Data Quality Analysis:

Column : Sex

I used both panda profile and openrefine to clean the data.

The first thing I noticed is the **sex** column, where the entries are marked as 'female' and 'male' seem to be correct, but there are typos that need to be fixed. Any such errors can be corrected to ensure consistency.

Additionally, I observed an age value of **40**, which appears to be a mistake. If this is a data entry error, it can be corrected by switching it to its intended value. Missing or blank entries in the dataset might also be filled in by leveraging the **name** column, as the titles (e.g., 'Mr.', 'Mrs.', 'Miss.') in the names can provide context for imputing the sex values or estimating other attributes like age.

This approach ensures that the sex column has accurate and complete values for all entries in the dataset.

Value	Count	Frequency (%)
male	570	64.1%
female	308	34.6%
F	1	0.1%
Female	1	0.1%
fem	1	0.1%
40	1	0.1%
Male	1	0.1%
emale	1	0.1%
M	1	0.1%
boy	1	0.1%
(Missing)	3	0.3%

Column : Age

Below I am showing how the openrefine show the values. Thus, as part of the process of fixing the sex column, I also resolved the shift below, ensuring that all data points are correctly aligned with their respective columns. However, the name column of this entity will stay blank.

All	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
346.	1	2	Miss.	40	0	0	13	

Column : Parents/Children Abroad

While reviewing the **parents/children aboard** column, I noticed 2 negative values.

Parents/Children Aboard

Real number (ℝ)

Zeros

Distinct	11	**Minimum**	-20
Distinct (%)	1.2%	**Maximum**	13
Missing	5	**Zeros**	666
Missing (%)	0.6%	**Zeros (%)**	74.9%
Infinite	0	**Negative**	2
Infinite (%)	0.0%	**Negative (%)**	0.2%
Mean	0.39479638	**Memory size**	7.1 KiB

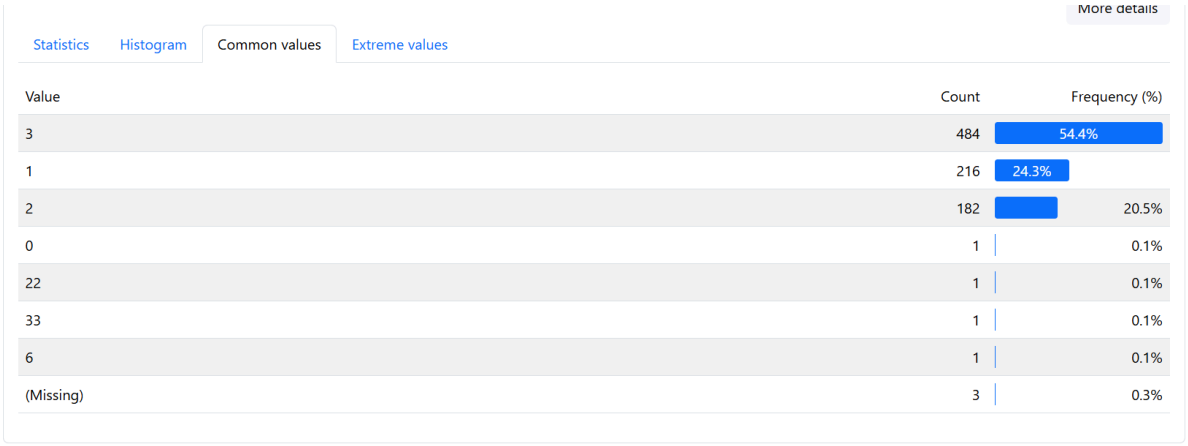
These are clearly errors. When analyzing the entities, it is impractical for a 21-year-old male to have 20 children or parents. To address this, I corrected these values to 2, assuming it represents a more reasonable count.

All	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
162.	0	3	Mr. John Hatfield Cribb	male	44	0		16.1
288.	1	2	Mr. Masabumi Hosono	male	42	0	-2	13
493.	0	3	Mr. Edward Roland Stanley	male	21	0	-20	8.05

This correction ensures the data is logical and aligns with practical expectations, improving its reliability for analysis.

Column : Pclass

While analyzing the **pclass** column, I observed that most entries correctly fall within the three Titanic classes: 1, 2, and 3. However, there are a few irregularities: three blank entries and unexpected values such as 0, 22, 33, and 6.

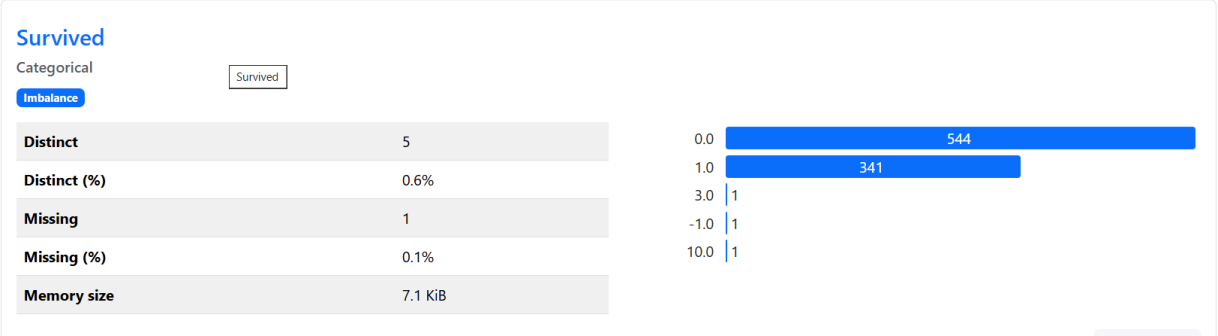


- The values 22 and 33 are likely errors, which can be reasonably assumed to represent 2 and 3, respectively.

The value 6 and 0 does not align with Titanic's three-class system, so it is more appropriate to mark it as blank for consistency and accuracy.

Column : Survived

The **survived** column should only contain 1 (indicating the passenger survived) or 0 (indicating the passenger did not survive). Upon inspection, I found some incorrect values:



- The value 10 can either be 0 or 1.
- The value 3 does not fit within the binary classification of survival and is difficult to interpret. Therefore, it is best to mark both as blank to maintain data accuracy.
- -1 can be fixed as 1.

Column : fare

When pay attention to fare column it is necessary to consider minimum and maximum 10 values.

Minimum 10 values | Maximum 10 values

Value	Count	Frequency (%)
152458	1	0.1%
5234.5542	1	0.1%
512.3292	3	0.3%
263	4	0.4%
262.375	2	0.2%
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%

However, for further analysis, I check the data and would like to suggest these corrections.
Minus fare should be a plus value.

239.	0	2	Mr. Frederick William Pengelly	male	19	0	0		
328.	1	3	Mrs. Frank John (Emily Alice Brown) Goldsmith	female	31	1	1	-20.525	
329.	1	1	Miss. Jean Gertrude Hippach	female	16	0	1		
370.	1	1	Mr. George Achilles Harder	male	25	1	0		

By looking at the '2' class values there should be a decimal place and below value should be 15. 2458

All	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
239.	0	2	Mr. Frederick William Pengelly	male	19	0	0	
256.	1	3	Mrs. Darwis (Hanne Youssef Razi) Touma	female	29	0	2	152458
329.	1	1	Miss. Jean Gertrude Hippach	female	16	0	1	
370.	1	1	Mr. George Achilles Harder	male	25	1	0	

Also below fares should be around 50s not 500s

All	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
239.	0	2	Mr. Frederick William Pengelly	male	19	0	0	
259.	1	1	Miss. Anna Ward	female	35	0	0	512.3292
329.	1	1	Miss. Jean Gertrude Hippach	female	16	0	1	
370.	1	1	Mr. George Achilles Harder	male	25	1	0	
678.	1	1	Mr. Thomas Drake Martinez Cardeza	male	36	0	1	512.3292
735.	1	1	Mr. Gustave J Lesurer	male	35	0	0	512.3292
869.	1	1	Mrs. Richard Leonard (Sallie Monypeny) Beckwith	female	47	1	1	5234.5542

Also based on the below statistics, 5% to 95% values are between 7 and 113.

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

Thus extreme ends values needed to analyze and correct if necessary. This can be done by comparing values for pclass and age. Minors may have low rates compare to adults. Need to analyze the data carefully to clean this column further.

Finally there are duplicate entities according to panda profile and we can remove

duplicate entities, since they are not necessary.

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
1	1.0	2.0	Miss. Emily Rugg	female	21.0	0	0.0	10.5	2