**Comparison of ALITE\_merge and Manual\_merge Datasets**

**Introduction**

For our class project, we analyzed two datasets, ALITE\_merge and Manual\_merge, created through different merging methods: automated merging (ALITE\_merge) and manual merging (Manual\_merge). Our objective was to evaluate the differences between these approaches in terms of dataset structure, coverage, data quality, and consistency, without performing any cleaning or processing. This comparison helps us understand the strengths and weaknesses of automated versus manual data integration, a key concept in data science. We used Python to compute metrics and generate four visualizations (Figures 1–4) to illustrate our findings. This report interprets the provided output, including summary metrics, country-year alignment, coverage, column overlap, missingness, data-type consistency, numeric comparisons, and the plots, to provide a clear analysis for our class project.

**1. Dataset Structure and Size**

The datasets differ significantly in their dimensions and completeness, as visualized in **Figure 1: Summary Metrics Comparison** (summary\_metrics\_comparison.png).

* **ALITE\_merge**: 16,080 rows, 77 columns, 20.30% average missingness, 11,041 unique country-year combinations.
* **Manual\_merge**: 11,040 rows, 60 columns, 18.29% average missingness, 11,040 unique country-year combinations.
* **Analysis**:
  + **Rows**: ALITE\_merge has 16,080 rows compared to Manual\_merge’s 11,040, a difference of 5,040 rows. This suggests the automated merge captured more observations, possibly by including additional data sources or less strict filtering.
  + **Columns**: ALITE\_merge has 77 columns, while Manual\_merge has 60, a difference of 17 columns. The automated approach likely pulled in more variables, as seen in Figure 1.
  + **Missingness**: Manual\_merge has lower average missingness (18.29% vs. 20.30%), indicating better data completeness per column despite its smaller size.
  + **Unique Country-Years**: ALITE\_merge has 11,041 unique country-years, just one more than Manual\_merge’s 11,040, showing nearly identical coverage of country-year pairs.
* **Insight**: Figure 1 highlights that ALITE\_merge is larger in both rows and columns, but Manual\_merge is more complete, with lower missingness, suggesting a trade-off between scale and quality in automated vs. manual merging.

**2. Country-Year Alignment**

The datasets are perfectly aligned in terms of country-year combinations:

* **Shared Country-Years**: 11,040 (all country-years in Manual\_merge).
* **Total Country-Years**: 11,041 (one additional country-year in ALITE\_merge).
* **Jaccard Index**: 1.0 (perfect overlap).
* **Analysis**:
  + The Jaccard Index of 1.0 indicates that Manual\_merge’s country-year combinations are a subset of ALITE\_merge’s, with ALITE\_merge having one additional country-year.
  + This near-perfect alignment suggests both merging methods target the same countries and years, with ALITE\_merge’s extra row count likely reflecting additional observations for the same country-years.
* **Insight**: The identical country-year coverage ensures that differences between the datasets stem from row counts, columns, or data quality, not from the scope of countries or years.

**3. Coverage of Countries and Years**

Both datasets cover the same countries and years:

* **Unique Countries**: 184 in both datasets, all common.
* **Unique Years**: 60 in both datasets, all common.
* **Analysis**:
  + The identical coverage (184 countries, 60 years) shows that both merging methods use the same criteria for selecting countries and time periods.
  + This consistency eliminates coverage as a source of difference, focusing our analysis on dataset structure and quality.
* **Insight**: The perfect overlap in countries and years simplifies comparisons, ensuring that observed differences are due to the merging process itself.

**4. Column Overlap**

The datasets share many columns but differ in unique variables, as shown in **Figure 2: Column Overlap** (column\_overlap.png).

* **Common Columns**: 56 columns shared between datasets.
* **ALITE\_only Columns**: 21 columns, including labor-related metrics (e.g., Labor force, total, Labor force, total\_64).
* **Manual\_only Columns**: 4 columns, including labor-specific metrics (e.g., Labor force, total\_x, Labor force, total\_y, unemployment metrics).
* **Analysis**:
  + ALITE\_merge has 21 unique columns compared to Manual\_merge’s 4, a difference of 17, as visualized in Figure 2. This reflects the automated merge’s ability to include a broader range of variables, likely from additional data sources like World Development Indicators.
  + Manual\_merge’s 4 unique columns are labor-focused, indicating a curated selection tailored to specific analyses, such as labor force studies.
  + The 56 common columns provide a strong foundation for direct comparisons, covering the majority of Manual\_merge’s variables.
* **Insight**: Figure 2 illustrates that ALITE\_merge’s automated approach results in a wider variable scope, while Manual\_merge prioritizes a smaller, labor-specific set of columns.

**5. Data Quality (Missingness)**

Data quality varies between the datasets, as shown in **Figure 3: Top-5 Missingness Comparison** (top5\_missingness\_comparison.png).

* **ALITE\_merge Top-5 Missing Columns**:
  + nan, Last Updated: 01/28/2025, Data from database: World Development Indicators: 100% missing (metadata/non-data columns).
  + avg\_hours\_per\_year: 69.96%.
  + tfp\_index: 48.63%.
* **Manual\_merge Top-5 Missing Columns**:
  + nan, Last Updated: 01/28/2025, Data from database: World Development Indicators: 100% missing (same metadata issues).
  + avg\_hours\_per\_year: 66.85%.
  + tfp\_index: 44.02%.
* **Analysis**:
  + Both datasets share the same metadata columns with 100% missingness, which are likely artifacts of the merging process and provide no analytical value.
  + ALITE\_merge has higher missingness for key variables like avg\_hours\_per\_year (69.96% vs. 66.85%) and tfp\_index (48.63% vs. 44.02%), as seen in Figure 3, indicating that Manual\_merge is more complete for these columns.
  + The overall lower missingness in Manual\_merge (18.29% vs. 20.30%) aligns with its smaller column count, suggesting a more curated dataset with fewer gaps in usable data.
* **Insight**: Figure 3 highlights Manual\_merge’s better data quality for key variables, while ALITE\_merge’s automated process introduces more missing data, particularly in metadata columns.

**6. Data-Type Consistency**

Data types are mostly consistent between the datasets:

* **Consistent Columns**: 54 out of 56 common columns.
* **Inconsistent Columns**: 2 columns with differing data types.
* **Analysis**:
  + The high consistency (54/56 columns) indicates that both merging methods produce compatible datasets, with only two columns showing discrepancies.
  + These inconsistencies may result from differences in automated (ALITE\_merge) vs. manual handling (e.g., numeric vs. string formats), but their small number suggests minimal impact on most analyses.
* **Insight**: The near-perfect data-type consistency ensures that most columns can be directly compared, with only two columns requiring further investigation.

**7. Numeric Columns Statistical Comparison**

The numeric columns reveal both similarities and significant differences, as visualized in **Figure 4: Distribution of Mean Absolute Differences in Numeric Columns** (numeric\_mad\_boxplot.png).

* **Key Observations**:
  + **Population Metrics** (e.g., Population, total: 1.83e+08 vs. 3.29e+07): ALITE\_merge reports much higher means, with large Mean Absolute Differences (MAD) (e.g., 2.21e+08) and low KS p-values (e.g., 3.17e-07), indicating significant distributional differences.
  + **Capital and GDP Metrics** (e.g., capital\_stock\_total: 9.61e+05 vs. 1.26e+06, real\_gdp\_ppp\_output: 2.91e+05 vs. 2.71e+05): ALITE\_merge shows differences in means, with notable MAD values (e.g., 1.83e+06 for capital\_stock\_total) and KS p-values suggesting distributional differences (e.g., 8.61e-04).
  + **Consistent Metrics** (e.g., kpriv\_rppp, GDP\_rppp, kppp\_n): Zero MAD and KS p-values of 1.0 indicate identical distributions, suggesting these columns are sourced identically in both datasets.
  + **Outliers**: Figure 4 highlights kppp\_n and population metrics as outliers with large MAD, indicating potential scaling or unit issues in ALITE\_merge.
* **Analysis**:
  + The large MAD values for population and capital metrics suggest ALITE\_merge may use different scales, units, or data sources compared to Manual\_merge, as confirmed by low KS p-values indicating distributional differences.
  + Metrics like kpriv\_rppp, GDP\_rppp, and kppp\_n show perfect alignment (MAD = 0, KS p-value = 1.0), indicating shared data sources or consistent processing for these variables.
  + Figure 4’s box plot shows that most columns have low or zero MAD, but outliers like kppp\_n and population metrics drive the differences, suggesting areas where the automated merge may introduce inconsistencies.
* **Insight**: ALITE\_merge’s automated process leads to discrepancies in population and capital metrics, while Manual\_merge is more consistent for these, but both align perfectly for many economic indicators.

**8. Visualizations**

The following plots, generated and saved as PNG files, support our analysis:

* **Figure 1: Summary Metrics Comparison** (summary\_metrics\_comparison.png): Shows ALITE\_merge’s higher row (16,080 vs. 11,040) and column counts (77 vs. 60), but Manual\_merge’s lower missingness (18.29% vs. 20.30%), summarizing structural differences.
* **Figure 2: Column Overlap** (column\_overlap.png): Highlights ALITE\_merge’s 21 unique columns vs. Manual\_merge’s 4, with 56 common columns, emphasizing the automated merge’s broader scope.
* **Figure 3: Top-5 Missingness Comparison** (top5\_missingness\_comparison.png): Contrasts ALITE\_merge’s higher missingness in key variables (e.g., avg\_hours\_per\_year: 69.96% vs. 66.85%) with Manual\_merge’s better completeness.
* **Figure 4: Distribution of Mean Absolute Differences** (numeric\_mad\_boxplot.png): Reveals that most numeric columns are consistent (low MAD), but outliers like kppp\_n and population metrics show significant differences, indicating potential scaling issues in ALITE\_merge.

**9. Conclusion**

This comparison highlights distinct trade-offs between automated (ALITE\_merge) and manual (Manual\_merge) merging approaches:

* **ALITE\_merge (Automated)**:
  + **Strengths**: Larger dataset (16,080 rows, 77 columns), broader variable scope (21 unique columns), and slightly more country-years (11,041 vs. 11,040).
  + **Weaknesses**: Higher missingness (20.30% vs. 18.29%), non-informative metadata columns (100% missing), and significant discrepancies in population and capital metrics (e.g., kppp\_n).
  + **Best Use**: Suitable for exploratory analyses requiring diverse variables, but requires careful filtering of non-data columns and verification of numeric scales.
* **Manual\_merge (Manual)**:
  + **Strengths**: Lower missingness (18.29%), more complete key variables (e.g., avg\_hours\_per\_year: 66.85% vs. 69.96%), and labor-specific columns (4 unique).
  + **Weaknesses**: Smaller dataset (11,040 rows, 60 columns), fewer variables, and still includes non-informative metadata columns.
  + **Best Use**: Ideal for targeted analyses, especially labor-related studies, due to its curated columns and better data quality.
* **Key Differences**:
  + **Scope**: ALITE\_merge’s automated approach captures more data but includes irrelevant columns, while Manual\_merge is leaner and more focused (Figures 1 and 2).
  + **Quality**: Manual\_merge has better completeness for key variables, but both suffer from metadata issues (Figure 3).
  + **Consistency**: Perfect country-year alignment (Jaccard Index: 1.0) and high data-type consistency (54/56 columns) ensure comparability, but numeric discrepancies in ALITE\_merge suggest scaling issues (Figure 4).
* **Recommendations for Class Project**:
  + **Choose ALITE\_merge** for broad, exploratory analyses (e.g., studying diverse economic indicators), but manually exclude metadata columns (e.g., nan) and verify population/capital metrics.
  + **Choose Manual\_merge** for labor-focused studies or when data quality is critical, leveraging its lower missingness and curated columns.
  + **Investigate Discrepancies**: Examine the two inconsistent data-type columns and the large MAD in kppp\_n and population metrics to understand scaling or source differences.
  + **Combine Datasets**: For comprehensive analyses, merge the 56 common columns, using Manual\_merge’s labor columns and ALITE\_merge’s unique variables, after resolving discrepancies.
* **Learning Takeaways**:
  + Automated merging offers scale but requires quality checks, while manual merging ensures focus but may miss broader data.
  + Visualizations (Figures 1–4) are essential for communicating data differences clearly in reports.
  + Understanding dataset structure, quality, and consistency is crucial for reliable data analysis, a key skill for our future projects.

**Acknowledgments**

We thank our instructor for guidance and our classmates for discussions that shaped this analysis. The Python script and plots were developed collaboratively, ensuring a robust comparison of the merging approaches.