## Module 22 home\_sales challenge - screen shots v1

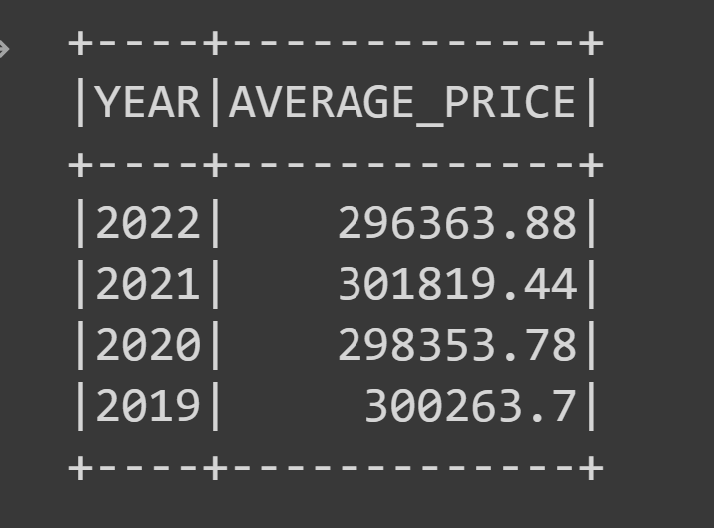
Raj Agrawal / SMU DS / Sep 2023

===

**Challenge –** To use, SparkSQL to determine key metrics about home sales data. Then use Spark to create temporary views, partition the data, cache and un-cache a temporary table, and verify that the table has been un-cached.

**## Deliverable :**

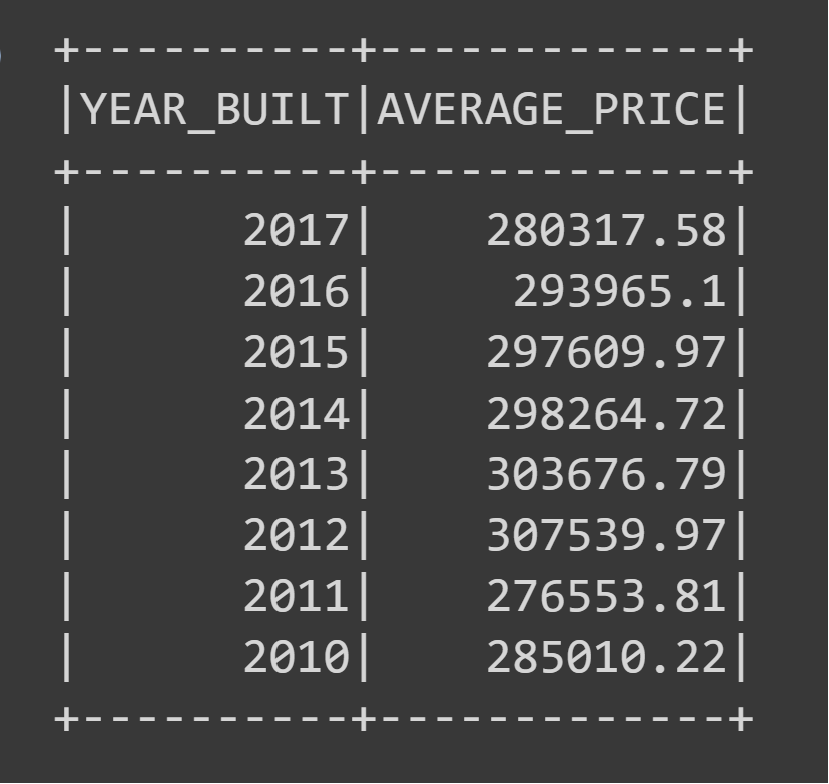
1. Rename the Home\_Sales\_starter\_code.ipynb file as **Home\_Sales.ipynb.**
2. Import the necessary PySpark SQL functions for this assignment.
3. Read the home\_sales\_revised.csv data in the starter code into a Spark DataFrame.
4. Create a temporary table called home\_sales.
5. Answer the following questions using SparkSQL:
   * What is the average price for a four-bedroom house sold for each year? Round off your answer to two decimal places.



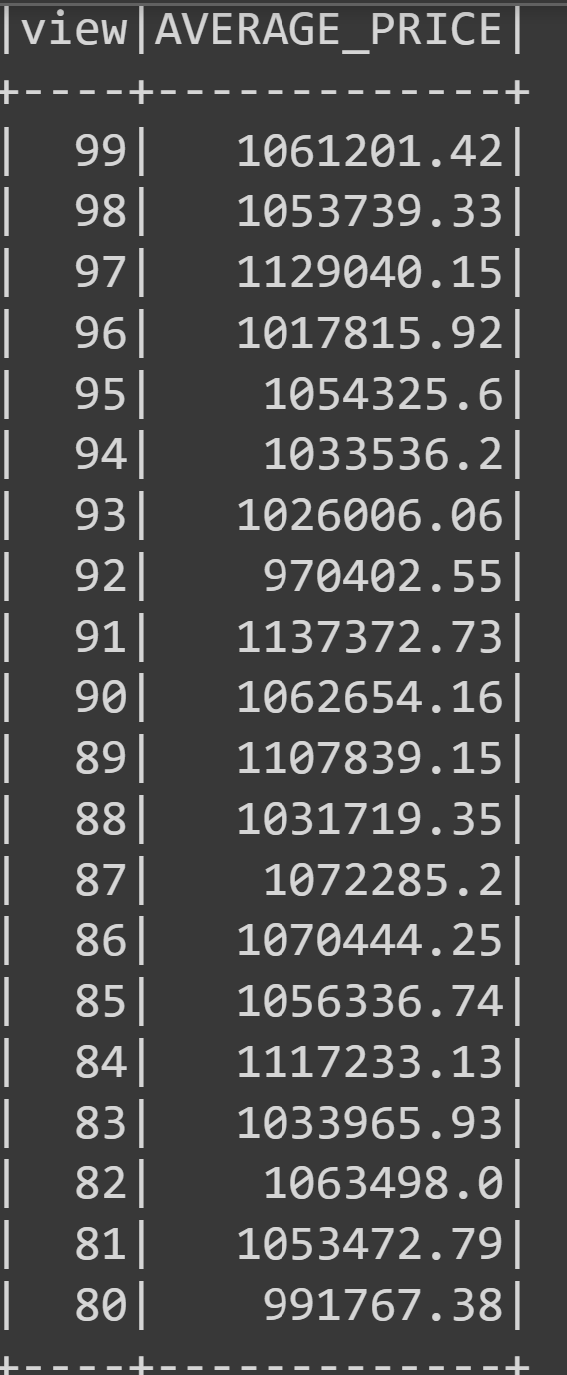
* + What is the average price of a home for each year it was built that has three bedrooms and three bathrooms? Round off your answer to two decimal places.



* + What is the average price of a home for each year that has three bedrooms, three bathrooms, two floors, and is greater than or equal to 2,000 square feet? Round off your answer to two decimal places.



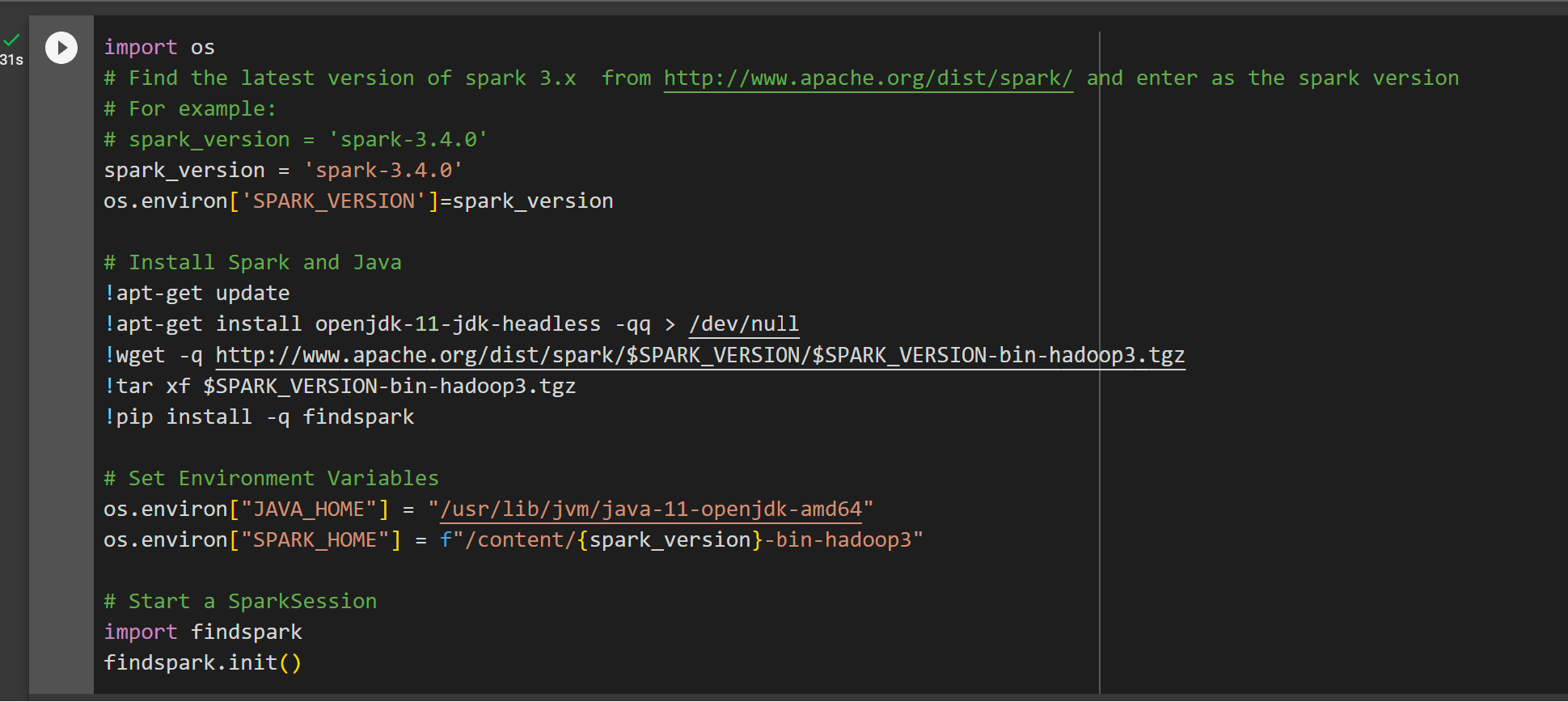
* + What is the "view" rating for homes costing more than or equal to $350,000? Determine the run time for this query, and round off your answer to two decimal places.

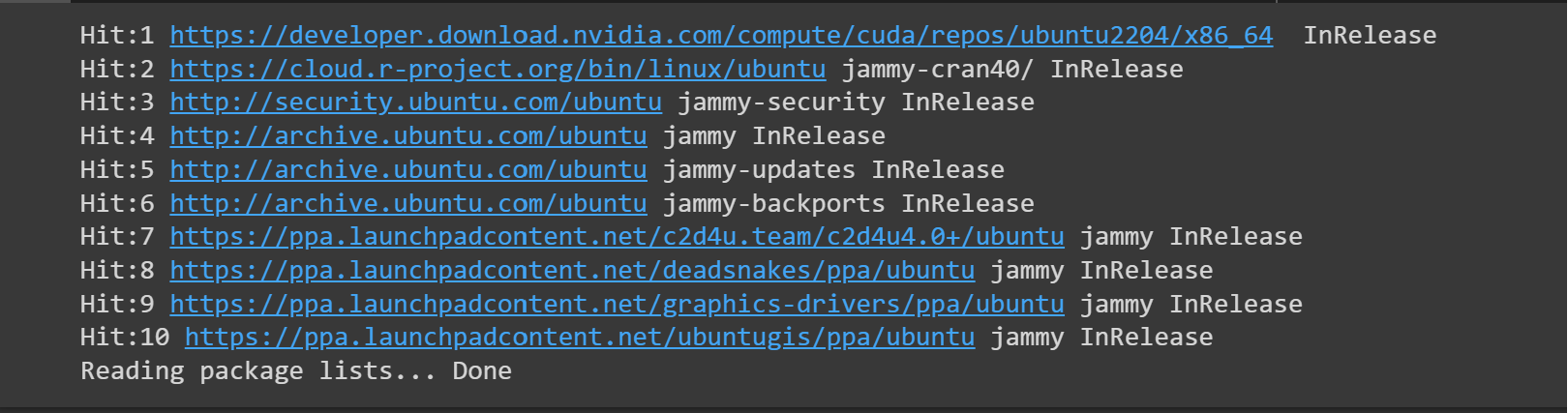


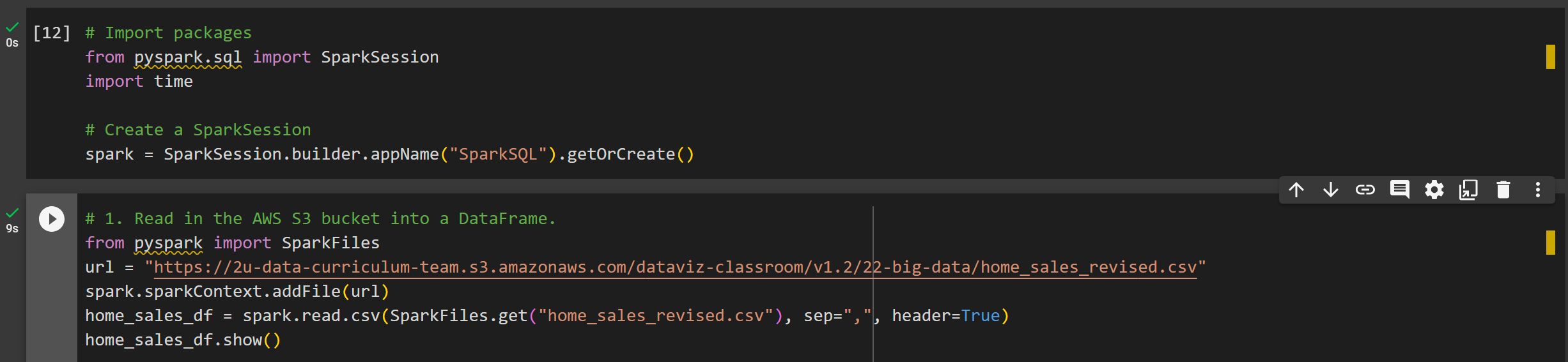
==================<<pls see below screen-shot for the answer>>>===========

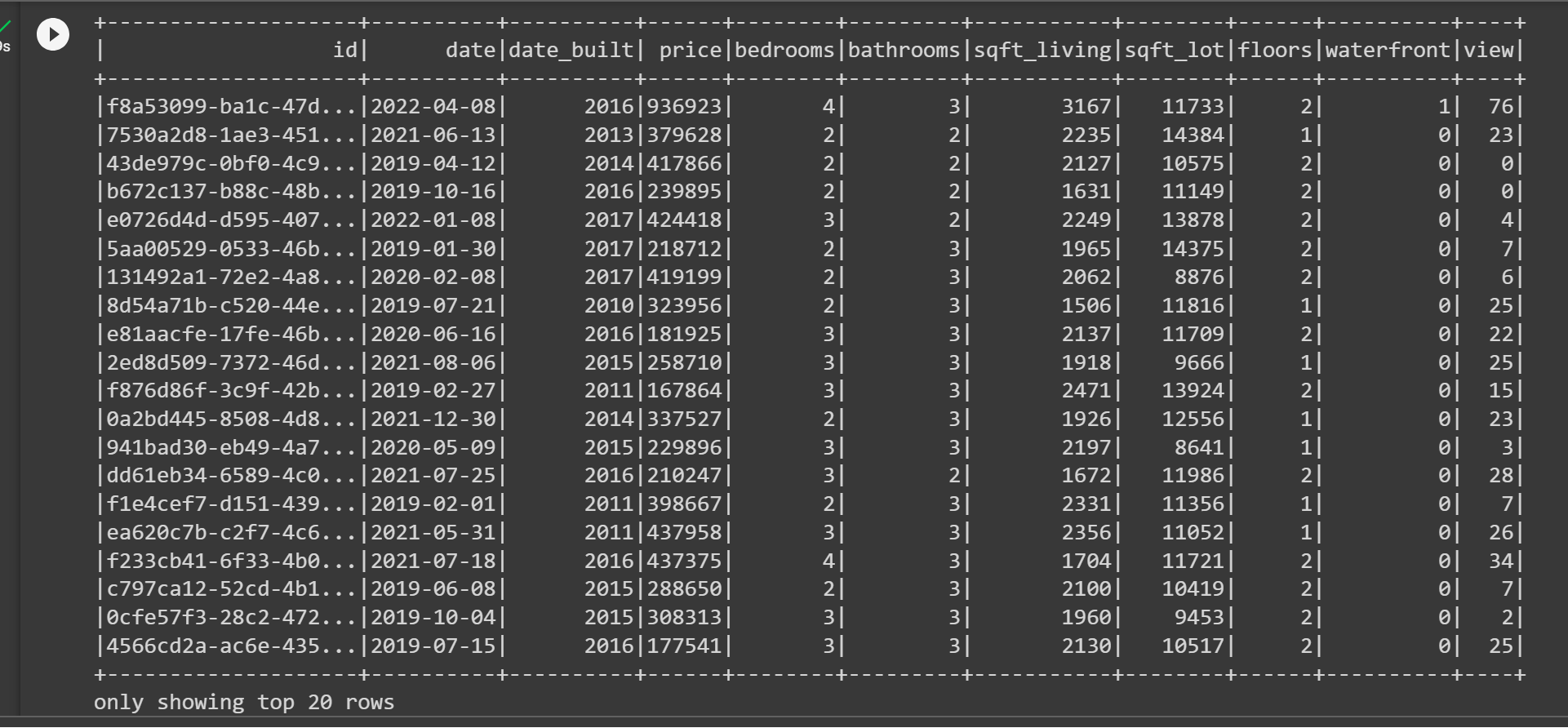
1. Cache your temporary table home\_sales.
2. Check if your temporary table is cached.
3. Using the cached data, run the query that filters out the view ratings with an average price of greater than or equal to $350,000. Determine the runtime and compare it to uncached runtime.
4. Partition by the "date\_built" field on the formatted parquet home sales data.
5. Create a temporary table for the parquet data.
6. Run the query that filters out the view ratings with an average price of greater than or equal to $350,000. Determine the runtime and compare it to uncached runtime.
7. Uncache the home\_sales temporary table.
8. Verify that the home\_sales temporary table is uncached using PySpark.
9. Download your Home\_Sales.ipynb file and upload it into your "Home\_Sales" GitHub repository.

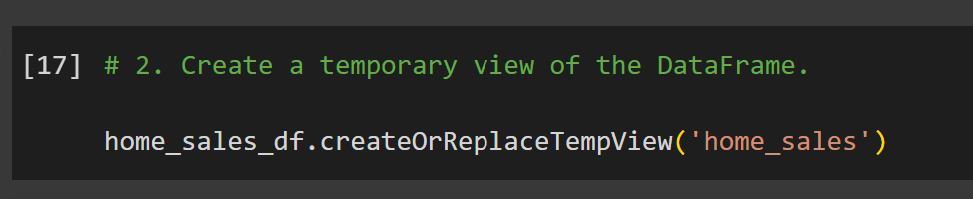
<<pls see below screen-shot for the answer>>>

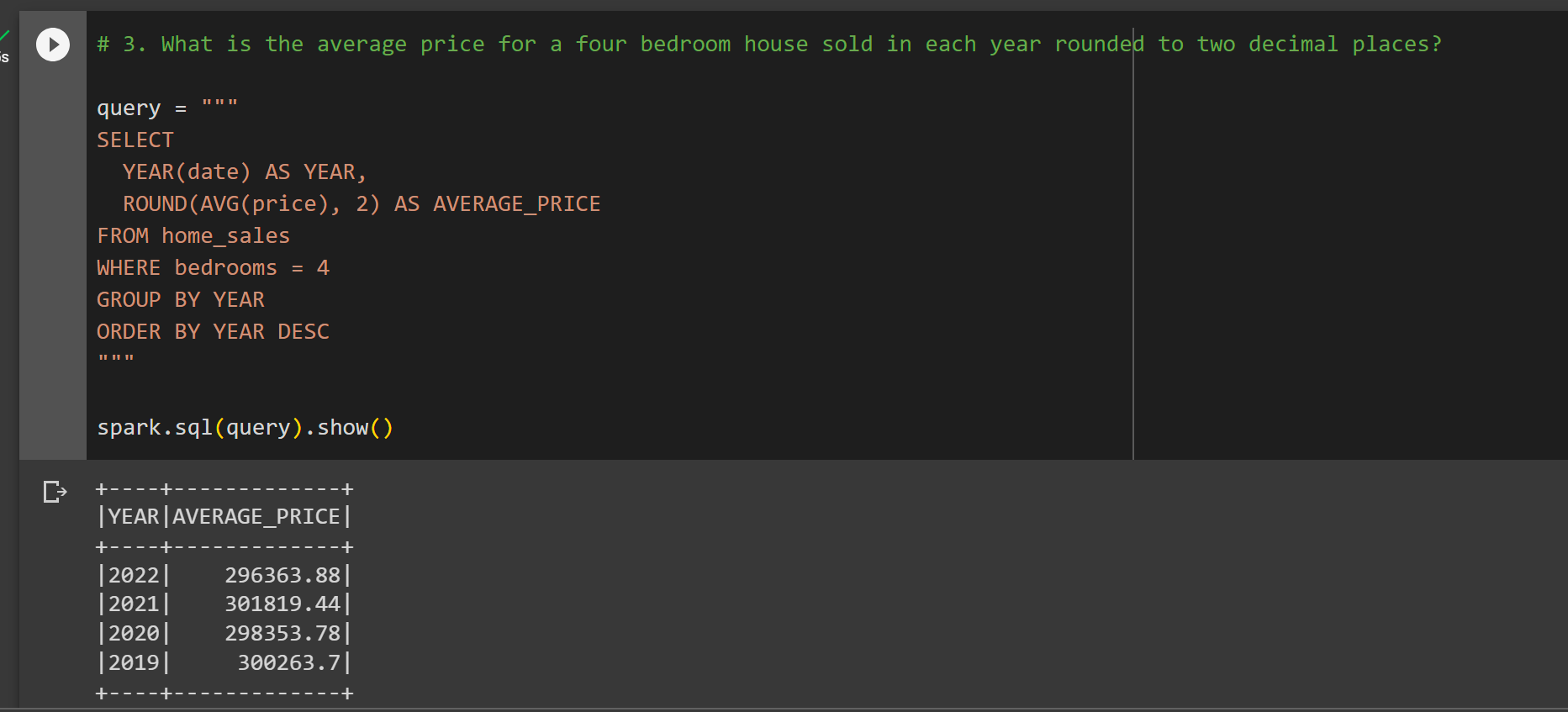
****

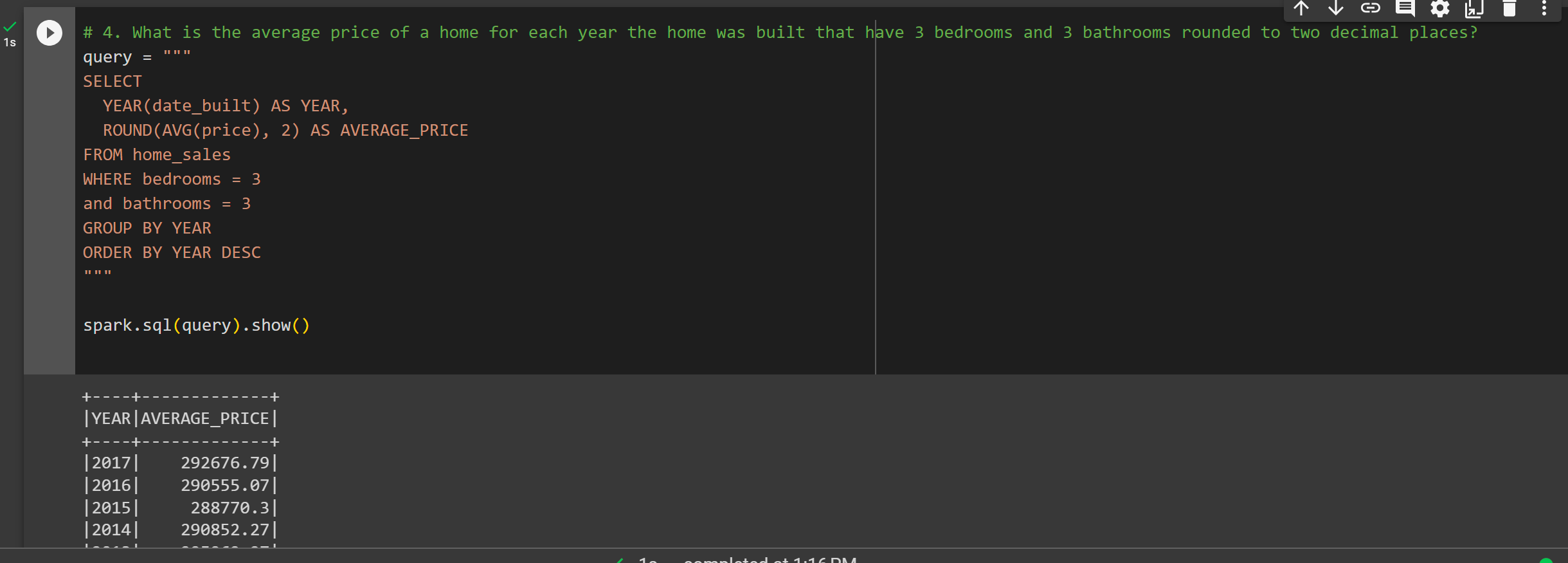
****

****

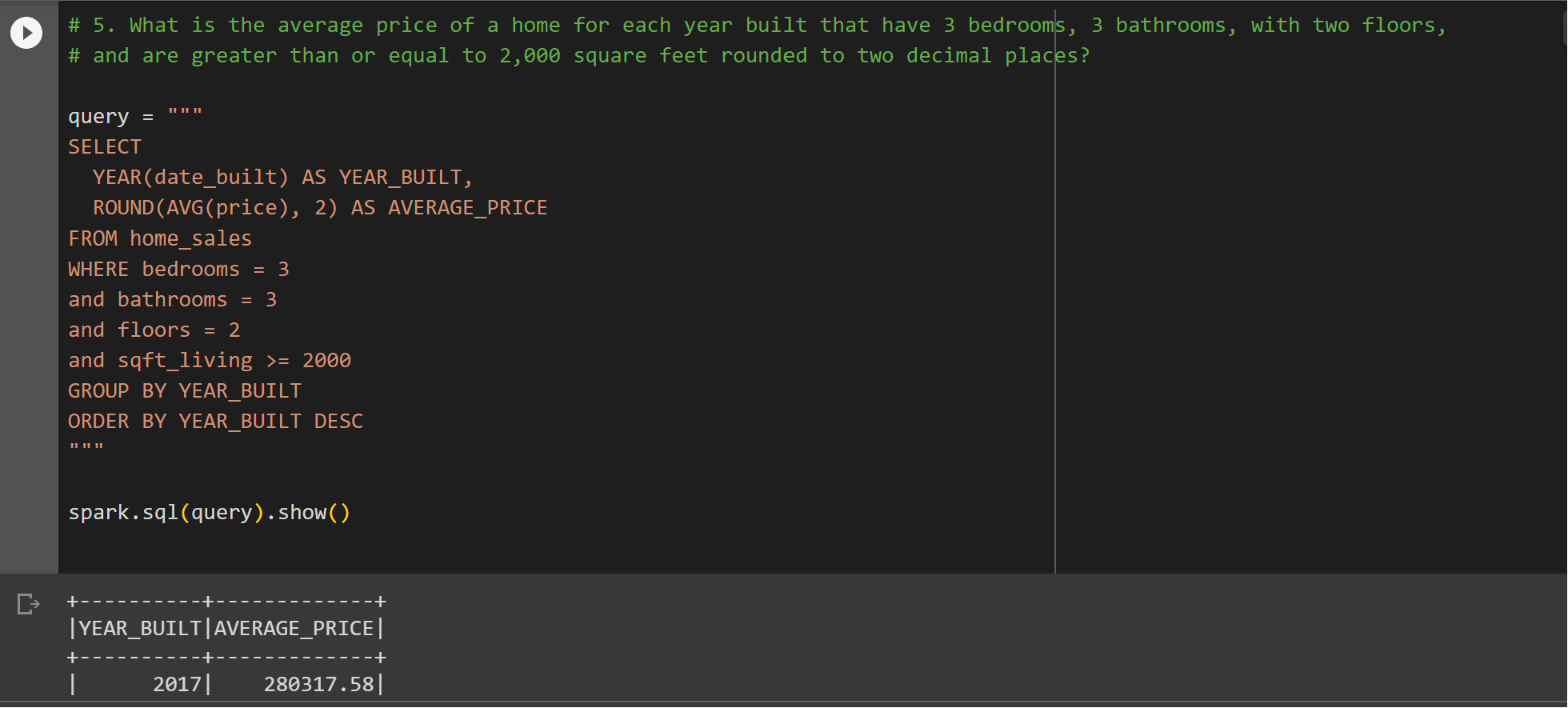
****

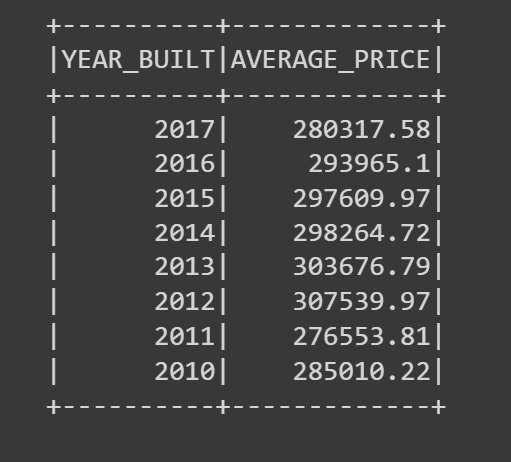
****

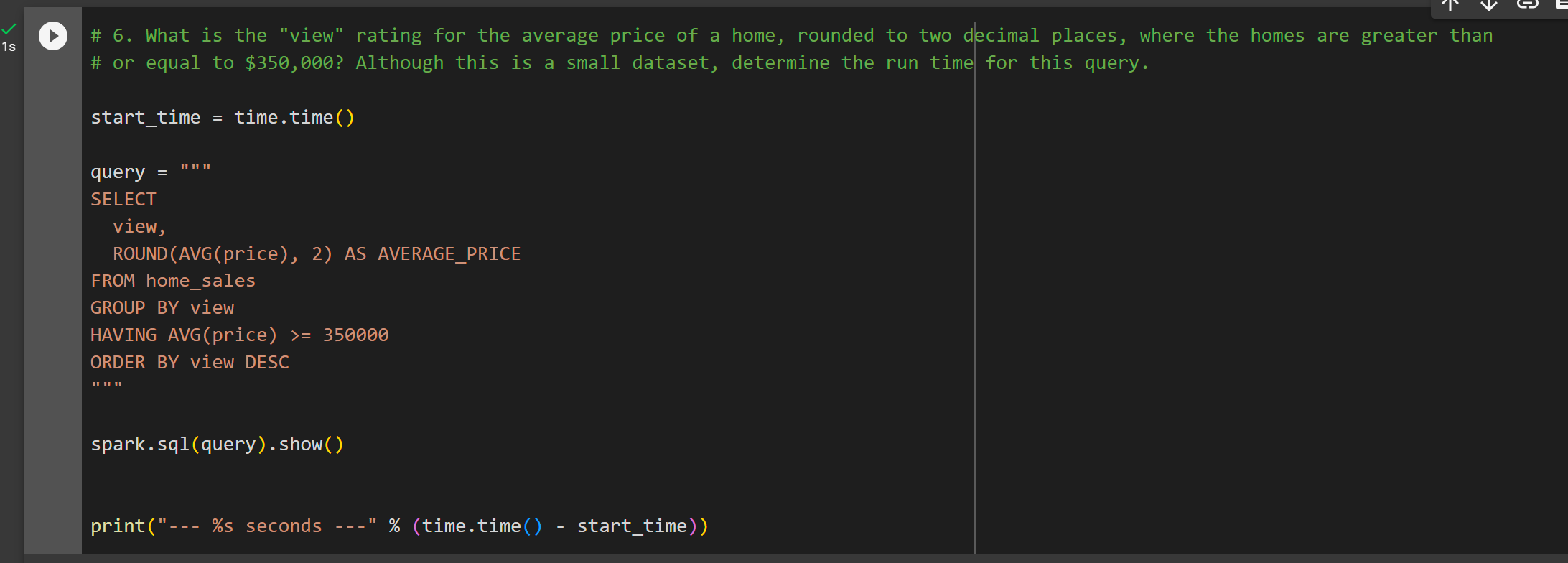
****

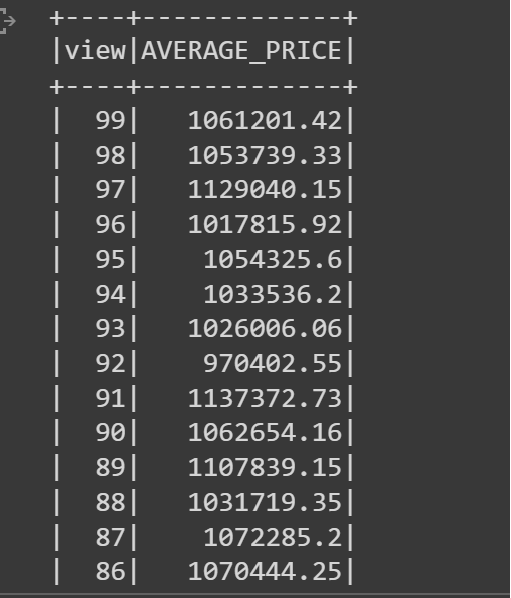
****

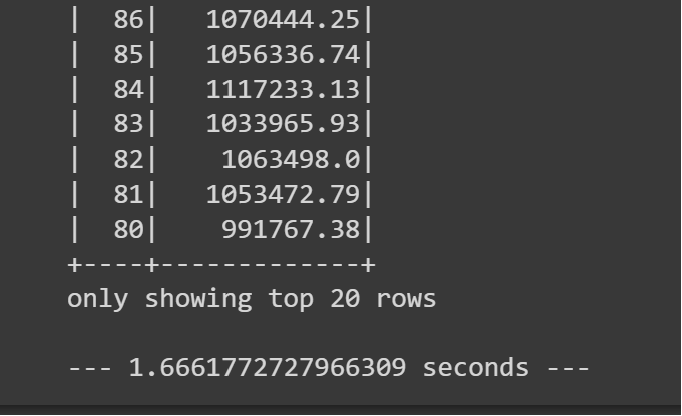
****

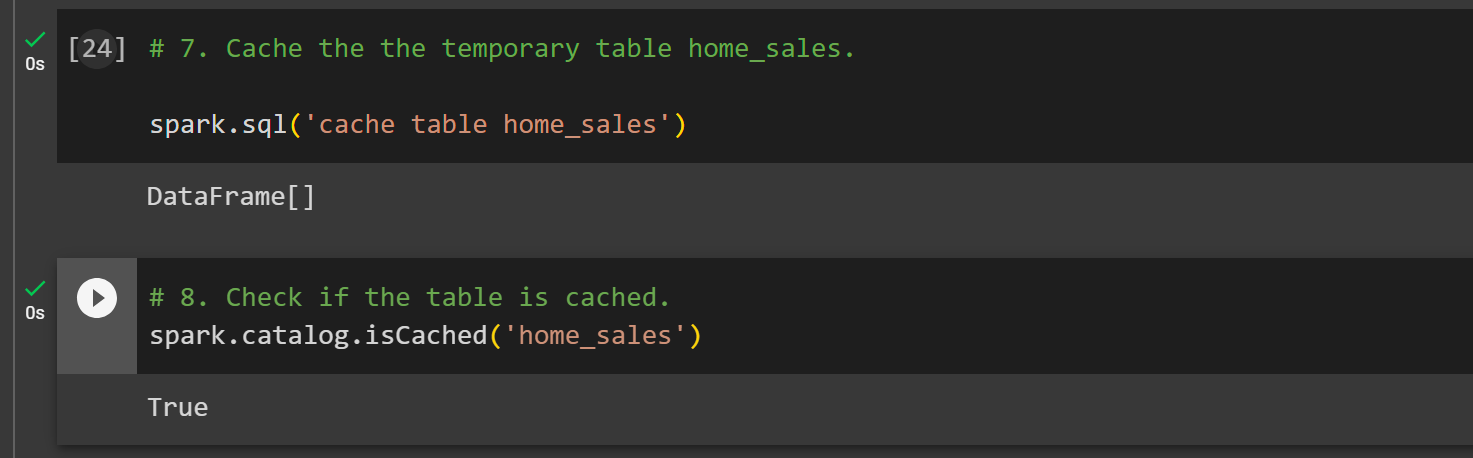
****

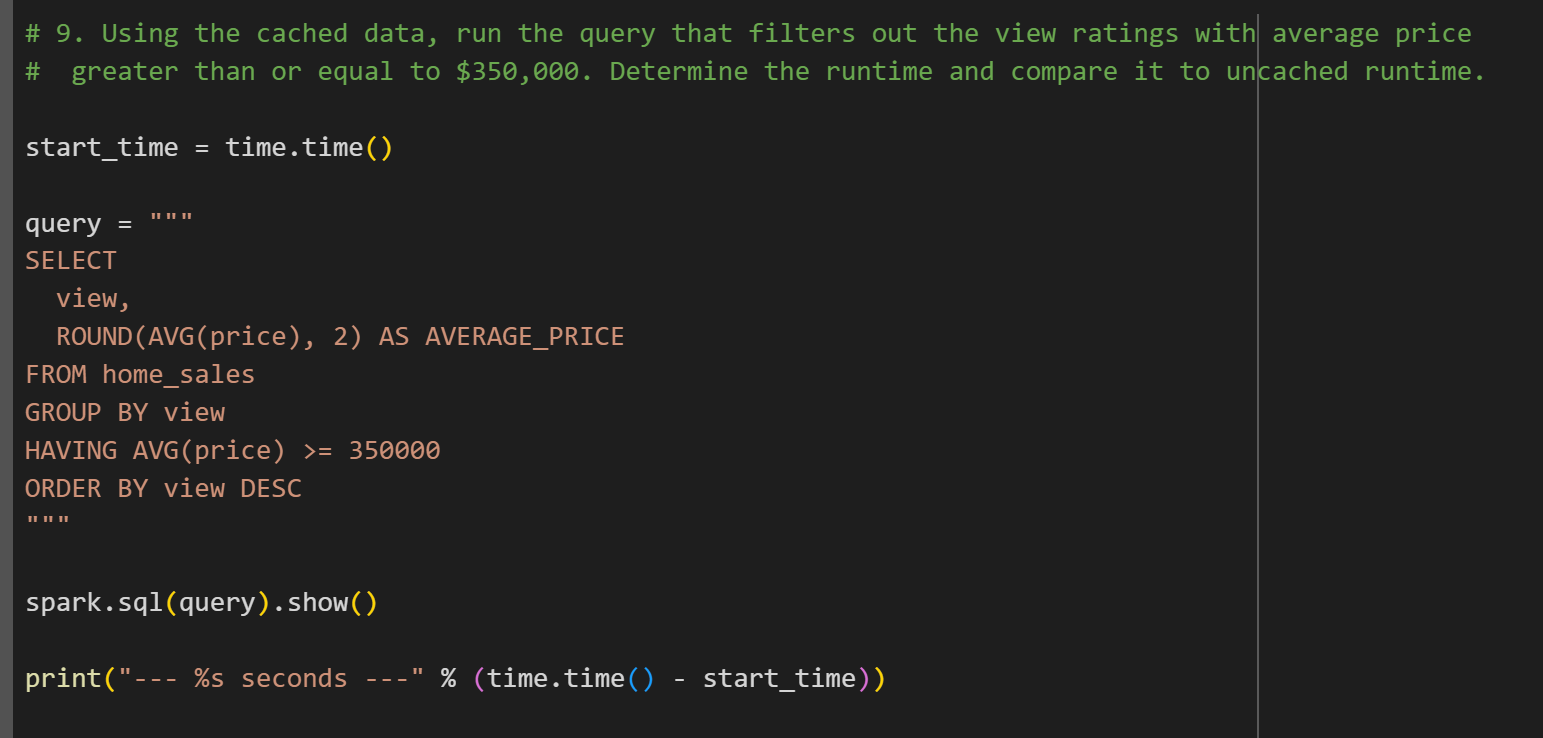
****

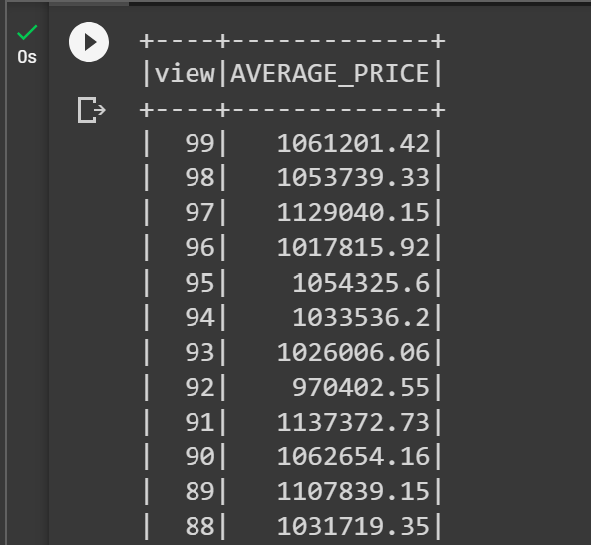
****

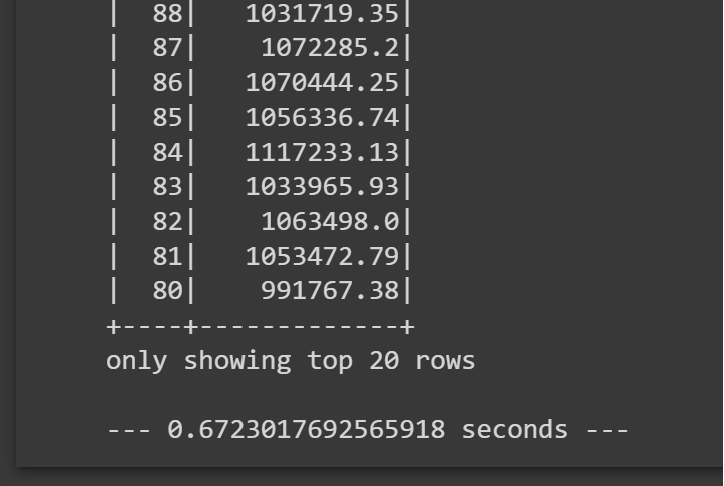
****

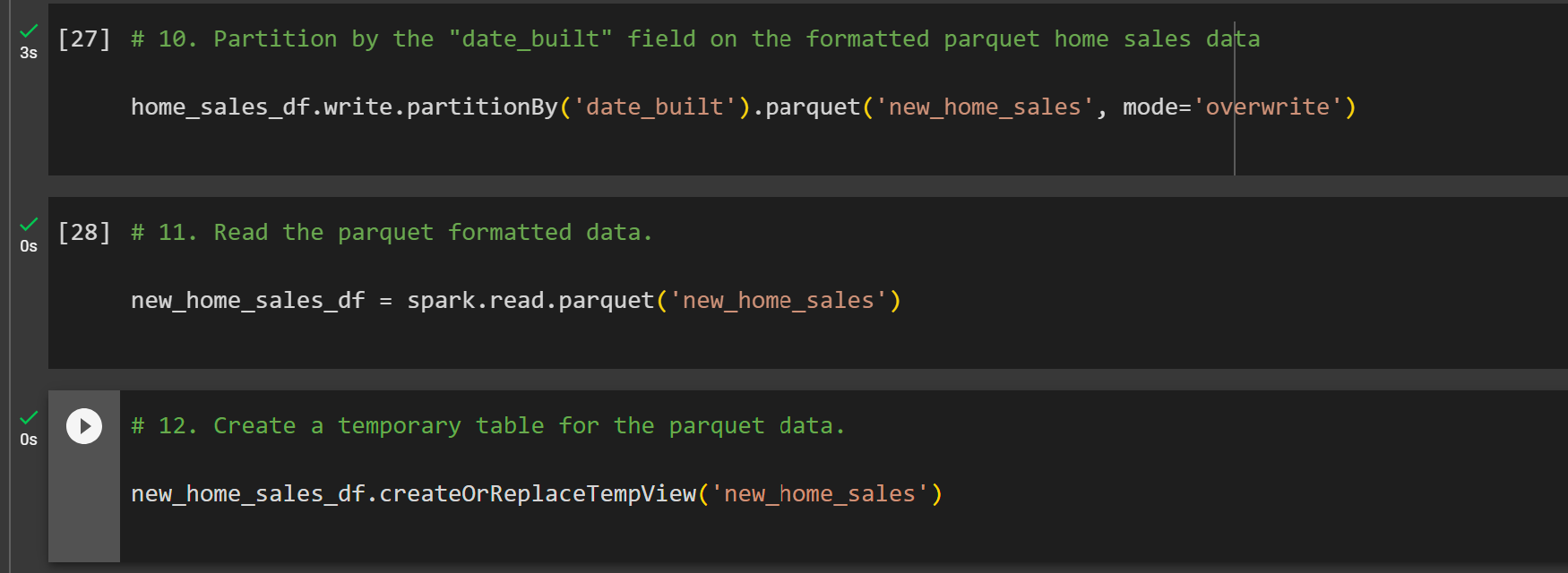
****

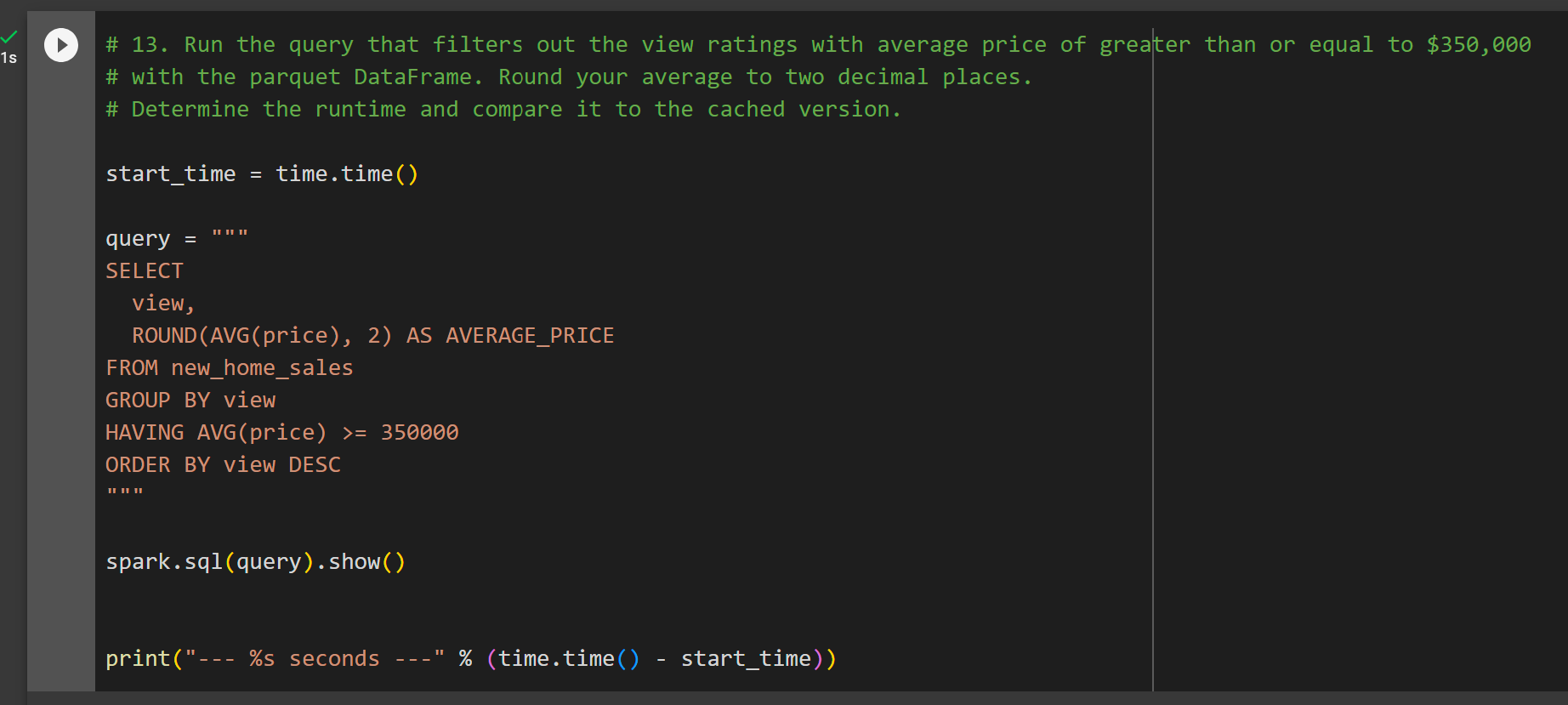
****

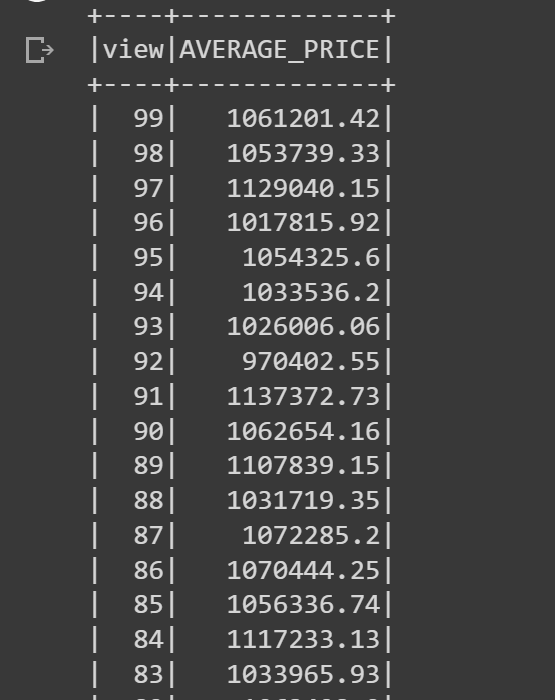
****

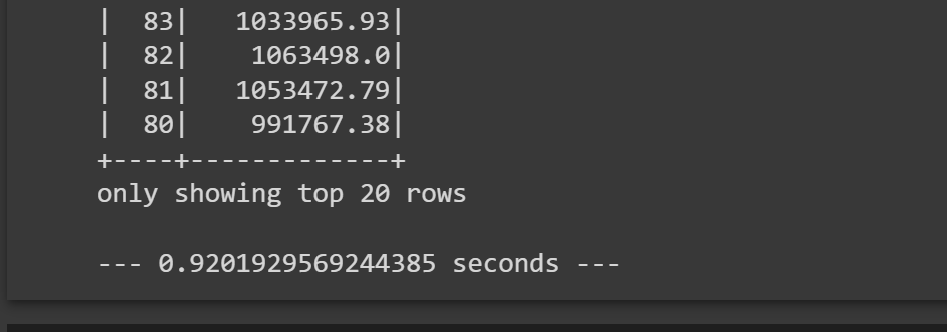
****

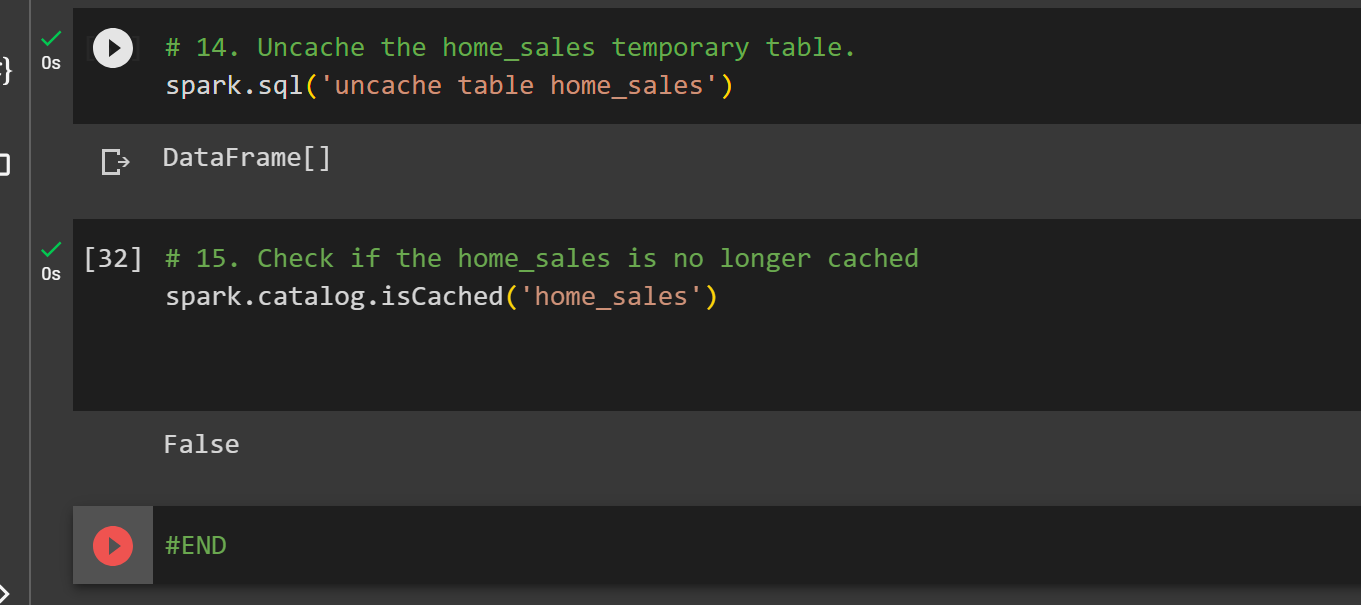
****

****

****

****

****

****

**#END**