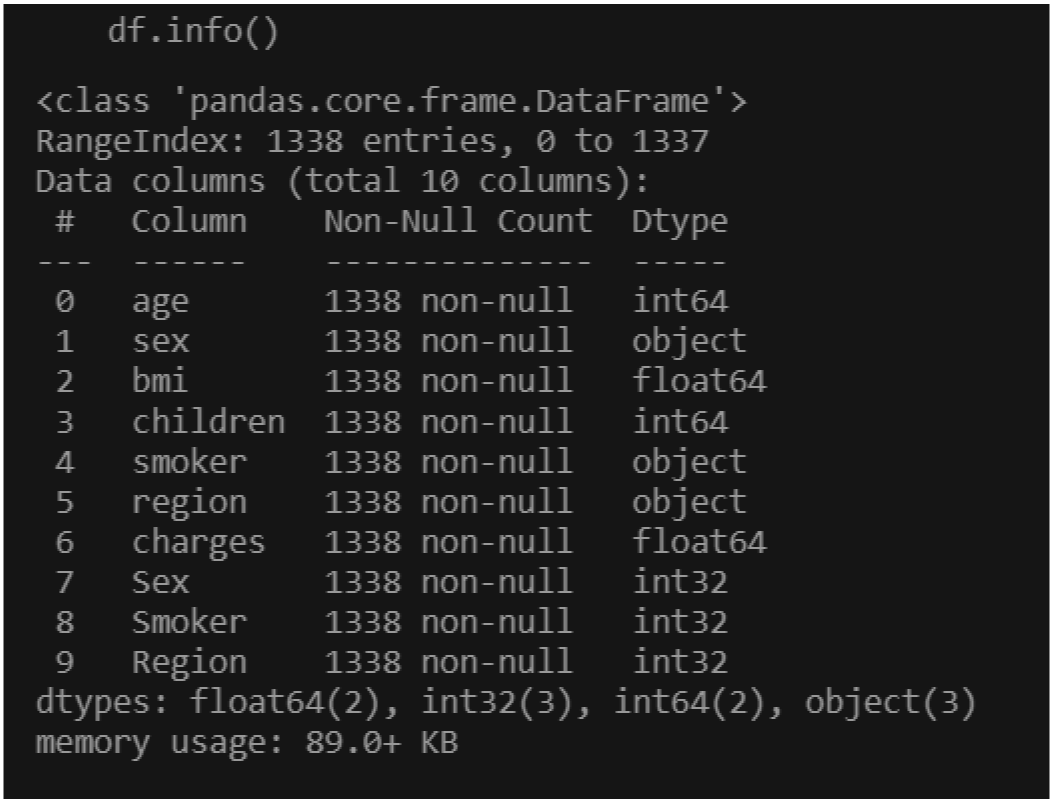
LAB REPORT 6

Mayank Raj (B19CSE053)

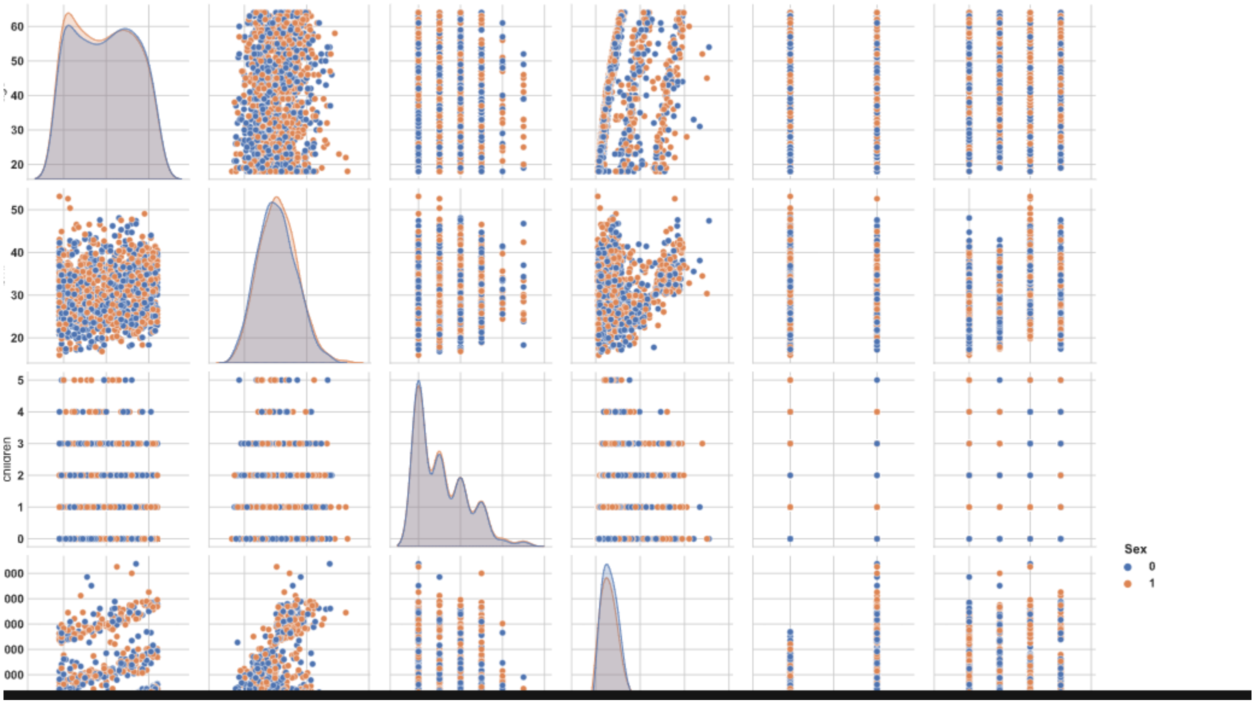
B.Tech CSE

1) Exploratory data analysis

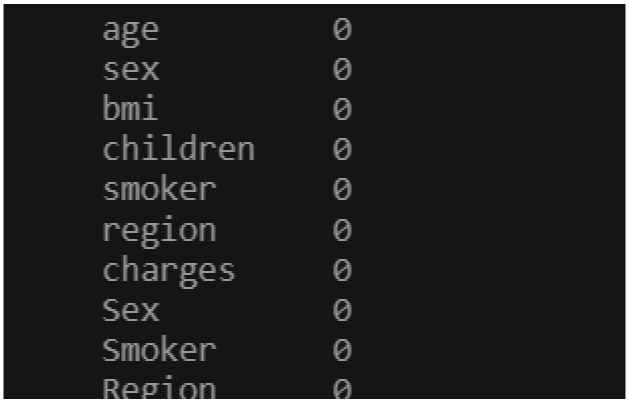


```
    df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
 7   Sex       1338 non-null   int32  
 8   Smoker    1338 non-null   int32  
 9   Region    1338 non-null   int32  
dtypes: float64(2), int32(3), int64(2), object(3)
memory usage: 89.0+ KB
```



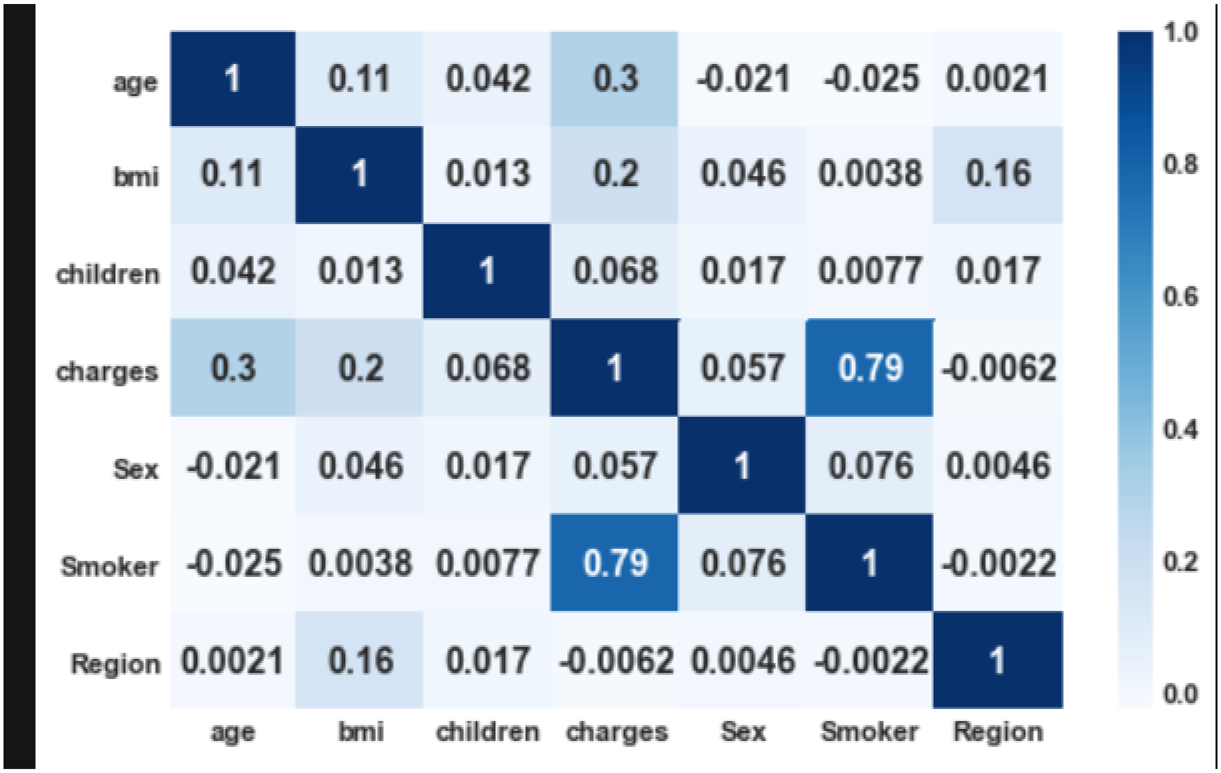
Null values:



```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
Sex         0
Smoker      0
Region      0
```

2)Correlation between variables

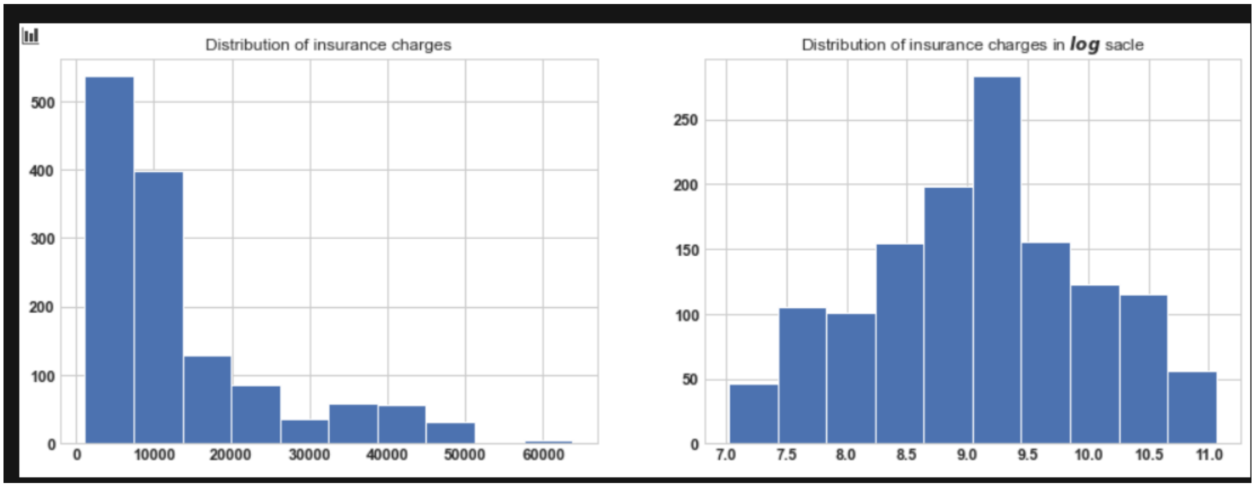
Correlations between the variables were calculated and a heatmap was plotted.



	age	bmi	children	charges	Sex	Smoker	Region
age	1	0.11	0.042	0.3	-0.021	-0.025	0.0021
bmi	0.11	1	0.013	0.2	0.046	0.0038	0.16
children	0.042	0.013	1	0.068	0.017	0.0077	0.017
charges	0.3	0.2	0.068	1	0.057	0.79	-0.0062
Sex	-0.021	0.046	0.017	0.057	1	0.076	0.0046
Smoker	-0.025	0.0038	0.0077	0.79	0.076	1	-0.0022
Region	0.0021	0.16	0.017	-0.0062	0.0046	-0.0022	1

3)Distribution of dependent variable and its log value was plotted

The plots show that the distribution is right skewed.



4)Conversion of categorical values into numerical

Categorical variables were converted into numerical values using Label encoder.

5)Data was splitted into train and test dataset with test size 0.3

6)Built the regression model using the given equation without any in-built library except numpy

Regression model was built using the equation $\theta=(X^TX)^{-1}X^Ty$ using numpy functionalities.

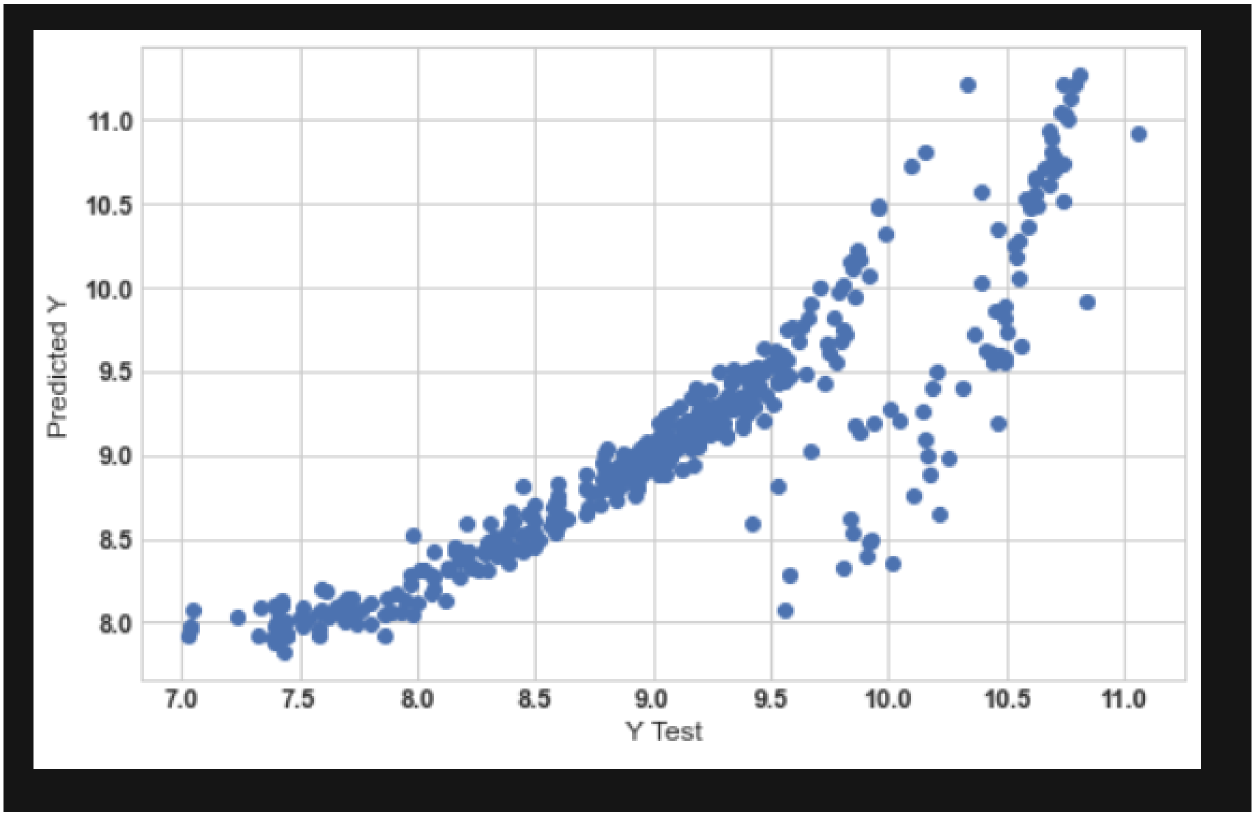
7)Trained the regression model using sklearn library

8)Comparison of parameters (coefficients) of both the models:

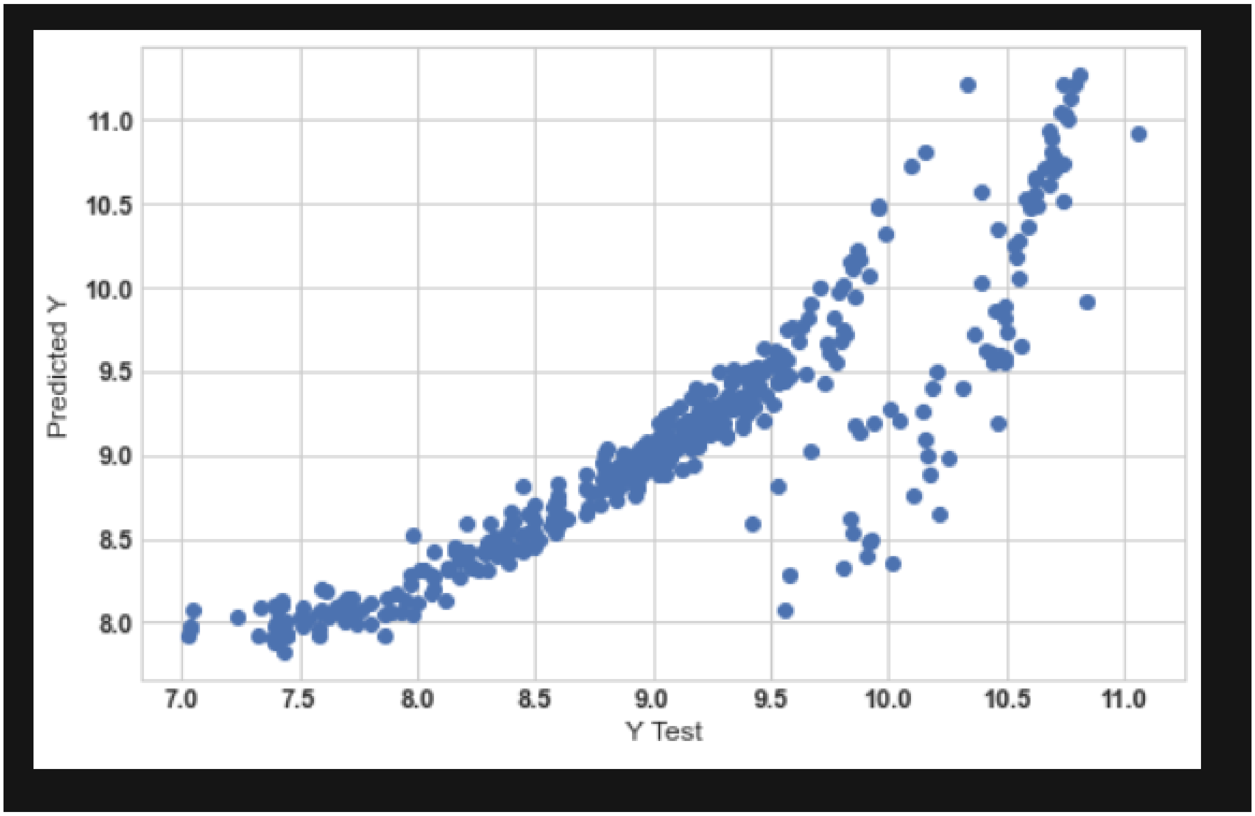
	age	bmi	children	sex	smoker	region
Manual	0.03325052	0.01407749	0.10057153	-0.04105018	1.52389362	-0.05391283
Using sklearn	0.03325052	0.01407749	0.10057153	-0.04105018	1.52389362	-0.05391283

9)Prediction were made using both the models:

Score for manual model: 0.7892774635828786

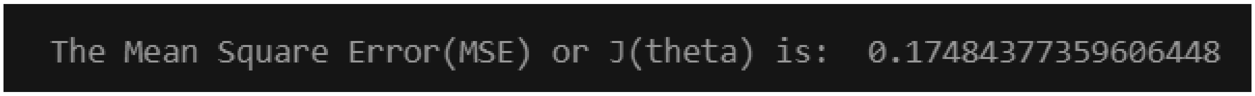


Score for sklearn model: 0.7892774635828786



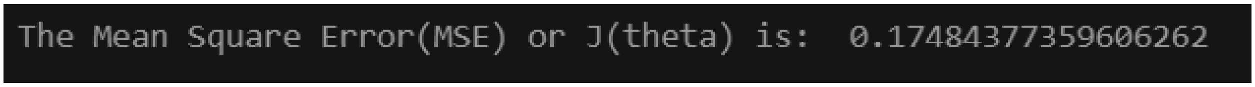
10)MSE was calculated using for both the models using manually made function and inbuilt function:

MSE using manual function :



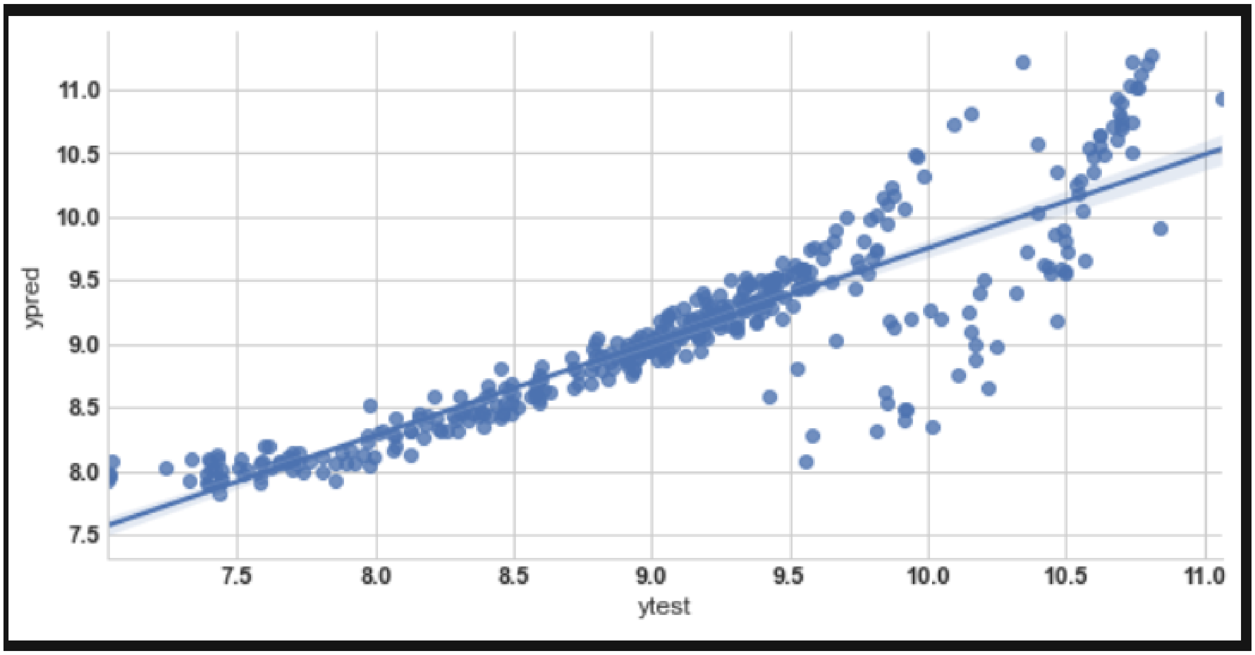
```
The Mean Square Error(MSE) or J(theta) is:  0.17484377359606448
```

MSE using inbuilt function from sklearn:



```
The Mean Square Error(MSE) or J(theta) is:  0.17484377359606262
```

11)Plotted the relationship between actual and predicted value:


ypred
11.0
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.5
8.0
8.5
9.0
9.5
10.0
10.5
11.0
ytest