# To: Supervisor Smith

**From: Brandon Hobbs**

**Date: January 25, 2023**

**Subject: Executive Summary and Actions for Wayne Supermarket Data Project**

To create an effective joining of Wayne and Bruce data, without sacrificing quality, the data was profiled with our current Business Rules in mind. Moreover, the attributes available within the Wayne data were reviewed for their possible mapping to the Bruce data.

Anomalies identified in the Wayne data, and a corrective action, are listed in the **Data Set Anomalies** table below. The differences in the attribute names and types, e.g., % vs. decimal, were identified and added to the **Data Types** table. A plan to clean and convert the Wayne data to conform to Bruce’s Business Rules are shown in the **Specific Transformations Needed to Join the Data** table. The business team may need to define a few more business rules to join the data, i.e., BitCoin may be valid for Wayne Enterprises but not for Bruce, should this be converted to “N/A” or “Cash”?

Validation tests, post-transformation, were conducted which show that transformation caused no change within the data Wayne data.

Both companies have the concept of *Product Line* but there may be a need to harmonize the spelling, capitalization, and meaning across individual products before any modelling is built upon this attribute.

Lastly, it was determined that three attributes available within the Bruce data are not available within the Wayne data. These missing attributes, and corrective actions, are shown in the **Attributes Missing from Wayne Data** table.

After cleaning and transformation, less than 10 objects will be rejected and there should be no loss of business value.

Please contact me if you have any questions or concerns about this analysis.

Brandon Hobbs

## Data Set Anomalies

As the Wayne data was analyzed a few anomalies were identified. The table below shows the anomalies identified and the action taken to resolve the issue.

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| -9999 (COGS) | Invoice 347-34-2234 has a negative COGS | Research correction or delete row |
| Myanmar (City) | City is missing for Invoice 704-48-3927 | Delete row |
| -43.19 (Unit Price) | Invoice 252-56-2699 has a negative Unit Price | Delete row |
| blank (Date) | Invoice 778-71-5554 has no date | Delete row |
| BitCoin (Payment) | Invoice 505-02-0892 shows BitCoin as payment | Convert to “N/A” or “Cash”. BitCoin may be valid for Wayne Ent., but not for Bruce |

## Data Types

## The Wayne data had different attributes (header names) available. The following table shows the mapping (and any notes) of the Wayne attributes to the Bruce attributes.

|  |  |  |
| --- | --- | --- |
| **Header Name from File** | **Equivalent Bruce Attribute** | **Data Types Note** |
| Invoice ID | InvoiceID | Ok. Check for uniqueness |
| City | Location | String slice for city and lookup using *lkupLocation* |
| Customer id | CustomerType | Ok. Convert to bool codes, i.e., 0 or 1 |
| Gender | Gender | Ok. Convert to 0, 1, 2 |
| Product line | ProductLine | Ok. May need to harmonize Wayne and Bruce names and spellings |
| Unit price | UnitPrice | Ok. Check for invalid prices, e.g., negatives |
| Quantity | Quantity | Ok. Check for floats or invalid quantities |
| Date | DateOfPurchase | Ok |
| Time | TimeOfDay | Convert to **TimeOfDay** using the range IDs |
| Payment | PaymentType | Ok. Need to establish a business rule on invalid methods, i.e., is BitCoin “N/A” or “Cash” |
| cogs | CostOfGoodsSold | Ok. Check for invalid values, e.g., negatives |
| gross margin percentage | GrossMarginPercentage | Ok. Wayne used % instead of a decimal representation. Values will need to be converted (divide by 100). However, this attribute could be considered for deletion as it is perfectly correlated with **CostOfGoodsSold** and **TotalOrder** and could cause issues in models later |
| Rating | Not used |  |

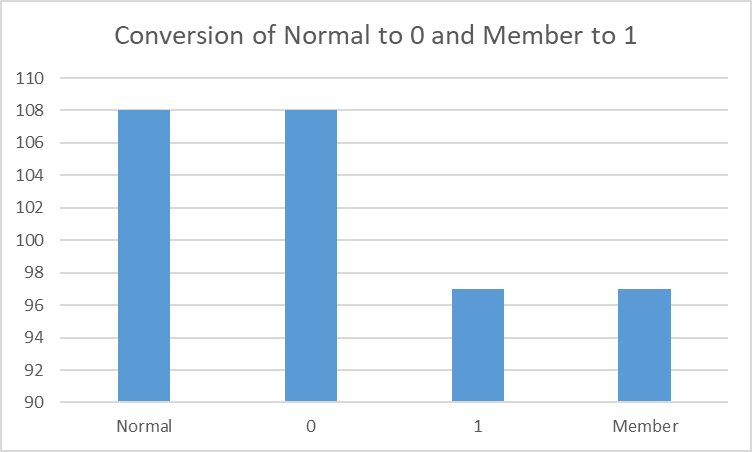
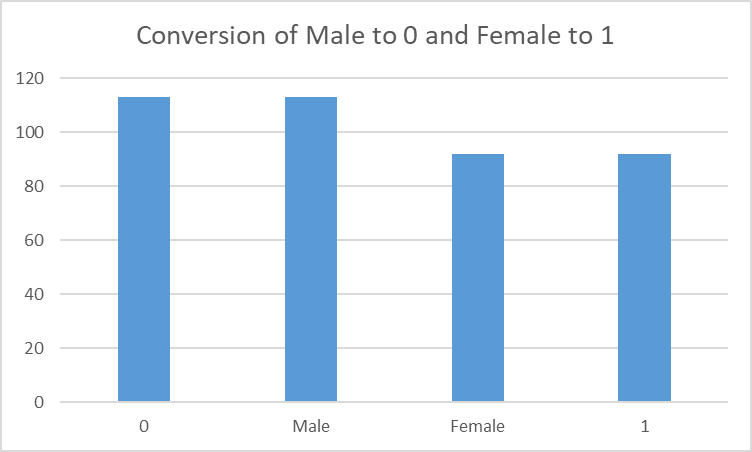
## Specific Transformations Needed to Join the Data

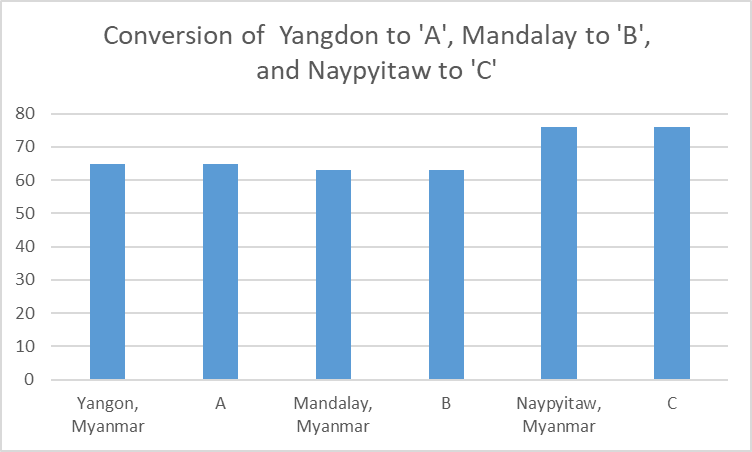
To perform the mapping of the Wayne data to the Bruce data, as per the previous table, some data transformations are needed. This table shows the method or formula used to perform the transformation.

Note: conversion of time is not straightforward and require an intermediate value, **TimeInMinutes**. The Time attribute in the object is converted to minutes, i.e., Hour(I2)\*60 + Minute(I2). This intermediate value is then classified.

| **Header Name From File** | **Excel Function One** | **Excel Function Two** | | **Excel Function Three** |
| --- | --- | --- | --- | --- |
| Invoice ID | PivotTable to identify duplicates |  | |  |
| City | FIND(",",B2)-1 | LEFT(B2,FIND(",",B2)-1) | | VLOOKUP(LEFT(B2,FIND(",",B2)-1),'SM Data Dictionary from Bruce'!$G$3:$H$5,2,FALSE) |
| Customer id | IF(C2=LOWER("normal"), 0, 1) | | | |
| Gender | IF(D2=LOWER("male"), 0, IF(D2=LOWER("female"),1,2)) | | | |
| Product line | None |  | |  |
| Unit price | None |  | |  |
| Quantity | None |  | |  |
| Date | None |  | |  |
| Time | **TimeInMinutes** = 60\*HOUR('I2)+MINUTE(I2) | | IF(AND(I2>=420, I2<=660),0,  IF(AND(I2>660, I2<=900),1,2)) | |
| Payment | None |  | |  |
| cogs | None |  | |  |
| gross margin percentage | L2/100 or 1 - **CostofGoodsSold** / **TotalOrder** | | | |
| Rating | Delete |  | |  |

Below are the tests performed after transformation to validate data integrity post-transformation.





|  |  |  |
| --- | --- | --- |
| **Hours of Time**  **Attribute** | **Count** | **Sum as if TimeOfDay** |
| 10 | 28 | 28 |
| 11 | 16 | 68 |
| 12 | 18 |
| 13 | 19 |
| 14 | 14 |
| 15 | 1 |
| 15+ | 17 | 109 |
| 16 | 17 |
| 17 | 14 |
| 18 | 17 |
| 19 | 28 |
| 20 | 16 |

|  |  |
| --- | --- |
| Conversion to **TimeOfDay** | Count |
| 0 | 28 |
| 1 | 68 |
| 2 | 109 |

## Attributes Missing from Wayne Data

Upon analyzing the Wayne data, it was notice that three attributes available in the Bruce data are not available in the Wayne data. These attributes and a possible resolution are shown in the table below.

|  |  |  |
| --- | --- | --- |
| **Attribute Available in Bruce but not Wayne** | **Plan for Resolution** | **Possible Transformation** |
| TaxApplied | Possibly redundant as it is perfectly correlated with **UnitPrice** | 0.05 \* **UnitPrice** |
| TotalOrder | Possibly redundant as it is perfectly correlated with **UnitPrice and Quantity** | **UnitPrice** \* **Quantity** |
| GrossIncome | Possibly redundant as it is perfectly correlated with **UnitPrice, Quantity, TaxApplied, and CostOfGoodsSold** | **UnitPrice** \* **Quantity + TaxApplied - CostOfGoodsSold** |