Project Two: Executive Summary Report

Nicole Hutto

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Professor Jennifer Johnson

September 29, 2025

# Data Types

The first two rows provide example answers. You will need to assess all variables in the data set. You may need to add or delete rows in the tables.

| **Variable Name** | **Data Types Note** |
| --- | --- |
| *Example: State* | *Example: Will use, requires profiling based on analysis* |
| *Example: Pay Rate* | *Example: Won’t Use* |
| |  | | --- | | Employee Number |  |  | | --- | |  | | |  | | --- | | Integer — Unique identifier; keep as-is. |  |  | | --- | |  | |
| Employee Name | String — Must be split into **First Name** and **Last Name** for merging. |
| Date of Birth (DOB) | Date — Standardize to YYYY-MM-DD format. |
| Marital Status | String → Integer Code — Recode values into numeric codes (e.g., 1 = Married, 0 = Single, 2 = N/A). |

# Anomalies

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

| **Variable Name** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| Employee Number | **Duplicate values or missing identifiers in some rows** | **Remove duplicates and require unique Employee Number for each record.** |
| Employee Name | Missing last names in some records | Delete rows with incomplete names if data cannot be recovered. |
| |  |  |  | | --- | --- | --- | | |  | | --- | | Date of Birth (DOB) |  |  | | --- | |  | |  |  | | --- | |  | | Inconsistent formats (MM/DD/YYYY vs. DD-MM-YY) or invalid dates (e.g., future dates, 01/01/1900) | Standardize to YYYY-MM-DD format; remove or flag invalid entries. |
| Marital Status | [Insert text.] Free-text values (e.g., “Married,” “single,” “NA”) instead of codes | Recode values into standardized integer codes. |

# Transforms

The variable names specified below are from the data set of the new firm being acquired. Specify the transformation required to merge their data with your firm’s database.

| **Variable Name** | **Required Transformation** |
| --- | --- |
| **Employee Number** | **Ensure uniqueness and validate as integer key; remove duplicates before merging.** |
| Employee Name | Split into **First Name** and **Last Name** fields to match company database schema. |
| DOB | Standardize all values into YYYY-MM-DD format. |
| Marital Status | Recode text values (“Married,” “Single,” “N/A”) into integer codes (e.g., 1 = Married, 0 = Single, 2 = N/A). |

# Summary

During the data profiling step we saw several quality problems that could affect how the new firm’s data joins our HR master file. First, some employee rows miss last names. Dates of birth and hire show mixed formats, like 12/31/2020 or 2020‑31‑12. Categorical fields such as marital status, race, and sex are entered as free‑text categories which makes analysis tricky. Also, zip codes appear invalid in many entries.

Our cleaning plan tries to fix these issues. We will standardize all dates into the YYYY‑MM‑DD form and replace free‑text categories with numeric codes. Records missing critical fields may be dropped, and duplicate rows will be removed. The employee name column will be broken into first‑name and last‑name columns, because a single field can cause matching errors later. Employee numbers will be checked for uniqueness, serving as the main key for joins.

These steps seem to follow what big‑data best practices recommend: accuracy, consistency, and reliability. Still, some might argue that dropping incomplete rows can lose useful information, so a more cautious approach could be to flag them instead.

Overall, cleaning the data should let us merge the new set with our existing database more smoothly. Errors should shrink, and stakeholders will get data that is more trustworthy for future reports and analysis. A brief appendix will list each correction, allowing supervisors to verify the clean data.

## References

**References**

IBM. (n.d.-a). *What is big data?* IBM Think. Retrieved October 2, 2025, from [https://www.ibm.com/think/topics/big-data](https://www.ibm.com/think/topics/big-data?utm_source=chatgpt.com)

IBM. (n.d.-b). *What is big data analytics?* IBM Think. Retrieved October 2, 2025, from [https://www.ibm.com/think/topics/big-data-analytics](https://www.ibm.com/think/topics/big-data-analytics?utm_source=chatgpt.com)