Analyzing Movie Ratings and Exploring Probability Distributions

1. Introduction

In this project, I will explain a dataset of movie ratings to understand patterns and variability. The analysis uses arrangement, bivariate relationships and linear regressions with visualization cavalry. Our aim is to get an understanding of the data and how movie ratings are distributed.

2. Data Description

The dataset consists of two CSV files:

movies.csv: Contains movie information with columns movieId, title, and genres.

ratings.csv: Contains user ratings with columns userId, movieId, rating, and timestamp.

3. Data Preprocessing

The datasets were loaded and combined to generate one comprehensive dataset for analysis. This in turn meant that we had to merge the two data sets on their common movieId column.

4. Data Visualization and Descriptive Statistics

4.1 Movie reviews distribution

Histogram and KDE (Kernel Density Estimate) plot showing the distribution of movie ratings This helps in understanding the distribution of ratings.

4.2 Descriptive Statistics

Movie rating was summarized using descriptive statistics. The characteristics describe the data and include key values such as mean, standard deviation, minimumm maximumumand quartiles.

Count: The total number of ratings in the dataset.

Mean: The average rating.

Standard Deviation: The measure of the amount of variation in the ratings.

Minimum and Maximum: The range of ratings.

25th, 50th (Median), and 75th Percentiles: The distribution of the ratings.

5. Linear Regression to Predict Average Movie Ratings

A linear regression model was used to predict the average movie rating based on the number of ratings a movie received. This analysis involved:

- Preparing the data by grouping and counting ratings per movie.
- Splitting the data into training and testing sets.
- Training a linear regression model.
- Evaluating the model's performance using mean squared error and visualizing the actual versus predicted ratings.

6. Conclusion

This project investigates the ratings distribution over movies and described this data using descriptive statistics, also tested it with simple linear regression.