Overview

Dataset stat	tistics		Variable types		
Number of va	riables	5	Text		4
Number of ob	servations	500	Categorical		1
Missing cells		227			
Missing cells	(%)	9.1%			
Duplicate row	/s	0			
Duplicate row	/s (%)	0.0%			
Total size in n	nemory	237.8 KiB			
Average reco	rd size in memory	487.0 B			
Alerts	has 227 (45.4%) missing values			Missing	
_id has uniqu				Unique	
id has unique	e values			Unique	
Reproduction	on				
Analysis started	2024-02-03 05:47:18.401625				
Analysis finished	2024-02-03 05:47:18.782324				
Duration	0.38 seconds				
Software version	ydata-profiling vv4.6.2 (https://github.co	om/ydataai/ydata-profiling)			
Download configuration	config.json (data:text/plain;charset=utf- 8,%7B%22title%22%3A%20%22Pand		%2C%20%22dataset%22%3A%	.20%7B%22description%22%3A%20%2	22%22%2C%20

Variables

Select Columns ✓

_id Text

UNIQUE

61e57fd65c5338f43c777f4e

Distinct	500
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	39.7 KiB



ength		Character	s and Uni	icode	Unique			Sample	
lax length	24	Total	12000		Unique	500	?	1st row	61e57fd65c5338f43c777f4a
ledian length	24	characters			Unique (%)	100.0%		2nd row	61e57fd65c5338f43c777f4c
lean length	24	Distinct characters	16					3rd row	61e57fd65c5338f43c777f4e
lin length	24	Distinct	2				?	4th row	61e57fd65c5338f43c777f50
		categories	(https://en.	wikipedia.org/wiki/L	Inicode_character_pr	operty#General_	Category)	5th row	61e57fd65c5338f43c777f52
		Distinct scripts	2 (https://en.v	wikipedia.org/wiki/S	cript_(Unicode)#List_	_of_scripts_in_U	? nicode)		
		Distinct blocks	1 (https://er	n.wikipedia.org/wiki	/Unicode_block)		?		
		properties to	each code	ssigns character point, which can tual variables.					
Value			Count	Frequency (%)					
61e57fd65c5338f43c7	77f4a		1	0.2%					
61e57fd65c5338f43c7	77760		1	0.2%					

0.2%

Value	Count	Frequency (%)
61e57fd65c5338f43c777f50	1	0.2%
61e57fd65c5338f43c777f52	1	0.2%
61e57fd65c5338f43c777f54	1	0.2%
61e57fd65c5338f43c777f56	1	0.2%
61e57fd65c5338f43c777f58	1	0.2%
61e57fd65c5338f43c777f5a	1	0.2%
61e57fd65c5338f43c777f5c	1	0.2%
Other values (490)	490	98.0%



Most occurring characters

Value	Count	Frequency (%)
7	2672	22.3%
f	1371	11.4%
6	1340	11.2%
5	1027	8.6%
1	876	7.3%
е	858	7.1%
d	663	5.5%
3	593	4.9%
8	421	3.5%
С	405	3.4%
Other values (6)	1774	14.8%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	8186	68.2%
Lowercase Letter	3814	31.8%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
7	2672	32.6%
6	1340	16.4%
5	1027	12.5%
1	876	10.7%
3	593	7.2%
8	421	5.1%
2	346	4.2%
9	322	3.9%
0	307	3.8%
4	282	3.4%

Lowercase Letter

Value	Count	Frequency (%)
f	1371	35.9%
е	858	22.5%
d	663	17.4%
С	405	10.6%
а	327	8.6%
b	190	5.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	8186	68.2%
Latin	3814	31.8%

Most frequent character per script

Common

Value	Count	Frequency (%)
7	2672	32.6%
6	1340	16.4%
5	1027	12.5%
1	876	10.7%
3	593	7.2%
8	421	5.1%
2	346	4.2%
9	322	3.9%
0	307	3.8%
4	282	3.4%

Latin

Value	Count	Frequency (%)
f	1371	35.9%
е	858	22.5%
d	663	17.4%
С	405	10.6%
a	327	8.6%
b	190	5.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	12000	100.0%

Most frequent character per block

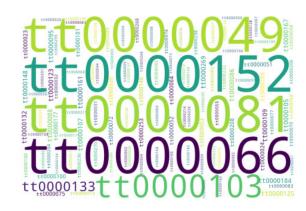
ASCII

Value	Count	Frequency (%)
7	2672	22.3%
f	1371	11.4%
6	1340	11.2%
5	1027	8.6%
1	876	7.3%
е	858	7.1%
d	663	5.5%
3	593	4.9%
8	421	3.5%
С	405	3.4%
Other values (6)	1774	14.8%

id Text

UNIQUE

Distinct	500
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	32.4 KiB



Length

Max length 9 Median length 9 Mean length 9 9 Min length

Characters and Unicode

blocks

Unique

Distinct	2	?
scripts	(https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)
Distinct	1 (https://en.wikinedia.org/wiki/Unicode_block)	?

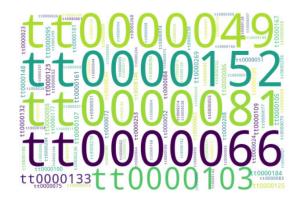
The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Value Count Frequency (%) tt0000081 0.2% 0.2% tt0000152 tt0000066 1 0.2% 0.2% tt0000049 tt0000103 1 0.2%

Sample

Total	4500	Unique	500	?	1st row	tt0000081
characters		Unique (%)	100.0%		2nd row	tt0000045
Distinct characters	11				3rd row	tt0000066
Distinct	2			?	4th row	tt0000049
categories	(https://en.wikipedia.org/wiki/Ur	nicode_character_prope	rty#General_0	Category)	5th row	tt0000103
Distinct	2			?		

Value	Count	Frequency (%)
tt0000133	1	0.2%
tt0000125	1	0.2%
tt0000184	1	0.2%
tt0000109	1	0.2%
tt0000086	1	0.2%
Other values (490)	490	98.0%



Most occurring characters

Value	Count	Frequency (%)
0	2204	49.0%
t	1000	22.2%
1	196	4.4%
2	188	4.2%
3	162	3.6%
4	146	3.2%
5	139	3.1%
6	136	3.0%
9	111	2.5%
8	109	2.4%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	3500	77.8%
Lowercase Letter	1000	22.2%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	2204	63.0%
1	196	5.6%
2	188	5.4%
3	162	4.6%
4	146	4.2%
5	139	4.0%
6	136	3.9%
9	111	3.2%
8	109	3.1%
7	109	3.1%

Lowercase Letter

Value	Count	Frequency (%)
t	1000	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	3500	77.8%
Latin	1000	22.2%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	2204	63.0%
1	196	5.6%
2	188	5.4%

Value	Count	Frequency (%)
3	162	4.6%
4	146	4.2%
5	139	4.0%
6	136	3.9%
9	111	3.2%
8	109	3.1%
7	109	3.1%

Latin

Value	Count	Frequency (%)
t	1000	100.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4500	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	2204	49.0%
t	1000	22.2%
1	196	4.4%
2	188	4.2%
3	162	3.6%
4	146	3.2%
5	139	3.1%
6	136	3.0%
9	111	2.5%

Value	Count	Frequency (%)
8	109	2.4%

titleText

Text

Distinct	492
Distinct (%)	98.4%
Missing	0
Missing (%)	0.0%
Memory size	69.6 KiB



row '__typename': 'TitleText'}

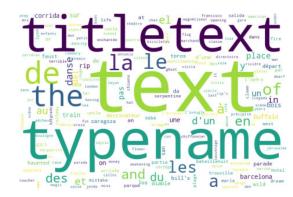
Length	
Max length	123
Median length	89
Mean length	60.658
Min length	44

Character	s and Unicode	Unique			Sam	ple
Total	30329	Unique	485 ? 1		1st	{'text': 'Les haleurs de bateaux',
characters		Unique (%)	Unique (%) 97.0%	row	'typename': 'TitleText'}	
Distinct characters	94				2nd row	{'text': 'Les blanchisseuses', 'typename': 'TitleText'}
Distinct categories	9 (https://en.wikipedia.org/w	/iki/Unicode_character_pr	operty#General_	? Category	0.0	{'text': 'Dessinateur: Von Bismark', 'typename':
Distinct	2			?		'TitleText'}
scripts	(https://en.wikipedia.org/w	/iki/Script_(Unicode)#List_	_of_scripts_in_U	nicode)	4th	{'text': 'Boxing Match; or, Glove
Distinct blocks	2 (https://en.wikipedia.org	/wiki/Unicode_block)		?	row	Contest', 'typename': 'TitleText'}
					5th	{'text': 'Plus fort que le maître',

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Value	Count	Frequency (%)
text	500	15.0%
typename	500	15.0%
titletext	500	15.0%
de	112	3.4%
the	108	3.2%

Value	Count	Frequency (%)
la	47	1.4%
le	44	1.3%
a	27	0.8%
les	25	0.7%
of	24	0.7%
Other values (1042)	1448	43.4%



Most occurring characters

Value	Count	Frequency (%)
•	3936	13.0%
e	3750	12.4%
t	2998	9.9%
	2835	9.3%
a	1364	4.5%
n	1111	3.7%
Т	1105	3.6%
i	1066	3.5%
х	1028	3.4%
:	1004	3.3%
Other values (84)	10132	33.4%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	17801	58.7%
Other Punctuation	5611	18.5%

Value	Count	Frequency (%)
Space Separator	2835	9.3%
Uppercase Letter	2054	6.8%
Connector Punctuation	1000	3.3%
Close Punctuation	503	1.7%
Open Punctuation	503	1.7%
Dash Punctuation	11	< 0.1%
Decimal Number	11	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	3750	21.1%
t	2998	16.8%
a	1364	7.7%
n	1111	6.2%
i	1066	6.0%
x	1028	5.8%
T	936	5.3%
p	655	3.7%
m	651	3.7%
r	632	3.6%
Other values (36)	3610	20.3%

Uppercase Letter

Value	Count	Frequency (%)
Т	1105	53.8%
L	141	6.9%
С	91	4.4%

Value	Count	Frequency (%)
В	80	3.9%
A	65	3.2%
S	64	3.1%
P	62	3.0%
D	60	2.9%
М	58	2.8%
R	49	2.4%
Other values (16)	279	13.6%

Other Punctuation

Value	Count	Frequency (%)
•	3936	70.1%
	1004	17.9%
,	515	9.2%
•	142	2.5%
	7	0.1%
1	4	0.1%
;	2	< 0.1%
?	1	< 0.1%

Decimal Number

Value	Count	Frequency (%)
1	4	36.4%
2	2	18.2%
8	1	9.1%
9	1	9.1%
0	1	9.1%
6	1	9.1%
5	1	9.1%

CI	ose	Pi	ın	ct	112	əti	in	r
		Γ	ai i	$\cup \iota$	uc		U	

Value	Count	Frequency (%)
}	500	99.4%
	3	0.6%

Open Punctuation

Value	Count	Frequency (%)
{	500	99.4%
	3	0.6%

Space Separator

Value	Count	Frequency (%)
	2835	100.0%

Connector Punctuation

Value	Count	Frequency (%)
	1000	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	11	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	19855	65.5%
Common	10474	34.5%

Most frequent character per script

Latin

Value	Count	Frequency (%)
е	3750	18.9%
t	2998	15.1%
а	1364	6.9%
n	1111	5.6%
Т	1105	5.6%
i	1066	5.4%
x	1028	5.2%
I	936	4.7%
p	655	3.3%
m	651	3.3%
Other values (62)	5191	26.1%

Common

Value	Count	Frequency (%)
•	3936	37.6%
	2835	27.1%
	1004	9.6%
-	1000	9.5%
,	515	4.9%
}	500	4.8%
{	500	4.8%
•	142	1.4%
-	11	0.1%
	7	0.1%
Other values (12)	24	0.2%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	30145	99.4%
None	184	0.6%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
•	3936	13.1%
е	3750	12.4%
t	2998	9.9%
	2835	9.4%
a	1364	4.5%
n	1111	3.7%
Т	1105	3.7%
i	1066	3.5%
х	1028	3.4%
:	1004	3.3%
Other values (64)	9948	33.0%

None

Value	Count	Frequency (%)
é	91	49.5%
à	16	8.7%
è	16	8.7%
í	14	7.6%
á	8	4.3%
ñ	5	2.7%
â	5	2.7%
ç	5	2.7%

Value	Count	Frequency (%)
ó	4	2.2%
ê	4	2.2%
Other values (10)	16	8.7%

releaseYear

Categorical

Distinct	20
Distinct (%)	4.0%
Missing	0
Missing (%)	0.0%
Memory size	56.3 KiB

Length	
Max length	58
Median length	58
Mean length	58
Min length	58

Character	s and Unicode	Unique			Sam	ple
Total	29000	Unique	2	?	1st	{'year': 1896, 'endYear': None,
characters		Unique (%)	0.4%		row	'typename': 'YearRange'}
Distinct	31	- 1 ()			2nd	{'year': 1896, 'endYear': None,
characters					row	'typename': 'YearRange'}
Distinct	8			?	3rd	{'year': 1896, 'endYear': None,
categories	(https://en.wikipedia.org/wiki/Ur	nicode_character_property	#General_	Category) row	'typename': 'YearRange'}
Distinct	2			?	4th	{'year': 1896, 'endYear': None,
scripts	(https://en.wikipedia.org/wiki/So	cript_(Unicode)#List_of_sc	ripts_in_Un	icode)	row	'typename': 'YearRange'}
Distinct	1 (https://en.wikipedia.org/wiki/	Unicode_block)		?	5th	{'year': 1896, 'endYear': None,
blocks					row	'typename': 'YearRange'}

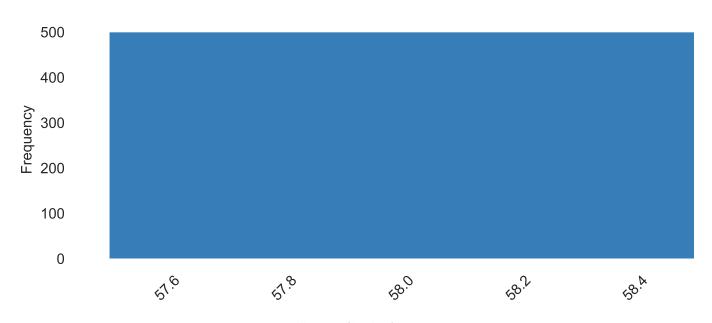
The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Common Values

Value	Count	Frequency (%)
{'year': 1896, 'endYear': None, 'typename': 'YearRange'}	101	20.2%
{'year': 1900, 'endYear': None, 'typename': 'YearRange'}	59	11.8%
{'year': 1898, 'endYear': None, 'typename': 'YearRange'}	39	7.8%
{'year': 1899, 'endYear': None, 'typename': 'YearRange'}	39	7.8%
{'year': 1909, 'endYear': None, 'typename': 'YearRange'}	38	7.6%
{'year': 1908, 'endYear': None, 'typename': 'YearRange'}	34	6.8%
{'year': 1897, 'endYear': None, 'typename': 'YearRange'}	33	6.6%
{'year': 1903, 'endYear': None, 'typename': 'YearRange'}	31	6.2%
{'year': 1901, 'endYear': None, 'typename': 'YearRange'}	23	4.6%

Value	Count	Frequency (%)
{'year': 1907, 'endYear': None, 'typename': 'YearRange'}	22	4.4%
Other values (10)	81	16.2%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
year	500	16.7%
endyear	500	16.7%
none	500	16.7%
typename	500	16.7%
yearrange	500	16.7%
1896	101	3.4%
1900	59	2.0%
1898	39	1.3%
1899	39	1.3%
1909	38	1.3%

Value	Count	Frequency (%)
Other values (15)	224	7.5%

Most occurring characters

Value	Count	Frequency (%)
	4000	13.8%
е	4000	13.8%
a	2500	8.6%
	2500	8.6%
n	2000	6.9%
r	1500	5.2%
:	1500	5.2%
Y	1000	3.4%
,	1000	3.4%
-	1000	3.4%
Other values (21)	8000	27.6%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	14000	48.3%
Other Punctuation	6500	22.4%
Space Separator	2500	8.6%
Uppercase Letter	2000	6.9%
Decimal Number	2000	6.9%
Connector Punctuation	1000	3.4%
Close Punctuation	500	1.7%
Open Punctuation	500	1.7%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
е	4000	28.6%
a	2500	17.9%
n	2000	14.3%
r	1500	10.7%
у	1000	7.1%
g	500	3.6%
t	500	3.6%
m	500	3.6%
p	500	3.6%
0	500	3.6%

Decimal Number

Value	Count	Frequency (%)
9	577	28.8%
1	526	26.3%
0	320	16.0%
8	311	15.6%
6	117	5.9%
7	55	2.8%
3	32	1.6%
5	30	1.5%
2	21	1.1%
4	11	0.5%

Other Punctuation

Valu	ue	Count	Frequency (%)
•		4000	61.5%

Uppercase Letter Value Count Fequency Y 100 50.0% N 50 50 Space Separator 7 Fequency Value Count Fequency Connector Punctuation 7 100.0% Value Count Fequency Losse Punctuation 100.0% 100.0% Copen Punctuation 100.0% 100.0% Oppen Punctuation 1 Fequency Value Count Fequency 1 50 100.0% Oppen Punctuation 1 Fequency Value 50 100.0% Most occurring scripts 5 Fequency Value 50 1 Fequency 4 50 1 Fequency 5 5 5 Fequency 6 5 5 5 6 5 5 5	Value	Count	Frequency (%)
Value Count Frequency Y 100 50.0% R 50 25.0% N 50 25.0% Space Separator Count Frequency Value Count Frequency Connector Functuation Total 100.0% Value Count Frequency Losse Punctuation Total Frequency Value Count Frequency Cooper Punctuation Total 100.0% Value Count Frequency 1 50 100.0% Walue Count Frequency 1 50 100.0% Count Frequency 100.0% Walue Count Frequency 1 50 100.0% Walue Count Frequency 2 100.0% 100.0% Walue Count Frequency 2 100.0% 100.0% Walue 100.	<u>:</u>	1500	23.1%
Value Country Speaker Superator Speaker Superator Feducation F	,	1000	15.4%
Y 100 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00 % 50.00	Uppercase Letter		
R 500 cm	Value	Count	Frequency (%)
N 500 december of the parameter of	Υ	1000	50.0%
Value Count Frequency Connector Punctuation 2500 100.0% Value Count Frequency Cilose Punctuation Value Count Frequency \(2\) Open Punctuation \(2\) Open Punctuation \(2\) Together Yellow \(2\) Together Yellow \(2\) Open Punctuation \(2\) Together Yellow \(2\) Together Yellow \(2\) Value \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow \(2\) World Yellow \(2\) Together Yellow \(2\) Together Yellow	R	500	25.0%
Value Count Frequency Connector Punctuation Yalue Count Frequency Ciose Punctuation Toul Frequency Value Count Frequency Yalue Count Frequency Value Count Frequency (50 100.0% Most occurring scripts Yalue Count Frequency Latin 1600 55.2%	N	500	25.0%
Connector Punctuation Frequency Value Count Frequency Close Punctuation Tequency Tequency Value Count Frequency } 500 100.0% Open Punctuation Tequency Tequency { 500 100.0% Most occurring scripts Tequency Tequency Value Count Frequency Latin 1600 55.2%	Space Separator		
Value Count Frequency Close Punctuation 100.0% Value Count Frequency } 500 100.0% Open Punctuation Count Frequency { 500 100.0% Most occurring scripts Count Frequency Latin 1600 55.2%	Value	Count	Frequency (%)
Value Country Frequency Close Punctuation Frequency Value Country Frequency Open Punctuation Country Frequency Value Country Frequency { 100.0% Wost occurring scripts Yelle Frequency Latin 100.0% 55.2%		2500	100.0%
Close Punctuation Yalue Count Frequency Yalue 500 100.0% Open Punctuation Frequency Yalue Count Frequency { 500 100.0% Most occurring scripts Yalue Count Frequency Latin 1600 55.2%	Connector Punctuation		
Value Count Punctuation Value Count Frequency Latin 1600 55.2%	Value	Count	Frequency (%
ValueCountFrequencyOpen PunctuationCountFrequencyValue500100.0%Most occurring scriptsCountFrequencyValueCountFrequencyLatin1600055.2%	_	1000	100.0%
ValueCountFrequencyOpen PunctuationCountFrequencyValue500100.0%Most occurring scriptsCountFrequencyValueCountFrequencyLatin1600055.2%	Close Punctuation		
Open Punctuation Count Frequency Value 500 100.0% Most occurring scripts Count Frequency Value Count Frequency Latin 16000 55.2%		Count	Frequency (%
ValueCountFrequency{500100.0%Most occurring scriptsValueCountFrequencyLatin1600055.2%			100.0%
Value Count Frequency Most occurring scripts Value Count Frequency Latin 1600 55.2%	On an Burnetunking		
Most occurring scripts Value Latin 500 100.0% Frequency 16000 555.2%		6	Eva
Most occurring scripts Value Latin Count Frequency 16000 55.2%			
ValueCountFrequencyLatin1600055.2%	1	500	100.0%
Value Count Frequency Latin 16000 55.2%	Most occurring scripts		
Latin 16000 55.2%		Count	Frequency (%)
DOMINON 1.3000 44 676	Common	13000	44.8%

Most frequent character per script

Common

Value	Count	Frequency (%)
•	4000	30.8%
	2500	19.2%
	1500	11.5%
,	1000	7.7%
_	1000	7.7%
9	577	4.4%
1	526	4.0%
}	500	3.8%
{	500	3.8%
0	320	2.5%
Other values (7)	577	4.4%
Latin		
Value	Count	Frequency (%)
Value e	Count 4000	Frequency (%) 25.0%
Value		
Value e	4000	25.0%
Value e a	4000 2500	25.0%
Value e a n	4000 2500 2000	25.0% 15.6% 12.5%
Value e a n r	4000 2500 2000 1500	25.0% 15.6% 12.5% 9.4%
e a n r	4000 2500 2000 1500 1000	25.0% 15.6% 12.5% 9.4% 6.2%
Value e a n r Y y	4000 2500 2000 1500 1000	25.0% 15.6% 12.5% 9.4% 6.2%
Value e a n r y g	4000 2500 2000 1500 1000 1000	25.0% 15.6% 12.5% 9.4% 6.2% 6.2% 3.1%
Value e a n r y y g t	4000 2500 2000 1500 1000 500 500	25.0% 15.6% 12.5% 9.4% 6.2% 6.2% 3.1% 3.1%

Most occurring blocks

Valu	ue	Count	Frequency (%)
ASC	CII	29000	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	4000	13.8%
е	4000	13.8%
a	2500	8.6%
	2500	8.6%
n	2000	6.9%
r	1500	5.2%
<u> </u>	1500	5.2%
Y	1000	3.4%
·	1000	3.4%
_	1000	3.4%
Other values (21)	8000	27.6%

releaseDate

Text

MISSING

Distinct	190
Distinct (%)	69.6%
Missing	227
Missing (%)	45.4%
Memory size	40.4 KiB



П	е	n	a	ŧ۱	h
_	c		ч	u	

Max length	71
Median length	68
Mean length	67.600733
Min length	65

Characters and Unicode

Character	s and Unicode	Unique			San
Total	18455	Unique	159	?	1st
characters		Unique (%)	58.2%		row
Distinct characters	33	,			2nd
Distinct categories	8 (https://en.wikipedia.org/	wiki/Unicode_character_pr	roperty#General_	? _Category	row
Distinct scripts	2 (https://en.wikipedia.org/	wiki/Script_(Unicode)#List	_of_scripts_in_U	? Inicode)	3rd row
Distinct blocks	1 (https://en.wikipedia.org	g/wiki/Unicode_block)		?	4th

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Sample

	Carrip	710
	1st row	{'day': None, 'month': None, 'year': 1896, 'typename': 'ReleaseDate'}
? Jory)	2nd row	{'day': None, 'month': 1, 'year': 1896, 'typename': 'ReleaseDate'}
?	3rd row	{'day': None, 'month': 5, 'year': 1899, 'typename': 'ReleaseDate'}
?	4th row	{'day': None, 'month': 9, 'year': 1896, 'typename': 'ReleaseDate'}
	5th row	{'day': None, 'month': None, 'year': 1896, 'typename': 'ReleaseDate'}

Value Count Frequency (%)

day	273	12.5%
month	273	12.5%
year	273	12.5%

Value	Count	Frequency (%)
typename	273	12.5%
releasedate	273	12.5%
none	177	8.1%
1896	61	2.8%
9	35	1.6%
1	35	1.6%
11	34	1.6%
Other values (49)	477	21.8%



Most occurring characters

Value	Count	Frequency (%)
•	2730	14.8%
е	2088	11.3%
	1911	10.4%
a	1365	7.4%
	1092	5.9%
	819	4.4%
у	819	4.4%
t	819	4.4%
n	723	3.9%
m	546	3.0%
Other values (23)	5543	30.0%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	8448	45.8%
Other Punctuation	4641	25.1%

Value	Count	Frequency (%)
Space Separator	1911	10.4%
Decimal Number	1640	8.9%
Uppercase Letter	723	3.9%
Connector Punctuation	546	3.0%
Close Punctuation	273	1.5%
Open Punctuation	273	1.5%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
е	2088	24.7%
a	1365	16.2%
у	819	9.7%
t	819	9.7%
n	723	8.6%
m	546	6.5%
0	450	5.3%
s	273	3.2%
I	273	3.2%
p	273	3.2%
Other values (3)	819	9.7%

Decimal Number

Value	Count	Frequency (%)
1	477	29.1%
9	363	22.1%
0	219	13.4%
8	193	11.8%

Value	Count	Frequency (%)
2	104	6.3%
6	95	5.8%
7	56	3.4%
5	55	3.4%
3	46	2.8%
4	32	2.0%
Other Punctuation		
Value	Count	Frequency (%)
•	2730	58.8%
	1092	23.5%
,	819	17.6%
Uppercase Letter		
Value	Count	Frequency (%)
D	273	37.8%
R	273	37.8%
N	177	24.5%
Space Separator		
Value	Count	Frequency (%)
	1911	100.0%
Connector Punctuation		
Value	Count	Frequency (%)
		400.00/
-	546	100.0%
	546	100.0%

273 100.0%

Open Punctuation

Value	Count	Frequency (%)
{	273	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	9284	50.3%
Latin	9171	49.7%

Most frequent character per script

Common

Value	Count	Frequency (%)
•	2730	29.4%
	1911	20.6%
	1092	11.8%
,	819	8.8%
-	546	5.9%
1	477	5.1%
9	363	3.9%
}	273	2.9%
{	273	2.9%
0	219	2.4%
Other values (7)	581	6.3%

Latin

Value	Count	Frequency (%)
е	2088	22.8%
a	1365	14.9%
у	819	8.9%

Value	Count	Frequency (%)
t	819	8.9%
n	723	7.9%
m	546	6.0%
0	450	4.9%
D	273	3.0%
s	273	3.0%
1	273	3.0%
Other values (6)	1542	16.8%

Most occurring blocks

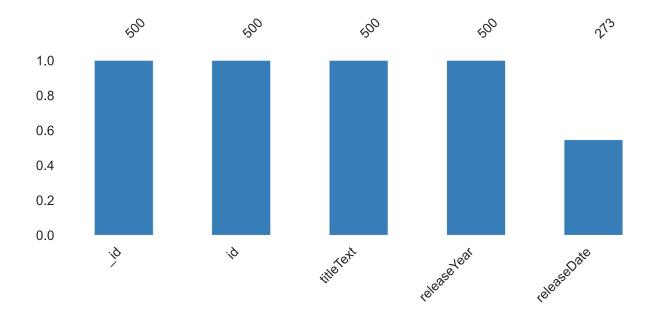
Value	Count	Frequency (%)
ASCII	18455	100.0%

Most frequent character per block

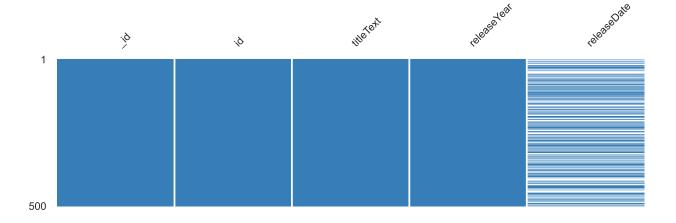
ASCII

Value	Count	Frequency (%)
•	2730	14.8%
е	2088	11.3%
	1911	10.4%
a	1365	7.4%
	1092	5.9%
,	819	4.4%
у	819	4.4%
t	819	4.4%
n	723	3.9%
m	546	3.0%
Other values (23)	5543	30.0%

Missing values



A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.

Sample

	_id	id	titleText	releaseYear
0	61e57fd65c5338f43c777f4a	tt0000081	{'text': 'Les haleurs de bateaux', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
1	61e57fd65c5338f43c777f4c	tt0000045	{'text': 'Les blanchisseuses', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
2	61e57fd65c5338f43c777f4e	tt0000066	{'text': 'Dessinateur: Von Bismark', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
3	61e57fd65c5338f43c777f50	tt0000049	{'text': 'Boxing Match; or, Glove Contest', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
4	61e57fd65c5338f43c777f52	tt0000103	{'text': 'Plus fort que le maître', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
5	61e57fd65c5338f43c777f54	tt0000133	{'text': 'La voiture du potier', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
6	61e57fd65c5338f43c777f56	tt0000125	{'text': 'The Terrible Railway Accident', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
7	61e57fd65c5338f43c777f58	tt0000184	{'text': 'Cripple Creek Bar-Room Scene', 'typename': 'TitleText'}	{'year': 1899, 'endYear': None, 'typename': 'YearRa
8	61e57fd65c5338f43c777f5a	tt0000109	{'text': 'Rip Meeting the Dwarf', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa
9	61e57fd65c5338f43c777f5c	tt0000086	{'text': 'Jetée et plage de Trouville (1er partie)', 'typename': 'TitleText'}	{'year': 1896, 'endYear': None, 'typename': 'YearRa

	_id	id	titleText	releaseYear
490	61e57fd7d8a2362a33777f84	tt0000469	{'text': 'The Sick Kitten', 'typename': 'TitleText'}	{'year': 1903, 'endYear': None, 'typename': 'YearRanç
491	61e57fd7d8a2362a33777f86	tt0000248	{'text': 'The Kiss in the Tunnel', 'typename': 'TitleText'}	('year': 1899, 'endYear': None, 'typename': 'YearRanç
492	61e57fd7d8a2362a33777f88	tt0000526	$\label{thm:constraint} \mbox{\ensuremath{$($'$} text': 'Pennsylvania Tunnel Excavation', '_typename': 'TitleText'} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$}} \mbox{\ensuremath{$($$$}} \mbox{\ensuremath{$($$$$}} \mbox{\ensuremath{$($$$}} \mbox{\ensuremath{$($$$$}} \ensuremath{$($$	('year': 1905, 'endYear': None, 'typename': 'YearRanç
493	61e57fd7d8a2362a33777f8a	tt0000218	{'text': "L'affaire Dreyfus", 'typename': 'TitleText'}	('year': 1899, 'endYear': None, 'typename': 'YearRanç
494	61e57fd7d8a2362a33777f8c	tt0000116	{'text': 'Sac au dos', 'typename': 'TitleText'}	('year': 1896, 'endYear': None, 'typename': 'YearRanç
495	61e57fd7d8a2362a33777f8e	tt0000234	{'text': 'Cléopâtre', 'typename': 'TitleText'}	('year': 1899, 'endYear': None, 'typename': 'YearRanç
496	61e57fd7d8a2362a33777f90	tt0000276	{'text': "Bataille d'oreillers", 'typename': 'TitleText'}	('year': 1900, 'endYear': None, 'typename': 'YearRanç
497	61e57fd7d8a2362a33777f92	tt0000566	{'text': 'Professorens Morgenavis', 'typename': 'TitleText'}	('year': 1906, 'endYear': None, 'typename': 'YearRanç
498	61e57fd7d8a2362a33777f94	tt0000328	{'text': 'La petite magicienne', 'typename': 'TitleText'}	{'year': 1900, 'endYear': None, 'typename': 'YearRanç
499	61e57fd7d8a2362a33777f96	tt0000488	{'text': 'The Land Beyond the Sunset', 'typename': 'TitleText'}	{'year': 1912, 'endYear': None, 'typename': 'YearRanç

Report generated by YData (https://ydata.ai/?utm_source=opensource&utm_medium=pandasprofiling&utm_campaign=report).