

**Class Task**

**Submitted by:**

**Name: Jameel Ahmad**

**Reg# : SP23-BCS-057**

**Submitted to:**

**Mr. Dr Fakher Mustafa**

**Subject:**

**Introduction to DataScience**

**Deadline:**

**Tue, 21 Oct**

**Computer Science Department**

**COMSATS University Islamabad, Sahiwal Campus**

**Q No. 1**

1. Discuss the significance of data science in enhancing decision-making processes across various industries. Provide specific examples from sectors such as healthcare and finance to illustrate your points.

**Solution:**

Data science plays a **vital role in improving decision-making** by transforming raw data into meaningful insights through statistical analysis, machine learning, and visualization. It helps organizations move from intuition-based decisions to **data-driven strategies** that improve accuracy, efficiency, and profitability.

### ****Examples:****

🔹 **Healthcare Sector:**

* Data science enables **predictive analytics** to forecast disease outbreaks and patient readmissions.
* Machine learning models analyze medical images (e.g., X-rays, MRIs) for **early diagnosis** of diseases such as cancer.
* Hospitals use patient data to **personalize treatment plans**, improving outcomes and reducing costs.

🔹 **Finance Sector:**

* Banks use data science for **fraud detection** by identifying unusual transaction patterns.
* Financial institutions use **risk assessment models** to evaluate loan applications and credit scoring.
* **Algorithmic trading** systems use real-time data to make faster and more accurate investment decisions.

1. Describe the typical steps in the data science process, including data collection, cleaning, exploration, and modeling. Explain the importance of each step in ensuring effective data analysis.

**Solution:**

The data science process involves several essential steps that ensure the accuracy, reliability, and usefulness of analytical results.

### ****1. Data Collection****

Gathering raw data from various sources such as databases, sensors, surveys, APIs, or web scraping.  
👉 Importance: The quality and relevance of collected data directly affect the accuracy of insights and predictions.

### ****2. Data Cleaning (Preprocessing)****

Removing errors, missing values, duplicates, and inconsistencies from the dataset.  
👉 Importance: Clean data ensures reliable results and prevents misleading conclusions during analysis.

### ****3. Data Exploration and Analysis (EDA)****

Using visualization and statistical tools to understand patterns, correlations, and trends in the data.  
👉 Importance: EDA helps form hypotheses and guides model selection by revealing data characteristics.

### ****4. Data Modeling****

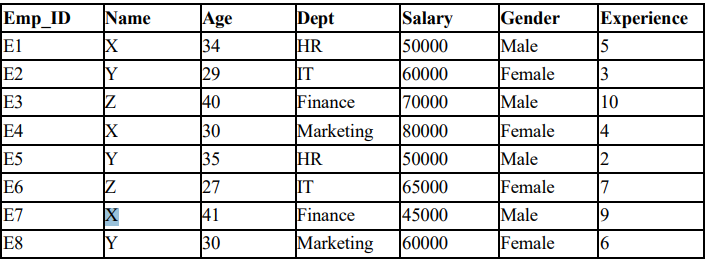
Applying machine learning or statistical models to make predictions or classifications.  
👉 Importance: Models translate insights into actionable decisions, such as predicting demand or detecting fraud.

### ****5. Model Evaluation and Deployment****

Assessing model performance using metrics (accuracy, precision, etc.) and deploying it for real-world use.

**Q No. 2**

Apply the concept of vector and list on below dataset, as well as answer the following questions.



1. **Extract the Salary column from the dataset as a vector and calculate the average salary.**

# Extract Salary as vector

salary\_vector <- employees$Salary

# Calculate average salary

avg\_salary <- mean(salary\_vector)

avg\_salary

1. **Use a vector to store the ages of employees. Find the minimum and maximum age among the employees.**

# Store ages in a vector

age\_vector <- employees$Age

# Find minimum and maximum

min\_age <- min(age\_vector)

max\_age <- max(age\_vector)

min\_age

max\_age

1. **Create a list to store the details of a single employee (e.g., Name, Department, Age, Salary). Display each element of the list.**

# Create a list for one employee (example: Employee 1)

employee1 <- list(

Name = "X",

Department = "HR",

Age = 34,

Salary = 50000

)

# Display each element

employee1$Name

employee1$Department

employee1$Age

employee1$Salary

1. **Explain how lists are advantageous over vectors when storing multiple types of information.**

# Mean and Standard Deviation

mean\_salary <- mean(employees$Salary)

sd\_salary <- sd(employees$Salary)

mean\_experience <- mean(employees$Experience)

sd\_experience <- sd(employees$Experience)

# Correlation between Salary and Experience

cor\_salary\_experience <- cor(employees$Salary, employees$Experience)

# Display results

mean\_salary

sd\_salary

mean\_experience

sd\_experience

cor\_salary\_experience

1. **Using the above-mentioned dataset in R, implement R code to calculate the mean, standard deviation, and correlation between two variables.**

# Mean and Standard Deviation

mean\_salary <- mean(employees$Salary)

sd\_salary <- sd(employees$Salary)

mean\_experience <- mean(employees$Experience)

sd\_experience <- sd(employees$Experience)

# Correlation between Salary and Experience

cor\_salary\_experience <- cor(employees$Salary, employees$Experience)

# Display results

mean\_salary

sd\_salary

mean\_experience

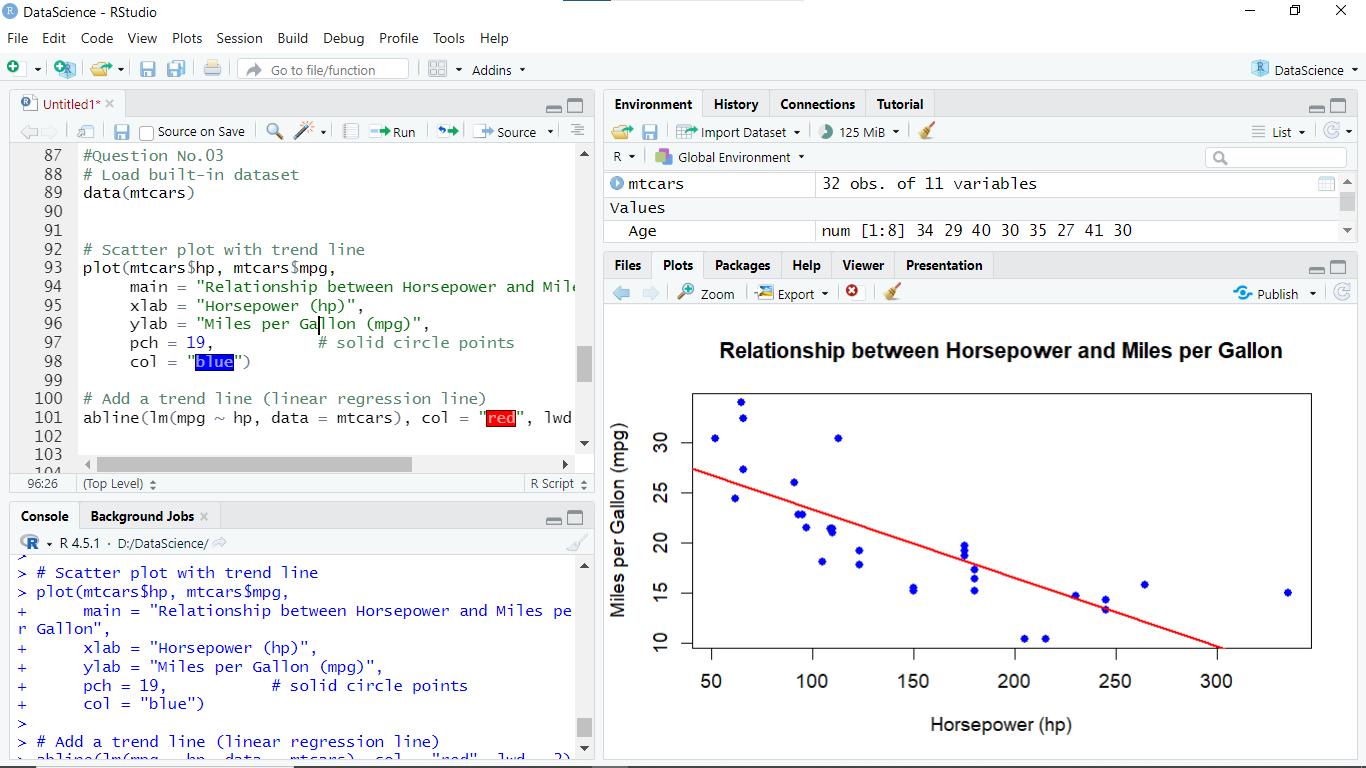
sd\_experience

cor\_salary\_experience

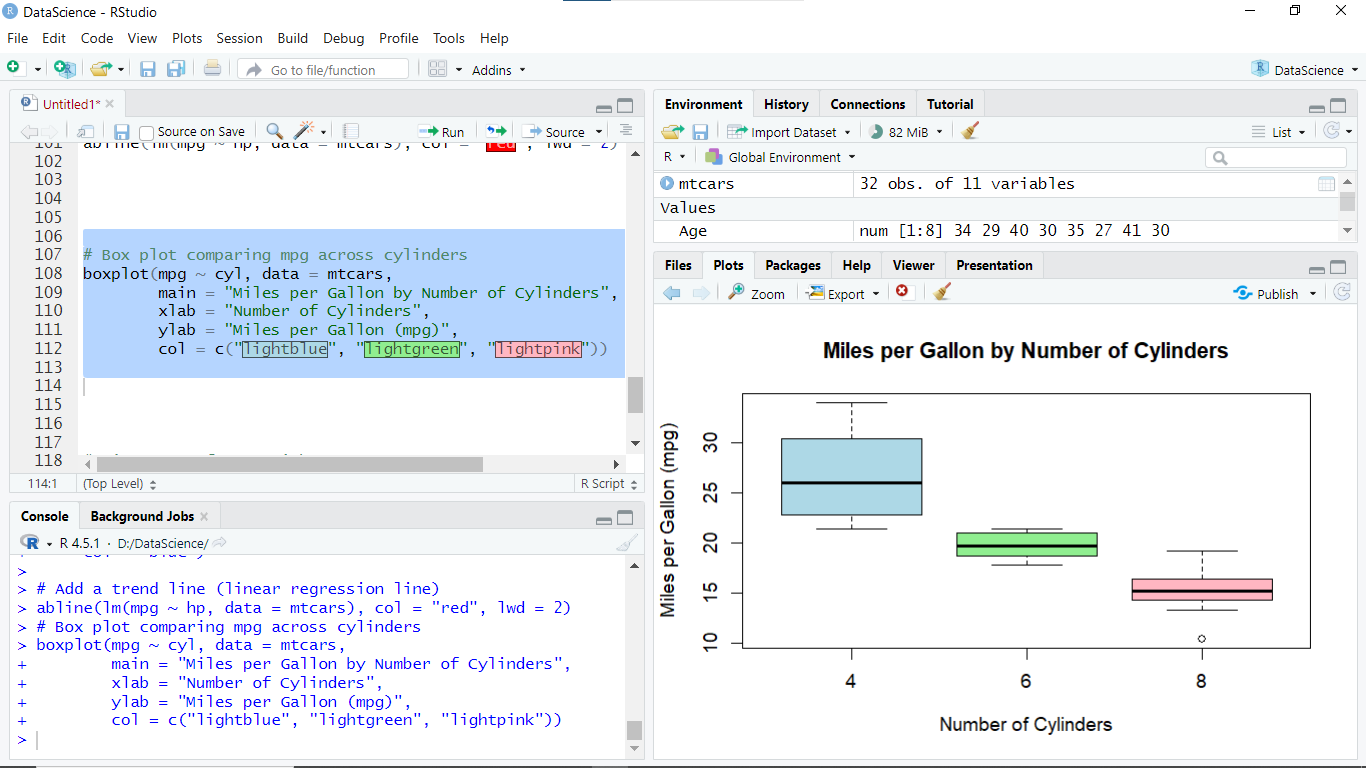
**Q No. 3**

**Analyze and visualize the mtcars dataset, which contains information about various car models, including their miles per gallon (mpg), number of cylinders, horsepower, and weight. You will create three different visualizations to explore relationships and distributions within the data.**

1. Construct a scatter plot to illustrate the relationship between horsepower (hp) and miles per gallon (mpg). Include a trend line to demonstrate the correlation between these two variables. Label the axes and provide a title for the plot.

****

1. Develop a box plot to compare the distribution of miles per gallon (mpg) across different numbers of cylinders (cyl). Ensure the plot is well-labeled with a clear title and axis labels to facilitate understanding of the data.



1. Create a histogram to depict the distribution of car weights (wt). Customize the number of bins to enhance the visualization and provide appropriate titles and axis labels.

