**Gwyn Reniers**

**Exercise 3.6**

**LINK TO EXCEL FILE:**

<https://docs.google.com/spreadsheets/d/12fBfiEVE-RmJaX701-ps21pvnvzNvXa6/edit?usp=sharing&ouid=103870910939658411214&rtpof=true&sd=true>

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).
2. SELECT title,
3. release\_year,
4. language\_id,
5. rental\_duration,
6. COUNT(\*)
7. FROM film
8. GROUP BY title,
9. release\_year,
10. language\_id,
11. rental\_duration
12. HAVING COUNT(\*) >1; --no result set means we have no duplicates

*Graphical user interface, application

Description automatically generated*

*There are no duplicate records appearing in the film table. If there were, I could run a view query (not delete) and notate that there were records removed from viewing due to duplication. If there were missing data for release year, a google search would allow me to fill in the information accurately, same for language. If there were missing rental duration information, I could use values of a movie with the same length to fill in the missing information, non- uniform values could be changed to uniform using the query: UPDATE/SET/WHERE/IN filling in the information I needed to change.*

SELECT customer\_id,

first\_name,

last\_name,

email,

address\_id,

COUNT(\*)

FROM customer

GROUP BY customer\_id,

first\_name,

last\_name,

email,

address\_id

HAVING COUNT(\*) >1; --no result set means we have no duplicates

Application

Description automatically generated with low confidence

*There are no duplicates in the customer table. I would follow the same protocol as mentioned above if data needed to be cleaned further.*

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.

*For the film table, numerical values I used the following query and switched rental\_rate out with: release\_year, rental\_duration, rental\_rate, length and replacement cost to find my information.*

SELECT MIN(rental\_rate) AS min\_rent,

MAX(rental\_rate) AS max\_rent,

AVG(rental\_rate) AS avg\_rent,

COUNT(rental\_rate) AS count\_rent\_values,

COUNT(\*) AS count\_rows

FROM film;

*For non-numerical values, I used the query and switched rating with title, description, last\_update and special\_feature to find the information.*

SELECT mode() WITHIN GROUP (ORDER BY rating)

AS modal\_value

FROM film;

*For the customer table, I followed the same steps as above, for numerical data I used store\_id, customer\_id, address\_id.*

SELECT MIN(store\_id) AS min\_store\_id,

MAX(store\_id) AS max\_store\_id,

AVG(store\_id) AS avg\_store\_id,

COUNT(store\_id) AS count\_store\_id,

COUNT(\*) AS count\_rows

FROM customer;

*For the non-numeric information, I used the same as above: using first\_name, last\_name, create\_date, last\_updated.*

SELECT mode() WITHIN GROUP (ORDER BY rating)

AS modal\_value

FROM film;

**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.

*SQL requires less instructions for retrieving the same data. On top of that, as the file grows, excel will slow down. There is also more room for errors as filtering in excel requires more steps than SQl, and each step is a chance for an error to be made. In SQL you are retrieving just the data you need, whereas in excel you are manipulating the whole data set, taking queries longer. SQL separates the data analysis from the data, reducing risk of data corruption by accident. Excel is good for working with data in a visual way, we were able to see all of our data on the spreadsheets in the Influenza task, making it easier to visualize, and we worked and interacted directly with the data, something that SQL does not do in the same way. For small amounts of data, excel is useful, but for larger data sets, SQL is much faster, and less prone to error/data corruption.*