Data and Artificial Intelligence
Cyber Shujaa Program

Week 2 Assignment
Data Wrangling

Student Name: Austin Githinji

Student ID: CS-EH03-25417

Introduction

The purpose of this assignment was to gain hands-on experience in data wrangling by using the Netflix dataset from Kaggle. Data wrangling is an essential skill in data science that involves cleaning, structuring, and enriching raw data into a format suitable for analysis and visualization. This project involved working with the Netflix Movies and TV Shows dataset by Shivam Bansal on Kaggle. The dataset includes metadata about Netflix titles such as type, director, cast, country, release year, duration, and genres.

Tasks Completed

1. Loading and Exploring the Dataset

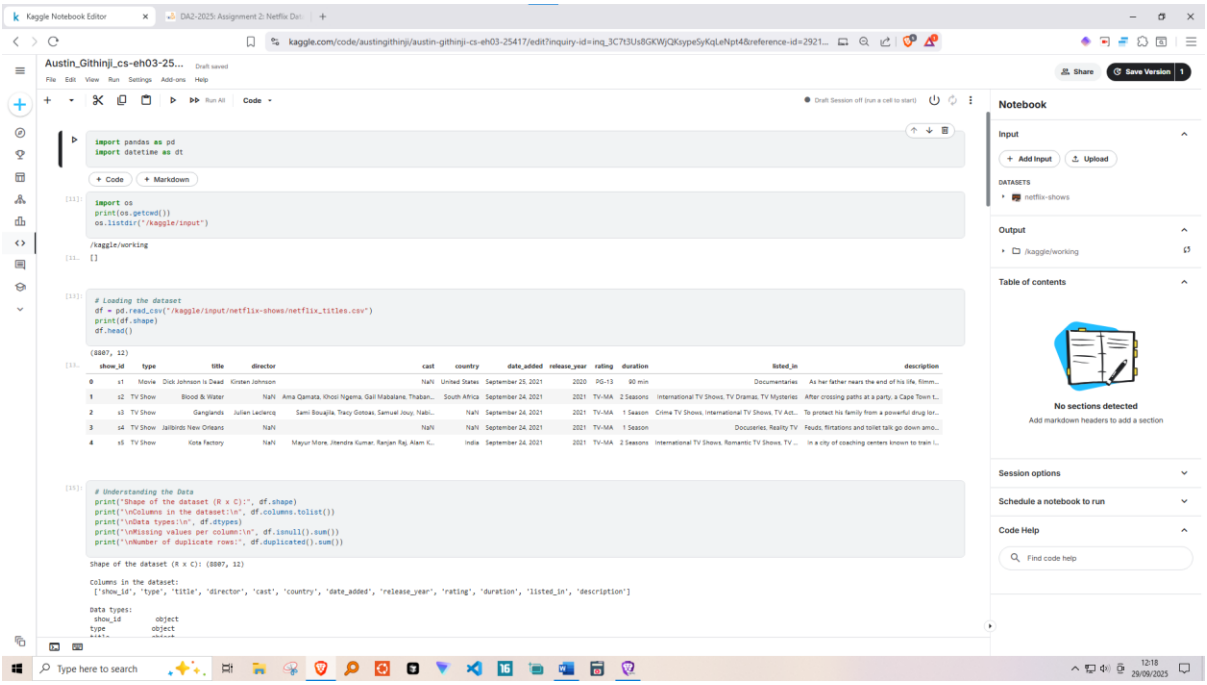
I loaded the dataset using pandas and checked its structure, shape, and missing values.

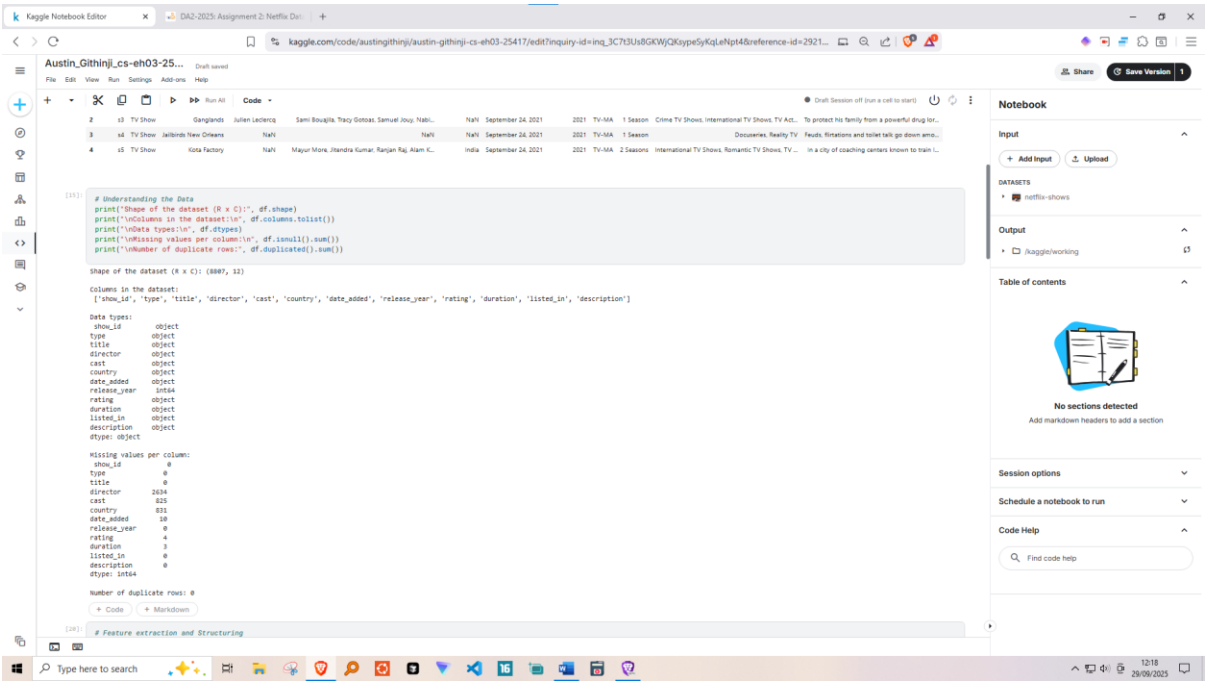
Code:

```
import pandas as pd

import datetime as dt

# Load dataset

df = pd.read_csv("/kaggle/input/netflix-shows/netflix_titles.csv")

# Explore structure

print("Shape of the dataset (R x C):", df.shape)

print("\nColumns in the dataset:\n", df.columns.tolist())

print("\nData types:\n", df.dtypes)
```

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

print("\nNumber of duplicate rows:", df.duplicated().sum())





2. Data Discovery

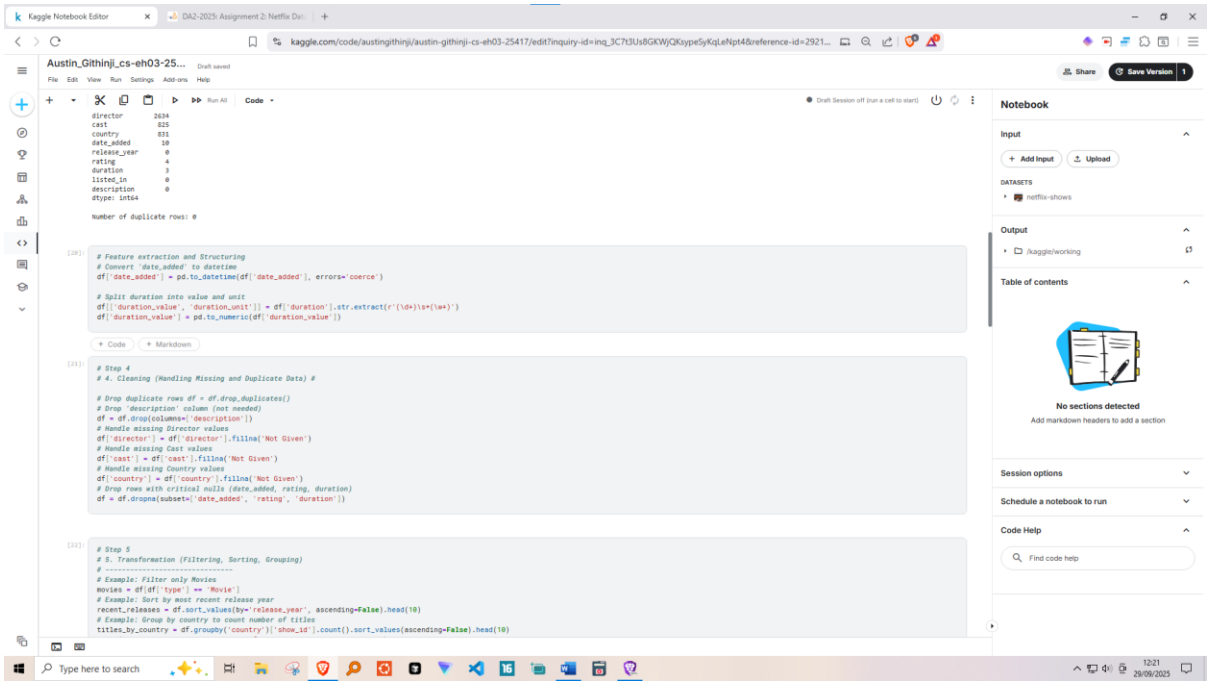
I checked data types, identified missing values, and looked for duplicates and inconsistencies.

Code:

df.info()

df.isnull().sum()

df.duplicated().sum()



3. Structuring the Data

I converted date_added to datetime format and extracted numeric values and units from the duration column.

Code:

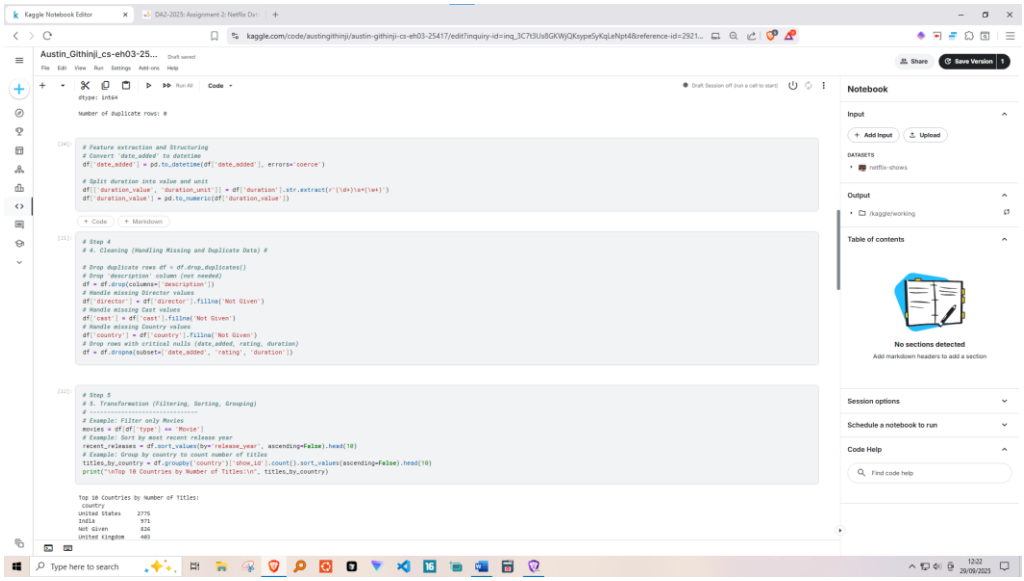
Convert 'date_added' to datetime

df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')

Split duration into value and unit

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

df['duration_value'] = pd.to_numeric(df['duration_value'])

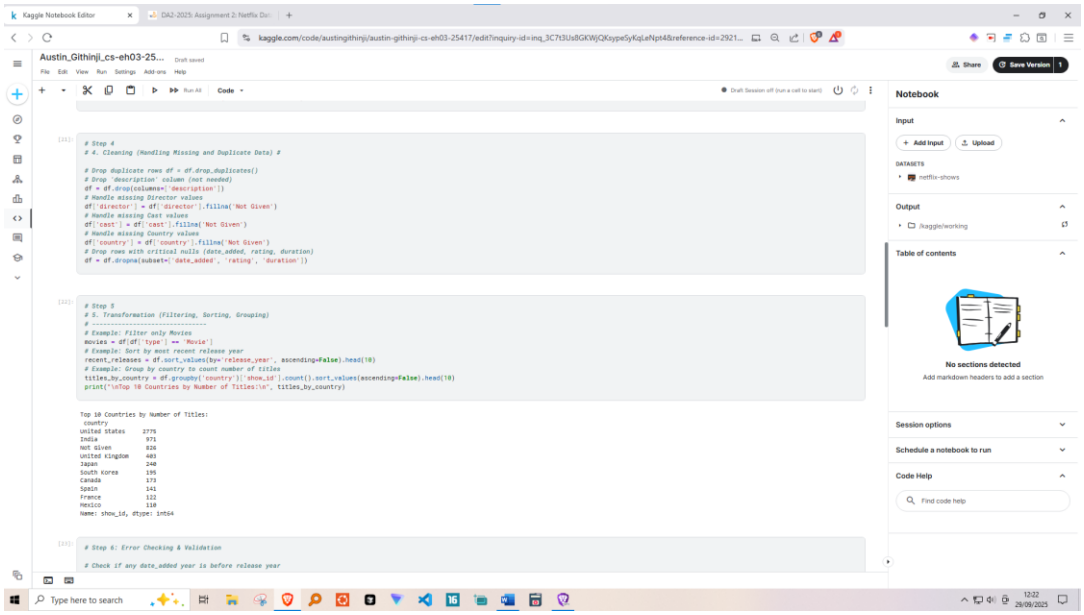


4. Cleaning the Data

I removed duplicate rows, dropped unnecessary columns, and handled missing values.

Code:

```
# Drop duplicates

df = df.drop_duplicates()

# Drop 'description' column

df = df.drop(columns=['description'])

# Fill missing values

df['director'] = df['director'].fillna('Not Given')

df['cast'] = df['cast'].fillna('Not Given')

df['country'] = df['country'].fillna('Not Given')

# Drop rows with critical nulls

df = df.dropna(subset=['date_added', 'rating', 'duration'])
```

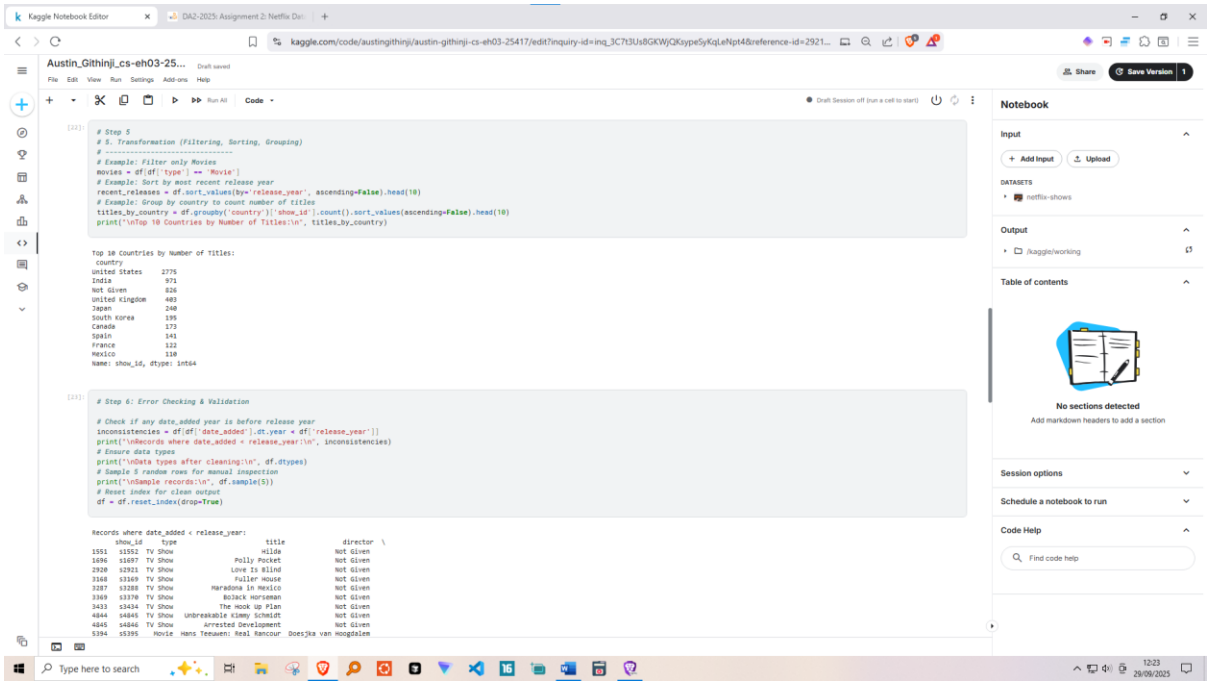


5. Transformation and Enrichment

I applied filtering, sorting, and grouping to gain insights.

Code:

```
# Filter only Movies
movies = df[df['type'] == 'Movie']
# Sort by most recent release year
recent_releases = df.sort_values(by='release_year', ascending=False).head(10)
# Group by country and count number of titles
titles_by_country = df.groupby('country')['show_id'].count().sort_values(ascending=False).head(10)
print(titles_by_country)
```

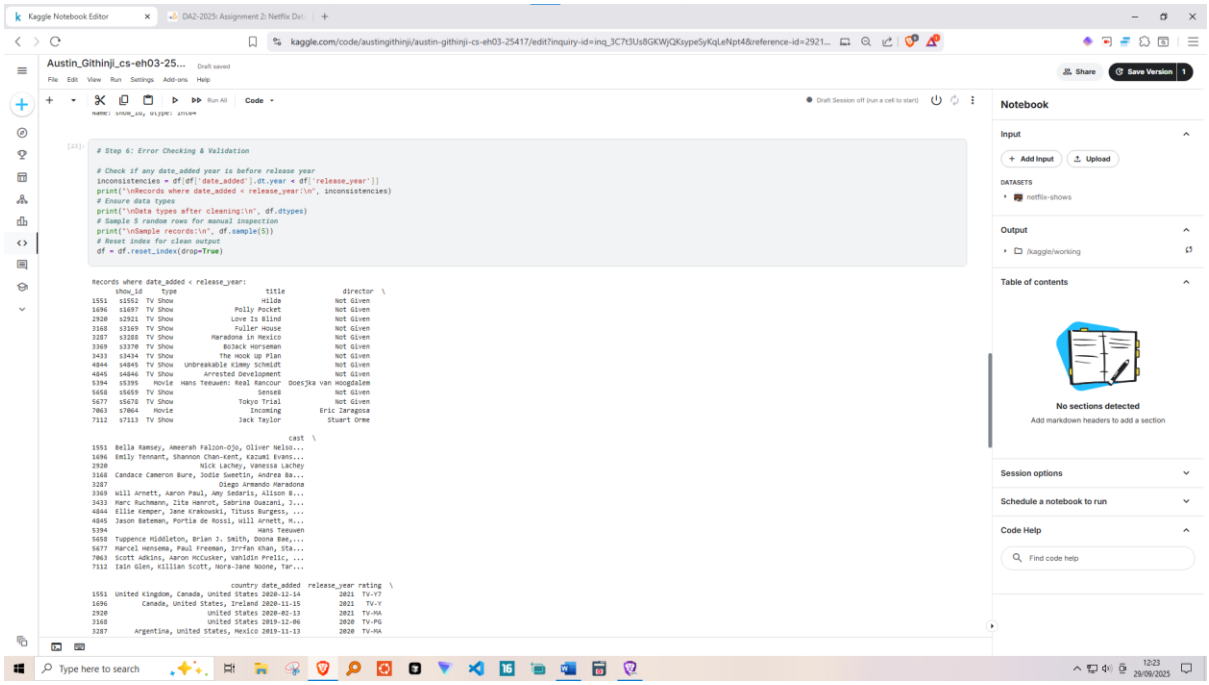


6. Validation

I checked for inconsistencies, confirmed data types, and sampled random rows.

Code:

```
# Check logical consistency

inconsistencies = df[df['date_added'].dt.year < df['release_year']]

print(inconsistencies)

# Check data types

print(df.dtypes)

# Sample 5 random rows

print(df.sample(5))

# Reset index

df = df.reset_index(drop=True)
```

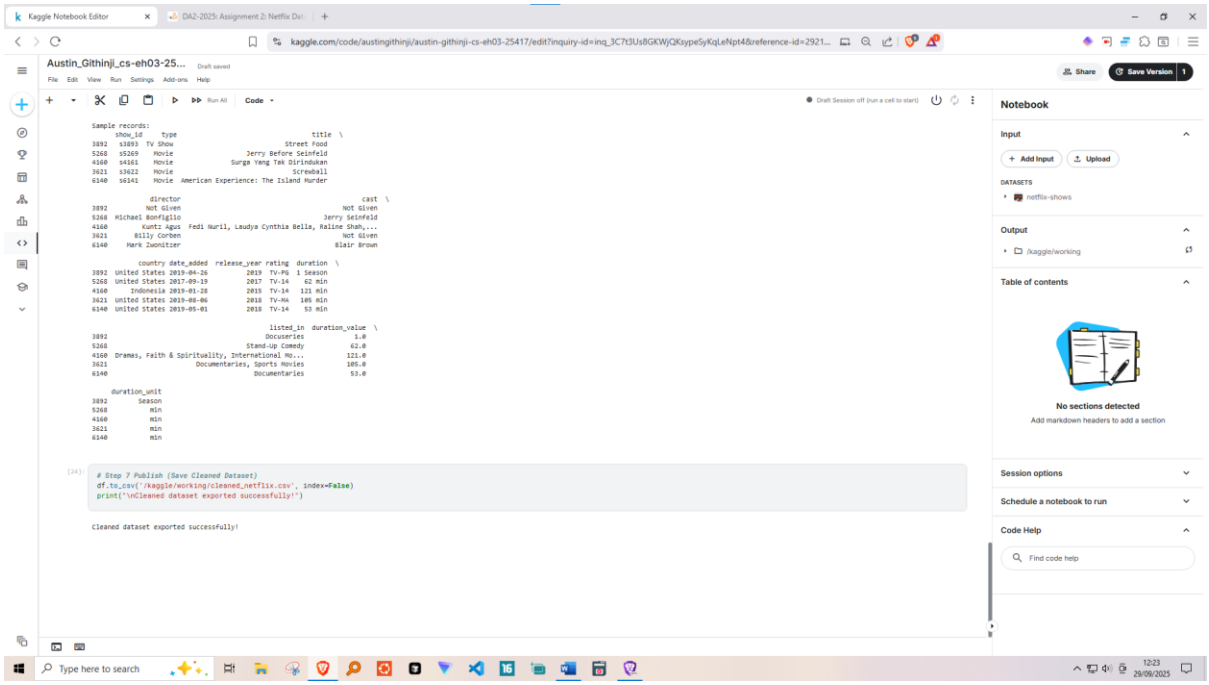


7. Exporting the Dataset

Finally, I exported the cleaned dataset as a CSV.

Code:

```
df.to_csv('/kaggle/working/cleaned_netflix.csv', index=False)

print("Cleaned dataset exported successfully!")
```



Conclusion

This project covered the full data wrangling workflow: loading, exploring, cleaning, structuring, transforming, validating, and exporting. I learned how to handle missing values, remove duplicates, split and reformat columns, and apply filtering, sorting, and grouping.

Kaggle Notebook Link

https://www.kaggle.com/code/austingithinji/austin-githinji-cs-eh03-25417

Conclusion

This week I gained a good grounding on the introductory concepts relating to data science and artificial intelligence. I am getting a better understanding that I can build on as we work on more advanced concepts in later weeks. I have posted my writeup on my blog and I look forward to building a portfolio that I can showcase on my CV as I look for jobs in Data and AI.