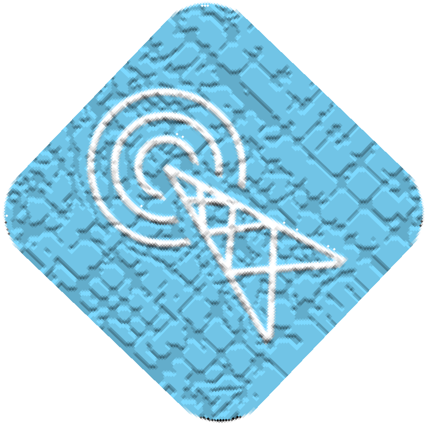
Data Science Using R

Lesson04–Data Visualization using R

Objective

After completing this lesson you will be able to:



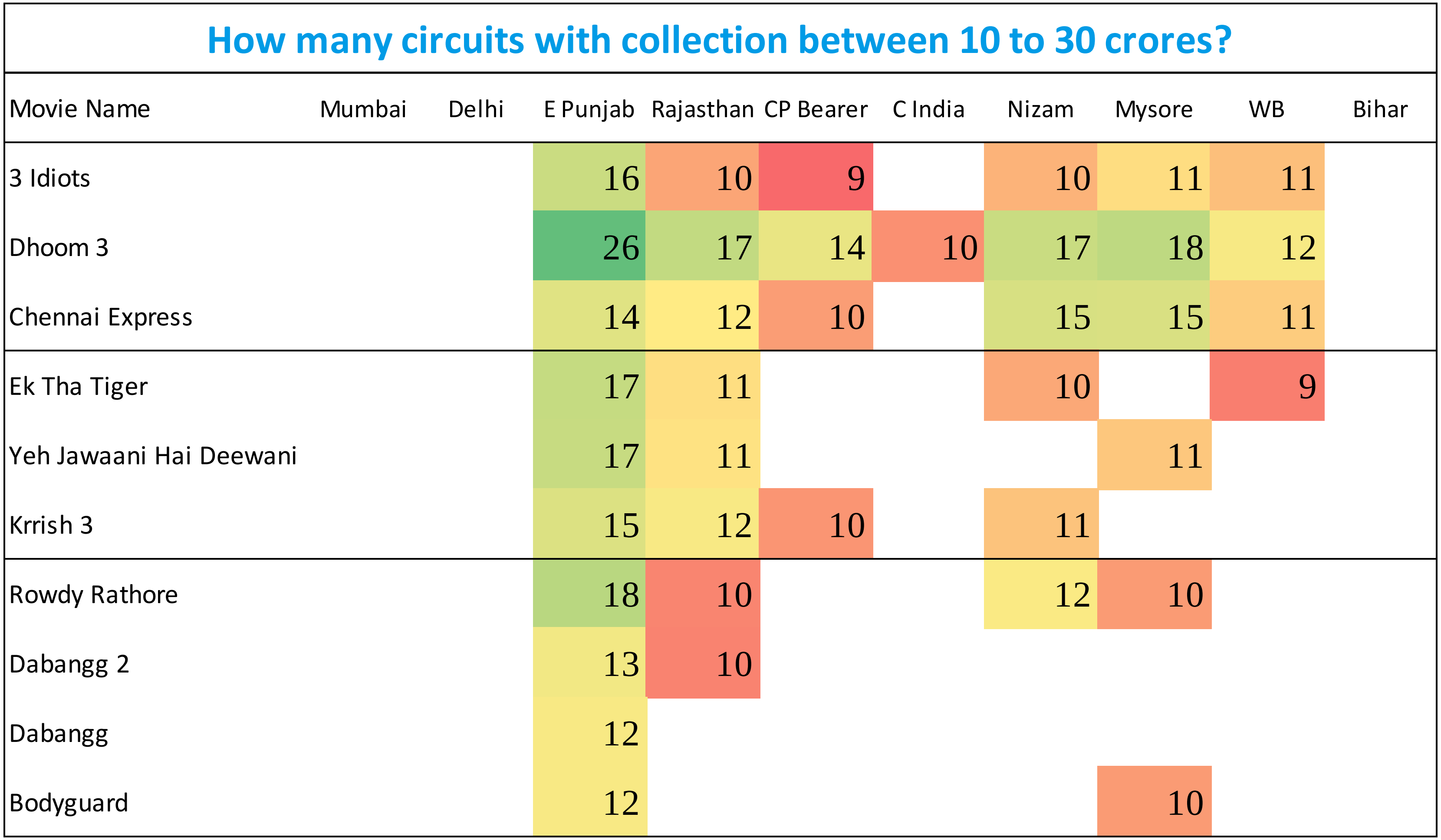
- Explain the importance of Data Visualization
- Create bar chart, pie chart, mosaic plot using R
- Create scatter plot, histogram and correlation plot in R
- Create box plot and other advanced plotting using R

Exercise 1

How many circuits with collection between 10 to 30 crores?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16	10	9	7	10	11	11	4
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39	14	12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11	9	7	10	9	9	4
Yeh Jawaani Hai Deewani	59	42	17	11	8	6	7	11	8	4
Krrish 3	58	33	15	12	10	7	11	9	9	5
Rowdy Rathore	51	38	18	10	9	6	12	10	5	4
Dabangg 2	49	32	13	10	9	7	8	4	7	4
Dabangg	55	33	12	8	8	5	7	6	7	4
Bodyguard	46	31	12	9	8	5	8	10	6	5

Exercise 1…



How many circuits with collection between 10 to 30 crores?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots			16	10	9		10	11	11	
Dhoom 3			26	17	14	10	17	18	12	
Chennai Express			14	12	10		15	15	11	
Ek Tha Tiger			17	11			10		9	
Yeh Jawaani Hai Deewani			17	11				11		
Krrish 3			15	12	10		11			
Rowdy Rathore			18	10			12	10		
Dabangg 2			13	10						
Dabangg			12							
Bodyguard			12					10		

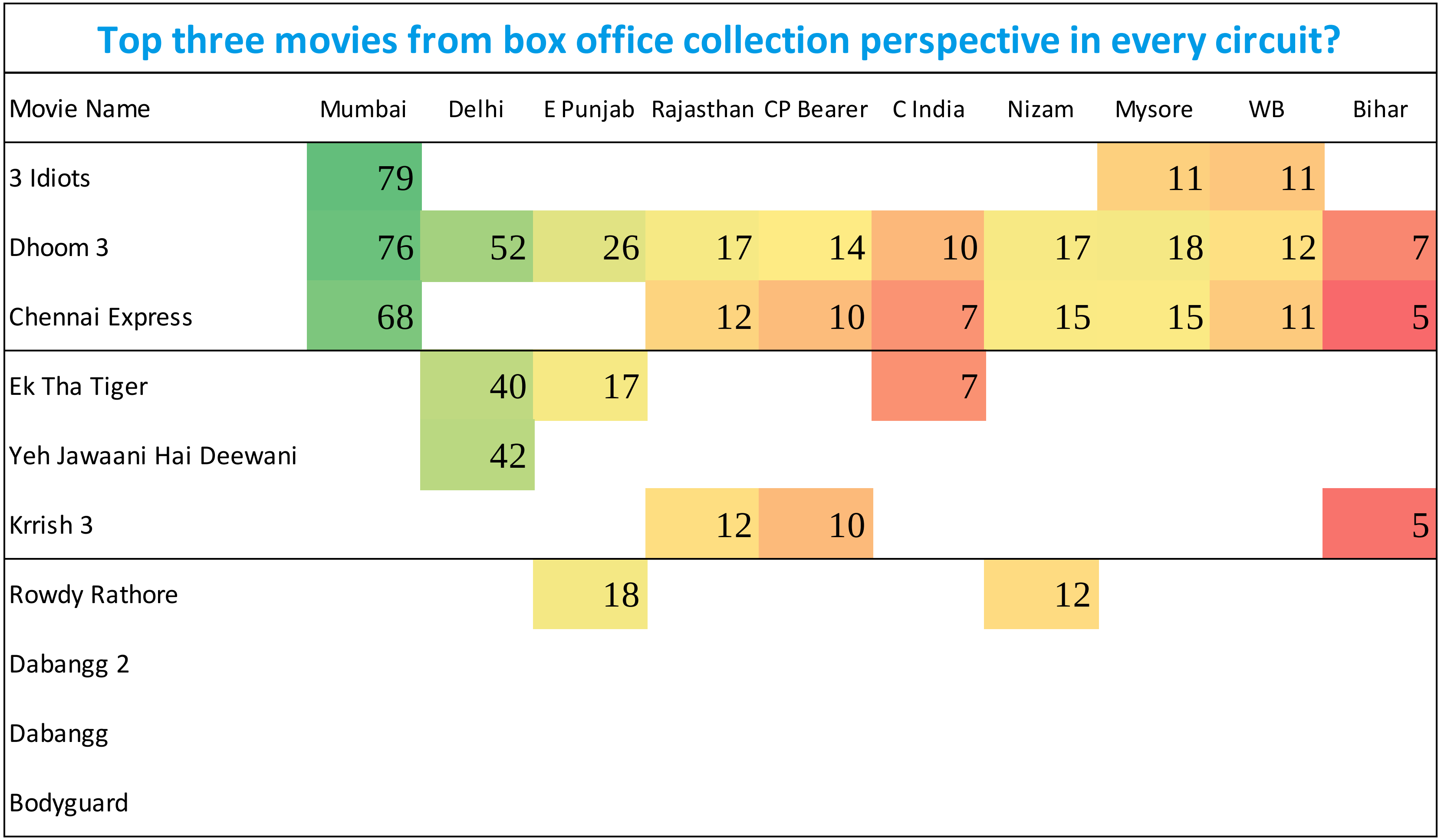
Conditional formatting on values between 10 and 30 crores

Exercise 2

Top three movies from box office collection perspective in every circuit?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16	10	9	7	10	11	11	4
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39	14	12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11	9	7	10	9	9	4
Yeh Jawaani Hai Deewani	59	42	17	11	8	6	7	11	8	4
Krrish 3	58	33	15	12	10	7	11	9	9	5
Rowdy Rathore	51	38	18	10	9	6	12	10	5	4
Dabangg 2	49	32	13	10	9	7	8	4	7	4
Dabangg	55	33	12	8	8	5	7	6	7	4
Bodyguard	46	31	12	9	8	5	8	10	6	5

Exercise 2…



Top three movies from box office collection perspective in every circuit?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79							11	11	
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68			12	10	7	15	15	11	5
Ek Tha Tiger		40	17			7				
Yeh Jawaani Hai Deewani		42								
Krrish 3				12	10					5
Rowdy Rathore			18				12			
Dabangg 2										
Dabangg										
Bodyguard										

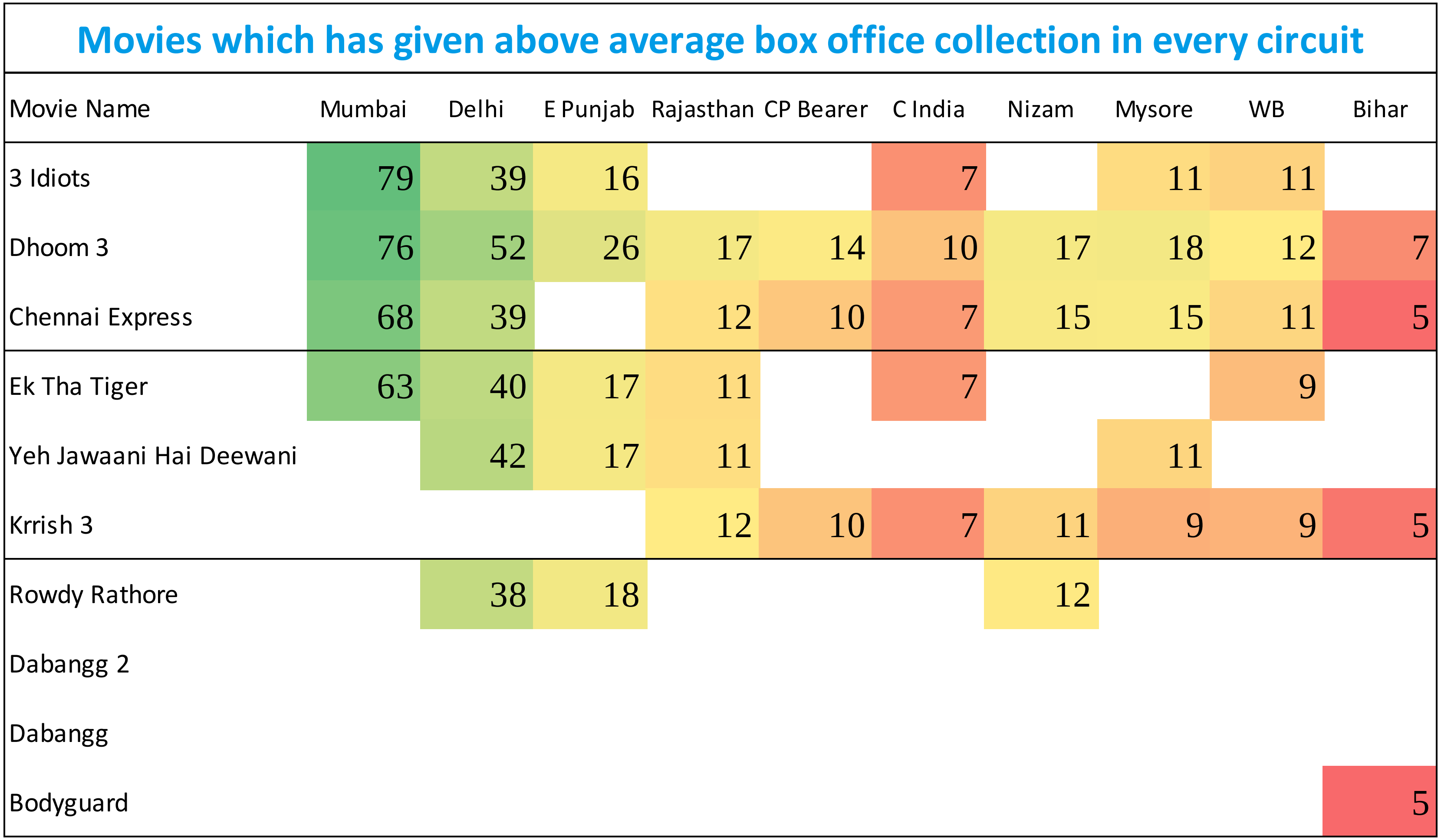
Conditional formatting on top 30% movies in different circuits.

Exercise 3

Movies which has given above average box office collection in every circuit

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16	10	9	7	10	11	11	4
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39	14	12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11	9	7	10	9	9	4
Yeh Jawaani Hai Deewani	59	42	17	11	8	6	7	11	8	4
Krrish 3	58	33	15	12	10	7	11	9	9	5
Rowdy Rathore	51	38	18	10	9	6	12	10	5	4
Dabangg 2	49	32	13	10	9	7	8	4	7	4
Dabangg	55	33	12	8	8	5	7	6	7	4
Bodyguard	46	31	12	9	8	5	8	10	6	5

Exercise 3…

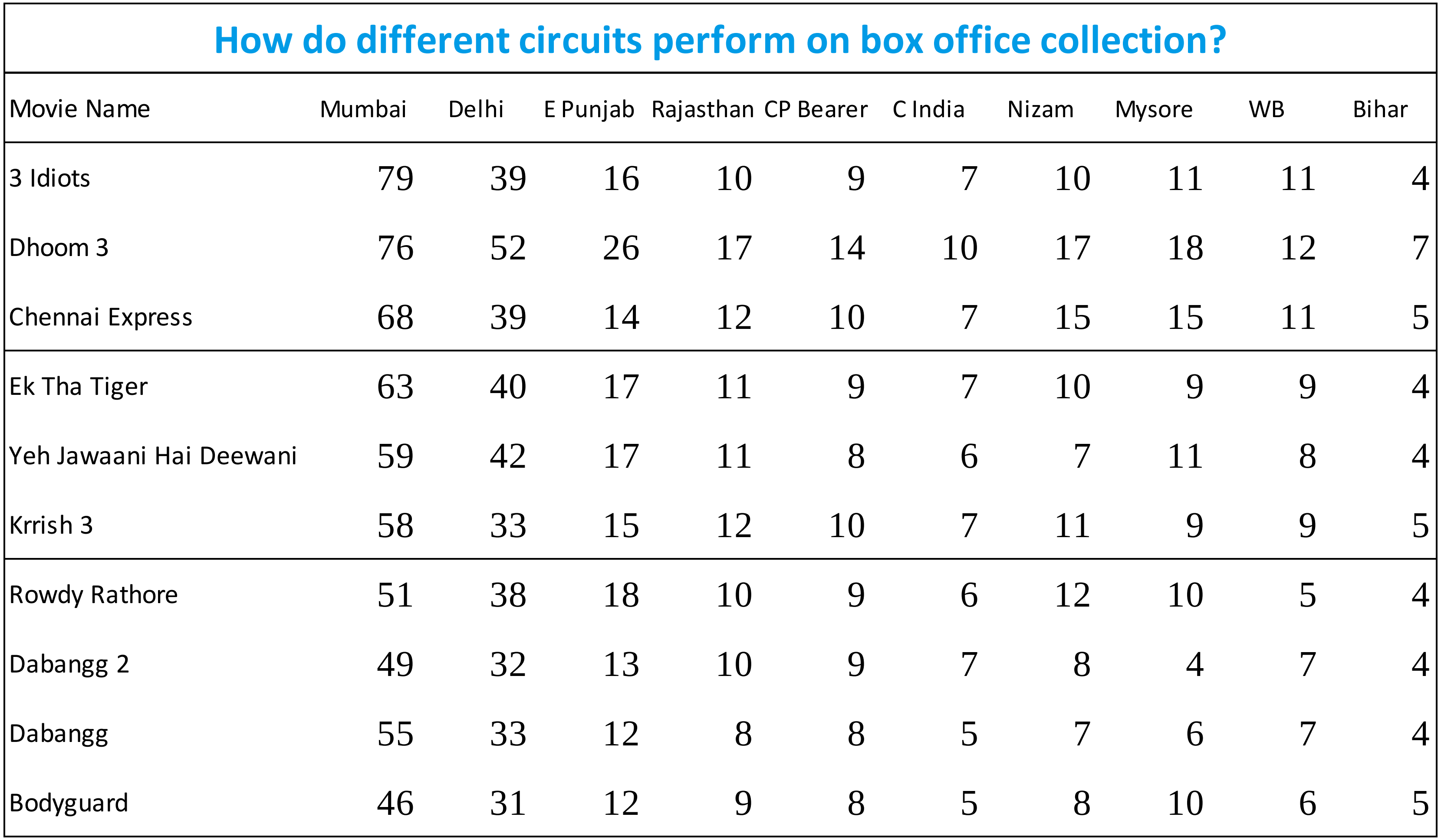


Movies which has given above average box office collection in every circuit

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16			7		11	11	
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39		12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11		7			9	
Yeh Jawaani Hai Deewani		42	17	11				11		
Krrish 3				12	10	7	11	9	9	5
Rowdy Rathore		38	18				12			
Dabangg 2										
Dabangg										
Bodyguard										5

Gradient fill of green on above average for each of the columns separately.

Exercise 4



How do different circuits perform on box office collection?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16	10	9	7	10	11	11	4
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39	14	12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11	9	7	10	9	9	4
Yeh Jawaani Hai Deewani	59	42	17	11	8	6	7	11	8	4
Krrish 3	58	33	15	12	10	7	11	9	9	5
Rowdy Rathore	51	38	18	10	9	6	12	10	5	4
Dabangg 2	49	32	13	10	9	7	8	4	7	4
Dabangg	55	33	12	8	8	5	7	6	7	4
Bodyguard	46	31	12	9	8	5	8	10	6	5

Exercise 4…

How do different circuits perform on box office collection?

Movie Name	Mumbai	Delhi	E Punjab	Rajasthan	CP Bearer	C India	Nizam	Mysore	WB	Bihar
3 Idiots	79	39	16	10	9	7	10	11	11	4
Dhoom 3	76	52	26	17	14	10	17	18	12	7
Chennai Express	68	39	14	12	10	7	15	15	11	5
Ek Tha Tiger	63	40	17	11	9	7	10	9	9	4
Yeh Jawaani Hai Deewani	59	42	17	11	8	6	7	11	8	4
Krrish 3	58	33	15	12	10	7	11	9	9	5
Rowdy Rathore	51	38	18	10	9	6	12	10	5	4
Dabangg 2	49	32	13	10	9	7	8	4	7	4
Dabangg	55	33	12	8	8	5	7	6	7	4
Bodyguard	46	31	12	9	8	5	8	10	6	5

Gradient fill from red (min of values) to green (max of the values)

Video: Best stats you have ever seen

Why Visualization

- Data visualization shifts the balance between seeing (perception) and thinking (cognition) to take maximum advantage of how brain functions.

- Studies in attention and memory have revealed that humans have limited ability to hold multiple items simultaneously in awareness.
  - Encoding information visually, allows more information to be chunked together into the limited slots available in working memory.
  - Several views of information in front of eyes at one time, extends ability to explore data from multiple dimension and from multiple perspectives.

More notes at: Data Visualization for human perception

Basic Points for Effective Visualization

- Human eye can read linear distances more effectively than circular distances.
- Human eyes are tuned to pick up red, green and blue colors instantly than any other color.
  - Coloring based on the gradient shades of green, blue or red brings more meaning to the data being represented.
- We live in a 3 dimensional space and thus are tuned to recognize 2 dimensional charts easily. But what after that?

- First two dimensions can be visualized through co-ordinates
- Color intensity may form the third dimension
- Size or length may form the fourth dimension
- Shape may form the fifth dimension
- Texture, angle…

- Numbers after decimals may not be needed when analyzing large data set.
- 3D rendering of charts often complicates comparison as perspective skews relative shape and size.
- Legends in graphs with many options/colors to select becomes non-intuitive.

Descriptive Statistics and Data Visualization

- Descriptive statistics is a field in analytics which caters to summarizing data and extracting information from the data.
- Data Visualization may form the building block for descriptive statistics.
- R provides the flexibility and robustness in data visualization. Some notable features of R which aids in data visualization are:

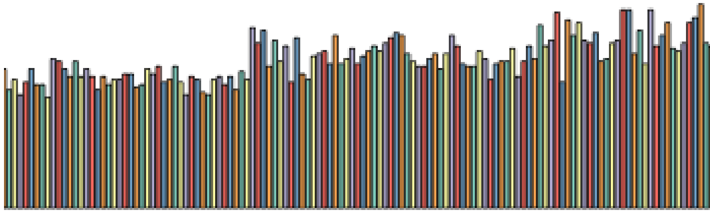
- Powerful environment for visualizing data
- Integrated graphics and statistics infrastructure
- Fully programmable and highly reproducible
- Vast number of R packages with graphics utilities

Data visualization is only successful to the degree to which it encodes information in a manner that our eyes can discern and our brains can understand.

Bar Charts

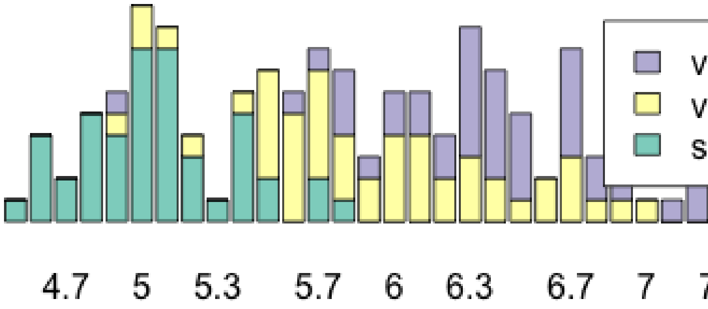
Used to show comparison of quantities over different categor

Examples to generate bar chart from Iris dataset.



```
# Simple bar charts . Uses graphics() library.
library(RColorBrewer)
barplot(iris$Sepal.Length,col  =
brewer.pal(3,"Set1"))

#stacked bar charts
library(RColorBrewer)
barplot(table(iris$Species,iris$Sepal.Le
ngth),col  = brewer.pal(4,"Set3"),
legend.text  = TRUE)
```



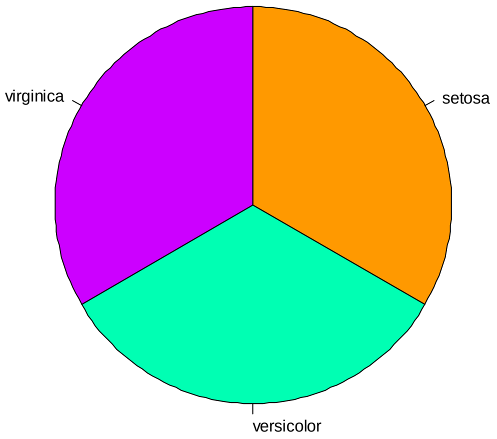
Pie Charts

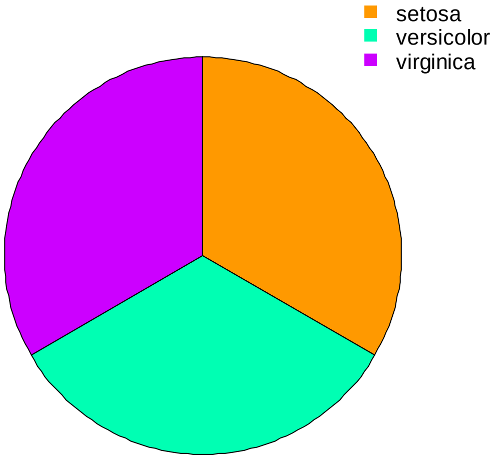
Used to show share of categorical variables in the overall dataset. Examples to generate pie chart from Iris dataset.

```
# Plots a simple pie chart. Uses graphics() library.
y <- table(iris$Species)
pie(y, col=rainbow(length(y), start=0.1,
end=0.8), main="Pie Chart", clockwise=T)

#plot a pie chart with legends
pie(y, col=rainbow(length(y), start=0.1,
end=0.8), labels=NA, main="Pie Chart",
clockwise=T); legend("topright",
legend=row.names(y), cex=1.3, bty="n",
pch=15, pt.cex=1.8,
col=rainbow(length(y), start=0.1,
end=0.8), ncol=1)
```

- Pie chart may not be a useful way to represent any data.



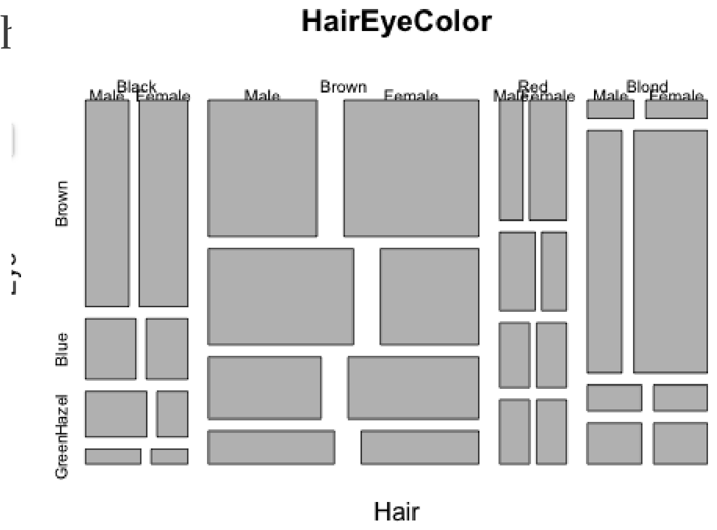


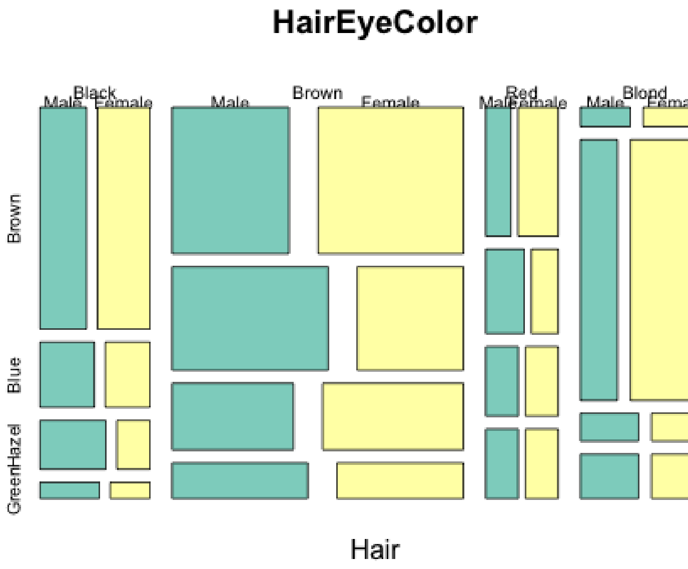
Mosaic Plot

Used for plotting large set of categorical data where area of tl
Examples to generate mosaic plot from Iris dataset.

```
# Mosaic plot without color. Uses graphics() library.
mosaicplot(HairEyeColor)

# Mosaic plot with color.
library(RColorBrewer)
mosaicplot(HairEyeColor,col  =
brewer.pal(6,"Set3"))
```

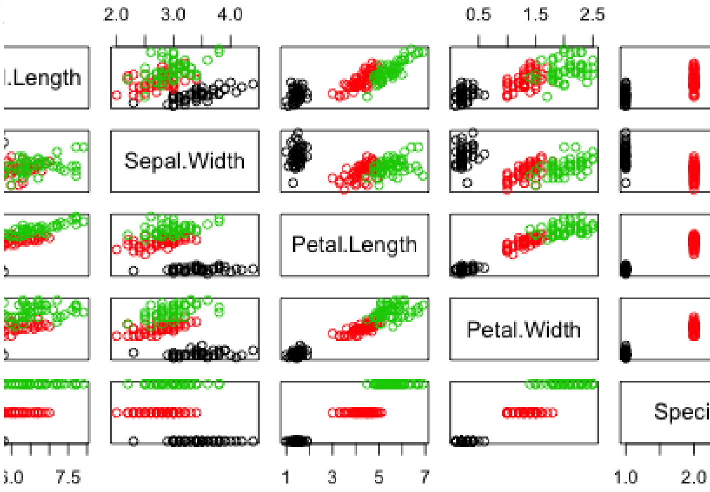


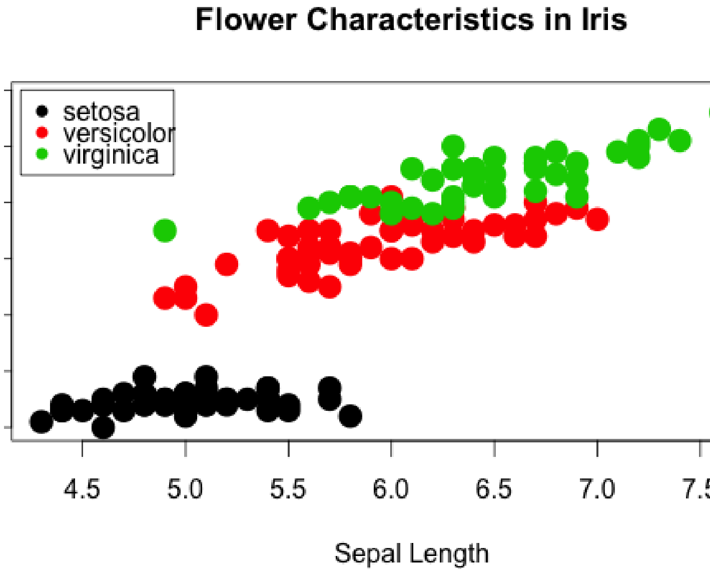


Pair plot or Scatter Plot

Used to show joint variation of numeric data which can be se
Examples to generate pair plot from Iris dataset.

```
# scatter plot matrix with iris dataset. Uses graphics()
#library.
data(iris)
pairs(iris, col = iris$Species) #pair
plot with color
#plot of all variables with color
plot(iris$Sepal.Length, iris$Petal.Length,  # x & y variable
col = iris$Species,                         # color by species
pch = 16,                                   # type of point to use
cex = 2,                                    # size of point to use
xlab = "Sepal Length",                      # x axis label
ylab = "Petal Length",                      # y axis label
main = "Flower Characteristics in Iris")    # plot title
legend (x = 4.2, y = 7, legend = levels(iris$Species), col =
c(1:3), pch = 16)
```



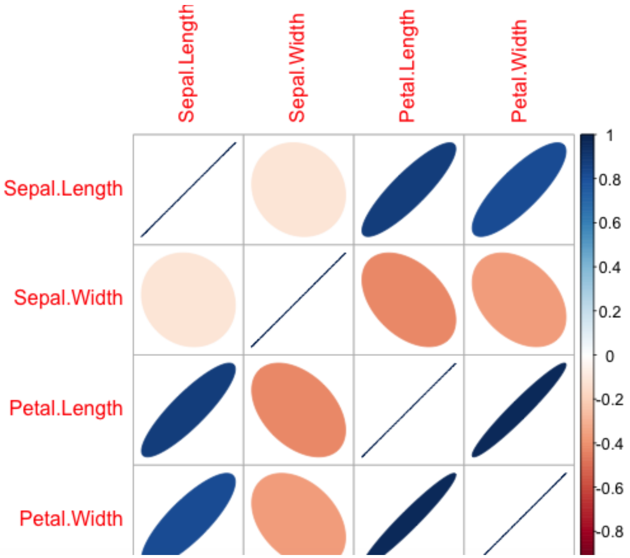


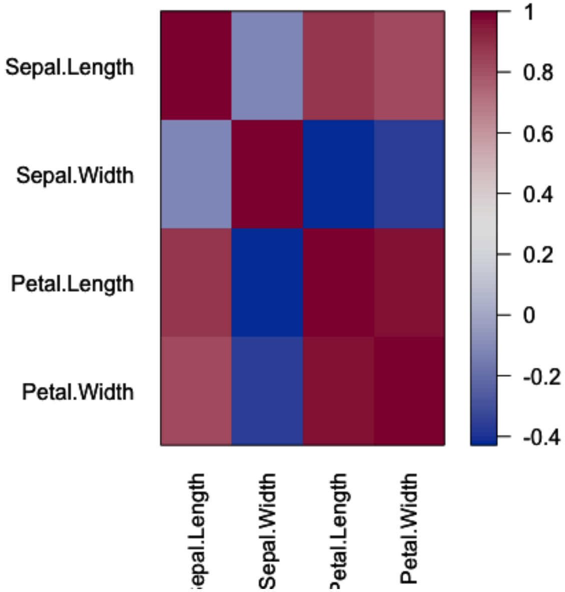
Correlation Plot

Correlation plot shows the degree of variation between two n
generate correlation plot.

```
#correlation plot with iris dataset
library(corrplot)
iris_matrix <- as.matrix(iris[,1:4])
corrplot(cor(iris_matrix),
method="ellipse")

#correlation plot with a different library
library(seriation)
iris_matrix <- as.matrix(iris[,1:4])
pimage(cor(iris_matrix), colorkey=TRUE,
range=c(-1,1), col=diverge_hcl(100))
```

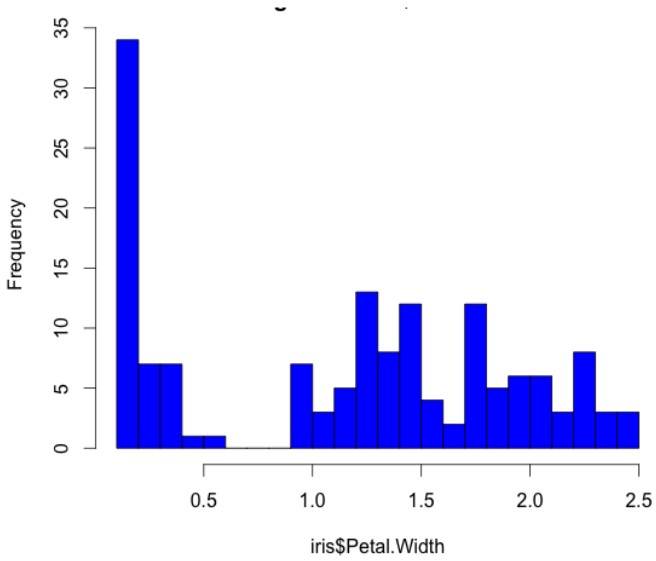




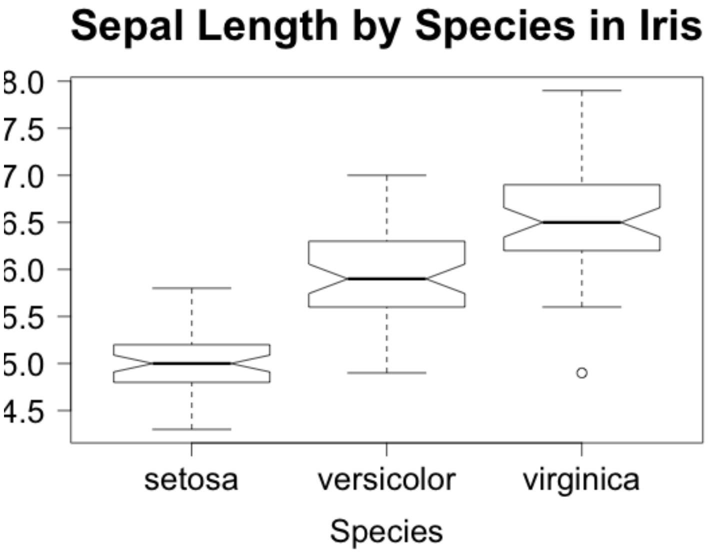
Histogram and Box Plot

Both used to summarize numeric data.

- Histogram is used to bin the data and understand the
- Boxplot can be used to identify outliers in the dataset and box plot.



```
#histogram plot with iris dataset. Uses graphics() library.
hist(iris$Petal.Width, breaks=20,
col=“blue")
#box plot of all variables
boxplot(iris$Sepal.Length ~ iris$Species,   #
x &y variable,
notch = T,    # Draw notch
las = 1,      # Orientate the axis tick labels
xlab = "Species",         # X-axis label
ylab = "Sepal Length",    # Y-axis label
main = "Sepal Length by Species in Iris",
cex.lab = 1.5,  # Size of axis labels
cex.axis = 1.5, # Size of the tick mark labels
cex.main = 2)   #Size of the plot title
```

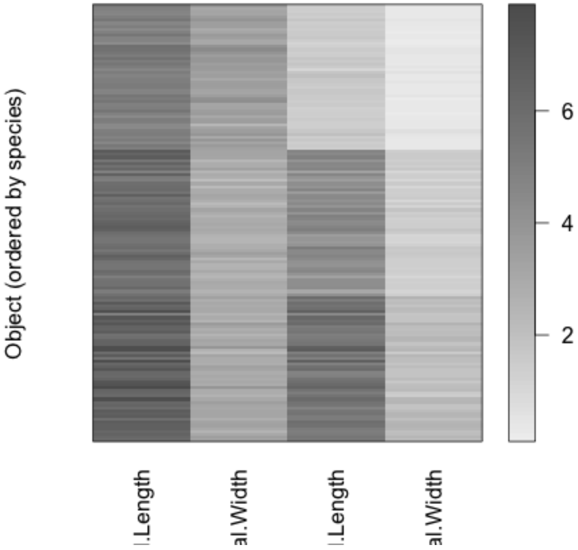


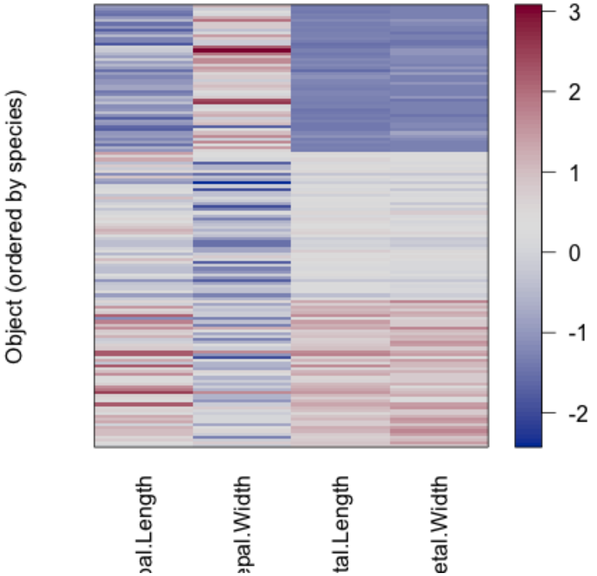
Data Matrix Visualization

Gradation of the color signifies the varied levels in the datase values of a dataset

```
#plotting individual values of the iris dataset
library(seriation) #for pimage
iris_matrix <- as.matrix(iris[,1:4])
pimage(iris_matrix, ylab="Object (ordered by
species)", main="Original values",
colorkey=TRUE)

#values smaller than the average are blue and larger ones
are red
library("colorspace") ### for diverge_hcl
library(seriation) #for pimage
iris_matrix <- as.matrix(iris[,1:4])
pimage(scale(iris_matrix), ylab="Object
 (ordered by species)",
main="Standard deviations from the feature
mean",
range=c(-3.5,3.5), col=diverge_hcl(100),
colorkey=TRUE)
```





Saving Graphs to Files

Saving graphs to a file follows a specific sequence of commands. Below are some of the examples:

```
# Saving a jpeg file in the working directory. The actual image data are not written to the file
#until the 'dev.off()' command is executed!
jpeg("test.jpeg"); plot(1:10, 1:10); dev.off()

# Same as above, but for pdf format. The pdf format provides often the best image quality,
#since it scales to any size.
pdf("test.pdf"); plot(1:10, 1:10); dev.off()

# Same as above, but for png format.
png("test.png"); plot(1:10, 1:10); dev.off()

# Same as above, but for PostScript format.
postscript("test.ps"); plot(1:10, 1:10); dev.off()
```

Graphical Parameters

The following options can be used inside the graph function to control text and symbol size in graphs.

option	description
cex	number indicating the amount by which plotting text and symbols should be scaled relative to the default. 1=default, 1.5 is 50% larger, 0.5 is 50% smaller, etc.
cex.axis	magnification of axis annotation relative to cex
cex.lab	magnification of x and y labels relative to cex
cex.main	magnification of titles relative to cex
cex.sub	magnification of subtitles relative to cex

Advanced Visualization Using R

Many libraries in R which provides the capability of advanced data visualization.

- `tabplotd3()` – *visualization for large dataset with both categorical and numeric variables*
- `metricsgraphics()` – *for advanced scatterplot*
- `dygraphs()` – *Time series plot with basic forecasting using holts winter technique*
- `d3heatmap()` – *heat map with clustering of similar groups*
- `treemap()` – *visualization of large dataset*
- `networkd3()` – *network graphs. Earlier it was d3network()*

More about networkd3() at: https://christophergandrud.github.io/networkD3/

Demo of Sales Dashboard

Summary

Summary of the topics covered in this lesson:

- Data visualization and Descriptive statistics goes hand in hand to summarize and extract useful information from data.
- R provides umpteen number of libraries which can be used to visualize any dataset.
- Scatter plot, box plot, histogram, correlation plot are some of the statistical plots useful in summarizing data.
- The graphs generated using the graph functions can be saved in different file formats using R commands.

QUIZ TIME

Quiz Question 1

Quiz 1

What will be plotted on x-axis and y-axis with the following command?
boxplot(iris$Sepal.Length ~ iris$Species)

a. Sepal Length on x-axis and Species on y-axis.

b. Sepal Length on y-axis and Species on x-axis.

c. Syntax incomplete. Graph will not be plotted.

d. Syntax complete but x and y axis plot not defined.

Quiz Question 1

Quiz 1

What will be plotted on x-axis and y-axis with the following command?
boxplot(iris$Sepal.Length ~ iris$Species)

a. Sepal Length on x-axis and Species on y-axis.

b. Sepal Length on y-axis and Species on x-axis.

c. Syntax incomplete. Graph will not be plotted.

d. Syntax complete but x and y axis plot not defined.

Correct answer is:

b

The first parameter in the boxplot represents y-axis variable and second parameter represents x-axis variable.

End of Lesson04–Data Visualization using R