GAUTHAM AJAY KANNAN - Batch Jan 23 Online

BUSINESS REPORT - INSURANCE CLAIM

GREAT LEARNING - DA

|  |  |  |
| --- | --- | --- |
| 1. NO. | HEADING | PAGE NO. |
| 1 | CATEGORICAL AND CONTINUOUS VARIABLES | 3 |
| 2 | HISTOGRAM ,BOX PLOT AND CORRELATION FOR CONTINUOUS VARIABLES | 3 |
| 3 | PIVOT TABLES AND CHARTS | 5 |
| 4 | MULTI LINEAR REGRESSION | 11 |

1.CATEGORICAL AND CONTINUOUS

VARIABLES.

CATEGORICAL VARIABLES .

Categorical variables are variables through which the observations or the data points could be grouped or categorized into sub groups.Through these categorical variables ,we could form smaller sub divisions.Here, in this data set, the following are identified as categorical variables,

1.Sex (Gender of the customer / claimant who has claimed the

1. insurance.) .
2. Smoker (whether the claimant smokes or not).
3. Region (Region to which the claimant belongs).
4. Children (No. Of children the claimant has / No. Of dependents ).

Here ,in this case we could consider the number of children as categorical since from the given observations,we could see the value of this field lies within the range between 0 (no childern) to 5 ( five children).

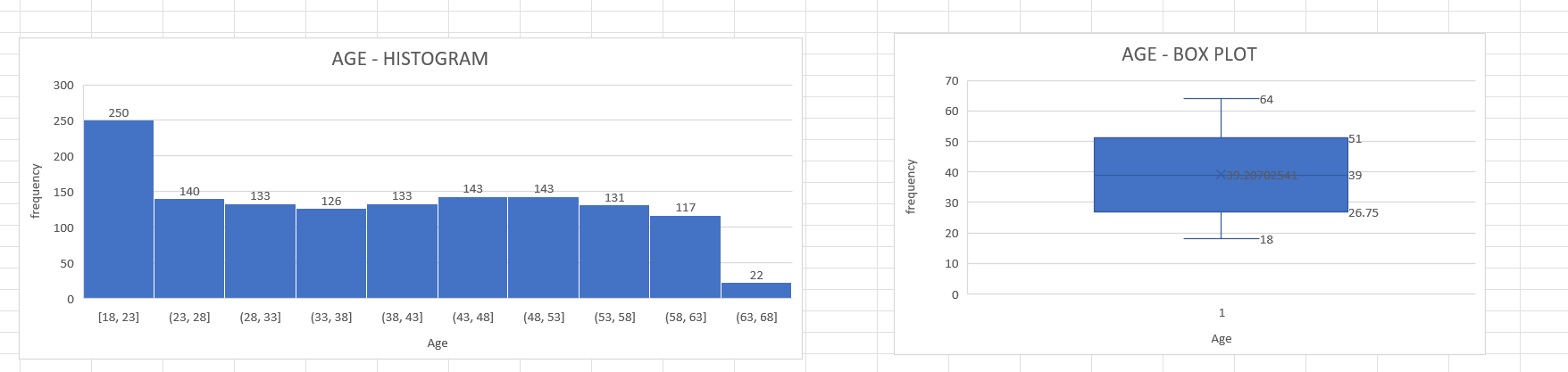
CONTINUOUS VARIABLES.

Continuous variables are those whose values can be refined or measured into finer levels or with exact measurement.Here, in this dataset,the following are identified as continuous variables.

1. Age (Age of the customer).
2. BMI (a health parameter to determine person’s obesity).
3. Charges (person’s claim from insurance).

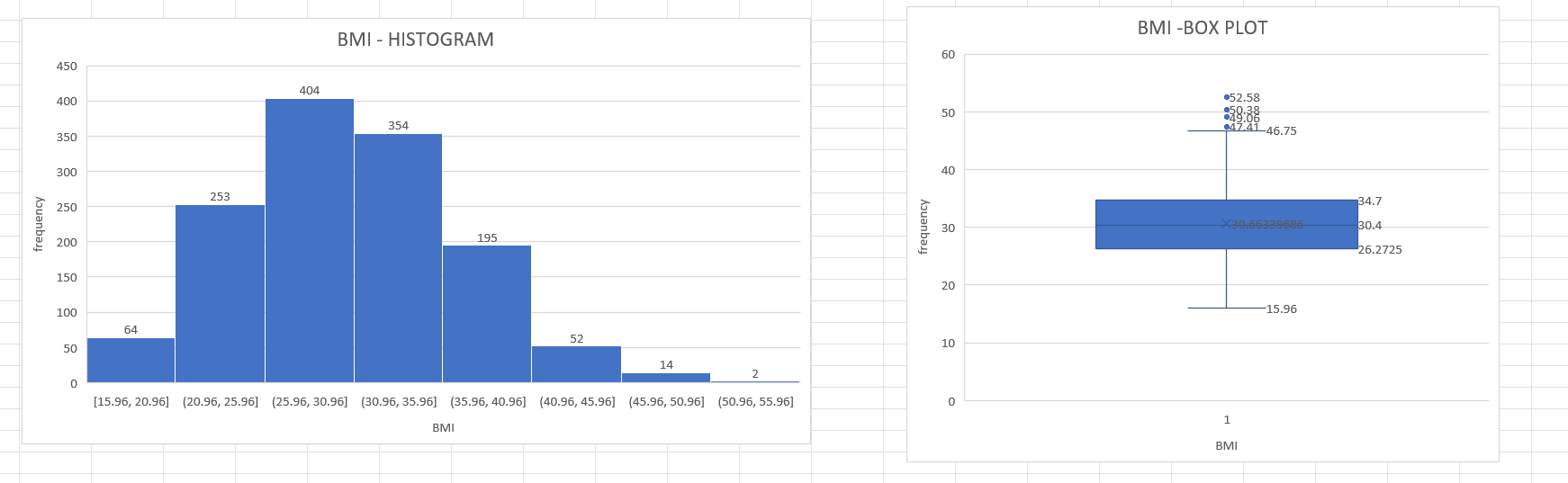
1. HISTOGRAM ,BOX PLOT AND CORRELATION FOR THE CONTINUOUS VARIABLES.

HISTOGRAM AND BOX PLOT FOR VARIABLE - AGE



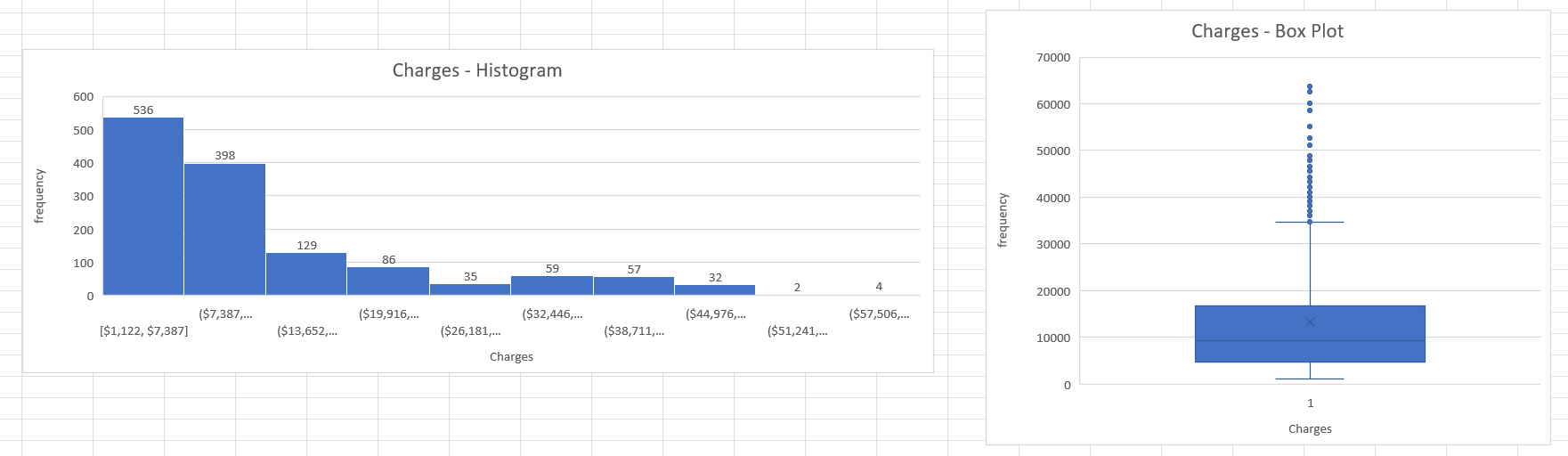
The above figure shows us the distribution of the observations or persons who have claimed insurance based on their age .From the figure ,we could see that most of the observations are distributed towards the left and the tail is formed at the right side .Hence it is positively skewed or right skewed.From the box plot ,we could see the quartile values and the maximum and minimum values .For the first 25% observations,the mean value of age is 18 and for the 50% of observations ,the mean is 39% and for 75 , it is 51 .The minimum and maximum value for the age variable is 18 and 64 respectively.

HISTOGRAM AND BOX PLOT FOR VARIABLE - BMI



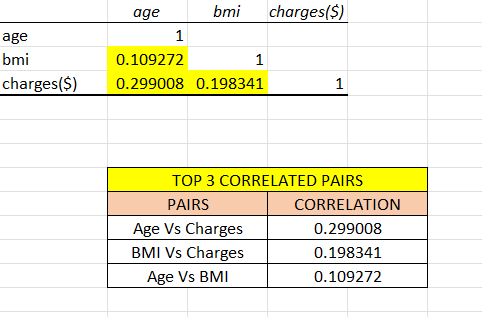
The above figure shows us the distribution of data or the observations based on BMI.From the histogram,we could see the data are distributed towards the left and tail is formed at the right side.Hence the data is positively skewed or right skewed with most of the observations are distributed in the range [25.96,30.96].From the box plot,we could see that first 25% of observations have mean BMI value of 26.27, 50% has 30.4 and 75% has 34.7.The minimum and maximum value considered by the box plot are 15.96 and 46.75 respectively.The remaining observations found above and below the whiskers are considered as outliers.

HISTOGRAM AND BOX PLOT FOR VARIABLE - CHARGES



The above figure shows us the distribution of data or the observations based on charges or claims by people.From the histogram,we could see the data are distributed towards the left and tail is formed at the right side.Hence the data is positively skewed or right skewed with most of the observations are distributed within the range [$1,122,$7,387].From the box plot,we could see that first 25% of observations have mean BMI value of $4733.63, 50% has $9382 and 75% has $16687.The minimum and maximum value considered by the box plot are $1121 and $34617 respectively.The remaining observations found above and below the whiskers are considered as outliers.

CORRELATION



From the above figure ,we could see the correlation for the continuous variables that we have identified from the given dataset.We could see that all the possible combinations of pairs are positively correlated.Hence ,there is an positive linear relationship between those pairs.

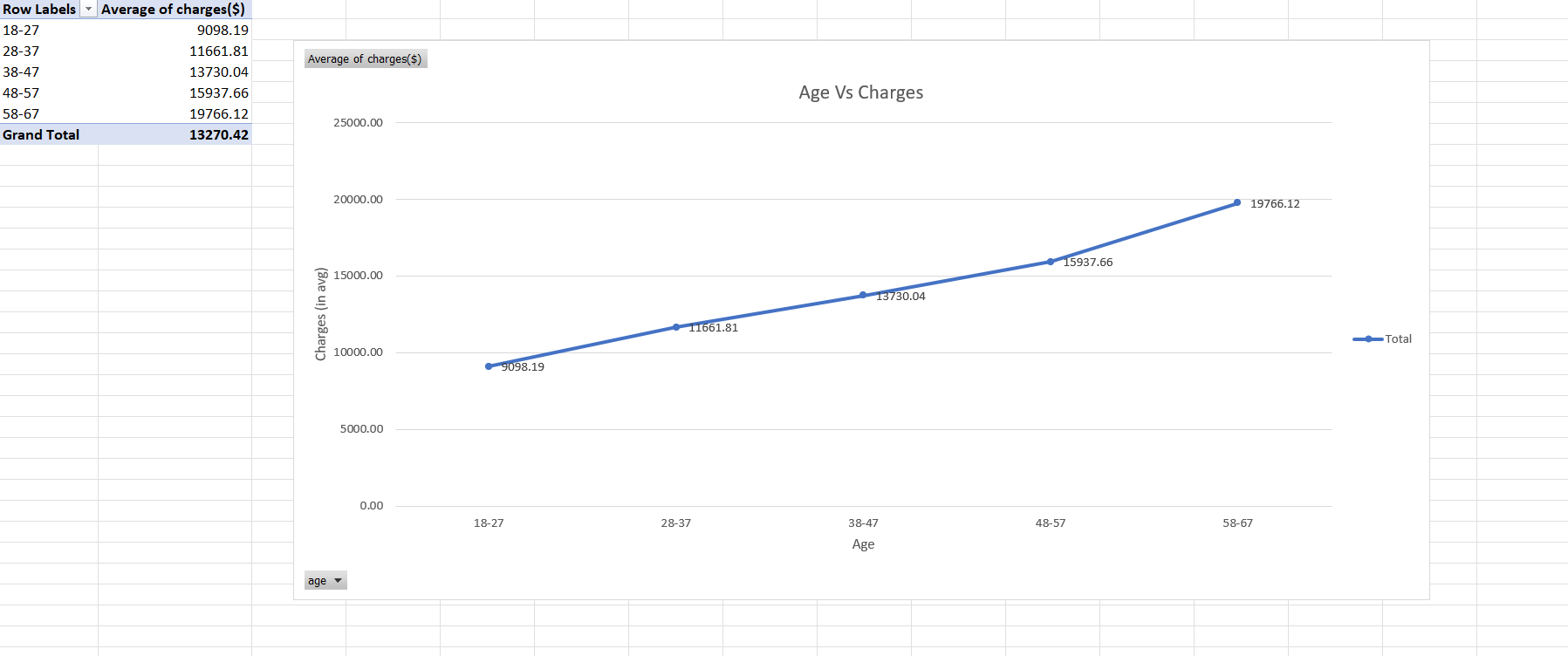
3.PIVOT TABLE AND CHARTS.

SMOKER COUNT BASED ON GENDER/SEX



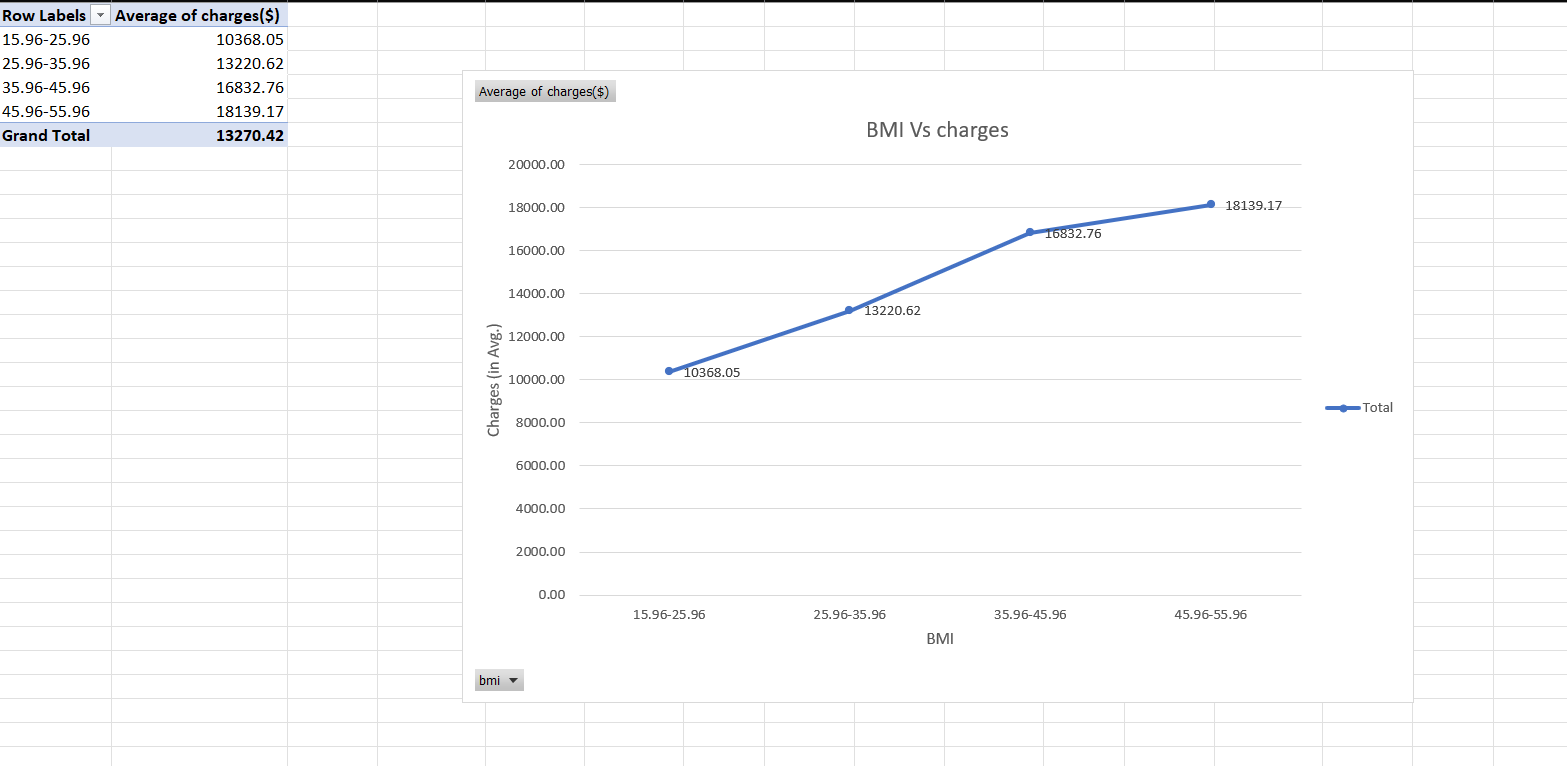
From the above table and chart,we could see that based on the given observations,males have higher smoker count when compared with females.In case of non smokers,females have higher count when compared with males.

CHARGES VS. AGE



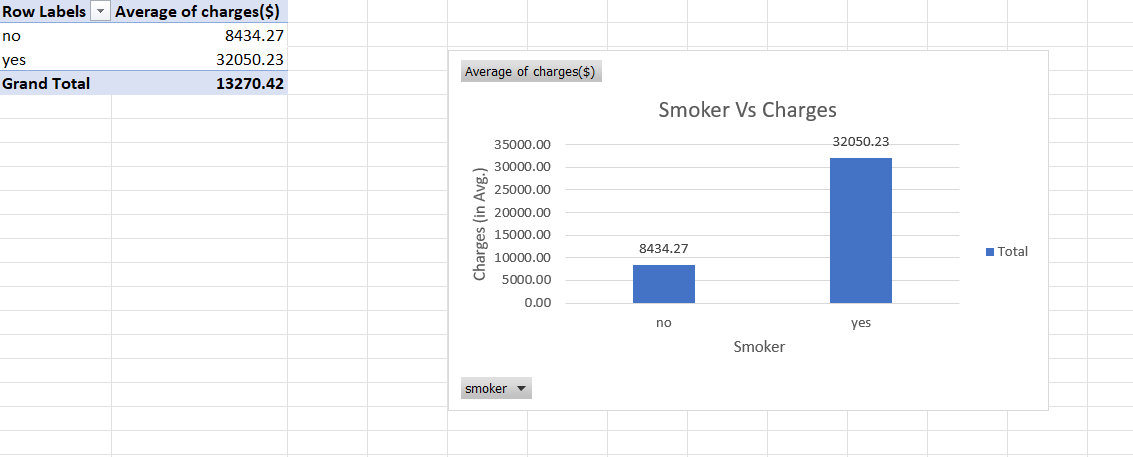
From the above chart and table,we could see that as range of the age increases ,the average charges or the claims by the respective people within the age range also increase steadily with person having age from 58 to 67 having the highest average claim as $19766.12.

CHARGES VS. BMI



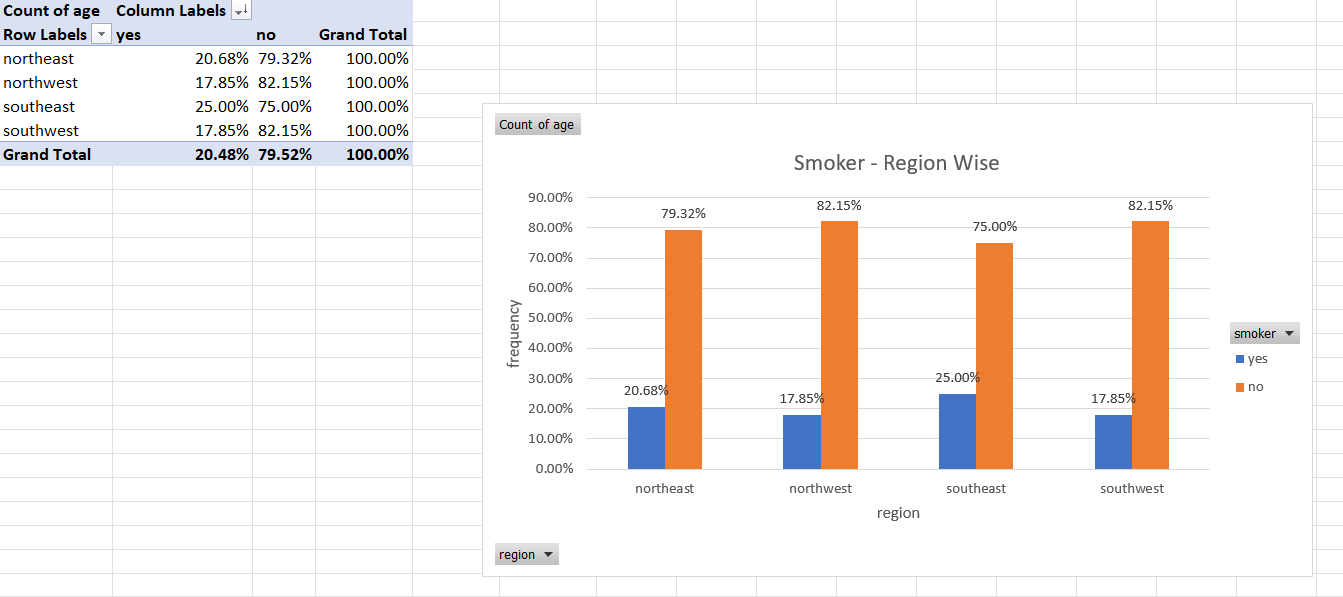
From the above chart and table,we could see that as range of the BMI of people increases ,the average charges or the claims by the respective people within the respective BMI range also increase steadily with person having BMI range 45.96-55.96 having the highest average claim as $18139.17.

CHARGES FOR SMOKERS VS. NON SMOKERS



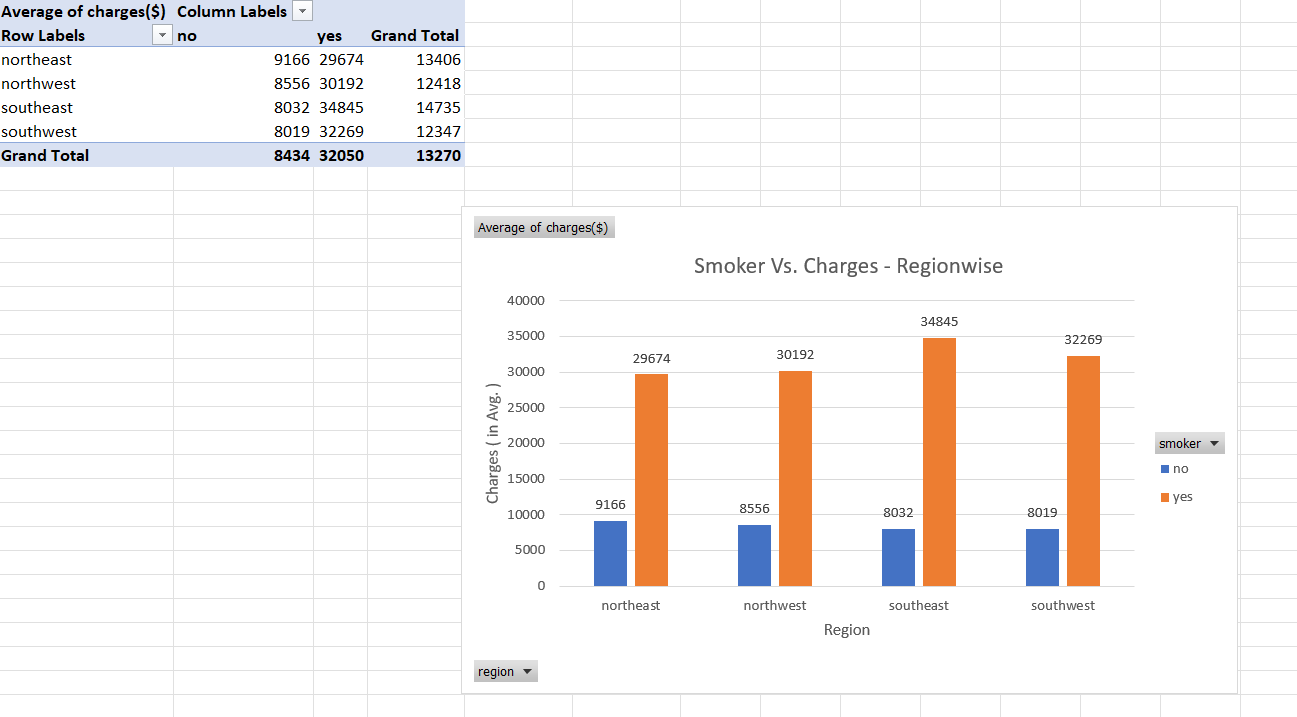
From the above chart and table,we could see the distribution of smokers and non smokers based on the charges or claims that they have obtained with people who smokes having the highest average claim as $32050.23.

SMOKER PERCENTAGE - REGIONWISE



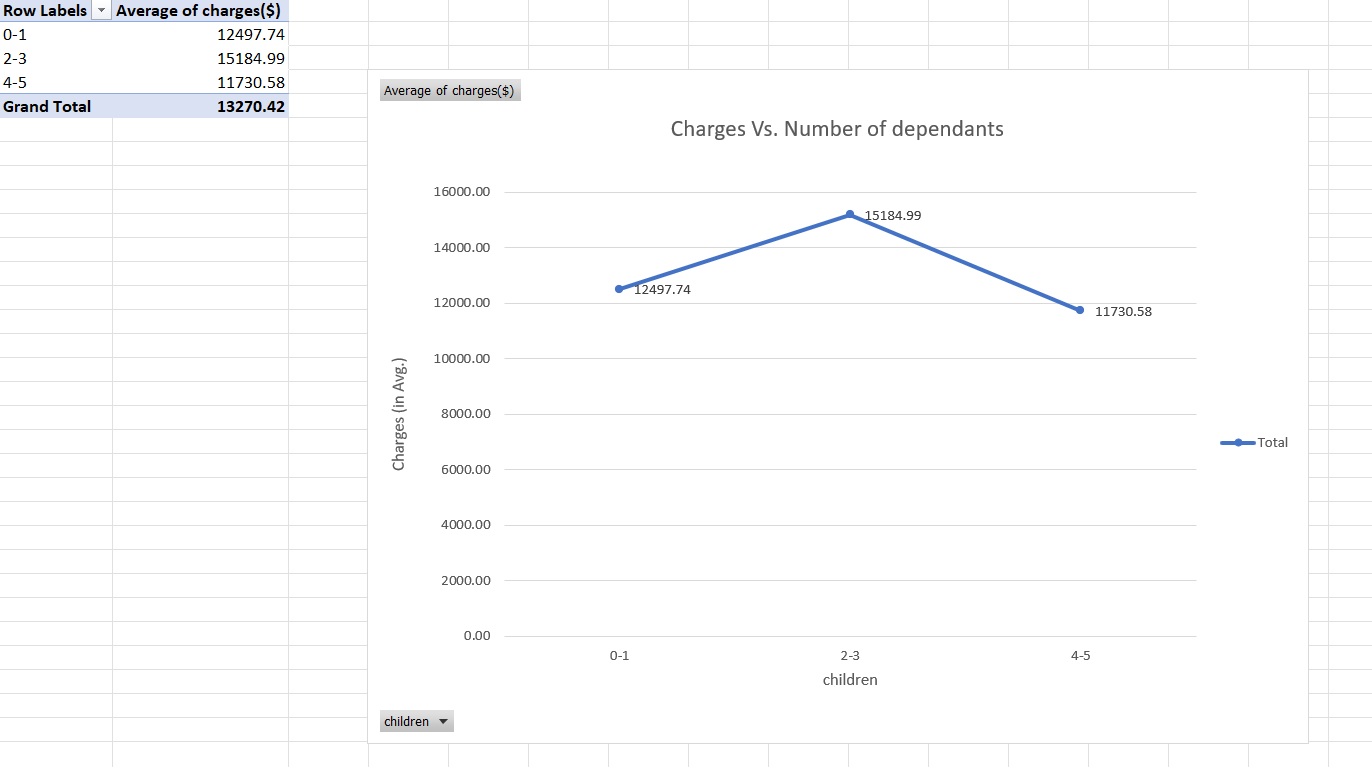
The above chart and table shows us the region wise distribution of smokers and non smokers in percentage .Comparing with other regions ,southeast region has the highest proportion of smokers and as well as lowest proportion of smokers.The overall smokers’ percentage is 20.48% and non smoker is 79.52%.

REGIONWISE CHARGES FOR SMOKER VS. NON SMOKER



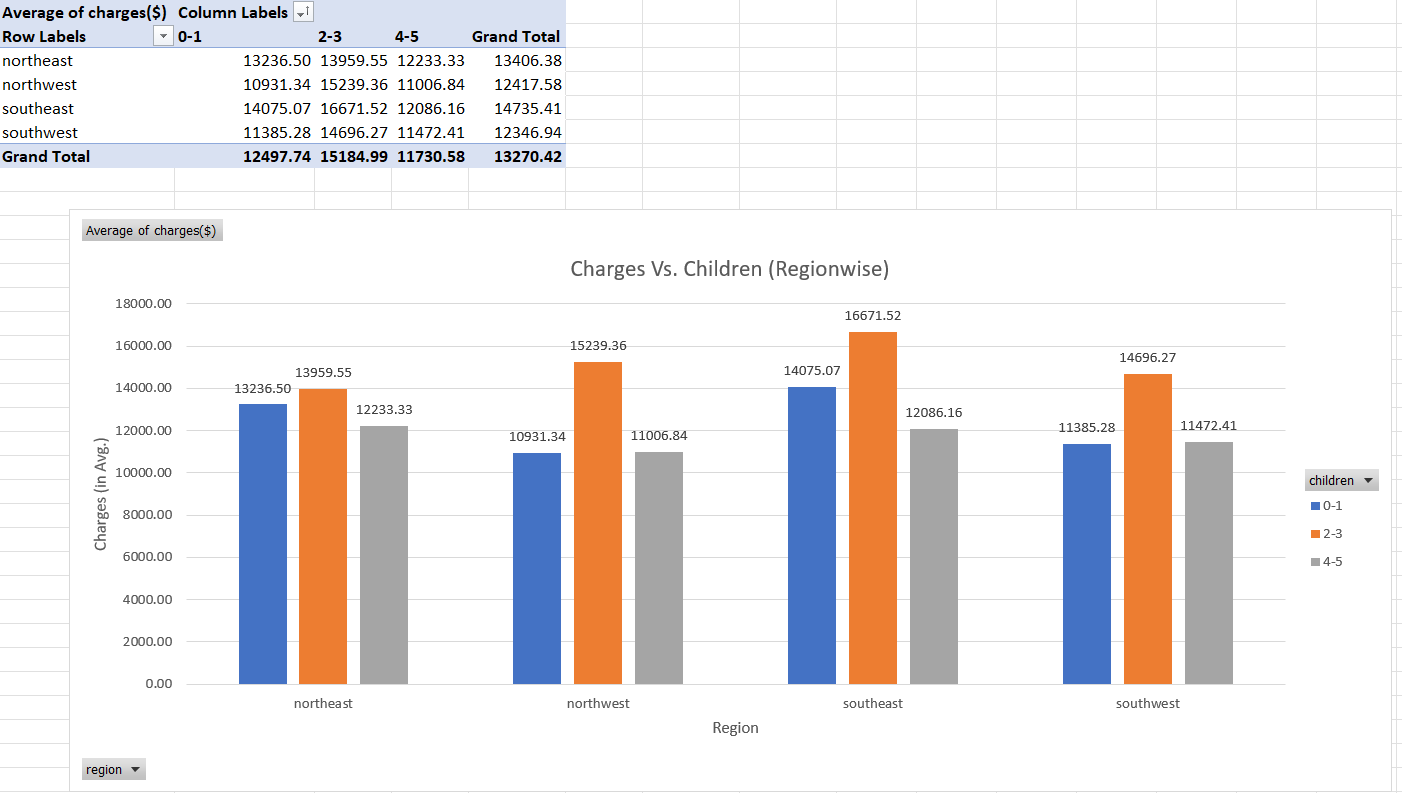
The above chart shows us the average charges or claims by smokers and non smokers in their respective regions.For smokers,the highest average claim ($34845) is found in southeast region and the lowest claim ($29674) is found in northeast region.For non smokers,the highest average claim is in northeast region and the lowest claim ($8019) is in southwest region.

CHARGES VS. NUMBER OF DEPENDENTS (CHILDREN)



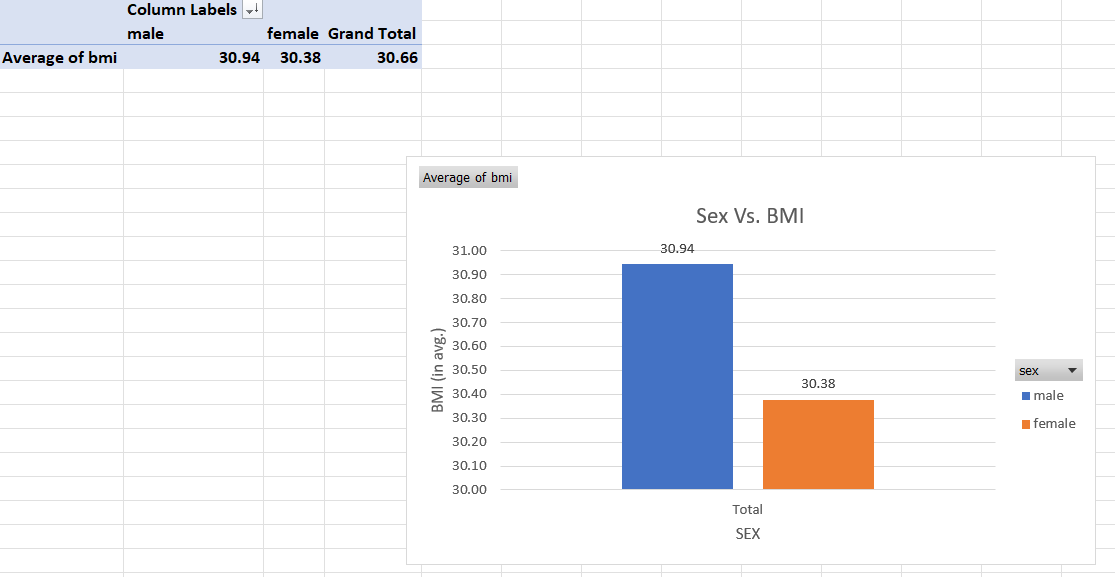
The above table and figure depicts the relationship between the number of children for the claimants versus their respective charges in average.Here, we could see that there is an steady increase in charges for the persons who have number of children or dependents between 0 to 3 and there is a sudden decrease for the charges claimed by persons who have children or dependents between 2 to 5.

CHARGES VS CHILDREN - REGIONWISE



The above table and figure depicts the relationship between the number of children for the claimants versus their respective charges in average region wise.Here, we could see that the highest average claim ($16671.52) is done by the people who have 2-3 children in southeast region and the lowest average claim ($10931.34) is done by people who have 0-1 children in northwest region.

SEX VS. BMI



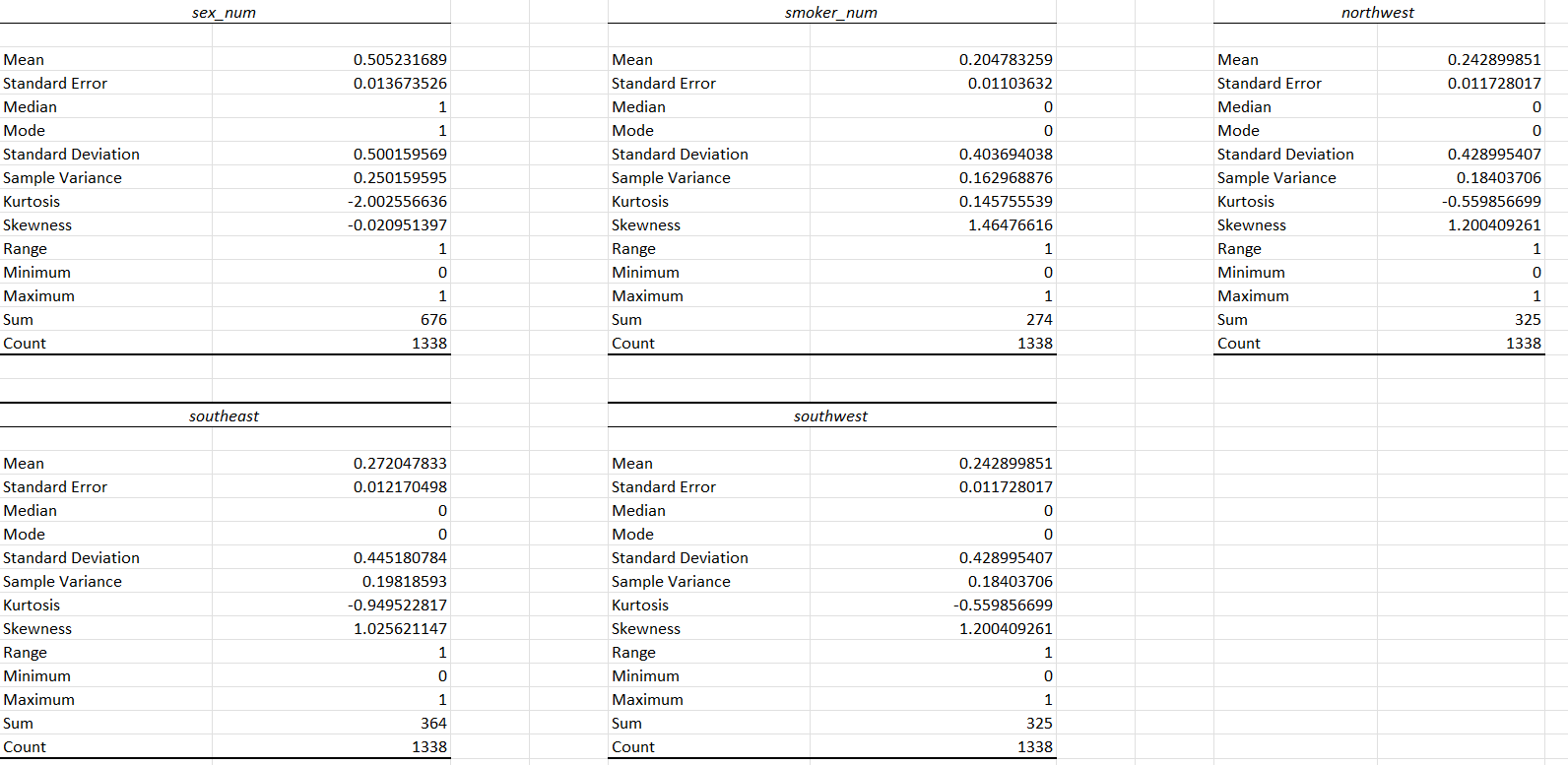
The above relation shows us the distribution of data based on their gender and BMI value.We could see that the highest average BMI is found in males when compared with females.

SEX VS BMI - REGIONWISE

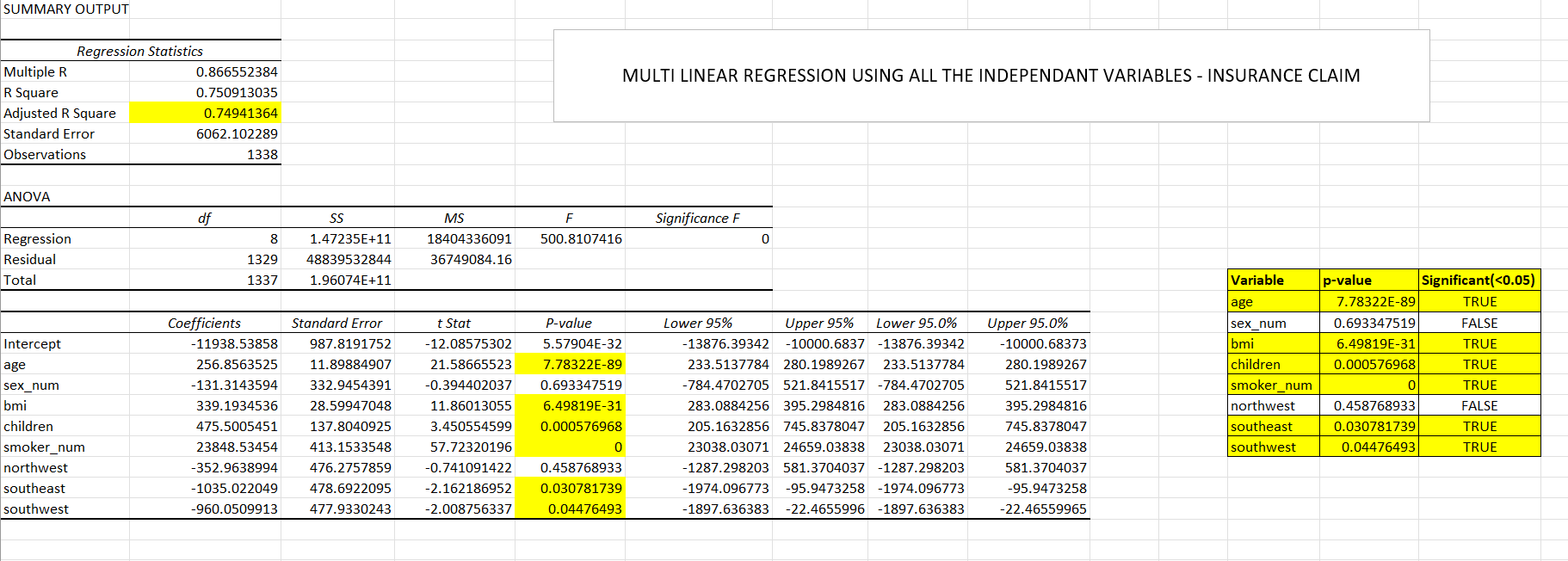


The above relation shows us the distribution of data based on their gender and BMI value regionwise.We could see that the highest average BMI is found in males in southeast region and the lowest is found in males in northeast region.

4.MULTI LINEAR REGRESSION

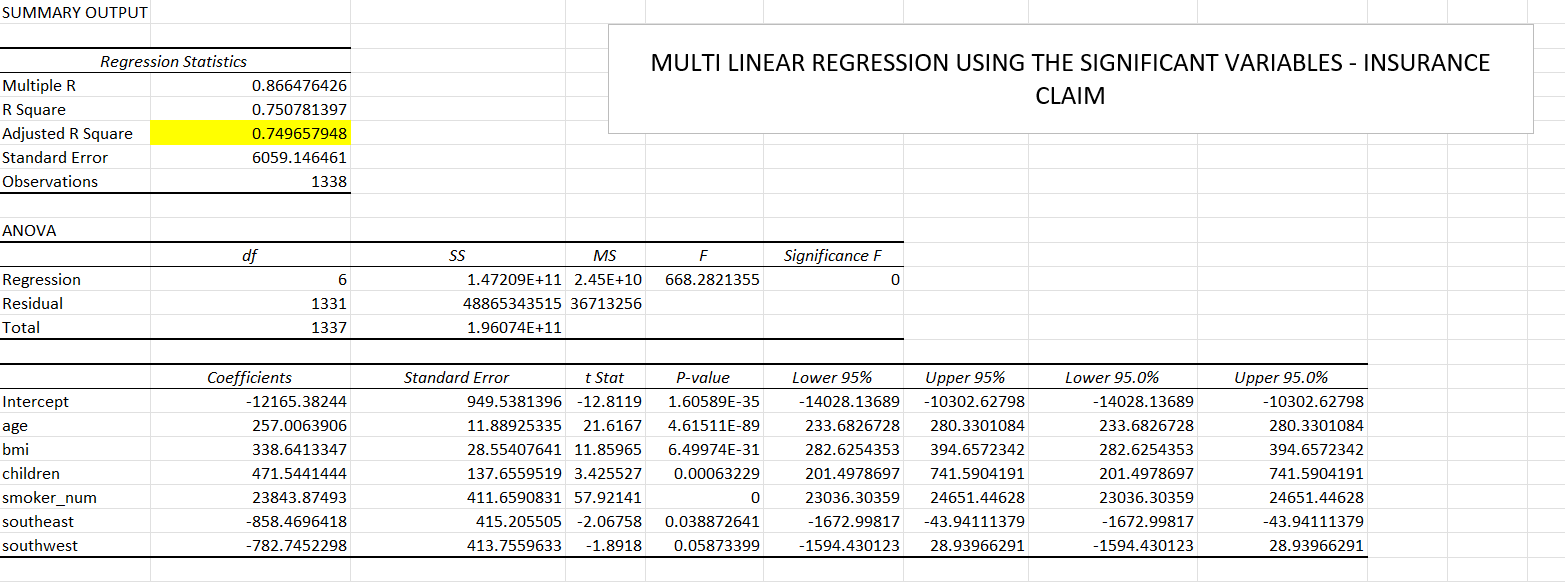


The above picture shows us the summary statistics of the edited or the categorical variables after converted to numerical.



The above picture shows us the created multi linear regression model after the categorical variables are converted to numerical and the identification of significant variables to further refine the model.Here the adjusted R2 value is 0.74941 or 75% which is closer to one.Hence ,based on this value,it could be considered as a good model.The value of the intercept is -11938.53 which is the least value that the charges could obtain when other dependent variables are zero.But this value is not feasible as charges or claims cannot be negative.The regression equation formed by the model is

Y=-11938.53+256.85X0-131.31X1+339.19X1+475.5X2+23848.5X3-352.96X4-1035.02X5-960X6



The above picture shows us the multi linear regression model based on the significant variables.The intercept value is -12165.38 which is the least possible value that the charges could obtain but it is not feasible.The adjusted R2 value is 0.74965 which is slightly higher than the previous model.The regression equation formed by the model is

Y=-12165.38+257X0+338.64X1+471.544X2+23843.87X3-858.46X4-782.74X5