514 Final Project

FinalProject_Group15

2023-11-30

Setup and Import of Libraries

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tigerstats)
#install.packages('MLmetrics')
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 4.3.2
```

```
#install.packages("Metrics")
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.3.2
```

```
library(MASS)
```

Part I: Data Preparation

Loading data into “MedicalCostPersonal” dataframe

```
medicalCostPersonal <- read.csv("C:/Users/bharg/OneDrive/Desktop/ITM Assignments/Prog For DA/Final Project/insurance.csv")
```

First 6 rows of the medicalCostPersonal dataframe

```
head(medicalCostPersonal)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

A summary of the distribution of data in each of the variables of medicalCostPersonal dataset

```
summary(medicalCostPersonal)
```

```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   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  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770  
```

Structure of Dataset

```
str(medicalCostPersonal)
```

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

> We can observe from above output that there are 1338 observations (Rows) of 7 variables (Columns) in the dataset.
> The bmi and charges are numerical variables.
> age and children are discrete variable as it takes on whole number values.
> sex, smoker and region are categorical variables.

Checking for anomalies in the dataset

Missing Values

```
missing_values <- colSums(is.na(medicalCostPersonal))
missing_values
```

```
##      age      sex      bmi  children   smoker   region  charges
##        0        0        0         0        0        0        0
```

> No missing values found in the dataset

Analyzing Sex Attribute

```
distinct_sex_values <- unique(medicalCostPersonal$sex)
distinct_sex_values
```

```
## [1] "female" "male"
```

> No abnormal entries in the sex attribute of the dataset.

Analyzing Smoker Attribute

```
distinct_smoke_values <- unique(medicalCostPersonal$smoker)
distinct_smoke_values
```

```
## [1] "yes" "no"
```

> No abnormal entries in the smoker attribute of the dataset.

Analyzing Region Attribute

```
distinct_region_values <- unique(medicalCostPersonal$region)
distinct_region_values
```

```
## [1] "southwest" "southeast" "northwest" "northeast"
```

> No abnormal entries in the region attribute of the dataset.

Analyzing for duplicate rows

```
duplicate_rows <- duplicated(medicalCostPersonal)
duplicate_entry <- medicalCostPersonal[duplicate_rows, ]
duplicate_entry
```

```
##     age  sex   bmi children smoker    region  charges
## 582  19 male 30.59        0     no northwest 1639.563
```

> Record no 582 found to be a duplicate entry.

Eliminating duplicate record and inserting into a new dataframe

```
MCPclean <- medicalCostPersonal[!duplicate_rows,]
total_in_medicalCostPersonal  <- nrow(medicalCostPersonal)
total_in_MCPclean  <- nrow(MCPclean)
```

> The total records in medicalCostPersonal dataset is 1338.
> The total records in MCPclean dataset is 1337.

```
duplicate_rows_count <- sum(duplicated(MCPclean))
duplicate_rows_count
```

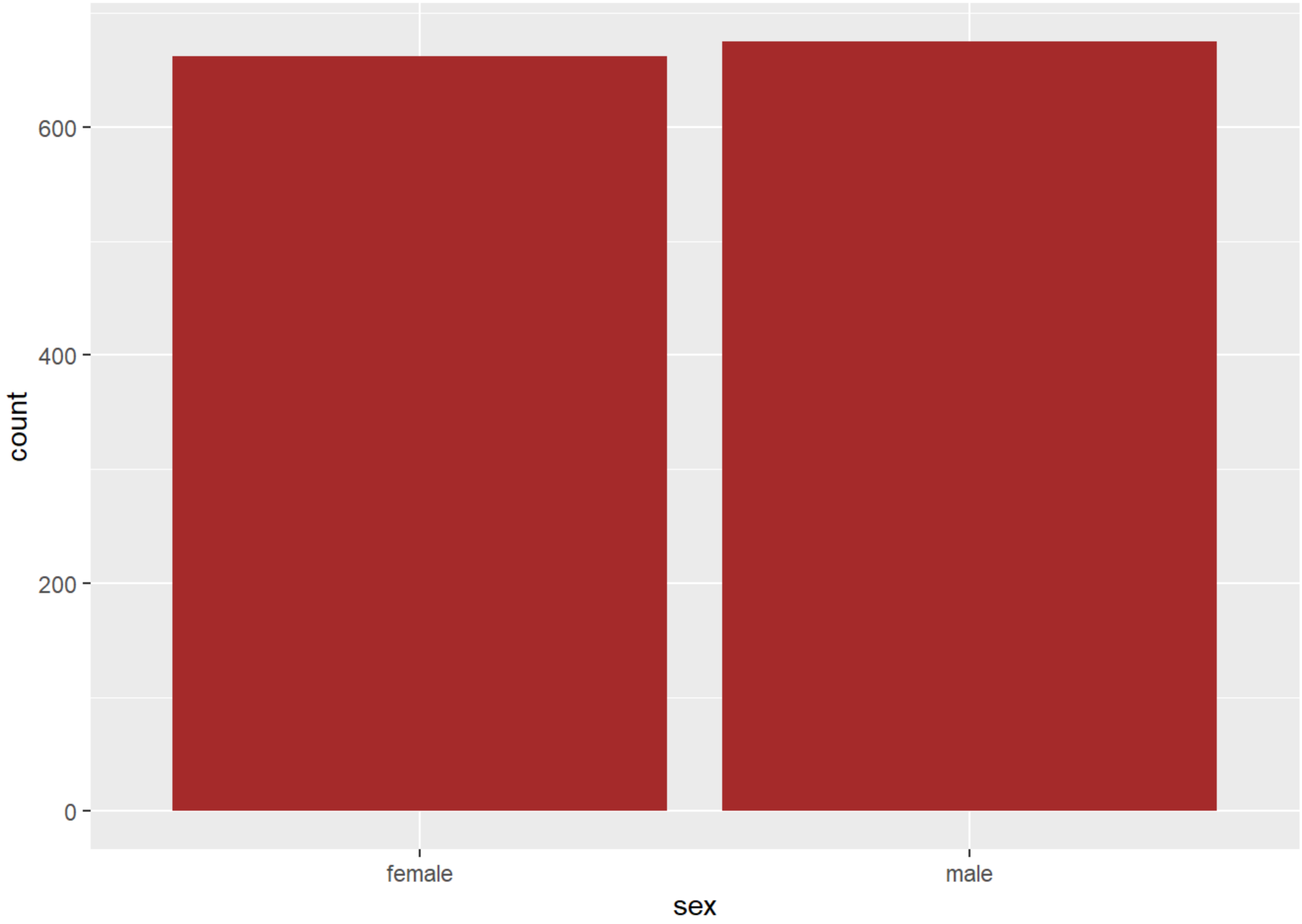
```
## [1] 0
```

> Hence, a new dataframe named ‘MCPclean’ has been created by eliminating the duplicate records from the original database.

Part II: Data Exploration(EDA)

Sex Distribution

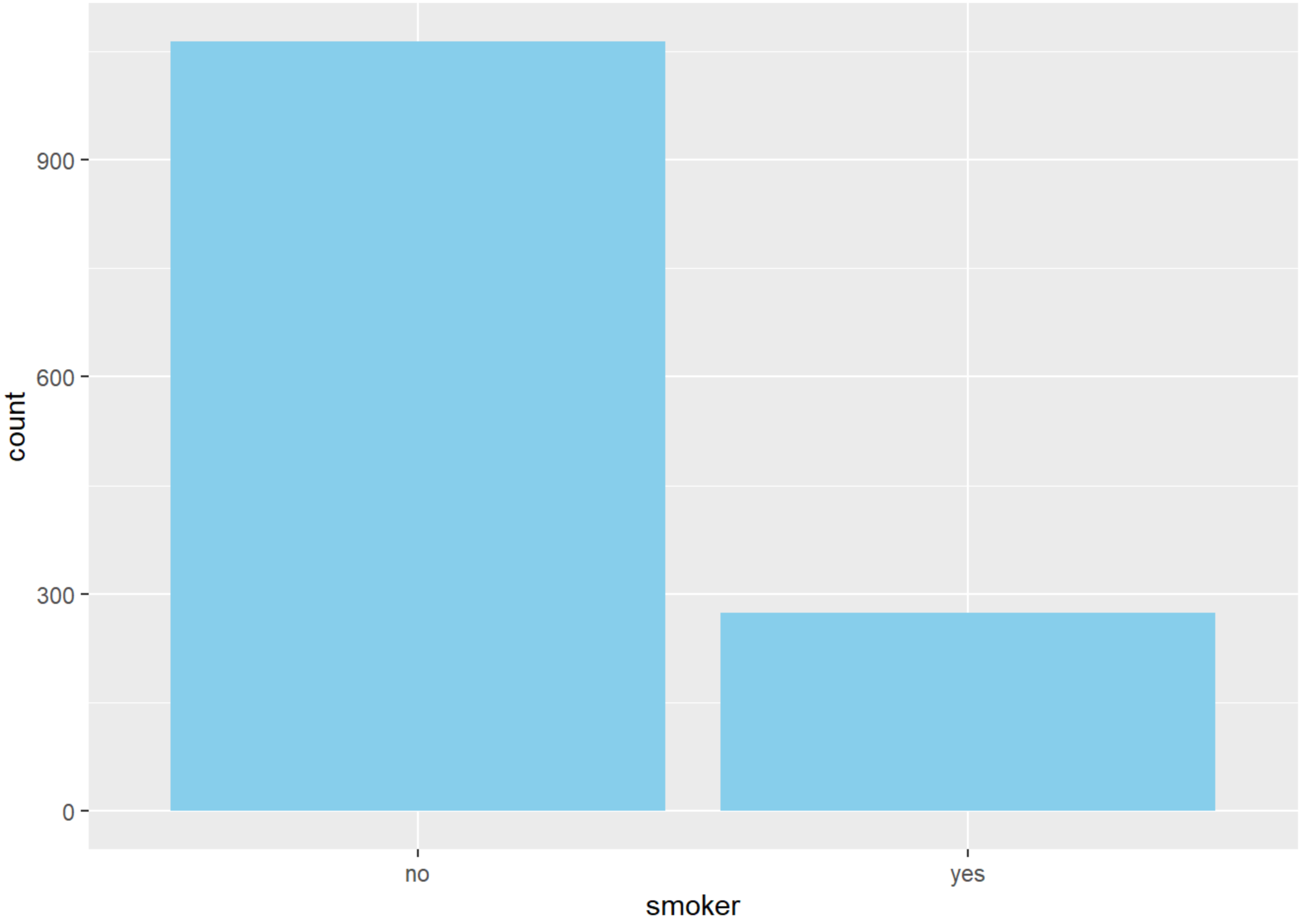
```
ggplot(data = MCPclean) + geom_bar(mapping = aes(x = sex), fill = "Brown")
```



> The distribution of Male and Female genders in the dataset is approximately equal.

Smoker Distribution

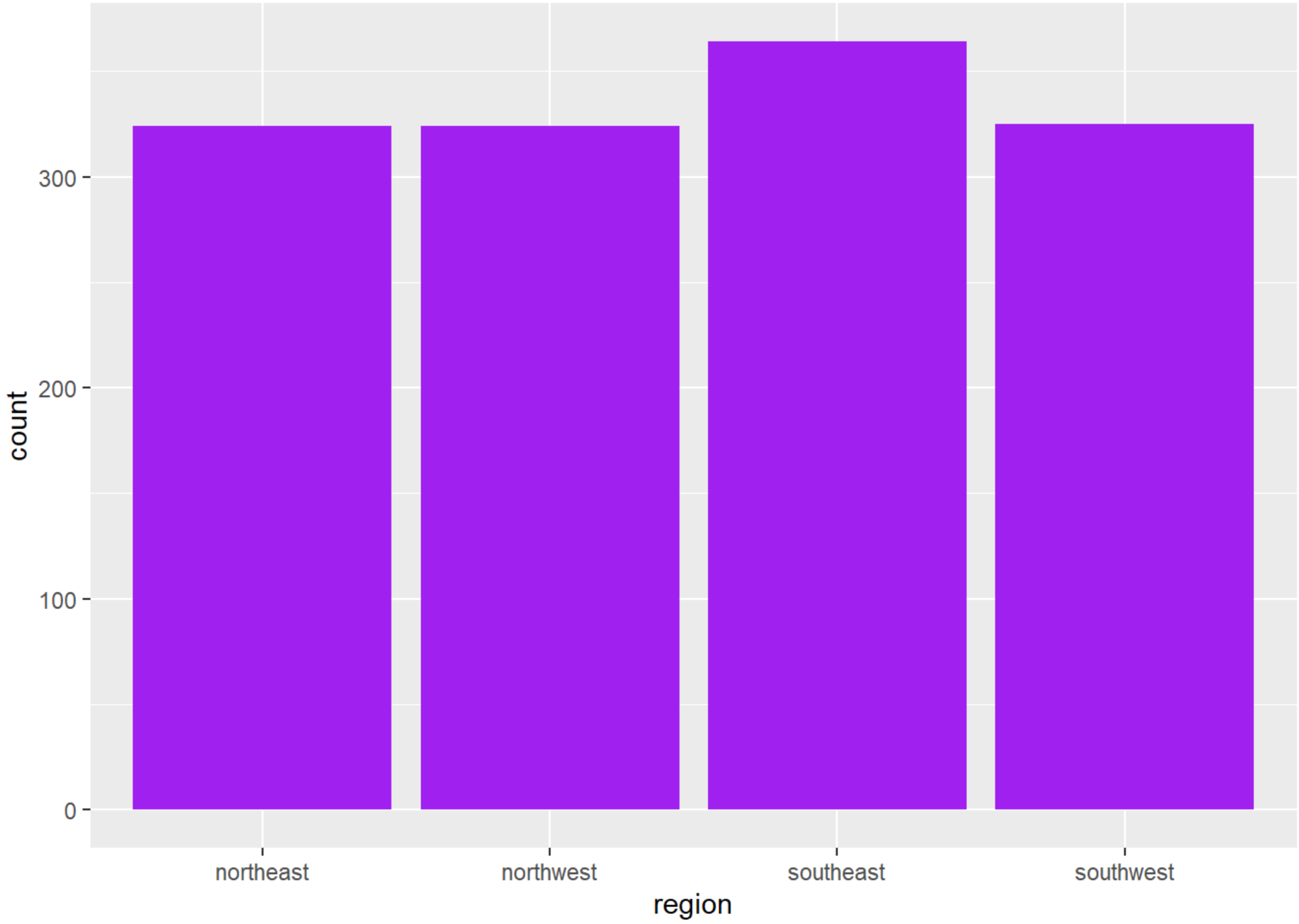
```
ggplot(data = MCPclean) + geom_bar(mapping = aes(x = smoker), fill = "skyblue")
```



> The ratio of smokers : non-smokers = 1:4 in the dataset

Region Distribution

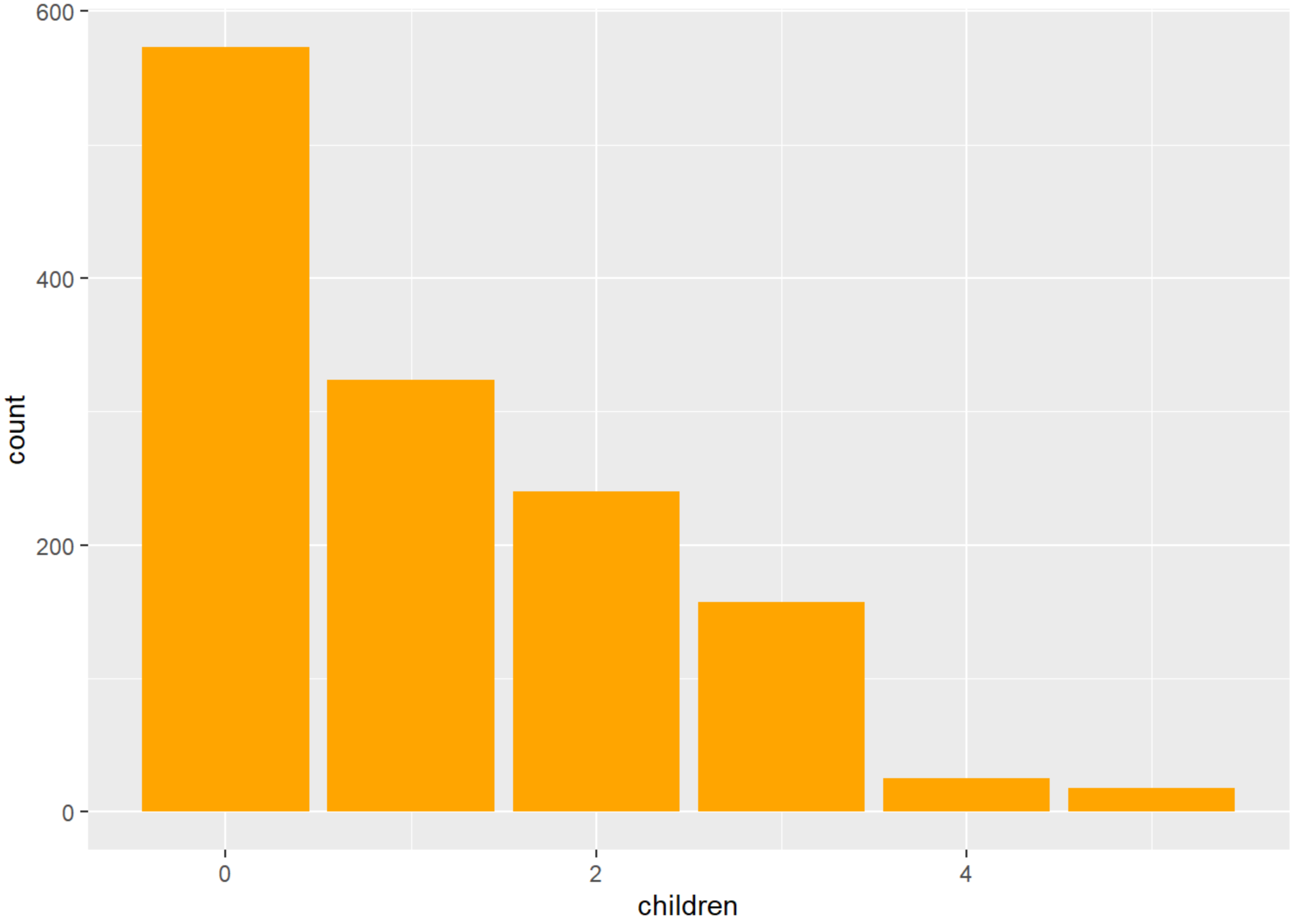
```
ggplot(data = MCPclean) + geom_bar(mapping = aes(x = region), fill = "purple")
```



> All the 4 regions are approximately equally and fairly distributed in the dataset.

Children Distribution

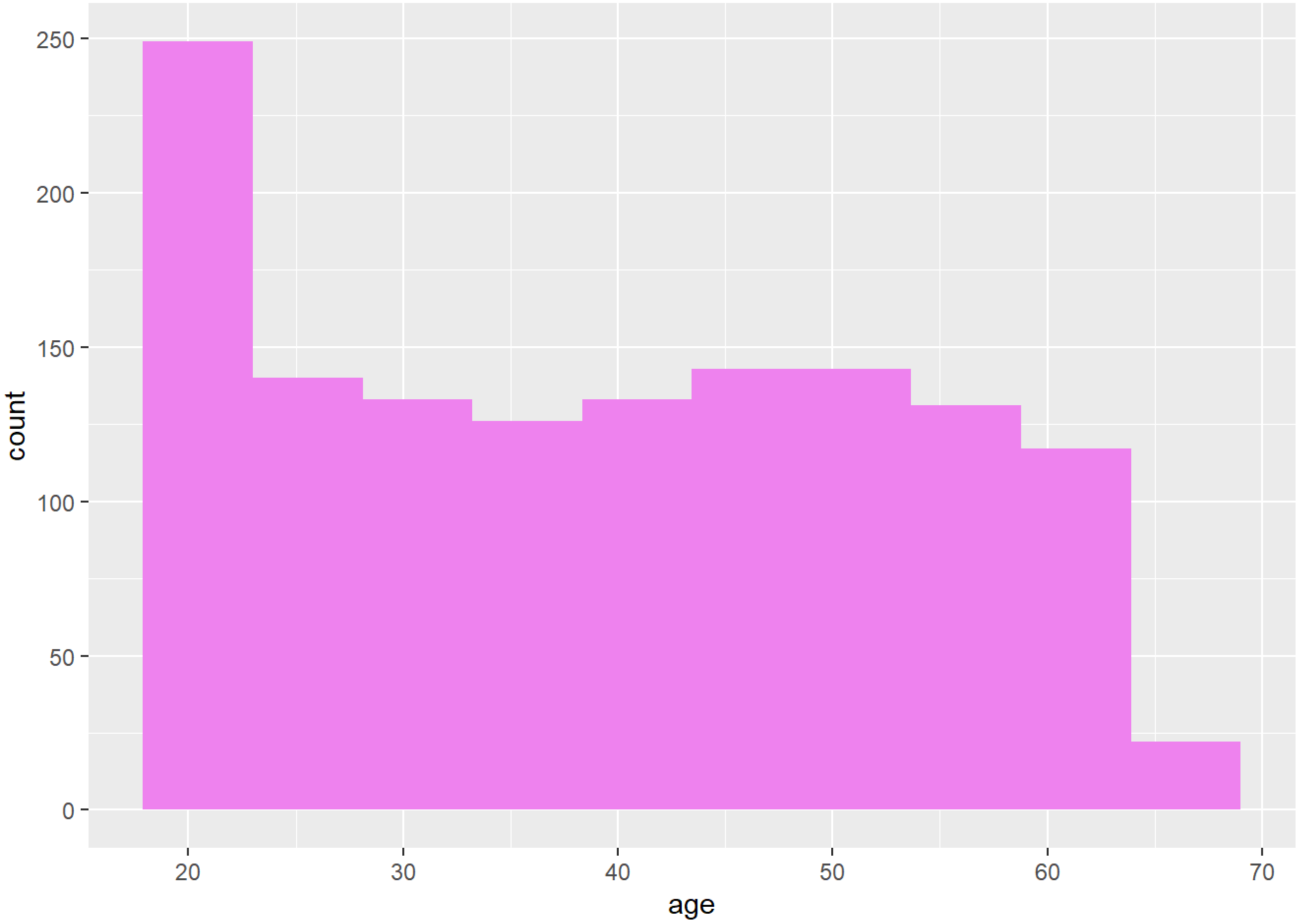
```
ggplot(data = MCPclean) + geom_bar(mapping = aes(x = children), fill = "Orange")
```



> The graph shows a clear trend, with a large number of people having no children, followed by a progressive reduction as the number of children grows, resulting in a significant drop for those with four or five children.

Age Distribution

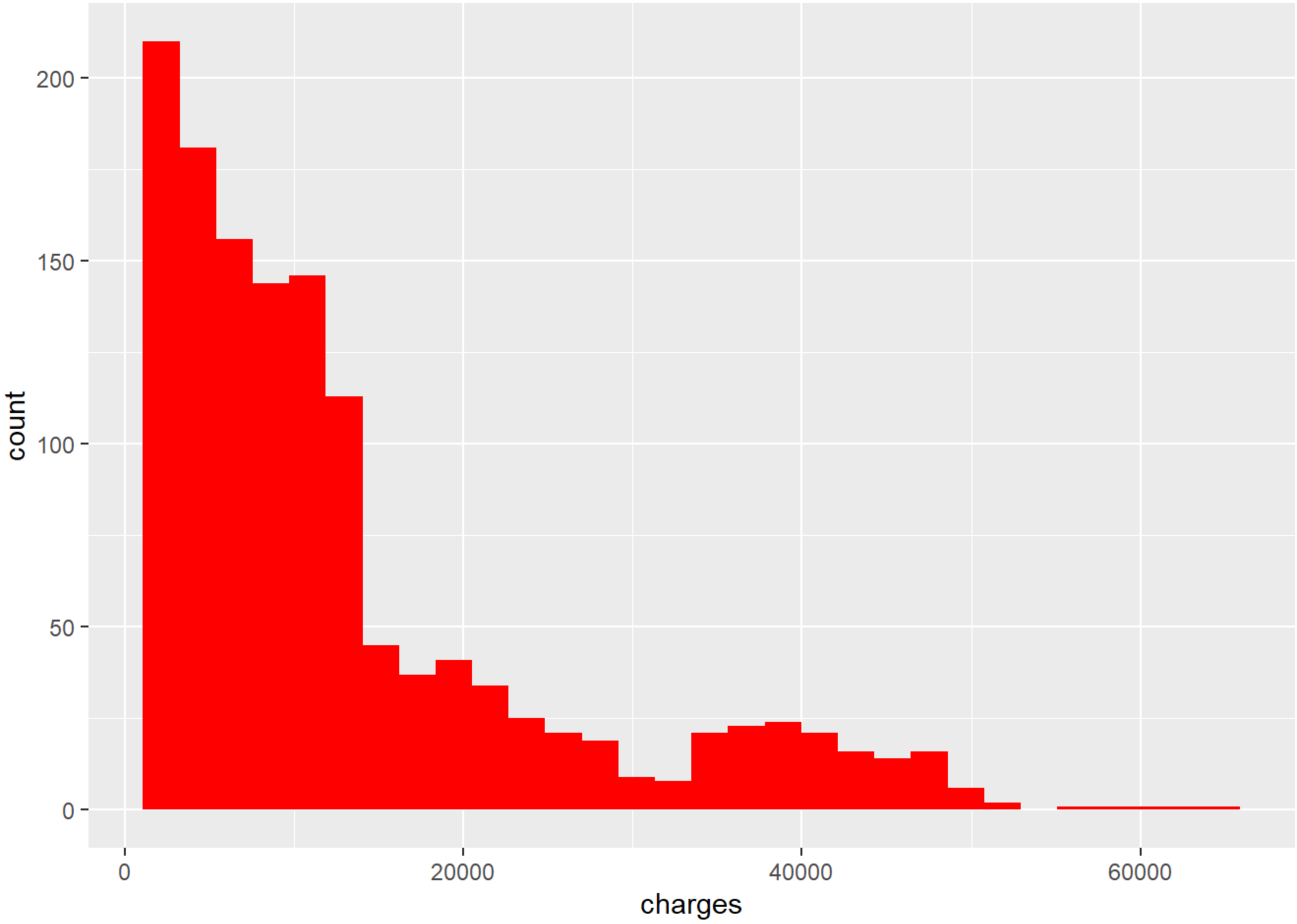
```
ggplot(data = MCPclean) + geom_histogram(mapping = aes(x = age),bins =10, fill = "violet")
```



> Individuals across all age groups have been fairly represented in the dataset, with peaks and troughs at extreme age groups.

Charges Distribution

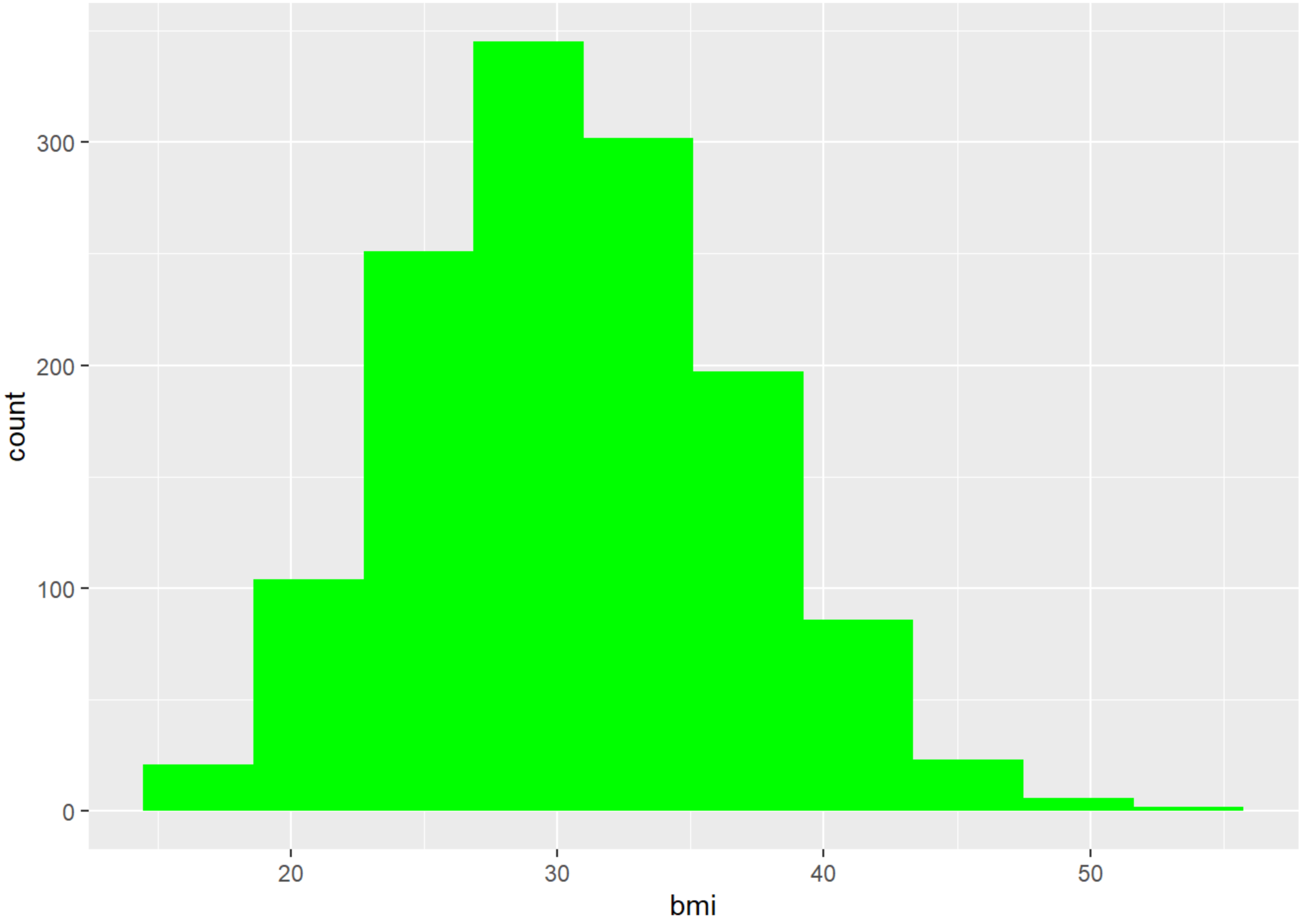
```
ggplot(data = MCPclean) + geom_histogram(mapping = aes(x = charges),bins =30, fill = "red")
```



> Most of the individuals in the dataset incur charges less than 20,000.

BMI Distribution

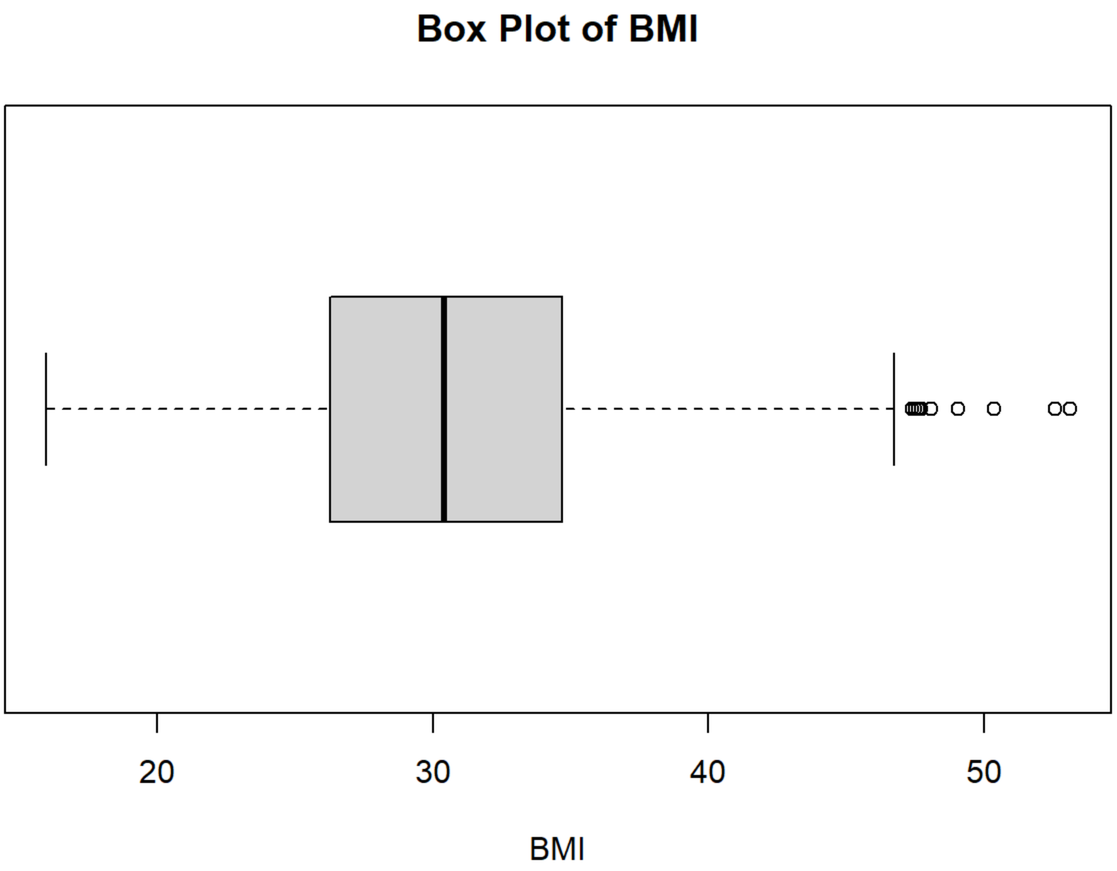
```
ggplot(data = MCPclean) + geom_histogram(mapping = aes(x = bmi),bins =10, fill = "green")
```



> The graph is approximately normally distributed, However, in extreme scenario it can be considered slightly skewed towards the right side.
> Let’s analyze it further.

Analyzing if the BMI attribute is normally distributed using ‘Boxplot’.

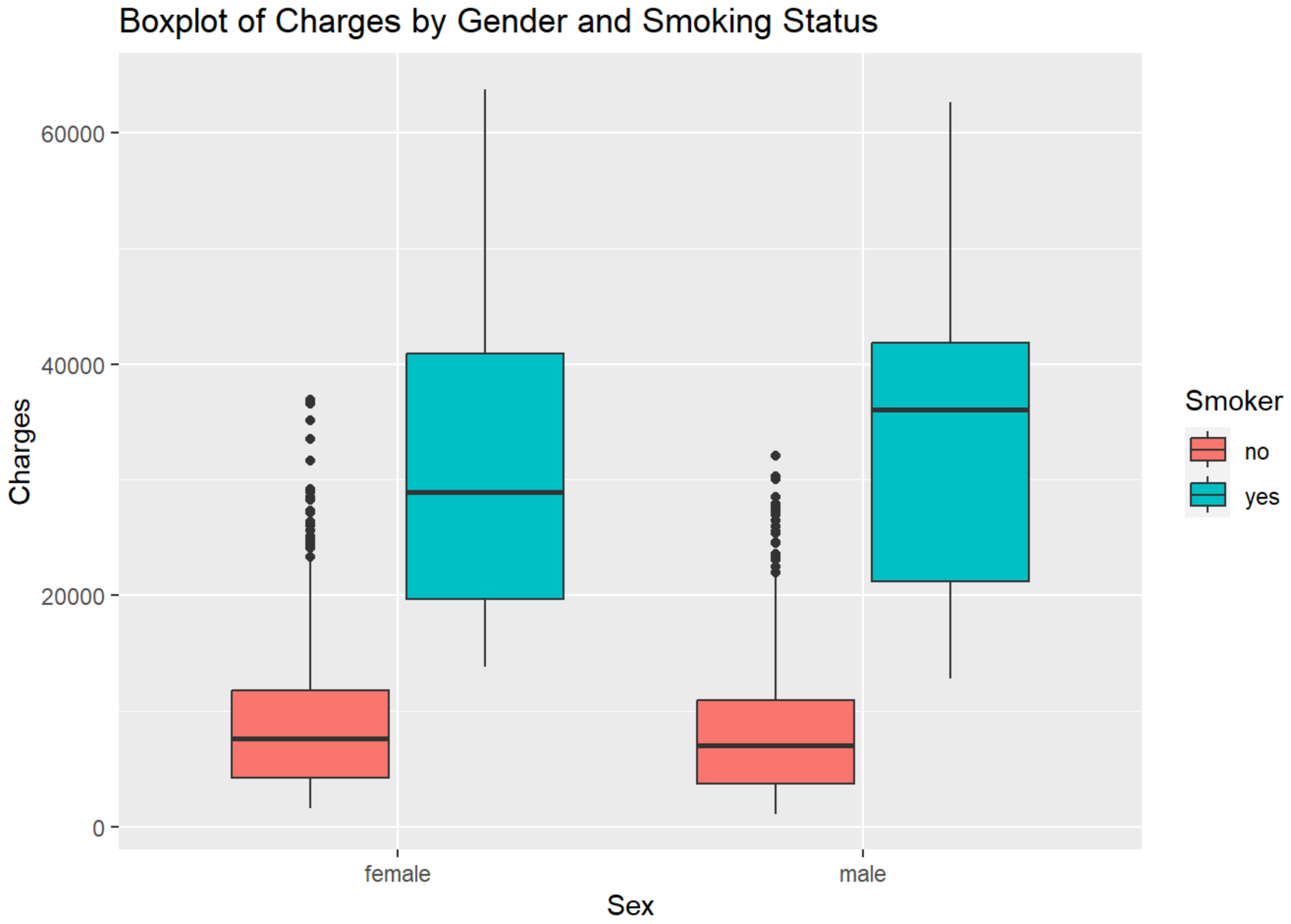
```
boxplot(MCPclean$bmi, horizontal = TRUE, main = "Box Plot of BMI", xlab = "BMI")
```



> After removing the outliers, It is evident from the above Boxplot, that the BMI data is in fact Normally Distributed for all practical purposes.

Compare medical changes based on gender and smoking status

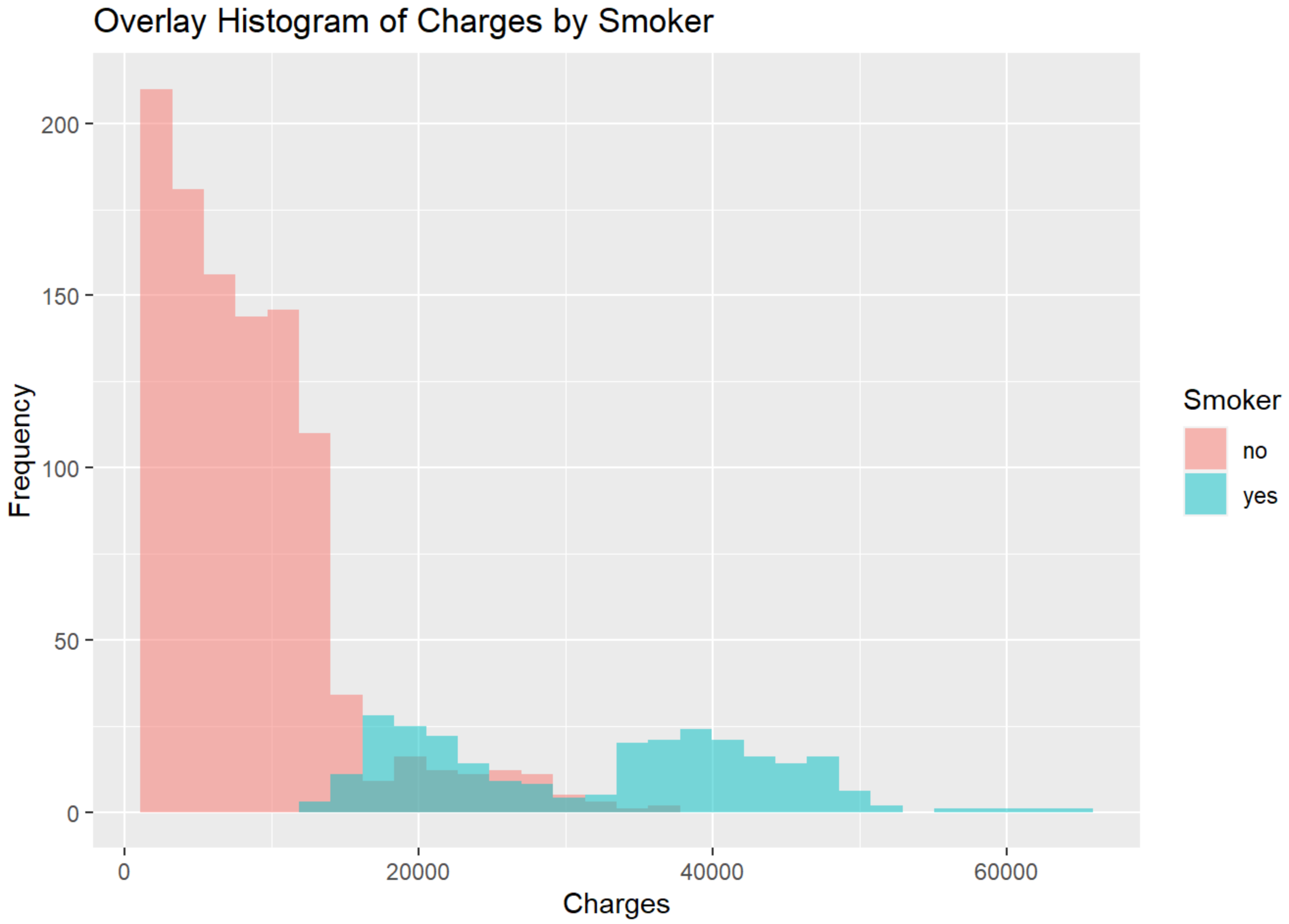
```
ggplot(MCPclean, aes(x = sex, y = charges, fill = smoker)) +
  geom_boxplot() +
  labs(title = "Boxplot of Charges by Gender and Smoking Status",
       x = "Sex",
       y = "Charges",
       fill = "Smoker")
```



> It is evident from the above analysis that medical charges are gender neutral as both male and female non-smokers have a similar distribution of charges. Likewise, smokers from both genders have their medical bills on the higher side.

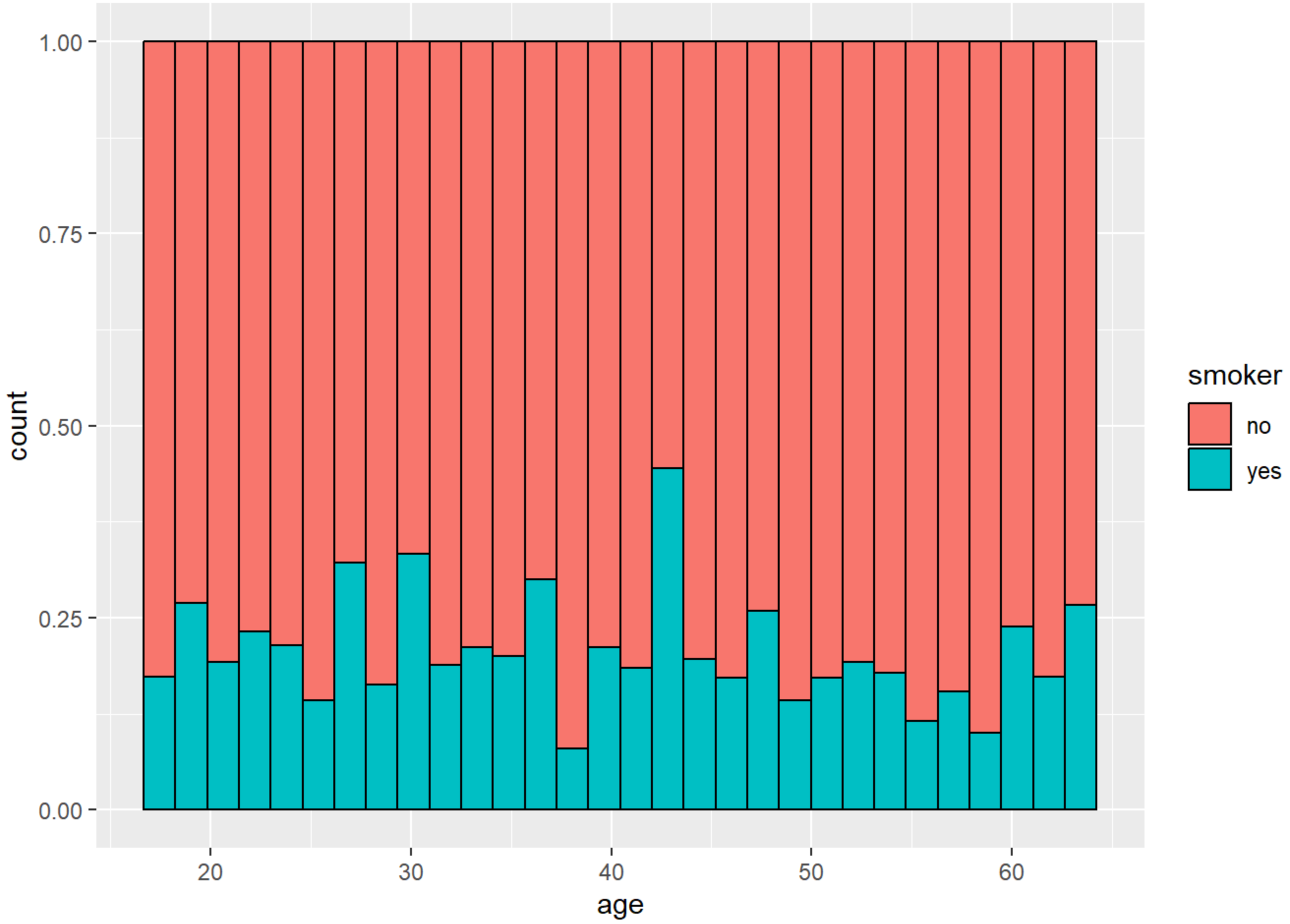
Comparing the medical charges of smokers with non-smokers

```
ggplot(MCPclean, aes(x = charges, fill = smoker)) +
  geom_histogram(alpha = 0.5, bins = 30, position = "identity") +
  labs(title = "Overlay Histogram of Charges by Smoker",
       x = "Charges",
       y = "Frequency",
       fill = "Smoker")
```



> It is clearly evident from the above graph that smokers tend to incur higher medical bills compared to non-smokers. Hence, it can be inferred that smoking adversely affects an individuals health and translates to higher medical costs.

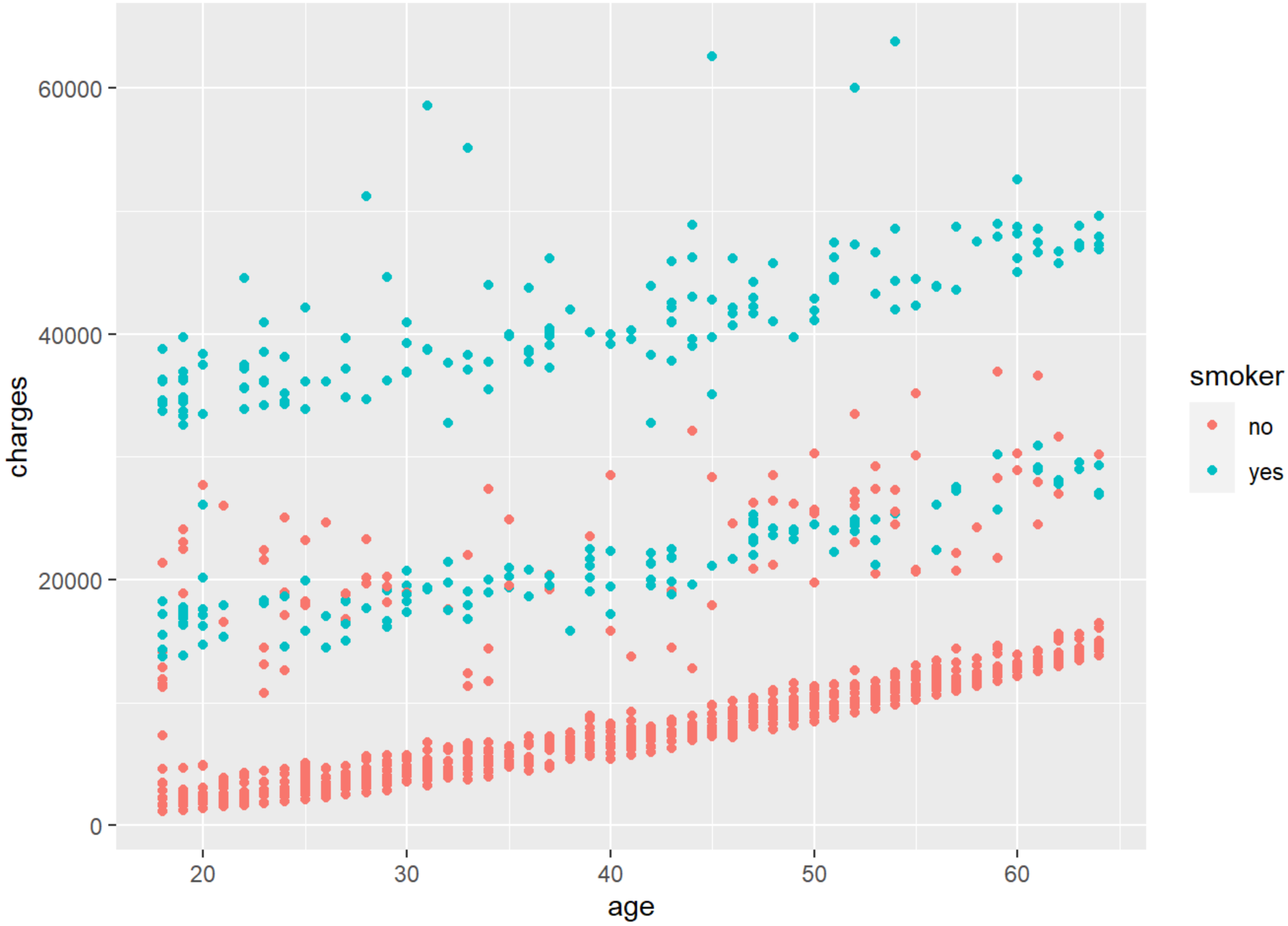
```
ggplot(data = MCPclean, aes(x = age)) + geom_histogram(aes(fill = smoker),bins = 30, color = "bl
ack", position = "fill")
```



###

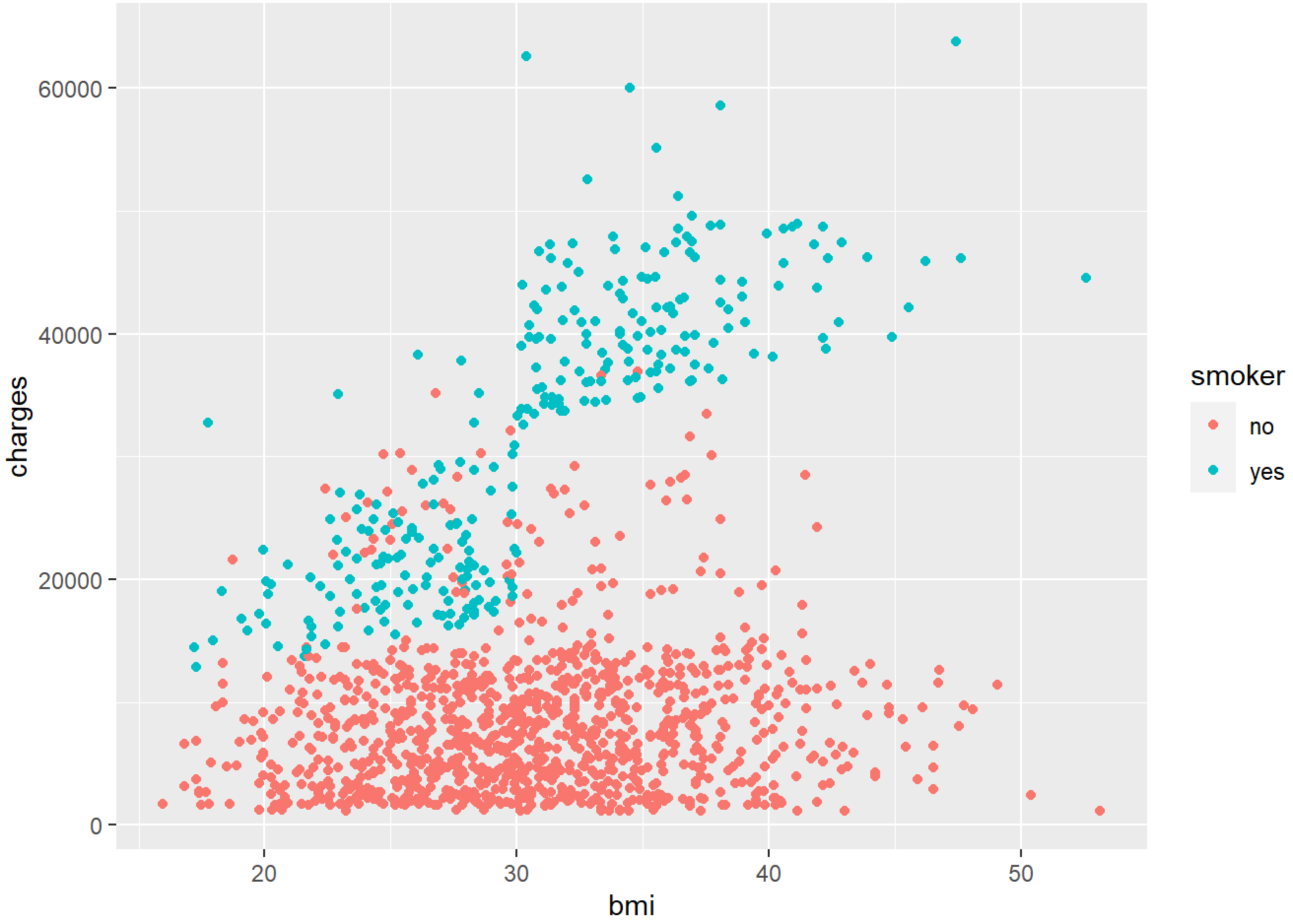
Analyzing the relation between medical charges and age

```
ggplot(data = MCPclean) + geom_point(mapping = aes(x = age, y = charges,colour = smoker))
```



> Distinct trend lines can be observed upon plotting the age against charges. As the age increases, the medical charges also increases correspondingly. This is true for both smokers and non-smokers as separate trend lines for both categories of people can be observed.

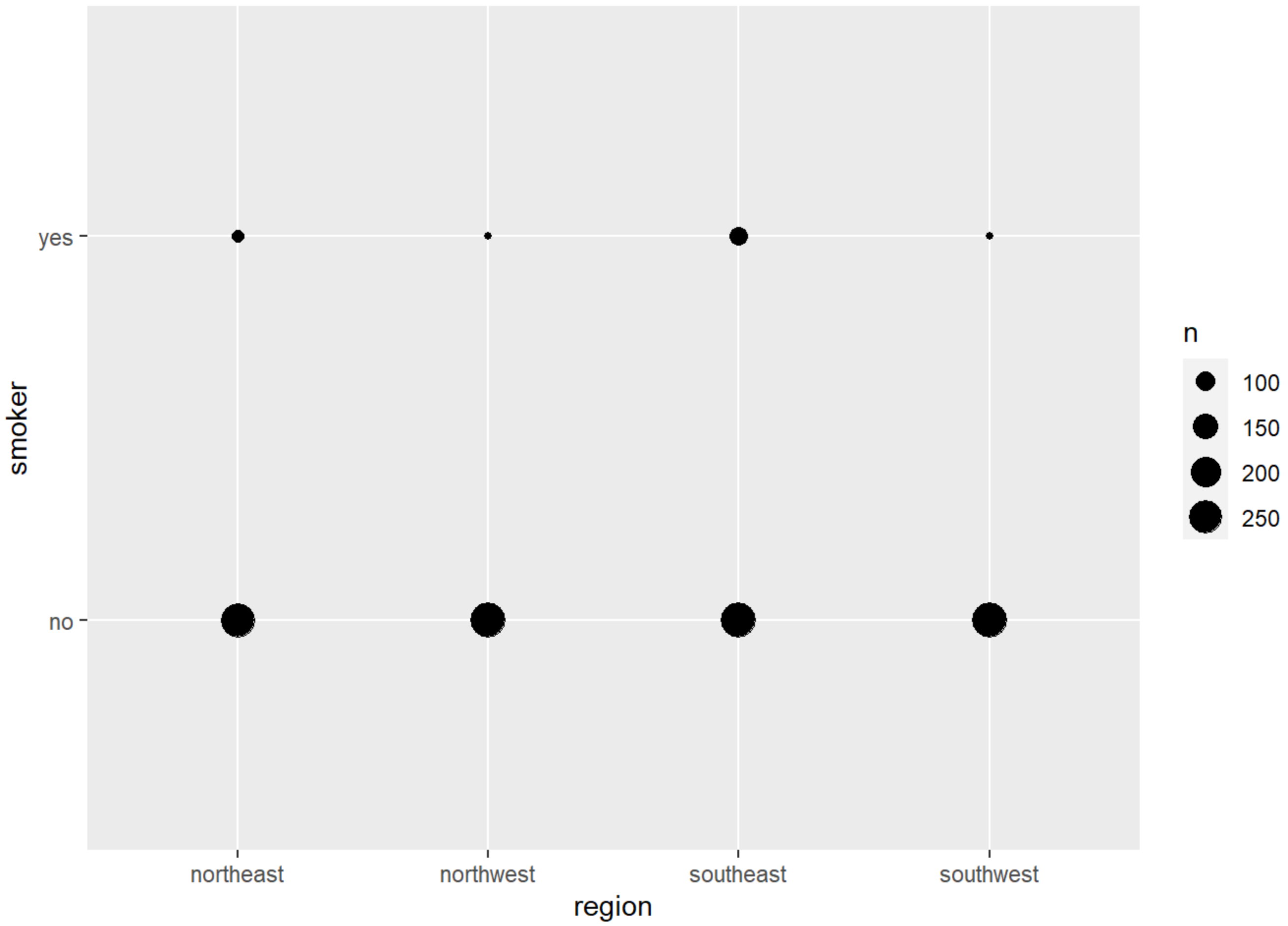
```
ggplot(data = MCPclean) + geom_point(mapping = aes(x = bmi, y = charges,colour = smoker))
```



> From the above analysis, it can be deduced that rising bmi coupled smoking habits, generally contribute to higher medical charges. If any of the two contributing factors is negative i.e. bmi is close to the healthy range (18 to 25) or the individual is non-smoker, it corresponds to lower medical charges.

Analyzing the distribution of smokers and non-smokers across the 4 regions

```
ggplot(data = MCPclean) + geom_count(mapping = aes(x = region, y = smoker))
```



> Non-smokers are fairly and equally distributed across all the 4 regions. However, upon comparative analysis, the smoking population is higher in the southeast region compared to the other regions.

Part III: Data Analysis

a) Hypothesis Testing

Comparing the variance of the medical charges of the smoking population with the non-smoking population

The hypothesis test for comparing the variance can be constricted as follows:
H0: variance of charges for smokers = variance of charges for non-smokers
H1: variance of charges for smokers != variance of charges for non-smokers

Determining the mean and standard deviation of the charges

```
mean(MCPclean$charges)
```

```
## [1] 13279.12
```

```
sd(MCPclean$charges)
```

```
## [1] 12110.36
```

> The mean of all the charges in the dataset is 13279 with a standard deviation of 12110

Creating separate data frames for smokers and non smokers for comparing their variance

```
smokers_charges <- MCPclean$charges[MCPclean$smoker == "yes"]
non_smokers_charges <- MCPclean$charges[MCPclean$smoker == "no"]
```

F test to compare varainces

```
var.test(smokers_charges, non_smokers_charges, alternative= "two.sided", conf.level= 0.95)
```

```
## 
##  F test to compare two variances
## 
## data:  smokers_charges and non_smokers_charges
## F = 3.7089, num df = 273, denom df = 1062, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.088374 4.501628
## sample estimates:
## ratio of variances 
##           3.708884
```

> As P value is less than alpha we can reject null hypothesis, From F test we can conclude that there is a statistically significant difference between the variances of the two samples.

Assuming,
Null Hypothesis (H0) : Mean charges for smokers is equal to Mean charges for non smokers
Alternative Hypothesis (HA) : Mean charges for smokers is not equal to Mean charges for non smokers.

```
t.test(smokers_charges, non_smokers_charges, var.equal = FALSE, conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  smokers_charges and non_smokers_charges
## t = 32.742, df = 311.88, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22190.79 25028.35
## sample estimates:
## mean of x mean of y
##  32050.23   8440.66
```

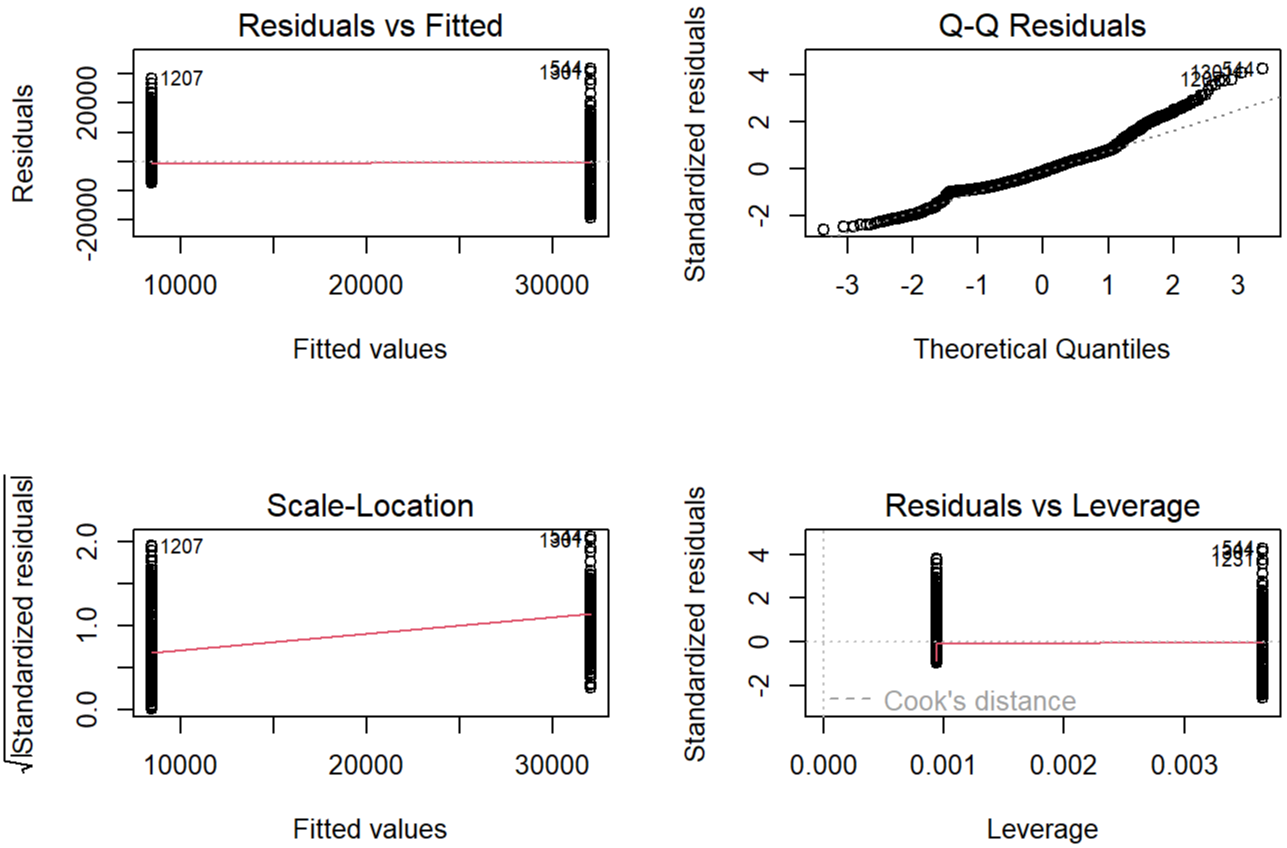
> As p value is less than alpha, we can reject null hypothesis. This implies that the mean of the charges for smokers is different from mean of charges for non smokers (Or) we can say difference in means is not equal to 0.

b) Linear Regression

```
#fitting linear model
fitlm <- lm(charges~smoker, data =MCPclean )

summary(fitlm)
```

```
##
## Call:
## lm(formula = charges ~ smoker, data = MCPclean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19221  -5048   -923   3702  31720
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8440.7      229.1   36.84   <2e-16 ***
## smokeryes    23609.6      506.2   46.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7471 on 1335 degrees of freedom
## Multiple R-squared:  0.6197, Adjusted R-squared:  0.6195
## F-statistic:  2176 on 1 and 1335 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fitlm)
```



```
# Convert categorical variables (sex, smoker, region) into factors
MCPclean$sex <- as.factor(MCPclean$sex)
MCPclean$smoker <- as.factor(MCPclean$smoker)
MCPclean$region <- as.factor(MCPclean$region)
```

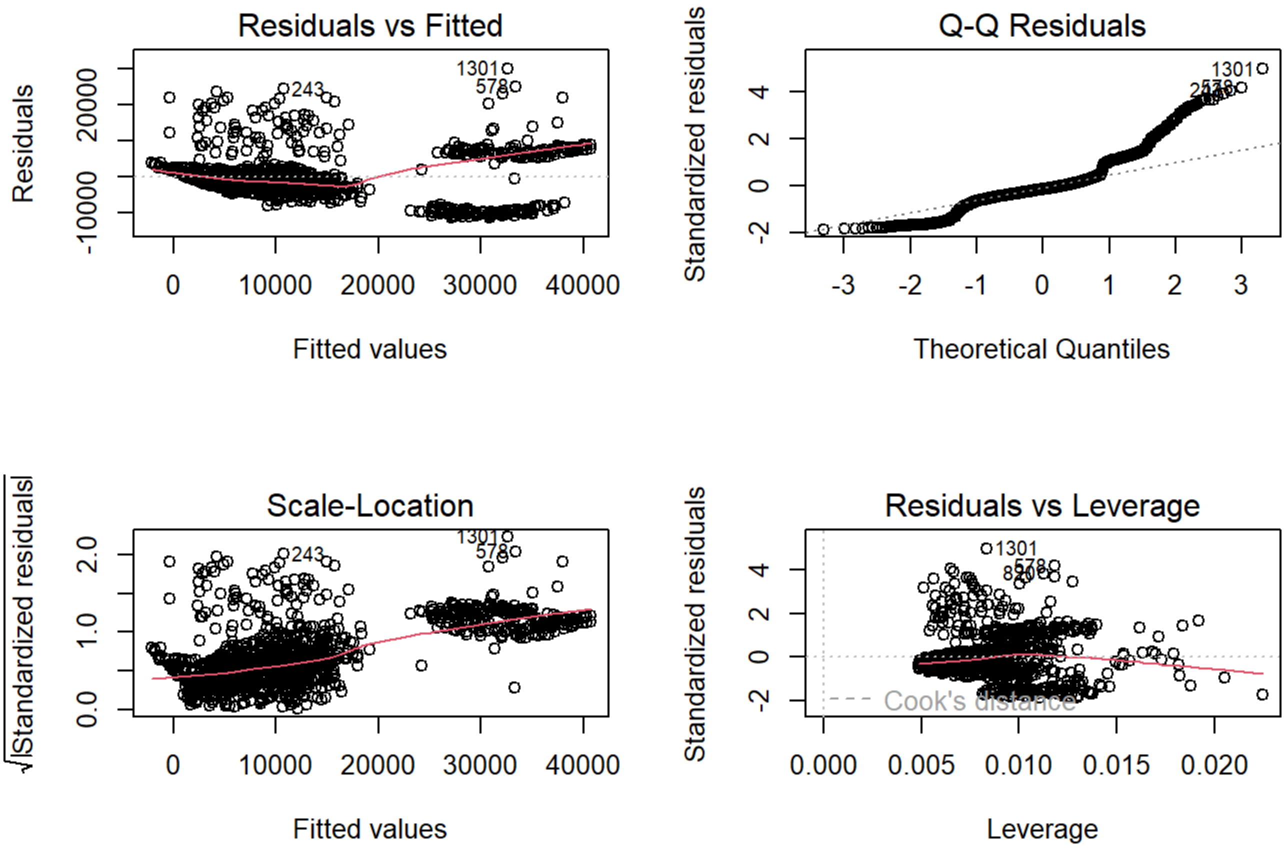
```
# Split the data into training and testing sets (80-20 split)
set.seed(42)
splitIndex <- sample(1:nrow(MCPclean), 0.8 * nrow(MCPclean))
trainData <- MCPclean[splitIndex, ]
testData <- MCPclean[-splitIndex, ]
```

```
# Train the linear regression model
model <- lm(charges ~ ., data = trainData)
```

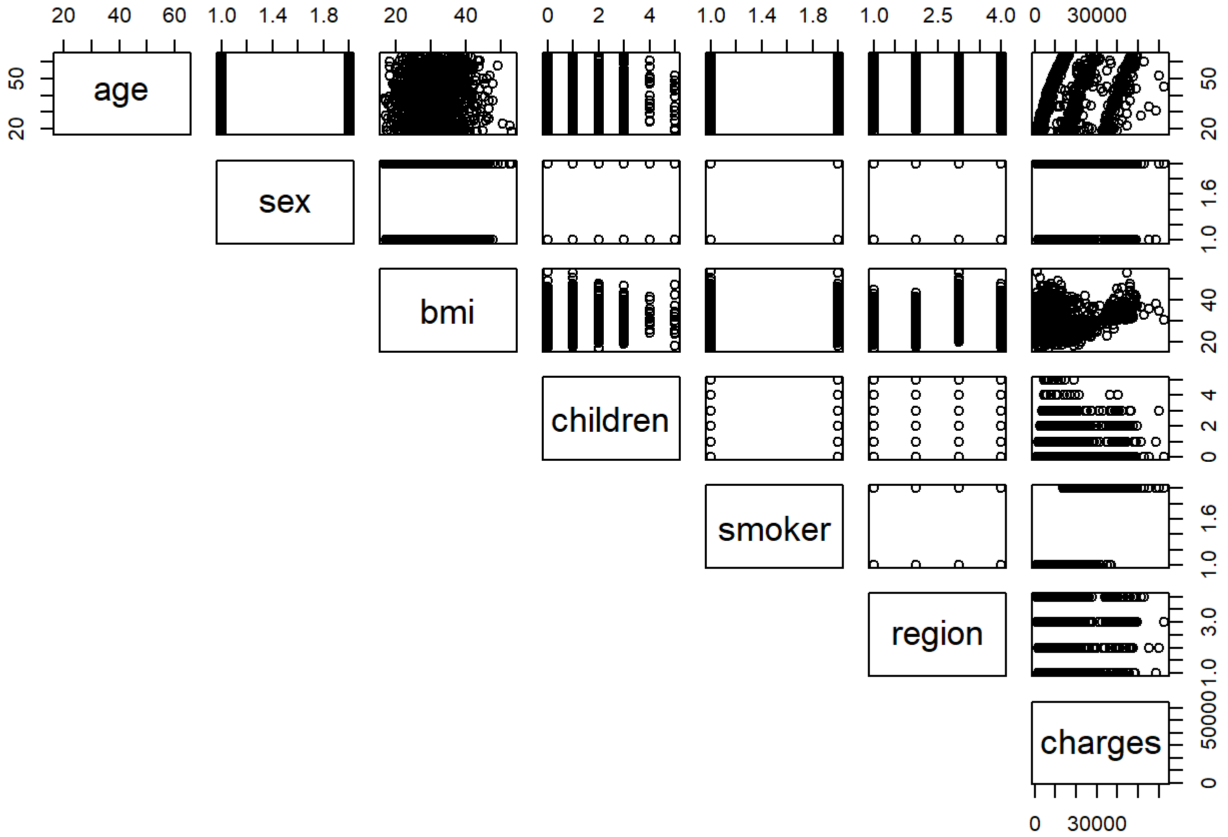
```
#summary of the model
summary(model)
```

```
## 
## Call:
## lm(formula = charges ~ ., data = trainData)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11363.9  -2875.2   -957.9   1502.7  30010.1 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     -11914.04    1139.93 -10.452   <2e-16 ***
## age                260.55      13.25  19.664   <2e-16 ***
## sexmale           -131.35     373.56  -0.352   0.7252    
## bmi                335.92      32.32  10.393   <2e-16 ***
## children           471.34     154.24   3.056   0.0023 ** 
## smokeryes        23995.87     458.58  52.327   <2e-16 ***
## regionnorthwest   -542.01     531.81  -1.019   0.3084    
## regionsoutheast  -1291.05     534.08  -2.417   0.0158 *  
## regionsouthwest  -1150.13     535.77  -2.147   0.0320 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6069 on 1060 degrees of freedom
## Multiple R-squared:  0.7538, Adjusted R-squared:  0.7519 
## F-statistic: 405.6 on 8 and 1060 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model)
```



```
pairs(trainData[,1:7], lower.panel = NULL)
```



```
# Predict on the test set
predictions <- predict(model, newdata = testData)

summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -1994    4833    9024   12001   14440   41061 
```

```
# Evaluate the model
mse <- mean((predictions - testData$charges)^2)
cat("Mean Squared Error on Test Set:", mse, "\n")
```

```
## Mean Squared Error on Test Set: 36644557
```

```
summary(predictions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -1994    4833    9024   12001   14440   41061 
```

```
# Feature Importance
feature_importance <- coef(model)[-1]  # Exclude intercept
feature_importance <- abs(feature_importance)
feature_importance <- sort(feature_importance, decreasing = TRUE)
```

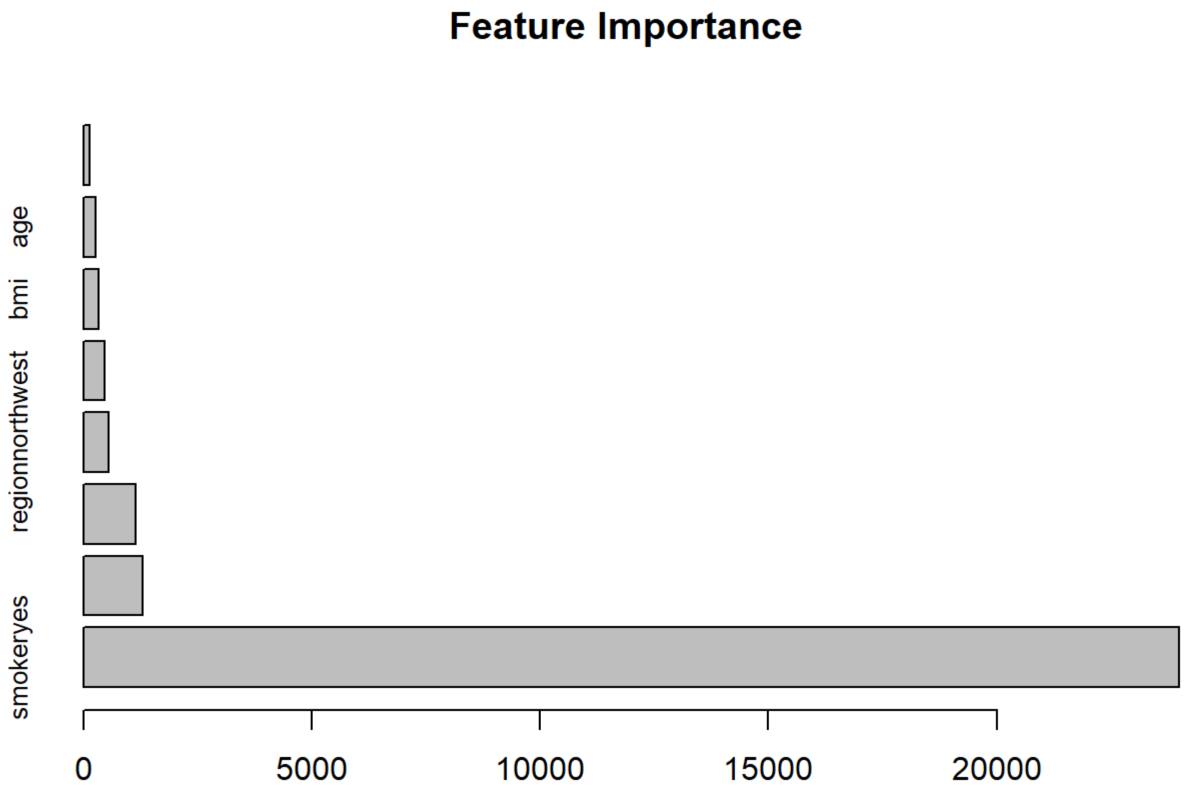
```
# Print feature importance
cat("Feature Importance:\n")
```

```
## Feature Importance:
```

```
print(feature_importance)
```

```
##       smokeryes regionsoutheast regionsouthwest regionnorthwest        children
##      23995.8704       1291.0459       1150.1298        542.0094        471.3367
##             bmi             age         sexmale
##        335.9191        260.5523        131.3502
```

```
# Plot feature importance
barplot(feature_importance, main = "Feature Importance", horiz = TRUE, cex.names = 0.8)
```



```
# Calculate Mean Squared Error (MSE) on the test set
mse <- mean((predictions - testData$charges)^2)
cat("Mean Squared Error on Test Set:", mse, "\n")
```

```
## Mean Squared Error on Test Set: 36644557
```

```
# Calculate Mean Absolute Error (MAE) on the test set
mae <- mean(abs(predictions - testData$charges))
cat("Mean Absolute Error on Test Set:", mae, "\n")
```

```
## Mean Absolute Error on Test Set: 4019.276
```

```
# Define the full model
full_model <- lm(charges ~ ., data = trainData)

# Forward Stepwise Regression
empty_model <- lm(charges ~ 1, data = trainData)

forward_model <- step(empty_model, direction = "forward", scope = formula(full_model))
```

```
## Start:  AIC=20115.23
## charges ~ 1
## 
##            Df  Sum of Sq        RSS   AIC
## + smoker    1 9.9368e+10 5.9204e+10 19064
## + age       1 1.3303e+10 1.4527e+11 20024
## + bmi       1 4.8131e+09 1.5376e+11 20084
## + region    3 1.2860e+09 1.5729e+11 20113
## + sex       1 6.9467e+08 1.5788e+11 20113
## + children  1 4.8008e+08 1.5809e+11 20114
## <none>                   1.5857e+11 20115
## 
## Step:  AIC=19064.02
## charges ~ smoker
## 
##            Df  Sum of Sq        RSS   AIC
## + age       1 1.5688e+10 4.3516e+10 18737
## + bmi       1 5.0096e+09 5.4194e+10 18972
## + children  1 4.4034e+08 5.8763e+10 19058
## <none>                   5.9204e+10 19064
## + sex       1 6.6924e+05 5.9203e+10 19066
## + region    3 1.3027e+08 5.9073e+10 19068
## 
## Step:  AIC=18736.92
## charges ~ smoker + age
## 
##            Df  Sum of Sq        RSS   AIC
## + bmi       1 3867834657 3.9648e+10 18639
## + children  1  338357359 4.3177e+10 18731
## <none>                   4.3516e+10 18737
## + sex       1    6691337 4.3509e+10 18739
## + region    3  118736431 4.3397e+10 18740
## 
## Step:  AIC=18639.41
## charges ~ smoker + age + bmi
## 
##            Df Sum of Sq        RSS   AIC
## + children  1 330121984 3.9318e+10 18633
## + region    3 256625290 3.9391e+10 18639
## <none>                  3.9648e+10 18639
## + sex       1   1178379 3.9647e+10 18641
## 
## Step:  AIC=18632.47
## charges ~ smoker + age + bmi + children
## 
##          Df Sum of Sq        RSS   AIC
## + region  3 267891148 3.9050e+10 18631
## <none>                3.9318e+10 18633
## + sex     1   3195311 3.9314e+10 18634
## 
## Step:  AIC=18631.16
## charges ~ smoker + age + bmi + children + region
```

```
##
##        Df Sum of Sq        RSS   AIC
## <none>              3.9050e+10 18631
## + sex   1   4554151 3.9045e+10 18633
```

```
# Predict on the test set
predictions_forward <- predict(forward_model, newdata = testData)

# Calculate Mean Squared Error (MSE) on the test set
mse_forward <- mean((predictions_forward - testData$charges)^2)
cat("Mean Squared Error (Forward Stepwise):", mse_forward, "\n")
```

```
## Mean Squared Error (Forward Stepwise): 36646634
```

```
# Calculate Mean Absolute Error (MAE) on the test set
mae_forward <- mean(abs(predictions_forward - testData$charges))
cat("Mean Absolute Error(Forward Stepwise):", mae_forward, "\n")
```

```
## Mean Absolute Error(Forward Stepwise): 4017.539
```

```
summary(forward_model)
```

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi + children + region,
##     data = trainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11429.5  -2835.9   -938.1   1524.2  29950.4
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11960.12    1131.90 -10.566  < 2e-16 ***
## smokeryes        23982.72     456.86  52.494  < 2e-16 ***
## age                260.65      13.24  19.683  < 2e-16 ***
## bmi                335.23      32.25  10.395  < 2e-16 ***
## children           469.21     154.06   3.046  0.00238 **
## regionnorthwest   -538.49     531.50  -1.013  0.31122
## regionsoutheast  -1286.84     533.73  -2.411  0.01608 *
## regionsouthwest  -1146.91     535.47  -2.142  0.03243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6067 on 1061 degrees of freedom
## Multiple R-squared:  0.7537, Adjusted R-squared:  0.7521
## F-statistic: 463.9 on 7 and 1061 DF,  p-value: < 2.2e-16
```

```
# Backward Stepwise Regression
full_model <- lm(charges ~ ., data = trainData)

backward_model <- step(full_model, direction = "backward")
```

```
## Start:  AIC=18633.04
## charges ~ age + sex + bmi + children + smoker + region
## 
##            Df  Sum of Sq        RSS   AIC
## - sex       1 4.5542e+06 3.9050e+10 18631
## <none>                   3.9045e+10 18633
## - region    3 2.6925e+08 3.9314e+10 18634
## - children  1 3.4396e+08 3.9389e+10 18640
## - bmi       1 3.9789e+09 4.3024e+10 18735
## - age       1 1.4244e+10 5.3289e+10 18964
## - smoker    1 1.0086e+11 1.3990e+11 19995
## 
## Step:  AIC=18631.16
## charges ~ age + bmi + children + smoker + region
## 
##            Df  Sum of Sq        RSS   AIC
## <none>                   3.9050e+10 18631
## - region    3 2.6789e+08 3.9318e+10 18633
## - children  1 3.4139e+08 3.9391e+10 18639
## - bmi       1 3.9772e+09 4.3027e+10 18733
## - age       1 1.4260e+10 5.3309e+10 18962
## - smoker    1 1.0142e+11 1.4047e+11 19998
```

```
# Predict on the test set
predictions_backward <- predict(backward_model, newdata = testData)

# Calculate Mean Squared Error (MSE) on the test set
mse_backward <- mean((predictions_backward - testData$charges)^2)
cat("Mean Squared Error (Backward Stepwise):", mse_backward, "\n")
```

```
## Mean Squared Error (Backward Stepwise): 36646634
```

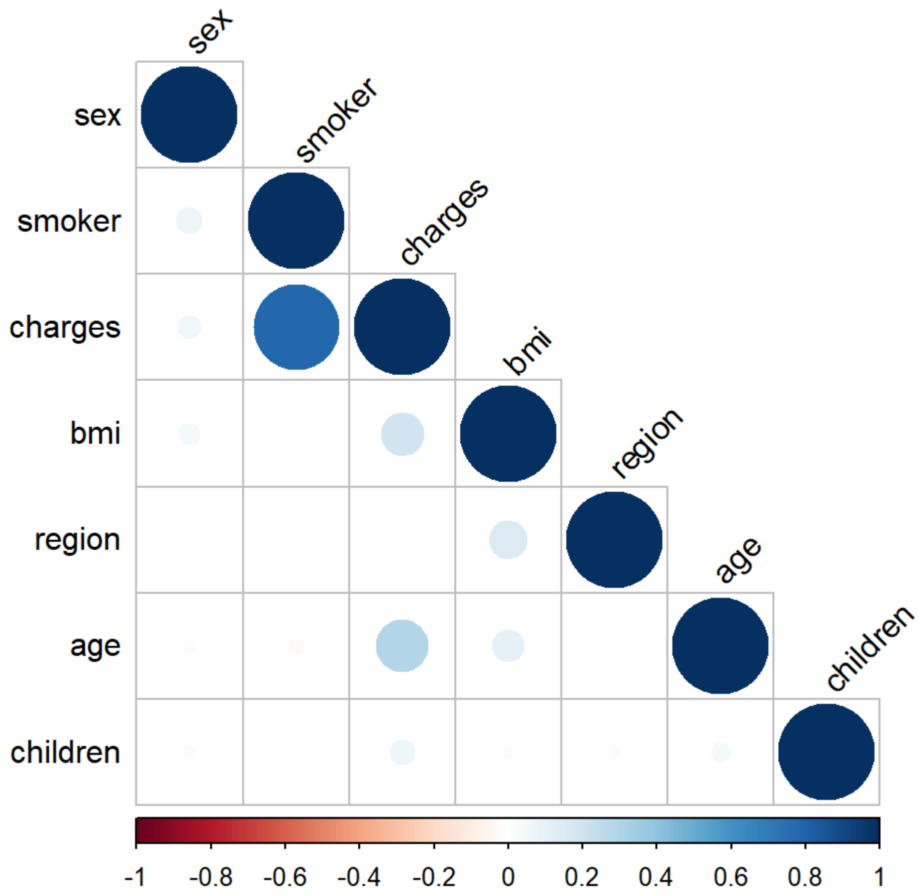
```
# Calculate Mean Absolute Error (MAE) on the test set
mae_backward <- mean(abs(predictions_backward - testData$charges))
cat("Mean Absolute Error(Backward Stepwise):", mae_backward, "\n")
```

```
## Mean Absolute Error(Backward Stepwise): 4017.539
```

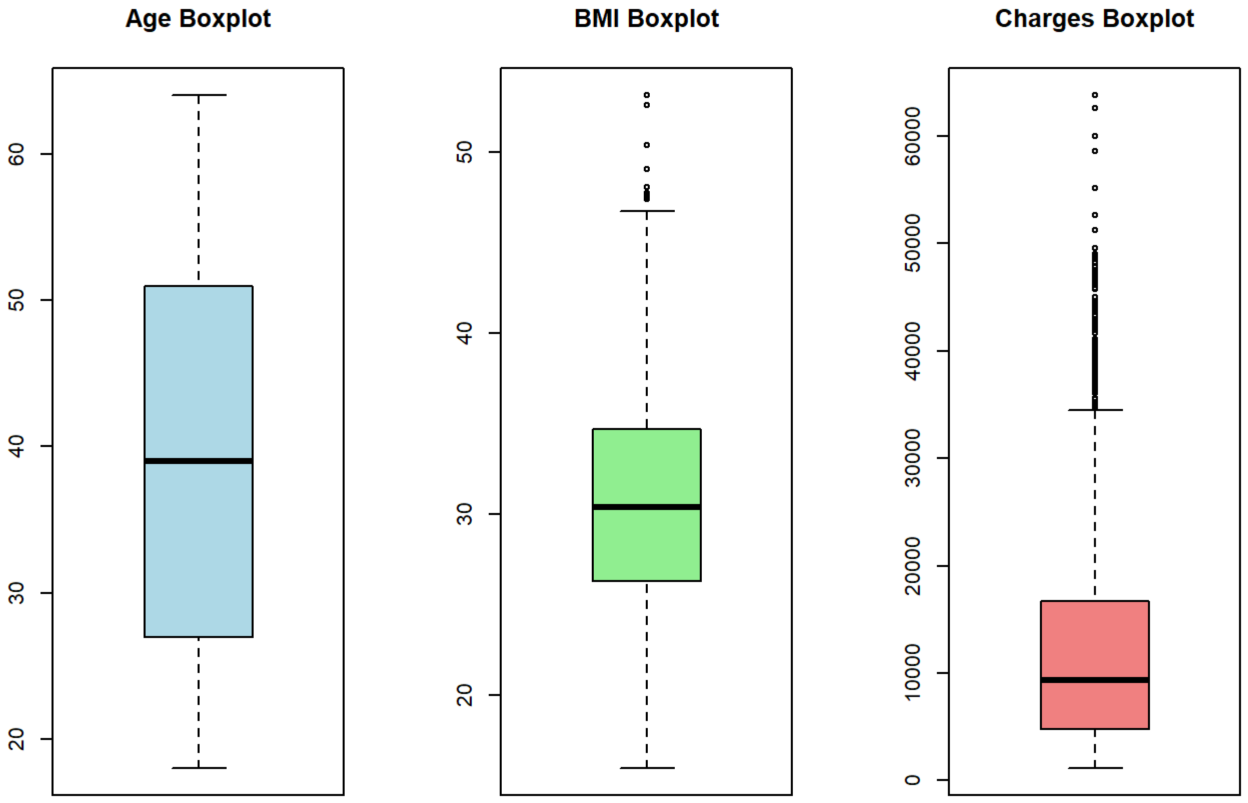
```
# Load the necessary libraries
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
## corrplot 0.92 loaded
```

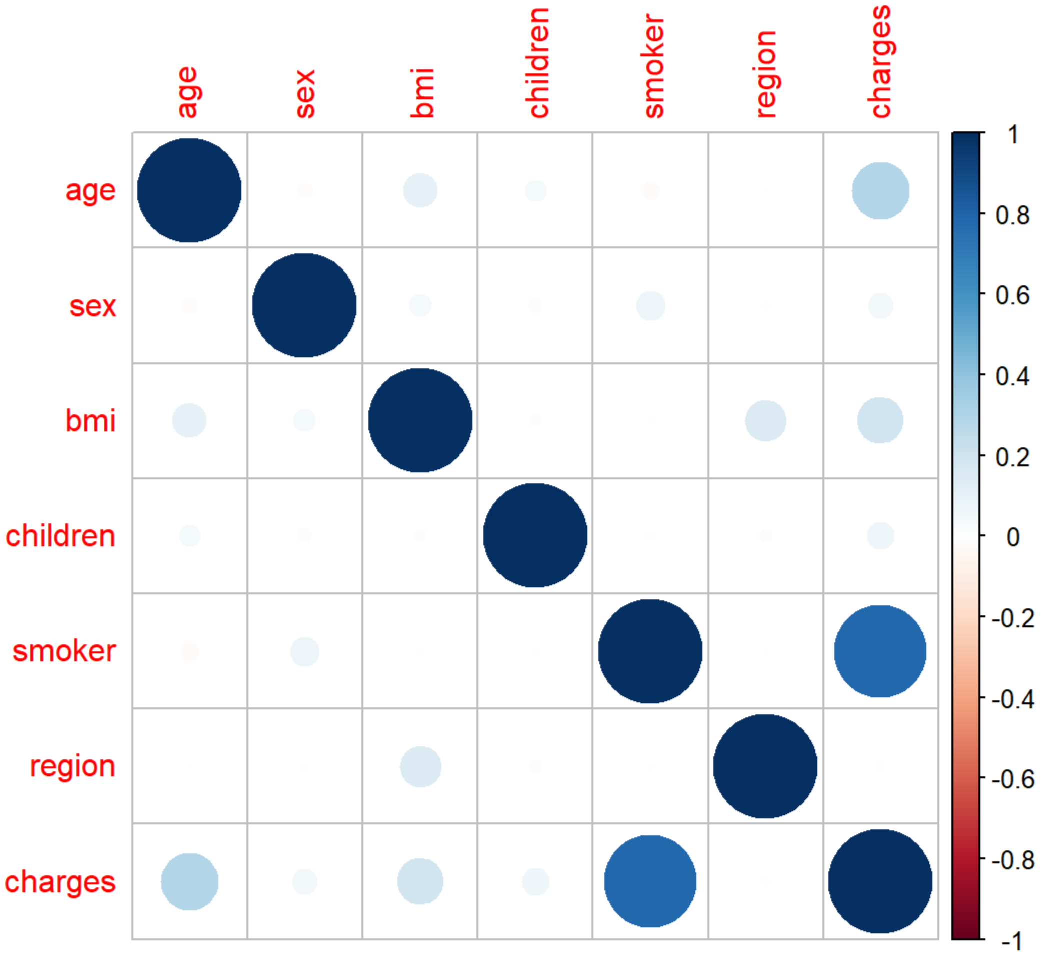
```
# Load the dataset
#df <- read.csv('insurance.csv')

# Convert categorical variables to factors
MCPclean$sex <- as.factor(MCPclean$sex)
MCPclean$smoker <- as.factor(MCPclean$smoker)
MCPclean$region <- as.factor(MCPclean$region)

# Convert factors to numeric
MCPclean$sex <- as.numeric(MCPclean$sex)
MCPclean$smoker <- as.numeric(MCPclean$smoker)
MCPclean$region <- as.numeric(MCPclean$region)

# Calculate the correlation matrix
cor_matrix <- cor(MCPclean)


# Draw the correlation matrix using corrplot
corrplot(cor_matrix, method = "circle", type = "lower", order = "hclust", tl.col = "black", tl.s
rt = 45)
```



```
par(mfrow=c(1,3))
boxplot(MCPclean$age, main="Age Boxplot", col="lightblue", border="black")
boxplot(MCPclean$bmi, main="BMI Boxplot", col="lightgreen", border="black")
boxplot(MCPclean$charges, main="Charges Boxplot", col="lightcoral", border="black")
```



```
library(corrplot)
corrplot(cor(MCPclean), method="circle")
```


age
sex
bmi
children
smoker
region
charges
age
sex
bmi
children
smoker
region
charges
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1