INSURANCE CLAIM

Reetesh V
reeteshvenki@gmail.com

Situation : - An insurance company in US is reviewing its insurance claims and tyring to do an effect analysis for their future business decisions.

Task:- To perform effect analysis on the data collected by the company for future decisions.

Action:-

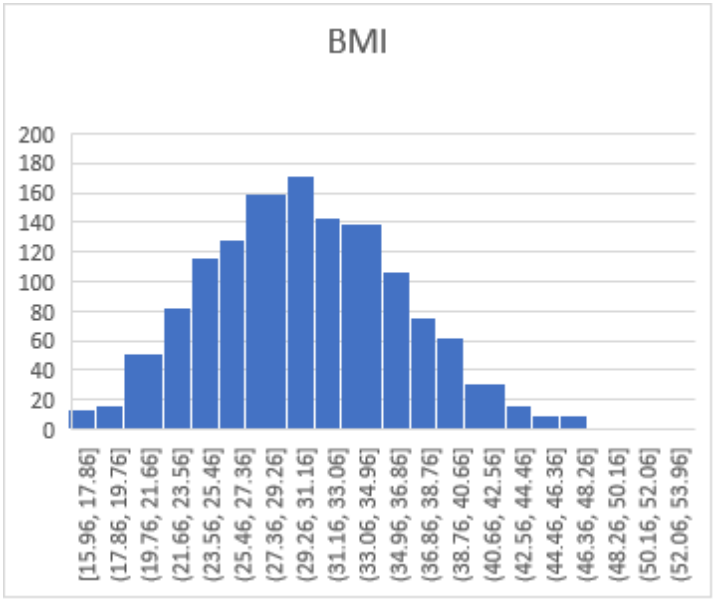
1. **a) Identify the categorical and continuous variables.**

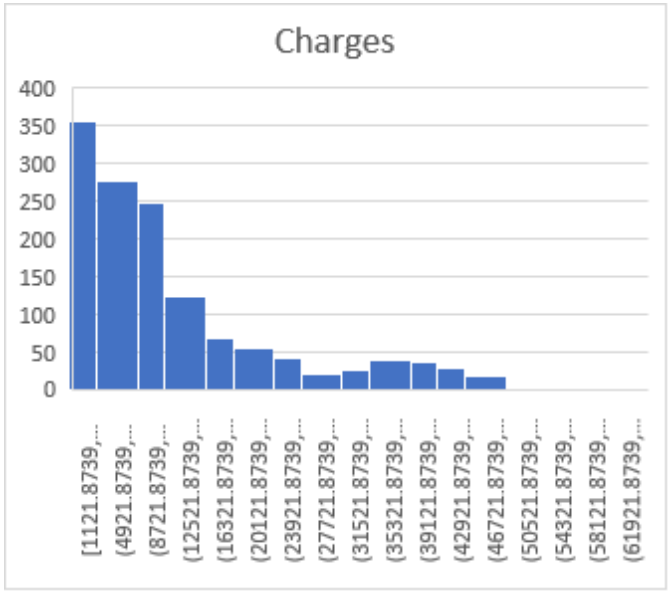
Categorical variable :- Sex, Smoker, Region, Children includes characters

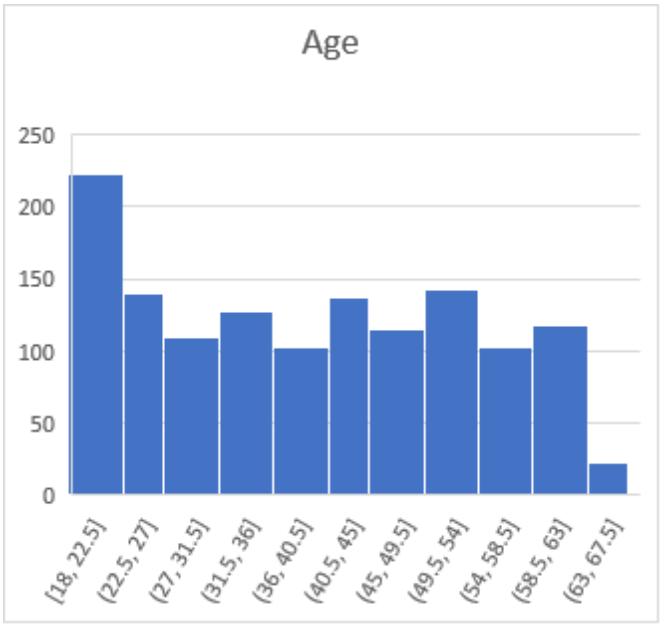
Continuous variable :- Age, BMI, and Charges includes numbers

b. Make Histograms and box plots for continuous variables, do a correlation analysis.

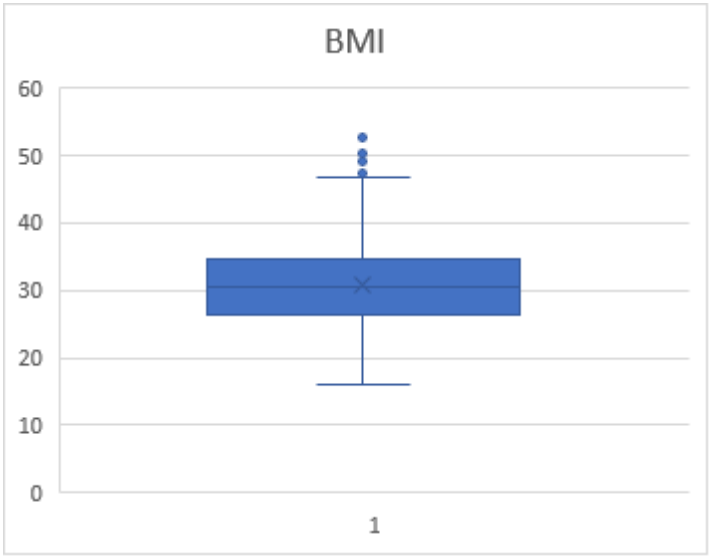
Histograms

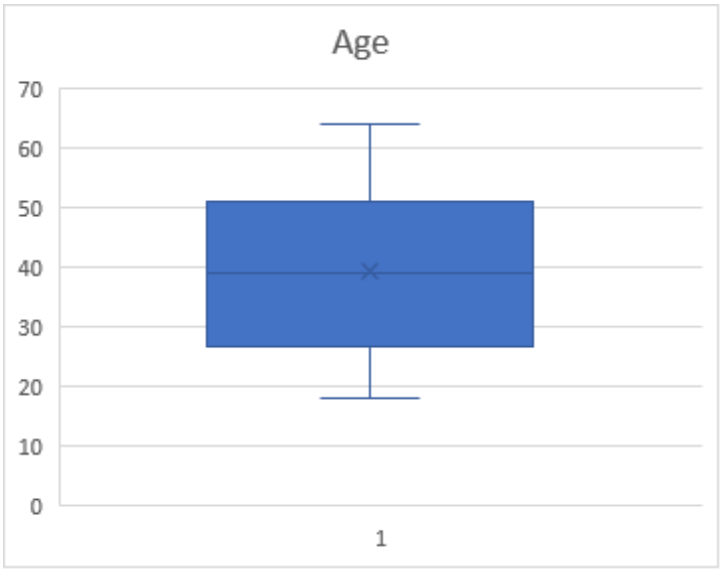


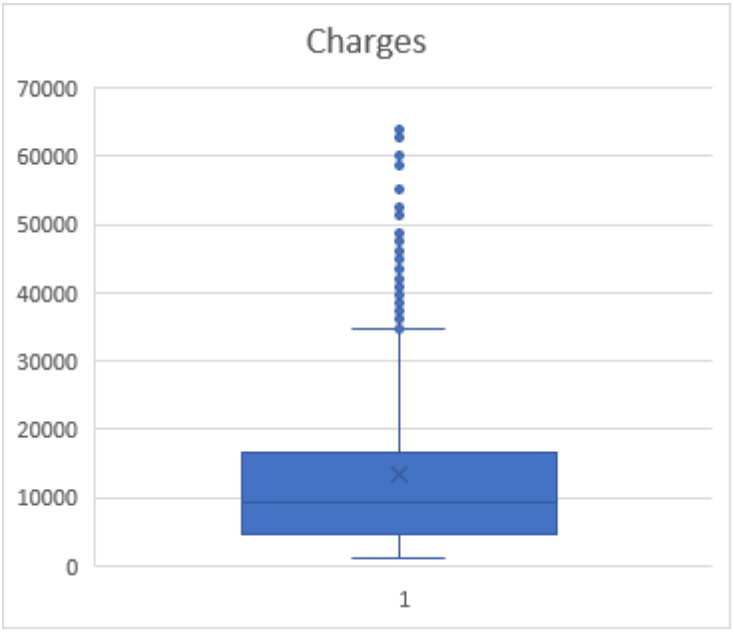



Age
250
200
150
100
50
0
[18, 22.5]
(22.5, 27]
(27, 31.5]
(31.5, 36]
(36, 40.5]
(40.5, 45]
(45, 49.5]
(49.5, 54]
(54, 58.5]
(58.5, 63]
(63, 67.5]

BOX PLOT


BMI
60
50
40
30
20
10
0
1


Age
70
60
50
40
30
20
10
0
1


Charges
70000
60000
50000
40000
30000
20000
10000
0
1

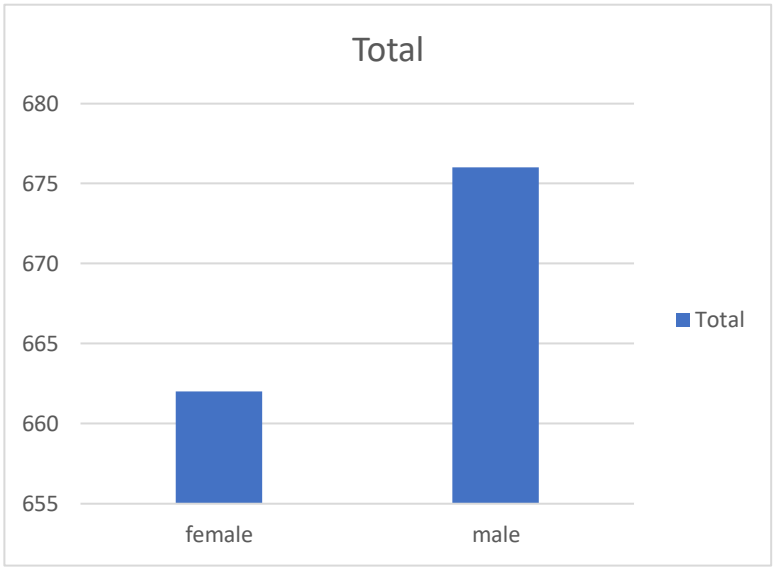
c. Make relevant Pivot tables and charts for :

1. Male/Female ratio and which gender has more smokers :

-

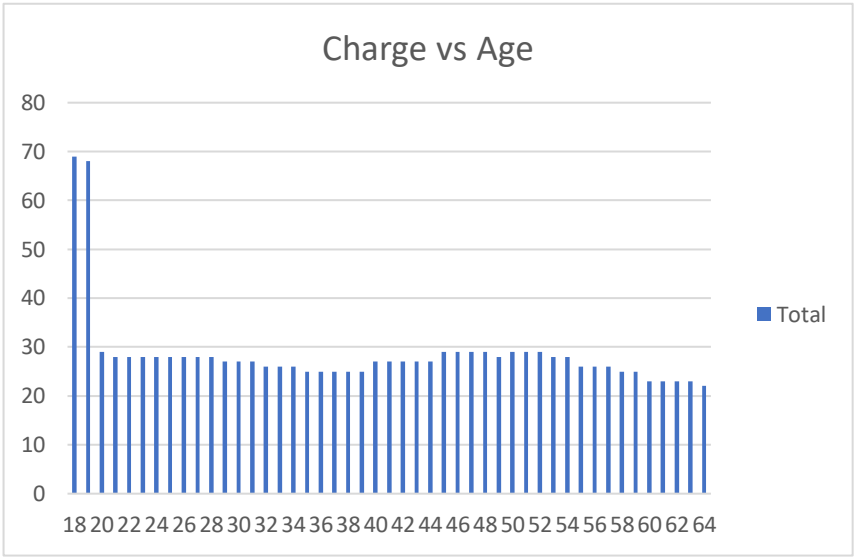
Row Labels	Count of smoker
female	662
male	676
Grand Total	**1338**

Here the Ratio of smokers is more for Male gender



2. Charge vs Age

Row Labels	Count of charges($)
18	69
19	68
20	29
21	28
22	28
23	28
24	28
25	28
26	28
27	28
28	28
29	27
30	27
31	27
32	26
33	26
34	26
35	25
36	25
37	25
38	25
39	25
40	27
41	27
42	27
43	27
44	27
45	29
46	29
47	29
48	29
49	28
50	29
51	29
52	29
53	28
54	28
55	26
56	26
57	26
58	25
59	25
60	23
61	23
62	23
63	23
64	22
Grand Total	**1338**

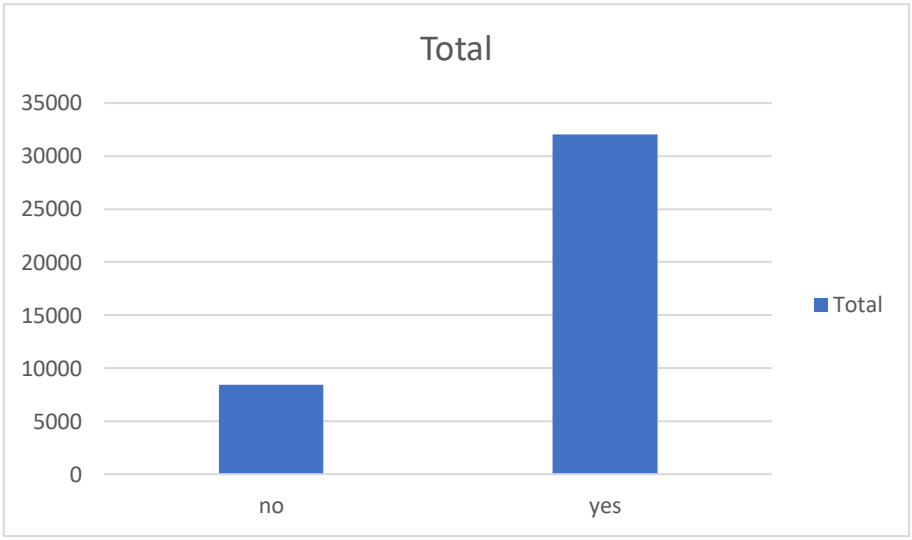


The charges is more for the people with 18 age category

3. Charges for Smokers vs Non Smokers

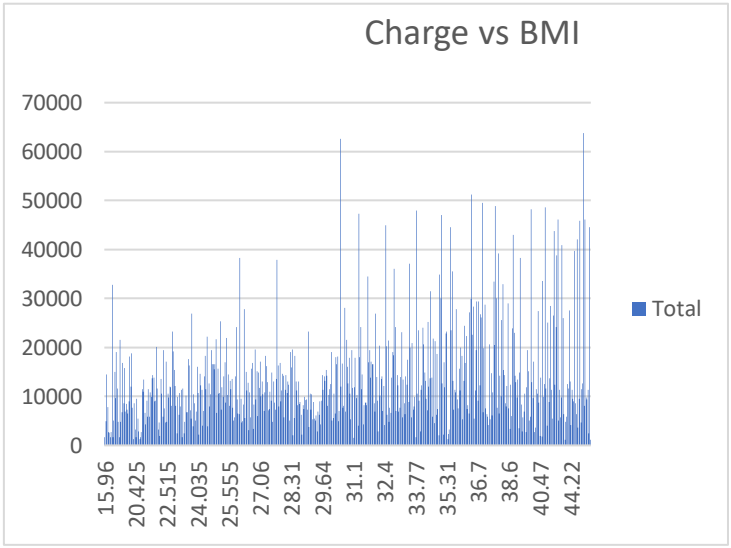
Row Labels	Average of charges($)
no	8434.268298
yes	32050.23183
Grand Total	**13270.42227**

The charges is more for the people who smoke



4. Charge vs BMI

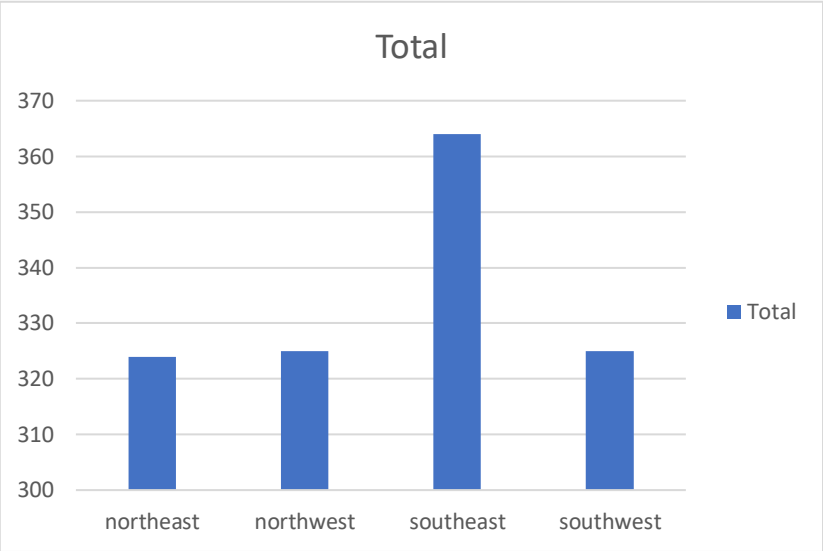
Row Labels	Average of charges($)
15.96	1694.7964
16.815	4904.00035
17.195	14455.64405
17.29	7813.353433
17.385	2775.19215
17.4	2585.269
17.48	1621.3402
17.67	2680.9493
17.765	32734.1863
17.8	1727.785
17.86	5116.5004
17.955	15006.57945
18.05	9644.2525
18.3	19023.26
18.335	11576.73198
18.5	4766.022
18.6	1728.897
18.715	21595.38229
18.905	4827.90495
19	6753.038
19.095	16776.30405
19.19	8627.5411
19.3	15820.699
19.475	6933.24225
19.57	8428.0693
19.8	7266.665667
19.855	6492.37645
19.95	9049.190833
20.045	18109.27455
20.1	12032.326



d. Region-wise Smokers vs non-smokers analysis with one or more pivot table and charts.

smoker	(All)

Row Labels	Count of smokers
northeast	324
northwest	325
southeast	364
southwest	325
Grand Total	**1338**



The people from Southeast region has more number of smokers compared to the people from other region.

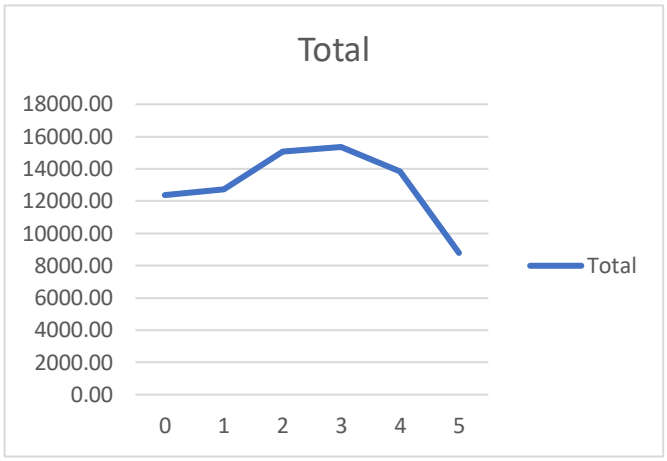
e.) Region-wise charges for smokers vs non-smokers ?

Average of charges($)	Column Labels		
Row Labels	no	yes	Grand Total
northeast	9165.531672	29673.53647	13406.38452
northwest	8556.463715	30192.00318	12417.57537
southeast	8032.216309	34844.99682	14735.41144
southwest	8019.284513	32269.06349	12346.93738
Grand Total	**8434.268298**	**32050.23183**	**13270.42227**

- The people from Southeast region has more number of smokers and are maximum charged.
- The people from Northeast region has smokers and are charged less.
- The people from northeast region doesn't smoke and are charged more.
- The people from southwest region doesn't smoke and are charged less.

f.) Has charges got something to do with no. of dependents ?

Row Labels	Average of charges($)
0	12365.98
1	12731.17
2	15073.56
3	15355.32
4	13850.66
5	8786.04
Grand Total	**13270.42**



- From the above graph we can infer that the charges increased from 0 children to 3 children.
- And the charges was decreased again for 4 and 5 th child.
- The Charges for high for the people of southeast region having 3 children
- The Charges are low for the people of northeast region having 5 children

g.) Do a similar dependents-charges analysis, Region-wise ?

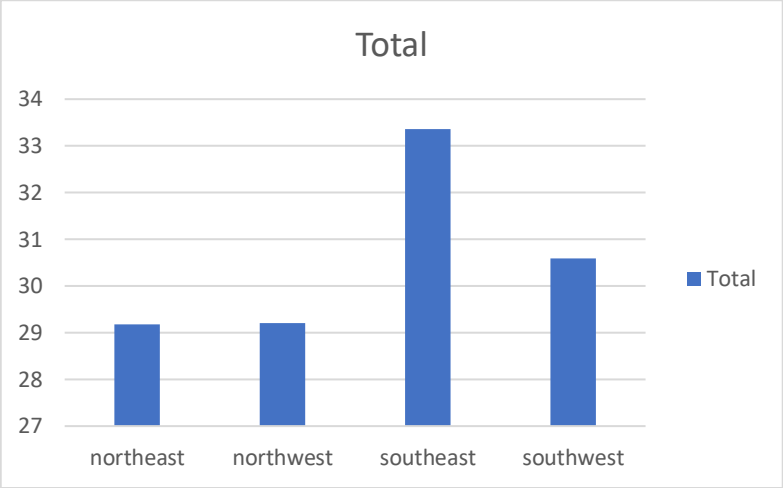
Average of charges($)	Column Labels				
Row Labels	northeast	northwest	southeast	southwest	Grand Total
0	11626.46266	11324.37092	14309.86838	11938.50499	12365.9756
1	16310.2064	10230.25631	13687.04197	10406.48495	12731.17183
2	13615.15272	13464.31469	15728.47062	17483.48556	15073.56373
3	14409.9133	17786.16067	18449.84602	10402.44226	15355.31837
4	14485.19312	11347.01873	14451.02397	14933.26053	13850.65631
5	6978.973483	8965.79575	10115.44154	8444.158625	8786.035247
Grand Total	**13406.38452**	**12417.57537**	**14735.41144**	**12346.93738**	**13270.42227**

In Northeast region people having 1 children has more charge and people having 5 children has less charges, In Northwest region people having 3 children has more charges and people having 5 children has less charges, in Southeast region

people having 2 children has more charges and people having 5 children has less charges.

h.) Region vs BMI

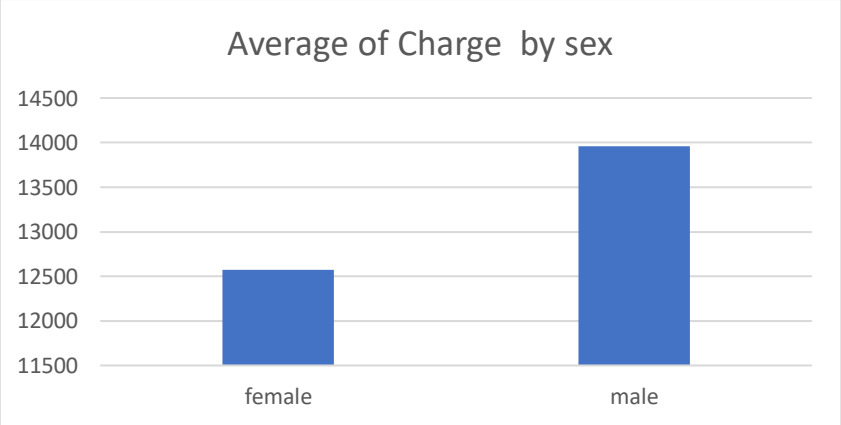
Row Labels	Average of bmi
northeast	29.17350309
northwest	29.19978462
southeast	33.35598901
southwest	30.59661538
Grand Total	**30.66339686**



The people from southeast region has more BMI and the people from northeast region has less BMI.

Sex vs Charge

sex	Average of charges($)
female	12569.57884
male	13956.75118



Male are charged more compared to Female.

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

Histogram of Age

- Maximum people are from the age category of 18 - 22.5 age.
- There are less people from the age category of 63 – 67.5 age.

Box Plot of Age

- There are no outliers in age box plot

Histogram of BMI

- The Maximum BMI lies between 29.26 – 31.16
- The Minimum BMI lies between 50.16 – 52.06

Box Plot of BMI

- The Minimum BMI is 15 and Maximum BMI is 45 and the Median is 30
- There are more number of outliers above 45.

Histogram of Charges

- The Maximum charge ranges from 18 to 22.5, minimum charge ranges from 63 to 67.5

Box Plot of Charges

- The Minimum charge is 18 and the Maximum charge is 65.
- There are no outliers.

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

1. The Male has more number of smokers that is 676 smokers and the has female up to 662 smokers which is less than the male ratio, the Male has more number of smokers.
2. From the chart of Charge vs Age, The people of age 64 are charged more and the people with age 26 are charged less

3. From the chart of Charge vs BMI, The people having maximum BMI that is 47.52 are Charged more and the people with BMI 29.48 are charged less.
4. From the chart of smokers and non smokers vs charges, The people who smoke are charged maximum compared to the people who doesn't smoke.

2. Do a descriptive summary analysis for the edited data.

age		*bmi*		*children*		*charges($)*		*southwest*	
Mean	39.20703	Mean	30.6634	Mean	1.094918	Mean	13270.422	Mean	0.2429
Standard Error	0.384102	Standard Error	0.166714	Standard Error	0.032956	Standard Error	331.0675	Standard Error	0.011728
Median	39	Median	30.4	Median	1	Median	9382.033	Median	0
Mode	18	Mode	32.3	Mode	0	Mode	1639.563	Mode	0
Standard Deviation	14.04996	Standard Deviation	6.098187	Standard Deviation	1.205493	Standard Deviation	12110.01	Standard Deviation	0.428995
Sample Variance	197.4014	Sample Variance	37.18788	Sample Variance	1.453213	Sample Variance	1.47E+08	Sample Variance	0.184037
Kurtosis	-1.24509	Kurtosis	-0.05073	Kurtosis	-0.20245	Kurtosis	1.606299	Kurtosis	-0.55986
Skewness	0.055673	Skewness	0.284047	Skewness	0.938380	Skewness	1.515885	Skewness	1.200409
Range	46	Range	37.17	Range	5	Range	62648.55	Range	1
Minimum	18	Minimum	15.96	Minimum	0	Minimum	1121.874	Minimum	0
Maximum	64	Maximum	53.13	Maximum	5	Maximum	63770.43	Maximum	1
Sum	52459	Sum	41027.63	Sum	1465	Sum	17755825	Sum	325
Count	1338	Count	1338	Count	1338	Count	1338	Count	1338

southeast		*northwest*		*northeast*		*smokers*		*gender*
Mean	0.272047833	Mean	0.242899851	Mean	0.242152466	Mean	0.204783259	Mean
Standard Error	0.012170498	Standard Error	0.011728017	Standard Error	0.011715738	Standard Error	0.011036321	Standard Error

Median	0	Median	0	Median	0	Median	0	Median	
Mode	0	Mode	0	Mode	0	Mode	0	Mode	
Standard Deviation	0.445180784	Standard Deviation	0.428995407	Standard Deviation	0.428546273	Standard Deviation	0.403694038	Standard Deviation	
Sample Variance	0.198185938	Sample Variance	0.184037068	Sample Variance	0.183651908	Sample Variance	0.162968876	Sample Variance	
	-0.949522817		-0.559856699		-0.548410009		-0.145755539		
Kurtosis	7	Kurtosis	9	Kurtosis	9	Kurtosis	9	Kurtosis	
Skewness	1.025621147	Skewness	1.200409261	Skewness	1.205160559	Skewness	1.464766161	Skewness	
Range	1	Range	1	Range	1	Range	1	Range	
Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	
Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	
Sum	364	Sum	325	Sum	324	Sum	274	Sum	
Count	1338	Count	1338	Count	1338	Count	1338	Count	

From the above Analysis we can infer that

- The kurtosis is at peak for Charges that is 1.6
- The kurtosis is low for Gender that is -2.0
- The skewness is high for Charges that is 1.51 and is right skewed.
- The skewness is low for Gender that is -0.02 and is left skewed.

2) 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.

SUMMARY OUTPUT - 4

Regression Statistics	
Multiple R	0.865849
R Square	0.749695
Adjusted R Square	0.748943
Standard Error	6067.787
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1.47E+11	3.67E+10	998.1232	0
Residual	1333	4.91E+10	36818042		
Total	1337	1.96E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12102.8	941.9839	-12.8482	1.05E-35	-13950.7	-10254.8	-13950.7	-10254.8
age	257.8495	11.89639	21.67461	1.75E-89	234.5118	281.1872	234.5118	281.1872
BMI	321.8514	27.37763	11.756	1.97E-30	268.1435	375.5593	268.1435	375.5593
children	473.5023	137.7917	3.436364	0.000608	203.1902	743.8145	203.1902	743.8145
smokers	23811.4	411.2197	57.90432	0	23004.69	24618.11	23004.69	24618.11

Result :-

From the above model we can infer that, BMI, children, Smokers are significant variables.

Charges is depending on age, BMI, children, smokers.

Thank you