Health Insurance Claim: Business Report

Report Done With STAR Framework

Situation:

Finding out the health parameters that effect health insurance claims.

Task:

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

Action:

1) Exploratory data analysis on the data.

a) 'BMI' and 'Charges' are continuous variable and 'Sex', 'Smoke', 'Region', are Categorical variable and 'Age', 'Children' are discrete variable.

b)

i)Histograms and box plots of continuous variables:-

BMI

This chart isn't available in your version of Excel.

Editing this shape or saving this workbook in a different file format will permanently break the chart.

This chart isn't available in your version of Excel.

Editing this shape or saving this workbook in a different file format will permanently break the chart.

Charges

This chart isn't available in your version of Excel.

Editing this shape or saving this workbook in a different file format will permanently break the chart.

This chart isn't available in your version of Excel.

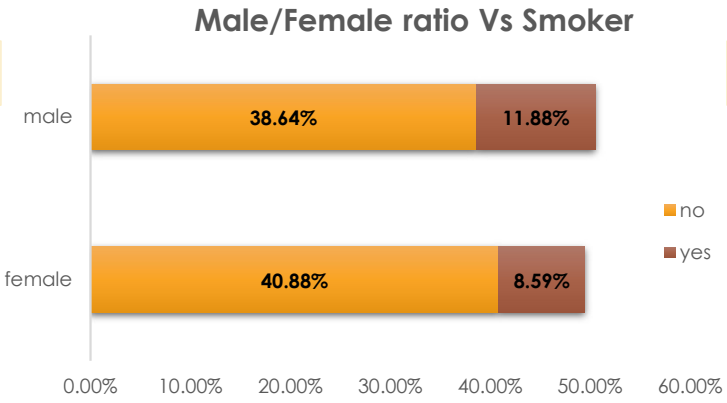
Editing this shape or saving this workbook in a different file format will permanently break the chart.

ii)Correlation analysis (multivariate):-

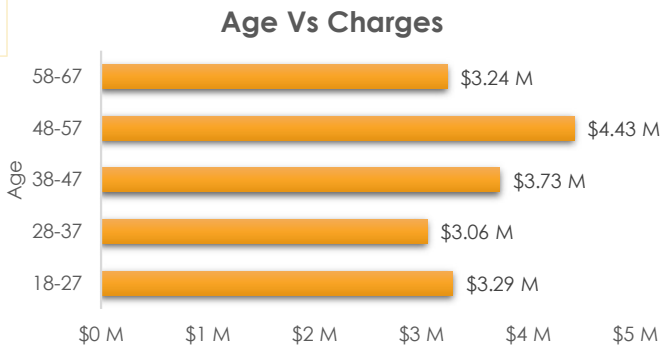
- Positive Correlation with Charges:
- Highest positive Correlation is of 'Smoking' and 'Charges'(0.78725143).
- 2nd highest positive Correlation is of 'Age' and 'Charges'(0.299008).
- 3rd highest positive Correlation is of 'bmi' and 'Charges' (0.198341).
- Negative Correlation with Charges:
- Highest Negative Correlation is of 'Southwest' and 'Charges'(-0.04321003).
- 2nd highest Negative Correlation is of 'Northwest' and 'Charges'(-0.03990486).

c) Pivot Charts:

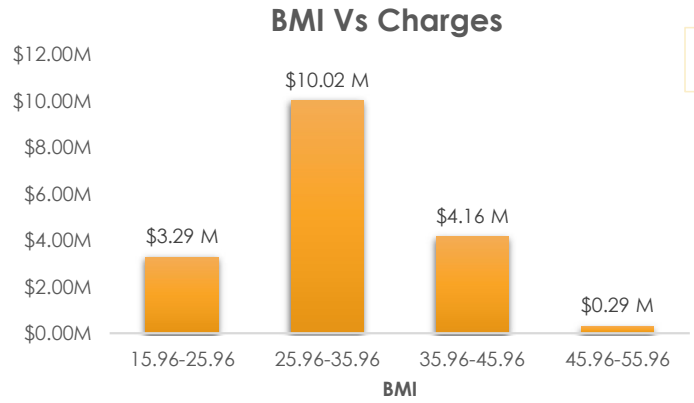
i)



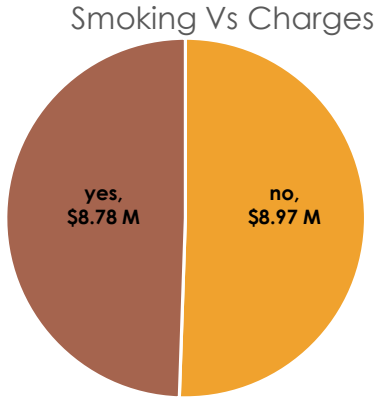
ii)



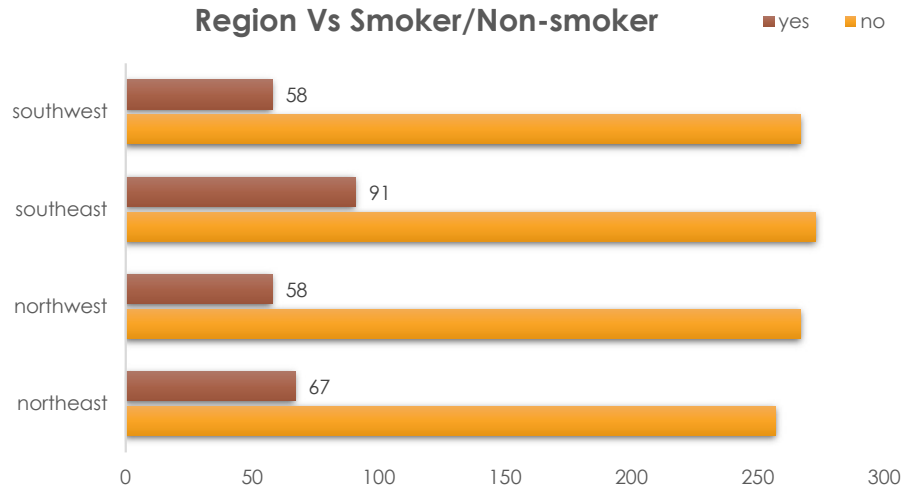
iii)



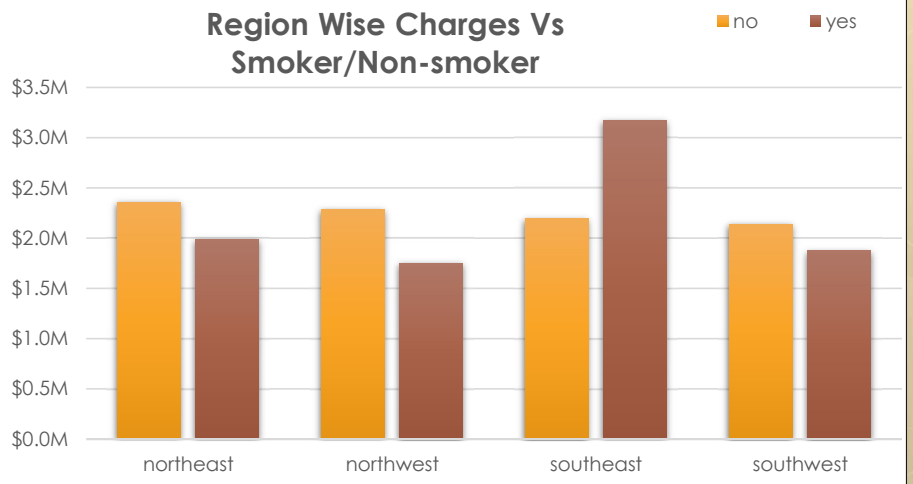
iv)



d) Region-wise smokers vs Non-smokers Pivot Chart:-



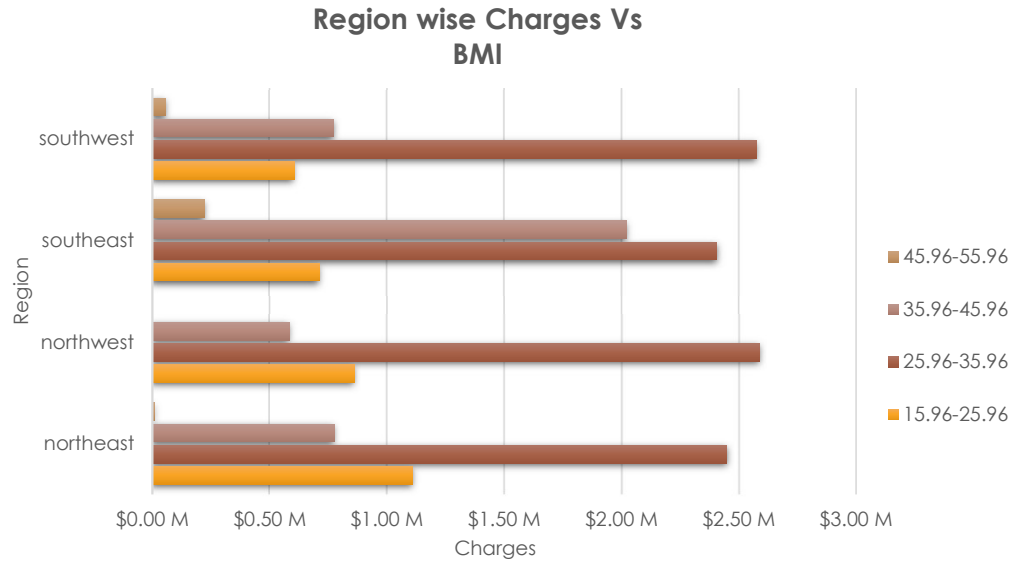
e) Region-wise charges for smokers vs non-smokers:-



f) Charges can be balanced on focusing on Smokers and people from south east as charges are high in southeast and people who smoke. Even if the smokers are less but their charges are high than Non-Smokers.

g) A similar dependants-charges analysis, Region-wise:-

- Region wise Charges Vs BMI:-

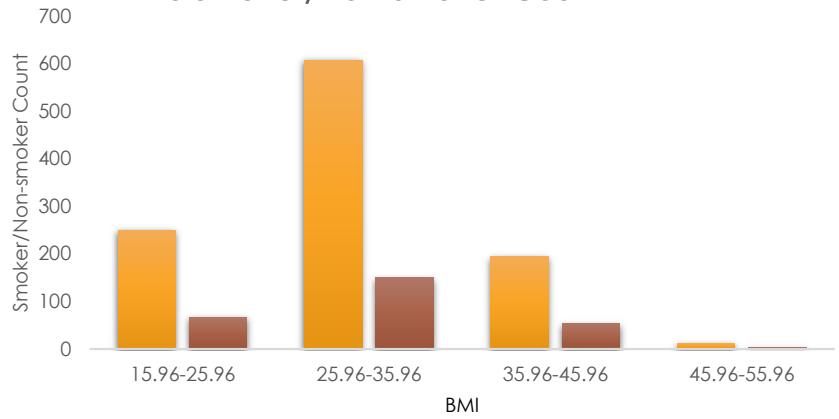


h) Additional Pivot Charts By Own Choice:-

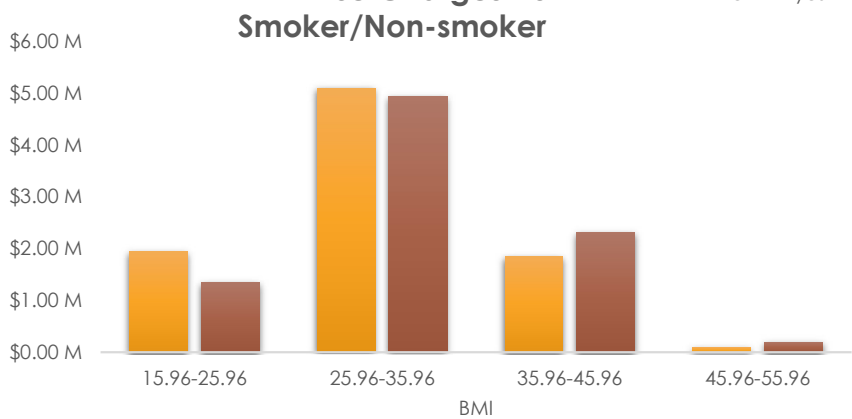
i) BMI Vs Smoker/Non-Smoker count

ii) BMI Wise Charges Vs Smoker/Non-Smoker

i)



ii)



i) Understanding from point (b):-

- BMI Histogram and Box plot:-
- Most (682) of the population falls near middle BMI Value (Median).
- Highest Frequency of BMI is from range 27-29.
- In BMI above 46,75 , all are outliers, Median is 30.4, Q1 is 26.27 and Q3 is 34.7 .
- Charges Histogram and Box plot:-
- Charges is following positively right skewed distribution that mean most of the People are more in lower end Charges and also the number of high charges is less (but more than a distribution of 'normal distribution').
- In Charges above 34672.4 all are outliers, Median is 9382.033, Q1 is 4733 and Q3 is 166687.

j) Interpretation for observations made in point (c):-

i. The number of Smokers in Male is more than Female.
ii. Highest Charges in range 48-57 Age.
iii. Highest Charges in range 25.96-35.96 BMI.
iv. The Charges of Non-smokers and Smokers are Close. But Smoker's numbers are but Charges are high.

2) All operations done in Excel Workbook.

3) Descriptive summary analysis for the edited data and all other significant Variables. A Multiple Linear Regression analysis to identify which variables decide the insurance charges/billed insurance claim:-

i. Descriptive Summary:- done in Excel worksheet.

ii. Regression analysis:-

- P-Value;
- Northwest:- 0.4587689, Southeast:- 0.0307817, Southwest:- 0.0447649, Age:- 7.783E-89, Smoker:- 0, Sex:- 0.6933475, Children:- 0.000577, BMI:- 6.498E-31
- Significant variables for Charges:- Southeast, Age, Smoker, Children, BMI. Even if the P-value of Southwest is less than 0.05 we eliminated it because it was not impacting that much (impacting very less insignificant) for Charges.
- Best Regression model formula:- Charges = [(-578)*(Southeast)] + [(257.1365)*(Age)] + [(23853.97)*(Smoker)] + [(468.0668)*(Children)] + [(333.4448)*(BMI)] + (-12275.6)

Ex:- (((-578)*(1))+((257.1365)*(28))+((23853.97*(1))+((468.0668)*(2))+((333.4448)*(33.77))+(-12275.6))) = 30396.76

Result:-

1) 'Smoker' and 'Charges' have a strong positive relationship, when Smokers are increasing Charges are also increasing that means Smoking is one of the biggest cause that causing Charges increase.
2) Range between 18-27 'Age' People are most Smokers and those Charges are highest among all Smokers in 'Age' group.
3) Smokers from range 25.96-35.96 are less in numbers as Non-smokers are 608 and Smoker are only 150. But Charges are so close to each other (Non-smoker's Charges = $50m and Smoker's Charges = $49m) and in range 35.96-45.96 BMI, Smokers (53) are less than Non-Smokers(194) but Charges are high of Smokers.(Smokers Charges = $2.3m and Non-smokers = $1.8m). So again we can say that main reason of Charges increase or Insurance claim are Smokers.
4) The Charges of Southeast is highest among all and Smoker's numbers and also Smoker's Charges are also highest in Southeast. Hence; we can say that the cause of Charges or Insurance claim in Southeast of US is because of Smokers.

The End

Thank you for reading my presentation.