# DAT 325 Project Two Template

# Executive Summary Report

## Data Set Anomalies

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| 1106026572 | Name- Last name missing | Leave as an empty string if this cannot be reconciled by matching the employee number to an employee last name in another table/database. If empty string violates the “not null” designation, substitute “ZZZZZ, William” for the employee name. |
| 1311063172 | Name- Extra space between last name and first | Trim on import |
| 1108028108 | Name- Extra space between last name and first | Trim on import |
| 1306058816 | Name- Extra space between last name and first | Trim on import |
| 1308060959 | Pay Rate- negative number | Remove negative by imparting abs() function |
| 1109029366 | Zip- invalid number and 6 digits | Cross-reference address and zip to correct, truncate to 5 digits if cannot be completed and leave as a dummy zip of 99999 |
| 1304055947 | Age/DOB- inconsistent (year of birth 1919, age 41) | Ignore Age and enter 1/1/1900 for DOB, to be corrected when cross-referenced |
| 1312063507 | MaritalStatus- missing | Enter “N/A” for marital status field |
| 1301052347 | MaritalStatus- missing | Enter “N/A” for marital status field |
| 1104025243 | MaritalStatus- missing | Enter “N/A” for marital status field |
| 1108027853 | MaritalStatus- missing | Enter “N/A” for marital status field |
| 1403066194 | CitizenDescrip- not matching data expectation, states “Eligible NonCitizen” | Change to Non-Citizen |
| 1011022863 | CitizenDescrip- not matching data expectation, states “Eligible NonCitizen” | Change to Non-Citizen |
| 1411071506 | CitizenDescrip- not matching data expectation, states “Eligible NonCitizen” | Change to Non-Citizen |
| 1206043417 | CitizenDescrip- not matching data expectation, states “Eligible NonCitizen” | Change to Non-Citizen |
| 1407069280 | Date of Hire- missing; Reason for Term/Days Employed- inconsistent, states Has Not Started Yet and Days Employed greater than 0 | Enter 1/1/1900 for date of hire and cross-reference for accurate date. Delete Reason for Termination. Ignore Days Employed as will not be imported. |
| 1211050782 | Date of Termination/Reason for Term- inconsistent, states still employed with a termination date | Delete Reason for Termination |
| 1211050782 | Date of Termination/Reason for Term- inconsistent, states still employed with a termination date | Delete Reason for Termination |
| 904013591 | Days Employed/Reason for Term- inconsistent, states Has not started yet with days employed greater than 0 | Delete Reason for Termination |
| 1010022337 | Days Employed/Reason for Term- inconsistent, states Has not started yet with days employed greater than 0 | Delete Reason for Termination |
| 1311063172 | Days Employed/Reason for Term- inconsistent, states Has not started yet with days employed greater than 0 | Delete Reason for Termination |
| 1412071562 | Reason for Term/Employment Status- inconsistent, states hours are reason and also terminated for cause | Delete Reason for Termination & delete Employment Status |
| 1102023965 | Reason for Term/Employment Status- inconsistent, states performance as reason and voluntarily terminated | Delete Reason for Termination & delete Employment Status |
| 1501072124 | Performance Score- undefined entry | Column of Performance Score data is not utilized in Bruce, Inc. database. Notification only and no action needed. |
| 1403066069;  1307060188;  1101023612;  1308060959;  1101023540;  808010278;  1110029732;  1212052023;  1003018246;  1301052124;  1301052347;  1411071506;  1105025718;  1102024173;  1311063172;  904013591;  1203032255;  1012023013;  1010022337;  1307059817 | Performance Score/Date of Hire- inconsistent, values do not match the number of dates employed. Possibly compliance issue with performance evals. | Column of Performance Score data is not utilized in Bruce, Inc. database. Notification only and no action needed. |

| **Header Name From File** | **Data Types Note** |
| --- | --- |
| Employee Name | Names need to be split into First (EmployeeFName) and Last (EmployeeLName) |
| Employee Number | Ok |
| Age | Will not use |
| Pay Rate | Round to nearest whole number instead of decimals |
| State | Ok |
| Zip | Ok |
| DOB | Ok |
| MaritalStatus | Change data type to INT, coding 0=Single, 1=Married, 2-N/A |
| Sex | Change data coding to 0=Male, 1=Female, 2=N/A |
| CitizenDesc | Change data coding to 0=Not a US Citizen, 1-US Citizen, 2=N/A |
| Hispanic/Latino | Will not use |
| RaceDesc | Change data coding to RaceDescription table |
| Date of Hire | Ok |
| Days Employed | Will not use |
| Date of Termination | Ok |
| Reason for Term | Ok |
| Employment Status | Ok |
| Department | Ok |
| Position | Ok |
| Manager Name | Will not use |
| Employee Source | Will not use |
| Performance Score | Will not use |

## Data Types

## Specific Transformations Needed to Join the Data

| **Header Name From File** | **Excel Function One** | **Excel Function Two** |
| --- | --- | --- |
| Employee Name | In EmployeeFName:  =TRIM(RIGHT(A2:A106,LEN(A2:A106)-FIND(",",A2:A106))) | In EmployeeLName: =TRIM(LEFT(A2:A106,FIND(",",A2:A106)-1)) |
| Employee Number | =B2:B106 |  |
| Pay Rate | =ROUND(D2:D106,0) |  |
| State | =E2:E106 |  |
| Zip | =F2:F106 |  |
| DOB | =G2:G106 |  |
| MaritalStatus | =IF(H2:H106="Married", 1, IF(H2:H106="NotMarried", 0, IF(H2:H106="N/A", 2))) |  |
| Sex | =IF(I2="Male", 0, IF(I2="Female", 1, IF(I2="N/A", 2))) |  |
| CitizenDesc | =IF(J2="Non-Citizen", 0, IF(J2="US Citizen", 1, IF(J2="N/A", 2))) |  |
| RaceDesc | =INDEX('HR Data Dictionary from Bruce '!H3:H8, MATCH('HR Data from Wayne Enterprise'!L2, 'HR Data Dictionary from Bruce '!G3:G8, 0)) |  |
| Date of Hire | =M2:M106 |  |
| Date of Termination | =O2:O106 |  |
| Reason for Term | =P2:P106 |  |
| Employment Status | =Q2:Q106 |  |
| Department | =R2:R106 |  |
| Position | =S2:S106 |  |

## Executive Summary

The data set from was evaluated across the core dimensions to determine the quality. With a guideline for evaluation of this data adapted from DAMA UK (Askham et al., 2013), how well each dimension was met was calculated, and a plan of action was determined. The dimensions were each rated on a scale of 1 to 4, with 1 being unacceptable, 2 needs improvement, 3 good, and 4 excellent. All scores below 4 will need to be examined more closely and the presence of multiple scores of 1 or 2 will raise questions about the usefulness of this data set. Measures of percentages of data that meet the quality dimension specifications are calculated based on 17 columns that will be used in the Bruce, Inc. database and the 105 rows of employee data.

The dimension of timeliness was difficult to ascertain, as there were numerous instances of the employee performance outcomes not documented in a way that matched the expected completion of the performance review. In 20 of the 105 employees, the performance review was either not completed in the timeframe expected or the database was not updated. While the category of Performance Score is not one that will be imported to our company’s database, it does raise questions about whether termination dates are updated quickly or if there are individuals who have been terminated, but it is not noted in this database. Timeliness is an area that needs improvement and moving forward the expectation would be that any changes to the employee database would be done in a timely fashion. The score for this dimension is 2.

The completeness of the data set varied across the different fields. 99% of the names were filled in completely with only 1% missing. This will be rectified with a dummy last name for the one missing until it can be cross-referenced. This is a field that is designated not null in our company’s database and will not pass without an entry and further missing data in the name field is not anticipated to be an issue. 96.2% of the marital statuses were filled in, with 3.8% missing. For the interim, the marital statuses that are blank will be set to “N/A”. This is also a not null field in our company’s data base and can remain as “N/A”. 99% of the hire dates are present, with one date missing. A dummy date for this will be entered until the correct data can be found to change. By substituting dummy or placeholder information, the data set will remain searchable and useable for the remaining fields. Since this is an HR database, deleting rows would mean deleting employee records, which may not be advisable. The remaining 14 columns for our company’s database are complete, indicating 99.7% of the data fields are completed. The completeness of this dataset is good and the score is 3.

The evaluation of the data set’s uniqueness indicated there were no duplications of data that were evident. There is no apparent repetition of inputs and the data points all appear to be unique. The score for this dimension is 4.

The validity, or integrity, of the data had various points where improvements could be noted. The data points that were determined to be invalid included a zip code of 6 digits, all of which were 9, when the expected input is 5 digits. 99% of the zip code data was valid, with this one exception. This one data point will be corrected to have 5 digits and the actual zip code will be found in cross-referencing the employee’s address. The pay rate field also indicated a negative hourly rate for one employee where the expectation is positive integers. The assumed correction for this will be to remove the minus from the data cell manually. 99% of the pay rate data is valid. Overall, 99.9% of the data set that will be imported is valid. The score for validity is 3.

The consistency of the data set was poor in different columns of data. The termination dates, reasons for termination, and number of days employed were inconsistent. 92.4% of the reasons for termination were consistent with data indicated elsewhere and 7.6% were inconsistent. For these occurrences, the reason for termination will be deleted and the number of days employed is not imported and will be ignored. The 2 termination dates that were inconsistent with other information, such as termination reason or days employed, will remain. In another instance the age and date of birth were inconsistent. For this, the date of birth will be changed to a dummy value of 1/1/1900 and the age will not be changed as it will not be imported to our company’s database. Overall, 99.4% of the data is consistent. The overall rating for consistency is 3, though it should be noted that information regarding dates of hire, termination dates, and reasons for termination should be examined much more closely for accuracy.

The accuracy of the data set is difficult to determine, though in conjunction with other dimensions of quality it can be determined that at least some of the data points are inaccurate. The lack of a last name for an employee, or a negative salary for another employee, are all examples of inaccuracies in this data set. A dummy zip code was entered and is not an accurate zip code for that employee. Multiple inaccuracies were also found in the dates of termination and/or reason for termination, though it is not clear which of these were inaccurate by looking at the data set. Performance score data, though not utilized by our company, was likely the most inaccurate, least updated column of data. This was excluded from this evaluation of data quality as it is not relevant to the data that will be imported. Overall, many of these inaccuracies will be deleted or a dummy value will be entered to hold the place of the cell so further analysis of the data set can proceed. Accurate information will be found by cross-referencing other sources of information in order to update with the actual data values. Based on the information available, 99.1% of the data is estimated to be accurate. The rating for accurace is 3.

The total score for the data set is 18 out of a possible 24 points. There are some areas that will require monitoring to ensure the data is useful, including information around termination dates and reasons. Once these corrections are made, in transforming some of the data to meet our company’s business needs, the data can be more trustworthy going forward. The transformations include separating the first and last names of the employees and placing them in the corresponding columns. Also, the string inputs will be changed to categorical values for marital status, sex, and citizenship. The string of an employee’s race will be matched to the corresponding values in the table for race in our company, which will provide uniform data on employee race. The pay rate will also be rounded to give the closest whole dollar amount, as is utilized in our company’s database. These transformations will incorporate the data from Wayne Enterprises into the database for Bruce, Inc. seamlessly.

References

Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., & Schwarzenbach, J. (2013, October). *THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT Defining Data*. DAMA UK. https://docplayer.net/3987248-The-six-primary-dimensions-for-data-quality-assessment.html