# Ramon Magallan

Data Immersion 3.6

Summarizing & Cleaning Data

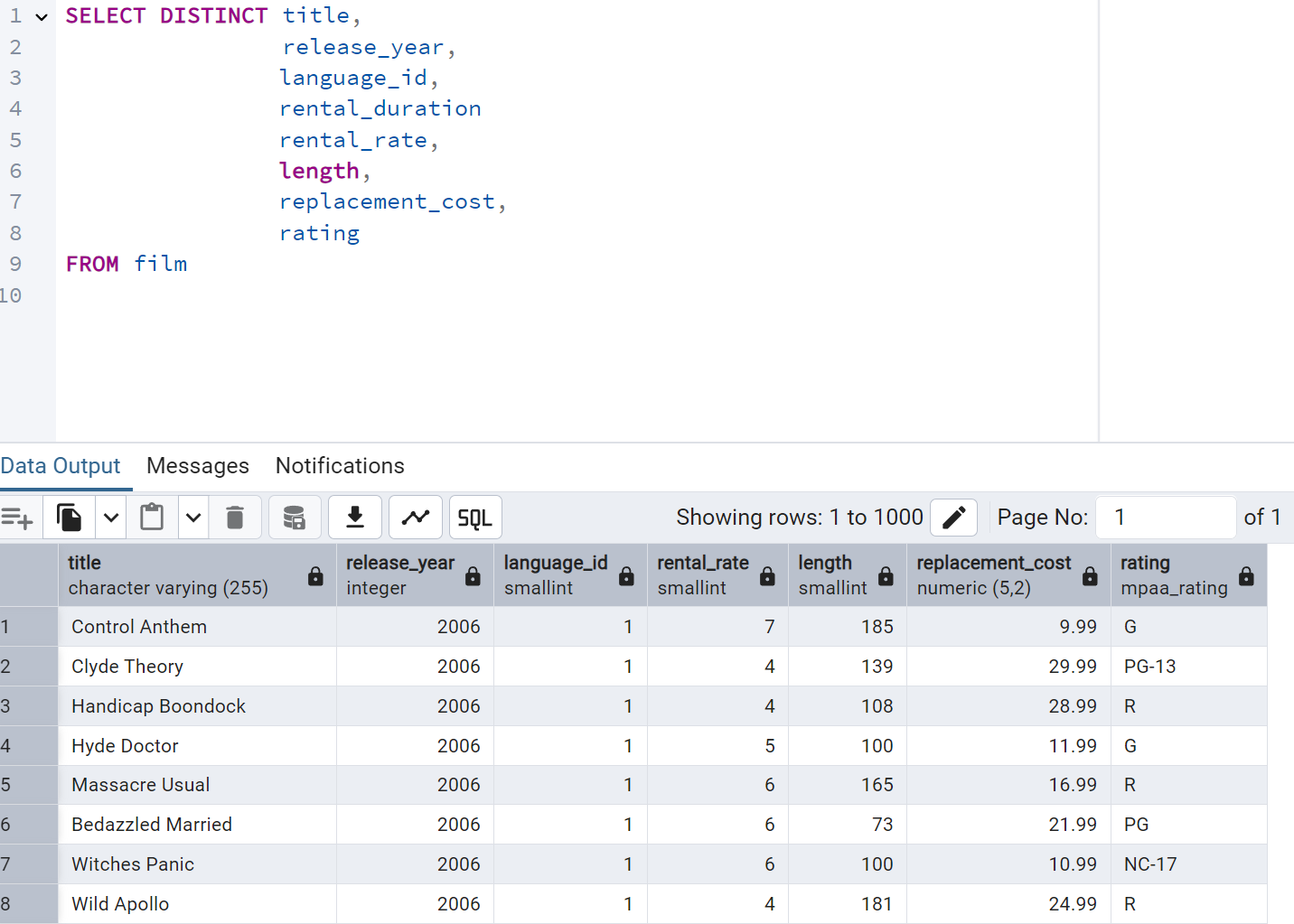
**1. Check For and Clean Dirty Data**

**Film table duplicate data:**

A screenshot of a computer

AI-generated content may be incorrect.   
There was no duplicate data found within the film table.

If there were duplicates, I would first create a virtual table. I would then select the unique records and delete the duplicate records from the view. After checking my work, I would execute the command within the table.

**Film non-uniform:  
**

The data appears to be fully uniform after a scroll through the table.  
  
If there were non-uniform entries, I would use the UPDATE command to correct any issues on the table.

**Film missing:**

**A computer screen with many different colored text

AI-generated content may be incorrect.**

There were no null values found

If there were columns with a large number of missing values, the column would be ignored.

**Customer duplicate data:**

**A screenshot of a computer

AI-generated content may be incorrect.**

There were no duplicates found within the customer table.

If there were duplicates, I would first create a virtual table. I would then select the unique records and delete the duplicate records from the view. After checking my work, I would execute the command within the table.

**Customer non-uniform:  
A screenshot of a computer

AI-generated content may be incorrect.**

There was no non-uniform data found in the customer table.  
  
If there were non-uniform entries, I would use the UPDATE command to correct any issues on the table.

**Customer missing:**

**A computer screen with text

AI-generated content may be incorrect.**

There were no nulls or missing data found within the customer table.

If there were columns with many missing values, the column would be ignored.

**2. Summarize Your Data**

Film

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure | Min | Max | Average | Count | Mode |
| Release Year | 2006 | 2006 | 2006 | 1000 |  |
| Rental Rate | 0.99 | 4.99 | 2.98 | 1000 |  |
| Rental Duration | 3 | 7 | 4.985 | 1000 |  |
| Length | 46 | 185 | 115.27 | 1000 |  |
| Replacement Cost | 9.99 | 29.99 | 19.984 | 1000 |  |
| Rating |  |  |  |  | PG-13 |
| language\_id |  |  |  |  | 1 |
| special\_features |  |  |  |  | {Trailers, Commentaries, "Behind the Scenes"} |

**Film numerical  
A screenshot of a computer

AI-generated content may be incorrect.  
A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Film non-numerical**

**A screenshot of a computer

AI-generated content may be incorrect.**

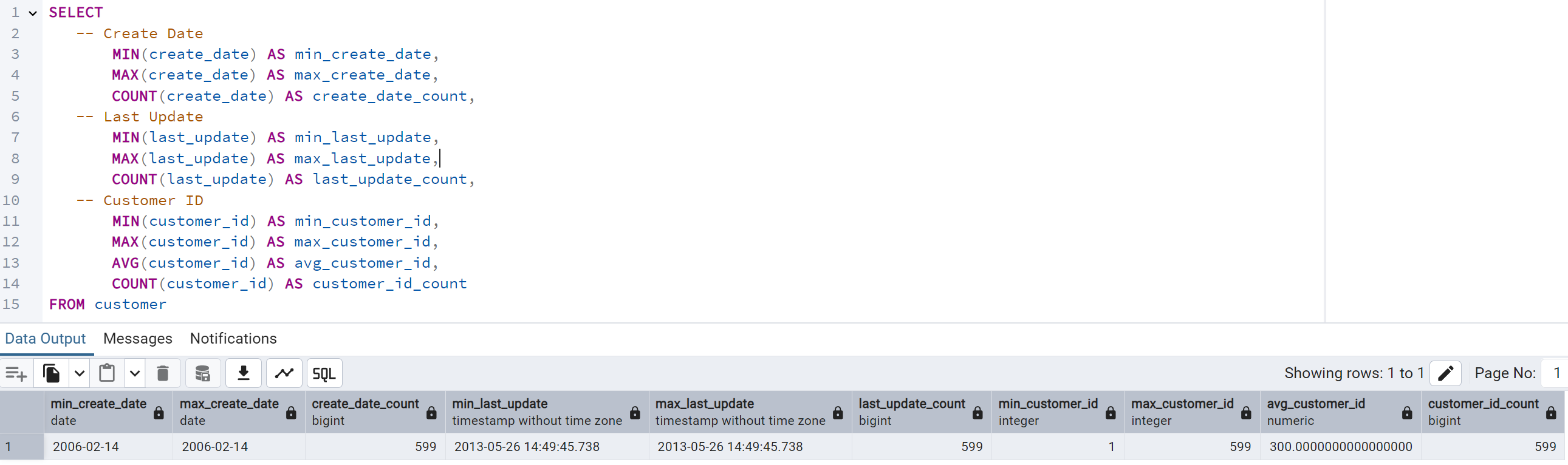
**Customer**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure | Min | Max | Average | Count | Mode |
| Store ID | 1 | 1/2/1900 | 1.456 | 599 |  |
| Address ID | 1/5/1900 | 8/27/1901 | 304.725 | 599 |  |
| Create Date | 2/14/2006 | 2/14/2006 |  | 599 |  |
| Last Update | 5/26/2013 | 5/26/2013 |  | 599 |  |
| Customer ID | 1 | 599 | 300 | 599 |  |
| Active |  |  |  |  | 1 |
| activebool |  |  |  |  | TRUE |
| First Name |  |  |  |  | Jamie |
| Last Name |  |  |  |  | Abney |
| email |  |  |  |  | [aaron.selby@sakilacustomer.org](mailto:aaron.selby@sakilacustomer.org) |

**Customer numerical**

**A screenshot of a computer

AI-generated content may be incorrect.**

****

**Customer non-numerical**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**3. Reflect on your work:**

After working with SQL and Excel I would see that SQL is a much more powerful tool when it comes to data profiling. Excel is very user-friendly and great for visualizing and quickly exploring small datasets. SQL is far more powerful when used with large datasets. Being able to compile a table pulling from multiple tables within the database is something that requires much preparation and time with excel. It can be done with just a few lines of query in SQL. SQL is also faster with filtering and summarizing through commands like SELECT, GROUP BY, and DISTINCT. Excel is a great tool for getting started or working with small sets of data, though SQL is the clear winner for large scale data jobs.