Health Insurance Claim Project

By

D. Harish Reddy

DA-March-2023

INDEX

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. --- 17

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. ---- 17

"Finding out the health parameters that affect health insurance claims" An insurance company in the US is reviewing its insurance claims/charges and is trying to do a cause and effect analysis for future business decisions. It has collected data for its customers who have made claims till recent time. The data-points collected are age, gender, BMI, number of children/dependents, smoking habit, region they belong to, charges/bills claimed under the insurance. This analysis would have a bearing on what premium should the company charge a customer availing an insurance policy. The insurance company has collected a dataset of 1338 customers-claims. Please refer to the data dictionary below:

Data Dictionary:

Attribute	Description
Age	Age of the customer/claimant who has claimed insurance for medical treatment charges
Sex	Gender of the customer/claimant
BMI	Health parameter: person's weight in kilograms divided by the square of height in meters
Children	No. of children the claimant has
Smoker	Whether the claimant smokes or not
Region	Region to which the claimant belongs
Charges	The exact medical charges for which the claimant has claimed insurance

Objective (Task):

- **To do a Cause and effect analysis on historic-data of insurance claims.**

You have been appointed as the "Analyst" for this project to achieve the objective of the study, your tasks are as under.

1. Perform the Exploratory Data Analysis on the data.

a. Identify the categorical and continuous variables.

- Categorical Variables: Sex, Children, Smoke and Region.
- Continuous Variables: Age, BMI and Charges.

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

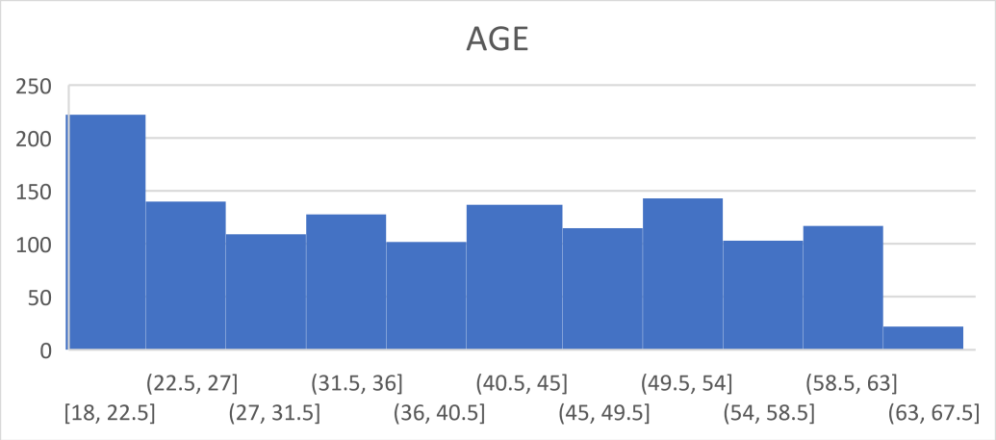

Table-1(b)

The distribution of the ages of the Health Insurance Claim is looks like a right skewed (positive). In these Graph it doesn't appear any outliers.

- The maximum number of people appearing to claim the insurance at the age is between 18 to 22.5.
- The Minimum Number of people appearing to claim the insurance at the age is between 63 to 67.5.

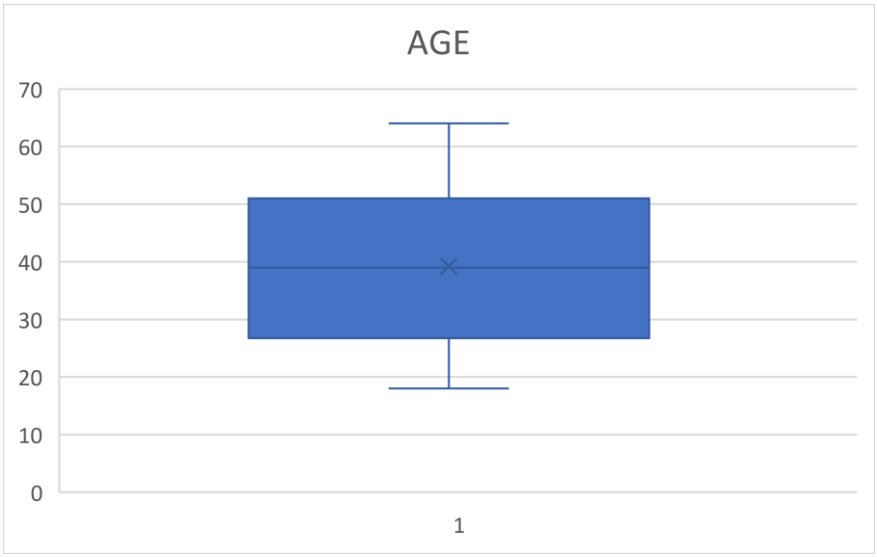

Table-1(b)

- The Box plot gives you a visualization Descriptive Statistics.
- The minimum age of the insurance claimant is 18, The maximum age is 64.
- The Quartile percentage Q1 (25%) is 26.75, here the median gives the Q2 (50%) is 39 the Q3(75%) is 51.
- The inter quartile range is Q3-Q1is 24.25. It doesn't have any outliers.

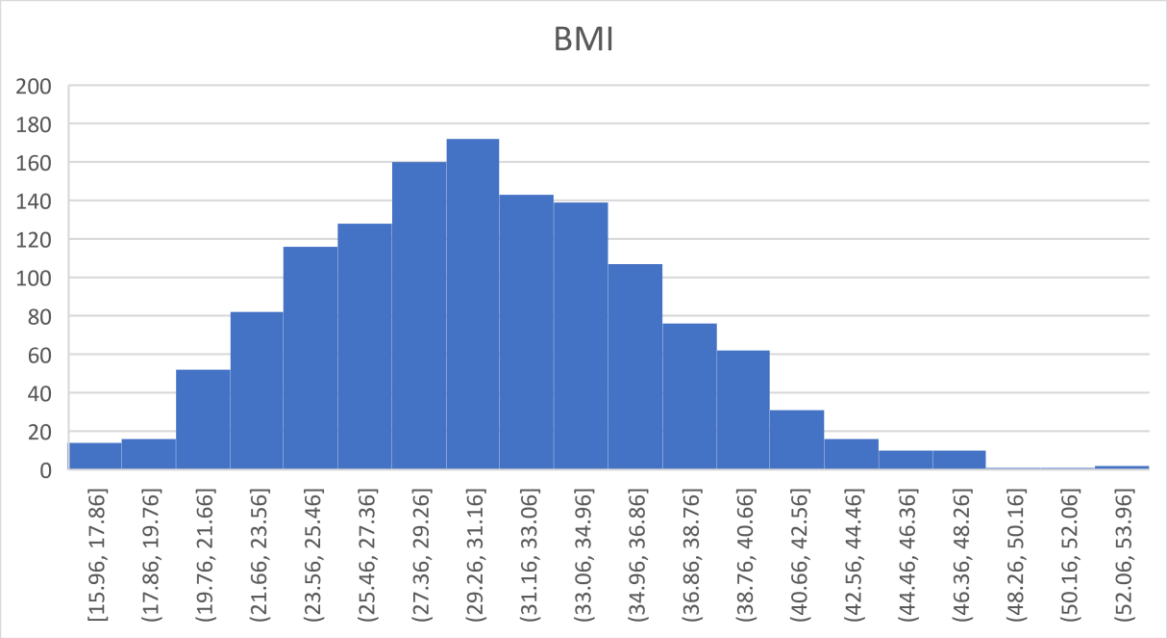

Table-1(b)

- The distribution of the BMI of the Health Insurance Claim the shape of graph is roughly symmetrical.
- In these Graph it has an Outliers. The maximum number of people appearing to claim the insurance at the age is between 18 to 22.5.
- The Minimum Number of people appearing to claim the insurance at the age is between 63 to 67.5

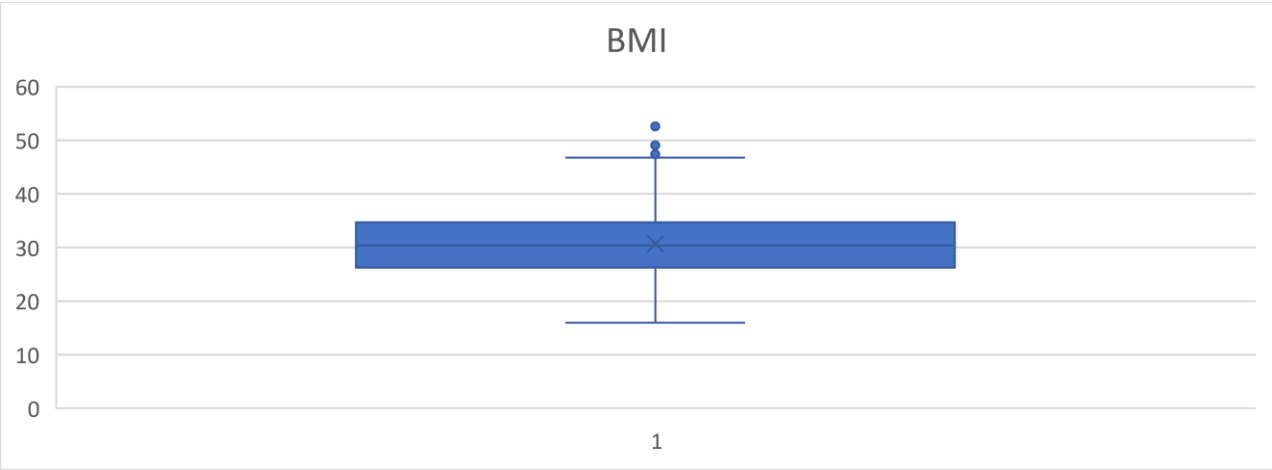

Table-1(b)

- The box plot gives you a visualization Descriptive Statistics.
- By observing the above Graph, it distributes the person's weight in kilograms divided by the square of height in meters of Health Insurance Claim.
- The Minimum weight and height of the insurance claimant is approximately 16, The Maximum weight and height is 46.75.

- The Quartile percentage Q1 (25%) is 26.27, here the median gives the Q2 (50%) is 30.4 and the Q3(75%) is 34.7.
- The inter quartile range is Q3-Q1is 8.43. It having a Four Outliers, The Range of these outliers are 47.41 to 52.58.

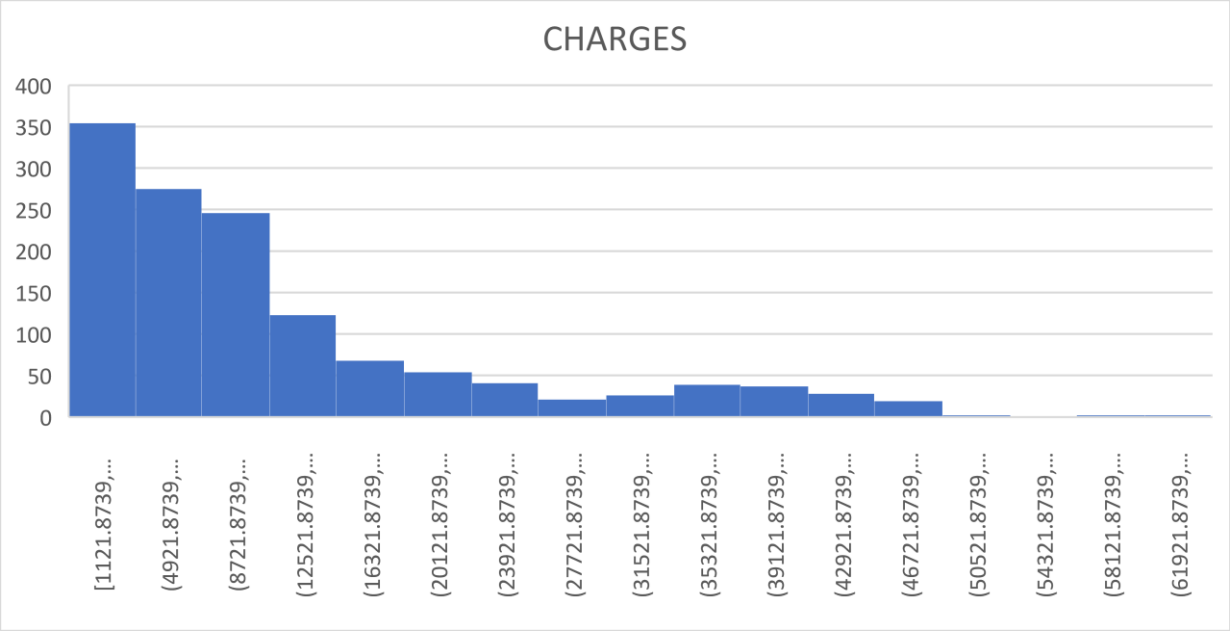

Table-1(b)

- The distribution of the charges of the Health Insurance Claim is looks like a right skewed (positive).
- In these Graph it having outliers, the maximum charges getting the customers from the health insurance is 34617.8.
- The Minimum charges getting the customers from the health insurance is 1121.8, the more customers are getting minimum charges.

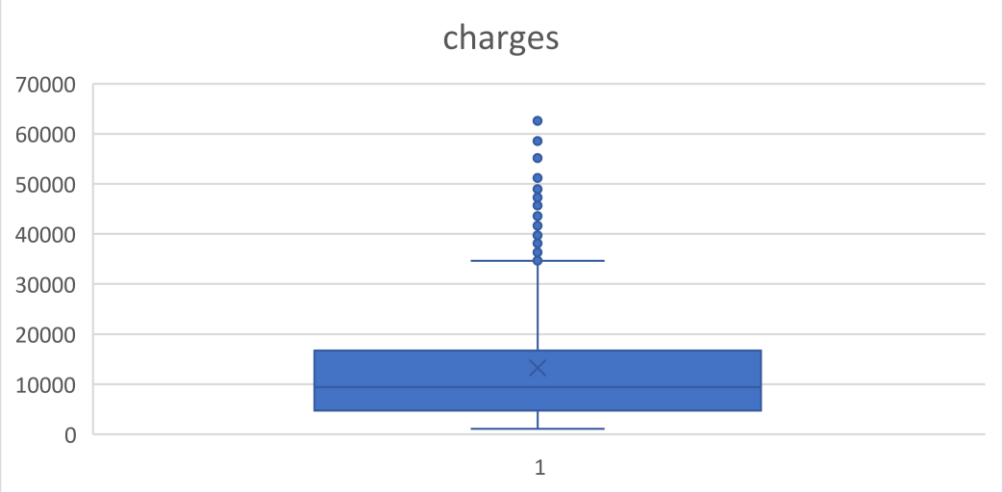

Table-1(b)

- The box plot gives you a visualization Descriptive Statistics.

- By Observing the above Graph, it distributes the Charges of the Health Insurance Claim.
- The minimum Charges of the insurance claimant is 1121.8.
- The maximum Charges is 34617.8, the Quartile percentage Q1 (25%) is 4733.6, here the median gives the Q2 (50%) is 9382.03 and the Q3(75%) is 16687.3.
- The inter quartile range is Q3-Q1is 7305.27. It having 17 outliers, the rage of these Outliers 34672.1 to 62592.8. Here the Outliers customers are getting more charges from the insurance.

Co Relation:

	AGE	*BMI*	*CHARGES ($)*
AGE	1		
BMI	0.109272	1	
CHARGES ($)	0.299008	0.198341	1

Table-1(b)

- The coefficient of correlation of Age is positively correlated with BMI and Charges.
- The BMI is positively correlated with the charges. The age is perfectly positive correlated with age, BMI is also perfectly positive correlated with BMI and the charges are perfectly positive correlated with charges.

c. Make relevant Pivot tables and charts for:

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

Count of smoker	**Column Labels**		
Row Labels	**no**	**yes**	**Grand Total**
female	82.63%	17.37%	100.00%
male	76.48%	23.52%	100.00%
Grand Total	**79.52%**	**20.48%**	**100.00%**

Table-C (1)

By observing, the Above Table the male smokers are more as compared to female smokers.

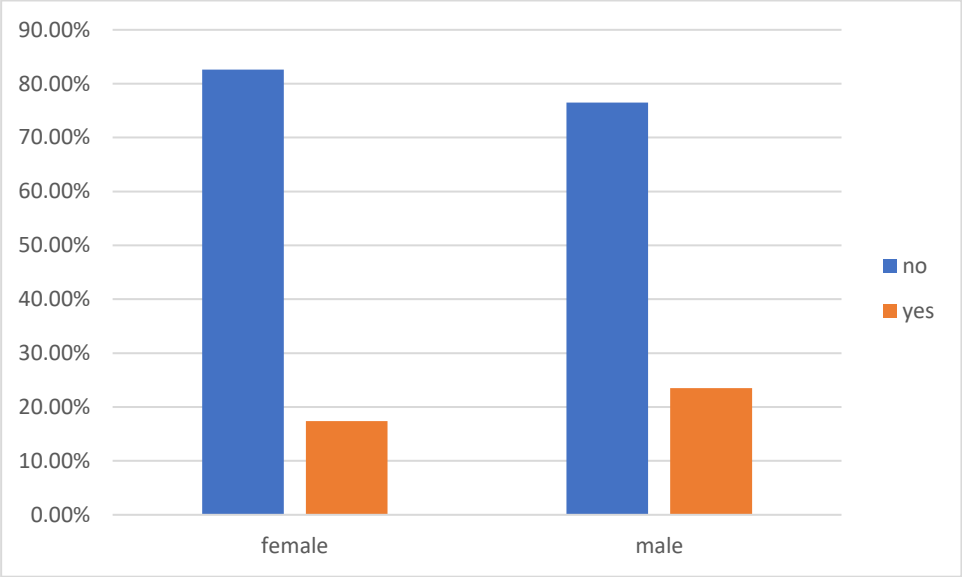

Table-C (1)

2.Charges Vs Age

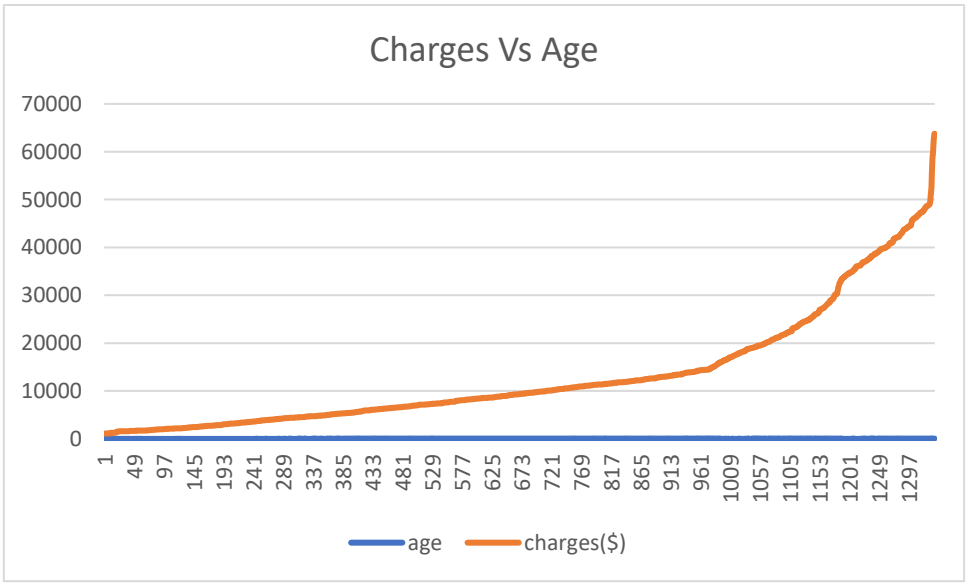

Table-C (2)

- From the graph, increasing the age the insurance charges also Gets increased. Here the ages and charges are having positive correlation.

3.Charges vs BMI

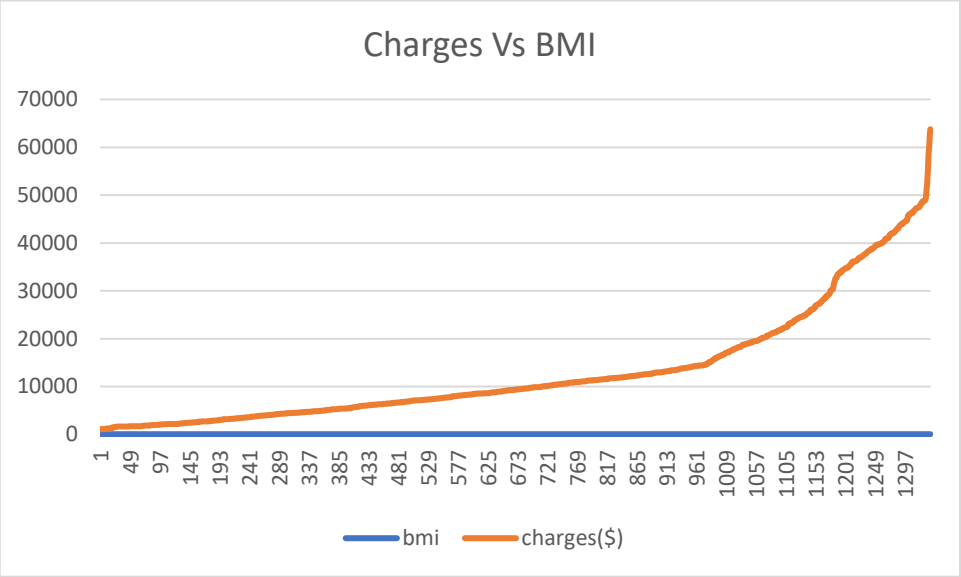

Table-C (3)

- There is low relation between the Charges and BMI.

4. Charges for Smokers vs Non-smokers.

Row Labels	Count of smoker
no	1064
yes	274
Grand Total	**1338**

Table-C (4)

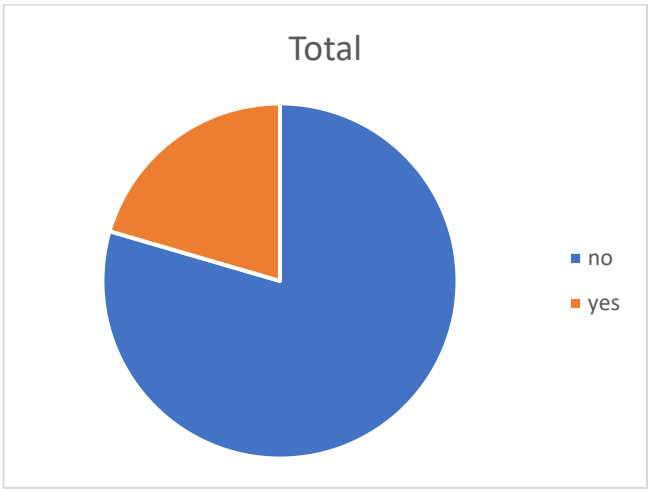

Table-C (4)

- The Non-Smokers are More than Four times of the Smokers For the Given Data.

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

Count of smoker	Column Labels	
Row Labels	**yes**	**Grand Total**
northeast	67	67
northwest	58	58
southeast	91	91
southwest	58	58
Grand Total	**274**	**274**

Table-D (1)

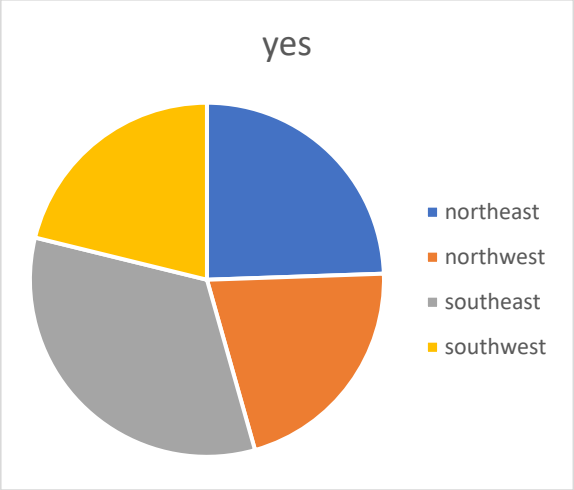

Table D (1)

Count of smoker	Column Labels	
Row Labels	**no**	**Grand Total**
northeast	257	257
northwest	267	267
southeast	273	273
southwest	267	267
Grand Total	**1064**	**1064**

Table-D (2)

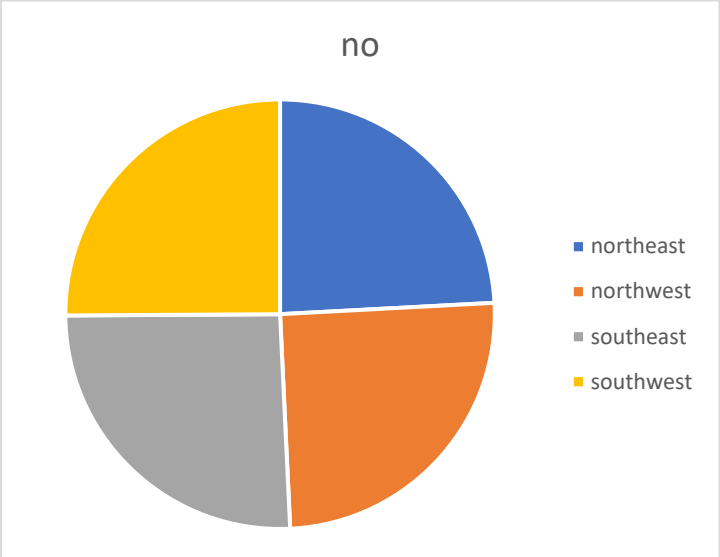

Table-D (2)

- By observing the above 2 graphs, Non-Smokers in each graph is approximately similar in all the regions.
- The smokers in southeast region is more as compared to remaining regions and the southwest and northwest regions have the same range of smokers.

e. Region-wise charges for Smokers 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**

Table-E (1)

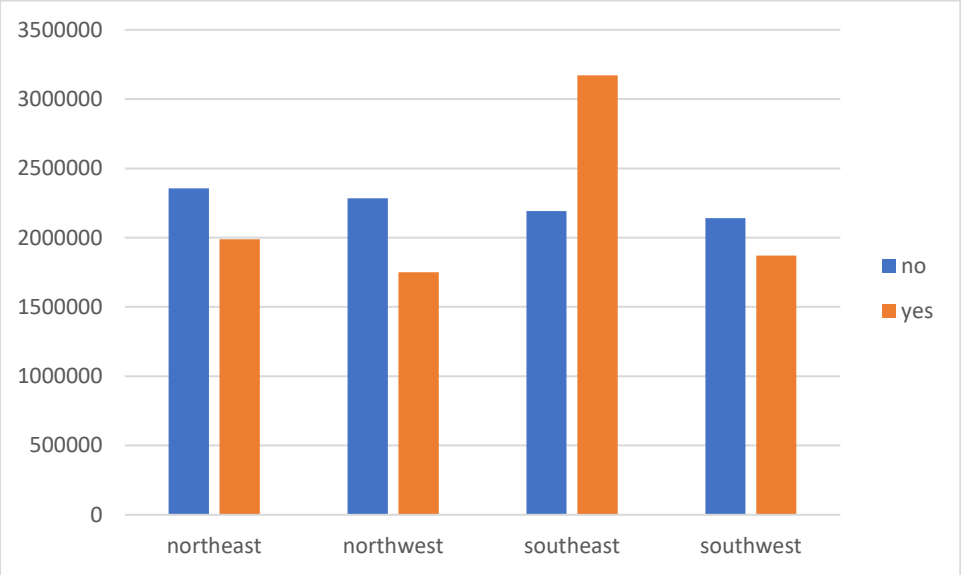

Table-E (2)

- From the table the Non-Smokers Average charges in each region is approximately similar.
- The Average Smokers charges in southeast region is more. The remaining regions average charges are approximately similar.
- The overall charges of Smokers is 4 times greater than non-smokers charges.

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

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**

Table-F (1)

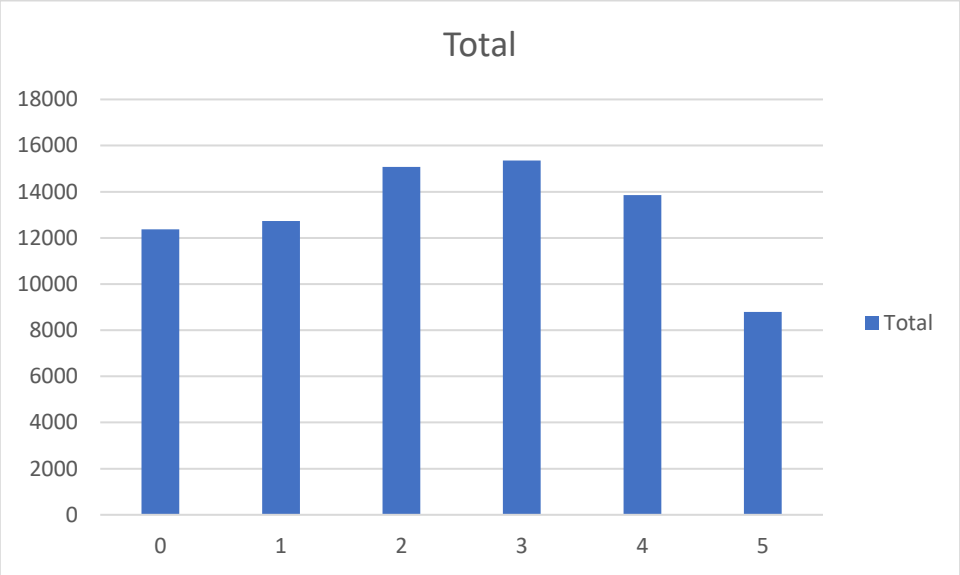

Table-f (2)

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

Count of region	Column Labels				
Row Labels	northeast	northwest	southeast	southwest	Grand Total
0	147	132	157	138	574
1	77	74	95	78	324
2	51	66	66	57	240
3	39	46	35	37	157
4	7	6	5	7	25
5	3	1	6	8	18
Grand Total	**324**	**325**	**364**	**325**	**1338**

Table-g (1)

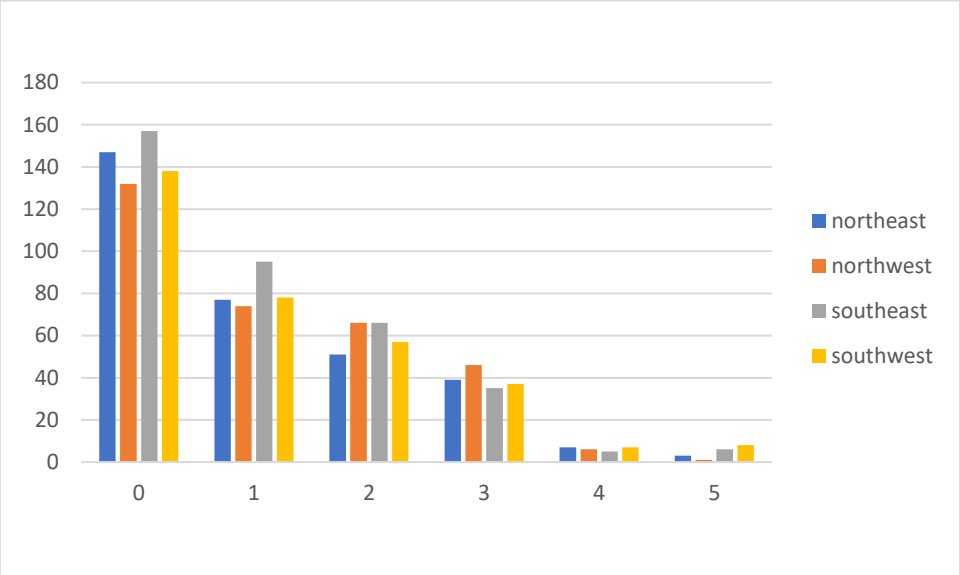

Table- (2)

- By observing the graph, in all the region who didn't have children they have high insurance charges. Who contain maximum number of children they have less insurance charges. who are have 2 to 3 children they are getting medium charges.
- The southeast region getting always maximum charges compared to other regions.

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

Row Labels	Sum of age	Sum of bmi	Sum of children
female	26151	20110.07	711
male	26308	20917.555	754
Grand Total	**52459**	**41027.625**	**1465**

Table-h (1)

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

From the pattern b:

- The distribution of the ages of the Health Insurance Claim is looks like a right skewed (positive). In these Graph it doesn't appear any outliers.
- The maximum number of people appearing to claim the insurance at the age is between 18 to 22.5.
- Minimum Number of people appearing to claim the insurance at the age is between 63 to 67.5.

- The inter quartile range is Q3-Q1is 24.25. It doesn't have any outliers. The distribution of the BMI of the Health Insurance Claim the shape of graph is roughly symmetrical.
- In these Graph it has an Outliers. The maximum number of people appearing to claim the insurance at the age is between 18 to 22.5.
- Minimum Number of people appearing to claim the insurance at the age is between 63 to 67.5. The inter quartile range is Q3-Q1is 8.43. It having a Four Outliers, the range of these Outliers are 47.41 to 52.58.
- The distribution of the charges of the Health Insurance Claim it looks like a right skewed (positive).in these Graph it having outliers.
- The maximum charges getting the customers from the health insurance is 34617.8. Minimum charges getting the customers from the health insurance is 1121.8, the more customers are getting minimum charges.
- The inter quartile range is Q3-Q1is 7305.27. It having 17 outliers, the range of these outliers are 34672.1 to 62592.8. Here the Outliers customers are getting more charges from the insurance.

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

From point c:

- The male smokers are more as compared to female smokers. Increasing the age the insurance charges also gets increased.
- Here the ages and charges are having positive correlation. From the chart we have a very low relation between the charges and BMI. Smokers charges are approximately 4 times greater than the non-smoker charges.

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

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.

- After replacing all the "Males" with "1" and "Females" with "0".

sex
0
1
1
1
1
0
0
0
1
0
1

0
1

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

smoker
1
0
0
0
0
0
0
0
0
0

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.

northwest	southeast	southwest
0	0	1
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	0

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.

Descriptive summary analysis for the edited data:

sex		*smoker*		*northwest*		*southeast*		*southwest*	
Mean	0.505232	Mean	0.204783	Mean	0.2429	Mean	0.272048	Mean	0.2429
Standard E	0.013674	Standard E	0.011036	Standard E	0.011728	Standard E	0.01217	Standard E	0.011728
Median	1	Median	0	Median	0	Median	0	Median	0
Mode	1	Mode	0	Mode	0	Mode	0	Mode	0
Standard D	0.50016	Standard D	0.403694	Standard D	0.428995	Standard D	0.445181	Standard D	0.428995
Sample Va	0.25016	Sample Va	0.162969	Sample Va	0.184037	Sample Va	0.198186	Sample Va	0.184037
Kurtosis	-2.00256	Kurtosis	0.145756	Kurtosis	-0.55986	Kurtosis	-0.94952	Kurtosis	-0.55986
Skewness	-0.02095	Skewness	1.464766	Skewness	1.200409	Skewness	1.025621	Skewness	1.200409
Range	1	Range	1	Range	1	Range	1	Range	1
Minimum	0	Minimum	0	Minimum	0	Minimum	0	Minimum	0
Maximum	1	Maximum	1	Maximum	1	Maximum	1	Maximum	1
Sum	676	Sum	274	Sum	325	Sum	364	Sum	325
Count	1338	Count	1338	Count	1338	Count	1338	Count	1338

Table-3 (1)

- From the above descriptive summery analysis, the count of the all variables are 1338.
- The mean values of the smokers, northwest, southwest and southeast are approximately same.
- The mean of the sex ratio is 50%. Here except smoker variable remaining all the variables are negatively kurtosis. Here only sex variable has negative skewness remaining all variables have positive skewness.
- The standard deviation except sex variable remaining all the variables are far away from the mean, sex variable lies on the mean.

Multiple Linear Regression analysis:

Regression Statistics	
Multiple R	0.866552
R Square	0.750913
Adjusted R	0.749414
Standard E	6062.102
Observatic	1338

ANOVA

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

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 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

Table-3 (2)

- By observing the linear regression analysis, Age, BMI, Children, smoker, southeast and southwest are significant to charges, sex and northwest variables insignificant to charges. In this we have strong multiple R value.
- The coefficient values of sex, northwest, southwest and southeast are having negative coefficient here the dependent variable gets decreased.
- The coefficient values of Age, BMI, Children and smoker are having positive coefficient, here the dependent variable gets increased. Here few of the variables have negative confidence it indicates the low-level confidence of independent variables. It effects the claiming of insurance charges.

Another set of regression analysis by dropping insignificant variables:

Regression Statistics	
Multiple R	0.866476
R Square	0.750781
Adjusted R	0.749658
Standard E	6059.146
Observatio	1338

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressio	6	1.47E+11	2.45E+10	668.2821	0
Residual	1331	4.89E+10	36713256		
Total	1337	1.96E+11			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-12165.4	949.5381	-12.8119	1.61E-35	-14028.1	-10302.6	-14028.1	-10302.6
age	257.0064	11.88925	21.6167	4.62E-89	233.6827	280.3301	233.6827	280.3301
bmi	338.6413	28.55408	11.85965	6.5E-31	282.6254	394.6572	282.6254	394.6572
children	471.5441	137.656	3.425527	0.000632	201.4979	741.5904	201.4979	741.5904
smoker	23843.87	411.6591	57.92141	0	23036.3	24651.45	23036.3	24651.45
southeast	-858.47	415.2055	-2.06758	0.038873	-1673	-43.9411	-1673	-43.9411
southwest	-782.745	413.756	-1.8918	0.058734	-1594.43	28.93966	-1594.43	28.93966

Table-3 (3)

- For the better analysis by dropping insignificant variables and took a regression model on significant variables.
- In this except southeast and southwest remaining all variables have positive coefficient values, these positive coefficient values helps to increase the value of dependent variable.
- The above and present regression model we have negative intercept value. There is no change in confidence level. As above regression model is better as compared to present regression model.

Thank you.