Snow Stormers — Data Appendix

Section 1: Raw Kaggle Dataset

Dataset Name: LetterboxdTop250-5000reviews.csv

Unit of Observation:

Each row represents a single film in the Letterboxd Top 250 dataset. Embedded within each row are list-formatted strings containing multiple user ratings and corresponding review texts.

Description

The raw dataset was obtained from Kaggle and includes film-level metadata along with embedded lists of ratings and reviews. Ratings are stored as star-symbol strings (e.g., ★★★½), and reviews are stored as text lists. These lists were parsed and transformed into review-level observations during cleaning.

Variables:

movie_rank: Integer. Rank of the film within the Letterboxd Top 250 list.

NAME: String. Title of the film.

YEAR: Integer. Year of film release.

DIRECTOR: String. Director of the film.

SYNOPSYS: String. Film synopsis.

RATINGS: List (stored as string). Embedded list of user star ratings in symbolic format (★ and ½).

REVIEWS: List (stored as string). Embedded list of user review texts.

Section 2: Cleaned Review-Level Dataset

Dataset Name: letterboxd_top250_reviews_clean.csv

Unit of Observation:

Each row represents a single user review of a film, paired with a numeric star rating.

Description

The raw dataset was transformed from film-level rows containing embedded lists into review-level observations. Ratings were converted from symbolic star format into numeric values (0.5–5.0). Reviews with missing ratings or empty text were removed. Non-breaking spaces and formatting artifacts were cleaned.

The final cleaned dataset contains 4,755 review-level observations.

Variables:

movie_rank: Integer. Rank of the film in the Letterboxd Top 250 list.

movie_title: String. Title of the film.

year: Integer. Year of release.

director: String. Director of the film.

synopsis: String. Film synopsis.

critic_id: Integer. Within-film index representing review order.

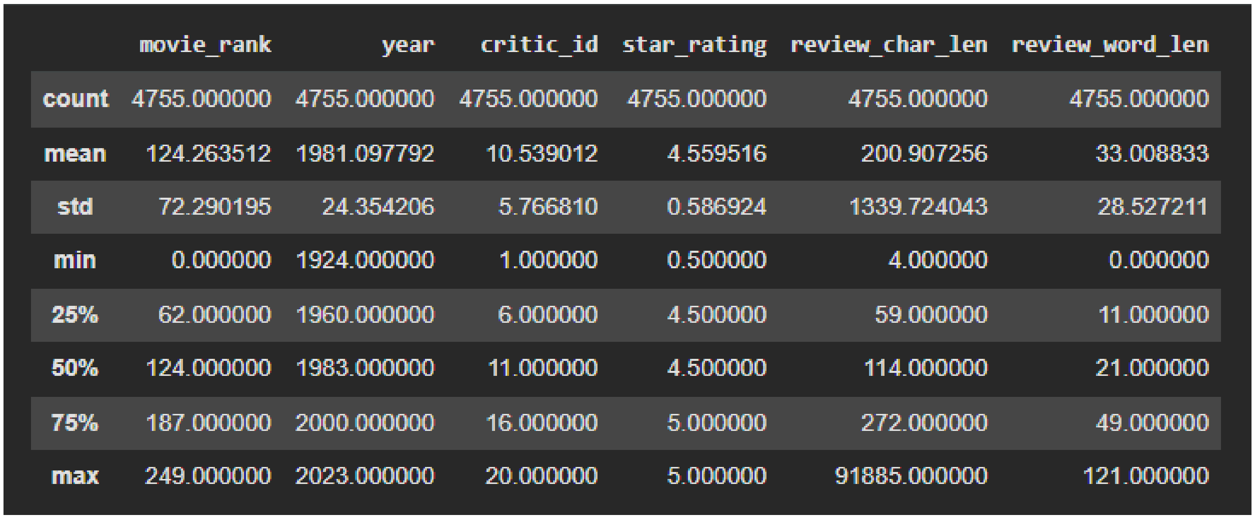
star_rating: Float. Numeric rating from 0.5 to 5.0 derived from symbolic star representation.

review_text: String. Cleaned body of the review.

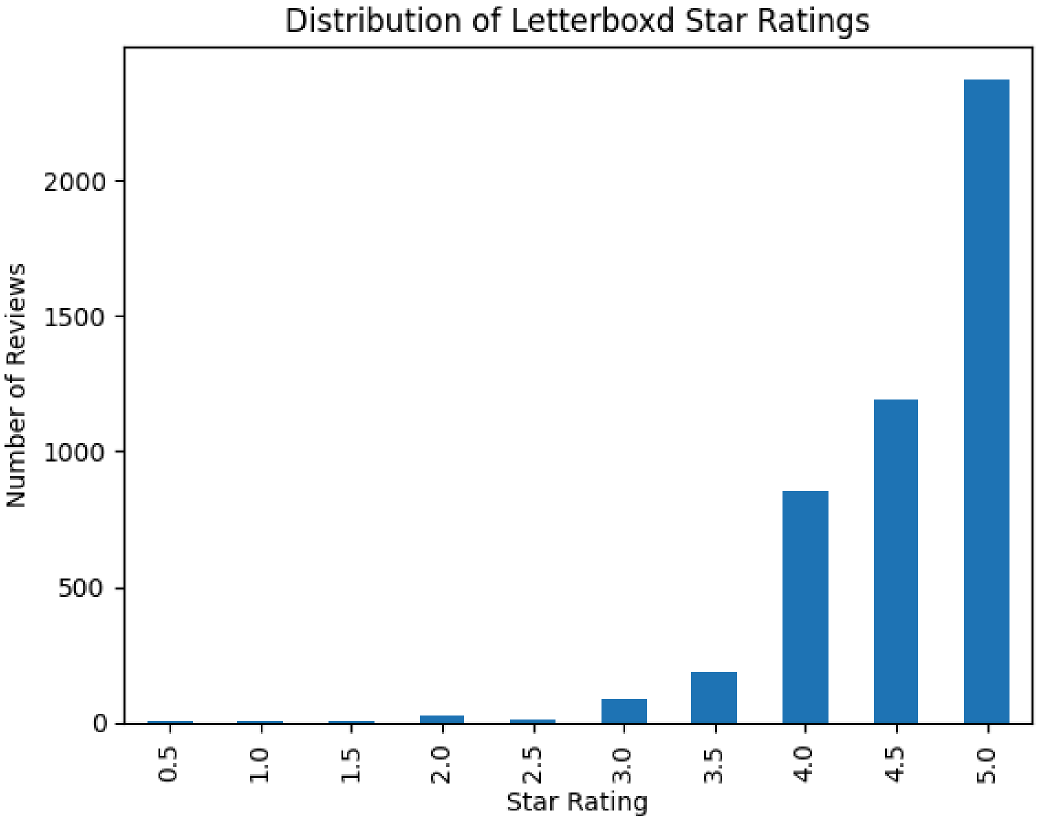
review_char_len: Integer. Number of characters in the cleaned review text.

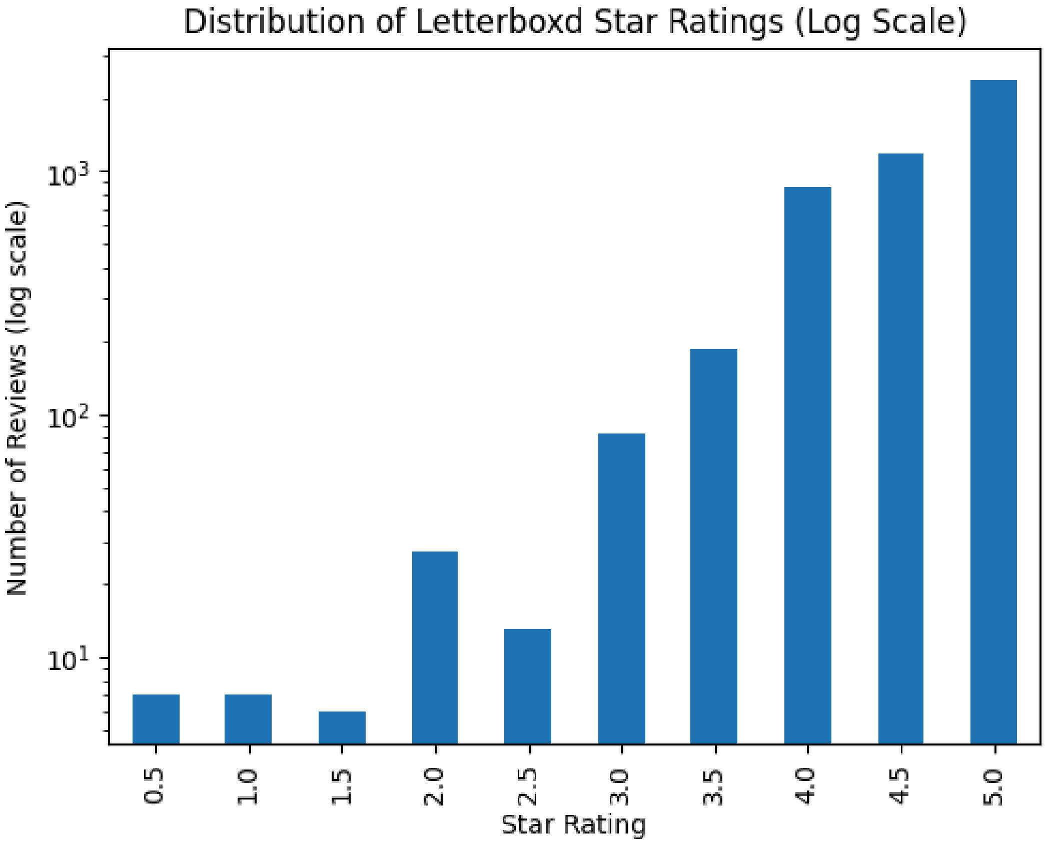
review_word_len: Integer. Number of word tokens in the review text.

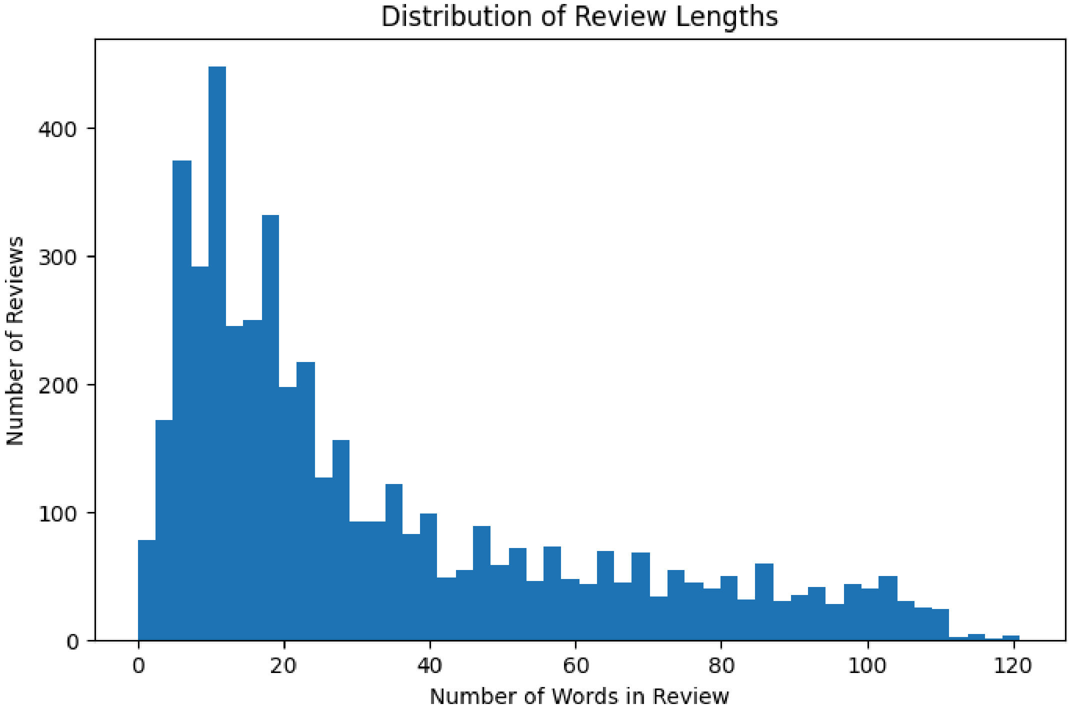
Summary Statistics

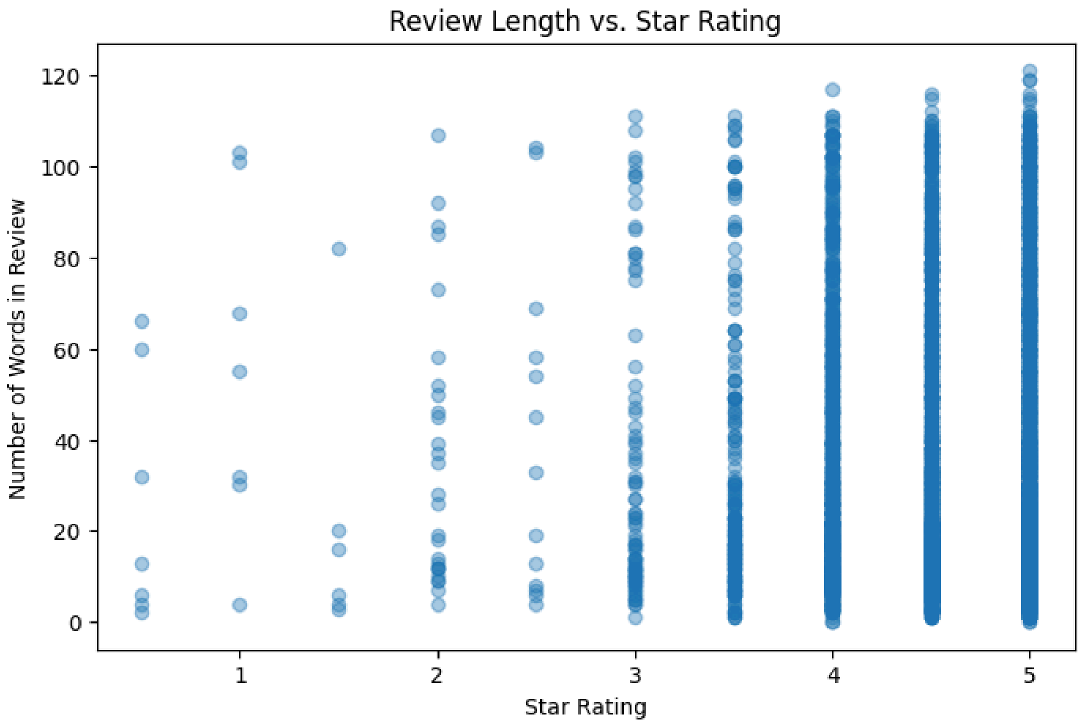


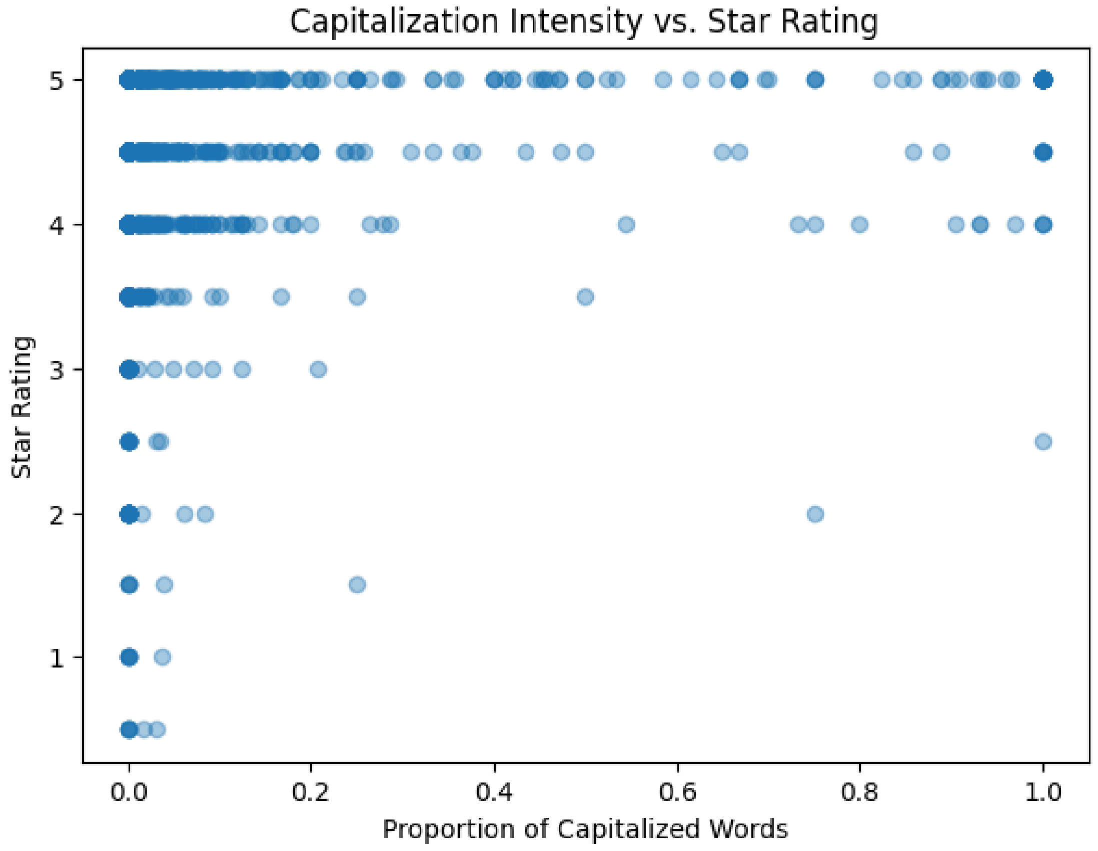
	movie_rank	year	critic_id	star_rating	review_char_len	review_word_len
count	4755.000000	4755.000000	4755.000000	4755.000000	4755.000000	4755.000000
mean	124.263512	1981.097792	10.539012	4.559516	200.907256	33.008833
std	72.290195	24.354206	5.766810	0.586924	1339.724043	28.527211
min	0.000000	1924.000000	1.000000	0.500000	4.000000	0.000000
25%	62.000000	1960.000000	6.000000	4.500000	59.000000	11.000000
50%	124.000000	1983.000000	11.000000	4.500000	114.000000	21.000000
75%	187.000000	2000.000000	16.000000	5.000000	272.000000	49.000000
max	249.000000	2023.000000	20.000000	5.000000	91885.000000	121.000000




Distribution of Letterboxd Star Ratings (Log Scale)
Number of Reviews (log scale)
10^3
10^2
10^1
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Star Rating


Distribution of Review Lengths
Number of Reviews
400
300
200
100
0
0
20
40
60
80
100
120
Number of Words in Review


Review Length vs. Star Rating


Capitalization Intensity vs. Star Rating