Challenge 22

I used SparkSQL to determine critical metrics for home sales data. Then, I created a temporary table named home sales and answered the following questions:

1. What is the average price for a four-bedroom house sold for each year?

2. What is the average price of a home for each year the home was built, that has three bedrooms and three bathrooms?

3. What is the average price of a home for each year the home was built that has three bedrooms, three bathrooms, two floors, and is greater than or equal to 2,000 square feet?

4. What is the average price of a home per "view" rating having an average home price greater than or equal to $350,000? Determine the run time for this query.

The runtime for this query was 0.56 seconds.

Next, I cached the temporary home sales table and ran the sparkSQL query for question#4 again but this time against the cached data, the runtime was 0.69 seconds. I partitioned the data on the parquet home sales data by the ‘date\_built’ and created a temporary table for the parquet data. I ran the same sparkSQL query for question#4 to determine the runtime with the parquet data. The runtime decreased to 0.52 seconds this time, just a little faster because data was partitioned.

Next, I uncached the home sales temporary table and verified that the table had been uncached.

Partitioning data reduces the running time.