For joining Parquet File 1 (large dataset) and Parquet File 2 (smaller dataset), it's best to avoid explicit joins due to the significant size difference.

Sort Parquet File 1 on the *geographical_location_oid* column to enable efficient co-partitioning with the broadcasted table later.

Since Parquet File 2 is significantly smaller (10000 rows), Spark can efficiently broadcast it to all worker nodes, avoiding shuffling a large amount of data.

By sorting the larger table and broadcasting the smaller one, a join-like operation can be achieved without the overhead of a full shuffle join, minimizing data movement and improving performance.