**Project**

**Student Name:** Saransh Arora **UID:** 24MCI10041

**Branch:** MCA(AIML) **Section/Group:** 1(B)

**Semester:** 1st **Date of Performance:** 20-10-2024

**Subject Name:** Statistical Techniques using R Lab **Subject Code:** 24CAP-614

1. **Aim:**

Choose a dataset from a repository like Kaggle or UCI Machine Learning Repository and perform exploratory data analysis using R. Explore the distribution of variables, identify outliers, and visualize relationships between variables using plots like histograms and boxplots.

1. **Task to be done:**

**Choosing a Dataset:**

First, you need to choose a dataset from repositories like Kaggle or the [UCI](https://archive.ics.uci.edu/ml/index.php) [Machine](https://archive.ics.uci.edu/ml/index.php) [Learning Repository.](https://archive.ics.uci.edu/ml/index.php)

Some popular datasets include the Iris dataset, Titanic dataset, Forest+Fires or the Wine Quality dataset.

**Loading the Dataset:**

After downloading the dataset, you can load it into your R environment using the read.csv() function or another appropriate method based on the dataset format (CSV, Excel, etc.).

**Basic Exploratory Data Analysis:**

**Summary Statistics:** Get a sense of your data by running basic summary statistics (mean, median, mode, etc.).

**Distribution of Variables:** Use histograms, boxplots, and density plots to understand the distribution of numeric variables.

**Identifying Outliers:** Boxplots and statistical tests can help in identifying outliers in your dataset.

**Visualization:**

**Histograms:** Show the distribution of individual numeric variables.

**Boxplots:** Identify outliers and visualize the spread of a variable.

1. **Code for experiment/ Output:**

**Unzip the File**

**Purpose**: Unzip the dataset and list the files.

**Steps**:

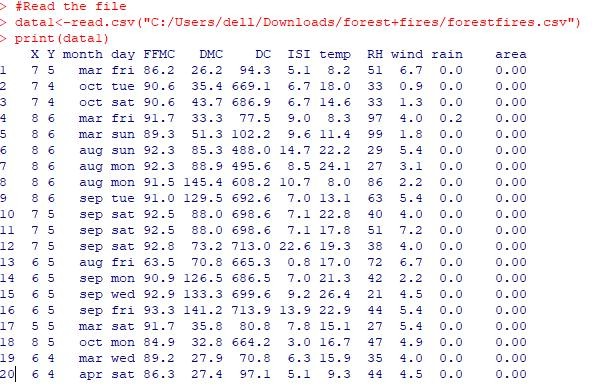
* + - * 1. Make sure the file path is correct.
        2. The unzip() function extracts the files and stores them in the specified directory.
        3. list.files() lists the unzipped files in that directory.

**Read the File**

**Purpose**: Load the CSV file into R.

**Steps**:

* + - 1. Use read.csv() to load the data into data1.
      2. print(data1) displays the dataset in the console.

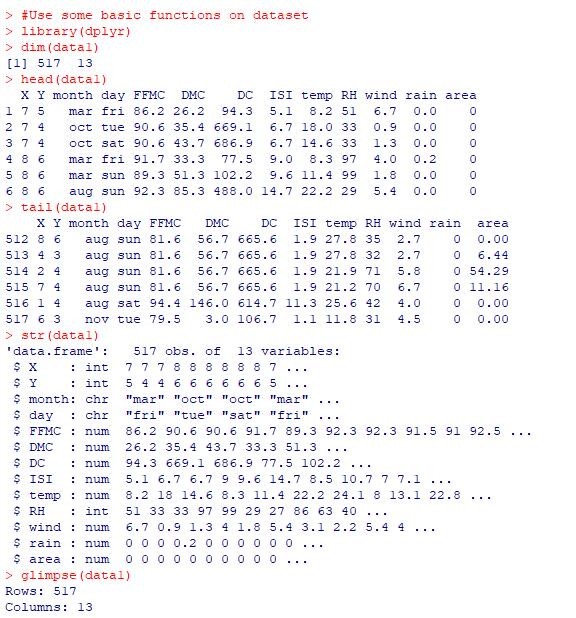


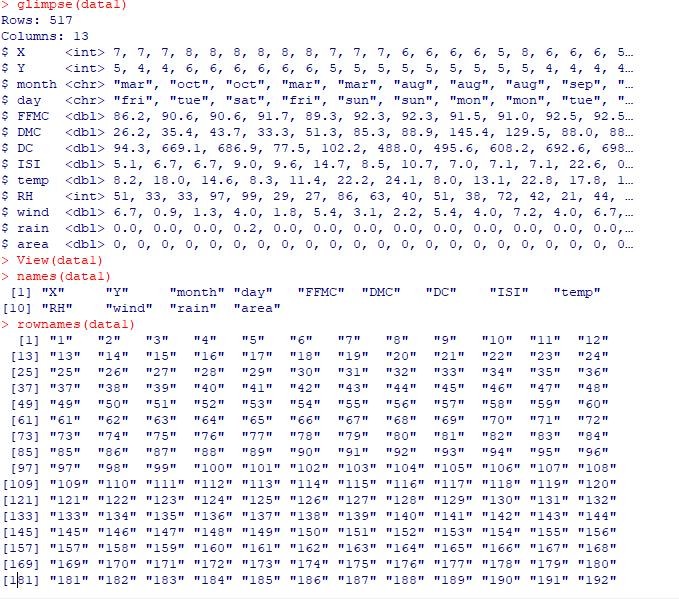
**Basic Dataset Functions**

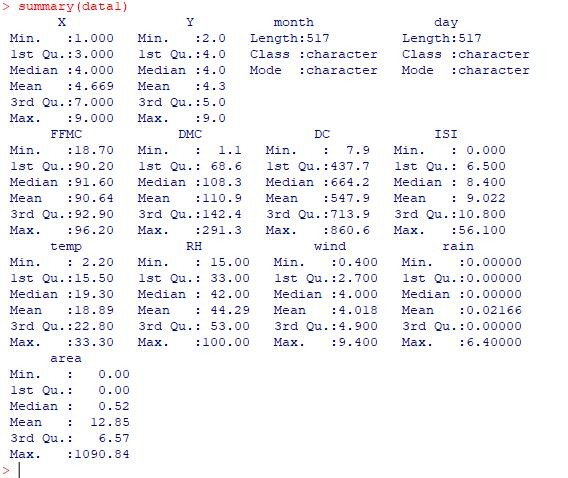
**Purpose**: Perform basic dataset operations and inspections.

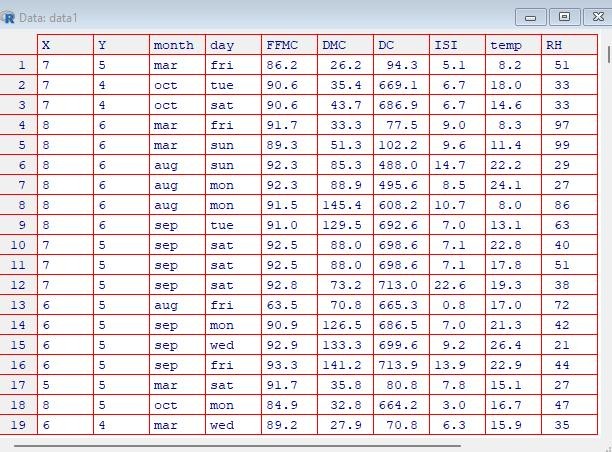
**Steps**:

* 1. dim() gives the dimensions of the dataset (number of rows and columns).
  2. head() and tail() show the first and last rows.
  3. str() and glimpse() provide an overview of the data structure.
  4. View() opens the dataset in a spreadsheet-like viewer in RStudio.
  5. names() and rownames() show column and row names.
  6. summary() provides basic descriptive statistics for each variable.







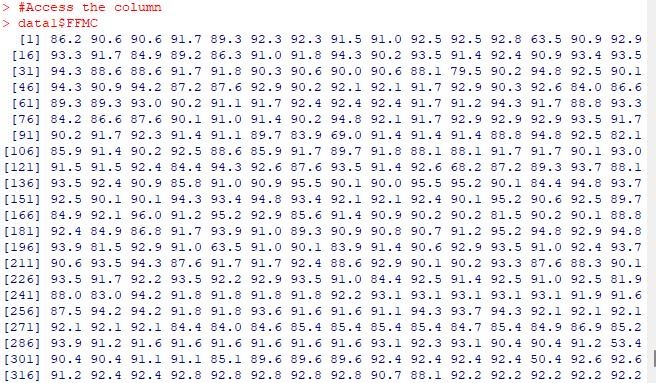


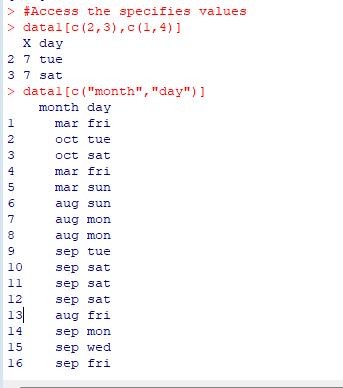
**Access Columns and Specific Values**

**Purpose**: Access specific columns and rows in the dataset.

**Steps**:

* 1. Use data1$FFMC to access the FFMC column.
  2. Use data1[c(2, 3), c(1, 4)] to access specific rows (2, 3) and columns (1, 4).
  3. data1[c("month", "day")] accesses multiple columns by name.



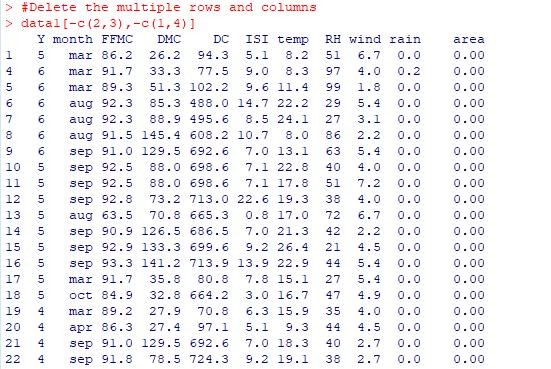


**Delete Rows and Columns**

**Purpose**: Remove specific rows and columns from the dataset.

**Steps**:

* 1. -c(2, 3) removes rows 2 and 3.
  2. -c(1, 4) removes columns 1 and 4.

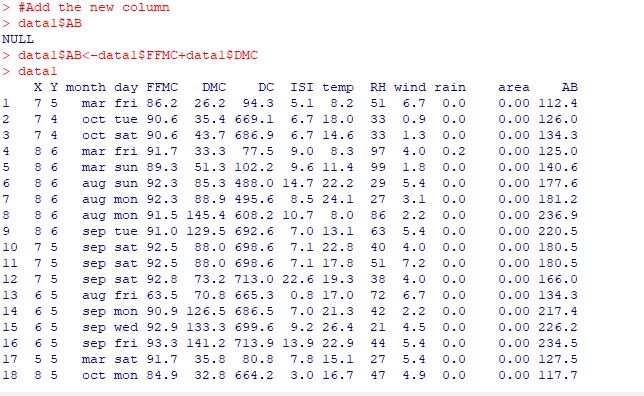


**Add New Column**

**Purpose**: Create a new column AB as a sum of two existing columns.

**Steps**:

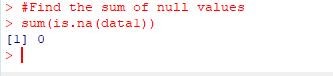
* 1. Create the new column by adding FFMC and DMC.
  2. Print the dataset to view the new column.

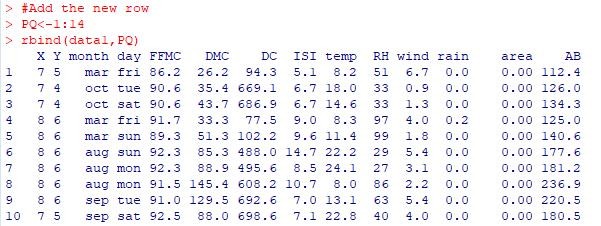


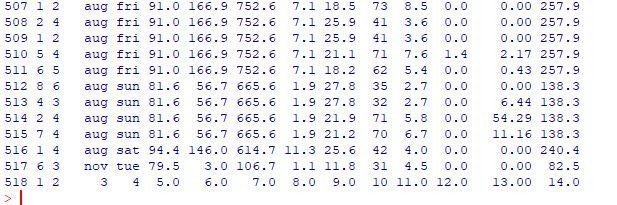
**Add New Row**

**Purpose**: Add a new row to the dataset using rbind().

**Steps**:

* 1. Create a vector PQ with 14 values (same as the number of columns in the dataset).
  2. Use rbind() to append the new row to data1.





**Sum of Missing Values**

**Purpose**: Find the total number of missing values in the dataset.

**Steps**:

* 1. is.na() checks for missing values in the dataset.
  2. sum() calculates the total number of missing values.

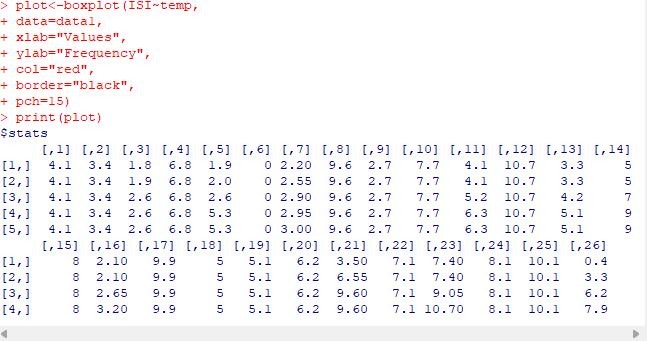
**Boxplot**

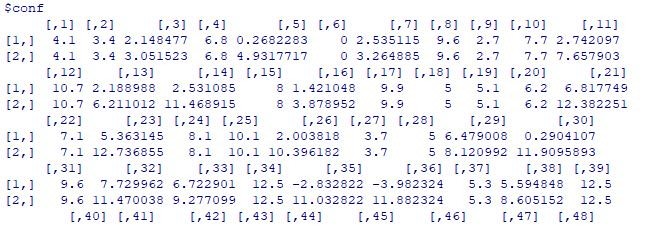
**Purpose**: Create a boxplot of ISI based on temp.

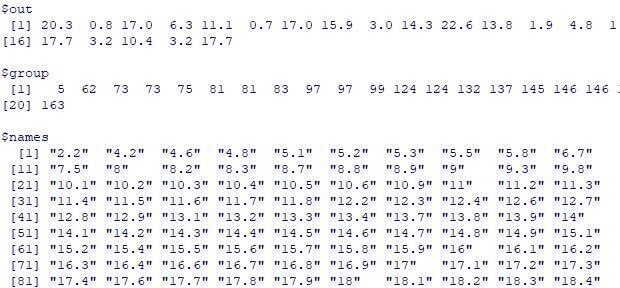
**Purpose**: Create a boxplot for RH.

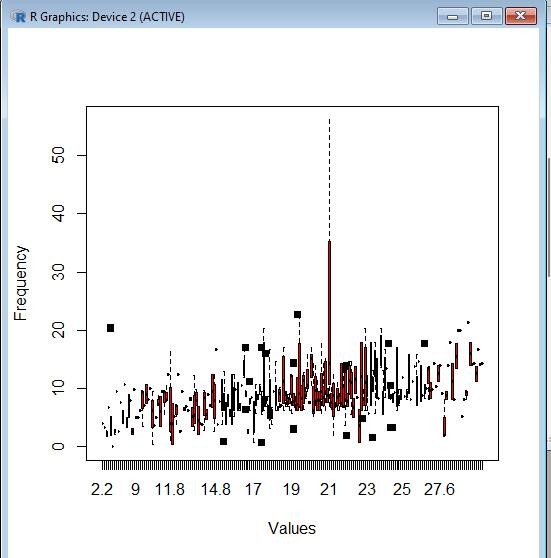
**Steps**:

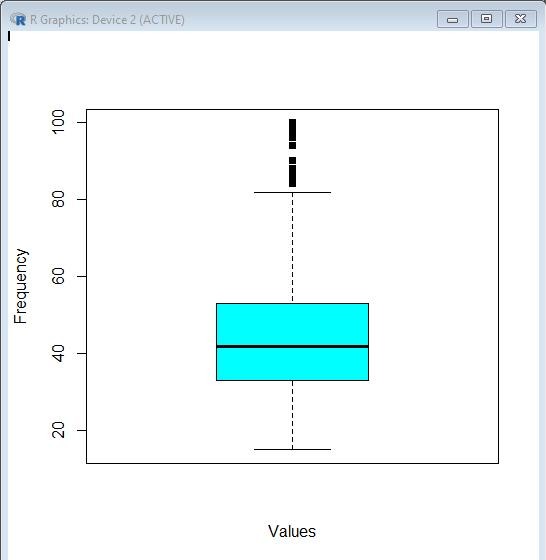
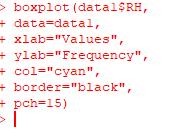
* 1. boxplot() creates a boxplot of ISI against temp.
  2. Customize the plot with labels, colors, and point characteristics (pch).
  3. print() to display the boxplot.
  4. boxplot() plots the boxplot for RH with customized colors and labels.









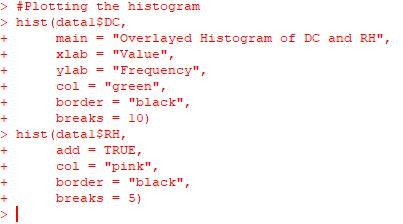


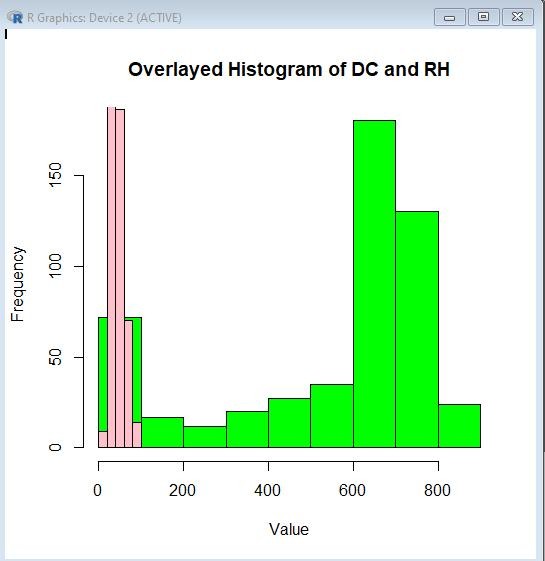
**Histograms**

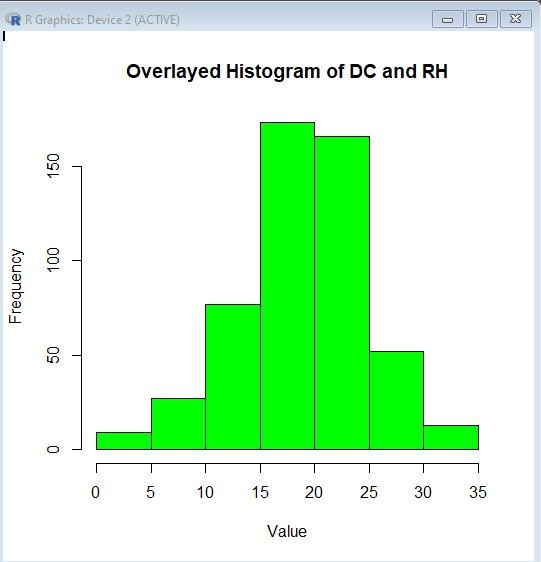
**Purpose**: Plot histograms for variables DC, RH, and temp.

**Steps**:

* + 1. Use hist() to create histograms for each variable.
    2. add = TRUE overlays the second histogram (RH) on the first (DC).







1. **LEARNING OUTCOMES:**

**Understand Dataset Structures**:

Learn how to inspect datasets and check data types, missing values, and basic statistics.

**Data Cleaning**:

Gain skills in identifying and handling missing values, outliers, and inconsistent data.

**Explore Distributions**:

Use histograms and boxplots to explore the distribution of numerical variables and identify skewness or outliers.

**Visualize Relationships**:

Use scatter plots and pair plots to visualize and interpret relationships between variables.

**Identify Outliers**:

Develop techniques to spot outliers using visual tools like boxplots and numerical methods like IQR.

**Perform Correlation Analysis**:

Compute and visualize correlations between variables to identify strong relationships.