Dataset Quality Report

Generated: 2025-09-26 02:10:42

HashPrep Version: 0.1.0-alpha

Executive Summary

- Critical Issues: 13
- Warnings: 38
- Rows: 891
- Columns: 12

Issues Overview

Category	Severity	Column	Description	Impact	Quick Fix
missing_values	critical	Cabin	77.1% missing values in 'Cabin'	high	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	critical	Name	Column 'Name' has 891 unique values (100.0% of rows)	high	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	warning	Ticket	Column 'Ticket' has 681 unique values (76.4% of rows)	medium	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).
high_cardinality	warning	Cabin	Column 'Cabin' has 147 unique values (16.5% of rows)	medium	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	warning	SibSp	Column 'SibSp' has 12 potential outliers (1.3% of non-missing values)	medium	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	warning	Parch	Column 'Parch' has 10 potential outliers (1.1% of non-missing values)	medium	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	warning	Fare	Column 'Fare' has 11 potential outliers (1.2% of non-missing values)	medium	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).

feature_correlation	critical	Name,Sex	Columns 'Name' and 'Sex' are highly associated (Cramer's V: 1.00)	high	Options: - Drop one feature: Avoids overfitting from high redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Extract common patterns (e.g., group categories) (Pros: Retains info; Cons: Requires domain knowledge). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	critical	Name,Ticket	Columns 'Name' and 'Ticket' are highly associated (Cramer's V: 1.00)	high	Options: - Drop one feature: Avoids overfitting from high redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Extract common patterns (e.g., group categories) (Pros: Retains info; Cons: Requires domain knowledge). - Retain and

					test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	critical	Name,Cabin	Columns 'Name' and 'Cabin' are highly associated (Cramer's V: 1.00)	high	Options: - Drop one feature: Avoids overfitting from high redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Extract common patterns (e.g., group categories) (Pros: Retains info; Cons: Requires domain knowledge). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	critical	Name,Embarked	Columns 'Name' and 'Embarked' are highly associated (Cramer's V: 1.00)	high	Options: - Drop one feature: Avoids overfitting from high redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Extract common patterns (e.g.,

					group categories) (Pros: Retains info; Cons: Requires domain knowledge). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Sex,Ticket	Columns 'Sex' and 'Ticket' are highly associated (Cramer's V: 0.86)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Group categories or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Sex,Cabin	Columns 'Sex' and 'Cabin' are highly associated (Cramer's V: 0.86)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust

					models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Group categories or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Ticket,Cabin	Columns 'Ticket' and 'Cabin' are highly associated (Cramer's V: 0.95)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Group categories or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	critical	Ticket,Embarked	Columns 'Ticket' and 'Embarked' are highly associated (Cramer's V: 1.00)	high	Options: - Drop one feature: Avoids overfitting from high redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Extract common

					patterns (e.g., group categories) (Pros: Retains info; Cons: Requires domain knowledge). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Cabin,Embarked	Columns 'Cabin' and 'Embarked' are highly associated (Cramer's V: 0.95)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Group categories or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	critical	Sex,Survived	Columns 'Sex' and 'Survived' show strong association (F: 372.41, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature:

					Transform categorical or numeric feature (Pros: Retains info; Cons: Adds complexity). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Sex,Pclass	Columns 'Sex' and 'Pclass' show strong association (F: 15.74, p: 0.0001)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Sex,Age	Columns 'Sex' and 'Age' show strong association (F: 6.25, p: 0.0127)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust

					models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Sex,SibSp	Columns 'Sex' and 'SibSp' show strong association (F: 11.84, p: 0.0006)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	critical	Sex,Parch	Columns 'Sex' and 'Parch' show strong association (F: 57.01, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Transform categorical or numeric

					feature (Pros: Retains info; Cons: Adds complexity). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	critical	Sex,Fare	Columns 'Sex' and 'Fare' show strong association (F: 30.57, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Transform categorical or numeric feature (Pros: Retains info; Cons: Adds complexity). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Ticket,Survived	Columns 'Ticket' and 'Survived' show strong association (F: 3.03, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info;

					Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Ticket,Age	Columns 'Ticket' and 'Age' show strong association (F: 1.72, p: 0.0007)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Ticket,SibSp	Columns 'Ticket' and 'SibSp' show strong association (F: 9.63, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy).

					- Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Ticket,Parch	Columns 'Ticket' and 'Parch' show strong association (F: 4.28, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	critical	Ticket,Fare	Columns 'Ticket' and 'Fare' show strong association (F: 12866198.63, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Transform categorical or numeric feature (Pros: Retains info; Cons: Adds complexity).

					- Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Cabin,PassengerId	Columns 'Cabin' and 'PassengerId' show strong association (F: 1.90, p: 0.0109)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Cabin,Age	Columns 'Cabin' and 'Age' show strong association (F: 2.48, p: 0.0012)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or

					encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Cabin,SibSp	Columns 'Cabin' and 'SibSp' show strong association (F: 10.23, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Cabin,Parch	Columns 'Cabin' and 'Parch' show strong association (F: 11.93, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently

					(Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Cabin,Fare	Columns 'Cabin' and 'Fare' show strong association (F: 5.13, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy; Cons: Adds complexity).
feature_correlation	warning	Embarked,Survived	Columns 'Embarked' and 'Survived' show strong association (F: 13.61, p: 0.0000)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces

					redundancy; Cons: Adds complexity).
feature_correlation	critical	Embarked,Pclass	Columns 'Embarked' and 'Pclass' show strong association (F: 46.51, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Transform categorical or numeric feature (Pros: Retains info; Cons: Adds complexity). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
feature_correlation	warning	Embarked,Parch	Columns 'Embarked' and 'Parch' show strong association (F: 3.23, p: 0.0402)	medium	Options: - Drop one feature: If less predictive (Pros: Simplifies model; Cons: Loses info). - Retain and test: Evaluate with robust models (Pros: Keeps info; Cons: Risk of redundancy). - Engineer feature: Transform or encode differently (Pros: Reduces redundancy;

					Cons: Adds complexity).
feature_correlation	critical	Embarked,Fare	Columns 'Embarked' and 'Fare' show strong association (F: 38.14, p: 0.0000)	high	Options: - Drop one feature: Avoids redundancy (Pros: Simplifies model; Cons: Loses info). - Engineer feature: Transform categorical or numeric feature (Pros: Retains info; Cons: Adds complexity). - Retain and test: Use robust models (e.g., trees) (Pros: Keeps info; Cons: May affect sensitive models).
high_zero_counts	warning	Survived	Column 'Survived' has 61.6% zero values	medium	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	warning	SibSp	Column 'SibSp' has 68.2% zero values	medium	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	warning	Parch	Column 'Parch' has 76.1% zero values	medium	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	warning	Age	Missingness in 'Age' correlates with	medium	Options: - Impute values: Use simple or domain-informed

			'Ticket' (p: 0.0000)		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).
missing_patterns	warning	Age	Missingness in 'Age' correlates with 'Embarked' (p: 0.0000)	medium	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).
missing_patterns	warning	Age	Missingness in 'Age' correlates with numeric 'Survived' (F: 7.62, p: 0.0059)	medium	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).
missing_patterns	warning	Age	Missingness in 'Age' correlates with numeric 'Pclass' (F: 27.41, p: 0.0000)	medium	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).
missing_patterns	warning	Age	Missingness in 'Age' correlates with numeric 'Parch' (F: 13.91, p: 0.0002)	medium	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).
missing_patterns	warning	Age	Missingness in 'Age' correlates with numeric 'Fare' (F: 9.11, p: 0.0026)	medium	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).
missing_patterns	warning	Cabin	Missingness in 'Cabin' correlates with 'Sex' (p: 0.0000)	medium	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).
missing_patterns	warning	Cabin	Missingness in 'Cabin' correlates with 'Embarked' (p: 0.0000)	medium	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).
missing_patterns	warning	Cabin	Missingness in 'Cabin' correlates with numeric 'Survived' (F: 99.25, p: 0.0000)	medium	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).
missing_patterns	warning	Cabin	Missingness in 'Cabin' correlates with numeric 'Pclass' (F: 988.15, p: 0.0000)	medium	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).
missing_patterns	warning	Cabin	Missingness in 'Cabin' correlates with numeric 'Age' (F: 47.36, p: 0.0000)	medium	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).
missing_patterns	warning	Cabin		medium	

			Missingness in 'Cabin' correlates with numeric 'Fare' (F: 269.15, p: 0.0000)		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).

Dataset Preview

Head

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05

Tail

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75

Sample

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
800	0	3	Van Impe, Mrs. Jean Baptiste (Rosalie Paula Govaert)	female	30.0	1	1	345773	24.15
429	0	3	Flynn, Mr. James	male	NaN	0	0	364851	7.750
346	1	2	Brown, Miss. Amelia "Mildred"	female	24.0	0	0	248733	13.00

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
394	1	1	Newell, Miss. Marjorie	female	23.0	1	0	35273	113.2
124	1	2	Webber, Miss. Susan	female	32.5	0	0	27267	13.00
760	1	1	Rothes, the Countess. of (Lucy Noel Martha Dyer-Edwards)	female	33.0	0	0	110152	86.50
469	0	3	Scanlan, Mr. James	male	NaN	0	0	36209	7.725
456	1	3	Jalsevac, Mr. Ivan	male	29.0	0	0	349240	7.895
161	0	3	Cribb, Mr. John Hatfield	male	44.0	0	1	371362	16.10
96	0	3	Shorney, Mr. Charles Joseph	male	NaN	0	0	374910	8.050

Variables

PassengerId


```
count: 891
histogram:
  bin_edges:
  - 1.0
  - 90.0
  - 179.0
  - 268.0
  - 357.0
  - 446.0
  - 535.0
  - 624.0
  - 713.0
  - 802.0
```

  - 891.0
  counts:
  - 89
  - 89
  - 89
  - 89
  - 89
  - 89
  - 89
  - 89
  - 89
  - 90
max: 891.0
mean: 446.0
min: 1.0
missing: 0
quantiles:
  25%: 223.5
  50%: 446.0
  75%: 668.5
std: 257.3538420152301
zeros: 0

Survived

count: 891
histogram:
  bin_edges:
  - 0.0
  - 0.1
  - 0.2
  - 0.30000000000000004
  - 0.4
  - 0.5
  - 0.6000000000000001
  - 0.7000000000000001
  - 0.8
  - 0.9
  - 1.0
  counts:
  - 549
  - 0
  - 0
  - 0
  - 0
  - 0
  - 0
  - 0
  - 0
  - 342
max: 1.0
mean: 0.3838383838383838

```
min: 0.0
missing: 0
quantiles:
  25%: 0.0
  50%: 0.0
  75%: 1.0
std: 0.4865924542648575
zeros: 549
```


Pclass


```
count: 891
histogram:
  bin_edges:
  - 1.0
  - 1.2
  - 1.4
  - 1.6
  - 1.8
  - 2.0
  - 2.2
  - 2.4000000000000004
  - 2.6
  - 2.8
  - 3.0
  counts:
  - 216
  - 0
  - 0
  - 0
  - 0
  - 184
  - 0
  - 0
  - 0
  - 491
max: 3.0
mean: 2.308641975308642
min: 1.0
missing: 0
quantiles:
  25%: 2.0
  50%: 3.0
  75%: 3.0
std: 0.836071240977049
zeros: 0
```


Name


```
avg_length: 26.9652076318743
char_freq:
```

' ': 2735
  M: 1128
  a: 1657
  e: 1703
  i: 1325
  l: 1067
  n: 1304
  o: 1008
  r: 1958
  s: 1297
common_lengths:
  18: 50
  19: 64
  25: 55
  26: 49
  27: 50
count: 891
max_length: 82.0
min_length: 12.0
missing: 0

Sex

avg_length: 4.704826038159371
char_freq:
  a: 891
  e: 1205
  f: 314
  l: 891
  m: 891
common_lengths:
  4: 577
  6: 314
count: 891
max_length: 6.0
min_length: 4.0
missing: 0

Age

count: 714
histogram:
  bin_edges:
  - 0.42
  - 8.378
  - 16.336000000000002
  - 24.294000000000004
  - 32.252
  - 40.21
  - 48.168000000000006
  - 56.126000000000005

  - 64.084
  - 72.042
  - 80.0
  counts:
  - 54
  - 46
  - 177
  - 169
  - 118
  - 70
  - 45
  - 24
  - 9
  - 2
max: 80.0
mean: 29.69911764705882
min: 0.42
missing: 177
quantiles:
  25%: 20.125
  50%: 28.0
  75%: 38.0
std: 14.526497332334042
zeros: 0

SibSp

count: 891
histogram:
  bin_edges:
  - 0.0
  - 0.8
  - 1.6
  - 2.4000000000000004
  - 3.2
  - 4.0
  - 4.800000000000001
  - 5.6000000000000005
  - 6.4
  - 7.2
  - 8.0
  counts:
  - 608
  - 209
  - 28
  - 16
  - 0
  - 18
  - 5
  - 0
  - 0
  - 7

max: 8.0
mean: 0.5230078563411896
min: 0.0
missing: 0
quantiles:
  25%: 0.0
  50%: 0.0
  75%: 1.0
std: 1.1027434322934317
zeros: 608

Parch

count: 891
histogram:
  bin_edges:
  - 0.0
  - 0.6
  - 1.2
  - 1.7999999999999998
  - 2.4
  - 3.0
  - 3.5999999999999996
  - 4.2
  - 4.8
  - 5.3999999999999995
  - 6.0
  counts:
  - 678
  - 118
  - 0
  - 80
  - 0
  - 5
  - 4
  - 0
  - 5
  - 1
max: 6.0
mean: 0.38159371492704824
min: 0.0
missing: 0
quantiles:
  25%: 0.0
  50%: 0.0
  75%: 0.0
std: 0.8060572211299483
zeros: 678

Ticket

avg_length: 6.750841750841751
char_freq:
  '0': 406
  '1': 689
  '2': 594
  '3': 746
  '4': 464
  '5': 387
  '6': 422
  '7': 490
  '8': 282
  '9': 328
common_lengths:
  4: 101
  5: 131
  6: 419
  8: 76
  10: 41
count: 891
max_length: 18.0
min_length: 3.0
missing: 0

Fare

count: 891
histogram:
  bin_edges:
  - 0.0
  - 51.23292
  - 102.46584
  - 153.69876
  - 204.93168
  - 256.1646
  - 307.39752
  - 358.63044
  - 409.86336
  - 461.09628
  - 512.3292
  counts:
  - 732
  - 106
  - 31
  - 2
  - 11
  - 6
  - 0
  - 0
  - 0

  - 3
max: 512.3292
mean: 32.204207968574636
min: 0.0
missing: 0
quantiles:
  25%: 7.9104
  50%: 14.4542
  75%: 31.0
std: 49.6934285971809
zeros: 15

Cabin

count: 204
missing: 687
most_frequent: B96 B98
top_values:
  B96 B98: 4
  C123: 2
  C22 C26: 3
  C23 C25 C27: 4
  C83: 2
  D: 3
  E101: 3
  F2: 3
  F33: 3
  G6: 4
unique: 147

Embarked

count: 889
missing: 2
most_frequent: S
top_values:
  C: 168
  Q: 77
  S: 644
unique: 3

Correlations

Numeric (Pearson)

{
  "PassengerId": {
    "PassengerId": 1.0,
    "Survived": -0.0050066607670665175,
    "Pclass": -0.03514399403038102,
    "Age": 0.036847197861327674,

    "SibSp": -0.0575268337844415,
    "Parch": -0.0016520124027188366,
    "Fare": 0.012658219287491099
  },
  "Survived": {
    "PassengerId": -0.0050066607670665175,
    "Survived": 1.0,
    "Pclass": -0.33848103596101514,
    "Age": -0.07722109457217756,
    "SibSp": -0.035322498885735576,
    "Parch": 0.08162940708348335,
    "Fare": 0.2573065223849626
  },
  "Pclass": {
    "PassengerId": -0.03514399403038102,
    "Survived": -0.33848103596101514,
    "Pclass": 1.0,
    "Age": -0.36922601531551735,
    "SibSp": 0.08308136284568686,
    "Parch": 0.018442671310748508,
    "Fare": -0.5494996199439076
  },
  "Age": {
    "PassengerId": 0.036847197861327674,
    "Survived": -0.07722109457217756,
    "Pclass": -0.36922601531551735,
    "Age": 1.0,
    "SibSp": -0.30824675892365666,
    "Parch": -0.1891192626320352,
    "Fare": 0.09606669176903912
  },
  "SibSp": {
    "PassengerId": -0.0575268337844415,
    "Survived": -0.035322498885735576,
    "Pclass": 0.08308136284568686,
    "Age": -0.30824675892365666,
    "SibSp": 1.0,
    "Parch": 0.41483769862015624,
    "Fare": 0.159651043242161
  },
  "Parch": {
    "PassengerId": -0.0016520124027188366,
    "Survived": 0.08162940708348335,
    "Pclass": 0.018442671310748508,
    "Age": -0.1891192626320352,
    "SibSp": 0.41483769862015624,
    "Parch": 1.0,
    "Fare": 0.21622494477076448
  },
  "Fare": {
    "PassengerId": 0.012658219287491099,
    "Survived": 0.2573065223849626,

```
    "Pclass": -0.5494996199439076,
    "Age": 0.09606669176903912,
    "SibSp": 0.159651043242161,
    "Parch": 0.21622494477076448,
    "Fare": 1.0
  }
}
```

Categorical (Cramer's V)

Pair	Value
Name__Sex	1.0
Name__Ticket	1.0
Name__Cabin	1.0
Name__Embarked	1.0
Sex__Ticket	0.86
Sex__Cabin	0.86
Sex__Embarked	0.12
Ticket__Cabin	0.95
Ticket__Embarked	1.0
Cabin__Embarked	0.95

Mixed

Pair	F-Stat	P-Value
Sex__PassengerId	1.64	0.2004
Sex__Survived	372.41	0.0
Sex__Pclass	15.74	0.0001
Sex__Age	6.25	0.0127
Sex__SibSp	11.84	0.0006
Sex__Parch	57.01	0.0
Sex__Fare	30.57	0.0
Ticket__PassengerId	1.05	0.3676

Ticket__Survived	3.03	0.0
Ticket__Age	1.72	0.0007
Ticket__SibSp	9.63	0.0
Ticket__Parch	4.28	0.0
Ticket__Fare	12866198.63	0.0
Cabin__PassengerId	1.9	0.0109
Cabin__Survived	1.26	0.2054
Cabin__Age	2.48	0.0012
Cabin__SibSp	10.23	0.0
Cabin__Parch	11.93	0.0
Cabin__Fare	5.13	0.0
Embarked__PassengerId	0.52	0.5941
Embarked__Survived	13.61	0.0
Embarked__Pclass	46.51	0.0
Embarked__Age	0.64	0.5294
Embarked__SibSp	2.18	0.1132
Embarked__Parch	3.23	0.0402
Embarked__Fare	38.14	0.0

Missing Values

Column	Count	Percentage
Age	177	19.87
Cabin	687	77.1
Embarked	2	0.22

Missing Patterns


```
{
  "Age": [
    5,
```

17,
19,
26,
28,
29,
31,
32,
36,
42,
45,
46,
47,
48,
55,
64,
65,
76,
77,
82,
87,
95,
101,
107,
109,
121,
126,
128,
140,
154,
158,
159,
166,
168,
176,
180,
181,
185,
186,
196,
198,
201,
214,
223,
229,
235,
240,
241,
250,
256,
260,
264,
270,

274,
277,
284,
295,
298,
300,
301,
303,
304,
306,
324,
330,
334,
335,
347,
351,
354,
358,
359,
364,
367,
368,
375,
384,
388,
409,
410,
411,
413,
415,
420,
425,
428,
431,
444,
451,
454,
457,
459,
464,
466,
468,
470,
475,
481,
485,
490,
495,
497,
502,
507,
511,

517,
522,
524,
527,
531,
533,
538,
547,
552,
557,
560,
563,
564,
568,
573,
578,
584,
589,
593,
596,
598,
601,
602,
611,
612,
613,
629,
633,
639,
643,
648,
650,
653,
656,
667,
669,
674,
680,
692,
697,
709,
711,
718,
727,
732,
738,
739,
740,
760,
766,
768,
773,

```
    776,
    778,
    783,
    790,
    792,
    793,
    815,
    825,
    826,
    828,
    832,
    837,
    839,
    846,
    849,
    859,
    863,
    868,
    878,
    888
  ],
  "Cabin": [
    0,
    2,
    4,
    5,
    7,
    8,
    9,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    22,
    24,
    25,
    26,
    28,
    29,
    30,
    32,
    33,
    34,
    35,
    36,
    37,
    38,
```

39,
40,
41,
42,
43,
44,
45,
46,
47,
48,
49,
50,
51,
53,
56,
57,
58,
59,
60,
63,
64,
65,
67,
68,
69,
70,
71,
72,
73,
74,
76,
77,
78,
79,
80,
81,
82,
83,
84,
85,
86,
87,
89,
90,
91,
93,
94,
95,
98,
99,
100,
101,

103,
104,
105,
106,
107,
108,
109,
111,
112,
113,
114,
115,
116,
117,
119,
120,
121,
122,
125,
126,
127,
129,
130,
131,
132,
133,
134,
135,
138,
140,
141,
142,
143,
144,
145,
146,
147,
149,
150,
152,
153,
154,
155,
156,
157,
158,
159,
160,
161,
162,
163,
164,

165,
167,
168,
169,
171,
172,
173,
175,
176,
178,
179,
180,
181,
182,
184,
186,
187,
188,
189,
190,
191,
192,
196,
197,
198,
199,
200,
201,
202,
203,
204,
206,
207,
208,
210,
211,
212,
213,
214,
216,
217,
219,
220,
221,
222,
223,
225,
226,
227,
228,
229,
231,

232,
233,
234,
235,
236,
237,
238,
239,
240,
241,
242,
243,
244,
246,
247,
249,
250,
253,
254,
255,
256,
258,
259,
260,
261,
264,
265,
266,
267,
270,
271,
272,
274,
276,
277,
278,
279,
280,
281,
282,
283,
285,
286,
287,
288,
289,
290,
293,
294,
295,
296,
300,

301,
302,
304,
306,
308,
312,
313,
314,
315,
316,
317,
320,
321,
322,
323,
324,
326,
328,
330,
333,
334,
335,
338,
342,
343,
344,
346,
347,
348,
349,
350,
352,
353,
354,
355,
357,
358,
359,
360,
361,
362,
363,
364,
365,
367,
368,
371,
372,
373,
374,
375,
376,

378,
379,
380,
381,
382,
383,
384,
385,
386,
387,
388,
389,
391,
392,
395,
396,
397,
398,
399,
400,
401,
402,
403,
404,
405,
406,
407,
408,
409,
410,
411,
413,
414,
415,
416,
417,
418,
419,
420,
421,
422,
423,
424,
425,
426,
427,
428,
431,
432,
433,
436,
437,

439,
440,
441,
442,
443,
444,
446,
447,
448,
450,
451,
454,
455,
458,
459,
461,
463,
464,
465,
466,
467,
468,
469,
470,
471,
472,
474,
476,
477,
478,
479,
480,
481,
482,
483,
485,
488,
489,
490,
491,
493,
494,
495,
497,
499,
500,
501,
502,
503,
506,
507,
508,

509,
510,
511,
513,
514,
517,
518,
519,
521,
522,
524,
525,
526,
528,
529,
530,
531,
532,
533,
534,
535,
537,
538,
541,
542,
543,
545,
546,
547,
548,
549,
551,
552,
553,
554,
555,
557,
559,
560,
561,
562,
563,
564,
565,
566,
567,
568,
569,
570,
573,
574,
575,

576,
578,
579,
580,
582,
584,
586,
588,
589,
590,
592,
593,
594,
595,
596,
597,
598,
600,
601,
602,
603,
604,
605,
606,
607,
608,
610,
611,
612,
613,
614,
615,
616,
617,
619,
620,
622,
623,
624,
626,
628,
629,
631,
633,
634,
635,
636,
637,
638,
639,
640,
642,

643,
644,
646,
648,
649,
650,
651,
652,
653,
654,
655,
656,
657,
658,
660,
661,
663,
664,
665,
666,
667,
668,
670,
672,
673,
674,
675,
676,
677,
678,
680,
682,
683,
684,
685,
686,
687,
688,
691,
692,
693,
694,
695,
696,
697,
702,
703,
704,
705,
706,
708,
709,

713,
714,
718,
719,
720,
721,
722,
723,
725,
726,
727,
728,
729,
731,
732,
733,
734,
735,
736,
738,
739,
743,
744,
746,
747,
749,
750,
752,
753,
754,
755,
756,
757,
758,
760,
761,
762,
764,
766,
767,
768,
769,
770,
771,
773,
774,
775,
777,
778,
780,
783,
784,

785,
786,
787,
788,
790,
791,
792,
793,
794,
795,
797,
798,
799,
800,
801,
803,
804,
805,
807,
808,
810,
811,
812,
813,
814,
816,
817,
818,
819,
821,
822,
824,
825,
826,
827,
828,
830,
831,
832,
833,
834,
836,
837,
838,
840,
841,
842,
843,
844,
845,
846,
847,

```
    848,
    850,
    851,
    852,
    854,
    855,
    856,
    858,
    859,
    860,
    861,
    863,
    864,
    865,
    866,
    868,
    869,
    870,
    873,
    874,
    875,
    876,
    877,
    878,
    880,
    881,
    882,
    883,
    884,
    885,
    886,
    888,
    890
  ],
  "Embarked": [
    61,
    829
  ]
}
```

Next Steps

- Address critical issues
- Handle warnings
- Re-analyze dataset

Generated by HashPrep