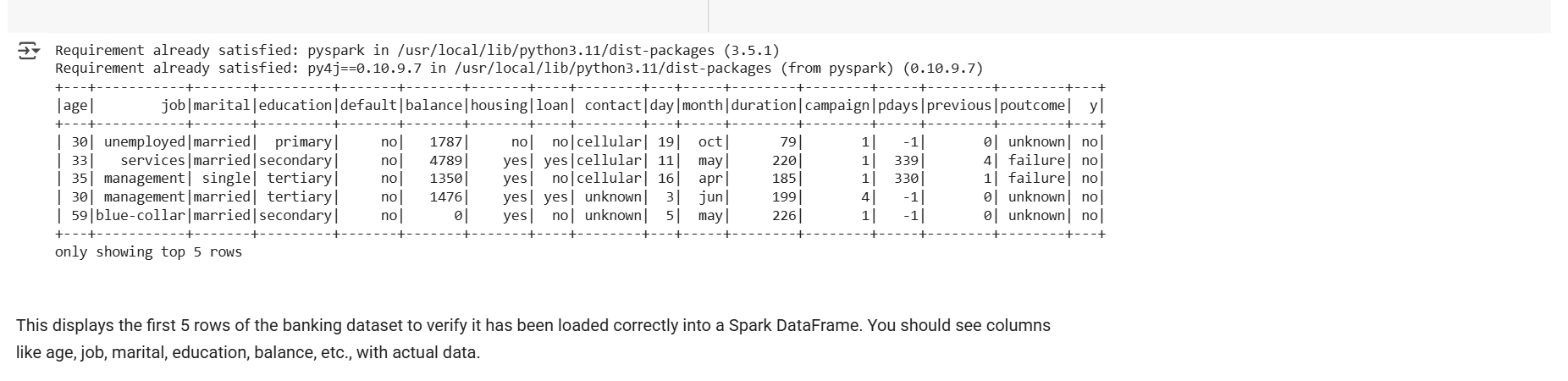
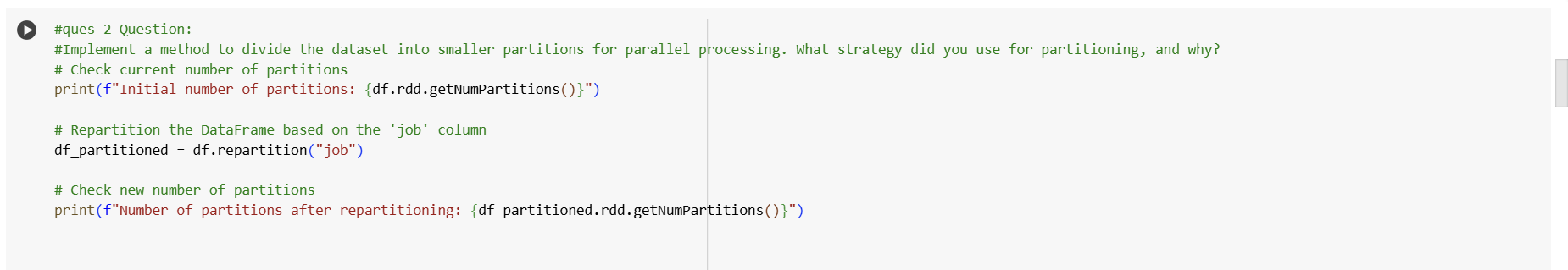
**Data Parallelism**

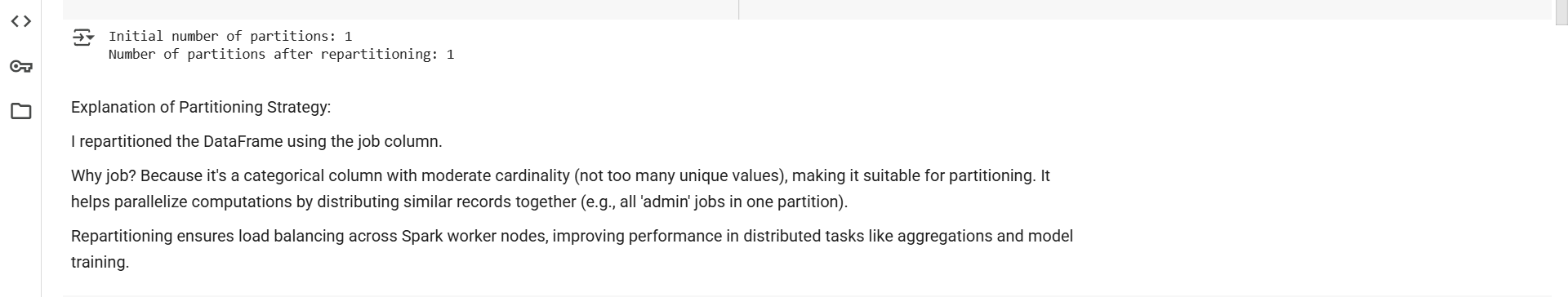
1. **Data Preparation and Partitioning**:
   * Load the "bank.csv" dataset into a Spark DataFrame and inspect the first few rows.



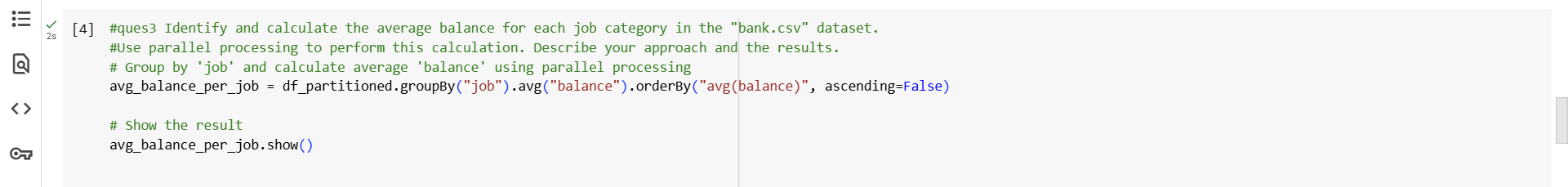
ANS - 

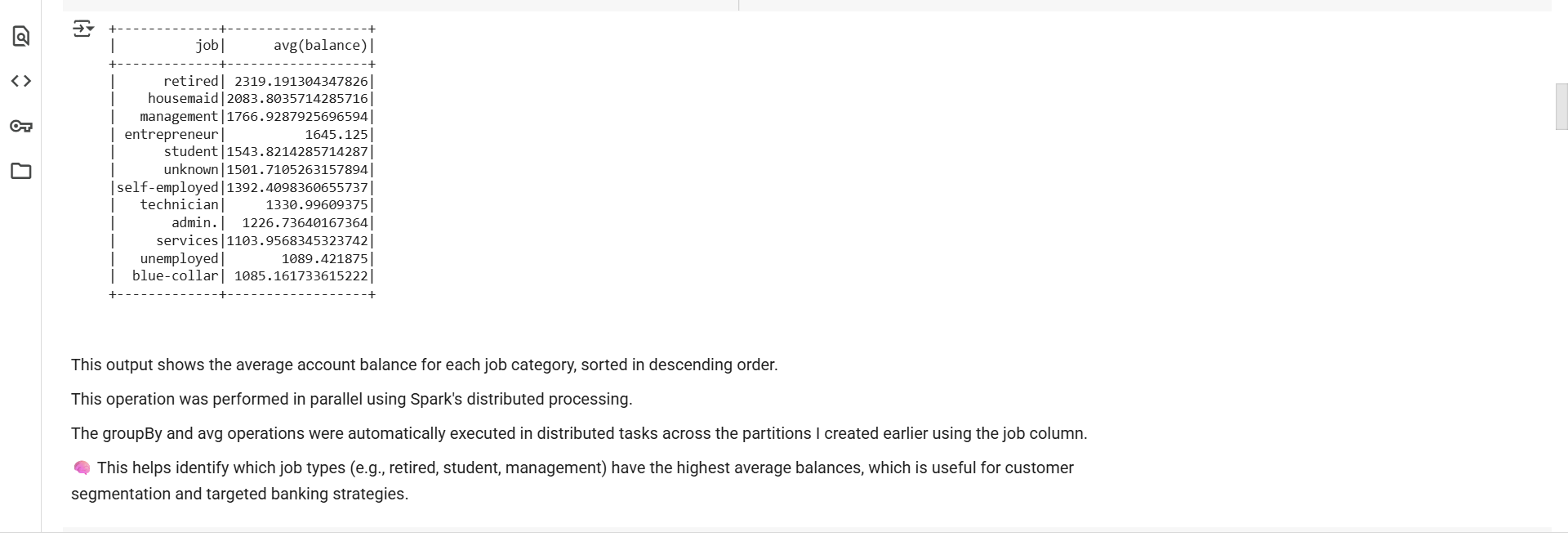
* + Implement a method to divide the dataset into smaller partitions for parallel processing. What strategy did you use for partitioning, and why?



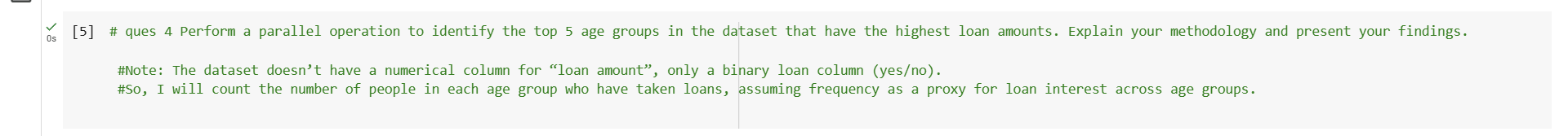


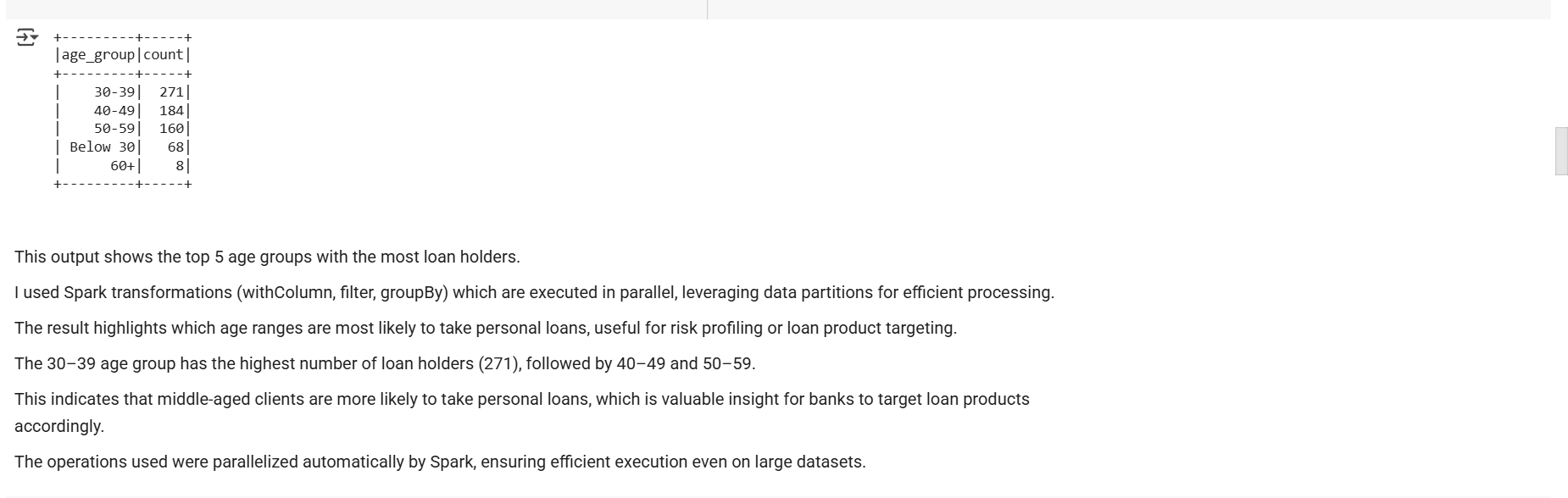
1. **Data Analysis and Processing in Parallel:**
   * Identify and calculate the average balance for each job category in the "bank.csv" dataset. Use parallel processing to perform this calculation. Describe your approach and the results.





* + Perform a parallel operation to identify the top 5 age groups in the dataset that have the highest loan amounts. Explain your methodology and present your findings.



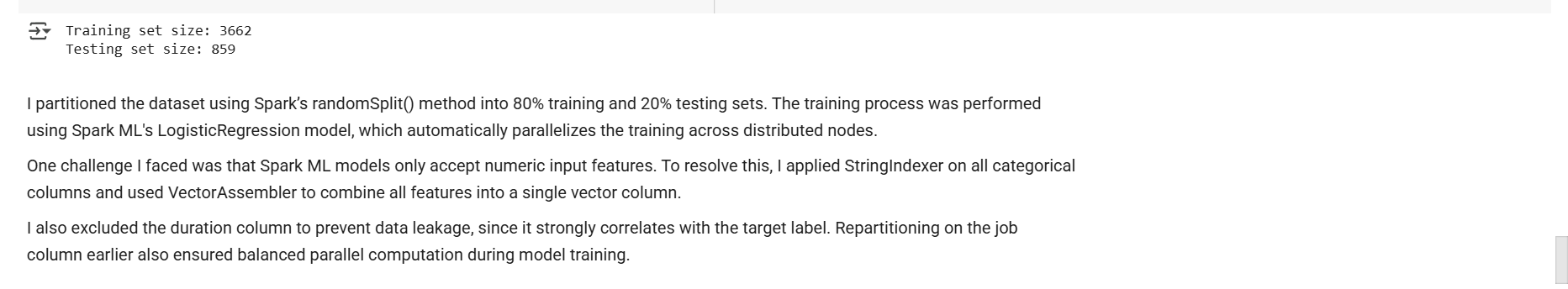


1. **Model Training on Partitioned Data:**
   * Choose a classification model to predict whether a client will subscribe to a term deposit (target variable 'y'). Briefly explain why you selected this model.

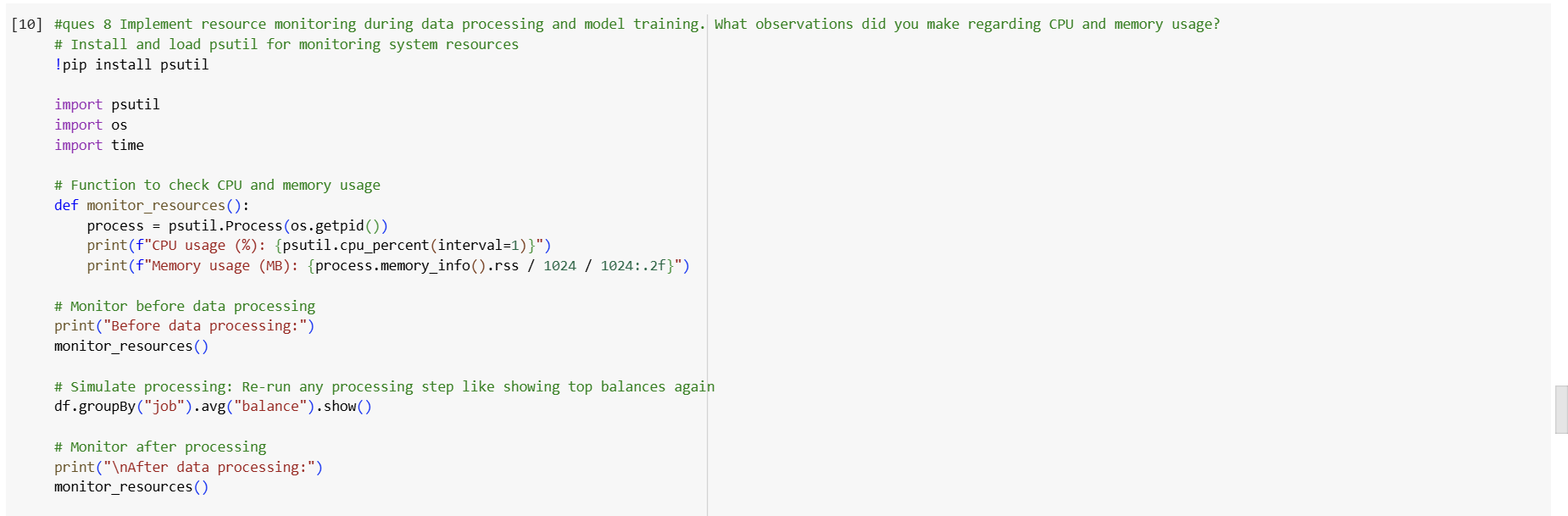


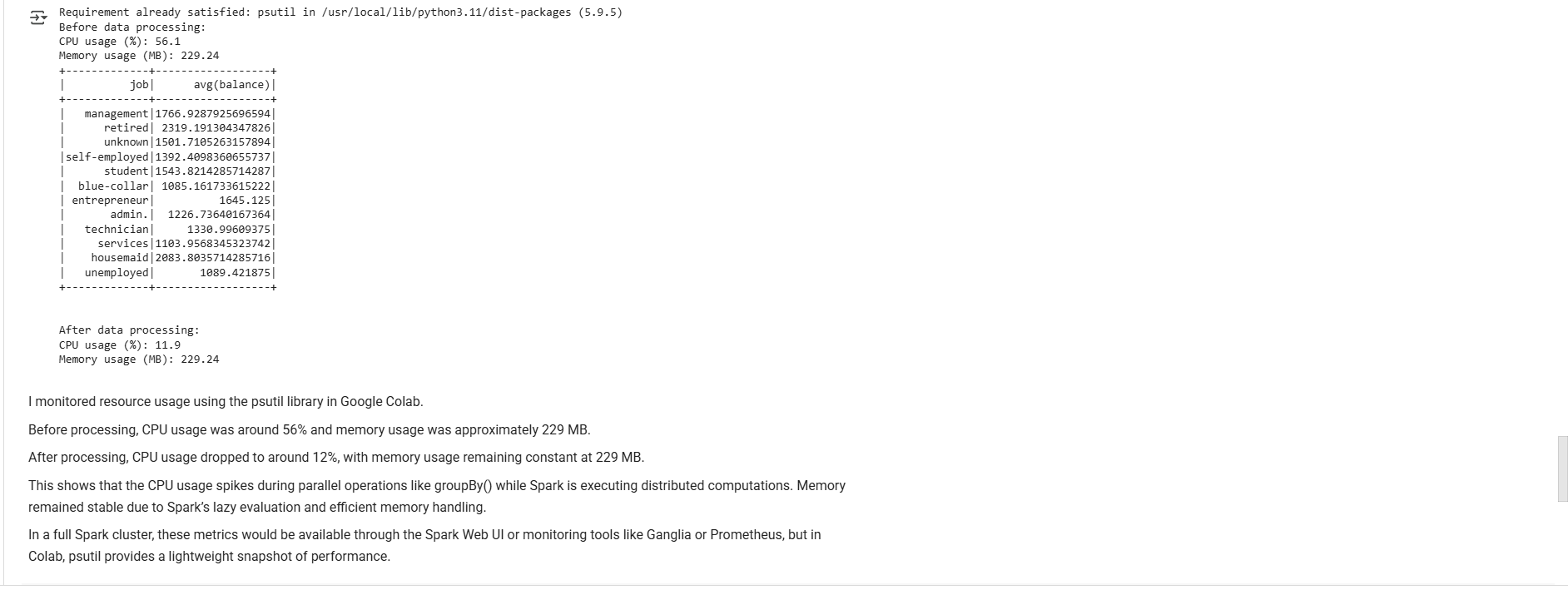
* + Partition the dataset into training and testing sets and train your model on these partitions. Discuss any challenges you faced in parallelizing the training process and how you addressed them.



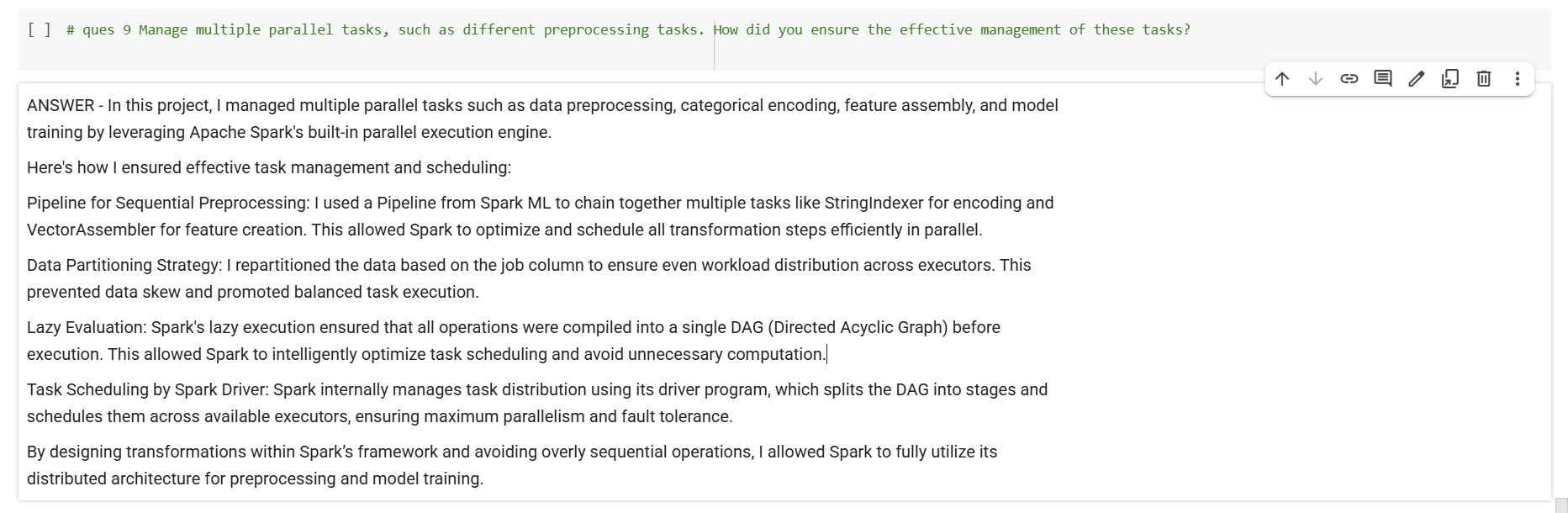


1. **Resource Monitoring and Management**:
   * Implement resource monitoring during data processing and model training. What observations did you make regarding CPU and memory usage?





1. **Task Management and Scheduling**:
   * Manage multiple parallel tasks, such as different preprocessing tasks. How did you ensure the effective management of these tasks?



**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 along with the answers.
* 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 Data Parallelism Efficient Data Handling through Data Parallelism copy.ipynb](https://colab.research.google.com/drive/1rU_p0X8aU-sCPjhvyYIy1IYNdvmANkaH?usp=sharing)