# DAT 325 Project Two

# Executive Summary Report

## Data Set Anomalies

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| 567767091579592000 | Negative AirlineScore\_Confidence | Change to a positive value |
| 567742500366872000 | Extremely high NegReason\_Confidence | Delete row/change decimal |
| 567730696656273000 | New airline name | Add Frontier to lkupAirline |
| 567670985403285000 | Unknown Airline Score | Change to Negative |

## Data Types

| **Header Name From File** | **Data Types Note** |
| --- | --- |
| tweet\_id | Change from scientific notation to standard form |
| airline\_sentiment | Change strings to ints |
| airline\_sentiment\_confidence | Looks good |
| negativereason | Change blanks to N/A |
| negativereason\_confidence | Change blanks to 99 |
| airline | Change values to keys from lkupAirline table |
| airline\_sentiment\_gold | Delete column. 503 null values |
| name | Delete column. Don’t need |
| negativereason\_gold | Delete column. 504 null values |
| retweet\_count | Values seem low but OK |
| text | Delete column. Don’t need |
| tweet\_coord | Delete column. Don’t need |
| tweet\_created | Lose the time stamp and make year 4 digits |
| tweet\_location | Use to determine the time zone for blank entries |
| user\_timezone | Change blank to N/A, get rid of parentheses |

## Specific Transformations Needed to Join the Data

| **Header Name From File** | **Excel Function One** | **Excel Function Two** | **Excel Function Three** |
| --- | --- | --- | --- |
| tweet\_id | Right-click format cells | Set to 0 decimal places |  |
| airline\_sentiment | Add a column to the left | SWITCH(B2:B506, ”negative”, -1, “neutral”, 0, “positive”) | Copy values into the original column as values. |
| airline\_sentiment\_confidence | None |  |  |
| negativereason | IF(D2:D506= “ “, “N/A”) |  |  |
| negativereason\_confidence | IF(E2:E506= “ “, 99) |  |  |
| airline | SWITCH(G2, "Delta", "DL01", "United", "UN01", "Southwest", "SW01", "US Airways", "US01", "Virgin America", "VA01", "Frontier", "FR01") | Add Frontier to lkupAirline as “FR01” |  |
| airline\_sentiment\_gold | Right click column | Delete column |  |
| name | Right click column | Delete column |  |
| negativereason\_gold | Right click column | Delete column |  |
| text | Right click column | Delete column |  |
| tweet\_coord | Right click column | Delete column |  |
| tweet\_created | Left click column | Under Number at the top, select “Short date” |  |
| tweet\_location | Click the city name for the rows with null time zones | Click geography under the data tab | Fill in the time zone column according to the result |
| user\_timezone | IF(D2:D506= “ “, “N/A”) | Find and select values within parentheses | Replace all values with empty strings |

## Executive Summary

The first step in cleaning the data from Wayne Enterprises is to look for anomalies in the dataset. Anomalies are values in a column of the dataset that are unusually high or low compared to the other values within that column. Four anomalies were found within the Wayne dataset which can be found in the first table above. To resolve these issues a variety of simple fixes have been proposed. The first issue is a simple change of a negative sign to a positive sign because the entry makes sense as positive, the negative sign was a mistype. Unlike the next outlier which had an entry of 732 negative reason confidence when the value should go no higher than 1. This might have also been a mistype and was supposed to be 0.732, but the entry also has a tweet location of “All over the world” with no time zone. This makes the entry even more suspect and without further information from the original data owner, it might be wise to delete this entry. Next, there was an entry in the Wayne dataset that had Frontier as the airline, which we do not currently have in our database. It would be wise to add this airline to our database table since they are a reputable airline company. Finally, the last anomaly has “Unknown” as the airline score, but upon further investigation, we can see there is an entry under the negative reason comment meaning this unknown entry should be changed to negative.

The next step in cleaning the data from Wayne Enterprises is to determine the data types in each column and what changes need to be made to match our database here at Bruce, Inc. Most of the columns will either need a small change or be deleted altogether as they are duplicated or contain irrelevant information. The only two columns that do not need any changes made to the datatypes are airline sentiment confidence and retweet count. The tweet id column needs to be changed from scientific form to standard form and the airline sentiment needs to be converted to ints. Negative reason and negative reason confidence both need their blank fields changed to N/A and 99 respectively to match our database standards. The airline column should be changed to reflect our codes from the lkupAirline table. Airline sentiment gold and negative reason gold should both be deleted because they are duplicate values with over 500 null values each. Since we do not track Twitter names or the text within the tweets these two columns can be deleted as well. The tweet-created column should be standardized to contain dates in shorthand form with 4-digit years and no timestamps. The user time zone should have the parentheses deleted and we can use the tweet location to determine some of the 168 time zones that are currently blank using the Excel geography tool. After the tweet location has been carefully used to determine missing time zones it can be deleted as well and the rest of the blanks in the time zone column can be changed to N/A.

Finally, the last step in cleaning the dataset from Wayne enterprises is to perform specific Excel transformations, so that the data cannot be added to our database. The tweet id can be formatted easily to include no decimals. Next, a simple SWITCH function can be used to reflect the values for airline sentiment and airlines that we use here at Bruce Inc. IF statements can be used to change the null values in negative reason, negative reason confidence, and user time zone to reflect the blank values we use in our database. The tweet created column can be changed by left-clicking and then selecting “short date” under the numbers tab. The Excel geography tool can then be used on the tweet location column to determine the time zones for some of the blank values. At last, all duplicate and unnecessary columns can be deleted.

This data quality plan looks to adhere to the data quality dimensions that we find necessary at Bruce, Inc. The dimension of completeness and validity are met by ensuring all fields have a value in them and the value is the same data type as the others in the column. All duplicate columns were deleted and/or combined into one ensuring uniqueness. The dataset will be consistent with other datasets in the database before being entered into the Bruce, Inc database. If the steps in this executive summary are followed then the dataset from Wayne enterprise can be confidently added to the Bruce, Inc. database as it will adhere to our data quality standards.