Overview

Dataset statistics

Number of variables	24
Number of observations	45471
Missing cells	25263
Missing cells (%)	2.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	8.3 MiB
Average record size in memory	192.0 B

Variable types

Unsupported	8
Categorical	14
Numeric	2

Alerts

`Genres` has a high cardinality: 4068 distinct values	High cardinality
`OriginalLanguage` has a high cardinality: 93 distinct values	High cardinality
`Overview` has a high cardinality: 44234 distinct values	High cardinality
`ProductionCompanies` has a high cardinality: 22667 distinct values	High cardinality
`ProductionCountries` has a high cardinality: 2390 distinct values	High cardinality
`ReleaseDate` has a high cardinality: 17334 distinct values	High cardinality
`SpokenLanguages` has a high cardinality: 1932 distinct values	High cardinality
`Tagline` has a high cardinality: 20269 distinct values	High cardinality
`Title` has a high cardinality: 42197 distinct values	High cardinality
`MovieCharacter` has a high cardinality: 40180 distinct values	High cardinality
`ActorName` has a high cardinality: 42678 distinct values	High cardinality
`Department` has a high cardinality: 23604 distinct values	High cardinality
`CrewJob` has a high cardinality: 26602 distinct values	High cardinality
`CrewName` has a high cardinality: 42945 distinct values	High cardinality
`OriginalLanguage` is highly imbalanced (67.4%)	Imbalance
`ProductionCountries` is highly imbalanced (58.3%)	Imbalance
`SpokenLanguages` is highly imbalanced (61.2%)	Imbalance
`Tagline` has 25073 (55.1%) missing values	Missing
`Tagline` is uniformly distributed	Uniform
`Title` is uniformly distributed	Uniform
`Budget` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`Popularity` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`Revenue` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`Runtime` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`VoteAverage` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`VoteCount` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`ReleaseYear` is an unsupported type, check if it needs cleaning or further analysis	Unsupported
`Return` is an unsupported type, check if it needs cleaning or further analysis	Unsupported

Reproduction

Analysis started	2023-06-09 06:55:24.462785
Analysis finished	2023-06-09 06:56:35.859227
Duration	1 minute and 11.4 seconds
Software version	pandas-profiling v3.6.6 (https://github.com/pandas-profiling/pandas-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Pandas%20Profiling%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2C%20%22c

Variables

Select Columns

~~Budget~~
~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Genres
Categorical

Distinct	4068
Distinct (%)	8.9%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	84
Median length	68
Mean length	15.951552
Min length	3

Characters and Unicode

Total characters	725333
Distinct characters	41
Distinct categories	4 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_bloc

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

Unique

Unique	2367 ?
Unique (%)	5.2%

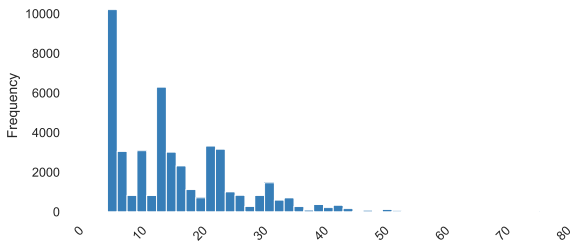
Sample

1st row	Animation, Comedy, Family
2nd row	Adventure, Fantasy, Family
3rd row	Romance, Comedy
4th row	Comedy, Drama, Romance
5th row	Comedy

Common Values

Value	Count	Frequency (%)
Drama	4998	11.0%
Comedy	3621	8.0%
Documentary	2713	6.0%
NoGenre	2476	5.4%
Drama, Romance	1301	2.9%
Comedy, Drama	1135	2.5%
Horror	974	2.1%
Comedy, Romance	930	2.0%
Comedy, Drama, Romance	593	1.3%
Drama, Comedy	532	1.2%
Other values (4058)	26198	57.6%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
drama	20255	20.8%
comedy	13181	13.5%
thriller	7619	7.8%
romance	6733	6.9%
action	6592	6.8%
horror	4670	4.8%
crime	4305	4.4%
documentary	3921	4.0%
adventure	3494	3.6%
science	3042	3.1%
Other values (37)	23535	24.2%

Most occurring characters

Value	Count	Frequency (%)
r	71558	9.9%
a	61822	8.5%
e	60738	8.4%
m	53101	7.3%
	51876	7.2%

Value	Count	Frequency (%)
o	51017	7.0%
,	48053	6.6%
i	39670	5.5%
n	38152	5.3%
y	28510	3.9%
Other values (31)	220836	30.4%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	524808	72.4%
Uppercase Letter	100596	13.9%
Space Separator	51876	7.2%
Other Punctuation	48053	6.6%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
r	71558	13.6%
a	61822	11.8%
e	60738	11.6%
m	53101	10.1%
o	51017	9.7%
i	39670	7.6%
n	38152	7.3%
y	28510	5.4%
c	27977	5.3%
t	26210	5.0%
Other values (12)	66053	12.6%

Uppercase Letter

Value	Count	Frequency (%)
D	24176	24.0%
C	17489	17.4%
A	12020	11.9%
F	9746	9.7%
T	8389	8.3%
R	6735	6.7%
H	6068	6.0%
M	4830	4.8%
S	3046	3.0%
G	2478	2.5%
Other values (7)	5619	5.6%

Space Separator

Value	Count	Frequency (%)
	51876	100.0%

Other Punctuation

Value	Count	Frequency (%)
,	48053	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	625404	86.2%
Common	99929	13.8%

Most frequent character per script

Latin

Value	Count	Frequency (%)
r	71558	11.4%
a	61822	9.9%
e	60738	9.7%
m	53101	8.5%
o	51017	8.2%
i	39670	6.3%
n	38152	6.1%
y	28510	4.6%
c	27977	4.5%
t	26210	4.2%
Other values (29)	166649	26.6%

Common

Value	Count	Frequency (%)
	51876	51.9%
,	48053	48.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	725333	100.0%

Most frequent character per block

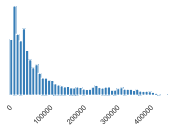
ASCII

Value	Count	Frequency (%)
r	71558	9.9%
a	61822	8.5%
e	60738	8.4%
m	53101	7.3%
	51876	7.2%
o	51017	7.0%
,	48053	6.6%
i	39670	5.5%
n	38152	5.3%
y	28510	3.9%
Other values (31)	220836	30.4%

Id

Real number (ℝ)

Distinct	45346	**Minimum**	2
Distinct (%)	99.9%	**Maximum**	469172
Missing	95	**Zeros**	0
Missing (%)	0.2%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	108027.1	**Memory size**	355.4 KiB

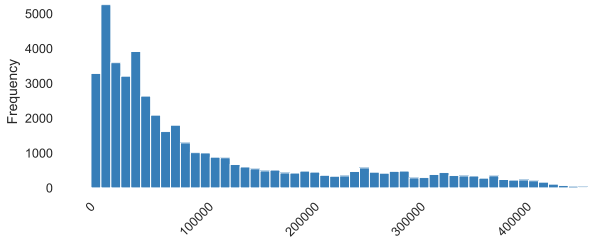


Quantile statistics

Minimum	2
5-th percentile	5348.75
Q1	26385.75
median	59857.5
Q3	156533.5
95-th percentile	357194.5
Maximum	469172
Range	469170
Interquartile range (IQR)	130147.75

Descriptive statistics

Standard deviation	112168.38
Coefficient of variation (CV)	1.0383355
Kurtosis	0.55951556
Mean	108027.1
Median Absolute Deviation (MAD)	44418.5
Skewness	1.2830689
Sum	4.9018378×10^9
Variance	1.2581745×10^{10}
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
141971	3	< 0.1%
159849	2	< 0.1%
168538	2	< 0.1%
109962	2	< 0.1%
97995	2	< 0.1%
119916	2	< 0.1%
132641	2	< 0.1%
84198	2	< 0.1%
10991	2	< 0.1%
18440	2	< 0.1%
Other values (45336)	45355	99.7%
(Missing)	95	0.2%

Value	Count	Frequency (%)
2	1	< 0.1%
3	1	< 0.1%
5	1	< 0.1%
6	1	< 0.1%
11	1	< 0.1%
12	1	< 0.1%
13	1	< 0.1%
14	1	< 0.1%
15	1	< 0.1%
16	1	< 0.1%

Value	Count	Frequency (%)
469172	1	< 0.1%
468707	1	< 0.1%
468343	1	< 0.1%
467731	1	< 0.1%

Value	Count	Frequency (%)
465044	1	< 0.1%
464819	1	< 0.1%
464207	1	< 0.1%
464111	1	< 0.1%
463906	1	< 0.1%
463800	1	< 0.1%

OriginalLanguage

Categorical

HIGH CARDINALITY IMBALANCE

Distinct	93
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	10
Median length	2
Mean length	2.0182754
Min length	2

Characters and Unicode

Total characters	91773
Distinct characters	35
Distinct categories	4 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)

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

Unique

Unique	20
Unique (%)	< 0.1%

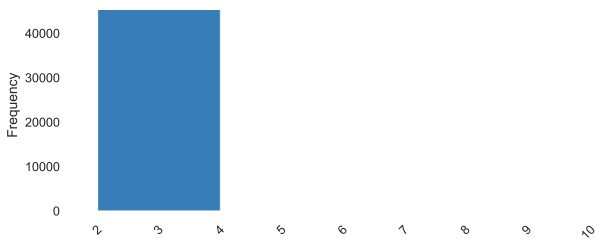
Sample

1st row	en
2nd row	en
3rd row	en
4th row	en
5th row	en

Common Values

Value	Count	Frequency (%)
en	32202	70.8%
fr	2437	5.4%
it	1528	3.4%
ja	1349	3.0%
de	1078	2.4%
es	992	2.2%
ru	822	1.8%
hi	508	1.1%
ko	444	1.0%
zh	408	0.9%
Other values (83)	3703	8.1%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
en	32202	70.8%
fr	2437	5.4%
it	1528	3.4%
ja	1349	3.0%
de	1078	2.4%
es	992	2.2%
ru	822	1.8%
hi	508	1.1%
ko	444	1.0%
zh	408	0.9%
Other values (83)	3703	8.1%

Most occurring characters

Value	Count	Frequency (%)
e	34630	37.7%

Value	Count	Frequency (%)
n	33013	36.1%
r	3630	4.0%
f	2835	3.1%
i	2388	2.6%
t	2250	2.5%
a	2045	2.2%
s	1652	1.8%
j	1350	1.5%
d	1323	1.4%
Other values (25)	6657	7.3%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	91554	99.8%
Uppercase Letter	206	0.2%
Decimal Number	10	< 0.1%
Other Punctuation	3	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	34630	37.8%
n	33013	36.1%
r	3630	4.0%
f	2835	3.1%
i	2388	2.6%
t	2250	2.5%
a	2045	2.2%
s	1652	1.8%
j	1350	1.5%
d	1323	1.4%
Other values (16)	6438	7.0%

Decimal Number

Value	Count	Frequency (%)
0	4	40.0%
8	2	20.0%
2	1	10.0%
6	1	10.0%
1	1	10.0%
4	1	10.0%

Uppercase Letter

Value	Count	Frequency (%)
N	103	50.0%
L	103	50.0%

Other Punctuation

Value	Count	Frequency (%)
.	3	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	91760	> 99.9%
Common	13	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	34630	37.7%

Value	Count	Frequency (%)
n	33013	36.0%
r	3630	4.0%
f	2835	3.1%
i	2388	2.6%
t	2250	2.5%
a	2045	2.2%
s	1652	1.8%
j	1350	1.5%
d	1323	1.4%
Other values (18)	6644	7.2%

Common

Value	Count	Frequency (%)
0	4	30.8%
.	3	23.1%
8	2	15.4%
2	1	7.7%
6	1	7.7%
1	1	7.7%
4	1	7.7%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	91773	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	34630	37.7%
n	33013	36.0%
r	3630	4.0%
f	2835	3.1%
i	2388	2.6%
t	2250	2.5%
a	2045	2.2%
s	1652	1.8%
j	1350	1.5%
d	1323	1.4%
Other values (25)	6657	7.3%

Overview

Categorical

Distinct	44234
Distinct (%)	97.3%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	1000
Median length	791
Mean length	316.15885
Min length	1

Characters and Unicode

Total characters	14376059
Distinct characters	429
Distinct categories	25 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	13 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	21 (https://en.wikipedia.org/wiki/Unicode_blo

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

Unique

Unique	44173
Unique (%)	97.1%

Sample

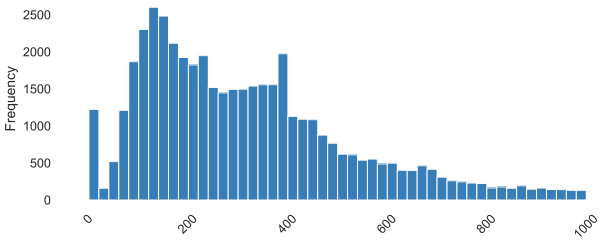
1st row	Led by Woody, Andy's toys live happily in his room until Andy's birthday brings Buzz Lightyear onto the scene. Afraid of losing his place in Andy's heart, Woody plots against Buzz. But when circumstances separate Buzz and Woody from their owner, the duo eventually learns to put aside their differences.	? ategory) ? code) ?
2nd row	When siblings Judy and Peter discover an enchanted board game that opens the door to a magical world, they unwittingly invite Alan -- an adult who's been trapped inside the game for 26 years -- into their living room. Alan's only hope for freedom is to finish the game, which proves risky as all three find themselves running from giant rhinoceroses, evil monkeys and other terrifying creatures.	
3rd row	A family wedding reignites the ancient feud between next-door neighbors and fishing buddies John and Max. Meanwhile, a sultry Italian divorcée opens a restaurant at the local bait shop, alarming the locals who worry she'll scare the fish away. But she's less interested in seafood than she is in cooking up a hot time with Max.	
4th row	Cheated on, mistreated and stepped on, the women are holding their breath, waiting for the elusive "good man" to break a string of less-than-stellar lovers. Friends and confidants Vannah, Bernie, Glo and Robin talk it all out, determined to find a better way to breathe.	

5th row	Just when George Banks has recovered from his daughter's wedding, he receives the news that she's pregnant ... and that George's wife, Nina, is expecting too. He was planning on selling their home, but that's a plan that -- like George -- will have to change with the arrival of both a grandchild and a kid of his own.

Common Values

Value	Count	Frequency (%)
NoOverview	1033	2.3%
No overview found.	133	0.3%
No Overview	7	< 0.1%
	5	< 0.1%
Released	3	< 0.1%
Recovering from a nail gun shot to the head and 13 months of coma, docto…	3	< 0.1%
King Lear, old and tired, divides his kingdom among his daughters, giving …	3	< 0.1%
No movie overview available.	3	< 0.1%
Adaptation of the Jane Austen novel.	3	< 0.1%
A few funny little novels about different aspects of life.	3	< 0.1%
Other values (44224)	44275	97.4%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
the	138082	5.6%
a	98889	4.0%
and	75259	3.1%
to	73321	3.0%
of	69574	2.8%
in	48143	2.0%
is	36500	1.5%
his	36165	1.5%
with	23902	1.0%
her	21484	0.9%
Other values (97092)	1828425	74.6%

Most occurring characters

Value	Count	Frequency (%)
	2406350	16.7%
e	1365862	9.5%
a	940505	6.5%
t	934766	6.5%
i	852547	5.9%
o	830906	5.8%
n	822601	5.7%
s	767854	5.3%
r	745307	5.2%

Value	Count	Frequency (%)
h	600810	4.2%
Other values (419)	4108551	28.6%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	11158346	77.6%
Space Separator	2406388	16.7%
Uppercase Letter	393031	2.7%
Other Punctuation	312824	2.2%
Decimal Number	42223	0.3%
Dash Punctuation	36767	0.3%
Close Punctuation	10100	0.1%
Open Punctuation	10077	0.1%
Final Punctuation	4556	< 0.1%
Initial Punctuation	882	< 0.1%
Other values (15)	865	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	1365862	12.2%
a	940505	8.4%
t	934766	8.4%
i	852547	7.6%
o	830906	7.4%
n	822601	7.4%
s	767854	6.9%
r	745307	6.7%
h	600810	5.4%
l	478816	4.3%
Other values (142)	2818372	25.3%

Uppercase Letter

Value	Count	Frequency (%)
A	42751	10.9%
T	35968	9.2%
S	31126	7.9%
M	23954	6.1%
B	23699	6.0%
C	22803	5.8%
H	19429	4.9%
W	18652	4.7%
I	16798	4.3%
D	16311	4.2%
Other values (77)	141540	36.0%

Other Letter

Value	Count	Frequency (%)
न	6	4.8%
र	6	4.8%
म	5	4.0%
ળ	4	3.2%
द	3	2.4%
प	3	2.4%
ಠ	3	2.4%
अ	3	2.4%
व	2	1.6%
۽	2	1.6%
Other values (76)	88	70.4%

Other Punctuation

Value	Count	Frequency (%)
,	133443	42.7%
.	124794	39.9%
'	31121	9.9%
"	11661	3.7%
:	3299	1.1%
?	2759	0.9%
;	2493	0.8%
!	1543	0.5%
/	765	0.2%
&	453	0.1%
Other values (12)	493	0.2%

Nonspacing Mark

Value	Count	Frequency (%)
◌́	4	12.1%
◌ি	4	12.1%
◌ং	3	9.1%
◌ী	3	9.1%
◌ু	3	9.1%
◌ঁ	3	9.1%
◌া	2	6.1%
◌ে	2	6.1%
◌ং	2	6.1%
◌ু	2	6.1%
Other values (4)	5	15.2%

Decimal Number

Value	Count	Frequency (%)
1	9748	23.1%
0	8265	19.6%
9	6405	15.2%
2	4251	10.1%
5	2440	5.8%
8	2379	5.6%
3	2342	5.5%
4	2176	5.2%
7	2131	5.0%
6	2086	4.9%

Spacing Mark

Value	Count	Frequency (%)
◌া	11	40.7%
◌ী	4	14.8%
◌ো	3	11.1%
◌ৗ	3	11.1%
◌ি	2	7.4%
◌্র	2	7.4%
◌ং	1	3.7%
◌ী	1	3.7%

Dash Punctuation

Value	Count	Frequency (%)
-	35244	95.9%
–	881	2.4%
—	633	1.7%
—	5	< 0.1%
-	4	< 0.1%

Other Symbol

Value	Count	Frequency (%)
®	45	70.3%

Value	Count	Frequency (%)
™	14	21.9%
¦	2	3.1%
°	2	3.1%
	1	1.6%

Math Symbol

Value	Count	Frequency (%)
~	20	50.0%
+	11	27.5%
=	6	15.0%
|	2	5.0%
−	1	2.5%

Open Punctuation

Value	Count	Frequency (%)
(	10024	99.5%
[	50	0.5%
{	2	< 0.1%
"	1	< 0.1%

Currency Symbol

Value	Count	Frequency (%)
$	317	96.4%
£	10	3.0%
₹	1	0.3%
€	1	0.3%

Space Separator

Value	Count	Frequency (%)
	2406350	> 99.9%
	36	< 0.1%
	2	< 0.1%

Close Punctuation

Value	Count	Frequency (%)
)	10048	99.5%
]	50	0.5%
}	2	< 0.1%

Final Punctuation

Value	Count	Frequency (%)
’	3847	84.4%
”	690	15.1%
»	19	0.4%

Initial Punctuation

Value	Count	Frequency (%)
“	672	76.2%
‘	192	21.8%
«	18	2.0%

Control

Value	Count	Frequency (%)
	106	96.4%
	3	2.7%
	1	0.9%

Modifier Symbol

Value	Count	Frequency (%)
´	25	65.8%
`	12	31.6%
¯	1	2.6%

Format

Value	Count	Frequency (%)
	31	60.8%
	20	39.2%

Other Number

Value	Count	Frequency (%)
½	8	50.0%
¹	8	50.0%

Connector Punctuation

Value	Count	Frequency (%)
_	19	100.0%

Line Separator

Value	Count	Frequency (%)
[illegible]	7	100.0%

Letter Number

Value	Count	Frequency (%)
Ⅱ	2	100.0%

Paragraph Separator

Value	Count	Frequency (%)
[illegible]	2	100.0%

Modifier Letter

Value	Count	Frequency (%)
ʼ	2	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	11546145	80.3%
Common	2824495	19.6%
Cyrillic	4587	< 0.1%
Greek	648	< 0.1%
Devanagari	77	< 0.1%
Telugu	30	< 0.1%
Hiragana	20	< 0.1%
Tamil	19	< 0.1%
Han	10	< 0.1%
Hangul	9	< 0.1%
Other values (3)	19	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	1365862	11.8%
a	940505	8.1%
t	934766	8.1%
i	852547	7.4%
o	830906	7.2%
n	822601	7.1%
s	767854	6.7%
r	745307	6.5%
h	600810	5.2%
l	478816	4.1%
Other values (132)	3206171	27.8%

Common

Value	Count	Frequency (%)
	2406350	85.2%

Value	Count	Frequency (%)
,	133443	4.7%
.	124794	4.4%
-	35244	1.2%
'	31121	1.1%
"	11661	0.4%
)	10048	0.4%
(	10024	0.4%
1	9748	0.3%
0	8265	0.3%
Other values (71)	43797	1.6%

Cyrillic

Value	Count	Frequency (%)
о	470	10.2%
е	404	8.8%
а	373	8.1%
н	323	7.0%
и	299	6.5%
т	265	5.8%
р	240	5.2%
с	218	4.8%
в	173	3.8%
л	161	3.5%
Other values (46)	1661	36.2%

Greek

Value	Count	Frequency (%)
α	60	9.3%
ο	55	8.5%
τ	43	6.6%
ι	36	5.6%
η	36	5.6%
ν	34	5.2%
ε	31	4.8%
ρ	31	4.8%
π	30	4.6%
ς	30	4.6%
Other values (33)	262	40.4%

Devanagari

Value	Count	Frequency (%)
ा	11	14.3%
न	6	7.8%
र	6	7.8%
म	5	6.5%
ी	4	5.2%
द	3	3.9%
ो	3	3.9%
ृ	3	3.9%
प	3	3.9%
अ	3	3.9%
Other values (21)	30	39.0%

Hiragana

Value	Count	Frequency (%)
の	4	20.0%
さ	1	5.0%
ん	1	5.0%
と	1	5.0%
そ	1	5.0%
め	1	5.0%

Value	Count	Frequency (%)
ひ	1	5.0%
ち	1	5.0%
ず	1	5.0%
か	1	5.0%
Other values (7)	7	35.0%

Telugu

Value	Count	Frequency (%)
ి	4	13.3%
ీ	3	10.0%
ొ	3	10.0%
ద	3	10.0%
ా	2	6.7%
న	2	6.7%
స	2	6.7%
మ	2	6.7%
ర	2	6.7%
ఐ	1	3.3%
Other values (6)	6	20.0%

Tamil

Value	Count	Frequency (%)
◌	3	15.8%
ம	2	10.5%
ர	2	10.5%
ா	2	10.5%
ட	2	10.5%
ன	1	5.3%
வ	1	5.3%
த	1	5.3%
ஆ	1	5.3%
ய	1	5.3%
Other values (3)	3	15.8%

Han

Value	Count	Frequency (%)
俣	1	10.0%
界	1	10.0%
患	1	10.0%
者	1	10.0%
世	1	10.0%
水	1	10.0%
鬼	1	10.0%
見	1	10.0%
難	1	10.0%
海	1	10.0%

Hangul

Value	Count	Frequency (%)
사	2	22.2%
회	1	11.1%
식	1	11.1%
주	1	11.1%
기	1	11.1%
찾	1	11.1%
랑	1	11.1%
첫	1	11.1%

Thai

Value	Count	Frequency (%)
◌	2	25.0%
ง	1	12.5%

Value	Count	Frequency (%)
ร	1	12.5%
พ	1	12.5%
แ	1	12.5%
ี	1	12.5%
ส	1	12.5%

Arabic

Value	Count	Frequency (%)
م	2	50.0%
ه	1	25.0%
ت	1	25.0%

Inherited

Value	Count	Frequency (%)
́	4	57.1%
̈	3	42.9%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	14358061	99.9%
Punctuation	7270	0.1%
None	5930	< 0.1%
Cyrillic	4587	< 0.1%
Devanagari	77	< 0.1%
Telugu	30	< 0.1%
Hiragana	20	< 0.1%
Tamil	19	< 0.1%
Letterlike Symbols	14	< 0.1%
CJK	10	< 0.1%
Other values (11)	41	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	2406350	16.8%
e	1365862	9.5%
a	940505	6.6%
t	934766	6.5%
i	852547	5.9%
o	830906	5.8%
n	822601	5.7%
s	767854	5.3%
r	745307	5.2%
h	600810	4.2%
Other values (82)	4090553	28.5%

Punctuation

Value	Count	Frequency (%)
’	3847	52.9%
–	881	12.1%
"	690	9.5%
“	672	9.2%
—	633	8.7%
…	303	4.2%
‘	192	2.6%
	31	0.4%
‍	7	0.1%
―	5	0.1%
Other values (4)	9	0.1%

None

Value	Count	Frequency (%)
é	1552	26.2%
ä	294	5.0%
á	293	4.9%
ö	250	4.2%
í	243	4.1%
è	209	3.5%
ü	178	3.0%
ı	165	2.8%
ó	164	2.8%
ç	158	2.7%
Other values (141)	2424	40.9%

Cyrillic

Value	Count	Frequency (%)
о	470	10.2%
е	404	8.8%
а	373	8.1%
н	323	7.0%
и	299	6.5%
т	265	5.8%
р	240	5.2%
с	218	4.8%
в	173	3.8%
л	161	3.5%
Other values (46)	1661	36.2%

Letterlike Symbols

Value	Count	Frequency (%)
™	14	100.0%

Devanagari

Value	Count	Frequency (%)
◌ॉ	11	14.3%
न	6	7.8%
र	6	7.8%
म	5	6.5%
◌ी	4	5.2%
द	3	3.9%
◌ो	3	3.9%
◌्	3	3.9%
प	3	3.9%
अ	3	3.9%
Other values (21)	30	39.0%

Alphabetic PF

Value	Count	Frequency (%)
ﬁ	4	100.0%

Hiragana

Value	Count	Frequency (%)
の	4	20.0%
さ	1	5.0%
ん	1	5.0%
と	1	5.0%
そ	1	5.0%
め	1	5.0%
ひ	1	5.0%
ち	1	5.0%
ず	1	5.0%
か	1	5.0%
Other values (7)	7	35.0%

Diacriticals

Value	Count	Frequency (%)
´	4	57.1%
¨	3	42.9%

Telugu

Value	Count	Frequency (%)
ి	4	13.3%
్	3	10.0%
ొ	3	10.0%
ద	3	10.0%
ా	2	6.7%
న	2	6.7%
స	2	6.7%
మ	2	6.7%
ర	2	6.7%
ఐ	1	3.3%
Other values (6)	6	20.0%

Tamil

Value	Count	Frequency (%)
்	3	15.8%
ம	2	10.5%
ர	2	10.5%
ா	2	10.5%
ட	2	10.5%
ன	1	5.3%
வ	1	5.3%
த	1	5.3%
ஆ	1	5.3%
ய	1	5.3%
Other values (3)	3	15.8%

Arabic

Value	Count	Frequency (%)
ه	2	50.0%
ة	1	25.0%
ت	1	25.0%

Hangul

Value	Count	Frequency (%)
사	2	22.2%
회	1	11.1%
식	1	11.1%
주	1	11.1%
기	1	11.1%
찾	1	11.1%
랑	1	11.1%
첫	1	11.1%

Number Forms

Value	Count	Frequency (%)
Ⅱ	2	100.0%

Modifier Letters

Value	Count	Frequency (%)
ʾ	2	100.0%

Thai

Value	Count	Frequency (%)
่	2	25.0%
ง	1	12.5%
ร	1	12.5%

Value	Count	Frequency (%)
พ	1	12.5%
แ	1	12.5%
ี	1	12.5%
ส	1	12.5%

CJK

Value	Count	Frequency (%)
俣	1	10.0%
界	1	10.0%
患	1	10.0%
者	1	10.0%
世	1	10.0%
水	1	10.0%
鬼	1	10.0%
見	1	10.0%
難	1	10.0%
海	1	10.0%

Math Operators

Value	Count	Frequency (%)
−	1	100.0%

Katakana

Value	Count	Frequency (%)
・	1	100.0%

Currency Symbols

Value	Count	Frequency (%)
₹	1	50.0%
€	1	50.0%

Specials

Value	Count	Frequency (%)
	1	100.0%

~~Popularity~~
~~Unsupported~~

REJECTED	UNSUPPORTED
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

ProductionCompanies
Categorical

Distinct	22667
Distinct (%)	49.8%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	609
Median length	476
Mean length	33.781289
Min length	2

Characters and Unicode

Total characters	1536069
Distinct characters	294
Distinct categories	17 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	6 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	6 (https://en.wikipedia.org/wiki/Unicode_bloc

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

Unique

Unique	20300	?
Unique (%)	44.6%	

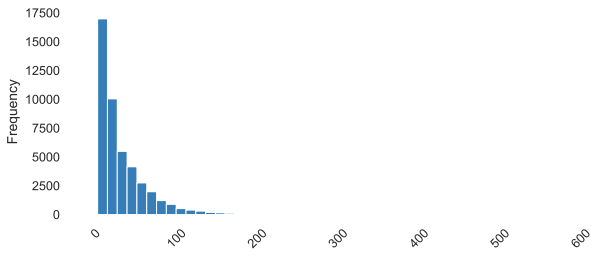
Sample

1st row	Pixar Animation Studios	
2nd row	TriStar Pictures, Teitler Film, Interscope Communications	?
3rd row	Warner Bros., Lancaster Gate	ategory) ?
4th row	Twentieth Century Fox Film Corporation	code) ?
5th row	Sandollar Productions, Touchstone Pictures	

Common Values

Value	Count	Frequency (%)
MissingValue	11891	26.2%
Metro-Goldwyn-Mayer (MGM)	742	1.6%
Warner Bros.	540	1.2%
Paramount Pictures	505	1.1%
Twentieth Century Fox Film Corporation	439	1.0%
Universal Pictures	320	0.7%
RKO Radio Pictures	247	0.5%
Columbia Pictures Corporation	207	0.5%
Columbia Pictures	146	0.3%
Mosfilm	145	0.3%
Other values (22657)	30289	66.6%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
missingvalue	11891	6.3%
films	9455	5.0%
pictures	9267	4.9%
productions	9059	4.8%
film	6679	3.5%
entertainment	5154	2.7%
corporation	2189	1.2%
company	1769	0.9%
warner	1478	0.8%

Value	Count	Frequency (%)
bros	1411	0.7%
Other values (18617)	131220	69.2%

Most occurring characters

Value	Count	Frequency (%)
	144110	9.4%
i	130720	8.5%
e	106535	6.9%
n	101860	6.6%
a	89034	5.8%
s	86449	5.6%
o	85292	5.6%
r	83547	5.4%
t	83433	5.4%
l	63155	4.1%
Other values (284)	561934	36.6%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	1105928	72.0%
Uppercase Letter	222747	14.5%
Space Separator	144115	9.4%
Other Punctuation	45099	2.9%
Decimal Number	4347	0.3%
Dash Punctuation	4331	0.3%
Open Punctuation	4328	0.3%
Close Punctuation	4327	0.3%
Math Symbol	662	< 0.1%
Other Letter	140	< 0.1%
Other values (7)	45	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
i	130720	11.8%
e	106535	9.6%
n	101860	9.2%
a	89034	8.1%
s	86449	7.8%
o	85292	7.7%
r	83547	7.6%
t	83433	7.5%
l	63155	5.7%
u	55642	5.0%
Other values (102)	220261	19.9%

Other Letter

Value	Count	Frequency (%)
스	9	6.4%
트	8	5.7%
인	6	4.3%
엔	5	3.6%
주	5	3.6%
터	5	3.6%
먼	5	3.6%
테	5	3.6%
픽	4	2.9%
로	3	2.1%

Value	Count	Frequency (%)
Other values (62)	85	60.7%

Uppercase Letter

Value	Count	Frequency (%)
P	27880	12.5%
F	26362	11.8%
M	25252	11.3%
C	20585	9.2%
V	14952	6.7%
S	11911	5.3%
E	9746	4.4%
A	9547	4.3%
T	9356	4.2%
B	9001	4.0%
Other values (52)	58155	26.1%

Other Punctuation

Value	Count	Frequency (%)
,	37354	82.8%
.	5671	12.6%
&	764	1.7%
/	645	1.4%
'	451	1.0%
"	133	0.3%
!	36	0.1%
%	18	< 0.1%
:	9	< 0.1%
@	5	< 0.1%
Other values (6)	13	< 0.1%

Decimal Number

Value	Count	Frequency (%)
2	1034	23.8%
1	712	16.4%
0	641	14.7%
3	556	12.8%
4	481	11.1%
9	205	4.7%
6	195	4.5%
5	178	4.1%
8	173	4.0%
7	172	4.0%

Open Punctuation

Value	Count	Frequency (%)
(	4318	99.8%
[	9	0.2%
（	1	< 0.1%

Close Punctuation

Value	Count	Frequency (%)
)	4317	99.8%
]	9	0.2%
）	1	< 0.1%

Space Separator

Value	Count	Frequency (%)
	144110	> 99.9%
	5	< 0.1%

Dash Punctuation

Value	Count	Frequency (%)
-	4329	> 99.9%

Value	Count	Frequency (%)
_	2	< 0.1%

Math Symbol

Value	Count	Frequency (%)
+	661	99.8%
|	1	0.2%

Other Symbol

Value	Count	Frequency (%)
°	23	92.0%
㉿	2	8.0%

Final Punctuation

Value	Count	Frequency (%)
’	3	50.0%
»	3	50.0%

Other Number

Value	Count	Frequency (%)
²	1	50.0%
½	1	50.0%

Control

Value	Count	Frequency (%)
	4	100.0%

Connector Punctuation

Value	Count	Frequency (%)
_	4	100.0%

Initial Punctuation

Value	Count	Frequency (%)
«	3	100.0%

Format

Value	Count	Frequency (%)
	1	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	1328272	86.5%
Common	207252	13.5%
Cyrillic	373	< 0.1%
Hangul	115	< 0.1%
Greek	31	< 0.1%
Han	26	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
i	130720	9.8%
e	106535	8.0%
n	101860	7.7%
a	89034	6.7%
s	86449	6.5%
o	85292	6.4%
r	83547	6.3%
t	83433	6.3%
l	63155	4.8%
u	55642	4.2%
Other values (99)	442605	33.3%

Hangul

Value	Count	Frequency (%)
스	9	7.8%
트	8	7.0%
인	6	5.2%
엔	5	4.3%
주	5	4.3%
터	5	4.3%
먼	5	4.3%
테	5	4.3%
픽	4	3.5%
로	3	2.6%
Other values (43)	60	52.2%

Common

Value	Count	Frequency (%)
	144110	69.5%
,	37354	18.0%
.	5671	2.7%
-	4329	2.1%
(	4318	2.1%
)	4317	2.1%
2	1034	0.5%
&	764	0.4%
1	712	0.3%
+	661	0.3%
Other values (37)	3982	1.9%

Cyrillic

Value	Count	Frequency (%)
и	34	9.1%
о	28	7.5%
а	26	7.0%
л	22	5.9%
н	20	5.4%
м	19	5.1%
т	17	4.6%
с	16	4.3%
е	16	4.3%
ь	16	4.3%
Other values (36)	159	42.6%

Greek

Value	Count	Frequency (%)
ο	3	9.7%
ν	3	9.7%
Ε	2	6.5%
λ	2	6.5%
η	2	6.5%
ι	2	6.5%
τ	2	6.5%
ρ	2	6.5%
Κ	2	6.5%
έ	1	3.2%
Other values (10)	10	32.3%

Han

Value	Count	Frequency (%)
北	2	7.7%
京	2	7.7%
司	2	7.7%
公	2	7.7%
限	2	7.7%
有	2	7.7%

Value	Count	Frequency (%)
影	2	7.7%
乐	1	3.8%
安	1	3.8%
电	1	3.8%
Other values (9)	9	34.6%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	1529839	99.6%
None	5711	0.4%
Cyrillic	373	< 0.1%
Hangul	113	< 0.1%
CJK	26	< 0.1%
Punctuation	7	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	144110	9.4%
i	130720	8.5%
e	106535	7.0%
n	101860	6.7%
a	89034	5.8%
s	86449	5.7%
o	85292	5.6%
r	83547	5.5%
t	83433	5.5%
l	63155	4.1%
Other values (77)	555704	36.3%

None

Value	Count	Frequency (%)
é	3176	55.6%
ó	416	7.3%
á	317	5.6%
í	173	3.0%
ü	154	2.7%
ñ	150	2.6%
ô	140	2.5%
ä	137	2.4%
è	136	2.4%
ö	132	2.3%
Other values (76)	780	13.7%

Cyrillic

Value	Count	Frequency (%)
и	34	9.1%
о	28	7.5%
а	26	7.0%
л	22	5.9%
н	20	5.4%
м	19	5.1%
т	17	4.6%
с	16	4.3%
е	16	4.3%
ь	16	4.3%
Other values (36)	159	42.6%

Hangul

Value	Count	Frequency (%)
스	9	8.0%

Value	Count	Frequency (%)
트	8	7.1%
인	6	5.3%
엔	5	4.4%
주	5	4.4%
터	5	4.4%
먼	5	4.4%
테	5	4.4%
픽	4	3.5%
로	3	2.7%
Other values (42)	58	51.3%

Punctuation

Value	Count	Frequency (%)
'	3	42.9%
–	2	28.6%
	1	14.3%
•	1	14.3%

CJK

Value	Count	Frequency (%)
北	2	7.7%
京	2	7.7%
司	2	7.7%
公	2	7.7%
限	2	7.7%
有	2	7.7%
影	2	7.7%
乐	1	3.8%
安	1	3.8%
电	1	3.8%
Other values (9)	9	34.6%

ProductionCountries

Categorical

HIGH CARDINALITY IMBALANCE

Distinct	2390
Distinct (%)	5.3%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	98
Median length	2
Mean length	4.5794023
Min length	2

Characters and Unicode

Total characters	208230
Distinct characters	42
Distinct categories	4 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)

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

Unique

Unique	1764
Unique (%)	3.9%

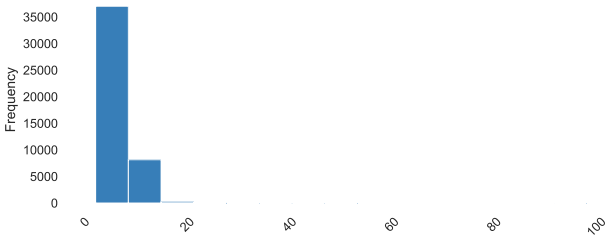
Sample

1st row	US
2nd row	US
3rd row	US
4th row	US
5th row	US

Common Values

Value	Count	Frequency (%)
US	17846	39.2%
Missing values	6214	13.7%
GB	2235	4.9%
FR	1653	3.6%
JP	1356	3.0%
IT	1029	2.3%
CA	840	1.8%
DE	749	1.6%
IN	735	1.6%
RU	734	1.6%
Other values (2380)	12080	26.6%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
us	21147	34.1%
values	6214	10.0%
missing	6214	10.0%
gb	4091	6.6%
fr	3939	6.4%
de	2254	3.6%
it	2168	3.5%
ca	1765	2.9%
jp	1648	2.7%
es	964	1.6%
Other values (154)	11524	18.6%

Most occurring characters

Value	Count	Frequency (%)
S	23041	11.1%

Value	Count	Frequency (%)
s	18734	9.0%
	16457	7.9%
i	12612	6.1%
,	10243	4.9%
R	6686	3.2%
M	6660	3.2%
u	6398	3.1%
n	6398	3.1%
Other values (32)	77977	37.4%

Most occurring categories

Value	Count	Frequency (%)
Uppercase Letter	105306	50.6%
Lowercase Letter	76224	36.6%
Space Separator	16457	7.9%
Other Punctuation	10243	4.9%

Most frequent character per category

Uppercase Letter

Value	Count	Frequency (%)
S	23041	21.9%
U	23024	21.9%
R	6686	6.3%
M	6660	6.3%
B	4982	4.7%
E	4752	4.5%
G	4448	4.2%
F	4342	4.1%
I	4010	3.8%
A	3136	3.0%
Other values (16)	20225	19.2%

Lowercase Letter

Value	Count	Frequency (%)
s	18734	24.6%
i	12612	16.5%
u	6398	8.4%
n	6398	8.4%
e	6306	8.3%
a	6214	8.2%
l	6214	8.2%
v	6214	8.2%
g	6214	8.2%
o	368	0.5%
Other values (4)	552	0.7%

Space Separator

Value	Count	Frequency (%)
	16457	100.0%

Other Punctuation

Value	Count	Frequency (%)
,	10243	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	181530	87.2%
Common	26700	12.8%

Most frequent character per script

Latin

Value	Count	Frequency (%)
S	23041	12.7%
U	23024	12.7%
s	18734	10.3%
i	12612	6.9%
R	6686	3.7%
M	6660	3.7%
u	6398	3.5%
n	6398	3.5%
e	6306	3.5%
a	6214	3.4%
Other values (30)	65457	36.1%

Common

Value	Count	Frequency (%)
	16457	61.6%
,	10243	38.4%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	208230	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
S	23041	11.1%
U	23024	11.1%
s	18734	9.0%
	16457	7.9%
i	12612	6.1%
,	10243	4.9%
R	6686	3.2%
M	6660	3.2%
u	6398	3.1%
n	6398	3.1%
Other values (32)	77977	37.4%

ReleaseDate

Categorical

Distinct	17334
Distinct (%)	38.1%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	13
Median length	10
Mean length	10.006268
Min length	10

Characters and Unicode

Total characters	454995
Distinct characters	20
Distinct categories	4 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category) ?
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode) ?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block) ?

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

Unique

Unique	8570 ?
Unique (%)	18.8%

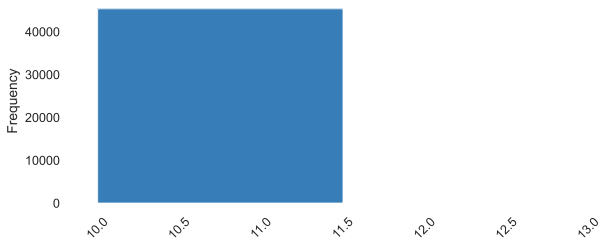
Sample

1st row	1995-10-30
2nd row	1995-12-15
3rd row	1995-12-22
4th row	1995-12-22
5th row	1995-02-10

Common Values

Value	Count	Frequency (%)
2008-01-01	136	0.3%
2009-01-01	121	0.3%
2007-01-01	118	0.3%
2005-01-01	111	0.2%
2006-01-01	101	0.2%
2002-01-01	96	0.2%
NoReleaseDate	95	0.2%
2004-01-01	90	0.2%
2001-01-01	84	0.2%
2003-01-01	76	0.2%
Other values (17324)	44443	97.7%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
2008-01-01	136	0.3%
2009-01-01	121	0.3%
2007-01-01	118	0.3%
2005-01-01	111	0.2%
2006-01-01	101	0.2%
2002-01-01	96	0.2%
noreleasedate	95	0.2%
2004-01-01	90	0.2%
2001-01-01	84	0.2%
2003-01-01	76	0.2%
Other values (17324)	44443	97.7%

Most occurring characters

Value	Count	Frequency (%)
0	97600	21.5%

Value	Count	Frequency (%)
-	90752	19.9%
1	84054	18.5%
2	52803	11.6%
9	39773	8.7%
3	15435	3.4%
8	15279	3.4%
6	15021	3.3%
5	14836	3.3%
7	14289	3.1%
Other values (10)	15153	3.3%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	363008	79.8%
Dash Punctuation	90752	19.9%
Lowercase Letter	950	0.2%
Uppercase Letter	285	0.1%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	97600	26.9%
1	84054	23.2%
2	52803	14.5%
9	39773	11.0%
3	15435	4.3%
8	15279	4.2%
6	15021	4.1%
5	14836	4.1%
7	14289	3.9%
4	13918	3.8%

Lowercase Letter

Value	Count	Frequency (%)
e	380	40.0%
a	190	20.0%
l	95	10.0%
s	95	10.0%
t	95	10.0%
o	95	10.0%

Uppercase Letter

Value	Count	Frequency (%)
N	95	33.3%
R	95	33.3%
D	95	33.3%

Dash Punctuation

Value	Count	Frequency (%)
-	90752	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	453760	99.7%
Latin	1235	0.3%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	97600	21.5%
-	90752	20.0%
1	84054	18.5%
2	52803	11.6%
9	39773	8.8%
3	15435	3.4%
8	15279	3.4%
6	15021	3.3%
5	14836	3.3%
7	14289	3.1%

Latin

Value	Count	Frequency (%)
e	380	30.8%
a	190	15.4%
N	95	7.7%
R	95	7.7%
l	95	7.7%
s	95	7.7%
D	95	7.7%
t	95	7.7%
o	95	7.7%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	454995	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	97600	21.5%
-	90752	19.9%
1	84054	18.5%
2	52803	11.6%
9	39773	8.7%
3	15435	3.4%
8	15279	3.4%
6	15021	3.3%
5	14836	3.3%
7	14289	3.1%
Other values (10)	15153	3.3%

~~Revenue~~

~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

~~Runtime~~

~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

SpokenLanguages

Categorical

HIGH CARDINALITY IMBALANCE

Distinct	1932
Distinct (%)	4.2%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	74
Median length	2
Mean length	4.0517693
Min length	2

Characters and Unicode

Total characters	184238
Distinct characters	32
Distinct categories	4 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)

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

Unique

Unique	1366
Unique (%)	3.0%

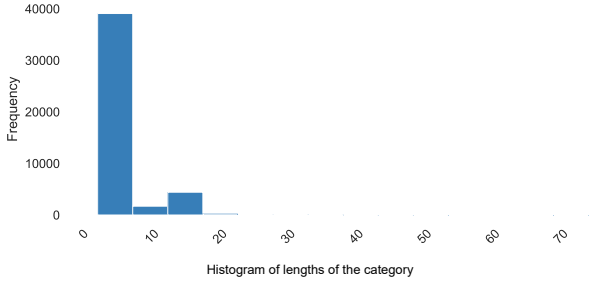
Sample

1st row	en
2nd row	en, fr
3rd row	en
4th row	en
5th row	en

Common Values

Value	Count	Frequency (%)
en	22380	49.2%
Missing values	3771	8.3%
fr	1852	4.1%
ja	1289	2.8%
it	1217	2.7%
es	901	2.0%
ru	807	1.8%
de	761	1.7%
en, fr	681	1.5%
en, es	572	1.3%
Other values (1922)	11240	24.7%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
en	28729	47.2%

Value	Count	Frequency (%)
fr	4194	6.9%
missing	3771	6.2%
values	3771	6.2%
de	2624	4.3%
es	2412	4.0%
it	2366	3.9%
ja	1758	2.9%
ru	1562	2.6%
zh	790	1.3%
Other values (126)	8931	14.7%

Most occurring characters

Value	Count	Frequency (%)
e	38314	20.8%
n	33761	18.3%
	15437	8.4%
s	15035	8.2%
,	11666	6.3%
i	11232	6.1%
a	6906	3.7%
r	6736	3.7%
u	5970	3.2%
l	5173	2.8%
Other values (22)	34008	18.5%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	153088	83.1%
Space Separator	15437	8.4%
Other Punctuation	11666	6.3%
Uppercase Letter	4047	2.2%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	38314	25.0%
n	33761	22.1%
s	15035	9.8%
i	11232	7.3%
a	6906	4.5%
r	6736	4.4%
u	5970	3.9%
l	5173	3.4%
f	4740	3.1%
v	4420	2.9%
Other values (16)	20801	13.6%

Uppercase Letter

Value	Count	Frequency (%)
M	3771	93.2%
N	92	2.3%
S	92	2.3%
L	92	2.3%

Space Separator

Value	Count	Frequency (%)
	15437	100.0%

Other Punctuation

Value	Count	Frequency (%)
,	11666	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	157135	85.3%
Common	27103	14.7%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	38314	24.4%
n	33761	21.5%
s	15035	9.6%
i	11232	7.1%
a	6906	4.4%
r	6736	4.3%
u	5970	3.8%
l	5173	3.3%
f	4740	3.0%
v	4420	2.8%
Other values (20)	24848	15.8%

Common

Value	Count	Frequency (%)
	15437	57.0%
,	11666	43.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	184238	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	38314	20.8%
n	33761	18.3%
	15437	8.4%
s	15035	8.2%
,	11666	6.3%
i	11232	6.1%
a	6906	3.7%
r	6736	3.7%
u	5970	3.2%
l	5173	2.8%
Other values (22)	34008	18.5%

Tagline

Categorical

HIGH CARDINALITY MISSING UNIFORM

Distinct	20269
Distinct (%)	99.4%
Missing	25073
Missing (%)	55.1%
Memory size	355.4 KiB

Length

Max length	297
Median length	204
Mean length	46.999314
Min length	1

Characters and Unicode

Total characters	958692
Distinct characters	170
Distinct categories	17 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	6 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	10 (https://en.wikipedia.org/wiki/Unicode_blo

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

Unique

Unique	20163	?
Unique (%)	98.8%	

Sample

1st row	Roll the dice and unleash the excitement!
2nd row	Still Yelling. Still Fighting. Still Ready for Love.
3rd row	Friends are the people who let you be yourself... and never let you forget it.
4th row	Just When His World Is Back To Normal... He's In For The Surprise Of His Life!
5th row	A Los Angeles Crime Saga

?

?

?

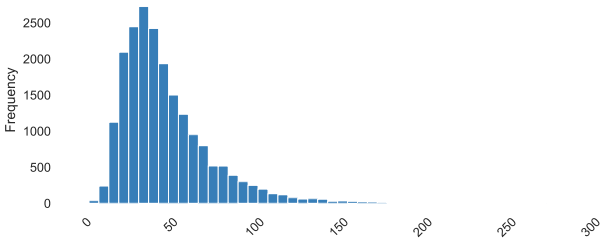
ategory)

code)

Common Values

Value	Count	Frequency (%)
Based on a true story.	7	< 0.1%
Trust no one.	4	< 0.1%
Be careful what you wish for.	4	< 0.1%
-	4	< 0.1%
How far would you go?	3	< 0.1%
Drama	3	< 0.1%
Classic Albums	3	< 0.1%
There are two sides to every love story.	3	< 0.1%
There is no turning back	3	< 0.1%
Documentary	3	< 0.1%
Other values (20259)	20361	44.8%
(Missing)	25073	55.1%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
the	10998	6.3%
a	6815	3.9%
of	4404	2.5%
to	3584	2.1%
is	2796	1.6%
in	2693	1.5%
and	2682	1.5%
you	2389	1.4%
	1582	0.9%
for	1523	0.9%
Other values (15100)	134470	77.3%

Most occurring characters

Value	Count	Frequency (%)
	153686	16.0%
e	94412	9.8%
t	57267	6.0%
o	56566	5.9%
a	51473	5.4%
n	47498	5.0%
i	46036	4.8%
r	44992	4.7%
s	42360	4.4%
h	37172	3.9%
Other values (160)	327230	34.1%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	680479	71.0%
Space Separator	153686	16.0%
Uppercase Letter	74991	7.8%
Other Punctuation	44585	4.7%
Decimal Number	2687	0.3%
Dash Punctuation	1944	0.2%
Final Punctuation	98	< 0.1%
Open Punctuation	56	< 0.1%
Close Punctuation	55	< 0.1%
Currency Symbol	37	< 0.1%
Other values (7)	74	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	94412	13.9%
t	57267	8.4%
o	56566	8.3%
a	51473	7.6%
n	47498	7.0%
i	46036	6.8%
r	44992	6.6%
s	42360	6.2%
h	37172	5.5%
l	30174	4.4%
Other values (43)	172529	25.4%

Other Letter

Value	Count	Frequency (%)
ല	1	2.9%
ಠ	1	2.9%
成	1	2.9%
劇	1	2.9%
熟	1	2.9%
ઠ	1	2.9%
ಹ	1	2.9%
時	1	2.9%
舞	1	2.9%
場	1	2.9%
Other values (24)	24	70.6%

Uppercase Letter

Value	Count	Frequency (%)
T	10009	13.3%
A	6874	9.2%

Value	Count	Frequency (%)
S	5652	7.5%
H	4402	5.9%
I	4387	5.9%
E	4306	5.7%
W	3681	4.9%
O	3477	4.6%
N	3195	4.3%
L	3194	4.3%
Other values (20)	25814	34.4%

Other Punctuation

Value	Count	Frequency (%)
.	26647	59.8%
!	5784	13.0%
'	5674	12.7%
,	4226	9.5%
?	1161	2.6%
"	582	1.3%
…	148	0.3%
:	138	0.3%
&	83	0.2%
*	42	0.1%
Other values (7)	100	0.2%

Decimal Number

Value	Count	Frequency (%)
0	802	29.8%
1	516	19.2%
2	299	11.1%
3	208	7.7%
9	208	7.7%
5	168	6.3%
4	140	5.2%
6	121	4.5%
7	121	4.5%
8	104	3.9%

Math Symbol

Value	Count	Frequency (%)
+	5	35.7%
=	5	35.7%
|	2	14.3%
~	1	7.1%
−	1	7.1%

Dash Punctuation

Value	Count	Frequency (%)
-	1927	99.1%
–	9	0.5%
—	8	0.4%

Final Punctuation

Value	Count	Frequency (%)
’	82	83.7%
”	15	15.3%
»	1	1.0%

Initial Punctuation

Value	Count	Frequency (%)
“	14	73.7%
‘	4	21.1%
«	1	5.3%

Open Punctuation

Value	Count	Frequency (%)
(	49	87.5%
[	7	12.5%

Close Punctuation

Value	Count	Frequency (%)
)	48	87.3%
]	7	12.7%

Other Number

Value	Count	Frequency (%)
½	2	66.7%
²	1	33.3%

Modifier Letter

Value	Count	Frequency (%)
ˌ	1	50.0%
ˈ	1	50.0%

Space Separator

Value	Count	Frequency (%)
	153686	100.0%

Currency Symbol

Value	Count	Frequency (%)
$	37	100.0%

Nonspacing Mark

Value	Count	Frequency (%)
◌̇	1	100.0%

Connector Punctuation

Value	Count	Frequency (%)
_	1	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	755470	78.8%
Common	203187	21.2%
Han	21	< 0.1%
Tamil	5	< 0.1%
Hiragana	5	< 0.1%
Katakana	4	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	94412	12.5%
t	57267	7.6%
o	56566	7.5%
a	51473	6.8%
n	47498	6.3%
i	46036	6.1%
r	44992	6.0%
s	42360	5.6%
h	37172	4.9%
l	30174	4.0%
Other values (73)	247520	32.8%

Common

Value	Count	Frequency (%)
	153686	75.6%

Value	Count	Frequency (%)
.	26647	13.1%
!	5784	2.8%
'	5674	2.8%
,	4226	2.1%
-	1927	0.9%
?	1161	0.6%
0	802	0.4%
"	582	0.3%
1	516	0.3%
Other values (42)	2182	1.1%

Han

Value	Count	Frequency (%)
成	1	4.8%
劇	1	4.8%
熟	1	4.8%
時	1	4.8%
舞	1	4.8%
場	1	4.8%
版	1	4.8%
蜜	1	4.8%
最	1	4.8%
后	1	4.8%
Other values (11)	11	52.4%

Tamil

Value	Count	Frequency (%)
ฉ	1	20.0%
◌	1	20.0%
๏	1	20.0%
ฐ	1	20.0%

Value	Count	Frequency (%)
ஆ	1	20.0%

Hiragana

Value	Count	Frequency (%)
は	1	20.0%
し	1	20.0%
て	1	20.0%
い	1	20.0%
る	1	20.0%

Katakana

Value	Count	Frequency (%)
ク	1	25.0%
ラ	1	25.0%
ナ	1	25.0%
ド	1	25.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	958262	> 99.9%
Punctuation	280	< 0.1%
None	110	< 0.1%
CJK	21	< 0.1%
Tamil	5	< 0.1%
Hiragana	5	< 0.1%
Katakana	4	< 0.1%
IPA Ext	2	< 0.1%
Modifier Letters	2	< 0.1%
Math Operators	1	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	153686	16.0%
e	94412	9.9%
t	57267	6.0%
o	56566	5.9%
a	51473	5.4%
n	47498	5.0%
i	46036	4.8%
r	44992	4.7%
s	42360	4.4%
h	37172	3.9%
Other values (78)	326800	34.1%

Punctuation

Value	Count	Frequency (%)
…	148	52.9%
’	82	29.3%
”	15	5.4%
“	14	5.0%
–	9	3.2%
—	8	2.9%
‘	4	1.4%

None

Value	Count	Frequency (%)
é	18	16.4%
ä	16	14.5%
ö	8	7.3%
á	6	5.5%

Value	Count	Frequency (%)
ó	6	5.5%
ü	5	4.5%
í	5	4.5%
ı	5	4.5%
·	4	3.6%
ć	3	2.7%
Other values (26)	34	30.9%

IPA Ext

Value	Count	Frequency (%)
ə	2	100.0%

Tamil

Value	Count	Frequency (%)
ஸ	1	20.0%
◌்	1	20.0%
ஶ	1	20.0%
ஜ	1	20.0%
ஆ	1	20.0%

CJK

Value	Count	Frequency (%)
成	1	4.8%
劇	1	4.8%
熟	1	4.8%
時	1	4.8%
舞	1	4.8%
場	1	4.8%
版	1	4.8%
蜜	1	4.8%
最	1	4.8%
后	1	4.8%
Other values (11)	11	52.4%

Katakana

Value	Count	Frequency (%)
ク	1	25.0%
ラ	1	25.0%
ナ	1	25.0%
ド	1	25.0%

Modifier Letters

Value	Count	Frequency (%)
ˌ	1	50.0%
ˈ	1	50.0%

Hiragana

Value	Count	Frequency (%)
は	1	20.0%
し	1	20.0%
て	1	20.0%
い	1	20.0%
る	1	20.0%

Math Operators

Value	Count	Frequency (%)
−	1	100.0%

Title

Categorical

HIGH CARDINALITY UNIFORM

Distinct	42197
Distinct (%)	92.8%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	105
Median length	79
Mean length	16.681511
Min length	1

Characters and Unicode

Total characters	758525
Distinct characters	287
Distinct categories	17 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	7 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	12 (https://en.wikipedia.org/wiki/Unicode_blo

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

Unique

Unique	39869	?
Unique (%)	87.7%	

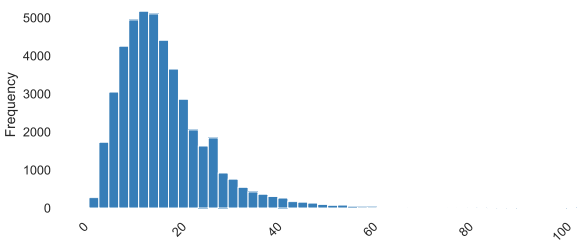
Sample

1st row	Toy Story	
2nd row	Jumanji	
3rd row	Grumpier Old Men	?
4th row	Waiting to Exhale	?
5th row	Father of the Bride Part II	?

Common Values

Value	Count	Frequency (%)
NoTitle	95	0.2%
Cinderella	11	< 0.1%
Alice in Wonderland	9	< 0.1%
Hamlet	9	< 0.1%
Les Misérables	8	< 0.1%
Beauty and the Beast	8	< 0.1%
Treasure Island	7	< 0.1%
A Christmas Carol	7	< 0.1%
The Three Musketeers	7	< 0.1%
Blackout	7	< 0.1%
Other values (42187)	45303	99.6%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
the	14555	10.7%
of	4930	3.6%
a	2241	1.6%
in	1693	1.2%
and	1631	1.2%
to	1054	0.8%
	757	0.6%
man	665	0.5%
love	664	0.5%
for	601	0.4%
Other values (24354)	107485	78.9%

Most occurring characters

Value	Count	Frequency (%)
	90827	12.0%
e	76346	10.1%
a	48940	6.5%
o	45766	6.0%
n	40817	5.4%
r	40018	5.3%
i	39859	5.3%
t	36817	4.9%
s	29519	3.9%
h	28516	3.8%
Other values (277)	281100	37.1%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	534609	70.5%
Uppercase Letter	117455	15.5%
Space Separator	90827	12.0%
Other Punctuation	10489	1.4%
Decimal Number	3850	0.5%
Dash Punctuation	981	0.1%
Close Punctuation	87	< 0.1%
Open Punctuation	85	< 0.1%
Final Punctuation	38	< 0.1%
Other Letter	25	< 0.1%
Other values (7)	79	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	76346	14.3%
a	48940	9.2%
o	45766	8.6%
n	40817	7.6%
r	40018	7.5%
i	39859	7.5%
t	36817	6.9%
s	29519	5.5%
h	28516	5.3%
l	26019	4.9%
Other values (121)	121992	22.8%

Uppercase Letter

Value	Count	Frequency (%)
T	16114	13.7%
S	10336	8.8%
M	8031	6.8%
B	7659	6.5%
C	7165	6.1%
A	6785	5.8%
D	6335	5.4%
L	5872	5.0%
H	5170	4.4%
W	5166	4.4%
Other values (65)	38822	33.1%

Other Letter

Value	Count	Frequency (%)
ج	2	8.0%
ه	2	8.0%
ى	2	8.0%

Value	Count	Frequency (%)
ک	2	8.0%
傅	1	4.0%
空	1	4.0%
時	1	4.0%
狗	1	4.0%
貓	1	4.0%
ª	1	4.0%
Other values (11)	11	44.0%

Other Punctuation

Value	Count	Frequency (%)
:	3717	35.4%
'	2505	23.9%
.	1603	15.3%
،	1134	10.8%
!	647	6.2%
&	458	4.4%
?	269	2.6%
/	79	0.8%
*	19	0.2%
#	13	0.1%
Other values (8)	45	0.4%

Decimal Number

Value	Count	Frequency (%)
2	861	22.4%
1	697	18.1%
0	616	16.0%
3	482	12.5%
9	230	6.0%
4	229	5.9%
5	225	5.8%
7	193	5.0%
8	161	4.2%
6	156	4.1%

Math Symbol

Value	Count	Frequency (%)
+	17	70.8%
×	3	12.5%
∞	1	4.2%
=	1	4.2%
→	1	4.2%
−	1	4.2%

Other Number

Value	Count	Frequency (%)
½	12	63.2%
²	3	15.8%
³	2	10.5%
⅓	1	5.3%
⁴	1	5.3%

Other Symbol

Value	Count	Frequency (%)
°	3	37.5%
☆	2	25.0%
™	1	12.5%
♡	1	12.5%
№	1	12.5%

Currency Symbol

Value	Count	Frequency (%)
$	18	85.7%
¢	2	9.5%
£	1	4.8%

Dash Punctuation

Value	Count	Frequency (%)
-	966	98.5%
–	15	1.5%

Close Punctuation

Value	Count	Frequency (%)
)	82	94.3%
]	5	5.7%

Open Punctuation

Value	Count	Frequency (%)
(	80	94.1%
[	5	5.9%

Final Punctuation

Value	Count	Frequency (%)
’	37	97.4%
”	1	2.6%

Initial Punctuation

Value	Count	Frequency (%)
‘	1	50.0%
“	1	50.0%

Space Separator

Value	Count	Frequency (%)
	90827	100.0%

Connector Punctuation

Value	Count	Frequency (%)
_	3	100.0%

Format

Value	Count	Frequency (%)
	2	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	651549	85.9%
Common	106436	14.0%
Cyrillic	346	< 0.1%
Greek	170	< 0.1%
Arabic	11	< 0.1%
Katakana	8	< 0.1%
Han	5	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	76346	11.7%
a	48940	7.5%
o	45766	7.0%
n	40817	6.3%
r	40018	6.1%
i	39859	6.1%
t	36817	5.7%

Value	Count	Frequency (%)
s	29519	4.5%
h	28516	4.4%
l	26019	4.0%
Other values (107)	238932	36.7%

Common

Value	Count	Frequency (%)
	90827	85.3%
:	3717	3.5%
'	2505	2.4%
.	1603	1.5%
‚	1134	1.1%
-	966	0.9%
2	861	0.8%
1	697	0.7%
!	647	0.6%
0	616	0.6%
Other values (50)	2863	2.7%

Cyrillic

Value	Count	Frequency (%)
е	32	9.2%
о	32	9.2%
а	29	8.4%
н	24	6.9%
и	23	6.6%
р	22	6.4%
к	17	4.9%
с	15	4.3%
в	14	4.0%
л	14	4.0%
Other values (38)	124	35.8%

Greek

Value	Count	Frequency (%)
α	20	11.8%
ι	14	8.2%
ο	14	8.2%
τ	9	5.3%
λ	8	4.7%
ά	8	4.7%
ρ	8	4.7%
ν	7	4.1%
π	6	3.5%
η	6	3.5%
Other values (32)	70	41.2%

Katakana

Value	Count	Frequency (%)
テ	1	12.5%
ポ	1	12.5%
ィ	1	12.5%
ス	1	12.5%
タ	1	12.5%
ン	1	12.5%
ァ	1	12.5%
フ	1	12.5%

Arabic

Value	Count	Frequency (%)
ۃ	2	18.2%
۔	2	18.2%

Value	Count	Frequency (%)
ی	2	18.2%
ک	2	18.2%
ں	1	9.1%
ا	1	9.1%
ج	1	9.1%

Han

Value	Count	Frequency (%)
傳	1	20.0%
空	1	20.0%
時	1	20.0%
狗	1	20.0%
貓	1	20.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	756960	99.8%
None	1124	0.1%
Cyrillic	346	< 0.1%
Punctuation	62	< 0.1%
Arabic	11	< 0.1%
Katakana	8	< 0.1%
CJK	5	< 0.1%
Misc Symbols	3	< 0.1%
Letterlike Symbols	2	< 0.1%
Math Operators	2	< 0.1%
Other values (2)	2	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	90827	12.0%
e	76346	10.1%
a	48940	6.5%
o	45766	6.0%
n	40817	5.4%
r	40018	5.3%
i	39859	5.3%
t	36817	4.9%
s	29519	3.9%
h	28516	3.8%
Other values (76)	279535	36.9%

None

Value	Count	Frequency (%)
é	218	19.4%
ä	127	11.3%
ö	55	4.9%
è	53	4.7%
ô	44	3.9%
ü	39	3.5%
ó	37	3.3%
á	35	3.1%
ı	35	3.1%
í	33	2.9%
Other values (108)	448	39.9%

Punctuation

Value	Count	Frequency (%)
'	37	59.7%

Value	Count	Frequency (%)
–	15	24.2%
…	5	8.1%
	2	3.2%
‘	1	1.6%
”	1	1.6%
“	1	1.6%

Cyrillic

Value	Count	Frequency (%)
е	32	9.2%
о	32	9.2%
а	29	8.4%
н	24	6.9%
и	23	6.6%
р	22	6.4%
к	17	4.9%
с	15	4.3%
в	14	4.0%
л	14	4.0%
Other values (38)	124	35.8%

Arabic

Value	Count	Frequency (%)
ة	2	18.2%
ه	2	18.2%
ى	2	18.2%
ك	2	18.2%
ص	1	9.1%
ا	1	9.1%
ج	1	9.1%

Misc Symbols

Value	Count	Frequency (%)
☆	2	66.7%
♡	1	33.3%

CJK

Value	Count	Frequency (%)
傅	1	20.0%
空	1	20.0%
時	1	20.0%
狗	1	20.0%
貓	1	20.0%

Number Forms

Value	Count	Frequency (%)
⅓	1	100.0%

Letterlike Symbols

Value	Count	Frequency (%)
™	1	50.0%
№	1	50.0%

Math Operators

Value	Count	Frequency (%)
∞	1	50.0%
−	1	50.0%

Katakana

Value	Count	Frequency (%)
テ	1	12.5%
ポ	1	12.5%
イ	1	12.5%

Value	Count	Frequency (%)
ス	1	12.5%
タ	1	12.5%
ン	1	12.5%
ァ	1	12.5%
フ	1	12.5%

Arrows

Value	Count	Frequency (%)
→	1	100.0%

~~VoteAverage~~

~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

~~VoteCount~~

~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

~~ReleaseYear~~

~~Unsupported~~

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

~~Return~~

~~Unsupported~~

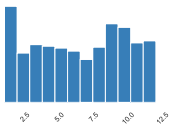
REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

ReleaseMonth

Real number (ℝ)

Distinct	12	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	12
Missing	95	**Zeros**	0
Missing (%)	0.2%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6.4590753	**Memory size**	355.4 KiB

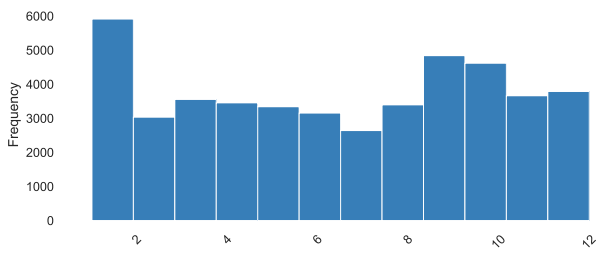


Quantile statistics

Minimum	1
5-th percentile	1
Q1	3
median	7
Q3	10
95-th percentile	12
Maximum	12
Range	11
Interquartile range (IQR)	7

Descriptive statistics

Standard deviation	3.6281605
Coefficient of variation (CV)	0.56171515
Kurtosis	-1.3247729
Mean	6.4590753
Median Absolute Deviation (MAD)	3
Skewness	-0.071880633
Sum	293087
Variance	13.163548
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=12)

Value	Count	Frequency (%)
1	5912	13.0%
9	4838	10.6%
10	4615	10.1%

Value	Count	Frequency (%)
12	3786	8.3%
11	3661	8.1%
3	3553	7.8%
4	3453	7.6%
8	3394	7.5%
5	3339	7.3%
6	3153	6.9%
Other values (2)	5672	12.5%

Value	Count	Frequency (%)
1	5912	13.0%
2	3032	6.7%
3	3553	7.8%
4	3453	7.6%
5	3339	7.3%
6	3153	6.9%
7	2640	5.8%
8	3394	7.5%
9	4838	10.6%
10	4615	10.1%

Value	Count	Frequency (%)
12	3786	8.3%
11	3661	8.1%
10	4615	10.1%
9	4838	10.6%
8	3394	7.5%
7	2640	5.8%
6	3153	6.9%
5	3339	7.3%
4	3453	7.6%
3	3553	7.8%

MovieCharacter

Categorical

Distinct	40180
Distinct (%)	88.4%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	6647
Median length	1773
Mean length	168.86416
Min length	2

Characters and Unicode

Total characters	7678422
Distinct characters	618
Distinct categories	20 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	12 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	14 (https://en.wikipedia.org/wiki/Unicode_blo

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

Unique

Unique	39945
Unique (%)	87.8%

Sample

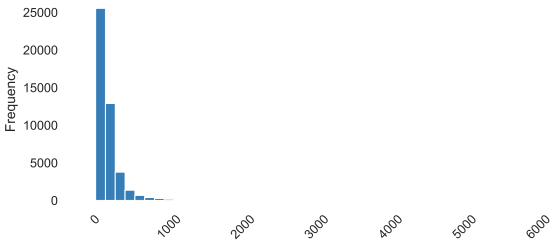
1st row	Woody (voice), Buzz Lightyear (voice), Mr. Potato Head (voice), Slinky Dog (voice), Rex (voice), Hamm (voice), Bo Peep (voice), Andy (voice), Sid (voice), Mrs. Davis (voice), Sergeant (voice), Hannah (voice), TV Announcer (voice)
2nd row	Alan Parrish, Samuel Alan Parrish / Van Pelt, Judy Sheperd, Peter Shepherd, Sarah Whittle, Nora Shepherd, Carl Bentley, Carol Anne Parrish, Alan Parrish (young), Sarah Whittle (young), Exterminator, Mrs. Thomas the Realtor, Benjamin, Caleb, Billy Jessup, Cop, Bum, Jim Shepherd, Martha Shepherd, Gun Salesman, Paramedic, Paramedic, Girl, Girl, Baker, Pianist
3rd row	Max Goldman, John Gustafson, Ariel Gustafson, Maria Sophia Coletta Ragetti, Melanie Gustafson, Grandpa Gustafson, Jacob Goldman
4th row	Savannah 'Vannah' Jackson, Bernadine 'Bernie' Harris, Gloria 'Glo' Matthews, Robin Stokes, Marvin King, Kenneth Dawkins, John Harris, Sr., Troy, Joseph, James Wheeler
5th row	George Banks, Nina Banks, Franck Eggelhoffer, Annie Banks-MacKenzie, Bryan MacKenzie, Matty Banks, Howard Weinstein, John MacKenzie, Joanna MacKenzie, Dr. Megan Eisenberg, Mr. Habib, Wife Mrs. Habib

Common Values

Value	Count	Frequency (%)
NoCharacter	2565	5.6%

Value	Count	Frequency (%)
Himself	315	0.4%
, , ,	211	0.5%
, , , ,	209	0.5%
, ,	141	0.3%
, , , , ,	129	0.3%
Narrator	124	0.3%
,	115	0.3%
, , , , , ,	107	0.2%
, , , , , , ,	85	0.2%
Other values (40170)	41269	90.8%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
	37208	3.5%
uncredited	19404	1.8%
himself	14232	1.3%
voice	13783	1.3%
the	10319	1.0%
dr	6831	0.6%
mrs	5580	0.5%
man	5252	0.5%
mr	5177	0.5%
girl	4420	0.4%
Other values (130038)	947604	88.6%

Most occurring characters

Value	Count	Frequency (%)
	1028323	13.4%
e	649941	8.5%
,	525726	6.8%
a	519505	6.8%
r	472126	6.1%
i	419713	5.5%
n	404639	5.3%
o	360968	4.7%
t	288075	3.8%
l	273828	3.6%
Other values (608)	2735578	35.6%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4977388	64.8%
Space Separator	1028323	13.4%
Uppercase Letter	950780	12.4%
Other Punctuation	608315	7.9%
Open Punctuation	42194	0.5%
Close Punctuation	42154	0.5%
Decimal Number	14368	0.2%
Dash Punctuation	13905	0.2%
Other Letter	632	< 0.1%
Final Punctuation	141	< 0.1%

Value	Count	Frequency (%)
Other values (10)	222	< 0.1%

Most frequent character per category

Other Letter

Value	Count	Frequency (%)
º	25	4.0%
ا	25	4.0%
ل	17	2.7%
ي	14	2.2%
ب	12	1.9%
د	9	1.4%
ה	9	1.4%
ر	9	1.4%
و	8	1.3%
ص	8	1.3%
Other values (274)	496	78.5%

Lowercase Letter

Value	Count	Frequency (%)
e	649941	13.1%
a	519505	10.4%
r	472126	9.5%
i	419713	8.4%
n	404639	8.1%
o	360968	7.3%
t	288075	5.8%
l	273828	5.5%
s	242202	4.9%
d	174401	3.5%
Other values (150)	1171990	23.5%

Uppercase Letter

Value	Count	Frequency (%)
M	93697	9.9%
S	80372	8.5%
C	75966	8.0%
B	61447	6.5%
D	57096	6.0%
H	55392	5.8%
P	52540	5.5%
A	50399	5.3%
G	44764	4.7%
L	44614	4.7%
Other values (95)	334493	35.2%

Other Punctuation

Value	Count	Frequency (%)
,	525726	86.4%
.	34275	5.6%
'	26738	4.4%
/	10433	1.7%
#	6037	1.0%
"	4257	0.7%
:	445	0.1%
&	268	< 0.1%
!	45	< 0.1%
?	31	< 0.1%
Other values (6)	60	< 0.1%

Decimal Number

Value	Count	Frequency (%)
1	5053	35.2%
2	4015	27.9%
3	1445	10.1%
4	748	5.2%
0	647	4.5%
9	568	4.0%
5	514	3.6%
8	490	3.4%
6	458	3.2%
7	430	3.0%

Nonspacing Mark

Value	Count	Frequency (%)
́	3	21.4%
̂	3	21.4%
̀	2	14.3%
̀	1	7.1%
̇	1	7.1%
̋	1	7.1%
̃	1	7.1%
̋	1	7.1%
̊	1	7.1%

Open Punctuation

Value	Count	Frequency (%)
(	42047	99.7%
[	121	0.3%
„	23	0.1%
‚	3	< 0.1%

Close Punctuation

Value	Count	Frequency (%)
)	42032	99.7%
]	121	0.3%
)	1	< 0.1%

Dash Punctuation

Value	Count	Frequency (%)
-	13876	99.8%
–	28	0.2%
—	1	< 0.1%

Final Punctuation

Value	Count	Frequency (%)
’	83	58.9%
»	47	33.3%
”	11	7.8%

Initial Punctuation

Value	Count	Frequency (%)
«	47	54.0%
“	33	37.9%
‘	7	8.0%

Other Symbol

Value	Count	Frequency (%)
°	24	92.3%
№	1	3.8%
®	1	3.8%

Math Symbol

Value	Count	Frequency (%)
|	6	50.0%

Value	Count	Frequency (%)
+	5	41.7%
<	1	8.3%

Modifier Symbol

Value	Count	Frequency (%)
`	28	59.6%
´	19	40.4%

Currency Symbol

Value	Count	Frequency (%)
$	13	86.7%
¢	2	13.3%

Control

Value	Count	Frequency (%)
	8	88.9%
□	1	11.1%

Format

Value	Count	Frequency (%)
	2	66.7%
	1	33.3%

Other Number

Value	Count	Frequency (%)
½	1	50.0%
²	1	50.0%

Space Separator

Value	Count	Frequency (%)
	1028323	100.0%

Connector Punctuation

Value	Count	Frequency (%)
_	7	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	5913891	77.0%
Common	1749608	22.8%
Cyrillic	14096	0.2%
Hangul	223	< 0.1%
Greek	212	< 0.1%
Arabic	156	< 0.1%
Han	117	< 0.1%
Hebrew	60	< 0.1%
Thai	26	< 0.1%
Katakana	23	< 0.1%
Other values (2)	10	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	649941	11.0%
a	519505	8.8%
r	472126	8.0%
i	419713	7.1%
n	404639	6.8%
o	360968	6.1%
t	288075	4.9%
l	273828	4.6%

Value	Count	Frequency (%)
s	242202	4.1%
d	174401	2.9%
Other values (150)	2108493	35.7%

Hangul

Value	Count	Frequency (%)
진	7	3.1%
영	6	2.7%
최	6	2.7%
동	5	2.2%
유	5	2.2%
이	5	2.2%
은	4	1.8%
정	4	1.8%
희	4	1.8%
사	4	1.8%
Other values (113)	173	77.6%

Han

Value	Count	Frequency (%)
大	5	4.3%
爸	4	3.4%
雄	4	3.4%
子	3	2.6%
蕭	2	1.7%
智	2	1.7%
心	2	1.7%
柏	2	1.7%
毒	2	1.7%
相	2	1.7%
Other values (77)	89	76.1%

Cyrillic

Value	Count	Frequency (%)
а	1497	10.6%
о	1125	8.0%
и	1040	7.4%
е	968	6.9%
н	924	6.6%
р	909	6.4%
т	631	4.5%
к	613	4.3%
л	600	4.3%
в	547	3.9%
Other values (55)	5242	37.2%

Common

Value	Count	Frequency (%)
	1028323	58.8%
,	525726	30.0%
(	42047	2.4%
)	42032	2.4%
.	34275	2.0%
'	26738	1.5%
-	13876	0.8%
/	10433	0.6%
#	6037	0.3%
1	5053	0.3%
Other values (50)	15068	0.9%

Greek

Value	Count	Frequency (%)
α	24	11.3%
ς	19	9.0%
ο	19	9.0%
ρ	14	6.6%
σ	9	4.2%
τ	8	3.8%
η	8	3.8%
ν	8	3.8%
ά	8	3.8%
λ	8	3.8%
Other values (32)	87	41.0%

Arabic

Value	Count	Frequency (%)
ا	25	16.0%
ل	17	10.9%
ي	14	9.0%
ب	12	7.7%
د	9	5.8%
ر	9	5.8%
و	8	5.1%
س	8	5.1%
ن	7	4.5%
ش	6	3.8%
Other values (17)	41	26.3%

Hebrew

Value	Count	Frequency (%)
ל	9	15.0%
א	7	11.7%
ו	7	11.7%
ה	5	8.3%
י	5	8.3%
ר	4	6.7%
ם	4	6.7%
ט	3	5.0%
ש	3	5.0%
נ	2	3.3%
Other values (9)	11	18.3%

Thai

Value	Count	Frequency (%)
เ	3	11.5%
ล	3	11.5%
ก	2	7.7%
ร	2	7.7%
น	2	7.7%
่	2	7.7%
ม	2	7.7%
็	1	3.8%
แ	1	3.8%
ค	1	3.8%
Other values (7)	7	26.9%

Katakana

Value	Count	Frequency (%)
ロ	4	17.4%
ヘ	4	17.4%
ニ	2	8.7%
ト	2	8.7%
ッ	2	8.7%

Value	Count	Frequency (%)
ク	2	8.7%
ラ	2	8.7%
マ	1	4.3%
ピ	1	4.3%
ゴ	1	4.3%
Other values (2)	2	8.7%

Inherited

Value	Count	Frequency (%)
́	3	42.9%
̂	3	42.9%
̀	1	14.3%

Hiragana

Value	Count	Frequency (%)
り	1	33.3%
お	1	33.3%
ん	1	33.3%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	7645646	99.6%
None	17870	0.2%
Cyrillic	14096	0.2%
Hangul	223	< 0.1%
Punctuation	191	< 0.1%
Arabic	156	< 0.1%
CJK	117	< 0.1%
Hebrew	60	< 0.1%
Thai	26	< 0.1%
Katakana	23	< 0.1%
Other values (4)	14	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	1028323	13.4%
e	649941	8.5%
,	525726	6.9%
a	519505	6.8%
r	472126	6.2%
i	419713	5.5%
n	404639	5.3%
o	360968	4.7%
t	288075	3.8%
l	273828	3.6%
Other values (80)	2702802	35.4%

None

Value	Count	Frequency (%)
é	4956	27.7%
è	1621	9.1%
ä	1166	6.5%
á	1004	5.6%
í	921	5.2%
ö	822	4.6%
ð	711	4.0%
ü	700	3.9%
ó	595	3.3%
ç	511	2.9%

Value	Count	Frequency (%)
Other values (149)	4863	27.2%

Cyrillic

Value	Count	Frequency (%)
а	1497	10.6%
о	1125	8.0%
и	1040	7.4%
е	968	6.9%
н	924	6.6%
р	909	6.4%
т	631	4.5%
к	613	4.3%
л	600	4.3%
в	547	3.9%
Other values (55)	5242	37.2%

Punctuation

Value	Count	Frequency (%)
’	83	43.5%
“	33	17.3%
–	28	14.7%
„	23	12.0%
”	11	5.8%
‘	7	3.7%
,	3	1.6%
	2	1.0%
—	1	0.5%

Arabic

Value	Count	Frequency (%)
ا	25	16.0%
ل	17	10.9%
ي	14	9.0%
ب	12	7.7%
د	9	5.8%
ر	9	5.8%
و	8	5.1%
س	8	5.1%
ن	7	4.5%
ش	6	3.8%
Other values (17)	41	26.3%

Hebrew

Value	Count	Frequency (%)
ל	9	15.0%
א	7	11.7%
ו	7	11.7%
ה	5	8.3%
י	5	8.3%
ר	4	6.7%
ם	4	6.7%
ט	3	5.0%
ש	3	5.0%
נ	2	3.3%
Other values (9)	11	18.3%

Hangul

Value	Count	Frequency (%)
진	7	3.1%
영	6	2.7%
최	6	2.7%
동	5	2.2%

Value	Count	Frequency (%)
유	5	2.2%
이	5	2.2%
은	4	1.8%
정	4	1.8%
희	4	1.8%
사	4	1.8%
Other values (113)	173	77.6%

CJK

Value	Count	Frequency (%)
大	5	4.3%
爸	4	3.4%
雄	4	3.4%
子	3	2.6%
蕭	2	1.7%
智	2	1.7%
心	2	1.7%
柏	2	1.7%
毒	2	1.7%
相	2	1.7%
Other values (77)	89	76.1%

Katakana

Value	Count	Frequency (%)
ロ	4	17.4%
ベ	4	17.4%
ニ	2	8.7%
ト	2	8.7%
ッ	2	8.7%
ク	2	8.7%
ラ	2	8.7%
マ	1	4.3%
ピ	1	4.3%
ゴ	1	4.3%
Other values (2)	2	8.7%

Diacriticals

Value	Count	Frequency (%)
◌́	3	42.9%
◌̂	3	42.9%
◌̀	1	14.3%

Thai

Value	Count	Frequency (%)
เ	3	11.5%
อ	3	11.5%
ก	2	7.7%
า	2	7.7%
น	2	7.7%
◌	2	7.7%
ม	2	7.7%
◌	1	3.8%
แ	1	3.8%
ค	1	3.8%
Other values (7)	7	26.9%

Hiragana

Value	Count	Frequency (%)
り	1	33.3%
お	1	33.3%
ん	1	33.3%

Letterlike Symbols

Value	Count	Frequency (%)
№	1	100.0%

Latin Ext Additional

Value	Count	Frequency (%)
ả	1	33.3%
à	1	33.3%
ỏ	1	33.3%

ActorName

Categorical

Distinct	42678
Distinct (%)	93.9%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	4551
Median length	1414
Mean length	187.80077
Min length	4

Characters and Unicode

Total characters	8539489
Distinct characters	395
Distinct categories	16 (https://en.wikipedia.org/wiki/Unicode_charac
Distinct scripts	9 (https://en.wikipedia.org/wiki/Script_(Unicode
Distinct blocks	10 (https://en.wikipedia.org/wiki/Unicode_blo

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

Unique

Unique	42472 ?
Unique (%)	93.4%

Sample

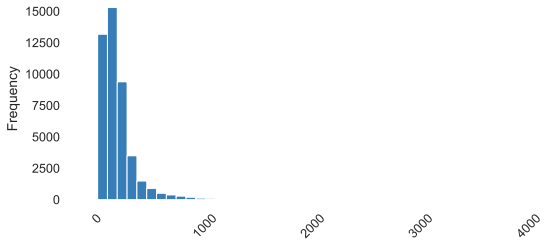
1st row	Tom Hanks, Tim Allen, Don Rickles, Jim Varney, Wallace Shawn, John Ratzenberger, Annie Potts, John Morris, Erik von Detten, Laurie Metcalf, R. Lee Ermey, Sarah Freeman, Penn Jillette
2nd row	Robin Williams, Jonathan Hyde, Kirsten Dunst, Bradley Pierce, Bonnie Hunt, Bebe Neuwirth, David Alan Grier, Patricia Clarkson, Adam Hann-Byrd, Laura Bell Bundy, James Handy, Gillian Barber, Brandon Obray, Cyrus Thiedeke, Gary Joseph Thorup, Leonard Zola, Lloyd Berry, Malcolm Stewart, Annabel Kershaw, Darryl Henriques, Robyn Driscoll, Peter Bryant, Sarah Gilson, Florica Vlad, June Lion, Brenda Lockmuller
3rd row	Walter Matthau, Jack Lemmon, Ann-Margret, Sophia Loren, Daryl Hannah, Burgess Meredith, Kevin Pollak
4th row	Whitney Houston, Angela Bassett, Loretta Devine, Lela Rochon, Gregory Hines, Dennis Haysbert, Michael Beach, Mykelti Williamson, Lamont Johnson, Wesley Snipes
5th row	Steve Martin, Diane Keaton, Martin Short, Kimberly Williams-Paisley, George Newbern, Kieran Culkin, BD Wong, Peter Michael Goetz, Kate McGregor-Stewart, Jane Adams, Eugene Levy, Lori Alan

Common Values

Value	Count	Frequency (%)
NoName	2413	5.3%
Georges Méliès	24	0.1%
Louis Theroux	15	< 0.1%
Mel Blanc	12	< 0.1%
Jimmy Carr	9	< 0.1%

Value	Count	Frequency (%)
Werner Herzog	8	< 0.1%
Louis C.K.	8	< 0.1%
George Carlin	8	< 0.1%
David Attenborough	8	< 0.1%
Trevor Noah	6	< 0.1%
Other values (42668)	42960	94.5%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
john	9809	0.8%
michael	7464	0.6%
david	6190	0.5%
robert	5725	0.5%
james	5693	0.5%
richard	4446	0.4%
paul	4320	0.4%
peter	3903	0.3%
william	3432	0.3%
george	3416	0.3%
Other values (112949)	1113657	95.3%

Most occurring characters

Value	Count	Frequency (%)
	1122712	13.1%
a	707745	8.3%
e	668082	7.8%
n	524439	6.1%
,	519745	6.1%
r	497639	5.8%
i	484270	5.7%
o	426424	5.0%
l	366664	4.3%
s	256009	3.0%
Other values (385)	2965760	34.7%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	5663878	66.3%
Uppercase Letter	1195902	14.0%
Space Separator	1122715	13.1%
Other Punctuation	542058	6.3%
Dash Punctuation	14112	0.2%
Other Letter	543	< 0.1%
Decimal Number	94	< 0.1%
Final Punctuation	83	< 0.1%
Initial Punctuation	23	< 0.1%
Open Punctuation	23	< 0.1%
Other values (6)	58	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
a	707745	12.5%
e	668082	11.8%
n	524439	9.3%
r	497639	8.8%
i	484270	8.6%
o	426424	7.5%
l	366664	6.5%
s	256009	4.5%
t	253361	4.5%
h	198021	3.5%
Other values (138)	1281224	22.6%

Other Letter

Value	Count	Frequency (%)
ا	32	5.9%
م	31	5.7%
ع	19	3.5%
ى	19	3.5%
ن	18	3.3%
松	17	3.1%
ر	17	3.1%
د	17	3.1%
ي	16	2.9%
美	12	2.2%
Other values (104)	345	63.5%

Uppercase Letter

Value	Count	Frequency (%)
M	109410	9.1%
S	92377	7.7%
C	84052	7.0%
J	83374	7.0%
B	82422	6.9%
A	70859	5.9%
R	67418	5.6%
D	65916	5.5%
L	61183	5.1%
G	54690	4.6%
Other values (81)	424201	35.5%

Decimal Number

Value	Count	Frequency (%)
5	37	39.4%
0	29	30.9%
2	8	8.5%
1	8	8.5%
9	4	4.3%
4	2	2.1%
3	2	2.1%
7	2	2.1%
6	1	1.1%
8	1	1.1%

Other Punctuation

Value	Count	Frequency (%)
,	519745	95.9%
.	16060	3.0%
'	6097	1.1%
"	129	< 0.1%
·	9	< 0.1%

Value	Count	Frequency (%)
:	6	< 0.1%
&	6	< 0.1%
!	5	< 0.1%
/	1	< 0.1%

Nonspacing Mark

Value	Count	Frequency (%)
́	10	58.8%
̃	2	11.8%
̈	1	5.9%
̀	1	5.9%
̄	1	5.9%
̂	1	5.9%
̆	1	5.9%

Final Punctuation

Value	Count	Frequency (%)
’	74	89.2%
”	6	7.2%
»	3	3.6%

Space Separator

Value	Count	Frequency (%)
	1122712	> 99.9%
	3	< 0.1%

Initial Punctuation

Value	Count	Frequency (%)
“	20	87.0%
«	3	13.0%

Open Punctuation

Value	Count	Frequency (%)
„	14	60.9%
(	9	39.1%

Format

Value	Count	Frequency (%)
	5	83.3%
	1	16.7%

Dash Punctuation

Value	Count	Frequency (%)
-	14112	100.0%

Control

Value	Count	Frequency (%)
	21	100.0%

Close Punctuation

Value	Count	Frequency (%)
)	9	100.0%

Currency Symbol

Value	Count	Frequency (%)
$	3	100.0%

Modifier Symbol

Value	Count	Frequency (%)
ˊ	2	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	6856696	80.3%
Common	1679148	19.7%
Cyrillic	3070	< 0.1%
Han	276	< 0.1%
Arabic	241	< 0.1%
Thai	27	< 0.1%
Greek	14	< 0.1%
Inherited	11	< 0.1%
Hangul	6	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
a	707745	10.3%
e	668082	9.7%
n	524439	7.6%
r	497639	7.3%
i	484270	7.1%
o	426424	6.2%
l	366664	5.3%
s	256009	3.7%
t	253361	3.7%
h	198021	2.9%
Other values (163)	2474042	36.1%

Han

Value	Count	Frequency (%)
松	17	6.2%
美	12	4.3%
田	11	4.0%
龙	11	4.0%
平	11	4.0%
长	11	4.0%
泽	11	4.0%
雅	11	4.0%
森	9	3.3%
杰	9	3.3%
Other values (55)	163	59.1%

Cyrillic

Value	Count	Frequency (%)
а	323	10.5%
и	315	10.3%
о	233	7.6%
н	229	7.5%
р	215	7.0%
е	174	5.7%
л	155	5.0%
к	136	4.4%
т	115	3.7%
с	109	3.6%
Other values (51)	1066	34.7%

Common

Value	Count	Frequency (%)
	1122712	66.9%
,	519745	31.0%
.	16060	1.0%

Value	Count	Frequency (%)
-	14112	0.8%
'	6097	0.4%
"	129	< 0.1%
’	74	< 0.1%
5	37	< 0.1%
0	29	< 0.1%
	21	< 0.1%
Other values (24)	132	< 0.1%

Arabic

Value	Count	Frequency (%)
ا	32	13.3%
م	31	12.9%
ع	19	7.9%
ى	19	7.9%
ن	18	7.5%
ر	17	7.1%
د	17	7.1%
ي	16	6.6%
ل	9	3.7%
ب	8	3.3%
Other values (18)	55	22.8%

Thai

Value	Count	Frequency (%)
ว	2	7.4%
น	2	7.4%
ง	2	7.4%
้	2	7.4%
ร	2	7.4%
า	2	7.4%
ี	1	3.7%
ส	1	3.7%
ด	1	3.7%
ธ	1	3.7%
Other values (11)	11	40.7%

Hangul

Value	Count	Frequency (%)
조	1	16.7%
병	1	16.7%
만	1	16.7%
강	1	16.7%
계	1	16.7%
열	1	16.7%

Greek

Value	Count	Frequency (%)
v	6	42.9%
Z	2	14.3%
α	2	14.3%
í	2	14.3%
o	2	14.3%

Inherited

Value	Count	Frequency (%)
́	10	90.9%
	1	9.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	8497394	99.5%

Value	Count	Frequency (%)
None	38289	0.4%
Cyrillic	3070	< 0.1%
CJK	276	< 0.1%
Arabic	241	< 0.1%
Punctuation	120	< 0.1%
Latin Ext Additional	56	< 0.1%
Thai	27	< 0.1%
Diacriticals	10	< 0.1%
Hangul	6	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	1122712	13.2%
a	707745	8.3%
e	668082	7.9%
n	524439	6.2%
,	519745	6.1%
r	497639	5.9%
i	484270	5.7%
o	426424	5.0%
l	366664	4.3%
s	256009	3.0%
Other values (66)	2923665	34.4%

None

Value	Count	Frequency (%)
é	9088	23.7%
á	4156	10.9%
í	2756	7.2%
ő	2332	6.1%
ö	2025	5.3%
ó	1882	4.9%
ü	1495	3.9%
ć	1360	3.6%
è	1243	3.2%
ä	996	2.6%
Other values (111)	10956	28.6%

Cyrillic

Value	Count	Frequency (%)
а	323	10.5%
и	315	10.3%
о	233	7.6%
н	229	7.5%
р	215	7.0%
е	174	5.7%
л	155	5.0%
к	136	4.4%
т	115	3.7%
с	109	3.6%
Other values (51)	1066	34.7%

Punctuation

Value	Count	Frequency (%)
’	74	61.7%
“	20	16.7%
„	14	11.7%
”	6	5.0%
	5	4.2%

Value	Count	Frequency (%)
	1	0.8%

Arabic

Value	Count	Frequency (%)
ا	32	13.3%
م	31	12.9%
ع	19	7.9%
ى	19	7.9%
ن	18	7.5%
ر	17	7.1%
د	17	7.1%
ي	16	6.6%
ل	9	3.7%
ب	8	3.3%
Other values (18)	55	22.8%

CJK

Value	Count	Frequency (%)
松	17	6.2%
美	12	4.3%
田	11	4.0%
龙	11	4.0%
平	11	4.0%
长	11	4.0%
泽	11	4.0%
雅	11	4.0%
森	9	3.3%
杰	9	3.3%
Other values (55)	163	59.1%

Latin Ext Additional

Value	Count	Frequency (%)
ẽ	15	26.8%
ạ	9	16.1%
ỳ	6	10.7%
ị	6	10.7%
ế	5	8.9%
à	4	7.1%
ỏ	4	7.1%
ề	4	7.1%
ầ	2	3.6%
ố	1	1.8%

Diacriticals

Value	Count	Frequency (%)
◌́	10	100.0%

Thai

Value	Count	Frequency (%)
ว	2	7.4%
น	2	7.4%
ง	2	7.4%
◌	2	7.4%
ร	2	7.4%
า	2	7.4%
◌	1	3.7%
ส	1	3.7%
ด	1	3.7%
ธ	1	3.7%
Other values (11)	11	40.7%

Hangul

Value	Count	Frequency (%)
조	1	16.7%
병	1	16.7%
만	1	16.7%
강	1	16.7%
계	1	16.7%
열	1	16.7%

Department
Categorical

Distinct	23604
Distinct (%)	51.9%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	4109
Median length	1694
Mean length	99.478789
Min length	3

Characters and Unicode

Total characters	4523400	
Distinct characters	35	
Distinct categories	5 (https://en.wikipedia.org/wiki/Unicode_charac ... ategory)	?
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode ... code)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_bloc	?

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

Unique

Unique	22441	?
Unique (%)	49.4%	

Sample

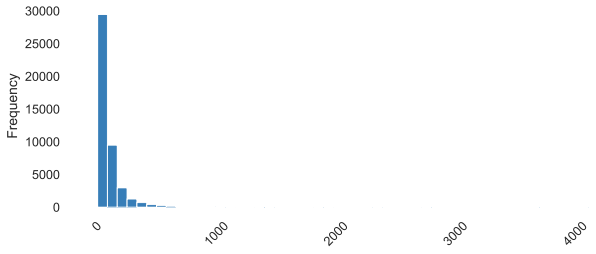
1st row	Directing, Writing, Writing, Writing, Writing, Production, Production, Production, Production, Editing, Art, Editing, Sound, Visual Effects, Sound, Sound, Editing, Visual Effects, Writing, Writing, Writing, Crew, Art, Visual Effects, Sound, Directing, Sound, Sound, Editing, Editing, Art, Crew, Sound, Writing, Crew, Sound, Production, Crew, Visual Effects, Lighting, Lighting, Visual Effects, Art, Visual Effects, Visual Effects, Visual Effects, Visual Effects, Visual Effects, Visual Effects, Visual Effects, Visual Effects, Directing, Art, Art, Art, Editing, Editing, Editing, Editing, Editing, Editing, Production, Production, Crew, Sound, Sound, Sound, Sound, Crew, Sound, Sound, Sound, Sound, Production, Production
2nd row	Production, Writing, Sound, Directing, Editing, Production, Visual Effects, Art, Production, Production, Production, Camera, Writing, Production, Writing, Writing

3rd row	Directing, Writing, Writing, Crew
4th row	Directing, Writing, Production, Production, Production, Writing, Production, Writing, Sound, Production
5th row	Sound, Camera, Writing, Production, Writing, Directing, Editing

Common Values

Value	Count	Frequency (%)
Directing	5014	11.0%
Directing, Writing	4270	9.4%
Directing, Writing, Writing	2318	5.1%
Writing, Directing	929	2.0%
NoDepartment	771	1.7%
Writing, Directing, Writing	731	1.6%
Directing, Writing, Writing, Writing	625	1.4%
Directing, Directing	569	1.3%
Writing, Writing, Directing	282	0.6%
Writing, Directing, Writing, Writing	220	0.5%
Other values (23594)	29742	65.4%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
production	94498	17.4%
writing	74831	13.8%
directing	58129	10.7%
sound	50605	9.3%
art	40694	7.5%
camera	33539	6.2%
crew	31605	5.8%
costume	30850	5.7%
	30850	5.7%
make-up	30850	5.7%
Other values (6)	65190	12.0%

Most occurring characters

Value	Count	Frequency (%)
	496170	11.0%
i	444635	9.8%
,	419609	9.3%
t	350102	7.7%
r	334086	7.4%
n	313512	6.9%
o	271241	6.0%
e	201376	4.5%
u	190814	4.2%
d	174934	3.9%
Other values (25)	1326921	29.3%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	3003509	66.4%
Uppercase Letter	542412	12.0%
Space Separator	496170	11.0%
Other Punctuation	450459	10.0%
Dash Punctuation	30850	0.7%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
i	444635	14.8%
t	350102	11.7%
r	334086	11.1%
n	313512	10.4%
o	271241	9.0%
e	201376	6.7%
u	190814	6.4%
d	174934	5.8%
g	172485	5.7%
c	167507	5.6%
Other values (9)	382817	12.7%

Uppercase Letter

Value	Count	Frequency (%)
C	95994	17.7%
P	94498	17.4%
W	74831	13.8%
D	58900	10.9%
S	50605	9.3%
E	44692	8.2%
A	40713	7.5%
M	30850	5.7%
U	30850	5.7%
V	14861	2.7%
Other values (2)	5618	1.0%

Other Punctuation

Value	Count	Frequency (%)
,	419609	93.2%
&	30850	6.8%

Space Separator

Value	Count	Frequency (%)
	496170	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	30850	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	3545921	78.4%
Common	977479	21.6%

Most frequent character per script

Latin

Value	Count	Frequency (%)
i	444635	12.5%
t	350102	9.9%

Value	Count	Frequency (%)
r	334086	9.4%
n	313512	8.8%
o	271241	7.6%
e	201376	5.7%
u	190814	5.4%
d	174934	4.9%
g	172485	4.9%
c	167507	4.7%
Other values (21)	925229	26.1%

Common

Value	Count	Frequency (%)
	496170	50.8%
,	419609	42.9%
&	30850	3.2%
-	30850	3.2%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	4523400	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	496170	11.0%
i	444635	9.8%
,	419609	9.3%
t	350102	7.7%
r	334086	7.4%
n	313512	6.9%
o	271241	6.0%
e	201376	4.5%
u	190814	4.2%
d	174934	3.9%
Other values (25)	1326921	29.3%

CrewJob
Categorical

Distinct	26602
Distinct (%)	58.5%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	6076
Median length	2806
Mean length	147.47954
Min length	5

Characters and Unicode

Total characters	6706042
Distinct characters	62
Distinct categories	8 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)
Distinct scripts	2 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)

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

Unique

Unique	25513
Unique (%)	56.1%

Sample

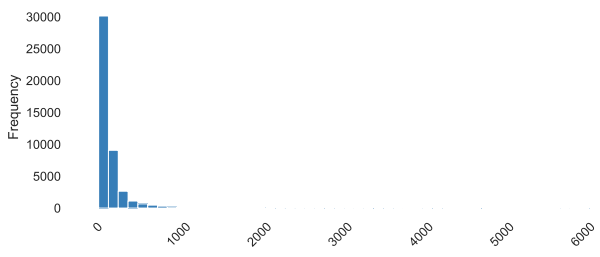
1st row	Director,
	Screenplay,
	Screenplay,
	Screenplay,
	Screenplay,
	Producer,
	Executive Producer,
	Producer,
	Executive Producer, Editor,
	Art Direction,
	Editor, Foley Editor,
	Animation, ADR Editor,
	Orchestrator,
	Color Timer, CG Painter, Original Story, Original Story, Original Story, Post Production Supervisor,
	Sculptor,
	Animation Director, Music,
	Layout,
	Orchestrator,
	Music Editor,
	Negative Cutter,
	Negative Cutter,
	Title Designer,
	Supervising Technical Director, Songs, Original Story,
	Supervising Animator, Sound Designer,
	Production Supervisor,
	Executive Music Producer,
	Animation Director,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Animation,
	Visual Effects Supervisor,
	Visual Effects,
	Visual Effects,
	Visual Effects,
	Visual Effects,
	Visual Effects,
	Lighting Supervisor,
	Lighting Supervisor, CG Painter, Sculptor,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Character Designer,
	Layout, Set Dresser, Set Dresser, Set Dresser,
	Editorial Manager,
	Assistant Editor,
	Assistant Editor,
	Assistant Editor,
	Assistant Editor,
	Editorial Coordinator,
	Production Coordinator,

	Production Coordinator, Unit Publicist, Sound Re-Recording Mixer, Sound Re-Recording Mixer, Supervising Sound Editor, Sound Effects Editor, Sound Design Assistant, Assistant Sound Editor, Assistant Sound Editor, Assistant Sound Editor, Assistant Sound Editor, Casting Consultant, ADR Voice Casting
2nd row	Executive Producer, Screenplay, Original Music Composer, Director, Editor, Casting, Animation Supervisor, Production Design, Producer, Executive Producer, Executive Producer, Director of Photography, Novel, Producer, Screenplay, Screenplay
3rd row	Director, Characters, Writer, Sound Recordist
4th row	Director, Screenplay, Producer, Producer, Producer, Screenplay, Executive Producer, Novel, Original Music Composer, Producer
5th row	Original Music Composer, Director of Photography, Screenplay, Producer, Screenplay, Director, Editor

Common Values

Value	Count	Frequency (%)
Director	5014	11.0%
Director, Writer	3541	7.8%
Director, Writer, Writer	1260	2.8%
NoJob	771	1.7%
Director, Director	559	1.2%
Writer, Director	539	1.2%
Director, Screenplay	521	1.1%
Screenplay, Director	341	0.7%
Writer, Director, Writer	267	0.6%
Director, Screenplay, Screenplay	209	0.5%
Other values (26592)	32449	71.4%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
director	76977	10.1%
producer	70782	9.3%
editor	39106	5.1%
writer	30866	4.1%
music	27392	3.6%
screenplay	25163	3.3%
design	21921	2.9%
of	21153	2.8%
photography	21140	2.8%
production	19300	2.5%
Other values (323)	408030	53.6%

Most occurring characters

Value	Count	Frequency (%)
r	721755	10.8%
	716359	10.7%
e	528513	7.9%
o	521394	7.8%
i	477721	7.1%
t	464426	6.9%
,	419609	6.3%
c	317118	4.7%
a	235240	3.5%
n	231527	3.5%
Other values (52)	2072380	30.9%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4808684	71.7%
Uppercase Letter	752229	11.2%
Space Separator	716359	10.7%
Other Punctuation	420724	6.3%
Dash Punctuation	7776	0.1%
Decimal Number	258	< 0.1%
Open Punctuation	6	< 0.1%
Close Punctuation	6	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
r	721755	15.0%
e	528513	11.0%
o	521394	10.8%
i	477721	9.9%
t	464426	9.7%
c	317118	6.6%
a	235240	4.9%
n	231527	4.8%
u	229274	4.8%
s	228183	4.7%
Other values (16)	853533	17.7%

Uppercase Letter

Value	Count	Frequency (%)
D	138116	18.4%
P	123731	16.4%
S	105349	14.0%
C	79019	10.5%

Value	Count	Frequency (%)
E	77287	10.3%
A	51569	6.9%
M	51311	6.8%
W	31501	4.2%
O	25064	3.3%
R	11383	1.5%
Other values (15)	57899	7.7%

Other Punctuation

Value	Count	Frequency (%)
,	419609	99.7%
&	977	0.2%
/	105	< 0.1%
'	33	< 0.1%

Decimal Number

Value	Count	Frequency (%)
3	205	79.5%
2	35	13.6%
4	18	7.0%

Space Separator

Value	Count	Frequency (%)
	716359	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	7776	100.0%

Open Punctuation

Value	Count	Frequency (%)
(	6	100.0%

Close Punctuation

Value	Count	Frequency (%)
)	6	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	5560913	82.9%
Common	1145129	17.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
r	721755	13.0%
e	528513	9.5%
o	521394	9.4%
i	477721	8.6%
t	464426	8.4%
c	317118	5.7%
a	235240	4.2%
n	231527	4.2%
u	229274	4.1%
s	228183	4.1%
Other values (41)	1605762	28.9%

Common

Value	Count	Frequency (%)
	716359	62.6%
,	419609	36.6%
-	7776	0.7%

Value	Count	Frequency (%)
&	977	0.1%
3	205	< 0.1%
/	105	< 0.1%
2	35	< 0.1%
'	33	< 0.1%
4	18	< 0.1%
(	6	< 0.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	6706042	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
r	721755	10.8%
	716359	10.7%
e	528513	7.9%
o	521394	7.8%
i	477721	7.1%
t	464426	6.9%
,	419609	6.3%
c	317118	4.7%
a	235240	3.5%
n	231527	3.5%
Other values (52)	2072380	30.9%

CrewName

Categorical

Distinct	42945
Distinct (%)	94.4%
Missing	0
Missing (%)	0.0%
Memory size	355.4 KiB

Length

Max length	6778
Median length	2186
Mean length	157.28851
Min length	3

Characters and Unicode

Total characters	7152066
Distinct characters	357
Distinct categories	15 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category) ?
Distinct scripts	8 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode) ?
Distinct blocks	9 (https://en.wikipedia.org/wiki/Unicode_block) ?

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

Unique

Unique	41823 ?
Unique (%)	92.0%

Sample

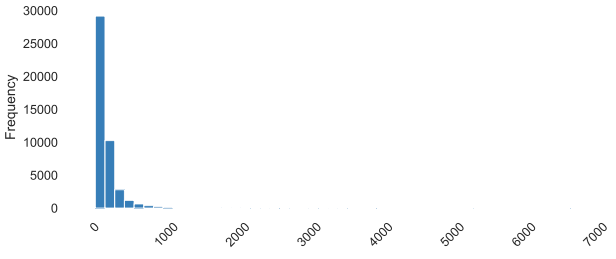
1st row	John Lasseter, Joss Whedon, Andrew Stanton, Joel Cohen, Alec Sokolow, Bonnie Arnold, Ed Catmull, Ralph Guggenheim, Steve Jobs, Lee Unkrich, Ralph Eggleston, Robert Gordon, Mary Helen Leasman, Kim Blanchette, Marilyn McCoppen, Randy Newman, Dale E. Grahn, Robin Cooper, John Lasseter, Pete Docter, Joe Ranft, Patsy Bouge, Norm DeCarlo, Ash Brannon, Randy Newman, Roman Figun, Don Davis, James Flamberg, Mary Beth Smith, Rick Mackay, Susan Bradley, William Reeves, Randy Newman, Andrew Stanton, Pete Docter, Gary Rydstrom, Karen Robert Jackson, Chris Montan, Rich Quade, Michael Berenstein, Colin Brady, Davey Crockett Feiten, Angie Glocka, Rex Grignon, Tom K. Gurney, Jimmy Hayward, Hal T. Hickel, Karen Kiser, Anthony B. LaMolinara, Guionne Leroy, Bud Luckey, Les Major, Glenn McQueen, Mark Oftedal, Jeff Pidgeon, Jeff Pratt, Steve Rabatich, Roger Rose, Steve Segal, Doug Sheppeck, Alan Sperling, Doug Sweetland, David Tart, Ken Willard, Thomas Porter, Mark Thomas Henne, Oren Jacob, Darwyn Peachey, Mitch Prater, Brian M. Rosen, Sharon Calahan, Galyn Susman, William Cone, Shelley Daniels Lekven, Bob Pauley, Bud Luckey, Andrew Stanton, William Cone, Steve Johnson, Dan Haskett, Tom Holloway, Jean Gillmore, Desirée Mourad, Kelly O'Connell, Sonoko Konishi, Ann M. Rockwell, Julie M. McDonald, Robin Lee, Tom Freeman, Ada Cochavi, Dana Mulligan, Deirdre Morrison, Lori Lombardo, Ellen Devine, Lauren Beth Strogoff, Gary Rydstrom, Gary Summers, Tim Holland, Pat Jackson, Tom Myers, J.R. Grubbs, Susan Sanford, Susan Popovic, Dan Engstrom, Ruth Lambert, Mickie McGowan

2nd row	Larry J. Franco, Jonathan Hensleigh, James Horner, Joe Johnston, Robert Dalva, Nancy Foy, Kyle Balda, James D. Bissell, Scott Kroopf, Ted Field, Robert W. Cort, Thomas E. Ackerman, Chris van Allsburg, William Teitler, Greg Taylor, Jim Strain
3rd row	Howard Deutch, Mark Steven Johnson, Mark Steven Johnson, Jack Keller
4th row	Forest Whitaker, Ronald Bass, Ronald Bass, Ezra Swerdlow, Deborah Schindler, Terry McMillan, Terry McMillan, Terry McMillan, Kenneth Edmonds, Caron K
5th row	Alan Silvestri, Elliot Davis, Nancy Meyers, Nancy Meyers, Albert Hackett, Charles Shyer, Adam Bernardi

Common Values

Value	Count	Frequency (%)
NoCrew	771	1.7%
Georges Méliès	34	0.1%
Christian I. Nyby II	13	< 0.1%
Gerald Thomas, Talbot Rothwell	13	< 0.1%
Frederick Wiseman	12	< 0.1%
Charlie Chaplin, Charlie Chaplin	12	< 0.1%
James Benning	10	< 0.1%
Stan Brakhage	10	< 0.1%
James H. White	9	< 0.1%
William K.L. Dickson , William Heise	9	< 0.1%
Other values (42935)	44578	98.0%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
john	10082	1.0%
david	8665	0.9%
michael	8236	0.8%
robert	6770	0.7%
james	5024	0.5%
paul	4535	0.5%
peter	4500	0.5%
richard	4365	0.4%
mark	4231	0.4%
william	3967	0.4%
Other values (90023)	929166	93.9%

Most occurring characters

Value	Count	Frequency (%)
	944117	13.2%
e	558512	7.8%
a	557624	7.8%
r	434711	6.1%
n	430921	6.0%
,	419756	5.9%
i	401738	5.6%
o	355718	5.0%
l	296182	4.1%
s	220942	3.1%
Other values (347)	2531845	35.4%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4723998	66.1%
Uppercase Letter	1009697	14.1%
Space Separator	944117	13.2%
Other Punctuation	463664	6.5%
Dash Punctuation	10100	0.1%
Other Letter	206	< 0.1%
Control	193	< 0.1%
Decimal Number	51	< 0.1%
Open Punctuation	11	< 0.1%
Close Punctuation	11	< 0.1%
Other values (5)	18	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	558512	11.8%
a	557624	11.8%
r	434711	9.2%
n	430921	9.1%
i	401738	8.5%
o	355718	7.5%
l	296182	6.3%
s	220942	4.7%
t	216278	4.6%
h	170645	3.6%
Other values (128)	1080727	22.9%

Other Letter

Value	Count	Frequency (%)
ㅣ	9	4.4%
진	8	3.9%
이	7	3.4%
ㅏ	7	3.4%
연	6	2.9%
정	6	2.9%
아	5	2.4%
ㅓ	5	2.4%
조	4	1.9%
ㄷ	4	1.9%
Other values (88)	145	70.4%

Uppercase Letter

Value	Count	Frequency (%)
M	90894	9.0%
S	84270	8.3%

Value	Count	Frequency (%)
J	74448	7.4%
B	69337	6.9%
C	67628	6.7%
R	60435	6.0%
A	60364	6.0%
D	57887	5.7%
L	50382	5.0%
G	48886	4.8%
Other values (80)	345166	34.2%

Other Punctuation

Value	Count	Frequency (%)
,	419756	90.5%
.	40328	8.7%
'	3522	0.8%
"	38	< 0.1%
&	8	< 0.1%
!	4	< 0.1%
/	3	< 0.1%
:	2	< 0.1%
·	2	< 0.1%
@	1	< 0.1%

Decimal Number

Value	Count	Frequency (%)
5	16	31.4%
0	12	23.5%
9	7	13.7%
8	5	9.8%
3	4	7.8%
7	3	5.9%
2	2	3.9%
1	2	3.9%

Dash Punctuation

Value	Count	Frequency (%)
-	10097	> 99.9%
–	3	< 0.1%

Final Punctuation

Value	Count	Frequency (%)
’	8	80.0%
”	2	20.0%

Nonspacing Mark

Value	Count	Frequency (%)
̃	2	50.0%
́	2	50.0%

Space Separator

Value	Count	Frequency (%)
	944117	100.0%

Control

Value	Count	Frequency (%)
	193	100.0%

Open Punctuation

Value	Count	Frequency (%)
(	11	100.0%

Close Punctuation

Value	Count	Frequency (%)
)	11	100.0%

Initial Punctuation

Value	Count	Frequency (%)
“	2	100.0%

Math Symbol

Value	Count	Frequency (%)
|	1	100.0%

Modifier Symbol

Value	Count	Frequency (%)
ˊ	1	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	5732655	80.2%
Common	1418162	19.8%
Cyrillic	1006	< 0.1%
Hangul	133	< 0.1%
Arabic	52	< 0.1%
Greek	33	< 0.1%
Han	21	< 0.1%
Inherited	4	< 0.1%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	558512	9.7%
a	557624	9.7%
r	434711	7.6%
n	430921	7.5%
i	401738	7.0%
o	355718	6.2%
l	296182	5.2%
s	220942	3.9%
t	216278	3.8%
h	170645	3.0%
Other values (145)	2089384	36.4%

Hangul

Value	Count	Frequency (%)
진	8	6.0%
이	7	5.3%
연	6	4.5%
정	6	4.5%
아	5	3.8%
조	4	3.0%
성	4	3.0%
모	4	3.0%
현	4	3.0%
박	4	3.0%
Other values (58)	81	60.9%

Cyrillic

Value	Count	Frequency (%)
и	116	11.5%
a	92	9.1%
p	72	7.2%
o	66	6.6%
e	58	5.8%

Value	Count	Frequency (%)
л	56	5.6%
н	54	5.4%
к	54	5.4%
с	45	4.5%
в	44	4.4%
Other values (42)	349	34.7%

Common

Value	Count	Frequency (%)
	944117	66.6%
,	419756	29.6%
.	40328	2.8%
-	10097	0.7%

Value	Count	Frequency (%)
'	3522	0.2%
	193	< 0.1%
"	38	< 0.1%
5	16	< 0.1%
0	12	< 0.1%
(	11	< 0.1%
Other values (20)	72	< 0.1%

Greek

Value	Count	Frequency (%)
ς	4	12.1%
η	3	9.1%
α	3	9.1%
μ	2	6.1%
ά	2	6.1%
Α	2	6.1%
ρ	2	6.1%
ι	2	6.1%
Γ	2	6.1%
Φ	1	3.0%
Other values (10)	10	30.3%

Arabic

Value	Count	Frequency (%)
ا	9	17.3%
م	7	13.5%
د	5	9.6%
ع	4	7.7%
ى	4	7.7%
ي	4	7.7%
ح	3	5.8%
ل	3	5.8%
ن	3	5.8%
و	2	3.8%
Other values (7)	8	15.4%

Han

Value	Count	Frequency (%)
塩	2	9.5%
谷	2	9.5%
直	2	9.5%
義	2	9.5%
玛	2	9.5%
莫	2	9.5%
森	2	9.5%
杰	2	9.5%
中	1	4.8%
村	1	4.8%
Other values (3)	3	14.3%

Inherited

Value	Count	Frequency (%)
̃	2	50.0%
́	2	50.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	7122717	99.6%
None	28111	0.4%
Cyrillic	1006	< 0.1%
Hangul	133	< 0.1%

Value	Count	Frequency (%)
Arabic	52	< 0.1%
CJK	21	< 0.1%
Punctuation	15	< 0.1%
Latin Ext Additional	7	< 0.1%
Diacriticals	4	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	944117	13.3%
e	558512	7.8%
a	557624	7.8%
r	434711	6.1%
n	430921	6.0%
,	419756	5.9%
i	401738	5.6%
o	355718	5.0%
l	296182	4.2%
s	220942	3.1%
Other values (65)	2502496	35.1%

None

Value	Count	Frequency (%)
é	7410	26.4%
á	3264	11.6%
í	2068	7.4%
ó	1800	6.4%
ö	1659	5.9%
ő	1414	5.0%
ü	965	3.4%
è	900	3.2%
ç	849	3.0%
ä	776	2.8%
Other values (113)	7006	24.9%

Cyrillic

Value	Count	Frequency (%)
и	116	11.5%
а	92	9.1%
р	72	7.2%
о	66	6.6%
е	58	5.8%
л	56	5.6%
н	54	5.4%
к	54	5.4%
с	45	4.5%
в	44	4.4%
Other values (42)	349	34.7%

Arabic

Value	Count	Frequency (%)
ا	9	17.3%
م	7	13.5%
د	5	9.6%
ع	4	7.7%
ى	4	7.7%
ي	4	7.7%
ح	3	5.8%
ل	3	5.8%
ن	3	5.8%

Value	Count	Frequency (%)
ﻭ	2	3.8%
Other values (7)	8	15.4%

Hangul

Value	Count	Frequency (%)
진	8	6.0%
이	7	5.3%
연	6	4.5%
정	6	4.5%
아	5	3.8%
조	4	3.0%
성	4	3.0%
모	4	3.0%
현	4	3.0%
박	4	3.0%
Other values (58)	81	60.9%

Punctuation

Value	Count	Frequency (%)
’	8	53.3%
–	3	20.0%
“	2	13.3%
”	2	13.3%

Latin Ext Additional

Value	Count	Frequency (%)
ě	5	71.4%
ả	1	14.3%
ạ	1	14.3%

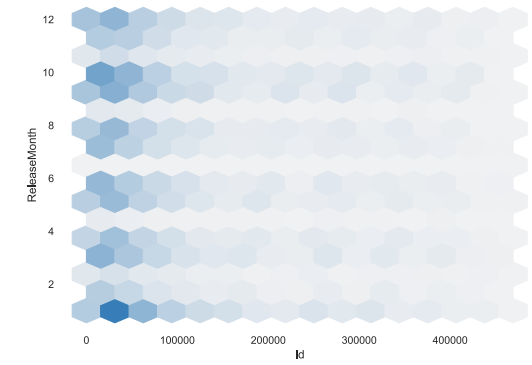
CJK

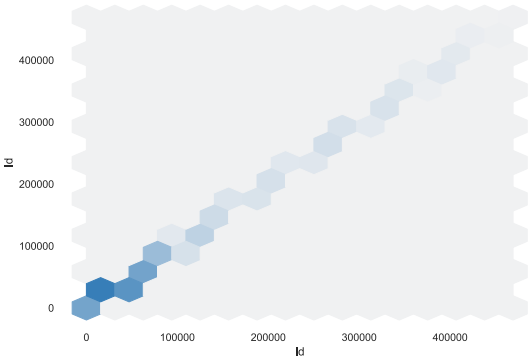
Value	Count	Frequency (%)
塩	2	9.5%
谷	2	9.5%
直	2	9.5%
義	2	9.5%
玛	2	9.5%
莫	2	9.5%
森	2	9.5%
杰	2	9.5%
中	1	4.8%
村	1	4.8%
Other values (3)	3	14.3%

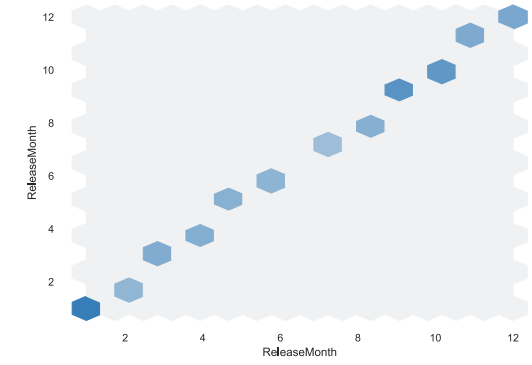
Diacriticals

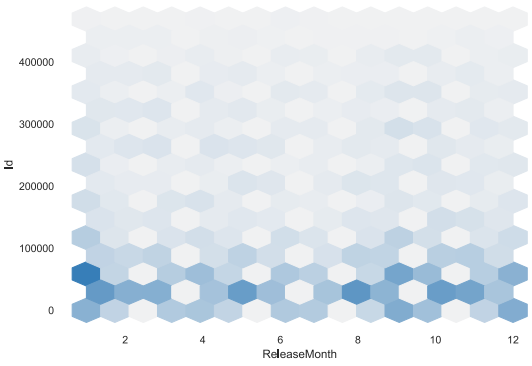
Value	Count	Frequency (%)
̃	2	50.0%
́	2	50.0%

Interactions

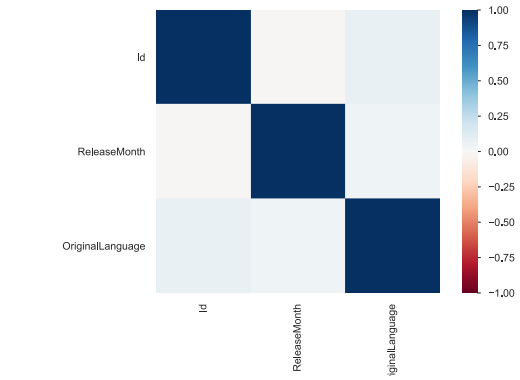






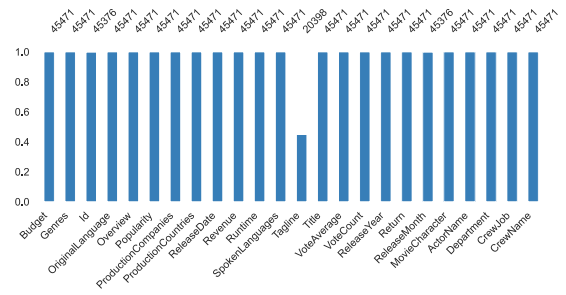


Correlations

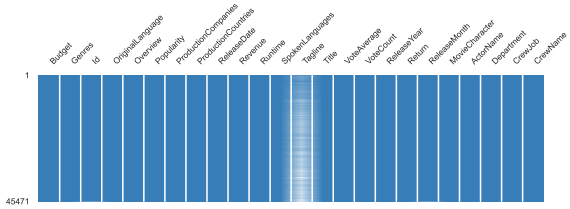


	Id	ReleaseMonth	OriginalLanguage
Id	1.000	-0.014	0.071
ReleaseMonth	-0.014	1.000	0.047
OriginalLanguage	0.071	0.047	1.000

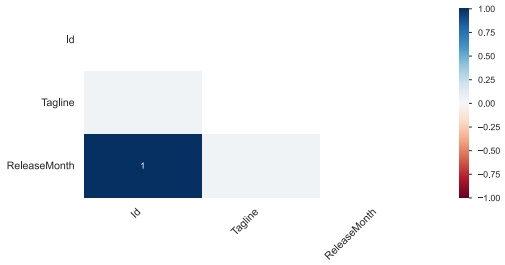
Missing values



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.



The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.

Sample

	Budget	Genres	Id	OriginalLanguage	Overview
0	30000000.0	Animation, Comedy, Family	862.0	en	Led by Woody, And
1	65000000.0	Adventure, Fantasy, Family	8844.0	en	When siblings Judy

	Budget	Genres	Id	OriginalLanguage	Overview
2	0.0	Romance, Comedy	15602.0	en	A family wedding r
3	16000000.0	Comedy, Drama, Romance	31357.0	en	Cheated on, mistre
4	0.0	Comedy	11862.0	en	Just when George
5	60000000.0	Action, Crime, Drama, Thriller	949.0	en	Obsessive master
6	58000000.0	Comedy, Romance	11860.0	en	An ugly duckling h
7	0.0	Action, Adventure, Drama, Family	45325.0	en	A mischievous you
8	35000000.0	Action, Adventure, Thriller	9091.0	en	International actior
9	58000000.0	Adventure, Action, Thriller	710.0	en	James Bond must

	Budget	Genres	Id	OriginalLanguage	Overview	Popularity	ProductionC
45461	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45462	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45463	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45464	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45465	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45466	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45467	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45468	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45469	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue
45470	NoBudget	NoGenre	NaN	NoLanguage	NoOverview	NoPopularity	MissingValue