# DAT 325 Project Three

# Data Validation

In order to integrate the Wayne and Bruce datasets smoothly and ensure accuracy, we carried out a detailed profiling process while adhering to our Business Rules. This process involved a meticulous examination of each attribute in the Wayne dataset to assess compatibility with the Bruce dataset.

**Key Steps Taken:**

1. **Data Profiling**:
   * We conducted a comprehensive analysis of the Wayne dataset.
   * Each attribute was scrutinized to ensure compatibility with the Bruce dataset attributes.
2. **Compatibility Check**:
   * Attributes in the Wayne data were reviewed for their compatibility with the Bruce data.
   * This step ensured that any differences or similarities were identified early on.
3. **Descriptive Statistics**:
   * To confirm the reliability of the data post-transformation, we gathered descriptive statistics.
   * These statistics provided insights into the data’s central tendency, dispersion, and overall structure.
4. **Object Count Establishment**:
   * The initial step in our data profiling was to establish the object count, which is displayed in Table 1.
   * This helped us understand the volume and scope of the data we were working with.

**Conclusion**: By following these steps, we ensured that the integration of the Wayne and Bruce datasets was both seamless and accurate, maintaining the integrity and reliability of the data throughout the process.

|  | **Excel Files** | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **Source File** | **Anomalies** | **Import Data** | **Existing Data** | **Merge Data** |
| **Count** | 205 | 5 | 203 | 796 | 996 |

**Data Set:** Supermarket

**Variable:** Gross Income by location

|  | **Excel** | | |
| --- | --- | --- | --- |
|  | **Import Data** | **Existing Data** | **Merge Data** |
| **MIN** | 10.96 | 10.08 | 10.08 |
| **MAX** | 99.96 | 99.96 | 99.96 |
| **AVERAGE** | 58.02 | 55.13 | 55.13 |

## Summary

**Summary of Data Import Validation**

To determine if an import has affected the existing data distribution, it’s essential to look at key statistical measures and visualize the data before and after the import.

**Summary of Findings**

I compared the existing and new data distributions to identify any significant changes.

**Key Observations**

* **Summary Statistics:** Gross Margin was computed for both Bruce and Wayne data, as well as the merged data. Notably, the minimum and maximum values for each dataset remained unchanged after merging. This indicates that the extremes of the data distribution are consistent before and after the merge.

By diligently following these validation steps in the project, we ensure that the data remains consistent, accurate, and reliable for future use. Regular validation checks help identify and correct anomalies, ensuring your dataset is error-free and trustworthy. This thorough approach maintains the integrity of your dataset, allowing for meaningful and reliable analysis. Regular validation also supports ongoing data quality improvements, promptly addressing any inconsistencies and keeping your data in top shape. This comprehensive approach ensures that the data is not only accurate and reliable but also ready for any future analysis.