# PRML LAB-6 REPORT

KSHITIZ
B19CSE111

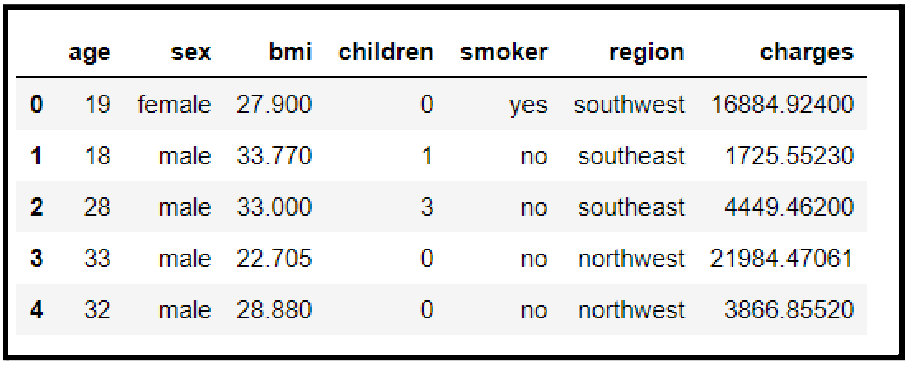
1. Problem statement:
    - Given a Medical Cost Dataset we need to build a Linear Regression Model.
    - The dataset consists of age, sex, BMI(body mass index), children, smoker and region feature, which are independent features while charge is the dependent feature(target variable)
    - The hypothesis function provided is as follows:

$$h_\theta(x_i)=\theta_0+\theta_1 age+\theta_2 sex+\theta_3 bmi+\theta_4 children+\theta_5 smoker+\theta_6 region$$

    - To predict the individual medical costs billed by the health insurance.
2. Linear regression
    - It is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables i.e, independent variable(x) and dependent variable(y)  respectively.
    - It utilises the principle of Mean Square Error(MSE) which aims to minimize the sum of square differences between observed dependent variables in the given dataset and the ones predicted by linear regression.
3. Import Necessary Dependencies and Dataset
    - We start by importing necessary libraries for our analysis.
    - For reading the dataset we use the read_csv() function of pandas library.

  - The independent variables in the dataset are age, sex, bmi, children, smoker, region while charges is the target variable.
  - Since more than one independent variable are used to predict the outcome, therefore we use Multiple Linear Regression.

## 4.Exploratory Data Analysis

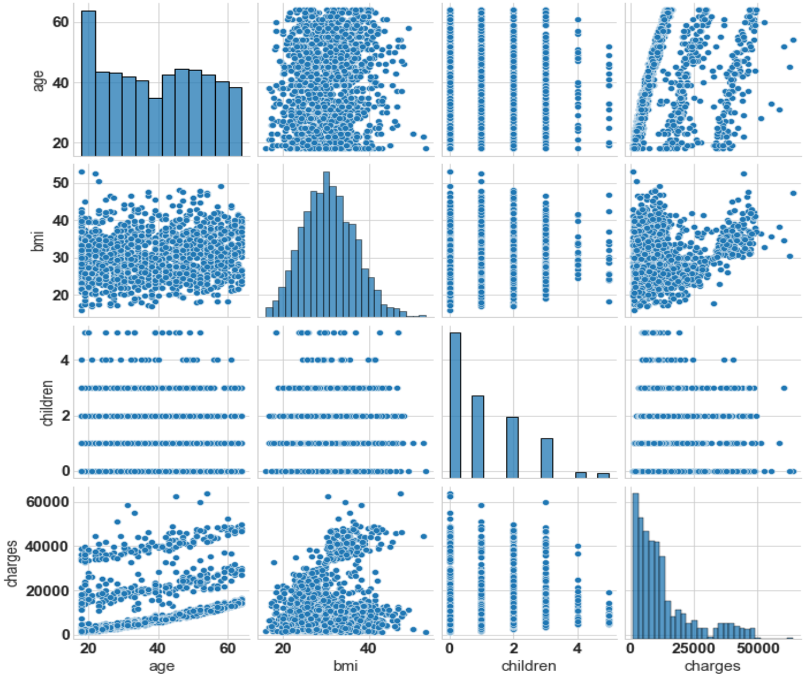


| | age | sex | bmi | children | smoker | region | charges |
|---|---|---|---|---|---|---|---|
| **0** | 19 | female | 27.900 | 0 | yes | southwest | 16884.92400 |
| **1** | 18 | male | 33.770 | 1 | no | southeast | 1725.55230 |
| **2** | 28 | male | 33.000 | 3 | no | southeast | 4449.46200 |
| **3** | 33 | male | 22.705 | 0 | no | northwest | 21984.47061 |
| **4** | 32 | male | 28.880 | 0 | no | northwest | 3866.85520 |

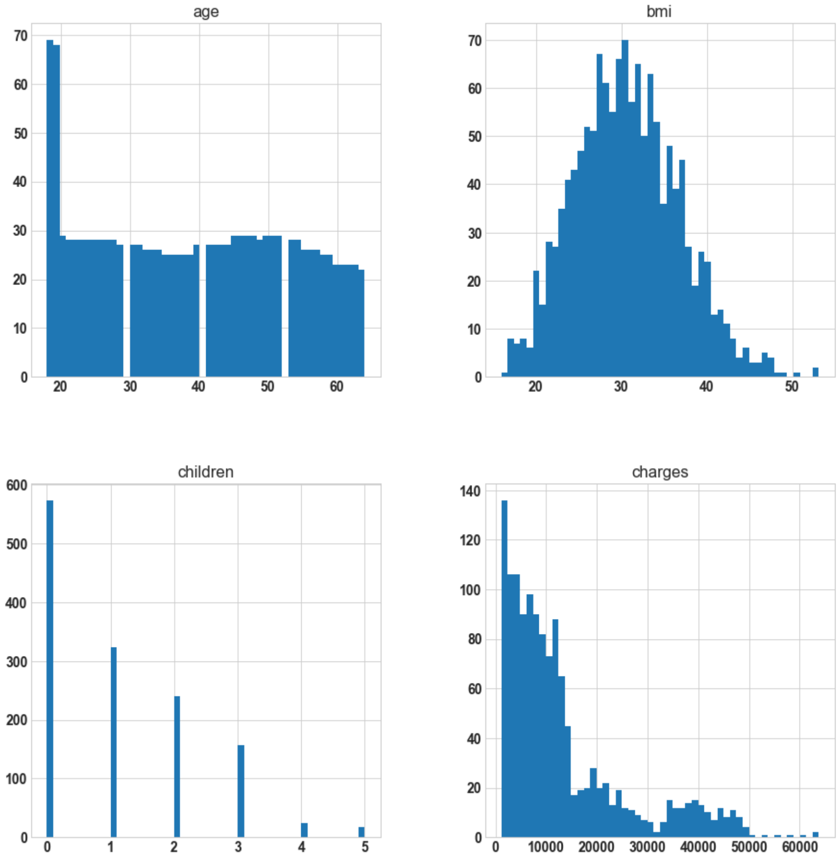
- The functions used in Exploratory Data Analysis are:
  - df.info - It prints the basic information about the DataFrame including the index dtype and memory usage.
  - df.shape - It prints the dimensions of the dataset.
  - In the given problem statement shape of the dataset is (1338,7)(i.e., 1338 - training examples and 7 independent variables)
  - df.columns - It prints the name of all the columns present in the dataset.
    - all_cols= ['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges']

- df.describe - It prints the statistical quantities for all the numerical features including mean, std deviation etc.
- df.nunique() - It prints the number of unique values present in each of the columns.
- In order to find the numerical columns, we use the list comprehension and then print those columns which do not have object datatype
- Similarly, for printing the categorical columns, we select those columns which have data type as Object(‘O’)

```
numerical_col = ['age', 'bmi', 'children', 'charges']
categorical_col = ['sex', 'smoker', 'region']
```

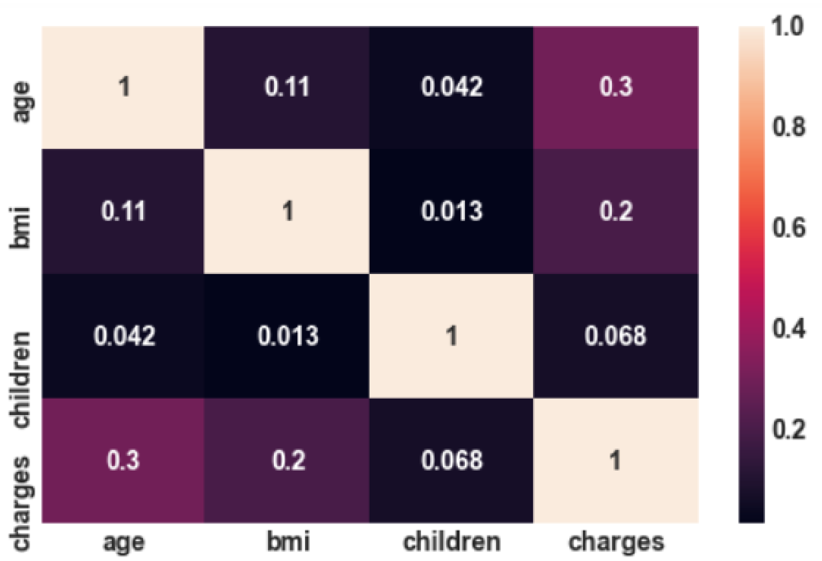


Pairplot

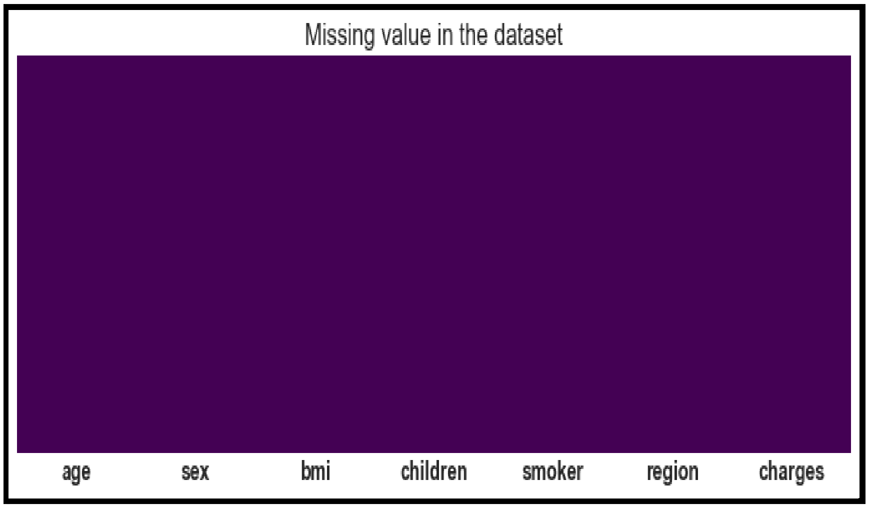


- From the histogram we observe that ‘bmi’ is almost numerically distributed while ‘charges’ are right-skewed.

5. Plot correlation between different variables and analyze whether there is correlation between any pairs of variables or not.
    - In order to find the correlation between the numerical columns, we use df.corr() function.Moreover, we also use the seaborn library to visualise the correlations using heatmap.
    - We observe that there does not exist musch correlation between independent features and hence we don't have the problem of multicollinearity.

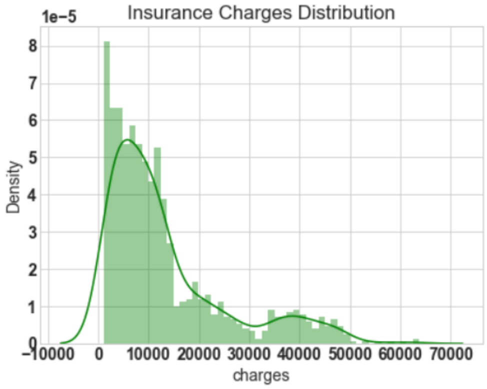


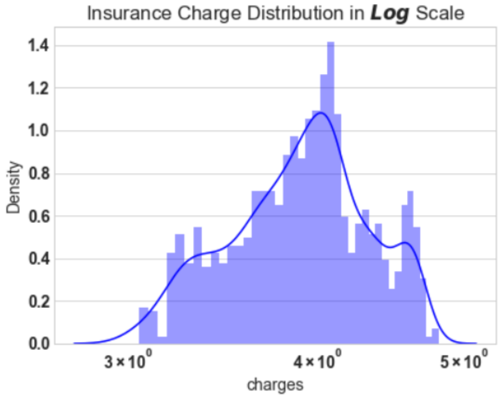
-

6. Check for missing values in the dataset
    - In order to find the missing values column wise, we use df.isnull().sum() function.
    - The above mentioned function prints the number of missing values in the data set.
    - We can therefore conclude that there are no missing values in the dataset.
    - 

7. Plot the distribution of the dependent variable and check for skewness (right or left skewed) in the distribution.
    - From the first plot we can infer that the charges vary from 1120 to 63500 and hence it is right skewed.

- Subsequently, in the second plot we apply natural log and observe that the plot approximately tends to be a normal plot.
- For further analysis we will apply log on target variable charges.





## 8. Box-Cox transformation

- A Box Cox transformation is a method to transform the non-normal dependent variables into a normal shape.
- In statistics, normality is an important assumption and therefore the transformation becomes important as we are able to multiple statistical tests as well.
- In order to perform the transformation we first find the lambda value and then apply the following rule to the variable.

$$\boxed{y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}}$$

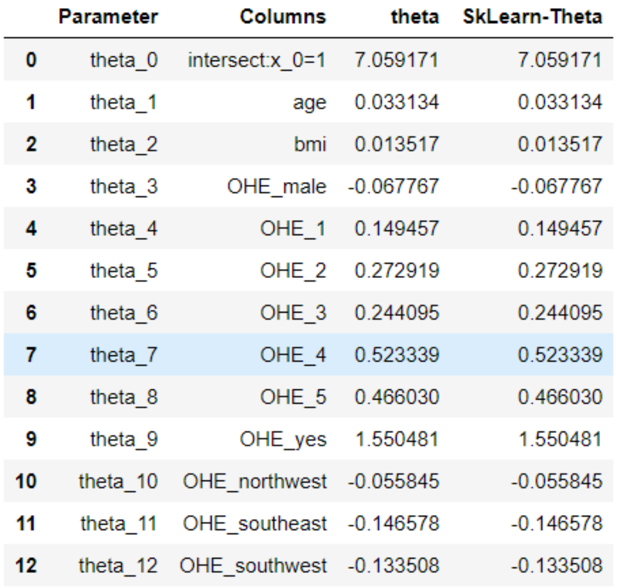
## 9. Converting categorical data into numbers.

- Machine learning algorithms cannot work with the categorical data directly and therefore in order to

    apply various algorithms,the categorical data must be converted into numbers.
    - There are mainly three types of encoding including:
        i. Dummy variable trap - The Dummy variable trap is a type of encoding in which the independent variables are multicollinear(i.e,more than 1 variables are highly correlated)
        ii. Label Encoding - It refers to converting the word labels into numerical form so that the algorithms can be able to operate on them.
        iii. One hot encoding - One hot encoding is a type of encoding method which allows the representation of categorical variables as binary vectors. It involves mapping the categorical values to integer values followed by representing them as a binary vector which is assigned all zero values except for the index of the integer, which is marked with 1.
    - In the given problem statement we have used a dummy variable trap using the get_dummies function from the pandas library
    - Moreover,setting the drop_first = True helps to reduce the extra column created during dummy variable creation thereby reducing the correlations created among dummy variables.

10. Splitting the data into training and testing sets with ratio 0.3
    - In order to split the dataset into a training and testing set with test_ratio = 0.3 we use the train_test_split() function from the model_selection module of sklearn library (param - test_size=0.3)
11. Linear Regression Model
    - Using Linear Regression Equation $\theta=(X^TX)^{-1}X^Ty$
    - Using Sklearn Library
12. Comparison of Parameters in 6(a) and 6(b) respectively

- The parameters obtained from both the models are the same which was expected.
- Thus we can conclude that our model was correctly built using the normal equations and further verified using the sklearn library.



| | Parameter | Columns | theta | SkLearn-Theta |
|---|---|---|---|---|
| **0** | theta_0 | intersect:x_0=1 | 7.059171 | 7.059171 |
| **1** | theta_1 | age | 0.033134 | 0.033134 |
| **2** | theta_2 | bmi | 0.013517 | 0.013517 |
| **3** | theta_3 | OHE_male | -0.067767 | -0.067767 |
| **4** | theta_4 | OHE_1 | 0.149457 | 0.149457 |
| **5** | theta_5 | OHE_2 | 0.272919 | 0.272919 |
| **6** | theta_6 | OHE_3 | 0.244095 | 0.244095 |
| **7** | theta_7 | OHE_4 | 0.523339 | 0.523339 |
| **8** | theta_8 | OHE_5 | 0.466030 | 0.466030 |
| **9** | theta_9 | OHE_yes | 1.550481 | 1.550481 |
| **10** | theta_10 | OHE_northwest | -0.055845 | -0.055845 |
| **11** | theta_11 | OHE_southeast | -0.146578 | -0.146578 |
| **12** | theta_12 | OHE_southwest | -0.133508 | -0.133508 |

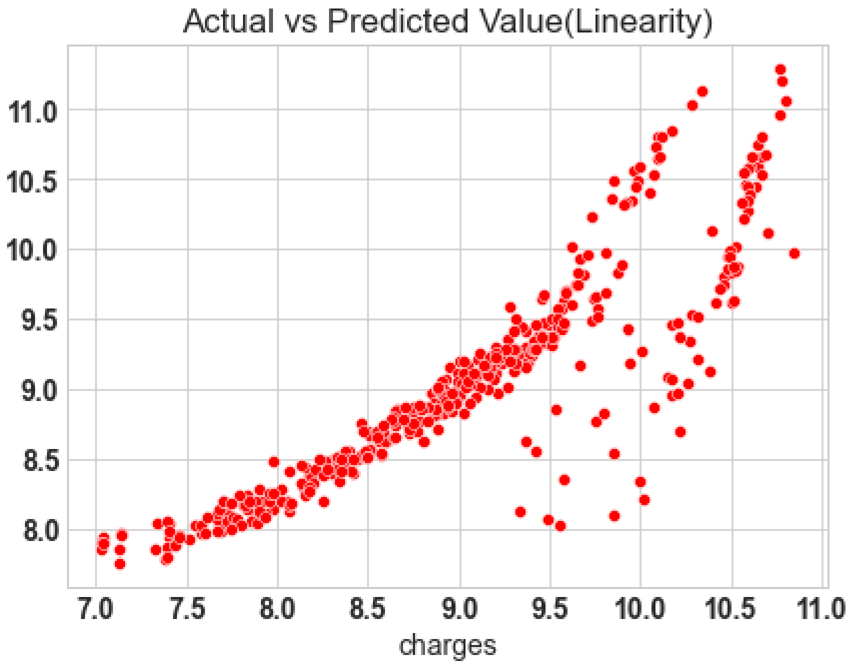
13. Prediction from both model
    - We observe that both the sklearn model and normal equation give almost the same values of R_Square and MSE.
    - Thus we can conclude that these models are closely related and their test predictions are also almost similar.
    - The R_square score using the normal equation is 0.7795687545055329
    - The R_square score using the sklearn library is is 0.779568754505532
14. Perform evaluation using the MSE of both models
    1. Using Linear Regression Equation
    2. Using Sklearn Library
    - MSE or Mean Squared Error measures the average of the squares of the errors(i.e., the average squared

difference between the estimated values and what is to be estimated).Accordingly,

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} (\underbrace{y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i}_{\text{actual value}})^2$$

- The Mean Square Error(MSE) or J(theta)using normal equation is 0.18729622322981812
- The Mean Square Error(MSE) or J(theta) using sklearn library is 0.18729622322981884

15. Plotting the actual and the predicted values
    - Relationship between the dependent and independent variable. (for both the models)



16. Model validation
    - For model validation various assumptions of the linear regression model needs to be verified.
    - The common assumptions for Linear Regression Model are as follows:
        i. The relationship between the dependent and independent variable has to be linear in Linear

Regression which can be verified by the scatter plot of Actual vs Predicted value

ii. The residual error plot should be normally distributed and the mean of the residual error should be 0 or close to 0.

iii. All the variables in Linear Regression should be multivariate normal. This assumption can be verified with a Q-Q plot.

iv. Linear regression assumes that there is little or no multicollinearity in the data. Generally, multicollinearity occurs only when the independent variables are too highly correlated with each other.

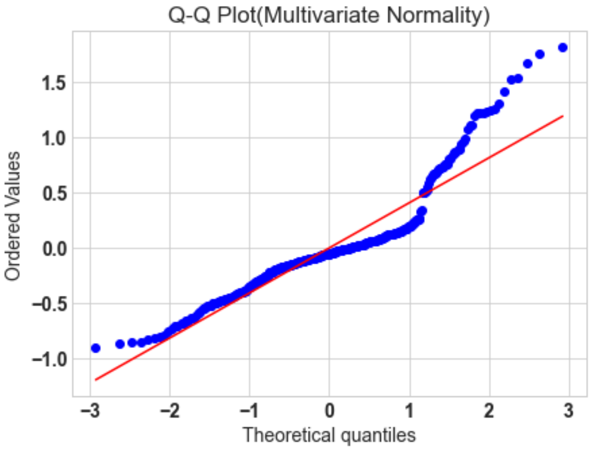
v. The variance inflation factor (VIF) identifies the correlation between independent variables as well as strength of that correlation.

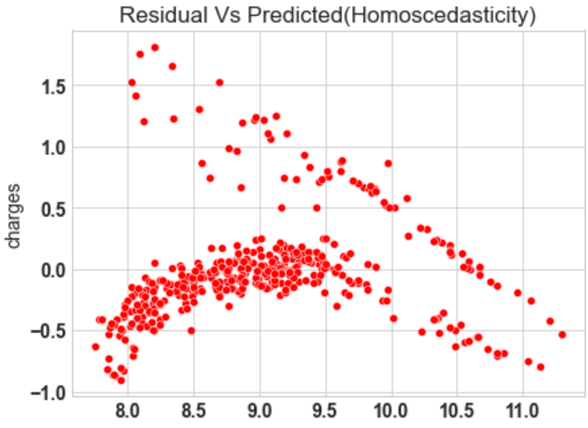
vi. It is defined as VIF = 1/ 1-(R^2)

vii. If the value of VIF belongs to (1,5) then it is termed as moderate correlation.

viii. The homoscedasticity of data refers to the fact that the residuals are equal across the regression line.From the plot we observe that the plot exhibits a funnel shape pattern.

ix.





17. We observe that the actual vs predicted plot in our model is curved, hence we can say that the linear assumption fails.
    - Similarly, the residual mean value is zero as well as the residual error plot is also right skewed.
    - From the Q-Q plot given above we can see that for values greater than 1.5 it tends to increase
    - Moreover, since the variance inflation factor(VIF) value < 5,therefore there is no multicollearity.