**A Linear Regression model to Predict medical charges**
**Project Report**
**IE7280**

**Shruthi Machimada**
**001880621**

## Objective:

Predicting the personalized health care costs for a user, based on based on factors such as Age, gender, BMI, number of children, smoking habits.

Insurance companies can use this to give suitable premiums to customers, based on their profile.

## Data:

The data can be found at : https://www.kaggle.com/mirichoi0218/insurance

**Input variables:**
Age
Sex
BMI
Children
Smoker
Region
**Outcome variable:**
Charges

## Exploring the data

Viewing the data types of each column, and the number of observations.

```
> glimpse(data)
Observations: 1,338
Variables: 7
$ age      <int> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 19, 52, 23, 56, 30, 60, 30, 18, 34, 37,
$ sex      <fct> female, male, male, male, male, female, female, female, male, female, male, female, male, female, m
$ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.440, 27.740, 29.830, 25.840, 26.220, 26.290, 34.
$ children <int> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 2, 3, 0, 2, 1, 2, 0, 0, 5,
$ smoker   <fct> yes, no, no, no, no, no, no, no, no, no, no, yes, no, no, yes, no, no, no, no, yes, no, no, no, yes
$ region   <fct> southwest, southeast, southeast, northwest, northwest, southeast, southeast, northwest, northeast,
$ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855, 3756.622, 8240.590, 7281.506, 6406.411, 28923.1
>
```

There are 7 variables in total.

The outcome variable is charges, which is a decimal number indicating the amount of medical charges a person incurs.

The input variables are:
Age and number of children are integer values.

**Distribution of data:**

```
> summary(data)
      age            sex            bmi           children       smoker         region        charges     
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740  
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382  
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270  
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640  
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770  
> 
```

The customers' Gender and Region are evenly distributed. There are 5 times more smokers than non-smokers and customers' Age ranges from 18 to 64 years.
The average charge is 13270, with a minimum cost of 1122 and a maximum cost of 63770.

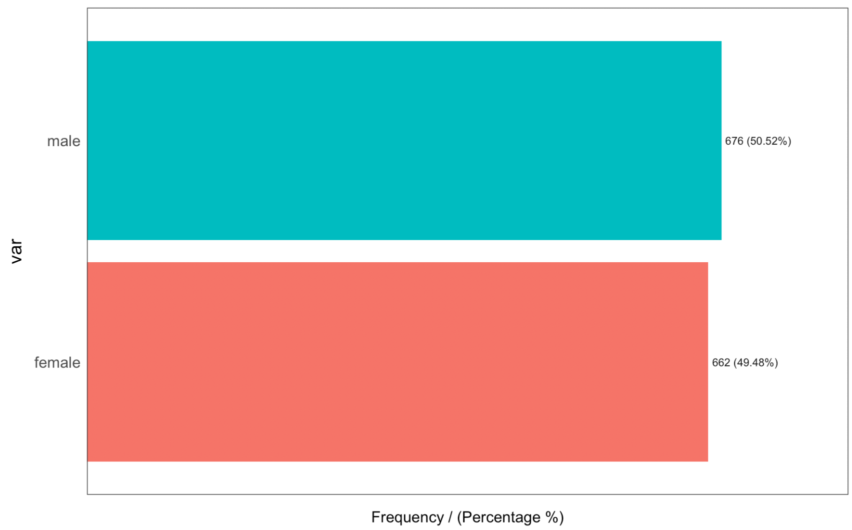
I then checked the quantity and percentage of zeros, NAs and infinite values, to handle the missing values.

```
> df_status(data)
  variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
1      age       0     0.0    0    0     0     0 integer     47
2      sex       0     0.0    0    0     0     0  factor      2
3      bmi       0     0.0    0    0     0     0 numeric    548
4 children     574    42.9    0    0     0     0 integer      6
5   smoker       0     0.0    0    0     0     0  factor      2
6   region       0     0.0    0    0     0     0  factor      4
7  charges       0     0.0    0    0     0     0 numeric   1337
>
```

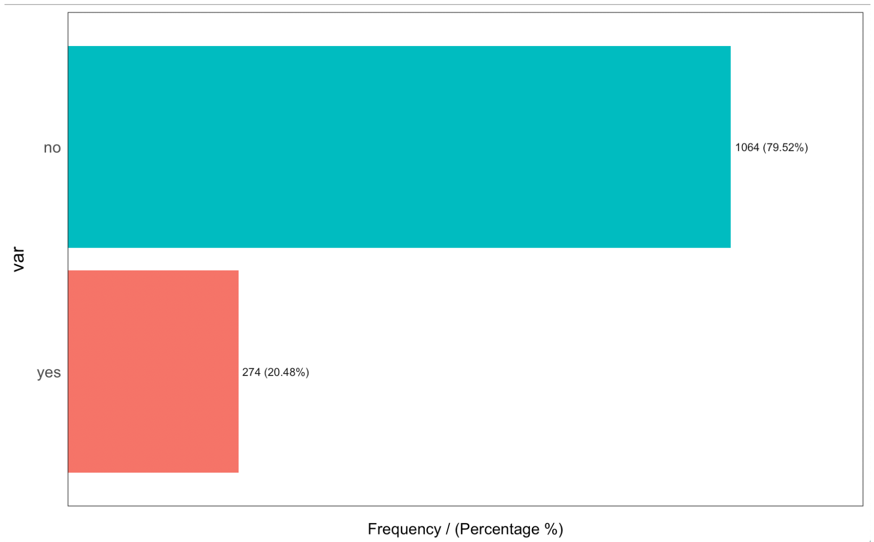
There are no missing values or NAs, so we do not need to clean this data.

## Distribution of Categorical variables

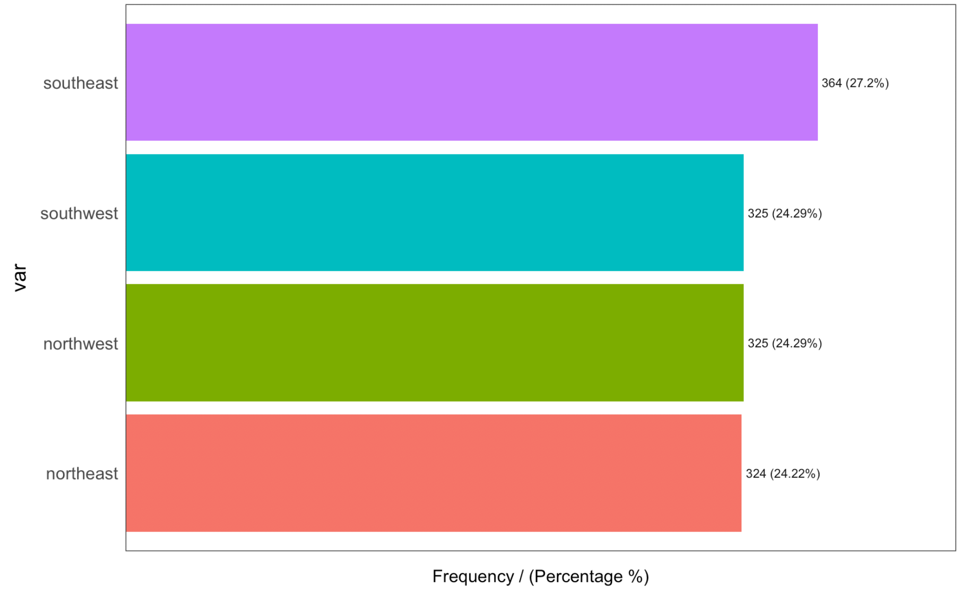
### Gender




## Smoker

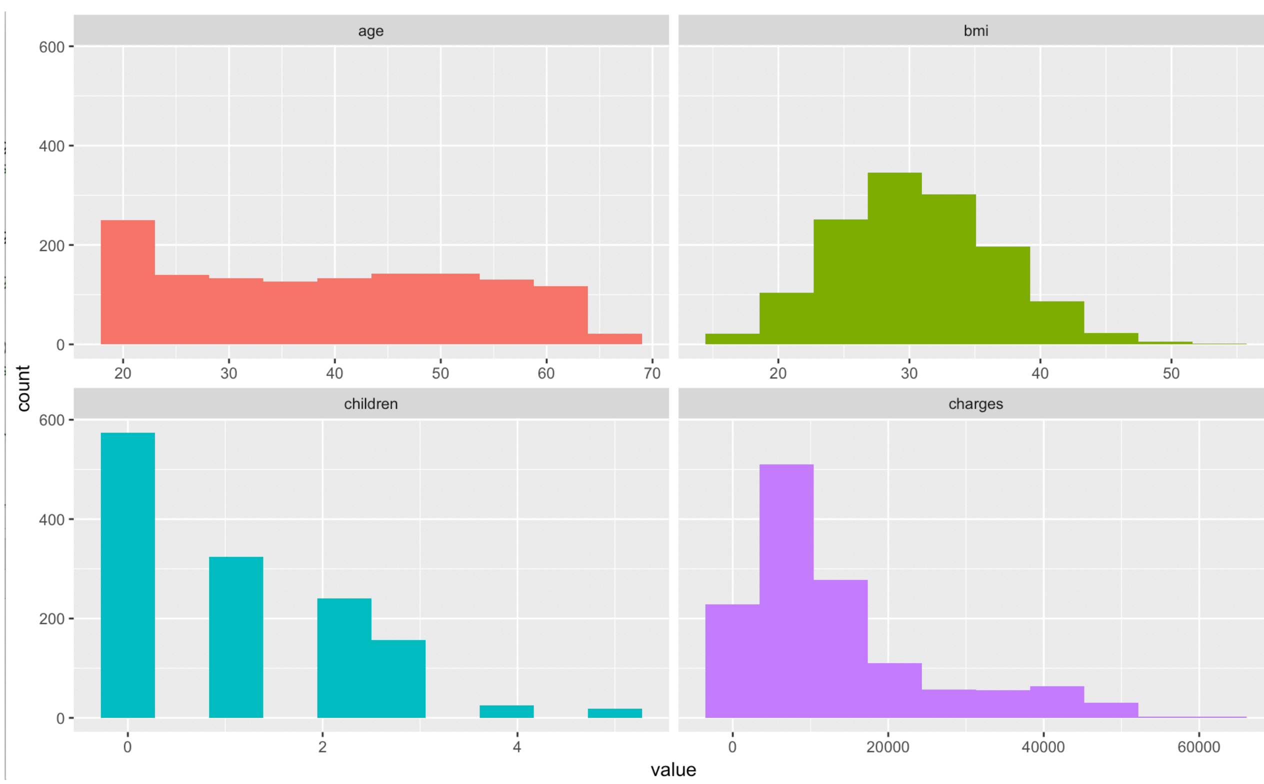



## Region

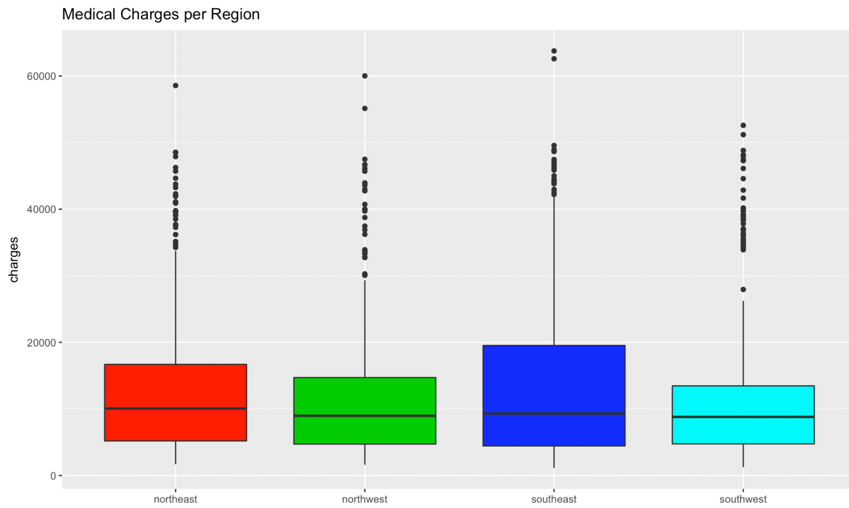


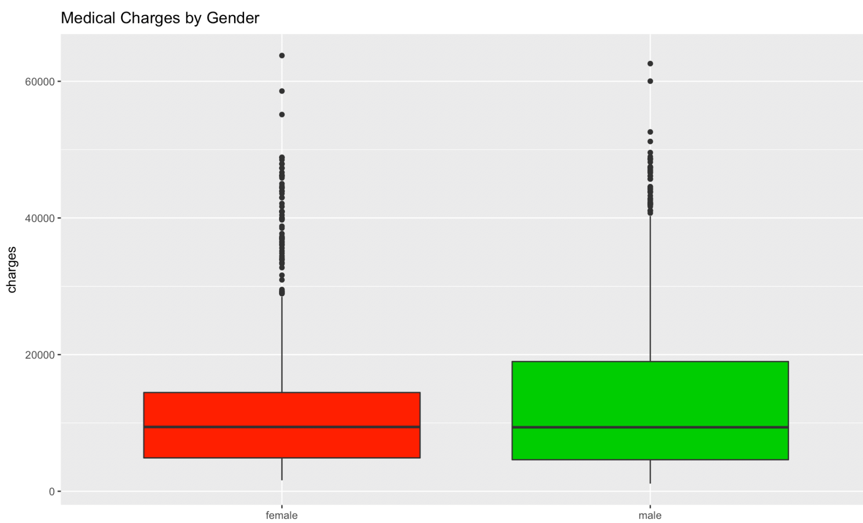

Sex and Region are evenly distributed, but the Smoker variable is distributed in the ratio 80:2.

## Distribution of the Numeric variables

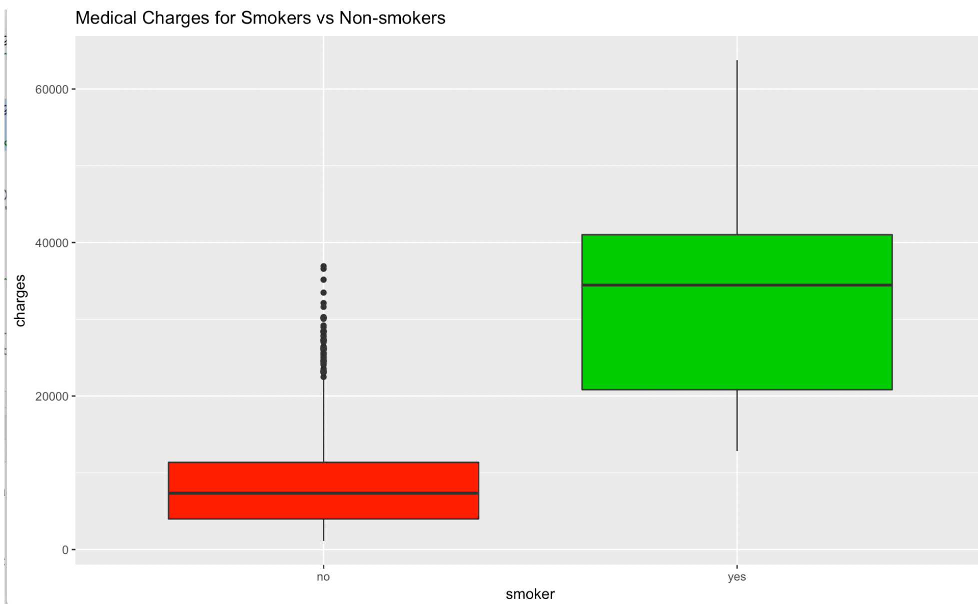



## Exploring relationships among variables



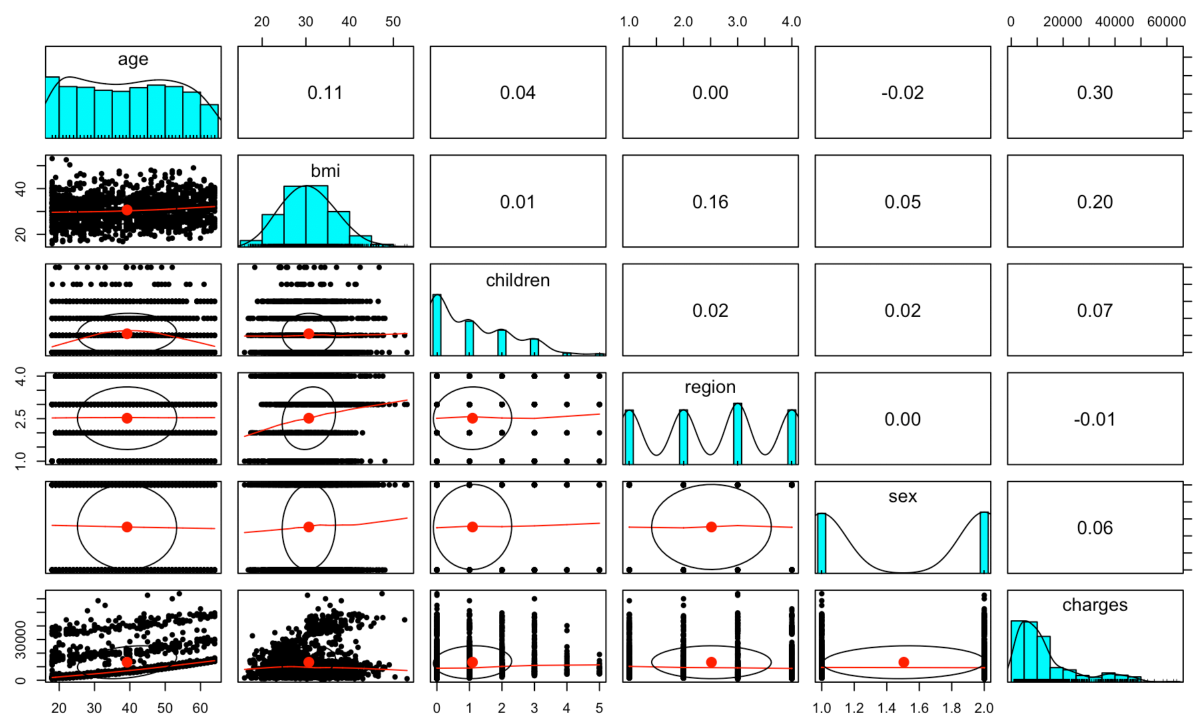




The charges are not affected by just the Region and Gender.




The charges for a Smoker is significantly higher than that of a non-smoker.

## Correlation




Age is mildly correlated to charges with a correlation coefficient of 0.3. All the other variables have negligible correlation coefficients.
On observing the distribution of age vs charges, we see that there is no clear linear relation – there are 3 levels of charges, across the distribution of age.

## Splitting the dataset

I split the data into training and test sets. 75% of the data will be in the training set, which will be used to fit the model, the remaining 25% will be used to evaluate the model's performance.

## Linear Models

### Model 1 – Using all 6 input variables to predict the Charges

```
charges= -11650.48 + (248)age – (194.51)sex + (342)bmi +
(483.95)children + (24212)smoker -
(539.55)RegionNW – (1137.52)RegionSE – (1095.81)RegionSW
```

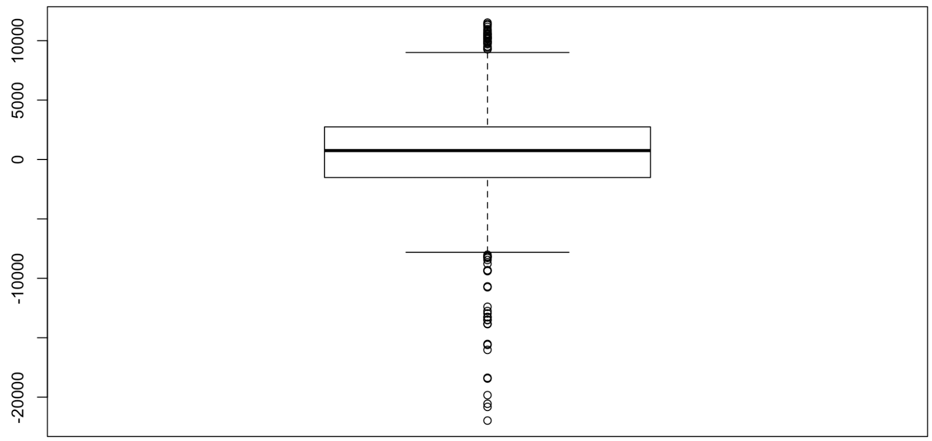
```
> linear_model6<-lm(charges~.,data=data_train)
> summary(linear_model6)

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

Residuals:
   Min     1Q Median     3Q    Max 
-11528  -2837  -1003   1445  29751 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11650.48    1169.62  -9.961  < 2e-16 ***
age                248.94      13.84  17.986  < 2e-16 ***
sexmale           -194.51     386.67  -0.503  0.61505    
bmi                342.89      33.38  10.273  < 2e-16 ***
children           483.95     159.22   3.040  0.00243 ** 
smokeryes        24212.35     485.21  49.900  < 2e-16 ***
regionnorthwest   -539.55     555.08  -0.972  0.33128    
regionsoutheast  -1137.52     562.58  -2.022  0.04345 *  
regionsouthwest  -1095.81     556.71  -1.968  0.04930 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6094 on 994 degrees of freedom
Multiple R-squared:  0.7504,	Adjusted R-squared:  0.7484 
F-statistic: 373.6 on 8 and 994 DF,  p-value: < 2.2e-16
```

### Use the model to Predict values in the Test dataset:

```
pred_6var <- predict(linear_model6, data_test)
pred_6var

#Evaluate the Model
residual<-pred_6var - data_test$charges
plot(residual)
boxplot(residual)
```

**Evaluating Model Performance**

**Residuals:**
A box plot of the residuals shows that the residuals are mostly concentrated around 0.




**R squared and Adjusted R squared:**

```
> r2 <- rSquared(pred_6var, resid = residual)
> r2
          [,1]
[1,] 0.6938541
>
> n<-nrow(data_test)
> n
[1] 335
> adj_r2<-r2 * ( (n- 1) / (n - 5))
> adj_r2
          [,1]
[1,] 0.7022644
>
```

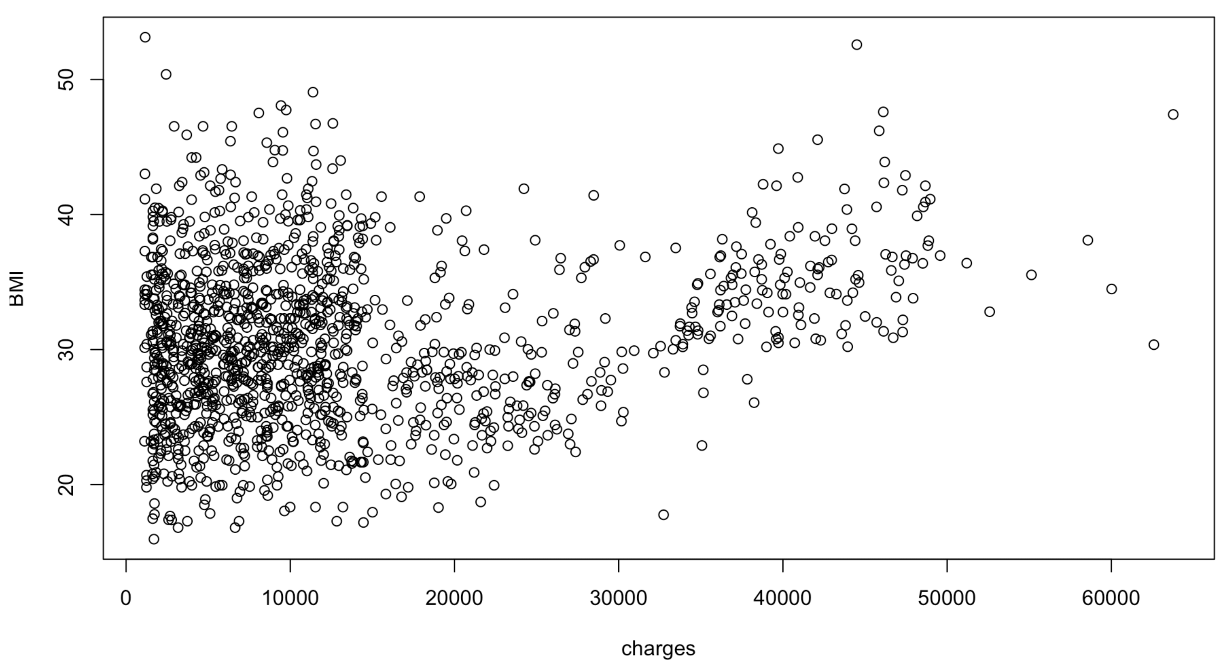
**Model 2**
Since the p value for Sex was 0.615, which is greater than 0.05, we Fail to Reject the null. The coefficient for Sex =0, so for the next model I deleted the variable.

```
charges= -11724.09 + (249.09)age + (341.87)bmi + (482.92)children +
(24196.26)smoker -
(532.59)RegionNW – (1127.50)RegionSE – (1087.85)RegionSW
```

```
> linear_model5<-lm(charges~age+bmi+children+smoker+region,data=data_train)
> summary(linear_model5)

Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max
-11627.3  -2804.3   -990.5   1470.6  29659.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11724.09    1159.99 -10.107  < 2e-16 ***
age                249.09      13.83  18.008  < 2e-16 ***
bmi                341.87      33.30  10.265  < 2e-16 ***
children           482.92     159.14   3.034  0.00247 **
smokeryes        24196.26     483.98  49.995  < 2e-16 ***
regionnorthwest   -532.59     554.70  -0.960  0.33722
regionsoutheast  -1127.50     562.02  -2.006  0.04511 *
regionsouthwest  -1087.85     556.27  -1.956  0.05079 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6092 on 995 degrees of freedom
Multiple R-squared:  0.7504,	Adjusted R-squared:  0.7486
F-statistic: 427.3 on 7 and 995 DF,  p-value: < 2.2e-16
```

**Model 3**

Model2 did not perform better than Model1, so I decided to include the Gender variable. Since BMI there was no clear relationship between BMI and charges, I fit the next model without BMI.



charges= -2261.46 + (261.76)age + (44.68)sex + (484.32)children + (24214.21)smoker - (423.91)RegionNW + (412.94)RegionSE – (624.57)RegionSW

```
> linear_modelBMI<-lm(charges~age+sex+children+smoker+region,data=data_train)
> summary(linear_modelBMI)

Call:
lm(formula = charges ~ age + sex + children + smoker + region, 
    data = data_train)

Residuals:
   Min     1Q Median     3Q    Max 
-16186  -1937  -1270   -288  28403 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2261.46     767.27  -2.947  0.00328 ** 
age               261.76      14.49  18.064  < 2e-16 ***
sexmale            44.68     405.74   0.110  0.91234    
children          484.32     167.37   2.894  0.00389 ** 
smokeryes       24214.21     510.07  47.473  < 2e-16 ***
regionnorthwest  -413.91     583.37  -0.710  0.47817    
regionsoutheast   412.94     569.72   0.725  0.46873    
regionsouthwest  -624.57     583.23  -1.071  0.28449    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6406 on 995 degrees of freedom
Multiple R-squared:  0.7239,	Adjusted R-squared:  0.722 
F-statistic: 372.7 on 7 and 995 DF,  p-value: < 2.2e-16
```

### Model 4 – No Gender and BMI variables

charges= -2238.11 + (261.73)age + (484.55)children + (24217.92)smoker - (415.43)RegionNW + (411.69)RegionSE – (626.08)RegionSW

```
> lm_noBMIGender<-lm(charges~age+children+smoker+region,data=data_train)
> summary(lm_noBMIGender)

Call:
lm(formula = charges ~ age + children + smoker + region, data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-16165.9  -1914.6  -1275.1   -303.4  28423.4 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2238.11     737.01  -3.037  0.00245 ** 
age               261.73      14.48  18.074  < 2e-16 ***
children          484.55     167.28   2.897  0.00385 ** 
smokeryes       24217.92     508.70  47.607  < 2e-16 ***
regionnorthwest  -415.43     582.92  -0.713  0.47621    
regionsoutheast   411.69     569.32   0.723  0.46978    
regionsouthwest  -626.08     582.78  -1.074  0.28295    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6403 on 996 degrees of freedom
Multiple R-squared:  0.7239,	Adjusted R-squared:  0.7223 
F-statistic: 435.3 on 6 and 996 DF,  p-value: < 2.2e-16
```

### Model 5 – No region

Since the **p value** for the region variables are all greater than 0.05, their coefficients are 0, and we can delete them.

charges= -11874.48 + (249.95)age – (162.68)sex + (325.44)bmi + (486.14)children + (24179.12)smoker

```
> lm_noRegion<-lm(charges~age+sex+bmi+children+smoker,data=data_train)
> summary(lm_noRegion)

Call:
lm(formula = charges ~ age + sex + bmi + children + smoker, data = data_train)

Residuals:
   Min     1Q Median     3Q    Max 
-12100  -2855  -1028   1437  29323 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11874.48    1123.06 -10.573  < 2e-16 ***
age            249.95      13.85  18.047  < 2e-16 ***
sexmale       -162.68     386.87  -0.421  0.67421    
bmi            325.44      31.88  10.208  < 2e-16 ***
children       486.14     159.30   3.052  0.00234 ** 
smokeryes    24179.12     482.10  50.153  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6102 on 997 degrees of freedom
Multiple R-squared:  0.7491,	Adjusted R-squared:  0.7478 
F-statistic: 595.2 on 5 and 997 DF,  p-value: < 2.2e-16

> 
```

### Model 6 – Polynomial regression for Age

Because of the non-linear relationship between Age and Charges, I modeled a polynomial regression with degree 2, for Age.

$$charges = -6390.096 - (57.217)age + (3.873)age^2 + (339.121)BMI - (217.756)sex + (637.733)children + (24277.59)smoker - (610.087)regionNW - (1152.644)regionSE - (1092.289)regionSW$$

```
> summary(lm_polyAge)

Call:
lm(formula = charges ~ age + I(age^2) + sex + bmi + children +
    smoker + region, data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max
-12204.2  -2825.7   -952.4   1264.7  30511.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6390.096   1974.456  -3.236  0.00125 **
age               -57.217     93.844  -0.610  0.54220
I(age^2)            3.873      1.174   3.298  0.00101 **
sexmale          -217.756    384.825  -0.566  0.57162
bmi               339.121     33.232  10.205  < 2e-16 ***
children          637.733    165.151   3.862  0.00012 ***
smokeryes       24277.590    483.227  50.241  < 2e-16 ***
regionnorthwest  -610.087    552.756  -1.104  0.26998
regionsoutheast -1152.664    559.825  -2.059  0.03976 *
regionsouthwest -1092.289    553.963  -1.972  0.04891 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6064 on 993 degrees of freedom
Multiple R-squared:  0.7531,    Adjusted R-squared:  0.7509
F-statistic: 336.6 on 9 and 993 DF,  p-value: < 2.2e-16
```

### Model 7 - Polynomial regression for Age and No Gender

Since I saw an increase in performance using Model 6 , I decided to use a Polynomial regression for Age and proceed with deleting Gender since the p value for Gender was greater than 0.05.

$$charges = -6489.010 - (56.083)age + (3.861)age^2 + (338.001)BMI + (636.095)children + (24259.379)smoker - (602.075)regionNW - (1141.404)regionSE - (1083.394)regionSW$$

```
Call:
lm(formula = charges ~ age + I(age^2) + bmi + children + smoker +
    region, data = data_train)
                                                    I(x)
Residuals:
     Min       1Q   Median       3Q      Max
-12298.6  -2801.9   -935.3   1327.2  30407.3

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6489.010   1966.031  -3.301 0.000999 ***
age               -56.083     93.790  -0.598 0.550002
I(age^2)            3.861      1.174   3.289 0.001039 **
bmi               338.001     33.162  10.193  < 2e-16 ***
children          636.095    165.069   3.854 0.000124 ***
smokeryes       24259.379    481.989  50.332  < 2e-16 ***
regionnorthwest  -602.075    552.386  -1.090 0.275998
regionsoutheast -1141.404    559.280  -2.041 0.041530 *
regionsouthwest -1083.394    553.550  -1.957 0.050607 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 994 degrees of freedom
Multiple R-squared:  0.7531,	Adjusted R-squared:  0.7511
F-statistic: 378.9 on 8 and 994 DF,  p-value: < 2.2e-16
```

**Model 8 - Polynomial regression for Age and No Gender and No Region**

Model 7 resulted in the best performance until now, and we see that p value is high for Region, so for Model 8 I deleted Gender and region variables.

$$charges = -6719.071 - (55.64)age + (3.868)age^2 + (321.124)BMI + (638.98)children + (24229.52)smoker$$

```
> lm_polyAgeNoRegionSex<-lm(charges~age+I(age^2)+bmi+children+smoker,data=data_train)
> summary(lm_polyAgeNoRegionSex)

Call:
lm(formula = charges ~ age + I(age^2) + bmi + children + smoker, 
    data = data_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11762.9  -2859.1   -989.2   1373.0  30004.4 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6719.071   1932.137  -3.478 0.000528 ***
age           -55.641     93.801  -0.593 0.553197    
I(age^2)        3.868      1.174   3.295 0.001019 ** 
bmi           321.124     31.684  10.135  < 2e-16 ***
children      638.985    165.171   3.869 0.000117 ***
smokeryes   24229.519    478.917  50.592  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6069 on 997 degrees of freedom
Multiple R-squared:  0.7517,	Adjusted R-squared:  0.7505 
F-statistic: 603.7 on 5 and 997 DF,  p-value: < 2.2e-16

> 
```

## Comparing the performance of all Models

```
r_values<-data.frame(model=c("All variables","- Gender","- BMI","- Gender and BMI","- region","poly Age","polyAge - Ge
                     rSquaredValue=c(r2,r2_noGender,r2_noBMI,r2_noBMIGender,r2_noRegion,r2_polyAge,r2_polyAgeNoGender,
                     adjrSquared=c(adj_r2,adj_r2_noGender,adj_r2_noBMI,adj_r2_noBMIGender,adj_r2_noRegion,adj_r2_polyA
r_values
```

```
                        model rSquaredValue adjrSquared
1               All variables     0.6938541   0.7022644
2                    - Gender     0.6940446   0.7003351
3                       - BMI     0.6512488   0.6571513
4            - Gender and BMI     0.6512032   0.6551261
5                    - region     0.6919551   0.6982266
6                    poly Age     0.7035139   0.7120414
7           polyAge - Gender     0.7037308   0.7101090
8 polyAge - Region - Gender     0.7018981   0.7061264
> 
```

## Conclusion

Even though there is not a lot of difference in the $R^2$ and Adjusted $R^2$ values between the models, the models with polynomial regression for Age perform better , and we get the best R squared and Adjusted R squared for a Model with the following input variables-

- Polynomial Regression for Age
- BMI
- Number of children
- Smoking habits

**Final Model**

$$\text{charges} = -6719.071 - (55.64)\text{age} + (3.868)\text{age}^2 + (321.124)\text{BMI} + (638.98)\text{children} + (24229.52)\text{smoker}$$

$R^2$ = 0.701 or 70%
Adjusted $R^2$ = 0.706 or 70.6%