SparkML Activity:

Problem Statement

The dataset is from a bank, data related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, to access if the product (bank term deposit) would be (or not) subscribed. The data and attribute description are in the folder.

Dataset Description

The dataset has the following attributes:
1 - unique sequence id
2 - age (numeric)
3 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
4 - marital_status : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
5 - education (categorical: "unknown","secondary","primary","tertiary")
6 - default: has credit in default? (binary: "yes","no")
7 - balance: average yearly balance, in euros (numeric)
8 - housing: has housing loan? (binary: "yes","no")
9 - loan: has personal loan? (binary: "yes","no")
10 - contact: contact communication type (categorical: "unknown","telephone","cellular")
11 - day: last contact day of the month (numeric)
12 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
13 - duration: last contact duration, in seconds (numeric)
14 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
15 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
16 - previous: number of contacts performed before this campaign and for this client (numeric)
17 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")
18 - opened_new_td_act_no_yes - has the client subscribed to a term deposit? (binary: "yes","no")

The idea is we have set of information about a person whether its their age the job they have the marital status whether they have loan with the bank or not etc. we're trying to decide here is whether this person open a new deposit account with this bank.

So the theory here is that if I can better target the right customers with the right offer at right time they'll convert over to actually open a deposit account with the bank so that I don't have to reach out whole population and waste a lot of money with ads or mailer or what ever have you but target only the perspective customers.

Questions:

1. **Define the schema for the dataset and use the schema to read the file bank_data.csv**
2. **Using the above schema read the data and the data frame.**
3. **Verify the schema**
4. **Check the datatypes**
5. **Cache the dataframe**
6. **Verify the first few records**
7. **Verify the total number of rows and columns**
8. **Verify the summary statistics**
9. **Find the maximum and minimum values in each column**
10. **Find if there are any negative balances in the columns**
11. **Replace the negative balances with zero.**
12. **Define a table/view on the spark dataframe created to run sql queries on the dataframe.**
13. **Verify the target distribution.**
14. **Find the pairwise frequency between target and loan columns**
15. **Find the term deposit opted for different job categories. (Plot a visualization for the same).**
16. **Find the term deposit not-opted for different job categories. (Plot a visualization for the same).**
17. **Find the term deposit not-opted for different education categories. (Plot a visualization for the same).**
18. **Find the term deposit not-opted for different marital status category. (Plot a visualization for the same).**
19. **Plot a heat map for the dataframe**
20. **Remove null values in the dataframe**
21. **Split the data into training and testing to make it ready for Spark ML pipeline.**