Multiple Linear Regression Model on Medical Cost Personal Dataset in R Shiny

Introduction:

Insurance companies nowadays are very much curious in predicting the insurance costs considering the various investment factors involved. Due to the increasing demand in predicting the appropriate cost of insurance, building models on this type of dataset will be very much useful and will definitely help insurance companies to take a second opinion and help predict the cost. Moreover, various types of visualizations will also help to better explore the dataset and predict more accurate results. It will also help us to answer the buzz question "Can you accurate predict insurance costs?" and the answer is "Yes". The model prediction in Shiny will give the business user the insurance costs value for the parameter considered and will give the result.

Regarding Shiny:

Shiny is an open - source R package that provides an elegant and powerful UI framework for building web-based applications in R. In my application, I have used Shiny package to make predictions on the new data based on significant parameters and provide predicted insurance cost value as output for the business. I have also showed the summary of the dataset, the dataset table, diagnostic plots, cross-validation plots, and various types of correlation and ggplots such as Scatter plots, Box plots, Frequency plots and Actual/ Predicted plots.

Dataset: (As seen in Shiny)

I had used the Medical Cost Personal Dataset available on Kaggle. My dataset consists of 1338 rows of 7 variables and size of 55 KB. My variable columns include Charges which is the dependent variable or target / response variable and rest of the variables (Age, Sex, BMI, Children, Smoker, Region) as the independent variables or the predictor variables.

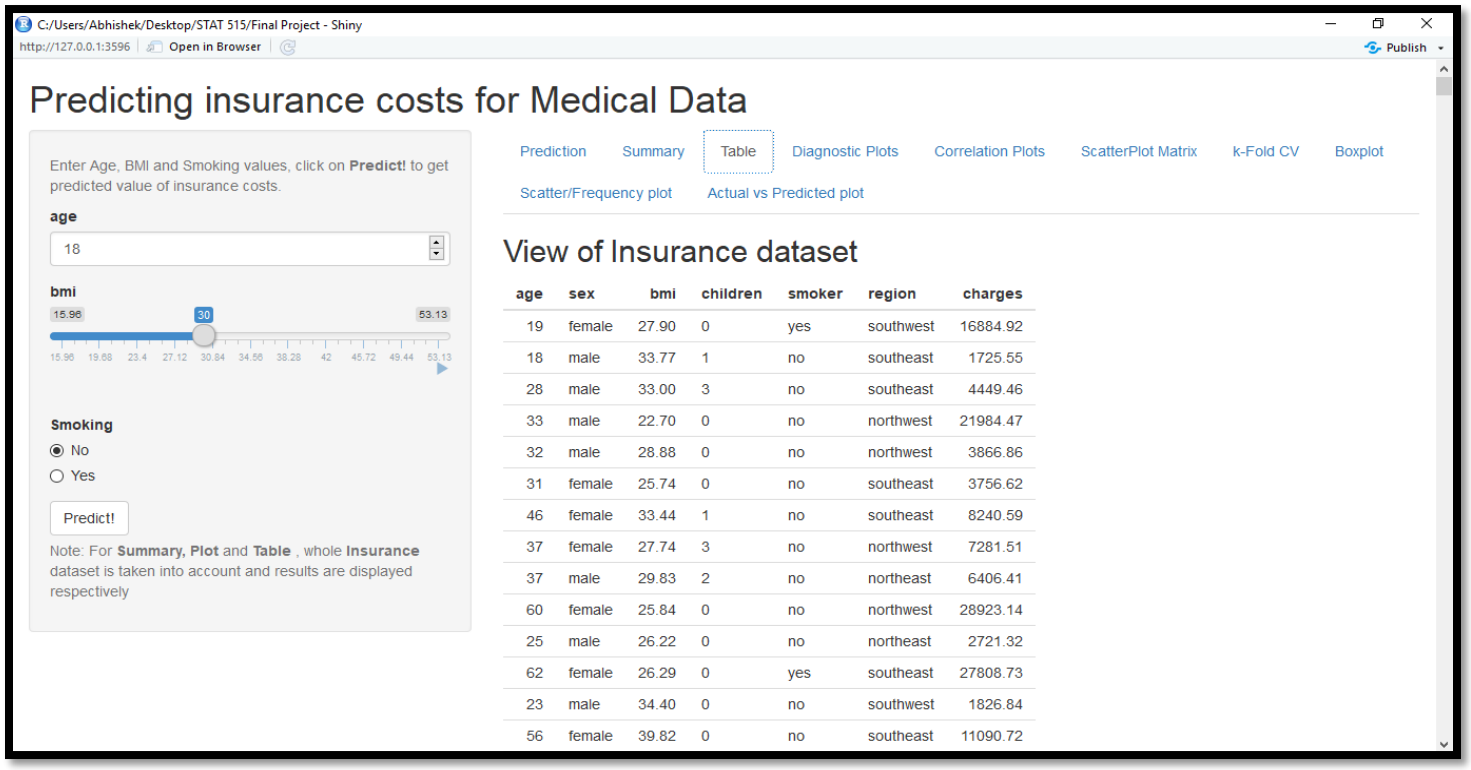

Fig. 1 – Insurance Data set as seen in RShiny Interface

Dataset Analysis:

Dataset analysis is the most important factor while studying the data and needs to be performed in order to understand the nature of the dataset at first glance. Having a look at the dataset using the Summary and the Structure command will give a more detail analysis on the data. The summary of the dataset will let us know the various variables available in the dataset and whether they are categorical or continuous variables. Also, it helps us know the distribution in the data and if there are any NA values present. For our dataset below, we can see that there are no NA values present and for the categorical columns such as Sex, Children, Smoker and Region, it gives us the count.

```
> summary(insurance)
      age            sex            bmi          children smoker           region       charges
 Min.   :18.0   female:662   Min.   :16.0   0:574    no :1064   northeast:324   Min.   : 1122
 1st Qu.:27.0   male  :676   1st Qu.:26.3   1:324    yes: 274   northwest:325   1st Qu.: 4740
 Median :39.0                Median :30.4   2:240               southeast:364   Median : 9382
 Mean   :39.2                Mean   :30.7   3:157               southwest:325   Mean   :13270
 3rd Qu.:51.0                3rd Qu.:34.7   4: 25                               3rd Qu.:16640
 Max.   :64.0                Max.   :53.1   5: 18                               Max.   :63770
>
```

Fig. 2 – Summary of Insurance Dataset

Structure as seen below gives us the class of the dataset. The number of rows and columns in the dataset, datatype and initial values for the columns present in the dataset.

```
> str(insurance)
'data.frame':   1338 obs. of  7 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: Factor w/ 6 levels "0","1","2","3",..: 1 2 4 1 1 1 2 4 3 1 ...
 $ 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 ...
>
```

Fig. 3 – Structure of the Insurance Dataset.

Dataset Exploration:

Firstly, as my dataset had lots of columns whose datatype is factor, I have converted the factor data to numeric for my analysis. Secondly, I have done splitting of the data into 70% train data and 30% test data for predicting on the dataset. I will be performing Multiple Linear Regression Analysis on the dataset to predict the cost. Thirdly, performed building model using train data based on significant variables and predicting the insurance cost on the test data and finally visualizing the data using various plots and graphs.

Visualization Analysis using Shiny:

1. Diagnostic Plots

The diagnostic plots help us to view the residuals in the dataset that will help us know if the model works well in the data. This visualization will not only will help us in building the model but also to improve it in a better way. The diagnostic plots help us to diagnose the data in 4 different ways.

1. Residual vs Fitted
   → This plot helps us to see if there are any non-linear pattern in the data. There can be a non-linear relationship between the outcome variables and the predictors and the pattern can be seen in the plot. If the data had equally spread residuals around a horizontal line without any pattern, then it will indicate us that we don't have non-linear relationships. Our plot below, since it's a red straight line is a good example.
2. Normal Q-Q
   →With this plot we can see if the residuals are normally distributed. In our plot, we can see that residuals follow the dotted plot for a large amount of data, but its tails deviate from the dotted line.
3. Scale-Location
   →This plot tells us about the spread of the data. It can be seen from the below graph, that the data is fairly spread along the range of predictors. However, the red line bends at the left tail which because of data being not equally spread at that point.
4. Residuals vs Leverage
   →This plot helps us to see if there are any outliers or influential cases which might had not been seen during analysis. In our plot, we can hardly see the Cook's distance line since all the cases are pretty well inside and there are no influential cases.

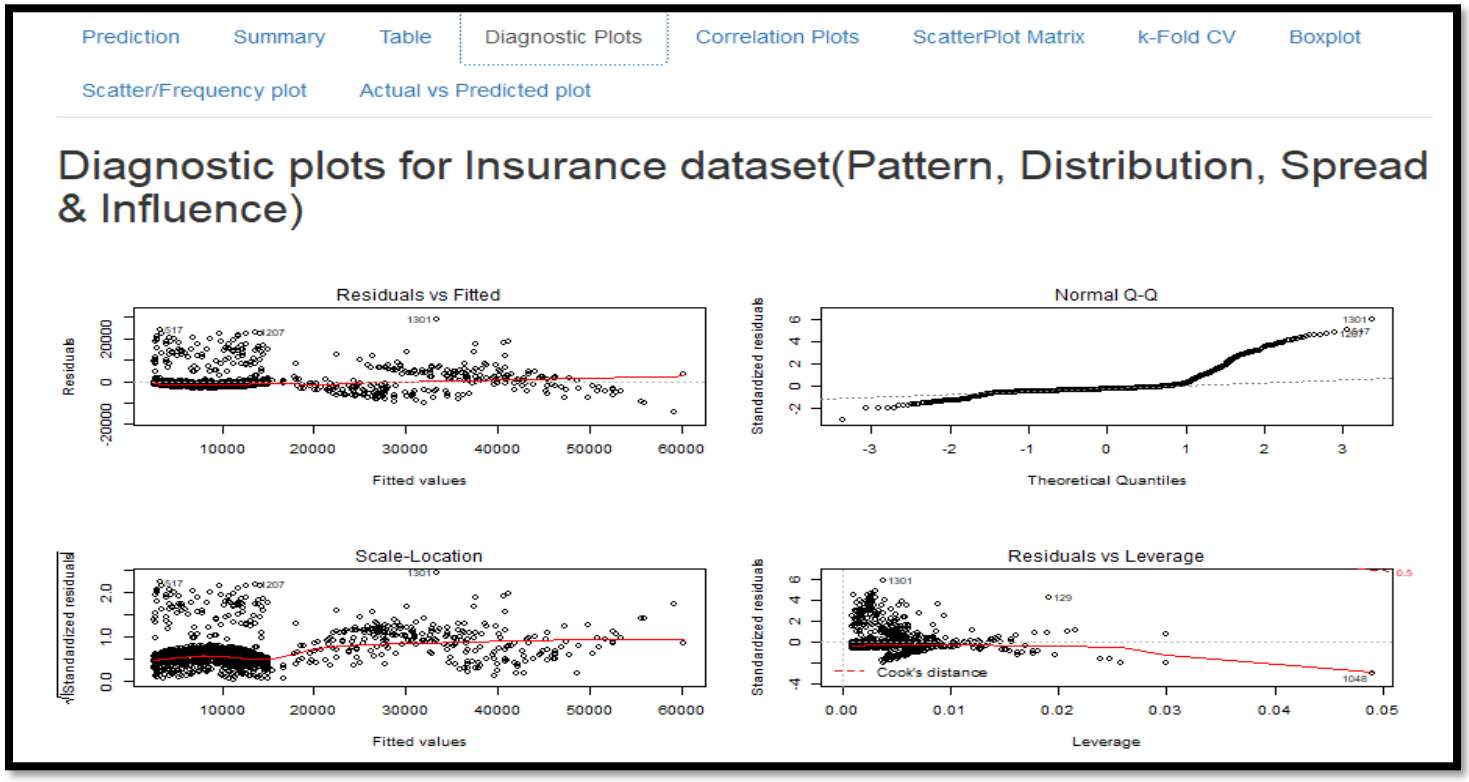

Fig. 4 – Diagnostic Plots

2. **Correlation Plots**

Correlation plots helps us to visualize relationship of one variable with all the other variables in the data. The higher the relationship between the variables, the higher is the correlation. By analysing the plot below, we can see that there are few parameters which are correlated and might show a stronger relationship when building the model.

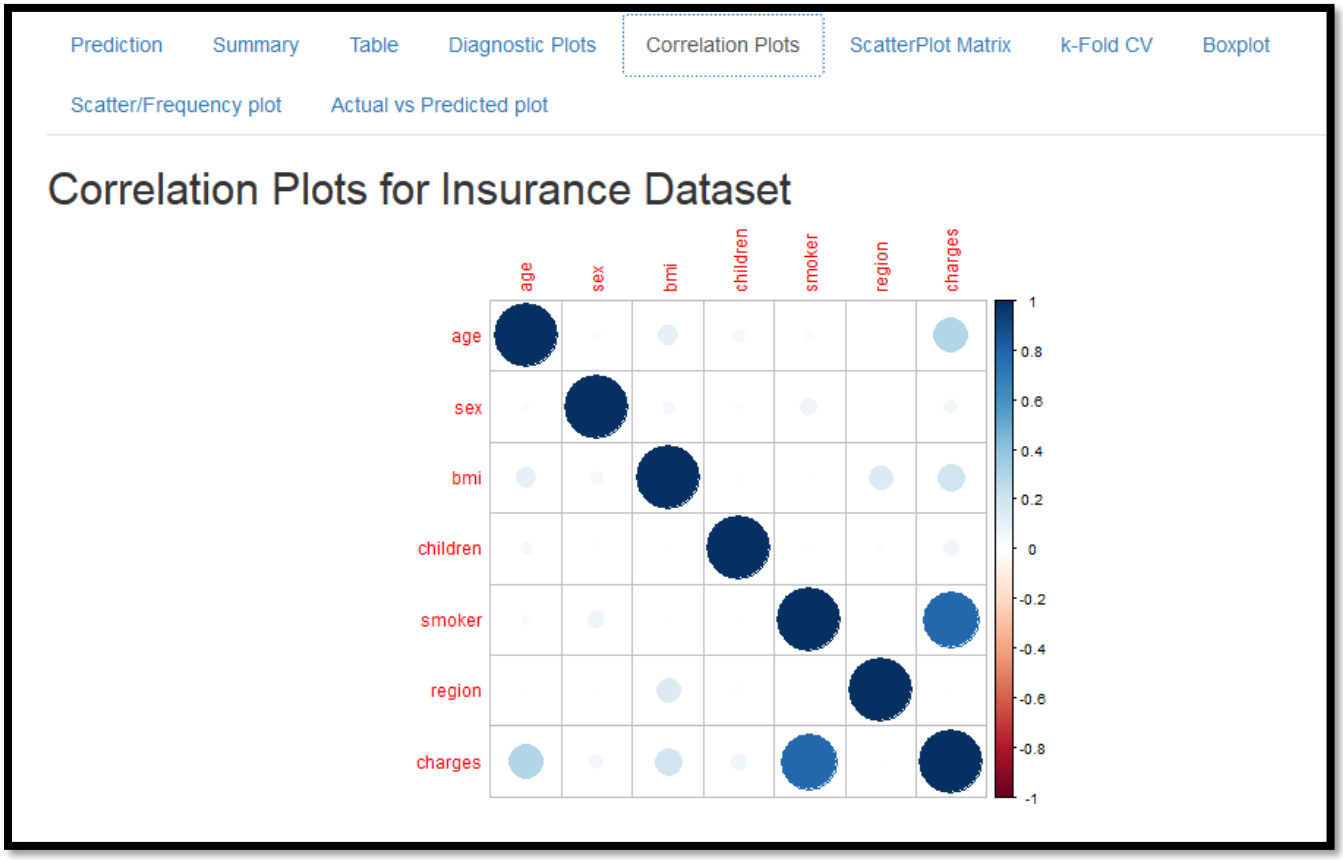

Fig 5- Correlation Plot

3. **Scatterplots and Frequency Plots**

→Scatterplots helps us to look out for relationships and frequency plots help us to have a count of the number of times the value is repeated. In our dataset, for analysis, I tried to visualize few scatterplots and frequency plots. In the first plot, we can see that as the age increases, the insurance charge too shows an increase. So, there is an upward trend in the best fit line. In the second plot, we can see that for BMI<30, the charges are somewhat on the lower side. However, for BMI>30, the charges vary a lot. In the third plot, we can see a scatterplot between number of children and charges. It can be analysed that lower number of children are more likely to incur charges whereas high number of children are less likely to incur. The final plot tells us the distribution of charges and its frequency. It can be seen that most of the insurance cost charges are low and very few were high.

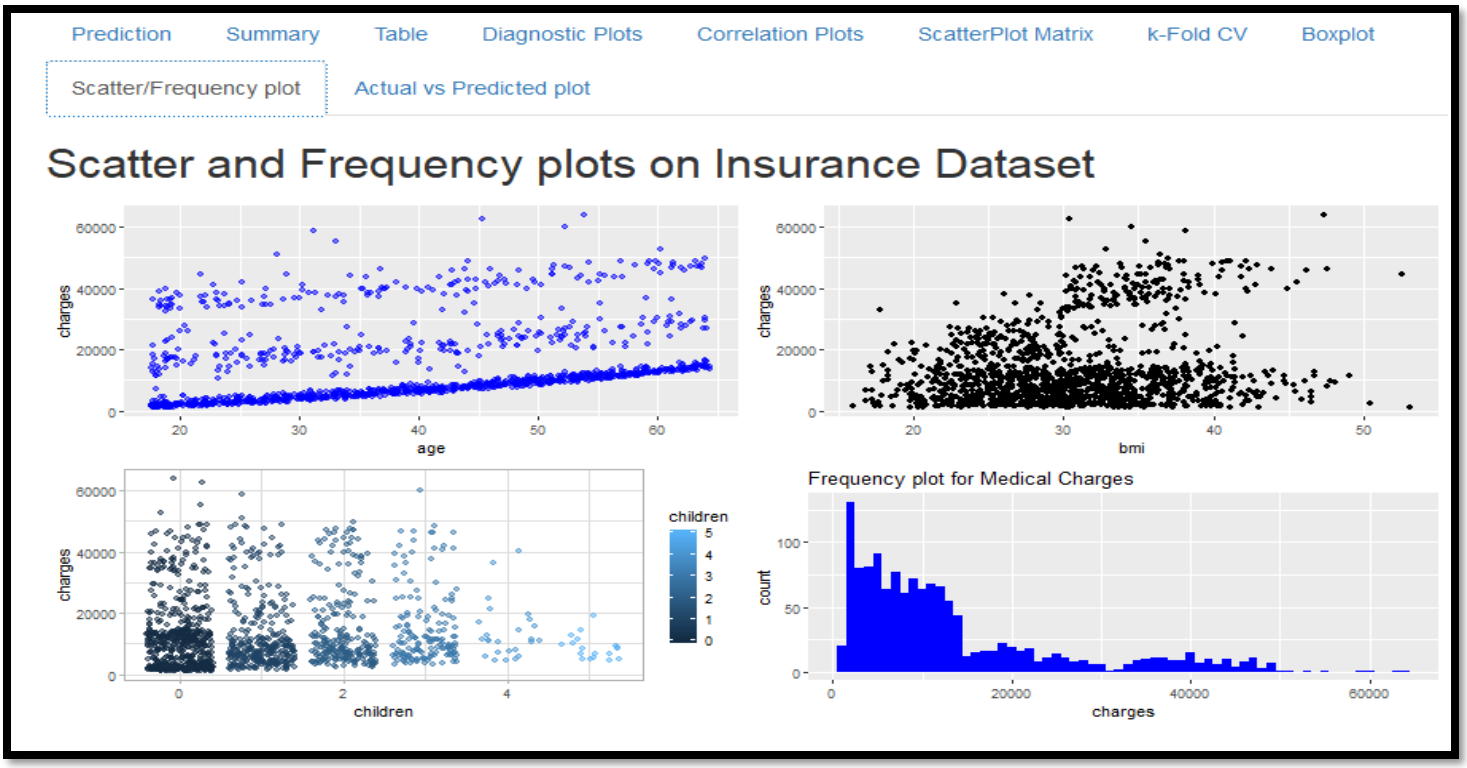

Fig 6 – Scatter and Frequency plot

4. **Boxplots**

→Boxplots helps us to visualize data and see if there are any outliers. Also, it helps us to see the distribution of the data across the various variable value. Even thought the first boxplot didn't showed any variation, we can see in the second boxplot that charges differ a lot for smoking equal to 'no' and smoking equal to 'yes'. In the third boxplot of charges vs children, we can see that number of children equal to five incur less charges. The fourth boxplot shows charges vs the region however there is not much of difference in the variable values.

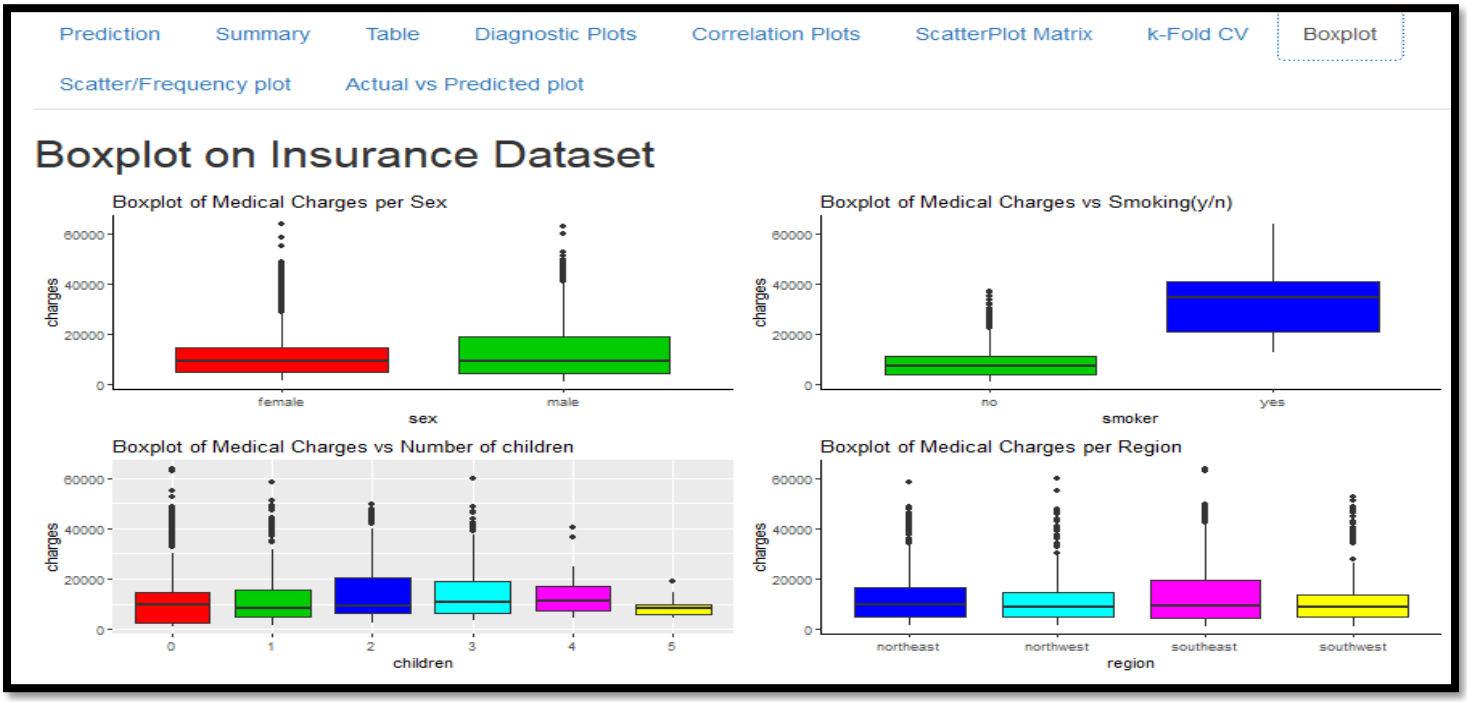

Fig 7 - Boxplots

Building Multiple Linear Regression Model:

The first step which involves in building the multiple linear regression model is to split the data into Train data and Test data. In my dataset, I have performed splitting keeping 70% as Train data and 30% as Test data. Then, I have performed a series of different combinations of Linear Regression Model to see which model gives the best accuracy and with least number of terms. A summary of the model which gives the highest adjusted R^2 value for my dataset can be found below. Here, for this model I have achieved a R square value of 0.836 which means that the model explains 83.6% of the variable predictions correctly. Moreover, since the p-value is very less, the model is highly significant and can be used for predicting values for a newer data.

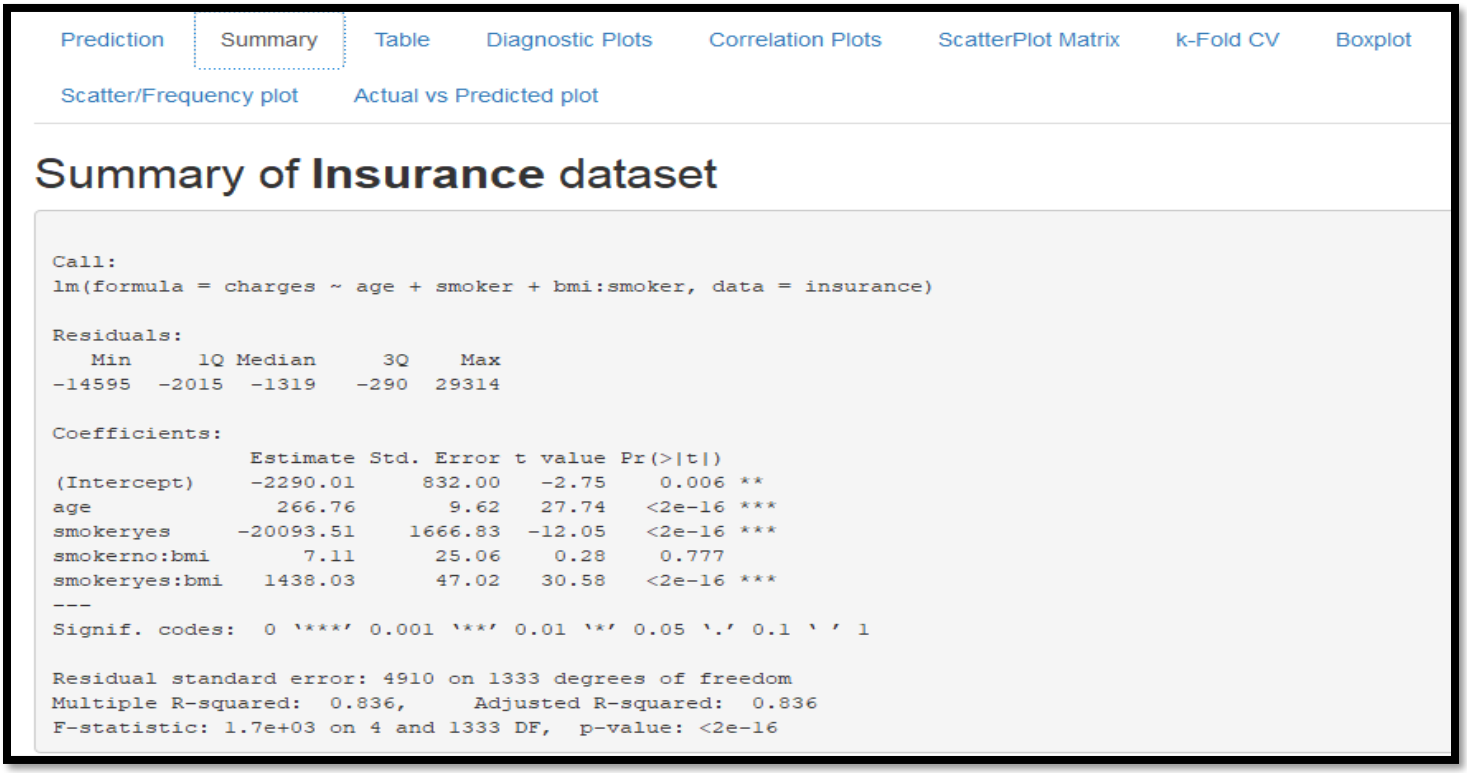

Fig 8- Summary of the Linear Regression Model

Making Predictions:

Here, I am using the model's significant variables as input to predict the insurance charges. This predicted Insurance costs charges will be somewhat similar to the actual charges which will be charged to the customer. Here, my Linear Regression created will accept the variable values as input and predict the cost based on the input values. Below, is a small example of my prediction on the data. Example, if age is 15, BMI as 30 and smoking as no, the insurance cost the customer would be charged would be somewhere near to 2725.

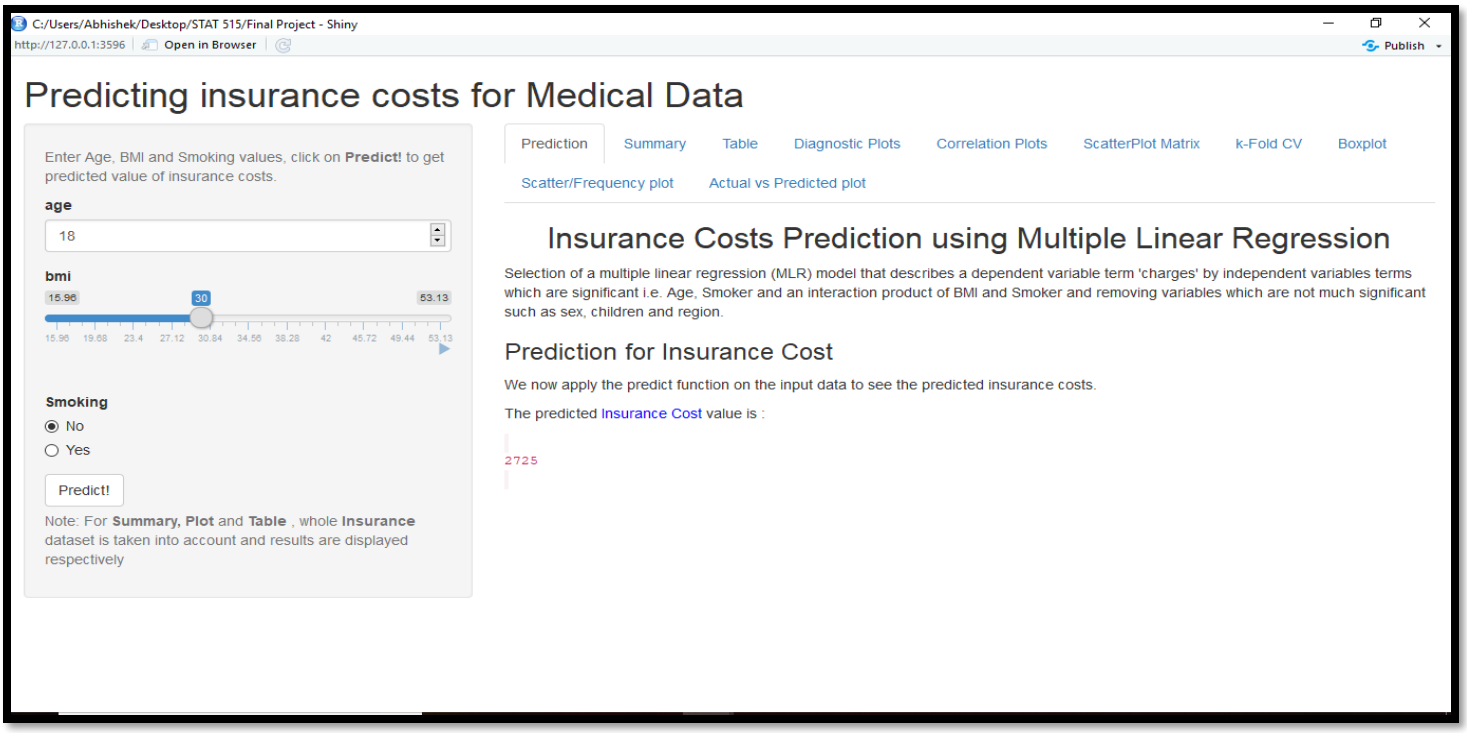

Fig 9 – Prediction of insurance cost in Shiny Interface

Making Cross Validation:

Sometimes it happens that data splitting provides an unbiased estimate of the test error; however, it can be possible that the data is biased and was unnoticed. So, we need to do Cross Validation on the data with multiple folds. In cross validation it happens that the data is equally partitioned into k-folds of equal sizes and the prediction error is calculated for each fold. The model's performance is calculated by taking average of error across the different test sets. In my model, I have performed 2-fold cross validation and visualized it in Shiny, it can be seen that the actual and predicted values are closed to the best fit line for each fold. It means that the model is performing well.

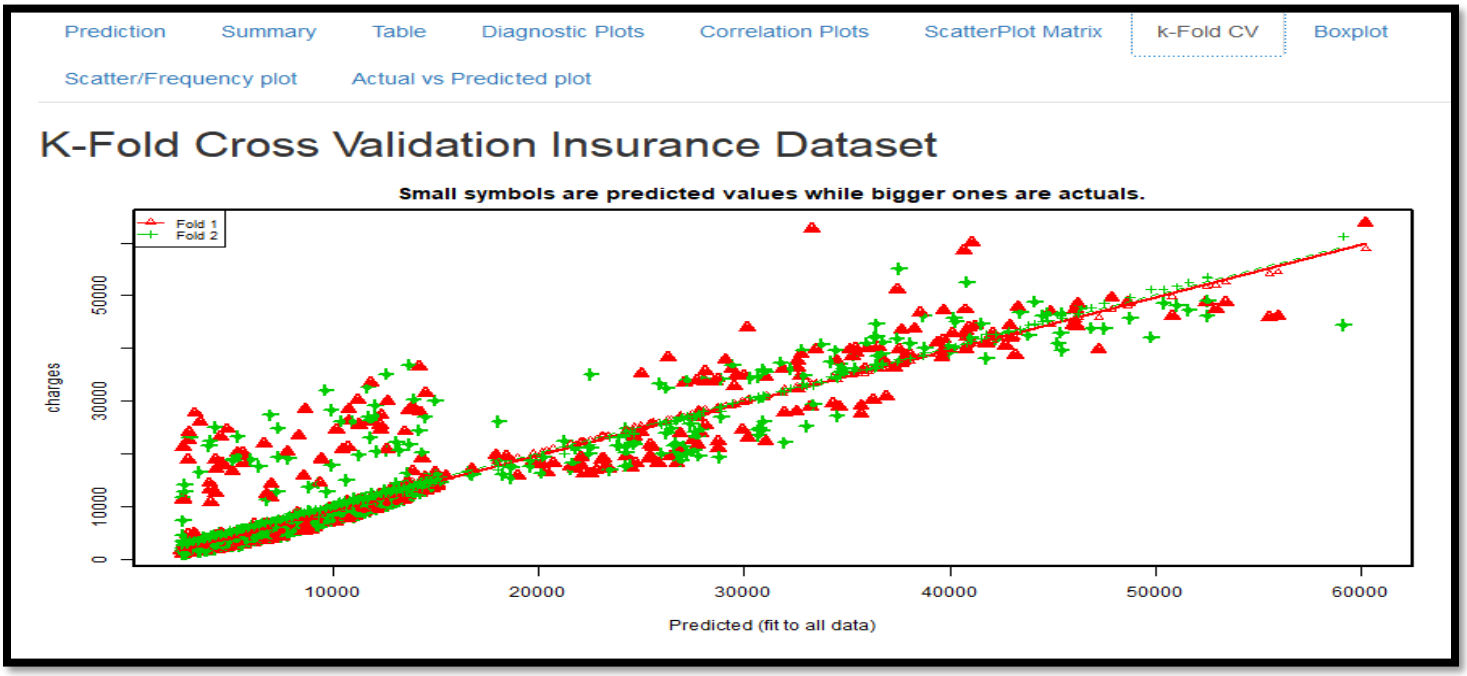

Fig 10 – 2-Fold Cross Validation on the dataset

Visualization for our model with all the significant variables:

Here, since our model used three predictors to predict the value of the response variable, it is always great to visualize the value of the outcome variable 'Charges' based on the three input variables. Here, I have shown a ggplot with Age on the x-axis, the target variable 'Charges' on the y-axis and BMI, Smoke as an aesthetic. For BMI, I have used the 'size' aesthetic and for Smoke, 'color' as an aesthetic.

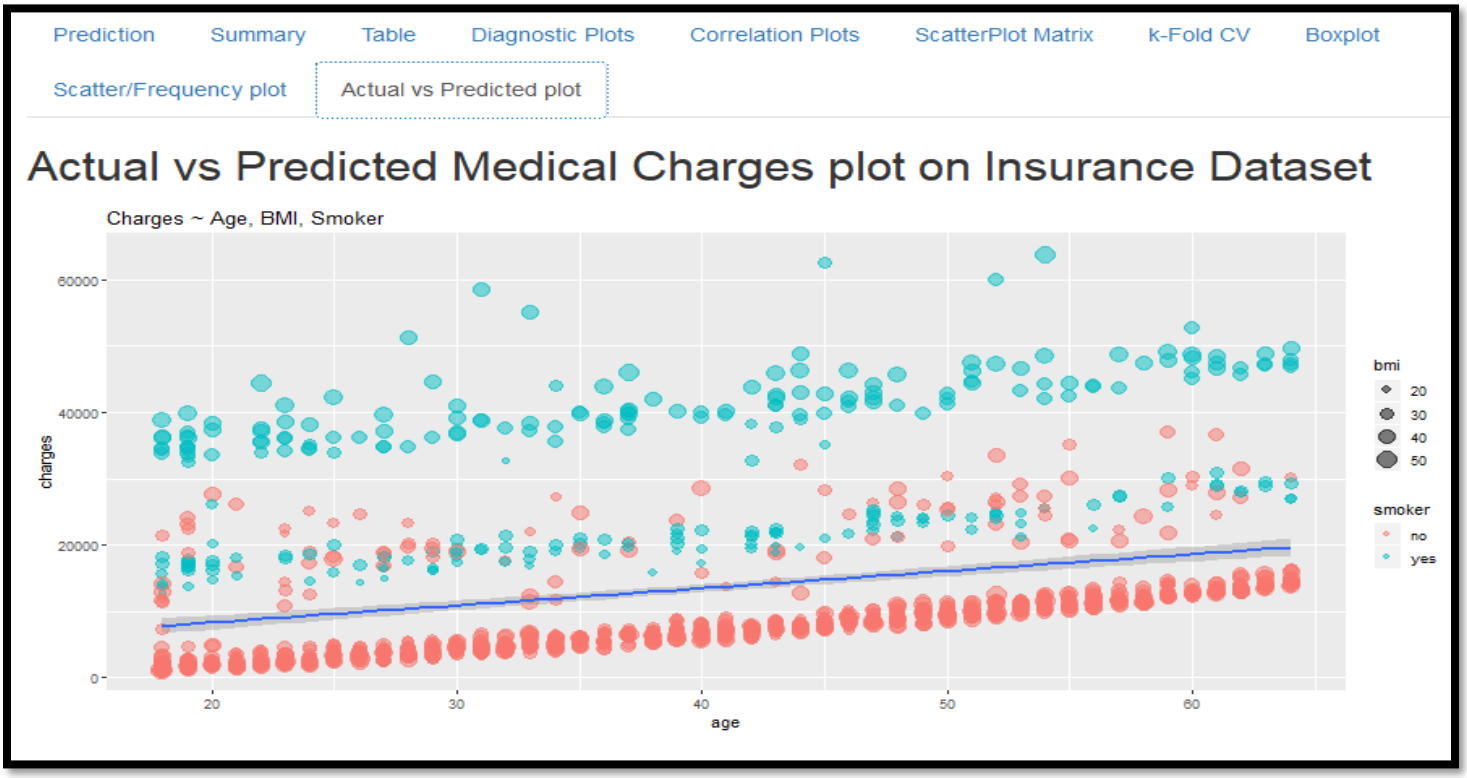

Fig. 11 – Showing Actual vs Predicted Insurance Cost for various dependent variables.