**Project**

**Question 1:**

Develop a question of your choice that can be addressed by identifying, collecting, and analyzing relevant data. You need relevant data by yourself, and describe the data such as the source, attributes, size, how the data were collected, is the dataset sample data or population data? etc. The dataset should have at least six distinct variables (i.e. columns) and a sample size (i.e. rows) of 500 or more. (3 points)

**Dataset**:

Medical Cost Personal Dataset

**Aim of the project**:

Prediction of Insurance using Linear Regression Model

**About Dataset**:

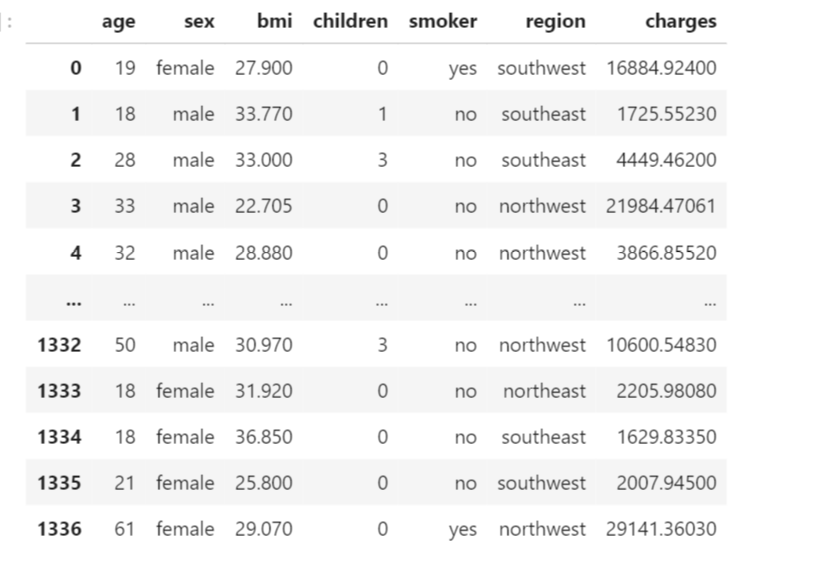
Total Number of Rows:1338

Total Number of Columns:7

1. Age: Age of the person who receives Insurance
2. Sex: Gender of the person who received Insurance
3. BMI: Body Mass Index is measure of body fat based on height and weight
4. Children: Number of children covered by health Insurance
5. Smoker: Is person smoker or not
6. Region: Residential areas of Persons who receives Insurance
7. Charges: Individual medical costs billed by health insurance

This is a Population data and it is collected from Kaggle.

Source: <https://www.kaggle.com/mirichoi0218/insurance>



**Question 2**:

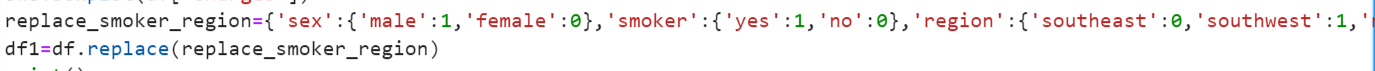
Perform exploratory data analysis (EDA). Describe the EDA process and result with at least four data visualizations. Explain whether the data sufficient to answer the question you developed based on EDA result. If it is not sufficient, how did you address the issue? (3 points)

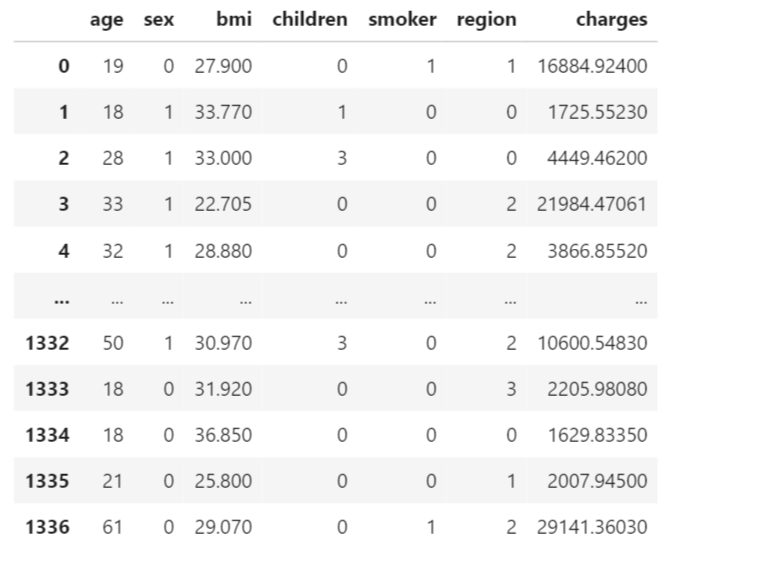
**Exploratory Data Analysis (EDA):**

Exploratory Data Analysis is very important in Data Science projects because it allows to get closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts. EDA is performed in order to define and refine the selection of feature variables that will be used for machine learning.



From Summary statistics we can observe that Mean Value is around 13279 and Standard Deviation is around 12110. Now we will find the correlation of dependent variables with Target Variable charges.When we observe data, we can find that some dependent variables are categorical variables. So, by use label encoding we will convert all categorical variables to numeric variable by giving labels 0’s and 1’s.

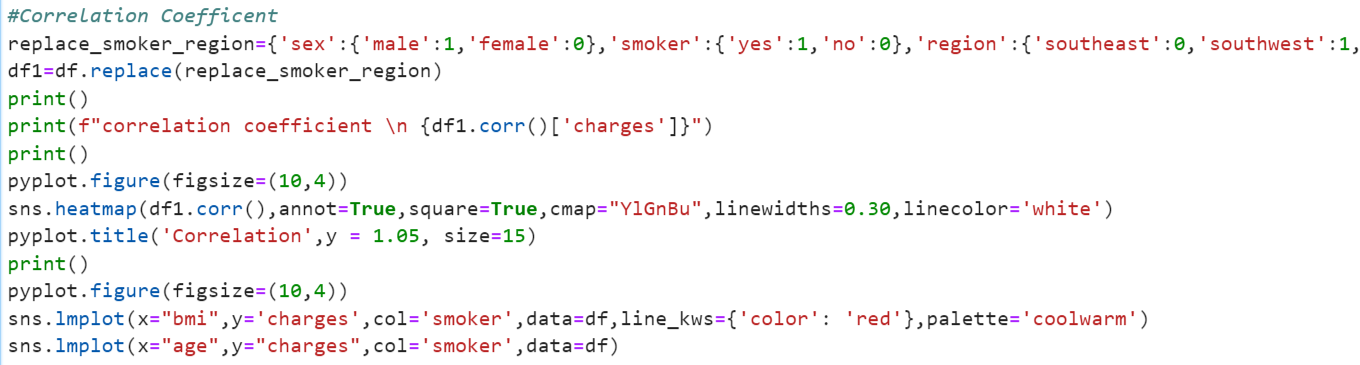




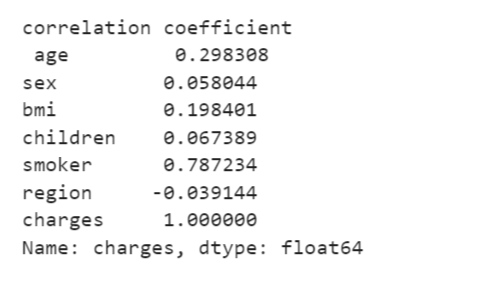
So, from above picture we can see that Sex, Smoker and region columns are converted from categorical columns to numeric by performing label encoding.

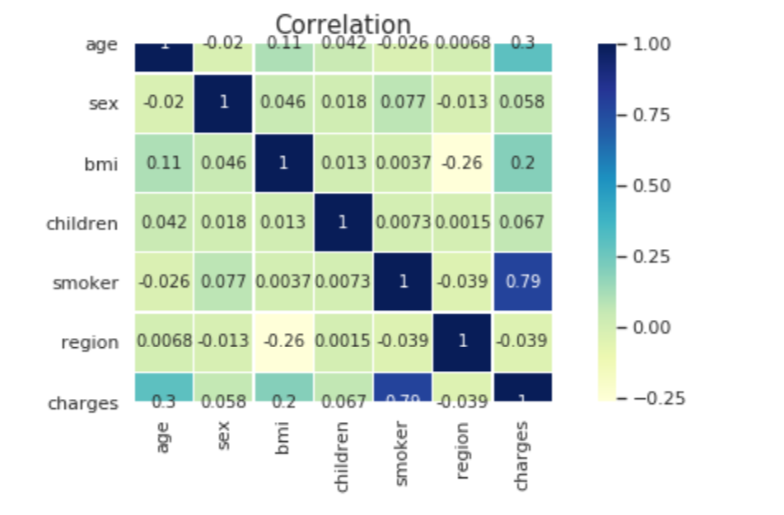
Now we will find correlation between dependent variables and Target variables

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables.



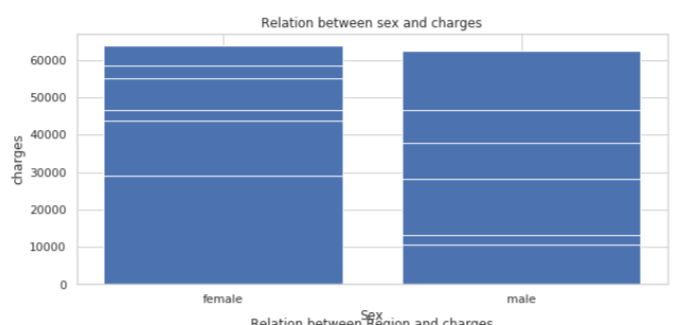
**Output 1:**

  
**Output 2**:



Among all dependent variables we can see that Smoker variable is having good correlation coefficient of 0.787234 with Target variable. We can also see that region variable has negative correlation of -0.039144 with Target variable. So, we can neglect region variable while building model.

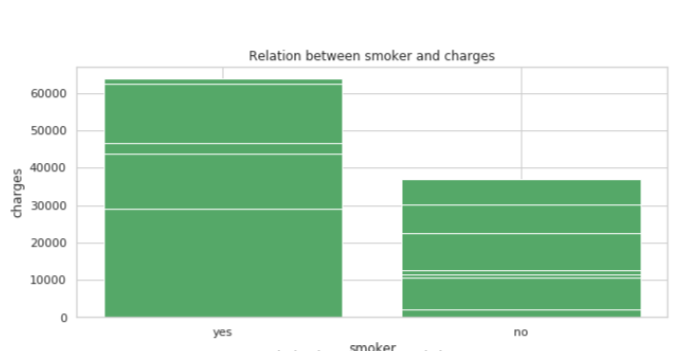
To understand more, now we will find relation between dependent variable sex and target variable charges by plotting bar graph





From the above Bar graph, we can say that there is no Gender Bias. Charges are almost same for both Male and Female.

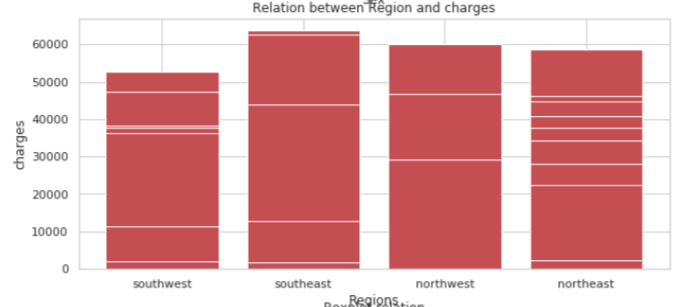
Now we will find relation between dependent variable smoker and target variable charges by plotting bar graph.





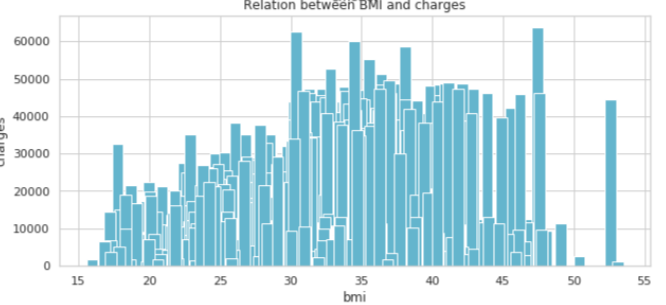
From above graph we can see that persons who smokes have high charges than people who never smoke.

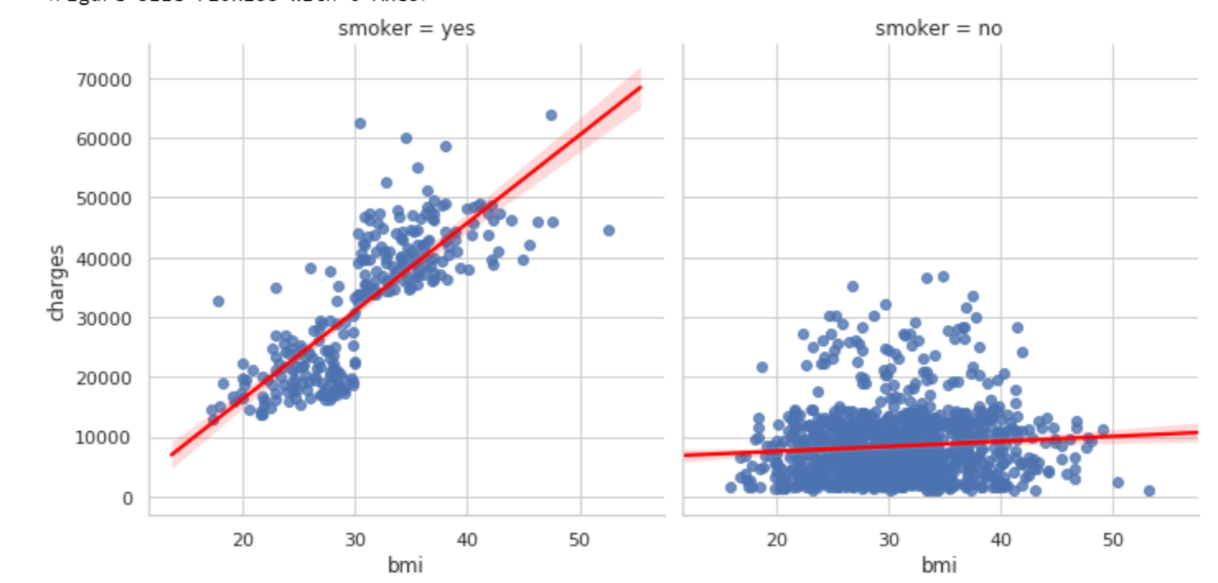
Now we will find relation between dependent variable smoker and target variable regions by plotting bar graph.



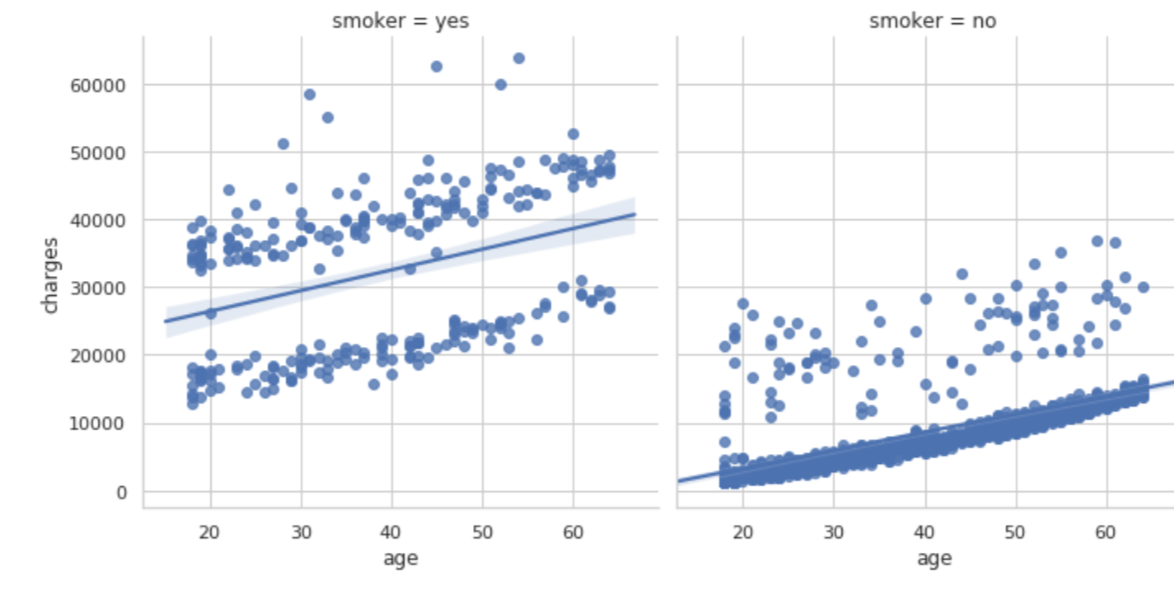
The above shows money spent by people on their**treatment** in different **regions**. From the plot we can say that the region doesn't have much impact on medical charges.

Now we will find relation between dependent variable BMI and target variable regions by plotting bar graph and regression graph.





Medical charge increases in case of **smoker** with the increasing **bmi**. But in case of **non-smokers** the increase in **bmi** doesn't have much impact on the medical charges.



From the above regression graph, we can see that in case of non-smokers, the cost of treatment increases with age, while in case of the non-smokers, there is no such impact. We can also see that age has positive correlation and has much affect after the smoker variable that leads to increase in the charges.

**Question 3:**

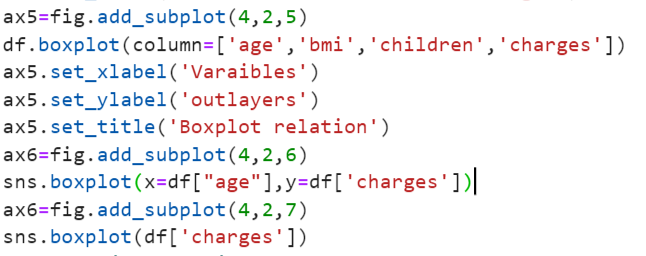
Describe any data cleaning or transformations that you perform and why they are motivated by your EDA? (2 point)

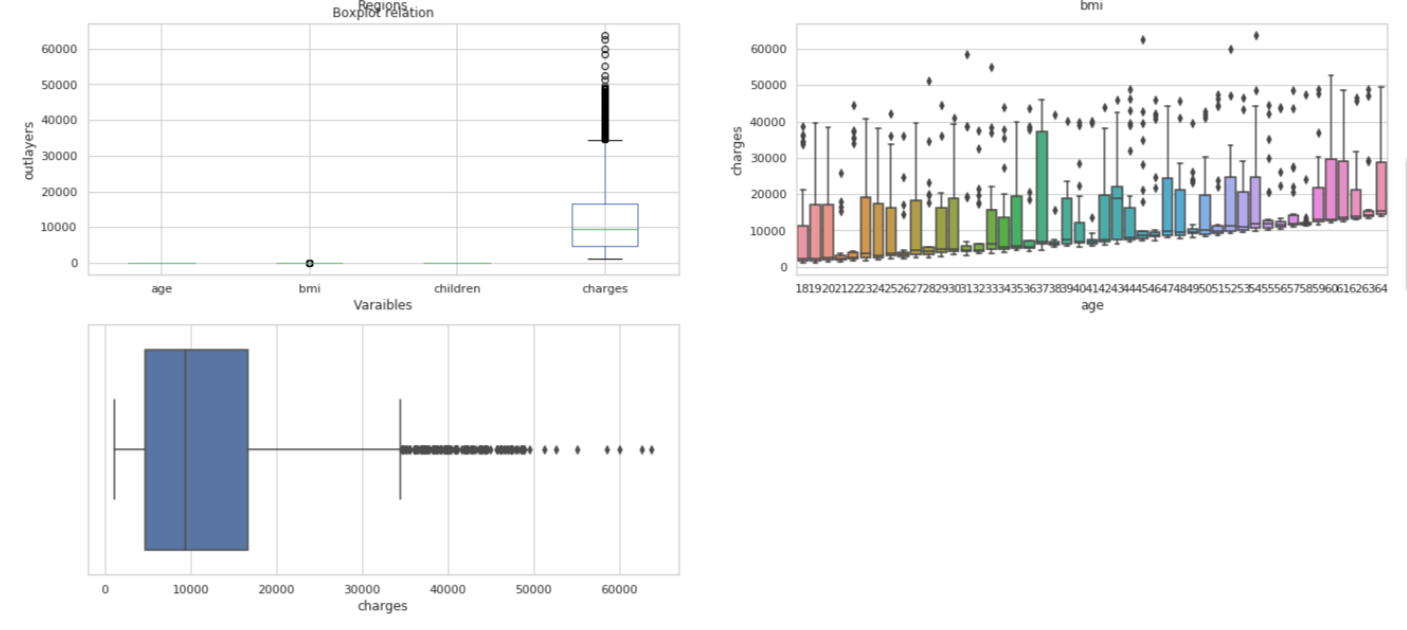
Checked if any Null values in Dataset and also Dropped duplicates in the Dataset.

We can see that there are no missing values in any features, hence no data missing steps required for this source data set.



Now, let us check if any outliers in the Dataset





From the above Boxplot graphs, we can see there are no outliers for dependent variables but for Target variable we can see so many outliers. If we remove outliers, then model accuracy will be good but there is a chance of Overfitting. So, in this case we cannot remove outliers for Charges variable.

**Question 4:**

Apply relevant inference or predication methods such as linear regression or K-nearest neighborhood (KNN) to analyze your processed ata, and validate the analysis results using cross-validation. Explain the training process, and the loss functions used in the analysis. Using examples (i.e. the values of the loss functions) to explain how the minimal value(s) of the loss function is/are found. (7 points)

In this project I used regression model to predict Target variable charges.

**Process:**

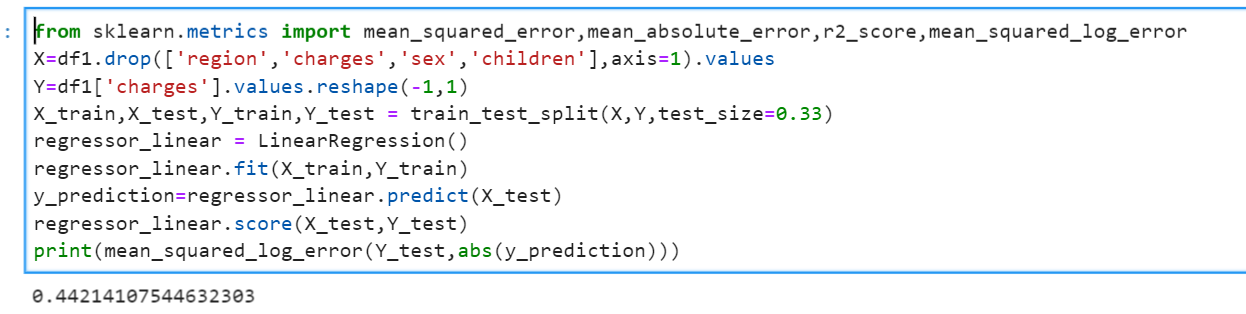
1. Splitting all dependent variables to one data frame and target variable to other data frame
2. Dividing data to X\_Train,X\_Test,Y\_Train,Y\_Test
3. Fit X\_train,Y\_train data into the model and then predict the X\_Test data.
4. Calculate Accuracy score and error rate.

Before going to the model prediction we will learn about two terms one is Score Accuracy and other one is Mean Square Error.

**Accuracy Score**: The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

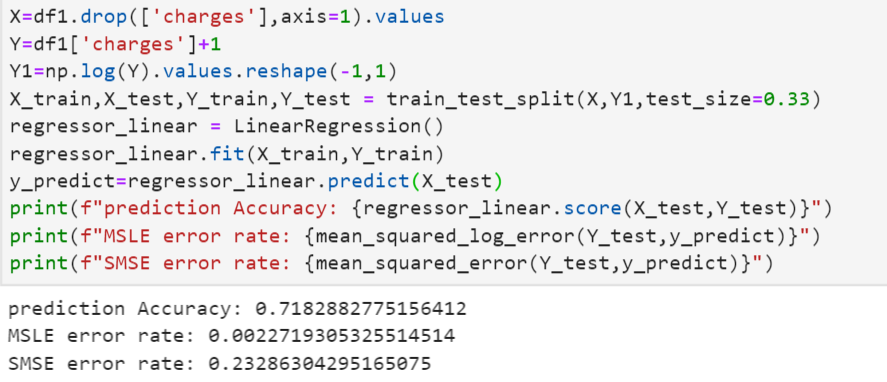
**Mean Square Error**: The average squared difference between the estimated values and the actual value.

Now, let us predict Target variable by using multiple variables before doing that we will drop region variable because we already observed that it has negative correlation with Target variable. So, we will drop region variable and check the error cost and prediction accuracy.



After performing simple linear regression model with only using smoker variable but we can see that Accuracy score is around 0.4421 which is too worst model.

So now we will build model with multiple variable and we will see if any increase in the Accuracy score.



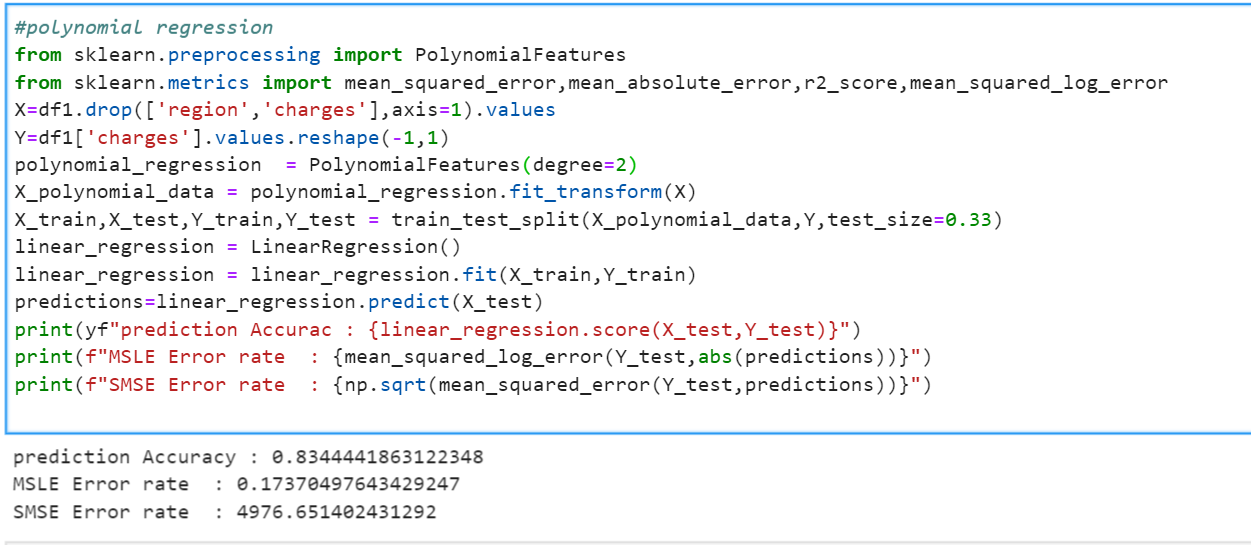
There is huge improvement in the model after using multiple variables.

We can see that Accuracy score improve a lot and it is about 71% and Error rate is about 0.23 and now we will use polynomial regression model to improve our model.

**Polynomial Regression Model**:

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in X.

We will use degree 2 and we will check how our model performs well now.

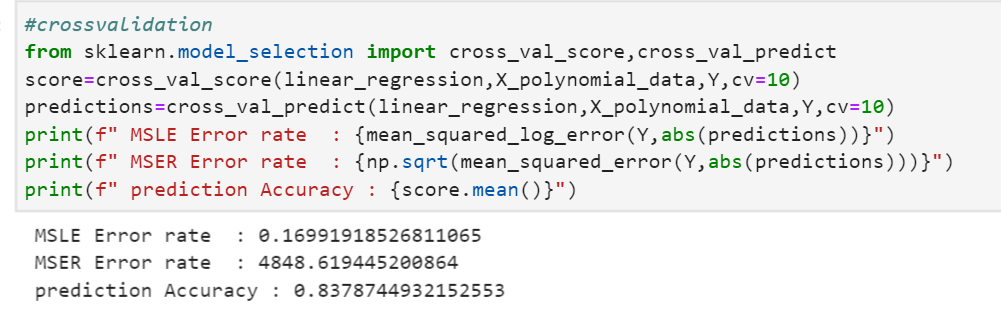


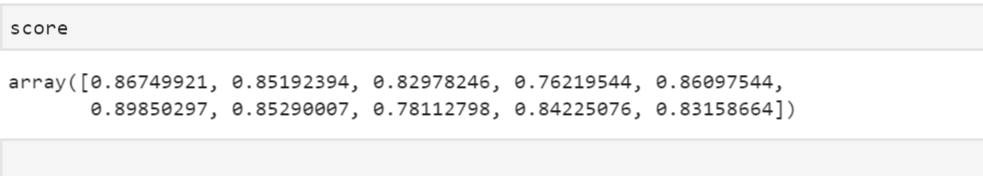
We can see that it produces good result when we compare with previos model.When we run this multiple times we will get multiple Accuracy score because Training data and Testing Data will be taken randomly basing on the percentage.So we will use Cross\_validation process to evaluate best model.

**Cross**-**validation**:

It is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data.

In this project I used K-fold cross validation Technique to evaluate model.



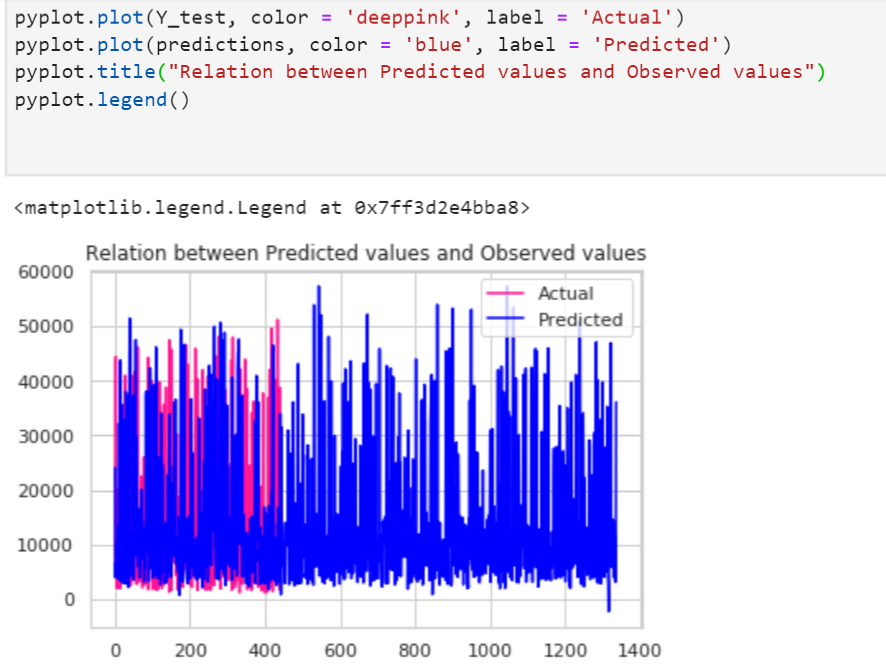


So, when we run 10 folds we got 10 Score values and we will mean all scores and we will get 0.8378. We can say that least score of the model is 0.7621 and best score of the model is 0.89850.

**Question 5:**

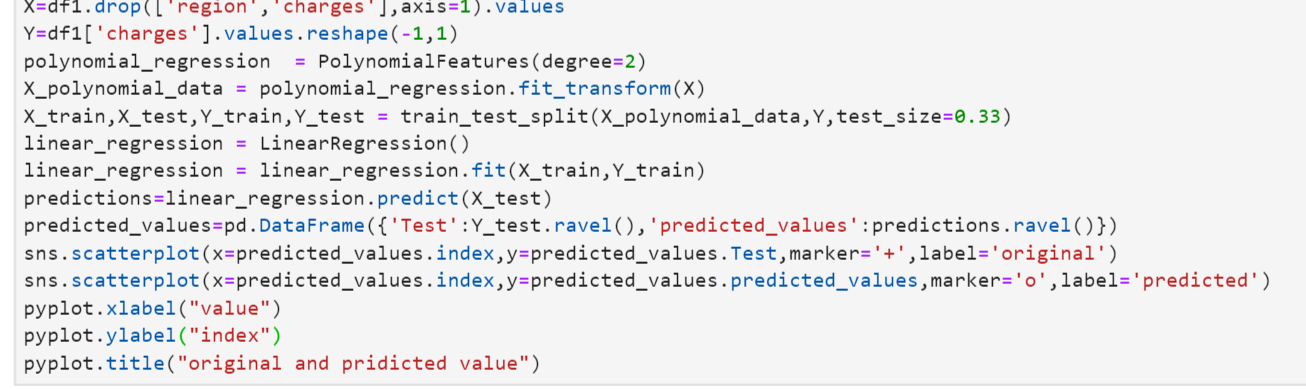
Summarize and interpret your results including at least four data visualizations. Provide an evaluation of your approach and discuss any limitations of the methods you used. (2 points)

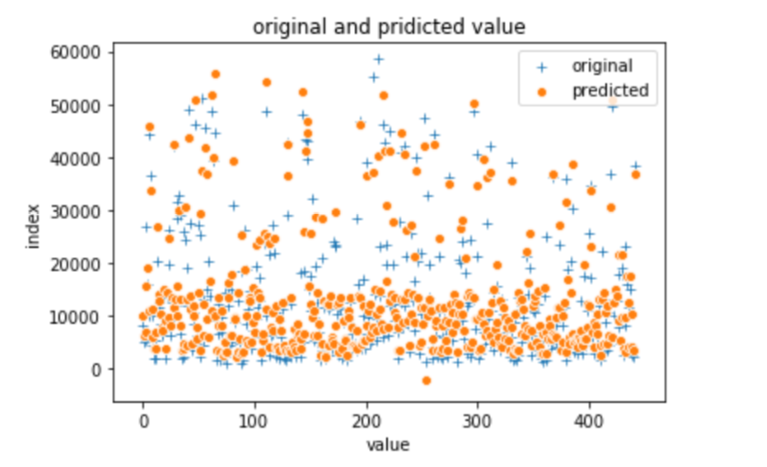
**Visualization:**

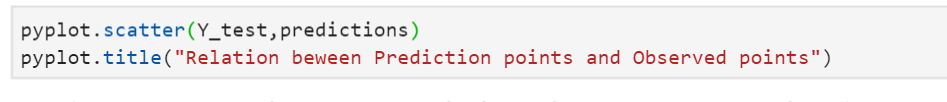


From above line graph we can observe that our model predicted good results for the medical charges of the people.

Now by visually we will see distance between our observed data points and predicted data points.







From above scatter plot graph we can observe that our model predicted good results for the medical charges of the people and Mean square error rate is also small.

