

Task: Build a Smart EDA & Statistics Python Library for AI and Data Science

In real-world data science and AI projects, datasets are rarely clean or ready for modeling. Data often contains missing values, outliers, irrelevant features, and unclear relationships with the target variable. Data scientists use mathematics, statistics, and exploratory data analysis (EDA) to understand data before building any AI or machine learning model.

Your task is to design and implement a reusable Python library that can perform statistical analysis and exploratory data analysis on any dataset. This library should help users quickly understand data quality, feature importance, and statistical relationships, and prepare datasets for AI/ML models.

What You Must Build

Create a Python library (not a single script) that contains the following functionality:

- Implement reusable Python functions to compute mean, median, and mode
- Implement functions to compute variance and standard deviation
- Implement z-score and basic probability calculations
- Implement normal distribution functions (PDF or CDF)
- Implement hypothesis testing functions: z-test, t-test, and chi-square test
- Ensure each function works on lists, NumPy arrays, or Pandas Series

- Detect missing values and calculate their percentage
- Handle missing values using mean, median, mode, or row deletion
- Detect duplicate rows

- Detect outliers using the Z-score or IQR method
- Handle outliers by removing or capping them

- Analyze the distribution of the target variable (mean, standard deviation, skewness)
- Normalize or transform the target variable if needed

- Analyze numerical features versus the target variable (correlation and strength of relationship)
- Analyze categorical features versus the target variable (group statistics and chi-square test)
- Identify which features have meaningful relationships with the target variable

- Rank and return the top 10 most important features with respect to the target variable using statistical measures

- Create a function that takes a dataset and target column as input
- Perform all EDA steps automatically
- Return a cleaned dataset, a summary of insights, and the most important features

Example:

```
clean_df, insights = full_eda(df, target="target_column")
```

You may use any real dataset (student performance, sales, medical, customer data, etc.). Your library must work for any dataset, not a specific one.

Final Deliverables

- A modular Python library with well-structured folders and files
- Clear and well-named functions
- Example usage of the library on a real dataset
- A short explanation of insights discovered from EDA