Data List and Analysis Report

|  |  |  |
| --- | --- | --- |
| Dataset Name | Source | Description |
| combined\_dataset | Kaggle - [Technician Scheduling Dataset](https://www.kaggle.com/datasets/ziya07/technician-scheduling-dataset) | This dataset contains task assignment and completion records for field technicians. Each row represents a unique task assigned to a technician, capturing key attributes related to both the task and the technician’s suitability for it. |
| enhanced\_combined\_dataset | Kaggle - [Technician Scheduling Dataset](https://www.kaggle.com/datasets/ziya07/technician-scheduling-dataset) | This dataset captures detailed information about completed task assignments and their outcomes, focusing on operational logistics, task requirements, and customer satisfaction. Each row represents a single task, including attributes that describe the context of the assignment, resource needs, and the results as perceived by the customer. |

Initial Analysis:

* Small rows of data – there are only 50 rows of data in combined\_dataset and 200 rows of data for enhanced\_combined\_dataset.
  + The 50 rows of data will be used as training data, so a small number of training data means the model will have less data to learn on so the chances of accuracy is lower.
* Lack contextual data
* Inconsistency in columns in both datasets
* No missing data
* Very clean, mostly numeric data type columns which means less standardization and transformation required for analysis.

Insights and Trends Analysis Report

**Purpose:**  
To synthesize your EDA findings, focusing on discoveries, patterns, and how the data aligns with project goals.

**Typical Contents:**

* **Key Trends:** Major patterns, distributions, outliers, correlations, or anomalies.
* **Insights:** What the data reveals relative to your project objectives.
* **Data Quality Assessment:** Is the data sufficient, clean, and reliable?
* **Alignment with Objectives:** Does the data support your initial hypotheses or project aims?
* **Additional Data Needs:** Identify if further data is needed.
* **Recommendations:** Suggest pivots, new approaches, or alternative data sources if necessary.
* **Stakeholder Value:** How findings address or could better address stakeholder interests.
* Expertise match vs task completion
  + If expertise match increases, does it really mean task completion also increases?
  + So does better technican skill to task requirement compatibility matter?
* Expertise match vs Penalty Cost
  + Does it cost the company more if there is low matching because it means technicians are not prepared for the task and causing delays and penalty costs
* Customer Rating vs Task Completed
  + Do customers rate higher when tasks are completed
* Penalty Cost vs Priority
  + Data source claims penalty is based on task priority. Testing this if this is true
* Task Complexity vs Equipment Required
  + Are tasks considered more complex if an equipment is involved? Does this mean the company needs to focus on getting all techs trained with equipment
* # techs in each complexity level
* # techs in eqpt trained
* # tasks in each complexity level
* # tasks in eqpt required
* # tasks in distance
* # tasks in priority
* # tasks in customer rating
* # tasks in penalty cost

Data Management and Transformation/Engineering/Architecture Document

**Purpose:**  
To document how the data will be handled, processed, and maintained throughout the project lifecycle.

**Typical Contents:**

* **Dataset Inventory:** Detailed list with descriptions, sources, and key variables.
* **Data Usage Plan:** How each dataset will be used, merged, and joined (e.g., on customer ID, date).
* **Data Processing:** Required transformations, aggregations, cleaning steps.
* **Quality Management:** How you’ll handle missing, inconsistent, or poor-quality data.
* **Technical Architecture:** Platforms, tools, and data storage solutions to be used.
* **Deployment Considerations:** How data will be managed from initial exploration to full production use.

|  |  |  |
| --- | --- | --- |
| Dataset | Purpose | Key Variables |
| Combined\_dataset.csv ([Technician Scheduling Dataset](https://www.kaggle.com/datasets/ziya07/technician-scheduling-dataset/data)) | Acts as historical records of tasks each technician has been assigned to | TechnicalID - Tech unique ID  TaskID – task unique ID  Expertise Match – whether task was assigned to a similar complex level as technican can handle (1 = Yes, 0 = No).  Task Priority - Task priority on a scale of 1 to 5, where a higher value indicates greater urgency.  Task Completed - whether tasks were successfully completed (1 = Yes, 0 = No).  Distance to Task - Distance (in kilometers) from the worker’s location to the task's location. |
| Enhanced\_combined\_dataset.csv ([Technician Scheduling Dataset](https://www.kaggle.com/datasets/ziya07/technician-scheduling-dataset/data)) | Acts as testing data that describes incoming task requests that technicians must be assigned to | Distance to Task (km): Distance (in kilometers) from the worker’s location to the task's location.  Priority: Task priority on a scale of 1 to 5, where a higher value indicates greater urgency.  Task Complexity: Complexity rating of the task on a scale of 1 to 10.  Equipment Required : Binary indicator showing whether equipment is needed for the task (1 = equipment required, 0 = no equipment required).  Customer Rating: Customer satisfaction rating on a scale of 1 to 5.  Penalty for Delay ($): Monetary penalty based on task priority, which increases with higher priority tasks. |

* What data aggregations, transformations need to be accomplished with each data set to render it useful to the analysis and models?
  + Data Processing: Drop columns
* How will you manage data quality issues?
* To calculate success metrics:
  + Penalty cost
    - If task is completed, then cost accrued = 0
  + Customer Rating
    - If task is completed, then rating increase by 1 but no greater than 5
  + Expertise Match
    - Determine what complexity each Tech can handle – create a column for complexity level
    - Determine if they are equipment trained – create a column for eqpt trained
    - If complexity & eqpt match with tech’s skills, then expertise match = 1
* What kind of architecture and platform is needed to support this project (initially and then in full deployment)