Data Analysis Project for Medical Insurance Forecast

By

Ayush Sharma 8A

Nittin Kakkar 21A

About Dataset

- **Despite knowing little about the insured population, insurance companies must determine premiums based on demographic trends to make a profit.**
- **Our objective is to estimate the cost that the insured will incur.**
- **Using Different Linear Regression and Methods available in R.**
- **The link to original dataset can be found here.**

DATASET

- This dataset consists of1338 observations on 7 variables
- Dependent variable - charges
- Independent variables – age, bmi, smoker ,etc.
- Also, the dataset contains some categorical variable like sex, children(number of children), and smoker, region.

DATASET

•Age: the age of the insured (recipients).

•Sex: sex of insured persons; “male” or “female”.

•bmi: body mass index, providing an understanding of the body, relatively high or low weights relative to height, objective body weight index (kg / m ^ 2) using the height / weight ratio.

•children: number of children covered by health insurance / number of dependents.

•smoker: does the insured smoke or not.

•region: the recipient’s residential area in the United States; northeast, southeast, southwest, northwest.

•charges: Individual medical costs billed by health insurance.

Summary of Data

```
> summary(dataset)
      age            sex                 bmi           children        smoker             region         
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338       
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character  
 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                                        
    charges     
 Min.   : 1122  
 1st Qu.: 4740  
 Median : 9382  
 Mean   :13270  
 3rd Qu.:16640  
 Max.   :63770  
> dim(dataset)
[1] 1338    7
> str(dataset)
'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 ...
> 
```

Objective

- Examine the relationship between the cost of insurance and the variables that affect it, including age, sex, BMI, the number of children covered by health insurance, smoking, and geographic location.

- We will examine and make predictions about the factors influencing health insurance premiums in this paper. Linear Regression will be used to achieve it. Data preparation, exploratory data analysis, model building, prediction alternative model, and conclusion are all steps in the process.

- **Response Variable**: Charges

- **Predictor Variables**: Age, BMI, children, smoker

DATA PRE - PROCESSING

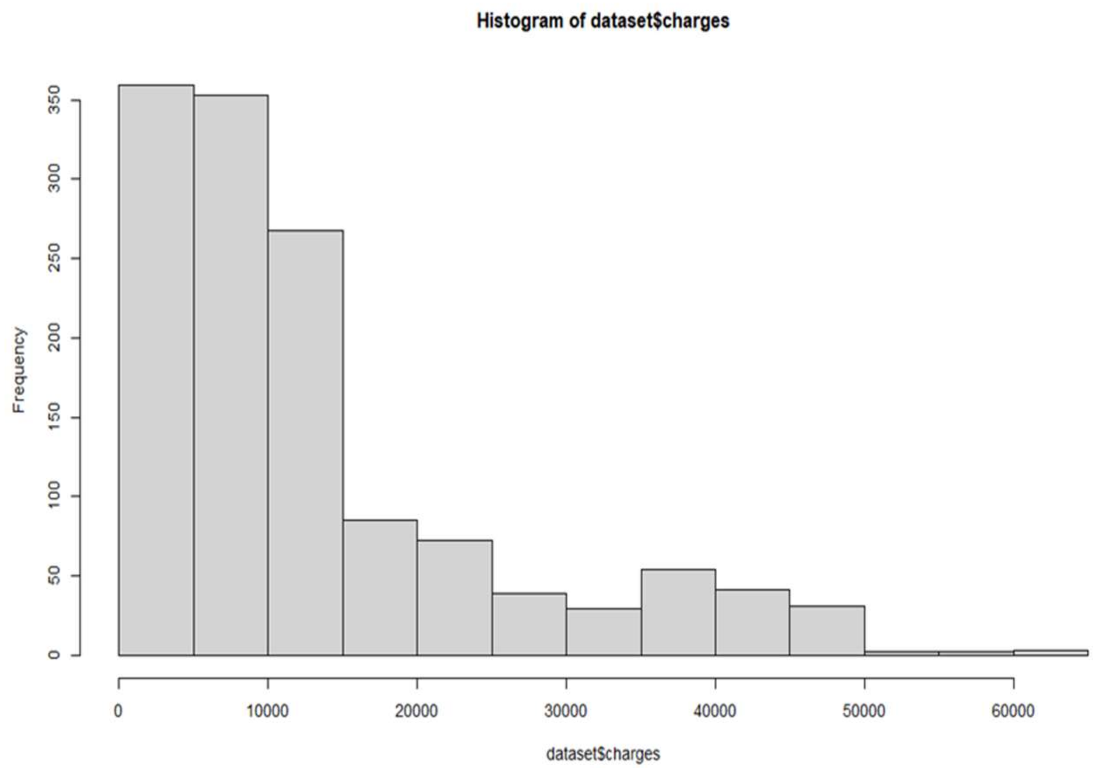
NO MISSING VALUES WERE FOUND

TRANSFORMED CHARACTER VARIABLES AS FACTORS AND CREATED NEW COLUMN TO CATEGORIZE BMI FOR BETTER ANALYSIS

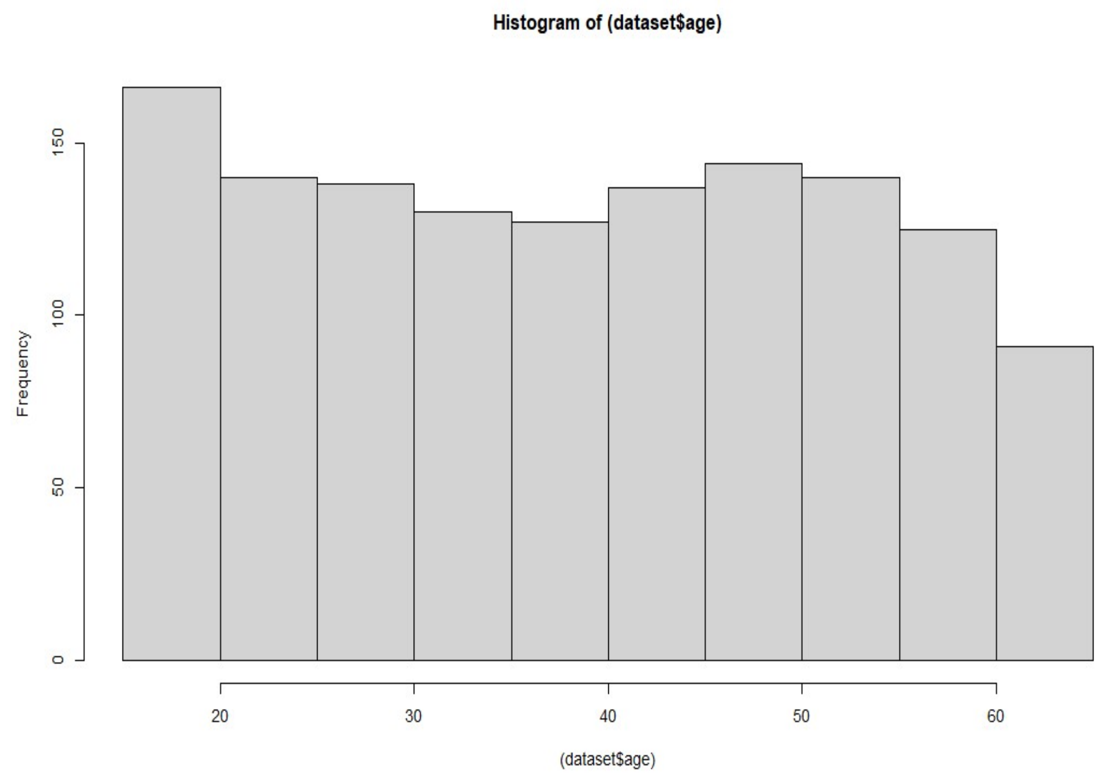
DATA ANALYSIS

- Scatter plots , co-relations plots
  are frequently used to identify any relationships between
  data since they make it simple to see any correlation.

Histogram ~ Charges



Histogram ~ Age



Distribution ~ Region wise Percentage

- **Gender is equally distributed**
- **20% of individuals are smokers**
- **54% of individuals have 1-3 children**

	region	count	percent
	<fct>	<int>	<dbl>
1	northeast	324	24
2	northwest	325	24
3	southeast	364	27
4	southwest	325	24

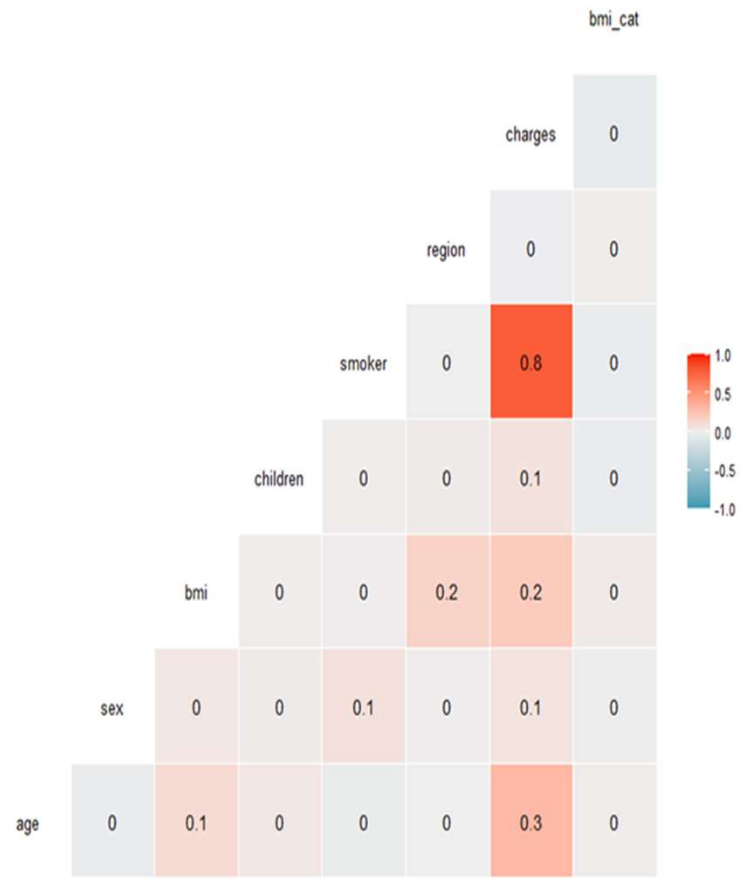
	smoker	count	percent
	<fct>	<int>	<dbl>
1	no	1064	80
2	yes	274	20

	sex	count	percent
	<fct>	<int>	<dbl>
1	female	662	49
2	male	676	51

	children	count	percent
	<int>	<int>	<dbl>
1	0	574	43
2	1	324	24
3	2	240	18
4	3	157	12
5	4	25	2
6	5	18	1

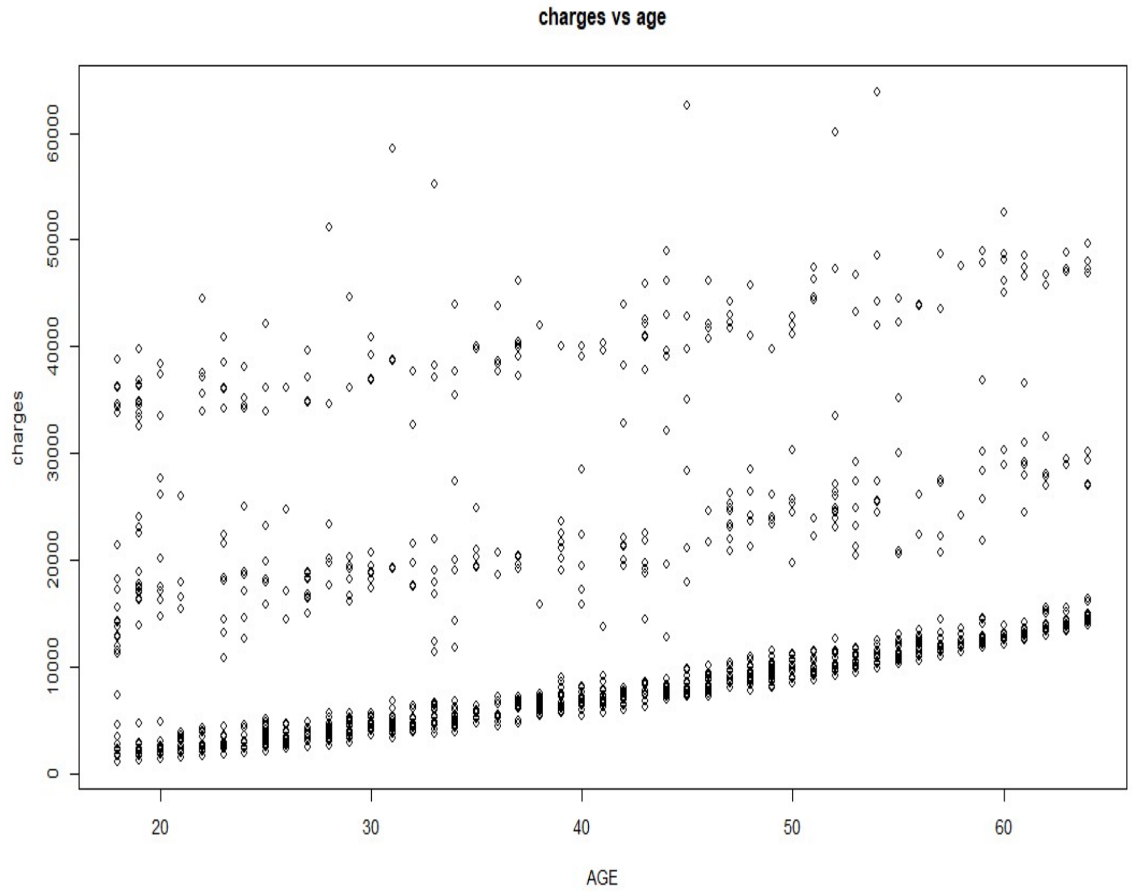
Correlation Charge vs Age vs BMI

As can be seen from the correlation plot, smoker, age and bmi are positively related to charges. So, we are going to analyse their relationship further.



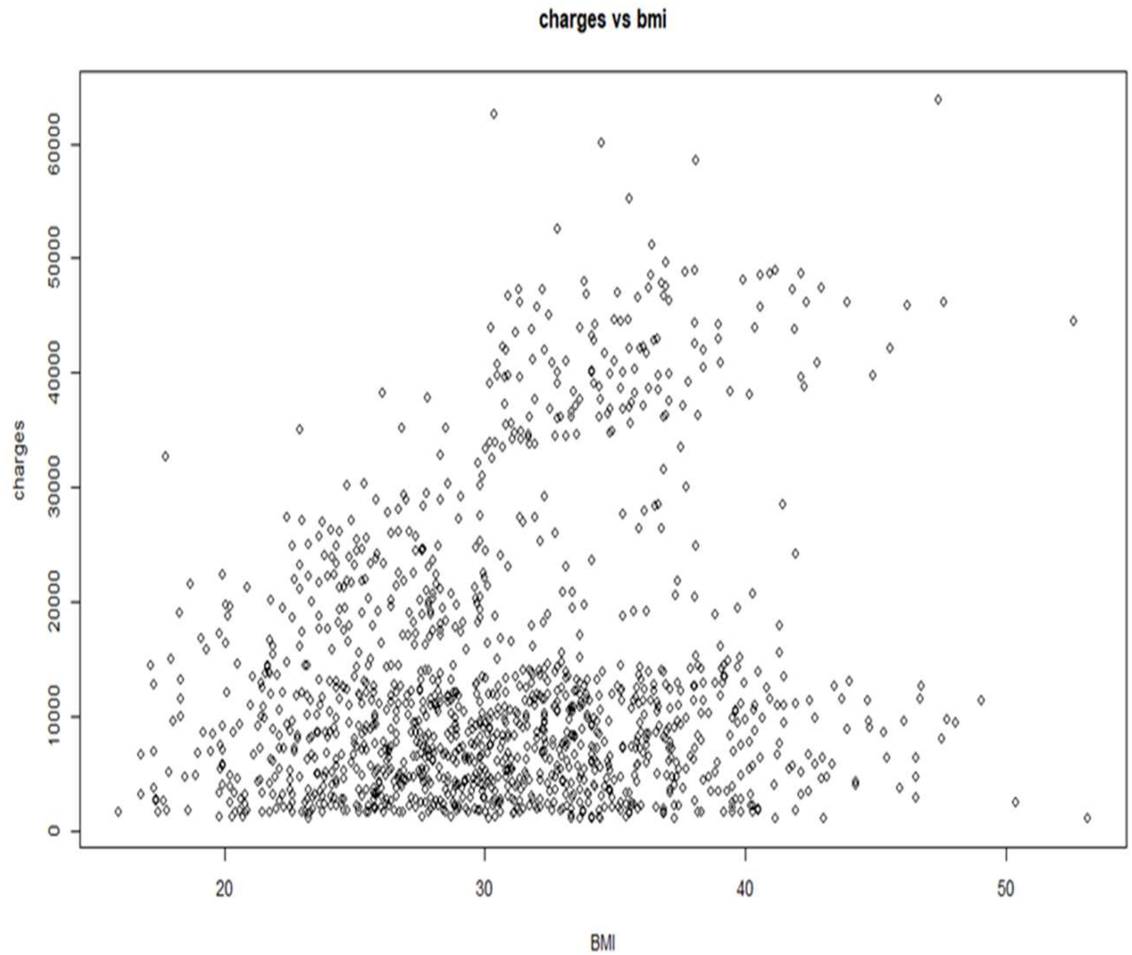
Scatter Plot Charge vs Age

- **According to the graph, charges slightly increases with increase in age**



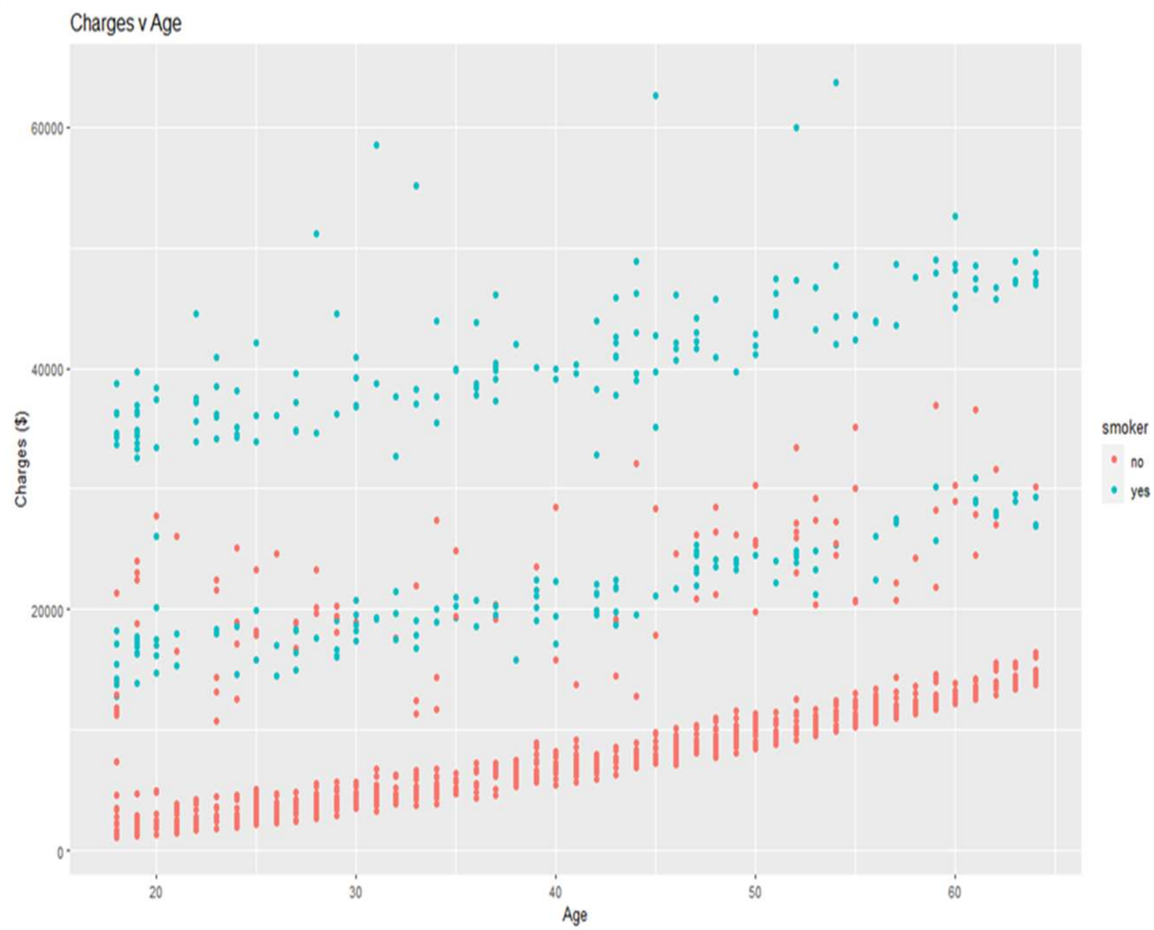
Scatter Plot Charge vs BMI

- According to the graph, BMI the relationship gets stronger after 30.



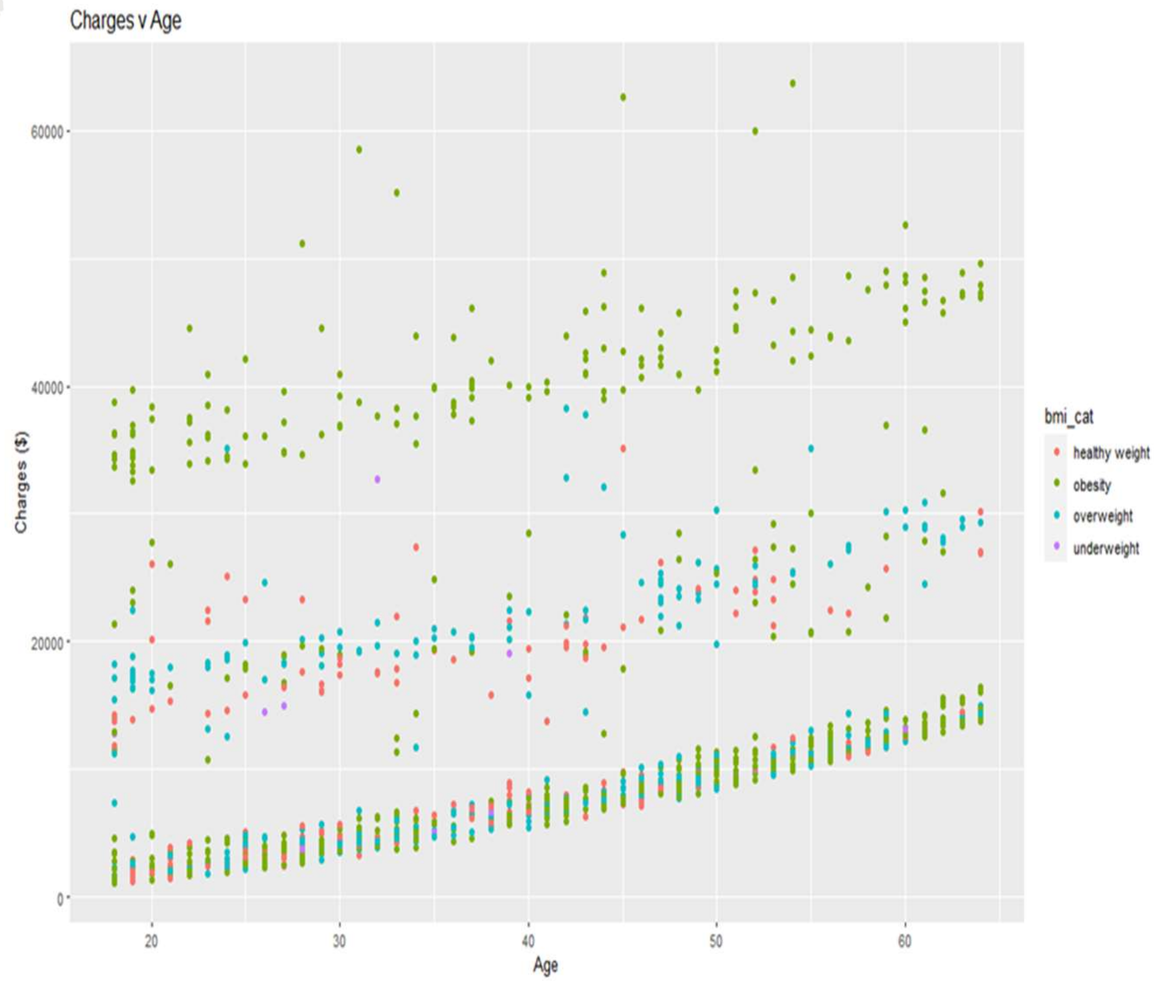
Scatter Plot Charge vs Age vs Smoker

It is clear from the plot that the charges rise in price with age. Additionally, it is evident that smokers paid more in fees than non-smokers did.



Scatter Plot Charge vs Age vs BMI

It is clear from the plot that the charges rise in price with age and obesity (higher bmi)



Regression Analysis:

- **The goal of linear regression is to mathematically represent a continuous variable Y as a function of one or more variables X, allowing us to use this regression model to predict the Y when only the X is known.**
- **Initial Linear regression model.**
  - ***Outcome:* Charges**
  - ***Predictors:* Age, Sex, BMI, Smokers**

Linear Regression Analysis:

Here, we attempt to build a model using one significant predictor at a time. Based on the correlation plot, we can see that age is the second important predictor, as the most significant predictor, smoking, is a categorical variable.

```
Call:
lm(formula = charges ~ age, data = training_set)

Residuals:
   Min     1Q Median     3Q    Max 
 -8106  -6704  -5933   5775  47338 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3381.74    1079.16   3.134  0.00178 ** 
age           253.27      25.79   9.821  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11420 on 1002 degrees of freedom
Multiple R-squared:  0.08781,	Adjusted R-squared:  0.0869 
F-statistic: 96.46 on 1 and 1002 DF,  p-value: < 2.2e-16

> |
```

We can see from the Pr(>|t|) column that the regression coefficient (2.2e-16) is significant from zero (p0.001) and that there should be an increase in charges of 253.27 for every year that a person gets older. The model only accounts for 8.78 percent of the variance in charges, which is quite low and suggests that it is not a very good model. The squared correlation between the actual and anticipated value is also known as the multiple R-squared. Because of how large the residual standard error (11420) is, it can be regarded as the typical error in estimating charges from age using this model.

Regression Analysis:

```
+           data = training_set)
> summary(rg)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-11484.8  -3615.8   -234.1   1625.2  25646.6 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -6147.85    1641.88  -3.744 0.000191 ***
age                  258.90      13.70  18.900  < 2e-16 ***
sexmale             -504.27     382.39  -1.319 0.187565    
bmi                   79.69      64.12   1.243 0.214248    
children             481.83     157.03   3.068 0.002210 ** 
smokeryes          23414.39     471.98  49.609  < 2e-16 ***
regionnorthwest     -567.62     547.94  -1.036 0.300499    
regionsoutheast    -1155.03     549.69  -2.101 0.035872 *  
regionsouthwest    -1086.28     553.28  -1.963 0.049886 *  
bmi_catobesity      4114.28     951.76   4.323  1.7e-05 ***
bmi_catoverweight    691.76     669.98   1.033 0.302086    
bmi_catunderweight   363.15    1656.70   0.219 0.826541    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6030 on 992 degrees of freedom
Multiple R-squared:  0.7485,	Adjusted R-squared:  0.7457 
F-statistic: 268.3 on 11 and 992 DF,  p-value: < 2.2e-16
```

Final Model:

- **Final model can be written as follow:**
- **charges= -12102.77 + 257.85(age) + 321.85(bmi) + 23811.40(smoker-yes) +473.50(children)**

```
> summary(model_3)

Call:
lm(formula = dataset$charges ~ dataset$age + dataset$bmi + dataset$smoker +
    dataset$children, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max
-11897.9  -2920.8   -986.6   1392.2  29509.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -12102.77     941.98 -12.848  < 2e-16 ***
dataset$age         257.85      11.90  21.675  < 2e-16 ***
dataset$bmi         321.85      27.38  11.756  < 2e-16 ***
dataset$smokeryes 23811.40     411.22  57.904  < 2e-16 ***
dataset$children    473.50     137.79   3.436 0.000608 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

- Using R, we fitted a linear regression model to the dataset and used the model to predict the values under various situations.
- With an Adjusted R-Squared value of 0.7489, the constructed linear regression model can account for 74.89 percent of the variance in the target variable (insurance charges).
- As a result, expanding the data collection may be necessary because there may not be enough data or predictors to fully explain the target variables.
- The most important factors in determining insurance charges, according to both models, are age, bmi, and smoker status. As a result, the beneficiary's insurance premiums will be higher if they smoke frequently, are older, have a higher BMI, or are obese.