

Netflix

web streaming platform

↳ Goals

- ① Get more users onboarded.      ② Retain existing customers  
# Get more & more new & updated content.

# Recommendation:

# ML Pipeline

- ① Gather data
- ② Data preprocess
- ③ EDA 

Visually

Code
- ④ Model
- ⑤ Predict
- ⑥ Evaluate & Remodel
- ⑦ Deploy -

→ Cleaning  
Missing value  
Transformation  
Scaling  
Encoding

## EDA

- ① Univariate
- ② Bivariate
- ③ Multivariate

- ① Missing Values ✓
- ② Inconsistencies ✓
- ③ Unnesting ?
- ④ Date Column ✓
- ⑤ Duplication ✓

Interact in some albums

Duration

Movies

( ) min

( ) season

↳ TVshow

A	B
●	●
●	●
	●

## Missing Values

① Drop

② Imputation

↳ find best possible guess

Data is very expensive

\* How much data is missing?

\* How sensitive data is?

## Categorical

Mode  
"Unknown"

## Numerical

Mean, median,

0, -1

\* unknown

\* absence of property

0  
0  
0  
0  
-1 ← missing

Paypal

1%

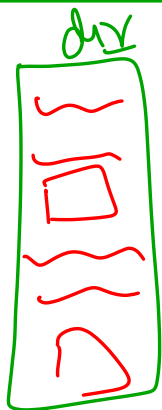


Amazon

5%



feedback




mode




- Christopher Nolan




grouping

<u>dir</u> count	
	A ✓
	A ✓
	A ✓

dir	country	genre	show type

  	B
	B
	B

	C C C
---	-------------

listed in "A1, A2, A3" =

split(",") → ["A", "B", "C"]

C1	C2	C3	C4	listed in
CD1	CD2	CD3	CD4	"A1, A2, A3"

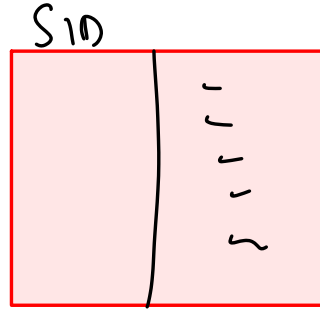
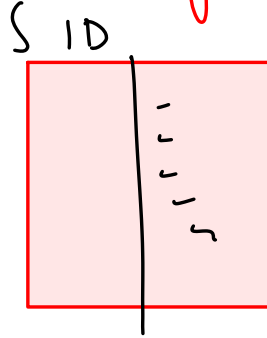
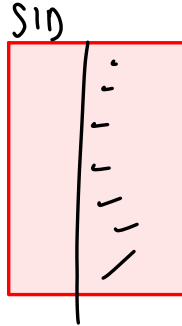
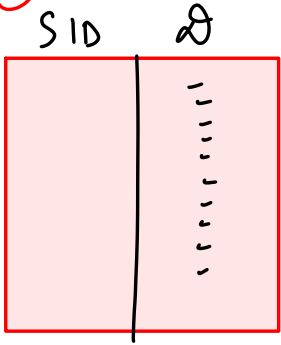
Unnesting

C1	C2	C3	C4	listed in
CD1	CD2	CD3	CD4	A1
CD1	CD2	CD3	CD4	A2
CD1	CD2	CD3	CD4	A3

pd. stack ( )

pd. explode ( )

(1) directors, Cast, Country, Listed In.

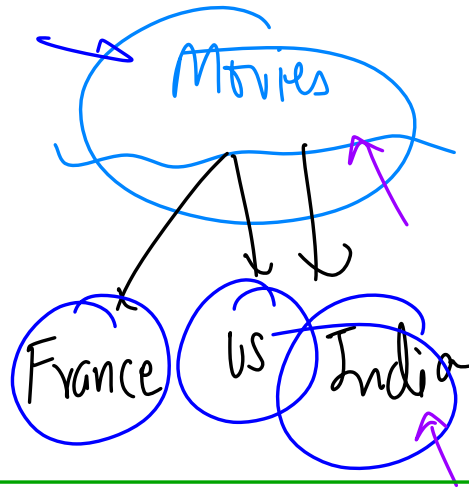
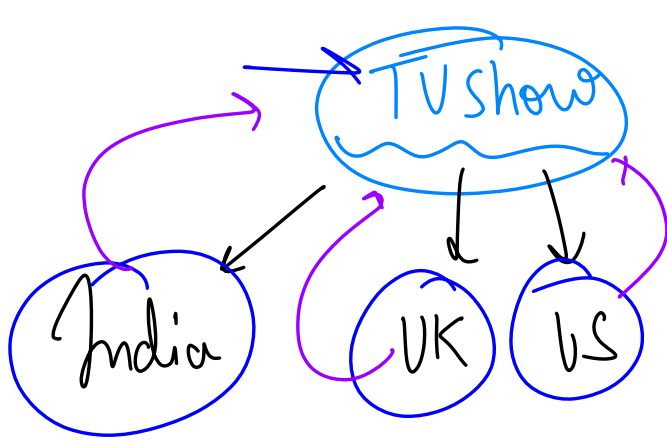


pd. merge,



Unique  
show id  
title

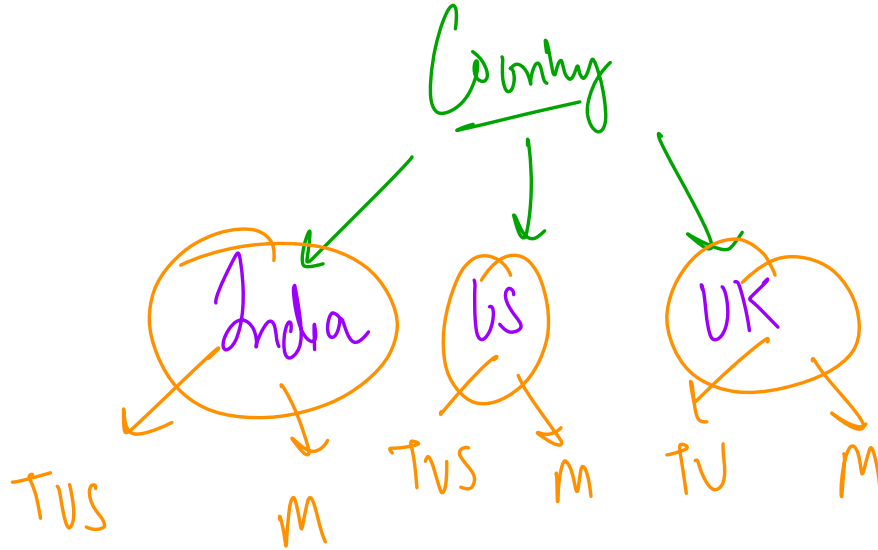




EIDA

①	UA	}
②	BA	
③	MA	

Visualize

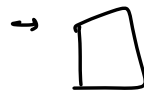


Suggestion

(1)



Code



Value / Char  
table

make down cell

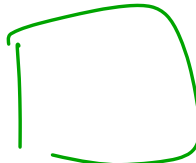
↳ All the important observations



Insight



Recommendation



ipynb → Run Cells → Insight + Recommendation.

pdf

ipynb as HTML page

HTML to pdf

