Online Influencer Product Recommendation System

Module 1: Data Processing and Exploratory Data Analysis

Group 23
Chen, Xinyao
Niravane, Anagha
Qiu, Emily
Peng, Yulin
Liu, Yijia

Table of Contents

Project Introduction

Background: With the rise of social media and e-commerce integration, influencers have become a key force in driving online sales through product promotions. Selecting the right products to promote remains a challenge, as influencers need to balance market trends with their personal brand and audience preferences.

Project Goal: Develop a **product recommendation system** tailored for **online influencers** on social media platforms.

- By analyzing both trending products and influencer characteristics, our system helps creators **identify the most suitable products to promote**, maximizing their sales potential and commission earnings.
- This data-driven approach not only enhances influencer revenue streams but also optimizes brand partnerships by ensuring more effective product placements.

Data Description and Acquisition

- The recommendation system is built based on **four datasets**, which provides data support from the aspects of influencers themselves, their created content, product feedback and product popularity on the internet.
- These datasets are the foundation for implementing **influencer-based collaborative filtering and content-based filtering**.

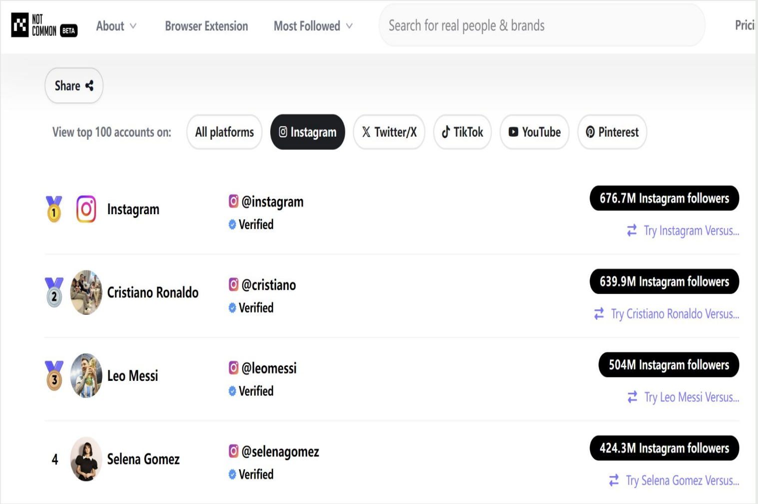
Influencers Profile Dataset	Influencers Post Dataset	Amazon Product Review Dataset	Desktop Search Keyword Dataset
detailed information about influencers (e.g. demographics, niche and follower count)	influencers content from Instagram, including post metadata, captions, engagement metrics, and sponsored tags	product reviews, ratings, helping identify trending and high-quality products in different categories.	From Dewey Dataset, Contain the keywords driving clicks to websites via organic searches

Influencer Dataset

The influencer profile dataset and post dataset were acquired through a combination of web crawling and the Instagram API.

Data Collection Process for instagram influencers profile and post datasets

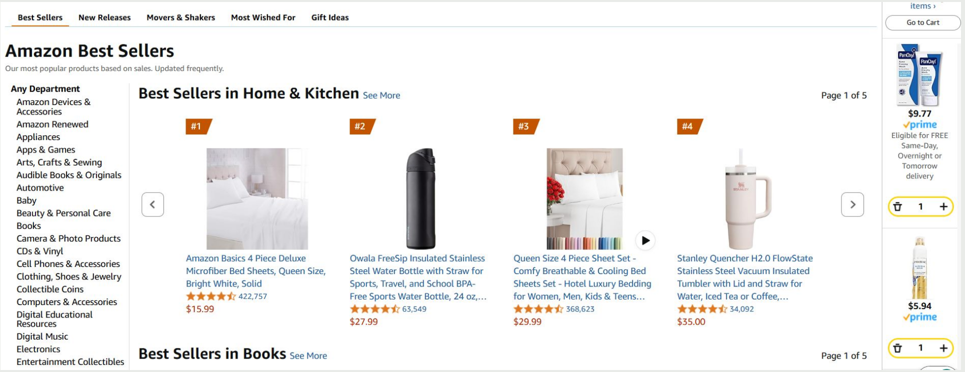
1. Used web crawler to extract the accounts of 8,000 influencers from Not Common, a platform that tracks the most-followed Instagram users.
2. Used Instagrapi API to obtain the detailed profiles of 3,000 influencers.
3. Filtered out business accounts and focused only on non-business influencers.
4. Retrieved up to 300 of the most recent posts for each non-business influencer.



Amazon Dataset

Data Collection Process for item metadata and user reviews

1. A large-scale Amazon Reviews dataset collected in 2023, made publicly available on HuggingFace by McAuley Lab (UCSD).
2. This dataset contains 48.19 million items, and 571.54 million reviews from 54.51 million users.
3. Items span across 33 varied categories - Software, Amazon Fashion, Automotive, Grocery and Gourmet Food, etc.



4. Source also provides subsets for preserving user-item ratings only for those users that have written at least 5 reviews, for improved reliability.

Source: [Amazon Reviews 2023 (huggingface.co)]

Amazon Dataset

1. Item Metadata

1) **Parent ASIN**: Unique identifier for item parent (Products with different colors, styles, sizes usually belong to the same parent ID).
2) **Title**: Name of the product.
3) **Average Rating**: Rating of the product shown on the product page.
4) **Rating Number**: Number of ratings in the product.
5) **Features**: Bullet-point format features of the product.
6) **Description**: Description of the product.
7) **Main Category**: Main category of the product.
8) **Price**: Price in US dollars (at time of crawling).
9) **Store**: Store name of the product.

2. User Reviews

1) **User ID**: Unique identifier of the reviewer.
2) **Parent ASIN**: Parent ID of the product.
3) **Rating**: Rating of the product (from 1.0 to 5.0).
4) **Title**: Title of the user review.
5) **Text**: Text body of the user review.
6) **Timestamp**: Time of the review (unix time).
7) **Helpful Vote:** Helpful votes of the review.
8) **Verified Purchase:** User purchase verification.

Usage in the recommendation system: This dataset provides the pool of items that influencers can choose from to promote useful and trending products that cater to their audiences. User reviews, ratings and product features can be used to perform content-based filtering and make more relevant recommendations.

Desktop Search Keyword Dataset

This dataset from Dewey provides insights into the keywords driving organic traffic to websites. With over **80 billion records**, it captures **search behavior and keyword trends**, making it a valuable resource for understanding product demand in online marketplaces.

- **Domain:** The website associated with the search.
- **Country:** Geographical region of the search.
- **Keyword:** The search term that led to a website visit.
- **Search Engine:** The platform used for the query.
- **SERP Type:** whether the click was organic or paid.
- **Search Volume Index:** how often a keyword is searched.
- **Date:** Timestamp of the search record.

- **Identify high-demand products** based on search frequency.
- **Analyze consumer interests and emerging trends** by tracking keyword performance.
- **Align product recommendations with search demand**, ensuring influencers promote products with high potential for sales.

DOMAIN	COUNTRY	KEYWORD	URL	SEARCH_ENGI...	SERP_TYPE	DESKTOP_ORG...	DATE
etsy.com	WW	grease summer...	etsy.com/uk/mar...	Google Search	organic	0	2023-08-12
etsy.com	WW	grease tbirds	etsy.com/market...	Google Search	organic	0.02	2023-08-12
etsy.com	WW	greaser gang	etsy.com/market...	Google Search	organic	0.02	2023-08-12
etsy.com	WW	greaser look	etsy.com/sg-...	Google Search	organic	0.1	2023-08-12

Source: Product Data Preview - Dewey (deweydata.io)

Profile Data Preprocessing - Data Type Conversion and Texts

Step 1: Drops business accounts and only keeps personal accounts for future analysis.

```
# Drop Business account as our system is for personal influencer
print(f"Original dataset size: {influencer_profile.shape[0]}")
influencer_profile = influencer_profile[influencer_profile["is_business"] == False].reset_index(drop=True)
print(f"Filtered dataset size (without business accounts): {influencer_profile.shape[0]}")
```

```
Original dataset size: 2778
Filtered dataset size (without business accounts): 1302
```

Step 2: Replace missing values with “other”, “unknown”, etc.

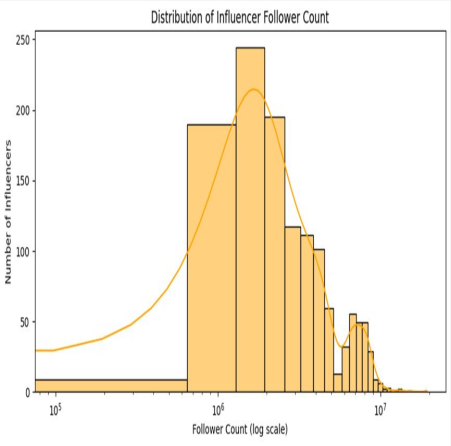
Step 3: Remove duplicated rows with the same user_id

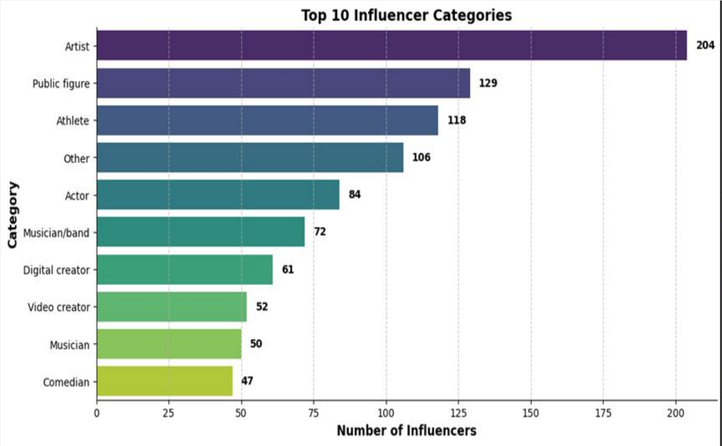
Step 4: Convert data types to string or categorical data for columns like “user_id”,”category”.

Step 5: Standardize text data. Convert to lowercase and remove special characters.

Profile Data EDA

- **Distribution of influencer followers:** Followers counts mostly range from 100k to 1M.
- **Number of influencers by category:** Artists, public figures, athletes and actor/actresses are the most populated categories.
- **Number of followers by category:** Influencers in hospitality, nonprofits and record labels have on average the most number of followers.





	Category	Average Follower Count
0	Hotel resort	8,873,933
1	Nonprofit organization	8,165,845
2	Record label	8,154,411
3	Business service	7,523,581
4	Movie/television studio	6,304,513
5	Sports	5,660,941
6	Song	5,151,953
7	Non-Governmental Organization (NGO)	5,107,679
8	Political Party	4,687,688
9	Community	4,639,152

Posts Data Preprocessing - Sponsor Tags and Text Cleaning

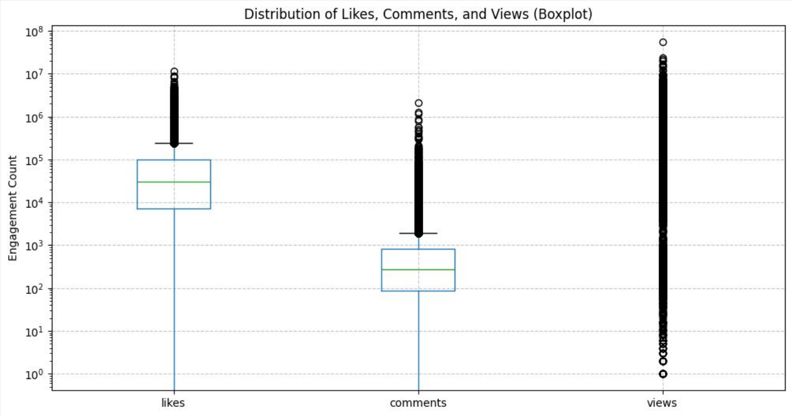
Step 1: Drops rows with missing values (less than 10 rows deleted).

Step 2: Cleans and extract date details.

Step 3: Create “is_sponsored” column to identify influencer marketing behaviors.

Step 4: Outlier Detection & Handling: Remove outliers in likes, views, and comments columns

Step 5: Text cleaning, standardization, tokenization, and non-English-text-translation for future NLP analysis, using the NLTK library.




```python
import re
import nltk
import time
from tqdm import tqdm
from langdetect import detect
from nltk.corpus import stopwords
from textblob import TextBlob
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
from deep_translator import GoogleTranslator

nltk.download("stopwords")
nltk.download("punkt")
```

```python
def translate_to_english(captions):
    translated = []
    for text in tqdm(captions):
        try:
            translated.append(GoogleTranslator(source="auto", target="en").translate(text))
            time.sleep(0.5)
        except:
            translated.append(text)
    return translated


df_posts_encoded["translated_caption"] = df_posts_encoded["clean_caption"].apply(translate_to_english)
```

✓ 764m 8.3s

Posts Data EDA - Feature Engineering

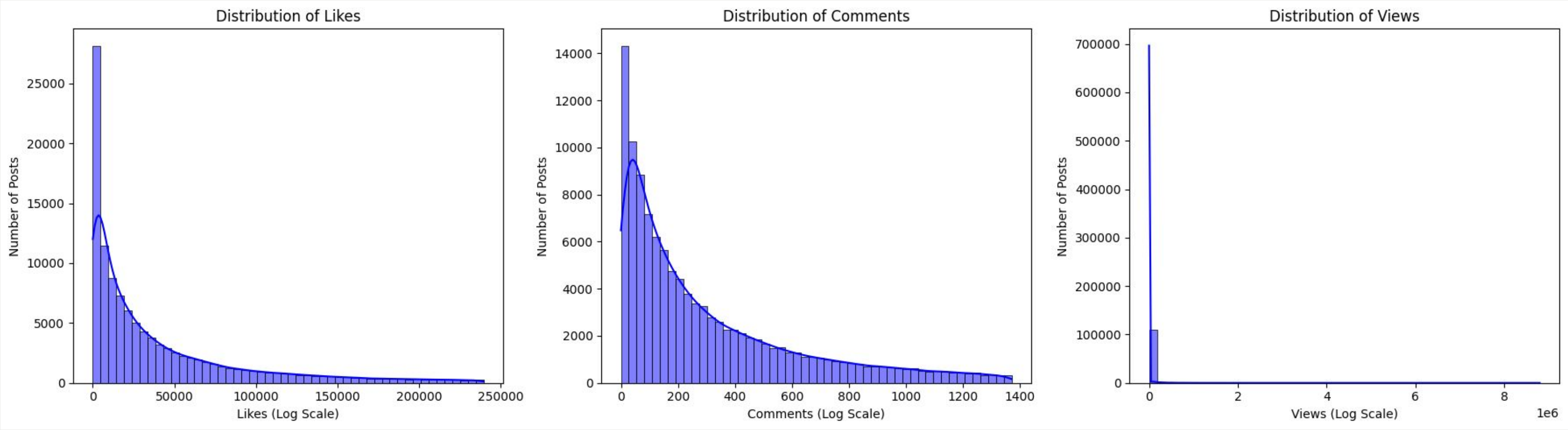
- Feature engineering:
  - **Engagement Rate** = (Total likes + Total comments) / follower_counts
  - **Average number of likes, comments and views per post.**

```
# Feature Engineering: Extract total likes, views, comments and average likes, views, comments for each user_id
df_engagement = df_posts[['user_id', 'likes', 'comments', 'views']].copy()

# Calculate total and average engagement metrics per user_id
df_engagement_aggregated = df_engagement.groupby('user_id').agg(
    total_likes = ('likes', 'sum'),
    total_comments = ('comments', 'sum'),
    total_views = ('views', 'sum'),
    avg_likes_per_post = ('likes', 'mean'),
    avg_comments_per_post = ('comments', 'mean'),
    avg_views_per_post = ('views', 'mean')
).reset_index()

df_engagement_aggregated.head(5)
```

	user_id	total_likes	total_comments	total_views	avg_likes_per_post	avg_comments_per_post	avg_views_per_post
0	1003414073	3632	137	0	3632.000000	137.0	0.0
1	1005579026	78694	894	0	39347.000000	447.0	0.0
2	1007076159	9043	87	0	9043.000000	87.0	0.0
3	10081325212	20999	240	0	6999.666667	80.0	0.0
4	10129804493	60673	512	0	60673.000000	512.0	0.0

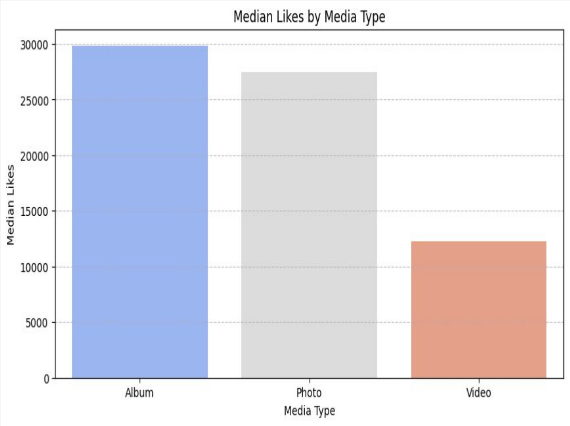
Posts Data EDA - Feature Engineering

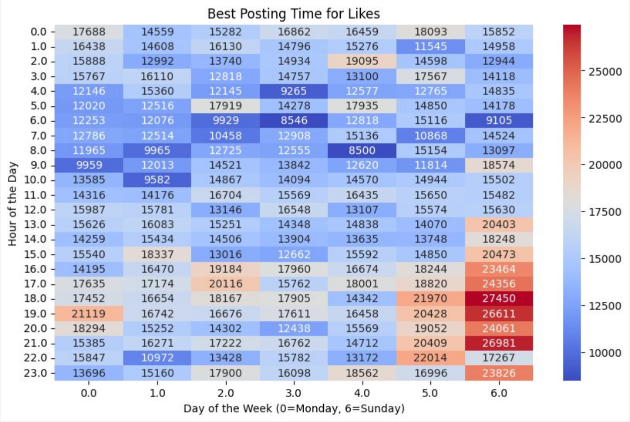
- **Feature engineering:**
  - **Engagement Rate** = (Total likes + Total comments) / follower_counts
  - **Average number of likes, comments and views per post.**

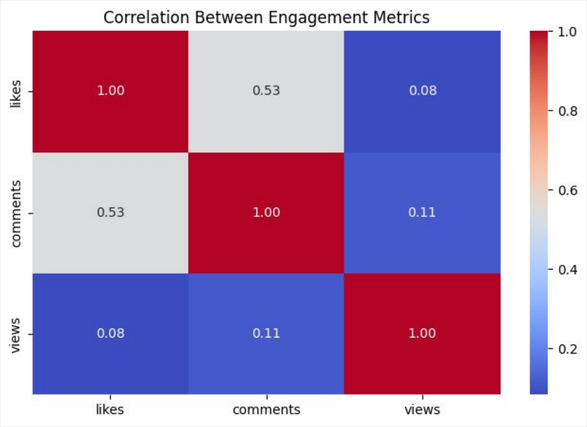


Posts Data EDA - Likes and Engagement

- **Most popular media type:** Albums and photos are still the most popular media format on Instagram, judging from median likes per post.
- **Best posting time for likes: When we post stuff on weekend afternoons and evenings, we get the most likes.**
- **Correlation between engagement metrics: Likes and comments are correlated.** Higher number of views don't necessarily mean higher likes and comments.







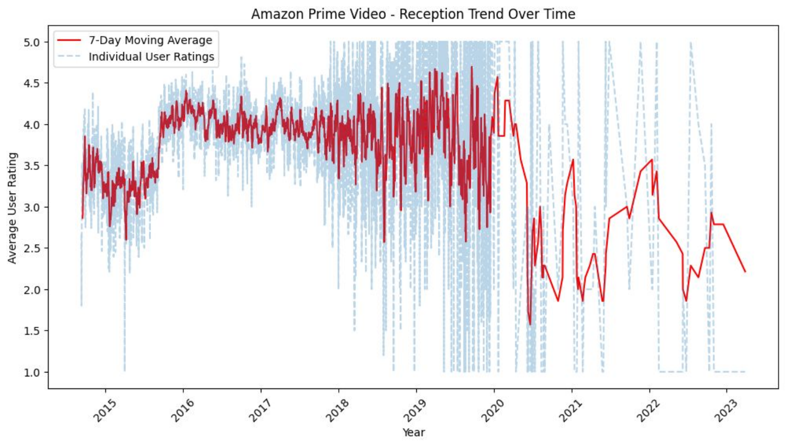
Amazon Reviews 2023: Data Processing

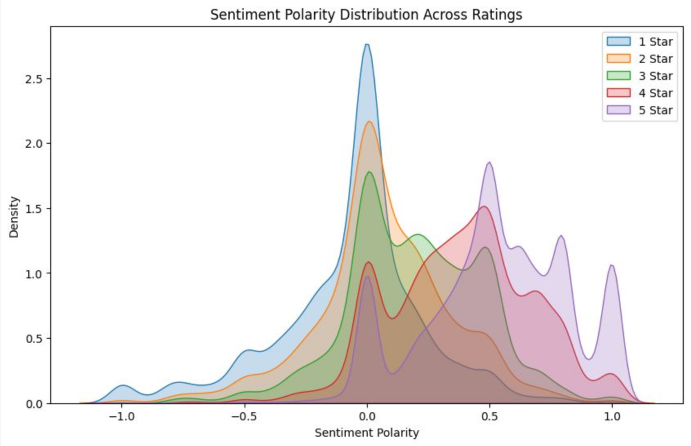
- **Text Processing for Reviews & Features:** Basic text cleaning operations for removal of punctuation, digits & stopwords, tokenization and stemming, followed by calculation review sentiment polarity (TextBlob), and vectorization (explore different embeddings).
- **Feature Engineering**: New attributes that account for temporal trends in ratings (e.g. moving averages), user sentiment, volume of reviews, seasonal variations, etc.
- **Datatype correction, handling null values:** Price, Timestamp, Ratings, # Ratings, and other categorical variables. Additionally, remove reviews containing less than 5 words.

	user_id	parent_asin	rating	title	text	review	review_processed	sentiment	bert_embedding
0	AGCI7FAH4GL5FI65HYLKWTMFZ2CQ	B0BQSK9QCF	1.0	malware	mcaffee IS malware	malware mcaffee IS malware	[malwar, mcaffe, malwar]	0.000000	[-0.035163686, 0.0103913015, 7.115581e-05, 0.0...
1	AHSPLDNW5OOUK2PLH7GXLACFBZNQ	B00CTQ6SIG	5.0	Lots of Fun	I love playing tapped out because it is fun to...	Lots of Fun I love playing tapped out because ...	[lot, fun, love, play, tap, fun, watch, town, ...	0.400000	[-0.04770106, -0.015891232, 0.03607068, -0.039...
2	AHSPLDNW5OOUK2PLH7GXLACFBZNQ	B0066WJLU6	5.0	Light Up The Dark	I love this flashlight app! It really illumin...	Light Up The Dark I love this flashlight app! ...	[light, dark, love, flashlight, app, realli, i...	0.280469	[-0.090288065, -0.018142018, -0.025567722, 0.0...
3	AH6CATODIVPVUOJEWHRSRCSKAOHA	B00KCYMAWK	4.0	Fun game	One of my favorite games	Fun game One of my favorite games	[fun, game, one, favorit, game]	0.133333	[-0.023053482, 0.058434553, -0.01570116, -0.09...
4	AEINY4XOINMMJCK5GZ3M6MMHBN6A	B00P1RK566	4.0	I am not that good at it but my kids are	Cute game. I am not that good at it but my kid...	I am not that good at it but my kids are Cute ...	[good, kid, cute, game, good, kid, love, nik, ...	0.425000	[-0.10075144, 0.014918627, -0.028490193, -0.08...
5	AEINY4XOINMMJCK5GZ3M6MMHBN6A	B00CWY76CC	4.0	good game	Made me think , variety of the puzzles kept it...	good game Made me think , variety of the puzzl...	[good, game, made, think, varieti, puzzl, kept...	0.220000	[-0.06449226, -0.02541187, -0.07404629, -0.087...

Amazon Data EDA - Software items

- **Trends in user ratings over time:** Accounting for trends in user ratings over time can help assess product reception and consequential impact on future sales.
- **Factual relevance of user reviews:** Sentiments expressed by users align with their ratings, suggesting that the reviews may be factual. However, concentration around 0 (neutral) implies that reviews may be short or lack strong emotional wording.

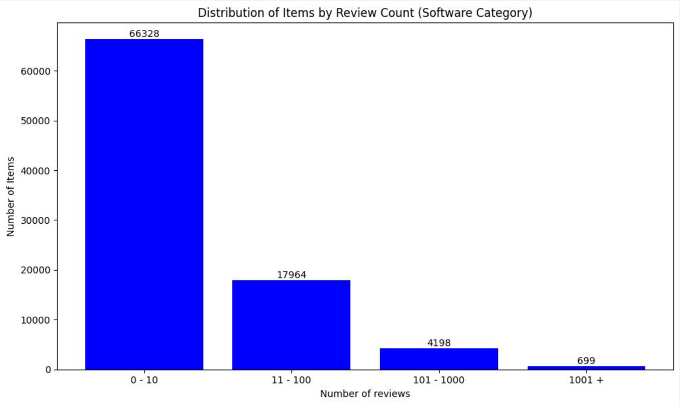




Performed for a subset of 4,880,181 reviews for 89,246 Software items, written by 2,589,466 different users.

Amazon Data EDA - Limitations

- **Pricing information mostly unavailable:** Pricing is a major factor that affects decision-making when a user is purchasing a product. (Unavailable (0.0/None) for ~81.84% of Software products)
- **Limited information about user:** While user reviews are available, there is little information about the user demographics, preferences, usage history, etc., which might be highly relevant in assessing product performance and review relevance.
- **Average Ratings and Number of Ratings unavailable**: Item metadata missing for a small portion of the data.
- **Skewed reviews per item:** Most items have a very low number of written user reviews (0-10), giving lower context about what users feel about the product.



THANK YOU!