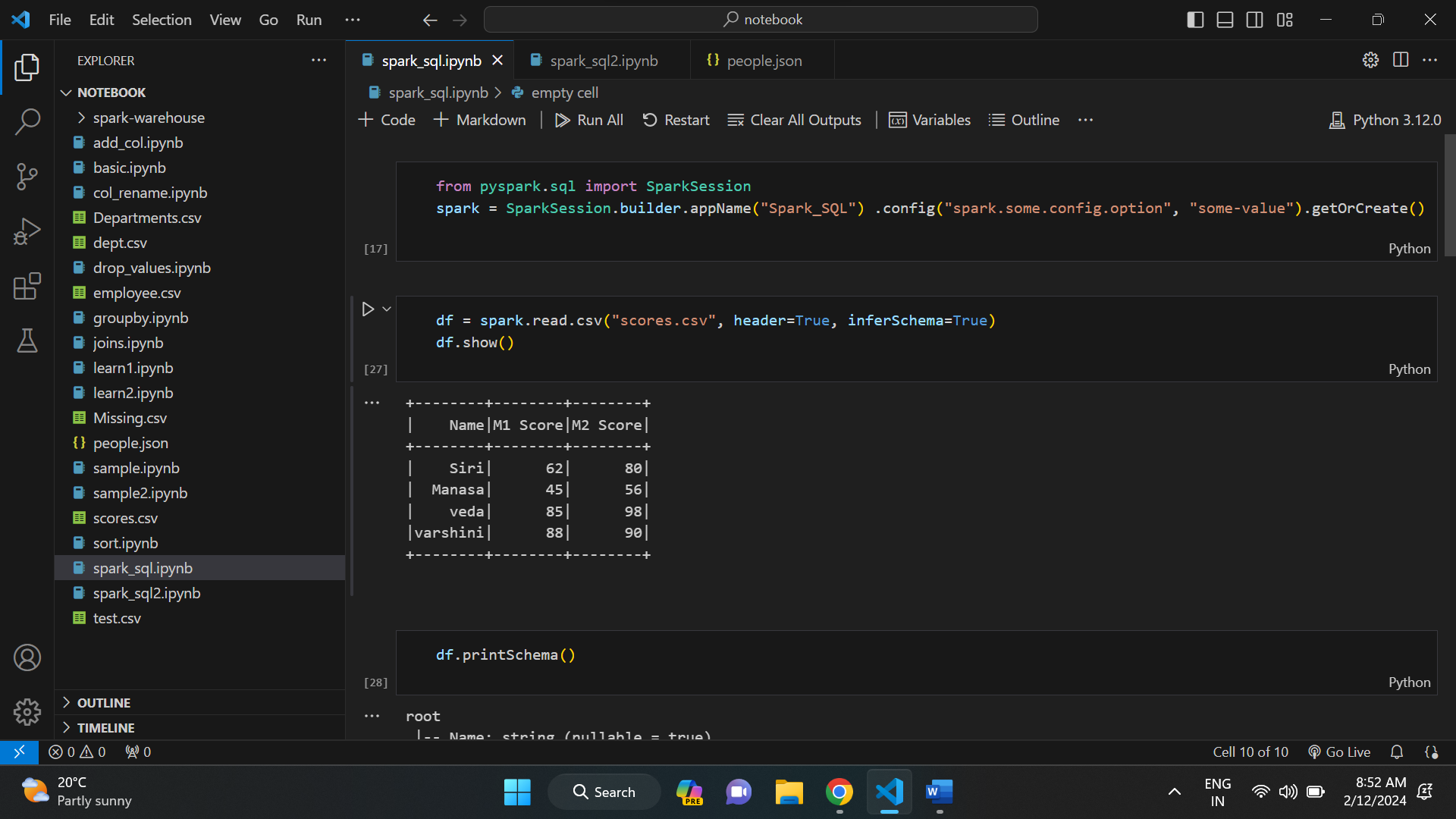
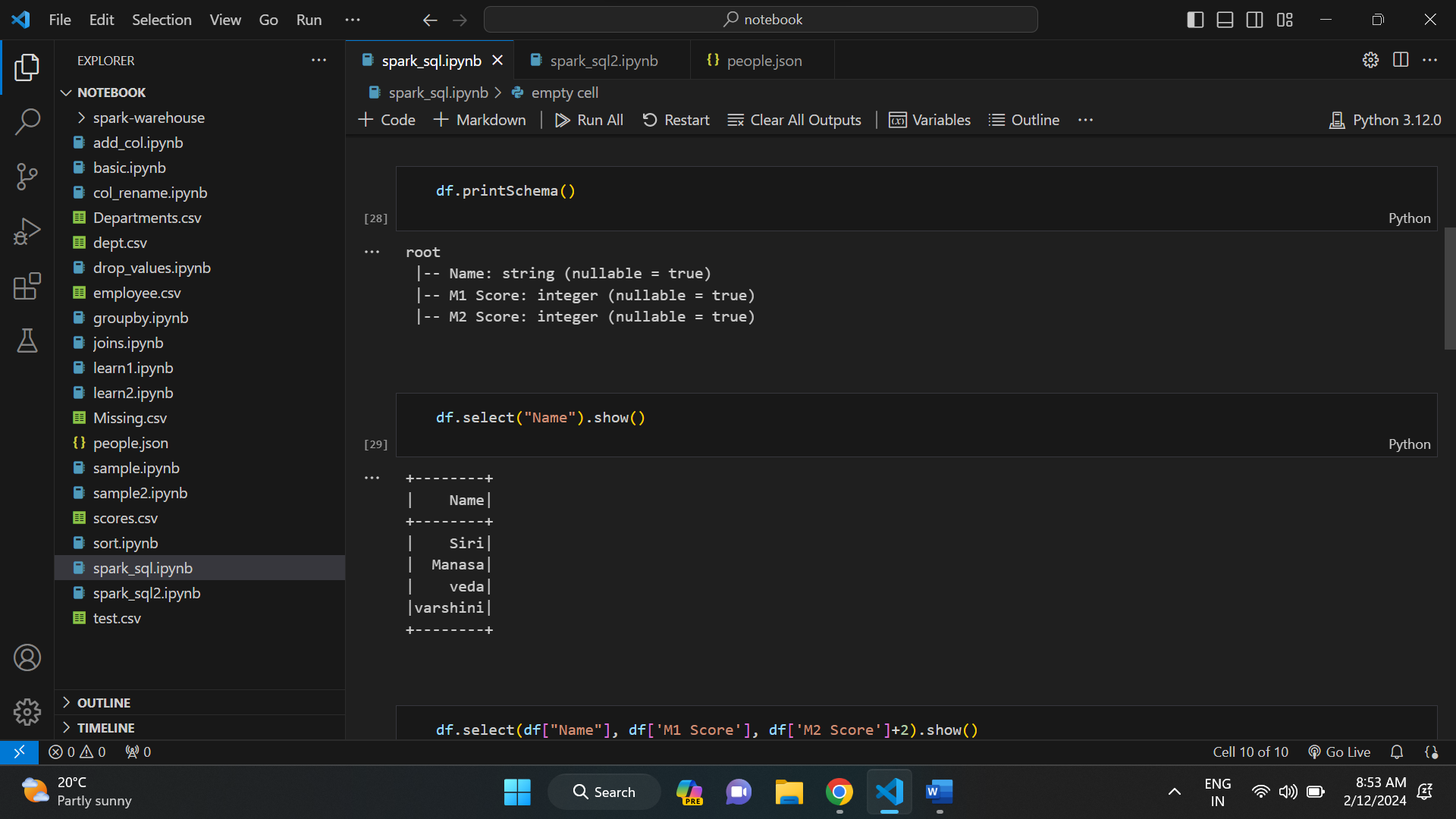
**Data Engineering Assignment – 15**

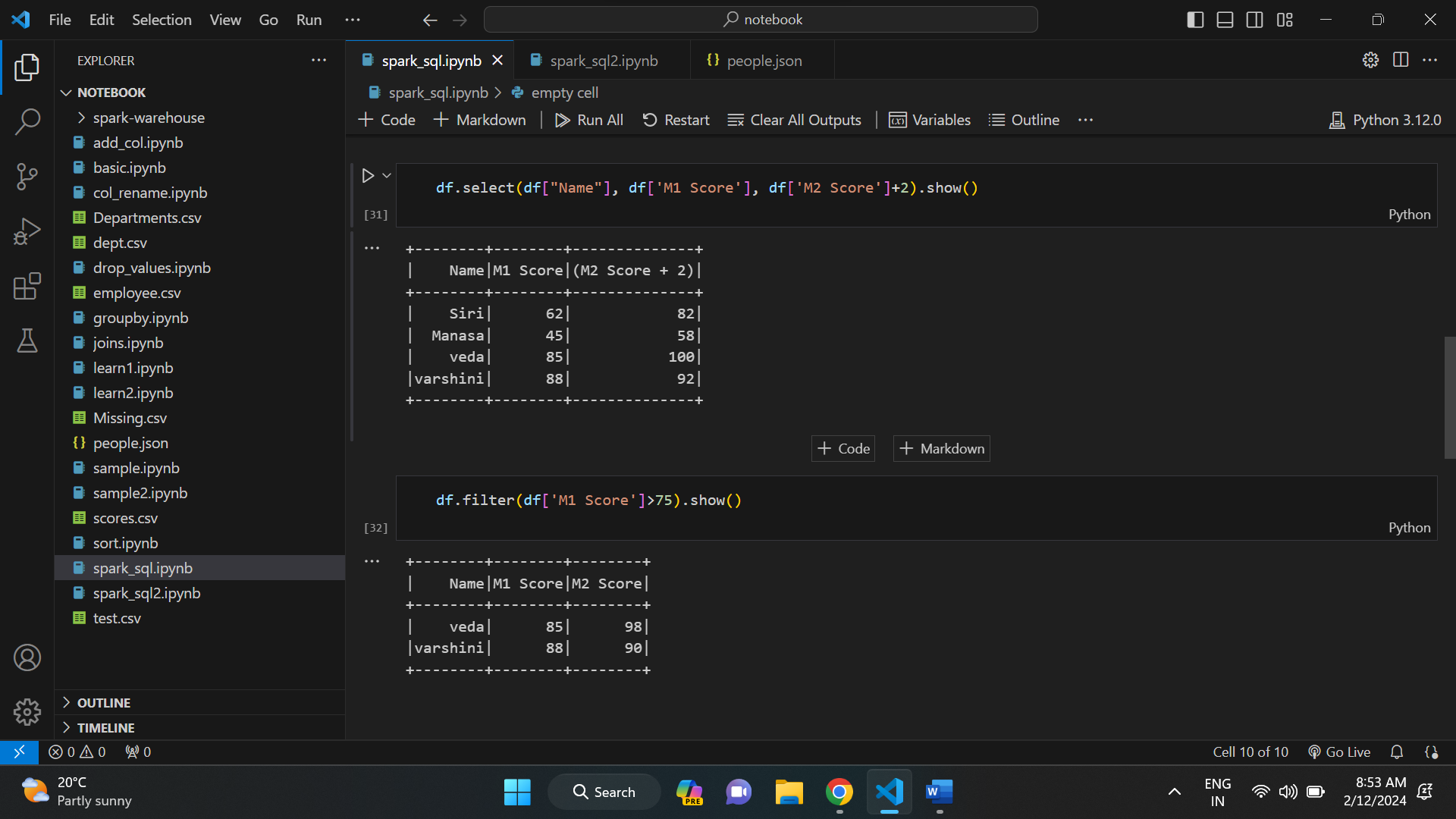
**Spark SQL:**

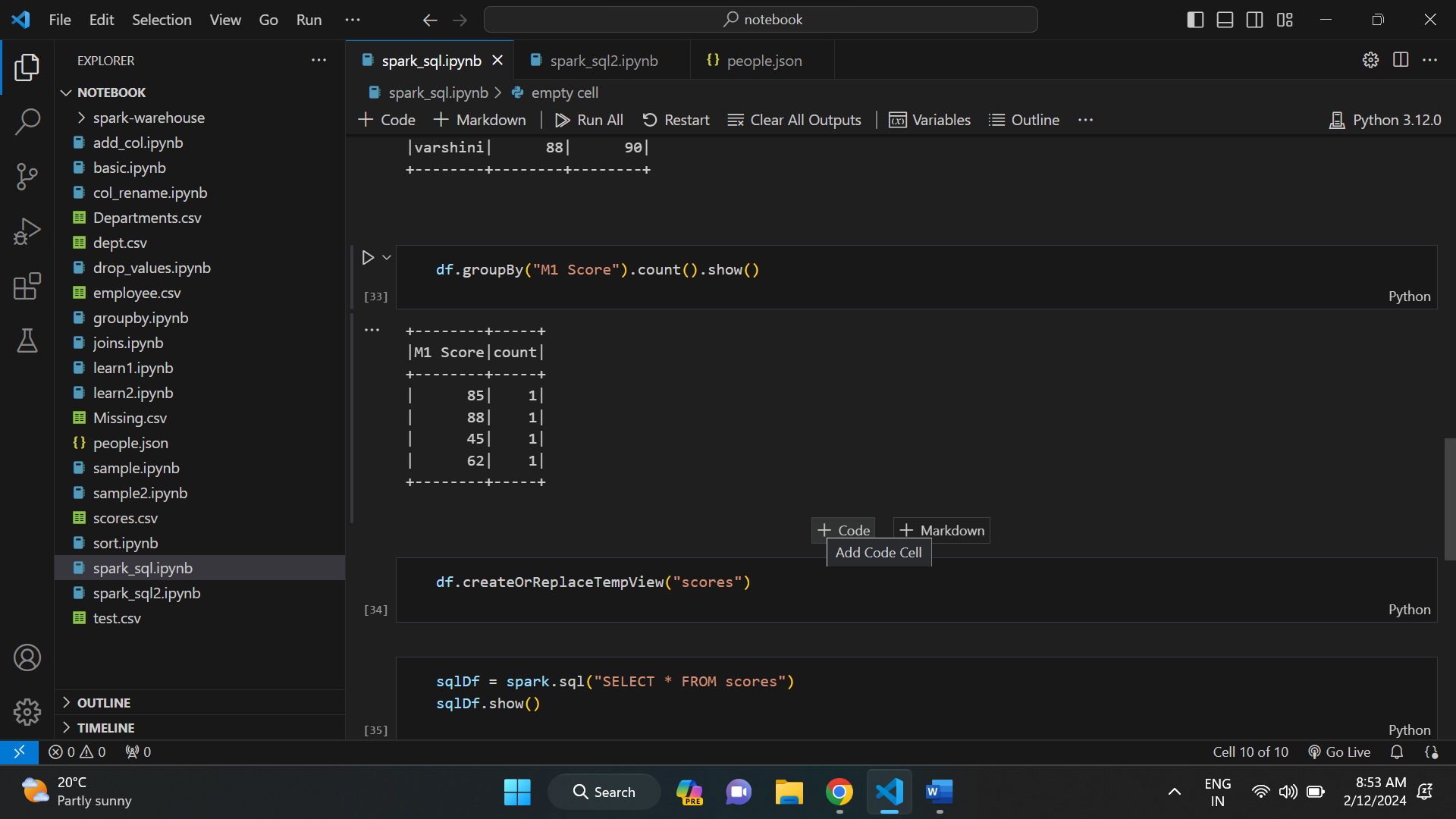
* It is a module for structured data processing.
* It is a component on top of spark core that introduces a new data abstraction called schema RDD.
* It enables unmodified Hadoop Hive queries to run up to 100x faster on existing deployments and data.
* Spark SQL comes with the use of Dataframes, a distributed data set. A data frame is used to organize columns adequately, and Dataframes can be easily constructed from different array sources such as Hive Tables or structured data files.

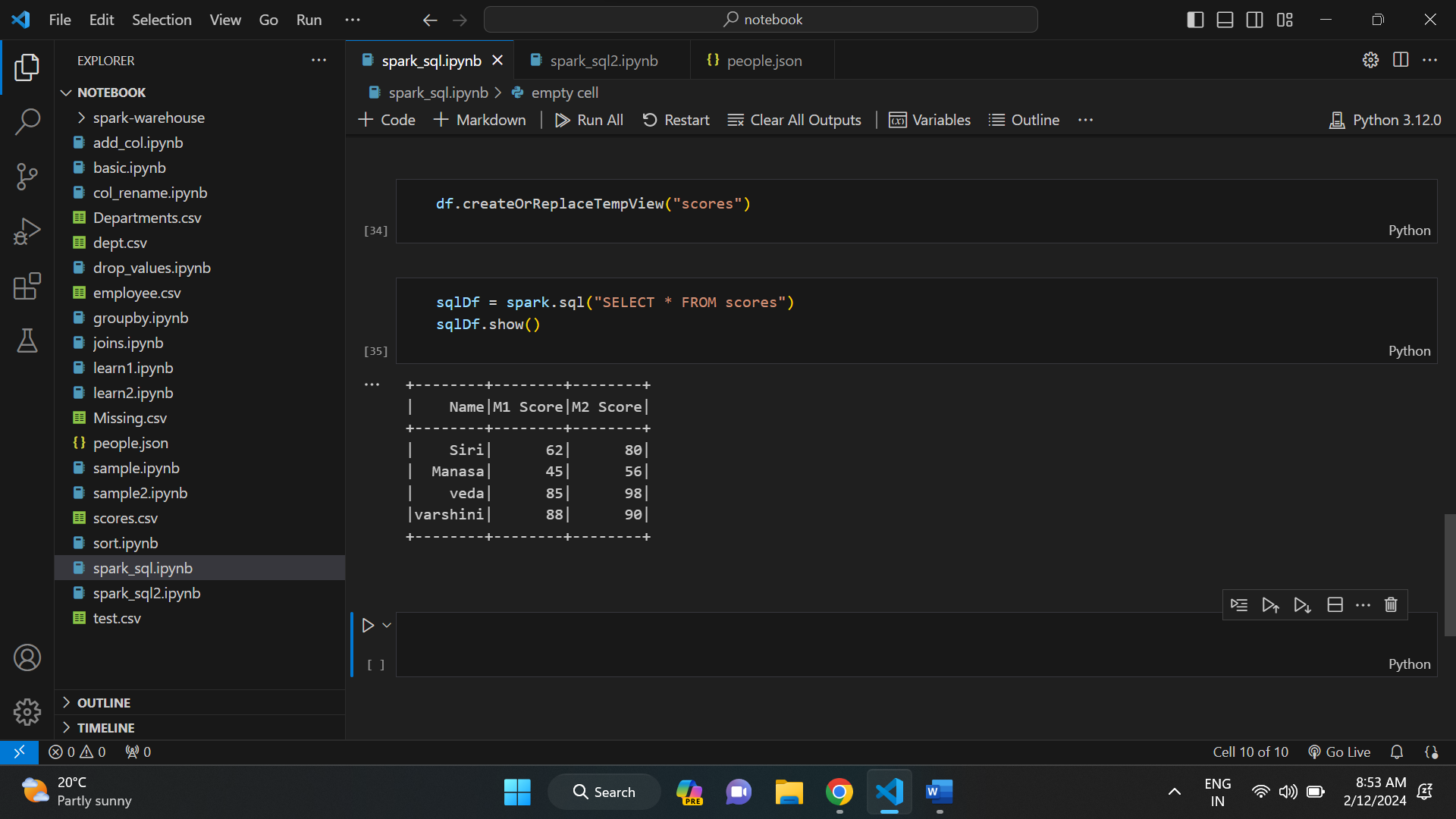
Execution:



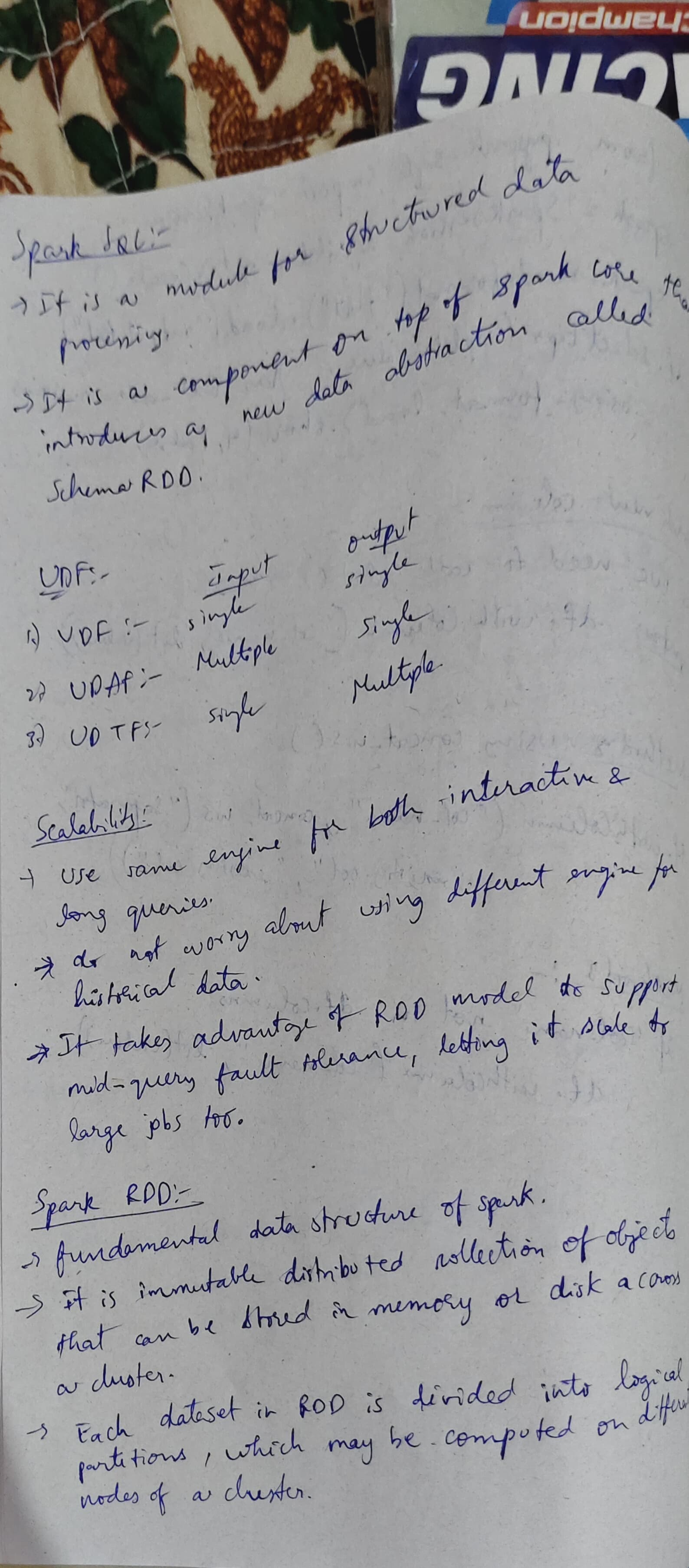


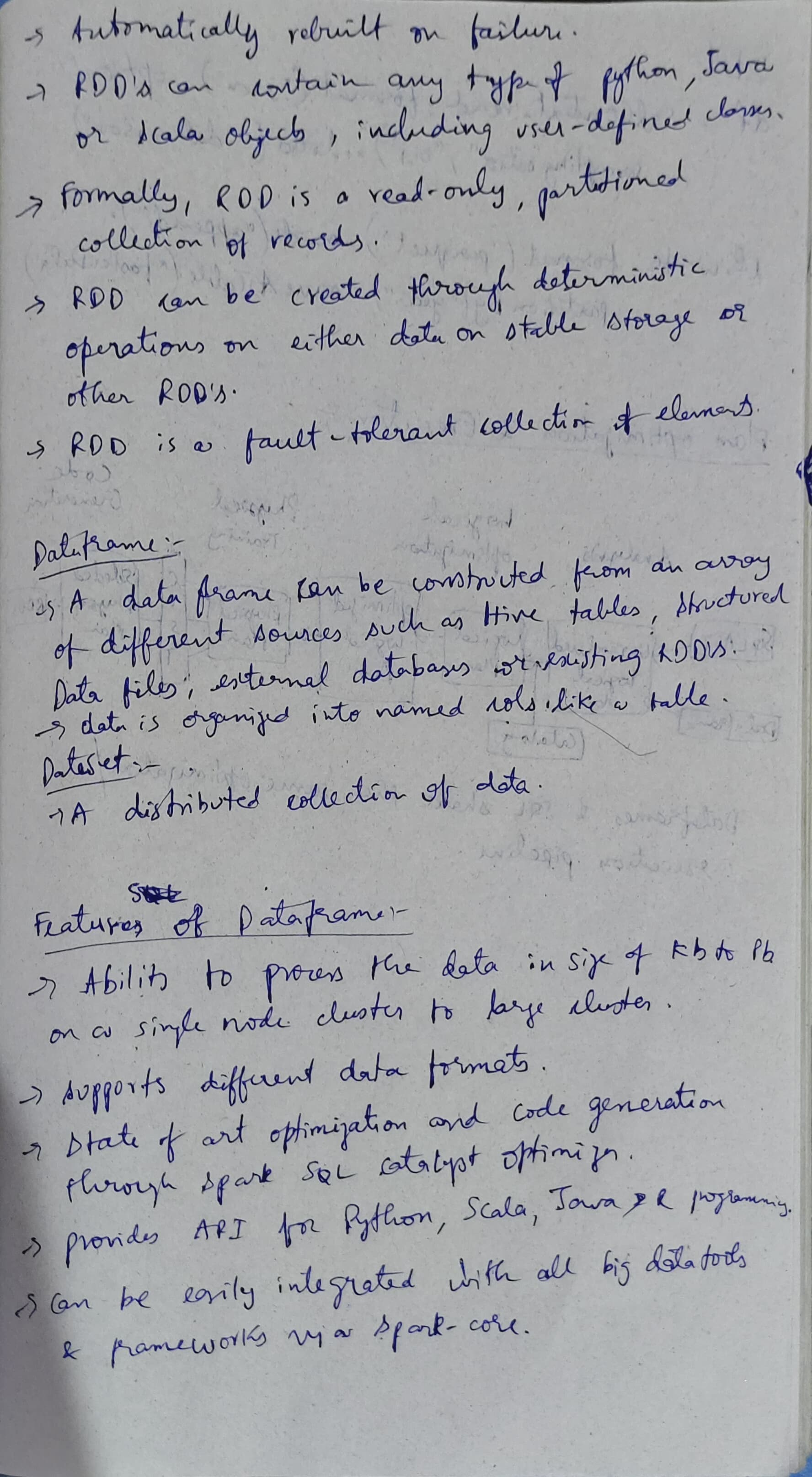






**HAND WRITTEN NOTES:**

****

****

