Project Two: Executive Summary Report

Rolesshania Jackson

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Sharmila Kundami

October 7, 2025

# Data Types

| **Variable Name** | **Data Types Note** |
| --- | --- |
| Marital\_Status | String; confirmed as categorical; standardized values to match approved categories(“Married”, “NotMarried”) |
| Citizen\_Desc | String; confirmed as categorical; standardized values and filled missing entries with mode |
| Age | Integer; confirmed as numeric; cleaned to remove negative values and fill missing entries with median |
| Days\_Employed | Float; confirmed as numeric; placeholder values -999 and negatives replaced with median |

# Anomalies

Specify the variable and anomaly or issue in the data. Provide a plan for resolution.

| **Variable Name** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| Age | Missing Values | Filled missing numeric using the median |
| Days\_Employed | Incorrect Placeholders | Filled missing numeric using the median. Replace placeholder and negative values witih valid medians |
| Citizen\_Desc | Non-standard | Validated all corrections using .describe(), .unique() |
| Marital\_Status | “NotMarried”, “Married” | Standardized values using .replace() to ensure consistency (attempt) |

# Transforms

| **Variable Name** | **Required Transformation** |
| --- | --- |
| Employee Name | Split from **Employee\_Name** using **.str.split(“,”)**  and **str[1]** |
| DOB | Convert to **datetime** using **pd.to\_determine()** |

# Summary

As part of our data quality initiative, I conducted a full profiling of the new firms dataset to prepart it for merging with our compamy’s master database. This process focused on identifying and resolving issues across four key variables:

* Age
* Days\_Employed
* Citizen\_Desc
* Marital\_Status

I was able to identify **Data Quality Issues** such as missing values in **Age**, **Days\_Employed,** and **Citizen\_Desc.** There were also incorrect placeholders involving Days\_Employed with negative value of -999 and negative values for Age containing invalid entries below zero. Additionally, Citizen\_Desc and Marital\_Status are non-standard categories and was inconsistent labels that did not match approved categories such as U.S Citizen or Married/ Not Married.

The step towards solution is filling in the missing mumeroc values using the median and categorical values using the mode, replacing the placeholder and negative values with statistically valud replacements, and standardize categorical fields using filters and .replace() to ensure the consistency. The cleaned dataset now aligns with out master schema and almost ready for integration. Most of the anomalies were addressed with data set validated against the provided data dictionary. Ensuring, mostly, that this remains smootha and accurate merge with minimal risk of data integrity issues.

## References

Smith, J. (2022). *Data quality completeness in enterprise systems*. *Journal of Data Management*, *18*(3), 45–59. <https://research.ebsco.com/c/ix3dnl/search/details/2undny6dsf>

Monte Carlo. (2022, March 15). *What is data completeness?* Monte Carlo Data. <https://www.montecarlodata.com/blog-what-is-data-completeness/>