

#### KAGGLE CHALLENGE

13/11/2023



## CONTENT

# kaggle

#### Variables:

- 28 explanatory variables
- 1 variable to be explained: 'averageRating'

Entrée [110]: origen.head()

Out[110]:

	Unnamed: 0	averageRating	numVotes	titleType	isAdult	startYear	endYear	runtimeMinutes	genres_x	din
0	0	4.4	15	movie	0.0	1951	0	91	Comedy, Musical	80mn
1	1	7.0	990	tvSeries	0.0	2007	2021	30	Action,Adventure,Animation	nm2291816,nm3088555,nm4930005,nm17
2	2	8.1	41	tvEpisode	0.0	2011	0	44	Documentary, History, War	nm04
3	3	4.6	48	movie	0.0	1969	0	84	Drama	nm29
4	4	5.6	28	movie	0.0	2010	0	130	Comedy,Drama	nm23

5 rows x 29 columns

Entrée [114]: origen.describe()

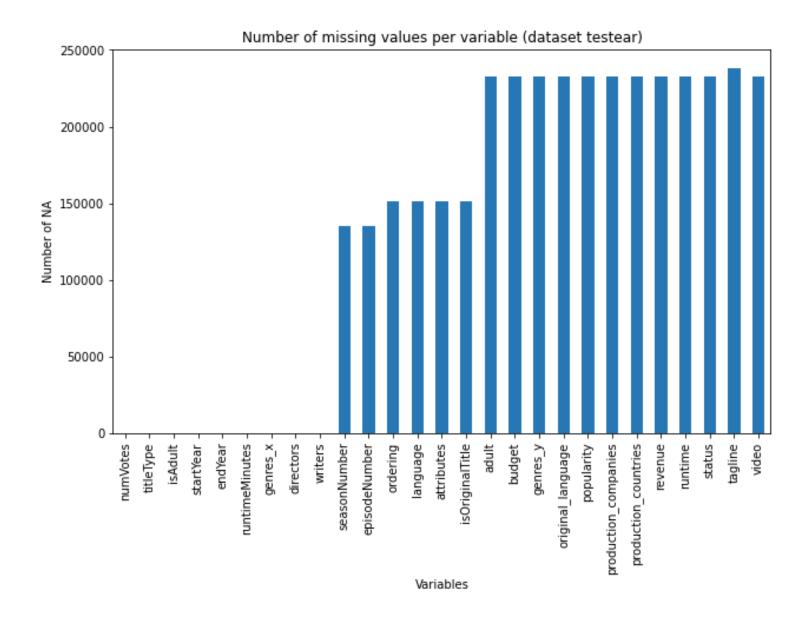
Out[114]:

	Unnamed: 0	averageRating	numVotes	isAdult	startYear	endYear	runtimeMinutes	seasonNumber	episodeNumber	orderi
cour	t 977541.000000	977541.000000	9.775410e+05	977541.000000	977541.000000	977541.000000	977541.000000	438243.000000	438243.000000	370623.0000
mea	n 488770.000000	6.881764	1.625621e+03	0.023017	1999.356151	58.196713	41.363622	4.061229	55.341327	3.4794
st	d 282191.924082	1.405315	2.509798e+04	2.888235	34.362292	336.455028	57.788808	12.336583	585.538414	5.1484
mi	n 0.000000	1.000000	5.000000e+00	0.000000	0.000000	0.000000	-22336.000000	0.000000	0.000000	1.0000
25%	6 244385.000000	6.100000	9.000000e+00	0.000000	1992.000000	0.000000	0.000000	1.000000	4.000000	1.0000
50%	488770.000000	7.100000	2.200000e+01	0.000000	2008.000000	0.000000	27.000000	2.000000	8.000000	2.0000
75%	6 733155.000000	7.900000	9.300000e+01	0.000000	2015.000000	0.000000	73.000000	4.000000	16.000000	3.0000
ma	x 977540.000000	10.000000	2.425542e+06	2020.000000	2021.000000	2022.000000	13319.000000	2012.000000	15762.000000	162.0000



#### Number of missing values per variable (dataset origen) 0.8 Number of NA 0.6 0.2 production\_countries -revenue -runtime language -attributes -isOriginalTitle startYear -endYear -runtimeMinutes -genres\_x budget -genres\_y status -tagline -video averageRating . popularity . writers ordering adult original\_language production\_companies episodeNumber seasonNumber Variables

### CLEANING





#### 017-0.059 0.1 0.0075-0.14-0.024-0.036-0.11-0.00260.049 0.1 0.1 0.15 0.00880.00510.014 0.075-0.00440.011 0.36 0.0002 0.6 0.56 0.71 0.16 isAdult -0.0590.0088 1 -0.015-0.022 0.08-0.00350.00130.0540.00090.00750.00970.00550.013 0.0077-0.051 0.018 0.015 -0.0870.000520.16 0.14 0.11 0.09 069 0.0024-0.012-0.012-0.016 0.32 endYear -0.0075 0.014 -0.0220.0077 - 0.6 runtimeMinutes - -0.14 0.075 0.08 -0.0510.0056 0.038-0.042 0.28 0.0019 0.16 0.14 0.13 seasonNumber -0.0240.00440.00350.018 1 -0.0140.004 0.12 0.12 episodeNumber -0.036-0.011-0.00130.015 .0097-0.11 - 0.4 ordering - -0.11 0.36 -0.054-0.087 0.069 0.28 -0.00440.012 1 0.0008 0.32 0.32 0.31 0.18 isOriginalTitle -0.0026.00028.0009.00052.00240.0019 1 -0.00040.00180.001-0.0002 - 0.2 budget -0.049 0.6 0.0075 0.16 -0.012 0.16 0.0810.0097 0.32 0.00047 1 0.49 0.79 0.16 0.1 0.56 0.0097 0.14 0.012 0.14 0.12 0.11 0.32 0.0018 0.49 0.56 0.15 - 0.0 0.1 0.71 -0.0055 0.11 -0.016 0.13 0.31 0.0011 0.79 0.56 0.044 -0.05 0.18-0.000260.16 0.15 0.14 runtime - 0.15 0.16 -0.011 0.098 0.32

### CLEANING

```
Variable 1
                                 Variable 2 Correlation
                   numVotes
                                    budget
                                               0.600867
                   numVotes
                                   revenue
                                               0.713653
             runtimeMinutes
                                               0.808981
                                   runtime
                     budget
                                   numVotes
                                               0.600867
                                               0.788960
                     budget
                                   revenue
                                               0.713653
                    revenue
                                   numVotes
                                               0.788960
                                    budget
                    revenue
                                              0.808981
                    runtime runtimeMinutes
Entrée [125]: for i in correlated_variables :
                 print("Missing ", i, " : ",origen[i].isnull().sum())
          Missing numVotes : 0
          Missing numVotes : 0
          Missing runtimeMinutes :
          Missing budget : 930169
          Missing budget: 930169
          Missing revenue : 930170
          Missing revenue: 930170
          Missing runtime: 930381
```



### MODELING

#### Choice of the Random Forest Regressor

- Why?
- What advantages ?
- What limits?

