Health Insurance claim

CAUSE AND EFFECT ANALYSIS

ANOOP E R | DATA ANALYTICS | 06-01-2023

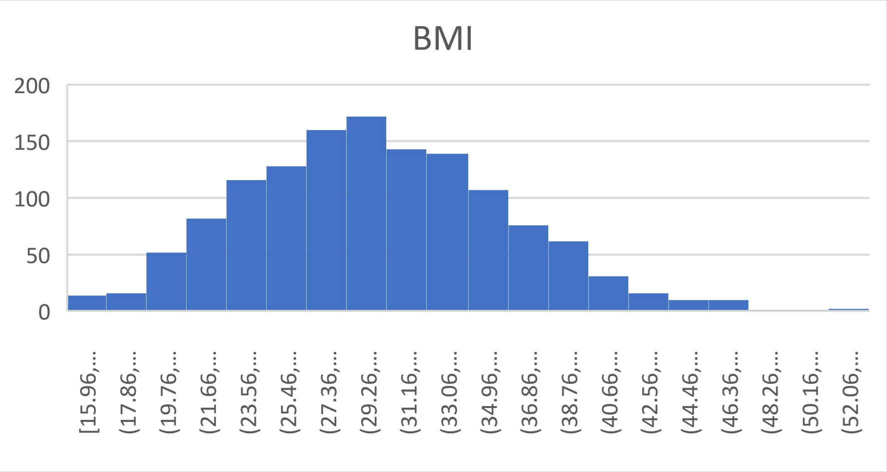
1) Perform the Exploratory Data Analysis on the data.

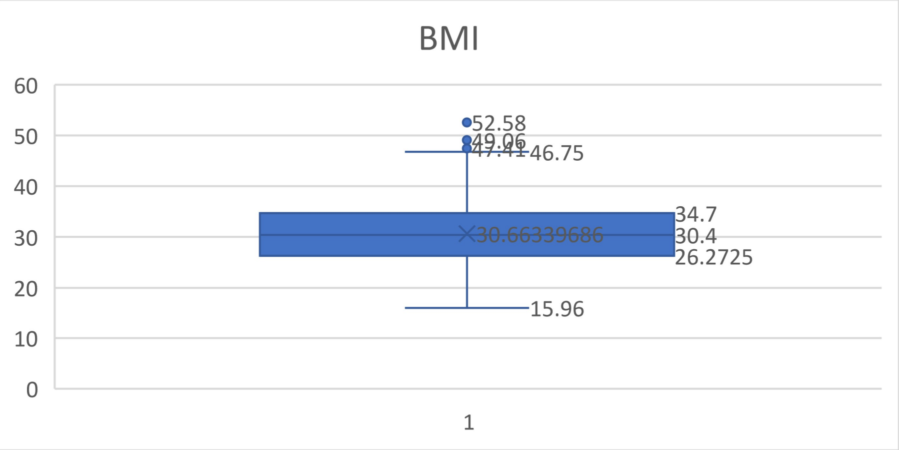
a) Identify the categorical and continuous variables

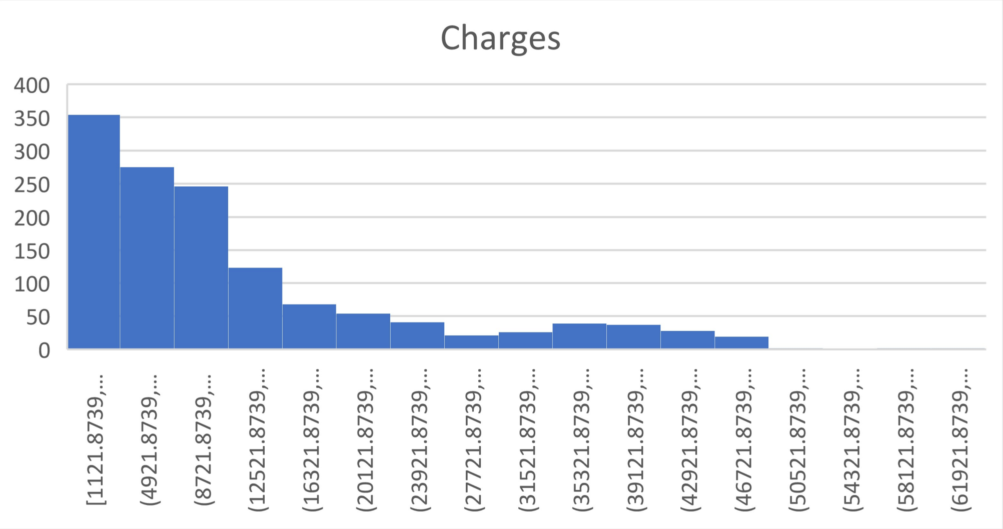
Categorical variables	Continuous variables
Sex	Bmi
Smoker	Charges
Region	

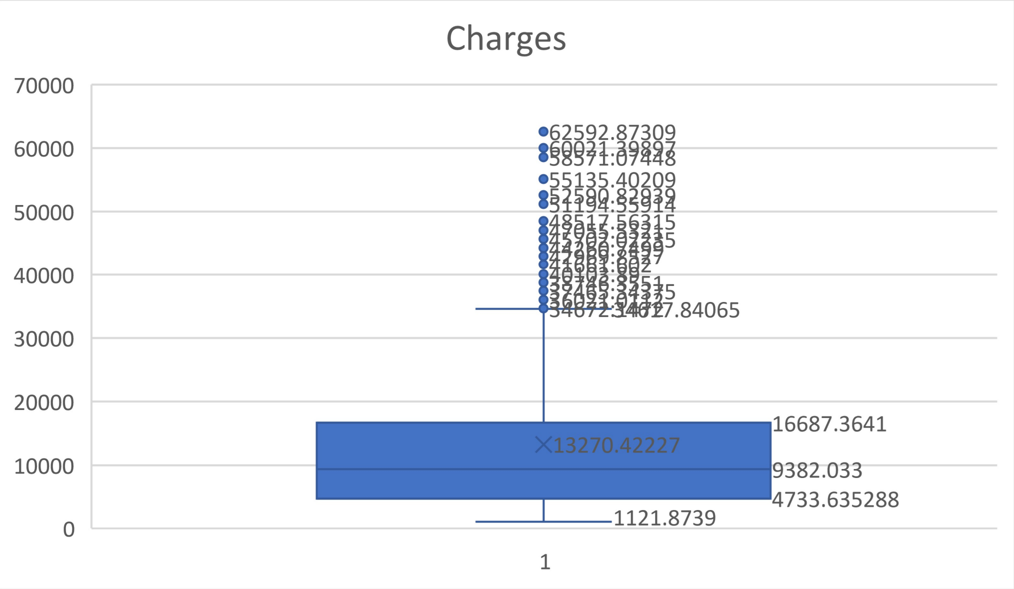
Age and **Children** is discrete so we separately place it in the category "discrete".

b) Make Histograms and box plots (univariate analysis) for continuous variables and do a correlation analysis (multivariate analysis)









Correlation analysis		
	bmi	*charges($)*
bmi	1	
charges($)	0.198340969	1

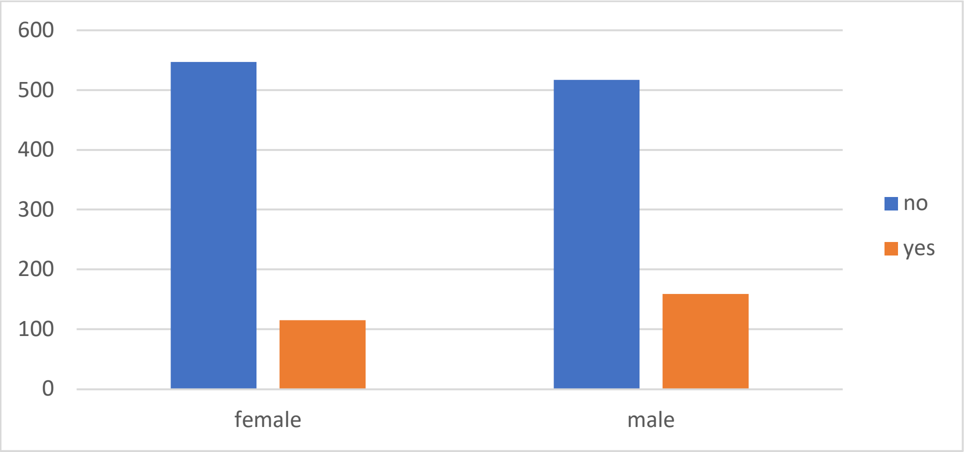
c) Make relevant Pivot tables and charts for:

1)Male/Female ratio and share information on which gender has more smokers

Count of smoker	Column Labels	
Sex	no	yes
female	547	115
male	517	159

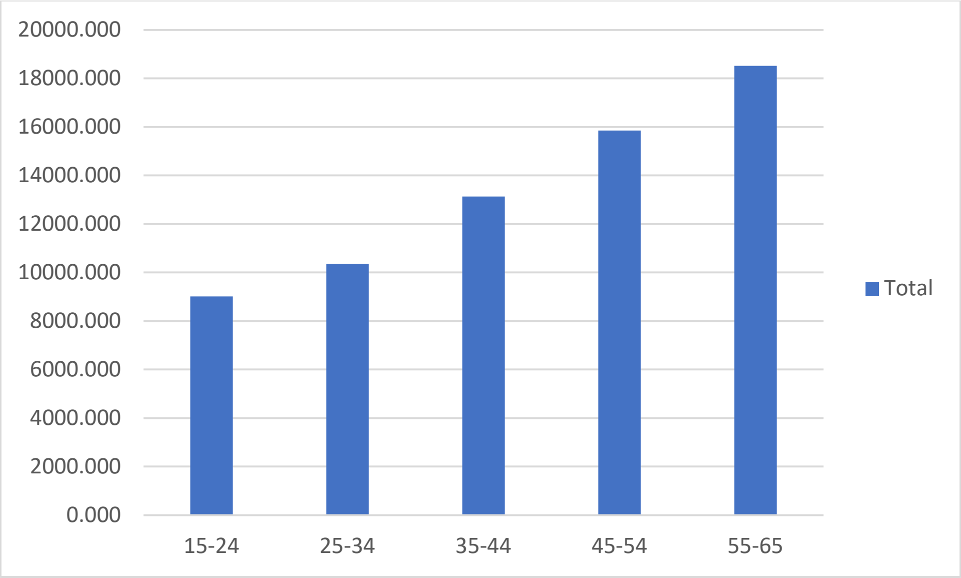
Male by Female Ratio = **1.382608696**

By examining the male by female ratio we see that the ratio is above 1,thus we can conclude that males has more smokers.



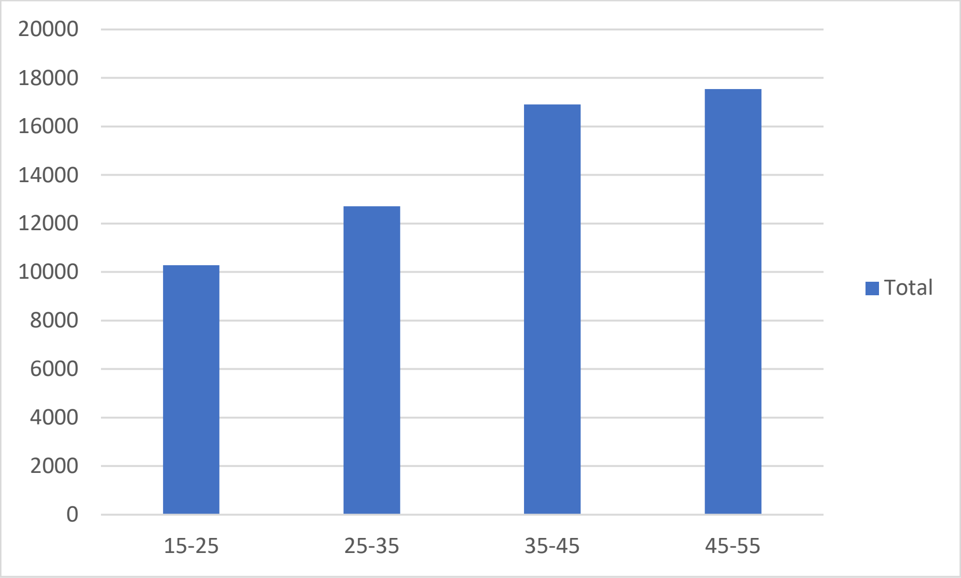
2)Charges vs Age

Age	Average of charges($)
15-24	9011.340
25-34	10352.393
35-44	13134.169
45-54	15853.928
55-65	18513.276
Grand Total	**13270.422**



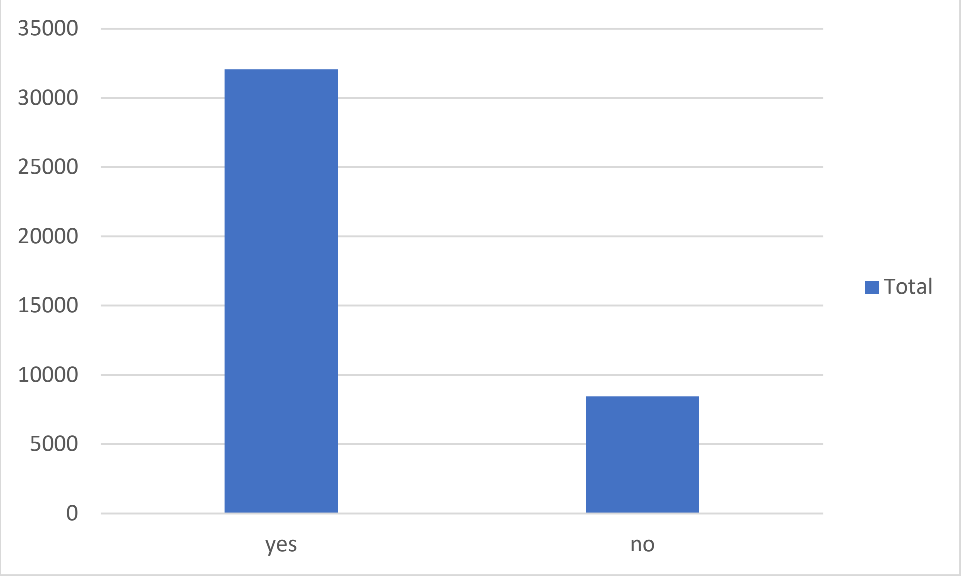
3)Charges vs BMI

BMI	Average of charges($)
15-25	10282.22447
25-35	12714.63543
35-45	16913.68151
45-55	17547.92675



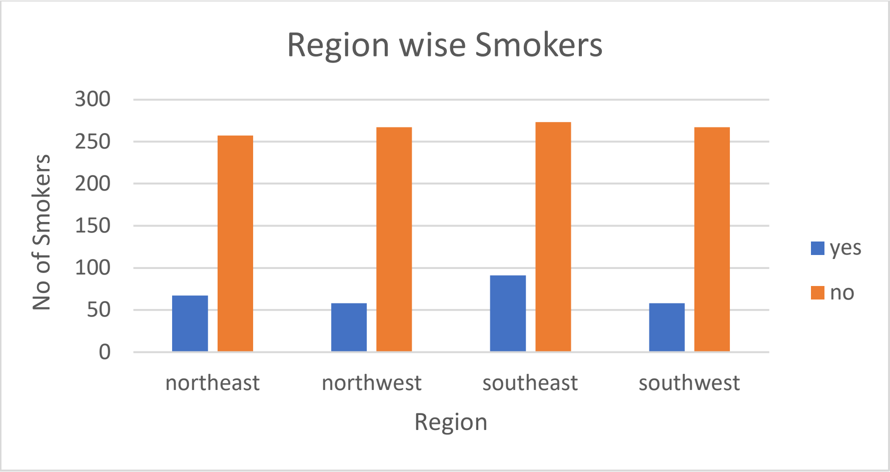
4) Charges for Smokers vs Non-smokers

Smokers	Average of charges($)
yes	32050.23183
no	8434.268298



d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

No of smoker	Region	
Row Labels	yes	no
northeast	67	257
northwest	58	267
southeast	91	273
southwest	58	267



region

northeast

northwest

southeast

southwest

smoker

no

yes

<span>Southeast</span> has more number of smokers and northwest along with southwest holds the less number of smokers

e) Region-wise charges for smokers vs non-smokers

Average of charges($)	Column Labels	
Row Labels	**no**	**yes**
northeast	9165.532	29673.536
northwest	8556.464	30192.003
southeast	8032.216	34844.997
southwest	8019.285	32269.063

f) Has charges got something to do with the number of dependents ?

Correlation between number of dependents and charges = 0.067998

Since we have a positive relation we can say that they are directly related. Thus we can say that as the value of no of dependents increase, charges also increase.

g) Do a similar dependants-charges analysis, Region-wise

Average of charges($)	Number of Students					
Region	**0**	**1**	**2**	**3**	**4**	**5**
northeast	11626.463	16310.206	13615.153	14409.913	14485.193	6978.973
northwest	11324.371	10230.256	13464.315	17786.161	11347.019	8965.796
southeast	14309.868	13687.042	15728.471	18449.846	14451.024	10115.442
southwest	11938.505	10406.485	17483.486	10402.442	14933.261	8444.159

h) Do at least one more pivot table and chart of your own choice on the remaining variables

Average of bmi	Sex	
Row Labels	**female**	**male**
no	30.53952468	30.77058027
yes	29.60826087	31.50418239

i) Give your understanding from the patterns observed in point (b)

<u>Interpretation for observations made in point (b)</u>

- **The datas in BMI is normally distributed with a median of 30.4.**
- **For BMI the first quartile data is under 26.272 and third quartile data is under 34.7.**
- **The datas in Charges are positively skewed with a median of 9382.033.**
- **The first quartile data is under 4733.635.**

j) Give your interpretation for observations made in point (c)

<u>Interpretation for observations made in point (c)</u>

- **Males has more number of smokers.**
- **The BMI range of 45-55 has highest average charge of 17547.92675.**
- **Average charges for smokers is four times the charges for non-smokers.**
- **The Age group 55-65 has the highest average charge of 18513.26.**

2) Edit the data as following, to obtain dummy variables:

a) Sex : Replace all the “Males” with “1” and “Females” with “0”, creating numerical entries for gender this way will help you do analysis further. You can use the “Replace with Match entire cell content” option. Do a replace all to save time.

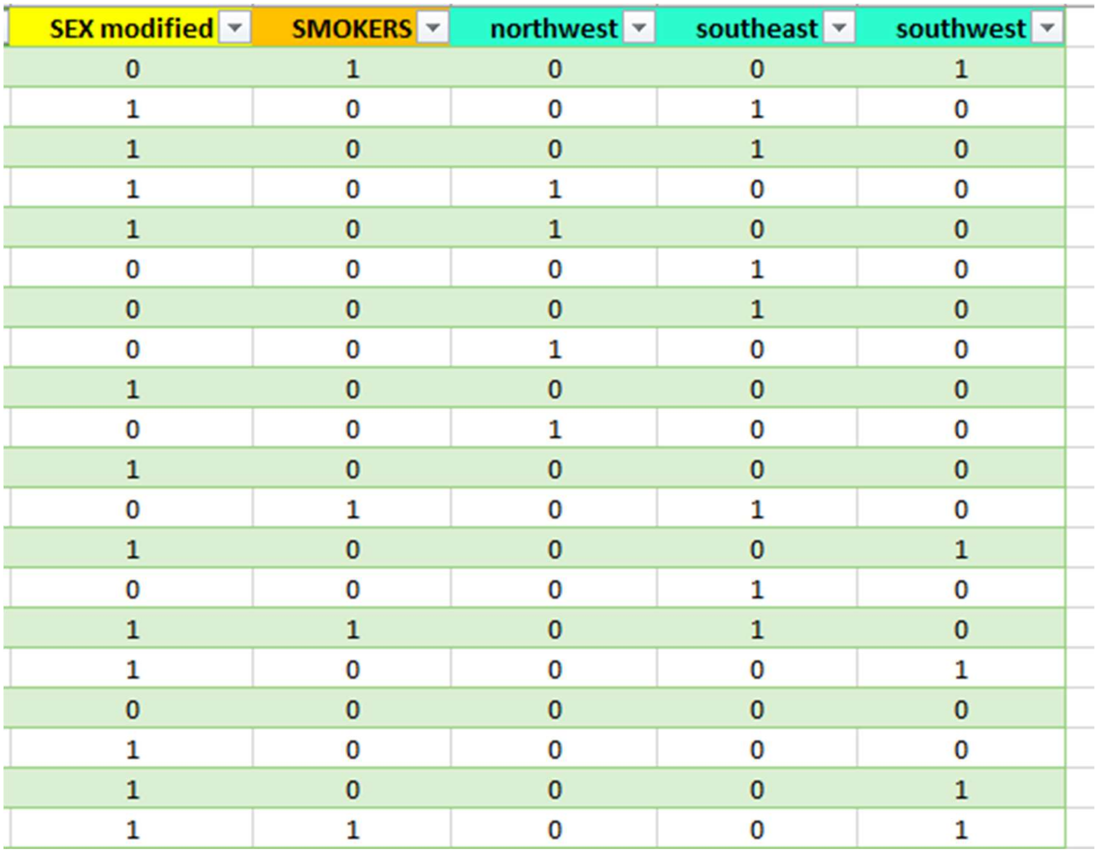
b) Smoker: Replace all the “Smokers” with “1” and “Non-smokers” with “0”.

c) Region: We always create one less category column for the dummy data w.r.t the categories available for that original variable. So for Region, we will create three dummy columns, assuming “Northeast” as zero and omit the column for it. Now create three columns for “northwest”, “Southeast”, “Southwest”. Whichever row has “northwest” region as an entry will take “1” as an entry otherwise “0” in “northwest” column. Similarly in the “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest of the column (Southeast). Do a similar operation on the “Southwest” column. Please refer to the below image for your understanding,

a) We use the if function to edit the data (=IF(Cell="male",1,0))

b) We use the if function to edit the data (=IF(Cell="yes",1,0))

c) We use the if function to edit the data (=IF(Cell=" northwest",1,0))
We use the if function to edit the data (=IF(Cell=" Southeast",1,0))
We use the if function to edit the data (=IF(Cell=" Southwest ",1,0))



SEX modified	SMOKERS	northwest	southeast	southwest
0	1	0	0	1
1	0	0	1	0
1	0	0	1	0
1	0	1	0	0
1	0	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	1	0	0
1	0	0	0	0
0	0	1	0	0
1	0	0	0	0
0	1	0	1	0
1	0	0	0	1
0	0	0	1	0
1	1	0	1	0
1	0	0	0	1
0	0	0	0	0
1	0	0	0	0
1	0	0	0	1
1	1	0	0	1

3) Do a descriptive summary analysis for the edited data. Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim. Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

Descriptive Summary Analysis of edited data

We use the summary statistics in the data analytics function

age		***bmi***	
Mean	**39.2070**	**Mean**	**30.6634**
Standard Error	**0.3841**	**Standard Error**	**0.1667**
Median	**39.0000**	**Median**	**30.4000**
Mode	**18.0000**	**Mode**	**32.3000**
Standard Deviation	**14.0500**	**Standard Deviation**	**6.0982**
Sample Variance	**197.4014**	**Sample Variance**	**37.1879**
Kurtosis	**-1.2451**	**Kurtosis**	**-0.0507**
Skewness	**0.0557**	**Skewness**	**0.2840**
Range	**46.0000**	**Range**	**37.1700**
Minimum	**18.0000**	**Minimum**	**15.9600**
Maximum	**64.0000**	**Maximum**	**53.1300**
Sum	**52459.0000**	**Sum**	**41027.6250**
Count	**1338.0000**	**Count**	**1338.0000**

children		*SEX*	
Mean	1.0949	Mean	0.5052
Standard Error	0.0330	Standard Error	0.0137
Median	1.0000	Median	1.0000
Mode	0.0000	Mode	1.0000
Standard Deviation	1.2055	Standard Deviation	0.5002
Sample Variance	1.4532	Sample Variance	0.2502
Kurtosis	0.2025	Kurtosis	-2.0026
Skewness	0.9384	Skewness	-0.0210
Range	5.0000	Range	1.0000
Minimum	0.0000	Minimum	0.0000
Maximum	5.0000	Maximum	1.0000
Sum	1465.0000	Sum	676.0000
Count	1338.0000	Count	1338.0000

SMOKERS		*northwest*		*southeast*	
Mean	0.2048	Mean	0.2429	Mean	0.2720
Standard Error	0.0110	Standard Error	0.0117	Standard Error	0.0122
Median	0.0000	Median	0.0000	Median	0.0000
Mode	0.0000	Mode	0.0000	Mode	0.0000
Standard Deviation	0.4037	Standard Deviation	0.4290	Standard Deviation	0.4452
Sample Variance	0.1630	Sample Variance	0.1840	Sample Variance	0.1982
Kurtosis	0.1458	Kurtosis	-0.5599	Kurtosis	-0.9495
Skewness	1.4648	Skewness	1.2004	Skewness	1.0256
Range	1.0000	Range	1.0000	Range	1.0000
Minimum	0.0000	Minimum	0.0000	Minimum	0.0000
Maximum	1.0000	Maximum	1.0000	Maximum	1.0000
Sum	274.0000	Sum	325.0000	Sum	364.0000
Count	1338.0000	Count	1338.0000	Count	1338.0000

southwest		*charges($)*	
Mean	0.2429	Mean	13270.4223
Standard Error	0.0117	Standard Error	331.0675
Median	0.0000	Median	9382.0330
Mode	0.0000	Mode	1639.5631
Standard Deviation	0.4290	Standard Deviation	12110.0112
Sample Variance	0.1840	Sample Variance	146652372.1529
Kurtosis	-0.5599	Kurtosis	1.6063
Skewness	1.2004	Skewness	1.5159
Range	1.0000	Range	62648.5541
Minimum	0.0000	Minimum	1121.8739
Maximum	1.0000	Maximum	63770.4280
Sum	325.0000	Sum	17755824.9908
Count	1338.0000	Count	1338.0000

We use the regression analysis in data analytics function from the data tab for Multiple Linear Regression analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866552384
R Square	0.750913035
Adjusted R Square	0.74941364
Standard Error	6062.102289
Observations	1338

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	1.47235E+11	18404336091	500.8107416	0
Residual	1329	48839532844	36749084.16		
Total	1337	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11938.53858	987.8191752	-12.08575302	5.57904E-32	-13876.39342	-10000.68373	-13876.39342	-10000.68373
age	256.8563525	11.89884907	21.58665523	7.78322E-89	233.5137784	280.1989267	233.5137784	280.1989267
bmi	339.1934536	28.59947048	11.86013055	6.49819E-31	283.0884256	395.2984816	283.0884256	395.2984816
children	475.5005451	137.8040925	3.450554599	0.000576968	205.1632856	745.8378047	205.1632856	745.8378047
SEX	-131.3143594	332.9454391	-0.394402037	0.693347519	-784.4702705	521.8415517	-784.4702705	521.8415517
SMOKERS	23848.53454	413.1533548	57.72320196	0	23038.03071	24659.03838	23038.03071	24659.03838
northwest	-352.9638994	476.2757859	-0.741091422	0.458768933	-1287.298203	581.3704037	-1287.298203	581.3704037
southeast	-1035.022049	478.6922095	-2.162186952	0.030781739	-1974.096773	-95.9473258	-1974.096773	-95.9473258
southwest	-960.0509913	477.9330243	-2.008756337	0.04476493	-1897.636383	-22.46559965	-1897.636383	-22.46559965

AVERAGE = 42.0353%

ACCURACY = 57.9647%

Interpretation for the above analysis

- **From this analysis we can observe that the insignificant variables are sex and southeast.**
- **The variable Smokers have a pvalue, i.e it is the most significant variable.**
- **This model has a accuracy of 57.964%.**

Observing p-value

Model created after removing the variables **sex** and **northwest.**

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.866476426
R Square	0.750781397
Adjusted R Square	0.749657948
Standard Error	6059.146461
Observations	1338

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.47209E+11	24534813009	668.2821355	0
Residual	1331	48865343515	36713255.83		
Total	1337	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12165.38244	949.5381396	-12.81189447	1.60589E-35	-14028.13689	-10302.62798	-14028.13689	-10302.62798
age	257.0063906	11.88925335	21.61669729	4.61511E-89	233.6826728	280.3301084	233.6826728	280.3301084
bmi	338.6413347	28.55407641	11.85964939	6.49974E-31	282.6254353	394.6572342	282.6254353	394.6572342
children	471.5441444	137.6559519	3.425526743	0.00063229	201.4978697	741.5904191	201.4978697	741.5904191
SMOKERS	23843.87493	411.6590831	57.92141097	0	23036.30359	24651.44628	23036.30359	24651.44628
southeast	-858.4696418	415.205505	-2.067577697	0.038872641	-1672.99817	-43.94111379	-1672.99817	-43.94111379
southwest	-782.7452298	413.7559633	-1.891804105	0.05873399	-1594.430123	28.93966291	-1594.430123	28.93966291

Observing co-orelation

Model created after removing the variables **Northwest** and **Southwest**.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.866105937
R Square	0.750139494
Adjusted R Square	0.74901315
Standard Error	6066.944607
Observations	1338

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.47083E+11	24513836220	665.9953865	0
Residual	1331	48991204250	36807816.87		
Total	1337	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12225.124	957.856346	-12.76300361	2.8118E-35	-14104.19667	-10346.05132	-14104.19667	-10346.05132
age	257.0213198	11.908064	21.58380403	7.8189E-89	233.6607002	280.3819393	233.6607002	280.3819393
bmi	333.9631442	28.48961163	11.72227788	2.84657E-30	278.0737084	389.8525801	278.0737084	389.8525801
children	468.9779152	137.8408603	3.402314192	0.000688007	198.5688968	739.3869335	198.5688968	739.3869335
SEX	-129.1910687	333.2080003	-0.387718988	0.698285997	-782.8611641	524.4790266	-782.8611641	524.4790266
SMOKERS	23866.02912	413.3256052	57.74147264	0	23055.18848	24676.86976	23055.18848	24676.86976
southeast	-579.0291828	388.5085342	-1.490389868	0.136358685	-1341.184984	183.1266187	-1341.184984	183.1266187