Capstone Project : Medical Cost Prediction
Final Report

Date Created : 8/1/2021

Contents

1. Problem Statement

This Project is to build a model that could predict the amount an individual needs to pay towards his Health insurance per year based on certain Health parameters.
The aim of this Project is to:

- To determine if there is a relationship between attributes and medical costs.
- To determine if there a significant difference in medical costs between different groups.
- To fit a multiple linear regression to predict costs.

Leveraging customer information is of paramount importance for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below can be crucial in making business decisions.

2. Data Set

Dataset for this Project has been taken from Kaggle – Medical Costs Dataset.

https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv
Columns:

- AGE: Age of primary beneficiary
- SEX: insurance contractor gender, female, male
- BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- CHILDREN: Number of children covered by health insurance / Number of dependents
- SMOKER: Whether Beneficiary is a Smoker
- REGION: Beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- CHARGES: Individual medical costs billed by health insurance.

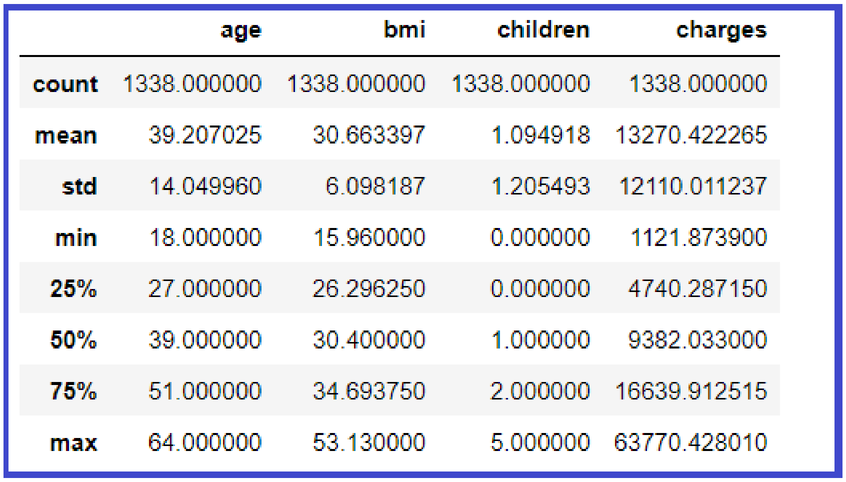
3. Data Wrangling

The Dataset used in the Project was pretty much clean and therefore no additional methods were applied to further clean the data.

4. Data Storytelling – Exploratory Data Analysis

Using the Medical Dataset we tried to answer the following questions:

- Are there more Male beneficaries ?
- Are there more smokers ?
- Which region has maximum , medical cost billed to health insurance ?
- What is age of beneficary?
- Do beneficary having more dependents had more medical cost billed ?



	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

Fig 1: Key Datapoints retrieved from the Dataset

Observations from the above resultset :

i) Average age of the primary beneficiary is 39.2 and maximum age is 64.
ii) Average BMI is 30.66, that is out of normal BMI range, Maximum BMI is 53.13
iii) Average medical costs billed to health insurance is 13270, median is 9382 and maxi mumum is 63770
iv) Median is less than mean in charges , indicating distrubution is postively skewed .
v) Customer on an average has 1 child.
vi) For Age, BMI, children , mean is almost equal to median , suggesting data is normally distrubuted.

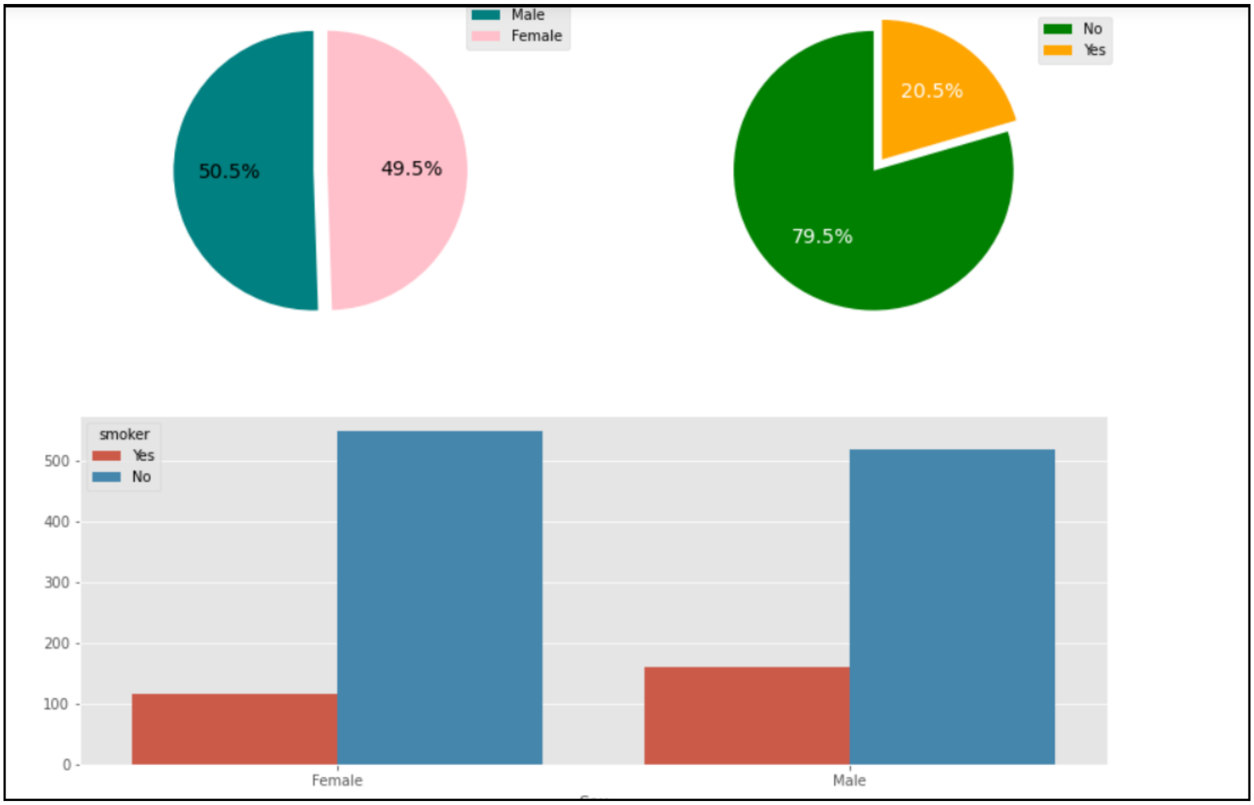


Fig 2: Proportion of Male vs Female smokers and non-smokers

From the Above figure we could conclude that:

- Among overall population, proportion of Male and Female smokers is almost same.
- Among overall population, almost 80% are Non-smokers
- This number further indicates that across Geneder group only 1/5 th of the population is non-smoker.

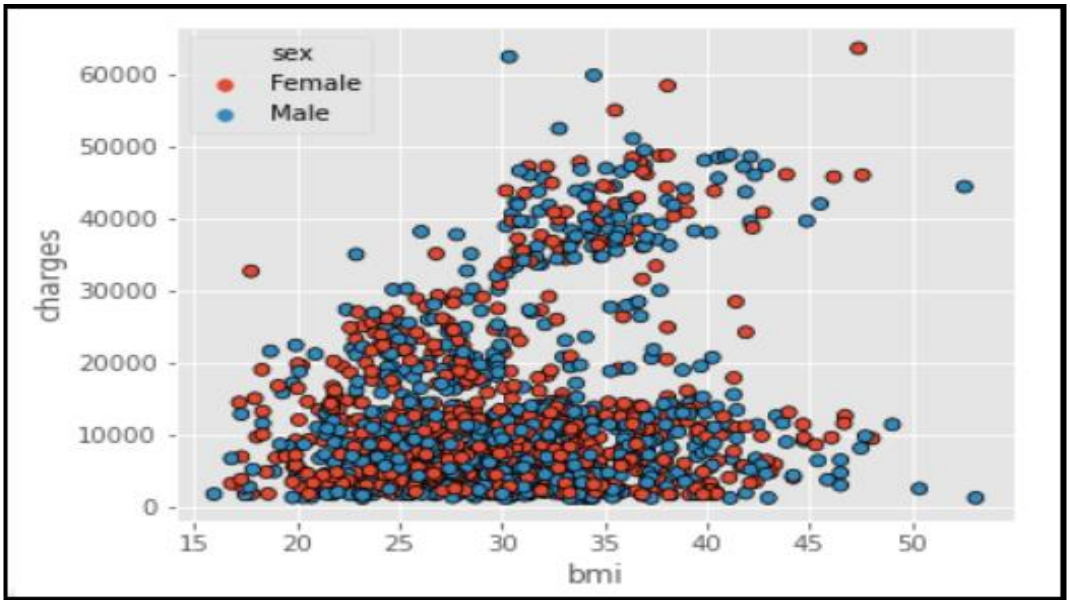


Fig 3: BMI vs Charges categorized for Male and Female

- Scatterplot for BMI vs Charges shows : Increased Medical Charges for higher BMI regardless of Gender.

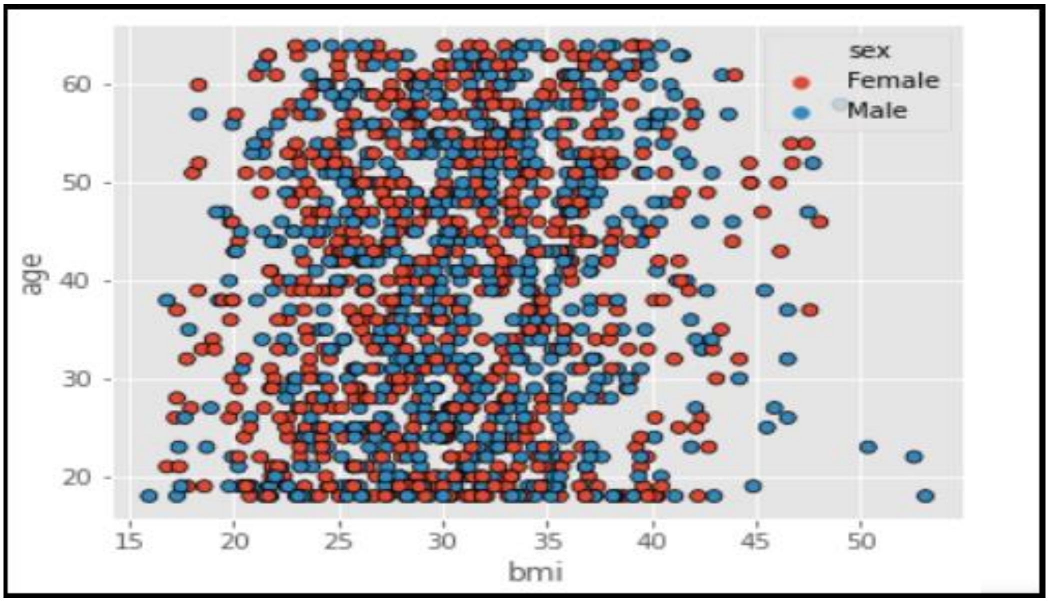
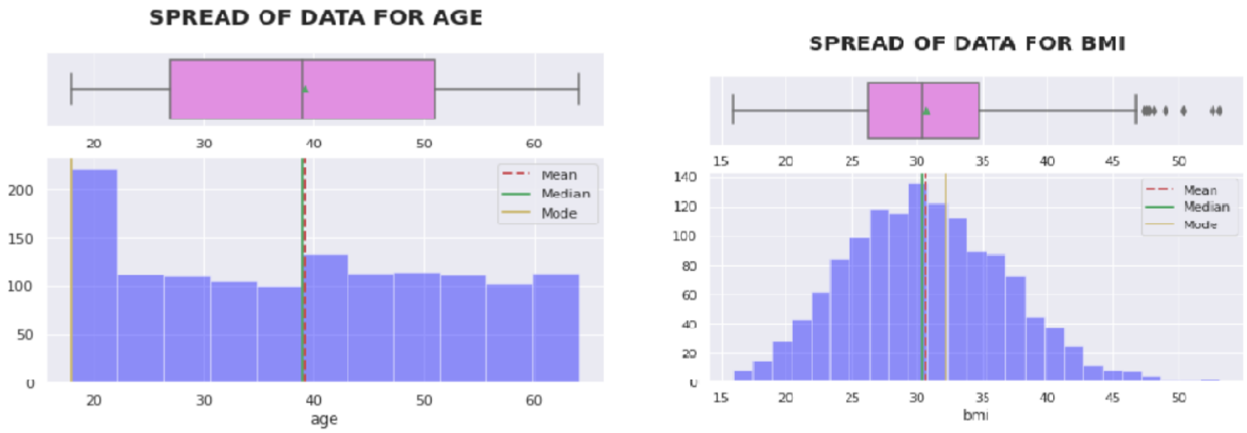


Fig 4: BMI vs Age categorized for Male and Female

- Scatterplot for BMI vs Age doesn't reflect any conclusive relationship.

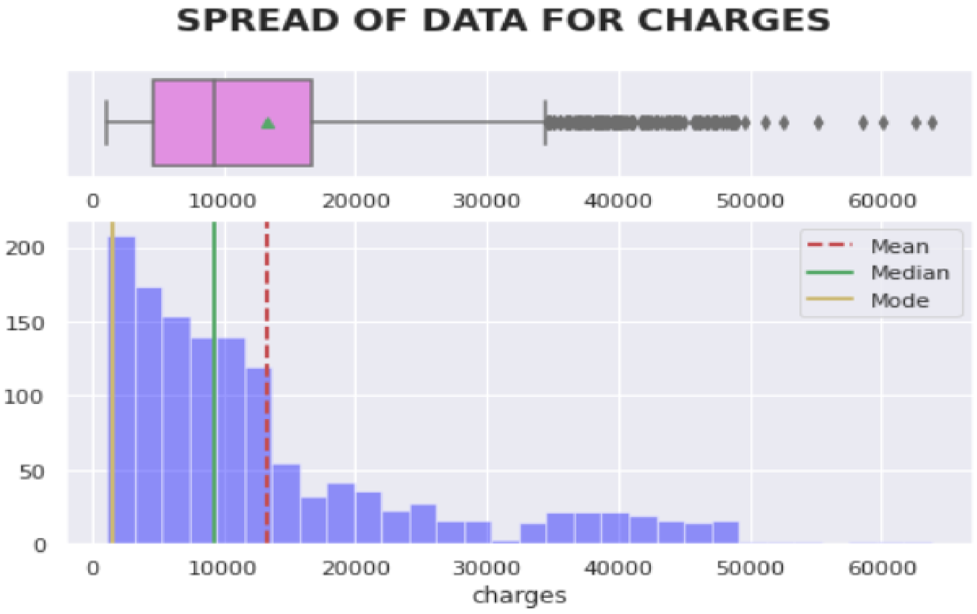


SPREAD OF DATA FOR AGE

SPREAD OF DATA FOR BMI

Fig 5: Spread of Data for Age and BMI

- Age of primary beneficary lies approximately between 20 - 65 . Average Age is aprrox. 40. Majority of customer are in range 18- 20's.
- BMI is normally distrubuted and Average BMI of beneficiary is 30.This BMI is outside the normal range of BMI. There are lot of outliers at upper end



SPREAD OF DATA FOR CHARGES

Fig 6: Spread of Data for Charges

- Charges distrubution is unimodal and is right skewed .Average cost incured to the insurance is appro. 130000 and highest charge is 63770.There are lot of outliers at upper end.

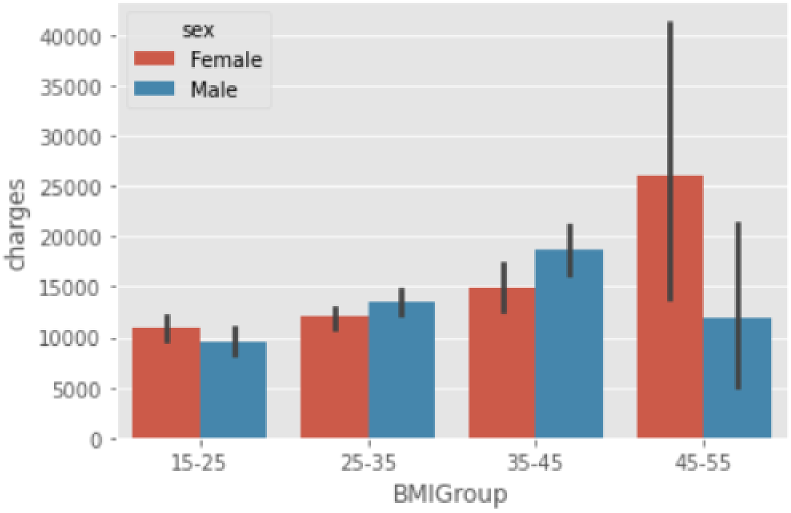


Fig 7: BMI Group vs Charges

- Females with most BMI has incured most charges to the insurance company.
- Noticable difference in BMI between Male and Female population.
- Beneficaries with higher BMI have incurred more cost to insurance.

5. Statistical Analysis

Hypothesis Test performed to get deeper insight into following:

- Whether the medical claims made by the people who smoke is greater than those who don't?

  H0:μ1<=μ2 The average charges of smokers is less than or equal to nonsmokers
  Ha:μ1>μ2 The average charges of smokers is greater than nonsmokers

  Conclusion: Since P value 1.080249501584019e-118 is less than alpha 0.05 , the NULL Hypothesis is Rejected.

  Therefore, Average charges for smokers are less than or equal to that of Non-smokers.

- Whether the BMI of females is different from that of males:

  H0:μ1−μ2=0 There is no difference between the BMI of Male and BMI of female.
  Ha:μ1−μ2<>0 There is difference between the BMI of Male and BMI of female.

  Mean BMI for Males --> 30.943128698224832
  Mean BMI for Females --> 30.377749244713023

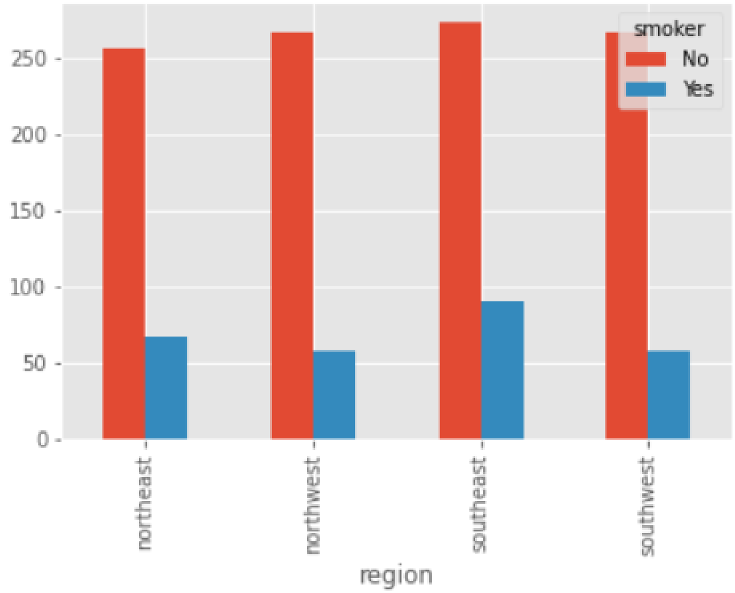
Resultant P-value: 0.08997637178984932

Conclusion:Since P value 0.08997637178984932 is greater than alpha 0.05
Failed to Reject Null Hypothesis that there is difference in BMI of Males and BMI of Females.

- Whether proportion of smokers significantly different across different regions:

H0: Smokers proportions is not significantly different across different regions.
Ha: Smokers proportions is different across different regions.

Since two different categorical variables are compared - smoker and different region- we'll perform a Chi-sq Test.



Using the Chi-square contingency test:

```
Chi-square statistic: 7.343477761407071 , P-value: 0.061719548391 70541
Degree of freedom: 3
Expected frequencies:    [[257.65022422  66.34977578]
                           [258.44544096  66.55455904]
                           [289.45889387  74.54110613]
                           [258.44544096  66.55455904]]
```

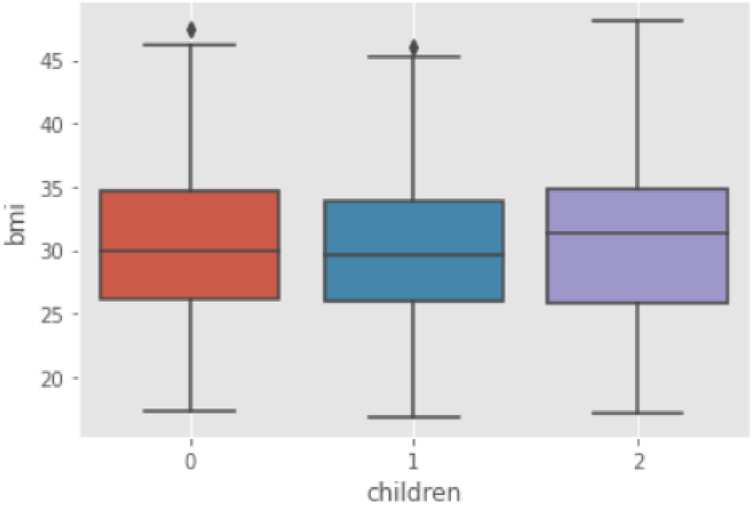
Conclusion: Since P value 0.06 is greater than alpha 0.05

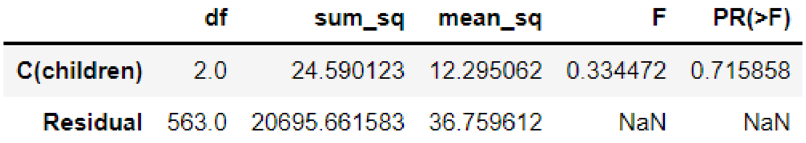
Failed to Reject Null Hypothesis and conclude that Smoker proportions is not significantly different across different regions.

- Whether mean BMI of women with no children, one child, and two children the same:

H0: µ1 = µ2 = µ3 The mean BMI of women with no children , one child,two children is same.
Ha: Atleast one of mean BMI of women is not same.

Significance Level: α = 0.05





	df	sum_sq	mean_sq	F	PR(>F)
C(children)	2.0	24.590123	12.295062	0.334472	0.715858
Residual	563.0	20695.661583	36.759612	NaN	NaN

Conclusion: *P value is 0.715858 and it is greater than aplha(0.05) ,we fail to reject the null hypothesis and conclude that mean BMI of women with no children,one children, two children are same.*

Recommendation: Based on Statistical Analysis

- *Based on EDA and statistical evidence it can be seen that customer who smoke or have higher BMI have more higher claims. We can encourage customers to quit smoking by providing them incentive points for talking to life coach, get help for improving lifestyle habits, Quit Tobacco- 28 day program. Give gift cards when customer accumulates specific number of points.*
- *We can have Active wellness programs which can help up reduce claims related to BMI.*
- *High BMI is primarily because of unhealthy life choices. We can provide customers with Diet plans and wellness health coaches which can help them to make right choices.*
- *Provide discount coupons for Gym or fitness devices encouraging customers to exercis*

6. Model Building:

- Model Evaluation

Predict value for target variable by using our model parameter for test data set. Then compare the predicted value with actual value in test set. We compute Mean Square Error using formula: J(θ)=1m∑i=1m(y^i−yi)2

Here y^ is predicted value and y¯ is mean value of y.

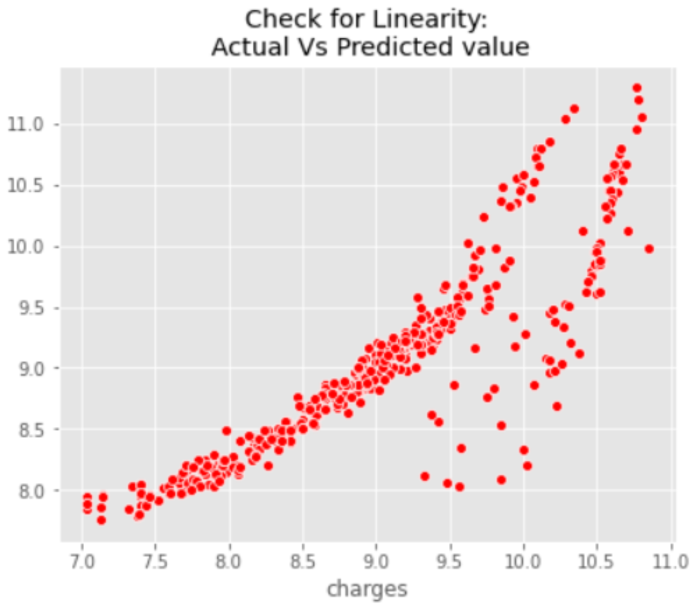
```
The Mean Square Error(MSE) or J(theta) is:  0.18729622322981895
 R square obtain for scikit learn library is : 0.7795687545055319
```

The model returns R Square value of **77.95%**, so it fit our data test very well, but still we can imporve the the performance of by diffirent technique. Please make a note that we have transformer out variable by applying natural log. When we put model into production antilog is applied to the equation.

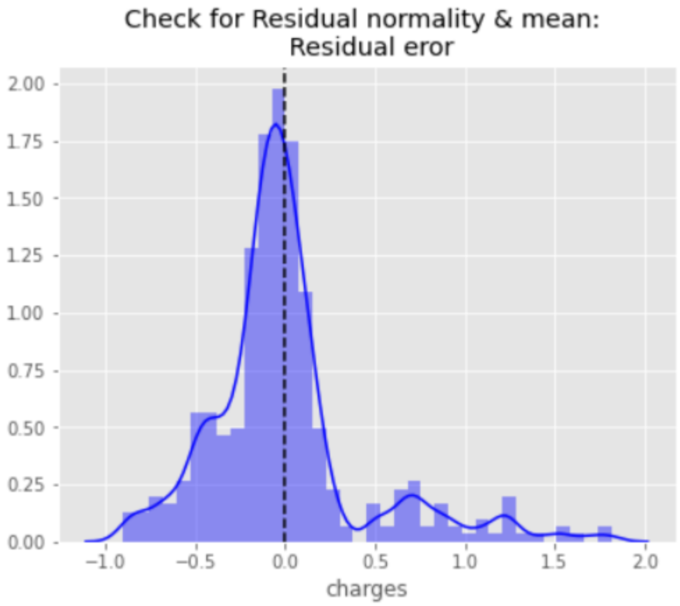
- Model Validaion

In order to validated model we need to check few assumption of linear regression model. The common assumption for Linear Regression model are following:

1. Linear Relationship: In linear regression the relationship between the dependent and independent variable to be linear. This can be checked by scatter ploting Actual value Vs Predicted value
2. The residual error plot should be normally distributed.
3. The mean of residual error should be 0 or close to 0 as much as possible.
4. The linear regression require all variables to be multivariate normal. This assumption can best checked with Q-Q plot.
5. Linear regession assumes that there is little or no Multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. The variance inflation factor VIF* identifies correlation between independent variables and strength of that correlation. **VIF=1/(1−R^2)** , If VIF >1 & VIF <5 moderate correlation, VIF < 5 critical level of multicollinearity.
6. Homoscedasticity: The data are homoscedastic meaning the residuals are equal across the regression line. We can look at residual Vs fitted value scatter plot. If heteroscedastic plot would exhibit a funnel shape pattern.



Check for Linearity:
Actual Vs Predicted value



Check for Residual normality & mean:
Residual eror

- Model Asuumption for Linear Regression:

1.In our model the actual vs predicted plot is curve so linear assumption fails.

2.The residual mean is zero and residual error plot right skewed.

3.Q-Q plot shows as value log value greater than 1.5 trends to increase.

4.The plot is exhibit heteroscedastic, error will insease after certian point.

5.Variance inflation factor value is less than 5, so no multicollearity.