# To: Supervisor Smith

**From: Brandon Hobbs**

**Date: February 6, 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.

To further validate the data, post transformation, descriptive statistics were captured. To start the data profiling, the count of objects was established, Table 1.

**Table 1: Descriptive Statistics of the Raw and Rejected Data**

|  | **Excel Files** | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Source Wayne File** | **Anomalies** | **Import Data** | **Existing Bruce Data** | **Merged Data** |
| **Count** | 205 | 5\*  \*BitCoin converted to “N/A” | 201 | 795 | 996 |

The data for the Gross Margin was computed in the Bruce data, Wayne data, and then the merged data, Table 2. It is important to note that the minimum of each individual minimum is not altered after merging. The same is true for the maximum. The average would need to be weighted for each dataset to have a sense of the correct value post-merge, but this has not been done at this time.

**Table 2: Descriptive Statistics of the Data Before and After Merging**

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Bruce Data** | **Merge Data** |
| **MIN** | $0.917 | $0.5085 | $0.5085 |
| **MAX** | $49.49 | $49.65 | $49.65 |
| **AVERAGE** | $17.63 | $14.77 | $15.349… |

## Summary

Validation tests, post-transformation, were conducted to prove that merging caused no change within the Wayne data. Particularly, the distribution of the transformed Wayne data and the raw Bruce data were determined. See Tables 3 through 5. Validation tests proving transformation caused no damage to the Wayne data were conducted and shown in the previous report.

**Table 3: Distribution of *Location* Attribute in the Wayne, Bruce, and Merged Datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Location** | **Wayne Data Count** | **Bruce Data Count** | **Expected Count** | **Merged Count** |
| A | 64 | 274 | 338 | 338 |
| B | 62 | 269 | 331 | 331 |
| C | 75 | 252 | 327 | 327 |

**Table 4:** **Distribution of *Customer* Attribute in the Wayne, Bruce, and Merged Datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer Type** | **Wayne Data Count** | **Bruce Data Count** | **Expected Count** | **Merged Count** |
| 0 | 107 | 391 | 498 | 498 |
| 1 | 94 | 404 | 498 | 498 |

**Table 5:** **Distribution of *Gender* Attribute in the Wayne, Bruce, and Merged Datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gender Type** | **Wayne Data Count** | **Bruce Data Count** | **Expected Count** | **Merged Count** |
| 0 | 110 | 386 | 496 | 496 |
| 1 | 91 | 409 | 500 | 500 |

By capturing the individual distributions, one can infer if the merge is performed correctly as its distribution will be the summation of the individual distributions.

However, because the transposition of two values can still produce the expected distribution it is important to look at other metrics, such as the MAX and MIN (see Table 2). MIN and MAX are useful as they are absolute and deterministic. With two datasets the MAX or MIN of the merged dataset must be the MAX or MIN of one of the individual datasets. If not, then an improper merge has occurred.

Average, mean, and mode is not as deterministic. These descriptors are sensitive to the range of the dataset and would need to be “weighted” to be useful for validation. This was shown in Table 2 – the merged average is not equal to Bruce’s average nor Wayne’s average.

By performing these validations one can be assured that the merged dataset still represents the data in a meaningful way. A poorly written formula to convert location from city to A, B, C could easily cause the nearly flat distribution to be skewed towards a location. This skew could cause our Executive Team to invest in new stores in the wrong city.

The tests from the prior report and this report prove that the transformed data was not altered in its meaning nor was and data missed or altered by the merging activities. It is my opinion that the merged Bruce and Wayne datasets maintain the quality that our Data Consumers require.

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

Brandon Hobbs