Python , SQL , Pandas , Numpy , Seaborn , Matplotlib.

Ritwik.

How to solve Netflix big case → Numpy , pandas

Solution oriented

trend for prices of an ecommerce company

tableau , SQL ←

Netflix Qiz

→ Explore

→ some insights, import observations

→ Recommendation

Concerns

✓ ① Missing value

✓ ② Duration

③ Date format

④ Nested data

Pre-processing

→ Cleaning

→ Sanitizing

Exploratory data analysis (EDA)

80%

Nested data

→ processing issues

Title	Cast	Director
ABC	M, N	D_1, D_2
3 idiots	Amir, Madhvan, Sharman Joshi	D_3
XYZ	N, M	D_1

Q Who's the most popular actor on Netflix platform

→ genre, country, Actor . . . .

M

Q which is the most popular Actor-director pair.

```
df.groupby(["cast","Director"])["Title"].nunique()
                                         .sort_value( )
```

Title	Cast_0	Cast_1	Cast_2
ABC	M	N	Null
3idi-	Anil	Madh	Sharma
XYZ	W	M	

① Split will not solve my problem

② Stack (unpivot)

(transpose)

Convert colums into rows

50%

X will not be able to solve

delete

Rename to cast

df

Title	Cat	Value	Directors	genre	Country
ABC	Cast_0	M	D1		
ABC	Cast_1	N	D2 D1		
3idicts	Cast_0	Amir	D2		
3idiots	Cast_1	Madh			
3idiots	Cast_2	Sharman			
XYZ					

df. groupby ("Cast") [Tittle] . cout ( )

nunique()

	Cast	Direct
ABC	M	D1
ABC	N	D1
ABC	M	D2
ABC	N	D2

~~M~~	~~2~~
~~N~~	~~2~~

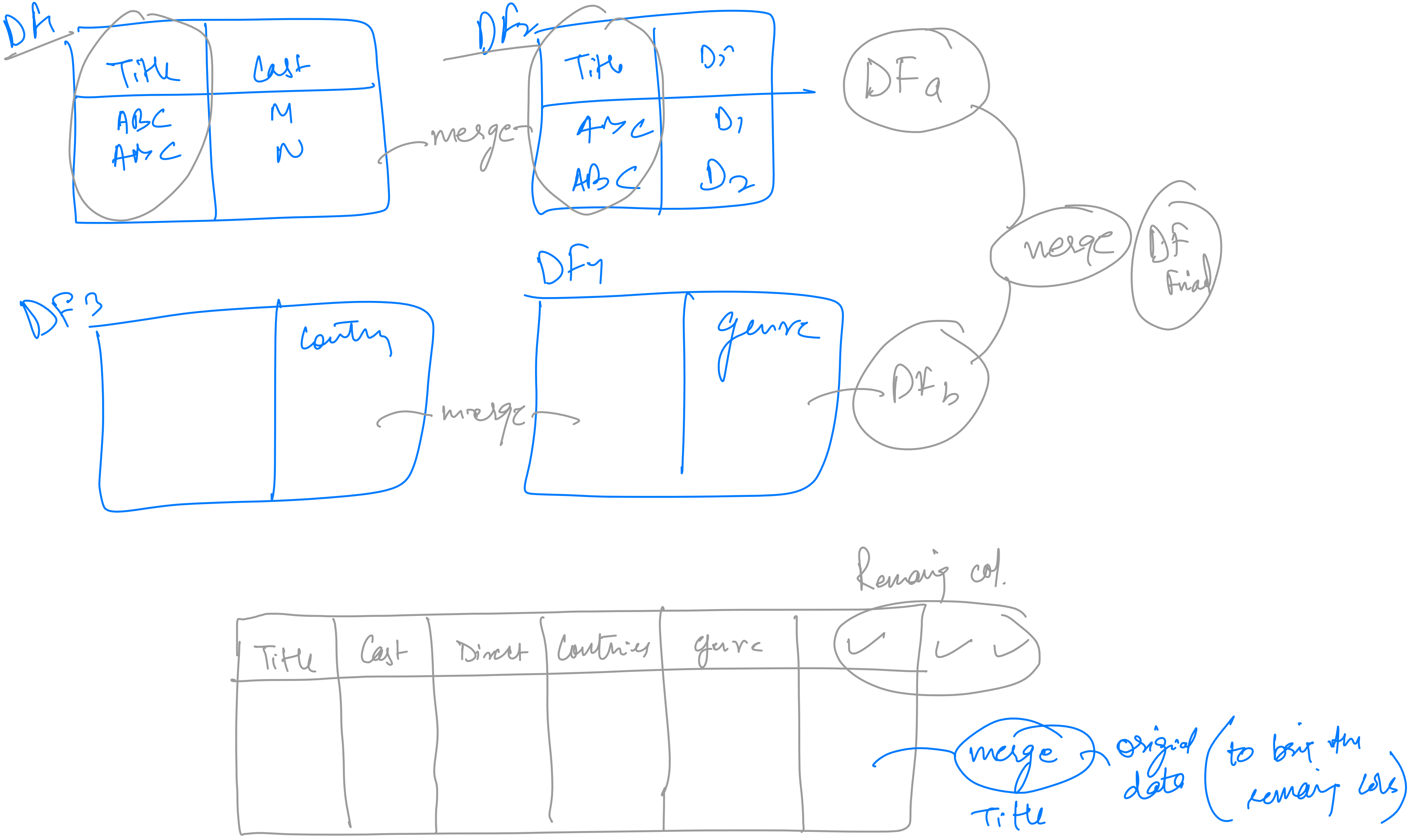
M	1
N	1

Unnest ① Split & Stack → Cast ⇒ df1

Unest ② → Director ⇒ df2

Unest ③ → Country ⇒ df3

④ → genre ⇒ df4


DF1
Title
Cast
ABC
AMC
M
N
merge
DF2
Title
Dir
AMC
ABC
D1
D2
DFa
merge
DF Final
DF3
Country
merge
DF4
genre
DFb
Remaing col.
Title
Cast
Direct
Countries
genre
merge
Title
original data
(to bring the remaing cols)

Summary:
step 1 : split into dataframes with titles into centre and other features as the specific dataframes
step 2: join these different dataframes using title

2 Missing values

1 Drop

2 Impute → mean/median * affected by outliers
→ mode

3 Drop the entire colm

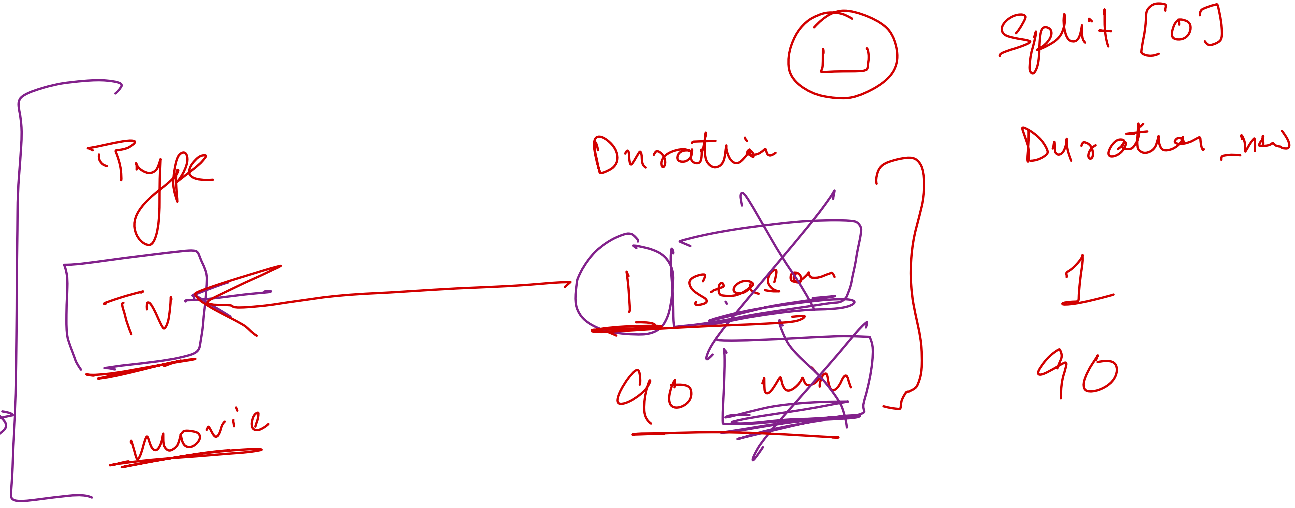
Medical

IQR

1.5 * IQR

3 * IQR

② Duration



→ Tell me the average rating of the movies in which SRK is present

→ get the average rating of all the actor & sort it in descending df. groupby ("Cast", "Type") [Duration] . average

mean ( ) . Sort valu

Anupm Kher	movie	100
Anupm Kher	TV	2

4 Date format

Pd. to_datetime (Date_added)

→ Extract year
month
day of the week }

Q: which is the most busy month for netflix

→ Increase server capacity before those months

Seaborn
matplotlib → Charts to visualize

Pick up the imp

Insight 2 give recomondatiās.

+ ipynb → pdf → upload to platform

Pls do come for review session

Cast ✓	
SRK ✓	✓
Salma	✓
SRK	✓
	✓

Numerical → Mean|Mean

Categorical → mode

SRK

```
x = df[cast].mode()
df[cast].fillna(x)
```

Simple imputation

Bonus

→ Multiple imputation

India

ajay.shenoy@scaler.com

+ Batch name

Module lead ←