Data and Artificial Intelligence
Cyber Shujaa Program

Week 2 Assignment
Data Wrangling in Python

Student Name: Faith Jeptoo

Student ID: CS-DA02-25005

Table of Contents

Introduction

This week's assignment was to practice data wrangling concepts using the Netflix dataset from Kaggle.

The objectives were:
- Load the dataset and explore its structure.
- Discover data types, missing values, and quality issues.
- Clean the dataset by handling duplicates, missing values, and formatting inconsistencies.
- Transform and enrich the dataset.
- Validate and export the final dataset.

Tasks Completed

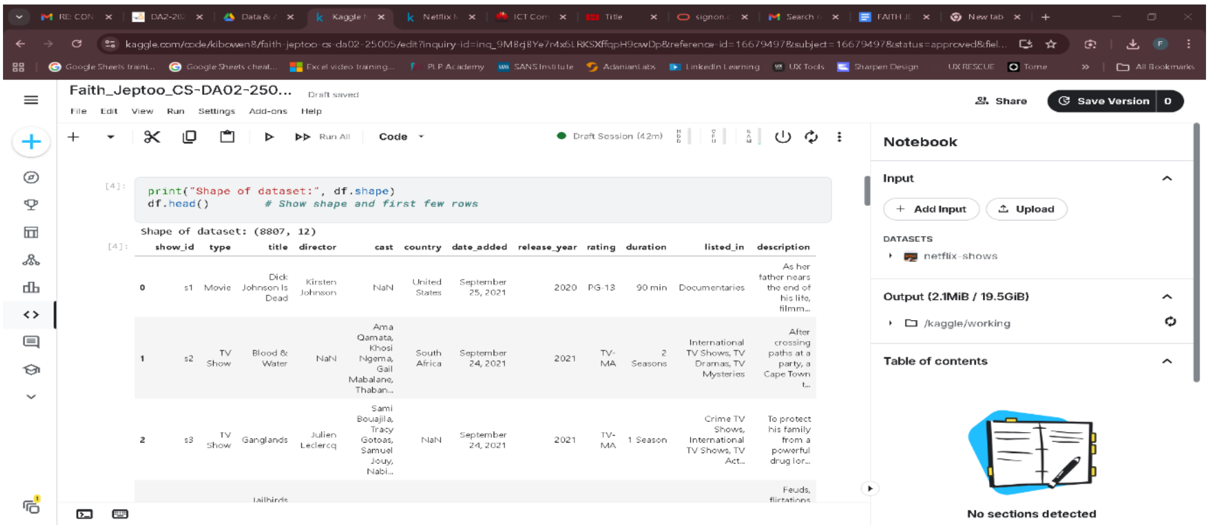
Step 1: Import Libraries and Load Dataset

```
import pandas as pd
import os

# Check current working directory
print(os.getcwd())

# Load Netflix dataset
filepath = '/kaggle/input/netflix-shows/netflix_titles.csv'
df = pd.read_csv(filepath)

# Show shape and first few rows
print("Shape of dataset:", df.shape)
df.head()
```



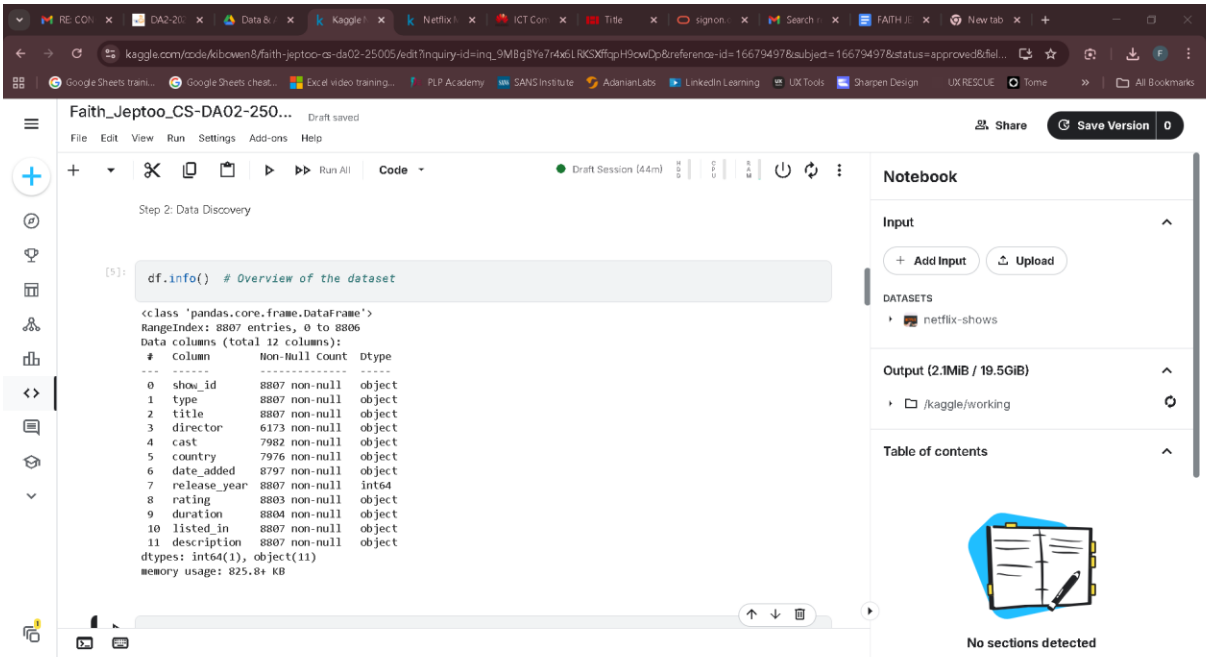
```
[4]: print("Shape of dataset:", df.shape)
     df.head()          # Show shape and first few rows
```

```
Shape of dataset: (8807, 12)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
			Jailbirds									Feuds, flirtations

Step 2: Data Discovery

```
# Overview of the dataset
df.info()
```



Step 2: Data Discovery

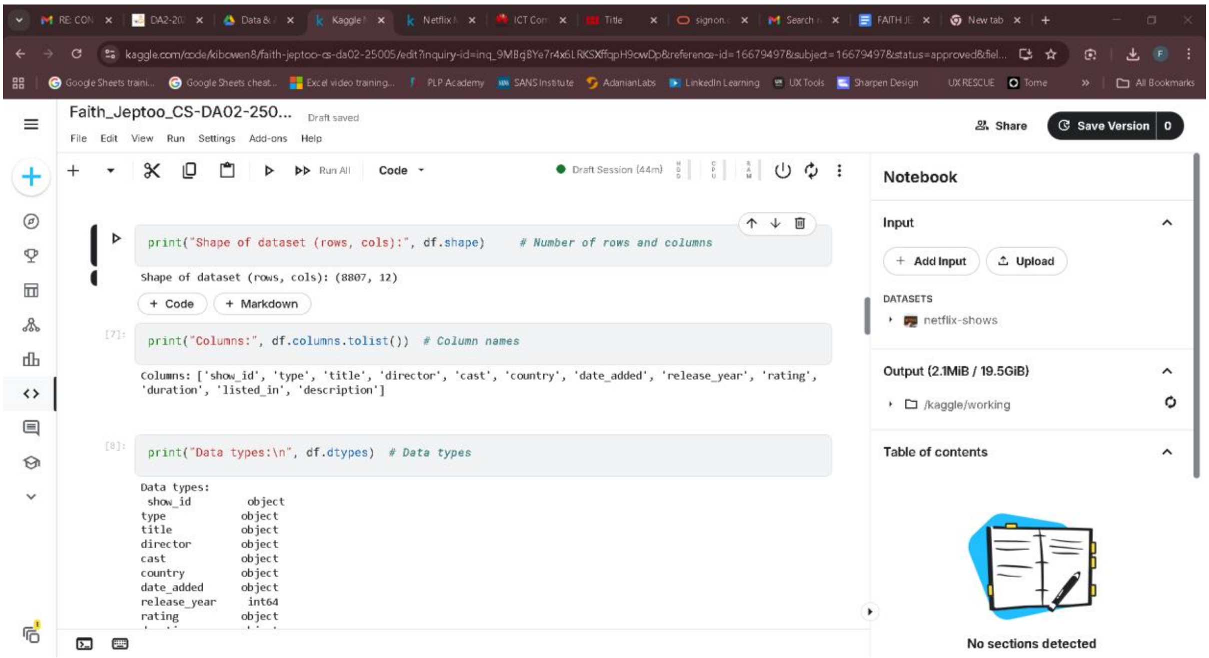
```
[5]: df.info()  # Overview of the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

```
# Number of rows and columns
print("Shape of dataset (rows, cols):", df.shape)

# Column names
print("Columns:", df.columns.tolist())

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



```
print("Shape of dataset (rows, cols):", df.shape)     # Number of rows and columns
```

```
Shape of dataset (rows, cols): (8807, 12)
```

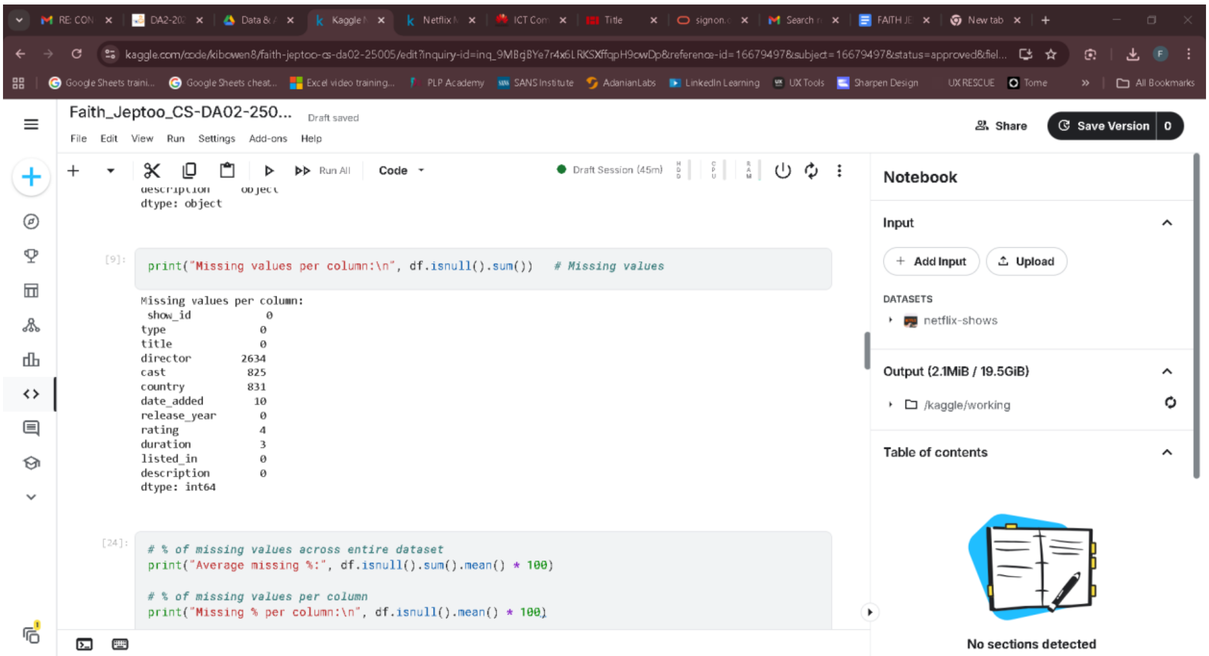
```
[7]: print("Columns:", df.columns.tolist())  # Column names
```

```
Columns: ['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating',
'duration', 'listed_in', 'description']
```

```
[8]: print("Data types:\n", df.dtypes)  # Data types
```

```
Data types:
 show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
```

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



Faith_Jeptoo_CS-DA02-250... Draft saved

```
description       object
dtype: object
```

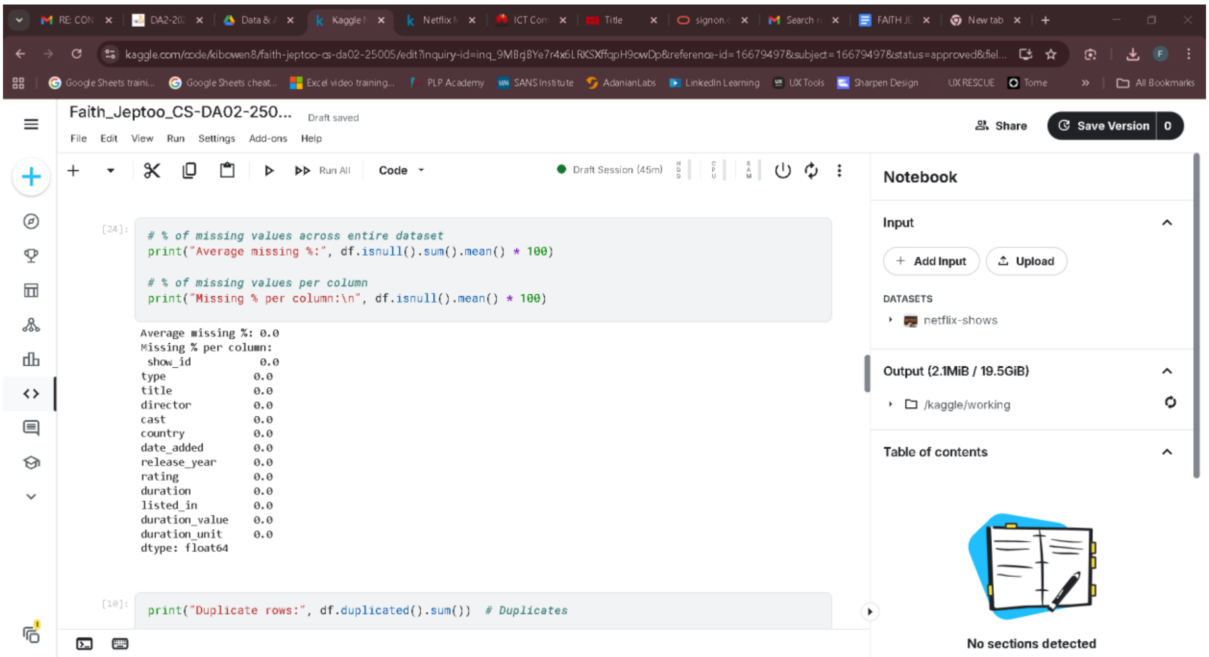
```
[9]: print("Missing values per column:\n", df.isnull().sum())   # Missing values
```

```
Missing values per column:
 show_id          0
type             0
title            0
director      2634
cast           825
country        831
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

```
[24]: # % of missing values across entire dataset
print("Average missing %:", df.isnull().sum().mean() * 100)

# % of missing values per column
print("Missing % per column:\n", df.isnull().mean() * 100)
```

```
# Missing values in percentage (overall and per column)
print("Average missing % across dataset:", df.isnull().sum().mean() * 100)
print("Missing % per column:\n", df.isnull().mean() * 100)
```



Faith_Jeptoo_CS-DA02-250... Draft saved

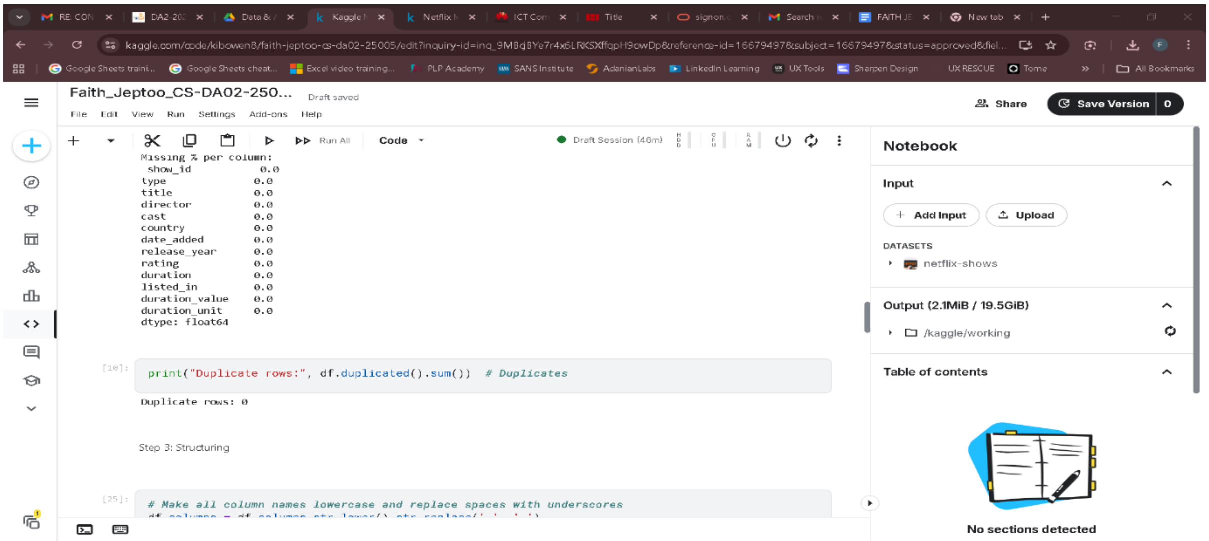
```
[24]: # % of missing values across entire dataset
print("Average missing %:", df.isnull().sum().mean() * 100)

# % of missing values per column
print("Missing % per column:\n", df.isnull().mean() * 100)
```

```
Average missing %: 0.0
Missing % per column:
 show_id          0.0
type            0.0
title           0.0
director        0.0
cast            0.0
country         0.0
date_added      0.0
release_year    0.0
rating          0.0
duration        0.0
listed_in       0.0
duration_value  0.0
duration_unit   0.0
dtype: float64
```

```
[10]: print("Duplicate rows:", df.duplicated().sum())  # Duplicates
```

```
# Duplicates
print("Duplicate rows:", df.duplicated().sum())
```


kaggle.com/code/kibowen8/faith-jeptoo-cs-da02-25005/edit?inquiry-id=inq_9MBqBYe7r4x6LRKSXffqpH9owDp&reference-id=16679497&subject=16679497&status=approved&fiel...
Faith_Jeptoo_CS-DA02-250...
Draft saved
File Edit View Run Settings Add-ons Help
Run All
Code
Draft Session (46m)
Share
Save Version
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.1MiB / 19.5GiB)
/kaggle/working
Table of contents
No sections detected

```
Missing % per column:
 show_id          0.0
type              0.0
title             0.0
director          0.0
cast              0.0
country           0.0
date_added        0.0
release_year      0.0
rating            0.0
duration          0.0
listed_in         0.0
duration_value    0.0
duration_unit     0.0
dtype: float64
```

[10]:

```
print("Duplicate rows:", df.duplicated().sum())  # Duplicates
```

```
Duplicate rows: 0
```

Step 3: Structuring

[25]:

```
# Make all column names lowercase and replace spaces with underscores
```

Step 3: Structuring

# Standardize column names: lowercase + replace spaces with underscores

df.columns = df.columns.str.lower().str.replace(' ', '_')

# Convert 'date_added' to datetime

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

# Separate duration into numeric value and unit

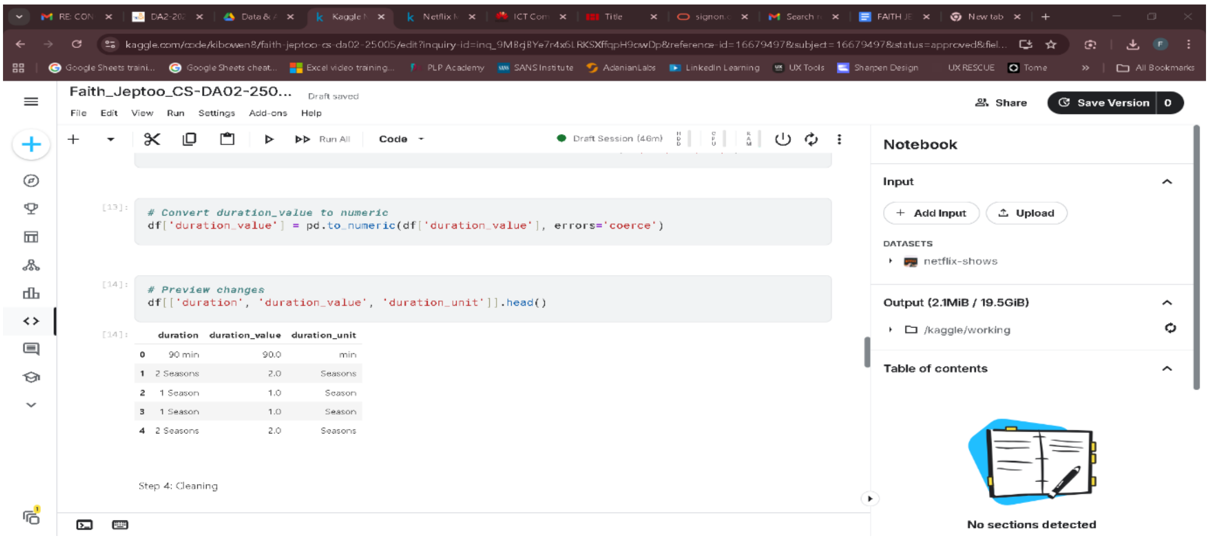
df[['duration_value', 'duration_unit']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')

Convert duration_value to numeric

df['duration_value'] = pd.to_numeric(df['duration_value'], errors='coerce')

Preview changes

df[['duration', 'duration_value', 'duration_unit']].head()


kaggle.com/code/kibowen8/faith-jeptoo-cs-da02-25005/edit?inquiry-id=inq_9MBqBYe7r4x6LRKSXffqpH9owDp&reference-id=16679497&subject=16679497&status=approved&fiel...
Faith_Jeptoo_CS-DA02-250...
Draft saved
File Edit View Run Settings Add-ons Help
Run All
Code
Draft Session (46m)
Share
Save Version
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.1MiB / 19.5GiB)
/kaggle/working
Table of contents
No sections detected

[13]:

```
# Convert duration_value to numeric
df['duration_value'] = pd.to_numeric(df['duration_value'], errors='coerce')
```

[14]:

```
# Preview changes
df[['duration', 'duration_value', 'duration_unit']].head()
```

[14]:

	duration	duration_value	duration_unit
0	90 min	90.0	min
1	2 Seasons	2.0	Seasons
2	1 Season	1.0	Season
3	1 Season	1.0	Season
4	2 Seasons	2.0	Seasons

Step 4: Cleaning

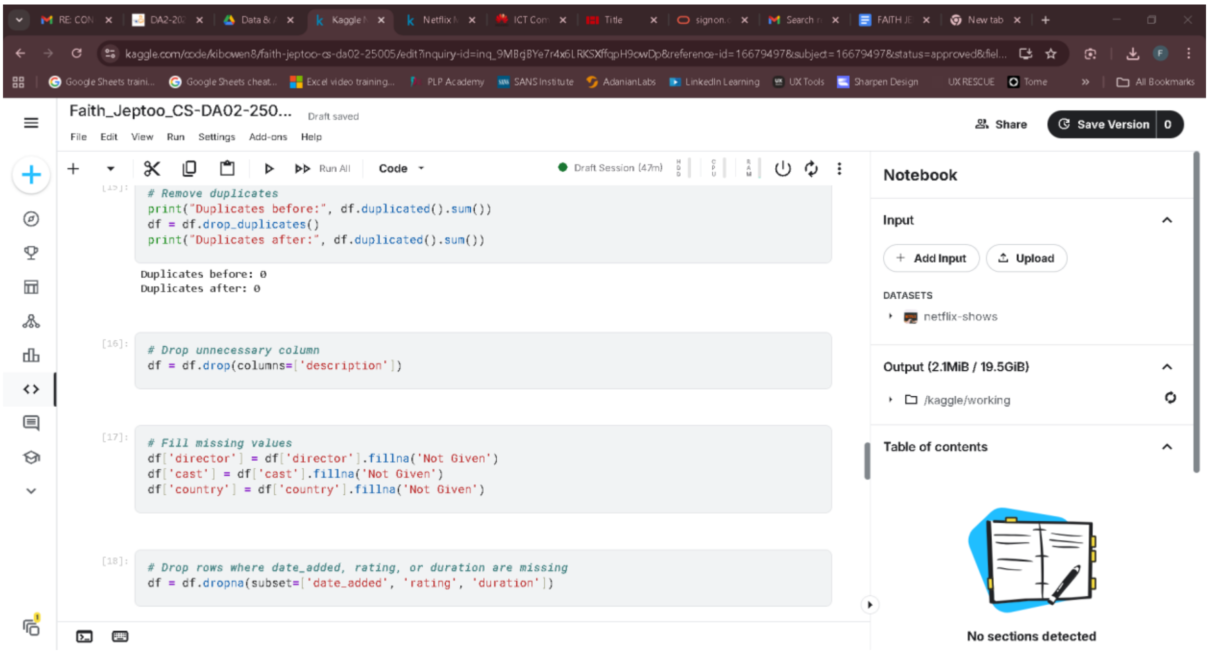
Step 4: Cleaning

```
# Remove duplicates
print("Duplicates before:", df.duplicated().sum())
df = df.drop_duplicates()
print("Duplicates after:", df.duplicated().sum())

# Drop unnecessary 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 where critical fields are missing
df = df.dropna(subset=['date_added', 'rating', 'duration'])
```



kaggle.com/code/kibowen8/faith-jeptoo-cs-da02-25005/edit?inquiry-id=inq_9MBqBYe7r4x6LRKSXffqpH9owDp&reference-id=16679497&subject=16679497&status=approved&fiel...
Faith_Jeptoo_CS-DA02-250...
Draft saved
File
Edit
View
Run
Settings
Add-ons
Help
Run All
Code
Draft Session (47m)
Share
Save Version
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.1MiB / 19.5GiB)
/kaggle/working
Table of contents
No sections detected

```
# Remove duplicates
print("Duplicates before:", df.duplicated().sum())
df = df.drop_duplicates()
print("Duplicates after:", df.duplicated().sum())
```

```
Duplicates before: 0
Duplicates after: 0
```

[16]:
```
# Drop unnecessary column
df = df.drop(columns=['description'])
```

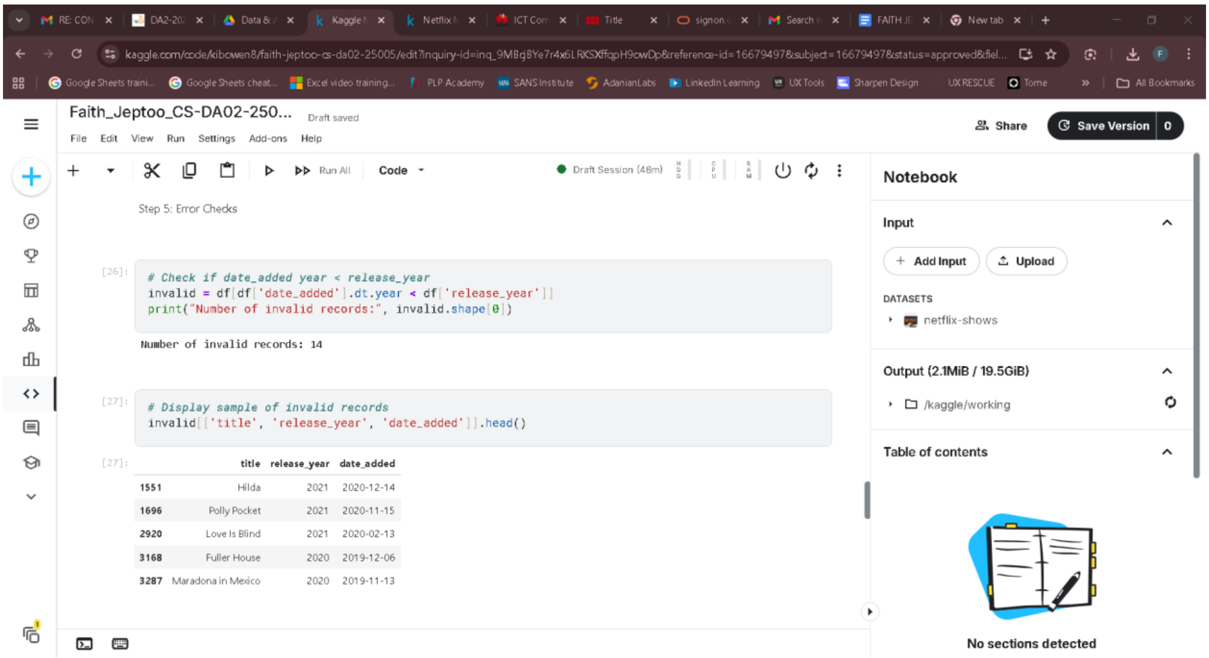
[17]:
```
# Fill missing values
df['director'] = df['director'].fillna('Not Given')
df['cast'] = df['cast'].fillna('Not Given')
df['country'] = df['country'].fillna('Not Given')
```

[18]:
```
# Drop rows where date_added, rating, or duration are missing
df = df.dropna(subset=['date_added', 'rating', 'duration'])
```

Step 5: Error Checks

```
# Check if date_added year < release_year
invalid = df[df['date_added'].dt.year < df['release_year']]
print("Number of invalid records:", invalid.shape[0])

# Display sample of invalid records
invalid[['title', 'release_year', 'date_added']].head()
```



Step 5: Error Checks

```
[26]:
# Check if date_added year < release_year
invalid = df[df['date_added'].dt.year < df['release_year']]
print("Number of invalid records:", invalid.shape[0])
```

```
Number of invalid records: 14
```

```
[27]:
# Display sample of invalid records
invalid[['title', 'release_year', 'date_added']].head()
```

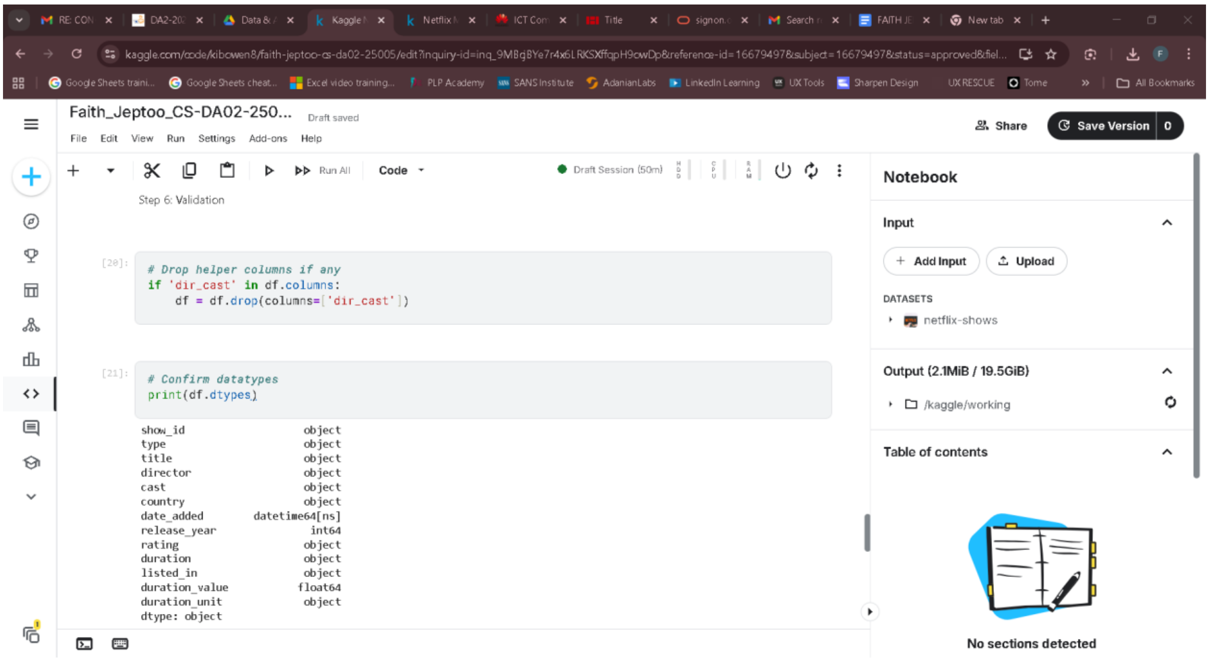
[27]:

	title	release_year	date_added
1551	Hilda	2021	2020-12-14
1696	Polly Pocket	2021	2020-11-15
2920	Love Is Blind	2021	2020-02-13
3168	Fuller House	2020	2019-12-06
3287	Maradona in Mexico	2020	2019-11-13

Step 6: Validation

```
# Drop helper columns if any
if 'dir_cast' in df.columns:
    df = df.drop(columns=['dir_cast'])
```

```
# Confirm datatypes
print(df.dtypes)
```



Step 6: Validation

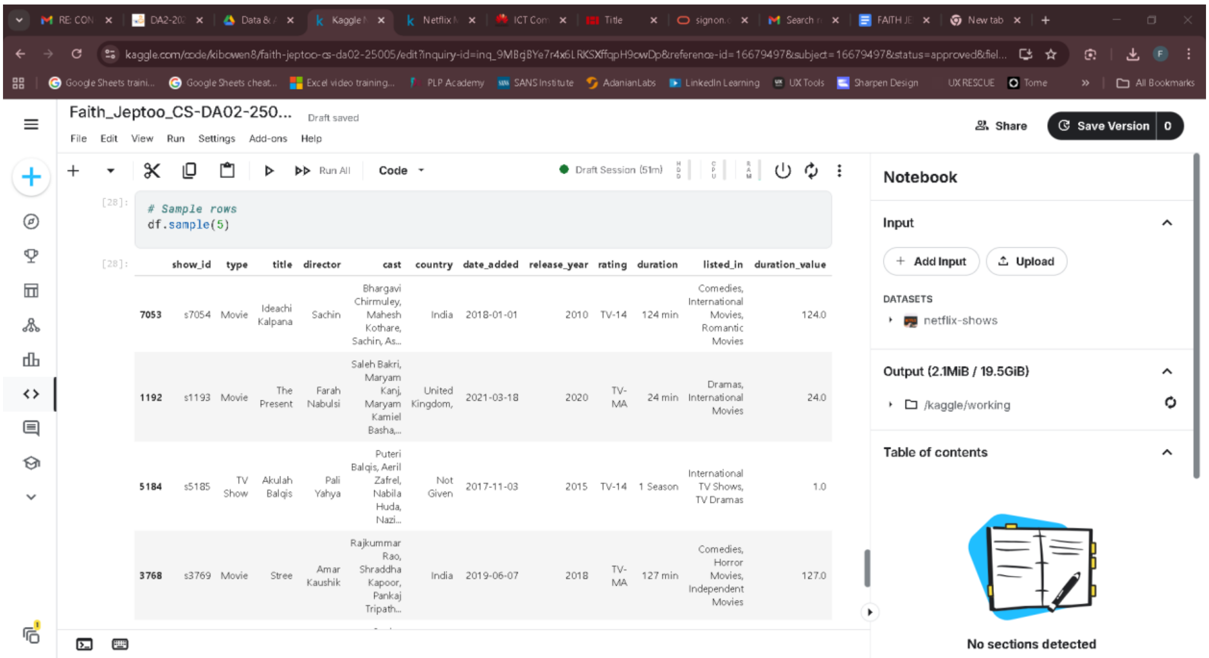
```
[20]:
# Drop helper columns if any
if 'dir_cast' in df.columns:
    df = df.drop(columns=['dir_cast'])
```

```
[21]:
# Confirm datatypes
print(df.dtypes)
```

```
show_id                  object
type                     object
title                    object
director                 object
cast                     object
country                  object
date_added       datetime64[ns]
release_year              int64
rating                   object
duration                 object
listed_in                object
duration_value          float64
duration_unit            object
dtype: object
```

```
# Check for missing values again
print("Missing values after cleaning:\n", df.isnull().sum())
```

# Sample few rows
df.sample(5)



Faith_Jeptoo_CS-DA02-250...

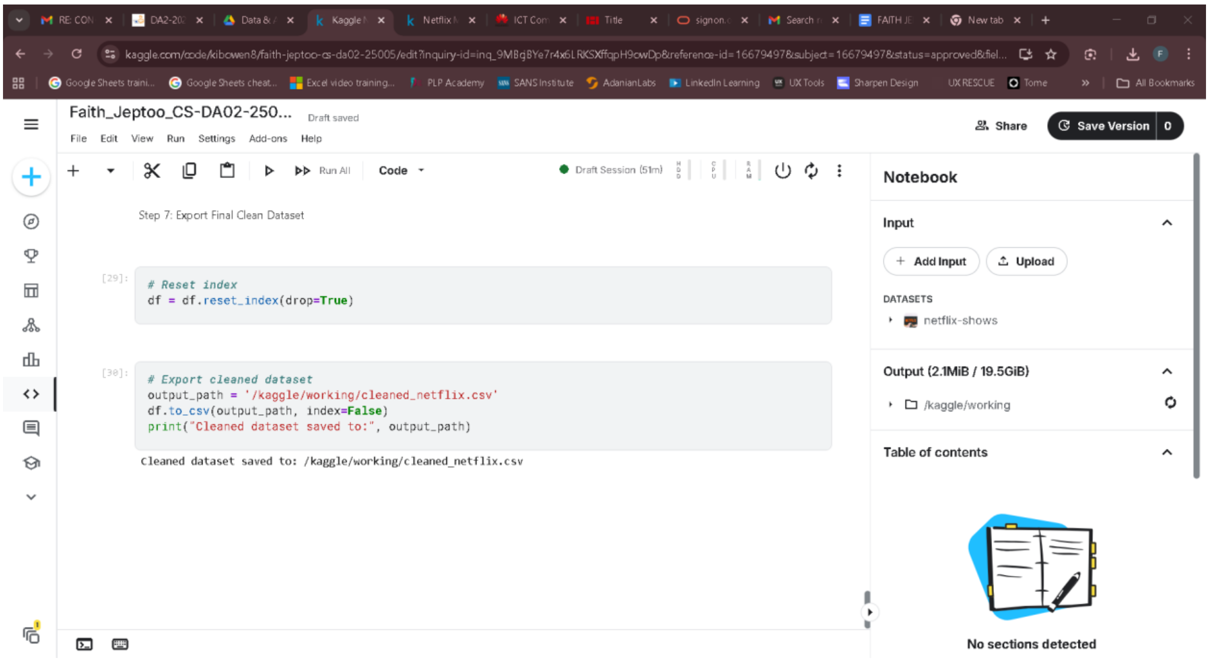
```
[28]:
# Sample rows
df.sample(5)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	duration_value
7053	s7054	Movie	Ideachi Kalpana	Sachin	Bhargavi Chirmuley, Mahesh Kothare, Sachin, As...	India	2018-01-01	2010	TV-14	124 min	Comedies, International Movies, Romantic Movies	124.0
1192	s1193	Movie	The Present	Farah Nabulsi	Saleh Bakri, Maryam Kanj, Maryam Kamiel Basha,...	United Kingdom,	2021-03-18	2020	TV-MA	24 min	Dramas, International Movies	24.0
5184	s5185	TV Show	Akulah Balqis	Pali Yahya	Puteri Balqis, Aeril Zafrel, Nabila Huda, Nazi...	Not Given	2017-11-03	2015	TV-14	1 Season	International TV Shows, TV Dramas	1.0
3768	s3769	Movie	Stree	Amar Kaushik	Rajkummar Rao, Shraddha Kapoor, Pankaj Tripath...	India	2019-06-07	2018	TV-MA	127 min	Comedies, Horror Movies, Independent Movies	127.0

Step 7: Export Final Dataset

Reset index
df = df.reset_index(drop=True)

Export cleaned dataset
output_path = '/kaggle/working/cleaned_netflix.csv'
df.to_csv(output_path, index=False)
print("Cleaned dataset saved to:", output_path)



Faith_Jeptoo_CS-DA02-250...

Step 7: Export Final Clean Dataset

```
[29]:
# Reset index
df = df.reset_index(drop=True)
```

```
[30]:
# Export cleaned dataset
output_path = '/kaggle/working/cleaned_netflix.csv'
df.to_csv(output_path, index=False)
print("Cleaned dataset saved to:", output_path)
```

```
Cleaned dataset saved to: /kaggle/working/cleaned_netflix.csv
```

Conclusion

This assignment provided hands-on experience in data wrangling. I learned how to explore, clean, structure, and validate real-world datasets. The final Netflix dataset is now ready for analysis and visualization.

Link to Notebook

https://www.kaggle.com/code/kibowen8/faith-jeptoo-cs-da02-25005-week-2