# **Data Quality HW1 Report**

**Author**: Aniket Tiwari  
 **Dataset Title:** Adult Dataset

**Ydata\_Profiling Report**

### **1. Dataset Overview**

* **Memory Usage:** ~27.7 MB
* **Average Record Size:** 567.81 bytes
* **Duplicate Records:** 28 (0.06%)
* **Missing Cells:** 2,203 (0.30%)
* **Variables with Missing Values:** 3
* **Data Types:**
  + *Numeric:* 6
  + *Categorical:* 9

### **2. Missing Data Analysis**

* **Percentage of Missing Data:** 0.30%
* **Variables with Missing Data:**
  + workclass: 963 missing values (~1.97%)
  + (Other variables with missing values are not explicitly listed in the snippet)

### **3. Notable Variable Profiles**

#### **a. Age**

* **Type:** Numeric
* **Range:** 17 – 90 years
* **Mean:** 38.64
* **Median:** 37
* **Standard Deviation:** 13.71
* **Skewness:** 0.56 (moderately right-skewed)
* **Kurtosis:** -0.18 (slightly platykurtic)
* **Distinct Values:** 74
* **Missing Values:** 0
* **Top Frequent Ages:** 36 (1,348), 35 (1,337), 33 (1,335)

#### **b. Workclass**

* **Type:** Categorical
* **Distinct Categories:** 9
* **Top Category:** *Private* (69.4%)
* **Missing Values:** 963 (~2%)
* **Special Category (? for missing/unknown):** Present
* **Imbalance Score:** 0.49 (high dominance of *Private*)

#### **c. fnlwgt**

* **Type:** Numeric
* **Distinct Values:** 28,523 (~58.4%)
* **No missing values detected**
* **Not normally distributed** — most values appear only a few times

### **4. Quality & Distributional Insights**

* **Data Quality:** Generally good with minimal missing values and low duplication
* **Distribution Issues:**
  + The variable workclass contains the "?" category, which should be treated as missing
  + A few numeric columns (like age) exhibit slight skewness, which could impact model assumptions requiring normality
* **Cardinality Issues:**
  + Variables like fnlwgt have high cardinality but may not carry predictive power without context

### **5. Recommendations**

1. **Handle Missing Values:**
   * Replace "?" entries in categorical fields with NaN and impute or analyze separately.
   * Consider imputation strategies or dropping variables with consistently missing information, based on downstream tasks.
2. **Normalize/Transform Skewed Distributions:**
   * Use log or Box-Cox transformations if needed for models sensitive to normality.
3. **Investigate High Cardinality Variables:**
   * Variables such as fnlwgt may require binning or removal depending on model use.
4. **Check for Data Consistency:**
   * Validate categorical categories and remove unlikely values (e.g., check that all values in workclass are valid and interpretable).

**PyDeequ Data Quality Report**

### **Note on Dataset Source**

**Important:** The dataset used for this PyDeequ analysis was **downloaded manually** and contains approximately **32,561 rows**. This differs from the datasets used in **YData Profiling** and **Great Expectations** analysis, which were based on data **fetched dynamically via code** from the **UCI Machine Learning Repository**, potentially leading to a slightly different number of rows and values due to source variations.

### **Summary of Constraint Checks**

| **Check** | **Status** | **Constraint** | **Result** |
| --- | --- | --- | --- |
| Size Check | ✅ Success | Total number of records in the dataset | 32,561 rows |
| Completeness Check on age | ✅ Success | No missing values in the age column | Passed |
| Completeness Check on income | ✅ Success | No missing values in the income column | Passed |
| Completeness Check on workclass | ⚠️ Warning | ~94.36% completeness in the workclass column | Partial Completeness |
| Uniqueness Check on fnlwgt | ❌ Failure | Only ~47.08% unique values found in fnlwgt (21,648 unique values out of 32,561 rows) | Failed |
| Non-Negative Check on hours\_per\_week | ✅ Success | All values in hours\_per\_week are non-negative | Passed |
| Domain Check on income | ✅ Success | Values only include >50K and <=50K | Passed |
| Domain Check on sex | ✅ Success | Values only include Male and Female | Passed |
| Domain Check on marital\_status | ✅ Success | Values conform to the expected set of marital status categories | Passed |
| Domain Check on workclass | ✅ Success | Values conform to the expected set of workclass categories | Passed |

### **Summary of Analyzers**

The following statistics and distinct value counts were derived using PyDeequ's built-in analyzers:

| **Metric** | **Column** | **Value** |
| --- | --- | --- |
| Total Rows | \* | 32,561 |
| Mean Age | age | 38.58 |
| Min Age | age | 17 |
| Max Age | age | 90 |
| Distinct Values (Income) | income | 2 |
| Distinct Values (Sex) | sex | 2 |
| Distinct Values (Race) | race | 5 |
| Distinct Values (Education) | education | 15 |
| Distinct Values (Occupation) | occupation | 14 |
| Distinct Values (Workclass) | workclass | 8 |
| Distinct Values (Country) | native\_country | 39 |
| Distinct Values (Marital Status) | marital\_status | 7 |
| Distinct Values (fnlwgt) | fnlwgt | 21,648 |
| Completeness (workclass) |  | 94.36% |

### 

### **Noteworthy Observation**

* The fnlwgt column **did not meet the uniqueness constraint**. This is **expected**, as the column serves as a **weighting factor** rather than a unique identifier.  
   Only **47.08%** of values were unique, aligning with its intended statistical use.

**Great Expectations Validation Report**

**Summary of Validation Results**

* **Total Expectations Evaluated:** 9
* **Successful Expectations:** 9
* **Failed Expectations:** 0
* **Validation Success Rate:** 100%

**Details of Expectations and Results**

1. **Age Column**
   * All values were non-null.
   * All values were within the expected range of **17 to 90**.
2. **Education Column**
   * No null values detected.
   * All values had string lengths between **1 and 20** characters.
3. **Income Column**
   * All values were within the defined set: **"<=50K"** and **">50K"**.
4. **Sex Column**
   * All values matched the allowed categories: **"Male"** and **"Female"**.
5. **Workclass Column**
   * No null values were present.
6. **Occupation Column**
   * No null values were present.
7. **Native-Country Column**
   * No null values were present.