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 *
from film
where film_id is null;
```

```
select *
from customer
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
    MIN(rental_duration) AS
min_rental_duration,
    MAX(rental_duration) AS
max_rental_duration,
    AVG(rental_duration) AS
avg_rental_duration,
    MIN(rental_rate) AS min_rental_rate,
    MAX(rental_rate) AS max_rental_rate,
    AVG(rental_rate) AS avg_rental_rate,
    MIN(replacement_cost) AS
min_replacement_cost,
    MAX(replacement_cost) AS
max_replacement_cost,
```

```
SELECT
  MIN(store_id) AS min_store_id,
  MAX(store_id) AS max_store_id,
  MIN(customer_id) AS min_customer_id,
  MAX(customer_id) AS max_customer_id,


FROM
  customer;
```

```
        AVG(replacement_cost) AS
avg_replacement_cost,
            mode() within group(order by rating) As
mode_value
      FROM
         film;
```

Outputs of descriptive analysis on customer and film tables.

min_rental_duration	max_rental_duration	avg_rental_duration	min_rental_rate	max_rental_rate	avg_rental_rate	min_replacement_cost	max_replacement_cost	avg_replacement_cost	mode_value
3	7	4.985000000000000	0.99	4.99	2.980000000000000	9.99	29.99	19.984000000000000	PG-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.