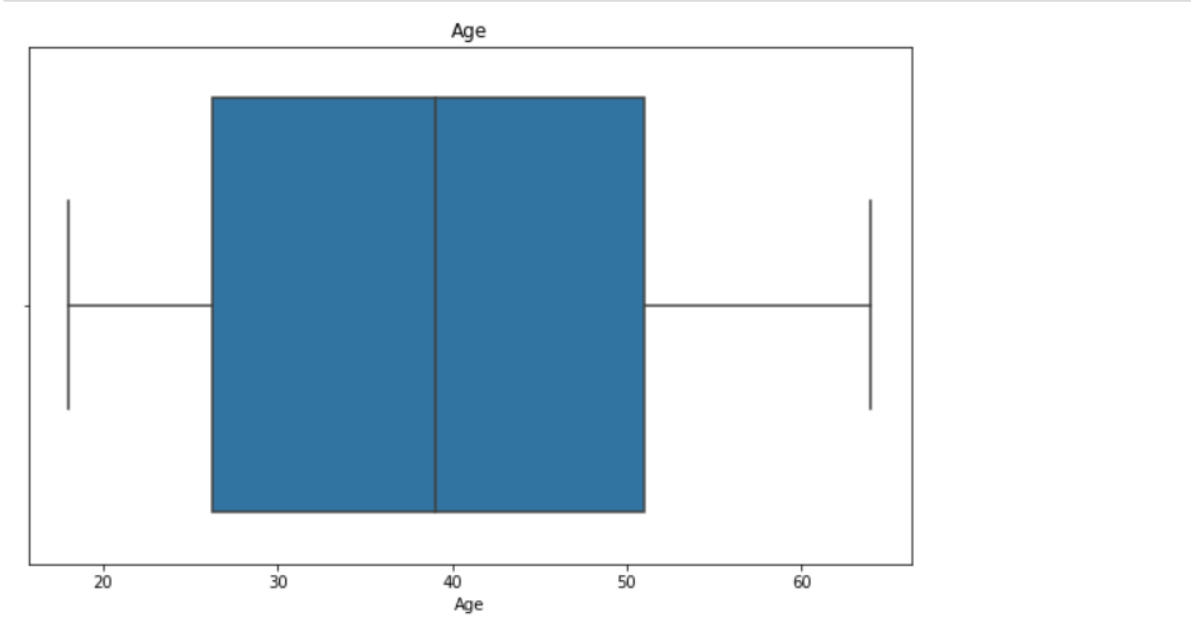
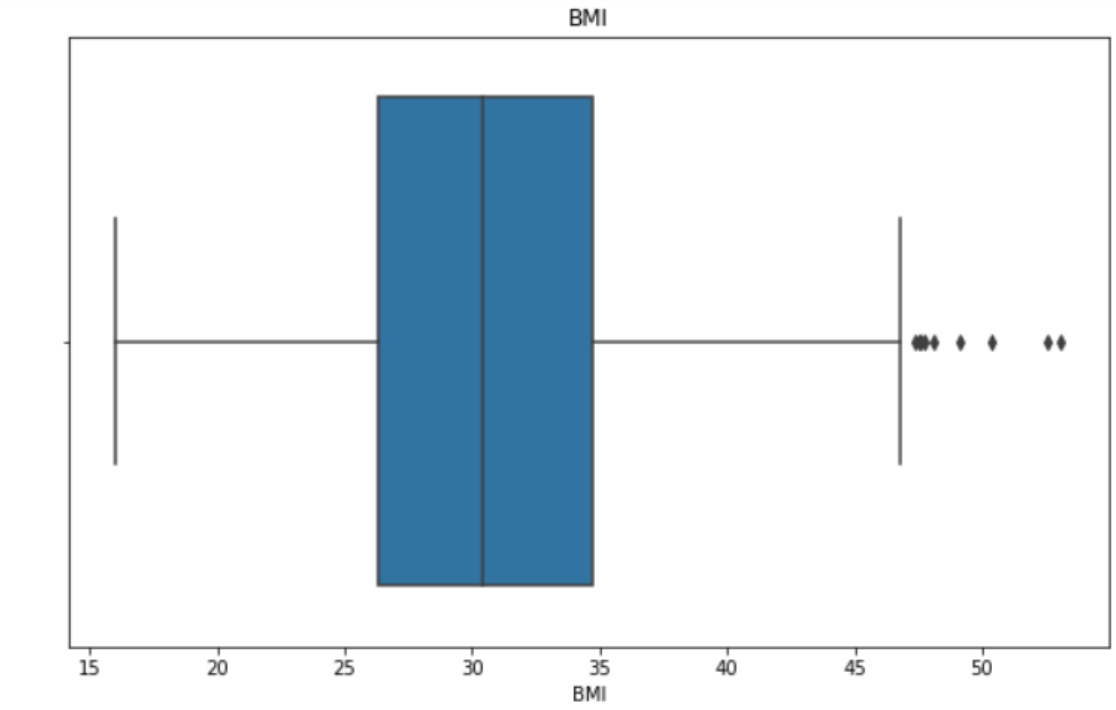
Final Assignment

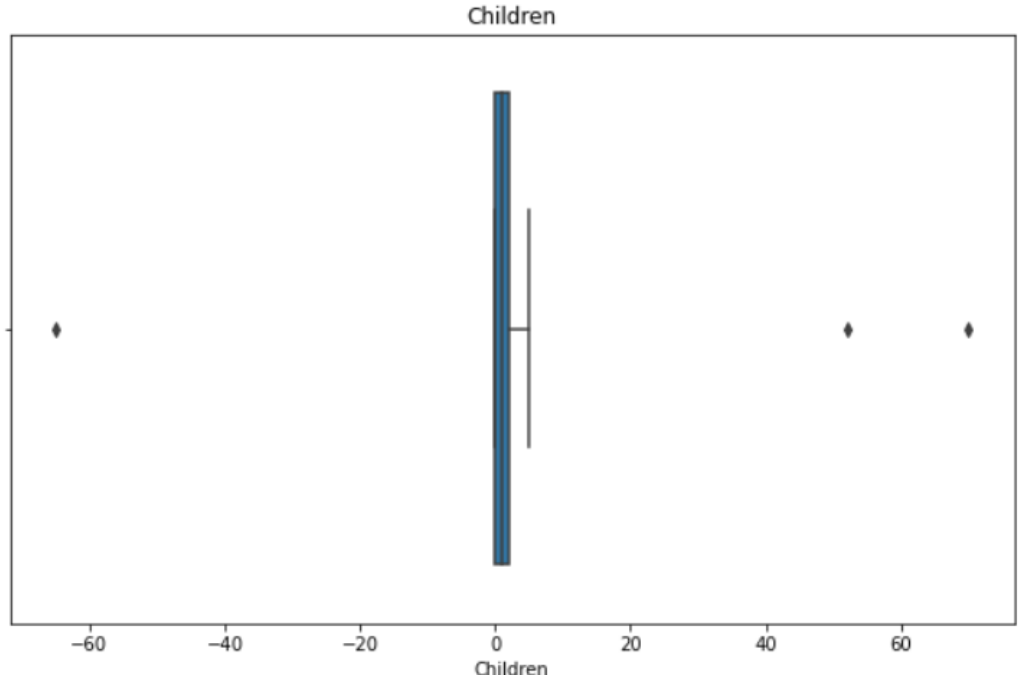
The dataset contains 7 columns: Age (numeric), Gender (categorical: male/female), BMI (floating point < 100), Children (integer), Smoker (categorical: yes/no), Region (categorical: northwest/northwest/southeast/southwest), and Expenses (floating point).

	Age	Gender	BMI	Children	Smoker	Region	Expenses
0	19.0	female	27.9	0	yes	southwest	16884.92
1	18.0	male	33.8	1	no	southeast	1725.55
2	28.0	male	33.0	3	no	southeast	4449.46
3	33.0	male	22.7	0	no	northwest	21984.47
4	32.0	male	28.9	0	no	northwest	3866.86

Outlier Detection and Removal:





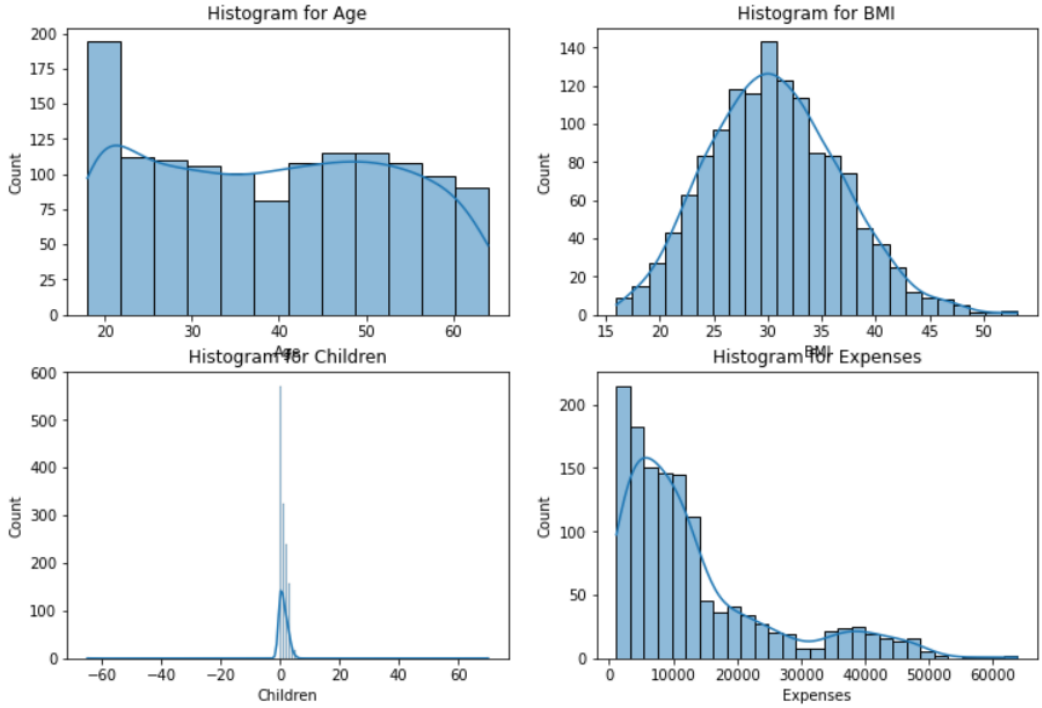


From this we get to know , the outliers are minimum and that too for the case of BMI and children , hence they are removed.

We had few missing values and they are filled up using mena and median

```
Missing Values:
Age         4
Gender      9
BMI         8
Children    0
Smoker      0
Region      2
Expenses    1
dtype: int64
Shape of the original DataFrame: (1338, 7)
Shape of the DataFrame after handling missing values: (1315, 7)
```

If we look at the histrogram of the numerical values in the numerical values

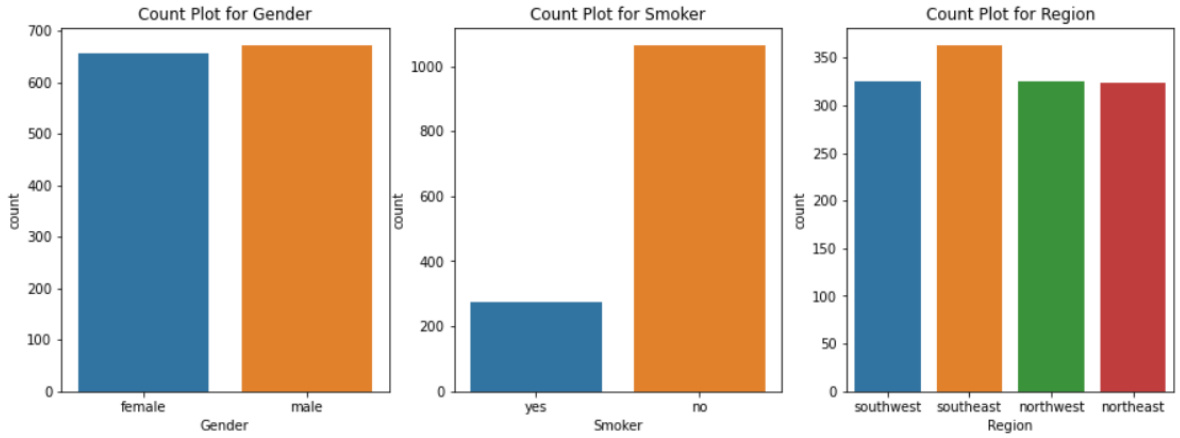


We see , the people in their 20's have more insurances and people in 40 with the least , followed by 60's.

We see , that between 31 and 34 , the maximum of the BMI lies for the people taken the insurance.

When is look into number of children , most of them have zero kids , followed by one.

when we look into other factors



Both male and female ratio is almost same , however male are silgtly higher

We see a lot of difference when we consider if the person is smoker or not , they is almost a ratio of 80 percent of them being non smoker and 20 percent of them being non – smoker. If we look into it much these smokers pay a lot of expenses.

After factor is the region , the southern , notherwest and northeast are same while southeast region is slightly more.

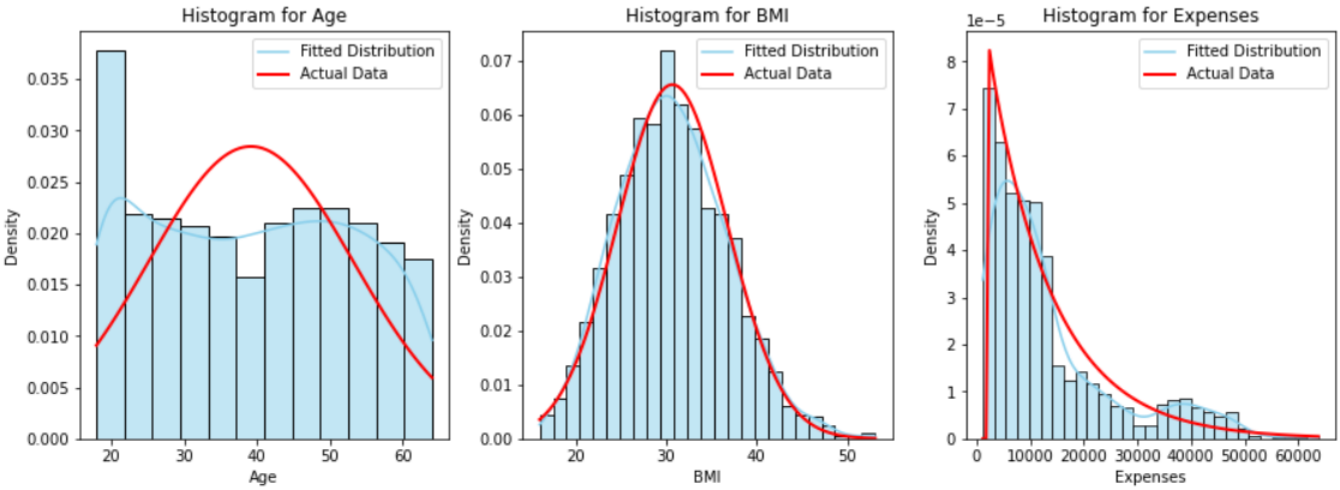
When we look into more of the descriptive analysis of the same:

```
Descriptive Statistics:
                Age          BMI     Children      Expenses
count  1338.000000  1338.000000  1338.000000   1337.000000
mean     39.176912    30.675262     1.136024  13273.306111
std      14.020347     6.076644     3.194662  12114.083012
min      18.000000    16.000000   -65.000000   1121.870000
25%      27.000000    26.300000     0.000000   4738.270000
50%      39.000000    30.400000     1.000000   9377.900000
75%      51.000000    34.600000     2.000000  16657.720000
max      64.000000    53.100000    70.000000  63770.430000

Variability Measures:
Age         1.965701e+02
BMI         3.692560e+01
Children    1.020587e+01
Expenses    1.467510e+08
dtype: float64

Mode Values:
Age           18.00
BMI           27.60
Children       0.00
Expenses    1639.56
Name: 0, dtype: float64
```

We notice here max age reported is 64 , while min is 18 and average is 39. Similarly max BMI reported is 53 , while 16 is 18 and average is 30.6. While most of the people has one child and they expenses has a standard deviation od 12114. The quartiles is as below:

```
Quartiles:
       Age   BMI  Children  Expenses
0.25  27.0  26.3       0.0   4738.27
0.50  39.0  30.4       1.0   9377.90
0.75  51.0  34.6       2.0  16657.72
```

When we look at the at the histrogram of the fitted data , it looks like below :



From the histogram we learn that there is a difference with respect to the age matrix , we see the actual data was projected to indicate the maximum value of around 40 , since the fitted data is projected the minimum values of around 60. While the peak curve with respect to the BMI and Expenses remain the same.

When the regression model is applied , we get the parameters as listed below:

```
Standardized DataFrame:
        Age       BMI  Children  Expenses
0 -1.439655 -0.456880 -0.355734  0.298245
1 -1.511006  0.514413 -0.042594 -0.953607
2 -0.797490  0.382713  0.583684 -0.728668
3 -0.440732 -1.312936 -0.355734  0.719363
4 -0.512084 -0.292254 -0.355734 -0.776779

Normalized DataFrame:
        Age       BMI  Children  Expenses
0  0.021739  0.320755  0.481481  0.251611
1  0.000000  0.479784  0.488889  0.009636
2  0.217391  0.458221  0.503704  0.053115
3  0.326087  0.180593  0.481481  0.333010
4  0.304348  0.347709  0.481481  0.043816
```

```
Model Coefficients:
Intercept: -4469.830982243126
Coefficients: {'Age': 224.45401898014998, 'BMI': 276.4528212031858, 'Children': 175.85641794451627}

Performance Metrics:
Mean Squared Error: 159928124.43836826
R-squared (R2): 0.1427458718787501
```

When is apply the same on the test parameter , we get

```
Model Coefficients:
Intercept: -4469.830982243126
Coefficients: {'Age': 224.45401898014998, 'BMI': 276.4528212031858, 'Children': 175.85641794451627}

Performance Metrics:
Mean Squared Error: 159928124.43836826
R-squared (R2): 0.1427458718787501
```

Reference:

https://realpython.com/linear-regression-in-python/