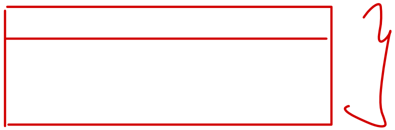
Data Cleaning → how to make the dataset ready to be consumed.

Analysis

Format

txt → csv



Count, groupby

right

Challenges

① Missing values

② Date format

③ Comma separated values (Nested values)

④ Duration

Title	Director	Cast	listed_in	Country
ABC	D1, D2	C1, C2, C3	G1, G2	Cont1, Cont2
XYZ	D2, D3	C2, C5	G1, G2	Cont1, Cont2

[Sort on the basis of movies directed by each director.

+ Director with most number of movies

Nested

Step 1 split (', '

	Cast0	Cast1	Cast2
ABC	C1	C2	C3

Step 2 Stack (col into rows)

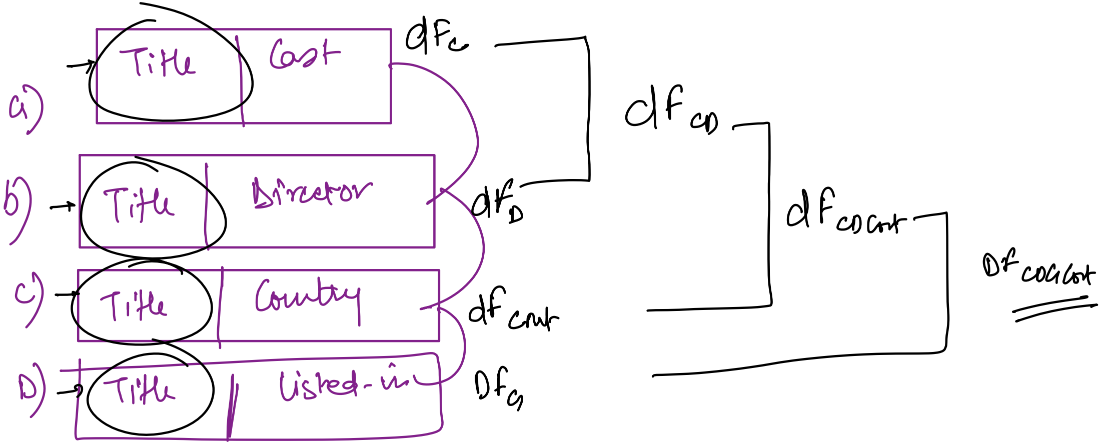
ABC	C1
ABC	C2
ABC	C3

df_{final}.groupby("Director")["Title"].nunique().sort_values(ascending=False)

Title	Director	Cast
ABC	D1	C1
ABC	D1	C2
ABC	D2	C1
ABC	D2	C2
ABC	D1	C3
ABC	D2	C3

Unnested

Step3



Step4 Merge → left, inner

Step5 Merge it back with remaing cols of the original DF using "Title"

df_{final}

- → Most popular actor & director pair.
- → Genre with most number of movies.

(*) {Date_added} → datetime → Extract month, year, day, day of week

↳ pd.to_datetime (___) →

What insights you can find if you fix this ?

→ month when most movies were added

year

→ day of week →

Duration

Split(' ')[0]

Title	Type	Duration	Duration[0]
ABC	Movie	90 min	90
XYZ	TV	2 season	2
MNO	Movie	110 minutes	110

Average runtime of movies & tv shows

Median runtime

↳ min

seasons

df.groupby('Type')(Duration[0]).median()

so we can say what is the most popular format in each country or genre..do people watch longer movies in some countries?

Missing values

⊛ Mode → most occurring Cat

1. Unknown
2. Delete ✓
3. Mode ✓

DFA

Actor	Director	Mode (Director)
A1	D1	D1
A1	D1	D1
A1	D1	D1
A1		D1
A1	D3	D1
A2	D5	D5
A2	D5	D5
A2	D5	D5
A1	D1	D1
A1	D1	D1
A2	D4	D5
A2		D5

Mode (Director) → D1

By Method → you are right ?

"Most Appropriate" imputation

How to do this

Reference column I will see → accordingly I will create mode & then inputs

df.groupby("Actor")[Director].mode()

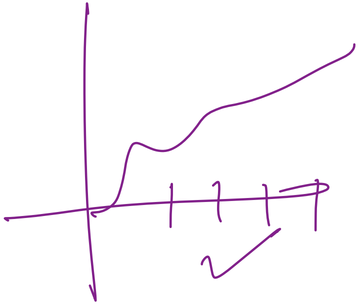
Actor	Mode(Director)
A1	D1
A2	D5

DfB

Soln

Merge DfA & DfB on the basis of Actor

Analysis



Insights