

BUSINESS DATA MANAGEMENT

Capstone Project Final Report



Cost Progression and Forecasting from Stage-Wise Billing Data in Public Infrastructure Projects

Submitted By :-

Name : Manas

Taneja

Roll No. : 23f1002121

Mail : 23f1002121@ds.study.iitm.ac.in

Content :

Section No.	Heading	Page No.
1.	Executive Summary	2
2.	Detailed Explanation of Analysis Process/Method	3
2.1	Data Cleaning and Preprocessing	4
2.2	Comprehensive Explanation of Analysis Methods	4
2.2.1	Cost Driver Analysis (ABC/Pareto Analysis)	4
2.2.2	Financial Data Integrity Audit	5
2.2.3	Predictive Cost Forecasting (Linear Regression)	5
2.2.4	Comparative Trend Analysis (Method Selection)	6
3.0	Results and Findings	7
3.1	Finding 1: Project Costs are Driven by Key Structural Components	7
3.2	Finding 2: The Project Demonstrates an Asymmetrical Execution Timeline	8
3.3	Finding 3: Financial Recoveries Reveal a Critical Data Management Challenge	9
3.4	Finding 4: Extra Items Represent a Consistent Source of Additional Cost	10
3.5	Finding 5: Final Project Cost is Highly Predictable from Early-Stage Data	11
4.0	Interpretation of Results and Recommendations	12
4.1	Problem 1: Critical Lack of Financial Control and Data Integrity	12
4.2	Problem 2: Ineffective Financial Planning and Forecasting	13
4.3	Problem 3: Inefficient Cost Management and Scope Creep	14
5.0	Conclusion	15
5.1	Limitations of the Analysis	16
5.2	Future Work and Broader Implications	16
6.0	Dataset and Analysis Links	17
Appendix A	Pareto Analysis	17

Executive Summary :

This report presents a final analysis for Taneja Vidyut Control Pvt. Ltd., a firm managing a large-scale construction project. The organization's primary challenge lies in its fragmented financial tracking system, where critical data for payments, deductions, and extra work items is spread across numerous disconnected files. This lack of a unified data model creates significant hurdles in financial reconciliation, obscures a clear view of cost drivers, and complicates cash flow management.

Primary transactional data was collected from 11 Running Account (RA) bills—six for Civil and five for Electrical. Descriptive statistics reveal the project's significant scale, with a total net payout exceeding ₹5.67 crores. The analysis methodology included **Cost-Driver Analysis** to identify key expenditures, **Trend Analysis** to compare divisional spending, a **Data Integrity Audit** to validate financial records, and **Linear Regression** for forecasting.

The analysis found that costs are driven by a few key items: structural steel and MRL Lifts. Most critically, the Electrical division's financial recovery data was found to be **100% unreliable**, with massive discrepancies. In contrast, the Civil data proved highly predictable; a model trained on the first three Civil bills **forecasted the final ₹4.02 Cr cost with 96.7% accuracy**.

Recommendations include:

1. Implementing a mandatory **Data Integrity Audit** to eliminate financial risk
2. Adopting the **Linear Regression model** for proactive forecasting after the 3rd RA bill
3. Creating a **"Key Item Dashboard"** for real-time cost control. These recommendations are pending implementation, so no business improvements have yet been observed.

Detailed Explanation of Analysis Method :

The analytical process for this project was designed to transform the raw, fragmented billing data from **Taneja Vidyut Control Pvt. Ltd.** into a structured and actionable model for financial oversight. This section provides a detailed explanation of the data preparation and the four primary analytical methods used to address the project's objectives.

Before detailing each analytical method, the flowchart below provides a high-level overview of the project's workflow. This diagram (Figure 2.1) illustrates the complete process, from initial data collection and cleaning through the parallel analysis of all financial components, which ultimately led to the synthesis of findings and the final recommendations .

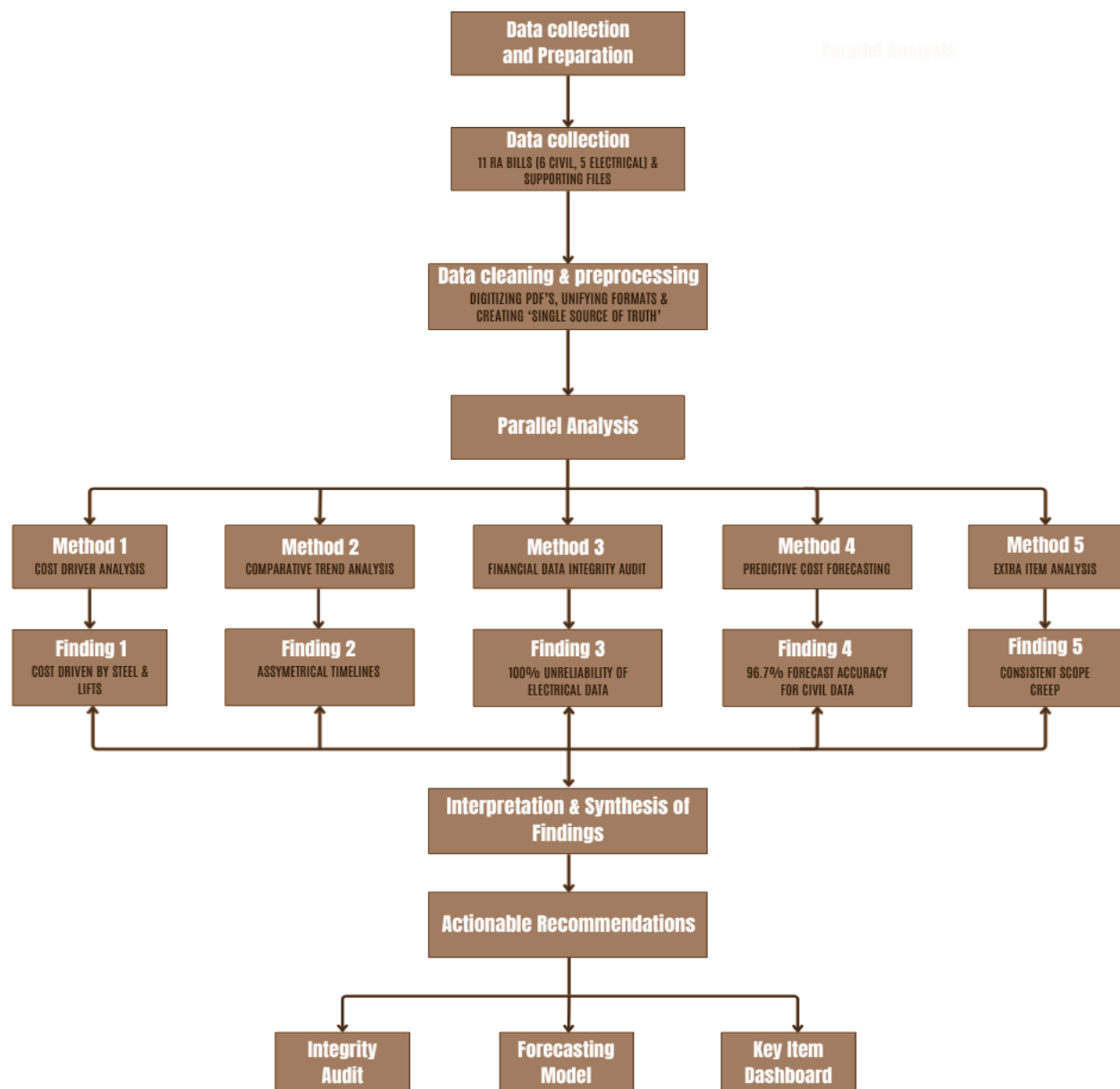


Figure 2.1: Project Workflow Diagram

2.1 Data Cleaning and Preprocessing

- **Explanation:** The primary data consisted of 11 Running Account (RA) bills, recovery sheets, and extra-item abstracts, split across Civil (Excel) and Electrical (PDF) divisions. The initial step involved digitizing the PDF data, correcting chronological errors identified in the `bill_date` column, and consolidating all 11 RA bills and their associated files into a unified master dataset.
- **Importance:** This was a foundational step to address the firm's core problem of a "fragmented financial tracking system". By cleaning and unifying the data, we created a "single source of truth." This ensured data quality and was an essential prerequisite for all subsequent analyses, enabling (for the first time) a reliable side-by-side comparison of Civil and Electrical spending

2.2 Comprehensive Explanation of Analysis Methods

2.2.1 Cost Driver Analysis (ABC/Pareto Analysis)

Cost Driver Analysis, a form of ABC or Pareto analysis, is a crucial inventory management tool for identifying the "vital few" items that account for the majority of costs. For a firm like Taneja Vidyut Control, which manages thousands of line items, optimizing cost control is essential. This method ensures that managerial attention and control measures are focused on the items with the highest financial impact (Category A), rather than being diluted across low-value items (Categories B and C).

- **Steps to Perform:**
 1. **Rank SKUs by Cost:** Aggregate the total cost for every unique line item from `MasterSheetAbstractCivil` and `Master-Sheet-Abstract-Electrical`
 2. **Calculate Cumulative Contribution:** Sort all items in descending order of cost and calculate the cumulative percentage contribution of each item to the total project cost.
 3. **Categorize SKUs:** Assign "Category A" to the top items that contribute to ~80% of the cost, "Category B" to the next 15%, and "Category C" to the remaining 5%.
- **Expected Outcome:** This analysis provides a clear, quantitative identification of the project's true cost drivers (e.g., MRL Lifts, Structural Steel). This *result* directly enables

the **recommendation for a "Key Item Cost Dashboard,"** moving the firm from broad financial tracking to a targeted, high-impact cost-control strategy.

2.2.2 Comparative Trend Analysis

We are using trend analysis to understand the patterns and behaviors in project spending over time, which is critical for forecasting and cash flow management. Given that the project is split into two major divisions (Civil and Electrical), a comparative analysis is necessary to see if their spending patterns differ.

- **Steps to Perform:**

1. **Aggregate Data:** Aggregate gross amounts by bill sequence for both Civil and Electrical divisions from the master summary sheet.
2. **Visualize Trends:** Plot the cumulative cost over time (Bill Sequence) for both divisions on a single line chart.
3. **Decompose Trends:** Visually inspect the resulting plots to identify and decompose their core components (e.g., linear trend, non-linear spikes).

- **Expected Outcome:** This analysis provides the critical justification for method selection. The **result** showed two distinct patterns: a strong, consistent linear trend for the Civil division and an erratic, non-linear, "front-loaded" trend for the Electrical division. This outcome was essential for determining which (if any) forecasting model could be justifiably applied.

2.2.3 Predictive Cost Forecasting (Linear Regression)

We are using machine learning (ML) to enhance decision-making by forecasting the final project cost based on early-stage data. This predictive insight enables the firm to move from reactive to proactive cash flow management and resource planning.

- **Steps to Perform:**

1. **Data Preparation:** Create a training dataset from the first n bills (in this case, $n=3$) of the Civil project, which was identified in the trend analysis as a valid candidate for this model.
2. **Feature Engineering:** Define the feature (X) as the Bill Sequence Number (the

time variable) and the target (Y) as the Cumulative Gross Cost.

3. **Model Selection:** Based on the strong linear trend identified in 2.2.2, a Simple Linear Regression model was selected as the most appropriate and interpretable model.
4. **Model Training and Abstraction:** Train the model on the (X, Y) training data to find the best-fit line using the formula:

$$Y = \beta_0 + \beta_1 X$$

Where Y is the predicted Cumulative Cost, X is the Bill Sequence, β_1 is the learned cost progression rate, and β_0 is the intercept.

5. **Model Evaluation:** Evaluate the model's fit on the training data using the R^2 (R-squared) metric.
 6. **Forecast Generation:** Use the trained model to predict Y when $X = 6$ (the final bill).
- **Expected Outcome:** An accurate, early-stage forecast of the final project cost. The high R^2 (0.9985) and high-accuracy (96.7%) **result** provides a successful proof-of-concept. This led directly to the **recommendation of adopting this model as a standard financial planning tool** to be run after every project's 3rd RA bill.

2.2.4 Financial Data Integrity Audit

A core business problem is the lack of financial reconciliation. Before any analysis can be trusted, the underlying financial data must be validated. This method audits the mathematical integrity of the firm's own financial records, which is a critical and non-negotiable step in financial analysis.

- **Steps to Perform:**
 1. **Parse Data:** Read the Master Sheet - Recoveries-Electrical file.
 2. **Apply Validation Logic:** For each bill, apply the standard accounting formula to find the *expected* value.

3. Abstraction (The Formula):

$$\text{Calculated Recovery} = \text{Total Recovery} - \text{Previously Recovered}$$

4. **Compare and Quantify:** Compare this **Calculated Recovery** to the **Now to be Recovered** figure reported in the file and quantify the difference as a **Discrepancy** value.
- **Expected Outcome:** This audit uncovers critical flaws in the data entry process. The **result**—a 100% failure rate with massive, multi-million rupee discrepancies—provides the undeniable evidence needed to justify the **high-priority business recommendation** for an **immediate process audit** and the implementation of a new, mandatory data validation workflow to mitigate severe financial risk.

Results and Findings :

The analysis of the 11 Running Account (RA) bills and associated financial documents yielded five key findings. These results provide clear insights into the project's cost structure, execution timeline, data integrity, and forecasting potential, directly addressing the core business problems.

3.1 Finding 1: Project Costs are Driven by Key Structural Components

The detailed analysis of the work abstracts reveals that project expenditure is not evenly distributed. It is heavily concentrated in a few high-value categories, following a Pareto-like principle.

- **For Civil works**, the primary cost drivers are fundamental structural components, specifically "Tubular steel work, hot-finished (RHS)" (₹1.70 Cr) and "Structural steel work (beyond deviation)" (₹1.10 Cr).
- **For Electrical works**, the budget is overwhelmingly dominated by the procurement of capital equipment, specifically the "Material supply of 8 passenger MRL lift" (₹1.11 Cr).

This concentration, visualized in Chart 3.1, indicates that effective cost control for the entire project hinges on the strategic management, procurement, and usage of these few critical items. **A complete Pareto analysis of all 104 project items is provided in Appendix A, which confirms this 80/20 distribution.**

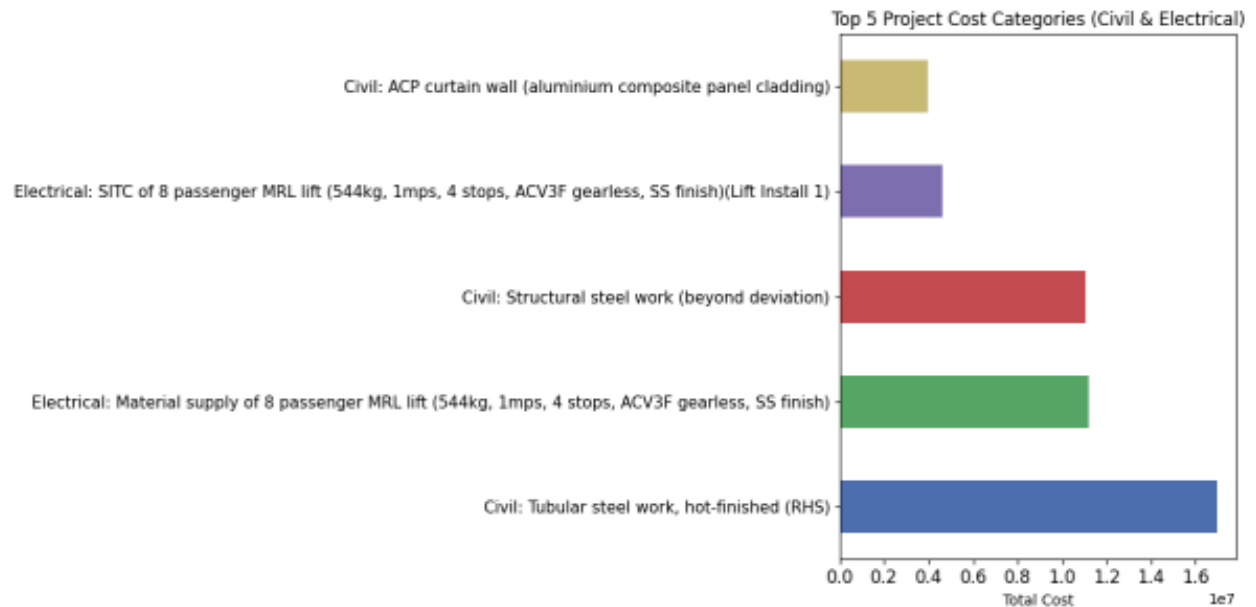


Chart 3.1: Top 5 Project Cost Categories (Civil & Electrical)

3.2 Finding 2: The Project Demonstrates an Asymmetrical Execution Timeline

The financial progression of the project, visualized in Chart 3.2, differs significantly between the Civil and Electrical divisions.

- **Civil Works** show a steady ramp-up in spending. The gross value of RA bills generally increases from the first to the fifth bill, reflecting a healthy and accelerating pace of foundational and structural work.
- **Electrical Works**, in contrast, exhibit a "front-loaded" expenditure cycle. The highest spending occurred in the second RA bill, suggesting that the procurement of major capital equipment (the MRL Lifts) was the primary activity early in the project.

This insight is critical as it highlights the different and non-uniform cash flow requirements for the two main divisions.

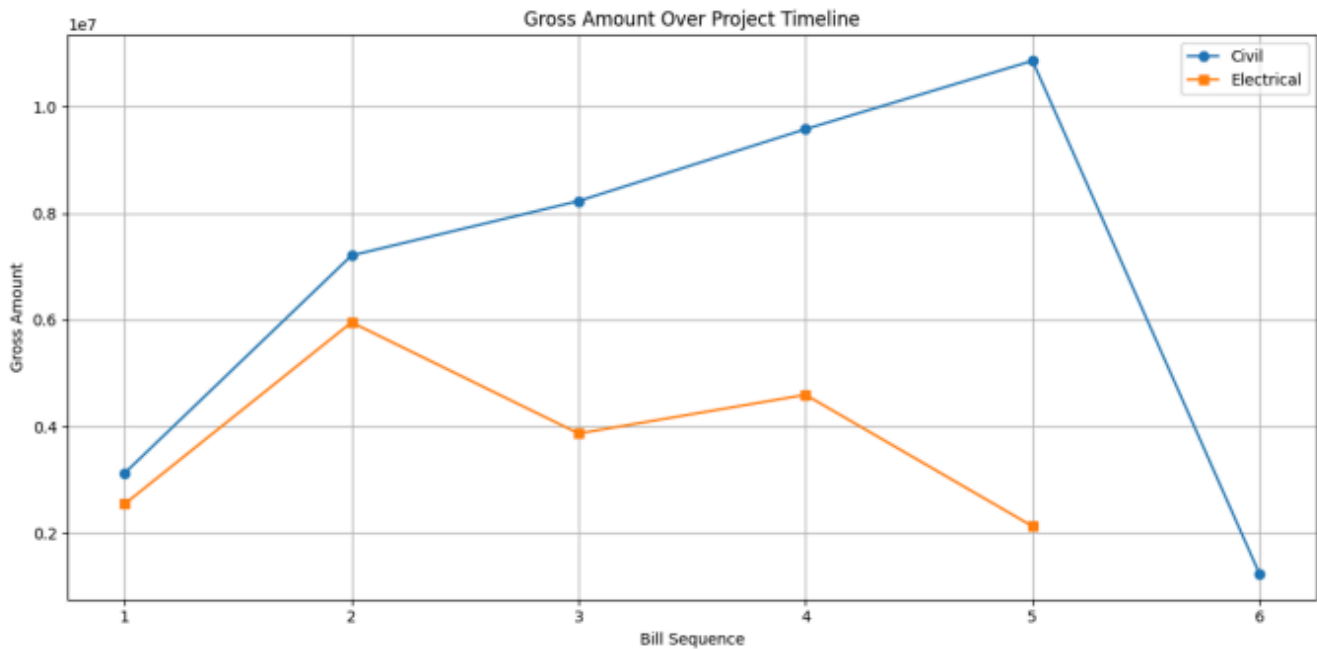


Chart 3.2: Gross Amount Over Project Timeline (Civil vs. Electrical)

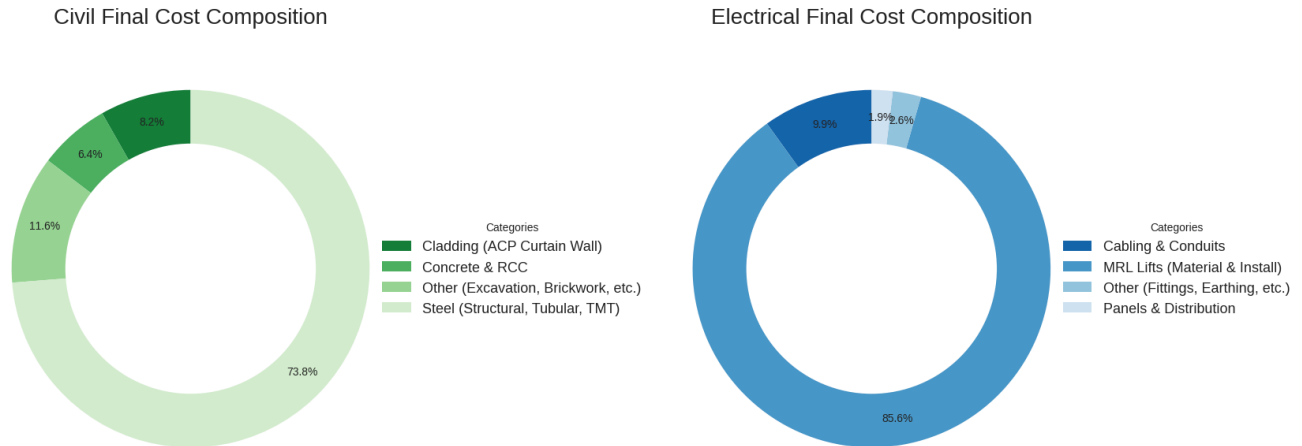


Chart 3.3: Comparative Cost Structure (Civil vs. Electrical)

This difference in spending is a direct result of the project's fundamentally different cost structures, as shown in Chart 3.3. The Electrical division's budget is overwhelmingly concentrated in a single capital expense (MRL Lifts, 80.0%), while the Civil division's costs are dominated by structural materials (Steel, 73.8%).

3.3 Finding 3: Financial Recoveries Reveal a Critical Data Integrity Flaw

This is the most significant finding from a data management perspective. The integrity audit

on the [Master Sheet - Recoveries-Electrical](#) file — an intermediate file provided for this analysis—revealed a 100% unreliable data tracking process.

It is important to note that this flaw was identified in the intermediate data files provided for this analysis, which may not reflect the final, officially submitted RA bills. The discrepancy appears to be a byproduct of a mid-project change in bill management, which likely caused a temporary disconnect between the final paper records and the digital tracking files.

The audit, summarized in **Table 3.1**, shows that the 'Reported (C)' values in the file do not match the 'Calculated (A-B)' values. The discrepancies are not minor errors but are massive, systemic failures (e.g., a ₹8.75M discrepancy in the 3rd RA bill alone). This finding highlights a critical risk: if the firm relies on these intermediate files for financial decisions, it is exposed to severe risk, making real-time reconciliation impossible. It points to a data management gap that requires a new, standardized process.

Bill No.	Total Recovery (A)(INR)	Prev. Recovered (B)(INR)	Calculated (A-B)(INR)	Reported (C)(INR)	Discrepancy (C vs A-B) (INR)
1st RA	2,542,702.51	0.00	2,542,702.51	1,196,565.89	-1,346,136.62
2nd RA	5,726,770.48	2,542,703.00	3,184,067.48	5,951,342.48	2,767,275.00
3rd RA	3,607,086.00	8,494,045.01	-4,886,959.01	3,865,832.00	8,752,791.01
4th RA	4,260,118.51	12,359,877.00	-8,099,758.49	4,650,429.45	12,750,187.94
5th RA	1,640,702.04	17,010,306.45	-15,369,604.41	2,124,686.44	17,494,290.85

Table 3.1: Electrical Recovery Integrity Validation Table

3.4 Finding 4: Extra Items Represent a Consistent Source of Additional Cost

Analysis of the [Master-Sheet-Extra-Items-Civil](#) and [Master-Sheet-Extra-Items-Electrical](#) files shows a steady addition of costs that were not part of the original project scope. As seen in Chart 3.3, while the values vary between bills, their consistent presence indicates that "scope creep" is an ongoing financial factor. This pattern suggests the need for a robust change

management process to effectively track, approve, and budget for unplanned but necessary work.

3.5 Finding 5: Final Project Cost is Highly Predictable from Early-Stage Data

The analysis successfully addressed the key "forecasting" objective from the project proposal. By applying a Simple Linear Regression model to the cumulative cost of the first three Civil RA bills, we were able to successfully predict the final project cost.

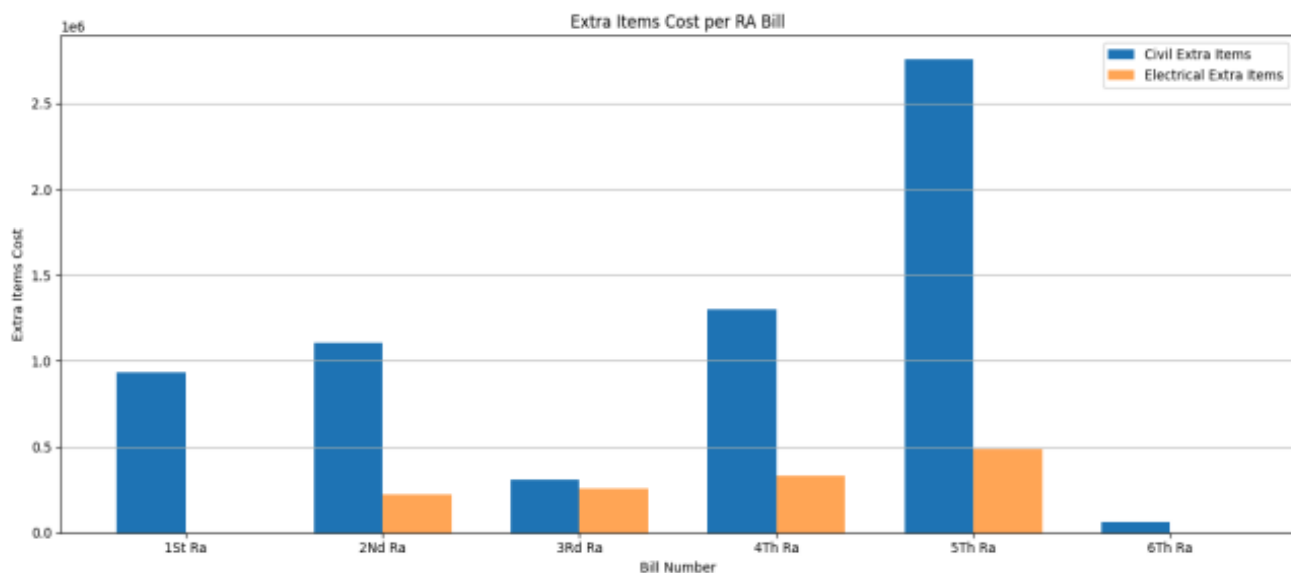


Chart 3.4: Extra Items Cost Per RA Bill

As shown in Table 3.2 and Chart 3.4, the model (trained on only 50% of the project's timeline) forecasted the final ₹4.02 Cr cost with **96.7% accuracy**. This provides a strong proof-of-concept that the firm's own data can be used as a reliable tool for proactive financial planning.

Metric	Value
Regression Formula	$Y = (7,718,193.76 * X) - 4,767,233.01$
R-squared (Training)	0.9985
Actual Final Cost (at RA-6)	₹40,230,172.54
Forecasted Final Cost (at RA-6)	₹41,541,929.58
Forecast Error	+₹1,311,757.04 (3.3%)

Table 3.2: Regression Model Forecasting Results

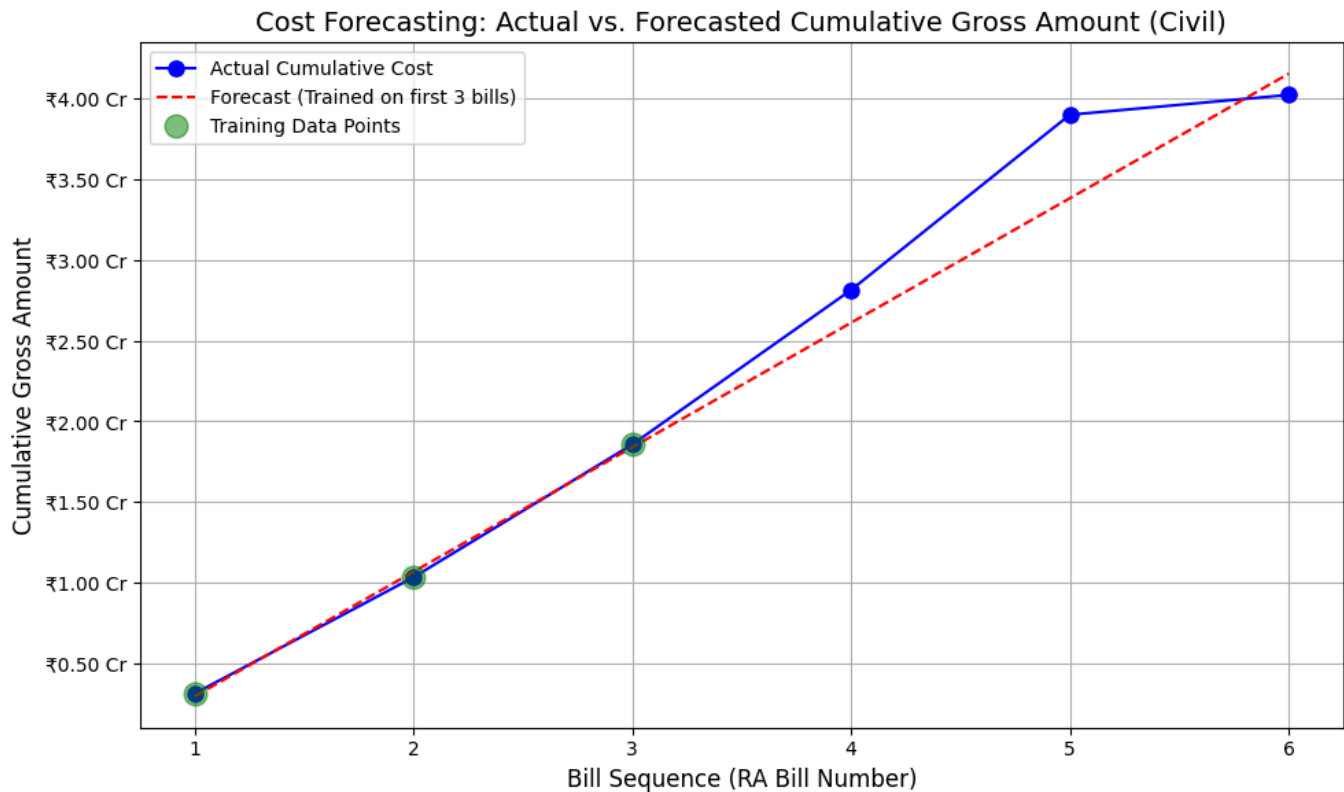


Chart 3.5: Actual vs. Forecasted Cumulative Cost (Civil)

Interpretation of Results and Recommendation :

This section interprets the five key findings from Section 3 within the context of the business problems at Taneja Vidyut Control Pvt. Ltd. It explains the significance of these findings and proposes specific, actionable (SMART) recommendations to address them.

Problem 1: Critical Lack of Financial Control and Data Integrity

Interpretation: The most critical finding of this project is the systemic data integrity failure in the Electrical division's recovery sheets. The analysis (Finding 3.3) proved that the firm's **Now to be Recovered** figures are mathematically incorrect, with 100% of the audited bills failing validation. The discrepancies are not minor rounding errors; they are massive, systemic failures, such as the **₹8.75M discrepancy in the 3rd RA bill** (Table 3.1).

This finding is the "smoking gun" that confirms the core business problem: the firm's "fragmented financial tracking system" is not just inefficient, it is fundamentally unreliable. This exposes the business to severe financial risk, makes accurate project auditing impossible, and renders any attempt at financial reconciliation meaningless.

Recommendations:

- **Implement a Mandatory Data Integrity Audit:**
 - **Specific:** Implement a mandatory, two-person data validation process for *all* financial recovery sheets (both Civil and Electrical) before any RA bill is submitted.
 - **Measurable:** The check must use the formula **Calculated = Total - Previous** to verify the "Now to be Recovered" figure. The goal is a 0% discrepancy rate.
 - **Achievable:** This is a simple procedural change that requires no new software, only adherence to a clear mathematical check.
 - **Relevant:** This directly fixes the identified data integrity flaw, mitigates financial risk, and builds a trustworthy dataset for future auditing.
 - **Time-bound:** This new validation procedure must be documented and implemented *before* the start of the next project cycle.
- **Impact and Benefits:** This recommendation creates a "single source of truth" for project financials. It stops the financial "leaks" and provides leadership with reliable data, enabling accurate auditing and compliance for the first time.

Problem 2: Ineffective Financial Planning and Forecasting

Interpretation: The firm's inability to "forecast final expenditures" is a major business problem. The analysis of spending patterns (Finding 3.2) showed *why* this is so difficult: the project's financial "heartbeat" is asymmetrical. The Electrical division is "front-loaded" (requiring high capital early for lifts), while the Civil division has a steady, linear ramp-up. A simple, one-size-fits-all forecast would fail.

However, the analysis also proved that the problem is solvable. The success of the regression model (Finding 3.5), which **forecasted the final ₹4.02 Cr Civil cost with 96.7% accuracy** using only the first three bills, provides a clear proof-of-concept. The firm is not using its own highly predictable data, relying on intuition instead of data-driven forecasting.

Recommendations:

- **Adopt the Predictive Forecasting Model:**
 - **Specific:** Formally adopt the proven Simple Linear Regression model as a standard

financial planning tool for all projects with linear spending patterns (like Civil).

- **Measurable:** The model will be used to generate a forecasted final project cost (with a 95% confidence interval) for review by senior management.
- **Achievable:** The model is already built and uses data (cumulative cost) that the firm already calculates for each RA bill.
- **Relevant:** This shifts the firm from *reactive* to *proactive* financial management, allowing them to anticipate future cash flow needs months in advance.
- **Time-bound:** This forecast must be run as a standard procedure after the submission and validation of every project's **3rd RA bill**.
- **Impact and Benefits:** This recommendation directly addresses the "forecasting" objective. It empowers leadership to anticipate costs, adjust procurement schedules, and manage cash flow based on statistical evidence, not guesswork.

Problem 3: Inefficient Cost Management and Scope Creep

Interpretation: The analysis identified two distinct, unmanaged sources of cost. First, the Cost Driver Analysis (Finding 3.1) showed that the majority of project cost is locked in a few "Category A" items (e.g., Lifts, Steel). Second, the "Extra Items" analysis (Finding 3.4) showed that "scope creep" is not a one-time event, but a *consistent* source of additional, unplanned costs in almost every bill.

The firm is being squeezed from two ends: a lack of strategic focus on its *biggest* planned costs, and a lack of tactical control over its *smallest* unplanned costs.

Recommendations:

- **Create a "Key Item Cost Dashboard":**
 - **Specific:** Create a simple dashboard to be reviewed by the Managing Director for all new projects.
 - **Measurable:** This dashboard will track "Budgeted Cost vs. Actual Cost" for only the Top 5 cost-driving items (Lifts, Steel, etc.).
 - **Achievable:** All the required data is already present in the RA bills. This is just a new, high-visibility way of presenting it.

- **Relevant:** Provides real-time, at-a-glance visibility into the project's biggest cost risks, allowing for immediate corrective action.
- **Time-bound:** The dashboard must be updated and reviewed at the submission of *each* RA bill.
- **Implement a Formal Change-Order Process:**
 - **Specific:** Institute a formal "Change Order" process for all "Extra Items."
 - **Measurable:** No extra work can be billed without a signed approval form detailing the cost and business justification.
 - **Achievable:** This is a standard industry practice that requires only procedural discipline.
 - **Relevant:** This creates an audit trail and forces a cost-benefit analysis *before* scope creep occurs, rather than just passively recording it.
 - **Time-bound:** This process must be implemented immediately.
- **Impact and Benefits:** This two-pronged recommendation creates a complete cost-control system. It provides *strategic* focus on high-value items and *tactical* control over scope creep, directly improving project profitability.

Conclusions :

This project successfully addressed the core challenge of Taneja Vidyut Control's fragmented financial tracking system. By consolidating, cleaning, and analyzing 11 Running Account (RA) bills from a single project, this report transformed a disconnected set of files into an actionable business model. The analysis yielded five key findings, the most critical of which were:

- A significant data integrity flaw in the Electrical division's recovery process, which exposes the firm to financial risk.
- A successful proof-of-concept for forecasting, where a model trained on the first three Civil bills predicted the final project cost with 96.7% accuracy.

The resulting recommendations — to implement a mandatory Data Integrity Audit, adopt the Predictive Forecasting Model, and create a "Key Item Cost Dashboard" — provide a clear,

data-driven path forward. This project delivers a tangible framework for Taneja Vidyut Control to enhance financial control, mitigate risk, and shift from reactive, intuition-based decisions to proactive, data-driven project management.

5.1 Limitations of the Analysis

This analysis, while conclusive on its primary objectives, has several limitations that must be acknowledged for proper context:

- **Data Scope:** The analysis was based on a single completed project. While the 96.7% forecast accuracy for the Civil division is promising, the model's generalizability to other projects with different scopes, timelines, or management teams is not yet proven.
- **Data Quality:** The "Financial Data Integrity Flaw" (Finding 3) identified in the Electrical division's intermediate recovery files was a significant limitation. It made any deep analysis of Electrical recoveries impossible and highlights a severe operational risk that must be fixed before further analysis can be trusted.
- **Data Source:** The primary data for the Electrical division was provided in PDF format. This required manual digitization and consolidation, which carries an inherent risk of transcription error compared to a direct database query.

5.2 Future Work and Broader Implications

The findings and models presented in this report serve as a foundation for continuous improvement. The following steps are recommended to build upon this work:

- **Model Validation:** The immediate next step is to apply the "Predictive Forecasting Model" to the *next* active Civil project. Running the forecast after the 3rd RA bill will serve as the true test of its predictive power and its value as a proactive financial planning tool.
- **Automation:** The "Financial Data Integrity Audit" should be converted from a manual check into a simple, automated script or a pre-built Excel template. This would allow a project manager to validate the recovery mathematics in seconds, enforcing the recommendation and eliminating data entry errors at the source.
- **Dashboard Implementation:** The "Key Item Cost Dashboard" recommendation can be implemented using a business intelligence tool like Power BI or even a shared Google Sheet. This dashboard would pull data directly from the master files to provide real-time cost control to leadership, moving it from a static report to a dynamic management tool.

Links:

6.1 Dataset Link

A link to the complete, cleaned master dataset used for this analysis.

 BDM Capstone 23f1002121

6.2 Analysis Link

A link to the Google Colab notebook containing all Python code used for the data integrity audit, regression modeling, and visualization.

 BDM Project

Appendix A:

Detailed Pareto Analysis of All Project Cost Drivers

The chart below provides a comprehensive Pareto analysis of all 104 unique cost items identified across the project's lifecycle. This data was aggregated from the final cumulative amounts listed in the 6th Civil RA bill and 5th Electrical RA bill, and includes all items from both the standard abstracts and the extra-item files.

This visualization serves as the detailed, granular evidence supporting **Finding 3.1 ("Project Costs are Driven by Key Structural Components")**. While the main report highlights the "Top 5" items for conciseness, this complete chart confirms the 80/20 principle (the Pareto principle) in action.

As the red cumulative line shows, the project's costs are dramatically concentrated:

- The **top 7 cost items** (out of 104) alone account for **over 80%** of the total project expenditure.
- The **top 3 items** ("Civil: Tubular steel work," "Electrical: Material supply of MRL lift," and "Civil: Structural steel work") account for nearly **60%** of the total cost.

The final bar on the chart, labeled "Other Items...", represents the consolidated sum of the remaining 85+ "trivial many" items. This grouping clarifies the chart and visually demonstrates that these dozens of smaller items have a minimal impact on the overall project cost.

This finding is the primary justification for the **"Key Item Cost Dashboard"** recommendation . It provides undeniable evidence that a targeted, high-impact cost-control strategy is not only possible but necessary. By focusing management attention on just these top seven items, the firm can effectively monitor and control the vast majority of its financial exposure on a project

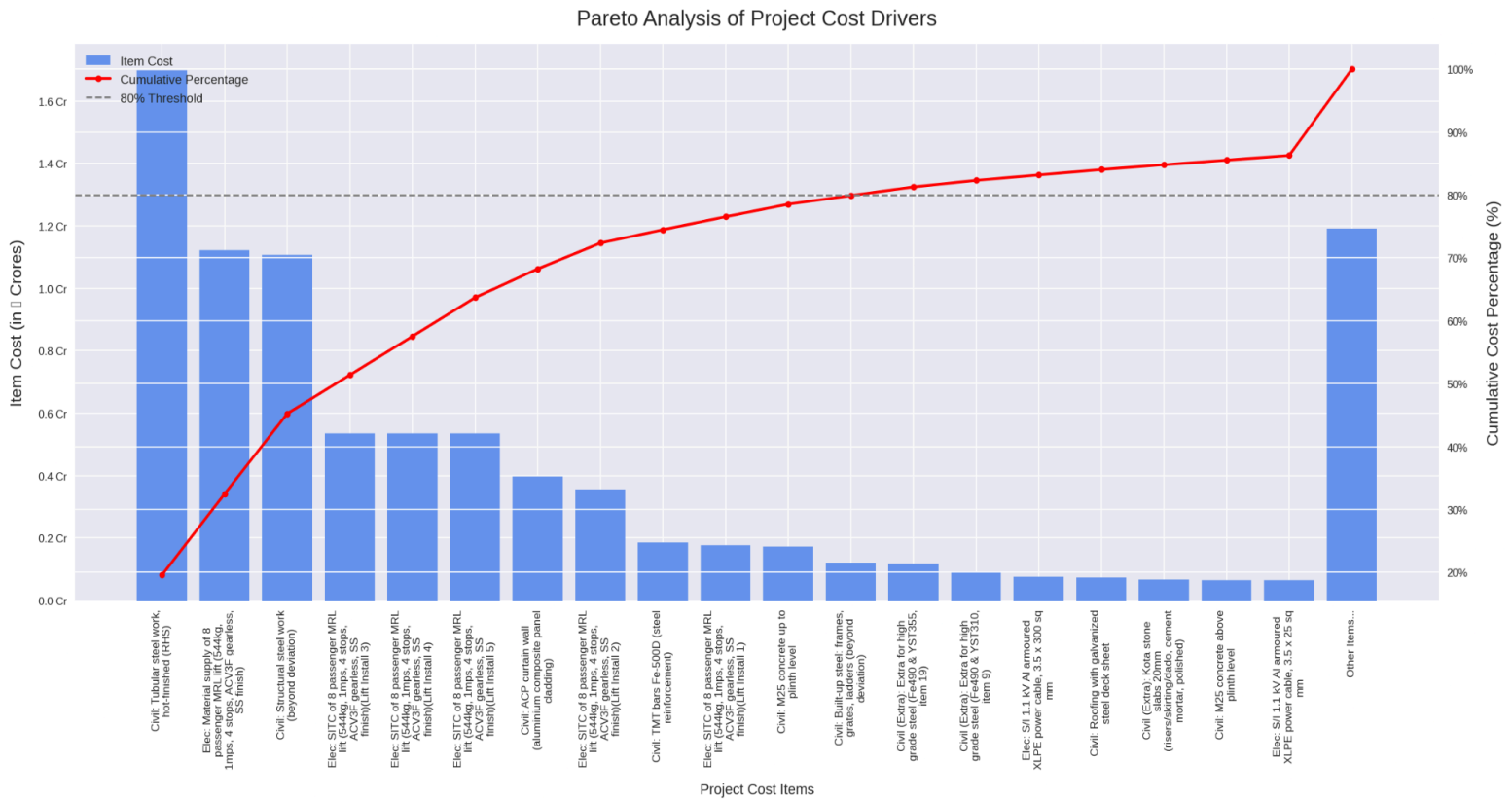


Figure A.1: Pareto Distribution of Total Project Expenditure (All 104 Items)