Medical Insurance Payout

OBJECTIVE: ACME Insurance Inc. offers affordable health insurance to thousands of customer all over the United States. Our task is to create an automated system to estimate the annual medical expenditure for new customers, using information such as their age, sex, BMI, children, smoking habits and region of residence.

Estimates from this system will be used to determine the annual insurance premium (amount paid every month) offered to the customer.

```
library(ggplot2)
library(gridExtra)
library(corrplot)

## corrplot 0.92 loaded
```

Load the data

```
expenses <- read.csv("/Users/srushtigupte/Desktop/R/expenses.csv", header =
TRUE)
head(expenses)

##   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
```

There are seven columns and all oof them are explained below. Age: Insurance contractor's age Sex: Insurance contractor's gender, [female, male] BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight(kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 Children: Number of children covered by health insurance / Number of dependents Smoker: Smoking, [yes, no] Region: The beneficiary's residential area in Bangladesh [northeast,southeast, southwest, northwest] Charges: Individual medical costs billed by health insurance

Adding a column of risk of higher insurance payout to indicate the charges were higher than $10,000 or not.

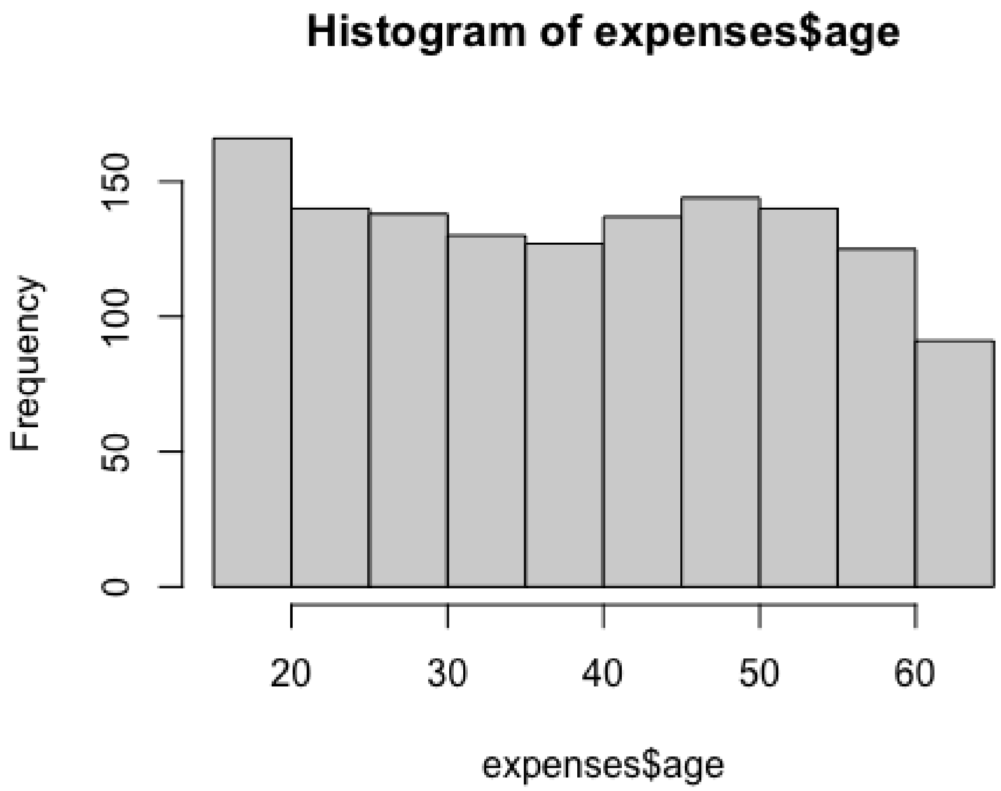
```
expenses["risk"] <- 0
expenses$risk <- ifelse(expenses$charges > 10000, "Yes", "No")
head(expenses)

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

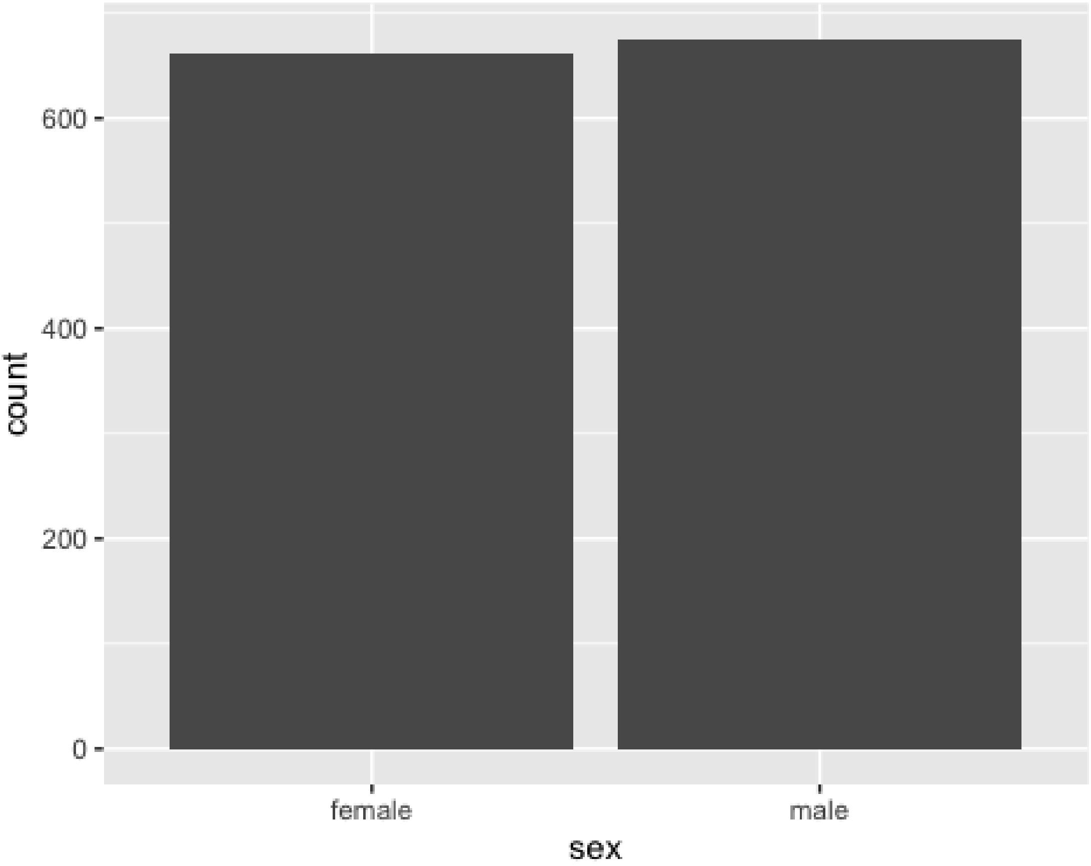
EDA

Distribution of all the variables

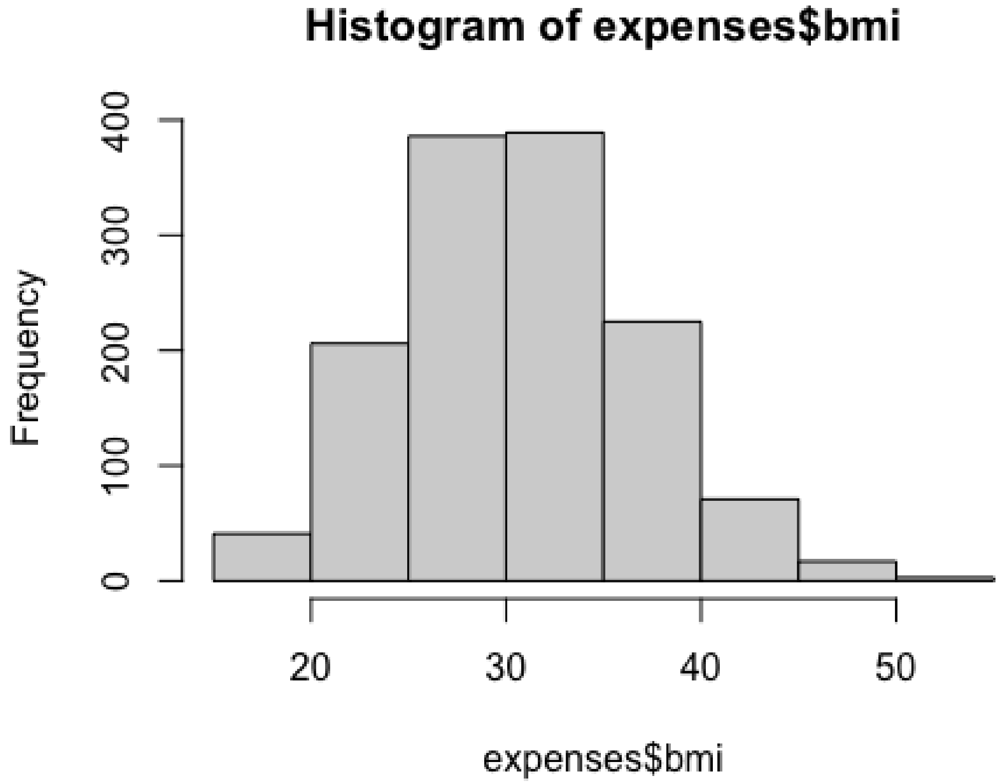
```
# Create individual ggplot plots
hist(expenses$age)
```



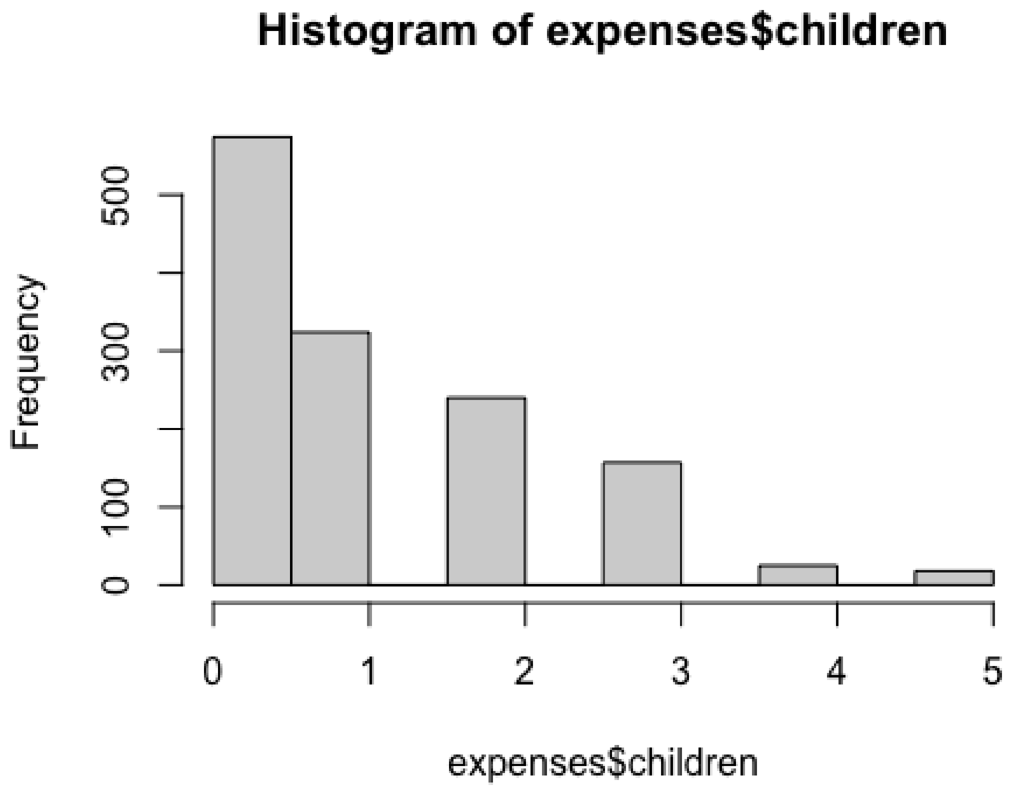
```
ggplot(expenses, aes(x = sex)) + geom_bar()
```



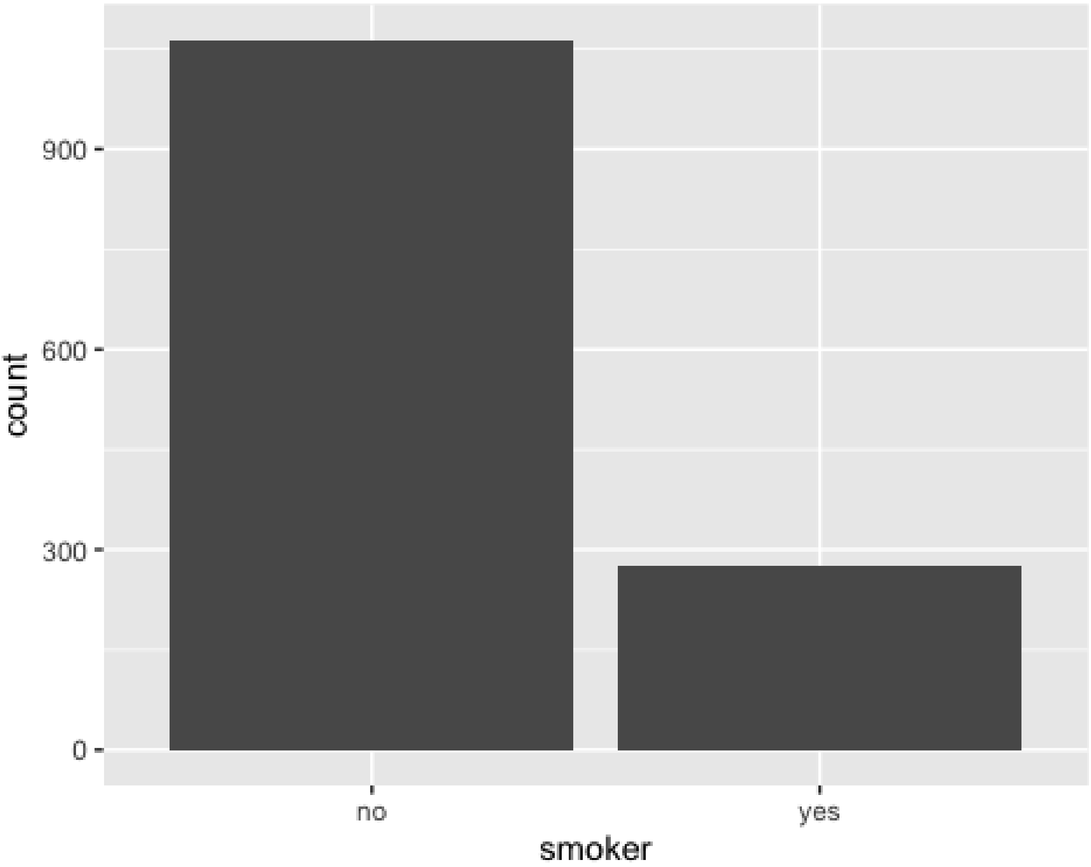
```r
hist(expenses$bmi)
```



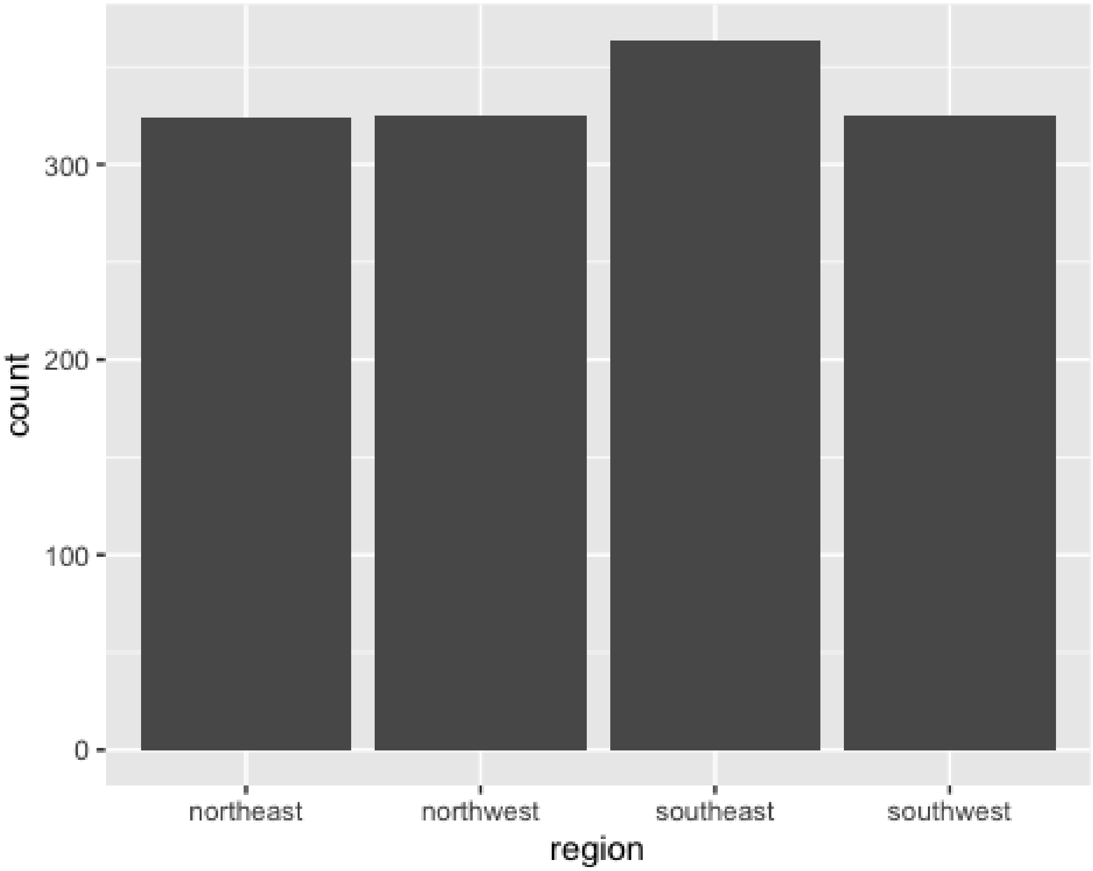
```
hist(expenses$children)
```



```
ggplot(expenses, aes(x = smoker)) + geom_bar()
```

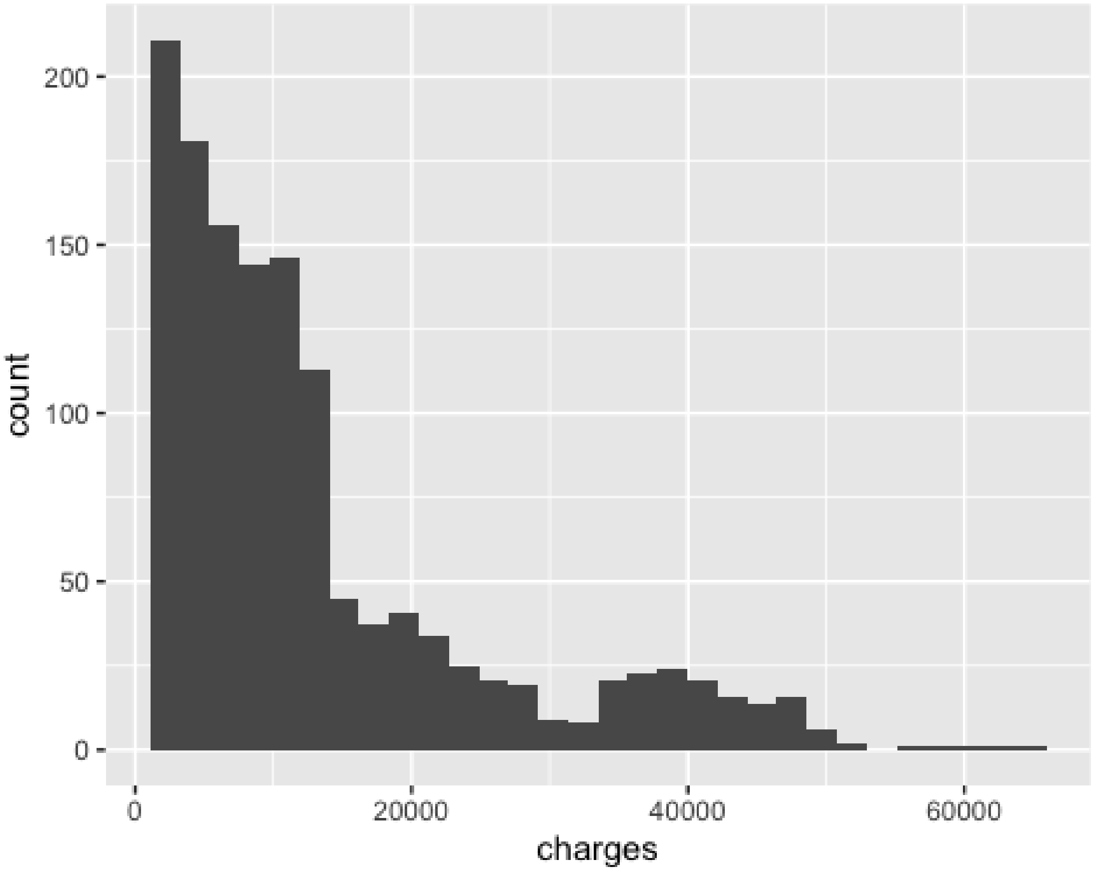


```
ggplot(expenses, aes(x = region)) + geom_bar()
```

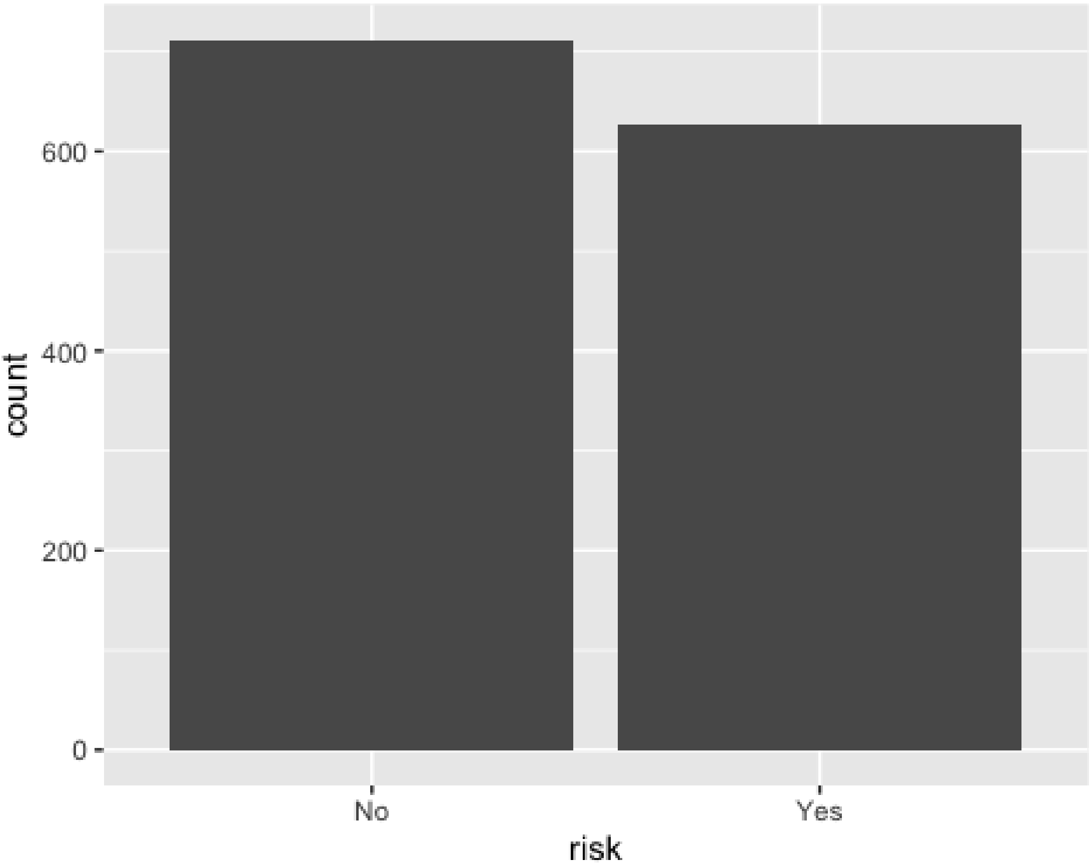


```
ggplot(expenses, aes(x = charges)) + geom_histogram()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

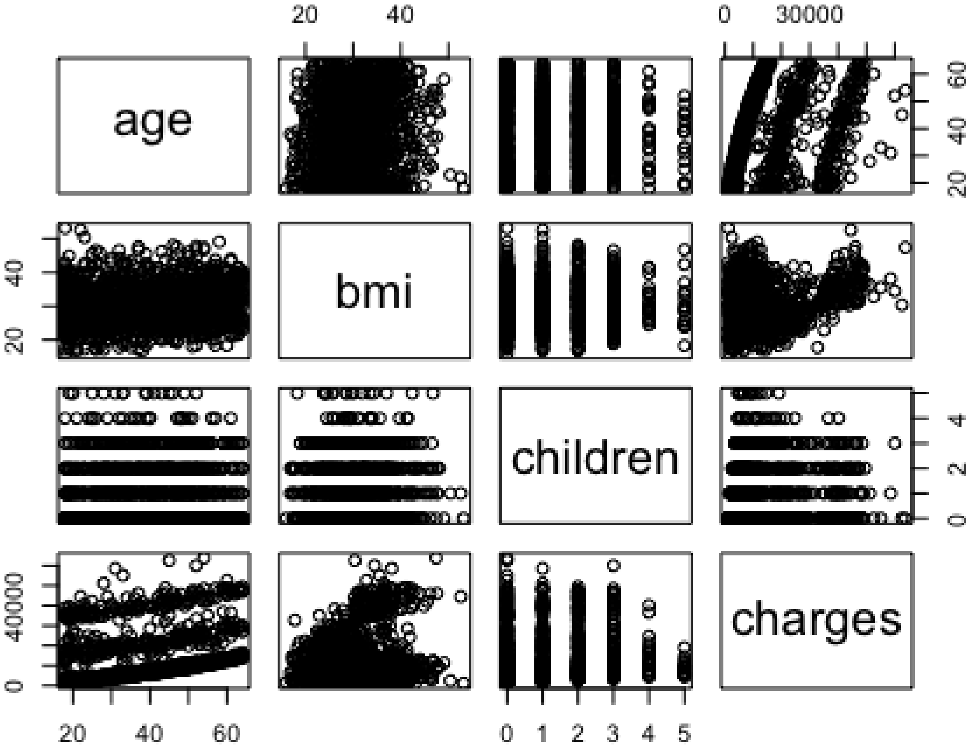


```r
ggplot(expenses, aes(x = risk)) + geom_bar()
```



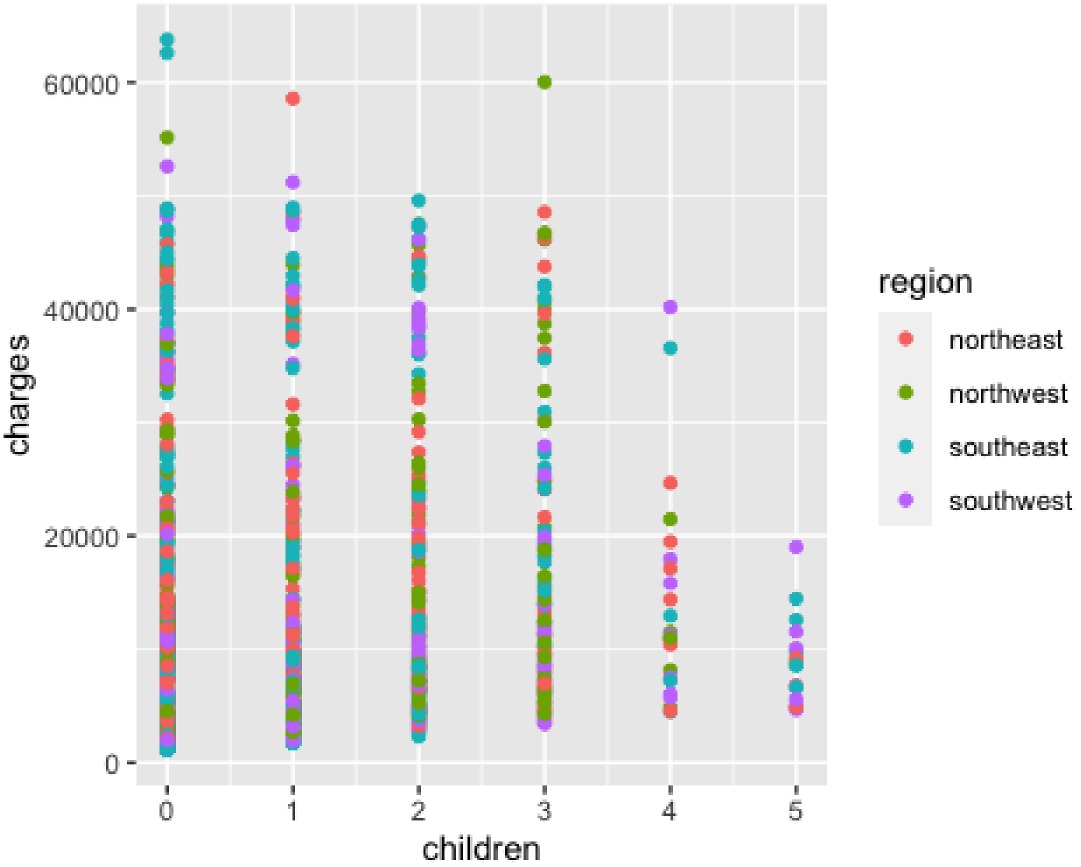
Let's create a plot to understand relation between all the numeric variables.

```
num_col <- unlist(lapply(expenses, is.numeric))
plot(expenses[,num_col])
```

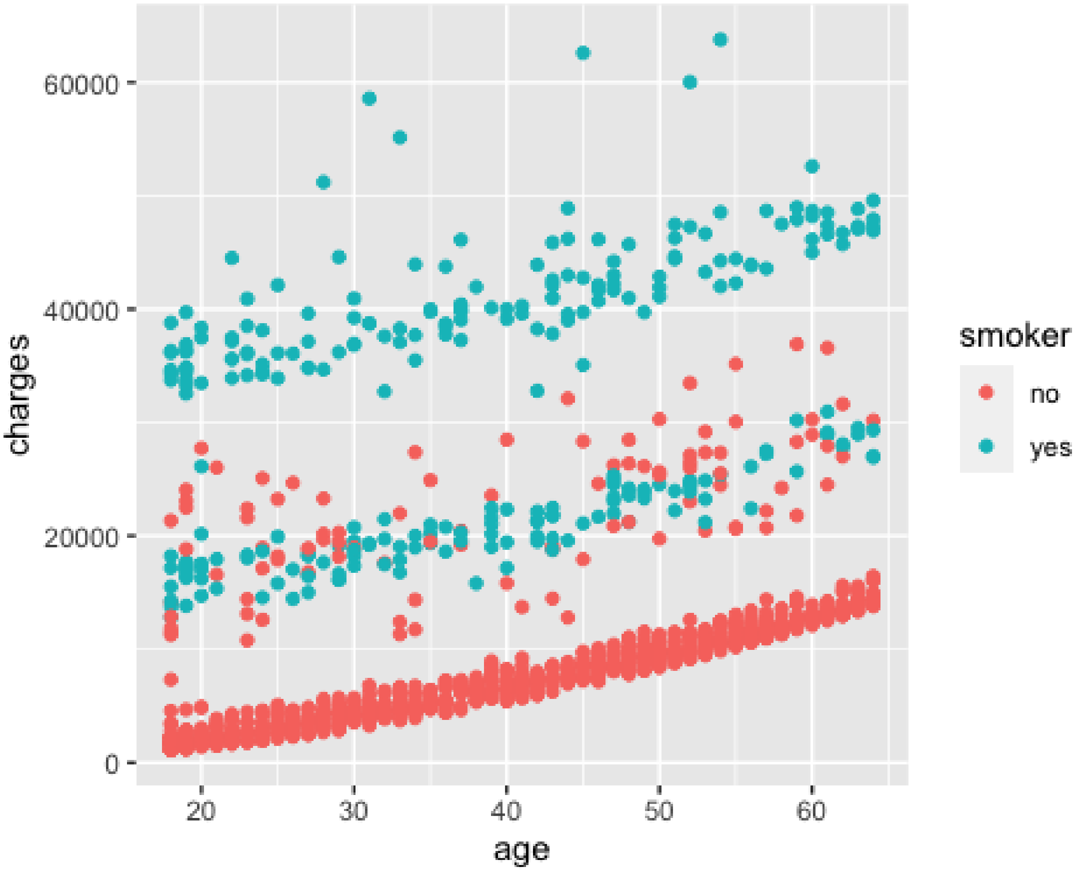


In this plot we can see some of the specific trends that can be useful later in the projects. 1. Interestingly, when the number of children is greater than 3 then charges is significantly low. 2. We can see three bands formation on the charges/age plot

```
ggplot(expenses, aes(x = children, y = charges)) +
  geom_point(aes(col = region))
```



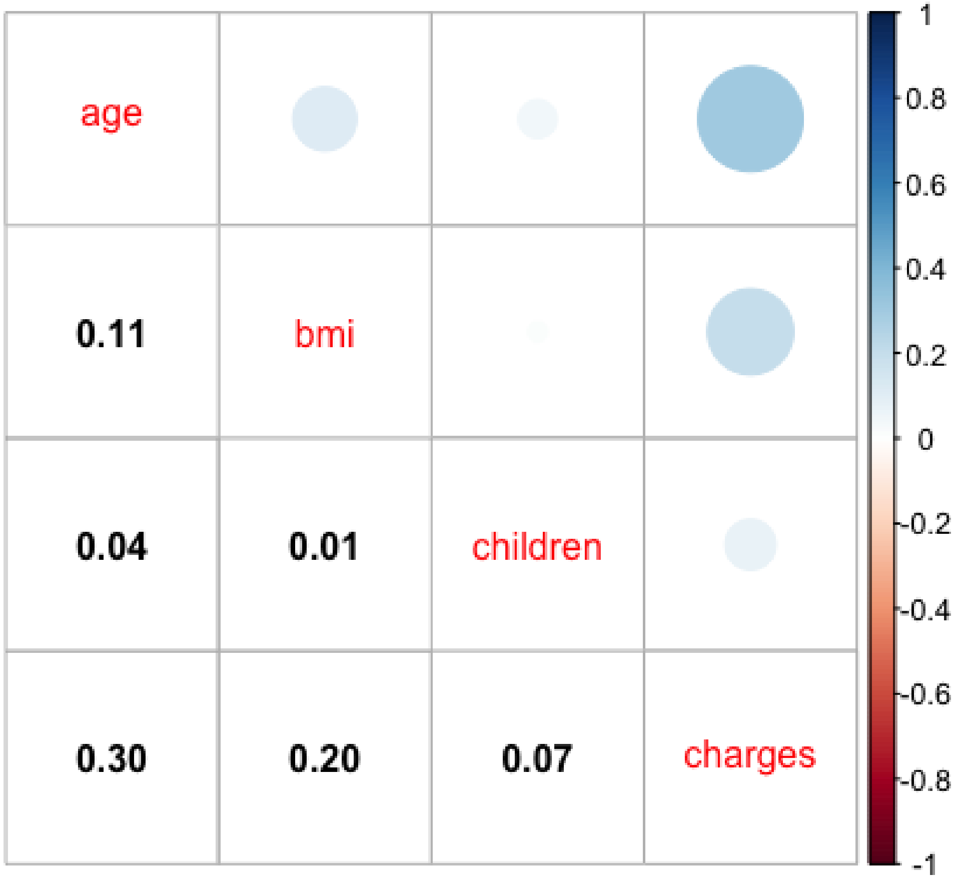
```r
ggplot(expenses, aes(x = age, y = charges)) +
  geom_point(aes(col = smoker))
```



In the first graph of children/charges there is no significant behavior. However, we can clearly see the three bands formation can be distinguished by smokers and non-smokers, smokers having higher charges and non-smokes having lower charges.

Correlation between continuous variables

```
corrplot.mixed(round(cor(expenses[,num_col]), 2), lower.col = "black")
```



Correlation between the charges and the other variables is not even greater than 50%, but still age is the factor with highest correlation of 0.30.

Distribution of Categorical Variables

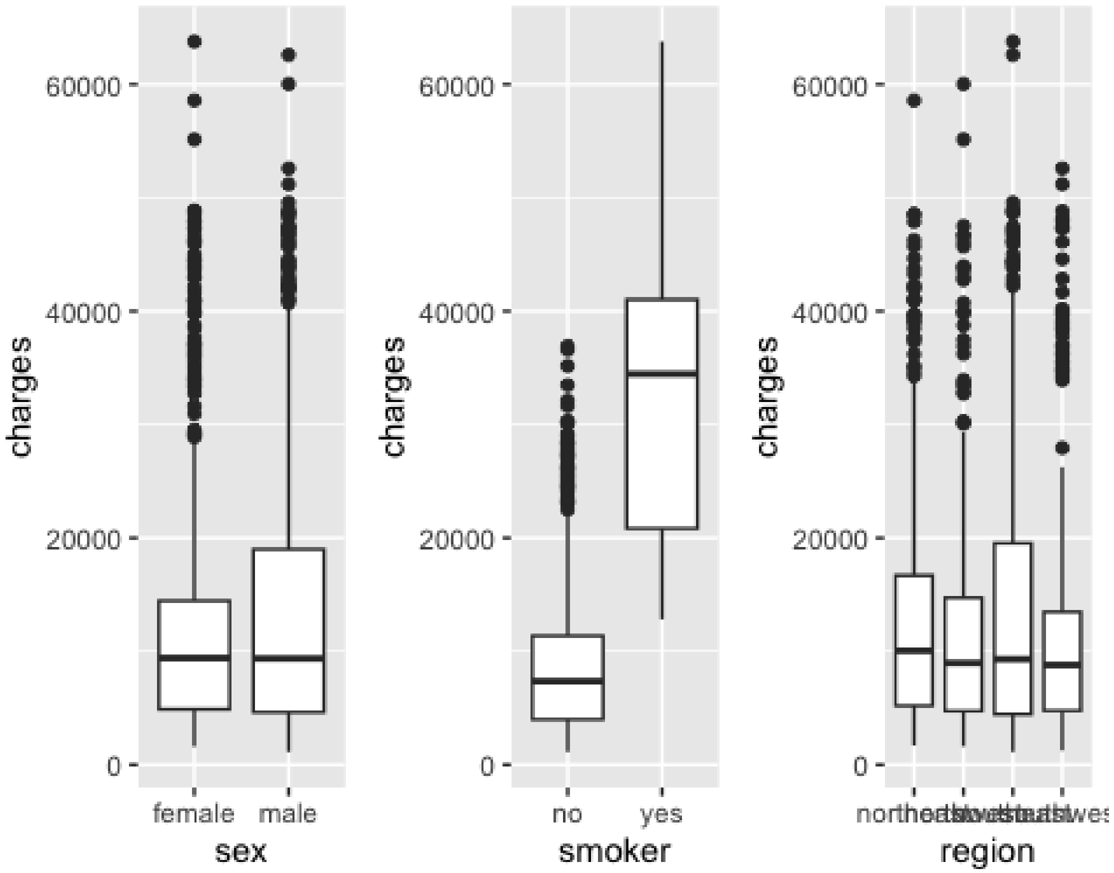
```
str(expenses)

## 'data.frame':    1338 obs. of  8 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 ...
##  $ risk    : chr  "Yes" "No" "No" "Yes" ...

expenses$sex <- as.factor(expenses$sex)
expenses$smoker <- as.factor(expenses$smoker)
expenses$region <- as.factor(expenses$region)
str(expenses)

## 'data.frame':    1338 obs. of  8 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3
2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
##  $ risk    : chr  "Yes" "No" "No" "Yes" ...
```

```
sex_boxplot <- ggplot(expenses, aes(x = sex, y = charges)) +
  geom_boxplot()
smoker_boxplot <- ggplot(expenses, aes(x = smoker, y = charges)) +
  geom_boxplot()
region_boxplot <- ggplot(expenses, aes(x = region, y = charges)) +
  geom_boxplot()
grid.arrange(sex_boxplot, smoker_boxplot, region_boxplot, ncol = 3)
```



Preparing dataset for Modelling

We will divide the dataset into 20-80 ratio. 20% for testing and 80% for training.

```
train_n <- round(0.8*nrow(expenses))
train_indices <- sample(1:nrow(expenses), train_n)
train_expenses <- expenses[train_indices, ]
test_expenses <- expenses[-train_indices, ]
```

Now will create a formula for our model to compare all the variable with charges

```
formula_1 <- as.formula("charges ~ age + sex + bmi + children + smoker +
region")
```

Building first Linear Regression Model

```
model_1 <- lm(formula_1, data = train_expenses)
summary(model_1)
```

```
##
## Call:
## lm(formula = formula_1, data = train_expenses)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11216.0  -2801.0   -964.8   1380.6  25846.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11464.15    1094.68 -10.473  < 2e-16 ***
## age                255.89      12.98  19.709  < 2e-16 ***
## sexmale           -120.82     368.39  -0.328 0.743000
## bmi                317.86      31.54  10.079  < 2e-16 ***
## children           566.12     151.85   3.728 0.000203 ***
## smokeryes        23581.09     457.08  51.591  < 2e-16 ***
## regionnorthwest   -431.98     527.84  -0.818 0.413314
## regionsoutheast   -546.15     530.74  -1.029 0.303693
## regionsouthwest  -1070.51     526.00  -2.035 0.042079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6005 on 1061 degrees of freedom
## Multiple R-squared:   0.75,  Adjusted R-squared:  0.7482
## F-statistic:   398 on 8 and 1061 DF,  p-value: < 2.2e-16
```

From the output above, following are the observations that can be made: Residuals:

The residuals represent the differences between the observed values and the values predicted by the model. The summary statistics for the residuals indicate that they range from -10,610 to 30,205, with quartiles at -3,057, -1,034, and 1,657, respectively. Coefficients:

Intercept: The intercept term (-12,584.74) represents the estimated value of the response variable when all predictor variables are zero. age: For each additional year of age, the estimated response variable increases by approximately $258.88. sexmale: The coefficient for the 'sex' variable indicates that being male is associated with a decrease in the estimated response variable by $115.15, but this effect is not statistically significant (p-value > 0.05). bmi: For each unit increase in BMI, the estimated response variable increases by $362.36. children: For each additional child, the estimated response variable increases by $430.89, and this effect is statistically significant at the 5% level (p-value < 0.05). smokeryes: Being a smoker is associated with a significant increase in the estimated response variable by $23,393.56. regionnorthwest, regionsoutheast, regionsouthwest: These coefficients represent the differences in the estimated response variable between the respective regions and the reference region (presumably northeast). However, only the coefficient for ‘regionsouthwest’ is statistically significant at the 5% level (p-value < 0.05). Residual standard error:

The residual standard error is approximately 6031, indicating the typical deviation of the observed values from the predicted values by the model. Multiple R-squared and Adjusted R-squared:

The multiple R-squared value of 0.7529 suggests that approximately 75.29% of the variance in the response variable is explained by the predictors included in the model. The adjusted R-squared value of 0.751 is similar, indicating that the model’s explanatory power remains high even after adjusting for the number of predictors. F-statistic:

The F-statistic tests the overall significance of the model. In this case, the F-statistic is 404 with a very low p-value (< 2.2e-16), indicating that the model as a whole is highly significant.

Overall, this model suggests that age, BMI, number of children, smoking status, and region (specifically southwest) are significant predictors of the response variable, while gender does not appear to be a significant predictor in this model. The model as a whole is highly significant in explaining the variance in the response variable.

For improving this model, we will eliminate the less significant variables here (sex)

```
formula_2 <- as.formula("charges ~ age + bmi + children + smoker + region")
model_2 <- lm(formula_2, data = train_expenses)
summary(model_2)

##
## Call:
## lm(formula = formula_2, data = train_expenses)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11271.5  -2768.2   -971.3   1404.4  25917.4
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11518.73    1081.50 -10.651  < 2e-16 ***
## age                256.06      12.97  19.747  < 2e-16 ***
## bmi                317.54      31.51  10.078  < 2e-16 ***
## children           565.16     151.76   3.724 0.000206 ***
## smokeryes        23572.86     456.20  51.673  < 2e-16 ***
## regionnorthwest   -429.90     527.58  -0.815 0.415336
## regionsoutheast   -547.65     530.50  -1.032 0.302145
## regionsouthwest  -1070.71     525.78  -2.036 0.041954 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6002 on 1062 degrees of freedom
## Multiple R-squared:   0.75,  Adjusted R-squared:  0.7484
## F-statistic: 455.2 on 7 and 1062 DF,  p-value: < 2.2e-16
```

Comparing both the Models

Adjusted R-squared: Both models have very similar adjusted R-squared values, indicating that they explain a similar proportion of variance in the response variable.

F-statistic: Model 2 has a higher F-statistic (462.1) compared to Model 1 (404), suggesting that Model 2 is better at explaining the variability in the response variable compared to Model 1.

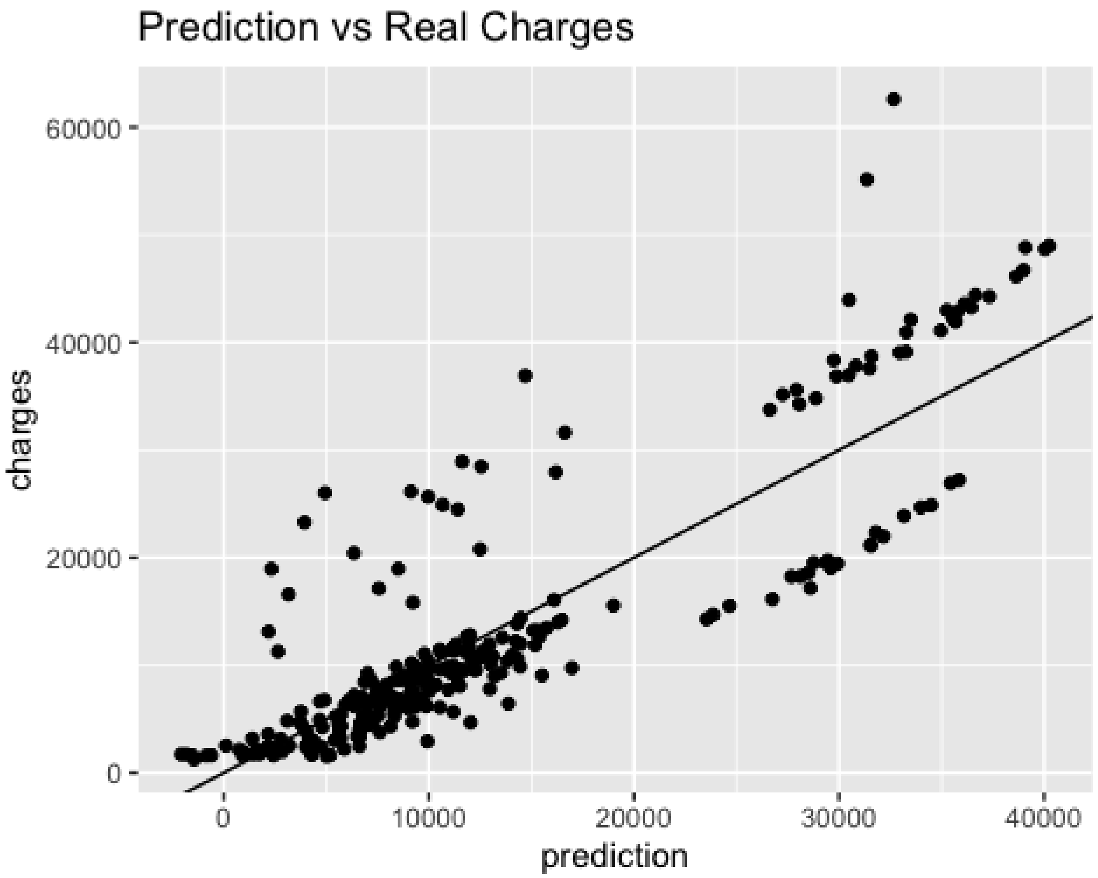
Residual standard error: Both models have similar residual standard errors, suggesting that they have similar levels of variation in their residuals.

Number of predictors: Model 1 includes an additional predictor ('sex') compared to Model 2.

Based on these comparisons, Model 2 appears to be slightly better than Model 1. It has a higher F-statistic, indicating a better overall fit to the data. Additionally, both models have similar adjusted R-squared values and residual standard errors, suggesting that Model 2 achieves comparable performance with fewer predictors, which is generally preferable as it may reduce complexity and overfitting.

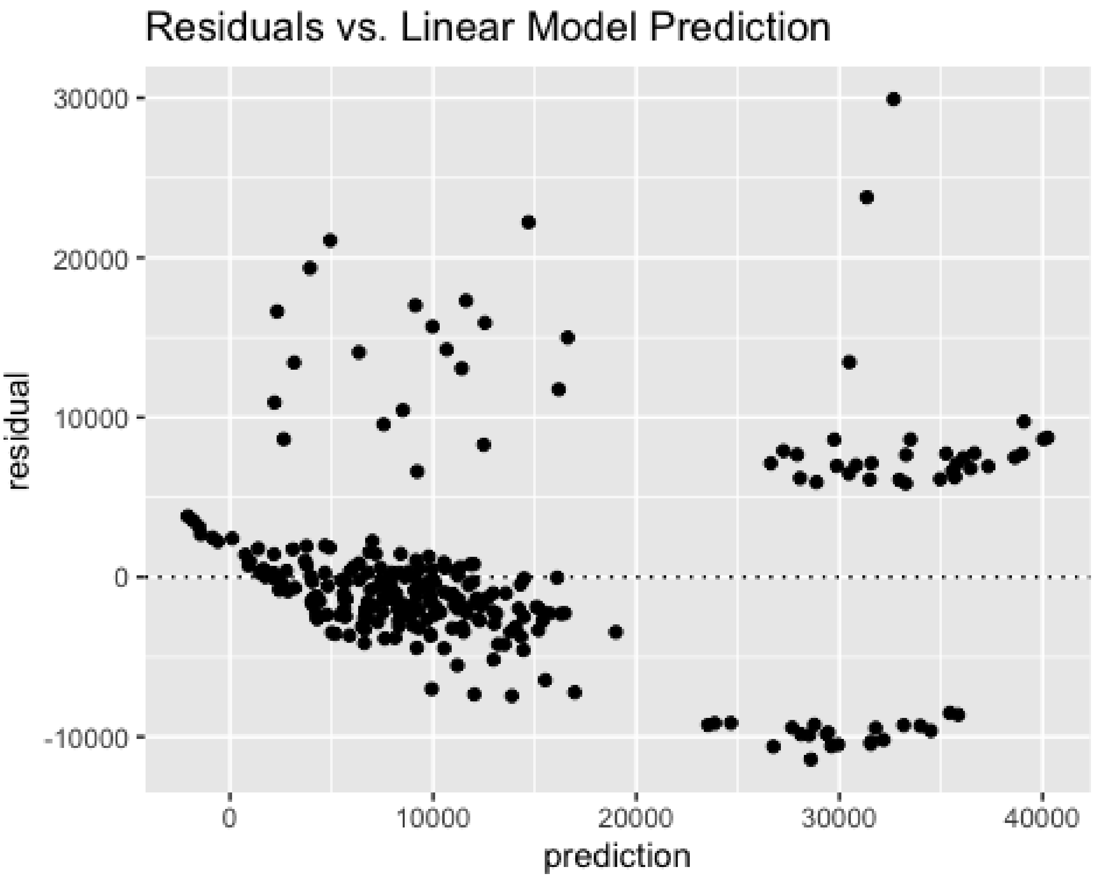
Therefore, Model 2 is considered better because it achieves similar or slightly better performance with fewer predictors, which can lead to a more parsimonious and interpretable model.

```r
test_expenses$prediction <- predict(model_2, newdata = test_expenses)
ggplot(test_expenses, aes(x = prediction, y = charges)) +
  geom_point() +
  geom_abline() +
  ggtitle("Prediction vs Real Charges")
```



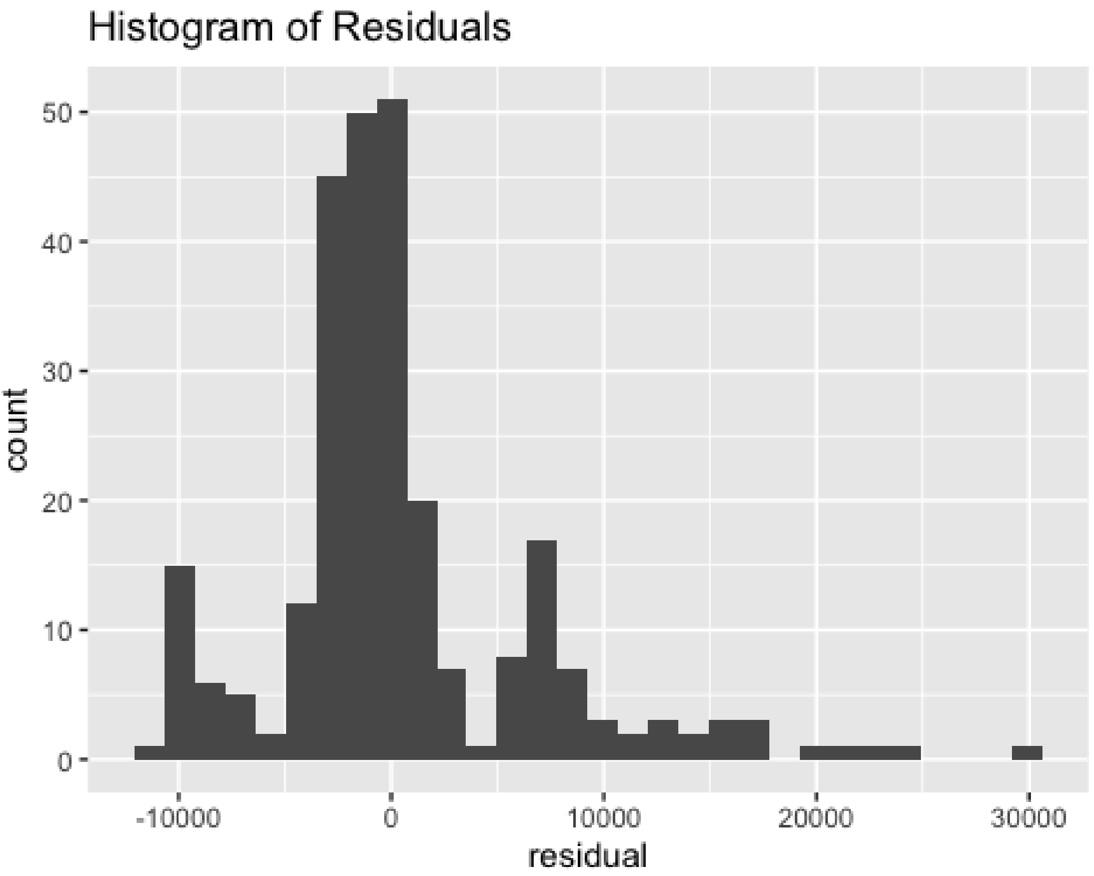
```r
test_expenses$residual <- test_expenses$charges - test_expenses$prediction

ggplot(test_expenses, aes(x = prediction, y = residual)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = 3) +
  ggtitle("Residuals vs. Linear Model Prediction")
```



```
ggplot(test_expenses, aes(residual)) +
  geom_histogram() +
  ggtitle("Histogram of Residuals")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Error in all the above graphs is close to zero, that means the linear model is giving out pretty accurate charges of healthcare based on the factors.

Now we will test the model with some dummy observations.

```r
dummy_data <- data.frame(age = 29,
                         bmi = 27,
                         children = 0,
                         smoker = "no",
                         region = "northeast")
dummy_data$predicted_charges <- round(predict(model_2, dummy_data), 2)
```

We have successfully created a data frame with the new dummy data and our predicting model to predict the charges.

```r
new_row <- data.frame(age = 30,
                      bmi = 25,
                      children = 2,
                      smoker = "yes",
                      region = "southeast")
new_row$predicted_charges <- round(predict(model_2, new_row), 2)
dummy_data <- rbind(dummy_data, new_row)

head(dummy_data)
```

```
##   age bmi children smoker    region predicted_charges
## 1  29  27        0     no northeast           4480.47
## 2  30  25        2    yes southeast          28256.97
```