3.6 Answers

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.

**Film Table:**

SELECT title,

release\_year,

language\_id,

rental\_duration,

COUNT(\*)

FROM film

GROUP BY title,

release\_year,

language\_id,

rental\_duration

HAVING COUNT(\*) >1;

There was no duplicates found. If there had been duplicates, I could deal with them by creating a new view with only unique values in it. The other option is to delete the duplicates, however typically it is better to not risk that and create a new view instead.

**Customer Table:**

SELECT customer\_id,

store\_id,

first\_name,

last\_name,

email,

address\_id,

COUNT (\*)

FROM customer

GROUP BY customer\_id,

store\_id,

first\_name,

last\_name,

email,

address\_id

HAVING COUNT(\*) >1;

There was no duplicates found. If there had been duplicates, I could deal with them the same way as above.

2. **Summarize your data:** Use SQL to calculate descriptive stats 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.

Summary for numeric columns in film table :

Graphical user interface, text

Description automatically generated

Summary for non-numeric columns in film table:

Graphical user interface, text, application, email

Description automatically generated

Summary of numeric columns in customer table:

Graphical user interface, text, application

Description automatically generated

Summary of non-numeric columns in customer table:

Graphical user interface, text

Description automatically generated

3. In my opinion SQL is a lot better than excel with large data sets. It is easier and quicker to gather the information needed and things feel more accessible and organized. Things like aliasing are much quicker with far less steps. Overall, SQL is essentially better with large data sets overall. Excel works well for smaller data sets.