**MSBC 5490**

**BUAN Experiential Project**

**Primoris**

**Data Cleanup Report**

**Submitted by Group 4:**

| **Name** | **Email Address** |
| --- | --- |
| Rafael Cintron | rafael.cintron@colorado.edu |
| Supria Deka | supria.deka@colorado.edu |
| Pranathi Manthri | pranathi.manthri@colorado.edu |
| Murali Prateek Manthri | murali.manthri@colorado.edu |
| Jinal Mehta | jinal.mehta@colorado.edu |

**TABLE OF CONTENTS**

[**1. Dataset Introduction 3**](#_mbqakkk6ca15)

[**Data Cleaning Overview 3**](#_cybb1lqeymor)

[**2. Data Cleaning Phases 4**](#_y8etunj9ntfp)

[2.1 Understanding the Data to Determine Necessary Sheets 4](#_o3k7mbc7j3sa)

[2.2 Removal 4](#_q1iuz1uqbvmz)

[2.3 Primary Key Identifier Standardization 4](#_ojaqjpttkwsh)

[2.4 Adjusting Data Types and Format 4](#_bgy42fsnfq7b)

[2.5 Handling Missing Values 5](#_84q5zq8cdase)

[2.6 Imputation for Reporting and Modeling 5](#_7s2gte7qvw15)

[**3. Feature Engineering 5**](#_12eb2jyijkxy)

[3.1 End Results 6](#_38752amvz6tr)

[**4. Data Merging 6**](#_ne43n7hsrs39)

[**5. Key Analytical Observations 7**](#_qemszdbex8jj)

[5.1. Most Impactful Feature: Promotion 7](#_2jjn90yq00qo)

[5.2. Departmental Findings 7](#_ih873buy8td2)

[5.3. Hierarchical Promotion Patterns 8](#_3ilnm17q9w51)

[5.4. Transfer Analysis 8](#_r0m7omisairb)

[**6. Conclusion 8**](#_2rv5jwu0gnd5)

[6.1 Inference 8](#_fbqmk9revaud)

[6.2 Impact on Data Quality 8](#_oexhrl1rpid7)

[**Recommendations for Future Analysis 8**](#_fr0eoqw1iua)

[**Appendix 9**](#_vdqc8lkhb7c1)

# **Dataset Introduction**

This report is based on a comprehensive analysis of the employee lifecycle dataset provided in the Excel file titled *NH-RH-Term Report for (Co 18) - All Time*. The dataset comprises multiple interconnected sheets that collectively capture various aspects of workforce movements, including *new hires, promotions, demotions, rehires, transfers, terminations, and job title changes*. The objective of this analysis is to clean the data, engineer meaningful features, and derive actionable insights that inform talent management decisions.

The data spans the entire employment journey of individuals, starting from their hiring and continuing through career advancements, internal movements, and exits. The rich structure of the dataset allows us to not only understand current workforce dynamics but also forecast future trends. Key focus areas in this report include identifying the most impactful factors influencing promotions, exploring department-specific patterns, and understanding hierarchical career movements. The analysis also highlights data gaps and addresses them through systematic cleaning and standardization.

Overall, this report aims to provide HR stakeholders with a clear, data-driven overview of employee progressions and transitions, supported by thorough visualization and metrics that can guide future talent strategies.

# **Data Cleaning Overview**

The process of data cleaning is one of the most crucial steps before performing any form of analysis or modeling. For this report, we thoroughly examined the Excel dataset to ensure that all anomalies, missing values, and inconsistencies were properly addressed. By systematically evaluating each column, we ensured data readiness for effective analysis and visualization.

We began by verifying the consistency of both numerical and categorical fields, ensuring that each variable was formatted correctly and contained reasonable values. Next, missing values were addressed by implementing standardized replacements, which helped to maintain consistency across records and prevent errors in later stages of analysis. Additionally, we conducted structural checks to remove irrelevant or redundant data, focusing on creating a dataset that was both lean and meaningful.

Additionally, *all sheets from the provided dataset were merged into a single comprehensive sheet.* This consolidation was done to improve reading efficiency and analytical performance, enabling smoother cross-referencing of employee data across various lifecycle stages. The unified dataset allowed for a more holistic view of workforce patterns and streamlined analysis workflows.

# **Data Cleaning Phases**

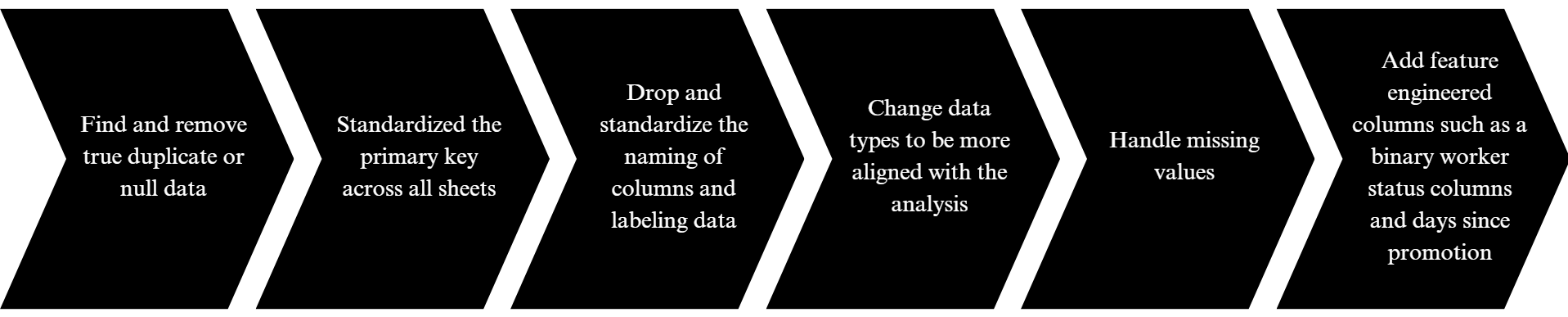


Fig 1: Overviews the steps of the data cleansing process.

## **2.1 Understanding the Data to Determine Necessary Sheets**

After examining the data and asking some questions, we determined that the Terms and All Job Changes sheets were unnecessary for the data. Through this analysis we discovered that T status means terminated, as in they are no longer with the company, not just fired employees.

## **2.2 Removal**

In this phase, unnecessary columns that did not add analytical value were removed from the dataset. Redundant records and duplicate entries were also eliminated, preventing skewed results and ensuring that the data accurately reflected the underlying patterns and behaviors. All columns across these datasets were standardized by converting column headers to lowercase and replacing spaces and special characters with underscores for consistency (e.g., EE # became ee\_number).

The following columns, if present, were dropped from all event-based datasets as they were not relevant for the final analysis:

* PR Dept Desc.
* Craft Class Desc
* Project
* Audit Date

## **2.3 Primary Key Identifier Standardization**

We identified *ee\_number* (originally *EE #*) as the common unique identifier across all datasets. This key was used for all merge operations to ensure consistency in tracking each employee's history across events.

## **2.4 Adjusting Data Types and Format**

There were some inconsistencies and inefficiencies within the data that were adjusted. There were IDs stored as floats which take up necessary space and create longer run times. There were also some areas where data was null, but was filled in with <None> which is still a string so we went through and made those all null.

## **2.5 Handling Missing Values**

Empty cells, especially in columns relating to Promotion, Demotion, TransferIn, and TransferOut, were identified and systematically filled with the placeholder “N/A.” and “Unknown” as per the context and assumption based on the titles as per the corporate structure. This step ensured that no columns contained null values, providing a clean foundation for statistical analysis and avoiding interpretation errors. After merging all the sheets into a single dataframe along with some feature engineering we have performed for better understanding, all *Days\_Since\_* columns, missing values were replaced by ‘NoEvent” to indicate that the employee had not experienced that specific event.

## **2.6 Imputation for Reporting and Modeling**

For further processing and model readiness, the following replacements were made:

* **Days\_Since\_ columns:** Missing or blank values were replaced with the text "NoEvent" to clearly indicate that the employee had not experienced that event. During model building, "NoEvent" will be replaced with 0 to maintain numeric consistency.
* **Job Title column:** Missing entries were replaced with "Unknown" and similarly will be mapped to 0 during modeling.
* **Project Code column:** Missing values were replaced with "Not Applicable" (interpreted as belonging to administrative or finance departments). This placeholder will also be converted to 0 during the modeling phase.

# **Feature Engineering**

Feature engineering added depth to the dataset. We introduced a new metric, “Time Between Promotions,” which enabled a clearer understanding of career progression across roles and departments. We also consolidated department classifications, simplifying the complexity of multiple categories and making the data more interpretable.

We started by consolidating multiple HR datasets, including promotions, demotions, rehires, transfer-ins, transfer-outs, and new hires, along with the all-employee master file. All datasets were cleaned by dropping unnecessary columns *(PR Dept Desc., Craft Class Desc., Project, and Audit Date).* Column names were standardized to lowercase and consistent formatting.

For each event dataset (promotions, demotions, rehires, transfer-ins, transfer-outs, new hires), a new feature called **Days\_Since\_Event** was engineered. This was calculated as:

1. **Days\_Since\_Event** = Reference Date (2024-12-31) – Event Date

This provided a time-based measure for how long ago each event occurred for every employee.

1. **FLSA Category Binary**

The FLSA Category column was transformed into a binary variable for modeling purposes, with the following mapping:

* Exempt → 0
* Non-Exempt → 1.

Hence we engineered two key features as mentioned.

## 

## **3.1 End Results**

After data cleaning and feature engineering, the resulting dataset was consistent, structured, and rich in actionable information. It was now well-prepared for visualization and further analysis, supporting clear and confident decision-making.

# 

# **4. Data Merging**

All event datasets were sequentially merged into the master employee file *“Final\_Megred\_Employee\_NH”* using *ee\_number a*s the join key. This resulted in one consolidated dataset containing each employee’s profile and the number of days since their last occurrence of each event type.

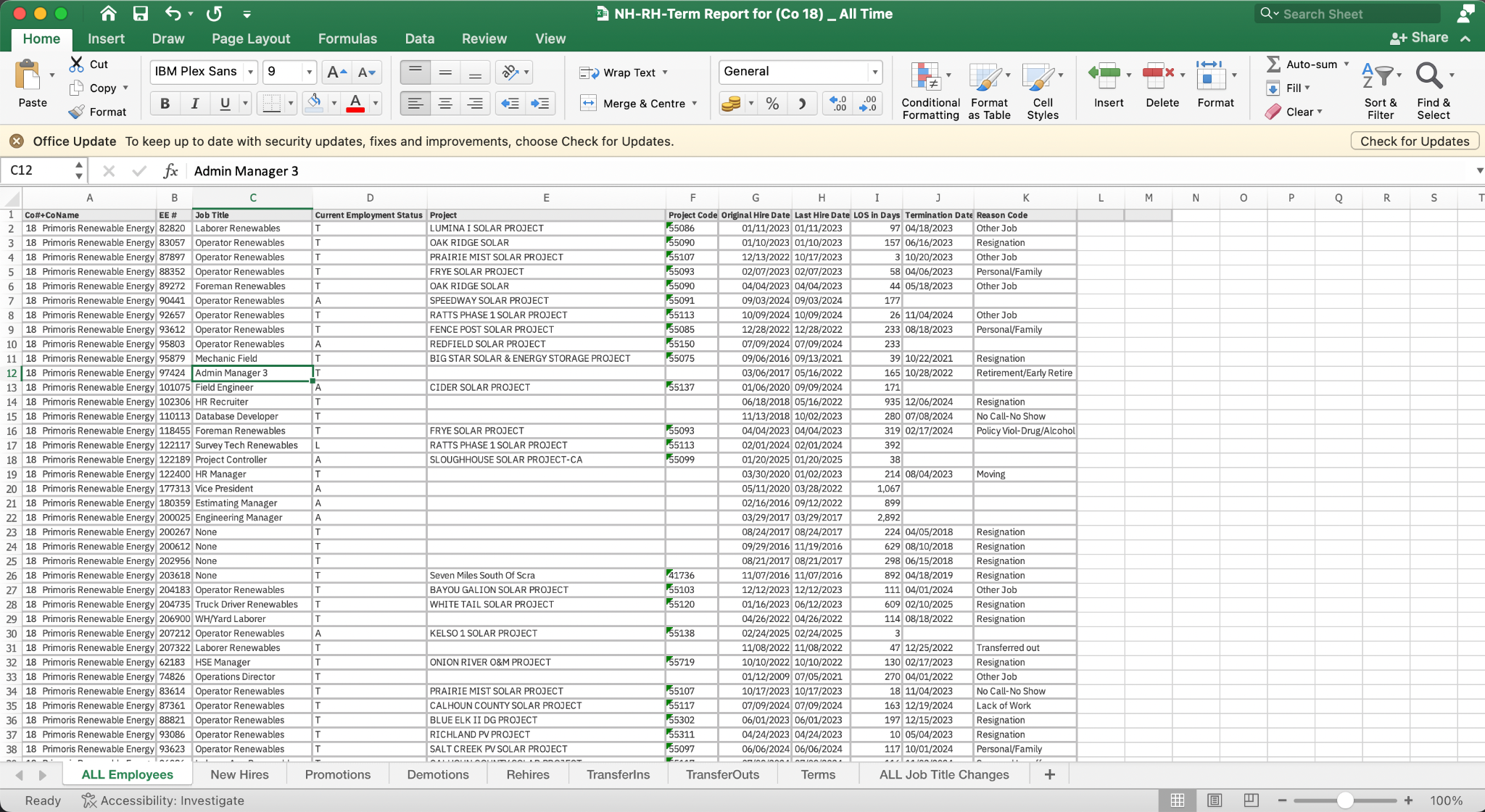
**Screenshot: Before Data clean Up**

Fig 2: Excel Sheet Consist of multiple tabs with missing values for each of the tabs.

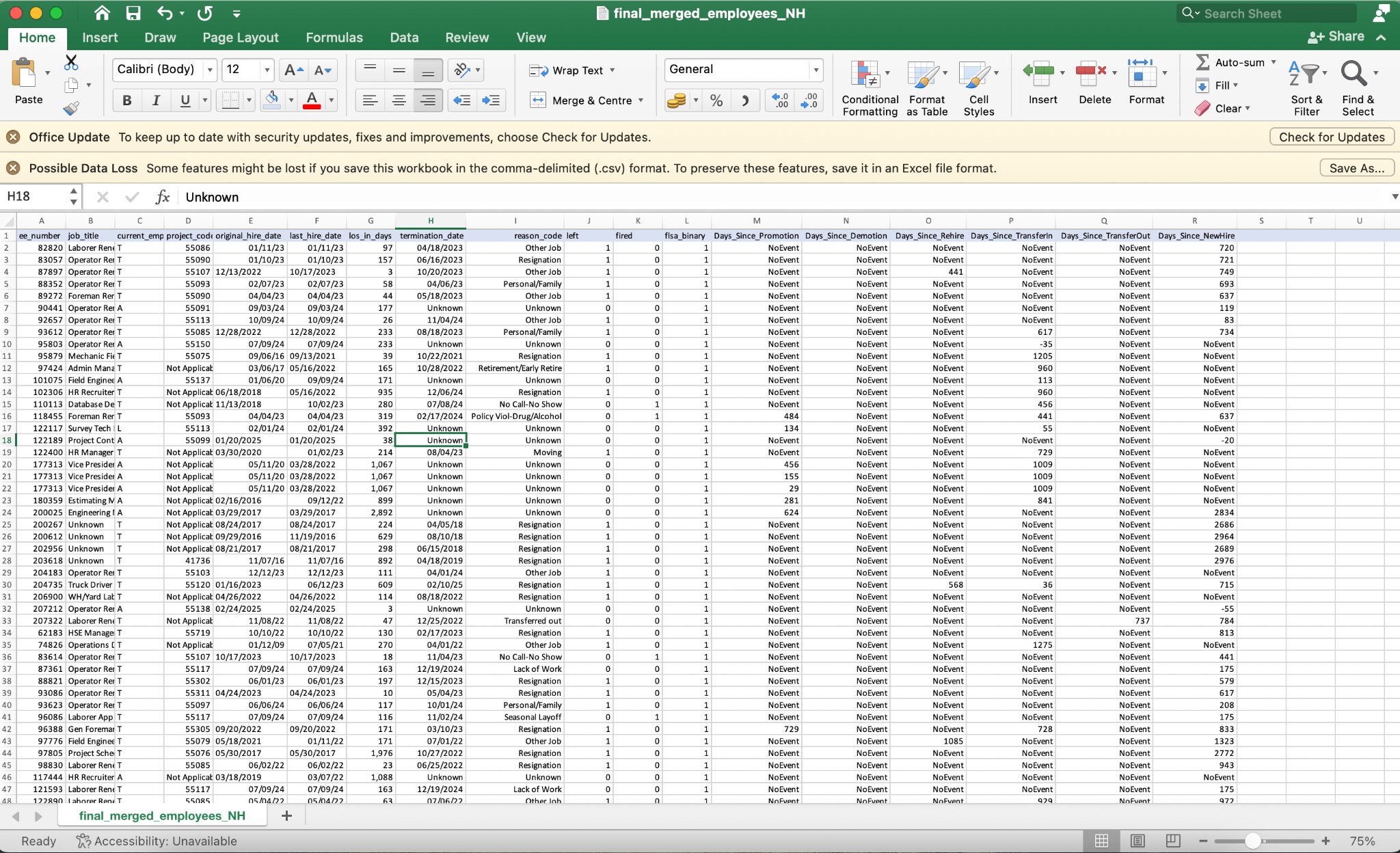
**Screenshot: After Data clean Up**

Fig 3: Single Excel Sheet: Missing Values Handled; Remove Unwanted Columns; Featured Engineered.

# **5. Key Analytical Observations**

## **5.1. Most Impactful Feature: Promotion**

The analysis revealed that Promotion is the most influential feature in understanding employee career progression. Employees who received frequent promotions were often involved in long-term projects and demonstrated consistent career movement. Promotions were largely clustered in senior positions, emphasizing hierarchical favoritism toward higher-level staff.

## **5.2. Departmental Findings**

We found that Admin and Finance departments operated in a static environment without project codes or project names, highlighting their support functions rather than project-based responsibilities. In contrast, the Operations department was the most dynamic, with the largest number of employees actively engaged in projects.

## **5.3. Hierarchical Promotion Patterns**

The report highlighted that senior-level employees experienced quicker promotions with fewer obstacles, benefiting from stability and project visibility. However, lower-level employees faced longer wait times for career advancement, indicating the presence of strict criteria or potential bottlenecks in lower hierarchies.

## **5.4. Transfer Analysis**

Transfers between departments or projects often coincided with promotions, suggesting that mobility within the organization contributed positively to career growth. Employees who had multiple transfers demonstrated quicker progression, particularly within Operations, where project reassignment was frequent.

# **6. Conclusion**

## **6.1 Inference**

The structured approach to data cleaning, combined with strategic feature engineering, allowed us to derive meaningful insights from the dataset. Promotion trends emerged as the strongest indicator of growth, with clear distinctions between senior and junior levels. The Operations department formed the core of project-based activity, while Admin and Finance roles remained stable and support-focused. The engineered metric “*Date Since \_ Columns*” stands as a valuable tool for future predictive modeling and workforce planning.

## **6.2 Impact on Data Quality**

The systematic data cleaning steps, from removal of irrelevant records to handling missing values and feature engineering, significantly enhanced the overall quality of the dataset. These actions ensured that the data became reliable, consistent, and analysis-ready. Merging multiple sheets into a single comprehensive view streamlined analysis performance and allowed for efficient cross-referencing. Altogether, these steps contributed to creating a robust dataset that supports informed decision-making and predictive workforce planning.

# **Recommendations for Future Analysis**

We recommend developing predictive models to estimate employee retention rate along with some other important measures to understand the probabilities based on historical data. Additionally, segmenting promotion timelines by project codes could help identify high-performing teams and areas of improvement. Finally, retention strategies should be considered for lower-level employees who face extended wait times for promotions based on the assumption and further clients based stands up.

# **Appendix**

1. Help me build a structure/skeleton for writing a professional data cleaning report.
2. How can I build a report which is clear, structured, and suitable for submission in a business or academic setting?
3. Should I include an Introduction section outlining the purpose of the report, the dataset source, size, features, and relevant business context?
4. What is the purpose of including a Data Overview section describing key variables, data types, and initial observations regarding missing data, duplicates, and outliers.
5. Help me understand the importance of handling or removing the missing values, their differences and their impact.
6. How to merge the xlsx to csv single format?
7. How to clean a dataset when all the variables are categorical?
8. What should be kept in mind when splitting the dataset into train and test dataset?

**Notes:**

1. Day\_since columns : blank is replaced as "NoEvent" this indicates no event occurred for that employee.
2. Job title : blank is replaced as "Unknown"
3. Project code : blank is replaced "Not Applicable" [Admin/finance dept]

**This will be replaced with 0 during model building to maintain the numeric constraint and integrity.**