**Project Handover Report: Nursing Home Staffing Analysis**

## **1. Introduction**

This report summarizes the work completed so far in the **Prepare Phase** and **Process Phase** of our Nursing Home Staffing Analysis. It details the key data preprocessing steps, validation checks, challenges faced, and next steps required for a smooth transition. The findings are critical for ensuring the accuracy of our staffing insights and aligning them with our SMART research questions.

## **2. Overview of the Data Pipeline**

### **2.1 Data Sources**

The project utilizes the following datasets:

* **PBJ Daily Nurse Staffing (pbj\_nurse)** - Registered Nurses (RNs), Licensed Practical Nurses (LPNs), and Certified Nursing Assistants (CNAs)
* **PBJ Daily Non-Nurse Staffing (pbj\_non\_nurse)** - Support and administrative staff
* **Quality Reporting Program Data (qrp\_provider)** - Facility quality metrics
* **Nursing Home Survey Data (nh\_survey)** - Deficiency and compliance records
* **MDS Quality Measures (nh\_quality\_mds)** - Resident care performance indicators
* **Nursing Home Ownership (nh\_ownership)** - Facility ownership structure
* **Health Citations (nh\_citations)** - Regulatory citations and penalties

These datasets were initially provided in **CSV format**, then preprocessed and converted into **Parquet format** for optimized performance.

## **3. Prepare Phase Summary**

### **3.1 Data Loading & Encoding Fixes**

* The raw datasets had encoding inconsistencies (UTF-8, Latin1, ISO-8859-1).
* Successfully loaded into pandas using an adaptive encoding approach.
* Converted all datasets into **Parquet format** for efficiency.

### **3.2 Data Quality Validation**

* **Missing Values:** Verified null values across datasets and imputed where necessary.
* **Duplicates:** Removed redundant rows that appeared due to data duplication errors.
* **Outliers:**
  + Found extreme RN staffing ratios (1% of facilities had near 100% reliance on contract RNs).
  + Identified top 1% of extreme staffing hours, confirming they are real rather than erroneous.

### **3.3 Data Standardization & Cleaning**

* Standardized column names for consistency.
* Converted date columns into **datetime format**.
* Normalized categorical values (e.g., facility states, staffing roles).

### **3.4 Key Fixes & Adjustments**

* **Date Conversion Issues:** Some datasets had **integer-based date values (YYYYMMDD)**, which required conversion to datetime.
* **Resolved Column Mismatches:** Some datasets had inconsistent column names and structures across different reports.

## **4. Process Phase Summary**

### **4.1 Key Metric Development**

The following metrics were derived to answer our SMART research questions:

* **RN Temporary Staffing Ratio** = hrs\_rn\_ctr / (hrs\_rn\_emp + hrs\_rn\_ctr + 1e-6)
* **Average Staffing Levels by Facility Type**
* **Intra-Quarter Staffing Variability**
* **Correlation Between Temporary Staffing & Costs**

### **4.2 Outlier Detection & Analysis**

* RN contract hours showed a heavy skew, with most facilities having zero contract hours but a small subset relying entirely on them.
* Scatter plots revealed distinct facility clusters based on staffing ratios.

### **4.3 Time-Based Analysis (Pending Fixes)**

* **Issue:** After grouping by year\_month, the dataset returned **empty results**.
* **Root Cause Identified:** The workdate column became **NaT** after reloading the prepared dataset.
* **Fix Required:** Ensure that workdate is properly saved in Parquet format and reloaded correctly.

### **4.4 Analysis of Three Key Variables**

The following additional analyses were proposed and are pending integration:

* **Extreme Outlier Detection**
  + Identified the top 1% extreme values for hrs\_rn\_emp, hrs\_rn\_ctr, and rn\_temp\_ratio.
  + Required adjustments in visualization to better capture distributions.
* **Exploratory Data Analysis (EDA)**
  + Generated histograms and boxplots for hrs\_rn\_emp, hrs\_rn\_ctr, and rn\_temp\_ratio.
  + Found that contract hours are highly concentrated in a small subset of facilities.
* **Time-Series Trends (Pending Fixes)**
  + Attempted to analyze trends over time but encountered an issue with workdate conversion.
  + Grouping by year\_month returned an empty DataFrame due to NaT values.
  + Next step: Ensure workdate column retains proper datetime format when reloaded.
* **Facility-Level Analysis**
  + Identified top facilities with the highest average RN temporary staffing ratio and total contract hours.
  + Provided deeper insights into which facilities rely most on contract RNs.
* **Correlation with Other Nursing Variables**
  + Analyzed correlations between RN staffing levels and other nursing roles (LPNs, CNAs).
  + Helped determine broader staffing patterns and interdependencies.
* **Advanced Outlier Profiling**
  + Segmented the top 1% of extreme values to analyze their facility type, location, and broader characteristics.

## **5. Next Steps & Open Issues**

### **5.1 Immediate Fixes Required**

* **Resolve Time-Based Analysis Issue:** Ensure that workdate is correctly stored and retrieved from the prepared dataset.
* **Confirm Column Integrity:** Validate that key computed metrics (like rn\_temp\_ratio) persist across sessions.
* **Run Final Validation Checks:** Before progressing to advanced analysis, recheck data consistency across all datasets.

### **5.2 Transition to Process Phase**

* Begin merging facility-level staffing trends with quality measures.
* Conduct correlation analysis between temporary staffing and quality/cost outcomes.
* Develop time-series visualizations for staffing fluctuations.
* Integrate facility ownership and regulatory data to assess broader impacts on staffing.

## **6. Handover Summary**

* **Data Preparation Complete:** Yes ✅ (with minor fixes needed for time-based grouping)
* **Key Metrics Defined:** Yes ✅
* **Outlier Analysis Conducted:** Yes ✅
* **Next Steps Identified:** Yes ✅ (see Section 5)
* **Pending Issues:** Time-based grouping fix required.

This document serves as a comprehensive handover to ensure that the next phase of analysis continues smoothly. Please refer to the process logs and prepared datasets for further validation.