HEALTH INSURANCE CLAIM PROJECT

K.PAVANSAI GOUD
GLCA March 2023

Contents

1. Perform the Exploratory Data Analysis on the data.

a) Identify the categorical and continuous variables.

continuous	categorical
Age	Sex
BMI	Children
Charges($)	Smoker
	Region

Contents

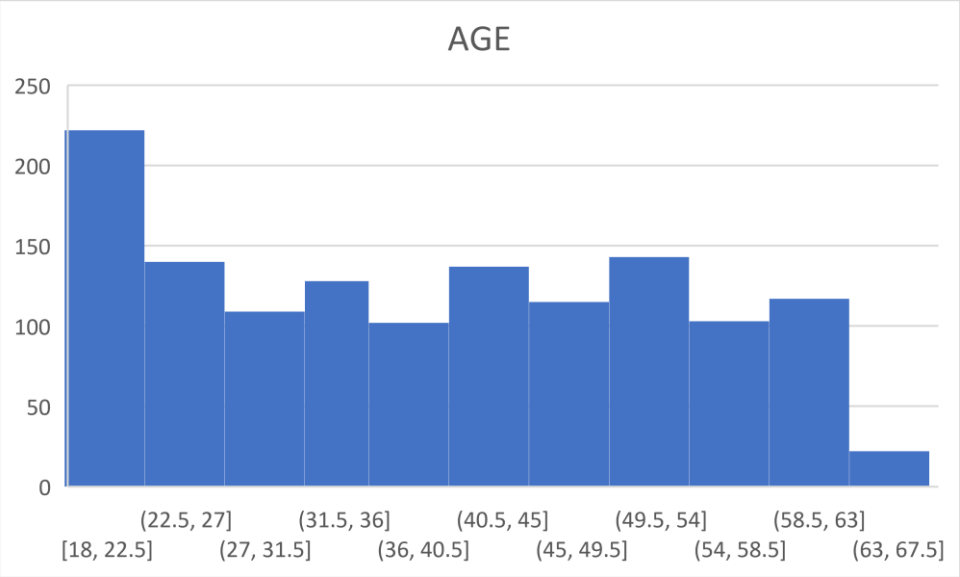
Continuous variable:

In statistics, a continuous variable is a kind of quantitative variable that can have any value within a range.

Categorical data:

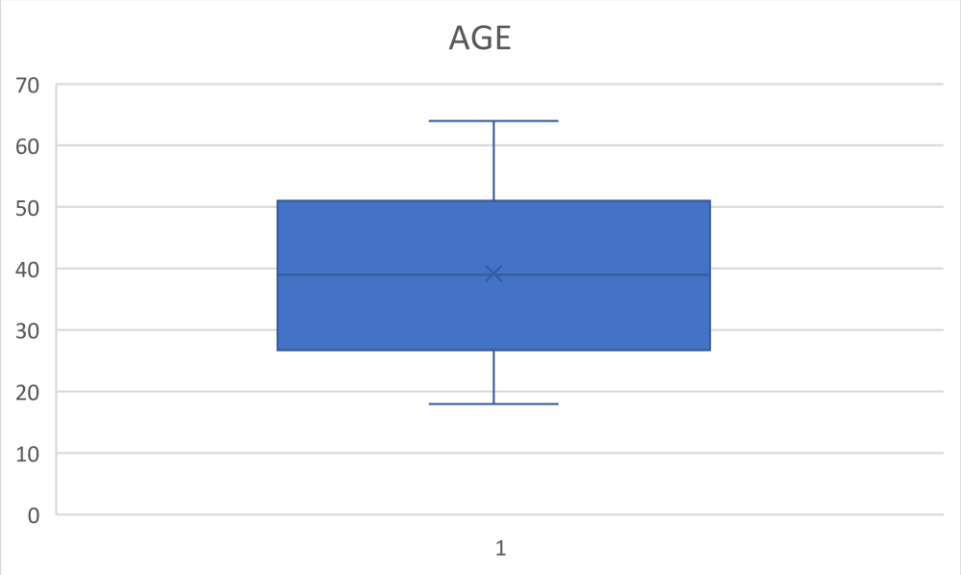
Data that represents qualities or characteristics rather than quantities is referred to as categorical data, also known as qualitative data.

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

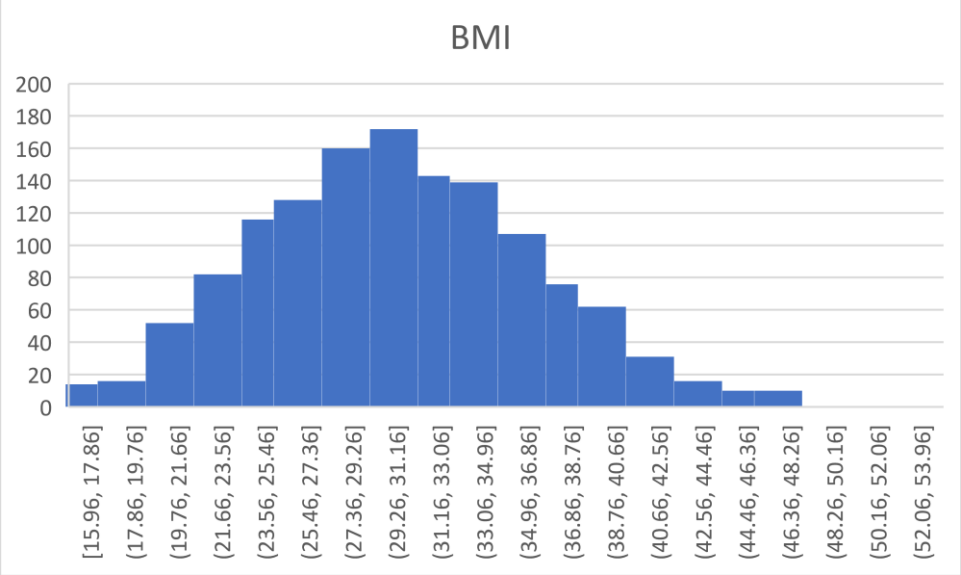


- A histogram is what the above graph is called. •
- The graph illustrates the various categorical age bins.
- Here, we can see that the highest age falls under the category of (18–22.5), and the next two highest ages fall into the ranges of (22.5–27) and (49.5–54), which are the closest in terms of numbers with only a three-point difference.

Contents

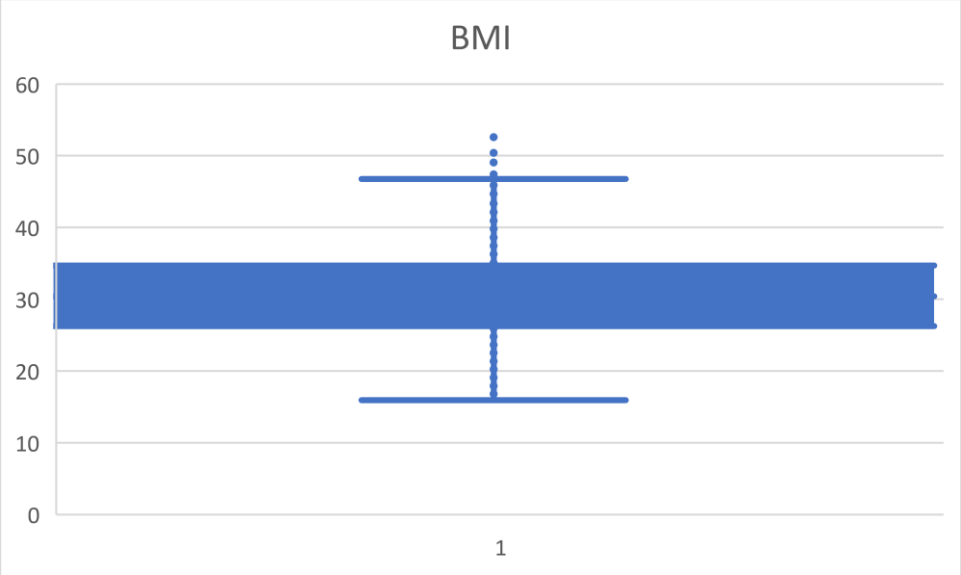


The above Boxplot Gives us the indication of the values which are present in the data.

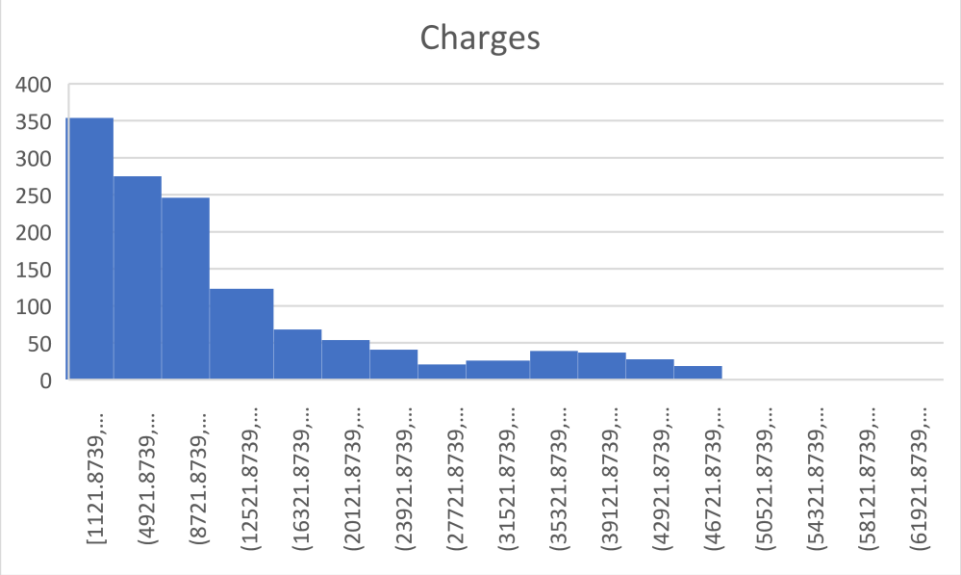


BMI: represents the BMI, which gives information about body weights that are high or low in relation to height. The graph above displays the highest and lowest ages along with weight.

Contents

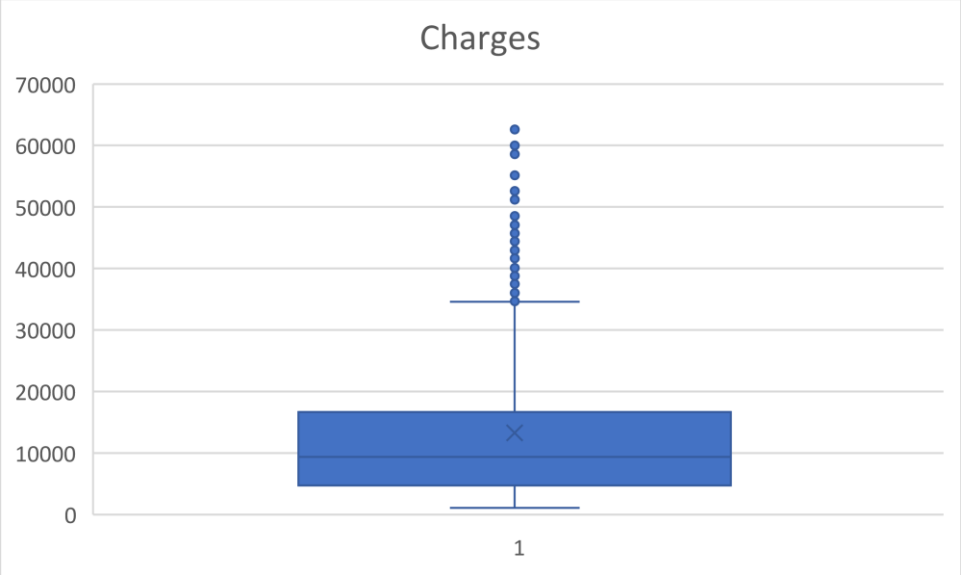


There are four outliers in the box plot above. And the whisker line contains most values. We increase the upper whisker with those upper outliers in accordance with box plot theory.



The histogram up top displays the count of values in various charge intervals. The graph has a 3800 interval and runs from 1121.87 to 65721.87.

Contents



Above box plot shown that there are many outliers are present in charges variable.

Correlation:

	age	*bmi*	*charges($)*
age	1		
bmi	0.109272	1	
charges($)	0.299008	0.198341	1

• The co-relation between age, BMI, and charges is depicted in the correlation table above.

• As I could see from the above table, there is a 10% correlation between BMI and age, and there is a 29% correlation between Charges-Age and Charges-BMI.

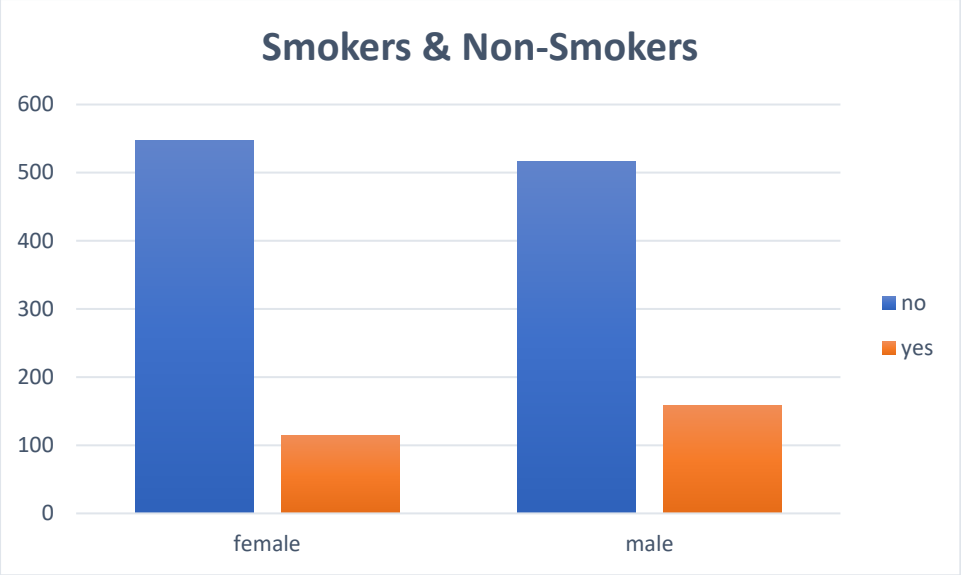
• All of the relationships, in my opinion, are nowhere near 50%. Age, BMI, and Charges do have a relationship, but it is not as strong as it could be.

c) Make relevant Pivot tables and charts for:

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

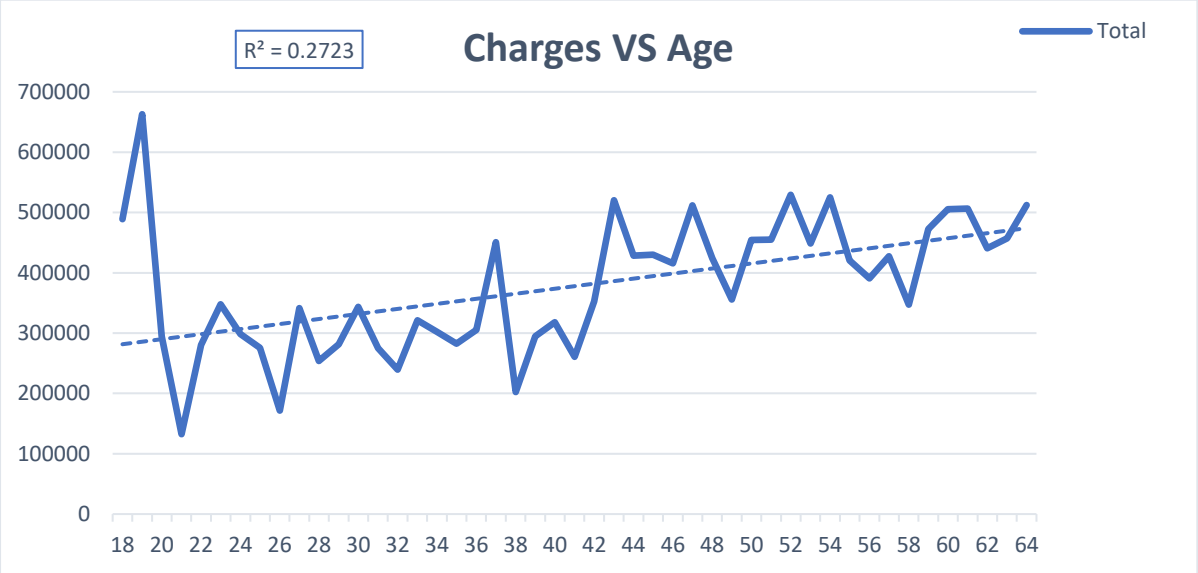
Count of smoker	Column Labels		
Row Labels	**no**	**yes**	**Grand Total**
female	547	115	662
male	517	159	676
Grand Total	**1064**	**274**	**1338**

Contents



- The above chart represents the smokers of male & Female.
- We can Males are higher in the smoking category when compared to females.

ii. Charges vs Age

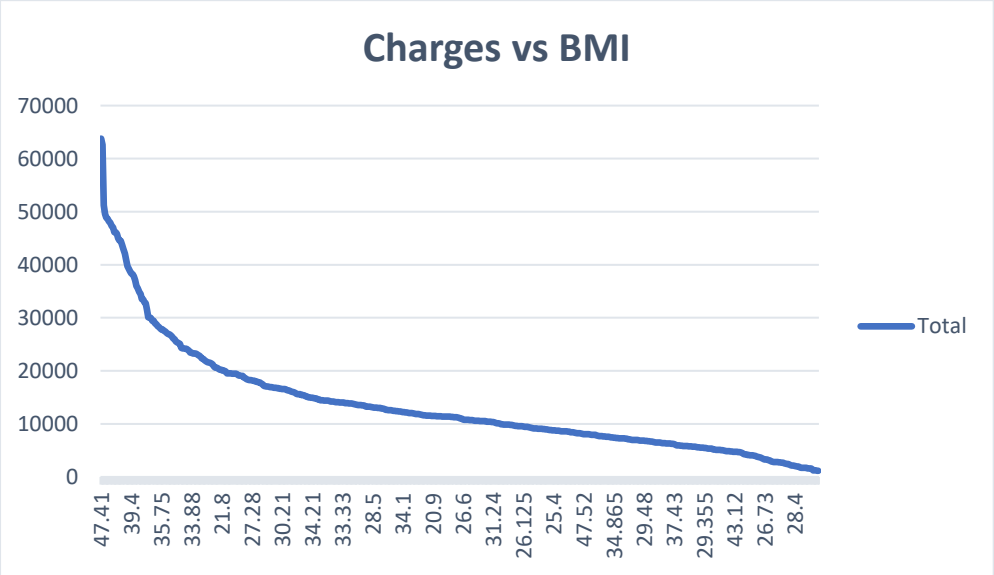


From the above plot as per my observation
The chargers are increasing as we can see the trend are moving to the upwards as the ages are increasing the charges.

Contents

Row Labels	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

iii. Charges vs BMI

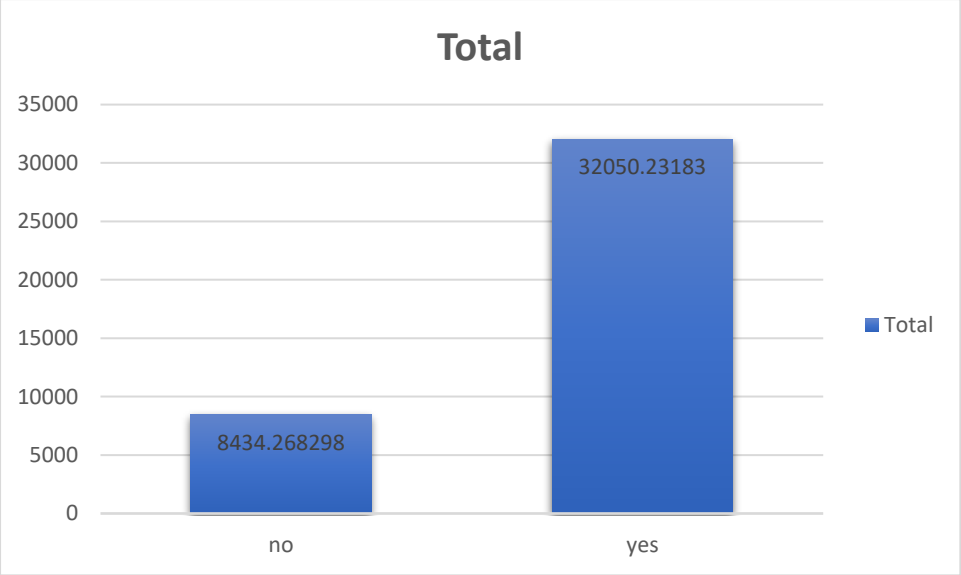


The above chart point BMI 47.41 charges are more and at BMI 28.4 we get less charges.

And I observe there is no such pattern between BMI and charges. But is was increased slightly by increasing of BMI.

Contents

iv. Charges for Smokers vs Non-smokers

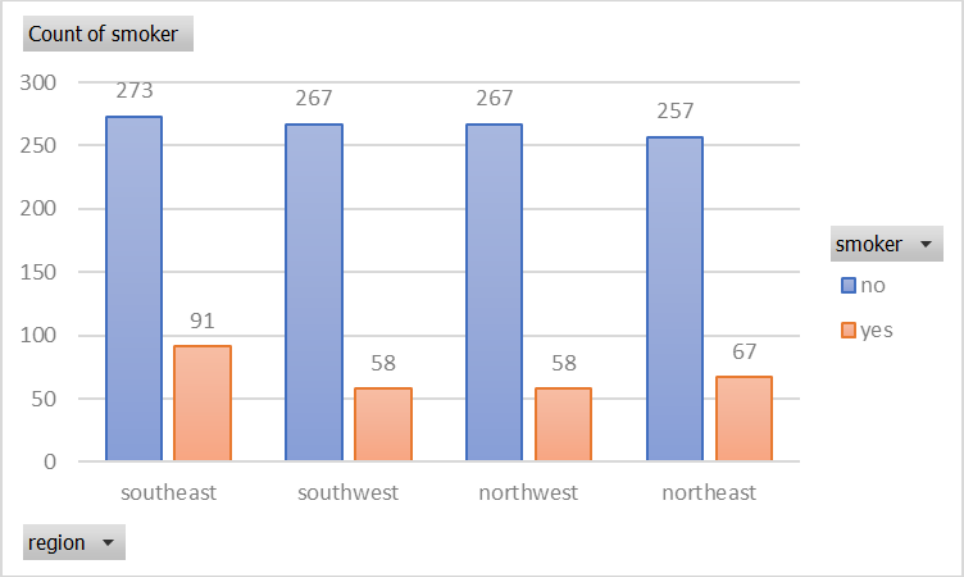


I have taken column chart for smoker's vs non-smokers. Non-smokers are more than smokers. But average charges are paid by smokers are high. But due to large members in non-smokers sum of charges are more for non-smokers.

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

Count of smoker	smoker(yes/no)		
Row Labels	no	yes	Grand Total
southeast	273	91	364
southwest	267	58	325
northwest	267	58	325
northeast	257	67	324
Grand Total	**1064**	**274**	**1338**

Contents

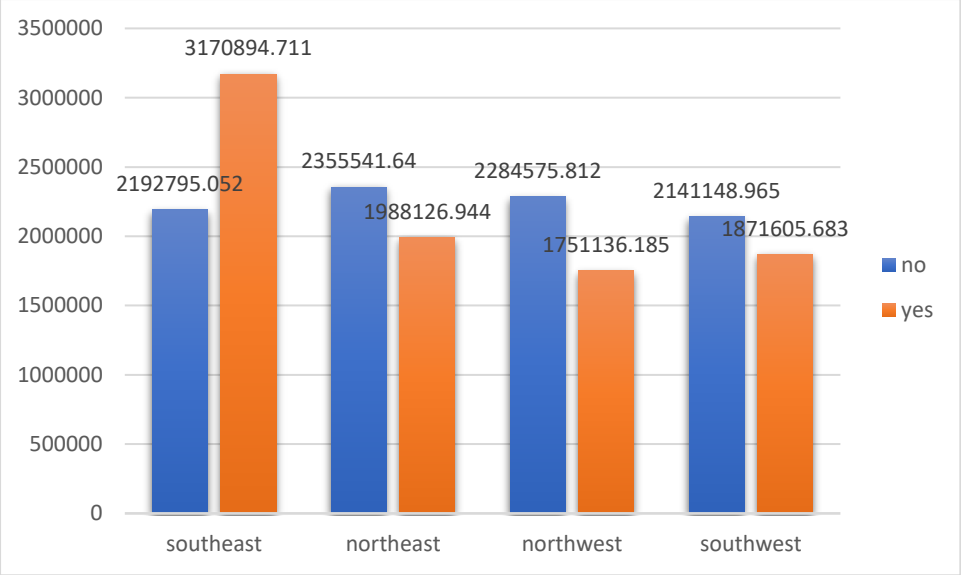


As per my observation
The non-smoker is higher than the smoker,
coming to region wise the sought east are having hinger smokers when compared to 4 regions.

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

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

Contents



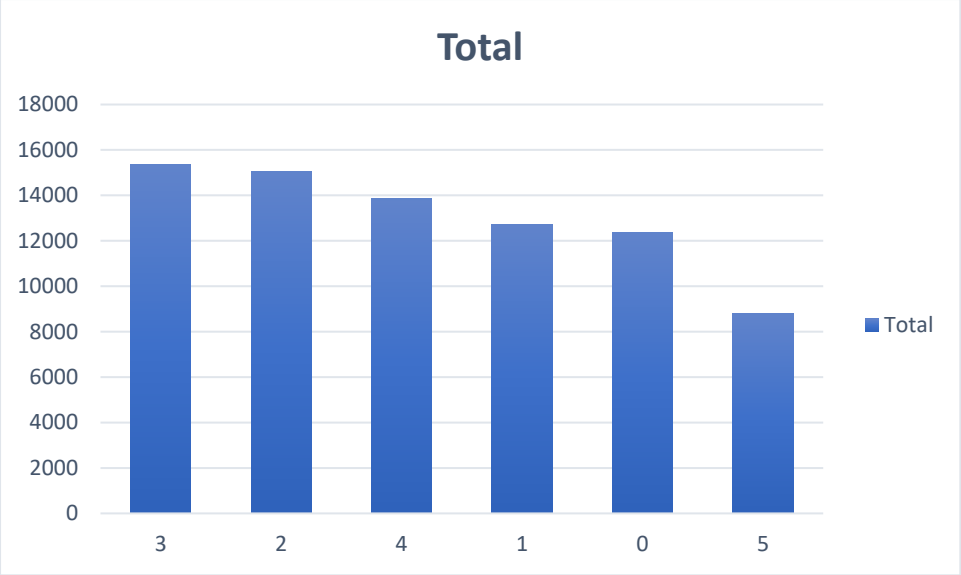
Here sum charges for southeast region is high and least was southwest. Coming to smokers' northwest was less charges compared to other.

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

Row Labels	Sum of charges ($)
5	158148.6345
4	346266.4078
3	2410784.984
2	3617655.296
1	4124899.673
0	7098069.995
Grand Total	**17755824.99**

	children	*charges($)*
children	1	
charges($)	0.067998	1

Contents



No, according to my observation I didn't find any strong relation between charges and no of dependents. Correlation of these shows nearly 6 percent only.

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

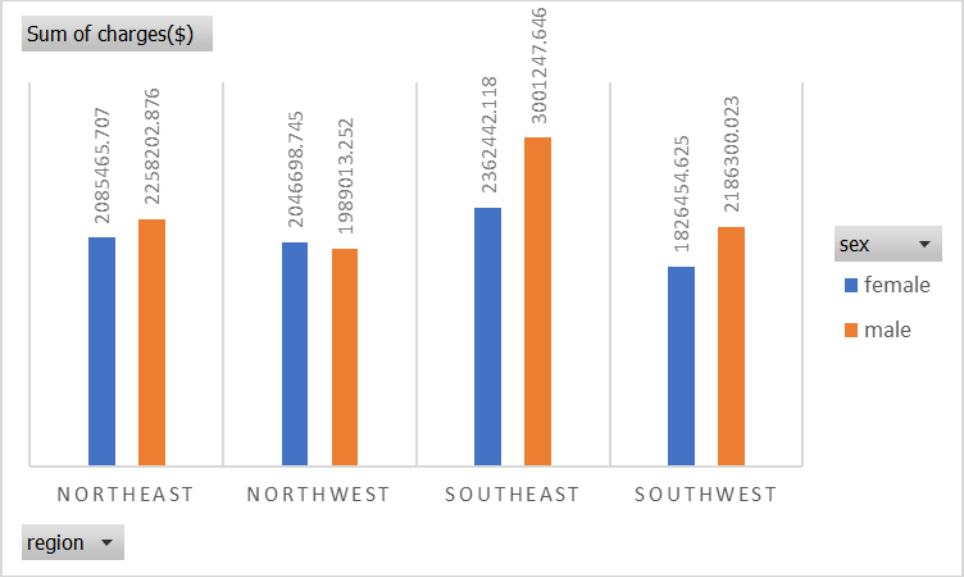
Sum of charges($)	Column Labels				
Row Labels	northeast	northwest	southeast	southwest	Grand Total
0	1709090.011	1494816.961	2246649.335	1647513.688	7098069.995
1	1255885.893	757038.9669	1300268.987	811705.8264	4124899.673
2	694372.7888	888644.7694	1038079.061	996558.6769	3617655.296
3	561986.6186	818163.3909	645744.6105	384890.3636	2410784.984
4	101396.3518	68082.11235	72255.11986	104532.8237	346266.4078
5	20936.92045	8965.79575	60692.64925	67553.269	158148.6345
Grand Total	**4343668.583**	**4035711.997**	**5363689.763**	**4012754.648**	**17755824.99**

When compared to all regions the charges spent higher to having 0 children's other than having children.

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

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**

Contents



The sum of charges for male is high in southeast region and low in in northwest region.

For female southeast has high charges and low in southwest region.

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

I have histograms and box plots. from this, the histograms show the count of persons in range of particular interval. And box plots show no. of outliers present in those variables. For BMI vs charges there is increase and decrease in pattern.

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

- As per my observation **Charger's** (column) is dependent value when we compared to order columns.
- **Age, Sex, BMI, Children's**, etc all columns are independent.
- **Non-smokers** are more than smokers.
- In smoker's category **Males** are higher when compared to **Females.**
- We can't any Strong correlation between the **Sex, Age & BMI**.
- When compared to all regions the charges spent higher to having 0 children's other than having children.
- Coming to region wise smokers & non- smokers
- Southeast has the highest no. of smokers
- Northwest has the lowest no. of smokers

Contents

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 have replaced male with 1 and female with 0. By using 'if' formula in excel.

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

We can replace by using **IF** condition.

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.

using 'if' formula, we can create the north & south west and southeast for three columns.

Contents

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.

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

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11938.53858	987.8191752	-12.0858	5.58E-32	-13876.39342	-10000.68373	-13876.39342	-10000.68373
age	256.8563525	11.89884907	21.58666	7.78E-89	233.5137784	280.1989267	233.5137784	280.1989267
sex	-131.3143594	332.9454391	-0.3944	0.693348	-784.4702705	521.8415517	-784.4702705	521.8415517
bmi	339.1934536	28.59947048	11.86013	6.5E-31	283.0884256	395.2984816	283.0884256	395.2984816
children	475.5005451	137.8040925	3.450555	0.000577	205.1632856	745.8378047	205.1632856	745.8378047
smoker	23848.53454	413.1533548	57.7232	0	23038.03071	24659.03838	23038.03071	24659.03838
northwest	-352.9638994	476.2757859	-0.74109	0.458769	-1287.298203	581.3704037	-1287.298203	581.3704037
southeast	-1035.022049	478.6922095	-2.16219	0.030782	-1974.096773	-95.9473258	-1974.096773	-95.9473258
southwest	-960.0509913	477.9330243	-2.00876	0.044765	-1897.636383	-22.46559965	-1897.636383	-22.46559965

According to the regression statistics, all of the independent variables listed in the table account for 75% of the variation in the charges.

This model's intercept is -119938.53.

- AGE: The expected charge increase is 256.85 units for every unit increase in AGE.
- Sex: The expected decrease in charges is -131.31 units for every unit increase in sex.
- BMI: The charges are anticipated to rise by 339.19 units for every unit increase in BMI. Children: The average price is predicted to rise by 475.5 units for every additional unit of children.
- Smoker: The charges are anticipated to rise by 23848.53 units for every additional smoker unit.
- Northwest: The charges are anticipated to increase for each additional unit in the Northwest. the fees are anticipated to drop by -352.96 units.
- Southeast: The expected decrease in charges is -1035.02 units for every unit increase in Southeast.

- Southwest: It is anticipated that the charges will go down by -960.05 units for every unit increase in Southwest.

  Every single variable in the model must have a p value of less than 0.05 before we can declare it significant for the y variable (charges).

Age, BMI, children, smokers, southeast, and southwest are all significant factors in this case.

Contents

Regression Statistics	
Multiple R	0.866476426
R Square	0.750781397
Adjusted R Square	0.749657948
Standard Error	6059.146461
Observations	1338

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-12165.38244	949.5381396	-12.81189447	1.60589E-35	-14028.13689	-10302.62798	-14028.1	-10302.6
age	257.0063906	11.88925335	21.61669729	4.61511E-89	233.6826728	280.3301084	233.6827	280.3301
bmi	338.6413347	28.55407641	11.85964939	6.49974E-31	282.6254353	394.6572342	282.6254	394.6572
children	471.5441444	137.6559519	3.425526743	0.00063229	201.4978697	741.5904191	201.4979	741.5904
smoker	23843.87493	411.6590831	57.92141097	0	23036.30359	24651.44628	23036.3	24651.45
southeast	-858.4696418	415.205505	-2.067577697	0.038872641	-1672.99817	-43.94111379	-1673	-43.9411
southwest	-782.7452298	413.7559633	-1.891804105	0.05873399	-1594.430123	28.93966291	-1594.43	28.93966