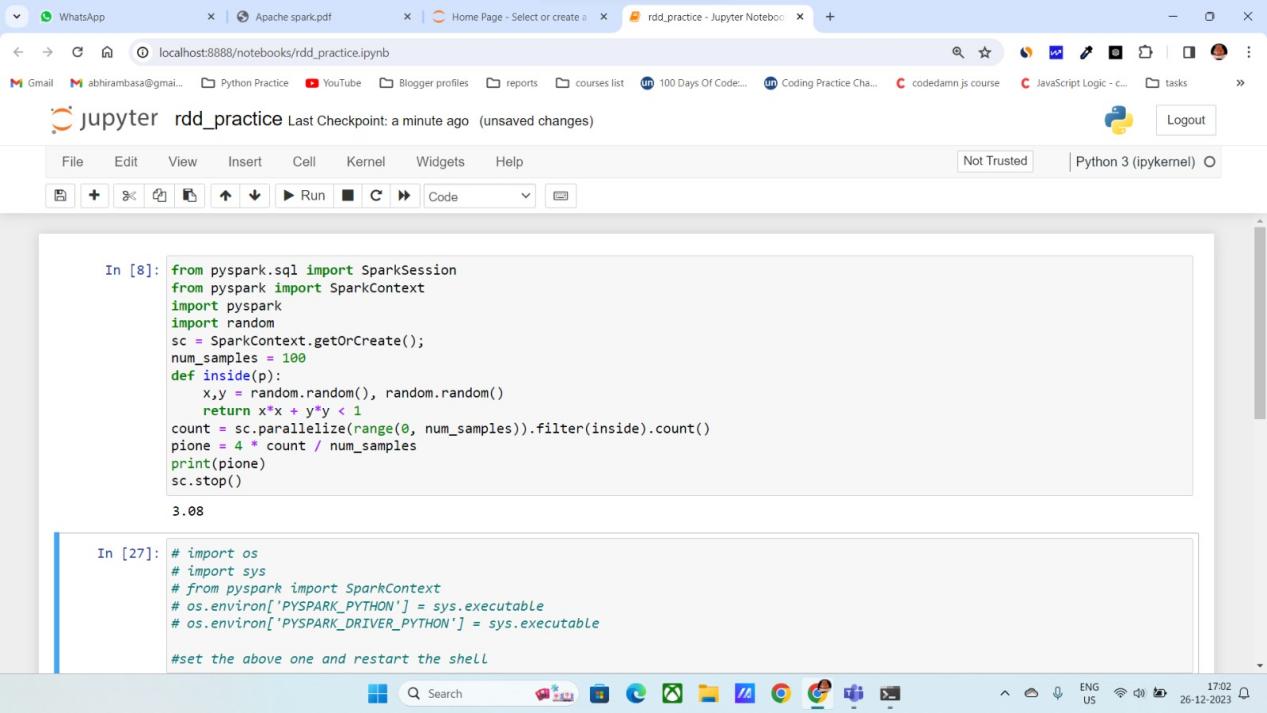
**PySpark Assessment 4 :** - Abhiram Basa

Today we started with sample programs in pyspark.

**Parallelize():** Here we used Parallelize() method to create an RDD for the program.

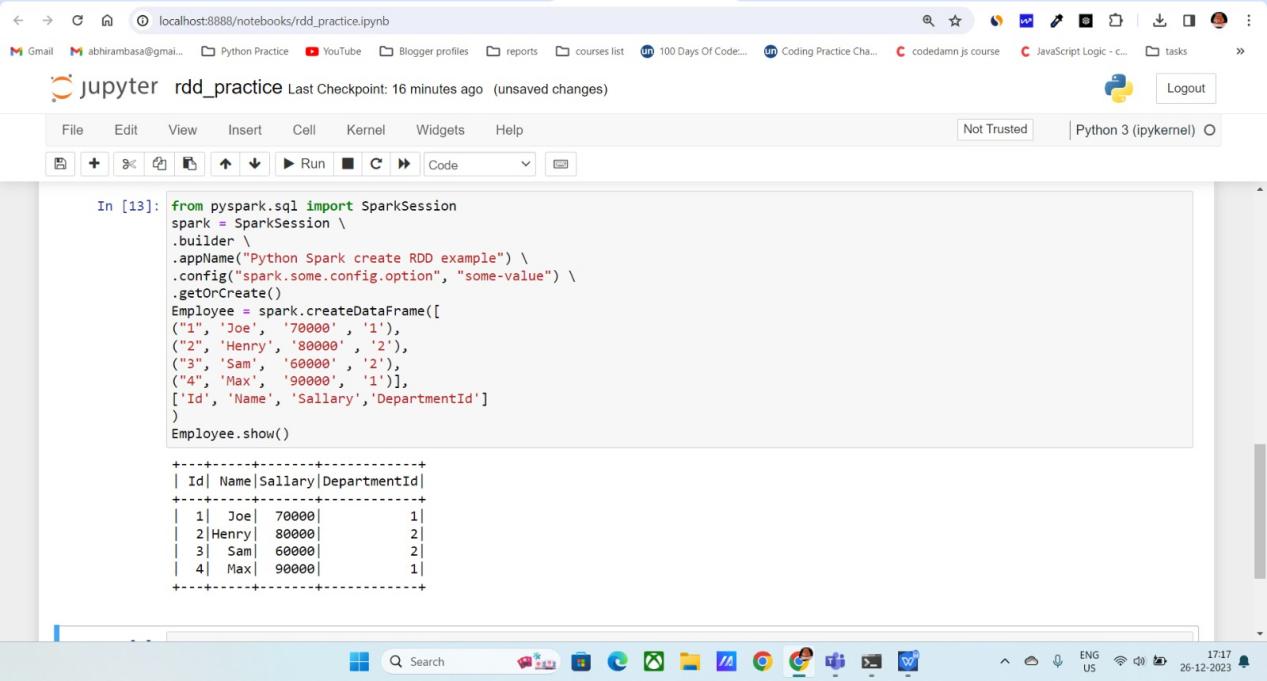
* First we created a spark session.
* Imported modules.
* We imported the **random module** which will give the random numbers.
* Below program returns the variables which are less than 1.
* At last we must stop the spark session.

**Output:**



In below program we create a data set using parallelize() method and print it.





**ETL :**

ETL stands for

**Extract**: extract the data from the different sources

**Transform**: Transform the unstructured data into structured data. Transformations like cleaning, manipulation, etc.

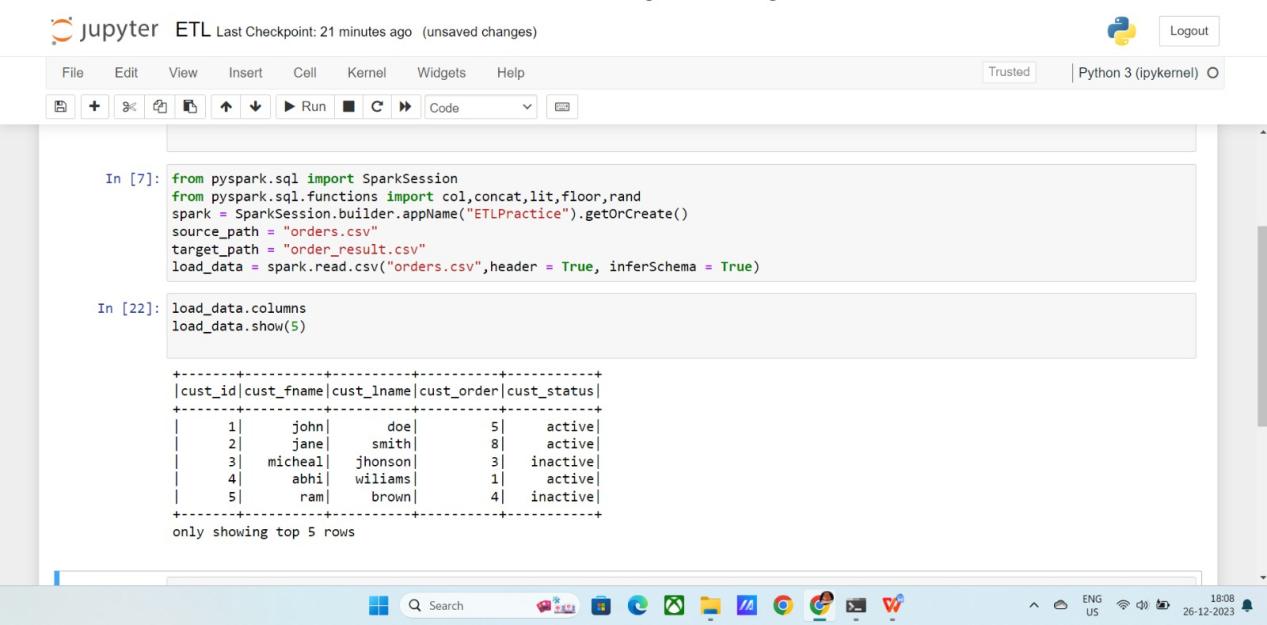
**Load**: Load the transformed data into a location or date warehouse.

Below we performed ETL on sample **CSV** data.

**OUTPUT:**

First I loaded the data using

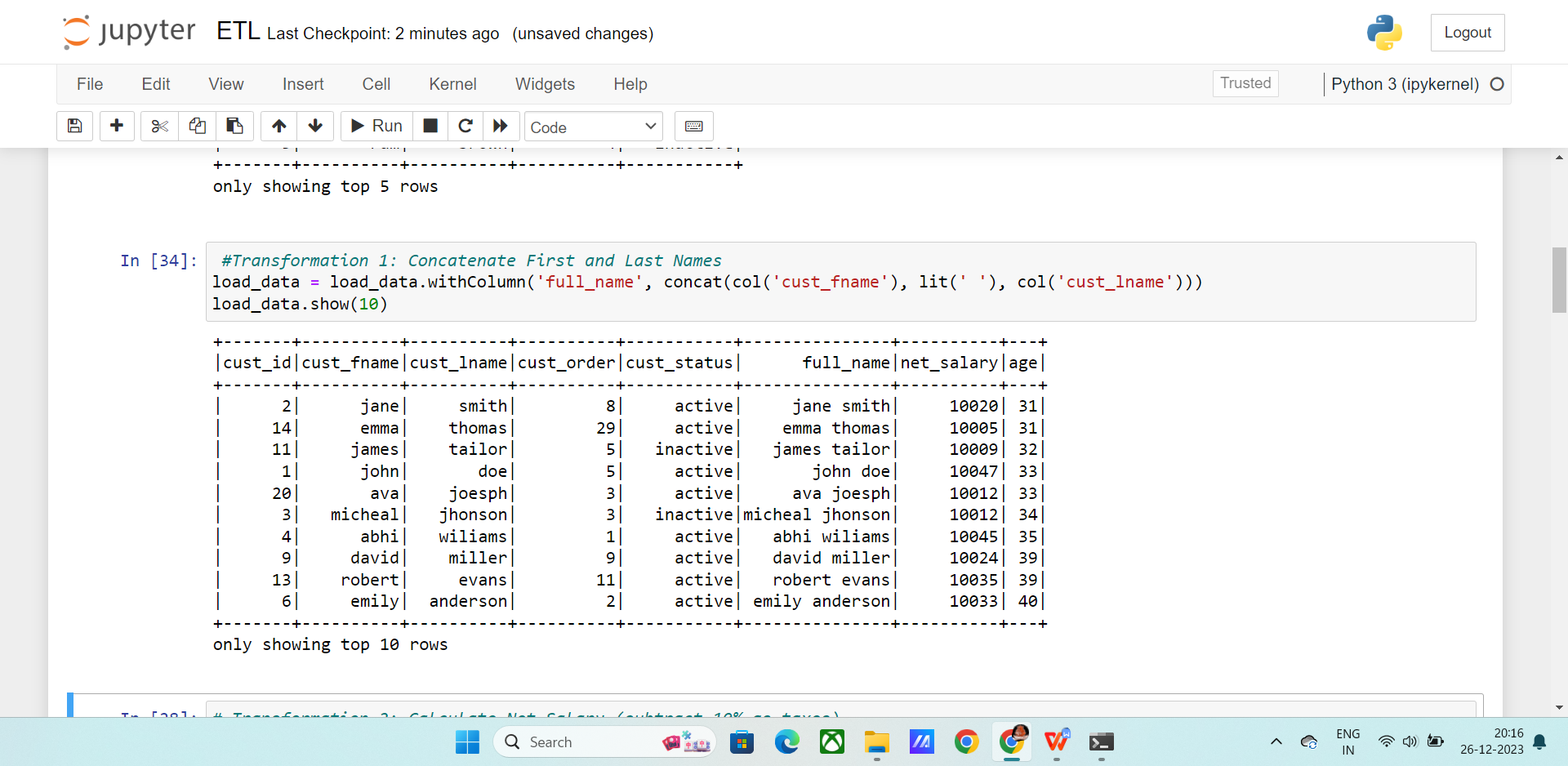
**Spark.read.csv(“filename.csv”)** and stored it in a variable.

****

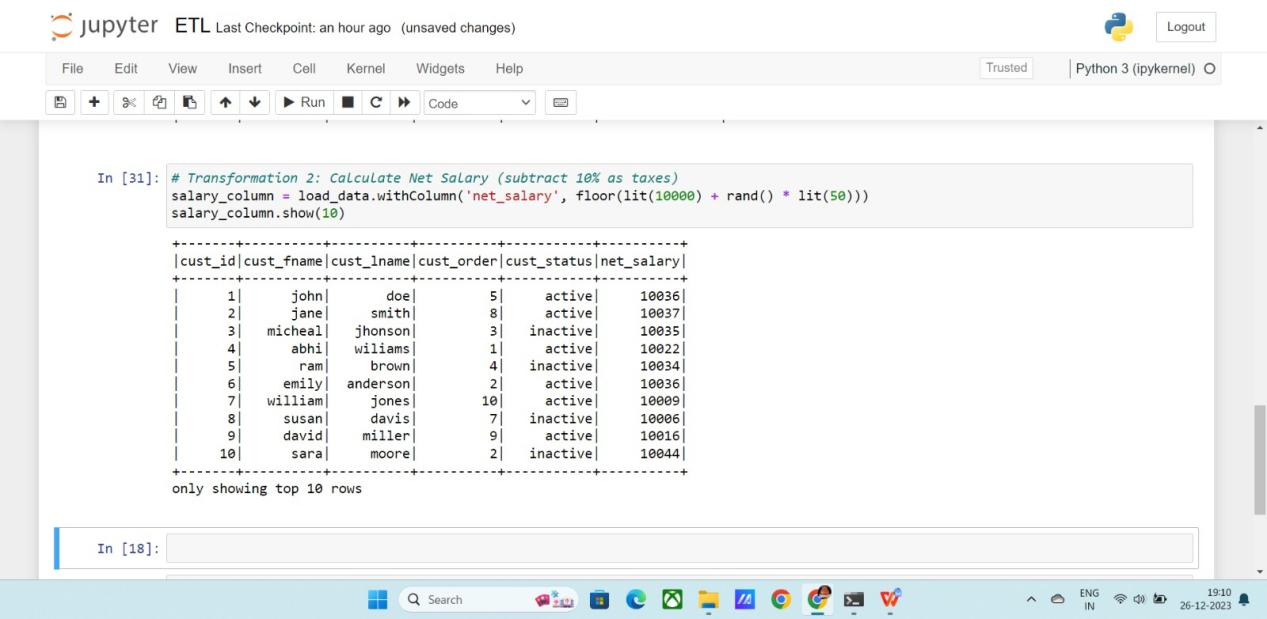
Added a column called “ net salary” using

**Withcolumn(“column\_name”,condition)**

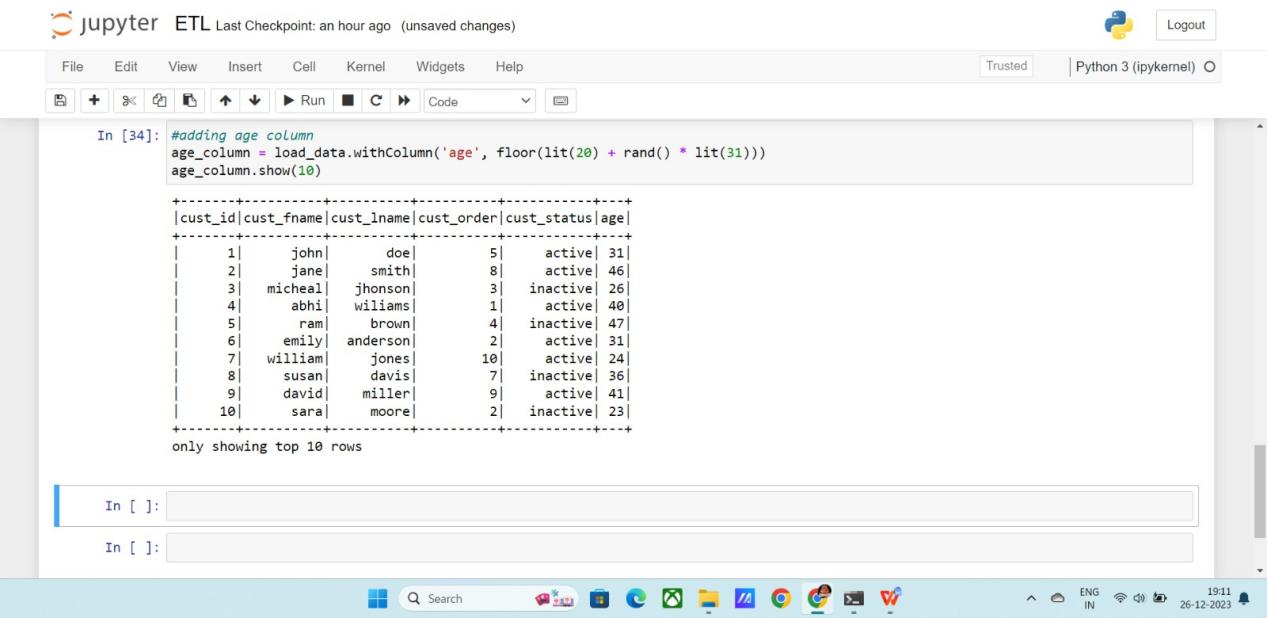
Here I added a full name column by concatenating the first name and last name of the users.



Here the salary is generated using **rand()** function and added as a new column as net\_salary.

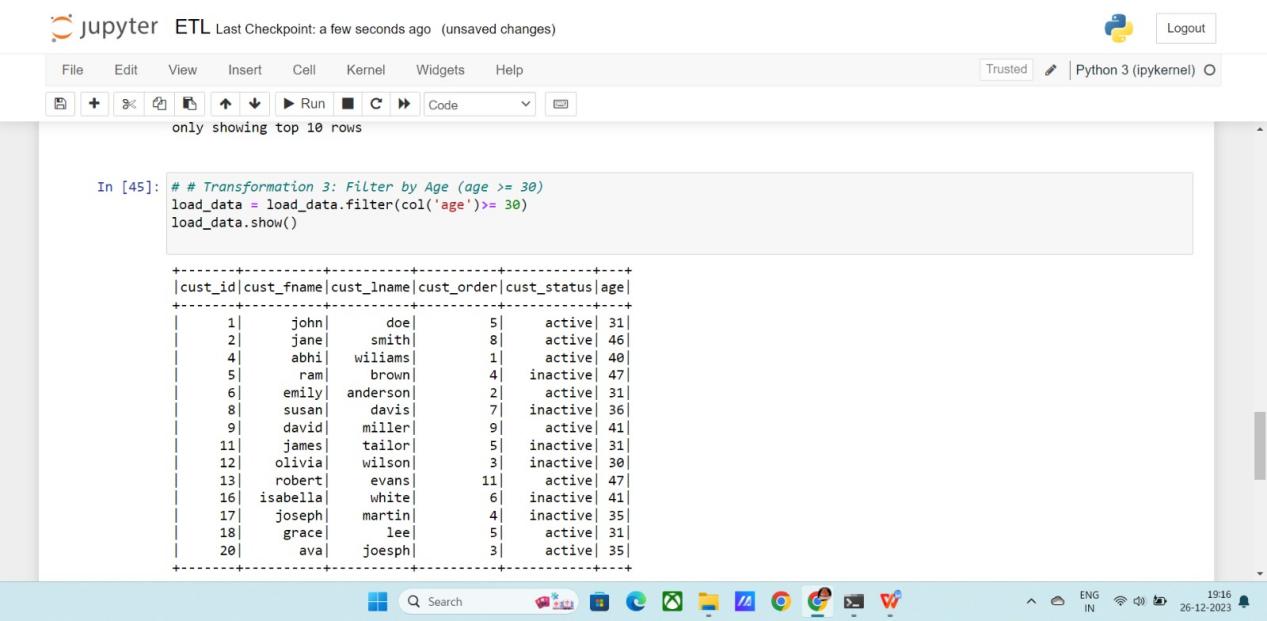


In the same way added “age” column using **withcolumn()** built in function.

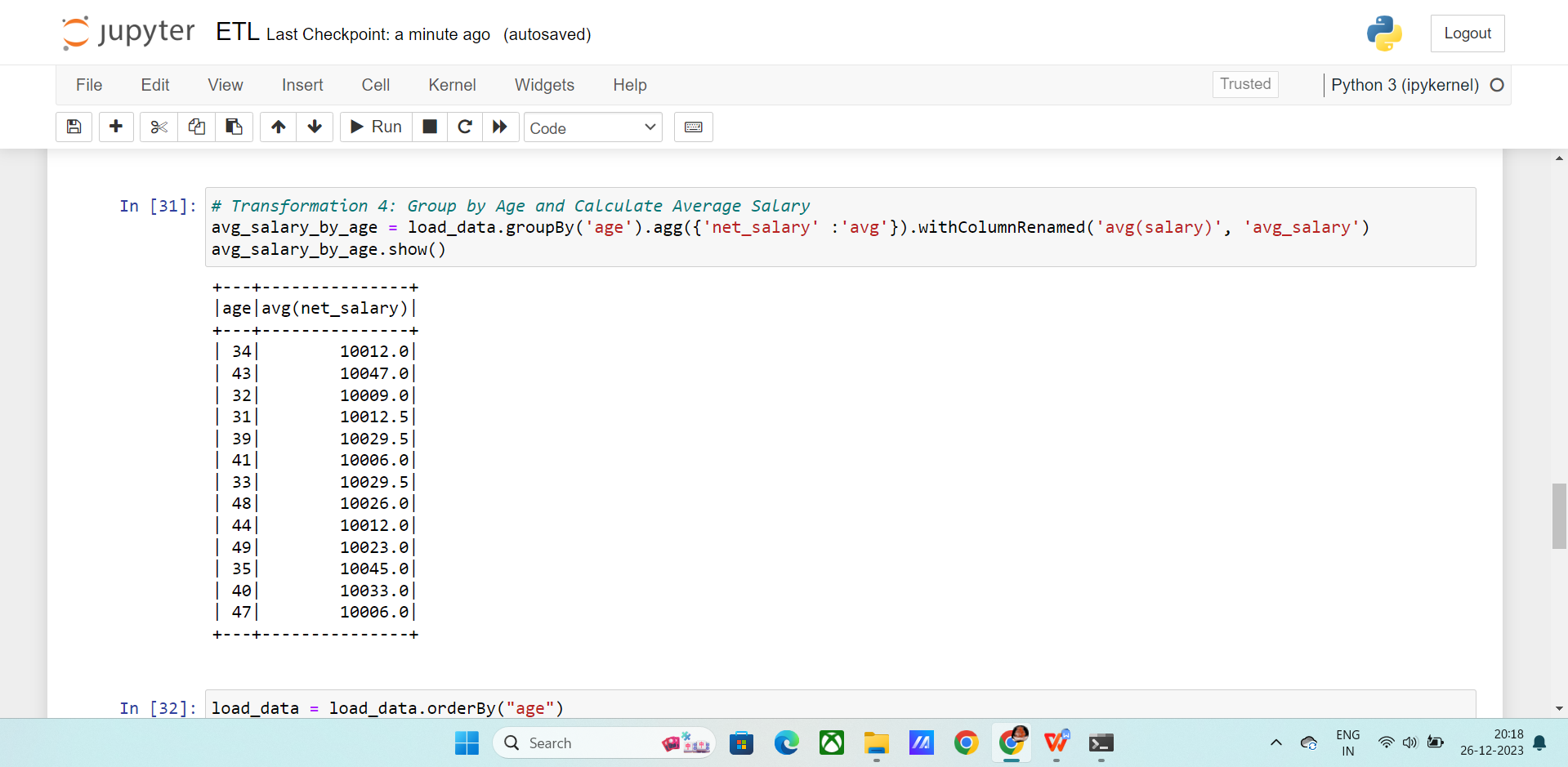


**Using filter:** Filter will filter the resultant data set based on the condition.

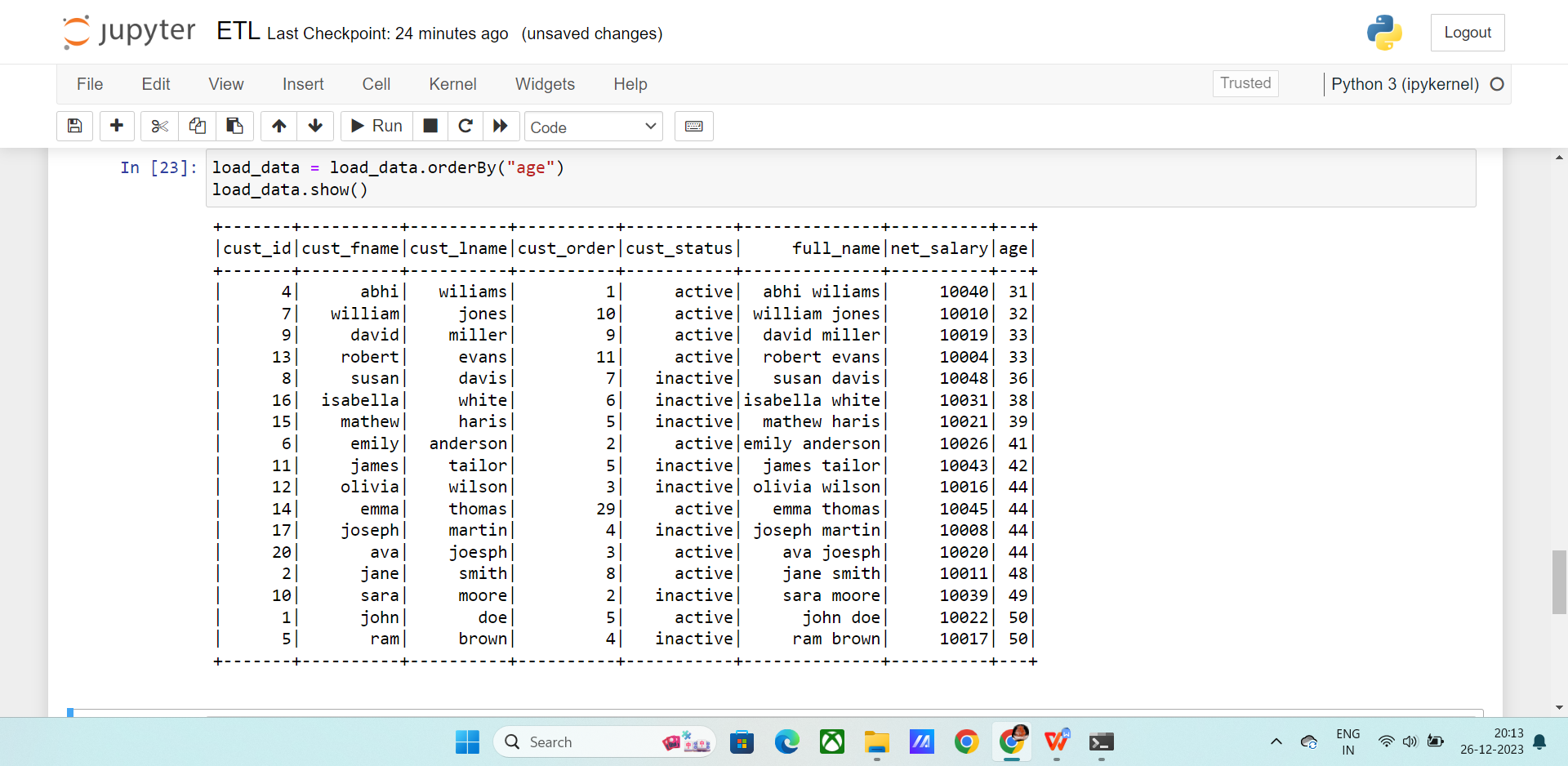
Here we are filtering data where age > 30.



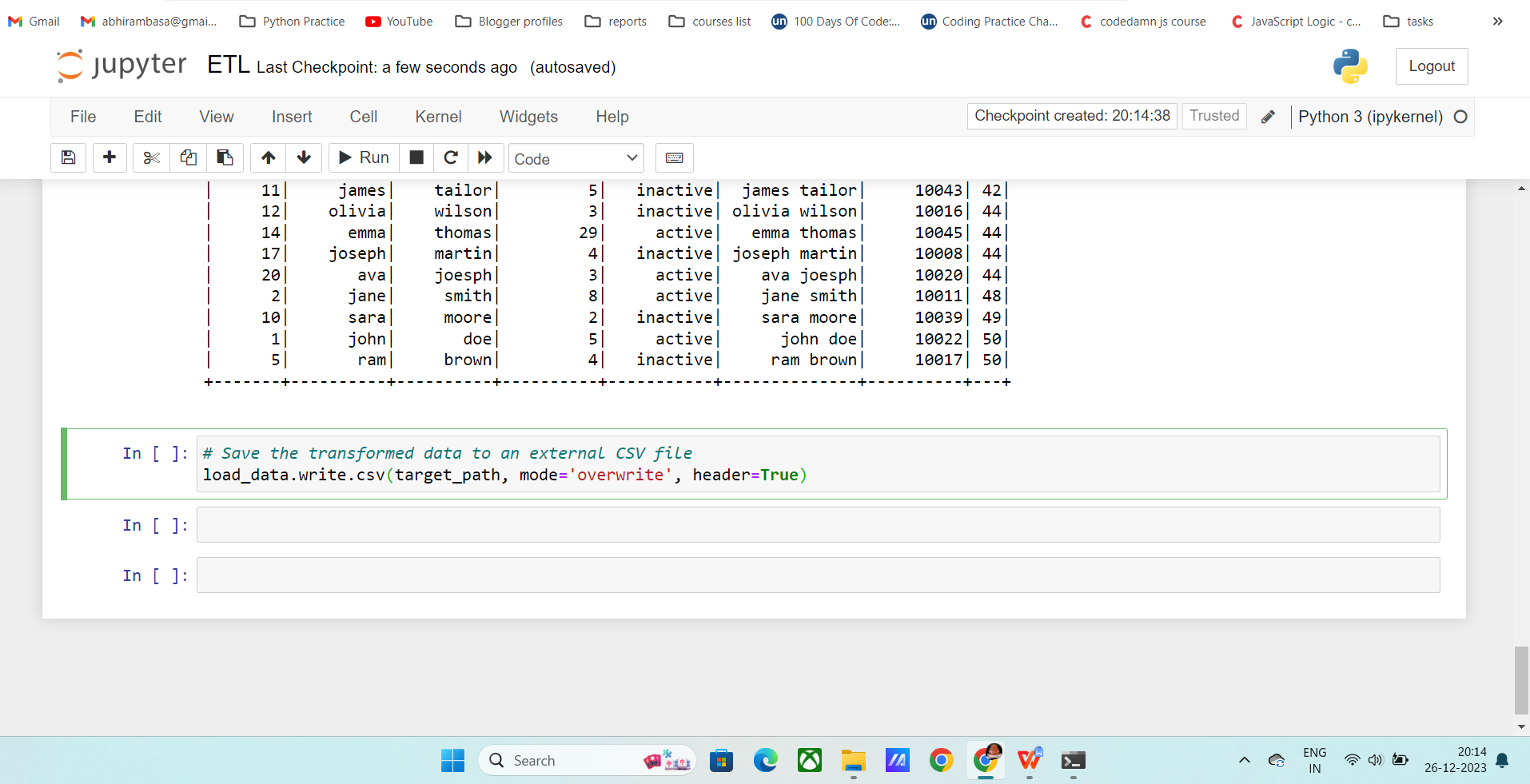
Here I’ve used **GroupBy()** with respect to net salary and age columns and below is the output



Next I ordered the data with age column



At last saving the data into another **CSV** file using writer method in spark.



Finally we completed the ETL exercise by working the ETL process on a data set.