from pyspark.sql import SparkSession

from pyspark.sql.functions import col

spark = SparkSession.builder.appName("Employee Data Analysis").getOrCreate()

data = [

    (1, 'Arjun', 'IT', 75000),

    (2, 'Vijay', 'Finance', 85000),

    (3, 'Shalini', 'IT', 90000),

    (4, 'Sneha', 'HR', 50000),

    (5, 'Rahul', 'Finance', 60000),

    (6, 'Amit', 'IT', 55000)

]

# Define schema

columns = ['EmployeeID', 'EmployeeName', 'Department', 'Salary']

# Create DataFrame

employee\_df = spark.createDataFrame(data, columns)

# Show the DataFrame

employee\_df.show()

# Tasks

# Task 1: Filter Employees by Salary\*\*

#    Filter the employees who have a salary greater than 60,000 and display the result.

employee\_df.filter(col("Salary")>60000).show()

# Task 2: Calculate the Average Salary by Department\*\*

#    Group the employees by department and calculate the average salary for each department

avg\_dept\_salary\_df =  employee\_df.groupBy("Department").avg("Salary").withColumnRenamed("avg(Salary)","AvgSalary")

avg\_dept\_salary\_df.show()

# Task 3: Sort Employees by Salary\*\*

#    Sort the employees in descending order of their salary

salary\_order\_df = employee\_df.orderBy(col("Salary").desc())

salary\_order\_df.show()

# Task 4: Add a Bonus Column\*\*

#    Add a new column called `Bonus` which should be 10% of the employee's salary.

bonus\_df = employee\_df.withColumn("Bonus", col("Salary") \* 1.10)

bonus\_df.show()