Report: Netflix Data Wrangling
Name: Daniel Muthama
Date: 20 May 2025
Description: Clean and prepare the Netflix dataset for analysis by addressing missing values, duplicates, and inconsistencies.

1. Introduction

This report documents the data wrangling process applied to the Netflix Titles dataset (Kaggle, 2025). The goal was to clean and structure the dataset for analysis by addressing missing values, duplicates, formatting errors, and logical inconsistencies. Key tasks included data discovery, structuring, cleaning, error checking, and validation. The final cleaned dataset is exported for further analysis.

2. Data Discovery

The dataset contains 8,807 rows and 12 columns with the following structure:

Key Findings:
Columns: show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description.

Missing Values:

director: 2,634 missing (29.9%)

cast: 825 missing (9.37%)

country: 831 missing (9.44%)

Minor missing values in date_added (0.11%), rating (0.05%), and duration (0.03%).

Duplicates: 0 duplicate rows identified.

Data Types:

date_added stored as object (needs conversion to datetime).

duration stored as strings (e.g., "90 min").

Code Output:

```
print("Missing values per column:\n", df.isnull().sum())
```

```
director      2634
cast           825
country        831
date_added      10
rating           4
duration         3
```

3. Structuring the Data
Actions Taken:
Convert date_added to datetime:

```
df['date_added'] = pd.to_datetime(df['date_added'], format='%B %d, %Y', errors='coerce')
```

Split duration into numeric and unit:

```
df[['duration_value', 'duration_unit']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')
df['duration_value'] = pd.to_numeric(df['duration_value'], errors='coerce')
```

Extract primary country (first country in the list):

```
df['primary_country'] = df['country'].str.split(',').str[0].str.strip()
```

Split listed_in into genres list:

```
df['genres'] = df['listed_in'].str.split(', ')
```

Fix for Unhashable List Error:
To resolve TypeError: unhashable type: 'list' during duplicate removal, lists were converted to tuples:

```
df['genres'] = df['genres'].apply(lambda x: tuple(x) if isinstance(x, list) else x)
```

4. Cleaning the Data
Actions Taken:
Drop duplicates and unused columns:

```
df.drop_duplicates(inplace=True)
```

```
df.drop(columns=['description'], inplace=True)
```

Impute missing directors:

Created a dir_cast column linking directors and cast.

Imputed directors appearing ≥3 times with the same cast.

```
df['director'].fillna('Not Given', inplace=True)
```
Impute missing countries using director-country relationships.

Handle remaining missing values:

cast: Filled with "Not Given".

Dropped rows with missing date_added, rating, or duration.

Output After Cleaning:

```
print("Missing values after cleaning:\n", df.isnull().sum())
director        0
cast            0
country         0
date_added      0
rating          0
duration        0
```

5. Error Checking

Identified Issues:

Date inconsistencies: 6 records had date_added years earlier than release_year.

```
invalid_dates = df[df['date_added'].dt.year < df['release_year']]  # Output: 6 rows
```
Invalid duration_unit values: Retained only "min", "Season", or "Seasons".

6. Validation

Final Checks:

Data Types:

```
print(df.dtypes)
date_added       datetime64[ns]
duration_value          float64
```

Business Rules:

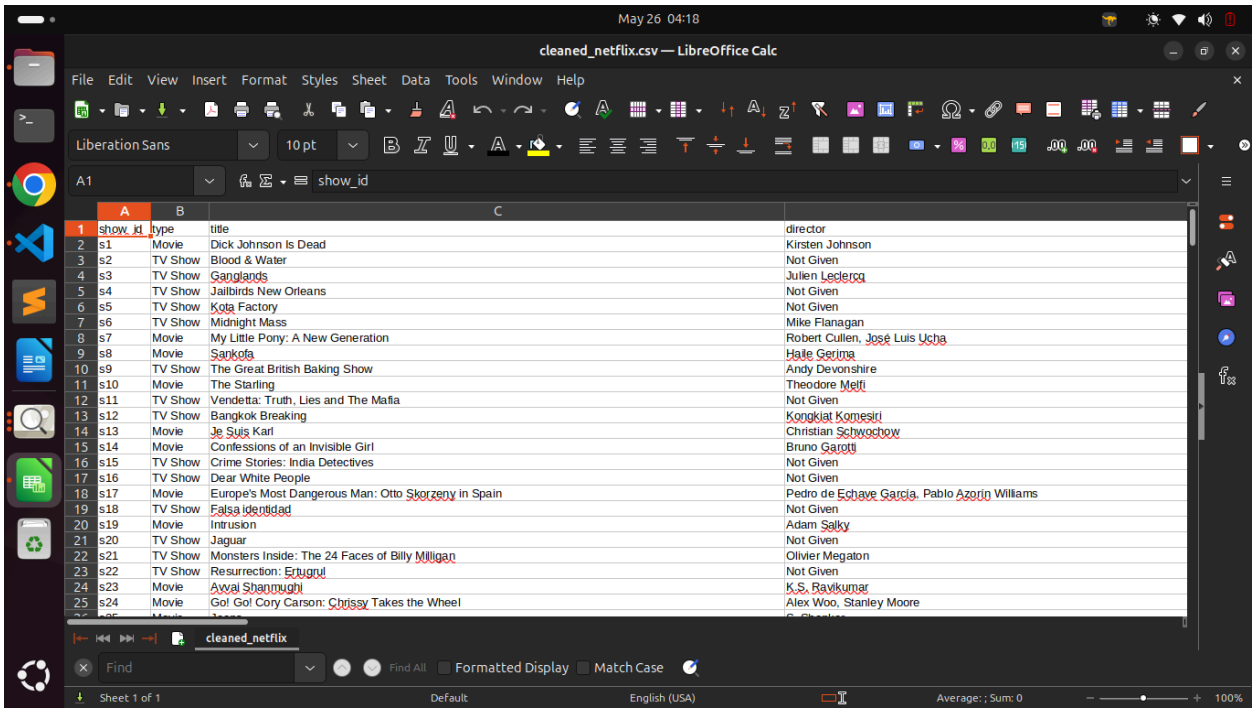
Entries before 1997: 407 retained (valid historical data).

Completeness:

Zero missing values in critical fields.

Sample Data:

- df.sample(3)



	A	B	C	
1	show_id	type	title	director
2	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson
3	s2	TV Show	Blood & Water	Not Given
4	s3	TV Show	Ganglands	Julien Leclercq
5	s4	TV Show	Jailbirds New Orleans	Not Given
6	s5	TV Show	Kota Factory	Not Given
7	s6	TV Show	Midnight Mass	Mike Flanagan
8	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha
9	s8	Movie	Sankofa	Haile Gerima
10	s9	TV Show	The Great British Baking Show	Andy Devonshire
11	s10	Movie	The Starling	Theodore Melfi
12	s11	TV Show	Vendetta: Truth, Lies and The Mafia	Not Given
13	s12	TV Show	Bangkok Breaking	Kongkiat Komesiri
14	s13	Movie	Je Suis Karl	Christian Schwochow
15	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti
16	s15	TV Show	Crime Stories: India Detectives	Not Given
17	s16	TV Show	Dear White People	Not Given
18	s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave Garcia, Pablo Azorin Williams
19	s18	TV Show	Falsa identidad	Not Given
20	s19	Movie	Intrusion	Adam Salky
21	s20	TV Show	Jaguar	Not Given
22	s21	TV Show	Monsters Inside: The 24 Faces of Billy Milligan	Olivier Megaton
23	s22	TV Show	Resurrection: Ertugrul	Not Given
24	s23	Movie	Avval Shanmughi	K.S. Ravikumar
25	s24	Movie	Go! Go! Cory Carson: Chrissy Takes the Wheel	Alex Woo, Stanley Moore

7. Export the Cleaned Data

Final Dataset:

Rows: 8,774 (after cleaning).

Columns: 14 (including structured fields).

Exported File: cleaned_netflix.csv.

df.to_csv('cleaned_netflix.csv', index=False)

8. Conclusion
The dataset is now cleaned and structured for analysis.

Key achievements include:

Imputed missing values using logical relationships (director-cast-country).

Removed invalid records (dates, durations).

Ensured consistency in data types and formatting.

Kaggle Notebook: Link:
GitHub: Link

Submitted by: Daniel Muthama