

Standardizing / Cleaning Dataset

Standardizing / Cleaning Dataset

Step 1: Identify and validate distinct company names

We ran a SQL query to list all distinct `company_name` values in the dataset, along with the number of projects associated with each name. This helps identify inconsistencies in spelling or formatting (e.g., “Emcor Group” vs. “EMCOR Group Inc.”) that can prevent accurate aggregation.

To ensure completeness, we also used the `WITH ROLLUP` function to include a total project count at the bottom, confirming that all rows in the dataset were accounted for.

Step 2: Standardize company names using a mapping table

We created a mapping table that pairs each raw company name in the dataset with a standardized version. This allowed us to consolidate variations in spelling, punctuation, or abbreviations (e.g., “ABM Tech Solns” vs. “ABM Technical Solutions”) into a single, consistent name.

We then ran an `UPDATE` query on the main dataset to replace all raw company names with their standardized equivalents using this mapping table as a reference. This ensured that all projects associated with the same company are correctly aggregated under a single name.

Finally, we verified the results by grouping the cleaned dataset by `company_name` and counting projects, confirming that all variations had been successfully consolidated and the data was ready for reliable analysis, such as market share calculations over time.

Step 3: Review and identify inconsistencies in categorical columns

To ensure accurate aggregation and analysis, we wanted to check for inconsistencies in several categorical columns, such as `client_name`, `client_industry`, `sector`, and `project_status`.

We ran a SQL query that grouped each column’s distinct values along with the count of associated projects. By using a `UNION ALL` approach, we combined multiple columns into a single result set with three fields:

- `column_name` – the column being analyzed
- `value` – the distinct value in that column
- `project_count` – the number of projects associated with that value

This allowed us to quickly scan multiple columns at once and identify inconsistent entries (for example, **Amazon Web Services** vs. **Amazon Web Services.** or **City of Chicago - Water** vs. **Chicago city water dept.**) without running separate queries for each column.

The results from this step provided a **comprehensive overview of all variations** in these categorical fields, enabling us to plan the next step of standardization and cleaning.

Step 4: Normalize monetary values to USD

Before running analyses on revenue, costs, and margins, we standardized all CAD amounts to USD by multiplying them by 1.4. This ensures consistency across the dataset, allowing accurate comparisons and aggregations. We applied this conversion to all relevant numeric columns, including contract values, recognized revenue, backlog, costs, gross margin, EBITDA, operating income, and average hourly rates, while leaving percentages and headcount columns unchanged.

Market Share and Revenue Trends

Step 1: Analyze top contractors and market share over time

To understand trends in market share and identify which companies dominate the MEP/construction space, we aggregated project-level revenue by company and year.

- We calculated the **total contract value per company per year** and derived each company’s **market share percentage** relative to the total revenue for that year.
- Using a ranking function, we selected the **top 3 companies for each year**, providing a focused view of the major players over the last decade.
- Revenue values were formatted in **currency (\$)** for readability, and market share percentages were rounded for clarity.
- The results were ordered with the **most recent years first**, making it easy to see current market leaders and observe how rankings have evolved over time.

Step 2: Identify top companies over the 10-year period

To understand overall market dominance, we aggregated **cumulative contract value for each company over the entire 10-year period**.

- We calculated **total revenue per company** and derived each company’s **market share as a percentage of the total market**.
- Using this, we identified the **top 3 companies by cumulative revenue**, providing a clear picture of which firms consistently commanded the largest share of the market.
- Revenue values were formatted in **currency (\$)** and market share percentages rounded for clarity.

	company_name	total_revenue_currency	market_share_pct_10yr
1	IES Holdings	\$1,868,870,684.00	9.06
2	Jacobs Mission Critical	\$1,782,264,996.00	8.64
3	Quanta Services	\$1,662,584,491.00	8.06

Step 3: Analyze revenue, project counts, and profitability by Client, Industry, and Subsector

We aggregated the dataset across three key dimensions — **client_name**, **client_industry**, and **subsector** — to understand revenue distribution, project volume, and financial performance. For each dimension, we calculated:

- **Total revenue** (normalized to USD) to measure absolute contribution.

- **Project count** to show relative volume and market activity.
- **Average revenue per project** to understand project size and client/industry/subsector significance.
- **Share of total revenue (%)** and **share of total project count (%)** to compare relative market positions.
- **Average gross margin, EBITDA margin, and project duration** to evaluate profitability and operational efficiency across clients, industries, and subsectors.

From this analysis, we observed that **profitability was relatively consistent across the dataset**, with no major differences between clients or industries. However, **private sector projects tended to be higher in average contract value**, and **large technology platforms such as Meta, Amazon, and Microsoft emerged as the highest-value customers**.

Step 4: Geography / Region Analysis

Columns: country, state, city, location_region

Process:

We analyzed MEP project trends by geography, looking at both state- and city-level activity. Using SQL, we aggregated total contract revenue and project counts by state and city. We also calculated the **share of revenue** and **share of projects** for each region to identify which areas contributed most to overall activity. Finally, we looked at trends over two multi-year periods (2016–2020 and 2021–2025) to identify emerging or declining regions. We also examined which companies dominated specific states to highlight regional leadership.

Findings:

- **Ontario** led in project activity from 2016–2020 but fell in the most recent period, replaced by **California** as the highest-revenue state.
- **New York** and **San Jose** were the cities with the highest project revenue in the period analyzed.
- Calculating revenue share by region helped highlight which companies were driving state-level activity and which regions contributed most to overall market share.

Step 5: Client Analysis & Revenue Concentration

We analyzed client performance across the last decade by evaluating total revenue, share of overall revenue, project volume, average project size, and EBITDA margin. This provided a clear picture of which customers drive the most value and whether profitability varies across the client base.

A key insight is that **Meta is the top client over the 10-year period**, generating the highest total revenue and the largest average project size. Despite the significantly higher dollar value

per engagement, **Meta's margins were broadly consistent with the rest of the dataset**, indicating that the higher revenue projects did not translate into disproportionately higher profitability. This suggests that scale-driven projects in the tech sector maintain similar margin structures to the broader MEP market.

Step 6 — Customer Concentration Summary

For each company, we calculated annual top-customer concentration based on **revenue dollars**, then averaged those percentages over the 10-year period. This metric reflects how dependent each business is on its single largest client in any given year.

Overall, the dataset shows **high top-customer concentration**, with many companies deriving a significant share of annual revenue from their largest customer. This indicates elevated customer dependency and potential revenue risk across the dataset.

Financial and Operational Trends

Step 1: Analyze Cost Structure by Company

We calculated the contribution of each cost category (labor, materials, equipment, subcontractor, overhead, SG&A, contingency) as a percentage of total project costs by company, including an overall average across all projects. The analysis shows that **labor and material costs make up over 70% of total average costs per project**, indicating that these are the primary drivers of spending. Across the dataset, **gross margins (~10%) and EBITDA margins (~6%) are low**, suggesting that most potential cost efficiencies in the MEP industry would need to come from **labor and material procurement** rather than other areas.

Step 2: Headcount & Productivity Analysis

We analyzed labor efficiency by company, calculating revenue and EBITDA per headcount and per labor hour. All monetary metrics were expressed in USD for consistency. The dataset shows a wide range of efficiency across companies, with **Mastec** achieving the highest revenue per headcount. This highlights differences in productivity and cost management, suggesting opportunities for optimizing field and office labor allocation across the industry.

Step 3: Time Analysis

We analyzed project activity and contract values across both quarters and years to identify temporal trends. First, we calculated the **10-year average by quarter**, observing that **Q3 consistently has the highest average revenue, most projects started, and shortest average duration**, indicating peak activity. Next, we examined **annual trends over the 10-year period**, tracking total revenue, project counts, and average duration to assess overall growth and fluctuations. These insights suggest that any client targeting MEP projects should focus sales efforts in Q3 to align with peak market demand.

Step 4: Competition & Joint Ventures

We analyzed project competitiveness and joint venture participation. Some companies consistently participate in highly competitive projects, while JV activity varies—certain firms take larger shares in partnerships. Overall, higher competition slightly correlates with lower margins, and JVs allow companies to share risk while maintaining project involvement.