

HashPrep Report

v0.1.0a4

2 Critical

17 Warnings

Overview

DATASET STATISTICS		VARIABLE TYPES	
Variables	12	Numeric	3
Observations	891	Categorical	6
Missing cells	866 (8.1%)	Text	3
Duplicate rows	0 (0.0%)		
Memory	315.0 KiB		
Avg record size	362.1 B		

Alerts

MISSING

- 77.1% missing values in 'Cabin'
- Missingness in 'Age' correlates with 6 columns (Pclass, Parch, Embarked)
- Missingness in 'Cabin' correlates with 6 columns (Pclass, Fare, Survived)

HIGH CARDINALITY

- Column 'Name' has 891 unique values (100.0% of rows)
- Column 'Ticket' has 681 unique values (76.4% of rows)
- Column 'Cabin' has 147 unique values (16.5% of rows)

OUTLIERS

- Column 'SibSp' has 12 potential outliers (1.3% of non-missing values)
- Column 'Parch' has 10 potential outliers (1.1% of non-missing values)
- Column 'Fare' has 11 potential outliers (1.2% of non-missing values)

ZEROS

- Column 'Survived' has 61.6% zero values
- Column 'SibSp' has 68.2% zero values
- Column 'Parch' has 76.1% zero values

SKEWNESS

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

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

UNIFORM

- 'PassengerId' is uniformly distributed and monotonic

UNIQUE

- 'PassengerId' has unique values
- 'Name' has unique values

CONSTANT LENGTH

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

HIGH CORRELATION

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

Reproduction

Analysis started	Analysis finished	Duration	Software version
2026-02-08T12:02:11	2026-02-08T12:02:15	4.35 seconds	hashprep v0.1.0a4

Variables

PassengerId

Numeric

Unique

DISTINCT

891

100.0%

MISSING

0

0.0%

MEAN

446

RANGE

1 – 891

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
12345678910
MAXIMUM
882883884885886887888889
890891

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%
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%

NameTextUnique

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

SexCategorical

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	RANGE 0.42 – 80
--------------------------------	--------------------------------	---------------------	---------------------------

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
0.420.670.750.750.830.830.92
111
MAXIMUM
656566707070.5717174
80

VALUE COUNTS

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

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

DISTINCT 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	RANGE 0 – 512.3
---------------------------------	-----------------------------	---------------------	---------------------------

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			EXTREME VALUES		
VALUE	COUNT	%	MINIMUM		
8.05	43	4.8%	0	0	0
13.0	42	4.7%	0	0	0
7.8958	38	4.3%	0	0	0
7.75	34	3.8%	0	0	0
26.0	31	3.5%	0	0	0
10.5	24	2.7%	0	0	0
7.925	18	2.0%	0	0	0
7.775	16	1.8%	0	0	0
7.2292	15	1.7%	0	0	0
26.55	15	1.7%	0	0	0
VALUE COUNTS			MAXIMUM		
Zeros	15	(1.7%)	247.5	262.4	262.4
Negative	0	(0.0%)	263	263	263
Infinite	0	(0.0%)	512.3	512.3	512.3

Cabin			
Text	77.1% missing		
DISTINCT	MISSING	MEMORY	LENGTH
147	687	33.7 KiB	1 – 15
72.1%	77.1%		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	MISSING	MEMORY	LENGTH
3	2	50.5 KiB	1 – 1
0.3%	0.2%		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

Numeric Correlations (Pearson - Top Pairs)

Feature 1	Feature 2	Correlation
Fare	Pclass	-0.549
Parch	SibSp	0.415
Age	Pclass	-0.369
Pclass	Survived	-0.338
Age	SibSp	-0.308
Fare	Survived	0.257
Fare	Parch	0.216
Age	Parch	-0.189
Fare	SibSp	0.160
Age	Fare	0.096
Pclass	SibSp	0.083
Parch	Survived	0.082
Age	Survived	-0.077
PassengerId	SibSp	-0.058
Age	PassengerId	0.037
SibSp	Survived	-0.035

Feature 1	Feature 2	Correlation
PassengerId	Pclass	-0.035
Parch	Pclass	0.018
Fare	PassengerId	0.013
PassengerId	Survived	-0.005

Categorical Correlations (Cramer's V - Top Pairs)

Pair	Value
Name__Sex	1.000
Name__Embarked	1.000
Name__Cabin	1.000
Name__Ticket	1.000
Ticket__Embarked	0.998
Ticket__Cabin	0.950
Cabin__Embarked	0.949
Sex__Ticket	0.860
Sex__Cabin	0.859
Sex__Embarked	0.123

Missing Values

Missing Values by Column

Column	Count	Percentage
Cabin	687	77.10%
Age	177	19.87%
Embarked	2	0.22%

Sample Data

Head (first 5 rows)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	None	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	None	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	None	S

Random Sample (10 rows)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
710	1	3	Moubarek, Master. Halim Gonios ("William George")	male	NaN	1	1	2661	15.2458	None	C

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
440	0	2	Kvillner, Mr. Johan Henrik Johannesson	male	31.0	0	0	C.A. 18723	10.5000	None	S
841	0	3	Alhomaki, Mr. Ilmari Rudolf	male	20.0	0	0	SOTON/O2 3101287	7.9250	None	S
721	1	2	Harper, Miss. Annie Jessie "Nina"	female	6.0	0	1	248727	33.0000	None	S
40	1	3	Nicola-Yarred, Miss. Jamila	female	14.0	1	0	2651	11.2417	None	C
291	1	1	Barber, Miss. Ellen "Nellie"	female	26.0	0	0	19877	78.8500	None	S
301	1	3	Kelly, Miss. Anna Katherine "Annie Kate"	female	NaN	0	0	9234	7.7500	None	Q
334	0	3	Vander Planke, Mr. Leo Edmondus	male	16.0	2	0	345764	18.0000	None	S
209	1	3	Carr, Miss. Helen "Ellen"	female	16.0	0	0	367231	7.7500	None	Q
137	1	1	Newsom, Miss. Helen Monypeny	female	19.0	0	2	11752	26.2833	D47	S

Tail (last 5 rows)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	None	S
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	None	S
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	None	Q