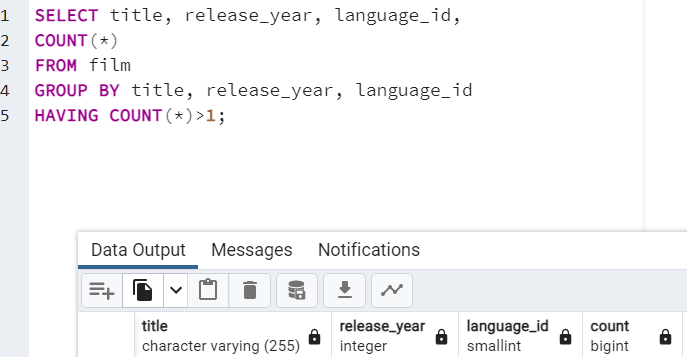
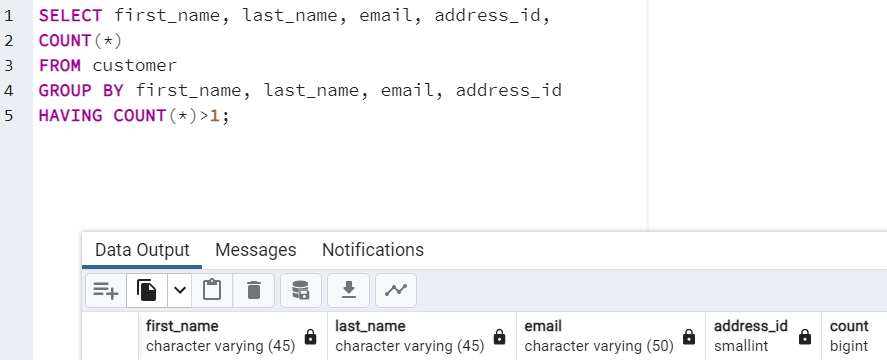
**3.6: Summarizing & Cleaning Data in SQL**

**Duplicates**

*Film table*



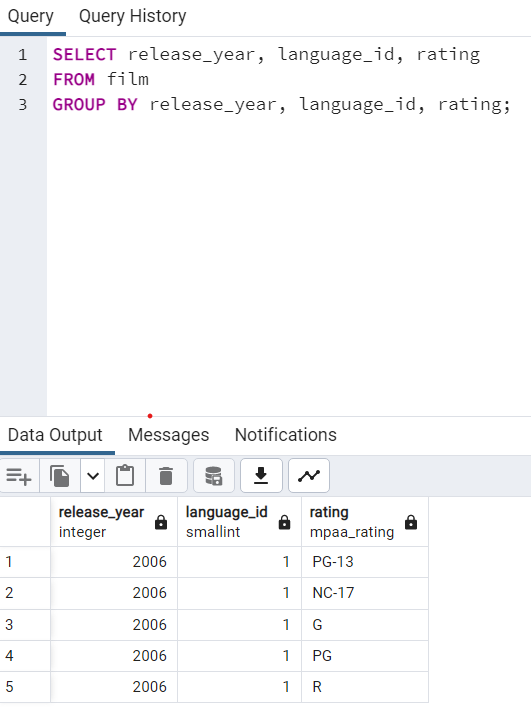
*Customer table*



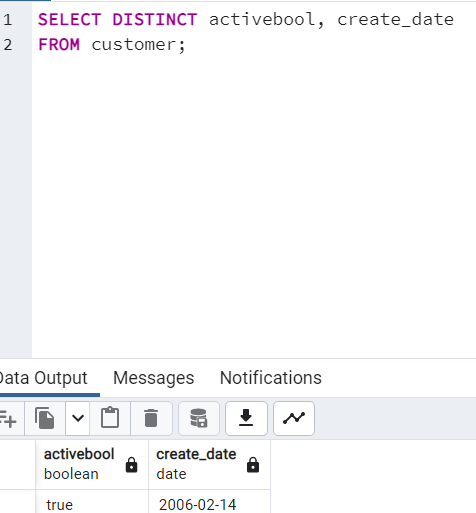
In none of the tables could be duplicates identified. However, in case that yes, I could create a VIEW with only unique records. Another option would be to remove duplicate data with DELETE. If I don’t have permission to do that, I`ll need to write a query that returns only unique records. There are 2 ways to do this – GROUP BY and DISTINCT.

**Non-Uniform Data**

*Film Table*

**

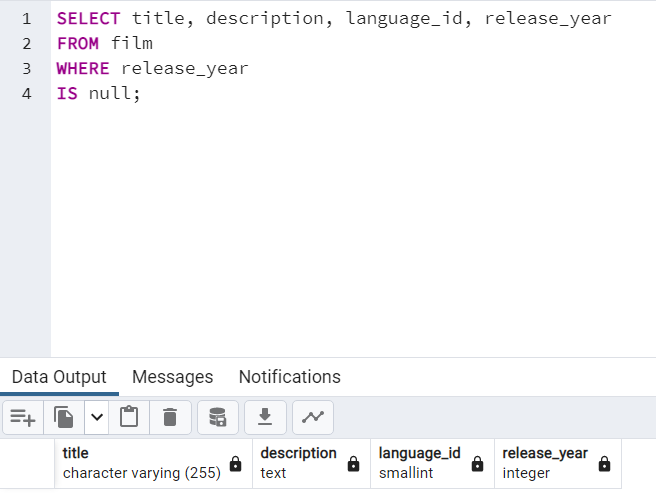
*Customer table*

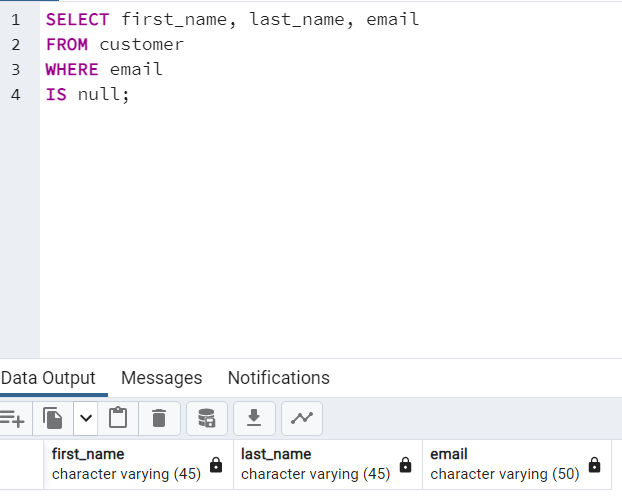
**

No non-uniform could be identified. In case that yes, the UPDATE can be used to make values consistent.

**Missing Values**

*Film Table*





No missing values could be identified. If a column contains a high number of missing values, the column should be excluded from the analyse by omitting it from the SELECT command. If a column contains a low amount of missing values, these missing values can be imputed with an estimate, such as column average.

**2. Descriptive Statistics**

*Film table*

****

*Customer table*

**

**3. Reflection**

SQL makes data profiling very simple and quick (once you get used to the Synthax and the logic). Especially when data set is large, SQL provides a very efficient way to work with a data set, especially when it’s about descriptive statistics. You can quickly identify the completenss and uniqueness of the dataset with DISTICNT and GROUP BY. However, since I am used to work with Excel already for several years, SQL is more complicated and slower for me. Especially the work with a smaller dataset would be faster with Excel. It’s literally like learning a new language – first you need to set a steady base (the Syntax) and then it’s practice and more practice to become fluent.