PREDICTING HEALTH INSURANCE PREMIUMS: A REGRESSION ANALYSIS

By

Shahareyar Hossain

Srijan Rit

Pritam Taldhi

Department of Computer Science

Ramakrishna Mission Vivekananda Educational and Research Institute

Contents

Introduction

In the complex fabric of modern life, health insurance stands as a fundamental thread, safeguarding individuals, families, and businesses against the financial turbulence of medical expenses. Accurately estimating premium rates, the bedrock of this security, remains a complex challenge. This study seeks to unravel this multifaceted phenomenon by employing the potent tool of regression analysis.

Through a rigorous examination of key features, including age, sex, body mass index, family composition, and lifestyle choices, we aim to shed light on the pricing dynamics driving health insurance premiums. Our investigation extends beyond mere quantification; it aspires to illuminate the interconnected web of factors influencing these crucial rates. This data-driven study holds profound implications, empowering individuals with insights for informed healthcare decisions and fostering a comprehensive understanding of cost structures for businesses, when the exponential rise in healthcare costs has become a paramount concern.

Data Description

The dataset comprises 1338 observations and 7 variables, with no missing values.

Three out of the seven variables are categorical, and the remaining four are numerical data.

The dependent variable is "Charges" representing individual medical costs billed by health insurance in United States Dollar.

The explanatory variables or covariates are described below.

- **Age:** Age of the primary beneficiary.
- **Sex:** Gender of the individual.
- **BMI:** Body Mass Index.
- **Children:** Number of dependents.
- **Smoker:** Smoking habit (yes/no).
- **Region:** Beneficiary's residential area in the US (Northeast, Southeast, Southwest, Northwest).

Data Source: https://www.kaggle.com/datasets/mirichoi0218/insurance/data

The summary statistics of the numeric columns are given in table 1.1 given below.

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Table 1.1 Descriptive Statistics of Numerical Columns

Exploratory Data Analysis

A thorough exploration of the dataset to uncover patterns, relationships, and potential outliers utilizing statistical and visualization tools to gain insights into the distribution and characteristics of the variables and identification of potential correlations between independent variables and the dependent variable.

The majority of people belongs to the range 30–50, wherein the above, we have already seen that the minimum age in the data is 18 and the maximum age is 64 as shown in Figure 1.1

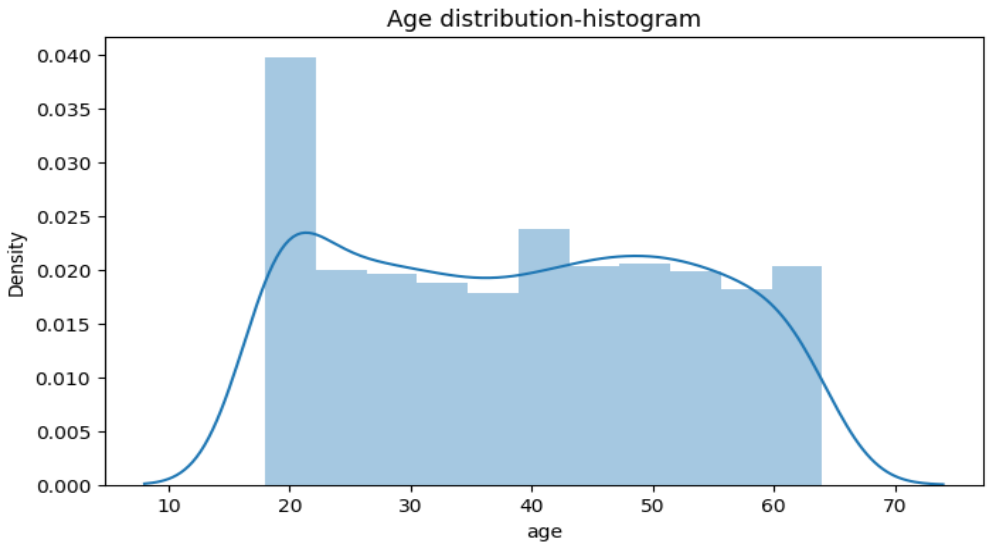

Figure 1.1 Distribution of age of the individuals

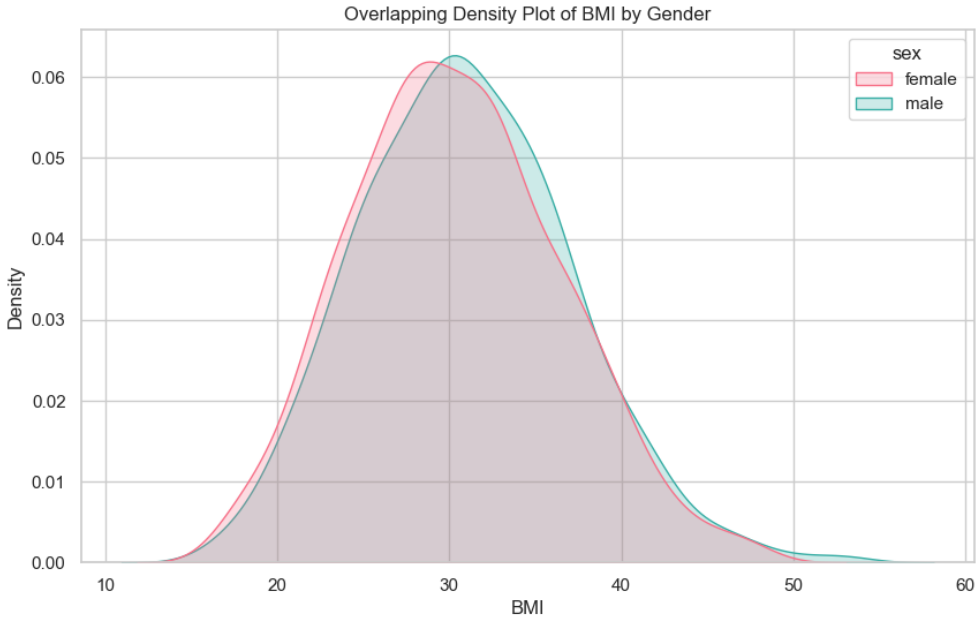

Figure 1.2 Density plot of Body Mass Index for both genders

As, can be seen from Figure 1.2, most persons have Body Mass Index (BMI) in the range 25-35. Males have slightly higher BMI than females. There is a small positive correlation between BMI and charges which is 0.1983 as shown in Figure 1.3

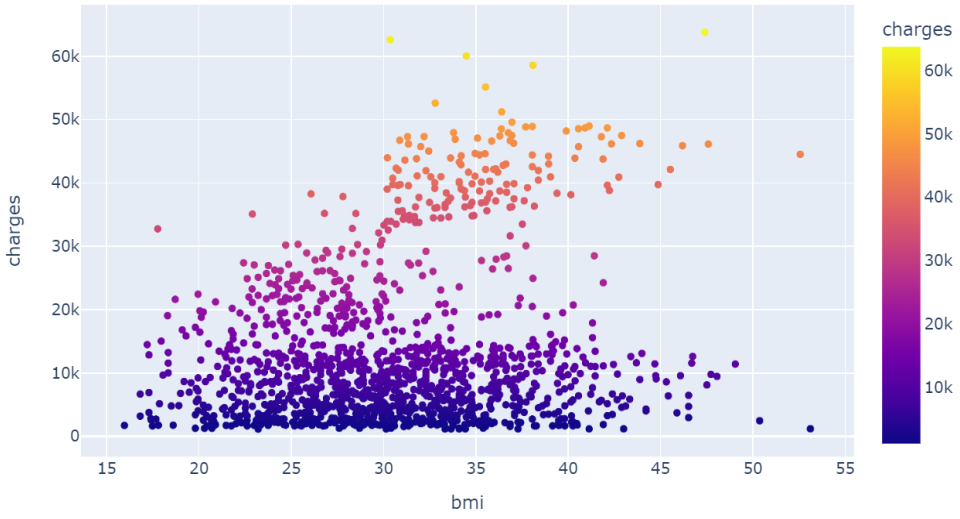

Figure 1.3 Scatterplot of Charges and BMI

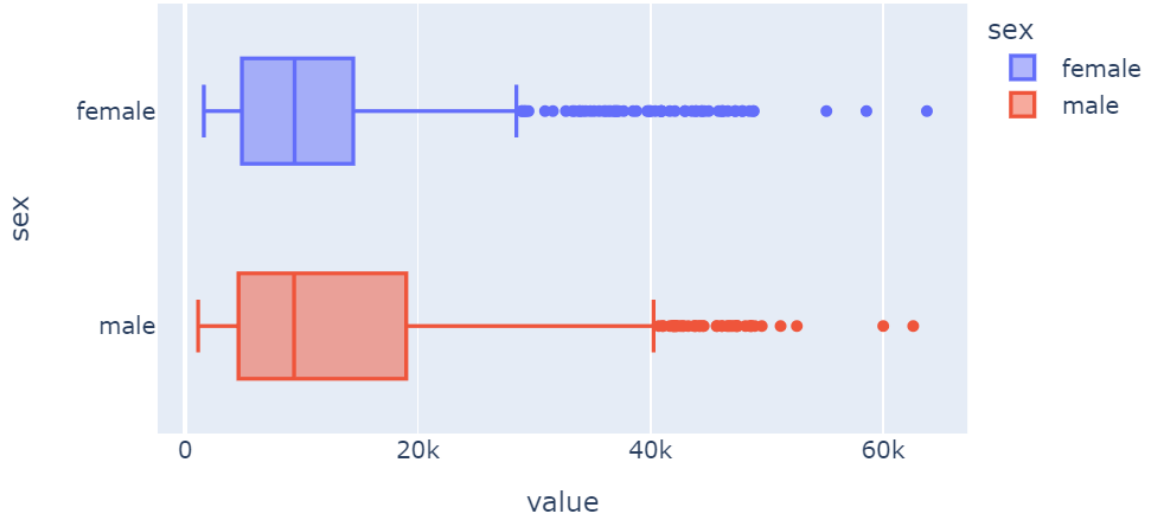

Figure 1.4 Box plot of Insurance charges by sex

Men's charges are slightly higher than women's as can be seen in Figure 1.4, where median charges for females are 9412.963 and that for males are 9369.616

It is visible from Figure 1.5 that medical charges for people who are habitual smokers are higher than for those who are not smoking. Males who smoke have a median increase of 29,099.713 in their charges. Females who smoke had an increase of 21,311.053 in the median charges.

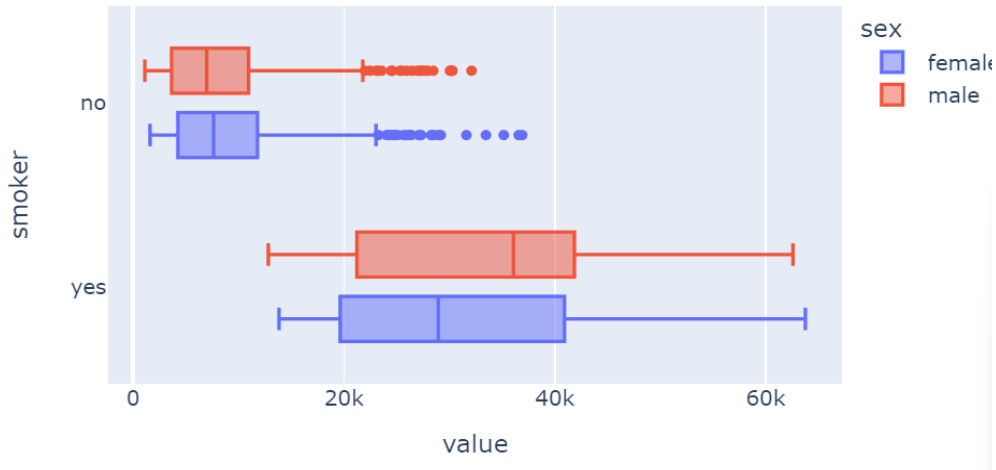

Figure 1.5 Box plot of charges by smoking status and gender

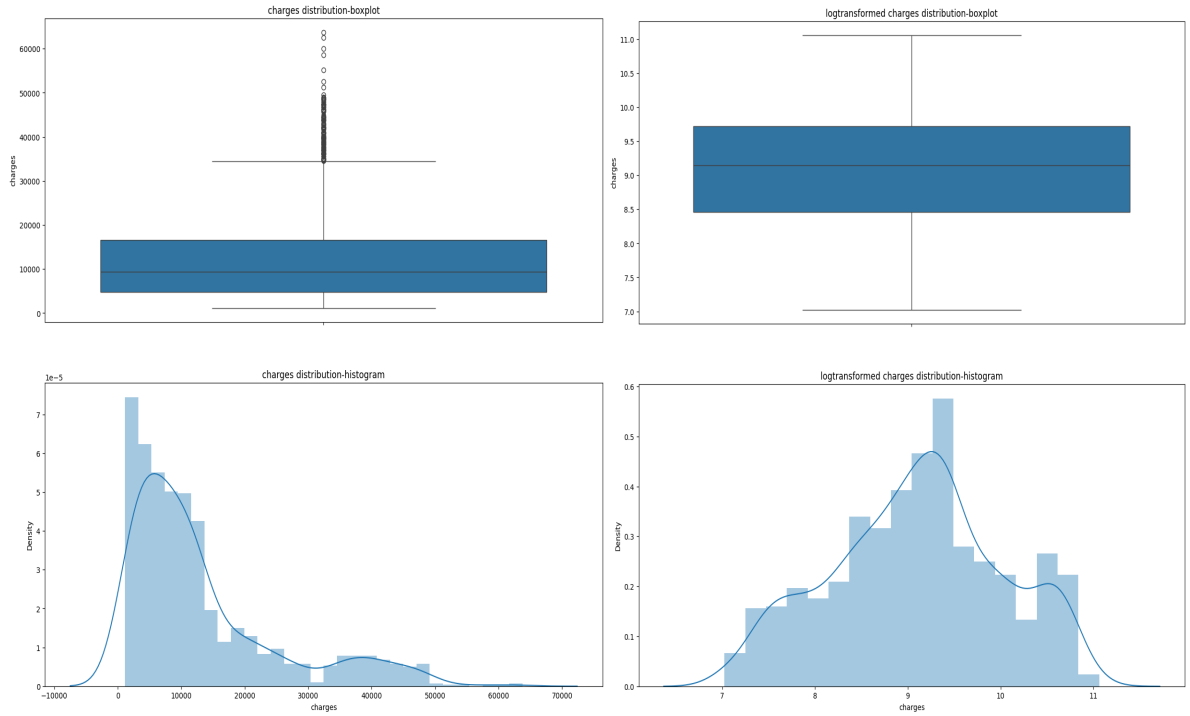

Figure 1.6 Distribution of charges(left) and log transformed charges (right)

The dependent variable Charges is right skewed as can be seen from Figure 1.6, so we need to make some appropriate transformation.

After log transformation, we can see that our dependent variable is not skewed like before.

Methodology

We can see from the matrix in Figure 1.7, we can see there is no multicollinearity in the dataset.

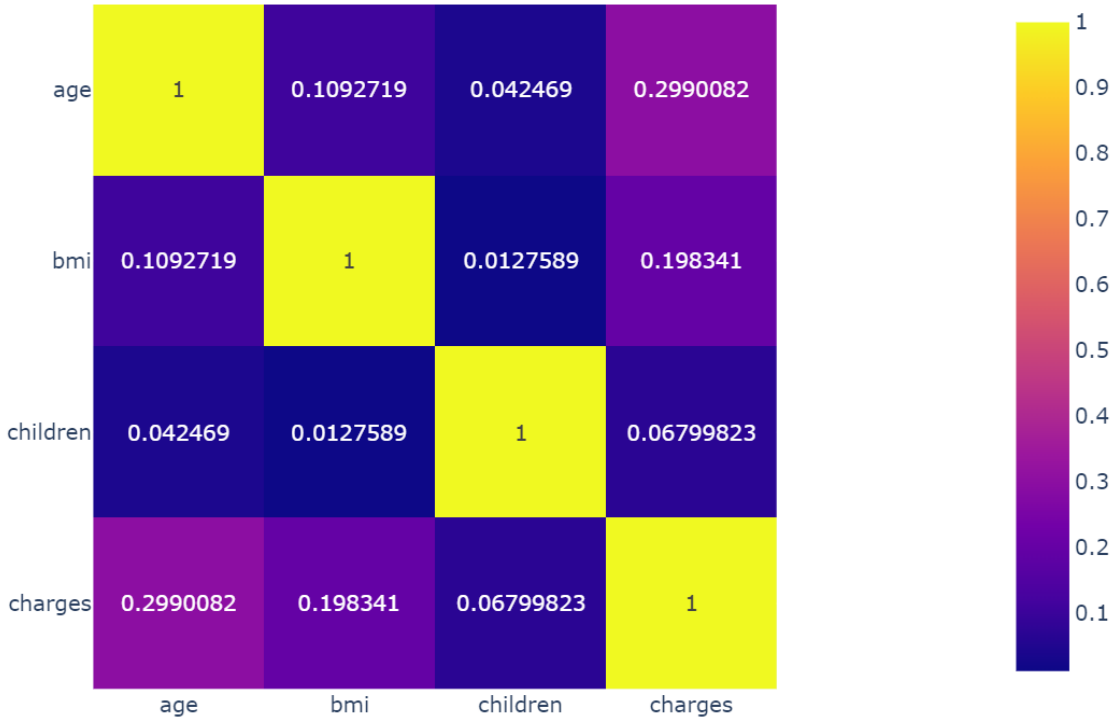

Figure 1.7 Correlation Matrix of Numerical Variables

One Hot Encoding is used to encode the categorical columns.

Then, we split train and test with ratio 80:20

Three assumptions of Ordinary Least Squares Regression i.e, linearity, normality, and independence are verified. For homoscedasticity, we will consider it at the time of modelling.

We have applied four models to the data:

1. Ordinary Least Squares
2. Ridge Regression
3. LASSO Regression
4. Random Forest Regression

The predicted and actual values have an almost similar trend. The residual plot is slightly right skewed. The Residuals vs Predicted represents heteroscedastic behaviours, so after a certain point, there will be increments in the errors as shown in Figure 1.8

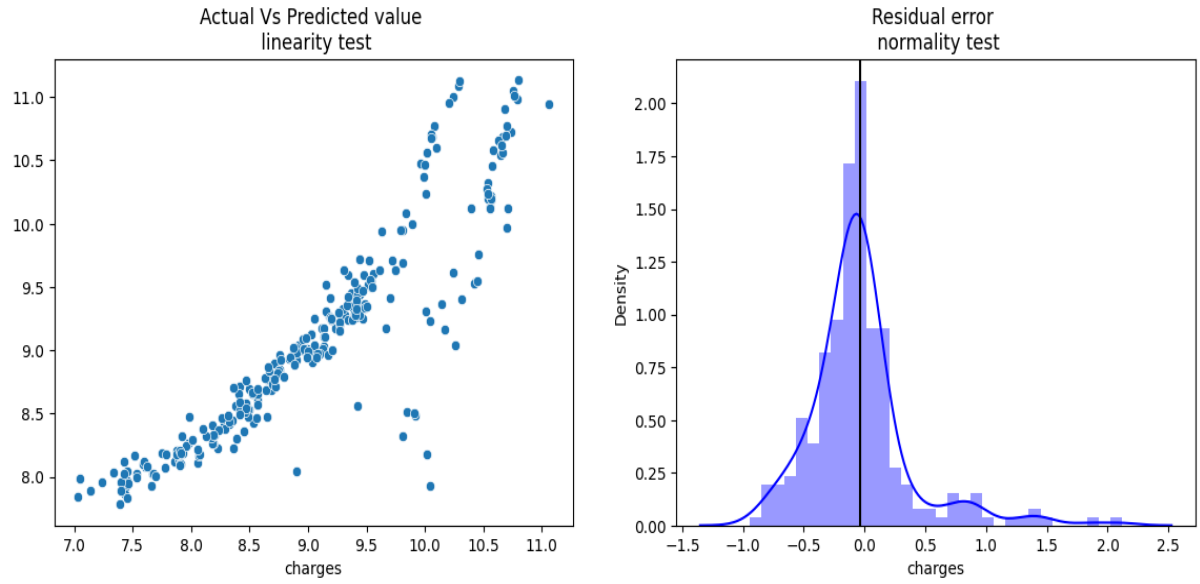

Figure 1.8 Predicted vs Actual Charges (Left), Distribution of Residuals

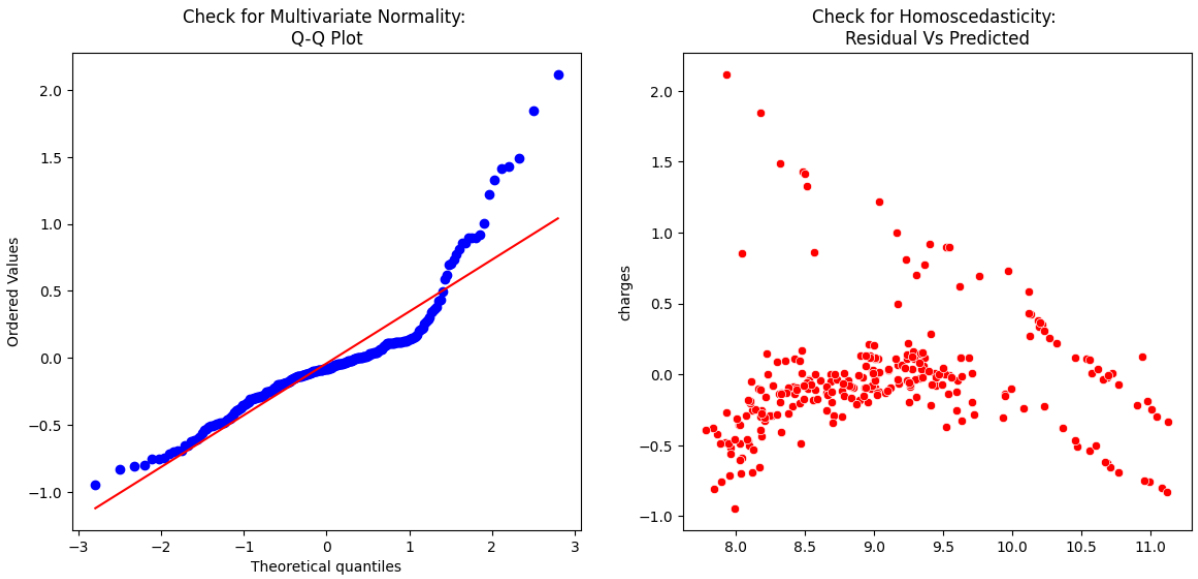

Figure 1.9 QQ Plot of Residuals(Left), Scatter plot of residuals vs charges

The Residuals vs Predicted in Figure 1.9 (Right) represents heteroscedastic behaviours, so after a certain point, there will be increments in the errors.

Hyper parameter for penalty term was taken to be 0.01 in case of Ridge Regression and 0.001 in case of LASSO Regression. For Random Forest Regression, the number of trees were taken to be 100.

Results

	Model	MAE	MSE	RMSE	R2 Score
3	Random Forest Regression	0.196990	0.141101	0.375635	0.843071
0	Linear Regression	0.274477	0.174396	0.417607	0.806042
1	Ridge Regression	0.274485	0.174399	0.417611	0.806038
2	Lasso Regression	0.275249	0.175999	0.419522	0.804259

Table 1.2 Evaluation Metrics for all Models

Random Forest outperforms Linear Regression with a higher R^2 score, as it can capture complex non-linear relationships within the data. It can also be seen that regularization is not much useful in linear regression in this case.

Future Scope of Work

To test for presence of heteroskedasticity, we applied the Breusch-Pagan test. Then, we fit Weighted Least Squares, however the results derived in Table 1.3 need to be investigated further.

WLS Regression Results

Dep. Variable:	charges	R-squared (uncentered):	0.980
Model:	WLS	Adj. R-squared (uncentered):	0.980
Method:	Least Squares	F-statistic:	4286.
Date:	Tue, 28 Nov 2023	Prob (F-statistic):	0.00
Time:	12:00:42	Log-Likelihood:	-4924.0
No. Observations:	1070	AIC:	9872.
Df Residuals:	1058	BIC:	9932.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	3.7104	0.108	34.259	0.000	3.498	3.923
bmi	-4.0889	0.098	-41.738	0.000	-4.281	-3.897
OHE_male	4.0108	0.073	55.190	0.000	3.868	4.153
OHE_1	-1.8057	0.165	-10.942	0.000	-2.130	-1.482
OHE_2	-2.1545	0.156	-13.788	0.000	-2.461	-1.848
OHE_3	3.4127	0.154	22.226	0.000	3.111	3.714
OHE_4	-6.0931	2.101	-2.900	0.004	-10.215	-1.971
OHE_5	4.7675	1.002	4.760	0.000	2.802	6.733
OHE_yes	9.9661	0.327	30.508	0.000	9.325	10.607
OHE_northwest	4.1995	0.086	48.738	0.000	4.030	4.369
OHE_southeast	5.6480	0.426	13.246	0.000	4.811	6.485
OHE_southwest	6.2928	0.151	41.686	0.000	5.997	6.589

Omnibus:	1097.523	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	245561.808
Skew:	4.286	Prob(JB):	0.00
Kurtosis:	76.719	Cond. No.	78.8

Table 1.3 Weighted Least Squares Regression Results

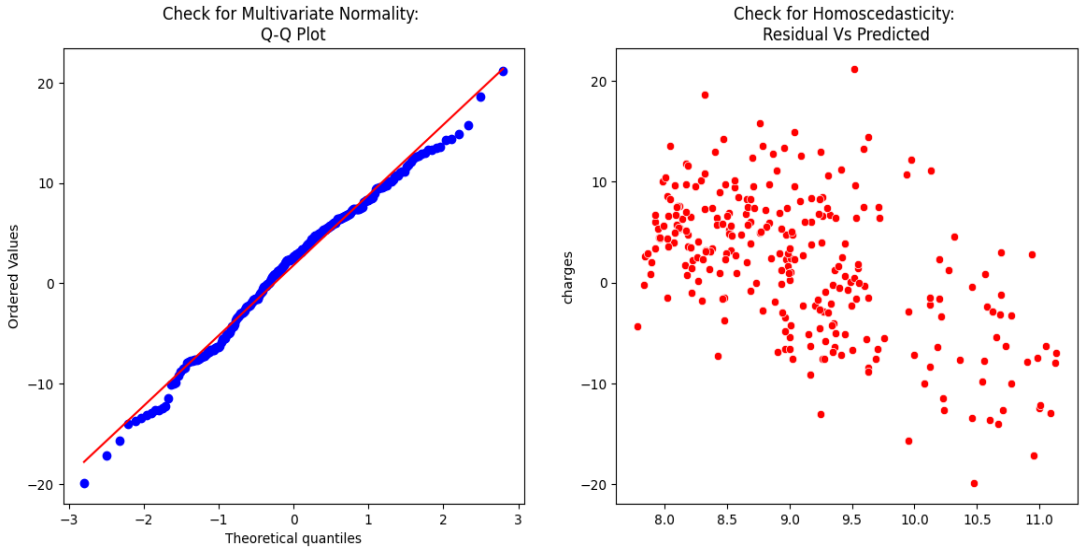

Figure 2.0 QQ Plot of Residuals(Left), Scatter plot of residuals vs charges

It might be useful to include some more covariates, for better explainability.

Some non-linear complex models like Artificial Neural Networks could be implemented, for improved accuracy.

An user friendly interface could be developed for easy access and deployment for utilization by stakeholders.