U.S INSURANCE COMPANY

BUSINESS REPORT

PREPARED BY

AMRUTH N

SUMMARY:

Performing cause and effect analysis on historic-data of insurance claims so that company can estimate what premium should the company charge a customer availing an insurance policy.

SCOPE:

The insurance company has collected a dataset of 1338 customers-claims.

FINDINGS:

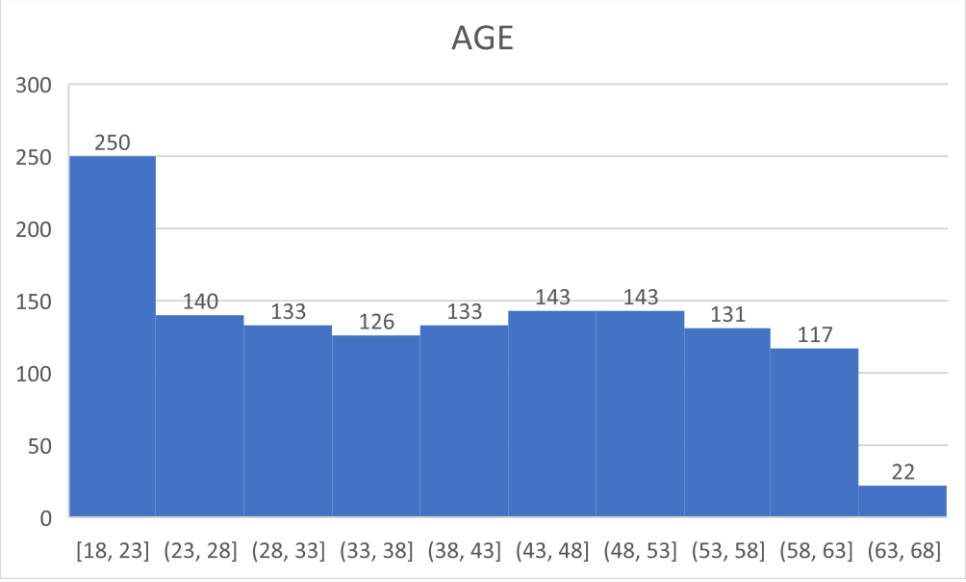
1) Exploratory Data Analysis on the data.

a) Identifying the categorical and continuous variables

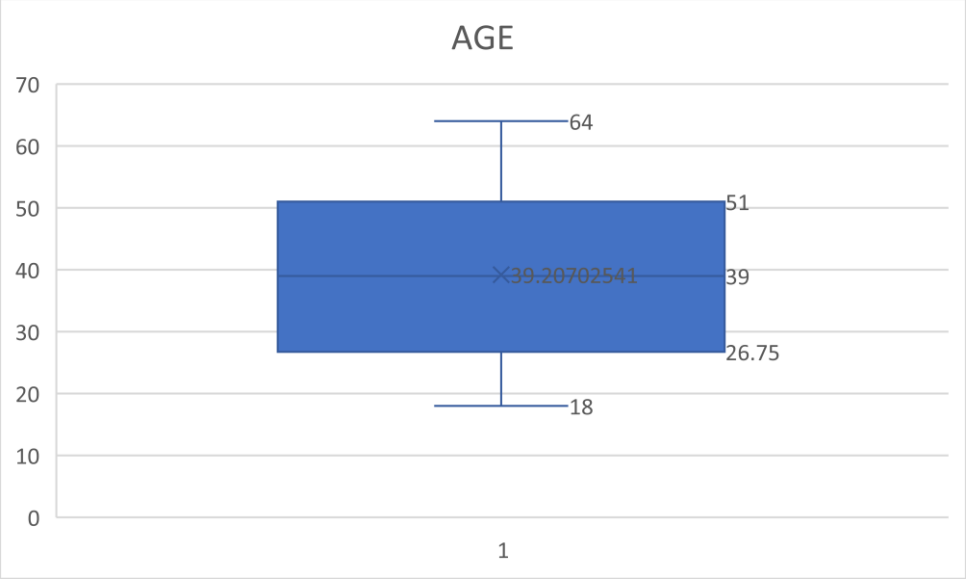
Categorical Variable	Continuous Variable
SEX	AGE
SMOKER	BMI
REGION	CHARGES ($)

b) Histogram and box plot for continuous variables:

i) AGE:

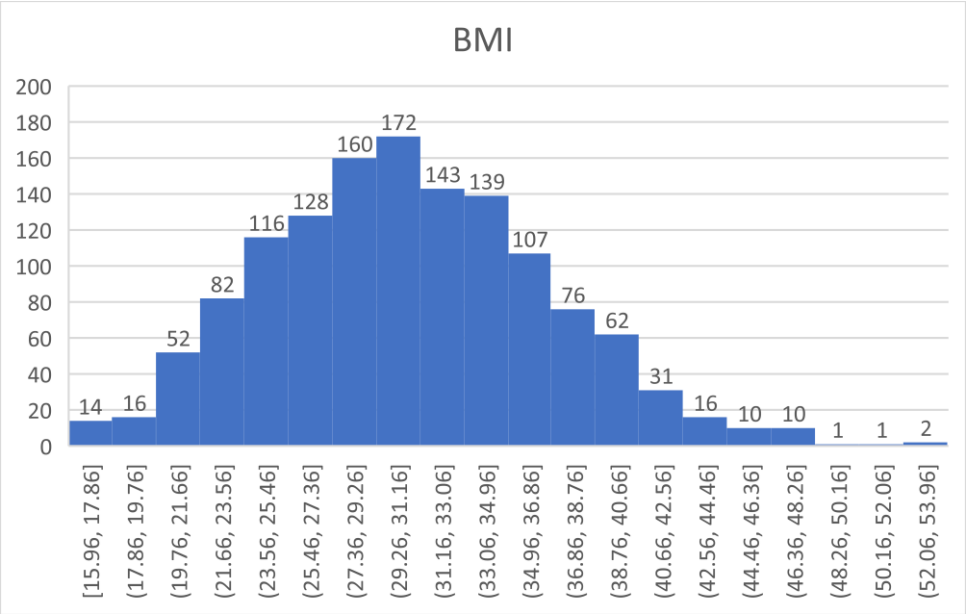


AGE: from the histogram we can infer that, 250 claimer's age is between 18-23 i.e 16.6% , and second largest age group i.e 143 is between 43-48 and 48-53 , and the third largest age group i.e 140 is between 23-28 i.e 10.5% and we can also observe that the least number is in the age group of 63 -68 which suggests senior citizens have not claimed much according to the given dataset or the sample size of this age group is less.

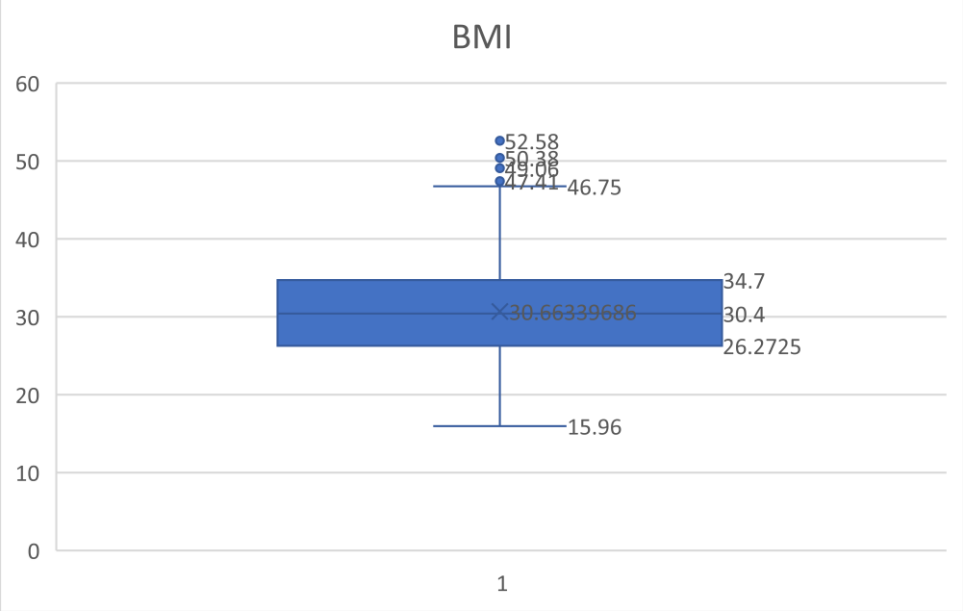


From the Box plot, we can infer that, minimum age who has claimed insurance is 18, maximum is 64 and the median i.e middle value is 39 , since the mean and median are close , this shows that the data is distributed evenly across the median and there are no outliers present in the data, which denotes that the given data is true.

ii) BMI



BMI: from the histogram, we can infer that, 172 claimer's BMI ranges from 29.26-31.6 i.e 12.3%, second largest group (160) has a BMI of 27.36-29.26, and the third largest group (143) has a BMI of 31.6-33.06. Interestingly, the least number of the claimer's BMI ranges from 48.26-52.06 which indicates the BMI of 48.26 and above have claimed lower than the remaining customers.



From the Box plot, we can infer that, minimum BMI of claimer is 15.96, maximum is 52.58, median i.e middle values is 30.4 , there are few values present outside the fourth quartile, which indicates that sample size of those values are less in numbers.

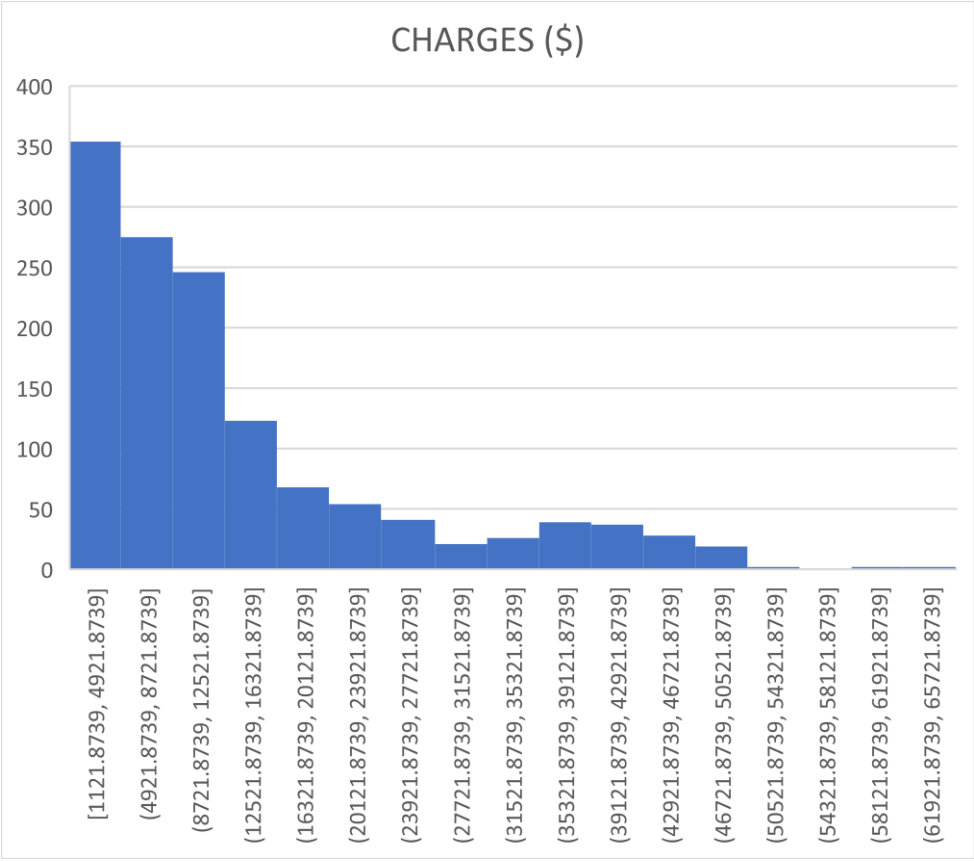
We can test if these values are outliers or not using formula,

Outlier = Q3+1.5IQR

IQR	8.4275
Q3	34.7
OUTLIER1	47.34125

Since the range of outlier is only up to 47.34, we can conclude that all values above this are outliers and they are obese.

iii) Charges($)

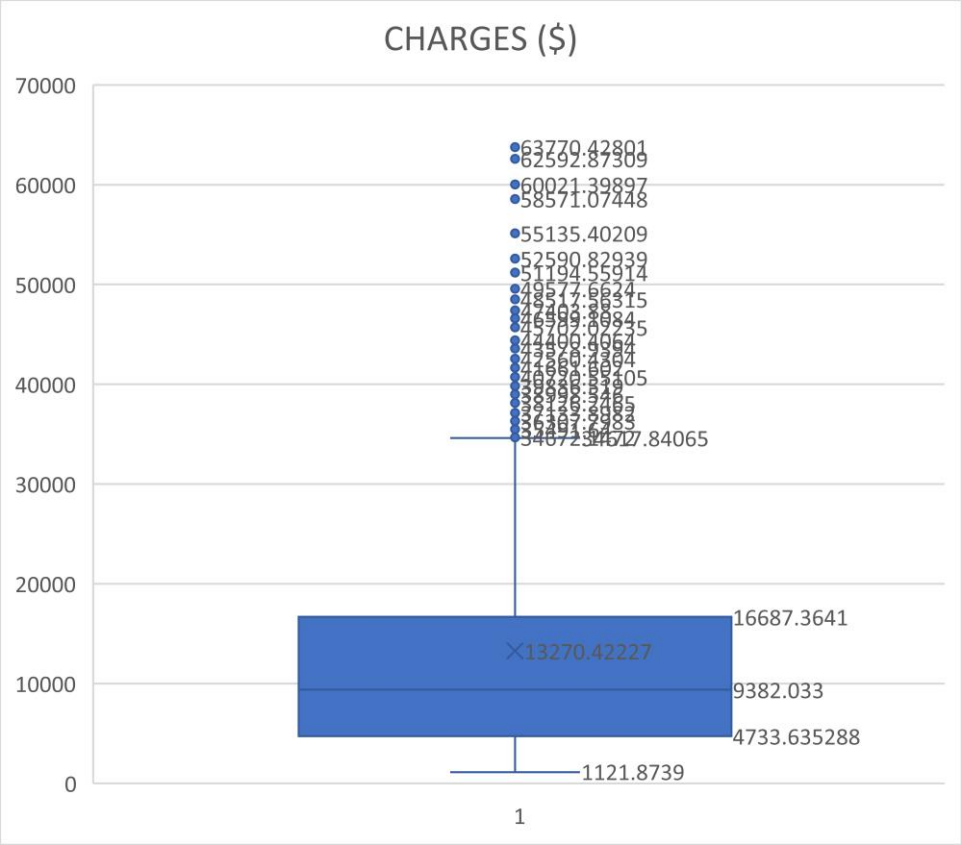


Charges($): From the histogram, we can infer that 354 customers have claimed an amount between 1121.8739 - 4921.8739 which is the least amount range, and 275 customers have claimed an amount between 4921.8739 - 87218739. The highest amount range is claimed by 1 or 2 customers only.

From the below box plot , we can infer that minimum amount claimed is 1121.8739 , Median is 9382.033 , mean is 13270.422 since the mean and median is far apart from each other, it indicates that amount claimed is not similar between customers.
we can test if these values are outliers or not using formula,
Outlier = Q3+1.5IQR

IQR	11953.73
Q3	16687.36
OUTLIER1	34617.96

All the values above 34618 are outliers and affects the data.



Correlation Matrix:

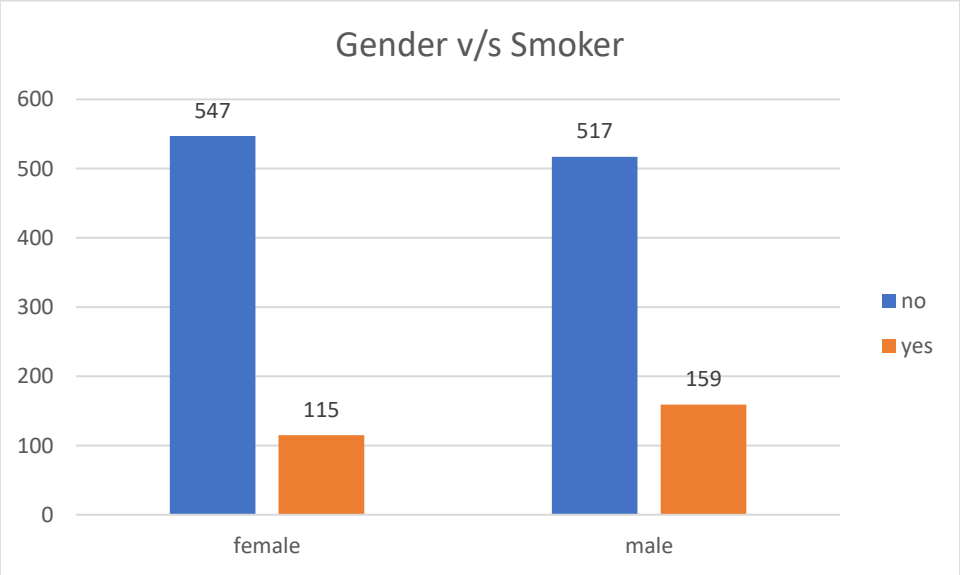
	age	*children*	*bmi*	*charges($)*
age	1			
children	0.042469	1		
bmi	0.109272	0.012759	1	
charges($)	0.299008	0.067998	0.198341	1

The matrix denotes that there is no real relation between any of the variables present in the data.

1)c. Pivot table and charts:

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

Count of sex	Column Labels		
SEX	no	yes	Grand Total
female	547	115	662
male	517	159	676
Grand Total	**1064**	**274**	**1338**



The male/female ratio is 676:662 i.e 1.02
Male gender has higher smoker ratio then female smoker. i.e 517:159, after simplifying, 5.17:1.59 whereas female ratio is 547:115 i.e 5.47:1.15

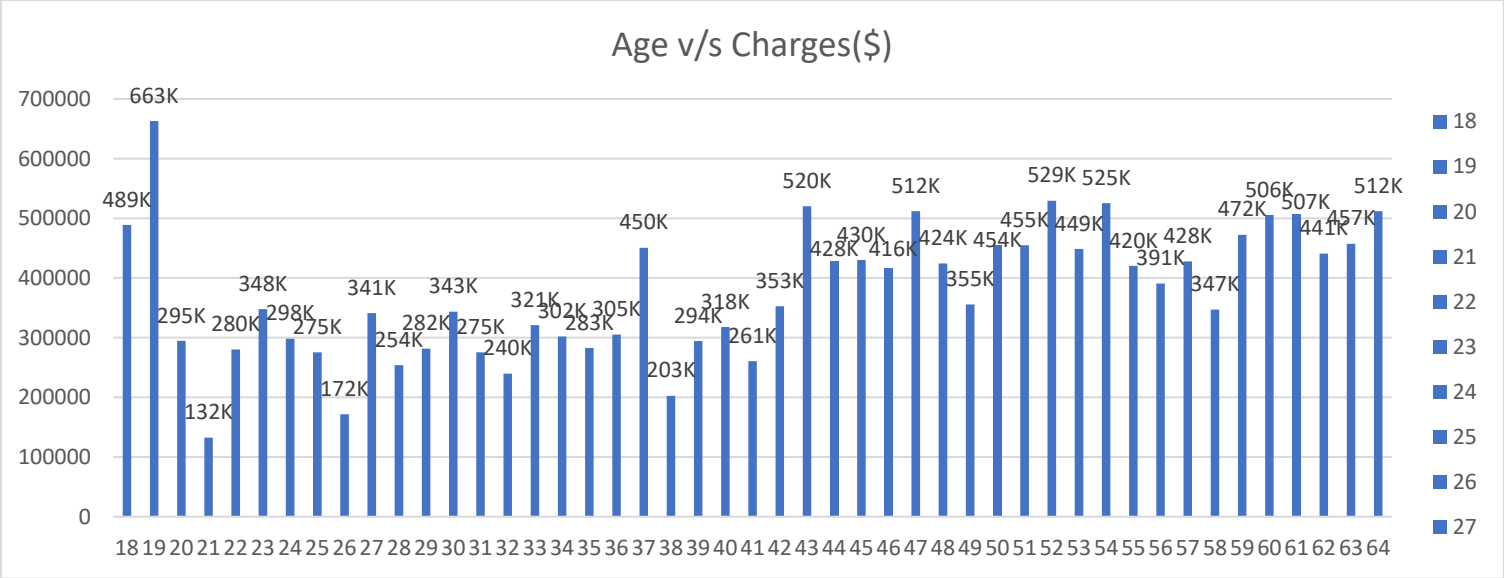
ii) Charges v/s Age:

Row Labels	Sum of charges($)
18	488949.0114
19	662857.8348
20	294631.2344
21	132453.0012
22	280362.1185
23	347754.9611
24	298144.4469
25	275474.2287
26	171747.1086
27	341171.6482

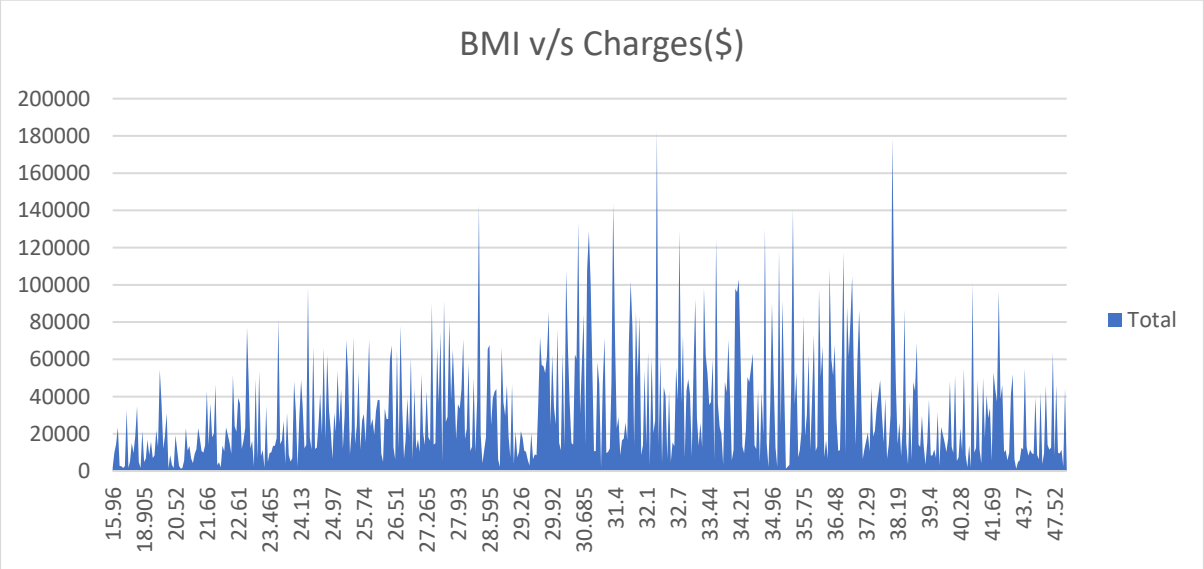
28	253937.2518
29	281614.2856
30	343415.9797
31	275318.4755
32	239727.8076
33	321139.8577
34	301951.7311
35	282679.5508
36	305111.9035
37	450497.7969
38	202568.3419
39	294456.0736
40	317850.7854
41	260651.1325
42	352648.0441
43	520216.5236
44	428203.7079
45	430075.7958
46	415935.1285
47	511965.9882
48	424342.5129
49	355488.1754
50	454227.0957
51	454785.4202
52	529431.8219
53	448586.0611
54	525239.3013
55	420278.1827
56	390663.4118
57	427626.8165
58	346973.2028
59	472396.7383
60	505526.6257
61	506562.525
62	440768.7012
63	457354.9646
64	512061.6784
Grand Total	**17755824.99**

Inference : The largest amount is claimed by customers whose age is 19 , i.e $663000, second largest claimed amount is $529000 belonging

to age 52, and third largest amount $525000 is claimed by age 54. Interestingly, the least sum of amount, $132000 is claimed by age 21.



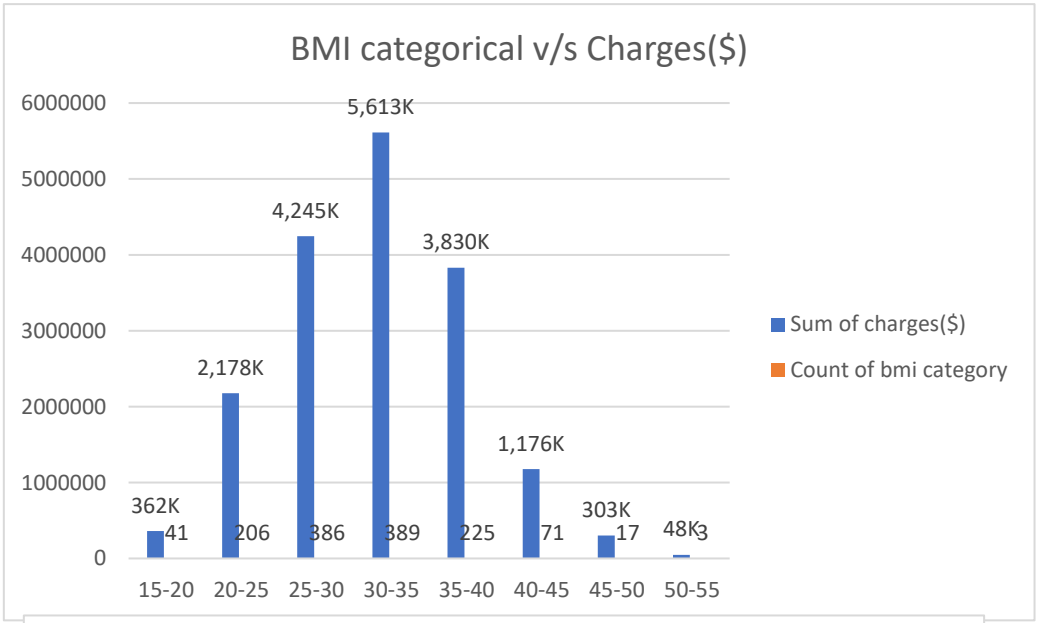
iii) Charges($) v/v BMI

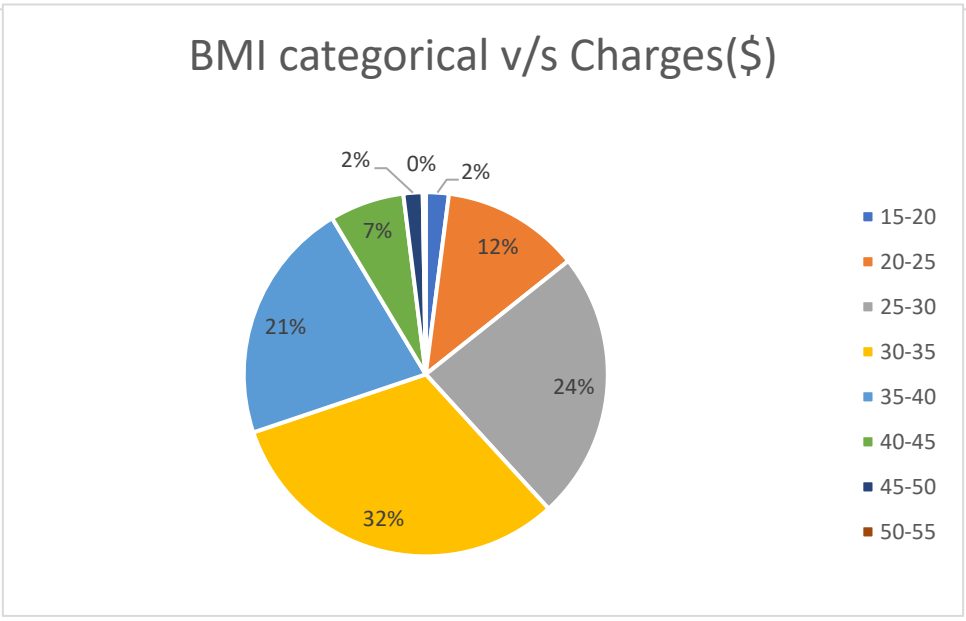


Since there are many Frequencies for bmi, we can group them in a category and analyse the data. i.e

BMI	Sum of charges($)
15-20	362381.0065
20-25	2177838.63
25-30	4245152.298
30-35	5613044.793
35-40	3830008.249
40-45	1176441.48
45-50	302855.6189

50-55	48102.9161
Grand Total	**17755824.99**

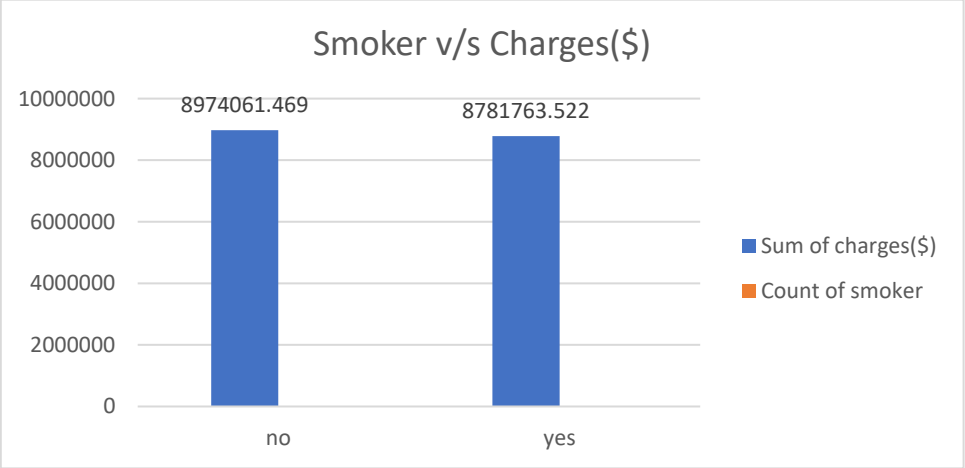




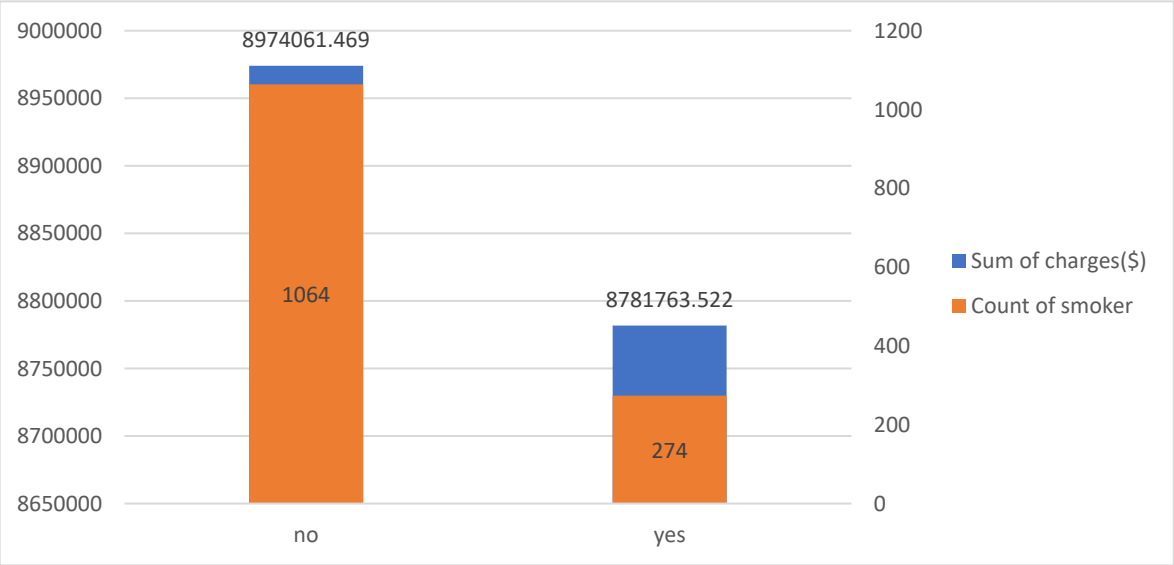
Inference : From the above sorted chart, it clearly denotes that BMI of customers ranging from 30-35 have claimed the highest sum($5613000), followed by BMI of 25-30 with a sum of amount,($4245000) , and third largest amount ($3830000) is of bmi group, 35-40.

iv) Charges for smokers v/s non -smokers

SMOKER	Sum of charges($)	Count of smoker
no	8974061.469	1064
yes	8781763.522	274
Grand Total	**17755824.99**	**1338**



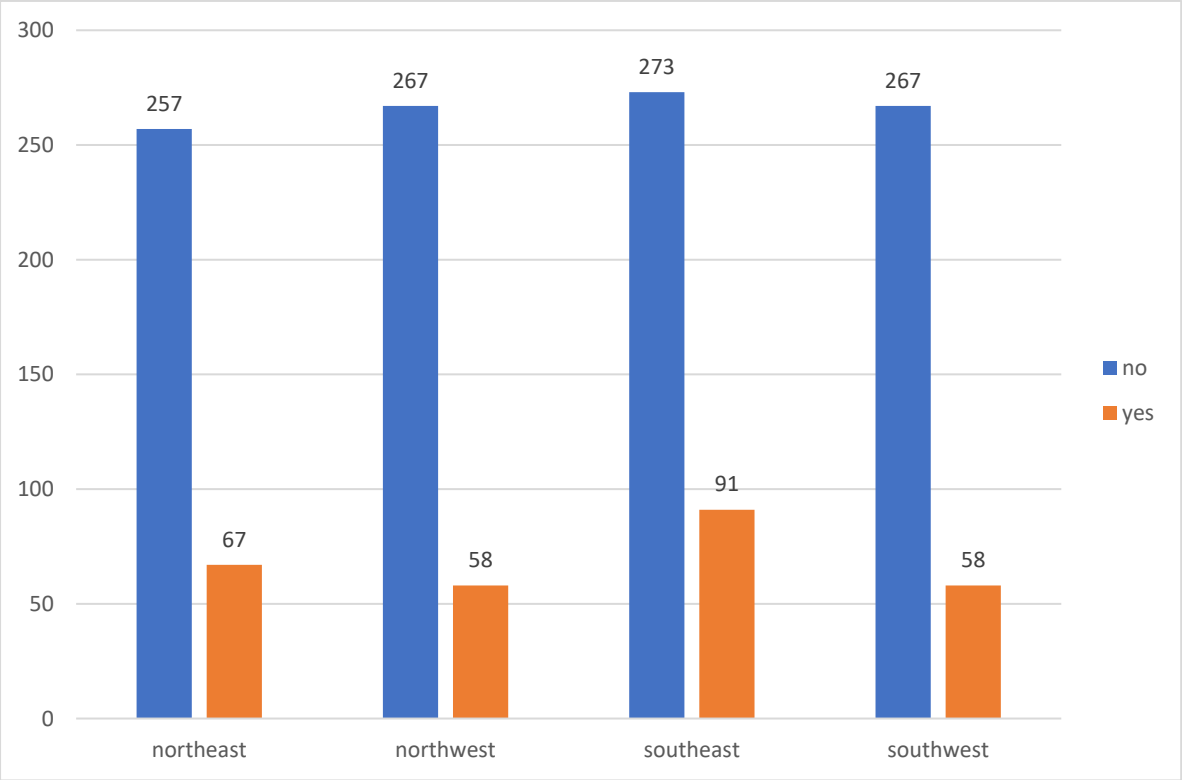
Inference: Although the graph shows that non-smokers have claimed much more than smokers, but we must inspect the ratio of non-smokers : smokers



Hence, after including the number of smokers and non - smokers , it is clear that even though smokers are less in numbers, the claimed amount that has a deficit of only $192297.95 w.r.t non-smokers. This indicates that Smokers tend to claim more amount than non-smokers.

1.d Region-wise smokers vs Non-smokers

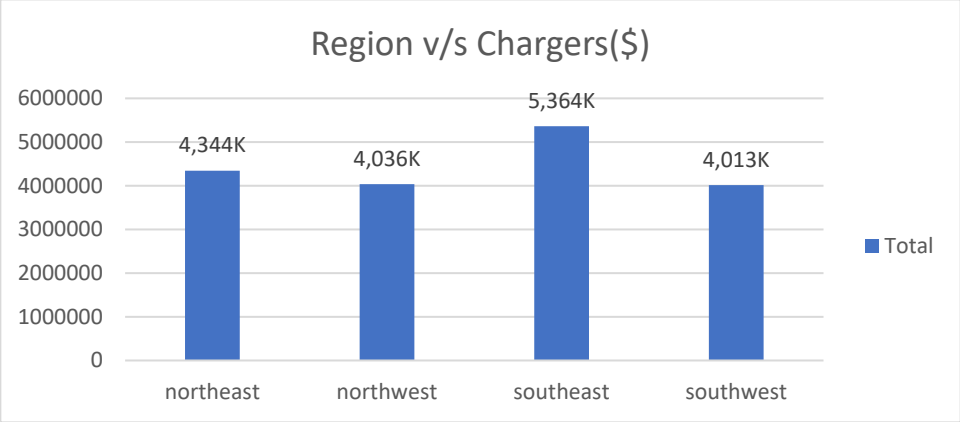
Count of smoker	Column Labels		
REGION	no	yes	Grand Total
northeast	257	67	324
northwest	267	58	325
southeast	273	91	364
southwest	267	58	325
Grand Total	**1064**	**274**	**1338**



Inference: from the above graph, we can infer that the non-smokers are distributed evenly across the regions. Southeast has an increased level smokers in that particular region. Whereas northwest and southwest have the least smokers.

1.e Region-wise charges for smokers vs non-smokers

Region	Sum of charges($)
northeast	4343668.583
northwest	4035711.997
southeast	5363689.763
southwest	4012754.648
Grand Total	**17755824.99**



From the above Chart we can infer that, customers living in southeast region have claimed more amount than any other region due to the fact that in this region, number of smokers are highest in southwest.

1.f Has charges got something to do with the number of dependents

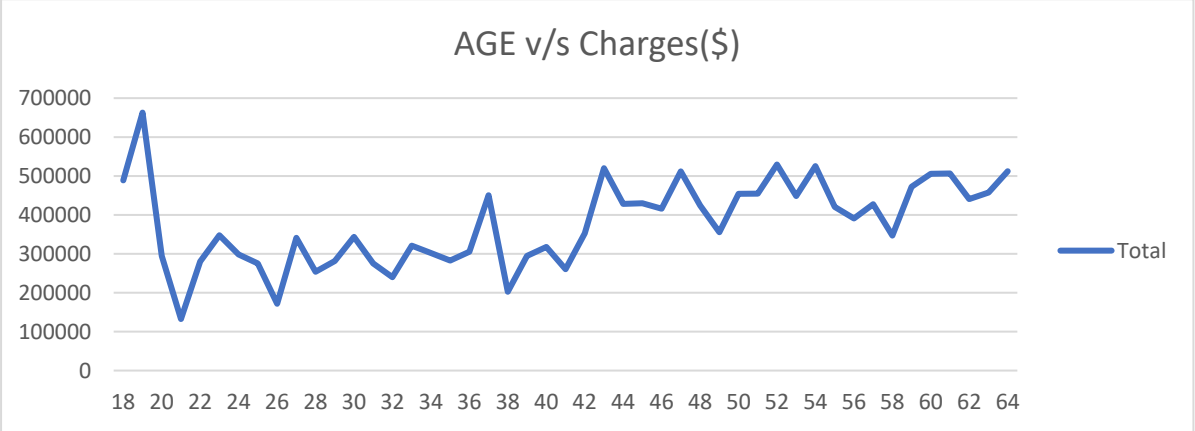
First, we shall do a regression model to analyse if there an any significant predictors.

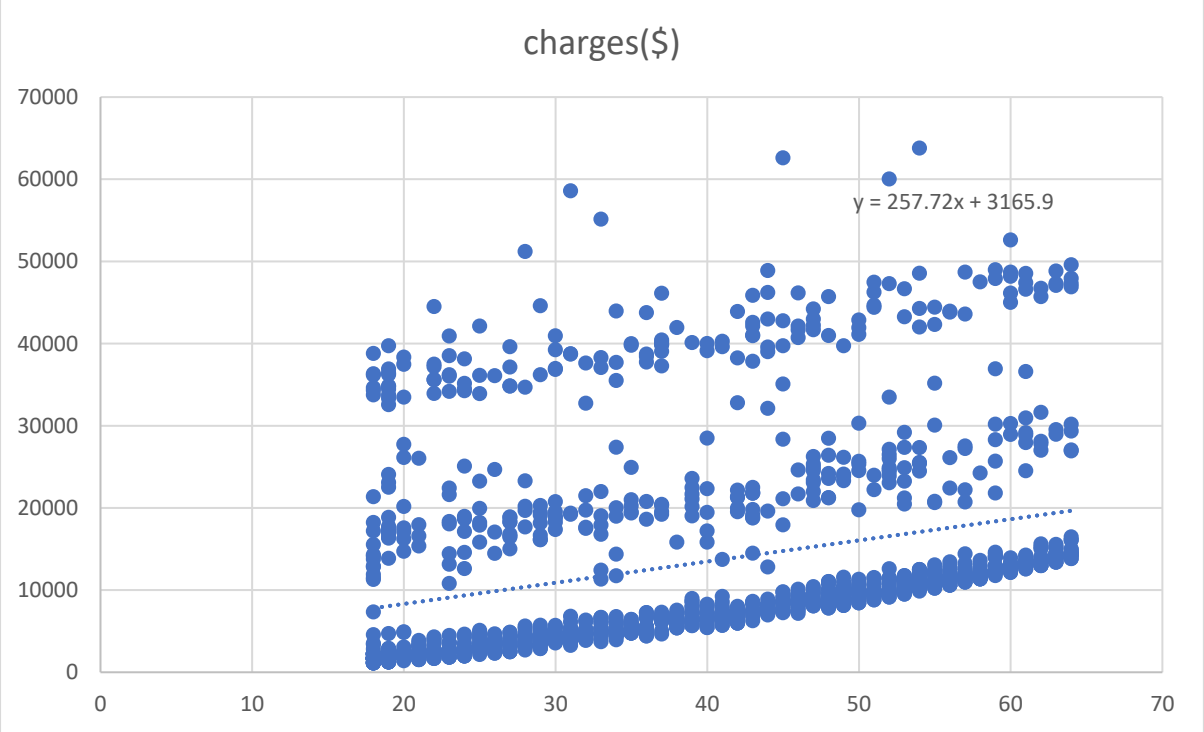
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-6916.24	1757.48	-3.93532	8.74E-05	-10364	-3468.52	-10364	-3468.52
age	239.9945	22.28888	10.76745	5.53E-26	196.2694	283.7195	196.2694	283.7195
children	542.8647	258.2413	2.102161	0.035726	36.26142	1049.468	36.26142	1049.468
bmi	332.0834	51.31046	6.47204	1.35E-10	231.4254	432.7414	231.4254	432.7414

Since the P-value of all the predictors are less than 0.05, we can infer that age, children, bmi are significant predictors of Charges($)

Now we shall analyse the same using pivot tables:

i) Age and charges:

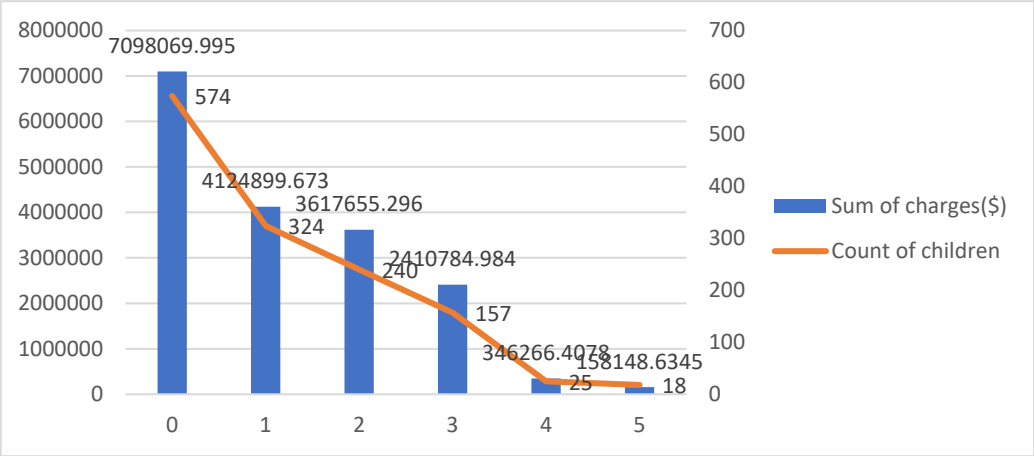


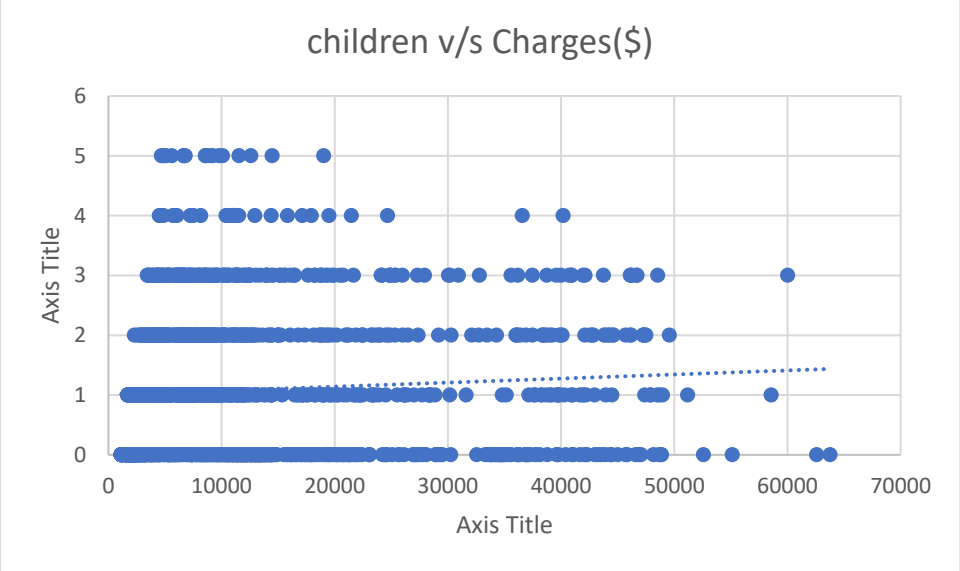


The above graph shows that there is a slight increase in charges as age increases, which is not strong enough.

ii) Children :

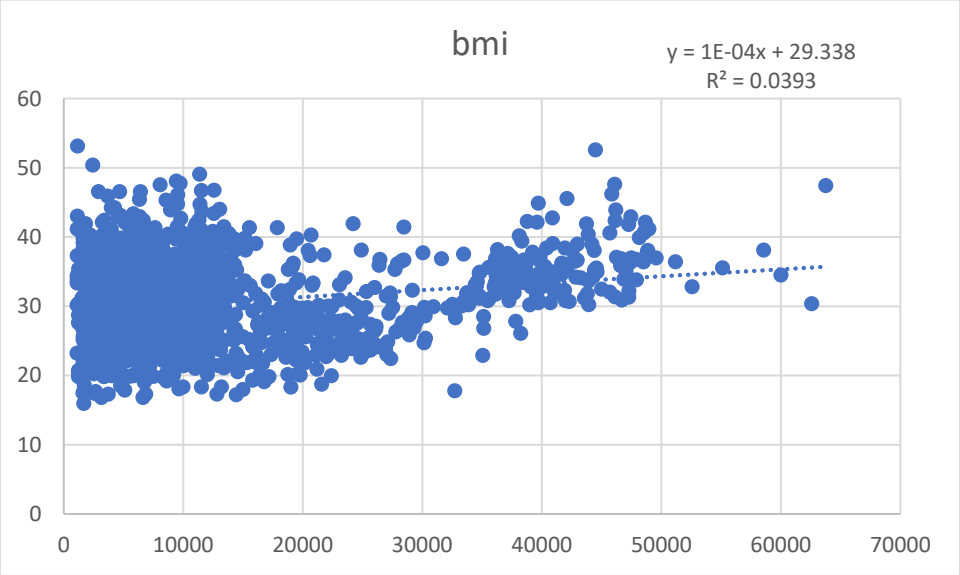
Children	Sum of charges($)	Count of children
0	7098069.995	574
1	4124899.673	324
2	3617655.296	240
3	2410784.984	157
4	346266.4078	25
5	158148.6345	18
Grand Total	**17755824.99**	**1338**





The Combination chart showes that customers with 0 children have claimed the most. But that is due to the huge sample size of number of customers with 0 children. Hence a scatter plot is done to verify the relationship, it shows that there is not strong relation between children and charges.

iii) Bmi and charges:



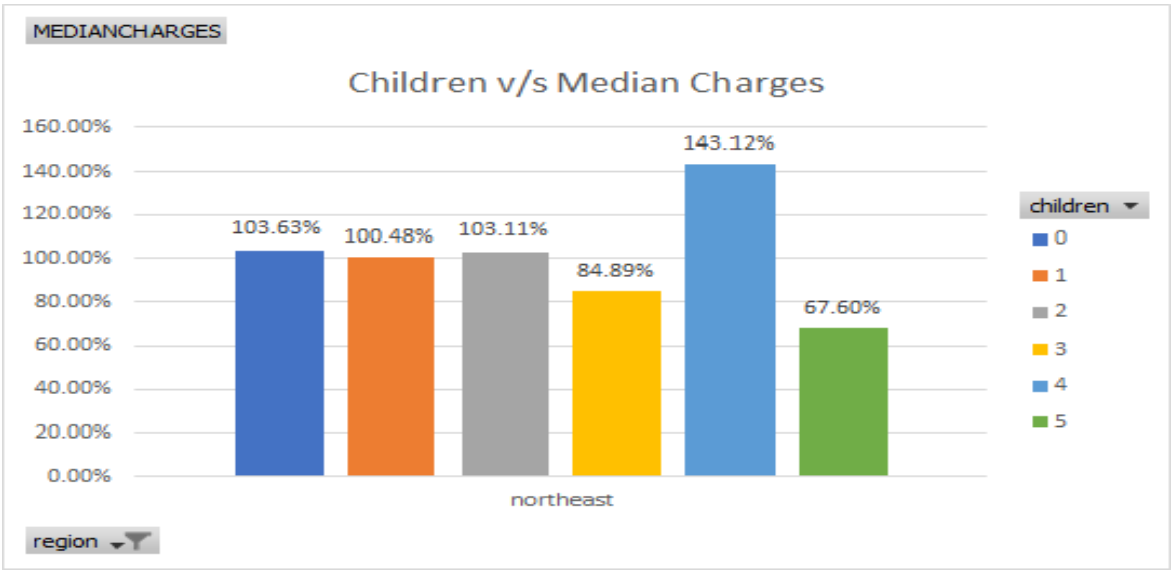
The above scatter plot shows no real relation between bmi and charges hence no strong evidence is found to conclude that charges is dependent on either of age, bmi, children.

1.g Dependent charges analysis, region-wise

i) Northeast region:

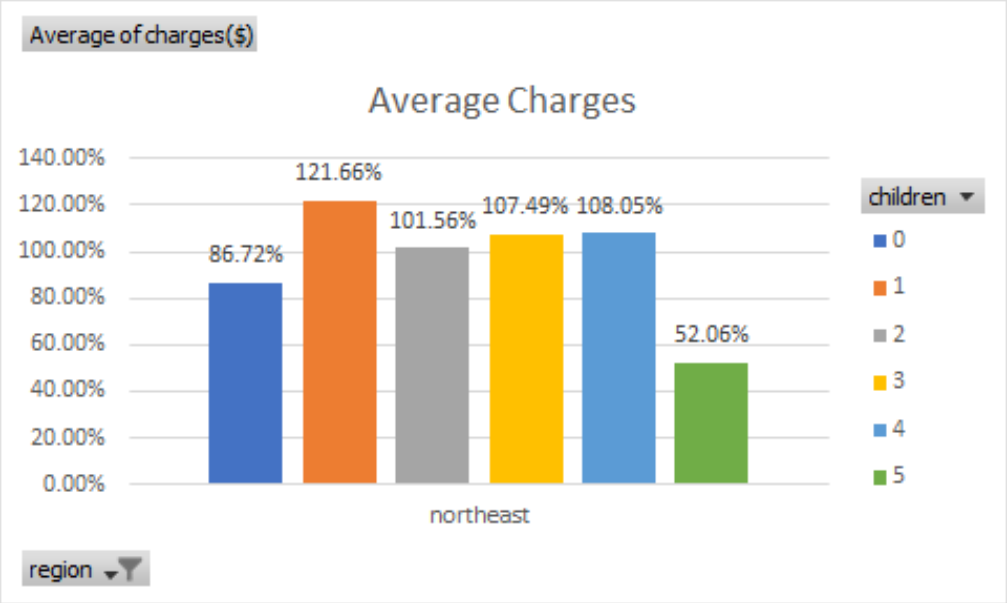
Analysis w.r.t age and median of charges:

MEDIANCHARGES	Column Labels						
Region	0	1	2	3	4	5	Grand Total

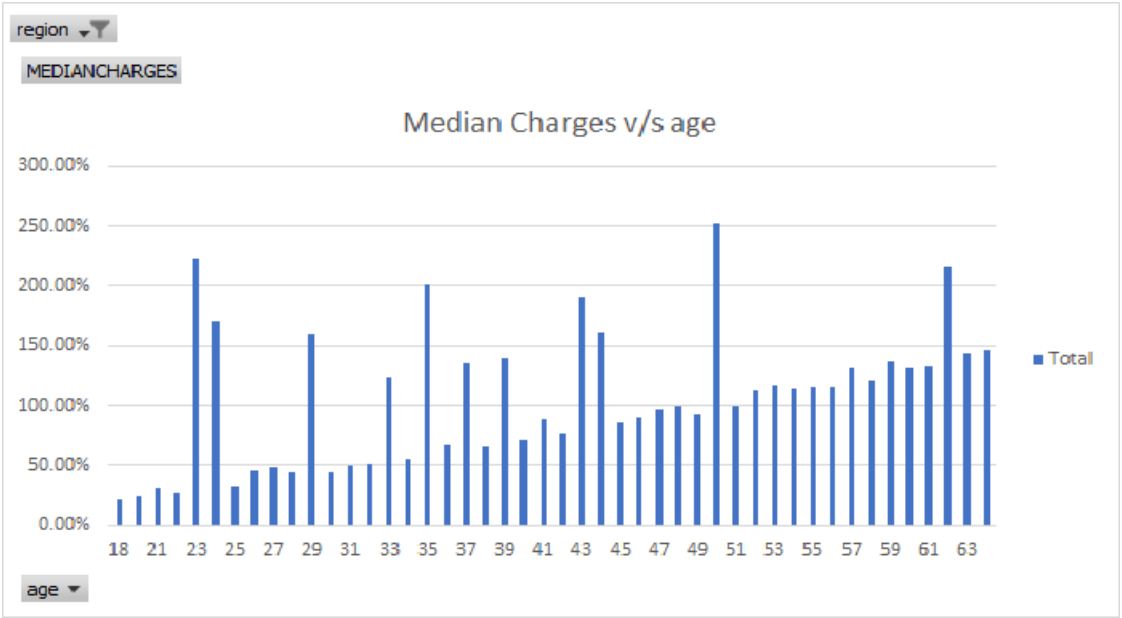


northeast	103.63%	100.48%	103.11%	84.89%	143.12%	67.60%	100.00%
Grand Total	**103.63%**	**100.48%**	**103.11%**	**84.89%**	**143.12%**	**67.60%**	**100.00%**

In northeast region, customers with 4 children are claiming more amount than other number of children i.e 43% more than median value. Interestingly, customers with 5 children have claimed the least.



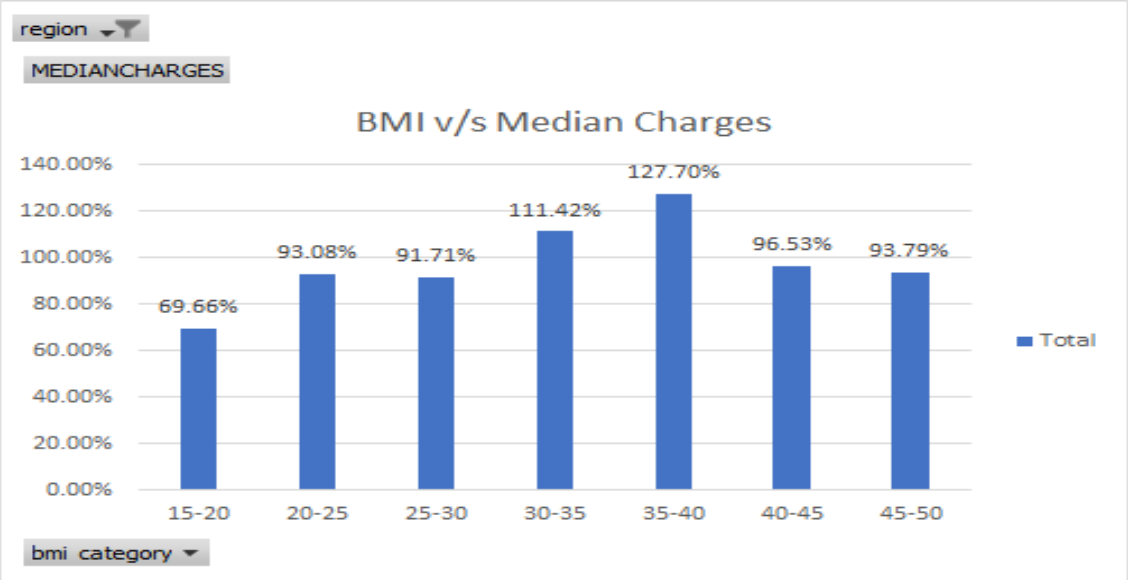
Analysis w.r.t Median Charges and age:



According to the chart, there are many customers whose claimed amount to well beyond median. Customer's age whose claimed amount more than double the amount i.e (>200%) are 23,35,50,62

(note: we are considering median of charges($) since there are clear outliers in charges variable.

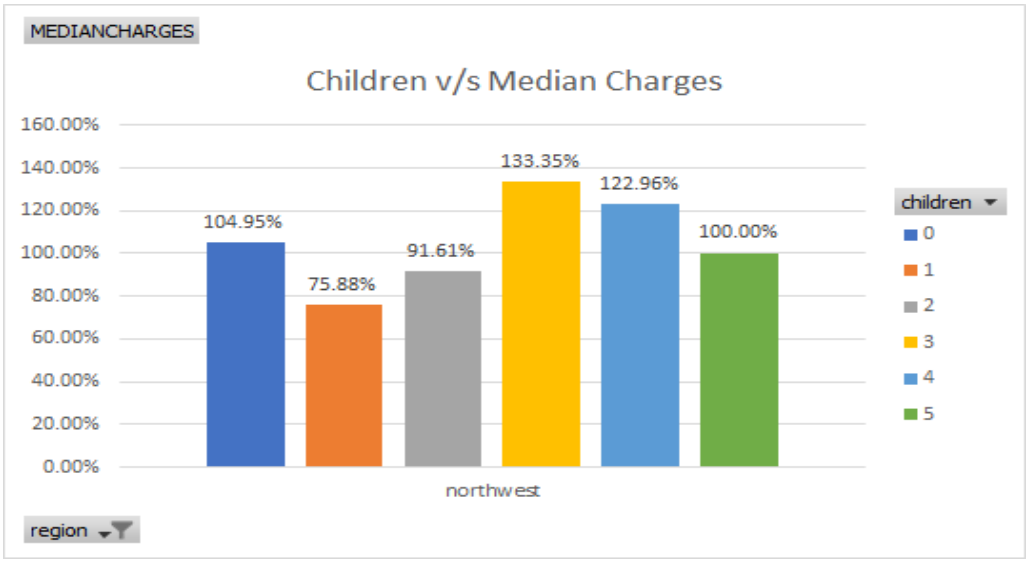
Analysis w.r.t bmi and median of charges:



Bmi of category 35-40 have claimed noticeably more amount than median value of charges.

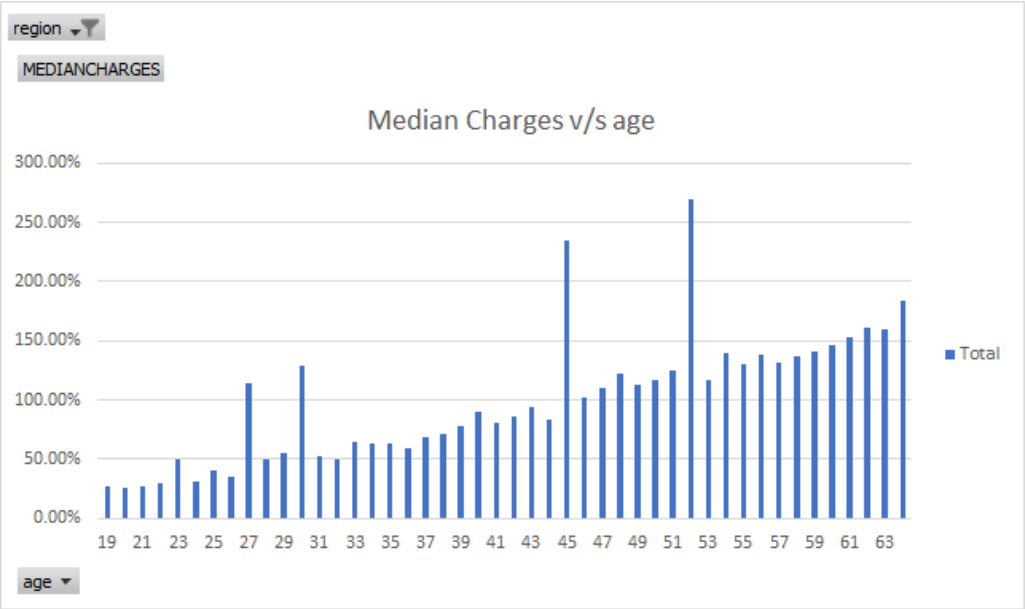
ii)Northwest region:

Analysis w.r.t children and Median Charges :



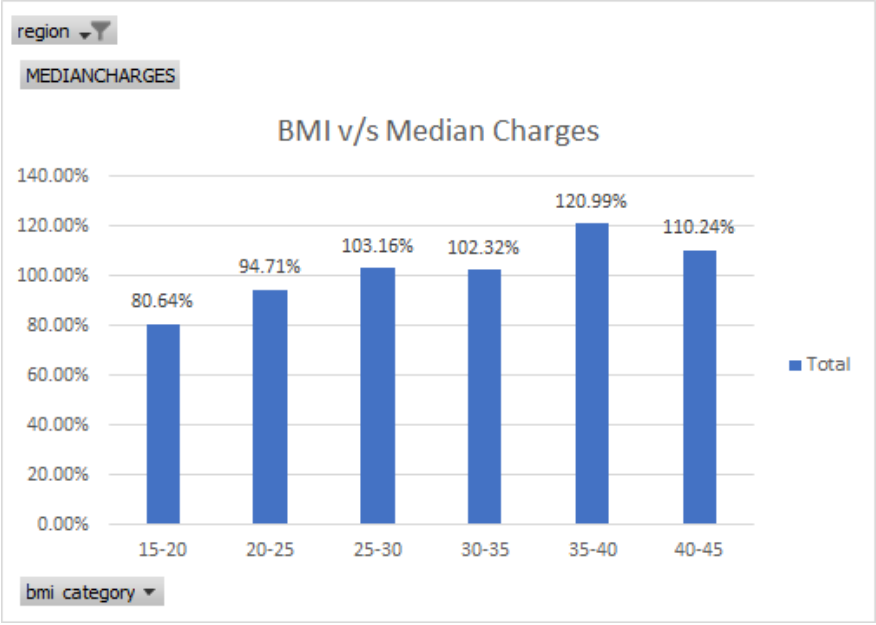
According to the chart, customers having children 3,4 have claimed more amount than median value of charges.

Analysis w.r.t age and median charges:



According to the chart, customers whose ages are 45,52 have claimed more than twice the median value of the charges($)

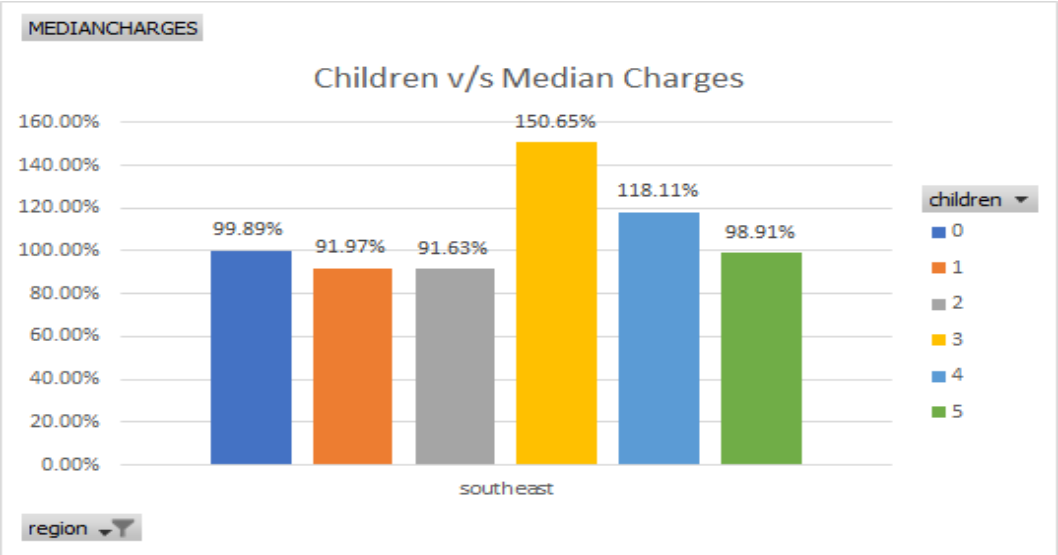
Analysis w.r.t bmi and median charges:



The above chart infers that bmi of category 35-40 have clamied more amount than the median value of charges($).

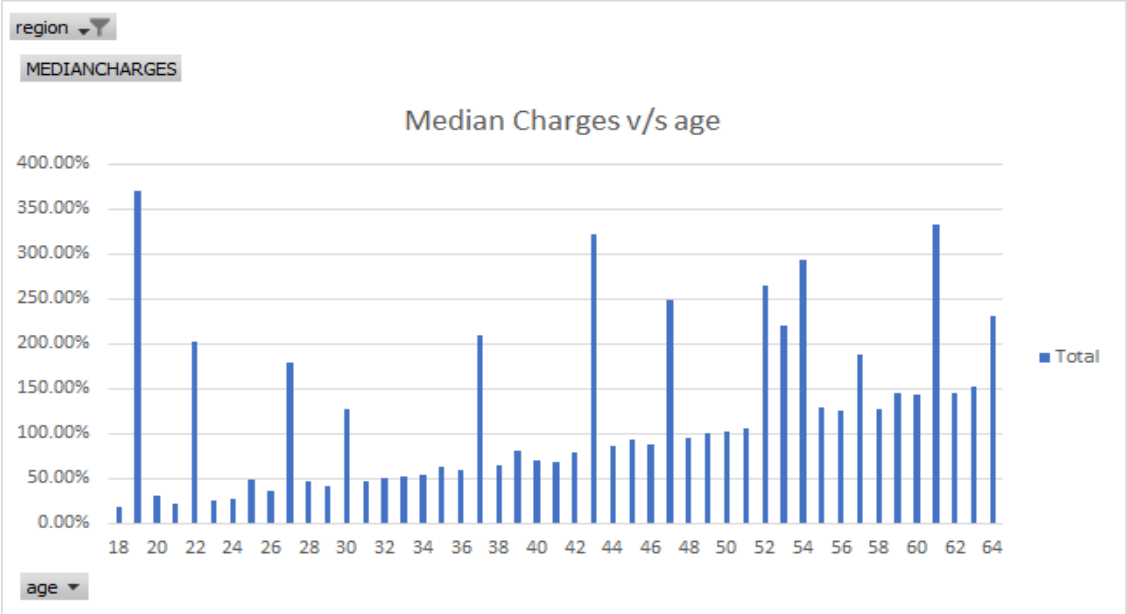
iii)Southeast region:

Analysis w.r.t Children and charges:



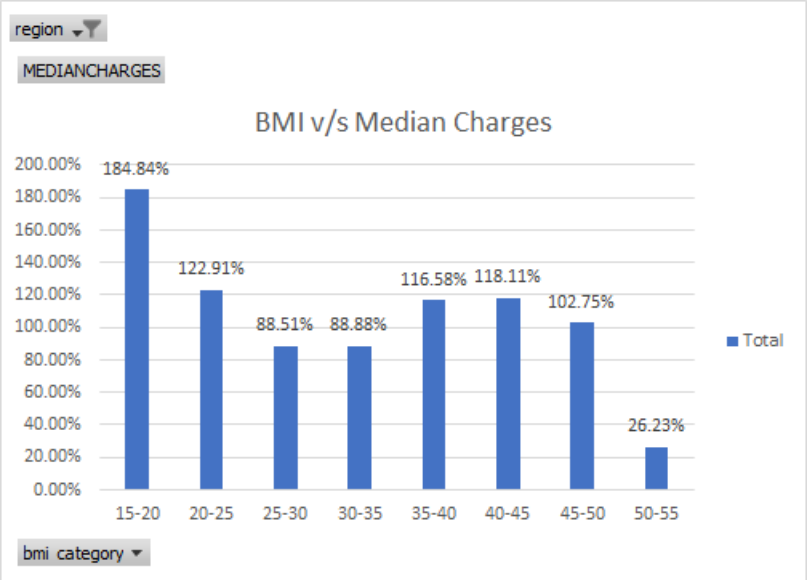
The above chart infers that customers who have children of 3 members have claimed more amount than median value of charges.

Analysis w.r.t Age and Median of Charges:



In the above chart, there are a lot of ages who have claimed significantly more than median. They are: 19,22,37,43,47,51,53,54,61,64.

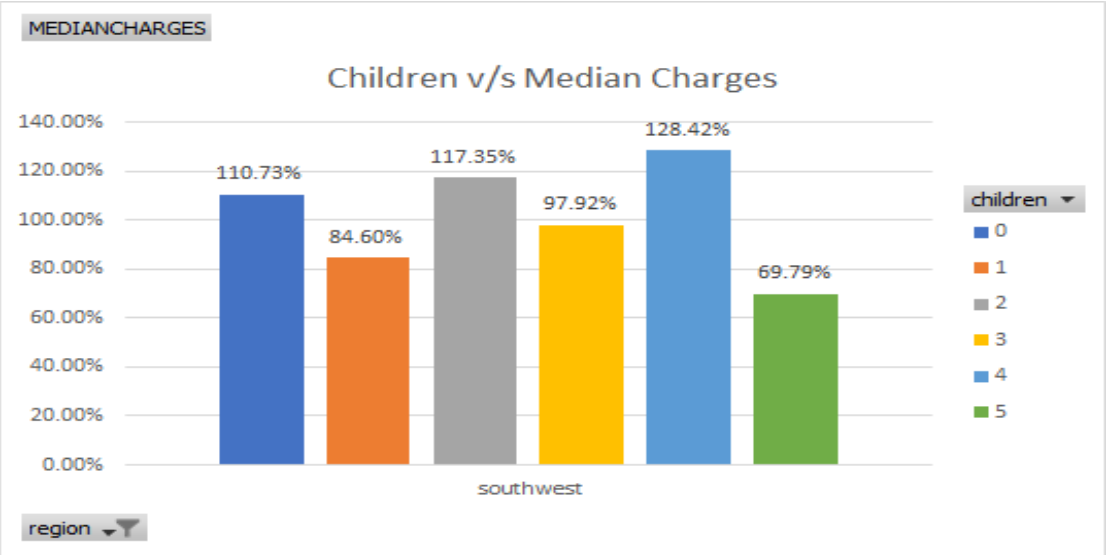
Analysis of bmi and median of charges:



According to the data, bmi whose range is between 15-20 have spent well more the median of charges($).

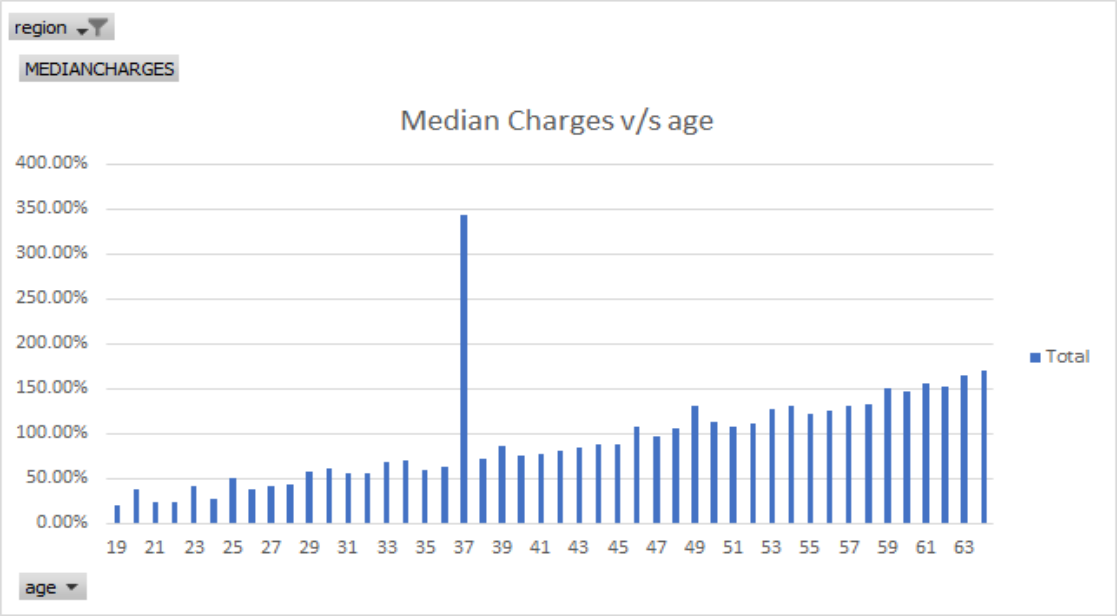
iv)Southwest region:

Analysis w.r.t children and charges:



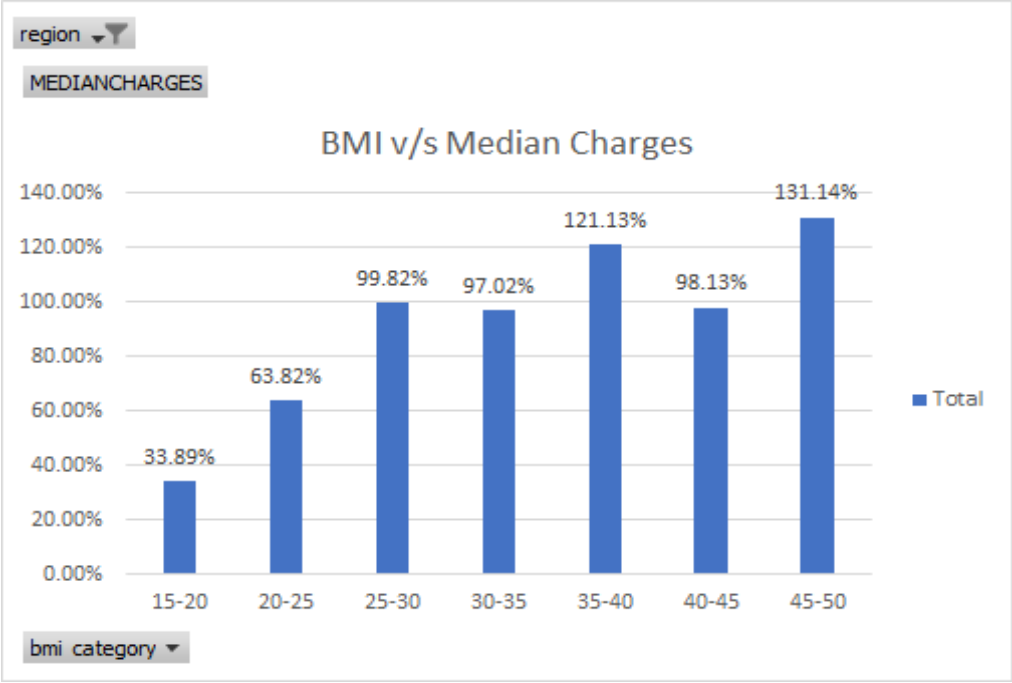
The above chart indicates that customers having 2,4 children have spent noticeably more than charges.

Analysis w.r.t age and median charges:



Customers whose age is 37 have claimed significantly large amount.

Analysis w.r.t bmi and median charges:



The above chart depicts that bmi of range 35-40 and 45-50 have claimed more amount than median value of charges($).

1.i NOTE: THE INFERENECE FOR QUESTION 1.B HAS BEEN GIVEN BELOW EVERY CHART RESPECTIVELY.

1.J NOTE: THE INFERENECE FOR QUESTION 1.C HAS BEEN GIVEN BELOW EVERY CHART RESPECTIVELY.

2.A,B,C:

NOTE: THE TABLE HAS BEEN UPDATED IN EXCEL FILE.

SS HAS BEEN ATTACHED BELOW SINCE COPY PASTING THE WHOLE TABLE IN WORD TAKES A LOT OF PAGES.

	A	B	C	D	E	F	G	H	I
1	age	children	bmi	smokerNu	southeast	southwest	northwest	sexNumeri	charges($)
2	18	0	53.13	0	1	0	0	1	1163.463
3	22	1	52.58	1	1	0	0	1	44501.4
4	23	1	50.38	0	1	0	0	1	2438.055
5	58	0	49.06	0	1	0	0	1	11381.33
6	46	2	48.07	0	0	0	0	0	9432.925
7	52	1	47.74	0	1	0	0	1	9748.911
8	37	2	47.6	1	0	1	0	0	46113.51
9	47	1	47.52	0	1	0	0	1	8083.92
10	54	0	47.41	1	1	0	0	0	63770.43
11	52	5	46.75	0	1	0	0	0	12592.53
12	54	2	46.7	0	0	1	0	0	11538.42
13	32	2	46.53	0	1	0	0	1	4686.389
14	37	3	46.53	0	1	0	0	1	6435.624
15	26	1	46.53	0	1	0	0	1	2927.065
16	43	0	46.2	1	1	0	0	0	45863.21
17	50	1	46.09	0	1	0	0	0	9549.565
18	27	2	45.9	0	0	1	0	1	3693.428
19	25	2	45.54	1	1	0	0	1	42112.24
20	39	2	45.43	0	1	0	0	1	6356.271
21	47	1	45.32	0	1	0	0	0	8569.862
22	19	0	44.88	1	1	0	0	1	39722.75
23	50	1	44.77	0	1	0	0	1	9058.73
24	50	0	44.745	0	0	0	0	0	9541.696
25	52	3	44.7	0	0	1	0	0	11411.69
26	32	0	44.22	0	1	0	0	0	3994.178
27	30	2	44.22	0	1	0	0	1	4266.166
28	61	0	44	0	0	1	0	0	13063.88
29	46	3	43.89	0	1	0	0	1	8944.115

3. Regression model:

Regression Statistics	
Multiple R	0.867
R Square	0.751
Adjusted R Square	0.749
Standard Error	6062.102
Observations	1338.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8.000	147234688724.445	18404336090.556	500.811	0.000
Residual	1329.000	48839532843.922	36749084.156		
Total	1337.000	196074221568.367			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11938.539	987.819	12.086	0.000	13876.393	10000.684	13876.393	10000.684
age	256.856	11.899	21.587	0.000	233.514	280.199	233.514	280.199
children	475.501	137.804	3.451	0.001	205.163	745.838	205.163	745.838
bmi	339.193	28.599	11.860	0.000	283.088	395.298	283.088	395.298
sexNumerical	-131.314	332.945	-0.394	0.693	-784.470	521.842	-784.470	521.842
smokerNumerical	23848.535	413.153	57.723	0.000	23038.031	24659.038	23038.031	24659.038
southeast	-1035.022	478.692	-2.162	0.031	-1974.097	-95.947	-1974.097	-95.947
southwest	-960.051	477.933	-2.009	0.045	-1897.636	-22.466	-1897.636	-22.466
northwest	-352.964	476.276	-0.741	0.459	-1287.298	581.370	-1287.298	581.370

Inference:

a) Adjusted R Square: The true R square value calculated by negating the insignificant variable with does not make sense in achieving a good/bad R square value and is always less than R square value. Here, the Adjusted R square is 0.749, which tells us that excel can truly predict the data which is 74.9% correct.
b) Intercept: It is the compensation value for the best fit line.
c) P-value: A predictor is said to be significant if its p-value is less than 0.05 In this case, the value is less than 0.05 for predictors: age, children, bmi, smoker-numerical, southeast and southwest.
d) Equation for best fit line:

age*256.856 + children * 475.501 + bmi*339.193 + sexNumerical * (-131.314) + smokerNumerical*23848.535 + Southeast * (-1035.022) + southwest * (-960.051) + northwest * (-352.964) + (-11938.539)

New regression model using only significant predictors:

Regression Statistics	
Multiple R	0.866
R Square	0.751
Adjusted R Square	0.750
Standard Error	6059.146
Observations	1338.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6.000	147208878053.492	24534813008.915	668.282	0.000
Residual	1331.000	48865343514.876	36713255.834		
Total	1337.000	196074221568.367			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12165.382	949.538	-12.812	0.000	-14028.137	-10302.628	-14028.137	-10302.628
age	257.006	11.889	21.617	0.000	233.683	280.330	233.683	280.330
children	471.544	137.656	3.426	0.001	201.498	741.590	201.498	741.590
bmi	338.641	28.554	11.860	0.000	282.625	394.657	282.625	394.657
smokerNumerical	23843.875	411.659	57.921	0.000	23036.304	24651.446	23036.304	24651.446
southeast	-858.470	415.206	-2.068	0.039	-1672.998	-43.941	-1672.998	-43.941
southwest	-782.745	413.756	-1.892	0.059	-1594.430	28.940	-1594.430	28.940

Inference:

a) This model is clearly better than the previous model in terms of R square, Adjusted R square and P-Value (only Considered significant predictors in the current model)

b) The Adjusted R square in the previous model was 0.749 i.e 74.9% , whereas in the current model, the Adjusted R square is 0.750 i.e 75%. Hence we can conclude that the Adjusted R square value has improved by 0.1% which is not a significant improvement.

c) Equation of current regression model:
age*257.006 + children * 741.511 + bmi * 338.641 + smokerNumerical * 23843.875 + southeast * (-858.470) + southwest * (-782.745) + (-12165.382)

Summary Statistics:

	age	children	bmi	mokerNumerica	southeast	southwest	northwest	sexNumerical	charges($)
Mean	39.207	1.095	30.663	0.205	0.272	0.243	0.243	0.505	13270.422
Standard Error	0.384	0.033	0.167	0.011	0.012	0.012	0.012	0.014	331.067
Median	39.000	1.000	30.400	0.000	0.000	0.000	0.000	1.000	9382.033
Mode	18.000	0.000	32.300	0.000	0.000	0.000	0.000	1.000	1639.563
Standard Deviation	14.050	1.205	6.098	0.404	0.445	0.429	0.429	0.500	12110.011
Sample Variance	197.401	1.453	37.188	0.163	0.198	0.184	0.184	0.250	146652372.153
Kurtosis	-1.245	0.202	-0.051	0.146	-0.950	-0.560	-0.560	-2.003	1.606
Skewness	0.056	0.938	0.284	1.465	1.026	1.200	1.200	-0.021	1.516
Range	46.000	5.000	37.170	1.000	1.000	1.000	1.000	1.000	62648.554
Minimum	18.000	0.000	15.960	0.000	0.000	0.000	0.000	0.000	1121.874
Maximum	64.000	5.000	53.130	1.000	1.000	1.000	1.000	1.000	63770.428
Sum	52459.000	1465.000	41027.625	274.000	364.000	325.000	325.000	676.000	17755824.991
Count	1338.000	1338.000	1338.000	1338.000	1338.000	1338.000	1338.000	1338.000	1338.000

Inference:

i)Mean and Median: if the mean and median are close to each other, it implies that the data is spread / distributed evenly. Here, the fields whose mean and median are far apart is charges($) hence we have to consider median value of charges for its analysis.

ii)Standard deviation: It gives the deflection of variable from its median value. Here the variable whose standard deviation is more than median value is Charges($).

iii)Kurtosis: It depicts the shape (peak) of the distribution. Ideally kurtosis must be 0 for normal distribution, Negative value indicates that the curve is flat, positive value indicates that it has a sharp peak. Here the variables whose kurtosis values are far from 0 are: negative kurtosis – age, southeast, southwest, northwest, sexNumerical. Positive kurtosis : Charges($)

iv)Skewness: It depicts the shape(spread) of the distribution. Ideally skewness must be 0 for normal distribution , negative skewness indicates that the data is trailing off on left , positive skewness indicates that data is trailing off on right. Here, variables whose skewness is far apart from 0 are: children, smokerNumerical, southeast, southwest, northwest, Charges($).

Conclusion:

The insurance company must increase the premium amount for following customers:

i) Smoker

ii) Children: Northeast – 4
Northwest – 3,4
Southeast – 3,4
Southwest – 3,4

iii) BMI: Northeast – 35-40
Northwest – 35-40
Southeast – 15-20,35-40
Southwest – 35-40,45-50