Project Two: Executive Summary Report

Creston Getz

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Ray Rabago

January 31, 2026

Data Types

Variable Name	Data Types Note
Age	Int
Citizen_Desc	VarChar
Days_Employed	Int
DOB	Date: *Dictionary* says date but in dataset this is an object (uCertify, 2026)
Employee_Name	VarChar
Hispanic_Latino	Varchar. *Should be changed to Category or Boolean*
Marital_Status	VarChar. *Should be changed to Boolean*
Race_Desc	VarChar

Anomalies

Variable Name	Description of Anomaly	Plan for Resolution
Age	Missing Values	Only one row with missing value. Verify that DOB is correct to determine age.
Age	Incorrect Values	Only one incorrect row. Verify that DOB is correct to determine age.

Variable Name	Description of Anomaly	Plan for Resolution
Age	Outlier Values	Only one incorrect row. Verify that DOB is correct to determine age.
Marital Status	Missing Values	4 rows with missing values. Verify with employees of marital status.
Marital Status	Incorrect Values	Only one row. Verify with employee of their marital status. Change status from female to Married or not.
Citizen_Desc	Missing Values	Only one row. Verify with employee of citizenship.
Citizen_Desc	Incorrect Values	Only one row. Male is included, verify with employee of true citizenship and remove wrong entry.
Days_Employed	Missing Values	Only one row. Verify with employee and cross check other records to determine days of employment.

Variable Name	Description of Anomaly	Plan for Resolution
Days_Employed	Incorrect Values	One row is negative. Verify with employee and cross check with other records to determine length of employment.
Days_Employed	Outliers	One row is an extreme outlier of 52 thousand days (>100 years). Verify with this employee and cross check to verify their true length of employment. 5 other rows have a length of employment around 8-10 years. Verify that this is less than the existence of the company.
DOB	Incorrect Values	One row has an incorrect DOB and age that does not line up with the rest of the data set. There are also DOB's that are outside the normal range. Verify with

Variable Name	Description of Anomaly	Plan for Resolution
		employee of their true age and DOB.
Employee_Name	Incorrect Values	The name column currently holds the first and last name separated by commas. This should be split into two columns first_name and last_name.
Employee_Name	Missing Values	Some of the employee names are missing their last name. One entry is missing a comma and only has the first name. This should be reviewed and cleaned for every entry.
Hispanic_Latino	Incorrect / Inconsistent values	Unique values are upper and lower case ‘yes’ and ‘no’. This should be changed to a boolean value.
Race_Desc	Incorrect / Inconsistent values	Hispanic is included as a race despite a separate

Variable Name	Description of Anomaly	Plan for Resolution
		column. Another unique value is two or more races which could be separated for more detail.

Transforms

Variable Name	Required Transformation
Employee Name	In our companies database the is separated into two columns. We will have to separate the new firms Employee Name column into two separate columns and address the missing values above. After we separate the column, we should verify it is still a var char.
DOB	The new firm's dataset uses a month/day/year format. We should verify our database uses the same and if not change the new firm dataset to match our format. We should also address the errors we mentioned above. We also need to change the data type of this column as mentioned.

Summary

Ensuring the new firm's data is high quality is a critical step before adding it to our existing dataset. To review the data, we used a Python tool called Pandas. We first verified the

data types of each column to ensure the overall structure was correct. During this review, we identified several changes that should be made. For example, the date of birth column was expected to be stored as a date but was not, despite what the firm's documentation stated (uCertify, 2026). We also found two columns; Hispanic/Latino and marital status, that should be stored as Boolean(yes/no) values instead of text.

After reviewing the data structure, we analyzed the actual content to look for errors, inconsistencies, and anomalies. Several issues were identified that need to be addressed to ensure a successful data merge. Some of these will require follow-up with the new firm's data owners and employees to correct (Herzberg, 2021). The employee's name column is inconsistent, with one record containing only the first name missing a commas. The date of birth column was stored as text instead of a standard date format, which could cause errors during analysis and must be corrected to align with our dataset. The age column also contained missing values, incorrect entries, and extreme outliers. In some cases, employee ages did not match their listed dates of birth. For example, one employee was recorded as 139 years old and another as -39 years old. In another case, the date of birth appeared valid, but the age did not align with the current year.

Similar issues were found in the days employed column. One employee was listed as employed for over 52,000 days, which equals roughly 142 years. Anomalies such as these can skew the accuracy of a workplace analysis and may lead to incorrect employee benefits/eligibility for programs (Roundy, 2024). These columns will need to be cleaned, and employee details such as date of birth and age will need to be verified with both the employees and the new firm's data owners (Herzberg, 2021). We also identified consistency issues in the Hispanic/Latino and marital status columns. These values were stored as mixed-case text instead

of a standard yes/no format. While these columns are not required for merging the datasets, correcting them would improve overall data quality and consistency (Roundy, 2024).

Addressing these issues is essential to maintaining our data quality and ensuring a smooth merger. Since we plan to join the datasets using employee name and date of birth, it is important that these columns are standardized. The data of birth column need to be changed to a Date type from the default type in Pandas(object). To avoid missing any employee records, we must also ensure that full names are separated into new columns and collected for employees with incomplete information. We must also change the datatypes of the Hispanic Latino, and Marital Status columns.

References

Herzberg, B. (2021, November 11). The Datamasters: Data Owners vs. Data Stewards vs. Data Custodians. Satori. https://blog.satoricyber.com/the-datamasters-data-owners-vs-data-stewards-vs-data-custodians/

Roundy, J. (2024, September 9). 6 dimensions of data quality boost data performance. Search Data Management. https://www.techtarget.com/searchdatamanagement/tip/6-dimensions-of-data-quality-boost-data-performance

uCertify. (2026). DAT–325 - Data validation, quality, and cleaning [Online course]. https://snhu.ucertify.com/app/?func=load_course&course=SNHU-DAT325.AJC1&class_code=09YKy