# Univariate Exploratory Data Analysis on New York Air Quality Dataset

*Arjun Dutta*

*31 Jan 2018*

## *A. About the Dataset*

### Data Set

New York Air Quality Measurements

### Description

Daily air quality measurements in New York, May to September 1973.

### Format

A data frame with 153 observations on 6 variables.

- **Ozone :** numeric Ozone (ppb)
- **Solar.R :** numeric Solar R (lang)
- **Wind :** numeric Wind (mph)
- **Temp :** numeric Temperature (degrees F)
- **Month :** numeric Month (1–12)
- **Day :** numeric Day of month (1–31)

### Details

Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.

- **Ozone:** Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
- **Solar.R:** Solar radiation in Langleys in the frequency band 4000-7700 Angstroms from 0800 to 1200 hours at Central Park
- **Wind:** Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
- **Temp:** Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.

### Source:

The data were obtained from the New York State Department of Conservation (**ozone data**) and the National Weather Service (**meteorological data**).

## *B. Analyzing the Structure of the Data*

Here we show the Data Structure of the dataset using the **class()** function, Dimension of the dataset by **dim()** and Data types of each variable by using **glimpse()** function from **base** and **dplyr** packages in R.Next identify the data type and category of the variables.

### DataStructure

```
[1] "data.frame"
```

### Dimension

```
[1] 153   6
```

### Data Types

```
Observations: 153
Variables: 6
$ Ozone   <int> 41, 36, 12, 18, NA, 28, 23, 19, 8, NA, 7, 16, 11, 14, ...
$ Solar.R <int> 190, 118, 149, 313, NA, NA, 299, 99, 19, 194, NA, 256,...
$ Wind    <dbl> 7.4, 8.0, 12.6, 11.5, 14.3, 14.9, 8.6, 13.8, 20.1, 8.6...
$ Temp    <int> 67, 72, 74, 62, 56, 66, 65, 59, 61, 69, 74, 69, 66, 68...
$ Month   <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ...
$ Day     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
```

The Data Structure of the dataset is a dataframe with Dimension of 153 rows and 6 columns and Data types are **int** and **dbl** i.e. integer and double(numeric) respectively.

### Category of the Variable

The variable with **int** and **dbl** data type are of the format of continuous variable.

## *C. Outlier Detection and Treatment*

### Outliers

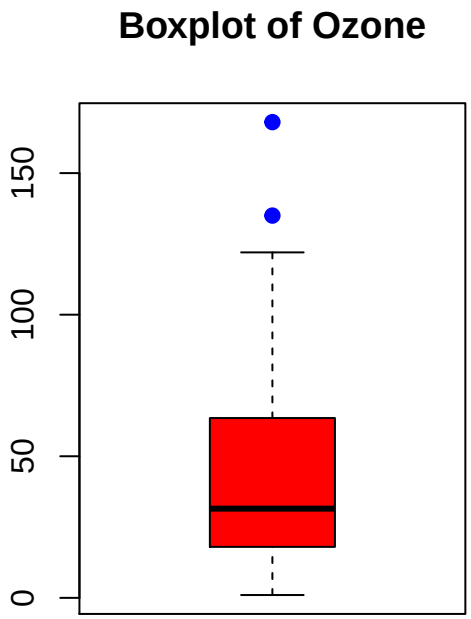
For a given continuous variable, outliers are those observations that lie outside 1.5 * IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles.

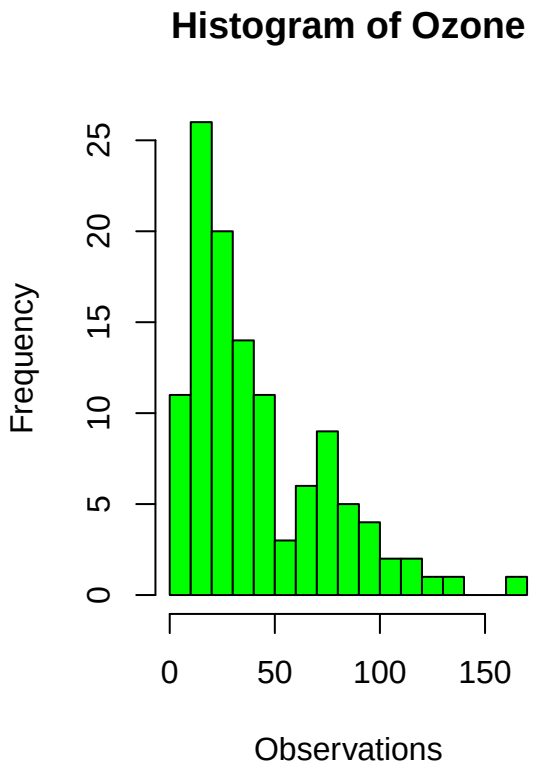
## Detecting Outliers

### Visualization and Mathematical Methods

For Visualization Methods Boxplot with range 1.5 and Histogram with break 15 is used to get a clear idea about the data. Quantile Capping Method is used to detect the outliers(Mathematically) in the data for each variable after Visualization.

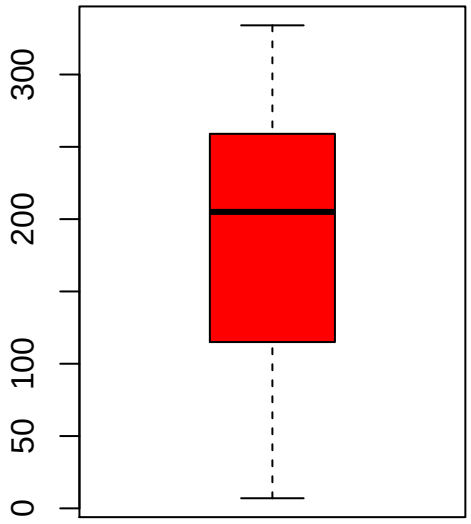
**Ozone**

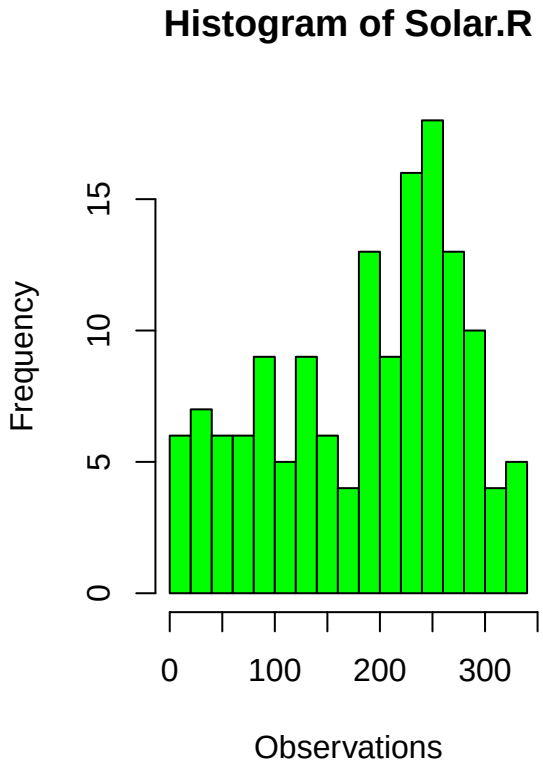



```
There are 0 observations below the 1st quantile
There are 2 observations above the 3rd quantile on rows 62 117 and the values are 135 168
```

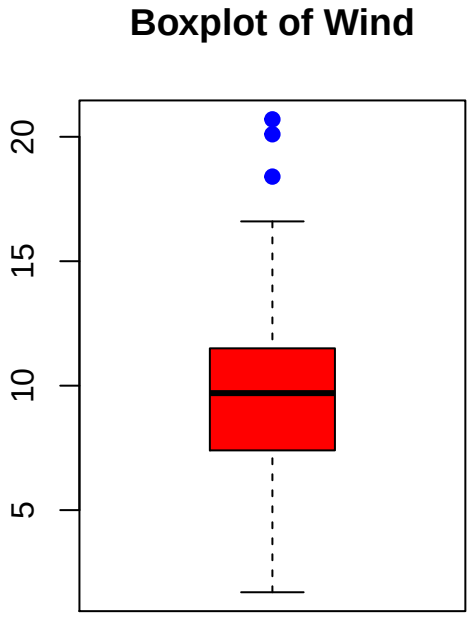
**Solar.R**

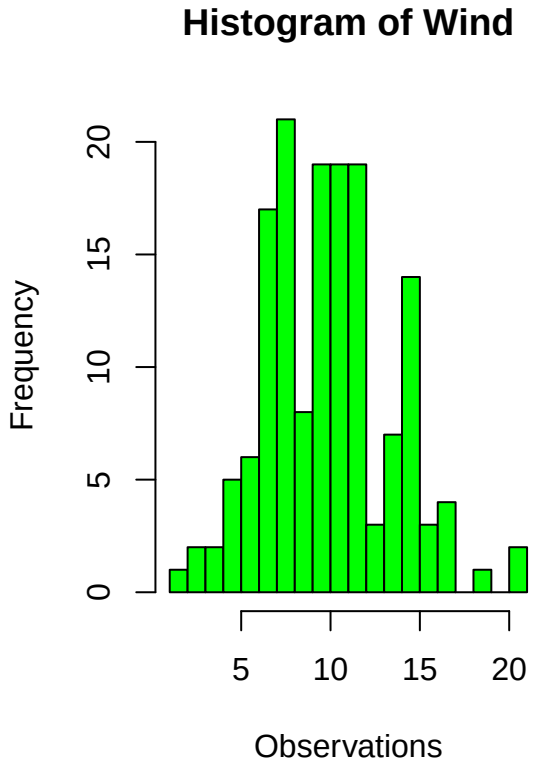



```
There are 0 observations below the 1st quantile
There are 0 observations above the 3rd quantile
```

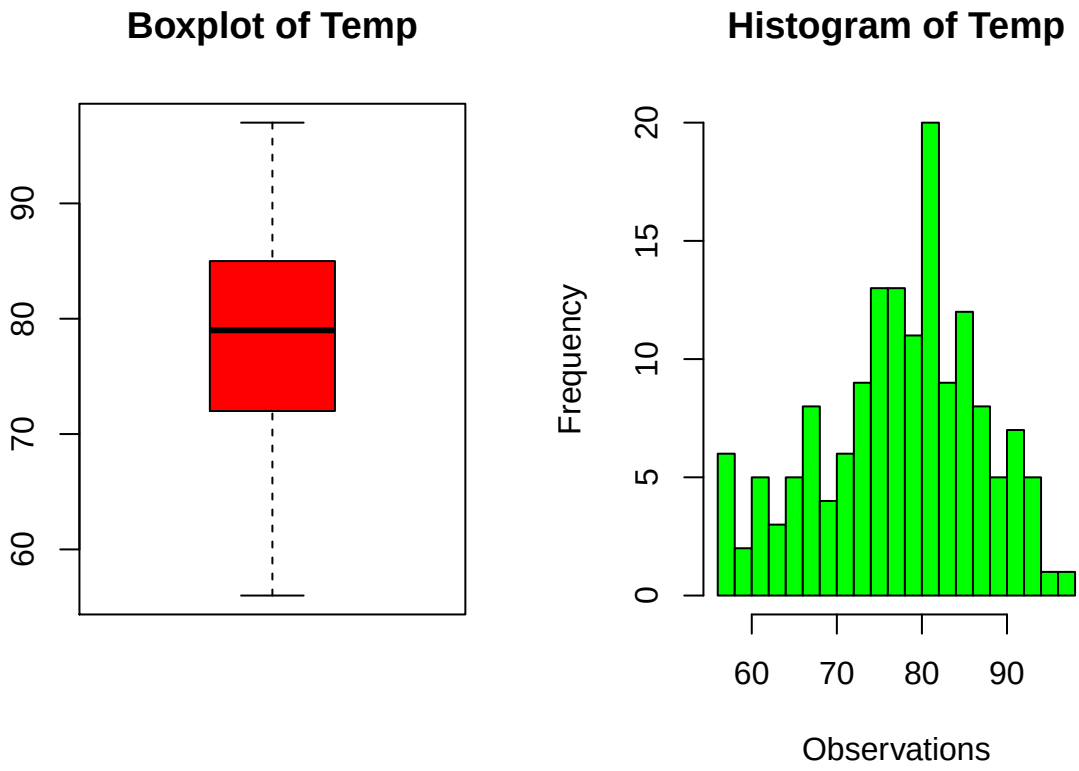
**Wind**



```
There are 0 observations below the 1st quantile
There are 3 observations above the 3rd quantile on rows 9 18 48 and the values are 20.1 18.4 2
```

**Temp**




```
There are 0 observations below the 1st quantile
There are 0 observations above the 3rd quantile
```

From the Above Boxplots we can see that in Ozone and Wind Variables there are **Outliers**(the blue dots). Also on Histogram of both the variables Ozone and Wind we can see that there is a gap between observations at extreme i.e. In Ozone Histogram there is one gap in the chart and in Wind Histogram there are two gaps in chart, so they are Outliers.

## Outliers Treatment

Since the number of outliers in the dataset is very small the best approach is to remove them and carry on with the analysis but capping method can also be used. Percentile Capping is a method of imputing the outlier values by replacing those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile of the same dataset.

```
    Ozone Wind
9     8.0 15.5
18    6.0 15.5
48   37.0 15.5
62  108.5  4.1
117 108.5  3.4
```

Now check the 9th, 18th and 48th value of Wind and 62nd 117th value of Ozone variable the Varibles are free of Outliers so we move to treat Missing Values.

# *D. Missing Value Detection and Treatment*

## Missing Values

Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

## Detecting Missing Values

### Mathematical Methods

To check for **Missing Value** call the **summary()** on the dataset.

**Summary of the Data**

```
     Ozone           Solar.R           Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 41.39   Mean   :185.9   Mean   : 9.875   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :122.00   Max.   :334.0   Max.   :16.600   Max.   :97.00  
 NA's   :37       NA's   :7                                       
```

See that on Ozone there are 37 NA's and on Solar.R there are 7 NA's.

**Names of the Columns which contains Missing Values**

```
[1] "Ozone"   "Solar.R"
```

**Percentage of Columns and Rows which contains Missing Values**

Assuming the data is **Missing Completly At Random**, too much missing data can be a problem too. A safe maximum threshold is 5% of the total for large datasets. If missing data for a certain Variable is more than 5% then leave that Variable out. Let's check the columns and rows where more than 5% of the data is missing using a simple function :-

**Columns**

```
    Ozone   Solar.R 
24.183007  4.575163 
```

**Rows**

```
  [1]  0.00000  0.00000  0.00000  0.00000 33.33333 16.66667  0.00000
  [8]  0.00000  0.00000 16.66667 16.66667  0.00000  0.00000  0.00000
 [15]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
 [22]  0.00000  0.00000  0.00000 16.66667 16.66667 33.33333  0.00000
 [29]  0.00000  0.00000  0.00000 16.66667 16.66667 16.66667 16.66667
```

```
 [36] 16.66667 16.66667  0.00000 16.66667  0.00000  0.00000 16.66667
 [43] 16.66667  0.00000 16.66667 16.66667  0.00000  0.00000  0.00000
 [50]  0.00000  0.00000 16.66667 16.66667 16.66667 16.66667 16.66667
 [57] 16.66667 16.66667 16.66667 16.66667 16.66667  0.00000  0.00000
 [64]  0.00000 16.66667  0.00000  0.00000  0.00000  0.00000  0.00000
 [71]  0.00000 16.66667  0.00000  0.00000 16.66667  0.00000  0.00000
 [78]  0.00000  0.00000  0.00000  0.00000  0.00000 16.66667 16.66667
 [85]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
 [92]  0.00000  0.00000  0.00000  0.00000 16.66667 16.66667 16.66667
 [99]  0.00000  0.00000  0.00000 16.66667 16.66667  0.00000  0.00000
[106]  0.00000 16.66667  0.00000  0.00000  0.00000  0.00000  0.00000
[113]  0.00000  0.00000 16.66667  0.00000  0.00000  0.00000 16.66667
[120]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
[127]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
[134]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
[141]  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000
[148]  0.00000  0.00000 16.66667  0.00000  0.00000  0.00000
```

We see that Ozone is missing almost 25% of the datapoints, therefore we might consider either dropping it from the analysis or gather more measurements. The Wind variables have below 5% threshold so we can keep them. As far as the Rows are concerned, missing just one feature leads to a 17% missing data per sample. Row Observations that are missing 2 or more Variables (>34%), should be dropped if possible.

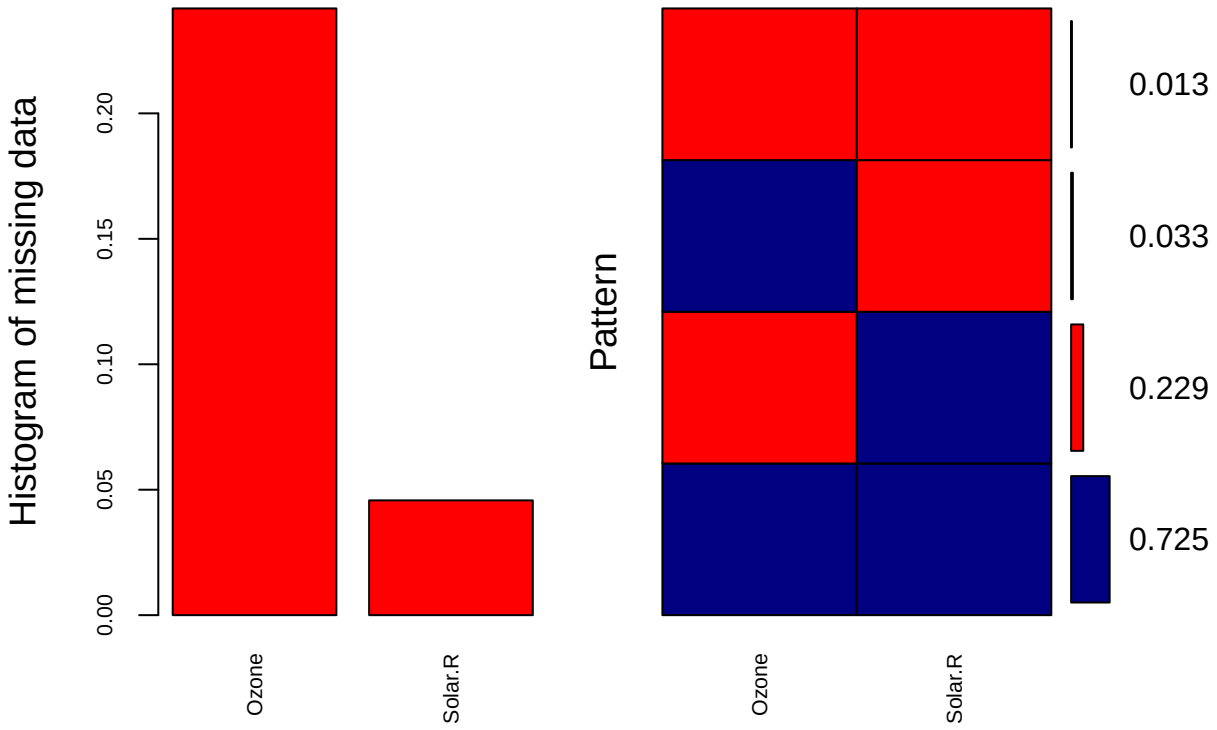
**Patterns and Visualizations of Missing Values**

The mice package provides a nice function md.pattern() to get a better understanding of the pattern of missing data. **Patterns**

```
    Wind Temp Month Day Solar.R Ozone
111    1    1     1   1       1     1  0
 35    1    1     1   1       1     0  1
  5    1    1     1   1       0     1  1
  2    1    1     1   1       0     0  2
       0    0     0   0       7    37 44
```

The output tells us that 44 samples are complete, 37 samples miss only the Ozone measurement, 7 samples miss only the Solar.R value.

A perhaps more helpful visual representation can be obtained using the VIM package as follows **Visual-**

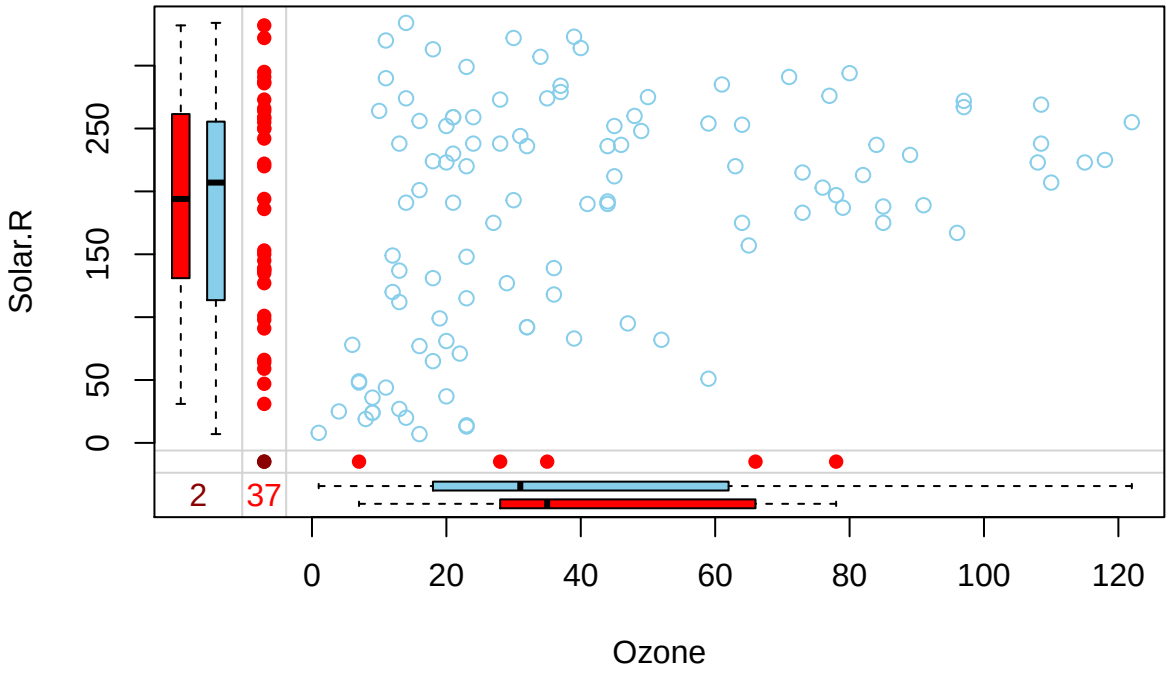



**izations**

```
Variables sorted by number of missings:
Variable      Count
   Ozone 0.24183007
 Solar.R 0.04575163
```

The plot helps us understanding that almost 72% of the samples are not missing any information, 24% are missing the Ozone value, and the remaining ones show other missing patterns. Through this approach the situation looks a bit clearer in my opinion.

Another helpful visual approach is a special box plot




Obviously here we are constrained at plotting 2 variables at a time only, but nevertheless we can gather some interesting insights. The red box plot on the left shows the distribution of Solar.R with Ozone missing while the blue box plot shows the distribution of the remaining datapoints. Likewise for the Ozone box plots at the bottom of the graph. If our assumption of MCAR data is correct, then we expect the red and blue box plots to be very similar.

## Missing Value Treatment

The mice() function takes care of the imputing process.PMM (Predictive Mean Matching) - technique is used because it is suitable for numeric variables.

### Summary of Missing Values

```
Multiply imputed data set
Call:
mice(data = Data, m = 5, method = "pmm", maxit = 50, seed = 500)
Number of multiple imputations:  5
Missing cells per column:
  Ozone Solar.R    Wind    Temp   Month     Day
     37       7       0       0       0       0
Imputation methods:
  Ozone Solar.R    Wind    Temp   Month     Day
  "pmm"   "pmm"   "pmm"   "pmm"   "pmm"   "pmm"
VisitSequence:
  Ozone Solar.R
      1       2
PredictorMatrix:
```

```
        Ozone Solar.R Wind Temp Month Day
Ozone       0       1    1    1     1   1
Solar.R     1       0    1    1     1   1
Wind        0       0    0    0     0   0
Temp        0       0    0    0     0   0
Month       0       0    0    0     0   0
Day         0       0    0    0     0   0
Random generator seed value:  500
```

**Summary of the Dataset(Before Treating Missing Values and Outlier)**

```
     Ozone           Solar.R
 Min.   :  1.00   Min.   :  7.0
 1st Qu.: 18.00   1st Qu.:115.8
 Median : 31.50   Median :205.0
 Mean   : 42.13   Mean   :185.9
 3rd Qu.: 63.25   3rd Qu.:258.8
 Max.   :168.00   Max.   :334.0
 NA's   :37       NA's   :7
```

**Summary of the Dataset(after Treating Missing Values and Outlier)**

```
     Ozone           Solar.R
 Min.   :  1.00   Min.   :  7.0
 1st Qu.: 16.00   1st Qu.:115.0
 Median : 30.00   Median :203.0
 Mean   : 40.15   Mean   :185.3
 3rd Qu.: 59.00   3rd Qu.:258.0
 Max.   :122.00   Max.   :334.0
```

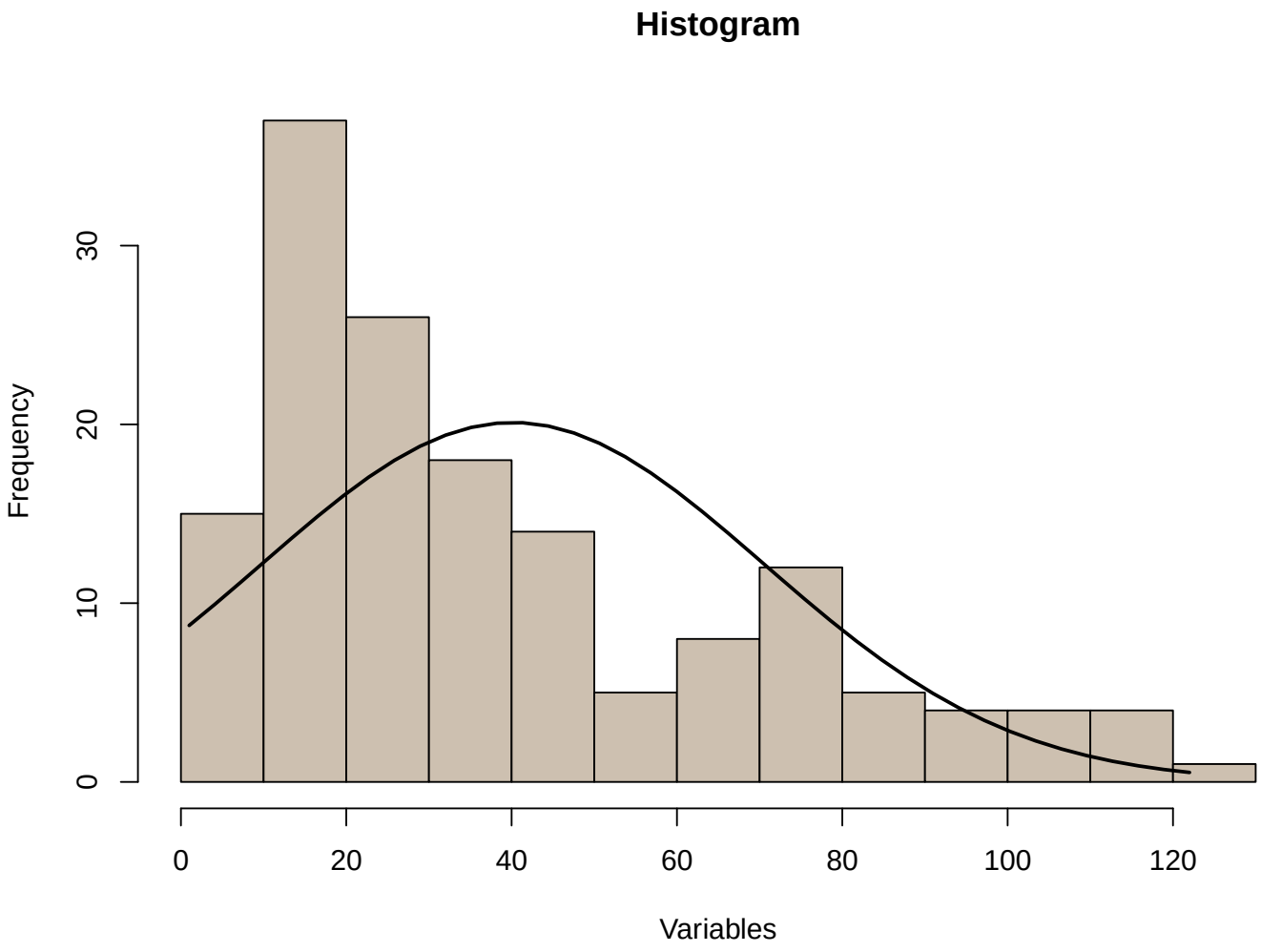
**Conclusions** If we compare the summaries of both the Datasets we can see that the values are not so much deviated from their respective summaries, so we can conclude that taking **Percentile Capping as Outlier Imputation** and **Predictive Mean Matching as Missing Value Imputation** was a right choice.

# *E. Feature Engineering and Analysis*

## Skewness Transformation

### For Ozone

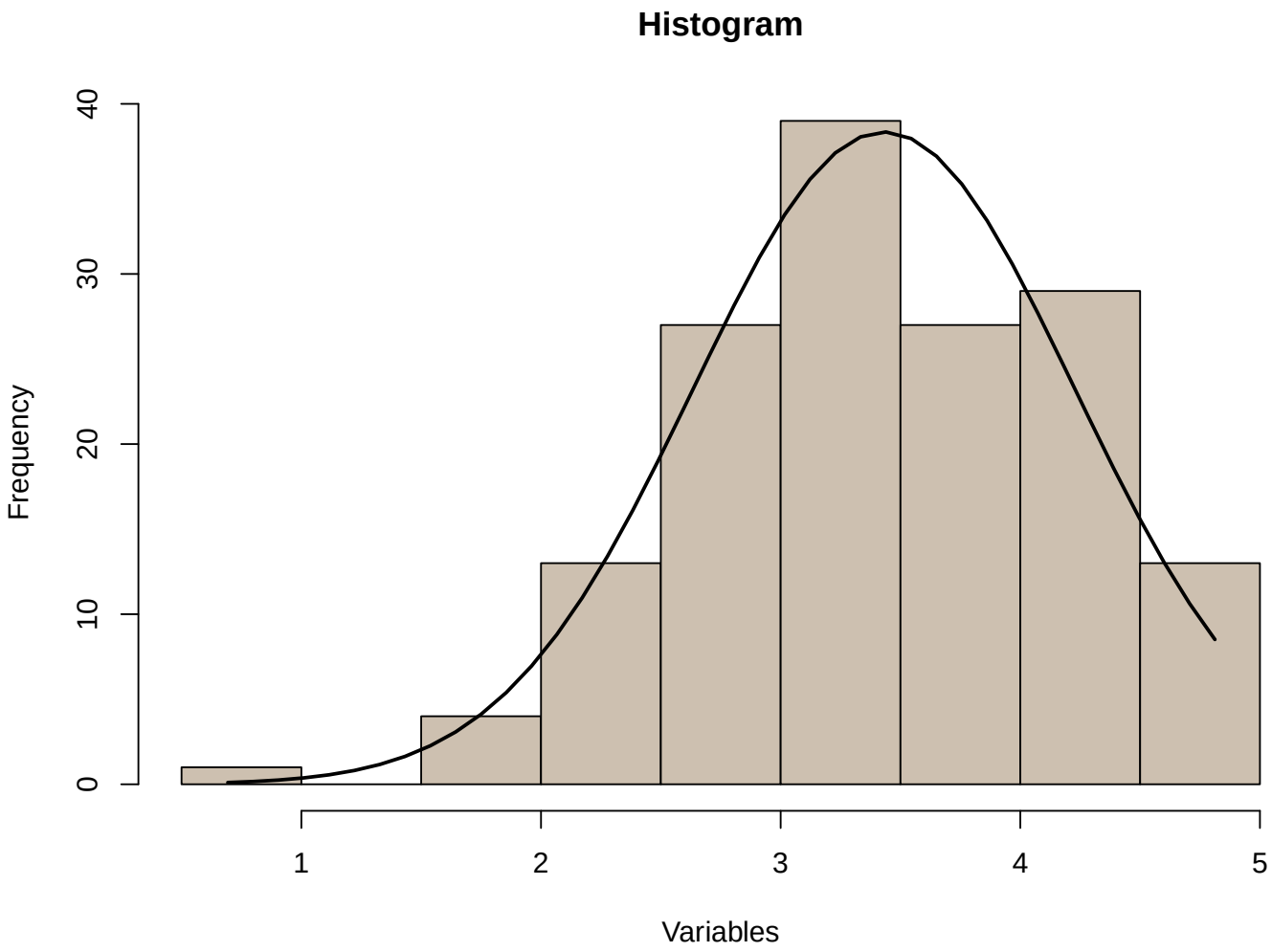
### Before Transformation Histogram




```
[1] 0.9819931
```

The histogram with the density curve of Ozone clearly shows that tail of the distribution lie towards right and thus the variable is Right Skewed.So we need to transform the Variable.

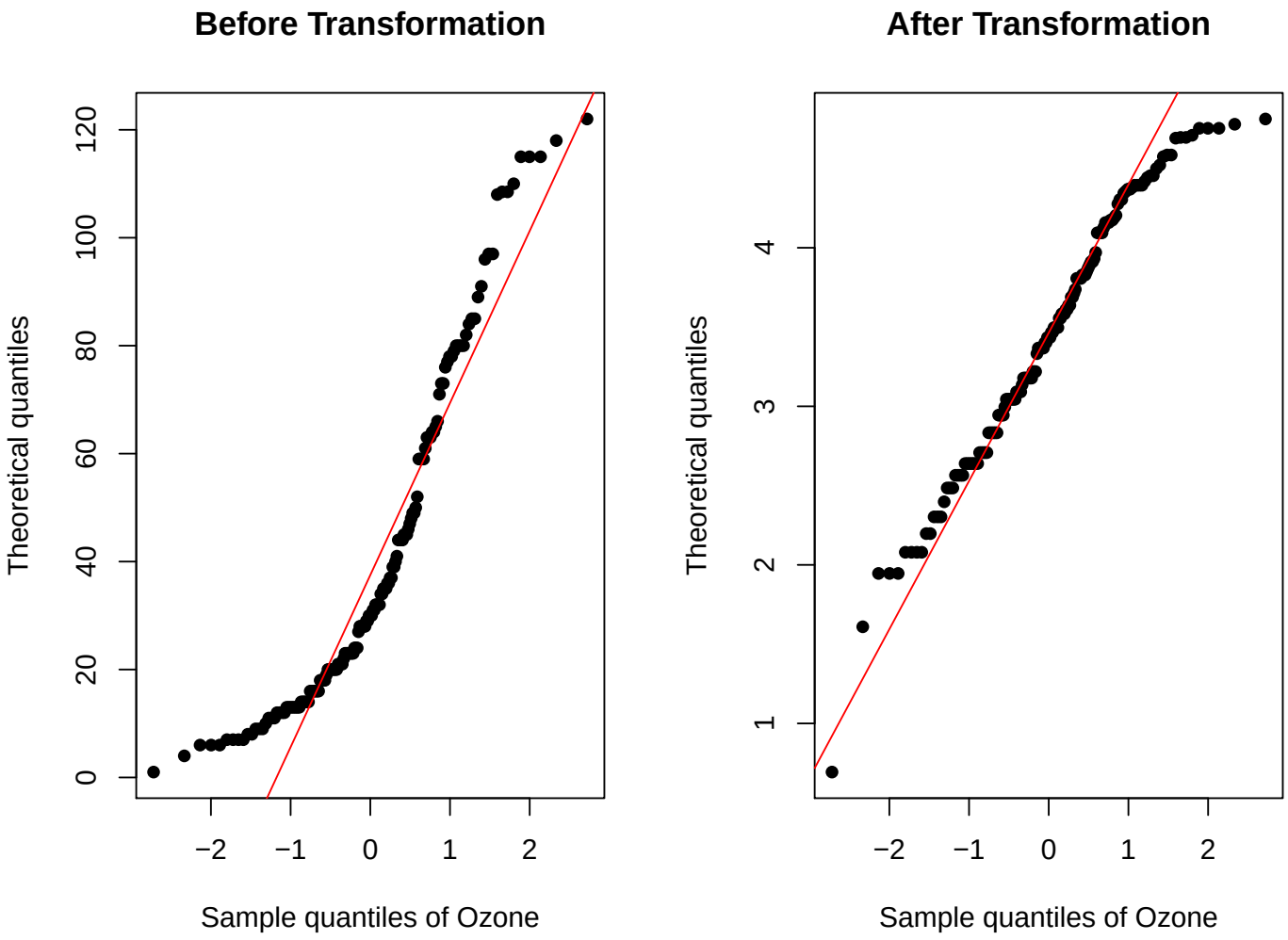
**After Transformation Histogram**




```
[1] -0.2845094
```

After **Log Transformation** the histogram with the density curve of Ozone clearly shows that maximum frequency of the values lie slightly towards left and thus the variable is nearly skewed and so the data is from normal population. Also the skewness is much close to 0.

**QQPlot**




**QQPlot** The above QQ plot clearly shows that most of the values lies above the normal line but more or less close to it. So we can interpret that the data is surely from a normal distribution.

**Hypothesis testing:**

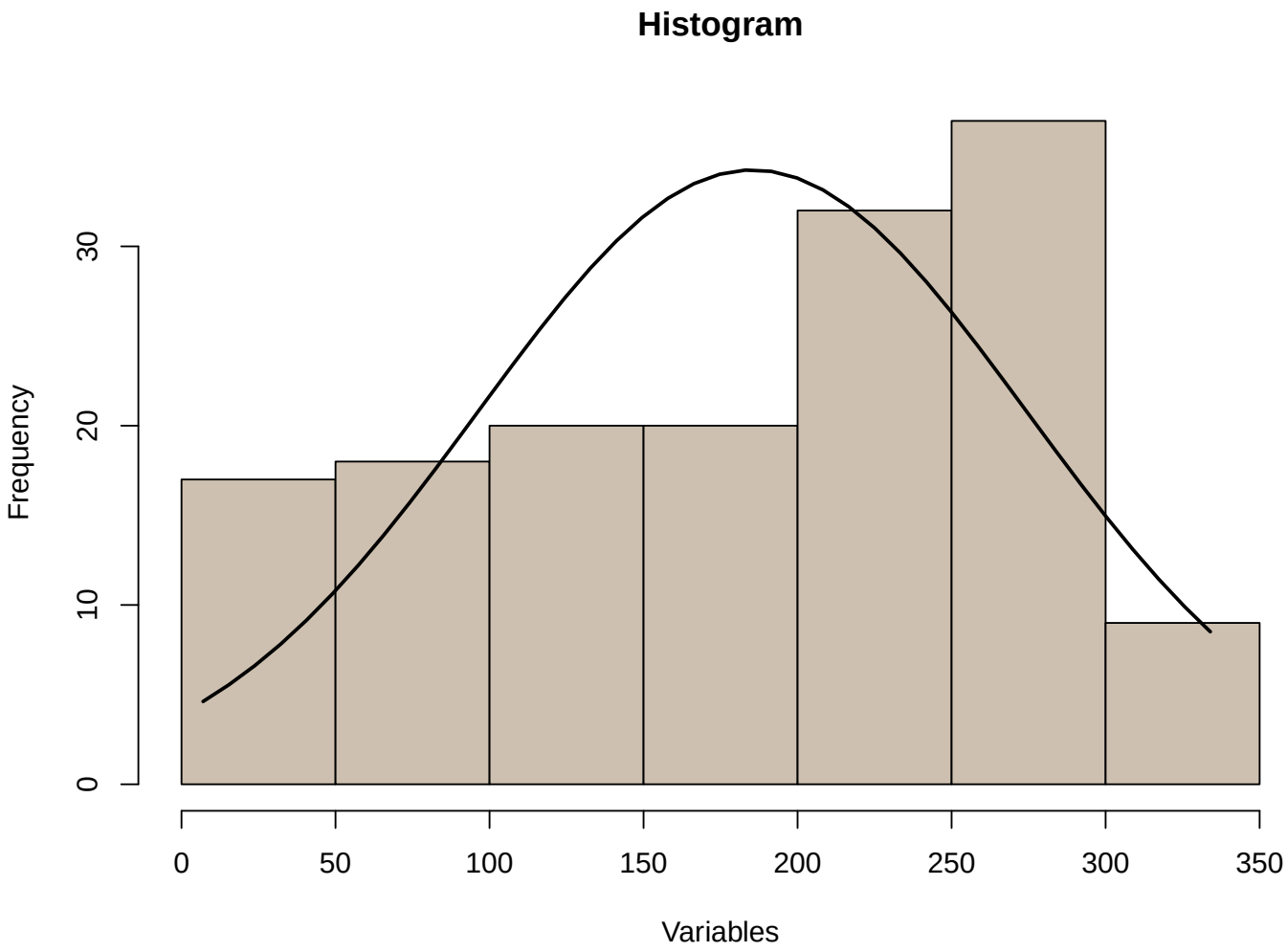
```
    Shapiro-Wilk normality test

data:  Ozone_T
W = 0.97695, p-value = 0.01139
```

If p is more than .01 $\{\alpha\}$, we can be 99% $\{(1-\alpha)100\%\}$ certain that the data are normally distributed.

**For Solar.R**

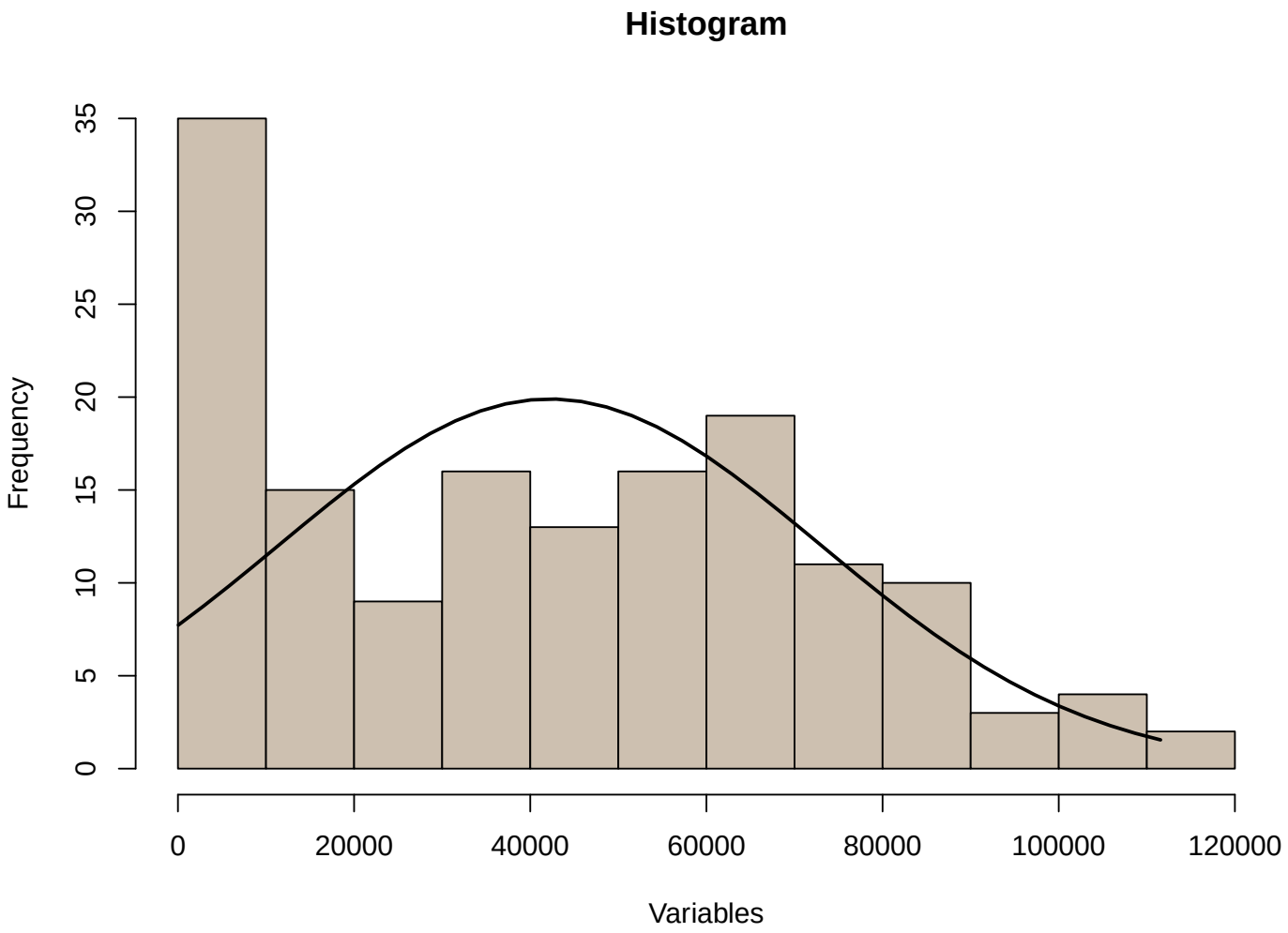
**Before Transformation Histogram**




```
[1] -0.4044689
```

The histogram with the density curve of Solar.R clearly shows that tail of the distribution lie towards left and thus the variable is Left Skewed.So we need to transform the Variable.

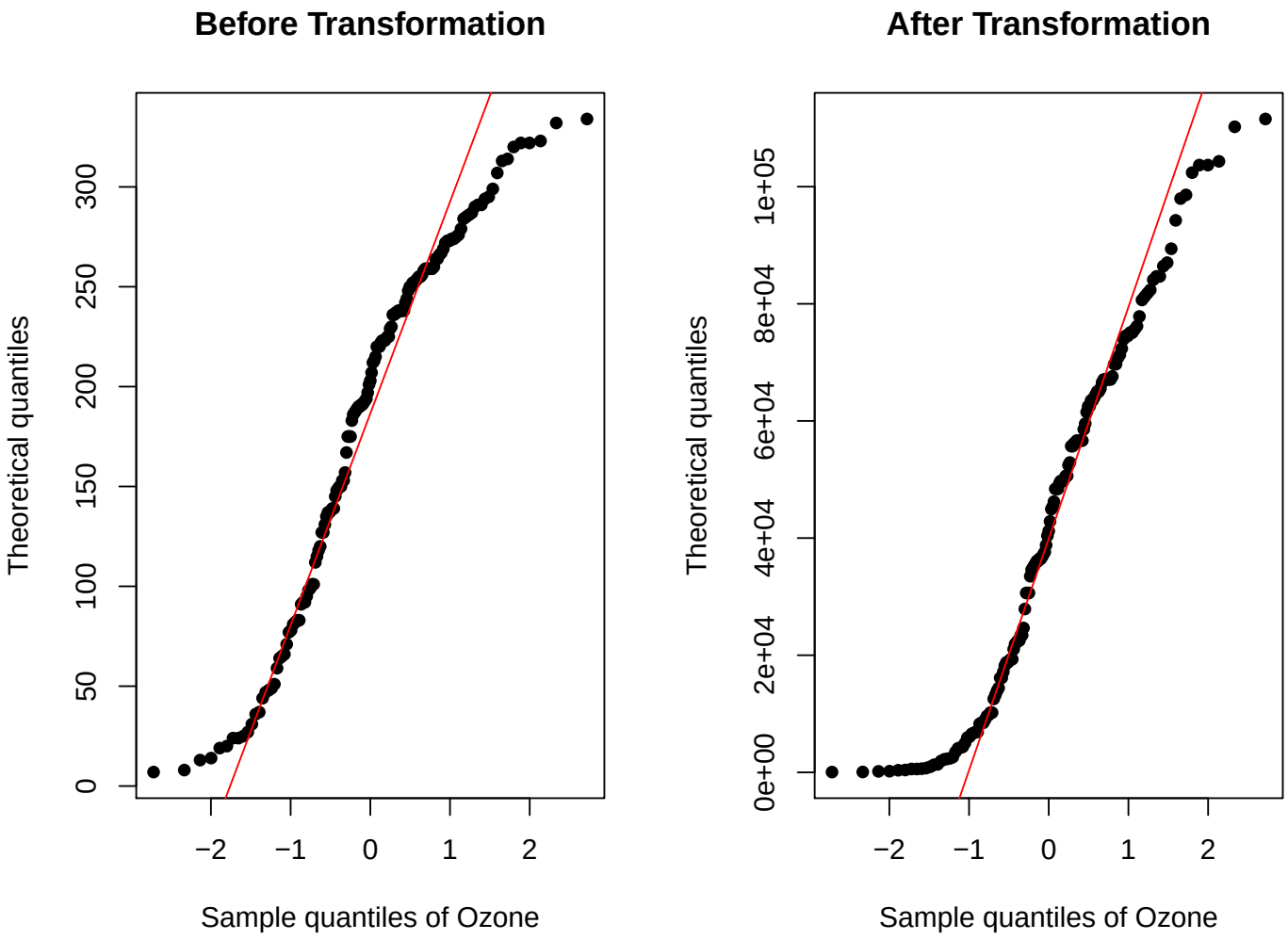
**After Transformation Histogram**




```
[1] 0.2455632
```

After **Square Transformation** the histogram with the density curve of Solar.R clearly shows that maximum frequency of the values lie slightly towards left and thus the variable is nearly skewed and so the data is from normal population. Also the skewness is much close to 0.

**QQPlot**




**QQPlot** The above QQ plot clearly shows that most of the values lies above the normal line but more or less close to it. So we can interpret that the data is surely from a normal distribution.

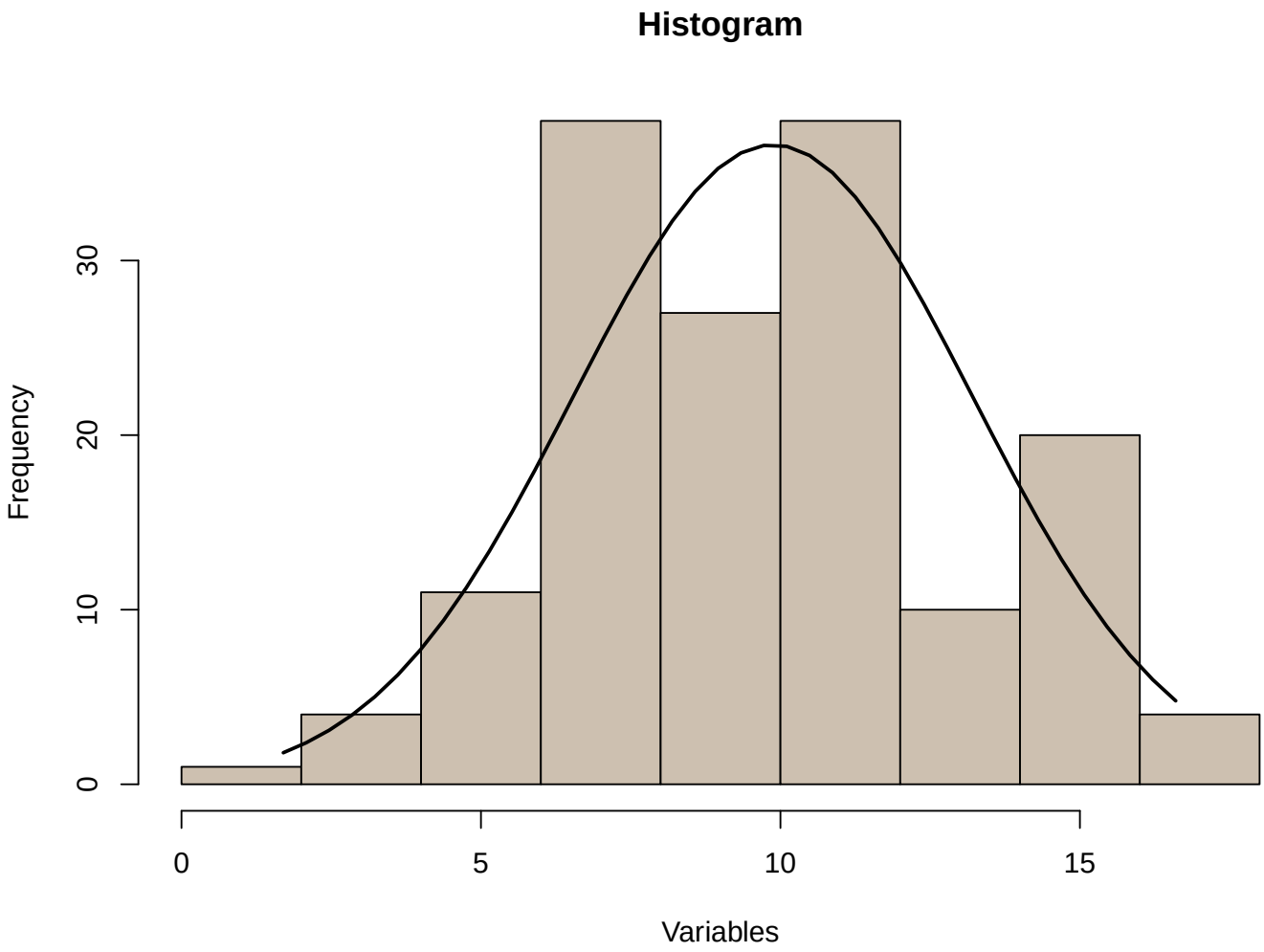
```
    Shapiro-Wilk normality test

data:  Solar_T
W = 0.94474, p-value = 1.015e-05
```

The distribution is slightly Skewed and is not normal even after transformation but it is not skewed like before.

**For Wind**

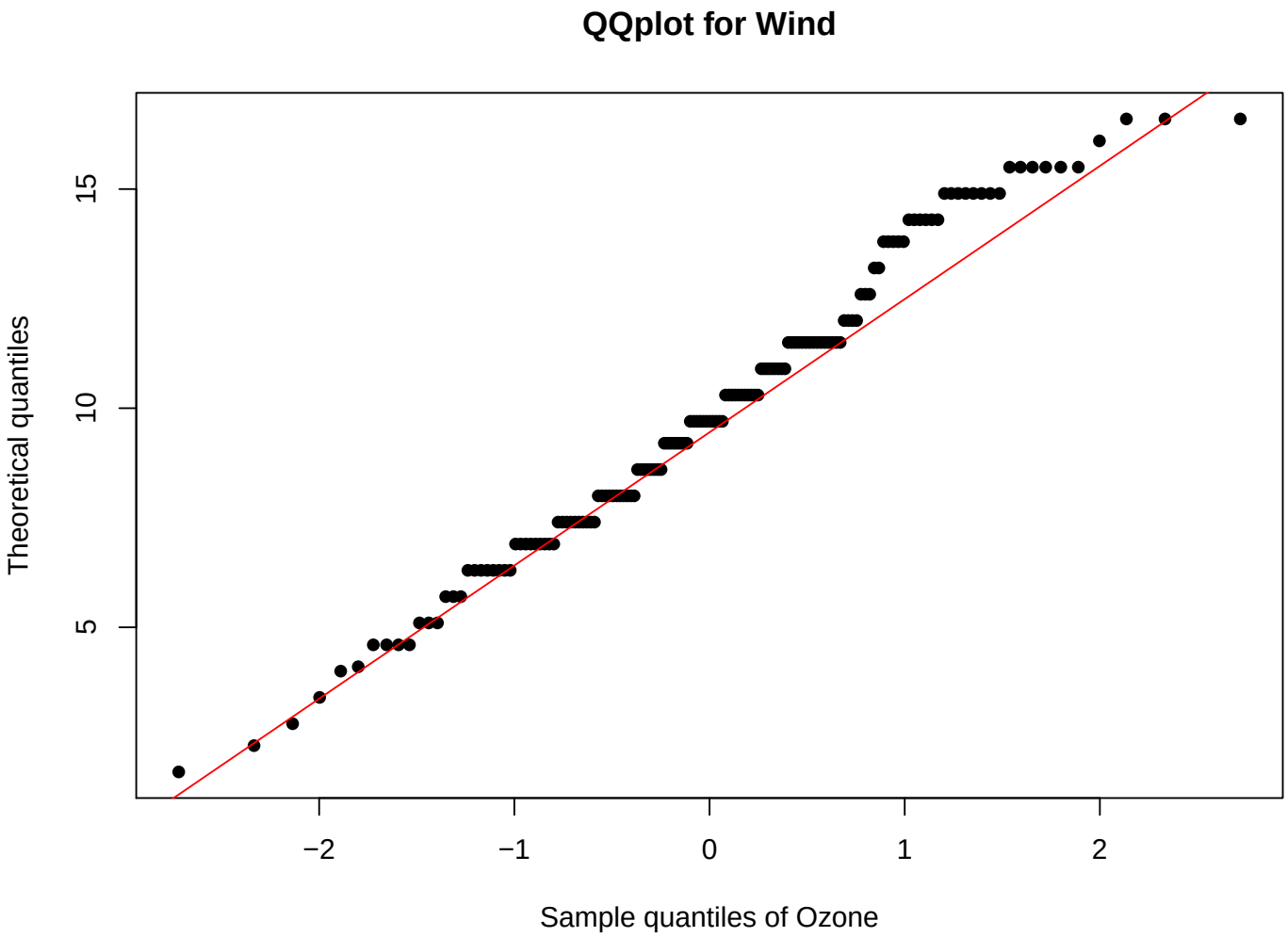
**Histogram**




```
[1] 0.05791221
```

The histogram with the density curve of Wind clearly shows that the distribution is normally distributed.

**QQPlot**




**QQPlot** The above QQ plot clearly shows that most of the values lies above the normal line but more or less close to it. So we can interpret that the data is surely from a normal distribution.

**Hypothesis testing:**

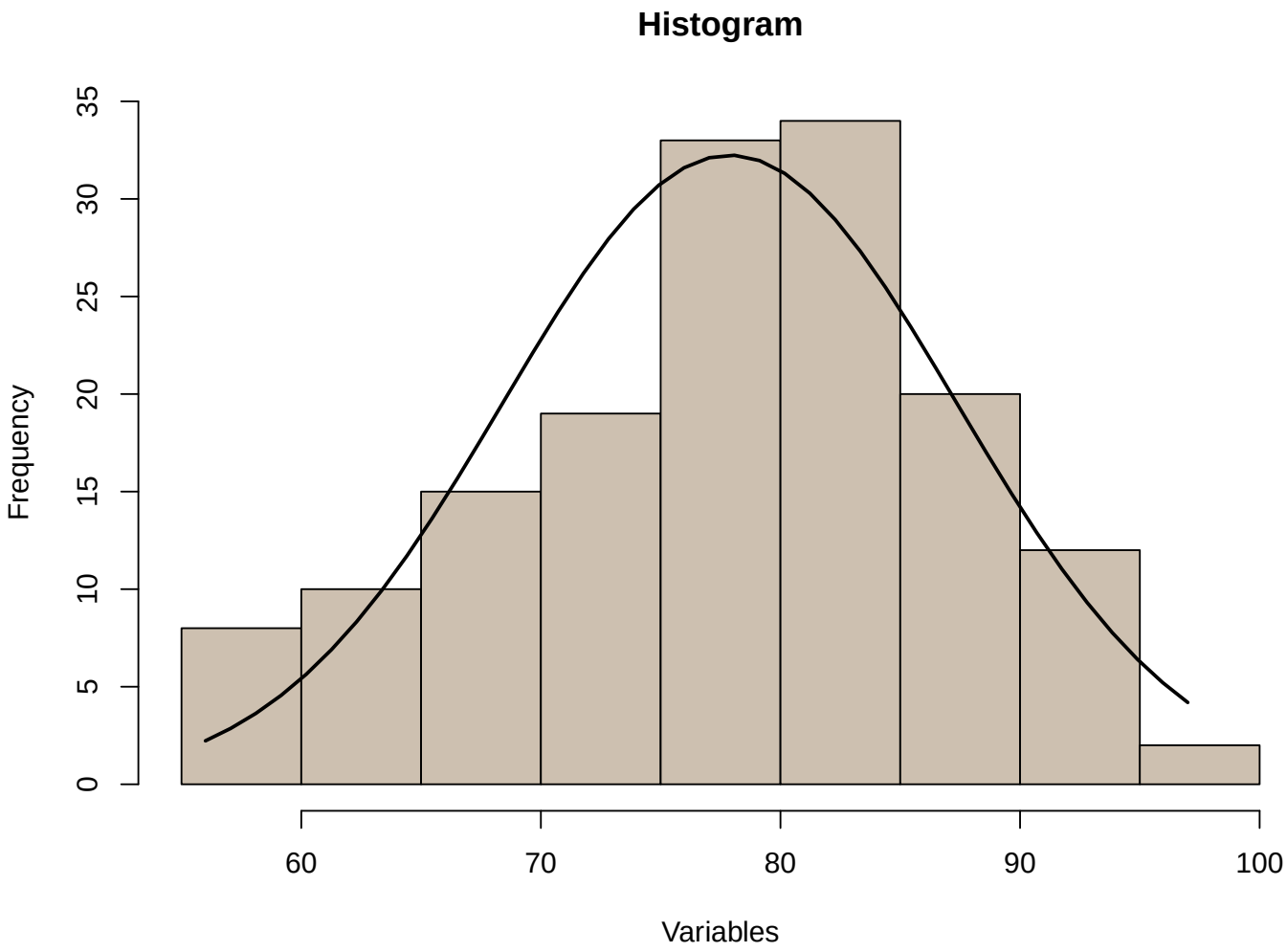
```
    Shapiro-Wilk normality test

data:  Data$Wind
W = 0.98177, p-value = 0.04044
```

If p is more than .01 $\{\alpha\}$, we can be 99% $\{(1-\alpha)100\%\}$ certain that the data are normally distributed.

**For Temp**

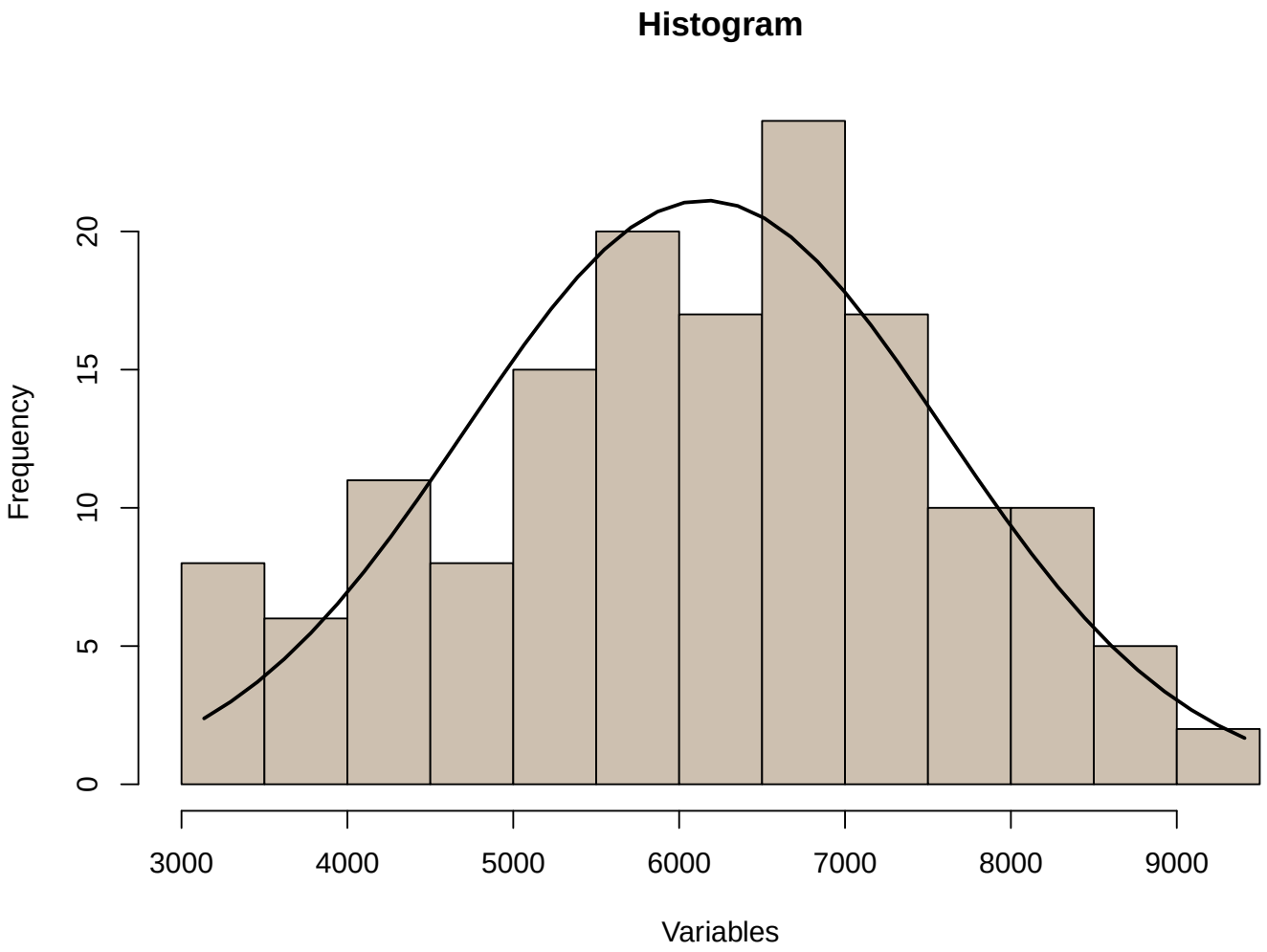
**Before Transformation Histogram**




```
[1] -0.3705073
```

The histogram with the density curve of Ozone clearly shows that tail of the distribution lie towards left and thus the variable is Left Skewed.So we need to transform the Variable.

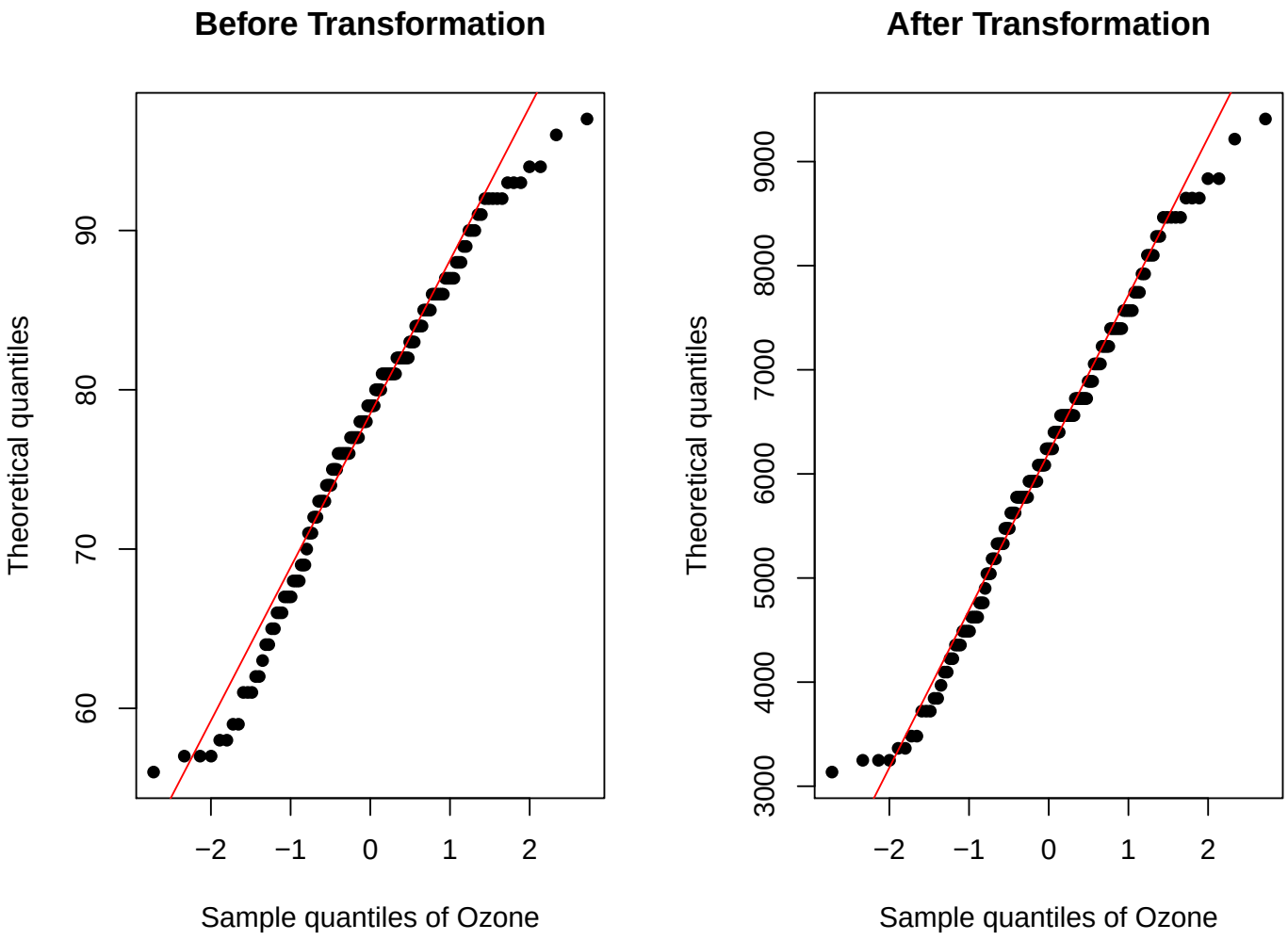
**After Transformation Histogram**




```
[1] -0.1080352
```

After **Square Transformation** the histogram with the density curve of Temp clearly shows that the variable is slightly skewed and so the data is from normal population. Also the skewness is much close to 0.

**QQPlot**




**QQPlot** The above QQ plot clearly shows that most of the values lies above the normal line but more or less close to it. So we can interpret that the data is surely from a normal distribution.

**Hypothesis testing:**

```
    Shapiro-Wilk normality test

data:  Temp_T
W = 0.98546, p-value = 0.1089
```

If p is more than .01 $\{\alpha\}$, we can be 99% $\{(1-\alpha)100\%\}$ certain that the data are normally distributed.

## Scale Transformation

This transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution.

```
##           Ozone    Solar.R        Wind        Temp Month Day
## 1   0.384942387 -0.1997429 -0.74252405 -1.15279397     5   1
## 2   0.225670508 -0.9227365 -0.56248256 -0.67179029     5   2
## 3  -1.088657976 -0.6528854  0.81783552 -0.46969953     5   3
```

```
## 4  -0.611804235  1.8173427  0.48775946 -1.59919308     5   4
## 5  -1.866521575 -0.6431372  1.32795308 -2.08919396     5   5
## 6  -0.080456709 -0.6135015  1.50799457 -1.24484216     5   6
## 7  -0.318251764  1.5380043 -0.38244107 -1.33550617     5   7
## 8  -0.547350827 -1.0571567  1.17791850 -1.85042235     5   8
## 9  -1.550728419 -1.3649245  1.68803606 -1.68432035     5   9
## 10 -1.088657976 -0.1496654 -0.38244107 -0.96454505     5  10
## 11 -1.698730572 -1.1520952 -0.89255863 -0.46969953     5  11
## 12 -0.751566582  0.7599450 -0.05236501 -0.96454505     5  12
## 13 -1.189236918  1.3651783 -0.20239958 -1.24484216     5  13
## 14 -0.908842327  1.0709731  0.30771797 -1.05936160     5  14
## 15 -0.611804235 -1.2389484  0.99787701 -1.93139707     5  15
## 16 -0.908842327  2.2603132  0.48775946 -1.42478599     5  16
## 17  0.155843325  1.6960613  0.63779403 -1.24484216     5  17
## 18 -1.866521575 -1.1783403  1.68803606 -2.01098761     5  18
## 19  0.003345403  2.0036662  0.48775946 -1.05936160     5  19
## 20 -1.189236918 -1.3135755 -0.05236501 -1.59919308     5  20
## 21 -3.440700881 -1.3746075 -0.05236501 -1.85042235     5  21
## 22 -1.189236918  1.9618045  2.01811212 -0.57143700     5  22
## 23 -2.289321135 -1.3563175 -0.05236501 -1.68432035     5  23
## 24  0.081906355 -1.1007463  0.63779403 -1.68432035     5  24
## 25 -1.550728419 -1.2346774  2.01811212 -2.01098761     5  25
## 26 -1.866521575  0.9301302  1.50799457 -1.93139707     5  26
## 27 -0.995536421  0.4700434 -0.56248256 -2.01098761     5  27
## 28 -0.318251764 -1.3711842  0.63779403 -1.15279397     5  28
## 29  0.499254427  0.6936967  1.50799457  0.28121990     5  29
## 30  1.661513600  0.2445969 -1.25264161  0.05975058     5  30
## 31  0.259180920  1.1611191 -0.74252405 -0.26207203     5  31
## 32  0.833127982  1.2900621 -0.38244107 -0.04890781     6   1
## 33 -0.751566582  1.3087433 -0.05236501 -0.46969953     6   2
## 34 -0.080456709  0.5326402  1.86807755 -1.15279397     6   3
## 35 -0.486042767 -0.2487771 -0.20239958  0.62380526     6   4
## 36  0.604028919  0.2012682 -0.38244107  0.74076875     6   5
## 37  0.003345403  0.8955715  1.32795308  0.05975058     6   6
## 38 -0.037857173 -0.8508479 -0.05236501  0.39403084     6   7
## 39  1.661513600  1.0531395 -0.89255863  0.97884828     6   8
## 40  1.062227044  1.3841204  1.17791850  1.34634894     6   9
## 41  0.323634328  2.0246948  0.48775946  0.97884828     6  10
## 42  1.210229198  0.8103159  0.30771797  1.72630725     6  11
## 43  1.210229198  0.6609638 -0.20239958  1.59827029     6  12
## 44 -0.318251764 -0.6625683 -0.56248256  0.39403084     6  13
## 45 -0.995536421  2.2168867  1.17791850  0.16979315     6  14
## 46 -1.698730572  2.0036662  0.48775946  0.05975058     6  15
## 47 -0.427587299 -0.1873213  1.50799457 -0.15618201     6  16
## 48  0.259180920  1.2528952  1.68803606 -0.67179029     6  17
## 49 -0.486042767 -1.3320612 -0.20239958 -1.33550617     6  18
## 50 -1.088657976 -0.9072177  0.48775946 -0.57143700     6  19
```

Now that the Data has been transformed and made consistent these are very important steps in Exploratory Data Analysis before Model Fitting. The quality and effort invested in data exploration can make a diffrence

in building a good model from bad model.