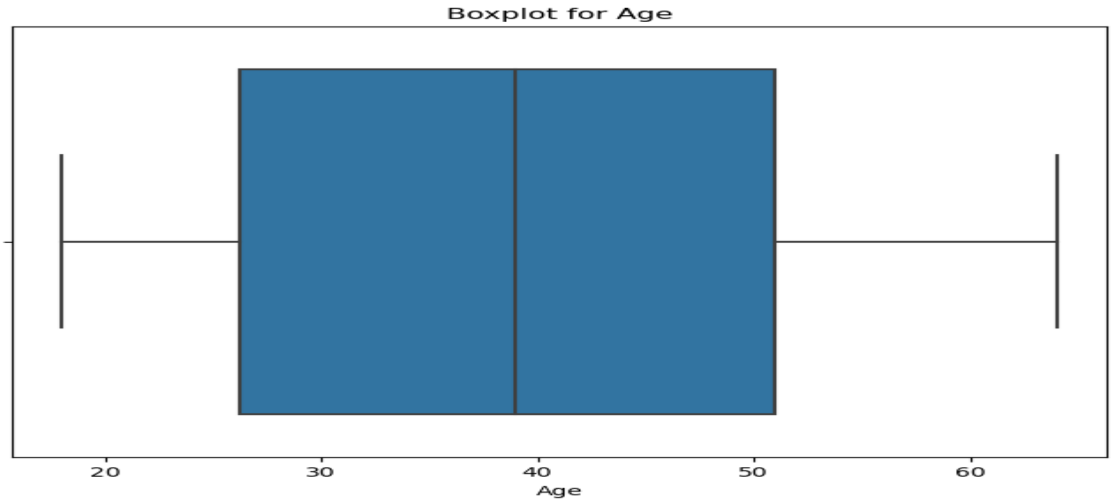
FINAL - ASSIGNMENT

Insurance Dataset Analysis:

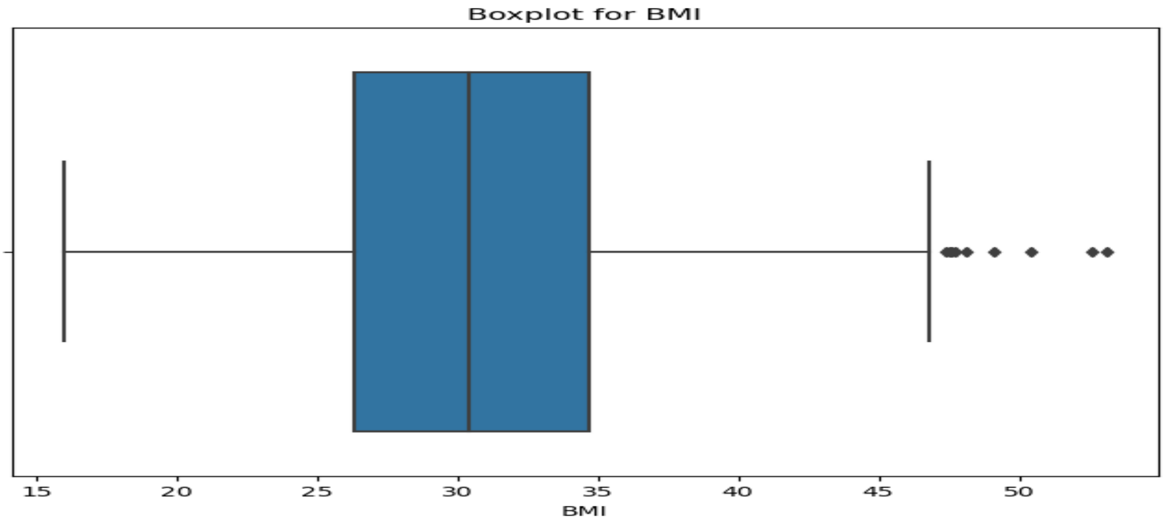
Outlier Detection and Removal:

The central box represents the middle 50% of the data, with the median (middle value) marked as a horizontal line. The whiskers extend to the 25th and 75th percentiles, and any points outside the whiskers are considered outliers.

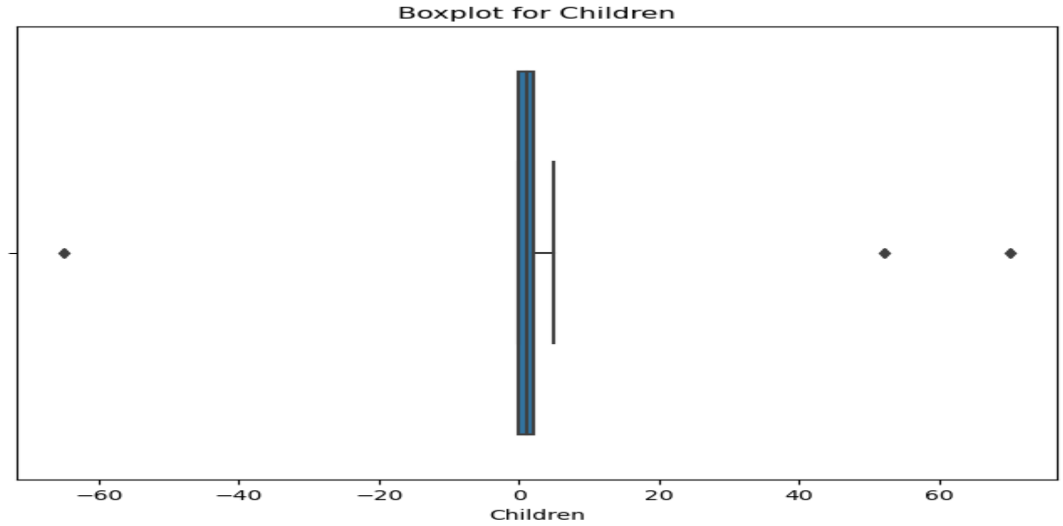
Below is the Boxplot for Age and we don’t see any outliers for age column in this insurance dataset. The median age is approximately 40 years.The interquartile range (IQR), which is the difference between the 75th and 25th percentiles, is approximately 10 years. This means that the middle 50% of the population is between 30 and 40 years old



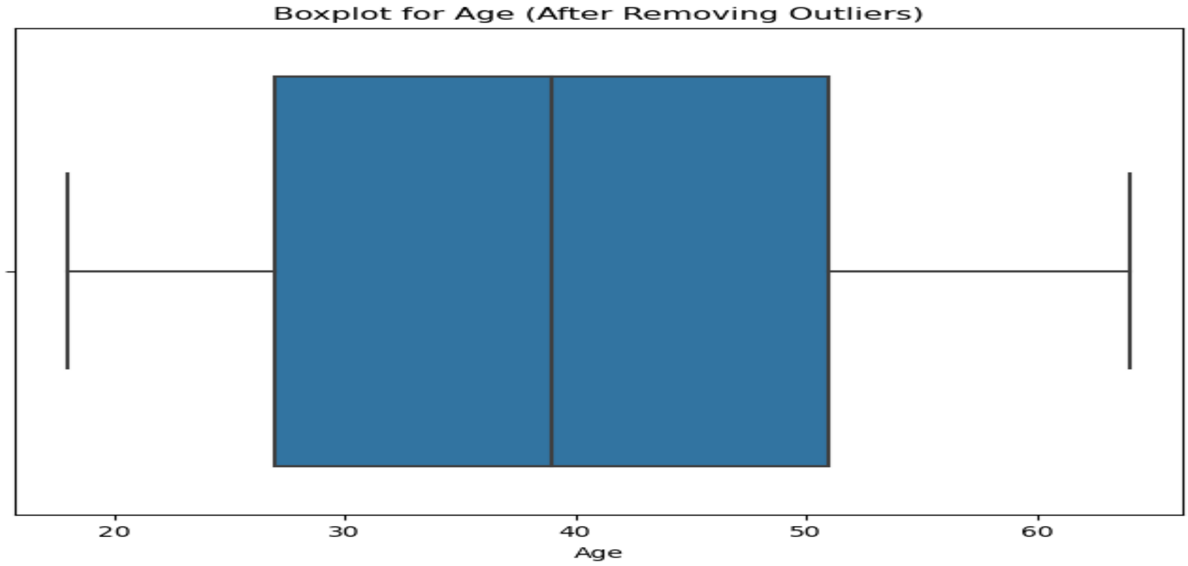
For the BMI box plot we can see the outliers and the median of the BMI value is approximately 32 and we can see the outliers after the value 48

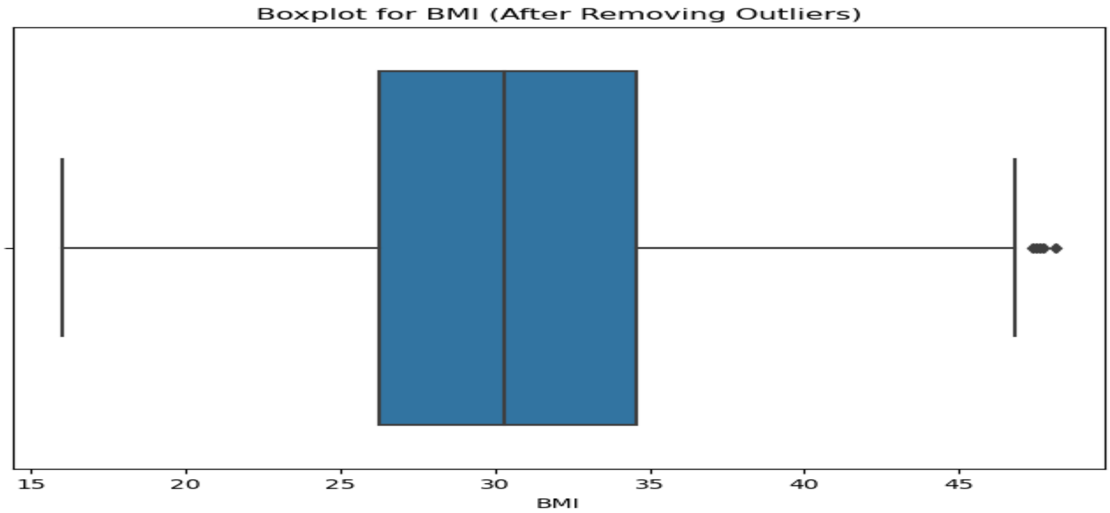


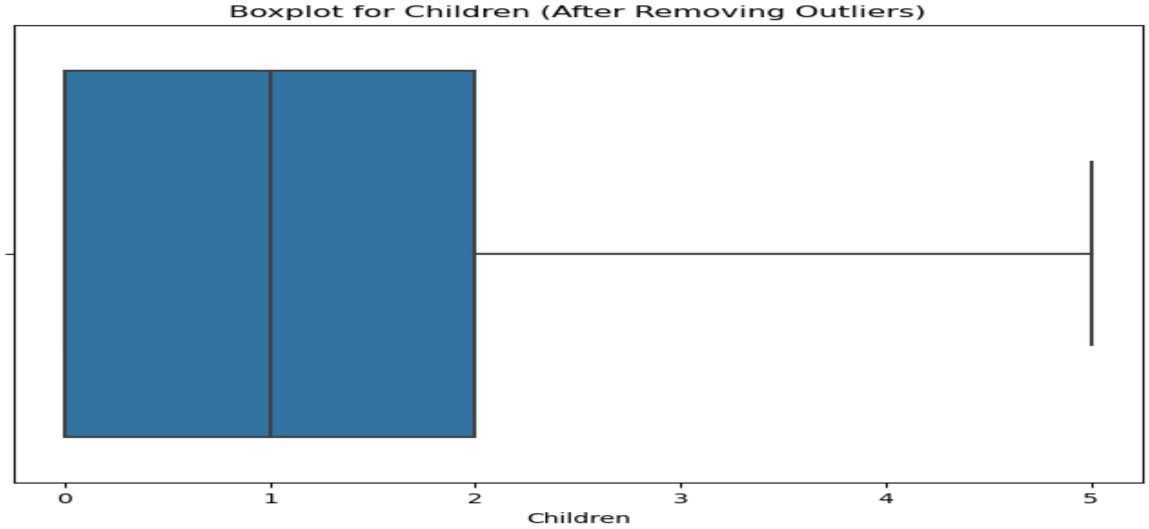
For the Boxplot for Children we can see the outliers and the median of the data seems to be nearer to 0



After the removal of the outliers the box plots of Age, BMI and Children are as below and we can still some outliers in BMI. Here we have removed the outliers which has exceeded the z_score of 3 i.e z_score>3







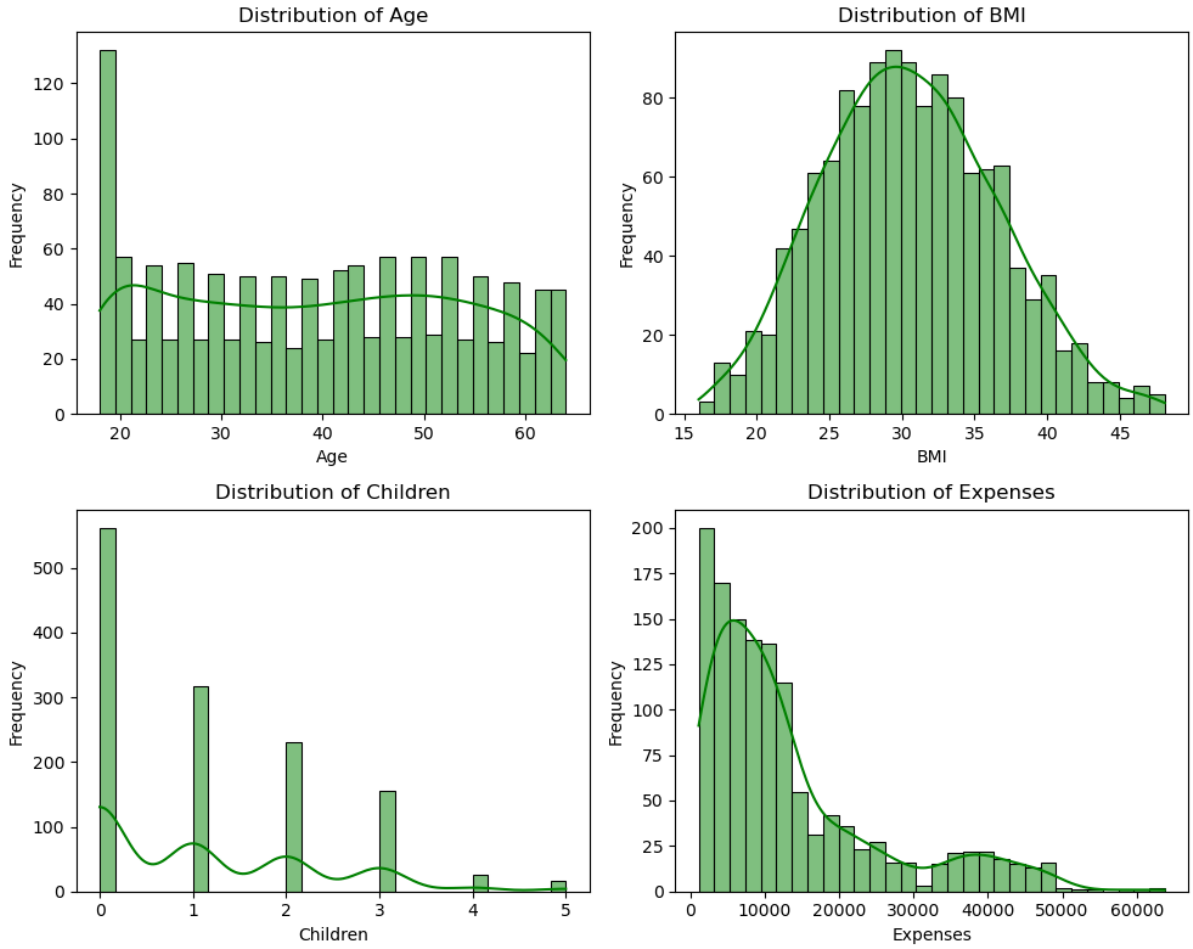
Data Distribution Visualization Explanation:

The distribution of age is skewed to the right, meaning that there are more people in the older age groups. The median age is approximately 50 years, and the interquartile range (IQR) is approximately 15 years. This means that the middle 50% of the population is between 35 and 50 years old. There are a few outliers, which are points that are more than 1.5 IQRs below or above the 25th and 75th percentiles, respectively. These outliers could be due to errors in data collection or could represent individuals who are significantly older or younger than the majority of the population.

The distribution of BMI is approximately normal, with a mean of 25.5 . This means that the average person in the population is overweight. There are a few outliers, which are points that are more than 2 standard deviations below or above the mean. These outliers could represent individuals who are underweight or obese.

The distribution of children is skewed to the left, meaning that there are more people with no children. The median number of children is 1, and the IQR is approximately 1. This means that the middle 50% of the population has between 0 and 2 children. There are a few outliers, which are points that are more than 1.5 IQRs below or above the 25th and 75th percentiles, respectively. These outliers could represent individuals who have no children or a large number of children.

The distribution of expenses is to the right, meaning that there are more people with higher expenses. The median expense is approximately $50,000, and the IQR is approximately $20,000. This means that the middle 50% of the population has expenses between $30,000 and $50,000. There are a few outliers, which are points that are more than 1.5 IQRs below or above the 25th and 75th percentiles, respectively. These outliers could represent individuals with very low or very high expenses.



Gender: There are more males than females in the population.

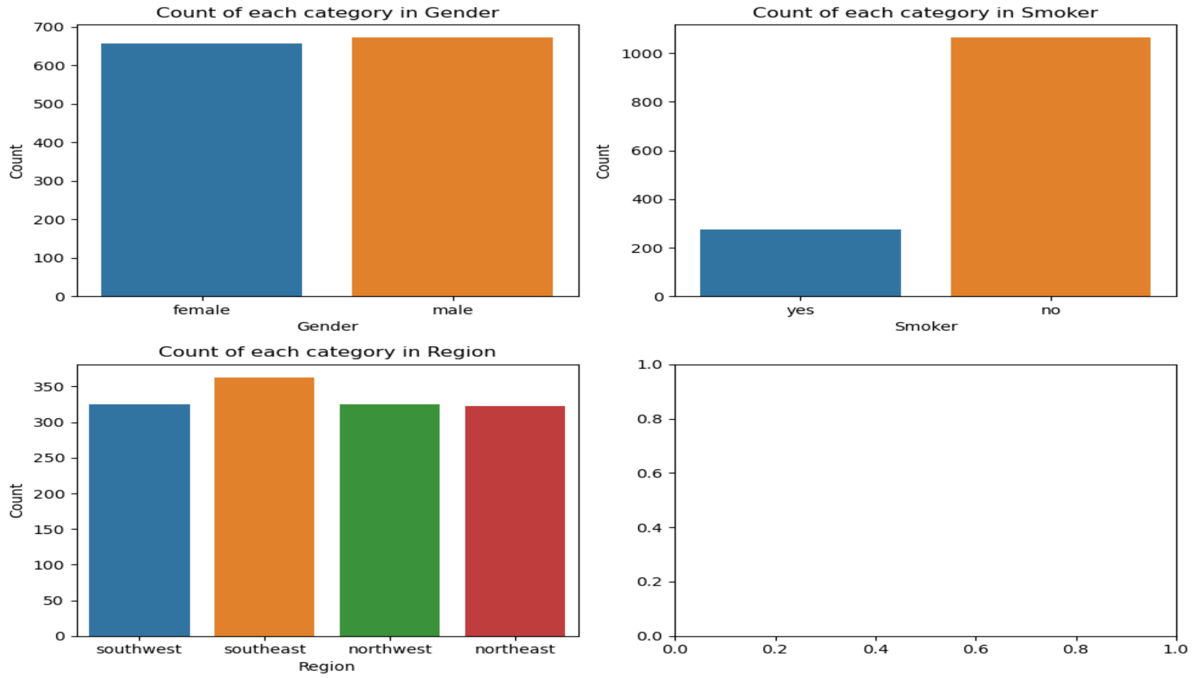
Region: The most common region is the southwest (350), followed by the northeast (300), southeast (250), and northwest (200).

Smoker: There are more non-smokers nearly (1000) than smokers (300 approximately) in the population.

Analysis:

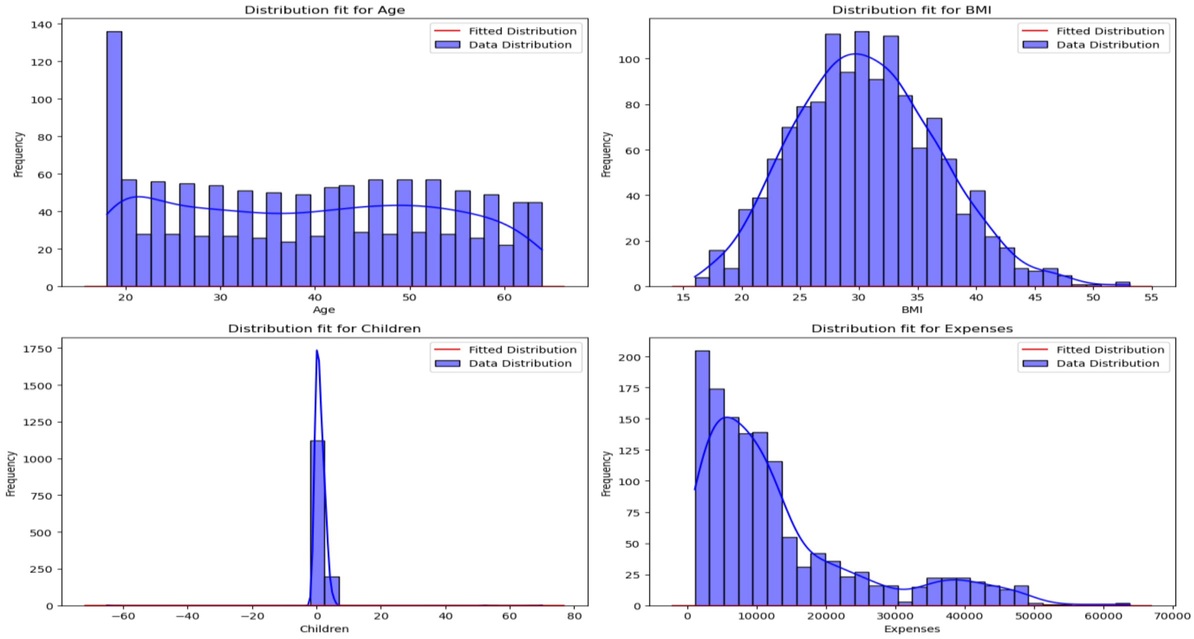
- The population is predominantly male and non-smoking.
- The southwest region has the most people, followed by the northeast, southeast, and northwest.

- There are more non-smokers than smokers in the population



Probability Distribution Modeling

Below is the probability distribution model plot for the Age , BMI , Children and Expenses column



Linear Regression Model Metrics:

```
Mean Squared Error: 0.01
R-squared: 0.69

Model Coefficients:
             Feature  Coefficient
0                Age     0.184887
1                BMI     0.176658
2           Children     0.038078
3      Gender_female    -0.001086
4        Gender_male     0.001086
5          Smoker_no    -0.191352
6         Smoker_yes     0.191352
7   Region_northeast     0.008167
8   Region_northwest     0.003222
9   Region_southeast    -0.006915
10  Region_southwest    -0.004473
Intercept: 0.13
```

Model Analysis

1. Mean Squared Error (MSE):

The MSE is a measure of the average squared difference between actual and predicted values. In above case, the MSE is 0.01. Lower MSE values indicate better model performance.

2. R-squared (R^2):

R-squared is a statistical measure of how well the regression predictions approximate the real data points. It ranges from 0 to 1, with 1 indicating a perfect fit. In above case, R-squared is 0.69, suggesting that the model explains about 69% of the variance in the dependent variable.

3. Model Coefficients:

These coefficients represent the weights assigned to each feature in the model. They indicate the strength and direction of the relationship between each feature and the predicted outcome.For example, a positive coefficient for "Age" suggests that as age increases, the predicted outcome also increases.

4. Intercept:

The intercept is the y-intercept of the regression line. In above case, it is 0.13. It represents the predicted outcome when all features are zero.

In summary, model seems to have a moderate level of explanatory power (R-squared = 0.69), and the coefficients provide insights into the relationships between the features and the predicted outcome. The negative and positive signs of the coefficients indicate the direction of the impact of each feature on the prediction.