Assignment – Day 17

-Sarthak Niranjan Kulkarni (Maverick)

- [sarthakkul2311@gmail.com](mailto:sarthakkul2311@gmail.com) - (+91) 93256 02791

**28/11/2024 (Thursday)**

**Practice EDA Analysis: -**

1. **Reading and Displaying Data from the Loan Table in Databricks**

🡪 data = spark.read.table("hive\_metastore.default.loan")

display(data)

**A screenshot of a computer

Description automatically generated**

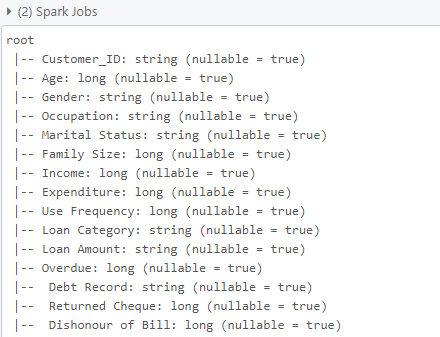
1. **Getting Row Count and Schema Information of the Data**

🡪 # Total row count

data.count()

# Schema information

data.printSchema()



1. **Displaying Summary Statistics for 'Income' Column**

🡪 # Summary statistics for 'Income'

data.describe(['Income']).show()

A screenshot of a computer program

Description automatically generated

1. **Counting Rows Grouped by Gender**

🡪 data.groupBy('gender').count().show()

**A screenshot of a computer

Description automatically generated**

1. **Displaying Top 5 Highest Incomes**

🡪 # Top 5 Highest Incomes

data.orderBy(data.Income.desc()).limit(5).show()

A screenshot of a computer program

Description automatically generated

1. **Grouping Employees by Salary Buckets and Counting**

🡪 # Salary Distribution

from pyspark.sql.functions import ceil, col

# Add salary buckets

data\_with\_buckets = data.withColumn('salary\_bucket', ceil(col('Income') / 20000) \* 20000)

# Count employees in each bucket

data\_with\_buckets.groupBy('salary\_bucket').count().orderBy('salary\_bucket').show()

A screenshot of a computer

Description automatically generated

**Summary of EDA Analysis: -**

I worked on a dataset from the hive\_metastore.default.loan table using PySpark in Databricks. First, I loaded the data into a Spark DataFrame and displayed it to get a view of the records. I calculated the total number of rows in the dataset with the count() function, which shows how many entries there are. Then, I examined the schema of the data to understand the structure of the table, such as the column names and data types.

I also performed summary statistics for the Income column, which gave me basic measures like the count, mean, and standard deviation. I grouped the data by gender to count how many records fall into each gender category. To further explore the data, I identified the top 5 highest incomes by sorting the data in descending order based on the Income column. Finally, I created salary buckets by dividing the Income into ranges and counted how many employees fall into each bucket, helping me understand the distribution of income within the dataset.