Spark ML Assignment

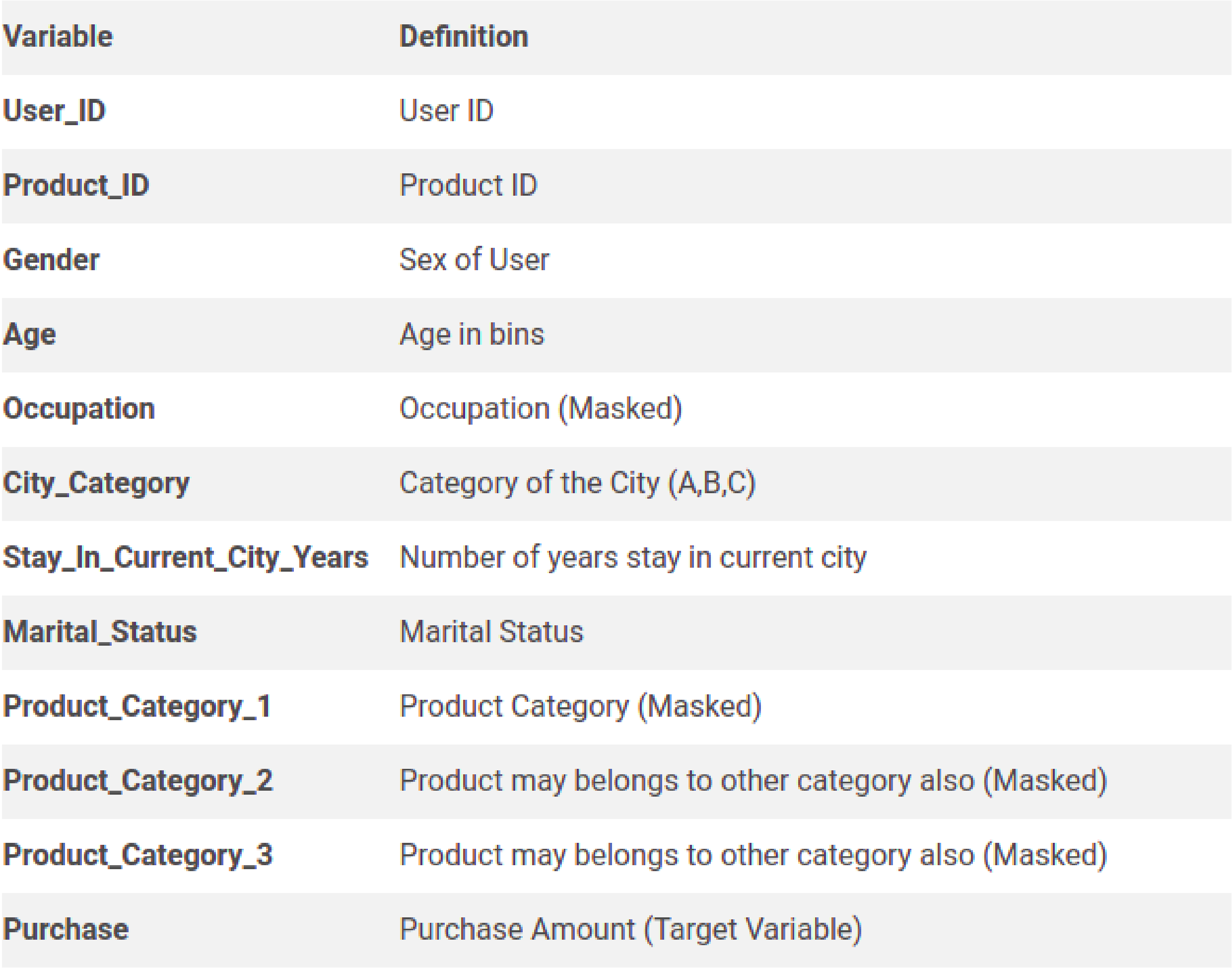
Problem statement:

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month.

The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Data:



Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Questions:

1. Average Purchase amount?
2. Counting and Removing null values
3. How many distinct values per column?
4. Count category values within each of the following column:
    - Gender
    - Age
    - City_Category
    - Stay_In_Current_City_Years

   - Marital_Status

5. Calculate average Purchase for each of the following columns:
   - Gender
   - Age
   - City_Category
   - Stay_In_Current_City_Years
   - Marital_Status
6. Label encode the following columns:
   - Age
   - Gender
   - Stay_In_Current_City_Years
   - City_Category
7. One-Hot encode following columns:
   - Gender
   - City_Category
   - Occupation
8. Build a baseline model using any of the ML algorithms.
9. Model improvement with Grid-Search CV
10. Create a Spark ML Pipeline for the final model.