Unit II: Data Visualization using R

Jamuna S Murthy
Assistant Professor
Department of CSE

Academic Year - 2025-26

1 Prerequisites : Loading Datasets

Loading Datasets

R provides multiple methods to load datasets, including CSV, Excel, and built-in datasets.

Loading CSV Files

```
data <- read.csv("data.csv")
print(head(data))
```

Loading Tabular Data with read.table()

```
data <- read.table("data.txt", header=TRUE, sep="\t")
print((head(data))
```

Loading Excel Files

Note: Requires `readxl` package.

```
library(readxl)
data <- read_excel("data.xlsx")
print(head(data))
```

Using Built-in Datasets

R has built-in datasets such as `mtcars` and `iris`.

```r
data(mtcars)
print(head(mtcars))

data(iris)
print(head(iris))
```

Preprocessing Data

Preprocessing is crucial to prepare data for analysis and visualization. In R, the $ operator is used to access a specific column (or component) from a dataset, dataframe, or list.

Handling Missing Values

Removing rows with missing values (example using `airquality`):

```r
data_clean <- na.omit(airquality)
print(head(data_clean))
```

Ignoring missing values in calculations:

```r
mean(airquality$Ozone, na.rm = TRUE)
```

Data Transformation

Transformations include scaling and normalization (example using `mtcars`):

```r
mtcars$scaled_mpg <- scale(mtcars$mpg)
print(head(mtcars))
```

Factorizing Categorical Variables

Convert categorical variables to factors (example using `iris`):

```r
iris$Species <- factor(iris$Species)
levels(iris$Species)
```

Subsetting Data

Subset data using specific conditions (example using `mtcars`):

```r
subset_data <- subset(mtcars, hp > 100)
print(head(subset_data))
```

2 Scatter Plots

A **scatter plot** visualizes relationships between two numeric variables.

2.1 Examples

```

Synatax: plot(x, y, main, xlab, ylab, col, pch, xlim, ylim)

x and y: Numeric vectors to plot.
main: Title of the plot.
xlab, ylab: Axis labels.
col: Color of points.
pch: Point shape (pch=19 is solid circle).
xlim, ylim: Range of x and y axes (optional).
```

```
Example 1: Iris dataset

plot(iris$Sepal.Length, iris$Petal.Length,
main="Sepal vs Petal Length", xlab="Sepal Length",
ylab="Petal Length", col="blue", pch=19)

Example 2: mtcars dataset

plot(mtcars$hp, mtcars$mpg, main="Horsepower vs MPG",
xlab="Horsepower", ylab="Miles per Gallon",
col="red", pch=17)

Example 3: airquality dataset

plot(airquality$Temp, airquality$Ozone, main="Temperature vs Ozone",
xlab="Temperature", ylab="Ozone", col="green", pch=15)
```

3 Box Plots

A **box plot** summarizes data distribution, highlighting medians, quartiles, and outliers.

3.1 Examples

```

Syntax: boxplot(x, main, xlab, ylab, col)

x: Numeric vector or formula (y ~ group).
Visualizes data spread, median, quartiles, outliers.
col: Color of the box.
```

```
Example 1: Iris dataset

boxplot(Sepal.Width ~ Species, data=iris,
main="Sepal Width by Species", col="lightblue")

Example 2: mtcars dataset

boxplot(mtcars$mpg, main="MPG Distribution", col="orange")

Example 3: airquality dataset

boxplot(airquality$Ozone ~ airquality$Month,
main="Ozone by Month", col="lightgreen")
```

4 Scatter Plots and Box-and-Whisker Plots Together

Combining scatter plots and box plots helps in comparative analysis.

4.1 Examples

```

Synatx: layout(matrix(c(1,2), nrow=1, ncol=2))
Use layout() to display multiple plots simultaneously.
```

```
Example 1: Iris dataset

layout(matrix(c(1,2), 1, 2))
boxplot(iris$Sepal.Length, main="Sepal Length", col="pink")
plot(iris$Sepal.Length, iris$Petal.Length, main="Sepal vs Petal Length",
col="purple", pch=19)

Example 2: mtcars dataset

layout(matrix(c(1,2), 1, 2))
boxplot(mtcars$mpg, main="MPG", col="cyan")
plot(mtcars$hp, mtcars$mpg, main="HP vs MPG", col="darkred", pch=17)

Example 3: airquality dataset

layout(matrix(c(1,2), 1, 2))
boxplot(airquality$Ozone, main="Ozone Levels", col="yellow")
plot(airquality$Temp, airquality$Ozone, main="Temp vs Ozone",
col="brown", pch=15)
```

5 Customize Plot Axes, Labels, Legends, and Colors

You can customize axes ranges (xlim, ylim) and labels (xlab, ylab. xlim and ylim control axes ranges explicitly.

5.1 Examples

```

Example 1: Customizing Axes and Labels (mtcars)

plot(mtcars$wt, mtcars$mpg, main="Weight vs MPG",
xlab="Weight (1000 lbs)", ylab="Miles per Gallon",
xlim=c(1,6), ylim=c(10,35), col="darkorange", pch=19)
```

```

Example 2: Adding Legends (Iris)

Syntax: legend(position, legend, col, pch, title)
position: "topright", "bottomleft", etc.
legend: Names of categories.
title: Optional title for the legend.

plot(iris$Sepal.Length, iris$Petal.Length,
col=c("red","blue","green")[iris$Species], pch=19,
main="Iris Species Sepal vs Petal Length")
legend("bottomright", legend=levels(iris$Species),
col=c("red","blue","green"), pch=19)
```

```
Example 3: Adding Colors (airquality)

month_factor <- factor(airquality$Month)
plot(airquality$Temp, airquality$Ozone,
col=rainbow(length(levels(month_factor)))[month_factor], pch=19,
main="Ozone Levels Colored by Month")
legend("topright", legend=levels(month_factor),
col=rainbow(length(levels(month_factor))), pch=19)
```

6 Practice Questions

Question 1: Iris Dataset

1. Load the built-in `iris` dataset.
2. Check for missing values and remove them if present.
3. Convert the `Species` column to a factor type.
4. Create a box plot of `Sepal.Length` grouped by `Species`.

Question 2: mtcars Dataset

1. Load the built-in `mtcars` dataset.
2. Scale the `mpg` (miles per gallon) column.
3. Subset the dataset to include only cars with `hp` (horsepower) greater than 100.
4. Visualize the relationship between `hp` and `mpg` using a scatter plot.

Question 3: airquality Dataset

1. Load the built-in `airquality` dataset.
2. Handle missing values appropriately.
3. Create a subset including observations where `Temp` ¿ 80.
4. Visualize the distribution of `Ozone` using a histogram.