import pandas as pd

# Create a sample CSV data

data = {

"name": ["John", "Jane", "Mike", "Emily", "Alex"],

"age": [28, 32, 45, 23, 36],

"gender": ["Male", "Female", "Male", "Female", "Male"],

"salary": [60000, 72000, 84000, 52000, 67000]

}

df = pd.DataFrame(data)

# Save the DataFrame as a CSV file

csv\_file\_path = "/content/sample\_people.csv"

df.to\_csv(csv\_file\_path, index=False)

# Confirm the CSV file is created

print(f"CSV file created at: {csv\_file\_path}")

from pyspark.sql import SparkSession

# Initialize Spark session

spark = SparkSession.builder.appName("Employee Salary ETL").getOrCreate()

# Load CSV data

employee\_df = spark.read.csv("sample\_people.csv", header=True, inferSchema=True)

employee\_df.show()

# Filter employees with age >= 30

filtered\_df = employee\_df.filter(employee\_df.age >= 30)

print("Filtered DataFrame:")

filtered\_df.show()

# Add a new column 'salary\_with\_bonus' (10% bonus on current salary)

from pyspark.sql.functions import col

transformed\_df = filtered\_df.withColumn("salary\_with\_bonus", col("salary") \* 1.10)

print("Transformed DataFrame:")

transformed\_df.show()

# Calculate average salary by gender

avg\_salary\_by\_gender = transformed\_df.groupBy("gender").avg("salary")

print("Average Salary by Gender:")

avg\_salary\_by\_gender.show()

# Save the transformed data in Parquet format

transformed\_df.write.parquet("transformed\_sample\_people.parquet")