Data Quality Homework Report

Author: Siddhesh Maheshwari MDS202347

# Objective

To evaluate and ensure the quality of datasets using three different tools—PyDeequ, YData Profiling, and Great Expectations. The objective is to explore how each tool helps in understanding the dataset, identifying potential issues, and maintaining high standards of data quality.

# 1. Dataset Overview

The dataset used across all three tools (YData Profiling, PyDeequ, and Great Expectations) consists of 48,842 entries and 15 columns, including a mix of numerical and categorical variables. It resembles the UCI Adult dataset, commonly used for income prediction tasks.

**Data Snapshot**

* Rows: 48,842
* Columns: 15
* Memory Usage: ~4.8 MB

**Column Types:**

* Numerical (int32): age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week
* Categorical (object): workclass, education, marital-status, occupation, relationship, race, sex, native-country, income

Highlights from the Statistical Summary:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Metric | age | fnlwgt | education-num | capital-gain | capital-loss | hours-per-week | | --- | --- | --- | --- | --- | --- | --- | | Mean | 38.64 | 189,664.1 | 10.08 | 1079.1 | 87.5 | 40.42 | | Std Dev | 13.71 | 105,604.0 | 2.57 | 7452.0 | 403.0 | 12.39 | | Min–Max | 17–90 | 12,285–1.49M | 1–16 | 0–99,999 | 0–4,356 | 1–99 | |

# 2. Tool 1: YData Profiling

## Introduction

YData Profiling (formerly Pandas Profiling) is an open-source tool that provides automated EDA (Exploratory Data Analysis). It generates detailed reports summarizing the structure, quality, and relationships within a dataset.

## Key Highlights

- Used ProfileReport from ydata\_profiling  
- Generated a comprehensive report with statistics on:  
 - Missing Values  
 - Variable Types (categorical, numerical, boolean)  
 - Correlations (Pearson, Spearman, Kendall)  
 - Value distributions  
 - Duplicate rows

## Observations

- The tool flagged some missing data.  
- Visualizations revealed skewness in some numerical distributions.  
- Detected duplicate rows which may impact downstream tasks.

## Strengths

- No-code, one-line generation of rich reports  
- Interactive and easy to interpret  
- Great for initial exploration

# 3. Tool 2: PyDeequ

## Introduction

PyDeequ is a Python wrapper for Deequ, a library built on Apache Spark for defining 'unit tests' for data. It helps build scalable, automated data quality checks.

## Key Highlights from the Notebook

- Integrated PyDeequ with PySpark to handle large-scale data  
- Defined analyzers and checks such as:  
 - Uniqueness of columns  
 - Completeness (non-null values)  
 - Minimum and maximum value constraints  
- Results stored in a VerificationResult object and displayed as a DataFrame

## Sample Rules Implemented

- "id" should be unique  
- "age" should not be negative  
- "salary" should not be null  
- Distribution constraints on specific columns

## Observations

- One column violated the uniqueness rule  
- Some null values in columns flagged for completeness  
- All other checks passed successfully

# 4. Tool 3: Great Expectations

## Introduction

Great Expectations is a powerful tool for data quality validation that focuses on building trust in data through explicit expectations and validation suites.

## Key Highlights

- Created expectation suites from the dataset  
- Included expectations such as:  
 - expect\_column\_values\_to\_not\_be\_null  
 - expect\_column\_mean\_to\_be\_between  
 - expect\_column\_values\_to\_be\_unique  
- Rendered interactive Data Docs for validation feedback  
- Demonstrated CLI as well as programmatic interface

## Expectations Setup

- Automatically profiled dataset to suggest initial expectations  
- Manually added checks for value ranges and types

## Observations

- Identified columns not meeting null expectations  
- Some columns were well within expected distributions  
- Validated data against defined constraints and flagged discrepancies

# 5. Key Findings

- Common Issues Identified:  
 - Null values in several columns  
 - Duplicate entries  
 - Columns with skewed distributions  
  
- Best Tool for EDA: YData Profiling  
- Best Tool for Scalable Validation: PyDeequ  
- Best Tool for Production Pipelines: Great Expectations

# 6. Conclusion

The combination of these tools offers a powerful suite for data quality management:  
- YData Profiling shines in quick data understanding.  
- PyDeequ ensures robust, scalable data validation for big data.  
- Great Expectations bridges quality assurance with documentation and collaboration.