Business Report

project- Insurance claim

Mohammed Toufiq

Case Study for presentation: An insurance company in US is reviewing its insurance claim/charges and trying to do a cause-and-effect analysis for future business decisions. It has collected data for its customers' age, gender, bmi, number of children/dependants, smoking habit, region they belong to, charges/bills claimed under the insurance.

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

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

- There are 1338 people who have claimed the insurance.
- The average Individual medical costs billed by health insurance is 13270.42
- The least claimed insurance charge is 1121.873 and the max is 63770.42

children	
Mean	1.094917788
Standard Error	0.032956155
Median	1
Mode	0
Standard Deviation	1.20549274
Sample Variance	1.453212746
Kurtosis	0.202454147
Skewness	0.93838044
Range	5
Minimum	0
Maximum	5
Sum	1465
Count	1338

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

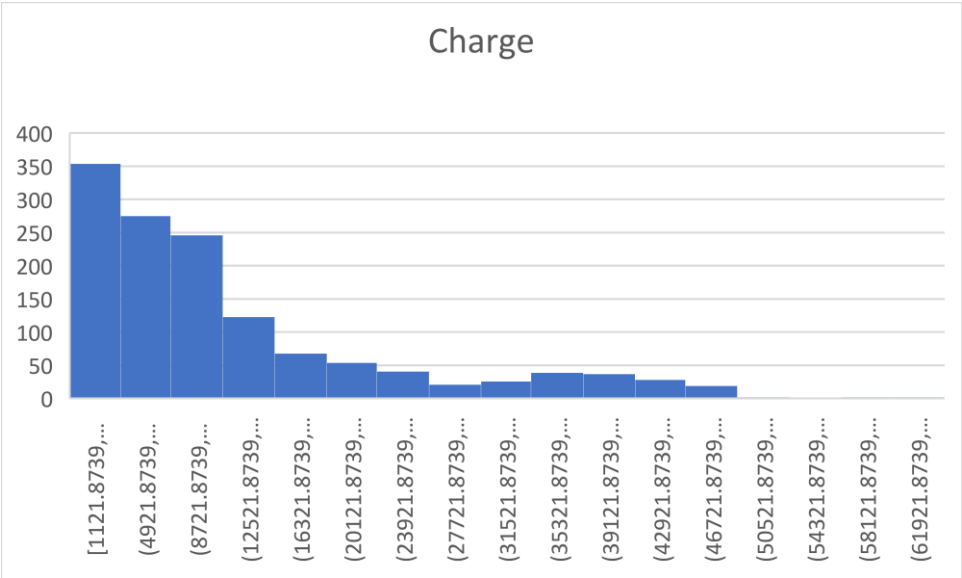
age	
Mean	39.20702541
Standard Error	0.384102419
Median	39
Mode	18
Standard Deviation	14.04996038
Sample Variance	197.4013867
Kurtosis	-1.245087653
Skewness	0.055672516
Range	46
Minimum	18
Maximum	64
Sum	52459
Count	1338

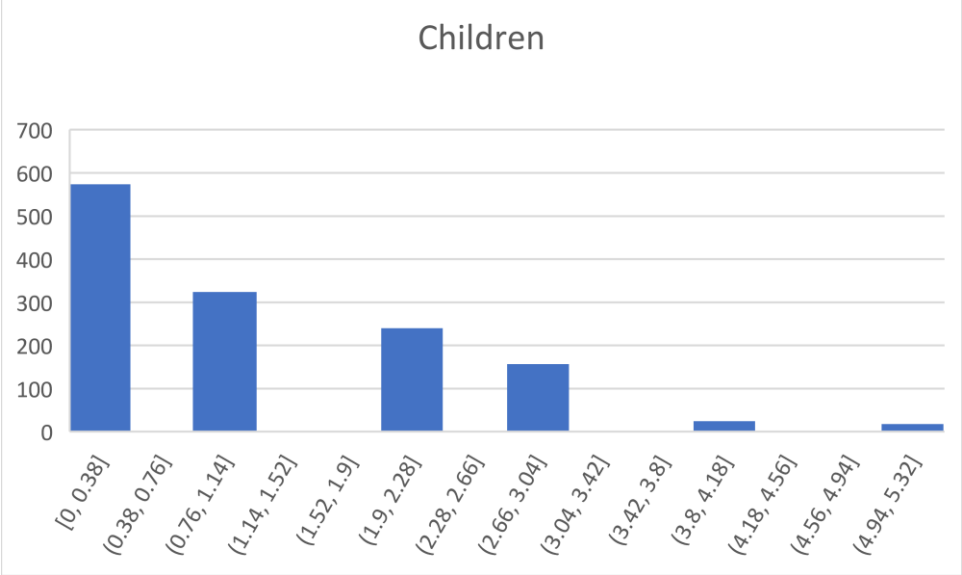
- The average age of the people claiming the insurance is 39
- Most of the people who have claimed the insurance is of age 18
- The least age of a person who have claimed insurance is 18, and the max aged person who have claimed the insurance is 64

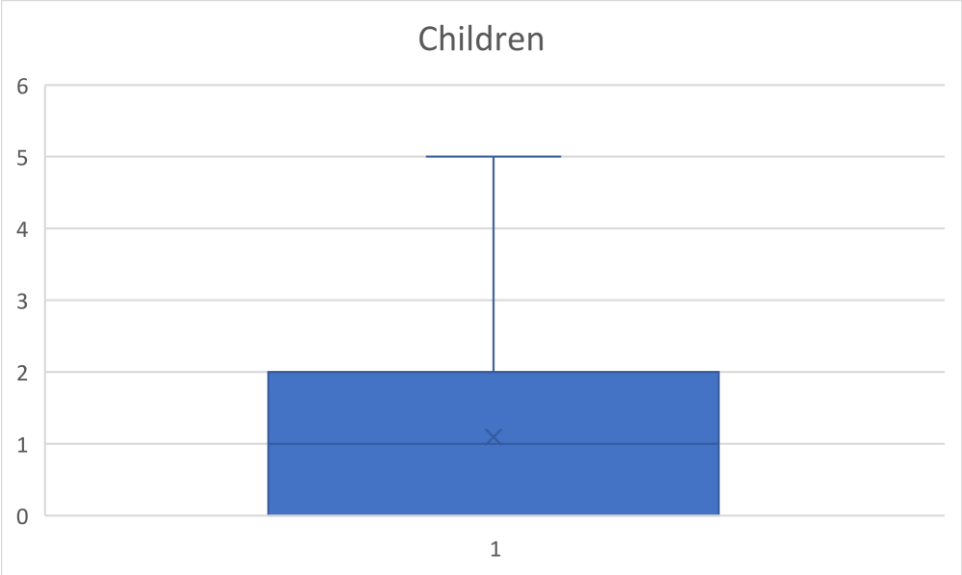
2 Identify the categorical and continuous variables

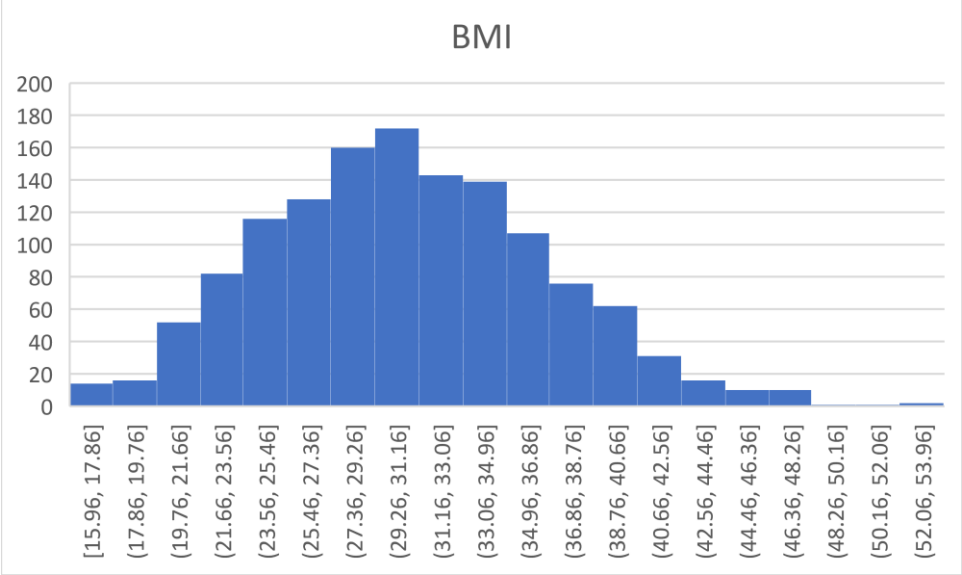
- sex, age, smoker, region are the categorical variables in the given set of data
- Bmi, children and charges are the continues variables in the given set of data

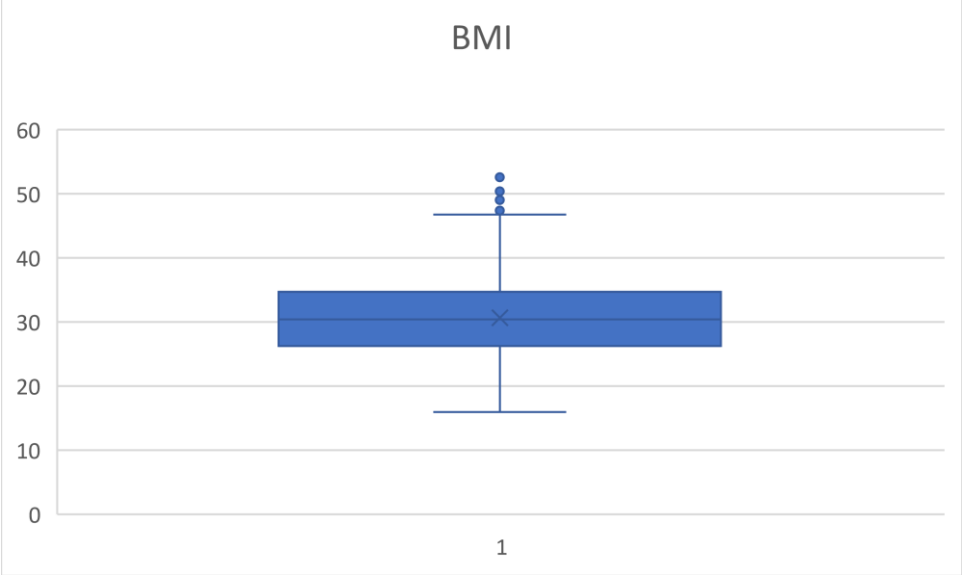
3 Make Histograms and box plots for continuous variables, do a correlation analysis.




Children
700
600
500
400
300
200
100
0
[0, 0.38]
(0.38, 0.76]
(0.76, 1.14]
(1.14, 1.52]
(1.52, 1.9]
(1.9, 2.28]
(2.28, 2.66]
(2.66, 3.04]
(3.04, 3.42]
(3.42, 3.8]
(3.8, 4.18]
(4.18, 4.56]
(4.56, 4.94]
(4.94, 5.32]


Children
6
5
4
3
2
1
0
1


BMI
200
180
160
140
120
100
80
60
40
20
0
[15.96, 17.86]
(17.86, 19.76]
(19.76, 21.66]
(21.66, 23.56]
(23.56, 25.46]
(25.46, 27.36]
(27.36, 29.26]
(29.26, 31.16]
(31.16, 33.06]
(33.06, 34.96]
(34.96, 36.86]
(36.86, 38.76]
(38.76, 40.66]
(40.66, 42.56]
(42.56, 44.46]
(44.46, 46.36]
(46.36, 48.26]
(48.26, 50.16]
(50.16, 52.06]
(52.06, 53.96]



	bmi	*children*	*charges($)*
bmi	1		
children	0.012758901	1	
charges($)	0.198340969	0.067998227	1

4. Make relevant Pivot tables and charts for:

1) Male/Female ratio and which gender has more smokers

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

Male gender has more smokers

2) Charges vs Age

Age	Average of charges($)	Sum of charges($)	Count of age
18	7086.217556	488949.0114	69
19	9747.909335	662857.8348	68
20	10159.69774	294631.2344	29
21	4730.46433	132453.0012	28
22	10012.9328	280362.1185	28
23	12419.82004	347754.9611	28
24	10648.01596	298144.4469	28
25	9838.365311	275474.2287	28
26	6133.825309	171747.1086	28
27	12184.70172	341171.6482	28
28	9069.187564	253937.2518	28
29	10430.15873	281614.2856	27
30	12719.11036	343415.9797	27
31	10196.98057	275318.4755	27
32	9220.300291	239727.8076	26
33	12351.53299	321139.8577	26
34	11613.52812	301951.7311	26
35	11307.18203	282679.5508	25
36	12204.47614	305111.9035	25
37	18019.91188	450497.7969	25
38	8102.733674	202568.3419	25
39	11778.24295	294456.0736	25
40	11772.25131	317850.7854	27
41	9653.74565	260651.1325	27
42	13061.03867	352648.0441	27
43	19267.27865	520216.5236	27
44	15859.39659	428203.7079	27
45	14830.19986	430075.7958	29
46	14342.59064	415935.1285	29
47	17653.99959	511965.9882	29
48	14632.50045	424342.5129	29
49	12696.00626	355488.1754	28
50	15663.0033	454227.0957	29
51	15682.25587	454785.4202	29
52	18256.26972	529431.8219	29
53	16020.93076	448586.0611	28
54	18758.54648	525239.3013	28
55	16164.54549	420278.1827	26
56	15025.51584	390663.4118	26
57	16447.18525	427626.8165	26
58	13878.92811	346973.2028	25
59	18895.86953	472396.7383	25
60	21979.41851	505526.6257	23
61	22024.45761	506562.525	23
62	19163.85657	440768.7012	23
63	19884.99846	457354.9646	23
64	23275.53084	512061.6784	22
Grand Total	**13270.42227**	**17755824.99**	**1338**

3) Charges vs BMI

BMI Category	Average of charges($)	Sum of charges($)	Count of bmi category
<20	8838.561135	362381.0065	41
>45	17547.92675	350958.535	20
20-25	10572.3725	2156763.99	204
25-30	10987.50989	4241178.818	386
30-35	14419.67497	5638092.913	391
35-40	17022.25888	3830008.249	225
40-45	16569.59831	1176441.48	71
Grand Total	**13270.42227**	**17755824.99**	**1338**

4) Charges for Smokers vs Non-smokers

Sum of charges($)	Column Labels		
Row Labels	no	yes	Grand Total
1121.8739	1121.8739		1121.8739
1131.5066	1131.5066		1131.5066
1135.9407	1135.9407		1135.9407
1136.3994	1136.3994		1136.3994
1137.011	1137.011		1137.011
1137.4697	1137.4697		1137.4697
1141.4451	1141.4451		1141.4451
1146.7966	1146.7966		1146.7966
1149.3959	1149.3959		1149.3959
1163.4627	1163.4627		1163.4627
1241.565	1241.565		1241.565
1242.26	1242.26		1242.26
1242.816	1242.816		1242.816
1252.407	1252.407		1252.407
1253.936	1253.936		1253.936
1256.299	1256.299		1256.299

5) Region-wise Smokers vs non-smokers analysis with one or more pivot table and 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**

the more smokers are from north east

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

Sum of charges($)	Column Labels		
Row Labels	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**

7) Have charges got something to do with no. of dependants?

	charges($)	*children*
charges($)	1	
children	0.067998227	1

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

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

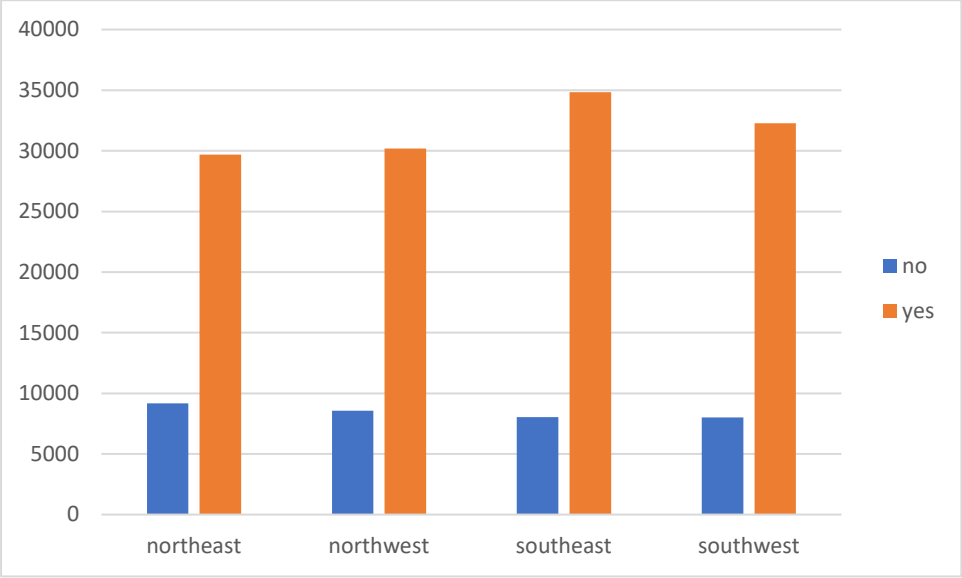
- Average charge in northeast region with 5 number of children is lowest among all regions
- Average charge in southeast region with 3 number of children is highest among all regions
- In northeast region Highest average charge is for 1 number of children and lowest is for 5 number of children
- In northwest region Highest average charge is for 3 number of children and lowest is for 5 number of children
- In southeast region Highest average charge is for 3 number of children and lowest is for 5 number of children
- In southwest region Highest average charge is for 2 number of children and lowest is for 5 number of children
- Lowest Average Charge for all regions has 5 number of children.

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

Average charge of among smokers and non-smokers

Region	Average charge Non-Smokers	Smokers	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**

Average charge Among smokers in various regions



Above is table and chart of region wise Average charge of among smokers and non-smokers.

Edit the data as following, to obtain dummy variables:

1. ***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.***
2. ***Smoker: Replace all the "Smokers" with "1" and "Non-smokers" with "0".***
3. ***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, which ever 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.***

charges($)	children	bmi	age	Sex Numerical representatin	Smokers numerical representation	North west	South East	South west	BMI_categorical
16884.924	0	27.9	19	0	1	0	0	1	25-30
1725.5523	1	33.77	18	1	0	0	1	0	30-35
4449.462	3	33	28	1	0	0	1	0	30-35
21984.47061	0	22.705	33	1	0	1	0	0	20-25
3866.8552	0	28.88	32	1	0	1	0	0	25-30
3756.6216	0	25.74	31	0	0	0	1	0	25-30
8240.5896	1	33.44	46	0	0	0	1	0	30-35
7281.5056	3	27.74	37	0	0	1	0	0	25-30
6406.4107	2	29.83	37	1	0	0	0	0	25-30
28923.13692	0	25.84	60	0	0	1	0	0	25-30
2721.3208	0	26.22	25	1	0	0	0	0	25-30
27808.7251	0	26.29	62	0	1	0	1	0	25-30
1826.843	0	34.4	23	1	0	0	0	1	30-35
11090.7178	0	39.82	56	0	0	0	1	0	35-40
39611.7577	0	42.13	27	1	1	0	1	0	40-45

Above is an example snip shot of the newly edited data.

3) 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.

Column1	*age*	*bmi*	*charges($)*
Mean	39.20703	30.6634	13270.4223
Standard E	0.384102	0.166714	331.067454
Median	39	30.4	9382.033
Mode	18	32.3	1639.5631
Standard D	14.04996	6.098187	12110.0112
Sample Va	197.4014	37.18788	146652372
Kurtosis	-1.24509	-0.05073	1.60629865
Skewness	0.055673	0.284047	1.51587966
Range	46	37.17	62648.5541
Minimum	18	15.96	1121.8739
Maximum	64	53.13	63770.428
Sum	52459	41027.63	17755825
Count	1338	1338	1338

SUMMARY OUTPUT

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

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
children	475.5005451	137.8040925	3.450554599	0.000576968	205.1632856	745.8378047	205.1632856	745.8378047
bmi	339.1934536	28.59947048	11.86013055	6.49819E-31	283.0884256	395.2984816	283.0884256	395.2984816
age	256.8563525	11.89884907	21.58665523	7.78322E-89	233.5137784	280.1989267	233.5137784	280.1989267
Sex Numerical rep	-131.3143594	332.9454391	-0.394402037	0.693347519	-784.4702705	521.8415517	-784.4702705	521.8415517
Smokers numeric	23848.53454	413.1533548	57.72320196	0	23038.03071	24659.03838	23038.03071	24659.03838
North west	-352.9638994	476.2757859	-0.741091422	0.458768933	-1287.298203	581.3704037	-1287.298203	581.3704037
South East	-1035.022049	478.6922095	-2.162186952	0.030781739	-1974.096773	-95.9473258	-1974.096773	-95.9473258
South west	-960.0509913	477.9330243	-2.008756337	0.04476493	-1897.636383	-22.46559965	-1897.636383	-22.46559965

From the above regression analysis:

- Adjusted R squared value is 0.749, which is very high and hence we can say that the linear model fits the data well
- P values for all the variables are less than 5% except sex and northwest region. Hence, we can call these variables insignificant and do regression analysis by ignoring these variables if needed.
- Also, we can be p value for smokers is very low and coefficient for smokers is very high. **Hence, we can say they Smokers is the most important variable which explains the behaviour of dependent variable (which is charge**s)
- The equation governing this regression line is
  Charge = 256.854*age + 339.19*bmi + 475.5*children - 131.31*sex + 23848*smoker - 352.964*northwest + 1035.02*southeast - 960.051*southwest - 11938.5