**Detailed README / Documentation:**

**Data Preprocessing Using SQL for Hotel Bookings (Kaggle Dataset)**

The goal of this project is to replicate the Python preprocessing for hotel booking data using the PostgreSQL dialect. The PostgreSQL GUI used for this purpose is DBeaver. Additionally, all locally set Python scripts from previous projects, which supported earlier KPI dashboards, have been migrated to Google Colab to improve execution speed and simplify GitHub version control. The PostgreSQL preprocessing steps are carefully compared with the corresponding Python preprocessing scripts for each step.

As a next step, the already preprocessed data are loaded to Power BI.

**Quick Check on Data Integrity in Power BI**

* Number of Columns: Click Report view → Expand Transform Data → Choose Transform Data → The Power Query Editor opens → Look at the bottom of the screen as the number of columns is displayed
* Number of Rows: Table view → Look at the bottom of the screen; the number of rows is displayed.
* Data Types: Open Power Query Editor with the steps described above and check next to each column header. Ensure that both the order of columns and the data types match the corresponding cleaned PostgreSQL table. Note that Power BI may use simpler data types for quick metric calculations.
* Null Values: Open Power Query Editor → Choose the View menu → Check Column Quality → Valid must be 100% for each column.
* Duplicates: Remember the initial number of rows (in this case, 117,316) → Open Power Query Editor → Choose Home → Remove Rows → Remove Duplicates → Return to Power BI Desktop → Apply changes → Check the row number after duplicates removal (in this case, 86,640).

**Note:** The duplicates in this dataset were generated due to simplifications during preprocessing. For example, all high-cardinality columns were dropped, and further preprocessing made some rows less unique. Only the arrival\_date column remains as a high-cardinality column, so bookings on the same day may appear as duplicates in the final cleaned dataset, but all rows still represent valid, unique bookings.