# DAT 325 Project Three Template

# Data Validation

## File Names Table

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 205 | 2 | 203 | 795 | 998 |

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | 0.9165 | 0.5085 | 0.5085 |
| **MAX** | 49.49 | 49.65 | 49.65 |
| **AVERAGE** | 17.6824 | 14.7733 | 15.365 |

## Summary

Using MIN, MAX, and AVERAGE function, it appeared that the import data only affected the average values in the merged data. The MIN and MAX value of the existing and merged data stayed the same. The existing data contains the lowest value for gross income, and it also contains the highest value for gross income. We can see that the minimum value in import is 0.9165 and the minimum value in existing is 0.5085. As well as the maximum for import is 49.49 and the existing data has 49.65. After merging, the existing data already contained the lowest (MIN) and the highest (MAX) values, so the import data did not have much of an impact. The average changed slightly because we added the import data to the existing data and the number of rows increased.

It is crucial we find the anomalies before we merge with the existing data. We want to ensure that all rows have been merged with the existing data and that no rows are missing. There were 2 rows I had to remove because it contained missing information and I cannot obtain the necessary information. We want to ensure accurate and valid information is in the data set. Another important thing to remember is to match the data sets. Ensure the header is the same and ensure the data type is the same. Otherwise, merging data will be unsuccessful.

Reference:

Bruni, E. (2021, June 18). *The Importance of Data Validation for Data Science*. FullStack. Retrieved April 17, 2022, from <https://fullstack.net/the-importance-of-data-validation-for-data-science/#:~:text=Data%20validation%20is%20a%20cost,for%20accurate%20prediction%20and%20analysis.&text=Data%20validation%20is%20essential%20to,early%20and%20avoid%20data%20errors>.