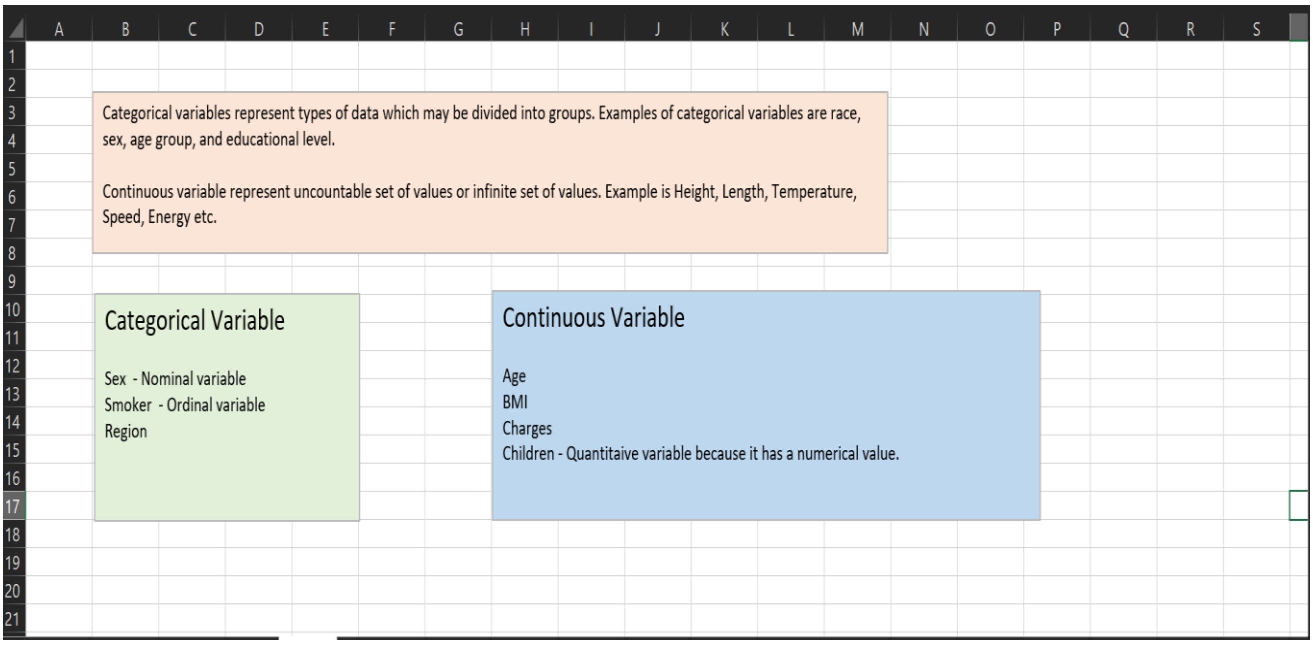
Business Report – Insurance Claim (Amit Varma)

Question 1(a): Categorical and Continuous Variable



Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level.

Continuous variable represent uncountable set of values or infinite set of values. Example is Height, Length, Temperature, Speed, Energy etc.

Categorical Variable

Sex - Nominal variable
Smoker - Ordinal variable
Region

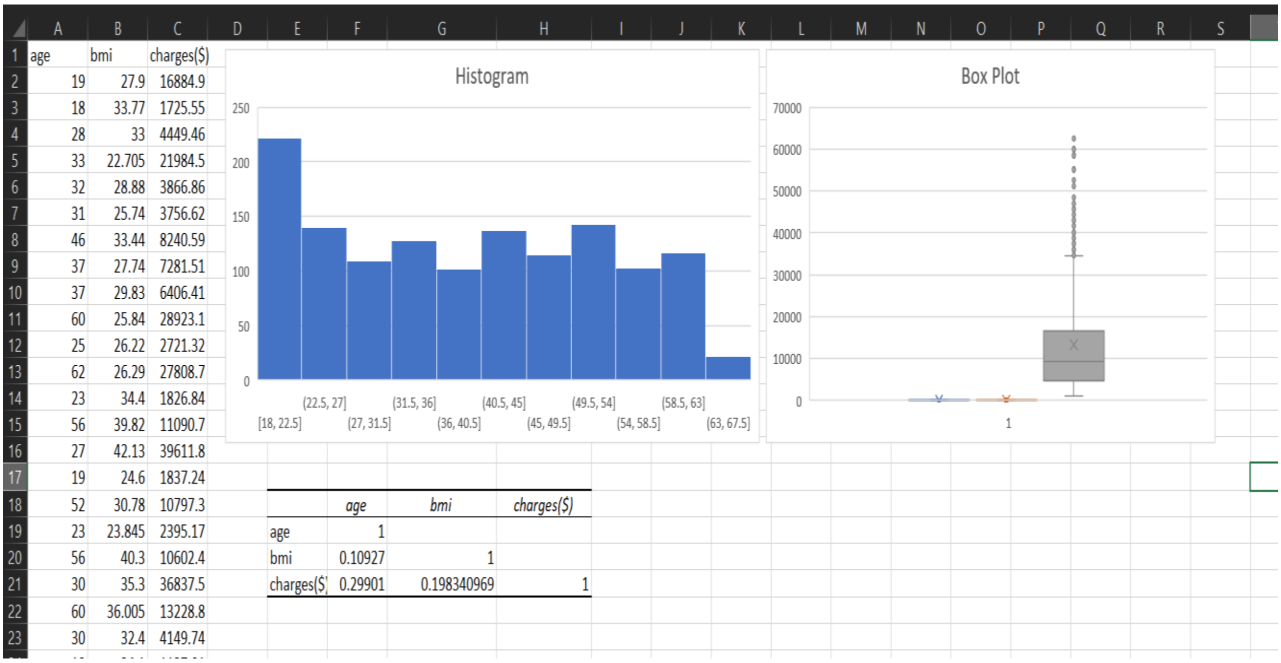
Continuous Variable

Age
BMI
Charges
Children - Quantitaive variable because it has a numerical value.

Here Sex, Smoker, Region is categorical variable because sex and region is nominal and ordinal variable this type of variable divided in to group.

Age, BMI, Charges and Children are continuous variable because he has numerical value.

Question 1(b): Histogram, Box plot and Correlation.



age	bmi	charges($)
19	27.9	16884.9
18	33.77	1725.55
28	33	4449.46
33	22.705	21984.5
32	28.88	3866.86
31	25.74	3756.62
46	33.44	8240.59
37	27.74	7281.51
37	29.83	6406.41
60	25.84	28923.1
25	26.22	2721.32
62	26.29	27808.7
23	34.4	1826.84
56	39.82	11090.7
27	42.13	39611.8
19	24.6	1837.24
52	30.78	10797.3
23	23.845	2395.17
56	40.3	10602.4
30	35.3	36837.5
60	36.005	13228.8
30	32.4	4149.74

	age	*bmi*	*charges($)*
age	1		
bmi	0.10927	1	
charges($	0.29901	0.198340969	1

Univariate analysis of continuous variable Age, BMI, Charges and multivariate analysis of correlation.

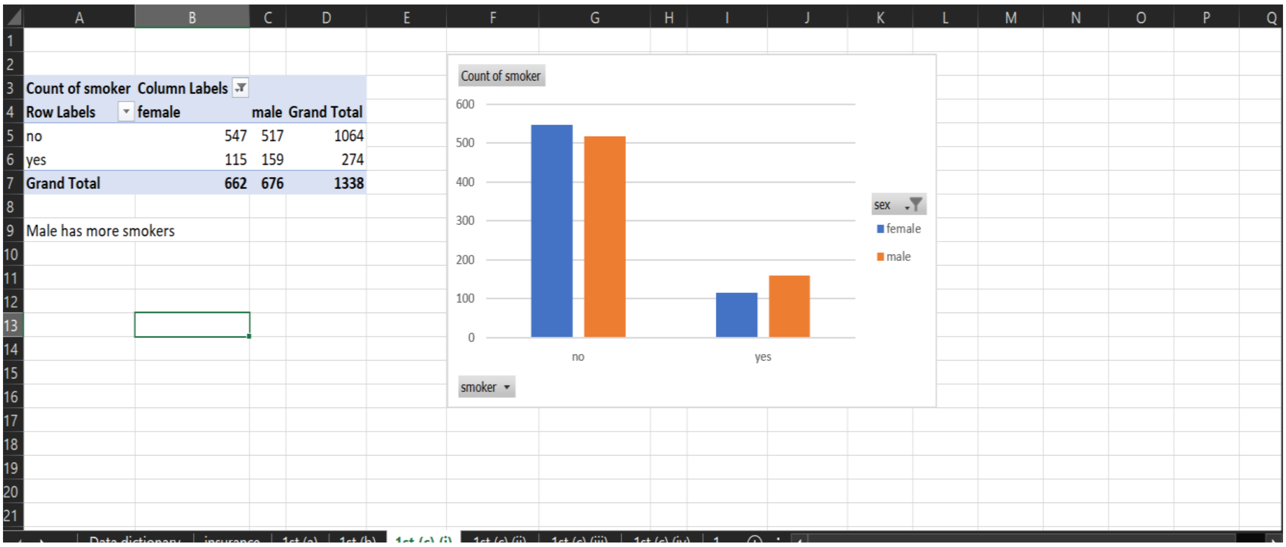
My understanding on point b is Age and Body mass index (18, 22.5)

charges are high and correlation between Age, BMI, Charges.

BMI has (0.10927) and charges (0.19834).

Question 1(c): Pivot table and Pivot chart

Male/Female ratio of smoker



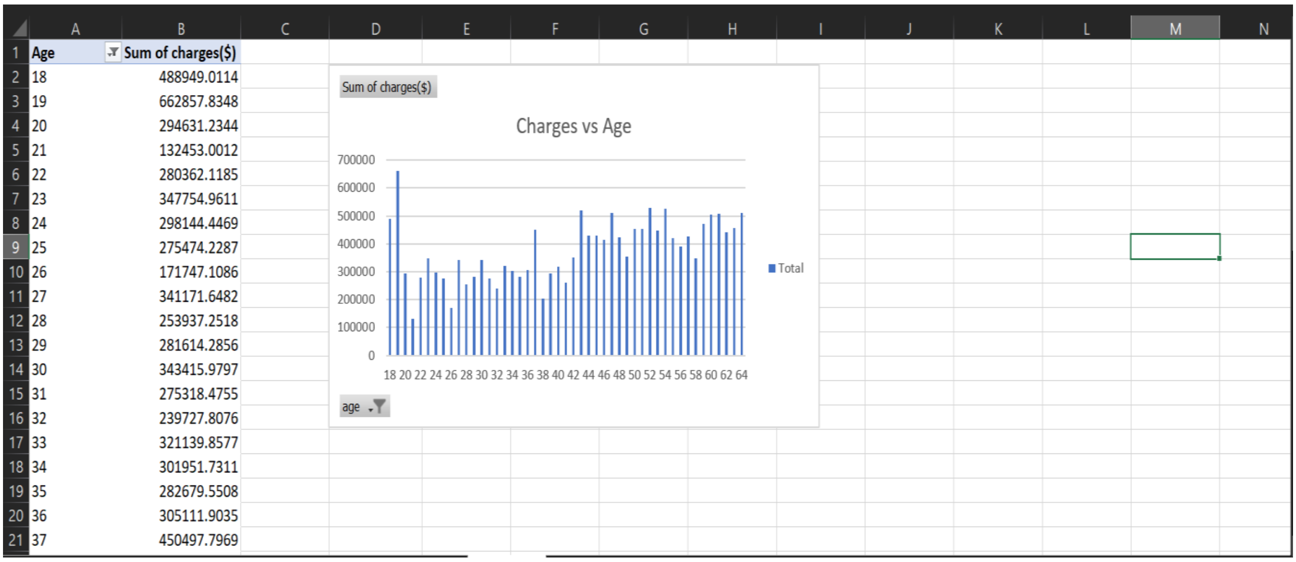
Count of smoker	Column Labels		
Row Labels	female	male	Grand Total
no	547	517	1064
yes	115	159	274
Grand Total	662	676	1338

Male has more smokers

Male/Female ratio of smoker – here ratio of male smoker has more compare to female smoker.

But the good thing is that the number of non-smokers male is also high.

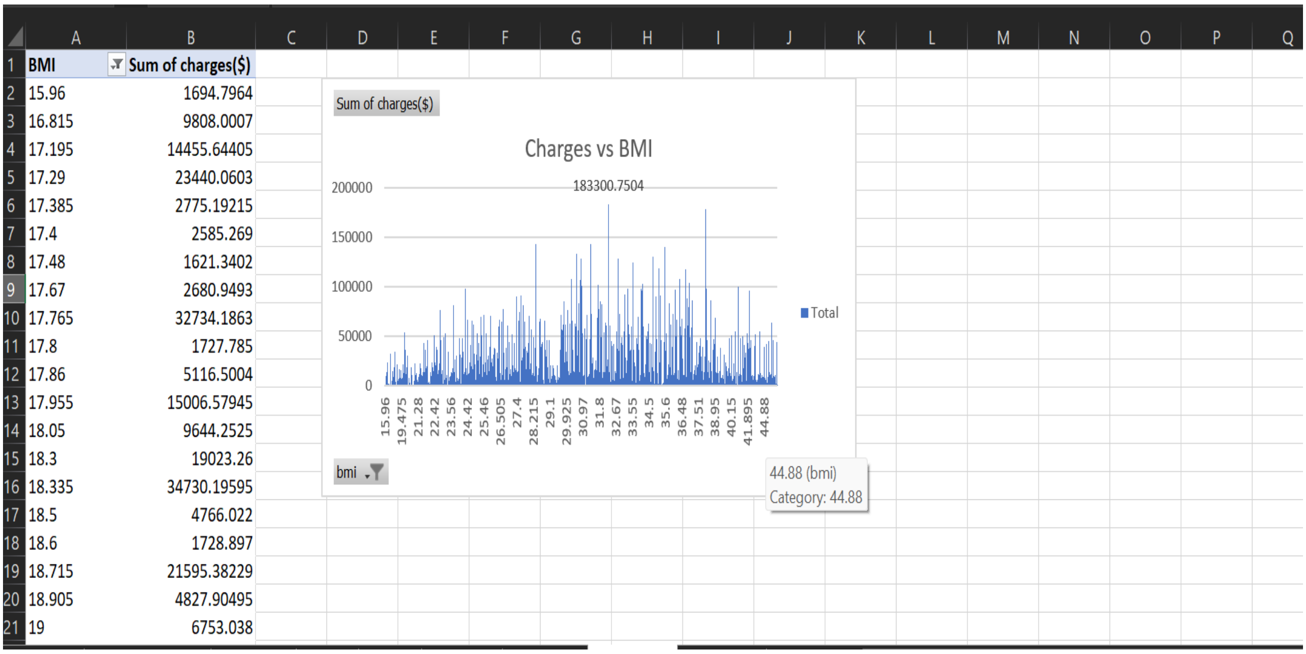
Charges vs Age



Age	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

Charges of age between 18 to 20 is too high to claim insurance.

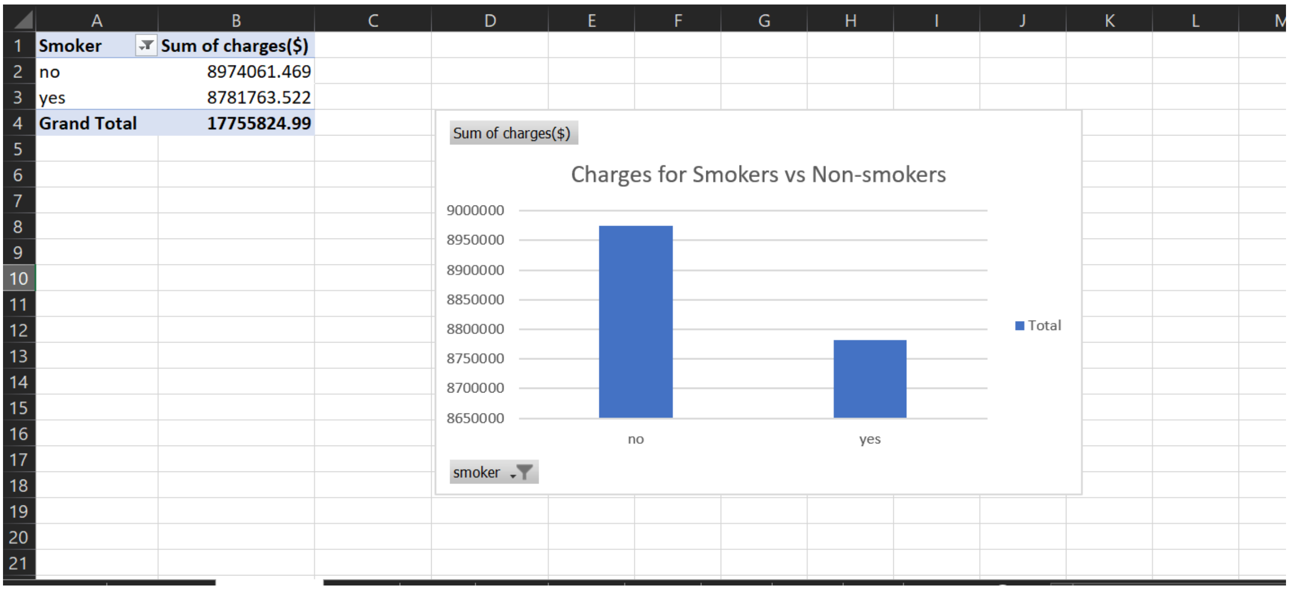
Charges vs BMI



BMI	Sum of charges($)
15.96	1694.7964
16.815	9808.0007
17.195	14455.64405
17.29	23440.0603
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	34730.19595
18.5	4766.022
18.6	1728.897
18.715	21595.38229
18.905	4827.90495
19	6753.038

Body mass index 31.8 to 32.67 charges is too high to claim insurance.

Charges for Smokers vs Non-smokers

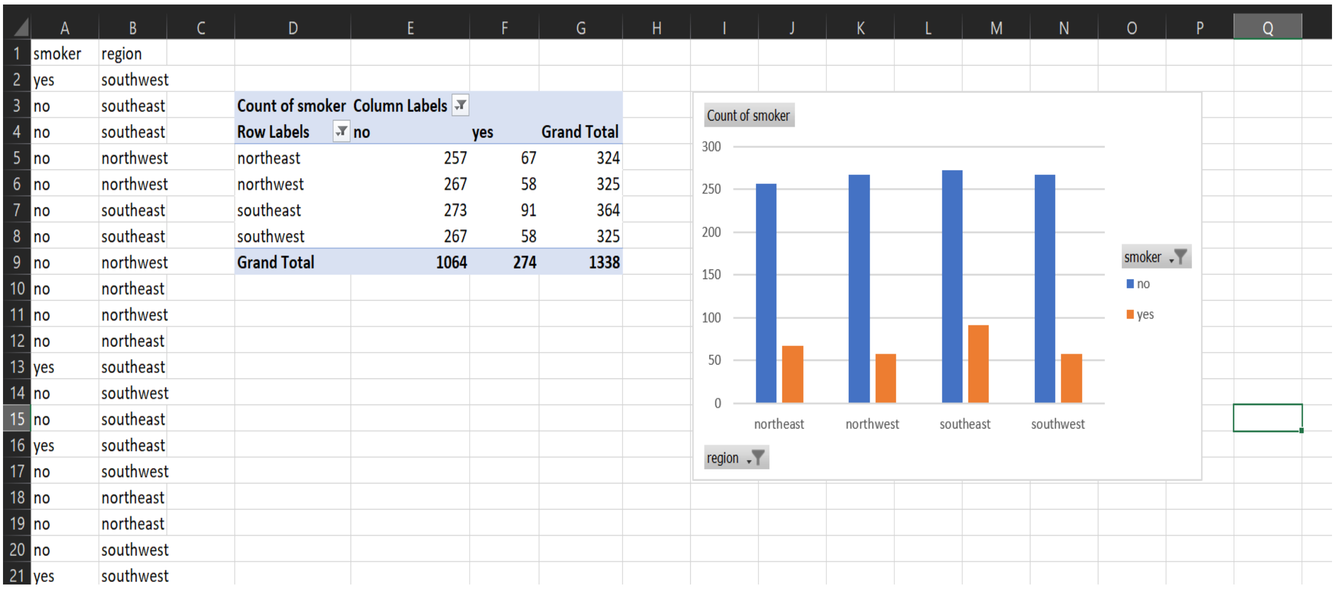


Smoker	Sum of charges($)
no	8974061.469
yes	8781763.522
Grand Total	**17755824.99**

Charges of non-smoker is high compare to smoker.

Question 1(d):

Region-wise smokers vs non-smokers



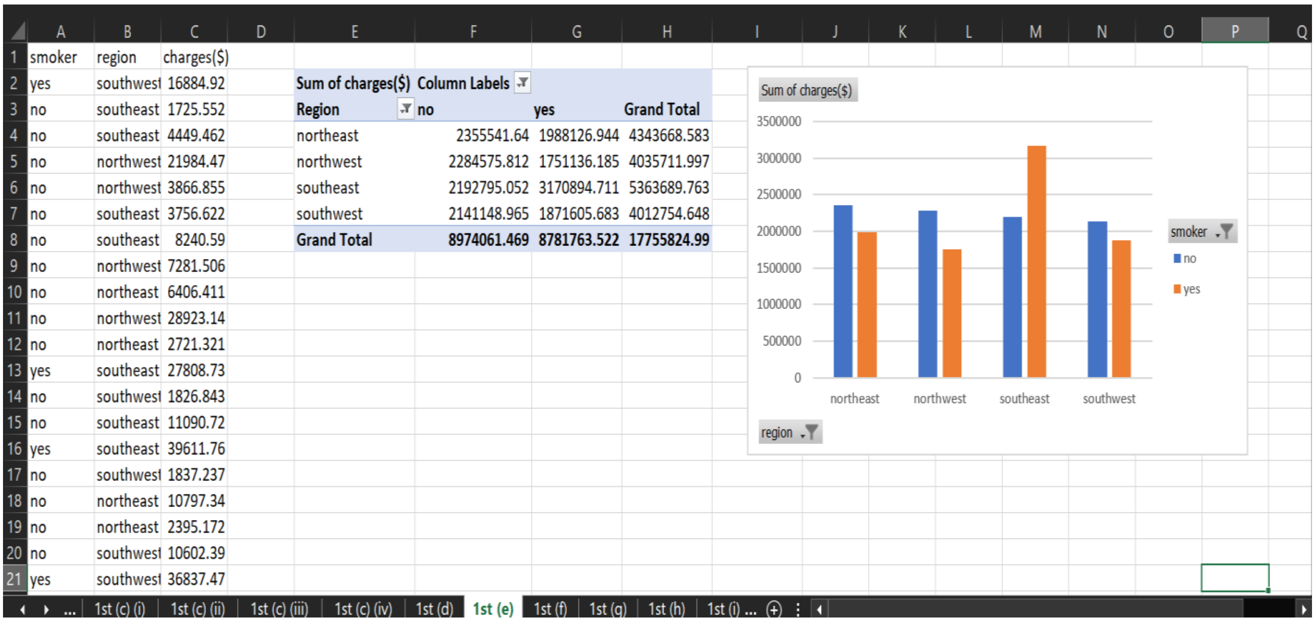
	smoker	region
2	yes	southwest
3	no	southeast
4	no	southeast
5	no	northwest
6	no	northwest
7	no	southeast
8	no	southeast
9	no	northwest
10	no	northeast
11	no	northwest
12	no	northeast
13	yes	southeast
14	no	southwest
15	no	southeast
16	yes	southeast
17	no	southwest
18	no	northeast
19	no	northeast
20	no	southwest
21	yes	southwest

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

Here southeast and northeast has more smokers and non-smoker has approx. equal to all region.

Question 1(e):

Region-wise charges for smokers vs non-smokers



	smoker	region	charges($)
2	yes	southwest	16884.92
3	no	southeast	1725.552
4	no	southeast	4449.462
5	no	northwest	21984.47
6	no	northwest	3866.855
7	no	southeast	3756.622
8	no	southeast	8240.59
9	no	northwest	7281.506
10	no	northeast	6406.411
11	no	northwest	28923.14
12	no	northeast	2721.321
13	yes	southeast	27808.73
14	no	southwest	1826.843
15	no	southeast	11090.72
16	yes	southeast	39611.76
17	no	southwest	1837.237
18	no	northeast	10797.34
19	no	northeast	2395.172
20	no	southwest	10602.39
21	yes	southwest	36837.47

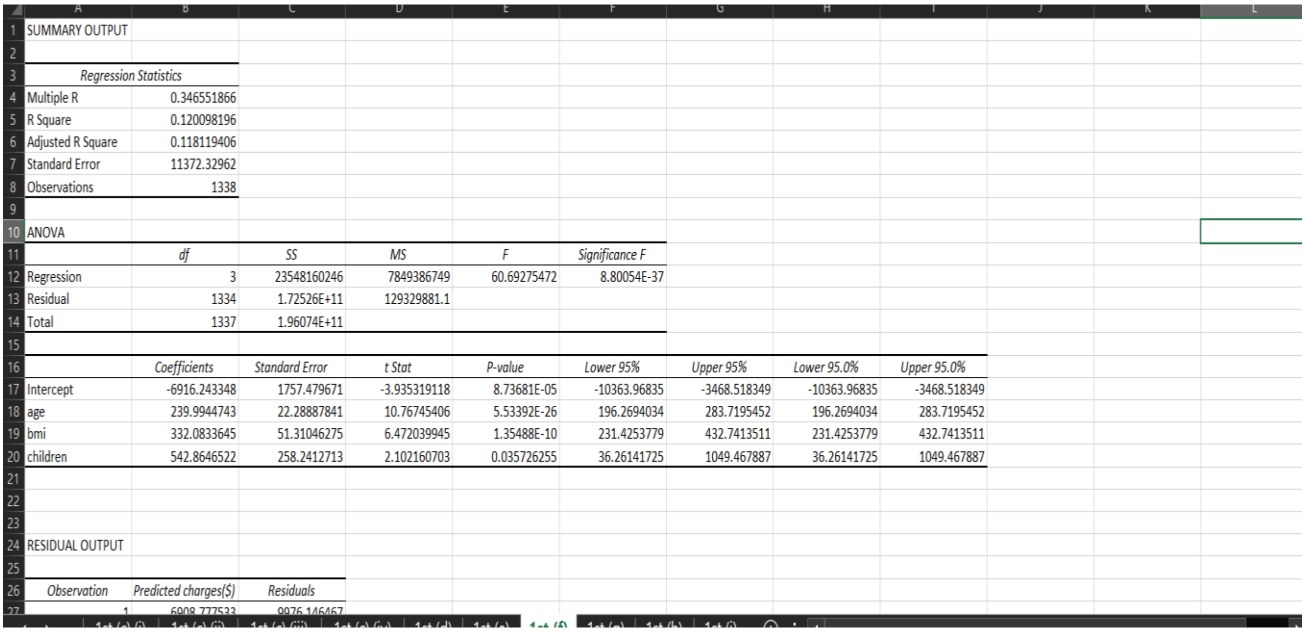
Sum of charges($)	Column Labels		
Region	no	yes	Grand Total
northeast	2355541.64	1988126.944	4343668.583
northwest	2284575.812	1751136.185	4035711.997
southeast	2192795.052	3170894.711	5363689.763
southwest	2141148.965	1871605.683	4012754.648
Grand Total	8974061.469	8781763.522	17755824.99

Here southeast has more smokers and his charges is 5363689.763 and second is northeast has more smokers and his charges is 4343668.583.

Northeast has also more non-smoker

Total charges of all region is 17755824.99.

Question 1(f):



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.346551866
R Square	0.120098196
Adjusted R Square	0.118119406
Standard Error	11372.32962
Observations	1338

ANOVA

	df	SS	MS	F	Significance F
Regression	3	23548160246	7849386749	60.69275472	8.80054E-37
Residual	1334	1.72526E+11	129329881.1		
Total	1337	1.96074E+11			

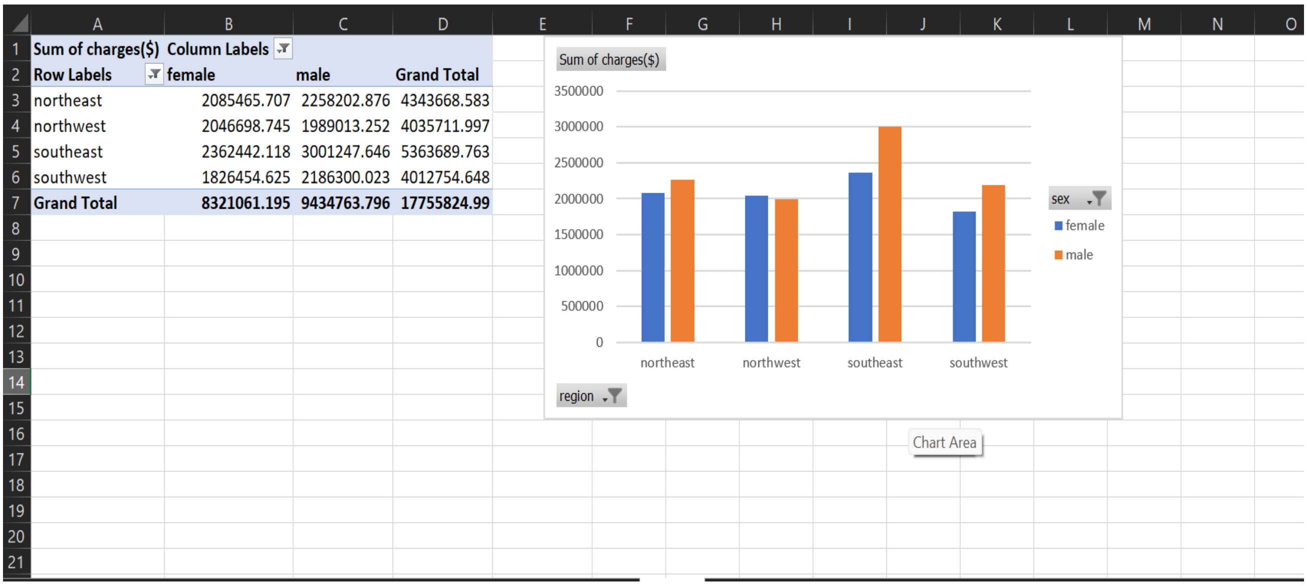
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-6916.243348	1757.479671	-3.935319118	8.73681E-05	-10363.96835	-3468.518349	-10363.96835	-3468.518349
age	239.9944743	22.28887841	10.76745406	5.53392E-26	196.2694034	283.7195452	196.2694034	283.7195452
bmi	332.0833645	51.31046275	6.472039945	1.35488E-10	231.4253779	432.7413511	231.4253779	432.7413511
children	542.8646522	258.2412713	2.102160703	0.035726255	36.26141725	1049.467887	36.26141725	1049.467887

RESIDUAL OUTPUT

Observation	Predicted charges($)	Residuals
1	6908.777533	9976.146467

Taking charges as y range and Age, BMI and children as x range and do regression analysis with residual plot.
and R Square and Adjusted R Square is approx. similar.

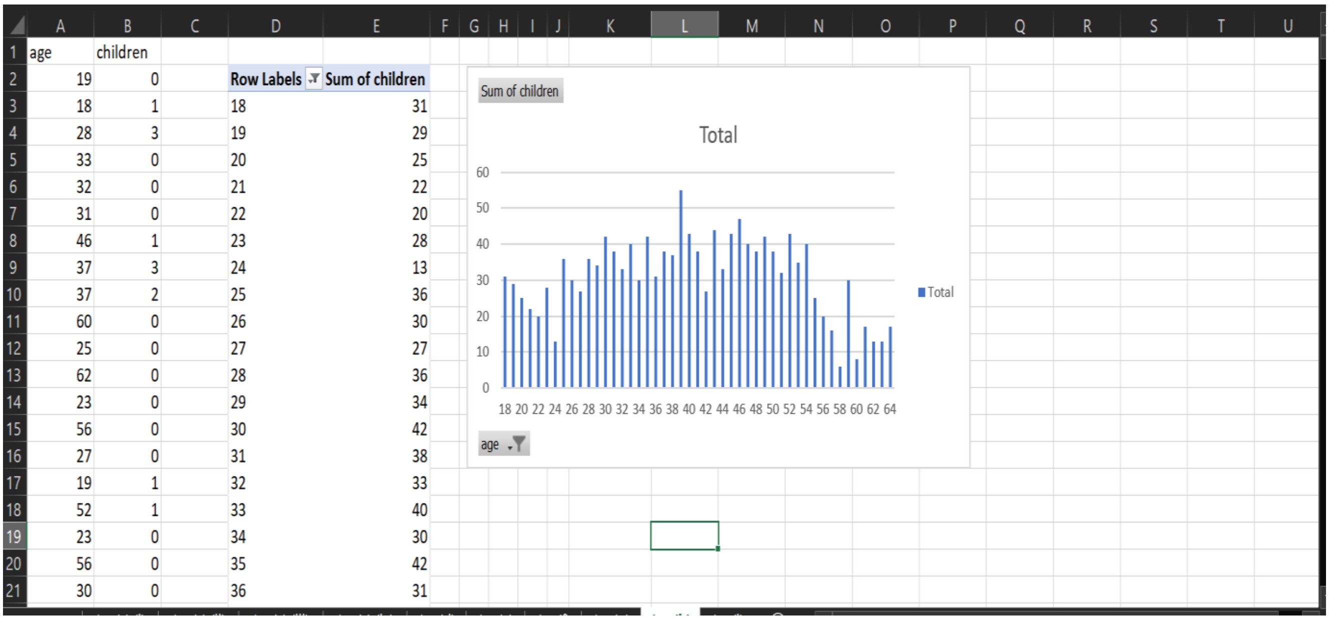
Question 1(g):



Sum of charges($)	Column Labels		
Row Labels	female	male	Grand Total
northeast	2085465.707	2258202.876	4343668.583
northwest	2046698.745	1989013.252	4035711.997
southeast	2362442.118	3001247.646	5363689.763
southwest	1826454.625	2186300.023	4012754.648
Grand Total	**8321061.195**	**9434763.796**	**17755824.99**

Taking all region, sex and charges and created pivot chart and table, region as column, sex as rows and charges as values and result is that – male is more compare to female and his charges are high.

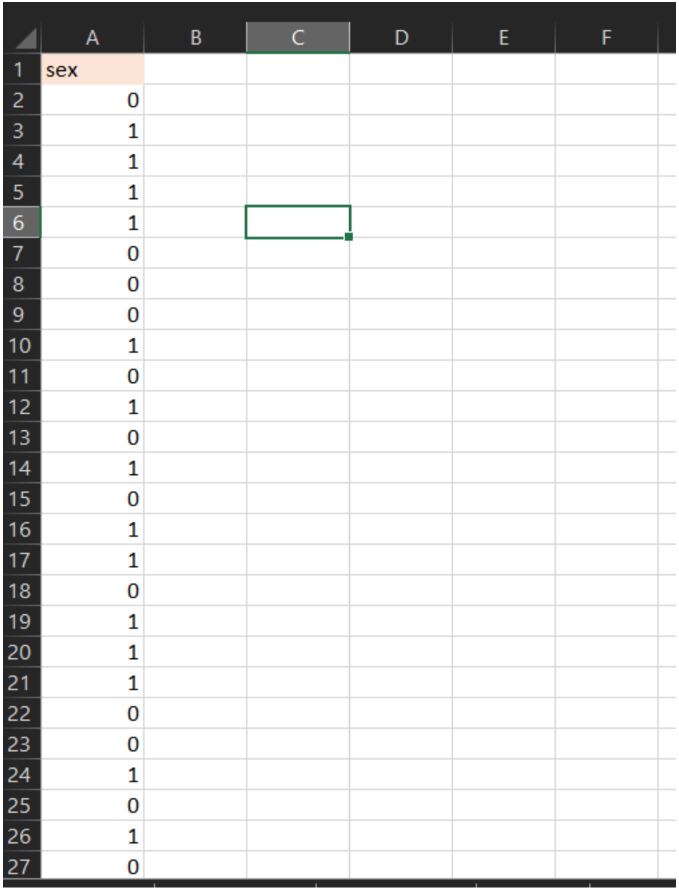
Question 1(h):



	A	B	C	D	E
1	age	children			
2	19	0		Row Labels	Sum of children
3	18	1		18	31
4	28	3		19	29
5	33	0		20	25
6	32	0		21	22
7	31	0		22	20
8	46	1		23	28
9	37	3		24	13
10	37	2		25	36
11	60	0		26	30
12	25	0		27	27
13	62	0		28	36
14	23	0		29	34
15	56	0		30	42
16	27	0		31	38
17	19	1		32	33
18	52	1		33	40
19	23	0		34	30
20	56	0		35	42
21	30	0		36	31

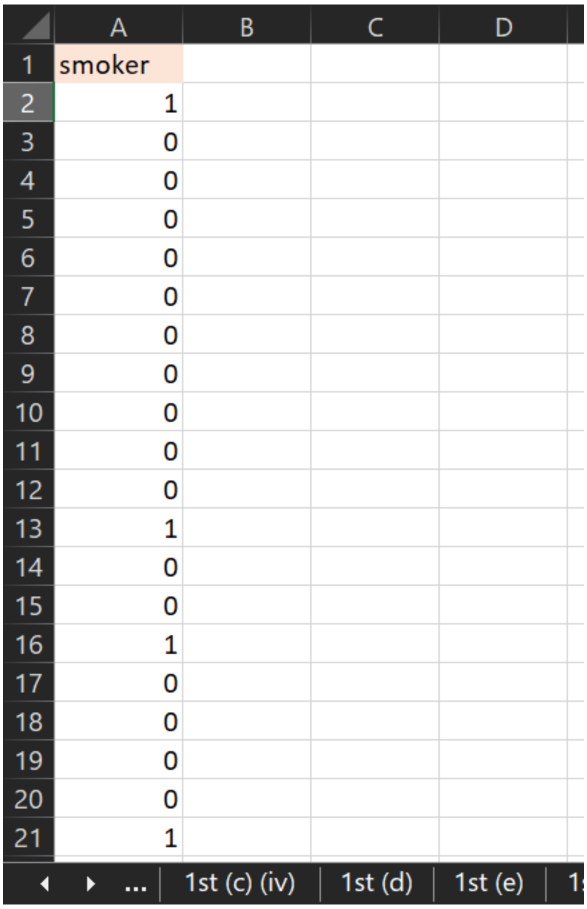
Question 2(a)

Replace all the male with 1 and female with 0.



	A
1	sex
2	0
3	1
4	1
5	1
6	1
7	0
8	0
9	0
10	1
11	0
12	1
13	0
14	1
15	0
16	1
17	1
18	0
19	1
20	1
21	1
22	0
23	0
24	1
25	0
26	1
27	0

Question 2(b):
Replace all the smokers with 1 and non-smokers with 0.



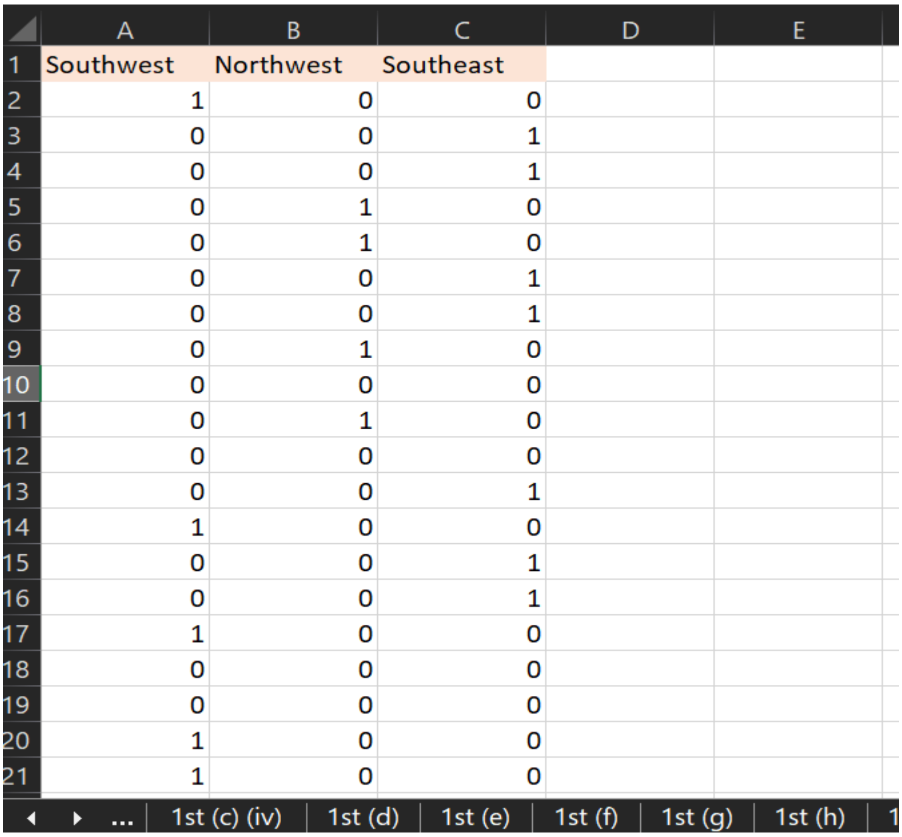
	A	B	C	D
1	smoker			
2	1			
3	0			
4	0			
5	0			
6	0			
7	0			
8	0			
9	0			
10	0			
11	0			
12	0			
13	1			
14	0			
15	0			
16	1			
17	0			
18	0			
19	0			
20	0			
21	1			

Question 2(c):

Replace whether northwest, southwest, southeast with 1 otherwise 0.

using conditional statement formula.

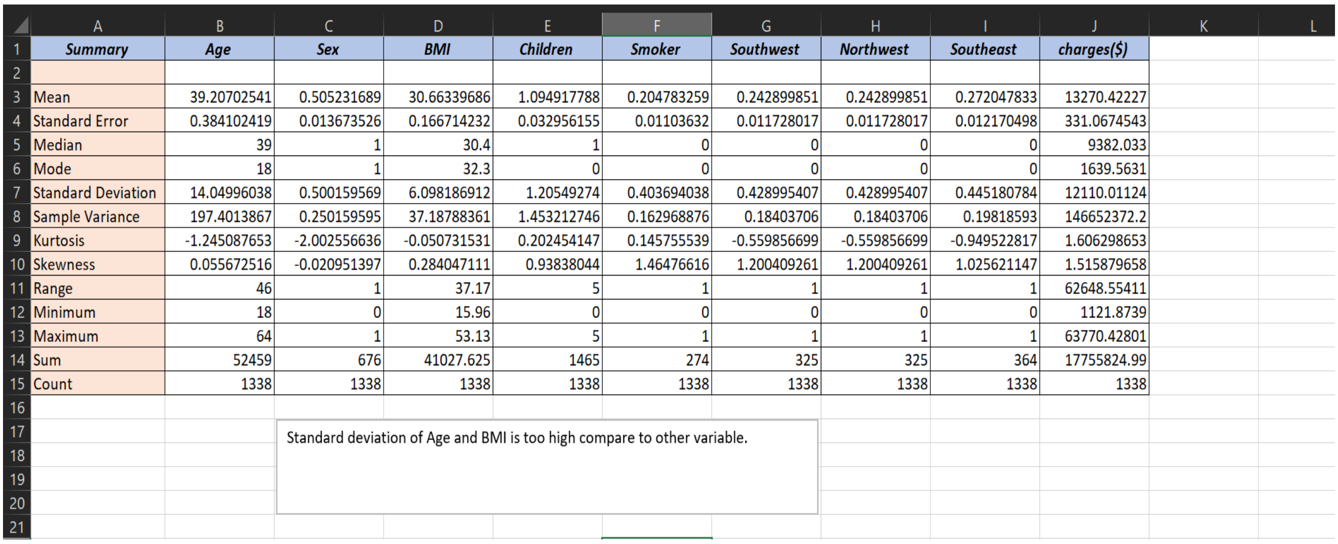
=IF(A2 = “northeast”, 1,0)



Southwest	Northwest	Southeast
1	0	0
0	0	1
0	0	1
0	1	0
0	1	0
0	0	1
0	0	1
0	1	0
0	0	0
0	1	0
0	0	0
0	0	1
1	0	0
0	0	1
0	0	1
1	0	0
0	0	0
0	0	0
1	0	0
1	0	0

Question 3:

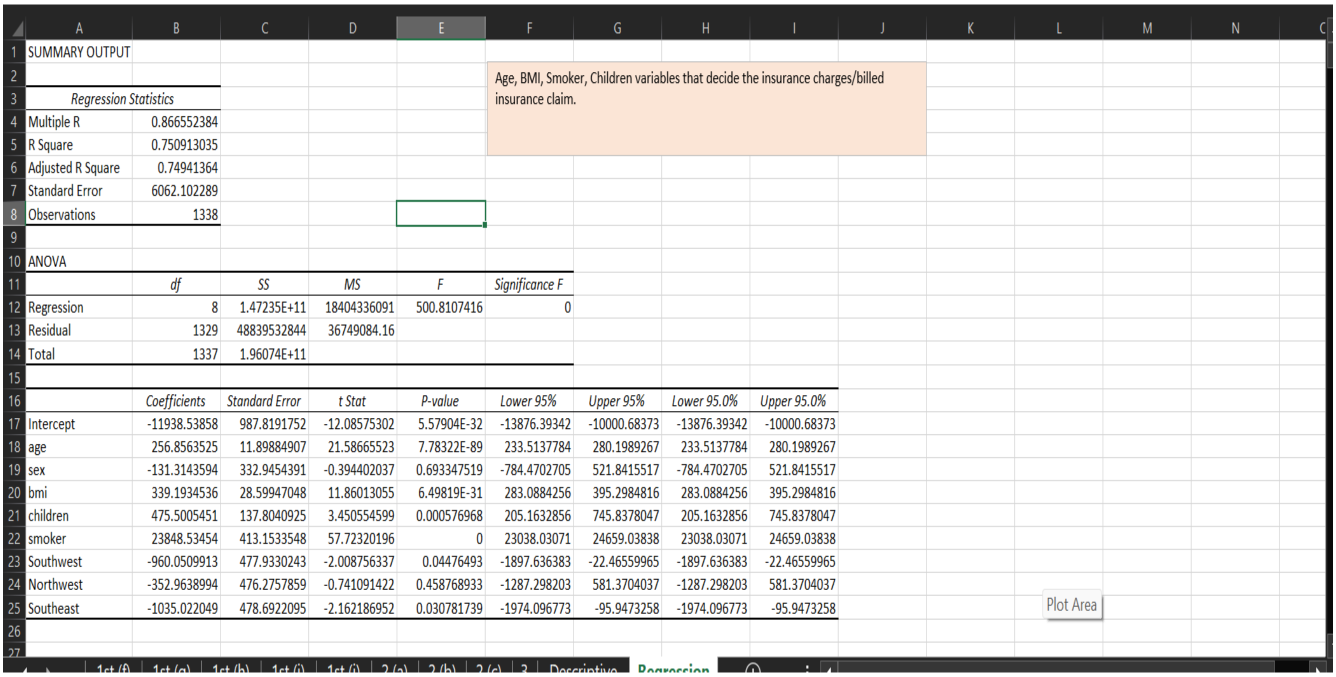
Descriptive summary analysis



Summary	Age	Sex	BMI	Children	Smoker	Southwest	Northwest	Southeast	charges($)
Mean	39.20702541	0.505231689	30.66339686	1.094917788	0.204783259	0.242899851	0.242899851	0.272047833	13270.42227
Standard Error	0.384102419	0.013673526	0.166714232	0.032956155	0.01103632	0.011728017	0.011728017	0.012170498	331.0674543
Median	39	1	30.4	1	0	0	0	0	9382.033
Mode	18	1	32.3	0	0	0	0	0	1639.5631
Standard Deviation	14.04996038	0.500159569	6.098186912	1.20549274	0.403694038	0.428995407	0.428995407	0.445180784	12110.01124
Sample Variance	197.4013867	0.250159595	37.18788361	1.453212746	0.162968876	0.18403706	0.18403706	0.19818593	146652372.2
Kurtosis	-1.245087653	-2.002556636	-0.050731531	0.202454147	0.145755539	-0.559856699	-0.559856699	-0.949522817	1.606298653
Skewness	0.055672516	-0.020951397	0.284047111	0.93838044	1.46476616	1.200409261	1.200409261	1.025621147	1.515879658
Range	46	1	37.17	5	1	1	1	1	62648.55411
Minimum	18	0	15.96	0	0	0	0	0	1121.8739
Maximum	64	1	53.13	5	1	1	1	1	63770.42801
Sum	52459	676	41027.625	1465	274	325	325	364	17755824.99
Count	1338	1338	1338	1338	1338	1338	1338	1338	1338

Standard deviation of Age and BMI is too high compare to other variable.

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



SUMMARY OUTPUT

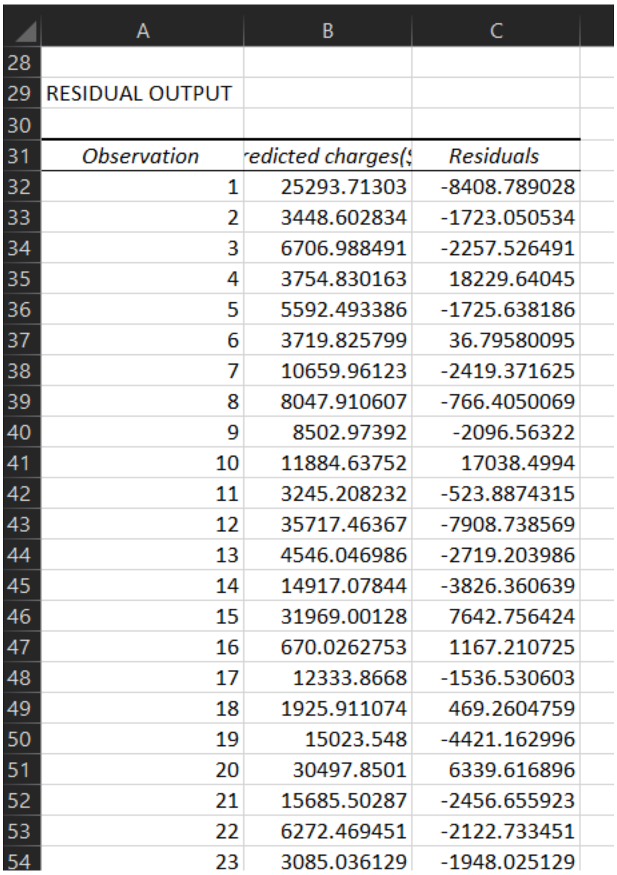
Regression Statistics	
Multiple R	0.866552384
R Square	0.750913035
Adjusted R Square	0.74941364
Standard Error	6062.102289
Observations	1338

Age, BMI, Smoker, Children variables that decide the insurance charges/billed insurance claim.

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
sex	-131.3143594	332.9454391	-0.394402037	0.693347519	-784.4702705	521.8415517	-784.4702705	521.8415517
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
smoker	23848.53454	413.1533548	57.72320196	0	23038.03071	24659.03838	23038.03071	24659.03838
Southwest	-960.0509913	477.9330243	-2.008756337	0.04476493	-1897.636383	-22.46559965	-1897.636383	-22.46559965
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



RESIDUAL OUTPUT

Observation	redicted charges(	Residuals
1	25293.71303	-8408.789028
2	3448.602834	-1723.050534
3	6706.988491	-2257.526491
4	3754.830163	18229.64045
5	5592.493386	-1725.638186
6	3719.825799	36.79580095
7	10659.96123	-2419.371625
8	8047.910607	-766.4050069
9	8502.97392	-2096.56322
10	11884.63752	17038.4994
11	3245.208232	-523.8874315
12	35717.46367	-7908.738569
13	4546.046986	-2719.203986
14	14917.07844	-3826.360639
15	31969.00128	7642.756424
16	670.0262753	1167.210725
17	12333.8668	-1536.530603
18	1925.911074	469.2604759
19	15023.548	-4421.162996
20	30497.8501	6339.616896
21	15685.50287	-2456.655923
22	6272.469451	-2122.733451
23	3085.036129	-1948.025129