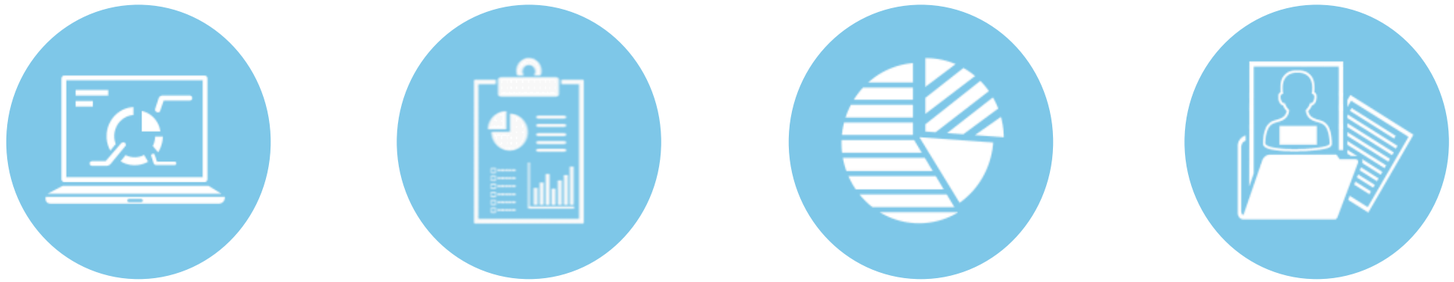
Data Science with R

Lesson 4—Data Visualization



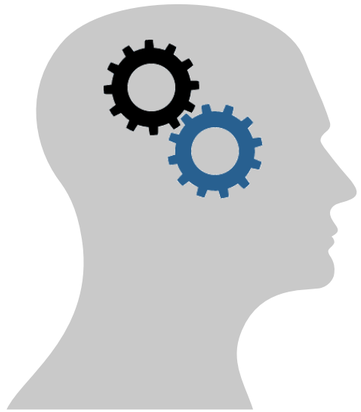
Learning Objectives

- Describe Data visualization
- List the graphics used for data visualization in R
- Explain ggplot with examples
- Discuss file formats of graphic outputs

Data Visualization

Topic 1—Introduction to Data Visualization

Solving Complex Challenges Using Data Visualization



GE specializes in solving complex challenges related to infrastructure, renewable energy, and affordable health care.

The marketing communications brand group was given the task of analyzing the causes of death of people.

https://hbr.org/2013/05/how-ge-uses-data-visualization

Solving Complex Challenges Using Data Visualization

The team separated people into groups based on gender and age.

The team used **data visualization** to simplify the information about causes of death for different age groups. For example, if you're in the age group of 24 to 36, you can use the table to understand the three things you are most likely to die of.

	15–24	25–34	35–44	45–54	55–64
WHITE	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 HEART DISEASE 2 CANCER 3 SUICIDE	1 HEART DISEASE 2 CANCER 3 SUICIDE	1 HEART DISEASE 2 CANCER 3 RESPIRATORY DISEASE
BLACK	1 SUICIDE 2 HEART DISEASE 3 CANCER	1 HIV 2 HEART DISEASE 3 SUICIDE	1 HEART DISEASE 2 HIV 3 CANCER	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 HEART DISEASE 2 CANCER 3 HYPERTENSION
ASIAN / PACIFIC ISLANDER	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 SUICIDE	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 HYPERTENSION
HISPANIC (ALL RACES)	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 HEART DISEASE 2 CANCER 3 HIV	1 HEART DISEASE 2 CANCER 3 LIVER DISEASE	1 CANCER 2 HEART DISEASE 3 LIVER DISEASE
NATIVE AMERICAN	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 SUICIDE 2 HEART DISEASE 3 LIVER DISEASE	1 HEART DISEASE 2 LIVER DISEASE 3 SUICIDE	1 CANCER 2 LIVER DISEASE 3 DIABETES	1 HEART DISEASE 2 CANCER 3 DIABETES
WHITE	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 CANCER 2 SUICIDE 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 SUICIDE	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 RESPIRATORY DISEASE
BLACK	1 HEART DISEASE 2 CANCER 3 HIV	1 HIV 2 CANCER 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 HIV	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 DIABETES
ASIAN / PACIFIC ISLANDER	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 CANCER 2 SUICIDE 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 HYPERTENSION
HISPANIC (ALL RACES)	1 CANCER 2 SUICIDE 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 SUICIDE	1 CANCER 2 HEART DISEASE 3 HIV	1 CANCER 2 HEART DISEASE 3 HYPERTENSION	1 CANCER 2 HEART DISEASE 3 DIABETES
NATIVE AMERICAN	1 SUICIDE 2 CANCER 3 HEART DISEASE	1 SUICIDE 2 CANCER 3 LIVER DISEASE	1 CANCER 2 LIVER DISEASE 3 HEART DISEASE	1 CANCER 2 HEART DISEASE 3 LIVER DISEASE	1 HEART DISEASE 2 DIABETES 3 LIVER DISEASE

SOURCE : Centers for Disease Control. NOTE These are the most likely causes of death for each race, sex, and age range in America, excepting homicide and accidental death. A collaboration between GOOD and Way Shape Form.

In partnership with

https://hbr.org/2013/05/how-ge-uses-data-visualization

What Is Data Visualization?

Data visualization is a modern equivalent of visual communication that involves the creation and study of the visual representation of data.

Data Visualization

Topic 2—Data Visualization using Graphics in R

Data Visualization in R

Data visualization in R can be done using the following graphics:

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Data Visualization in R

- Bar Plots
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Bar plots are horizontal or vertical bars used to show comparisons between categorical values. They represent length, frequency, or proportion of categorical values.

Syntax: *barplot(x)*

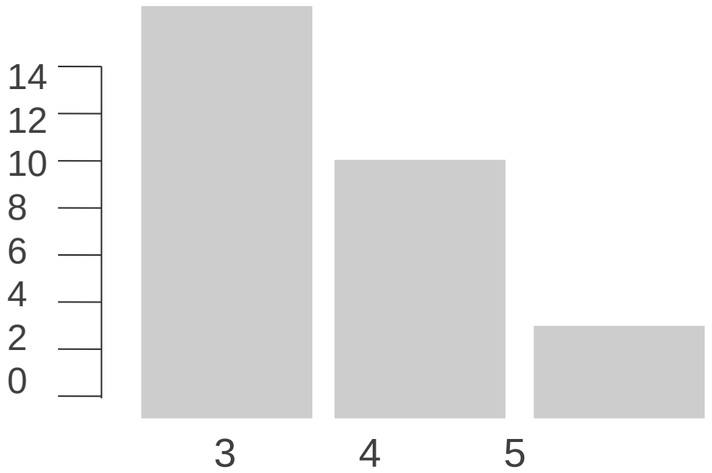
Data Visualization in R

CREATING BAR CHARTS IN R

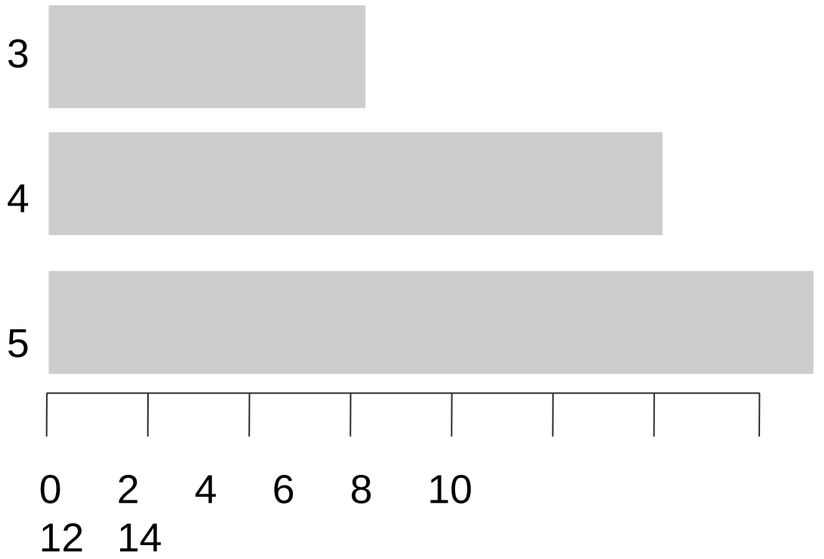
- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Use the mtcars dataset (inbuilt in R) to create simple and horizontal barplots:

```
counts <- table(mtcars$gear)
barplot(counts)
```




```
#horizontal bar chart
barplot(counts, horiz=TRUE)
```



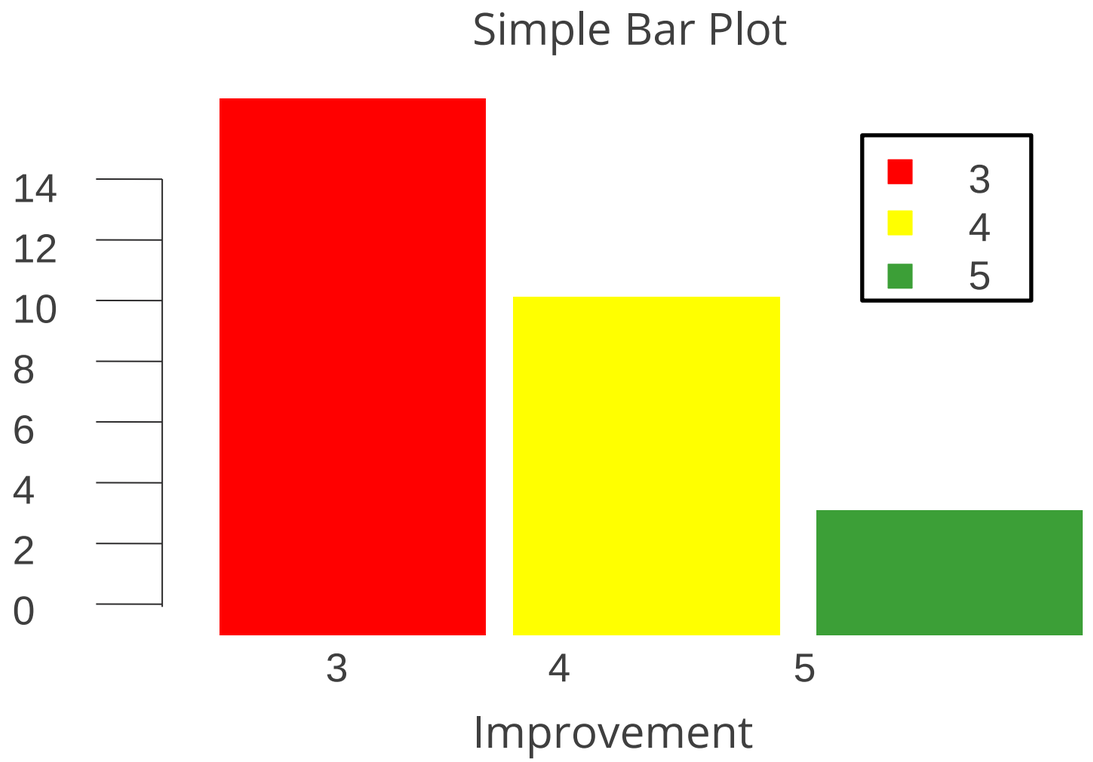
Data Visualization in R

EDITING BAR CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Titles, legends, and colors can be added to a simple bar chart using the following code:

```
counts <- table(mtcars$gear)
barplot(counts,
    main="Simple Bar Plot",
    xlab="Improvement",
    ylab="Frequency",
    legend=rownames(counts),
    col=c("red", "yellow", "green")
)
```



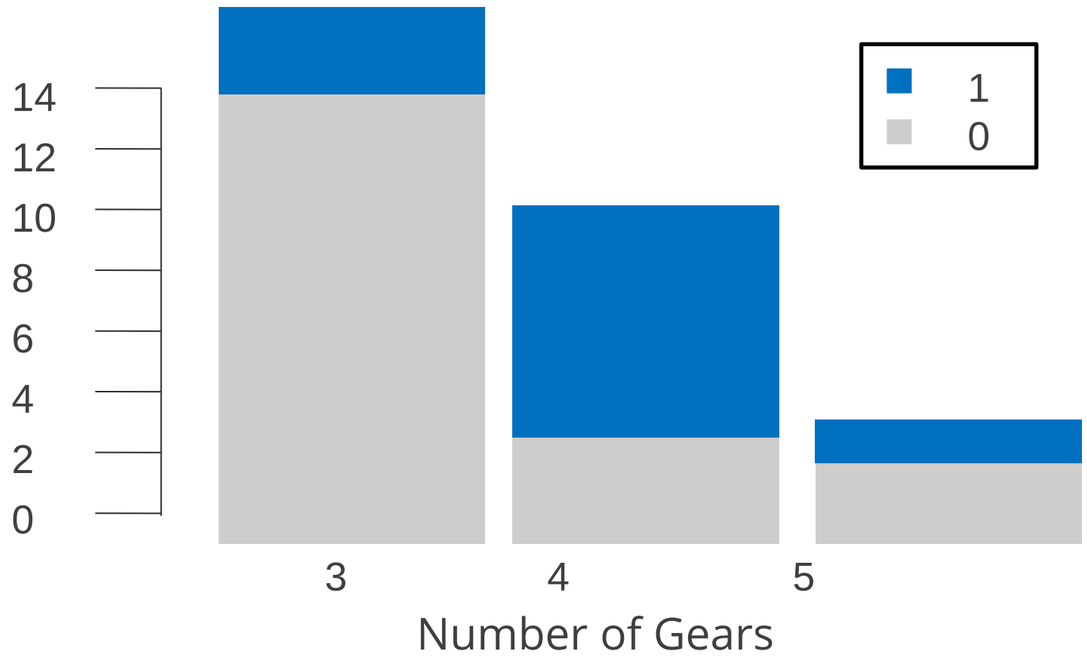
Data Visualization in R

EDITING BAR CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A stacked bar plot with colors and legends can be created using the following code:

```
counts <- table(mtcars$vs,
mtcars$gear)
barplot(counts,
    main="Car Distribution by Gears
and VS",
    xlab="Number of Gears",
    col=c("grey","cornflowerblue"),
    legend = rownames(counts))
```



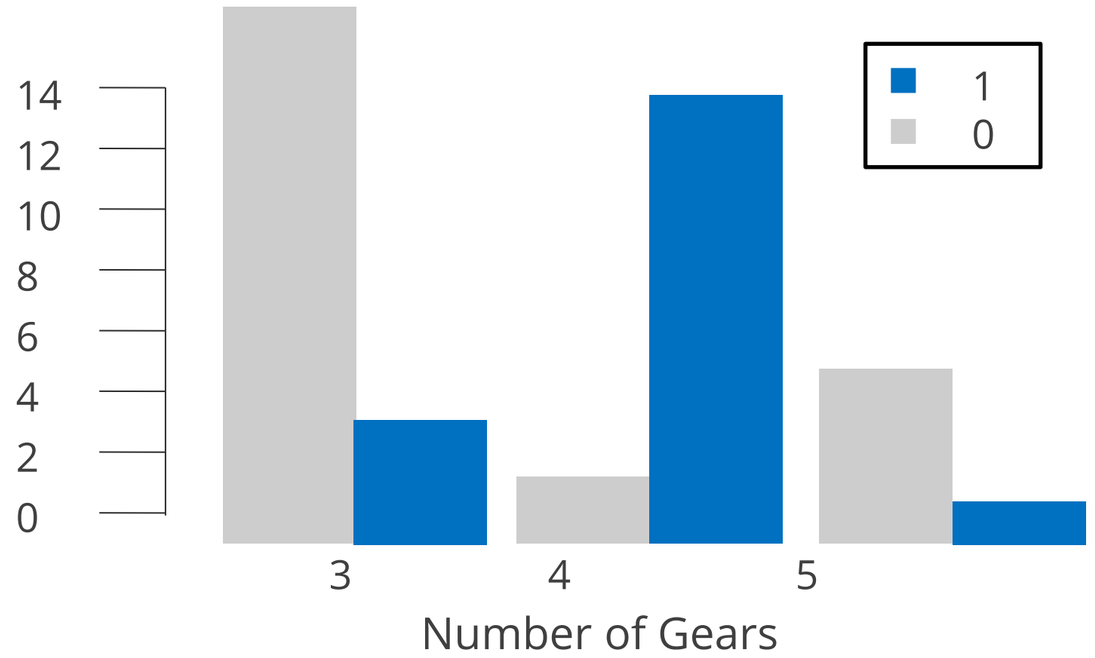
Data Visualization in R

EDITING BAR CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A grouped bar plot can be created using the following code:

```
counts <- table(mtcars$vs,
mtcars$gear)
barplot(counts,
   main="Car Distribution by
Gears and VS",
    xlab="Number of Gears",
   col=c("grey","cornflowerblue"
),
    legend = rownames(counts),
   beside=TRUE)
```



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A pie chart is a graph in which a circle is divided into sectors, each representing a proportion of the whole.

Syntax: *pie(attributes)*

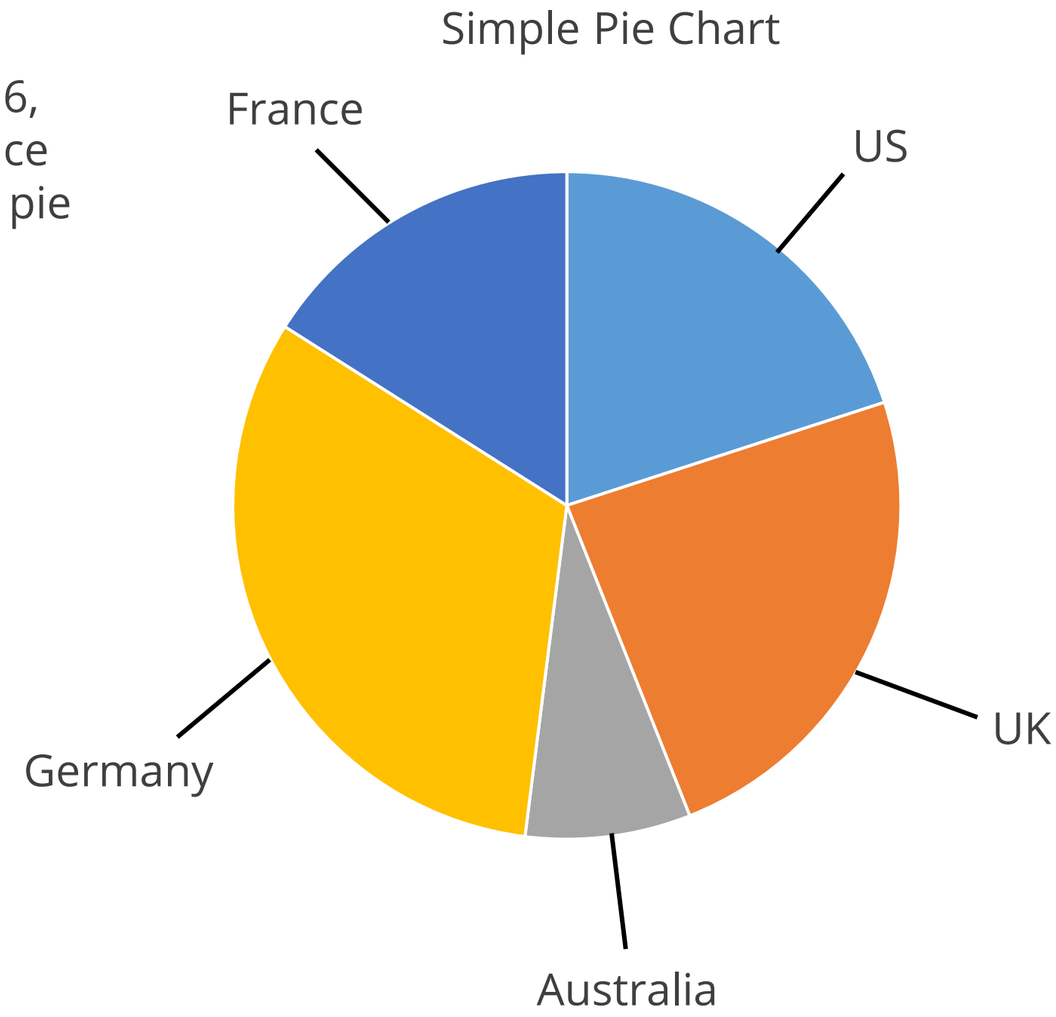
Data Visualization in R

CREATING PIE CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Consider a pie chart that contains values 10, 12, 4, 16, 8 as slices and US, UK, Australia, Germany, and France as labels. Use **pie(x, labels =)** function to create the pie chart:

```
slices <- c(10, 12,4, 16, 8)
lbls <- c("US", "UK",
"Australia", "Germany",
"France")
pie( slices, labels = lbls,
main="Simple Pie Chart")
```



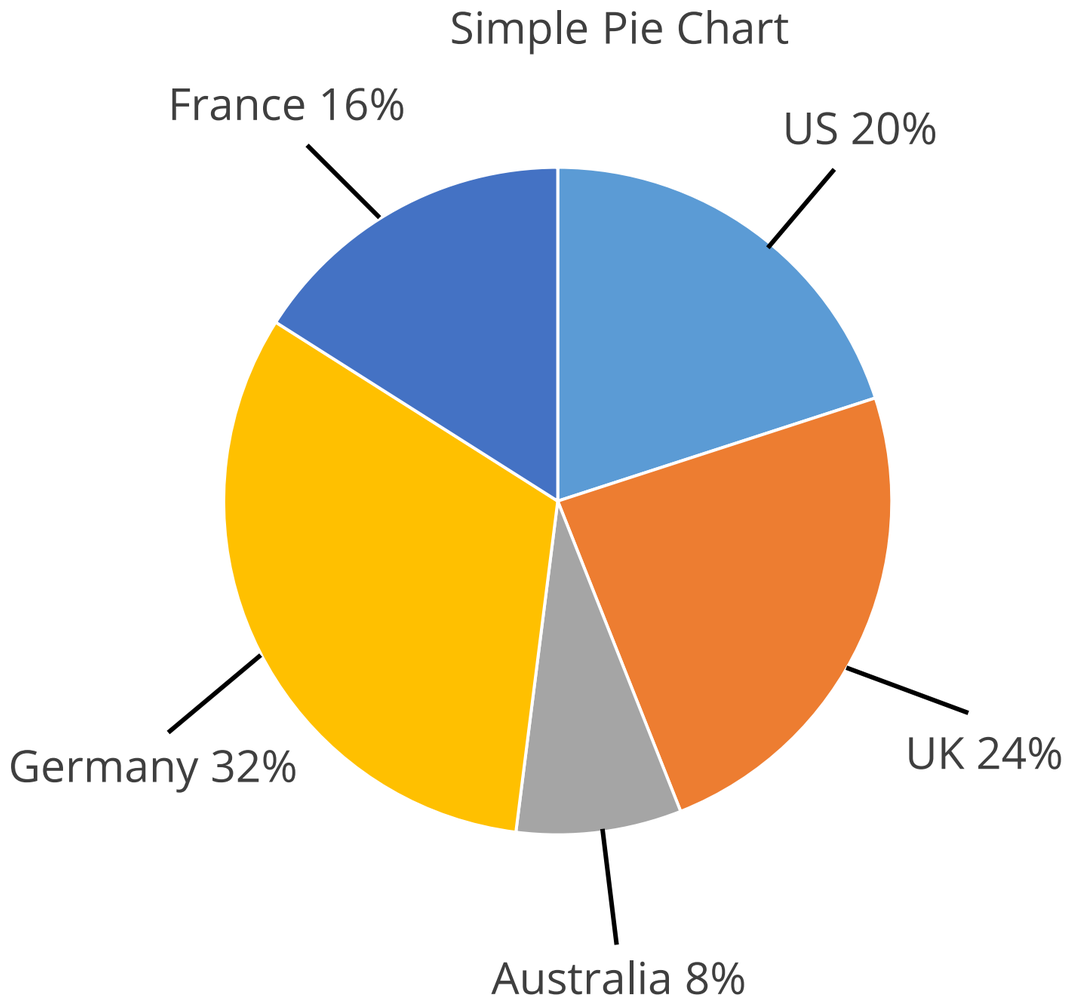
Data Visualization in R

EDITING PIE CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Percentages can be added to a pie chart using the following code:

```
slices <- c(10, 12,4, 16, 8)
pct <-
round(slices/sum(slices)*100)
lbls <- paste(c("US", "UK",
"Australia",
"Germany", "France"), " ", pct,
"%", sep="")
pie(slices, labels=lbls2,
col=rainbow(5),main="Pie Chart
with Percentages")
```



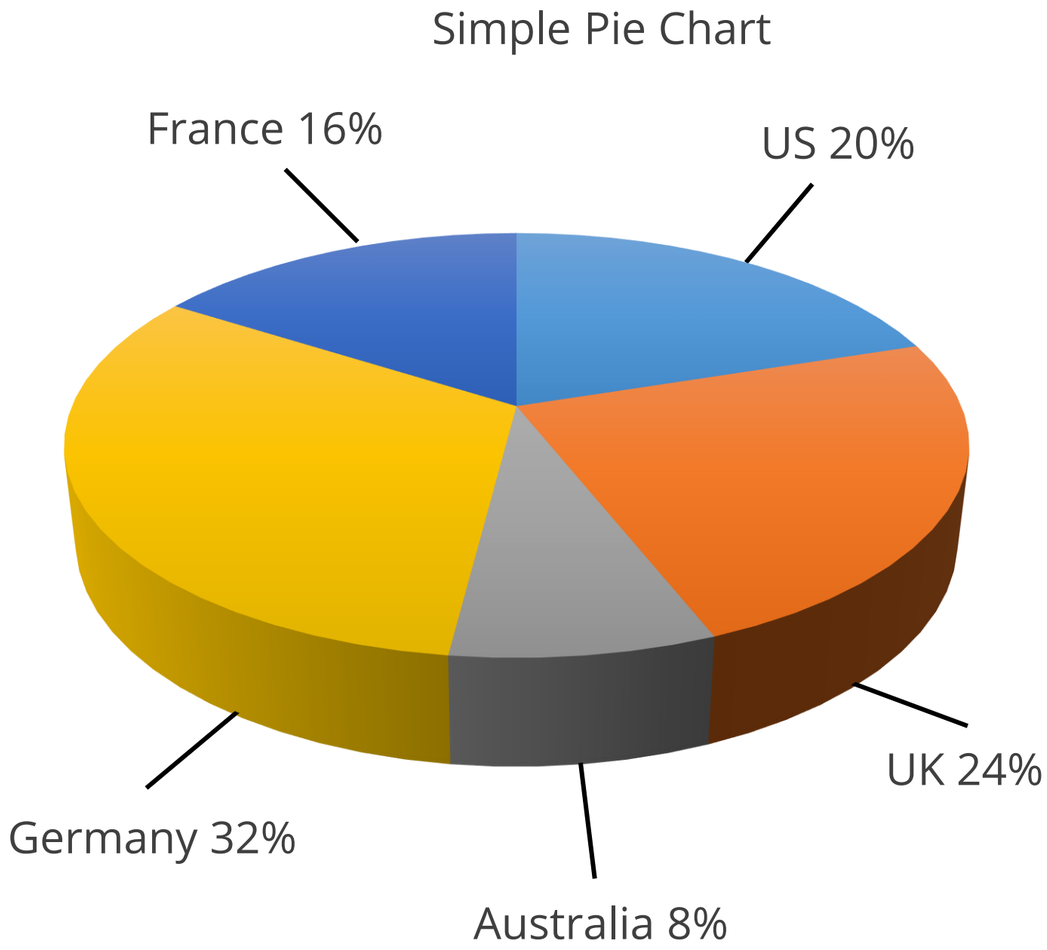
Data Visualization in R

EDITING PIE CHARTS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A 3-dimensional pie chart can be created as shown:

```
library(plotrix)
slices <- c(10, 12,4, 16, 8)
lbls <- paste(
c("US", "UK", "Australia",
"Germany", "France"),
 " ", pct, "%", sep="")
pie3D(slices,
labels=lbls,explode=0.0,
 main="3D Pie Chart")
```



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A histogram represents the distribution of a continuous variable and the frequency of values bucketed into ranges.

Syntax: *hist(x)*

Data Visualization in R

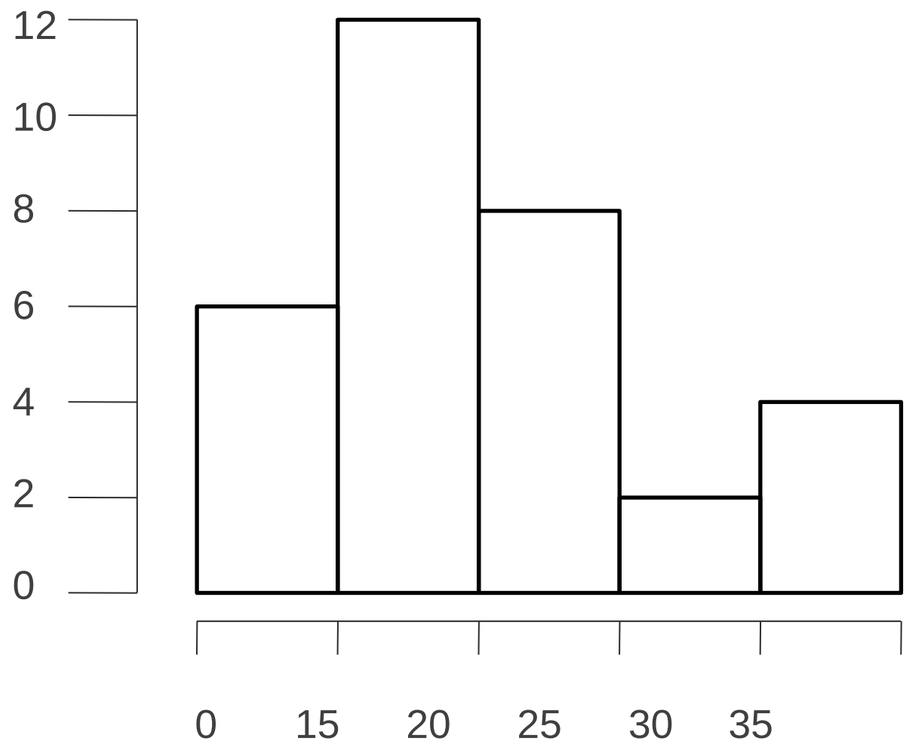
CREATING HISTOGRAMS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Creating a simple histogram using the mtcars dataset:

The first step is to “bin” the range of values, i.e., divide the entire range of values into a series of intervals and then count how many values fall into each interval. Next, use the following code:

```
mtcars$mpg #miles per gallon data
hist(mtcars$mpg)
```



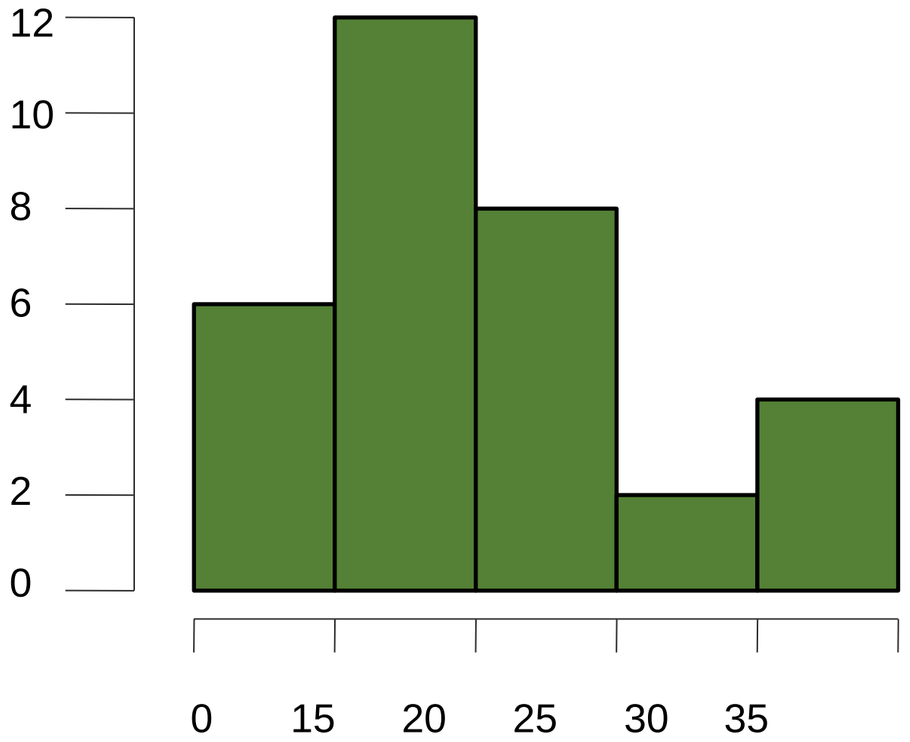
Data Visualization in R

EDITING HISTOGRAMS IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To color histograms with different number of bins, use the following code:

```
# Colored Histogram with
Different Number of Bins
hist(mtcars$mpg, breaks=8,
col="darkgreen")
```



The function **break =** controls the number of bin.

Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A Kernel density plot shows the distribution of a continuous variable.

Syntax: *plot(density(x))*

The Histogram is not a great method for determining the shape of a distribution because it depends on the number of bins used. To aid this, Kernel density plots are used over histograms.

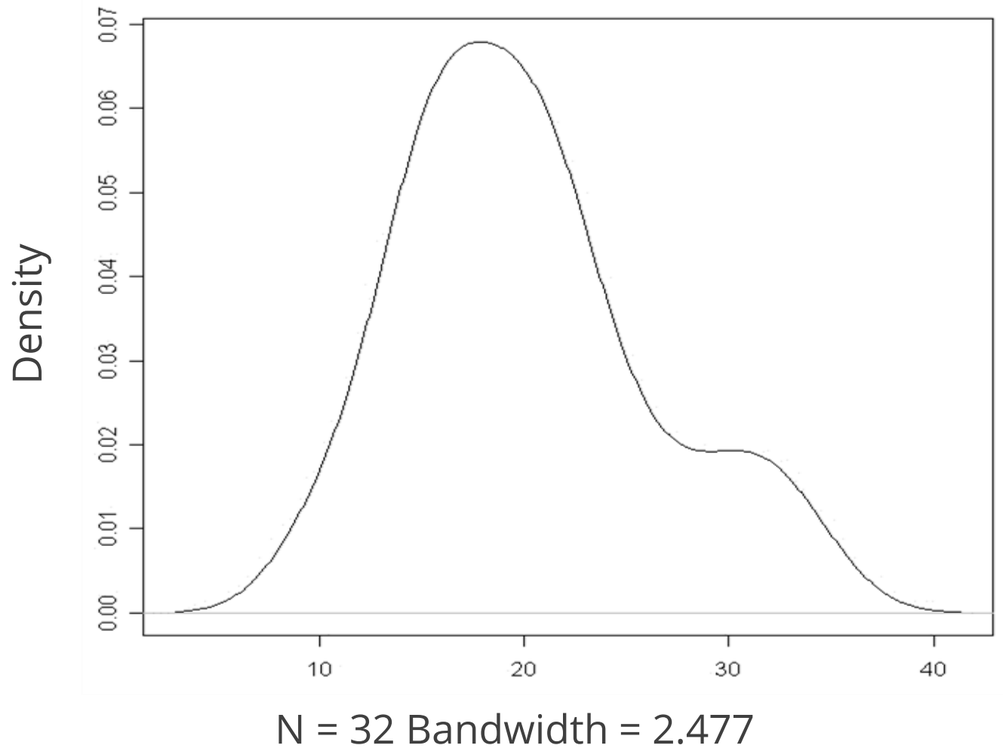
Data Visualization in R

CREATING A KERNEL DENSITY PLOT IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

The plot can be created using **plot(density(x)),** where x is a numeric vector. Use the mtcars dataset in R.

```
# kernel Density Plot
density_data <- density(mtcars$mpg)
plot(density_data)
```



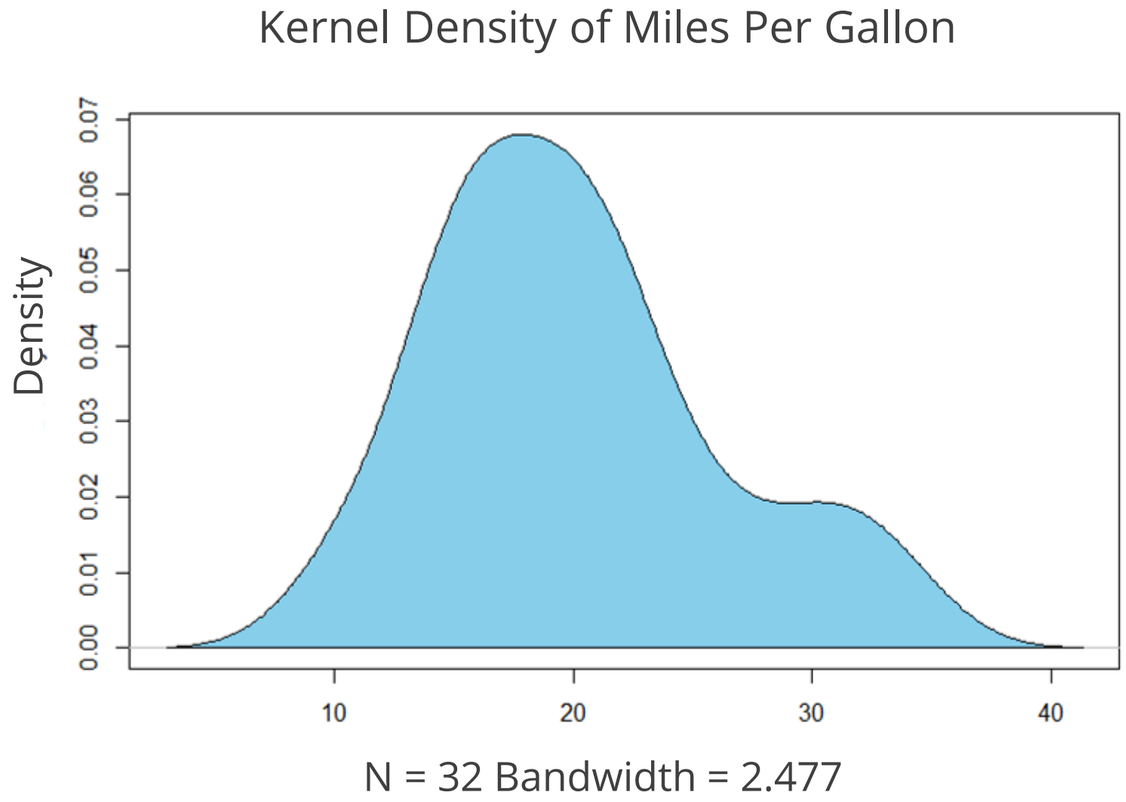
Data Visualization in R

EDITING A KERNEL DENSITY PLOT IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To add color and border to the plot, use the following codes:

```
# Filling density Plot with
color
density_data <-
density(mtcars$mpg)
plot(density_data, main="Kernel
Density of Miles Per Gallon")
polygon(density_data,
col="skyblue", border="black")
```



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A Line chart is used to represent a series of data points connected by a straight line. It helps visualize data that changes over time.

Syntax: *lines(x, y,type=)*

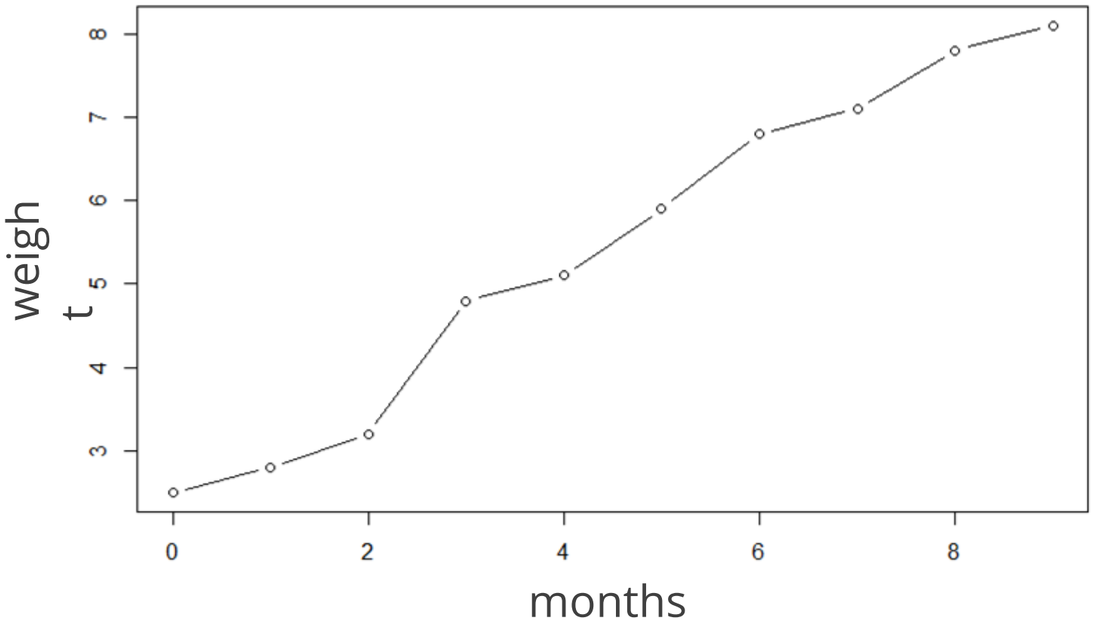
Data Visualization in R

CREATING A LINE CHART IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To create a line chart using plot() function by plotting body weight against months, use the following code:

```
weight <- c(2.5, 2.8, 3.2, 4.8, 5.1,
 5.9, 6.8, 7.1, 7.8,8.1)
months <- c(0,1,2,3,4,5,6,7,8,9)
plot(months,
weight, type = "b",
main="Baby Weight Chart")
```



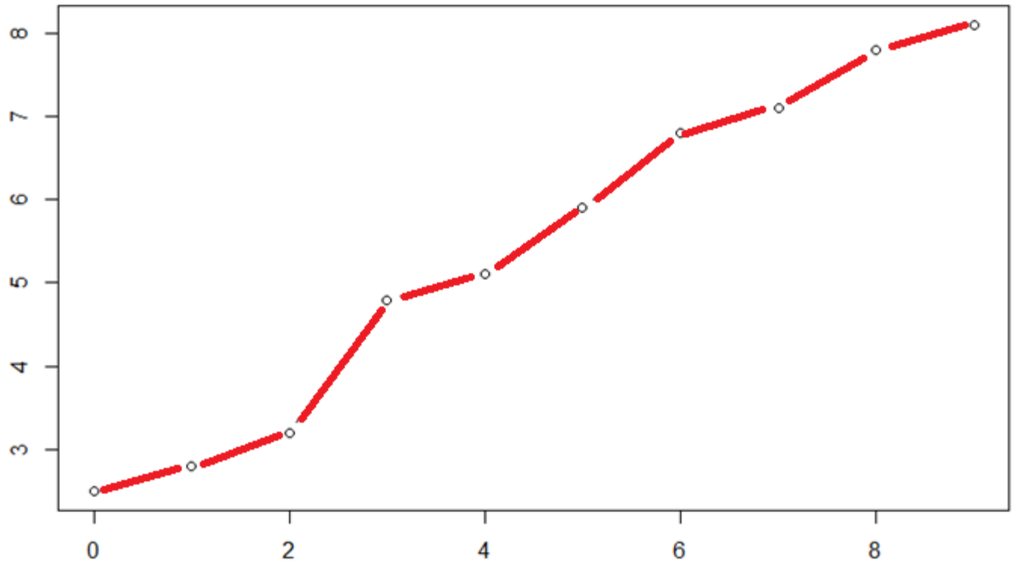
Data Visualization in R

EDITING A LINE CHART IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To change the color of the plot, use the following code:

```
Plot months, weight, type = “b”,
color = Red
```



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- **Box plot**
- Heat map
- Word cloud

Box plot, also called whisker diagram, displays the distribution of data based on the five-number summary:

- Minimum
- First quartile
- Median
- Third quartile
- Maximum

Syntax: *boxplot(data)*

Data Visualization in R

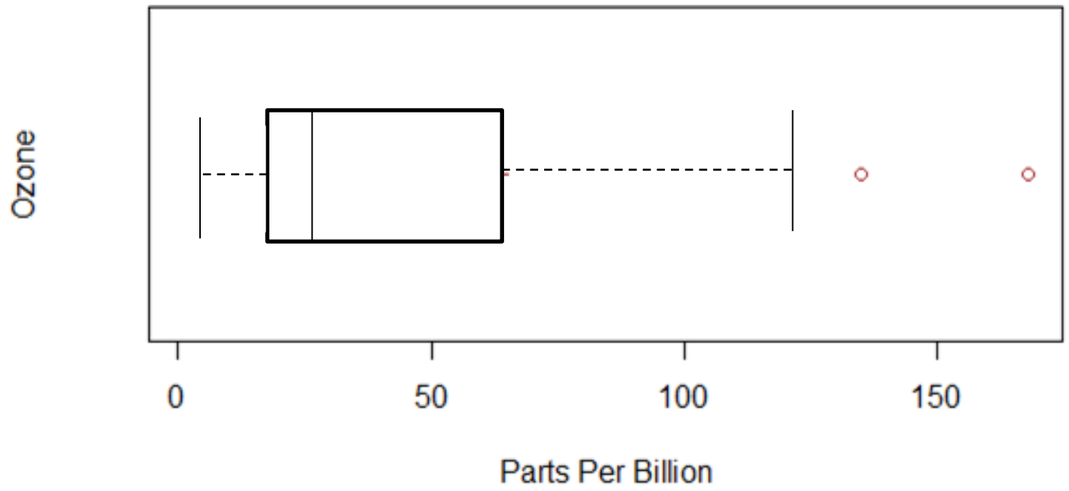
CREATING A BOX PLOT IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Use the following code to create a box plot using the inbuilt R dataset “airquality”:

```
boxplot(airquality$Ozone,
main="Mean Ozone in parts per
billion at Roosevelt Island",
xlab="Parts Per Billion",
ylab="Ozone",
horizontal=TRUE)
```

Mean ozone in parts per billion at Roosevelt Island



Data Visualization in R

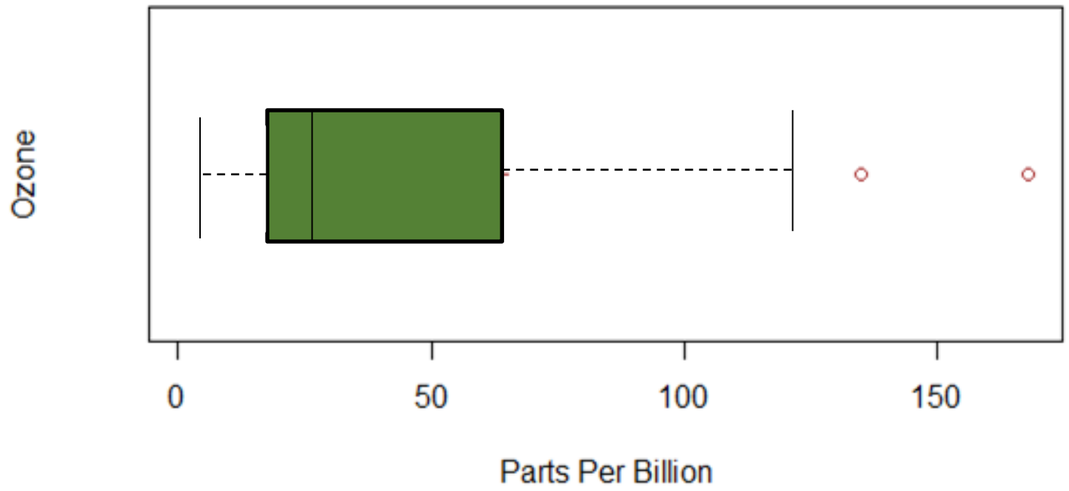
EDITING BOX PLOT IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To change the color of the plot, use the following code:

```
boxplot(airquality$Ozone,
main="Mean Ozone in parts per
billion at Roosevelt Island",
xlab="Parts Per Billion",
ylab="Ozone",
col="green",
horizontal=TRUE)
```

Mean ozone in parts per billion at Roosevelt Island



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A heat map is a two-dimensional representation of data that uses colors to represent the values. The two types of heat maps are:

1. Simple Heat Map: Provides an immediate visual summary of information

1. Elaborate Heat Map: Helps in understanding complex data sets

Syntax: *heatmap(data,Rowv=NA, Colv=NA)*

Data Visualization in R

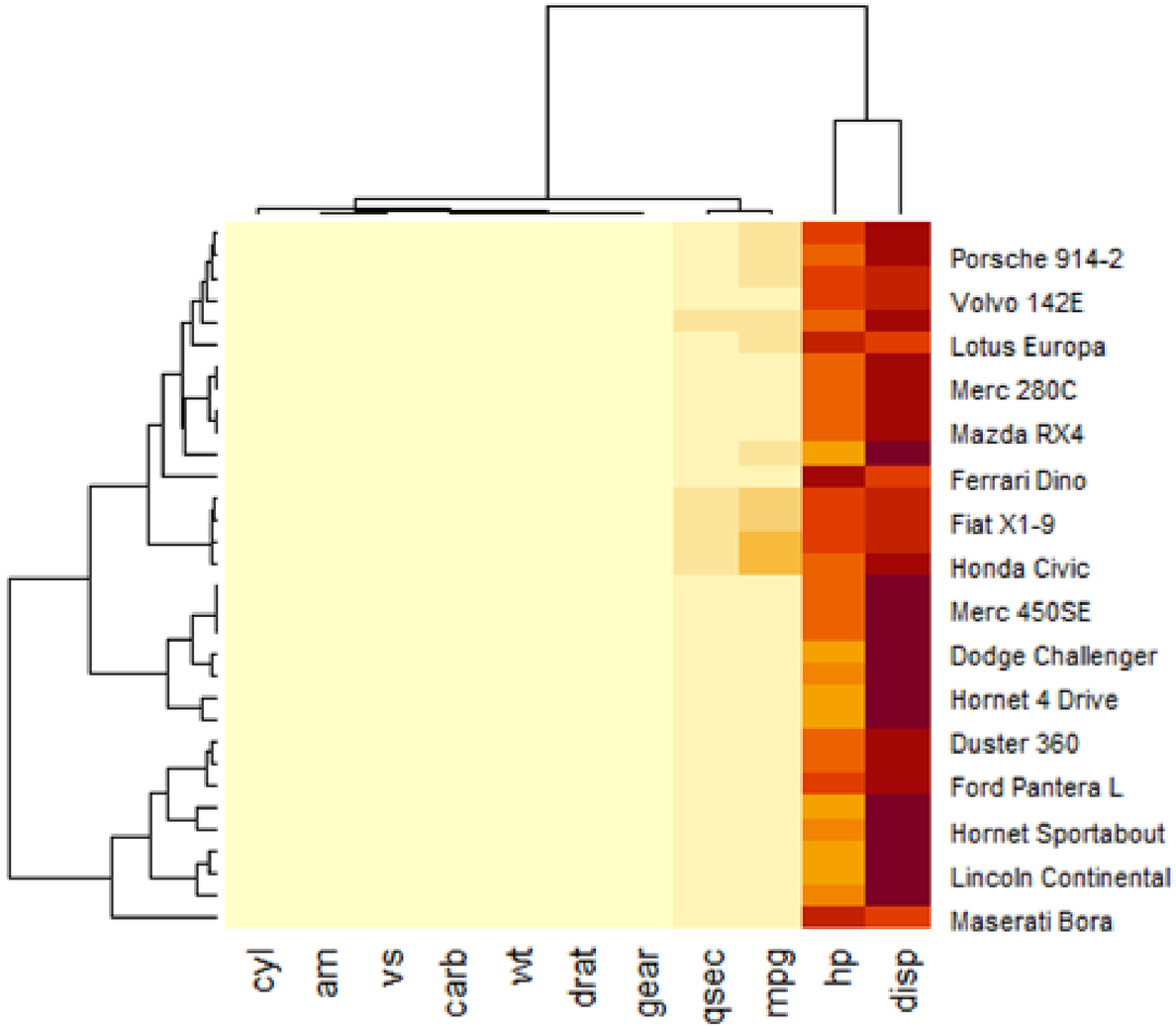
CREATING HEAT MAP IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To generate a simple heatmap, use the following code:

```
mat<-as.matrix(mtcars);
heatmap(mat);
```

Certain variables with relatively high values absorb all the variance.



Data Visualization in R

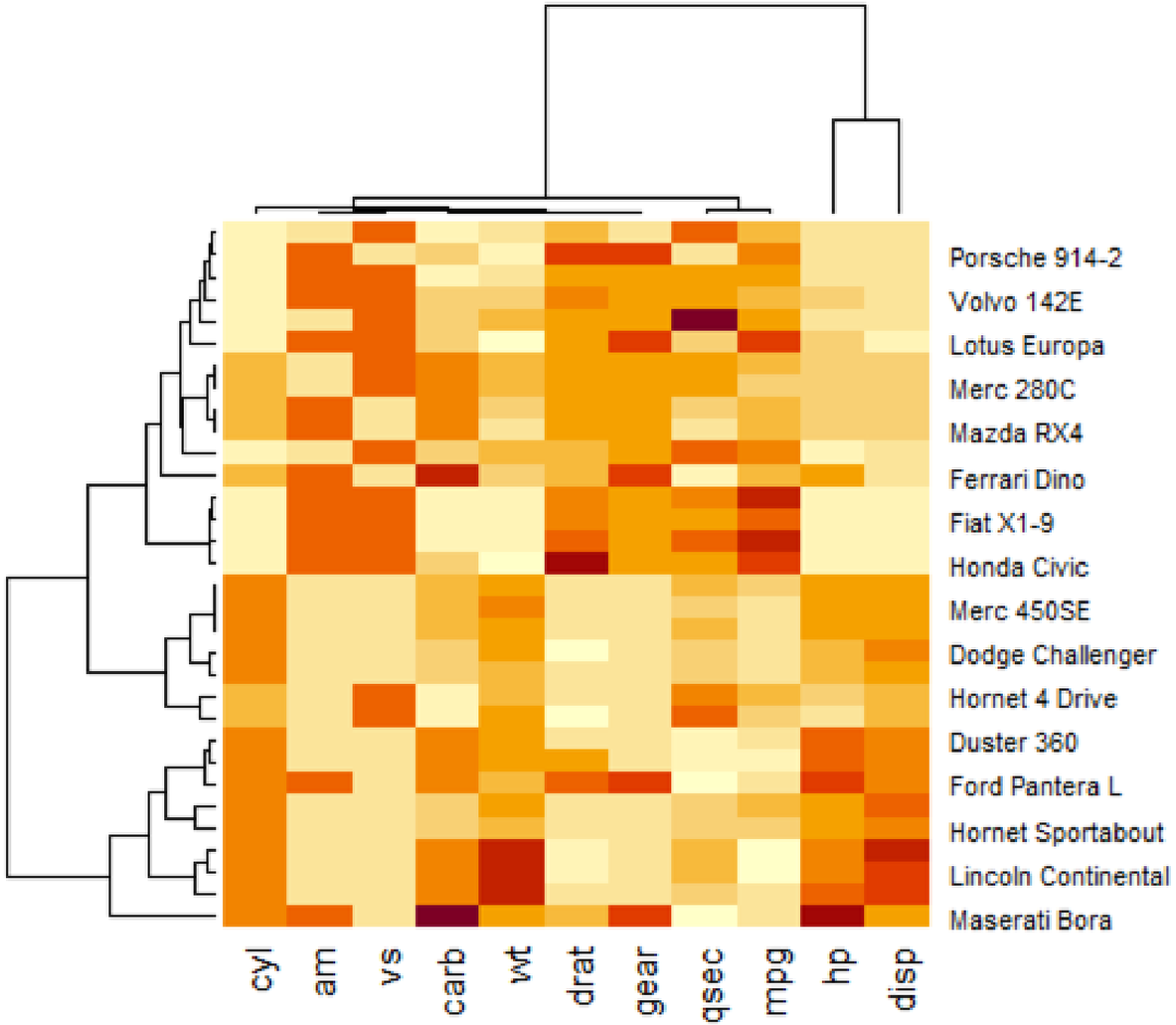
EDITING HEAT MAP IN R - NORMALIZATION

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

The **scale** argument of the heatmap is used to normalize the data matrix, as shown below:

```
heatmap(mat, scale="column");
```

In order to adjust the variation between columns, we may set the value of **scale** as **column** in the heatmap.



Data Visualization in R

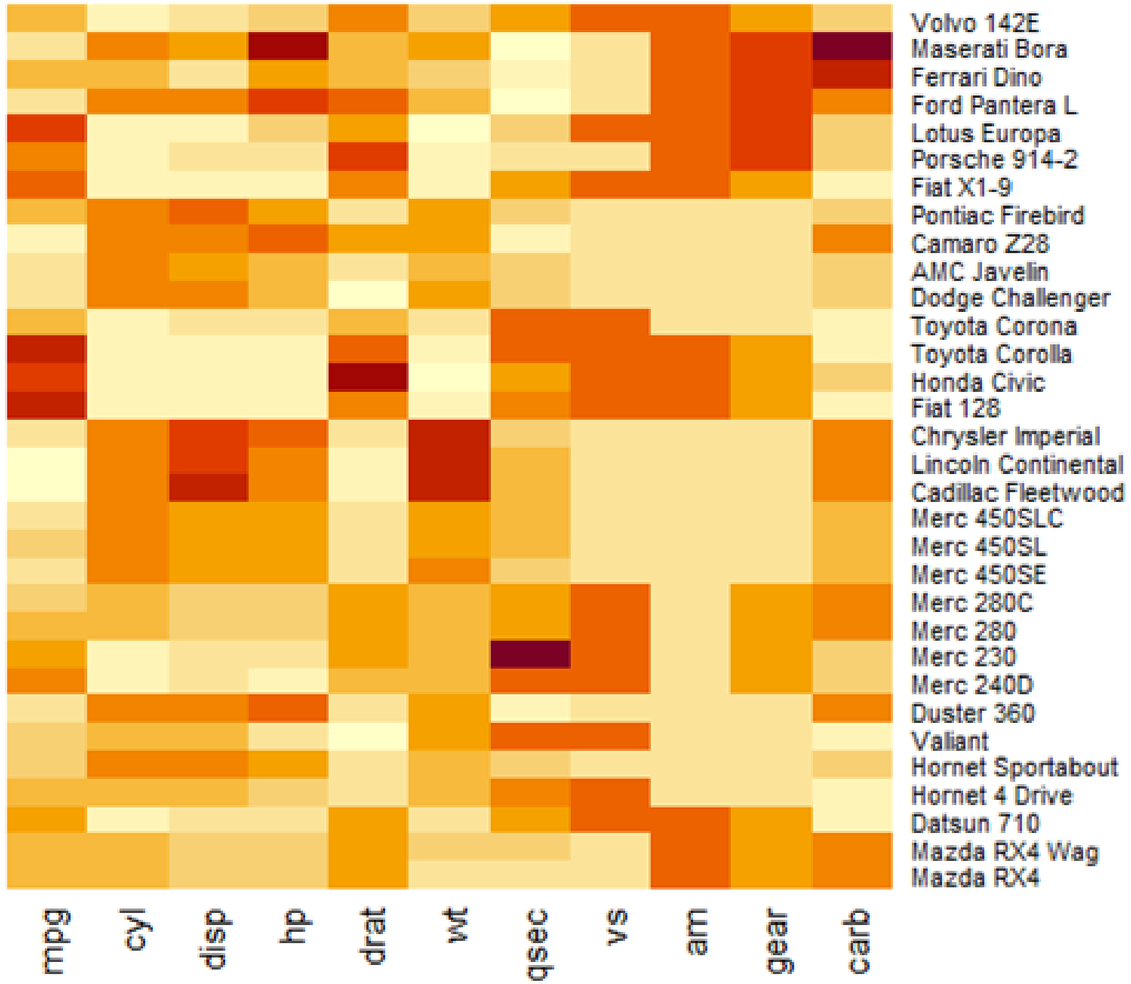
EDITING HEAT MAP IN R – DENDOGRAM AND REORDERING

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

A clustering algorithm sorts the order of rows and columns differently in the heatmap based on similarity.

The raw data matrix can be visualized and normalized without reordering columns or utilizing the dendrograms with the following code:

```
heatmap(mat, Colv = NA, Rowv
= NA, scale="column");
```



Data Visualization in R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

Word cloud (also called tag clouds) highlights the most commonly cited words in a text using a quick visualization.

Syntax: *wordcloud(words = data, freq =freq,min.freq = 2,)*

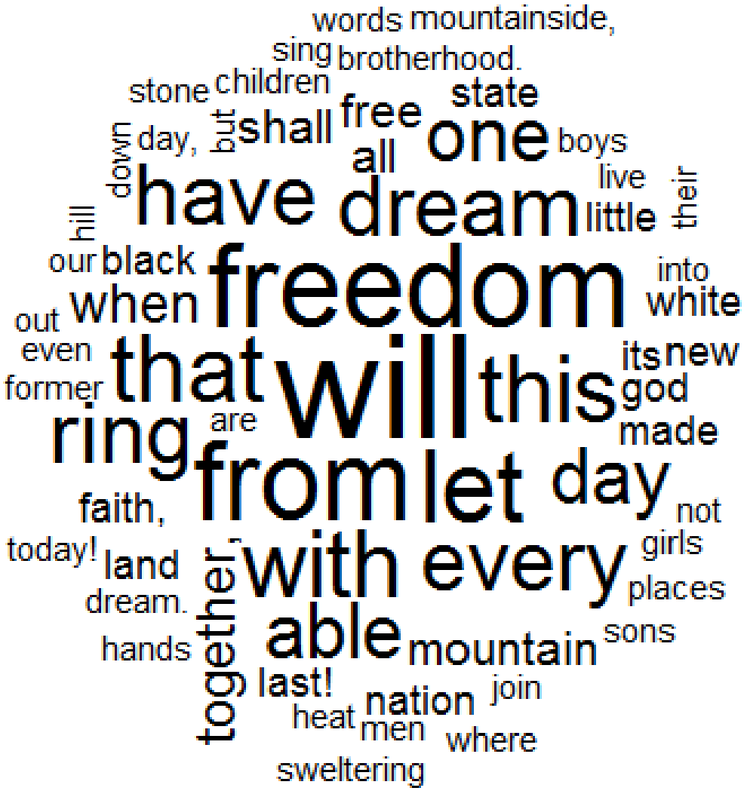
Data Visualization in R

CREATING WORD CLOUD IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- Word cloud

To create a word cloud, load the .csv data followed by the required library as shown below:

```
install.packages("wordcloud")
library("wordcloud")
data <- read.csv("TEXT.csv",header
= TRUE)
head(data)
wordcloud(words = data$word,
freq = data$freq, min.freq = 2,
 max.words=100, random.order=FALSE)
```



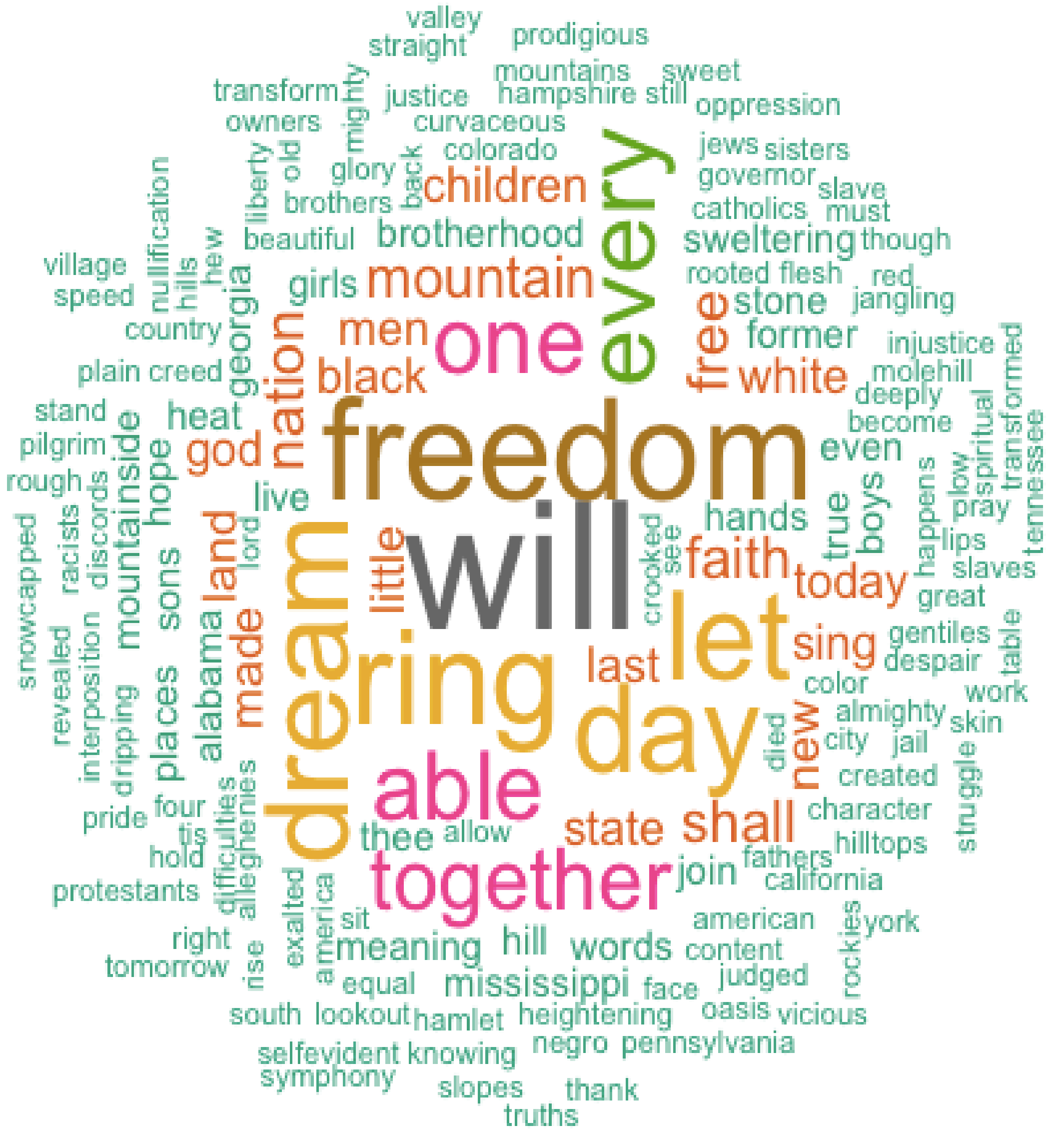
Data Visualization in R

EDITING WORD CLOUD IN R

- Bar chart
- Pie chart
- Histogram
- Kernel density plot
- Line chart
- Box plot
- Heat map
- **Word cloud**

For an attractive and colorful word cloud, use the code below:

```
install.packages("wordcloud")
library("wordcloud")
data <- read.csv("TEXT.csv",header =
TRUE)
head(data)
wordcloud(words = data$word,
freq = data$freq, min.freq = 2,
 max.words=100, random.order=FALSE,
rot.per=0.35, colors=brewer.pal(8,
“Dark2”))
```



simplilearn

Data Visualization in R

EXAMPLE

Problem statement

Study

Outcome

A real estate company is planning to launch apartments in Boston city and needs to analyze the factors affecting the price. The following data is available:

# of rooms in a House	Average price (in $ '000)
4.3	14.7
5.0	18.9
5.2	12.2
5.4	17.0
5.5	12.2
5.6	16.9
5.6	19.3
5.6	17.6
5.7	15.9
5.7	17.2
5.8	19.9
5.8	16.9
5.9	17.1
5.9	23.0
5.9	19.4

Data Visualization in R

EXAMPLE

Problem statement

Study

Outcome

The available data should be visualized to understand the interpretation easily. Graphics in R can be used to visualize the data.

Data Visualization in R

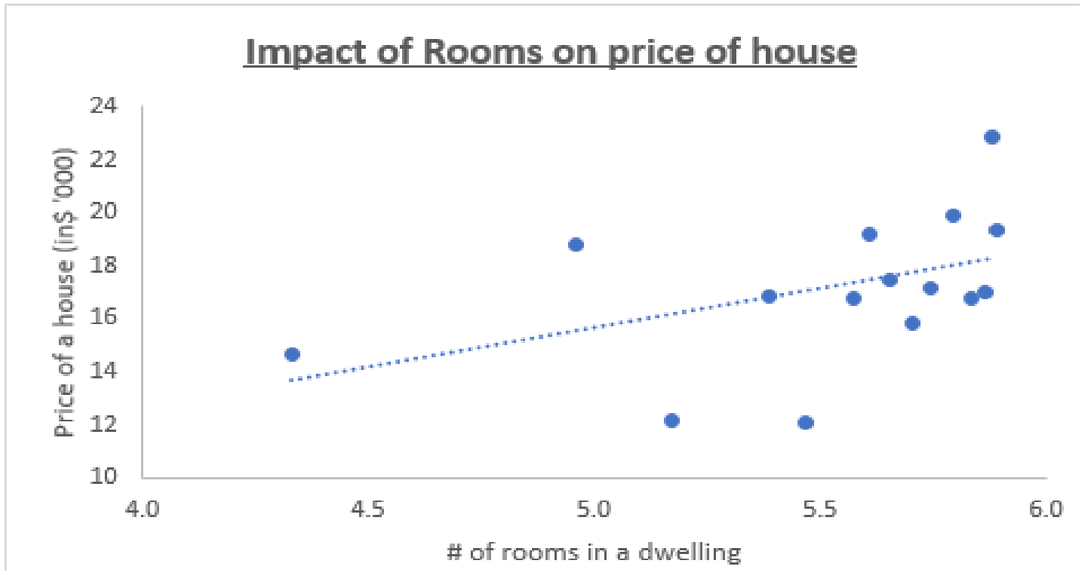
EXAMPLE

Problem statement

Study

Outcome

An X and Y scatter plot that shows the correlation between number of rooms and Average price of house can be depicted with an upward trend:



The price of house increases with the increase in number of rooms

Data Visualization in R

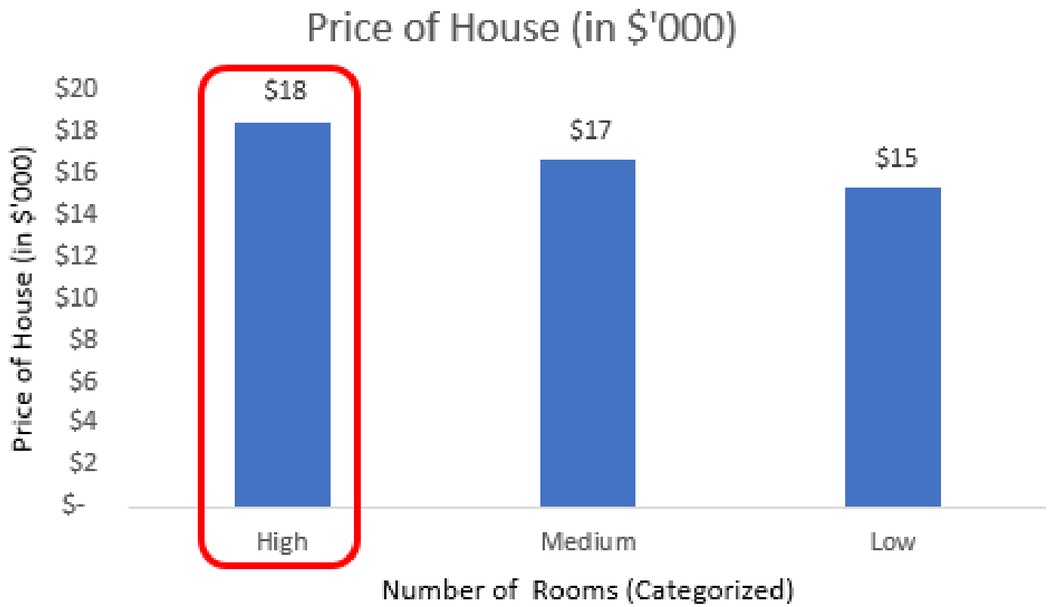
EXAMPLE

Problem statement

Study

Outcome

Categorization of number of rooms in bar chart:



Data Visualization in R

GRAPHICS LIMITATIONS

- Plots cannot be saved as objects
- Multivariate exploration is complex
- Layers are not supported
- Merging graphics is not supported

To overcome these challenges, ggplot2 is used.

Data Visualization in R

Topic 3—ggplot2

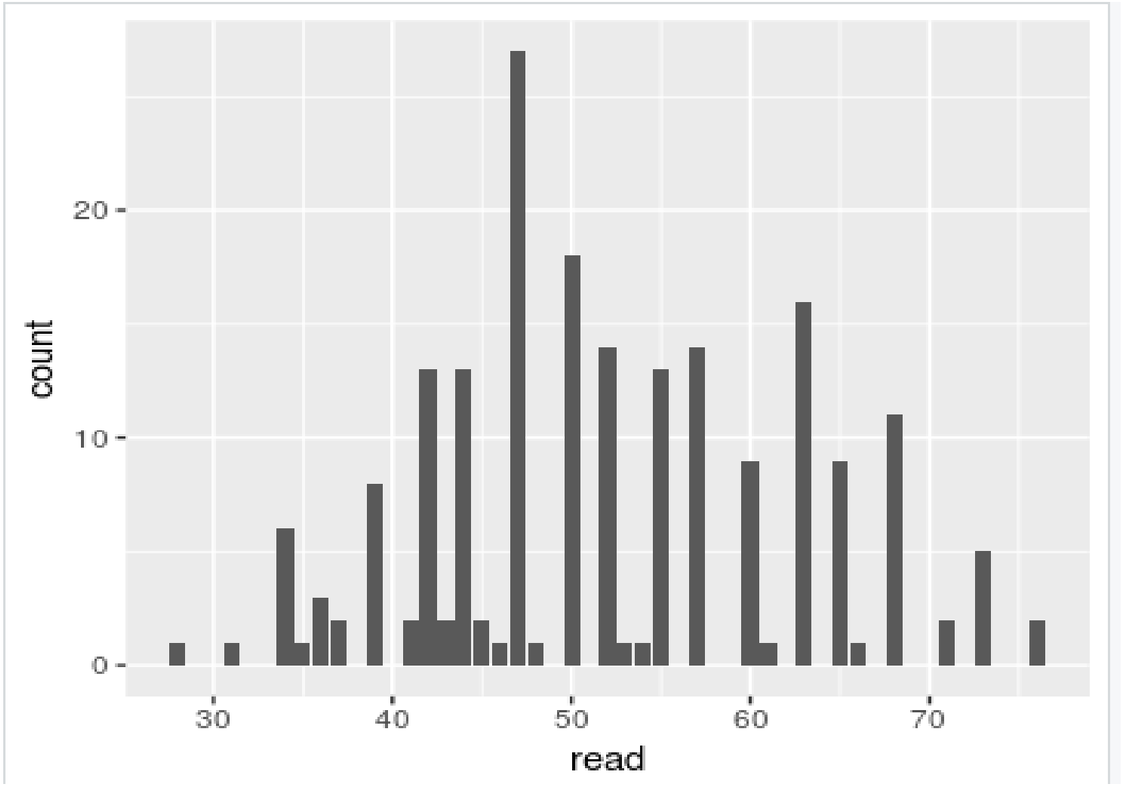
What Is ggplot2?

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.

ggplot2: Example 1

Creating a bar plot with just one variable with bars (In ggplot, the frequency need not be calculated):

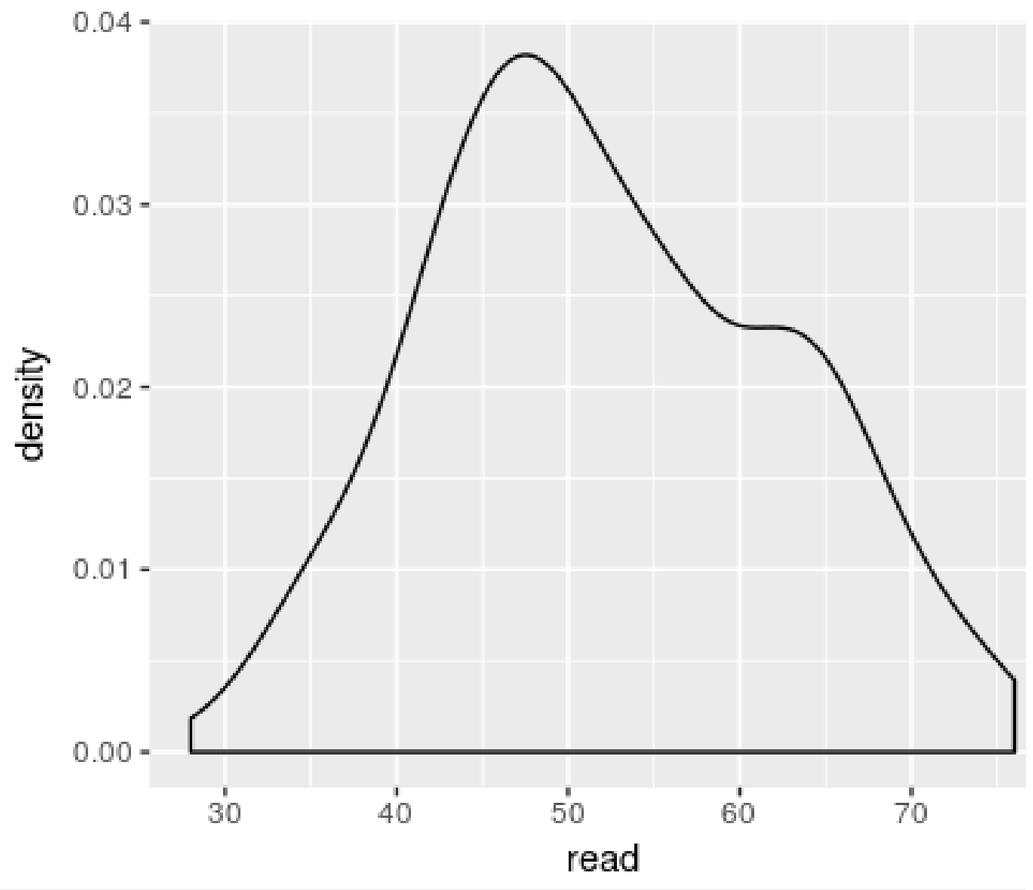
```
library("ggplot2")

ggplot(hsv,aes(x=read)) + geom_bar()
```



ggplot2: Example 2

Creating a Kernel density plan with one variable with a curve line:

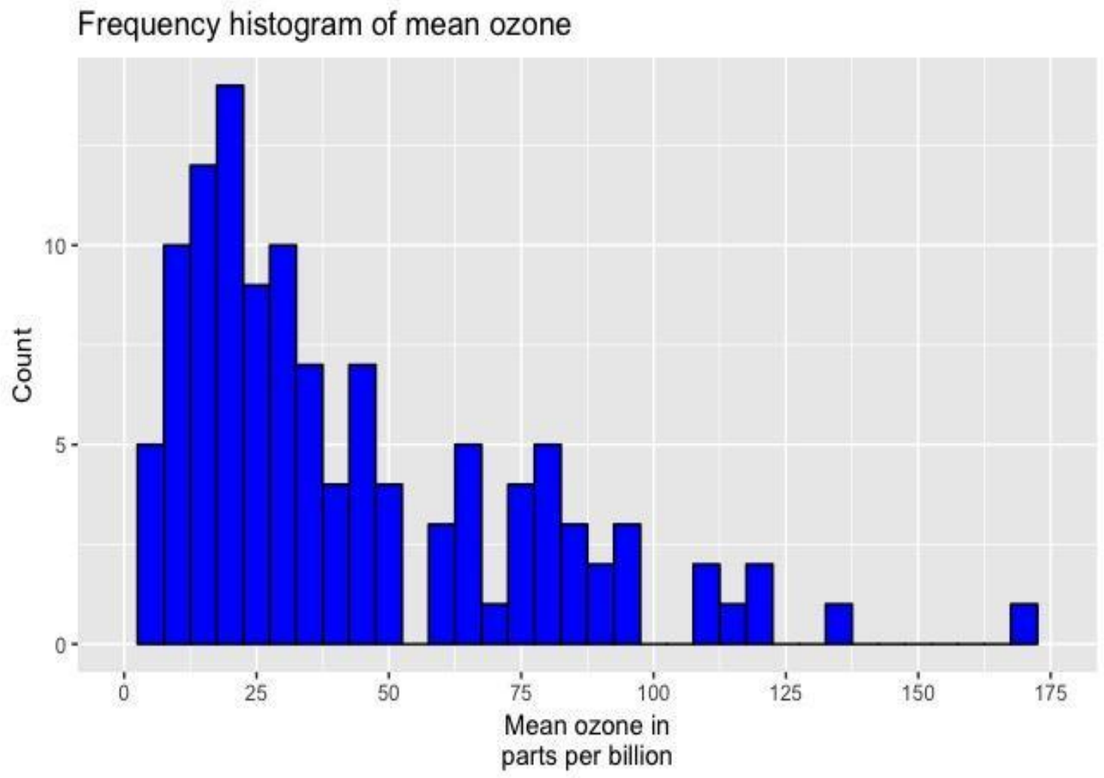
```
ggplot(hsv,aes(x=read)) + geom_density()
```



ggplot2: Example 3

Creating a Histogram using the “airquality” dataset:

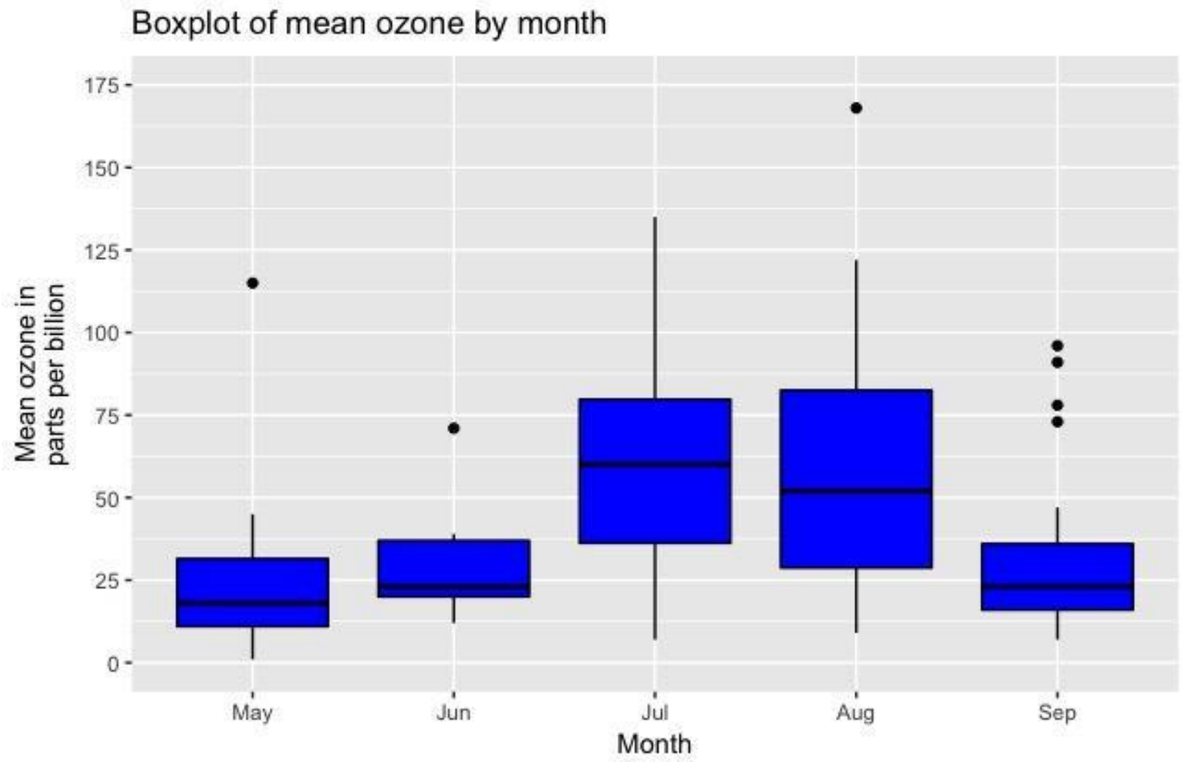
```
ggplot(airquality, aes(x = Ozone))
+geom_histogram(aes(y = ..count..),
binwidth = 5,  colour = "black", fill
= "blue")
+  scale_x_continuous(name = "Mean
ozone in\nparts per billion", breaks
= seq(0, 175, 25), limits=c(0, 175))
 +  scale_y_continuous(name =
"Count")
 +  ggtitle("Frequency histogram of
mean ozone")
```



ggplot2: Example 4

Creating a box plot using the “airquality” dataset:

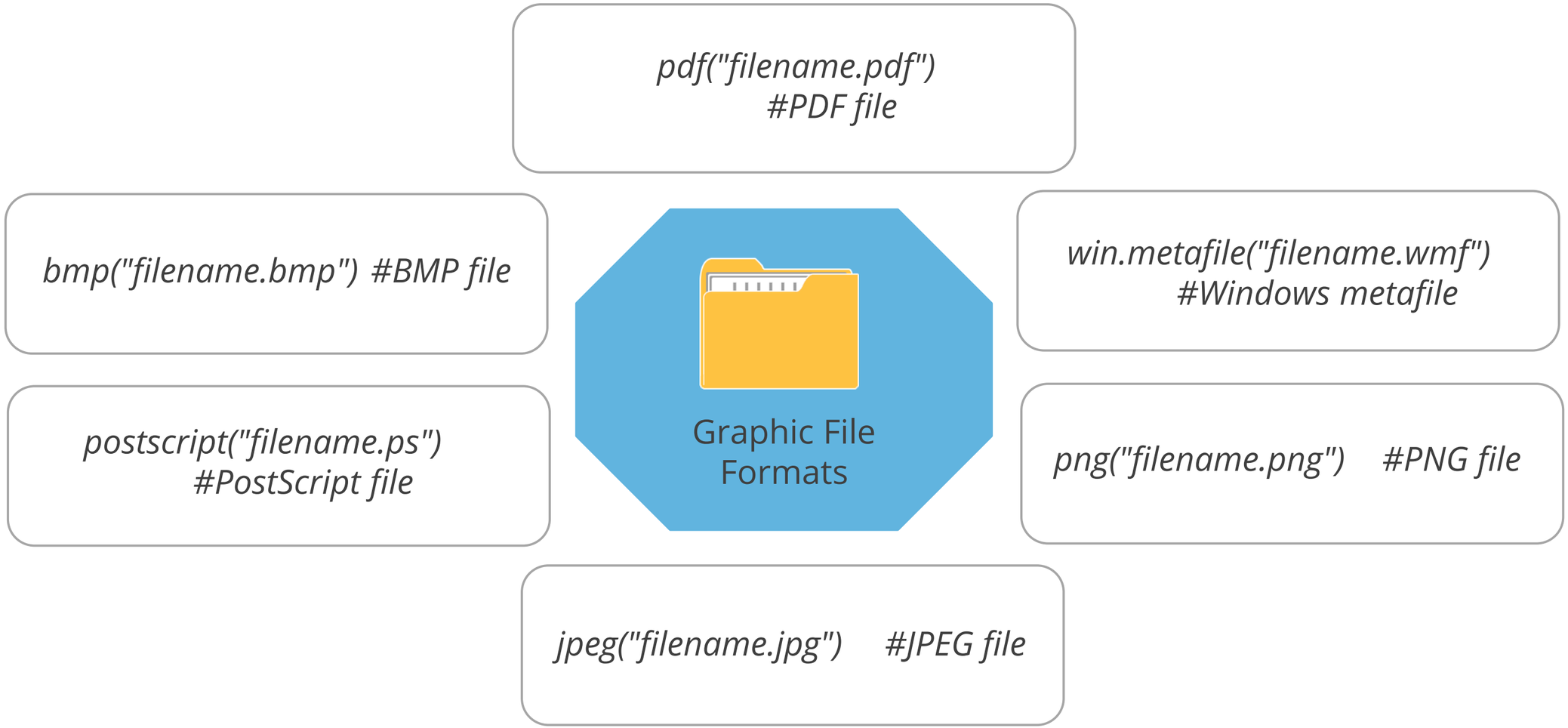
```
airquality$Month <-factor(airquality$Month,
labels = c("May", "Jun", "Jul", "Aug",
"Sep"))
ggplot(airquality, aes(x = Month, y = Ozone))
+  geom_boxplot(fill = "blue", colour =
"black")
+  scale_y_continuous(name = "Mean ozone
in\nparts per billion", breaks = seq(0, 175,
25), limits=c(0, 175))
+  scale_x_discrete(name = "Month") +
ggtitle("Boxplot of mean ozone by month")
```



Data Visualization in R

Topic 4—File Formats of Graphic Outputs

File Formats of Graphic Outputs

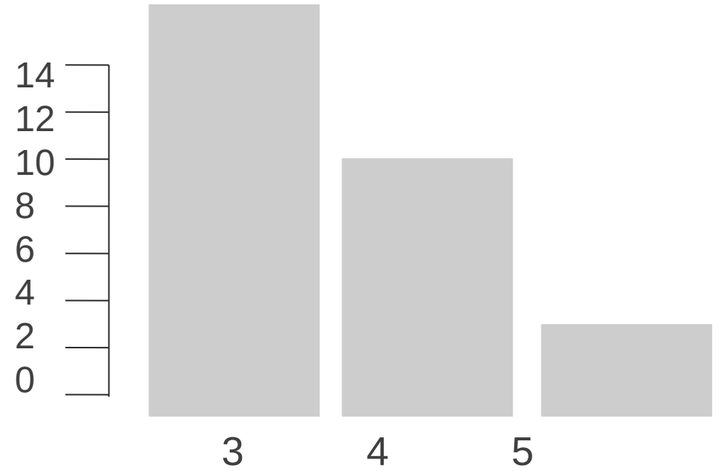


Saving a Graphic Output as a File

Example:

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

```
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 a modern equivalent of visual communication that involves the creation and study of the visual representation of data.
- R includes powerful packages of graphics that help in data visualization:
  - Bar chart
  - Pie chart
  - Histogram
  - Kernel density plot
  - Line chart
  - Box plot
  - Heat map
  - Word cloud
- 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.

Quiz

QUIZ

1

Which of the following graphics represents lengths, frequency, or proportion of categorical values?

a. Line chart

b. Bar plot

c. Bar chart

d. Kernel density plot

QUIZ

1

Which of the following graphics represents lengths, frequency, or proportion of categorical values?

a. Line chart

b. Bar plot

c. Bar chart

d. Kernel density plot

The correct answer is **c**

Bar chart represents lengths, frequency, or proportion of categorical values.

QUIZ

2

Graphic outputs can be saved as files using the _____ function.

a. save("filename.png")

b. write.table("filename.png")

c. write.file("filename.png")

d. png("filename.png")

QUIZ

2

Graphic outputs can be saved as files using the _____ function.

a. save("filename.png")

b. write.table("filename.png")

c. write.file("filename.png")

d. png("filename.png")

The correct answer is **d.**

Graphic outputs can be saved as files using the png("filename.png") function.

This concludes “Data Visualization”

The next lesson is “Statistics for Data Science – I.”