EDA Toolkit: A Lightweight Python Library for Exploratory Data Analysis and Reporting

30 July 2025

Summary

EDA Toolkit is a lightweight Python package for conducting and visualizing exploratory data analysis (EDA). It provides flexible plotting tools, profiling summaries, and exportable Excel reports tailored for both academic and industry workflows. Designed to be highly customizable and easily integrated into notebooks or pipelines, EDA Toolkit helps users rapidly understand and communicate data characteristics.

Statement of need

Exploratory Data Analysis (EDA) is a critical step in any data science project. It helps researchers understand the structure of a dataset, detect data quality issues, identify meaningful patterns, and shape the direction of analysis. While tools such as `pandas_profiling` and `sweetviz` provide automated reports, they often lack the flexibility, transparency, and formatting needed in professional or academic settings.

EDA Toolkit addresses these limitations by offering a modular and customizable set of tools designed for clarity, reproducibility, and high-quality presentation. It supports both academic research and applied data science use cases, with an emphasis on producing outputs that are publication-ready and easy to interpret.

Several examples throughout the documentation and figures in this paper are based on the Adult Income dataset from the UCI Machine Learning Repository [@uci_adult; @kohavi1996census]. This real-world tabular dataset offers a rich foundation for testing demographic segmentation, group comparisons, and reporting tools, particularly in contexts involving income classification and social variables. It serves as a practical benchmark for demonstrating the capabilities of functions such as `generate_table1()` and outlier visualization utilities within the EDA Toolkit.

Key Features

Table 1 Generation

The `generate_table1()` function allows users to produce clean, formatted descriptive tables often used in clinical and academic research. The output includes summaries by group and supports filtering by data type, making it easier to communicate sample characteristics without relying on external tools like Excel.

Table 1: Group-wise descriptive statistics using the `generate_table1()` function on the UCI Adult Income dataset.

Variable	Count	Proportion (%)	<=50K (n = 37155)	>50K (n = 11687)
age_group	48842	100.00	37155	11687
age_group = 18-29	13920	28.50	13174 (35.46%)	746 (6.38%)
age_group = 30-39	12929	26.47	9468 (25.48%)	3461 (29.61%)
age_group = 40-49	10724	21.96	6738 (18.13%)	3986 (34.11%)
age_group = 50-59	6619	13.55	4110 (11.06%)	2509 (21.47%)
age_group = 60-69	3054	6.25	2245 (6.04%)	809 (6.92%)
age_group = 70-79	815	1.67	668 (1.80%)	147 (1.26%)
age_group = < 18	595	1.22	595 (1.60%)	0 (0.00%)
age_group = 80-89	131	0.27	115 (0.31%)	16 (0.14%)
age_group = 90-99	55	0.11	42 (0.11%)	13 (0.11%)
age_group = 100 +	0	0.00	0 (0.00%)	0 (0.00%)
marital-status	48842	100.00	37155	11687
marital-status = Married-civ-spouse	22379	45.82	12395 (33.36%)	9984 (85.43%)
marital-status = Never-married	16117	33.00	15384 (41.40%)	733 (6.27%)

Variable	Count	Proportion (%)	<=50K (n = 37155)	>50K (n = 11687)
marital-status = Divorced	6633	13.58	5962 (16.05%)	671 (5.74%)
marital-status = Separated	1530	3.13	1431 (3.85%)	99 (0.85%)
marital-status = Widowed	1518	3.11	1390 (3.74%)	128 (1.10%)
marital-status = Married-spouse-absent	628	1.29	570 (1.53%)	58 (0.50%)
marital-status = Married-AF-spouse	37	0.08	23 (0.06%)	14 (0.12%)

Outlier and anomaly detection support

The library includes functions to identify and visualize outliers based on distributional thresholds or robust statistics. This capability supports detecting data quality issues early, understanding variable spreads, and guiding preprocessing decisions.