HashPrep v0.1.0a4

Overview

Variables

Correlations

Missing

Sample

2 Critical 17 Warnings

Overview

Dataset Statistics

Number of variables	12
Number of observations	891
Missing cells	866
Missing cells (%)	8.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	315.0 KiB
Average record size	362.1 B

Variable Types

Numeric	3
Categorical	6
Text	3

Alerts

2 Critical 17 Warnings

Missing

- 77.1% missing values in 'Cabin'

  Missing

- Missingness in 'Age' correlates with 6 columns (Pclass, Parch, Embarked)

  Missing

- Missingness in 'Cabin' correlates with 6 columns (Pclass, Fare, Survived)

Missing

High Cardinality

- Column 'Name' has 891 unique values (100.0% of rows)

  High Cardinality

- Column 'Ticket' has 681 unique values (76.4% of rows)

  High Cardinality

- Column 'Cabin' has 147 unique values (16.5% of rows)

  High Cardinality

Outliers

- Column 'SibSp' has 12 potential outliers (1.3% of non-missing values)

  Outliers

- Column 'Parch' has 10 potential outliers (1.1% of non-missing values)

  Outliers

- Column 'Fare' has 11 potential outliers (1.2% of non-missing values)

  Outliers

Zeros

- Column 'Survived' has 61.6% zero values

  Zeros

- Column 'SibSp' has 68.2% zero values

  Zeros

- Column 'Parch' has 76.1% zero values

  Zeros

Skewness

- Column 'SibSp' is highly skewed (skewness: 3.70)

  Skewness

- Column 'Fare' is highly skewed (skewness: 4.79)

  Skewness

Uniform

- 'PassengerId' is uniformly distributed and monotonic

  Uniform

Unique

- 'PassengerId' has unique values

  Unique

- 'Name' has unique values

  Unique

Constant Length

- 'Embarked' has constant length (1 chars for 100.0% of values)

  Constant Length

High Correlation

- Categorical columns 'Survived' and 'Sex' highly associated (Cramer's V: 0.540)

  High Correlation

Reproduction

Analysis started
    2026-02-06T16:31:10
Analysis finished

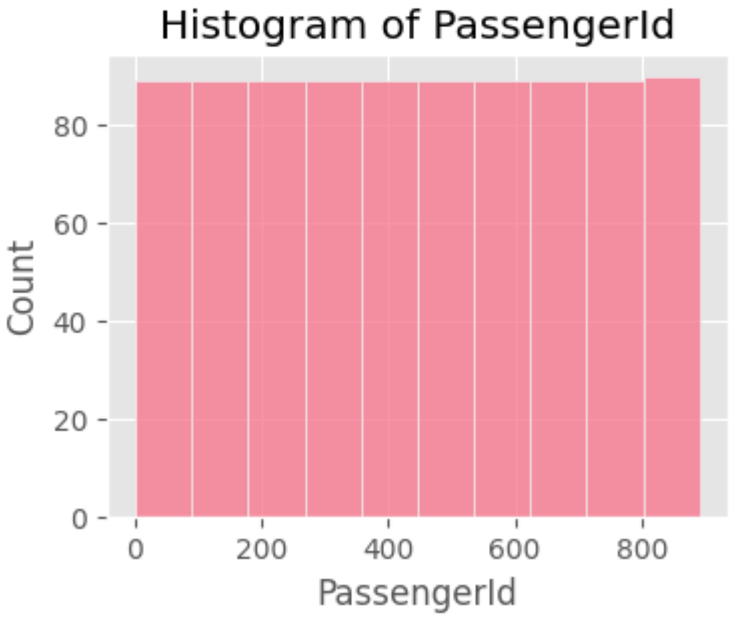
2026-02-06T16:31:17
Duration
6.81 seconds
Software version
hashprep v0.1.0a4

Variables

PassengerId

Numeric Unique

Distinct
891
100.0%
Missing
0
0.0%
Mean
446
avg value
Range
1 - 891
min - max



Quantile Statistics

Minimum	1
5th percentile	45.5
Q1 (25%)	223.5
Median (50%)	446
Q3 (75%)	668.5
95th percentile	846.5
Maximum	891
Range	890
IQR	445

Descriptive Statistics

Mean	446
Std deviation	257.354
Variance	66231
CV	0.577027
Skewness	0
Kurtosis	-1.2
MAD	223
Sum	397386
Monotonicity	Increasing

Common Values

Value	Count	%
891	1	0.1%
1	1	0.1%
2	1	0.1%
3	1	0.1%
4	1	0.1%
5	1	0.1%
6	1	0.1%
7	1	0.1%
8	1	0.1%
9	1	0.1%

Extreme Values

Minimum values
1 2 3 4 5 6 7 8 9 10
Maximum values
882 883 884 885 886 887 888 889 890 891

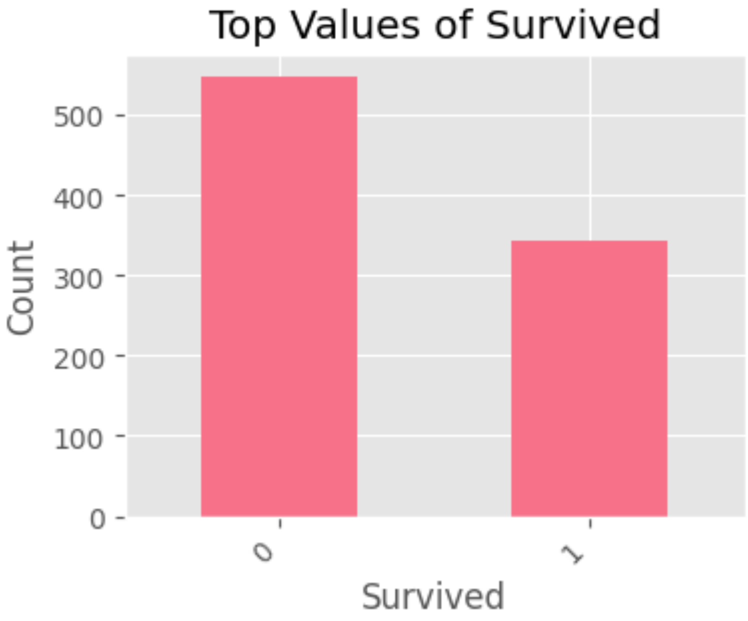
Value Counts

Zeros	0 (0.0%)
Negative	0 (0.0%)
Infinite	0 (0.0%)

Survived

Categorical

Distinct
2
0.2%
Missing
0
0.0%
Memory
7.1
KiB
Length
1 - 1
chars



Length Statistics

Min length	1
Max length	1
Mean length	1.00
Median length	1.00

Character Statistics

Total characters	891
Distinct characters	2
Distinct categories	1

Common Values

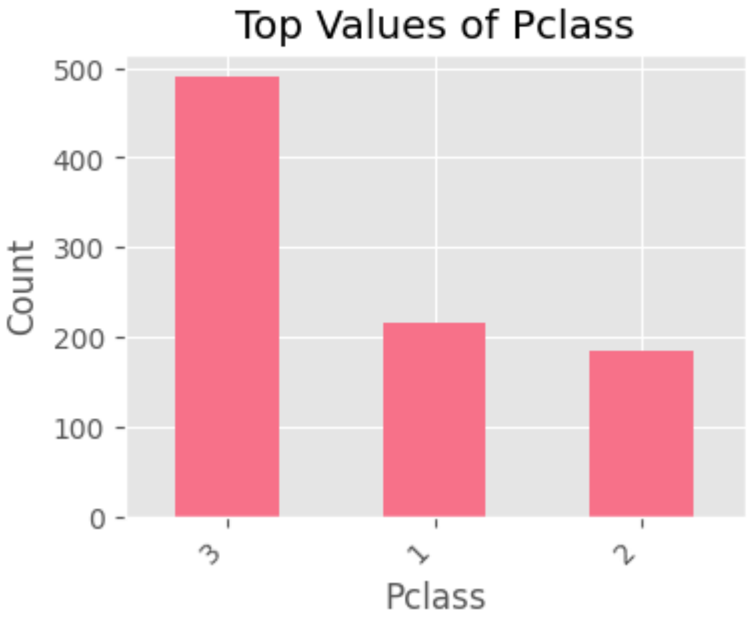
Value	Count	%
0	549	61.6%

Value	Count	%
1	342	38.4%

Pclass

Categorical

Distinct
3
0.3%
Missing
0
0.0%
Memory
7.1
KiB
Length
1 - 1
chars



Length Statistics

Min length	1
Max length	1
Mean length	1.00
Median length	1.00

Character Statistics

Total characters	891
Distinct characters	3
Distinct categories	1

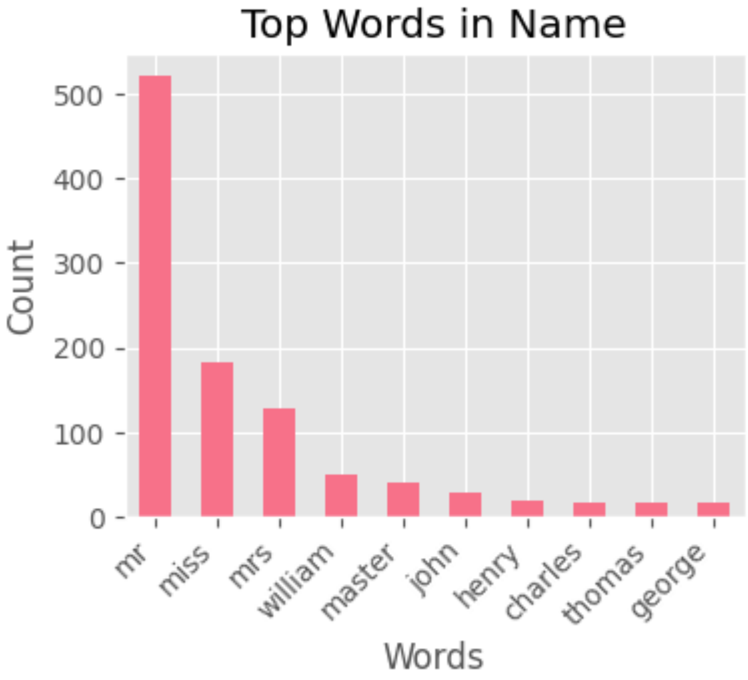
Common Values

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

Name

Text Unique

Distinct
891
100.0%
Missing
0
0.0%
Memory
73.2
KiB
Length
12 - 82
chars



Length Statistics

Min length	12
Max length	82
Mean length	26.97
Median length	25.00

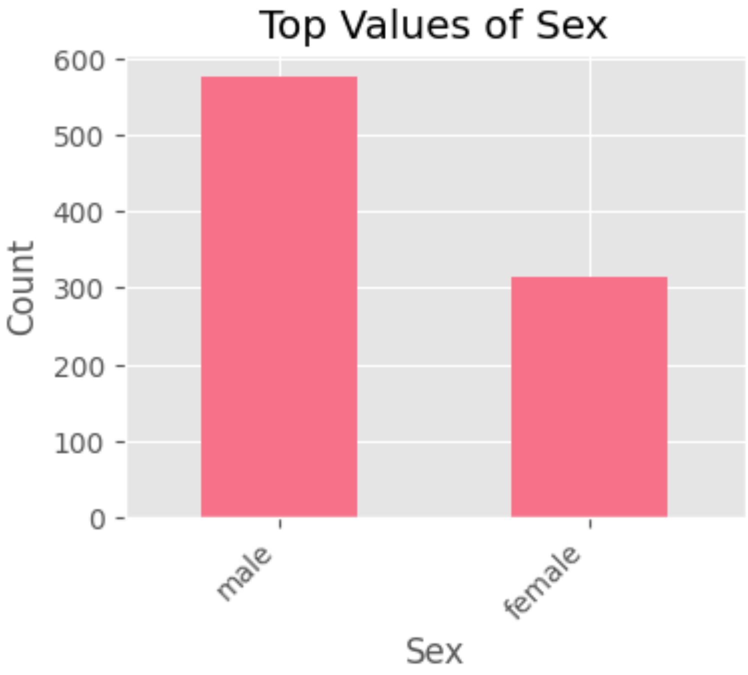
Character Statistics

Total characters	24026
Distinct characters	60
Distinct categories	7

Sex

Categorical

Distinct
2
0.2%
Missing
0
0.0%
Memory
53.8
KiB
Length
4 - 6
chars



Length Statistics

Min length	4
Max length	6
Mean length	4.70
Median length	4.00

Character Statistics

Total characters	4192
Distinct characters	5
Distinct categories	1

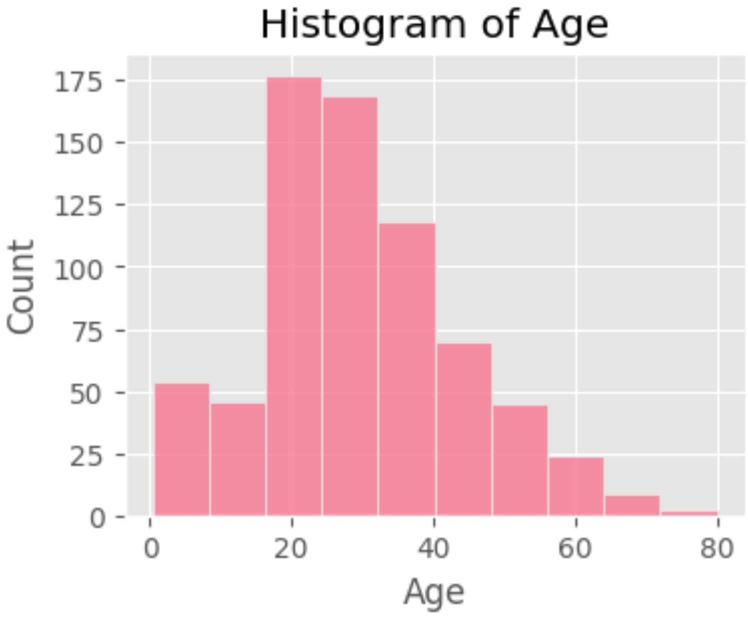
Common Values

Value	Count	%
male	577	64.8%
female	314	35.2%

Age

Numeric 19.9% missing

Distinct
88
12.3%
Missing
177
19.9%
Mean
29.7
avg value
Range
0.42 - 80
min - max



Quantile Statistics

Minimum	0.42

5th percentile	4
Q1 (25%)	20.125
Median (50%)	28
Q3 (75%)	38
95th percentile	56
Maximum	80
Range	79.58
IQR	17.875

Descriptive Statistics

Mean	29.6991
Std deviation	14.5265
Variance	211.019
CV	0.489122
Skewness	0.389108
Kurtosis	0.178274
MAD	9
Sum	21205.2
Monotonicity	None

Common Values

Value	Count	%
24.0	30	4.2%
22.0	27	3.8%
18.0	26	3.6%
28.0	25	3.5%
30.0	25	3.5%
19.0	25	3.5%
21.0	24	3.4%
25.0	23	3.2%
36.0	22	3.1%
29.0	20	2.8%

Extreme Values

Minimum values
0.42 0.67 0.75 0.75 0.83 0.83 0.92 1 1 1
Maximum values
65 65 66 70 70 70.5 71 71 74 80

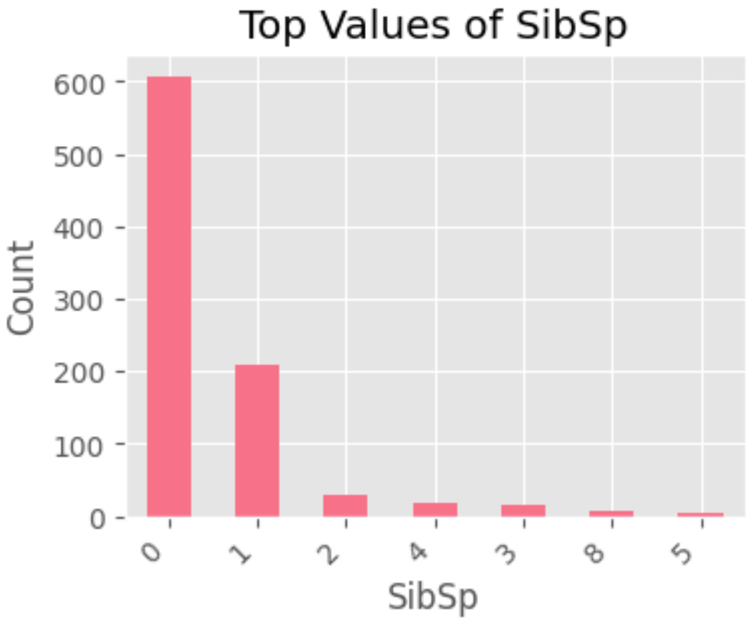
Value Counts

Zeros	0 (0.0%)
Negative	0 (0.0%)
Infinite	0 (0.0%)

SibSp

Categorical

Distinct
7
0.8%
Missing
0
0.0%
Memory
7.1
KiB
Length
1 - 1
chars



Length Statistics

Min length	1
Max length	1
Mean length	1.00
Median length	1.00

Character Statistics

Total characters	891
Distinct characters	7
Distinct categories	1

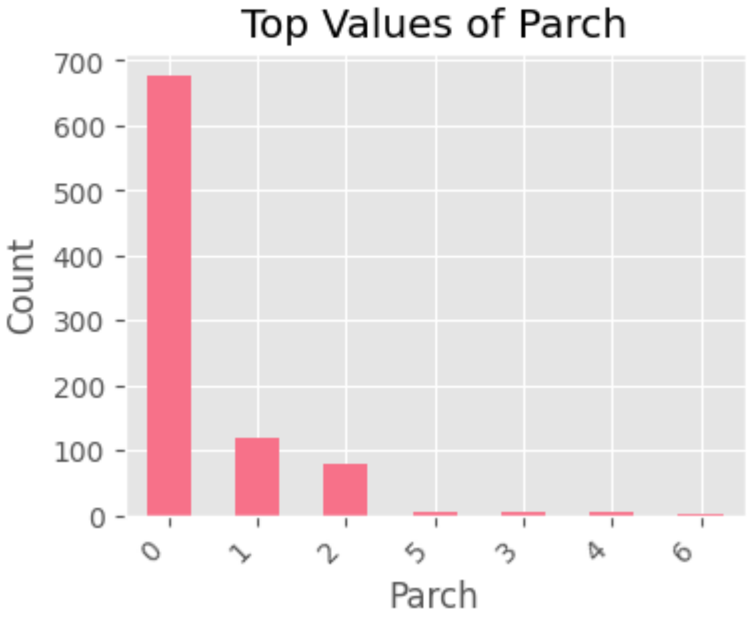
Common Values

Value	Count	%
0	608	68.2%
1	209	23.5%
2	28	3.1%
4	18	2.0%
3	16	1.8%
8	7	0.8%
5	5	0.6%

Parch

Categorical

Distinct
7
0.8%
Missing
0
0.0%
Memory
7.1
KiB
Length
1 - 1
chars



Length Statistics

Min length	1

Max length	1
Mean length	1.00
Median length	1.00

Character Statistics

Total characters	891
Distinct characters	7
Distinct categories	1

Common Values

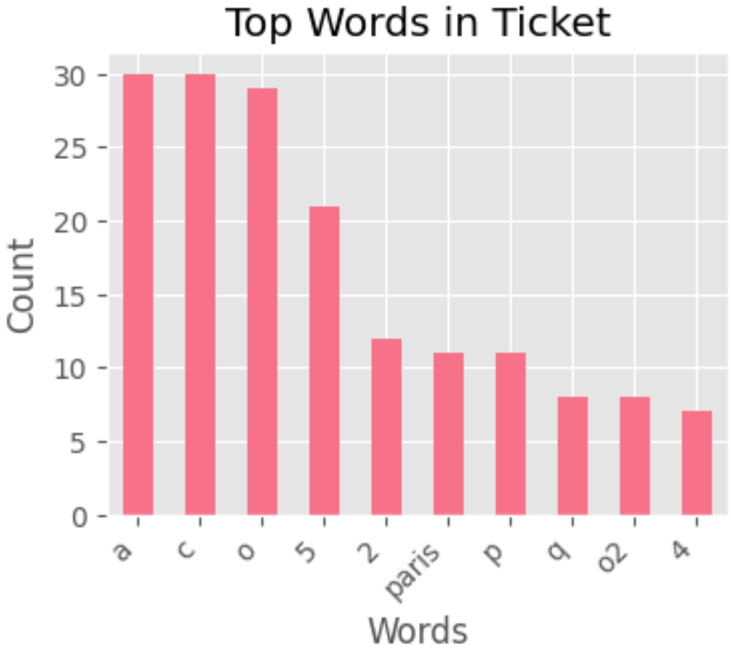
Value	Count	%
0	678	76.1%
1	118	13.2%
2	80	9.0%
5	5	0.6%
3	5	0.6%
4	4	0.4%
6	1	0.1%

Ticket

Text

Distict
681
76.4%
Missing
0
0.0%

Memory
55.6
KiB
Length
3 - 18
chars



Length Statistics

Min length	3
Max length	18
Mean length	6.75
Median length	6.00

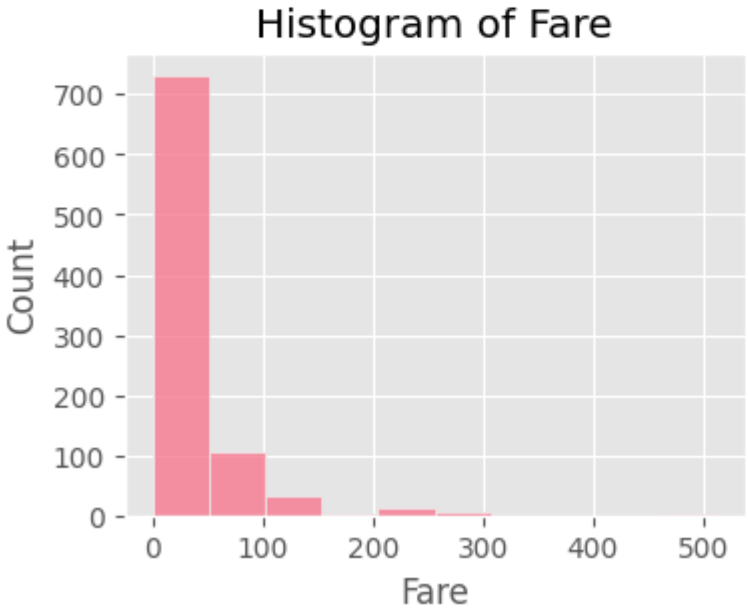
Character Statistics

Total characters	6015
Distinct characters	35
Distinct categories	5

Fare

Numeric

Distinct
248
27.8%
Missing
0
0.0%
Mean
32.2
avg value
Range
0 - 512.3
min - max



Quantile Statistics

Minimum	0

5th percentile	7.225
Q1 (25%)	7.9104
Median (50%)	14.4542
Q3 (75%)	31
95th percentile	112.079
Maximum	512.329
Range	512.329
IQR	23.0896

Descriptive Statistics

Mean	32.2042
Std deviation	49.6934
Variance	2469.44
CV	1.54307
Skewness	4.78732
Kurtosis	33.3981
MAD	6.9042
Sum	28693.9
Monotonicity	None

Common Values

Value	Count	%
8.05	43	4.8%
13.0	42	4.7%
7.8958	38	4.3%
7.75	34	3.8%
26.0	31	3.5%
10.5	24	2.7%
7.925	18	2.0%
7.775	16	1.8%
7.2292	15	1.7%
26.55	15	1.7%

Extreme Values

Minimum values
0 0 0 0 0 0 0 0 0 0
Maximum values
247.5 262.4 262.4 263 263 263 263 512.3 512.3 512.3

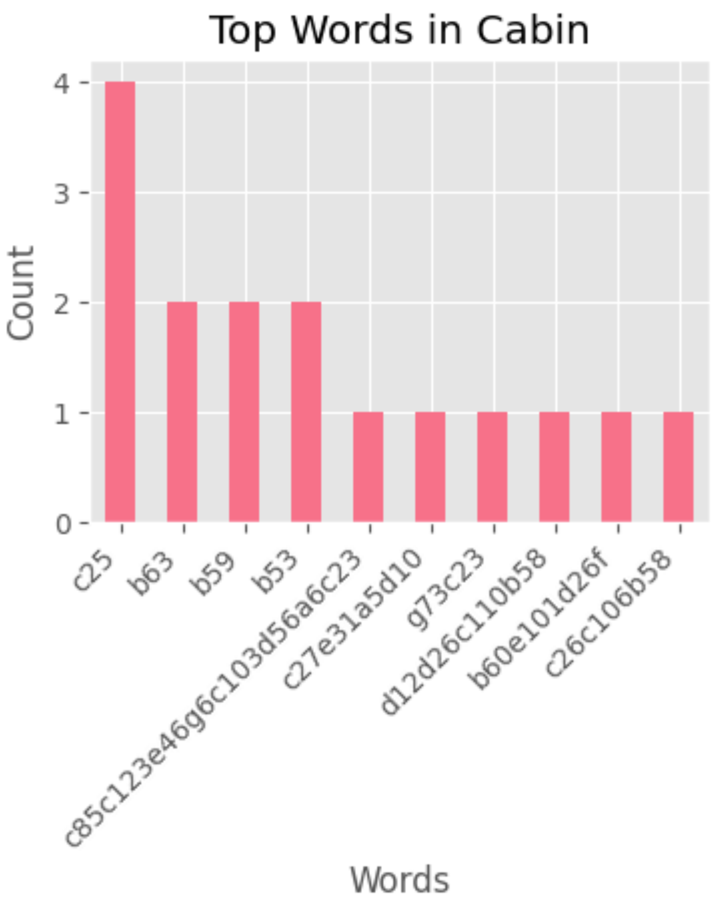
Value Counts

Zeros	15 (1.7%)
Negative	0 (0.0%)
Infinite	0 (0.0%)

Cabin

Text 77.1% missing

Distinct
147
72.1%
Missing
687
77.1%
Memory
33.7
KiB
Length
1 - 15
chars



Length Statistics

Min length	1
Max length	15
Mean length	3.59
Median length	3.00

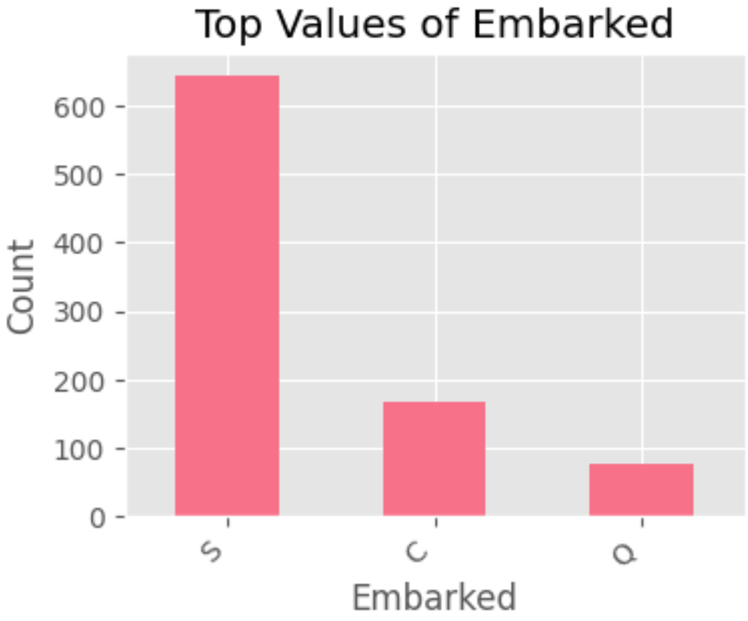
Character Statistics

Total characters	732
Distinct characters	19
Distinct categories	3

Embarked

Categorical 0.2% missing

Distinct
3
0.3%
Missing
2
0.2%
Memory
50.5
KiB
Length
1 - 1
chars



Length Statistics

Min length	1

Max length	1
Mean length	1.00
Median length	1.00

Character Statistics

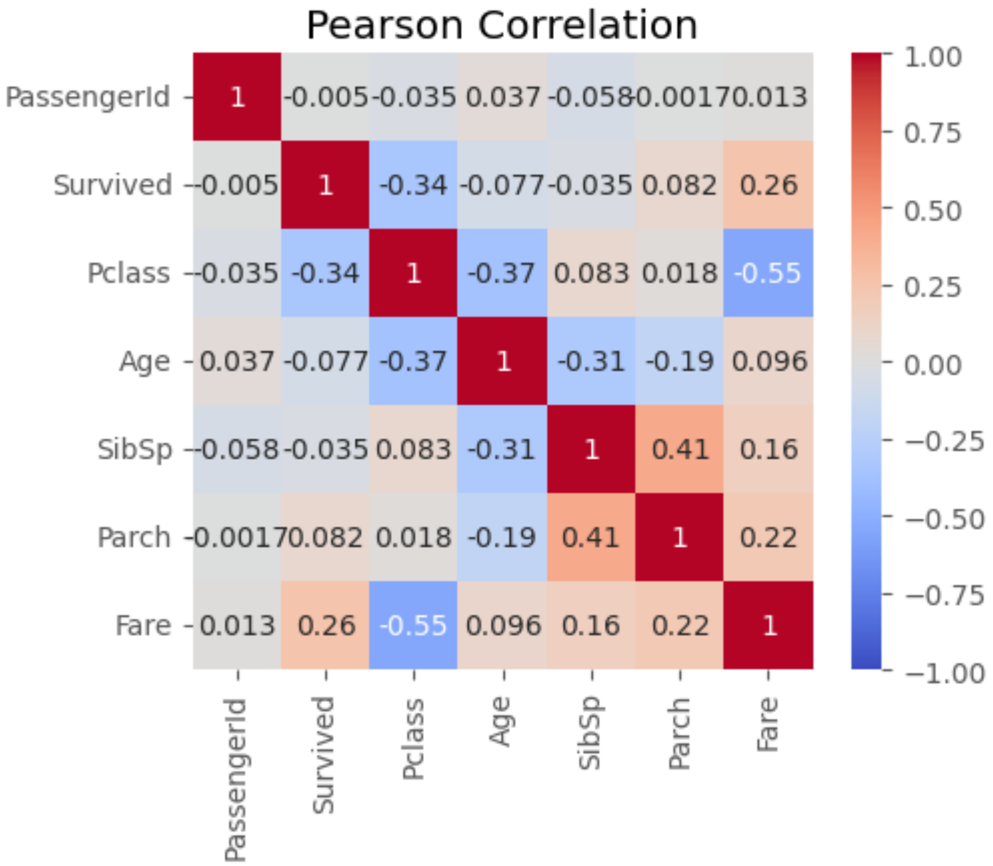
Total characters	889
Distinct characters	3
Distinct categories	1

Common Values

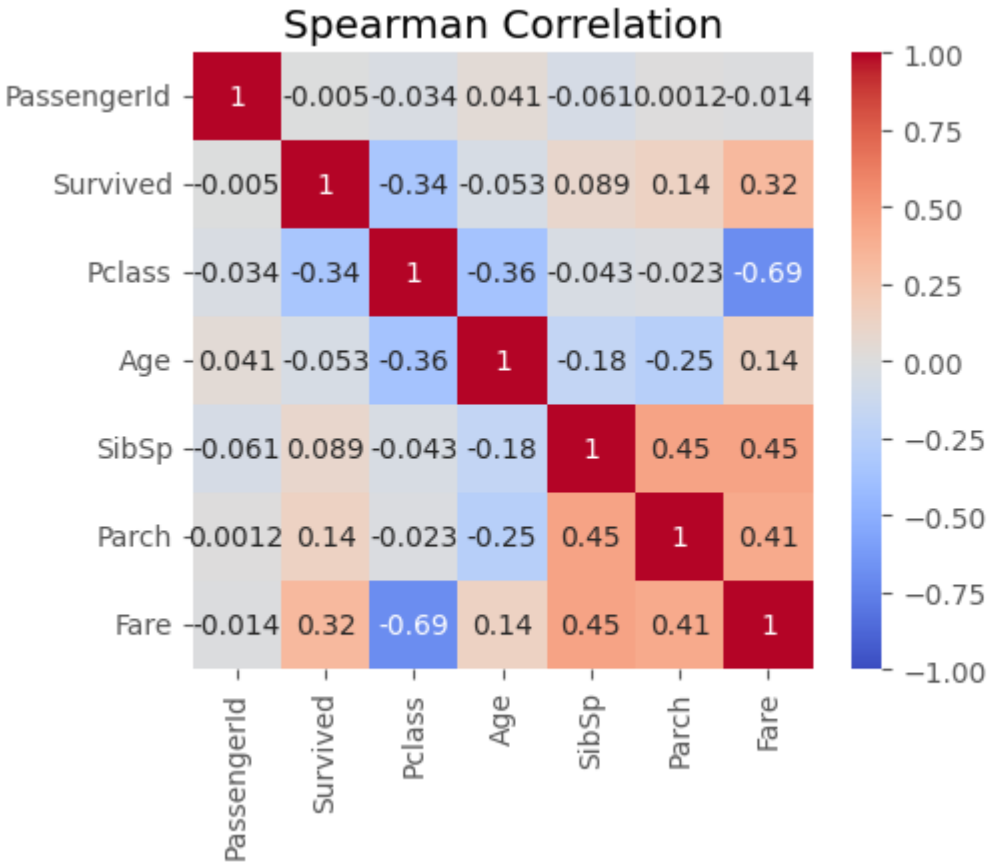
Value	Count	%
S	644	72.4%
C	168	18.9%
Q	77	8.7%

Correlations

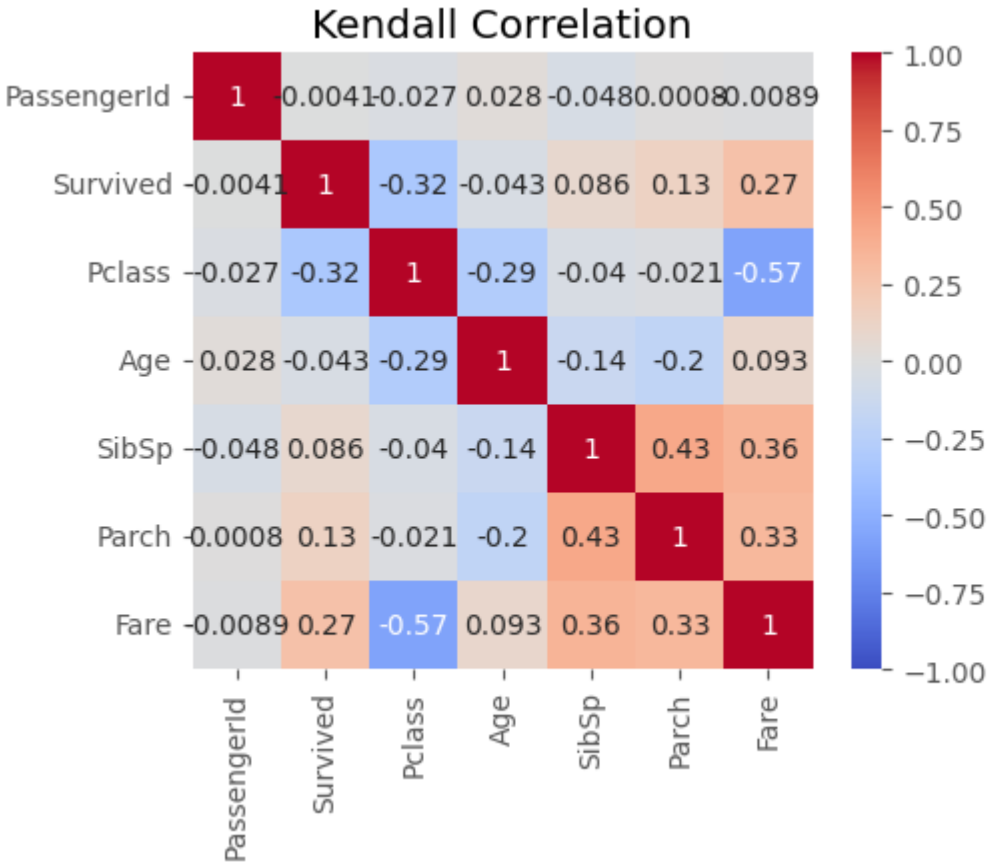
pearson Correlation



spearman Correlation

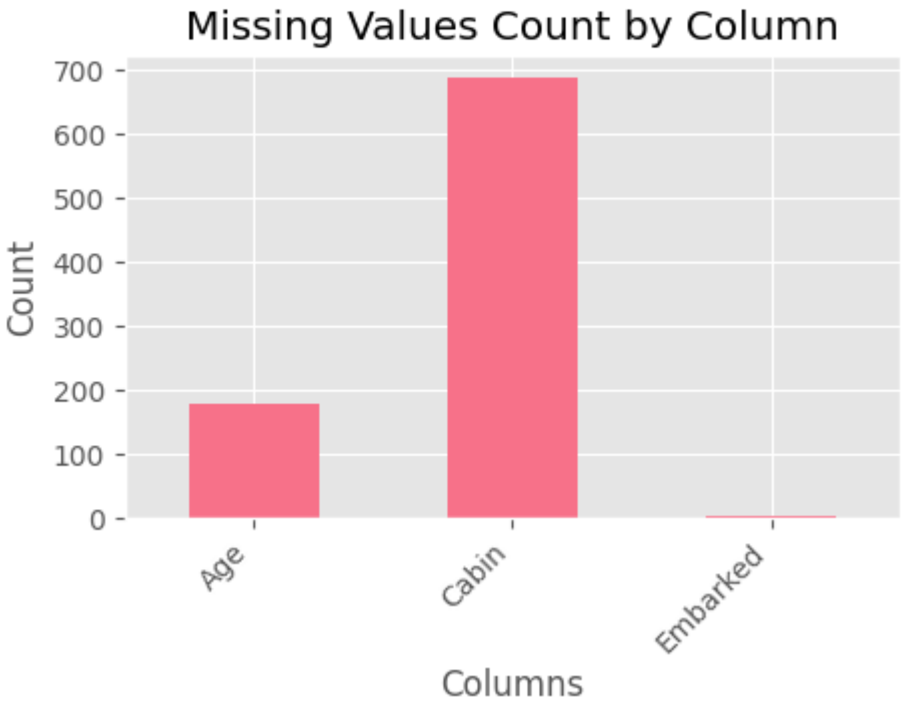


kendall Correlation

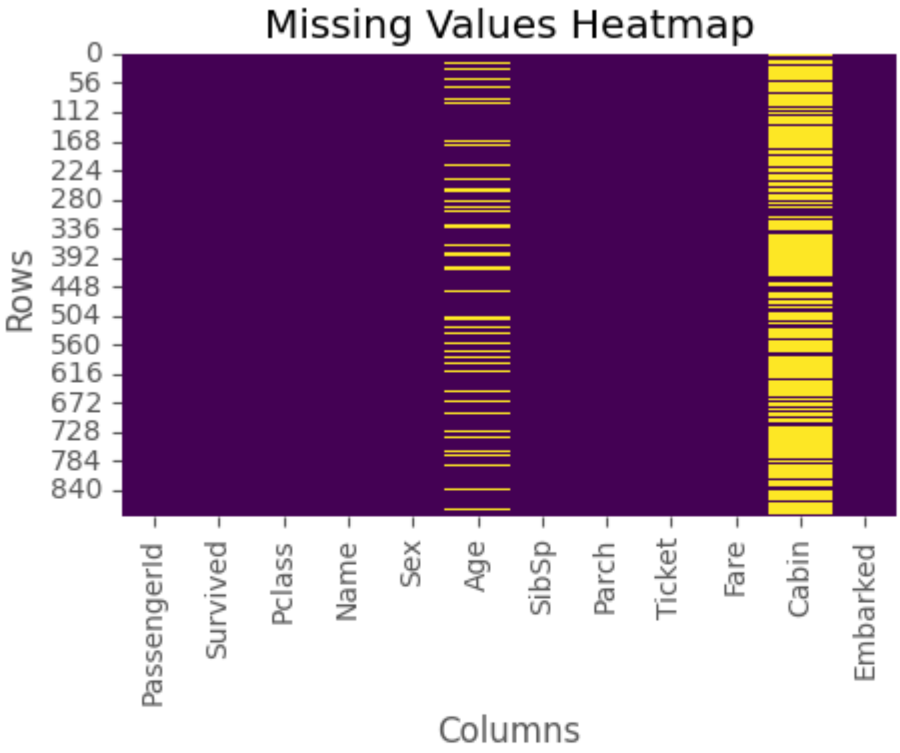


Missing Values

Missing Count per Column



Missing Pattern Matrix



Sample Data

PASSENGERID	SURVIVED	PCLASS	NAME	S
1	0	3	Braund, Mr. Owen Harris	m
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	fe
3	1	3	Heikkinen, Miss. Laina	fe
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	fe
5	0	3	Allen, Mr. William Henry	m

Built with HashPrep