**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.**

Data Science plays a crucial role in modern decision-making by helping organizations make smart, data-driven choices. It uses techniques such as data mining, machine learning, and statistical analysis to discover patterns and insights from large volumes of data.

**Importance of Data Science in Decision Making:**

* **Improved Accuracy**: Data science helps companies base their decisions on real evidence rather than guesswork.
* **Faster Decisions**: Automated systems powered by machine learning can make quick decisions in real-time.
* **Cost Reduction**: By predicting future trends, companies can reduce waste and save money.
* **Customer Understanding**: Businesses can understand customer behavior and preferences better, leading to improved services.

**Industry Examples:**

**Healthcare**:

* **Predictive Analysis**: Hospitals use patient data to predict disease outbreaks or risks (e.g., diabetes or cancer prediction using machine learning).
* **Diagnosis Support**: AI models help doctors diagnose conditions faster using medical imaging (e.g., X-rays, MRI).
* **Personalized Treatment**: Data helps in creating treatment plans suited to individual patients based on their health records.

**Finance**:

* **Fraud Detection**: Banks analyze transaction data to detect unusual patterns and prevent fraud in real-time.
* **Risk Assessment**: Credit scoring models use historical data to evaluate if a customer is eligible for a loan.
* **Stock Market Predictions**: Investment firms use data science to forecast stock price movements.

**Conclusion:**

In summary, data science provides powerful tools for enhancing decision-making in various sectors. Whether it's diagnosing diseases or detecting fraud, it helps organizations act with confidence and efficiency.

**2. 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.**

The data science process follows a structured set of steps to ensure that meaningful insights can be extracted from raw data.

**Steps in the Data Science Process:**

|  |  |  |
| --- | --- | --- |
| **1** | Data Collection | The first step involves gathering data from different sources such as sensors, web APIs, databases, or surveys. High-quality data is the foundation of any good analysis. |
| **2** | Data Cleaning | Raw data often has missing, duplicate, or incorrect values. In this step, we clean the data by handling missing values, removing outliers, and correcting errors. |
| **3** | Data Exploration | Also called Exploratory Data Analysis (EDA). We use visualizations (like histograms, scatter plots) and statistical summaries to understand patterns,statistical summaries to understand patterns, relationships, and distributions in the data. |
| **4** | Data Modeling | Here we apply machine learning algorithms or statistical models to build prediction or classification systems |
| **5** | Evaluation | We test how well our model is performing using metrics like accuracy, precision, recall, etc. If the model doesn’t perform well, we may go back and improve earlier steps. |
| **6** | Deployment | The final model is integrated into the business system where it provides real-time predictions or analysis. E.g., recommendation systems on Netflix or fraud detection systems in banks. |

**Importance of Each Step:**

1. If data collection is poor, the entire process becomes weak.
2. Cleaning ensures the analysis is accurate and trustworthy.
3. Exploration helps identify hidden trends or patterns.
4. Modeling is where real intelligence is built.
5. Evaluation helps us trust our model.
6. Deployment brings the model into real-world use and delivers value to the organization.

**Conclusion:**

Each step of the data science process is critical and contributes to the success of the final output. Skipping or poorly performing any step may lead to incorrect or misleading results.

## Q No. 2 – Apply the concept of Vector and List (in ****R****)

## ****Dataset Given (Employee Table)****

**# Step 1**:

Create the employee\_data dataset

employee\_data <- data.frame(

Emp\_ID = c("E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8"),

Name = c("X", "Y", "Z", "X", "Y", "Z", "X", "Y"),

Age = c(34, 29, 40, 30, 35, 27, 41, 30),

Dept = c("HR", "IT", "Finance", "Marketing", "HR", "IT", "Finance", "Marketing"),

Salary = c(50000, 60000, 70000, 80000, 50000, 65000, 45000, 60000),

Gender = c("Male", "Female", "Male", "Female", "Male", "Female", "Male", "Female"),

Experience = c(5, 3, 10, 4, 2, 7, 9, 6)

)

**# Step 2:** Print the dataset

print(employee\_data)

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

***# Employee Salary Vector***

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

**# Average Salary**

avg\_salary <- mean(salary)

# Output

print(paste("Average Salary is:", avg\_salary))

**(ii) Use a vector to store the ages of employees. Find the minimum and maximum age.**

# Employee Ages Vector

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

# Minimum and Maximum Age

min\_age <- min(ages)

max\_age <- max(ages)

# Output

print(paste("Minimum Age is:", min\_age))

print(paste("Maximum Age is:", max\_age))

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

Let’s take employee E1 as an example:

# Employee E1 Details as a List

employee\_E1 <- list(

Name = "X",

Department = "HR",

Age = 34,

Salary = 50000

)

# Displaying each element

print(paste("Name:", employee\_E1$Name))

print(paste("Department:", employee\_E1$Department))

print(paste("Age:", employee\_E1$Age))

print(paste("Salary:", employee\_E1$Salary))

**(iv) Explain how lists are advantageous over vectors when storing multiple types of information**

Vectors in R can only store one type of data (all numbers, or all characters).

Lists can store multiple data types (numbers, strings, vectors, even other lists).

For example, one employee's data includes:

* Name (character)
* Age (numeric)
* Salary (numeric)
* Department (character)

So we use a list, not a vector, to store such mixed-type information.

**(v) Using the above dataset in R, implement R code to calculate the mean, standard deviation, and correlation between two variables**

**Let’s take Age and Salary for this:**

# Vectors

age <- c(34, 29, 40, 30, 35, 27, 41, 30)

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

# Mean

mean\_age <- mean(age)

mean\_salary <- mean(salary)

# Standard Deviation

sd\_age <- sd(age)

sd\_salary <- sd(salary)

# Correlation between Age and Salary

correlation <- cor(age, salary)

# Output

print(paste("Mean Age:", mean\_age))

print(paste("Mean Salary:", mean\_salary))

print(paste("SD of Age:", sd\_age))

print(paste("SD of Salary:", sd\_salary))

print(paste("Correlation between Age and Salary:", correlation))

**Question #03**

mtcars dataset, specifically:  
1.Scatter plot (horsepower vs mpg)  
2. Box plot (mpg vs number of cylinders)  
3.Histogram (distribution of weights

**R code solution**

# Load the mtcars dataset

data(mtcars)

# View first few rows (optional)

head(mtcars)

# **i. Scatter plot:** horsepower vs miles per gallon

plot(mtcars$hp, mtcars$mpg,

main = "Relationship between Horsepower and Miles per Gallon",

xlab = "Horsepower (hp)",

ylab = "Miles per Gallon (mpg)",

pch = 19, col = "darkblue")

# Add a regression line

abline(lm(mpg ~ hp, data = mtcars), col = "red", lwd = 2)

# **ii. Box plot:** miles per gallon across different cylinder counts

boxplot(mpg ~ cyl, data = mtcars,

main = "Miles per Gallon by Number of Cylinders",

xlab = "Number of Cylinders",

ylab = "Miles per Gallon (mpg)",

col = c("lightblue", "lightgreen", "lightpink"))

**# iii. Histogram**: distribution of car weights

hist(mtcars$wt,

breaks = 8, # Customize number of bins

main = "Distribution of Car Weights",

xlab = "Weight (1000 lbs)",

ylab = "Frequency",

col = "lightcoral", border = "white")

Explanation

**i. Scatter Plot**

plot() creates the scatter plot.

abline() adds a red regression line to show correlation between horsepower and mpg.

**ii. Box Plot**

boxplot(mpg ~ cyl, ...) compares mpg for 4, 6, and 8 cylinder cars.

Colors added for better visual distinction.

**iii. Histogram**

hist() visualizes the distribution of weights (wt).

breaks = 8 sets number of bins for clarity.