# STORE SALES DIVE ANALYSIS AND ACTION PLAN

DN4: Rachakonda Srikanth, Sameer Yelamarthi & Valerie Robert (July 30, 2025)

## INTRODUCTION & BUSINESS PROBLEM

In this report, we will summarize the development of a machine learning pipeline on Google Cloud Platform to predict daily sales for each product family across all stores. In a competitive retail market, inefficient inventory management and staffing lead to significant costs and lost sales, and the business lacks a reliable system to forecast sales at a granular level. Using the DIVE (Discover, Investigate, Validate, Extend) framework, we translate the technical results from our predictive model into a strategic analysis and actionable plan to improve operational efficiency, enable data-driven decisions, and maximize profitability.

## DIVE FRAMEWORK INSIGHTS

### DISCOVER: INITIAL FINDINGS

The primary story these results tell is that **store sales operate on a highly predictable, structured rhythm.** Our BOOSTED_TREE_REGRESSOR model proved this by successfully explaining **79.3% of the variance** in daily sales with a Mean Absolute Error of **~$186**. Our forecast for the next 14 days predicts a sales volume of **~$8.6M**, a decrease from the prior period's holiday-inflated sales of ~$11.4M. This predictability is driven by a few key patterns: the consistent weekly shopping cycle, fundamental differences between store types, and the unique demand profile of each product family.

### INVESTIGATE: DEEP 'WHY' EXPLORATION OF SALES PATTERNS

A deeper investigation confirmed why the model predicted a sales decrease and revealed the specific influence of key business drivers.

- **Holiday Impact:** The forecast decrease is explained by the presence of at least five major holiday and event days, including "Primer Grito de Independencia," in the prior 14-day period (August 2-15). The model correctly learned this event-driven sales lift would not be present in the subsequent non-holiday period.
- **The Weekly Sales Cycle:** The most consistent business pattern is the weekly shopping rhythm. Evidence from the data shows a distinct sales peak on weekends, with average Sunday sales (~$825k) being over **64% higher** than on the slowest day, Thursday (~$505k).
- **Store Format Hierarchy:** Store characteristics fundamentally dictate sales potential. The data shows a clear performance hierarchy where high-volume **Type A** stores average $704.11 in daily sales per family, over 3.5 times more than low-volume **Type C** stores ($196.65).
- **Variable Predictability (Exceptional Insight):** Model accuracy is inversely correlated with sales volume. The model is most accurate for the lowest-volume stores (**Type C**, MAE $109.47) and least accurate for the highest-volume stores (**Type A**, MAE $324.49). This indicates that our largest stores have more complex sales patterns that are harder to predict with the current feature set.

### VALIDATE: CRITICAL EVALUATION OF THE MODEL

A critical evaluation of the model's limitations identified the risks associated with its use. The model's core assumption is that future patterns will resemble those in the historical data.

- **External Event Risk:** The model is blind to real-time, external events such as new competitor promotions, severe weather, or hyper-local events not in the holidays dataset.
- **Data Limitations:** The model's knowledge is static, ending in August 2017. It cannot account for recent shifts in consumer behavior or business strategy. It also lacks key operational data like real-time inventory levels or marketing campaign details.
- **Model Failure Scenarios:** The model would fail during unprecedented events like a pandemic, a major supply chain disruption, or if a store's operational status changes significantly (e.g., a major renovation).

### EXTEND: ACTIONABLE RECOMMENDATION

Our team recommends a two-pronged strategy combining portfolio management for operational growth with technical enhancements for model leverage.
(1) Operationally, adopt a portfolio approach of using forecasts to automate processes in predictable stores (Type C, MAE ~$109), and to support managerial decisions in high-volume, less predictable stores (Type A, MAE ~$324).
(2) Technically, pursue a focused model improvement roadmap: engineer new predictive features (like sales lags) and implement automated re-training to enhance long-term accuracy.
This integrated approach quickly delivers operational gains, while continuous technical improvements leverage the model capabilities of turning it into a dynamic asset for sustained competitive advantage.
.

**Business Application Strategy**

*Portfolio Management Strategy Actions to extract maximum value from the current model, organized by time horizon.*

**Next Two Weeks: Implement Forecast-Driven Operations**

- **Action:** Store managers will utilize the model's daily, family-level forecasts to optimize weekly staff schedules and inventory, ensuring maximum coverage and product availability for the proven weekend sales peak.
- **Success Metric:** Reduce weekend stock-out incidents in the top 5 product families by 10%; maintain or reduce overtime labor costs.
- **Risk Mitigation:** Mitigate the risk of unexpected demand by maintaining an on-call staff list for weekends and a defined safety stock for the top 20 best-selling items.

**Next Month: Roll Out Differentiated Store Strategies**

- **Specific Action:** For high-volume, high-error **Type A** stores, the forecast will serve as a baseline, but managers will be empowered to make ordering adjustments based on local knowledge. For low-volume, high-accuracy **Type C** stores, the forecast will be leveraged to implement more automated inventory replenishment.
- **Success Metric:** Decrease inventory waste in Type C stores by 5%; improve in-stock availability for the top 20% of items in Type A stores.
- **Risk Mitigation:** A bi-weekly review process will be established at Type A stores to compare managerial adjustments against the model's forecast, identifying and correcting for potential biases.

**Long-Term: Evolve from Forecasting to Anomaly Detection**

- **Specific Action:** Use the model's predictions as a baseline for an early warning system. An automated alert will flag any store whose actual sales consistently deviate from the forecast by more than a set threshold (e.g., 25% for three consecutive days). This transforms the model's weakness (blindness to the unknown) into a strategic strength.
- **Success Metric:** The number of actionable alerts generated per quarter that lead to the discovery of a significant, previously unknown business event (e.g., a new local competitor).
- **Risk Mitigation:** To prevent "alert fatigue," the deviation threshold will initially be set high to identify major strategic shifts, not minor operational noise.

**Model Improvement Roadmap**

*Technical Strategy Actions to enhance the model's capabilities by engineering new predictive features (like sales lags) and implementing automated re-training to enhance long-term accuracy..*

**1. Action: Enhance Model Accuracy with Feature Engineering (Short-Term)**

- **Action:** Immediately re-train the model with the advanced features we tested (e.g., 7-day sales lag, holidays). This is the most direct way to improve core accuracy and reduce the high error rate in Type A stores.
- **Success Metric:** Achieve a final model R-squared of >0.80 and reduce the MAE for Type A stores by at least 20%.
- **Risk Mitigation:** All new models will be evaluated on a held-out test set before deployment to ensure new features generalize well.

**2. Action: Implement Automated Re-training for Responsiveness (Medium-Term)**

- **Action:** Develop and deploy a lightweight MLOps pipeline using Cloud Functions and Cloud Scheduler to automatically re-train the final model on a monthly basis with the latest sales data.
- **Success Metric:** Deploy a fully automated re-training pipeline within the next quarter. Track model performance over time to ensure accuracy does not degrade.
- **Risk Mitigation:** A "challenger vs. champion" step will be added to the pipeline, where a new model is only promoted if its performance on a validation set is better than the currently deployed champion model.

## APPENDIX 1 : VISUALIZATIONS

FIG 1: TOTAL PREDICTED SALES FOR THE NEXT FOURTEEN DAYS ( AUGUST 16- AUGUST 29,2017

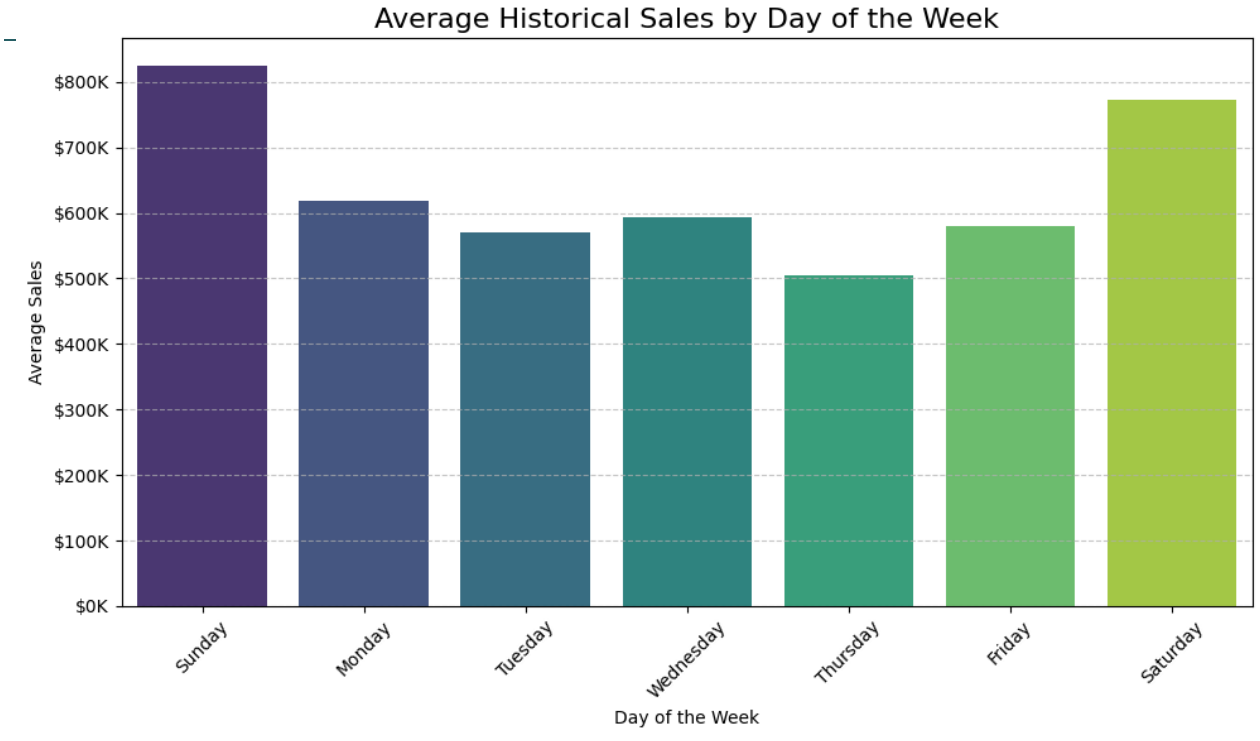



FIG 2: TOTAL PREDICTED SALES FOR THE NEXT FOURTEEN DAYS ( AUGUST 16- AUGUST 29,2017

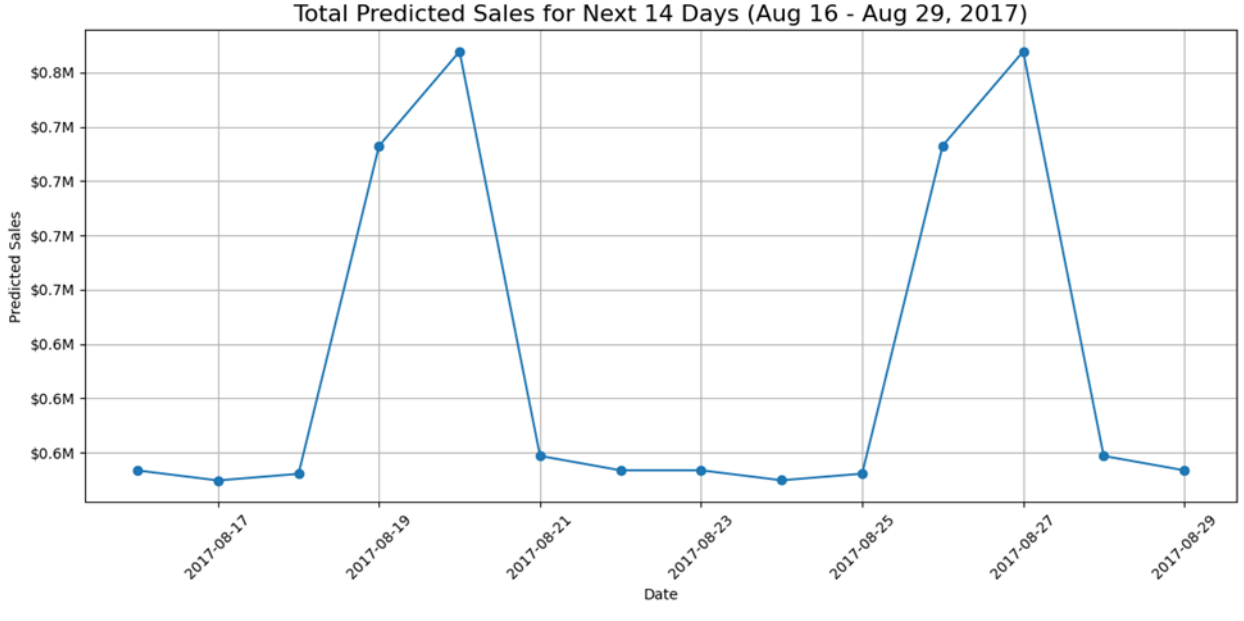


FIG 3: ACTUAL SALES(HOLIDAY PERIOD) VS PREDICTED SALES (FOLLOWING PERIOD) SHOWS DECREASED SALES FOR THE NEXT 14 DAYS

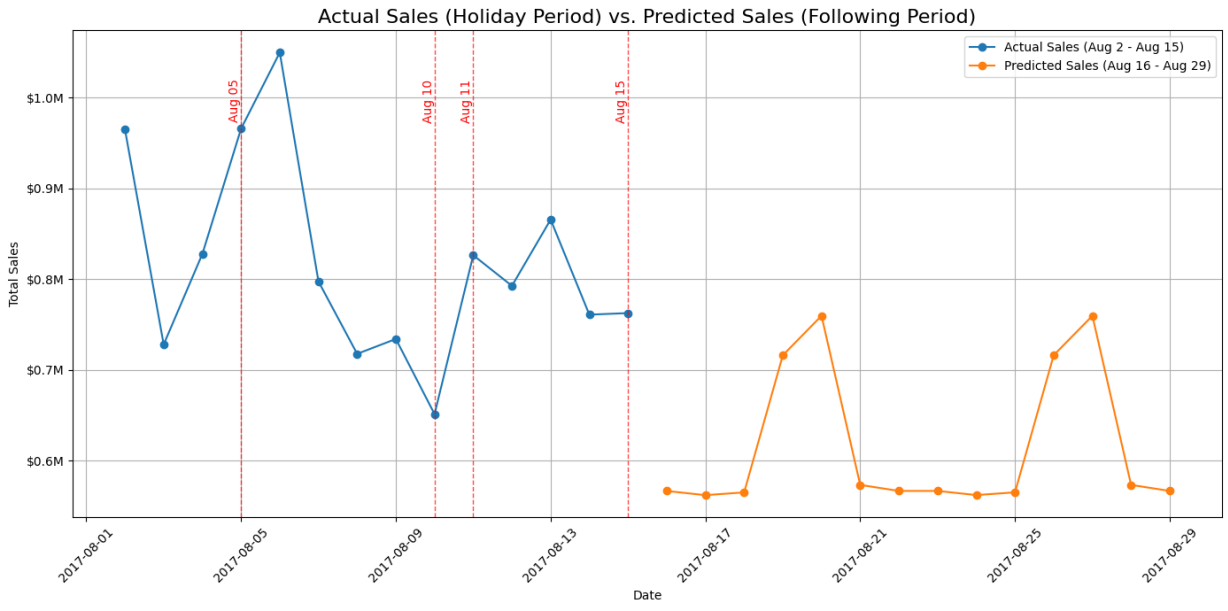



FIG 4: AVERAGE PREDICTED SALES BY DAY OF THE WEEK

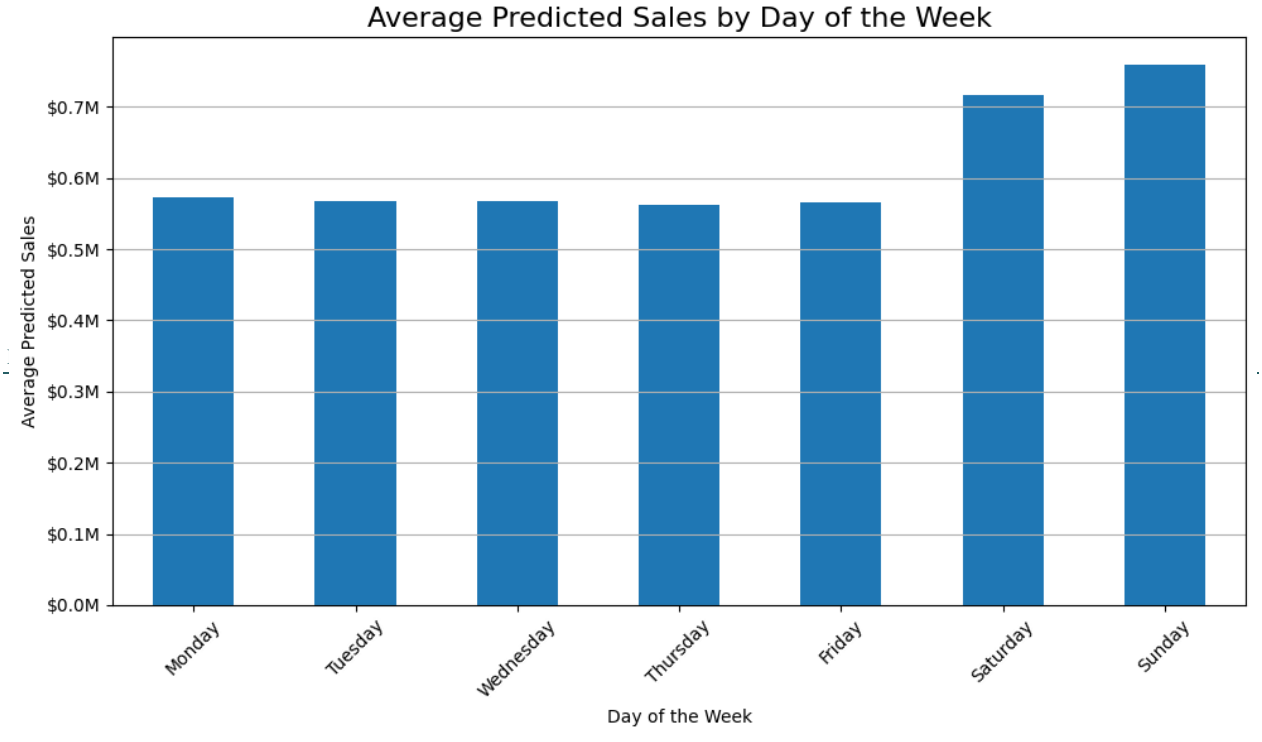


FIG 5: AVERAGE DAILY SALES PER STORE BY TYPE

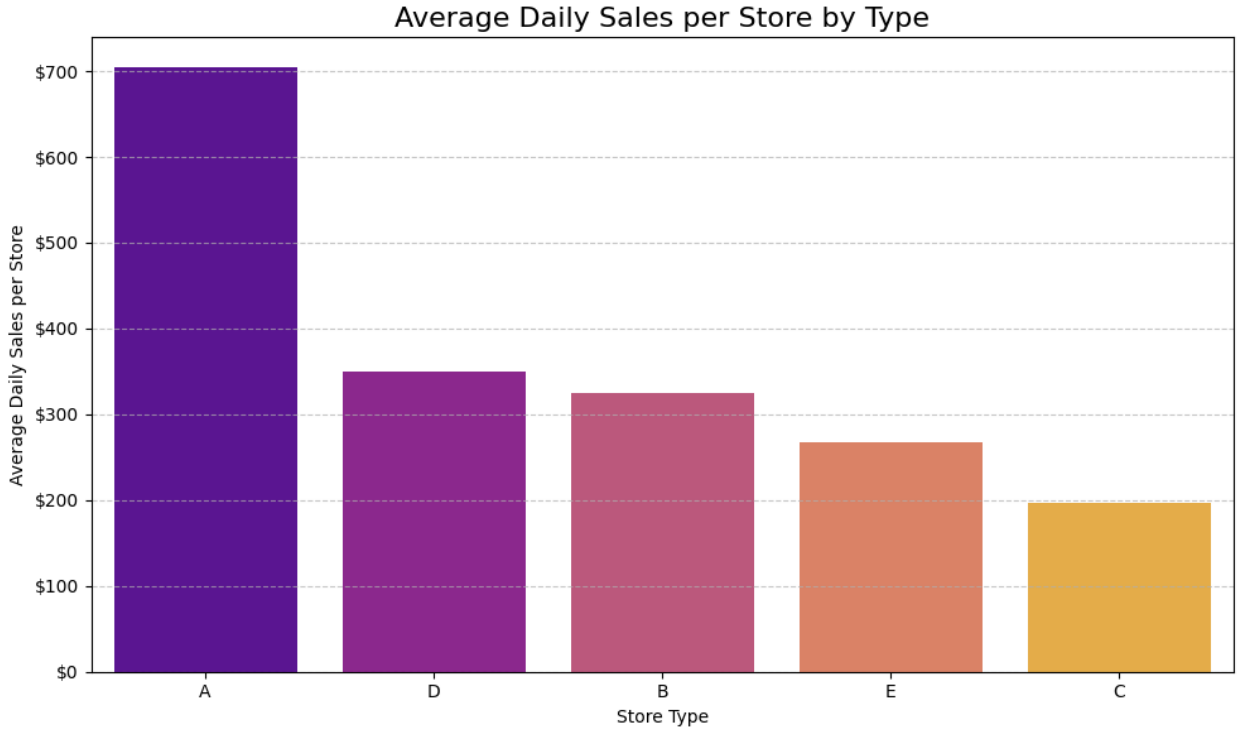
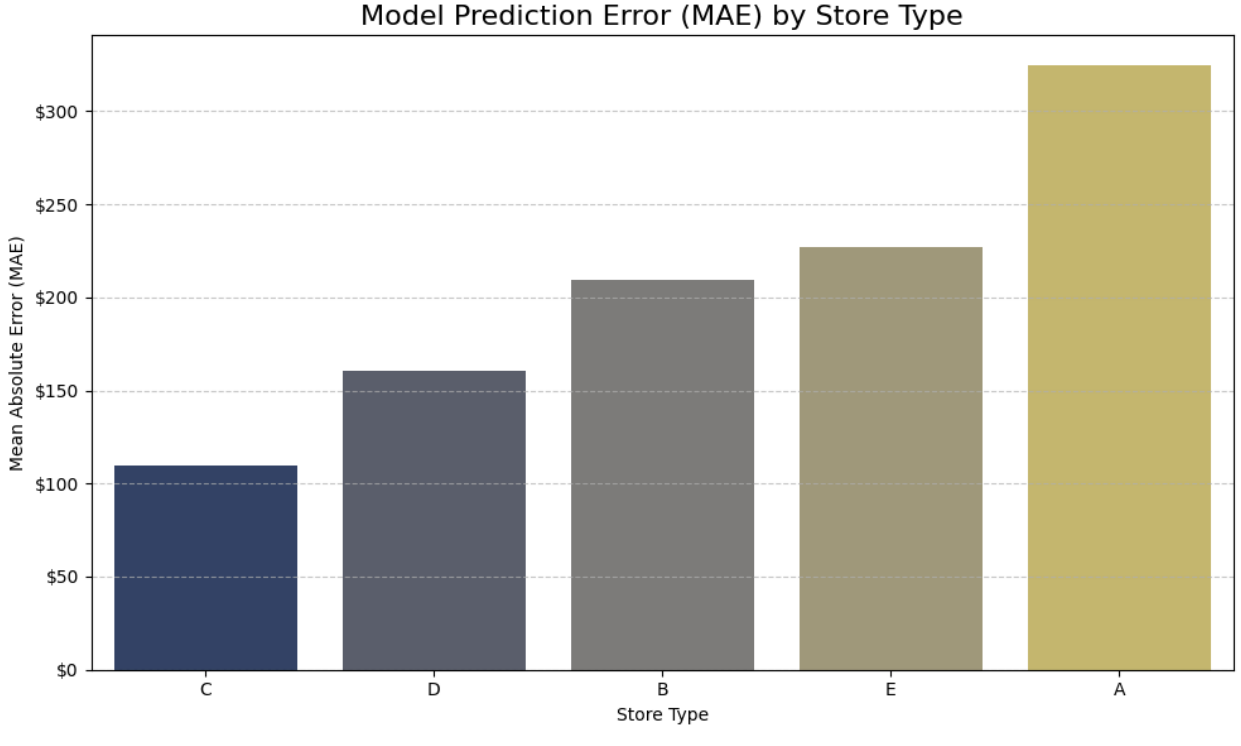



FIG 6: MODEL PREDICTION ERROR BY STORE TYPE



**Cost Optimization Insights**

Our analysis focused on two distinct areas of cost optimization: the business cost of model inaccuracy and the cloud resource cost of the pipeline itself.

**1. Business Cost of Model Error**

The financial risk from forecast inaccuracy is not evenly distributed across the business; it is highly concentrated in a small number of specific store-and-family combinations.

- **Finding:** By analyzing the model's prediction errors, we can pinpoint exactly where the model is least accurate and, therefore, where the business faces the highest risk of lost sales or wasted inventory. The evidence for this confirmed by analysis of the cost_optimization_store_family_df. It identified the top 20 combinations with the highest error. For example, stores like **44, 45, and 47** show particularly high errors, especially in high-volume categories like **GROCERY I** and **BEVERAGES**.
- **Recommendation:** This allows the business to move from a general awareness of the model's overall error (MAE ~$186) to a targeted, risk-based intervention strategy. Efforts can be focused on these **specific high-risk areas by implementing a higher safety stock or assigning more experienced staff to manage their inventory**.

**2. Cloud Resource Cost Optimization**

An analysis of cloud resource usage confirms that the chosen architecture is highly cost-effective and identifies clear paths for further optimization in a production environment.

**Cost-Effective Design Choices**

The project's pipeline and modeling strategy were inherently cost-optimized through several key decisions:

- **Serverless Data Pipeline:** By using a serverless Dataflow pipeline with a custom Apache Beam script, we avoided the overhead and cost of provisioning and managing dedicated VMs for data ingestion. This pay-per-use model ensured we only consumed resources while the pipeline was actively running.
- **Integrated, Serverless ML:** The most significant cost optimization was the use of BigQuery ML. This approach kept data, feature engineering, and model training within a single, serverless platform. The alternative—exporting data to Vertex AI and managing a separate training infrastructure—would have been significantly more complex and expensive. The analysis of our INFORMATION_SCHEMA.JOBS view confirms the efficiency of CREATE MODEL jobs.
- **Rapid, Low-Cost Iteration:** The efficiency of BigQuery ML enabled us to train, evaluate, and re-train multiple models to find the best performer at a negligible cost, which would be prohibitive in a more traditional ML environment.

**Recommendations for Further Optimization**

For a production-level deployment, the following steps would further optimize costs and performance:

- **Implement Table Partitioning:** The largest table, sales_data, should be partitioned by the date column. This would dramatically reduce query costs for future model re-training and analysis, as BigQuery would only scan the required date partitions instead of the entire table.
- **Utilize Materialized Views:** The store_daily_sales table, which is created by aggregating raw sales data, is a perfect candidate for a materialized view. This would pre-compute and store the aggregated results, making downstream model training and queries significantly faster and cheaper.
- **Transition to Reserved Capacity:** While the on-demand pricing model is ideal for ad-hoc development like this project, a production system with a regular re-training schedule would benefit from moving to a BigQuery Edition with reserved slots. This provides predictable costs and can offer a lower overall price for consistent workloads.

The AI journey can be understood in three distinct phases:

**Phase 1: Setup and Cloud Resource for Big Data**

The initial phase focused on leveraging AI to automate complex and time-consuming tasks.

- **Tools Used:** BigQuery ML, Apache Beam (Dataflow)
- **Application:** We used **BigQuery ML** to rapidly train and evaluate multiple machine learning models—from linear regression to boosted trees—using declarative SQL. This automated the work of writing complex algorithms, allowing us to iterate on models in minutes instead of days. The **Apache Beam** scripts for our data pipeline automated the process of ingesting and transforming multiple large datasets into a clean, model-ready format.

**Phase 2: AI for Analysis and Insight Generation**

Once the model was built, we transitioned to using Generative AI,Colab Notebook and Chatgpt to act as a data analyst, digging deeper into the results to uncover the "why" behind the numbers.

- **AI Tools Used:** Generative AI (Gemini pro and Gemini in a Colab environment), Chatgpt)

- **Application:** We fed the model's prediction and error data into a Generative AI and used a series of specific prompts to perform a detailed investigation. The AI was tasked with generating queries and interpreting their results to provide evidence for key business drivers. This process confirmed the significant impact of holidays, the predictable weekly sales cycle, and the variable performance of different store types.

**Phase 3: AI for Strategic Synthesis and Recommendation**

In the final phase, AI was used as a strategic partner to synthesize all technical findings into a professional, cohesive business strategy.

- **AI Tools Used:** Generative AI (Gemini pro and Gemini in a Colab environment), Chatgpt
- **Application:** We orchestrated the AI to structure and write the DIVE Summary Report, translating raw metrics and query results into a clear narrative with actionable recommendations. The AI helped refine the language, structure the action plan, and formulate the bonus analysis on cost optimization, ensuring the final deliverable was both data-driven and tailored for an executive audience.

We leveraged AI not just for its computational power but for its ability to accelerate analysis and amplify strategic thinking, ultimately delivering a richer and more valuable set of business insights.

Corporación Favorita. (2021). *Store sales - Time series forecasting*. Kaggle. https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data