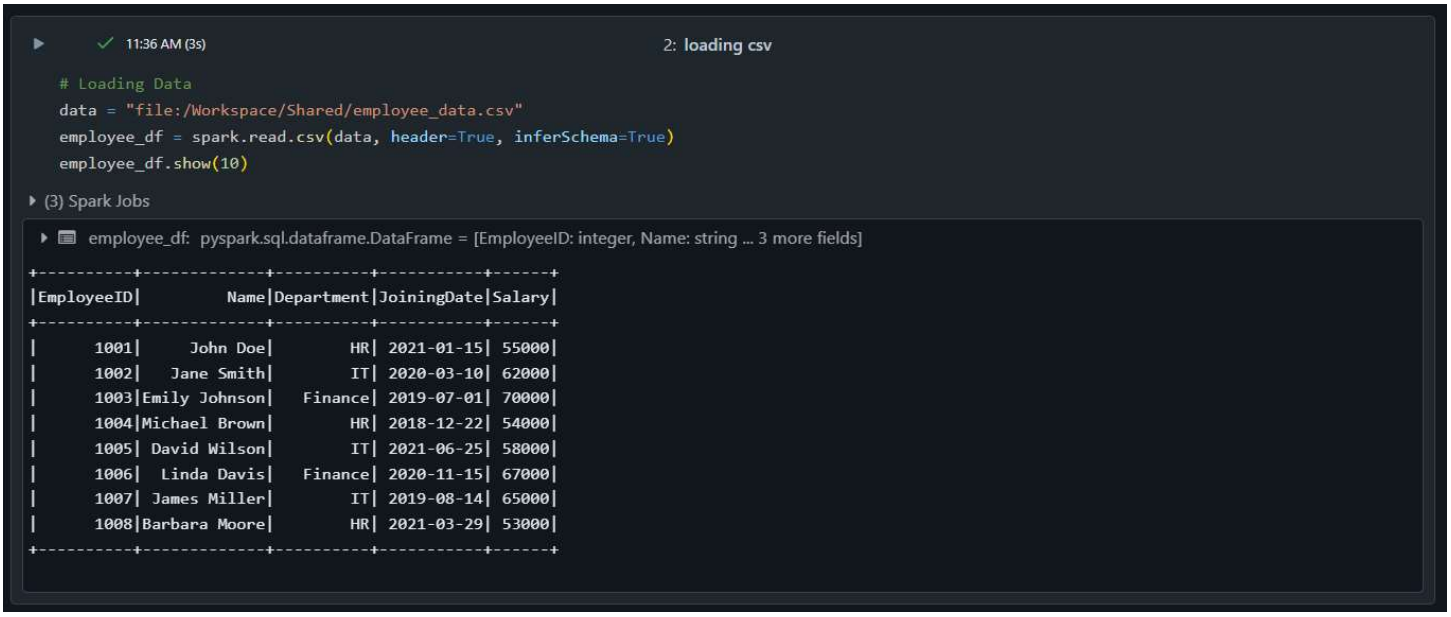
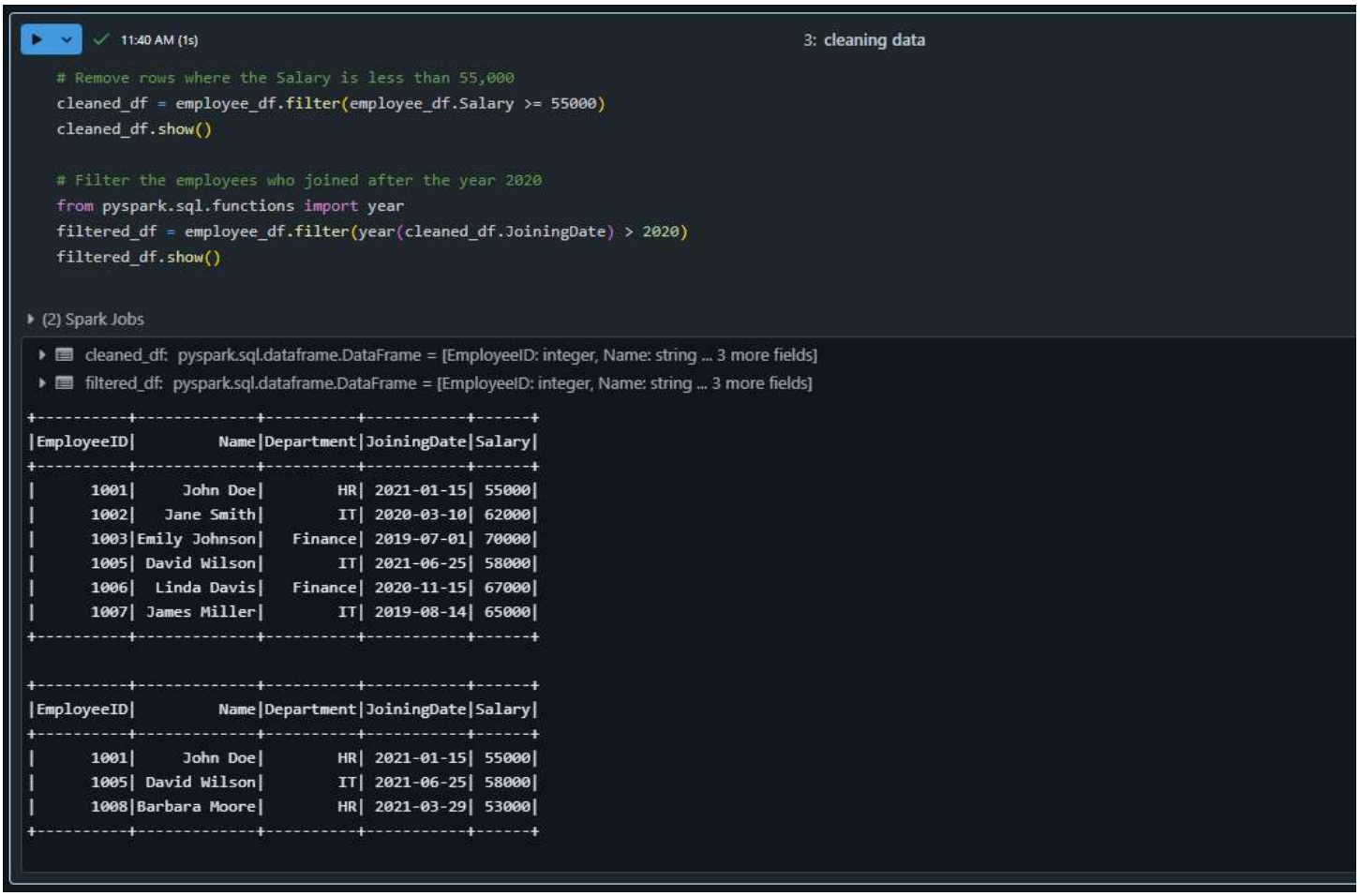
Azure DataBricks Assignment – 12th September

SaiPrabath Chowdary S

Assignment 1





11:40 AM (1s) 4: Data Aggregation

```
from pyspark.sql.functions import avg, count

# Find the average salary by Department
avg_salary_df = employee_df.groupBy("Department").agg(avg("Salary").alias("Average Salary"))
avg_salary_df.show()

# Count the number of employees in each Department
employee_count_df = employee_df.groupBy("Department").agg(count("EmployeeID").alias("Employee Count"))
employee_count_df.show()
```

▸ (4) Spark Jobs

▸ avg_salary_df: pyspark.sql.dataframe.DataFrame = [Department: string, Average Salary: double]

▸ employee_count_df: pyspark.sql.dataframe.DataFrame = [Department: string, Employee Count: long]

```
+----------+------------------+
|Department|    Average Salary|
+----------+------------------+
|        HR|           54000.0|
|   Finance|           68500.0|
|        IT|61666.666666666664|
+----------+------------------+

+----------+--------------+
|Department|Employee Count|
+----------+--------------+
|        HR|             3|
|   Finance|             2|
|        IT|             3|
+----------+--------------+
```

12:50 PM (3s) 5: loading data Python

```
# Save the cleaned data to a new CSV file
filtered_df.write.csv("file:/Workspace/Shared/filtered_employee_data.csv", header=True)
```

▸ (1) Spark Jobs