Data List and Analysis Report

**Purpose:**  
To catalog all datasets you intend to use and provide an initial evaluation.

**Typical Contents:**

* **List of Data Sources:** Where each dataset comes from (with links or references).
* **Dataset Descriptions:** What each dataset contains, scope, and relevance.
* **Data Gaps & Risks:** Identify missing data, incomplete records, or limitations.
* **Impact of Data Gaps:** How these gaps or risks might affect your project.
* **Initial Analysis:** Early observations about quality or usability.
* **Aggregation of Module Files:** Combining previous modules’ work into one document/slides.

**Focus:**  
What data do you have? Where does it come from? Are there any initial concerns or limitations?

In your report you should identify all target datasets, and view the data and identify any data gaps, associated risks, the impact of these gaps on your final project etc. You can aggregate all of the Module 4 files into a slide deck or one document file, and I'll grade this across all of the 3 separate deliverables this week/module and place it in this folder please.

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.

**Focus:**  
What have you learned from the data so far? Are you on track to answer your project’s core questions?

Describe what trends and insights your data exploration has identified. Is this analysis supporting your expectations with respect to the project objective? Is the data of good quality and sufficient to merge and transform to support your efforts? Are other or additional datasets needed? Are correlations intuitive and is the data sufficient to support the initial project charter? Do you need to consider a different approach, different data, pivoting in scope to provide value related to the project objective to navigate any challenges that have been identified? If so how might you do this and what options that you have not yet considered might be viable that adequately address the stakeholder interest in the topic.

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.
* Describe the various datasets and provide links or files for each of these. List the dataset, source, what the dataset includes and key variables expected to be used from it.
  + 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
    - Key Variables include: TechnicalID, TaskID, Expertise Match, Task Priority, Task Completed, Distance to Task
  + 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
    - Key Variables include:
      * 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?
* How will you manage data quality issues?
* What kind of architecture and platform is needed to support this project (initially and then in full deployment)