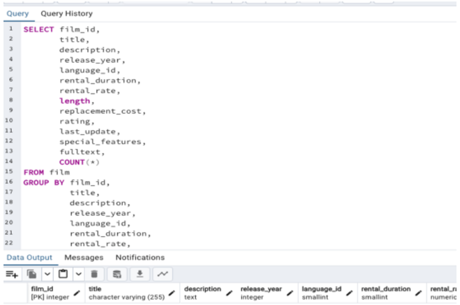
Katie Goyal

Data Analytics Immersion 3.6

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).

Duplicate data:

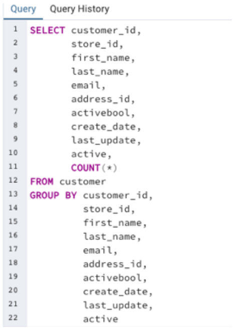
Film Table



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

film_id [PK] integer	title character varying (255)	description text	release_year integer	language_id smallint	rental_duration smallint	rental_r numeric

Customer Table

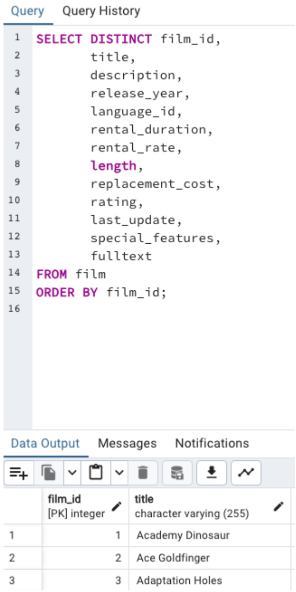


```
SELECT customer_id,
       store_id,
       first_name,
       last_name,
       email,
       address_id,
       activebool,
       create_date,
       last_update,
       active,
       COUNT(*)
FROM customer
GROUP BY customer_id,
         store_id,
         first_name,
         last_name,
         email,
         address_id,
         activebool,
         create_date,
         last_update,
         active
```

- ***There were no duplicates found when running the query. If duplicates where found one would just need to run a query to delete the duplicates***

Non_uniform:

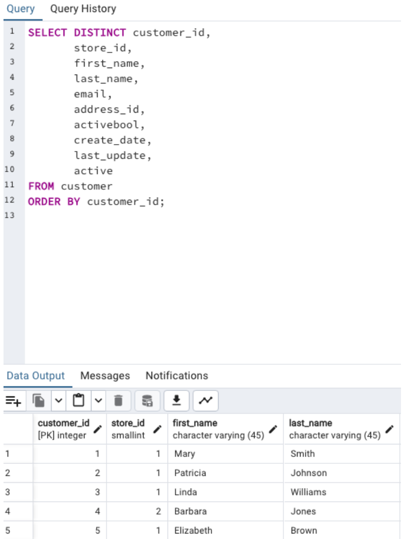
Film Table



```sql
SELECT DISTINCT film_id,
       title,
       description,
       release_year,
       language_id,
       rental_duration,
       rental_rate,
       length,
       replacement_cost,
       rating,
       last_update,
       special_features,
       fulltext
FROM film
ORDER BY film_id;
```

	film_id [PK] integer	title character varying (255)
1	1	Academy Dinosaur
2	2	Ace Goldfinger
3	3	Adaptation Holes

Customer Table



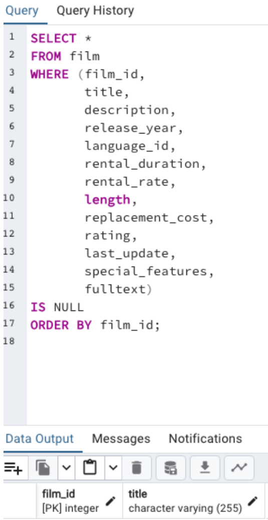
```sql
SELECT DISTINCT customer_id,
       store_id,
       first_name,
       last_name,
       email,
       address_id,
       activebool,
       create_date,
       last_update,
       active
FROM customer
ORDER BY customer_id;
```

	customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)
1	1	1	Mary	Smith
2	2	1	Patricia	Johnson
3	3	1	Linda	Williams
4	4	2	Barbara	Jones
5	5	1	Elizabeth	Brown

- ***There were no non_uniform data in the tables. If there were to be non_uniform data one would the UPDATE Statement instead of select to filter and uniform the data.***

Missing:

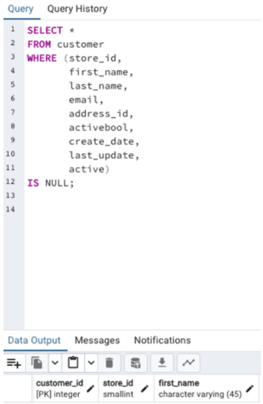
Film Table



```
SELECT *
FROM film
WHERE (film_id,
       title,
       description,
       release_year,
       language_id,
       rental_duration,
       rental_rate,
       length,
       replacement_cost,
       rating,
       last_update,
       special_features,
       fulltext)
IS NULL
ORDER BY film_id;
```

film_id [PK] integer	title character varying (255)

Customer Table



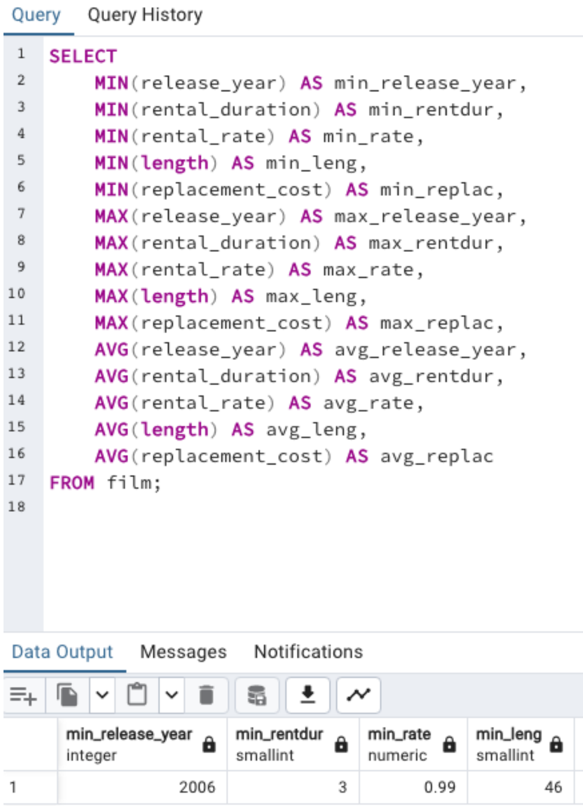
```
SELECT *
FROM customer
WHERE (store_id,
       first_name,
       last_name,
       email,
       address_id,
       activebool,
       create_date,
       last_update,
       active)
IS NULL;
```

customer_id [PK] integer	store_id smallint	first_name character varying (45)

- ***There was no missing data found in the tables above. If there was missing data I would filter clean and update the data to fill in the missing values.***

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.

Numerical

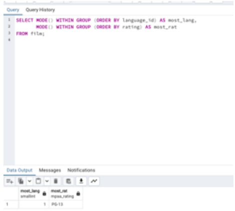


```
SELECT
    MIN(release_year) AS min_release_year,
    MIN(rental_duration) AS min_rentdur,
    MIN(rental_rate) AS min_rate,
    MIN(length) AS min_leng,
    MIN(replacement_cost) AS min_replac,
    MAX(release_year) AS max_release_year,
    MAX(rental_duration) AS max_rentdur,
    MAX(rental_rate) AS max_rate,
    MAX(length) AS max_leng,
    MAX(replacement_cost) AS max_replac,
    AVG(release_year) AS avg_release_year,
    AVG(rental_duration) AS avg_rentdur,
    AVG(rental_rate) AS avg_rate,
    AVG(length) AS avg_leng,
    AVG(replacement_cost) AS avg_replac
FROM film;
```

	min_release_year integer	min_rentdur smallint	min_rate numeric	min_leng smallint
1	2006	3	0.99	46

- ***There is no numerical data that would help in understanding this data. Knowing the number of customers who rented would have been more beneficial***

Nonnumerical



```
SELECT MODE() WITHIN GROUP (ORDER BY language_id) AS most_lang,
       MODE() WITHIN GROUP (ORDER BY rating) AS most_rat
FROM film;
```

	most_lang smallint	most_rat mpaa_rating
1	1	PG-13

- ***When looking at this data we are given a numerical number for most languages rented and not a language which does not help us in***

understanding what was the most rented movie by language. We can understand that the most rented movies are PG 13.

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.

    - ***When working with data in SQL versus Excel we can control, manage, edit, and manipulate the data much faster and easier than if we were to do it in Excel. The amount of time it takes to write and run a query I would have only finished one column in Excel.***