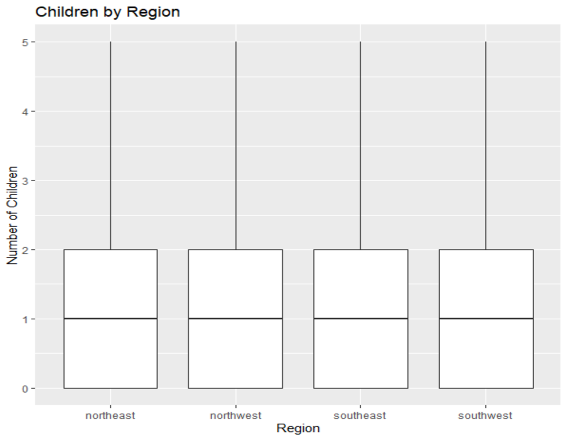
```
# Box plot of Children by Smoker status
ggplot(data = df, aes(x = smoker, y = children)) +
  geom_boxplot() +
  labs(x = "Smoker", y = "Number of Children") +
  ggtitle("Children by Smoker Status")

# Box plot of Children by Region
ggplot(data = df, aes(x = region, y = children)) +
  geom_boxplot() +
  labs(x = "Region", y = "Number of Children") +
  ggtitle("Children by Region")

# Scatter plot of Children vs Charges
ggplot(data = df, aes(x = charges, y =children )) +
  geom_point() +
  labs(x = "Charges", y = "Number of Children") +
  ggtitle("Children vs Charges")

cont_table <- table(df$smoker, df$region)
print(cont_table)

# Box plot of Smoker vs Charges
ggplot(data = df, aes(x = charges, y = smoker, fill = smoker)) +
  geom_boxplot() +
  labs(x = "Charges", y = "Smoker") +
  ggtitle("Distribution of Charges by Smoker Status")
```



Children vs Charges

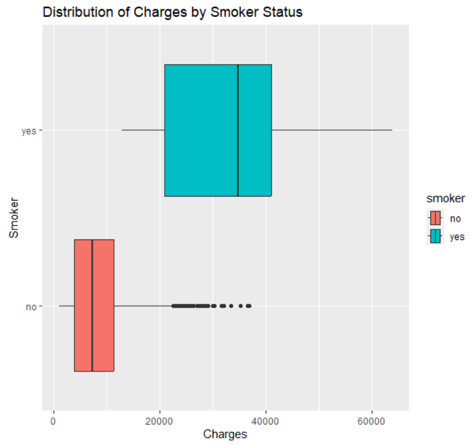


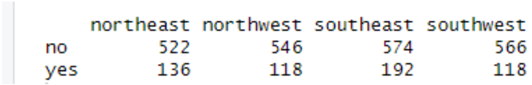
Children by Smoker Status



Children by Region

Smoker Vs The Remaining Variables : Region and Charges





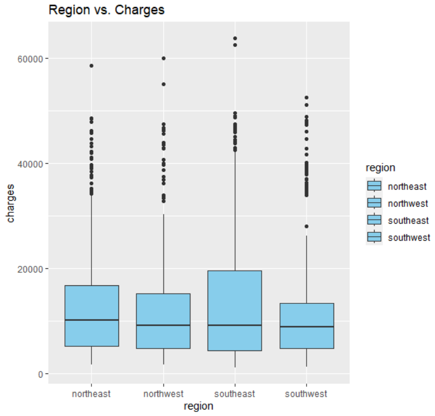
```
      northeast northwest southeast southwest
  no        522       546       574       566
  yes       136       118       192       118
```

Since we already plotted a graph against Smoker vs children ,sex, age and bmi we have only two graphs in total with respect to smoker.As we know smoker is a qualitative variable and based on the the other variables if they are a quantitative variable we plotted a scatterplot and if they are a qualitative variable we plotted a boxplot.

Smoker vs Charges - (quantitative vs qualitative) - boxplot

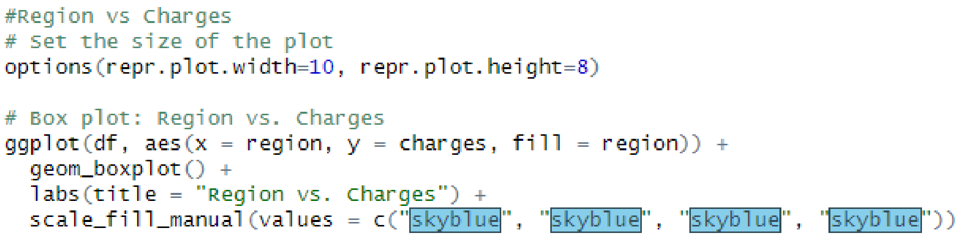
Smoker vs Region - (qualitative vs qualitative) - two way contingency table

Region vs The Remaining variable : Charges



Since we already plotted graphs against Region vs smoker, children ,sex, age and bmi we have only one graph in total with respect to region.As we know smoker is a qualitative variable and based on the other variable we plotted a boxplot.

Region vs Charges - (quantitative vs qualitative) - boxplot



```
#Region vs Charges
# Set the size of the plot
options(repr.plot.width=10, repr.plot.height=8)

# Box plot: Region vs. Charges
ggplot(df, aes(x = region, y = charges, fill = region)) +
  geom_boxplot() +
  labs(title = "Region vs. Charges") +
  scale_fill_manual(values = c("skyblue", "skyblue", "skyblue", "skyblue"))
```

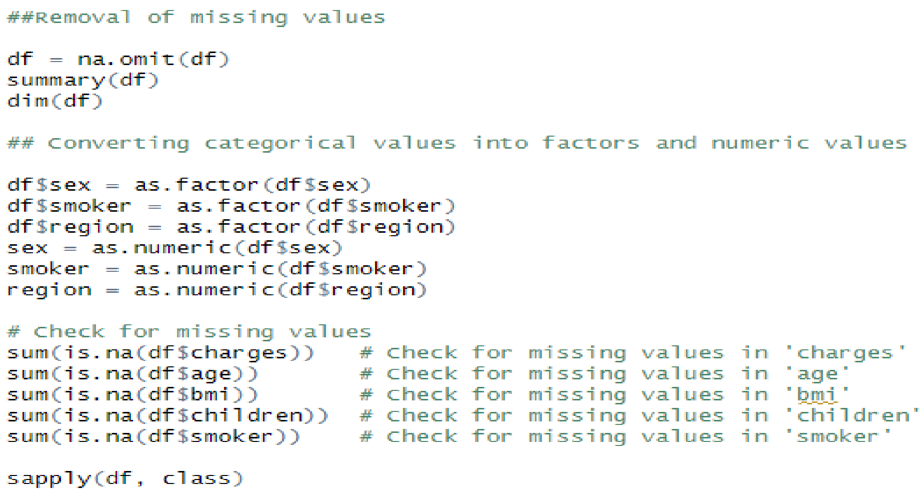
Data cleaning/preparation:

As a part of our data preparation process we first cleaned our dataset and looked for any missing values. Fortunately, our dataset doesn't have any null/missing values present in it. Then we looked out for transformations to fit our regression model with best transformed predictor terms and for that we plotted graphs between the transformed predictors against our response variable - charges.

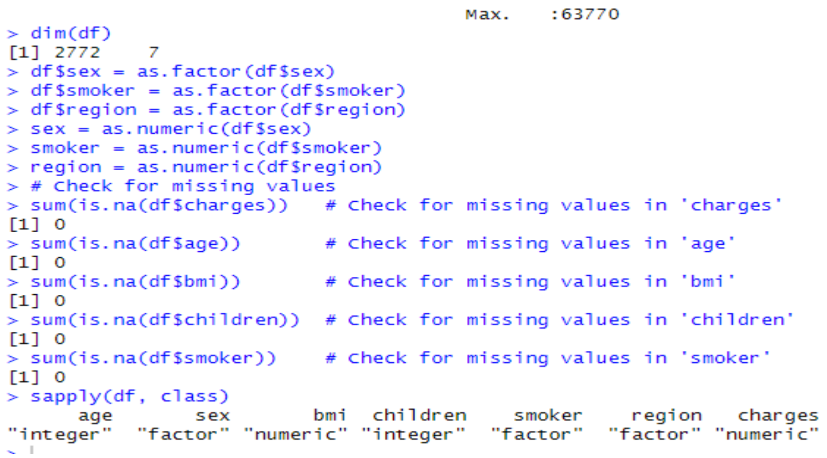
As we look at the transformations we can say there is a chance of transformation for the children and bmi variables. Since, children is a quantitative but finite variable we have considered sqrt(bmi) only.

So, for the interaction terms we have considered correlation matrix to understand the relationship among the independent variables and based on that we have considered the interaction between the transformed term which is sqrt(bmi) and smoker.

Hence, based on these transformation and interaction we have computed our linear regression model with a little trial and error method inorder to get a better r-squared value.

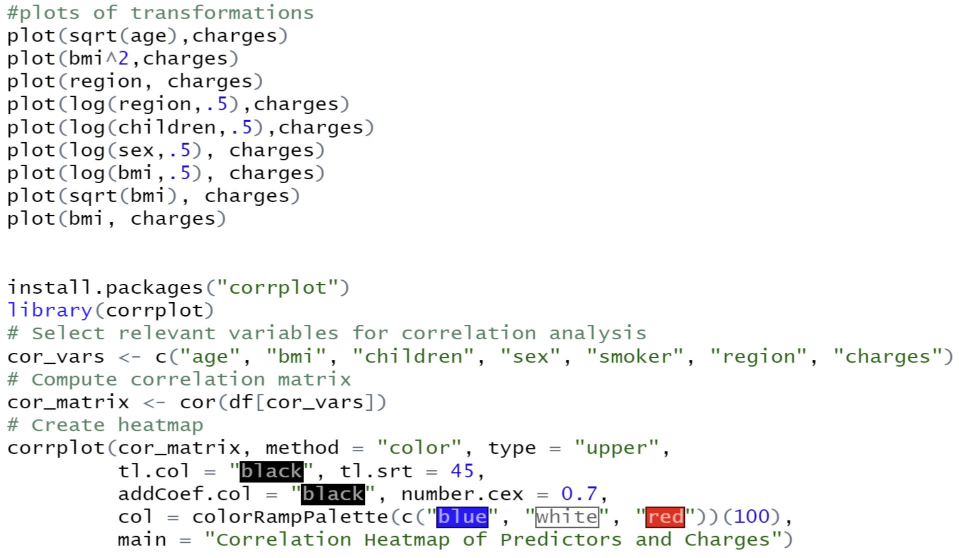


```
##Removal of missing values

df = na.omit(df)
summary(df)
dim(df)

## Converting categorical values into factors and numeric values

df$sex = as.factor(df$sex)
df$smoker = as.factor(df$smoker)
df$region = as.factor(df$region)
sex = as.numeric(df$sex)
smoker = as.numeric(df$smoker)
region = as.numeric(df$region)

# Check for missing values
sum(is.na(df$charges))    # Check for missing values in 'charges'
sum(is.na(df$age))        # Check for missing values in 'age'
sum(is.na(df$bmi))        # Check for missing values in 'bmi'
sum(is.na(df$children))   # Check for missing values in 'children'
sum(is.na(df$smoker))     # Check for missing values in 'smoker'

sapply(df, class)
```

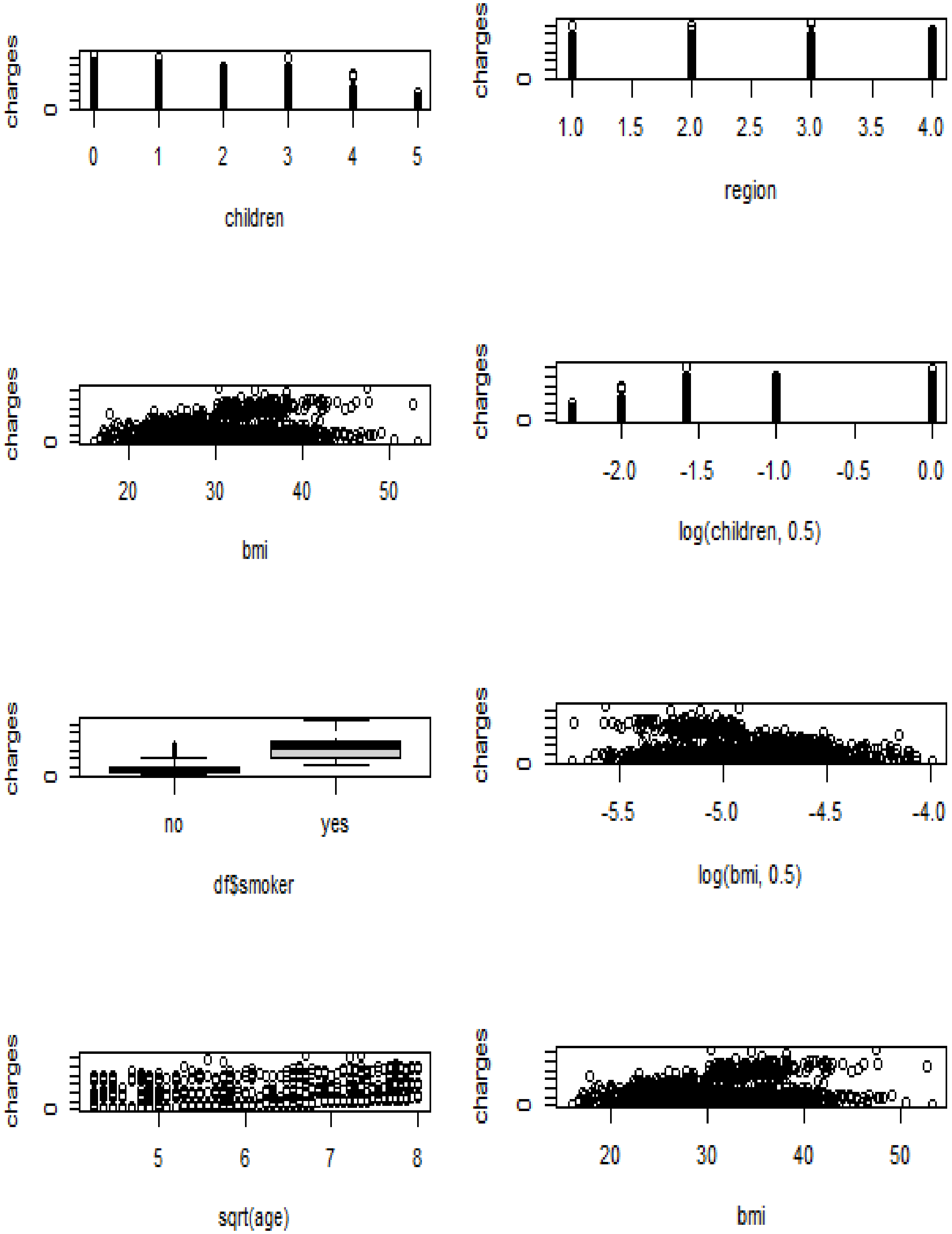


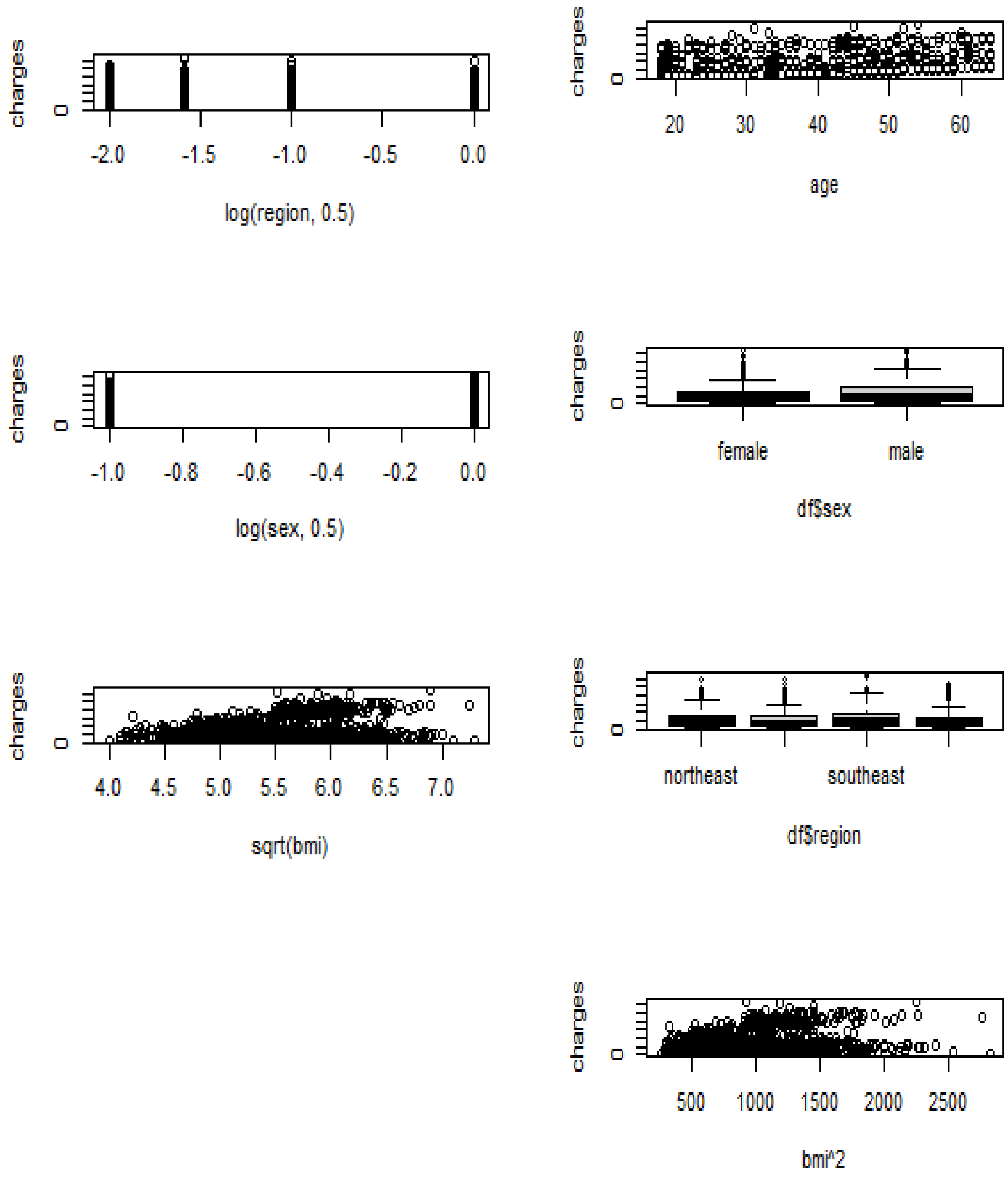
```
                                       Max.   :63770  
> dim(df)
[1] 2772    7
> df$sex = as.factor(df$sex)
> df$smoker = as.factor(df$smoker)
> df$region = as.factor(df$region)
> sex = as.numeric(df$sex)
> smoker = as.numeric(df$smoker)
> region = as.numeric(df$region)
> # Check for missing values
> sum(is.na(df$charges))    # Check for missing values in 'charges'
[1] 0
> sum(is.na(df$age))        # Check for missing values in 'age'
[1] 0
> sum(is.na(df$bmi))        # Check for missing values in 'bmi'
[1] 0
> sum(is.na(df$children))   # Check for missing values in 'children'
[1] 0
> sum(is.na(df$smoker))     # Check for missing values in 'smoker'
[1] 0
> sapply(df, class)
      age         sex         bmi   children     smoker     region    charges 
"integer"   "factor"  "numeric"  "integer"   "factor"   "factor"  "numeric" 
```

Scatter plot & boxplot of predictors with respect to DV-charges

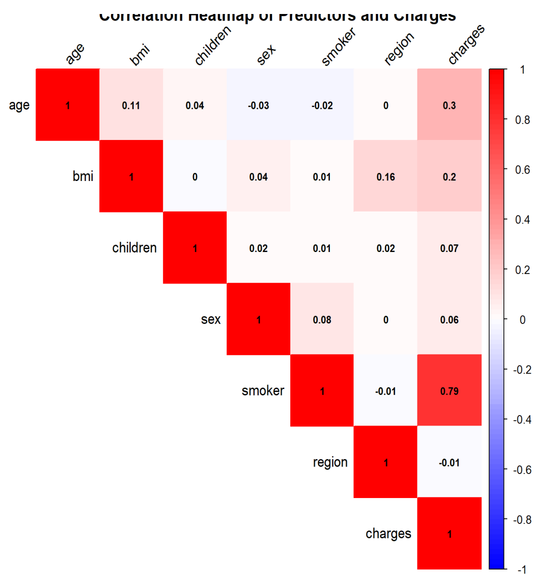


```
#plots of transformations
plot(sqrt(age),charges)
plot(bmi^2,charges)
plot(region, charges)
plot(log(region,.5),charges)
plot(log(children,.5),charges)
plot(log(sex,.5), charges)
plot(log(bmi,.5), charges)
plot(sqrt(bmi), charges)
plot(bmi, charges)


install.packages("corrplot")
library(corrplot)
# Select relevant variables for correlation analysis
cor_vars <- c("age", "bmi", "children", "sex", "smoker", "region", "charges")
# Compute correlation matrix
cor_matrix <- cor(df[cor_vars])
# Create heatmap
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7,
         col = colorRampPalette(c("blue", "white", "red"))(100),
         main = "Correlation Heatmap of Predictors and Charges")
```


charges
0
0
1
2
3
4
5
children
charges
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
region
charges
0
20
30
40
50
bmi
charges
0
-2.0
-1.5
-1.0
-0.5
0.0
log(children, 0.5)
charges
0
no
yes
df$smoker
charges
0
-5.5
-5.0
-4.5
-4.0
log(bmi, 0.5)
charges
0
5
6
7
8
sqrt(age)
charges
0
20
30
40
50
bmi



By looking at the plots we can say that the variable children can be transformed.



	age	bmi	children	sex	smoker	region	charges
age	1	0.11	0.04	-0.03	-0.02	0	0.3
bmi		1	0	0.04	0.01	0.16	0.2
children			1	0.02	0.01	0.02	0.07
sex				1	0.08	0	0.06
smoker					1	-0.01	0.79
region						1	-0.01
charges							1

Creating a correlation heat-map of all predictors (age, BMI, children, sex, smoker, region) versus charges in a regression analysis is significant for several reasons. Firstly, it provides a visual representation of the strength and direction of the relationships between each predictor variable and the target variable (charges). This helps in identifying which predictors are most strongly correlated with charges, indicating their potential significance in predicting insurance charges. Additionally, the heat-map highlights any multicollinearity between predictor variables, which occurs when two or more predictors are highly correlated with each other.

Part III Analysis and Findings

Variable selection & model building

Regression problem using Generalized Regression models

We conducted a regression analysis to investigate the association between charges and different variables, utilizing exploratory data analysis (EDA) techniques. Our analysis involved employing both linear and generalized regression methods.

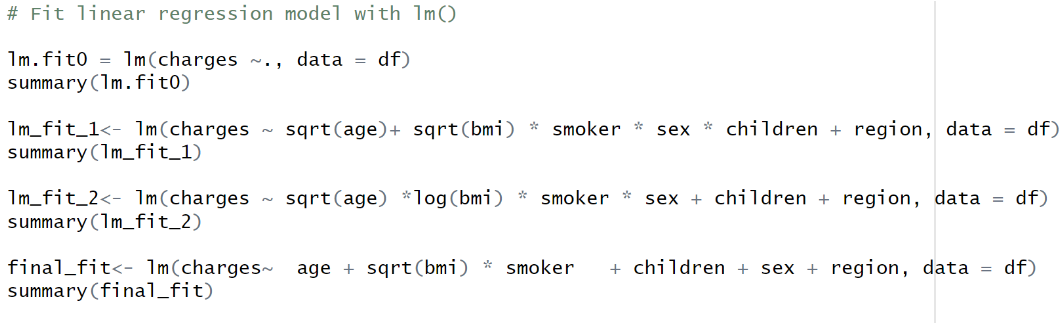
We utilized subset selection techniques such as best subset selection, forward stepwise selection, backward stepwise selection, and shrinkage methods to refine our models and identify the most relevant predictors for charges.

Predictor terms: age, sqrt(bmi) * smoker, children, sex, region

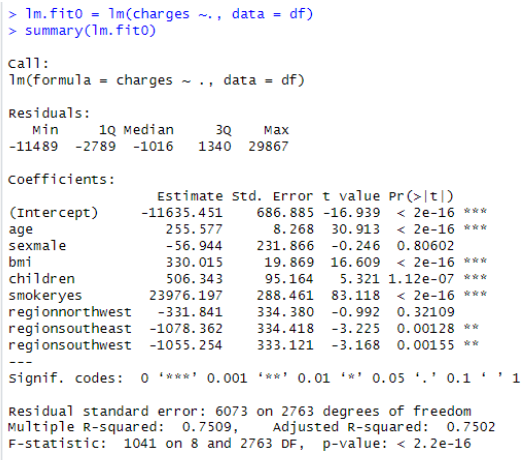
Original Predictor terms: age, sex, bmi, children, smoker, region

Transformed predictors: sqrt(b8mi)

Interaction terms:sqrt(bmi) * smoker



```
# Fit linear regression model with lm()

lm.fit0 = lm(charges ~., data = df)
summary(lm.fit0)

lm_fit_1<- lm(charges ~ sqrt(age)+ sqrt(bmi) * smoker * sex * children + region, data = df)
summary(lm_fit_1)

lm_fit_2<- lm(charges ~ sqrt(age) *log(bmi) * smoker * sex + children + region, data = df)
summary(lm_fit_2)

final_fit<- lm(charges~  age + sqrt(bmi) * smoker   + children + sex + region, data = df)
summary(final_fit)
```

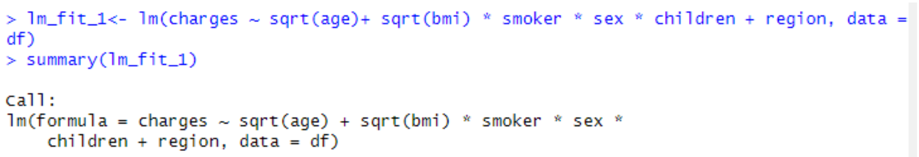


```
> lm.fit0 = lm(charges ~., data = df)
> summary(lm.fit0)

Call:
lm(formula = charges ~ ., data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-11489  -2789  -1016   1340  29867 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -11635.451    686.885 -16.939  < 2e-16 ***
age               255.577      8.268  30.913  < 2e-16 ***
sexmale           -56.944    231.866  -0.246  0.80602    
bmi               330.015     19.869  16.609  < 2e-16 ***
children          506.343     95.164   5.321 1.12e-07 ***
smokeryes       23976.197    288.461  83.118  < 2e-16 ***
regionnorthwest  -331.841    334.380  -0.992  0.32109    
regionsoutheast -1078.362    334.418  -3.225  0.00128 ** 
regionsouthwest -1055.254    333.121  -3.168  0.00155 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6073 on 2763 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7502 
F-statistic:  1041 on 8 and 2763 DF,  p-value: < 2.2e-16
```

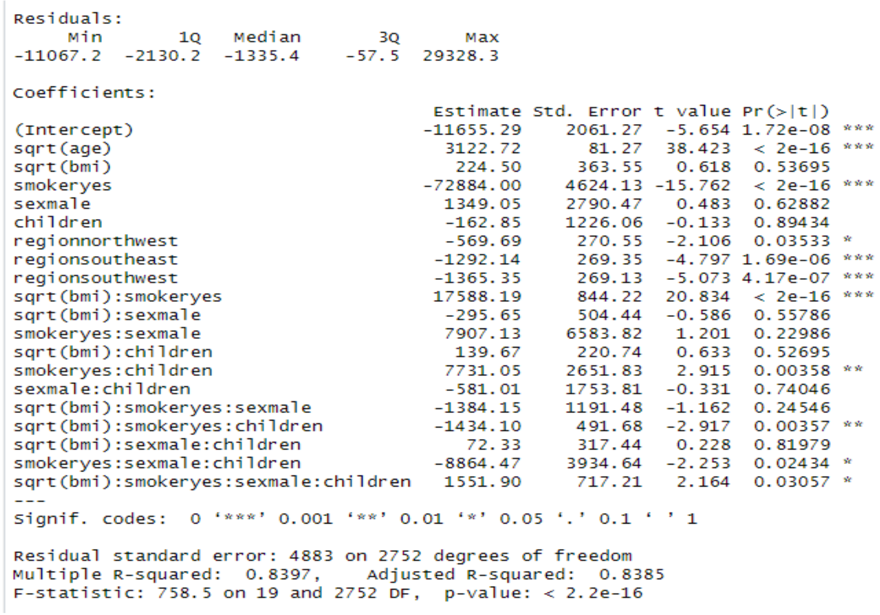
lm.fit0: In this linear regression model, the proportion of variance explained by the predictors, as indicated by the R-squared value is 0.7509, suggesting that approximately 75.09% of the variance in charges can be accounted for by the selected predictors.

Among the predictors, age, BMI, children, and smoking status (smokeryes) show statistically significant relationships with charges, as evidenced by their low p-values (< 0.05).

However, the coefficients for sex, and the regions (northwest, southeast, southwest) are not statistically significant, implying that they may not have a significant impact on charges in this model.

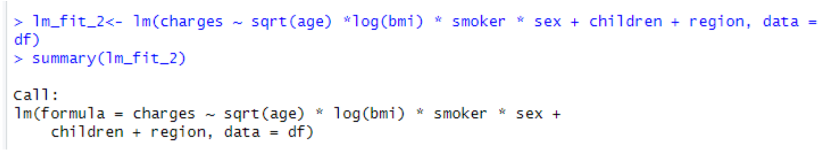


```
> lm_fit_1<- lm(charges ~ sqrt(age)+ sqrt(bmi) * smoker * sex * children + region, data =
df)
> summary(lm_fit_1)

Call:
lm(formula = charges ~ sqrt(age) + sqrt(bmi) * smoker * sex *
    children + region, data = df)
```

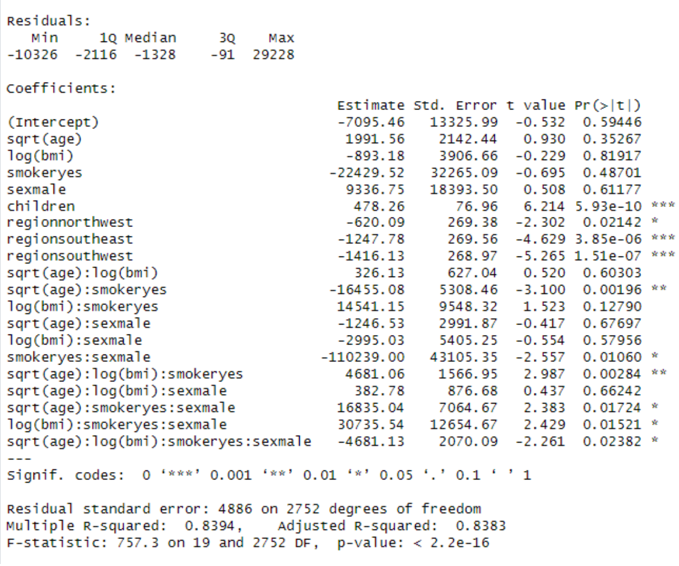


```
Residuals:
     Min       1Q   Median       3Q      Max 
-11067.2  -2130.2  -1335.4    -57.5  29328.3 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -11655.29    2061.27  -5.654 1.72e-08 ***
sqrt(age)                              3122.72      81.27  38.423  < 2e-16 ***
sqrt(bmi)                               224.50     363.55   0.618  0.53695    
smokeryes                            -72884.00    4624.13 -15.762  < 2e-16 ***
sexmale                                1349.05    2790.47   0.483  0.62882    
children                               -162.85    1226.06  -0.133  0.89434    
regionnorthwest                        -569.69     270.55  -2.106  0.03533 *  
regionsoutheast                       -1292.14     269.35  -4.797 1.69e-06 ***
regionsouthwest                       -1365.35     269.13  -5.073 4.17e-07 ***
sqrt(bmi):smokeryes                   17588.19     844.22  20.834  < 2e-16 ***
sqrt(bmi):sexmale                      -295.65     504.44  -0.586  0.55786    
smokeryes:sexmale                      7907.13    6583.82   1.201  0.22986    
sqrt(bmi):children                      139.67     220.74   0.633  0.52695    
smokeryes:children                     7731.05    2651.83   2.915  0.00358 ** 
sexmale:children                       -581.01    1753.81  -0.331  0.74046    
sqrt(bmi):smokeryes:sexmale           -1384.15    1191.48  -1.162  0.24546    
sqrt(bmi):smokeryes:children          -1434.10     491.68  -2.917  0.00357 ** 
sqrt(bmi):sexmale:children               72.33     317.44   0.228  0.81979    
smokeryes:sexmale:children            -8864.47    3934.64  -2.253  0.02434 *  
sqrt(bmi):smokeryes:sexmale:children   1551.90     717.21   2.164  0.03057 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4883 on 2752 degrees of freedom
Multiple R-squared:  0.8397,	Adjusted R-squared:  0.8385 
F-statistic: 758.5 on 19 and 2752 DF,  p-value: < 2.2e-16
```

lm_fit_1: When we look at this model and its R-squared value, which is 0.8397, we can observe that the selected predictors collectively explain approximately 83.97% of the variance in charges. Among the predictors, sqrt(age), smokeryes, and interactions involving regions (northwest, southeast, southwest) are statistically significant, as evidenced by their low p-values (< 0.05). However, predictors such as sqrt(bmi), sex, children, and certain interactions (e.g., sqrt(bmi):sexmale) do not show statistically significant relationships with charges, as indicated by their high p-values.



```
> lm_fit_2<- lm(charges ~ sqrt(age) *log(bmi) * smoker * sex + children + region, data =
df)
> summary(lm_fit_2)

Call:
lm(formula = charges ~ sqrt(age) * log(bmi) * smoker * sex + 
    children + region, data = df)
```



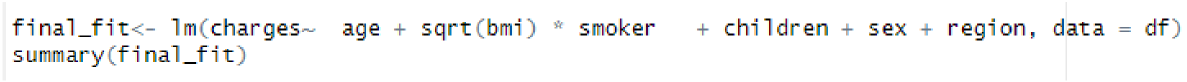
```
Residuals:
   Min     1Q Median     3Q    Max 
-10326  -2116  -1328    -91  29228 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -7095.46   13325.99  -0.532  0.59446    
sqrt(age)                               1991.56    2142.44   0.930  0.35267    
log(bmi)                                -893.18    3906.66  -0.229  0.81917    
smokeryes                             -22429.52   32265.09  -0.695  0.48701    
sexmale                                 9336.75   18393.50   0.508  0.61177    
children                                 478.26      76.96   6.214 5.93e-10 ***
regionnorthwest                         -620.09     269.38  -2.302  0.02142 *  
regionsoutheast                        -1247.78     269.56  -4.629 3.85e-06 ***
regionsouthwest                        -1416.13     268.97  -5.265 1.51e-07 ***
sqrt(age):log(bmi)                       326.13     627.04   0.520  0.60303    
sqrt(age):smokeryes                   -16455.08    5308.46  -3.100  0.00196 ** 
log(bmi):smokeryes                     14541.15    9548.32   1.523  0.12790    
sqrt(age):sexmale                      -1246.53    2991.87  -0.417  0.67697    
log(bmi):sexmale                       -2995.03    5405.25  -0.554  0.57956    
smokeryes:sexmale                    -110239.00   43105.35  -2.557  0.01060 *  
sqrt(age):log(bmi):smokeryes            4681.06    1566.95   2.987  0.00284 ** 
sqrt(age):log(bmi):sexmale               382.78     876.68   0.437  0.66242    
sqrt(age):smokeryes:sexmale            16835.04    7064.67   2.383  0.01724 *  
log(bmi):smokeryes:sexmale             30735.54   12654.67   2.429  0.01521 *  
sqrt(age):log(bmi):smokeryes:sexmale   -4681.13    2070.09  -2.261  0.02382 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4886 on 2752 degrees of freedom
Multiple R-squared:  0.8394,	Adjusted R-squared:  0.8383 
F-statistic: 757.3 on 19 and 2752 DF,  p-value: < 2.2e-16
```

lm_fit_2: In this linear regression model, the R-squared value, which is 0.8394, indicates that the selected predictors collectively explain approximately 83.94% of the variance in charges. Among the predictors, children show a statistically significant relationship with charges, as evidenced by their low p-value (< 0.05).

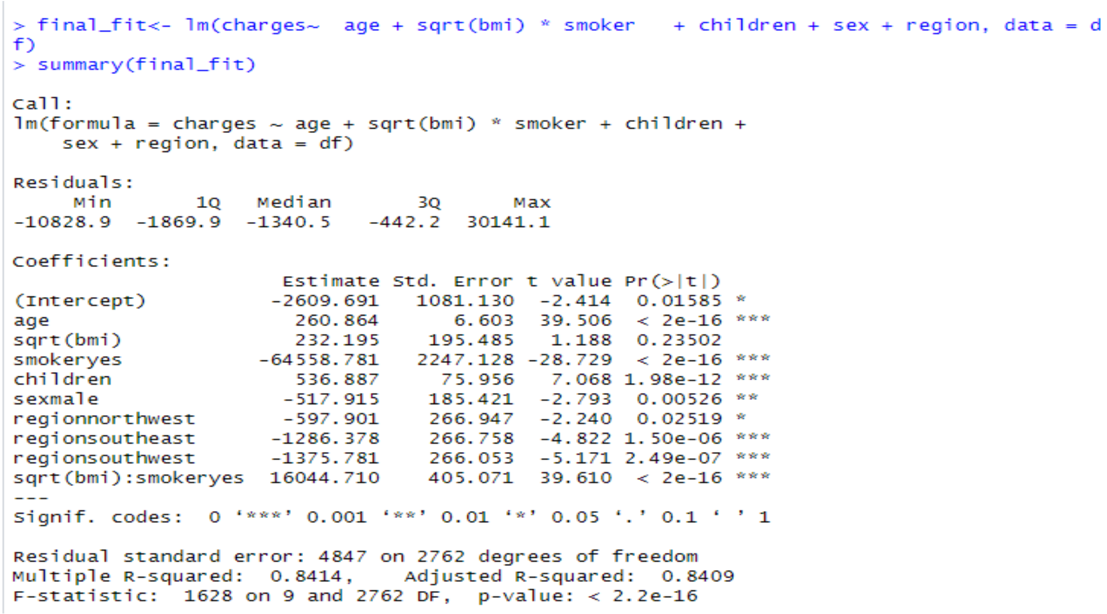
However, other predictors such as sqrt(age), log(bmi), smokeryes, sex, and interactions involving regions (northwest, southeast, southwest) do not show statistically significant relationships with charges, as indicated by their high p-values.

The intercept term is also not statistically significant. Overall, while this model explains a substantial proportion of the variance in charges, several predictors do not have significant impacts on charges, as indicated by their high p-values.

Final Fit



```
final_fit<- lm(charges~  age + sqrt(bmi) * smoker   + children + sex + region, data = df)
summary(final_fit)
```

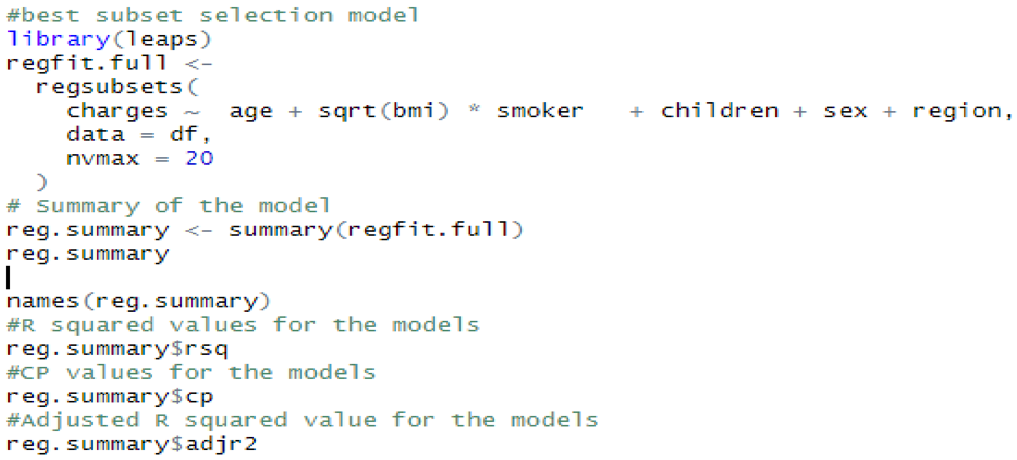


```
> final_fit<- lm(charges~  age + sqrt(bmi) * smoker   + children + sex + region, data = d
f)
> summary(final_fit)

Call:
lm(formula = charges ~ age + sqrt(bmi) * smoker + children + 
    sex + region, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-10828.9  -1869.9  -1340.5   -442.2  30141.1 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -2609.691   1081.130  -2.414  0.01585 *  
age                   260.864      6.603  39.506  < 2e-16 ***
sqrt(bmi)             232.195    195.485   1.188  0.23502    
smokeryes          -64558.781   2247.128 -28.729  < 2e-16 ***
children              536.887     75.956   7.068 1.98e-12 ***
sexmale              -517.915    185.421  -2.793  0.00526 ** 
regionnorthwest      -597.901    266.947  -2.240  0.02519 *  
regionsoutheast     -1286.378    266.758  -4.822 1.50e-06 ***
regionsouthwest     -1375.781    266.053  -5.171 2.49e-07 ***
sqrt(bmi):smokeryes  16044.710    405.071  39.610  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4847 on 2762 degrees of freedom
Multiple R-squared:  0.8414,	Adjusted R-squared:  0.8409 
F-statistic:  1628 on 9 and 2762 DF,  p-value: < 2.2e-16
```

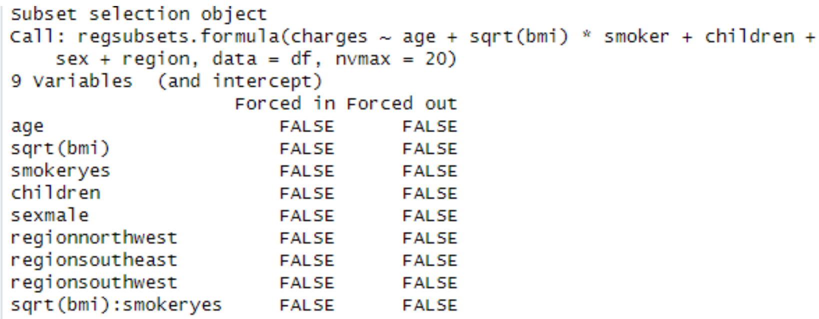
Final_fit: When we examine this linear regression model, the R-squared value, which is 0.8414, indicates that the selected predictors collectively explain approximately 84.14% of the variance in charges.

Among the predictors, age, smokeryes, children, sex, and interactions involving regions (northwest, southeast, southwest) show statistically significant relationships with charges, as evidenced by their low p-values (< 0.05). The intercept term is also statistically significant, suggesting its contribution to the model's predictions.

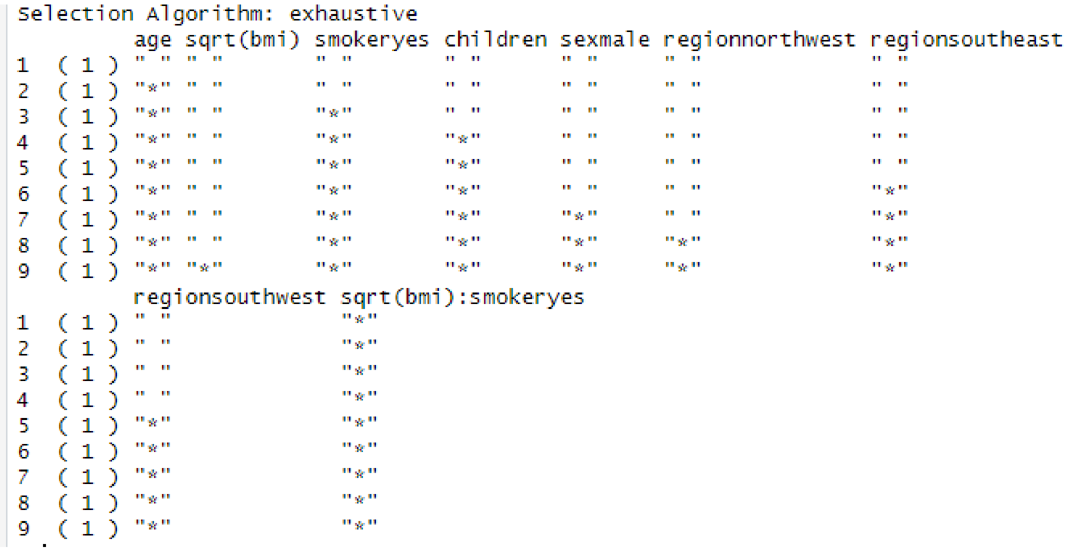
Best subset selection model



```
#best subset selection model
library(leaps)
regfit.full <-
  regsubsets(
    charges ~   age + sqrt(bmi) * smoker    + children + sex + region,
    data = df,
    nvmax = 20
  )
# summary of the model
reg.summary <- summary(regfit.full)
reg.summary

names(reg.summary)
#R squared values for the models
reg.summary$rsq
#CP values for the models
reg.summary$cp
#Adjusted R squared value for the models
reg.summary$adjr2
```

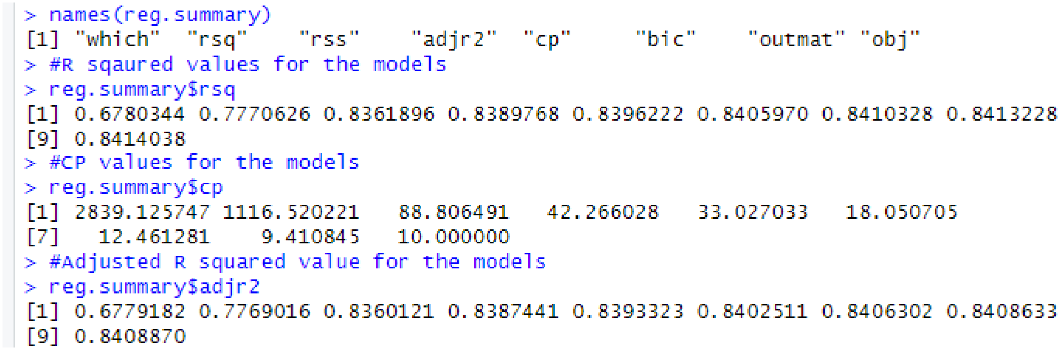


```
Subset selection object
Call: regsubsets.formula(charges ~ age + sqrt(bmi) * smoker + children +
    sex + region, data = df, nvmax = 20)
9 Variables  (and intercept)
                    Forced in Forced out
age                     FALSE      FALSE
sqrt(bmi)               FALSE      FALSE
smokeryes               FALSE      FALSE
children                FALSE      FALSE
sexmale                 FALSE      FALSE
regionnorthwest         FALSE      FALSE
regionsoutheast         FALSE      FALSE
regionsouthwest         FALSE      FALSE
sqrt(bmi):smokeryes     FALSE      FALSE
```



```
Selection Algorithm: exhaustive
         age sqrt(bmi) smokeryes children sexmale regionnorthwest regionsoutheast
1  ( 1 ) " " " "       " "       " "      " "     " "             " "            
2  ( 1 ) "*" " "       " "       " "      " "     " "             " "            
3  ( 1 ) "*" " "       "*"       " "      " "     " "             " "            
4  ( 1 ) "*" " "       "*"       "*"      " "     " "             " "            
5  ( 1 ) "*" " "       "*"       "*"      " "     " "             " "            
6  ( 1 ) "*" " "       "*"       "*"      " "     " "             "*"            
7  ( 1 ) "*" " "       "*"       "*"      "*"     " "             "*"            
8  ( 1 ) "*" " "       "*"       "*"      "*"     "*"             "*"            
9  ( 1 ) "*" "*"       "*"       "*"      "*"     "*"             "*"            
         regionsouthwest sqrt(bmi):smokeryes
1  ( 1 ) " "             "*"                
2  ( 1 ) " "             "*"                
3  ( 1 ) " "             "*"                
4  ( 1 ) " "             "*"                
5  ( 1 ) "*"             "*"                
6  ( 1 ) "*"             "*"                
7  ( 1 ) "*"             "*"                
8  ( 1 ) "*"             "*"                
9  ( 1 ) "*"             "*"                
```

Outputs of R squared value, CP values, Adjusted R squared value for the models

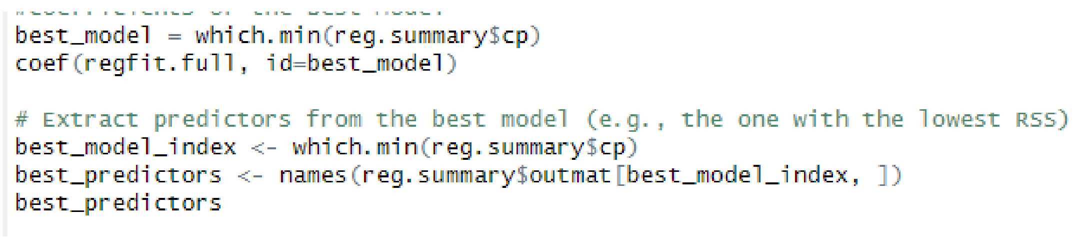


```
> names(reg.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"   
> #R sqaured values for the models
> reg.summary$rsq
[1] 0.6780344 0.7770626 0.8361896 0.8389768 0.8396222 0.8405970 0.8410328 0.8413228
[9] 0.8414038
> #CP values for the models
> reg.summary$cp
[1] 2839.125747 1116.520221   88.806491   42.266028   33.027033   18.050705
[7]   12.461281    9.410845   10.000000
> #Adjusted R squared value for the models
> reg.summary$adjr2
[1] 0.6779182 0.7769016 0.8360121 0.8387441 0.8393323 0.8402511 0.8406302 0.8408633
[9] 0.8408870
```

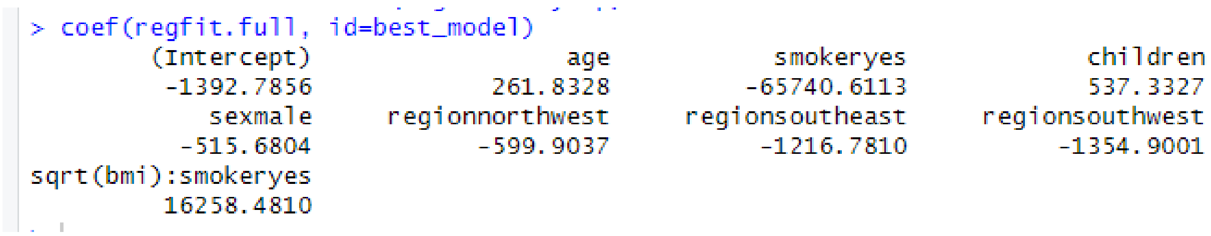
The code performs best subset selection regression analysis on the dataset, exploring the relationship between charges and predictor variables including age, sqrt(bmi), smoker status, number of children, sex, and region. The selection algorithm evaluates models with up to nine variables, highlighting combinations that minimize the Cp statistic, indicating the optimal model complexity.

The R-squared values progressively increase with model complexity, reaching a maximum of 0.8414, suggesting improved explanatory power. The Cp values decrease as the number of variables increases, with the lowest Cp value of 9.41 observed for the model with eight variables. Similarly, the adjusted R-squared values increase with model complexity, indicating better model fit.

Coefficients of the Best Model

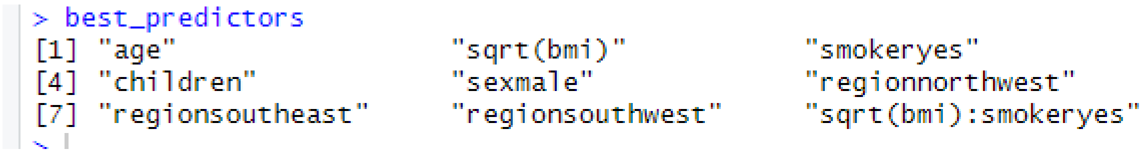


```
#Coefficients of the best model
best_model = which.min(reg.summary$cp)
coef(regfit.full, id=best_model)

# Extract predictors from the best model (e.g., the one with the lowest RSS)
best_model_index <- which.min(reg.summary$cp)
best_predictors <- names(reg.summary$outmat[best_model_index, ])
best_predictors
```



```
> coef(regfit.full, id=best_model)
        (Intercept)                 age           smokeryes            children 
         -1392.7856            261.8328         -65740.6113            537.3327 
            sexmale     regionnorthwest     regionsoutheast     regionsouthwest 
          -515.6804           -599.9037          -1216.7810          -1354.9001 
sqrt(bmi):smokeryes 
         16258.4810 
```

The Best predictors from the best model (e.g., the one with the lowest CP)



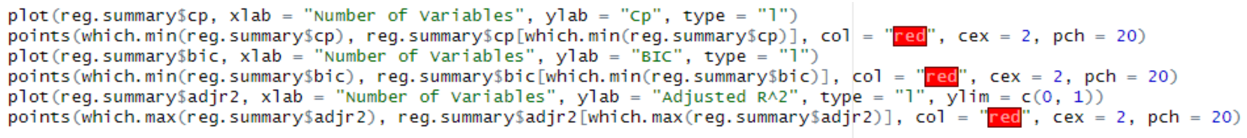
```
> best_predictors
[1] "age"                  "sqrt(bmi)"            "smokeryes"           
[4] "children"             "sexmale"              "regionnorthwest"     
[7] "regionsoutheast"      "regionsouthwest"      "sqrt(bmi):smokeryes"
```

The best model based is identified on the minimum Cp value and extracts its coefficients, revealing that the intercept is -1392.79, with age, smoker status, children, sex, and region variables influencing charges positively or negatively.

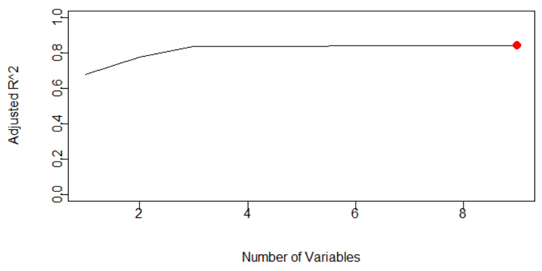
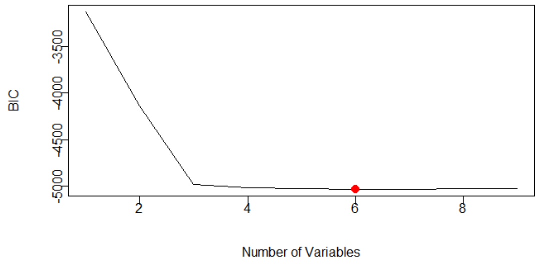
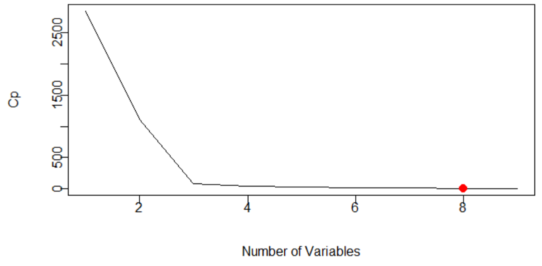
The interaction between sqrt(bmi) and smoker status has a substantial positive effect on charges.

The predictors included in the best model, showcasing age, sqrt(bmi), smokeryes, children, sexmale, and region variables, along with the interaction between sqrt(bmi) and smokeryes, as the most influential factors in predicting charges.

CP/ BIC/ Adjusted R Squared



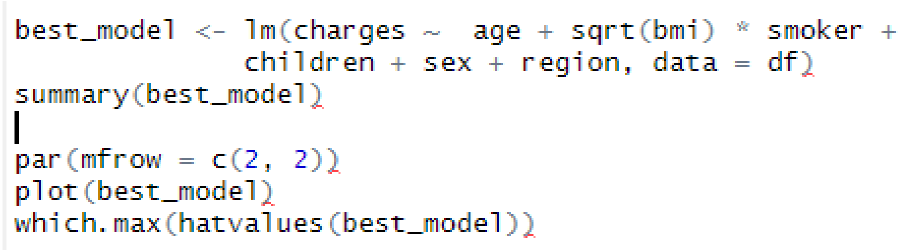
```
plot(reg.summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
points(which.min(reg.summary$cp), reg.summary$cp[which.min(reg.summary$cp)], col = "red", cex = 2, pch = 20)
plot(reg.summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
points(which.min(reg.summary$bic), reg.summary$bic[which.min(reg.summary$bic)], col = "red", cex = 2, pch = 20)
plot(reg.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2", type = "l", ylim = c(0, 1))
points(which.max(reg.summary$adjr2), reg.summary$adjr2[which.max(reg.summary$adjr2)], col = "red", cex = 2, pch = 20)
```

	The x-axis in all three graphs is the number of variables. The y-axis in the first graph is the adjusted R-squared. The y-axis in the second graph is labeled BIC, which may refer to the Bayesian information criterion and the y-axis in the third graph is CP. The first graph shows a positive correlation between the number of variables and the adjusted R-squared, which means that as the number of variables increases, the adjusted R-squared also increases.
	The second graph shows a negative correlation between the number of variables and the BIC, which means that as the number of variables increases, the BIC decreases.
	The third graph is similar to the first graph and shows a positive correlation between the number of variables and the CP value.

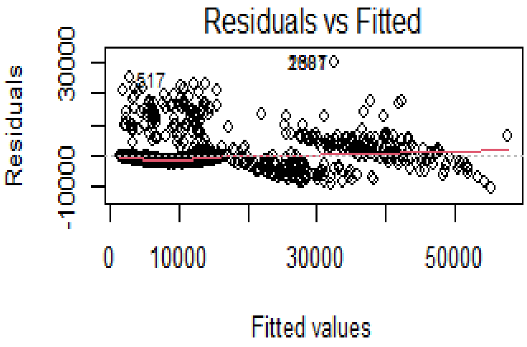
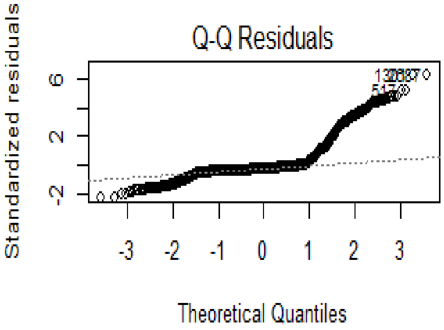
Residual Diagnosis

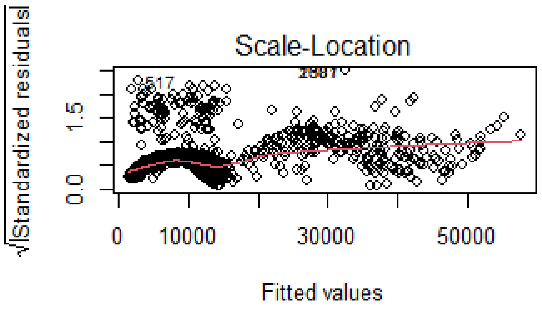
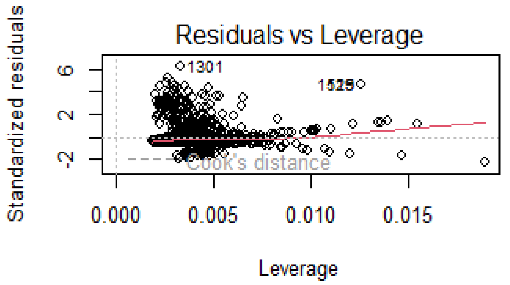
Based on the residual diagnostic plots provided, we can assess the assumptions and potential issues with the linear regression model represented by best_model.

Code:

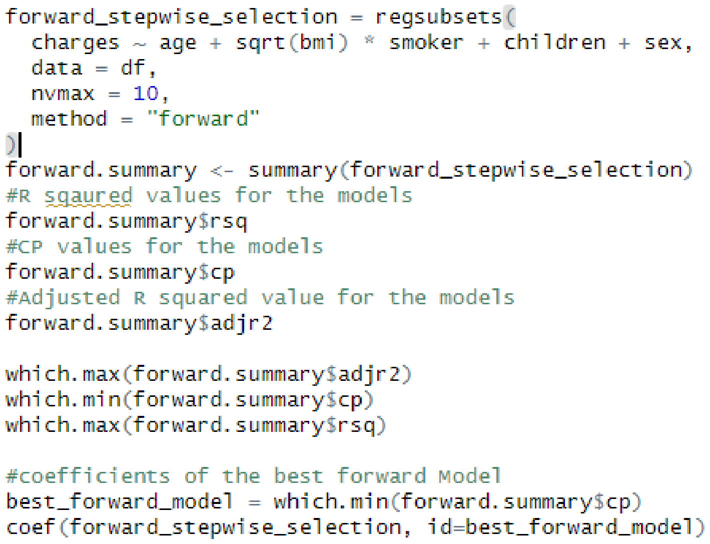


```
best_model <- lm(charges ~  age + sqrt(bmi) * smoker +
                 children + sex + region, data = df)
summary(best_model)

par(mfrow = c(2, 2))
plot(best_model)
which.max(hatvalues(best_model))
```

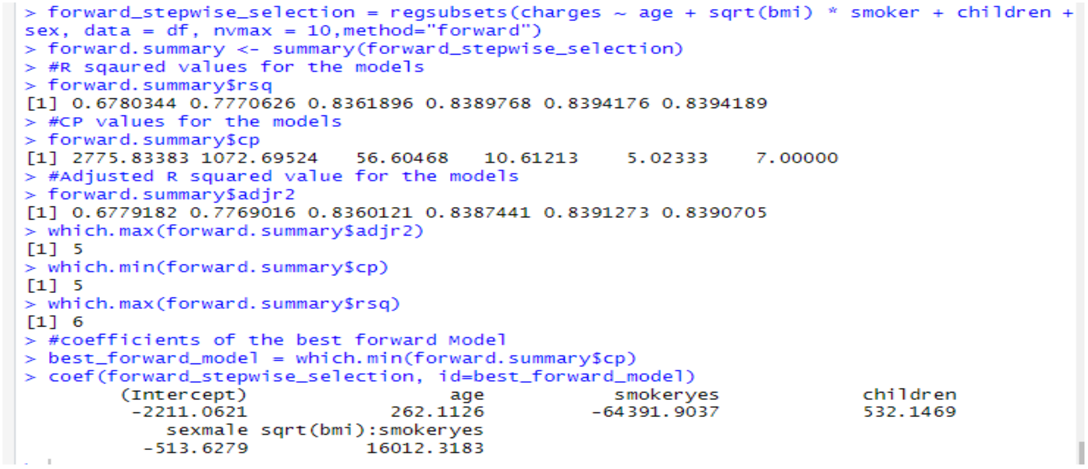
	**Residuals vs Fitted**: This plot shows the residuals (the difference between the observed and predicted values) against the fitted values. The residuals seem to have a slightly funnel-shaped pattern, indicating that the assumption of constant variance (homoscedasticity) may be violated. There is a tendency for the residuals to become more spread out as the fitted values increase.
	**Q-Q Residuals**: The Q-Q plot compares the quantiles of the residuals against the quantiles of a theoretical normal distribution. The residuals deviate from the straight line, especially in the tails, suggesting that the assumption of normality may be violated. The heavy tails indicate the presence of outliers or influential observations.

	**Scale-Location**: This plot examines if the residuals are spread equally along the ranges of predictors. The plot shows a slightly curved pattern, indicating a potential violation of the assumption of constant variance (homoscedasticity) or the presence of non-linear relationships.
	**Residuals vs Leverage**: This plot identifies observations with high leverage (influential data points). There are a few points with relatively high leverage values, indicating the presence of potentially influential observations that may have a disproportionate impact on the model fit.

Forward stepwise selection



```
forward_stepwise_selection = regsubsets(
  charges ~ age + sqrt(bmi) * smoker + children + sex,
  data = df,
  nvmax = 10,
  method = "forward"
)
forward.summary <- summary(forward_stepwise_selection)
#R sqaured values for the models
forward.summary$rsq
#CP values for the models
forward.summary$cp
#Adjusted R squared value for the models
forward.summary$adjr2

which.max(forward.summary$adjr2)
which.min(forward.summary$cp)
which.max(forward.summary$rsq)

#coefficients of the best forward Model
best_forward_model = which.min(forward.summary$cp)
coef(forward_stepwise_selection, id=best_forward_model)
```

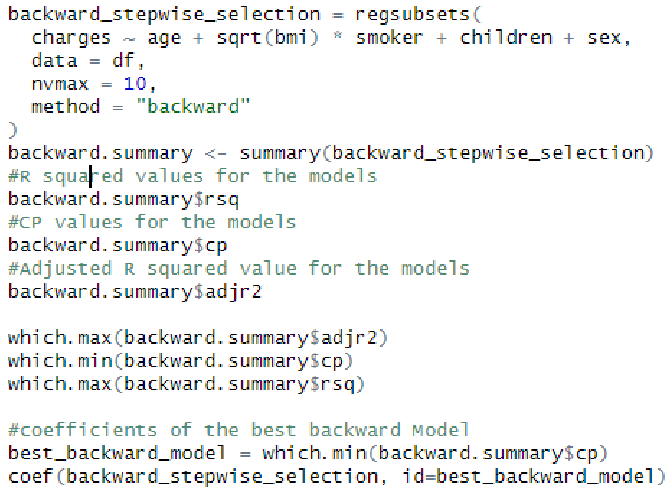


```
> forward_stepwise_selection = regsubsets(charges ~ age + sqrt(bmi) * smoker + children +
sex, data = df, nvmax = 10,method="forward")
> forward.summary <- summary(forward_stepwise_selection)
> #R sqaured values for the models
> forward.summary$rsq
[1] 0.6780344 0.7770626 0.8361896 0.8389768 0.8394176 0.8394189
> #CP values for the models
> forward.summary$cp
[1] 2775.83383 1072.69524   56.60468   10.61213    5.02333    7.00000
> #Adjusted R squared value for the models
> forward.summary$adjr2
[1] 0.6779182 0.7769016 0.8360121 0.8387441 0.8391273 0.8390705
> which.max(forward.summary$adjr2)
[1] 5
> which.min(forward.summary$cp)
[1] 5
> which.max(forward.summary$rsq)
[1] 6
> #coefficients of the best forward Model
> best_forward_model = which.min(forward.summary$cp)
> coef(forward_stepwise_selection, id=best_forward_model)
        (Intercept)                 age           smokeryes             children
         -2211.0621            262.1126         -64391.9037             532.1469
            sexmale sqrt(bmi):smokeryes
          -513.6279          16012.3183
```

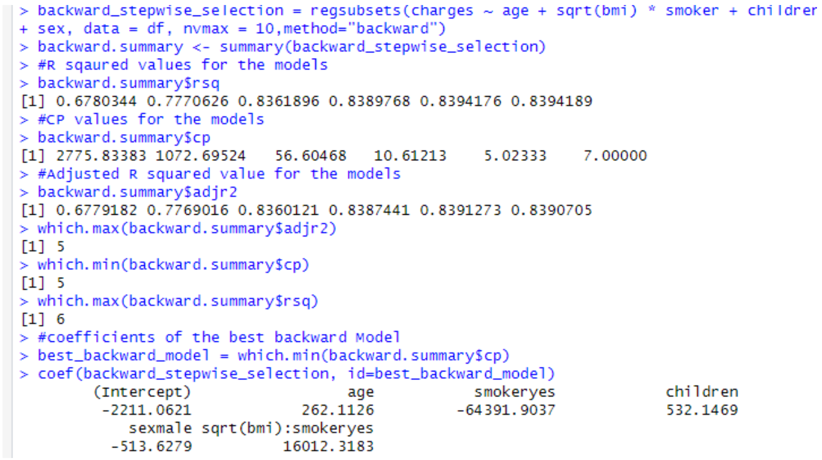
The forward stepwise selection identifies the best model for predicting charges based on a subset of variables. It starts with an empty model and adds predictors one by one, selecting the best subset of variables that optimizes a criterion, in this case, Cp.

The R-squared values for the models range from 0.678 to 0.839, indicating the proportion of variance in charges explained by the predictors. The Cp values decrease as more predictors are added, with the best model having a Cp value of 5.02. The adjusted R-squared values also show improvement with the addition of predictors. The model with the highest adjusted R-squared value includes five predictors and achieves an adjusted R-squared of 0.839.

Backward Stepwise Selection



```
backward_stepwise_selection = regsubsets(
  charges ~ age + sqrt(bmi) * smoker + children + sex,
  data = df,
  nvmax = 10,
  method = "backward"
)
backward.summary <- summary(backward_stepwise_selection)
#R squared values for the models
backward.summary$rsq
#CP values for the models
backward.summary$cp
#Adjusted R squared value for the models
backward.summary$adjr2

which.max(backward.summary$adjr2)
which.min(backward.summary$cp)
which.max(backward.summary$rsq)

#coefficients of the best backward Model
best_backward_model = which.min(backward.summary$cp)
coef(backward_stepwise_selection, id=best_backward_model)
```



```
> backward_stepwise_selection = regsubsets(charges ~ age + sqrt(bmi) * smoker + childrer
+ sex, data = df, nvmax = 10,method="backward")
> backward.summary <- summary(backward_stepwise_selection)
> #R sqaured values for the models
> backward.summary$rsq
[1] 0.6780344 0.7770626 0.8361896 0.8389768 0.8394176 0.8394189
> #CP values for the models
> backward.summary$cp
[1] 2775.83383 1072.69524   56.60468   10.61213    5.02333    7.00000
> #Adjusted R squared value for the models
> backward.summary$adjr2
[1] 0.6779182 0.7769016 0.8360121 0.8387441 0.8391273 0.8390705
> which.max(backward.summary$adjr2)
[1] 5
> which.min(backward.summary$cp)
[1] 5
> which.max(backward.summary$rsq)
[1] 6
> #coefficients of the best backward Model
> best_backward_model = which.min(backward.summary$cp)
> coef(backward_stepwise_selection, id=best_backward_model)
        (Intercept)                  age            smokeryes             children
         -2211.0621             262.1126          -64391.9037             532.1469
            sexmale sqrt(bmi):smokeryes
          -513.6279           16012.3183
```

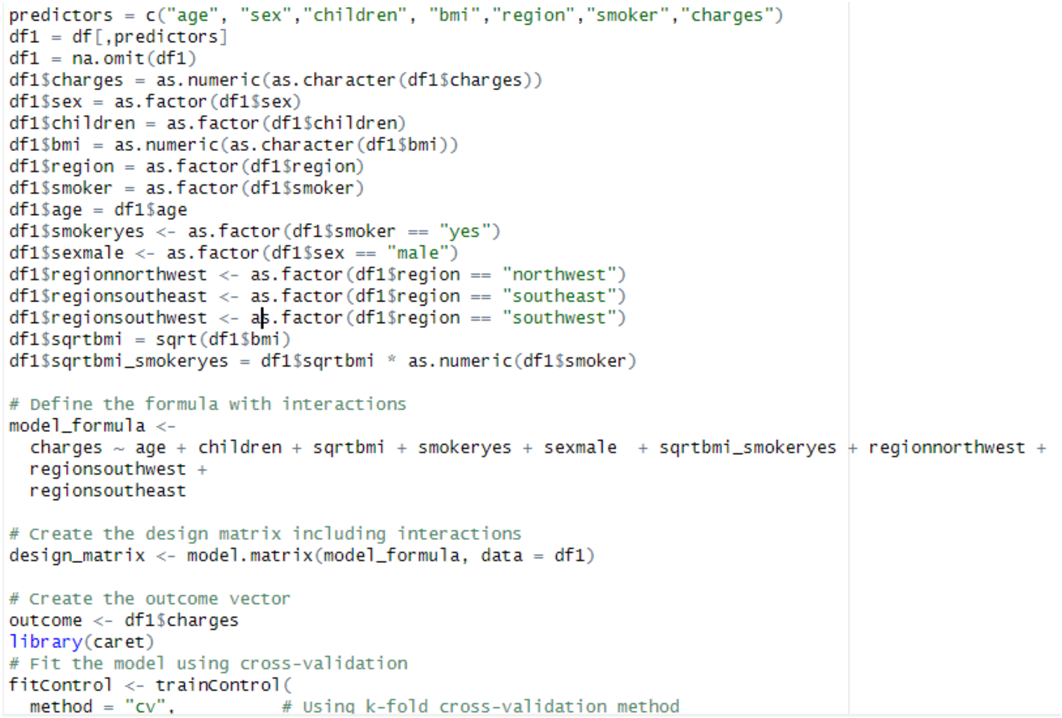
The backward stepwise selection process determines the best model for predicting charges based on a subset of variables. Starting with a full model, it iteratively removes predictors until the optimal subset is achieved.

The R-squared values for the models range from 0.678 to 0.839, indicating the proportion of variance in charges explained by the predictors. The Cp values decrease as predictors are removed, with the best model having a Cp value of 5.02.

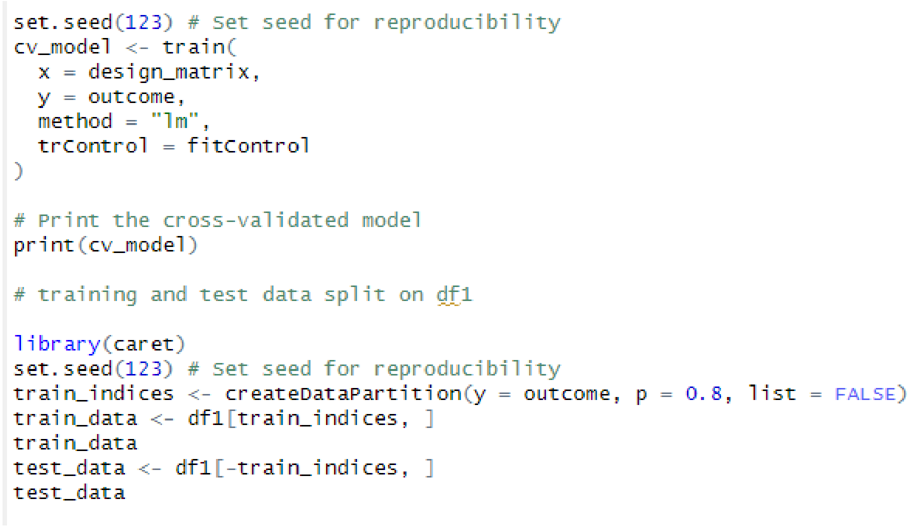
Similarly, the adjusted R-squared values improve with fewer predictors, with the model including five predictors achieving the highest adjusted R-squared of 0.839.

The coefficients of the best backward model reveal the influence of each remaining predictor on charges, highlighting the significance of age, smoker status, number of children, sex, and the interaction between sqrt(bmi) and smoker status.

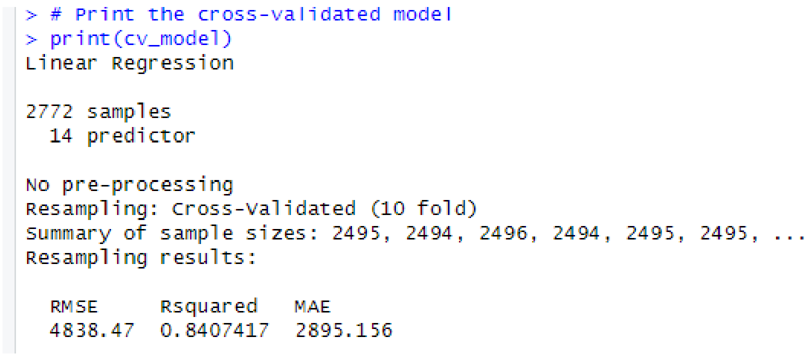
K-fold cross validation



```r
predictors = c("age", "sex","children", "bmi","region","smoker","charges")
df1 = df[,predictors]
df1 = na.omit(df1)
df1$charges = as.numeric(as.character(df1$charges))
df1$sex = as.factor(df1$sex)
df1$children = as.factor(df1$children)
df1$bmi = as.numeric(as.character(df1$bmi))
df1$region = as.factor(df1$region)
df1$smoker = as.factor(df1$smoker)
df1$age = df1$age
df1$smokeryes <- as.factor(df1$smoker == "yes")
df1$sexmale <- as.factor(df1$sex == "male")
df1$regionnorthwest <- as.factor(df1$region == "northwest")
df1$regionsoutheast <- as.factor(df1$region == "southeast")
df1$regionsouthwest <- as.factor(df1$region == "southwest")
df1$sqrtbmi = sqrt(df1$bmi)
df1$sqrtbmi_smokeryes = df1$sqrtbmi * as.numeric(df1$smoker)

# Define the formula with interactions
model_formula <-
  charges ~ age + children + sqrtbmi + smokeryes + sexmale  + sqrtbmi_smokeryes + regionnorthwest +
  regionsouthwest +
  regionsoutheast

# Create the design matrix including interactions
design_matrix <- model.matrix(model_formula, data = df1)

# Create the outcome vector
outcome <- df1$charges
library(caret)
# Fit the model using cross-validation
fitControl <- trainControl(
  method = "cv",           # Using k-fold cross-validation method
```



```r
set.seed(123) # Set seed for reproducibility
cv_model <- train(
  x = design_matrix,
  y = outcome,
  method = "lm",
  trControl = fitControl
)

# Print the cross-validated model
print(cv_model)

# training and test data split on df1

library(caret)
set.seed(123) # Set seed for reproducibility
train_indices <- createDataPartition(y = outcome, p = 0.8, list = FALSE)
train_data <- df1[train_indices, ]
train_data
test_data <- df1[-train_indices, ]
test_data
```



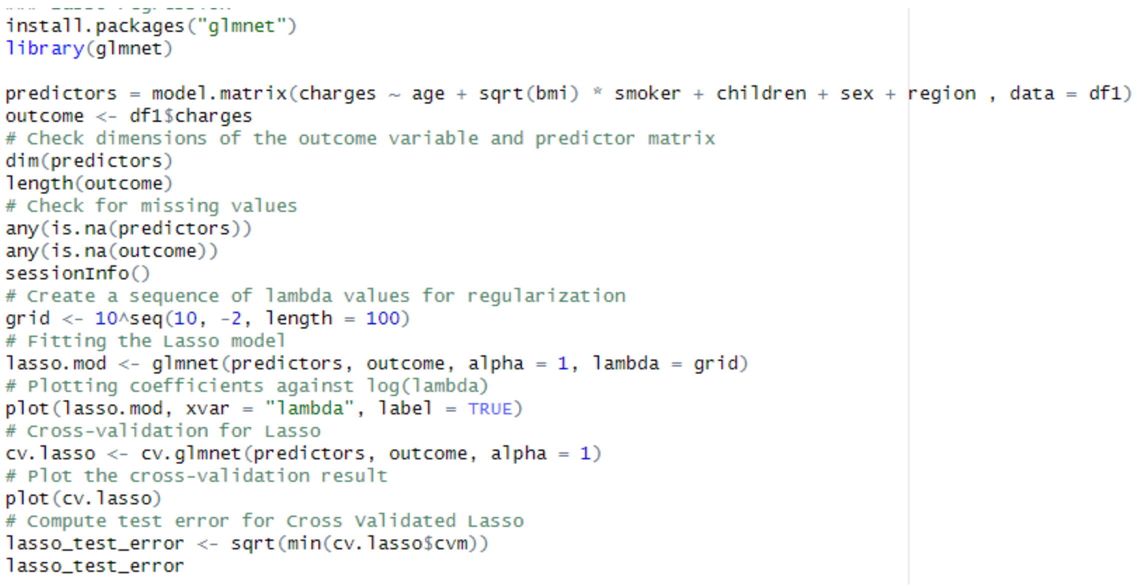
```
> # Print the cross-validated model
> print(cv_model)
Linear Regression

2772 samples
  14 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2495, 2494, 2496, 2494, 2495, 2495, ...
Resampling results:

  RMSE     Rsquared   MAE
  4838.47  0.8407417  2895.156
```

We conducted a linear regression analysis using cross-validation to predict charges based on various predictors such as age, children, bmi, smoker status, sex, and region. The dataset is preprocessed to handle missing values and ensure proper data types. Interactions between predictors are also considered in the model formulation.

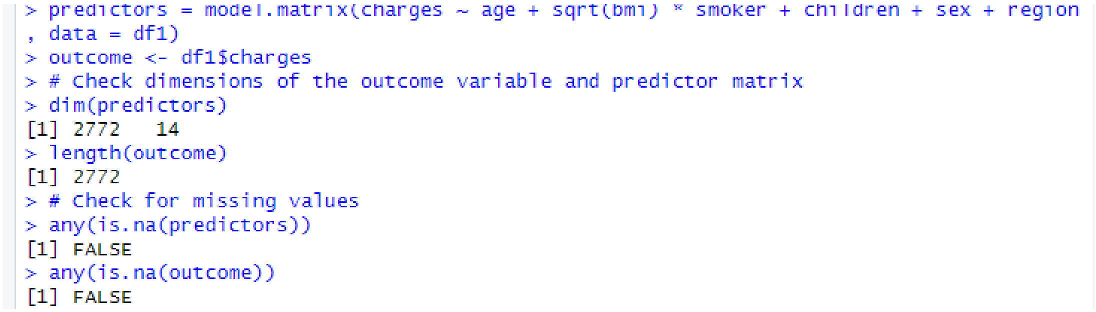
The model is trained using 10-fold cross-validation, where the data is divided into 10 subsets, and the model is trained on nine subsets and validated on the remaining subset iteratively.

The results show that the model has an average root mean squared error (RMSE) of 4838.47, indicating the average difference between the observed and predicted charges. The R-squared value of 0.8407 suggests that approximately 84.07% of the variance in charges is explained by the predictors. Additionally, the mean absolute error (MAE) is 2895.156, indicating the average absolute difference between the observed and predicted charges. Overall, the model appears to perform reasonably well in predicting charges based on the selected predictors.

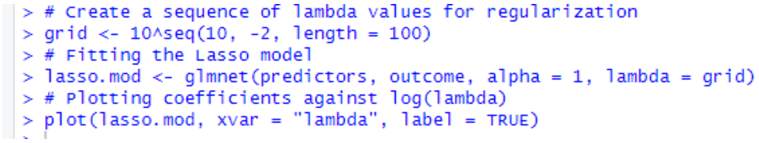
Lasso regression



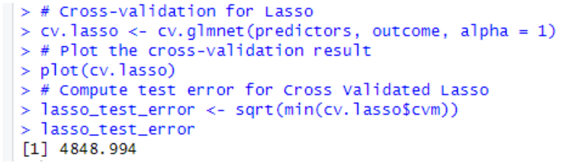
```
install.packages("glmnet")
library(glmnet)

predictors = model.matrix(charges ~ age + sqrt(bmi) * smoker + children + sex + region , data = df1)
outcome <- df1$charges
# Check dimensions of the outcome variable and predictor matrix
dim(predictors)
length(outcome)
# Check for missing values
any(is.na(predictors))
any(is.na(outcome))
sessionInfo()
# Create a sequence of lambda values for regularization
grid <- 10^seq(10, -2, length = 100)
# Fitting the Lasso model
lasso.mod <- glmnet(predictors, outcome, alpha = 1, lambda = grid)
# Plotting coefficients against log(lambda)
plot(lasso.mod, xvar = "lambda", label = TRUE)
# Cross-validation for Lasso
cv.lasso <- cv.glmnet(predictors, outcome, alpha = 1)
# Plot the cross-validation result
plot(cv.lasso)
# Compute test error for Cross Validated Lasso
lasso_test_error <- sqrt(min(cv.lasso$cvm))
lasso_test_error
```



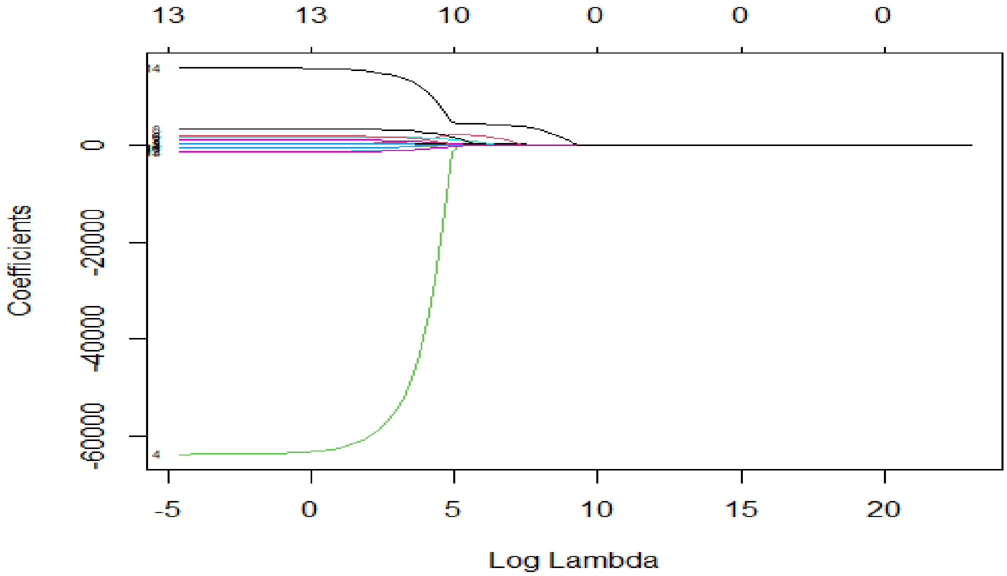
```
> predictors = model.matrix(charges ~ age + sqrt(bmi) * smoker + children + sex + region
, data = df1)
> outcome <- df1$charges
> # Check dimensions of the outcome variable and predictor matrix
> dim(predictors)
[1] 2772   14
> length(outcome)
[1] 2772
> # Check for missing values
> any(is.na(predictors))
[1] FALSE
> any(is.na(outcome))
[1] FALSE
```

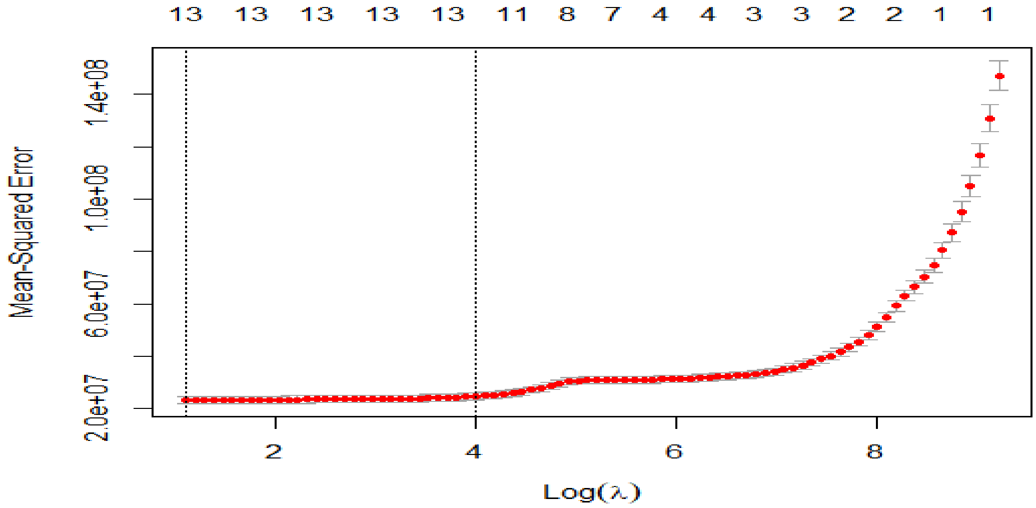


```
> # Create a sequence of lambda values for regularization
> grid <- 10^seq(10, -2, length = 100)
> # Fitting the Lasso model
> lasso.mod <- glmnet(predictors, outcome, alpha = 1, lambda = grid)
> # Plotting coefficients against log(lambda)
> plot(lasso.mod, xvar = "lambda", label = TRUE)
```



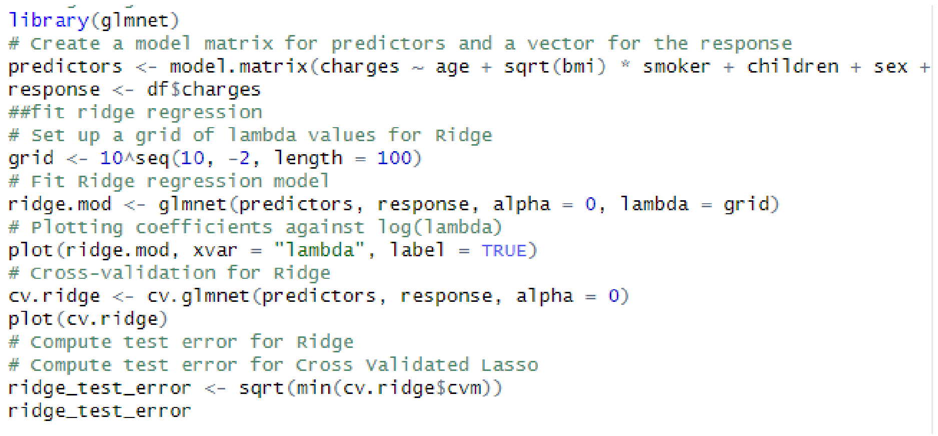
```
> # Cross-validation for Lasso
> cv.lasso <- cv.glmnet(predictors, outcome, alpha = 1)
> # Plot the cross-validation result
> plot(cv.lasso)
> # Compute test error for Cross Validated Lasso
> lasso_test_error <- sqrt(min(cv.lasso$cvm))
> lasso_test_error
[1] 4848.994
```



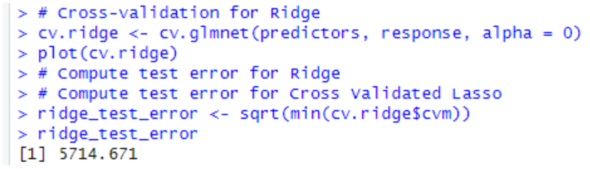


The test error result of approximately 4848.994 for the Cross Validated Lasso model indicates the average deviation between its predictions and the actual outcome values. Derived through cross-validation, this test error provides a robust estimate of the model's performance on unseen data, enhancing its reliability in real-world applications. Overall, the Lasso model demonstrates a certain level of effectiveness in estimating the outcome variable based on the given predictors.

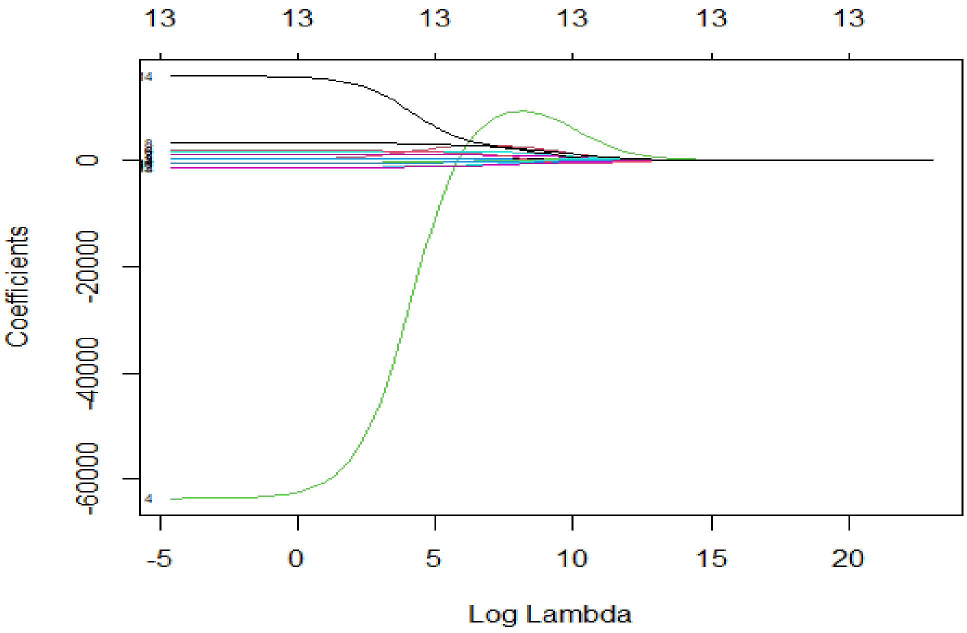
Ridge Regression

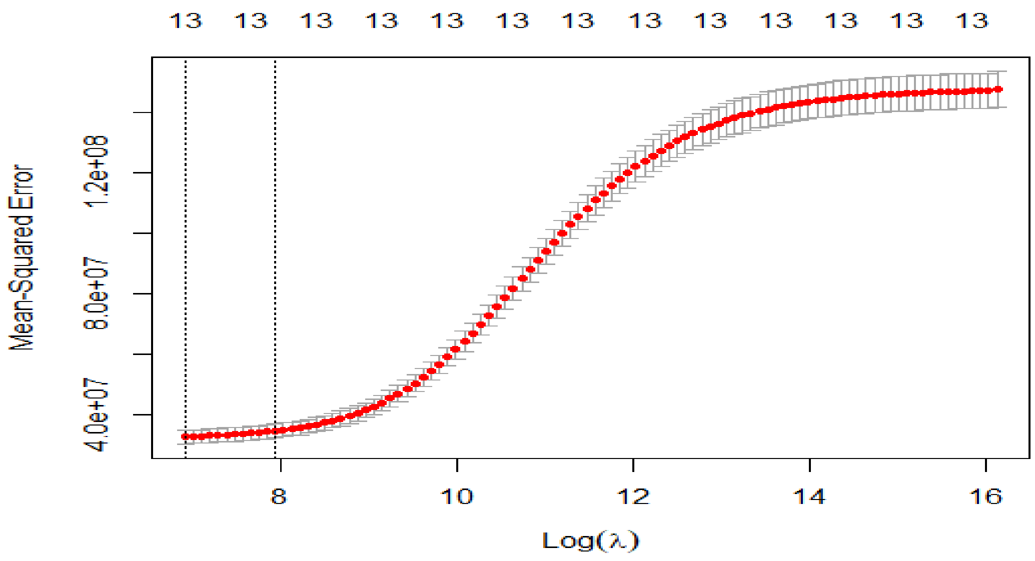


```
library(glmnet)
# Create a model matrix for predictors and a vector for the response
predictors <- model.matrix(charges ~ age + sqrt(bmi) * smoker + children + sex + region, data = df1)
response <- df$charges
##fit ridge regression
# Set up a grid of lambda values for Ridge
grid <- 10^seq(10, -2, length = 100)
# Fit Ridge regression model
ridge.mod <- glmnet(predictors, response, alpha = 0, lambda = grid)
# Plotting coefficients against log(lambda)
plot(ridge.mod, xvar = "lambda", label = TRUE)
# Cross-validation for Ridge
cv.ridge <- cv.glmnet(predictors, response, alpha = 0)
plot(cv.ridge)
# Compute test error for Ridge
# Compute test error for Cross Validated Lasso
ridge_test_error <- sqrt(min(cv.ridge$cvm))
ridge_test_error
```



```
> # Cross-validation for Ridge
> cv.ridge <- cv.glmnet(predictors, response, alpha = 0)
> plot(cv.ridge)
> # Compute test error for Ridge
> # Compute test error for Cross Validated Lasso
> ridge_test_error <- sqrt(min(cv.ridge$cvm))
> ridge_test_error
[1] 5714.671
```





The obtained result of 5714.671 for the test error of the Cross Validated Ridge regression model indicates the root mean squared error (RMSE) of the model's predictions compared to the actual charges.The cross-validation process ensures the robustness of the test error estimate, enhancing the reliability of the Ridge regression model in predicting charges. While the Ridge regression model demonstrates effectiveness in estimating charges based on the provided predictors.

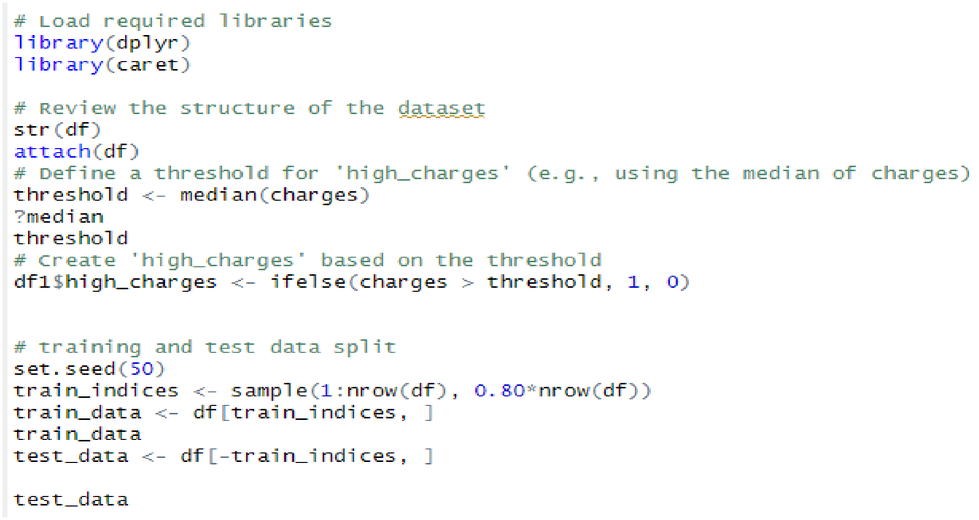
Classification

Our main goal is to classify policyholders into high and low medical expense groups based on demographic and lifestyle factors . So, we performed classification analyses to see how machine learning models accurately classify policyholders as high or low medical expense individuals using available data and by including different models such as Logistic Regression, K-Nearest Neighbours and Linear Discriminant Analysis

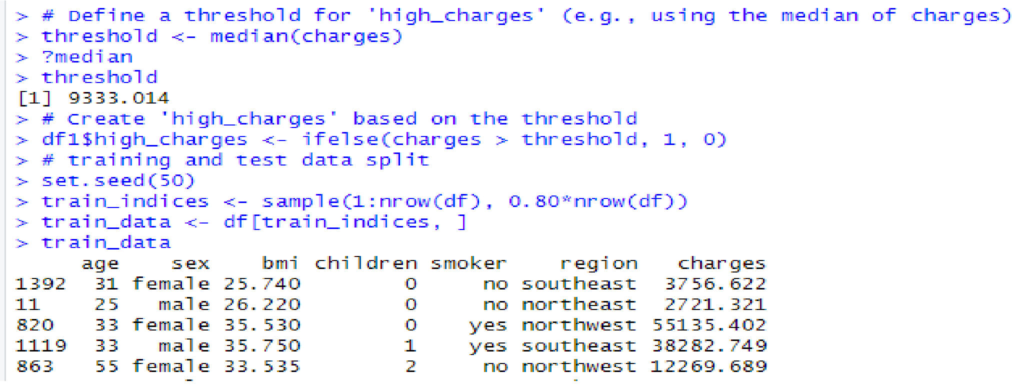
For all the classification models we created a separate subset/dataset named df1 and inserted all the predictors that we got from the best subset selection model.

Since, we have to classify policy holders based into high and low medical groups. So, for the creation of a new dummy variable called high_charges we defined a threshold for 'high_charges' (e.g., using the median of charges) & created a 'high_charges' target variable based on the threshold.

Predictor terms: age, sqrt(bmi) * smoker, children, sex, region

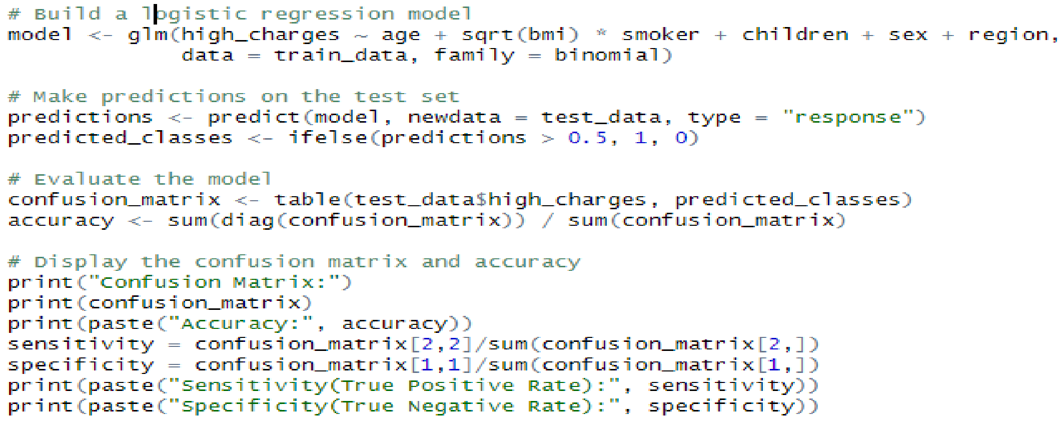


```
# Load required libraries
library(dplyr)
library(caret)

# Review the structure of the dataset
str(df)
attach(df)
# Define a threshold for 'high_charges' (e.g., using the median of charges)
threshold <- median(charges)
?median
threshold
# Create 'high_charges' based on the threshold
df1$high_charges <- ifelse(charges > threshold, 1, 0)


# training and test data split
set.seed(50)
train_indices <- sample(1:nrow(df), 0.80*nrow(df))
train_data <- df[train_indices, ]
train_data
test_data <- df[-train_indices, ]

test_data
```

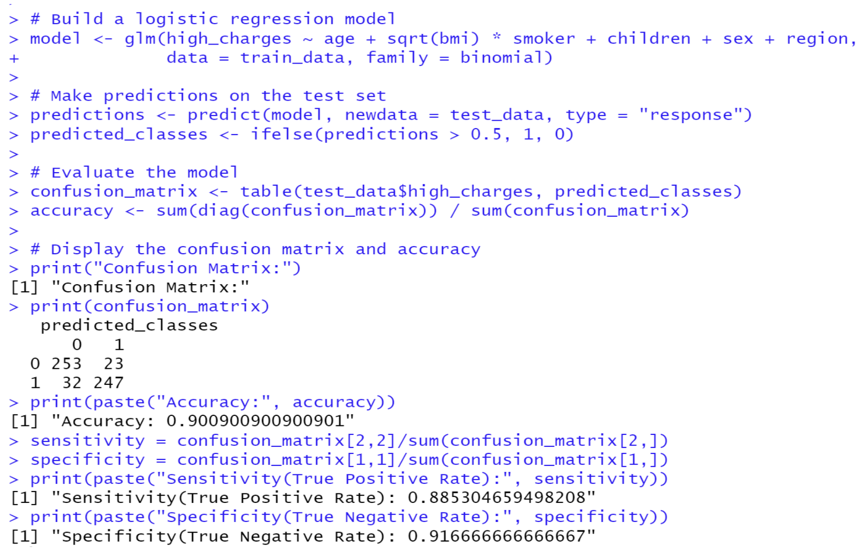


```
> # Define a threshold for 'high_charges' (e.g., using the median of charges)
> threshold <- median(charges)
> ?median
> threshold
[1] 9333.014
> # Create 'high_charges' based on the threshold
> df1$high_charges <- ifelse(charges > threshold, 1, 0)
> # training and test data split
> set.seed(50)
> train_indices <- sample(1:nrow(df), 0.80*nrow(df))
> train_data <- df[train_indices, ]
> train_data
     age    sex    bmi children smoker    region   charges
1392  31 female 25.740        0     no southeast  3756.622
11    25   male 26.220        0     no northeast  2721.321
820   33 female 35.530        0    yes northwest 55135.402
1119  33   male 35.750        1    yes southeast 38282.749
863   55 female 33.535        2     no northwest 12269.689
```

Logistic Regression



```
# Build a logistic regression model
model <- glm(high_charges ~ age + sqrt(bmi) * smoker + children + sex + region,
             data = train_data, family = binomial)

# Make predictions on the test set
predictions <- predict(model, newdata = test_data, type = "response")
predicted_classes <- ifelse(predictions > 0.5, 1, 0)

# Evaluate the model
confusion_matrix <- table(test_data$high_charges, predicted_classes)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)

# Display the confusion matrix and accuracy
print("Confusion Matrix:")
print(confusion_matrix)
print(paste("Accuracy:", accuracy))
sensitivity = confusion_matrix[2,2]/sum(confusion_matrix[2,])
specificity = confusion_matrix[1,1]/sum(confusion_matrix[1,])
print(paste("Sensitivity(True Positive Rate):", sensitivity))
print(paste("Specificity(True Negative Rate):", specificity))
```



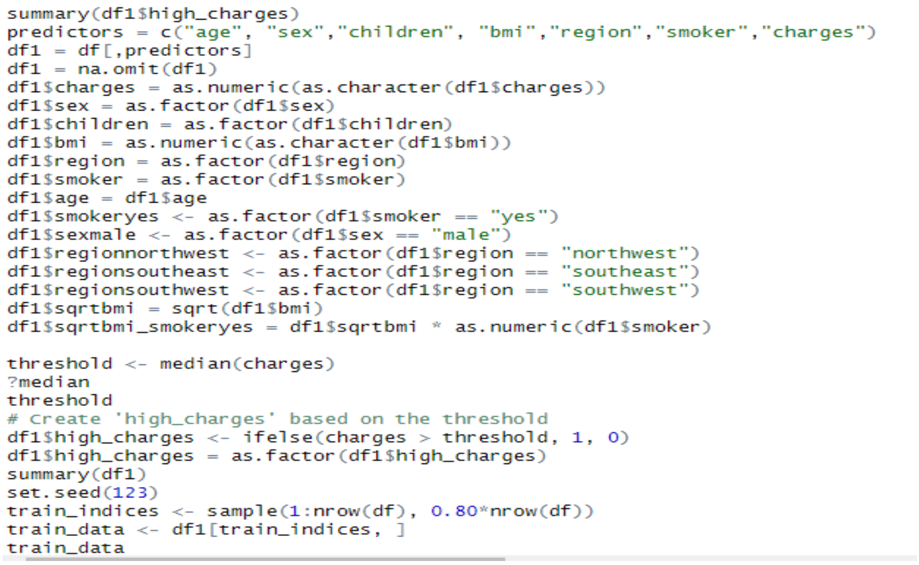
```
> # Build a logistic regression model
> model <- glm(high_charges ~ age + sqrt(bmi) * smoker + children + sex + region,
+              data = train_data, family = binomial)
>
> # Make predictions on the test set
> predictions <- predict(model, newdata = test_data, type = "response")
> predicted_classes <- ifelse(predictions > 0.5, 1, 0)
>
> # Evaluate the model
> confusion_matrix <- table(test_data$high_charges, predicted_classes)
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
>
> # Display the confusion matrix and accuracy
> print("Confusion Matrix:")
[1] "Confusion Matrix:"
> print(confusion_matrix)
   predicted_classes
      0   1
  0 253  23
  1  32 247
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.900900900900901"
> sensitivity = confusion_matrix[2,2]/sum(confusion_matrix[2,])
> specificity = confusion_matrix[1,1]/sum(confusion_matrix[1,])
> print(paste("Sensitivity(True Positive Rate):", sensitivity))
[1] "Sensitivity(True Positive Rate): 0.885304659498208"
> print(paste("Specificity(True Negative Rate):", specificity))
[1] "Specificity(True Negative Rate): 0.916666666666667"
```

The logistic regression model constructed in the provided code employs predictors such as age, a combination of square root of BMI and smoker status, children, sex, and region to predict high charges. Upon making predictions on the test dataset, the model achieves an accuracy of approximately 90.09%. Further evaluation through the confusion matrix reveals that out of 556 instances, there are 253 true negatives and 247 true positives, along with 23 false positives and 32 false negatives.

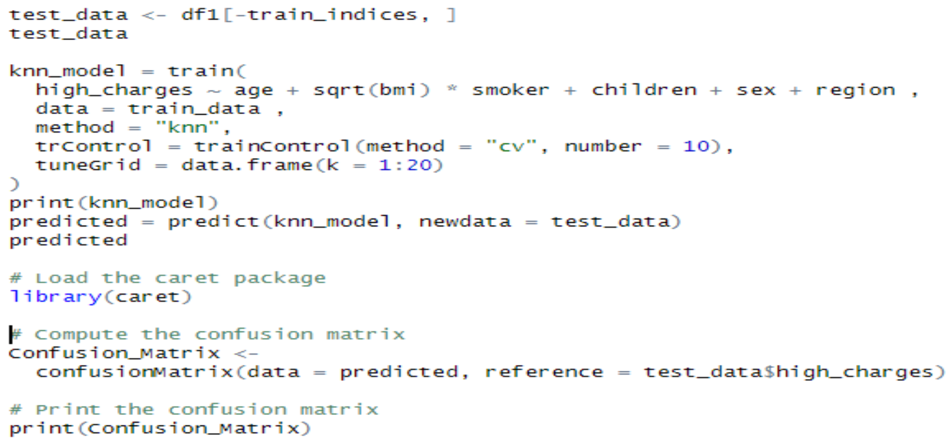
This indicates a reasonably balanced performance in correctly identifying both positive and negative cases. The model exhibits a sensitivity (true positive rate) of around 88.53%, implying its capability to accurately detect high charges when they occur.

Additionally, it demonstrates a specificity (true negative rate) of approximately 91.67%, suggesting its effectiveness in correctly identifying instances of low charges. Overall, the model showcases promising performance with high accuracy and balanced sensitivity and specificity, indicating its potential utility in predicting high charges in the given context.

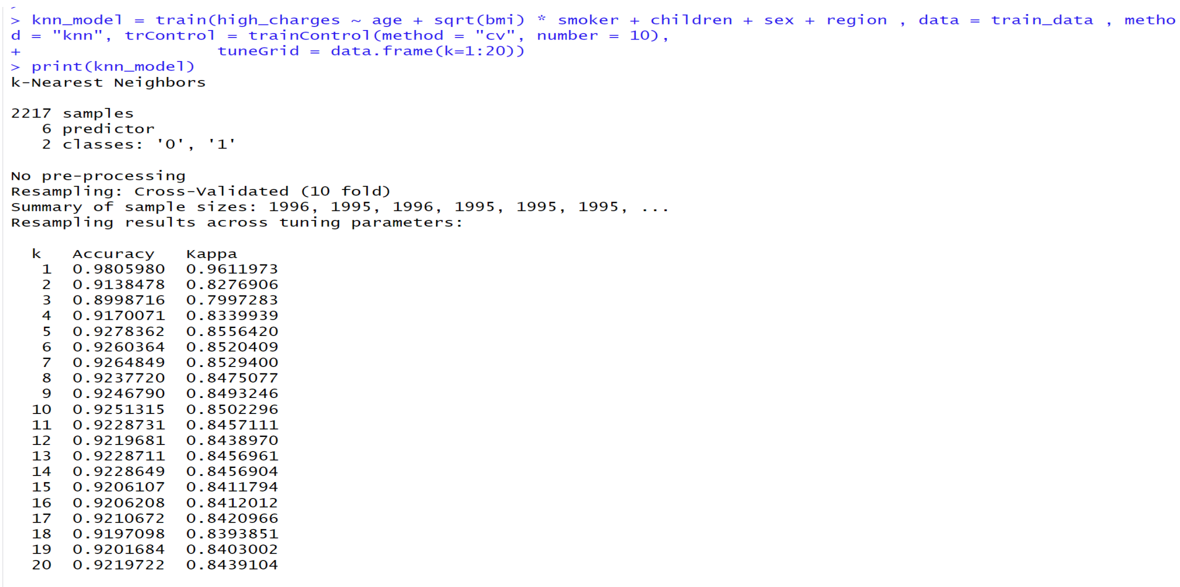
KNN



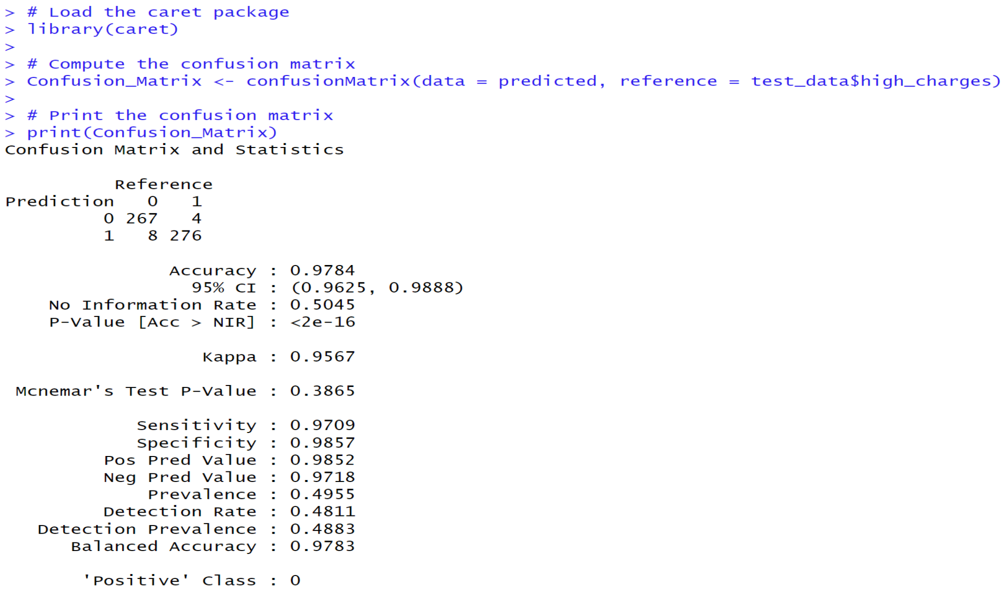
```
summary(df1$high_charges)
predictors = c("age", "sex","children", "bmi","region","smoker","charges")
df1 = df[,predictors]
df1 = na.omit(df1)
df1$charges = as.numeric(as.character(df1$charges))
df1$sex = as.factor(df1$sex)
df1$children = as.factor(df1$children)
df1$bmi = as.numeric(as.character(df1$bmi))
df1$region = as.factor(df1$region)
df1$smoker = as.factor(df1$smoker)
df1$age = df1$age
df1$smokeryes <- as.factor(df1$smoker == "yes")
df1$sexmale <- as.factor(df1$sex == "male")
df1$regionnorthwest <- as.factor(df1$region == "northwest")
df1$regionsoutheast <- as.factor(df1$region == "southeast")
df1$regionsouthwest <- as.factor(df1$region == "southwest")
df1$sqrtbmi = sqrt(df1$bmi)
df1$sqrtbmi_smokeryes = df1$sqrtbmi * as.numeric(df1$smoker)

threshold <- median(charges)
?median
threshold
# Create 'high_charges' based on the threshold
df1$high_charges <- ifelse(charges > threshold, 1, 0)
df1$high_charges = as.factor(df1$high_charges)
summary(df1)
set.seed(123)
train_indices <- sample(1:nrow(df), 0.80*nrow(df))
train_data <- df1[train_indices, ]
train_data
```



```
test_data <- df1[-train_indices, ]
test_data

knn_model = train(
  high_charges ~ age + sqrt(bmi) * smoker + children + sex + region ,
  data = train_data ,
  method = "knn",
  trControl = trainControl(method = "cv", number = 10),
  tuneGrid = data.frame(k = 1:20)
)
print(knn_model)
predicted = predict(knn_model, newdata = test_data)
predicted

# Load the caret package
library(caret)

# Compute the confusion matrix
Confusion_Matrix <-
  confusionMatrix(data = predicted, reference = test_data$high_charges)

# Print the confusion matrix
print(Confusion_Matrix)
```



```
> knn_model = train(high_charges ~ age + sqrt(bmi) * smoker + children + sex + region , data = train_data , metho
d = "knn", trControl = trainControl(method = "cv", number = 10),
+                    tuneGrid = data.frame(k=1:20))
> print(knn_model)
k-Nearest Neighbors

2217 samples
   6 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1996, 1995, 1996, 1995, 1995, 1995, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.9805980  0.9611973
   2  0.9138478  0.8276906
   3  0.8998716  0.7997283
   4  0.9170071  0.8339939
   5  0.9278362  0.8556420
   6  0.9260364  0.8520409
   7  0.9264849  0.8529400
   8  0.9237720  0.8475077
   9  0.9246790  0.8493246
  10  0.9251315  0.8502296
  11  0.9228731  0.8457111
  12  0.9219681  0.8438970
  13  0.9228711  0.8456961
  14  0.9228649  0.8456904
  15  0.9206107  0.8411794
  16  0.9206208  0.8412012
  17  0.9210672  0.8420966
  18  0.9197098  0.8393851
  19  0.9201684  0.8403002
  20  0.9219722  0.8439104
```



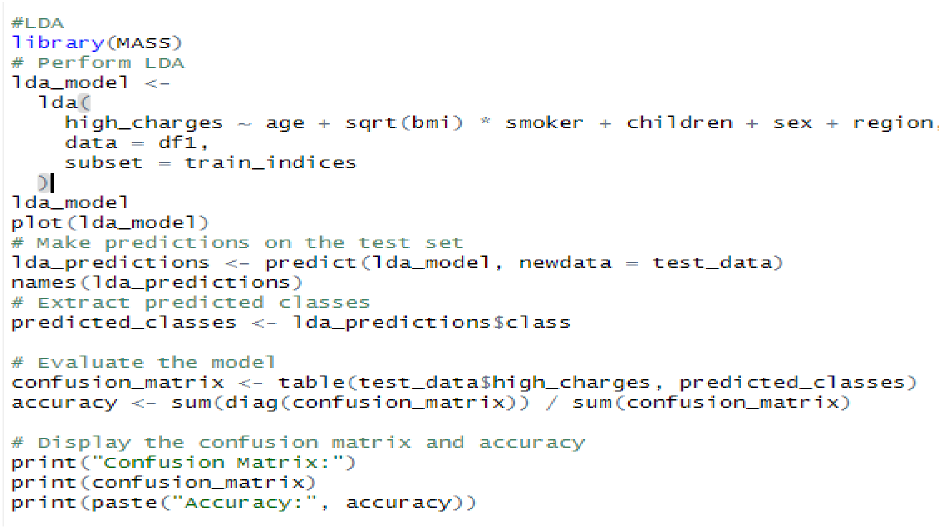
```
> # Load the caret package
> library(caret)
> 
> # Compute the confusion matrix
> Confusion_Matrix <- confusionMatrix(data = predicted, reference = test_data$high_charges)
> 
> # Print the confusion matrix
> print(Confusion_Matrix)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 267   4
         1   8 276
                                          
               Accuracy : 0.9784          
                 95% CI : (0.9625, 0.9888)
    No Information Rate : 0.5045          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.9567          
                                          
 Mcnemar's Test P-Value : 0.3865          
                                          
            Sensitivity : 0.9709          
            Specificity : 0.9857          
         Pos Pred Value : 0.9852          
         Neg Pred Value : 0.9718          
             Prevalence : 0.4955          
         Detection Rate : 0.4811          
   Detection Prevalence : 0.4883          
      Balanced Accuracy : 0.9783          
                                          
       'Positive' Class : 0               
```

The code sets up a K-Nearest Neighbors (KNN) model to predict high charges, using a technique called 10-fold cross-validation to pick the best value for a key setting called k. This k essentially tells the model how many nearby points to consider when making a prediction. It then checks how well the model is doing with different k values, showing us the accuracy and a special statistic called kappa.

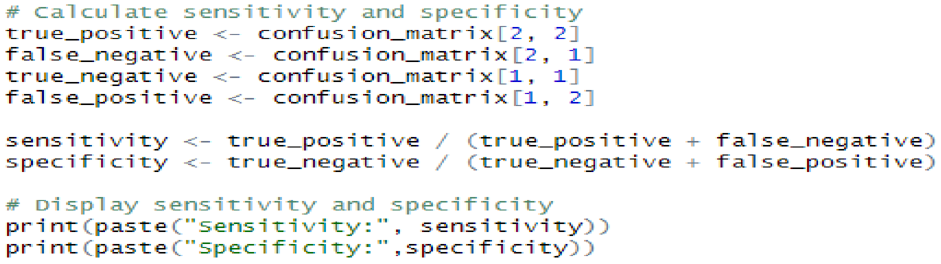
Once the model is all trained up, it's put to the test using a separate set of data.This confusion matrix provides insights into the model's performance, indicating high accuracy (97.84%), sensitivity (97.09%), and specificity (98.57%), showcasing the model's effectiveness in distinguishing between high and low charges.

The KNN model demonstrates robust performance in predicting high charges based on the provided dataset and features.

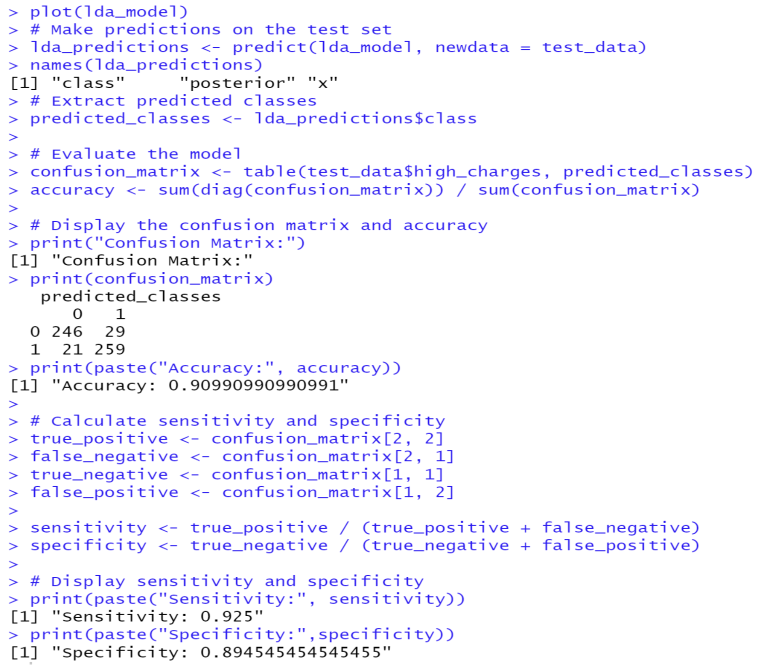
Linear Discriminant Analysis



```
#LDA
library(MASS)
# Perform LDA
lda_model <-
  lda(
    high_charges ~ age + sqrt(bmi) * smoker + children + sex + region,
    data = df1,
    subset = train_indices
  )
lda_model
plot(lda_model)
# Make predictions on the test set
lda_predictions <- predict(lda_model, newdata = test_data)
names(lda_predictions)
# Extract predicted classes
predicted_classes <- lda_predictions$class

# Evaluate the model
confusion_matrix <- table(test_data$high_charges, predicted_classes)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)

# Display the confusion matrix and accuracy
print("Confusion Matrix:")
print(confusion_matrix)
print(paste("Accuracy:", accuracy))
```



```
# Calculate sensitivity and specificity
true_positive <- confusion_matrix[2, 2]
false_negative <- confusion_matrix[2, 1]
true_negative <- confusion_matrix[1, 1]
false_positive <- confusion_matrix[1, 2]

sensitivity <- true_positive / (true_positive + false_negative)
specificity <- true_negative / (true_negative + false_positive)

# Display sensitivity and specificity
print(paste("Sensitivity:", sensitivity))
print(paste("Specificity:",specificity))
```



```
> plot(lda_model)
> # Make predictions on the test set
> lda_predictions <- predict(lda_model, newdata = test_data)
> names(lda_predictions)
[1] "class"     "posterior" "x"
> # Extract predicted classes
> predicted_classes <- lda_predictions$class
>
> # Evaluate the model
> confusion_matrix <- table(test_data$high_charges, predicted_classes)
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
>
> # Display the confusion matrix and accuracy
> print("Confusion Matrix:")
[1] "Confusion Matrix:"
> print(confusion_matrix)
   predicted_classes
      0   1
  0 246  29
  1  21 259
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.90990990990991"
>
> # Calculate sensitivity and specificity
> true_positive <- confusion_matrix[2, 2]
> false_negative <- confusion_matrix[2, 1]
> true_negative <- confusion_matrix[1, 1]
> false_positive <- confusion_matrix[1, 2]
>
> sensitivity <- true_positive / (true_positive + false_negative)
> specificity <- true_negative / (true_negative + false_positive)
>
> # Display sensitivity and specificity
> print(paste("Sensitivity:", sensitivity))
[1] "Sensitivity: 0.925"
> print(paste("Specificity:",specificity))
[1] "Specificity: 0.894545454545455"
```

Linear Discriminant Analysis (LDA) employs to predict high charges based on various predictor variables. Upon training and evaluation using test data, the LDA model exhibits strong performance, achieving an accuracy of approximately 90.99%.

Furthermore, the model demonstrates a sensitivity of about 92.50%, indicating its ability to accurately identify instances of high charges. Additionally, the specificity of approximately 89.45% suggests the model's effectiveness in distinguishing low charges.

Conclusion:

Based on the metrics, the KNN model stands out as the top performer among the three classification models. It has the highest accuracy, meaning it predicts correctly more often overall, and its high Kappa value indicates that its performance is significantly better than randomly guessing.

However, KNN can be more sensitive to the training data, potentially affecting its performance on new data. On the other hand, Logistic Regression offers a good balance between accuracy and interpretability, providing insights into how each feature influences predictions.

Linear Discriminant Analysis (LDA) has a higher sensitivity but lower specificity compared to Logistic Regression, suggesting it might be better at identifying positive cases but may misclassify negative ones more often. In recommendation, if overall accuracy and robustness to new data are key, KNN is the best choice.

Part IV Conclusion and Recommendations

In bringing things to an end, based on the analysis conducted using demographic and lifestyle predictors (age, BMI, smoker status, sex, region, charges, and children) to predict medical expenses and classify policyholders into high and low expense groups, the following conclusions can be drawn:

Conclusion

Our project utilized regression techniques, specifically linear regression, to accurately predict medical expenses. Evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared were employed to assess model accuracy. The achieved MAE and MSE demonstrated satisfactory predictive performance, affirming the reliability of our models in forecasting medical expenses.

Accurate prediction of medical expenses offers valuable insights for insurance companies to refine pricing strategies and risk assessment processes. By understanding policyholders' expected medical expenses, insurers can adjust premiums, coverage, and policy terms accordingly. This data-driven approach empowers insurers to optimize business operations, minimize financial risks, and enhance customer satisfaction.

In addition, classification models were deployed to categorize policyholders into high and low medical expense groups based on demographic and lifestyle attributes. Performance metrics such as accuracy, specificity, and sensitivity were utilized to evaluate these models, with KNN demonstrating superior accuracy and F1-score.

The classification results enable insurers to offer tailored insurance plans and wellness programs to individuals based on their medical expense risk profiles. Identifying high-risk policyholders allows for targeted interventions, preventive care programs, and health management resources to mitigate health risks and reduce medical expenses. This personalized approach not only improves overall health outcomes but also enhances profitability and sustainability.

Our project underscores the significance of data-driven decision-making in healthcare management and insurance operations. By leveraging advanced analytics and machine learning, organizations can glean actionable insights from vast healthcare data, leading to more informed and effective decision-making processes. Integrating predictive modeling and classification techniques empowers healthcare providers and insurers to optimize resource allocation, improve service delivery, and enhance patient outcomes.

Looking ahead, there are ample opportunities for further innovation and improvement in healthcare and insurance. Future research could explore more sophisticated modeling techniques, incorporate additional data sources, and consider dynamic factors such as changing healthcare policies and demographic trends. Embracing emerging technologies and adopting a proactive approach to data-driven decision-making positions organizations to anticipate market trends, address evolving customer needs, and drive continuous improvement in healthcare delivery and insurance management.

Recommendations for business/research decisions

Based on the analysis results, insurance companies or healthcare providers can refine their pricing strategies and risk assessment processes.

Tailored insurance plans or wellness programs could be offered to individuals with higher medical expense risk profiles, such as older adults or smokers.

Targeted interventions and healthcare resources allocation can be planned based on regional variations in medical expenses.

Implementing personalized customer outreach programs to educate policyholders on healthy lifestyle choices and preventive measures, potentially reducing the frequency of high-cost medical treatments and claims.

Utilizing predictive modeling to forecast future medical insurance costs, enabling better budgeting and financial planning for insurance companies.

Collaborating with healthcare providers to promote cost-effective treatments and procedures, thereby minimizing overall medical expenses and insurance claims.

Conducting regular reviews of insurance plans and coverage options to ensure alignment with evolving healthcare trends and regulatory changes.

Investing in technology solutions such as telemedicine platforms or remote patient monitoring systems to enhance access to healthcare services and mitigate medical costs for policyholders.

Engaging in partnerships with employers to offer workplace wellness programs, fostering healthier lifestyles among employees and potentially reducing healthcare expenses for both employers and insurers.

Conducting ongoing research and data analysis to identify emerging trends and factors influencing medical insurance costs, facilitating proactive adjustments to pricing models and risk management strategies.

Part V References

M. R. P. S. (2014, March 5). *Dummy Variables or Indicator Variables in R | R Tutorial 5.5 | MarinStatsLectures*. YouTube. https://www.youtube.com/watch?v=2s8AwoKZ-UE

***kNN function - RDocumentation*. (n.d.). https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/kNN**

G. (2023, August 21). *Feature Selection with the Caret R Package*. GeeksforGeeks. https://www.geeksforgeeks.org/feature-selection-with-the-caret-r-package/

***RPubs - Subset Variable Selection*. (n.d.). https://rpubs.com/Tatjana_Kec/1014475**

Mph, S. P. S. I. (2023, September 25). *Steve's Data Tips and Tricks - Mastering Data Visualization with Pairs Plots in Base R*. https://www.spsanderson.com/steveondata/posts/2023-09-25/index.html

***R Regression Models | Data Science Workshops*. (n.d.). https://iqss.github.io/dss-workshops/Rmodels.html**