**Business Situation:**

An insurance company is reviewing its insurance claims and charges and is seeking to conduct a cause and effect analysis for future business decisions. They have collected data for customers who have made claims, including various health parameters and insurance-related information.

**Task:**

The task is to analyze the collected dataset, consisting of 1,338 customer claims, and identify the health parameters that have an impact on health insurance claims. This analysis will help the insurance company determine the appropriate premium to charge customers who avail insurance policies.

**Action:**

**Data:** sex, smoker, region, age, bmi, children, charges($), Cate\_smoker, Cate\_sex, Northwest, Southeast, Southwest

**Descriptive Analysis:**

**Categorical Variables:**

1. Sex

2. Smoker

3. Region

**Continuous variables:**

1. Age

2. BMi

3. Children

4. Chargers($)

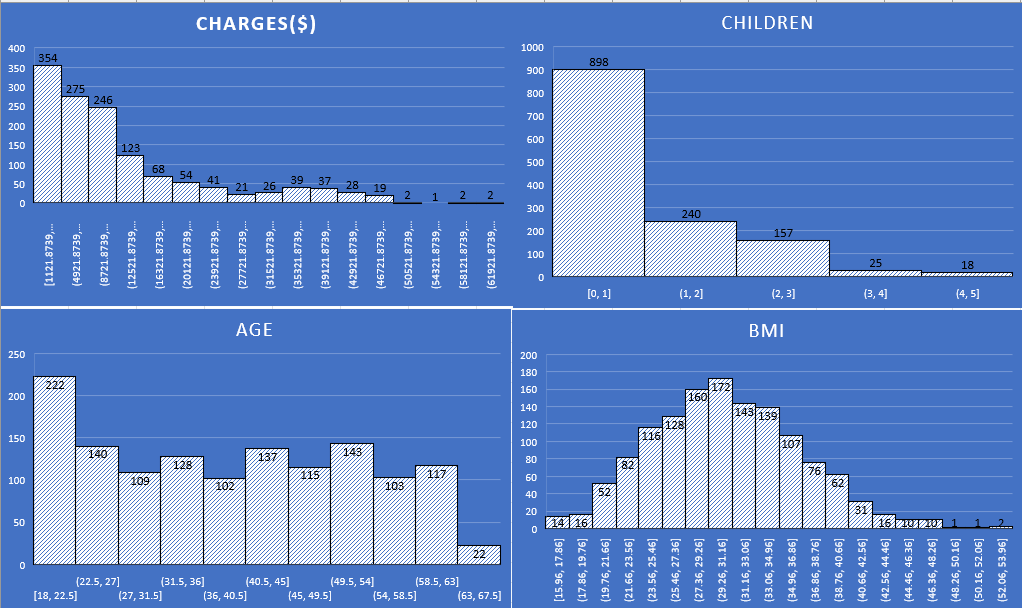
**Below are the histogram charts of charges($), Children, Age, BMI:**

The charges($) has a positively skewed where most of the charges lie in first few bars and leading to low count of charges with having high amount in the last.

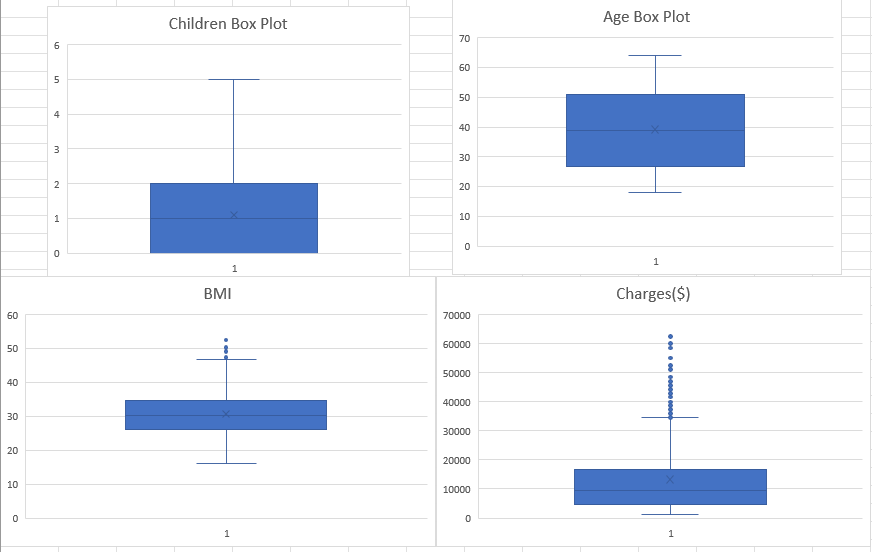
In children histogram we can that it is also the positively skewed where the age of 0-1 lie more than compare to other ages and the least is lie between 4-5

The BMI is in normal distribution and most of the data lie in and near mean of BMI

In age in also some what positive skewed where the most of the data lie between 18-22.5



**Below are the Box plots charts of** **charges($), Children, Age, BMI:**



**Top 3 Positive Correlation:**

1. Charges($) vs Age: 0.299008193
2. Charges($) vs BMI: 0.198340969
3. BMI vs Age: 0.109271882

**Information on which gender has more smokers:**

The sum of Cat\_smoker for females is 115.

The sum of Cat\_smoker for males is 159.

The grand total of Cat\_smoker is 274.

So here we can say that the Male Population as more smokers than females

**Charges vs Age:**

Age 19 as the maximum sum of charges 662857.8345 applied and following with 52, 54, 43 and 64

Ages.

**Charges vs BMI:**

Based on the provided data, here are the top 5 BMI that appear most frequently with highest charges:

32.3: 13 occurrences

28.31: 9 occurrences

34.1: 8 occurrences

31.35: 8 occurrences

**Charges for Smokers vs Non-smokers:**

The sum of charges ($) for the category "no" is 8,974,061.469.

The sum of charges ($) for the category "yes" is 8,781,763.522.

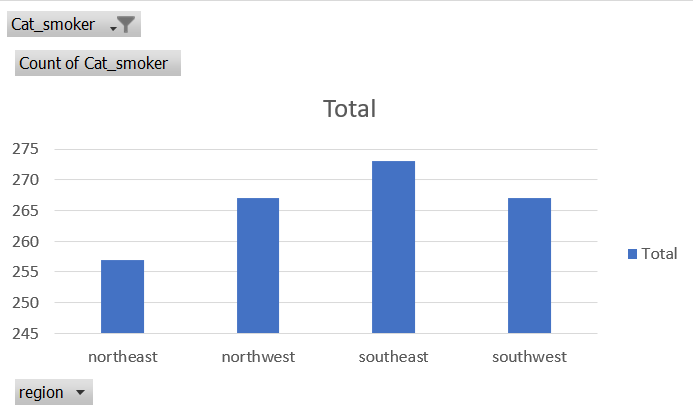
The grand total of charges ($) is 17,755,824.99.

Here we can see Non-smokers was claimed more insurance than Smokers considering charges.

**Region wise smokers:**

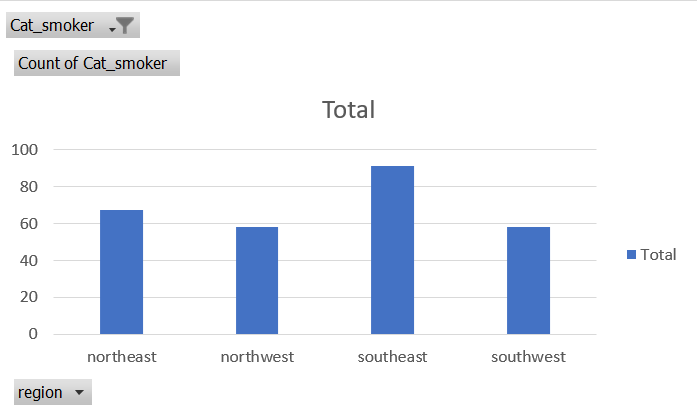
**The population having smoke**

In below chat we can see that the total amount of smokers is high in southeast region and then continuing with northeast

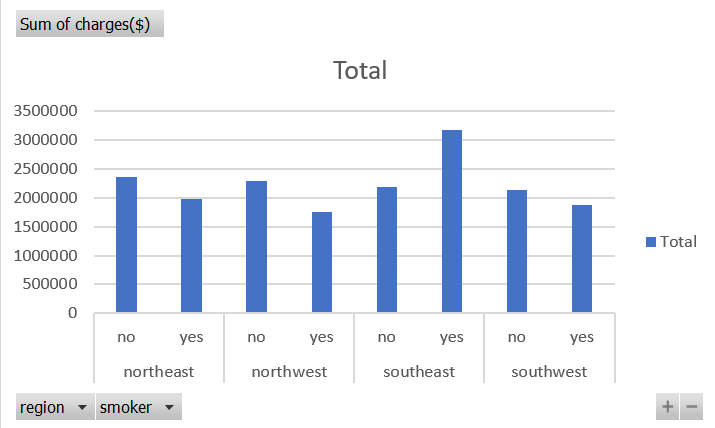


**The population not having smoke**

Here also we can see that it is the same data as above. The least smokers are in southeast region and next is northeast.



**Region-wise charges for smokers vs non-smokers:**

****

The total charges for smokers (yes) across all regions amount to $8,781,763.542, while the total charges for non-smokers (no) amount to $8,914,461.469. This indicates that, overall, smokers have slightly lower total charges compared to non-smokers.

Among the regions, the southeast region has the highest total charges, both for smokers ($3,170,894.711) and non-smokers ($2,192,795.052). On the other hand, the northeast region has the lowest total charges for smokers ($1,988,126.944) and non-smokers ($2,355,541.64).

In each region, non-smokers tend to have slightly higher total charges compared to smokers, except for the northwest region where smokers have higher total charges ($1,751,136.185) compared to non-smokers ($2,284,575.812).

Overall, it appears that smoking habit does not necessarily correlate with higher total charges. The variation in charges between smokers and non-smokers is not consistent across all regions. Other factors such as age, BMI, and number of children/dependents may also contribute to the differences in total charges. Further analysis and exploration of the data would be required to gain a more comprehensive understanding of the health parameters that affect health insurance claims.

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

Based on the correlation coefficient of 0.067998227 and an R Square value of 0.004623759, it can be concluded that the relationship between the number of dependents and charges is weak. The low R Square value indicates that the model does not provide a good fit. Additionally, the p-value of 0.012852129, which is less than the significance level of 0.05, suggests that the null hypothesis can be rejected in favor of the alternative hypothesis. Therefore, it can be inferred that the charges are not purely dependent on the number of dependents