# DAT 325 Project Three Template

# Data Validation

**Ashley Littles**

## File Names Table

Perform the following data validation on each of the import, existing, and merge data tables.

* The COUNT of rows in each table. You should reconcile these counts to make sure the number of rows you inserted made it into the merged table.

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 506 | 4 | 502 | 14,136 | 14638 |

* Use the identified variable for **your chosen data set** as listed below to determine MIN, MAX and AVERAGE for each table:

**Data Set:** Twitter

**Variable:** Airline Score by Airline name

Determine the MIN, MAX, and AVERAGE in your uploaded table, the MIN, MAX, and AVERAGE in the existing table before you inserted the new rows, and the MIN, MAX, and AVERAGE in the existing table after you insert new rows.

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | -0.6895 | 0.335 | 0.335 |
| **MAX** | 1 | 1 | 1 |
| **AVERAGE** | 0.91160 | 0.9003 | 0.9003 |

## Summary

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

The distribution data allows me to have more insight on the dataset as a whole to ultimately determine how successful the data intergation is. Having proper understanding of the distribution data can help the organization as a whole make better decisions and predictions with the dataset. When looking at the distribution data I see that the import didn’t not drastically change the existing distribution. I was able to correct any anomolies within the data from Wayne Enterprises and combine the cleaned data with the existing Bruce data.

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

The step of using validation within the dataset helps guarentee that the data is of highest quality and accuracy in its final form. Validation makes sure that every column is the exat way we entend it to be and lets us know of any errors that could still be remaining even after cleaning the data. Using the validation steps is similar to having another pair of eyes check over your work to make sure nothing was missed especially with datasets it can be easy to miss something just because there is so much to look at.