Course Project

Topic 2: Predicting Medical Insurance Cost Using Linear Regression

IE7280:Statistical Methods of Engineering
Professor Nasser Fard

Aditya Trivedi
Ravi Patel

trivedi.adi@northeastern.edu
patel.ravikh@northeastern.edu

Topic 2: Predicting Medical Insurance Cost Using Linear Regression:

Overview:

The project "Predicting Medical Insurance Cost using Linear Regression" attempts to solve the problem of comprehending and forecasting health insurance costs by considering many important variables. Many factors go into determining premiums in the health insurance market, and the goal of this research is to apply a linear regression technique to predict the correlations between particular variables and the corresponding insurance prices. Important variables included in the dataset include age, gender, body mass index (BMI), number of dependents, smoking status, and the insurance contractor's location. Through the investigation of these variables, the project aims to create a prediction model that will provide stakeholders in the insurance industry and people with important information about the factors affecting the cost of medical insurance.

Problem Statement:

A person trying to control their healthcare costs must grasp the links between the many lifestyle and demographic factors that affect the cost of health insurance. The goal of this research is to build a solid linear regression model that can accurately forecast health insurance expenses by taking into account the following crucial variables:

1. Age: An important factor in calculating insurance rates is the primary beneficiary's age, with older people often paying more.

2. Gender: This study aims to quantify the association between being male or female and the related insurance charges. Gender can have an influence on insurance prices.

3. Body Mass Index, or BMI, is a measure of body weight in relation to height that is objective. This factor looks into the relationship between changes in BMI and changes in health insurance premiums.

4. Number of Dependents: One important consideration is the number of dependents, or children, that the insurance covers. Larger families may pay higher insurance costs.

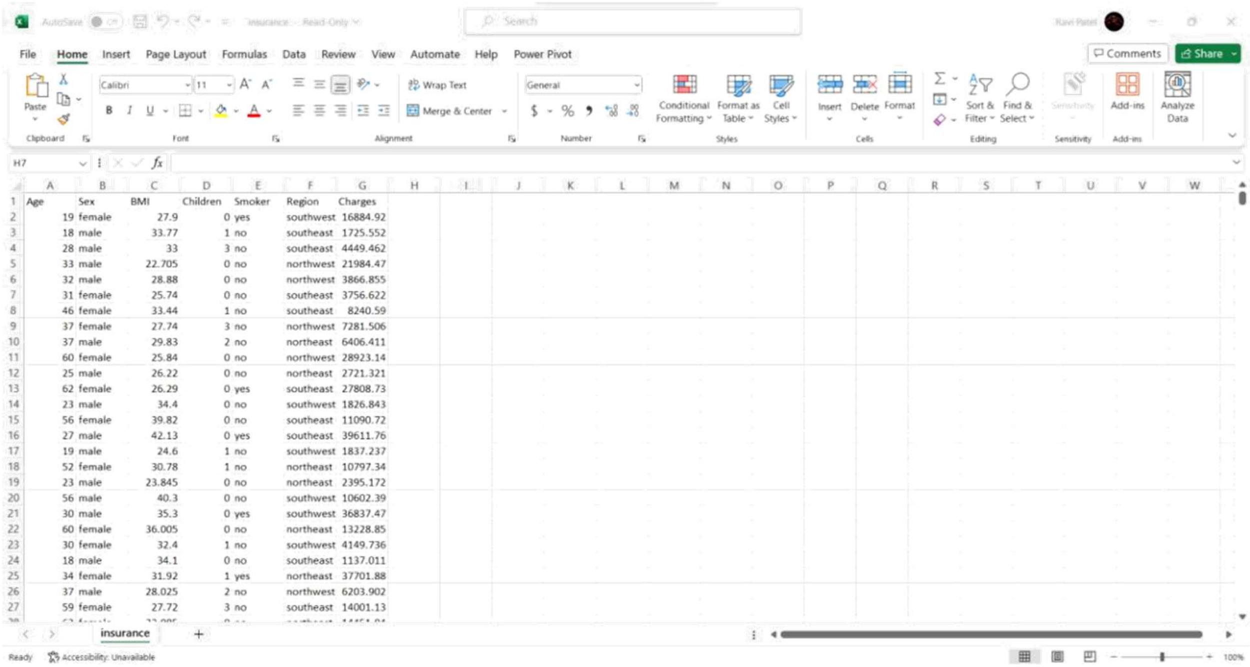
5. Smoking Habits: It is well known that smoking increases the risk of a number of illnesses. The purpose of thi factor is to calculate the effect of smoking on health insurance premiums.

6. Geographic Region: Regional differences in healthcare prices are introduced by the insurance contractor's residence location in the US (northeast, southeast, southwest, northwest). The goal of this factor is to determine how a person's location affects insurance costs.

A dataset including historical data on people's insurance costs and the associated values of these influencing variables will be used to train the prediction model created by linear regression. The model will be assessed for accuracy and dependability in estimating insurance costs, offering a useful tool for people to project their future medical bills. The knowledge gathered from this model can also help insurance companies and legislators understand the dynamics of premium drivers and perhaps direct the development of cost-control and cost-optimization plans.

Dataset:

The Dataset snapshot is below, and it has 6 features.



Age	Sex	BMI	Children	Smoker	Region	Charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321
62	female	26.29	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.72
27	male	42.13	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.34
23	male	23.845	0	no	northeast	2395.172
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47
60	female	36.005	0	no	northeast	13228.85
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.88
37	male	28.025	2	no	northwest	6203.902
59	female	27.72	3	no	southeast	14001.13

Exploratory Data Analysis:

The data has all 6 features as its main because they all are correlated to charges. The dataset has 6 features with 1338 records.

Loading the data set. Getting to know the data type variable and the shape of the data.

Section 1: Data Collection and Data Processinng

```
#Loading the data using pandas.
insurance_df=pd.read_csv('insurance.csv')
in_df=pd.read_csv('insurance.csv')
```

```
#Getting data types of each column.
insurance_df.dtypes
```

```
Age           int64
Sex          object
BMI         float64
Children      int64
Smoker       object
Region       object
Charges     float64
dtype: object
```

```
#Getting shape of the data.
insurance_df.shape
```

```
(1338, 7)
```

Converting the categorical values into numerical values using label encoder and checking if there are any null values.

```
#Converting categorical values to numerical values.
label_encoder = preprocessing.LabelEncoder()
insurance_df['Sex'] = label_encoder.fit_transform(insurance_df['Sex'])
insurance_df['Smoker'] = label_encoder.fit_transform(insurance_df['Smoker'])
insurance_df['Region'] = label_encoder.fit_transform(insurance_df['Region'])

insurance_df.head()
```

	Age	Sex	BMI	Children	Smoker	Region	Charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

```
[ ] #Checking if there is any null values.
    insurance_df.isna().any()
```

```
Age         False
Sex         False
BMI         False
Children    False
Smoker      False
Region      False
Charges     False
dtype: bool
```

Summary of the data.

By describing the dataset, we can know that if the dataset has outliers in it or not. We can see that there are some outliers and the data isn't following normal distribution, so for this reason we will be using Inter Quartile Range(IQR).

```
[ ] #Describing the data which is basically summary of the data.
    insurance_df.describe()
```

	Age	Sex	BMI	Children	Smoker	Region	Charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.505232	30.663397	1.094918	0.204783	1.515695	13270.422265
std	14.049960	0.500160	6.098187	1.205493	0.403694	1.104885	12110.011237
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1121.873900
25%	27.000000	0.000000	26.296250	0.000000	0.000000	1.000000	4740.287150
50%	39.000000	1.000000	30.400000	1.000000	0.000000	2.000000	9382.033000
75%	51.000000	1.000000	34.693750	2.000000	0.000000	2.000000	16639.912515
max	64.000000	1.000000	53.130000	5.000000	1.000000	3.000000	63770.428010

Using Inter Quartile Range to remove outliers.

```
[ ] #Using Inter Quartile range for removing outliers where the First Quartile is 25% and the Third Quartile is 75%.
    def find_outliers_IQR(insurance_df):
       Quartile1=insurance_df.quantile(0.25)
       Quartile3=insurance_df.quantile(0.75)
       IQR=Quartile3-Quartile1
       ins_outliers = insurance_df[((insurance_df<(Quartile1-1.5*IQR)) | (insurance_df>(Quartile3+1.5*IQR)))]
       return ins_outliers
```

```
[ ] #Finding outliers BMI column.
    outliers_bmi = find_outliers_IQR(insurance_df['BMI'])
    print('Number of outliers: ' + str(len(outliers_bmi)))
    print('Maximum outlier value: ' + str(outliers_bmi.max()))
    print('Minimum outlier value: '+ str(outliers_bmi.min()))

    Number of outliers: 9
    Maximum outlier value: 53.13
    Minimum outlier value: 47.41
```

```
[ ] #Finding outliers in Charges column.
    outliers_charge = find_outliers_IQR(insurance_df['Charges'])
    print('Number of outliers: ' + str(len(outliers_charge)))
    print('Maximum outlier value: ' + str(outliers_charge.max()))
    print('Minimum outlier value: '+ str(outliers_charge.min()))

    Number of outliers: 139
    Maximum outlier value: 63770.42801
    Minimum outlier value: 34617.84065
```

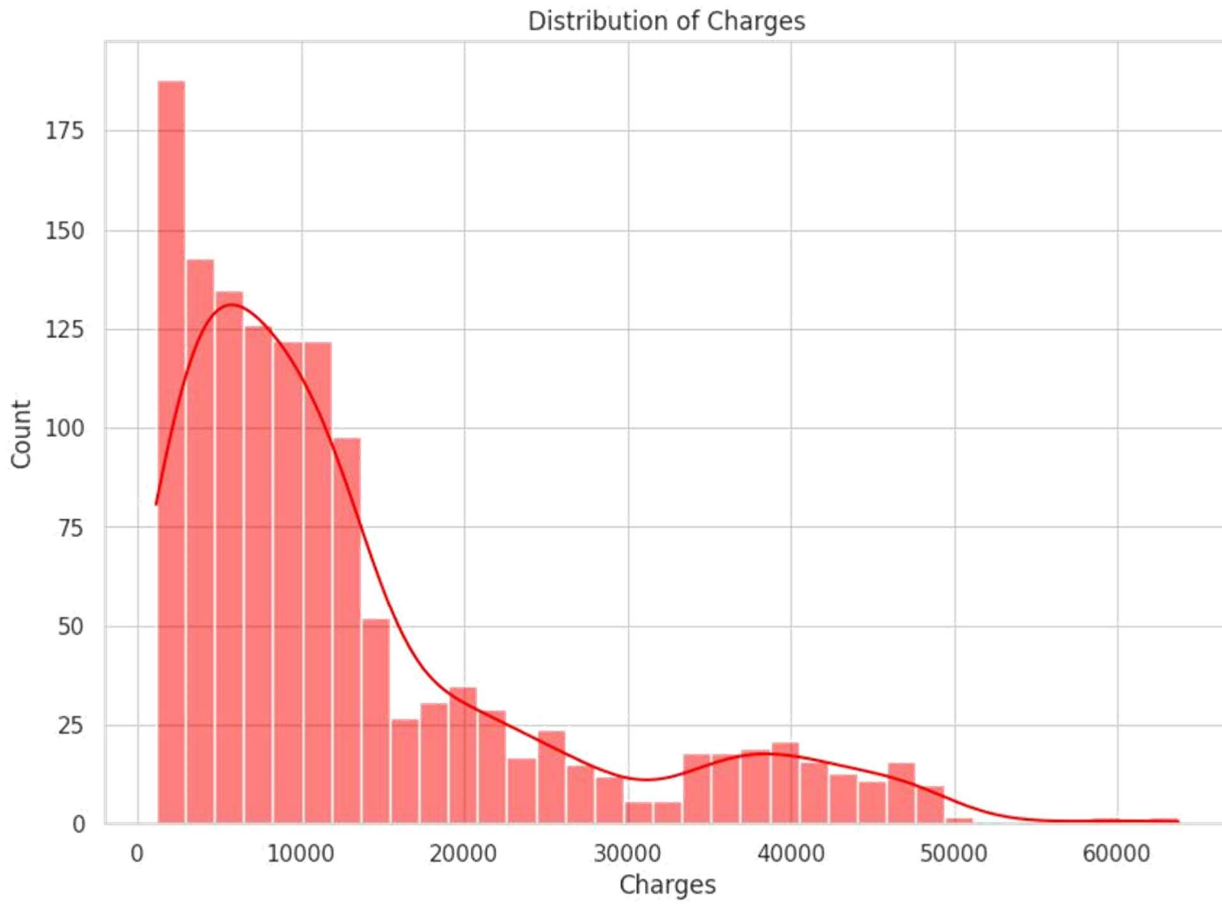
```
[ ] #Dropping the outliers.
    def drop_outliers_IQR(df):
       Quartile1=insurance_df.quantile(0.25)
       Quartile3=insurance_df.quantile(0.75)
       IQR=Quartile3-Quartile1
       not_outliers = insurance_df[~((insurance_df<(Quartile1-1.5*IQR)) | (insurance_df>(Quartile3+1.5*IQR)))]
       return not_outliers
```

```
[ ] #Data after dropping the outliers.
    insurance_df = drop_outliers_IQR(insurance_df)
    insurance_df
```

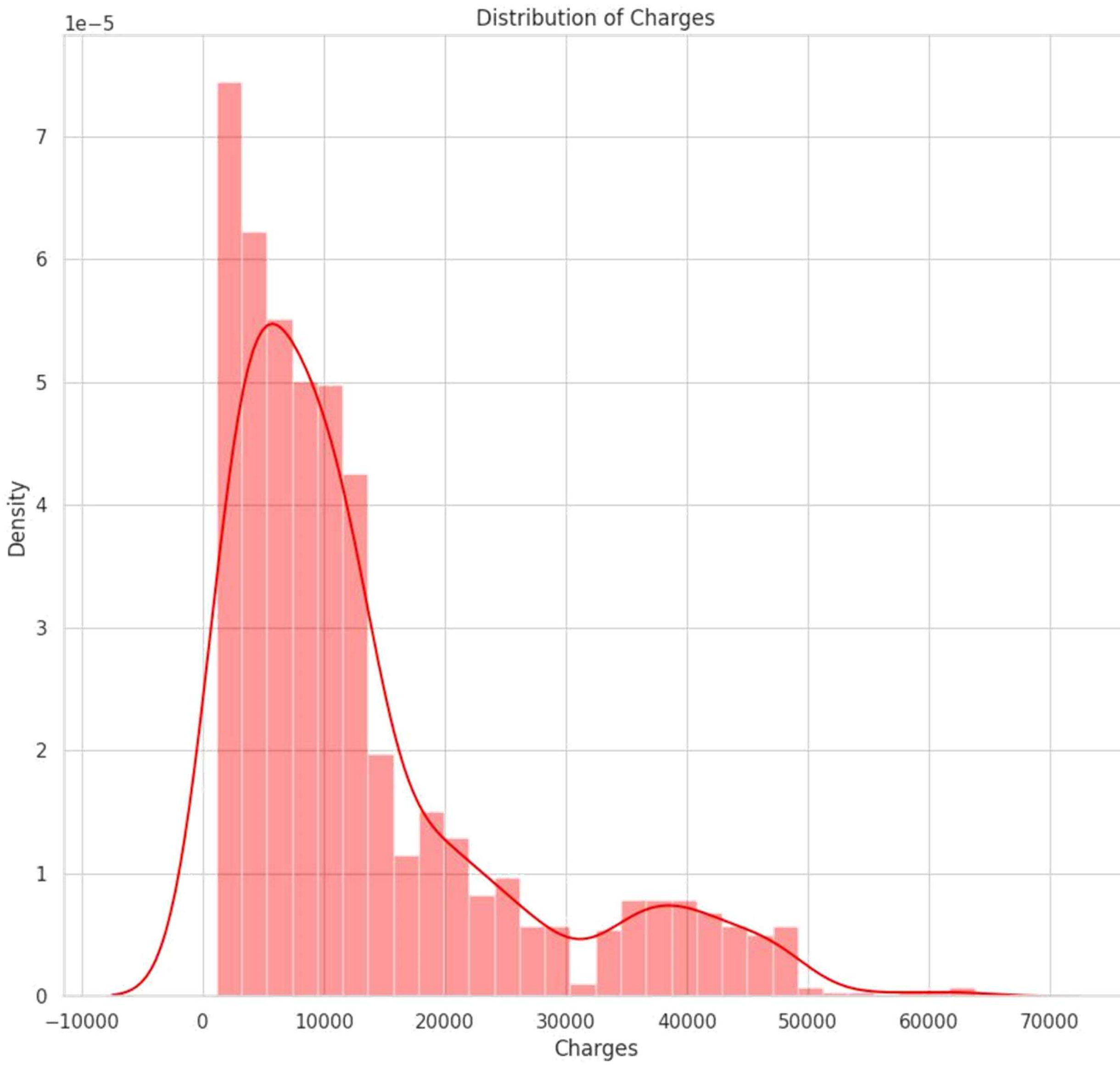
	Age	Sex	BMI	Children	Smoker	Region	Charges
0	19	0	27.900	0	NaN	3	16884.92400
1	18	1	33.770	1	0.0	2	1725.55230
2	28	1	33.000	3	0.0	2	4449.46200
3	33	1	22.705	0	0.0	1	21984.47061
4	32	1	28.880	0	0.0	1	3866.85520
...	...	...	...	...	...	...	...
1333	50	1	30.970	3	0.0	1	10600.54830
1334	18	0	31.920	0	0.0	0	2205.98080
1335	18	0	36.850	0	0.0	2	1629.83350

Data Visualization:

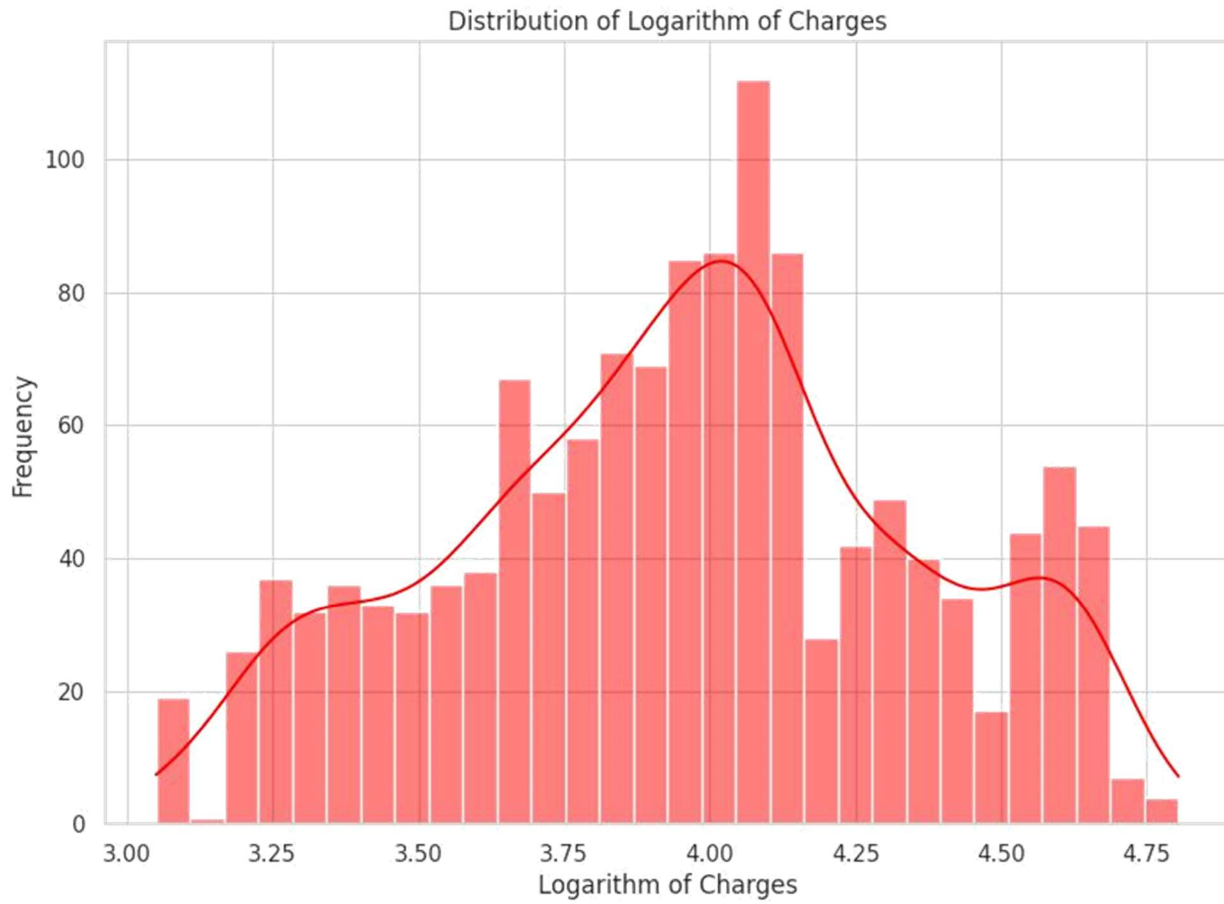
Plotting a graph for distribution of charges.



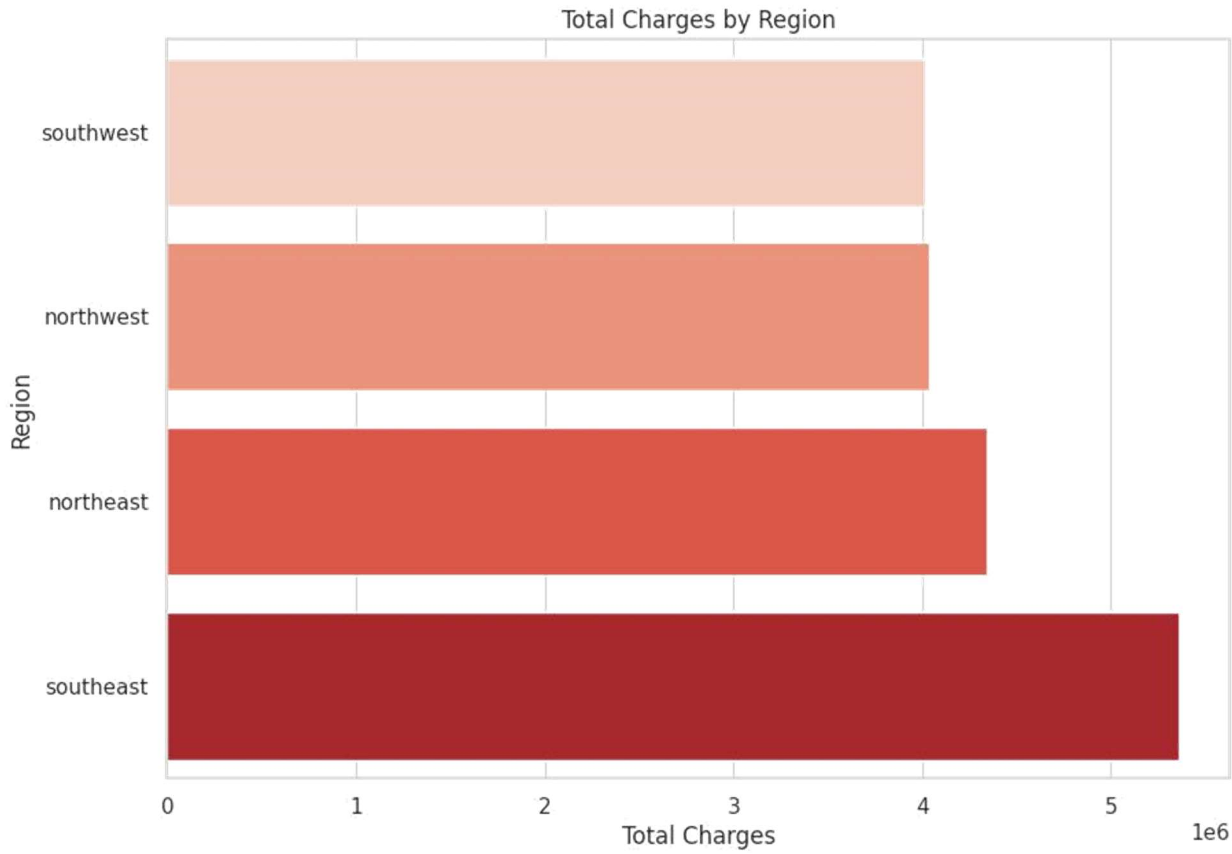
Plotting a smoothen graph for distribution of charges.



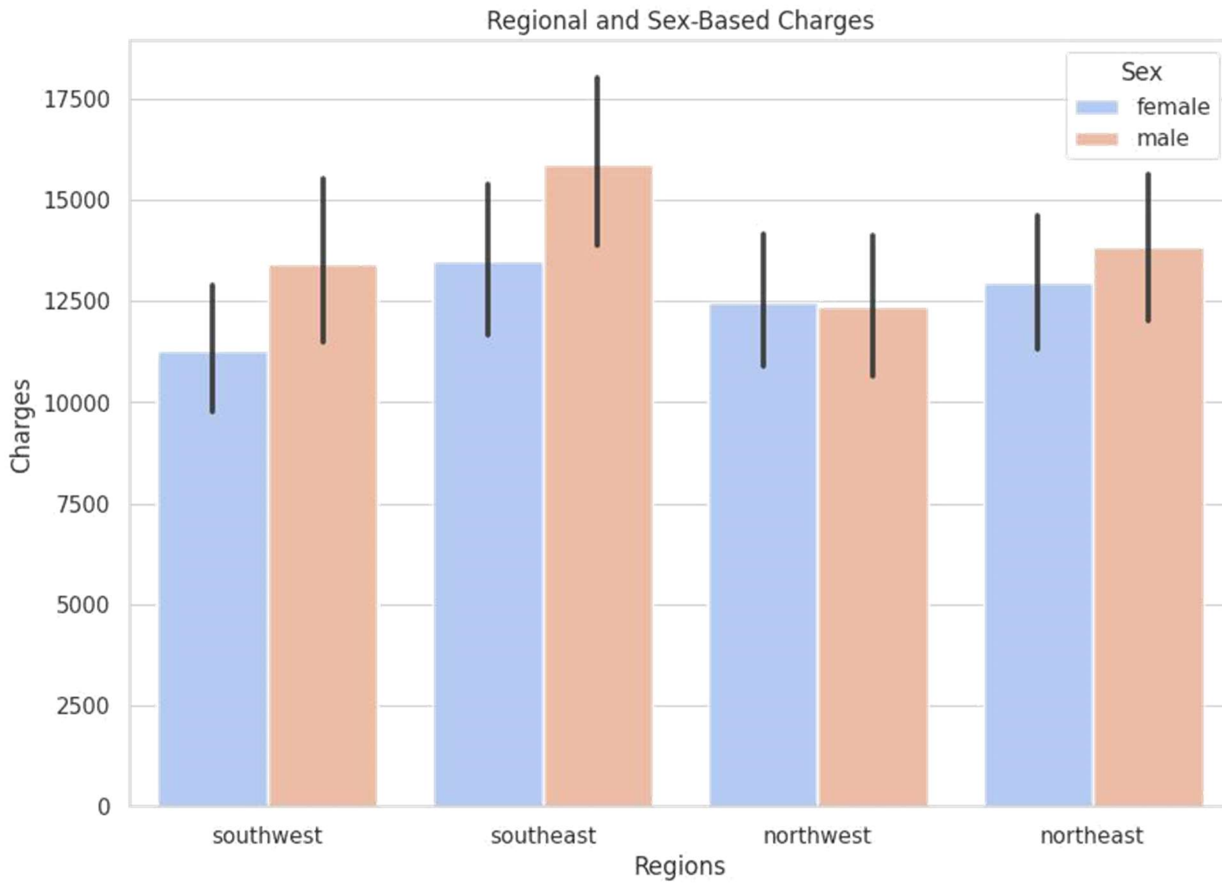
Plotting a logarithmic graph for Distribution of Logarithm of Charges.



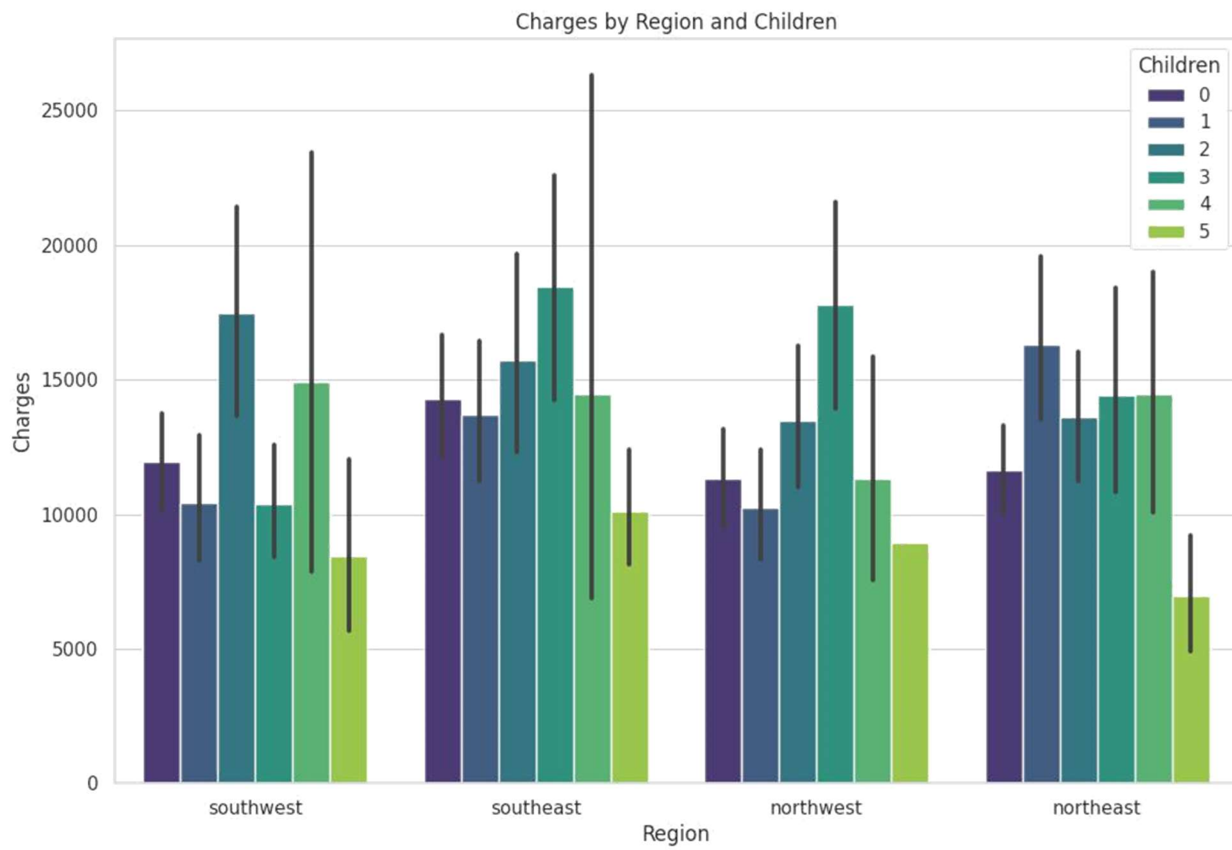
Plotting a graph for total charges by region.



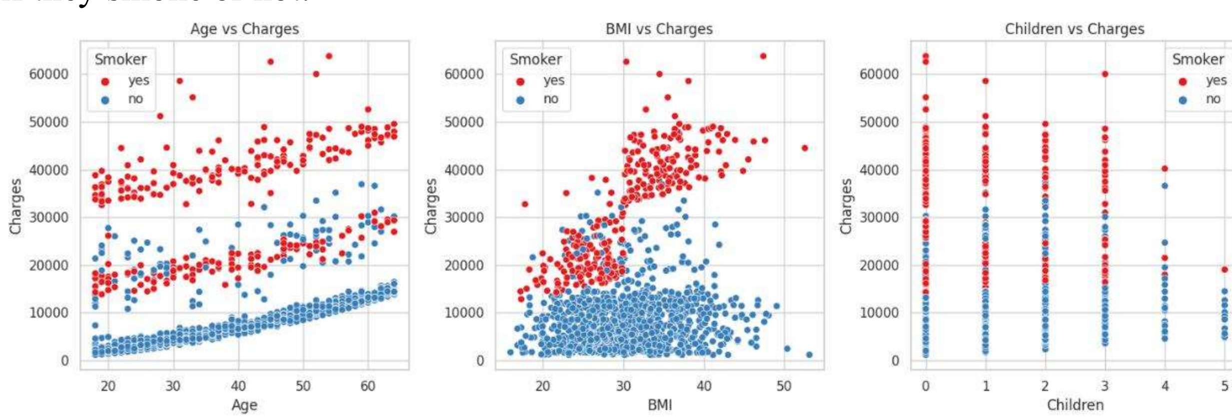
Plotting a graph for comparing different region with the gender.



Plotting a graph for medical insurance charges for families in different regions having different number of children.



Plotting a graph to compare charges with age, bmi and children with respect to if they smoke or not.



<u>Pearson Correlation:</u>

Model Exploration and Selection:

Splitting the data into Training, Testing and Validating.

```
y = in_df['Charges']
X = in_df.drop(['Charges'],axis=1)

X_train, X_1, y_train, y_1 = train_test_split(X, y , test_size=0.4, random_state=9)
X_val, X_test, y_val, y_test = train_test_split(X_1, y_1 , test_size=0.375, random_state=9)
```

```
[29] print(len(y), len(y_train), len(y_val), len(y_test))

1338 802 335 201
```

Performed Linear Regression on the data and the accuracy for the model is 74%.

Linear Regression

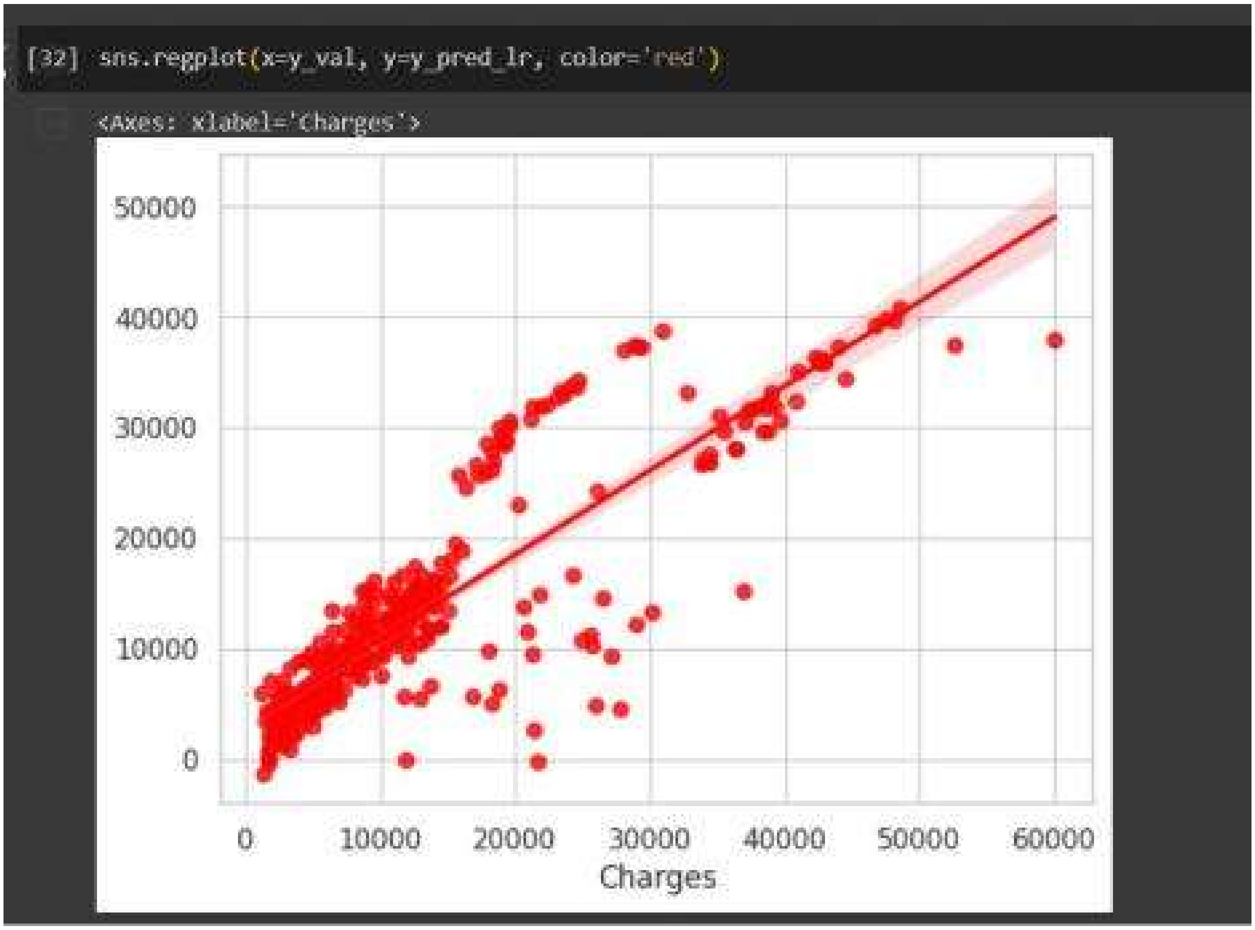
```
[30] linear_regression = LinearRegression()
     linear_regression.fit(X_train, y_train)
     y_pred_lr = linear_regression.predict(X_val)
```

```
[31] score_lr = linear_regression.score(X_val, y_val)
     mse_lr = mean_squared_error(y_val, y_pred_lr)
     r2_lr = r2_score(y_val, y_pred_lr)
     mae_lr = mean_absolute_error(y_val, y_pred_lr)

     print("Accuracy of Model: ","{:}".format(score_lr*100),'%')
     print("MAE: ","{:}".format(mae_lr))
     print("MSE: ","{:}".format(mse_lr))
     print("R-Squared: ","{:}".format(r2_lr*100),'%')
```

```
Accuracy of Model:  74.26809356314793 %
MAE:  4092.938596373554
MSE:  35183914.26199526
R-Squared:  74.26809356314793 %
```

Plotted a graph of the data after performing Linear Regression on the data.



Performance Evaluation and Performance Interpretation:

We tried using the model for predicting the Medical Insurance Price and we got an accuracy of 78%.

```
[33] y_pred_test = linear_regression.predict(X_test)
```

```
[34] score_test = linear_regression.score(X_test, y_test)
     mse_test = mean_squared_error(y_test, y_pred_test)
     r2_test = r2_score(y_test, y_pred_test)
     mae_test = mean_absolute_error(y_test, y_pred_test)

     print("Accuracy: ","{:}".format(score_test*100),'%')
     print("MSE: ","{:}".format(mse_test))
     print("MAE: ","{:}".format(mae_test))
     print("R-Squared: ","{:}".format(r2_test*100),'%')
```

```
Accuracy:  76.15303669682743 %
MSE:  30994849.343204744
MAE:  3800.411549901234
R-Squared:  76.15303669682743 %
```

Summary:

We find that we need more people's data regarding their location, smoking etc. So that we can see the Medical Insurance Cost and then can compare it for predicting with accordance to the factor that it is linked with. We can see that BMI and Smoker are the features are the biggest factors which increase the Medical Insurance Cost the highest as compared to other factor. Charges is correlated with each factor, but it is mostly related with BMI and Smoker. Each factor has an indirect influence on the increase in Medical Insurance Prices.