Question 1:

<u>Check for and clean dirty data</u>: Find out if the film table and the customer table contain any dirty data, specifically non-uniform or duplicate data, or missing values. Next to each query write 2 to 3 sentences explaining how you would clean the data (even if the data is not dirty).

Table: film

(A) Data Accuracy

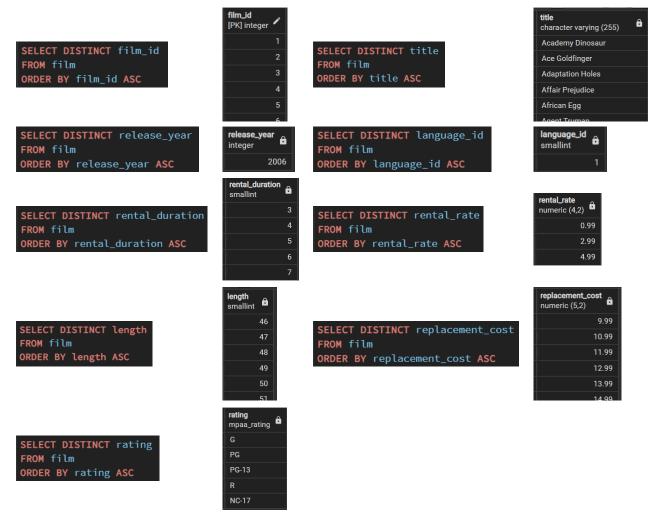
• Check whether the data is correct or not using MAX, MIN, and AVG for each column (applicable to quantitative data: release_year, rental_duration, rental_rate, length, replacement_cost).



- No issue was found in data accuracy. All the quantitative data is within a reasonable range.
- ➤ If there are any incorrect data, we have to refer to the source owner or another source recording the same data to find out the actual value. Sometimes we might be able to know the correct value by checking the other values of that column. For instance, in a year column with values of 2010, 2011, 2012, 20133, 2014 and 2015, we know that the year 20133 is incorrect and it should be 2013. After knowing the correct value, we could correct the data using the UPDATE command.

(B) Data Consistency

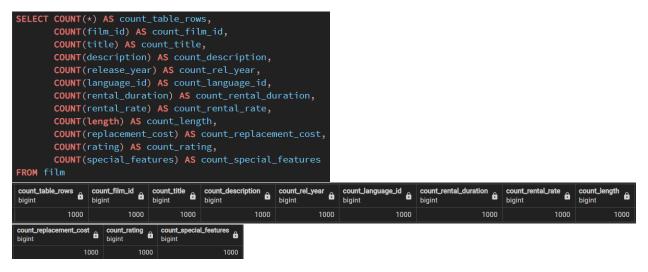
• Use DISTINCT to check if there are any format/spelling/naming issues in the values of each column.



- No issue was found in data consistency. All the columns have a consistent format with their values.
- ➤ If there are any format inconsistencies in the values, we could use the UPDATE command to correct the wrong format.

(C) Data Completeness

- Use DISTINCT to check if there are any values with null, 'NA', 'missing', 'unknown', or dummy values in each column (The queries are the same as those in the data consistency section, so I'm not going to paste the queries here again).
- Use COUNT to compare the number of values in each column to the total number of rows in the table.



- There is no null, dummy or missing value found in all the columns. The number of values of each column also tally with the total number of rows in the film table.
- If there is a column with a high percentage of missing values, we could simply omit that particular column in our SELECT command and leave a comment aside. If the missing values are less than 5% of the column, we could impute the missing values with estimates, say average, using the UPDATE command.

(D) Data Uniqueness

- Define the data grain of the table first. Then SELECT the data grain columns and COUNT all the rows. Show the rows that have more than one appearance.
- Data grain: title, release_year, language_id, rental_duration, rental_rate, length, replacement_cost, rating

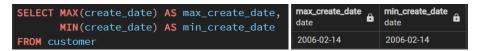
```
SELECT title,
      language_id,
      rental_duration,
ROM film
        language_id,
        length,
HAVING COUNT(\star) > 1 --no result set means we have no duplicates
                   language_id rental_duration
                                                           rental_rate
                                                                                  replacement_cost
                                                                                               mpaa_rating bigint
character varying (255)
                                                          numeric (4,2) a smallint
                                                                                                                     â
                                            smallint
                                                                                 numeric (5,2)
```

- From the result, there are no duplicates in the film table.
- If there are any duplicates, we could create a virtual table, known as a "view," where we select only unique records using the DISTINCT or GROUP BY clause.
- We could also delete the duplicates from the table using the DELETE command only if we have permission to do so.

Table: customer

(A) Data Accuracy

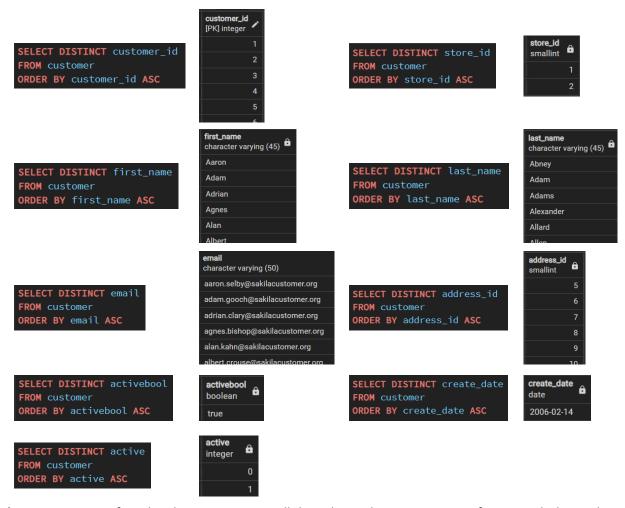
 Check whether the data is correct or not using MAX, MIN, and AVG for each column (applicable to quantitative data: create_date).



No issue was found in data accuracy. All the quantitative data is within a reasonable range.

(B) Data Consistency

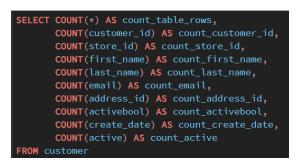
Use DISTINCT to check if there are any format/spelling/naming issues in the values of each column.



> No issue was found in data consistency. All the columns have a consistent format with their values.

(C) Data Completeness

- Use DISTINCT to check if there are any values with null, 'NA', 'missing', 'unknown', or dummy values in each column (The queries are the same as those in the data consistency section, so I'm not going to paste the queries here again).
- Use COUNT to compare the number of values in each column to the total number of rows in the table.

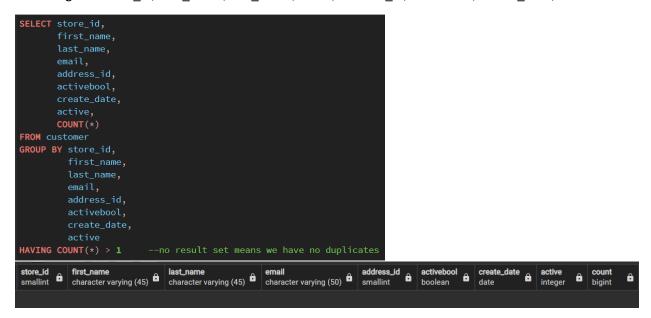




There is no null, dummy or missing value found in all the columns. The number of values of each column also tally with the total number of rows in the customer table.

(D) Data Uniqueness

- Define the data grain of the table first. Then SELECT the data grain columns and COUNT all the rows. Show the rows that have more than one appearance.
- Data grain: store id, first name, last name, email, address id, activebool, create date, active



From the result, there are no duplicates in the customer table.