import pandas as pd  
from pyspark.sql.functions import col  
from pyspark.sql import functions as F  
  
# 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\_people1.csv"  
df.to\_csv(csv\_file\_path, index=False)  
   
# Confirm the CSV file is created  
print(f"CSV file created at: {csv\_file\_path}")  
  
  
df\_etl = spark.read.format("csv").option("header","true").option("inferSchema","true").load(csv\_file\_path)  
  
# Show the Data frame  
df\_etl.show()  
  
# 2. \*\*Transform\*\*:  
# - \*\*Filter\*\*: Only include employees aged 30 and above in the analysis.  
  
df\_etl = df\_etl.filter(col("age") > 30)  
df\_etl.show()  
  
# - \*\*Add New Column\*\*: Calculate a 10% bonus on the current salary for each employee and add it as a new column (`salary\_with\_bonus`).  
  
df\_etl = df\_etl.withColumn("salary\_with\_bonus", col("salary") \* 1.10)  
df\_etl.show()  
  
# - \*\*Aggregation\*\*: Group the employees by gender and compute the average salary for each gender.  
  
df\_group\_etl = df\_etl.groupBy("gender").agg(F.avg("salary").alias("average\_salary"))  
df\_group\_etl.show()  
  
# 3. \*\*Load\*\*:  
# - Save the transformed data (including the bonus salary) in a Parquet file format for efficient storage and retrieval.  
# - Ensure the data can be easily accessed for future analysis or reporting.  
  
df\_etl.write.parquet("/content/people.parquet")  
  
df\_group\_etl.write.parquet("/content/gender\_average.parquet")