Project Three: Data Validation Report

Creston Getz

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Raymond Rabago

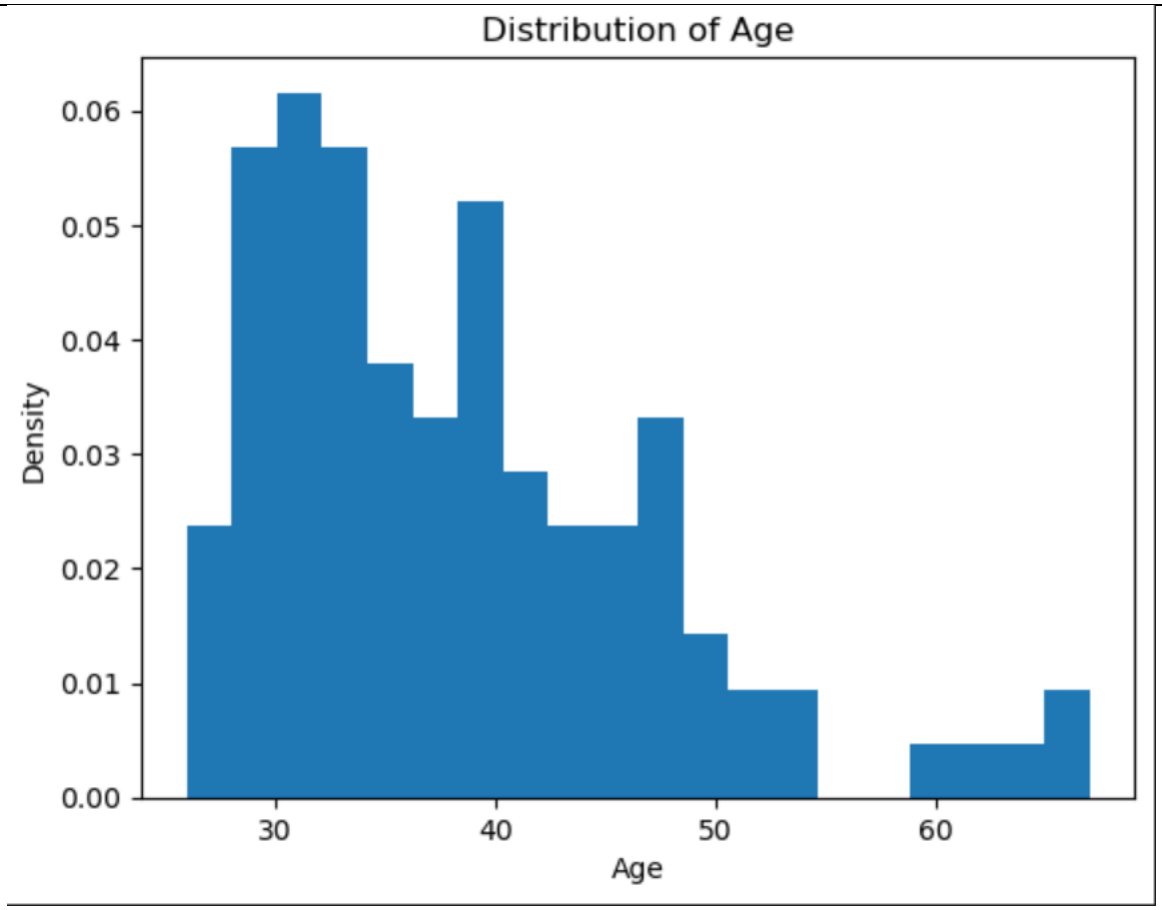
February 16, 2026

Validations of Data Sets

Data Set	Row Count	Column Count
Cleaned firm's data set	93	8 (10 after separating name)
Your company's data set	105	17
Merged data set	105	24

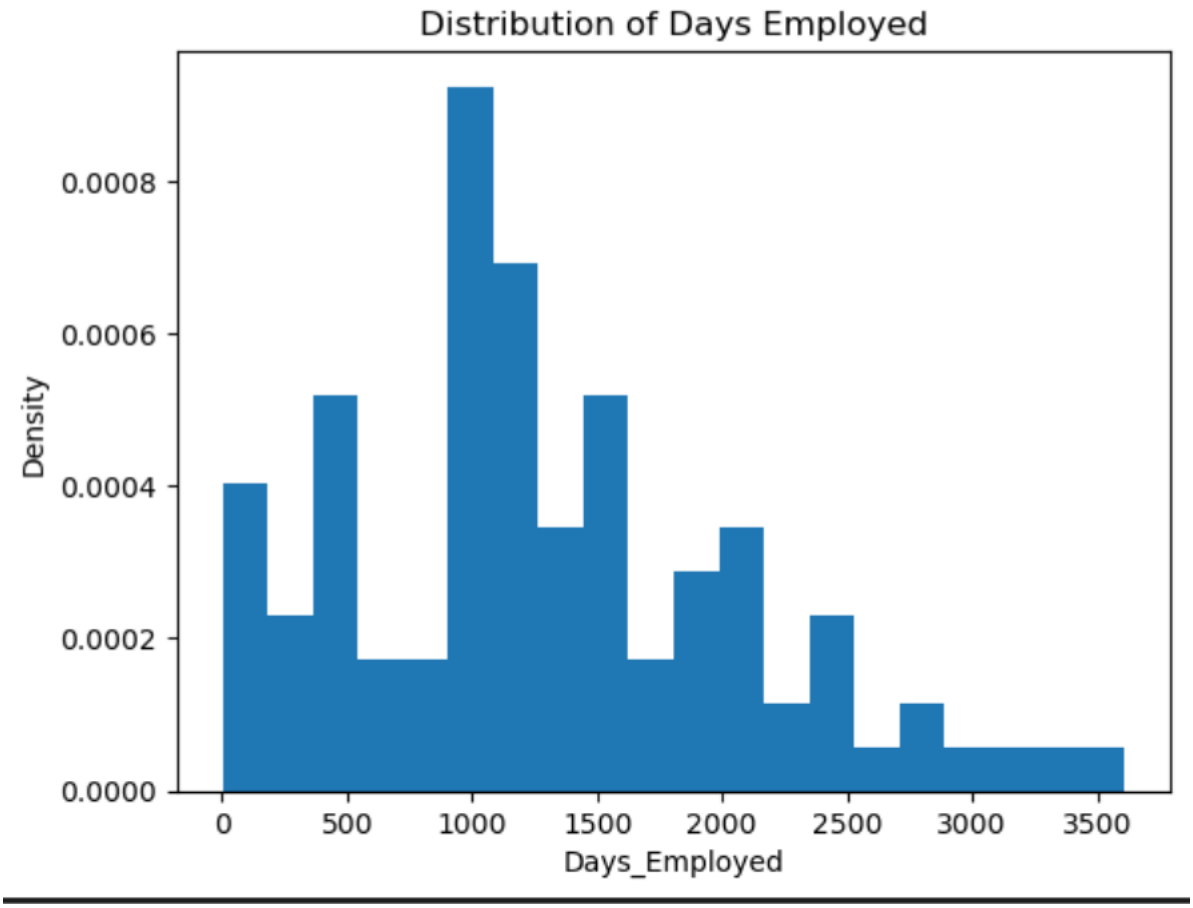
Data Set	Minimum	Maximum	Average
Age	26.00	67.00	38.66
Days Employed	2.00	3611.00	1303.03

Age



Outlier value of 136 was removed from the dataset.

Days Employed



Outlier value of 52,246 was removed. And incorrect value of –514 was also removed.

Summary

In this project, we reviewed and cleaned the new firm's dataset to address various issues and anomalies identified in our data profile. By cleaning the dataset, we have improved the six dimensions of data quality, thereby enhancing the overall data quality of our dataset (Roundy, 2024). For quantitative columns, age and days_employed, we addressed missing values, outliers, and incorrect values found in the data profile. We replaced the missing values with the median to keep the overall distribution without skewing the data from extreme values. Rows with incorrect values, such as negative numbers, were removed from the dataset, as someone cannot be younger

than 0. Lastly, any extreme values were removed from the dataset to prevent the data from being skewed. For example, one row recorded an employee being employed for over 50,000 days, which would have been 140 years of employment. The min, max, average, and plots of the quantitative variables are shown above.

We also addressed the issues with the categorical variables we found in our data profile to ensure high data quality. The columns of marital status and citizen desc had missing values that we removed. Due to the sensitive nature of these columns, we found it best to remove them instead of trying to fill the value. The column Hispanic Latino had some inconsistent values, such as ‘yes’ and ‘Yes’. Having inconsistencies like this can affect aggregate methods, such as trying to find the sum of all Hispanic and Latino people in our dataset. We addressed these values and recommend in the future these be changed to a boolean data type. Citizens desc had a similar inconsistency, that we removed. In one of the rows, it had a value of ‘Male’ which was likely mis-entered. In the date of birth column, we temporarily converted it to a datetime type to perform some analysis. We found in our data profile one row showed a birth year of 1919 paired with an age of 41. By cross-checking other dates of births and ages, we determined this dataset should be in 2017-2018. Based on this, we removed this row because it is impossible. Being born in 1919 and 41 years old would be 1960 not 2017. These validation steps ensure our data is consistent and of high quality.

After we addressed the anomalies we found in our data profile, we merged the new firm's data with our own. To start, we separated the old firm's name column into two separate columns to match our dataset. We then checked that the data types for all columns matched the join. Finally, we performed a left join using the last name, first name, and DOB columns. After the join, we verified that all original company records were kept, and the new firm's data was added

by reviewing the row and column counts listed above. Duplicate checks confirmed no recorded where duplicated during the merge.

The steps we took to clean and transform the new firm's dataset ensure that the dataset is free from invalid/incorrect values, free from outliers, and consistent for future analysis. We listed two histograms above for age and days employed to highlight the distribution of the data and ensure no outliers were found. Data quality is a key step in making sure data is usable. It helps improve decision-making, increases productivity, and business performance (uCerify, 2026). As a result, the merged HR datasets can now confidently be used for these tasks and more, such as reporting and performance evaluations, without the risk of data quality issues.

References

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