# DAT 325 Project Three Template

# Data Validation

Replace the bracketed text with the relevant information.

## 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** | 205 rows | 3 rows | 202 rows | 795 rows | 997 rows |

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

**Data Set:** Supermarket

**Variable:** Gross Income by location

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** | 91.65 | 0.5085 | 0.5085 |
| **MAX** | 4949 | 49.65 | 4949 |
| **AVERAGE** | 1754.72302 | 14.77325849 | 369.3267708 |

## Summary

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

The import most certainly does change how this data is distributed. The average goes from 1754.7 and 14.8 for imported and existing data respectively, to 369.3, which is a wild swing any way you look at it. The Minimum and Maximum values between the imported and existing data are also wildly different.

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

The most important reason that validation helps, especially in this case, is that you can at least rule out errors in the data you are bringing in, eliminating it as a cause of error down the road. For instance, when looking at the distributions above both before and after the merge, and while witnessing such a big differential in these distributive tendencies, if validation has been performed and quality ensured, then you can conclude that there may be an error in the existing data. Or perhaps you could conclude that there was a fundamental difference in the way that the data were measured in the first place. Taking a quick look at the GrossMarginPercentages in the existing *and* imported data, these values are all identical in both data sets. It is not explained why in either case. Another observation I made was in comparing the measures of central tendency above. Looking at the min/max and average of both data sets, the values aren’t *necessarily* that far apart if we were to shift a decimal point to the left by two places on the import data. This leads me to believe that there may be an error in the way that gross income has been calculated between the data sets, specifically the Bruce Inc. dataset.