ITBA
Instituto Tecnológico de Buenos Aires

KAGGLE CHALLENGE

13/11/2023

Thomas SIMON (65931)

82.05 - Análisis Predictivo - Secondo Parcial

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	nm08
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

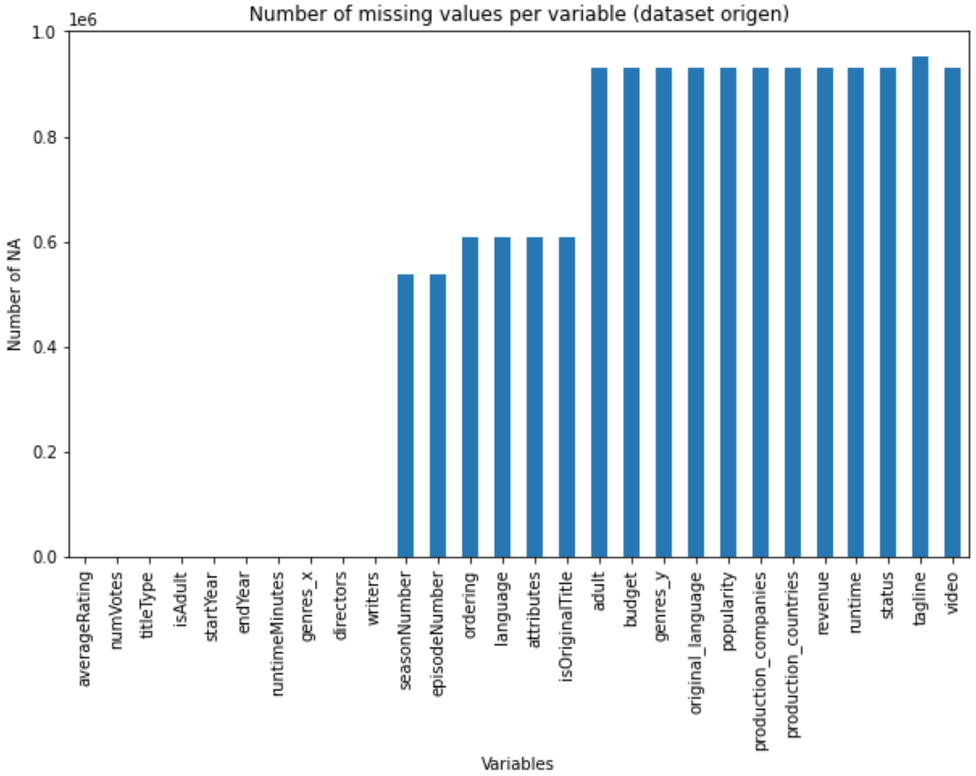
5 rows × 29 columns

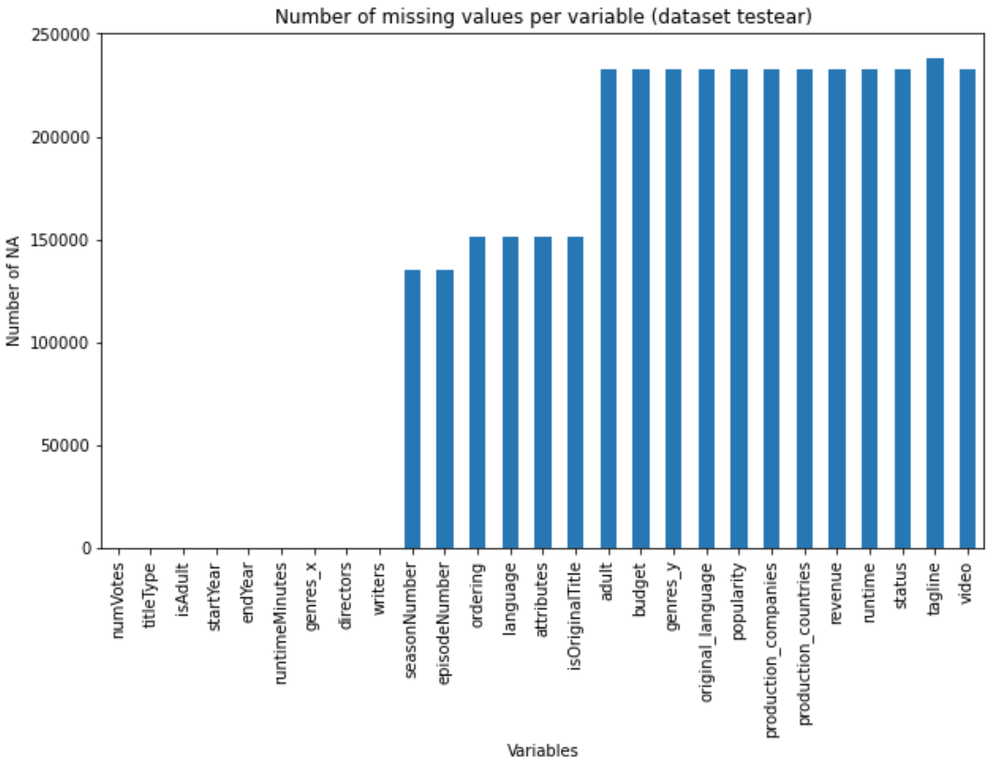
```
Entrée [114]: origen.describe()
```

Out[114]:

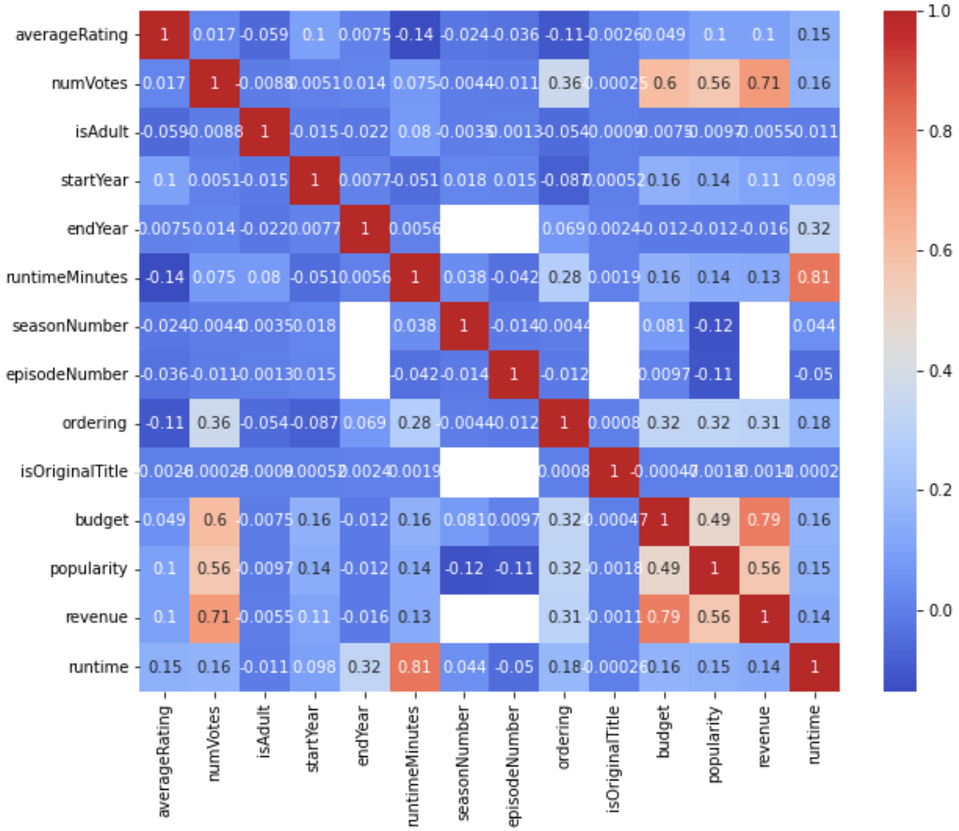
	Unnamed: 0	averageRating	numVotes	isAdult	startYear	endYear	runtimeMinutes	seasonNumber	episodeNumber	orderi
count	977541.000000	977541.000000	9.775410e+05	977541.000000	977541.000000	977541.000000	977541.000000	438243.000000	438243.000000	370623.0000
mean	488770.000000	6.881764	1.625621e+03	0.023017	1999.356151	58.196713	41.363622	4.061229	55.341327	3.4794
std	282191.924082	1.405315	2.509798e+04	2.888235	34.362292	336.455028	57.788808	12.336583	585.538414	5.1484
min	0.000000	1.000000	5.000000e+00	0.000000	0.000000	0.000000	-22336.000000	0.000000	0.000000	1.0000
25%	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%	733155.000000	7.900000	9.300000e+01	0.000000	2015.000000	0.000000	73.000000	4.000000	16.000000	3.0000
max	977540.000000	10.000000	2.425542e+06	2020.000000	2021.000000	2022.000000	13319.000000	2012.000000	15762.000000	162.0000

CLEANING


Number of missing values per variable (dataset origen)
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Number of NA
averageRating
numVotes
titleType
isAdult
startYear
endYear
runtimeMinutes
genres_x
directors
writers
seasonNumber
episodeNumber
ordering
language
attributes
isOriginalTitle
adult
budget
genres_y
original_language
popularity
production_companies
production_countries
revenue
runtime
status
tagline
video
Variables


Number of missing values per variable (dataset testear)
250000
200000
150000
100000
50000
0
Number of NA
numVotes
titleType
isAdult
startYear
endYear
runtimeMinutes
genres_x
directors
writers
seasonNumber
episodeNumber
ordering
language
attributes
isOriginalTitle
adult
budget
genres_y
original_language
popularity
production_companies
production_countries
revenue
runtime
status
tagline
video
Variables

CLEANING



	Variable 1	Variable 2	Correlation
0	numVotes	budget	0.600867
1	numVotes	revenue	0.713653
2	runtimeMinutes	runtime	0.808981
3	budget	numVotes	0.600867
4	budget	revenue	0.788960
5	revenue	numVotes	0.713653
6	revenue	budget	0.788960
7	runtime	runtimeMinutes	0.808981

Entrée [125]:
```
for i in correlated_variables :
    print("Missing ", i, " : ",origen[i].isnull().sum())
```

```
Missing  numVotes  :  0
Missing  numVotes  :  0
Missing  runtimeMinutes  :  0
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 ?