# DAT 325 Project Three Template

# Data Validation

## Human Resources Table

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 105 | 0 | 105 | 205 | 310 |

**Data Set:** Human Resources

**Variable:** Pay Rate by Department

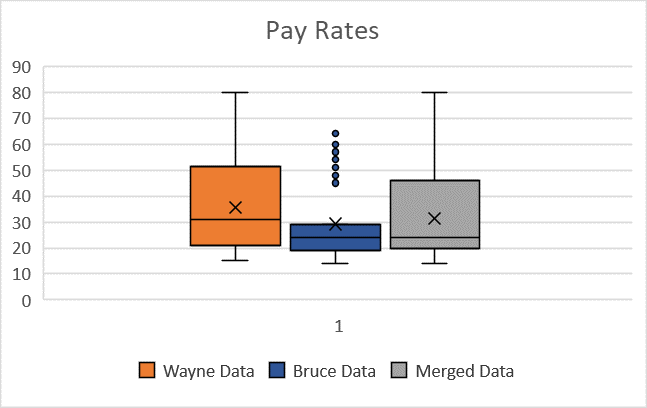
|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | 15 | 14 | 14 |
| **MAX** | 80 | 64 | 80 |
| **AVERAGE** | 35.55 | 29.2 | 31.35 |

## Summary

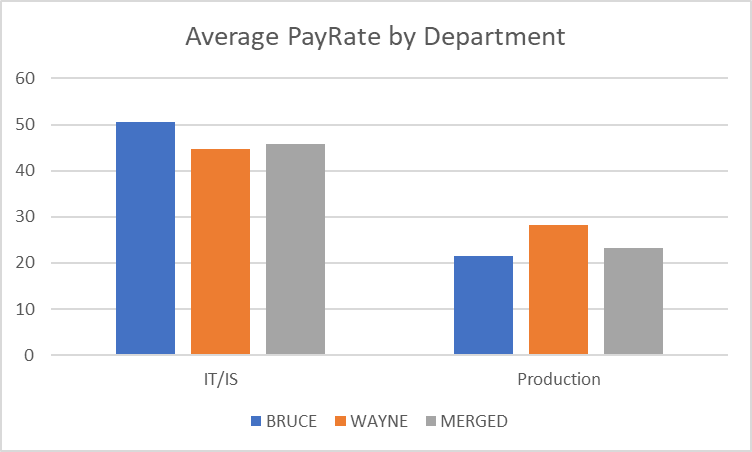
Describe how the distribution data tells you if the import changes the existing distribution.

The distribution data shows that the range of the pay rates is incorporated in the merged data. The data imported from Wayne, Enterprises has a range of pay rates from 15 to 80 where Bruce, Inc. has a range of 14 to 64. The merged data now has a range of 14 to 80, taking the smallest minimum and the largest maximum as its new range. Additionally, the average for the merged data now falls between the averages for each of the individual data sets, as would be expected.

These data sets can also be visualized with box plots, as depicted below, which show the ranges of the data, as well as the mean and median for each of the data sets. The box plot shows where half of the pay rates are for each of the data sets, as noted by the shaded areas. The data from Bruce, Inc. shows a significant number of outlying data points, which when the data sets are merged are no longer outliers. The mean and median for the merged data is also shown to fall between each of the individual company’s data sets in the box plot, which reinforces what is seen with the chart of numbers above.



The two companies have only two departments in common, IT/IS and Production. Breaking the data down by department also shows that within each of the departments the combined data for average pay rate in the respective departments falls between each of the individual companies’ averages.



Why do the validation steps help to ensure that the final import data is clean and of high quality?

The more each of the dimensions of data quality can be satisfied, the higher quality the data will be. Validation allows specific dimensions to be examined regarding the “structure and content” of the data, as noted by Essnet Validat Foundation (2018), meaning that the accuracy, consistency, and validity are the most important aspects to ensuring the data set is worth using.

The validation steps allow the data to be trusted. Knowing that the data imported has impacted the overall data set in a certain way, modifying the range of numbers and the average, shows that it is included in the merged data set. When these numbers are skewed or give unanticipated results, such as a range that does not match the smallest minimum and largest maximum of all data, or an average that changes more than expected or does not change at all, these are significant clues that the data imported is not trustworthy. Issues can include columns of data not aligning or duplication or deletion of data, and validations highlight the need for the merged data set to be explored further to identify where the error has occurred.

References

Essnet Validat Foundation. (2018). *Methodology for data validation 1.1*. https://ec.europa.eu/eurostat/cros/system/files/methodology\_for\_data\_validation\_v1.0\_rev-2016-06\_final.pdf