Netflix Data Cleaning and Analysis

Tools Used: SQL & Python

Objective: Clean the Netflix dataset, handle inconsistencies, and extract insights for better content understanding and decision-making.

1. Project Objective

To clean and normalize the Netflix dataset, resolve inconsistencies, and perform insightful analysis to support content-based decisions.

2. SQL-Python Integration

- Connected SQL Server with Python to handle raw data efficiently.
- Optimized data types in the netflix_raw table:
 - Changed varchar to nvarchar for handling foreign characters.
 - Limited maximum length to improve performance.

3. Data Cleaning in SQL

Duplicate Handling

- Checked for duplicates based on title and type.
- Removed 3 exact duplicates using ROW_NUMBER().
- Final count: **8807** → **8804** rows.

Table Normalization

Split multi-valued columns into separate tables using STRING_SPLIT():

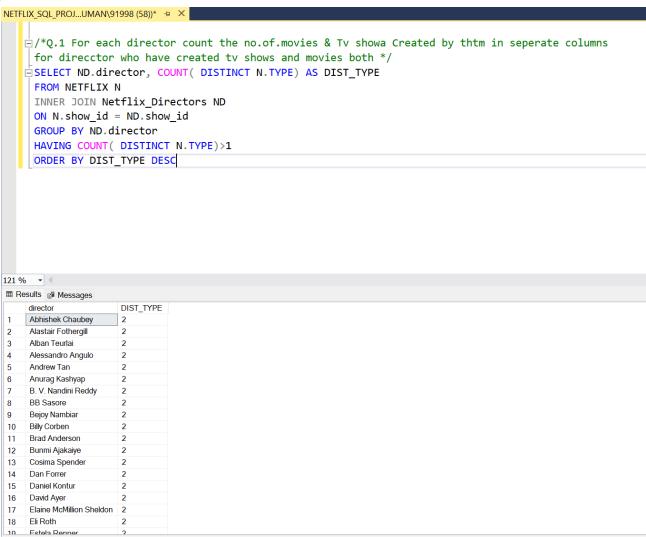
- Netflix_Directors
- Netflix_Cast
- Netflix_Country
- Netflix_Genre

Missing Values

- Country: Populated missing values using existing director-country relationships.
- **Duration:** Corrected misplaced values from rating to duration column (3 records).
- Final table created as Netflix.

📊 4. Data Analysis & Insights

Q1. Directors who created both Movies & TV Shows



- 83 directors identified.
- Used conditional aggregation to get counts:
 - NO_OF_MOVIES
 - o NO_OF_TVSHOWS

📌 Q2. Country with Highest Number of Comedy Movies

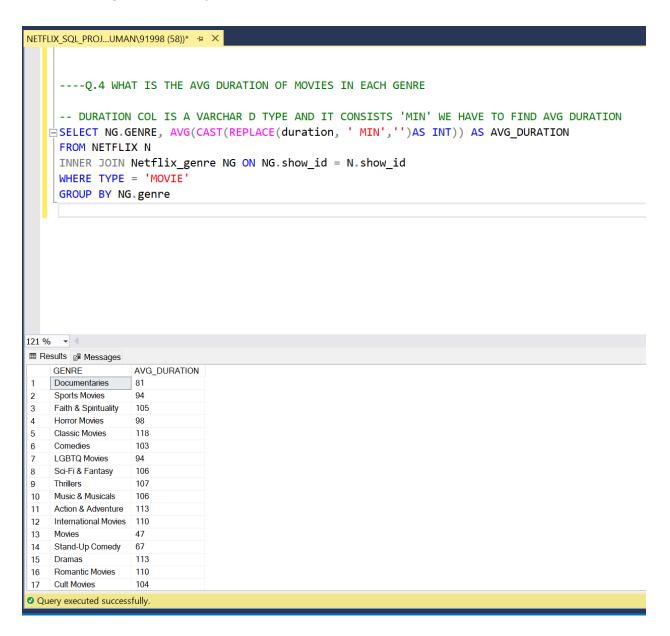
- Filtered by genre = 'Comedies' and type = 'Movie'
- Top contributing country identified.

📌 Q3. Most Active Directors by Year

```
NETFLIX_SQL_PROJ...UMAN\91998 (58))* ≠ ×
     --Q.3 FOR EACH YEAR (AS PER ADDED TO NETFLIX), WHICH DIRECTOR HAS MAX NO.OF.MOVIES RELEASED
   ⊟WITH CTE AS (
     SELECT ND.director, YEAR(date_added) AS YEAR,
     COUNT(DISTINCT N.show_id) AS NO_OF_MOVIES
     FROM Netflix_Directors ND
     INNER JOIN NETFLIX N ON N.show_id=ND.show_id
     WHERE TYPE = 'MOVIE'
     GROUP BY ND.director, YEAR(DATE_ADDED)
     CTE2 AS (
     SELECT *,
     ROW_NUMBER() OVER (PARTITION BY YEAR ORDER BY NO_OF_MOVIES DESC, DIRECTOR) AS RN
     --ORDER BY YEAR, NO_OF_MOVIES DESC
     SELECT * FROM CTE2 WHERE RN = 1
121 % -
■ Results ■ Messages
                   YEAR NO_OF_MOVIES RN
    director
   Sorin Dan Mihalcescu 2008
   Joe Dante
                   2009
3
    Jim Monaco
                    2010
   Arthur Allan Seidelman 2011 1
   Constance Marks 2012 1
   Dave Higby
                   2013 1
                   2014 2
    Leo Riley
                   2015 2
   Jay Karas
   Jan Suter
                   2016 4
10 Jay Chapman11 Jan Suter
                   2017 7
2018 12
12 Cathy Garcia-Molina
                   2019 7
13 Youssef Chahine
                   2020 11
14 Rajiv Chilaka
                   2021 17
```

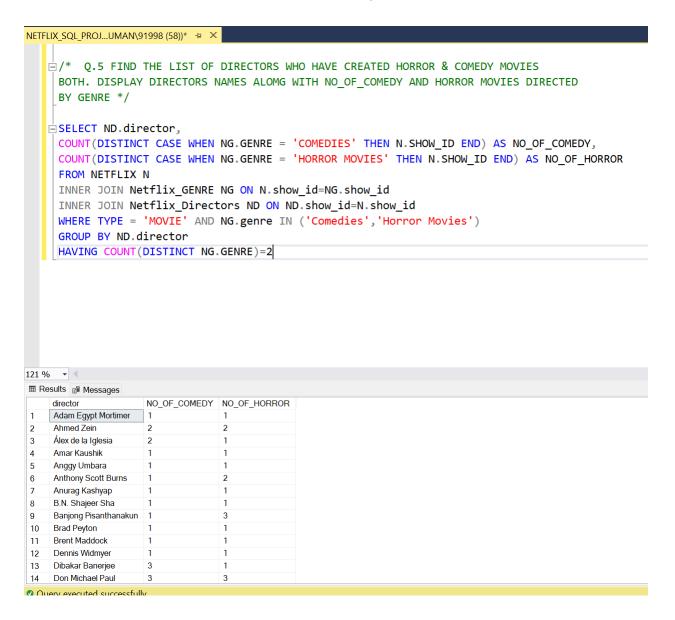
- Used YEAR(date_added)
- Applied ROW_NUMBER() to rank top directors per year.

📌 Q4. Average Duration by Genre



- Cleaned duration field (removed 'min' and cast to INT).
- Calculated average duration grouped by genre.

♣ Q5. Directors Who Made Both Horror & Comedy Movies



- Identified using genre IN ('Comedies', 'Horror Movies')
- Listed directors with counts of both genres.

5. Final Outcome

- Cleaned and normalized dataset ready for advanced analytics.
- Multiple business questions answered through SQL analysis.
- Project showcased data wrangling, transformation, and insight generation.