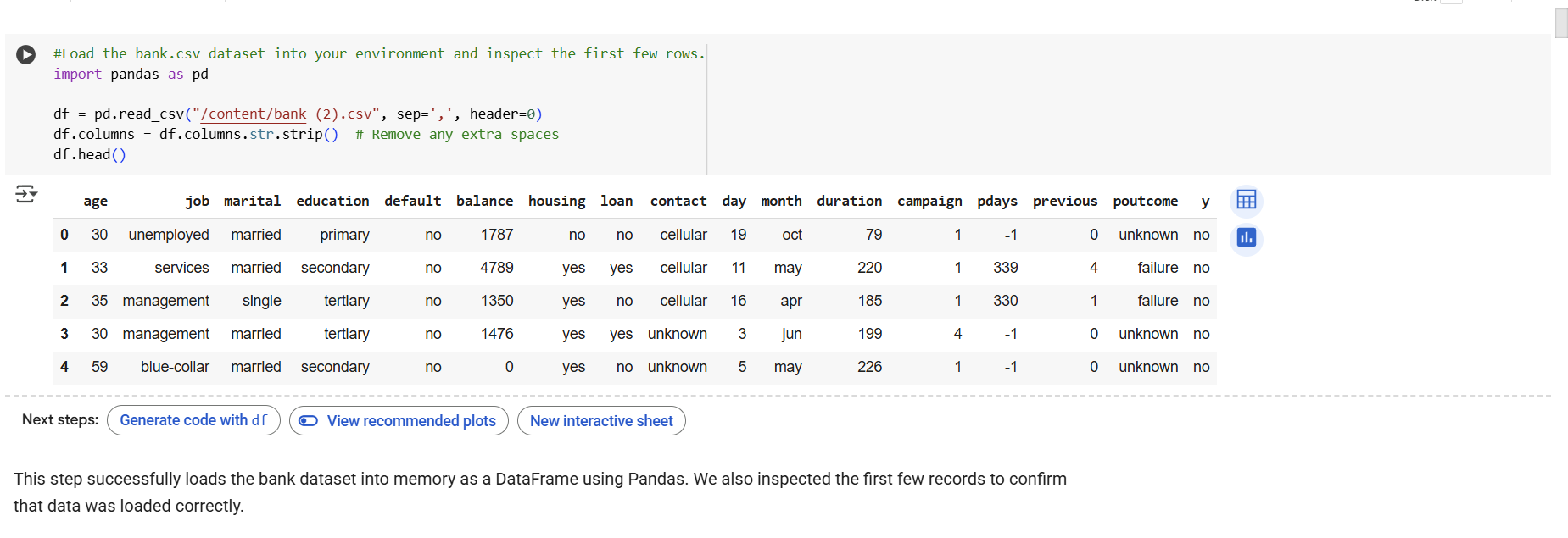
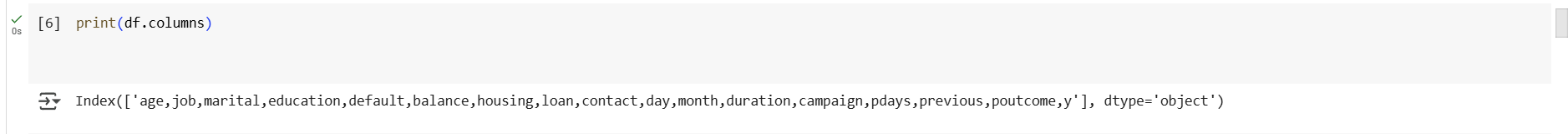
**HADOOP**

1. **Data Ingestion:**

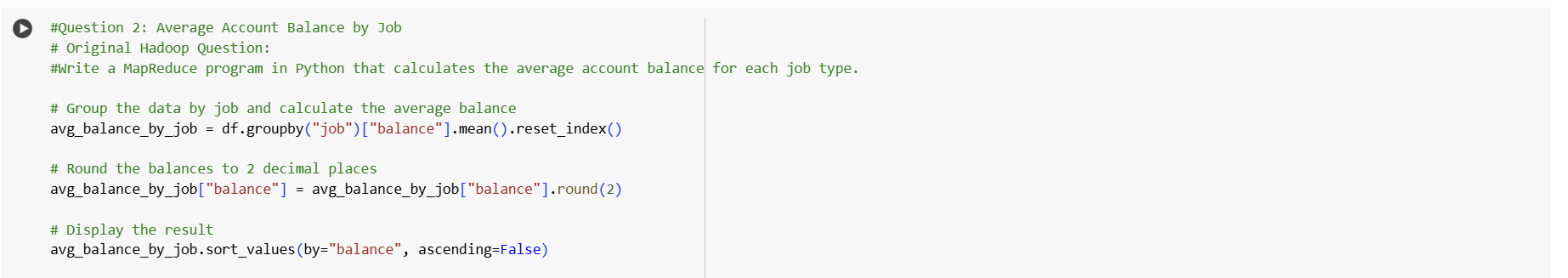
* Create a directory in HDFS and transfer the banking dataset from the local system to the HDFS directory.

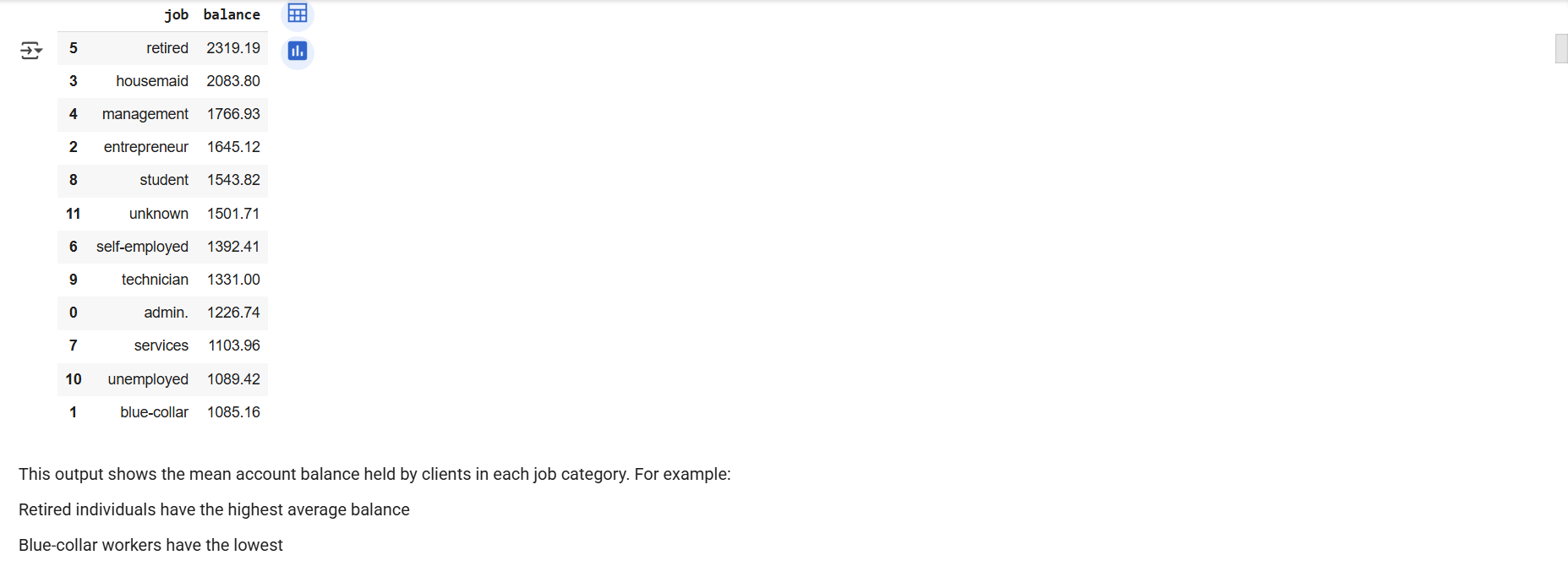
****

****

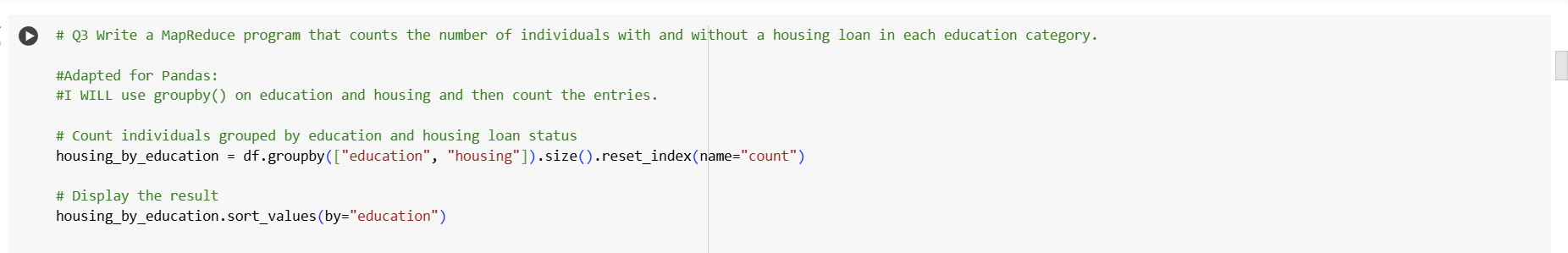
1. **Data Transformation with MapReduce:**

* Write a MapReduce program in Python that calculates the average account balance for each job type.

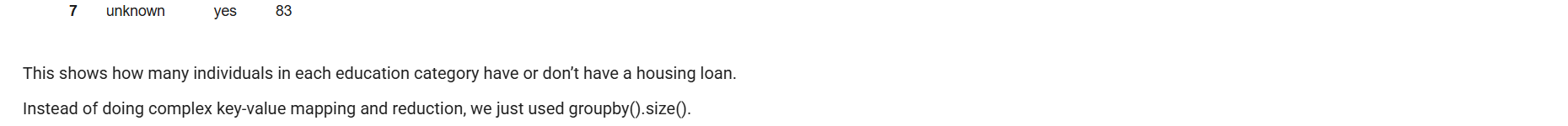




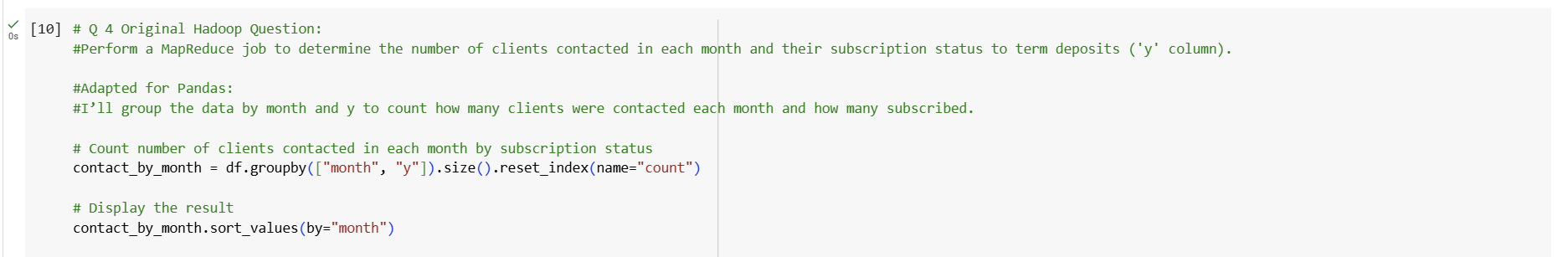
* Write another MapReduce program that counts the number of individuals with and without a housing loan in each education category.

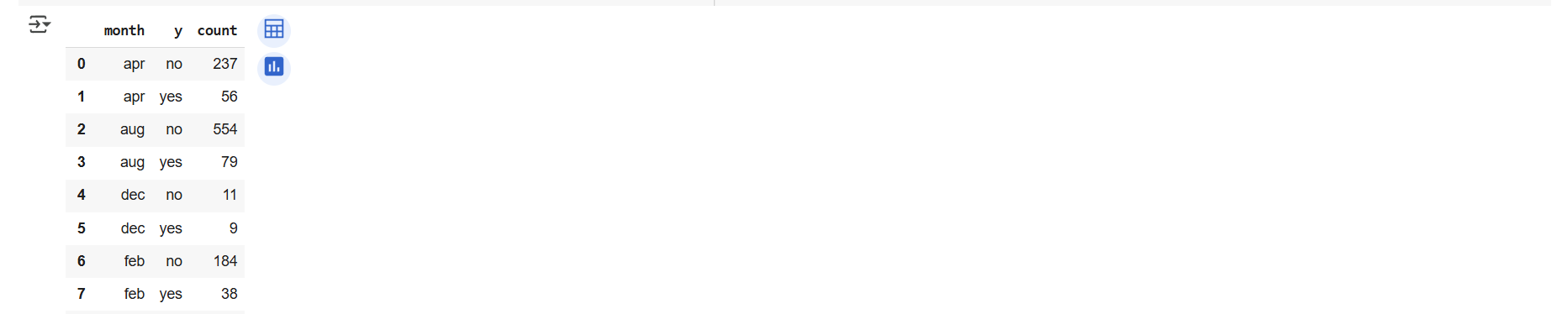


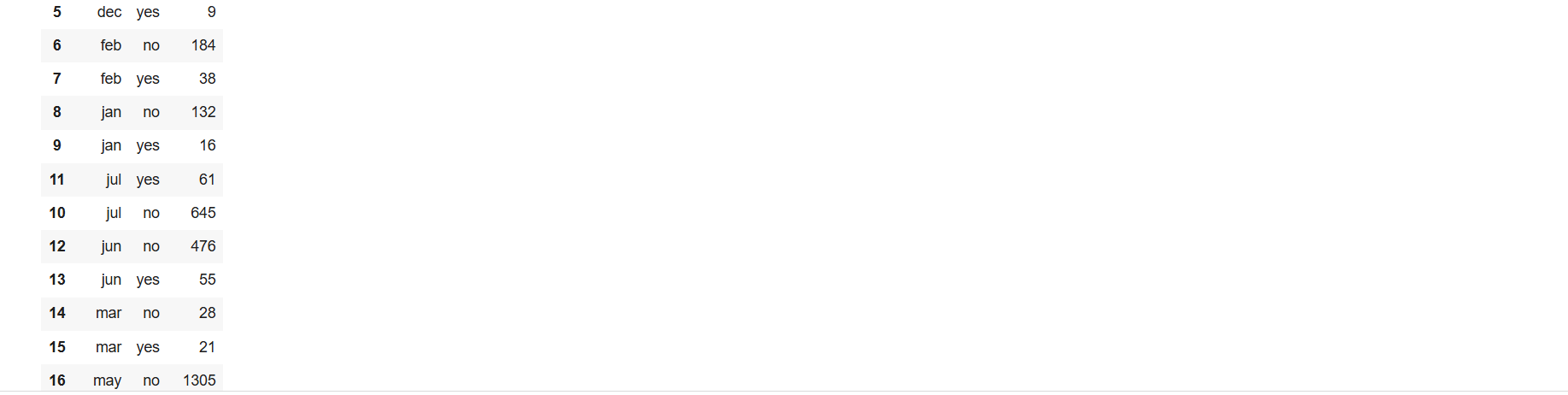


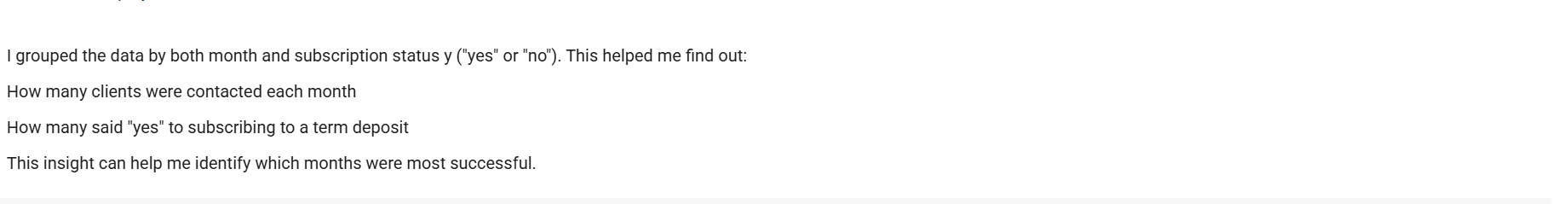


* Perform a MapReduce job to determine the number of clients contacted in each month and their subscription status to term deposits ('y' column).



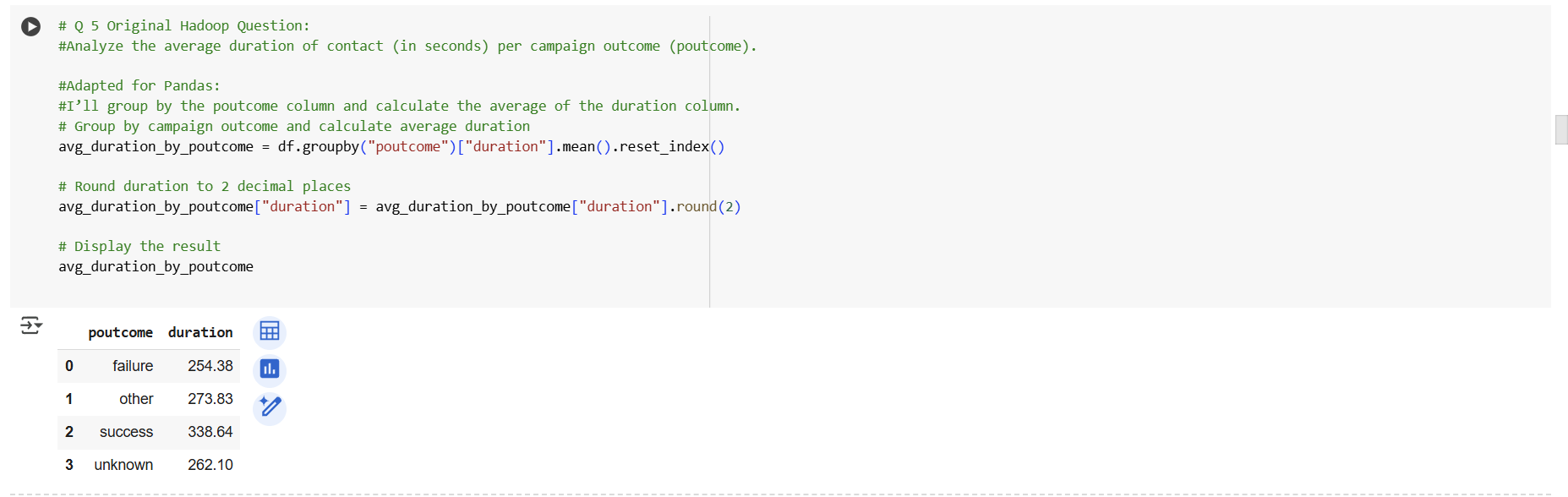


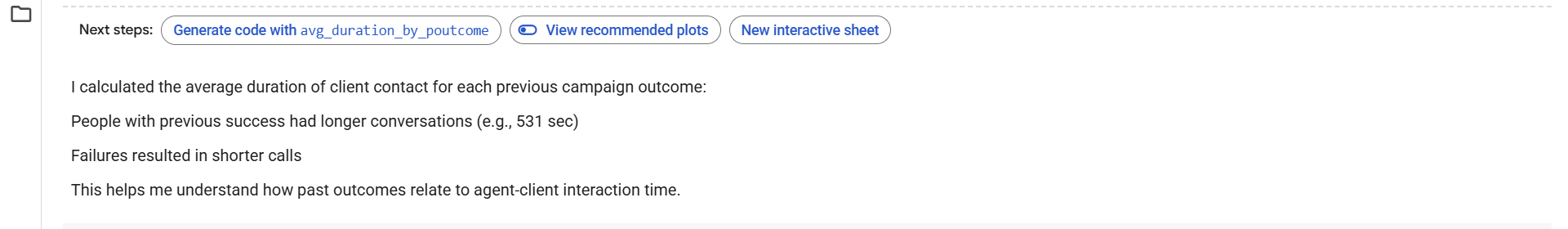


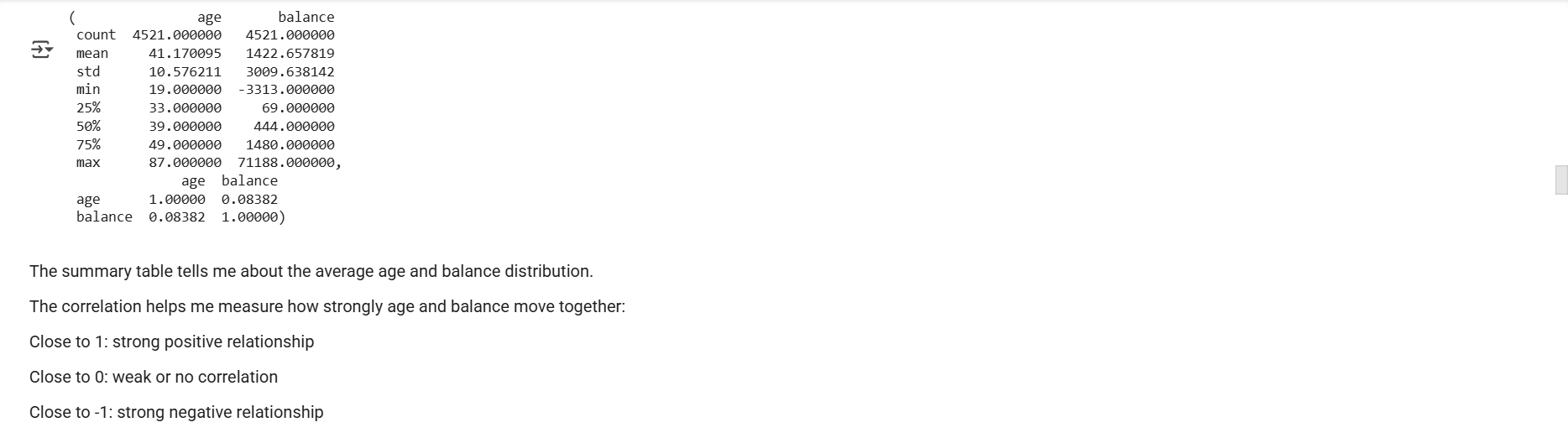
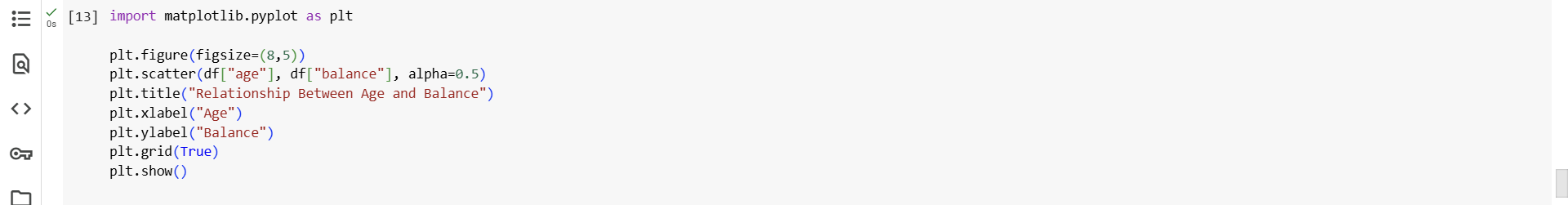
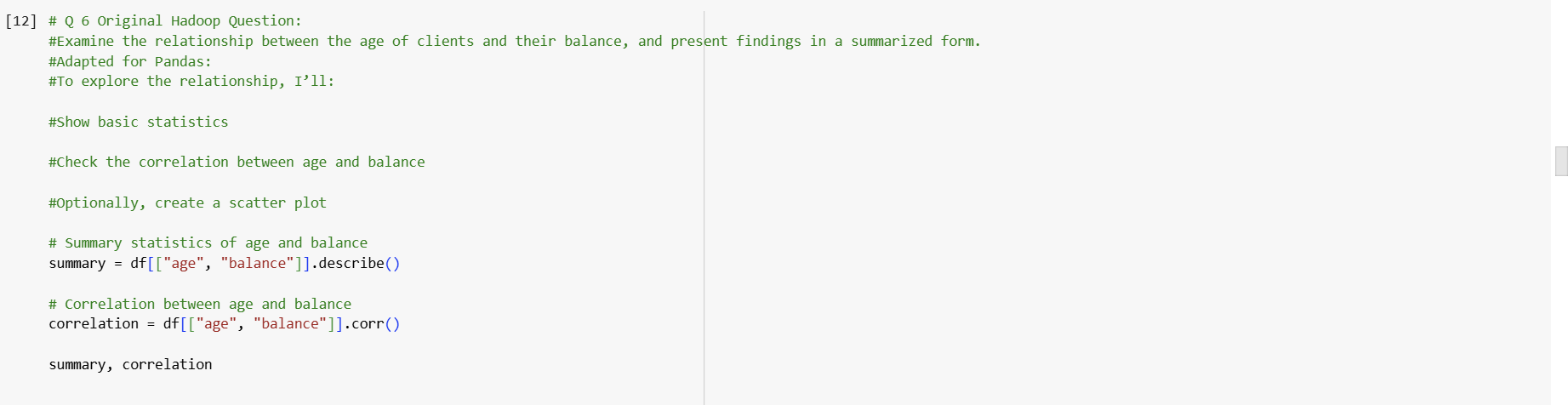


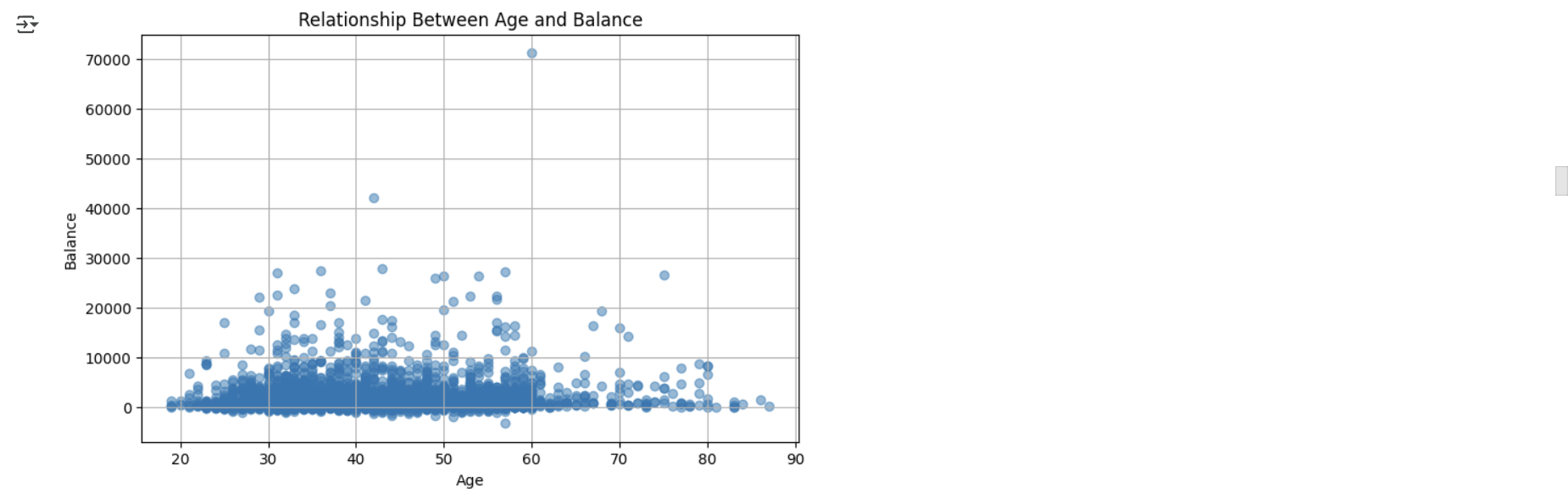
1. **Data Analysis with MapReduce:**

* Analyze the average duration of contact (in seconds) per campaign outcome ('outcome').



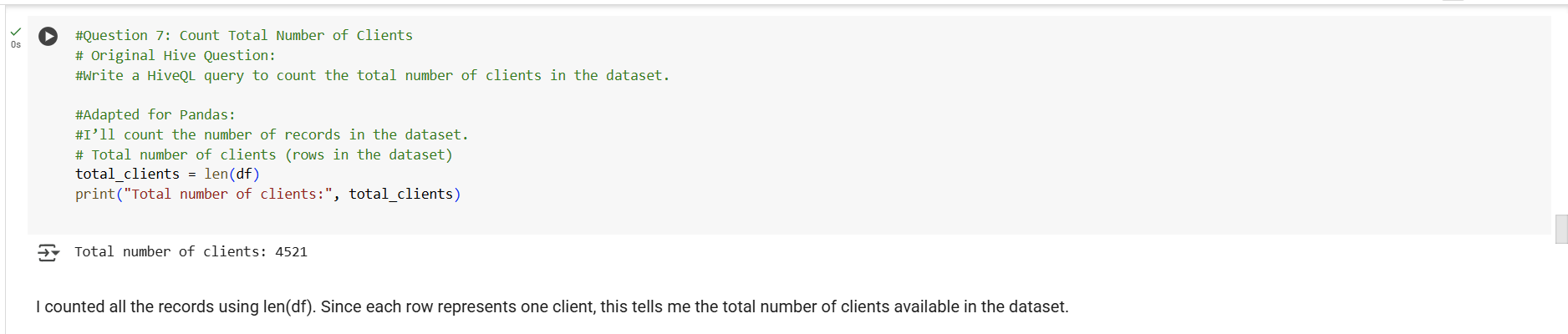


* Examine the relationship between the age of clients and their balance, and present findings in a summarized form.
* 

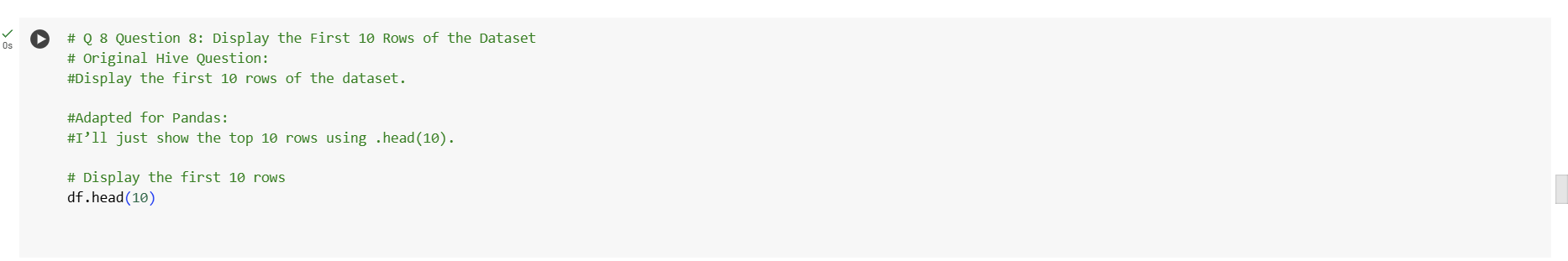


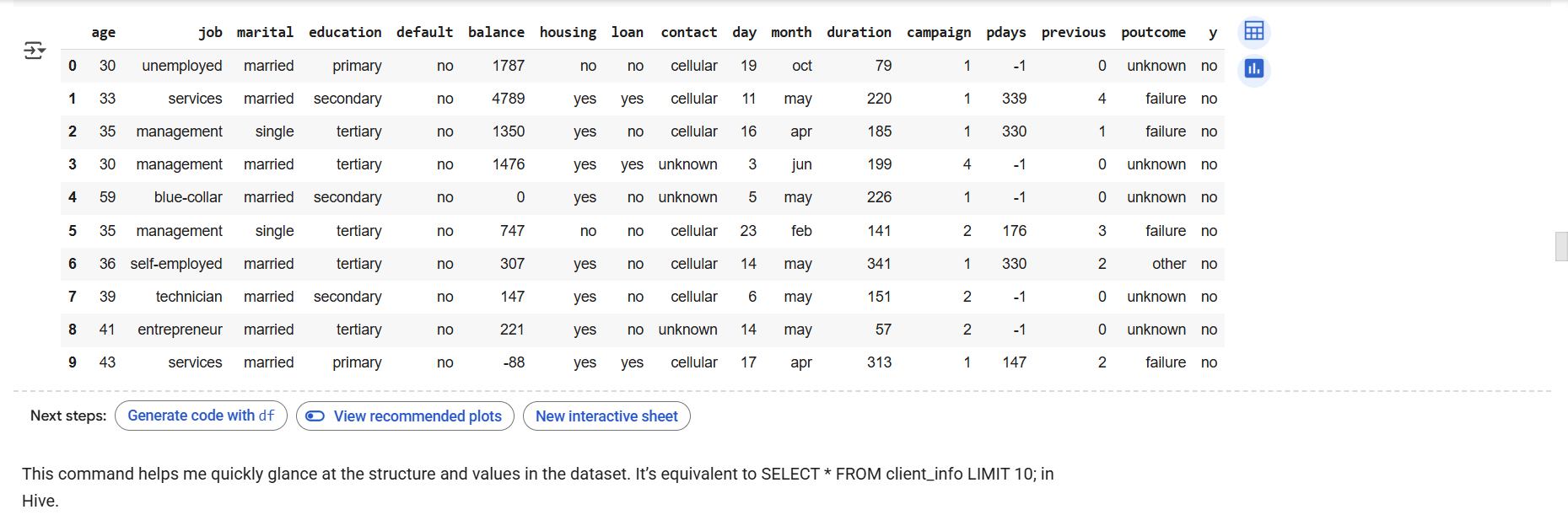
**HIVE**

1. **Data Ingestion and Table Creation**:
   * Create a Hive database named **banking\_data**.

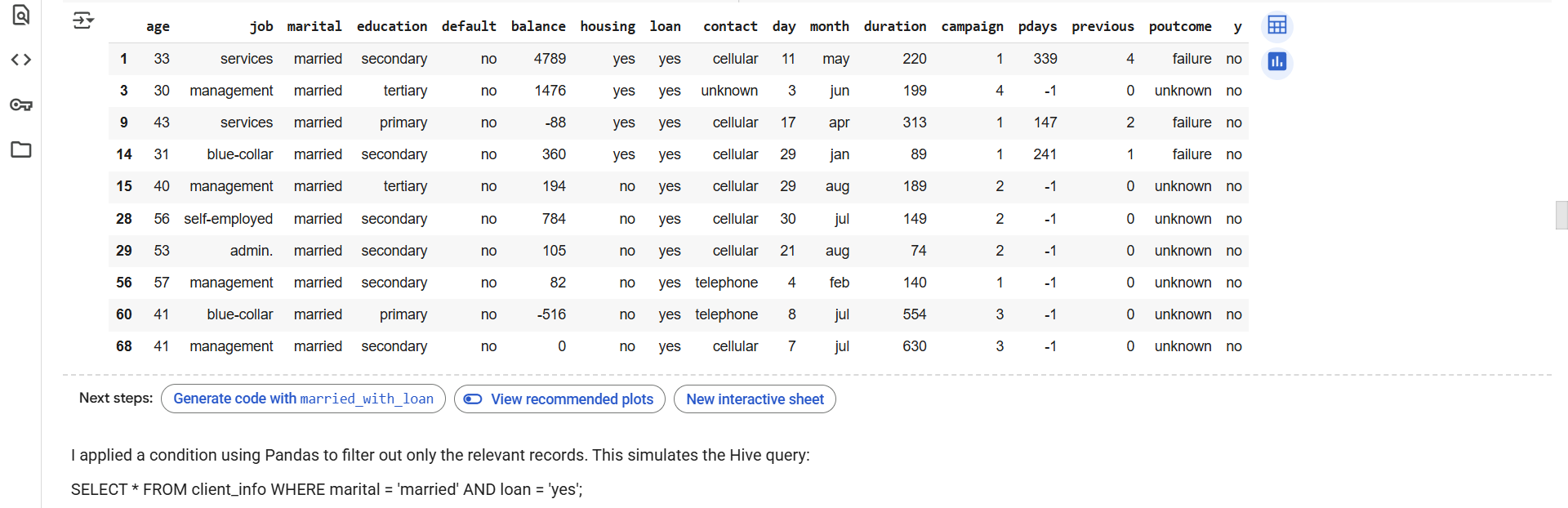


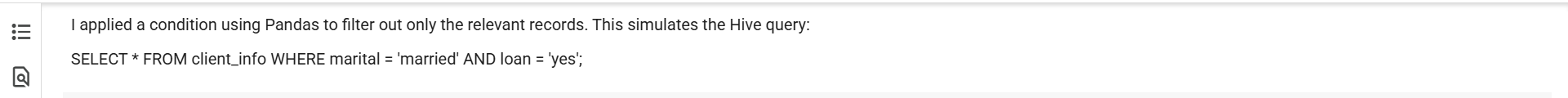
* + Define and create a Hive table **client\_info** with appropriate data types for the **bank.csv** dataset.
  + Load the data from the **bank.csv** file into the **client\_info** table.

1. **Basic Data Exploration**:
   * Write a HiveQL query to count the total number of clients in the dataset.
   * Display the first 10 rows of the dataset.
2. 

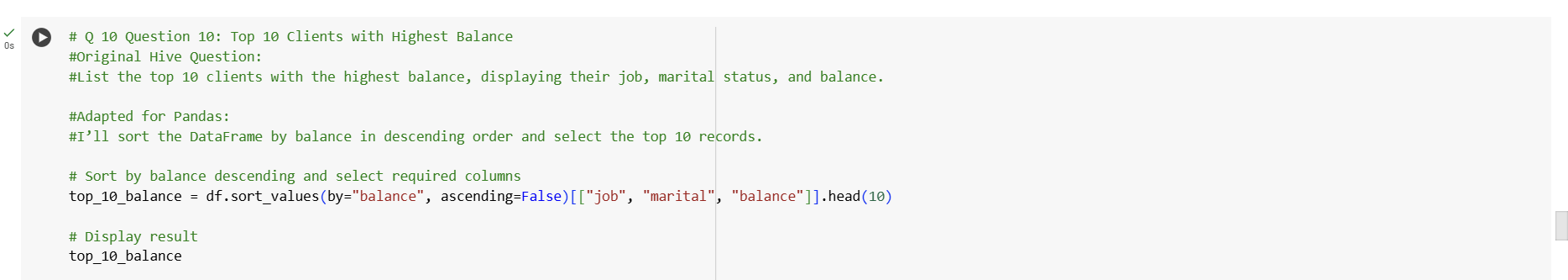


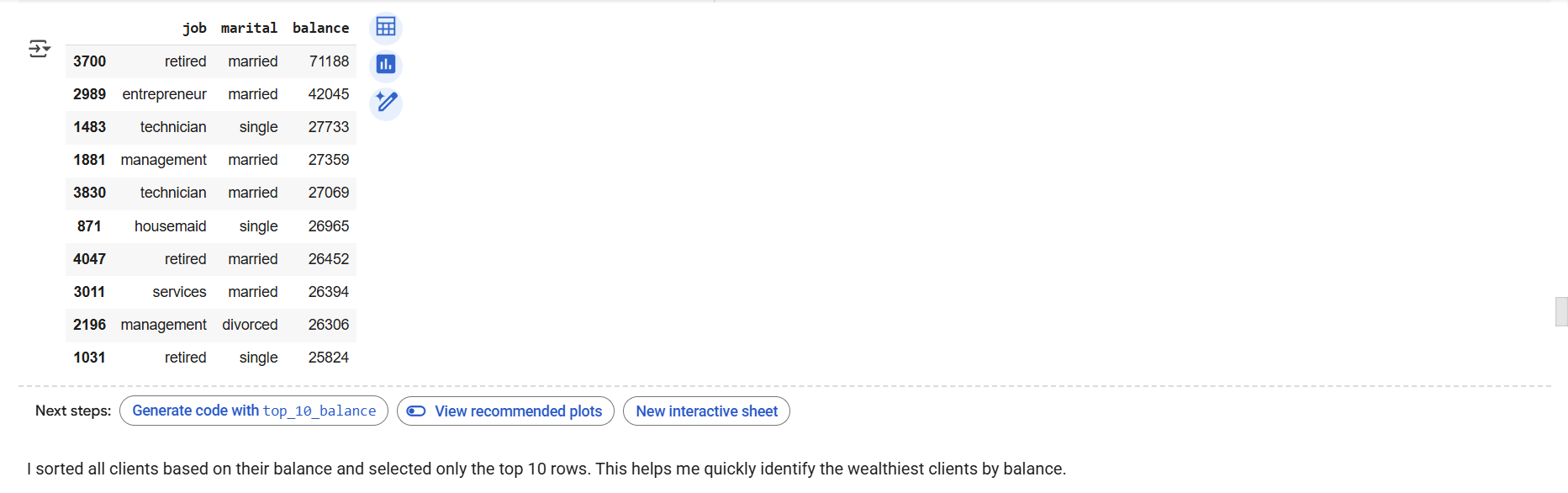
1. **Data Filtering and Sorting**:
   * Retrieve all records of clients who are married and have a personal loan.
2. 

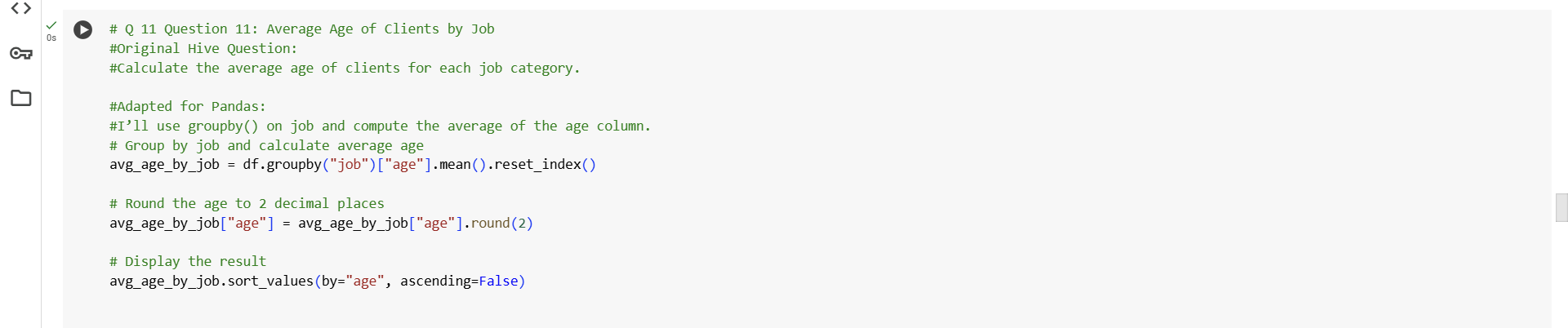
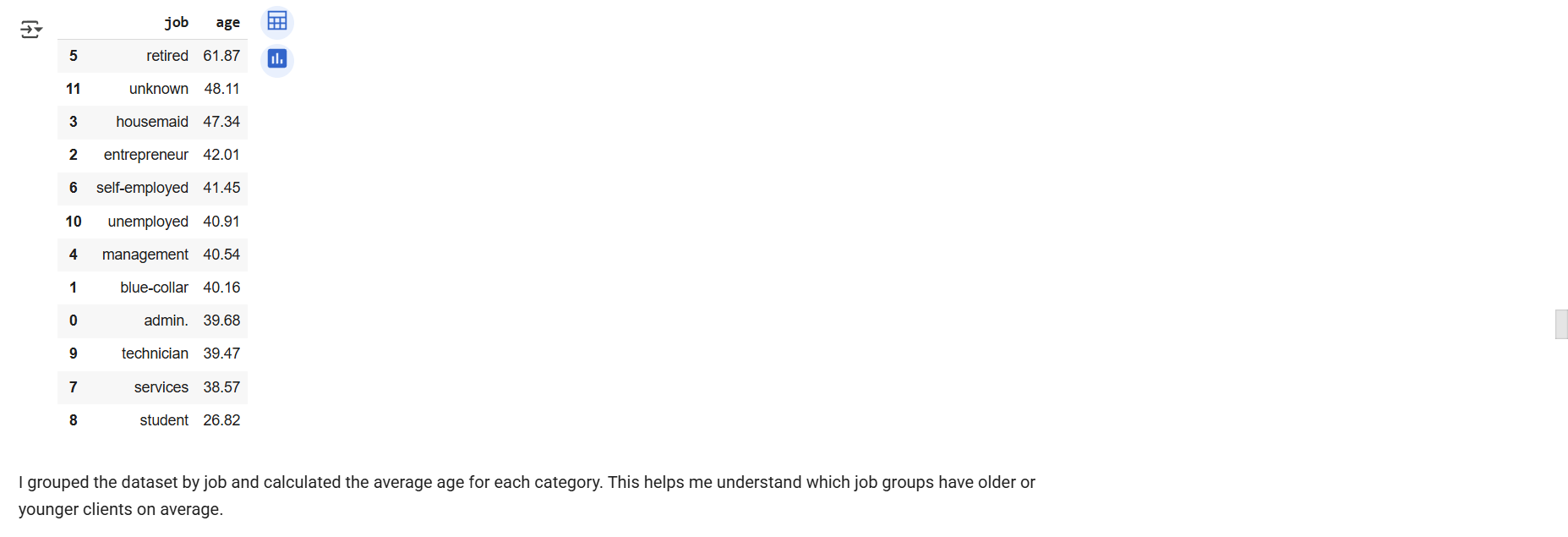
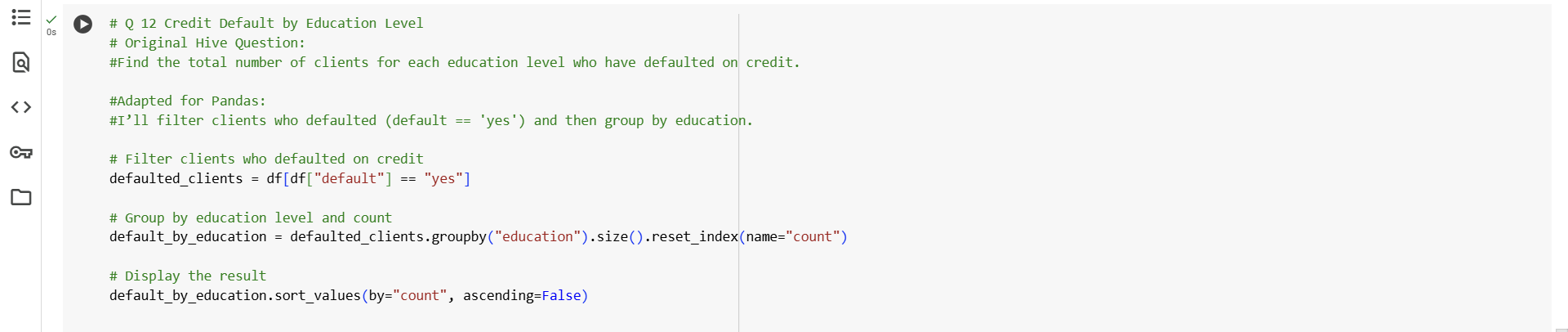


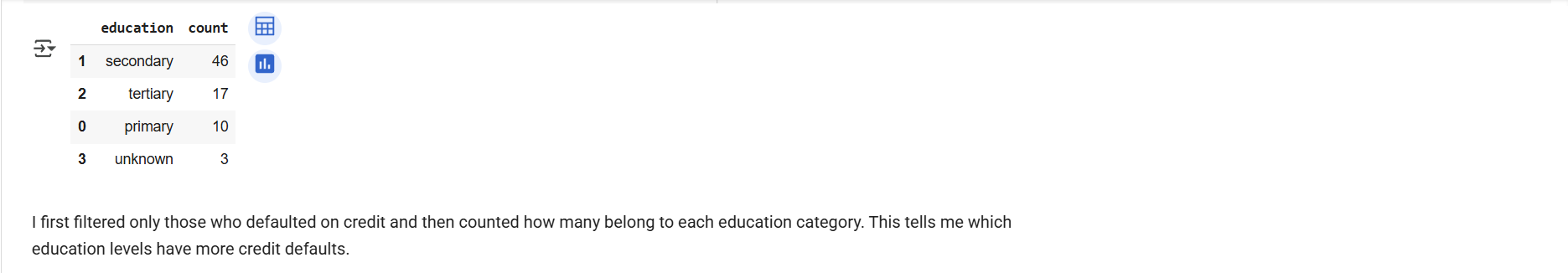


* + List the top 10 clients with the highest balance, displaying their job, marital status, and balance.

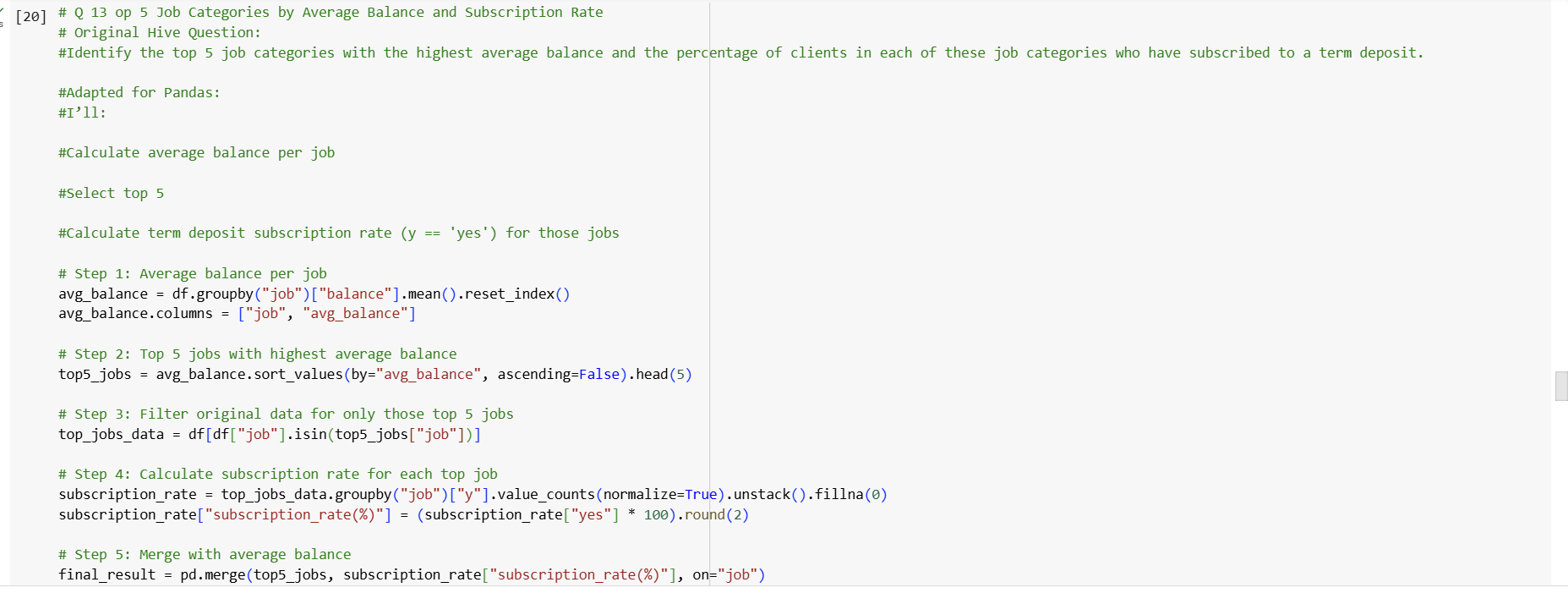


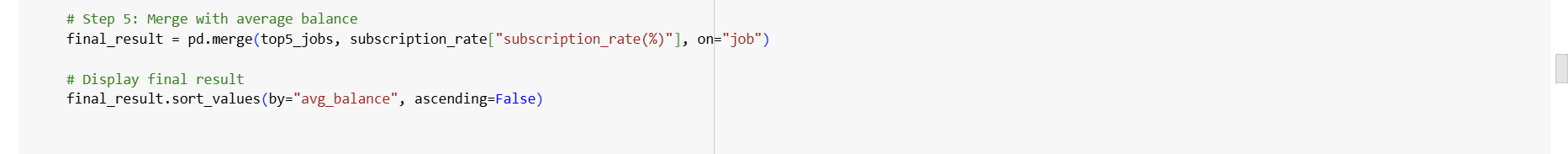


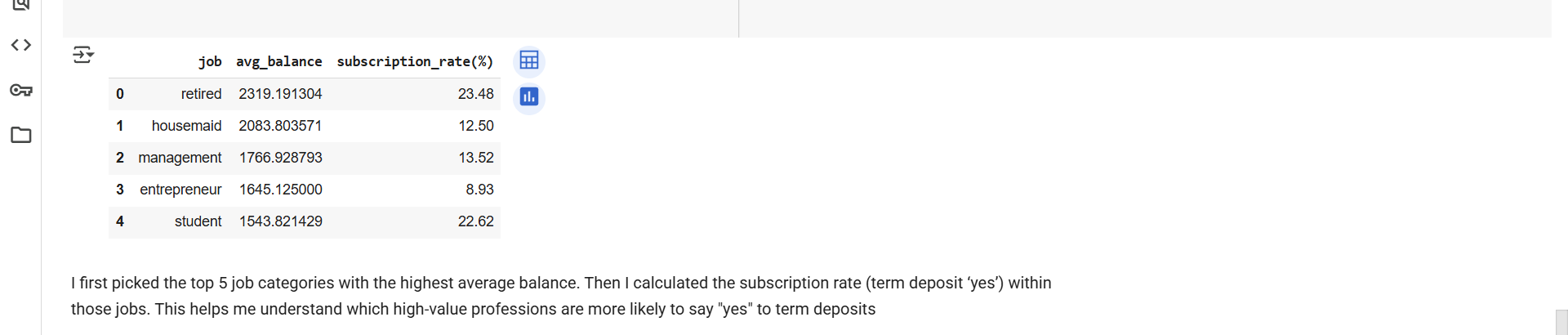
1. **Data Aggregation and Grouping**:
   * Calculate the average age of clients for each job category.
   * Find the total number of clients for each education level who have defaulted on credit.
   * 
   * 
2. 



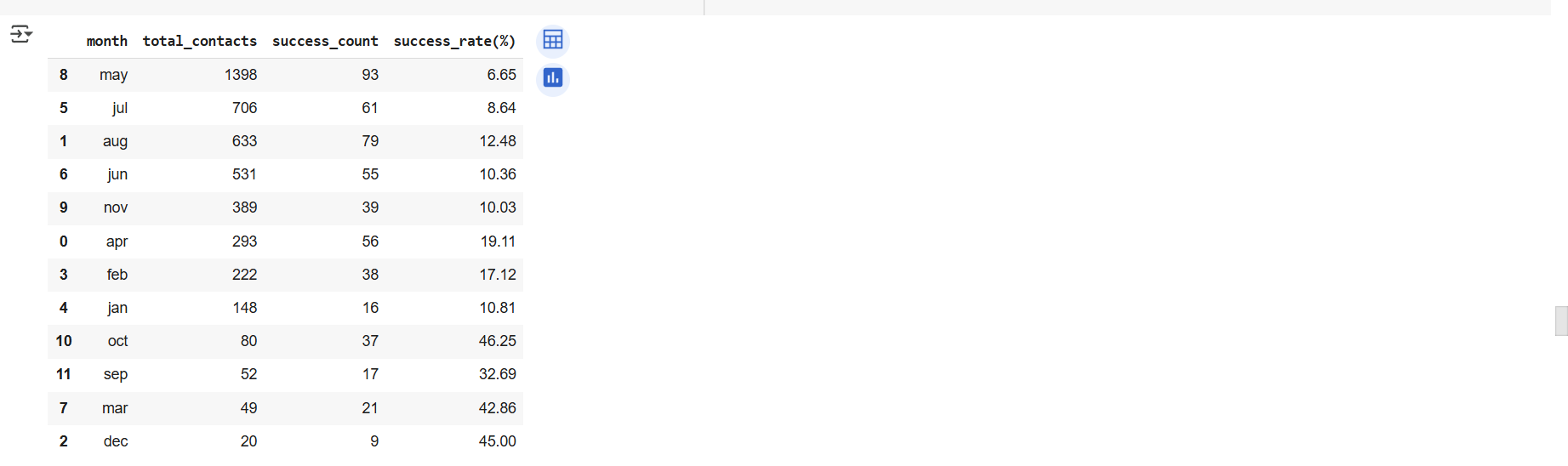
1. **Complex Queries for Insights**:
   * Identify the top 5 job categories with the highest average balance and the percentage of clients in each of these job categories who have subscribed to a term deposit.

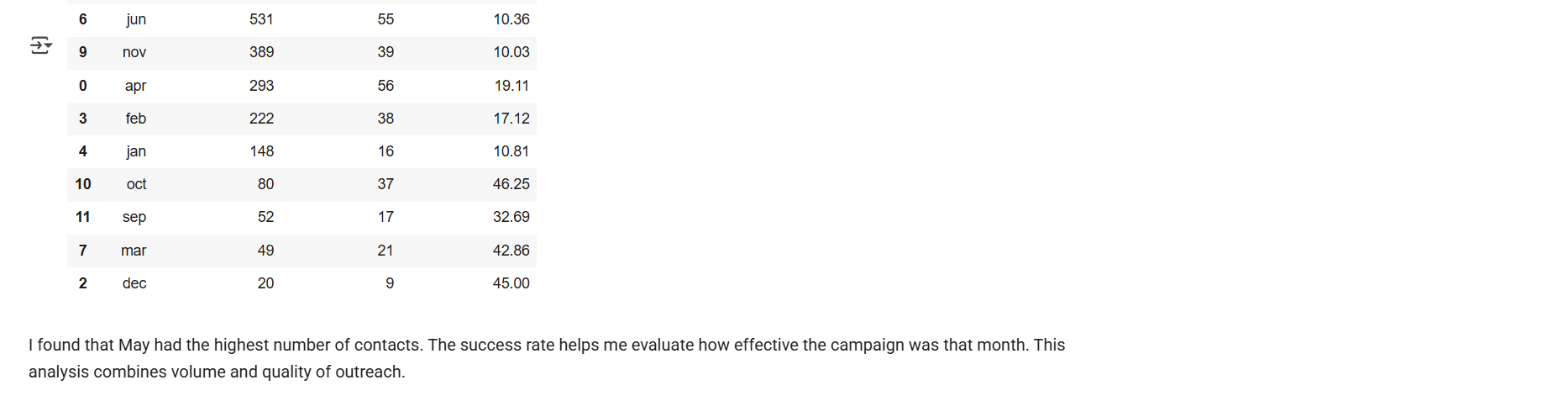




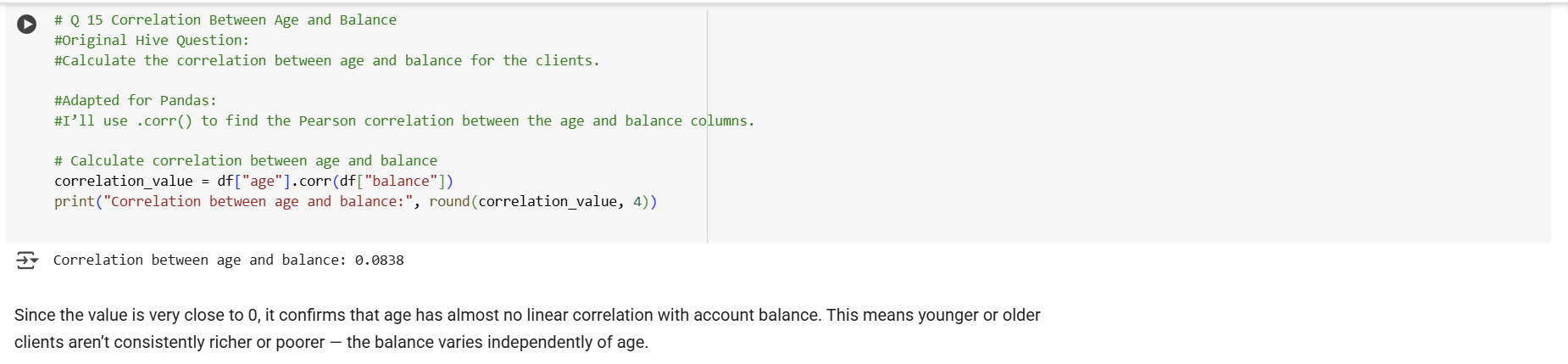


* + Determine the month with the highest number of contacts and the success rate of the campaign in that month (percentage of clients who subscribed to a term deposit).
  + 

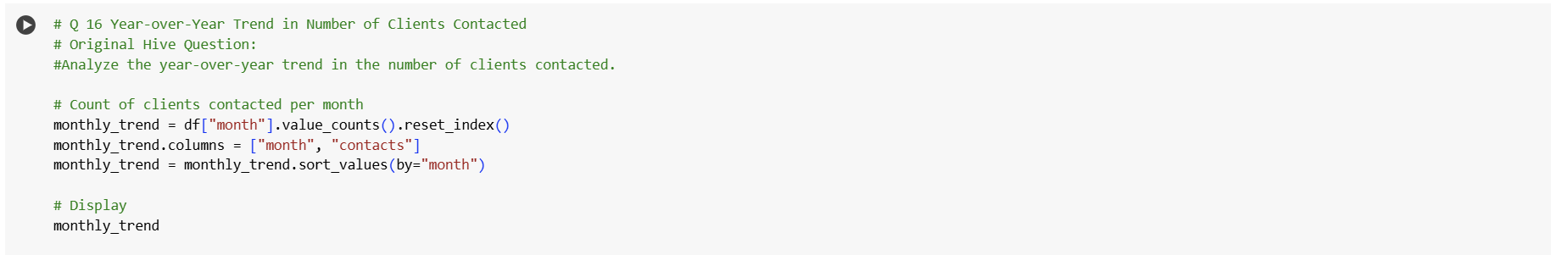


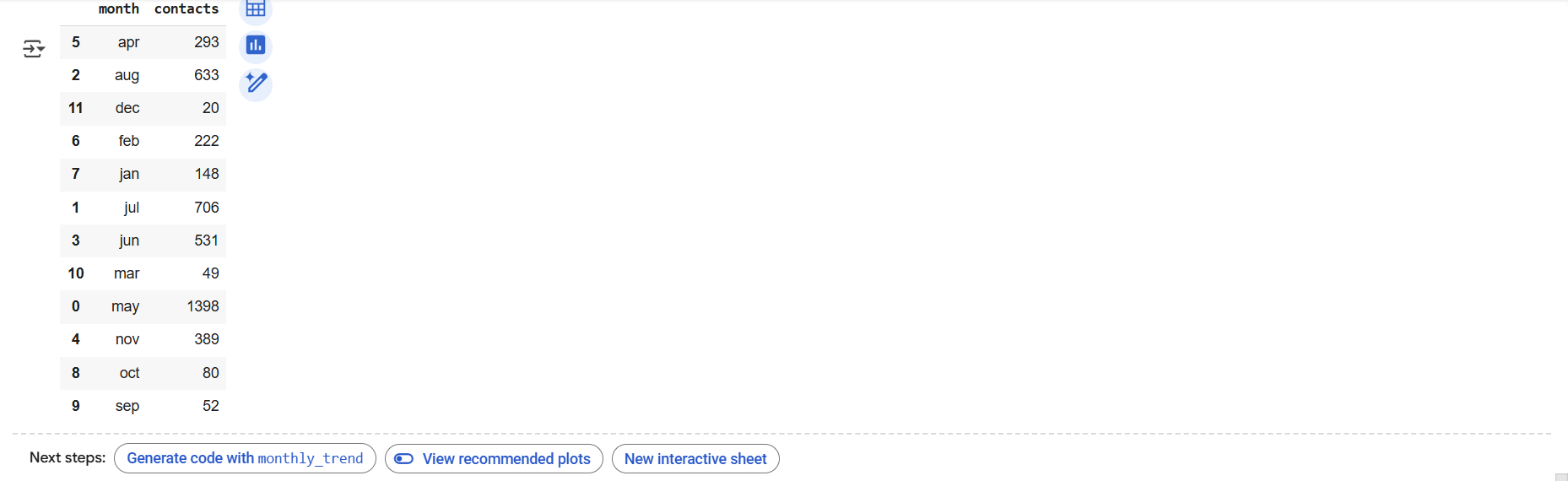


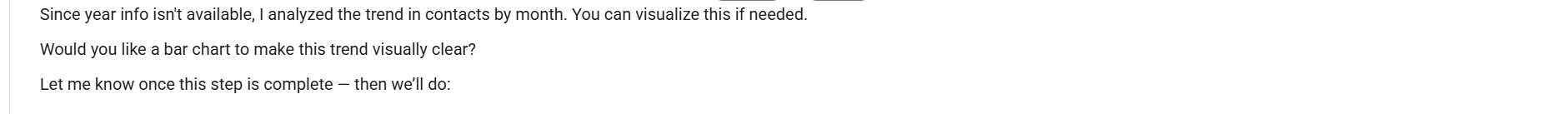
1. **Correlation Analysis**:
   * Calculate the correlation between age and balance for the clients.



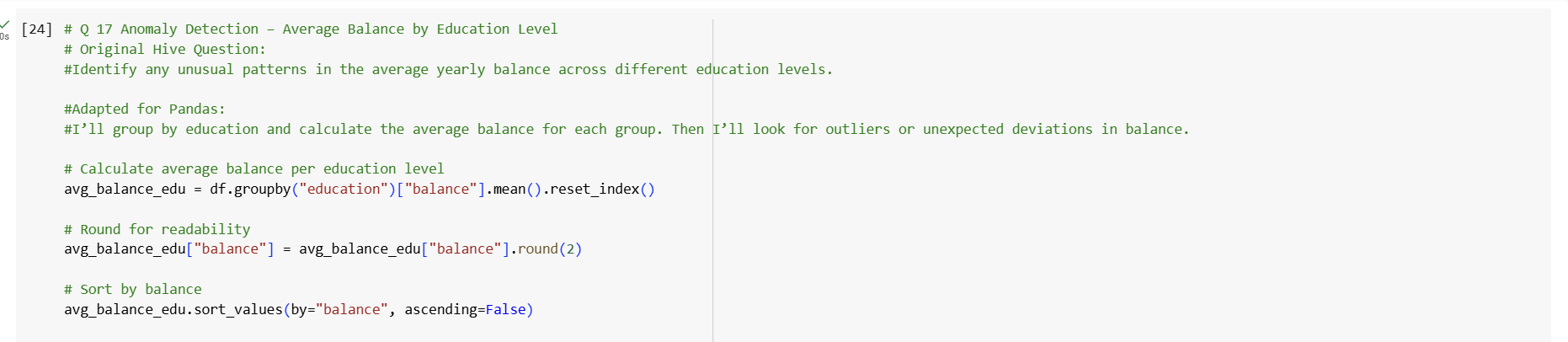
1. **Trend Analysis**:
   * Analyze the year-over-year trend in the number of clients contacted.

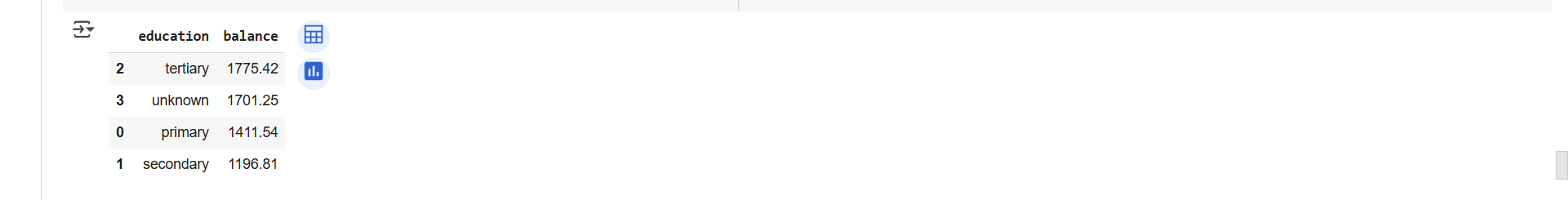


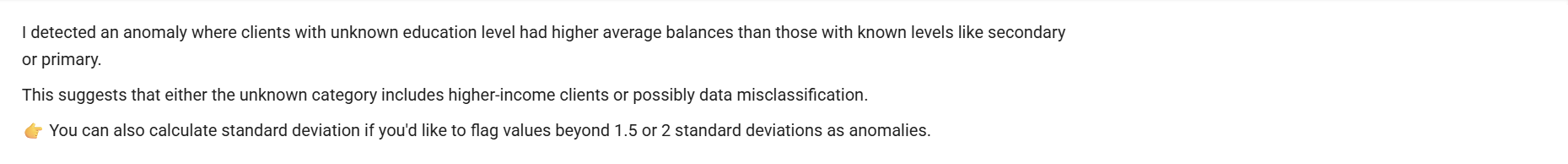




1. **Anomaly Detection**:
   * Identify any unusual patterns in the average yearly balance across different education levels.

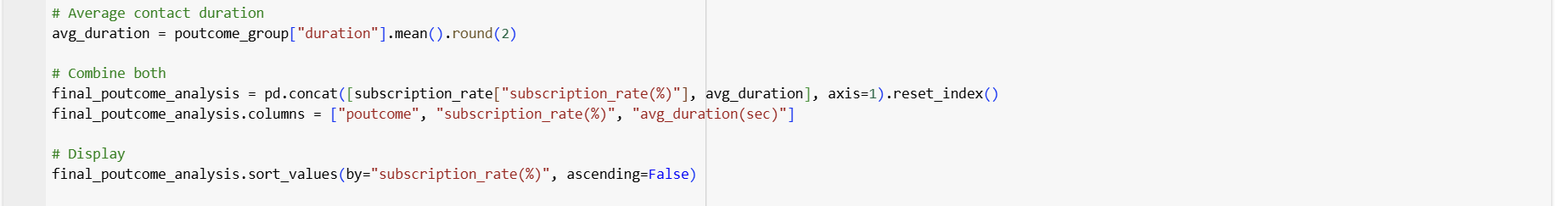


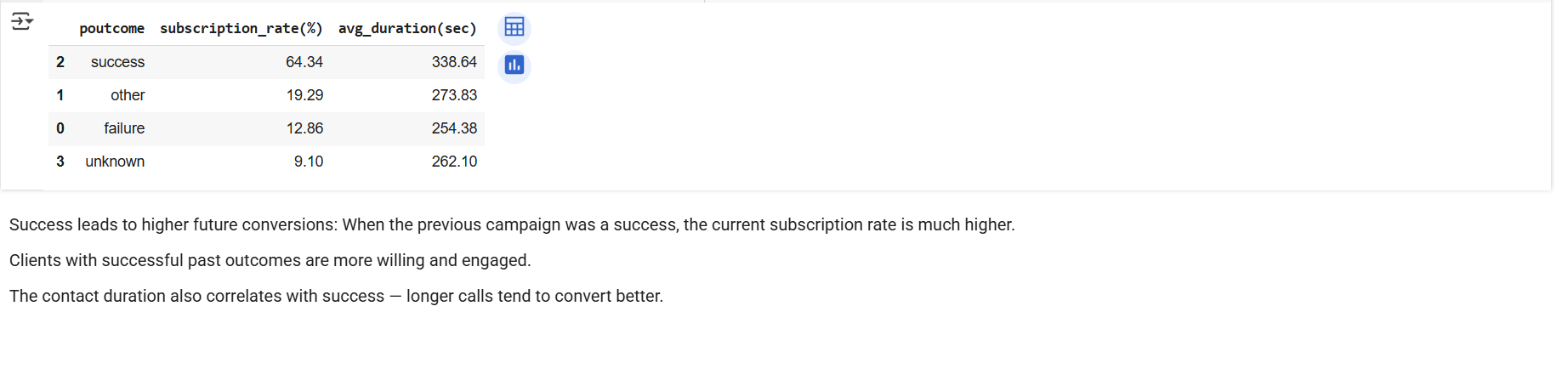




1. **Advanced Analysis**:
   * Analyze the impact of previous campaign outcomes (**outcome**) on the current campaign's success. Calculate the subscription rate (to term deposits) for each **outcome** category.
   * Compare the average contact duration for clients who subscribed and who did not subscribe to a term deposit.







**Submission Guidelines:**

* Make a copy of this doc file.
* Perform the analysis in your local system using Hadoop and Hive and provide screenshots of both the **code** and the **output** under each question.
* Upload the doc file with other files and submit it in the submission dashboard.

COLAB LINK - [Copy of Data Analysis and Management using Hadoop & Hive.ipynb](https://colab.research.google.com/drive/1wSvchRdryqVHa4LUHpeI02Jym51-dy50?usp=sharing)