Overview

Dataset statistics

Number of variables	10
Number of observations	1000386
Missing cells	1239
Missing cells (%)	< 0.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	116.2 MiB
Average record size in memory	121.8 B

Variable types

Numeric	5
Categorical	5

Alerts

`Zip-code` has a high cardinality: 3439 distinct values	High cardinality
`Title` has a high cardinality: 3883 distinct values	High cardinality
`Genres` has a high cardinality: 301 distinct values	High cardinality
`UserID` is highly overall correlated with `Timestamp`	High correlation
`Timestamp` is highly overall correlated with `UserID`	High correlation
`Occupation` has 130499 (13.0%) zeros	Zeros

Reproduction

Analysis started	2023-02-18 17:58:05.885449
Analysis finished	2023-02-18 17:58:27.465266
Duration	21.58 seconds
Software version	pandas-profiling v0.0.dev0 (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%22creator%22%3A%20%22%22%2C%20%22author%22%3A%20%22%22%2

Variables

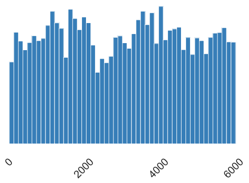
Select Columns ▼

UserID

Real number (ℝ)

Distinct	6040
Distinct (%)	0.6%
Missing	177
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	3024.5123

Minimum	1
Maximum	6040
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	15.3 MiB

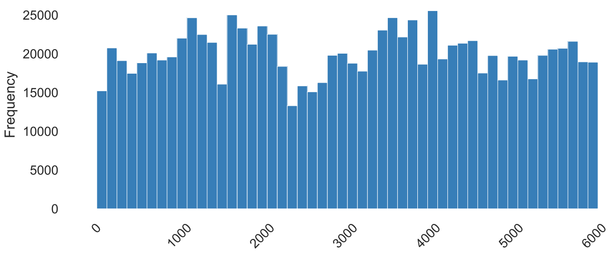


Quantile statistics

Minimum	1
5-th percentile	331
Q1	1506
median	3070
Q3	4476
95-th percentile	5740
Maximum	6040
Range	6039
Interquartile range (IQR)	2970

Descriptive statistics

Standard deviation	1728.4127
Coefficient of variation (CV)	0.57146822
Kurtosis	-1.2009951
Mean	3024.5123
Median Absolute Deviation (MAD)	1465
Skewness	0.0057345591
Sum	3.0251445×10^9
Variance	2987410.4
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
4169	2314	0.2%
1680	1850	0.2%
4277	1743	0.2%
1941	1595	0.2%
1181	1521	0.2%
889	1518	0.2%
3618	1344	0.1%
2063	1323	0.1%
1150	1302	0.1%
1015	1286	0.1%
Other values (6030)	984413	98.4%

Value	Count	Frequency (%)
1	53	< 0.1%
2	129	< 0.1%
3	51	< 0.1%
4	21	< 0.1%
5	198	< 0.1%
6	71	< 0.1%
7	31	< 0.1%
8	139	< 0.1%
9	106	< 0.1%
10	401	< 0.1%

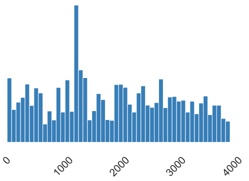
Value	Count	Frequency (%)
6040	341	< 0.1%
6039	123	< 0.1%
6038	20	< 0.1%
6037	202	< 0.1%
6036	888	0.1%
6035	280	< 0.1%
6034	21	< 0.1%
6033	60	< 0.1%
6032	104	< 0.1%
6031	51	< 0.1%

MovieID

Real number (ℝ)

Distinct	3883
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1865.5259

Minimum	1
Maximum	3952
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	15.3 MiB

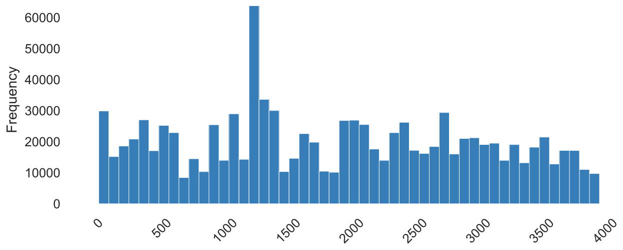


Quantile statistics

Minimum	1
5-th percentile	172
Q1	1030
median	1835
Q3	2770
95-th percentile	3675
Maximum	3952
Range	3951
Interquartile range (IQR)	1740

Descriptive statistics

Standard deviation	1096.03
Coefficient of variation (CV)	0.58751799
Kurtosis	-1.1110215
Mean	1865.5259
Median Absolute Deviation (MAD)	884
Skewness	0.092488983
Sum	1.866246×10^9
Variance	1201281.9
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
2858	3428	0.3%
260	2991	0.3%
1196	2990	0.3%
1210	2883	0.3%
480	2672	0.3%
2028	2653	0.3%
589	2649	0.3%
2571	2590	0.3%
1270	2583	0.3%
593	2578	0.3%
Other values (3873)	972369	97.2%

Value	Count	Frequency (%)
1	2077	0.2%
2	701	0.1%
3	478	< 0.1%
4	170	< 0.1%
5	296	< 0.1%
6	940	0.1%
7	458	< 0.1%
8	68	< 0.1%
9	102	< 0.1%
10	888	0.1%

Value	Count	Frequency (%)
3952	388	< 0.1%
3951	40	< 0.1%
3950	54	< 0.1%
3949	304	< 0.1%
3948	862	0.1%
3947	55	< 0.1%
3946	100	< 0.1%
3945	43	< 0.1%
3944	9	< 0.1%
3943	96	< 0.1%

Rating

Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	177
Missing (%)	< 0.1%
Memory size	15.3 MiB

Length

Max length	3
Median length	3
Mean length	3
Min length	3

Characters and Unicode

Total characters	3000627	
Distinct characters	7	
Distinct categories	2 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	**?**
Distinct scripts	1 (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	0	**?**
Unique (%)	0.0%	

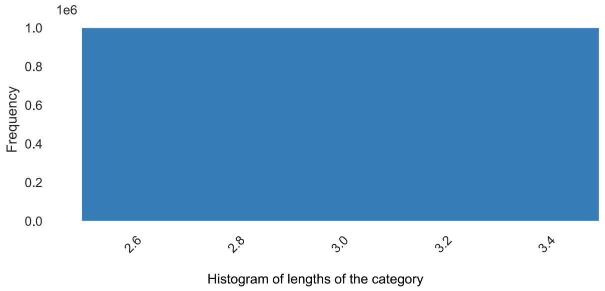
Sample

1st row	5.0
2nd row	5.0
3rd row	4.0
4th row	4.0
5th row	5.0

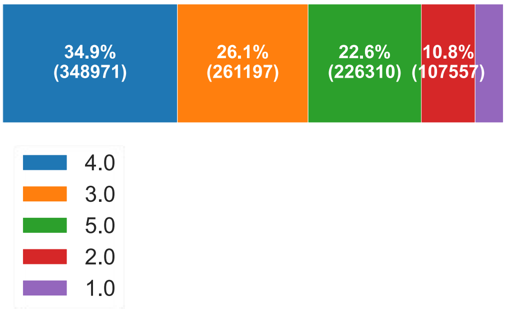
Common Values

Value	Count	Frequency (%)
4.0	348971	34.9%
3.0	261197	26.1%
5.0	226310	22.6%
2.0	107557	10.8%
1.0	56174	5.6%
(Missing)	177	< 0.1%

Length



Common Values (Plot)



Value	Count	Frequency (%)
4.0	348971	34.9%
3.0	261197	26.1%
5.0	226310	22.6%
2.0	107557	10.8%
1.0	56174	5.6%

Most occurring characters

Value	Count	Frequency (%)
.	1000209	33.3%
0	1000209	33.3%
4	348971	11.6%
3	261197	8.7%
5	226310	7.5%
2	107557	3.6%
1	56174	1.9%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	2000418	66.7%
Other Punctuation	1000209	33.3%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	1000209	50.0%
4	348971	17.4%
3	261197	13.1%
5	226310	11.3%
2	107557	5.4%
1	56174	2.8%

Other Punctuation

Value	Count	Frequency (%)
.	1000209	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	3000627	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
.	1000209	33.3%
0	1000209	33.3%
4	348971	11.6%
3	261197	8.7%
5	226310	7.5%
2	107557	3.6%
1	56174	1.9%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	3000627	100.0%

Most frequent character per block

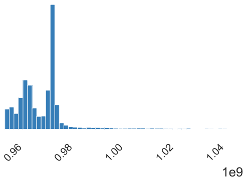
ASCII

Value	Count	Frequency (%)
.	1000209	33.3%
0	1000209	33.3%
4	348971	11.6%
3	261197	8.7%
5	226310	7.5%
2	107557	3.6%
1	56174	1.9%

Timestamp

Real number (ℝ)

Distinct	458455
Distinct (%)	45.8%
Missing	177
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	9.722437×10^8
Minimum	9.5670393×10^8
Maximum	1.0464546×10^9
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	15.3 MiB

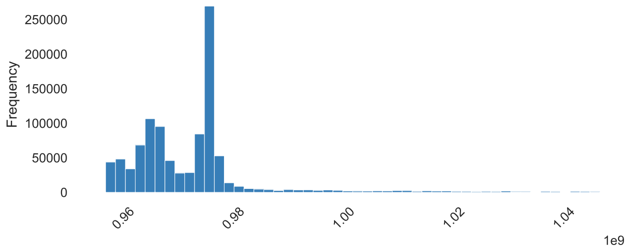


Quantile statistics

Minimum	9.5670393×10^8
5-th percentile	9.5870409×10^8
Q1	9.6530264×10^8
median	9.7301801×10^8
Q3	9.7522094×10^8
95-th percentile	9.9307415×10^8
Maximum	1.0464546×10^9
Range	89750658
Interquartile range (IQR)	9918302

Descriptive statistics

Standard deviation	12152559
Coefficient of variation (CV)	0.012499499
Kurtosis	10.949978
Mean	9.722437×10^8
Median Absolute Deviation (MAD)	5308808
Skewness	2.7656912
Sum	9.7244689×10^{14}
Variance	1.4768469×10^{14}
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
975528402	30	< 0.1%
975527781	28	< 0.1%
975440712	28	< 0.1%
1025585635	27	< 0.1%
975528243	27	< 0.1%
975528115	26	< 0.1%
975280276	26	< 0.1%
1025036288	25	< 0.1%
975280390	25	< 0.1%
974698015	24	< 0.1%
Other values (458445)	999943	> 99.9%
(Missing)	177	< 0.1%

Value	Count	Frequency (%)
956703932	1	< 0.1%
956703954	2	< 0.1%
956703977	2	< 0.1%
956704056	5	< 0.1%
956704081	1	< 0.1%
956704191	3	< 0.1%
956704219	1	< 0.1%
956704257	3	< 0.1%
956704305	1	< 0.1%
956704448	1	< 0.1%

Value	Count	Frequency (%)
1046454590	1	< 0.1%
1046454548	2	< 0.1%
1046454443	1	< 0.1%
1046454338	1	< 0.1%
1046454320	1	< 0.1%
1046454282	1	< 0.1%
1046454260	1	< 0.1%
1046444711	1	< 0.1%
1046437932	1	< 0.1%
1046437879	1	< 0.1%

Gender
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	177
Missing (%)	< 0.1%
Memory size	15.3 MiB

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Characters and Unicode

Total characters	1000209	
Distinct characters	2	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (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	0	?
Unique (%)	0.0%	

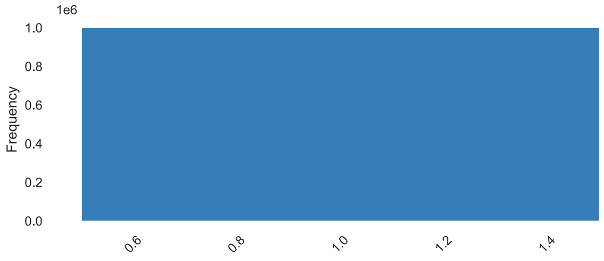
Sample

1st row	F
2nd row	M
3rd row	M
4th row	M
5th row	M

Common Values

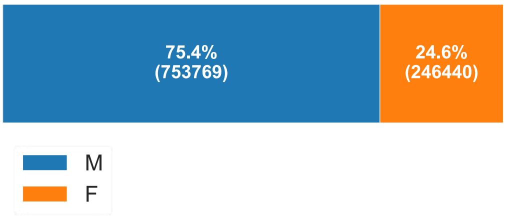
Value	Count	Frequency (%)
M	753769	75.3%
F	246440	24.6%
(Missing)	177	< 0.1%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
m	753769	75.4%
f	246440	24.6%

Most occurring characters

Value	Count	Frequency (%)
M	753769	75.4%
F	246440	24.6%

Most occurring categories

Value	Count	Frequency (%)
Uppercase Letter	1000209	100.0%

Most frequent character per category

Uppercase Letter

Value	Count	Frequency (%)
M	753769	75.4%
F	246440	24.6%

Most occurring scripts

Value	Count	Frequency (%)
Latin	1000209	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
M	753769	75.4%
F	246440	24.6%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	1000209	100.0%

Most frequent character per block

ASCII

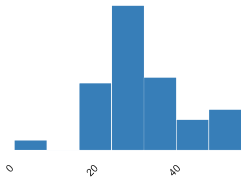
Value	Count	Frequency (%)
M	753769	75.4%
F	246440	24.6%

Age

Real number (ℝ)

Distinct	7
Distinct (%)	< 0.1%
Missing	177
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	29.738314

Minimum	1
Maximum	56
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	15.3 MiB

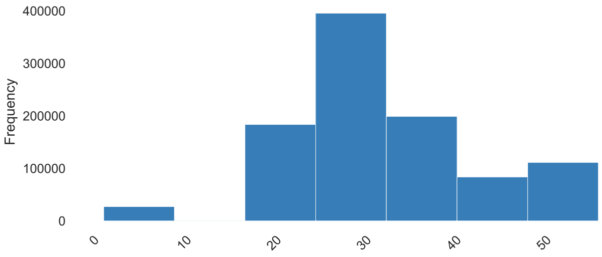


Quantile statistics

Minimum	1
5-th percentile	18
Q1	25
median	25
Q3	35
95-th percentile	50
Maximum	56
Range	55
Interquartile range (IQR)	10

Descriptive statistics

Standard deviation	11.751983
Coefficient of variation (CV)	0.39517986
Kurtosis	0.019044421
Mean	29.738314
Median Absolute Deviation (MAD)	7
Skewness	0.3984714
Sum	29744529
Variance	138.10909
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=7)

Value	Count	Frequency (%)
25	395556	39.5%
35	199003	19.9%
18	183536	18.3%
45	83633	8.4%
50	72490	7.2%
56	38780	3.9%
1	27211	2.7%
(Missing)	177	< 0.1%

Value	Count	Frequency (%)
1	27211	2.7%
18	183536	18.3%
25	395556	39.5%
35	199003	19.9%
45	83633	8.4%
50	72490	7.2%
56	38780	3.9%

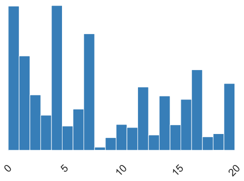
Value	Count	Frequency (%)
56	38780	3.9%
50	72490	7.2%
45	83633	8.4%
35	199003	19.9%
25	395556	39.5%
18	183536	18.3%
1	27211	2.7%

Occupation

Real number (ℝ)

Distinct	21
Distinct (%)	< 0.1%
Missing	177
Missing (%)	< 0.1%
Infinite	0
Infinite (%)	0.0%
Mean	8.0361384

Minimum	0
Maximum	20
Zeros	130499
Zeros (%)	13.0%
Negative	0
Negative (%)	0.0%
Memory size	15.3 MiB

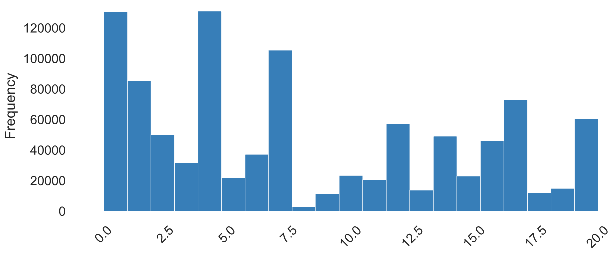


Quantile statistics

Minimum	0
5-th percentile	0
Q1	2
median	7
Q3	14
95-th percentile	20
Maximum	20
Range	20
Interquartile range (IQR)	12

Descriptive statistics

Standard deviation	6.5313358
Coefficient of variation (CV)	0.81274555
Kurtosis	-1.2170057
Mean	8.0361384
Median Absolute Deviation (MAD)	6
Skewness	0.40436252
Sum	8037818
Variance	42.658347
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=21)

Value	Count	Frequency (%)
4	131032	13.1%
0	130499	13.0%
7	105425	10.5%
1	85351	8.5%
17	72816	7.3%
20	60397	6.0%
12	57214	5.7%
2	50068	5.0%
14	49109	4.9%
16	46021	4.6%
Other values (11)	212277	21.2%

Value	Count	Frequency (%)
0	130499	13.0%
1	85351	8.5%
2	50068	5.0%
3	31623	3.2%
4	131032	13.1%
5	21850	2.2%
6	37205	3.7%
7	105425	10.5%
8	2706	0.3%
9	11345	1.1%

Value	Count	Frequency (%)
20	60397	6.0%
19	14904	1.5%
18	12086	1.2%
17	72816	7.3%
16	46021	4.6%
15	22951	2.3%
14	49109	4.9%
13	13754	1.4%
12	57214	5.7%
11	20563	2.1%

Zip-code

Categorical

Distinct	3439
Distinct (%)	0.3%
Missing	177
Missing (%)	< 0.1%
Memory size	15.3 MiB

Length

Max length	10
Median length	5
Mean length	5.071958
Min length	5

Characters and Unicode

Total characters	5073018	
Distinct characters	11	
Distinct categories	2 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (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	0	?
Unique (%)	0.0%	

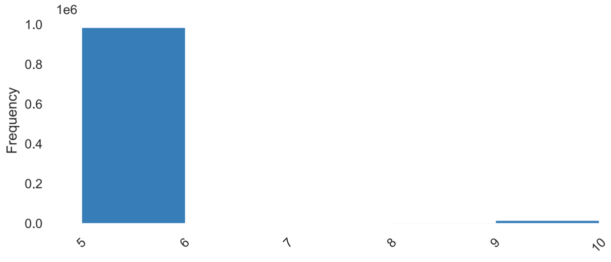
Sample

1st row	48067
2nd row	70072
3rd row	32793
4th row	22903
5th row	95350

Common Values

Value	Count	Frequency (%)
94110	3802	0.4%
60640	3430	0.3%
98103	3204	0.3%
95616	3079	0.3%
02138	3019	0.3%
55408	2787	0.3%
48135	2725	0.3%
97401	2663	0.3%
10025	2632	0.3%
10024	2594	0.3%
Other values (3429)	970274	97.0%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
94110	3802	0.4%
60640	3430	0.3%
98103	3204	0.3%
95616	3079	0.3%
02138	3019	0.3%
55408	2787	0.3%
48135	2725	0.3%
97401	2663	0.3%
10025	2632	0.3%
10024	2594	0.3%
Other values (3429)	970274	97.0%

Most occurring characters

Value	Count	Frequency (%)
0	873885	17.2%
1	685622	13.5%
2	559558	11.0%
5	499853	9.9%
4	489803	9.7%
3	453488	8.9%
9	422031	8.3%
6	383841	7.6%
7	349915	6.9%
8	341796	6.7%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	5059792	99.7%
Dash Punctuation	13226	0.3%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	873885	17.3%
1	685622	13.6%
2	559558	11.1%
5	499853	9.9%
4	489803	9.7%
3	453488	9.0%
9	422031	8.3%
6	383841	7.6%
7	349915	6.9%
8	341796	6.8%

Dash Punctuation

Value	Count	Frequency (%)
-	13226	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Common	5073018	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	873885	17.2%
1	685622	13.5%
2	559558	11.0%
5	499853	9.9%
4	489803	9.7%
3	453488	8.9%
9	422031	8.3%
6	383841	7.6%
7	349915	6.9%
8	341796	6.7%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	5073018	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	873885	17.2%
1	685622	13.5%
2	559558	11.0%
5	499853	9.9%
4	489803	9.7%
3	453488	8.9%
9	422031	8.3%
6	383841	7.6%
7	349915	6.9%
8	341796	6.7%

Title

Categorical

Distinct	3883
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	15.3 MiB

Length

Max length	82
Median length	68
Mean length	23.330858
Min length	8

Characters and Unicode

Total characters	23339864	
Distinct characters	98	
Distinct categories	10 (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	2 (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	291	?
Unique (%)	< 0.1%	

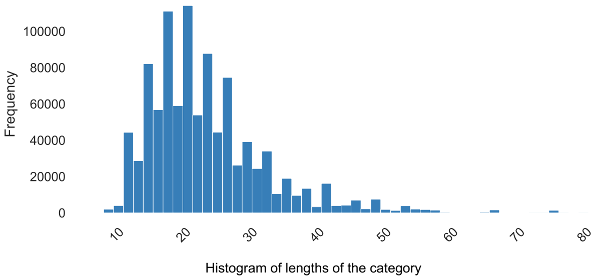
Sample

1st row	One Flew Over the Cuckoo's Nest (1975)
2nd row	One Flew Over the Cuckoo's Nest (1975)
3rd row	One Flew Over the Cuckoo's Nest (1975)
4th row	One Flew Over the Cuckoo's Nest (1975)
5th row	One Flew Over the Cuckoo's Nest (1975)

Common Values

Value	Count	Frequency (%)
American Beauty (1999)	3428	0.3%
Star Wars: Episode IV - A New Hope (1977)	2991	0.3%
Star Wars: Episode V - The Empire Strikes Back (1980)	2990	0.3%
Star Wars: Episode VI - Return of the Jedi (1983)	2883	0.3%
Jurassic Park (1993)	2672	0.3%
Saving Private Ryan (1998)	2653	0.3%
Terminator 2: Judgment Day (1991)	2649	0.3%
Matrix, The (1999)	2590	0.3%
Back to the Future (1985)	2583	0.3%
Silence of the Lambs, The (1991)	2578	0.3%
Other values (3873)	972369	97.2%

Length



Value	Count	Frequency (%)
the	330676	8.5%
1999	86845	2.2%
of	79250	2.0%
1998	68246	1.8%
1997	65413	1.7%
1995	60784	1.6%
1996	59415	1.5%
1994	52974	1.4%
1993	46322	1.2%
2000	41954	1.1%
Other values (4646)	2984074	77.0%

Most occurring characters

Value	Count	Frequency (%)
	2875567	12.3%
9	1636077	7.0%
e	1615225	6.9%
a	1034617	4.4%
1	1033072	4.4%
)	1026696	4.4%
(	1026696	4.4%
o	924101	4.0%
r	876088	3.8%
n	864706	3.7%
Other values (88)	10427019	44.7%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	11320460	48.5%
Decimal Number	4076867	17.5%
Space Separator	2875567	12.3%
Uppercase Letter	2552866	10.9%
Close Punctuation	1026696	4.4%
Open Punctuation	1026696	4.4%
Other Punctuation	436209	1.9%
Dash Punctuation	23541	0.1%
Other Number	925	< 0.1%
Currency Symbol	37	< 0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	1615225	14.3%
a	1034617	9.1%
o	924101	8.2%
r	876088	7.7%
n	864706	7.6%
i	831653	7.3%
t	811245	7.2%
h	621921	5.5%
s	621905	5.5%
l	511032	4.5%
Other values (32)	2607967	23.0%

Uppercase Letter

Value	Count	Frequency (%)
T	354990	13.9%
S	234116	9.2%
M	177056	6.9%
B	171131	6.7%
A	151363	5.9%
D	133135	5.2%
C	131377	5.1%
P	122627	4.8%
F	118280	4.6%
W	113338	4.4%
Other values (19)	845453	33.1%

Other Punctuation

Value	Count	Frequency (%)
,	247192	56.7%
:	63981	14.7%
.	49641	11.4%
'	45577	10.4%
!	8760	2.0%
&	8504	1.9%
?	4923	1.1%
/	4200	1.0%
*	3375	0.8%
;	28	< 0.1%

Decimal Number

Value	Count	Frequency (%)
9	1636077	40.1%
1	1033072	25.3%
8	343039	8.4%
0	200543	4.9%

7	196868	4.8%
6	155772	3.8%
5	140827	3.5%
2	138916	3.4%
4	128670	3.2%
3	103083	2.5%

Space Separator

Value	Count	Frequency (%)
	2875567	100.0%

Close Punctuation

Value	Count	Frequency (%)
)	1026696	100.0%

Open Punctuation

Value	Count	Frequency (%)
(	1026696	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	23541	100.0%

Other Number

Value	Count	Frequency (%)
³	925	100.0%

Currency Symbol

Value	Count	Frequency (%)
$	37	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	13873326	59.4%
Common	9466538	40.6%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	1615225	11.6%
a	1034617	7.5%
o	924101	6.7%
r	876088	6.3%
n	864706	6.2%
i	831653	6.0%
t	811245	5.8%
h	621921	4.5%
s	621905	4.5%
l	511032	3.7%
Other values (61)	5160833	37.2%

Common

Value	Count	Frequency (%)
	2875567	30.4%
9	1636077	17.3%
1	1033072	10.9%
)	1026696	10.8%
(	1026696	10.8%
8	343039	3.6%
,	247192	2.6%
0	200543	2.1%
7	196868	2.1%
6	155772	1.6%
Other values (17)	725016	7.7%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	23335252	> 99.9%
None	4612	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
	2875567	12.3%
9	1636077	7.0%
e	1615225	6.9%
a	1034617	4.4%
1	1033072	4.4%
)	1026696	4.4%
(	1026696	4.4%
o	924101	4.0%
r	876088	3.8%
n	864706	3.7%
Other values (68)	10422407	44.7%

None

Value	Count	Frequency (%)
è	1475	32.0%
é	1375	29.8%
³	925	20.1%
ü	242	5.2%
ê	141	3.1%
ö	104	2.3%
î	89	1.9%
í	74	1.6%
à	43	0.9%
É	39	0.8%
Other values (10)	105	2.3%

Genres

Categorical

Distinct	301
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	15.3 MiB

Length

Max length	47
Median length	40
Mean length	14.648918
Min length	3

Characters and Unicode

Total characters	14654572	
Distinct characters	30	
Distinct categories	5 (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	1	**?**
Unique (%)	< 0.1%	

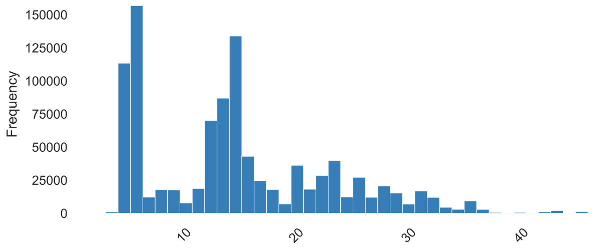
Sample

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

Common Values

Value	Count	Frequency (%)
Comedy	116905	11.7%
Drama	111507	11.1%
Comedy|Romance	42716	4.3%
Comedy|Drama	42254	4.2%
Drama|Romance	29173	2.9%
Action|Thriller	26759	2.7%
Horror	22566	2.3%
Drama|Thriller	18250	1.8%
Thriller	17852	1.8%
Action|Adventure|Sci-Fi	17783	1.8%
Other values (291)	554621	55.4%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
comedy	116905	11.7%
drama	111507	11.1%
comedy|romance	42716	4.3%
comedy|drama	42254	4.2%
drama|romance	29173	2.9%
action|thriller	26759	2.7%
horror	22566	2.3%
drama|thriller	18250	1.8%
thriller	17852	1.8%
action|adventure|sci-fi	17783	1.8%
Other values (291)	554621	55.4%

Most occurring characters

Value	Count	Frequency (%)
r	1404457	9.6%
e	1202962	8.2%
|	1101645	7.5%
a	1090698	7.4%
i	1078124	7.4%
m	1007823	6.9%
o	983878	6.7%
n	762640	5.2%
c	611757	4.2%
d	562759	3.8%
Other values (20)	4847829	33.1%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	11027595	75.3%
Uppercase Letter	2277588	15.5%
Math Symbol	1101645	7.5%
Dash Punctuation	175557	1.2%
Other Punctuation	72187	0.5%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
r	1404457	12.7%
e	1202962	10.9%
a	1090698	9.9%
i	1078124	9.8%
m	1007823	9.1%
o	983878	8.9%
n	762640	6.9%
c	611757	5.5%
d	562759	5.1%
t	539805	4.9%
Other values (6)	1782692	16.2%

Uppercase Letter

Value	Count	Frequency (%)
C	508355	22.3%
A	434713	19.1%
D	362566	15.9%
F	211858	9.3%
T	189687	8.3%
S	157296	6.9%
R	147535	6.5%
W	89213	3.9%
M	81714	3.6%
H	76390	3.4%

Math Symbol

Value	Count	Frequency (%)
|	1101645	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	175557	100.0%

Other Punctuation

Value	Count	Frequency (%)
'	72187	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	13305183	90.8%
Common	1349389	9.2%

Most frequent character per script

Latin

Value	Count	Frequency (%)
r	1404457	10.6%
e	1202962	9.0%
a	1090698	8.2%
i	1078124	8.1%
m	1007823	7.6%
o	983878	7.4%
n	762640	5.7%
c	611757	4.6%
d	562759	4.2%
t	539805	4.1%
Other values (17)	4060280	30.5%

Common

Value	Count	Frequency (%)
|	1101645	81.6%
-	175557	13.0%
'	72187	5.3%

Most occurring blocks

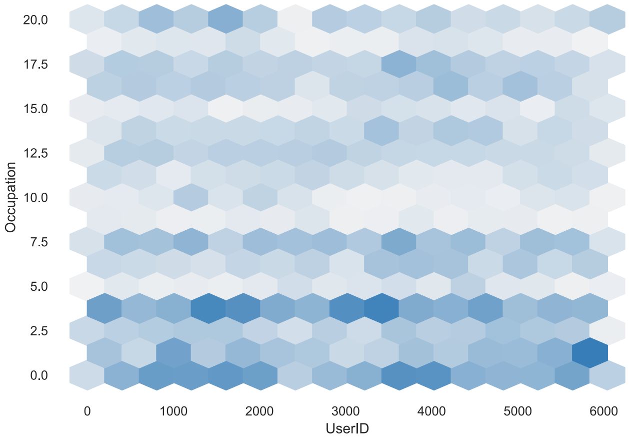
Value	Count	Frequency (%)
ASCII	14654572	100.0%

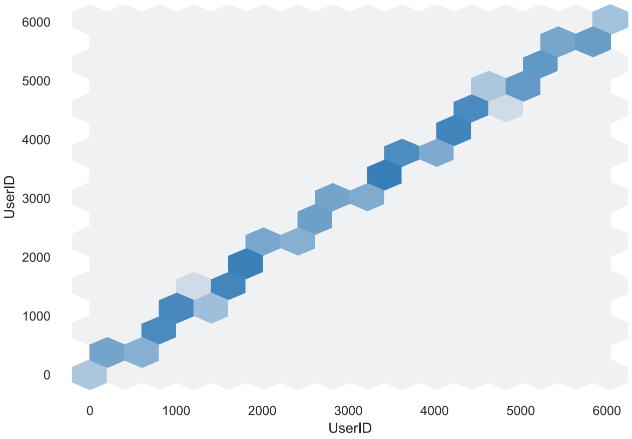
Most frequent character per block

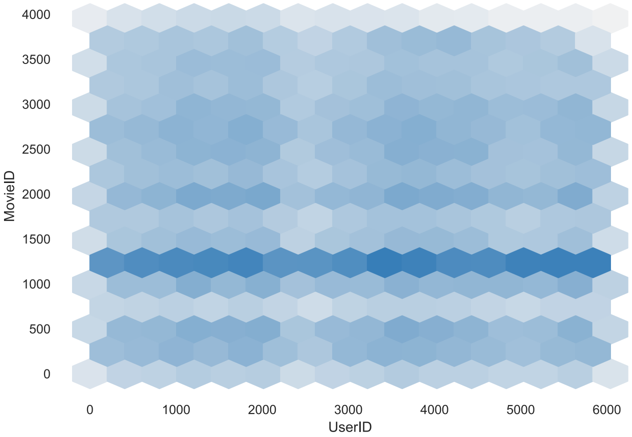
ASCII

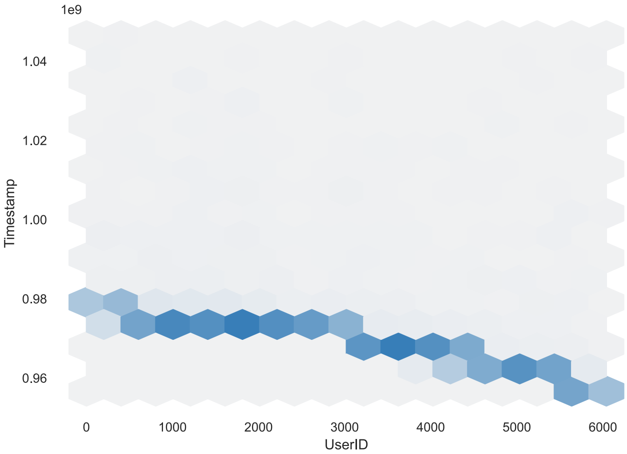
Value	Count	Frequency (%)
r	1404457	9.6%
e	1202962	8.2%
|	1101645	7.5%
a	1090698	7.4%
i	1078124	7.4%
m	1007823	6.9%
o	983878	6.7%
n	762640	5.2%
c	611757	4.2%
d	562759	3.8%
Other values (20)	4847829	33.1%

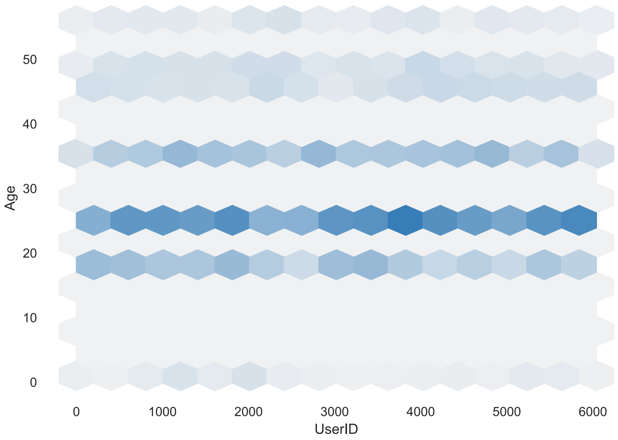
Interactions

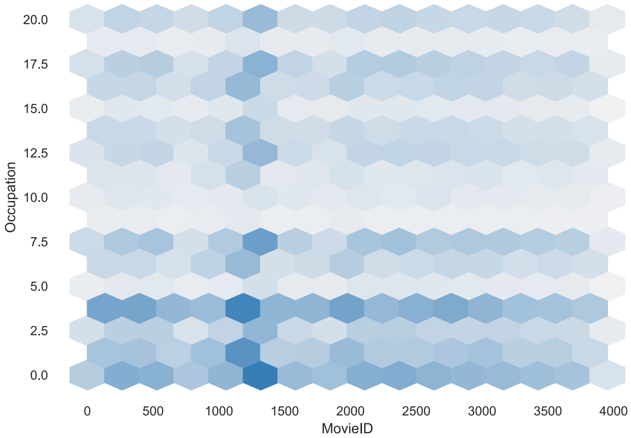


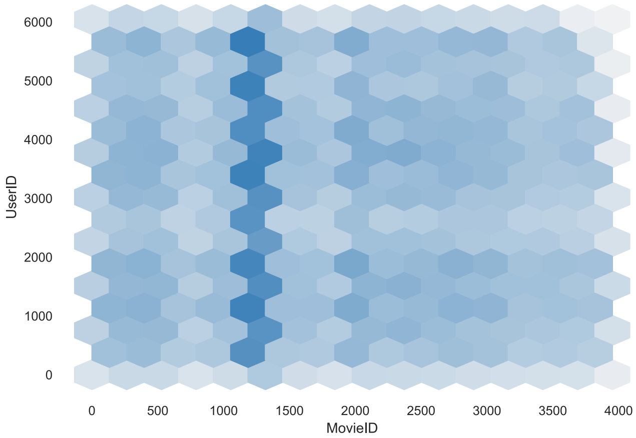

6000
5000
4000
3000
2000
1000
0
UserID
0
1000
2000
3000
4000
5000
6000
UserID

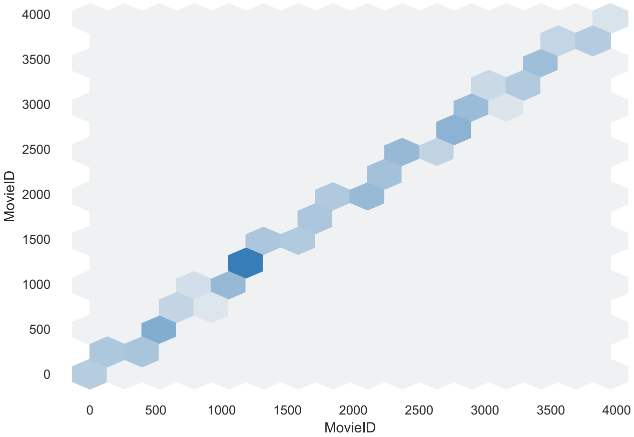

4000
3500
3000
2500
2000
1500
1000
500
0
MovieID
0
1000
2000
3000
4000
5000
6000
UserID

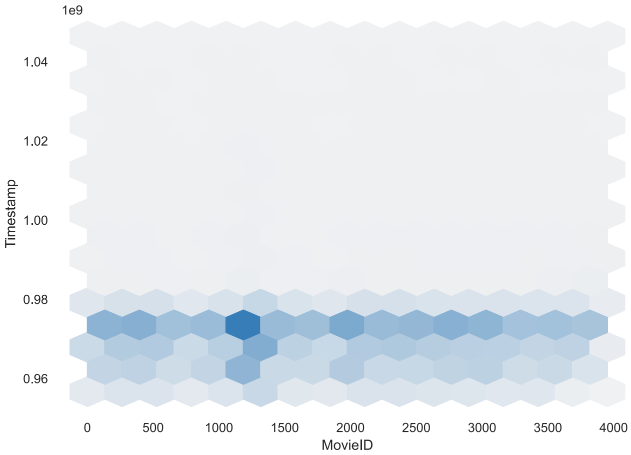

1e9
1.04
1.02
1.00
0.98
0.96
Timestamp
0
1000
2000
3000
4000
5000
6000
UserID

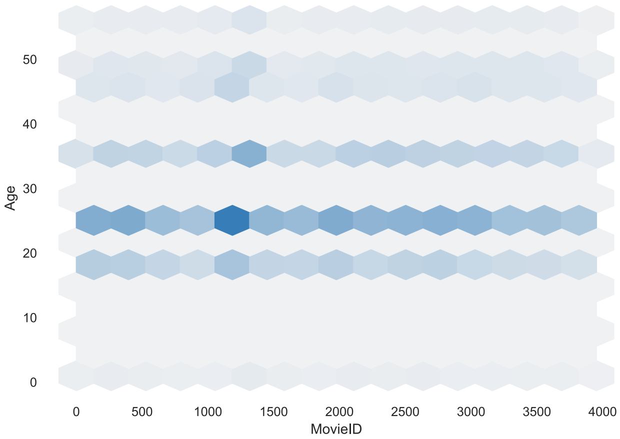

Age
50
40
30
20
10
0
0
1000
2000
3000
4000
5000
6000
UserID

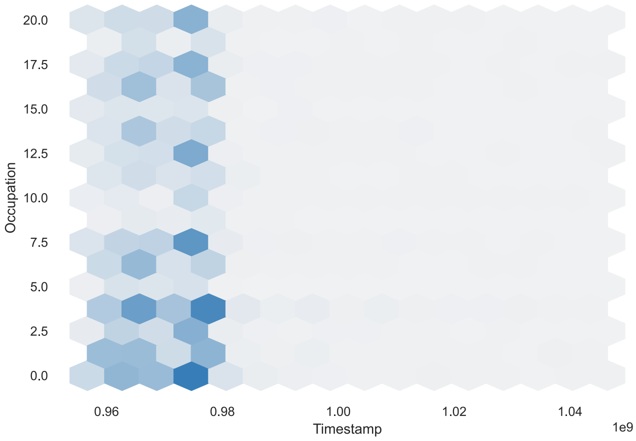

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Occupation
0
500
1000
1500
2000
2500
3000
3500
4000
MovieID

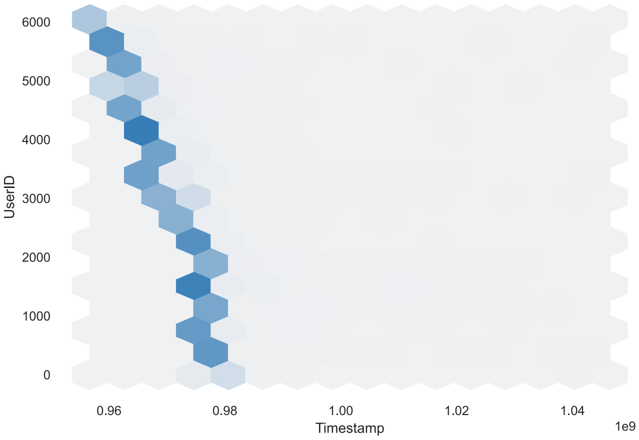

6000
5000
4000
3000
2000
1000
0
UserID
0
500
1000
1500
2000
2500
3000
3500
4000
MovieID

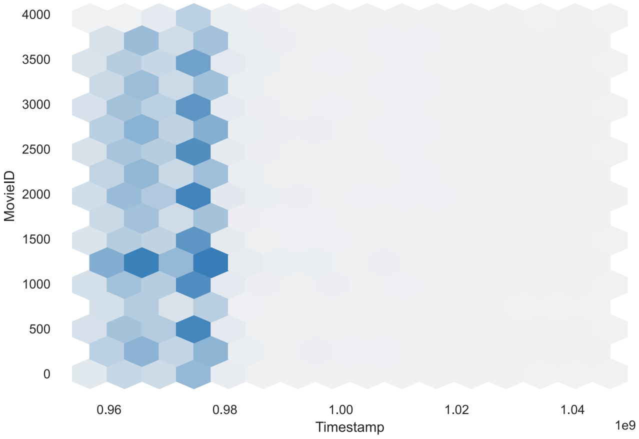

4000
3500
3000
2500
2000
1500
1000
500
0
MovieID
0
500
1000
1500
2000
2500
3000
3500
4000
MovieID

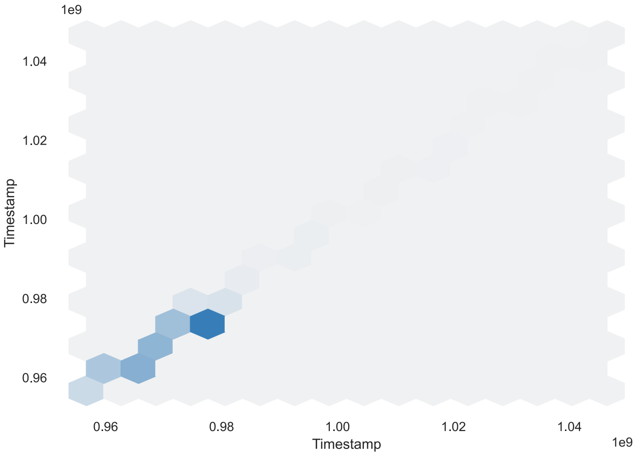

1e9
1.04
1.02
1.00
0.98
0.96
Timestamp
0
500
1000
1500
2000
2500
3000
3500
4000
MovieID

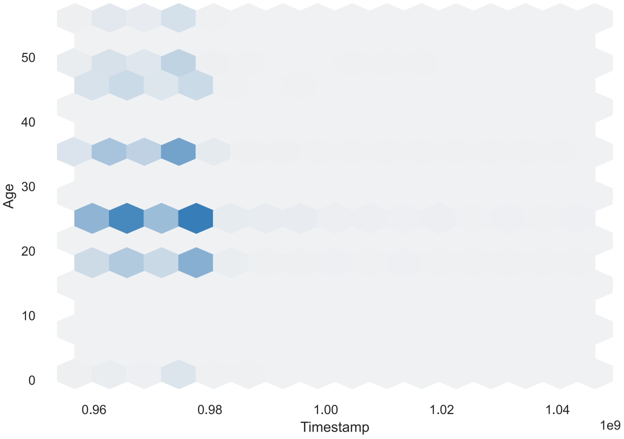

50
40
30
20
10
0
Age
0
500
1000
1500
2000
2500
3000
3500
4000
MovieID

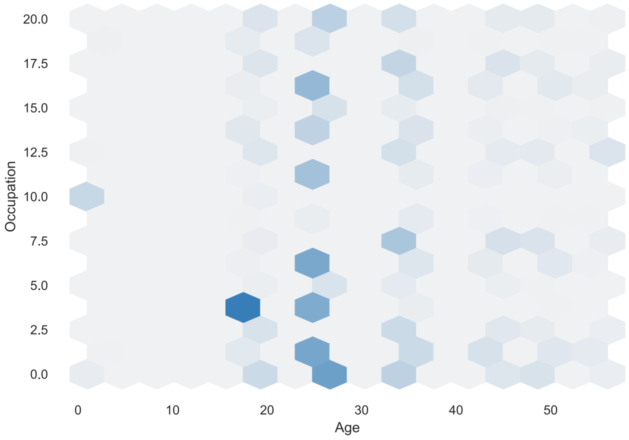

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Occupation
0.96
0.98
1.00
1.02
1.04
Timestamp
1e9

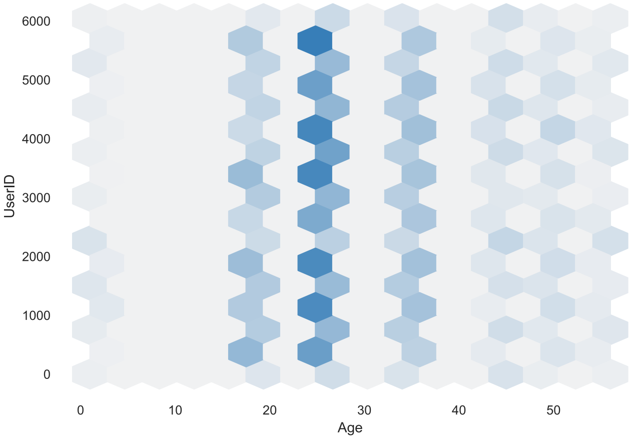

6000
5000
4000
3000
2000
1000
0
UserID
0.96
0.98
1.00
1.02
1.04
Timestamp
1e9

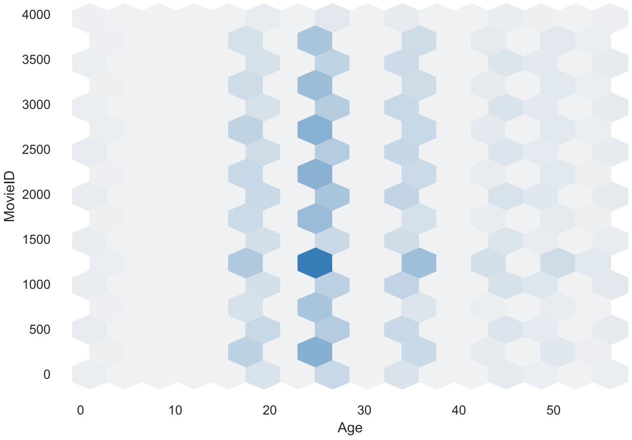

4000
3500
3000
2500
2000
1500
1000
500
0
MovieID
0.96
0.98
1.00
1.02
1.04
Timestamp
1e9

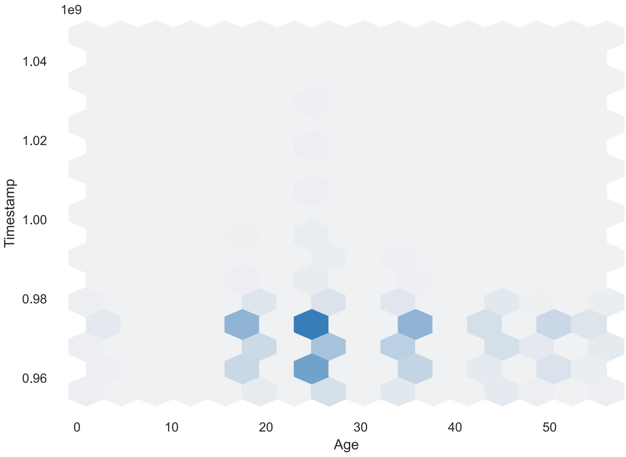

1e9
1.04
1.02
1.00
0.98
0.96
Timestamp
0.96
0.98
1.00
1.02
1.04
Timestamp
1e9

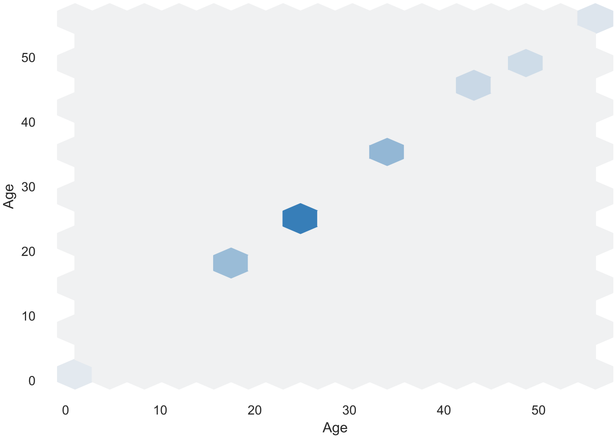

Age
0
10
20
30
40
50
0.96
0.98
1.00
1.02
1.04
1e9
Timestamp

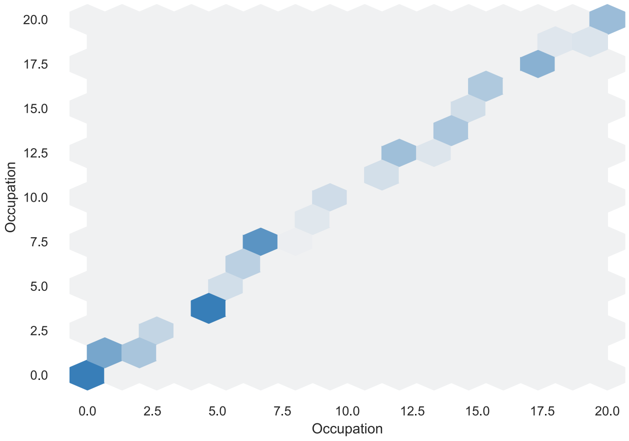

20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Occupation
0
10
20
30
40
50
Age

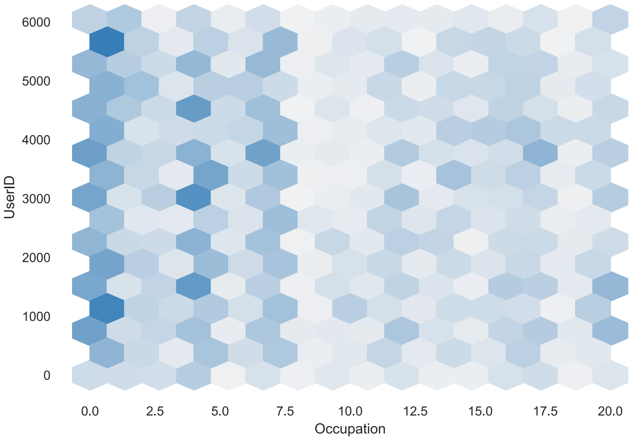

6000
5000
4000
3000
2000
1000
0
UserID
0
10
20
30
40
50
Age

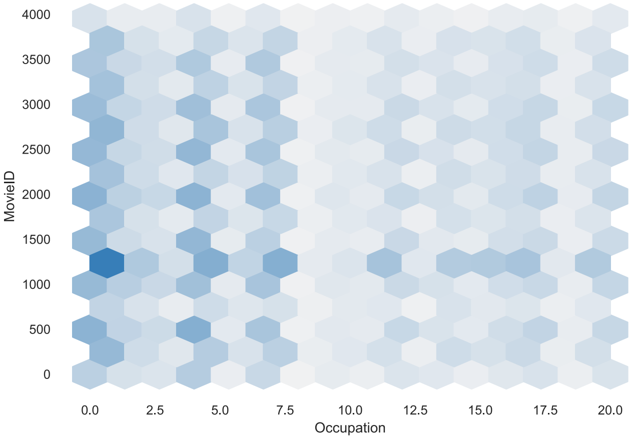

4000
3500
3000
2500
2000
1500
1000
500
0
MovieID
0
10
20
30
40
50
Age

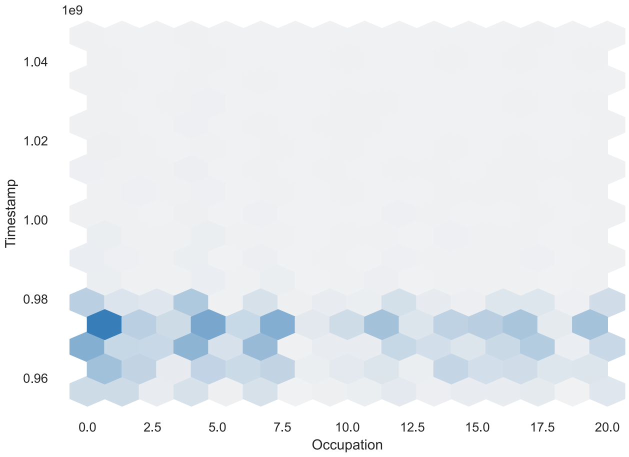

1e9
1.04
1.02
1.00
0.98
0.96
Timestamp
0
10
20
30
40
50
Age

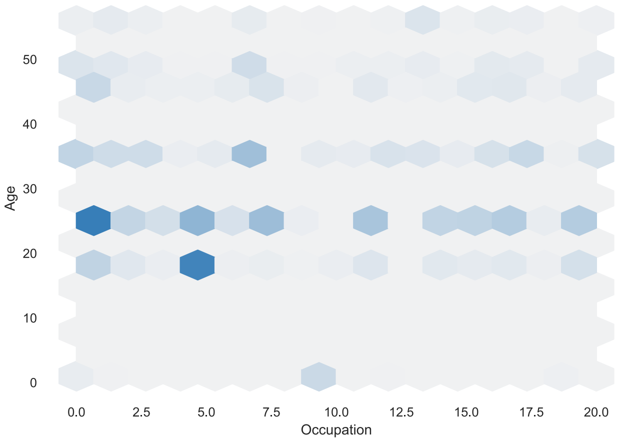

0
10
20
30
40
50
Age
0
10
20
30
40
50
Age


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Occupation
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Occupation

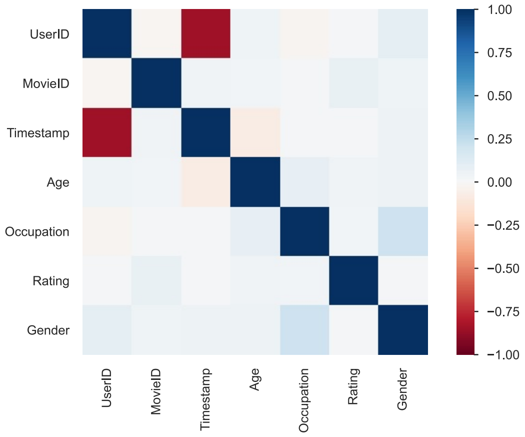

6000
5000
4000
3000
2000
1000
0
UserID
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Occupation


4000
3500
3000
2500
2000
1500
1000
500
0
MovieID
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Occupation


1e9
1.04
1.02
1.00
0.98
0.96
Timestamp
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Occupation

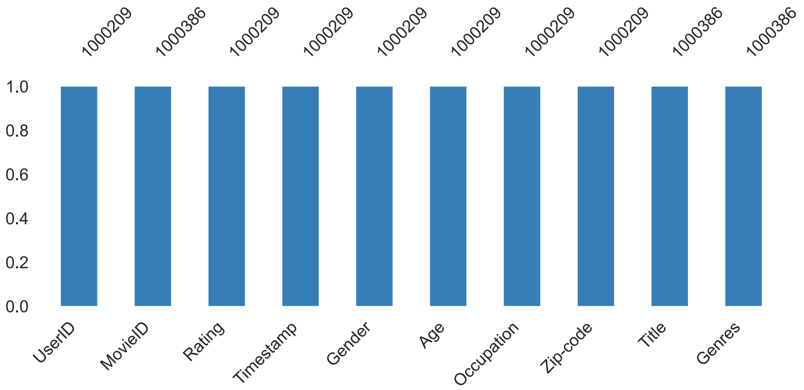

Age
50
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Occupation

Correlations

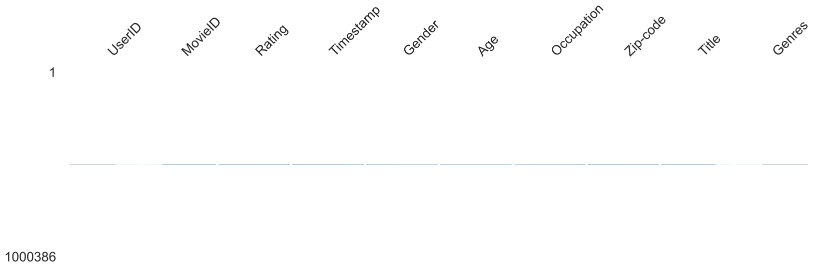


	UserID	MovieID	Timestamp	Age	Occupation	Rating	Gender
UserID	1.000	-0.016	-0.844	0.044	-0.026	0.020	0.095
MovieID	-0.016	1.000	0.038	0.028	0.009	0.076	0.040
Timestamp	-0.844	0.038	1.000	-0.072	0.020	0.021	0.052
Age	0.044	0.028	-0.072	1.000	0.085	0.037	0.047
Occupation	-0.026	0.009	0.020	0.085	1.000	0.027	0.209
Rating	0.020	0.076	0.021	0.037	0.027	1.000	0.021
Gender	0.095	0.040	0.052	0.047	0.209	0.021	1.000

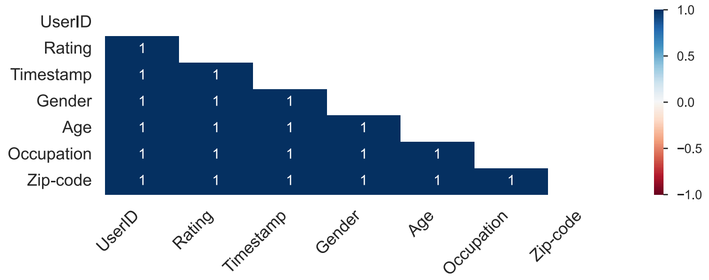
Missing values



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

	UserID	MovieID	Rating	Timestamp	Gender	Age	Occupation	Zip-code	Title	Genres
0	1.0	1193	5.0	978300760.0	F	1.0	10.0	48067	One Flew Over the Cuckoo's Nest (1975)	Drama
1	2.0	1193	5.0	978298413.0	M	56.0	16.0	70072	One Flew Over the Cuckoo's Nest (1975)	Drama
2	12.0	1193	4.0	978220179.0	M	25.0	12.0	32793	One Flew Over the Cuckoo's Nest (1975)	Drama
3	15.0	1193	4.0	978199279.0	M	25.0	7.0	22903	One Flew Over the Cuckoo's Nest (1975)	Drama
4	17.0	1193	5.0	978158471.0	M	50.0	1.0	95350	One Flew Over the Cuckoo's Nest (1975)	Drama
5	18.0	1193	4.0	978156168.0	F	18.0	3.0	95825	One Flew Over the Cuckoo's Nest (1975)	Drama
6	19.0	1193	5.0	982730936.0	M	1.0	10.0	48073	One Flew Over the Cuckoo's Nest (1975)	Drama
7	24.0	1193	5.0	978136709.0	F	25.0	7.0	10023	One Flew Over the Cuckoo's Nest (1975)	Drama
8	28.0	1193	3.0	978125194.0	F	25.0	1.0	14607	One Flew Over the Cuckoo's Nest (1975)	Drama
9	33.0	1193	5.0	978557765.0	M	45.0	3.0	55421	One Flew Over the Cuckoo's Nest (1975)	Drama

	UserID	MovieID	Rating	Timestamp	Gender	Age	Occupation	Zip-code	Title	Genres
1000376	NaN	3561	NaN	NaN	NaN	NaN	NaN	NaN	Stacy's Knights (1982)	Drama
1000377	NaN	3582	NaN	NaN	NaN	NaN	NaN	NaN	Jails, Hospitals & Hip-Hop (2000)	Drama
1000378	NaN	3583	NaN	NaN	NaN	NaN	NaN	NaN	Black Tights (Les Collants Noirs) (1960)	Drama
1000379	NaN	3589	NaN	NaN	NaN	NaN	NaN	NaN	Kill, Baby... Kill! (Operazione Paura) (1966)	Horror
1000380	NaN	3630	NaN	NaN	NaN	NaN	NaN	NaN	House of Exorcism, The (La Casa dell'esorcismo) (1974)	Horror
1000381	NaN	3650	NaN	NaN	NaN	NaN	NaN	NaN	Anguish (Angustia) (1986)	Horror
1000382	NaN	3750	NaN	NaN	NaN	NaN	NaN	NaN	Boricua's Bond (2000)	Drama
1000383	NaN	3829	NaN	NaN	NaN	NaN	NaN	NaN	Mad About Mambo (2000)	Comedy|Romance
1000384	NaN	3856	NaN	NaN	NaN	NaN	NaN	NaN	Autumn Heart (1999)	Drama
1000385	NaN	3907	NaN	NaN	NaN	NaN	NaN	NaN	Prince of Central Park, The (1999)	Drama

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