# **Comprehensive Report on the Data Preparation Phase**

## **1. Introduction**

This report details the entire Prepare Phase for our staffing data project. Our objective was to ensure that all datasets were accurately loaded, cleaned, validated, and standardized prior to further analysis. This process included:

* Loading preprocessed Parquet files from the interim folder.
* Performing detailed data quality checks.
* Validating and converting data types (with a focus on proper date formatting).
* Standardizing column names for consistency.
* Verifying the integrity of the preprocessed datasets by comparing them against the raw CSV files—where we addressed encoding challenges.

## **2. Environment Setup and Logging**

* **Logging Configuration:** We configured Python logging to capture all output (both print statements and errors) in a log file (prepare\_phase\_parquet\_output.log). This allowed us to trace every step of the process.
* **Data Paths:**
  + Preprocessed Parquet files were loaded from ./data/interim/.
  + Raw CSV files were stored in ./data/raw/.

## **3. Loading and Validating Preprocessed Data (Parquet Files)**

### **3.1 Data Loading**

We loaded the following preprocessed datasets:

* **pbj\_nurse:** 1,325,324 rows × 33 columns.
* **pbj\_non\_nurse:** 1,325,324 rows × 82 columns.
* **qrp\_provider:** 710,016 rows × 16 columns.
* **nh\_survey:** 44,189 rows × 41 columns.
* **nh\_quality\_mds:** 251,464 rows × 23 columns.
* **nh\_ownership:** 144,651 rows × 13 columns.
* **nh\_citations:** 406,789 rows × 23 columns.

The log confirmed that each file was loaded successfully with the expected shapes and column structures.

### **3.2 Data Quality Checks**

For each dataset, we conducted:

* **Missing Value Analysis:** Computed the count of missing values per column. Most critical fields had zero missing values; a few columns (e.g., some deficiency counts in NH Survey and NH Quality MDS) had minimal missing values, which were noted for later review.
* **Duplicate Checks:** Verified that no duplicate rows were present in any dataset.
* **Descriptive Statistics:** Generated basic statistics (mean, standard deviation, min, max, quartiles) for numeric columns to identify outliers or anomalies.

## **4. Data Type Validation and Conversion**

We ensured that all critical fields were correctly formatted:

* **Date Conversion:**
  + **PBJ Datasets (Nurse and Non-Nurse):** The WorkDate field, originally stored as an integer (YYYYMMDD), was converted to a datetime format.
  + **QRP Provider:** The Start Date and End Date columns were converted to datetime objects.
  + **NH Survey:** Converted Health Survey Date, Fire Safety Survey Date, and Processing Date to datetime.
  + **NH Quality MDS:** Converted Measure Period and Processing Date to datetime.
  + **NH Ownership:** Converted Association Date and Processing Date to datetime using an explicitly specified format (%Y-%m-%d), ensuring consistency.
  + **NH Citations:** Converted Survey Date, Correction Date, and Processing Date to datetime.

Logs confirmed that all these conversions succeeded, with the appropriate datetime64[ns] types shown for each converted column.

## **5. Standardization and Basic Cleaning**

To achieve consistent naming conventions across datasets:

* **Renaming Columns:** All column names were converted to lowercase and spaces (and hyphens) were replaced with underscores (e.g., "Provider Name" became "provname").
* **Additional Cleaning:** Trimming extra spaces and ensuring uniformity across all datasets.

The resulting standardized columns were verified in the log output, ensuring clarity and consistency.

## **6. Verification Against Raw Datasets**

To ensure that the preprocessing had not altered the underlying data unexpectedly, we compared the preprocessed Parquet files to the raw CSV files:

### **6.1 Loading Raw CSV Files**

* The raw CSV files (located in ./data/raw/) for each dataset were loaded with special attention to encoding:
  + Files that could not be read with UTF-8 (e.g., PBJ datasets) were loaded using alternative encodings such as Latin-1.
* Each raw file was successfully loaded, with the raw shapes matching those of the preprocessed datasets.

### **6.2 Quality and Shape Comparison**

* **Shape Comparison:** Each raw dataset’s shape exactly matched its corresponding preprocessed version (e.g., pbj\_nurse\_raw was (1,325,324, 33), identical to the preprocessed version).
* **Encoding Adjustments:** The raw datasets required careful handling of encodings—attempting multiple encodings until the correct one was found—ensuring that no data was lost or misinterpreted during the initial loading.
* **Data Quality Checks on Raw Files:** Missing value counts, duplicate checks, and numeric statistics for the raw datasets confirmed consistency with the preprocessed versions.

These comparisons provided high confidence that our preprocessing steps (including re-encoding) preserved the original data’s integrity.

## **7. Decision on Merging**

After thorough review:

* **Merging Decision:** We decided not to merge the datasets because the PBJ nurse and non-nurse files contain distinct variables (as defined in their respective data dictionaries). Keeping them separate avoids potential confusion and preserves the contextual definitions of each role.

## **8. Final Saving of Prepared Data**

The final cleaned and standardized datasets were saved individually as Parquet files in the ./data/prepared/ folder. This allows for:

* Easy access to each dataset during analysis.
* A clear audit trail of the data preparation steps.
* Assurance that the datasets are ready for downstream analysis.

## **9. Conclusion**

The Prepare Phase was executed successfully through the following steps:

1. **Logging Setup and Environment Configuration:** Ensured full traceability by capturing all notebook outputs and errors.
2. **Loading of Preprocessed Datasets:** Verified the shape, columns, and basic statistics for each dataset.
3. **Data Quality Validation:** Checked for missing values, duplicates, and outlier statistics.
4. **Data Type Validation and Conversion:** Converted key date fields to proper datetime formats, ensuring consistency.
5. **Standardization and Cleaning:** Standardized column names to a uniform naming convention.
6. **Comparison with Raw Datasets:** Loaded raw CSV files (handling encoding differences) and confirmed that their shapes and quality metrics match the preprocessed datasets.
7. **Decision Not to Merge:** Opted to keep datasets separate to maintain clarity based on their distinct definitions.
8. **Final Save:** Saved the prepared datasets for further analysis.

This comprehensive process ensures that our data is accurate, consistent, and fully validated—providing a solid foundation for the subsequent analysis phase.