Project Three: Data Validation Report

Danielle Franklin

Southern New Hampshire University

DAT 325: Data Validation: Quality and Cleaning

Valisa R Brown

August 16th 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 | 105 | 8 |
| 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 | 26.000000 | 67.000000 | 38.669903 |
| Days Employed | 2.000000 | 52246.000000 | 1828.938144 |

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

|  |
| --- |
| **Age**    There was an outlier that was removed. There was an original maximum age value of 139. After removing the outlier, this graph now shows a distribution that reflects the true employee age distribution. |
| **Days Employed**    There was an outlier that was removed. There was an original maximum value of days employed was 52,246 days. After removing the outlier, this graph now shows a distribution that reflects the true days employed distribution. |

# Summary

There were several changes that were made to prepare the firm’s data set to merge with the company’s. The first change was fixing missing values in the variables Age and Days Employed to be filled using the meidan value to maintina consistency. Negative and unrealistic values were identified as outliers and removed from the variable Age and Days Employed. The name variable was stripped of extra spaces and separated into first name and last name to align with the company’s data structure. These validation steps help to ensure that the final import data is clean and reliable. The steps taken help to reduce errors, improve data accuracy, and support high-quality integration. Clean data minimizes inconsistencies durimg merging and improves acuracy of furture analysis.

## References

uCertify. (2025). DAT 325: Data validation: Quality and cleaning