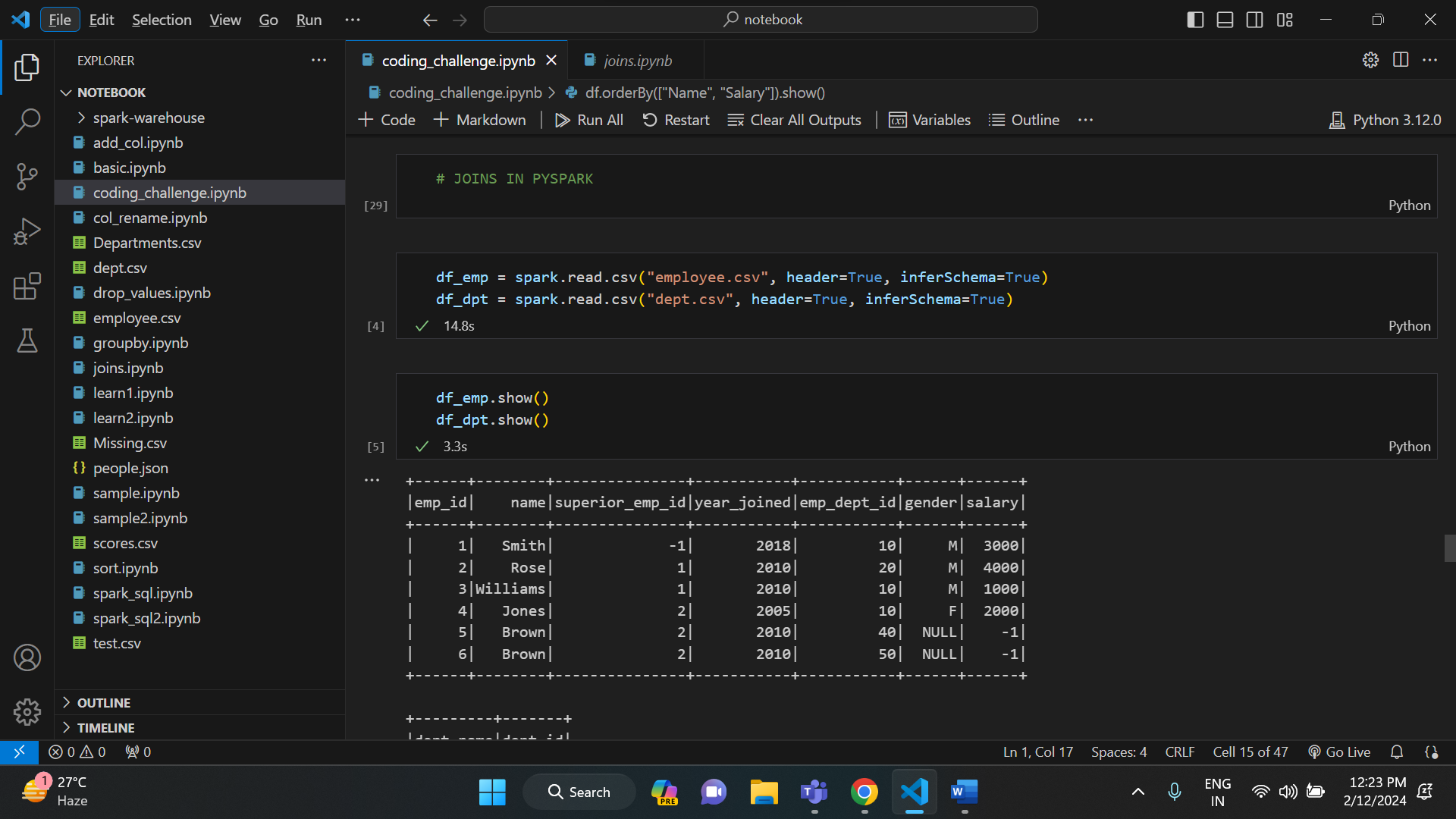
**PYSPARK CODING CHALLENGE QUESTION – 2**

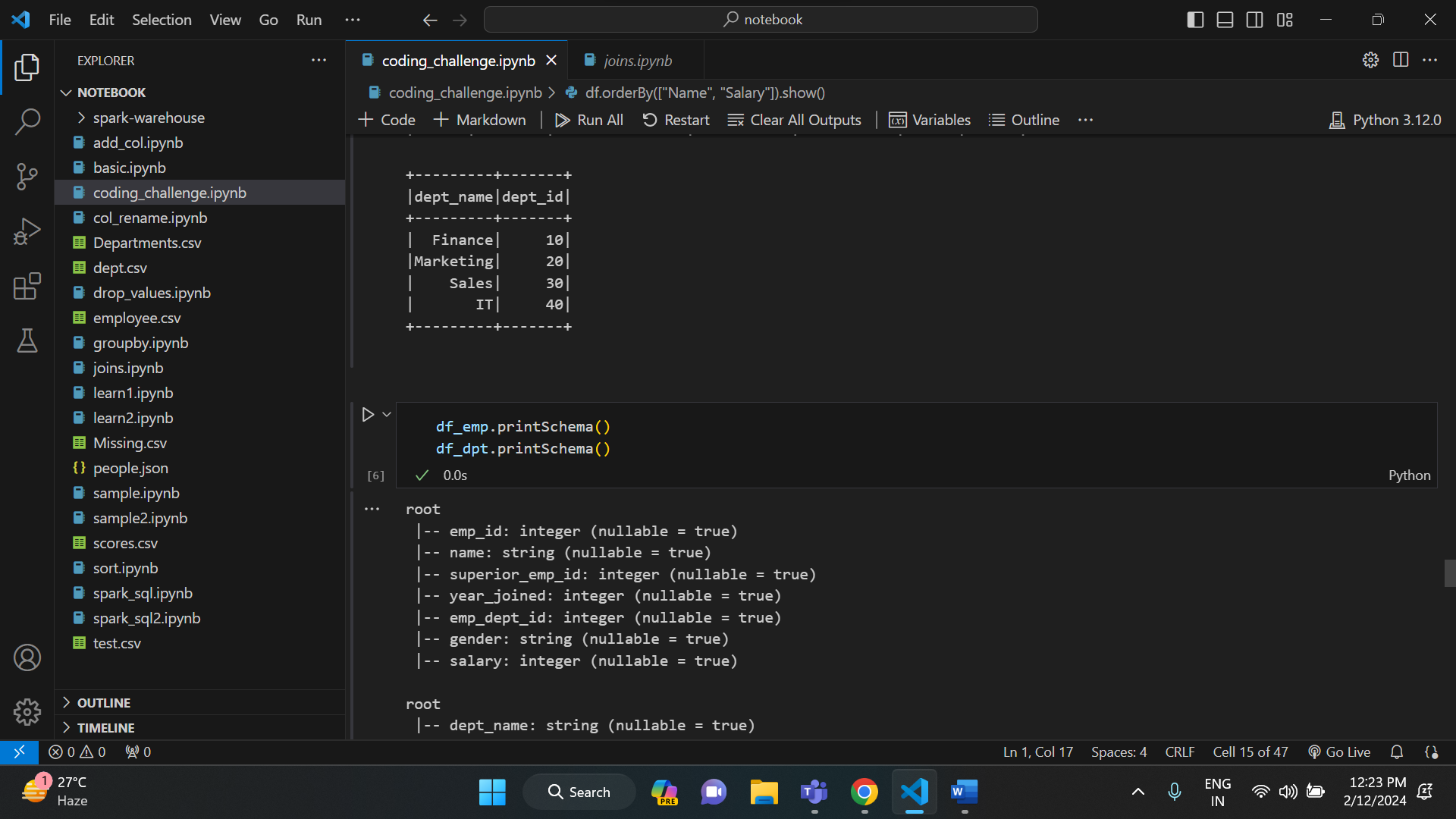
* **PYSPARK SQL JOINS:**

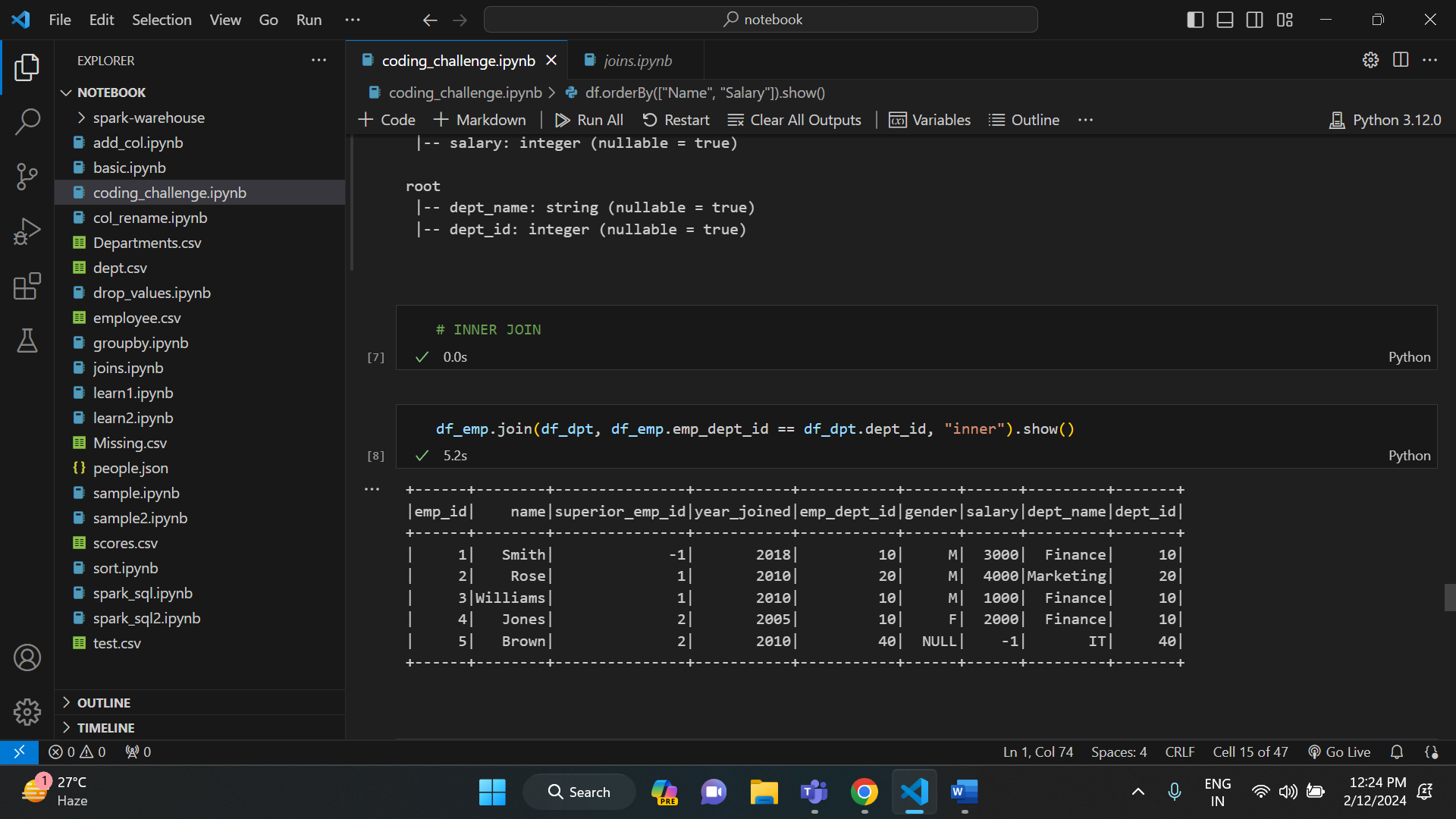
Pyspark SQL joins are similar to joins in SQL. Pyspark joins basically are used to join two dataframes and get the required output out of it. They are very useful when querying large datasets. In pyspark we have wide variety of joins. For the execution of joins, I have taken two csv files. One is employee.csv and other is dept.csv. They are:

Execution of Joins:

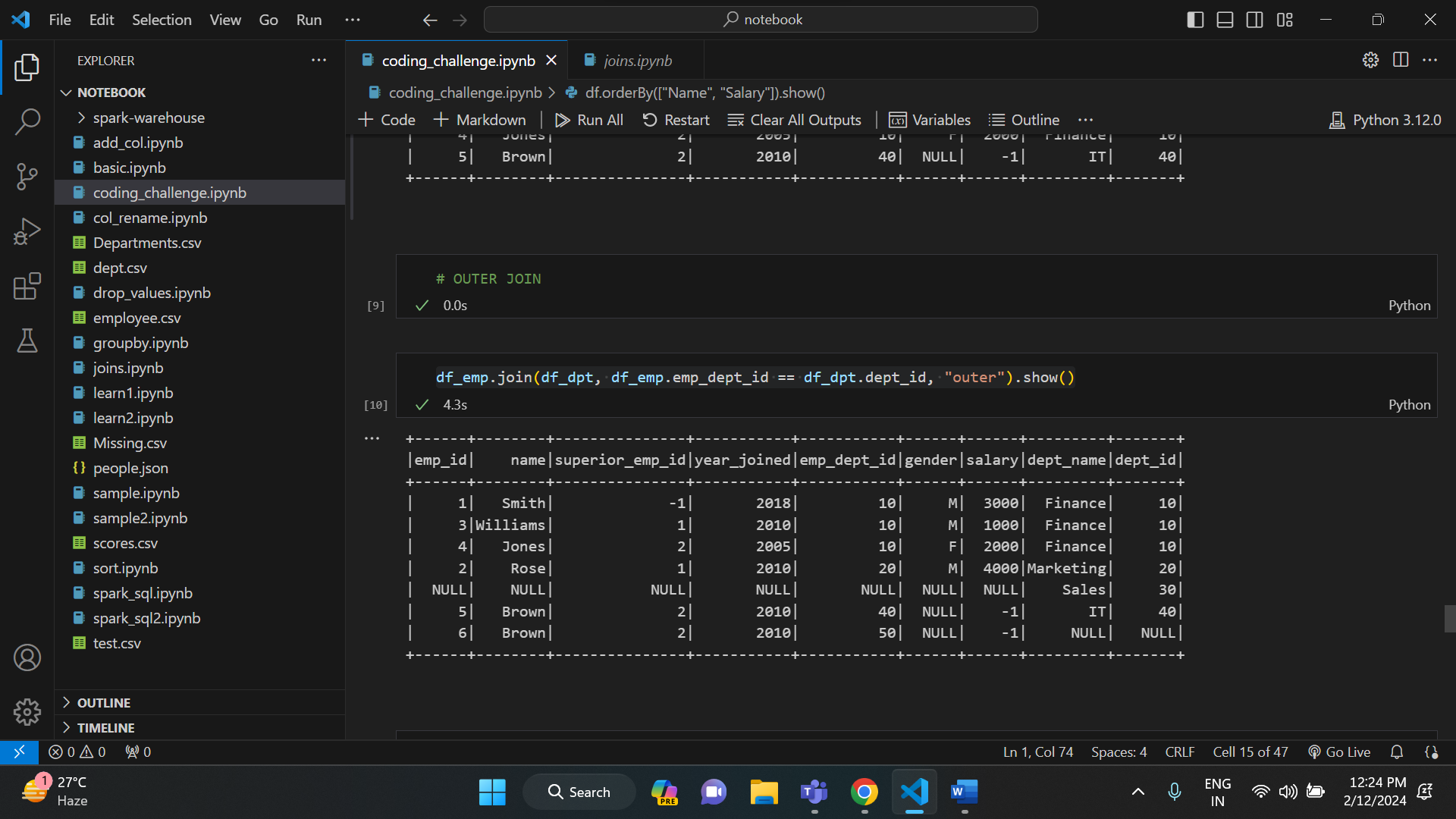
* **Inner Join:** It joins records when key columns are matched and drops when they are not matched.



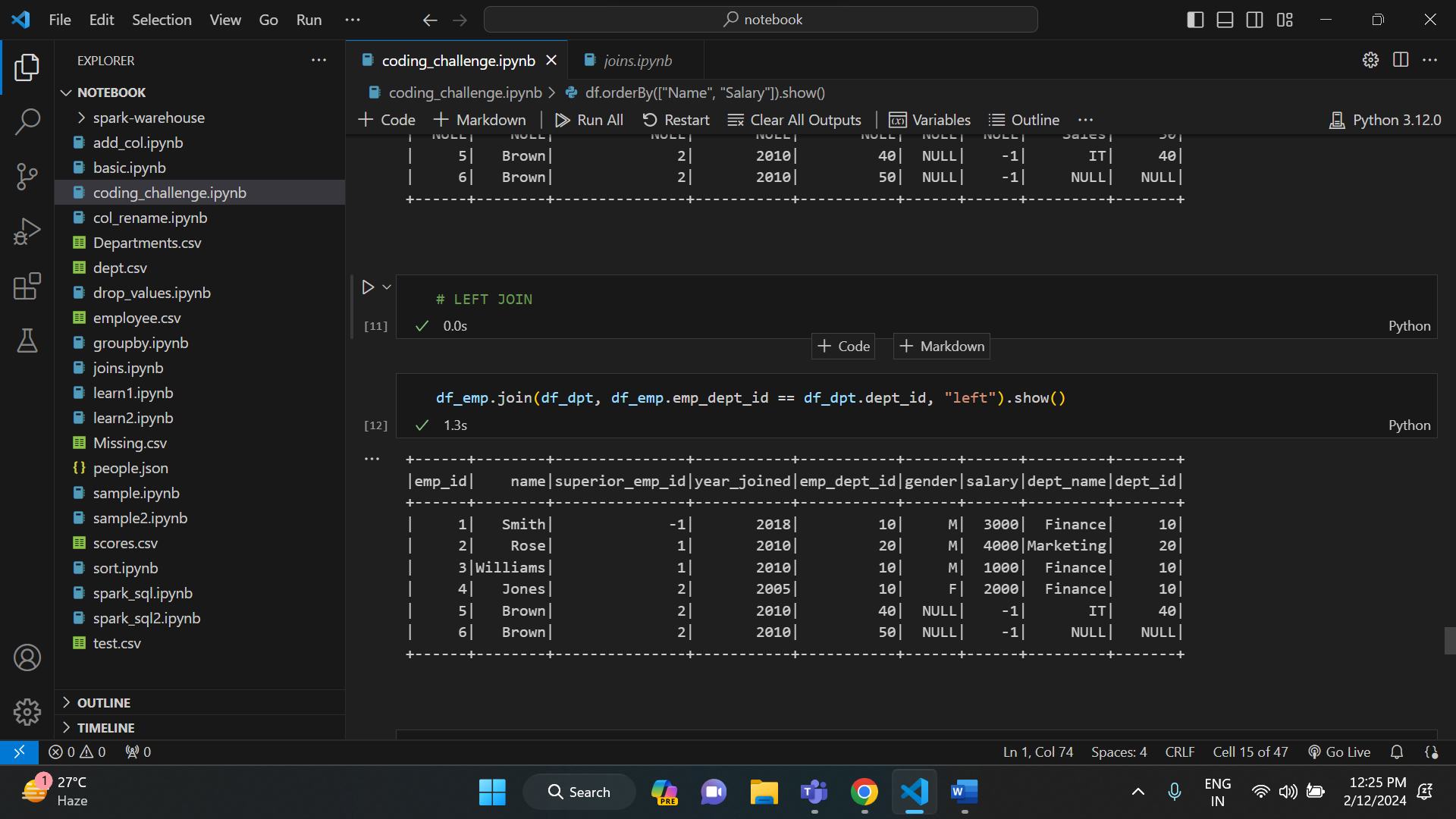




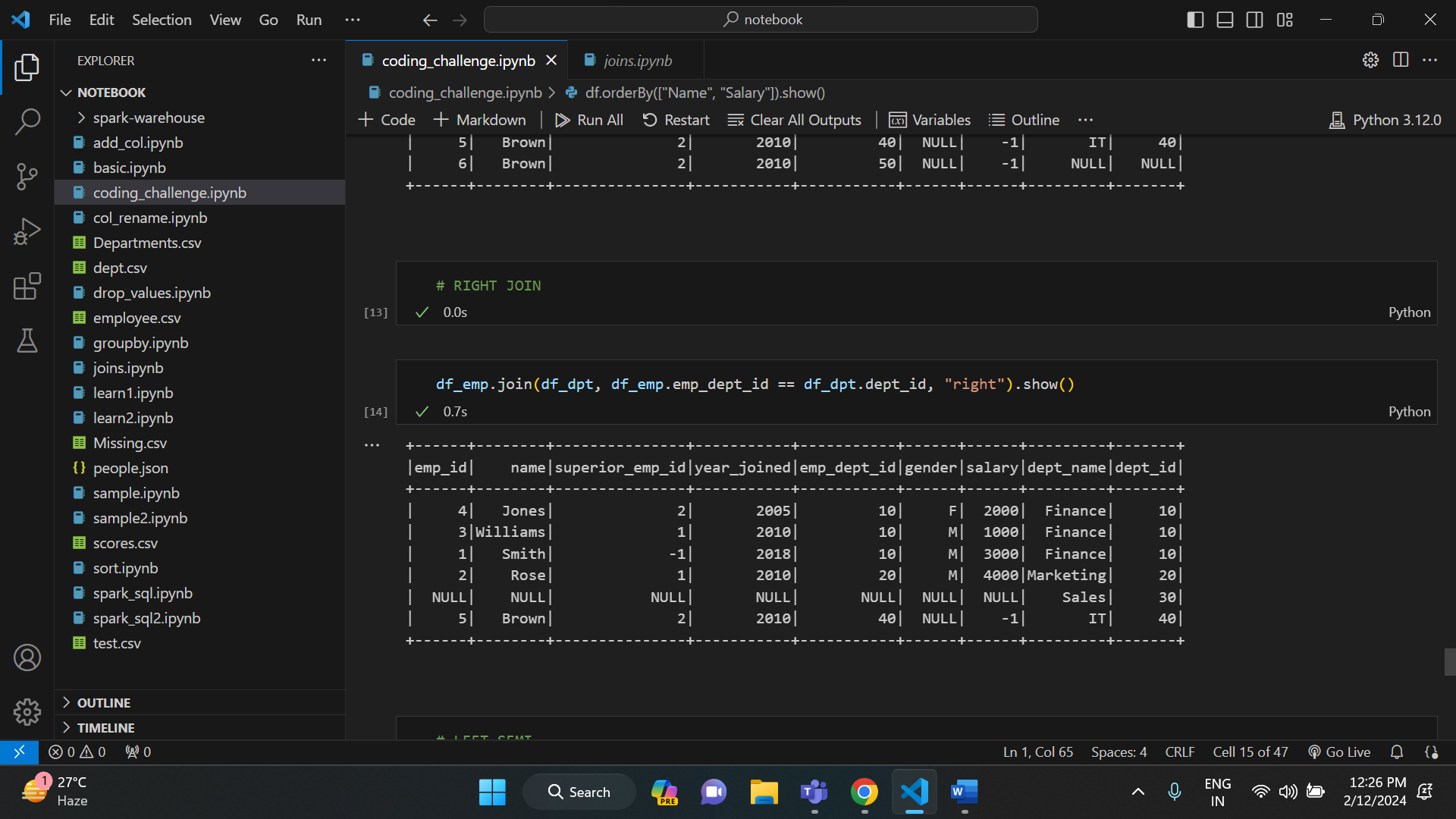
* **Outer join:** It returns all rows from both datasets, where join expression doesn’t match it returns null or respective columns.



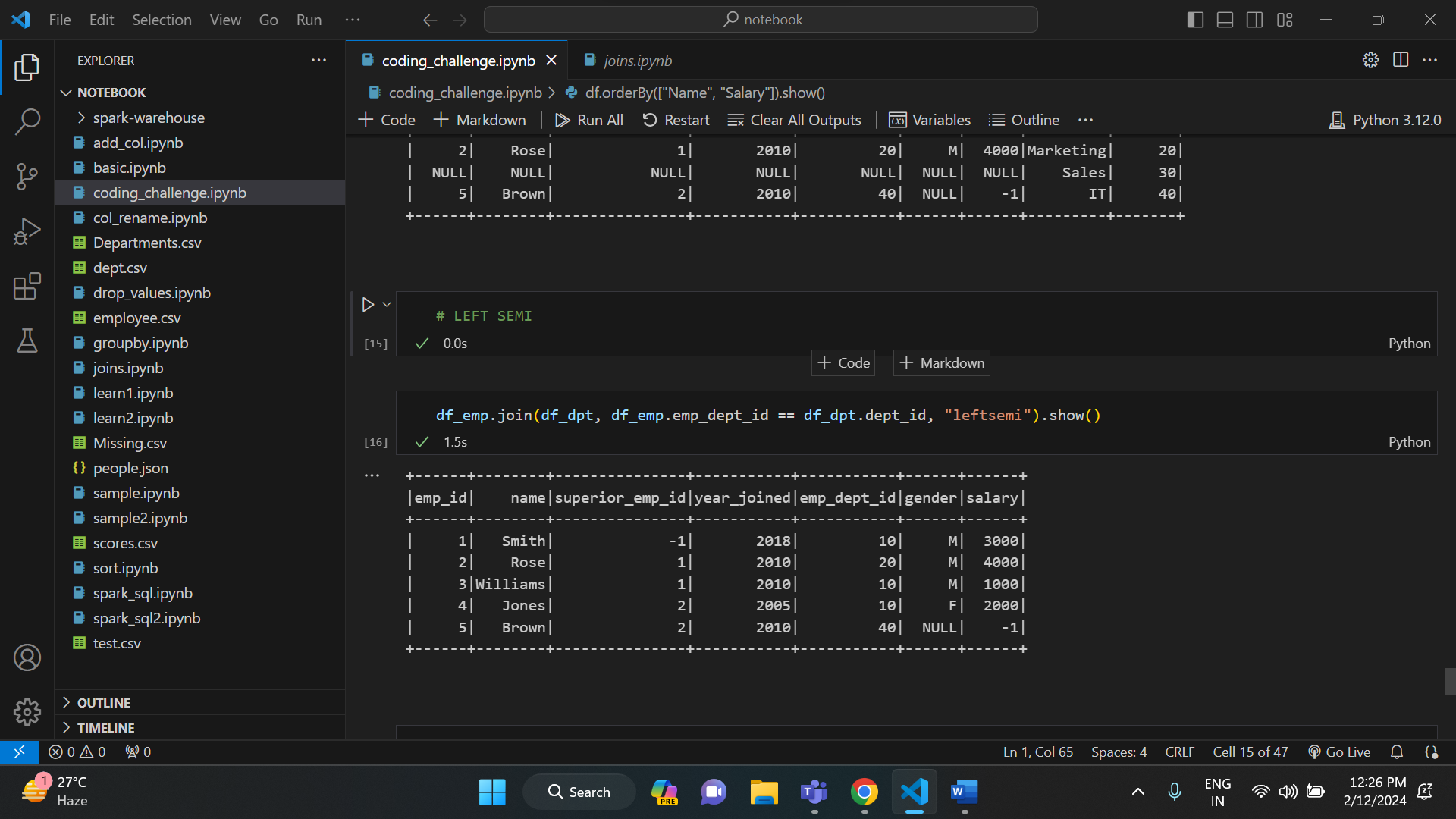
* **Left join:** It returns all rows from left dataset irrespective of match found on right dataset, when join doesn’t matches it assigns null for that record.



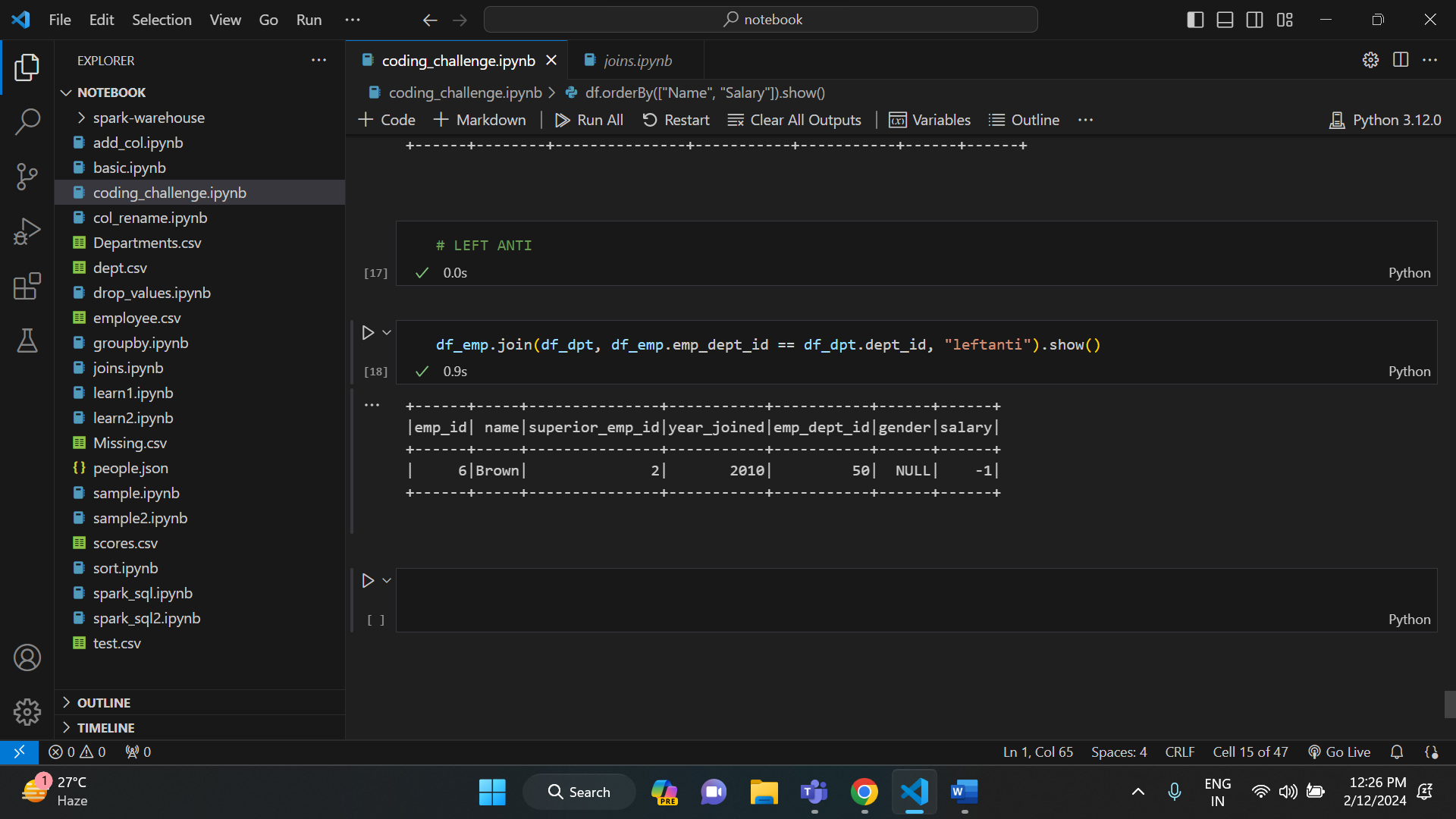
* **Right join:** It returns all rows from right dataset regardless of match found on left dataset. When join doesn’t match it assigns null for that record.



* **Left semi join:** It returns columns from only left dataset for the matched records in the right dataset on join expression.



* **Left anti join**: It returns columns from left dataset for non-matched records.



* **FUNCTIONS IN PANDAS DATAFRAME:**

There are several functions in pandas dataframe. Some of them are like

DataFrame (), head (), tail (), type (), drop () etc.

Below is the execution of some of the functions in pandas dataframe.

Execution:

