Data Visualization with R Part I

Md Mominul Islam

Descriptive Statistics

It's difficult to discover trends and explain them to other experts when a dataset is very large. The goal of descriptive statistics is to properly summarize the dataset and uncover trends. Descriptive statistics, in particular, are used to effectively convey data by Summary statistics, contingency tables, and graphs are used to convey information.

Summary statistics are used to more succinctly convey the qualities of a sample of continuous data, whereas contingency tables are used for samples of nominal or ordinal data for the same reason. Graphs are used to visualize data attributes and are employed in the vast majority of statistical approaches.

Measures that summarize data and may be calculated for quantitative variables comprise summary statistics. These measurements are divided into three categories based on how they convey a sample's properties:

a. measures of central tendency or measures of location
b. measures of dispersion, and
c. measures of shape

Measures of central tendency give information concerning the location of the center of the distribution of the data if we arrange it along an axis. Measures of dispersion, as their name suggests, show how dispersed the data is around the midpoint along this axis. Measures of shape relate to the shape of the distribution of the data, that is, they show how the data is arranged around a central value.

If we organize the data along an axis, measures of central tendency provide information about the location of the center of the distribution of the data. Measure of dispersion, as the name implies, reveal how spread the data is around the axis's midpoint. Measures of shape tells us how the data is grouped around a central value.

```
#Summary Statistics Table
library(knitr)
summary_st <-
  data.frame(
              'Measures of central tendency' = c("Mean", "Median","Trimmed Mean","Mode"),
              'Measures of Dispersion' = c("Variance", "Standard Derviation",
                                           "Range", "Interqaurtile Range"),
              'Measures of Shape' = c("Skewness","Kurtosis", "NA", "NA"))

kable(summary_st)
```

Measures.of.central.tendency	Measures.of.Dispersion	Measures.of.Shape
Mean	Variance	Skewness
Median	Standard Derviation	Kurtosis
Trimmed Mean	Range	NA

Measures.of.central.tendency	Measures.of.Dispersion	Measures.of.Shape
Mode	Interqaurtile Range	NA

We made a simple table above. Measures of central tendency is measured by means of mean, median and mode from the data set. With the help of variance, standard deviation, Interquartile range, and range we can determine the measures of dispersion. Lastly with skewness and kurtosis we can have measures of shape.

Measures of shape

Sample values are distributed around the mean value. This distribution may be symmetric or asymmetric. The measures used to describe the shape of distribution of the data, that is, how the data is arranged around a central value, are skewness and kurtosis. Skewness shows how symmetric the distribution of the data around a central value (mean or median) is, while kurtosis shows how pointy this distribution is.

- When skewness = 0, the distribution of values around the mean is symmetric.
- When skewness < 0, the distribution is asymmetric and extends more to the left, which means that most values in the sample are smaller than the mean, while when skewness > 0, the distribution is asymmetric and extends more to the right.
- When kurtosis = 3, the distribution is neither too narrow nor too broad, when kurtosis < 3, the distribution is too broad, and when kurtosis > 3, the distribution is too narrow. In a normal distribution the value of skewness is 0 and kurtosis is equal to 3

Functions for Summary Statistics

summary(x), describe(x) and describeBy(x,group)

- The function summary(x) estimates the minimum and maximum value, first and third quartile, median, and mean.
- The functions describe(x) and describeBy(x, group) belong to the *psych* package and, therefore, in order to use them, we should have first installed and subsequently loaded *packagepsych*. The function describe(x) estimates the minimum and maximum value, the range, median, mean, trimmed mean, standard deviation, skewness and kurtosis.
- If in the dataset there is a grouping variable and we want to estimate the descriptive statistics per group, we can use the function describeBy(x, group), where group is the grouping variable.
- You can also use kable() in the knitr package on some of the variables produced by summary(model) to create nice looking output of the information.

```
#csv data set reading
water <- read.csv('https://umich.instructure.com/files/399172/download?download_frd=1',
                  header=T)
colnames(water) = c("Year","Region","Country", "Residence Area",
                    "Drirnking Water Population","Sanitation Population")

#Summary Statictics
summary(water)
```

```
##       Year         Region            Country          Residence Area
##  Min.   :1990   Length:3331        Length:3331        Length:3331
##  1st Qu.:1995   Class :character   Class :character   Class :character
##  Median :2005   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2002
##  3rd Qu.:2010
##  Max.   :2012
##
##  Drirnking Water Population Sanitation Population
##  Min.   :  3.0              Min.   :  0.00
##  1st Qu.: 77.0              1st Qu.: 42.00
##  Median : 93.0              Median : 81.00
##  Mean   : 84.9              Mean   : 68.87
##  3rd Qu.: 99.0              3rd Qu.: 97.00
##  Max.   :100.0              Max.   :100.00
##  NA's   :32                 NA's   :135
```

```
library ("psych")
# We have grouped our region based on Africa only then used describeBy() function
describeBy(water$`Sanitation Population`,water$Region=='Africa')
```

```
##
##  Descriptive statistics by group
## group: FALSE
##    vars    n  mean   sd median trimmed   mad min max range  skew kurtosis   se
## X1    1 2411 80.23 23.9     91   84.63 13.34   0 100   100 -1.31     0.74 0.49
## ------------------------------------------------------------
## group: TRUE
##    vars   n  mean    sd median trimmed   mad min max range skew kurtosis  se
## X1    1 785 33.99 25.21     29   30.75 23.72   0  99    99 0.94     0.11 0.9
```

Graphical Visualization Using R

Boxplot consist of a rectangle with two antennas, one at the lower base and the other at the upper one. The antennas are T and inverse T-shaped, respectively. The lower base of the rectangle lies on the first quartile (Q1) and the upper base marks the third quartile (Q3). The median is represented by a horizontal line in the interior of the rectangle. The antennas are called whiskers and extend up to Q3 + 1.5(Q3-Q1) and Q1-1.5(Q3-Q1). If the maximum and minimum sample values lie within this range, then the whiskers shift to the maximum and minimum values. If there are outliers, these appear as points beyond the whiskers.

I want to look at the insect spray data; according to the description, it's "The counts of insects in agricultural experimental units treated with different insecticides."

```
library("datasets")
#better to rename it different than the library for avoiding accidents
## choosing the data
bug_dat <- InsectSprays
str(bug_dat)
```

```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
#using character instead of factors
#looking at the data
head(bug_dat)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Basic plotting and ggplot plotting

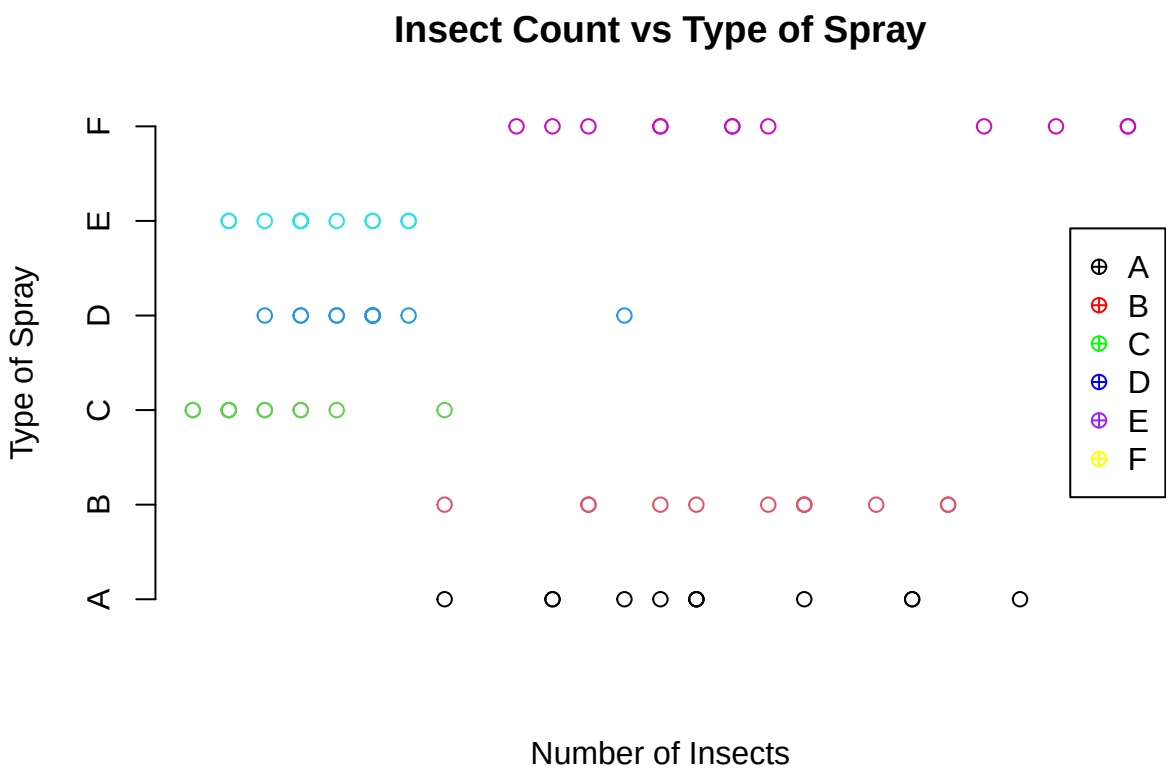
The basic version of R includes the plot() function, which can create a wide variety of graphs. For bar plots, pie plots, boxplots, and histograms we may use the functions barplot(), pie(), boxplot(), and hist(), respectively. Details for the use of these functions may be found in many websites as well as via the commands ?barplot, ?pie, ?boxplot, and ?hist

```r
library(ggplot2)
```

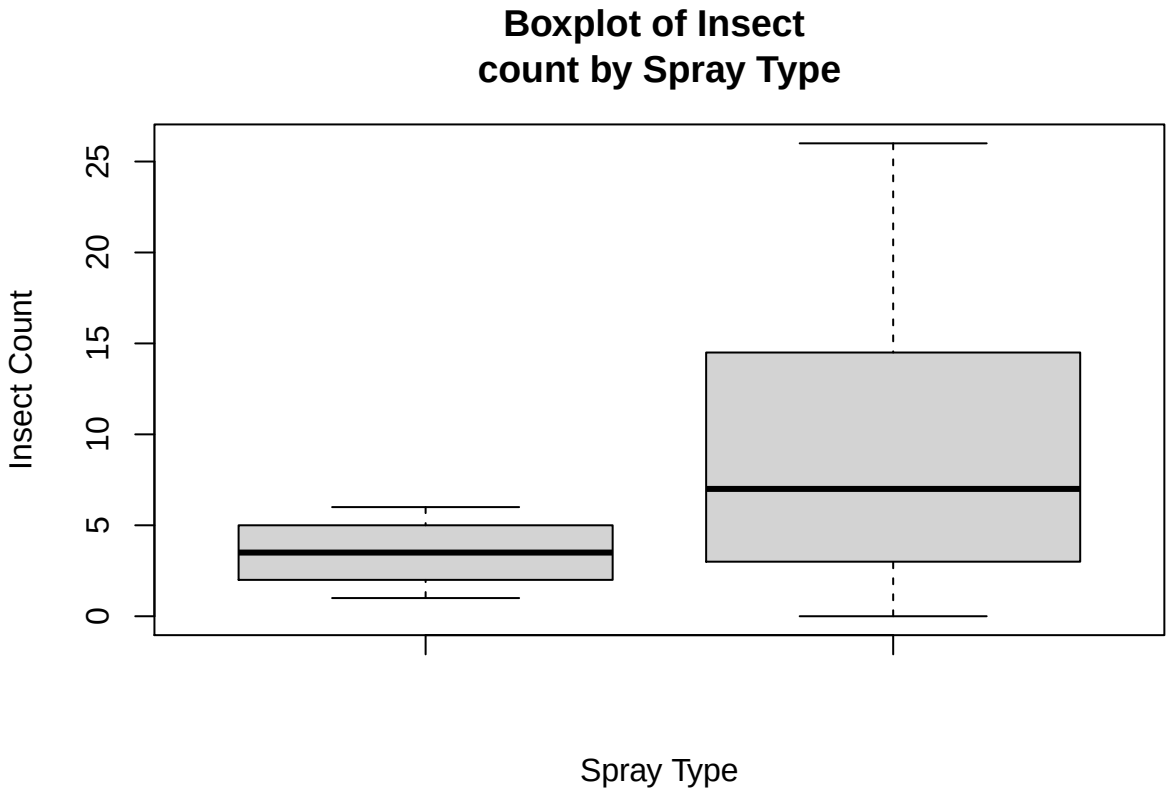
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
plot(bug_dat[,1],bug_dat[,2],
     main = "Insect Count vs Type of Spray",
     xlab = "Number of Insects",
     ylab = "Type of Spray",
     col = bug_dat[,2],
     axes = F) #y-axis removed

#Adding tick mark based on spray
axis (side = 2,
      at = c(1,2,3,4,5,6),
      labels = c("A","B","C","D","E","F"))

legend (x = "right",
        legend = unique(bug_dat[,2]),
        col = c("black","red","green","blue","purple","yellow"),
        pch= 10)
```

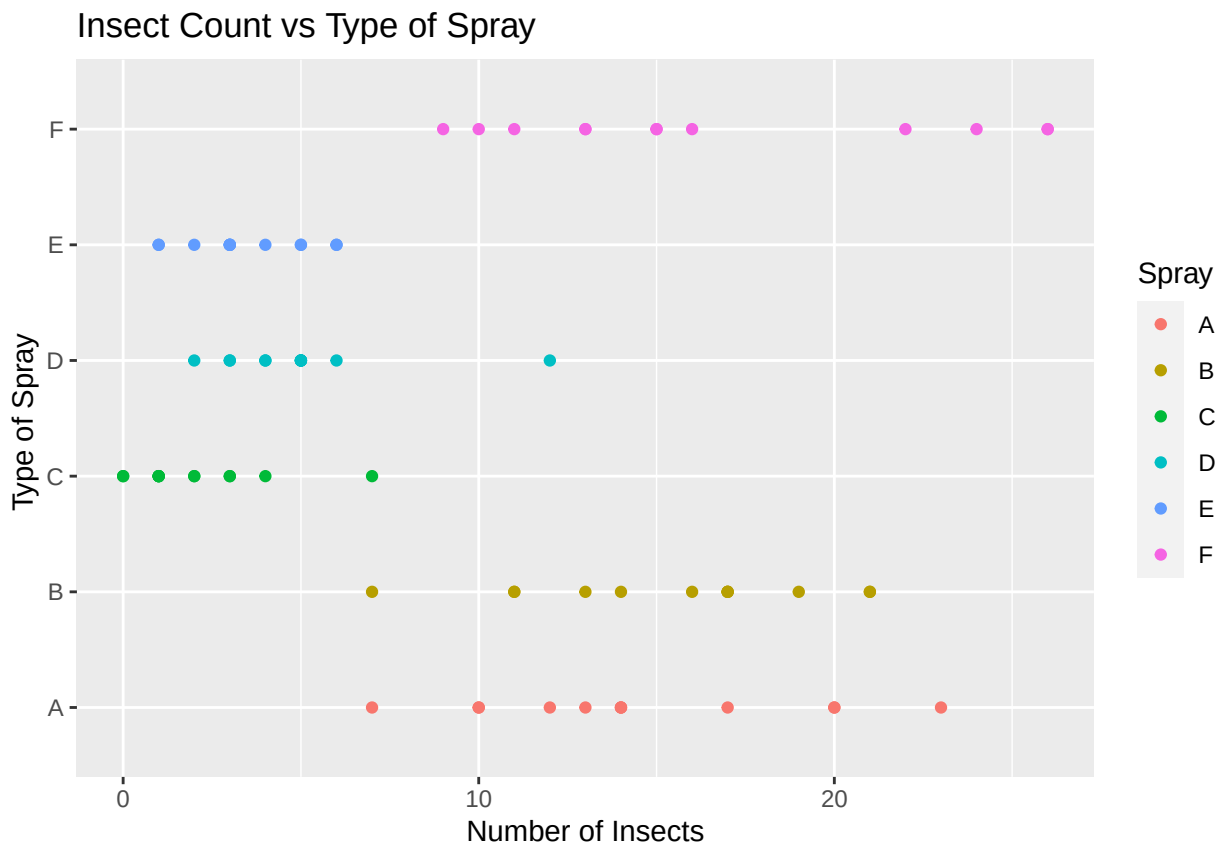



```
# to wrap long title we can use  \n command

boxplot(bug_dat[,2],bug_dat[,1],
main = "Boxplot of Insect \ncount by Spray Type",
xlab = "Spray Type",
ylab = "Insect Count")
```

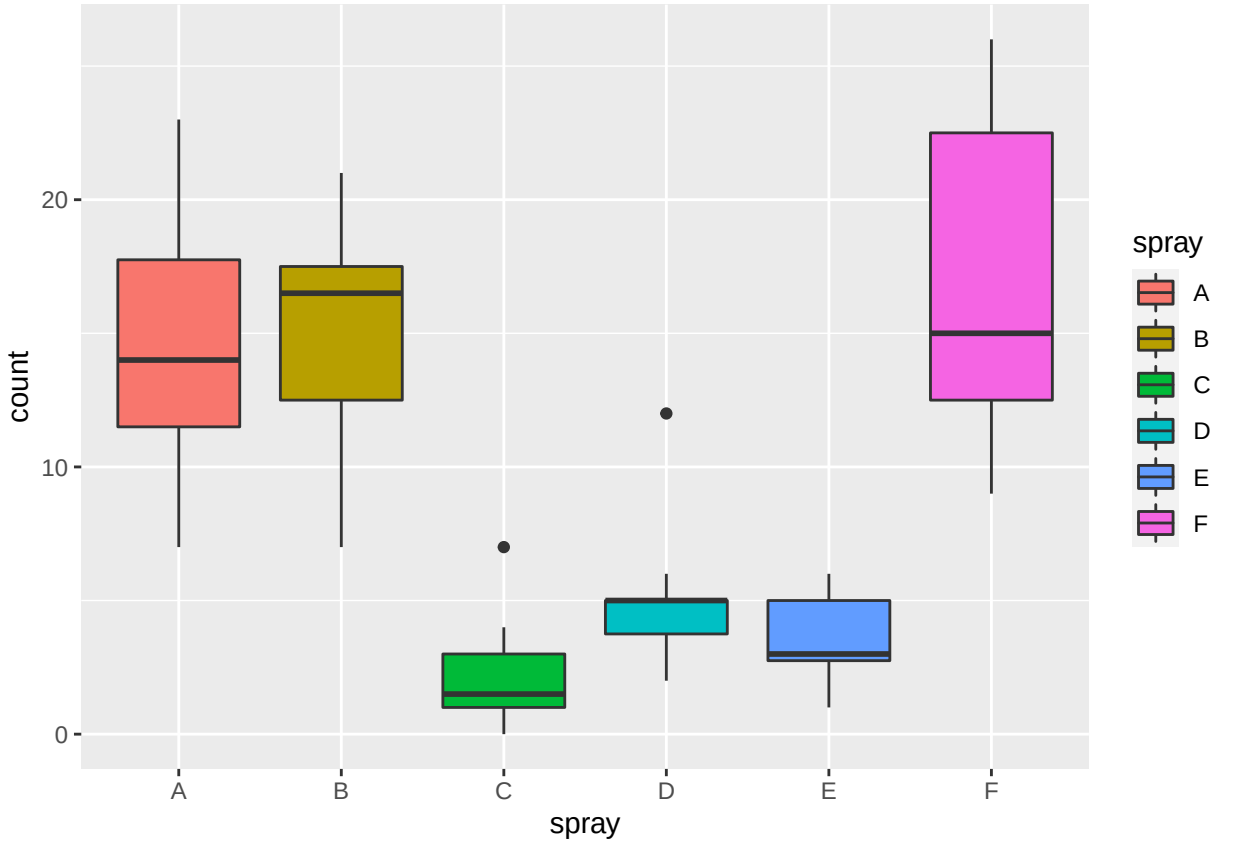


Box plots are used to display general responses for a group. They're a great strategy to see the range and other features of a large group's responses. We can say from the above figure that medians are same for these two although their distribution are different.

```
#convert the data into a data frame
bug_dat2 <- data.frame(bug_dat)
ggplot(data = bug_dat2) +
  geom_point(aes(x = count, y  = spray, color  = spray))+
  ggtitle("Insect Count vs Type of Spray")+
  xlab("Number of Insects") +
  ylab("Type of Spray ") +
  guides(color = guide_legend(title = "Spray"))
```




```
ggplot(data = bug_dat2) +
  geom_boxplot(aes(x = spray, y  = count, fill  = spray))
```

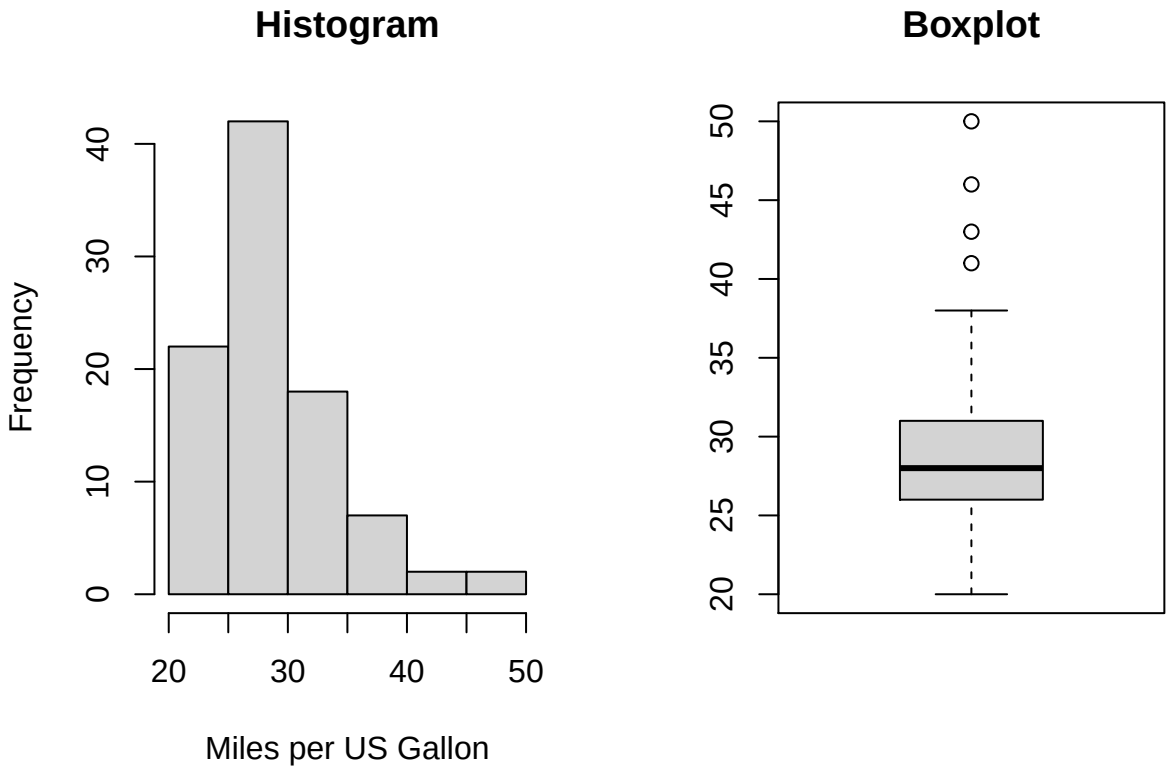


You can read more about ggplot tool in here: http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/

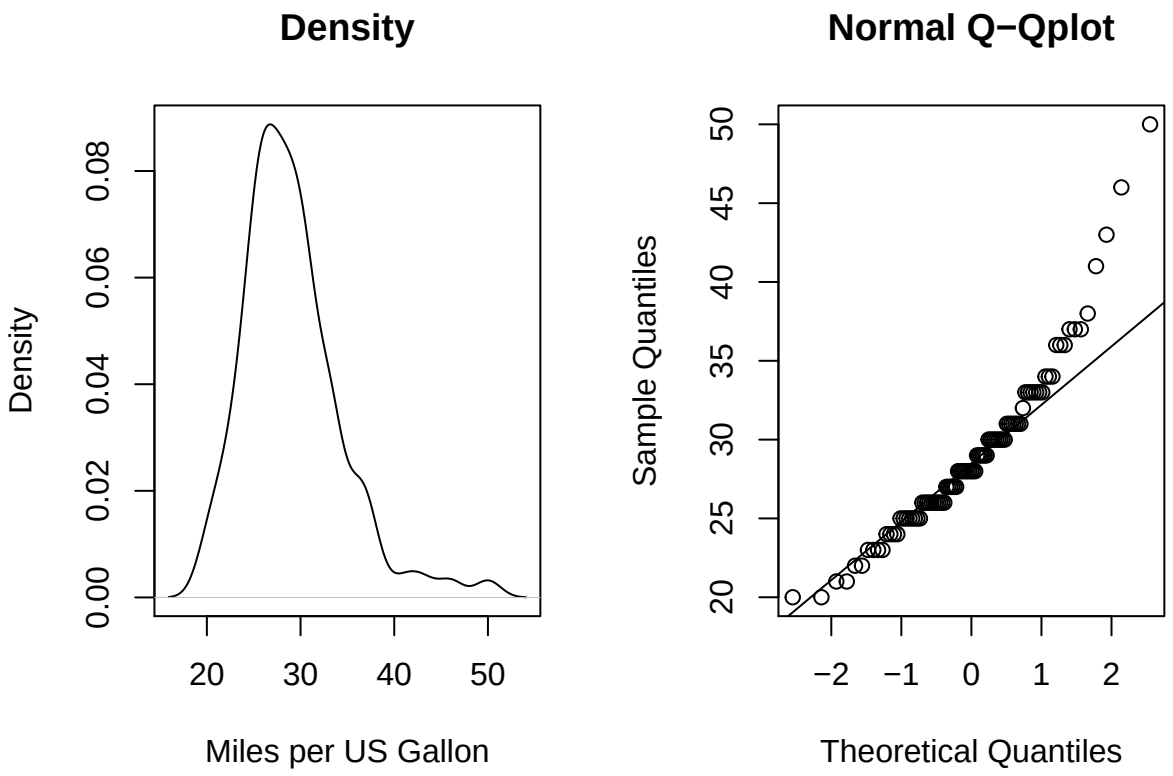
Exploratory Data Analysis for Univariate Data

The Cars93 dataset will be used to illustrate some of these plots. The following script sets up a 2×2 plotting region and produces a histogram, boxplot, density plot and Normal scores plot of the MPG.highway vector.

```
data(Cars93,package="MASS")
par(mfrow=c(1,2))
# Histogram
hist(Cars93$MPG.highway,
     xlab="Miles per US Gallon",
     main="Histogram")
# Boxplot
boxplot(Cars93$MPG.highway,
        main="Boxplot")
```




```
# Density
plot(density(Cars93$MPG.highway),type="l",
     xlab="Miles per US Gallon",
     main="Density")
# Q-Q Plot
qqnorm(Cars93$MPG.highway,main="Normal Q-Qplot")
qqline(Cars93$MPG.highway)
```



Results shows a distribution that is skewed heavily towards the right. This is visible in all four plots.

Boxplot Interpretation

If the distance between the minimum value and the first quartile exceeds 1.5 × IQR then the whisker extends from the lower quartile to the smallest value within 1.5 × IQR. Extreme points, representing those beyond this limit are indicated by points. A similar procedure is adopted for distances between the maximum value and the third quartile.

Let's look at the summary statistics to interpret the boxplot in a more precise way.

```
summary(Cars93$MPG.highway)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   26.00   28.00   29.09   31.00   50.00
```

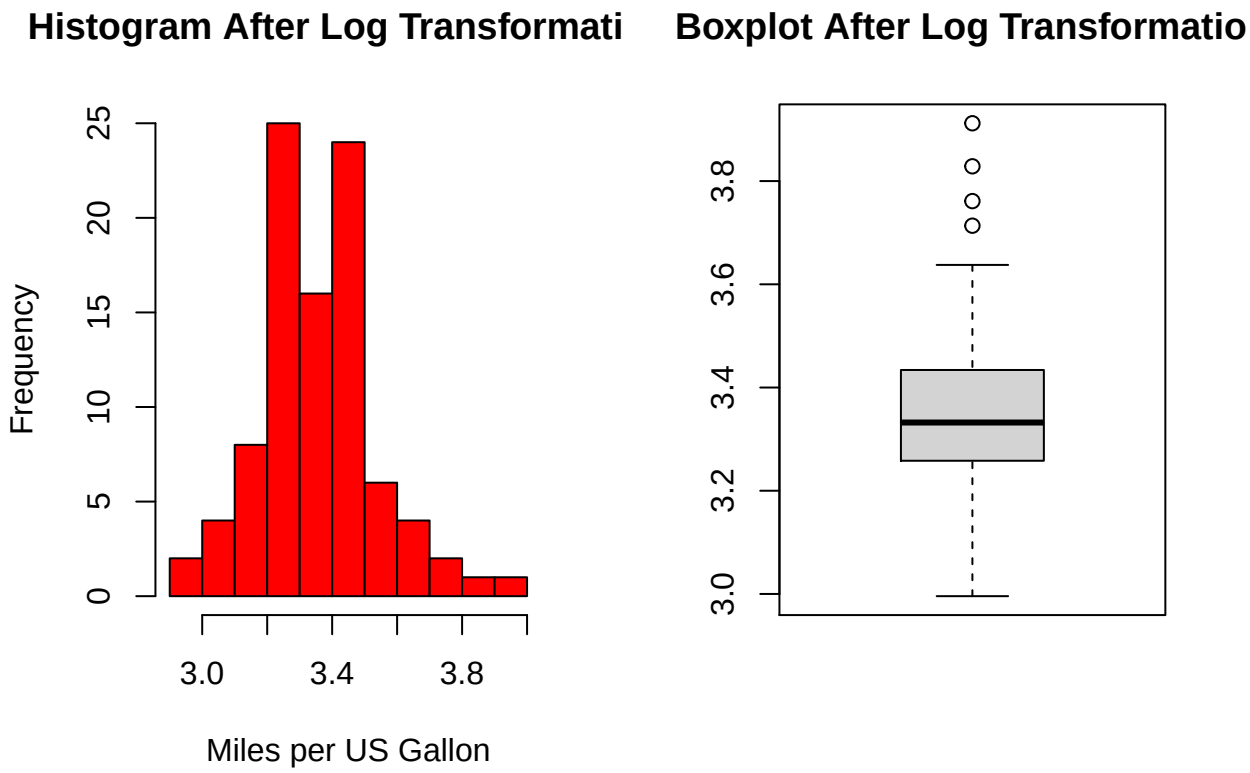
Here, $IQR = Q_3 - Q1 = 31 - 26 = 5$

So we can say that distance between the minimum value and the first quartile does not exceed 1.5 × IQR. One way to define an outlier is

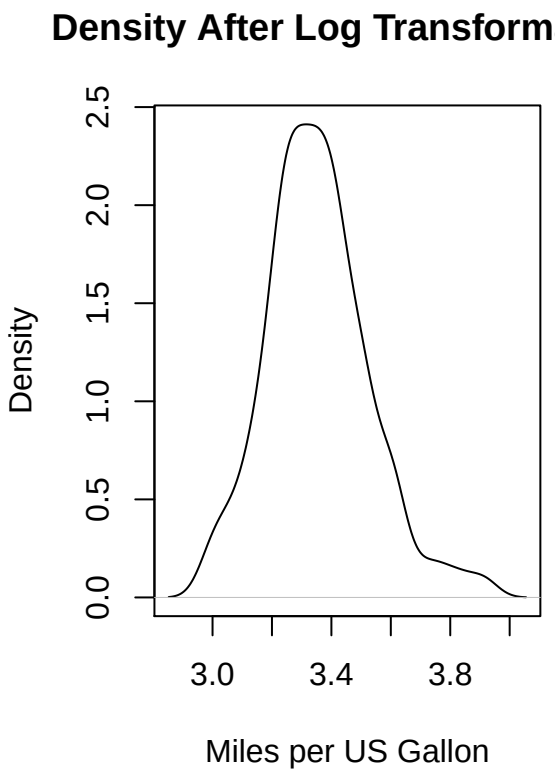
- Anything below Q1 – 1.5 IQR
- Anything above Q3 + 1.5 IQR

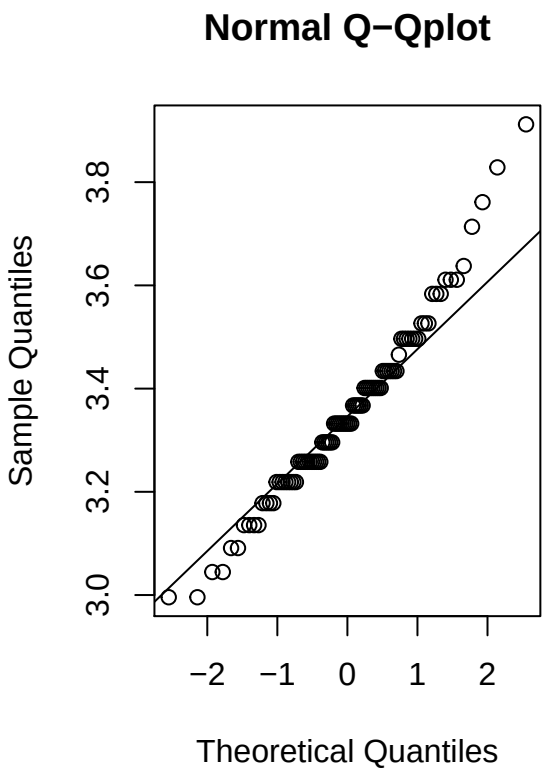
So here 31+7.5 = 38.5. So, the points above this value are termed as outliers.

```r
#log transfromation to remove skewness in the data set
par(mfrow=c(1,2))
data(Cars93,package="MASS")
a<- Cars93$MPG.highway
# Histogram
hist(log(a),
     col="red",
     xlab="Miles per US Gallon",
     main="Histogram After Log Transformation")
# Boxplot
boxplot(log(a),
        main="Boxplot After Log Transformation")
```




```r
# Density
plot(density(log(a)),type="l",
     xlab="Miles per US Gallon",
     main="Density After Log Transformation")
# Q-Q Plot
qqnorm(log(a),main="Normal Q-Qplot")
qqline(log(a))
```

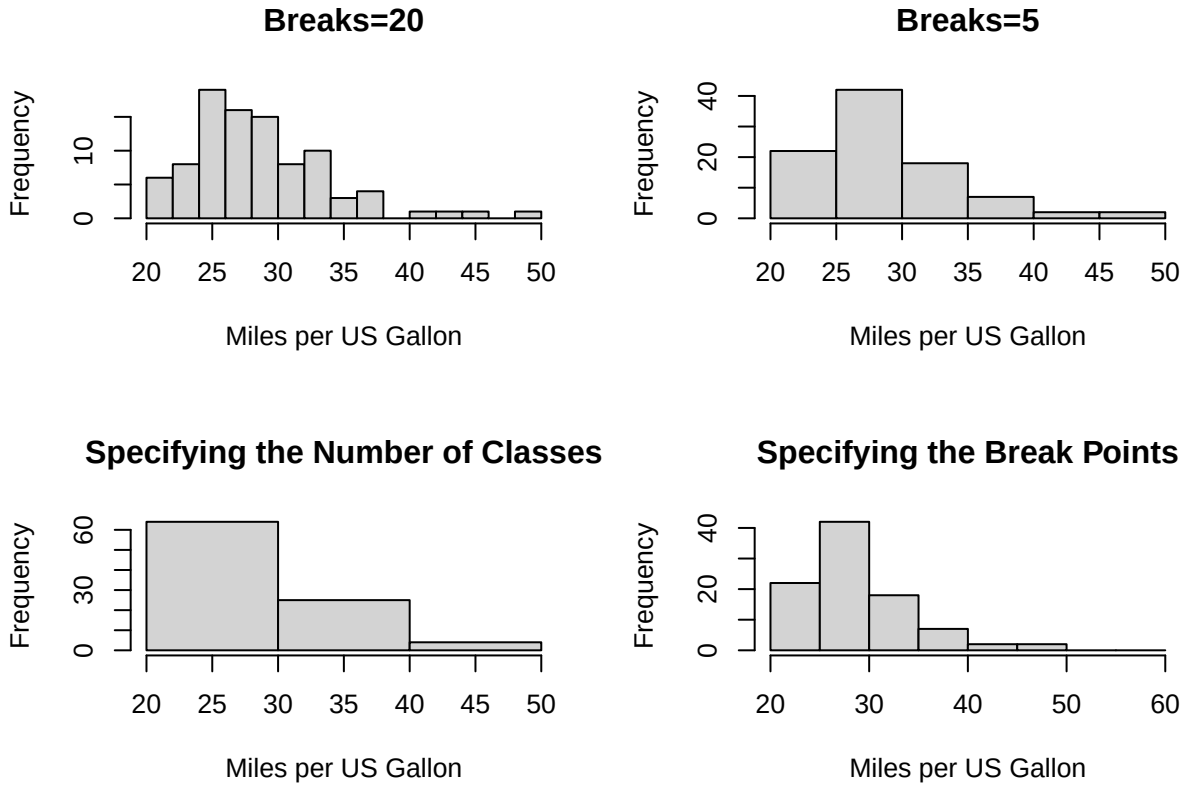




Because of its numerous advantages, the normal distribution is a probability and statistical concept that is commonly employed in scientific investigations. To highlight a few of these advantages, normal distribution is straightforward. It has the same mean, median, and mode values, and it can be defined using only two parameters: mean and variance. The Central Limit Theorem, for example, is one of the most fundamental mathematical implications. When our continuous data does not follow the bell curve, we may log convert it to make it as "normal" as feasible, increasing the validity of the statistical analysis results.After log transformation, we can see that skewness is removed. Now we can say that data is normally distributed.

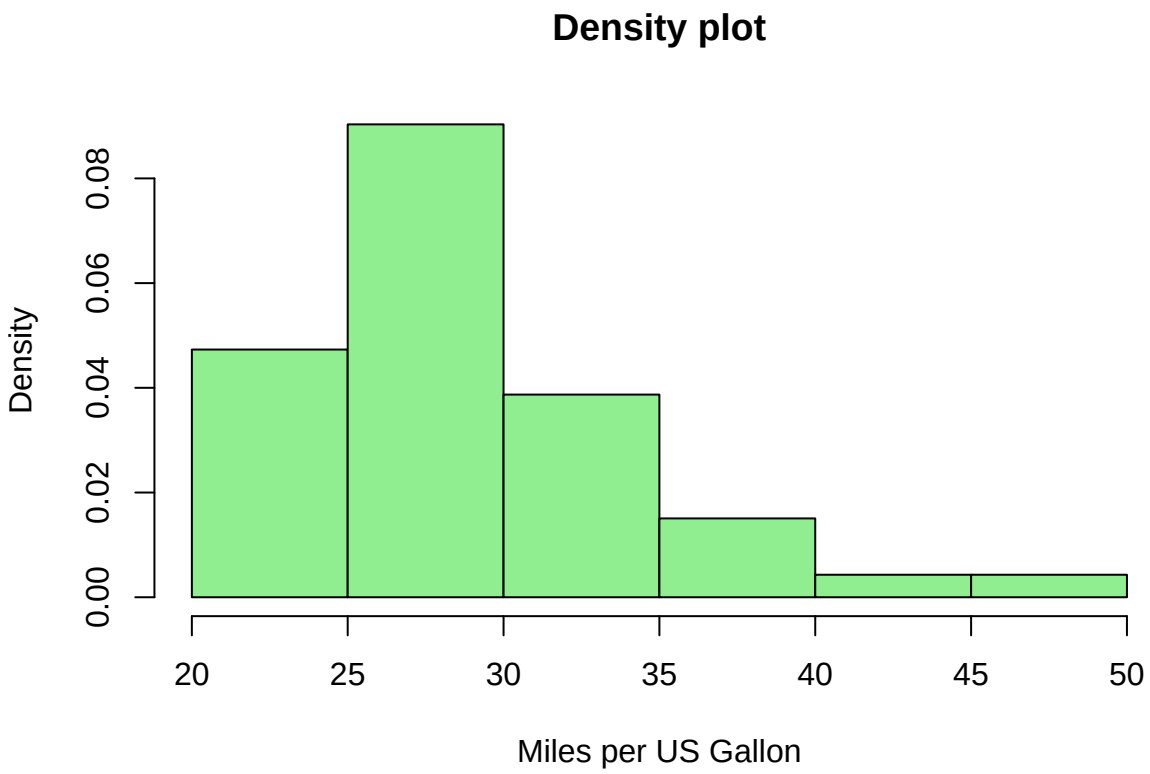
Histogram Using Breaks You can use the breaks() option to change Histogram in a number of ways. An easy way is just to give it one number that gives the number of cells for the histogram:

```
par(mfrow=c(2,2))
hist(a, breaks=20, xlab="Miles per US Gallon",
     main="Breaks=20")
hist(a, xlab="Miles per US Gallon",
     breaks=5, main="Breaks=5")
hist(a,nclass=4,xlab="Miles per US Gallon",
     main="Specifying the Number of Classes")
hist(a,xlab="Miles per US Gallon",
     breaks=seq(from=20,to=60,by=5),
     main="Specifying the Break Points")
```



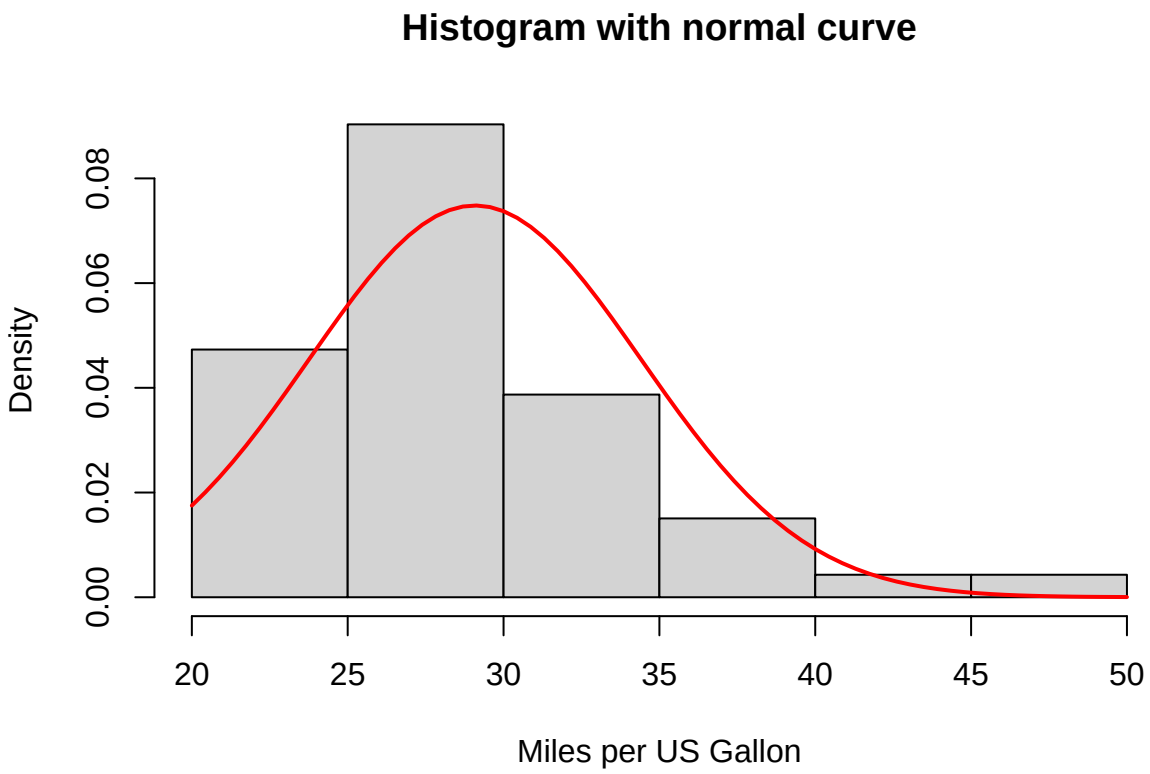
Because frequency is proportional to sample size, we are often more interested in density than frequency. Using the freq=FALSE option, R may report the probability densities instead of counting the number of datapoints each bin.

```
hist(a, xlab="Miles per US Gallon",
     col="lightgreen",
     freq=FALSE, main="Density plot")
```



Histograms with Normal Curve

```
hist(a, prob = TRUE, xlab="Miles per US Gallon",
     main = "Histogram with normal curve")
x <- seq(min(a), max(a), length = 70)
f <- dnorm(x, mean = mean(a), sd = sd(a))
lines(x, f, col = "red", lwd = 2)
```



A normal curve is drawn in histogram to have an idea about the distribution.