v2 Visualisations

using ggplot2 in R

Contents

What is ggplot2?

- The ggplot2 package, created by Hadley Wickham, offers a powerful graphics language for creating elegant and complex plots.
- Originally based on Leland Wilkinson's The Grammar of Graphics, ggplot2 allows you to create graphs that represent both univariate and multivariate, numerical and categorical data in a straightforward manner.
- Grouping can be represented by color, symbol, size, and transparency.
- This package is available from CRAN via
  - `install.packages("ggplot2")`
  - `library(ggplot2)`

*Website: http://ggplot2.org (better documentation)

Case Study - 1

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 look at the distribution of customer database
To see how the Calls and Amount are distributed across customers

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	Active usage of telecom	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

Diagrams in R

```
# Importing Data

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

# Installing and calling the package

install.packages("ggplot2")
library(ggplot2)
```

Simple Bar Chart in R

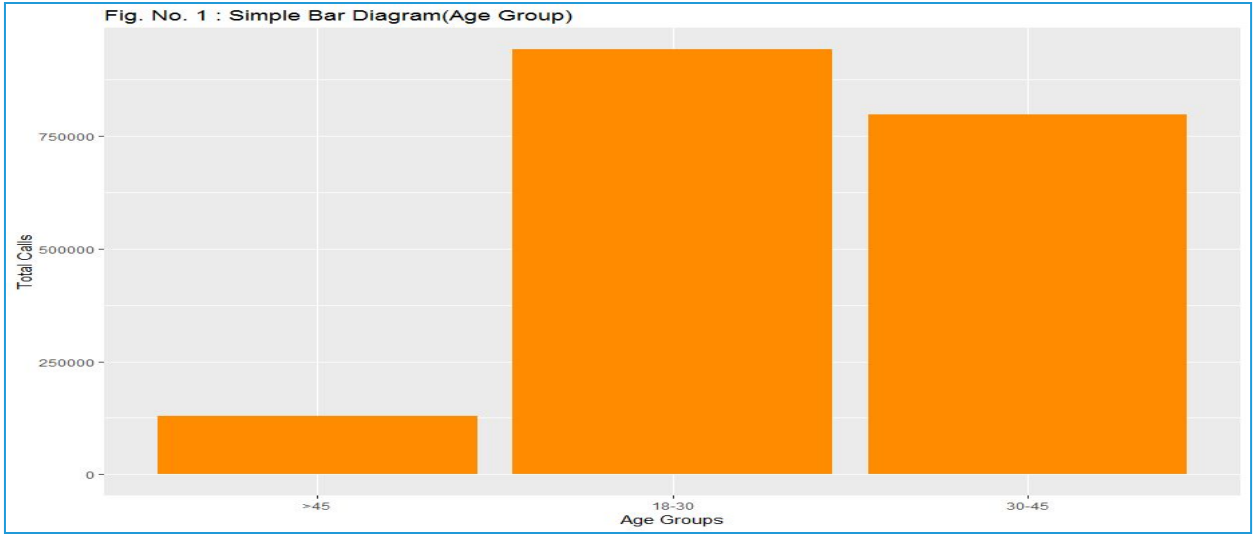
```
# Simple Bar Chart (Age Group)

ggplot(telecom,aes(x=Age_Group,y=Calls))+
geom_bar(stat="identity",fill="darkorange")+labs(x="Age Groups",y="Total
Calls",title="Fig. No. 1 : Simple Bar Diagram(Age Group)")
```

- **ggplot()** is a function in ggplot2 which yields different types of plots
- **telecom** is the data that is used
- **aes()** specifies the variables to be used on each axis
- **geom_bar()** makes the height of the bar proportional to the number of cases in each group
- **stat="identity"** is used to represent the height of the bar which represent values in the data
- **labs()** is used to label the various features of the graph

* In `geom_bar(), stat="bin"` is the default function which represents count of cases in each group

Simple Bar Chart in R

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

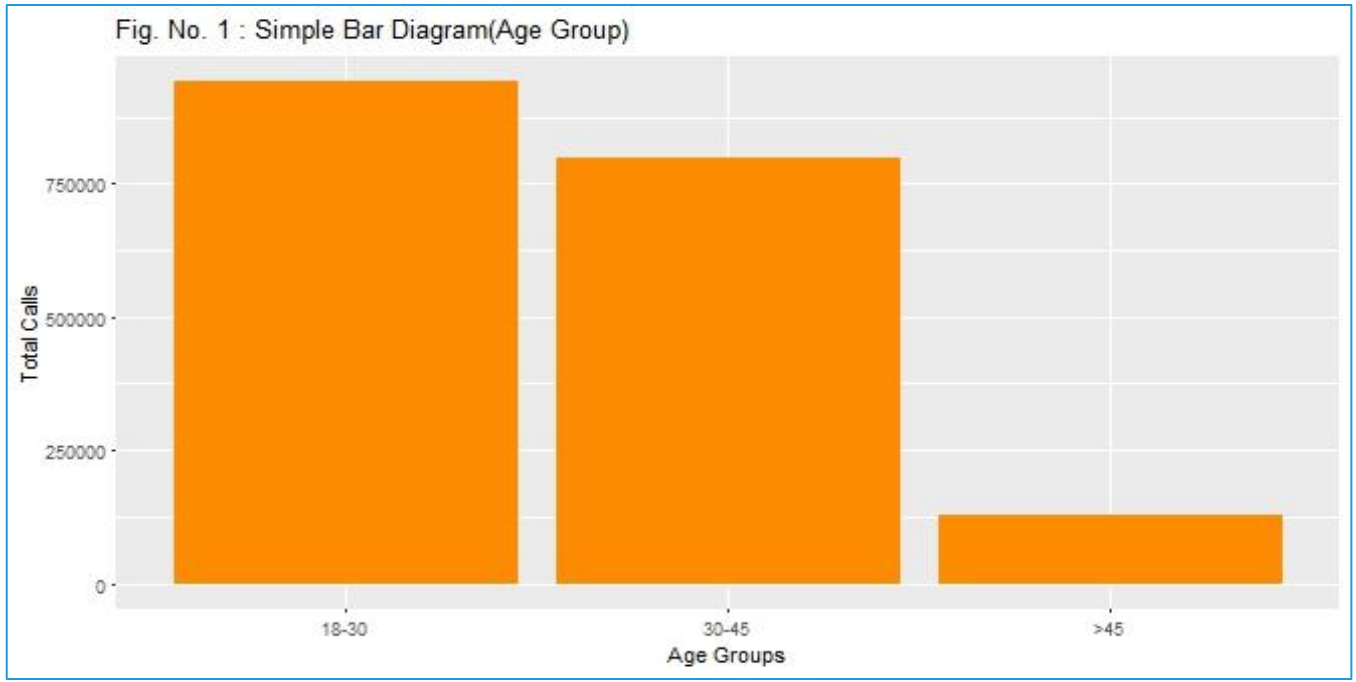


To get the bars in proper order, we will have to re-order the levels of column "Age_Group" in telecom data as follows & then run the same ggplot code :

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

Simple Bar Chart in R

```
# Output is a ordered bar graph :
```



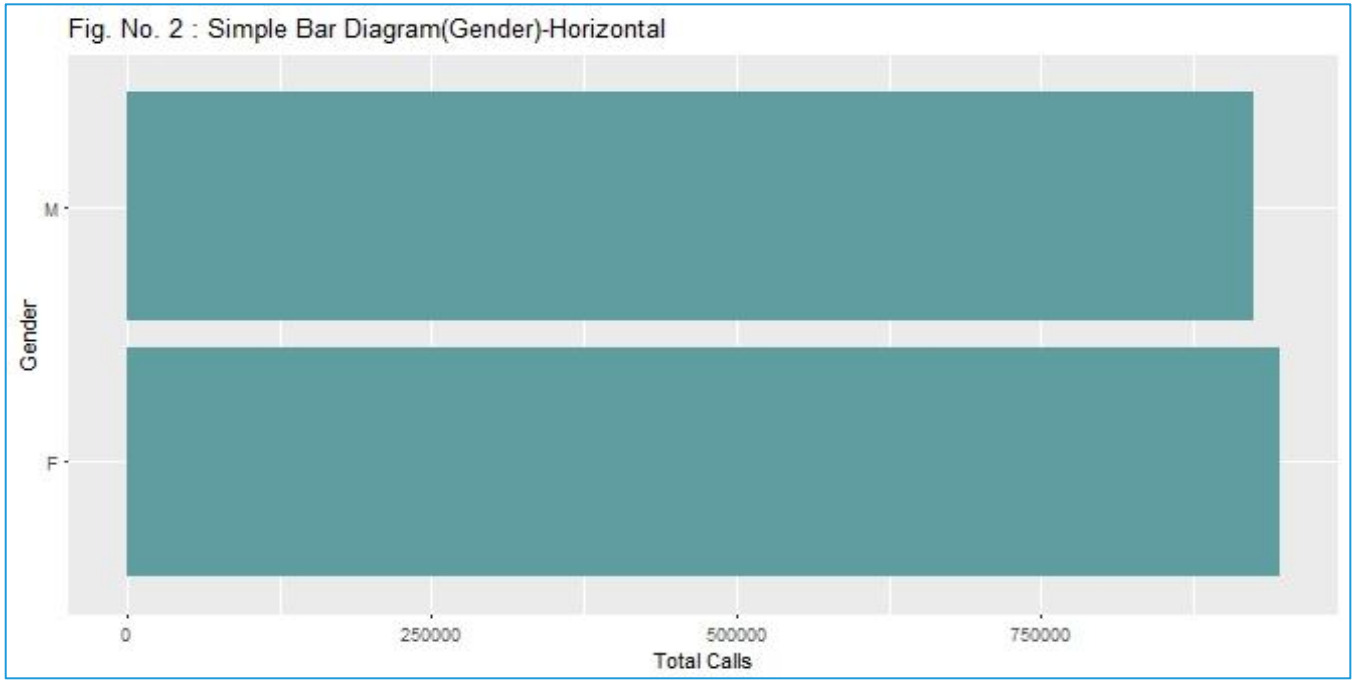
Simple Bar Chart in R

```
# Simple Bar Chart (Gender) - Horizontal

ggplot(telecom, aes(x=Gender, y=Calls))+
   geom_bar(stat="identity",fill="cadetblue")+
    labs(x="Gender",y="Total Calls",
      title="Fig. No. 2 : Simple Bar Diagram(Gender)-Horizontal")+
         coord_flip()
```

- **coord_flip()** gives us horizontal bars by flipping the co-ordinates.

Simple Bar Chart in R

```
# Output
```



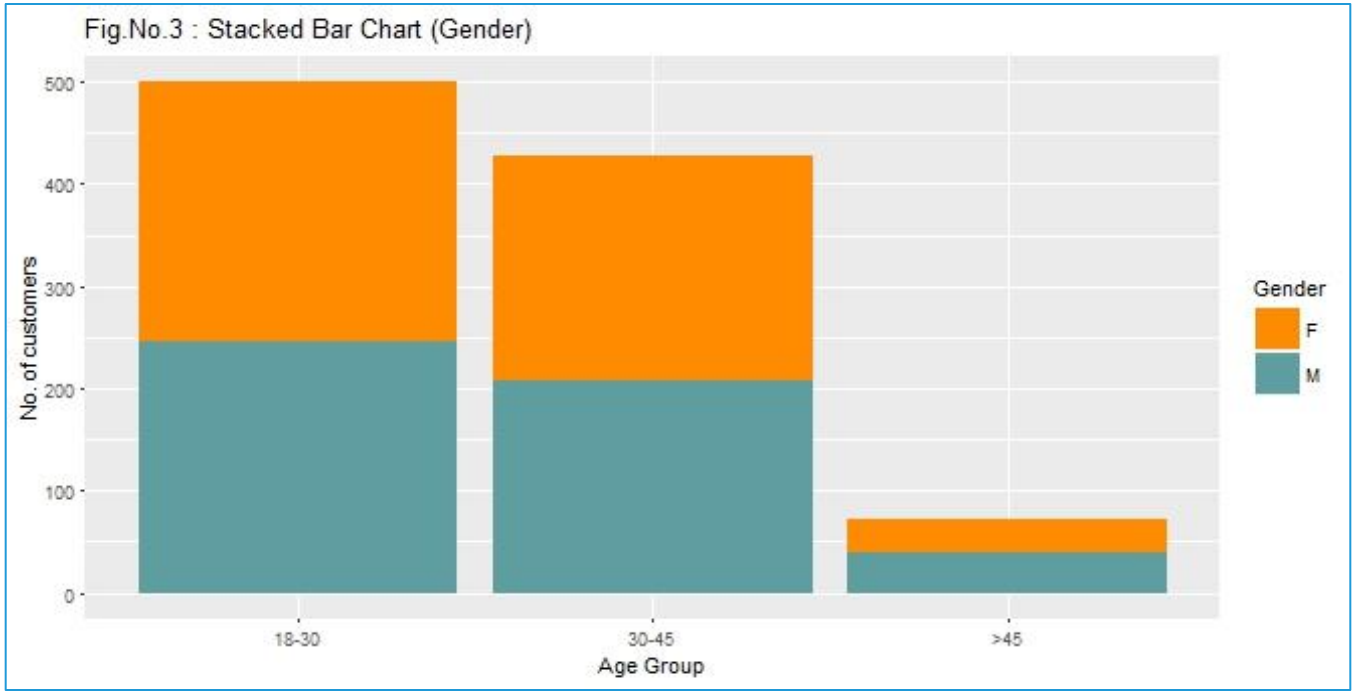
Stacked Bar Chart in R

```
# Stacked (or Sub-Divided) Bar Chart
ggplot(telecom, aes(x=Age_Group))+ geom_bar(aes(fill=Gender))+
labs(x="Age Group", y="No. of customers",title="Fig.No.3 : Stacked Bar
Chart (Gender)")+scale_fill_manual(values=c("darkorange","cadetblue"))
```


- **aes() function in geom_bar()** divides each bar as per the input variable using **fill=** Gender
- **scale_fill_manual()** allows to use the user defined colors for the sub divided bar

Stacked Bar Chart in R

```
# Output
```



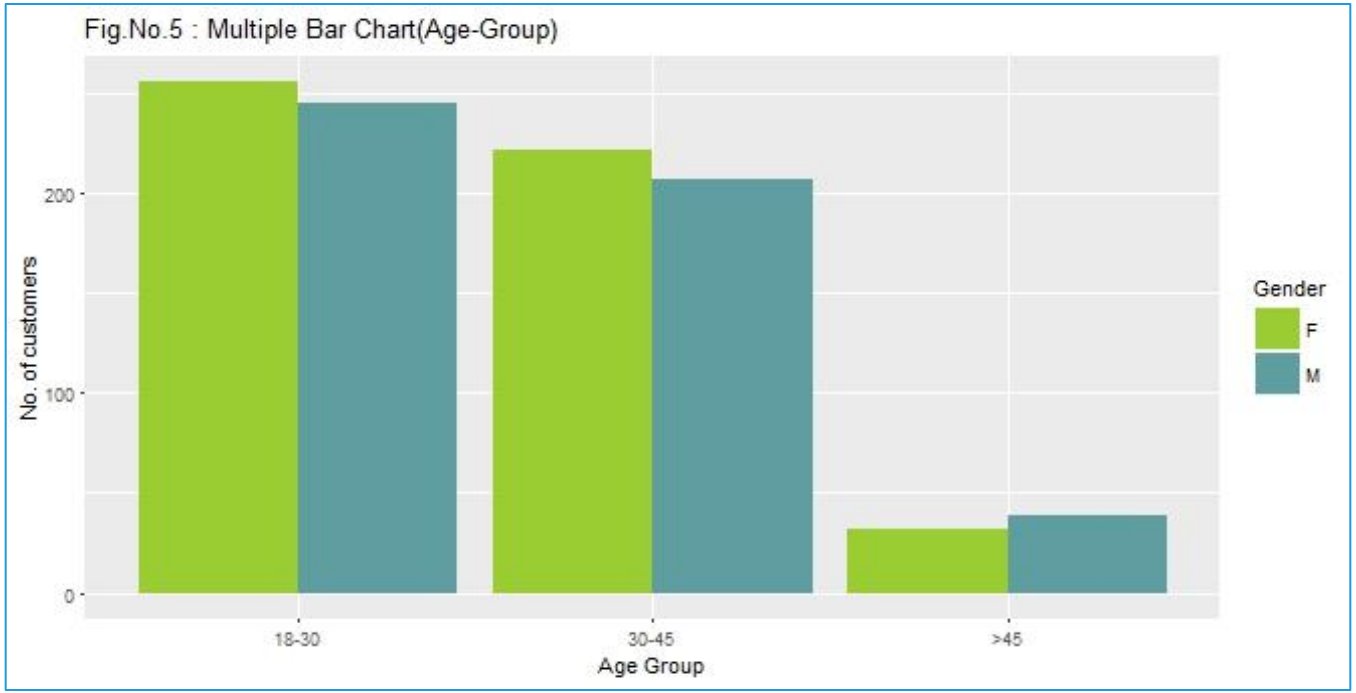
Multiple Bar Chart in R

```
# Multiple (or Grouped) Bar Chart

ggplot(telecom, aes(x=Age_Group))+geom_bar(aes(fill=Gender),position="dodge")
+ labs(x="Age Group", y="No. of customers",title="Fig.No.5 : Multiple Bar
Chart(Age-Group)")+ scale_fill_manual(values=c("yellowgreen","cadetblue"))
```

- **position="dodge"** gives us the divided bars one beside the other

Multiple Bar Chart in R

```
# Output
```



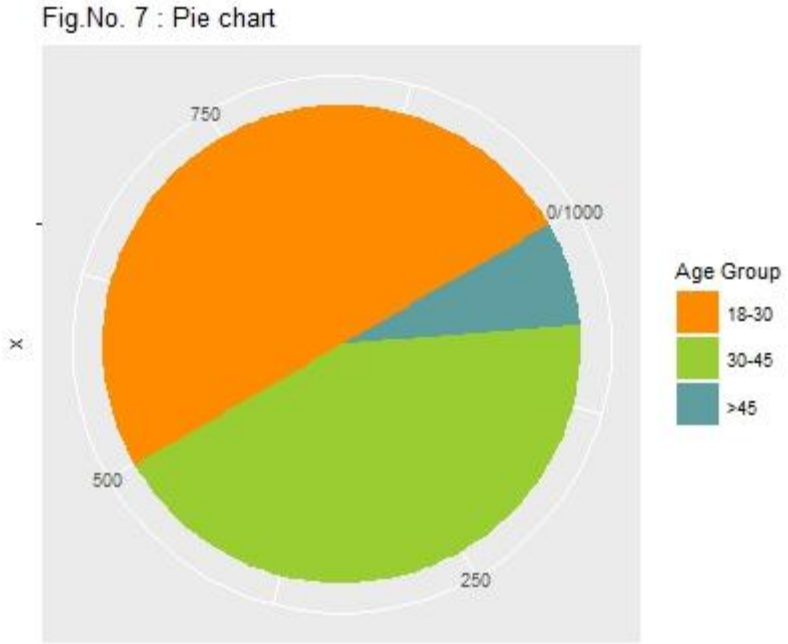
Pie Chart in R

```
# Pie Chart
ggplot(telecom, aes(x="", fill=Age_Group))+ geom_bar(width=1)+
coord_polar(theta="y", start=pi/3)+labs(title="Fig.No. 7 : Pie chart",
fill="Age Group")+scale_fill_manual(values=c("darkorange","yellowgreen",
"cadetblue"))
```

- **coord_polar()** it transforms stacked bar charts into circular pie chart
- **theta="y"** uses Y axis scale for proportion
- **start=pi/3** it starts the first proportion of pie from pi/3 angle

Pie Chart in R

```
# Output
```



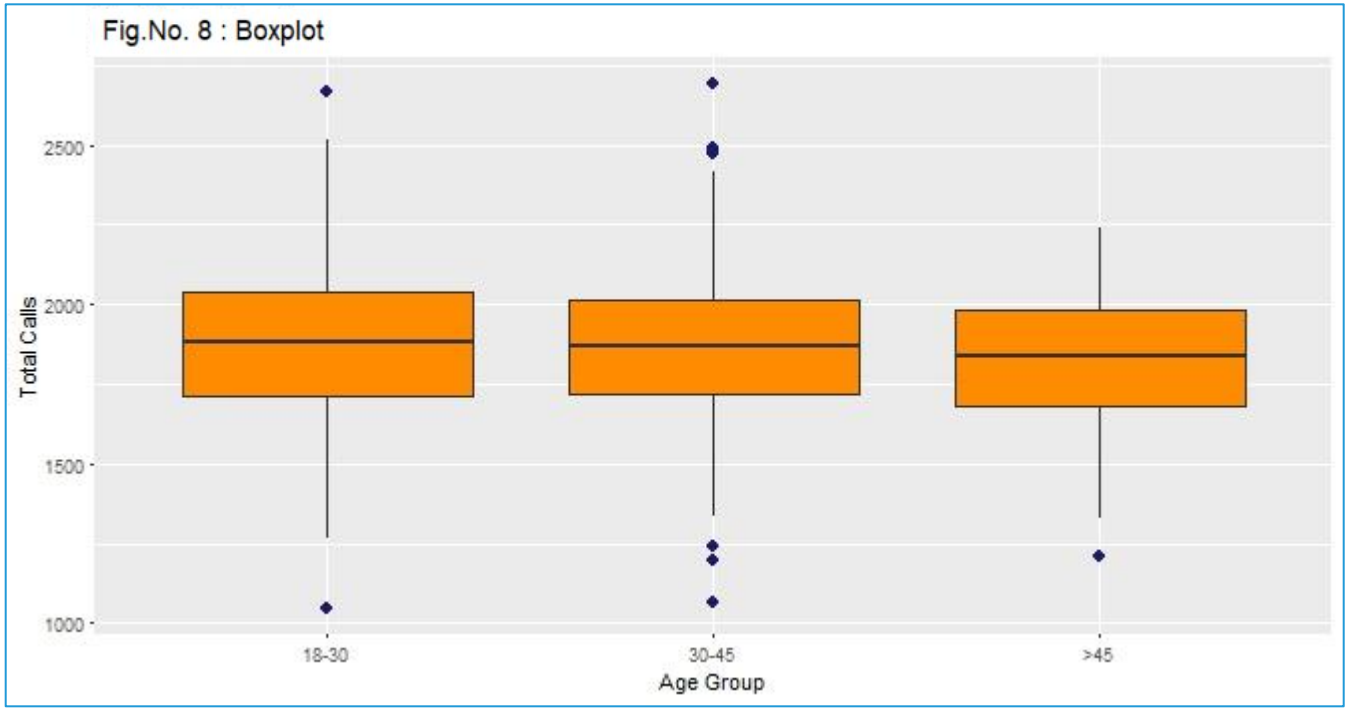
Box Plot in R

```
# Box Plot
ggplot(telecom, aes(x=Age_Group, y=Calls))+ geom_boxplot(fill="darkorange",
outlier.colour="midnightblue", outlier.size=2.5)+labs(x="Age Group", y="Total
Calls", title="Fig.No. 8 : Boxplot")
```

- **geom_boxplot ()** calls the boxplot function

Box Plot in R

```
# Output
```



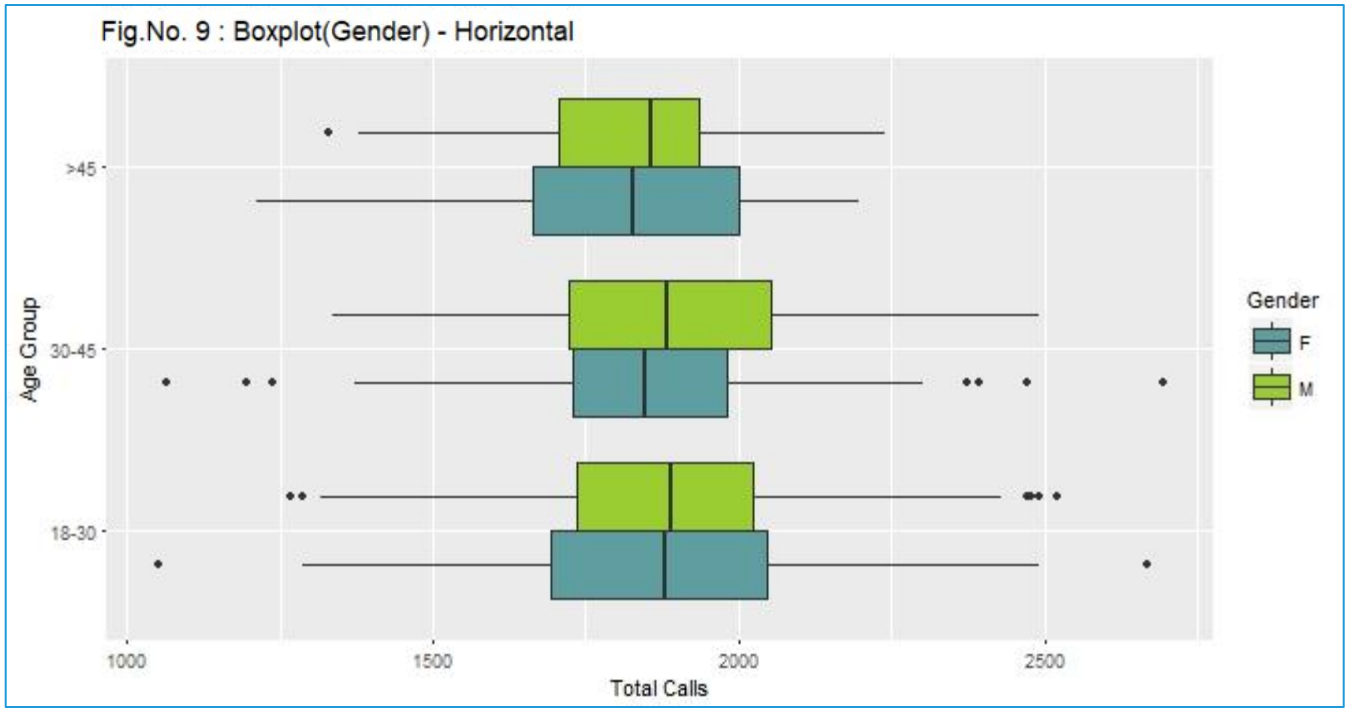
Box Plot in R

```
# Box Plot (Gender) - Horizontal
ggplot(telecom, aes(x=Age_Group, y=Calls))+ geom_boxplot(aes(fill=Gender))+
labs(y="Total Calls", x="Age Group",title="Fig.No. 9 : Boxplot(Gender) -
Horizontal")+scale_fill_manual(values=c("cadetblue","yellowgreen"))
+coord_flip()
```

- **aes()function in geom_boxplot()** gives multiple boxplot one beside the other using **fill=** Gender

Box Plot in R

```
# Output
```



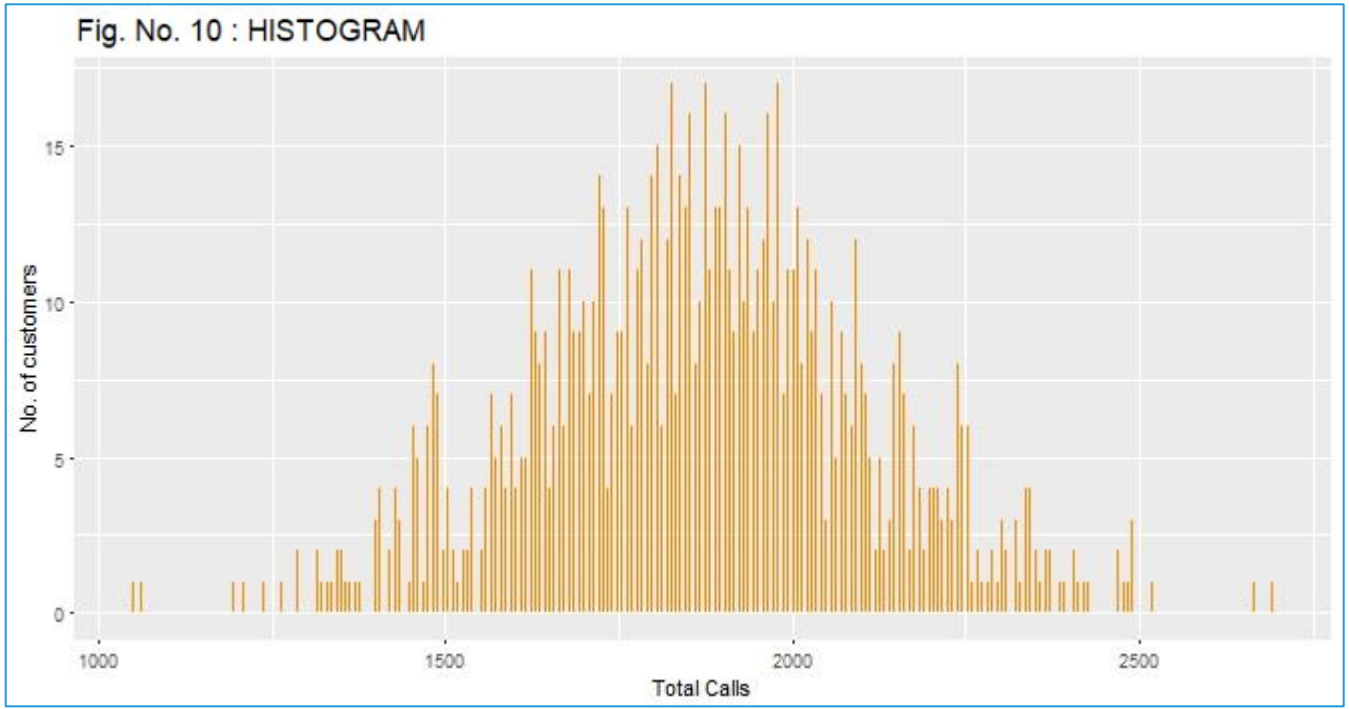
Histogram in R

```
# Histogram
ggplot(telecom, aes(x=Calls))+ geom_histogram(binwidth=2,
fill="darkorange")+labs(x="Total Calls", y="No. of customers", title="Fig. No.
10 : HISTOGRAM")
```

- **geom_histogram()** is used to plot histogram
- **binwidth=** gives size to each bar in the graph

Histogram in R

```
# Output
```



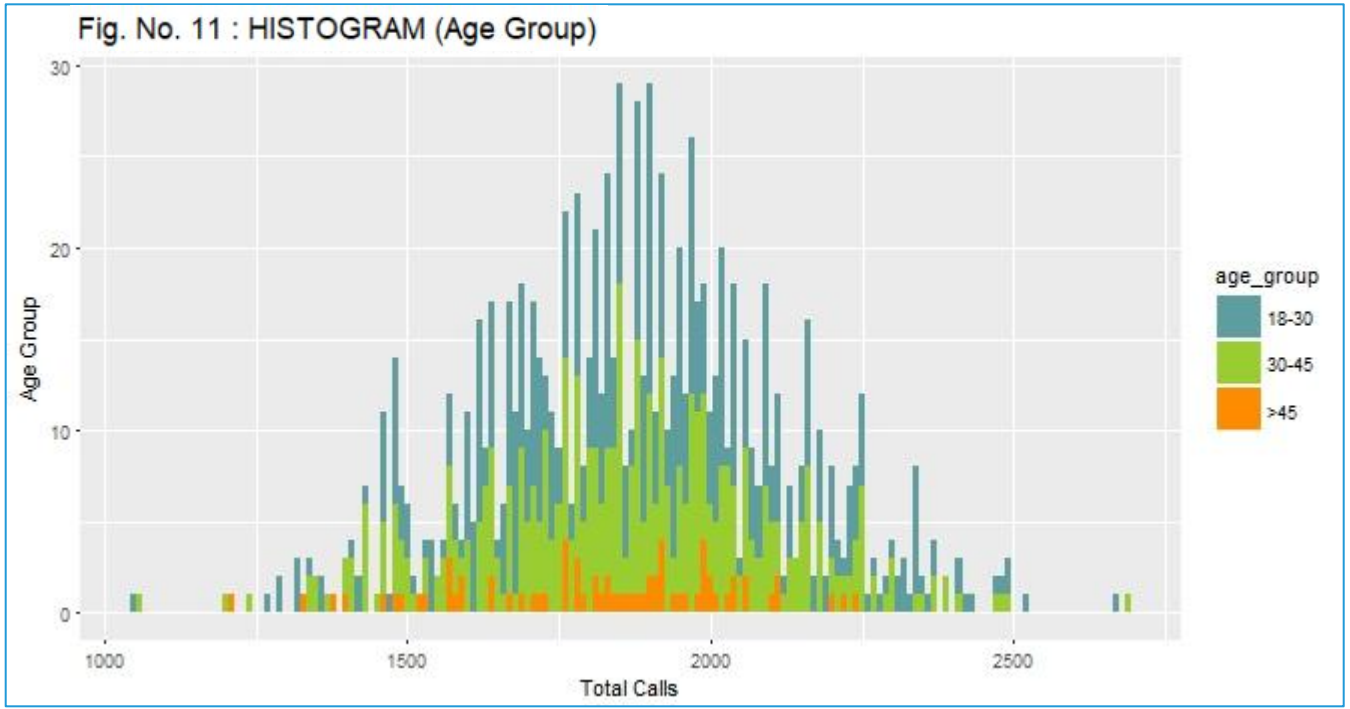
Histogram in R

```
# Histogram with Age Group
ggplot(telecom, aes(x=Calls))+ geom_histogram(aes(fill=Age_Group),
binwidth=10)+ labs(x="Total Calls", y="Age Group",title="Fig. No. 11 :
HISTOGRAM (Age Group)", colour="Age Group")+
scale_fill_manual(values=c("cadetblue","yellowgreen", "darkorange"))
```

- **aes() function in geom_histogram()** gives multiple bar one over the other using **fill=** Age_Group

Histogram in R

```
# Output
```



Data Snapshot

JOB PROFICIENCY DATA

Variables

empno	aptitude	testofen	tech_	g_k_	job_prof
1	86	110	100	87	88
2	62	62	99	100	80
3	110	107	103	103	96

Observations

Columns	Description	Type	Measurement	Possible values
empno	Employee No	Numeric	-	-
aptitude	Aptitude	Numeric	-	positive values
testofen	Test of English	Numeric	-	positive values
tech_	Technical Score	Numeric	-	positive values
g_k_	General Knowledge	Numeric	-	positive values
job_prof	Job Proficiency	Numeric	-	positive values

Case Study - 2

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

Background

A company has the scores of various attribute tests of their employees

Objective

To study the correlation between Aptitude and Job Proficiency.

Sample Size

25

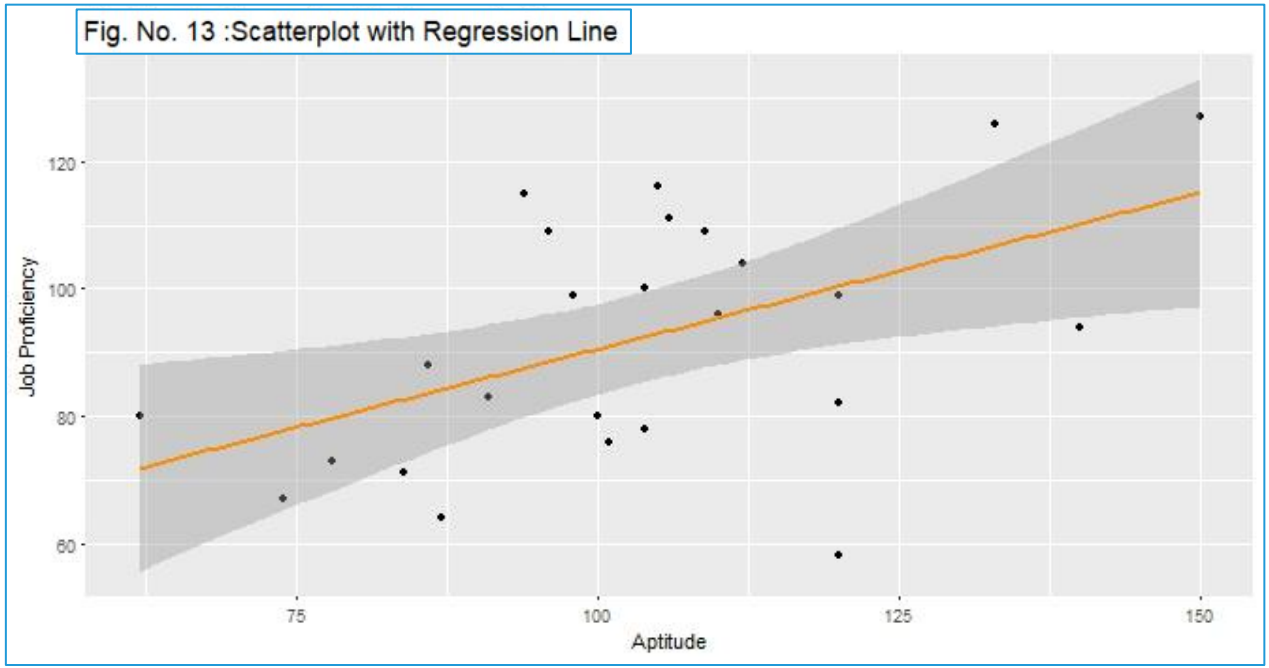
ScatterPlot with Regression Line in R

```
# Importing Data
job<-read.csv("JOB PROFICIENCY DATA.csv", header=TRUE)

# Scatterplot with Regression Line
ggplot(job, aes(x=aptitude, y=job_prof))+
geom_point()+geom_smooth(method="lm",col="darkorange")+
labs(x="Aptitude", y="Job Proficiency",title="Fig. No. 13 :Scatterplot
with Regression Line")
```

- **geom_point ()** is used to plot the data points, in this case it's a scatter plot
- **geom_smooth ()** is used to plot the curve
- **method="lm"** is used to get a linear regression line

ScatterPlot with Regression Line in R

```
# Output
```



Data Snapshot

Plotting a trendline requires time-element.
Consider the following two datasets. Week can be taken as the time element.

TelecomData_CustDemo

Variables

Observations

CustID	Age	Gender	PinCode	Active
1001	29	F	186904	Yes

Columns	Description	Type	Measurement	Possible values
CustID	Customer ID	Numeric	-	-
Age	Age	Numeric	-	18-51
Gender	Gender	Categorical	-	M,F
PinCode	Area's PinCode	Numeric	-	-
Active	Active usage of telecom	Categorical	-	Y,N

TelecomData_WeeklyData

Variables

Observations

CustID	Week	Calls	Minutes	Amt
1001	1	56	392	78.4

Columns	Description	Type	Measurement	Possible values
CustID	Customer ID	Numeric	-	-
Week	Week no.	Numeric	-	1-24
Calls	No. of Calls	Numeric	-	positive values
Minutes	Total Minutes	Numeric	Minutes	positive values
Amt	Amount Charged	Numeric	Rs.	positive values

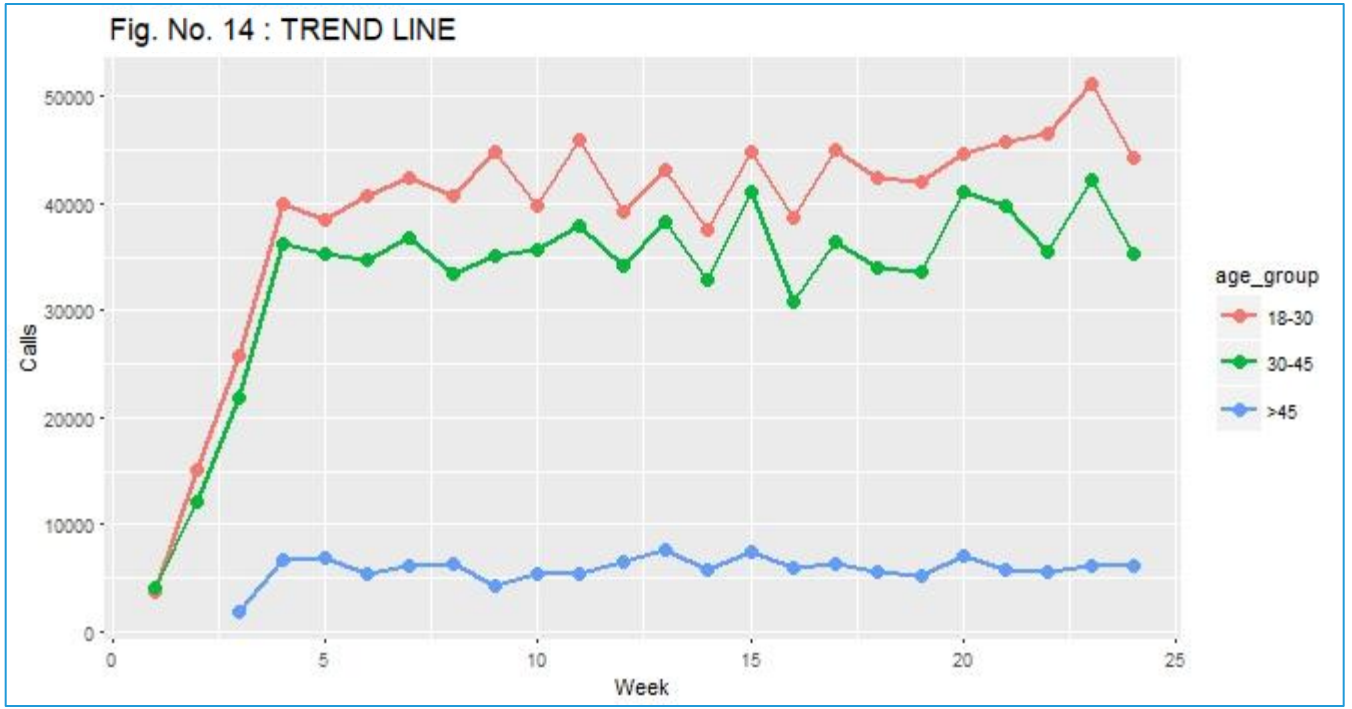
Trend Line in R

```
# Importing Data

demographic<-read.csv("TelecomData_CustDemo.csv", header=TRUE)
transaction<-read.csv("TelecomData_WeeklyData.csv", header=TRUE)

# Merging and Formatting Data
# Creating new variable Age_Group & aggregating

working<-merge(demographic, transaction, by=("CustID"),all=TRUE)
working$Age_Group<-cut(working$Age, breaks= c(0,30,45,Inf), labels= c("18-30","30-45",">45"))
trend<-aggregate(Calls~Week+Age_Group, data=working, FUN=sum)

# Observing Age_group wise Trend

ggplot(trend, aes(x=Week, y=Calls, colour=Age_Group))+
geom_line(size=1)+ geom_point(size=3)+labs(y="Calls", title="Fig. No. 14 : TREND LINE")
```

- **geom_line()** is used to call the trend line
- **geom_point()** is used to plot the data points

Trend Line in R

```
# Output
```



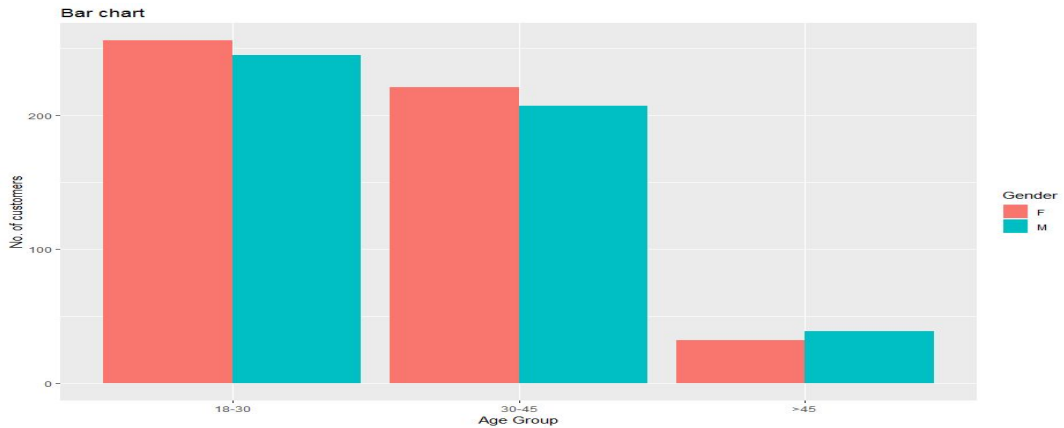
Get an Edge!

Multiple Bar Chart in R (CAUTION)

```
ggplot(telecom, aes(x=Age_Group,fill=Gender))+geom_bar(position="dodge",
fill="darkorange")+labs(x="Age Group", y="No. of customers", title="Bar
chart")
```

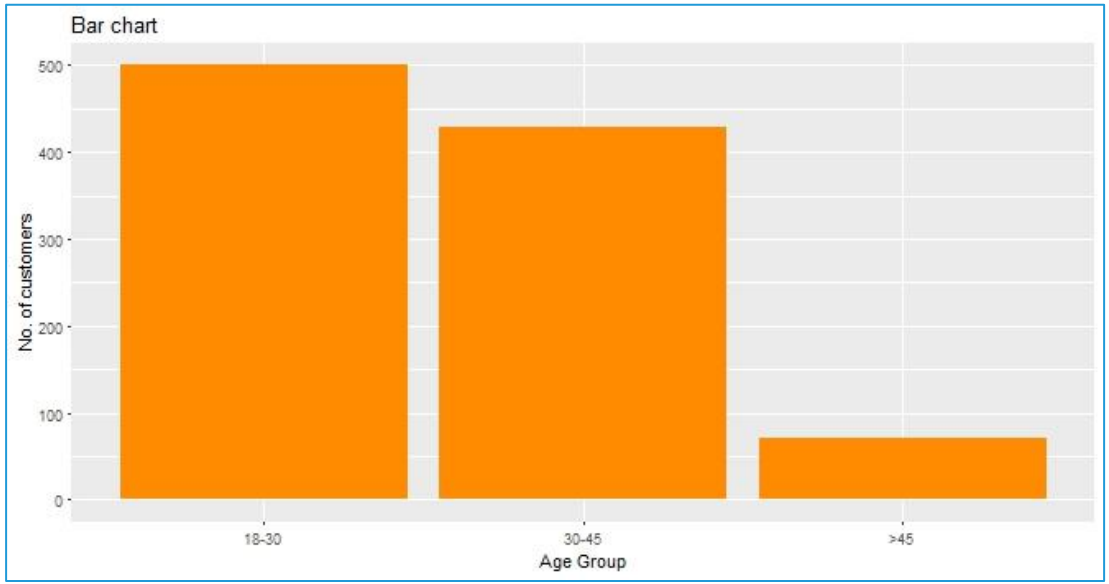
Caution: fill="darkorange" in geom_bar() overrides the fill=Gender in ggplot()

So instead of getting this output :



Get an Edge!

You get this output only because of fill = “darkorange” argument in geom_bar() function.



Quick Recap

Using ggplot package in R

i. Bar Charts
ii. Pie Chart
iii. Box-Whisker Plot
iv. Histogram
v. Scatterplot with Regression Line
vi. Trend Line