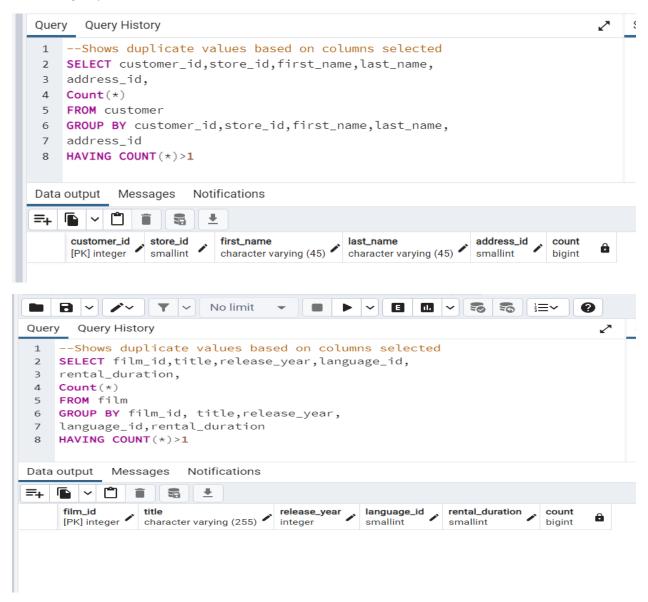
# ANSWERS 3.6 - Summarizing & Cleaning Data in SQL

# 1a. Finding Duplicate values



There are no duplicates, but to remove a duplicate, one would either create a view with unique records which is also known as a virtual table or the other is to delete the duplicates. But I will rather create a virtual table by selecting unique records and if I have to delete any duplicates, I will better use a DSP model.

### 1b. Finding non-uniform values



There are not any non-uniform values but to fix this if there are any, I will need to update any values that represent the real values in different formats to be in the exact format as the real value format e.g GEN, g, would have to be updated to be G.

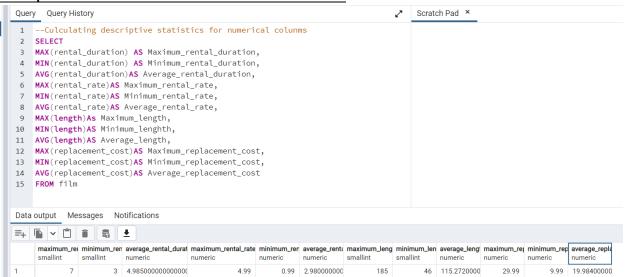
### 1c. Finding Missing Values





If there are a lot of missing values in a column, I would better ignore that column to minimize the impact on the analysis. If there are just a few missing values and I know what the values are I can fill it in or just leave it as it is.

2. <u>Descriptive statistics for numerical value in the film table</u>

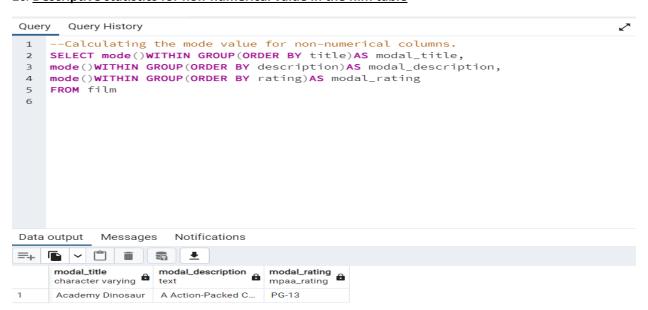


### 2b. Descriptive statistics for the numerical column in the customer table

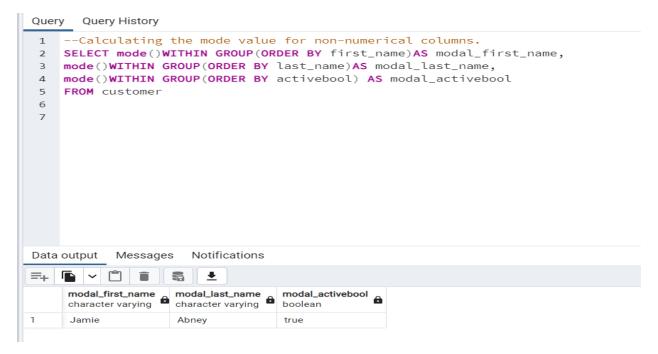
I didn't calculate the Average for these numerical values as we are with "id" so I sticked with Max and Minimum to know the highest and lowest value for the ids.



#### 2c. Descriptive statistics for non-numerical value in the film table



## 2d. <u>Descriptive statistics for non-numerical value in the customer table</u>



3. Using a pivot table for profiling in excel is quite easy when it comes to profiling big data, I think SQL is much easier and faster as much as one knows how to write the syntax and command properly. With Excel, you must go through various stages, like creating the table first, then following some procedures if you want to aggregate the data but with SQL one just needs to know the SQL syntax, write, and click execute and the answer is right there. Overall, using SQL is much easy and faster when it comes to large data and excel is also great when the data isn't large.