Dataset Quality Report

Generated on 2026-01-18 12:57:32 • HashPrep v0.1.0a3

Critical 2
Warnings 13

Issues Overview

Category	Column	Description	Fix Suggestion
Missing values	Cabin	77.1% missing values in 'Cabin'	Options: - Drop column: Reduces bias from missing data (Pros: Simplifies model; Cons: Loses potential info). - Impute values: Use domain-informed methods (e.g., median, mode, or predictive model) (Pros: Retains feature; Cons: May introduce bias). - Create missingness indicator: Flag missing values as a new feature (Pros: Captures missingness pattern; Cons: Adds complexity).
High cardinality	Name	Column 'Name' has 891 unique values (100.0% of rows)	Options: - Drop column: Avoids overfitting from unique identifiers (Pros: Simplifies model; Cons: Loses potential info). - Engineer feature: Extract patterns (e.g., titles from names) (Pros: Retains useful info; Cons: Requires domain knowledge). - Use hashing: Reduce dimensionality (Pros: Scalable; Cons: May lose interpretability).
High cardinality	Ticket	Column 'Ticket' has 681 unique values (76.4% of rows)	Options: - Group rare categories: Reduce cardinality (Pros: Simplifies feature; Cons: May lose nuance). - Use feature hashing: Map to lower dimensions (Pros: Scalable; Cons: Less interpretable). - Retain and test: Evaluate

Category	Column	Description	Fix Suggestion
			feature importance (Pros: Data-driven; Cons: Risk of overfitting).
High cardinality	Cabin	Column 'Cabin' has 147 unique values (16.5% of rows)	Options: - Group rare categories: Reduce cardinality (Pros: Simplifies feature; Cons: May lose nuance). - Use feature hashing: Map to lower dimensions (Pros: Scalable; Cons: Less interpretable). - Retain and test: Evaluate feature importance (Pros: Data-driven; Cons: Risk of overfitting).
Outliers	SibSp	Column 'SibSp' has 12 potential outliers (1.3% of non-missing values)	Options: - Investigate outliers: Verify if valid or errors (Pros: Ensures accuracy; Cons: Time-consuming). - Transform: Use log/sqrt to reduce impact (Pros: Retains data; Cons: Changes interpretation). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
Outliers	Parch	Column 'Parch' has 10 potential outliers (1.1% of non-missing values)	Options: - Investigate outliers: Verify if valid or errors (Pros: Ensures accuracy; Cons: Time-consuming). - Transform: Use log/sqrt to reduce impact (Pros: Retains data; Cons: Changes interpretation). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
Outliers	Fare	Column 'Fare' has 11 potential outliers (1.2% of non-missing values)	Options: - Investigate outliers: Verify if valid or errors (Pros: Ensures accuracy; Cons: Time-consuming). - Transform: Use log/sqrt to reduce impact (Pros: Retains data; Cons: Changes interpretation).

Category	Column	Description	Fix Suggestion
			- Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
High zero counts	Survived	Column 'Survived' has 61.6% zero values	Options: - Transform: Create binary indicator for zeros (Pros: Captures pattern; Cons: Adds complexity). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: May skew results). - Investigate zeros: Verify validity (Pros: Ensures accuracy; Cons: Time-consuming).
High zero counts	SibSp	Column 'SibSp' has 68.2% zero values	Options: - Transform: Create binary indicator for zeros (Pros: Captures pattern; Cons: Adds complexity). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: May skew results). - Investigate zeros: Verify validity (Pros: Ensures accuracy; Cons: Time-consuming).
High zero counts	Parch	Column 'Parch' has 76.1% zero values	Options: - Transform: Create binary indicator for zeros (Pros: Captures pattern; Cons: Adds complexity). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: May skew results). - Investigate zeros: Verify validity (Pros: Ensures accuracy; Cons: Time-consuming).
Missing patterns	Age	Missingness in 'Age' correlates with 6 columns (Pclass, Parch, Embarked)	Options: - Impute values: Use simple or domain-informed methods (Pros: Retains feature; Cons: Risk of bias). - Drop column: If less critical (Pros: Simplifies model; Cons: Loses info). - Test impact: Evaluate model with/without feature (Pros: Data-

Category	Column	Description	Fix Suggestion
			driven; Cons: Requires computation).
Missing patterns	Cabin	Missingness in 'Cabin' correlates with 6 columns (Pclass, Fare, Survived)	Options: - Impute values: Use simple or domain-informed methods (Pros: Retains feature; Cons: Risk of bias). - Drop column: If less critical (Pros: Simplifies model; Cons: Loses info). - Test impact: Evaluate model with/without feature (Pros: Data-driven; Cons: Requires computation).
Skewness	SibSp	Column 'SibSp' is highly skewed (skewness: 3.70)	Options: - Square root transform: Reduces moderate skew. - Monitor: Evaluate model performance on skewed data.
Skewness	Fare	Column 'Fare' is highly skewed (skewness: 4.79)	Options: - Square root transform: Reduces moderate skew. - Monitor: Evaluate model performance on skewed data.
Feature correlation	Survived,Sex	Categorical columns 'Survived' and 'Sex' highly associated (Cramer's V: 0.540)	Options: - Monitor redundancy. - Re-encode.

Variable Analysis

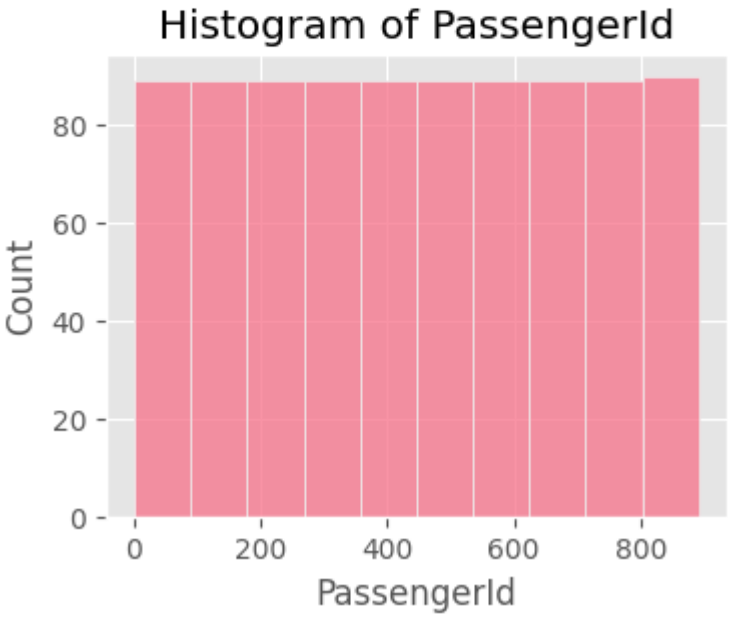
PassengerId

Numeric
Missing: 0 (0.0%)
Distinct: 891

Statistics

```
descriptive:
  coefficient_of_variation: 0.5770265516036549
  kurtosis: -1.1999999999999997
  mad: 223.0
  mean: 446.0
  monotonicity: increasing
  skewness: 0.0
  standard_deviation: 257.3538420152301
  sum: 397386.0
  variance: 66231.0
quantiles:
  iqr: 445.0
  maximum: 891.0
  median: 446.0
  minimum: 1.0
  p5: 45.5
  p95: 846.5
  q1: 223.5
  q3: 668.5
  range: 890.0
```

histogram



Survived

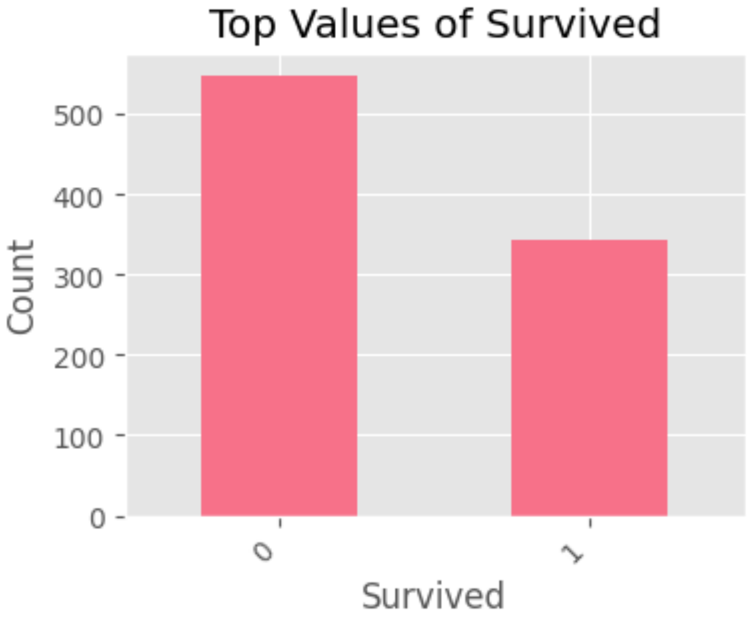
Categorical
Missing: 0 (0.0%)
Distinct: 2

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 2
  distinct_scripts: null
  total_characters: 891
```

```
length:
  max_length: 1
  mean_length: 1.0
  median_length: 1.0
  min_length: 1
sample:
- '0'
- '1'
- '1'
- '1'
- '0'
```

common values bar



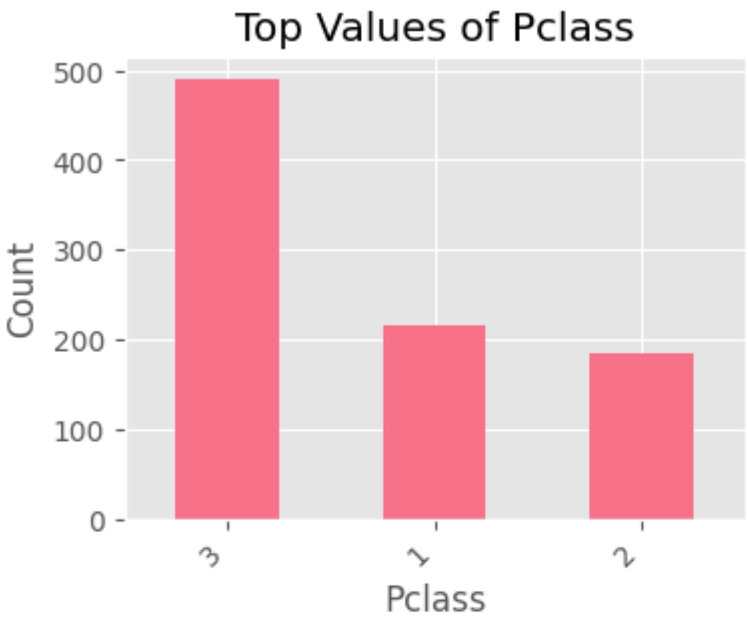
Pclass

Categorical
Missing: 0 (0.0%)
Distinct: 3

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 3
  distinct_scripts: null
  total_characters: 891
length:
  max_length: 1
  mean_length: 1.0
  median_length: 1.0
  min_length: 1
sample:
- '3'
- '1'
- '3'
- '1'
- '3'
```

common values bar



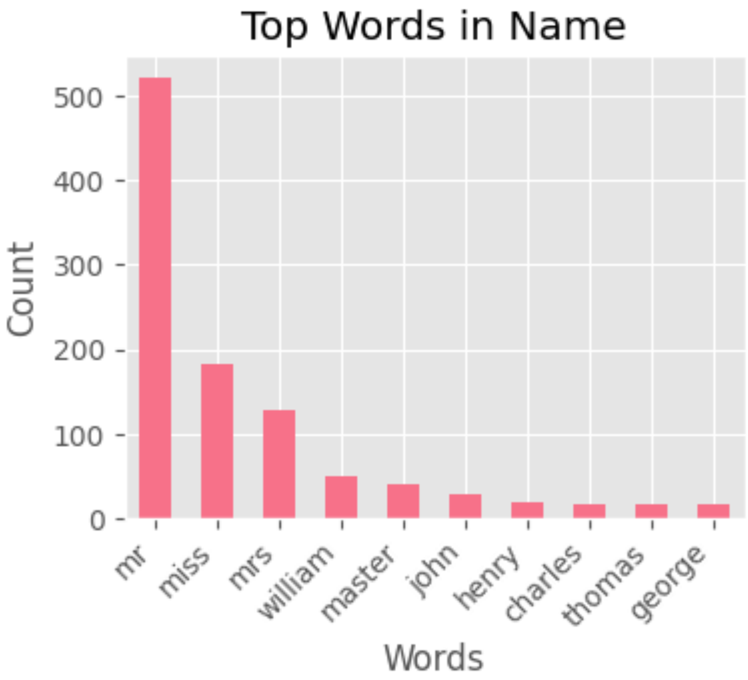
Name

Text
Missing: 0 (0.0%)
Distinct: 891

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 7
  distinct_characters: 60
  distinct_scripts: null
  total_characters: 24026
length:
  max_length: 82
  mean_length: 26.9652076318743
  median_length: 25.0
  min_length: 12
sample:
- Braund, Mr. Owen Harris
- Cumings, Mrs. John Bradley (Florence Briggs Thayer)
- Heikkinen, Miss. Laina
- Futrelle, Mrs. Jacques Heath (Lily May Peel)
- Allen, Mr. William Henry
```

word bar



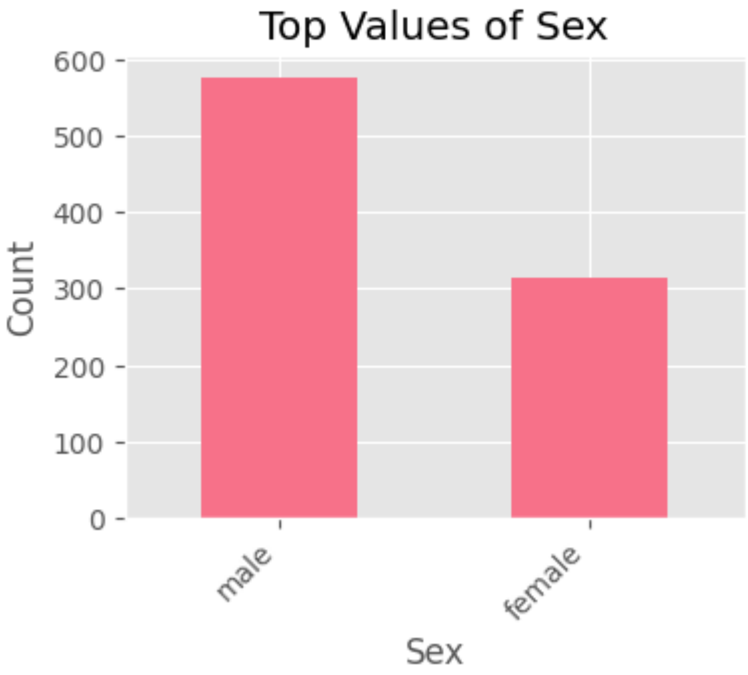
Sex

Categorical
Missing: 0 (0.0%)
Distinct: 2

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 5
  distinct_scripts: null
  total_characters: 4192
length:
  max_length: 6
  mean_length: 4.704826038159371
  median_length: 4.0
  min_length: 4
sample:
- male
- female
- female
- female
- male
```

common values bar



Age

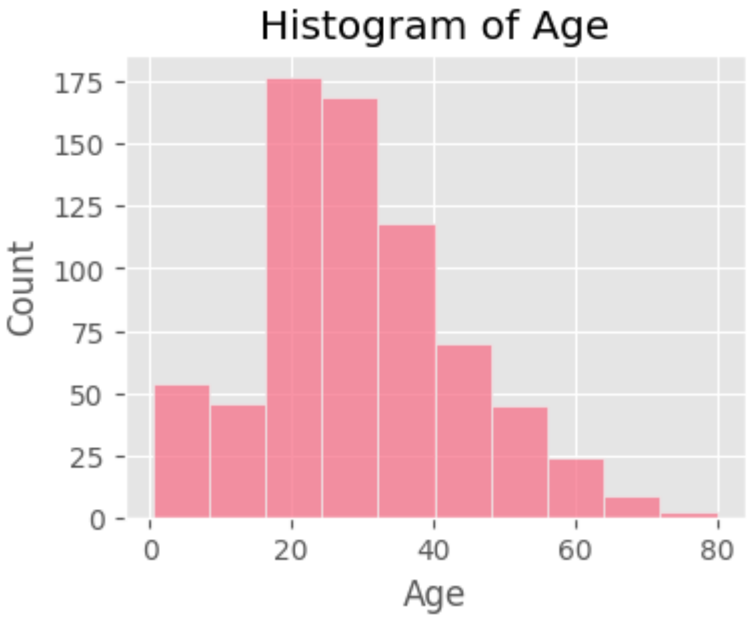
Numeric
Missing: 177 (19.865319865319865%)
Distinct: 88

Statistics

```
descriptive:
  coefficient_of_variation: 0.4891221855465675
  kurtosis: 0.1782741536421022
  mad: 9.0
  mean: 29.69911764705882
  monotonicity: none
  skewness: 0.38910778230082693
  standard_deviation: 14.526497332334042
  sum: 21205.17
  variance: 211.01912474630805
```

```
quantiles:
  iqr: 17.875
  maximum: 80.0
  median: 28.0
  minimum: 0.42
  p5: 4.0
  p95: 56.0
  q1: 20.125
  q3: 38.0
  range: 79.58
```

histogram



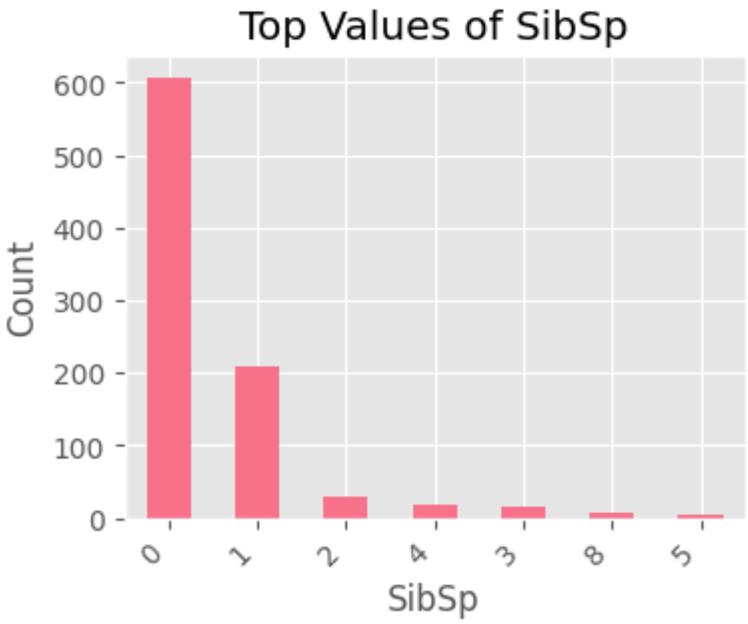
SibSp

Categorical
Missing: 0 (0.0%)
Distinct: 7

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 7
  distinct_scripts: null
  total_characters: 891
length:
  max_length: 1
  mean_length: 1.0
  median_length: 1.0
  min_length: 1
sample:
- '1'
- '1'
- '0'
- '1'
- '0'
```

common values bar



Parch

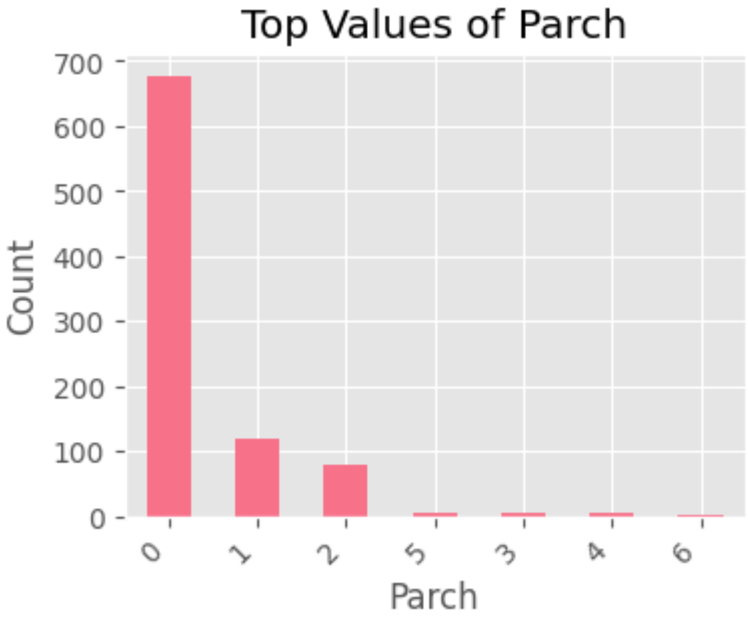
Categorical
Missing: 0 (0.0%)
Distinct: 7

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 7
  distinct_scripts: null
  total_characters: 891
length:
  max_length: 1
  mean_length: 1.0
  median_length: 1.0
  min_length: 1
```

```
sample:
- '0'
- '0'
- '0'
- '0'
- '0'
```

common values bar



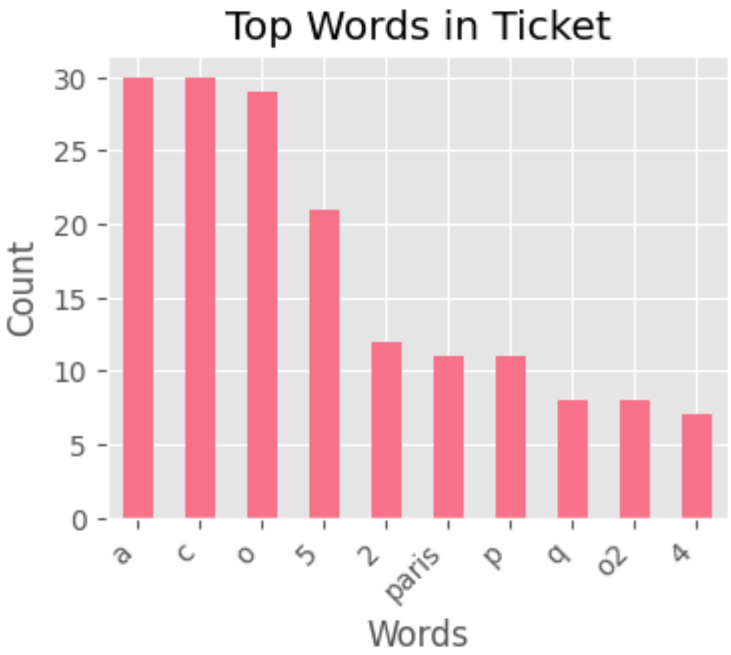
Ticket

Text
Missing: 0 (0.0%)
Distinct: 681

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 5
  distinct_characters: 35
  distinct_scripts: null
  total_characters: 6015
length:
  max_length: 18
  mean_length: 6.750841750841751
  median_length: 6.0
  min_length: 3
sample:
- A/5 21171
- PC 17599
- STON/O2. 3101282
- '113803'
- '373450'
```

word bar



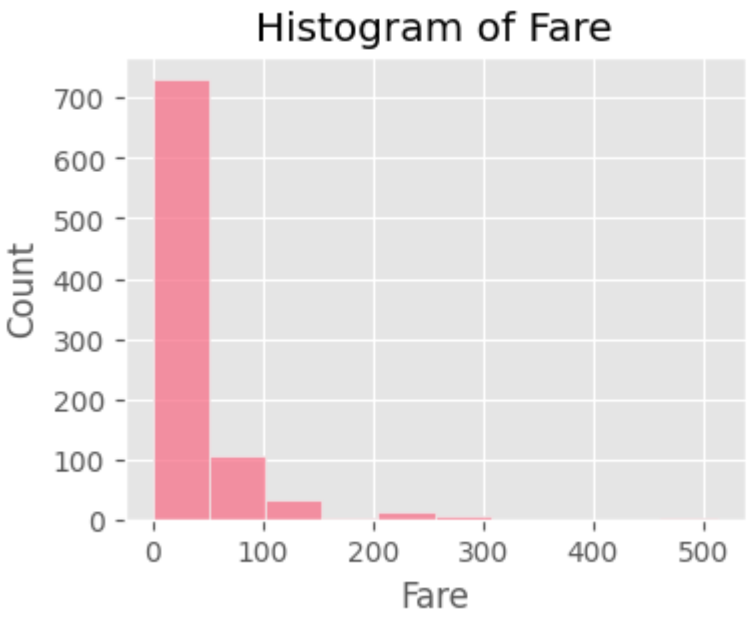
Fare

Numeric
Missing: 0 (0.0%)
Distinct: 248

Statistics

```
descriptive:
  coefficient_of_variation: 1.5430725278408497
  kurtosis: 33.39814088089868
  mad: 6.9042
  mean: 32.204207968574636
  monotonicity: none
  skewness: 4.787316519674893
  standard_deviation: 49.6934285971809
  sum: 28693.9493
  variance: 2469.436845743116
quantiles:
  iqr: 23.0896
  maximum: 512.3292
  median: 14.4542
  minimum: 0.0
  p5: 7.225
  p95: 112.07915
  q1: 7.9104
  q3: 31.0
  range: 512.3292
```

histogram



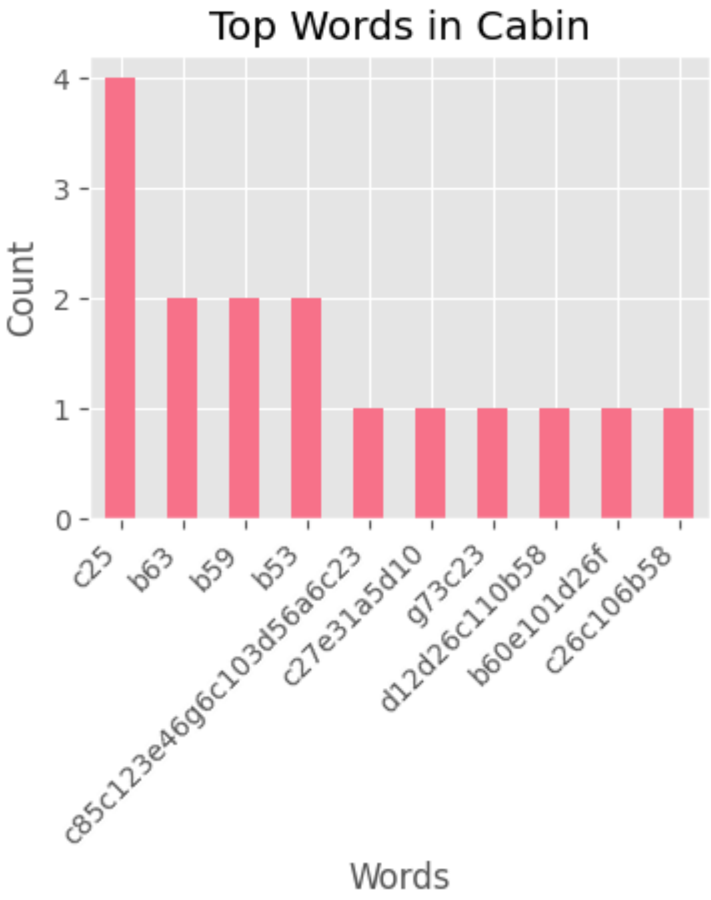
Cabin

Text
Missing: 687 (77.10437710437711%)
Distinct: 147

Statistics

```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 3
  distinct_characters: 19
  distinct_scripts: null
  total_characters: 732
length:
  max_length: 15
  mean_length: 3.588235294117647
  median_length: 3.0
  min_length: 1
sample:
- C85
- C123
- E46
- G6
- C103
```

word bar



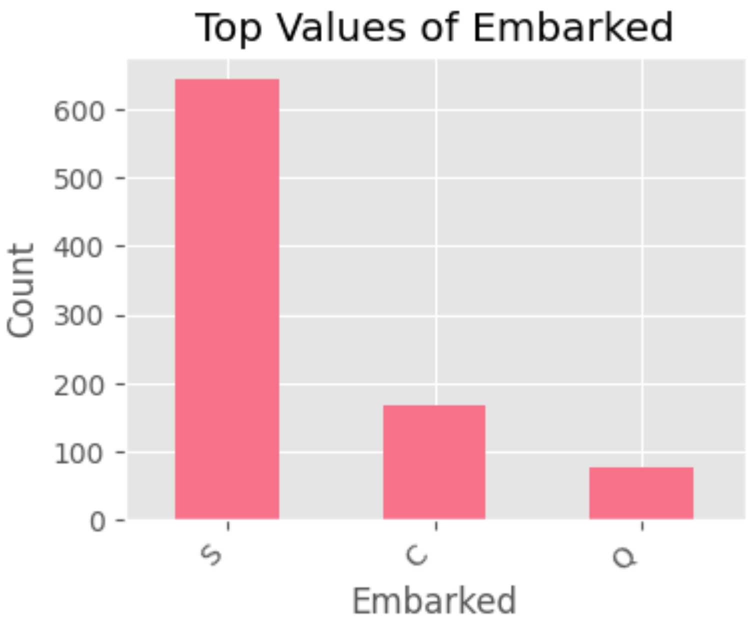
Embarked

Categorical
Missing: 2 (0.22446689113355783%)
Distinct: 3

Statistics

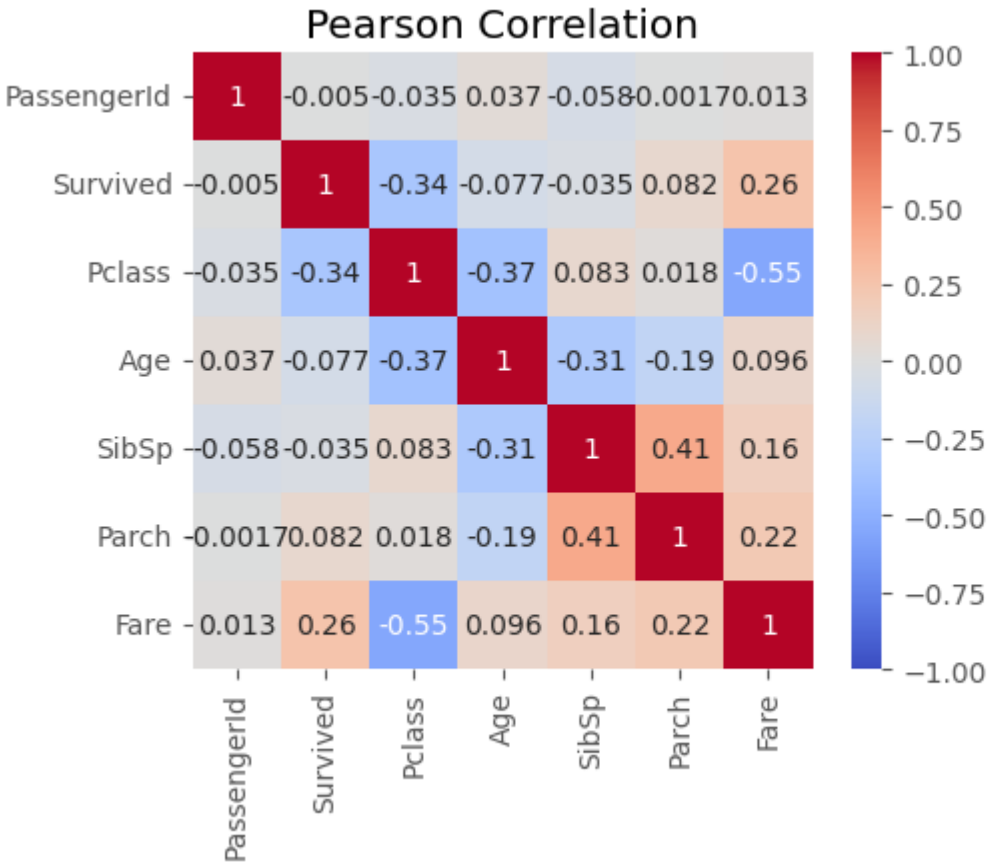
```
characters_and_unicode:
  distinct_blocks: null
  distinct_categories: 1
  distinct_characters: 3
  distinct_scripts: null
  total_characters: 889
length:
  max_length: 1
  mean_length: 1.0
  median_length: 1.0
  min_length: 1
sample:
- S
- C
- S
- S
- S
```

common values bar

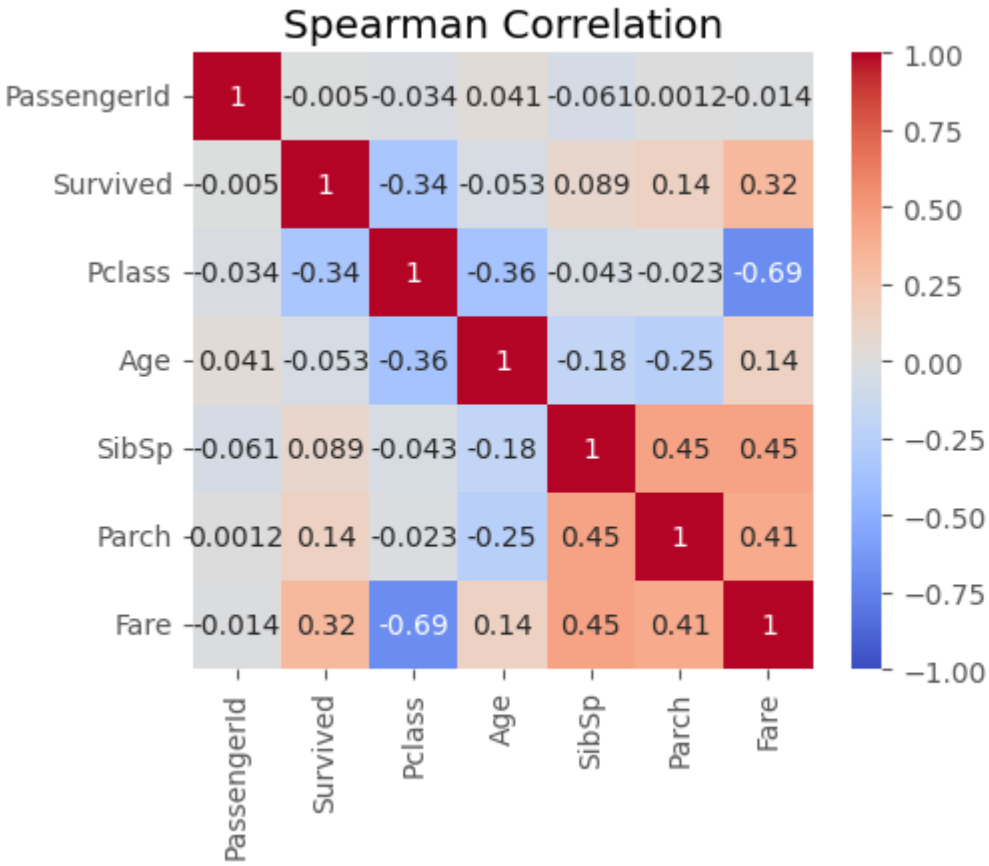


Correlations

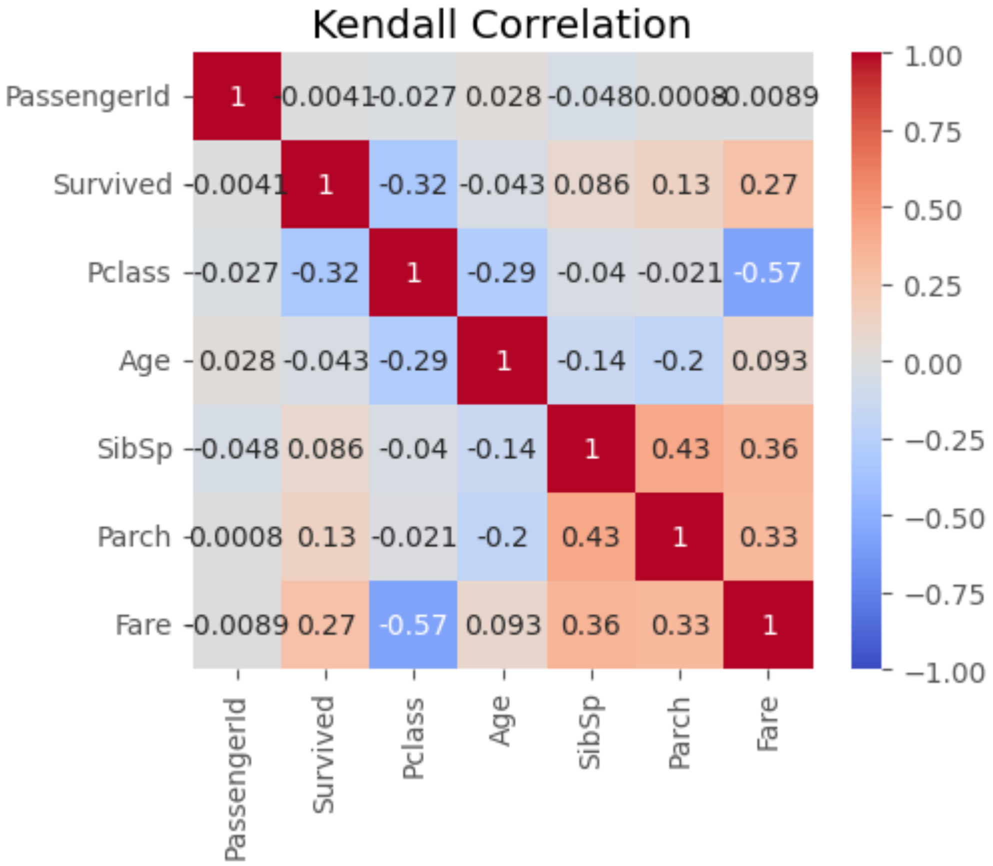
pearson Correlation



spearman Correlation

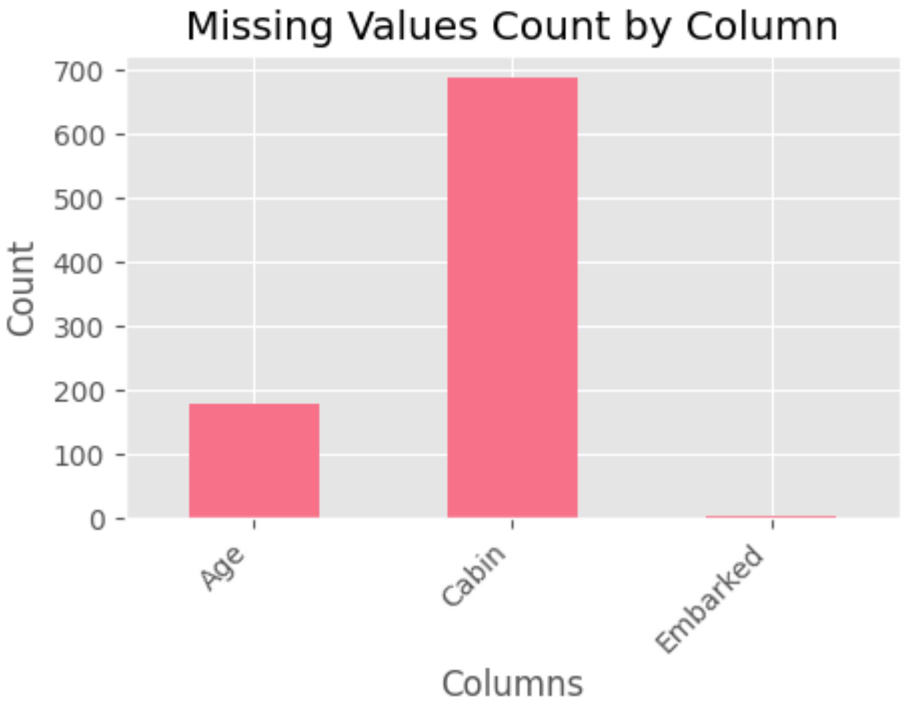


kendall Correlation

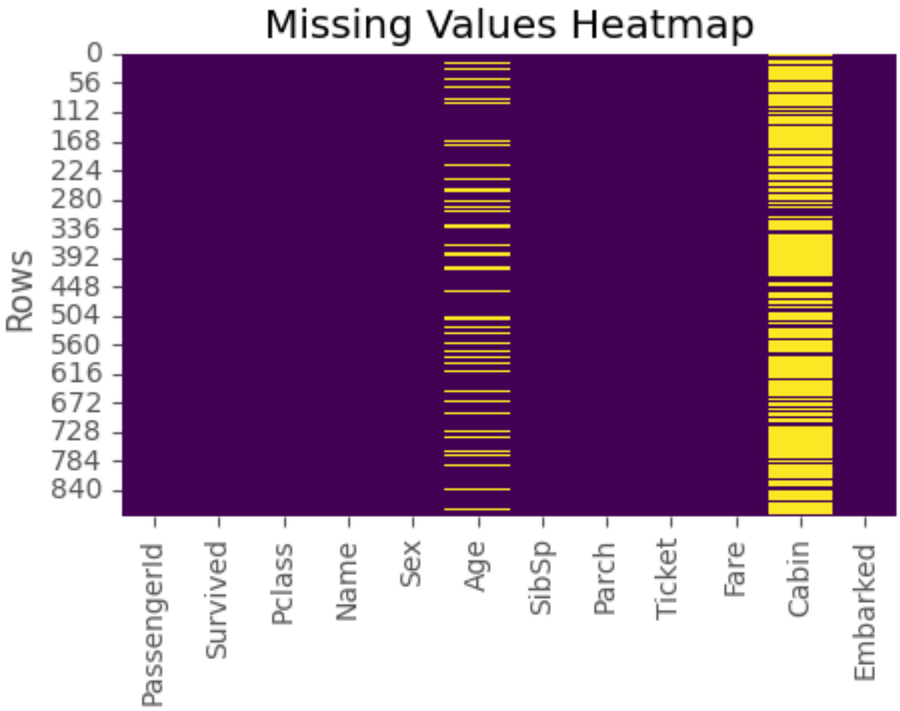


Missing Values

Missing Count



Missing Pattern



Dataset Profile

Rows 891
Columns 12
Total Issues 15

Sample Data

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

Built with HashPrep • Open Source Dataset QA