DECEMBER 18, 2025

3.6 SUMMARIZING & CLEANING DATA IN SQL

JANELLA VALENCIA

Table of Contents

STEP 1- CHECK FOR AND CLEAN DIRTY DATA: 2

FILM DUPLICATES 2
FILM NON-UNIFORM DATA 2
FILM MISSING DATA 3
CUSTOMER DUPLICATES 3
CUSTOMER NON-UNIFORM DATA 4
CUSTOMER MISSING DATA 4

STEP 2- SUMMARIZE YOUR DATA: 5

NUMERIC FILM COLUMNS 5
NON-NUMERIC FILM COLUMN 5
NUMERIC CUSTOMER COLUMNS 5
NON-NUMERIC CUSTOMER COLUMN 5

STEP 3- REFLECT ON YOUR WORK: 6

Step 1- **Check for and clean dirty data:**

Film Duplicates

```sql
SELECT title, release_year,language_id,rental_duration,
COUNT(*)
FROM film
GROUP BY title, release_year,language_id,rental_duration
HAVING COUNT(*) >1
```

Data Output Messages Notifications

	title character varying (255)	release_year integer	language_id smallint	rental_duration smallint	count bigint

- No records showed at duplicates under the film table. In order to delete duplicate records I would create a view in order to visualize the unique records. I would then delete the records in question. Depending on my position in the company I would either be responsible for the deletion OR I wouldn't be, in that case I would use GROUP BY or DISTINCT to show unique records for analyzing purposes.

Film Non-Uniform Data

```sql
SELECT release_year,rating,
COUNT(*)
FROM film
GROUP BY release_year,rating
ORDER BY release_year,rating;
```

Data Output Messages Notifications

	release_year integer	rating mpaa_rating	count bigint
1	2006	G	178
2	2006	PG	194
3	2006	PG-13	223
4	2006	R	195
5	2006	NC-17	210

- No records showed as non-uniform data. I checked for release year and rating because I believed in the rating is where we may see the most non-uniform data. If there would have been non-uniformed data I would have used the UPDATE command to make the values uniformed.

Film Missing Data

```sql
SELECT *
FROM film
WHERE title IS NULL
   OR release_year IS NULL
   OR language_id IS NULL
   OR rental_duration IS NULL
   OR rental_rate IS NULL
   OR rating IS NULL;
```

Data Output Messages Notifications

film_id [PK] integer	title character varying (255)	description text	release_year integer	language_id smallint	rental_duration smallint	rental_rate numeric (4,2)	length smallint	replacement_cost numeric (5,2)	rating mpaa_rating	last_update timestamp witho

- There's no missing data for the above columns in film table. But if there was missing data I would check to see how much of it was missing. If 0%-5% missing it's okay to correct either by imputing the Avg of the data to complete the missing data. if 20%-30% or more missing data, it's better to ignore it and make a note of it in your analysis.

Customer Duplicates

```sql
SELECT first_name, last_name, email,
COUNT(*)
FROM customer
GROUP BY first_name, last_name, email
HAVING COUNT(*) >1
```

Data Output Messages Notifications

first_name character varying (45)	last_name character varying (45)	email character varying (50)	count bigint

- No duplicate records appeared. I searched for the first, last and email- this helped make a record unique enough in order to know if there are any duplicates. The email is most helpful to search for. If there was duplicate records, I would DELETE records, if allowed. If not I would use the GROUP BY or DISTINCT commands.

Customer Non-Uniform Data

```sql
SELECT customer_id, store_id,first_name,last_name, email, create_date, activebool
FROM customer
GROUP BY customer_id
ORDER BY customer_id
```

Data Output Messages Notifications

Showing rows: 1 to 599 Page No: 1

	customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)	email character varying (50)	create_date date	activebool boolean
1	1	1	Mary	Smith	mary.smith@sakilacustomer.org	2006-02-14	true
2	2	1	Patricia	Johnson	patricia.johnson@sakilacustomer.org	2006-02-14	true
3	3	1	Linda	Williams	linda.williams@sakilacustomer.org	2006-02-14	true
4	4	2	Barbara	Jones	barbara.jones@sakilacustomer.org	2006-02-14	true
5	5	1	Elizabeth	Brown	elizabeth.brown@sakilacustomer.org	2006-02-14	true
6	6	2	Jennifer	Davis	jennifer.davis@sakilacustomer.org	2006-02-14	true
7	7	1	Maria	Miller	maria.miller@sakilacustomer.org	2006-02-14	true
8	8	2	Susan	Wilson	susan.wilson@sakilacustomer.org	2006-02-14	true
9	9	2	Margaret	Moore	margaret.moore@sakilacustomer.org	2006-02-14	true
10	10	1	Dorothy	Taylor	dorothy.taylor@sakilacustomer.org	2006-02-14	true
11	11	2	Lisa	Anderson	lisa.anderson@sakilacustomer.org	2006-02-14	true
12	12	1	Nancy	Thomas	nancy.thomas@sakilacustomer.org	2006-02-14	true
13	13	2	Karen	Jackson	karen.jackson@sakilacustomer.org	2006-02-14	true
14	14	2	Betty	White	betty.white@sakilacustomer.org	2006-02-14	true
15	15	1	Helen	Harris	helen.harris@sakilacustomer.org	2006-02-14	true
16	16	2	Sandra	Martin	sandra.martin@sakilacustomer.org	2006-02-14	true
17	17	1	Donna	Thompson	donna.thompson@sakilacustomer.org	2006-02-14	true
18	18	2	Carol	Garcia	carol.garcia@sakilacustomer.org	2006-02-14	true
19	19	1	Ruth	Martinez	ruth.martinez@sakilacustomer.org	2006-02-14	true
20	20	2	Sharon	Robinson	sharon.robinson@sakilacustomer.org	2006-02-14	true

- There is no non- uniform data for customers. If there were non-uniformed data I would UPDATE command to make the consistent

Customer Missing Data

```sql
SELECT *
FROM customer
WHERE first_name IS NULL
   OR last_name IS NULL
   OR email IS NULL
   OR store_id IS NULL
   OR address_id IS NULL;
```

Data Output Messages Notifications

customer_id [PK] integer	store_id smallint	first_name character varying (45)	last_name character varying (45)	email character varying (50)	address_id smallint	activebool boolean	create_date date	last_update timestamp without time zone	active integer

- There is no missing data for customer table. I checked for the main columns. If there was missing data i would check to see within what percentage it fell. If it was 5% I would go ahead and use the AVG of the data if it was more than 30% I wouldn't use that information.

Step 2- **Summarize your data:**

- Use SQL to calculate descriptive statistics for both the film table and the customer table.

Numeric Film Columns

```sql
SELECT
MIN(rental_duration) AS min_rental,
MAX(rental_duration) AS max_rental,
AVG(rental_duration) AS avg_rental,
MIN(length) AS min_length,
MAX(length) AS max_length,
AVG(length) AS avg_length,
MIN(rental_rate) AS min_rate,
MAX(rental_rate) AS max_rate,
AVG(rental_rate) AS avg_rate,
MIN(replacement_cost) AS min_cost,
MAX(replacement_cost) AS max_cost,
AVG(replacement_cost) AS avg_cost
FROM film;
```

Data Output Messages Notifications

Showing rows: 1 to 1 Page No: 1

	min_rental smallint	max_rental smallint	avg_rental numeric	min_length smallint	max_length smallint	avg_length numeric	min_rate numeric	max_rate numeric	avg_rate numeric	min_cost numeric	max_cost numeric	avg_cost numeric
1	3	7	4.9850000000000000	46	185	115.2720000000000000	0.99	4.99	2.9800000000000000	9.99	29.99	19.9840000000000000

Non-Numeric Film Column

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY language_id)
AS modal_value
FROM film;
```

Data Output Messages Notifications

	modal_value smallint
1	1

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY rating)
AS modal_value
FROM film;
```

Data Output Messages Notifications

	modal_value mpaa_rating
1	PG-13

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY title)
AS modal_value
FROM film;
```

Data Output Messages Notifications

	modal_value character varying
1	Academy Dinosaur

Numeric Customer Columns

- No non numeric columns in customer table

Non-Numeric Customer column

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY first_name)
AS modal_value
FROM customer;
```

Data Output Messages Notifications

	modal_value character varying
1	Jamie

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY address_id)
AS modal_value
FROM customer;
```

Data Output Messages Notifications

	modal_value smallint
1	5

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY email)
AS modal_value
FROM customer;
```

Data Output Messages Notifications

	modal_value character varying
1	aaron.selby@sakilacustomer....

```sql
SELECT MODE() WITHIN GROUP
(ORDER BY activebool)
AS modal_value
FROM customer;
```

Data Output Messages Notifications

	modal_value boolean
1	true

Step 3- **Reflect on your work:**

- 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.
    - In excel we were able to go a step above and find stand deviations and correlation coefficient ; not to say we can't do that in SQL, but I haven't searched or learned it yet in the program. In excel we had to manually find the MIN,MAX,AVG for each section we needed. The advantage was that we were able to work with different sheets at once.
    - For SQL, I found it much quicker to get information like MIN,MAX or AVG in one SQL command for multiple columns. It was all arranged in a chart and easy to read.