

# EDA AUTOMATION TOOL

```
In [1]: import pandas as pd
import numpy as np
```

```
In [2]: df=pd.read_csv("D:/Downloads/archive (20)/tested.csv")
```

```
In [4]: df.head()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	



## ProfileReport

```
In [ ]: from ydata_profiling import ProfileReport

ProfileReport(df, title="My Data Profile").to_file("report.html")
```

## SWEETVIZ

```
In [ ]: pip install sweetviz
```

```
In [ ]: import sweetviz as sv
```

```
In [ ]: df=pd.read_csv("D:/Downloads/archive (20)/tested.csv")
my_report = sv.analyze(df)
my_report.show_html()
```

#### Benefits of EDA Automation Tools:

**Efficiency:** Automation tools can significantly reduce the time and effort required to perform EDA tasks, especially for large and complex datasets. They can quickly generate summary statistics, visualizations, and insights, allowing analysts to focus on interpreting the results rather than repetitive tasks.

**Consistency:** Automated EDA processes ensure consistency in analysis across different datasets and users. By following standardized procedures and algorithms, these tools minimize human errors and biases that may arise from manual analysis.

**Scalability:** Automation tools are well-suited for handling large volumes of data, making them scalable to diverse applications and datasets. They can efficiently process extensive datasets that may be impractical to analyze manually.

**Exploration:** Automation tools can facilitate exploratory data analysis by providing interactive visualizations and intuitive interfaces for data exploration. They enable users to interactively explore the data, identify patterns, and gain insights in real-time.

#### Challenges and Limitations:

**Overreliance:** There's a risk of overreliance on automation tools, leading to a lack of critical thinking and domain expertise in data analysis. Users may become dependent on automated results without fully understanding the underlying assumptions or nuances of the data.

**Black Box Nature:** Some automation tools operate as "black boxes," meaning that users may not have full visibility into the algorithms and methodologies used. This lack of transparency can make it challenging to interpret the results accurately or troubleshoot issues effectively.

**Customization:** Automation tools may lack flexibility or customization options to accommodate specific analysis requirements or domain-specific knowledge. Users may find it challenging to tailor the automated processes to suit their unique needs or preferences.

**Quality Assurance:** Automated EDA processes require rigorous quality assurance to ensure the accuracy and reliability of the results. Users must validate the outputs against manual analyses and domain knowledge to verify their correctness.

## Overview

OverviewAlerts7Reproduction

Dataset statistics

Number of variables	12
Number of observations	418
Missing cells	414
Missing cells (%)	8.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	39.3 KiB
Average record size in memory	96.3 B

Variable types

Numeric	5
Categorical	4
Text	3

## Variables

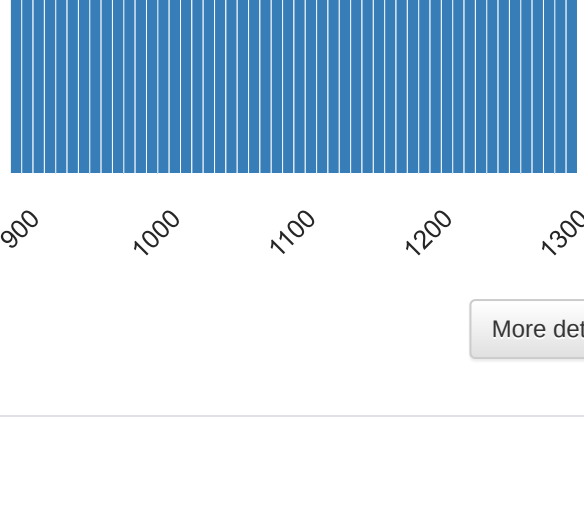
Select Columns

PassengerId

Real number (R)

UNIFORMUNIQUE

Distinct	418	Minimum	892
Distinct (%)	100.0%	Maximum	1309
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1100.5	Memory size	3.4 KiB




More details

Survived

Categorical

Distinct	2
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB

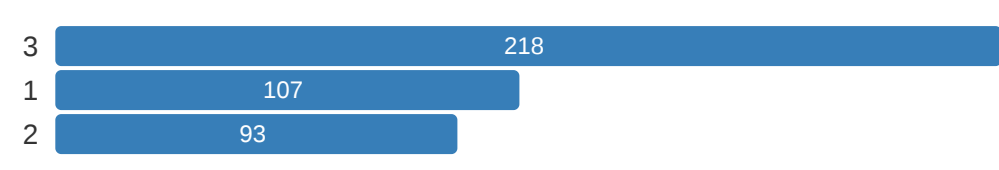


More details

Pclass

Categorical

Distinct	3
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB




More details

Name

Text

UNIQUE

Distinct	418
Distinct (%)	100.0%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB




More details

Sex

Categorical

Distinct	2
Distinct (%)	0.5%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB



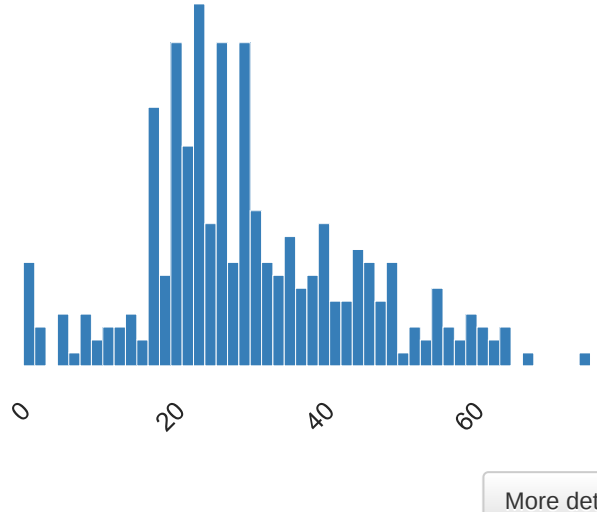
More details

Age

Real number (R)

MISSING

Distinct	79	Minimum	0.17
Distinct (%)	23.8%	Maximum	76
Missing	86	Zeros	0
Missing (%)	20.6%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	30.27259	Memory size	3.4 KiB



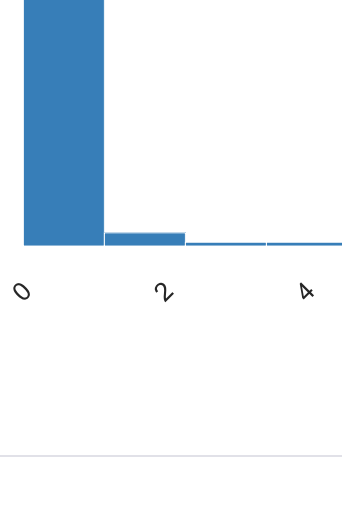
More details

SibSp

Real number (R)

ZEROS

Distinct	7	Minimum	0
Distinct (%)	1.7%	Maximum	8
Missing	0	Zeros	283
Missing (%)	0.0%	Zeros (%)	67.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.44736842	Memory size	3.4 KiB



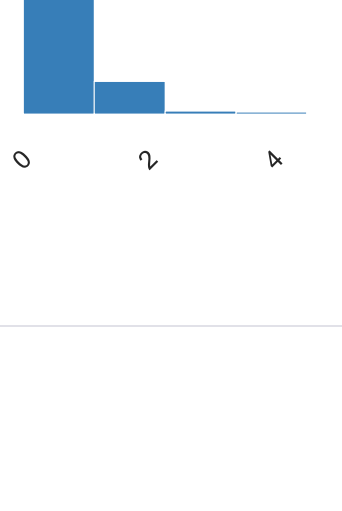
More details

Parch

Real number (R)

ZEROS

Distinct	8	Minimum	0
Distinct (%)	1.9%	Maximum	9
Missing	0	Zeros	324
Missing (%)	0.0%	Zeros (%)	77.5%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.3923445	Memory size	3.4 KiB

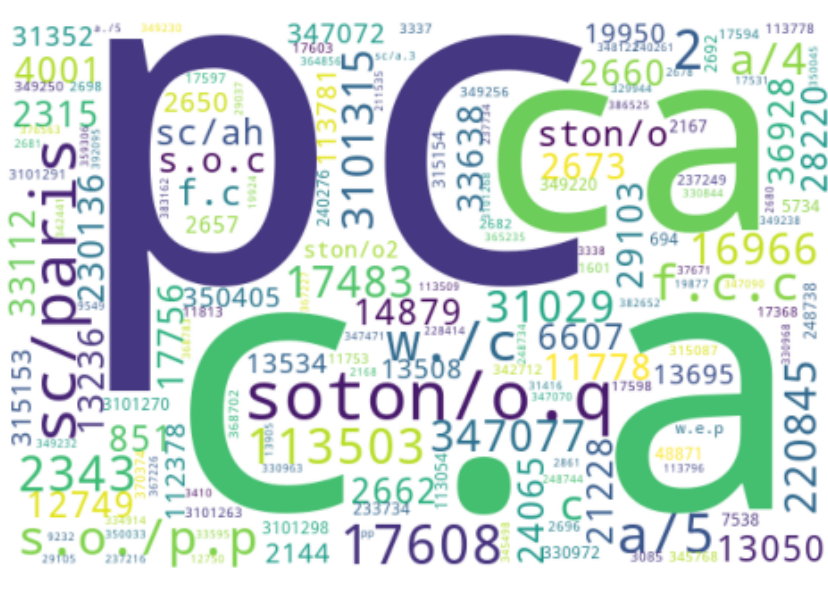


More details

Ticket

Text

Distinct	363
Distinct (%)	86.8%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB

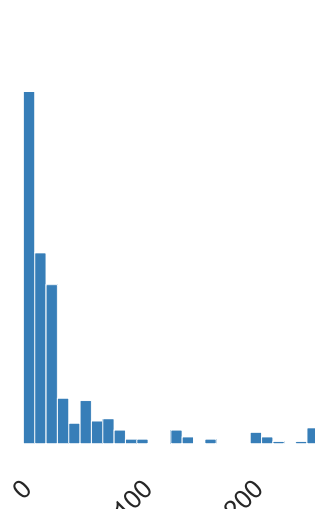


More details

Fare

Real number (R)

Distinct	169	Minimum	0
Distinct (%)	40.5%	Maximum	512.3292
Missing	1	Zeros	2
Missing (%)	0.2%	Zeros (%)	0.5%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	35.627188	Memory size	3.4 KiB




More details

Cabin

Text

MISSING

Distinct	76
Distinct (%)	83.5%
Missing	327
Missing (%)	78.2%
Memory size	3.4 KiB




More details

Embarked

Categorical

Distinct	3
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	3.4 KiB



More details

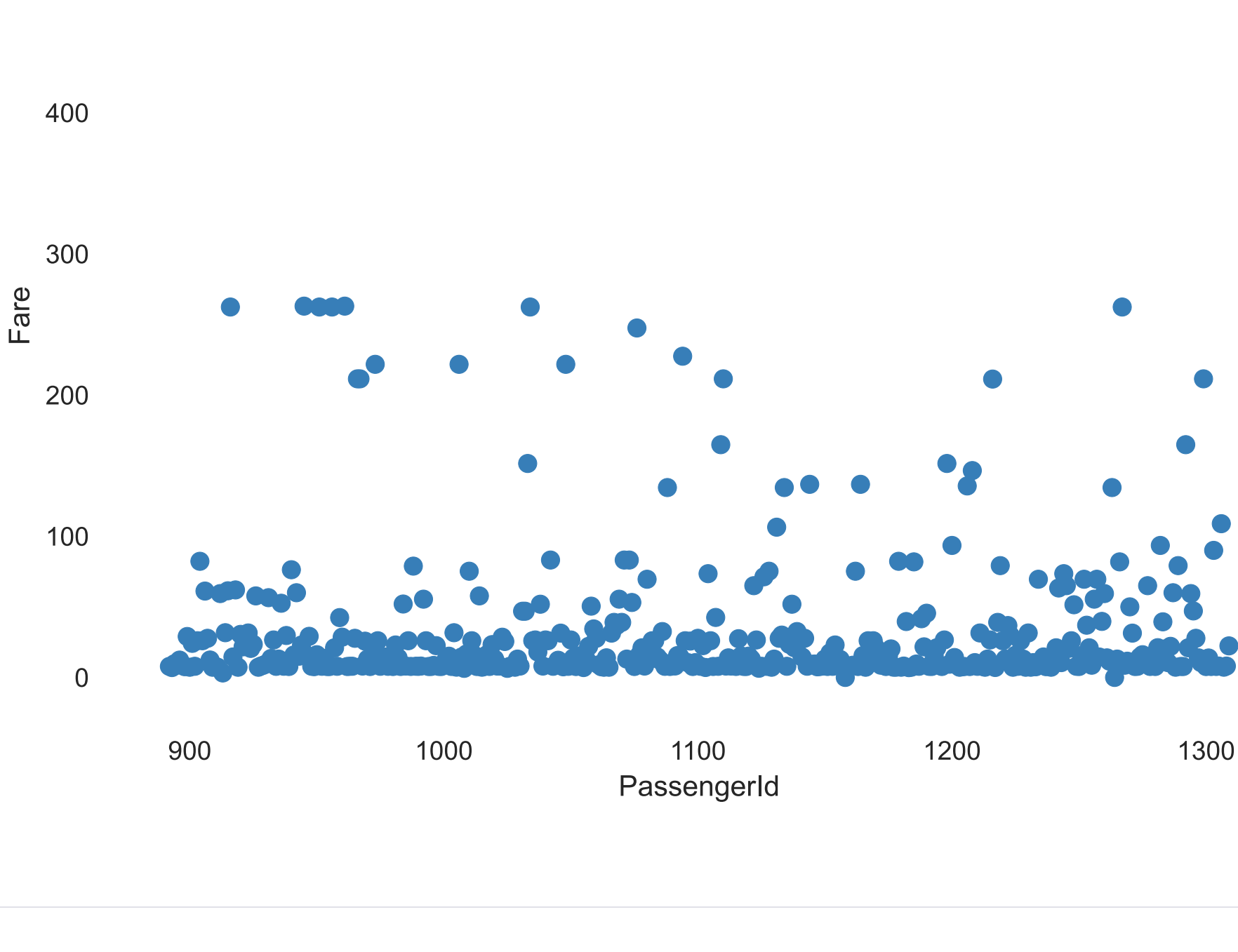
## Interactions

PassengerId

AgeSibSpParchFare

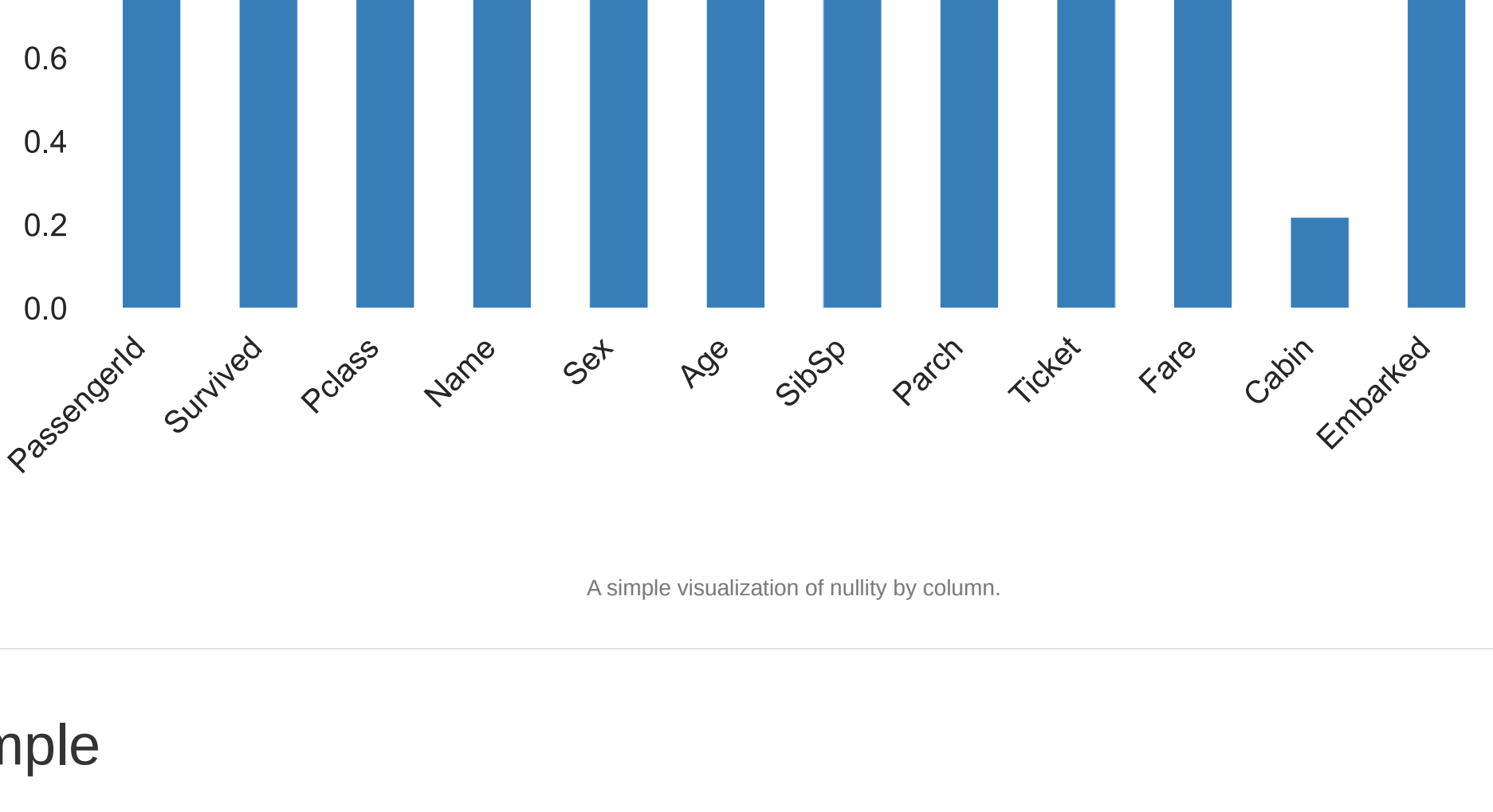
Fare

PassengerIdAgeSibSpParch



## Missing values

CountMatrix



Variable	Count
PassengerId	418
Survived	418
Pclass	418
Name	418
Sex	418
Age	332
SibSp	418
Parch	418
Ticket	418
Fare	417
Cabin	91
Embarked	418

A simple visualization of nullity by column.

## Sample

First rowsLast rows

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S
5	897	0	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S
6	898	1	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q
7	899	0	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S
8	900	1	3	Abraham, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C
9	901	0	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S

Report generated by YData.



DataFrame

NO COMPARISON TARGET

418 ROWS  
0 DUPLICATES  
151.6 kb RAM  
12 FEATURES  
6 CATEGORICAL  
3 NUMERICAL  
3 TEXT

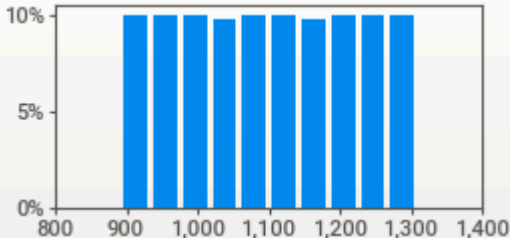
ASSOCIATIONS

DataFrame

PassengerId

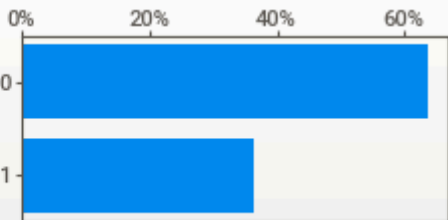
VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 418 (100%)  
ZEROES: ---

MAX	1,309	RANGE	417
95%	1,288	IQR	208
Q3	1,205	STD	121
MEDIAN	1,100	VAR	14,595
AVG	1,100		
Q1	996	KURT.	-1.20
5%	913	SKEW	0.00
MIN	892	SUM	460k



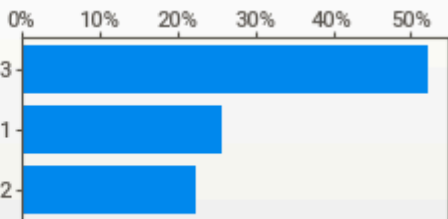
Survived

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 2 (<1%)



Pclass

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 3 (<1%)



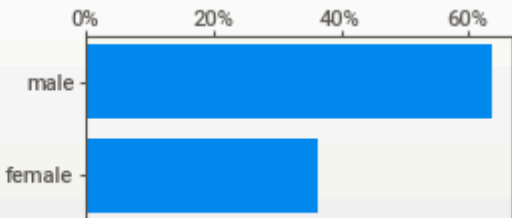
Name

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 418 (100%)

1	<1%	Kelly, Mr. James
1	<1%	Carr, Miss. Jeannie
1	<1%	Dennis, Mr. William
1	<1%	Rosblom, Miss. Salli Helena
1	<1%	Touma, Miss. Maria Youssef
1	<1%	Fleming, Miss. Honora
1	<1%	Peacock, Master. Alfred Edward
411	98%	(Other)

Sex

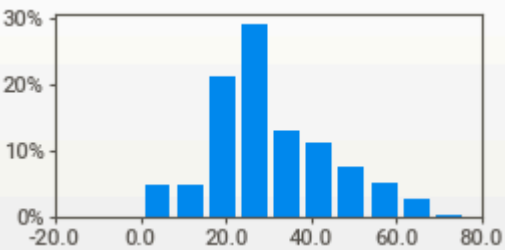
VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 2 (<1%)



Age

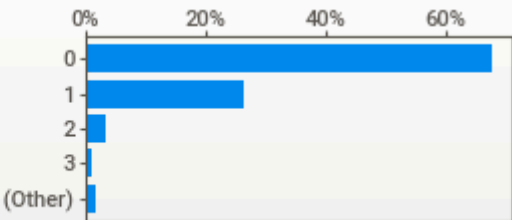
VALUES: 332 (79%)  
MISSING: 86 (21%)  
DISTINCT: 79 (19%)  
ZEROES: ---

MAX	76.0	RANGE	75.8
95%	57.0	IQR	18.0
Q3	39.0	STD	14.2
AVG	30.3	VAR	201
MEDIAN	27.0		
Q1	21.0	KURT.	0.084
5%	8.0	SKEW	0.457
MIN	0.2	SUM	10,050



SibSp

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 7 (2%)



Parch

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 8 (2%)



Ticket

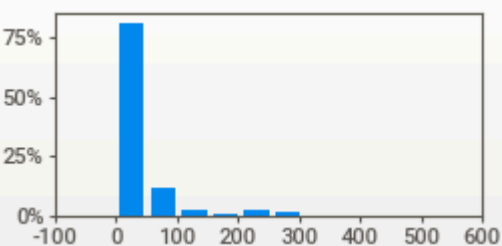
VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 363 (87%)

5	1%	PC 17608
4	<1%	CA. 2343
4	<1%	113503
3	<1%	PC 17483
3	<1%	220845
3	<1%	347077
3	<1%	SOTON/O.Q. 3101315
393	94%	(Other)

Fare

VALUES: 417 (>99%)  
MISSING: 1 (<1%)  
DISTINCT: 169 (40%)  
ZEROES: 2 (<1%)

MAX	512	RANGE	512
95%	152	IQR	23.6
Q3	32	STD	55.9
AVG	36	VAR	3,126
MEDIAN	14		
Q1	8	KURT.	17.9
5%	7	SKEW	3.69
MIN	0	SUM	14,857



Cabin

VALUES: 91 (22%)  
MISSING: 327 (78%)  
DISTINCT: 76 (18%)

3	3%	B57 B59 B63 B66
2	2%	C89
2	2%	C116
2	2%	C80
2	2%	C55 C57
2	2%	C101
2	2%	A34
76	84%	(Other)

Embarked

VALUES: 418 (100%)  
MISSING: ---  
DISTINCT: 3 (<1%)

