

# **10 Automated EDA** **Tools That Will** **Save You Hours** **Of (Tedious) Work**

**Avi Chawla**

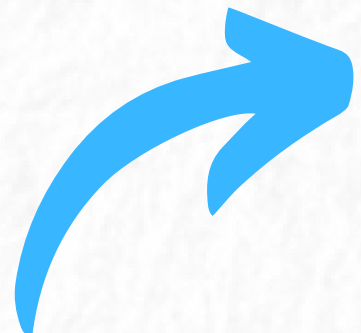


[avichawla.substack.com](http://avichawla.substack.com)

**EDA is a vital but  
time-consuming  
task in a data  
project.**



Here are **10 open-source  
tools** that generate  
an EDA report  
**in seconds.**



**1.**

**SweetViz**



In-depth EDA report  
in two lines  
of code.

Covers information about  
missing values, data  
statistics, etc.

Creates a variety of data  
visualizations.

Integrates with Jupyter  
Notebook.

**2.**

# Pandas-Profiling

| Overview                      | Variables  | Interactions | Correlations | Missing values | Sample | Duplicate rows |
|-------------------------------|------------|--------------|--------------|----------------|--------|----------------|
| Overview                      | Alerts (6) | Reproduction |              |                |        |                |
| Number of variables           | 5          |              | Numeric      | 4              |        |                |
| Number of observations        | 150        |              | Categorical  | 1              |        |                |
| Missing cells                 | 0          |              |              |                |        |                |
| Missing cells (%)             | 0.0%       |              |              |                |        |                |
| Duplicate rows                | 1          |              |              |                |        |                |
| Duplicate rows (%)            | 0.7%       |              |              |                |        |                |
| Total size in memory          | 6.0 KiB    |              |              |                |        |                |
| Average record size in memory | 40.9 B     |              |              |                |        |                |

- ➔ Generate a high-level **EDA report** of your data in no time.
- ➔ Covers info about **missing values, data statistics, correlation** etc.
- ➔ Produces **data alerts**.
- ➔ Plots data **feature interactions**.



**3.**

**DataPrep**

| Dataset Statistics         |                                | Dataset Insights  |                  |
|----------------------------|--------------------------------|---|------------------|
| Number of Variables        | 12                             | <code>PassengerId</code> is uniformly distributed               | Uniform          |
| Number of Rows             | 891                            | <code>Age</code> has 177 (19.87%) missing values                | Missing          |
| Missing Cells              | 866                            | <code>Cabin</code> has 687 (77.1%) missing values               | Missing          |
| Missing Cells (%)          | 8.1%                           | <code>Fare</code> is skewed                                     | Skewed           |
| Duplicate Rows             | 0                              | <code>Name</code> has a high cardinality: 891 distinct values   | High Cardinality |
| Duplicate Rows (%)         | 0.0%                           | <code>Ticket</code> has a high cardinality: 681 distinct values | High Cardinality |
| Total Size in Memory       | 315.0 KB                       | <code>Cabin</code> has a high cardinality: 147 distinct values  | High Cardinality |
| Average Row Size in Memory | 362.1 B                        | <code>Survived</code> has constant length 1                     | Constant Length  |
| Variable Types             | Numerical: 3<br>Categorical: 9 | <code>Pclass</code> has constant length 1                       | Constant Length  |
|                            |                                | <code>SibSp</code> has constant length 1                        | Constant Length  |

Supports **Pandas** and **Dask** DataFrames.

Interactive Visualizations.

**10x Faster** than Pandas based tools.

Covers info about **missing values, data statistics, correlation** etc.

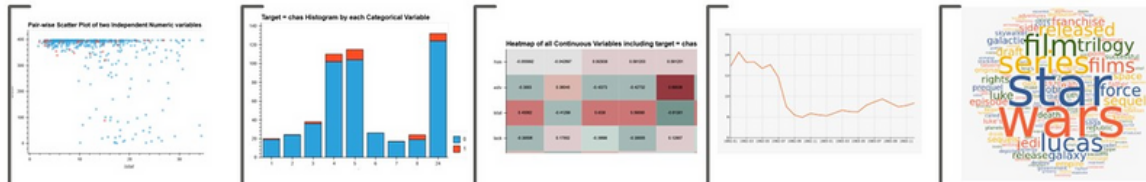
Plots data **feature interactions**.



**4.**

**AutoViz**

**YOU CAN NOW VISUALIZE ALL YOUR CHARTS IN EITHER A BROWSER OR A JUPYTER NOTEBOOK – ANY DATA SET, ANY SIZE, ANY TYPE**



Scatter Plots

Bar Charts

Heat Maps

Time Series

Word Cloud

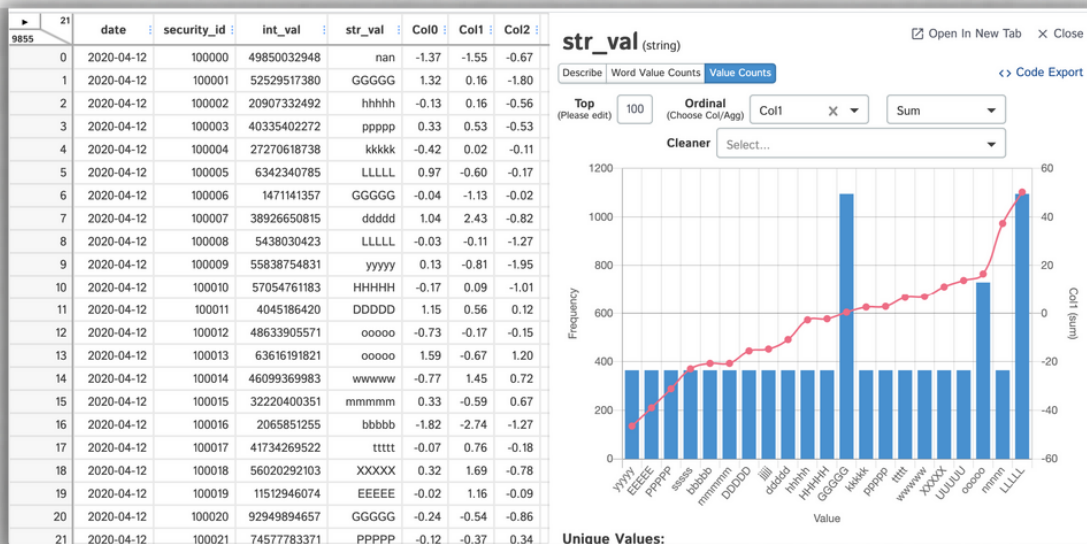
Supports **CSV, TXT**, and **JSON**.

Interactive **Boken** charts.

Covers info about **missing values, data statistics, correlation** etc.

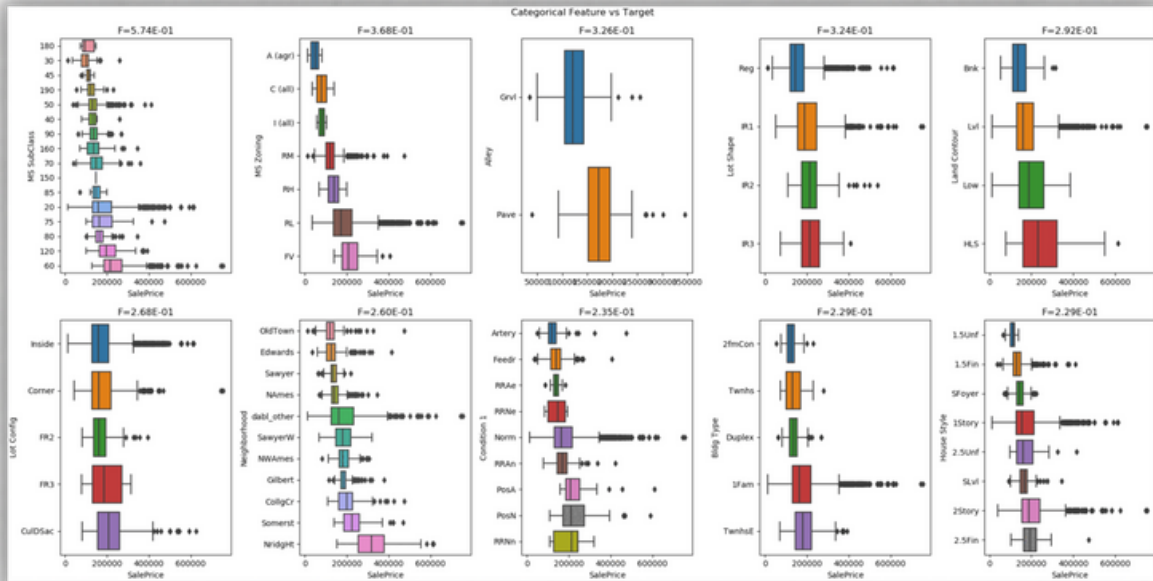
Presents **data cleaning** suggestions.

# **5.** **D-Tale**



- Runs common Pandas operation with **no-code**.
- Exports code** of analysis.
- Integrates with **Jupyter Notebook**.
- Covers info about **missing values, data statistics, correlation** etc.
- Highlights **duplicates, outliers**, etc.

**6.**  
**dabl**



Primarily provides  
**visualizations.**

Covers **wide range**  
of plots.

# Target distribution.

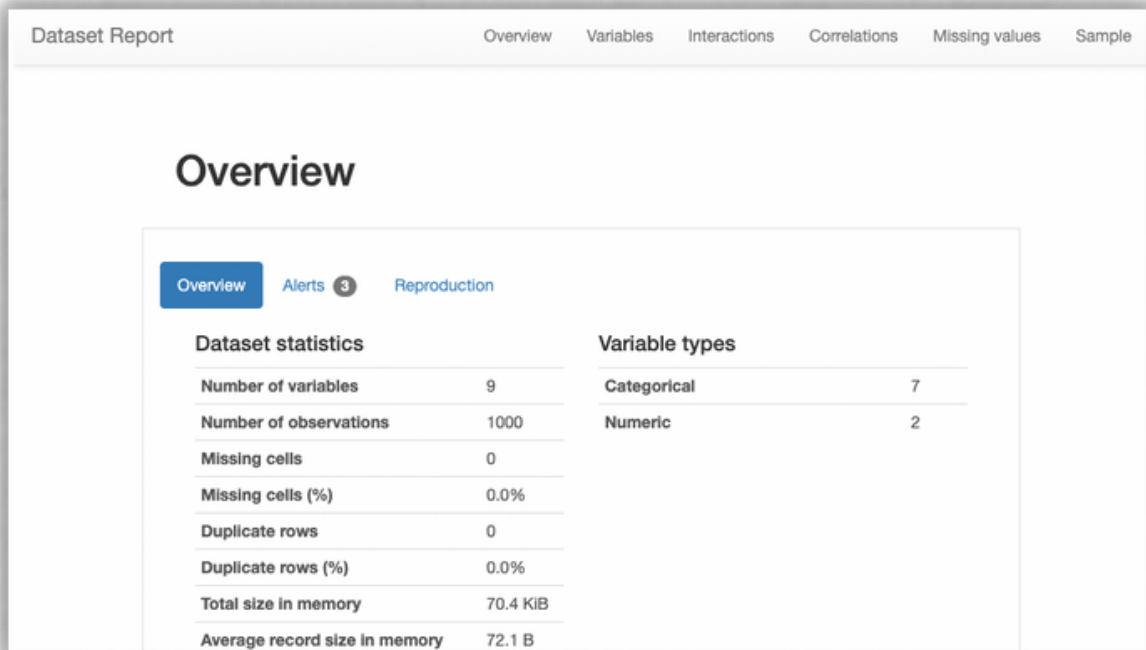
## Scatter pair plots.

# Histograms.



**7.**

**QuickDA**



The screenshot shows the 'Dataset Report' interface with a navigation bar at the top containing 'Overview', 'Variables', 'Interactions', 'Correlations', 'Missing values', and 'Sample'. The 'Overview' tab is selected, showing a sub-navigation bar with 'Overview', 'Alerts 3', and 'Reproduction'. The main content area is divided into two tables: 'Dataset statistics' and 'Variable types'.

| Dataset statistics            |          | Variable types |   |
|-------------------------------|----------|----------------|---|
| Number of variables           | 9        | Categorical    | 7 |
| Number of observations        | 1000     | Numeric        | 2 |
| Missing cells                 | 0        |                |   |
| Missing cells (%)             | 0.0%     |                |   |
| Duplicate rows                | 0        |                |   |
| Duplicate rows (%)            | 0.0%     |                |   |
| Total size in memory          | 70.4 KiB |                |   |
| Average record size in memory | 72.1 B   |                |   |

- Get overview report of dataset.
- Covers info about **missing values, data statistics, correlation** etc.
- Produces **data alerts**.
- Plots data **feature interactions**.

**8.**

**Datatile**

|              | age     | job         | marital     | education   | default | balance | housing | loan  | contact     | day     | month       | duration | campaign | pdays   | previous | poutcome    | y     |
|--------------|---------|-------------|-------------|-------------|---------|---------|---------|-------|-------------|---------|-------------|----------|----------|---------|----------|-------------|-------|
| count        | 45211   | NaN         | NaN         | NaN         | NaN     | 45211   | NaN     | NaN   | NaN         | 45211   | NaN         | 45211    | 45211    | 45211   | 45211    | NaN         | NaN   |
| mean         | 40.9362 | NaN         | NaN         | NaN         | NaN     | 1362.27 | NaN     | NaN   | NaN         | 15.8064 | NaN         | 258.163  | 2.76384  | 40.1976 | 0.580323 | NaN         | NaN   |
| std          | 10.6188 | NaN         | NaN         | NaN         | NaN     | 3044.77 | NaN     | NaN   | NaN         | 8.32248 | NaN         | 257.528  | 3.09802  | 100.129 | 2.30344  | NaN         | NaN   |
| min          | 18      | NaN         | NaN         | NaN         | NaN     | -8019   | NaN     | NaN   | NaN         | 1       | NaN         | 0        | 1        | -1      | 0        | NaN         | NaN   |
| 25%          | 33      | NaN         | NaN         | NaN         | NaN     | 72      | NaN     | NaN   | NaN         | 8       | NaN         | 103      | 1        | -1      | 0        | NaN         | NaN   |
| 50%          | 39      | NaN         | NaN         | NaN         | NaN     | 448     | NaN     | NaN   | NaN         | 16      | NaN         | 180      | 2        | -1      | 0        | NaN         | NaN   |
| 75%          | 48      | NaN         | NaN         | NaN         | NaN     | 1428    | NaN     | NaN   | NaN         | 21      | NaN         | 319      | 3        | -1      | 0        | NaN         | NaN   |
| max          | 95      | NaN         | NaN         | NaN         | NaN     | 102127  | NaN     | NaN   | NaN         | 31      | NaN         | 4918     | 63       | 871     | 275      | NaN         | NaN   |
| counts       | 45211   | 45211       | 45211       | 45211       | 45211   | 45211   | 45211   | 45211 | 45211       | 45211   | 45211       | 45211    | 45211    | 45211   | 45211    | 45211       | 45211 |
| uniques      | 77      | 12          | 3           | 4           | 2       | 7168    | 2       | 2     | 3           | 31      | 12          | 1573     | 48       | 559     | 41       | 4           | 2     |
| missing      | 0       | 0           | 0           | 0           | 0       | 0       | 0       | 0     | 0           | 0       | 0           | 0        | 0        | 0       | 0        | 0           | 0     |
| missing_perc | 0%      | 0%          | 0%          | 0%          | 0%      | 0%      | 0%      | 0%    | 0%          | 0%      | 0%          | 0%       | 0%       | 0%      | 0%       | 0%          | 0%    |
| types        | numeric | categorical | categorical | categorical | bool    | numeric | bool    | bool  | categorical | numeric | categorical | numeric  | numeric  | numeric | numeric  | categorical | bool  |

Extends Pandas'  
**describe()**.

Mostly **statistical  
information**.

Provides **column stats**.

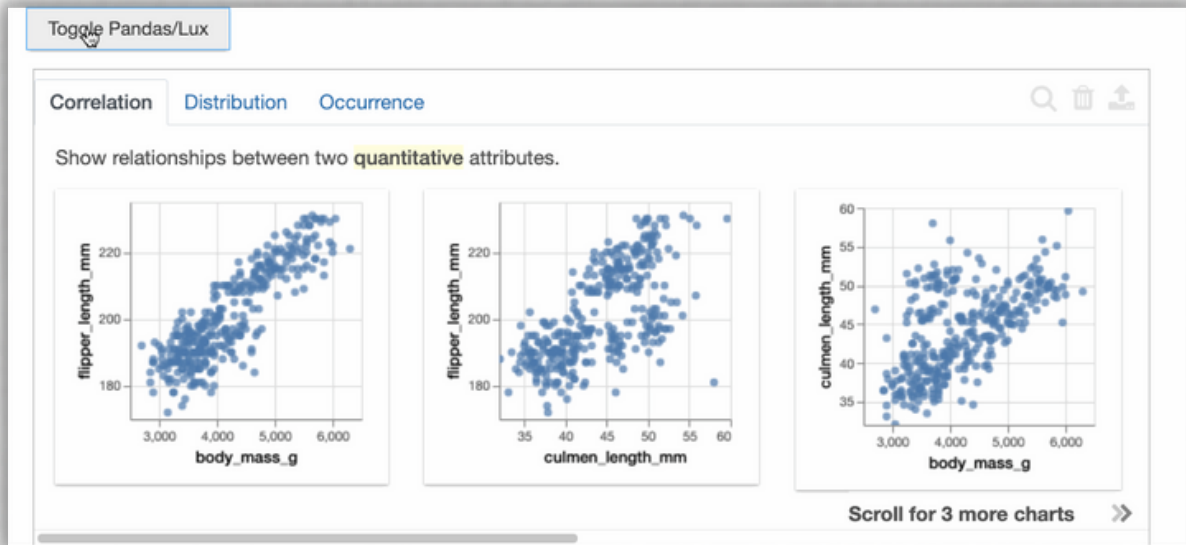
Counts, missing, etc.

Column datatype.

Column type count.

**9.**

**Lux**

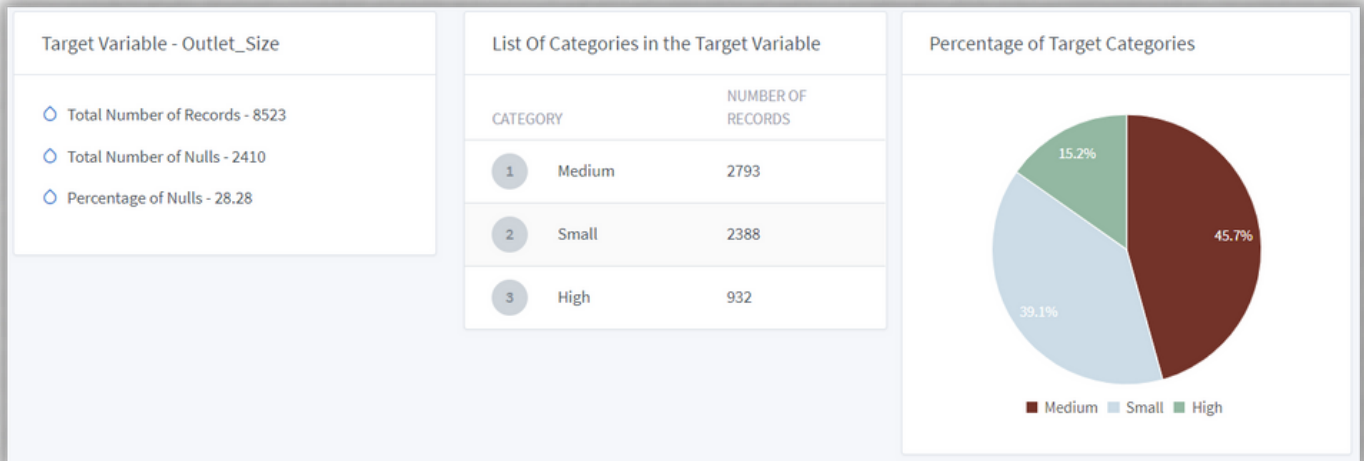


- Integrates with **Jupyter Notebook**.
- Provides **visualization recommendations**.
- Supports EDA on a **subset of columns**.
- **Exports code** of analysis.



# 10.

# ExploriPy



→ Covers info about **missing values, data statistics, correlation** etc.

→ Performs **statistical testing**.

→ Column type-wise **distribution**.

→ Continuous

→ Categorical

# Hope that helped.



Checkout my daily newsletter to learn something new about **Python** and **Data Science** everyday.

 [avichawla.substack.com](https://avichawla.substack.com)

Connect with me on **LinkedIn**.

<https://www.linkedin.com/in/avi-chawla>

