ASSESSMENT - 19 | PySpark - 04

*“PYSPARK ETL COMMANDS”*



Submitted By

VAIBHAV PATIDAR

IPS ACADEMY, INDORE (M.P.)

**Date :** 26-12-2023 | **Day 19**

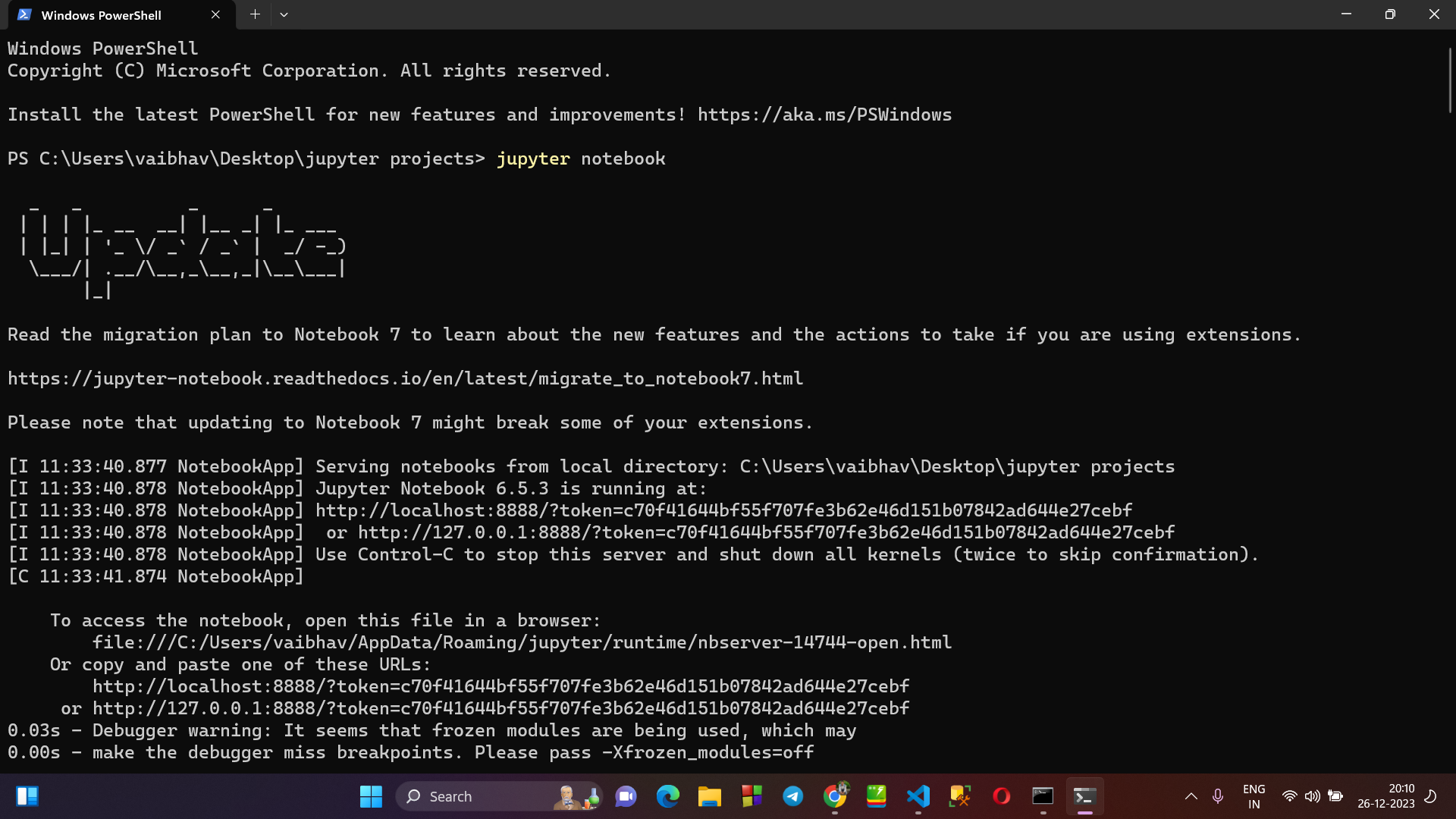
**Week 4** | **Day - 01**

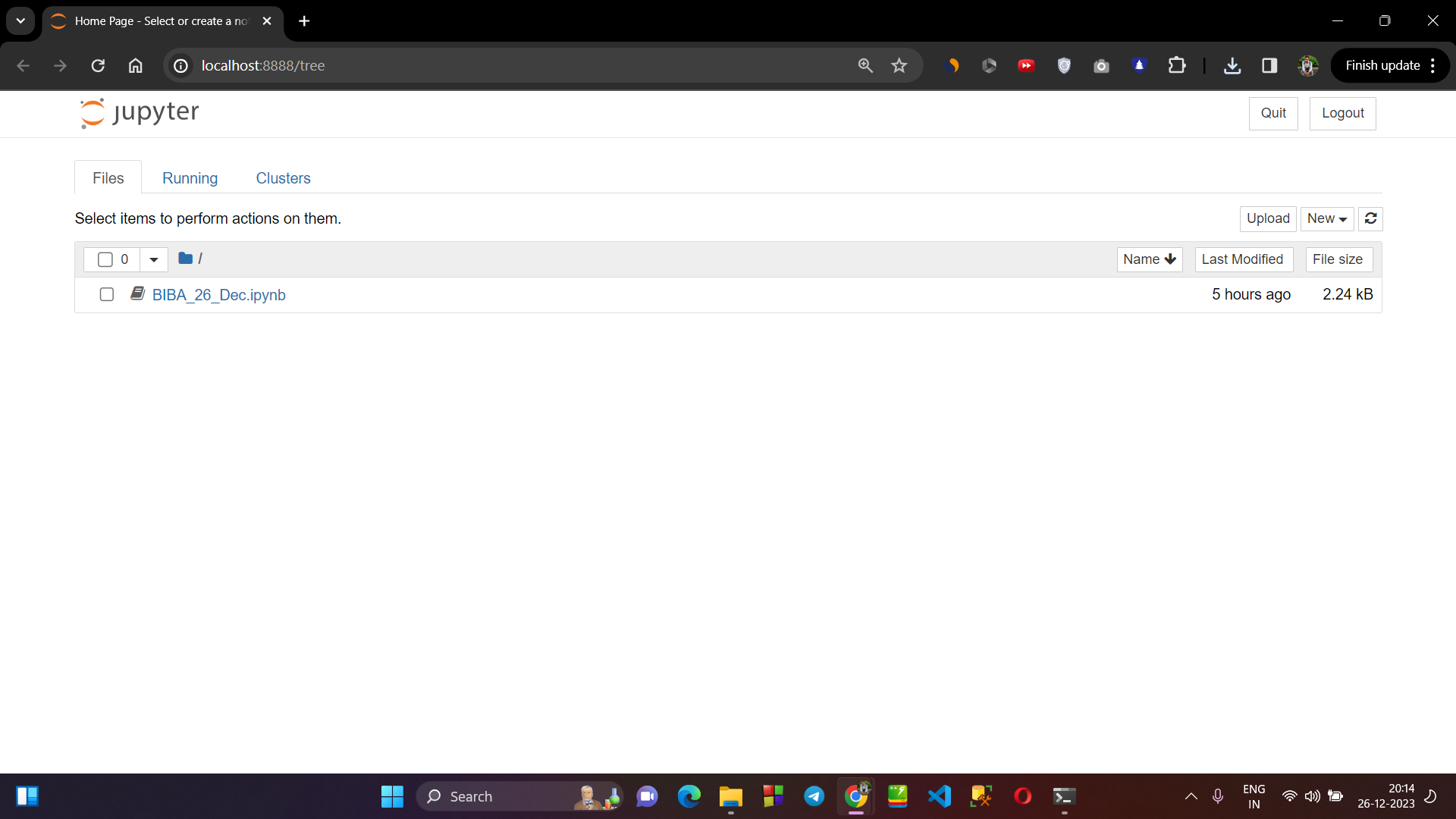
**→ What is Jupyter Notebook :**

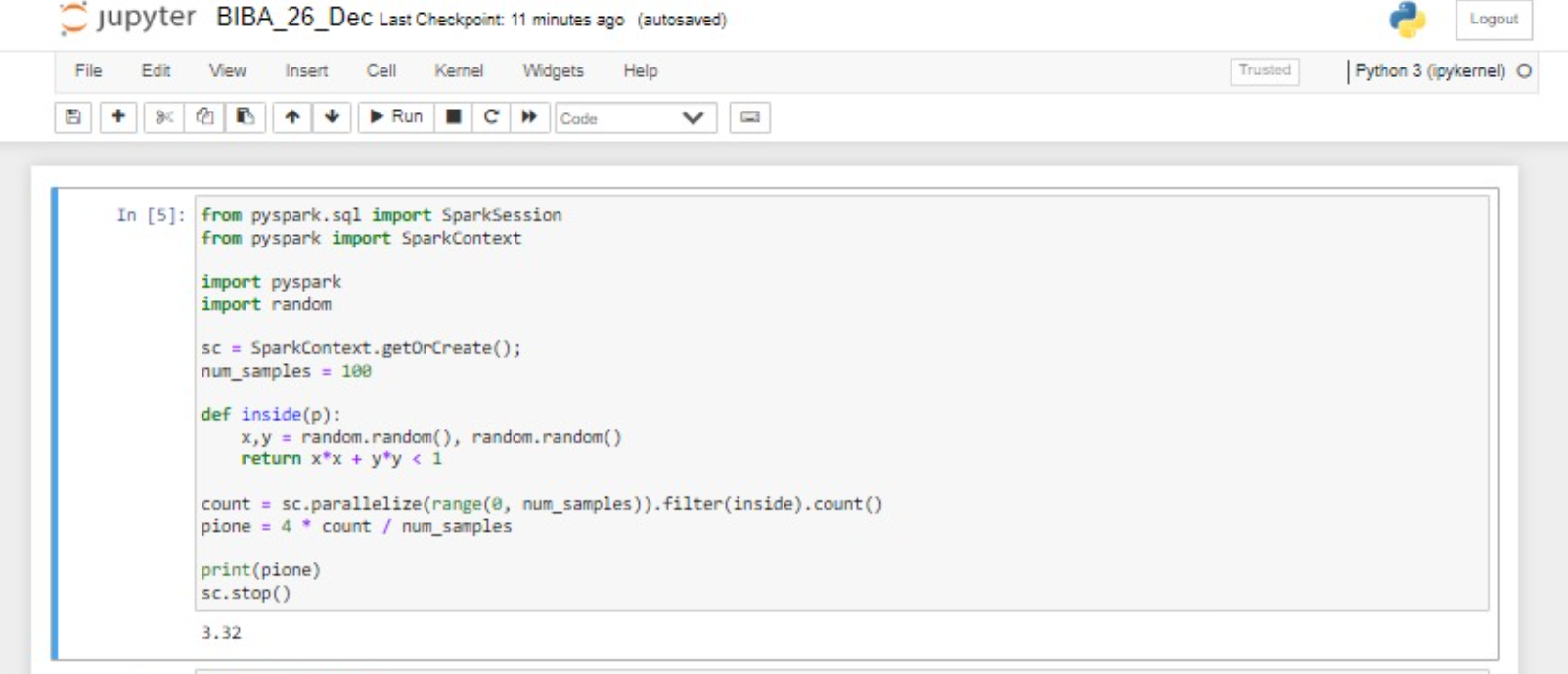
* Jupyter Notebook is a user-friendly tool for interactive computing.
* Combines code, text, and visualizations.
* Great for data analysis, research, and learning programming.
* Supports various programming languages.

**→ Downloading Jupyter Notebook :**

1. First of all, make sure we have python and spark installed in our system.
2. Then open command prompt, type → pip install jupyter → press enter.
3. Now, installation will take 2-3 minutes.
4. Now create a folder and open it , right click and click on “**open in terminal**”
5. Now , in terminal type → **“jupyter notebook”** and press enter.
6. It will automatically redirect to the localhost.





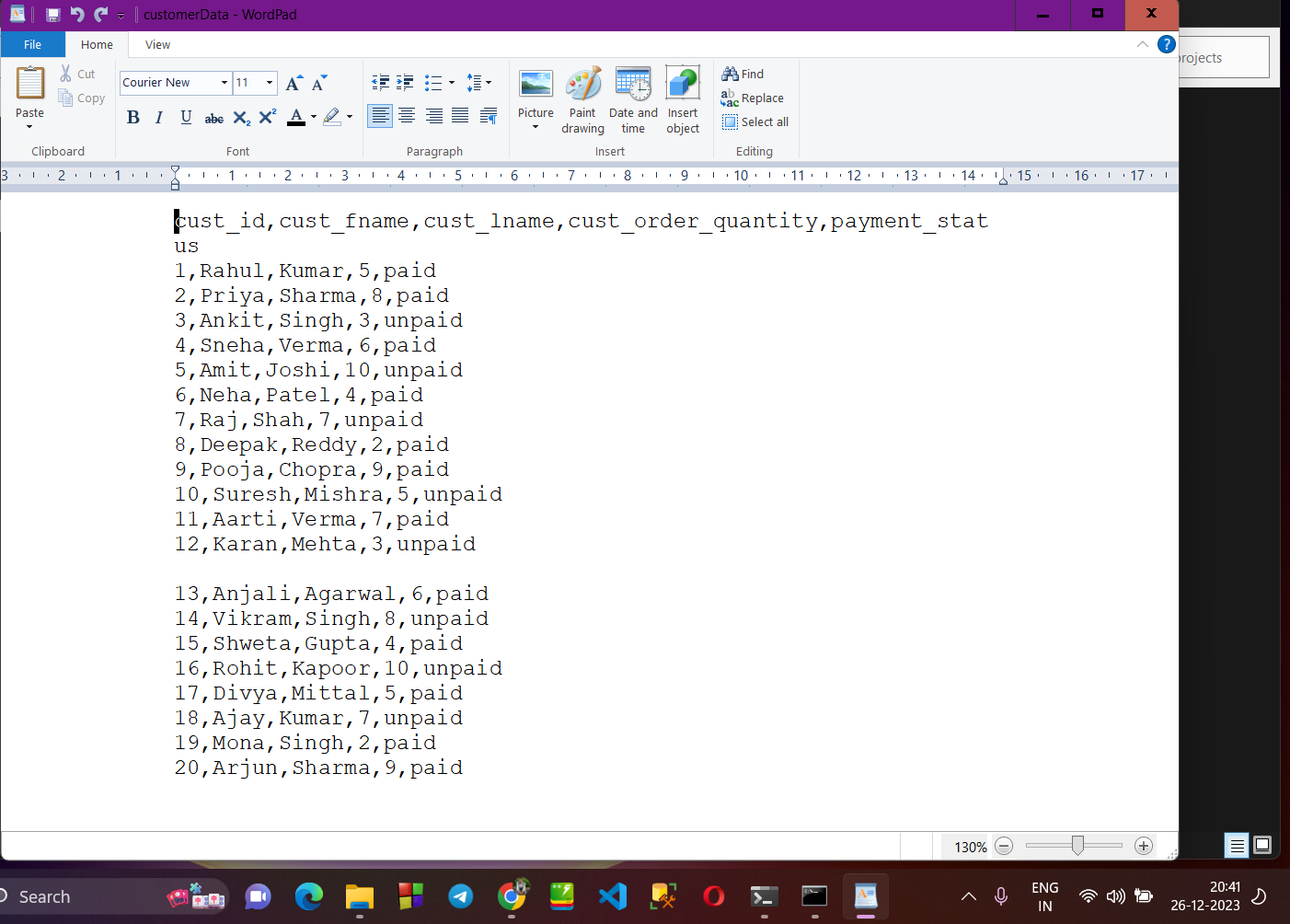


**→ What is ETL?**

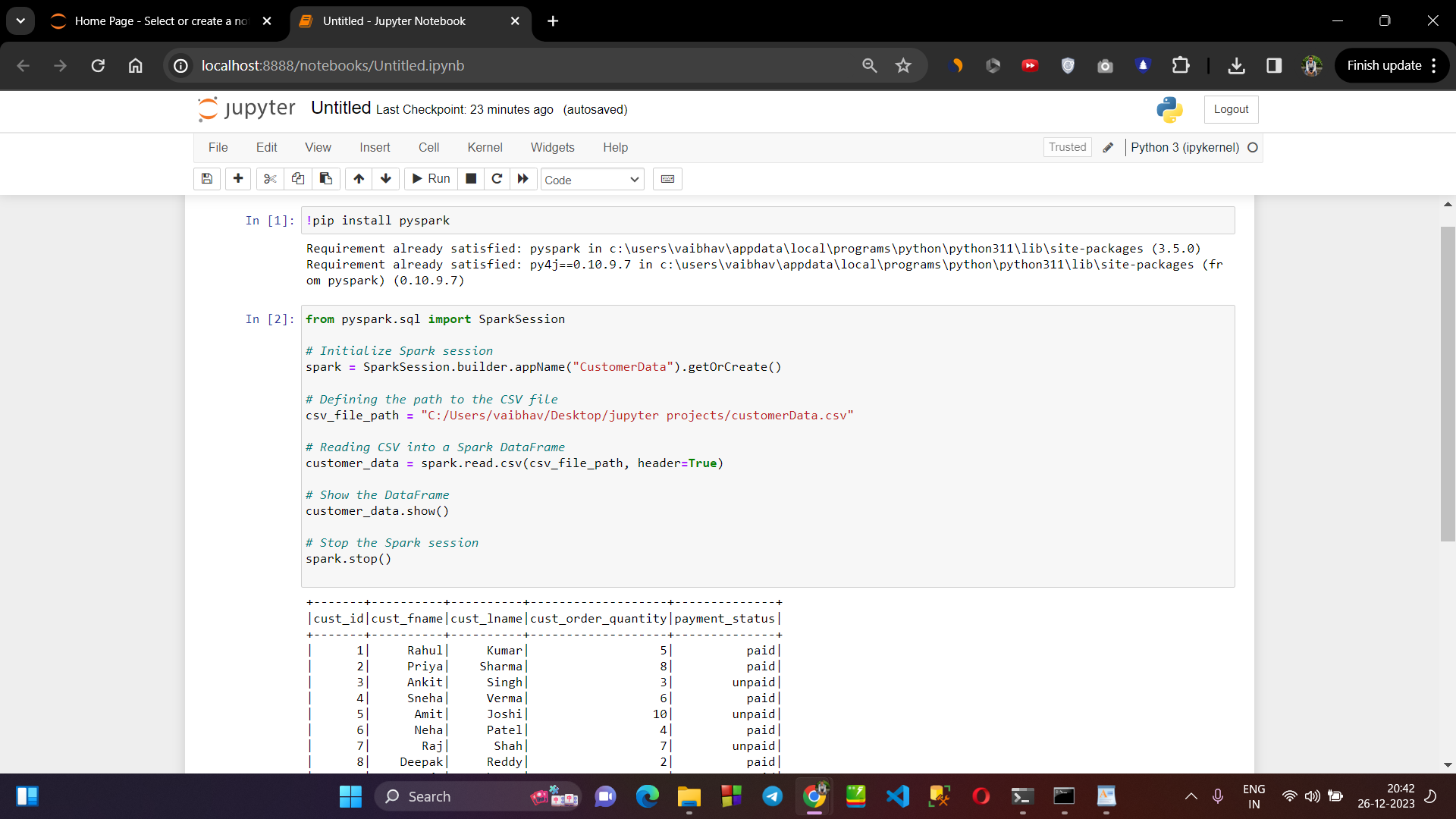
* The ETL stands for ***“Extract Transform & Load”.***
* Extract: Retrieve data from various sources like databases, files, or APIs.
* Transform: Clean, aggregate, and manipulate data to fit your analysis needs.
* Load: Store the transformed data into a database or data warehouse for analysis.

→ **Reading data of a CSV File into the jupyter notebook using pyspark :**

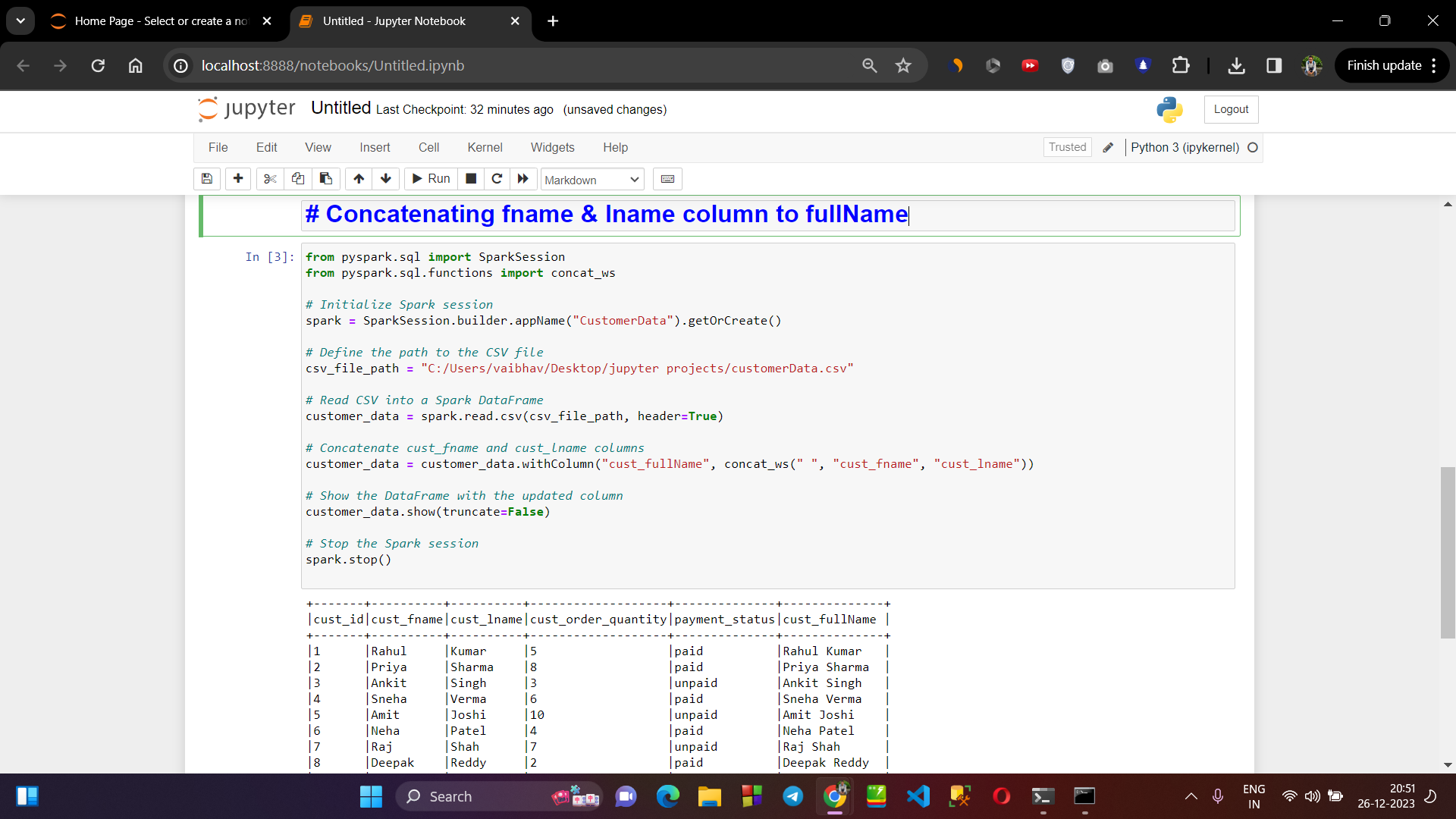
**>** **Creating a csv file :**



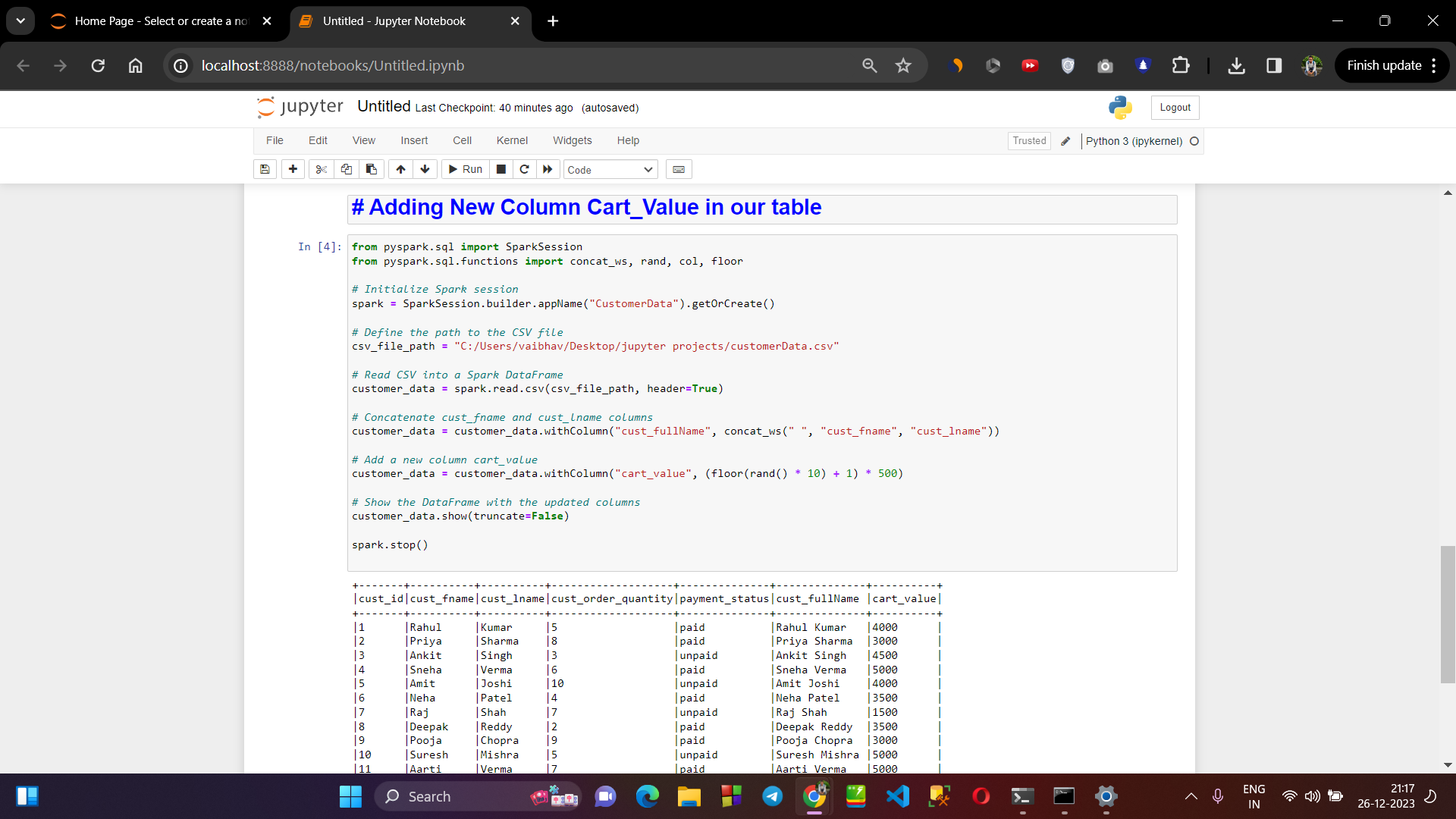
**> Now reading its data using SPARK.READ.CSV() in jupyter notebook :**



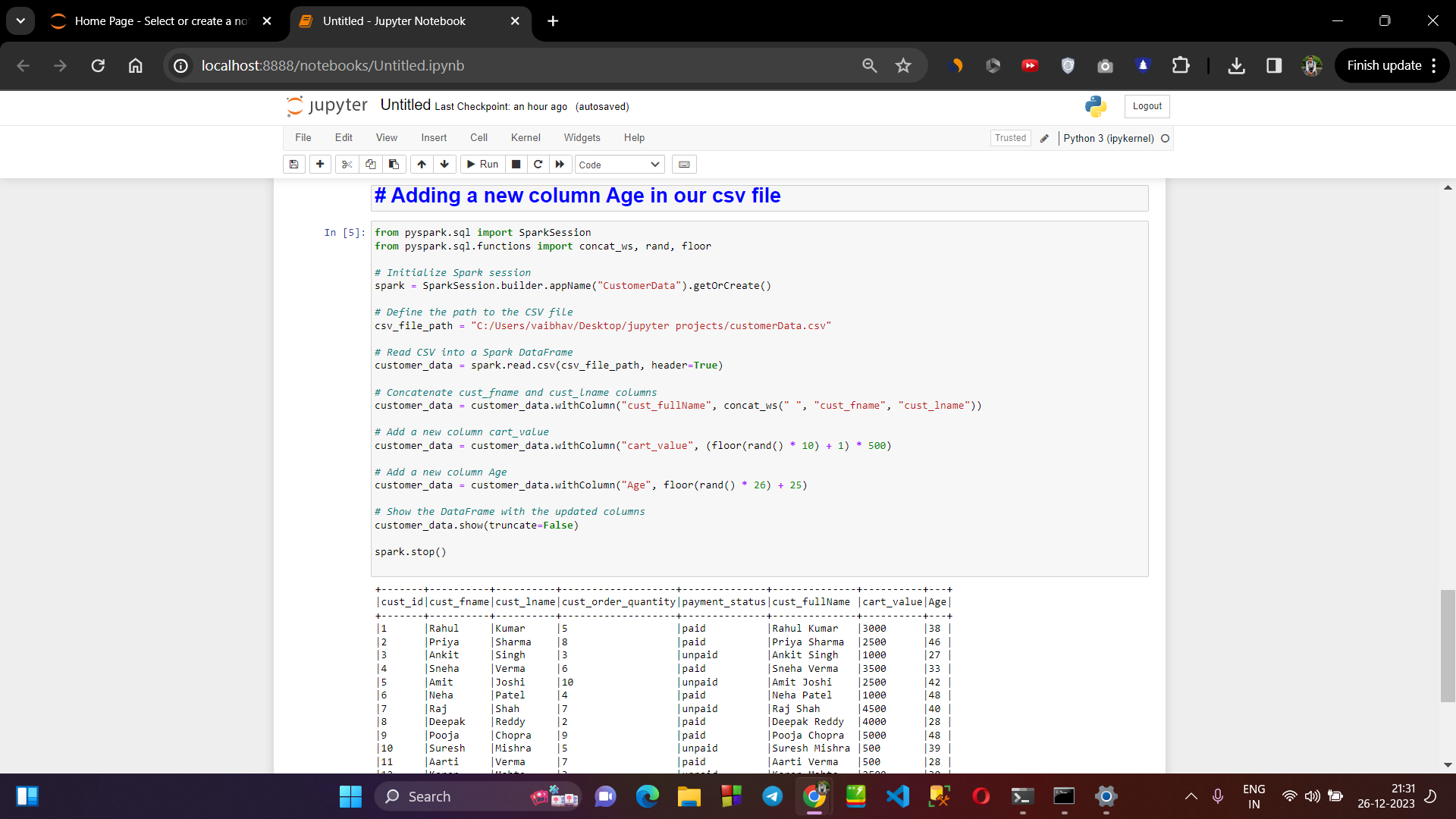
**>Transformation 1 : Creating a new column cust\_fullName by adding cust\_fname and cust\_lname :**

****

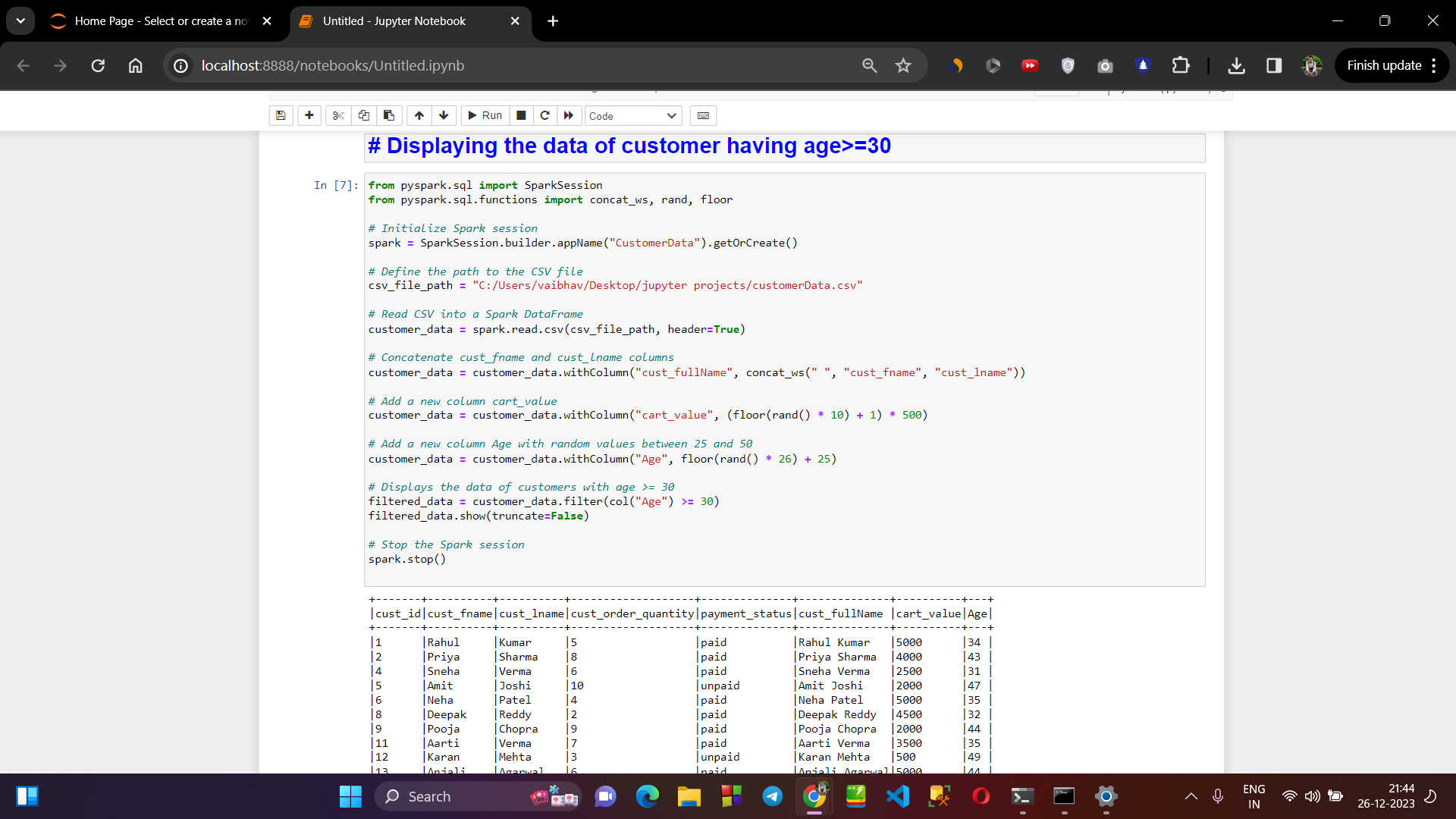
**>Transformation 2 : Adding a new column Cart\_Value in our table :**

****

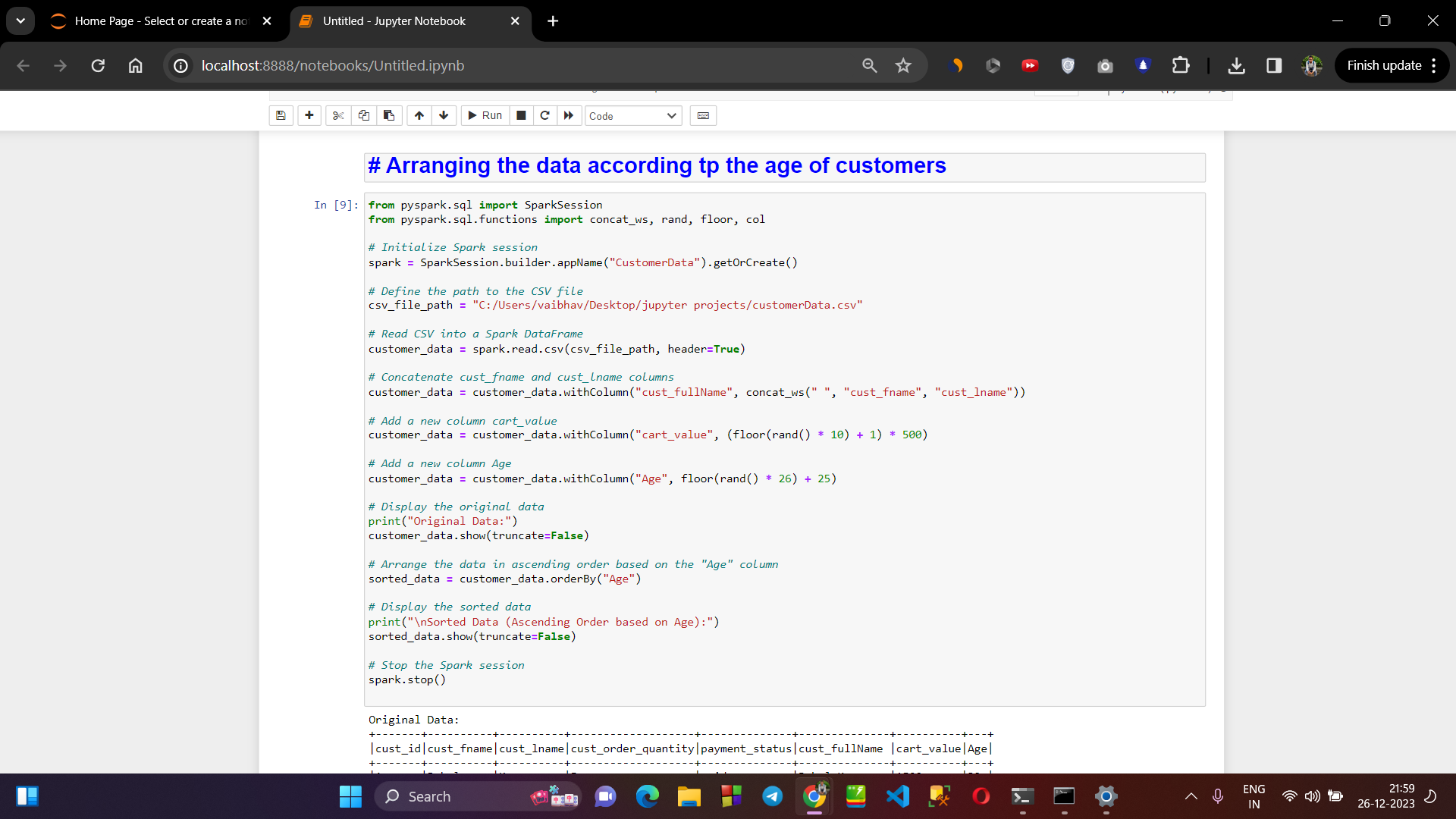
**> Transformation 3 : Adding a new column Age in our table :**

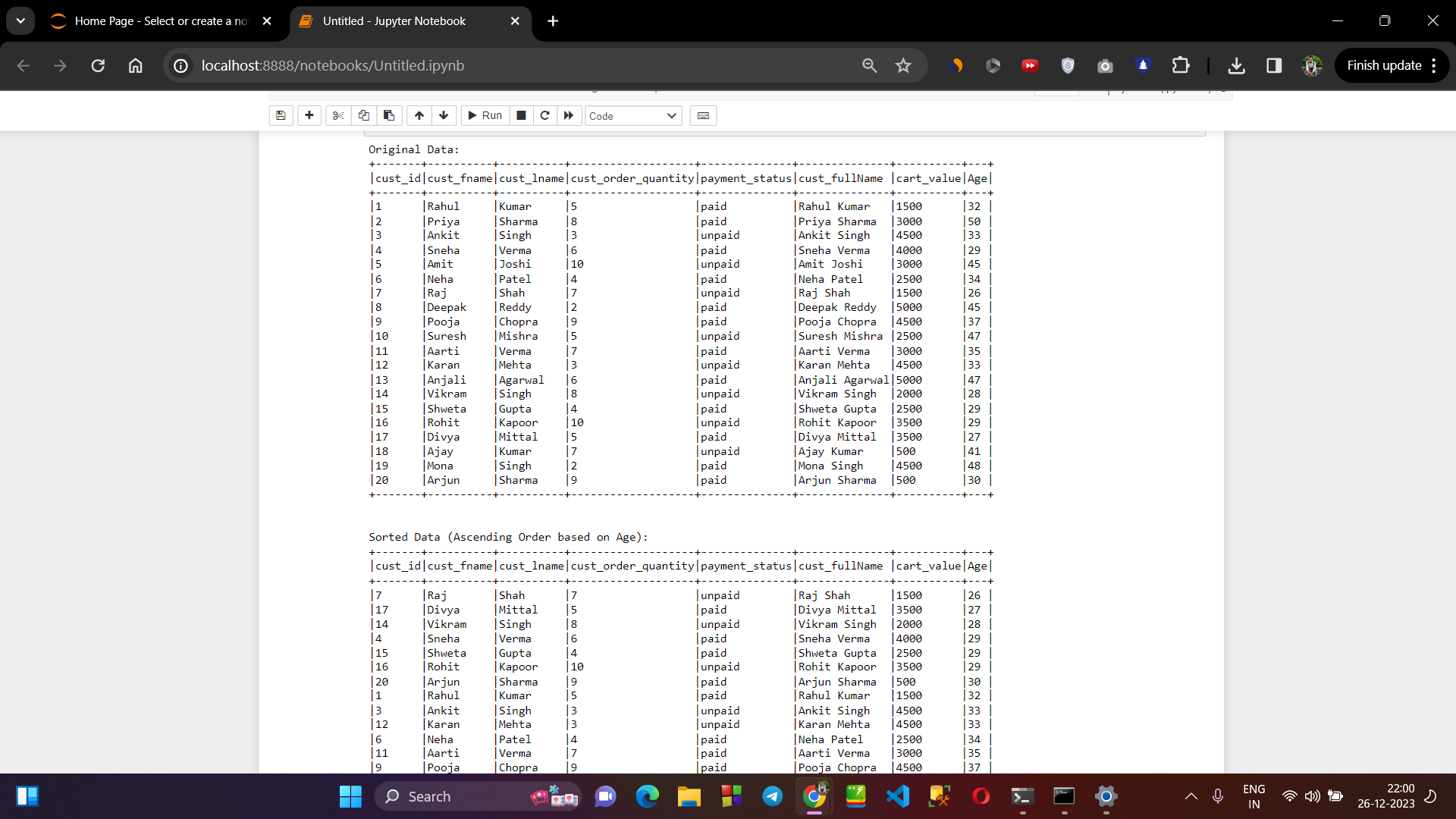
****

**>Transformation 4 : Displaying the data of customer having age>=30**

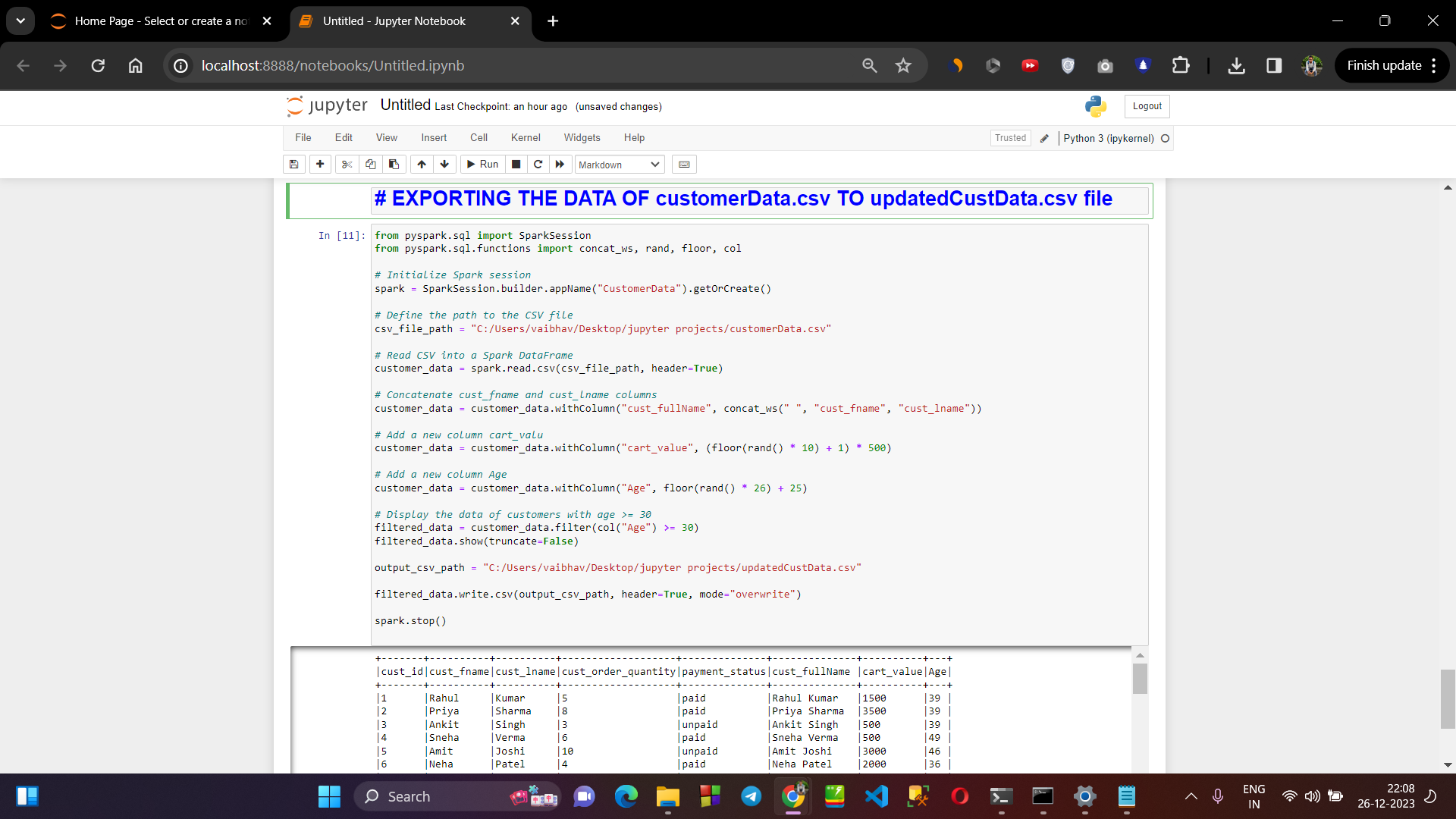
****

**> Transformation 5 : Arrange the data in ascending order based on the age of customers :**

****

****

**> Transformation 6 : Exporting the data into a new file :**

****