Achievement 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).

Non-uniform

```
1 SELECT DISTINCT rating
2 FROM film
3 GROUP BY rating

1 SELECT DISTINCT store_id
2 FROM customer
3 GROUP BY store_id
```

Using DISTINCT to check if any data are entered differently. If so, for example,

The above query will group all "gen" "g" and "General" and change them into "G" in col1.

Duplicate Data

```
1
     SELECT title,
 2
              release_year,
 3
              language_id,
 4
              rental_duration,
 5
              COUNT(*)
 6
     FROM film
 7
     GROUP BY title,
 8
                release_year,
 9
                language_id,
10
                rental_duration
     HAVING COUNT(*) >1
11
Data output
                          Notifications
              Messages
=+
                            release_year
                                          language_id
                                                       rental_duration
                                                                       count
                                                                               â
      character varying (255)
                            integer
                                          smallint
                                                       smallint
                                                                       bigint
     SELECT customer_id,
 2
              store_id,
 3
              first_name,
 4
              last_name,
 5
              COUNT(*)
 6
     FROM customer
 7
     GROUP BY customer_id,
              store_id,
 8
 9
              first_name,
10
              last_name
11
     HAVING COUNT(*) >1
Data output
             Messages
                          Notifications
=+
                 store_id
                              first_name
      customer_id
                                                   last_name
                                                                        count
                              character varying (45)
      [PK] integer
                   smallint
                                                   character varying (45)
                                                                         bigint
```

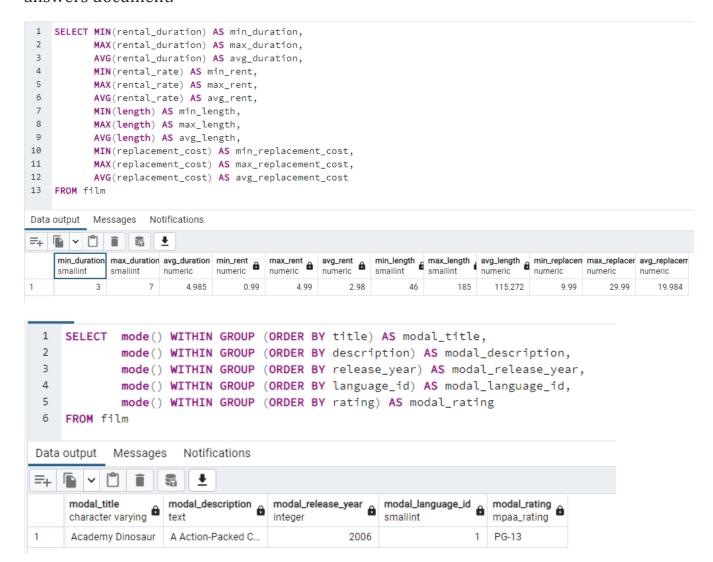
If there are duplicated data, we can create a unique view table by choosing the right column (CREATE VIEW viewname AS) or we can delete the data with query.

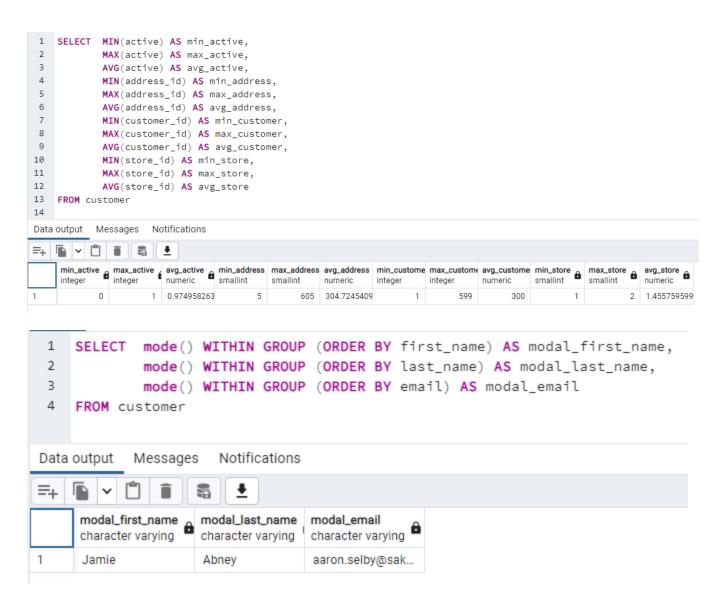
Missing Values

 If there are many missing values in a coloumn, we can just omit the column, and if there are a few missing data we can impute values such as using the AVG query.

```
1  UPDATE tablename
2  SET = AVG(col1)
3  WHERE col1 IS NULL
```

2. **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.





- 3. **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.
- It's more effective to use SQL for data profiling and getting information because SQL is able to collect all data from the database whereas in Excel you need to know what data to include first. Also, there can be many mistakes in Excel such as deleting or modifying data by accidents, whereas SQL will just show you the query is wrong, and you just need to adjust the query. Both programs need to have good understanding of the software to perform well. Overall, SQL is a better tool for company with lots of data and constantly pulling data from the database without making much mistakes.