STATISTICAL ANALYSIS ON PREDICTING MEDICAL INSURANCE COST OF AN INDIVIDUAL

- BY DIYA PATEL

CONTENT

INTRODUCTION

Consumer healthcare cost forecasting has become a critical tactic for enhancing healthcare accountability. The healthcare industry generates a lot of data about patients, illnesses, and diagnoses, but due to incomplete evaluation, this data does not, in addition to the high cost of patient care, give the relevance that it should.

With the help of many health-related factors, like a person's BMI, age, whether they smoke, whether they have children, where they live, etc., this report aims to estimate a person's insurance expenses. With the use of these factors, we may determine whether a person's living situation could influence their long-term health and, consequently, the insurance rates that are available to them**.**

Columns Description :

- Age: Age of primary beneficiary

- Sex: Primary beneficiary’s gender

- BMI: Body mass index (providing an understanding of the body, weights that are relatively high or low relative to height)

- Children: Number of children covered by health insurance / Number of dependents

- Smoker: Smoking (yes, no)

- Region: Beneficiary’s residential area in the US (northeast, southeast, southwest, northwest)

- Charges: Individual medical costs billed by health insurance

CHARGES is the dependant variable to be studied based on the input variables.

UNDERSTANDING THE DATA

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

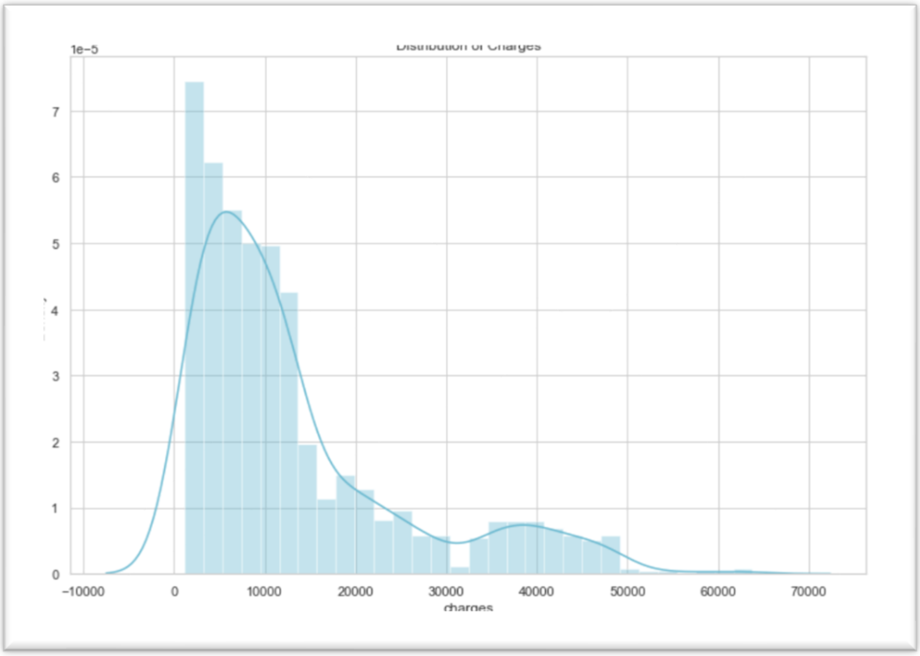
- These are the top 10 observations from my data frame out of 1338 rows.
- The data is unbiased as we have taken the ages from 18-64.
- No null values are present in the dataset. Dataset is clean.
- Now let's do EDA with some cool graphs!

```
df.describe()
```

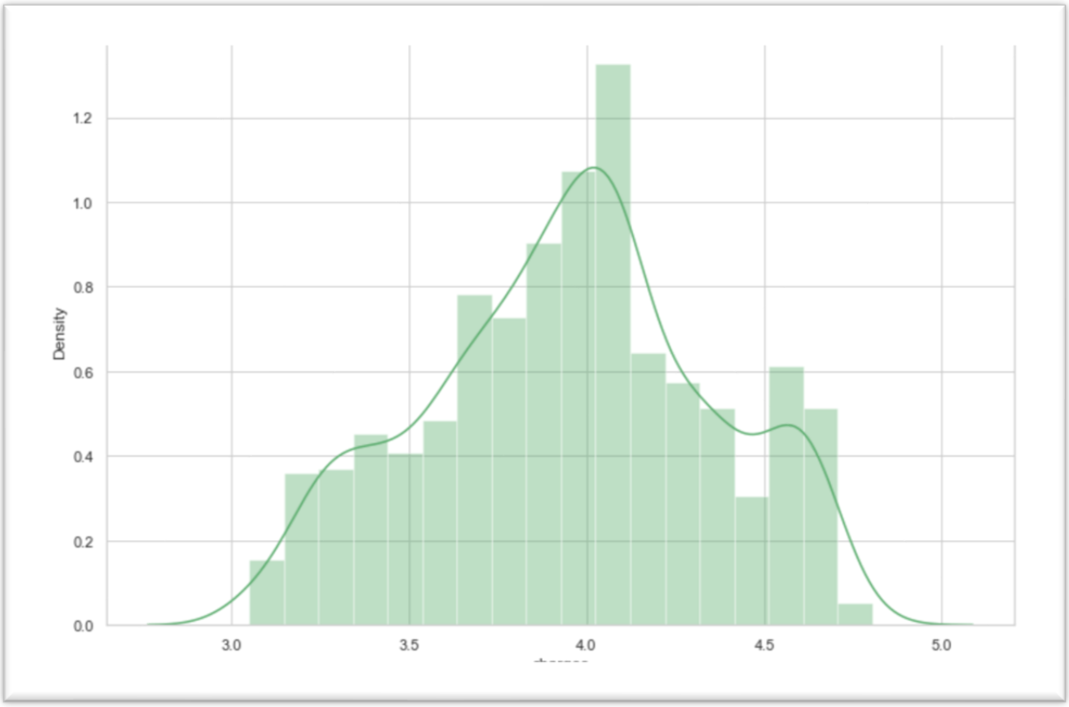
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000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

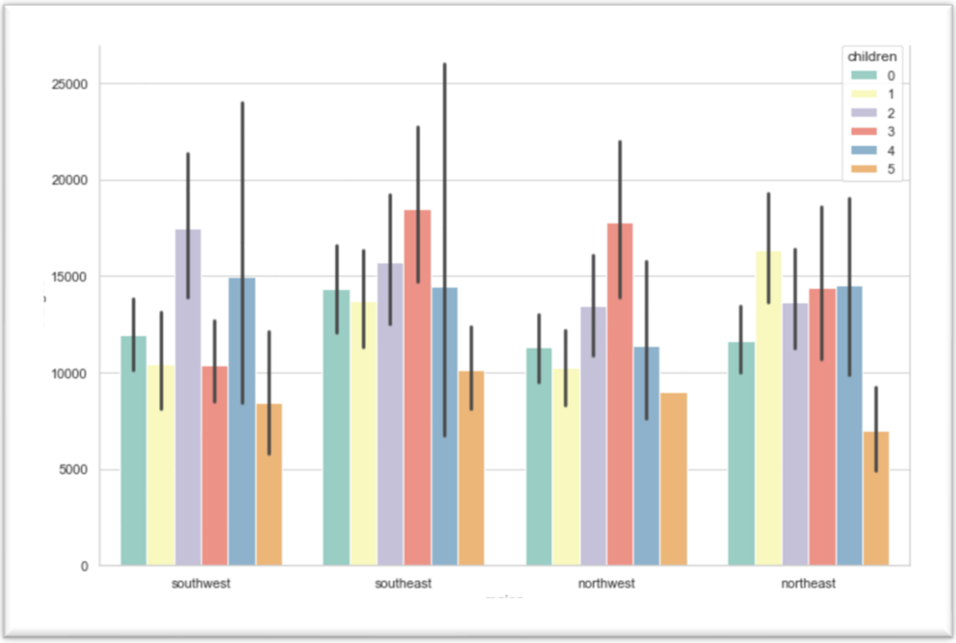
- These are the key parameters of our dataset like mean, interquartile range, min, max, etc using df.describe( ).

HISTOGRAMS & BOX PLOT

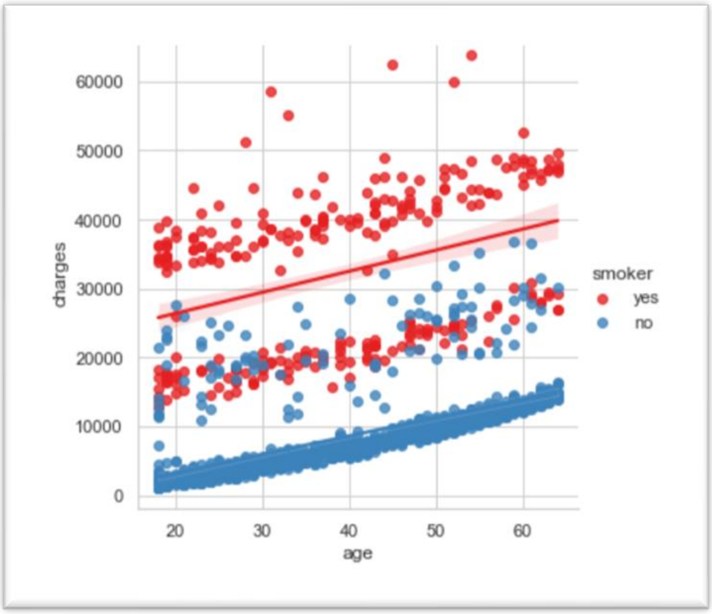


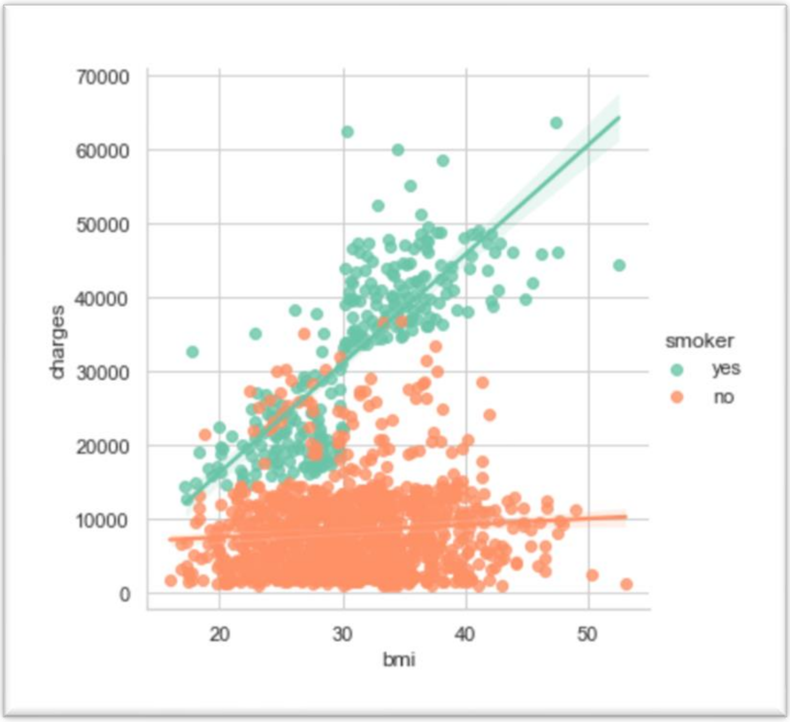
This distribution is right-skewed. To make it closer to normal we can apply natural log

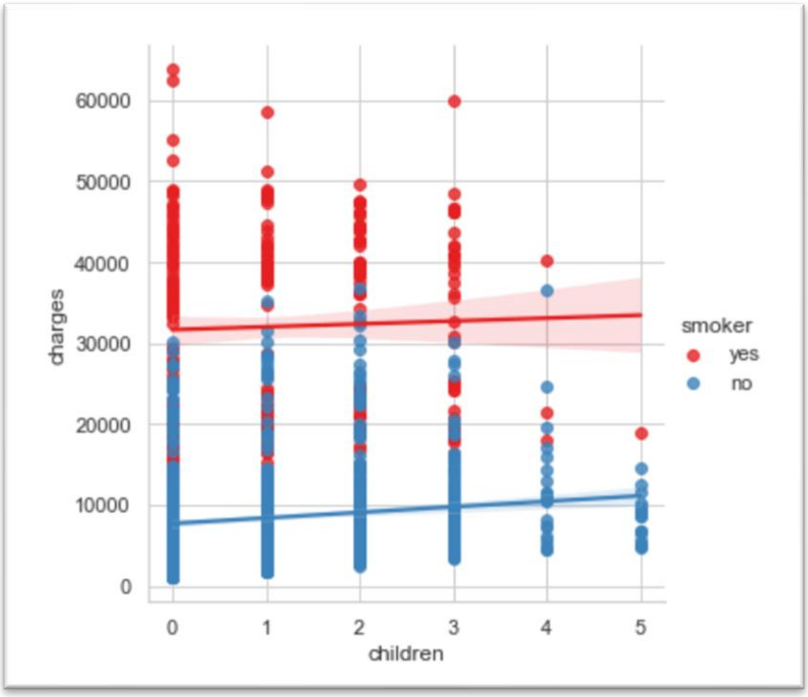




The plot show that the Southeast continues to have the greatest smoking-related charges, while the Northeast has the lowest rates. Although persons in the Northeast have higher charges overall than those in the Southwest and Northwest overall, people in the Southwest generally smoke more than those in the Northeast. Additionally, the overall expense of healthcare is typically higher for families with children.





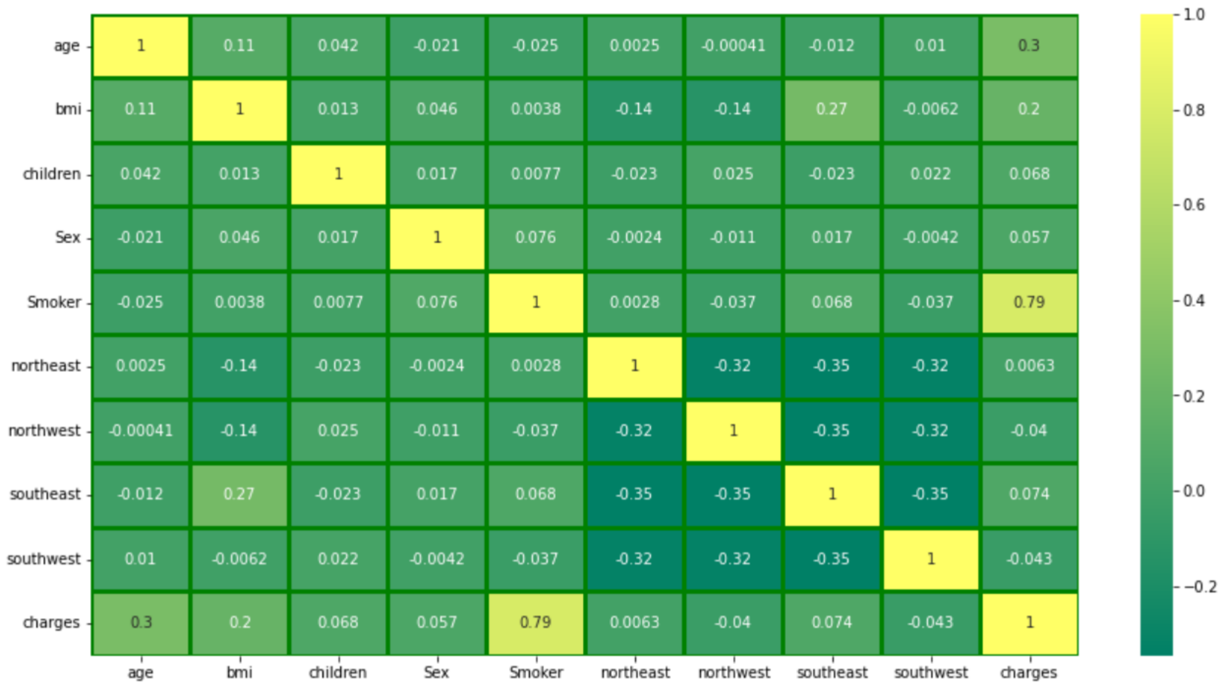


Smoking has the highest impact on medical costs, even though the costs are growing with age, bmi and children. Also, people who have children generally smoke less.

Correlation Scatterplot and Correlation Matrix

A heatmap that displays a 2D correlation matrix between two discrete dimensions and uses coloured cells to represent data from typically a monochromatic scale is called a correlation heatmap. The first dimension's values are displayed as the table's rows, while the second dimension's values are displayed as columns. The percentage of measurements that match the dimensional value is shown in the cell's colour. Because they show differences and variance in the same data and make patterns easy to comprehend, correlation heatmaps are perfect for data analysis. A colorbar helps a correlation heatmap, like a conventional heatmap, by making the data more legible and understandable.

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$



With the help of heat map, we can see that smoking has the highest correlation.

SPLITTING THE DATA FOR TRAINING & TESTING

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2)

X_train.shape,y_train.shape

((1070, 6), (1070,))

X_test.shape,y_test.shape

((268, 6), (268,))
```

- **Data is split in the 80:20 ratio for training & testing where 80% is the training size & remaining 20% is the test size.**

- **As you can see from the code, out of 1338 rows, 1070 has been occupied for training & the remaining 268 rows have been occupied for testing.**

- **For future analysis we will be using training data only.**

- **In the end, we will check the accuracy of our model.**

PERFORMING REGRESSION

X variables in the test:-

- AGE
- SEX
- BMI
- REGION
- SMOKER
- CHILDREN

Y variable in the test:-

- CHARGES

Some of the variables in which are independent can bring the accuracy score of the model down, hence with the help of P-value, we will remove such variables to enhance the predicting efficiency of our model.

We will be doing 2 trials for our model to get a better accuracy and to make a better model!

Here, X variables are the independent variables and Y (charges) is the dependent variable.

Let us proceed with our first trial!

REGRESSION MODEL TRIAL 1:-

Fitting multiple Linear Regression

```
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     500.8
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        22:59:57   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1329   BIC:                         2.716e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.002e+04    781.640    -12.820      0.000   -1.16e+04   -8487.055
age          256.8564     11.899     21.587      0.000     233.514     280.199
bmi          339.1935     28.599     11.860      0.000     283.088     395.298
children     475.5005    137.804      3.451      0.001     205.163     745.838
Sex         -131.3144    332.945     -0.394      0.693    -784.470     521.842
Smoker      2.385e+04    413.153     57.723      0.000      2.3e+04    2.47e+04
northeast  -1918.1003    333.386     -5.753      0.000   -2572.121   -1264.080
northwest  -2271.0642    333.477     -6.810      0.000   -2925.263   -1616.865
southeast  -2953.1224    384.752     -7.675      0.000   -3707.910   -2198.335
southwest  -2878.1513    350.871     -8.203      0.000   -3566.473   -2189.830
==============================================================================
Omnibus:                      300.366   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.887
Skew:                           1.211   Prob(JB):                    7.86e-157
Kurtosis:                       5.651   Cond. No.                     2.39e+17
==============================================================================
```

The result we got is good enough, but we can try to improve it a bit by reducing unimportant features later as we got the value of R-square as 75%

The multiple linear equation for the given data would be:

charges = -10020 + 256.8564 * age + 339.1935 * bmi + 475.5005 * children - 131.3144 * Sex + 23850 * Smoker - 1918.1003 * northeast - 2271.0642 * northwest - 2953.1224 * southeast - 2878.1513 * southwest

The link between the independent variables (age, bmi, children, Sex, Smoker, northeast, northwest, southeast, and southwest) and the dependent variable (charges) in the multiple regression model is shown by this equation. The coefficients, while leaving all other independent variables constant, represent the estimated change in the dependent variable for a one-unit change in each independent variable. Keeping all other factors fixed, the predicted rise in costs, for a one-unit increase in age, would be 256.8564.

Since there are many variables in one equation, we perform ‘P-TEST’ and reduce the dimension for this equation.

Our model is subjected to anova to determine the p-value for each variable. We keep the variables whose P value is less than 0.05 since they are reliable at predicting the outcome. We eliminate the remaining variables from the equation.

Anova for model 1:

	sum_sq	df	F	PR(>F)
age	1.712447e+10	1.0	465.983684	7.783217e-89
Sex	5.716429e+06	1.0	0.155553	6.933475e-01
bmi	5.169225e+09	1.0	140.662697	6.498194e-31
children	4.375466e+08	1.0	11.906327	5.769682e-04
Smoker	1.224468e+11	1.0	3331.968045	0.000000e+00
northeast	1.216449e+09	1.0	33.101478	1.084431e-08
northwest	1.704405e+09	1.0	46.379531	1.471672e-11
southeast	2.164947e+09	1.0	58.911593	3.177130e-14
southwest	2.472742e+09	1.0	67.287165	5.483105e-16
Residual	4.883953e+10	1329.0	NaN	NaN

The 'Sex' variable has a p-value greater than 0.05, which means that it is not statistically significant at the 5% significance level. All other variables have p-values less than 0.05, which means they are statistically significant at the 5% significance level.

Model 2 after removing insignificant values:-

X variables in the test:-

- AGE
- ~~SEX~~
- BMI
- REGION
- SMOKER
- CHILDREN

Y variable in the test:-

- CHARGES

Removing variables having P-value greater than 0.05 significance level will help us achieve a better model.

Regression model trial 2:-

Summary of the new model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.750
Method:                 Least Squares   F-statistic:                     572.7
Date:                Thu, 20 Apr 2023   Prob (F-statistic):               0.00
Time:                        00:38:59   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1330   BIC:                         2.715e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.006e+04    774.464    -12.991      0.000   -1.16e+04   -8542.095
age          256.9736     11.891     21.610      0.000     233.646     280.301
bmi          338.6646     28.559     11.858      0.000     282.639     394.690
children     474.5665    137.740      3.445      0.001     204.355     744.778
Smoker      2.384e+04    411.856     57.875      0.000      2.3e+04    2.46e+04
northeast  -1928.8706    332.160     -5.807      0.000   -2580.486   -1277.255
northwest  -2281.0527    332.409     -6.862      0.000   -2933.155   -1628.950
southeast  -2963.2307    383.776     -7.721      0.000   -3716.102   -2210.359
southwest  -2888.2453    349.825     -8.256      0.000   -3574.515   -2201.976
==============================================================================
Omnibus:                      300.735   Durbin-Watson:                   2.089
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              720.516
Skew:                           1.212   Prob(JB):                    3.48e-157
Kurtosis:                       5.654   Cond. No.                     2.44e+17
==============================================================================
```

New multiple regression is:-

charges = -10060 + 256.97 * age + 338.66 * bmi + 474.57 * children + 23840 * Smoker - 1928.87 * northeast - 2281.05 * northwest - 2963.23 * southeast - 2888.25 * southwest

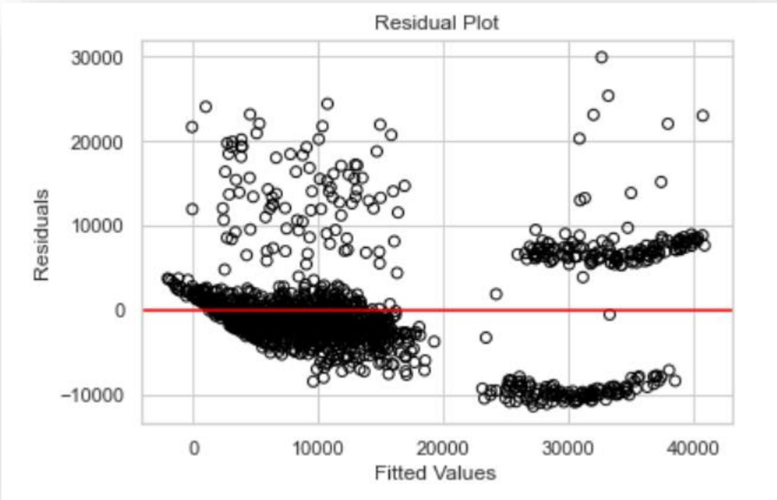
Anova model for trial 2:-

	df	sum_sq	mean_sq	F	PR(>F)
age	1.0	1.753019e+10	1.753019e+10	477.326986	1.148293e-90
bmi	1.0	5.446449e+09	5.446449e+09	148.300547	2.020630e-32
children	1.0	5.715190e+08	5.715190e+08	15.561807	8.401791e-05
Smoker	1.0	1.234476e+11	1.234476e+11	3361.336575	0.000000e+00
northeast	1.0	1.441528e+08	1.441528e+08	3.925115	4.777562e-02
northwest	1.0	8.811518e+07	8.811518e+07	2.399275	1.216294e-01
southeast	1.0	9.328792e+05	9.328792e+05	0.025401	8.733956e-01
southwest	1.0	6.358737e+07	6.358737e+07	1.731411	1.884575e-01
Residual	1330.0	4.884525e+10	3.672575e+07	NaN	NaN

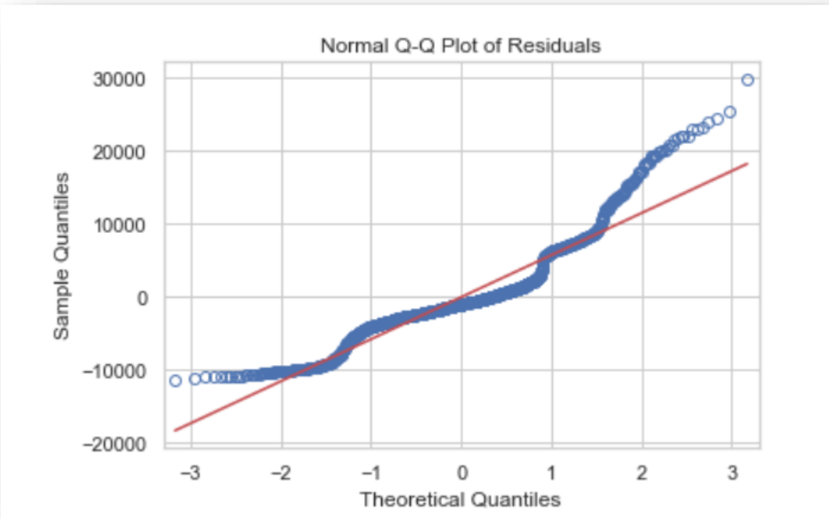
- **As seen, the P value < 0.05 for the above variables indicating that Model-2 holds good for predicting the output.**
- **Hence, the above results convey that all the p values are significant.**
- **From our original model, we have removed the column “Sex”.**
- **Now that our final model is ready with the final Regression Equation, we can do ‘Residual Analysis’**

RESIDUAL ANALYSIS:-

In statistical analysis, a residual plot is a graphical tool used to assess how well a regression model fits the data. The residuals in a regression analysis are the discrepancies between the predicted values and the actual values of the dependent variable.



A graphical tool used in statistical analysis to evaluate the normality of the residuals in a regression model is called a normal probability plot or a normal Q-Q plot of residuals.



Predictions

Each row in the resulting Data Frame, or "result," represents one observation in the dataset, and it has two columns, "Actual" and "Predicted." The Data Frame's top 10 rows are shown using the head() method.

The table makes it simple to compare the actual target values and the predicted values the model produced. By comparing how closely the predicted values match the actual values, this data can be used to assess the model's performance.

```
y_pred = model.predict(X)

# Create a DataFrame with actual and predicted values
results = pd.DataFrame({'Actual': y, 'Predicted': y_pred})

# Print the table
print(results.head(10))
```

```
        Actual     Predicted
0  16884.92400  25217.897406
1   1725.55230   3512.165759
2   4449.46200   6770.262752
3  21984.47061   3827.056827
4   3866.85520   5661.337382
5   3756.62160   3658.778822
6   8240.58960  10595.666738
7   7281.50560   7983.827016
8   6406.41070   8569.251751
9  28923.13692  11827.057193
```

The R-squared score is a statistical indicator that shows what percentage of the variance in the target variable(the dependent variable) in a regression model can be predicted from the independent variables (the features).

An R-score of 0.86 signifies a good model!

```
print(r2_score(y_test,pred))
```

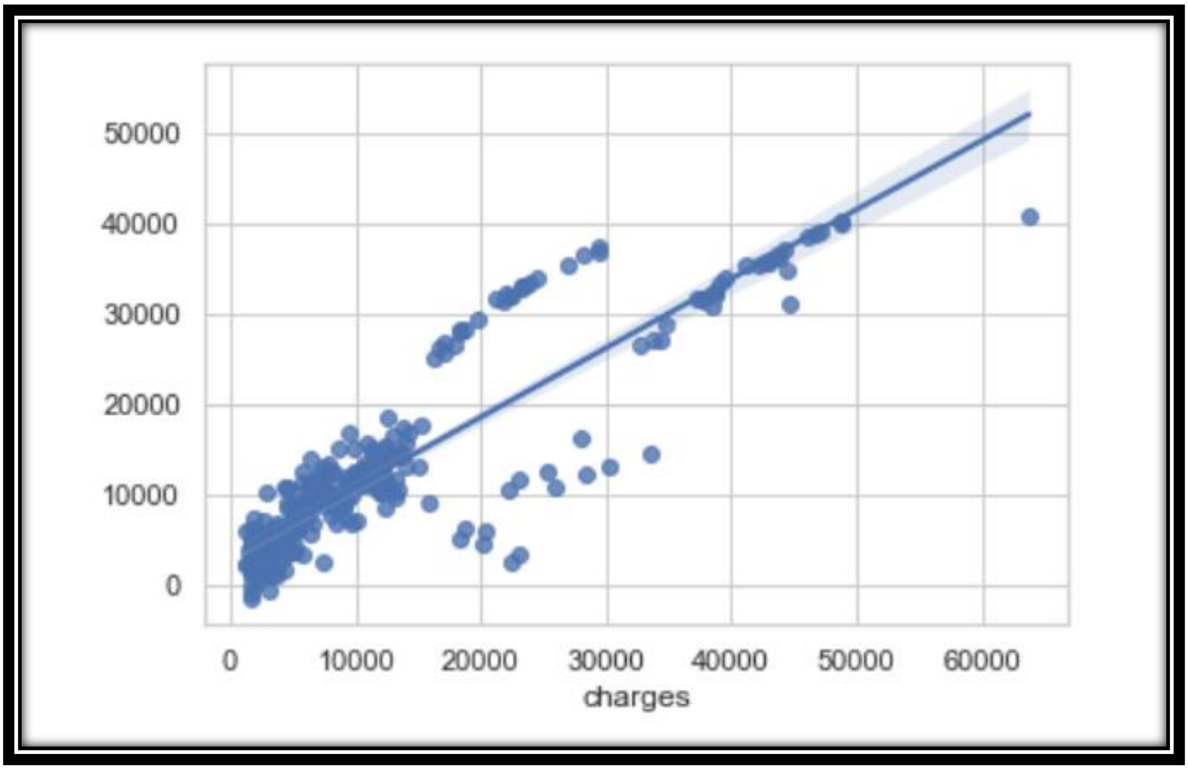
```
0.8673711669559139
```

ACTUAL VS PREDICTED GRAPH

- This is our actual vs predicted graph where X axis contains y_test values & Y axis contains Predicted values.

- As we can see that most of the points are near the line, indicating a good model :)

```
sns.regplot(x=y_test,y=pred)
```



REFERENCES

Medical cost personal datasets insurance forecast by using Linear Regression
https://www.kaggle.com/mirichoi0218/insurance

K Swathi and R Anuradha (2017), Health insurance in India- An overview

Suman Devi and Dr. Vazir Singh Nehra (2015), The problems with health insurance sector in India.12. Shatakshi Chatterjee, Dr. ArunangshuGiri, Dr. S.N. Bandyopadhyay (2018), Health insurance sector in India: A study.

Types of health insurance from reliance general
https://www.reliancegeneral.co.in/Insurance/KnowledgeCenter/Insurance-Reads/Types-Of-Health-Insurance-Covers.Aspx

International journal of creative research thoughts.

THANK YOU!