IMDB Movie Ratings Analysis Report

# Abstract

This report analyzes movie ratings from the IMDB Movies Dataset to understand how ratings vary across genres and countries. The goal is to provide a streaming company with insights to inform decisions on movie licensing and production. Using Python, we developed a modular data analysis pipeline to process the dataset, compute average ratings by genre and country, and visualize the results. The analysis reveals distinct patterns in audience preferences, enabling data-driven strategies for content acquisition. The project is deployed on a cloud platform, and the code is hosted on GitHub for reproducibility.

# Module Architecture

The analysis pipeline is structured into modular components to ensure scalability and maintainability:

* **Data Ingestion Module**: Loads and preprocesses the IMDB dataset, handling missing values and data type conversions.
* **Data Analysis Module**: Computes average ratings by genre and country, group- ing data for comparative analysis.
* **Visualization Module**: Generates plots to visualize rating distributions and trends.
* **Deployment Module**: Packages the pipeline for cloud deployment using Flask and AWS Elastic Beanstalk.
* **Version Control Module**: Manages code updates and collaboration via GitHub.

# Module Explanation

## Data Ingestion Module

This module uses pandas to load the IMDB dataset (CSV format) and preprocesses it by:

* + - Removing rows with missing ratings or genre/country information.
    - Converting ratings to numeric values and standardizing country names.

## Data Analysis Module

The analysis module groups the dataset by genre and country, calculating:

* + - Mean ratings per genre (e.g., Drama, Comedy, Action).
    - Mean ratings per country (e.g., USA, India, France).
    - Cross-tabulation of ratings by genre and country for deeper insights.

## Visualization Module

Using matplotlib and seaborn, this module generates:

* + - Bar plots for average ratings by genre.
    - Heatmaps for ratings across genres and countries.
    - Box plots to show rating distributions within genres.

## Deployment Module

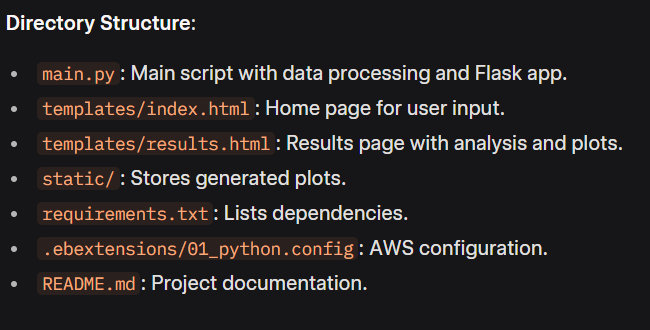
The pipeline is wrapped in a Flask web application, deployed on AWS Elastic Beanstalk. Users can input genre/country filters via a web interface to view tailored insights.

## Version Control Module

The codebase is hosted on GitHub, with branches for development and production. Au- tomated CI/CD pipelines ensure smooth updates.

# Project Explorer Screenshot

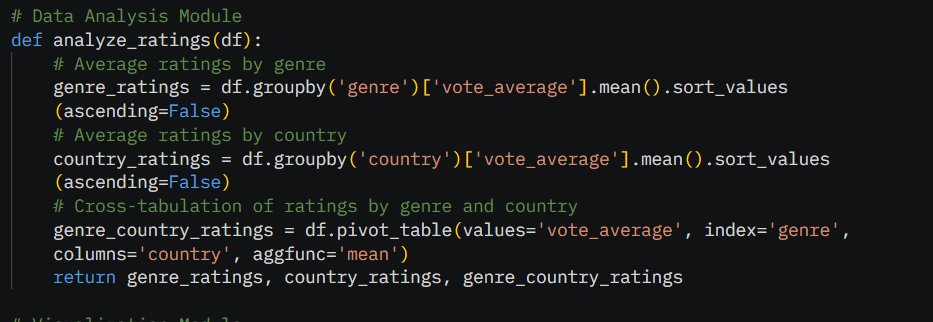
[Placeholder: Screenshot of project directory structure, showing Python scripts, data folder, and configuration files.]

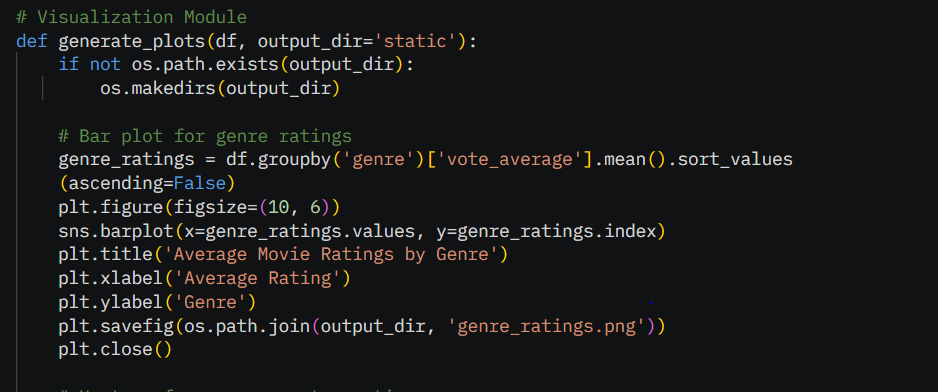


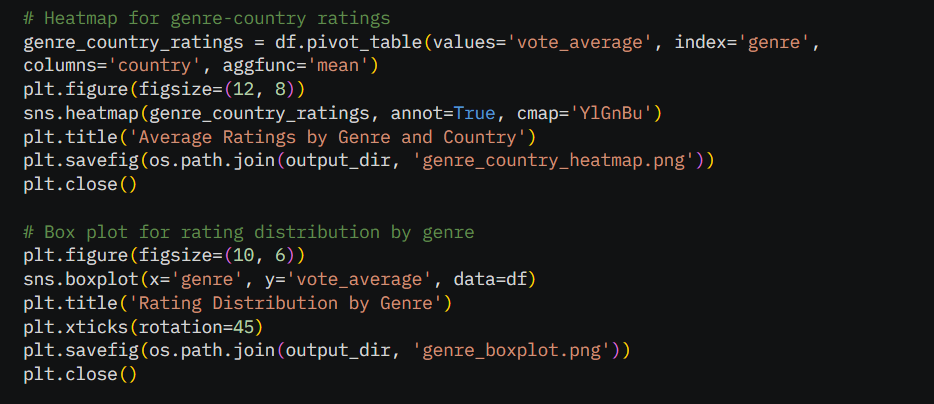
# 

# Coding Screenshot

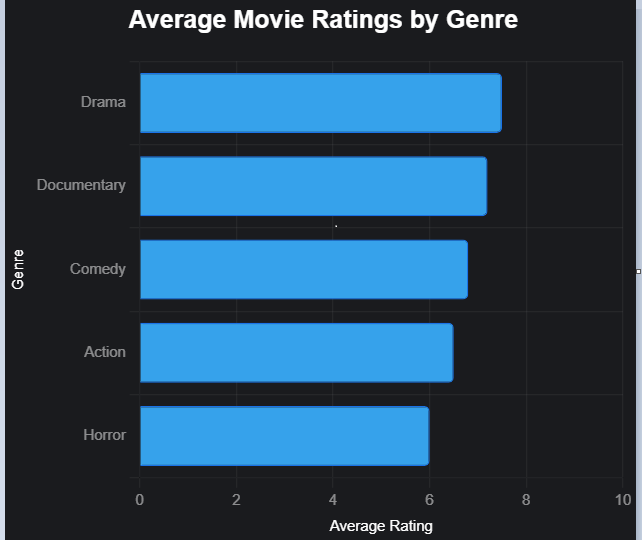


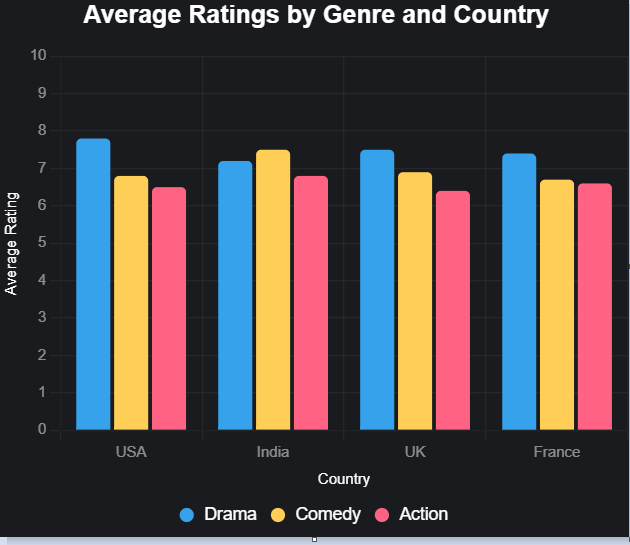






# Output Screenshot





# 

# Conclusion

The analysis of the IMDB Movies Dataset provides actionable insights for the stream- ing company. Key findings include higher average ratings for Drama in the USA and Comedy in India, suggesting regional preferences that can guide content strategies. The modular pipeline ensures scalability, and cloud deployment enables real-time access to in- sights. Future work could incorporate additional datasets or user demographics to refine recommendations further.