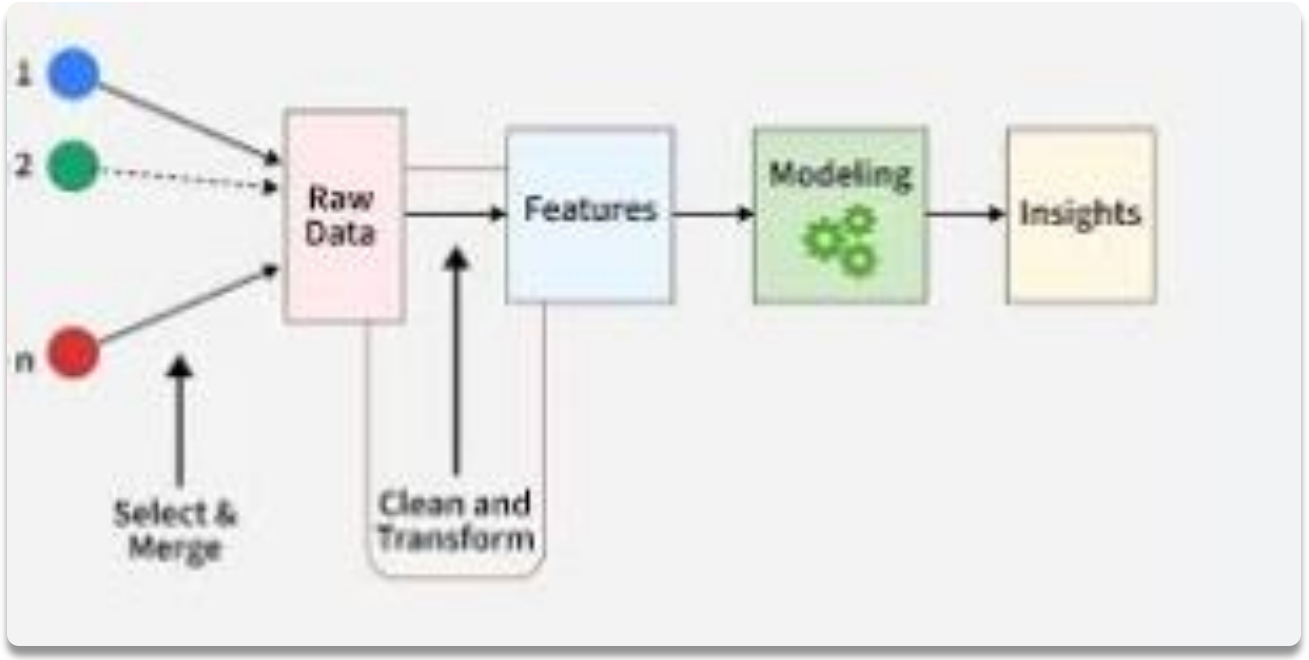
NETFLIX DATASET – **Feature Engineering & Data Transformation**(Week_4)

DATA SCIENCE INTERNSHIP-FUTUREXCEL

ANALYTIC BY :**SUMBUL ABDUL.WAHID**

Introduction:

- In this task, feature engineering was applied to the Netflix dataset to enhance the quality and usability of the data. The main objective was to create meaningful features from existing variables and prepare the dataset for future analytical or machine learning tasks.



Dataset Used

- The cleaned Netflix dataset prepared in Week 1 was used for this task. Using a pre-cleaned dataset ensured consistency and allowed the focus to remain on feature transformation rather than data cleaning.

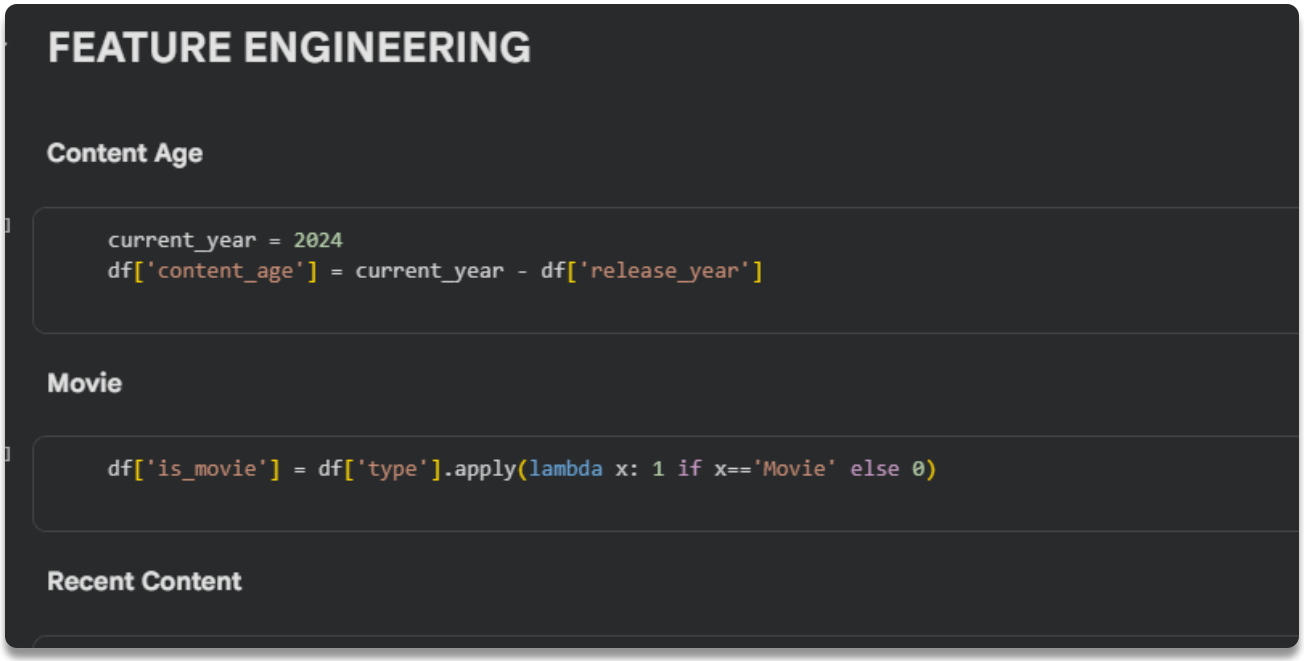
```
import pandas as pd

df = pd.read_csv("/content/cleaned_netflix_data (4).csv")
df.head()
```

	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	Not Available	United States	2021-09-25	2020.0	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021.0	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...	Unknown	2021-09-24	2021.0	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	Unknown	Not Available	Unknown	2021-09-24	2021.0	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
		TV	Kota		Mayur More, Jitendra Kumar					2	International TV	In a city of

Feature Engineering

- Several new features were created from existing columns to better represent the underlying patterns in the data. These features include content age, content type indicators, simplified rating groups, and duration in numerical form. Feature creation helped convert raw information into more useful analytical variables.



Feature Engineering

```
df.head()
```

type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	content_age	is_movie	is_recent	d
Movie	Dick Johnson Is Dead	Kirsten Johnson	Not Available	United States	2021-09-25	2020.0	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	4.0	1	1	
TV Show	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021.0	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	3.0	0	1	
TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	2021-09-24	2021.0	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	3.0	0	1	
TV	Jailbirds		Not						Docuseries,	Feuds, flirtations and				

Duration

```
df['duration_minutes'] = df['duration'].str.extract('(\d+)').astype(float)
```

```
<>:1: SyntaxWarning: invalid escape sequence '\d'
<>:1: SyntaxWarning: invalid escape sequence '\d'
/tmp/ipython-input-3957158811.py:1: SyntaxWarning: invalid escape sequence '\d'
  df['duration_minutes'] = df['duration'].str.extract('(\d+)').astype(float)
```

Rating Group

```
def rating_group(r):
    if r in ['G','TV-Y']:
        return 'Kids'
    elif r in ['PG','PG-13']:
        return 'Teen'
    else:
        return 'Adult'

df['rating_group'] = df['rating'].apply(rating_group)
```

Encoding and Scaling

ENCODING

```
[ ]  from sklearn.preprocessing import LabelEncoder

     le = LabelEncoder()
     df['rating_group_encoded'] = le.fit_transform(df['rating_group'])

[ ]  df[['rating_group','rating_group_encoded']].head()
```

	rating_group	rating_group_encoded
0	Teen	2
1	Adult	0
2	Adult	0
3	Adult	0
4	Adult	0

- Categorical variables were converted into numerical format using encoding techniques. Numerical features were scaled to bring them onto a common range. This step improves data consistency and is essential for model readiness.

Scaling:

SCALING

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
df['content_age_scaled'] = scaler.fit_transform(df[['content_age']])
```

```
df[['content_age','content_age_scaled']].head()
```

	content_age	content_age_scaled
0	4.0	-0.975877
1	3.0	-1.240022
2	3.0	-1.240022
3	3.0	-1.240022
4	3.0	-1.240022

Before and After Comparison

- A comparison of the dataset before and after transformation shows improved structure and standardized values. The engineered dataset is more suitable for advanced analysis.

BEFORE vs AFTER

```
df[['content_age','content_age_scaled']].describe()
```

	content_age	content_age_scaled
count	8088.000000	8.088000e+03
mean	7.694486	-2.811247e-17
std	3.786044	1.000062e+00
min	3.000000	-1.240022e+00
25%	5.000000	-7.117330e-01
50%	7.000000	-1.834444e-01
75%	9.000000	3.448442e-01
max	20.000000	3.250431e+00

Conclusion :

- This task strengthened my understanding of feature engineering techniques and highlighted their importance in transforming raw data into a model-ready format.