v2 Basic Data Visualisation in R

Contents

About Data Visualisation

What is Data Visualisation?

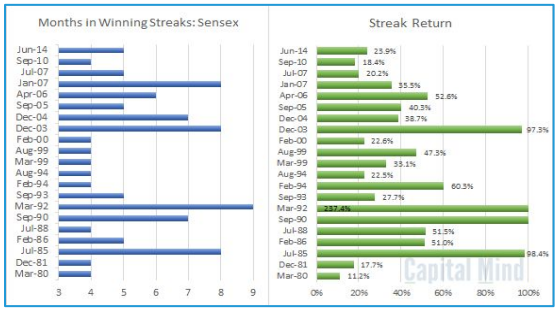
It is the visual representation of data in the form of graphs and plots.

Why is it important?

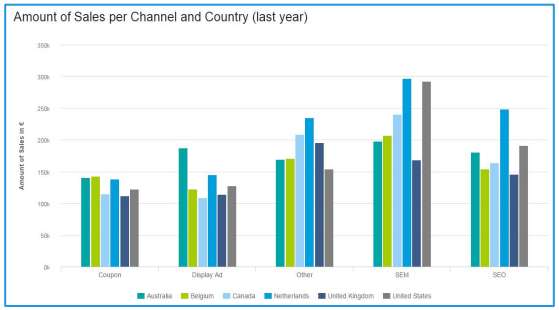
It enables us to

- See the data and get insights in one glance
- Allows us to grasp difficult / complex data in an easy manner
- Helps us to identify patterns or trends easily. Also shows distribution, correlation and causality in data.

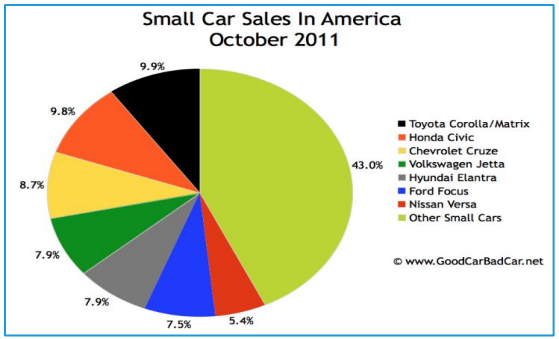
Application Areas



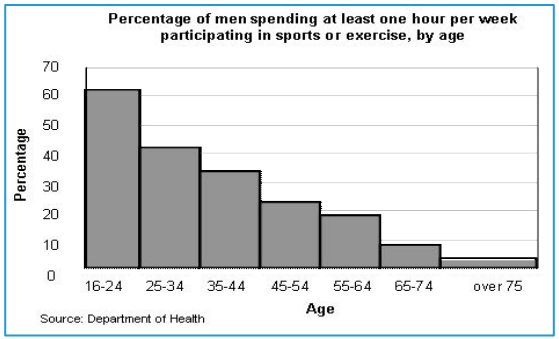
Sensex Charts



Sales Charts



Sales Charts



Survey Results

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

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

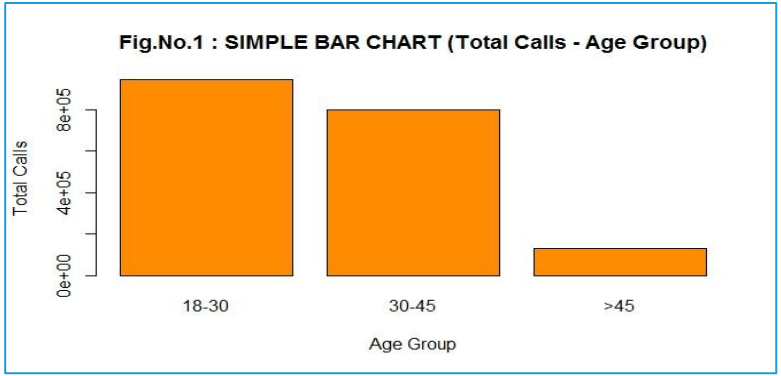
Simple Bar Diagram

A Bar Chart is the simplest and the most basic form of graph. In this graph, for each data item, we simply draw a ‘bar’ showing its value.

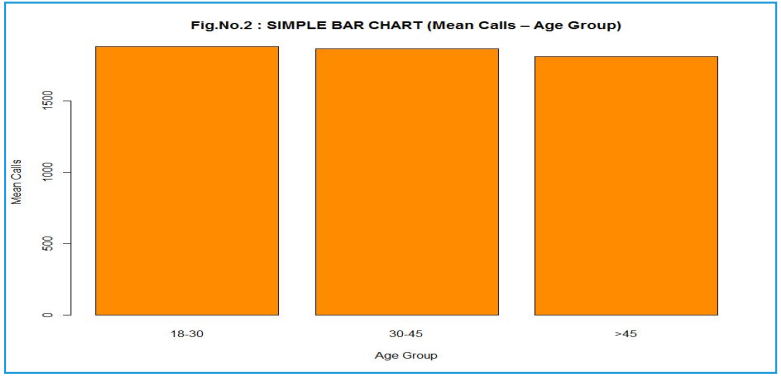
Simple Bar Chart: It is a type of chart which shows the values of different categories of data as rectangular bars with different lengths. The values are generally :

- Frequency
- Mean
- Totals
- Percentages

Simple Bar Diagram



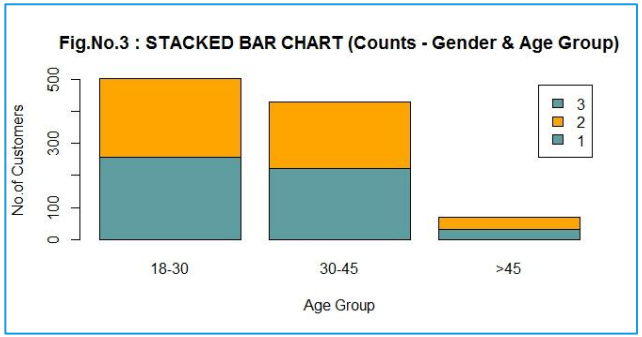
This graph simply gives the total number of calls for each age group.



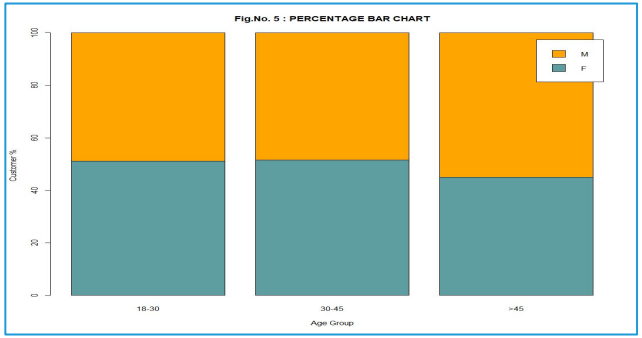
By plotting the average calls we can see that, though there is quite a difference in total calls in each age group, the average number of calls across age groups is similar.

Stacked Bar Diagram

Sub Divided or Stacked Bar Chart: It further divides the bar into different categories within the variable.



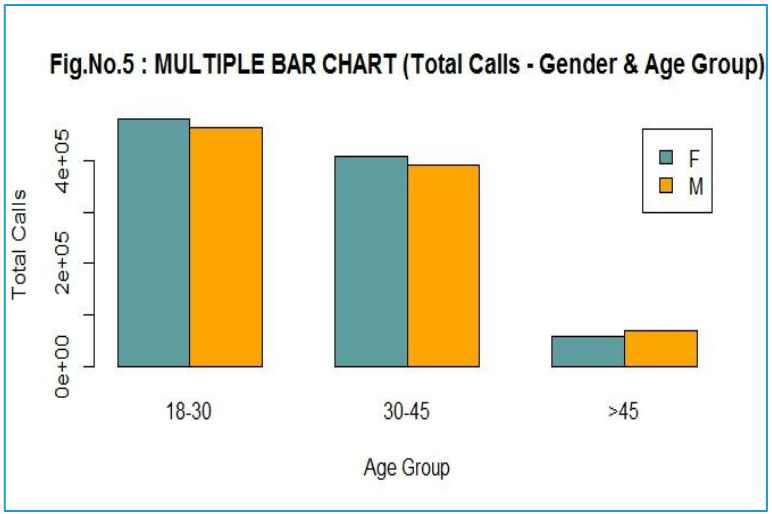
This graph divides the number of customers in each age group by Gender. However, this graph has absolute counts and is difficult to compare the gender wise distribution across the Age Groups.



Plotting a percentage stacked graph makes it efficient to compare the gender wise distribution of the number of customers across the Age Groups.

Multiple Bar Diagram

Multiple or Grouped Bar Chart: It divides the bar into different categories within the variable and places it one besides the other. By multiple bars diagram two or more sets of inter-related data are represented.



This bar chart can be used when we wish to see the gender-wise distribution of number of calls across age groups.

Diagrams in R

```
#Importing Data
telecom<-read.csv("telecom.csv", header=TRUE)

#Aggregating Data
telecom1<-aggregate(Calls~Age_Group,data = telecom, FUN=sum)
telecom1
```

```
   Age_Group   Calls
1        >45  128870
2      18-30  943187
3      30-45  798721
```

For plotting a bar chart in R, it is important to aggregate the data to get required vector/matrix

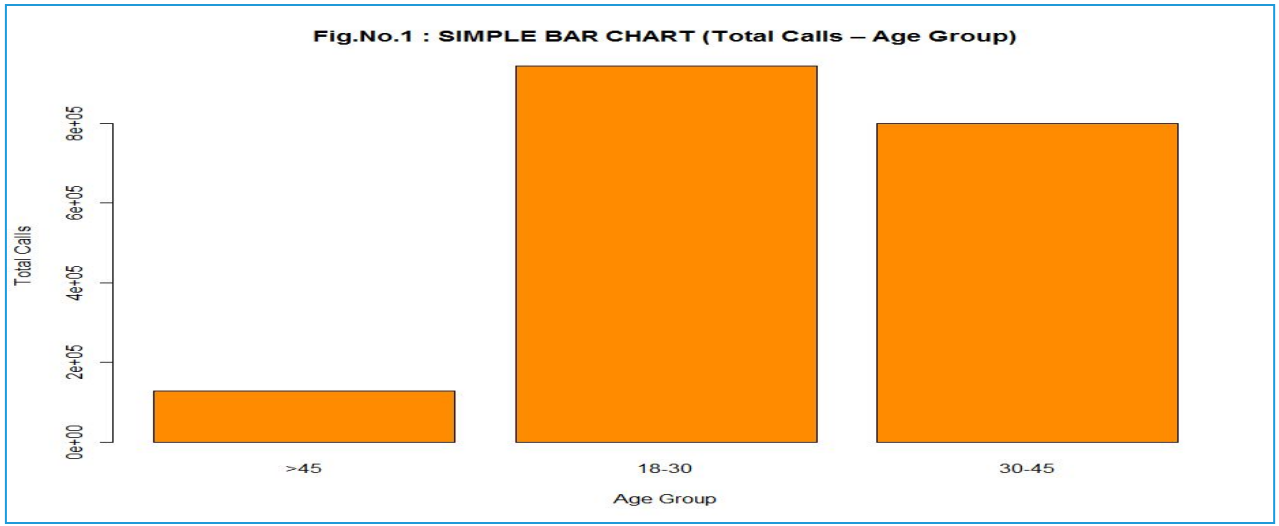
Simple Bar Chart in R

```
#Simple Bar Chart – Total Calls for different Age Groups
barplot(telecom1$Calls, main= "Fig.No.1 : SIMPLE BAR CHART (Total Calls – Age Group)", names.arg = telecom1$Age_Group,
        xlab = "Age Group", ylab="Total Calls", col = "darkorange")
```

- **barplot()** in base R yields different types of bar chart
- **telecom1$Calls** has to be a vector or matrix for which the bar chart needs to be plotted
- **main=** provides the user defined name of the chart. It has to be put in double quotes
- **names.arg=** specifies the names given to each bar
- **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

Simple Bar Chart in R

This is the output that you get on running the previous code

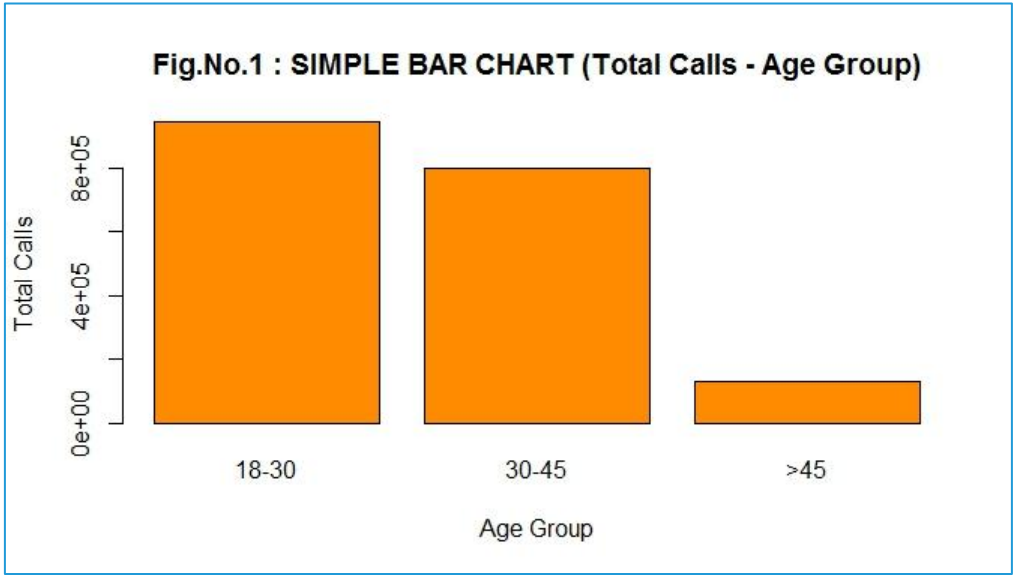


To get the bars in proper order, we will have to re-order the levels of column “Age_Group” in telecom data as follows & then run the same R codes again :

```
telecom$Age_Group <- factor(telecom$Age_Group, levels = c("18-30","30-45",
">45"))
```

Simple Bar Chart in R

This graph simply gives the distribution of the **Total number of calls** across different **Age Groups**.



Interpretation :

- Number of calls made by young age group (18-30) is slightly higher than mid age group (30-45) and very high than age group >45.

Simple Bar Chart in R

```
# Simple Bar Chart – Mean Calls for different Age Groups
telecom2<-aggregate(Calls~Age_Group,data = telecom, FUN=mean)
Telecom2
```

```
 Age_Group    Calls
1     18-30 1882.609
2     30-45 1866.171
3       >45 1815.070
```

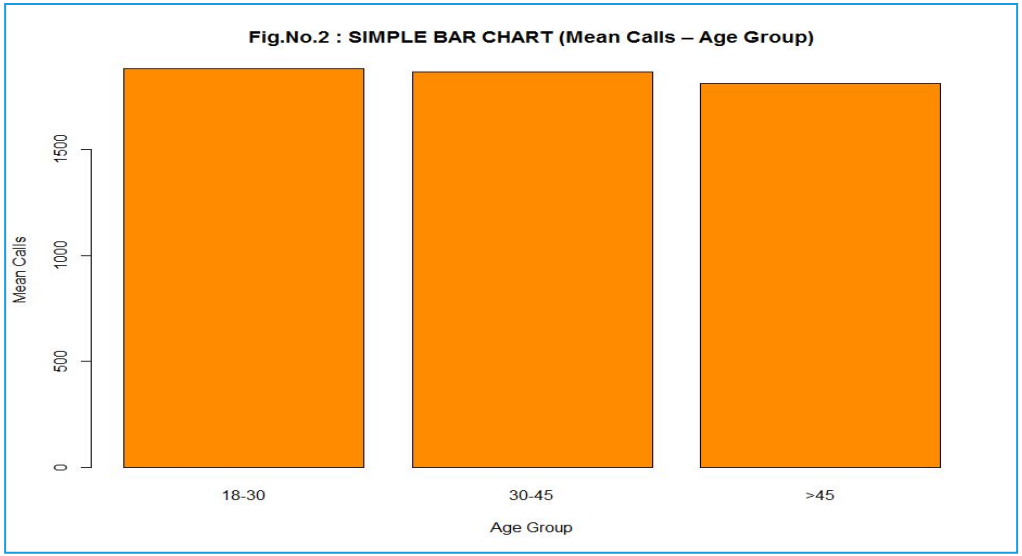
```
barplot(telecom2$Calls, main= "Fig.No.2 : SIMPLE BAR CHART (Mean Calls – Age Group)", names.arg = telecom2$Age_Group, xlab = "Age Group", ylab="Mean Calls", col = "darkorange")
```

Note :

- The barplot code remains the same with respect to previous barplot code, the only difference is while aggregating the data.
- In previous plot aggregation function was **“sum”** & in this plot aggregation function is **“mean”**.

Simple Bar Chart in R

This graph simply gives the distribution of the **Mean calls** across different **Age Groups**.



Interpretation :

- By plotting the average calls we can see that, though there is quite a difference in total calls in each age group, **the average number of calls across age groups is similar**.

Simple Bar Chart in R

```
# Simple Bar Chart in Horizontal orientation
tele1<-table(telecom$Age_Group)
tele1
```

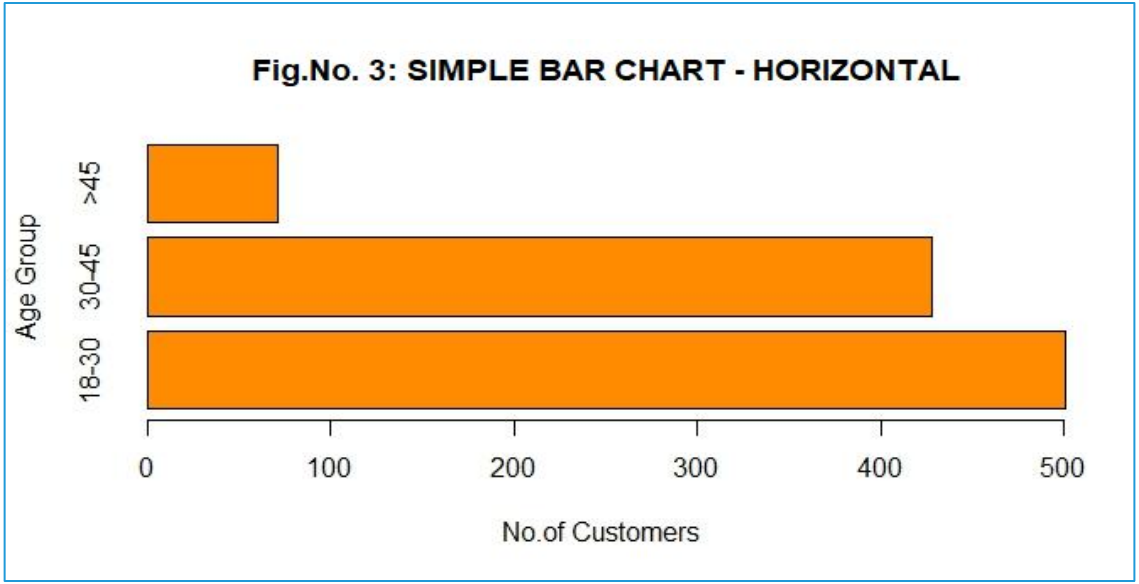
```
18-30   30-45  >45
  501    428    71
```

```
barplot(tele1, main= "Fig.No. 3: SIMPLE BAR CHART - HORIZONTAL",
xlab="No.of Customers",ylab = "Age Group",col = "darkorange",horiz = TRUE)
```

- **horiz =** gives horizontal orientation to the bars. It takes the frequency on the X axis

Simple Bar Chart in R

This graph displays the number of customers across age group.



Interpretation :

- This is horizontal view, which indicates that there are very few customers for age group >45 as compared to other two age groups.
- This graph is generally useful when there are negative frequency values in the data.

Stacked Bar Chart in R

```
# Stacked Bar Chart
telecom3<-table(telecom$Gender,telecom$Age_Group)
telecom3
```

```
  18-30  30-45  >45
F   256    221   32
M   245    207   39
```

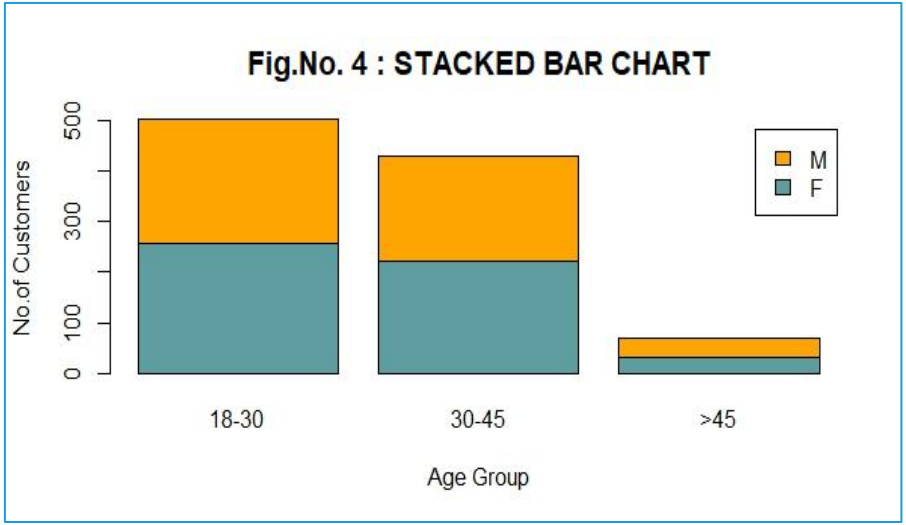
- **table()** inputting two variables gives a matrix having their counts in each category

```
barplot(telecom3,main="Fig.No. 4 : STACKED BAR CHART",
xlab ="Age Group",ylab ="No.of Customers", col=c("cadetblue","orange"),
legend=rownames(telecom3))
```

- **legend=rownames()** displays the legend on the graph output

Stacked Bar Chart in R

This graph divides the number of customers in each age group by Gender.



Interpretation :

- This graph shows that, though there are more young customers in data but, almost equal number of Males and Females are present in each age group.

Percentage Bar Chart in R

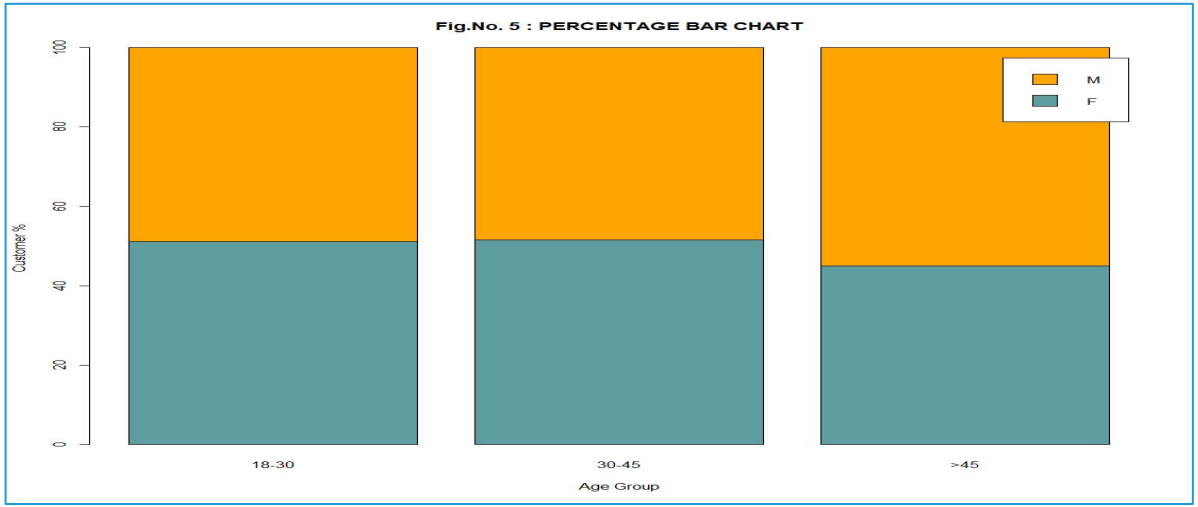
```
# Percentage Bar Chart

telecom4<-prop.table(telecom3,2)
telecom4

      18-30       30-45       >45
F   0.5109789   0.5163551   0.45070
M   0.4890220   0.4836449   0.54929

barplot(telecom4*100, main= "Fig.No. 5 : PERCENTAGE BAR CHART",
xlab = "Age Group", ylab="Customer %", col = c("cadetblue","orange"),
legend= rownames(telecom4))
```

- **prop.table()** helps us create data frame with percentage values
- **(,2)** gives percentage as per column count

- **telecom4*100** has to be a vector or matrix for which the bar chart needs to be plotted. *100 would display percentage scale on y-axis.

Percentage Bar Chart in R

```
# Output for gender wise distribution of number of customers across the
Age Groups.
```



Interpretation :

- Data contains almost equal proportion of Male and Female callers across three different age groups.
- Plotting a percentage stacked graph makes it efficient to compare the gender wise distribution of the number of customers across the Age Groups.

Multiple Bar Chart in R

```
# Multiple Bar Chart
telecom5<-aggregate(Calls~Age_Group+Gender, data= telecom, FUN = sum)
telecom5
telecom6<-xtabs(Calls~Gender+Age_Group,telecom5)
telecom6
```

```
    18-30    30-45     45
F  480235   408184   583
M  462952   390537   705
```

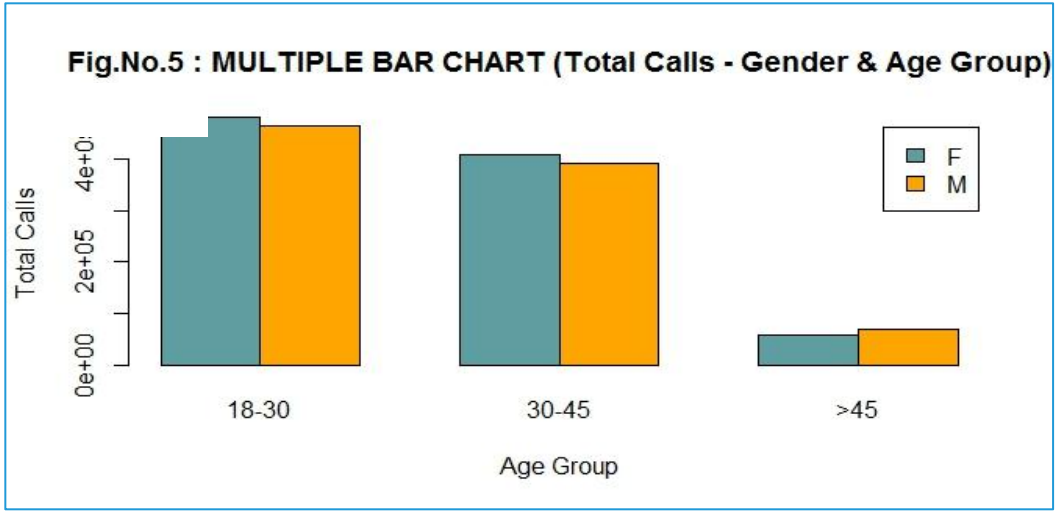
- **xtabs()** is used to cross tabulate the categories of more than one variables using another numeric variable which results in total of each category

```
barplot(telecom6,main="Fig.No.6 : MULTIPLE BAR CHART (Total Calls - Gender
& Age Group)", xlab ="Age Group", ylab="Total Calls",
col=c("cadetblue","orange"), legend=rownames(telecom6), beside = TRUE)
```

- **beside=TRUE** enables us to show the different class of the same bar one beside the other

Multiple Bar Chart in R

```
# Output for gender-wise distribution of number of calls across age groups
```



Interpretation :

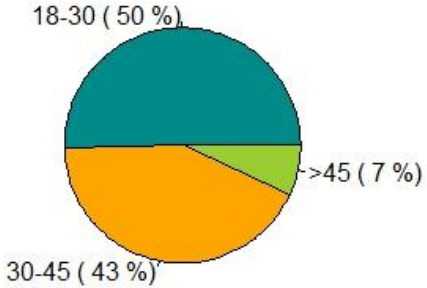
- There is no significant difference between Male and Female in terms of number of calls made across three different age groups, the only difference is that, age group >45 has slightly more male customers than female customers as compared to other age groups.
- This can be used as an alternative way of representing a stacked bar graph.

Pie Chart

Pie charts are generally used to show percentage or proportional data.

In this graph the entire circle (pie) is sliced proportional to the values of each category.

Fig.No. 6 : Pie Chart with Percentage



The above Pie Chart show how the total number of Calls are proportionally distributed amongst age groups.

Pie Chart in R

```
# Pie Chart
telecom7<-aggregate(Calls~Age_Group,data = telecom, FUN=sum)
telecom7$pct <- round(telecom7$Calls/sum(telecom7$Calls)*100)
telecom7
```

```
    Age_Group    Calls    pct
1       18-30   943187     50
2       30-45   798721     43
3         >45   128870      7
```

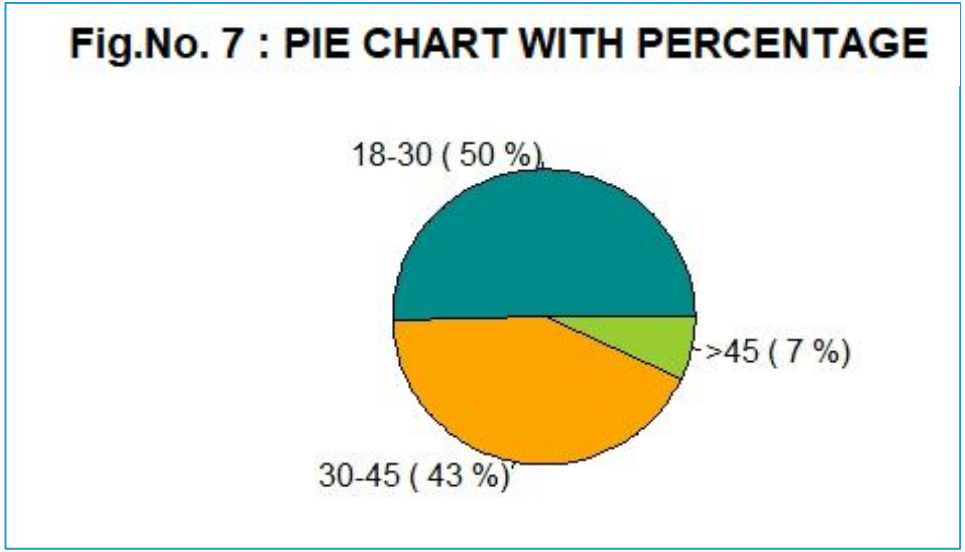
Here, we calculate the proportions for each category using formula

```
pie(telecom7$Calls,labels = paste(telecom7$Age_Group,"(",telecom7$pct,"%)"),
col=c("darkcyan","orange","yellowgreen"),main="Fig.No. 7 : PIE CHART WITH
PERCENTAGE")
```

- **pie()** in base R yields a pie chart
- **telecom7$Calls** has to be a vector or matrix for which the pie chart needs to be plotted
- **labels=** provides a user defined label for the variable on X axis
- **paste()** labels each category using string values separated by commas
- **col=** can be used to input your choice of color to the bars

Pie Chart in R

```
# Output of Pie chart with percentage
```



Interpretation :

- **50%** of calls are made by Age_Group 18-30, **43%** by 30-45 & **only 7%** by >45 Age_Group.

Quick Recap

In this session, we learnt data visualisation using basic graphs

Chart Types and Functions in R	• Bar Diagrams - `barplot()` • Pie Chart - `pie()`