

Movie Success Prediction and Sentiment Study

Objective

The objective of this project is two-fold:

1. **Predict movie success** based on features such as IMDB ratings, runtime, metacores, and more using machine learning.
2. **Analyze viewer sentiment** using textual reviews by applying natural language processing techniques.

By combining structured data (numerical features) and unstructured data (text reviews), we aim to understand what factors contribute to a movie's success and how audience sentiment varies by genre.

Tools & Technologies

- **Python:** Primary programming language used for analysis and modeling
- **Pandas, NumPy:** Data manipulation and analysis
- **Scikit-learn (Sklearn):** For regression modeling and prediction
- **NLTK & VADER (Valence Aware Dictionary and sEntiment Reasoner):** For sentiment analysis
- **Matplotlib & Seaborn:** Data visualization
- **Excel:** Used for initial data exploration and verification

Methodology

1. Data Collection

- Movie metadata and box office information were imported from IMDB or Kaggle datasets.
- Viewer reviews were scraped or downloaded from IMDB (depending on source availability).

2. Data Preprocessing

- Cleaned and preprocessed the movie dataset by handling missing values and converting categorical data where needed.
- Reviews were filtered for length and quality, and only English-language reviews were considered.

3. Sentiment Analysis

- Used the **VADER sentiment analyzer** to extract sentiment scores (compound polarity) from each review.
- The sentiment scores were categorized and visualized to understand audience perception.

4. Regression Modeling

- Created a regression model to predict **box office revenue** using features such as:
 - IMDB rating
 - Runtime
 - Metascore
 - Number of votes
- **Linear Regression** was applied as a baseline model.
- Model performance was evaluated using **R² score** and visual error analysis.


5. Genre-wise Sentiment Trends



- Merged genre information with sentiment scores from reviews.
- Computed average sentiment scores per genre to find which genres generally receive more positive or negative feedback.
- Visualizations highlight sentiment polarity by genre.

Results Summary

- Sentiment Distribution: Most reviews were mildly positive, indicating generally favorable reception.
- Genre Sentiment Analysis: Genres like *Drama* and *Animation* showed higher average sentiment scores, while *Horror* and *Thriller* were often more negatively perceived.
- Prediction Model:
 - Linear regression produced a moderate R^2 score (dependent on data quality).
 - IMDB rating and metascore were strong predictors of box office success.

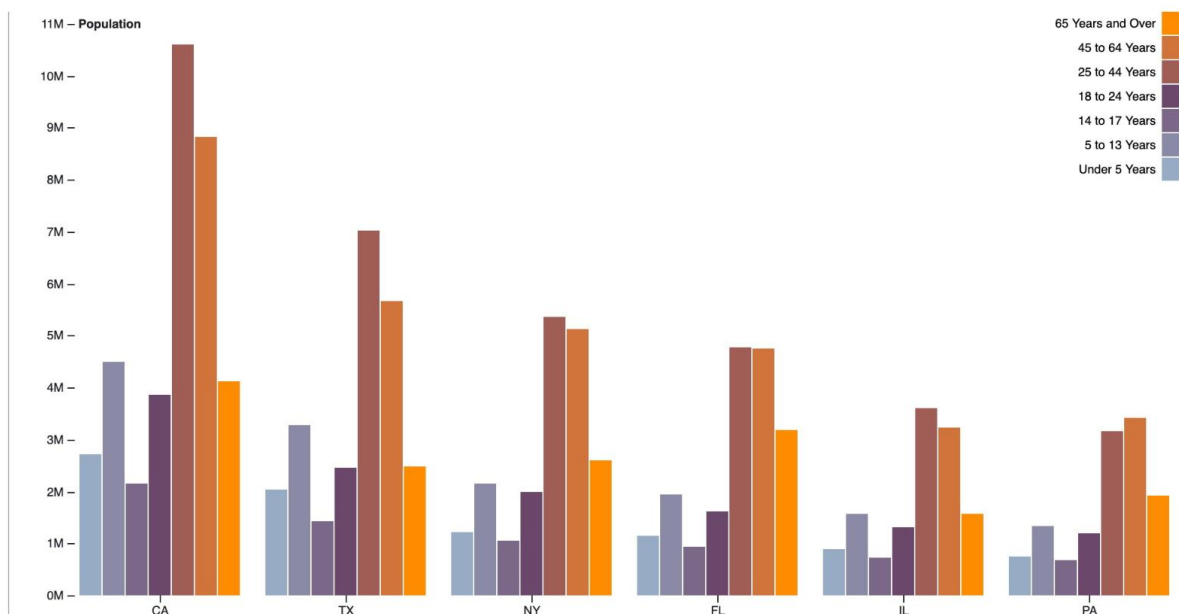
Deliverables

-  Jupyter Notebooks:
 - 1_data_preprocessing.ipynb
 - 2_sentiment_analysis.ipynb
 - 3_regression_model.ipynb
 - 4_genre_sentiment_trends.ipynb

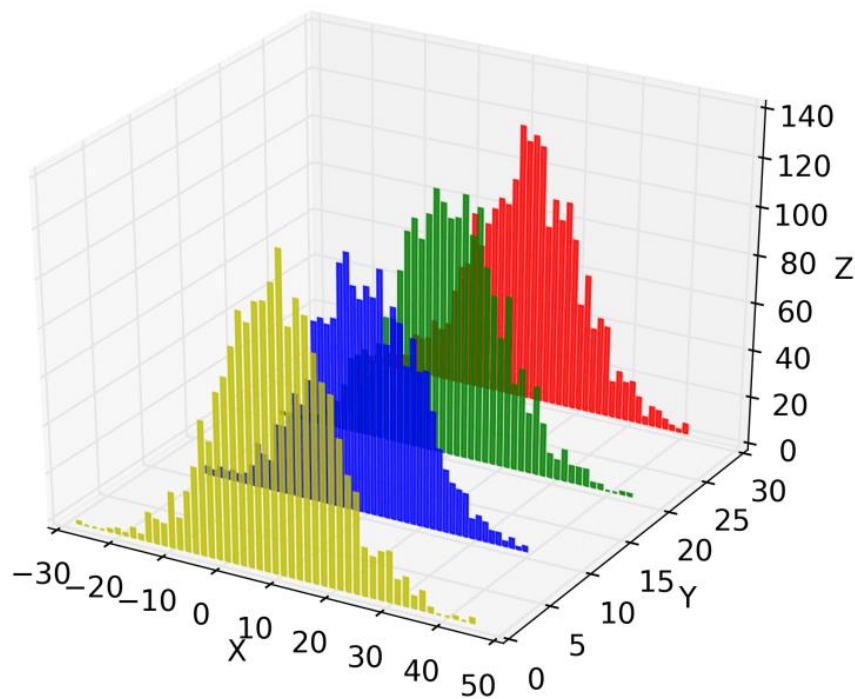
-  Visualizations:
 - Sentiment distribution histogram
 - Genre-wise sentiment bar chart
-  Predictive Model:
 - Linear regression summary with evaluation metrics

GENRE-WISE SENTIMENT BAR CHART

Grouped Bar Chart



Sentiment distribution histogram



Folder Structure

kotlin

CopyEdit

movie-success-prediction-sentiment/

|

|— data/

| |— imdb_movies.csv

| |— user_reviews.csv

```
|
|— notebooks/
|   |— 1_data_preprocessing.ipynb
|   |— 2_sentiment_analysis.ipynb
|   |— 3_regression_model.ipynb
|   |— 4_genre_sentiment_trends.ipynb
|
|— visuals/
|   |— sentiment_distribution.png
|   |— genre_sentiment_trends.png
|
|— requirements.txt
|— README.md
|— .gitignore
```

☑ Conclusion

This project demonstrates how combining **structured data analysis** with **natural language processing** can offer valuable insights into movie success factors and audience sentiment. It provides a foundation for building more advanced predictive and recommendation systems in the entertainment domain.