

Overview

Dataset statistics

Number of variables	13
Number of observations	18635
Missing cells	20473
Missing cells (%)	8.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.8 MiB
Average record size in memory	104.0 B

Variable types

Numeric	2
Categorical	4
Text	4
DateTime	2
Boolean	1

Alerts

suffix is highly imbalanced (64.1%)	Imbalance
party is highly imbalanced (60.3%)	Imbalance
middlename has 3536 (19.0%) missing values	Missing
suffix has 16937 (90.9%) missing values	Missing

Reproduction

Analysis started	2023-09-09 20:02:16.787225
Analysis finished	2023-09-09 20:02:17.918715
Duration	1.13 second
Software version	ydata-profiling vv4.5.1 (https://github.com/ydataai/ydata-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Profiling%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2C%20%22create

Variables

Select Columns ↕

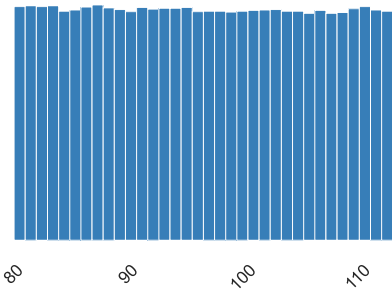
congress

Real number (ℝ)

Distinct	34	Minimum	80
----------	----	---------	----

Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	96.445989

Maximum	113
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	145.7 KiB

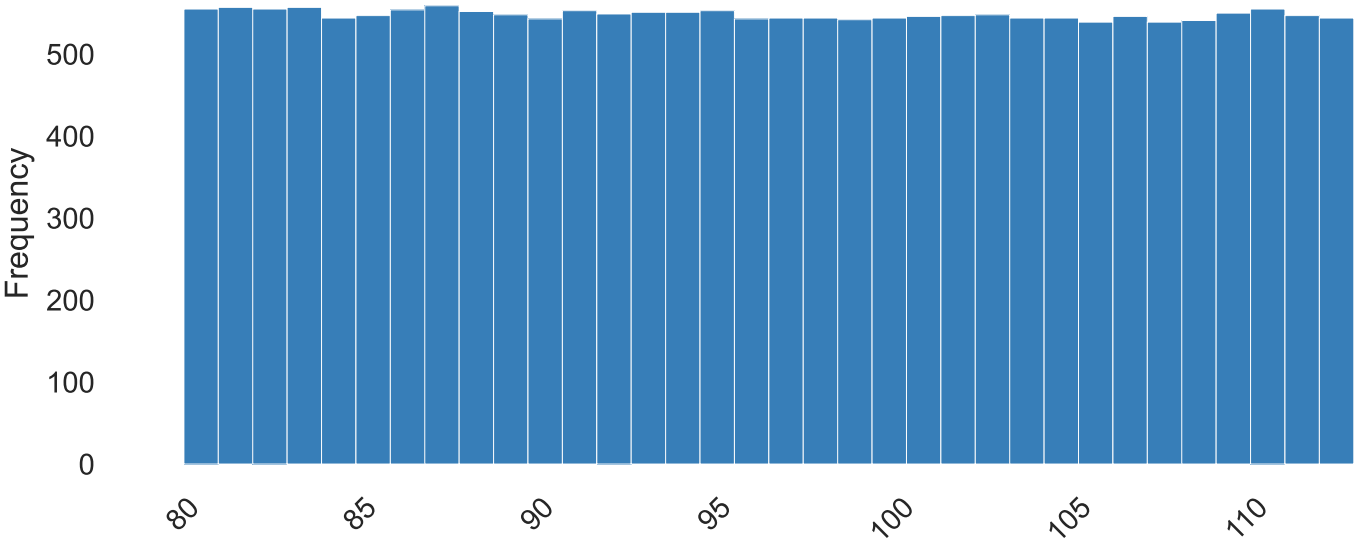


Quantile statistics

Minimum	80
5-th percentile	81
Q1	88
median	96
Q3	105
95-th percentile	112
Maximum	113
Range	33
Interquartile range (IQR)	17

Descriptive statistics

Standard deviation	9.8234288
Coefficient of variation (CV)	0.1018542
Kurtosis	-1.2041551
Mean	96.445989
Median Absolute Deviation (MAD)	9
Skewness	0.0077050102
Sum	1797271
Variance	96.499752
Monotonicity	Increasing



Histogram with fixed size bins (bins=34)

Value	Count	Frequency (%)
87	559	3.0%
83	557	3.0%
81	557	3.0%
80	555	3.0%
111	555	3.0%
82	555	3.0%

86	554	3.0%
95	553	3.0%
91	553	3.0%
88	552	3.0%
Other values (24)	13085	70.2%

Value	Count	Frequency (%)
80	555	3.0%
81	557	3.0%
82	555	3.0%
83	557	3.0%
84	544	2.9%
85	547	2.9%
86	554	3.0%
87	559	3.0%
88	552	3.0%
89	548	2.9%

Value	Count	Frequency (%)
113	544	2.9%
112	547	2.9%
111	555	3.0%
110	550	3.0%
109	541	2.9%
108	539	2.9%
107	546	2.9%
106	539	2.9%
105	544	2.9%
104	544	2.9%

chamber

Categorical

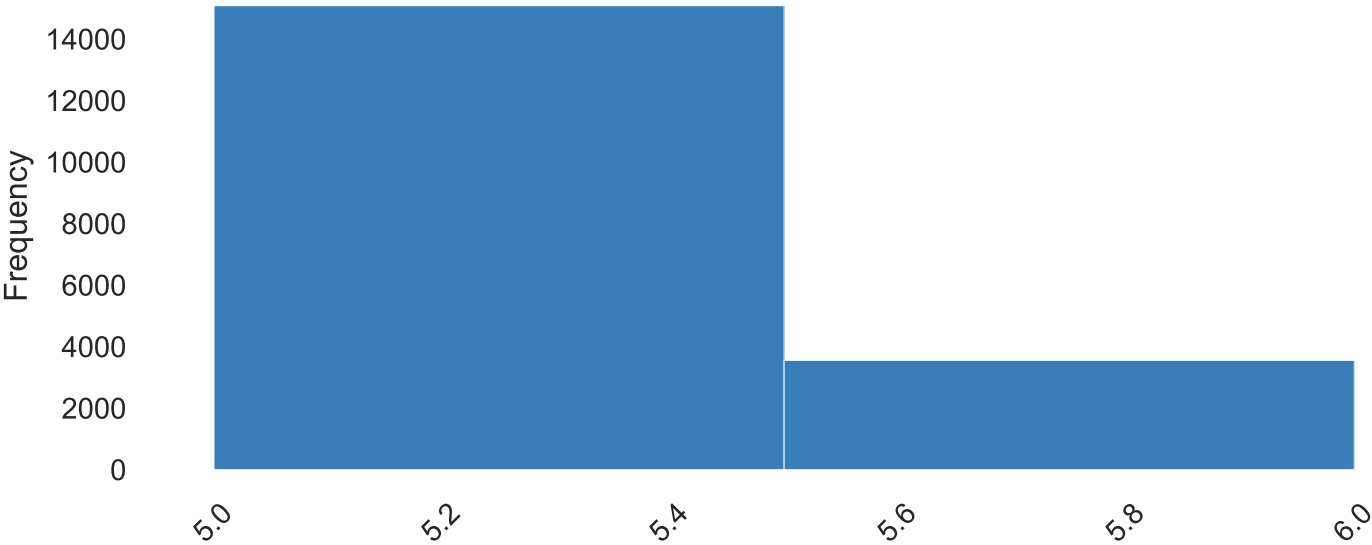
Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB

Length		Characters and Unicode	Unique		Sample
Max length	6	Total 96727 characters	Unique 0	?	1st row house
Median length	5		Unique (%) 0.0%		2nd row house
Mean length	5.1906091	Distinct 8 characters			3rd row house
Min length	5	Distinct 1 categories (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)		?	4th row house
		Distinct 1 scripts (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)		?	5th row house
		Distinct 1 blocks (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.					

Common Values

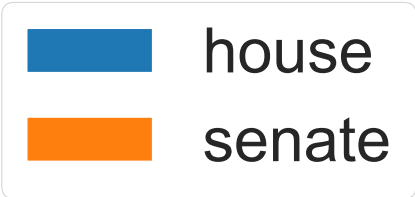
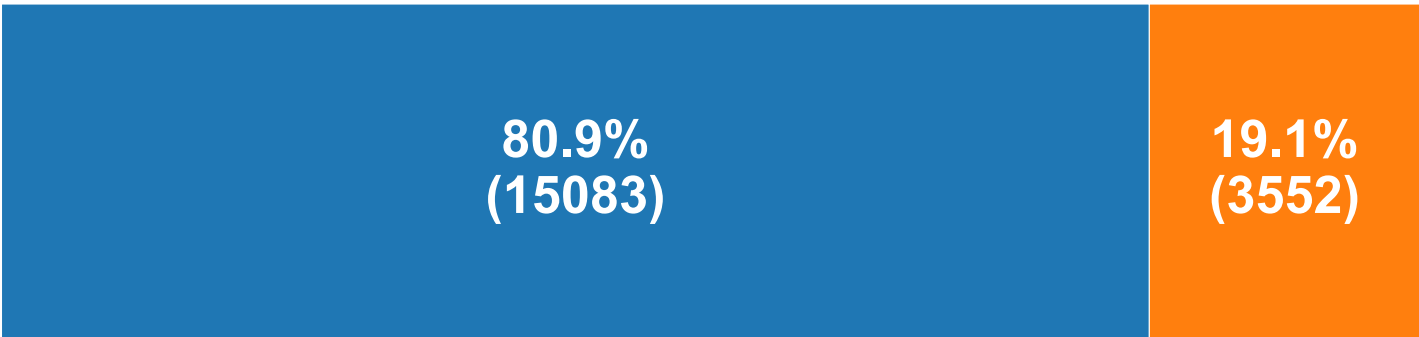
Value	Count	Frequency (%)
house	15083	80.9%
senate	3552	19.1%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
house	15083	80.9%
senate	3552	19.1%

Most occurring characters

Value	Count	Frequency (%)
e	22187	22.9%
s	18635	19.3%

h	15083	15.6%
o	15083	15.6%
u	15083	15.6%
n	3552	3.7%
a	3552	3.7%
t	3552	3.7%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	96727	100.0%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	22187	22.9%
s	18635	19.3%
h	15083	15.6%
o	15083	15.6%
u	15083	15.6%
n	3552	3.7%
a	3552	3.7%
t	3552	3.7%

Most occurring scripts

Value	Count	Frequency (%)
Latin	96727	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	22187	22.9%
s	18635	19.3%
h	15083	15.6%
o	15083	15.6%
u	15083	15.6%

n	3552	3.7%
a	3552	3.7%
t	3552	3.7%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	96727	100.0%

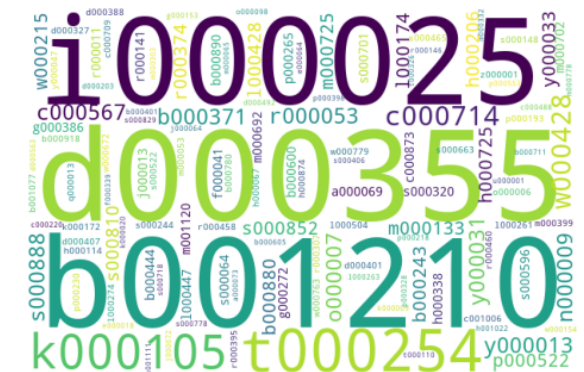
Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	22187	22.9%
s	18635	19.3%
h	15083	15.6%
o	15083	15.6%
u	15083	15.6%
n	3552	3.7%
a	3552	3.7%
t	3552	3.7%

Text

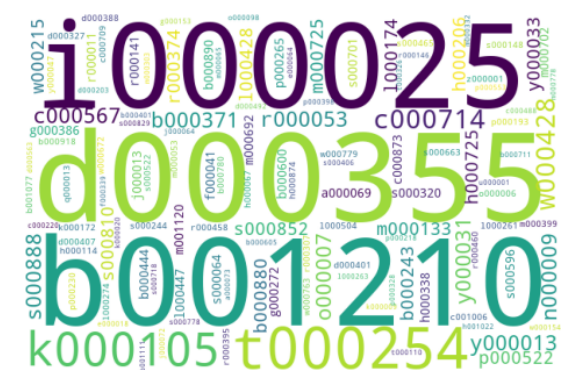
Distinct	3192
Distinct (%)	17.1%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB



Length		Characters and Unicode		Unique			Sample	
Max length	7	Total	130445	Unique	574	?	1st row	M000112
Median length	7	characters		Unique (%)	3.1%		2nd row	D000448
Mean length	7	Distinct	35				3rd row	S000001
Min length	7	characters					4th row	E000023
		Distinct	2			?	5th row	L000296
		categories	(https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)					
		Distinct	2			?		
		scripts	(https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)					
		Distinct	1 (https://en.wikipedia.org/wiki/Unicode_block)			?		
		blocks						

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

Value	Count	Frequency (%)
d000355	30	0.2%
b001210	29	0.2%
i000025	27	0.1%
t000254	26	0.1%
k000105	25	0.1%
c000714	25	0.1%
w000428	24	0.1%
y000013	24	0.1%
y000031	22	0.1%
r000053	22	0.1%
Other values (3182)	18381	98.6%



Most occurring characters

Value	Count	Frequency (%)
0	61913	47.5%
1	8961	6.9%
2	6652	5.1%
4	5801	4.4%
3	5791	4.4%
5	5384	4.1%
6	4636	3.6%
7	4535	3.5%
8	4271	3.3%
9	3866	3.0%
Other values (25)	18635	14.3%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	111810	85.7%
Uppercase Letter	18635	14.3%

Most frequent character per category*Uppercase Letter*

Value	Count	Frequency (%)
M	1964	10.5%
B	1936	10.4%
S	1748	9.4%
C	1448	7.8%
H	1388	7.4%
R	1048	5.6%
D	1033	5.5%
G	886	4.8%
W	885	4.7%
L	838	4.5%
Other values (15)	5461	29.3%

Decimal Number

Value	Count	Frequency (%)
-------	-------	---------------

0	61913	55.4%
1	8961	8.0%
2	6652	5.9%
4	5801	5.2%
3	5791	5.2%
5	5384	4.8%
6	4636	4.1%
7	4535	4.1%
8	4271	3.8%
9	3866	3.5%

Most occurring scripts

Value	Count	Frequency (%)
Common	111810	85.7%
Latin	18635	14.3%

Most frequent character per script

Latin

Value	Count	Frequency (%)
M	1964	10.5%
B	1936	10.4%
S	1748	9.4%
C	1448	7.8%
H	1388	7.4%
R	1048	5.6%
D	1033	5.5%
G	886	4.8%
W	885	4.7%
L	838	4.5%
Other values (15)	5461	29.3%

Common

Value	Count	Frequency (%)
0	61913	55.4%
1	8961	8.0%
2	6652	5.9%

4	5801	5.2%
3	5791	5.2%
5	5384	4.8%
6	4636	4.1%
7	4535	4.1%
8	4271	3.8%
9	3866	3.5%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	130445	100.0%

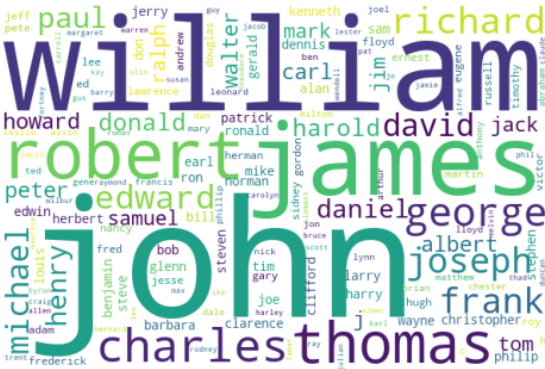
Most frequent character per block

ASCII

Value	Count	Frequency (%)
0	61913	47.5%
1	8961	6.9%
2	6652	5.1%
4	5801	4.4%
3	5791	4.4%
5	5384	4.1%
6	4636	3.6%
7	4535	3.5%
8	4271	3.3%
9	3866	3.0%
Other values (25)	18635	14.3%

firstname
Text

Distinct	810
Distinct (%)	4.3%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB



Length	
Max length	12
Median length	10
Mean length	5.4948752
Min length	2

Characters and Unicode		Unique		Sample	
Total	102397	Unique	95	?	1st row
characters		Unique (%)	0.5%		2nd row
Distinct	57				3rd row
characters					4th row
Distinct	5			?	5th row
categories	(https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)				
Distinct	2			?	
scripts	(https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)				
Distinct	2	(https://en.wikipedia.org/wiki/Unicode_block)		?	
blocks					

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

Value	Count	Frequency (%)
john	1453	7.8%
william	935	5.0%
james	855	4.6%
robert	753	4.0%
thomas	512	2.7%
charles	488	2.6%
george	355	1.9%
richard	333	1.8%
joseph	314	1.7%
frank	308	1.7%
Other values (797)	12344	66.2%



Most occurring characters

Value	Count	Frequency (%)
e	10027	9.8%
a	9932	9.7%
r	8224	8.0%
l	6930	6.8%
o	6800	6.6%
n	6715	6.6%
i	6233	6.1%
h	4639	4.5%
s	3623	3.5%
J	3506	3.4%
Other values (47)	35768	34.9%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	83449	81.5%
Uppercase Letter	18667	18.2%
Other Punctuation	263	0.3%
Space Separator	15	< 0.1%
Dash Punctuation	3	< 0.1%

Most frequent character per category*Lowercase Letter*

Value	Count	Frequency (%)
e	10027	12.0%
a	9932	11.9%
r	8224	9.9%
l	6930	8.3%
o	6800	8.1%
n	6715	8.0%
i	6233	7.5%
h	4639	5.6%
s	3623	4.3%
m	3486	4.2%
Other values (18)	16840	20.2%

Uppercase Letter

Value	Count	Frequency (%)
J	3506	18.8%
R	1682	9.0%
C	1398	7.5%
W	1346	7.2%
D	1066	5.7%
M	983	5.3%
T	968	5.2%
H	934	5.0%
E	909	4.9%
G	803	4.3%
Other values (16)	5072	27.2%

Other Punctuation

Value	Count	Frequency (%)
.	263	100.0%

Space Separator

Value	Count	Frequency (%)
	15	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	3	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	102116	99.7%
Common	281	0.3%

Most frequent character per script*Latin*

Value	Count	Frequency (%)
e	10027	9.8%

a	9932	9.7%
r	8224	8.1%
l	6930	6.8%
o	6800	6.7%
n	6715	6.6%
i	6233	6.1%
h	4639	4.5%
s	3623	3.5%
J	3506	3.4%
Other values (44)	35487	34.8%

Common

Value	Count	Frequency (%)
.	263	93.6%
	15	5.3%
-	3	1.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	102363	> 99.9%
None	34	< 0.1%

Most frequent character per block*ASCII*

Value	Count	Frequency (%)
e	10027	9.8%
a	9932	9.7%
r	8224	8.0%
l	6930	6.8%
o	6800	6.6%
n	6715	6.6%
i	6233	6.1%
h	4639	4.5%
s	3623	3.5%
J	3506	3.4%
Other values (45)	35734	34.9%

None

Value	Count	Frequency (%)
é	26	76.5%
ú	8	23.5%

middlename

Text

MISSING

Distinct	955
Distinct (%)	6.3%
Missing	3536
Missing (%)	19.0%
Memory size	145.7 KiB



Length		Characters and Unicode		Unique		Sample		
Max length	16	Total characters	76342	Unique	133	?	1st row	Jefferson
Median length	15	Distinct characters	57	Unique (%)	0.9%		2nd row	Lee
Mean length	5.0560964						3rd row	Joachim
Min length	1	Distinct categories	6 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)			?	4th row	Aubrey
							5th row	A.
		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.

Value	Count	Frequency (%)
e	423	2.8%
j	387	2.5%
joseph	384	2.5%
a	380	2.5%
m	330	2.2%
william	328	2.1%
edward	309	2.0%
l	290	1.9%
h	236	1.5%
james	227	1.5%
Other values (940)	12016	78.5%



Most occurring characters

Value	Count	Frequency (%)
e	6964	9.1%
a	5647	7.4%
r	5633	7.4%
n	4782	6.3%
l	4673	6.1%
o	4354	5.7%
.	4132	5.4%
i	3888	5.1%
s	3128	4.1%
t	2736	3.6%
Other values (47)	30405	39.8%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	56513	74.0%
Uppercase Letter	15466	20.3%
Other Punctuation	4150	5.4%
Space Separator	211	0.3%
Open Punctuation	1	< 0.1%
Close Punctuation	1	< 0.1%

Most frequent character per category*Lowercase Letter*

Value	Count	Frequency (%)
e	6964	12.3%
a	5647	10.0%
r	5633	10.0%
n	4782	8.5%
l	4673	8.3%
o	4354	7.7%
i	3888	6.9%
s	3128	5.5%
t	2736	4.8%

d	2426	4.3%
Other values (17)	12282	21.7%
<i>Uppercase Letter</i>		
Value	Count	Frequency (%)
J	1500	9.7%
E	1274	8.2%
L	1145	7.4%
M	1124	7.3%
A	1106	7.2%
H	1063	6.9%
W	953	6.2%
B	910	5.9%
C	844	5.5%
R	799	5.2%
Other values (15)	4748	30.7%
<i>Other Punctuation</i>		
Value	Count	Frequency (%)
.	4132	99.6%
'	18	0.4%
<i>Space Separator</i>		
Value	Count	Frequency (%)
	211	100.0%
<i>Open Punctuation</i>		
Value	Count	Frequency (%)
(1	100.0%
<i>Close Punctuation</i>		
Value	Count	Frequency (%)
)	1	100.0%
<i>Most occurring scripts</i>		
Value	Count	Frequency (%)
Latin	71979	94.3%
Common	4363	5.7%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	6964	9.7%
a	5647	7.8%
r	5633	7.8%
n	4782	6.6%
l	4673	6.5%
o	4354	6.0%
i	3888	5.4%
s	3128	4.3%
t	2736	3.8%
d	2426	3.4%
Other values (42)	27748	38.6%

Common

Value	Count	Frequency (%)
.	4132	94.7%
	211	4.8%
'	18	0.4%
(1	< 0.1%
)	1	< 0.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	76336	> 99.9%
None	6	< 0.1%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	6964	9.1%
a	5647	7.4%
r	5633	7.4%
n	4782	6.3%

l	4673	6.1%
o	4354	5.7%
.	4132	5.4%
i	3888	5.1%
s	3128	4.1%
t	2736	3.6%
Other values (46)	30399	39.8%

None

Value	Count	Frequency (%)
é	6	100.0%

Distinct	2289
Distinct (%)	12.3%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB



Length		Characters and Unicode		Unique		Sample		
Max length	19	Total	121102	Unique	339	?	1st row	Mansfield
Median length	17	characters		Unique (%)	1.8%		2nd row	Doughton
Mean length	6.4986316	Distinct	58				3rd row	Sabath
Min length	2	characters					4th row	Eaton
		Distinct	5			?	5th row	Lewis
		categories	(https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)					
		Distinct	2			?		
		scripts	(https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)					
		Distinct	2 (https://en.wikipedia.org/wiki/Unicode_block)			?		
		blocks						

Value	Count	Frequency (%)
smith	185	1.0%
johnson	117	0.6%
milller	115	0.6%
davis	93	0.5%
young	92	0.5%
brown	90	0.5%
jones	82	0.4%
rogers	71	0.4%
wilson	71	0.4%
thompson	66	0.4%
Other values (2286)	17812	94.8%



Most occurring characters

Value	Count	Frequency (%)
e	12294	10.2%
n	10140	8.4%
a	9051	7.5%
r	9043	7.5%
o	8510	7.0%
l	7464	6.2%
i	6277	5.2%
s	5974	4.9%
t	5402	4.5%
h	3559	2.9%
Other values (48)	43388	35.8%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	101102	83.5%
Uppercase Letter	19621	16.2%
Space Separator	167	0.1%
Other Punctuation	148	0.1%
Dash Punctuation	64	0.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	12294	12.2%
n	10140	10.0%
a	9051	9.0%
r	9043	8.9%
o	8510	8.4%
l	7464	7.4%
i	6277	6.2%
s	5974	5.9%
t	5402	5.3%
h	3559	3.5%
Other values (19)	23388	23.1%

Uppercase Letter

Value	Count	Frequency (%)
M	2031	10.4%
B	2002	10.2%
S	1763	9.0%
C	1698	8.7%
H	1454	7.4%
D	1082	5.5%
R	1056	5.4%
G	995	5.1%
W	900	4.6%
L	886	4.5%
Other values (15)	5754	29.3%

Other Punctuation

Value	Count	Frequency (%)
'	121	81.8%
.	27	18.2%

Space Separator

Value	Count	Frequency (%)
	167	100.0%

Dash Punctuation

Value	Count	Frequency (%)
-	64	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	120723	99.7%
Common	379	0.3%

Most frequent character per script*Latin*

Value	Count	Frequency (%)
-------	-------	---------------

e	12294	10.2%
n	10140	8.4%
a	9051	7.5%
r	9043	7.5%
o	8510	7.0%
l	7464	6.2%
i	6277	5.2%
s	5974	4.9%
t	5402	4.5%
h	3559	2.9%
Other values (44)	43009	35.6%

Common

Value	Count	Frequency (%)
	167	44.1%
'	121	31.9%
-	64	16.9%
.	27	7.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	120999	99.9%
None	103	0.1%

Most frequent character per block**ASCII**

Value	Count	Frequency (%)
e	12294	10.2%
n	10140	8.4%
a	9051	7.5%
r	9043	7.5%
o	8510	7.0%
l	7464	6.2%
i	6277	5.2%
s	5974	4.9%
t	5402	4.5%

h	3559	2.9%
Other values (45)	43285	35.8%

None

Value	Count	Frequency (%)
á	55	53.4%
é	38	36.9%
í	10	9.7%

suffix

Categorical

IMBALANCE MISSING

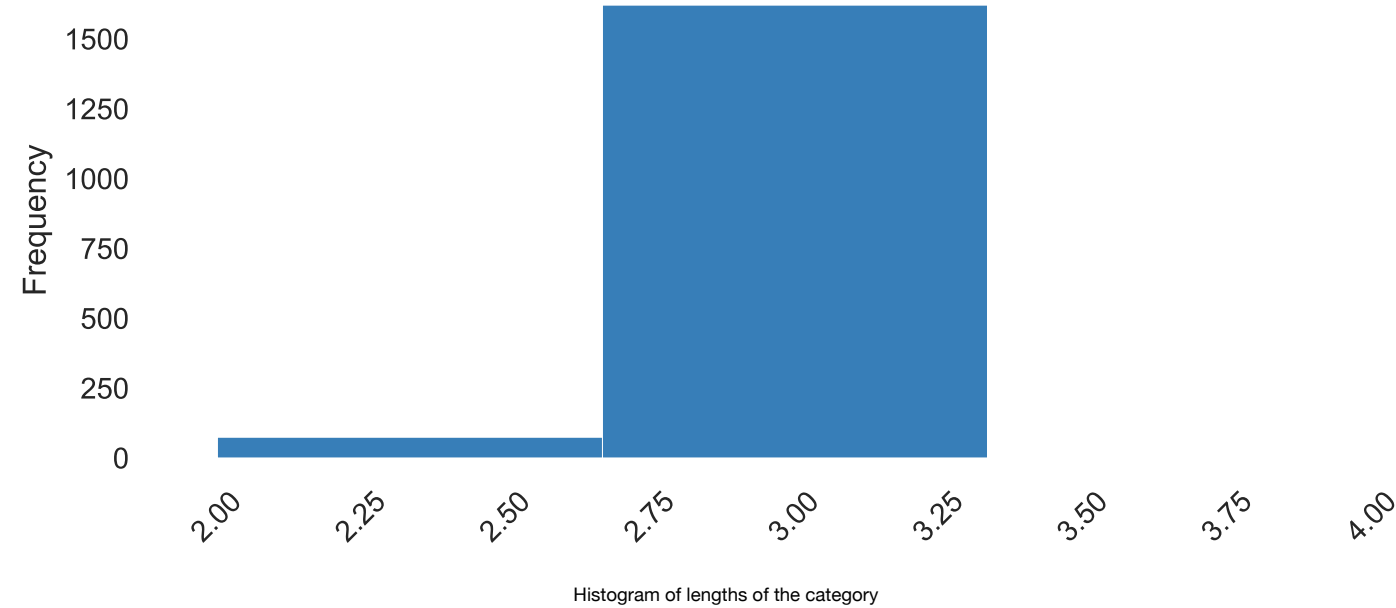
Distinct	7
Distinct (%)	0.4%
Missing	16937
Missing (%)	90.9%
Memory size	145.7 KiB

Length		Characters and Unicode		Unique		Sample		
Max length	4	Total	5023	Unique	1	?	1st row	Jr.
Median length	3	characters		Unique (%)	0.1%		2nd row	Jr.
Mean length	2.9581861	Distinct	7				3rd row	Jr.
Min length	2	characters					4th row	Jr.
		Distinct	4			?	5th row	Jr.
		categories	(https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)					
		Distinct	2			?		
		scripts	(https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)					
		Distinct	1	(https://en.wikipedia.org/wiki/Unicode_block)		?		
		blocks						
The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.								

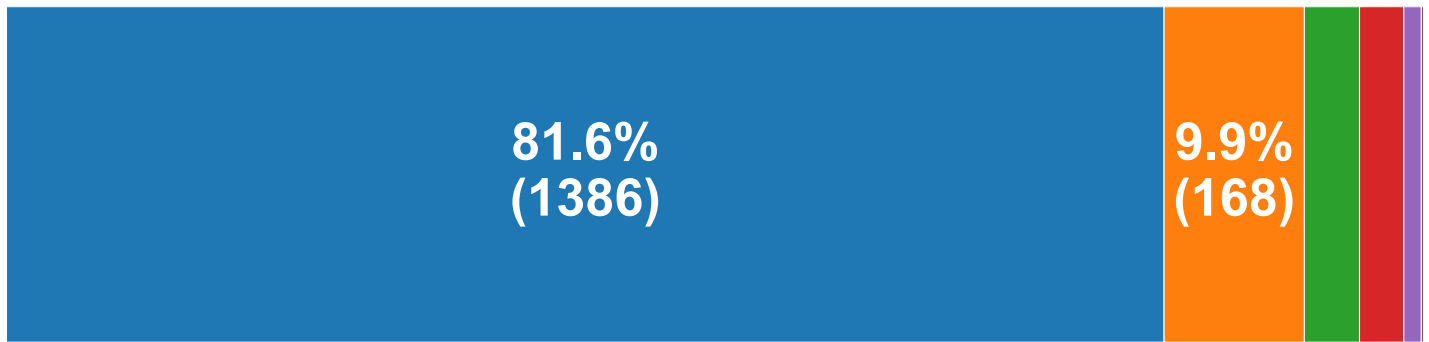
Common Values

Value	Count	Frequency (%)
Jr.	1386	7.4%
III	168	0.9%
Sr.	66	0.4%
II	53	0.3%
IV	21	0.1%
Jr.	3	< 0.1%
IV	1	< 0.1%
(Missing)	16937	90.9%

Length



Common Values (Plot)



	Jr.
	III
	Sr.
	II
	IV
	Jr.
	IV

Value	Count	Frequency (%)
jr	1389	81.8%
iii	168	9.9%
sr	66	3.9%
ii	53	3.1%
iv	22	1.3%

Most occurring characters

Value	Count	Frequency (%)
r	1455	29.0%
.	1455	29.0%
J	1389	27.7%
I	632	12.6%

S	66	1.3%
V	22	0.4%
	4	0.1%

Most occurring categories

Value	Count	Frequency (%)
Uppercase Letter	2109	42.0%
Lowercase Letter	1455	29.0%
Other Punctuation	1455	29.0%
Space Separator	4	0.1%

Most frequent character per category

Uppercase Letter

Value	Count	Frequency (%)
J	1389	65.9%
I	632	30.0%
S	66	3.1%
V	22	1.0%

Lowercase Letter

Value	Count	Frequency (%)
r	1455	100.0%

Other Punctuation

Value	Count	Frequency (%)
.	1455	100.0%

Space Separator

Value	Count	Frequency (%)
	4	100.0%

Most occurring scripts

Value	Count	Frequency (%)
Latin	3564	71.0%
Common	1459	29.0%

Most frequent character per script*Latin*

Value	Count	Frequency (%)
r	1455	40.8%
J	1389	39.0%
I	632	17.7%
S	66	1.9%
V	22	0.6%

Common

Value	Count	Frequency (%)
.	1455	99.7%
	4	0.3%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	5023	100.0%

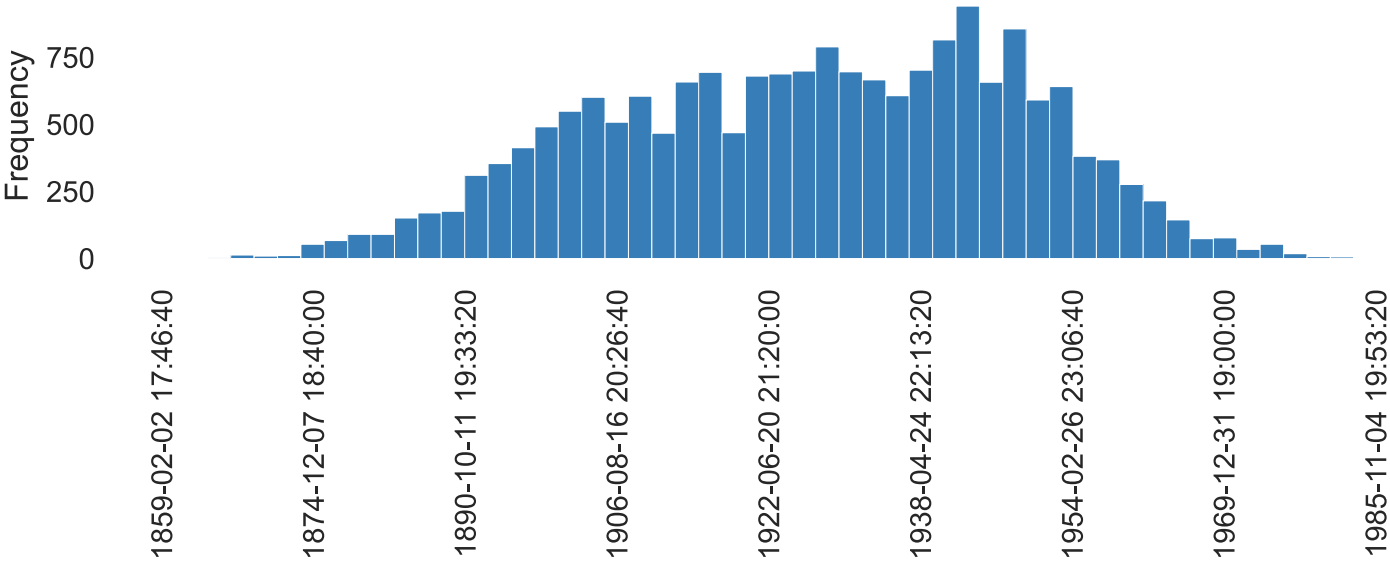
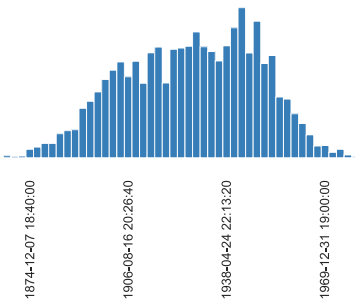
Most frequent character per block*ASCII*

Value	Count	Frequency (%)
r	1455	29.0%
.	1455	29.0%
J	1389	27.7%
I	632	12.6%
S	66	1.3%
V	22	0.4%
	4	0.1%

birthday
Date

Distinct	3064
Distinct (%)	16.4%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB

Minimum	1861-02-09 00:00:00
Maximum	1983-03-30 00:00:00



Histogram with fixed size bins (bins=50)

state
Categorical

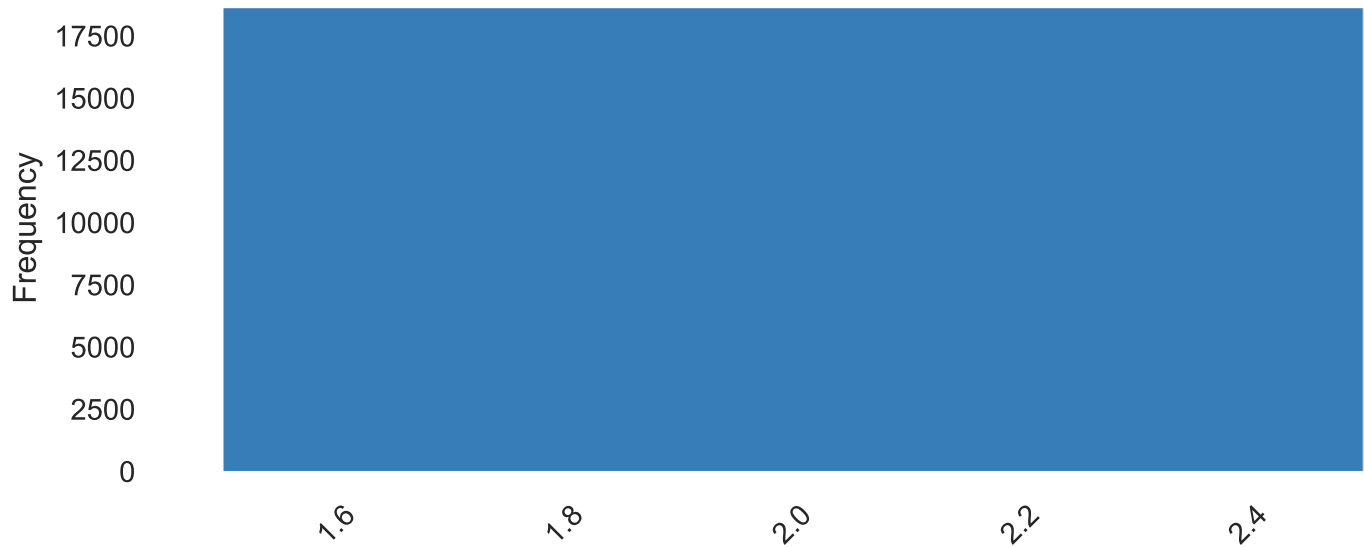
Distinct	50
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB

Length		Characters and Unicode	Unique	Sample
Max length	2	Total characters 37270	Unique 0 ?	1st row TX
Median length	2		Unique (%) 0.0%	2nd row NC
Mean length	2	Distinct characters 24		3rd row IL
Min length	2	Distinct categories 1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)		4th row NJ
		Distinct scripts 1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)		5th row KY
		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.				

Common Values

Value	Count	Frequency (%)
CA	1534	8.2%
NY	1347	7.2%
TX	985	5.3%
PA	939	5.0%
IL	849	4.6%
OH	809	4.3%
MI	670	3.6%
FL	629	3.4%
NJ	556	3.0%
NC	481	2.6%
Other values (40)	9836	52.8%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
ca	1534	8.2%
ny	1347	7.2%
tx	985	5.3%
pa	939	5.0%
il	849	4.6%
oh	809	4.3%
mi	670	3.6%
fl	629	3.4%
nj	556	3.0%
nc	481	2.6%
Other values (40)	9836	52.8%

Most occurring characters

Value	Count	Frequency (%)
A	5662	15.2%
N	4279	11.5%
I	3013	8.1%
M	2934	7.9%
C	2818	7.6%
L	2155	5.8%
T	2026	5.4%
O	1975	5.3%

Y	1778	4.8%
H	1074	2.9%
Other values (14)	9556	25.6%

Most occurring categories

Value	Count	Frequency (%)
Uppercase Letter	37270	100.0%

Most frequent character per category

Uppercase Letter

Value	Count	Frequency (%)
A	5662	15.2%
N	4279	11.5%
I	3013	8.1%
M	2934	7.9%
C	2818	7.6%
L	2155	5.8%
T	2026	5.4%
O	1975	5.3%
Y	1778	4.8%
H	1074	2.9%
Other values (14)	9556	25.6%

Most occurring scripts

Value	Count	Frequency (%)
Latin	37270	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
A	5662	15.2%
N	4279	11.5%
I	3013	8.1%
M	2934	7.9%
C	2818	7.6%

L	2155	5.8%
T	2026	5.4%
O	1975	5.3%
Y	1778	4.8%
H	1074	2.9%
Other values (14)	9556	25.6%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	37270	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
A	5662	15.2%
N	4279	11.5%
I	3013	8.1%
M	2934	7.9%
C	2818	7.6%
L	2155	5.8%
T	2026	5.4%
O	1975	5.3%
Y	1778	4.8%
H	1074	2.9%
Other values (14)	9556	25.6%

party
Categorical

IMBALANCE

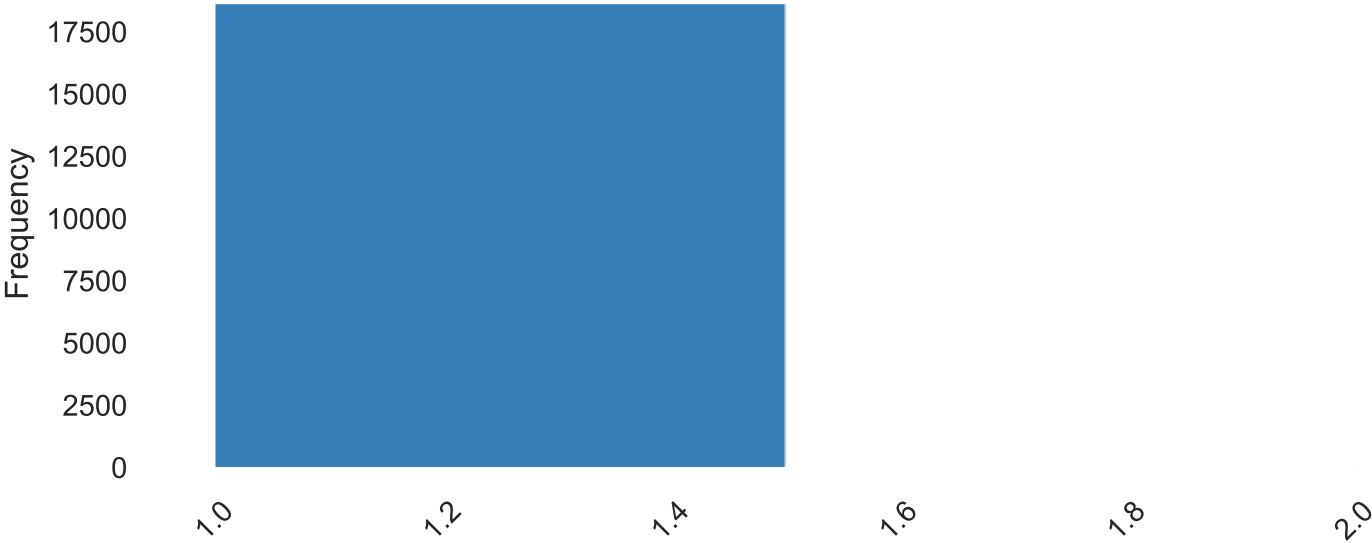
Distinct	6
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB

Length		Characters and Unicode	Unique	Sample
Max length	2	Total characters 18642	Unique 1 ?	1st row D
Median length	1		Unique (%) < 0.1%	2nd row D
Mean length	1.0003756	Distinct characters 5		3rd row D
Min length	1	Distinct categories 1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)		4th row R
		Distinct scripts 1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)		5th row R
		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.				

Common Values

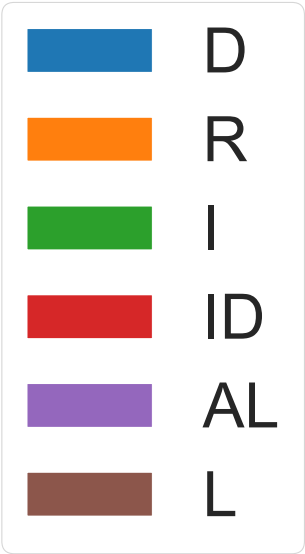
Value	Count	Frequency (%)
D	10290	55.2%
R	8274	44.4%
I	63	0.3%
ID	4	< 0.1%
AL	3	< 0.1%
L	1	< 0.1%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
-------	-------	---------------

d	10290	55.2%
r	8274	44.4%
i	63	0.3%
id	4	< 0.1%
al	3	< 0.1%
l	1	< 0.1%

Most occurring characters

Value	Count	Frequency (%)
D	10294	55.2%
R	8274	44.4%
I	67	0.4%
L	4	< 0.1%
A	3	< 0.1%

Most occurring categories

Value	Count	Frequency (%)
Uppercase Letter	18642	100.0%

Most frequent character per category

Uppercase Letter

Value	Count	Frequency (%)
D	10294	55.2%
R	8274	44.4%
I	67	0.4%
L	4	< 0.1%
A	3	< 0.1%

Most occurring scripts

Value	Count	Frequency (%)
Latin	18642	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
D	10294	55.2%
R	8274	44.4%
I	67	0.4%
L	4	< 0.1%
A	3	< 0.1%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	18642	100.0%

Most frequent character per block

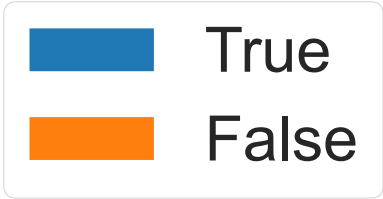
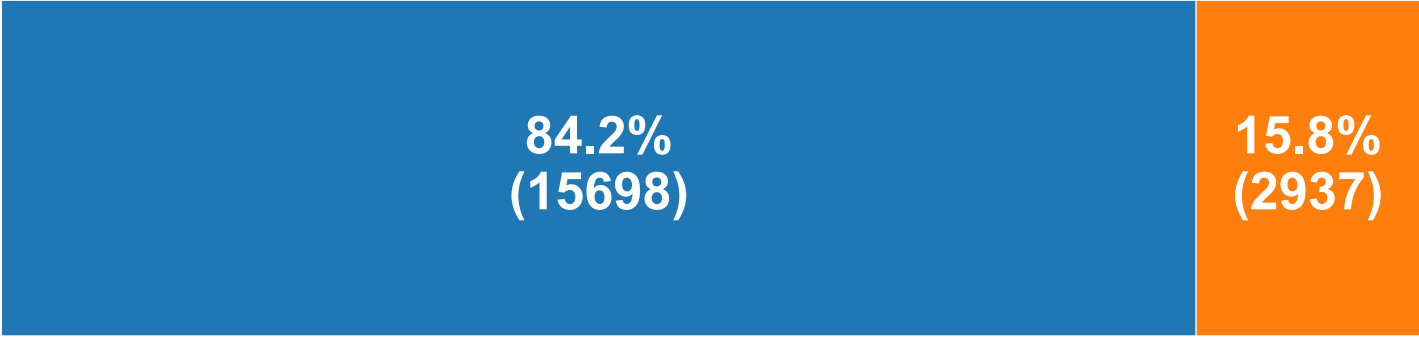
ASCII

Value	Count	Frequency (%)
D	10294	55.2%
R	8274	44.4%
I	67	0.4%
L	4	< 0.1%
A	3	< 0.1%

incumbent
Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	18.3 KiB

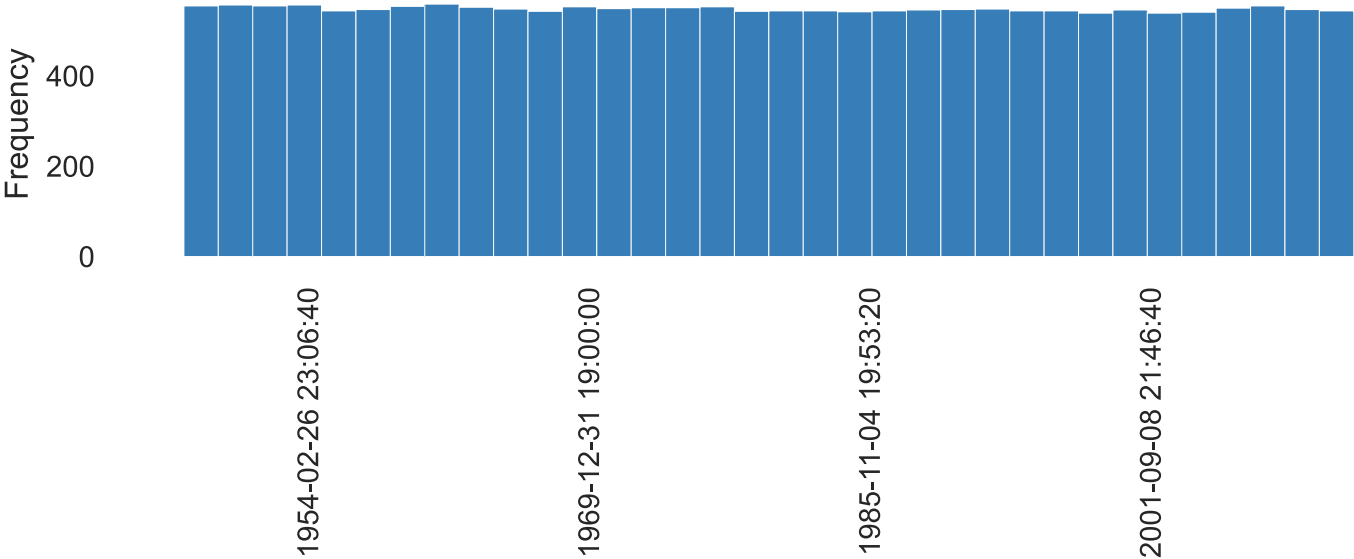
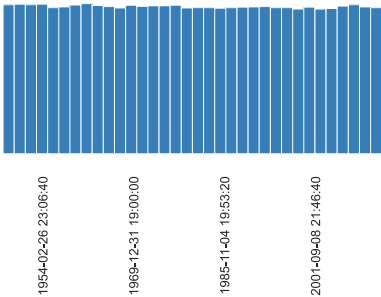
Value	Count	Frequency (%)
True	15698	84.2%
False	2937	15.8%



termstart
Date

Distinct	34
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	145.7 KiB

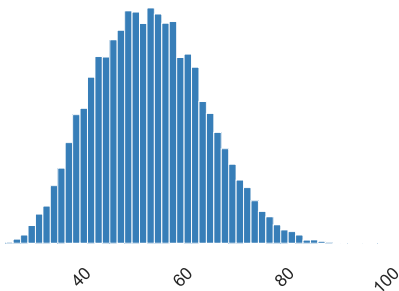
Minimum	1947-01-03 00:00:00
Maximum	2013-01-03 00:00:00



Histogram with fixed size bins (bins=34)

age
Real number (\mathbb{R})

Distinct	605	Minimum	25
Distinct (%)	3.2%	Maximum	98.1
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	53.313732	Memory size	145.7 KiB

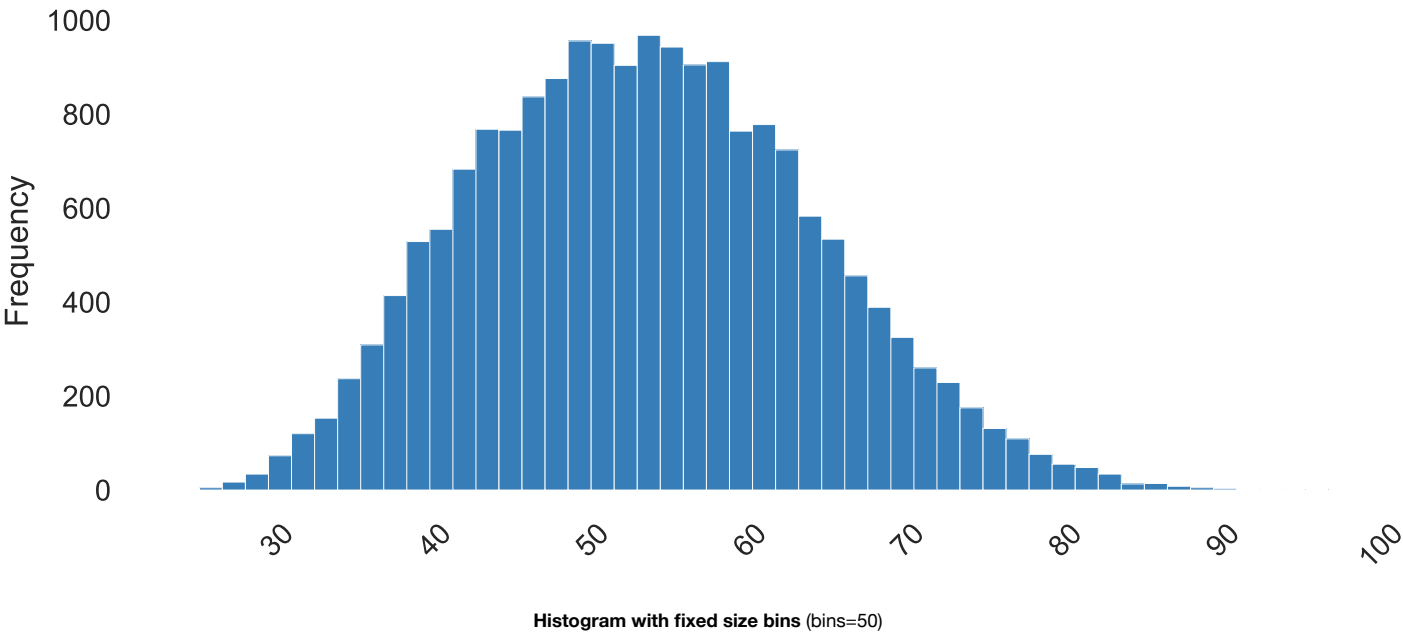


Quantile statistics

Minimum	25
5-th percentile	36.6
Q1	45.4
median	53
Q3	60.55
95-th percentile	71.5
Maximum	98.1
Range	73.1
Interquartile range (IQR)	15.15

Descriptive statistics

Standard deviation	10.678469
Coefficient of variation (CV)	0.20029491
Kurtosis	-0.26081871
Mean	53.313732
Median Absolute Deviation (MAD)	7.6
Skewness	0.23155383
Sum	993501.4
Variance	114.0297
Monotonicity	Not monotonic



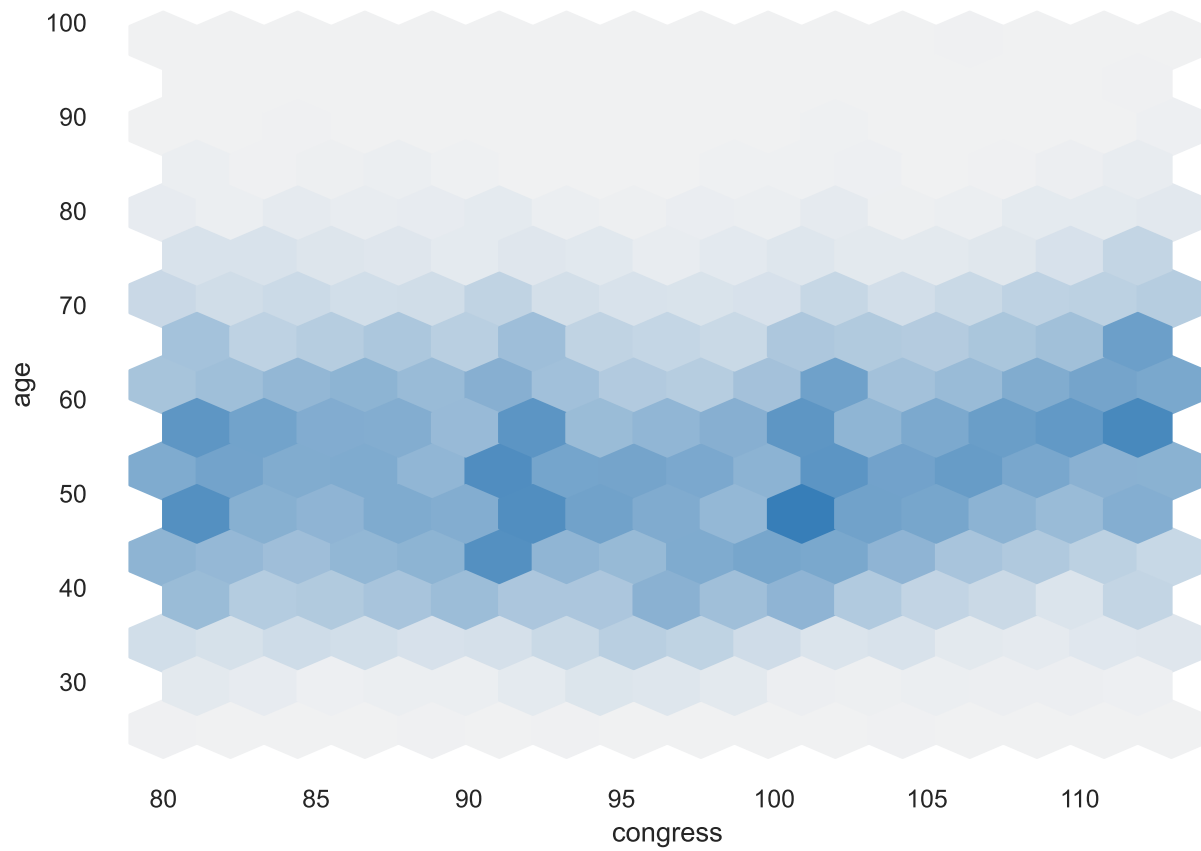
Value	Count	Frequency (%)
56.3	84	0.5%
51.3	80	0.4%
53.3	78	0.4%

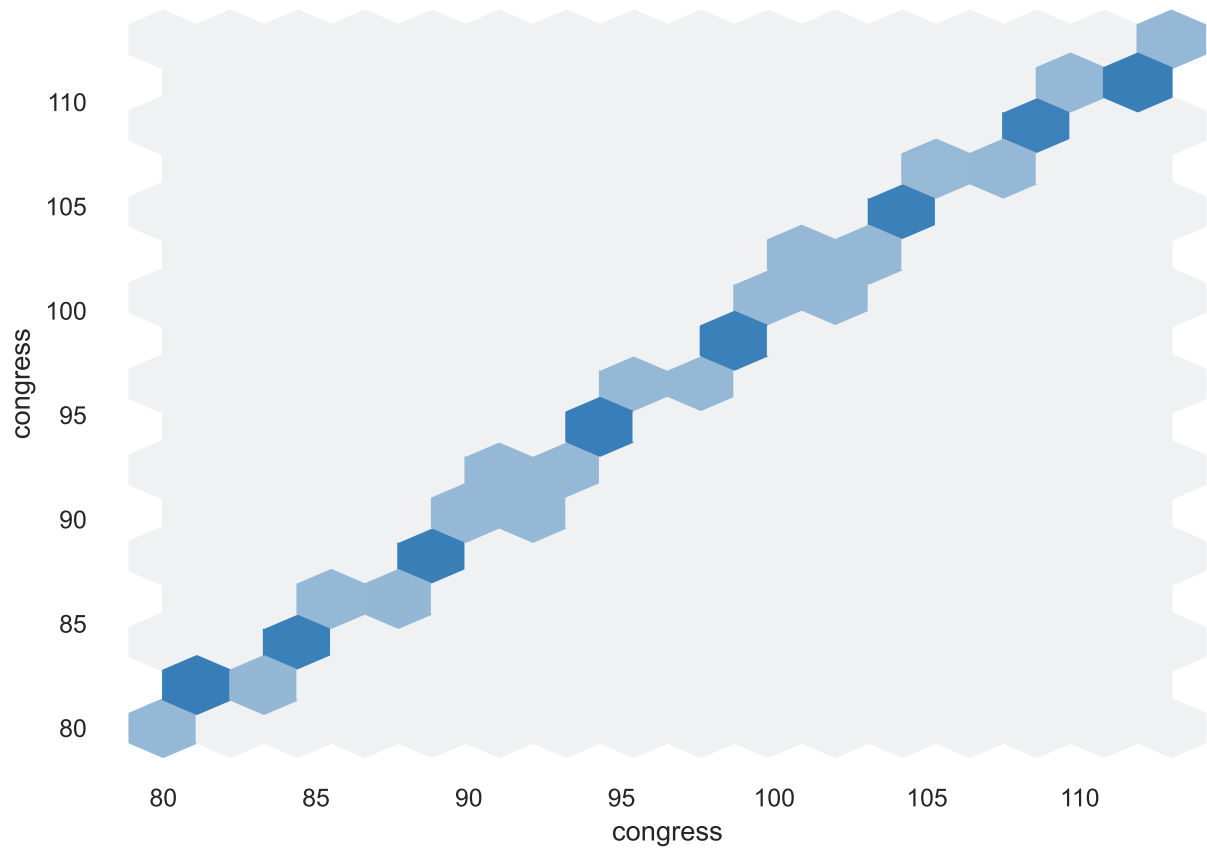
58.3	77	0.4%
52.3	77	0.4%
49.3	77	0.4%
50.5	76	0.4%
54.2	76	0.4%
55.3	76	0.4%
52.2	76	0.4%
Other values (595)	17858	95.8%

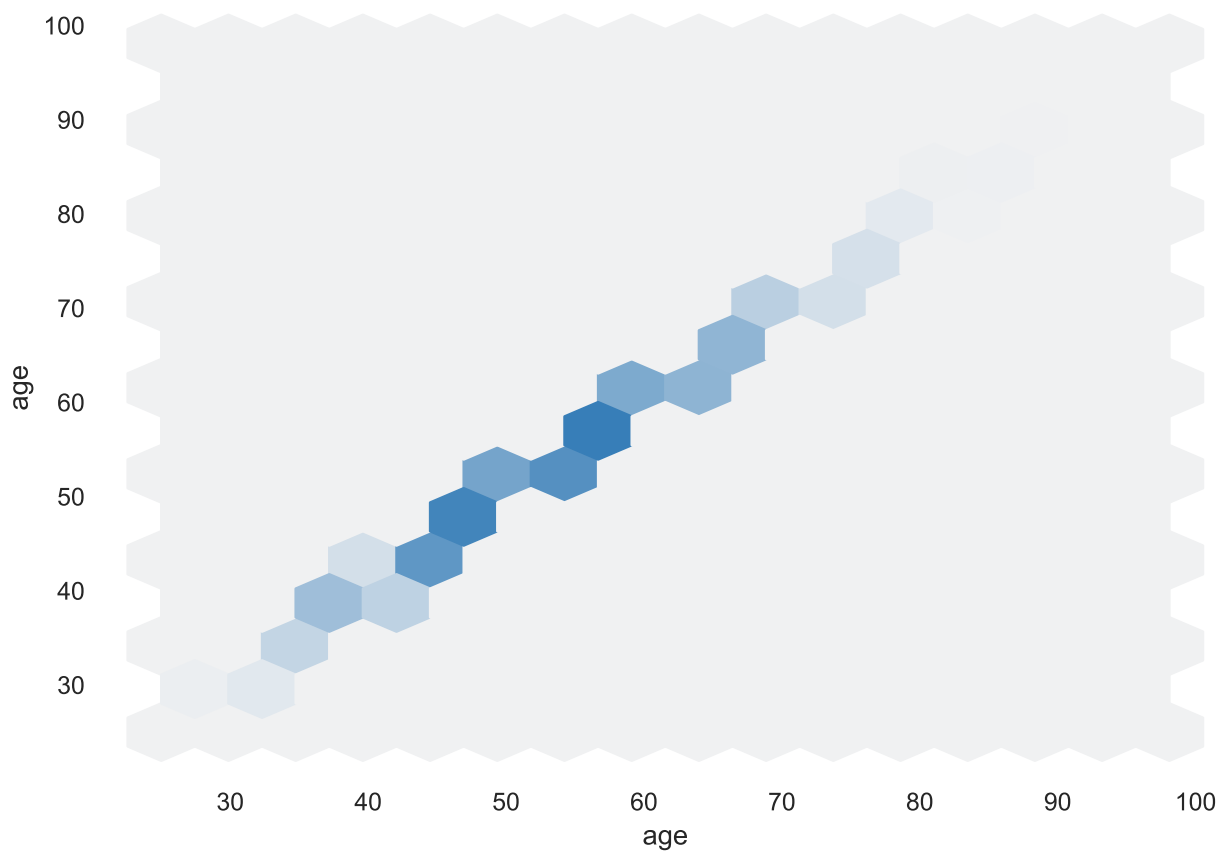
Value	Count	Frequency (%)
25	1	< 0.1%
25.9	1	< 0.1%
26	1	< 0.1%
26.4	2	< 0.1%
26.5	1	< 0.1%
26.7	2	< 0.1%
26.8	1	< 0.1%
26.9	1	< 0.1%
27.2	1	< 0.1%
27.3	1	< 0.1%

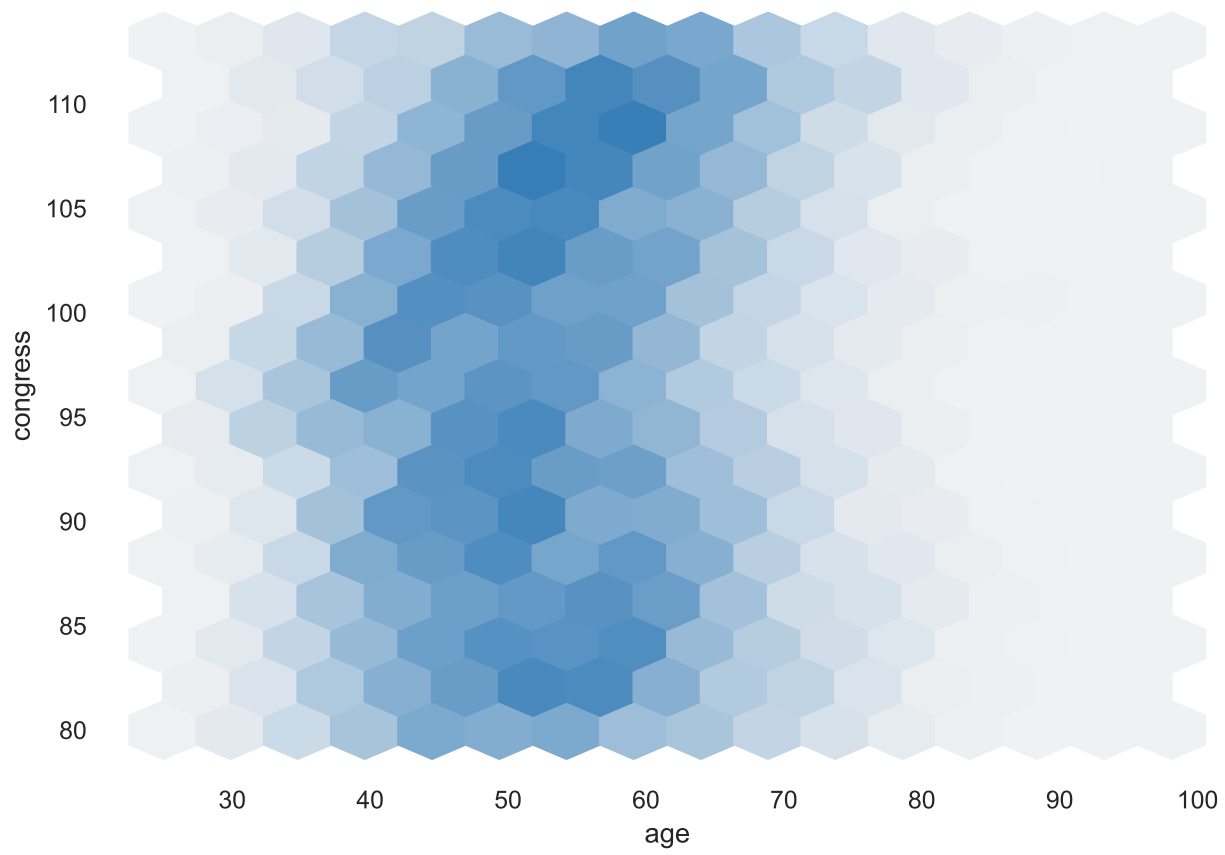
Value	Count	Frequency (%)
98.1	1	< 0.1%
96.1	1	< 0.1%
95.1	1	< 0.1%
94.1	1	< 0.1%
93	1	< 0.1%
92.1	1	< 0.1%
91.3	1	< 0.1%
90.3	1	< 0.1%
90.1	1	< 0.1%
89.7	1	< 0.1%

Interactions

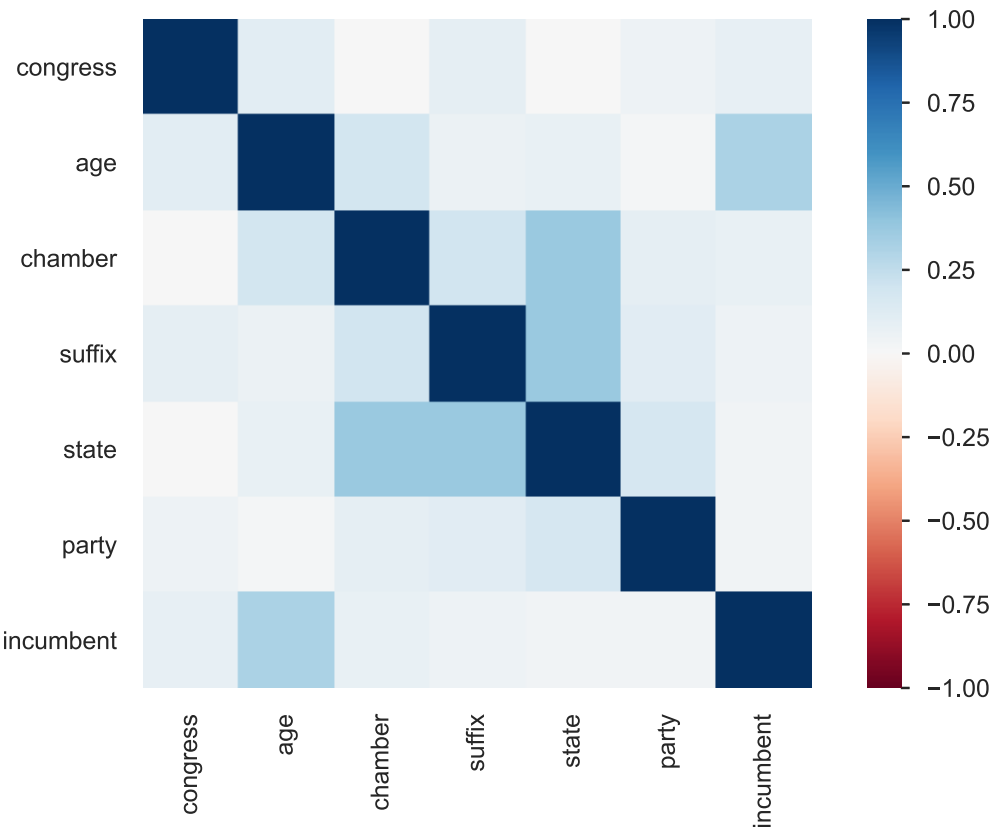






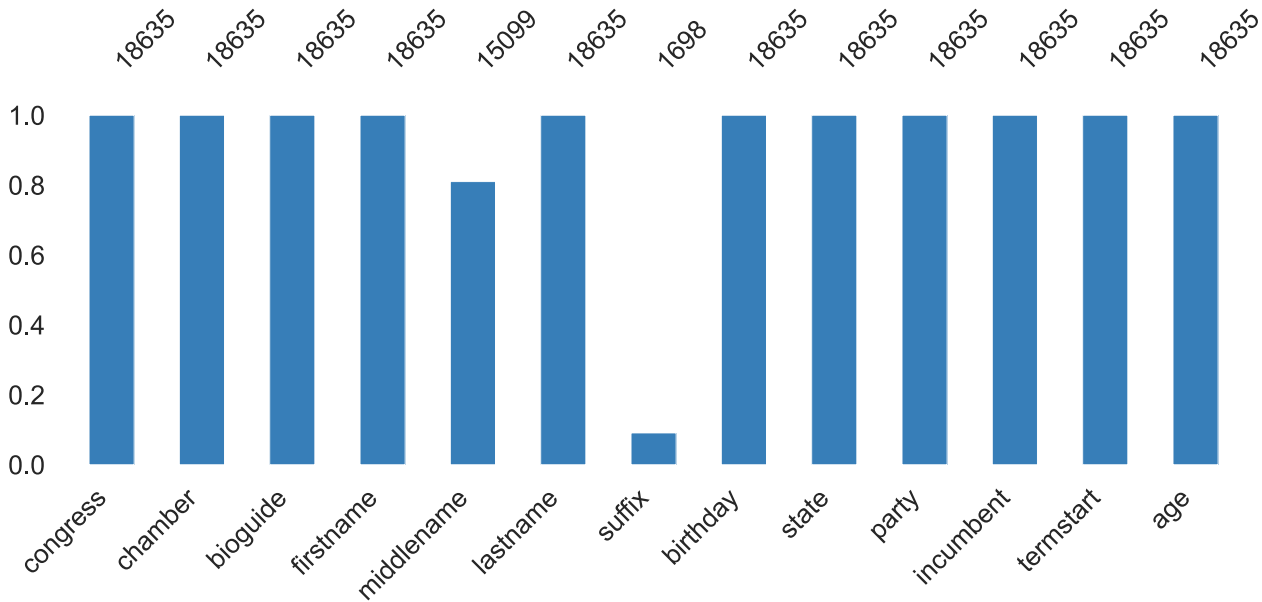


Correlations

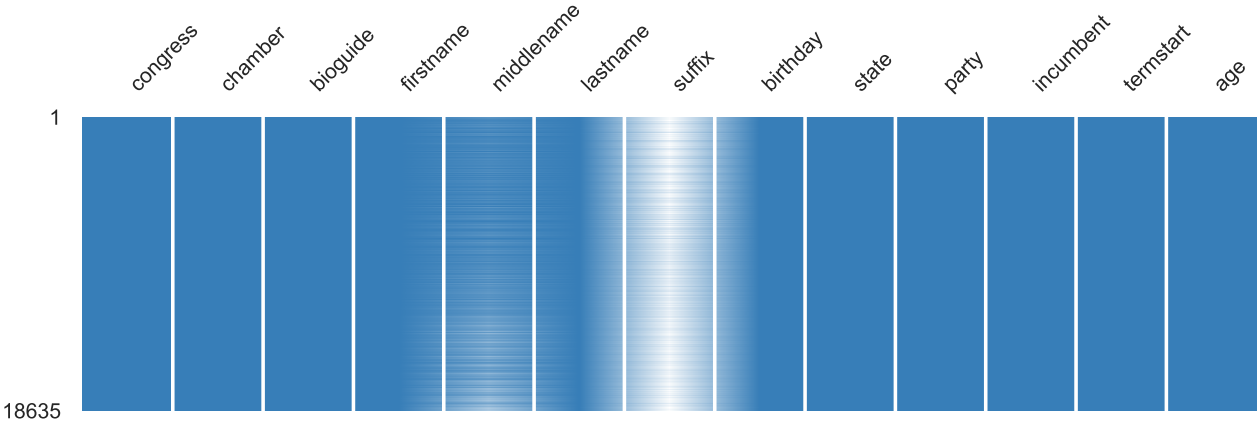


	congress	age	chamber	suffix	state	party	incumbent
congress	1.000	0.105	0.000	0.089	0.000	0.050	0.081
age	0.105	1.000	0.183	0.058	0.071	0.021	0.315
chamber	0.000	0.183	1.000	0.191	0.368	0.090	0.074
suffix	0.089	0.058	0.191	1.000	0.372	0.117	0.047
state	0.000	0.071	0.368	0.372	1.000	0.170	0.037
party	0.050	0.021	0.090	0.117	0.170	1.000	0.034
incumbent	0.081	0.315	0.074	0.047	0.037	0.034	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

congress		chamber	bioguide	firstname	middlename	lastname	suffix	birthday	state	party	incumbent	termstart	age
0	80	house	M000112	Joseph	Jefferson	Mansfield	NaN	1861-02-09	TX	D	Yes	1947-01-03	85.9
1	80	house	D000448	Robert	Lee	Doughton	NaN	1863-11-07	NC	D	Yes	1947-01-03	83.2
2	80	house	S000001	Adolph	Joachim	Sabath	NaN	1866-04-04	IL	D	Yes	1947-01-03	80.7
3	80	house	E000023	Charles	Aubrey	Eaton	NaN	1868-03-29	NJ	R	Yes	1947-01-03	78.8
4	80	house	L000296	William	NaN	Lewis	NaN	1868-09-22	KY	R	No	1947-01-03	78.3
5	80	house	G000017	James	A.	Gallagher	NaN	1869-01-16	PA	R	No	1947-01-03	78.0
6	80	house	W000265	Richard	Joseph	Welch	NaN	1869-02-13	CA	R	Yes	1947-01-03	77.9
7	80	house	B000565	Sol	NaN	Bloom	NaN	1870-03-09	NY	D	Yes	1947-01-03	76.8
8	80	house	H000943	Merlin	NaN	Hull	NaN	1870-12-18	WI	R	Yes	1947-01-03	76.0
9	80	house	G000169	Charles	Lacelle	Gifford	NaN	1871-03-15	MA	R	Yes	1947-01-03	75.8

congress		chamber	bioguide	firstname	middlename	lastname	suffix	birthday	state	party	incumbent	termstart	age
18625	113	senate	R000595	Marco	NaN	Rubio	NaN	1971-05-28	FL	R	Yes	2013-01-03	41.6
18626	113	senate	L000577	Mike	NaN	Lee	NaN	1971-06-04	UT	R	Yes	2013-01-03	41.6
18627	113	senate	H001046	Martin	NaN	Heinrich	NaN	1971-10-17	NM	D	Yes	2013-01-03	41.2
18628	113	senate	S001194	Brian	Emanuel	Schatz	NaN	1972-10-20	HI	D	Yes	2013-01-03	40.2
18629	113	senate	M001169	Christopher	S.	Murphy	NaN	1973-08-03	CT	D	Yes	2013-01-03	39.4
18630	113	senate	B000243	Max	S.	Baucus	NaN	1937-12-10	MT	D	Yes	2013-01-03	75.1
18631	113	senate	C001099	William	NaN	Cowan	NaN	1965-04-03	MA	D	No	2013-01-03	47.8

18632	113	senate	C001100	Jeffrey	Scott	Chiesa	NaN	1961-06-21	NJ	R	No	2013-01-03	51.5
18633	113	senate	K000148	John	Forbes	Kerry	NaN	1939-12-10	MA	D	Yes	2013-01-03	73.1
18634	113	senate	L000123	Frank	R.	Lautenberg	NaN	1920-01-22	NJ	D	Yes	2013-01-03	93.0

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