Project Three: Data Validation Report

Nicole Hutto

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Professor Jennifer Johnson

October 2025

# Validations of Data Sets

Provide counts of rows and columns in all data sets.

| **Data Set** | **Row Count** | **Column Count** |
| --- | --- | --- |
| Cleaned firm’s data set | 95 | 10 |
| Your company’s data set | 105 | 17 |
| Merged data set | 105 | 24 |

Report the minimum, maximum, and average values of the quantitative variables (after removing negative values and outliers).

| **Data Set** | **Minimum** | **Maximum** | **Average** |
| --- | --- | --- | --- |
| Age | 18 | 64 | ~38.7 |
| Days Employed | 2 | 3,611 | 1,303.76 |

Include data distribution plots for the quantitative variables and confirm if the outlier was removed. You should also specify the outlier value.

|  |
| --- |
| **Age**    The distribution of the Age variable is right-skewed, with most employees between 25 and 45 years old. The histogram shows no extreme outliers. The data values all fall within a realistic human employment range of 18 to 64 years old. Therefore, no outliers were identified or removed for this variable. |
| **Days Employed**    The distribution of Days Employed shows a concentration of employees with less than 2,000 days of employment, gradually tapering off toward the higher end. One extreme outlier at 52,246 days was identified and removed from the dataset. After removing this outlier, the maximum value decreased to 3,611 days, producing a realistic and smoother distribution of employment duration among employees. |

# Summary

**Age**

Most workers seem to be between 25 and 45, with a lot clustering near the early 30s. An odd, far‑out age was spotted and tossed out during cleaning, likely to keep the numbers realistic. Without that weird case, the age spread looks more even, maybe reflecting a normal labor market. Still, the sample could miss some genuine older staff if the cut was too harsh.

**Days Employed**

The record for days on the job once showed a massive outlier – 52,246 days, which is about 143 years, impossible. That point was dropped during quality checks. After its removal, the highest value fell to 3,611 days (about 9.9 years). Most staff have fewer than 2,000 days, and frequency declines as tenure increases. Could this trimming hide long‑term veterans? But it gives a smoother picture.