Overview

Brought to you by YData (https://ydata.ai/?utm_source=opensource&utm_medium=ydataprofiling&utm_campaign=report)

Dataset statistics

Number of variables	7
Number of observations	891
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	57
Duplicate rows (%)	6.4%
Total size in memory	48.9 KiB
Average record size in memory	56.1 B

Variable types

Categorical	4
Numeric	3

Alerts

Dataset has 57 (6.4%) duplicate rows	Duplicates
`Fare` is highly overall correlated with `famnp`	High correlation
`Sex` is highly overall correlated with `Survived`	High correlation
`Survived` is highly overall correlated with `Sex`	High correlation
`famnp` is highly overall correlated with `Fare`	High correlation
`Fare` has 15 (1.7%) zeros	Zeros
`famnp` has 537 (60.3%) zeros	Zeros

Reproduction

Analysis started	2024-08-16 21:23:13.430392
Analysis finished	2024-08-16 21:23:19.815448
Duration	6.39 seconds
Software version	ydata-profiling vv4.9.0 (https://github.com/ydataai/ydata-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%2

Variables

Select Columns

Survived

Categorical

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Sex)

Distinct	2
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	7.1 KiB

Length

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

Characters and Unicode

Total characters	891
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%

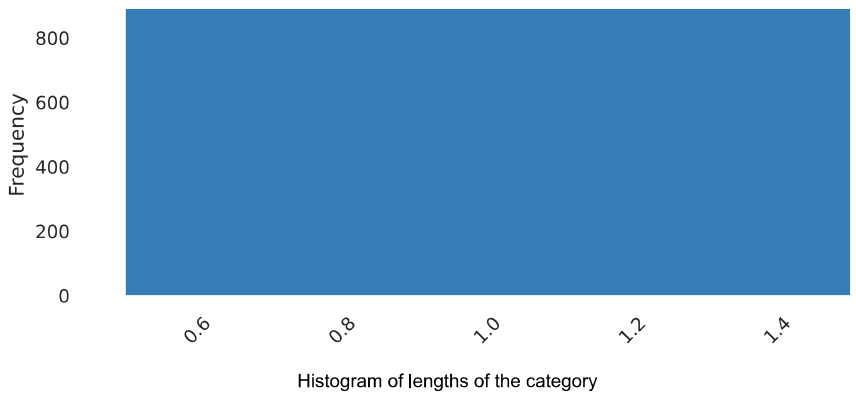
Sample

1st row	0
2nd row	1
3rd row	1
4th row	1
5th row	0 ?

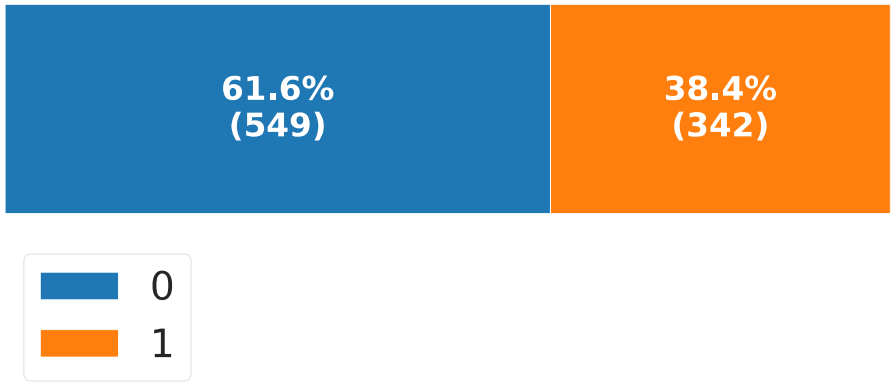
Common Values

Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Length



Common Values (Plot)



Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Most occurring characters

Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
0	549	61.6%
1	342	38.4%

Pclass

Categorical

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	7.1 KiB

Length

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

Characters and Unicode

Total characters	891
Distinct characters	3
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%

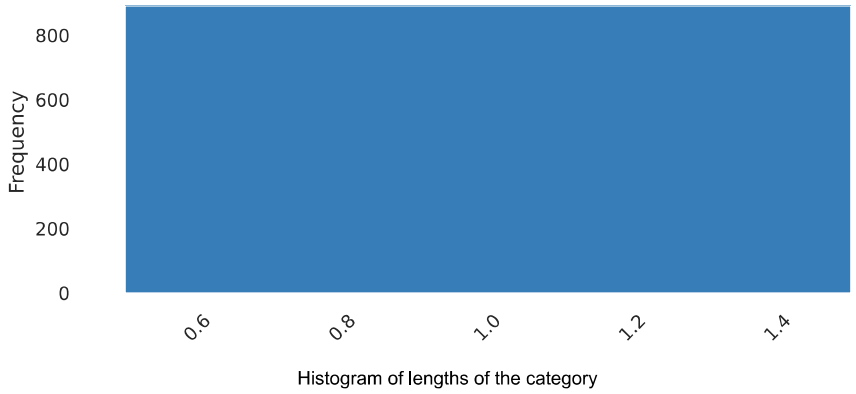
Sample

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

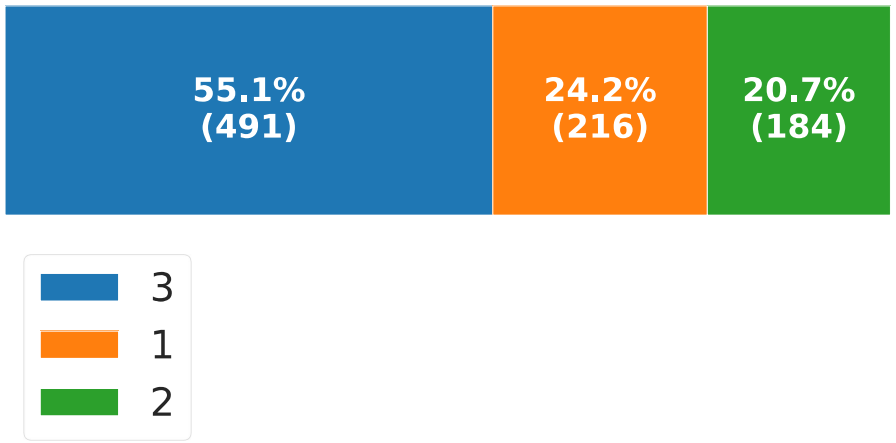
Common Values

Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%
2	184	20.7%

Length



Common Values (Plot)



Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%
2	184	20.7%

Most occurring characters

Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%

Value	Count	Frequency (%)
2	184	20.7%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%
2	184	20.7%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%
2	184	20.7%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
3	491	55.1%
1	216	24.2%
2	184	20.7%

Sex

Categorical

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Survived)

Distinct	2
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	7.1 KiB

Length

Max length	6
Median length	4
Mean length	4.704826
Min length	4

Characters and Unicode

Total characters	4192
Distinct characters	5
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%

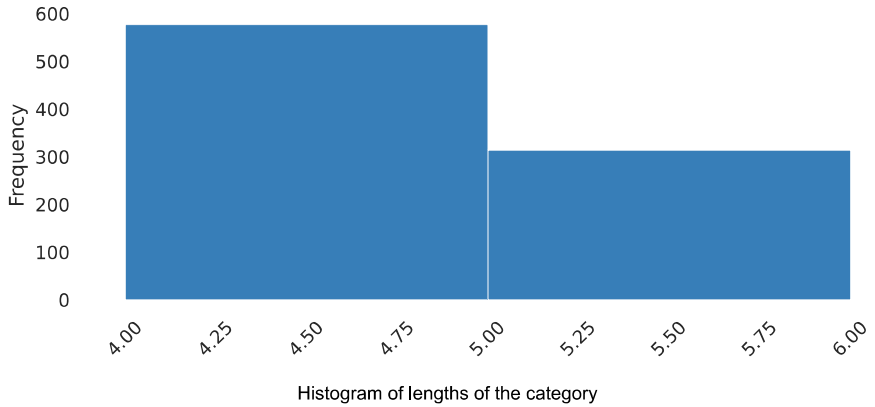
Sample

1st row	male
2nd row	female
3rd row	female
4th row	female
5th row	male

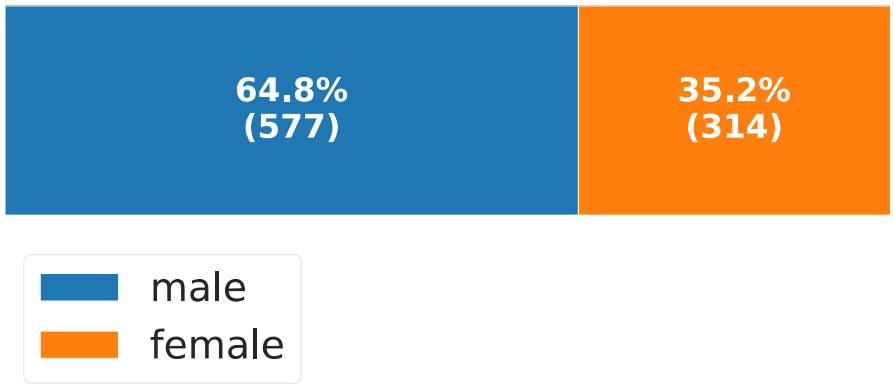
Common Values

Value	Count	Frequency (%)
male	577	64.8%
female	314	35.2%

Length



Common Values (Plot)



Value	Count	Frequency (%)
male	577	64.8%
female	314	35.2%

Most occurring characters

Value	Count	Frequency (%)
e	1205	28.7%
m	891	21.3%
a	891	21.3%
l	891	21.3%

Value	Count	Frequency (%)
f	314	7.5%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	4192	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
e	1205	28.7%
m	891	21.3%
a	891	21.3%
l	891	21.3%
f	314	7.5%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	4192	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
e	1205	28.7%
m	891	21.3%
a	891	21.3%
l	891	21.3%

Value	Count	Frequency (%)
f	314	7.5%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	4192	100.0%

Most frequent character per block

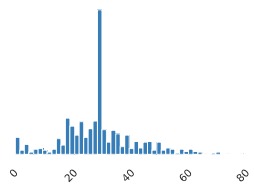
(unknown)

Value	Count	Frequency (%)
e	1205	28.7%
m	891	21.3%
a	891	21.3%
l	891	21.3%
f	314	7.5%

Age

Real number (ℝ)

Distinct	89	**Minimum**	0.42
Distinct (%)	10.0%	**Maximum**	80
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	29.699118	**Memory size**	7.1 KiB

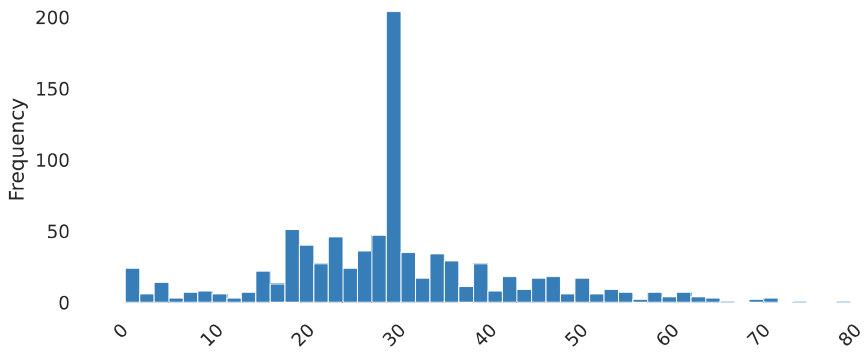


Quantile statistics

Minimum	0.42
5-th percentile	6
Q1	22
median	29.699118
Q3	35
95-th percentile	54
Maximum	80
Range	79.58
Interquartile range (IQR)	13

Descriptive statistics

Standard deviation	13.002015
Coefficient of variation (CV)	0.4377913
Kurtosis	0.9662793
Mean	29.699118
Median Absolute Deviation (MAD)	6.3008824
Skewness	0.43448809
Sum	26461.914
Variance	169.0524
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
29.69911765	177	19.9%
24	30	3.4%
22	27	3.0%
18	26	2.9%
28	25	2.8%
30	25	2.8%
19	25	2.8%
21	24	2.7%
25	23	2.6%
36	22	2.5%
Other values (79)	487	54.7%

Value	Count	Frequency (%)
0.42	1	0.1%
0.67	1	0.1%
0.75	2	0.2%
0.83	2	0.2%
0.92	1	0.1%
1	7	0.8%
2	10	1.1%
3	6	0.7%
4	10	1.1%
5	4	0.4%

Value	Count	Frequency (%)
80	1	0.1%
74	1	0.1%
71	2	0.2%

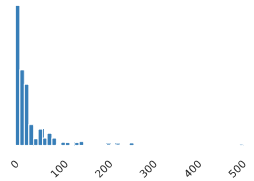
Value	Count	Frequency (%)
70.5	1	0.1%
70	2	0.2%
66	1	0.1%
65	3	0.3%
64	2	0.2%
63	2	0.2%
62	4	0.4%

Fare

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: famnp) ZEROS

Distinct	248	**Minimum**	0
Distinct (%)	27.8%	**Maximum**	512.3292
Missing	0	**Zeros**	15
Missing (%)	0.0%	**Zeros (%)**	1.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	32.204208	**Memory size**	7.1 KiB

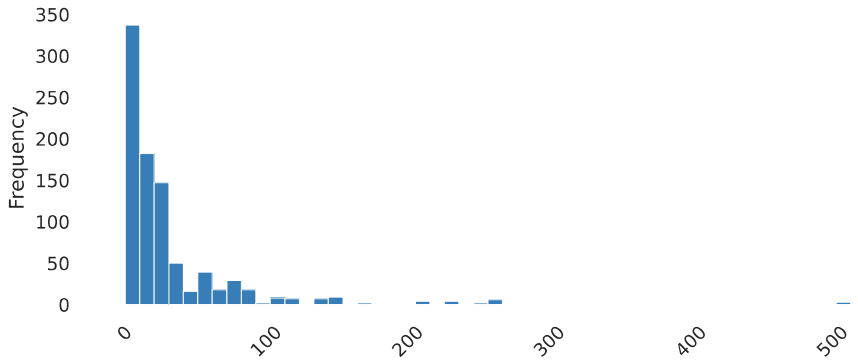


Quantile statistics

Minimum	0
5-th percentile	7.225
Q1	7.9104
median	14.4542
Q3	31
95-th percentile	112.07915
Maximum	512.3292
Range	512.3292
Interquartile range (IQR)	23.0896

Descriptive statistics

Standard deviation	49.693429
Coefficient of variation (CV)	1.5430725
Kurtosis	33.398141
Mean	32.204208
Median Absolute Deviation (MAD)	6.9042
Skewness	4.7873165
Sum	28693.949
Variance	2469.4368
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
8.05	43	4.8%
13	42	4.7%
7.8958	38	4.3%
7.75	34	3.8%
26	31	3.5%
10.5	24	2.7%
7.925	18	2.0%
7.775	16	1.8%
7.2292	15	1.7%
0	15	1.7%
Other values (238)	615	69.0%

Value	Count	Frequency (%)
0	15	1.7%
4.0125	1	0.1%
5	1	0.1%
6.2375	1	0.1%
6.4375	1	0.1%
6.45	1	0.1%
6.4958	2	0.2%
6.75	2	0.2%
6.8583	1	0.1%
6.95	1	0.1%

Value	Count	Frequency (%)
512.3292	3	0.3%
263	4	0.4%
262.375	2	0.2%

Value	Count	Frequency (%)
247.5208	2	0.2%
227.525	4	0.4%
221.7792	1	0.1%
211.5	1	0.1%
211.3375	3	0.3%
164.8667	2	0.2%
153.4625	3	0.3%

Embarked

Categorical

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	7.1 KiB

Length

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

Characters and Unicode

Total characters	891
Distinct characters	3
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%

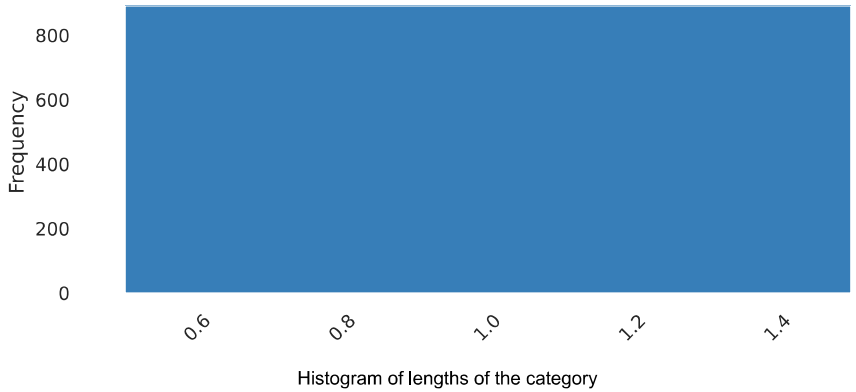
Sample

1st row	S
2nd row	C
3rd row	S
4th row	S
5th row	S ?

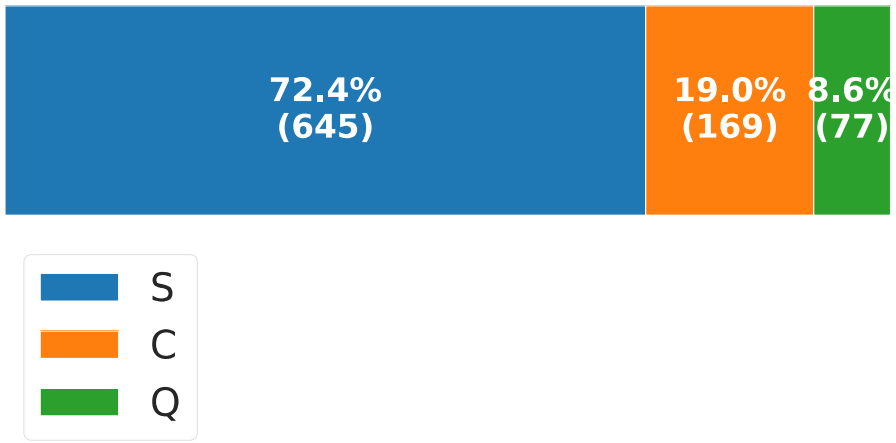
Common Values

Value	Count	Frequency (%)
S	645	72.4%
C	169	19.0%
Q	77	8.6%

Length



Common Values (Plot)



Value	Count	Frequency (%)
s	645	72.4%
c	169	19.0%
q	77	8.6%

Most occurring characters

Value	Count	Frequency (%)
S	645	72.4%
C	169	19.0%

Value	Count	Frequency (%)
Q	77	8.6%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
S	645	72.4%
C	169	19.0%
Q	77	8.6%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
S	645	72.4%
C	169	19.0%
Q	77	8.6%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	891	100.0%

Most frequent character per block

(unknown)

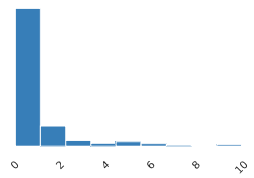
Value	Count	Frequency (%)
S	645	72.4%
C	169	19.0%
Q	77	8.6%

famnp

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Fare) ZEROS

Distinct	9	**Minimum**	0
Distinct (%)	1.0%	**Maximum**	10
Missing	0	**Zeros**	537
Missing (%)	0.0%	**Zeros (%)**	60.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.90460157	**Memory size**	7.1 KiB

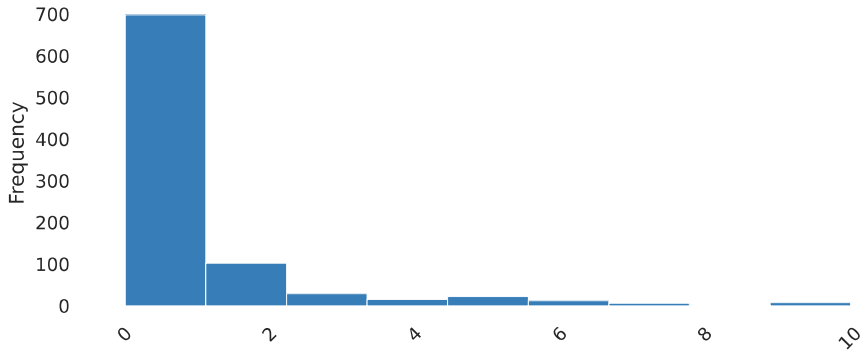


Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	1
95-th percentile	5
Maximum	10
Range	10
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	1.6134585
Coefficient of variation (CV)	1.7836124
Kurtosis	9.159666
Mean	0.90460157
Median Absolute Deviation (MAD)	0
Skewness	2.7274415
Sum	806
Variance	2.6032485
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=9)

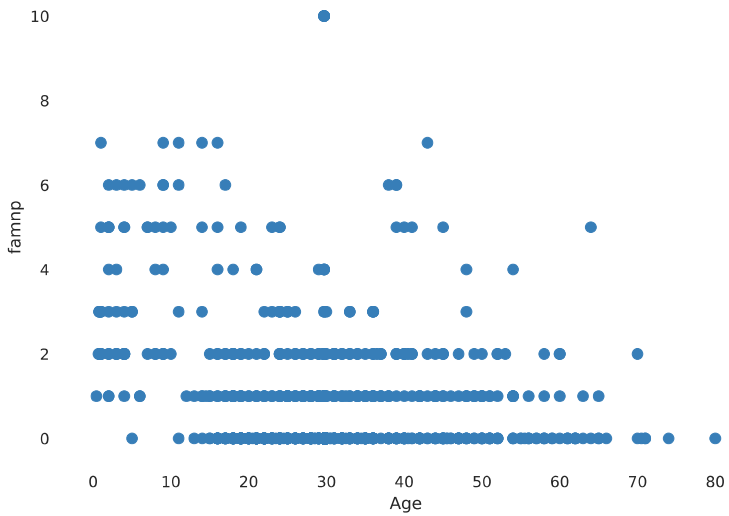
Value	Count	Frequency (%)
0	537	60.3%
1	161	18.1%
2	102	11.4%
3	29	3.3%
5	22	2.5%
4	15	1.7%
6	12	1.3%
10	7	0.8%
7	6	0.7%

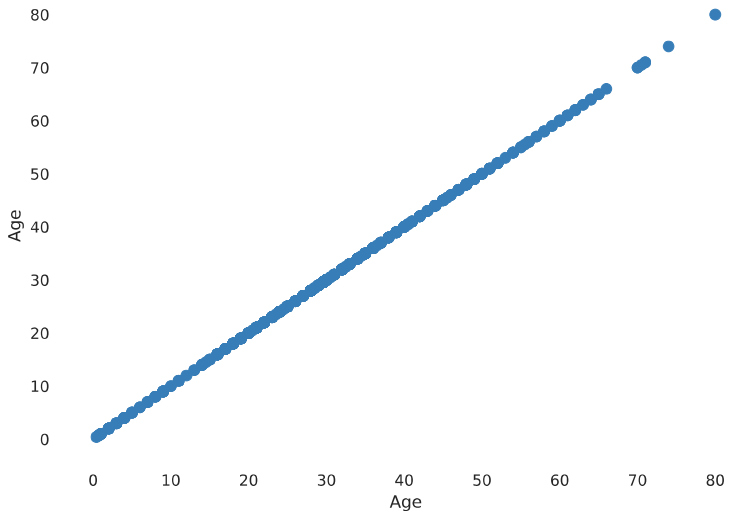
Value	Count	Frequency (%)
0	537	60.3%
1	161	18.1%
2	102	11.4%
3	29	3.3%
4	15	1.7%
5	22	2.5%
6	12	1.3%
7	6	0.7%
10	7	0.8%

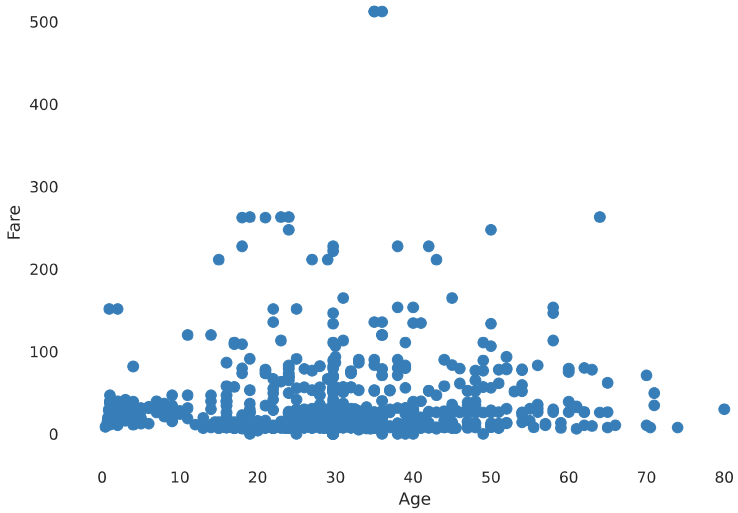
Value	Count	Frequency (%)
10	7	0.8%
7	6	0.7%
6	12	1.3%
5	22	2.5%
4	15	1.7%
3	29	3.3%

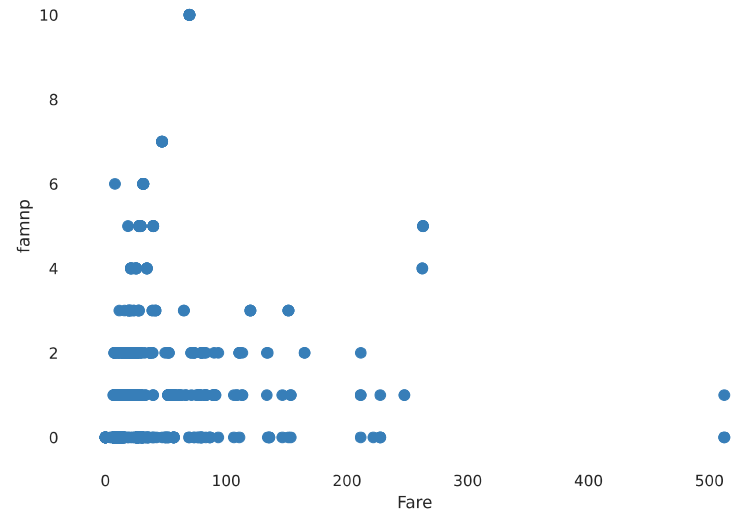
Value	Count	Frequency (%)
2	102	11.4%
1	161	18.1%
0	537	60.3%

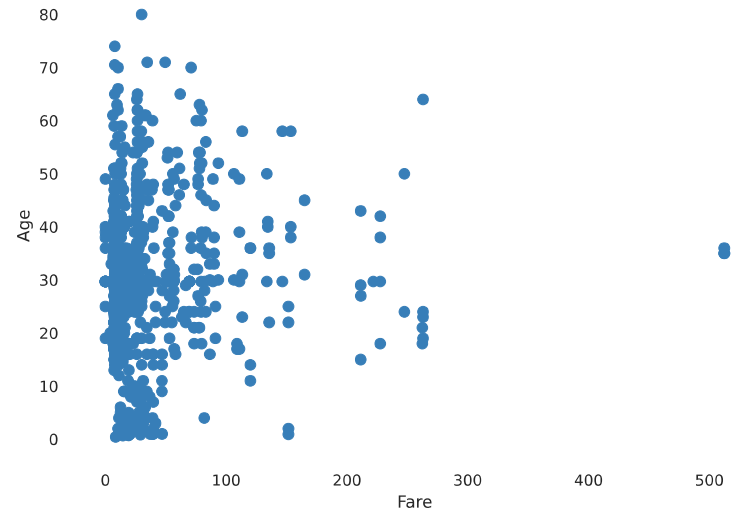
Interactions

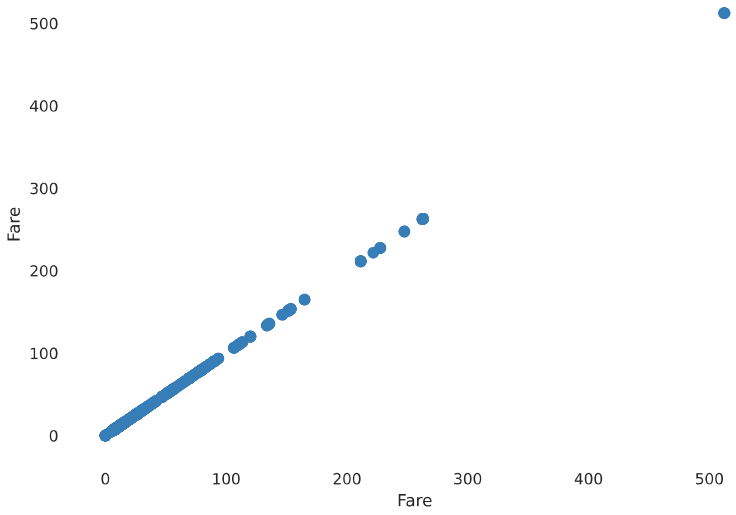


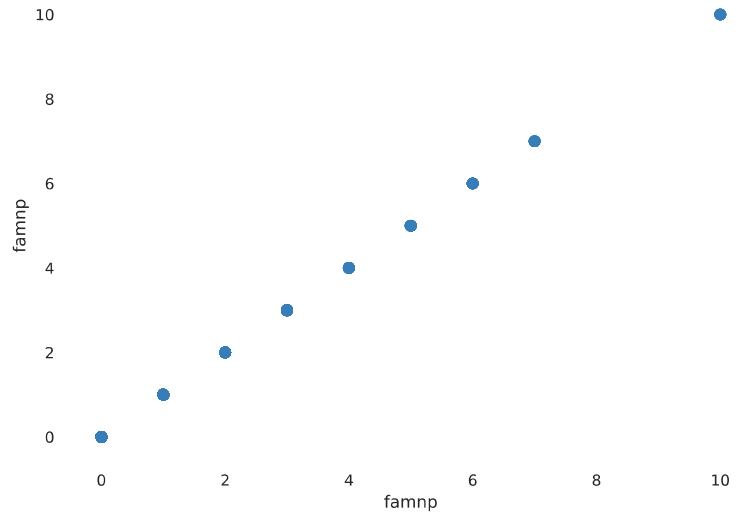


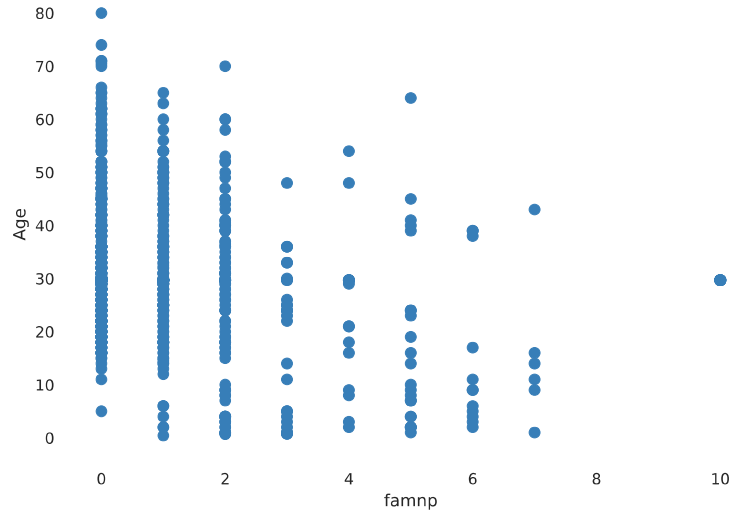

500
400
300
200
100
0
Fare
0
10
20
30
40
50
60
70
80
Age

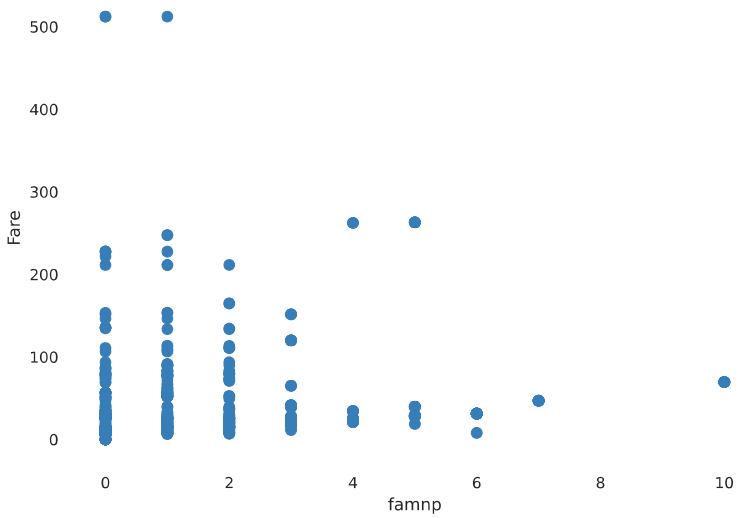

10
8
6
4
2
0
famnp
0
100
200
300
400
500
Fare


80
70
60
50
40
30
20
10
0
Age
0
100
200
300
400
500
Fare

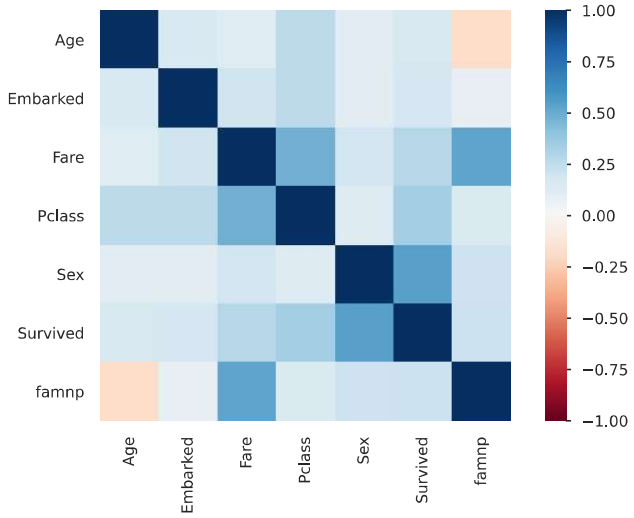

500
400
300
200
100
0
Fare
0
100
200
300
400
500
Fare


famnp
0
2
4
6
8
10
famnp


Age
0
10
20
30
40
50
60
70
80
0
2
4
6
8
10
famnp

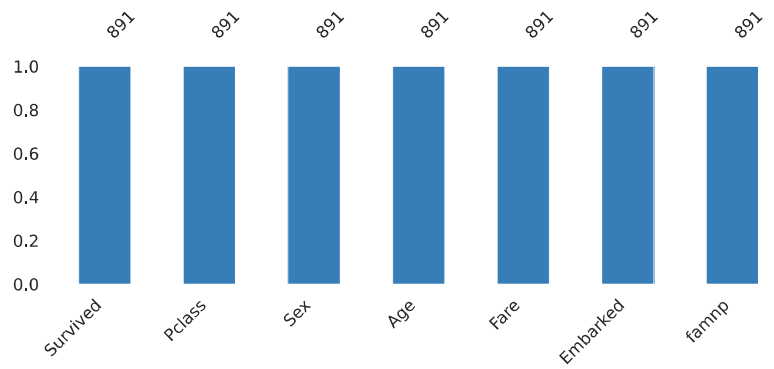


Correlations

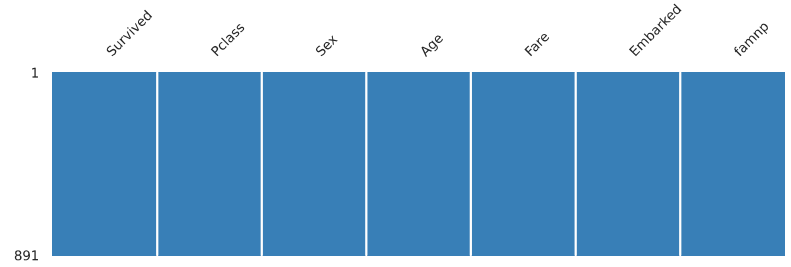


	Age	Embarked	Fare	Pclass	Sex	Survived	famnp
Age	1.000	0.151	0.119	0.265	0.106	0.158	-0.186
Embarked	0.151	1.000	0.197	0.261	0.114	0.168	0.083
Fare	0.119	0.197	1.000	0.479	0.189	0.283	0.529
Pclass	0.265	0.261	0.479	1.000	0.130	0.337	0.137
Sex	0.106	0.114	0.189	0.130	1.000	0.540	0.205
Survived	0.158	0.168	0.283	0.337	0.540	1.000	0.215
famnp	-0.186	0.083	0.529	0.137	0.205	0.215	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.

Sample

	Survived	Pclass	Sex	Age	Fare	Embarked	famnp
0	0	3	male	22.000000	7.2500	S	1
1	1	1	female	38.000000	71.2833	C	1
2	1	3	female	26.000000	7.9250	S	0
3	1	1	female	35.000000	53.1000	S	1
4	0	3	male	35.000000	8.0500	S	0
5	0	3	male	29.699118	8.4583	Q	0

	Survived	Pclass	Sex	Age	Fare	Embarked	famnp
6	0	1	male	54.000000	51.8625	S	0
7	0	3	male	2.000000	21.0750	S	4
8	1	3	female	27.000000	11.1333	S	2
9	1	2	female	14.000000	30.0708	C	1

	Survived	Pclass	Sex	Age	Fare	Embarked	famnp
881	0	3	male	33.000000	7.8958	S	0
882	0	3	female	22.000000	10.5167	S	0
883	0	2	male	28.000000	10.5000	S	0
884	0	3	male	25.000000	7.0500	S	0
885	0	3	female	39.000000	29.1250	Q	5
886	0	2	male	27.000000	13.0000	S	0
887	1	1	female	19.000000	30.0000	S	0
888	0	3	female	29.699118	23.4500	S	3
889	1	1	male	26.000000	30.0000	C	0
890	0	3	male	32.000000	7.7500	Q	0

Duplicate rows

Most frequently occurring

	Survived	Pclass	Sex	Age	Fare	Embarked	famnp	# duplicates
32	0	3	male	29.699118	7.8958	S	0	13
33	0	3	male	29.699118	8.0500	S	0	12
30	0	3	male	29.699118	7.7500	Q	0	9
49	1	3	female	29.699118	7.7500	Q	0	7
7	0	2	male	29.699118	0.0000	S	0	6

	Survived	Pclass	Sex	Age	Fare	Embarked	famnp	# duplicates
27	0	3	male	29.699118	7.2250	C	0	5
28	0	3	male	29.699118	7.2292	C	0	5
37	0	3	male	29.699118	69.5500	S	10	4
4	0	2	male	23.000000	13.0000	S	0	3
5	0	2	male	25.000000	13.0000	S	0	3