Data Appendix

Analysis Data File 1

- Unit of Observation: Each row in the dataset represents an individual movie review.
- Total Observations: 1,444,963

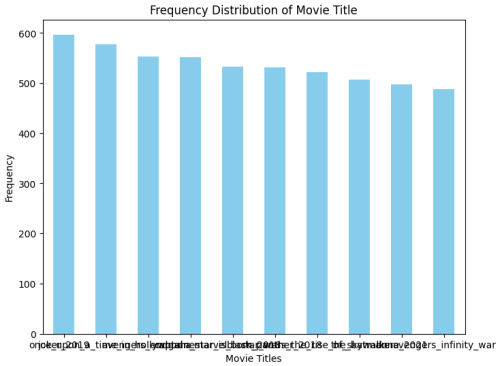
Variables

1. Variable name: movieTitle

- Original variable name: id
- Type: Categorical
- Description: This variable serves as a unique identifier for movie each review
- Observations: 1,444,963(1,444,963)
- Transformations: The variable was renamed from id to movieTitle to enhance readability
- Frequency table for the top 10 movie titles:
    -

movieTitle	Count
joker_2019	597
once_upon_a_time_in_hollywood	573
avengers_endgame	553
captain_marvel	552
a_star_is_born_2018	533
black_panther_2018	531
star_wars_the_rise_of_skywalker	522
the_batman	507
dune_2021	498
avenger_infinity_war	488

    -
- Bar chart of frequency distribution:



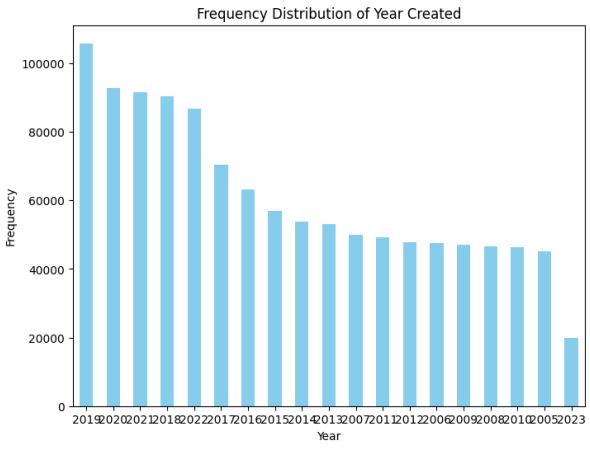
---

2. Variable name: year_created

- Original variable name: creationDate
- Type: Datetime
- Original Type: object
- Description: Year when the review was created
- Original description: Date when the review was created
- Observations: 1,444,963(1,444,963)
- Transformations: The variable was converted from an object to datetime. Then, the year was extracted from the datetime variable, creating a new column named year_created and dropping the original creationDate variable. Removed rows made before the year 2004 as the number of reviews began to rise during the start of the 21st century and older reviews might not reflect the same positive or negative sentiments as they do in more recent reviews.
- Frequency table:

year_created	Count
2019	105881
2020	92879
2021	91644
2018	90530
2022	86776
2017	70335
2016	63300

2015	56988
2014	53747
2013	53094
2007	50082
2011	49369
2012	47721
2006	47552
2009	47142
2008	46664
2010	46430
2005	45271
2023	19955

- Bar chart of frequency distribution:
  - 

---

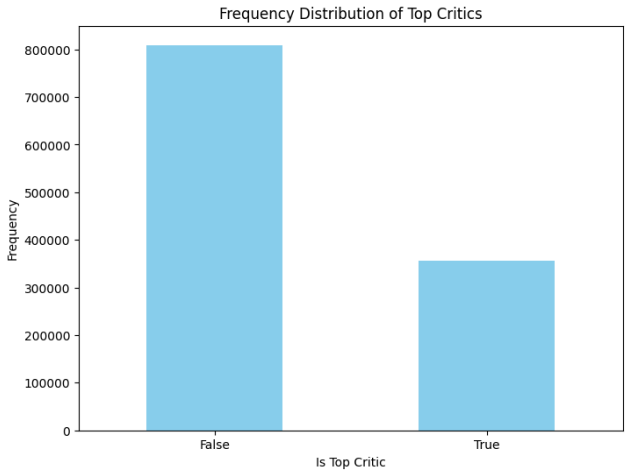
3. Variable Name: isTopCritic

- Type: Boolean
- Description: Boolean flag indicating if the critic is recognized as a top critic.
- Observations: 1,444,963(1,444,963)
- Transformations: There were no transformations performed on this variable.
- Frequency table:

-

isTopCritic	count
False	809173
True	356187

- Bar chart of frequency distribution:



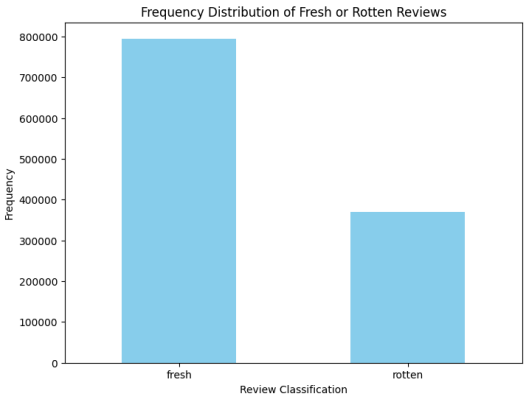
---

4. Variable Name: fresh_or_rotten

- Original variable name: reviewState
- Type: object
- Description: Review classification state (e.g., “fresh” or “rotten”)
- Observations: 1,444,963(1,444,963)
- Transformations: The variable was renamed from reviewState to fresh_or_rotten to enhance readability.
- Frequency table:

Fresh	794920
Rotten	370440

- Bar chart of frequency distribution:



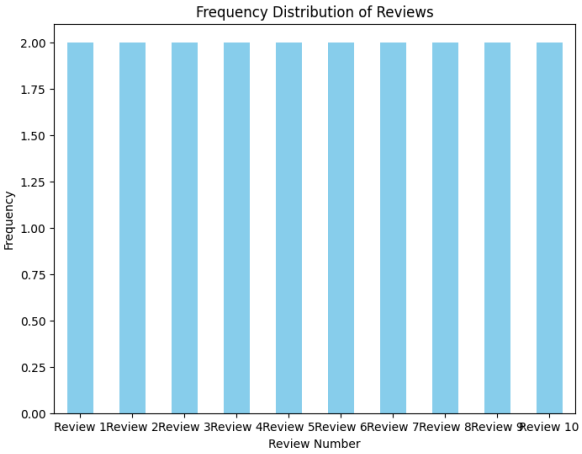
---

5. Variable Name: review

- Original variable name: reviewText
- Type: object
- Description: Full text of the review
- Observations: 1,444,963(1,375,738)
- Transformations: The variable was renamed from reviewText to review to enhance readability. All missing observations and duplicate reviews in this variable were dropped from the dataset.
- Frequency table for top 10 movie reviews:

review	count
A decidedly underwhelming sequel…	2
Nothing makes sense.	2
…watchable but entirely forgettable…	2
It's a mess.	2
Tepid.	2
…breezy, light-hearted…	2
A stunning misfire.	2
A stone-cold masterpiece.	2
Pretty lousy.	2
A beautiful bore.	2

- Bar chart of frequency distribution:



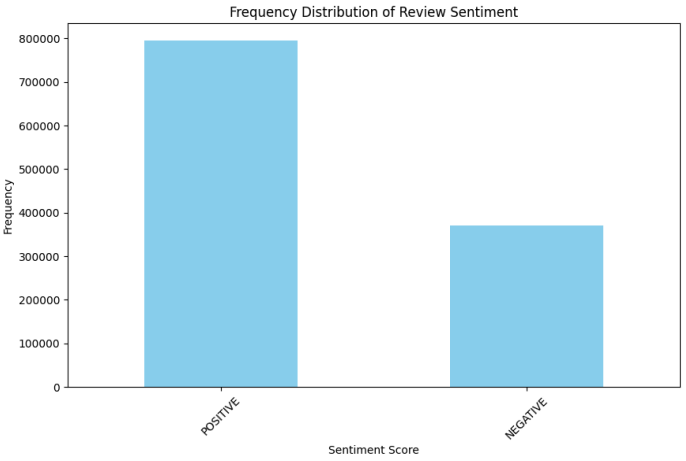
---

6. Variable Name: scoreSentiment

- Type: categorical
- Description: sentiment associated with the review score (e.g., “POSITIVE” or “NEGATIVE”)
- Observations: 1,444,963(1,444,963)
- Transformations: There were no transformations performed on this variable.
- Frequency table:

scoreSentiment	count
POSITIVE	794864
NEGATIVE	370416

- Bar chart of frequency distribution:



---

7. Variable Name: reviewId

- Type: quantitative
- Description: unique identifier for each review (integer)
- Observations: 1,444,963(1,444,963)
- Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

8. Variable Name: criticName

- Type: object
- Description: Name of the critic providing the review
- Observations: 1,444,963(1,444,963)
- Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

9. Variable Name: originalScore

- Type: object
- Description: the original rating score given by the critic (may include null values)
- Observations: 1,444,963(1,009,745)
- Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

10. Variable Name: publicatioName

- Type: object
- Description: Name of the publication where the review was published
- Observations: 1,444,963(1,444,963)
- Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis

---

11. Variable Name: reviewUrl

- Type: object
- Description: URL linking to the full review
- Observations: 1,444,963(1,234,038)
- Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.