**Axion Ray**

**Task 2**

1. **Primary Key Identification:**
2. The Primay Key I have identified is Order No.
3. I justify my selection of Order No as the Primary Key for several reasons:

**Uniqueness**: The Order No column appears to be unique in both datasets, with no duplicate values.

**Consistency**: The Order No column is consistently formatted in both datasets, with no missing or inconsistent values.

**Relevance**: The Order No column is relevant to both datasets, as it identifies the specific repair order or work order being referenced.

However, there are potential challenges in identifying a unique key:

**Data quality issues**: If the data is not accurate or complete, it may be difficult to identify a unique key.

**Data inconsistencies**: If the data is inconsistent in terms of formatting or naming conventions, it may be challenging to identify a unique key.

**Multiple potential keys**: If there are multiple columns that could potentially serve as a unique key, it may be difficult to determine which one is the most suitable.

**Data duplication**: If there are duplicate values in the dataset, it may be challenging to identify a unique key.

1. Data Cleaning:

* It inspects the dataset structure and checks for missing values.
* It fills missing values using forward fill and removes duplicates.
* It formats the Revenue and Cost columns to ensure they are of type float.
* If needed, it applies language translation to the Complaint column.

1. Data Integration:

* The two datasets are merged using an inner join on the identified primary key.

**Justification for Using Inner Join:**

1. **Complete Records**: An inner join only includes records that have matching values in both datasets. This ensures that only complete and relevant records are retained in the merged dataset, which is crucial for accurate analysis.
2. **Data Integrity**: By using an inner join, we avoid introducing incomplete or irrelevant data into the analysis.
3. **Focus on Relevant Data**: If the goal is to analyse only those orders that have been processed and have corresponding entries in both datasets, an inner join is the most appropriate choice.

**Implications of Using Other Join Types**

1. **Left Join**: A left join includes all records from the left dataset (repair data) and the matched records from the right dataset (work order data). If there is no match, the result is NULL on the right side.
2. **Right Join**: A right join includes all records from the right dataset (work order data) and the matched records from the left dataset (repair data). It is less common in this context unless the work order data is the primary focus and you want to ensure all work orders are included.
3. **Outer Join**: An outer join includes all records from both datasets, with NULLs in places where there is no match.

**Task 3**

1. **Trend Analysis:**

* **Total Revenue Over Time –** A line plot to show the Total Revenue over a time period.

A graph with numbers and a dot

AI-generated content may be incorrect.

* **Cost vs Actual Hours-**  A scatter plot showing the Cost with respect to Actual Hours

A graph with blue dots

AI-generated content may be incorrect.

* Heatmap of Failure Conditions vs. Costs

A white paper with black text

AI-generated content may be incorrect.

1. **Root Cause Identification:**

* Counts how many times each unique failure condition-component pair and fix condition-component pair happens (frequency).
* Helps stakeholders see which failures and fixes happen most and which are the most expensive.

**a. Failure Condition Insights:**

* **The costliest failure observed was "Leak - Fuel Filter", with an average cost of ₹15,887 despite occurring only once. This indicates a high impact but rare failure that needs urgent attention to prevent expensive downtime.**
* **Other significant failures by cost include:**
  + **"Oil Leak - No Component Mentioned, Damaged - Oil Seal" with an average cost around ₹8,557 (3 occurrences),**
  + **"Not Latched - Door" costing approximately ₹5,775 (1 occurrence),**
  + **"Broken - Harness" with moderate frequency (5 times) and an average cost of ₹5,136.**
* **Some failure conditions show negative average costs (around -₹59), which may indicate data entry errors, adjustments, or refunds that require data validation before further analysis.**

**b. Fix Condition Insights:**

* **The most expensive fixes align closely with costly failure conditions, such as:**
  + **"Quick Connect Software - Reassembled, O Ring - Replaced" averaging ₹8,557 (3 occurrences),**
  + **"NCV Harness - Replaced" averaging ₹6,434 (4 occurrences),**
  + **"Latch - Replaced, Latch - Realigned" averaging ₹5,775 (1 occurrence).**
* **Fixes with fewer counts but high average costs highlight critical repair processes that should be evaluated for efficiency or alternative solutions.**
* **Like failure data, some fix conditions show negative costs, indicating potential data inconsistencies needing review.**

**c. Recommendations:**

* **Prioritize prevention and early detection of costly failures like fuel filter leaks and oil leaks to reduce expensive repairs and downtime.**
* **Review and optimize repair procedures for expensive fixes such as quick connect software reassembly and harness replacements to improve cost-effectiveness.**
* **Investigate data quality issues reflected by negative cost values to ensure accurate reporting and reliable decision-making.**
* **Use the heatmap visualization (if shared) to identify specific failure-fix pairs driving the highest costs and tailor interventions accordingly.**