Data and Artificial Intelligence

Cyber Shujaa Program

Week 2 Assignment
Netflix Data Wrangling

Student Name: Salome Kungu

Student ID: CS-DA03-26054

Contents

Introduction

The objective of this assignment was to apply data wrangling techniques to a real-world dataset sourced from Kaggle, specifically the Netflix Movies and TV Shows dataset. The task focused on developing practical skills in data discovery, structuring, cleaning, transformation, validation, and publication using Python and the Pandas library within a Kaggle Notebook environment.
Through this assignment, I gained hands-on experience in identifying data quality issues such as missing values, duplicates, and inconsistent formatting, and resolving them using systematic and reproducible methods. The final outcome is a cleaned and structured dataset suitable for further analysis and visualization.

The key objectives of the assignment were to:

1. Load and explore the Netflix dataset.
2. Perform data discovery and assess data quality.
3. Clean and structure the dataset using data wrangling techniques.
4. Validate the dataset for consistency and accuracy.
5. Export and publish the final cleaned dataset and notebook on Kaggle.

Tasks Completed

The assignment began with importing the Netflix dataset from a CSV file into a Pandas DataFrame. An initial data discovery phase was conducted to understand the dataset's structure, including the number of records, column data types, missing values, and duplicate entries.

The dataset was then structured by converting date fields into appropriate datetime formats and extracting numeric and categorical components from complex fields such as the duration column. Additional formatting improvements were applied to ensure consistency across columns.

Data cleaning was performed by removing duplicate records, dropping unnecessary columns, and handling missing values. Missing director values were inferred using recurring director–cast relationships, while country values were populated using director–country associations where possible. Remaining missing values were clearly labeled as "Not Given" to maintain dataset completeness without introducing false assumptions. Records with critical missing fields such as date_added, rating, and duration were removed.

Logical validation checks were carried out to identify inconsistencies, such as cases where the date added to Netflix occurred before the release year. The dataset was further validated by confirming correct data types, ensuring completeness, and sampling records for visual inspection. Temporary columns created during wrangling were removed, and the index was reset.

Finally, the cleaned dataset was exported as a CSV file and the complete Kaggle Notebook was published with public access enabled.

Ensure you allow 'Anyone with the link' to be a Viewer

Link to Code: https://www.kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054

Conclusion

This assignment provided a comprehensive introduction to practical data wrangling using Python and Pandas. By working with a real-world dataset, I developed a deeper understanding of data quality challenges and the importance of systematic cleaning and validation processes. The skills acquired during this exercise—particularly in handling missing data, enforcing logical consistency, and preparing datasets for analysis—form a strong foundation for future work in data analytics and artificial intelligence. The completed notebook and dataset also contribute to my growing professional portfolio.

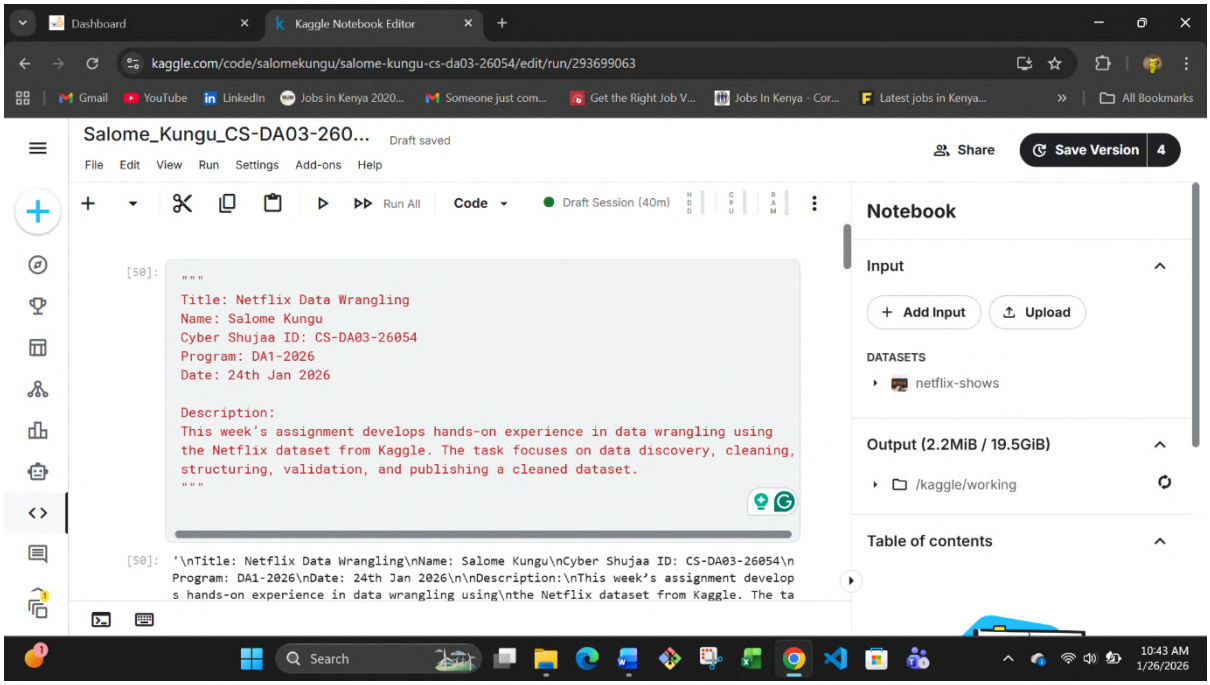
Screenshots of the project

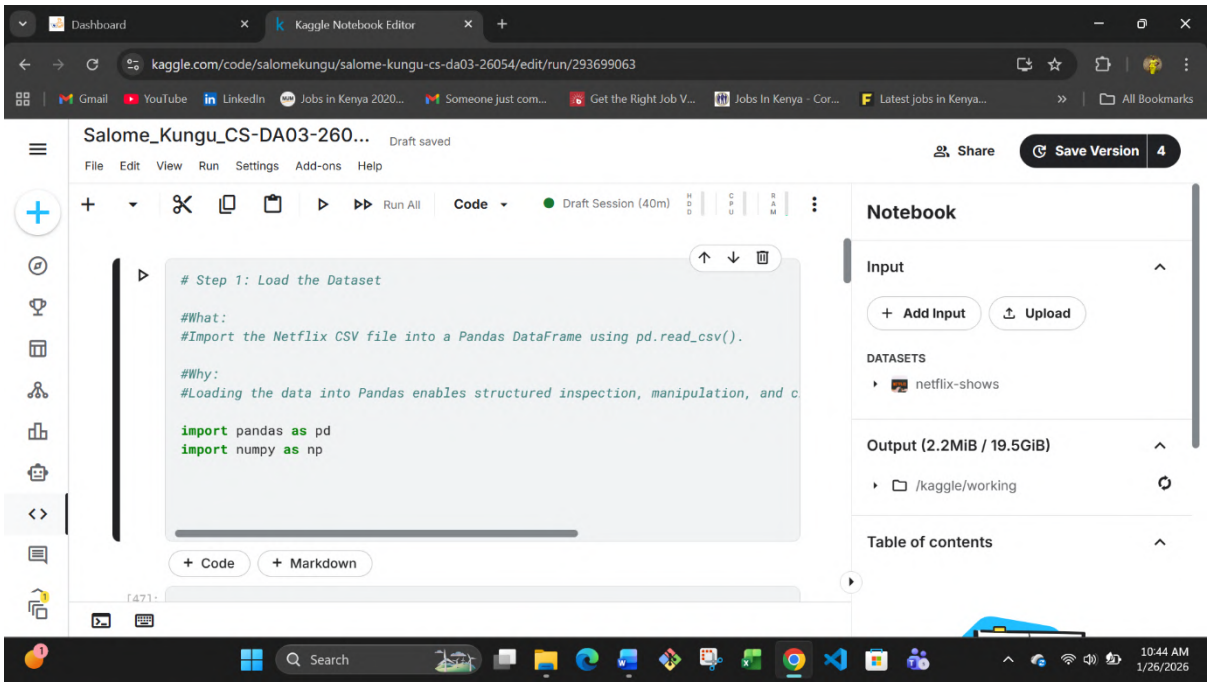
The following screenshots are taken from my Kaggle Notebook and document the data wrangling process applied to the Netflix movies dataset. They show key steps such as data cleaning, handling missing values, filtering and transforming variables, and verifying the dataset structure. The accompanying explanations demonstrate my active engagement with the lab tasks and confirm the successful completion of the assignment.

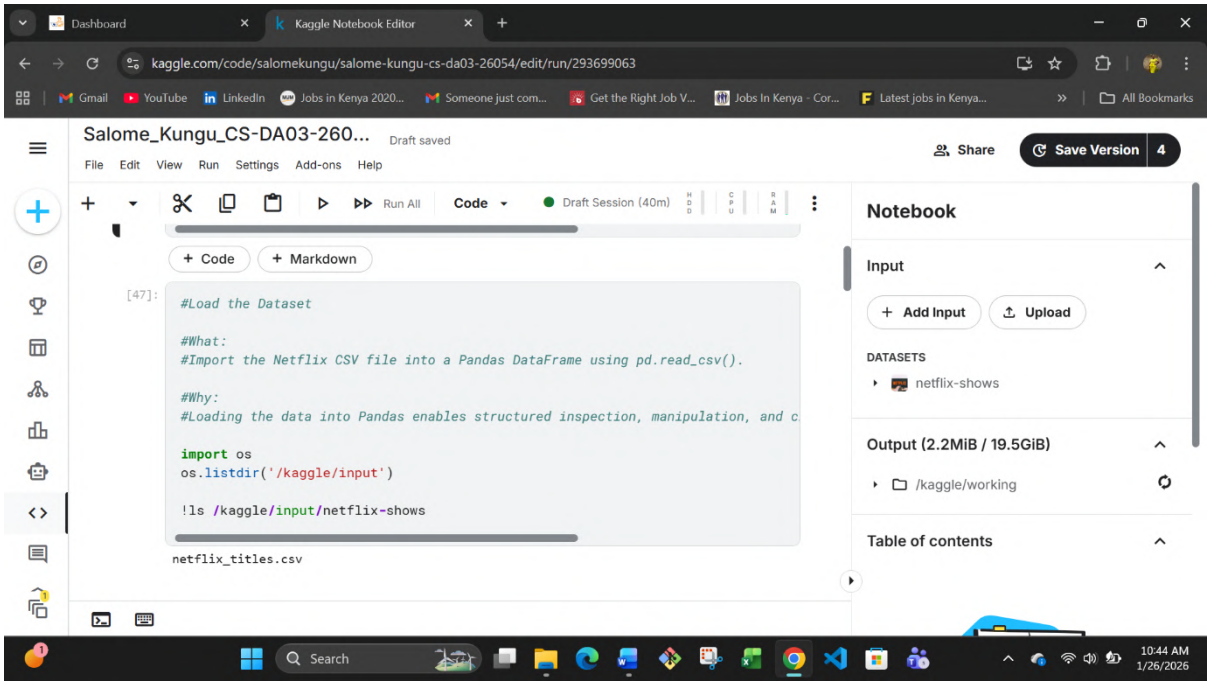
Step 1: Load the Dataset

What:
Import the Netflix CSV file into a Pandas DataFrame using pd.read_csv().

Why:
Loading the data into Pandas enables structured inspection, manipulation, and cleaning using efficient data-handling tools.




Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (40m)
Step 1: Load the Dataset
#What:
#Import the Netflix CSV file into a Pandas DataFrame using pd.read_csv().
#Why:
#Loading the data into Pandas enables structured inspection, manipulation, and c
import pandas as pd
import numpy as np
+ Code
+ Markdown
Notebook
Input
+ Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:44 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (40m)
+ Code
+ Markdown
[47]:
#Load the Dataset
#What:
#Import the Netflix CSV file into a Pandas DataFrame using pd.read_csv().
#Why:
#Loading the data into Pandas enables structured inspection, manipulation, and c
import os
os.listdir('/kaggle/input')
!ls /kaggle/input/netflix-shows
netflix_titles.csv
Notebook
Input
+ Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:44 AM
1/26/2026

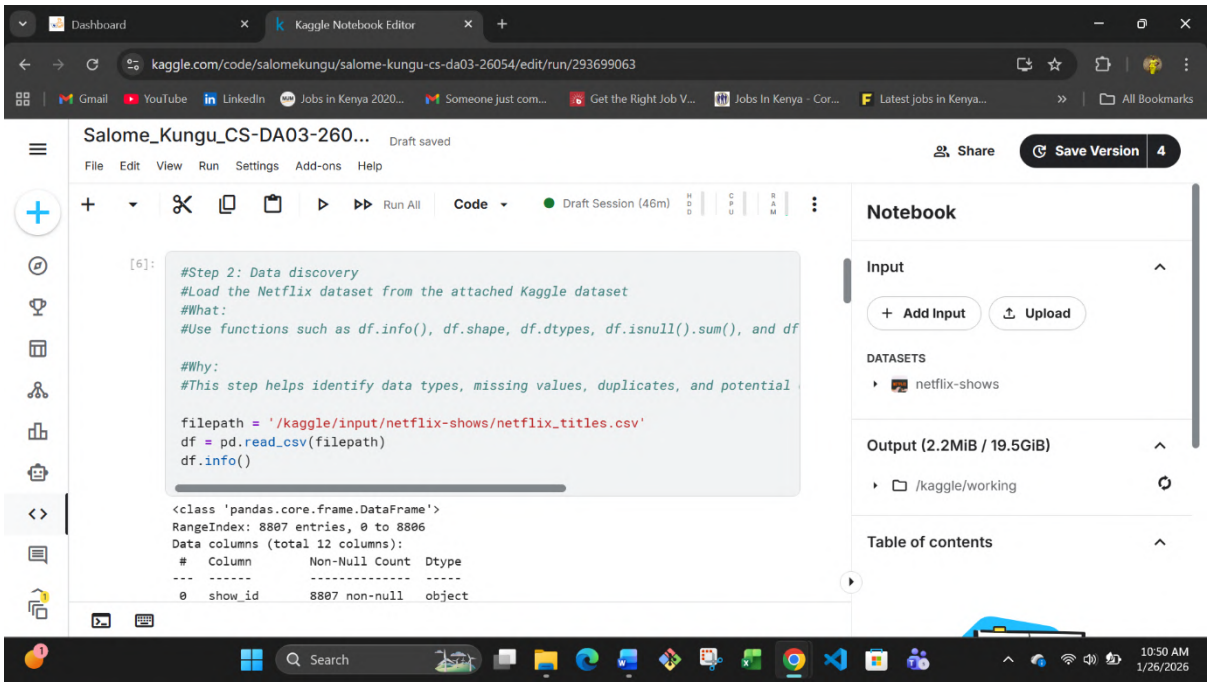
Step 2: Data Discovery

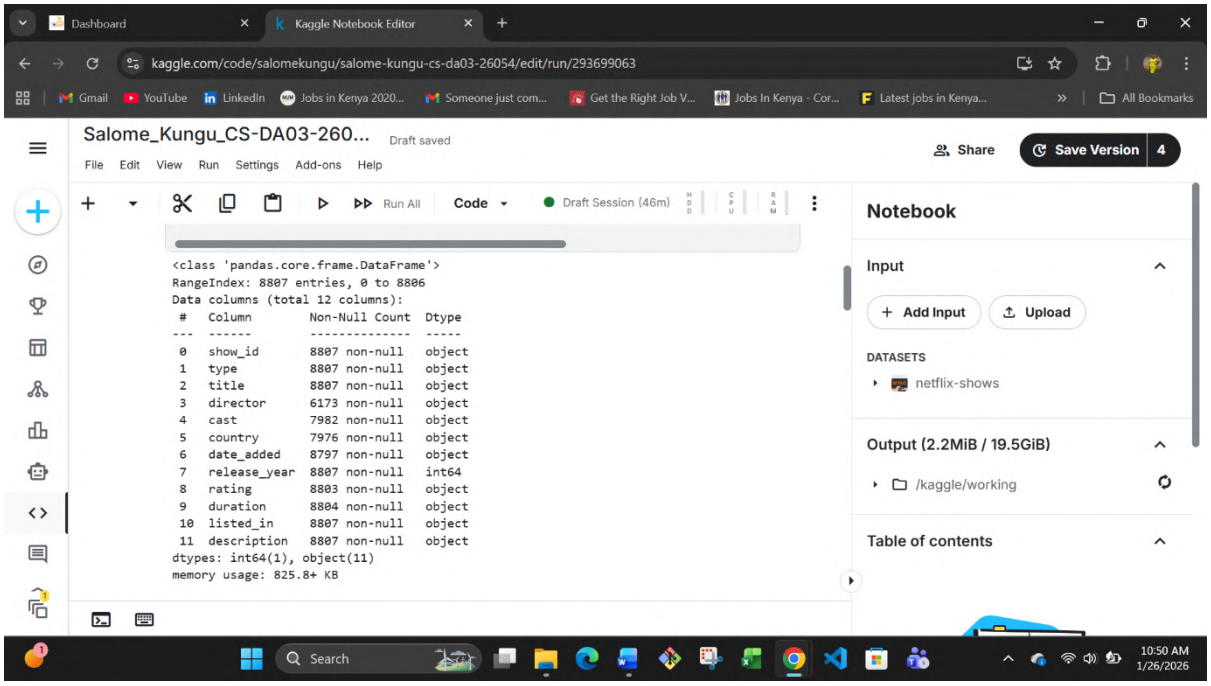
What:

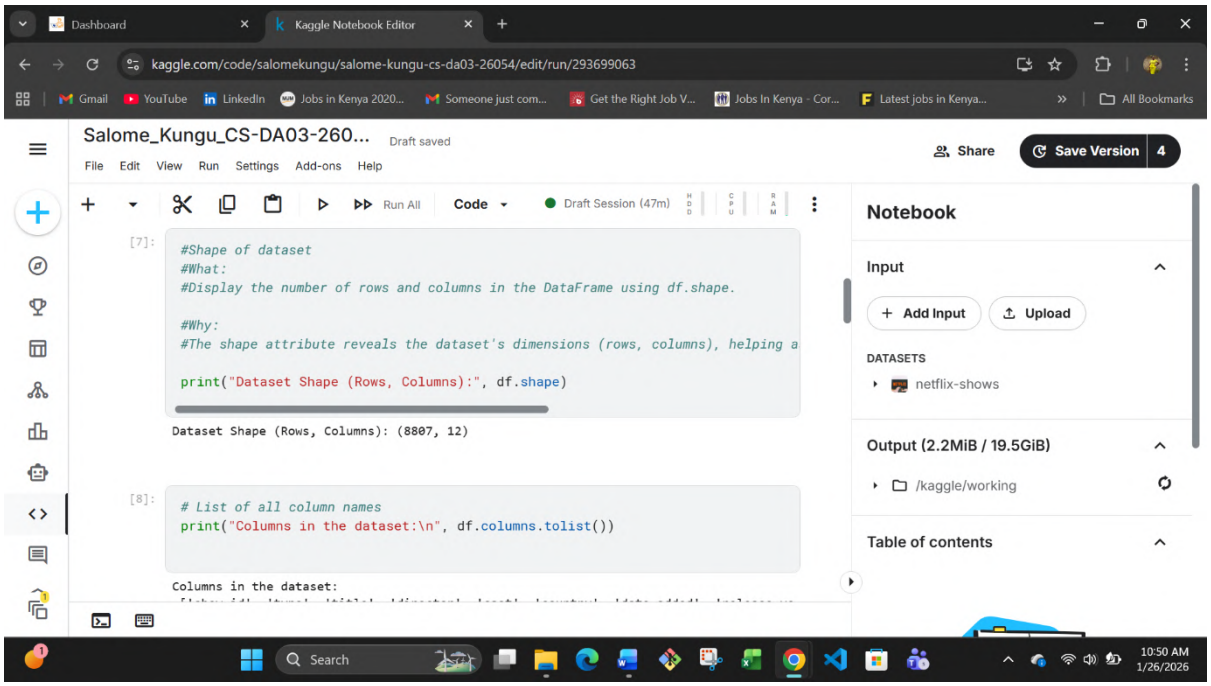
Use functions such as df.info(), df.shape, df.dtypes, df.isnull().sum(), and df.duplicated().sum() to explore the dataset.

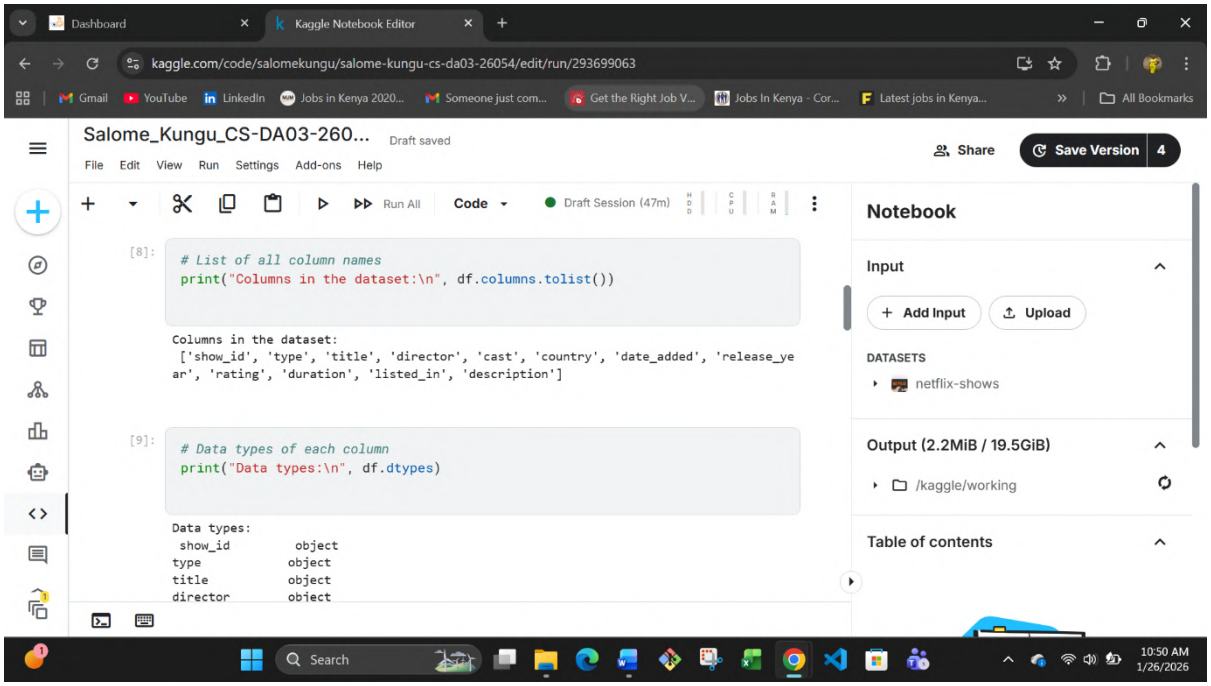
Why:

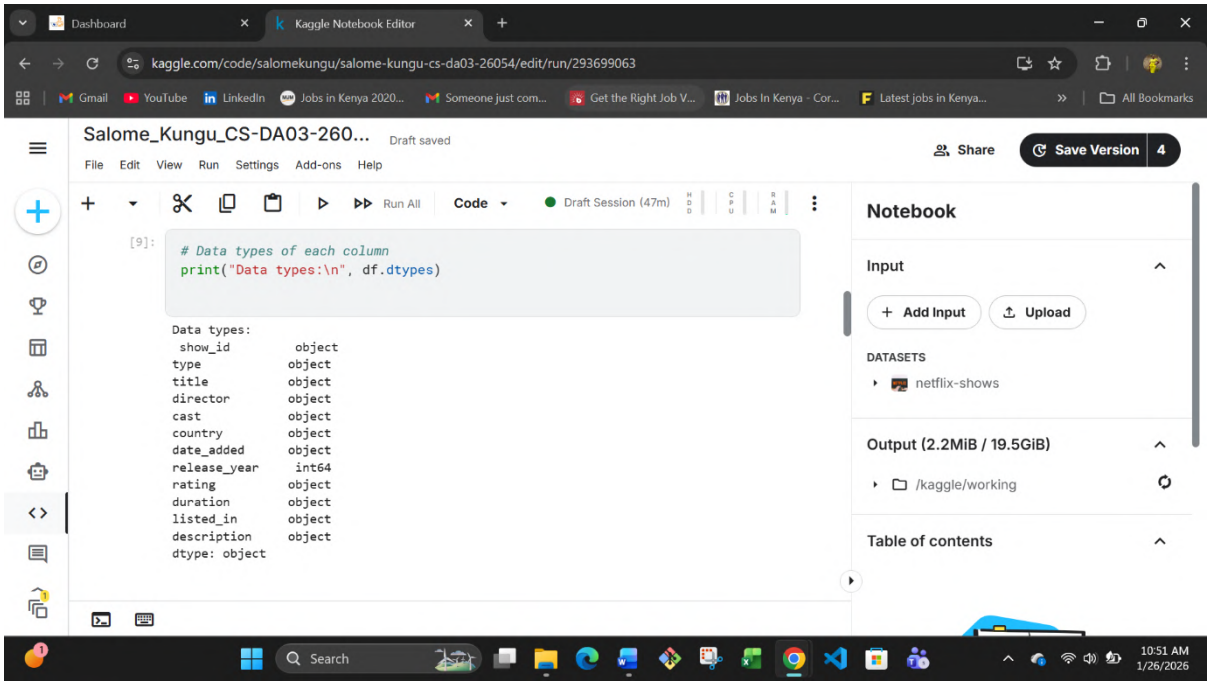
This step helps identify data types, missing values, duplicates, and potential quality issues before any cleaning is performed.

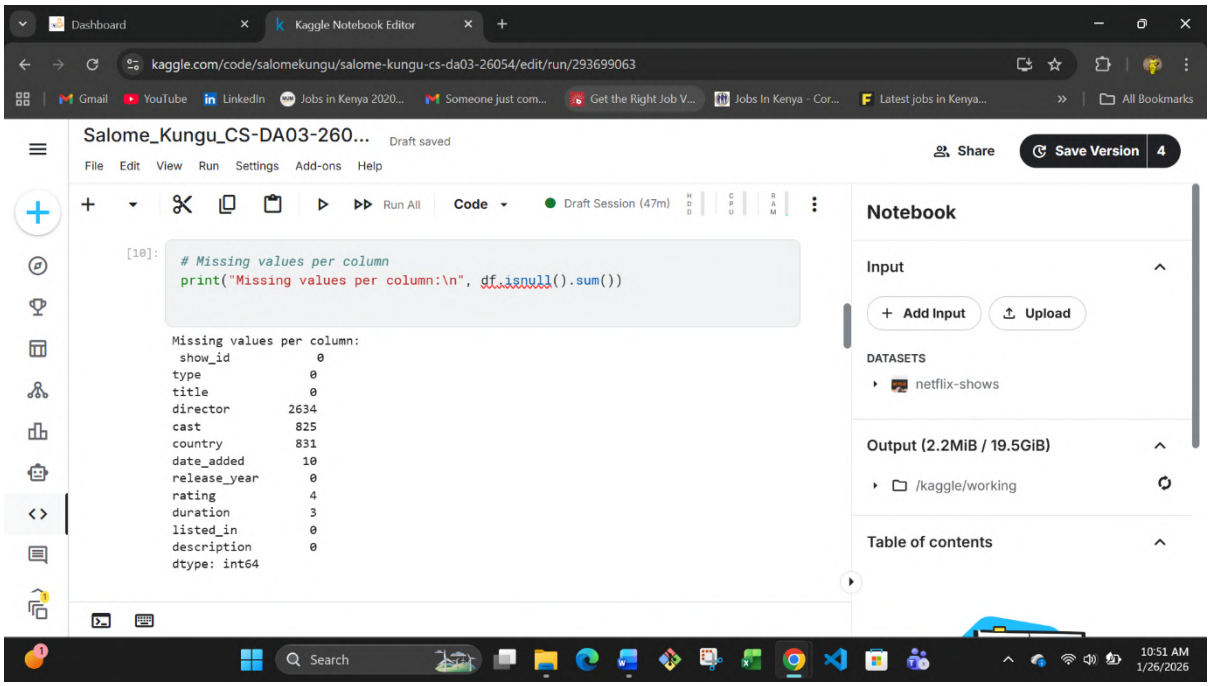


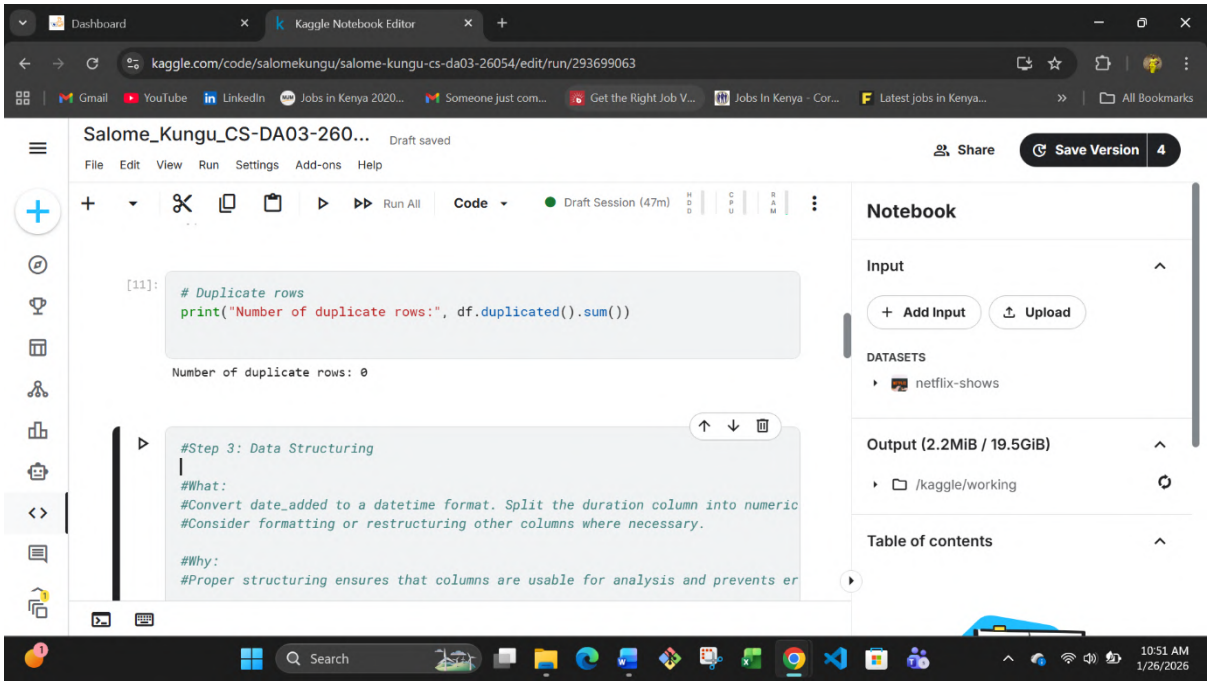



Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (47m)
[7]:
#Shape of dataset
#What:
#Display the number of rows and columns in the DataFrame using df.shape.
#Why:
#The shape attribute reveals the dataset's dimensions (rows, columns), helping a
print("Dataset Shape (Rows, Columns):", df.shape)
Dataset Shape (Rows, Columns): (8807, 12)
[8]:
# List of all column names
print("Columns in the dataset:\n", df.columns.tolist())
Columns in the dataset:
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:50 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (47m)
[8]:
# List of all column names
print("Columns in the dataset:\n", df.columns.tolist())
Columns in the dataset:
['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_ye
ar', 'rating', 'duration', 'listed_in', 'description']
[9]:
# Data types of each column
print("Data types:\n", df.dtypes)
Data types:
show_id object
type object
title object
director object
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:50 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (47m)
[9]:
# Data types of each column
print("Data types:\n", df.dtypes)
Data types:
show_id object
type object
title object
director object
cast object
country object
date_added object
release_year int64
rating object
duration object
listed_in object
description object
dtype: object
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:51 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (47m)
[10]:
# Missing values per column
print("Missing values per column:\n", df.isnull().sum())
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
Notebook
Input
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
10:51 AM
1/26/2026



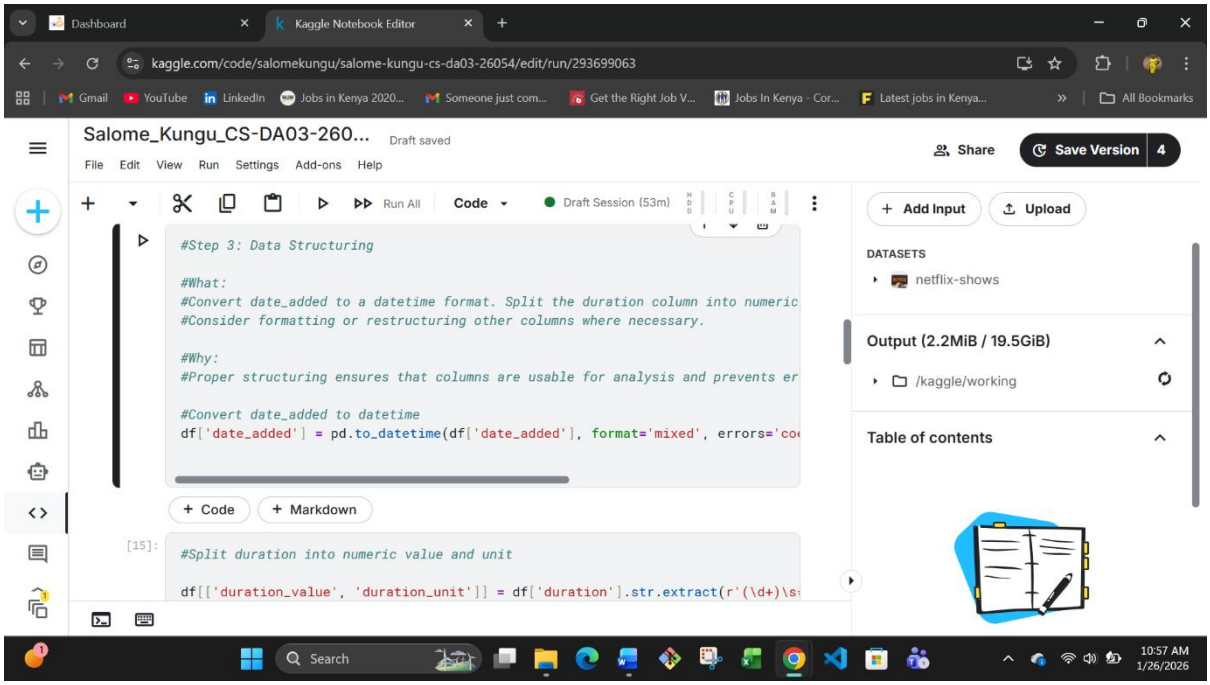
Step 3: Data Structuring

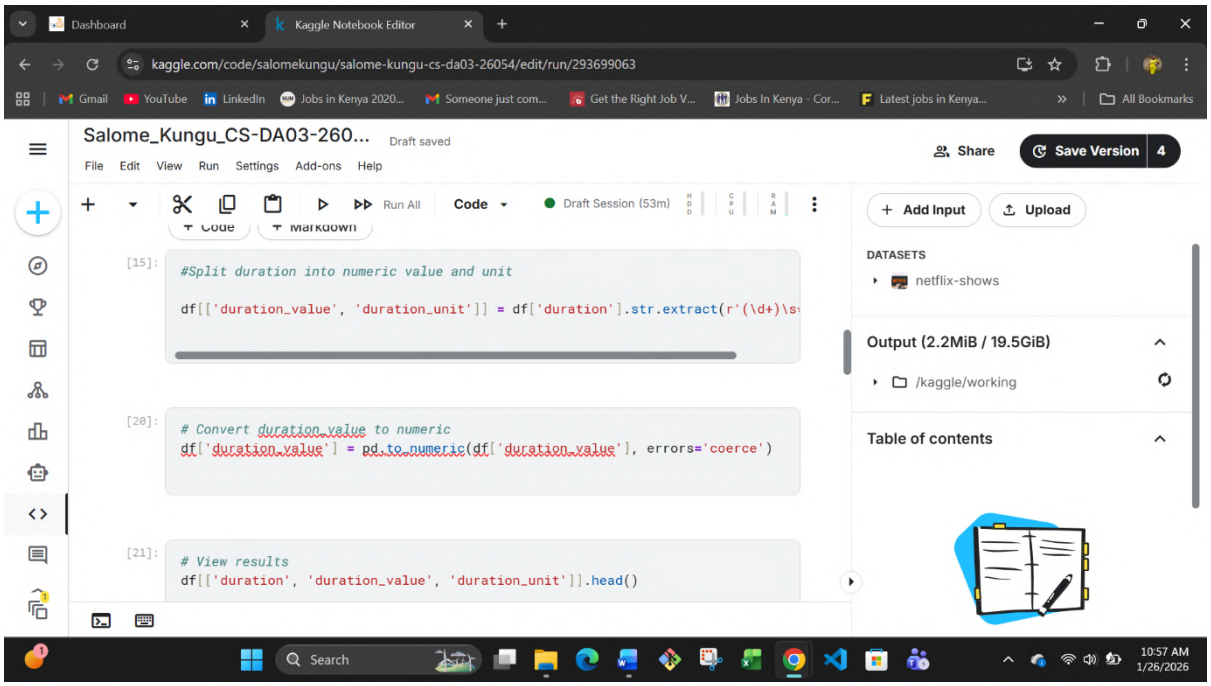
What:

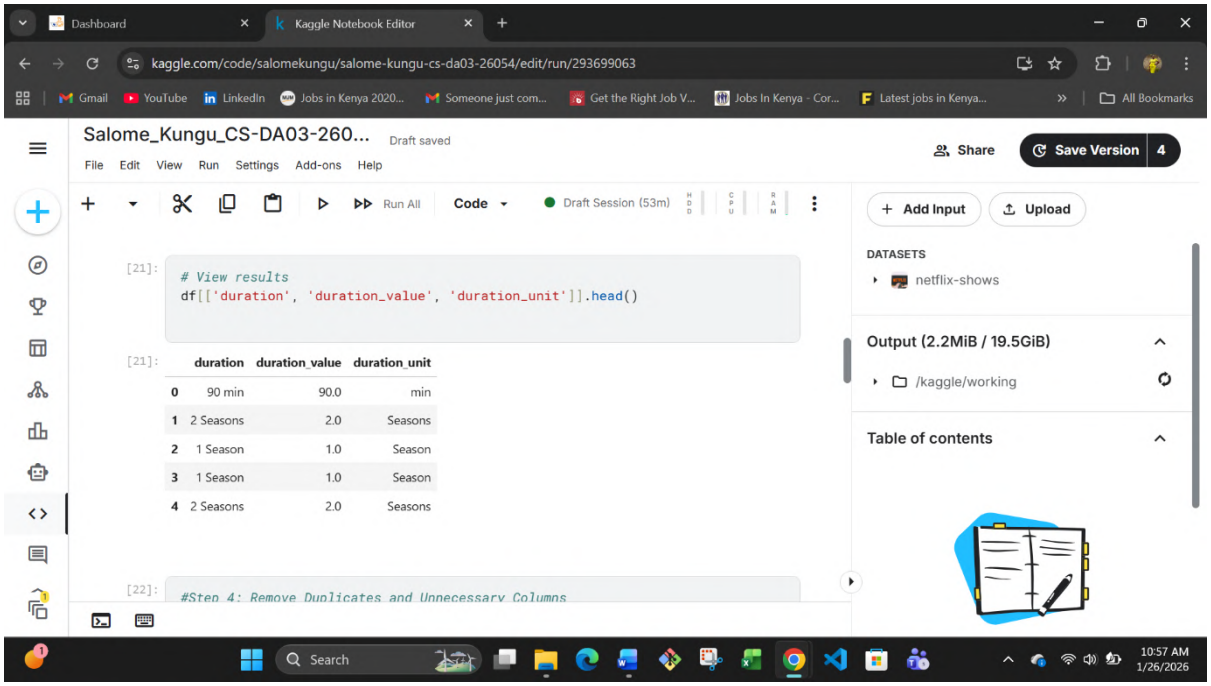
- Convert date_added to a datetime format.
- Split the duration column into numeric and unit components.
- Consider formatting or restructuring other columns where necessary.

Why:

Proper structuring ensures that columns are usable for analysis and prevents errors during filtering, grouping, or validation.





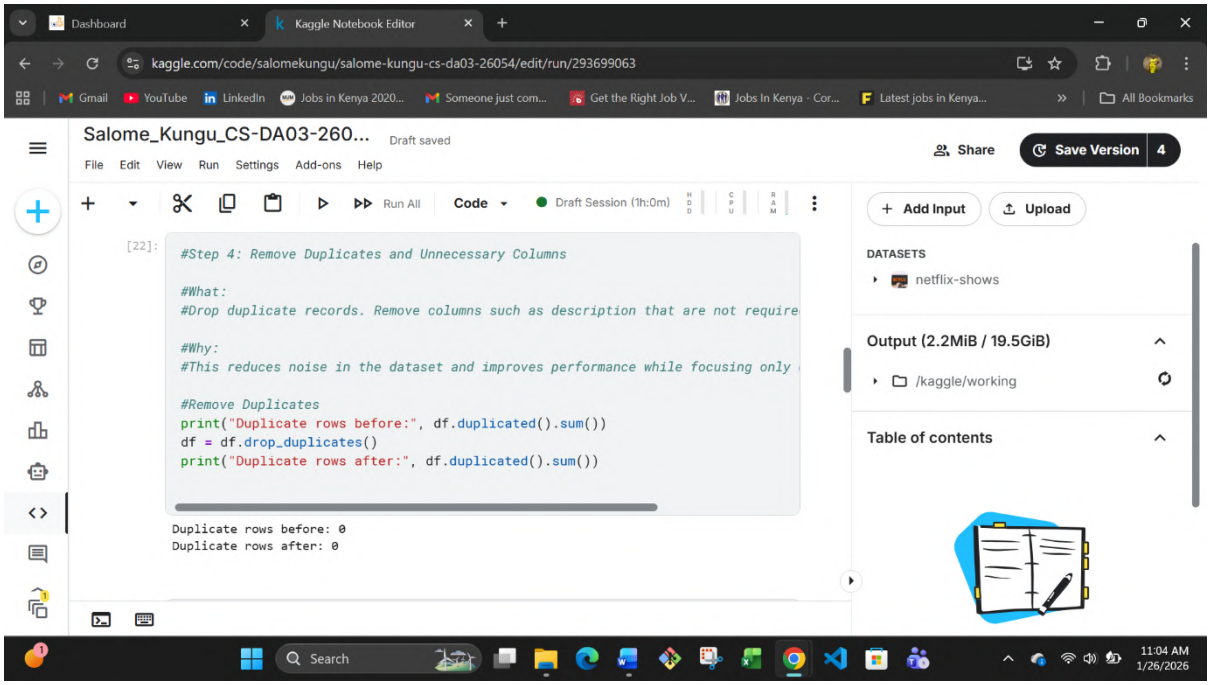


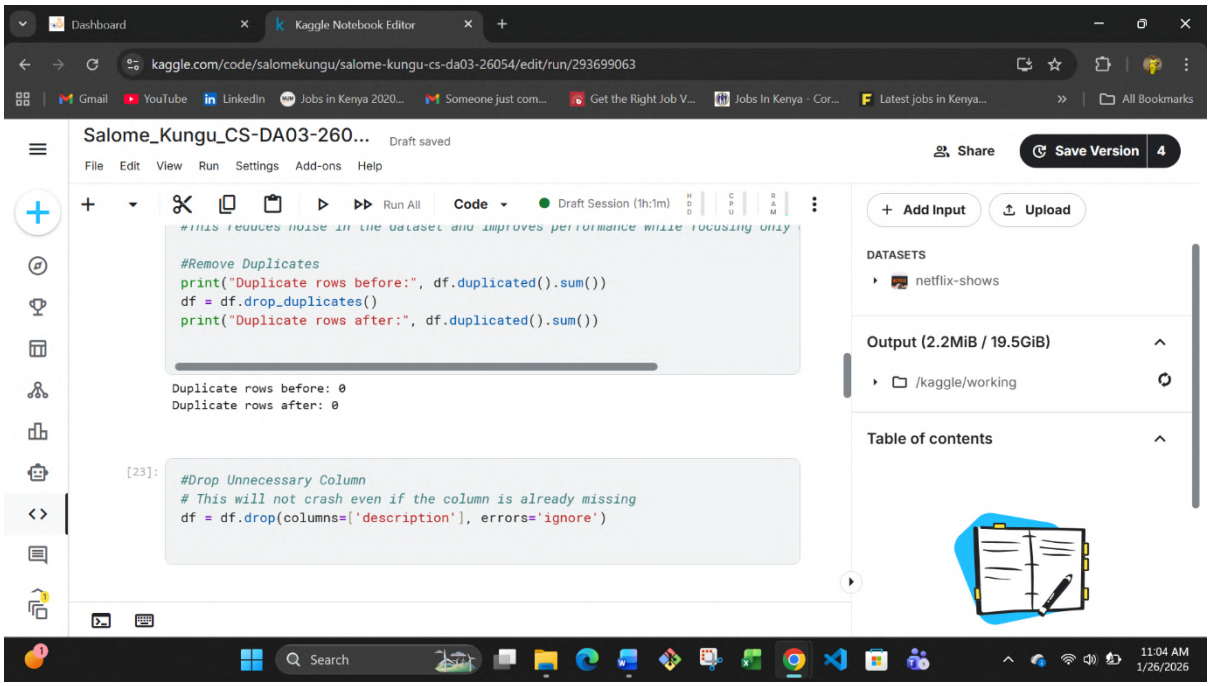
Step 4: Remove Duplicates and Unnecessary Columns

What:

- Drop duplicate records.
- Remove columns such as description that are not required for analysis.

Why:
This reduces noise in the dataset and improves performance while focusing only on relevant attributes.





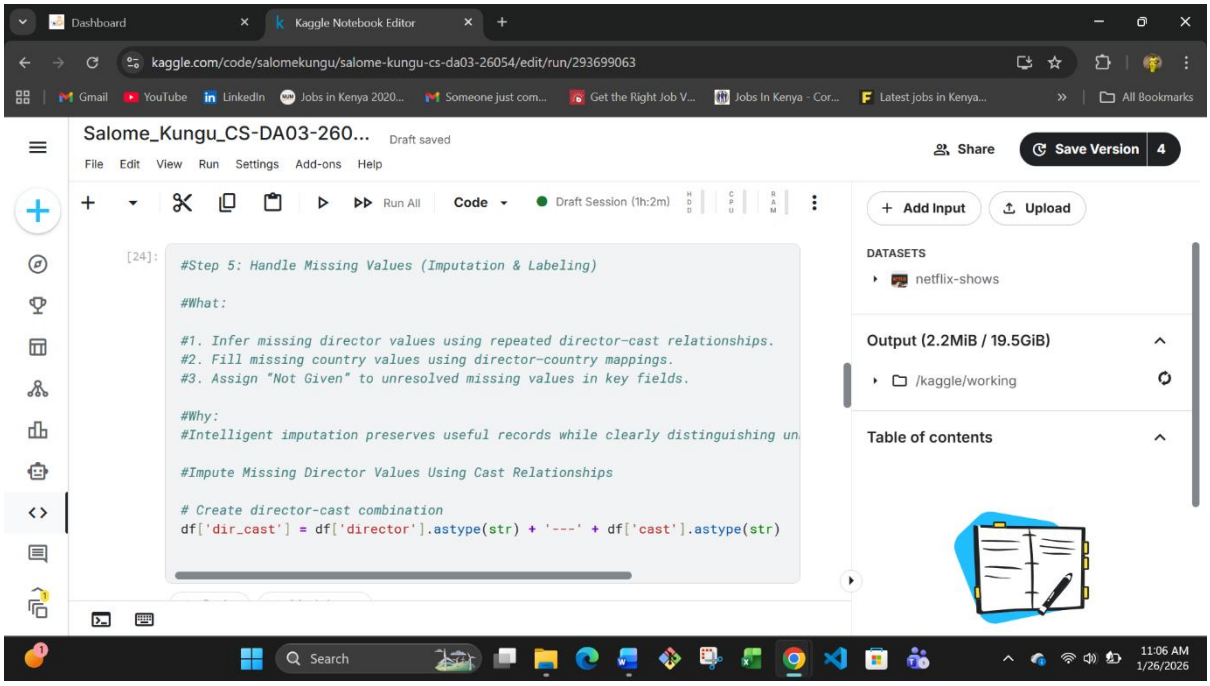
Step 5: Handle Missing Values (Imputation & Labeling)

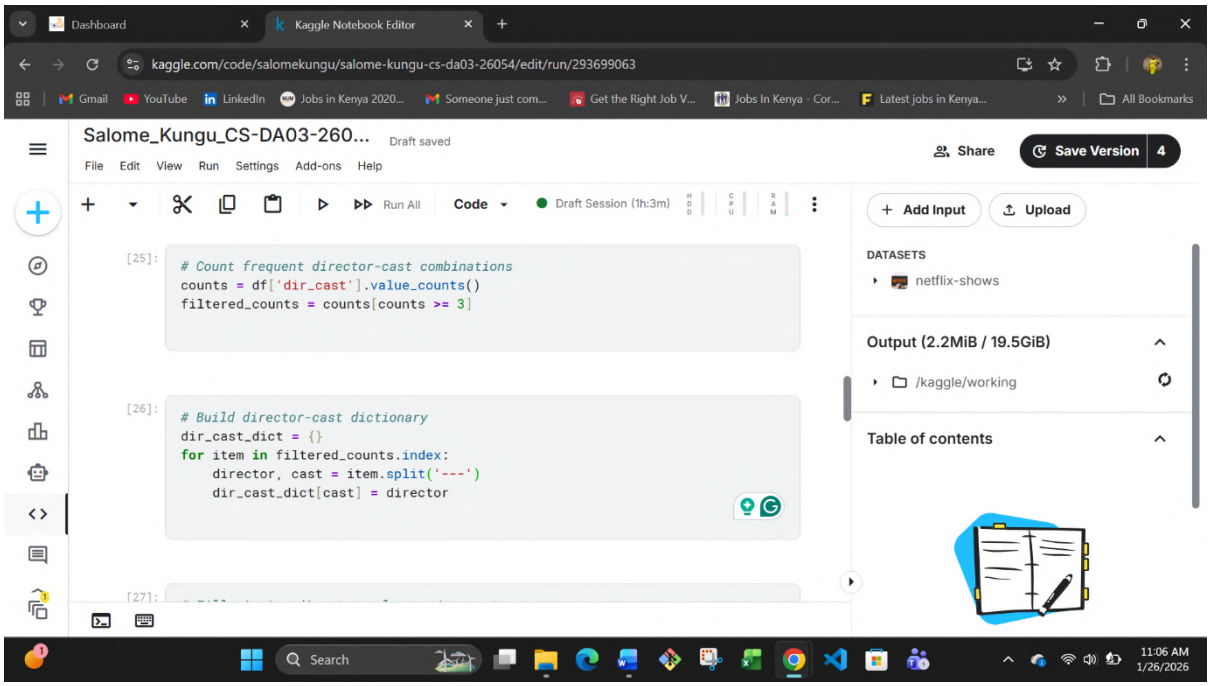
What:

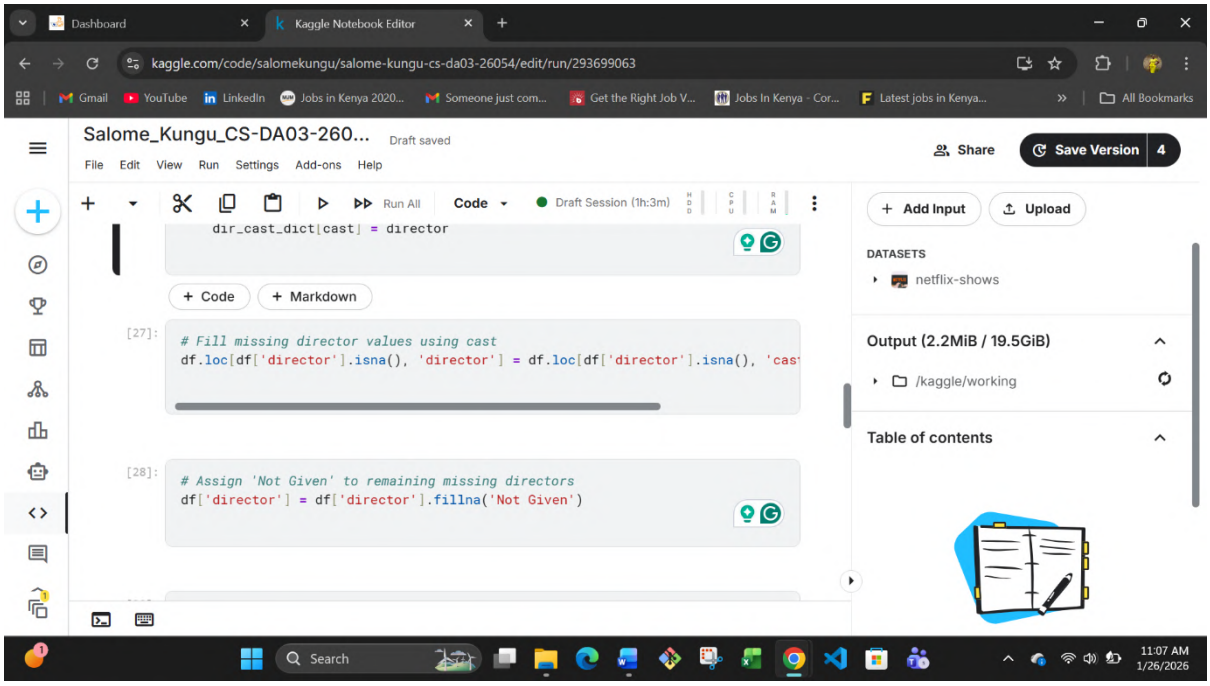
- Infer missing director values using repeated director–cast relationships.
- Fill missing country values using director–country mappings.
- Assign “Not Given” to unresolved missing values in key fields.

Why:

Intelligent imputation preserves useful records while clearly distinguishing unknown data from valid values.







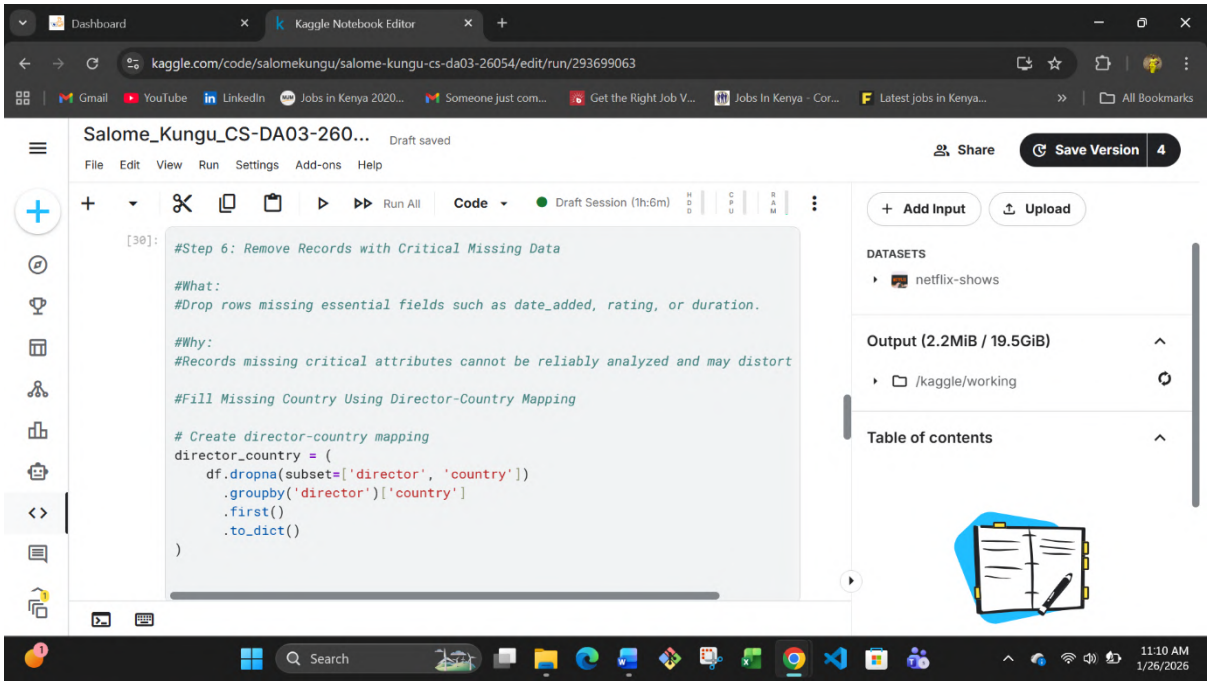
Step 6: Remove Records with Critical Missing Data

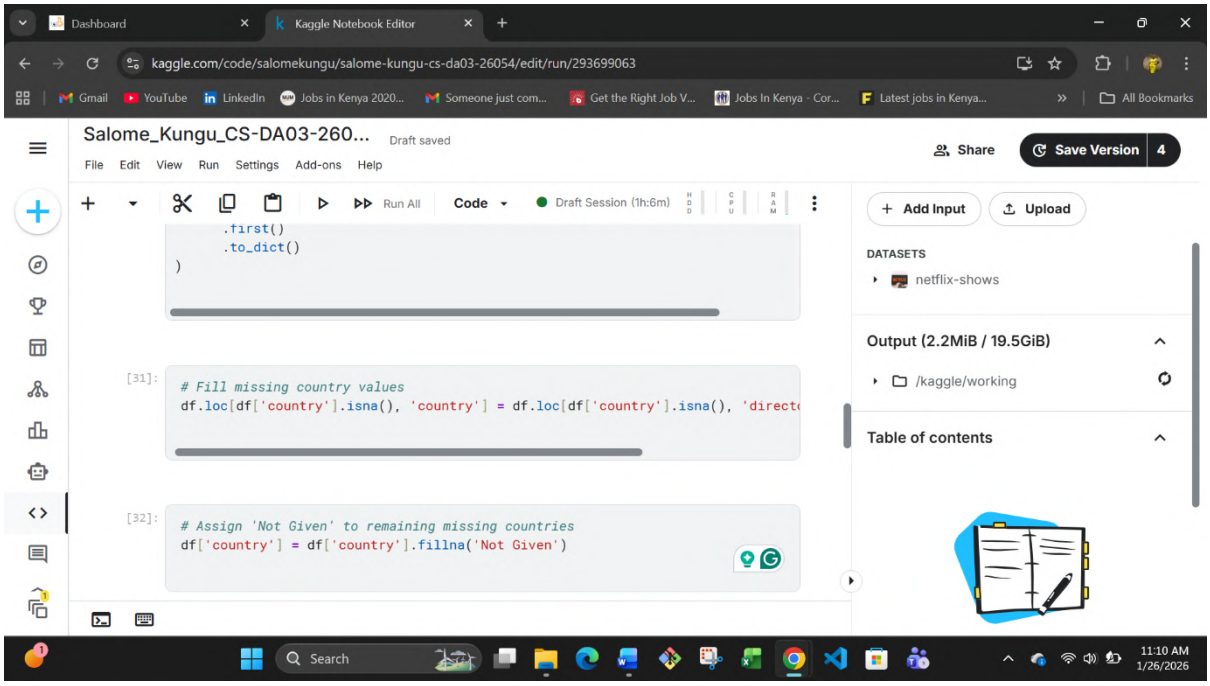
What:

Drop rows missing essential fields such as date_added, rating, or duration.

Why:

Records missing critical attributes cannot be reliably analyzed and may distort insights.





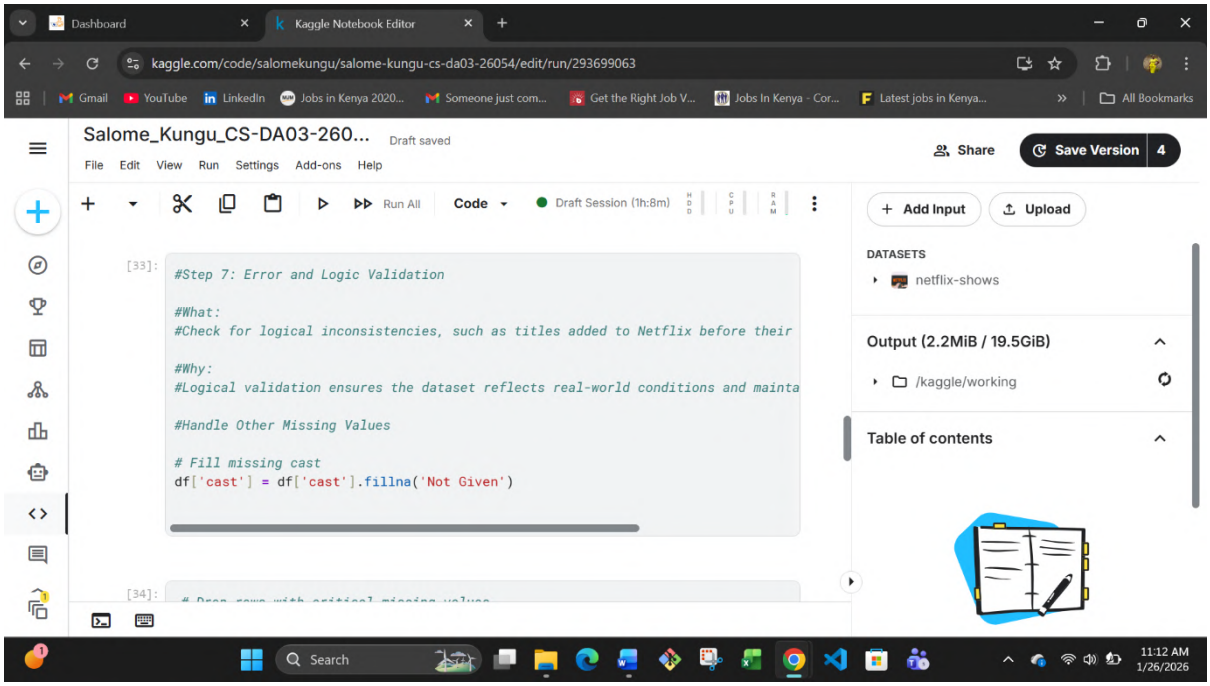
Step 7: Error and Logic Validation

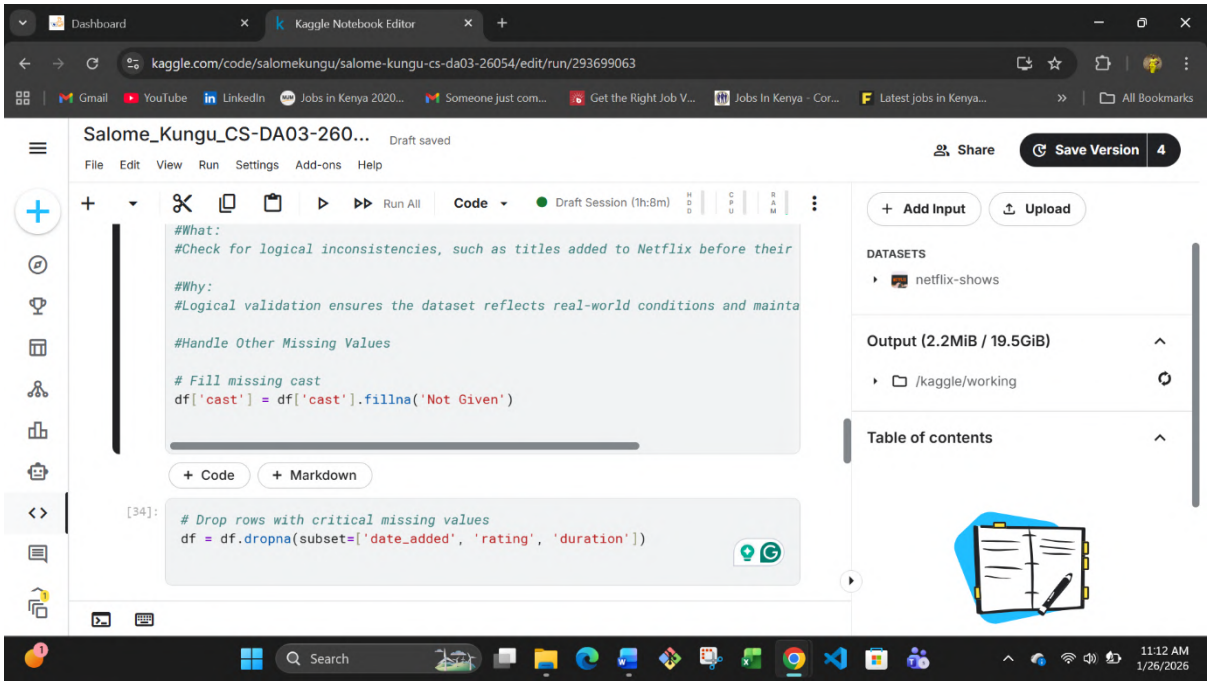
What:

Check for logical inconsistencies, such as titles added to Netflix before their release year.

Why:

Logical validation ensures the dataset reflects real-world conditions and maintains analytical integrity.





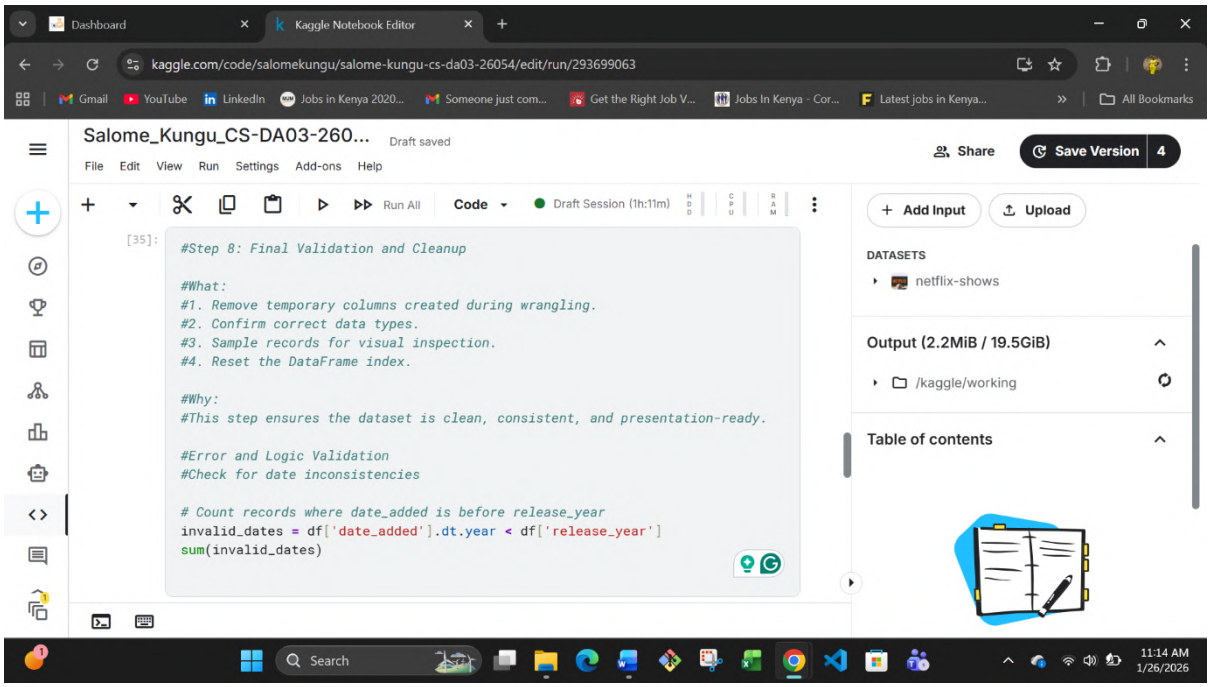
Step 8: Final Validation and Cleanup

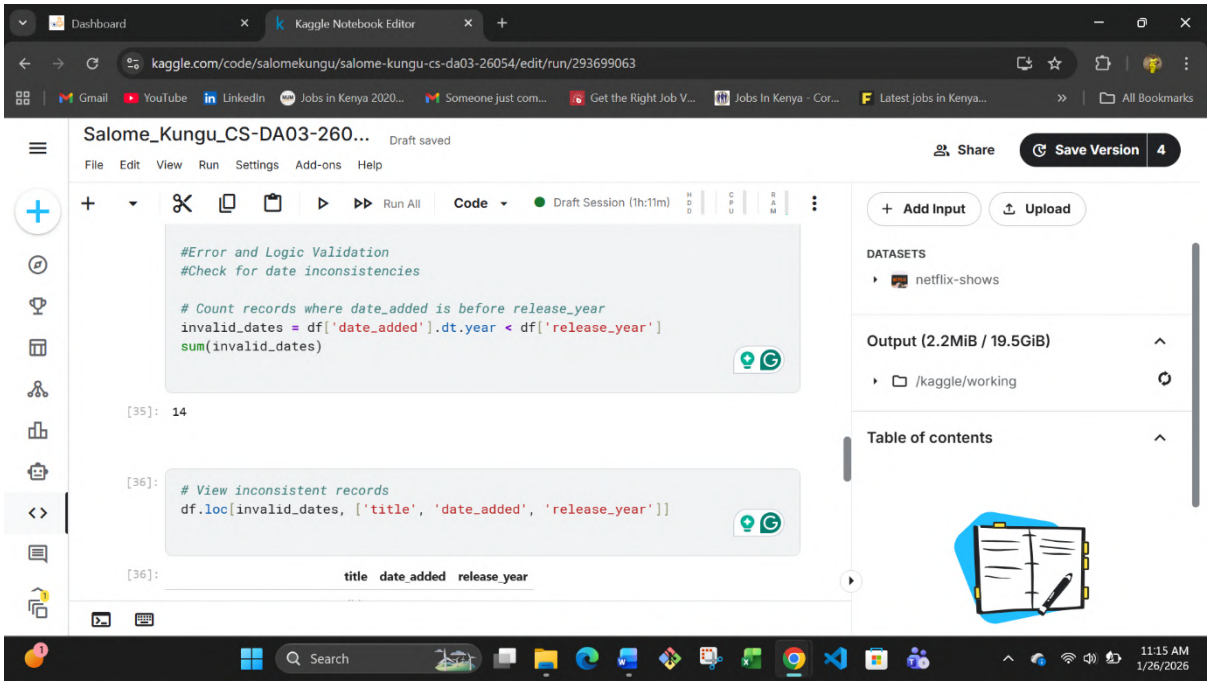
What:

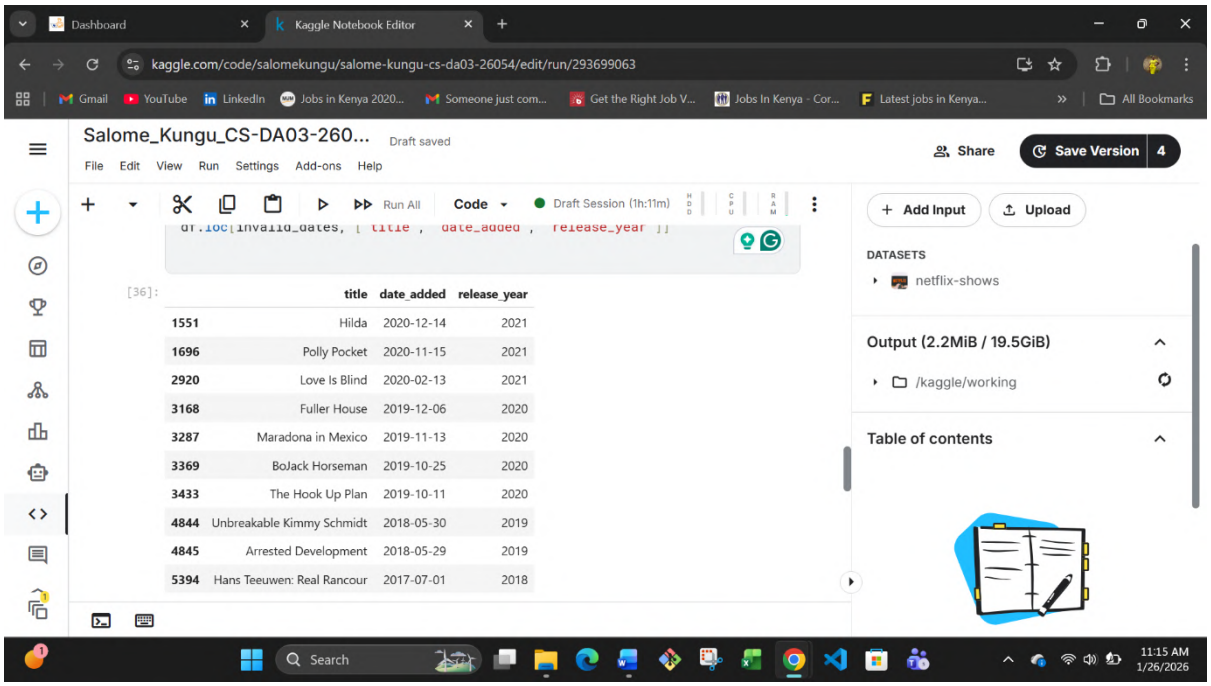
- Remove temporary columns created during wrangling.
- Confirm correct data types.
- Sample records for visual inspection.
- Reset the DataFrame index.

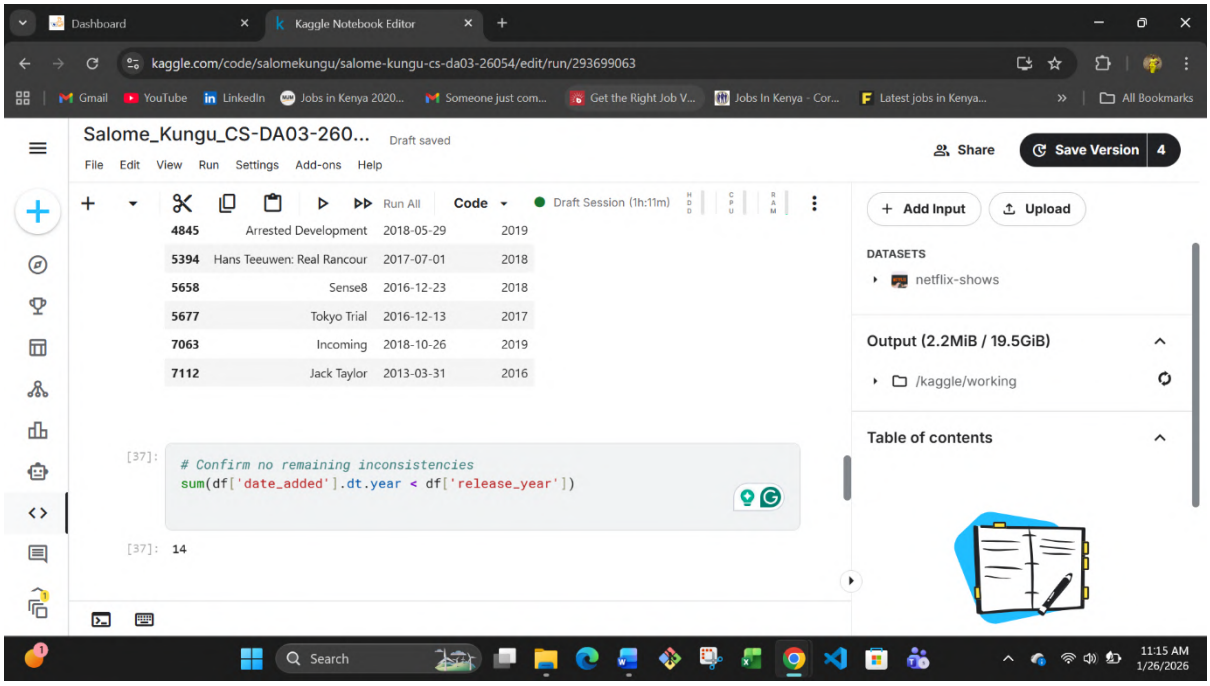
Why:

This step ensures the dataset is clean, consistent, and presentation-ready.



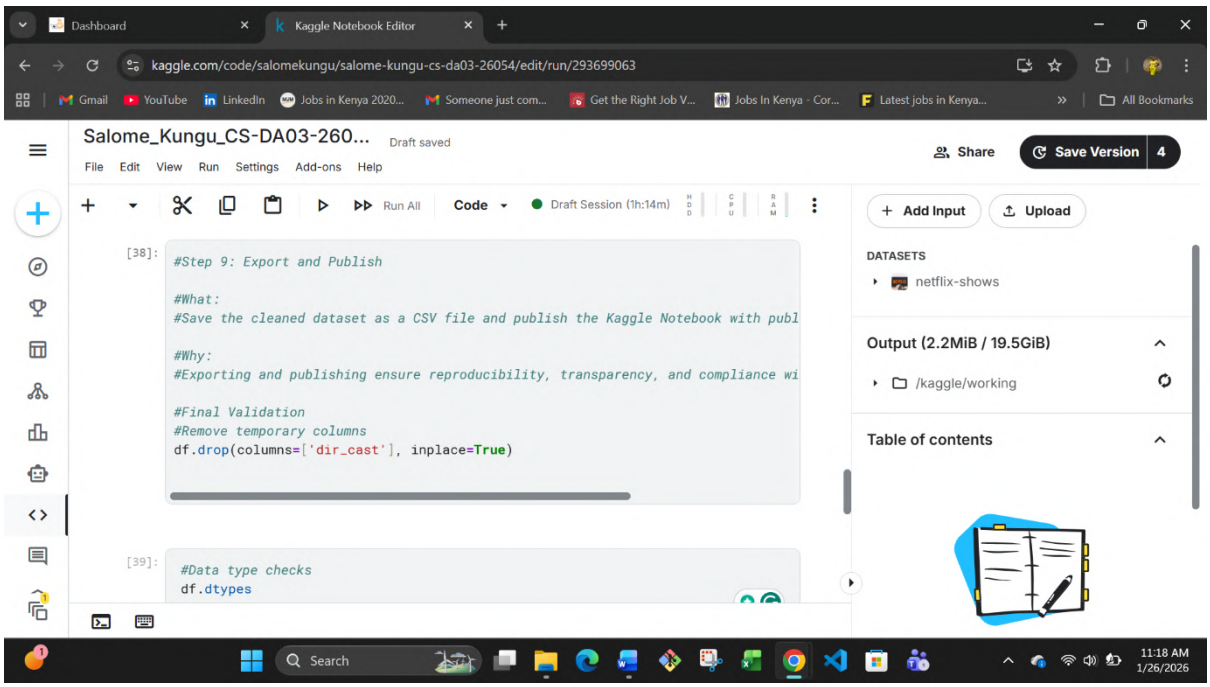

Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:11m)
#Error and Logic Validation
#Check for date inconsistencies
# Count records where date_added is before release_year
invalid_dates = df['date_added'].dt.year < df['release_year']
sum(invalid_dates)
[35]: 14
[36]:
View inconsistent records
df.loc[invalid_dates, ['title', 'date_added', 'release_year']]
[36]: title date_added release_year
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
11:15 AM
1/26/2026

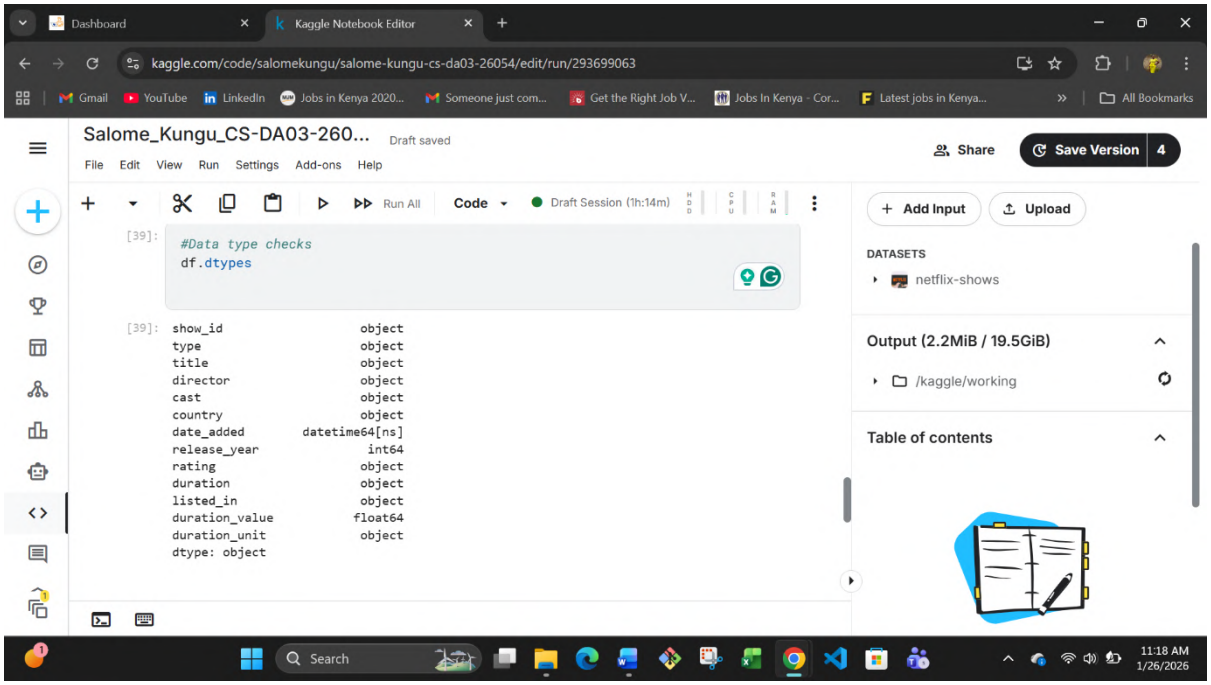

Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:11m)
[36]:
title date_added release_year
1551 Hilda 2020-12-14 2021
1696 Polly Pocket 2020-11-15 2021
2920 Love Is Blind 2020-02-13 2021
3168 Fuller House 2019-12-06 2020
3287 Maradona in Mexico 2019-11-13 2020
3369 BoJack Horseman 2019-10-25 2020
3433 The Hook Up Plan 2019-10-11 2020
4844 Unbreakable Kimmy Schmidt 2018-05-30 2019
4845 Arrested Development 2018-05-29 2019
5394 Hans Teeuwen: Real Rancour 2017-07-01 2018
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
Search
11:15 AM
1/26/2026

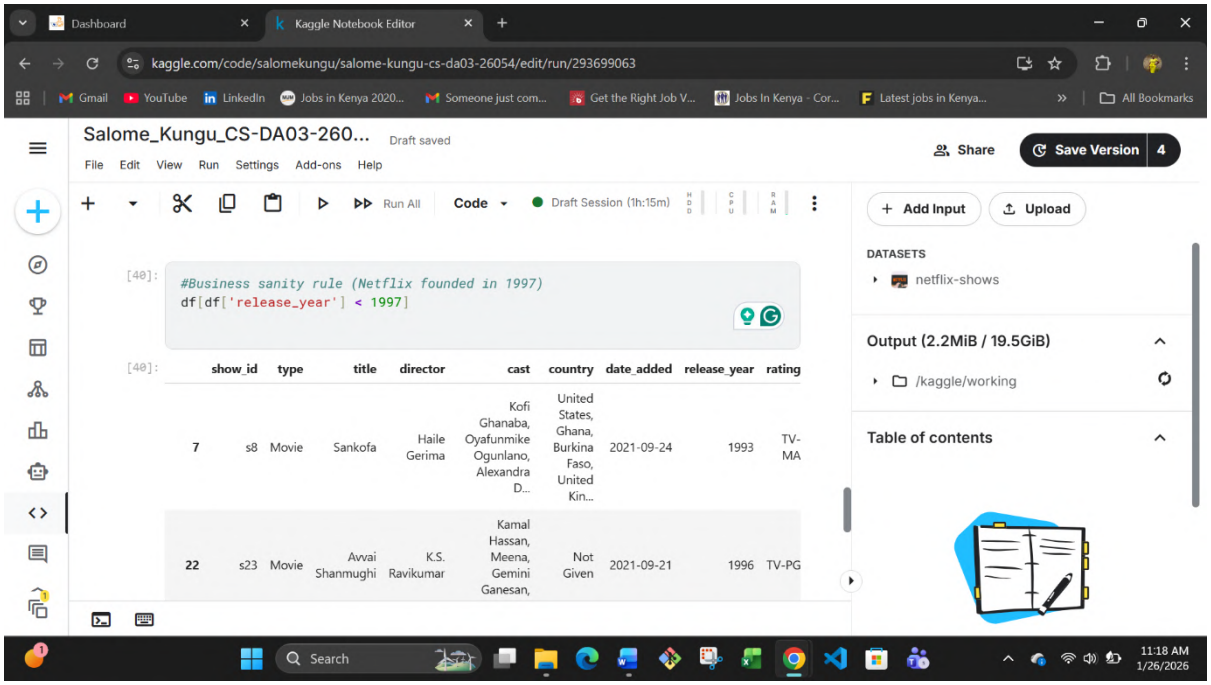


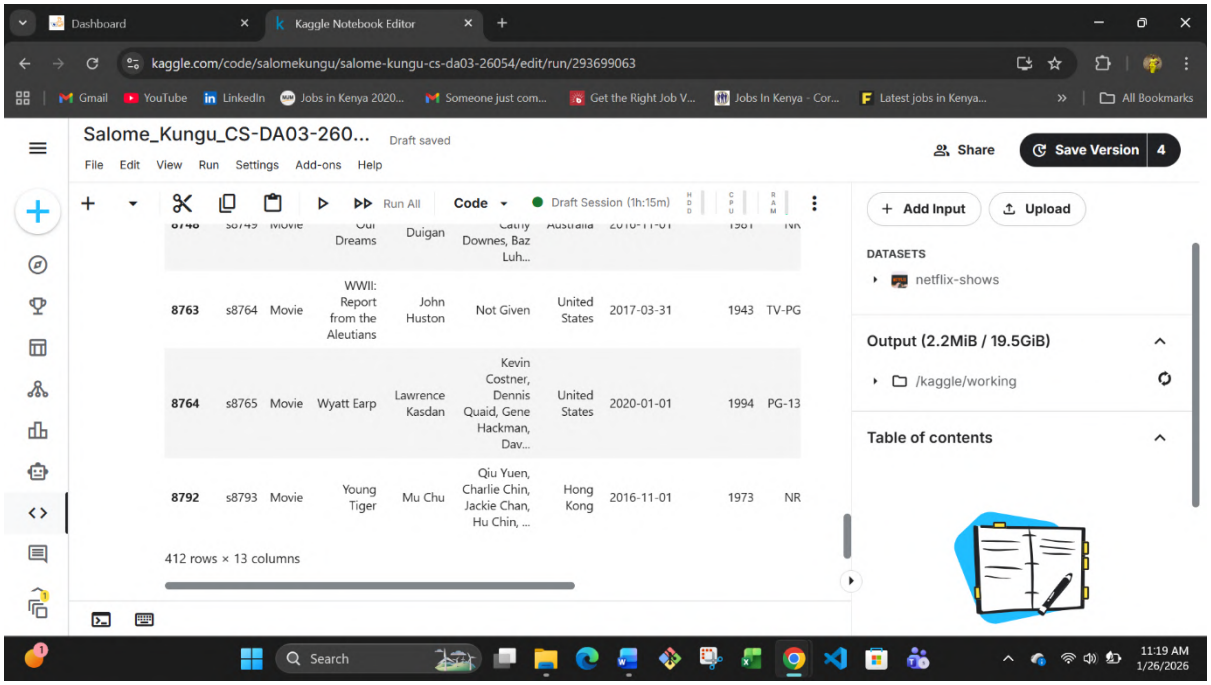
Step 9: Export and Publish

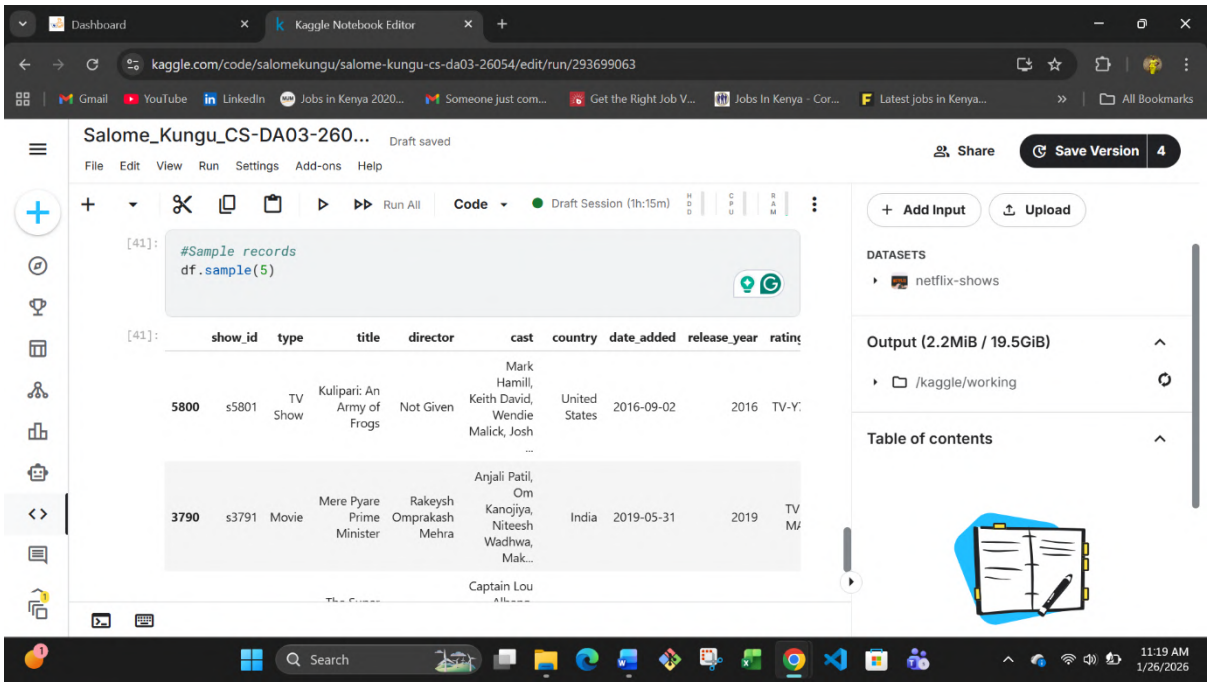
What:
Save the cleaned dataset as a CSV file and publish the Kaggle Notebook with public access enabled.
Why:
Exporting and publishing ensures reproducibility, transparency, and compliance with submission requirements.

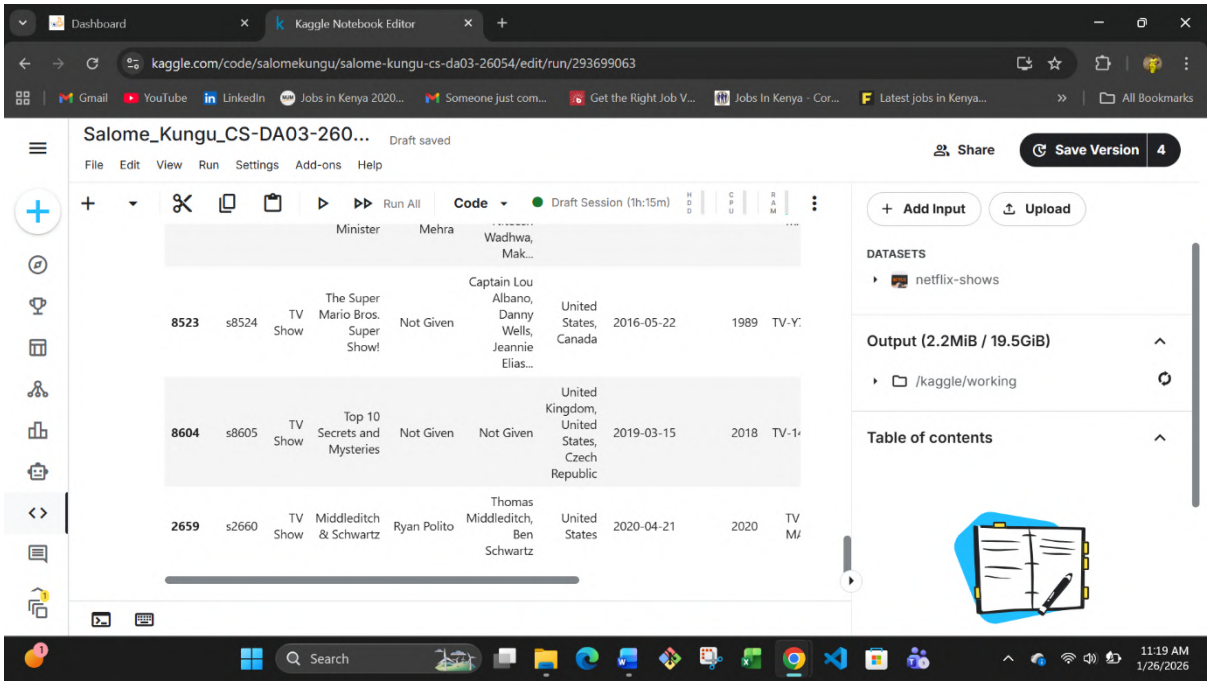


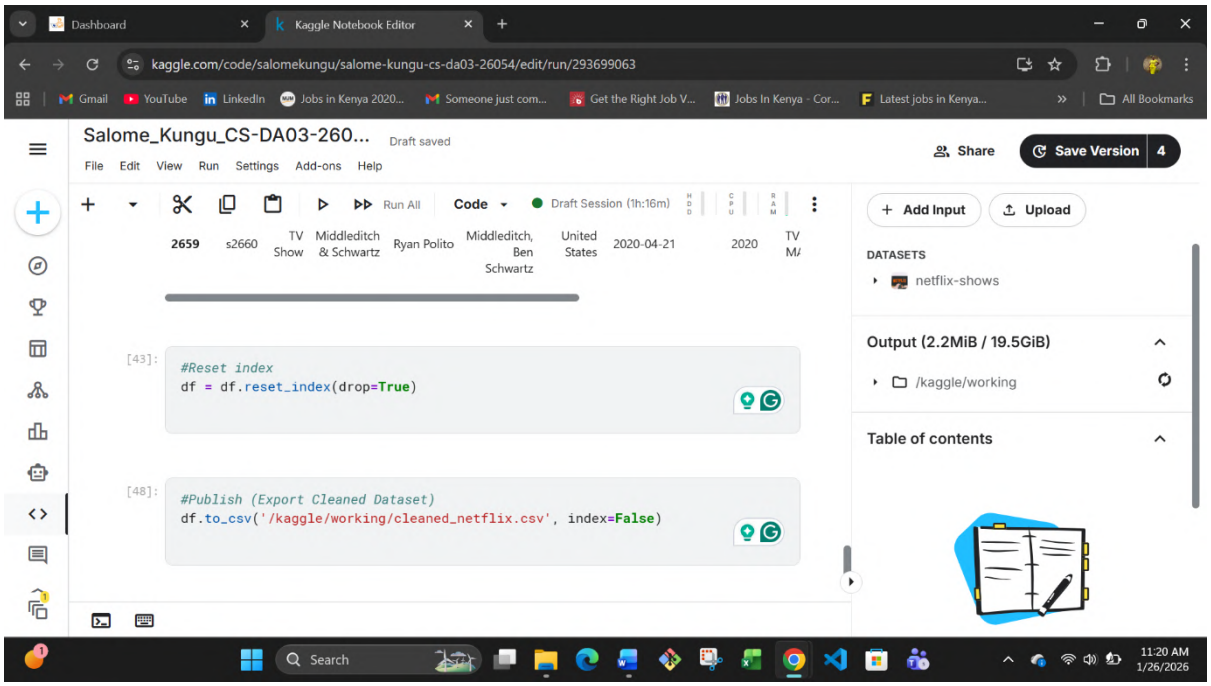

Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:14m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
[39]:
#Data type checks
df.dtypes
[39]:
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
Search
11:18 AM
1/26/2026

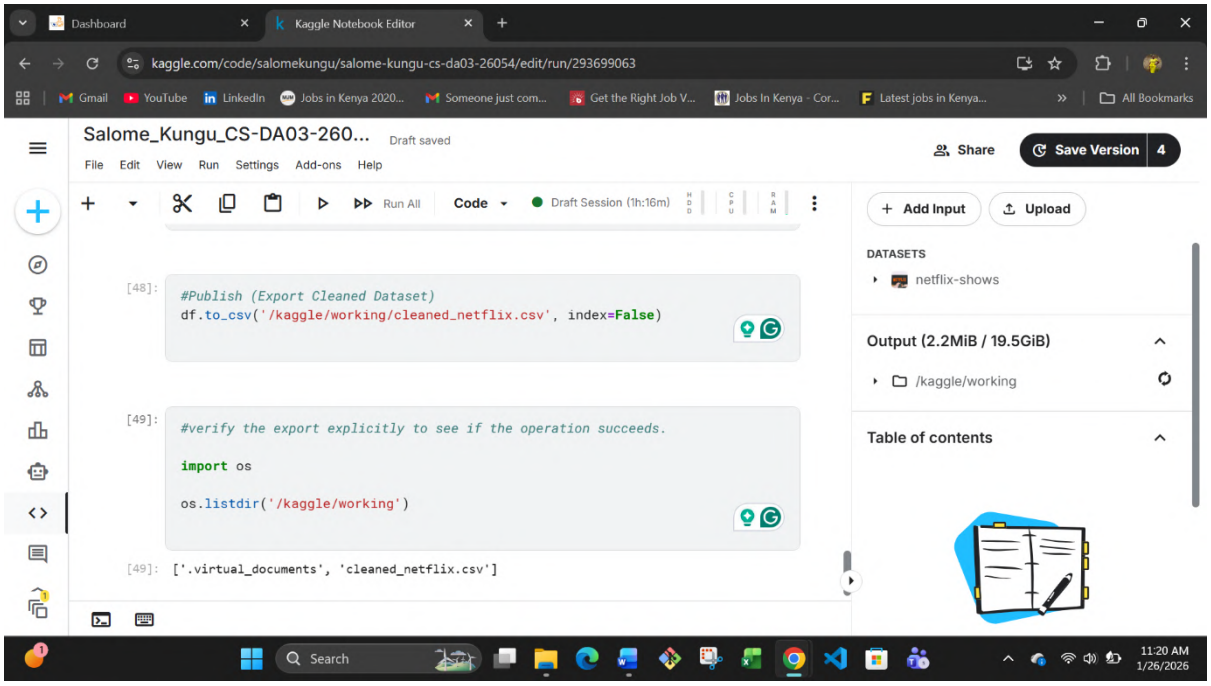

Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:15m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
[40]:
#Business sanity rule (Netflix founded in 1997)
df[df['release_year'] < 1997]
[40]:
show_id type title director cast country date_added release_year rating
7 s8 Movie Sankofa Haile Gerima Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D... United States, Ghana, Burkina Faso, United Kin... 2021-09-24 1993 TV-MA
22 s23 Movie Avvai Shanmughi K.S. Ravikumar Kamal Hassan, Meena, Gemini Ganesan, Not Given 2021-09-21 1996 TV-PG
Search
11:18 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:15m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
8763 s8764 Movie WWII: Report from the Aleutians John Huston Not Given United States 2017-03-31 1943 TV-PG
8764 s8765 Movie Wyatt Earp Lawrence Kasdan Kevin Costner, Dennis Quaid, Gene Hackman, Dav... United States 2020-01-01 1994 PG-13
8792 s8793 Movie Young Tiger Mu Chu Qiu Yuen, Charlie Chin, Jackie Chan, Hu Chin, ... Hong Kong 2016-11-01 1973 NR
412 rows × 13 columns
Search
11:19 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:15m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
[41]:
#Sample records
df.sample(5)
[41]:
show_id type title director cast country date_added release_year
5800 s5801 TV Show Kulipari: An Army of Frogs Not Given Mark Hamill, Keith David, Wendie Malick, Josh ... United States 2016-09-02 2016
3790 s3791 Movie Mere Pyare Prime Minister Rakeysh Omprakash Mehra Anjali Patil, Om Kanojiya, Niteesh Wadhwa, Mak... India 2019-05-31 2019
Captain Lou
Search
11:19 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:15m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
8523 s8524 TV Show The Super Mario Bros. Super Show! Not Given Captain Lou Albano, Danny Wells, Jeannie Elias... United States, Canada 2016-05-22 1989
8604 s8605 TV Show Top 10 Secrets and Mysteries Not Given Not Given United Kingdom, United States, Czech Republic 2019-03-15 2018
2659 s2660 TV Show Middleditch & Schwartz Ryan Polito Thomas Middleditch, Ben Schwartz United States 2020-04-21 2020
Search
11:19 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Salome_Kungu_CS-DA03-260...
Draft saved
File Edit View Run Settings Add-ons Help
Share
Save Version 4
Run All
Code
Draft Session (1h:16m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
2659 s2660 TV Show Middleditch & Schwartz Ryan Polito Middleditch, Ben Schwartz United States 2020-04-21 2020
[43]:
#Reset index
df = df.reset_index(drop=True)
[48]:
#Publish (Export Cleaned Dataset)
df.to_csv('/kaggle/working/cleaned_netflix.csv', index=False)
Search
11:20 AM
1/26/2026


Dashboard
Kaggle Notebook Editor
kaggle.com/code/salomekungu/salome-kungu-cs-da03-26054/edit/run/293699063
Gmail
YouTube
LinkedIn
Jobs in Kenya 2020...
Someone just com...
Get the Right Job V...
Jobs In Kenya - Cor...
Latest jobs in Kenya...
All Bookmarks
Salome_Kungu_CS-DA03-260...
Draft saved
File
Edit
View
Run
Settings
Add-ons
Help
Share
Save Version 4
Run All
Code
Draft Session (1h:16m)
Add Input
Upload
DATASETS
netflix-shows
Output (2.2MiB / 19.5GiB)
/kaggle/working
Table of contents
[48]:
#Publish (Export Cleaned Dataset)
df.to_csv('/kaggle/working/cleaned_netflix.csv', index=False)
[49]:
#verify the export explicitly to see if the operation succeeds.
import os
os.listdir('/kaggle/working')
[49]:
['.virtual_documents', 'cleaned_netflix.csv']
Search
11:20 AM
1/26/2026