



INSURANCE CLAIM PROJECT

BY: S VIGNESH

1. Perform the basic Exploratory Data Analysis on the sample data.

AGE

- ▶ The Average age is around 39.20
- ▶ Negative Kurtosis gives us a flatter distribution for this variable.
- ▶ Skewness is nearly 0, which says curve follows normal distribution.

AGE	
Mean	39.20702541
Standard Error	0.384102419
Median	39
Mode	18
Standard Deviation	14.04996038
Sample Variance	197.4013867
Kurtosis	-1.245087653
Skewness	0.055672516
Range	46
Minimum	18
Maximum	64
Sum	52459
Count	1338

1. Perform the basic Exploratory Data Analysis on the sample data.

BMI

- ▶ The Average of BMI is around 30.6
- ▶ The data gives us slightly negative kurtosis but since the value is small so we can say it is a Normal distribution.
- ▶ Skewness of about 0.28 gives a right tailed distribution.

BMI	
Mean	30.66339686
Standard Error	0.166714232
Median	30.4
Mode	32.3
Standard Deviation	6.098186912
Sample Variance	37.18788361
Kurtosis	0.050731531
Skewness	0.284047111
Range	37.17
Minimum	15.96
Maximum	53.13
Sum	41027.625
Count	1338

1. Perform the basic Exploratory Data Analysis on the sample data

CHILDREN

- ▶ The average no. of children is around 39.
- ▶ The data gives 0.20 kurtosis but since the value is small so we can say follows Normal distribution.
- ▶ Skewness of about 0.93 gives a right tailed distribution.

CHILDREN	
Mean	1.094918
Standard Error	0.032956
Median	1
Mode	0
Standard Deviation	1.205493
Sample Variance	1.453213
Kurtosis	0.202454
Skewness	0.93838
Range	5
Minimum	0
Maximum	5
Sum	1465
Count	1338

1. Perform the basic Exploratory Data Analysis on the sample data

CHARGES

- ▶ The average charges is around 13270.
- ▶ Positive Kurtosis gives us a shape curve than normal curve
 - saying more values are concentrated near to median.
- ▶ Skewness of about 1.51 gives a right tailed distribution.

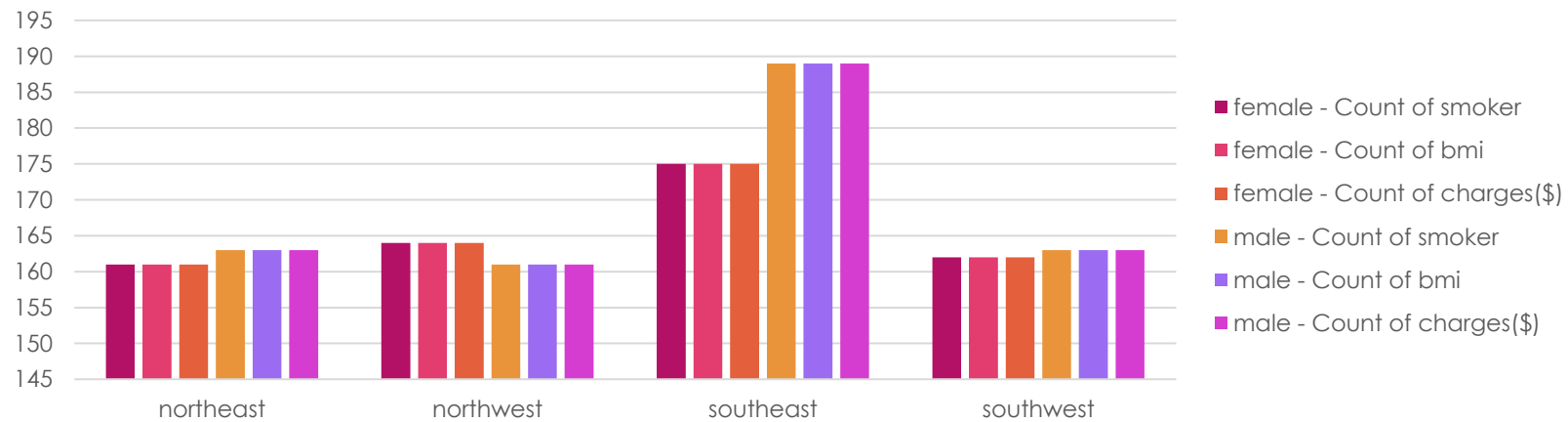
CHARGES(\$)	
Mean	13270.42227
Standard Error	331.0674543
Median	9382.033
Mode	1639.5631
Standard Deviation	12110.01124
Sample Variance	146652372.2
Kurtosis	1.606298653
Skewness	1.515879658
Range	62648.55411
Minimum	1121.8739
Maximum	63770.42801
Sum	17755824.99
Count	1338

2. Identify the categorical and continuous variables.

Continuous Variables	Categorical Variables
Age	Sex
BMI	Smoker
Charges	Region
	Children

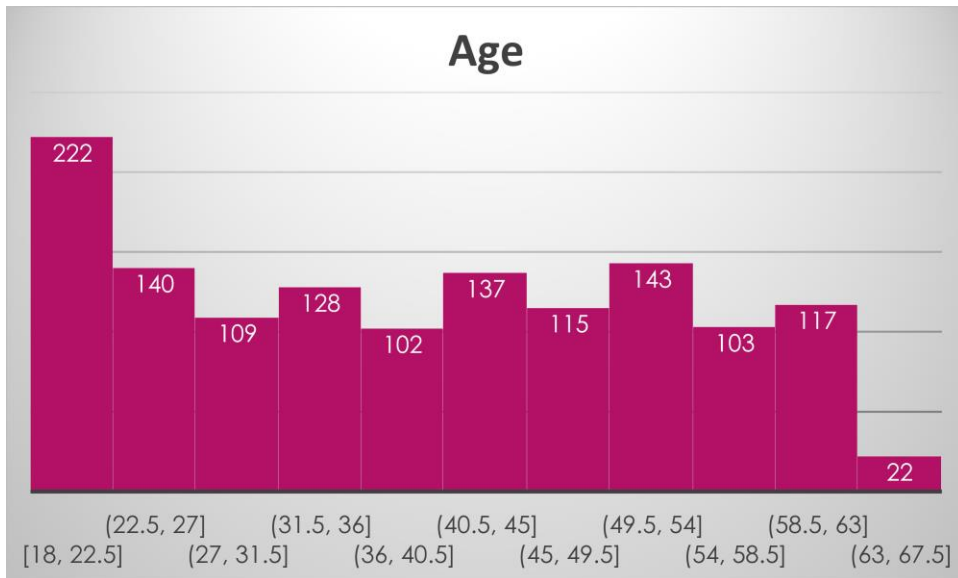


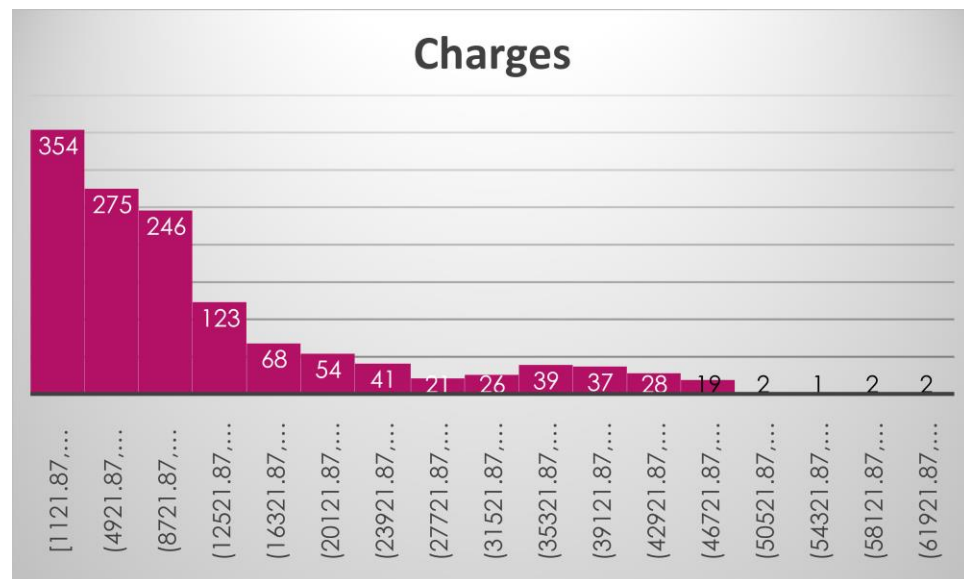
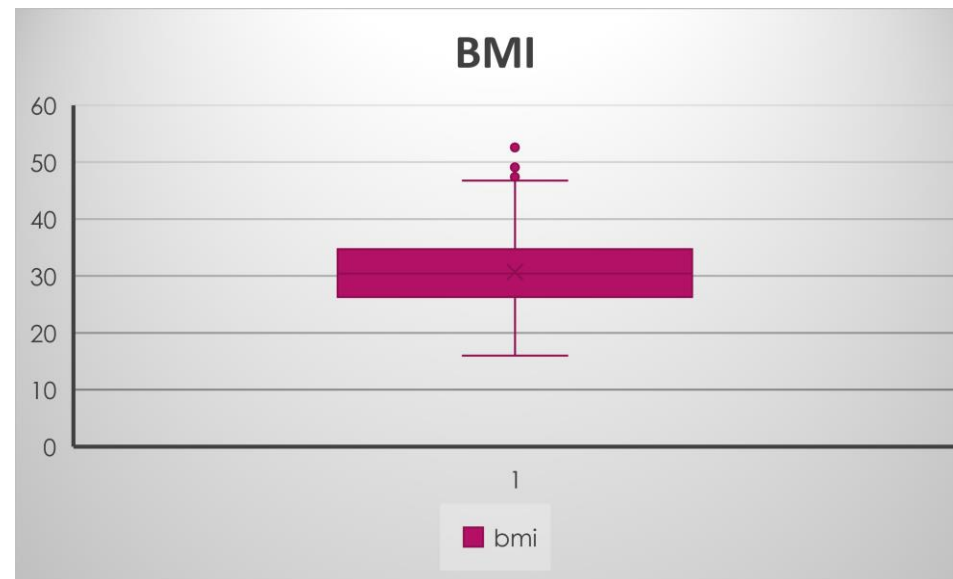
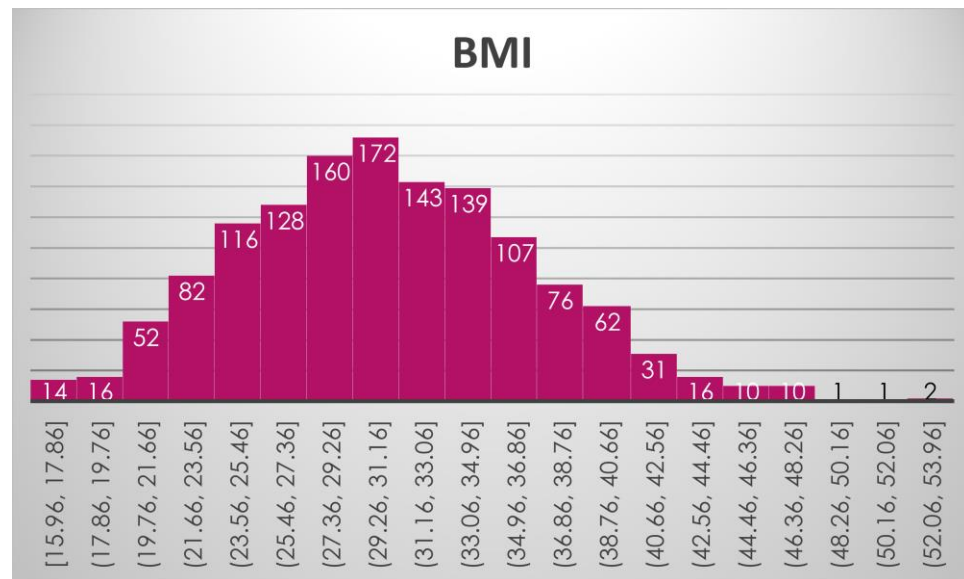
Row Labels	female			male			Total Count of smoker	Total Count of bmi	Total Count of charges(\$)
	Count of smoker	Count of bmi	Count of charges(\$)	Count of smoker	Count of bmi	Count of charges(\$)			
	Column Labels								
northeast	161	161	161	163	163	163	324	324	324
northwest	164	164	164	161	161	161	325	325	325
southeast	175	175	175	189	189	189	364	364	364
southwest	162	162	162	163	163	163	325	325	325
Grand Total	662	662	662	676	676	676	1338	1338	1338



3. Make Histograms and box plots for Continuous Variables, do a Correlation analysis

AGE





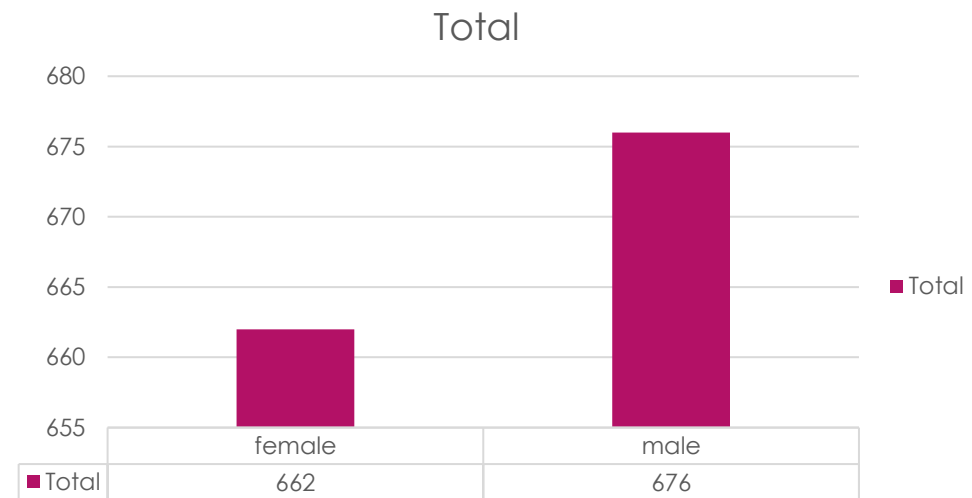
CORRELATION ANALYSIS

	BMI	CHILDREN	CHARGES(\$)	AGE
BMI	1			
CHILDREN	0.012759	1		
CHARGES(\$)	0.198341	0.067998	1	
AGE	0.109272	0.042469	0.299008	1

4. Make relevant Pivot tables and charts for:

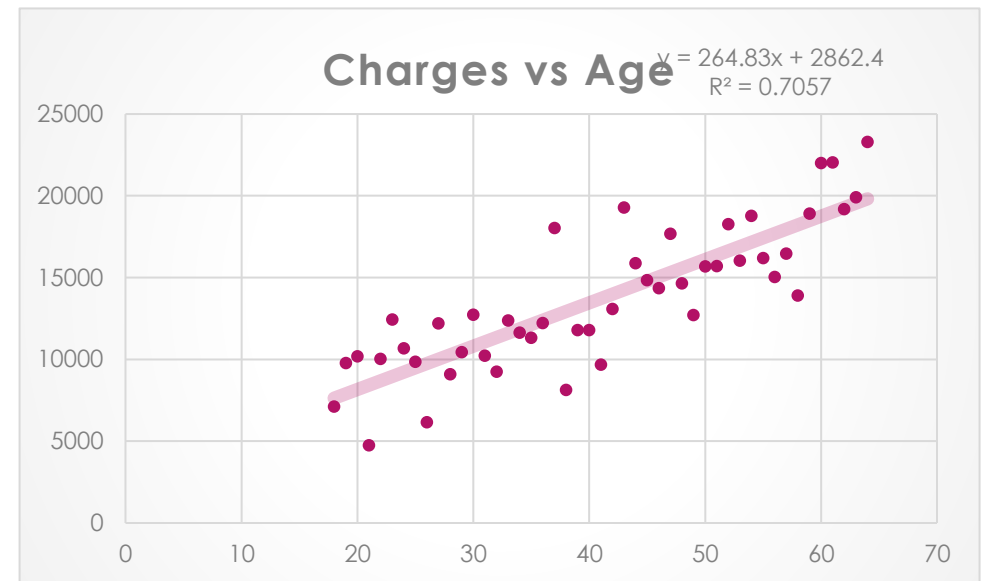
- Male/Female ratio & which gender has more smokers

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



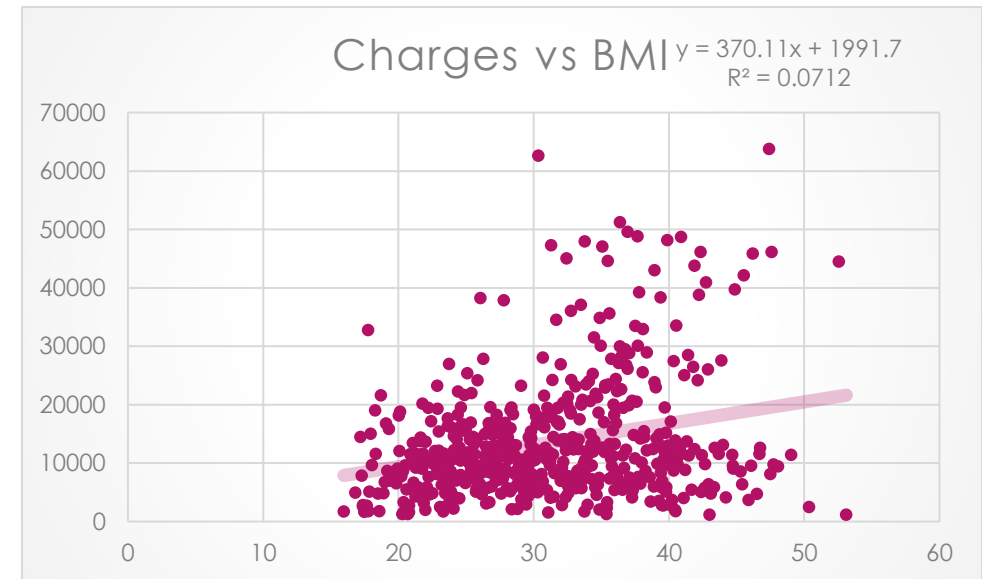
➤ CHARGES VS AGES

Row Labels	Average of charges(\$)
64	23275.53084
61	22024.45761
60	21979.41851
63	19884.99846
43	19267.27865
62	19163.85657
59	18895.86953
54	18758.54648
52	18256.26972
37	18019.91188
47	17653.99959
57	16447.18525
55	16164.54549
53	16020.93076
44	15859.39659



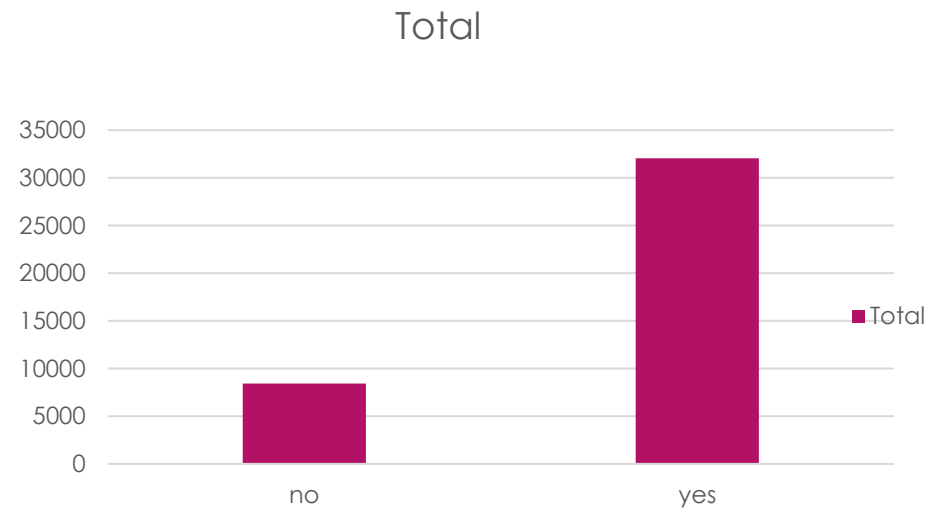
➤ CHARGES VS BMI

Row Labels	Average of charges(\$)
47.41	63770.42801
30.36	62592.87309
36.4	51194.55914
36.96	49577.6624
37.7	48824.45
40.92	48673.5588
39.9	48173.361
33.8	47928.03
31.3	47291.055
35.09	47055.5321
42.35	46151.1245
47.6	46113.511
46.2	45863.205
32.45	45008.9555
35.5	44585.45587
52.58	44501.3982
41.895	43753.33705
38.95	42983.4585
45.54	42112.2356



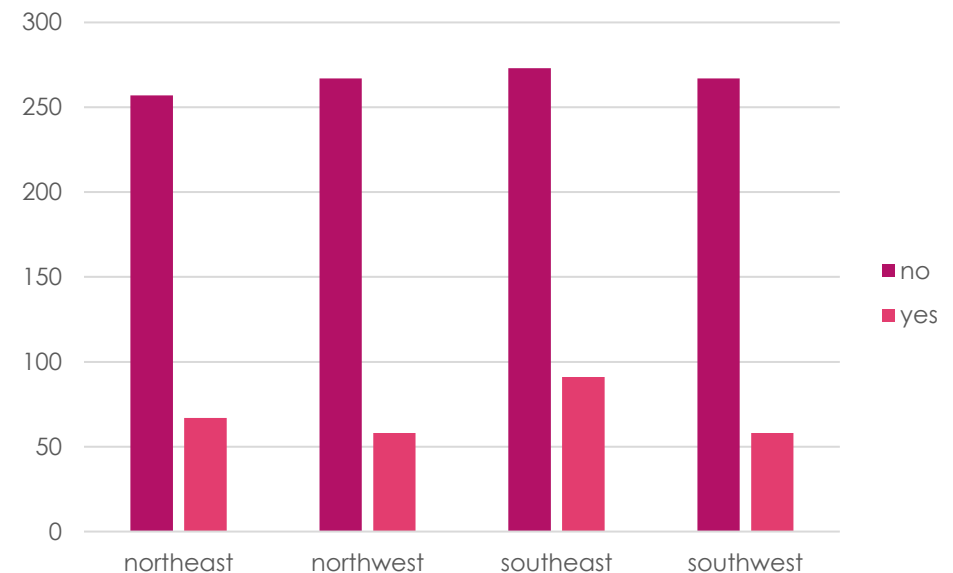
➤ CHARGES FOR SMOKERS VS NON-SMOKERS

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



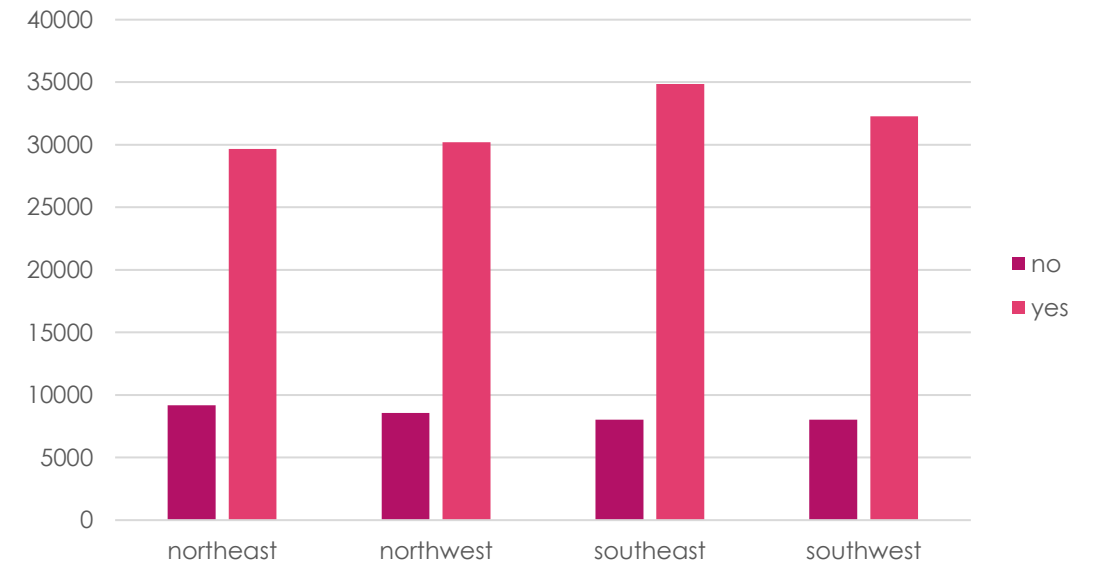
5.Region-Wise Smokers vs Non-Smokers analysis with one or more pivot table & charts

Count of smoker	Column Labels		
Row Labels	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



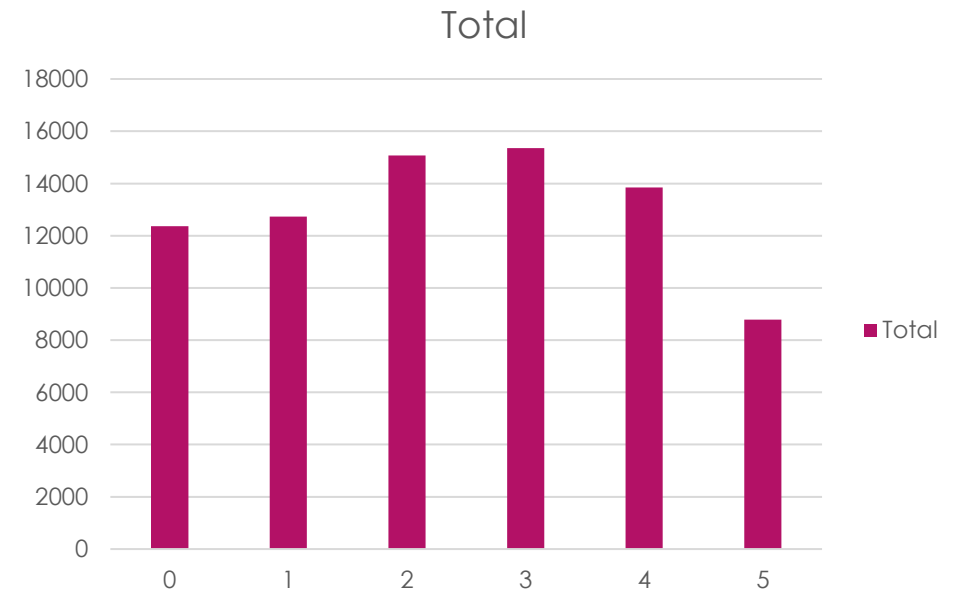
6. Region-Wise Charges for smokers vs non-smokers

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



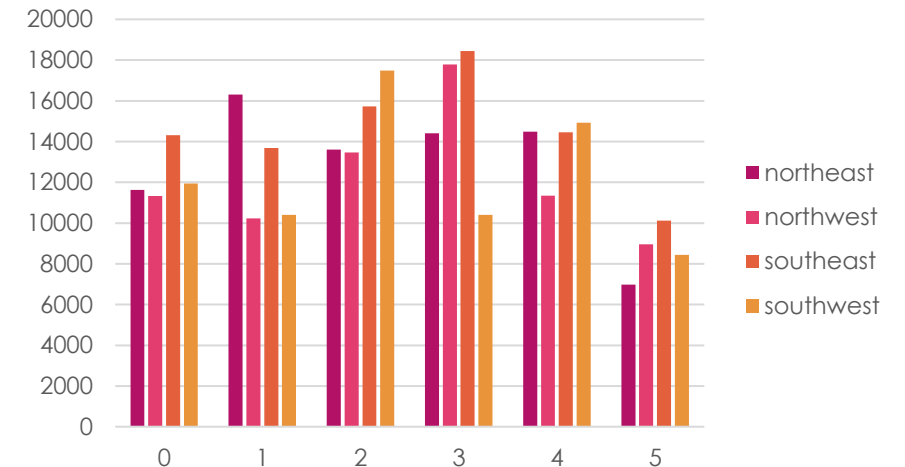
7.Has Charges got something to do with no. of dependants?

Row Labels	Average of charges(\$)
0	12365.9756
1	12731.17183
2	15073.56373
3	15355.31837
4	13850.65631
5	8786.035247
Grand Total	13270.42227



8. Do a similar dependants-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



9) Do at least one more pivot table and chart of your own choice, if needed

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

2. Give your interpretation for observations made in point (c)

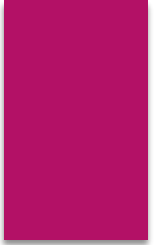
3. Edit the data as following, to obtain dummy variables:

4. 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 Replace with “Match entire cell content” option. Do a replace all to save time.

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

6. 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 “Southeast” column, whichever row had “southeast” as an entry will take “1” as the new entry and “0” for the rest columns. Do a similar operation on “Southwest” column.

	A	B	C	D	E	F	G	H	I
1	age	sex	bmi	children	smoker	northwest	southeast	southwest	charges(\$)
2	19	0	27.9	0	1	0	0	1	16884.924
3	18	1	33.77	1	0	0	1	0	1725.5523
4	28	1	33	3	0	0	1	0	4449.462
5	33	1	22.705	0	0	1	0	0	21984.47061
6	32	1	28.88	0	0	1	0	0	3866.8552
7	31	0	25.74	0	0	0	1	0	3756.6216
8	46	0	33.44	1	0	0	1	0	8240.5896
9	37	0	27.74	3	0	1	0	0	7281.5056
10	37	1	29.83	2	0	0	0	0	6406.4107
11	60	0	25.84	0	0	1	0	0	28923.13692
12	25	1	26.22	0	0	0	0	0	2721.3208
13	62	0	26.29	0	1	0	1	0	27808.7251
14	23	1	34.4	0	0	0	0	1	1826.843
15	56	0	39.82	0	0	0	1	0	11090.7178
16	27	1	42.13	0	1	0	1	0	39611.7577
17	19	1	24.6	1	0	0	0	1	1837.237
18	52	0	30.78	1	0	0	0	0	10797.3362
19	23	1	23.845	0	0	0	0	0	2395.17155
20	56	1	40.3	0	0	0	0	1	10602.385
21	30	1	35.3	0	1	0	0	1	36837.467
22	60	0	36.005	0	0	0	0	0	13228.84695
23	30	0	32.4	1	0	0	0	1	4149.736
24	18	1	34.1	0	0	0	1	0	1137.011
25	34	0	31.92	1	1	0	0	0	37701.8768
26	37	1	28.025	2	0	1	0	0	6203.90175
27	59	0	27.72	3	0	0	1	0	14001.1338
28	63	0	23.085	0	0	0	0	0	14451.83515



10) 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.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.866552
R Square	0.750913
Adjusted R Square	0.749414
Standard Error	6062.102
Observations	1338

ANOVA

	df	SS	MS	F	Significance F
Regression	8	1.47E+11	1.84E+10	500.8107	0
Residual	1329	4.88E+10	36749084		
Total	1337	1.96E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-11938.5	987.8192	-12.0858	5.58E-32	-13876.4	-10000.7	-13876.4	-10000.7
age	256.8564	11.89885	21.58666	7.78E-89	233.5138	280.1989	233.5138	280.1989
sex	-131.314	332.9454	-0.3944	0.693348	-784.47	521.8416	-784.47	521.8416
bmi	339.1935	28.59947	11.86013	6.5E-31	283.0884	395.2985	283.0884	395.2985
children	475.5005	137.8041	3.450555	0.000577	205.1633	745.8378	205.1633	745.8378
smoker	23848.53	413.1534	57.7232	0	23038.03	24659.04	23038.03	24659.04
northwest	-352.964	476.2758	-0.74109	0.458769	-1287.3	581.3704	-1287.3	581.3704
southeast	-1035.02	478.6922	-2.16219	0.030782	-1974.1	-95.9473	-1974.1	-95.9473
southwest	-960.051	477.933	-2.00876	0.044765	-1897.64	-22.4656	-1897.64	-22.4656