The data from the Titanic dataset contains the following columns:

1. **Survived:** Indicates whether the passenger survived (1) or not (0).

2. **Pclass:** Passenger class (1st, 2nd, or 3rd class).

3. **Name:** Name of the passenger.

4. **Sex:** Gender of the passenger (male/female).

5. **Age:** Age of the passenger.

6. **Siblings/Spouses Aboard:** Number of siblings or spouses aboard.

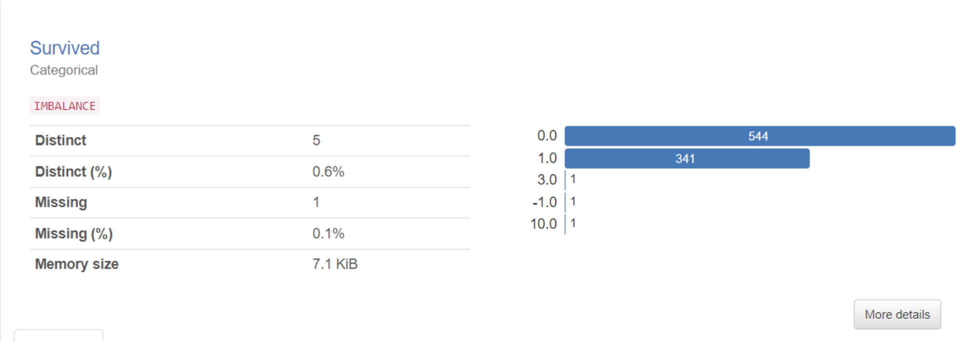
7. **Parents/Children Aboard:** Number of parents or children aboard.

8. **Fare:** Ticket fare.

To investigate the data quality and identify possible errors, I will perform an initial analysis on each column. This includes checking for invalid values, inconsistencies, and missing data. Let's start by examining each column individually.

Based on the initial analysis, here are the potential issues and validation criteria for each column:

1. Survived:



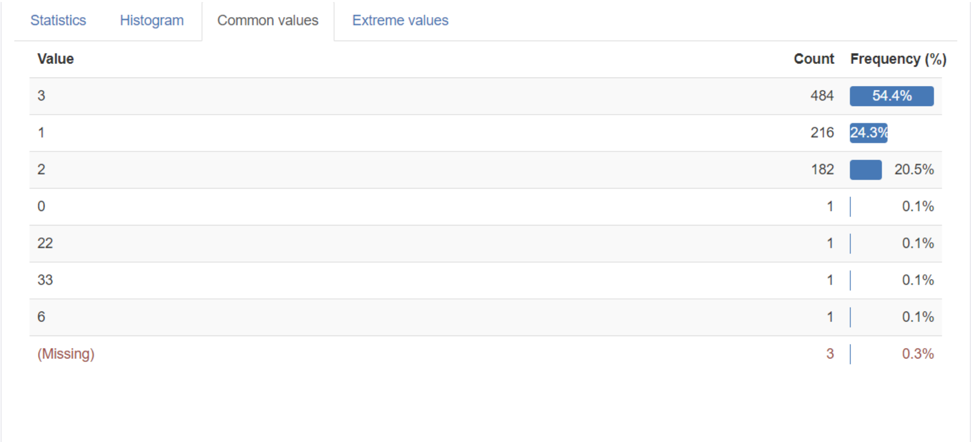
Survived
Categorical

IMBALANCE

Distinct	5
Distinct (%)	0.6%
Missing	1
Missing (%)	0.1%
Memory size	7.1 KiB

It appears that the data representing survival has been categorized as either 0 or 1. However, there seem to be 5 distinct values in this column which is not accurate as the value should be either 0 or 1. This could be due to a typo during manual data entry.

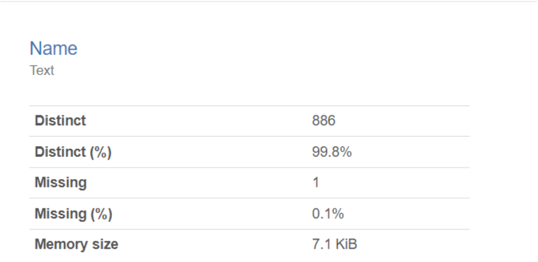
2. Pclass:



Value	Count	Frequency (%)
3	484	54.4%
1	216	24.3%
2	182	20.5%
0	1	0.1%
22	1	0.1%
33	1	0.1%
6	1	0.1%
(Missing)	3	0.3%

There may have been a manual entry error for the additional classes 0, 22, 33, and 6. Some class values are also missing. We can replace 22 and 33 with 2 and 3 Pclasses.

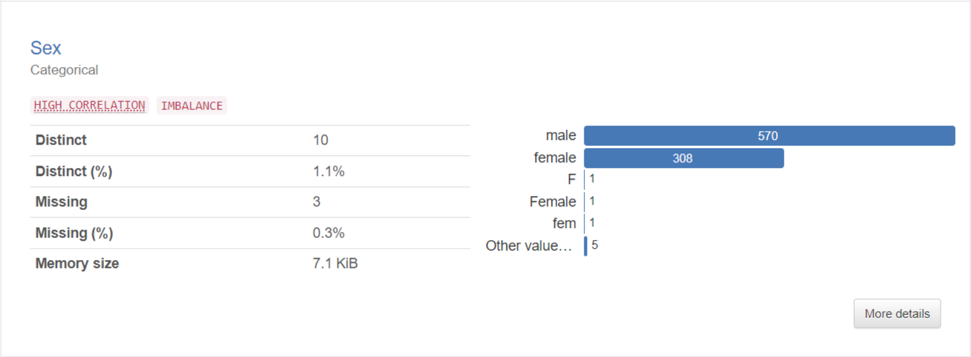
3. Name:


Name
Text

Distinct	886
Distinct (%)	99.8%
Missing	1
Missing (%)	0.1%
Memory size	7.1 KiB

Missing values in Name column.

4. Sex:

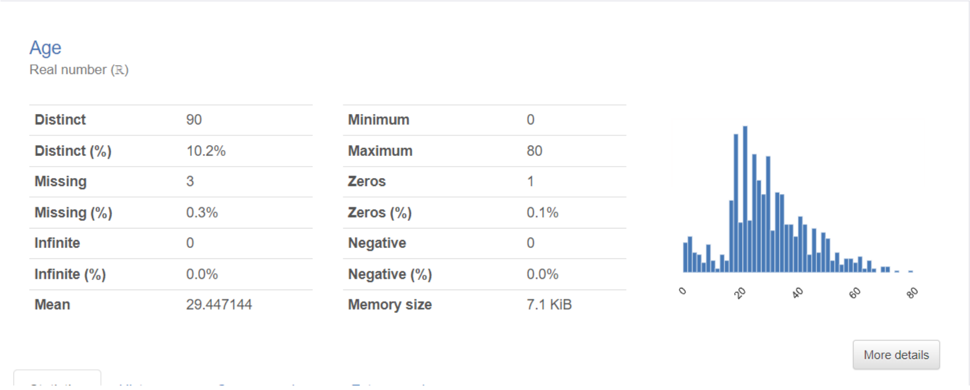

Sex
Categorical

HIGH CORRELATION IMBALANCE

Distinct	10
Distinct (%)	1.1%
Missing	3
Missing (%)	0.3%
Memory size	7.1 KiB

There are 10 distinct values in the sex column which are different versions of 'female' and 'male'. Additionally, there are about 3 missing values.

5. Age:



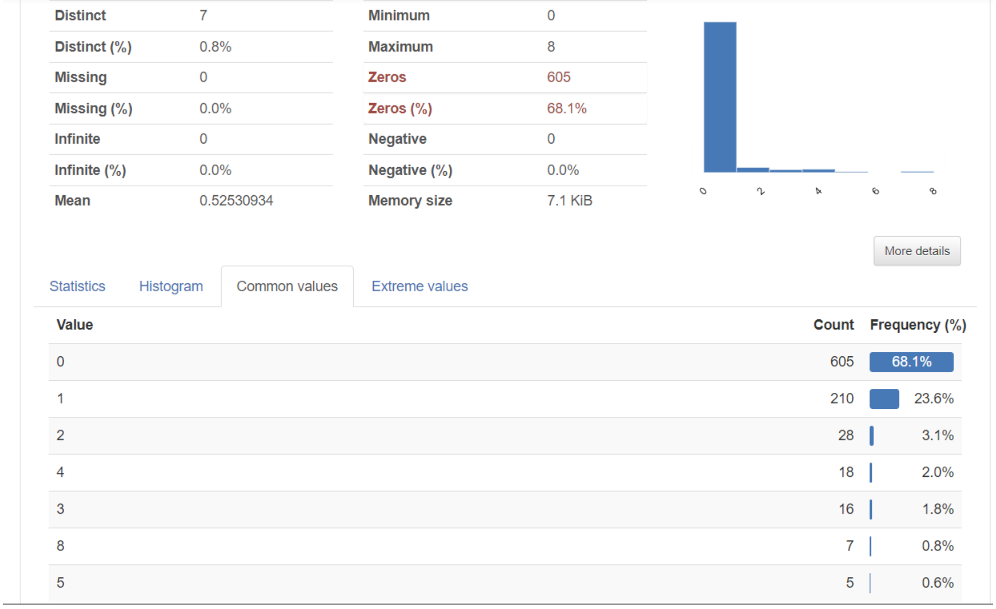
Age

Real number (ℝ)

Distinct	90	Minimum	0
Distinct (%)	10.2%	Maximum	80
Missing	3	Zeros	1
Missing (%)	0.3%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	29.447144	Memory size	7.1 KiB

The age column seems to be fine with just some missing values and normally distributed

6. Siblings/Spouses Aboard:

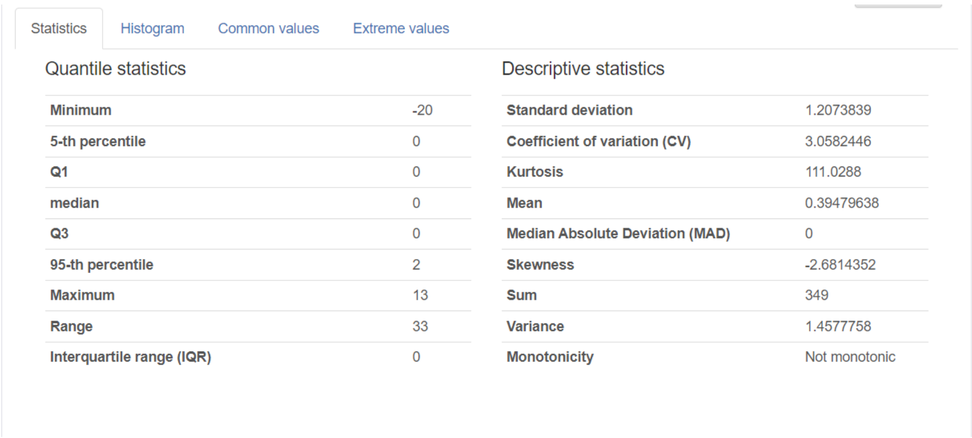


Distinct	7	Minimum	0
Distinct (%)	0.8%	Maximum	8
Missing	0	Zeros	605
Missing (%)	0.0%	Zeros (%)	68.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.52530934	Memory size	7.1 KiB

Value	Count	Frequency (%)
0	605	68.1%
1	210	23.6%
2	28	3.1%
4	18	2.0%
3	16	1.8%
8	7	0.8%
5	5	0.6%

The Siblings/Spouses column seems to be fine and there are no missing values.

7. Parents/Children Aboard:

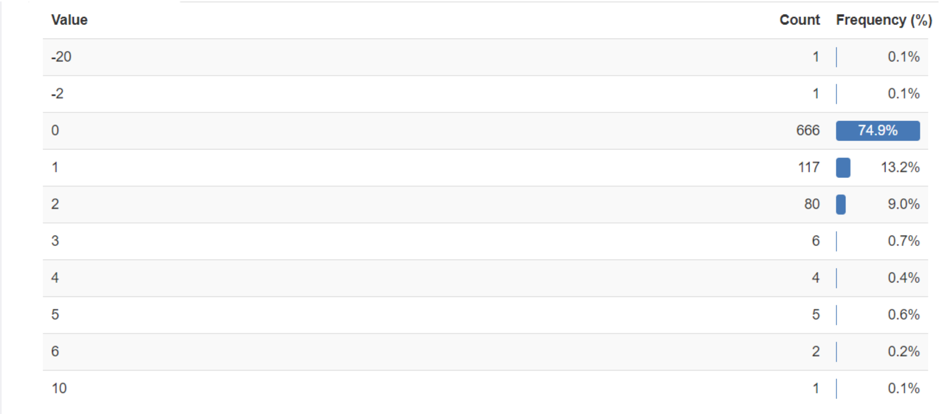


Quantile statistics

Minimum	-20
5-th percentile	0
Q1	0
median	0
Q3	0
95-th percentile	2
Maximum	13
Range	33
Interquartile range (IQR)	0

Descriptive statistics

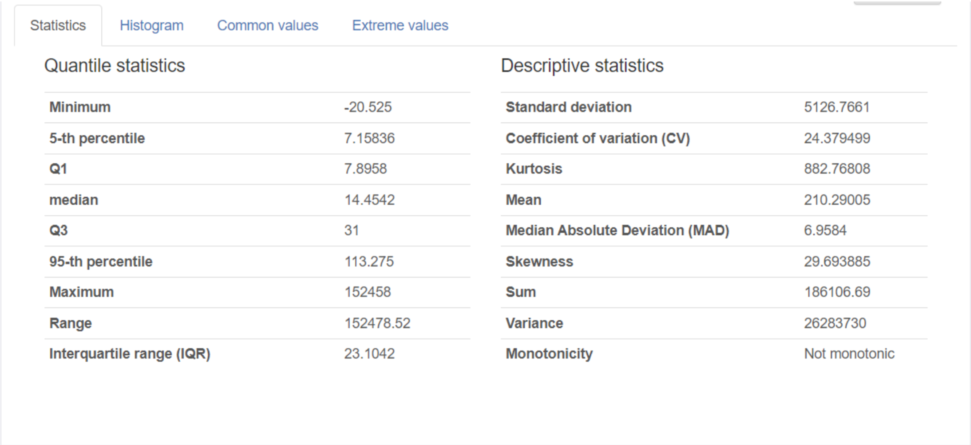
Standard deviation	1.2073839
Coefficient of variation (CV)	3.0582446
Kurtosis	111.0288
Mean	0.39479638
Median Absolute Deviation (MAD)	0
Skewness	-2.6814352
Sum	349
Variance	1.4577758
Monotonicity	Not monotonic



Value	Count	Frequency (%)
-20	1	0.1%
-2	1	0.1%
0	666	74.9%
1	117	13.2%
2	80	9.0%
3	6	0.7%
4	4	0.4%
5	5	0.6%
6	2	0.2%
10	1	0.1%

In the Parents/Children Aboard column, there are outliers such as -20 and -2. These negative values are certainly errors in the data.

8. Fare:



Quantile statistics

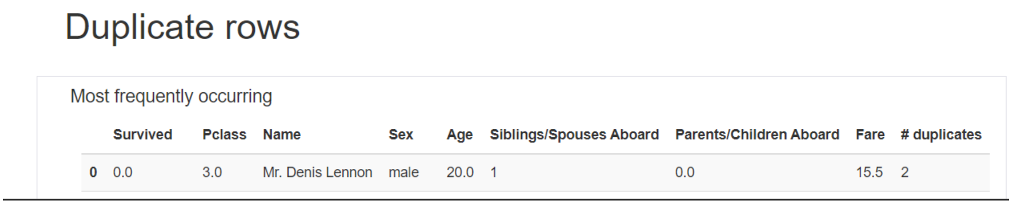
Minimum	-20.525
5-th percentile	7.15836
Q1	7.8958
median	14.4542
Q3	31
95-th percentile	113.275
Maximum	152458
Range	152478.52
Interquartile range (IQR)	23.1042

Descriptive statistics

Standard deviation	5126.7661
Coefficient of variation (CV)	24.379499
Kurtosis	882.76808
Mean	210.29005
Median Absolute Deviation (MAD)	6.9584
Skewness	29.693885
Sum	186106.69
Variance	26283730
Monotonicity	Not monotonic

The Fare column is highly skewed, and the data type needs to be changed to currency. The minimum value of -20 is incorrect as there should not be any negative values in this column. Additionally, the column has too many zeros, accounting for almost 15 or 1.7% of total values, which indicates either missing values or incorrect data. The maximum value of "152458" could potentially be an extreme value outlier as it is significantly different from other values and is influencing the skewed data distribution and the mean value of this column.

Duplicate rows



Duplicate rows

Most frequently occurring

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	# duplicates
0	0.0	3.0	Mr. Denis Lennon	male	20.0	1	0.0	15.5	2

and one duplicate entry has been noticed the data set which has occurred 2 times we can delete 1 entry to remove the duplicates entry .

To fix these issues, typically, I would:

- Replace or remove invalid or inconsistent entries.

- Impute or remove missing values based on the context.

- Standardize the entries in categorical columns like 'Sex'.

- Investigate and potentially correct outliers in columns like and 'Fare'.