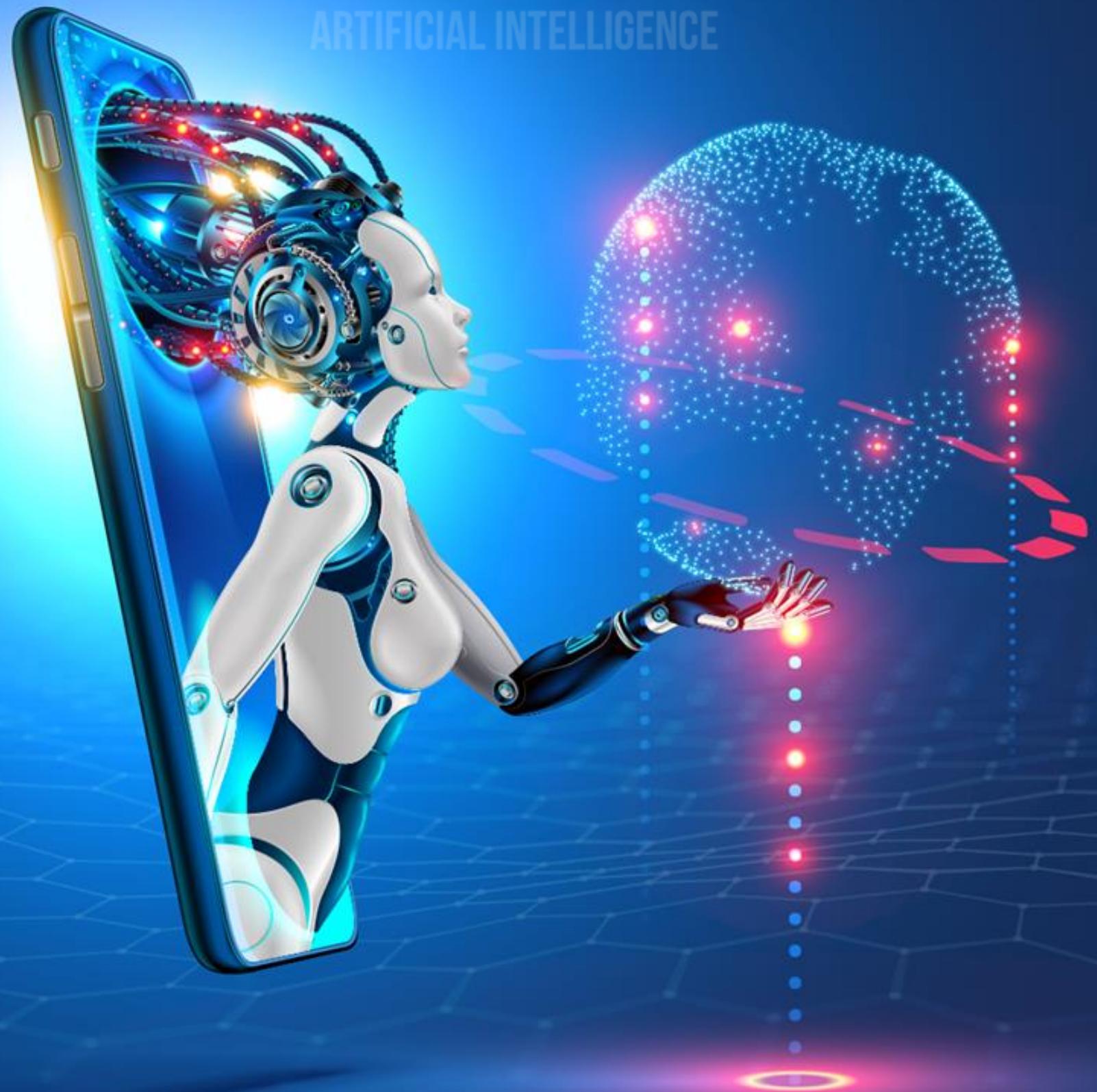


DATA AND ARTIFICIAL INTELLIGENCE



Data Analytics with R

DATA AND ARTIFICIAL INTELLIGENCE

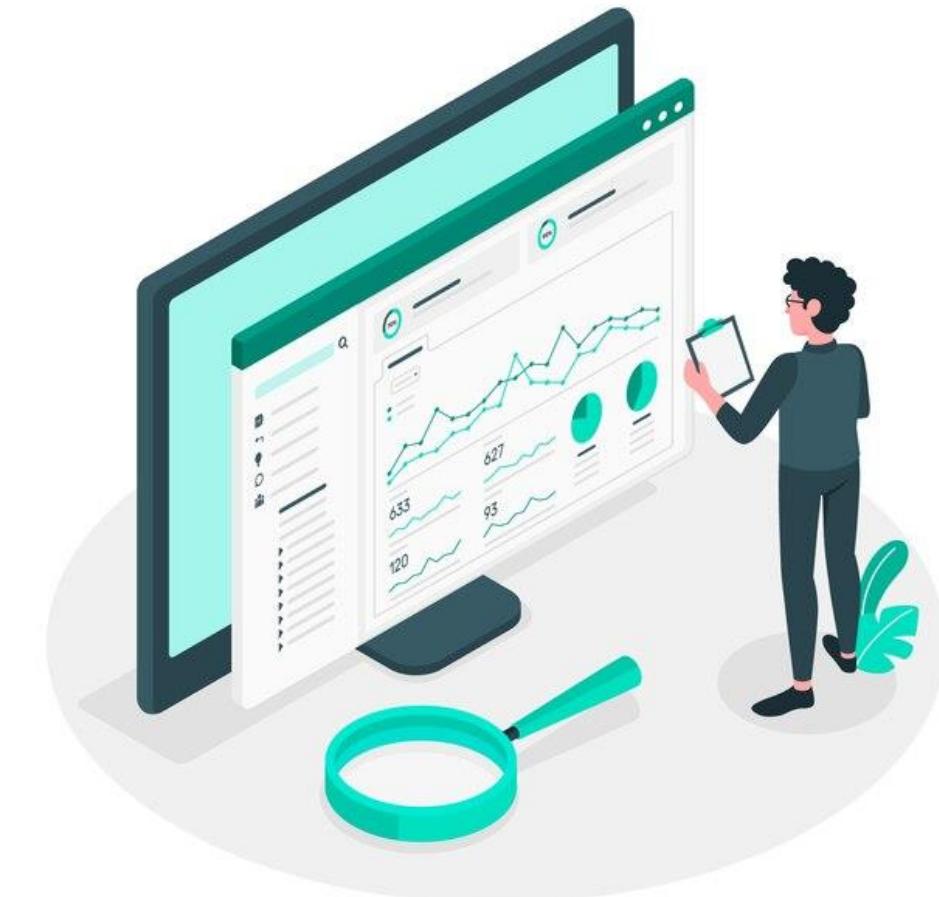


Data Visualization in R

Business Scenario

- Anna has been hired by a firm for a reporting role. The firm uses R for all its analytical purposes. She has to create monthly reports with data in the most useful and easily understandable format for the management.

Approach: To excel at the job, she needs to understand and learn various methods of creating impactful visualizations using the given data.



Learning Objectives

By the end of this lesson, you will be able to:

- Explain the importance of data visualization
- Create different plots in R using the base packages
- Use ggplot2 package to create plots for different types of data
- Export R plots to different file formats



Need for Data Visualization

Why Is Data Visualization Important?

Data visualization is representation of huge strips of data in a visual format, such as graphs, charts, and maps.



Data visualizations provide a quick and effective way of communicating the insights the data is holding.



Data visualization strengthens the impact by allowing the comprehension of a large amount of data at a glance.

“The simple graph has brought more information to the data analyst’s mind than any other device.” — John Tukey

Why R for Data Visualization?

R is designed for statistical computing and data analysis.



It provides flexibility and minimum coding for the creation of high-quality visuals for impactful communication of the insights.



It helps to prepare any type of graph, chart, or visualization.

Packages for Data Visualization

In addition to the "graphics" package, R also offers packages for visualization that include:

Function	Description
ggplot2	To create elegant data visualization using 'The Grammar of Graphics'
plotly	To create interactive web graphics
lattice	To implement trellis graphs in R
rgl	To create 3D visualizations and data plots
plotrix	To use various plotting functions
leaflet	To create interactive web maps
shiny	To build interactive web apps straight from R for visualization of data models

Visual Analysis of Data in R

Different kind of data requires different plots for representation. Analysis may be:

Univariate

It is used to analyze patterns or describing the data in a single variable.

Bi-variate

It is used to analyze two variables to determine the empirical relationship between them.

Multivariate

It is used to analyze more than two variables to determine the empirical relationship between them.

Based on the type of data, such as categorical or continuous, the type of chart is decided in each analysis.

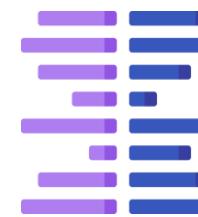
Creating Plots in R

Plots in R

Data visualization can be done using:



Pie chart



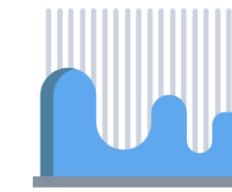
Bar plot



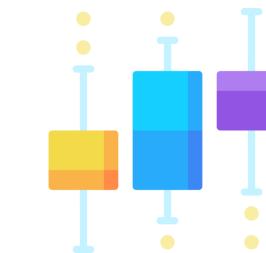
Histogram



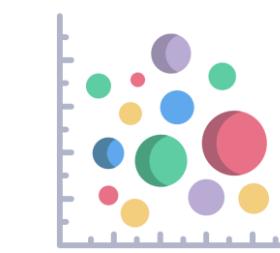
Line chart



Kernel density plot



Box plot



Scatter plot

Pie Chart

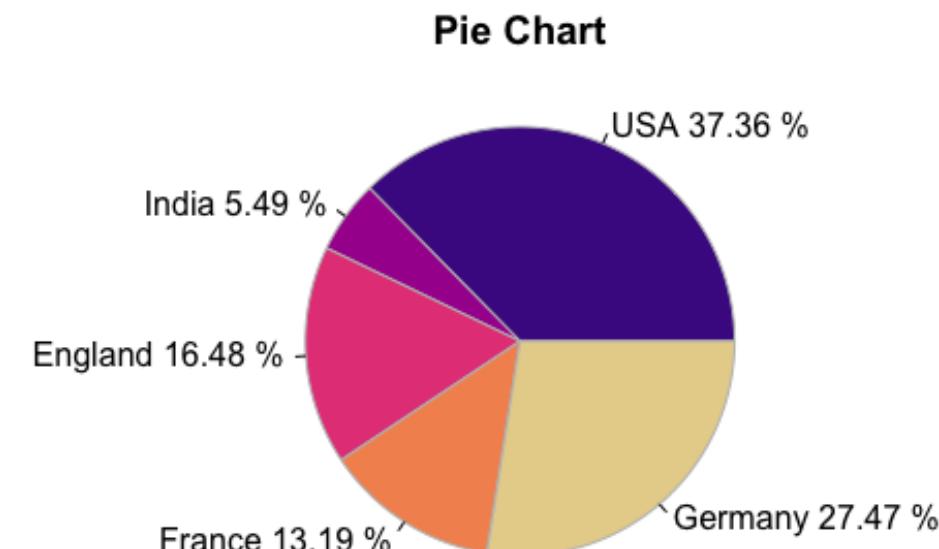
A pie chart is used to represent data in numerical proportions.

Pie chart in R is created using `pie()` function.

Code :

```
> # consider the data for no of participants from different  
countries in a symposium.  
> country <- c('USA', 'India', 'England', 'France', 'Germany' )  
> count <- c(34, 5, 15, 12, 25 )  
> # calculating percentage participation  
> perc <- round(count/sum(count)* 100, 2)  
> # add frequency or proportion to country names to create labels  
> labels <- paste(country, perc,'%')  
> pie(count, labels = labels,  
+ radius = 1, col = hcl.colors(n = 5, palette = 'ag_Sunset'),  
+ border = 'gray', main = "Pie Chart"  
+ )
```

Output:



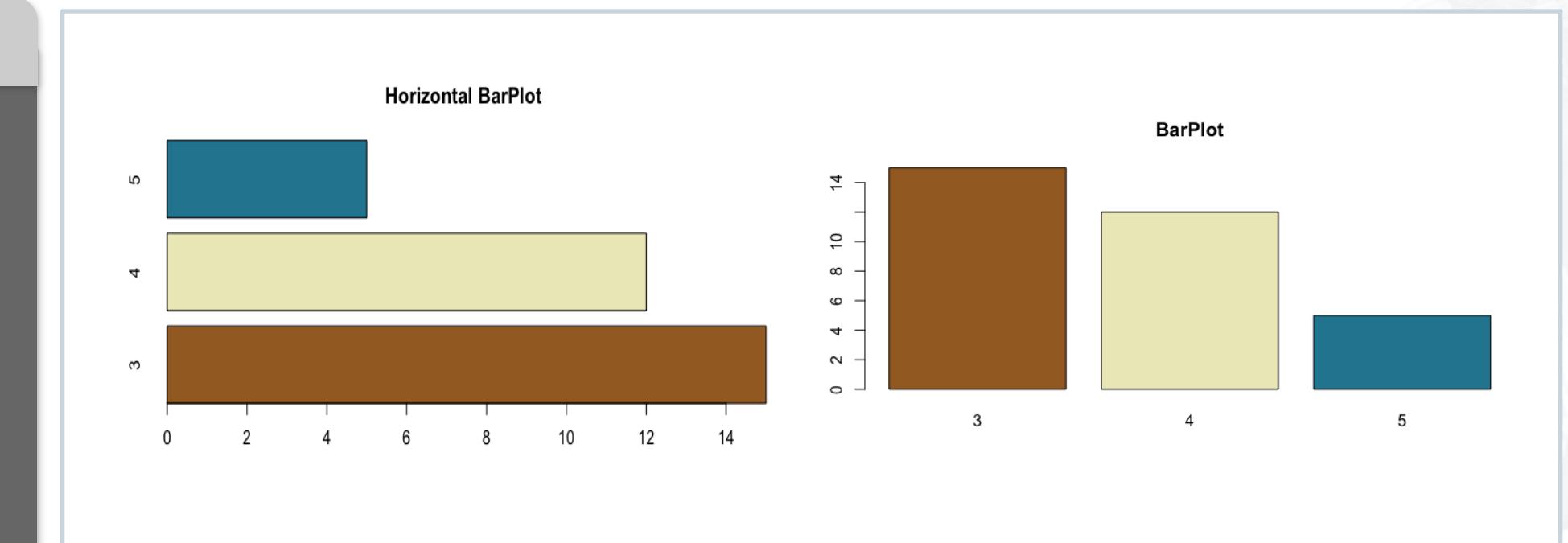
Bar Chart

A bar chart is used to present the frequency of categorical data using vertical or horizontal bars.

Code:

```
> # consider the no. of gears in mtcars data
> par(mfrow = c(1,2))
> freq_table <- table(mtcars$gear)
> # for horizontal barchart
> barplot(freq_table,
+ col = hcl.colors(3, palette = 'Earth'),
+ main = 'Horizontal Bar Chart', horiz = TRUE)
> # for vertical barchart
> barplot(freq_table,
+ col = hcl.colors(3, palette = 'Earth'),
+ main = 'Vertical Bar Chart')
```

Output:



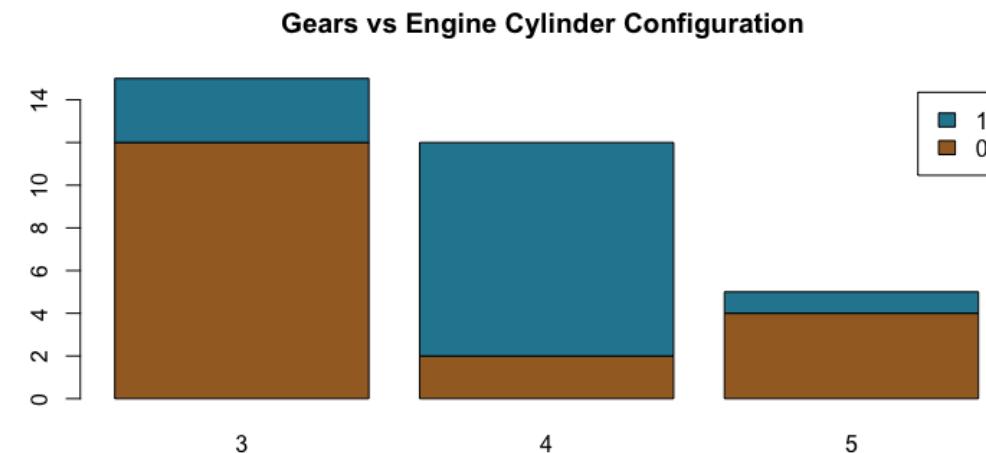
Bar Chart

Bar charts can also be used to perform bi-variate analysis as stacked bar charts or grouped bar charts.

Code :

```
> # stacked bar chart for no. of gears vs V/S variable  
> barplot(table(mtcars$vs, mtcars$gear),  
+ col = hcl.colors(2, palette = 'Earth'),  
+ main = 'Gears vs Engine Cylinder Configuration ',  
+ legend = unique(mtcars$vs), beside = FALSE)
```

Output :



Histogram

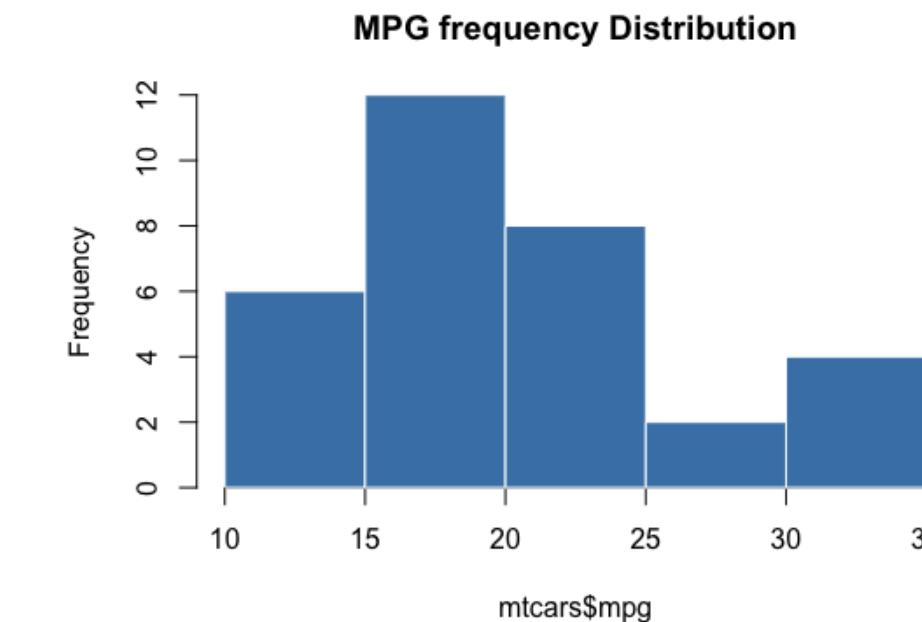
Histogram represents frequency distribution of a continuous variable using vertical bars.

The data is arranged into bins or ranges.

Code :

```
> # histogram showing frequency distribution of  
"mpg".  
> hist(mtcars$mpg,  
+ main = 'MPG frequency Distribution',  
+ col = 'steelblue', border = 'white')
```

Output :



Kernel Density Plots

Density plots use Kernel Density Estimate to present the probability density function of a random variable.

```
> density(mtcars$mpg)

Call:
  density.default(x = mtcars$mpg)

Data: mtcars$mpg (32 obs.)      Bandwidth 'bw' = 2.477
      x          y
Min. : 2.97 Min. :6.481e-05
1st Qu.:12.56 1st Qu.:5.461e-03
Median :22.15 Median :1.926e-02
Mean  :22.15 Mean :2.604e-02
3rd Qu.:31.74 3rd Qu.:4.530e-02
Max. :41.33 Max. :6.795e-02
```

`density(x)` computes kernel density estimate for a random variable.

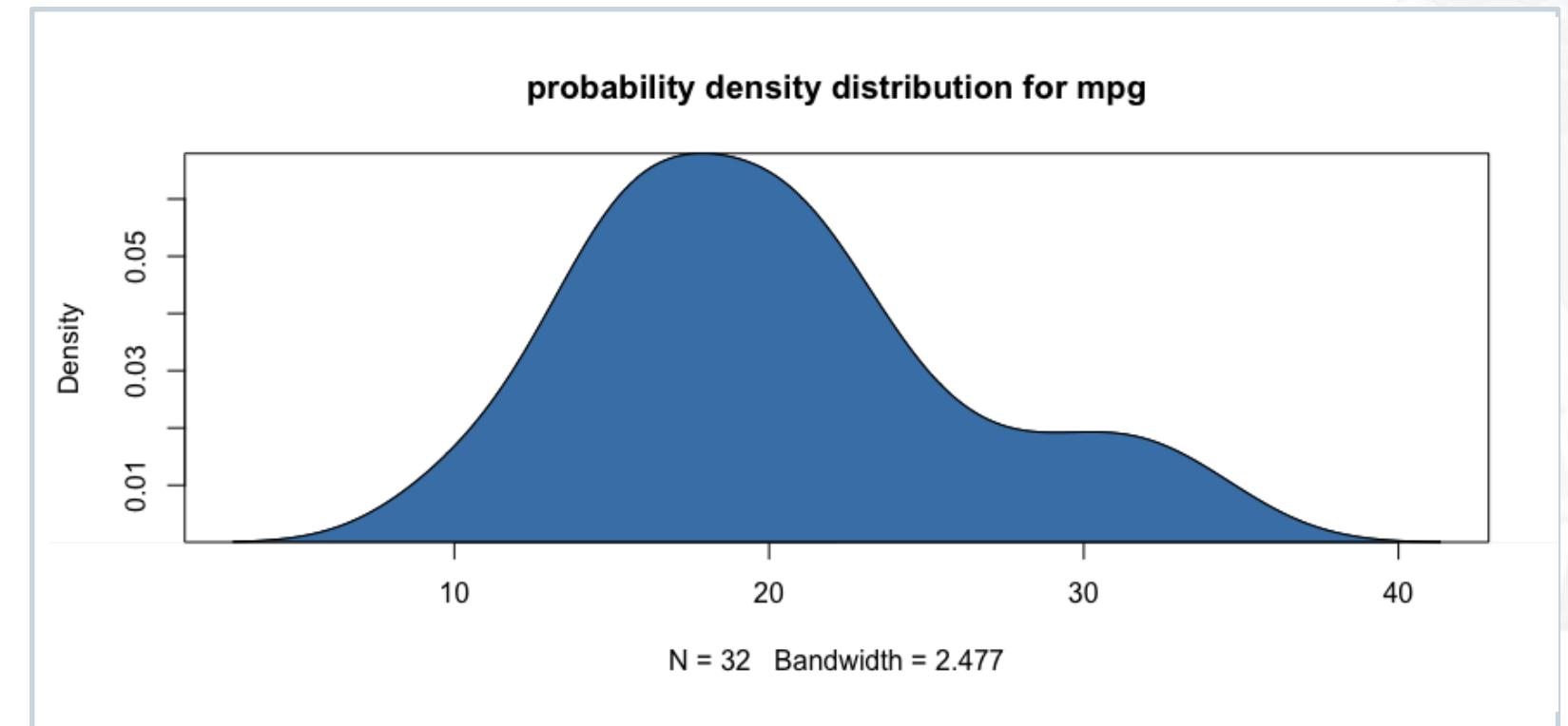
Kernel Density Plots

Polygon is an accessory to plot function as it helps to fill in the color.

Code :

```
> plot(density(mtcars$mpg),  
+ main = "probability density distribution for mpg")  
> polygon(density(mtcars$mpg), col = c('steelblue'))
```

Output :

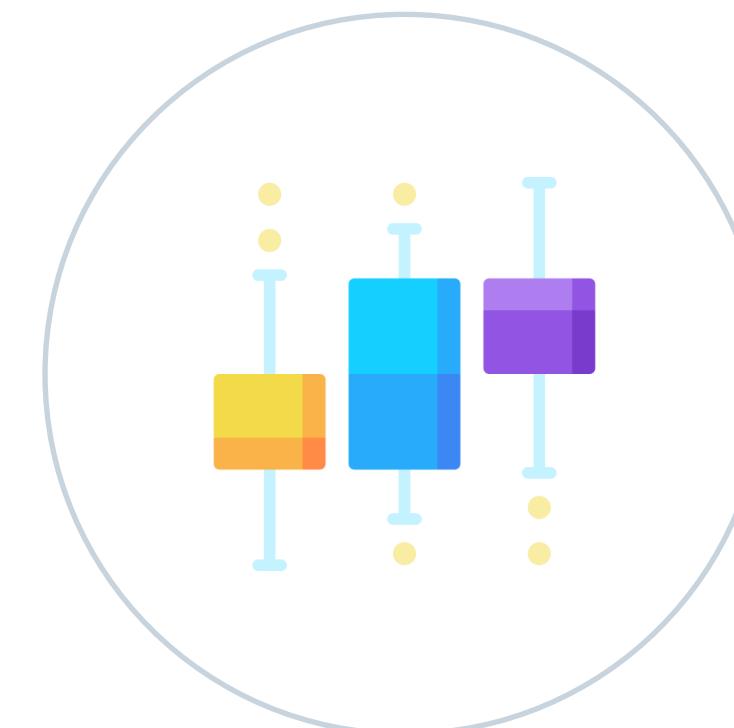


Box and Whisker Plot

A box and whisker plot is a convenient way of presenting data distributions.

Box represents the inter quartile range while the whiskers indicate variability outside the upper and lower quartiles.

Box and whisker plots present the outlier information in the form of individual dots that are in-line with the whiskers.

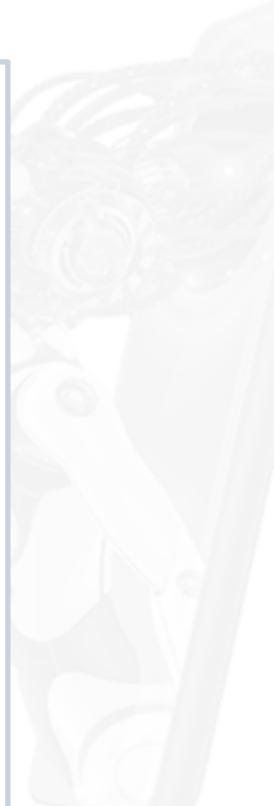
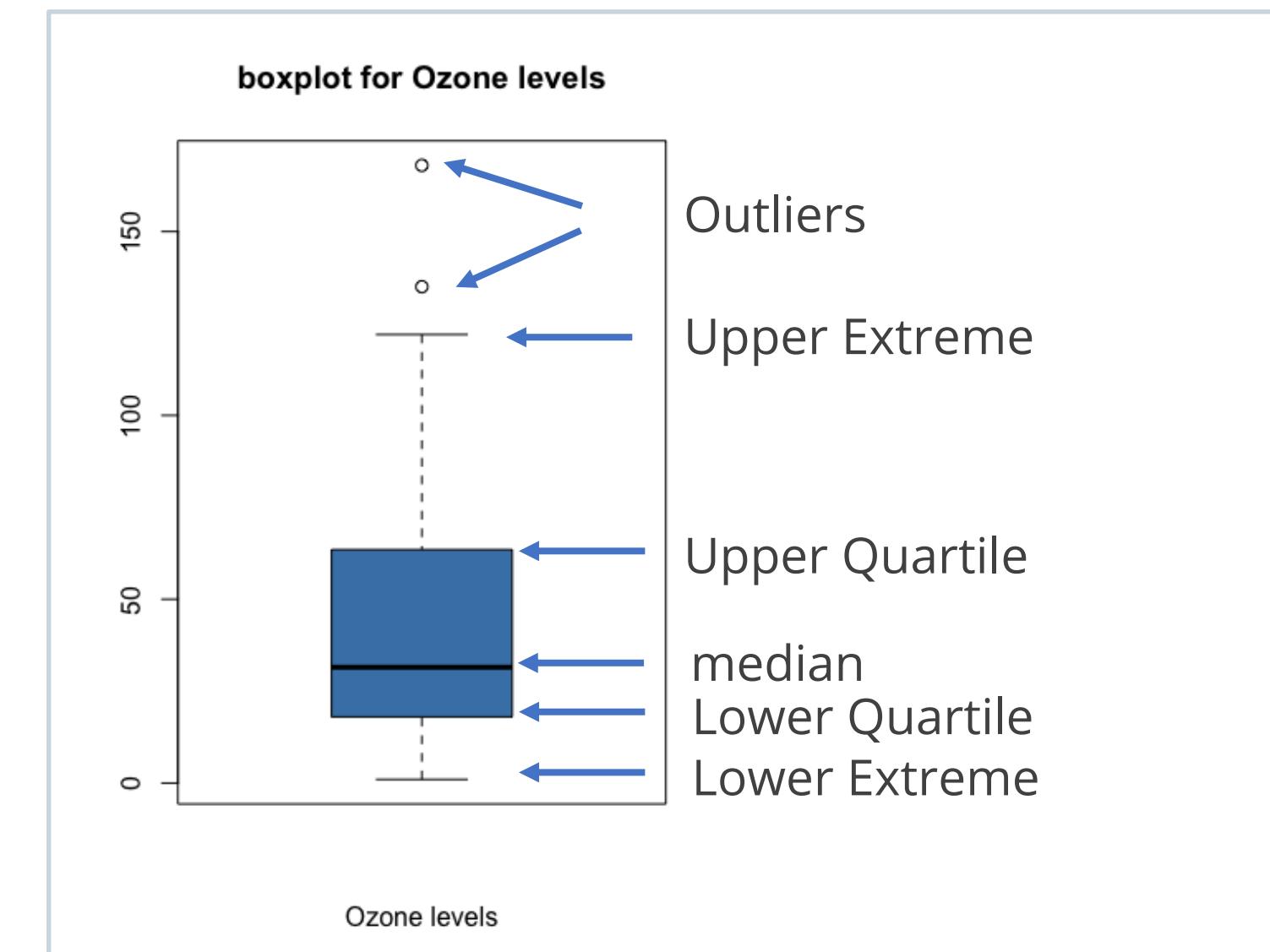


Box and Whisker Plot

Consider this boxplot for the Ozone levels in the "airquality" data available in R.

```
> boxplot(airquality$Ozone , col = 'steelblue',  
+ main = "boxplot for Ozone levels",  
+ xlab = "Ozone levels")
```

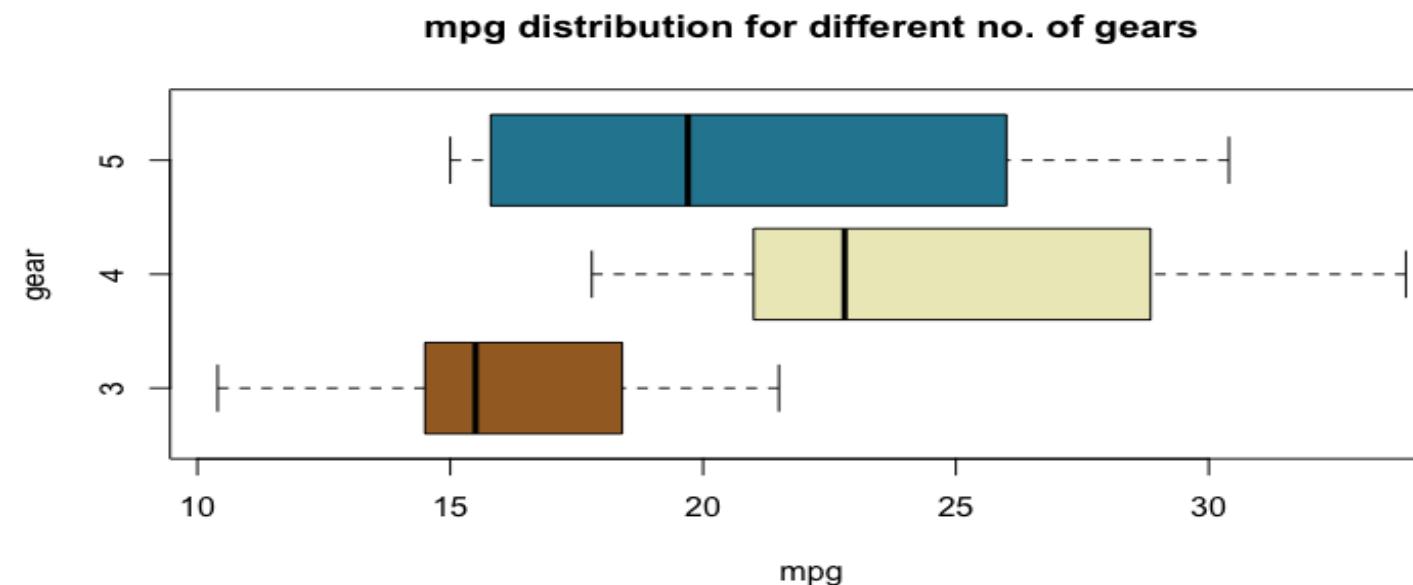
Anatomy of a BoxPlot



Comparison of Distribution for Multiple Groups

Data distribution can be studied across groups using a box plot.

```
> boxplot(mpg ~ gear, data = mtcars,  
+ col= hcl.colors(3, "Earth"),  
+ horizontal = TRUE,  
+ main="mpg distribution for different no. of gears")
```

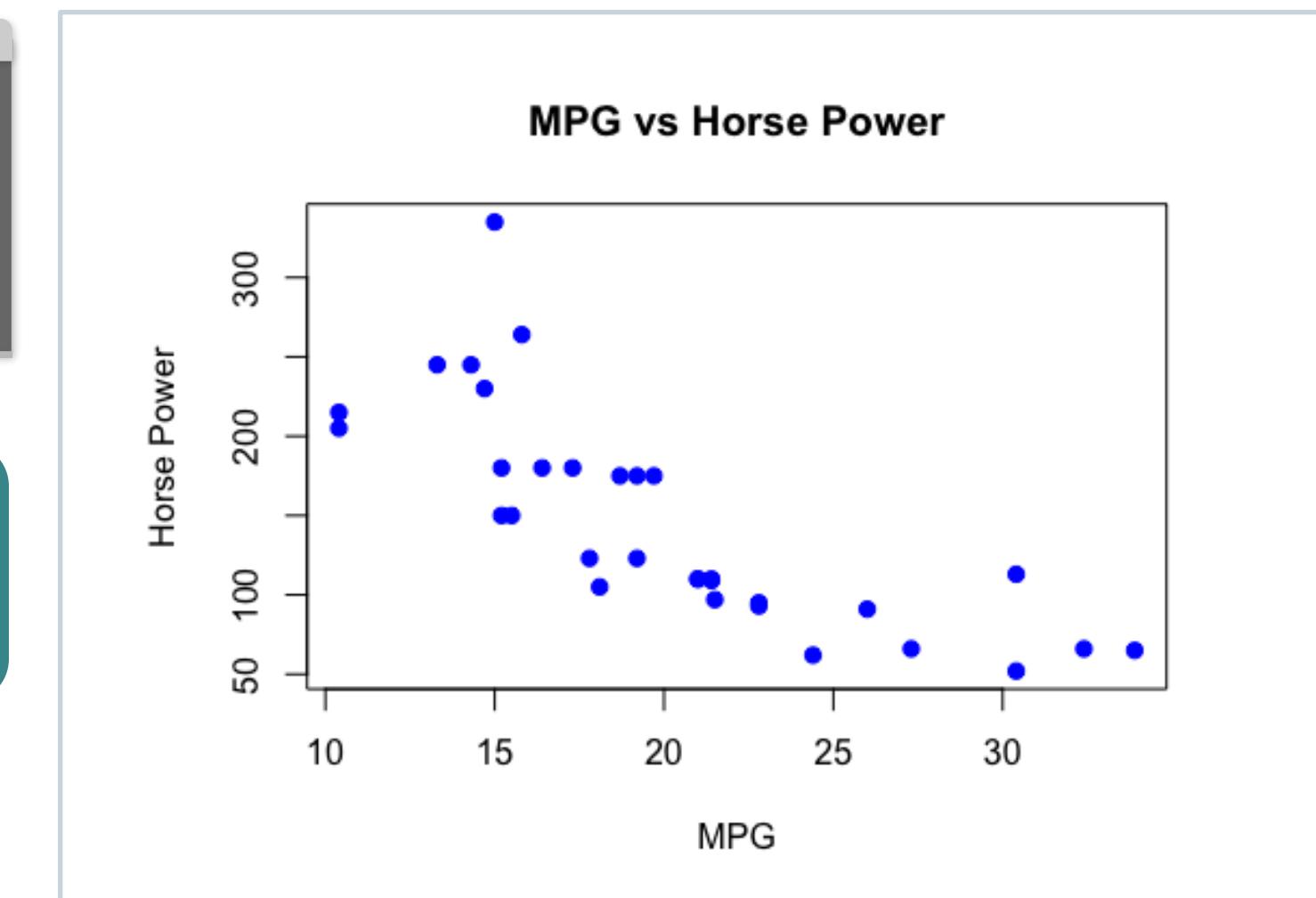


Scatter Plot

Scatter plot represents two numerical data variables in the form of points along the XY axis. It helps to understand the relationship between the variables used.

```
> plot(mtcars$mpg, mtcars$hp, pch = 19, col = 'blue',
+       xlab = " MPG ", ylab = "Horse Power",
+       main = "MPG vs Horse Power")
```

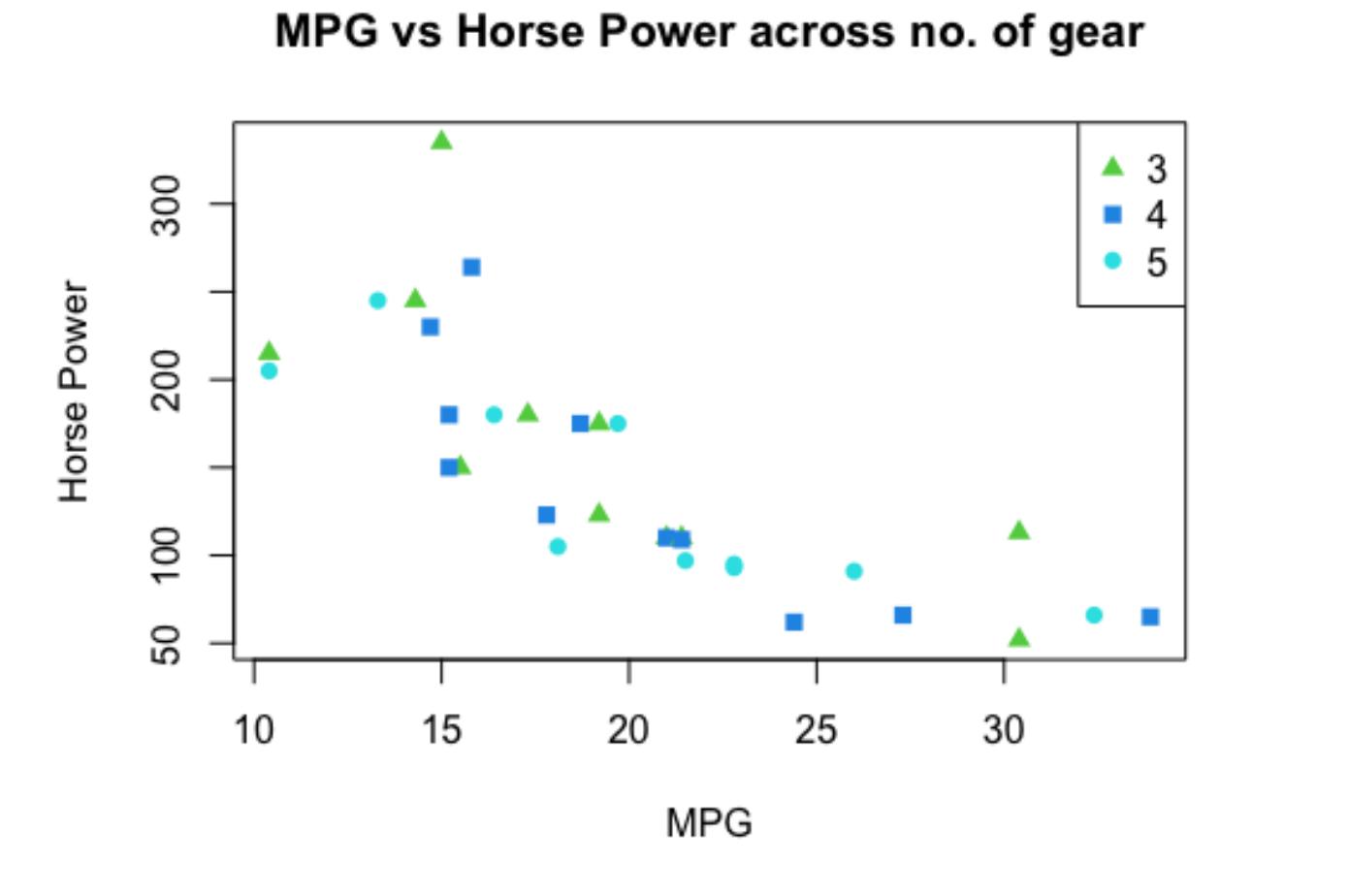
Mileage and horsepower are inversely related.



Scatter Plot

Data points in scatter plot can be differentiated for a group by using color or plotting symbol ("pch" argument).

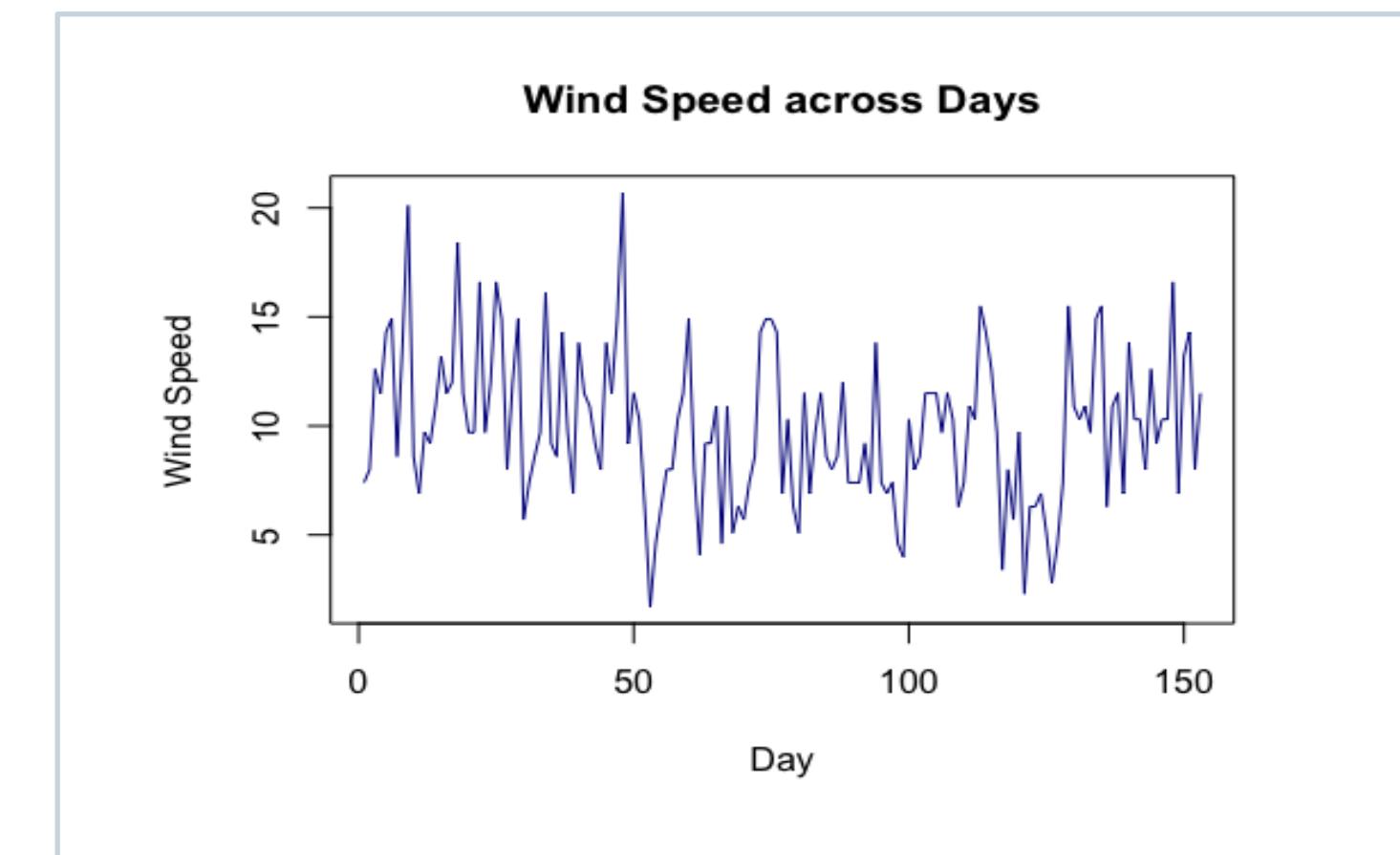
```
> mtcars$gear <- as.factor(mtcars$gear)
> plot(mtcars$mpg, mtcars$hp,
+ col = levels(mtcars$gear), pch = c(17, 15, 16),
+ xlab = " MPG ", ylab = "Horse Power",
+ main = "MPG vs Horse Power across no. of gear")
> legend("topright", legend = levels(mtcars$gear),
+ col = levels(mtcars$gear), pch = c(17, 15, 16))
```



Line Chart

A line chart represents a series of data points connected through a continuous line.

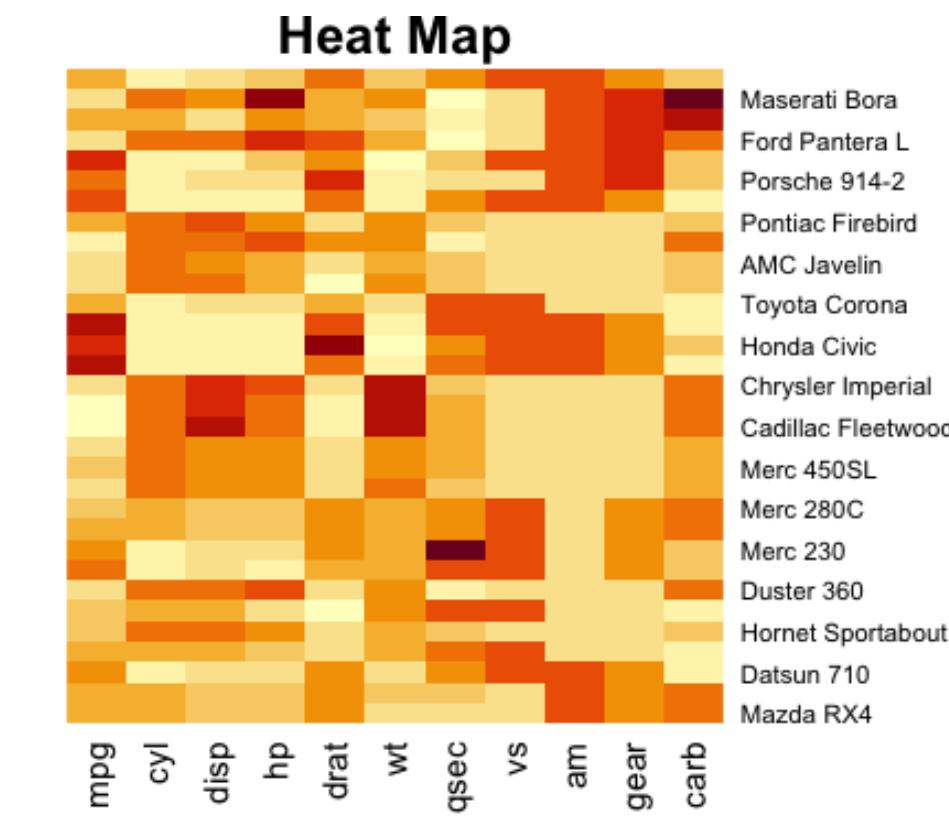
```
> wind_data <- na.omit(airquality$Wind)
> plot(wind_data, type = 'l',
+       ylab = 'Wind Speed',
+       xlab = 'Day',
+       main = 'Wind Speed across Days',
+       col = 'dark blue')
```



Heat Map

Heat map is a technique that presents data in two-dimensional format where colors represent the magnitude of the variables.

```
> heatmap(as.matrix(mtcars),  
+ Colv = NA, Rowv = NA,  
+ main = "Heat Map", scale = "column")
```



Word Cloud

Word cloud is a graphical representation of word frequency in text data. It is a mining technique that allows a quick look at the most frequent keywords.



In R the function to create a word cloud is available in the *wordcloud* package.

Syntax:

```
wordcloud(words, freq)
```

words and freq argument specify the words and their corresponding frequencies respectively

Accessory Functions for Plot

Function	Description
legend()	Adds legend to the plots
abline()	Adds straight lines to the plots
lines()	Joins coordinate using line segments
text()	Adds strings to the plot
points()	Draws a sequence of points at the specified coordinates to existing plot

Creating a Word Cloud



Duration: 5 minutes

Problem Scenario: Consider the data from a restaurant aggregator company. A researcher wants to use this data to look at the restaurant chain having the most franchises.

The researcher wants to use a word cloud as a quick way to achieve that.

Create a word cloud to help the researcher achieve their desired results.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

ggplot for Plotting

What is ggplot2?

ggplot2 is a data visualization package of R that provides a general scheme for data visualization.

It breaks up graphs into semantic components such as scales and layers. It is an alternative for the basic graphics of R.



Building a ggplot2 Graph

ggplot2 builds graphs in layers. It divides the plot into three parts:

Data

It is the dataframe that contains data to be plotted.

Aesthetics

These are the variables mapping to the visual properties of the plot.

Geometry

It refers to the type of graph that is used, such as bar graph or histogram.

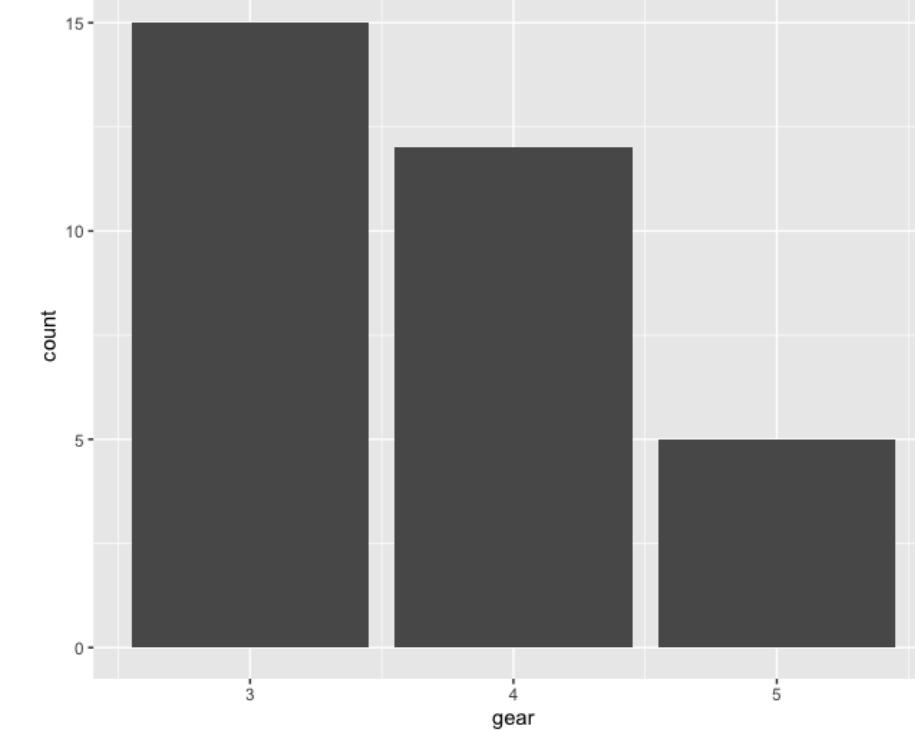
Building a ggplot2 Graph

Building a basic bar plot using ggplot():

```
> library(ggplot2)
> ggplot(mtcars, aes( x = gear)) +
+ geom_bar()
```

Geometry

Data and aesthetics

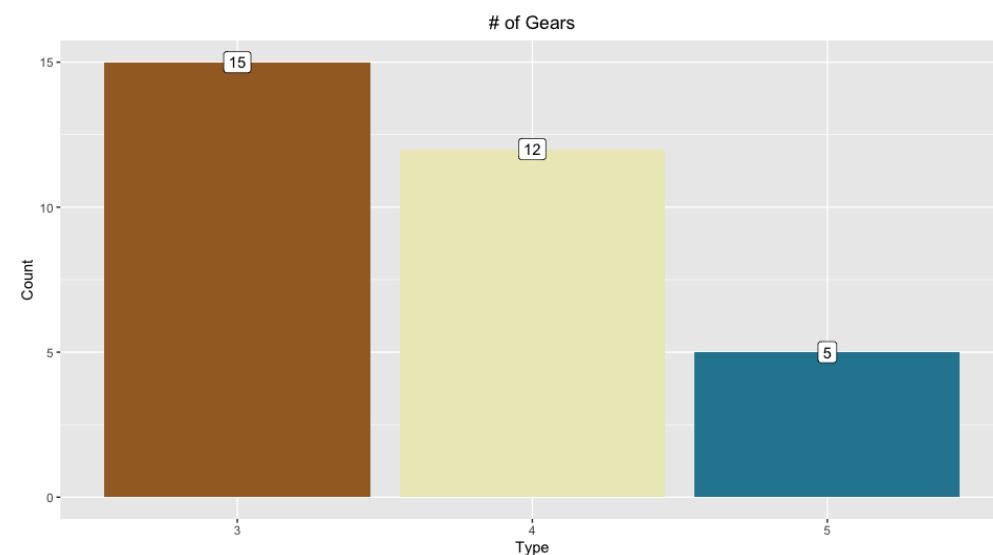


Bar Plot Using ggplot2

For a color-coded and visually pleasing chart, add more elements as shown:

```
> freq_df<- as.data.frame(table(mtcars$gear))  
> ggplot(freq_df, aes( x = Var1, y = Freq)) +  
+ geom_bar(stat = 'identity',fill = hcl.colors(3, 'Earth'))+  
+ labs(x = 'Type', y = 'Count')+  
+ ggtitle('Bar Plot') +  
+ theme(plot.title = element_text(hjust = 0.5) )+  
+ geom_label(aes(label =Freq, y = Freq ), fill = 'white', colour ='black')
```

- Frequency data as dataframe
- Data and aesthetics
- Geometry
- For x-axis label and y-axis label
- For title of the chart
- Location or placement of the title
- For assigning labels to each bar



Geometry Functions of ggplot2

Multiple functions are available in R to create different geometry in ggplot2 package.

Plot type	Syntax with default arguments	Special note
Bar plot	geom_bar(stat = "count", position = "stack")	stat : Is used to identify the mode (use "identity" when using frequency data) position : Specifies stack or dodge for the bar plot in multivariate
Pie chart	coord_polar(theta = "x", start = 0, direction = 1) Used with geom_bar()	theta : Is a variable to map angle to x or y start : Offsets starting point from 12 o'clock in radians direction : Uses 1 for clockwise or -1 for anticlockwise
Histogram	geom_histogram(bins)	bins : Specifies number of bins or ranges

Geometry Functions of ggplot2

Multiple functions are available in R to create different geometry in ggplot2 package.

Plot type	Syntax with default arguments	Special Notes
Density Plot	geom_density(kernel = "gaussian")	Kernel: Specifies the distribution which may be "gaussian", "rectangular", "triangular", "cosine", "optcosine", etc.
Box & Whisker plot	geom_boxplot(coef = 1.5, orientation = NA)	Coef: Length of the whiskers as multiple of IQR Orientation: Orientation of the layer "x" or "y"
Scatter plot	geom_point(aes(shape, color, size))	These optional aesthetics can be assigned variables from the data to decide color shape & size of the plotting character.
Line plot	geom_line() and geom_step()	geom_line() is best suited for time series data. geom_step() can be used to study the change in a variable.



Grouped Density Distribution Plot



Duration: 5 minutes

Problem Statement: Create a density distribution plot for the Mileage (mpg) variable of the **mtcars** data using the ggplot2 package.

Also, study the density distribution of mileage across different number of gears.

Note: Please download the data set and the solution document from the **Course Resources** section and follow the steps given in the document

ASSISTED PRACTICE

File Formats of Graphic Outputs

File Formats of Graphic Outputs

pdf("filename.pdf")
#PDF file

bmp("filename.bmp")
#BMP file

postscript("filename.ps")
#PostScript file

jpeg("filename.jpg")
#JPEG file

win.metafile("filename.wmf")
#Windows metafile

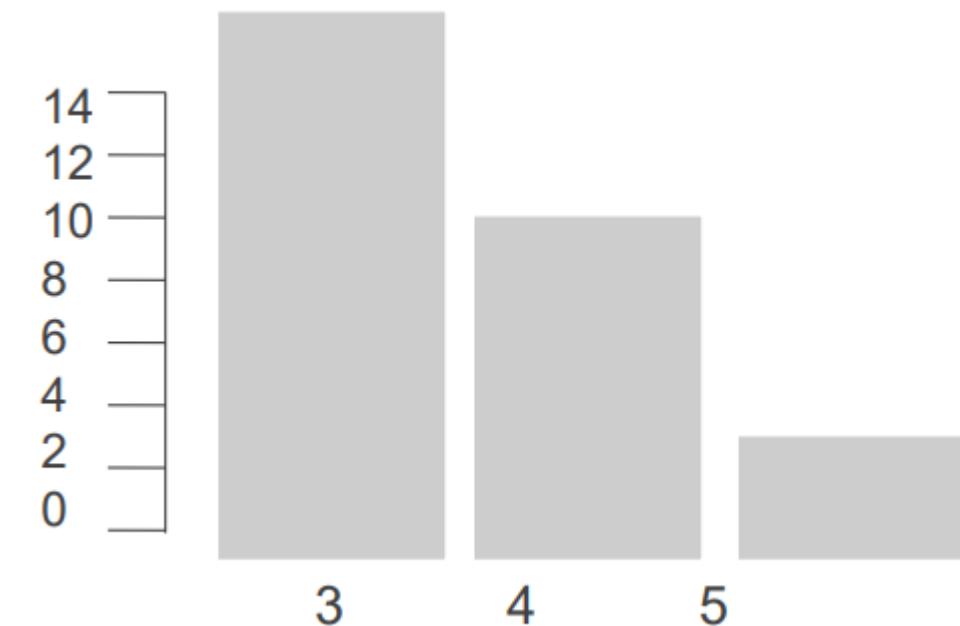
png("filename.png")
#PNG file



Saving a Graphic Output as a File

Example: To save a graphic output as a file, the following code can be used.

```
> # creating a bar plot and saving it as an image.  
> jpeg("myplot.jpg")  
> counts <- table(mtcars$gear)  
> barplot(counts)  
> dev.off()
```



The `dev.off()` function returns the control back to the terminal.

Key Takeaways

- Data visualization is the representation of huge strips of data in a visual format, such as graphs, charts, and maps.
- R provides multiple packages for creation of interactive plots.
- ggplot2 is one of the most used packages for data visualization in R and it builds plots in layers.
- R provides functions to export the charts and plots to external file formats, such as JPEG and PDF.



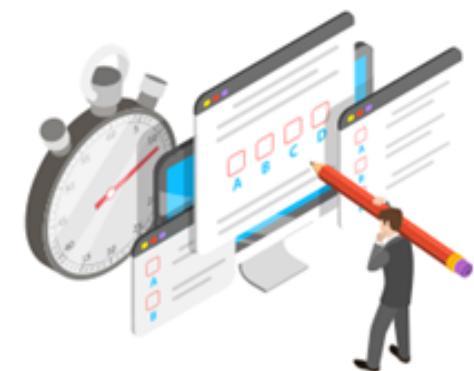


Knowledge Check

**Knowledge
Check
1**

Which of the following charts can be used to study distribution of continuous variable through their quartiles?

- A. Bar Plot
- B. Box Plot
- C. Scatter Plot
- D. Line Plot



**Knowledge
Check
1**

Which of the following charts can be used to study distribution of continuous variable through their quartiles?

- A. Bar Plot
- B. Box Plot
- C. Scatter Plot
- D. Line Plot



The correct answer is **B**

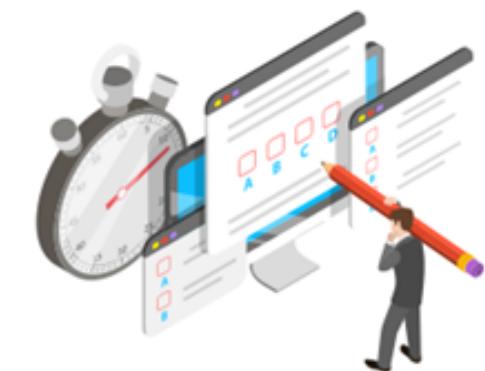
Box plot represents continuous data through its quartiles.

**Knowledge
Check**
2

Complete the below code to create a Scatter Plot.

ggplot(data, aes(x = V1, y = V2)) +

- A. geom_scatter()
- B. geom_scatter_plot()
- C. geom_point()
- D. geom_points()



**Knowledge
Check
2**

Complete the below code to create a Scatter Plot.

```
ggplot(data, aes(x = V1, y = V2)) +
```

- A. geom_scatter()
- B. geom_scatter_plot()
- C. geom_point()
- D. geom_points()



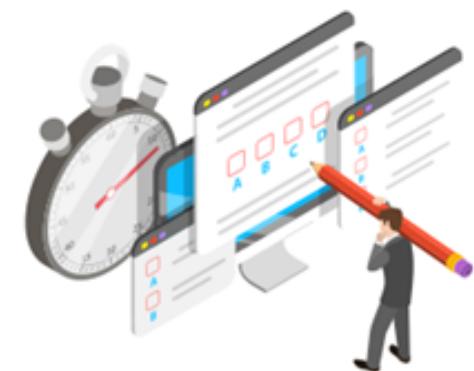
The correct answer is **C**

In ggplot2 package, geom_point() is used to create scatter plots.

**Knowledge
Check
3**

Graphic outputs can be saved as PNG files using _____.

- A. `save("filename.png")`
- B. `write.table("filename.png")`
- C. `write.file("filename.png")`
- D. `png("filename.png")`



**Knowledge
Check
3**

Graphic outputs can be saved as PNG files using _____.

- A. `save("filename.png")`
- B. `write.table("filename.png")`
- C. `write.file("filename.png")`
- D. `png("filename.png")`



The correct answer is **D**

`png("filename.png")` can be used to save plot in the PNG format.