Netflix

Practice

medium to tough

Programing → Python ≤ numpy, Pandas, Seaborn

Follow the steps :

Assumptions working good for netflix

Insights & recommendation + trend.

Challenges

1. Missing (NaN ....) → Mode imputation, median / mean
2. Duration col. → 90 min, 1 seas → Avg runtime for actors
3. ↳, ↳, ↳, ↳

challenge : why ?

Actors	movies
A, B, C	M1
A, B	M2
B	M3
X, Y, A	M4
D	M5

A → Rajpal yadav

Q) Get me the no. of movies every actor has worked in.

Group by A → now

Nested → Unnested data

Datatype issues . pd.Datetime(...)

- dt.year
- dt.month
- dt.week

Nested data

→ Cast, Directors, Country, listed_in

Le if we do not fix it.

- Most popular actor in our DB
- Best Actor - Director combo.
- Famous actor in a country.

Title	Cast	Director	Country
ABC	C1, C2	D1	
XYZ	C3, C4	D2, D3	
Mano	C5, C6	D4	

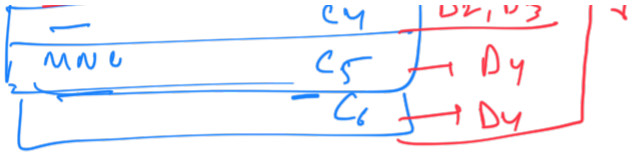
Unnesting → How many rows in total.

ABC	C1	D1
ABC	C2	D1
XYZ	C3	D2, D3

Unnesty

Cast Direct and good

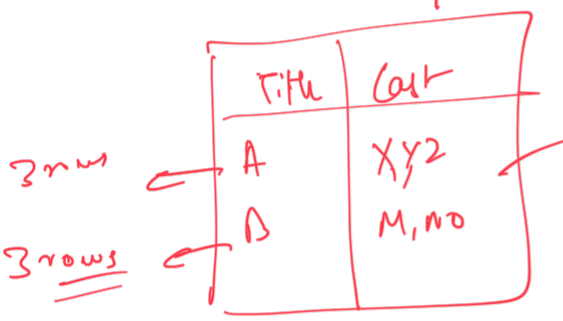
1	ABC		D1
2	ABC		D1
3	XYZ	C3	D2
4	XYZ	C3	D3



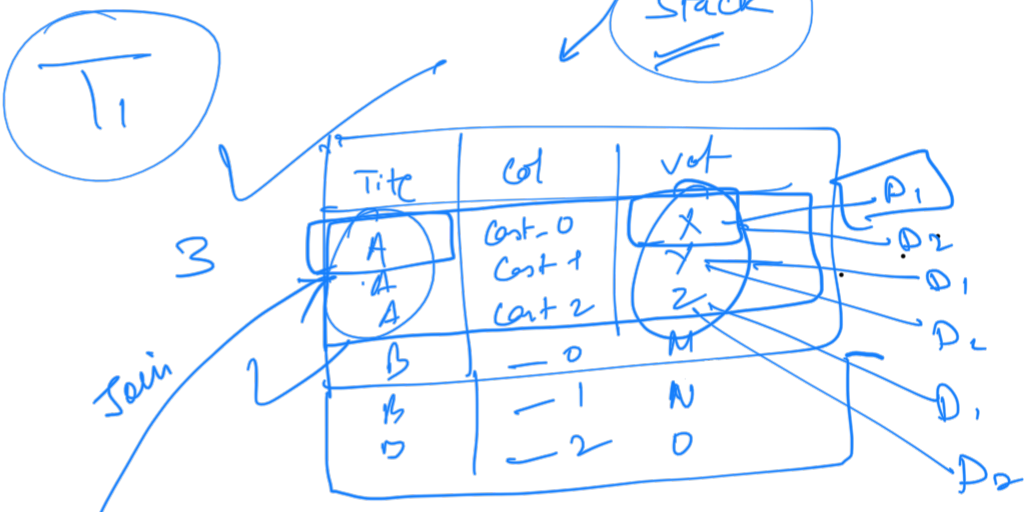
5	XYZ	C4	D2
6	XY2	C4	D3
7	MNO	C5	D4
8	MN1	C6	D4

Steps to Solve Unnesting

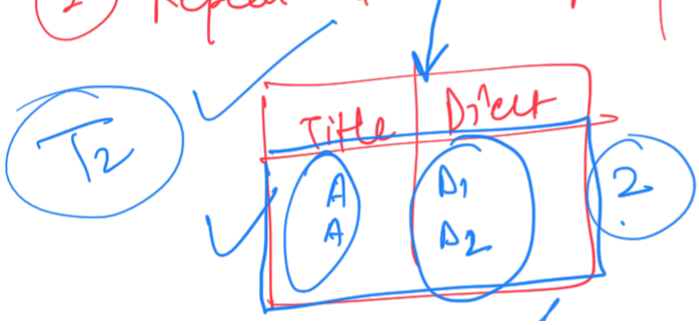
(1) Treat every column differently

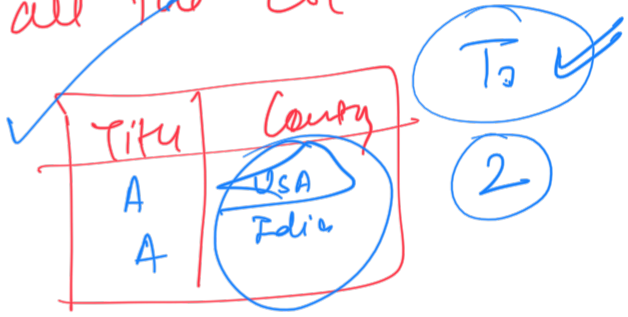


Title	Cost_0	Cost_1	Cost_2
A	X	Y	2
B	M	N	O

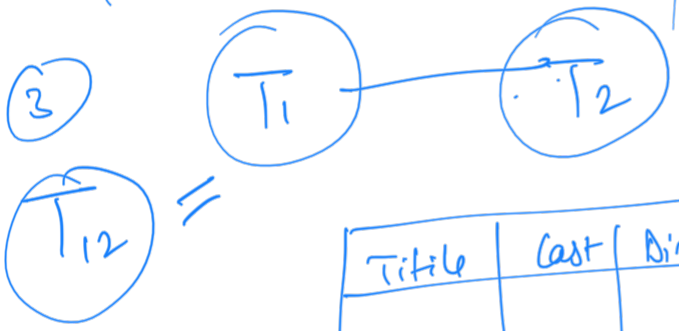


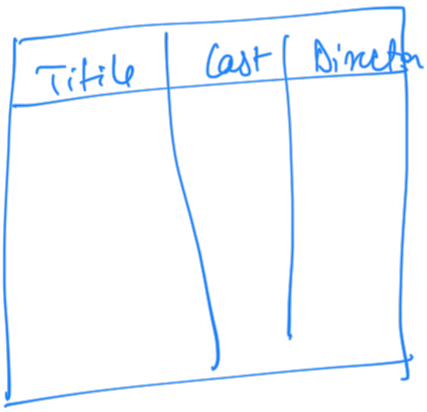
(2) Repeat this step for all the Col

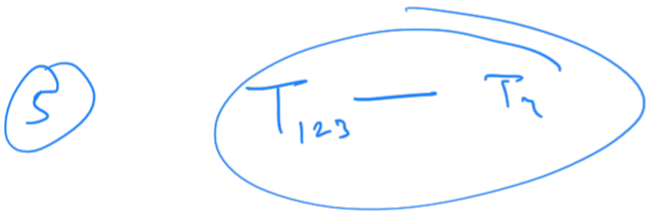




Separate table for each cd

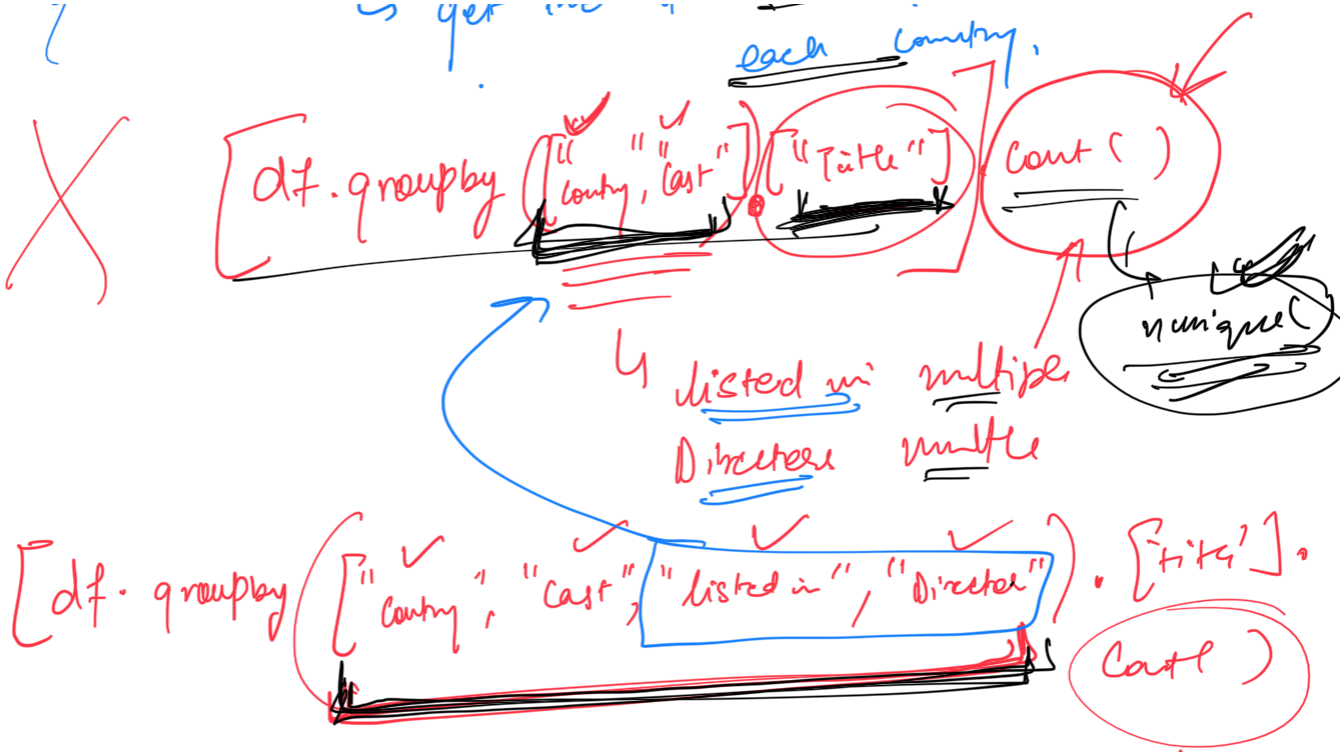






Tconsolidated

- Q. Is T_C unique at each row level.
  - No, the most popular actor in



(4) Join with original dataset on 'title' & fetch the remaining column.

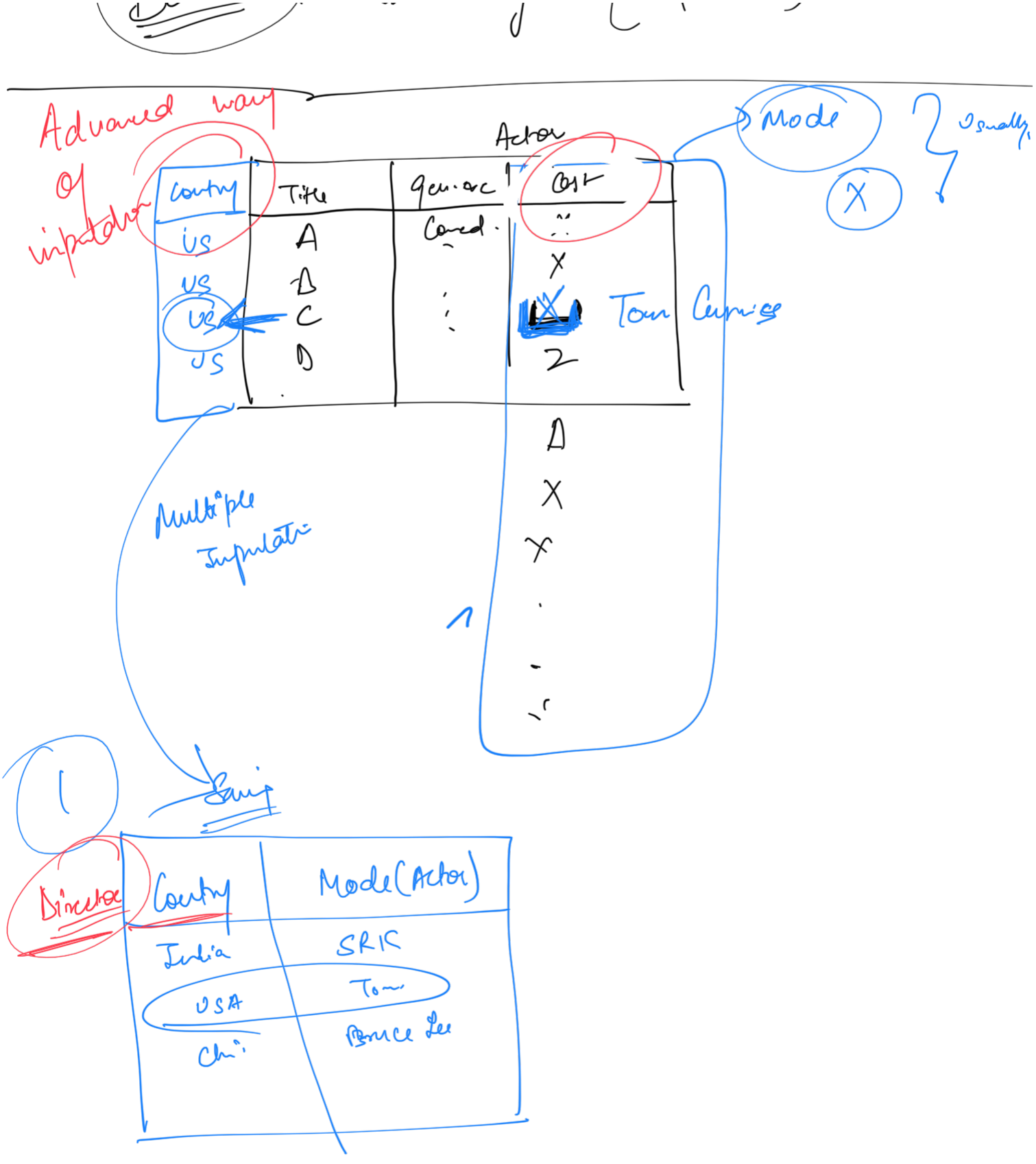
Easy

Questions

Missing value imputation

(1) Mean / median / mode. ✓ Right.

Roles → industry. (Optional)



Processing of the data is complete

EDA

EDA

1. Univariate Analysis
   - Displots
   - Histograms
   - Box plot

eg, year, month (Date_Added)

Q. Which is most popular year for netflix

Q. In which month, netflix adds most no. of movies.

2. Bivariate analysis → X

→ ←

Insights & Recommendations.