**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 Data for Film table

Graphical user interface, text, application, email

Description automatically generated

Duplicate Data for Customer Table

Graphical user interface, text, application, email

Description automatically generated

Non-uniform data for film table

Graphical user interface, text, application, Word

Description automatically generated

The rating section has too many of the same variables but in different format. For example, “PG” and “PG-13” are the same rating but in different format. To fix fix this you would first find out which of the two variables listed above have the most records in the data set. Then you would use the SQL script:

UPDATE film

SET rating = “PG-13”

WHERE rating IN (“G”)

**Missing values from film table**

Missing values are often null, empty, or replaced with a dummy or default value instead. To fix this first you would, find the average of the column with the missing value and use it to fill in the missing records. If there is over 90% of the data missing, keeping that data out of the analysis would be wise. To replace the data above follow the command below. A lot of missing data can be fixed with logic.

UPDATE tablename

SET = AVG (col1)

WHERE col1 IS NULL

1. **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.

**Summary for the film table**

**Graphical user interface, text, application

Description automatically generated**

**Summary for the film table**

Graphical user interface, text, application, email

Description automatically generated

**Summary for the customer table**

Graphical user interface, text, application

Description automatically generated

1. **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.

Data profiles help analysts organize some of the results and insights. These profiles also help better spot data quality and data integrity. After using SQL in theses last 6 tasks, I feel using excel would be easier to profile data. Building tables and entering in data is a lot easier and simpler in excel. In contrary, if the data being used is more that an excel spreadsheet can handle then using SQL to profile data would be a much better choice.