NAME : AKULA SHARATH CHANDRA
DATE : 12-02-2024
BATCH : DATA ENGINEERING

PY SPARK CODING ASSESSMENT

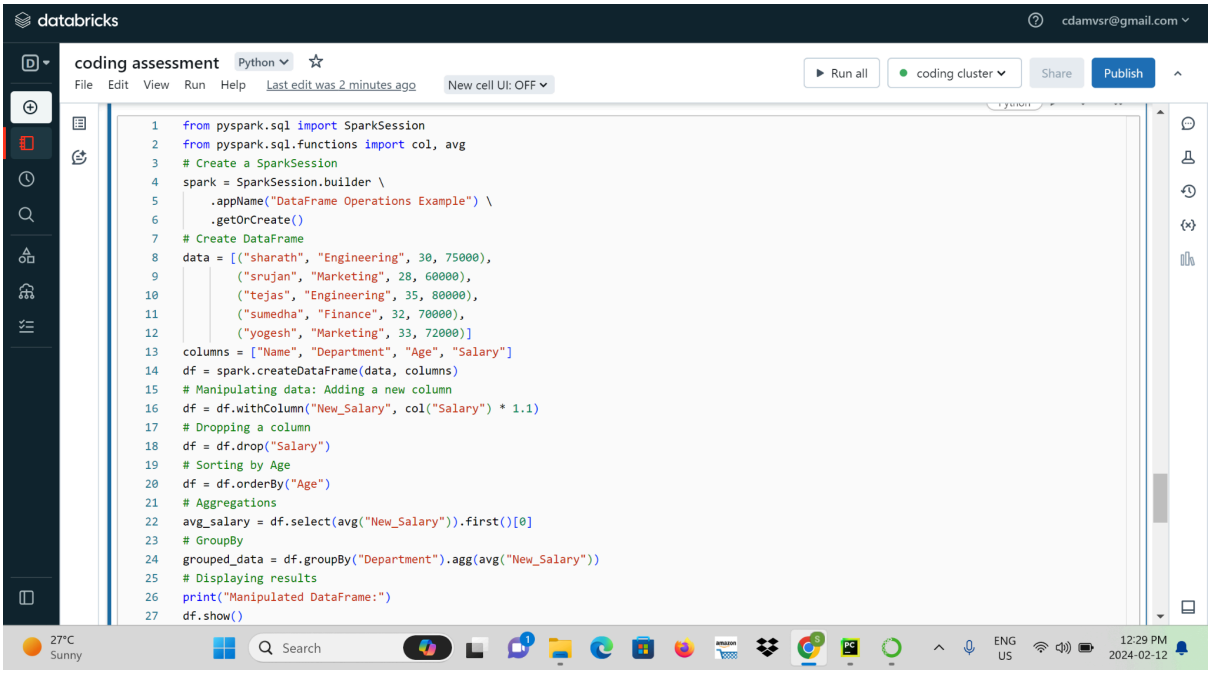
Questions :
1)Execute Manipulating, Dropping, Sorting, Aggregations, Joining, GroupeBy DataFrames
2)Execute Pyspark -sparksql joins & Applying Functions in a Pandas DataFrame

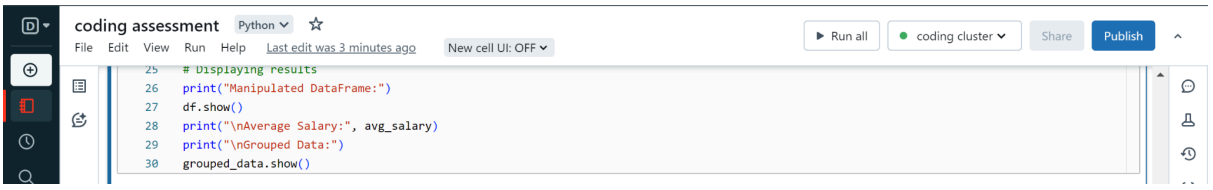
1)Execute Manipulating, Dropping, Sorting, Aggregations, Joining, GroupeBy DataFrames

ANSWER:

DataFrame manipulation in PySpark involves various operations such as filtering, dropping unnecessary columns, sorting, aggregating data, joining multiple DataFrames, and grouping data for analysis. Each operation helps in transforming and preparing the data for analysis or visualisation.

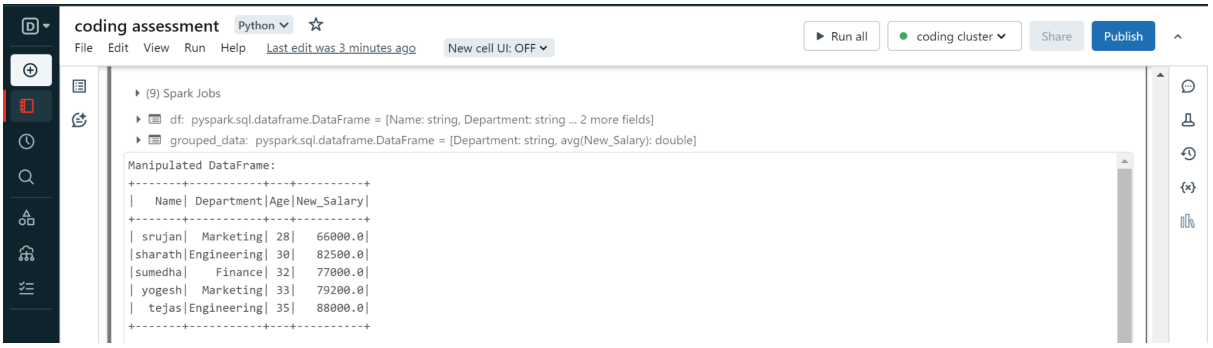
EXAMPLE :





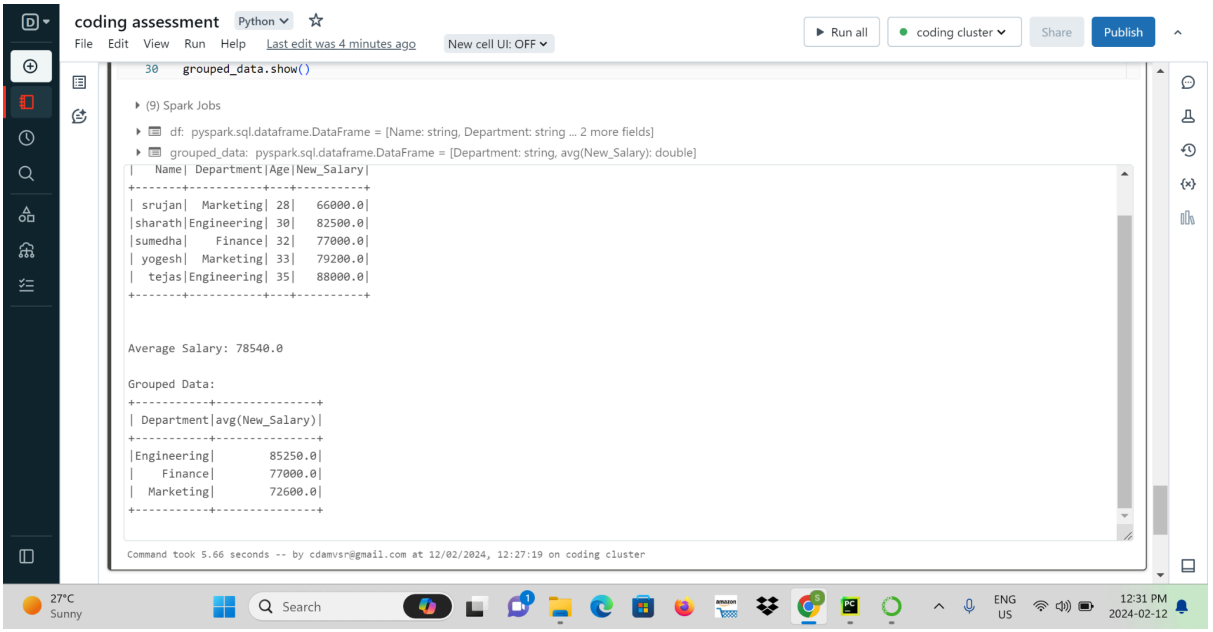
Outputs:

i) manipulated dataframe:



ii)average salary:

iii)grouped data:



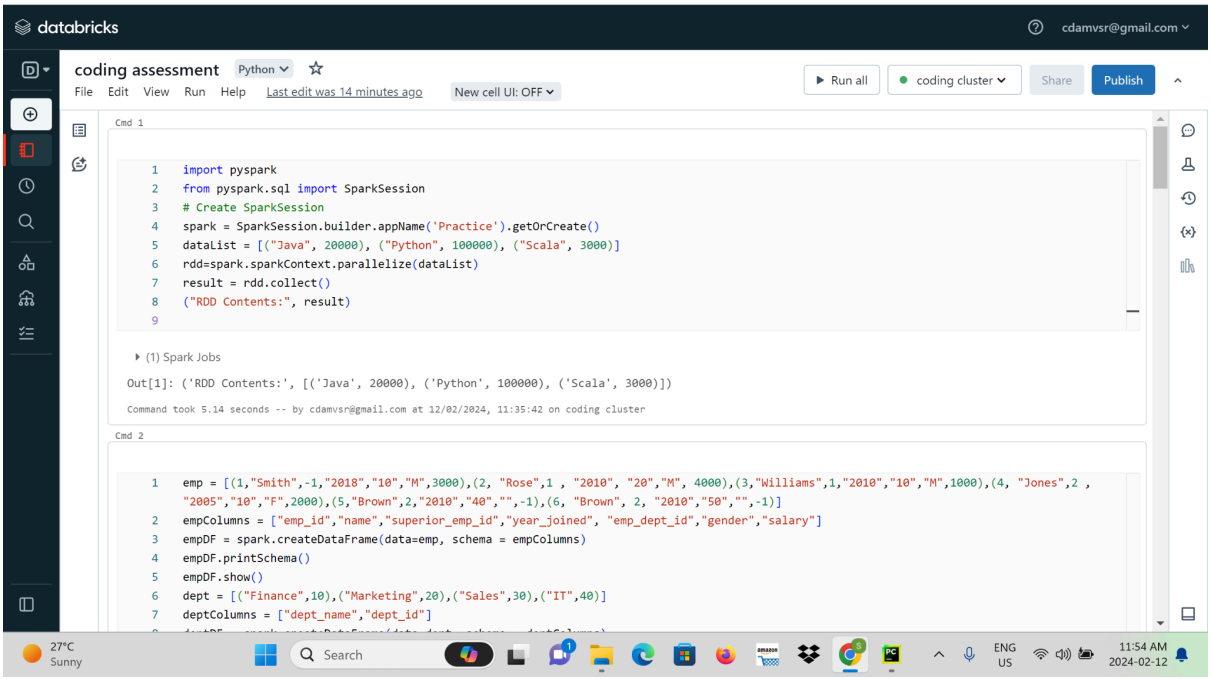
2)Execute Pyspark -sparksql joins & Applying Functions in a Pandas DataFrame.

ANSWER:

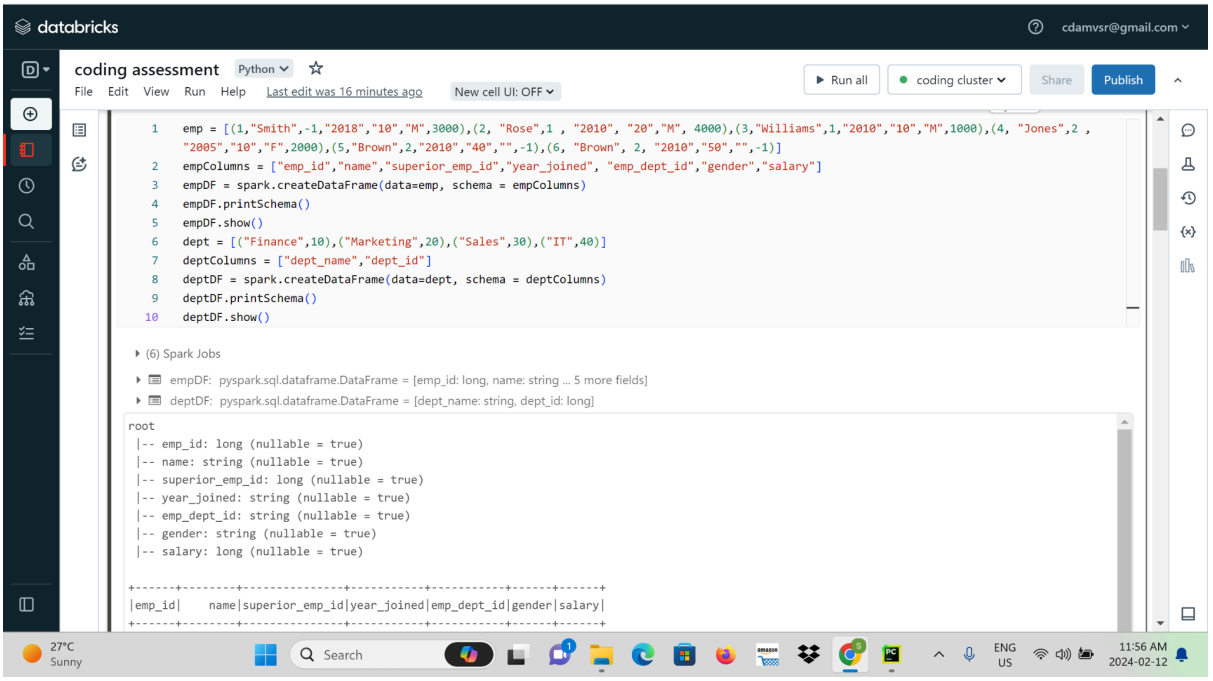
JOINS():PySpark Join is used to combine two DataFrames and by chaining these you can join multiple DataFrames

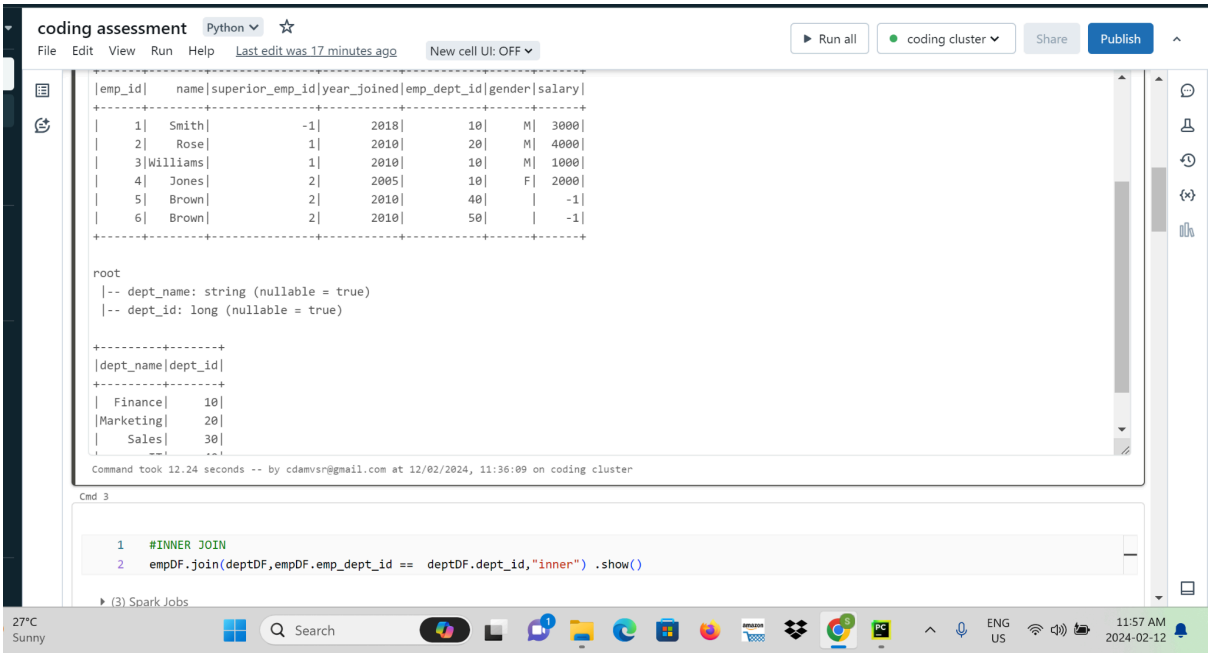
—>**PERFORMING JOINS:**

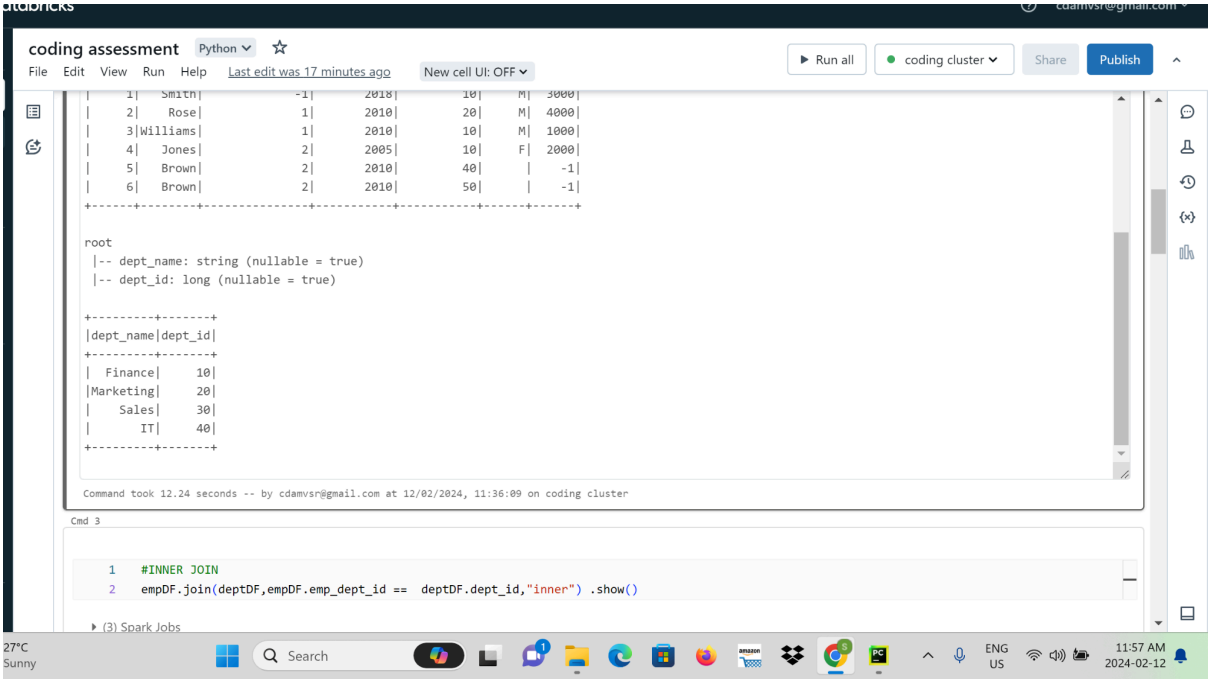
IMPORTING PY SPARK IN DATABRICKS



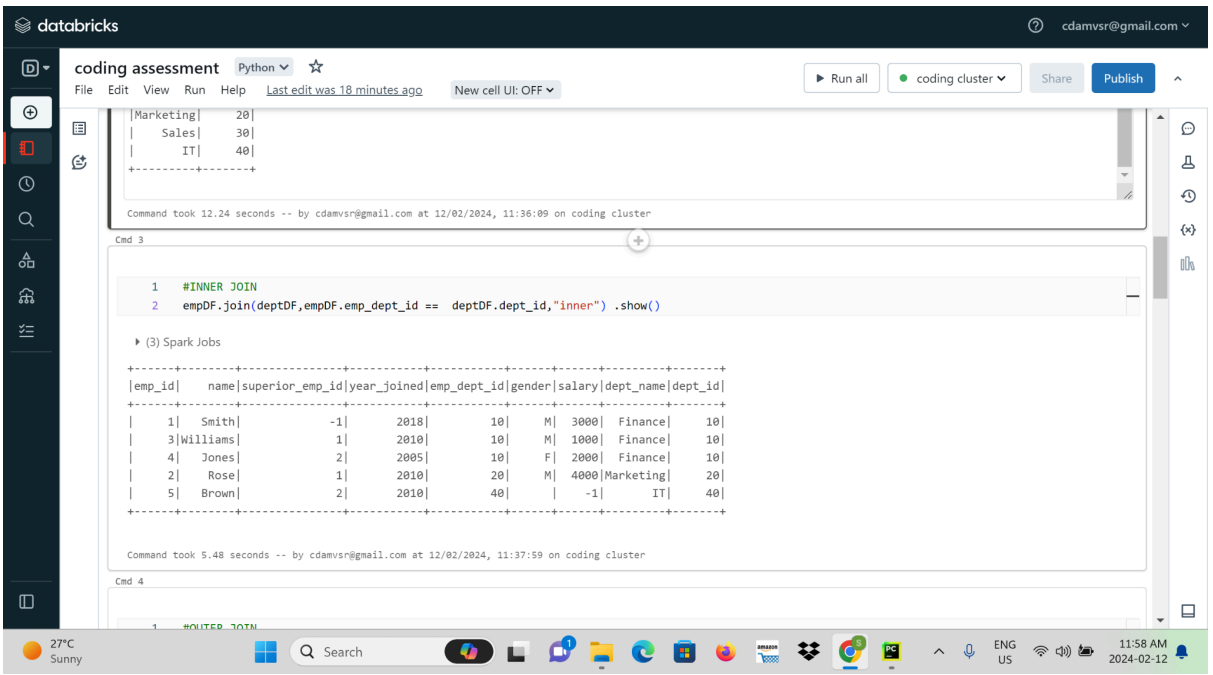
CREATING A DATA FRAME AND DISPLAYING IT :



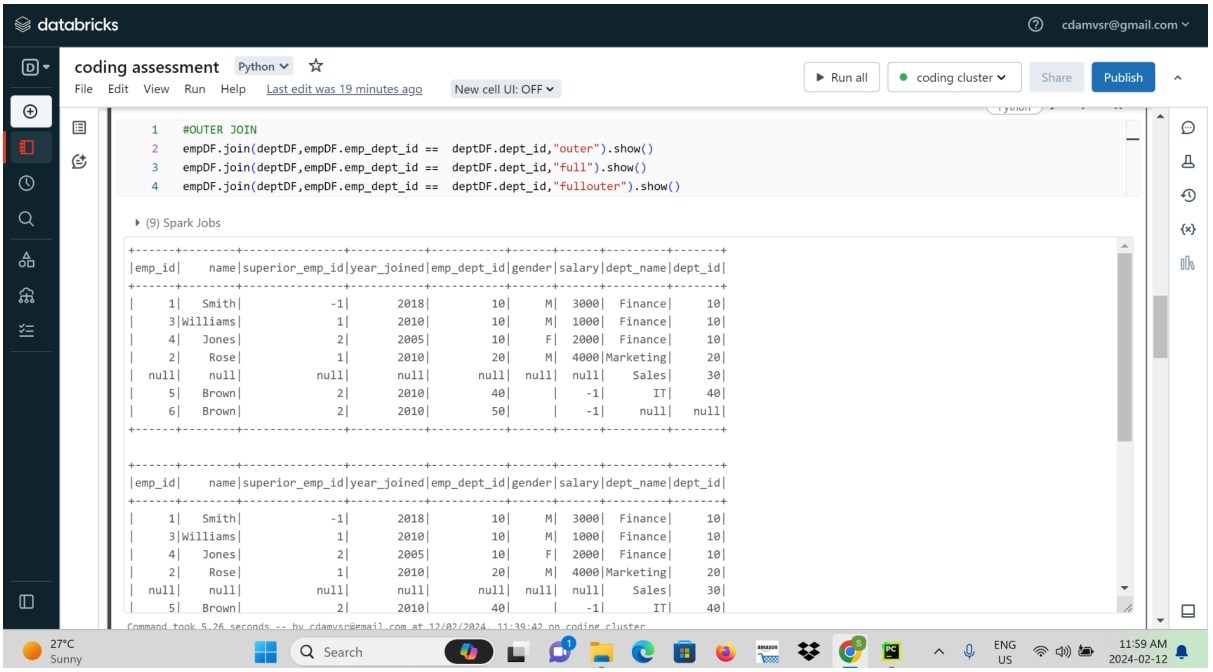


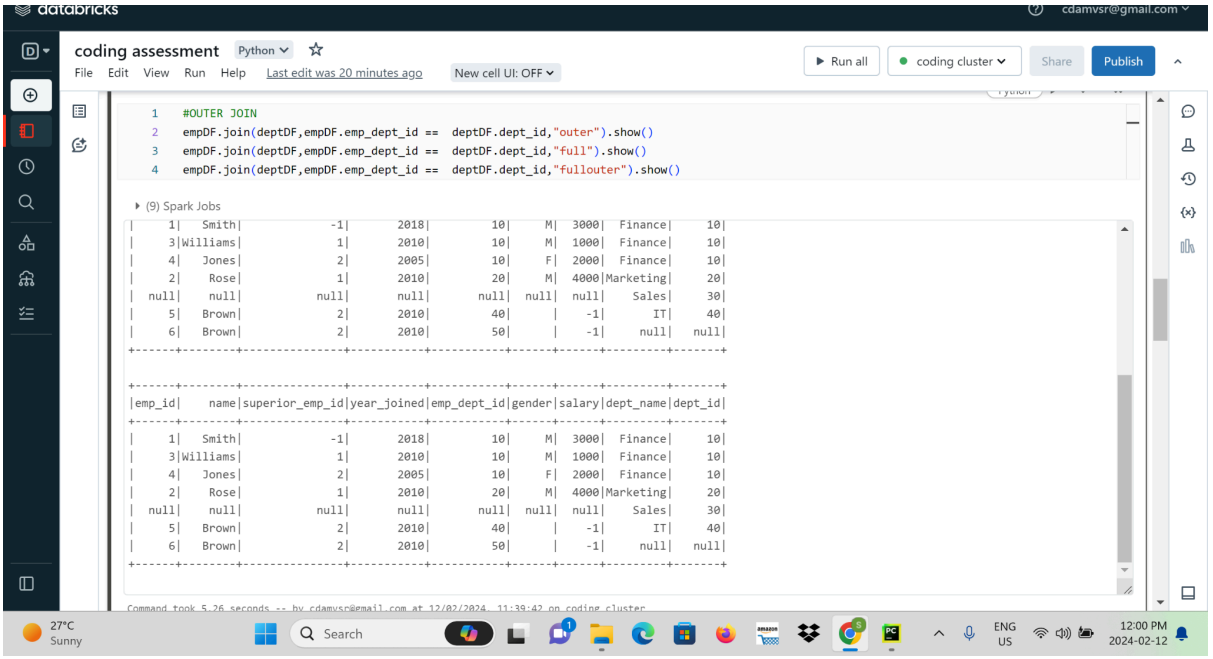


i) **INNER JOIN**:Join records when key columns are matched, and dropped when they are not matched.

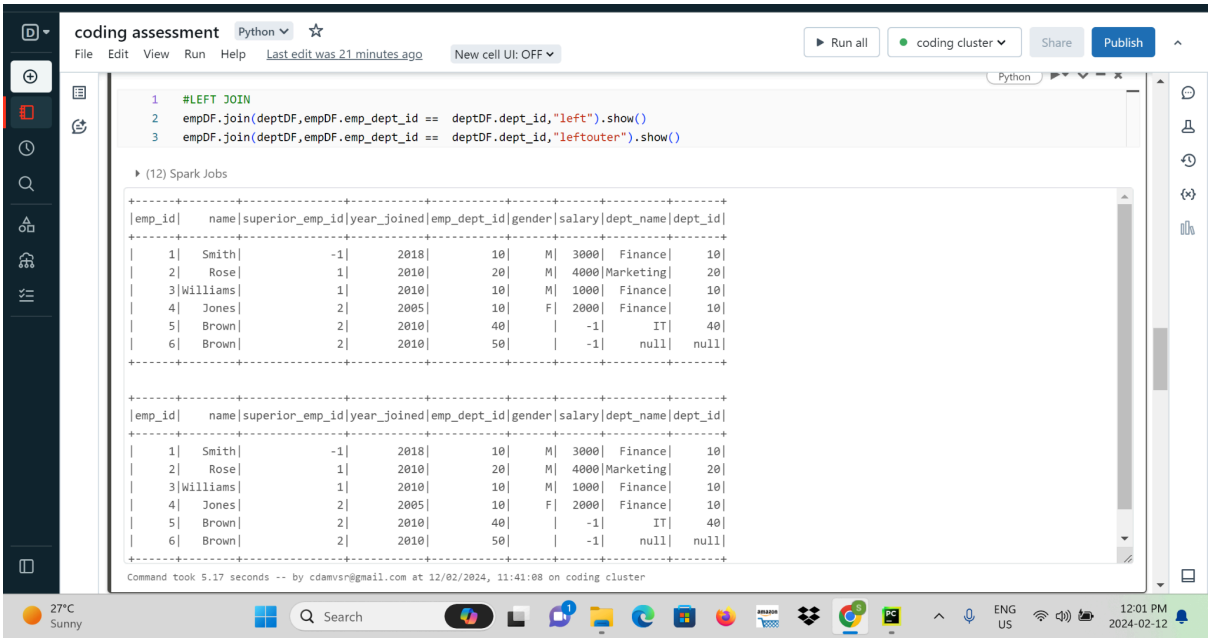


ii)**OUTER JOIN:**Returns all rows from both datasets, where the Join expression doesn't match it returns null or respective columns.

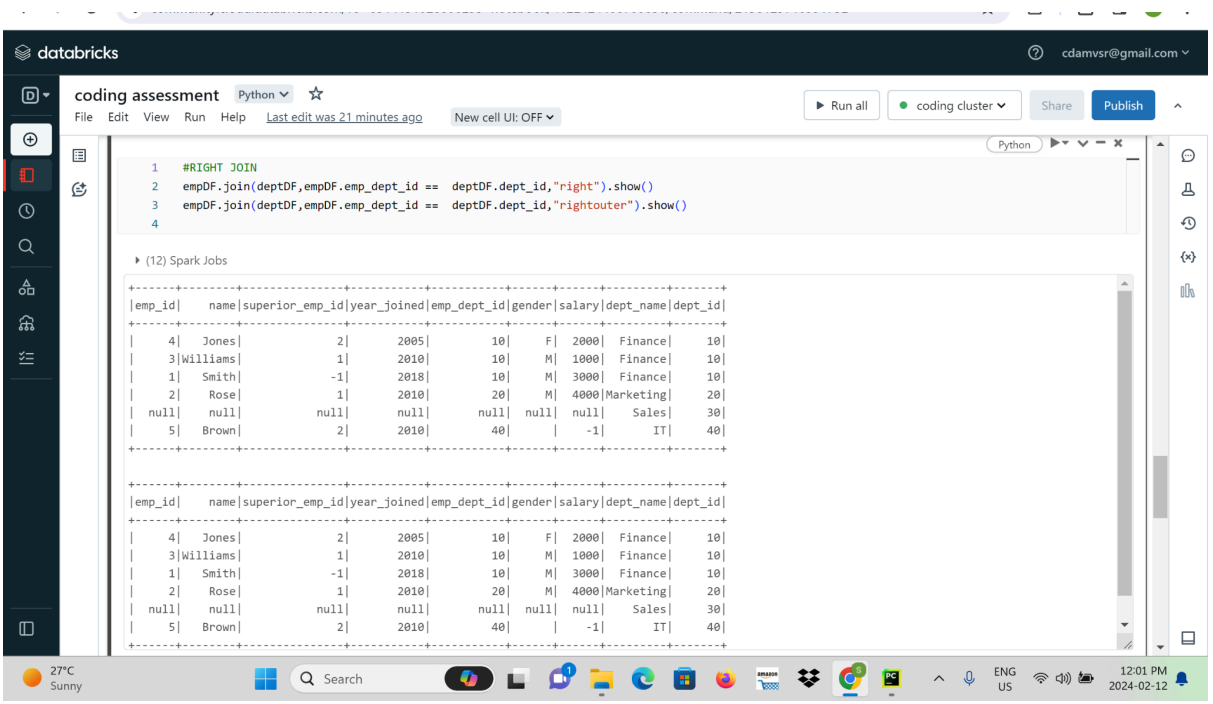




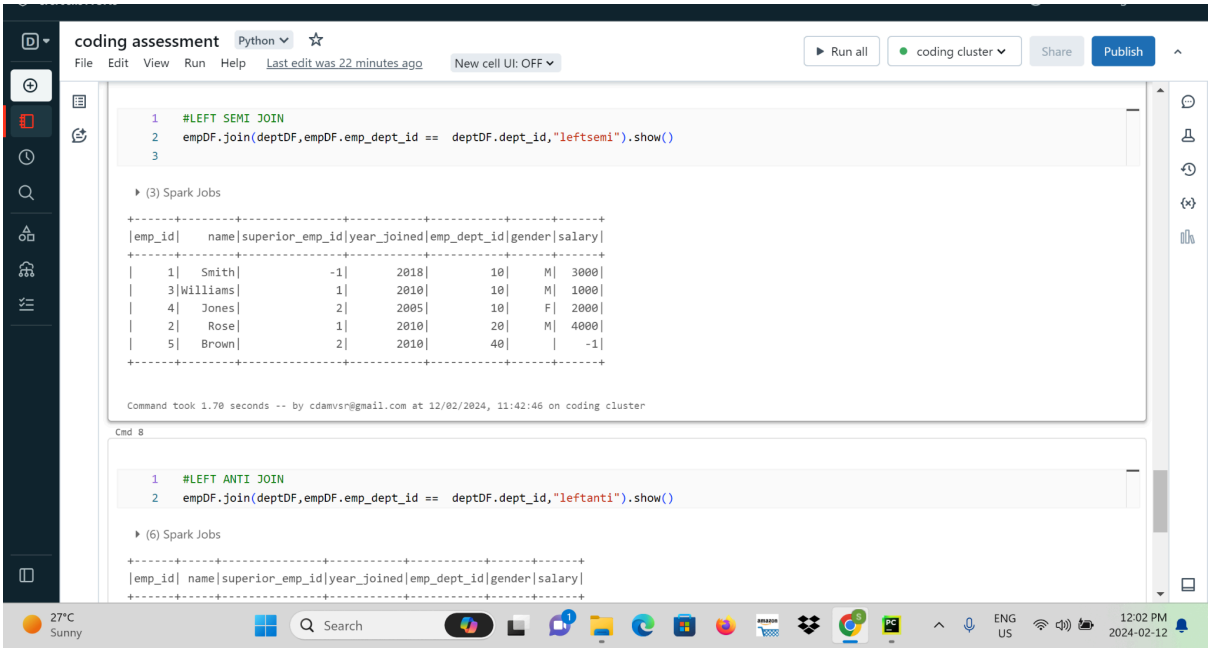
iii)**LEFT JOIN/ LEFT OUTER JOIN:** Returns all rows from left dataset regardless of match found on right dataset, when join doesn't match - it assigns null for that record.



iv)**RIGHT JOIN/ RIGHT OUTER JOIN**: Returns all rows from Right dataset regardless of match found on left dataset,when Join doesn't match - it assigns null for that record.



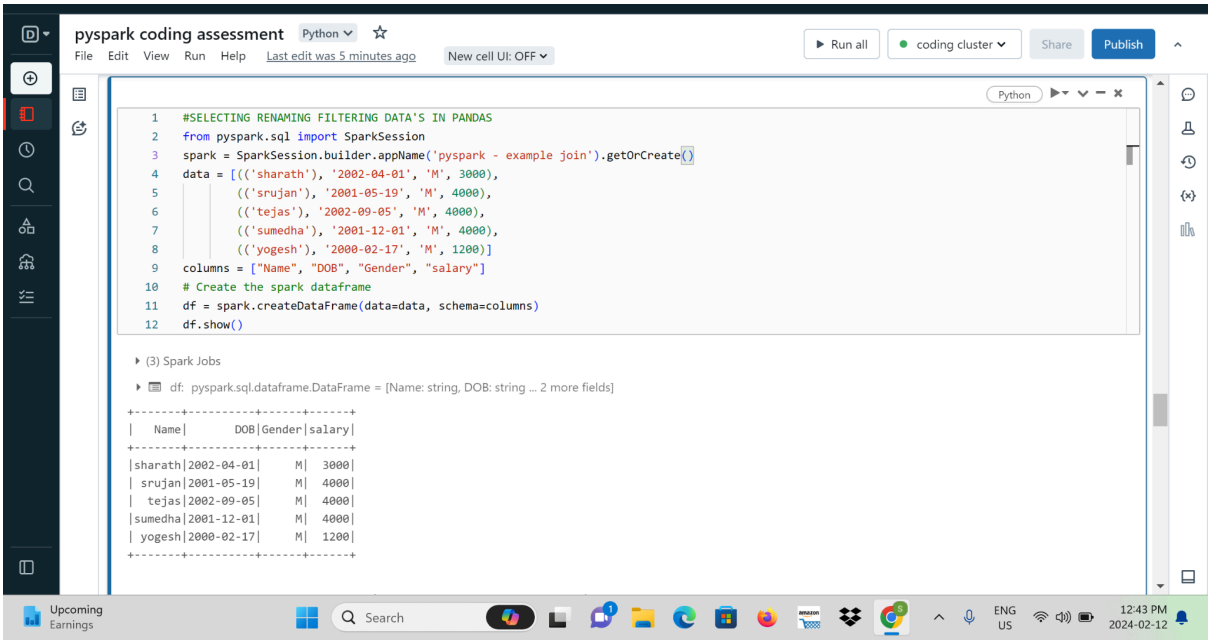
v)**LEFT SEMI JOIN**:Returns columns from the only left dataset for the matched records in the right dataset on join expression.



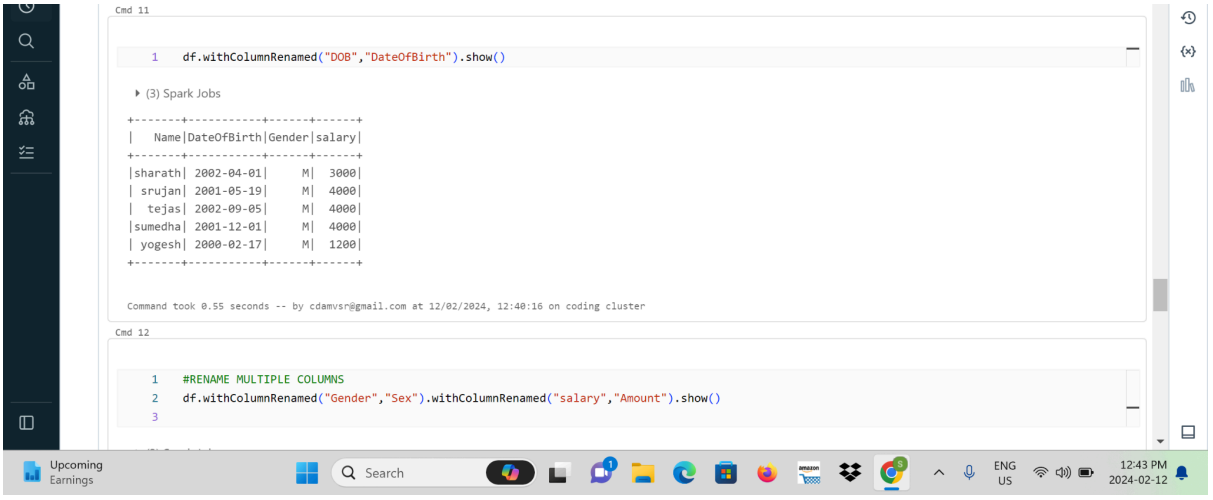
vi)**LEFT ANTI JOIN**: Returns only columns from the left dataset for non-matched records.

—>APPLYING FUNCTIONS IN A PANDAS DATAFRAME:

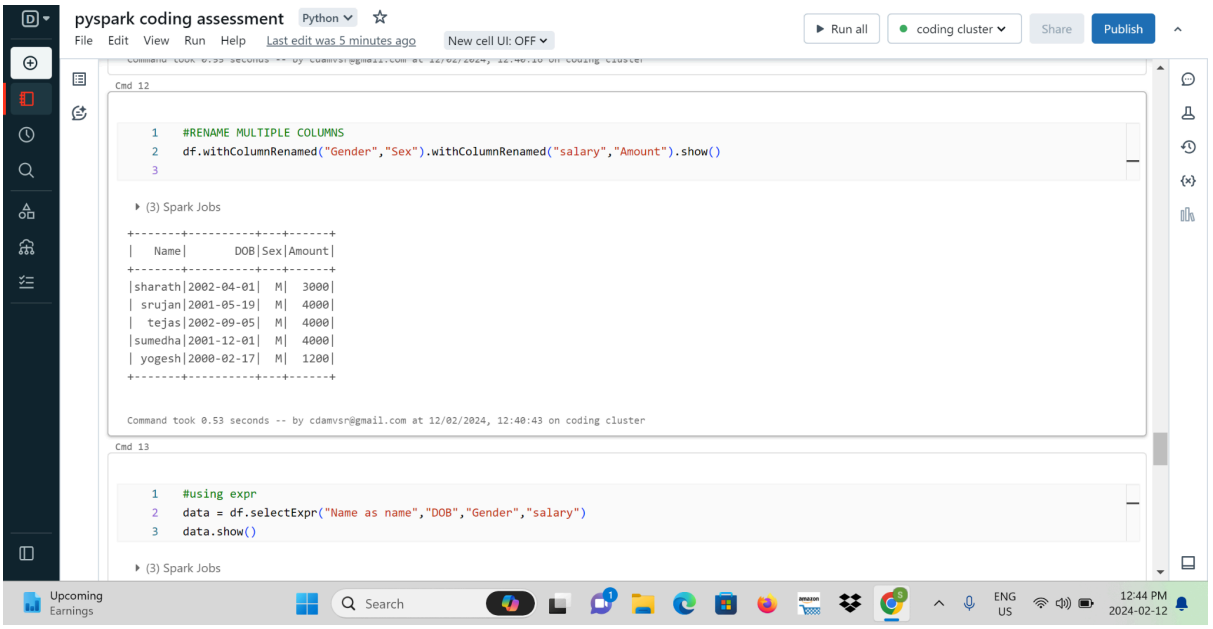
i)selecting renaming filtering data frames :



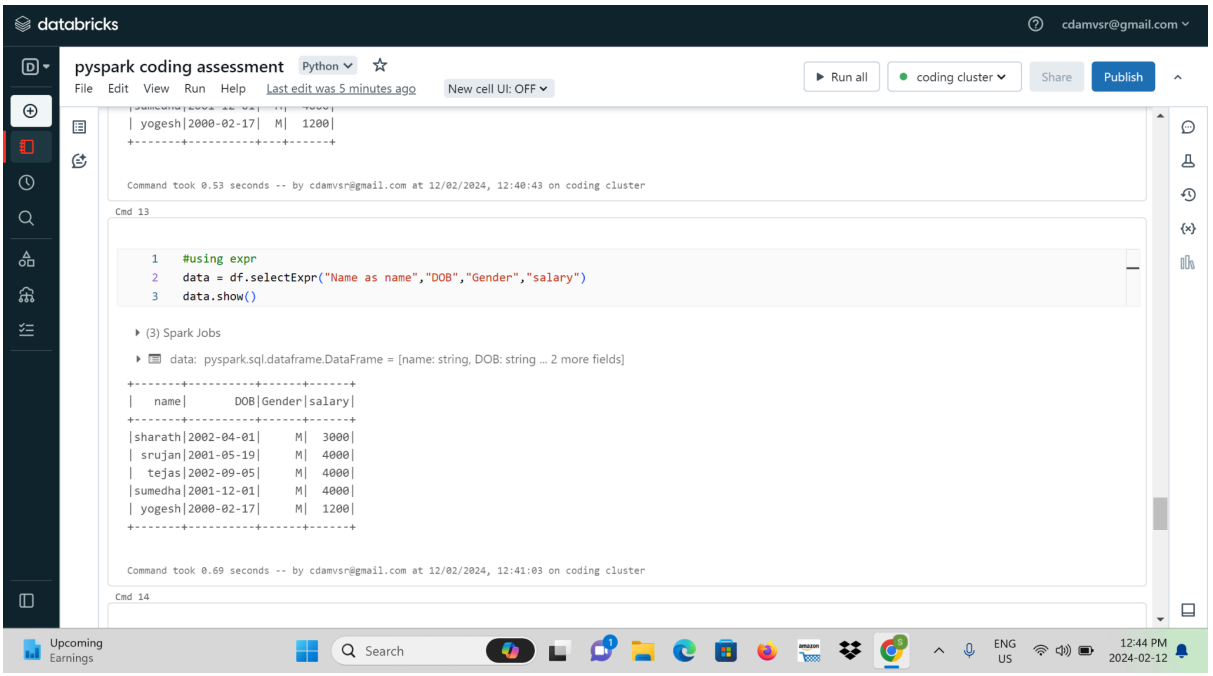
ii)ColumnRename():



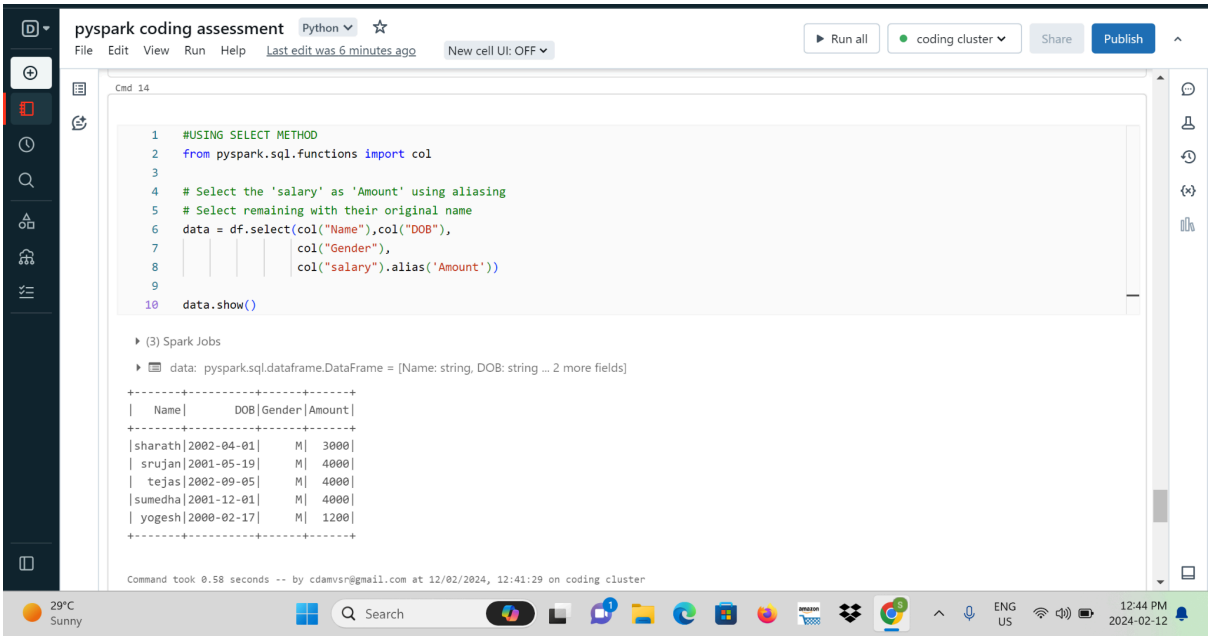
iii)renaming multiple column:



iv)using select expr():



v)using select():



vi)using data frame: