University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

CSE 4/587 - Project Phase - 1

Members : Chandrahas Gurram , UBID : cgurram , Ub# : 50538624
Sai Venkata Aditya Arepalli, UBID: sarepall, Ub#: 50536170
Gowtham Bobbili, UBID: gbobbili, Ub#: 50540347

Problem Statement:

Problem : Trend and sentiment analysis of movie metadata.

1) a)

More content than ever before is being produced by the film business, which involves a lot of data, including viewer reviews and movie facts. To make better selections, they must comprehend what the viewers desire. There is a great demand for computer-based techniques to analyze movie trends and consumer sentiment because the amount of data is too large to handle manually.

The quick growth of the film industry has produced a great amount of data, including user-generated reviews and information about movies. Although this spike makes analysis more difficult, it also provides new chances to learn about viewer preferences and industry patterns. Processing such a large dataset to extract relevant attitudes and patterns—which are essential for marketers, distributors, and content creators—is a problem. To modify their tactics in response to audience expectations, they require these insights. The volume and complexity of the data make manual analysis not practical, emphasizing the need for automated, advanced data processing methods.

1) b)
This study uses advanced algorithms for data-intensive computing, including machine learning (ML) and natural language processing (NLP), to perform a thorough review of movie metadata and sentiment learned from user reviews. The project intends to decipher complex trends in genre popularity, theme evolution, and audience reception over time by utilizing machine

learning techniques including clustering for trend analysis and natural language processing for sentiment extraction. Sentiment analysis techniques allow the measurement of subjective viewer opinions, resulting in a more sophisticated understanding of customer behavior and preferences.

Additionally, by presenting innovative methods to the analysis of cultural objects, this initiative promotes the computational social sciences. It provides a standard for multidisciplinary research that may have an impact on upcoming investigations in media analytics, cultural studies, and entertainment economics by bridging the gap between technical data analysis and creative content appraisal. The project aims to enhance understanding of the digital media environments by combining technical precision with creative analysis, providing competitive advantages in the entertainment sector.

Data Sources:

Data set -
https://www.kaggle.com/datasets/bharatnatrayn/movies-dataset-for-feature-extracion-prediction?select=movies.csv

The "movies.csv" dataset is a comprehensive collection of data about different movies that includes important details like the title of the film, the year of release, the genre, the rating, a synopsis, the lead actor or actress, the number of votes or ratings the film received, its runtime, and its gross earnings. With the help of this dataset, one can analyze and comprehend the workings of the film industry and forecast the level of success and popularity of individual films based on their attributes.

Understanding the elements that influence a film's success and popularity is the main goal of this dataset analysis. Our goal is to create predictive models that can gauge a film's level of audience engagement and commercial performance by looking at the correlations between cast, genre, rating, and runtime. Producers, distributors, and filmmakers are just a few of the industry participants who can benefit from this analysis's insightful advice on production, marketing, and distribution tactics.

Number of Rows: 10000
Number of Columns: 9

Below mentioned are the features of the dataset

- MOVIES: Name of the movie.
- YEAR: Release year of the movie.
- GENRE: Genre(s) of the movie.
- RATING: Average rating of the movie.
- ONE-LINE: A short description or summary of the movie.
- STARS: Lead actors or actresses in the movie.
- VOTES: Number of votes/ratings the movie received.
- RunTime: Duration of the movie in minutes.
- Gross: Gross earnings of the movie.

Predicting the popularity of movies based on different features like genre, rating, lead stars, runtime, and gross earnings is the aim of the analysis using the "Movies" dataset. This model can help producers, distributors, and film studios better understand the elements that go into making a successful film.

Data cleaning and preprocessing

The data set contained a total of 10000 entries in rows and 9 columns in total. Out of them the following cleaning was done to take out and filter the meaning less data. The cleaning procedures done are :

1) Removing duplicate values :
The number of duplicates are a detrimental factor to the model and removing them is very important
After removing the duplicates the number of rows came down to
Number of rows: 9568
Number of columns: 9

2) Breaking the words into individual values else called tokenization but manually.
It is the process of tokenizing the genre column to get text from
“ drama scifi thriller “ to “drama” ”scifi” and “thriller”.

3) CONVERTING ALL THE VARIOUS TYPES OF DATA INTO A SINGLE TYPE
It is convert the objects to float to perform computation and get statistics on the

```
MOVIES      object
YEAR        object
GENRE       object
RATING     float64
```

```
ONE-LINE    object
STARS       object
VOTES       object
RunTime     float64
Gross       object
```

To

```
MOVIES      object
YEAR        object
GENRE       object
RATING      float64
ONE-LINE    object
STARS       object
VOTES       float64
RunTime     float64
Gross       object
```

4) Removing null values

```
MOVIES         0
YEAR         542
GENRE         78
RATING      1400
ONE-LINE       0
STARS          0
VOTES       1400
RunTime     2560
Gross       9108
```

Number of rows with missing values for the RATING, VOTES and RunTime columns: 1178

And they have been removed from the dataset.

5) Filling in the data with other measures

Maintaining dataset completeness and guaranteeing robustness in later studies can be achieved by filling in missing data with appropriate metrics. This cleaning method reduces the influence of missing values on statistical analyses and model performance, hence improving the dependability of insights derived from the data.

6)Setting all the numeric characters to the same precision to avoid inconsistency:

Accurate comparisons and analyses are made easier by keeping uniformity and consistency in data representation by setting all numeric characters to the same precision. This cleaning method improves the dependability and interpretability of numerical data by reducing the possibility of disparities brought about by different degrees of precision.

7) Convert all text data into lowercase data.

Lowercase text data conversion standardizes the text format, lowers the possibility of discrepancies, and makes text processing jobs easier to complete later. By guaranteeing consistency in text representation, this cleaning raises the accuracy of text-based algorithms and models and increases the efficiency of text analysis.

8) Identifying and removing outliers based on interquartile range

We find the interquartile values for the numeric columns and then we work out the outliers that disturb the distribution and remove keep to keep the model safe but not removing all of them so that we do keep the model generalized and prone to sudden values.

9) Removing the trailing whitespaces in the text data to avoid and inconsistencies.

Inorder to remove the inconsistencies from the text data we remove any extra white spaces to avoid any problems during the tokenization and analysis of the data and making all the values in each row consistent.

10)removing the punctuation and special characters

In order to simplify the analysis process and normalize the text, we eliminate special characters and punctuation from text data. Tasks involving text processing, like tokenization, sentiment analysis, and natural language processing, become more accurate as a result of this cleansing.

11) Removing numbers from our text data columns

For tasks like text analysis and natural language processing, it is necessary to remove numbers from text data columns since, in a textual context, numerical digits frequently carry little semantic significance. By keeping the emphasis on important textual content, this cleaning helps to improve the accuracy of later text processing activities.

12) Removing the extra white spaces for consistency

Excessive white spaces can distort the appearance of text and result in inconsistent data processing, therefore removing them from text data maintains consistency and improves readability. By standardizing text layout, this cleaning aids to facilitate precise text analysis and manipulation.

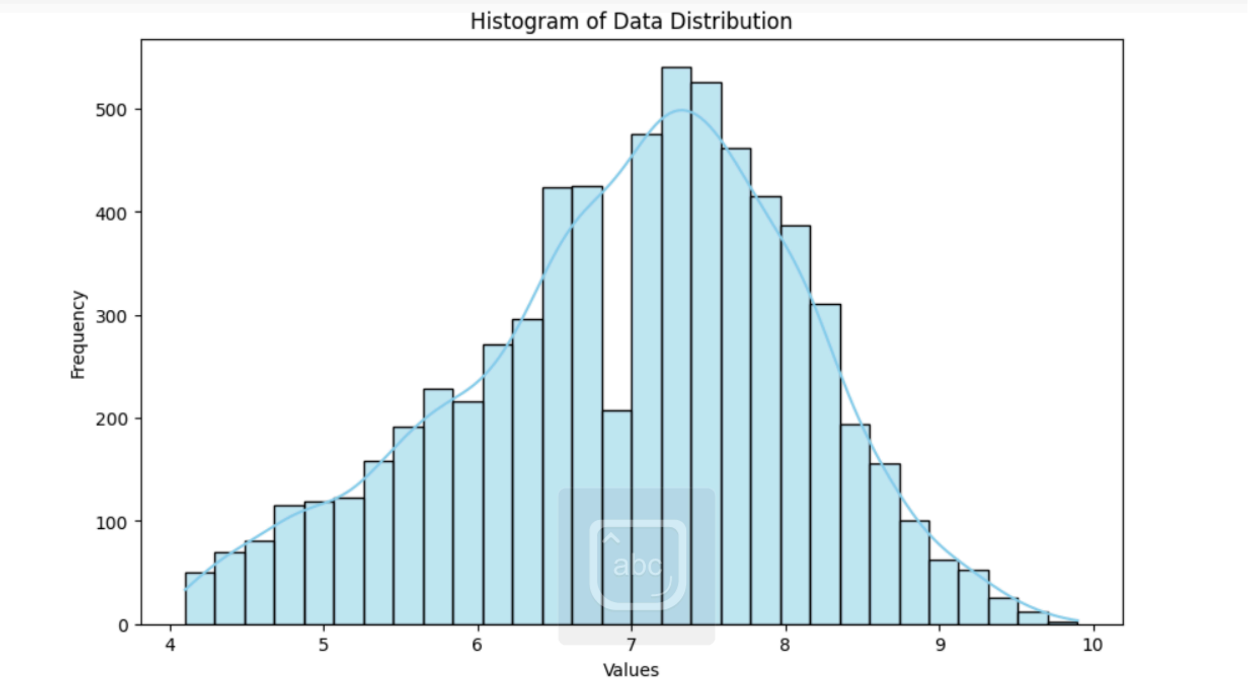
13) removing the emojis and all emoticons from the text data

By guaranteeing that analysis and processing concentrate exclusively on the textual content, removing emoticons and emojis from text data helps prevent data interruption. Text-based tasks like sentiment analysis and natural language processing benefit from this cleaning since it increases accuracy and consistency.

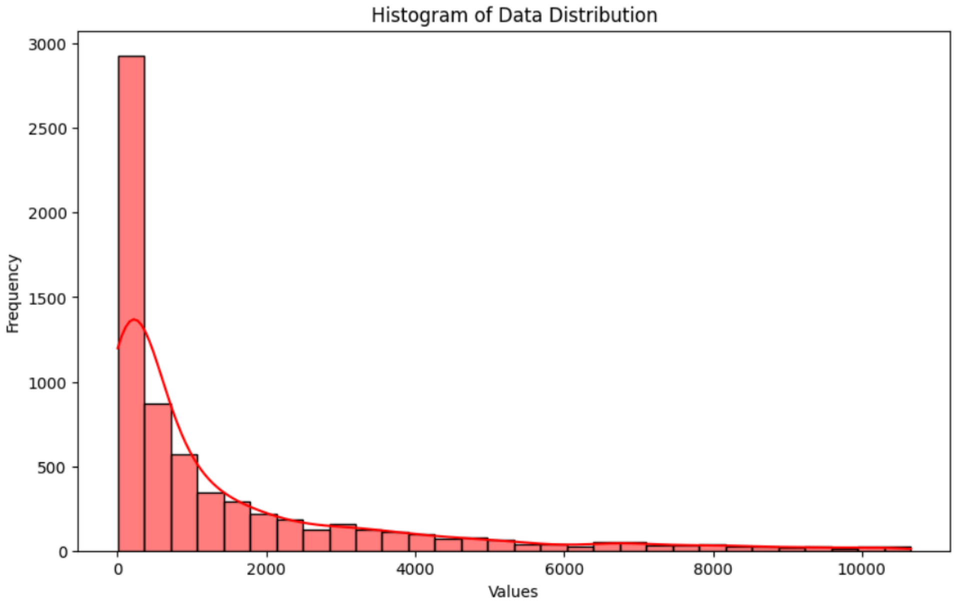
14) expanding the contractions to avoid data inconsistencies

By restoring abbreviated forms to their complete expressions, expanding contractions in text data reduces ambiguity and boosts the precision of text analysis operations performed later on. This cleaning improves the readability of written content and boosts the efficiency of algorithms used in natural language processing.

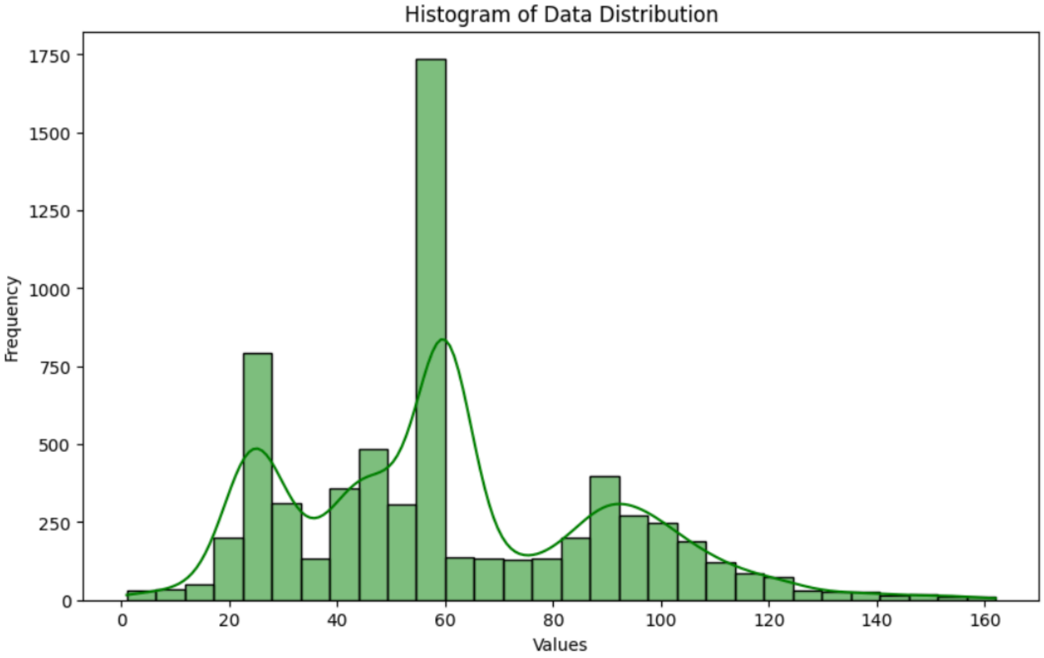
Data exploratory analysis :



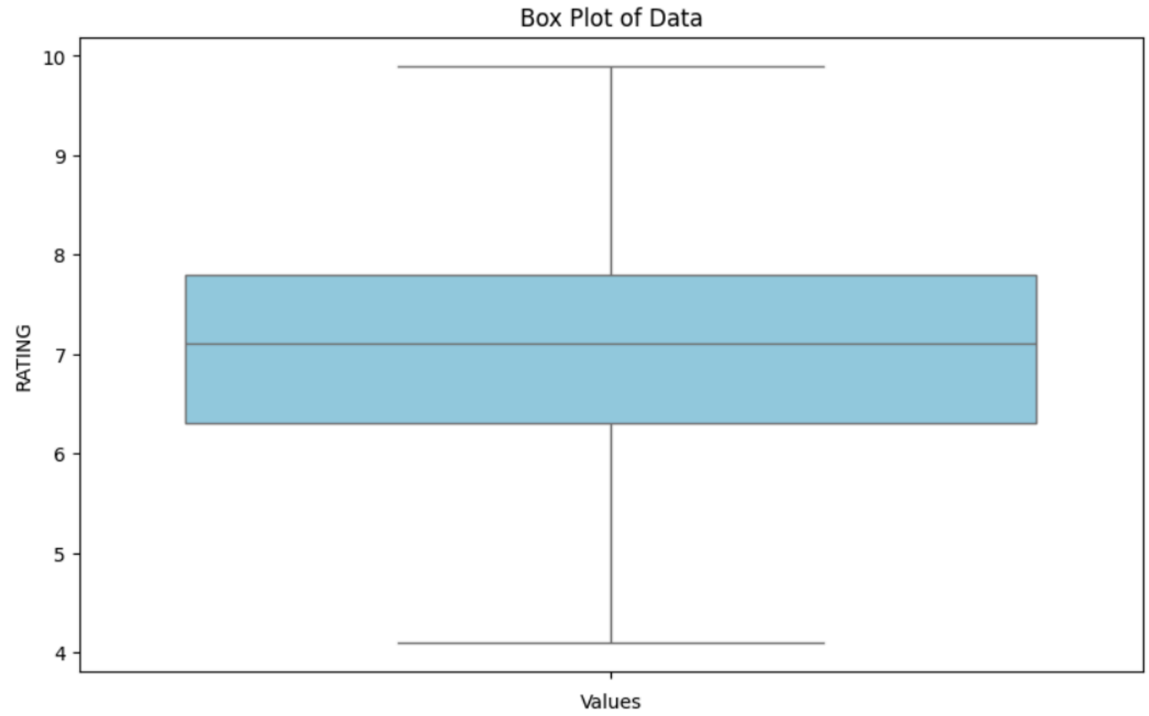
- The above histogram depicts the distribution of movie ratings in the dataset, highlighting the frequency of ratings across a range of values.



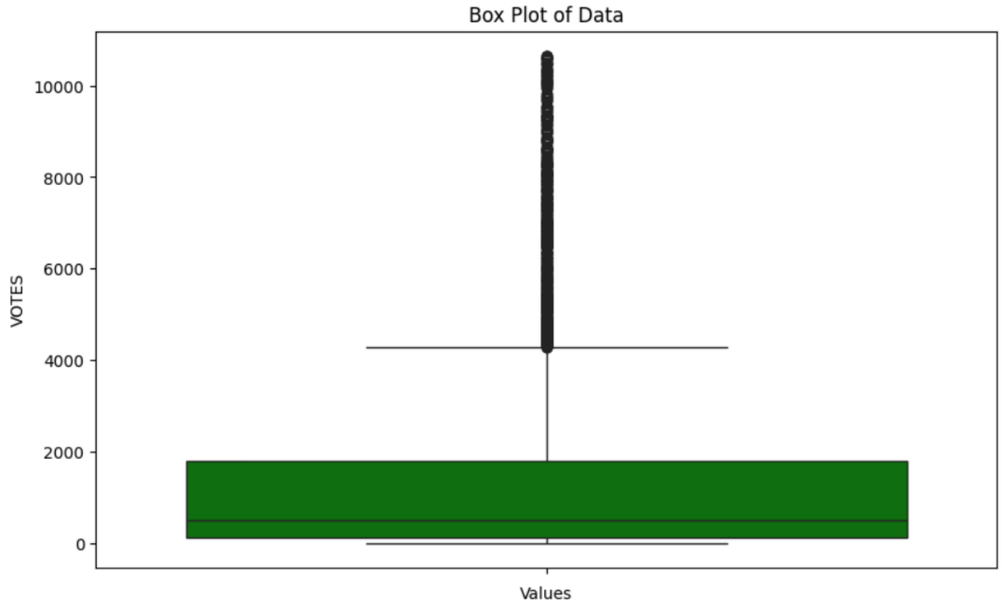
- The histogram showcases the distribution of the number of votes/ratings received by movies in the dataset, providing insights into the frequency of different levels of audience interest.



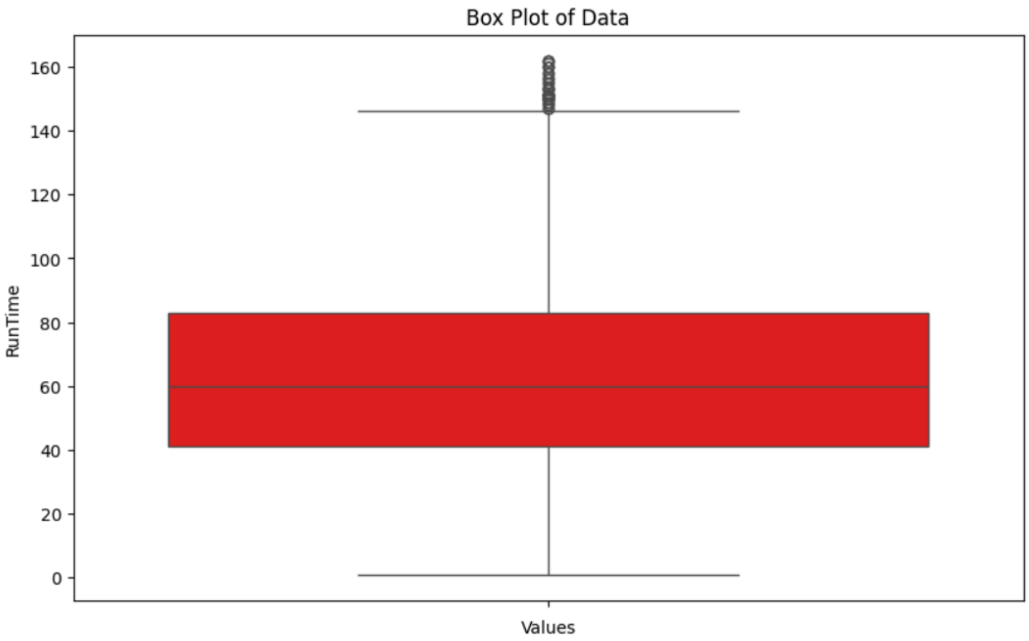
- The histogram visualizes the distribution of movie runtimes in the dataset, offering an overview of the frequency of different duration intervals.



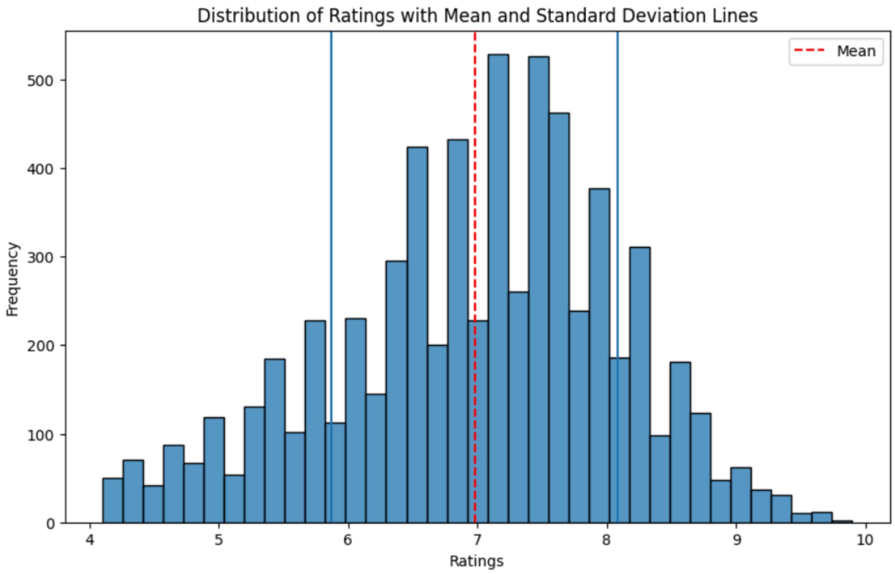
- The box plot presents the distribution of movie ratings, showcasing the central tendency, spread, and presence of outliers within the dataset.



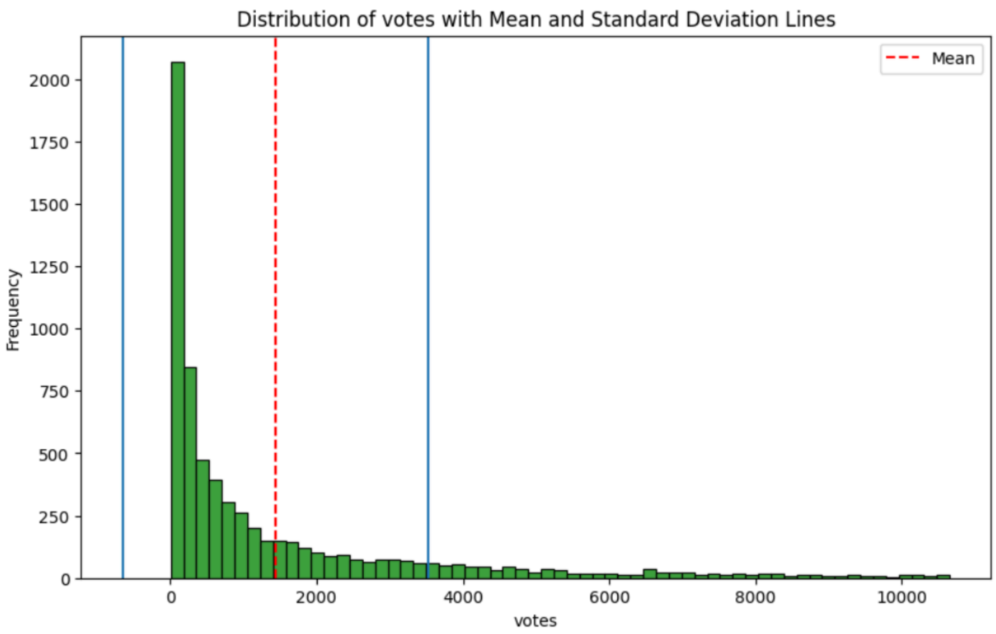
- This box plot illustrates the distribution of the number of votes/ratings received by movies, providing insights into central tendency, variability, and the presence of potential outliers in the dataset.



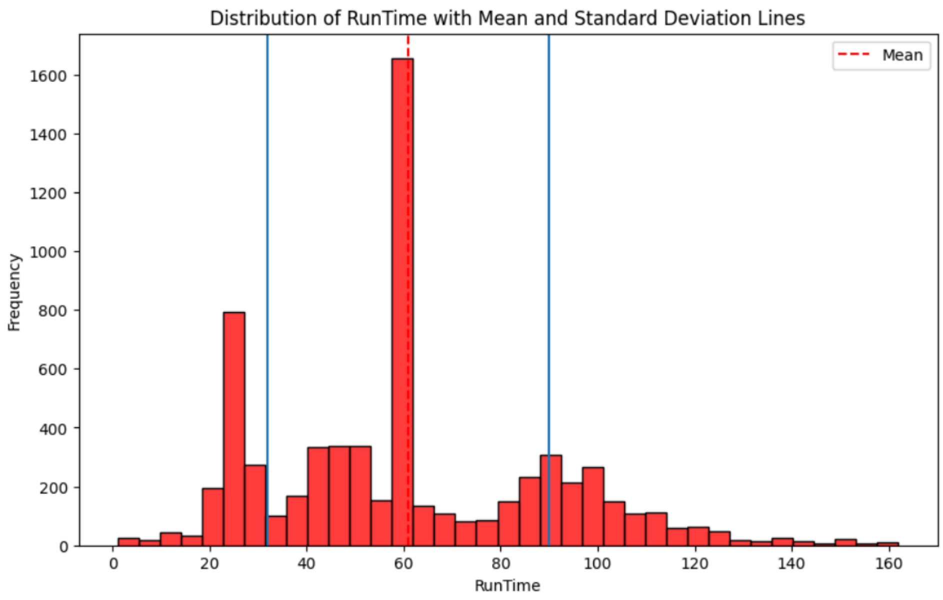
- This box plot represents the distribution of movie runtimes, displaying key statistical measures such as median, quartiles, and potential outliers, allowing for the identification of central tendency and variability in the dataset.



- The histogram visualizes the distribution of movie ratings, with dashed lines indicating the mean rating (red) and one standard deviation from the mean (blue), providing insights into the central tendency and dispersion of ratings in the dataset.



- Above histogram with dashed lines denoting the mean number of votes (red) and one standard deviation from the mean (blue), the histogram shows the distribution of the number of votes/ratings that movies received. This information can be used to understand the dataset's central tendency and variability in audience engagement.



- The histogram demonstrates the distribution of movie runtimes, with dashed lines indicating the mean runtime (red) and one standard deviation from the mean (blue), providing insights into the central tendency and variability of movie durations in the dataset.

References :

1) https://www.kaggle.com/datasets/bharatnatrayn/movies-dataset-for-feature-extracion-prediction?select=movies.csv
2) https://www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/
3) https://www.upwork.com/resources/data-cleaning-techniques
4) https://www.geeksforgeeks.org/how-to-conduct-a-two-sample-t-test-in-python/amp/
5) https://www.itl.nist.gov/div898/handbook/eda/section4/eda43.htm
6) https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf