**Linear Regression and Data Visualization for the Medical Cost Data**

This is the Data that represents the medical cost which supported by age, sex, BMI, children, smoker and region. In this table the independent variable is age, sex, BMI, children, smoker and region whereas the dependent variable is the charges data.

**Regression Model**

* We have made this scatter chart before the regression model to detect whether the BMI and age are related or not, if the both are related, we will have to eliminate one of the variables from the regression model.
* As we can see that the scattergram clearly indicates that the trendline we achieved is a horizontal trendline which indicates that there is no correlation between the BMI and age due to which they both can be used as an independent variable in our Regression Model.



In the regression Model we were able to predict the change in medical cost associated with the change in per unit of the Age, Sex, BMI, Children and Smoker. In this data we eliminated the region for data analysis as region data was not relevant to the analysis that was carried out.

**Intercept**

Explanation: The intercept value indicates the expected value of the dependent variable (Y) that is the medical cost= $12181.10 when all independent variables are set to zero.

**Variable 1 (Age)**

Explanation: The value in the Co-efficient column for variable 1 indicates that for every 1-year increase in age, the medical cost would increase by approximately $257.73, having the other variables constant. Getting the p-value= 2.59E-.

**Variable 2 (Sex)**

Explanation: The value for co-efficient of Sex is $128.63 which indicated that the medical cost for female is $128 more than the male. As we can see that the p-value is high which indicates that gender may not be statistically significant in predicting medical cost and the standard error is 333.36 due to which we can't definitely depend on this data achieved.

**Variable 3 (BMI)**

Explanation: The value for co-efficient of BMI is $322.36 which indicates that each unit increase in BMI increase the medical cost by $322.36 having other variables constant. Here, the p-value is 1.95E-30 which is very close to zero which indicates that BMI is highly statistically significant in predicting medical cost.

**Variable 4 (Children)**

Explanation: The value of co-efficient for Children is $474.41 which predicts that each additional children increases the medical cost by $474.41, having the other variables constant. The p-value = 0.0005 indicates that the number of children is statistically significant in predicting medical cost.

**Variable 5 (Smoker)**

Explanation: The value of co-efficient for Smoker is $23823.39 which predicts that smoking increases the medical cost by $23823.39 as compared to non-smokers. The p-value = 0 indicating that smoking is significant predictor of medical cost.

**Multiple R**

It represents the strength and direction of the linear relationship between the dependant variable and the independent variable

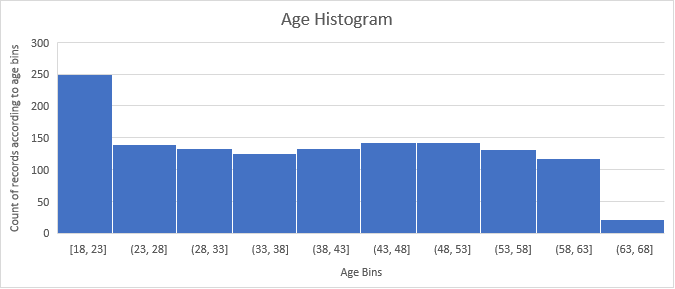
**R Square**

It indicates the proportion of the variance in the dependent variable that is justified by the independent variable in the regression model. With the value closer to 1 indicates the model to be a fit for the data.

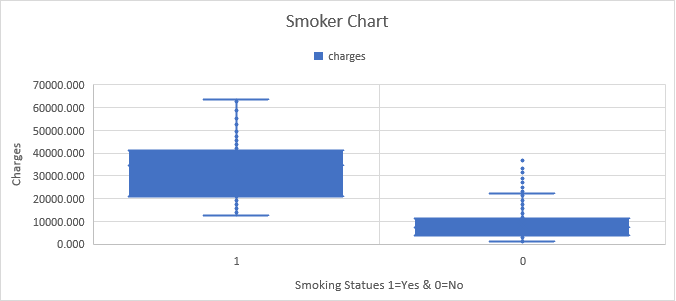
**Standard Error**

It measures the average distance between the observed value and the predicted values from the regression model having lower values indicates the model is a better fit for the data.

**Data Visualization**



* This is a Histogram Chart for the age-related data for the medical cost. Due to the Histogram Chart, we can see that the highest data for the medical cost we have is for the age between 18 to 23 years. Whereas, the lowest data we have for the medical cost is between 63 to 68 years.
* We can adjust the bin size by clicking on the x-axis bar and adjust the bin size in the setting.
* We can now the exact count of records of the age bins by moving the cursor to the age bars in the excel sheet.
* This is a scattergram for the BMI data which we can visualize that as we have less data for the higher ages, but we can clearly visualize that increase in BMI tends to increase in medical cost.
* We can visualize that the Linear Trendline indicates that the medical cost increase as the BMI increases.
* The equation y = 393.87x + 1192.9 represents a linear regression model where y is the predicted medical cost, and x is the BMI value. In this equation:
* The coefficient 393.87 represents the estimated increase in medical cost for each unit increase in BMI. The constant term 1192.9 represents the estimated medical cost when BMI is 0.
* The R² value of 0.0393 indicates the proportion of the variance in medical costs that is explained by the linear relationship with BMI. In other words, approximately 3.93% of the variability in medical costs can be attributed to changes in BMI according to this model. This suggests that BMI alone may not be a strong predictor of medical costs, as the model explains only a small portion of the variability.



* For this chart we can see that the mean of the people who smoke is not close to the mean average of people who not smoke.
* This Box and whisker chart indicates that the smoking status of an individual make a huge change in the medical cost on which can use this data for the regression modelling.