Project Report - Phase 2: Data Collection and Preprocessing

Author: Abhay Date: August 6, 2025 Project: Employee Productivity Prediction

2.1. Data Sourcing and Initial Exploration

The foundation of this project is the garments_worker_productivity.csv dataset. This dataset contains 1,197 records and 15 distinct attributes related to the work of employees in a garment factory. An initial exploratory data analysis (EDA) was conducted to understand the structure, data types, and statistical properties of the dataset. This involved examining the distribution of key variables and identifying any immediate data quality issues.

2.2. Data Cleaning and Transformation

To prepare the data for machine learning, several cleaning and transformation steps were performed:

- **Standardization of Categorical Data: The department column contained inconsistencies such as "sweing" and "finishing ", which were corrected to "sewing" and "finishing" respectively. This ensures that the model treats these categories correctly.**
- **Handling of Missing Values: The wip (work in progress) column was identified as having a significant number of missing values. For the final model, this column was dropped from the dataset to maintain data integrity.**
- **Feature Engineering: The date column was leveraged to create a new, more useful month feature. This was done to capture any potential seasonal effects on productivity. After this, the original date colusmn was removed.**

2.3. Categorical Data Encoding

Machine learning models require all input features to be numerical. Therefore, the categorical columns (quarter, department, and day) were converted into a numerical format. The Label Encoding technique was used for this purpose, as implemented in the training notebook. This method assigns a unique integer to each category within a feature.

2.4. Final Dataset Preparation

After the preprocessing steps, the dataset was finalized and prepared for the model development phase. The data was split into features (X) and the target variable (y, which is actual_productivity). This clean and structured dataset formed the basis for training and evaluating the machine learning models in the next phase.