Filtering data

cleaning data

duplicate data

```
select customer_id,
count(*)
from customer
group by customer_id
having count(*)>1;
select film_id,rating,title,
count(*)
from film
group by film_id,rating,title
having count(*)>1;
```

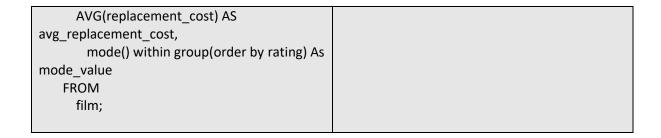
missing data

| select * | select * |
|------------------------|----------------------------|
| from film | from customer |
| where film_id is null; | where customer_id is null; |
| | |

There is no duplicate, non-uniform, missing, incorrect data in both customer and film tables.

Descriptive analysis: customer table & film table

| SELECT | SELECT |
|--------------------------------------|--------------------------------------|
| MIN(rental_duration) AS | MIN(store_id) AS min_store_id, |
| min_rental_duration, | MAX(store_id) AS max_store_id, |
| MAX(rental_duration) AS | MIN(customer_id) AS min_customer_id, |
| max_rental_duration, | MAX(customer_id) AS max_customer_id, |
| AVG(rental_duration) AS | |
| avg_rental_duration, | |
| MIN(rental_rate) AS min_rental_rate, | FROM |
| MAX(rental_rate) AS max_rental_rate, | customer; |
| AVG(rental_rate) AS avg_rental_rate, | |
| MIN(replacement_cost) AS | |
| min_replacement_cost, | |
| MAX(replacement_cost) AS | |
| max_replacement_cost, | |



Outputs of descriptive analysis on customer and film tables.

| min_re ntal_duratio n | _ | avg_re ntal_duratio n | min _rental_ra te | max _rental_ra te | avg_re ntal_rate | min_re placement_co st | max_re placement_co st | avg_rep lacement_cos t | m ode_val ue |
|-----------------------------|---|-----------------------------|-------------------------|-------------------------|----------------------------|------------------------------|------------------------------|------------------------------|--------------------|
| 3 | 7 | 4.9850 0000000000 00 | 0.99 | 4.99 | 2.9800 0000000000 00 | 9.99 | 29.99 | 19.9840 00000000000 0 | P G-13 |

| min_store_id | max_store_id | min_customer_id | max_customer_id |
|--------------|--------------|-----------------|-----------------|
| 1 | 2 | 1 | 599 |

Reflect on excel vs sql use case in filtering data

In the process of data cleaning, SQL offers superior efficiency. more important key point is a Data analyst should exercise caution and have solid justification when deleting duplicates, and not treat it as a default action. Excel simplifies the process of duplicate removal, but SQL provides an alternative perspective, allowing for the examination of unique record sets without immediate deletion. However, when it comes to data summarization, there is some complexity with SQL. some descriptive analysis requiring lengthy queries. In contrast, Excel's pivot table feature presents a more streamlined and effective method for such summarization tasks, bypassing the need for complex SQL queries.