# DAT 325 Project Three

# Data Validation

## File Names Table

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 505 | 4 got fixed | 505 | 14,135 | 14,640 |

**Data Set:** Twitter

**Variable:** Airline Score by Airline name

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | -0.6316 (US Airways) | -0.6866 (US Airways) | -0.6845 (US Airways) |
| **MAX** | 1 (Frontier) | -0.0617(VirginAmerica) | 1 (Frontier) |
| **AVERAGE** | -0.4198 | -0.4671 | -0.4655 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Count of AirlineScore** | **AirlineScore** |  |  |  |
| **AirlineName** | **-1** | **0** | **1** | **Grand Total** |
| VA01 | 181 | 171 | 152 | 504 |
| US01 | 2263 | 381 | 269 | 2913 |
| UN01 | 2633 | 697 | 492 | 3822 |
| SW01 | 1186 | 664 | 570 | 2420 |
| FR01 |  |  | 1 | 1 |
| DL01 | 955 | 723 | 543 | 2221 |
| AA01 | 1960 | 463 | 336 | 2759 |
| **Grand Total** | **9178** | **3099** | **2363** | **14640** |

## Summary

The merger of the Wayne Enterprises dataset into the Bruce Inc. database has been completed. This merger of datasets was completed and validated using the Excel functions that were outlined in the Executive Summary. To err on the side of caution, anomalies were fixed and not deleted because there were reasonable solutions for each. The distribution data shows that the imported data has slightly changed the existing distribution. This can be seen by the average of the airline scores by airline names. Our existing dataset had an average airline score of -0.4671 and the merged dataset had an average of -0.4655. This is because the import dataset from the Wayne dataset had an average of -0.4198. Their dataset had a lower overall distribution of airline scores which slightly lowered the overall distribution. It is also encouraging that each dataset had US Airways as their lower average airline score by airline name, showing consistency between the datasets. The max average airline score, Virgin American would have been the largest for all three, if it were not for Frontier having only one entry in the database. This is also good because it shows the distribution of the data was the consistent amount the datasets. This also allows for future additions to the database to include Frontier airlines, an airline we had previously not tracked. The validation steps helped to ensure the final import data is clean and of high quality. This is done by making sure the counts of rows in the merged dataset are the sum of rows in the imported and existing dataset. This ensures that no rows were duplicated or deleted in the merging of data. In addition, the min, max, and average airline scores in the merged data do not significantly differ from the other two datasets showing that the data values or columns were not changed in the merging process. The merged dataset has been entered into the Bruce Inc. database and concludes the merger with Wayne Enterprises.