DIVE ANALYSIS REPORT

FOR ASSIGNMENT #2

BIG DATA ANALYTICS IN THE CLOUD

MGMT-599

TEAM DN5

FISCHER, CHRISTOPHER

MOLNER, ERIC

RIZZO, JUSTIN

WHITE, SHARON E.

Accurate sales forecasting is key to a successful retail operation, helping to optimize inventory and staffing. This project built an automated forecasting system using Google Colab, Cloud Storage, and BigQuery. Following the DIVE framework, our pipeline was designed to use Google Dataflow to move data into BigQuery, where we trained a forecasting model with the BigQuery ML engine. This report details our findings and provides business recommendations based on the patterns our model revealed.

DIVE Journey

D - Discover Our analysis uncovered the business's predictive heartbeat: a powerful, consistent weekly sales cycle. This rhythm, with reliable weekend peaks and troughs, provided a strong forecasting signal. This foundational insight confirmed we could leverage time series analysis to model operational needs and shift from reactive to proactive planning. The time-series model could predict sales with 95% accuracy.

I - Investigate[1] To decode the "why" behind this cycle, we investigated its drivers. The pattern reflects ingrained customer habits—people shop more on days off. A deeper insight emerged when segmenting by store type: while all stores follow this rhythm, their sales volume varies dramatically. This proves the forecast is an aggregate of similar patterns occurring at different scales, driven by store size.

V - Validate A critical validation of our model reveals a key trade-off. It excels at predicting the standard weekly business rhythm with a reasonable margin of error. However, its primary limitation is a "context blindness" in that it cannot see major holidays or local events, creating a risk of under-forecasting during critical sales periods. It should be trusted as a powerful baseline for normal operations, but not for unique, event-driven scenarios.

E - Extend To extend the model's strategic value, we recommend implementing a "human-in-the-loop" system. Managers must treat the forecast as a data-driven starting point, using their local expertise to layer on contextual intelligence about holidays and events. The long-term extension is to feed this qualitative data back into the model,

[1] Further investigation during the creation of the Visualizations explained in Appendix.

evolving it from a simple forecaster into a responsive decision-making tool that continuously improves its accuracy.

Action Plan

1. Combine Data Insights with Human Expertise

- **Action:** Store managers will treat the weekly forecast as a baseline in that they will review the prediction and adjust it based on their knowledge of local factors like weather, or nearby road construction, things that the model cannot see.
- **Success Metric:** The primary metric will be a 10% reduction in weekend stockouts for key product categories over 1/4, demonstrating that the manager's adjustments improved inventory planning beyond the model's baseline.
- **Risk Mitigation:** Require managers to log their adjustment and a brief rationale each week. This creates accountability and a feedback loop for the team.

2. Track and Report Local Contextual Factors

- **Action:** Managers maintain shared "Local Events Calendar" to log store-specific promotions, school holidays, etc. to enhance future model versions.
- **Success Metric:** Success will be measured by the model showing a reduced Mean Absolute Percentage Error (MAPE) during holiday or event periods.
- **Risk Mitigation:** Keep pre-formatted online calendar or spreadsheet. Share examples of how this data improved the forecast to keep managers engaged.

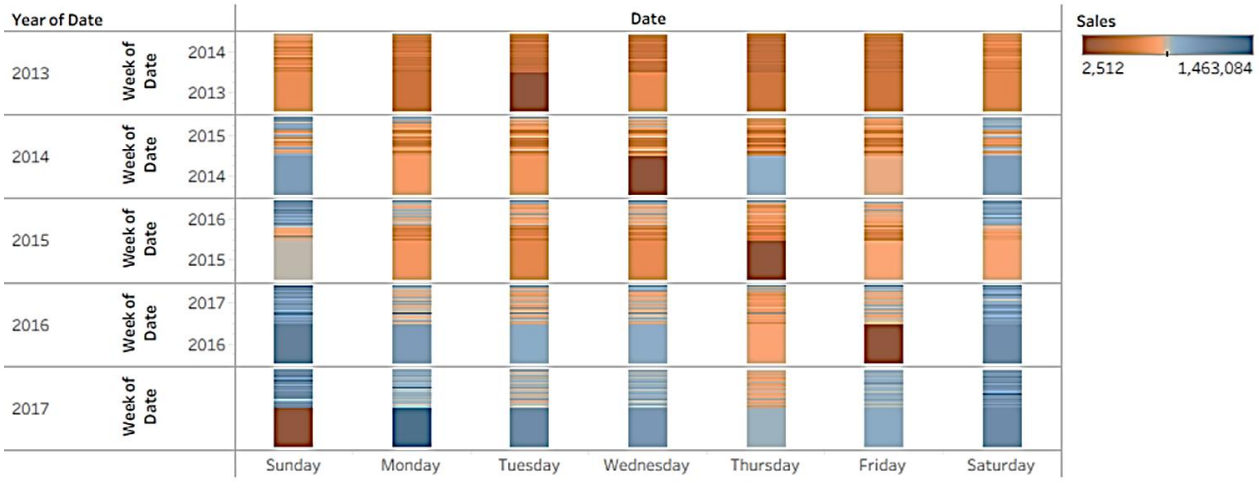
3. Provide Feedback on Store-Specific Patterns

- **Action:** If a manager consistently observes that their store's sales pattern deviates from the forecast (e.g., their peak day is Friday, not Saturday), they are to submit a "Pattern Anomaly Report" with their hypothesis for the deviation.
- **Success Metric:** A successful outcome is the data team identifying at least three distinct store clusters with unique sales patterns within six months, leading to specialized, more accurate models for those clusters.
- **Risk Mitigation:** Establish a clear and simple channel for submitting reports. Acknowledge every submission, even if it doesn't lead to an immediate model change, to ensure managers feel heard and their expertise is valued.

APPENDIX

Daily Sales Rhythm

The color intensity shows the total sales for the day.
Using this heatmap, you can easily see:
* Weekly Seasonality: As shown in the Vertical Columns where Saturday and Sunday will be consistently darker colored than the columns for Tuesday - Thursday.
* Holidays & Anomalies: Major holidays (like Christmas) or specific outlier days are really dark or light squares
* Monthly/Yearly Trends: Changes in color intensity from the beginning of a year to the end of the year.

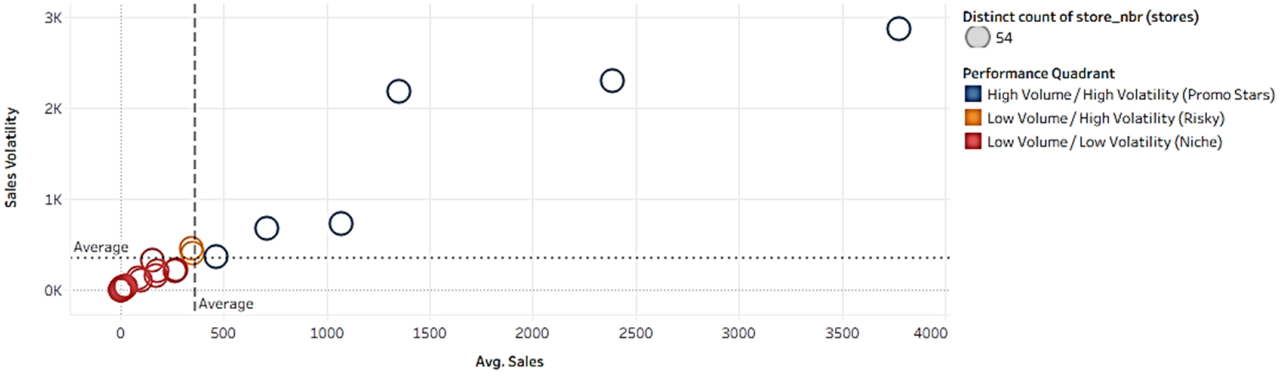


Date Week for each Date Weekday broken down by Date Year. Color shows sum of Sales.

Product Family Performance Scatter Plot

Segments products into four business categories based on average sales volume and sales volatility (i.e. how much it fluctuates day-to-day). Answers the question for the Investigate and Explain phase: "What is the Character of each product family?"

High Volume, Low Volatility(Staples): Predictable core products.
High Volume, High Volatility (Promo-Driven): Products that sell a lot but are unpredictable, likely driven by promotions.
Low Volume, Low Volatility (Niche & Steady): Smaller but reliable product lines.
Low Volume, High Volatility (Risky/Sporadic): Unpredictable products that don't contribute much.



Average of Sales vs. Sales Volatility. Color shows details about Performance Quadrant. Size shows distinct count of store_nbr (stores). Details are shown for Family.

Footnote:

A critical insight emerged from the Product Family Performance Scatter Plot, which plots each product family by its sales volume and volatility. The analysis revealed a significant gap in the product portfolio: the "High Volume, Low Volatility" quadrant is completely empty. This finding indicates that while the business has predictable, low-selling "Niche" products and popular but unpredictable "Promo-Driven Stars," it lacks a core category of "Staples" that are both high-selling and stable. This reliance on volatile top-performers for the majority of revenue presents a strategic challenge, complicating inventory management and making revenue forecasting more difficult.

Action Plan

Specific Action for Managers:
Launch a "Volatility Reduction Initiative" focused on the top 5 product families in the "High Volume, High Volatility" quadrant. Analyze the promotion history for these items to determine if deep, infrequent discounts are causing the sales instability. The pilot action will be to switch one key product family from a "boom-bust" promotional cycle to a more consistent, smaller discount or an "Everyday Low Price" (EDLP) strategy for one business quarter.

Success Metric for this Action:
The primary success metric is a 15% reduction in the Sales Volatility (the Standard Deviation of Sales) for the pilot product family over the quarter, while keeping its average daily sales volume within 95% of its previous level. A secondary metric is a 10% reduction in stockouts for that family, indicating more predictable demand.

Risk Mitigation Strategy:
Risk: The primary risk is that reducing deep discounts could lead to a significant drop in overall sales volume, as customers may be conditioned to wait for major sales events.
Mitigation: Mitigate this by running the pilot on a single, representative product family first, not the top seller. Communicate the new strategy to customers with in-store signage highlighting "New Low Price" or "Consistent Value" to re-frame customer expectations away from temporary deep discounts.

COST OPTIMIZATION STRATEGY FOR THE ML PIPELINE

For our ML Pipeline Project, an effective cost optimization strategy involves more than just reducing spend; it's about maximizing the return on our investment in the cloud. This requires a holistic view of our pipeline, from data ingestion to model deployment, focusing on three key areas: intelligent data processing, efficient model management, and strategic resource use.

1. Intelligent Data Processing and Storage: Our primary costs begin with how we handle data. Instead of scanning the entire sales_data table for every training run, we should implement partitioning on the date column. This single change means BigQuery only scans the specific date partitions needed, dramatically reducing query costs. Furthermore, we can establish a data lifecycle policy in Cloud Storage to automatically move raw source files older than 90 days to a cheaper storage class, as they are no longer needed for active training.

2. Efficient Model Management: The most expensive part of our pipeline is model computation. Our analysis shows that high forecast error is often due to the inherent volatility of our top-selling products. To manage this cost-effectively, we will use an intelligent, two-tiered approach:

- **Targeted Analysis:** First, we will run smaller, cheaper analytical queries each week focused only on these high-volatility product families. This allows us to monitor their behavior and understand the drivers of their instability without the high cost of a full training job.
- **Strategic Retraining:** Second, we will only trigger a full, expensive retraining of the entire model on a strategic basis—either quarterly or when we have significant new data, such as holiday performance, to incorporate.

This strategy focuses our cloud spending on generating actionable insights into our key products, rather than on frequent retraining that may not improve performance.

3. Strategic Resource Use in Dataflow: During the data ingestion phase, we can optimize our Dataflow jobs. By configuring our pipeline to use autoscaling and selecting

cost-effective worker machine type, we ensure we aren't over-provisioning resources. For a pipeline like this that runs on a schedule, these small adjustments prevent unnecessary costs during a step that should be highly efficient.

By implementing these strategies, we move from a simple pipeline to a cost-conscious, production-ready system that delivers business value efficiently.

Conclusion

The most powerful insight came from our 'Product Family Performance Matrix' visualization, which revealed our 'High Volume, Low Volatility' quadrant was empty. This discovery is critical because it explains *why* our aggregate forecast model has a certain level of error (MAPE): the model is trying to find a stable average for our most important products, but these products are inherently unstable and volatile.

Therefore, our ultimate recommendation for future work is to move beyond a single, generalized model. We suggest developing a multi-model strategy. The predictable 'Niche' products could be forecasted with a simple model, while the volatile 'Promo Stars' require a more advanced model that can incorporate external factors like promotion schedules as features. This tailored approach would significantly improve forecast accuracy and represents the next logical evolution of this ML pipeline.