EXPERIMENT NO: 07

AIM: Perform exploratory data analysis (EDA) using R by importing, cleaning, and visualizing data to extract insights and understand data distributions.(na,summary,plot,hist,boxplot)

What is EDA?

Exploratory Data Analysis (EDA) is the process of examining and summarizing data to understand its structure, detect patterns, and identify anomalies before applying statistical models. It involves data cleaning, summarization, and visualization to extract useful insights.

Steps in EDA

EDA typically involves the following steps:

1. **Importing Data**
   Load the dataset into R using functions like read.csv() for structured data files.
2. **Data Cleaning**
   - Handling **missing values** (NA)
   - Detecting and removing **duplicates**
   - Converting data types (e.g., factor to numeric)
   - Removing **outliers** if necessary
3. **Descriptive Statistics**
   - **Summary statistics** (summary()) provide a numerical overview of data, including mean, median, min, max, and quartiles.
   - **Structure of data** (str()) helps in understanding variable types and dimensions.
4. **Data Visualization**
   - Helps in identifying trends, relationships, and distributions.
   - Common types:
     - **Histogram (hist())** → Shows the distribution of a single numeric variable.
     - **Boxplot (boxplot())** → Detects outliers and visualizes spread.
     - **Scatter plot (plot())** → Shows relationships between two numerical variables.
     - **Bar plot (barplot())** → Displays categorical variable distributions.
5. **Correlation Analysis**
   - Determines how two numeric variables are related.

- The correlation coefficient (cor()) helps measure the strength and direction of relationships.

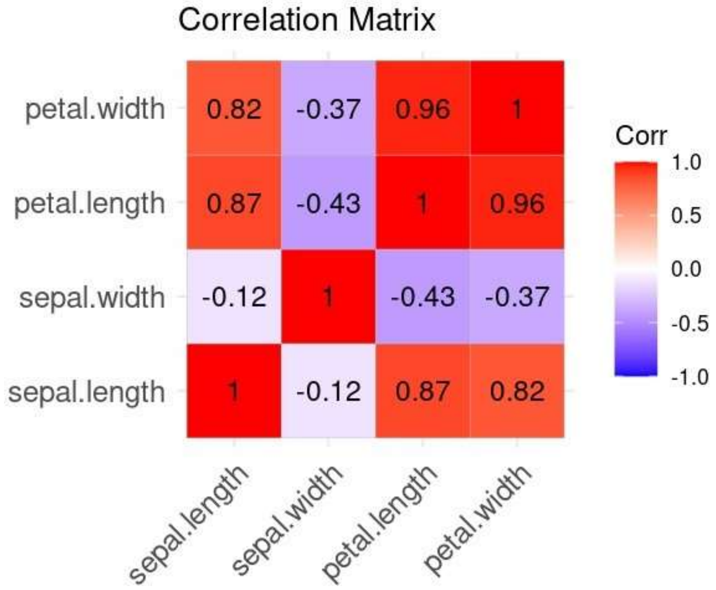
Importance of EDA

- Detects missing values and anomalies.
- Identifies data distributions and relationships.
- Helps in feature selection for machine learning models.
- Provides insights into trends and patterns before modeling.

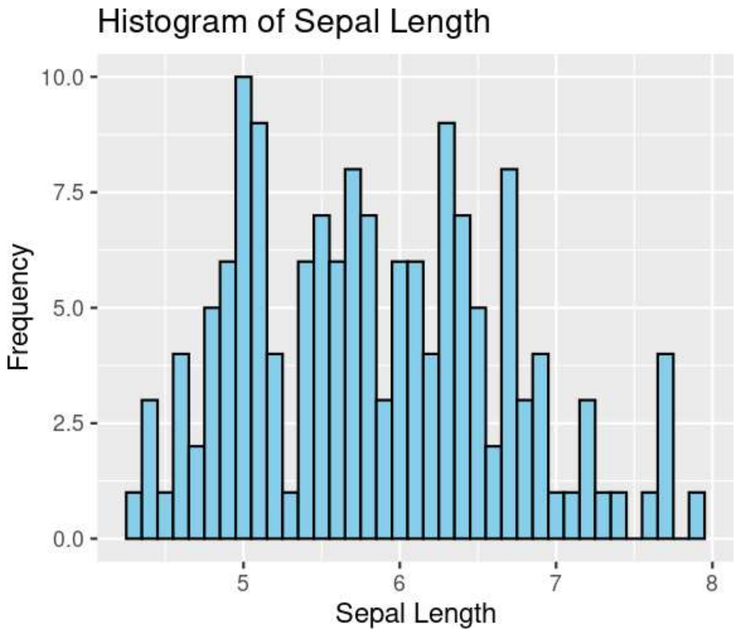
CODE:

```
> library(ggcorrplot)
> library(ggplot2)
> library(dplyr)
> library(tidyr)
> library(summarytools)
> library(ggcorrplot)
> setwd("/home/a/aman/")
> data <- read.csv("iris.csv")
> head(data)
  sepal.length sepal.width petal.length petal.width variety

1          5.1         3.5          1.4         0.2  Setosa
2          4.9         3.0          1.4         0.2  Setosa
3          4.7         3.2          1.3         0.2  Setosa
4          4.6         3.1          1.5         0.2  Setosa
5          5.0         3.6          1.4         0.2  Setosa
6          5.4         3.9          1.7         0.4  Setosa
>
> #checking for missing values
> sum(is.na(data))
[1] 0
>
> #remove rows with missing values
> data <-na.omit(data)
> data$sepal.length[is.na(data$sepal.length)]<-mean(data$sepal.length,na.rm= TRUE)
> summary(data)
  sepal.length    sepal.width     petal.length    petal.width
        variety Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.
        :0.100   Length:150
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character
 Median          Median          Median          Median              Mode
 :5.800          :3.000          :4.350          :1.300            :character
 Mea    :5.843   Mea    :3.057   Mea    :3.758   Mea    :1.199
 n               n               n               n
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
> #visualizing data
> #correlation plot
> cor_matrix<-cor(data[,1:4])
> ggcorrplot(cor_matrix,lab = TRUE, title= "Correlation Matrix")
```



```
> ggplot(data, aes(x = sepal.length)) +
+   geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
+   labs(title = "Histogram of Sepal Length", x = "Sepal Length", y = "Frequency")
```



CONCLUSION: Hence, we successfully implemented EDA is a crucial step in data analysis, ensuring data quality and providing insights before further statistical analysis or modelling. In R, functions like summary(), hist(), boxplot(), plot(), and cor() help in performing effective EDA.