TASK 3.6 Jada Myrie

1. **Check for and clean dirty data:** Find out if the film table and the customer table contain any dirty data, specifically non-uniform or duplicate data, or missing values. Create a new "Answers 3.6" document and copy-paste your queries into it. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

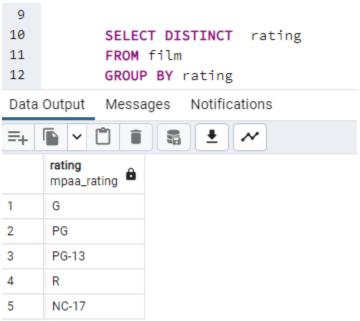
```
1
   SELECT customer_id,
2
           first_name,
3
           last_name,
4
           count(*)
5
   FROM customer
6
   GROUP BY customer_id,
7
             first_name,
8
             last_name
9
   HAVING COUNT(*) > 1; -- no results set means that we have no duplicates
10
```

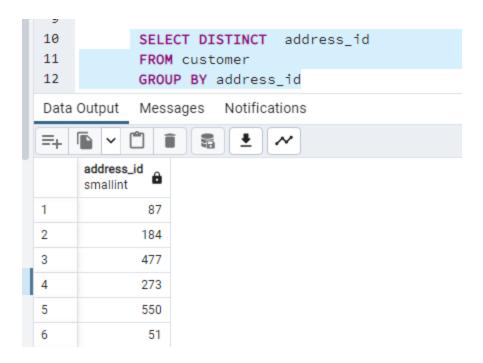
```
SELECT film_id,
 1
 2
           title,
 3
           length,
 4
           count(*)
 5
   FROM film
 6
   GROUP BY film_id,
 7
              title,
 8
             length
 9
   HAVING COUNT(*) > 1; -- no result set means that we have no duplicates
10
```

I would clean the data by removing any duplicate data using a duplicate query.

Also, if data is ununiform I would use a Group by or distinct query to determine the data I want to format all other data to match. Then after seeing the format/language used I would UPDATE the others.

For example



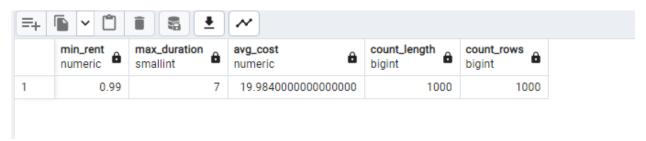


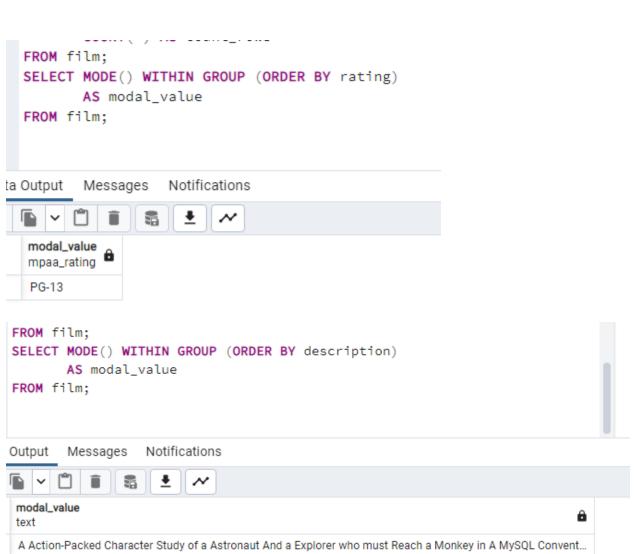
Missing data can be imputed using an aggregate statement in a query. However, it is best to do this if there is only a small amount of data missing. I can also use an ignore statement if large amounts of data are missing.

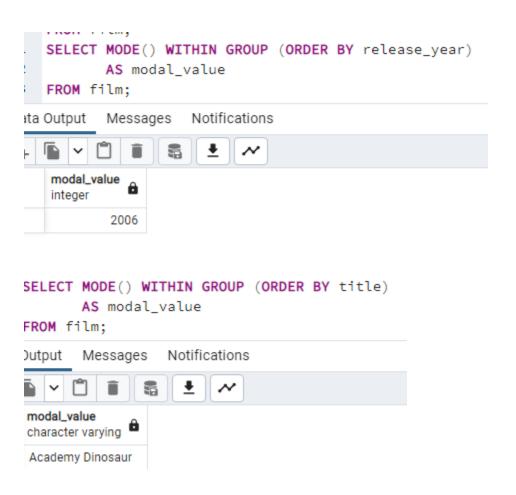
2.Summarize your data: Use SQL to calculate descriptive statistics for both the film table and the customer table. For numerical columns, this means finding the minimum, maximum, and average values. For non-numerical columns, calculate the mode value. Copy-paste your SQL queries and their outputs into your answers document.

Film Table Attempt

```
SELECT MIN(rental_rate) AS min_rent,
MAX(rental_duration) AS max_duration,
AVG(replacement_cost) AS avg_cost,
COUNT(length) AS count_length,
COUNT(*) AS count_rows
FROM film;
```

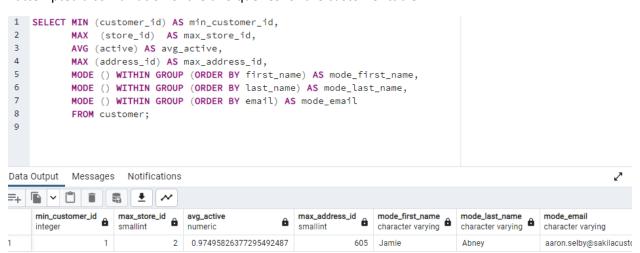






Customer Table Attempt

I attempted a combination of the two queries for the customer table.



4. **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? Consider their respective functions, ease of use, and speed. Write a short paragraph in the running document that you have started.

For data profiling I believe Excel is easier because you can navigate the table and immediately notice the difference in data, for example if PG-13 is listed as PG-thirteen for some cells. This is easy to pick up with the naked eye. Then you could use SQL to write queries to then update and or delete certain data. I believe SQL handles data profiling as well as making cleaning easier. In excel when you write functions you mainly cater to dealing with one column at a time besides when an IF function is used. In SQL you can run multiple functions at once on an entire table.