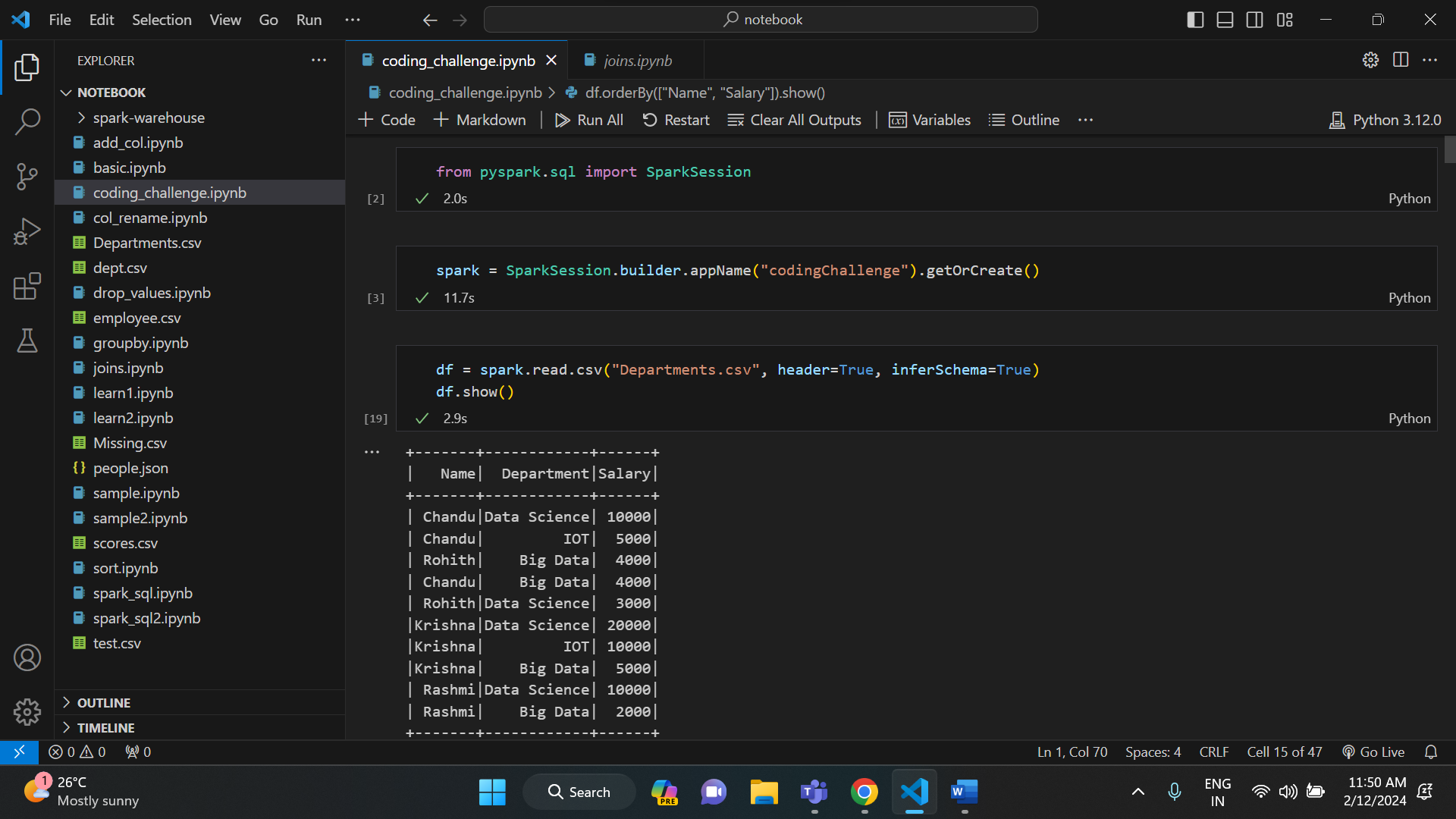
**PYSPARK CODING CHALLENGE QUESTION –1**

* **GROUP BY IN PYSPARK:**

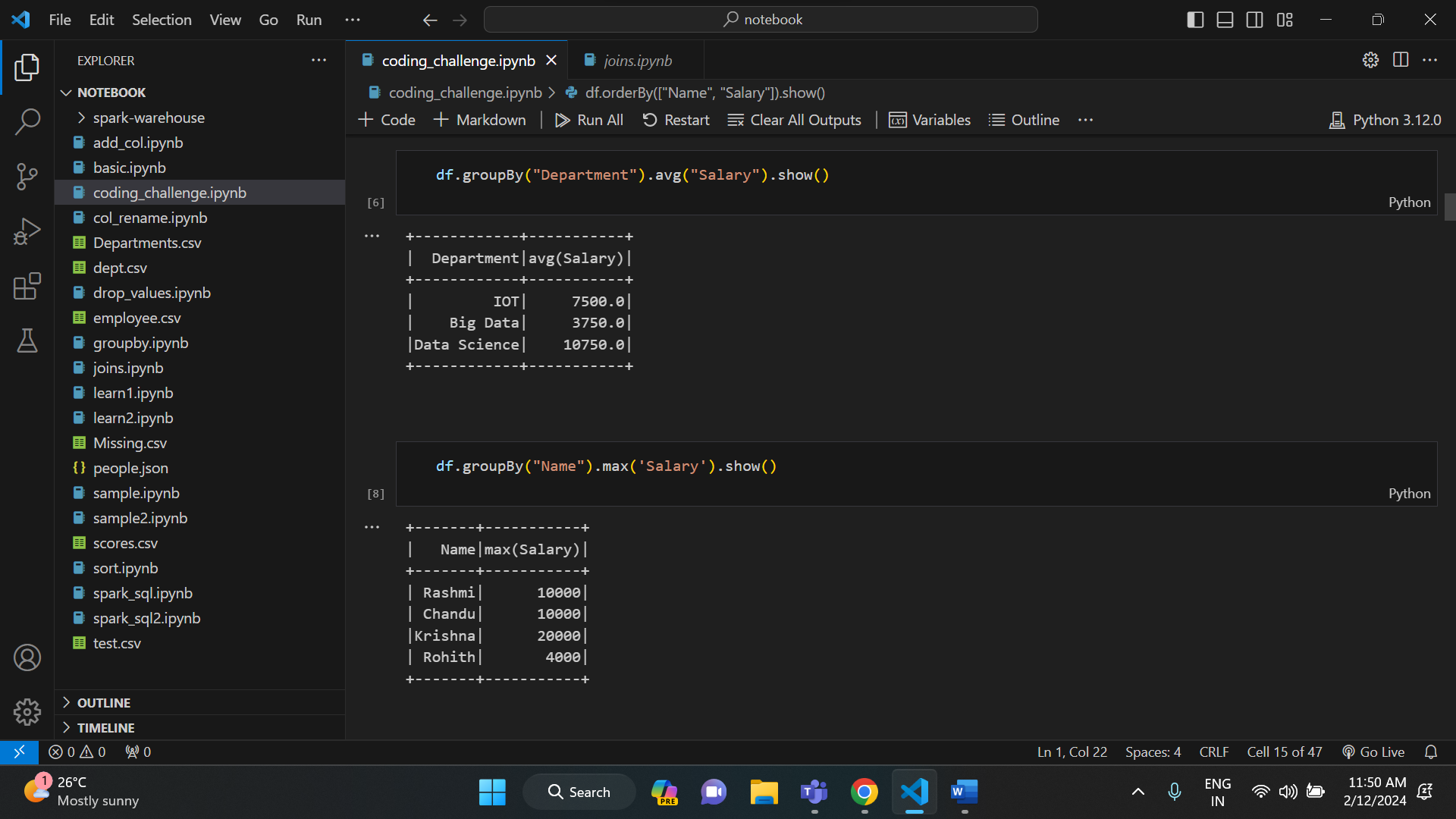
The group by in pyspark works similar to group by in SQL. The main purpose of group by is to group similar data.

For the execution purpose, I have used departments.csv file.

Execution of group by:



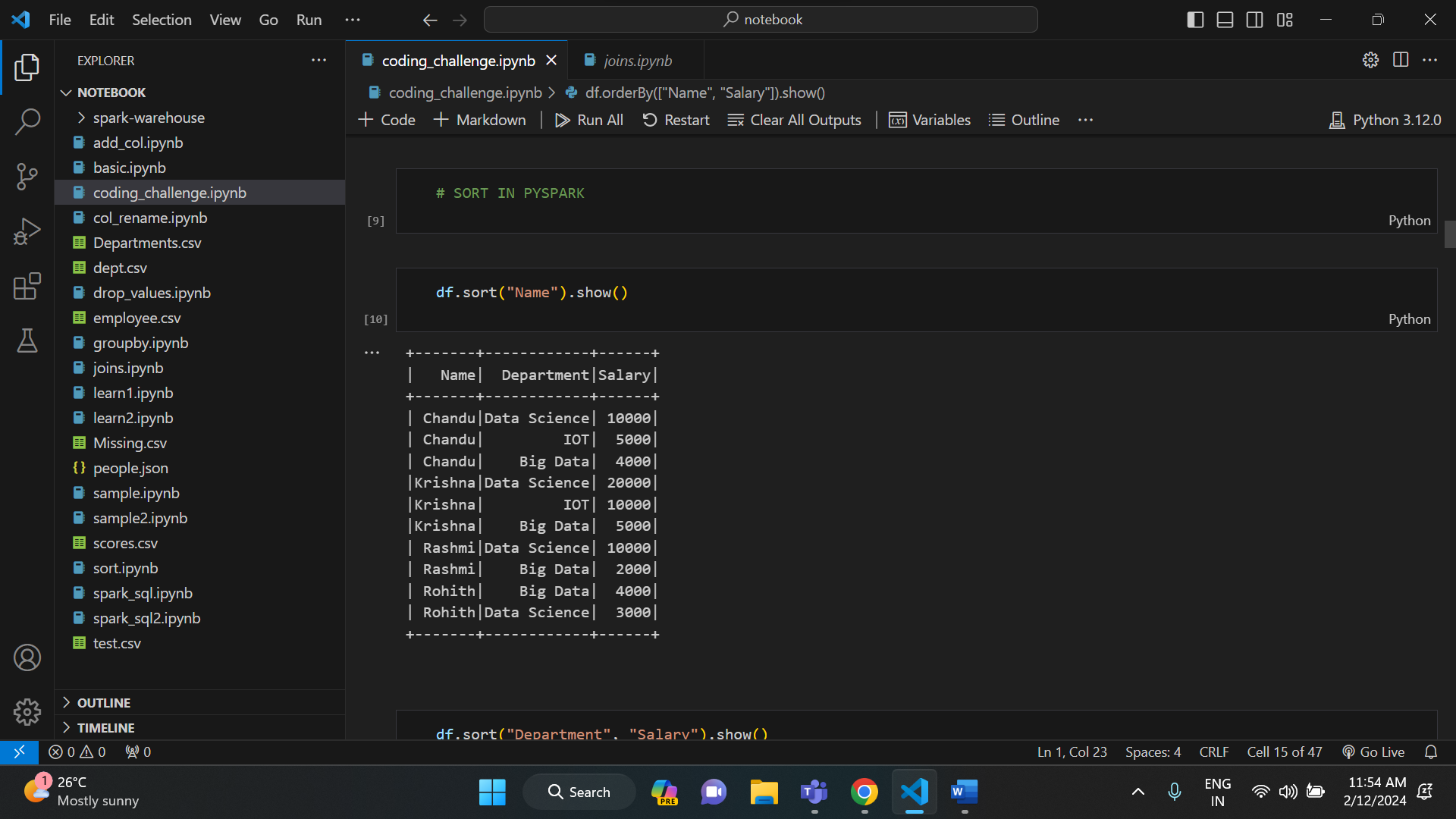


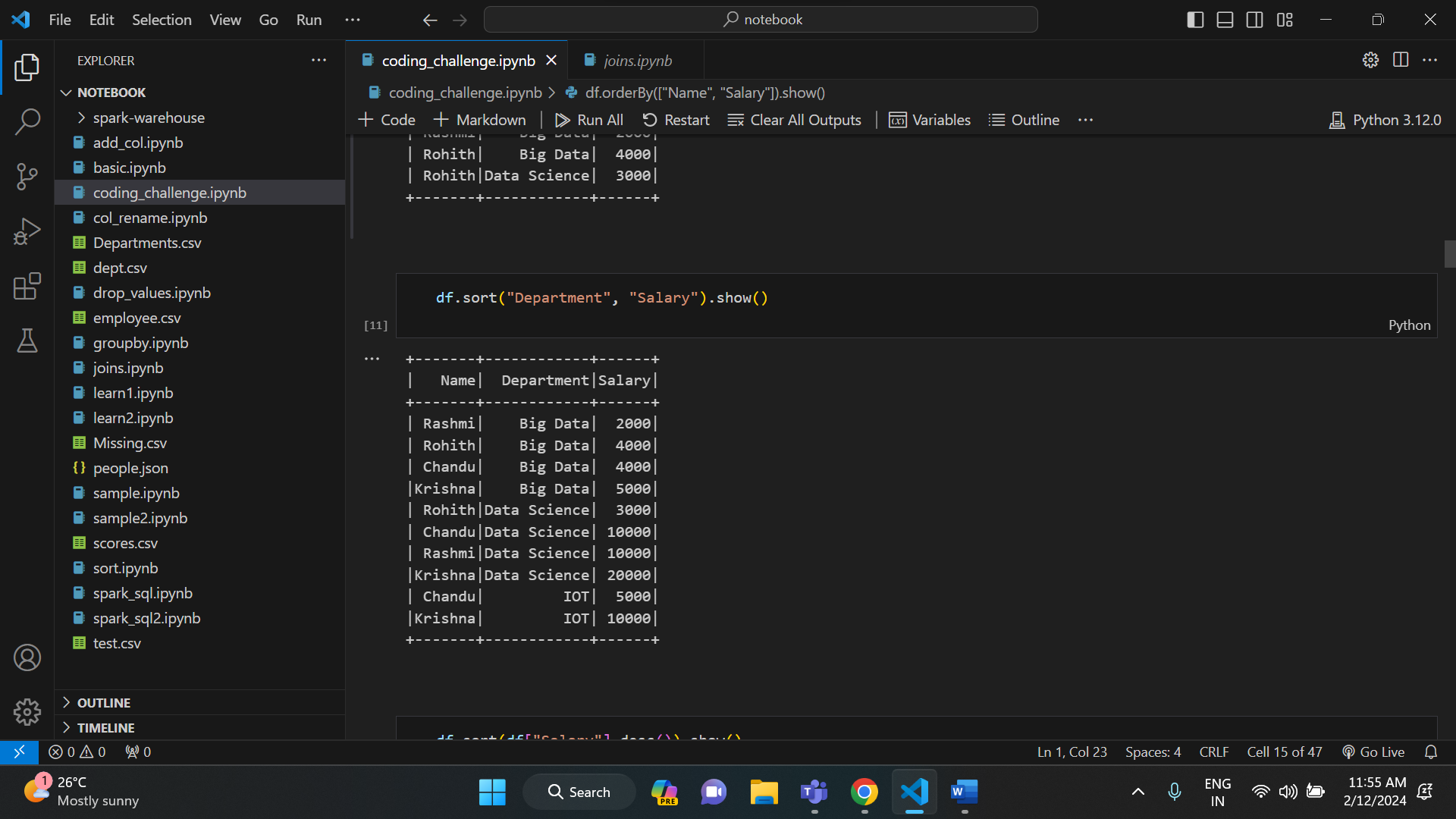


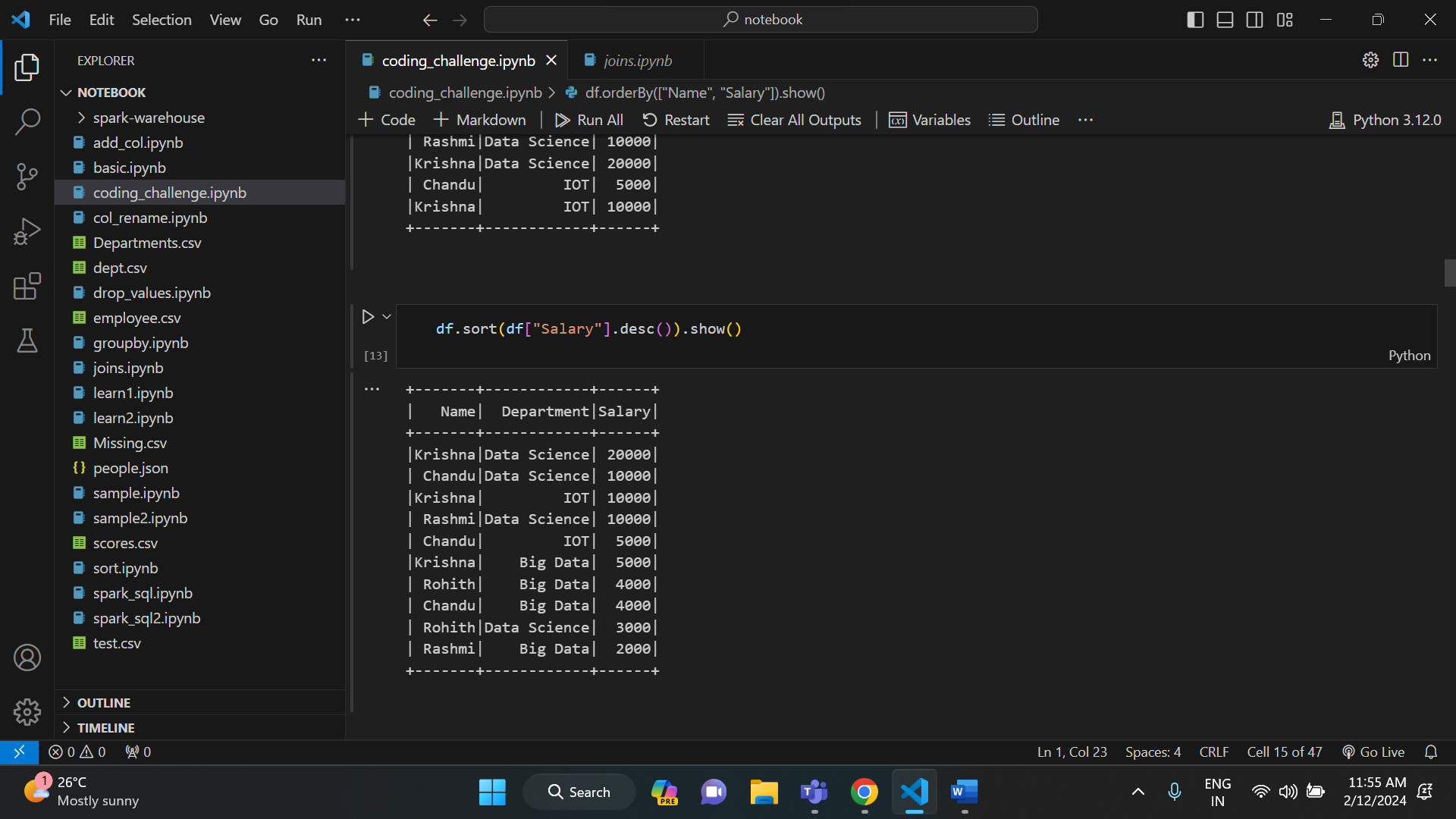
* **SORT IN PYSPARK:**

The sort () in pyspark is used to sort the data based on the parameter given inside the parenthesis. If we give salary inside the parenthesis, it sorts the data based on salary. For the execution purpose, I have used departments.csv file.

Execution of sort ():

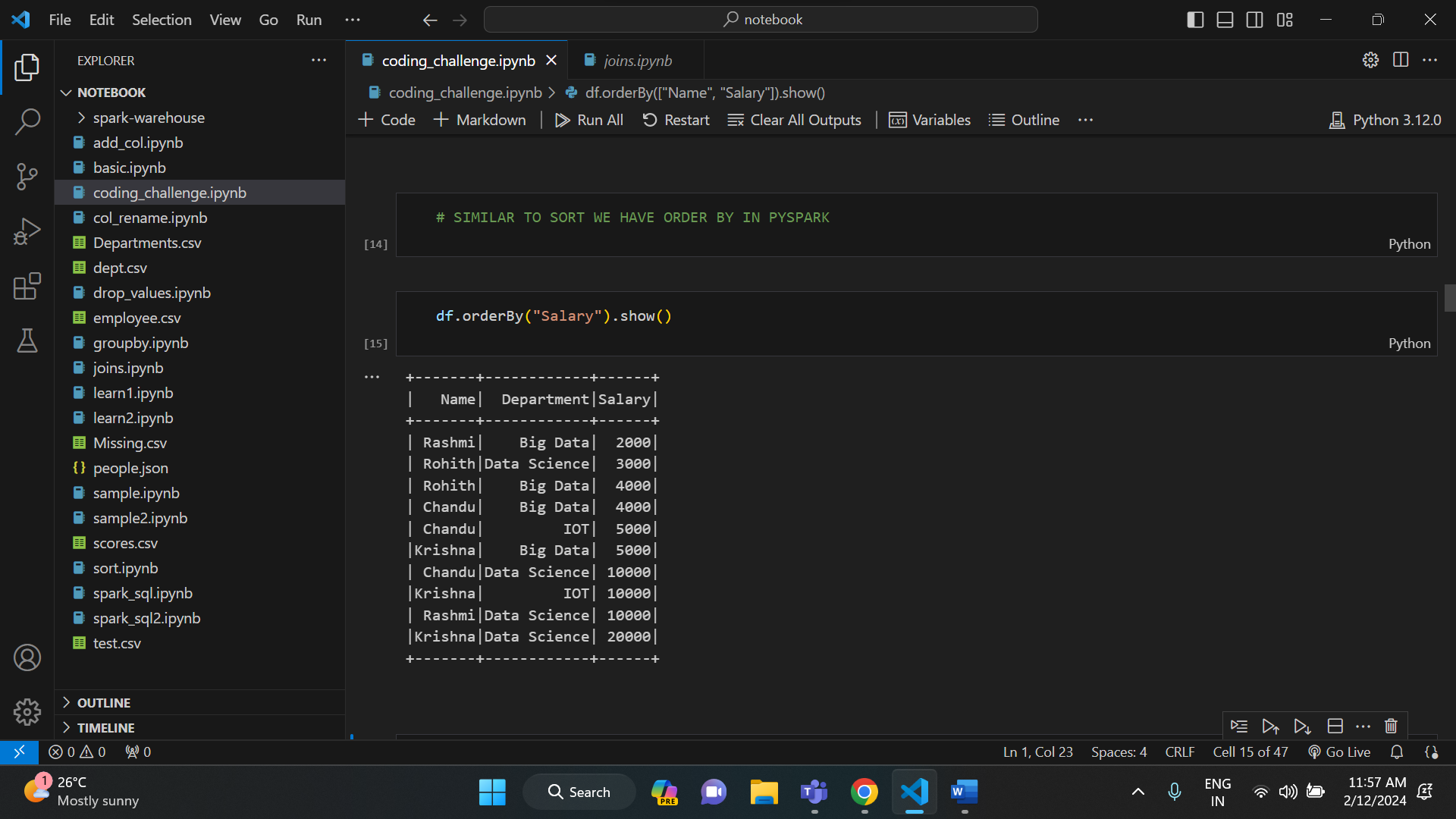


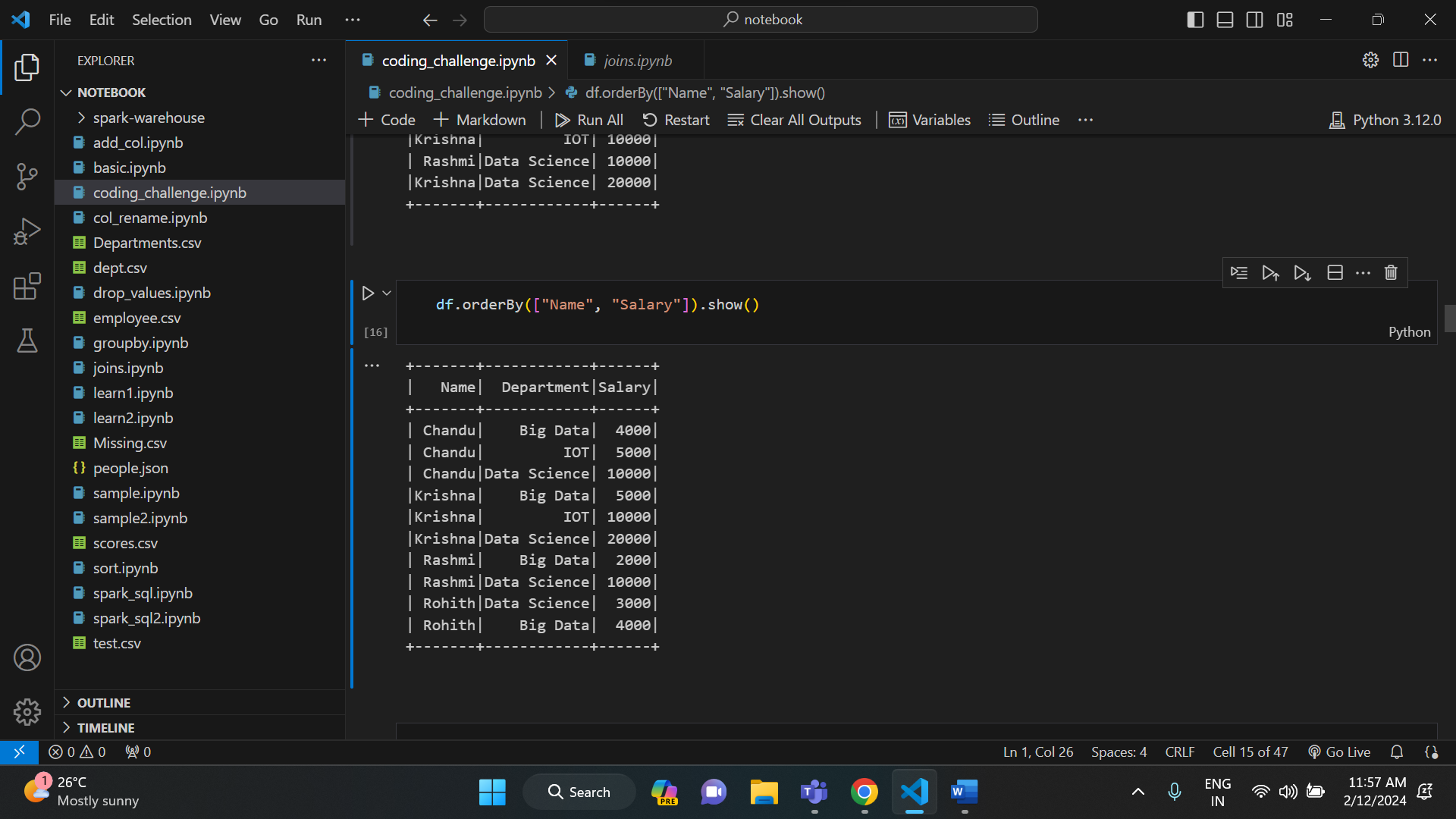




In the above example, I have sorted data in descending order based on salary.

Similar to sort (), we have order by function in pyspark. It is also used to sort the data in a particular order.



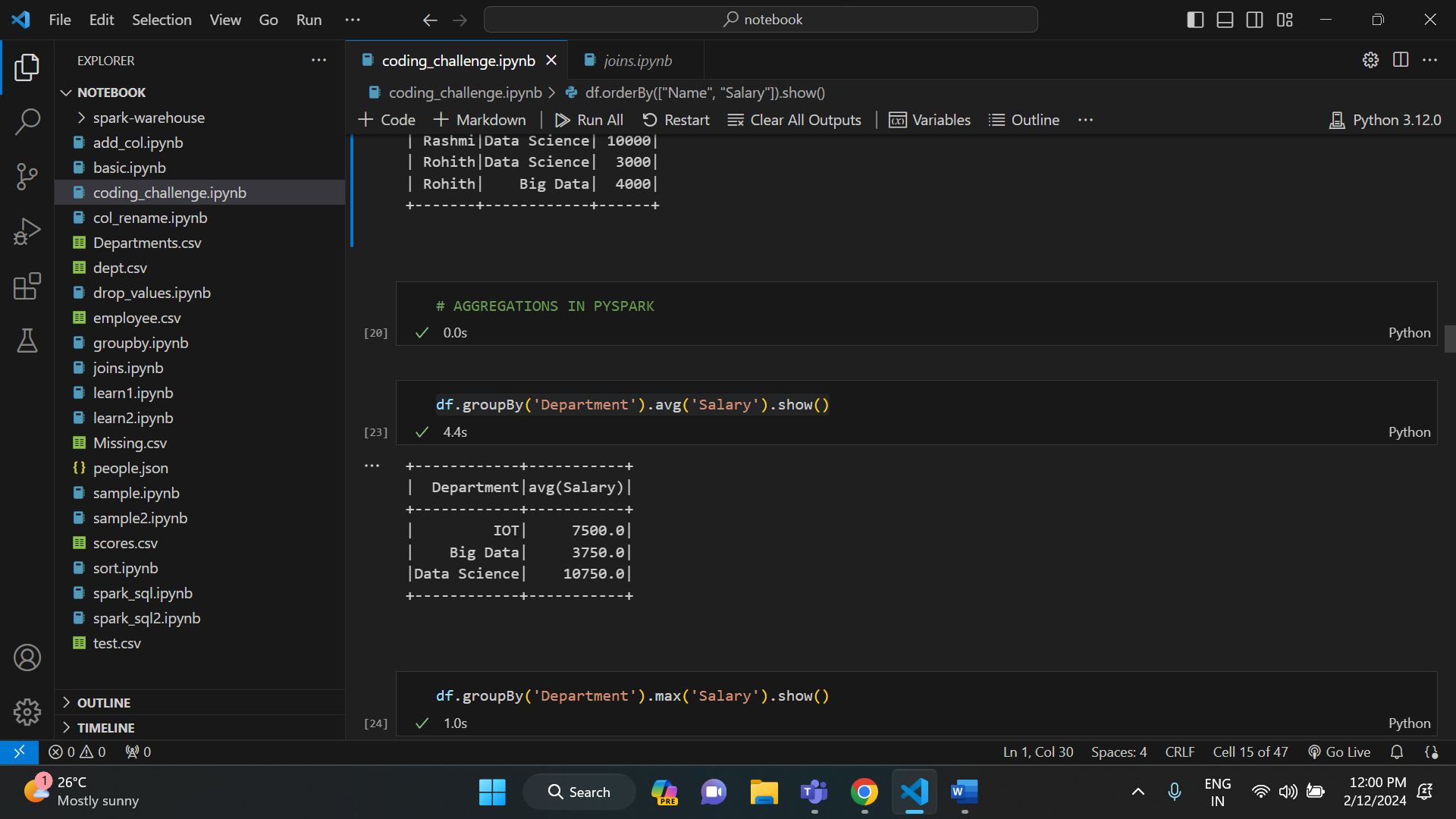


* **AGGREGATIONS IN PYSPARK:**

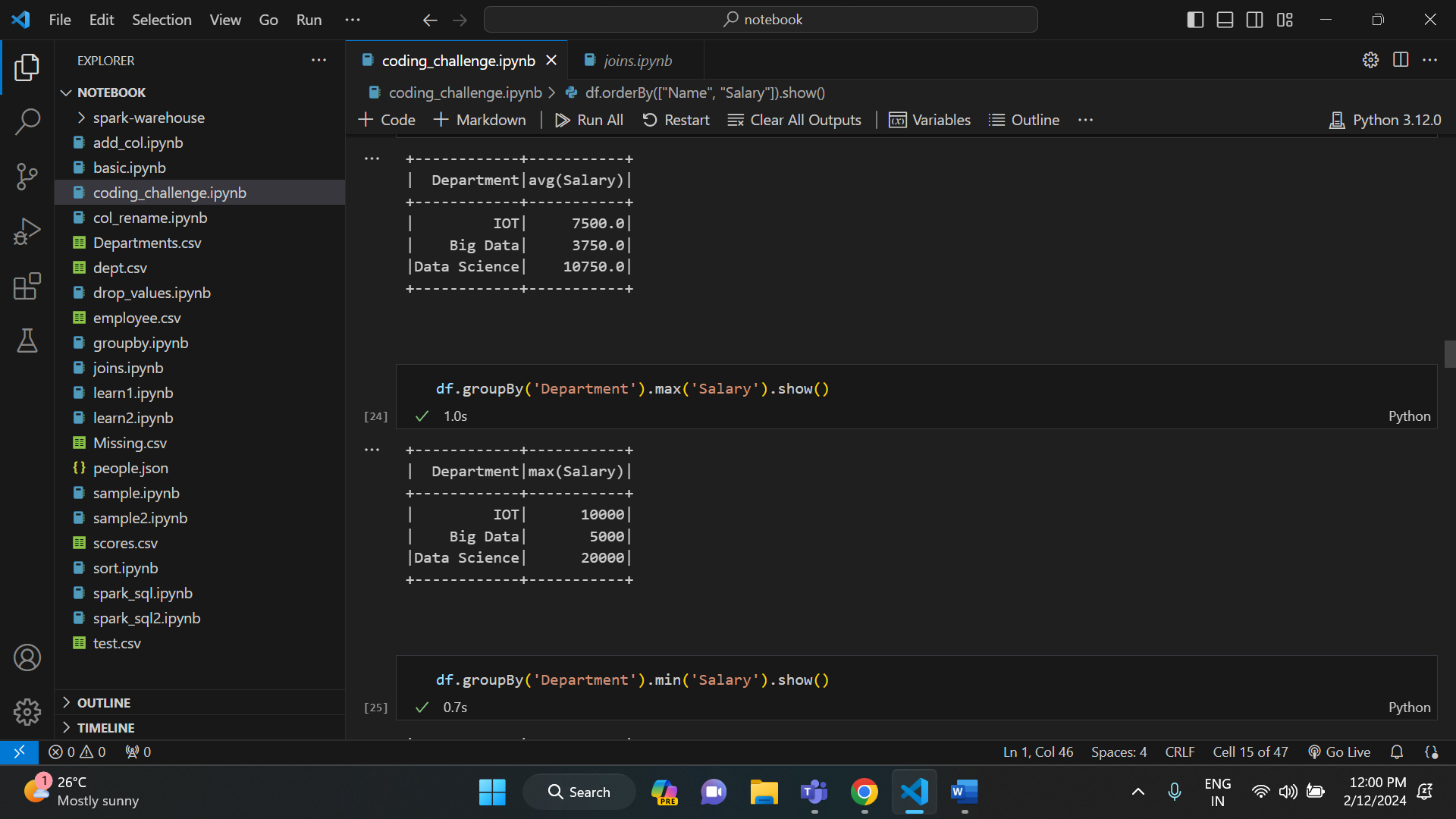
Similar to SQL, we have some aggregate functions in pyspark. They are sum, avg, mean, min, max etc.

Execution of Aggregate Functions in pyspark:

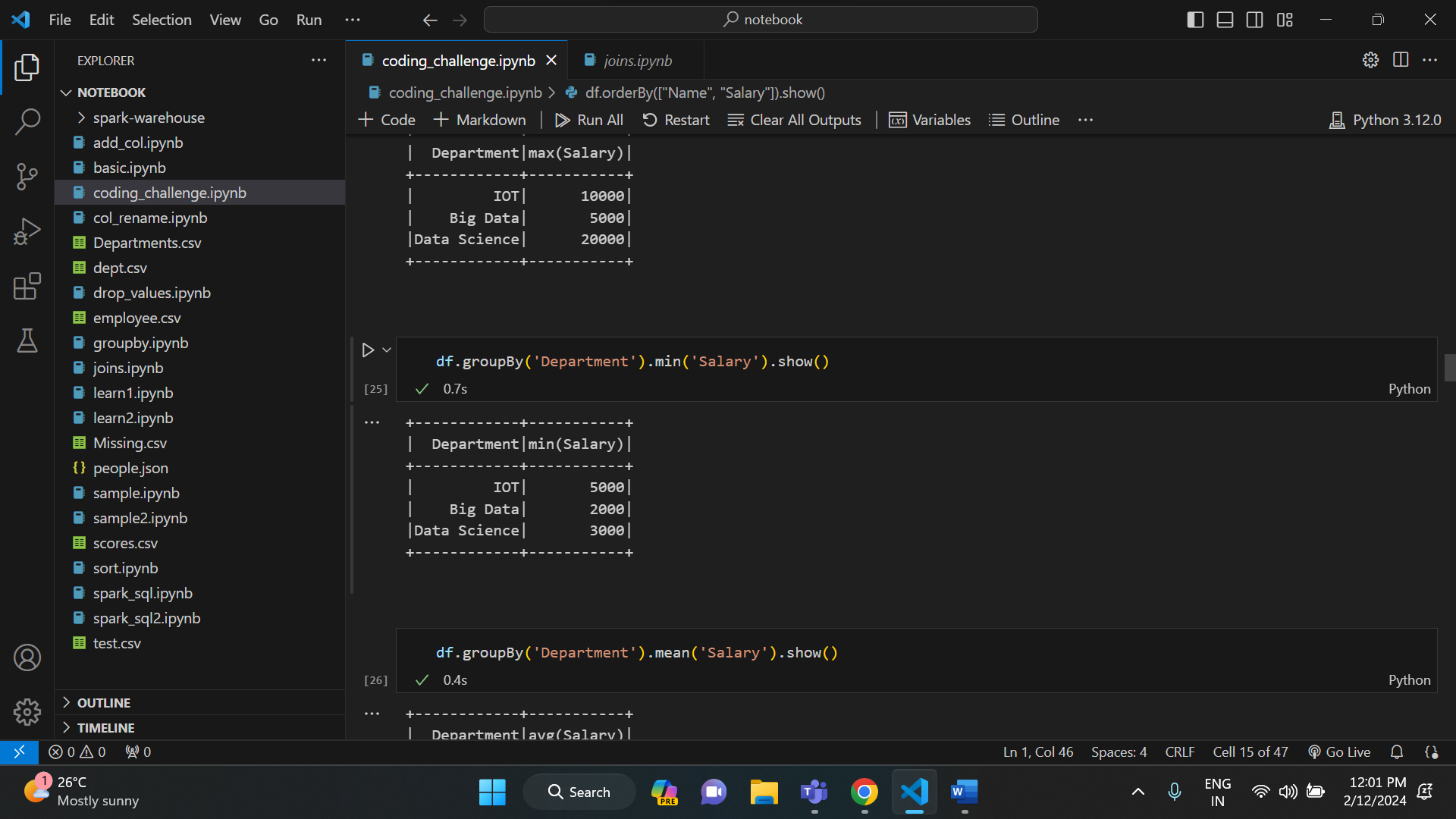
* Avg ()



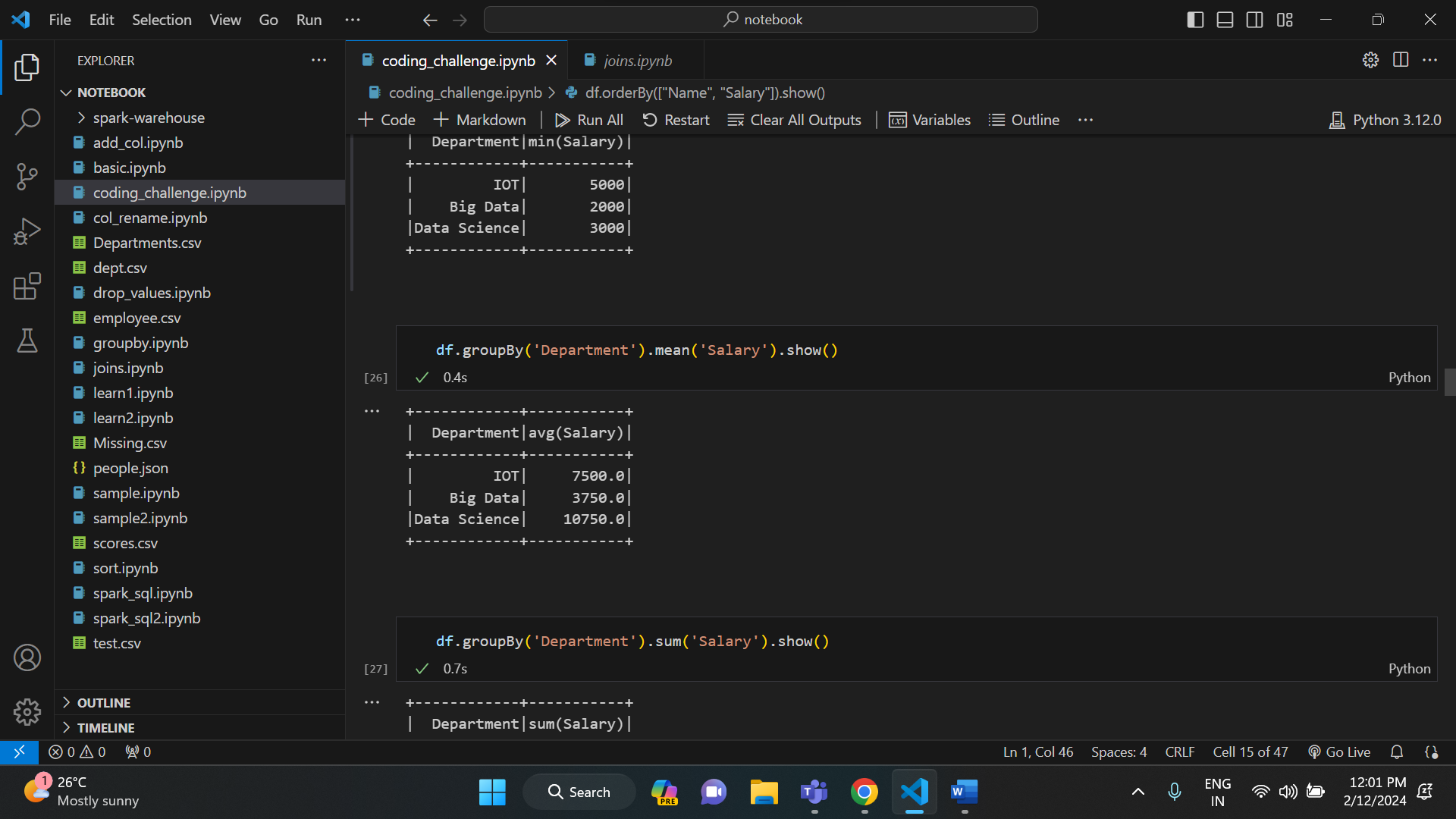
* Max ():



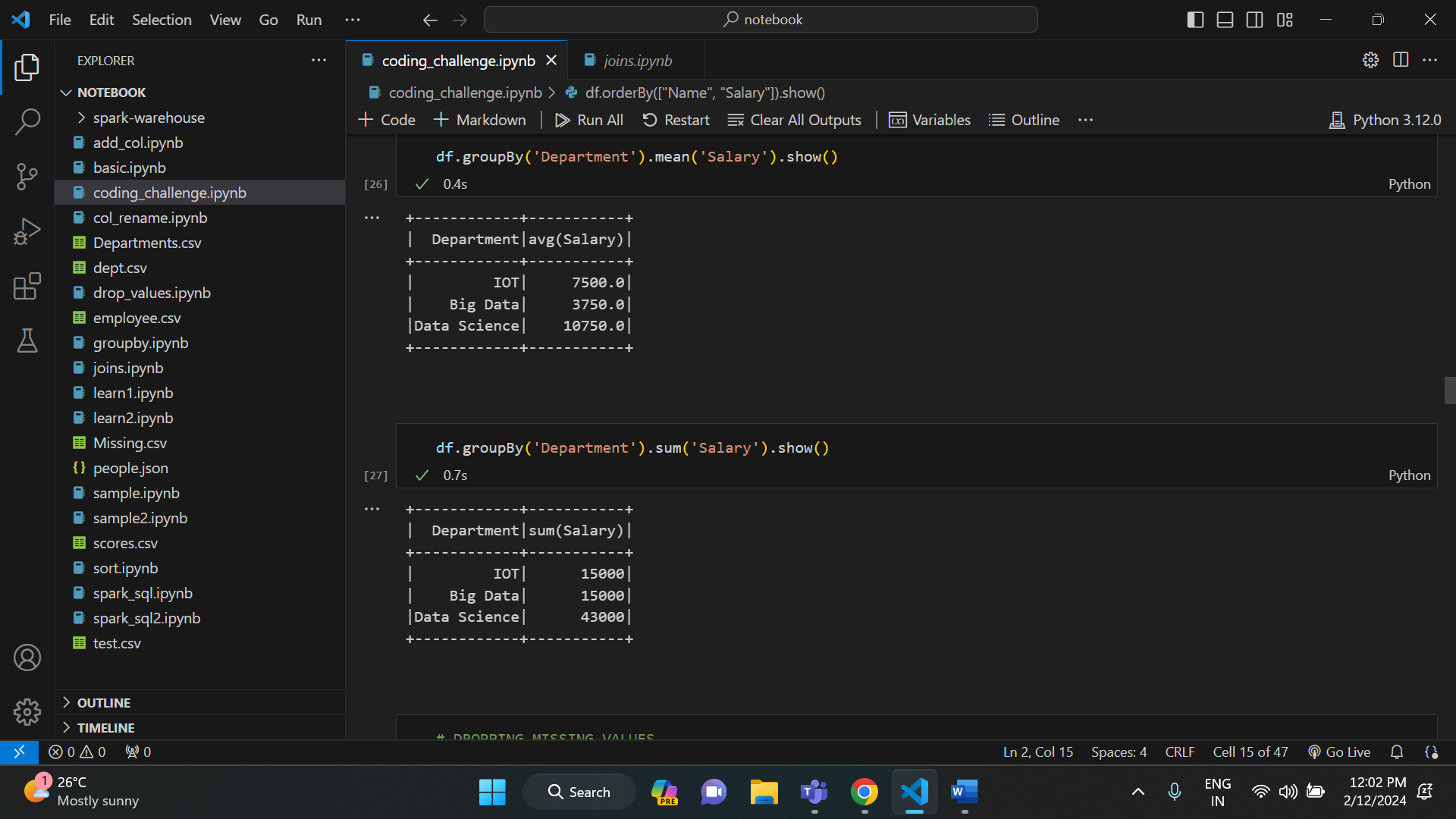
* Min ():



* Mean ():



* Sum ():

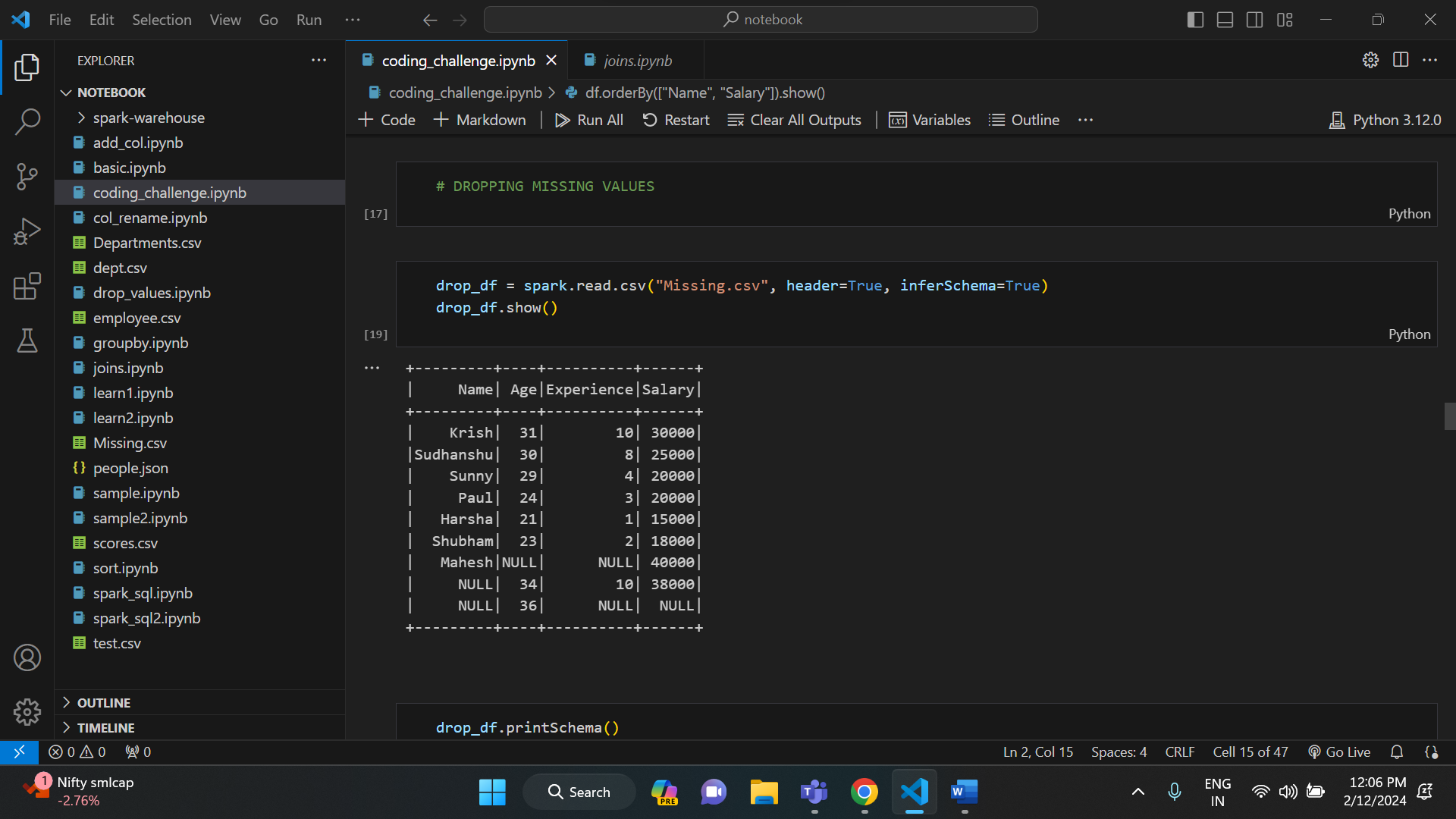


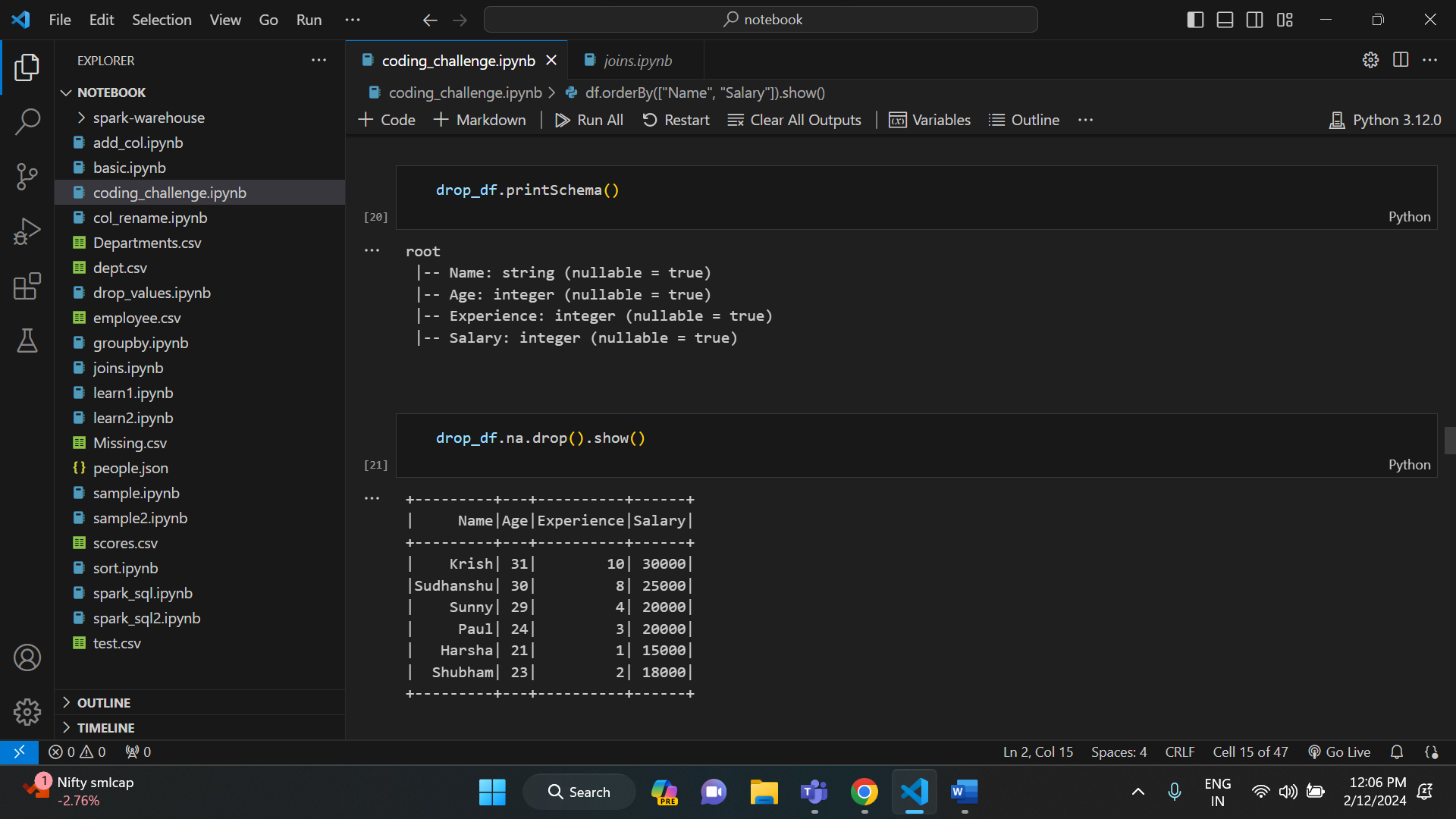
* **DROPPING MISSING VALUES IN PYSPARK:**

If there are any missing values in the dataset, then we can handle them using some functions in pyspark. We have functions like drop () to drop all the rows which has missing values and fill () to fill all the missing rows.

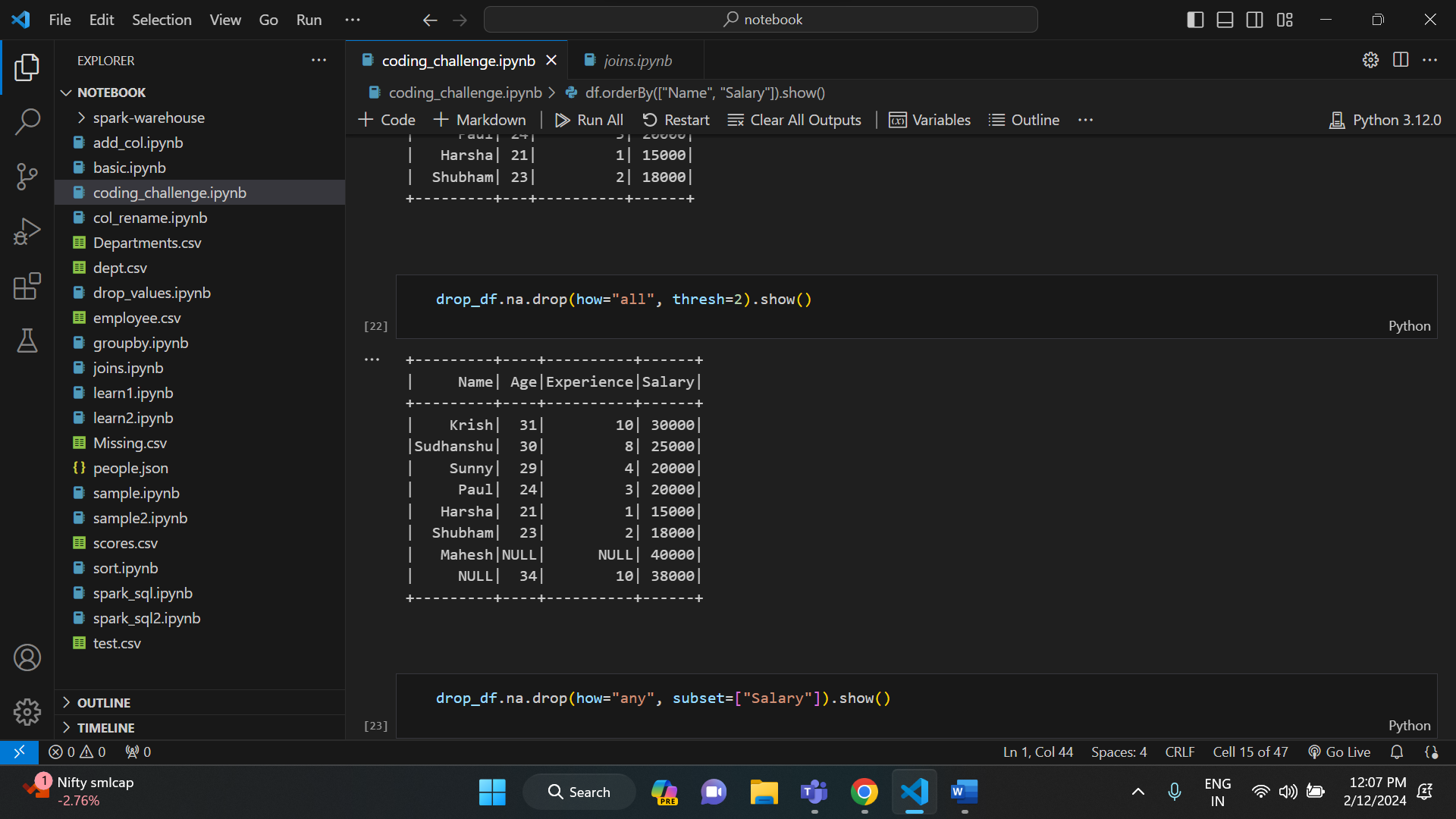
Execution of dropping values:

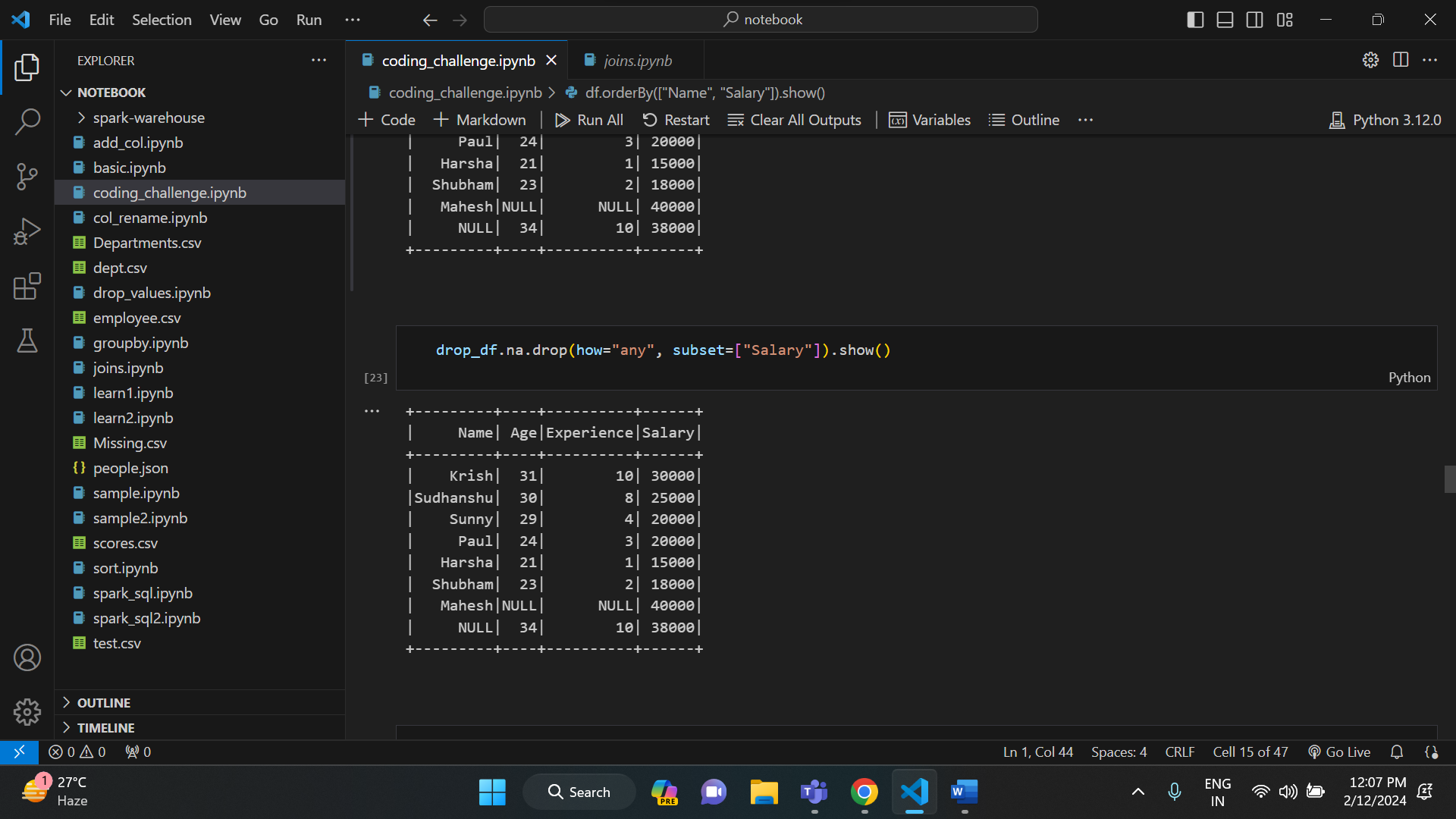
For the execution of drop and fill functions, I have used different dataset called Missing.csv. It has some missing values in the rows.





We have some parameters in drop () function like ‘how=”’, ‘thresh=’ etc.

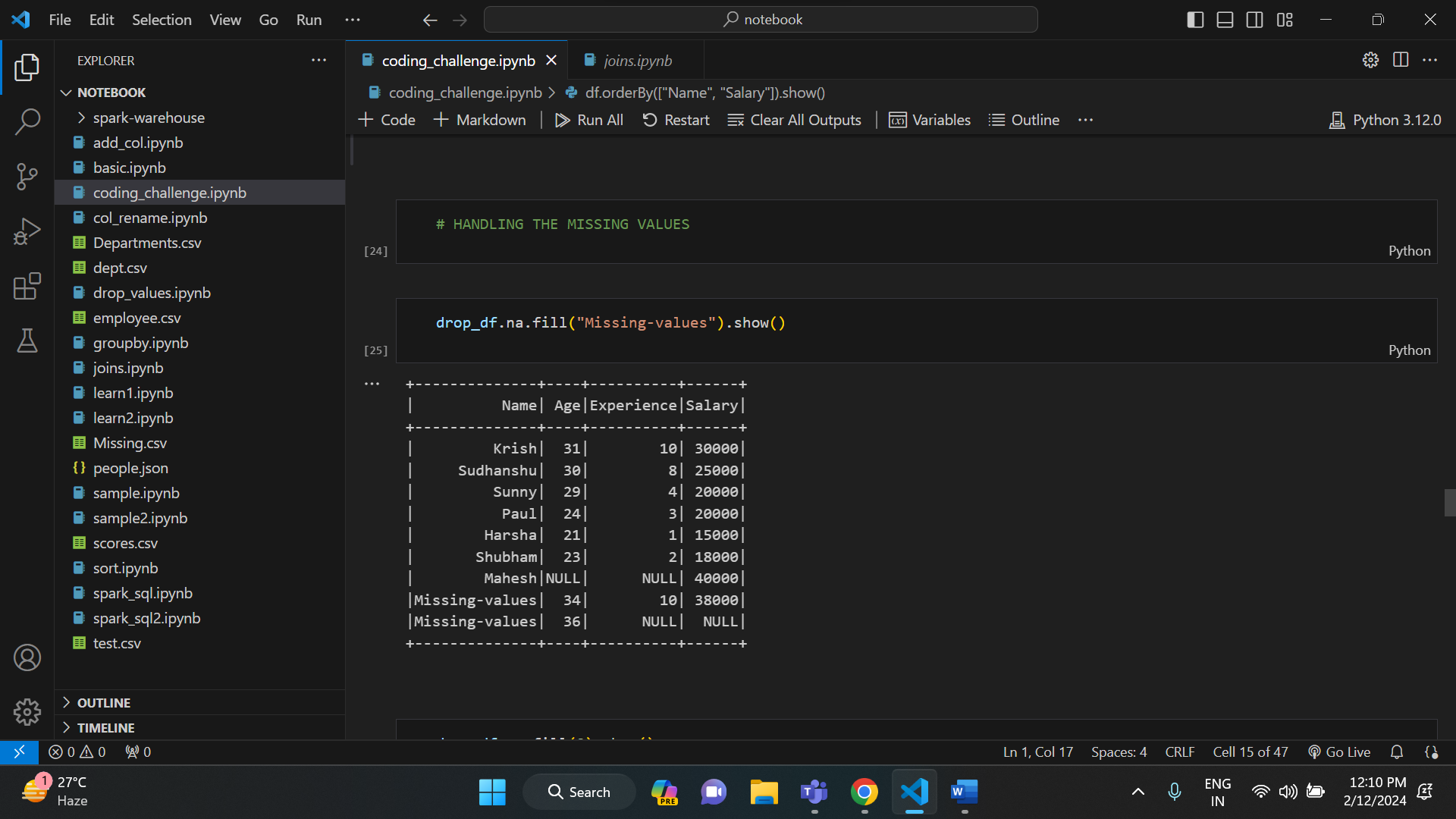




* **HANDLING MISSING VALUES IN PYSPARK:**

In order to handle the missing values, we have a function called fill (). It is used to fill the missing values in rows.

Execution:



Here, the value inside the fill () function is string. So, this fill () function fills the missing values with string type.

In order to fill the missing values with integer type, we need to pass the integer value.



* **JOINS IN PYSPARK:**

Similar to SQL joins, pyspark also has join in it. Joins in pyspark are used to join different dataframes and get the required output out of it. We have joins like inner, outer, left, right, left-semi, left-anti joins in pyspark. All the type of joins are discussed in second question.

Sample Example of join:

