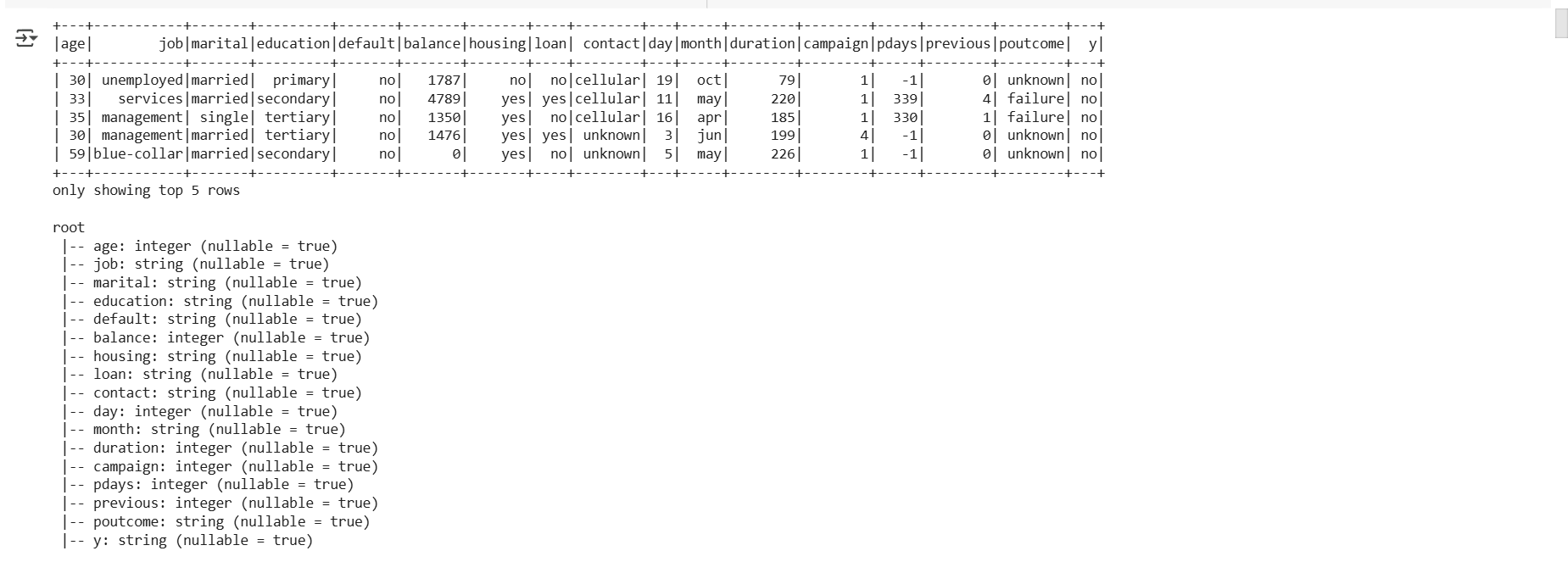
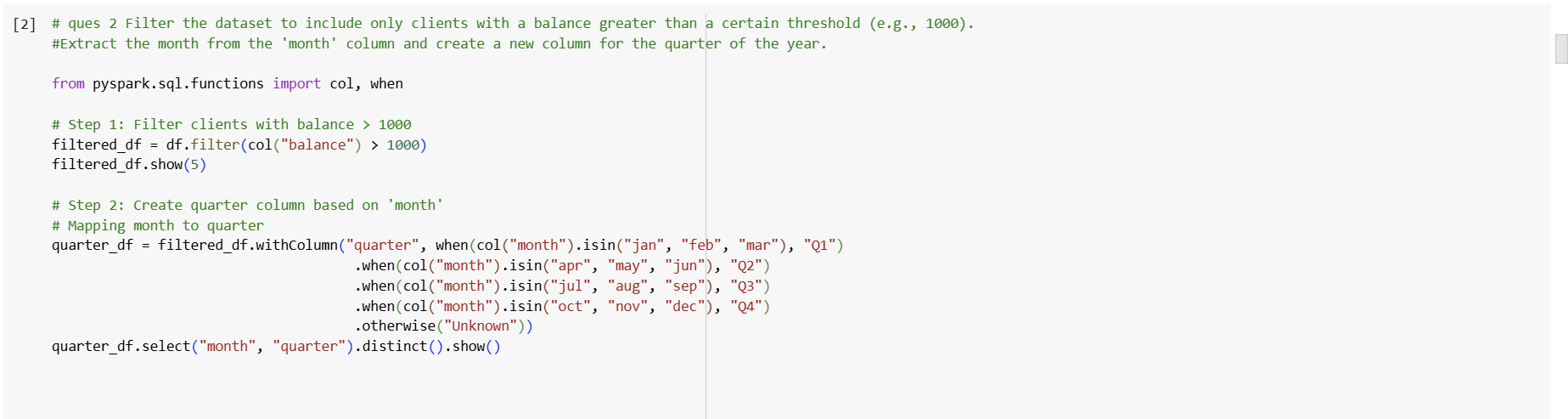
**SPARK Data Processing**

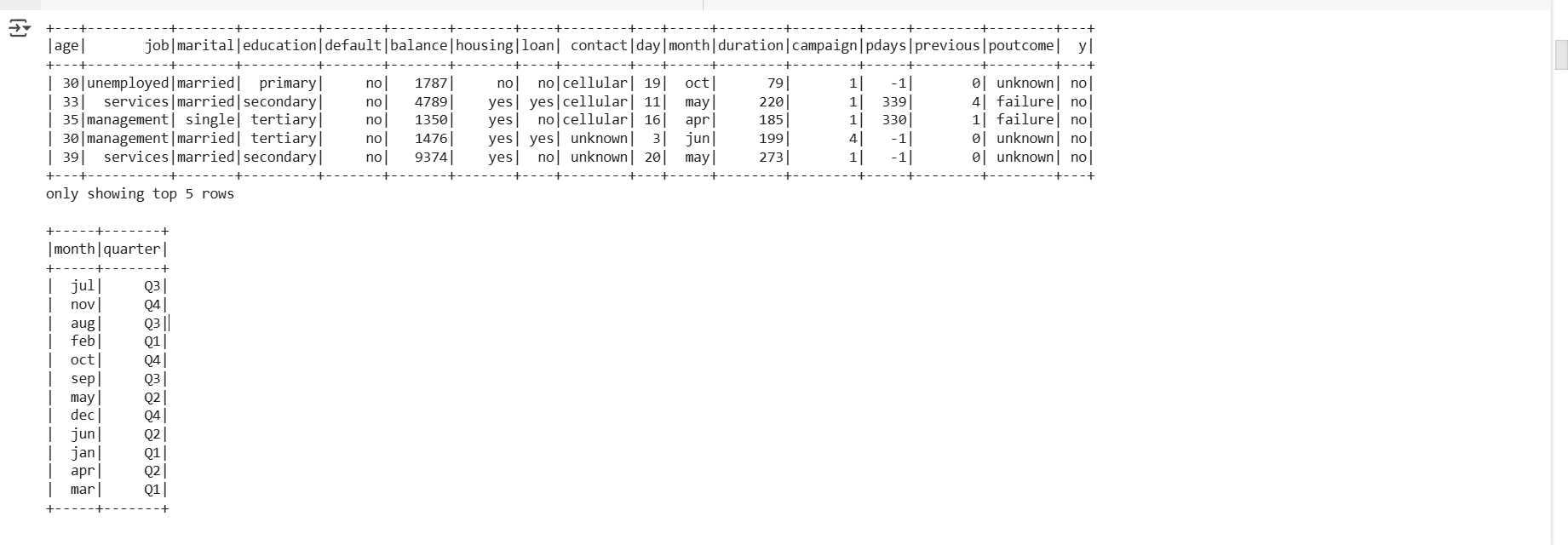
1. **Data Loading and Basic Inspection**:
   * Load the **bank.csv** dataset into a Spark DataFrame.
   * Perform basic data inspection tasks such as showing the first few rows, printing the schema, and displaying a summary of numerical columns.

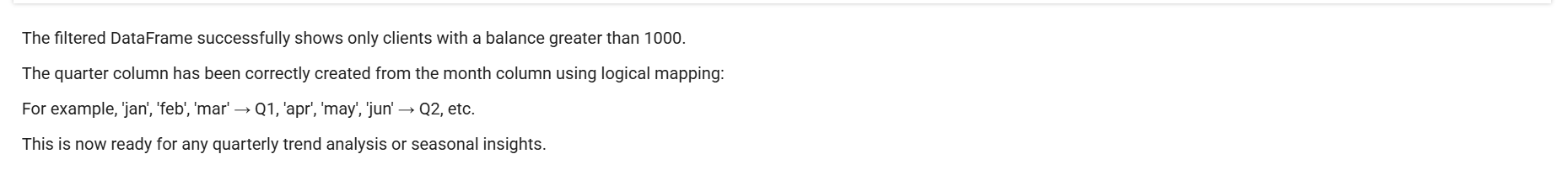




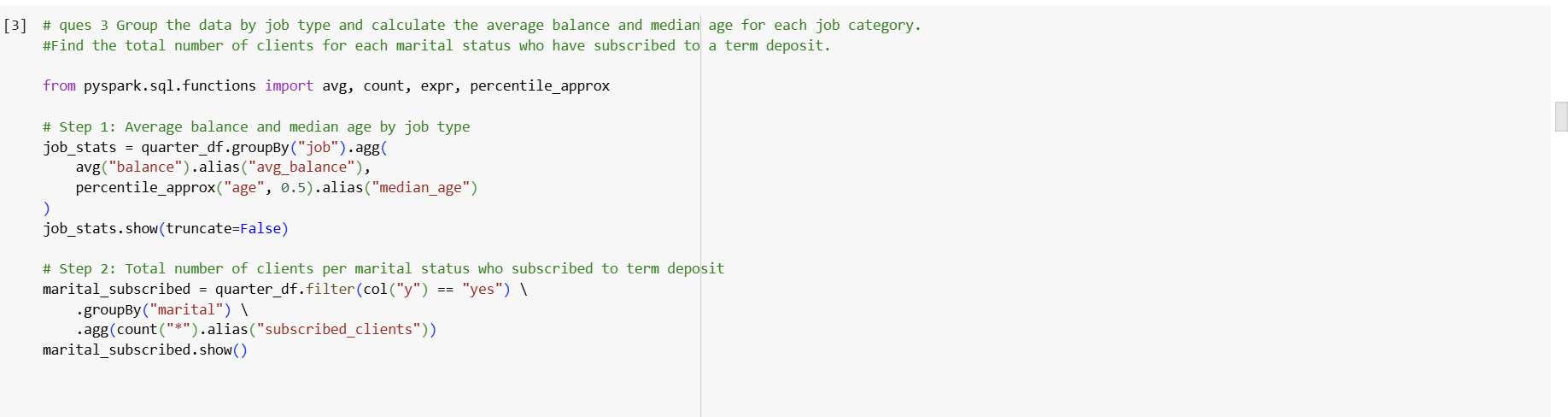
1. **Data Filtering and Column Operations**:
   * Filter the dataset to include only clients with a balance greater than a certain threshold (e.g., 1000).
   * Extract the month from the 'month' column and create a new column for the quarter of the year.

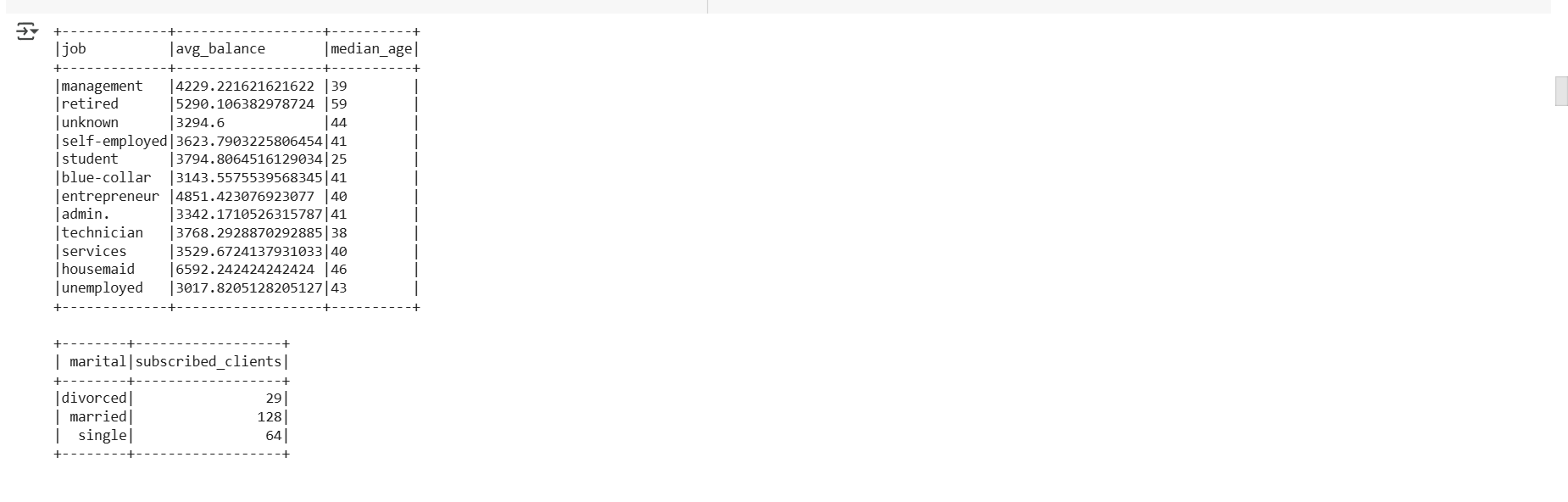






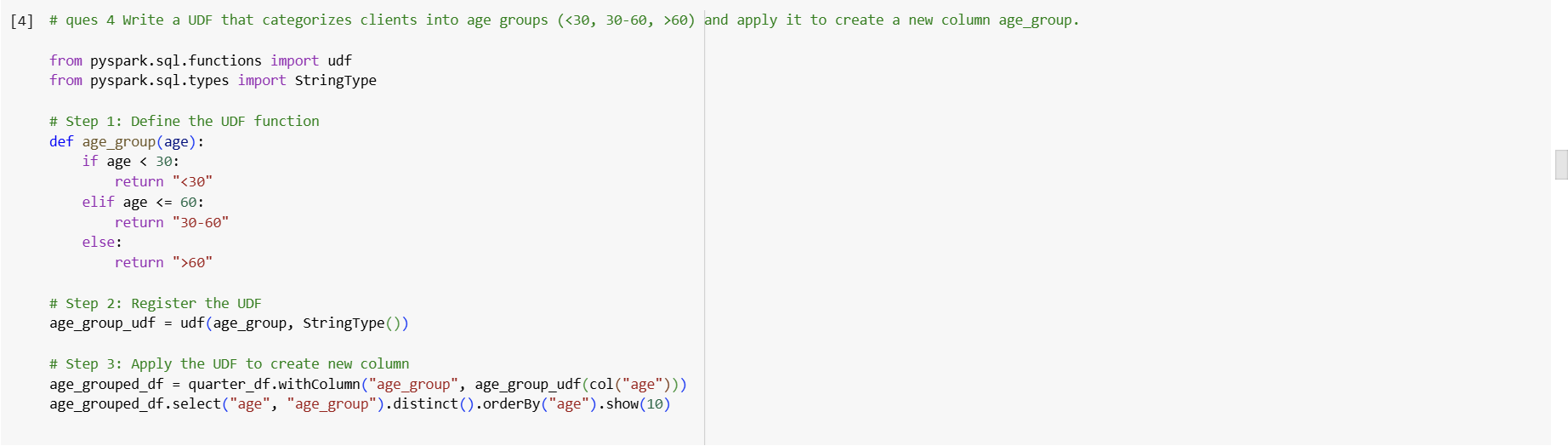
1. **GroupBy and Aggregation**:
   * Group the data by job type and calculate the average balance and median age for each job category.
   * Find the total number of clients for each marital status who have subscribed to a term deposit.

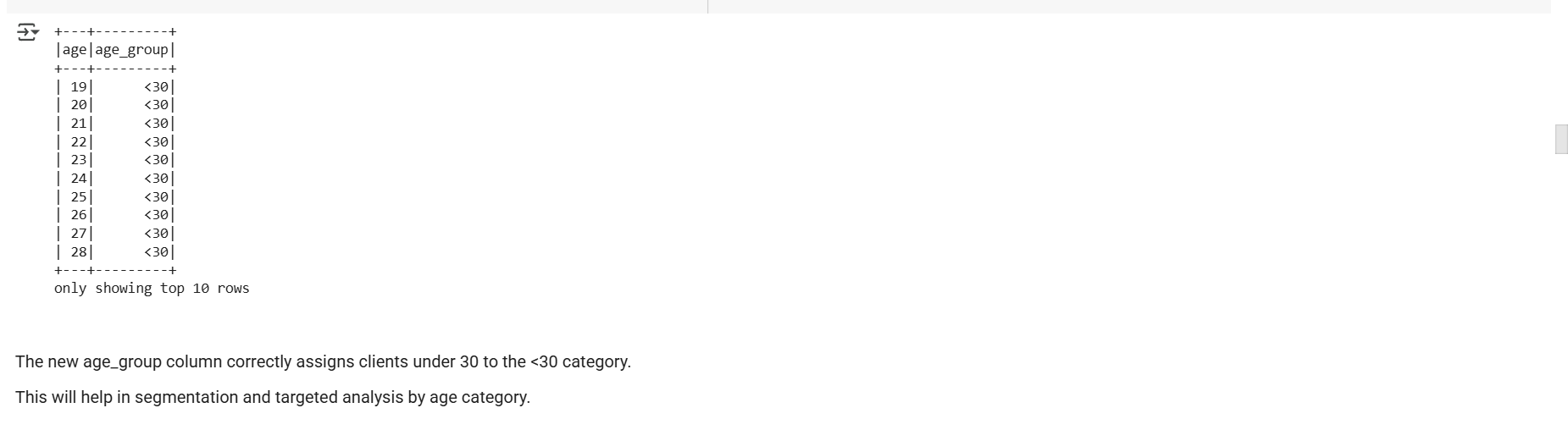






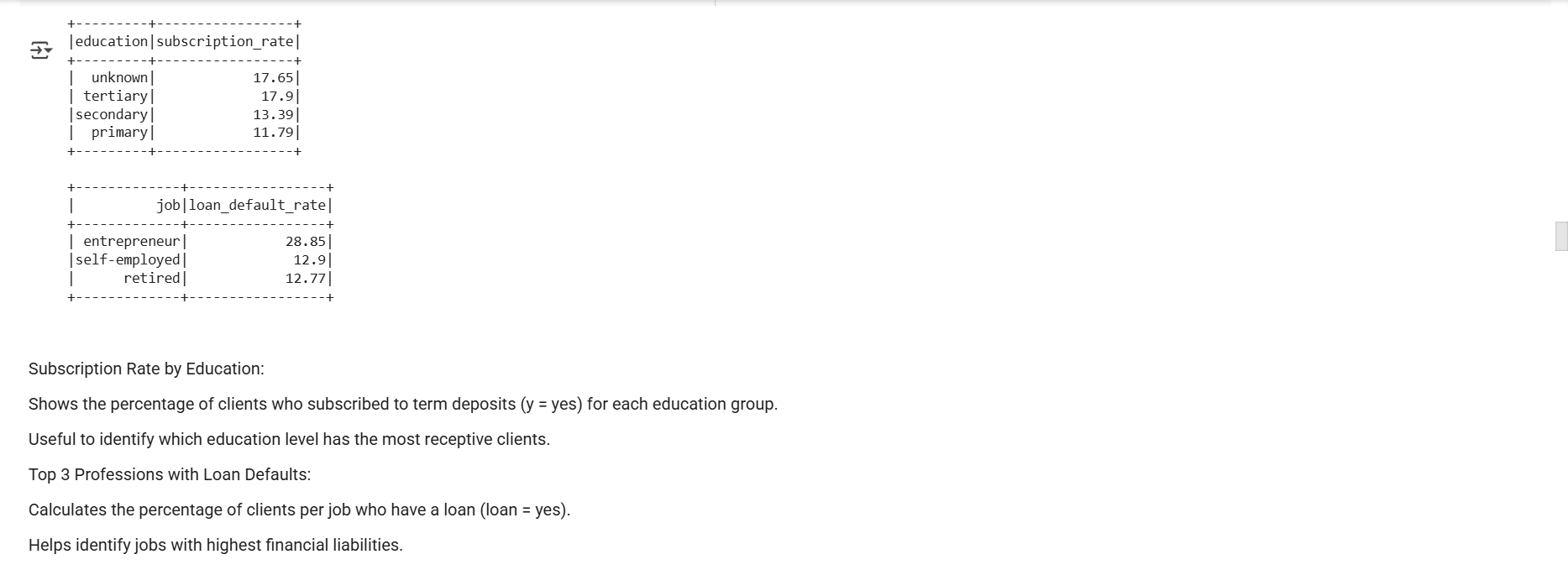
1. **Use UDF to Categorize Age Groups**:
   * Write a UDF that categorizes clients into age groups ('<30', '30-60', '>60') and apply it to create a new column **age\_group**.





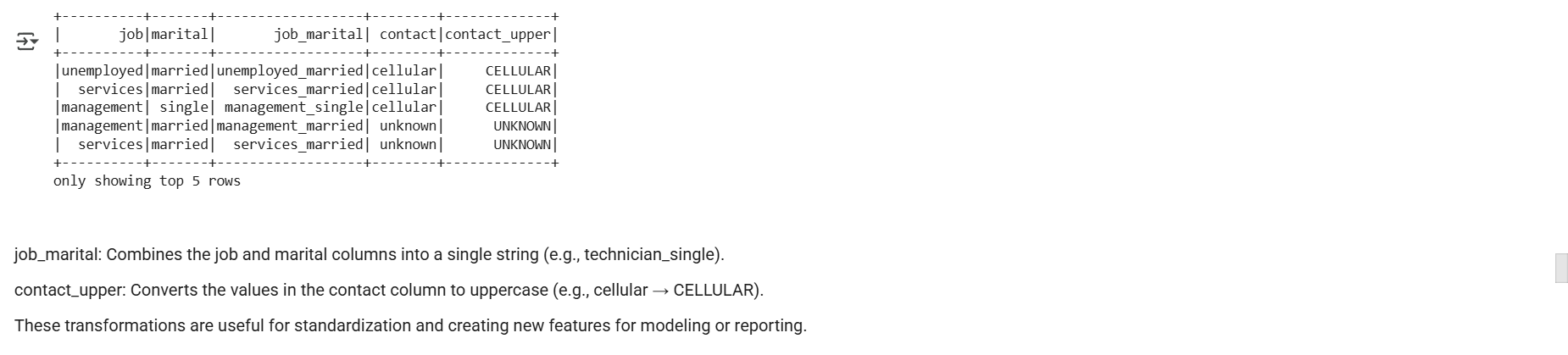
1. **Advanced Data Transformations**:
   * Calculate the subscription rate (percentage of clients who subscribed to a term deposit) for each education level.
   * Identify the top 3 professions that have the highest loan default rate.



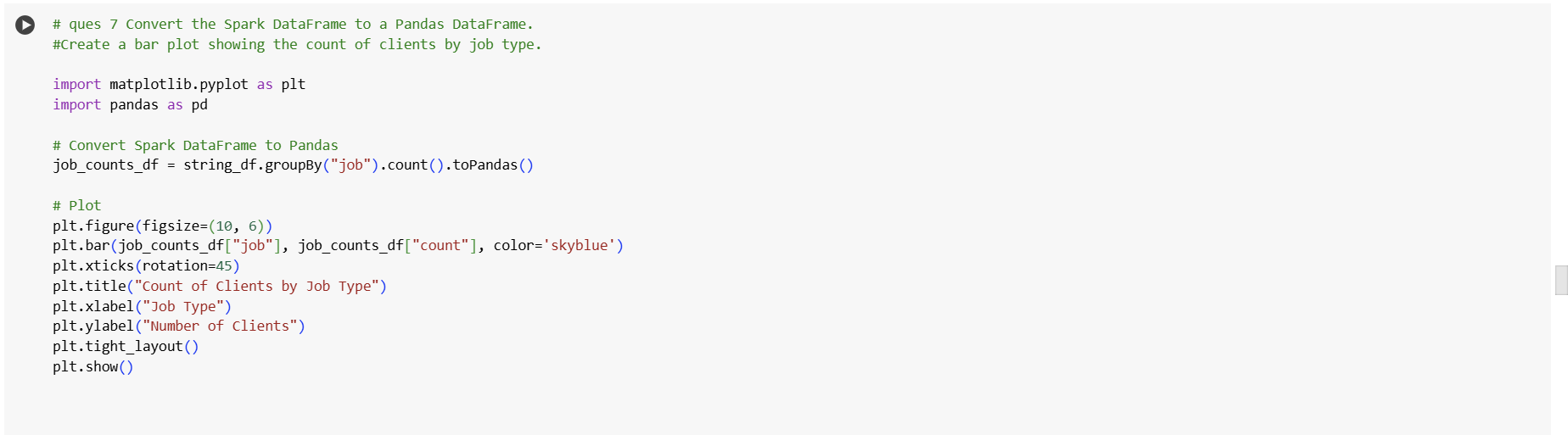


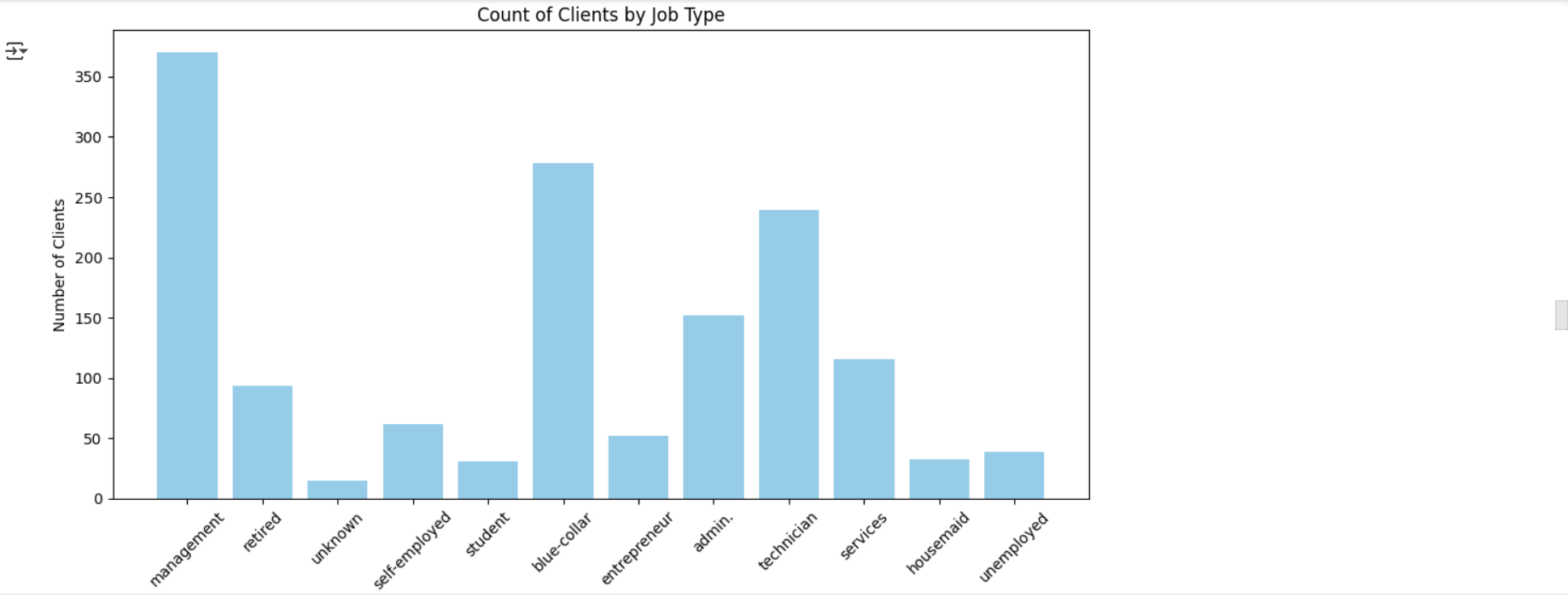
1. **String Manipulation and Date Functions**:
   * Concatenate the 'job' and 'marital' columns into a new column 'job\_marital'.
   * Use string functions to convert the 'contact' column to uppercase.





1. **Data Visualization** 
   * Convert the Spark DataFrame to a Pandas DataFrame.
   * Create a bar plot showing the count of clients by job type.





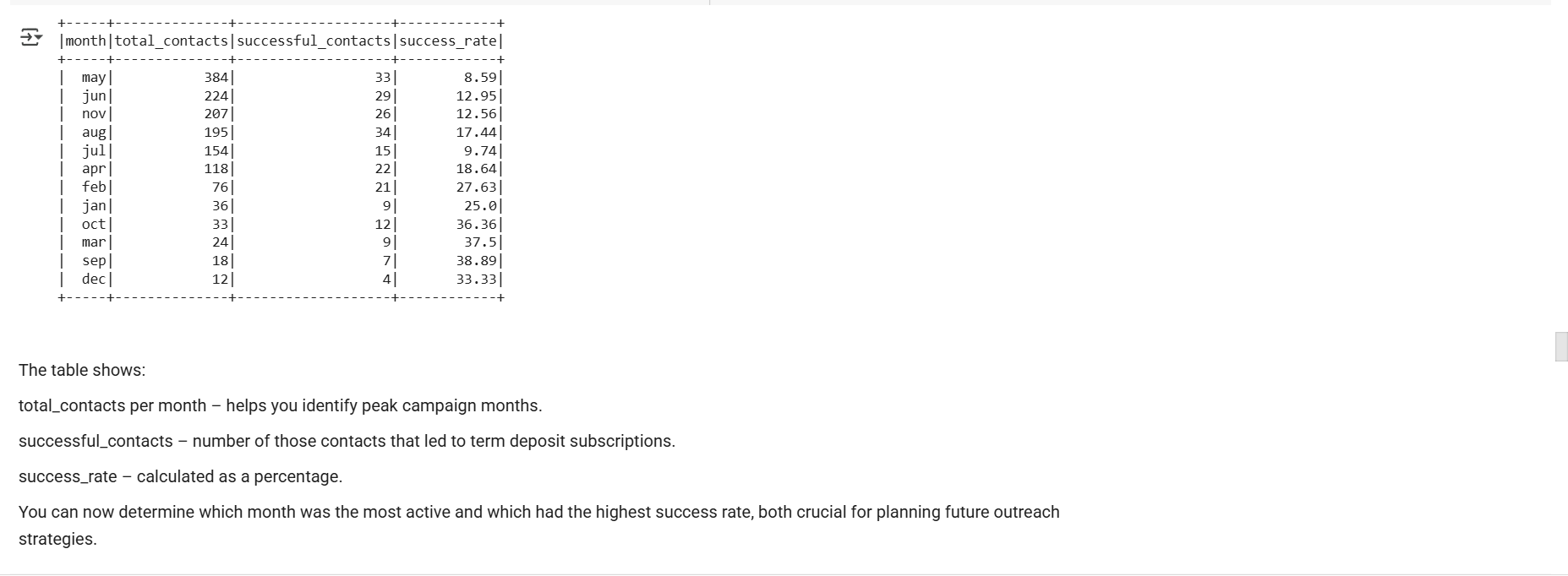
The bar plot displays how many clients belong to each job type.

You can now visually identify the most common professions in your dataset (e.g., blue-collar, admin., technician).

This helps prioritize high-volume segments for campaign focus.

1. **Complex Queries for Insights**:
   * Analyze which month of the year has the highest number of clients contacted, and determine the success rate of the campaign in that month.



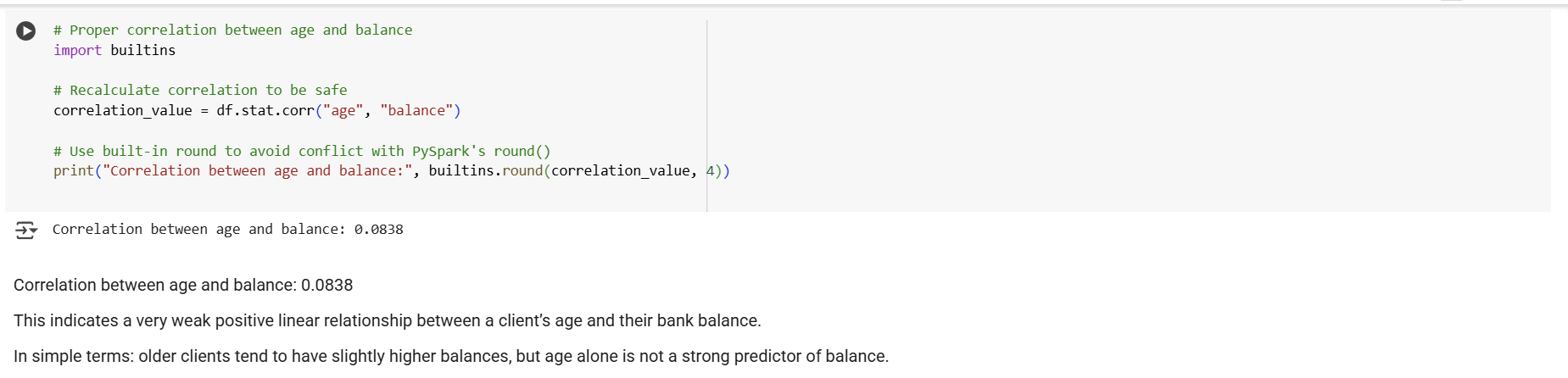


* + Find the average duration of contact for clients who subscribed versus those who did not.

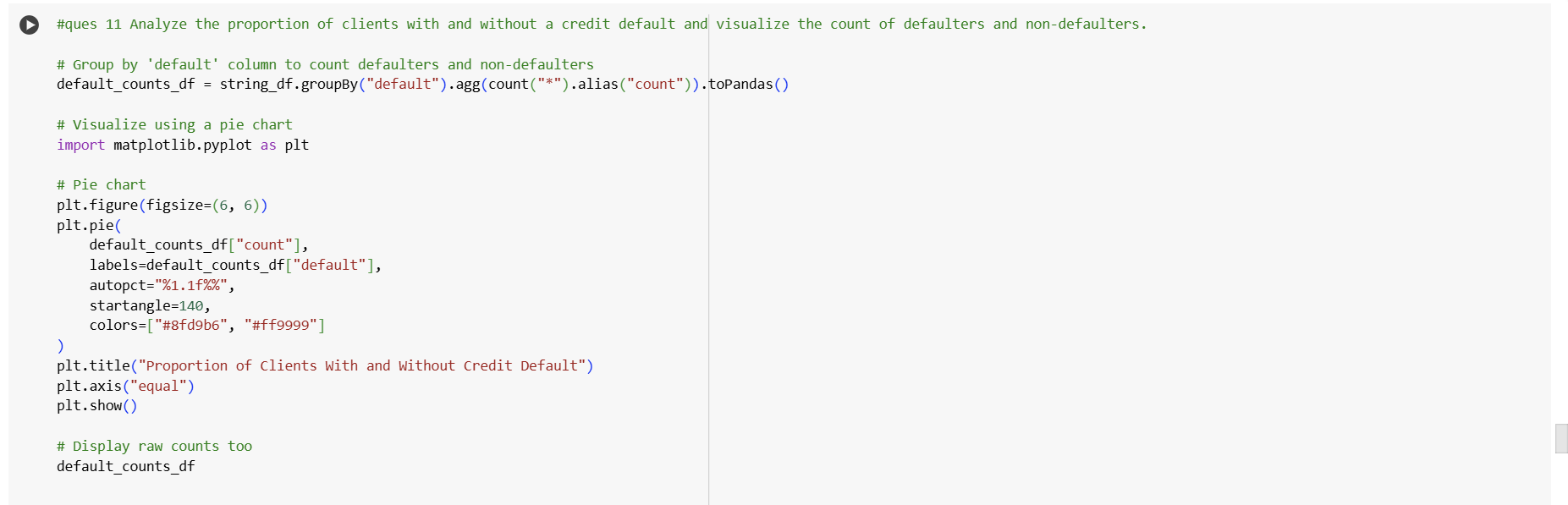


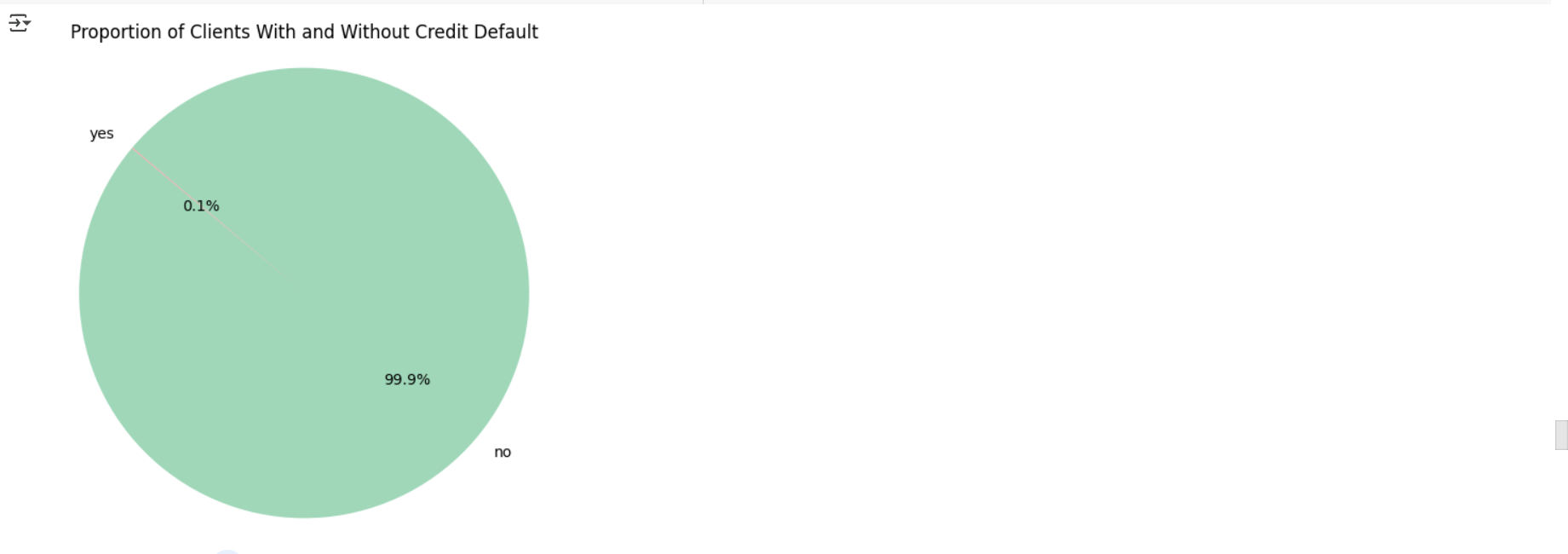
1. **Correlation Between Age and Balance**:
   * Calculate the correlation coefficient between 'age' and 'balance' to see if there’s any linear relationship between these two variables.

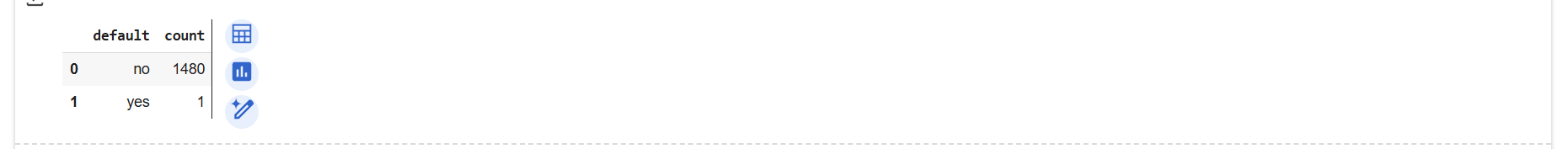




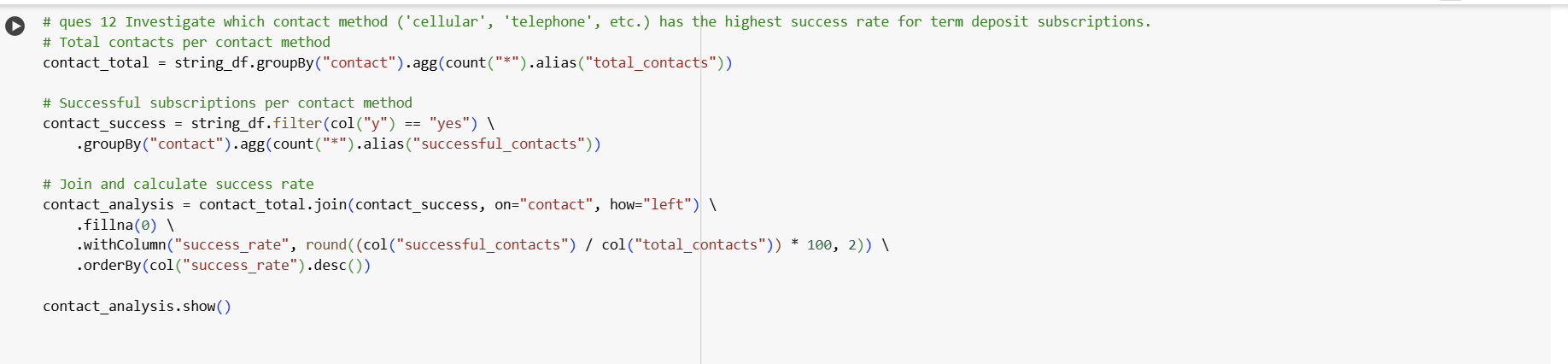
1. **Exploring Loan Defaults**:
   * Analyze the proportion of clients with and without a credit default and visualize the count of defaulters and non-defaulters.

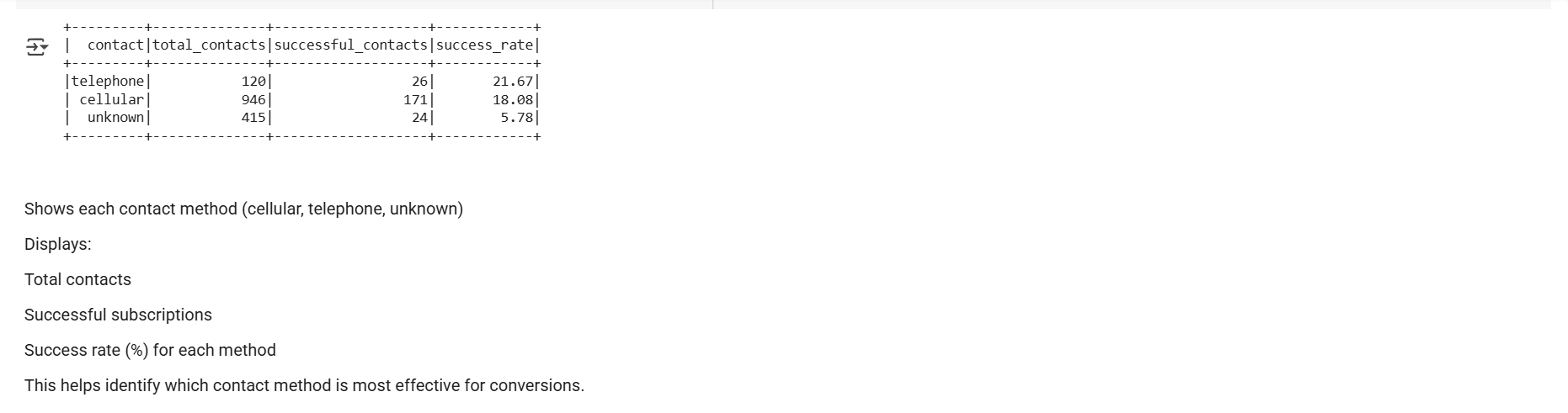






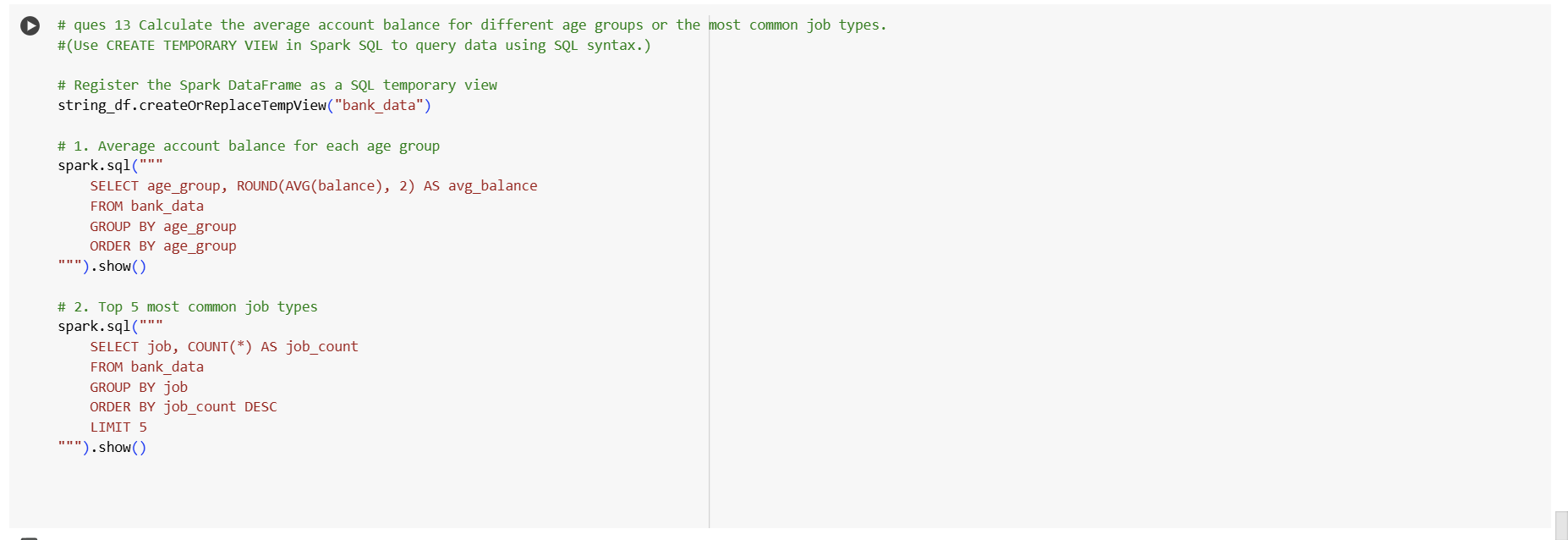
1. **Contact Method Analysis**:
   * Investigate which contact method ('cellular', 'telephone', etc.) has the highest success rate for term deposit subscriptions.

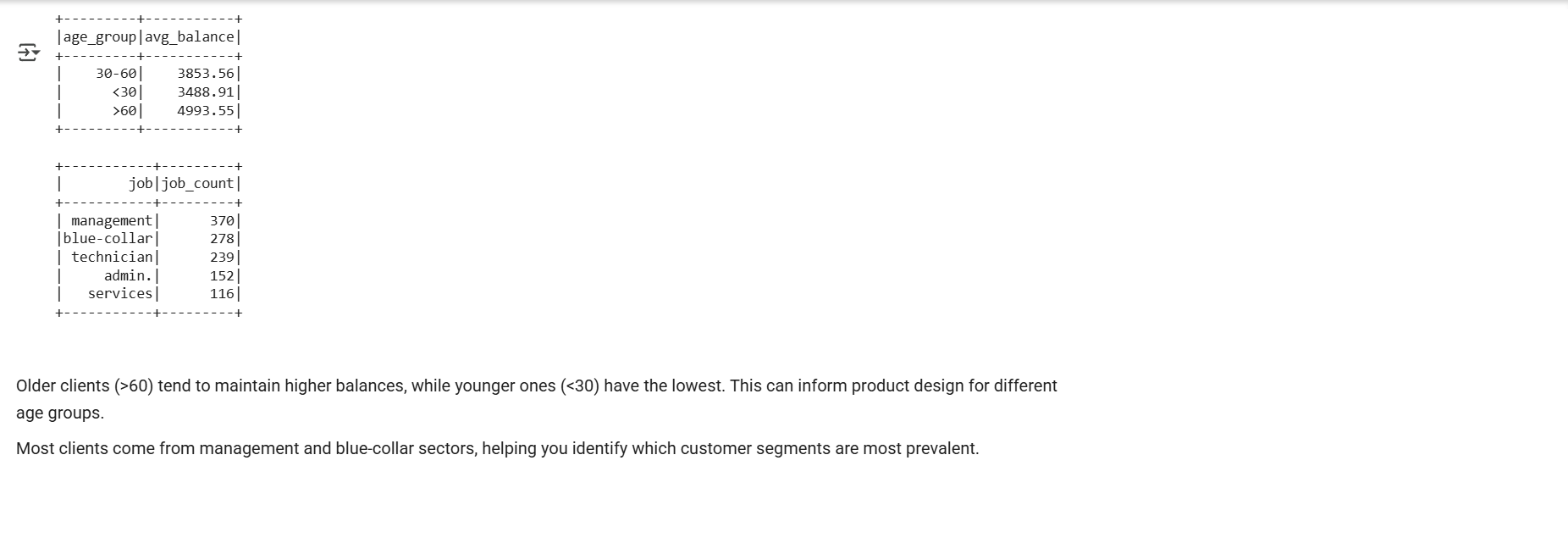




1. **Data Processing with Spark SQL**:

* Calculate the average account balance for different age groups or the most common job types. (Use **CREATE TEMPORARY VIEW** in Spark SQL to query data using SQL syntax.)





**Submission Guidelines:**

* You have the flexibility to perform the Spark data analysis either in Google Colab or in your local environment. Choose the option that is most convenient for you.
* If you choose to work in a local environment, please ensure you take screenshots of both the code and the output for each question. Make a copy of this doc file and include the screenshots in your copied document.
* If you opt for Google Colab, ensure to include the Colab file directly as part of your submission.
* Once your project is complete and your documentation is ready, please upload the document along with any other required files to the submission dashboard.

COLAB LINK - [Copy of Exploratory Data Analysis (EDA) with Spark.ipynb](https://colab.research.google.com/drive/19bTDr6yzzL8cmRDRfgiPqVlkVdaA1mnM?usp=sharing)