Group Members: Sri Ram Theerdh Manikyala, Sai Hemanth Kilaru

Course Title: INFO511 Foundations in Data Science

Term name and year: Spring 2025

Submission: Final Project Report

Instructor’s Name: Angela Cruze

Date of Submission: 05-11-2025

Introduction

In today's fast-paced digital marketplace, understanding what drives consumer choices in electronics is crucial for brands aiming to stay ahead. This project explores the purchasing patterns and trends surrounding consumer electronics, using data-driven techniques to uncover actionable insights. From how product reviews and ratings influence recommendations to how seasonal trends and pricing fluctuations affect sales; we dive deep into what makes consumers say “yes” to a product.

Our analysis combines statistical methods, sentiment analysis, and forecasting models to not only describe current behaviours but also predict future ones. By leveraging tools like Pandas, Plotly, and advanced models such as ARIMA and BART for sentiment evaluation, this study bridges the gap between raw consumer data and meaningful strategic decisions. Ultimately, the goal is to provide businesses with smarter ways to align product positioning, timing, and marketing with actual consumer behaviour.

Dataset Information

The dataset used for this project includes 1,000 consumer electronics entries, each enriched with a variety of product-related attributes. These include final sale prices, customer ratings, review counts, textual sentiment from customer feedback, and recommendation indicators. Together, these data points form a well-rounded snapshot of both product performance and consumer perception.

What makes this dataset particularly suitable for analysis is its rich metadata structure. Each entry offers multiple dimensions for exploration—quantitative fields like price and rating are complemented by qualitative insights drawn from textual reviews and sentiment scores. This dual-layered data format allows for nuanced analysis, making it ideal for pilot studies, market trend evaluations, and behavioral research.

In addition, the dataset serves as a strong foundation for data-driven decision-making. By examining the interactions between variables such as sentiment, price, and recommendation likelihood, businesses can tailor their strategies more effectively. Whether it’s optimizing product descriptions, adjusting pricing strategies, or timing product launches, the insights gained from this dataset can guide more informed, evidence-based actions.

Methodology

1. Data Collection

The dataset was sourced and managed using Databricks, which enabled efficient data access and integration. By loading the CSV files directly into a cloud-based pipeline, we ensured consistent and scalable ingestion, which is particularly useful when expanding the scope of analysis or integrating additional data streams.

2. Data Cleaning and Preprocessing

To prepare the data for analysis, we conducted thorough cleaning using Pandas and Regex. This step involved:

- Handling missing or null values
- Standardizing formats (e.g., price and date entries)
- Removing unwanted characters from text fields
  This careful preprocessing ensured that the dataset was reliable and analytics-ready, minimizing the risk of skewed results due to data quality issues.

3. Analytical Techniques

The project employed several analytical techniques to extract insights:

- Sentiment Analysis: Using the BART transformer model from Hugging Face, we classified customer reviews into positive, neutral, or negative sentiment categories. This provided a rich layer of qualitative insight.
- Correlation and Visualization: Relationships between variables such as price, ratings, and recommendation percentages were explored through heatmaps and scatter plots.
- Time Series Forecasting: An ARIMA model was used to capture seasonal trends and forecast future sales behavior based on historical data patterns.

Research Question 1

What factors most influence product recommendations in consumer electronics?

To explore this question, we examined how various features—like product ratings, sentiment from customer reviews, and review counts—correlate with the likelihood of a product being recommended.

Correlation Analysis

We began with a correlation heatmap to identify which features had the strongest linear relationships with the recommendation percentage. Among all variables, **customer rating** showed the strongest positive correlation, indicating that higher-rated products are more likely to be recommended. This validated the intuitive assumption that satisfied customers tend to endorse products they like.

Rating vs Recommendation

A detailed rating-recommendation plot further emphasized this pattern. Products with ratings above 4.5 consistently had higher recommendation rates, while those with lower ratings saw a noticeable drop. This highlights the role of perceived quality in influencing word-of-mouth and trust-based purchasing behaviour.

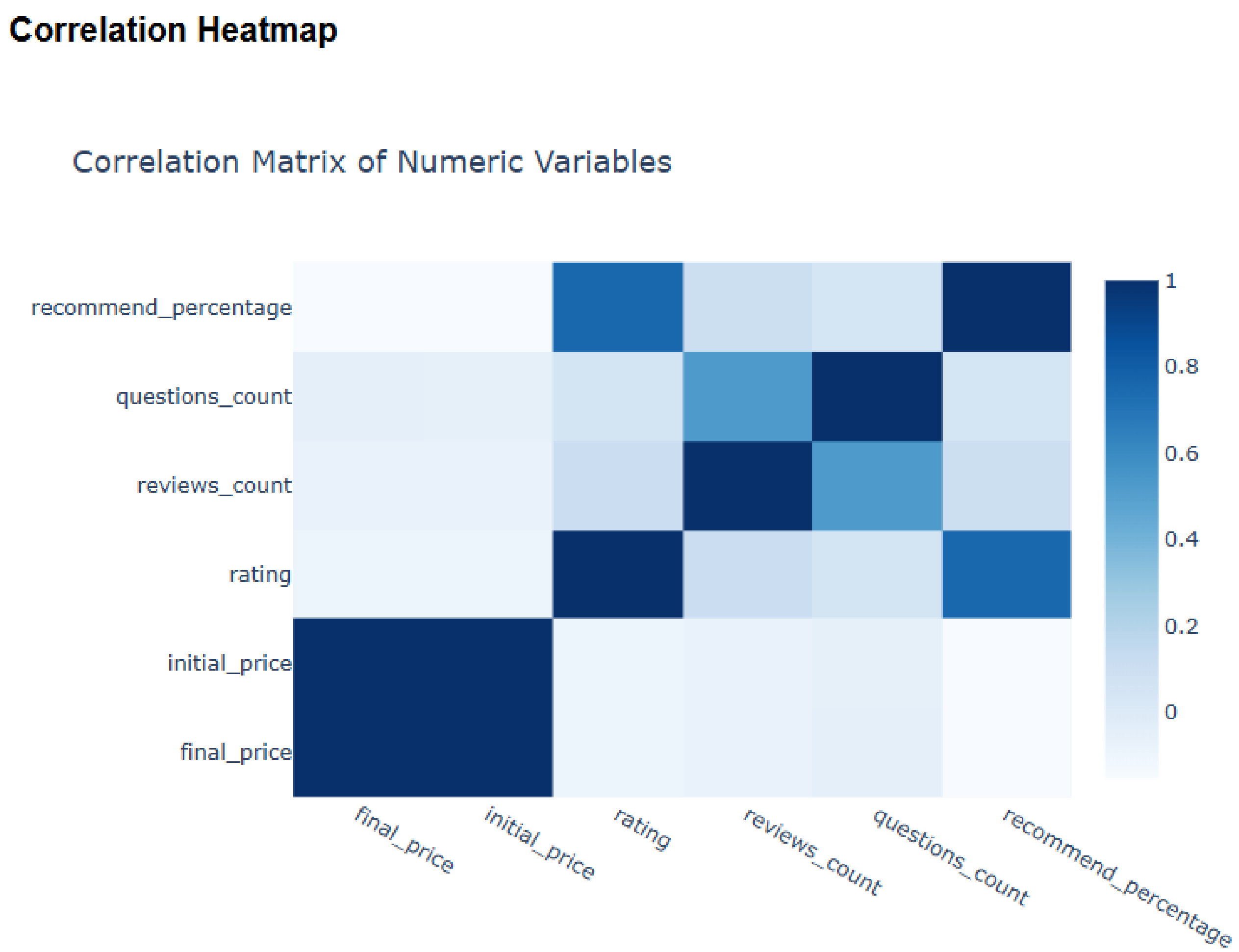
Role of Sentiment

Next, we analysed the sentiment extracted from customer reviews using the BART model. A tree map visualization illustrated how **negative sentiment sharply reduced the chances of a product being recommended**, regardless of other factors. Even moderately priced or feature-rich products were less likely to be recommended if customer sentiment was poor. This finding underscores the powerful role that emotional tone and user experience play in shaping consumer behaviour.

Dashboard Analysis – Recommendation Insights

To visually support the findings from Research Question 1, an interactive dashboard titled **"What Factors Affect Recommendation Percentage?"** was created. This dashboard showcases four key visualizations that help explain the relationships between product features and recommendation trends.

Plot 1 : Correlation Heatmap

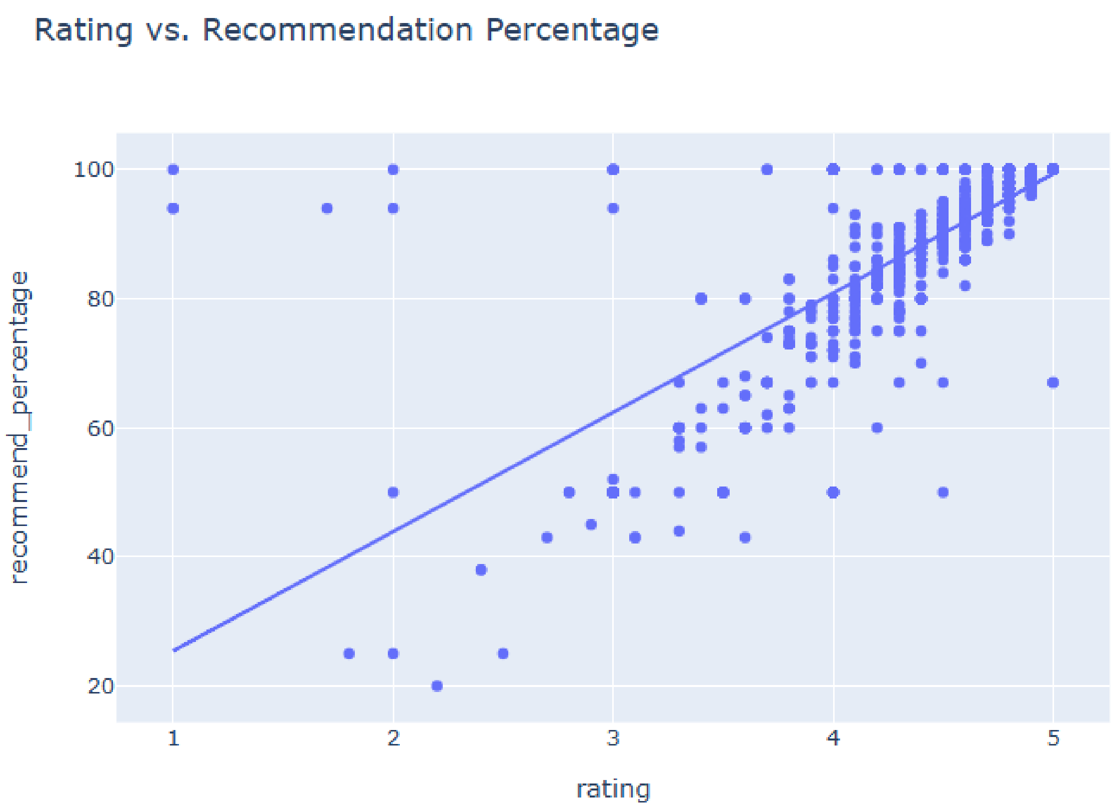


Correlation Heatmap

- **Rating** shows a strong positive correlation (~0.76) with **recommendation percentage**, confirming that higher ratings significantly drive recommendations.
- **Final and Initial Price** both have a mild negative correlation with recommendation, indicating that pricing alone doesn't positively influence consumer endorsement.
- **Reviews Count** and **Questions Count** show only weak correlations, suggesting that while engagement matters, it's not as decisive as sentiment or ratings.

Plot 2 : Rating vs Recommendation Plot

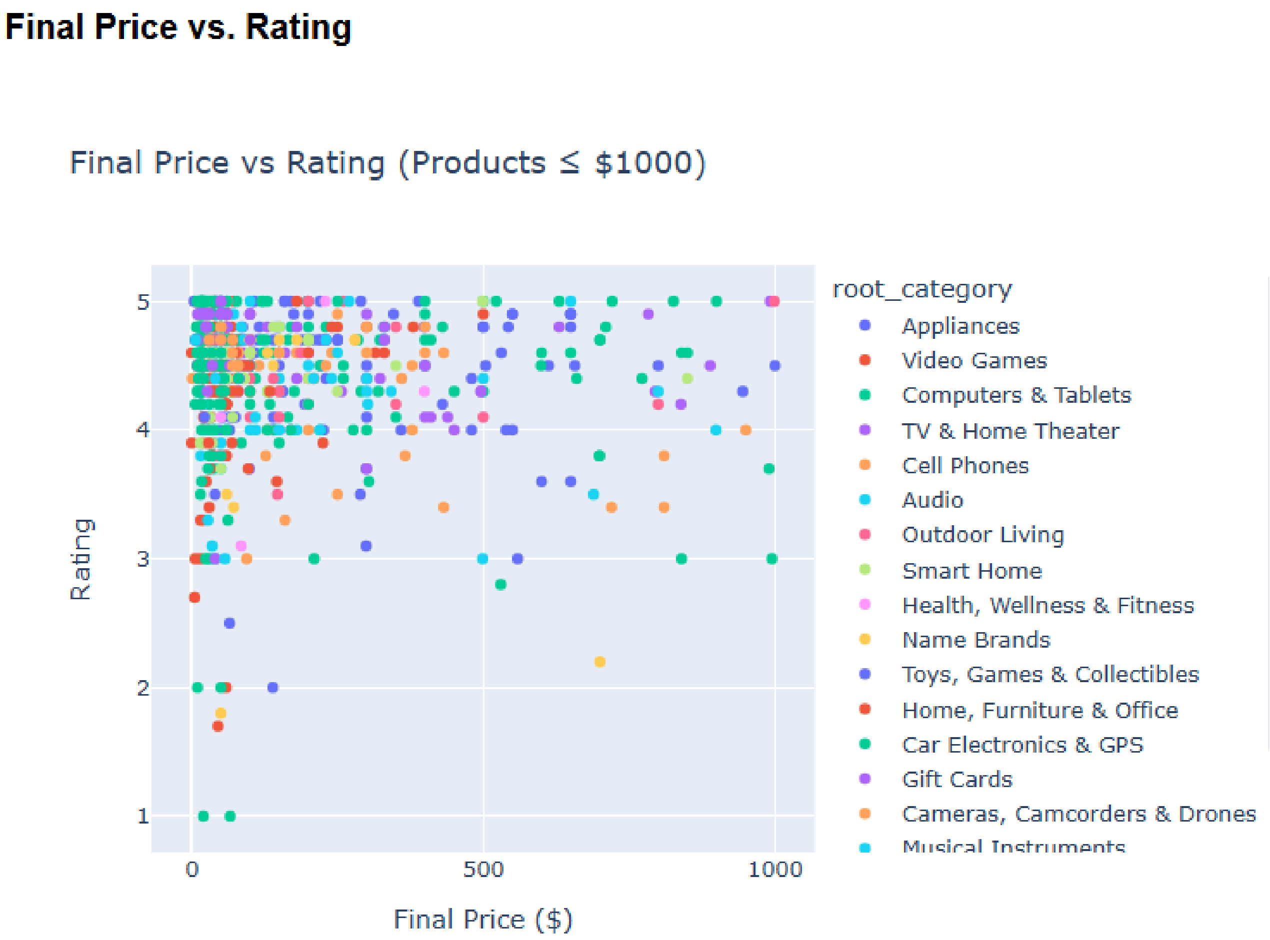
Rating vs. Recommendation %



Rating vs Recommendation Plot

- Products with a **rating above 4.5** show a steep increase in recommendation likelihood, often crossing **90% recommendation**.
- There's a visible threshold effect: **ratings below 4.0** correspond with noticeably fewer recommendations, hinting at a trust cutoff for many buyers.
- The trend line clearly reflects a positive slope, reaffirming rating as a core recommendation driver.

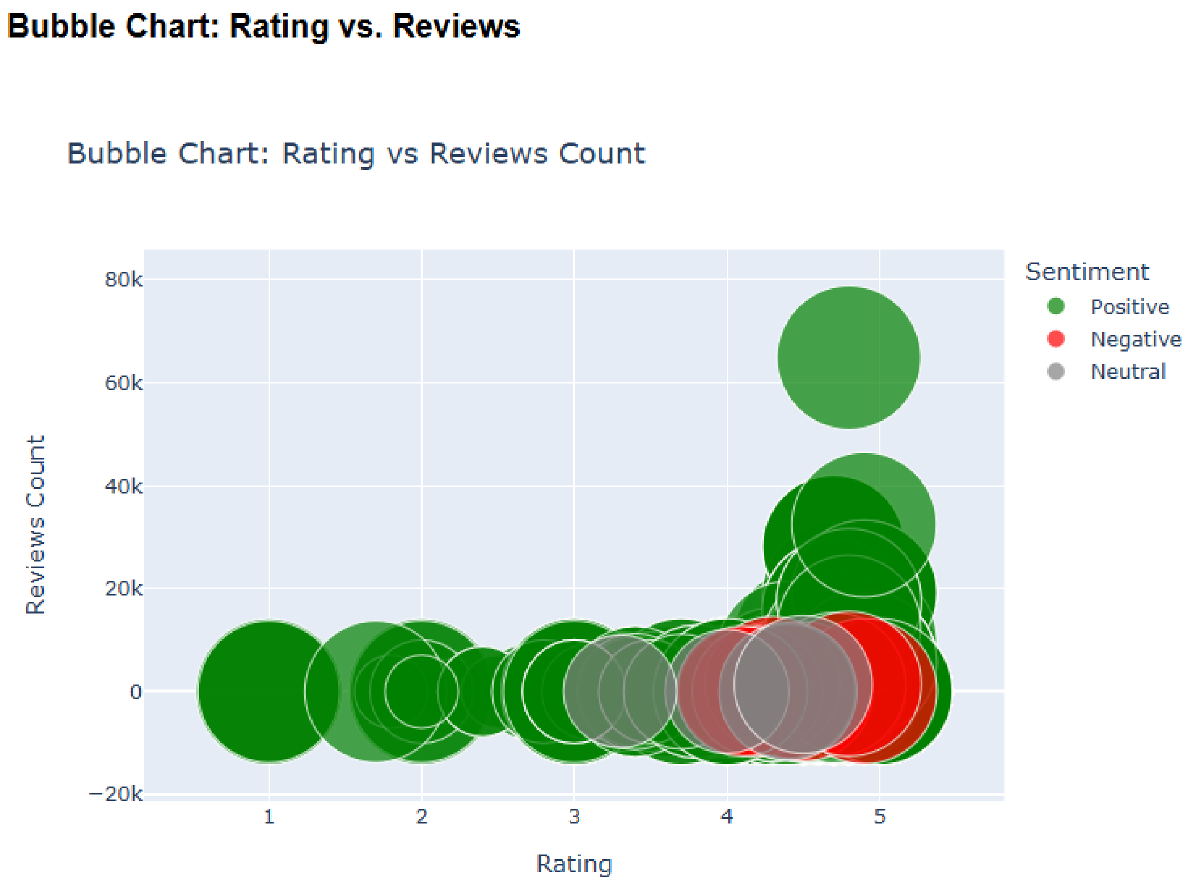
Plot 3 : Final Price vs. Rating



Final Price vs. Rating

- The relationship between **final price and rating** appears **weak or scattered**, with no clear linear trend.
- High-priced products are present across both high and low rating ranges, indicating that **price doesn't guarantee satisfaction**.
- Several **mid-priced products (around $30–$100)** tend to cluster in the higher rating range, suggesting a **sweet spot** for price-to-value perception.

Plot 4 : Bubble Chart: Rating vs. Reviews



Bubble Chart: Rating vs. Reviews

- Each bubble's size represents the number of reviews, offering a third dimension of insight.
- Products with **higher ratings and large review counts** indicate strong customer satisfaction and wide engagement—ideal for promotion.
- A few products with **moderate ratings** still have **large bubbles**, meaning they're frequently reviewed but not necessarily well-liked—potential flag for quality concerns.
- Some **highly rated products with low review counts** suggest emerging or niche items that are performing well in early stages.

Research Question 2

How do seasonal trends and pricing fluctuations influence consumer purchasing behaviour in electronics?

Understanding when and why consumers buy is just as important as knowing what they buy. This part of the analysis focused on uncovering patterns tied to seasonal spikes, pricing over time, and customer satisfaction across the calendar year.

Seasonal Purchase Trends

Using monthly data on product releases and sales volume, we observed a clear **seasonal spike in December**, aligning with major holiday shopping periods like Black Friday and Christmas. Other moderate peaks occurred in September and November, possibly tied to back-to-school and early holiday promotions. These patterns suggest the importance of **timing product launches** to coincide with high-demand seasons.

Yearly Price Variations

The **average final price per year** varied significantly over time. While earlier years showed steady growth, there was an unusual price spike around **2019**, possibly reflecting high-value product launches or inflation effects. Recent years have shown a normalization trend, with prices stabilizing in a more accessible range. This trend helps in understanding **consumer price tolerance** and sets benchmarks for competitive pricing strategies.

Rating Stability Over Time

Interestingly, **average ratings remained relatively stable across years**, staying close to 4.5 out of 5. This implies that regardless of pricing or volume changes, **product satisfaction has been consistently high**, which may indicate better product-market fit and improved quality control in recent product lines.

Forecasting Future Sales

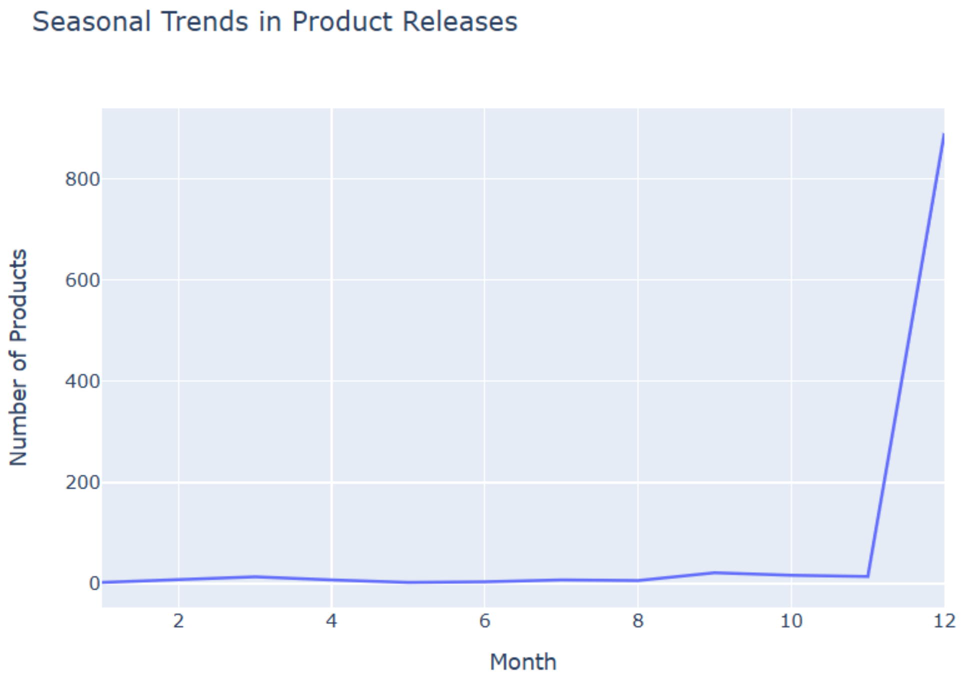
To predict future purchasing behaviour, an **ARIMA model** was applied, considering seasonality and historical sales data. The forecast aligns with past spikes, predicting higher activity in Q4. This reinforces the need for **early inventory planning and targeted marketing** during peak shopping windows.

Dashboard Analysis – Seasonal & Pricing Trends

An interactive dashboard titled **"Seasonal & Price Trends in Consumer Electronics"** was created to support Research Question 2. It visualizes key patterns in product release timing, price changes, customer ratings, and seasonal demand. These insights help businesses optimize launch schedules, adjust pricing strategies, and plan around consumer buying cycles more effectively.

Plot 1 : Seasonal Trends in Product Releases

Seasonal Trends in Product Releases

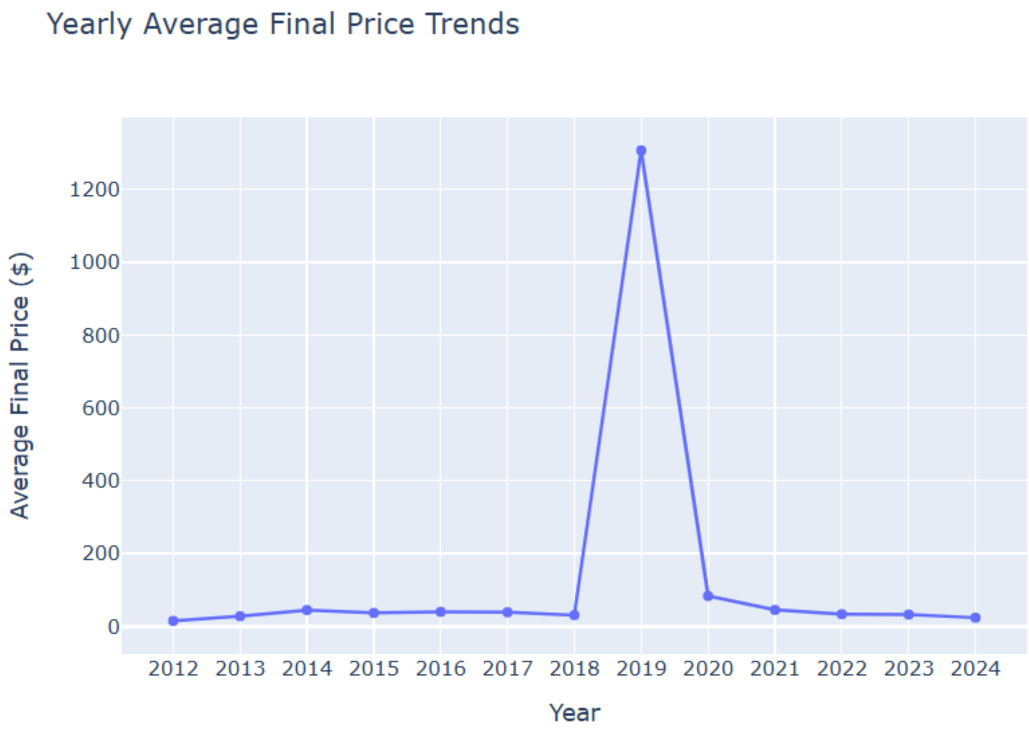


Seasonal Trends in Product Releases

- A dramatic **spike in product releases occurs in December**, confirming strong alignment with holiday sales campaigns.
- Other noticeable increases are seen in **October and November**, which aligns with pre-holiday shopping and promotional seasons.
- The lowest product activity occurs during **Q1 (January–March)**, indicating it may not be the most strategic time for launching new electronics.

Plot 2 : Yearly Average Final Price

Yearly Average Final Price

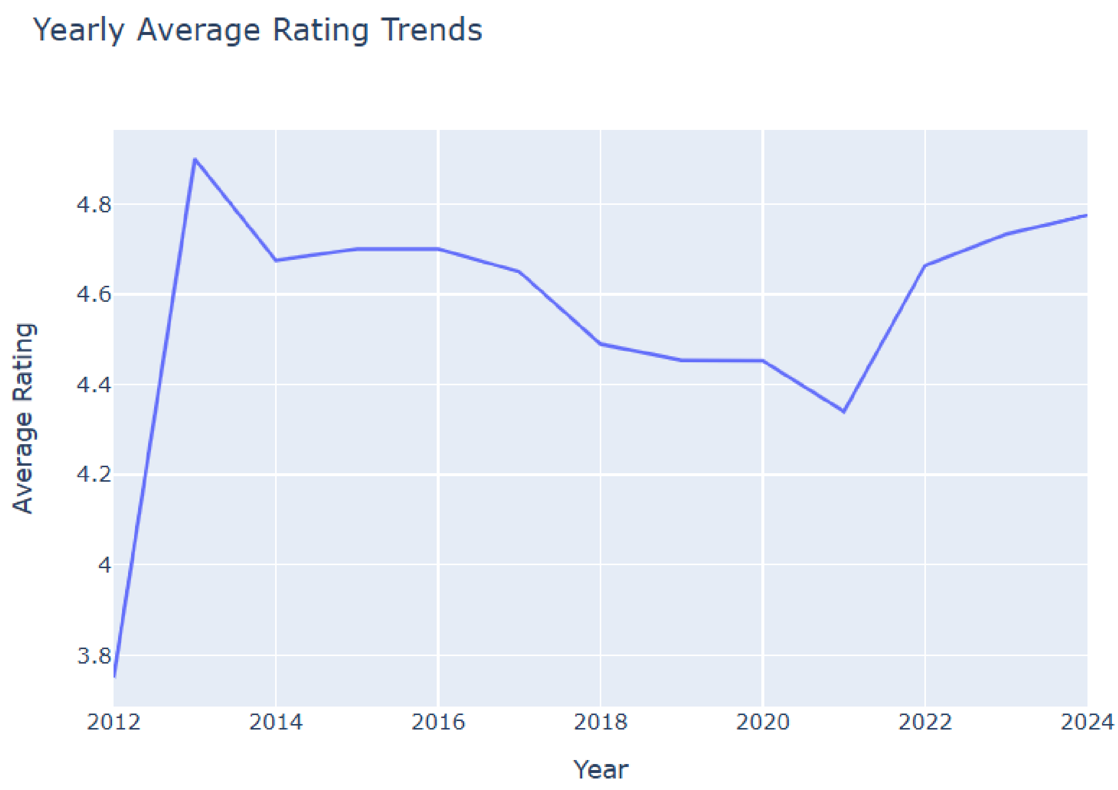


Yearly Average Final Price

- Prices showed **fluctuations over the years**, with a sharp and unusual **price surge in 2019**—possibly due to premium product introductions or a shift in data quality.
- After 2020, prices appear to **stabilize in the $30–$80 range**, suggesting a more competitive and consumer-friendly pricing environment.
- This helps brands identify **years with pricing anomalies** and reassess pricing strategies accordingly.

Plot 3 : Yearly Average Rating

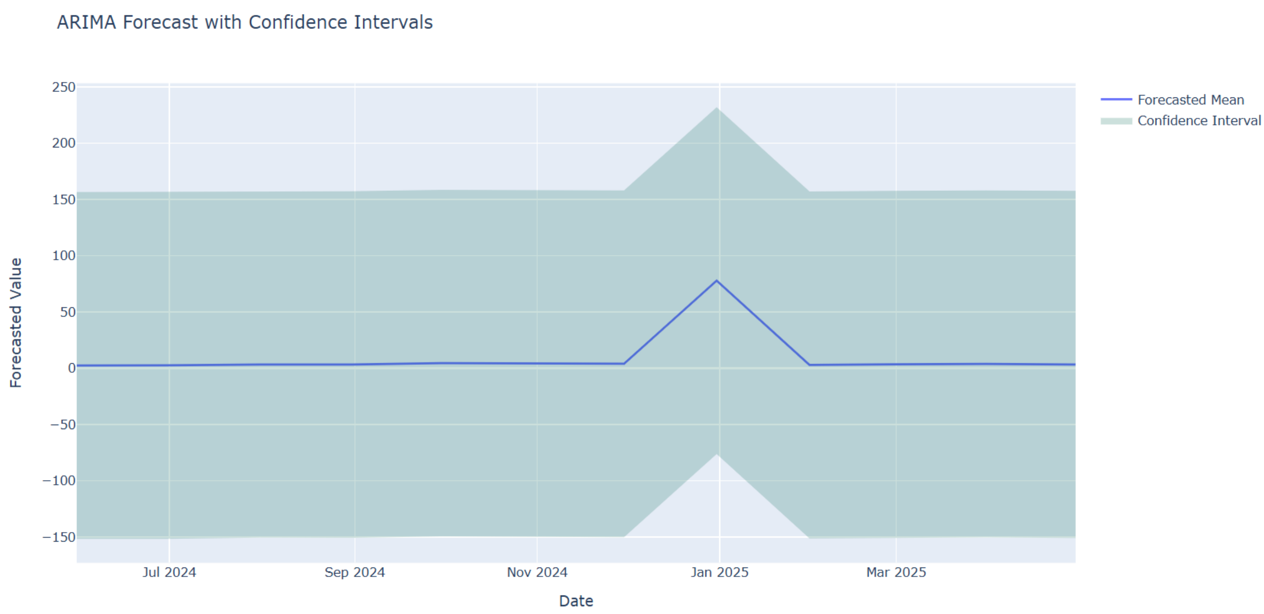
Yearly Average Rating



Yearly Average Rating

- Despite fluctuations in price, **average product ratings remain consistently high**, mostly between **4.4 to 4.9** across all years.
- This indicates that **consumer satisfaction hasn't been significantly affected by pricing trends**, and that quality perception has been largely positive.
- The steady rating trend suggests that **customer expectations are being met** or exceeded consistently over time.

Plot 4 : ARIMA Forecast Plot (Sales Forecasting)



ARIMA Forecast Plot (Sales Forecasting)

- The ARIMA model predicts future **seasonal sales spikes**, especially in **Q4**, aligning well with historical holiday trends.
- Forecasted dips in early quarters highlight **off-season months**, useful for planning maintenance cycles, marketing boosts, or bundling strategies during low-traffic periods.
- These insights help inform **inventory and advertising planning**, enabling more targeted campaign execution around projected peaks.

Challenges Faced

While the project yielded valuable insights, several challenges had to be addressed along the way.

- **Data Quality Issues**: The raw dataset contained inconsistencies such as missing values, unusual price spikes, and irregular text formatting in reviews. Cleaning and standardizing this data required significant preprocessing with Pandas and Regex to ensure accurate analysis.
- **Limited Dataset Size**: With only 1,000 entries, the dataset offered a strong starting point but limited the depth of some advanced modeling techniques. Broader generalizations were made cautiously, and some patterns—especially in forecasting—would benefit from a larger sample.
- **Sentiment Classification Complexity**: Implementing BART for sentiment analysis brought high accuracy but interpreting nuanced or ambiguous text proved challenging. Some neutral or mixed-tone reviews didn't fit neatly into sentiment categories, potentially affecting correlation strength.
- **ARIMA Model Tuning**: Building a reliable ARIMA forecast required multiple iterations to select the right seasonal parameters. The short time range and limited data points added complexity to modelling long-term trends.

Despite these hurdles, thoughtful workarounds and tool optimizations ensured that the analysis remained robust, insightful, and relevant.

Results Summary

The analysis revealed several key factors that influence consumer purchasing decisions in the electronics sector. The most consistent finding was the **strong link between product ratings and recommendation likelihood**—higher ratings almost always led to more recommendations. This affirms the importance of maintaining product quality and encouraging satisfied customers to leave reviews.

Another major insight came from the sentiment analysis. **Negative sentiment in reviews significantly reduced the chance of a product being recommended**, even when the product had reasonable ratings or features. This underlines the need for companies to monitor and manage customer feedback closely, as emotional tone directly affects brand reputation and buyer confidence.

From a pricing perspective, **moderately priced products tended to receive higher satisfaction scores**, suggesting that consumers often seek a balance between value and performance rather than simply going for the cheapest or most expensive option.

Finally, **seasonal patterns** played a major role in influencing purchasing behavior. The data showed a clear surge in product launches and sales around December, indicating the strategic importance of the holiday season. The ARIMA forecast further confirmed this, predicting recurring Q4 demand spikes. This makes it essential for businesses to align marketing, inventory, and promotional efforts with these seasonal cycles.

Rare Insight

Beyond the expected trends, the project uncovered some less obvious yet valuable insights that could significantly impact strategic decision-making.

One standout observation was the **disproportionate impact of negative sentiment**. Even when a product had a high rating or strong feature set, a few strongly negative reviews could noticeably reduce its recommendation percentage. This suggests that customers may give more weight to bad experiences than positive ones—a reminder that **managing negative feedback is just as critical as promoting positive reviews**.

Another subtle insight was that **product price alone doesn’t drive satisfaction**. Some of the most expensive products did not receive the highest ratings or recommendation rates. In contrast, **mid-priced items often hit the sweet spot**, offering perceived value without compromising on performance. This challenges the assumption that premium pricing always equates to consumer approval.

Lastly, the analysis showed that **timing matters more than expected**. Products launched during low-activity months often went unnoticed, regardless of quality or features. This highlights the importance of aligning product releases with periods of high consumer attention, especially the Q4 holiday season.

These insights, though not immediately obvious, offer actionable value for improving product strategies, pricing models, and release schedules.

Conclusion

This project set out to uncover what drives consumer behaviour in the electronics market, and the findings delivered both expected patterns and deeper strategic insights. At the core, **product ratings emerged as a key influencer**, strongly tied to whether consumers would recommend a product. Similarly, **customer sentiment**, especially when negative, played a crucial role in shaping perception and influencing future buyers.

We also found that **mid-range pricing tends to correlate with higher satisfaction**, suggesting consumers value a balance of affordability and performance more than extreme pricing at either end. Meanwhile, the **impact of seasonal trends** became impossible to ignore—most product activity and purchasing spikes occurred around the holiday season, emphasizing the value of timing in product launches and promotions.

Through a mix of data exploration, sentiment modelling, and time series forecasting, this study highlights the multifaceted nature of consumer decision-making. These insights not only help us understand existing trends but also equip businesses with the tools to better plan, adapt, and connect with their audience.

Future Scope

While this project provided a solid foundation for understanding consumer electronics trends, there's ample opportunity to build on these insights with more advanced tools and broader datasets.

- **Scaling the Dataset**: Expanding the dataset beyond 1,000 entries would allow for more robust pattern detection, stronger statistical confidence, and better generalization across different product categories and time periods.
- **Incorporating Advanced Models**: Future iterations could integrate models like **XGBoost** or **deep learning architectures** to enhance prediction accuracy, especially for recommendation scoring and sentiment classification.
- **Real-Time Sentiment Monitoring**: Building a live dashboard that tracks customer sentiment in real-time could give businesses an edge by allowing them to respond quickly to shifts in public perception.
- **Geographic and Demographic Segmentation**: Adding user demographic data or regional sales patterns could uncover location-specific trends, helping tailor marketing and pricing strategies more precisely.
- **Promotional Impact Analysis**: Future work could also explore how discount campaigns, bundling, or limited time offers affect purchase and recommendation behaviour over time.

References

1. **Pandas Documentation** – Data manipulation and cleaning
   https://pandas.pydata.org/docs/
2. **Regex (Regular Expressions)** – Pattern matching in text preprocessing
   https://docs.python.org/3/library/re.html
3. **Hugging Face Transformers** – BART model for sentiment analysis
   https://huggingface.co/docs/transformers
4. **Plotly** – Interactive data visualization
   https://plotly.com/python/
5. **ARIMA Model (Statsmodels)** – Time series forecasting
   https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html
6. **Databricks** – Unified analytics platform for data ingestion and exploration
   https://www.databricks.com/
7. **Consumer Behavior Theory** – Understanding purchase decision-making
   Solomon, M. R. (2018). *Consumer Behavior: Buying, Having, and Being* (12th ed.). Pearson.
8. **Sentiment Analysis in Practice** – Application of NLP in e-commerce reviews
   Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.