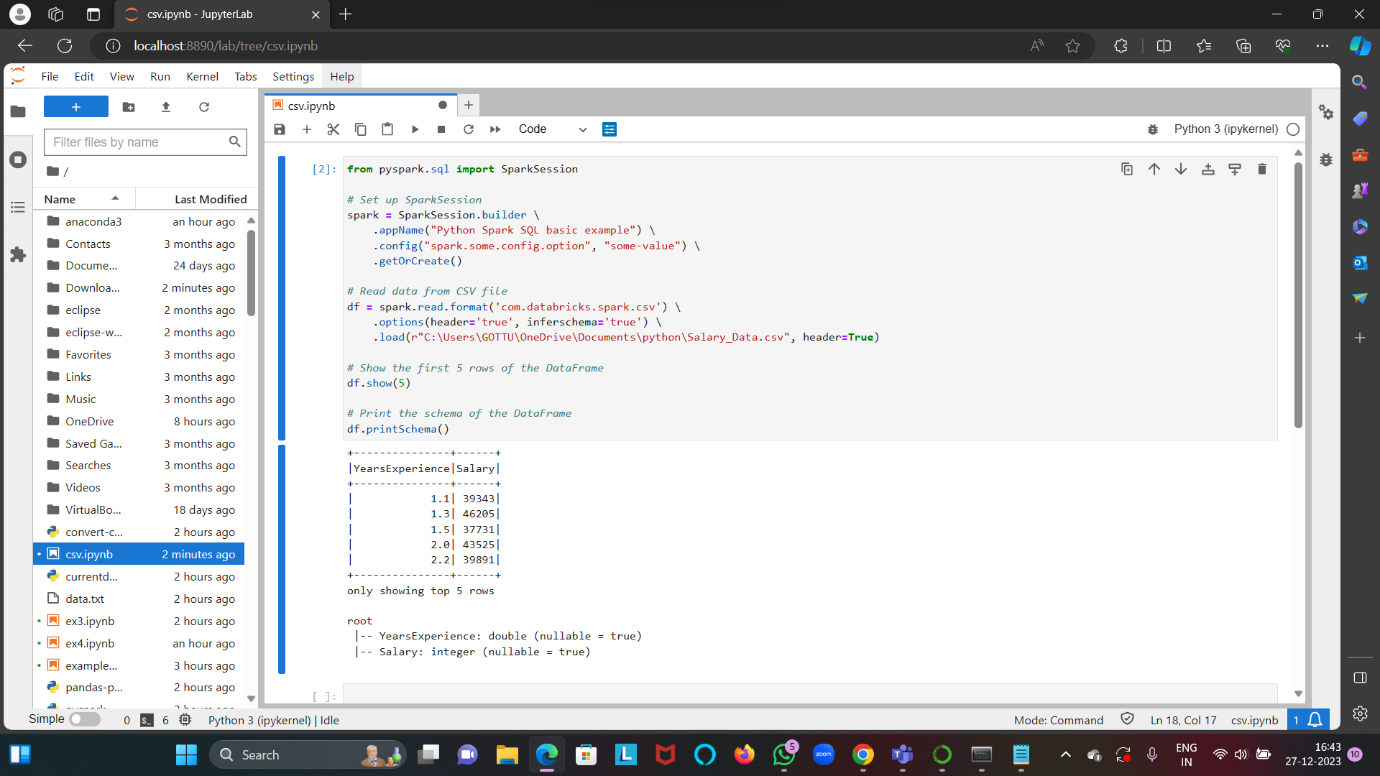
G. Rupa Manasa 27-12-2023

PySpark Coding Assessment

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

Loading the data using csv file and showing the first 5 rows.

****

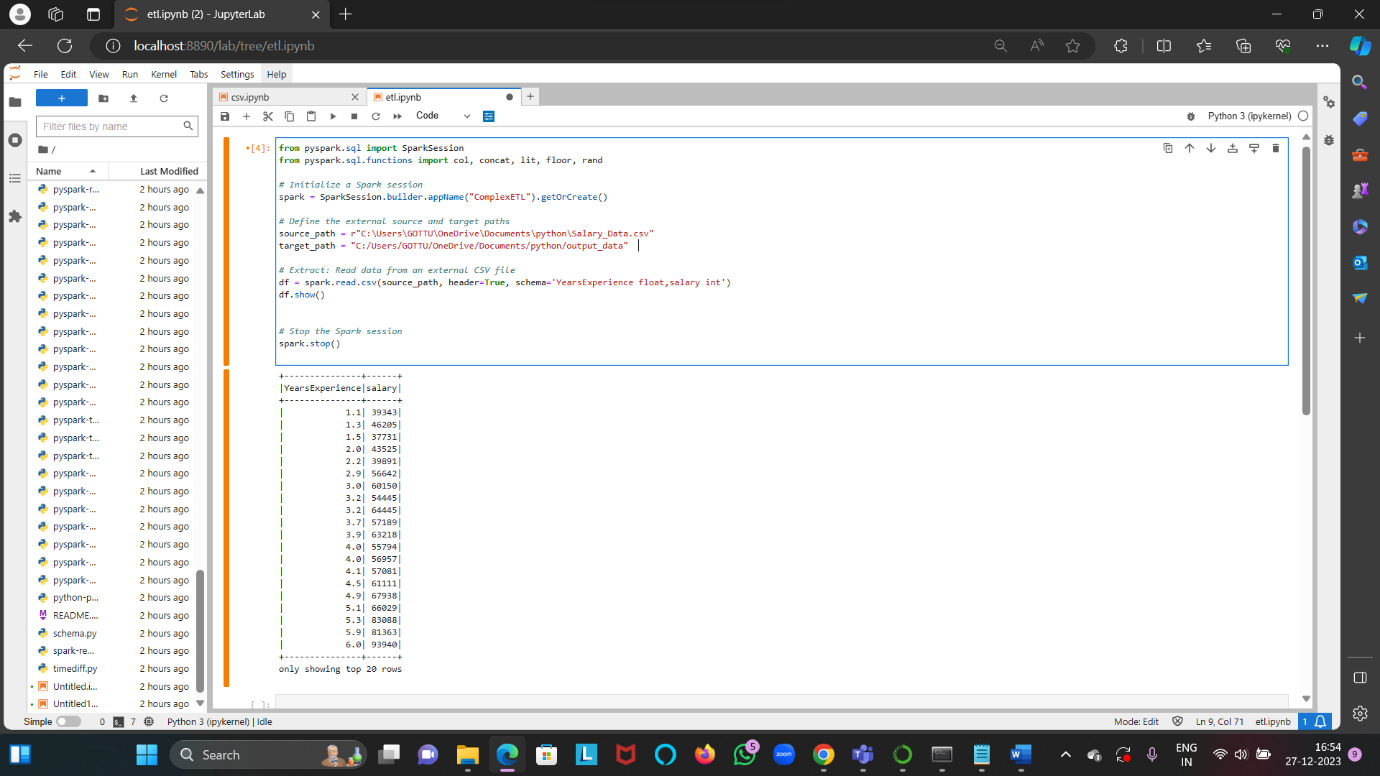
**Explain ETL (Extract, Transform, Load) with PySpark:**

Extract: Retrieve data from various sources like databases, files, or APIs.

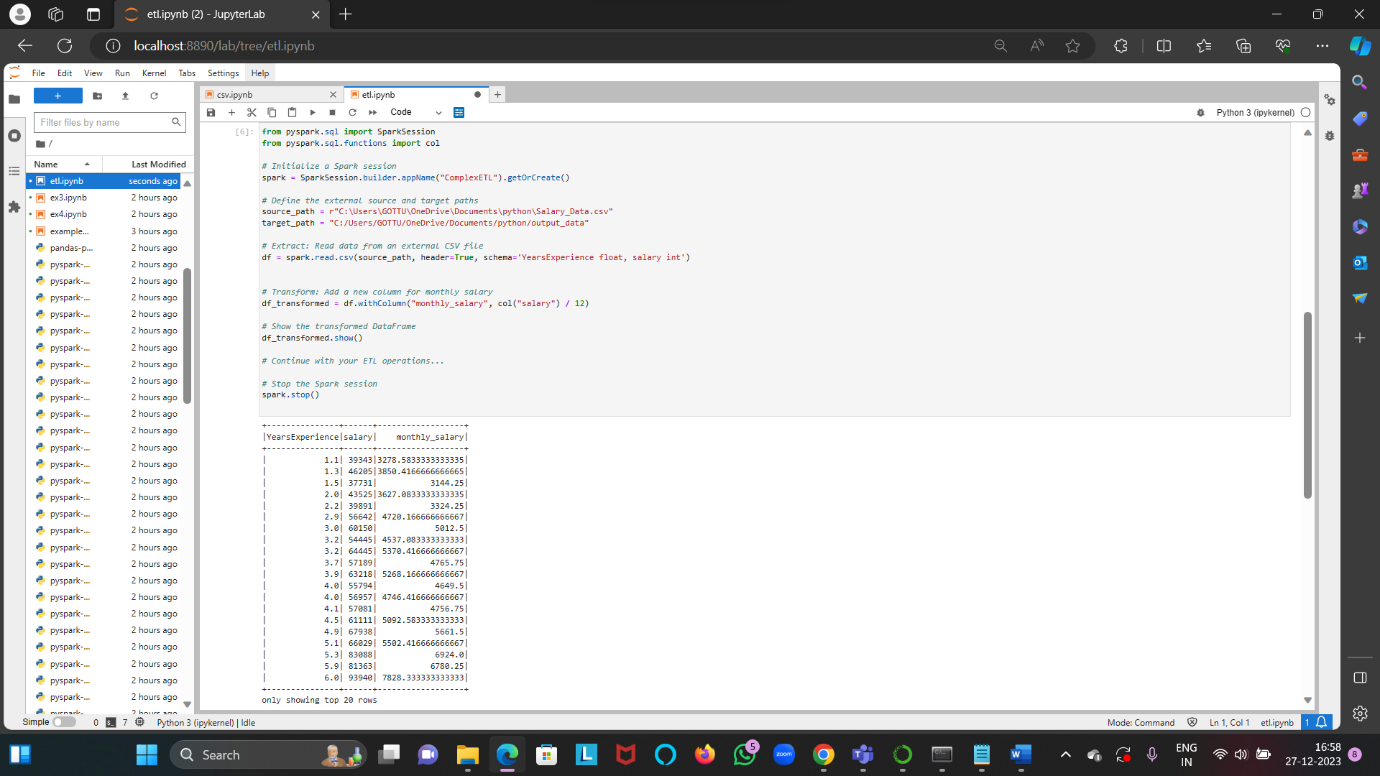
Transform: Clean, aggregate, and manipulate data to fit your analysis needs.

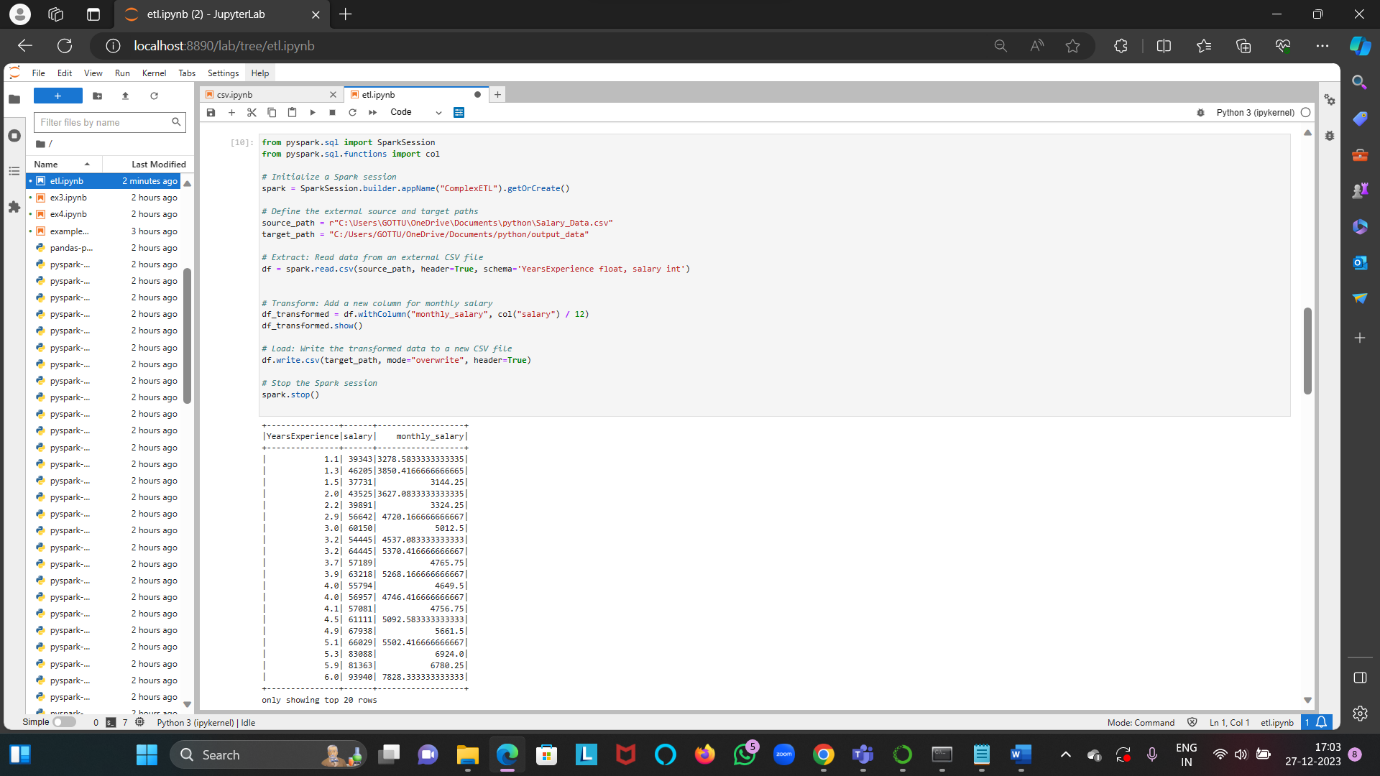
Load: Store the transformed data into a database or data warehouse for analysis.

Extracting the csv file



Transform: adding a new column “monthly\_salary”

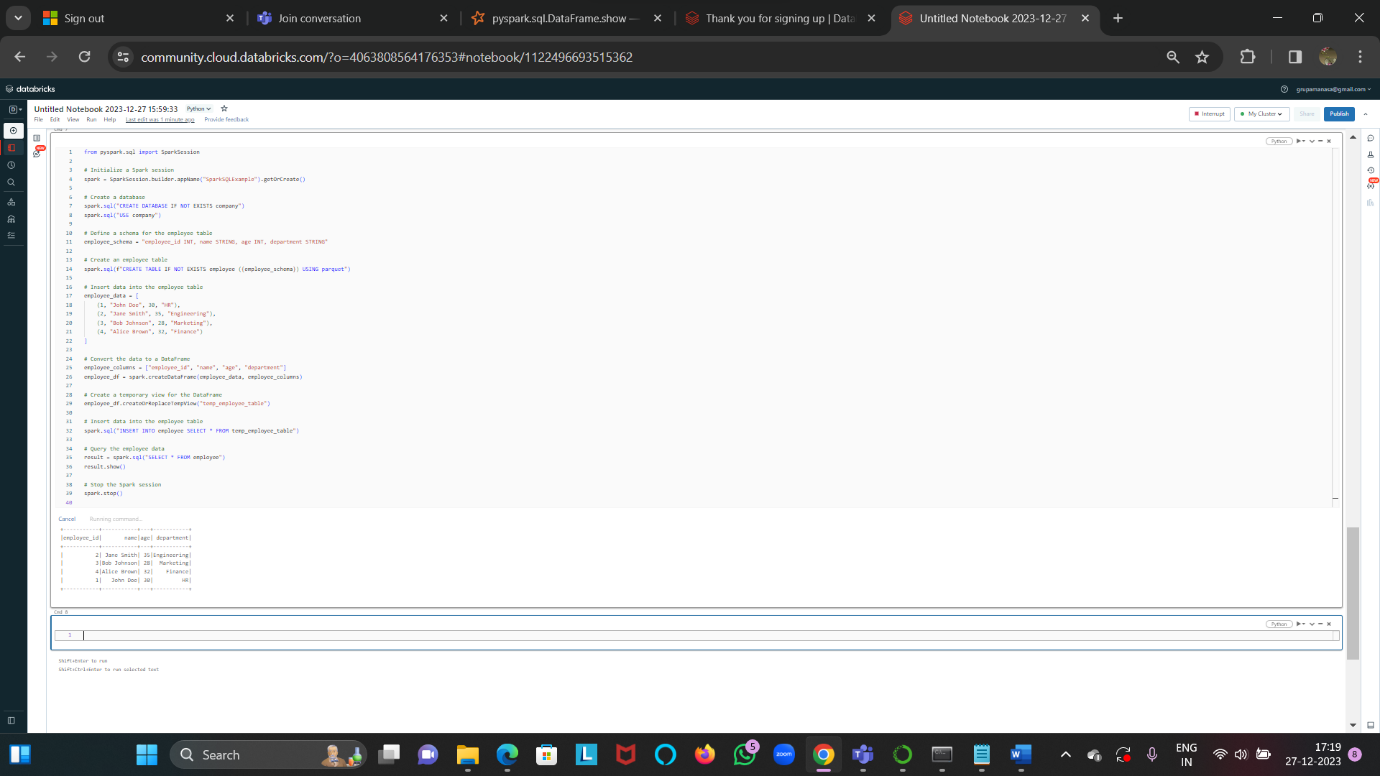


Loading the file into another location.

**Using Spark SQL - Creating databases, tables:**

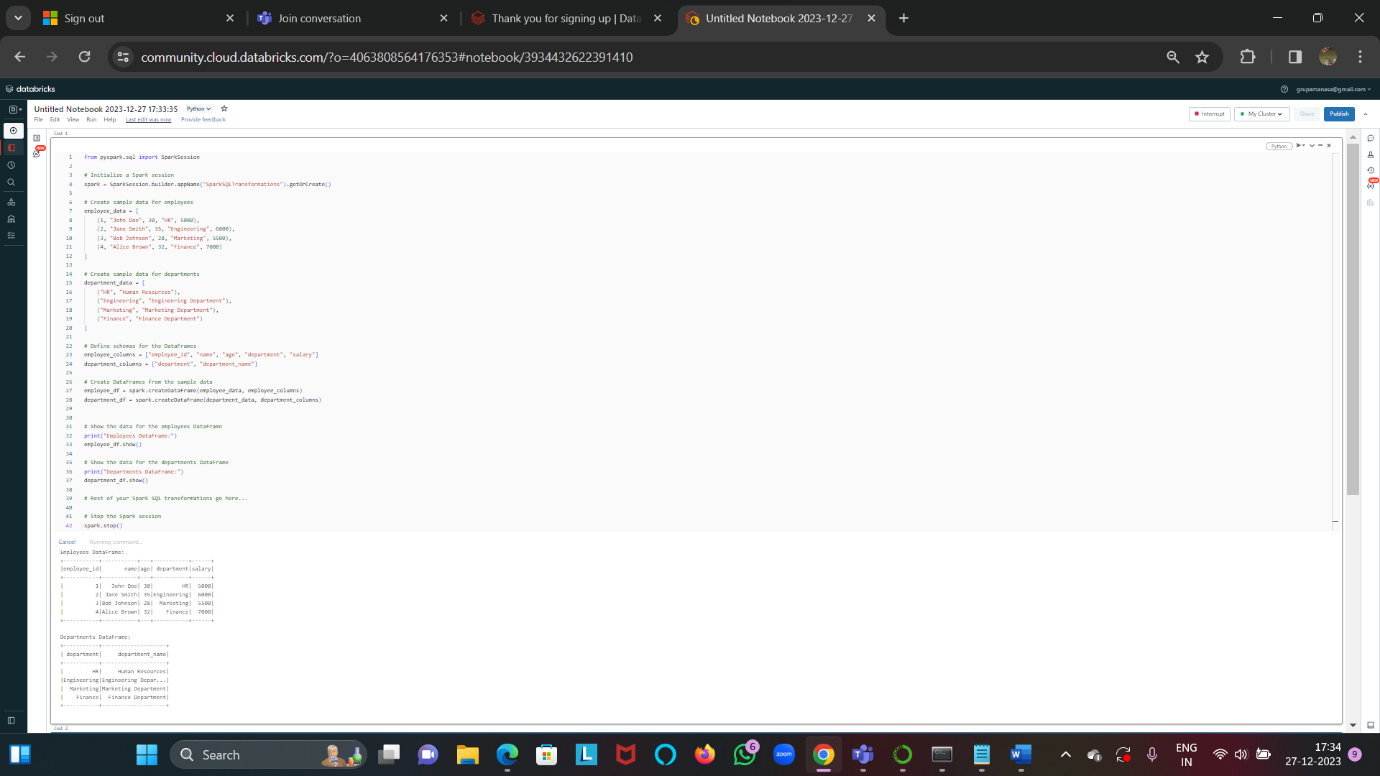
Created a new database called “company” and table called “employee\_data”.

Table contains columns (employee\_id, name, age , department)

****

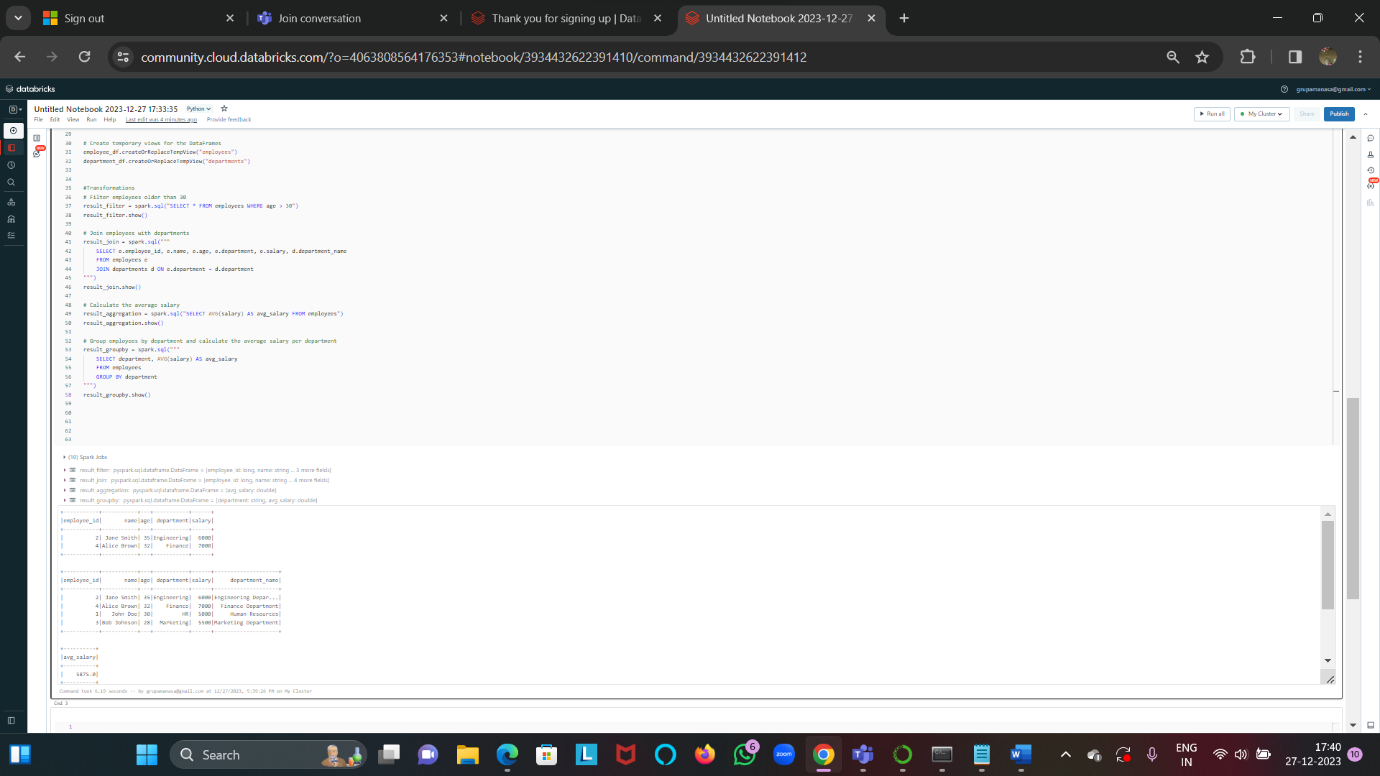
**Using Spark SQL - Transformations such as Filter, Join, Simple Aggregations, GroupBy:**

Created two tables employee\_data and department\_data.

****

**Filter, Join, Simple Aggregations, GroupBy:**

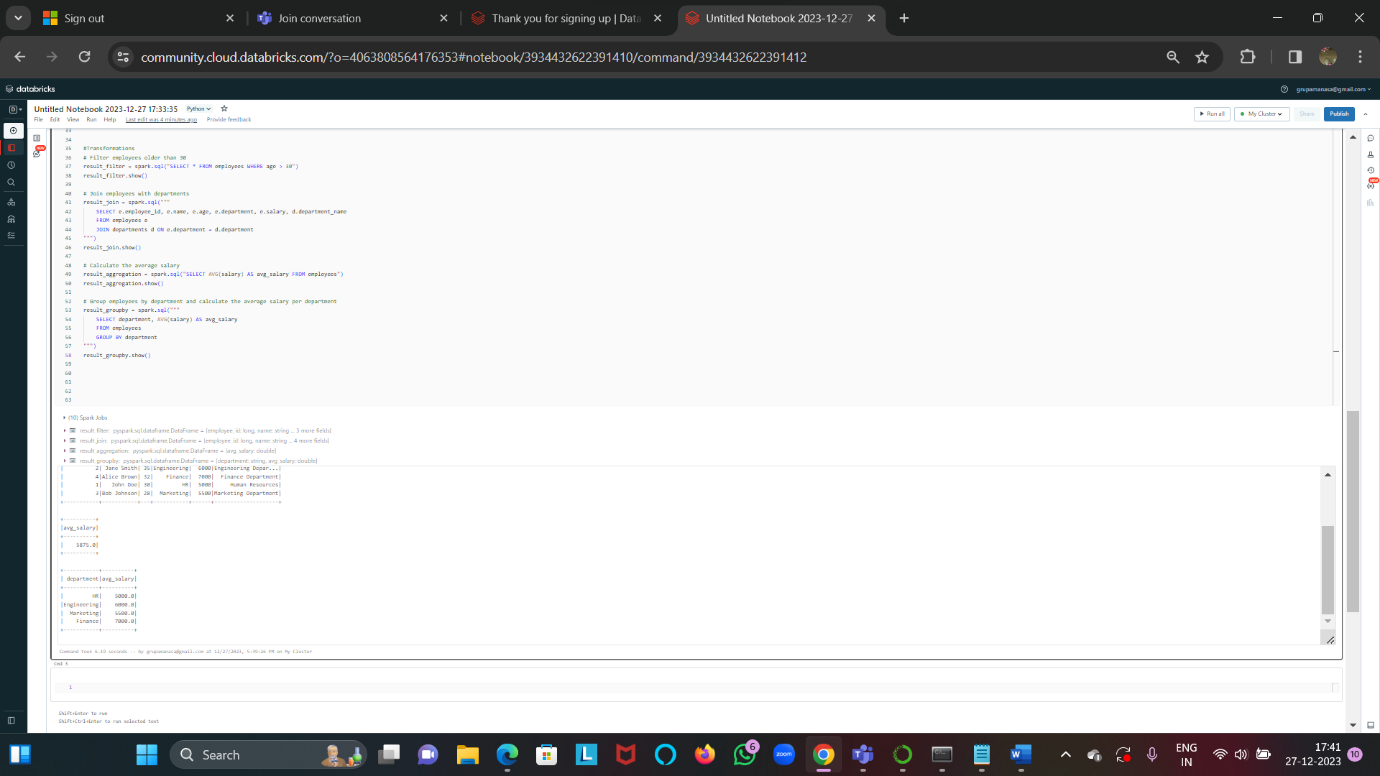
Filtering employees older than 30

****

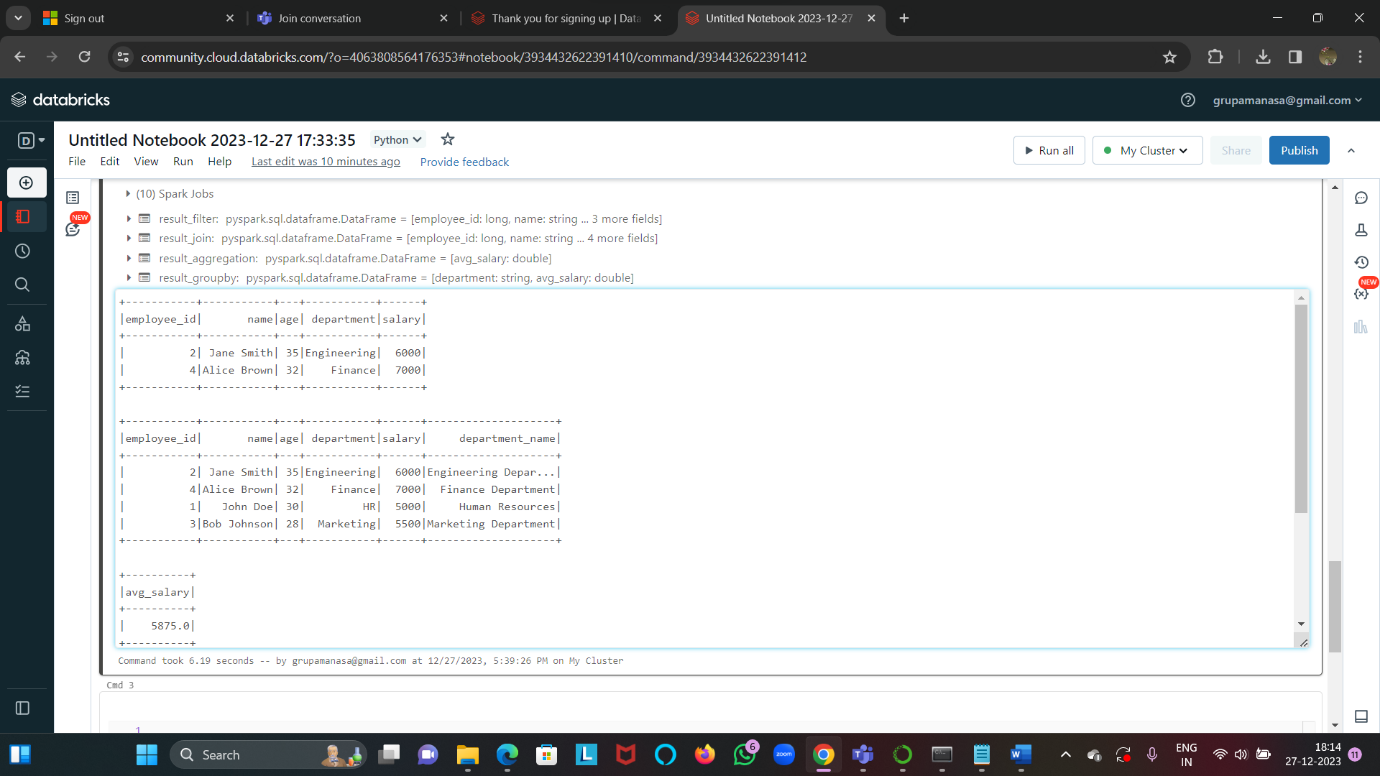
Joining employees with departments

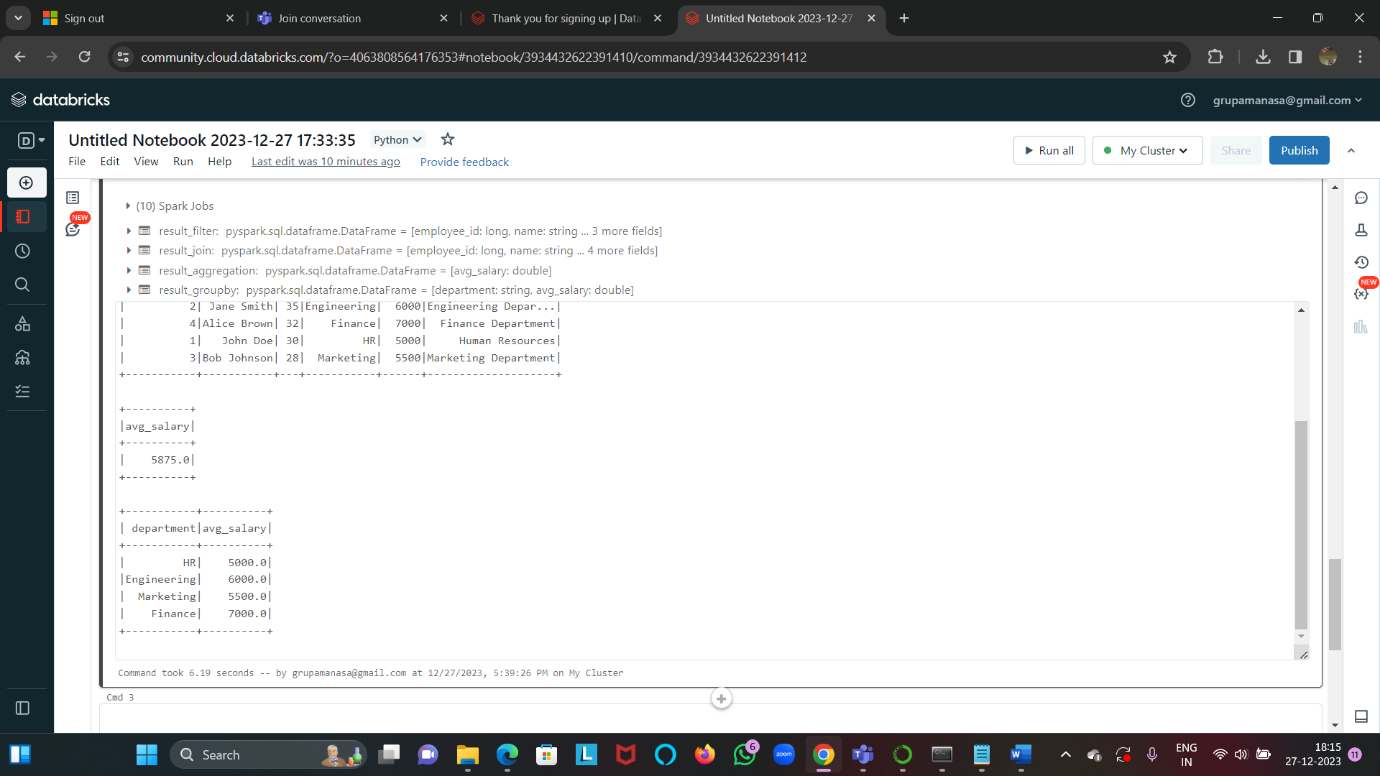
Calculating Average salary(aggregate function)

Group employees by department and calculate the average salary per department(groupby)

****

Outputs of Transformations

****

****