**Naan Mudhalvan Project**

**TN Marginal Workers Assessment**

**Phase 3**

**TEAM MEMBERS**

PRAVEEN V 2021115076

RATISH MS 2021115085

SANJAY R 2021115091

SARAVANAN S 2021115094

HEMANTH P 2021115305

**Data Preprocessing Report:**

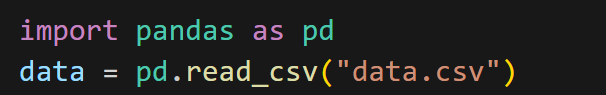
***Executive Summary:***

This report outlines the data preprocessing steps undertaken for the "TN Marginal Workers Assessment" project. The primary objective of this preprocessing is to prepare the dataset for subsequent analysis, ensuring data consistency and completeness.

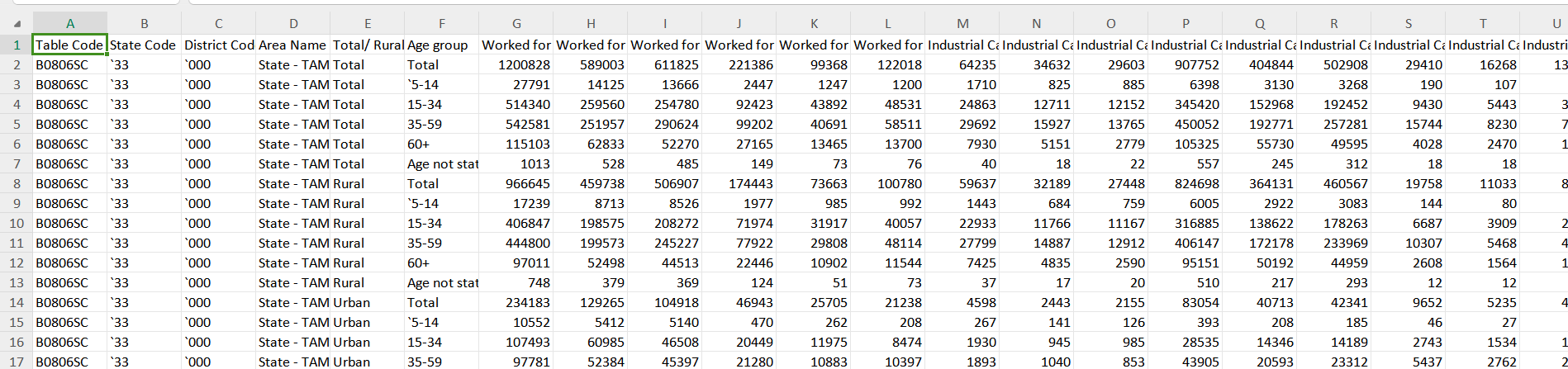
***Data Preprocessing Steps***

**Step 1:** Data Loading

- The dataset was loaded from the provided CSV file.

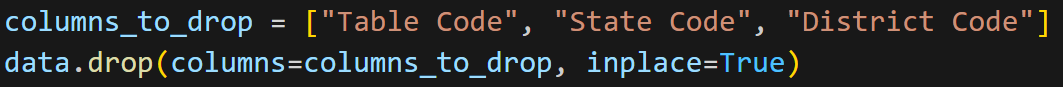


**INITIAL DATASET**

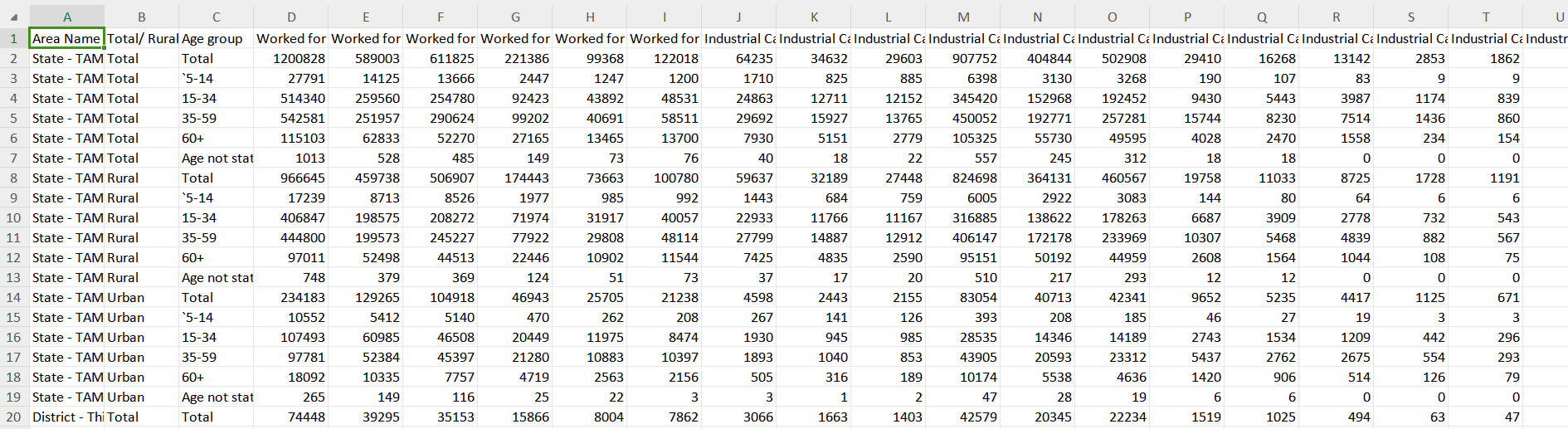


**Step 2:** Column Removal

- Unnecessary columns, namely "Table Code," "State Code," and "District Code," were removed to simplify the dataset and focus on relevant attributes.

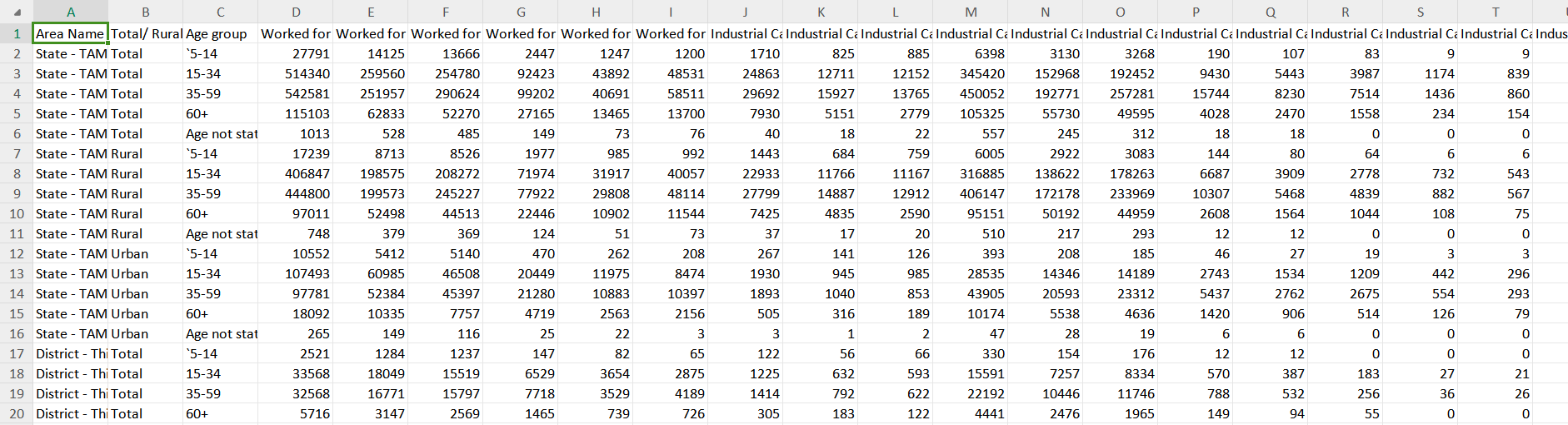


**OUTPUT:**



**Step 3:** Removal of Rows with "Total" in Age Group

- Rows with "Total" in the "Age Group" column were eliminated as they represent aggregated data and do not pertain to individual age groups.



**Step 4**: Age Group Mapping

- The "Age Group" column was standardized to ensure a consistent format.

- Specific mappings:

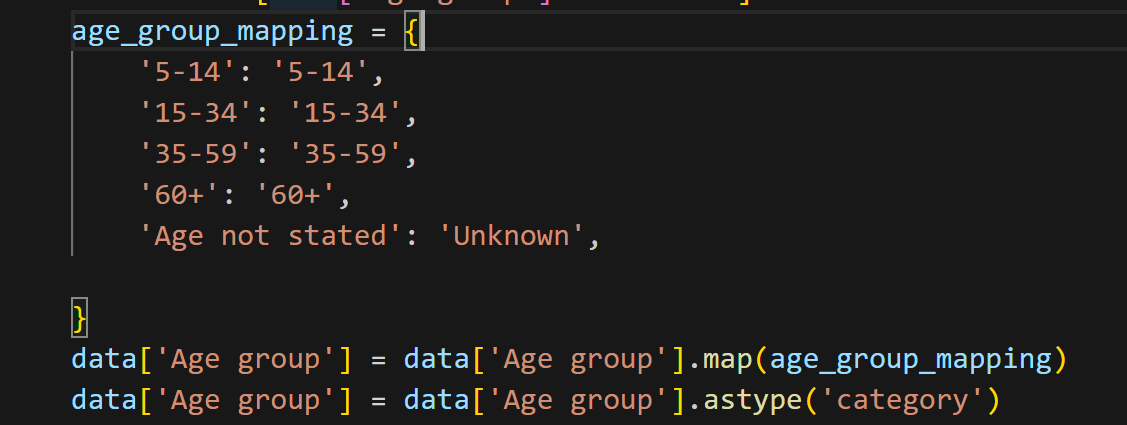
- "5-14" remains "5-14"

- "15-34" remains "15-34"

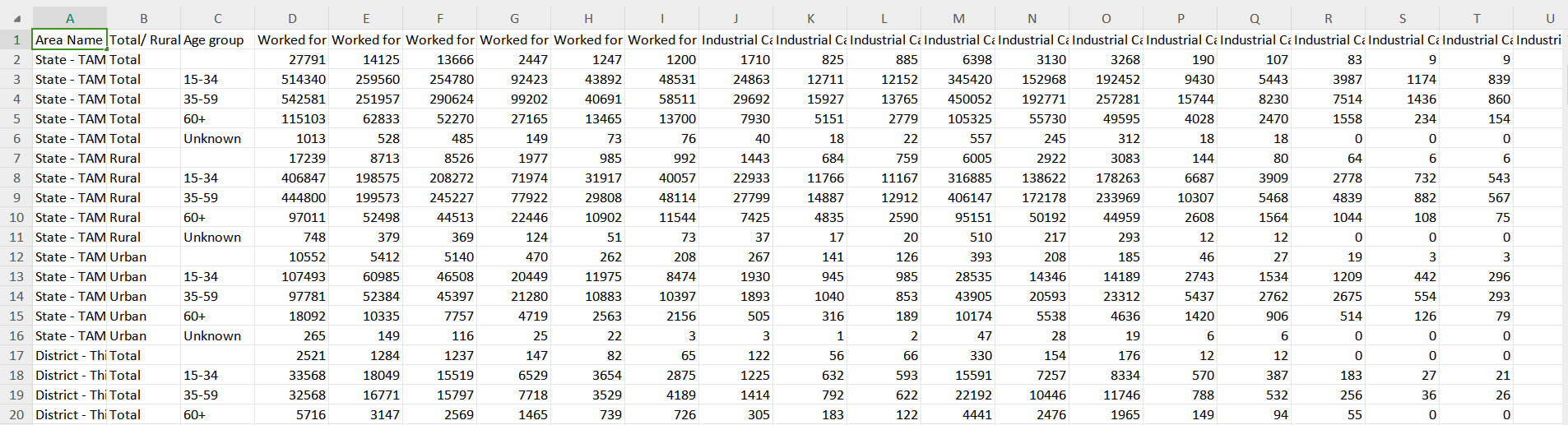
- "35-59" remains "35-59"

- "60+" remains "60+"

- "Age not stated" has been replaced with "Unknown" to address missing or unspecified age group entries.



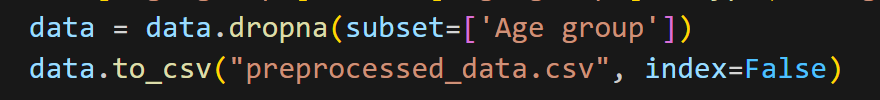
**OUTPUT:**

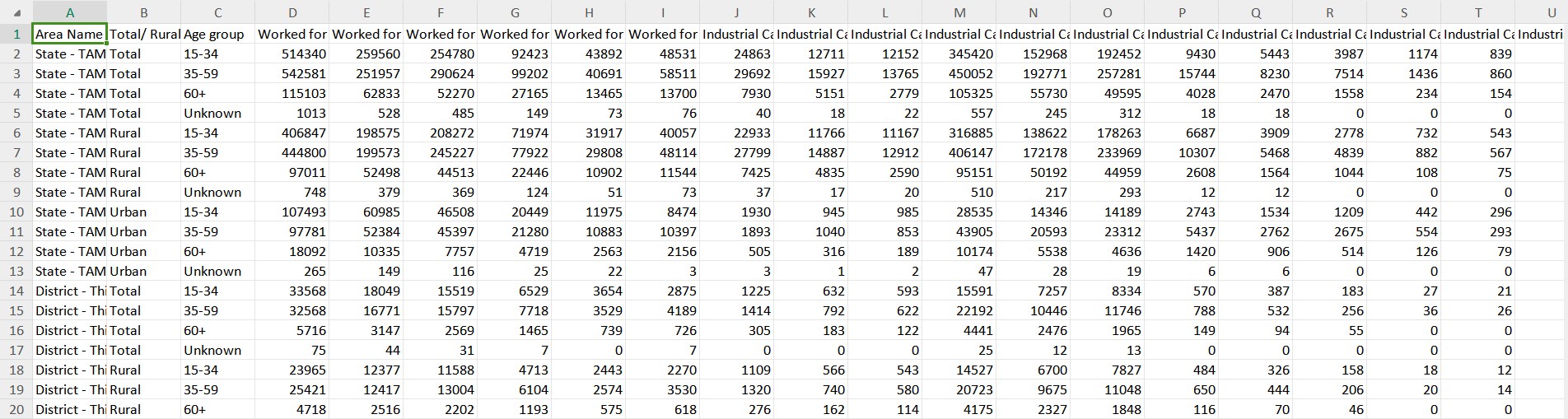


**Step 5:** Removal of Rows with Empty Entries in Age Group and Data Saving

- Rows containing empty or missing entries in the "Age Group" column were removed to maintain data consistency.

- The preprocessed data was saved to a new CSV file, "preprocessed\_data.csv," making it suitable for further analysis.





***Conclusion***

The data preprocessing steps have been executed successfully, resulting in a clean and well-organized dataset. This processed data is now prepared for various analytical tasks, such as statistical analysis, data visualization, and more in the context of the "TN Marginal Workers Assessment" project. The dataset exhibits enhanced data quality, ensuring reliable and informative results for future analysis.