Task 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

Film

Query | Query History

```
1  SELECT title,
2         release_year,
3         language_id,
4         rental_duration,
5         COUNT(*)
6  FROM film
7  GROUP BY title,
8           release_year,
9           language_id,
10          rental_duration
11 HAVING COUNT(*) >1;
```

Customer

Query | Query History

```
1  SELECT first_name,
2         last_name,
3         email,
4         activebool,
5         COUNT(*)
6  FROM customer
7  GROUP BY first_name,
8           last_name,
9           email,
10          activebool
11 HAVING COUNT(*) >1;
```

If we had duplicated values in these tables, I could use one of 2 techniques the first one is creating a view to analyze just the unique values, or I would delete the duplicated values, but this is usually not the roll of a junior data analyst.

Non-Uniform

Film

Query | Query History

```
1  SELECT rating
2  FROM film
3  GROUP BY rating
```

Customer

Query | Query History

```
1  SELECT active
2  FROM customer
3  GROUP BY active
```

We can use this query to keep checking the columns in the table to figure out if there is any non-uniform data in these tables and to solve this, we can use the update function to change the entries that have different spelling or data type

Missing values

To know if we have missing values, we can get a visualization of the whole table and do a quick check to see if there is any null or empty entry, which can be solved by setting a default value when there is an empty entry or if the missing value are a considerable amount, a good course of action could be to not use that column for the analysis

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

Query | Query History

```sql
SELECT MIN(rental_rate) AS min_rent,
       MAX(rental_rate) AS max_rent,
       AVG(rental_rate) AS avg_rent,
       MIN(release_year) AS min_release_year,
       MAX(release_year) AS max_release_year,
       AVG(release_year) AS avg_release_year,
       MIN(rental_duration) AS min_rental_duration,
       MAX(rental_duration) AS max_rental_duration,
       AVG(rental_duration) AS avg_rental_duration,
       mode() WITHIN GROUP (ORDER BY special_features)
       AS special_features_mode
FROM film;
```

Customer

Query | Query History

```sql
SELECT MIN(create_date) AS min_createD,
       MAX(create_date) AS max_createD,
       MIN(last_update) AS min_update,
       MAX(last_update) AS max_update,
       mode() WITHIN GROUP (ORDER BY active)
       AS active
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.

I think SQL is a more specialized program than Excel which makes it more fitted to do profiling when we have more specific requests about the data and give us the chance to work with just the necessary amount of data from the whole data set which in some cases could be really wide and almost impossible to profile, this making it easier to work with.