

Data and Artificial Intelligence

Cyber Shujaa Program

Week 2 Assignment

Netflix Data Wrangling

Student Name: Cherotich Mercy

Student ID: CS-DA01-25091

Introduction

This week's assignment was data wrangling. I was not that new to the tools we were introduced to. I had written Python code before using vscode and Jupiter notebook. I had an issue with Kaggle notebook so I opted to use Jupiter notebook because I had it already installed in my laptop. Later I uploaded my code and the final dataset to my GitHub account ,shared the link to allow access.

The objectives of the assignment were:

1. Load the Netflix dataset from a CSV file and explore its structure using pandas.
2. Perform data discovery to assess data types, missing values, and quality issues.
3. Clean the dataset by handling duplicates, missing values, and formatting inconsistencies.
4. Transform and enrich the dataset using techniques like filtering, sorting, grouping, and feature extraction.
5. Validate the final dataset by checking consistency, completeness, and logical accuracy.
6. Export the final cleaned dataset to a .csv file ready for analysis or visualization.

Task

I downloaded the data from Kaggle. Imported pandas and finally loaded the data which was in a csv file .

```
import pandas as pd
```

```
df=pd.read_csv(r'C:\Users\ALLAN\Desktop\github\data cleaning\netflix_titles.csv')
```

Checked the data overview

```
#overview of the data  
df.info()
```

```
<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
```

Checked the number of rows and columns and also the list of all column names

```
#number of rows and columns  
df.shape
```

```
(8807, 12)
```

```
#list of all column names  
df.columns.tolist()
```

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

Checked the data types of each column

```
# datatype of each column
df.dtypes
```

```
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
```

Grouped and counted the number of missing values in each column

```
#Grouping and counting missing values in each column
df.isnull().sum()
```

```
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
```

Checked the duplicate values in each column

```
##Grouping and counting duplicate values in each column
df.duplicated().sum()
```

```
0
```

Converted date_added to datetime and separated duration into duration_value and duration_unit giving the below result.

```
#Structuring
#convert 'date_added'to datetime
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')

#separating duration into numeric value
df[['duration_value','duration_unit']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')
df
```

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	duration_value	duration_unit
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Given	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	90	min
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	Seasons
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas	France, Belgium	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International	To protect his family from a	1	Season

Converted duration value to numeric

```
#converting duration_value to numeric
df['duration_value'] = pd.to_numeric(df['duration_value'])
df
```

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	duration_value	duration_unit	dir_c	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Given	United States	2021-09-25	2020	PG-13	90 min	Documentaries	90.0	min	Kirs Johnson Not Gi
1	s2	TV Show	Blood & Water	Not Given	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	2.0	Seasons	N
				Sami Bouajila.						Crime TV			Ju	

Viewed the final result

```
#Viewing resulting columns
print(df[['duration_value','duration_unit']])
```

	duration_value	duration_unit
0	90.0	min
1	2.0	Seasons
2	1.0	Season
3	1.0	Season
4	2.0	Seasons
...
8802	158.0	min
8803	2.0	Seasons
8804	88.0	min
8805	88.0	min
8806	111.0	min

[8807 rows x 2 columns]

Started data cleaning ,Check duplicates, drop available duplicates, drop the description column

```
#Cleaning
#check duplicates/
df.duplicated().sum()
```

0

```
#drop duplicated rows
df=df.drop_duplicates()
```

```
#dropdescription column
df=df.drop(columns=['description'])
df
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	duration_value	duration_unit
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Given	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	90.0	min
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabani...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	2.0	Seasons
					Sami Bouailla						Crime TV		

Impute director values by using the relationship between cast and directors ,counted the unique values, checked if it repeated more than three times.

```
#impute Director values by using relationship between cast and director
df['dir_cast']=df['director']+'--'+ df['cast']
counts = df['dir_cast'].value_counts()
```

```
#counts unique values
filtered_counts =counts[counts>=3]
```

```
#check if repeated 3 or more times
filtered_values=filtered_counts.index
```

```
#getting the values
lst_dir_cast= list(filtered_values)
dict_direcast=dict()
for i in lst_dir_cast:
    director,cast = i.split('--')
    dict_direcast[director]= cast
for i in range(len(dict_direcast)):
    df.loc[(df['director'].isna())&(df['cast']==list(dict_direcast.items())[1]),'director'] = list(dict_direcast.items())[1][0]
```

```
#Assign not given to all other director
df.loc[df['director'].isna(), 'director'] = 'Not Given'
```

```
#directors to fill missing countries
directors = df['director']
countries = df['country']
```

```
#pairing each directors with their country
pairs = zip(directors,countries)

#converting the list of tuples into a dictionary
dir_cntry =dict(list(pairs))

#director matched to country values to fill in null country values
for i in range (len(dir_cntry)):
    df.loc[(df['country'].isna())&(df['director']==list(dir_cntry.items())[i][0]),'country']=list(dir_cntry.items())[i][1]

#assigning NOT GIVEN to all other country field
df.loc[df['country'].isna(),'country'] = 'Not Given'

##assigning NOT GIVEN to all field
df.loc[df['cast'].isna(),'cast'] = 'Not Given'

#dropping other row records with null
df.drop(df[df['date_added'].isna()].index,axis)
df.drop(df[df['rating'].isna()].index,axis = 0,input =True)
df.drop(df[df['duration'].isna()].index,axis=0,input =True)
```

Checked if any added_dates come before the release_year

```
#checking if any added_date that comes before release_year
import datetime as dt
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
sum(df['date_added'].dt.year < df['release_year'])
df.loc[(df['date_added'].dt.year < df['release_year']),
['date_added','release_year']]
```

	date_added	release_year
1551	2020-12-14	2021
1696	2020-11-15	2021
2920	2020-02-13	2021
3168	2019-12-06	2020
3287	2019-11-13	2020
3369	2019-10-25	2020
3433	2019-10-11	2020
4844	2018-05-30	2019
4845	2018-05-29	2019
5394	2017-07-01	2018
5658	2016-12-23	2018
5677	2016-12-13	2017

Found the sum and finally dropped the dir_cast column

```
sum(df['date_added'].dt.year <
df['release_year'])
```

14

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

Generated five samples of the data

```
df.sample(5)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	duration_value	duration_unit
3848	s3849	TV Show	Cinta Teruna Kimchi	Not Given	Nazim Othman, Johan As'ari, Nur Risteena, Nien...	Not Given	2019-05-08	2016	TV-PG	1 Season	International TV Shows, TV Dramas	1.0	Season
4901	s4902	Movie	The Clapper	Dito Montiel	Ed Helms, Amanda Seyfried, Tracy Morgan, Brend...	United States	2018-05-01	2017	R	90 min	Comedies, Dramas, Independent Movies	90.0	min
1274	s1275	TV Show	Canine Intervention	Not Given	Jas Leverette	United States	2021-02-24	2021	TV-PG	1 Season	Reality TV	1.0	Season
3336	s3337	TV Show	Haikyuu!!	Not Given	Ayumu Murase, Kaito Ishikawa, Satoshi Hino, Mi...	Japan	2019-11-01	2015	TV-14	2 Seasons	Anime Series, International TV Shows, Teen TV ...	2.0	Seasons
3752	s3753	TV Show	Marvel's Jessica Jones	Not Given	Krysten Ritter, David Tennant, Rachael Taylor,...	United States	2019-06-14	2019	TV-MA	3 Seasons	Crime TV Shows, TV Action & Adventure, TV Dramas	3.0	Seasons

Reset the index

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

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	duration_value	duration_unit
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Given	United States	2021-09-25	2020	PG-13	90 min	Documentaries	90.0	min
1	s2	TV Show	Blood & Water	Not Given	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	2.0	Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabil...	France, Belgium	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, TV International TV Shows, TV Act...	1.0	Season
3	s4	TV Show	Jailbirds New Orleans	Not Given	Not Given	Not Given	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	1.0	Season
4	s5	TV Show	Kota Factory	Not Given	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	2.0	Seasons

Published and saved as csv file

```
df.to_csv(r'C:\Users\ALLAN\Desktop\github\data cleaning\netflix_titles.csv', index=False)
```

Link to Code: <https://github.com/Chero-dev/Data-wrangling-week-2-assignment.git>

Conclusion

This week I gained a lot on data wrangling which involves understanding the data, cleaning and doing data validation . I have posted my project on my GitHub account so as to build a portfolio that I can showcase on my CV as I look for jobs in Data and AI.