PYSPARK ASSESSMENT 2

Q) sparksql joins & Applying Functions in a Pandas DataFrame

**SPARK SQL :** it is a Distributed framework for structured data processing. It can perform better optimization. It is a Spark Module fro structured data processing. It uses same execution engine while computing an output.

**FEATURES OF SPARK SQL:**

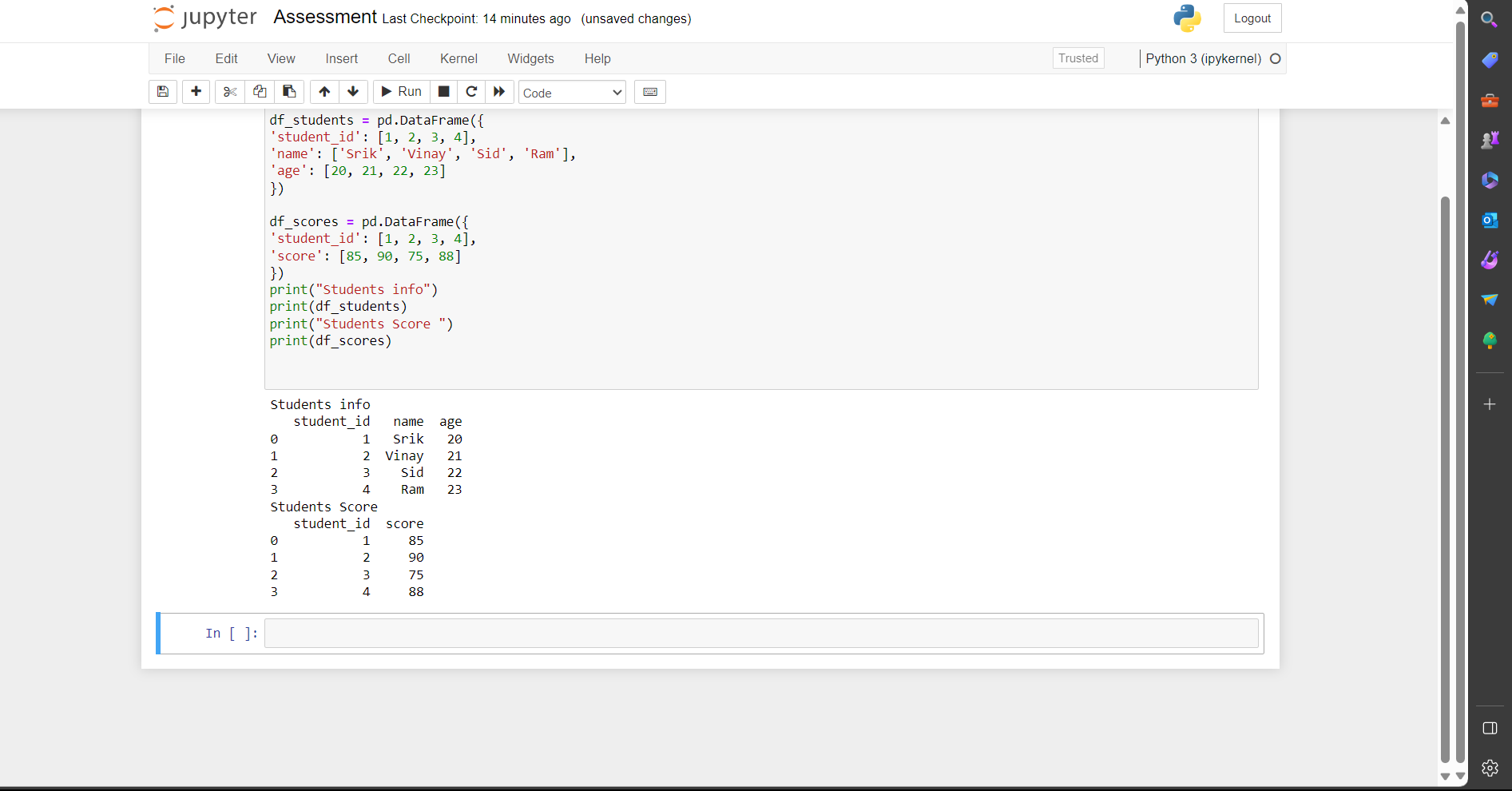
* It is cost based optimizer
* It is Mid query Fault-Tolerance
* It has full compatibility with existing Hive data

Steps to perform joins and Applying functions to the Panda dataframes :

STEP1: First we need to import Panda into the kernel

Step2: We need to add dataframes and these should contain data In it.

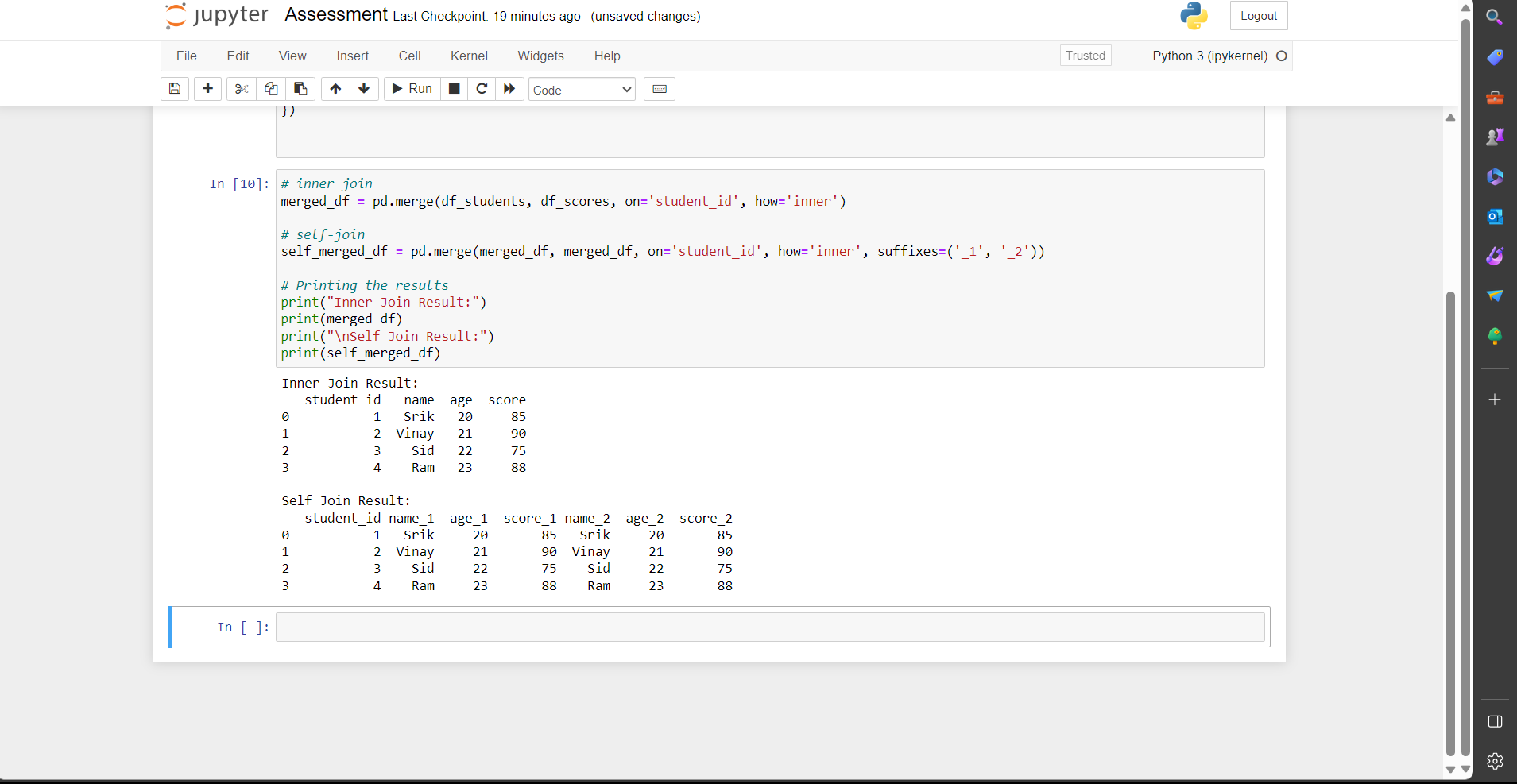
Step3: Lets display the dataframes we have designed.



Now we are performing “INNER JOIN AND SELF JOIN” for the dataframes and executing them :

We perform an inner join between “df\_students” and “df\_scores” to merge the data and

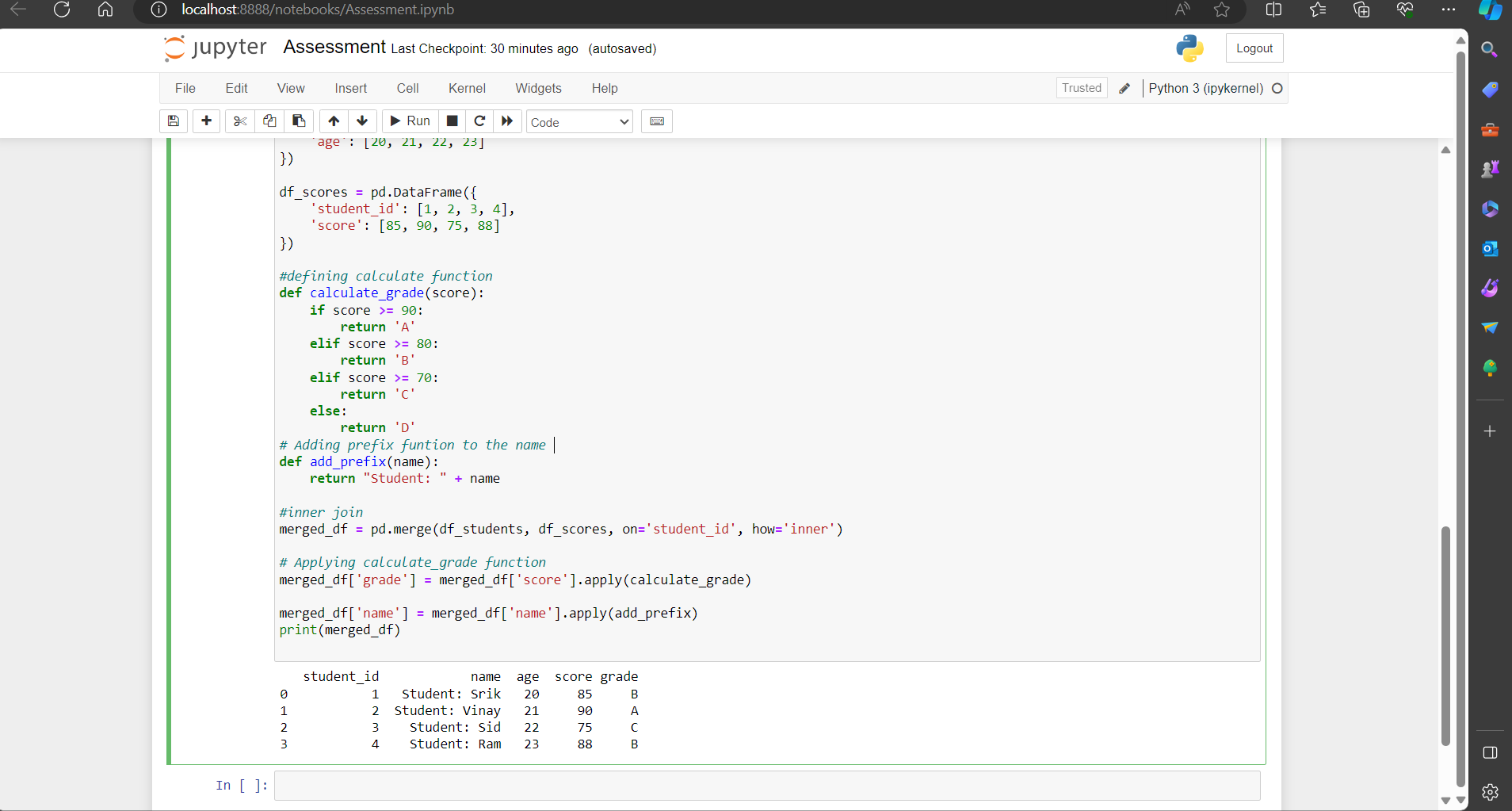
We perform a self-join on the merged DataFrame to join it with itself based on the 'student\_id' column.



Now lets use function to calculate grade and we also use adding prefix to the name and lets display them :

“calculate\_grade” function to calculate grades based on scores.

“add\_prefix” function to add a prefix to student names.



Here we added another function which is “double\_score” to display the scores of the students which is doubled using this function and displaying that.

