## **Introduction To Data Science**

### Logo, company name Description automatically generated

Submitted By

**Rao Mehtab Ahmed**

**FA22-BCS-102**

Submitted to

**Fakhar Mustafa**

**Department Of Computer Science**

**COMSATS University Islamabad**

**Sahiwal Campus**

**Q No. 1 (3 × 2 = 6) – CLO1-C2-GA1**

**Q1. Discuss the significance of data science in enhancing decision-making processes across various industries.**

Data science plays a vital role in improving decision-making by turning raw data into actionable insights. It enables organizations to make evidence-based and data-driven decisions that increase efficiency and accuracy.  
  
Healthcare: Data science helps analyze patient records, predict disease risks (e.g., heart disease, diabetes), and improve diagnosis through AI-based image analysis. Hospitals also use predictive models to manage staff and resources efficiently.  
Finance: Banks and fintech companies use data science for fraud detection, credit risk scoring, and automated trading strategies. Predictive analytics improves investment decisions and customer targeting.  
Retail: Data-driven recommendations (e.g., on Amazon) and demand forecasting improve customer satisfaction and inventory control.  
Manufacturing: Machine learning models predict machinery failures and optimize production through predictive maintenance.  
  
Overall, data science transforms intuition-based decision-making into analytical, transparent, and measurable processes.

**Q2. Describe the typical steps in the data science process.**

1. Data Collection: Gathering raw data from reliable sources (databases, sensors, APIs, or surveys).  
2. Data Cleaning: Removing duplicates, handling missing values, and correcting inconsistencies.  
3. Data Exploration (EDA): Using descriptive statistics and visualizations to identify trends, outliers, and patterns.  
4. Feature Engineering: Creating meaningful variables or transformations.  
5. Modeling: Applying algorithms (e.g., regression, classification, clustering) to train and evaluate models.  
6. Evaluation: Measuring performance using metrics like accuracy, precision, recall, or RMSE.  
7. Deployment & Monitoring: Implementing models in production and tracking their performance.  
Each step ensures that the data-driven solution is accurate and efficient.

**Q No. 2 (2 × 5 = 10) — Apply vectors and lists on the given dataset**

Dataset includes Emp\_ID, Name, Age, Dept, Salary, Gender, and Experience.

**(i) Extract Salary column and calculate average salary**

R Code:

salary <- c(50000, 60000, 70000, 80000, 50000, 65000, 45000, 60000)  
mean\_salary <- mean(salary)  
mean\_salary

Result: Average Salary = 60000

**(ii) Use a vector to store ages. Find min and max age**

ages <- c(34, 29, 40, 30, 35, 27, 41, 30)  
min(ages)  
max(ages)

Minimum Age = 27, Maximum Age = 41

**(iii) Create a list to store one employee’s details**

emp\_E3 <- list(Name = "Z", Department = "Finance", Age = 40, Salary = 70000)  
emp\_E3

**(iv) Explain how lists are advantageous over vectors**

Vectors can store only one type of data, while lists can hold multiple data types together. Lists are better for complex data such as employee records that contain both numeric and text fields.

**(v) Calculate mean, standard deviation, and correlation**

salary <- c(50000, 60000, 70000, 80000, 50000, 65000, 45000, 60000)  
ages <- c(34, 29, 40, 30, 35, 27, 41, 30)  
mean\_salary <- mean(salary)  
sd\_salary <- sd(salary)  
mean\_age <- mean(ages)  
sd\_age <- sd(ages)  
corr\_age\_salary <- cor(ages, salary, method = "pearson")

Results: Mean(Salary)=60000, SD(Salary)=11649.65, Correlation(Age,Salary)=-0.40

**Q No. 3 (3 × 3 = 9) — Data Visualization on mtcars Dataset**

**(i) Scatter plot: Horsepower vs MPG**

library(ggplot2)  
data(mtcars)  
ggplot(mtcars, aes(x = hp, y = mpg)) +  
 geom\_point(size = 2, color = "blue") +  
 geom\_smooth(method = "lm", se = TRUE, color = "red") +  
 labs(title = "Relationship between Horsepower and MPG",  
 x = "Horsepower (hp)",  
 y = "Miles per Gallon (mpg)") +  
 theme\_minimal()

**(ii) Box plot: MPG by Cylinders**

mtcars$cyl <- as.factor(mtcars$cyl)  
ggplot(mtcars, aes(x = cyl, y = mpg, fill = cyl)) +  
 geom\_boxplot() +  
 labs(title = "MPG Distribution by Cylinder Count",  
 x = "Number of Cylinders",  
 y = "Miles per Gallon (mpg)") +  
 theme\_minimal()

**(iii) Histogram: Distribution of Car Weights**

ggplot(mtcars, aes(x = wt)) +  
 geom\_histogram(bins = 10, color = "black", fill = "lightblue") +  
 labs(title = "Distribution of Car Weights (in 1000 lbs)",  
 x = "Weight (1000 lbs)",  
 y = "Frequency") +  
 theme\_minimal()

**Summary of Key Results**

Average Salary = 60000  
Min Age = 27  
Max Age = 41  
SD (Salary) = 11649.65  
Correlation (Age vs Salary) = −0.40