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)

Film Table (Identifying Duplicates)

```
SELECT title, description, release_year, language_id, rental_duration, rental_rate, length, replacement_cost, rating, last_update, special_features, COUNT (*)
FROM film
GROUP BY film_id, title, description, release_year, language_id, rental_duration, rental_rate, length, replacement_cost, rating, last_update, special_features
HAVING COUNT (*) >1
```



Customer Table (Identifying Duplicates)

SELECT store_id, first_name, last_name, email, address_id, activebool, active, COUNT(*)
FROM customer
GROUP BY customer_id, store_id, first_name, last_name, email, address_id, activebool, active HAVING COUNT (*) >1

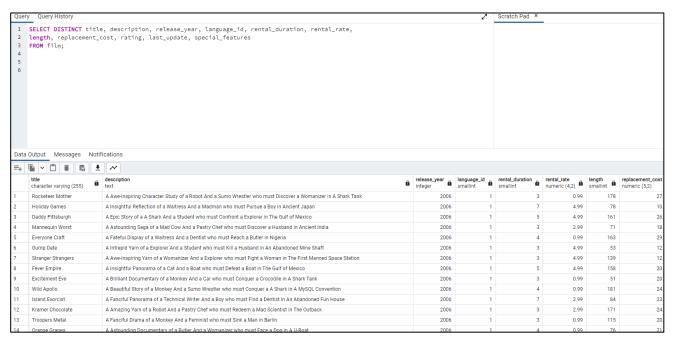
```
Query Query History
    SELECT store_id, first_name, last_name, email, address_id, activebool, active,
 1
    COUNT(*)
 3
    FROM customer
    GROUP BY customer_id, store_id, first_name, last_name, email, address_id, activebool, active
 4
    HAVING COUNT (*) >1
 5
 6
                        Notifications
Data Output
            Messages
=+
                                                                             address_id
     store_id
                first_name
                                     last_name
                                                         email
                                                         character varying (50)
                character varying (45)
                                    character varying (45)
     smallint
                                                                              smallint
                                                                                          boolean
                                                                                                      integer
                                                                                                                 bigint
```

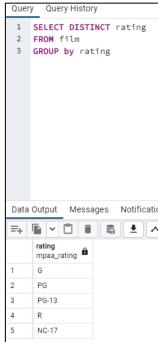
The film table and customer table thankfully do not have any duplicates, but if there were any, I would first identify and confirm it is indeed a duplicate (or dirty in general), then utilize the ALTER function in SQL to adjust it to conform with the existing items, or I could DELETE the whole record altogether – this is assuming I have editing permissions as an analyst. If not, I would flag the duplicate / dirty data for IT or the data engineers to fix for me. In the meantime, I could use the VIEW feature to adjust what I need to adjust to perform my analysis.

Film Table (Non-Uniform Data)

SELECT DISTINCT title, description, release_year, language_id, rental_duration, rental_rate, length, replacement_cost, rating, last_update, special_features
FROM film;

SELECT DISTINT rating FROM film GROUP BY rating

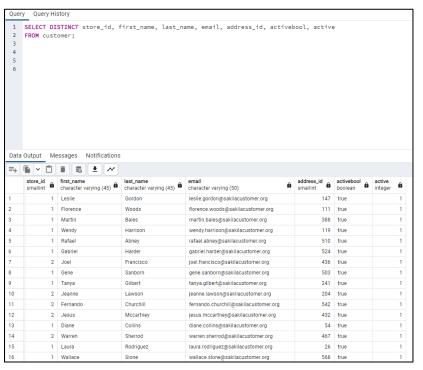


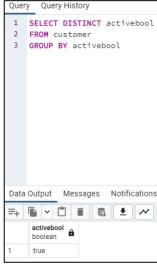


Customer Table (Non-Uniform Data)

SELECT DISTINCT store_id, first_name, last_name, email, address_id, activebool, active FROM customer;

SELECT DISTINCT activebool FROM customer GROUP BY activebool





Using SELECT DISTINT instead of SELECT would provide us with all unique data points, so we can quickly identify any records that aren't uniformed with the others. I first ran it for all of the columns in the respective tables, then picked a column I thought may have non-uniformed records to test them out. Again, thankfully there aren't any non-uniformed records, but if there were, once confirmed, I can use the ALTER or UPDATE function to fix them assuming I have editing permissions. If not, I would flag them and report them to IT and/or the data engineer(s).

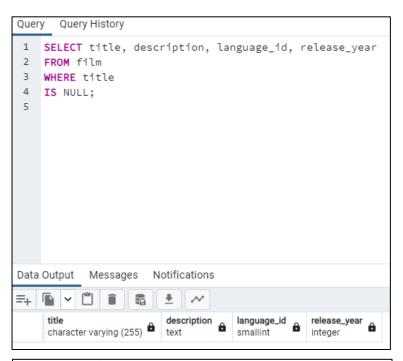
Film Table (Missing Values)

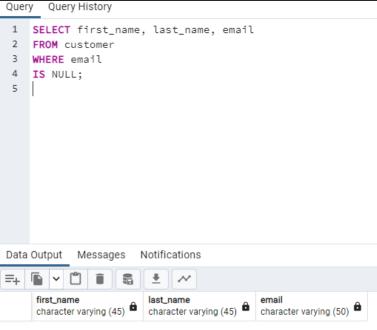
SELECT title, description, language_id, release_year FROM film WHERE title IS NULL;

Customer Table (Missing Values)

SELECT first_name, last_name, email FROM customer WHERE email IS NULL;

I ran this portion multiple times changing the portion after WHERE to different columns to see if any showed up. Thankfully, like before, there aren't any missing values. If there were, it'd be harder to fix than other types of dirty data since we can't be completely sure what was supposed to go in there. In this very particular example, we may be able to deduce someone's missing first or last name from their email if they happen to use their first and last name in their email address; however, not every column is going to be as easy to deduce from. As analysts, we could remove the records with missing data from our analysis or count them as "N/A" and follow-up with the respective teams to try to get the missing data filled in.

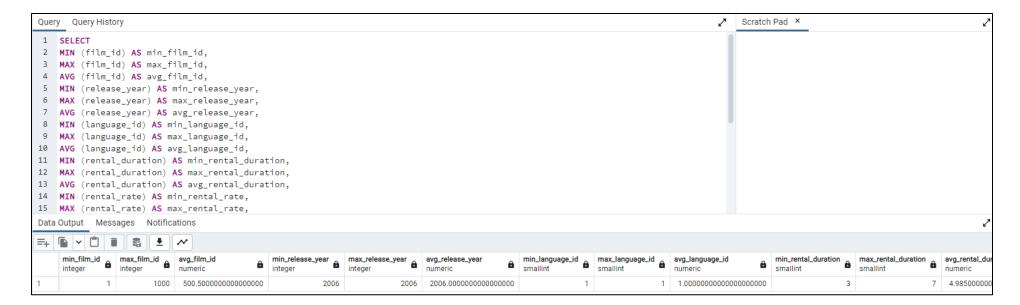




2. Summarize your data: Use SQL to calculate descriptive statistics for both the film table and the customer table. For numerical columns, this means finding the minimum, maximum, and average values. For non-numerical columns, calculate the mode value. Copy-paste your SQL queries and their outputs into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

Film Table

```
SELECT
MIN (film id) AS min film id,
MAX (film id) AS max film id,
AVG (film id) AS avg film id,
MIN (release year) AS min release year,
MAX (release year) AS max release year,
AVG (release year) AS avg release year,
MIN (language id) AS min language id,
MAX (language_id) AS max_language_id,
AVG (language id) AS avg language id,
MIN (rental duration) AS min rental duration,
MAX (rental duration) AS max rental duration,
AVG (rental_duration) AS avg_rental_duration,
MIN (rental rate) AS min rental rate,
MAX (rental_rate) AS max_rental_rate,
AVG (rental rate) AS avg rental rate,
MIN (length) AS min length,
MAX (length) AS max length,
AVG (length) AS avg length,
MIN (replacement cost) AS min replacement cost,
MAX (replacement cost) AS max replacement cost,
AVG (replacement cost) AS avg replacement cost,
MODE () WITHIN GROUP (ORDER BY title) AS mode title,
MODE () WITHIN GROUP (ORDER BY description) AS mode description,
MODE () WITHIN GROUP (ORDER BY rating) AS mode rating,
MODE () WITHIN GROUP (ORDER BY special features) AS mode special features,
MODE () WITHIN GROUP (ORDER BY fulltext) AS mode fulltext
FROM film:
```



Customer Table

SELECT

MIN (customer_id) AS min_customer_id,

MAX (customer_id) AS max_customer_id,

AVG (customer_id) AS avg_customer_id,

MIN (store_id) AS min_store_id,

MAX (store_id) AS max_store_id,

AVG (store_id) AS avg_store_id,

MIN (address_id) AS min_address_id,

MAX (address_id) AS max_address_id,

AVG (address_id) AS avg_address_id,

MIN (active) AS min_active,

MAX (active) AS max_active,

AVG (active) AS avg_active,

MODE () WITHIN GROUP (ORDER BY first_name) AS mode_first_name,

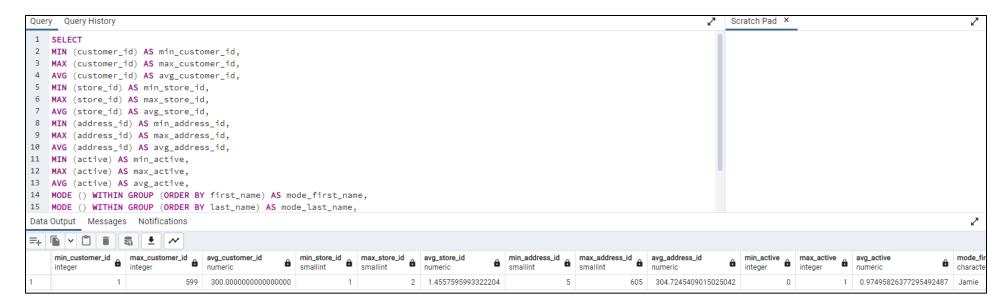
MODE () WITHIN GROUP (ORDER BY last_name) AS mode_last_name,

MODE () WITHIN GROUP (ORDER BY email) AS mode_email,

MODE () WITHIN GROUP (ORDER BY active) AS mode_active,

MODE () WITHIN GROUP (ORDER BY activebool) AS mode_activebool

FROM customer;



3. 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.

As I'm still more comfortable using Excel, it was easier for me to perform the data profiling using it versus SQL. It took me a bit to manually type out the SQL to accomplish what I needed, but once I had it laid out, I know I have it ready for the next time. Unless I'm using Excel VBA, SQL would be the quicker option for repetition versus manually doing it over and over in Excel. Once the SQL is written out, pressing the play/run button would be the quicker option versus writing out the respective formulas in Excel and flash-filling the rest of the column. Currently, I'm still on the fence between the two options. I can see that SQL would be the quicker option, especially with bigger datasets with a lot more columns, but as I'm still learning it, it's taking me much longer to type everything out, have it in the correct order, and becoming fluent in it as I am in Excel.