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).

```sql
SELECT customer_id,
       first_name,
       last_name,
       count(*)
FROM customer
GROUP BY customer_id,
         first_name,
         last_name
HAVING COUNT(*) > 1; -- no results set means that we have no duplicates

```

```sql
SELECT film_id,
       title,
       length,
       count(*)
FROM film
GROUP BY film_id,
          title,
         length
HAVING COUNT(*) > 1; -- no result set means that we have no duplicates

```

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

```
SELECT DISTINCT  rating
FROM film
GROUP BY rating
```

Data Output Messages Notifications

	rating mpaa_rating
1	G
2	PG
3	PG-13
4	R
5	NC-17

```sql
SELECT DISTINCT  address_id
FROM customer
GROUP BY address_id
```

Data Output Messages Notifications

	address_id smallint
1	87
2	184
3	477
4	273
5	550
6	51

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

```sql
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;
```

	min_rent numeric	max_duration smallint	avg_cost numeric	count_length bigint	count_rows bigint
1	0.99	7	19.9840000000000000	1000	1000

```
FROM film;
SELECT MODE() WITHIN GROUP (ORDER BY rating)
       AS modal_value
FROM film;
```

ta Output Messages Notifications

modal_value mpaa_rating
PG-13

```
FROM film;
SELECT MODE() WITHIN GROUP (ORDER BY description)
       AS modal_value
FROM film;
```

Output Messages Notifications

modal_value text
A Action-Packed Character Study of a Astronaut And a Explorer who must Reach a Monkey in A MySQL Convent...

```
SELECT MODE() WITHIN GROUP (ORDER BY release_year)
       AS modal_value
FROM film;
```

ata Output Messages Notifications

modal_value integer
2006

```
SELECT MODE() WITHIN GROUP (ORDER BY title)
       AS modal_value
FROM film;
```

Output Messages Notifications

modal_value character varying
Academy Dinosaur

Customer Table Attempt

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

```
SELECT MIN (customer_id) AS min_customer_id,
       MAX  (store_id)  AS max_store_id,
       AVG (active) AS avg_active,
       MAX (address_id) AS max_address_id,
       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
       FROM customer;
```

Data Output Messages Notifications

	min_customer_id integer	max_store_id smallint	avg_active numeric	max_address_id smallint	mode_first_name character varying	mode_last_name character varying	mode_email character varying
1	1	2	0.97495826377295492487	605	Jamie	Abney	aaron.selby@sakilacust

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.