Amazon Product Reviews Analysis

1. Project Overview

The objective is to apply **Big Data analytics concepts** in a real-world dataset to test a behavioral hypothesis related to online reviews. The focus is on: - Data understanding - Methodology - Correct analytical reasoning - Interpretation of results

---

2. Research Motivation & Study Rationale

Why This Project Exists

Online reviews strongly influence consumer perception, yet there is limited empirical evidence on whether *early reviews* truly shape long-term outcomes.

This project was created to: - Experimentally test a common assumption in e-commerce - Practice large-scale data analysis techniques - Demonstrate correct experimental design using observational data

---

3. Research Question

Main Question:
Does the first rating of a product significantly affect its future sales performance?

Supporting Questions: - Does a bad first review reduce customer engagement? - Do products recover after a poor start? - Can the first rating predict future product quality?

---

3. Data Understanding

3.1 Data Source

- Dataset: Amazon Fine Food Reviews
- Time Period: 1999 – 2012
- Data Type: Customer reviews and ratings

3.2 Data Size

- Total Reviews: ~568,000
- Unique Products: ~74,000
- Unique Users: ~256,000

This is a **large-scale dataset**, which required distributed processing tools.

4. Tools & Technologies Used

- **PySpark**: For handling large-scale data efficiently
- **Spark SQL & DataFrames**: For data manipulation
- **Python (SciPy)**: For statistical analysis

These tools were chosen to ensure scalability, speed, and reliability.

---

5. Notebook Walkthrough (Code Explained in Simple Terms)

This section explains **what each major part of the notebook does**, why it exists, and how it contributes to answering the business question. The goal is understanding, not code syntax.

5.1 Environment Setup & Spark Initialization

What we did: - Initialized a Spark session - Configured Spark to handle large datasets efficiently

Why this matters: The dataset is too large for traditional in-memory tools (like Pandas). Spark allows us to process hundreds of thousands of reviews reliably and quickly.

Business value: Ensures results are scalable and trustworthy for real-world datasets.

5.2 Loading the Dataset

What we did: - Loaded the Amazon Fine Food Reviews CSV file into Spark DataFrames

Why this matters: DataFrames provide structured, SQL-like access to the data, making transformations safer and easier.

Business value: Allows consistent handling of all products and reviews without sampling or manual filtering.

5.3 Data Cleaning

What we did: - Converted Unix timestamps into readable dates - Validated rating values (1–5 only) - Removed incomplete or invalid records

Why this matters: Time-based analysis requires accurate dates, and invalid ratings can distort conclusions.

Business value: Guarantees that insights are based on clean and reliable data.

5.4 Identifying the First Review Per Product

What we did: - Used Spark window functions to sort reviews chronologically per product - Extracted the very first rating each product received

Why this matters: The entire research question depends on identifying the *true first impression*.

Business value: Ensures that the "first rating" is not approximated or guessed, but precisely calculated.

5.5 Product Filtering (Minimum Reviews Rule)

What we did: - Filtered out products with fewer than 5 total reviews

Why this matters: Products with very few reviews do not provide enough data to measure future performance.

Business value: Prevents misleading conclusions based on noise or one-time interactions.

5.6 Feature Engineering

What we did: For each product, we computed: - Total number of reviews - Reviews per month (review velocity) - Normalized review rate (adjusted for product age) - Average rating after the first review

Why this matters: Raw review counts alone are unfair—older products naturally have more reviews.

Business value: Allows fair comparison between products launched at different times.

5.7 Grouping Products by First Rating

What we did: - Categorized products into Low (1–2), Medium (3), and High (4–5) first-rating groups

Why this matters: Grouping simplifies comparisons and highlights behavioral patterns.

Business value: Transforms raw data into clear segments that decision-makers can reason about.

5.8 Statistical Testing Logic (Conceptual Explanation)

What we did: - Used non-parametric tests to compare groups - Measured correlations between first rating and future performance

Why this matters: Review data is skewed and noisy—traditional assumptions do not hold.

Business value: Ensures conclusions are statistically valid and not artifacts of poor assumptions.

5.9 Validation Checks

What we did: - Checked whether product age is correlated with first rating

Why this matters: If older products systematically had higher or lower first ratings, results would be biased.

Business value: Confirms that findings are driven by ratings, not timing effects.

---

6. Feature Engineering (What We Built From the Data)

For each product, we calculated:

- **First Rating**: The earliest rating the product received
- **Total Reviews**: Total number of reviews
- **Review Velocity**: Reviews per month
- **Normalized Review Rate**: Adjusted for product age
- **Future Average Rating**: Average rating after the first review

These metrics allow us to compare products fairly, even if they were launched at different times.

---

7. Grouping Strategy

Products were grouped based on their **first rating**:

- **Low**: 1–2 stars
- **Medium**: 3 stars

- **High**: 4–5 stars

This makes it easier to compare performance patterns across different starting impressions.

---

8. Analysis Approach (Without Heavy Statistics)

We focused on answering three simple questions:

1. Do products with high first ratings get more reviews?
2. Do products with low first ratings fail?
3. Does the first rating reflect true product quality over time?

We used non-parametric statistical methods because: - Review data is not normally distributed - Results are more robust and realistic

---

9. Key Findings (What the Data Told Us)

9.1 Impact on Sales (Review Volume)

- Products with **low first ratings** receive almost the **same number of reviews** as products with high first ratings
- No meaningful difference in review growth rate

Conclusion: First rating does NOT affect sales performance

9.2 Product Quality Over Time

- Products with bad first ratings **improve significantly** over time
- Products with high first ratings tend to stabilize or slightly decrease

First rating has a **weak but real relationship** with future product quality

---

9.3 The "Controversy Effect" (Important Insight)

- Products that start with **3-star ratings** receive the **highest engagement**

Ambiguous ratings create curiosity and more customer interaction

---

10. Business Insights

For Sellers

- A bad first review does NOT kill a product
- Focus on improving quality instead of panicking

For Platforms

- First reviews should not be heavily weighted in ranking algorithms
- Rating trends over time are more informative

For Decision Makers

- True product quality matters more than first impressions
- Controversy can drive engagement

---

11. Final Answer to the Business Question

Does the first rating affect future sales?

NO

The data clearly shows that the first product rating does not significantly affect sales performance.

However: - It weakly predicts future product quality - Products can recover from bad starts

---

12. Final Recommendation

- Do not overreact to early reviews
- Monitor rating trends, not single ratings
- Focus on long-term quality improvements

---

13. Project Conclusion

This project demonstrates how large-scale data analysis can challenge common assumptions.

Key Takeaway: > First impressions matter less than sustained quality.

---

14. Study Design & Experimental Setup (Important for Defense)

This project is a **pilot experimental study** based on **observational data**, not a controlled experiment.

- No variables were manipulated.
- No causal claims are made.
- The goal is to **test hypotheses and observe patterns**, not to prove cause-and-effect.

Variables Definition

- **Independent Variable:** First product rating
- **Dependent Variables (Outcomes):**
  - Review volume (sales proxy)
  - Review velocity
  - Future average rating
- **Control Considerations:**
  - Product age
  - Time normalization

This design is suitable for exploratory studies and hypothesis validation in real-world data.

---

15. Hypotheses & Testing Strategy

Hypothesis 1: First rating affects future sales performance

- **Metric Used:** Normalized review rate
- **Test Used:** Mann-Whitney U Test + Spearman Correlation
- **Why:**
  - Review data is skewed
  - No normality assumptions

Result: No statistically significant difference

Decision: Hypothesis rejected

Hypothesis 2: First rating predicts future product quality

- **Metric Used:** Future average rating
- **Test Used:** Spearman Rank Correlation

Result: Weak but significant relationship

Decision: Hypothesis accepted (weak effect)

Hypothesis 3: Medium ratings create higher engagement

- **Metric Used:** Review velocity
- **Test Used:** Kruskal-Wallis + Pairwise Mann-Whitney

Result: 3-star products show highest engagement

Decision: Evidence supports the "Controversy Effect"

---

16. Decision Framework (How Results Became Conclusions)

Question	Test Used	Key Result	Final Decision
Does first rating affect sales?	Mann-Whitney, Spearman	$p = 0.83$, $\rho \approx 0$	No effect
Does first rating predict quality?	Spearman	$\rho = 0.187$	Weak prediction
Which products get more engagement?	Kruskal-Wallis	$p < 0.01$	Medium ratings win

All decisions were made based on: - Statistical significance - Effect size - Practical interpretation

---

17. How to Defend This Project (Talking Points)

- This is an **exploratory pilot study**, not production analytics
- Review volume is a **proxy for sales**, commonly used in research
- Non-parametric tests were chosen due to data distribution
- Results challenge intuition, which adds research value
- Findings are limited but directionally informative

---

18. Final Study Conclusion

This pilot study demonstrates that:

- First impressions do not drive long-term engagement
- Product quality reveals itself over time
- Early ratings are weak signals, not destiny

The project successfully meets its academic objective: > Designing, executing, and defending an experimental data study using large-scale real-world data.

End of Report