v2 Basics of Data Visualisation in R

Other Basic Graphs with R

Contents

Box – Whisker Plot

Box and Whisker plot summarizes data graphically using 5 measures:

- Minimum
- The Three Quartiles : Q1, Q2 (i.e. Median) and Q3
- Maximum

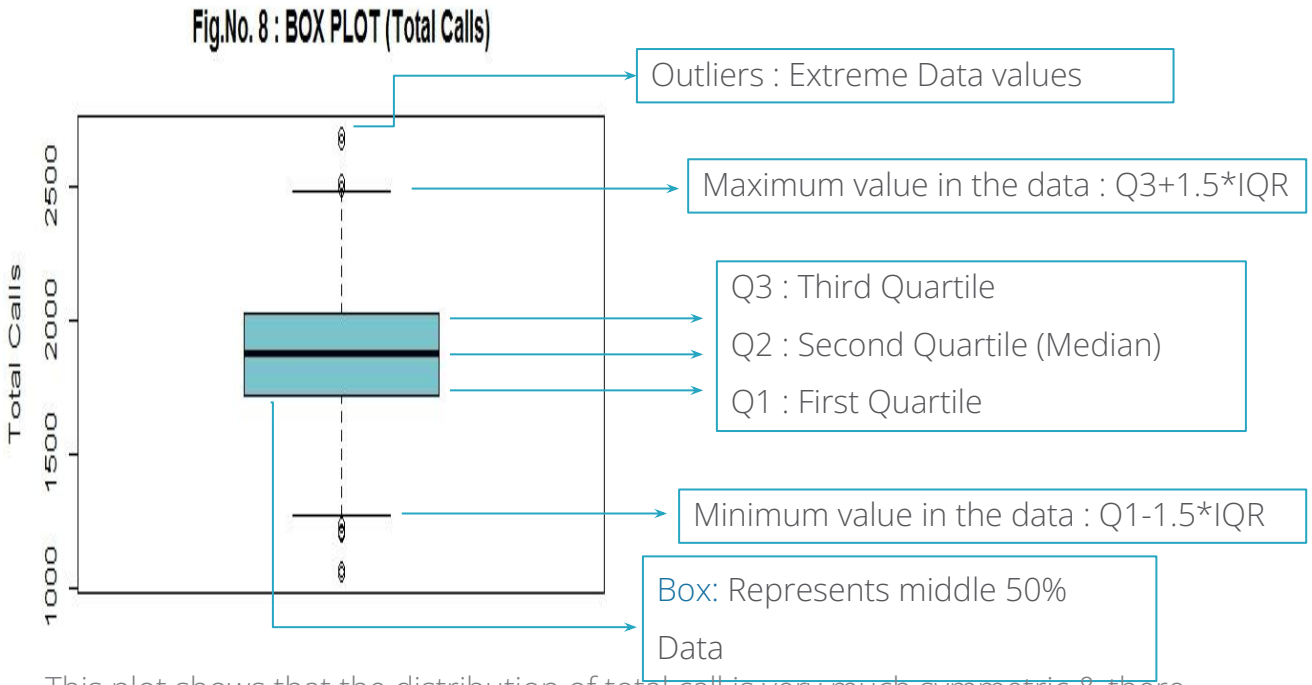
Describing a Box-Plot :

- The rectangle (box) in the middle represents the middle 50% of the data (between the values that are ¼ and ¾ of the way through the data).
- The lines (whiskers) extend from the box to the smallest and largest values.
- The diagram also shows the middle value (i.e. The Median).
- The outliers which are plotted outside the plot (The observations which are outside 1.5 times the interquartile range above the upper quartile and below the lower quartile)

Advantages of a Box Plot :

- A boxplot is particularly effective when comparing two sets of data
- It shows us the shape of the data.

Box – Whisker Plot

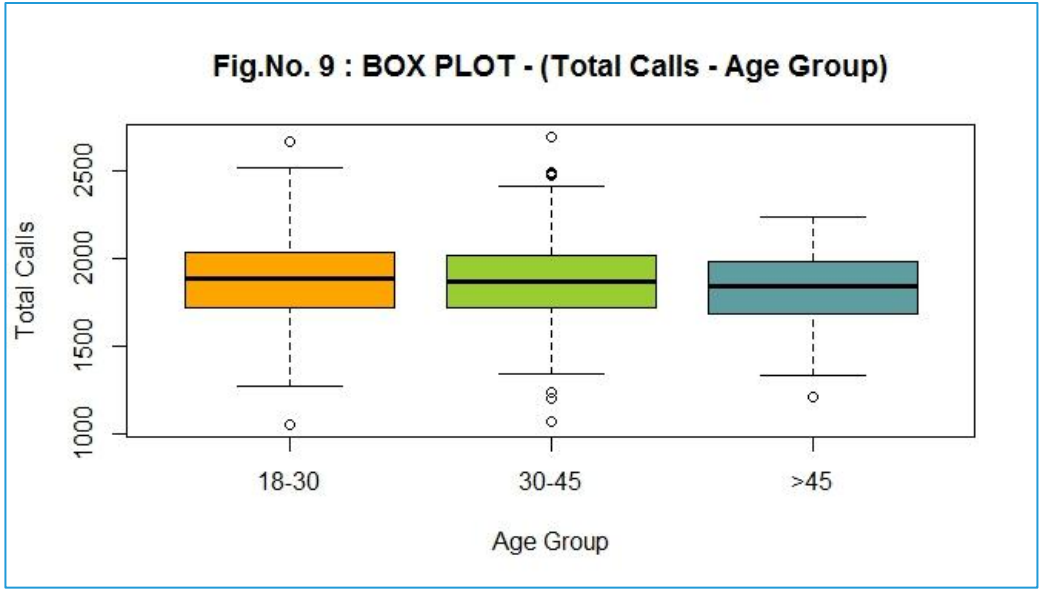


This plot shows that the distribution of total call is very much symmetric & there exists few outliers in the data.

* The minimum and maximum values are the ones excluding the outliers

Box – Whisker Plot

Here, plotting box plots for each categories of a variable gives us a good comparison of how 'total calls' is distributed for various age groups.



- We can observe that the distribution is almost symmetric amongst various age groups, but variability is least in >45 age group.

- Also, the age group 30-45 has many outliers.

Case Study

To get a better understanding of the subject, we shall consider the below case as an example.

Background

A telecom service provider has the Demographic and Transactional information of their customers

Objective

To visualize the data using usage variables and customer demographic information for generating business insights.

Sample Size

1000

* Here we continue to use previous data for our further analysis.

Data Snapshot

telecom data

Variables

CustID	Age	Gender	PinCode	Active	Calls	Minutes	Amt	AvgTime	Age_Group
1001	29	F	186904	Yes	2247	18214	3168.76	8.105919	18-30

Observations

Columns	Description	Type	Measurement	Possible values
CustID	Customer ID	Numeric	-	-
Age	Age of the Customer	Numeric	-	-
Gender	Gender of the Customer	Categorical	M, F	2
PinCode	Pincode of area	Numeric	-	-
Active	Age of the Customer	Categorical	Yes, No	2
Calls	Number of Calls made	Numeric	-	positive values
Minutes	Number of minutes spoken	Numeric	minutes	positive values
Amt	Amount charged	Continuous	Rs.	positive values
AvgTime	Mean Time per call	Continuous	minutes	positive values
Age_Group	Age Group of the Customer	Categorical	18-30, 30-45, >45	3

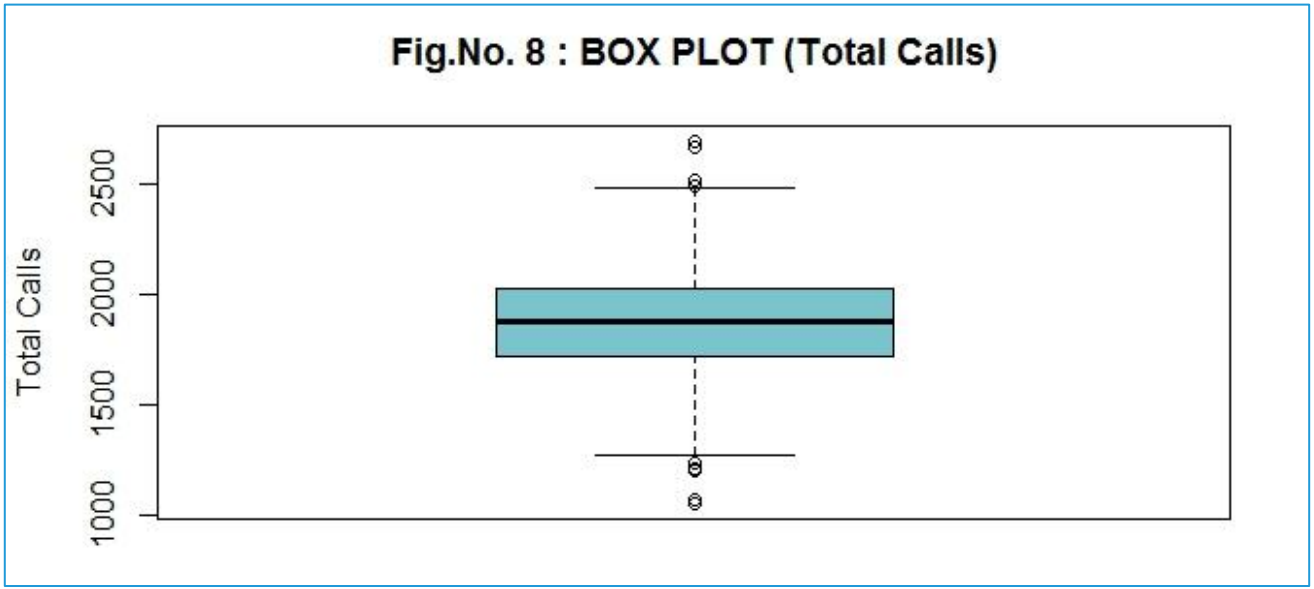
Box Plot in R

```
#Importing Data

telecom<-read.csv("telecom.csv", header=TRUE)

#BoxPlot – Total Calls

boxplot(telecom$Calls, data= telecom, main="Fig.No.8 : BOX PLOT (Total
Call$)", ylab= "Total Calls", col= "cadetblue3")
```

- **boxplot()** in base R yields a different types of box chart
- **telecom$Calls** specfies vector (variable) for which the box plot needs to be plotted
- **data=** calls the data out of which the variable needs to be plotted
- **main=** provides the user defined name of the chart. It has to be put in double quotes
- **ylab=** provides a user defined label for the variable on Y axis
- **col=** can be used to input your choice of color to the bodies of the box plots.

Box Plot in R

```
# Output
```



Interpretation :

- While we see a few outliers , the data of the number of calls overall is symmetric

Box Plot in R

```
#BoxPlot for different categories of Age_Group

boxplot(Calls~Age_Group,data=telecom,
        main="Fig.No. 9 : BOX PLOT – (Total Calls – Age Group)",
        xlab="Age Group",ylab="Total Calls",
        col=c("orange","green","cadetblue"))
```

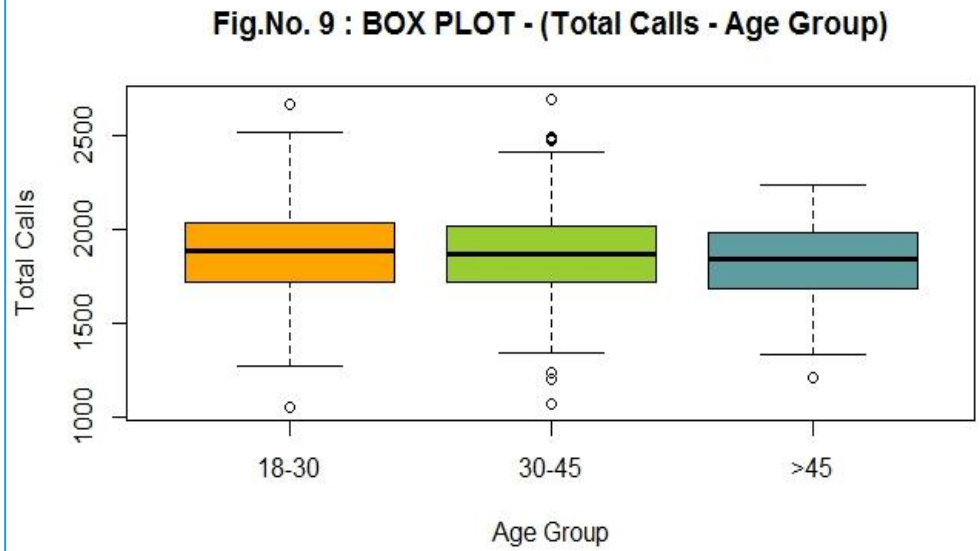
Difference between previous boxplot & this boxplot code is,

- **Calls~Age_Group** specfies vector (variable) for which the box plot for different categories of a variable is to be plotted.
- **xlab=** provides a user defined label for the variable on X axis .
- **col=** can be used to input your choice of color to the bodies of the box plots. Here we have mentioned 3 colors as the variable has 3 categories.

Note : Re – order the levels of Variable Age_Group as explained in previous ppt before you execute the boxplot code, Age_Group wise.

Box Plot in R

```
# Output
```

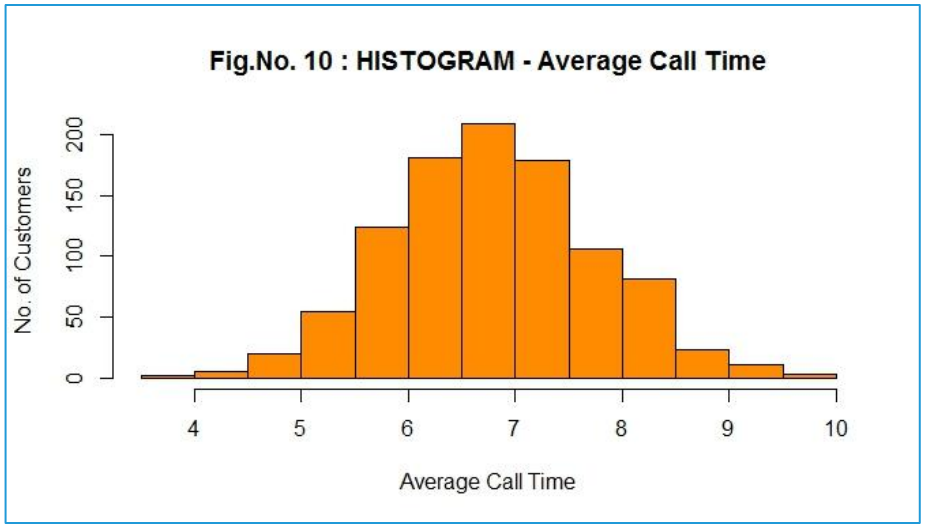


Interpretation :

- Here we can observe that the spread of total calls is higher in the age group 18-30.
- The number of outliers is higher in 30 – 45 age group.
- However, there is symmetry between all age groups.

Histogram

- A Histogram shows frequency for each bin or bucket created based on range of values of a variable. The histogram is recommended for a continuous variable and is generally used to check the Normality of the data.



- This plot shows that the distribution of Average Call Time is very much symmetric.

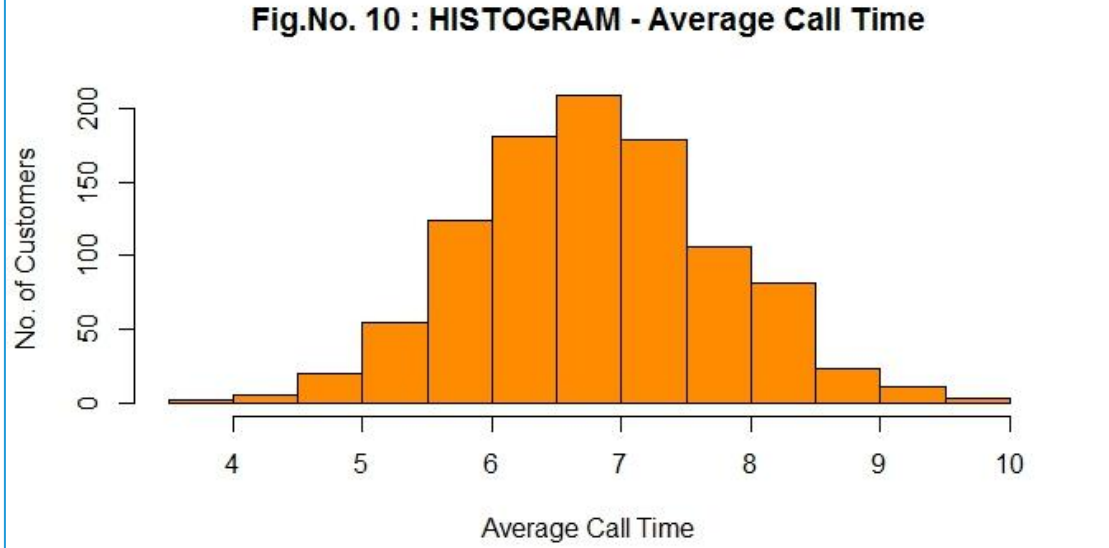
Histogram in R

```
# Histogram – Average Call Time

hist(telecom$AvgTime, breaks=12, main = "Fig.No. 10 : HISTOGRAM – Average
Call Time", xlab = "Average Call Time", ylab = "No. of Customers",
col="darkorange")
```

- **hist()** in base R yields a histogram
- **telecom$AvgTime** specfies vector (variable) for which the histogram needs to be plotted
- **breaks=** specifies the number of bins in the histogram
- **main=** provides the user defined name of the chart. It is to be put in double quotes
- **xlab=** provides a user defined label for the variable on X axis
- **ylab=** provides a user defined label for the variable on Y axis
- **col=** can be used to input your choice of color to the bars

Histogram in R

This plot shows the distribution of Average Call Time

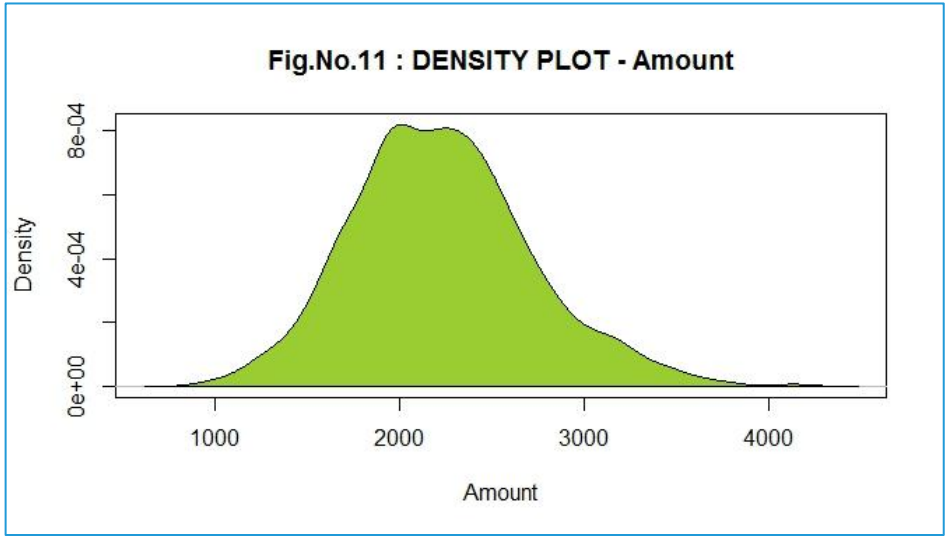


Interpretation :

- This plot shows that the distribution of Average Call Time is quite symmetric.

Density Plot

- A Density Plot is similar to a histogram which plots the probability.
- It is generally used to check the Normality of the data when there are higher data points.



- This plot shows that the distribution of amount is slightly positively skewed.

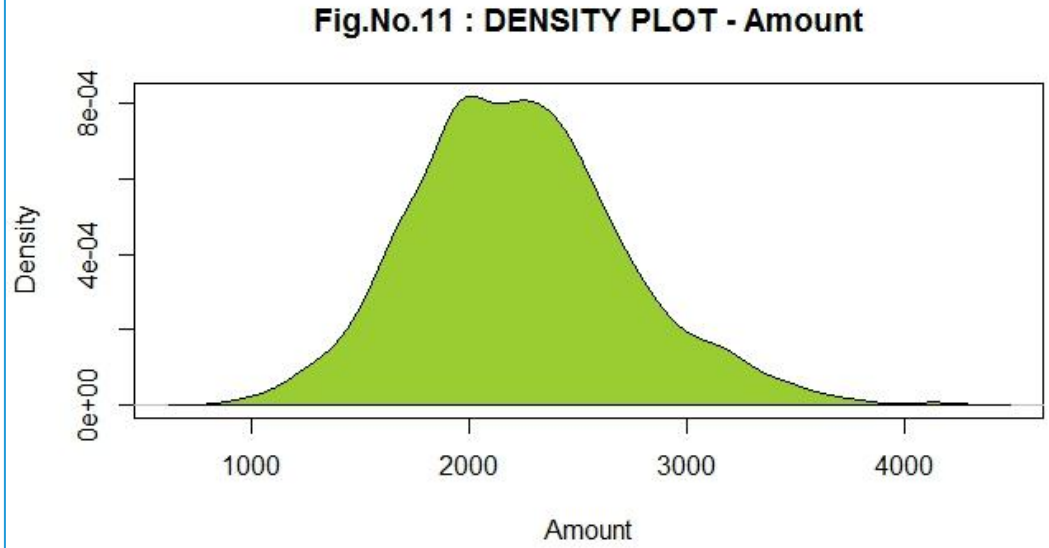
Density Plot in R

```
# Density Plot - Amount
Telecom_den<-density(telecom$Amt)
plot(telecom_den, main="Fig.No.11 : DENSITY PLOT - Amount",xlab="Amount")
polygon(telecom_den, col="yellowgreen")
```

- **density()** returns the density values of the variable
- **plot()** plots the line graph taking object returned by function density
- **main=** provides the user defined name of the chart. It is to be put in double quotes
- **xlab=** provides a user defined label for the variable on X axis
- **polygon()** shows the area covered under the curve
- **col=** can be used to input your choice of color to the polygon

Density in R

This plot shows the distribution of Amount



Interpretation :

- This plot shows that the distribution of Amount is slightly positively skewed.

Stem and Leaf Plot

- A Stem and Leaf diagram can, again, be an alternative to a histogram.
- It is a special table where each numeric value is split into a stem (First digit(s)) and a leaf (last Digit)
- Stem and leaf diagrams show the shape of the distribution (like bar charts) but have the advantage of not losing the detail of the original data.
- Arranging the leaves in numerical order, will allow us to use the diagram to find the middle value (the median) and the values that are a quarter and three-quarters of the way through the data (the lower and upper quartiles).

```
The decimal point is 2 digit(s) to the right of the |

10 | 56
11 |
12 | 014799
13 | 2223444556778
14 | 000111122333344456666666666678888888888889999999
15 | 00111111233334444556666777777788888888889999
16 | 000000000001111122222222222222233333333344444444555555555555666666+34
17 | 00000000001111111222222222222222222222233333333333334444444444555+53
18 | 000000000000001111111111111111111112222222222223333333333333333333+93
19 | 00000000000000000000000000001111111111122222222233333333333333333+89
20 | 0000000000000000000000111111111111222222222222222333333334444+54
21 | 000000011111111111122333334444455555555666666666666666778888888888
22 | 00001111111122233333334444444455555555555556778899
23 | 000011222344445555556777799
24 | 112237789999
25 | 2
26 | 7
27 | 0
```

Stem and Leaf Plot in R

```
# Stem and Leaf Plot in R
stem(telecom$Calls)
```

- **stem()** in base R yields a stem and leaf chart
- **telecom$Calls** specfies vector (variable) for which the stem and leaf plot needs to be plotted

Stem and Leaf Plot in R

```
# Output
```

```
The decimal point is 2 digit(s) to the right of the |

10 | 56
11 |
12 | 014799
13 | 2223444556778
14 | 000111122333344456666666666678888888888888999999
15 | 001111112333344445566667777777888888888889999
16 | 000000000001111122222222222222223333333334444444455555555555555666666+34
17 | 00000000001111111222222222222222222222222233333333333333444444444444555+53
18 | 0000000000000011111111111111111111112222222222223333333333333333333333+93
19 | 000000000000000000000000000001111111111122222222223333333333333333333+89
20 | 00000000000000000000001111111111111122222222222222222222233333333334444+54
21 | 0000000011111111111122333334444455555555566666666666666667788888888888
22 | 000011111111222333333344444444555555555555556778899
23 | 00001122234444555555677779 9
24 | 112237789999
25 | 2
26 | 7
27 | 0
```

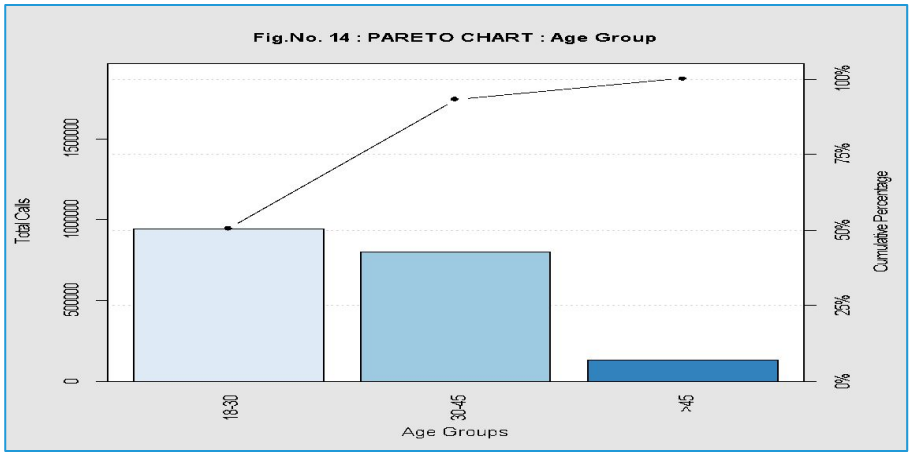
Interpretation :

- The stem and leaf plot of overall calling data shows that, calls values are symmetrically distributed and there exists few outliers also in the data.

* The output of the stem and leaf diagram in produced in the R console

Pareto Chart

- Pareto chart, named after Vilfredo Pareto, is a type of chart that contains both a bar and a line graph, where individual values are represented in descending order by bars. In this way the chart visually depicts which categories are more significant. The cumulative total is represented by the line.
- There needs to be at least one categorical variable to plot this chart.



- From the above chart we can interpret that 50% of the Total calls made come from age group 18-30.
- Another 42% calls are made by age group 30-45, only 8% calls are made by customers > 45 .

Get an Edge!

RColorBrewer

RColorBrewer is a package that uses www.colorbrewer2.org to help choose colour schemes for graphics in R

The colours are split into 3 groups :

1. Sequential : Light colours for low data, dark for high data
2. Diverging : Light colours for mid-range data, low and high contrasting dark colours
3. Qualitative: Colours designed to give maximum visual difference between classes

```
install.packages("RColorBrewer")
library(RColorBrewer)

col=brewer.pal(n,"palette")
```

We need to install the "RColorBrewer" package to use the color brewer in R

- **brewer.pal()** is the function to be used in **"col="** argument.
- **n** specifies the number of colors to be used
- **palette** is the name of the color palette which can be chosen by running **display.brewer.all()** function

* Refer to www.stat.auckland.ac.nz for details

Pareto Chart in R

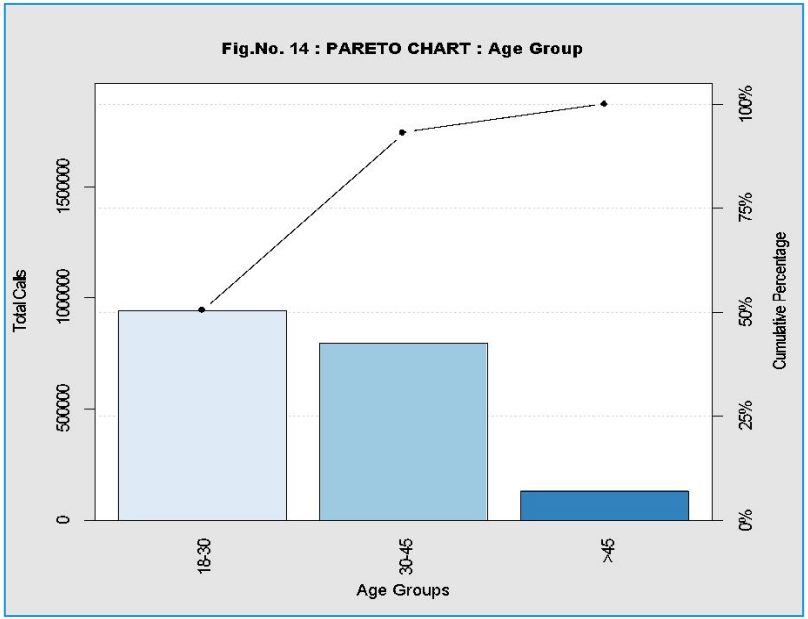
```
# Pareto Chart – Age Group

install.packages("qcc")
library(qcc)
library(RColorBrewer)

telecom1<-aggregate(Calls~Age_Group,data = telecom, FUN=sum)

pareto.chart(telecom1$Calls, xlab= "Age Groups" ,ylab= "Total Calls" ,
main = "Fig.No. 14 : PARETO CHART : Age Group",
col=brewer.pal(3,"Blues"), names.arg=telecom1$Age_Group)
```

Using "qcc" package is the easiest way to plot a Pareto Chart in R

- **pareto.chart()** is the function in "qcc" package used to plot a Pareto Chart
- **telecom1$Calls** specifies vector (variable) for which the Pareto chart needs to be plotted
- **names.arg=** is the argument that allows the bars to be named according the row names in the variable mentioned
- **main=** provides the user defined name of the chart. It has to be put in double quotes
- **xlab=** provides a user defined label for the variable on X axis
- **col=** can be used to input your choice of color to the bars
- **brewer.pal** uses the RColorBrewer to colour the bars

Pareto Chart in R

```
# Output
```



Interpretation :

- 50% of the Total calls made come from age group 18-30.
- Another 42% calls are made by age group 30-45, only 8% calls are made by customers > 45 years of age

Quick Recap

In this session, we learnt data visualisation using basic graphs

Chart Types and Functions in R

- Box-Whisker Plot - `boxplot()`
- Histogram - `hist()`
- Density Plot - `plot() + polygon()`
- Stem and Leaf Diagram - `stem()`
- Pareto Chart - `pareto.chart()` in package "qcc"