

# TEAM SAFETY DANCE


 Mariah  
★★★★☆ sharp edges  
Reviewed in the United States on August 18, 2024  
Color: Silver | Size: 24 Pieces | **Verified Purchase**

 Victor Sarmento  
★★★★☆ Product defect claim with explosion risk - NGP STORE USA  
Reviewed in the United States on June 4, 2016


 Amazon Customer  
★★★★☆ OVERHEATING/BATTERY  
Reviewed in the United States on September 18, 2024  
**Verified Purchase**

**Verified Purchase**

 Taylor H.  
★★★★☆ CONTAINS LEAD! DO NOT BUY!  
Reviewed in the United States on March 26, 2016  
Size: 20 LB 270 YDS | Unit Count: 1.0 | **Verified Purchase**

 Kyle Henderson  
★★★★☆ Not safe, do not let child use unattended.  
Reviewed in the United States on July 28, 2023  
Style: Original | **Verified Purchase**


 Sarah V.  
★★★★☆ Possible fire hazard  
Reviewed in the United States on May 24, 2023  
Color: Brushed Silver | **Verified Purchase**

 Desiree  
★★★★☆ Irritated my face on the first use  
Reviewed in the United States on October 2, 2024  
Size: 1 Fl Oz (Pack of 1) | **Verified Purchase**


 Luanne  
★★★★☆ Dangerous Product  
Reviewed in the United States on June 7, 2024  
**Verified Purchase**




# PROACTIVE IDENTIFICATION OF PRODUCT SAFETY ISSUES

 Kerry Carey  
★★★★☆ BEWARE!! TOXIC  
Reviewed in the United States on December 19, 2021  
Color: Multi Color With Silver | **Verified Purchase**

★★★★☆ Smelled like it was going to catch fire  
Reviewed in the United States on July 11, 2024  
Color: Brushed Silver | **Verified Purchase**

 SB  
★★★★☆ Stomach hurts and no refunds for unused bottles as well  
Reviewed in the United States on September 7, 2022  
Size: 60 Count (Pack of 1) | Pattern Name: Vitamins | **Verified Purchase**

 Katanajo  
★★★★☆ Caught Fire Melted Bucket  
Reviewed in the United States on May 12, 2024  
Color: Black | **Verified Purchase**


 Rileigh Shanks  
★★★★☆ I had an allergic reaction...  
Reviewed in the United States on January 10, 2024  
Size: 1.69 Fl Oz (Pack of 1) | **Verified Purchase**

 gn  
★★★★☆ Dangerous malfunction  
Reviewed in the United States on February 10, 2022  
Color: Blue | **Verified Purchase**

 DaniW  
★★★★☆ The smell screams: Toxic!  
Reviewed in the United States on September 15, 2020  
Color: Multi Color | **Verified Purchase**

 Meuy  
★★★★☆ Danger- electric shocks  
Reviewed in the United States on April 5, 2021  
**Verified Purchase**

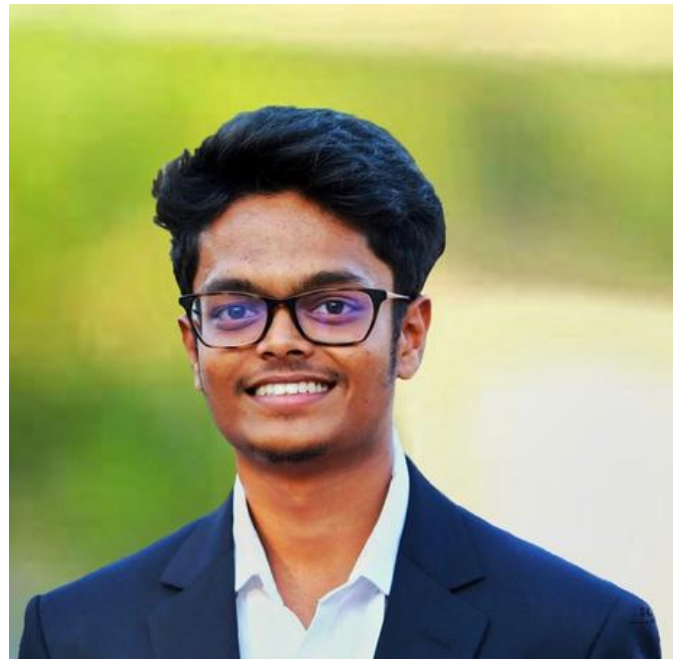
 Celeste  
★★★★☆ Disappointing Experience for Oily Skin: Caused Breakouts  
Reviewed in the United States on January 7, 2024  
Size: 1.69 Fl Oz (Pack of 1) | **Verified Purchase**

 linda\_alexandrias  
★★★★☆ Caused CHEMICAL BURN on my BABY! REGRETFUL/REMORSEFUL  
Reviewed in the United States on April 9, 2021  
Size: Newborn | Unit Count: 31 | **Verified Purchase**

 Sarah Weissman  
★★★★☆ CHOKING HAZARD  
Reviewed in the United States on August 12, 2024  
Style: Original | **Verified Purchase**



# MEET TEAM SAFETY DANCE



**Aakash Boenal**

Scrum Master



**Jonathan King**

Product Owner



**Utkarsh Ganjihal**

Developer



**Saiphani Vuppala**

Developer

# PROJECT PARTNER

## NIRA, INC.

- **Dr. Wen Zhu**, Chief Architect
- **John Oh**, .NET Developer

- NIRA, Inc. is a woman-owned small business and certified 8(a) firm.
- Specializes in Information Technology services for public sector clients.
- Works closely with customers throughout the project lifecycle.
- Key Clients: Federal agencies like CPSC, FDA, FAA, and DTIC.



# NIRA



# PROBLEM CONTEXT

## Domain of problem:

Retail and product safety across various categories.

## Importance of problem:

Unsafe products in the market can lead to serious risks leading to costly recalls and reputational damage for companies.

A proactive approach is essential to avoid these issues.

# PROBLEM STATEMENT

- **The current recall process not proactive**
  - After products already in tens of thousands of consumers' hands.
- Focused on using crowdsourcing, historical recall/incident data, and advanced analytics to identify potential safety risks
- **Baby products used as approach baseline**
  - Pipeline can be applied to many other product categories
- By detecting risks early, aim to reduce defective products reaching too far into the market,
  - Enhance consumer safety
  - Minimize impact of recalls





# DATASETS



## Amazon

**Reviews** - 6M ~ 2.2 GB (Baby products only)  
**Metadata** - Approximate 217K records ~ 600MB



## CPSC

**Incidents** - Approximate 59K records ~ 84MB  
**Recalls** - 9100 records ~ 11MB

**Dataset Location:** Internet

### **Dataset Access:**

<https://amazon-reviews-2023.github.io/>  
<https://www.saferproducts.gov/SPDB.zip>

**Storage & Computing:** GMU ORC Hopper Cluster

### Specification:

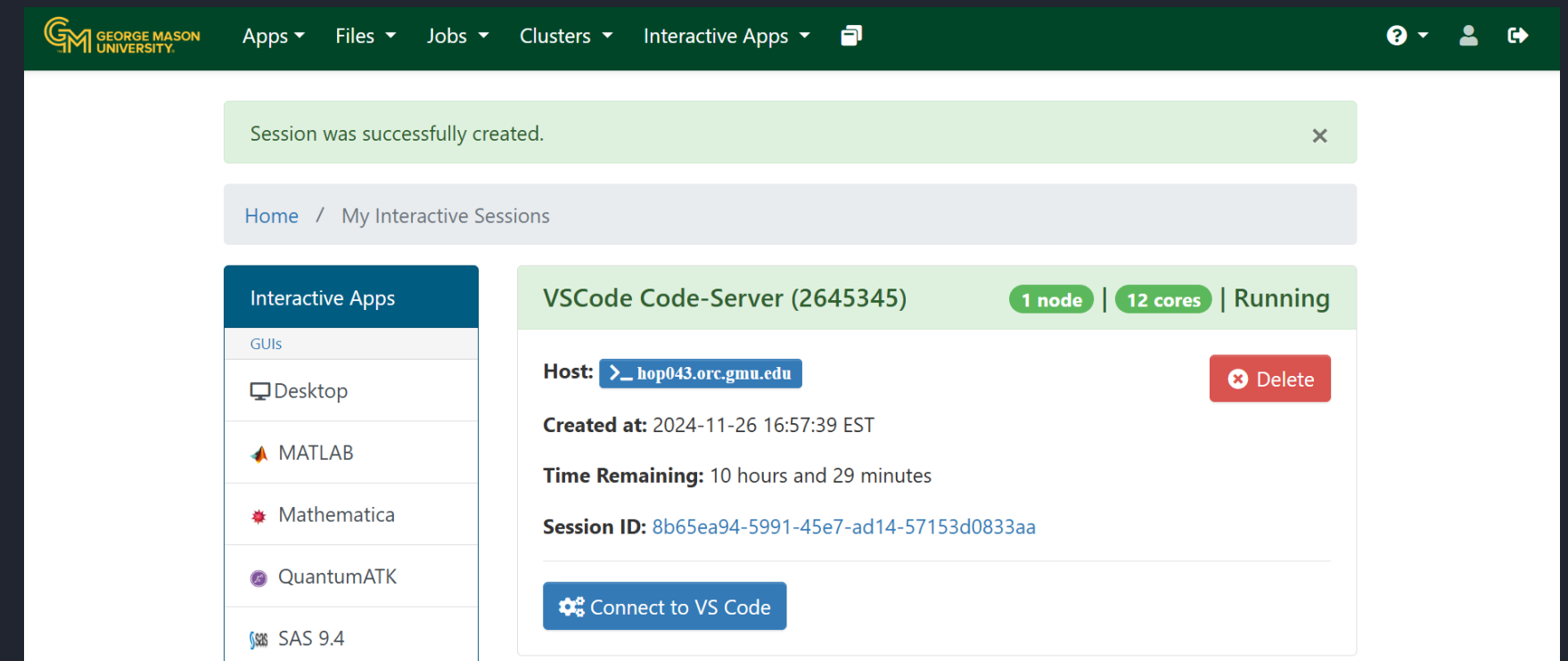
**Node Type:** AMD, Intel

**Partition:** Interactive, Normal, Contrib, GPU, Contrib GPU, BigMemory

**Time limit (in hours):** MAX (12 hours)

**Number of cores:** 1-12

**Memory GBs/core:** 4-8



# WORKFLOW



Downloading raw review data for 'raw\_review\_Baby\_Products'...

Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)...

Downloading Review Data: 100% |██████████| 6028884/6028884 [08:27<00:00, 11873.48records/s]

```
Downloading raw metadata for 'raw_review_Baby_Products'...
Downloading Metadata: 100%|██████████| 217724/217724 [00:49<00:00, 4371.27records/s]
Loaded 217724 metadata records.
```

Data download complete for 'Baby Products' category!

```
Detecting language: 100%|██████████| 5951569/5951569 [1:00:23<00:00, 1642.55it/s]
Final number of rows after removing non-English reviews: 5073260
Percentage removed in language filter: 14.757604255281255
```

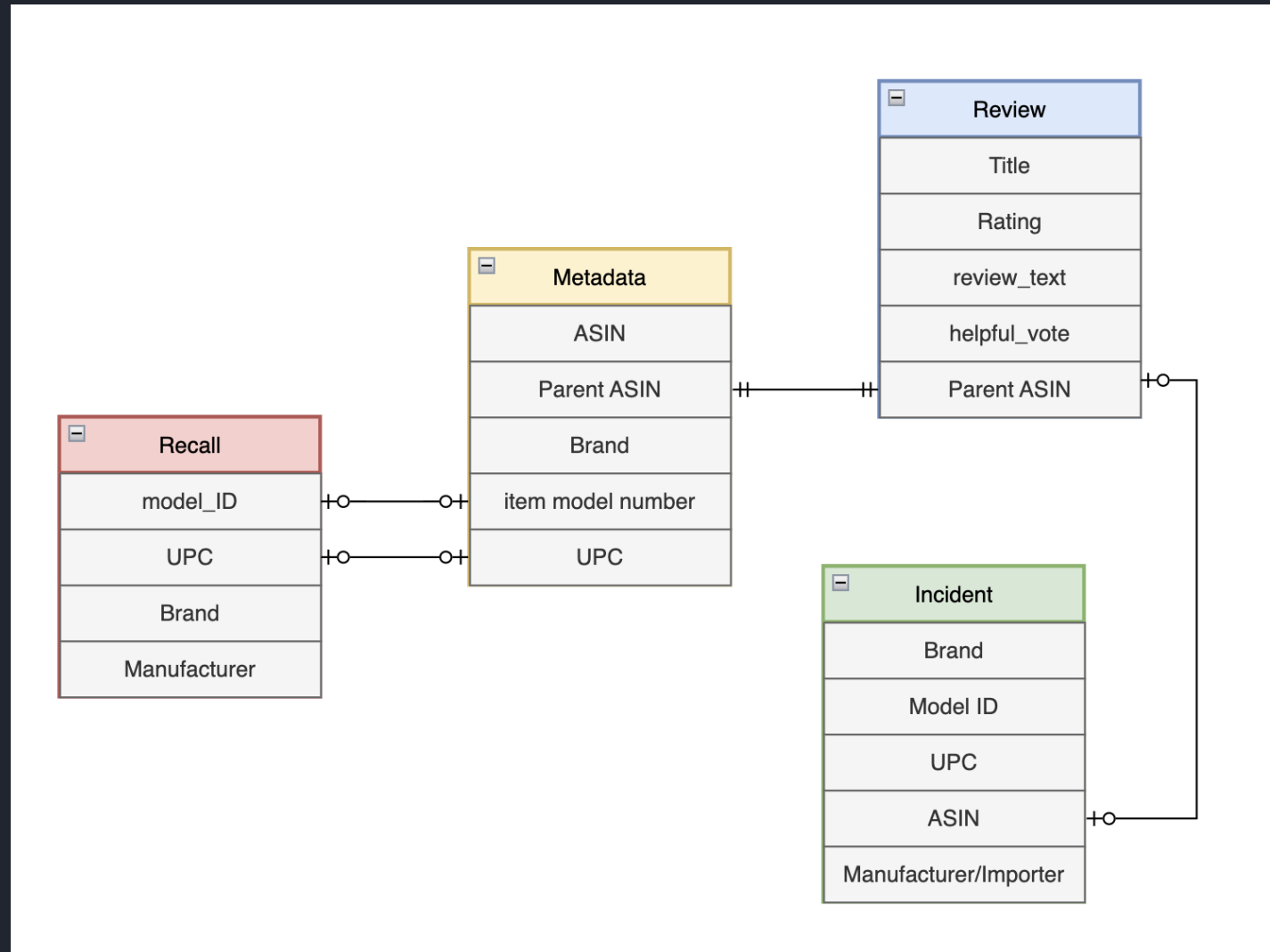
```
item_model_number
True      4997035
False      0
Name: count, dtype: int64
```

## Data Cleaning:

- Both the incidents and the recalls were cleaned manually which includes:
  - Extracting model\_id, UPC codes, ASIN from the model id column in the Incident dataset.
  - Extracting model\_id, UPC codes, Brand, Manufacturers in the Recall dataset

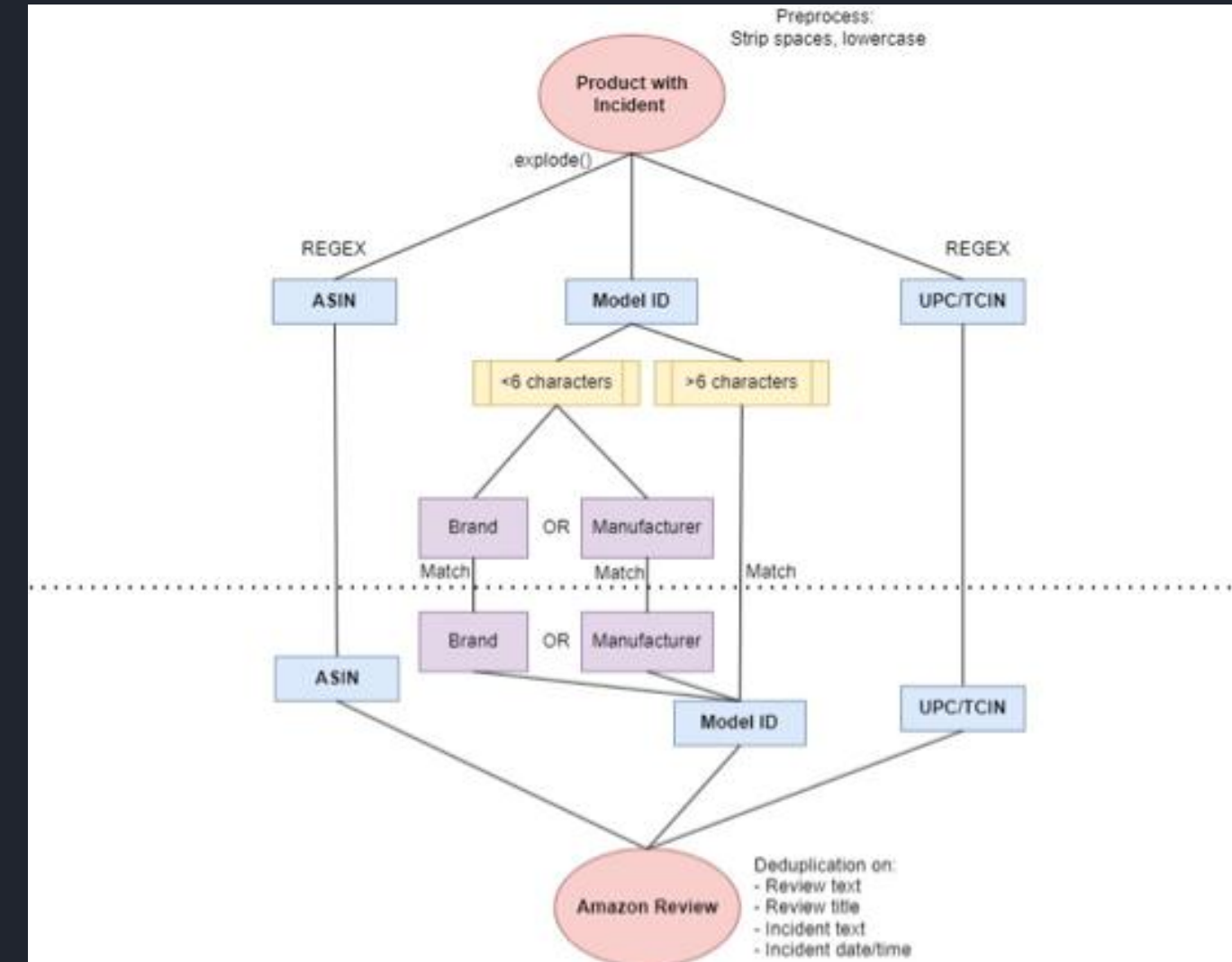
Model Name or Number	Extracted model_id	UPC codes	model.id	brand_name	retail_name	manufacturer	Title
10FM02Q	10FM02Q			IKEA	IKEA		IKEA Offers Free Wall Anchoring Repair Kit for Chests and Dressers Due to Tip-over Hazard After Two Children Died
2705FE	2705FE		RS2300	Delta	Delta		Delta Cycle Recalls Ceiling Hoists with Straps Due to Injury Hazard
68EE967B7DDD	68EE967B7DDD		PS-1000	HALO			HALO 1000 Portable Power Stations Recalled Due to Serious Fire and Burn Hazards; One Death Reported; Imported by ZAGG; Sold by ACG, QVC and Z
Electric Range Model NE63	NE63T8111SS/AA		GJD49	Fisher-Price	-		Fisher-Price Recalls Dumbbell Toy in Baby Biceps Gift Sets Due to Choking Hazard
2 gun safe			TBSD150-20	Head			Head Rush Technologies Recalls TRUBLUE Speed Auto Belay Devices Due to Fall Hazard
NE595ROABSR	NE595ROABSR		702053, 768152				Shawshank LEDz Recalls Squeeze Plush Ball Monsters Toys and Easter Squeezable Toys Due to Injury Hazard; Sold Exclusively at Ace Hardware
DVE50R8500V/A3	DVE50R8500V/A3		4061464174788, 40614641748		ALDI		ALDI Recalls Ambiano Single Serve Coffee Makers Due to Burn Hazard
Model 39N7A	39N7A		E2037,E2038	IKEA	IKEA		IKEA Recalls VARMFRONT Power Banks Due to Fire Hazard
Model P4010DCS-W	P4010DCS-W		DCMWP234U2,DCMWP600X2	Stanley			Stanley Black & Decker Recalls DeWALT Battery Walk-Behind Mowers Due to Laceration Hazard

# DATA MERGE



```
--- Merge Complete. Incident Dataset Merge: Summary Statistics ---
Total records processed: 4997035
Total affected records (involved in an incident): 49538
Percentage of records with incidents: 0.99%
Total ASIN matches: 4423
Total model number matches: 41363
Total UPC matches: 6080

Total unique products involved in incidents: 414
Done with some cleanup
```



```
--- Merge Complete. Recall Dataset Merge: Summary Statistics ---
Total records processed: 4964294
Total affected reviews (involved in a recall): 1714
Percentage of records with recalls: 0.03%
Total ASIN matches: 0
Total model number matches: 1714
Total UPC matches: 0
Total unique products involved in a recall: 36
Filtered dataset with recalls saved as 'checkpoint5_recalls.csv'.
```



# SENTIMENT ANALYSIS DISTRIBUTION

01

Conducted Sentiment Analysis on Raw Amazon Reviews dataset

02

Mapped product ratings to sentiment categories (positive, negative, neutral)

03

Used Vader-Sentiment Analysis model to evaluate the sentiment score of text (text and incident\_description) producing a compound score and sentiment

04

Modified sentiment scores based on rating categories for finer analysis

## RATING DISTRIBUTION

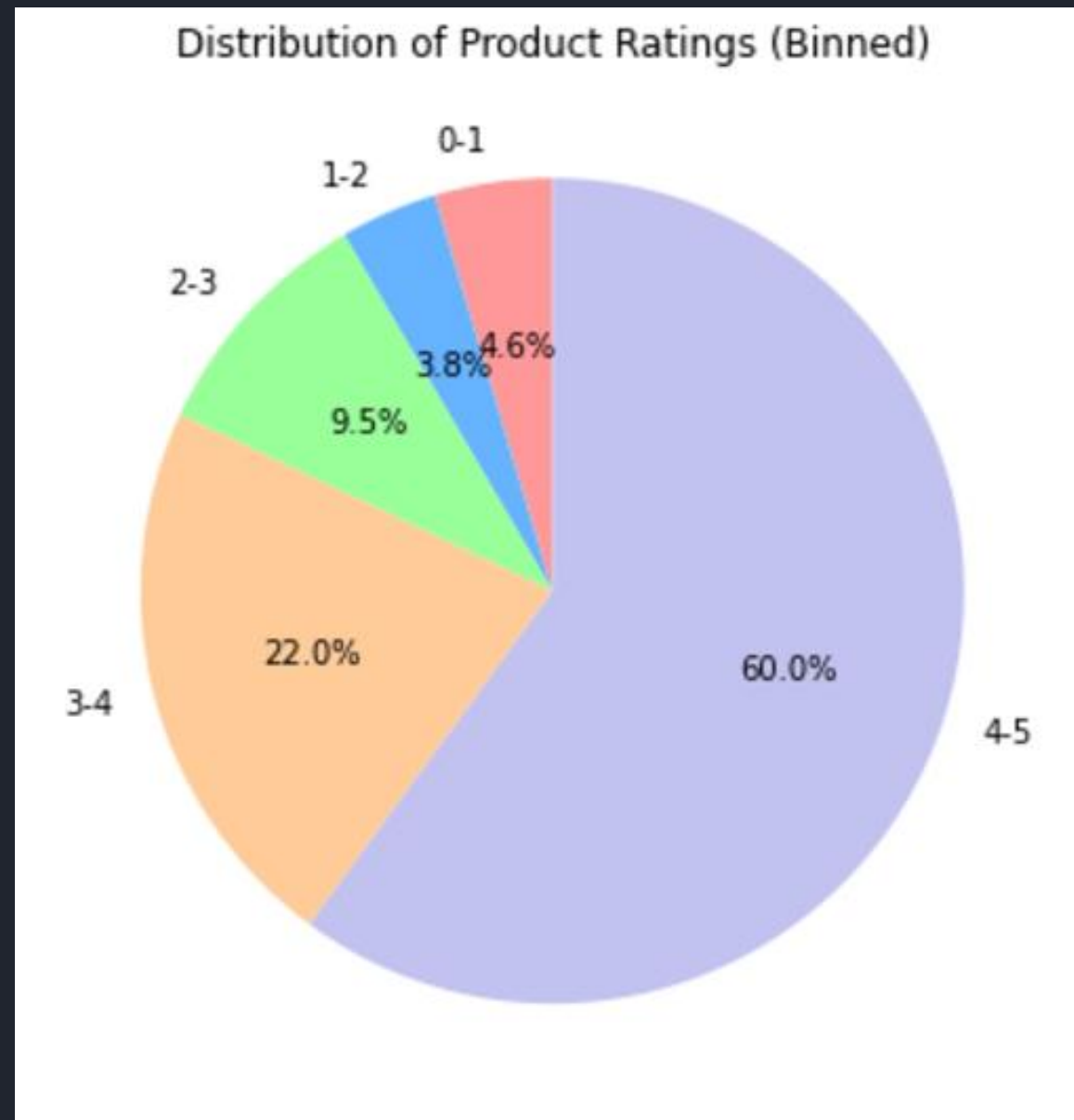


Fig: Pie chart of Product Ratings Distribution

- **Highlights:**

- Proportions of 1-star to 5-star ratings,
  - Comparisons with sentiment analysis
- General customer sentiment towards products

## SENTIMENT ANALYSIS DISTRIBUTION

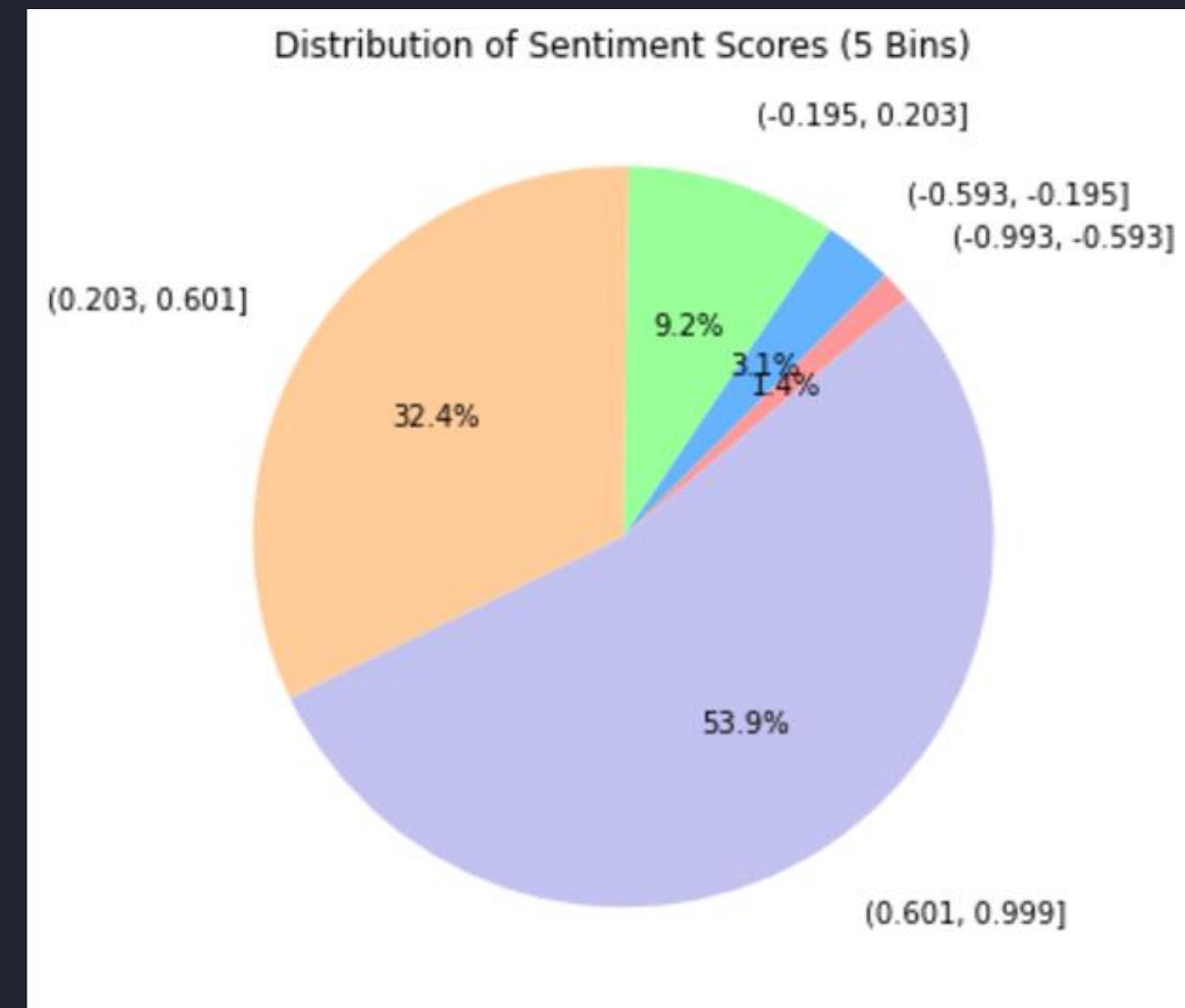


Fig: Pie Chart of Sentiment Scores grouped into bins

- **Highlight:**

- Focus points are negative & neutral sentiment.
- General polarity scores

# SENTIMENT ANALYSIS VISUALIZATION 2

- The histogram shows the distribution of average sentiment scores across all product reviews.
- Sentiment score ranges from -1 to 1
- The predominance of positive sentiment highlights general satisfaction with the products.
- Negative sentiments, although minimal, may correspond to specific product defects or incidents.
- Products with frequent negative sentiment scores should be prioritized for further investigation

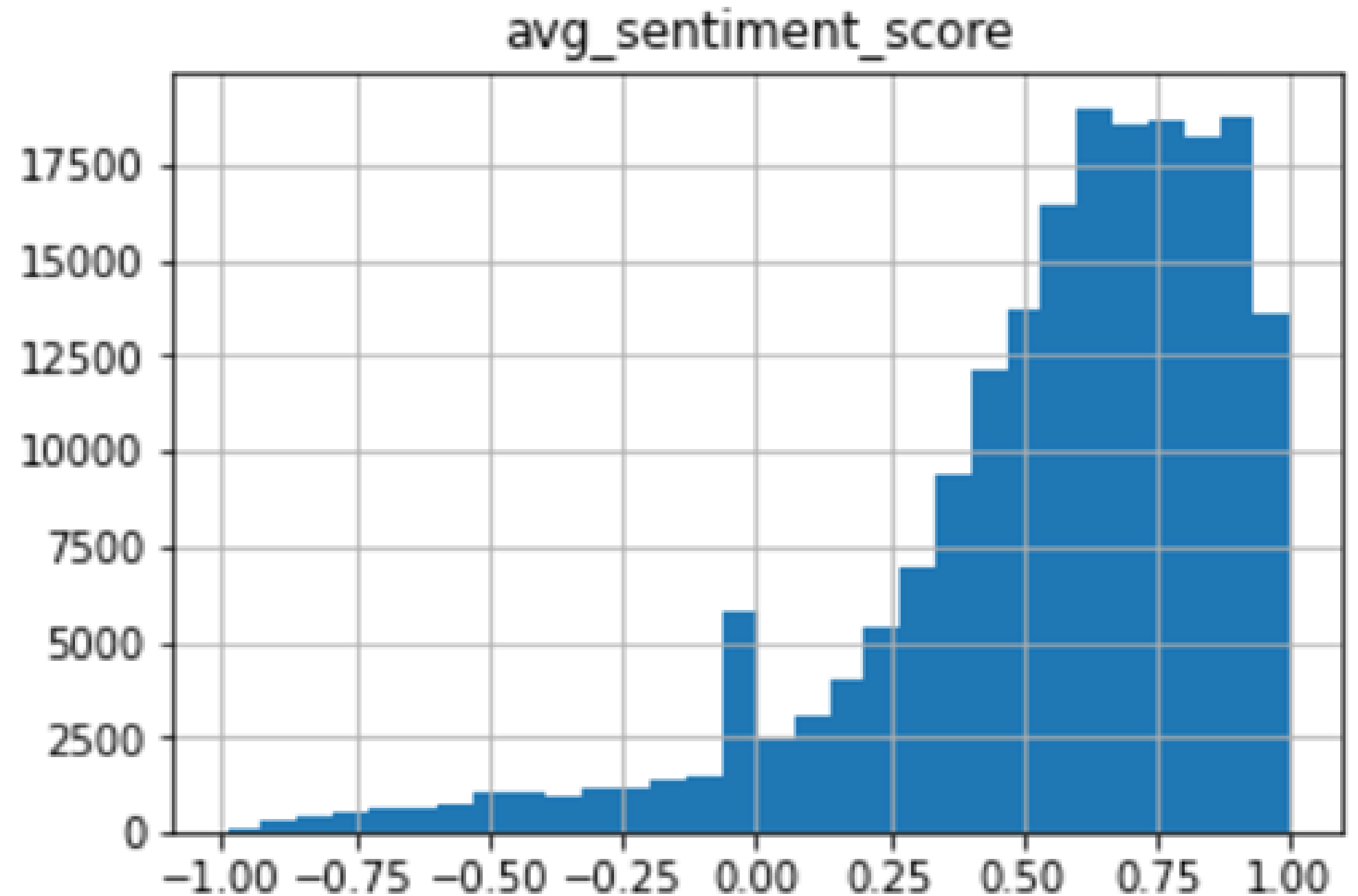


Fig: Histogram of sentiment scores



# CORRELATION ANALYSIS

## Key Observations:

### Strong Positive Correlations:

- safety\_term\_count and review\_count (**0.95**):  
More reviews uncover more safety-related terms.
- avg\_rating and avg\_sentiment\_score (**0.66**):  
Positive sentiment aligns with higher ratings.

### Weak Correlations:

- avg\_rating and safety\_term\_density (**-0.07**):  
Safety term density has minimal impact on ratings.
- avg\_sentiment\_score and incident\_review\_match\_rate (**-0.00**):  
No significant relationship observed.

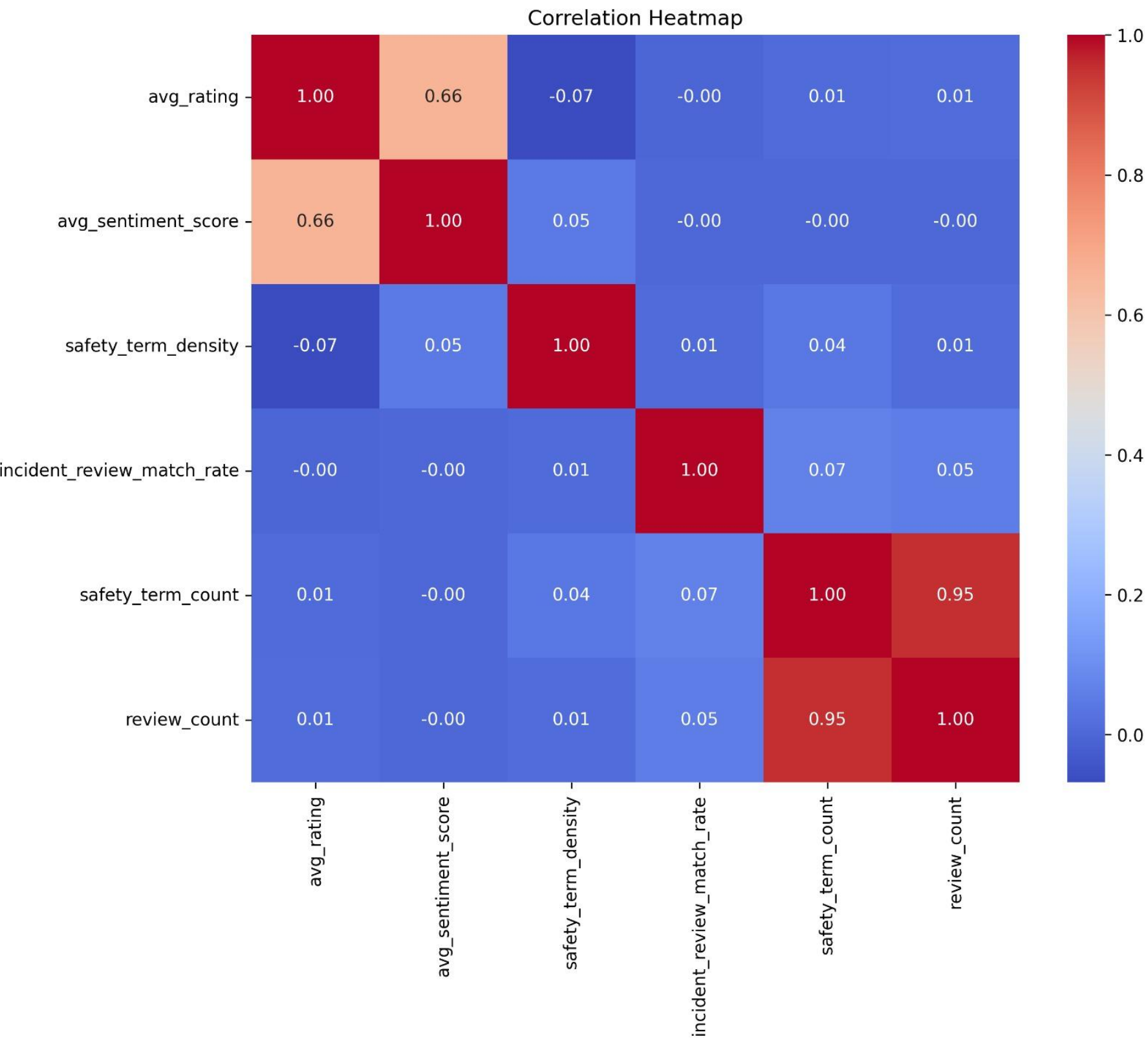


Fig: Heatmap of Correlation Analysis

# TOP 10 REVIEWED PRODUCTS

- **Visualization:** Table of top 10 products by review count and rating
- **Columns:** Product title, review count, average rating, involvement in incidents, safety terms.
- **Key Insight:** "High-review products generally have high ratings, but some are linked to incidents."

Top 10 Reviewed Unique Products:			
	title	review_count	\
parent_asin			
b00echytbi	Infant Optics DXR-8 Video Baby Monitor, Non-Wi...	27282	
b075qq8vzw	iHealth No-Touch Forehead Thermometer, Digital...	14693	
b0bg6jynqx	Regalo Easy Step 38.5-Inch Wide Walk Thru Baby...	13907	
b0bb84jxs9	VTech Sit-to-Stand Learning Walker (Frustratio...	9197	
b07xm9dx9h	Diapers Newborn/Size 1 (8-14 lb), 84 Count - P...	7959	
b0bngy8dwb	Shynerk Baby Car Mirror, Safety Car Seat Mirro...	7814	
b0c5jmrgrd	Regalo Easy Step Extra Tall Walk Thru Baby Gat...	7788	
b00oo9k5qm	Regalo My Cot Portable Travel Bed, Includes Fi...	7255	
b09b8lct53	Summer Multi-Use Decorative Extra Tall Safety ...	6981	
b00oqczavw	Baby Banana Yellow Banana Infant Toothbrush, E...	6944	
	avg_rating	involved_in_incident	safety_terms
parent_asin			
b00echytbi	4.198629	False	[]
b075qq8vzw	3.618186	False	[]
b0bg6jynqx	4.025814	False	[]
b0bb84jxs9	4.639121	False	[]
b07xm9dx9h	4.099510	False	[]
b0bngy8dwb	4.845278	False	[]
b0c5jmrgrd	4.026965	False	[]
b00oo9k5qm	4.310682	False	[]
b09b8lct53	4.317720	False	[]
b00oqczavw	4.644297	False	[]

	text_safety_terms
parent_asin	
b00echytbi	[overheating, trap, fracture, injury, unstable...
b075qq8vzw	[overheating, trap, injury, fracture, unstable...
b0bg6jynqx	[trap, fracture, unstable, injury, damaged, br...
b0bb84jxs9	[trap, fracture, unstable, injury, damaged, mi...
b07xm9dx9h	[trap, fracture, unstable, injury, damaged, mi...
b0bngy8dwb	[trap, fracture, unstable, injury, damaged, mi...
b0c5jmrgrd	[trap, injury, unstable, fracture, damaged, mi...
b00oo9k5qm	[trap, injury, fracture, unstable, damaged, mi...
b09b8lct53	[trap, fracture, injury, unstable, damaged, br...
b00oqczavw	[trap, fracture, injury, damaged, toxic, missi...



# IDENTIFYING RISKY PRODUCTS

- **Visualization:**
  - Table of the lowest-rated products with incidents and safety terms.
- **Columns:**
  - Product title, review count, average rating, safety-related terms.
- **Key Insights:**
  - "Products with low ratings and high safety terms indicate areas of concern."

Lowest Reviewed Products with Incidents and Safety Terms:			
	parent_asin	product_avg_rating	\
45689	b004vl2vro	3.475957	
24712	b004vl2vro	3.475957	
35181	b01ad3l1jc	4.121472	
31422	b000wjlkqm	4.434572	
46697	b003am8cm8	3.500000	
12955	b004vl2vro	3.475957	
47951	b00e8kjync	3.649015	
26779	b002a9iz0q	4.264286	
19776	b001i463g2	4.736842	
35021	b01ad3l1jc	4.121472	
	title	sentiment_score	\
45689	Motorola MBP36 Remote Wireless Video Baby Moni...	-0.49830	
24712	Motorola MBP36 Remote Wireless Video Baby Moni...	-0.49715	
35181	Evenflo Exersaucer Triple Fun Active Learning ...	-0.49515	
31422	Graco Doorway Bumper Jumper, Little Jungle	-0.49475	
46697	OXO Tot Sprout High Chair, Green/Walnut	-0.49425	
12955	Motorola MBP36 Remote Wireless Video Baby Moni...	-0.49330	
47951	Baby Brezza Formula Pro	-0.49280	
26779	Storkcraft Tuscany 5-in-1 Convertible Crib (Es...	-0.49260	
19776	Joovy Room2 Portable Playard	-0.49220	
35021	Evenflo Exersaucer Triple Fun Active Learning ...	-0.49195	
	incident_description_safety_terms		
45689	[faulty, damaged, wound, injury, trauma, damag...		
24712	[faulty, damaged, wound, injury, trauma, damag...		
35181	[fire]		
31422	[faulty, death, break, detached, defective, fa...		
46697	[stuck, lead, flaw, caught, risk, defect, hazard]		
12955	[faulty, damaged, wound, injury, trauma, damag...		
47951	[faulty, missing, severe, blood, break, fuse, ...		
26779	[severe, rash, scratch, serious]		



# SAFETY WORD ANALYSIS

## Key Takeaways:

- Frequent terms reveal common safety concerns, such as:
  - **Structural issues** (e.g., loose parts, cracks).
  - **Critical incidents** (e.g., severe injuries, hot surfaces).
- Insights guide **quality control** and **risk mitigation** efforts.

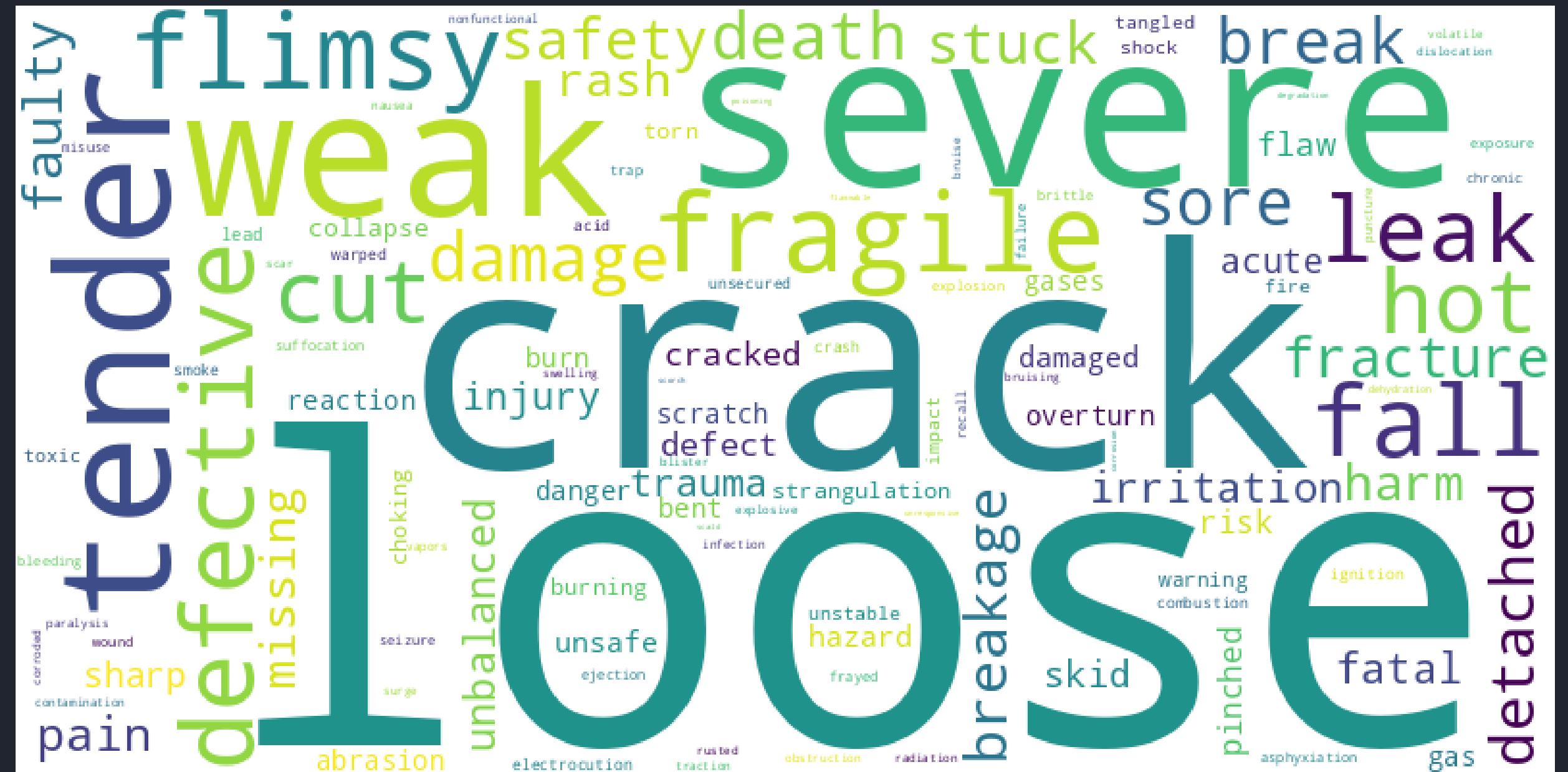


Fig: Word Cloud for Safety Terms

# ML REGRESSION MODELS USED

## LOGISTIC REGRESSION

[19]: `%run -i /scratch/jking47/scripts_part2/logistic_regression.py`

Classification Report with Adjusted Threshold:

	precision	recall	f1-score	support
False	1.00	0.98	0.99	39527
True	0.00	0.14	0.00	7
accuracy			0.98	39534
macro avg	0.50	0.56	0.50	39534
weighted avg	1.00	0.98	0.99	39534

Confusion Matrix:

```
[[38594  933]
 [    6    1]]
```

Summary of potentially unsafe predictions:

predicted\_unsafe

0 192884

1 4782

Name: count, dtype: int64

Fig: Performance Metrics of Logistic Regression

## XG BOOST

[25]: `%run -i /scratch/jking47/scripts_part2/xgboost.py`

XGBoost Classification Report with Adjusted Threshold:

	precision	recall	f1-score	support
False	1.00	1.00	1.00	39527
True	0.00	0.00	0.00	7
accuracy			1.00	39534
macro avg	0.50	0.50	0.50	39534
weighted avg	1.00	1.00	1.00	39534

Confusion Matrix:

```
[[39525    2]
 [    7    0]]
```

Summary of potentially unsafe predictions:

predicted\_unsafe

0 197640

1 26

Name: count, dtype: int64

Fig: Performance Metrics of XG Boost

# ML REGRESSION MODELS USED

## DECISION TREE

```
... Classification Report:
      precision    recall  f1-score   support

   False      1.00      0.91      0.95    39527
   True       0.00      0.14      0.00         7

 accuracy      0.91    39534
 macro avg     0.50    39534
weighted avg     1.00    39534

Confusion Matrix:
[[35842  3685]
 [    6    1]]

Summary of potentially unsafe predictions:
predicted_unsafe
False    179337
True     18329
Name: count, dtype: int64
```

Fig: Performance Metrics of Decision Tree

## K-NEAREST NEIGHBOR

```
Classification Report:
      precision    recall  f1-score   support

   False      1.00      1.00      1.00    39527
   True       0.00      0.00      0.00         7

 accuracy      1.00    39534
 macro avg     0.50    39534
weighted avg     1.00    39534

Confusion Matrix:
[[39527    0]
 [    7    0]]

Summary of potentially unsafe predictions:
predicted_unsafe
False    197666
Name: count, dtype: int64
```

Fig: Performance Metrics of K-Nearest Neighbor

## NAÏVE BAYES CLASSIFIER

```
product_df = pd.read_csv("checkpoint6.csv")
Classification Report:
      precision    recall  f1-score   support

   False      1.00      1.00      1.00    39527
   True       0.00      0.00      0.00         7

 accuracy      1.00    39534
 macro avg     0.50    39534
weighted avg     1.00    39534

Confusion Matrix:
[[39452    75]
 [    7     0]]

Summary of potentially unsafe predictions:
predicted_unsafe
False    197351
True       315
Name: count, dtype: int64
```

Fig: Performance Metrics of Naïve Bayes Classifier



# HEURISTIC MODEL

- We have performed Feature Engineering with some aggregations to create features.
- Since, the merged dataset is not well labelled, we use this features to label the product is potentially safe.
- Threshold values use to classify the product as potentially unsafe
  - Sentiment\_threshold < 25 Percentile
  - Rating\_threshold <25 percentile
  - Safety\_term\_density > 75 percentile
  - Incident\_review\_match\_rate >75 percentile
  - Involved\_in\_incident = 1

```
Reloading dataset and beginning aggregation
Performing heuristic search

Thresholds used:
Rating threshold (<=): 3.625
Sentiment score threshold (<=): 0.6361
Safety term density threshold (>=): 3.0
Incident-review match rate threshold (>=): 0.0

Sample of Potentially Unsafe Products saved as heuristic_unsafe_products.csv
```

parent_asin	avg_rating	review_count	avg_sentiment_score	title	item_model_number	recall_summary	incident_safety_terms	text_safety_terms	involved_in_incident	involved_in_recall	safety_term_count	safety_term_density	incident_review_match_rate	potentially_unsafe
b0009exoeg	1	2	-0.0218	Safety 1st Essentials Childproofing	hs145		['trauma', 'damag	['impact', 'loose', 'stuck', 'torn', 'f	TRUE	FALSE	7	3.5	0	TRUE
b000gchmwq	3.333333333	3	0.2308333333	Lansinoh Breastmilk Storage Bag, 2	thomaswi		['fall', 'bent', 'fall',	['leakage', 'leak', 'cut', 'fragile', 'f	TRUE	FALSE	17	5.666666666	0	TRUE
b00192jizq	2.525773195	97	0.5114159793	Safety 1st 2 Pack Grip n' Go Cabinet	48386		['impact', 'loose',	['safety', 'collapsing', 'torn', 'coll	TRUE	FALSE	432	4.45360824	0.8	TRUE
b001cwqt46	1	2	0.1243499999	Safety 1st 2 Pack Grip n' Go Cabinet	48386		['impact', 'loose',	['lead', 'traction', 'loose', 'stuck',	TRUE	FALSE	8	4	0.4	TRUE
b002wcu6wq	1	1	-0.4233	Gerber 2nd Foods Yogurt Blends - E	4764		['pinched', 'lead']	['stuck', 'caught', 'break']	TRUE	FALSE	3	3	0	TRUE
b0032am8k0	3.333333333	3	0.2262166666	Earth Mama A Little Something for E	10341		['death', 'bruising']	['death', 'fire', 'crack', 'serious', 'l	TRUE	FALSE	10	3.333333333	0.5454545	TRUE
b005cb8ice	1.333333333	3	0.1766666666	Safety 1st Essentials Childproofing	hs145		['trauma', 'damag	['fracture', 'breakage', 'severe', 'c	TRUE	FALSE	19	6.333333333	0	TRUE
b0073jdj7c	1	1	-0.0115	Safety 1st 2 Pack Grip n' Go Cabinet	48386		['impact', 'loose',	['severe', 'loose', 'stuck', 'caught'	TRUE	FALSE	4	4	0.2	TRUE
b007hdfllm	2.8	5	0.23843	Fisher Price 3-in-1 Child Booster Hi	y9463		['fracture', 'collap	['fracture', 'impact', 'severe', 'ble	TRUE	FALSE	20	4	0.5	TRUE
b007ryu4jy	2.625	8	0.52493125	Gerber Graduates Waffle Wheels - F	00015000049614		['fall', 'fall', 'fall',	['stuck', 'caught', 'fire', 'burn', 'm	TRUE	FALSE	25	3.125	0	TRUE
b007y47m7o	1	1	-0.48865	Gerber Graduates Yogurt Melts, Mix	none		['pinched', 'lead']	['death', 'fall', 'serious']	TRUE	FALSE	3	3	0	TRUE
b008s1y31k	1	1	-0.37345	Lansinoh Breastmilk Storage Bag, 2	thomaswi		['fall', 'bent']	['death', 'damage', 'stuck', 'skid',	TRUE	FALSE	11	11	0.5	TRUE
b00cyrse5a	3.25	4	0.4906500000	Gerber Chicken & Vegetables Stars	ns724		['acute', 'sharp', 'e	['stuck', 'caught', 'serious', 'loose	TRUE	FALSE	12	3	0	TRUE
b00krb0t3u	2	4	0.3364375	Gerber Graduates Pasta Pick-Ups S	015000009083		['acute', 'sharp', 'e	['rusted', 'corroded', 'corrode', 's	TRUE	FALSE	24	6	0	TRUE
b00mfbsn60	1	1	-0.2202	Gerber Snacks for Baby Fruit & Veg	none		['pinched', 'lead']	['stuck', 'caught', 'break']	TRUE	FALSE	3	3	0	TRUE
b0777n3vwb	3.157894736	19	0.546478947	Graco Dream Suite Bassinet, Lulla	1957236		['substance', 'suff	['severe', 'recall', 'loose', 'fire', 'se	TRUE	FALSE	63	3.31578947	0	TRUE
b0786m488z	3.25	8	0.55721875	The Tour+ (Zoe XL1) - Best Everyday	xl1bestv2		['fall', 'collision',	['crack', 'missing', 'break', 'choki	TRUE	FALSE	52	6.5	0.5	TRUE
b07zsh7jlc	2	4	0.5432	SUNTRADE Adjustable Visor Cartoo	none		['mold', 'mold', 'm	['severe', 'stuck', 'caught', 'loose'	TRUE	FALSE	12	3	0	TRUE
b083bqq9k4	2	1	0.4439999999	Gerber Natural for Baby, 1st Foods,	none		['acute', 'sharp']	['rusted', 'corroded', 'corrode']	TRUE	FALSE	3	3	0	TRUE
b08v9dxdfc	3	1	0.4604	Grosimimi Vacuum Insulated Sippy	none		['lead']	['loose', 'fall', 'pain']	TRUE	FALSE	3	3	0	TRUE



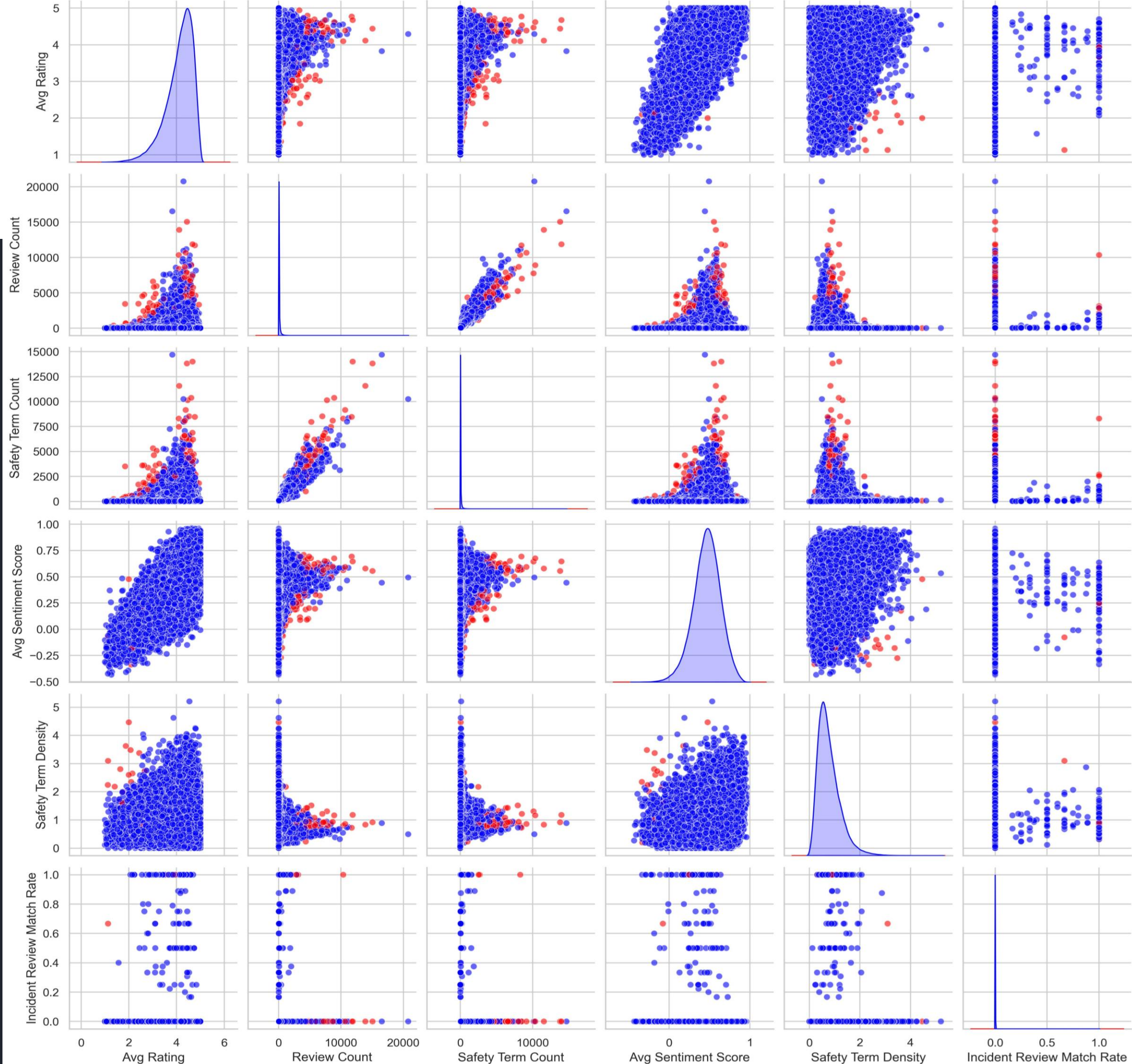
# ISOLATION FOREST

- Used to detect outliers within a given dimensionality by utilizing random partitioning

- **Filtered:** Products with >15 reviews
- **Number of Trees:** 100
- **Threshold:** .0005

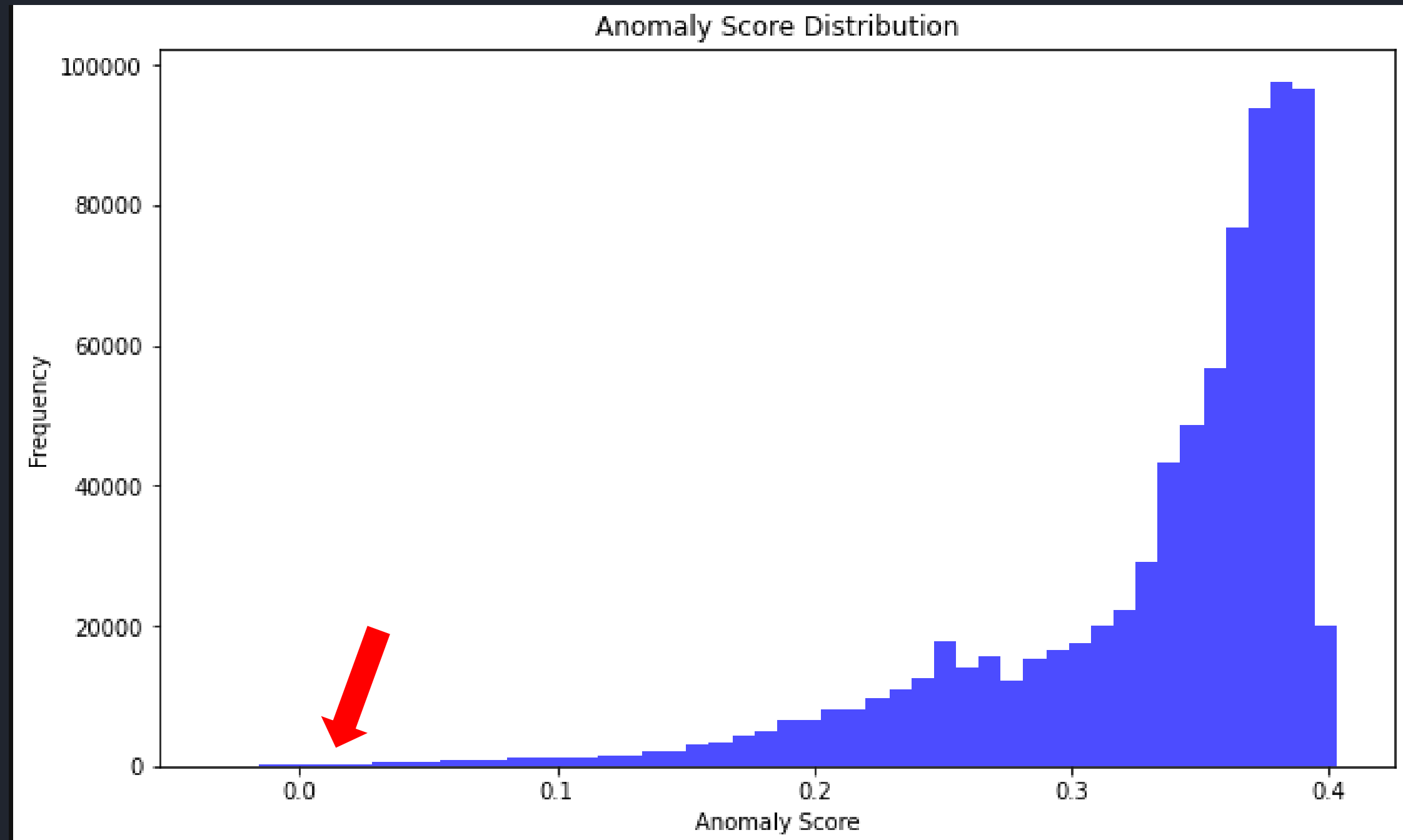
- **Features chosen (6 dimensions):**

- Average Rating
- Review Count
- Safety Term Count
- Safety Term Density
- Average Sentiment Score
- Incident Review Match Rate



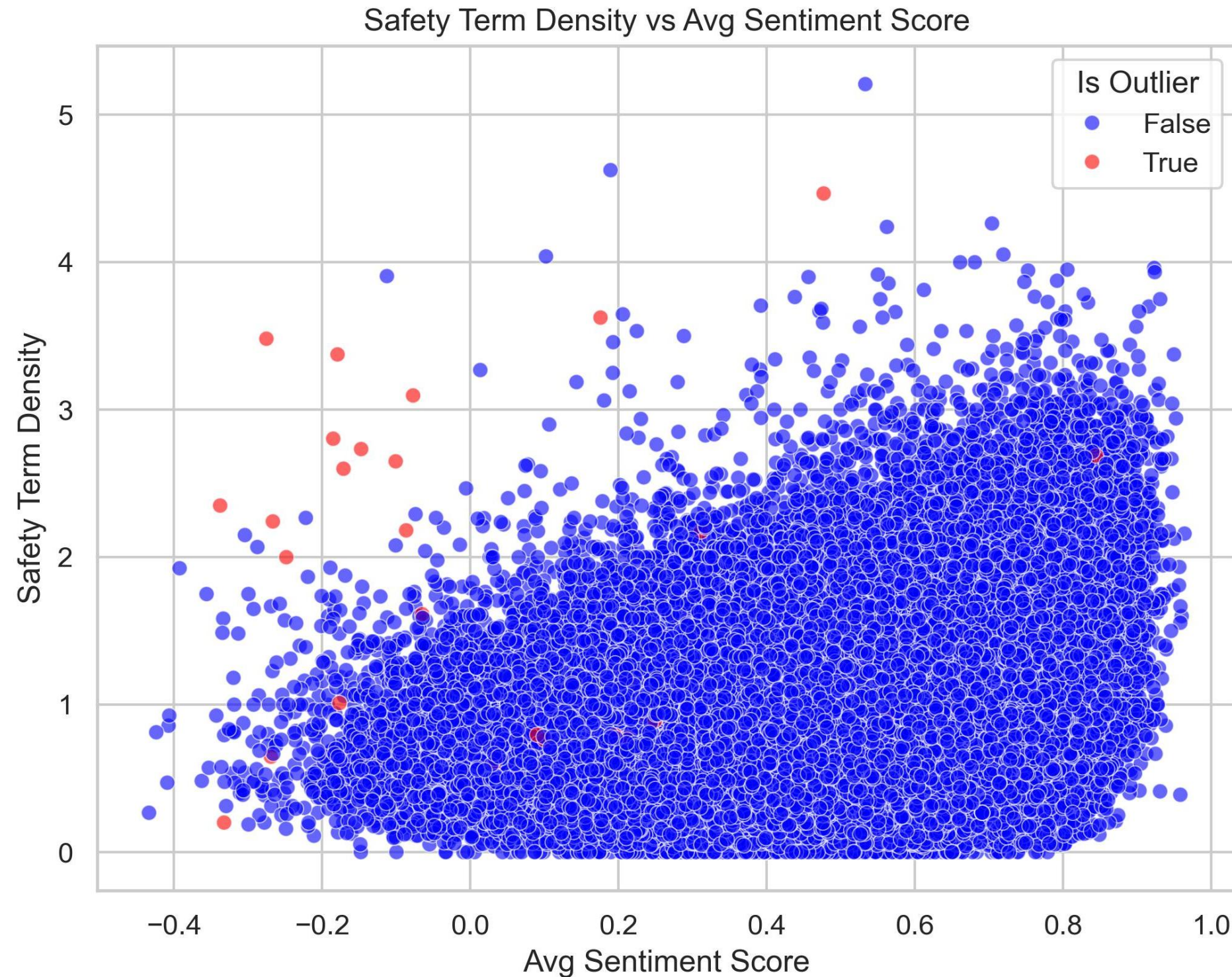
# ISOLATION FOREST

- Used to detect outliers within a given dimensionality by utilizing random partitioning
  - **Filtered:** Products with >15 reviews
  - **Number of Trees:** 100
  - **Threshold:** .0005





# SUMMARY



**With some reservations**, hypothesis is deemed valid to detect potentially unsafe products using Amazon review data.

- Combined approach of sentiment analysis, "dirty word" search, and anomaly detection (or heuristic search)
- Methodology limitations:
  - Limited to flagging for human review
    - False positive rate ~30%
  - Unlabeled data
    - (Processed labelled data deemed unviable/imbalanced)
  - Non-time series
  - Relies on aggregation of a manually extracted merged incident dataset



# FURTHER AREAS OF STUDY

## Supervised Learning

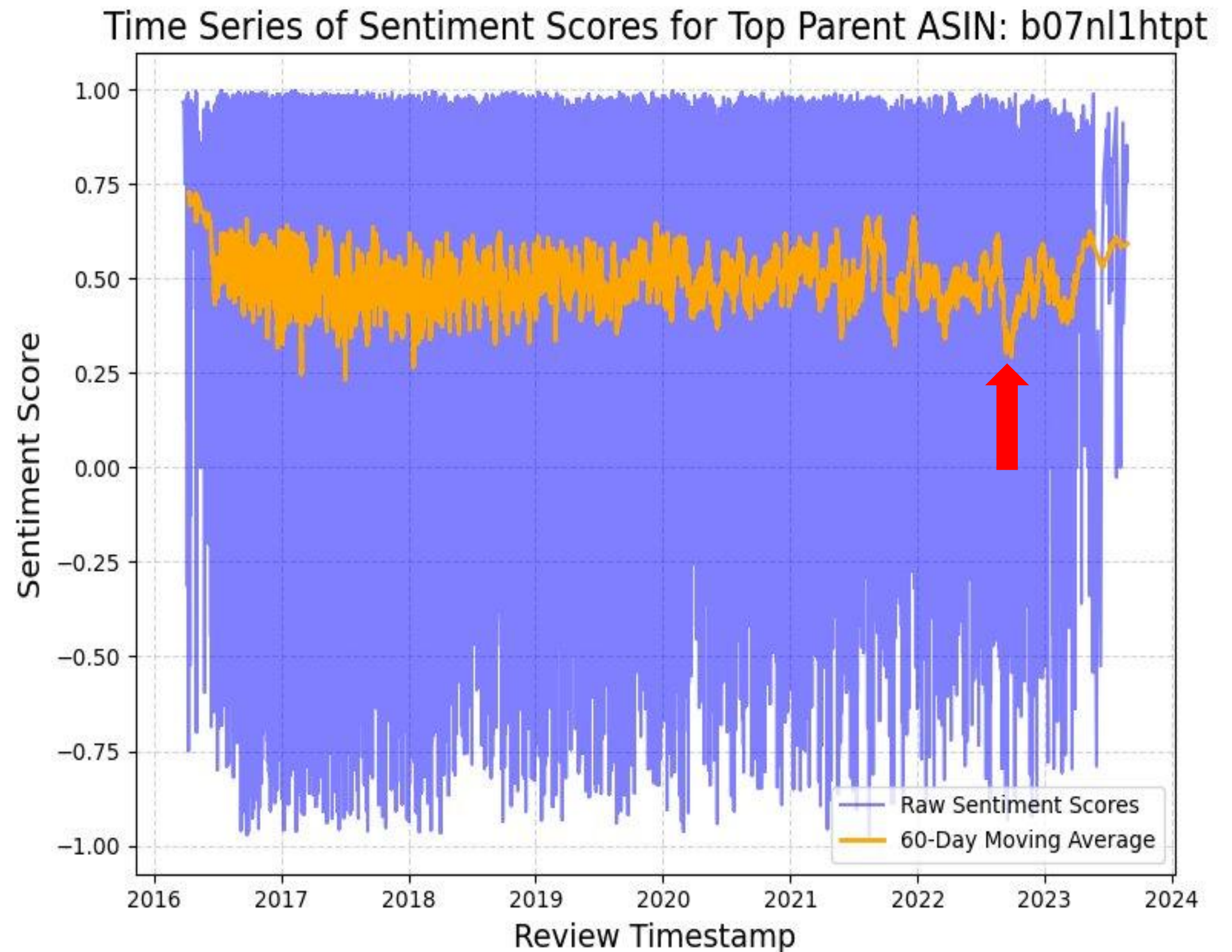
- Requires a labeled dataset with appropriate proportionality (1%)

## Time Series

- Moving average indicator of problems with product
  - Outside of standard deviations?

## Word Matching: Semantic Similarity Scoring

- Better than standard "dirty word" searching
  - Scores word usage intent inside of the parts of speech.
- Highly processing intensive



# TWO PROCESSING SOLUTIONS

Welcome to the Safety Dance Amazon Review Processor! (Data by McAuley Lab)

Available categories:

1. All Beauty
2. Toys and Games
3. Cell Phones and Accessories
4. Industrial and Scientific
5. Gift Cards
6. Musical Instruments
7. Electronics
8. Handmade Products
9. Arts Crafts and Sewing
10. Baby Products
11. Health and Household
12. Office Products
13. Digital Music
14. Grocery and Gourmet Food
15. Sports and Outdoors
16. Home and Kitchen
17. Subscription Boxes
18. Tools and Home Improvement
19. Pet Supplies
20. Video Games
21. Kindle Store
22. Clothing Shoes and Jewelry
23. Patio Lawn and Garden
24. Books
25. Automotive
26. CDs and Vinyl
27. Beauty and Personal Care
28. Amazon Fashion
29. Magazine Subscriptions
30. Software
31. Health and Personal Care
32. Appliances
33. Movies and TV

Select a category by number (e.g., 1 for All Beauty):

CUDA not detected. Using CPU for processing.

Loading synonyms: 100%|██████████| 193/193 [00:00<00:00, 55762.26it/s]

64

Pre-processing incident dataset

Preprocessing Recall Dataset.

Executing the Big Loop....

Processing category: Toys and Games

Downloading raw review data for 'raw\_review\_Toys\_and\_Games'...

Downloading Review Data: 0%|██████████| 76591/16260406 [00:05<19:45, 13655.88records/s]

Fig: Single "Big Loop" Function against all categories

Fig: Command Line Interface

# Q&A

## STATISTICS:

- **Number of scripts:** 6
- **Lines of code:** 3,252  
(1,806 in data conditioning, 1,449 in main loop)
- **Number of models used/attempted:** 15
- **Datasets processed:** 25 out 34  
(Other categories were intangibles)
- **Dataset size:** 565mil rows, 337gb
- **Total time to process:** 244 hours
- **Open-source methodology:** 100%



THANK YOU