Answers 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).

**Film Table**

**Duplicate Data:** SELECT title, release\_year, language\_id, COUNT(\*) FROM film GROUP BY title, release\_year, language\_id HAVING COUNT(\*) >1;

No Duplicates were found

**Non-Uniform Data**: --Returns only unique records from the film table SELECT DISTINCT release\_year, language\_id, rating FROM film GROUP BY release\_year, language\_id, rating

No non-uniform instances were found.

**Missing Data**: -- looking for missing values. SELECT DISTINCT title, release\_year,language\_id, rental\_duration FROM film WHERE title IS NULL;

No missing values were found.

**Customer Table**

**Duplicate Data:** SELECT first\_name, last\_name, address\_id, email, COUNT(\*) FROM customer GROUP BY first\_name, last\_name, address\_id, email HAVING COUNT(\*) >1;

No duplicates were found.

**Non-Uniform Data**: --Returns only unique records from the film table SELECT DISTINCT first\_name, last\_name, address\_id, email FROM customer;

Nonon-uniform data was found

**Missing Data**: -- looking for missing values. SELECT DISTINCT first\_name, last\_name, email, address\_id FROM customer WHERE address\_id IS NULL;

No missing data was found.

**How to deal with each of these situations:**

* In the case of missing **duplicate data**, I could create a virtual table called a view or delete the duplicate record from the table or view.
* In The case of **non-uniform data**, I could use the UPDATE command to fix it.
* In the case of **missing data**, I could choose to ignore that particular column or use the AVG command to input the missing values.

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.

**Film analysis**

SELECT

MIN(rental\_rate) AS min\_rent,

MAX(rental\_rate) AS max\_rent,

AVG(rental\_rate) AS avg\_rent,

MIN(rental\_duration) AS min\_duration,

MAX(rental\_duration) AS max\_duration,

AVG(rental\_duration) AS avg\_duration,

MIN(rental\_rate) AS min\_rent,

MAX(rental\_rate) AS max\_rent,

AVG(rental\_rate) AS avg\_rent,

MIN(replacement\_cost) AS min\_cost,

MAX(replacement\_cost) AS max\_cost,

AVG(replacement\_cost) AS avg\_cost,

MIN(length) AS min\_length,

MAX(length) AS max\_length,

AVG(length) AS avg\_length,

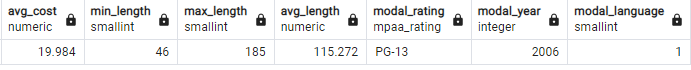
MODE() WITHIN GROUP (ORDER BY rating) AS modal\_rating,

MODE() WITHIN GROUP (ORDER BY release\_year) AS modal\_year,

MODE() WITHIN GROUP (ORDER BY language\_id) AS modal\_language

FROM film;

****



**Customer Analysis**

SELECT

MIN(customer\_id) AS min\_customer\_id,

MAX(customer\_id) AS max\_customer\_id,

MIN(store\_id) AS min\_duration,

MAX(store\_id) AS max\_duration,

MIN(address\_id) AS min\_address\_id,

MAX(address\_id) AS max\_address\_id,

MIN(create\_date) AS min\_date,

MAX(create\_date) AS max\_date,

MIN(last\_update) AS min\_update,

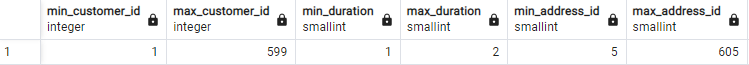
MAX(last\_update) AS max\_update,

MODE() WITHIN GROUP (ORDER BY first\_name) AS modal\_first,

MODE() WITHIN GROUP (ORDER BY last\_name) AS modal\_last,

MODE() WITHIN GROUP (ORDER BY activebool) AS modal\_active

FROM customer;



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.

**I think that SQL is a lot easier and quicker when it comes to analyzing data. In Excel theses certain commands may have to be done manually.**