# DAT 325 Project Two Template

# Executive Summary Report

## Data Set Anomalies

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| 704-48-3927 | Missing city location data | Recommended Row Deletion |
| 252-56-2699 | Negative price data, should always be positive | The negative symbol could be a typo, the data in this field doesn’t look suspicious. Converting this to a positive price is likely sufficient. |
| 778-71-5554 | Missing date data | Recommended Row Deletion |
| 505-02-0892 | Invalid payment type (Bitcoin) | N/A is an available type, if row deletion isn’t favorable, the bitcoin value could be swapped for N/A. |
| 347-34-2234 | Invalid COGS value range, should not be negative, value is invalid (-9999) | Recommended Row Deletion |

## Data Types

| **Header Name From File** | **Data Types Note** |
| --- | --- |
| Invoice ID | Appears OK |
| City | Needs to be transformed from varchar to varchar (1) with proper city designation format |
| Customer Type | Transform from varchar to INT type |
| Gender | Transform from varchar to INT type |
| Product Line | Appears OK |
| Unit Price | Appears OK (proper formatting needed (0.00 instead of 0.0) |
| Quantity | Appears OK |
| Date | Appears OK |
| Time | Transform data from date type to INT type |
| Payment | Appears OK |
| COGS | Appears OK |
| Gross Margin Percentage | Appears OK |
| Rating | Not Relevant/Needed? |
| **MISSING** Tax Applied | Needs to be created |
| **MISSING** Total Order | Needs to be created |
| **MISSING** Gross Income | Needs to be created |

## Specific Transformations Needed to Join the Data

| **Header Name From File** | **Excel Function One** | **Excel Function Two** | **Excel Function Three** |
| --- | --- | --- | --- |
| City | =substitute(B2,”Yangon, Myanmar”,”A”) | =substitute(B11,”Mandalay, Myanmar”,”B”) | =substitute(B3,”Naypyitaw, Myanmar”,”C”) |
| Customer Type | =substitute(C2,”Member”,1) | =substitute(C3,”Normal”,0) | none |
| Gender | =substitute(D2,”Female”,1) | =substitute(D4, “Male”,0) | Even though there is an option to give fields with “N/A” a value of 2, no such entries exist in this data set. |
| Time | =IF(AND(I2>=TIME(07,00,00),I2<=TIME(11,00,00)),"0","") | =IF(AND(I3>TIME(11,000,00),I2<=TIME(15,00,00)),"1","") | =IF(AND(I4>TIME(15,00,00),I4<TIME(07,00,00)),"2","") |

## Executive Summary

After a quick review, this data doesn’t appear to need a lot of work, however some small inconsistencies and formatting issues need to be addressed. Some small anomalies are very apparent from the start such as missing data, invalid data ranges (negative where it should be positive, or way too high or both), and invalid data types. Many of these small issues can and probably should be solved simply by deleting that row entry, other issues are simple single replacements. A more complicated adjustment to make will be to transform data types for proper compatibility. Many of these can be completed via the substitute function and just drag copying the formula to the following rows. The issue of transforming the time data from HH:MM format to INT (0,1,2) is a little bit more complex. If not for the limit on syntax length, this could be done with one formula (if/and, if/and, else), however it will have to completed via three different implementations of the same if/and formula. Most of the data appears to be usable in its current format, although for consistencies sake data such as unit price may need to be adjusted to be the proper length (IE: $15.00 instead of $15.0). There is also the matter of data that is missing or not needed. For data that is missing, it is advisable to add the columns in the necessary places and perform the calculations needed to populate them, such as Tax Applied, *before* the data migration occurs. This may help to prevent compatibility issues in the data migration later. Then the issue of the apparently useless “Rating” data. We do not track rating data on our end, and no method explaining how this data is calculated has been given. Without a method to calculate rating we can not modify data on our end prior to migration, or after, to include rating. It is therefore advisable that the rating data either be deleted before import, or a blank field be added on our end that can be populated by placeholder values should this data be deemed important.