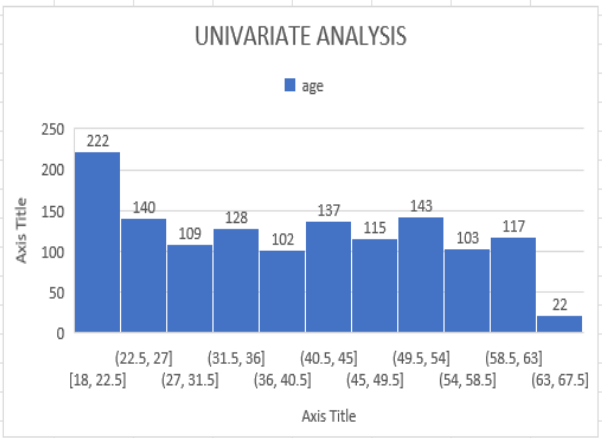
1. Perform the Exploratory Data Analysis on the data.

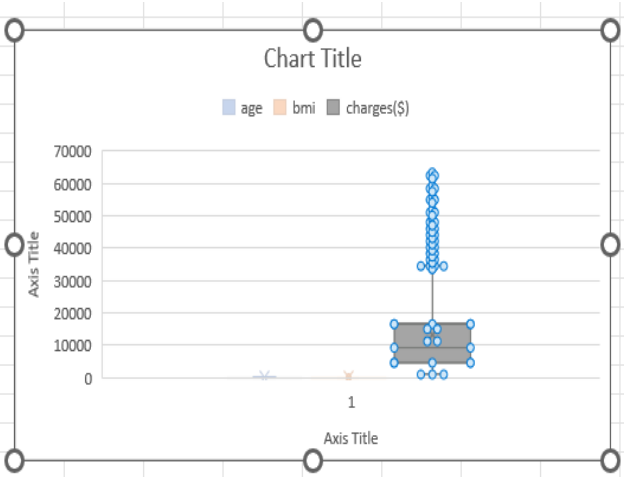
a) Identify the categorical and continuous variables

categorical variables are - sex, smoker, region, children.

continuous variable is - age, BMI, charges.

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





	age	bmi	charges($)
age	1.00		
bmi	0.11	1.00	
charges($)	0.30	0.20	1.00

Firstly, select age, BMI, charges move into insert select histogram and box plot and plot into excel.

The above corelation analysis is explanation is below.

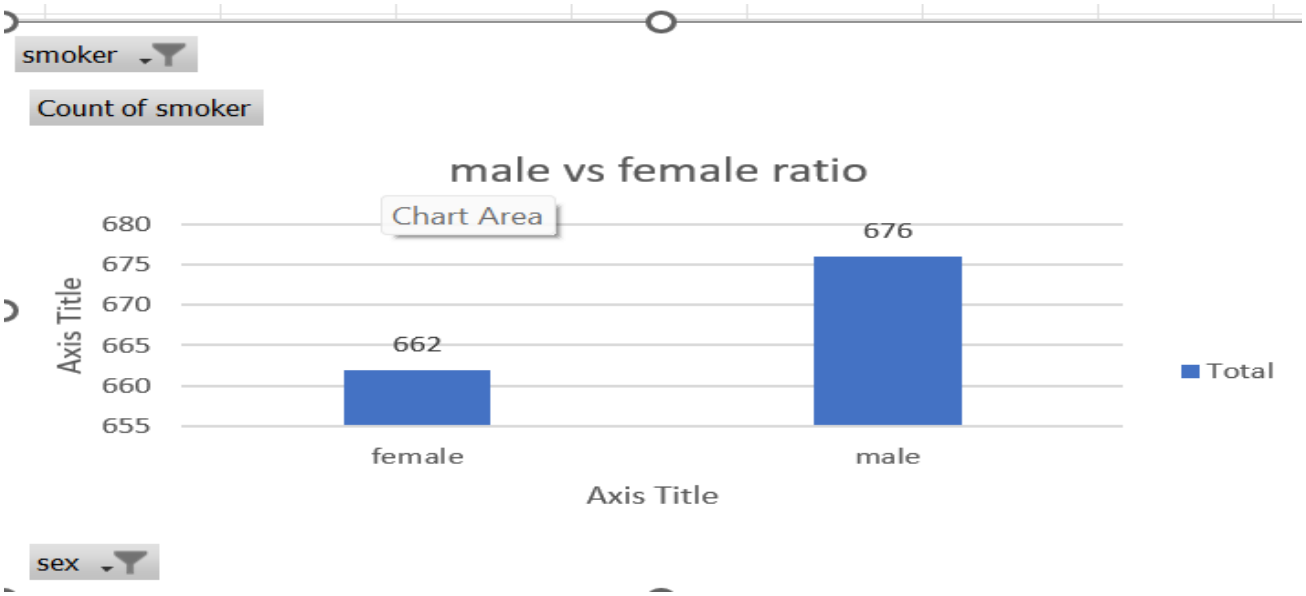
Select data analysis tool kit select correlation BMI, age, charges then ok.

BMI and age are positively correlated

Charges and BMI, age both are strongly correlated. therefore, there are no negative correlations.

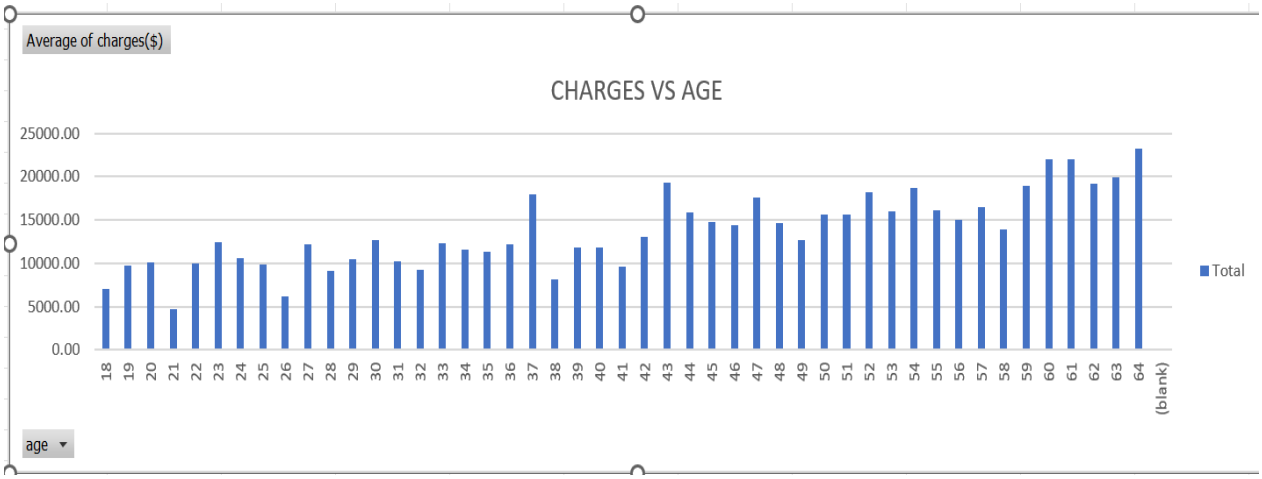
c) Make relevant Pivot tables and charts for:

I. Male/Female ratio and share information on which gender has more smokers.



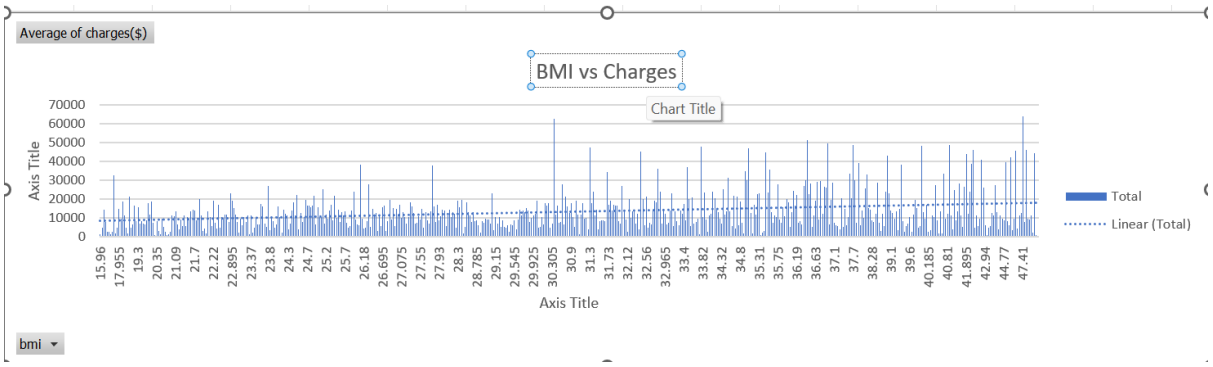
The above graph represents that male ratio as high ratio of 676 members are smokers.

ii. Charges vs Age



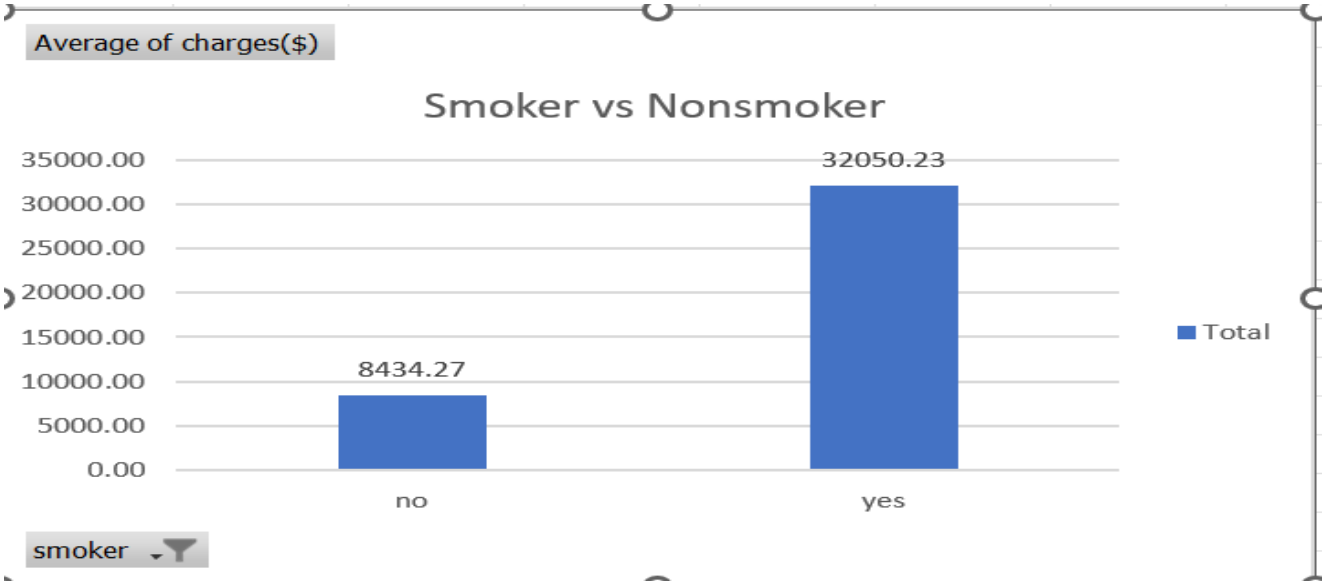
The above graph represents that age of 64 are more charged when compared to other age persons.

iii. Charges vs BMI



The above graph represents that the BMI between 30.2 & 30.685, between 45.32 & 47.52 for these criteria of BMI charges more when compared to other BMI'S.

iv. Charges for Smokers vs Non-smokers



The above graph represents that the more charges for smokers whereas they are into more smoking aspects they have certain impacts in health issues, so they have more charges for them.

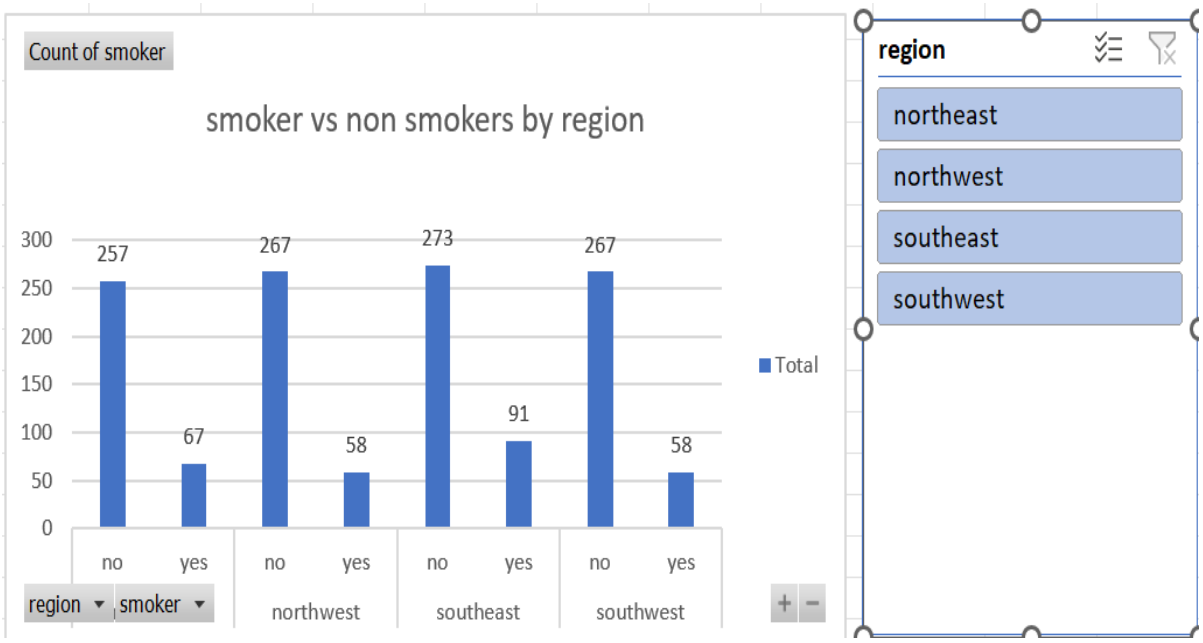
d) Region-wise smokers vs Non-smokers analysis with one or more pivot table and charts

I selected the insert -> pivot chart.

I have taken region, smoker.

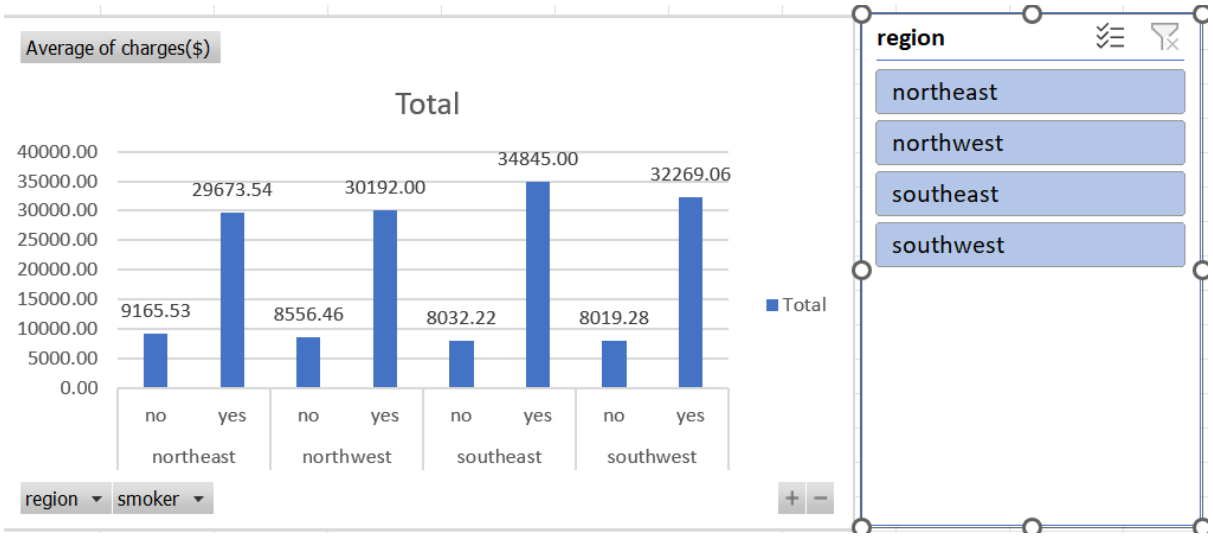
I selected insert slicer for region.

We can select in inside the insert slider individual region we can plot graph individually.



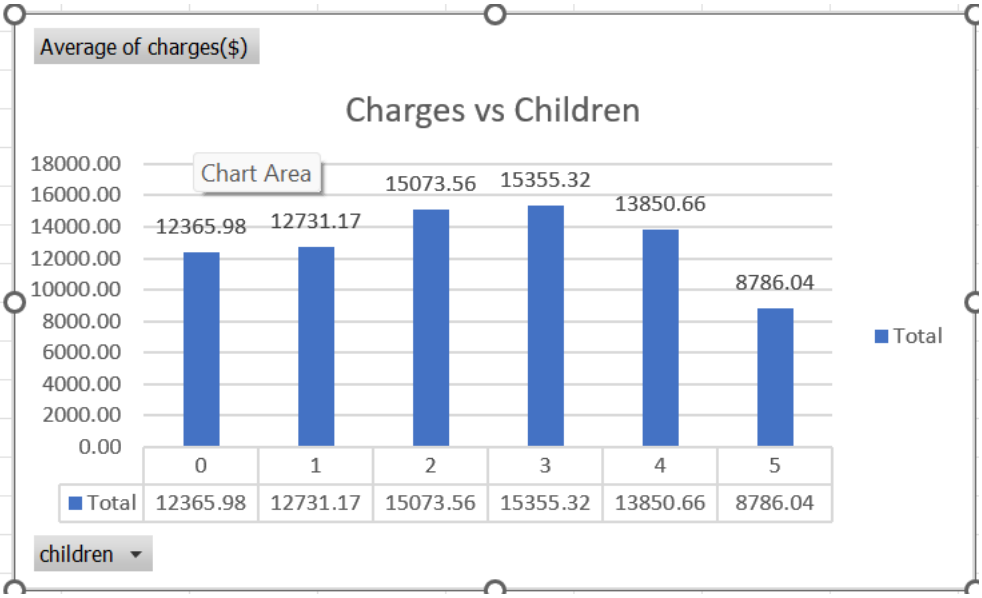
The above graph represents that smoker vs non smokers based on region the more smokers are in southeast is 273 regions compare to other regions.

e) Region-wise charges for smoker's vs non-smokers



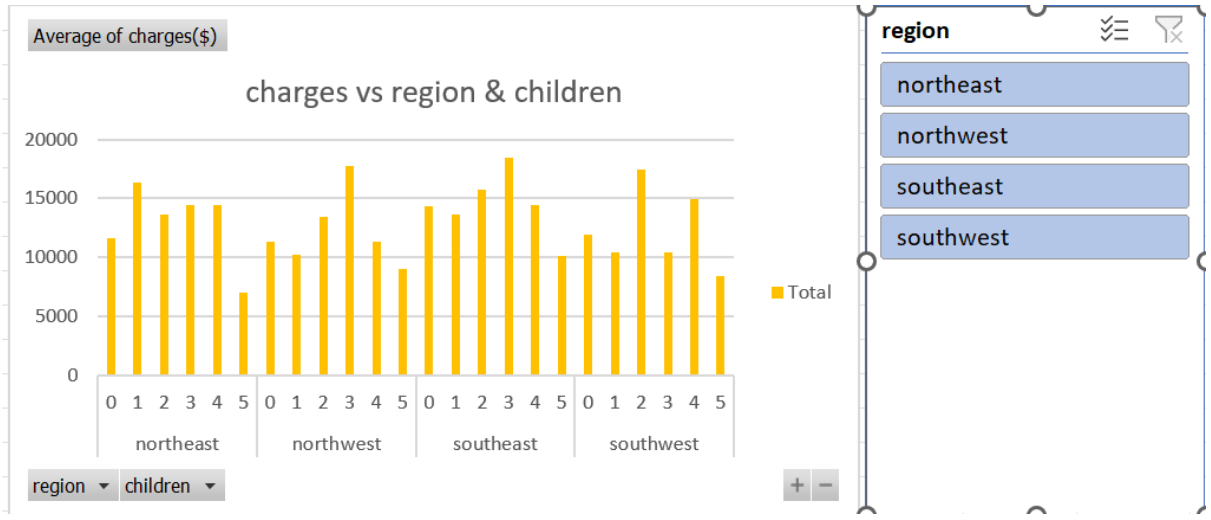
The above graph represented as southeast region as more smokers and more charges are by southeast .so I used to insert slicer which we can plot individual graph for every region.

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



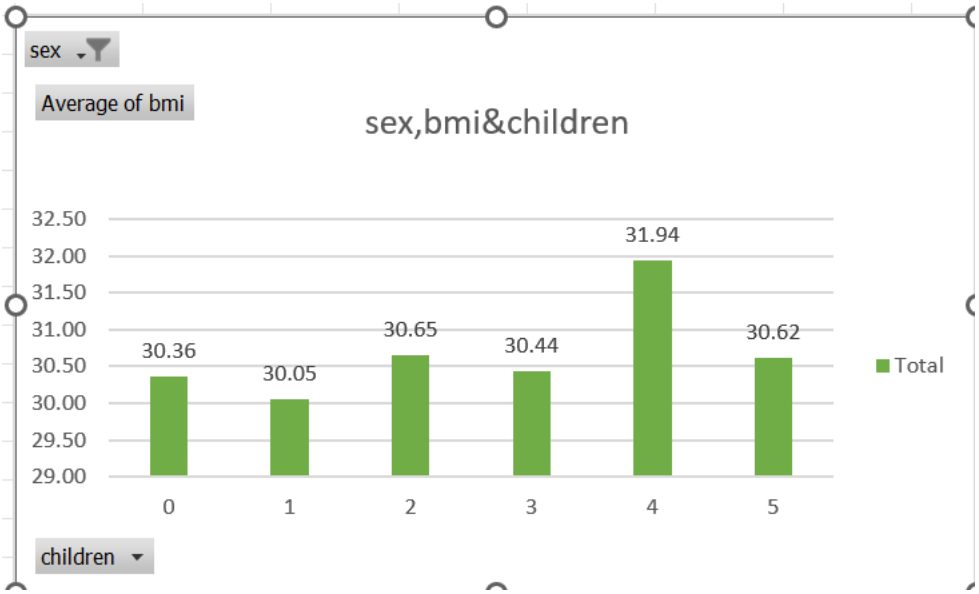
The above graph represents those 3 children of the customer has its charges of 15355.2 charges its high in the graph.

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

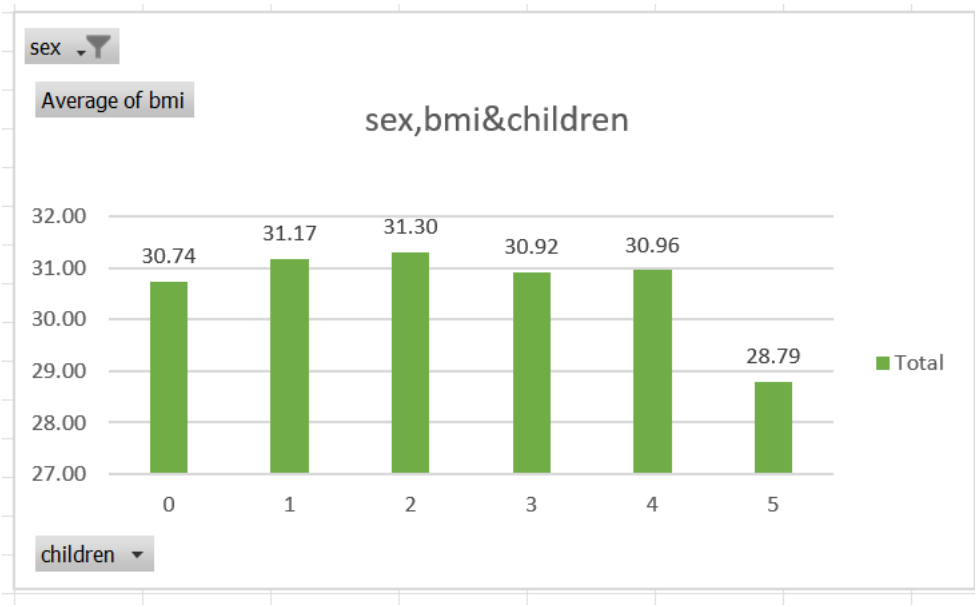


The above graph represents that south-east region charges more whereas, 3 children of a per customer charges more when compared to others. I used to insert slicer which we can represent graph in individual plots.

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



The above graph represents that BMI of 4 children of per female customers as high in graph represents that 31.94 when compared to others.



The above graph represents that BMI of 4 children of per male customers as high in graph represents that 30.96 when compared to others.

So, both graph represents that slightly differences male and female customers of BMI children but both has 4 children per customer BMI as slightly differs.

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

The histogram says that 18 ,22.5 has 222 range has more when compared to others.

The box plots represent that age and BMI are related to each other and plots is based on charges which moderate range of the charges occurred.

BMI and age are positively correlated.

Charges and BMI, age both are strongly correlated. therefore, there are no negative correlations.

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

male ratio as high ratio of 676 members is smokers.

age of 64 are more charged when compared to other age persons.

the BMI between 30.2 & 30.685, between 45.32 & 47.52 for these criteria of BMI charges more when compared to other BMI'S.

the more charges for smokers whereas they are into more smoking aspects they have certain impacts in health issues, so they have more charges for them.

The smoker has more charges as been charged more when compared to female.

2. Edit the data as following, to obtain dummy variables: (5 marks) 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.

I had changed the female and male with numbers replace 1 with male and female with 0.

I put a short cut key ctrl + h.

The find and replace dialog box appear.

In find option male in replace option as 1 so replaced male as 1 same for female also.

sex	sex	
female	0	
male	1	
male	1	
male	1	
male	1	
female	0	
female	0	
female	0	
male	1	
female	0	
male	1	
female	0	
male	1	
female	0	
male	1	
male	1	
female	0	
male	1	
male	1	
male	1	

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

I had used if condition statement for replace of 0 & 1 in smoker and non-smoker.

=IF(A2="yes",1,0)

smoker	smoker in numbers
yes	1
no	0
no	0
no	0
no	0
no	0
no	0
no	0
no	0
no	0
no	0
yes	1

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

in this also I had use if condition statements

=IF($A2=B$1,1,0)

Which is the table is below

region	northeast	northwest	southeast	southwest
southwest	0	0	0	1
southeast	0	0	1	0
southeast	0	0	1	0
northwest	0	1	0	0
northwest	0	1	0	0
southeast	0	0	1	0
southeast	0	0	1	0
northwest	0	1	0	0
northeast	1	0	0	0
northwest	0	1	0	0
northeast	1	0	0	0
southeast	0	0	1	0
southwest	0	0	0	1
southeast	0	0	1	0
southeast	0	0	1	0
southwest	0	0	0	1
northeast	1	0	0	0
northeast	1	0	0	0

3. a) Do a descriptive summary analysis for the edited data.

b) Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim.

c) Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

a) descriptive summary analysis for edited data

Summary output	*sex*	*smoker*	*northeast*	*northwest*	*southeast*	*southwest*
Mean	0.51	0.20	0.24	0.24	0.27	0.24
Standard Error	0.01	0.01	0.01	0.01	0.01	0.01
Median	1.00	0.00	0.00	0.00	0.00	0.00
Mode	1.00	0.00	0.00	0.00	0.00	0.00
Standard Deviation	0.50	0.40	0.43	0.43	0.45	0.43
Sample Variance	0.25	0.16	0.18	0.18	0.20	0.18
Kurtosis	-2.00	0.15	-0.55	-0.56	-0.95	-0.56
Skewness	-0.02	1.46	1.21	1.20	1.03	1.20
Range	1	1	1	1	1	1
Minimum	0	0	0	0	0	0
Maximum	1	1	1	1	1	1
Sum	676	274	324	325	364	325
Count	1338	1338	1338	1338	1338	1338

the set of the data includes sex smoke regions are in binary numbers. The mean of sex indicates that the slightly more than half 50% the mean of smokers as the 20% of the individual smokers. The mean value of 4 regions is relatively like all individuals but southeast as 27%

standard error as has every variable as similar numbers 0.01.

standard deviation as all data has been spread out through all the variables.

sample variance as square root of standard deviation

skewness as the sex as negatively skewed rather than others has positively skewed.

b) Perform a Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78760514
R Square	0.620321857
Adjusted R Square	0.618145888
Standard Error	7475.94221
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	1.21629E+11	20271520881	435.2469215	0
Residual	1332	74445096282	55889711.92		
Total	1338	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8246.438523	467.6641265	17.63325014	1.03981E-62	7329.000032	9163.877015	7329.000032	9163.877015
sex	-71.47576133	410.0070399	-0.174328132	0.861634083	-875.8056618	732.8541391	-875.8056618	732.8541391
smoker	23571.15645	509.2980382	46.28165569	1.4868E-279	22572.04277	24570.27012	22572.04277	24570.27012
northeast	321.6221259	587.0934677	0.547820992	0.583906609	-830.1064653	1473.350717	-830.1064653	1473.350717
northwest	0	0	65535	#NUM!	0	0	0	0
southeast	633.2962167	571.7351159	1.107674164	#NUM!	-488.3031806	1754.895614	-488.3031806	1754.895614
southwest	-70.19814584	586.4665846	-0.119696753	0.904741423	-1220.696951	1080.30066	-1220.696951	1080.30066

Multilinear regression analysis is performed to identify the which variables as significant impact on the insurance charges/billed insurance claimed the result suggest that sex and all regions are statistically insignificant($p>0.05$). but smoker has highly significant in the insurance charges($p<0.05$) this indicates that individual who smokes have higher insurance charges compared to those who do not smoke.

c) Give your interpretation for the above analysis, do another set of regression analysis by dropping insignificant variables, if needed.

The above table regression analysis is below.

The R squared indicates that 0.62 which is of 62%of variance in insurance charges.

The adjusted R square is 0.61 which is of 61%

The significant are only smokers.

The insignificant variables are sex and regions.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.78725143
R Square	0.619764815
Adjusted R Square	0.619480208
Standard Error	7470.216208
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.2152E+11	1.2152E+11	2177.614868	8.2714E-283
Residual	1336	74554317947	55804130.2		
Total	1337	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	8434.268298	229.0141716	36.8285868	0.00	7985.001758	8883.534838	7985.001758	8883.534838
smoker	23615.96353	506.0752904	46.66492117	0.00	22623.17478	24608.75229	22623.17478	24608.75229

The above table represents that dropping all the insignificant only significant variable is in regression analysis.

The results says that the smoker has high positive associate with insurance charges with co-efficient of 23615.96. this means the smokers have insurance charges that are $23615.96 higher than non-smokers.

The r squared value is 0.62 models are same, but the smoker are remains significant value 23615.96.

Thank you