27/12/2023

Chakradhar Bhogapurapu

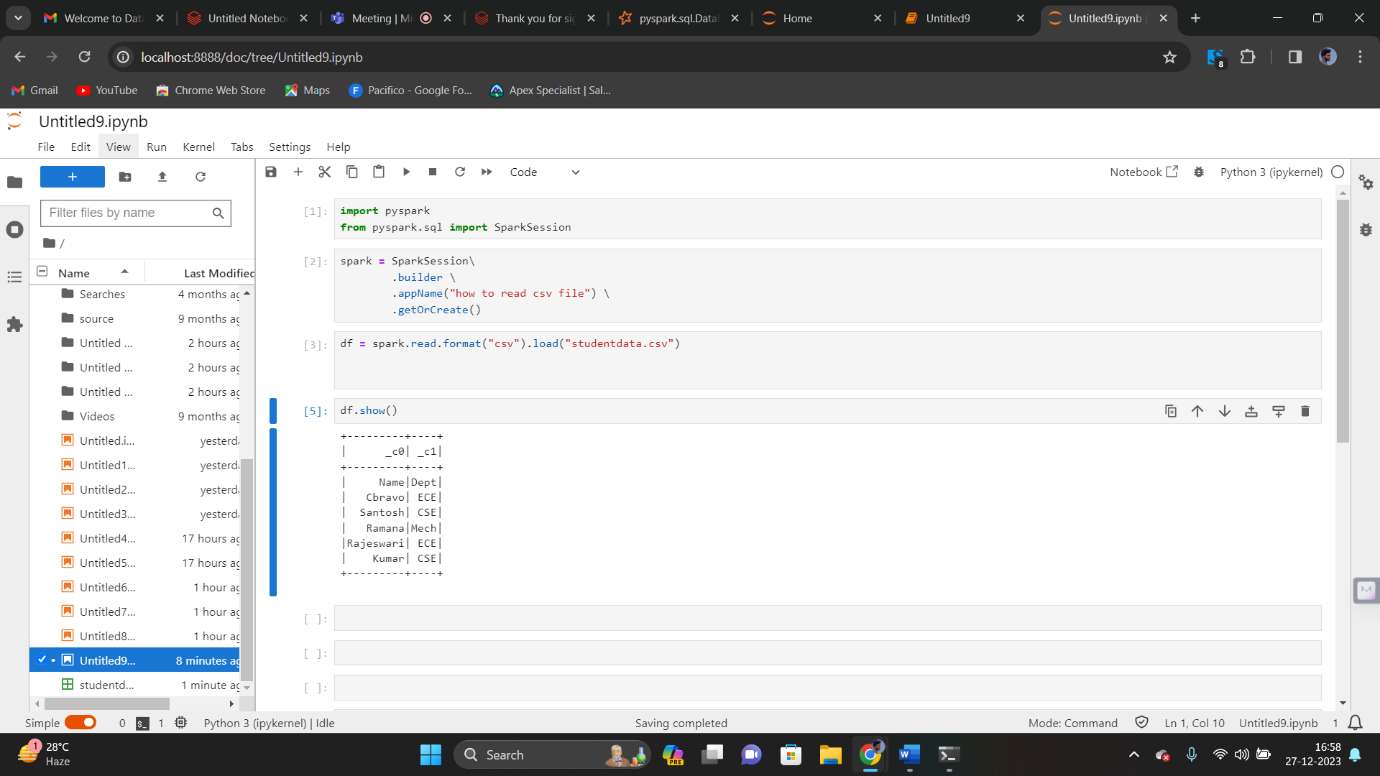
Spark and Pyspark Coding Assessment

Question 1

**Implement Processing JSON and CSV data with PySpark**

**Ans**

**🡪Implementing processing CSV data with pyspark**



**Here, studentdata.csv is a csv file that was uploaded.**

**We need to import pyspark from pyspark.sql**

**Then we need to create the spark session. If we didn’t create spark session, then we will get error like “spark” variable is not define.**

**After creating spark session, we need to load the csv file in read format.**

**Then we have a command as “df.show()”. This method “show()” will displays the output that are present in the csv file.**

**🡪Implementing processing json data with pyspark**

**jsonStr = “{“id”:1, “name”: “abcd”}”**

**df= spark.createDataFrame([(1,jsonStr),[“id”,”name”]])**

**Question 2**

**Explain ETL (Extract, Transform, Load) with PySpark** **Using Spark SQL - Creating databases, tables**  
**Using Spark SQL - Transformations such as Filter, Join, Simple Aggregations, GroupBy.**

**Ans**

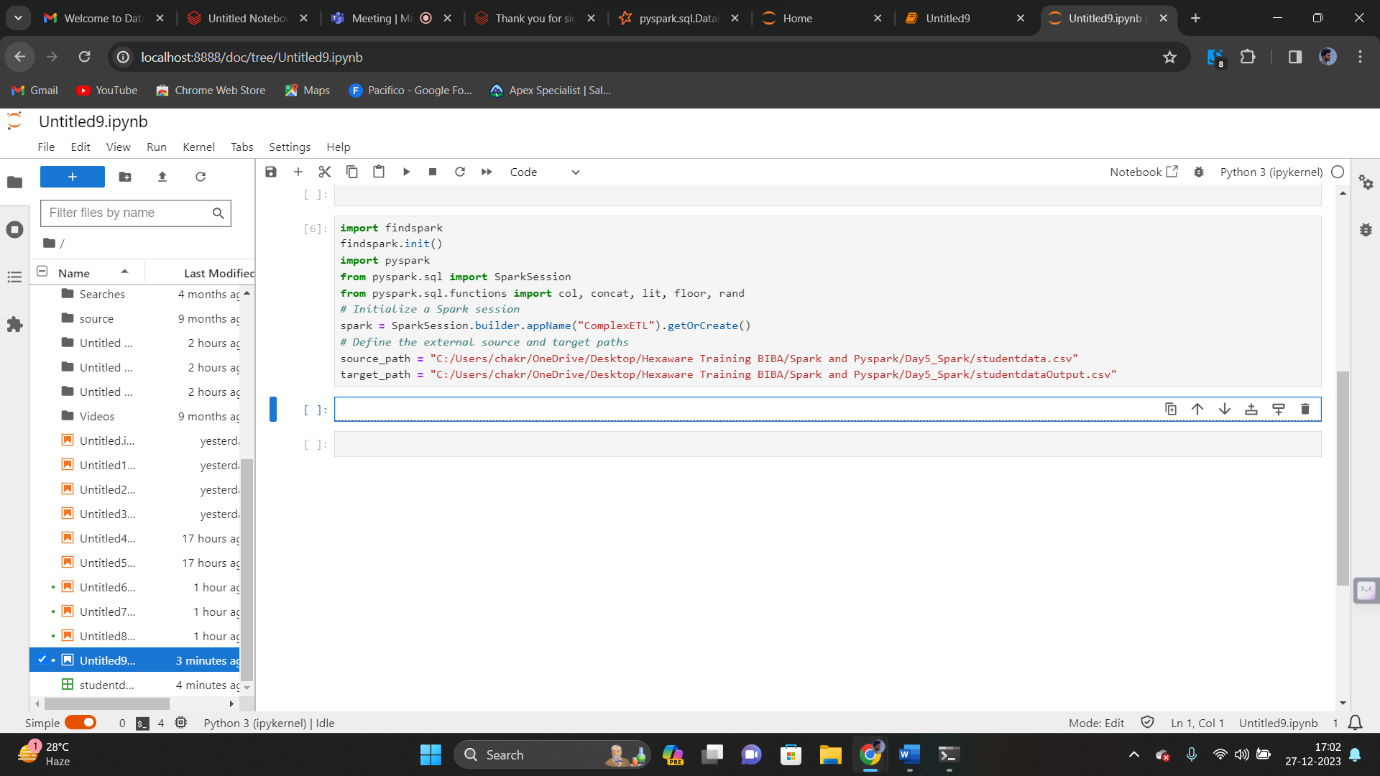
ETL full form is Extract, Transform and Load

Extract : The data will be extracted from various sources like csv files, etc

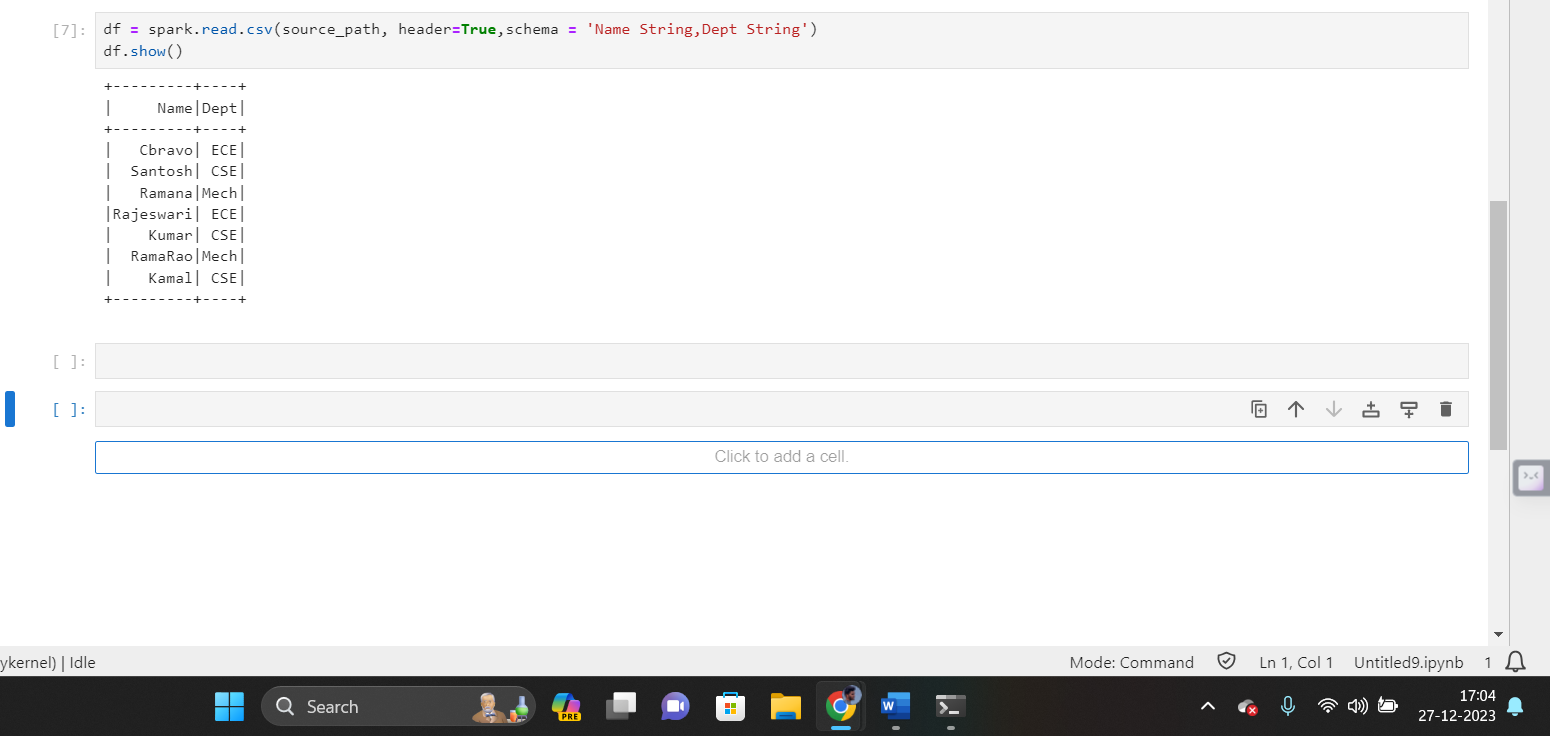
Transform : Transforming data means, to manipulate data like addind data or cleaning data, etc.

Load : After all the operations, the data will be stored in the database.

🡪Create databases and tables

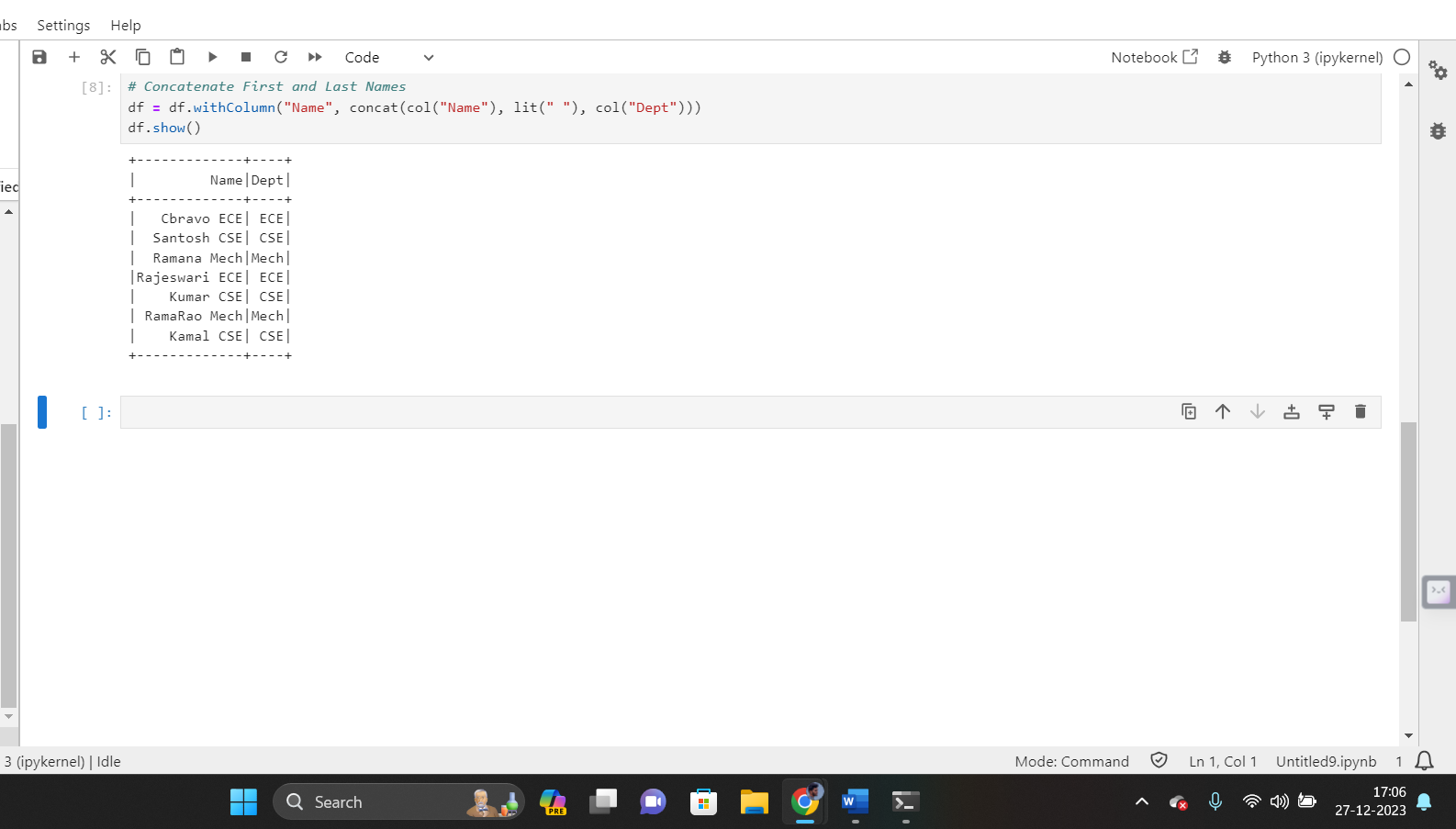


Read data

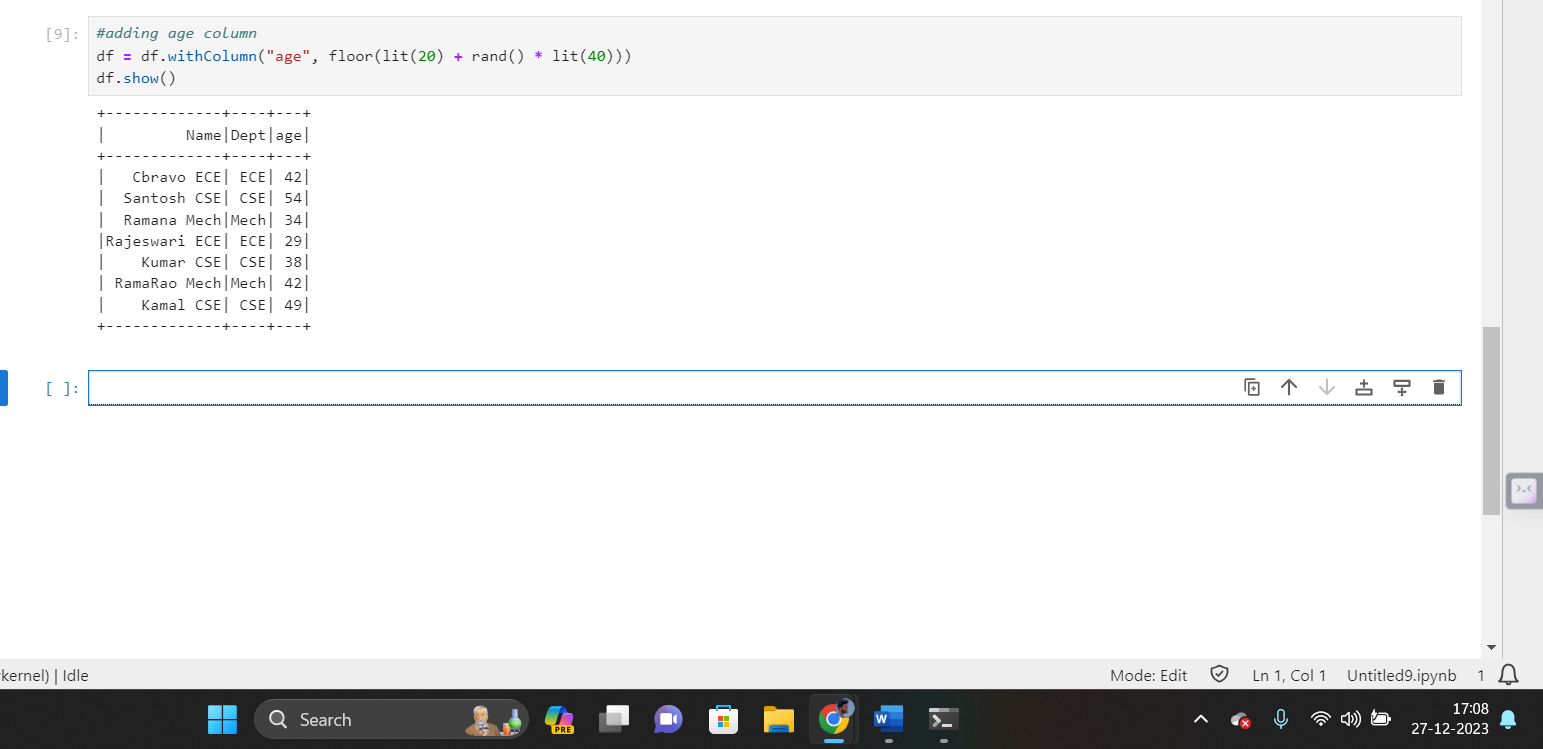


Transformation:  
🡪Concatinating the rows

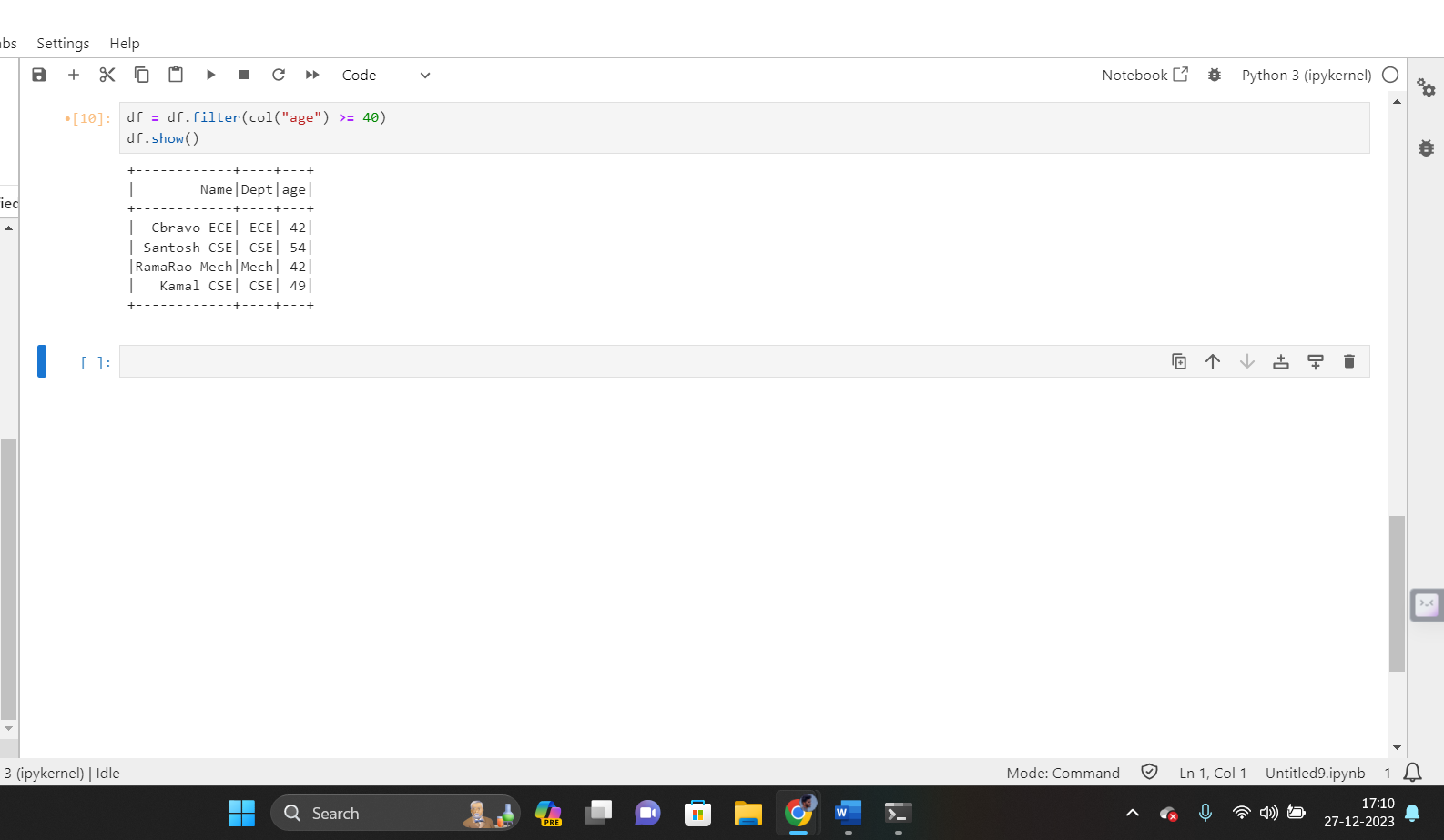
Selected rows will be concatinated to the result table. Lit() is a function that is used to append the fixed value or variable to the output.



🡪Adding age column



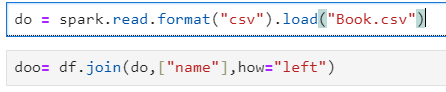
🡪Filtering

Filtering the data what what we want as an output by giving conditions.

🡪Join

Table1= Table1.join(Table2,[column\_names],how=’right join’)

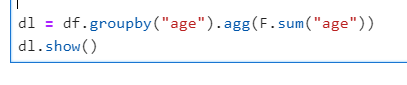
Two tables are left joined from the above command.

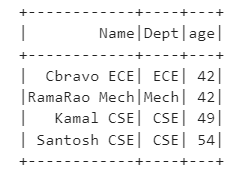


🡪GroupBy

GroupBy() function uses aggregation funcitons with it. It will group data based on the f=given condition.

Syntax : groupBy().agg(sum())





🡪OrderBy

