# Analysis of Insurance Charges

**Group 7**

## Introduction

This project is based on observing the trends of the Annual Medical Cost associated with the Primary Insurance Holder, based on multiple factors such as their *Age* in years, *Sex*, *BMI*, No. of *Children*, and *Smoking* tendency.

This data is simulated based on Census responses and was taken from Kaggle at: "https://www.kaggle.com/mirichoi0218/insurance".

The population of study is the general population of the United States.

This dataset and population were selected for analysis because medical charges in the United States are a hot topic for most adults in the country. The goal of this study is to identify relationships between certain demographic factors and the amount of medical charges that are incurred as a result.

**The relationships to be analysed in this paper are:**

1) The effect of *Age* on medical charges

2) The effect of *BMI* on medical charges.

A multiple linear regression will be conducted at the end to try to identify the variables with the most impact on medical charges.

## Methods

This analysis is an observational study.

The data doesn't signify any control over extraneous variables on the insurance holders such as: how healthy they have been, their eating habits, their exercising frequencies etc.

The sampling unit for this observational study is an individual with insurance (Primary Insurance Holder) that was canvased during the Census.

### Variables

There are 6 variables in the data considered, in which 2 are continuous variables and 3 are categorical variables. The descriptions of the variables are mentioned below:

***Independent Variables***

- **Continuous Predictors:**
    - *Age (discrete)*: Age of Primary Insurance Holder in **years**.
    - *BMI (continuous)*: Body Mass Index, henceforth referred to as BMI, is the ratio of the weight of an individual to the square of their height. The units are in Metric system (**kg / m ^ 2**).

- **Categorical Predictors:**
    - *Sex (categorical)*: Gender of the Primary Insurance Holder, Male or Female.
    - *Smoker (categorical)*: If the Primary Insurance Holder is a Smoker or Non-Smoker.
    - *Children (categorical)*: Number of children covered by health insurance; ranging from 0 to 5.

***Dependent Variable***

- *Charges (continuous)*: **Dollar** amount of medical costs for each Primary Insurance Holder charged by Health Insurers.

### Data Pre-processing:

The data does not contain any missing values, hence no imputation of the data was required. The following data describes the

**Age**: None
**Sex**: Converted from strings to factors with levels: Male/Female
**BMI**: None
**Children**: Converted from integer to factors with levels: 0,1,2,3,4,5
**Smoker**: Converted from string format to factors with levels: Yes/No
**Charges**: None

## Results

Three simple linear regression analysis is performed to study the effect of age of the primary insurance holder and the effect of the BMI of the primary insurance holders on the amount of charges.

The total number of observations in the data set are 1338.

**Table 1. Summary Statistics of Independent and response Variables**

```
        Count     Mean  Median    St.Dev.           Pearson Correlation w/ Charges
Age      1338    39.21   39.00     14.05                           0.2990
BMI      1338    30.66   30.40      6.10                           0.1983
Charges  1338 13270.42 9382.03   12110.01                          1.0000
```

**Table 2. Sex**

```
sex
  male female
   676    662
```

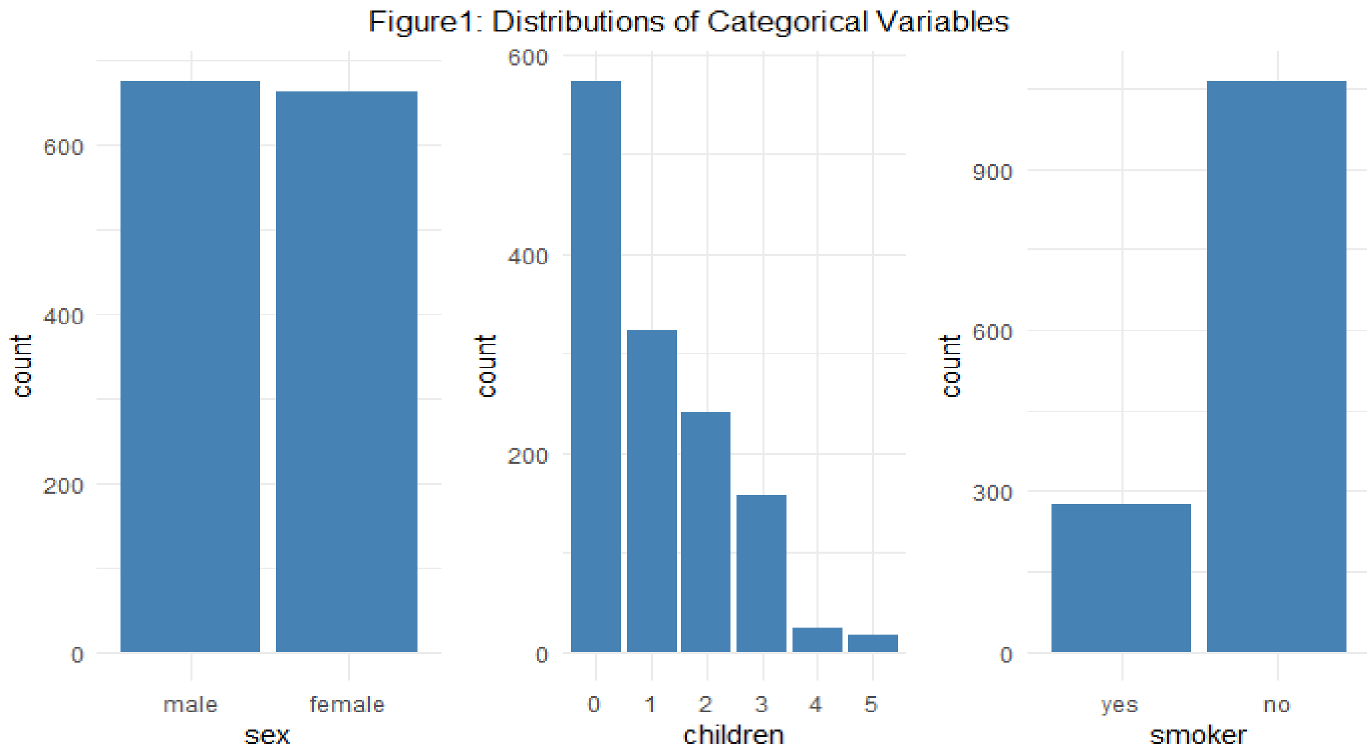
**Table 3. Smoker**

```
smoker
 yes   no
 274 1064
```

**Table 4: Children**

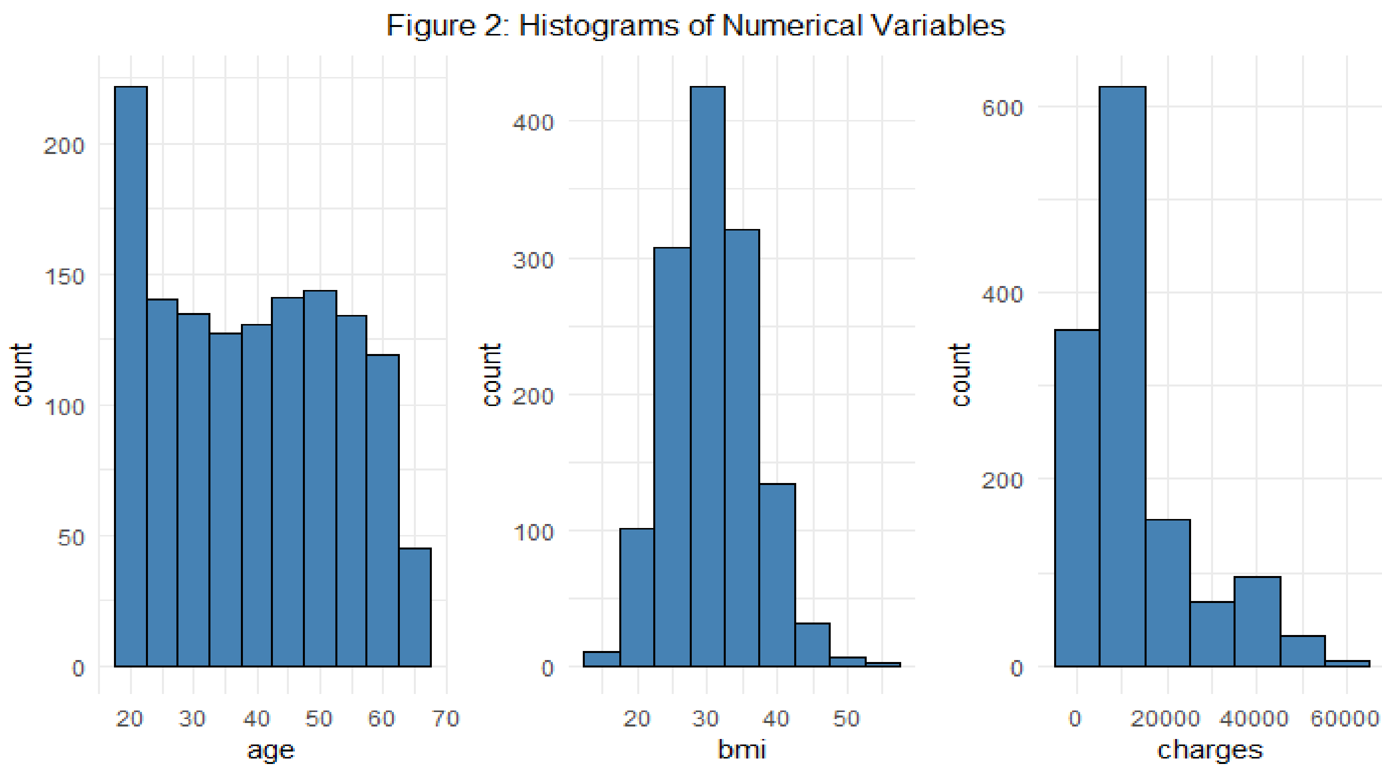
```
children
  0   1   2   3   4   5
574 324 240 157  25  18
```

**Distributions:**




From Figure 1 – The distributions of the categorical variables, we can observe that the variable Sex is balanced, Children and Smoker are unbalanced.

**Histograms:**




From Figure 2 – The Histograms of the Numerical variables, it can be observed that the variable Age and BMI seem normal, but the response variable Charges seems to be right skewed.

**Hypothesis Test 1: A simple linear regression analysis on Age and Charges –**

As *Age* increases, does the amount of *Charges* of the Primary Insurance Holder increase.

**a priori hypothesis**: Medical *Charges* of the older Primary Insurance Holders will be higher than younger Primary Insurance Holders.

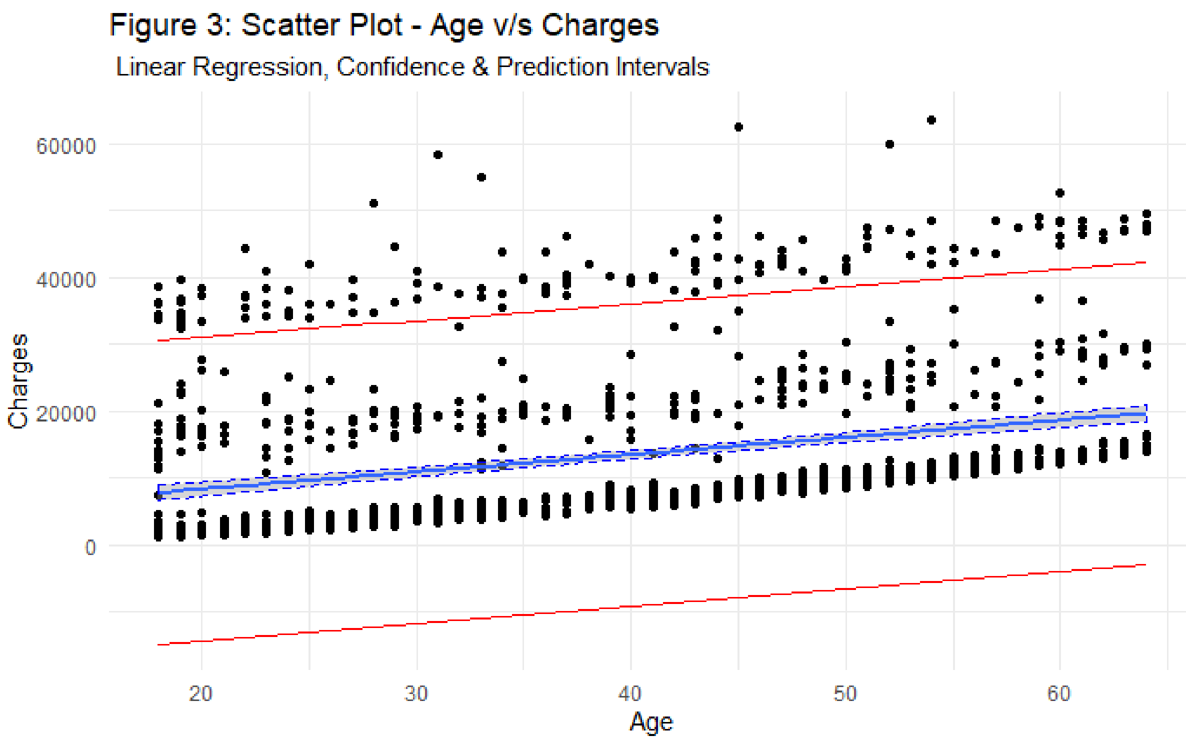
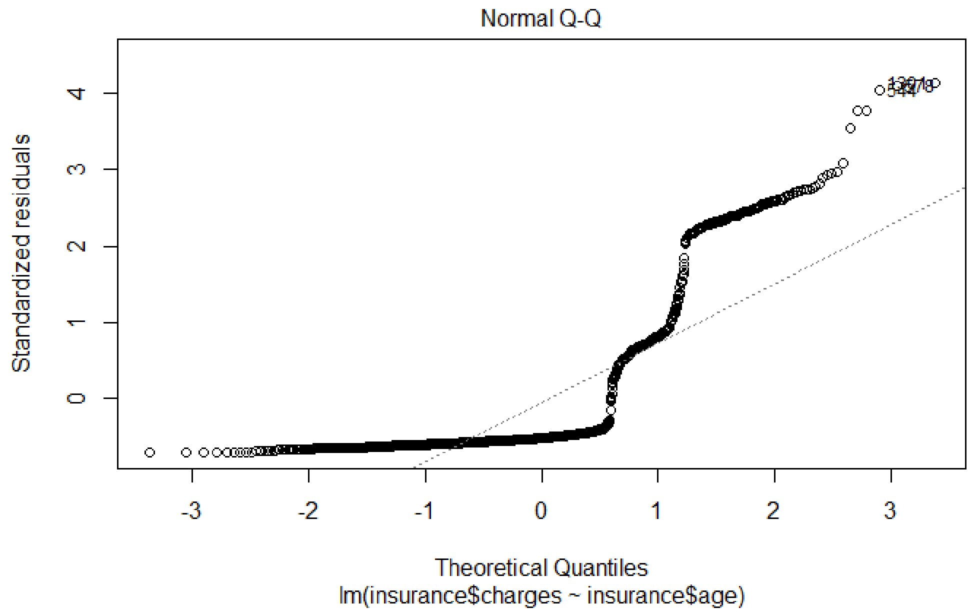



Figure 3 shows the scatter plot between the Dependent variable – Charges and the Independent variable – Age.

**Table 5:** Summary of Linear Model between Age and Charges:

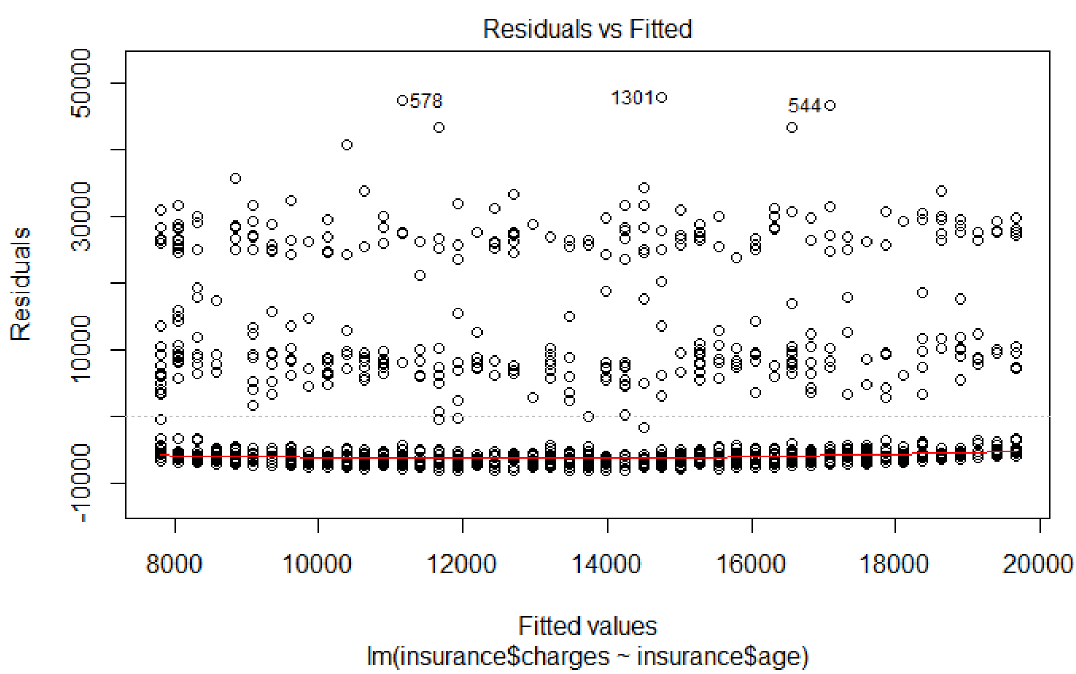
```
                   2.5 %    97.5 %
(Intercept)    1327.4403 5004.3297
insurance$age   213.5788  301.8665

Call:
lm(formula = insurance$charges ~ insurance$age)

Residuals:
   Min     1Q Median     3Q    Max 
 -8059  -6671  -5939   5440  47829 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3165.9      937.1   3.378 0.000751 ***
insurance$age    257.7       22.5  11.453  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1336 degrees of freedom
Multiple R-squared:  0.08941,	Adjusted R-squared:  0.08872 
F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

**Figure 4:** Normal Q-Q Plot of the Residuals:




As the points do not lie on the linear regression line, we can conclude that Normality assumption is violated.

**Figure 5:** Residuals v/s Fitted Plot:




From Figure 5, we can conclude that the Homoscedasticity assumption is not satisfied.

**Hypothesis Test 2: A simple linear regression analysis on the BMI of insurance holders and charges**

As *BMI* increases in Primary Insurance Holders so do medical *Charges*. **a priori hypothesis** this is true.

Figure 6: Scatter Plot - BMI v/s Charges

Linear Regression, Confidence & Prediction Intervals

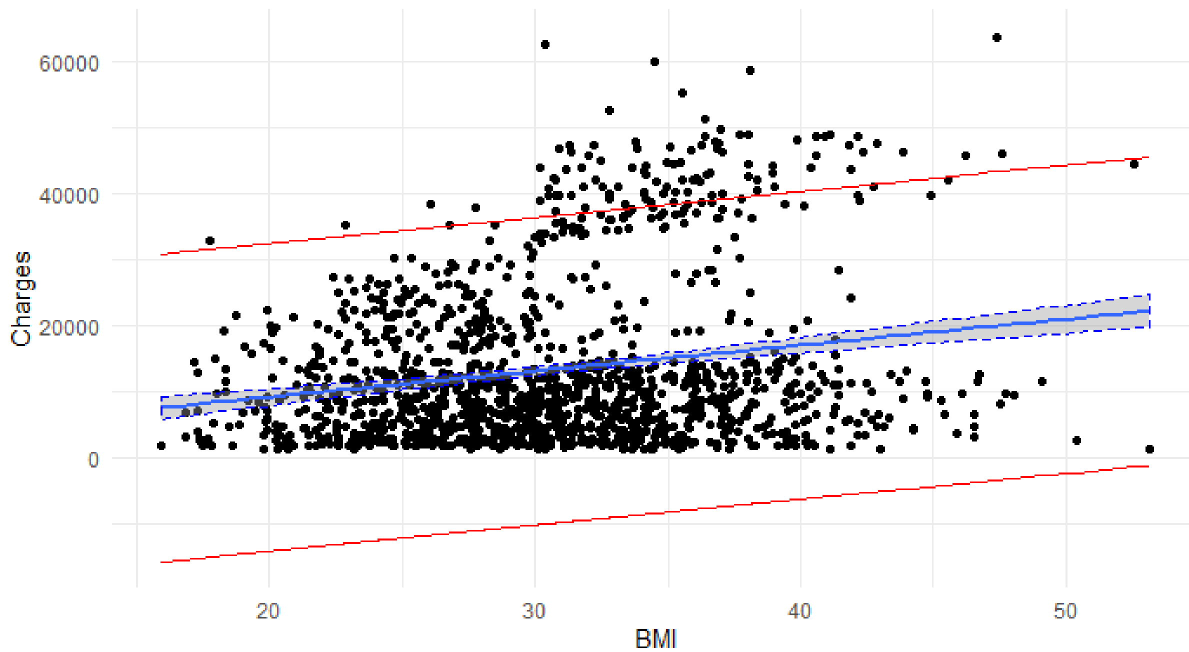
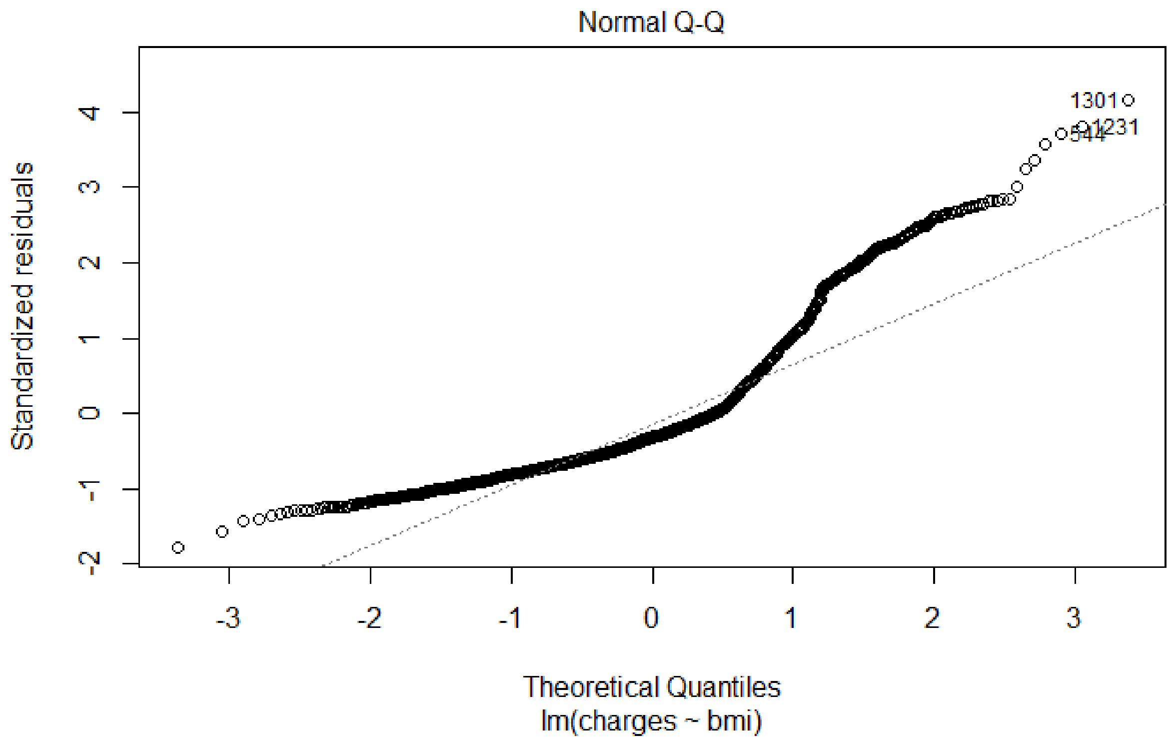



Figure 6 shows the scatter plot between the Dependent variable – Charges and the Independent variable – BMI.

**Table 6:** Summary of Linear Model between BMI and Charges:

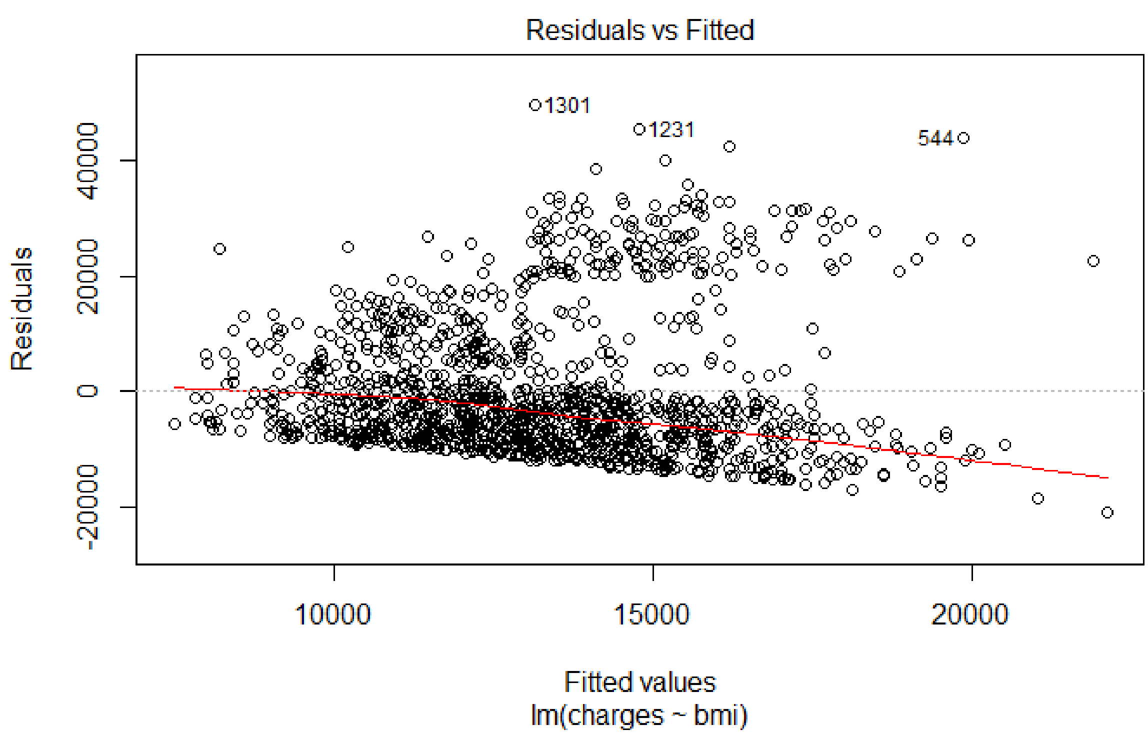
```
Call:
lm(formula = charges ~ bmi)

Residuals:
   Min     1Q Median     3Q    Max
-20956  -8118  -3757   4722  49442

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1192.94    1664.80   0.717    0.474
bmi           393.87      53.25   7.397 2.46e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11870 on 1336 degrees of freedom
Multiple R-squared:  0.03934,   Adjusted R-squared:  0.03862
F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

**Figure 7:** Normal Q-Q Plot of the Residuals:




From Figure 7, it can be observed that the points do not lie on the linear regression line, we can conclude that Normality assumption is violated.

**Figure 8:** Residuals v/s Fitted Plot




From Figure 8, we can conclude that the Homoscedasticity assumption is not satisfied.

**Transformations of the Variables:**

For the continuous variables, which are Age, BMI and Charges the skewness is found to be:

**Age** – 0.05554775

**BMI** – 0.2834106

**Charges** – 1.512483

As both Age and BMI have an almost symmetric distribution, there is no transformation required for these variables. Since Charges variable is not symmetric, log transformation is applied to the response variable – Charges.

The skewness value found after applying the transformation to the Charges is -0.08989561, which is approximately symmetric.

**Hypothesis Test 1 - Transformed:**
**A simple linear regression analysis on Age and transformed Charges**

As *Age* increases, does the amount of *Charges* of the Primary Insurance Holder increase.

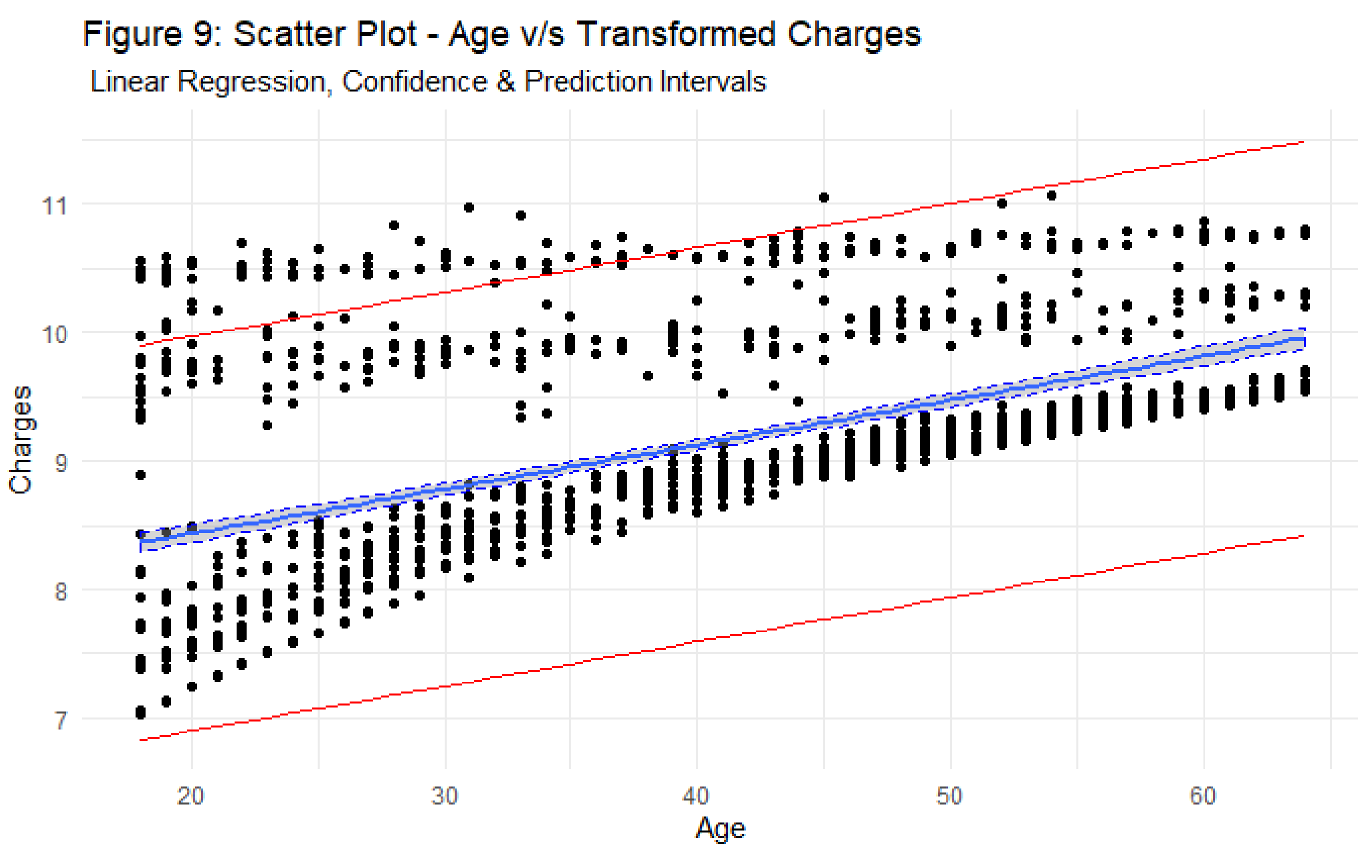



Figure 9 shows the scatter plot between the transformed Dependent variable – Log Charges and the Independent variable – Age.

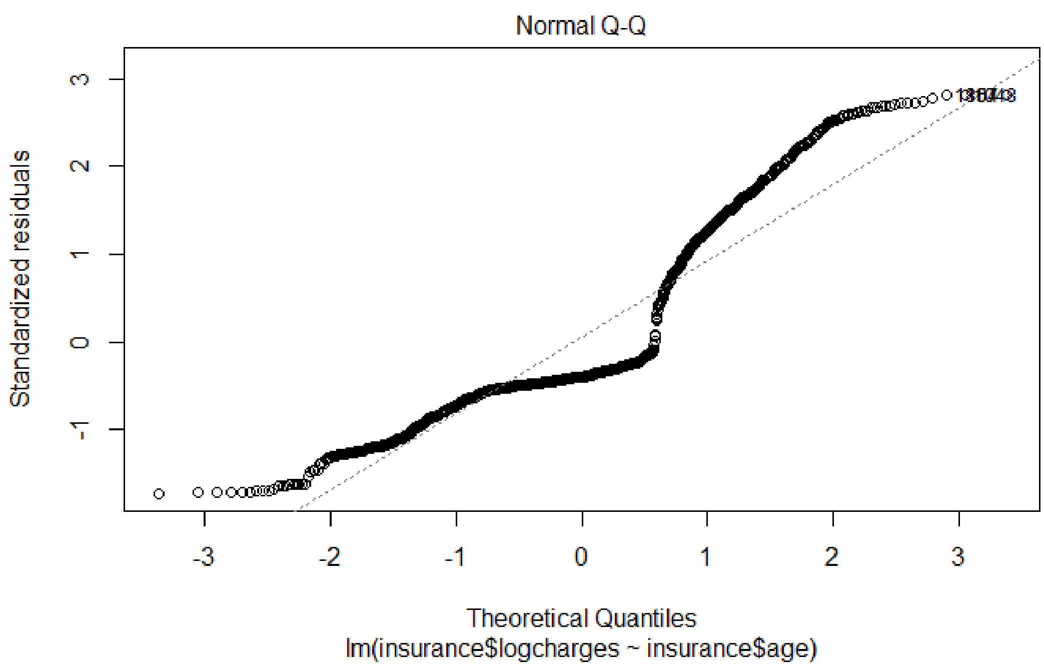
**Table 7:** Summary of Linear Model between Age and Log Charges:

```
Call:
lm(formula = insurance$logcharges ~ insurance$age)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3433 -0.4166 -0.3094  0.5000  2.1999

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.744247   0.063336  122.27   <2e-16 ***
insurance$age 0.034545   0.001521   22.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7813 on 1336 degrees of freedom
Multiple R-squared:  0.2786,    Adjusted R-squared:  0.2781
F-statistic:   516 on 1 and 1336 DF,  p-value: < 2.2e-16
```

**Inference of the Summary Statistics:**

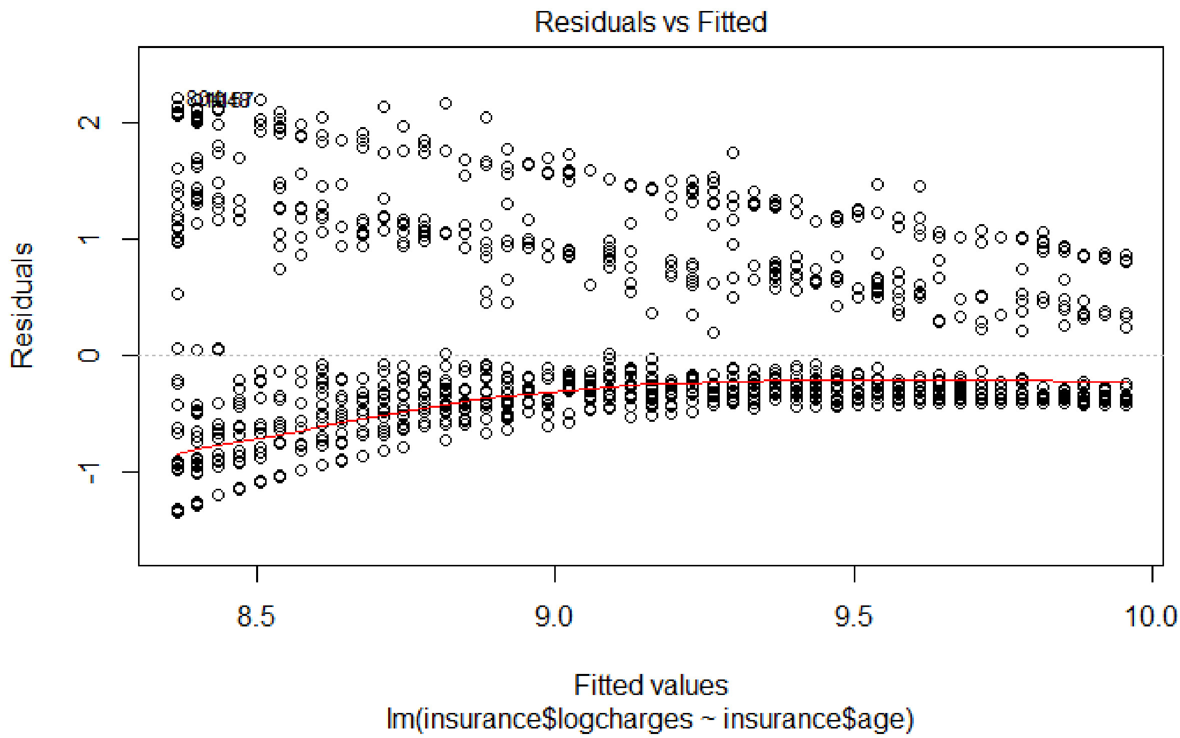
- The P-value of the predictor Age is found to be less than 0.05 and this means that Age is significantly affecting the dependent variable – Charges.
- Since the slope is found to be positive, it can be inferred that as the Age increases, the amount of Charge increases.

**Figure 10:** Normal Q-Q Plot of the Residuals:




The normality assumption is roughly satisfied as the point lie around the linear line.

**Figure 11**: Residuals v/s Fitted Plot:




From Figure 11, we can conclude that the Homoscedasticity violation is reduced than the previous non-transformed data.

**Hypothesis Test 2 - Transformed:**
**A simple linear regression analysis on the BMI of insurance holders and charges**

As *BMI* increases in Primary Insurance Holders so do medical *Charges*. **a priori hypothesis** this is true.

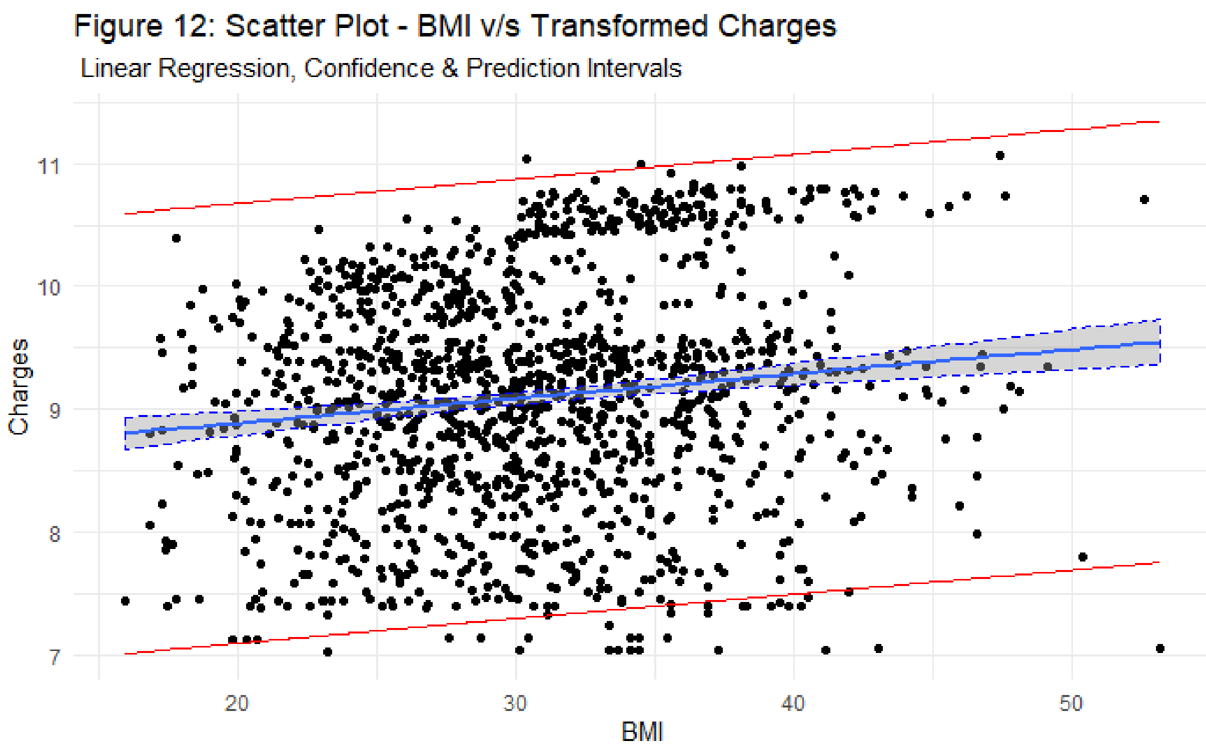



Figure 12 shows the scatter plot between the transformed Dependent variable – Log Charges and the Independent variable – BMI.

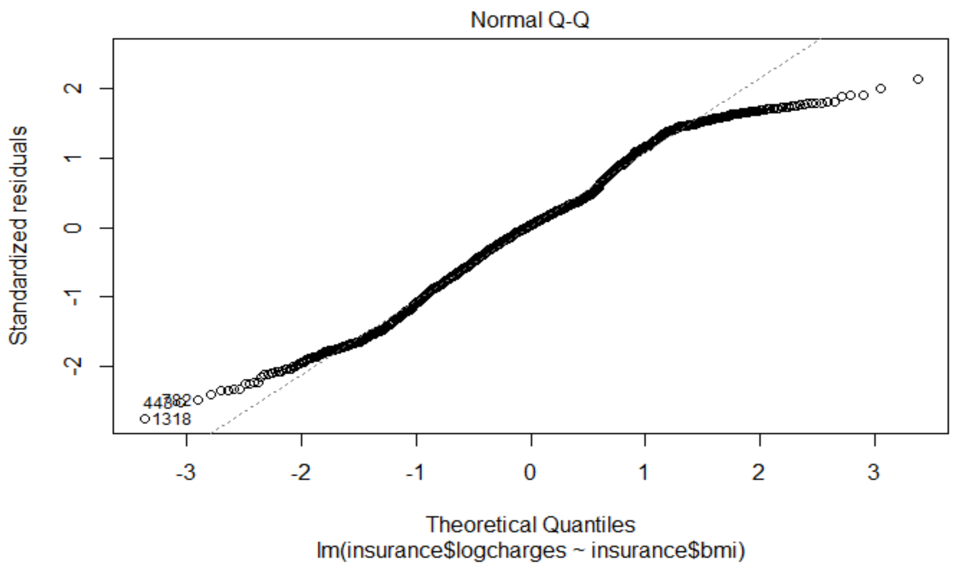
**Table 8:** Summary of Linear Model between BMI and Log Charges:

```
Call:
lm(formula = insurance$logcharges ~ insurance$bmi)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.48894 -0.63536  0.03136  0.68007  1.95182 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.485243   0.127833  66.378  < 2e-16 ***
insurance$bmi 0.020005   0.004089   4.892 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9117 on 1336 degrees of freedom
Multiple R-squared:  0.0176,	Adjusted R-squared:  0.01687 
F-statistic: 23.94 on 1 and 1336 DF,  p-value: 1.117e-06
```

**Inference of the Summary Statistics:**

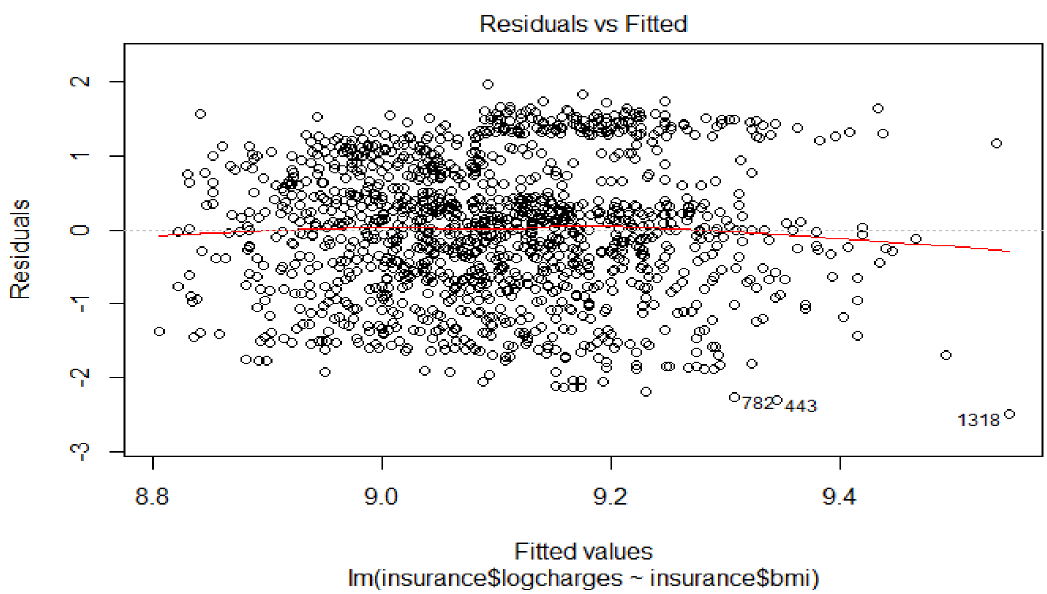
- The P-value of the predictor BMI is found to be less than 0.05 and this means that BMI is significantly affecting the dependent variable – Charges.
- Since the slope is found to be positive, it can be inferred that as the BMI increases, the amount of Charge increases.

**Figure 13:** Normal Q-Q Plot of the Residuals:




From Figure 13, it can be observed that the points mostly lie on the linear regression line, we can conclude that Normality assumption is roughly satisfied.

**Figure 14:** Residuals v/s Fitted Plot




From Figure 14, we can conclude that the Homoscedasticity assumption is satisfied.

## Multiple Linear Regression:

For finding the effects of the multiple variables on the transformed response variable, Multiple Linear Regression is modeled.

**Table 9:** Summary of Multiple Linear Model for Log Charges:

```
Call:
lm(formula = insurance$logcharges ~ insurance$age + insurance$bmi +
    insurance$children + insurance$smoker + insurance$sex)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05897 -0.20500 -0.05249  0.07884  2.10863

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.4741318  0.0744164 113.874  < 2e-16 ***
insurance$age       0.0347496  0.0008769  39.628  < 2e-16 ***
insurance$bmi       0.0107164  0.0020170   5.313 1.26e-07 ***
insurance$children1  0.1436991  0.0310272   4.631 3.99e-06 ***
insurance$children2  0.2795804  0.0343352   8.143 8.82e-16 ***
insurance$children3  0.2502409  0.0403144   6.207 7.19e-10 ***
insurance$children4  0.5249055  0.0912392   5.753 1.09e-08 ***
insurance$children5  0.4035001  0.1069707   3.772 0.000169 ***
insurance$smokerno  -1.5508632  0.0304129 -50.994  < 2e-16 ***
insurance$sexfemale  0.0752367  0.0245148   3.069 0.002191 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4463 on 1328 degrees of freedom
Multiple R-squared:  0.7661,    Adjusted R-squared:  0.7645
F-statistic: 483.2 on 9 and 1328 DF,  p-value: < 2.2e-16
```

**Inferences of Summary Table 9:**

It can be observed that all the independent variables are significantly affecting the amount of Charges.

The most significant predictors effecting the amount of Charges are Age and Non – Smokers as they have the least P-Value.

As the Multiple R-Squared Value is found to be 0.7661, it can be concluded that there is a decent linear relationship between the Dependent and Independent Variables.