Assignment 2: ML Pipeline Project

From Kaggle to Predictions with BigQuery ML

Group: DN3

Members: James Fazecas, Kyasha Williams, Yuxiao Wang, Michael Yudanin

D - Discover (Initial Exploration)

The model has good predictive ability: predictions are on average within 19% of the actual sales, the model explains 83% of the variation in sales per store when making predictions.

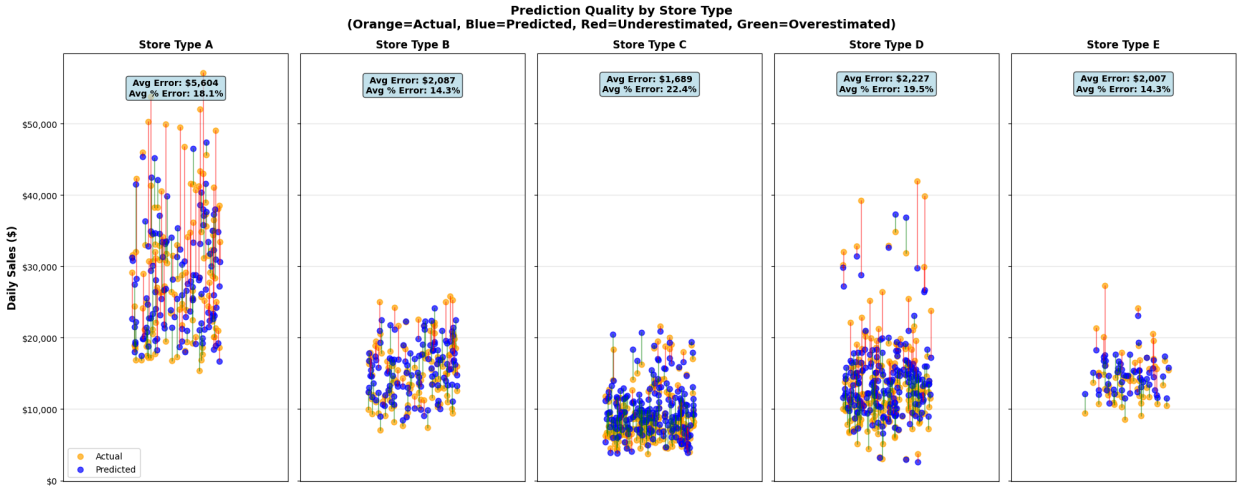
1. **There is room for improvement**:

   (1) On average, the model is within 19% of the actual sales. This suggests the need for further tune-up, perhaps, training on a larger dataset or morr sophisticated consideration of temporal variables.

   (2) Predictive performance is worse than training. This suggests overfitting and the need to improve the model.

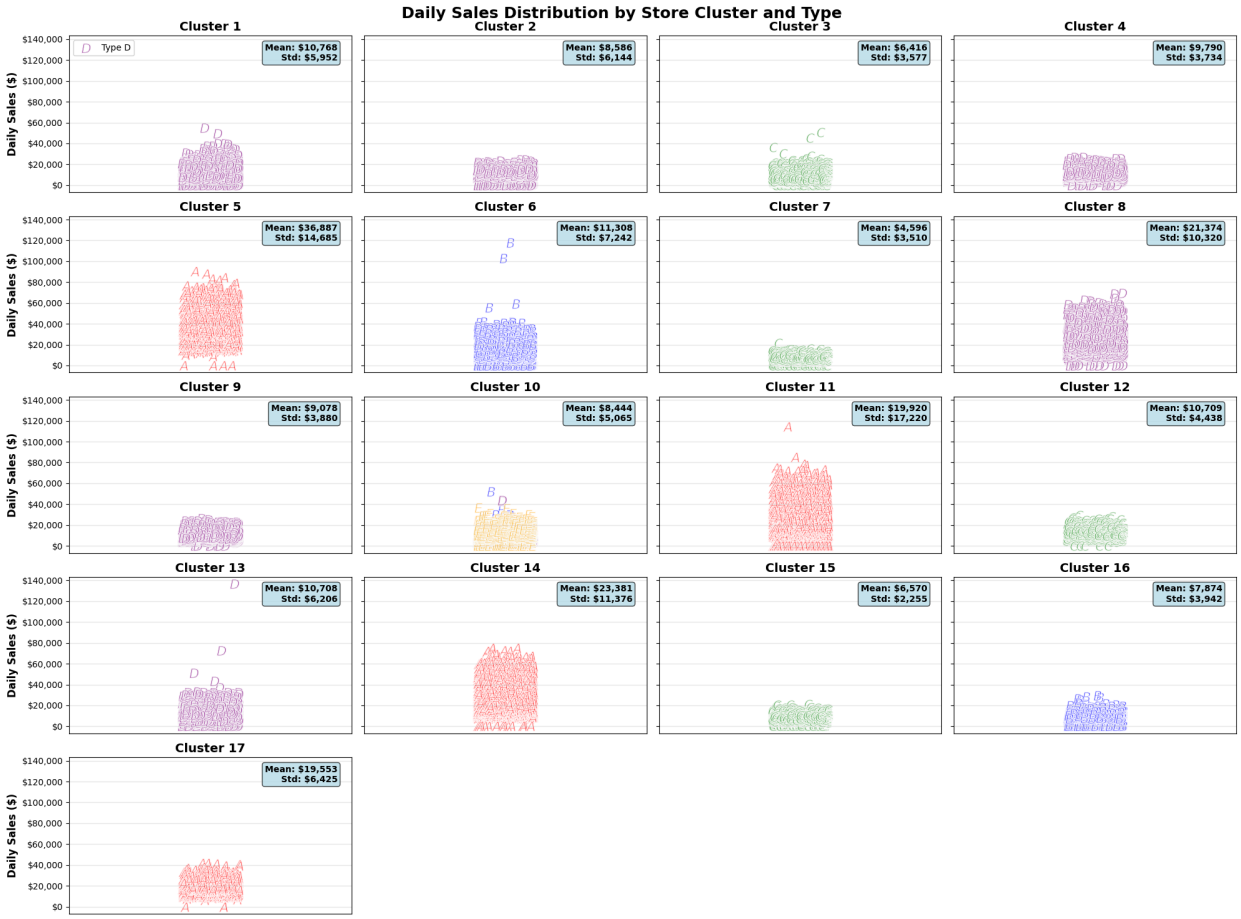
---

3. **Store Type Significantly Influences Sales and the model's predictive abilities**, with stores type A having the highest sales but the hardest to predict using the model, and type C stores have low average sales and a high average percentage error (22.25%)



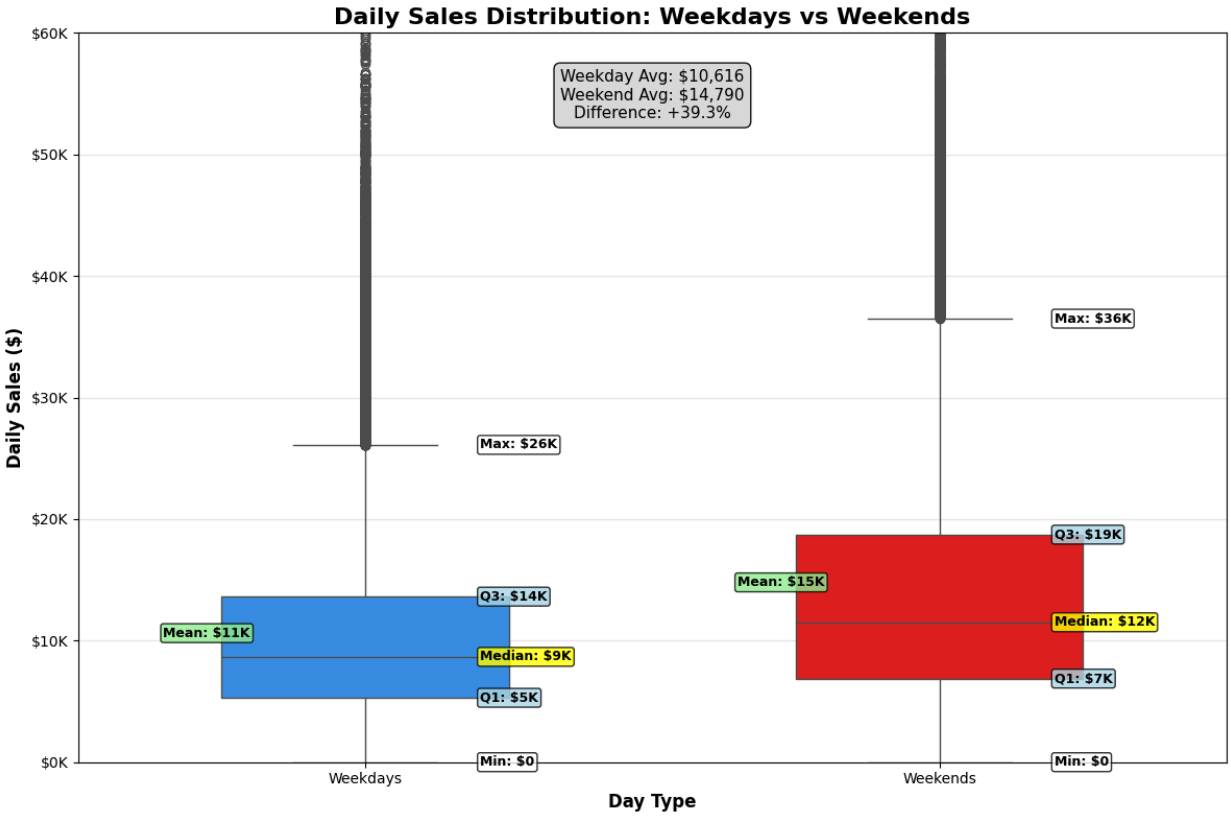
We can see that the dispersement of the actual results for type A is also higher. It is larger in actual sales than in predictions.

4. **Store Performance Relative to Cluster:** is an important factor.

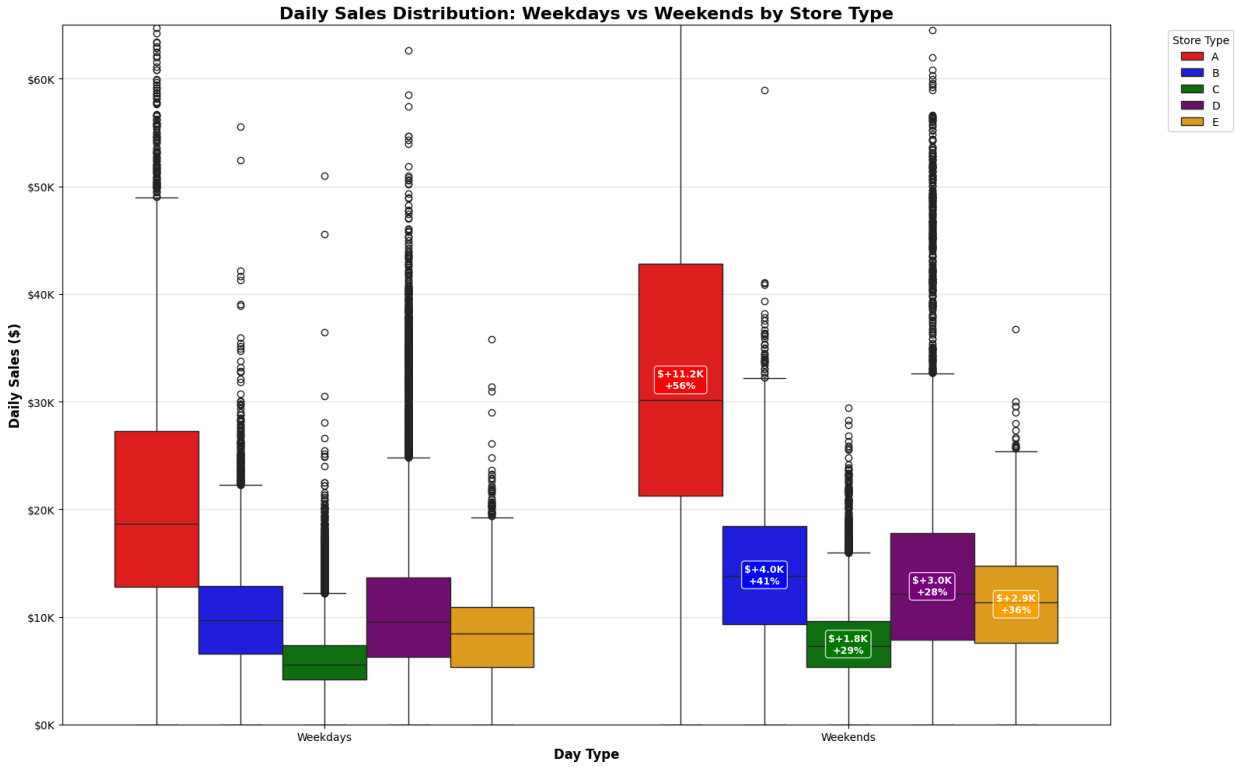


Most clusters show similar sales performance by their stores and sport only one store type. The exception is clusters with type A and one cluster that has type D stores (cluster 12) that show a wide range of sales results.

5. **Weekend impacts performance** - on average, stores sell more on weekends.

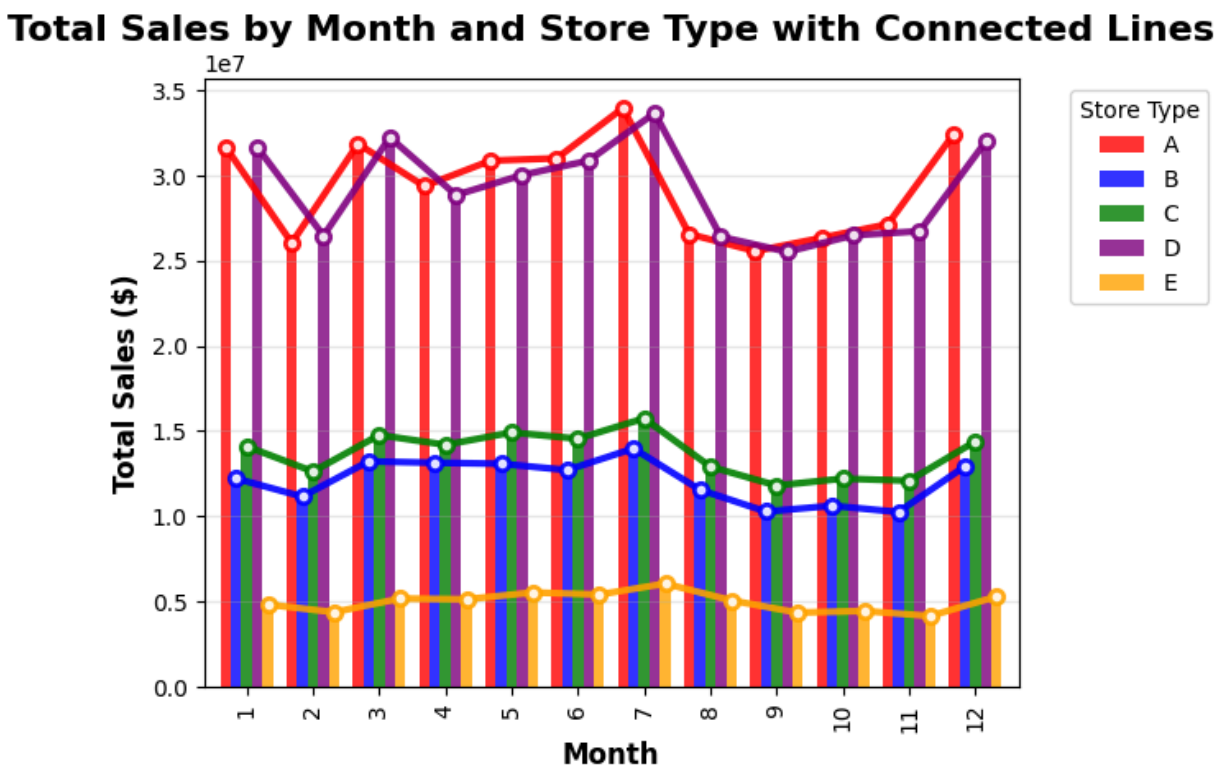


Dependence on type:



As we can see, the impact of the weekend sales bump is not equal across different store types, with type A having a bump of $11.2K / 56% on average while type D - $3K / 28%.

6. **Month impacts performance** - which needs to be investigated further.



As we can see, the month-to-month pattern if the came for all types: peaks in July, December, January, and March, and lows in February, August, September, October, and November.

I - Investigate (Deep Analysis)

1. Weekend influence is modulated by store type.
2. Customer demographics, product assortments, and location migth be underlying the impact of store type on sales. Store size migth have an impact as well.

3. Seasonality plays a significant role.
4. Past sales are poor predictor of sales performance.
5. The model suffers from overfitting.

Several **hypotheses** that can be formulated based on these insights:

H1: Store Role Hierarchy Hypothesis

Theory: Type A stores serve as "anchor stores" that drive traffic to entire shopping areas, while types C and D are "convenience/specialty" stores that benefit from spillover traffic.

Testable Prediction:

- Type A stores will show stronger correlation between their sales and other stores' sales within the same cluster When type A stores have promotions/events, types C and D in the same cluster will see 5-15% sales increases within 1-2 days
- Type A stores' sales variance will predict cluster-wide sales variance better than any other type

Why this matters: This explains why Type A has higher but more variable sales - they're not just bigger stores, they're ecosystem drivers.

H2: Consumer Mission-Based Shopping Hypothesis

Theory: Weekend shopping represents fundamentally different consumer missions (leisure/family vs weekday convenience/necessity), and store types capture different missions rather than just demographics.

Testable Prediction:

- Weekend sales increases will be disproportionately higher for stores selling discretionary/experience goods vs necessity goods
- The weekend effect should vary significantly within store types based on product mix (even among Type A stores) Stores with longer average transaction times will show larger weekend premiums
- Holiday weekends vs regular weekends will show different type-specific patterns

Why this matters: This explains why the weekend effect varies by type - it's about shopping behavior, not just available time.

H3: Local Market Saturation Threshold Hypothesis

Theory: The relationship between cluster performance and individual store performance follows a non-linear saturation curve, where stores only benefit from high-performing clusters up to a threshold.

Testable Prediction:

- Stores' performance relative to cluster average will show diminishing returns above a certain cluster performance level
- In the highest-performing clusters, individual store characteristics (type, size) will matter less than in average clusters
- New stores entering high-performing clusters will struggle more than those entering average clusters
- The "sales_vs_cluster_avg" feature should have different coefficients for stores in top quartile vs bottom quartile clusters

Why this matters: This explains why cluster effects aren't linear - there are market capacity constraints.

H4: Temporal Routine Disruption Hypothesis

Theory: Sales patterns are driven by disruptions to consumer routines, and different store types serve different routine vs disruption needs.

Testable Prediction:

- Sales will spike significantly during "routine disruption" periods: school holidays, weather events, local events, paydays
- Type A stores will be more sensitive to positive disruptions (events, holidays), while types C and D will be more sensitive to negative disruptions (bad weather, economic stress)
- The predictive power of lagged sales will be stronger during "routine periods" and weaker during "disruption periods"
- Stores near schools/offices will show different disruption sensitivity than those in residential areas

Why this matters: This explains why simple historical patterns don't predict well - consumer behavior is routine-based with disruptions.

H5: Cross-Store Cannibalization Network Hypothesis

Theory: Stores don't compete uniformly - there are specific cannibalization networks based on consumer substitution patterns that create predictable sales transfers.

Testable Prediction:

- When stores temporarily close/reduce hours, sales increases in other stores will follow predictable patterns based on distance + store type combinations
- Promotional events at one store will cause measurable sales decreases at specific other stores (not uniformly across the cluster)
- The effect of new store openings will vary dramatically based on the type-distance combination of existing stores Stores that consistently over/under-perform their cluster average will have identifiable "competitor shadows" - specific other stores that perform inversely

Why this matters: This explains why cluster effects are complex - it's not about average performance, but about specific competitive relationships.

V - Validate (Quality Assurance)

- Variations within store types that are relevant to sales. This migth be particularly important for type A stores.
- Ignopring lag sales impact: the lags chosen for analysis, 1 and 7 days, might be too short.
- Interactions between features: the model, being Linear Regression, focuses on individuals features. It well migght be that interactions betwene features, e.g., store type and weekend, have significant impact on sales.

E - Extend (Strategic Application)

1. Next Week:

- **Targeted promotions**
    - **Type A Stores:**
        - **Action:** Implement targeted weekend promotions focusing on high-margin discretionary items. This could include bundled deals, flash sales, or experiential events.
        - **Metrics:** Weekend sales uplift (compared to the previous weekend and to weekdays), conversion rates, margin per transaction, customer feedback.
    - **Types C & D Stores:**
        - **Action:** Ensure adequate staffing during peak weekend hours to handle potential increased traffic spillover from Type A stores subject to promotions, if inclose proximity. Optimize product placement for impulse purchases, e.g., for aisle buy.
        - **Metrics:** Weekend sales uplift, average transaction value, customer satisfaction, inventory turnover for promoted items.
- **Common metrics:** Week-over-week sales growth, promotion redemption rate, conversion rate (sales/traffic), average transaction value.
- **Inventory optimization**

    - **Action:** Adjust inventory levels based on predicted demand for the upcoming week.
    - **Metrics:** Inventory turnover rate, stockout rate, days of inventory on hand, wherever relevant - product disposal due to expiration.

2. Next Month:

- **Monthly Promotional Calendar**
    - **Actions:** Prepare a promotional calendar. Consider local events that might impact sales. sales trends, and competitor activities. Taylor promotions accordingly.
    - **Metrics:** Sales lift during the event period, event-related product sales, social media engagement, Month-over-month sales growth, overall promotional ROI, customer acquisition cost.
- **Local Marketing:**
    - **Actions:** Explore local marketing opportunities to reach the target audience for each store type. This could include partnerships with local businesses, community events, or targeted advertising.
    - **Metrics:** Website traffic, social media engagement, new customer acquisition.
    - **Type A Stores:**
        - **Action:** Plan and execute a month-long marketing campaign focusing on a specific theme or product category. Coordinate with other stores in the cluster to create a cohesive experience.

Experiment with small-scale events or promotions mid-week to drive traffic beyond weekends.

        - **Metrics:** Month-over-month sales growth, customer acquisition cost, campaign ROI, website traffic, social media engagement.
    - **Types C & D Stores:**
        - **Action:** Tailor product displays and promotions to align with the Type A store's monthly campaign. Offer complementary products or services to capitalize on cross-promotion opportunities.
        - **Metrics:** Sales of promoted items, cross-selling rates, customer basket size, customer feedback on promotions.

3. Long-Term Planning:

- **Store-Specific Strategies**
    - **Actions:** Develop individualized strategies for each store, considering its unique location, customer base, and competitive landscape. This might involve tailoring the product mix, adjusting pricing, or implementing specific marketing initiatives.
    - **Metrics:** Store-level sales growth, customer lifetime value, market share within the store's catchment area.
    - **Type A Stores:**
        - **Actions:**
            - Invest in market research to better understand the local customer base and their evolving needs.

        - Explore potential partnerships with complementary businesses to enhance the customer experience.
        - Develop a long-term strategy for managing sales variability and optimizing inventory management.
        - Evaluate the potential for store expansion or format changes within the cluster.
    - **Metrics:** Market share, customer lifetime value, customer satisfaction, inventory turnover, return on investment for store improvements.
- **Types C & D Stores:**
    - **Action:** Focus on building strong customer relationships and loyalty programs. Specialize in niche products or services to differentiate from competitors. Explore opportunities for online sales or delivery services to expand reach.
    - **Metrics:** Customer retention rate, customer lifetime value, online sales growth, delivery service efficiency.

- **All Stores:**
    - **Actions:**
        - Implement a robust data analytics system to track key performance indicators (KPIs) and monitor the effectiveness of sales strategies.
        - Build a web-based dashboard available to all management levels.

        - Rotate employees between well-performing and underperforming stores and monitor impact.
        - Experiment mixing type A and other stores in the same cluster.
    - **Metrics:** Overall sales growth, profitability, employee satisfaction, customer satisfaction, environmental impact metrics.