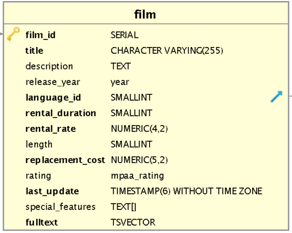
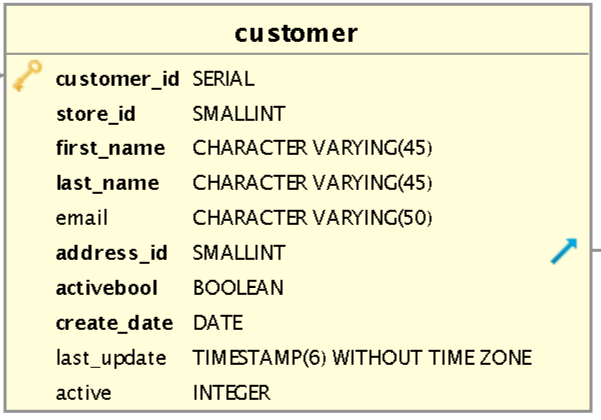
**ANSWERS 3.6**

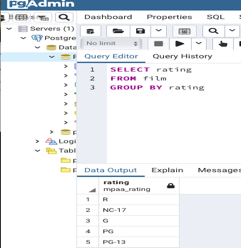
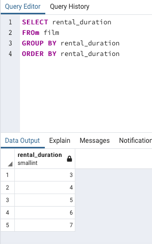
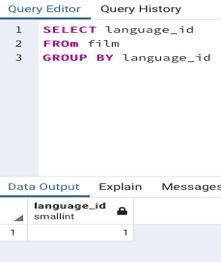
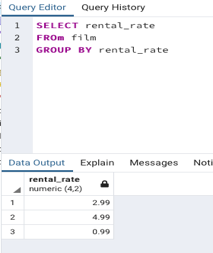
For this exercise I believe it will be useful to look at the ERD for both tables:

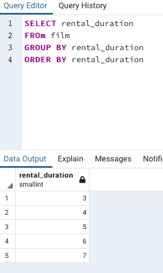
 

1. **Check for clean and dirty data.**

-Non-uniform data. According to the text *“…it's good practice to go through a few random values to check for inconsistencies*”.

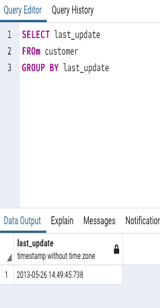
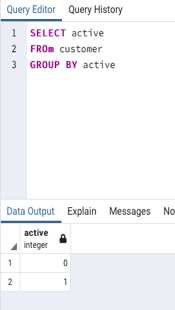
For the ***film*** table I have checked the following columns which contain different data types. In all cases data types are all consistent.



For the ***customer*** table I checked the following tables (with different data types) with the same results—no inconsistencies. For last\_update all records have the same date. For create\_date all records have the same date.

For active all records are either 0 or 1. Finally, for store\_id all records are either 1 or 2.

Should have been any non-uniform data in any columns from those two tables I would have fixed them by updating values to be consistent. The command would be:

UPDATE table\_name

SET column1 = ‘X’

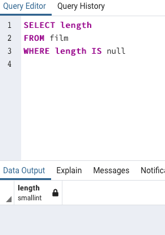
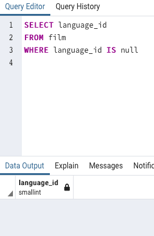
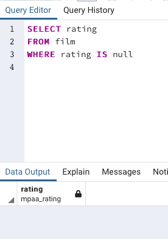
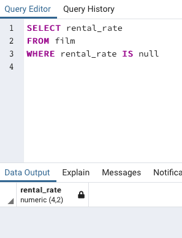
WHERE column1 IN (‘x’, ‘xi’, ‘xii’…)

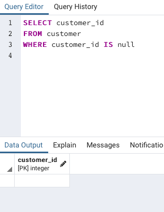
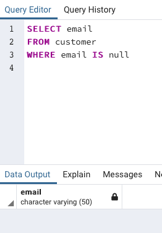
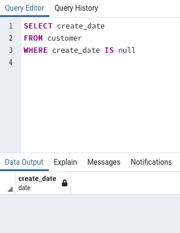
-Duplicated data

I have not found any duplicate values in neither of both tables. Should have any duplicate values been found I would have created a virtual table, known as “view,” with only unique records since other analysts might need those records in the future.

-Missing values

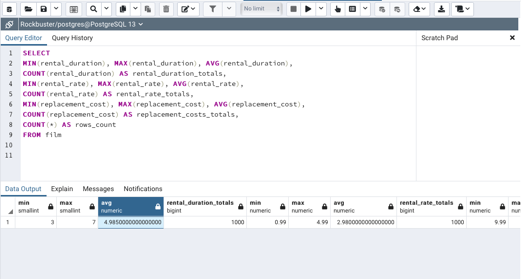
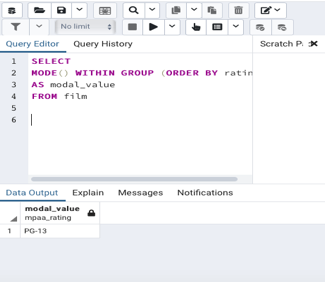
   

For **film** and **customer** tables I run few queries looking for *null* on few columns returning no records which means there are not missing values.

In this case I would choose to ignore those columns with missing values if the percentage of them was very high. Otherwise, I would impute the missing values by using statistical methods.

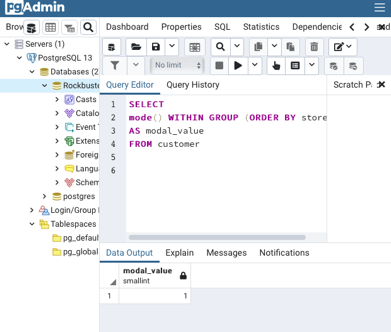
1. **Summarize your data.**

For the film table:

**** 

For the customer table there are not numerical columns to calculate descriptive statistics from. Even though *customer\_id* or date might have numbers on them it wouldn’t make sense to calculate max, min or avg from those columns. Therefore, I will only calculate the mode.

The only column where values appear more than once is store\_id so I calculate the mode for that.



1. **Reflect on your work.**

**Back in Achievement 1 you learned about data profiling in Excel. Based on your previous experience, which tool (Excel or SQL) do you think is more effective for data profiling, and why.**

Where it comes to data profiling, I would say that SQL is a more effective wat to profile data due to the fact that it can be done much faster than in EXCEL. When trying to find patterns with SQL it is a matter of writing a query while in EXCEL the process has more steps.