data-extraction-cleaning

April 20, 2025

1 Data Extraction and Cleaning

1.1 Import Essential Libraries

[49]:
```
# Basic Libraries
import numpy as np
import pandas as pd
```

1.2 Import Data

[50]:
```
df = pd.read_csv('datasets/full-twitch-data.csv')
df.head()
```

[50]:

	Channel	Watch time(Minutes)	Stream time(minutes)	Peak viewers	Average viewers	Followers	Followers gained	Views gained	Partnered	Mature	Language
0	xQcOW	6196161750	215250	222720	27716	3246298	1734810	93036735	True	False	English
1	summit1g	6091677300	211845	310998	25610	5310163	1370184	89705964	True	False	English
2	Gaules	5644590915	515280	387315	10976	1767635	1023779	102611607	True	True	Portuguese
3	ESL_CSGO	3970318140	517740	300575	7714	3944850	703986	106546942	True	False	English
4	Tfue	3671000070	123660	285644	29602	8938903	2068424	78998587	True	False	English

1.3 Clean Data

Lets first rename the column values to make it easier on us when retrieving specific items, and add a bit more metrics from our original values.

[51]:
```
new_cols = []
for col in df.columns:
  new_cols.append(col.replace(' ', '_').replace('(', '_').replace(')', '').
 ↪lower())
df.columns = new_cols
df.head()
```

[51]:

	channel	watch_time_minutes	stream_time_minutes	peak_viewers	\
0	xQcOW	6196161750	215250	222720	
1	summit1g	6091677300	211845	310998	
2	Gaules	5644590915	515280	387315	
3	ESL_CSGO	3970318140	517740	300575	
4	Tfue	3671000070	123660	285644	

	average_viewers	followers	followers_gained	views_gained	partnered	\
0	27716	3246298	1734810	93036735	True	
1	25610	5310163	1370184	89705964	True	
2	10976	1767635	1023779	102611607	True	
3	7714	3944850	703986	106546942	True	
4	29602	8938903	2068424	78998587	True	

	mature	language
0	False	English
1	False	English
2	True	Portuguese
3	False	English
4	False	English

[52]:
```
df['watch_time_hours'] = df['watch_time_minutes'] / 60
df['stream_time_hours'] = df['stream_time_minutes'] / 60
df['followers_per_hour'] = df['followers_gained'] / df['stream_time_hours']
df['views_per_follower'] = df['views_gained'] / df['followers_gained'] # how␣
 ↪many views gained per one follower gained
df['engagement_rate'] = df['average_viewers'] / df['followers'] # how many␣
 ↪views gained per one follower gained
df.head()
```

[52]:

	channel	watch_time_minutes	stream_time_minutes	peak_viewers	\
0	xQcOW	6196161750	215250	222720	
1	summit1g	6091677300	211845	310998	
2	Gaules	5644590915	515280	387315	
3	ESL_CSGO	3970318140	517740	300575	
4	Tfue	3671000070	123660	285644	

	average_viewers	followers	followers_gained	views_gained	partnered	\
0	27716	3246298	1734810	93036735	True	
1	25610	5310163	1370184	89705964	True	

```
2            10976    1767635           1023779     102611607       True
3             7714    3944850            703986     106546942       True
4            29602    8938903           2068424      78998587       True

   mature    language  watch_time_hours  stream_time_hours  \
0   False     English      1.032694e+08            3587.50
1   False     English      1.015280e+08            3530.75
2    True  Portuguese      9.407652e+07            8588.00
3   False     English      6.617197e+07            8629.00
4   False     English      6.118333e+07            2061.00

   followers_per_hour  views_per_follower  engagement_rate
0          483.570732           53.629351         0.008538
1          388.071656           65.470013         0.004823
2          119.210410          100.228279         0.006209
3           81.583729          151.348098         0.001955
4         1003.602135           38.192647         0.003312
```

[53]:
```python
# Check for any null values
print("\nNull Values:")
print(df.isnull().sum())

# Check what each column's data type is
print("\nData Types:")
print(df.dtypes)

# Check for any duplicate rows
print("\nDuplicate Rows:")
print(df.duplicated().sum())
```

```
Null Values:
channel                0
watch_time_minutes     0
stream_time_minutes    0
peak_viewers           0
average_viewers        0
followers              0
followers_gained       0
views_gained           0
partnered              0
mature                 0
language               0
watch_time_hours       0
stream_time_hours      0
followers_per_hour     0
views_per_follower     0
engagement_rate        0
```

```
dtype: int64

Data Types:
channel                 object
watch_time_minutes       int64
stream_time_minutes      int64
peak_viewers             int64
average_viewers          int64
followers                int64
followers_gained         int64
views_gained             int64
partnered                 bool
mature                    bool
language                object
watch_time_hours       float64
stream_time_hours      float64
followers_per_hour     float64
views_per_follower     float64
engagement_rate        float64
dtype: object

Duplicate Rows:
0
```

This is good! There are no null values in any of the rows and the data types look to be correct for each associated column.

[54]:
```
df.to_csv('datasets/twitch-data-cleaned.csv', index=False)
```