

# Cover Page

Course Title: Machine Learning

Section: \_\_\_\_\_

Student Name / ID:

- Mohamed Mostafa : 23101594
- Marwan Khaled : 23101599
- Mohamed Adel : 23101899

Generated on: 2025-12-20 11:39

# **Section 1: Problem Domain**

Streaming-catalog triage for international mid-/long-tail films.

Tasks: engagement classification + revenue regression.

## Section 2: Project Summary

Dataset rows/cols: 10014 / 13

Cleaning: missing values, formatting fixes, duplicate removal, outlier removal (IQR).

Models: Logistic Regression, Random Forest, SVM, KNN (classification + regression variants).

Compared preprocessing: none vs StandardScaler vs MinMaxScaler.

## **Section 3: Source Code**

Notebook: MI\_Project\_Full\_Rebuild.ipynb

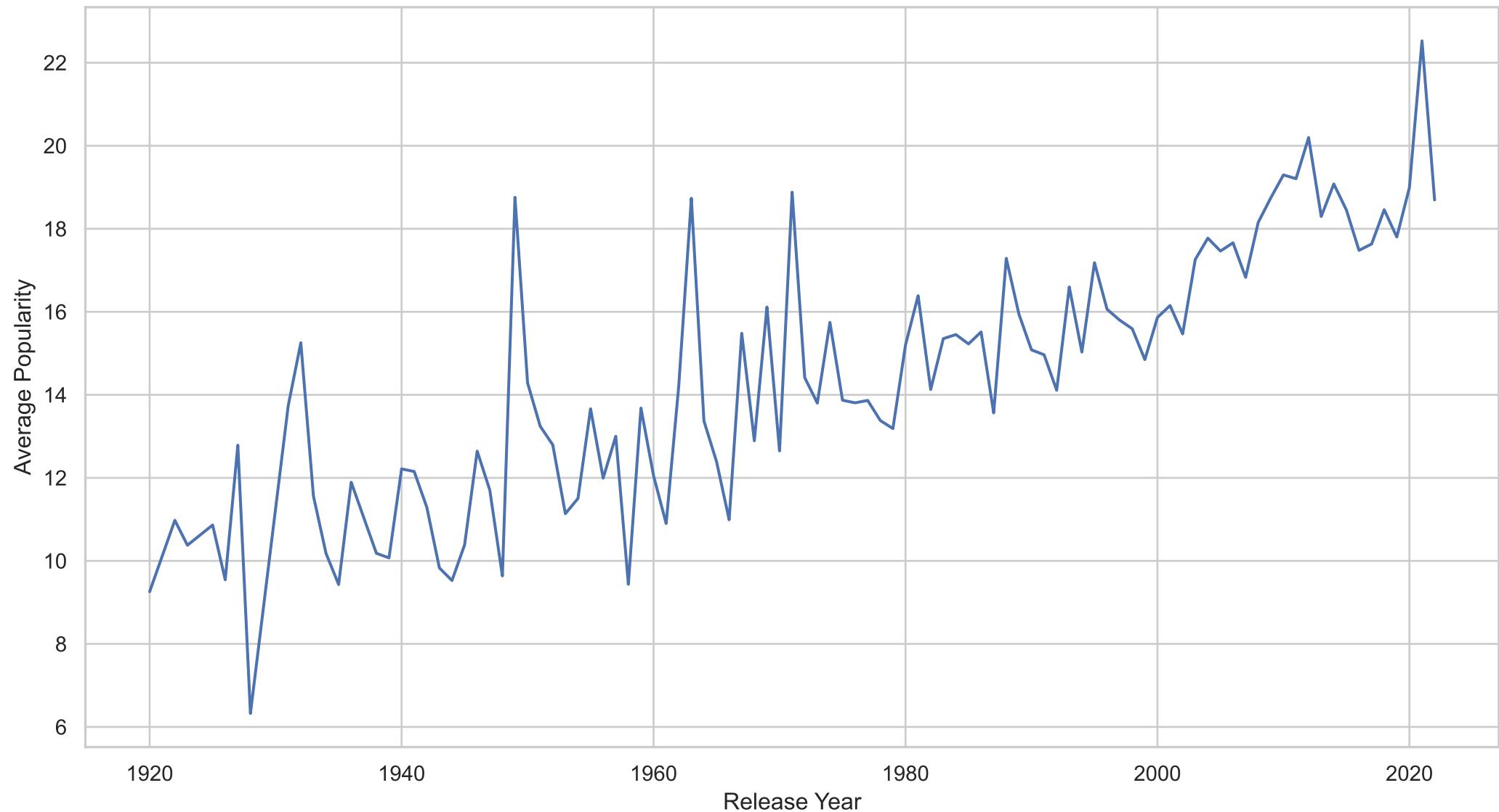
Dataset: Top\_10000\_Movies.csv

Download template: download\_dataset.py

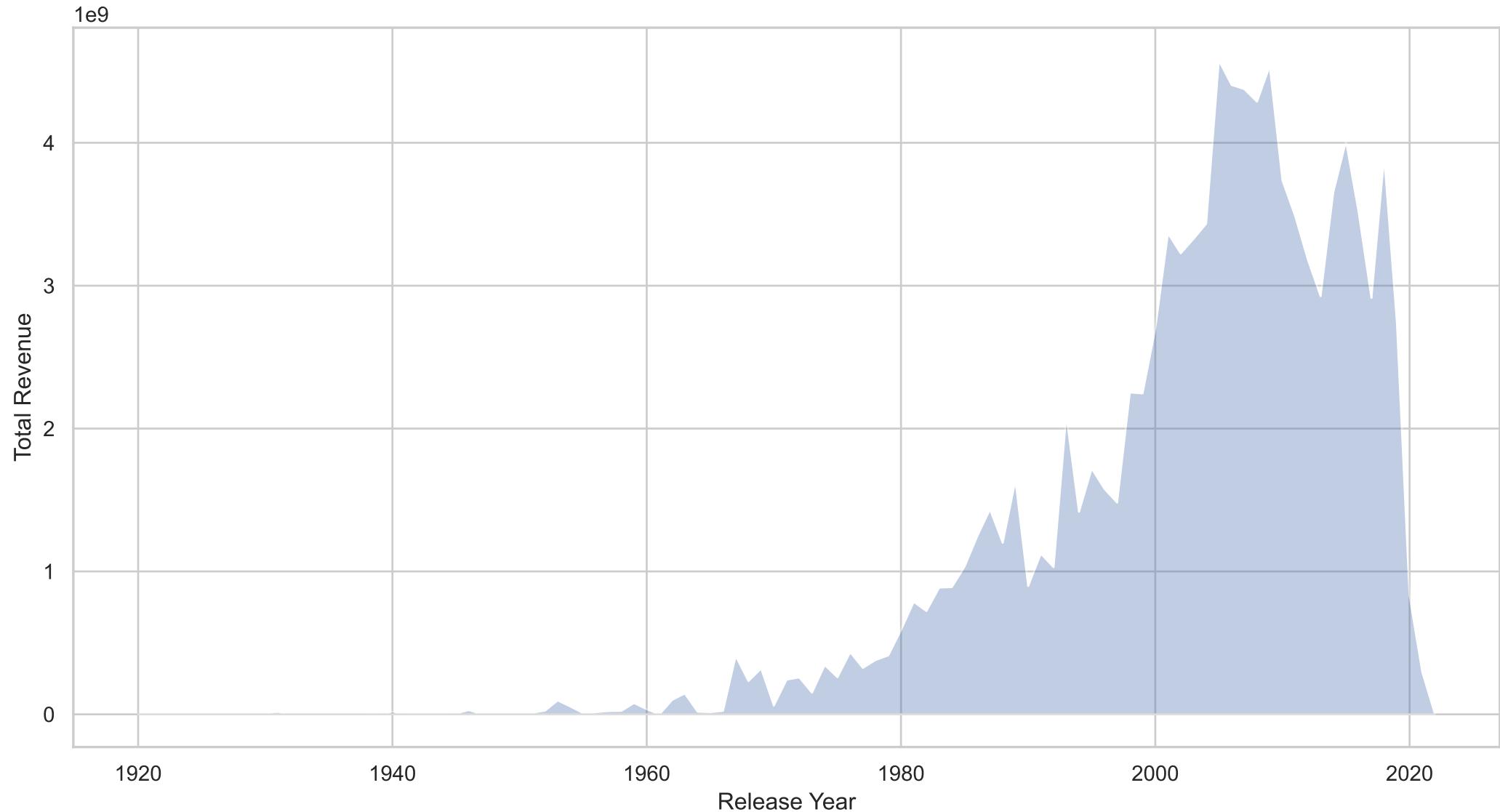
## **Section 4: Visualization Snapshots**

All graphs generated in the notebook are included in the following pages.

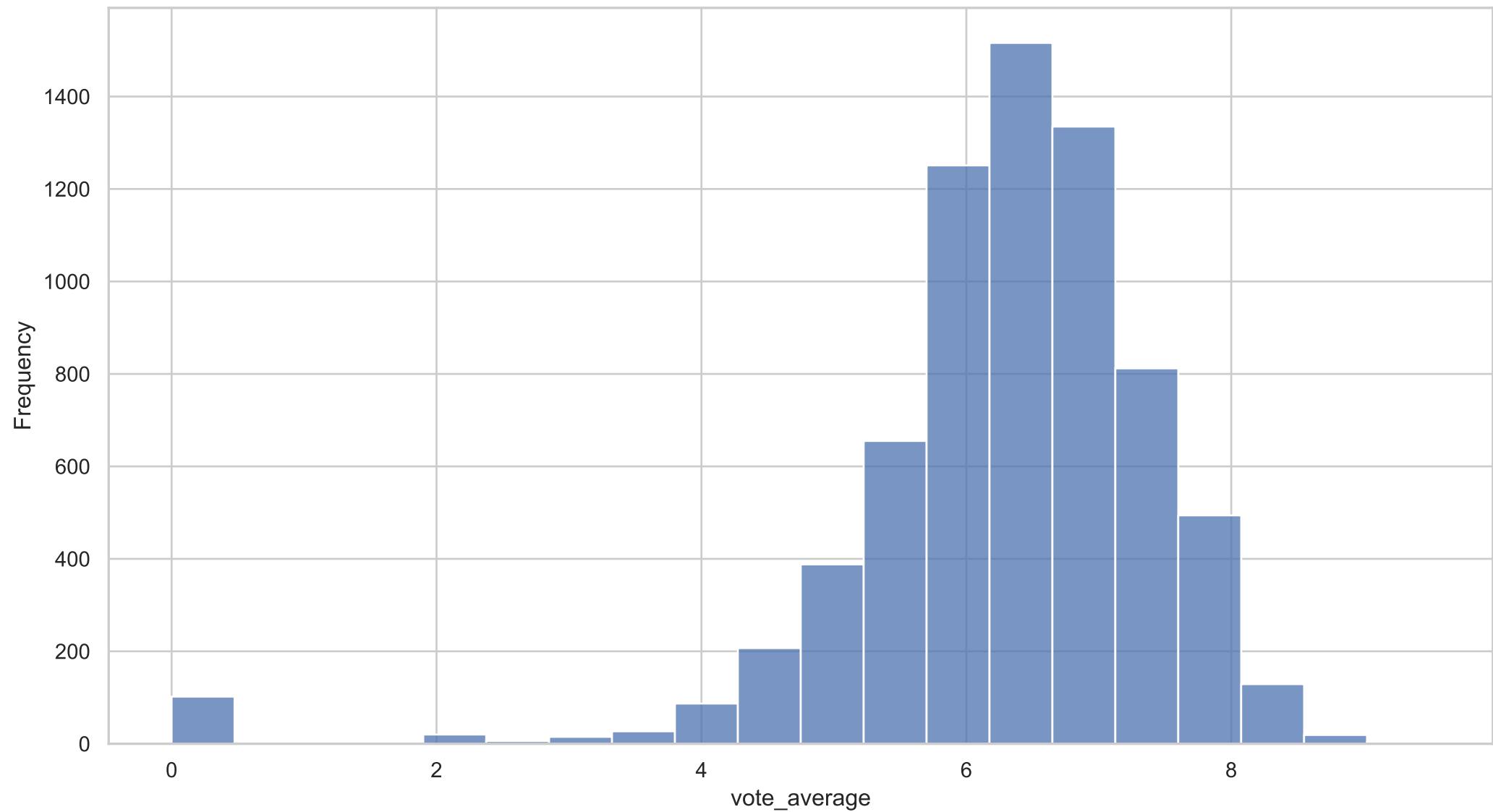
## Line Plot: Average Popularity by Release Year



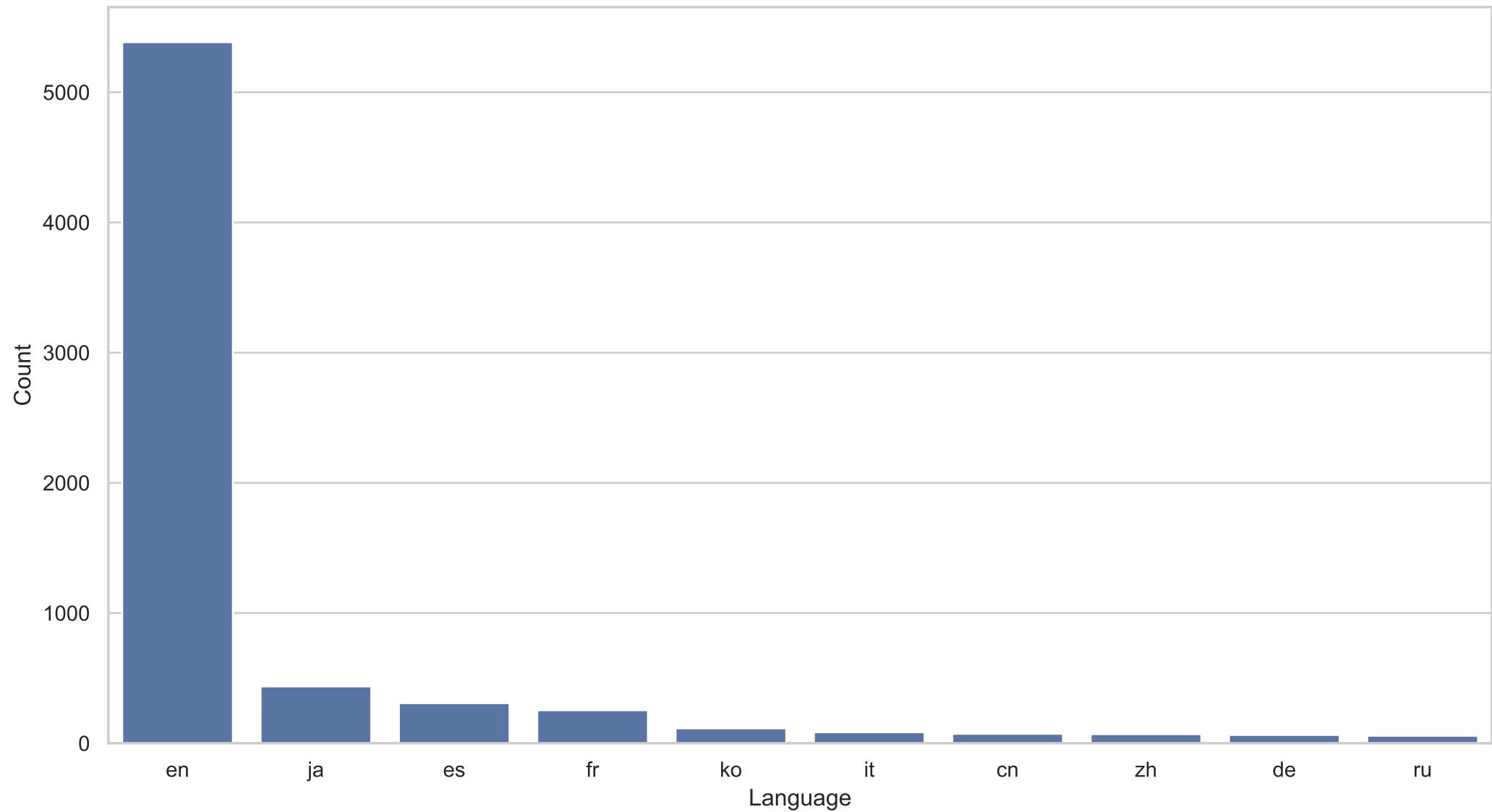
## Area Plot: Total Revenue by Release Year



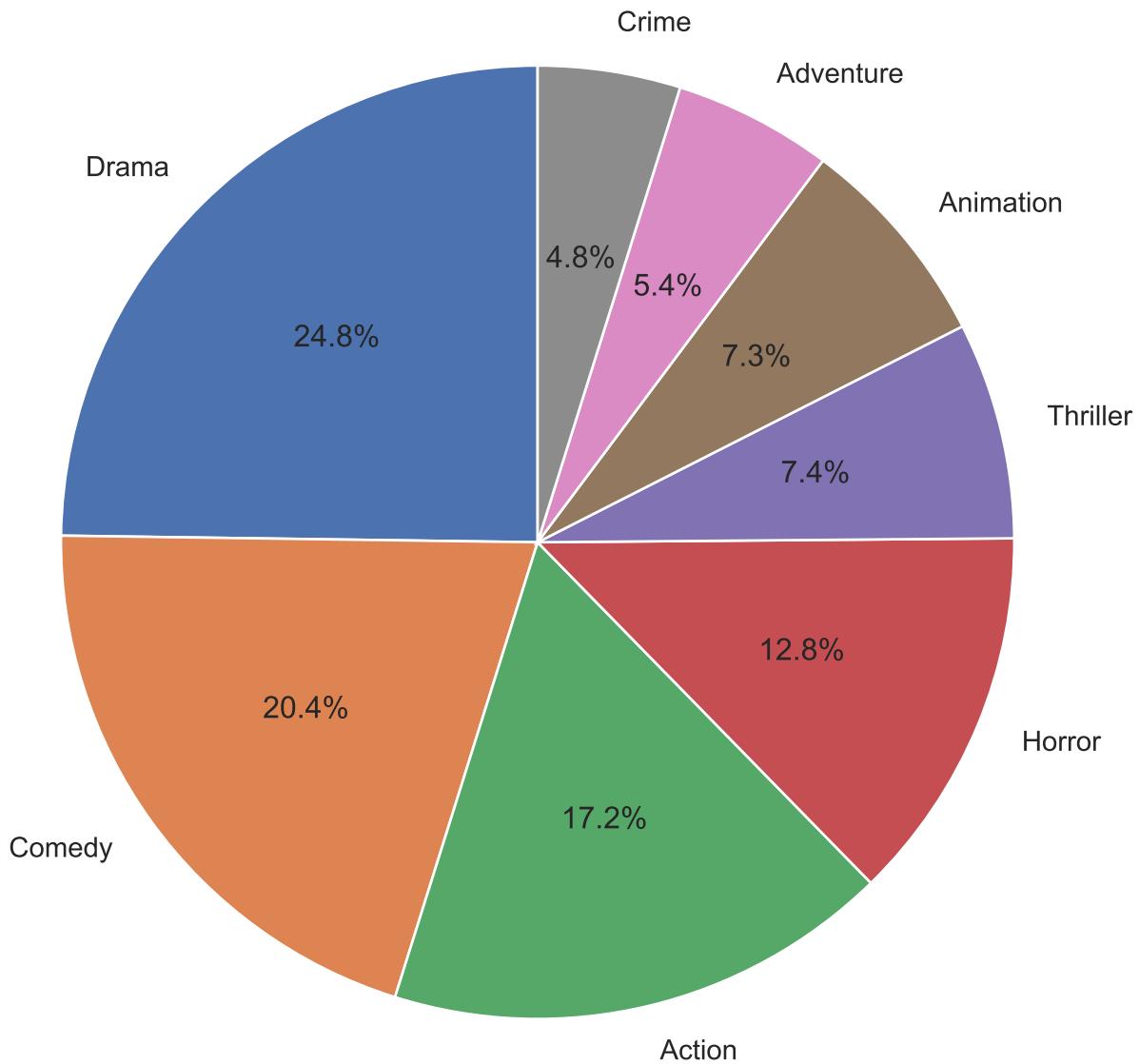
# Histogram: Vote Average Distribution



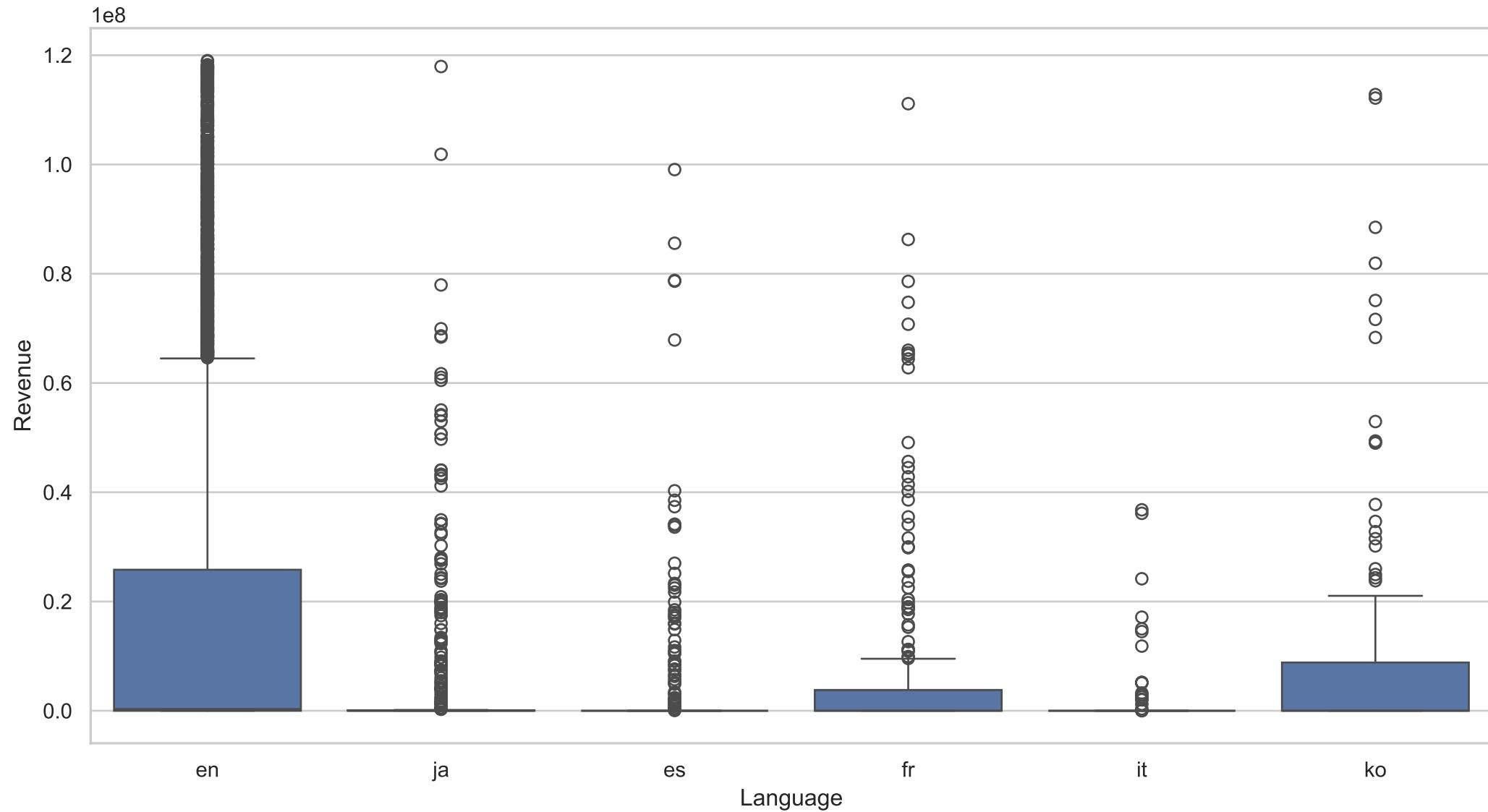
## Bar Chart: Top 10 Original Languages



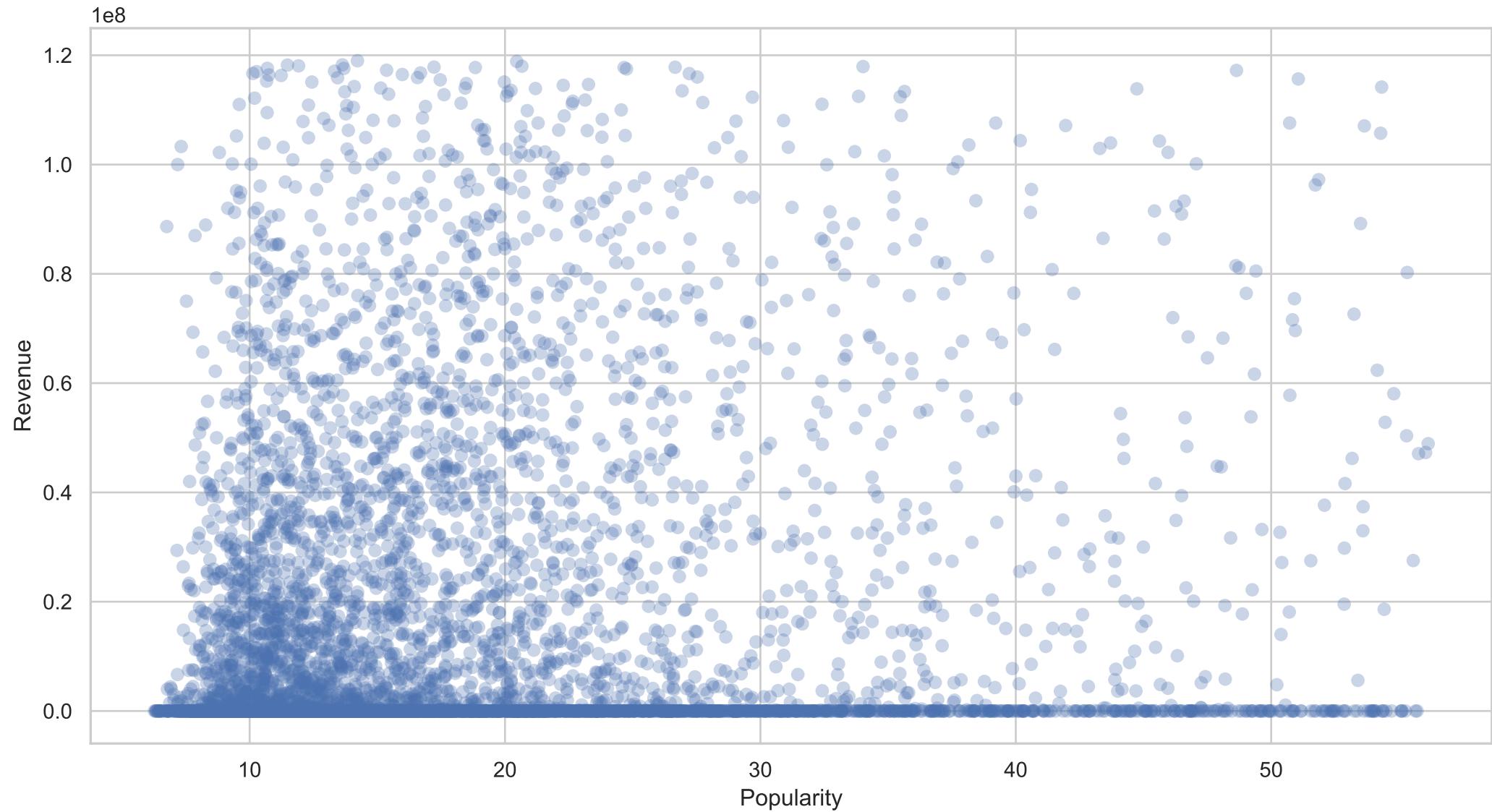
# Pie Chart: Top Genres (Main Genre)



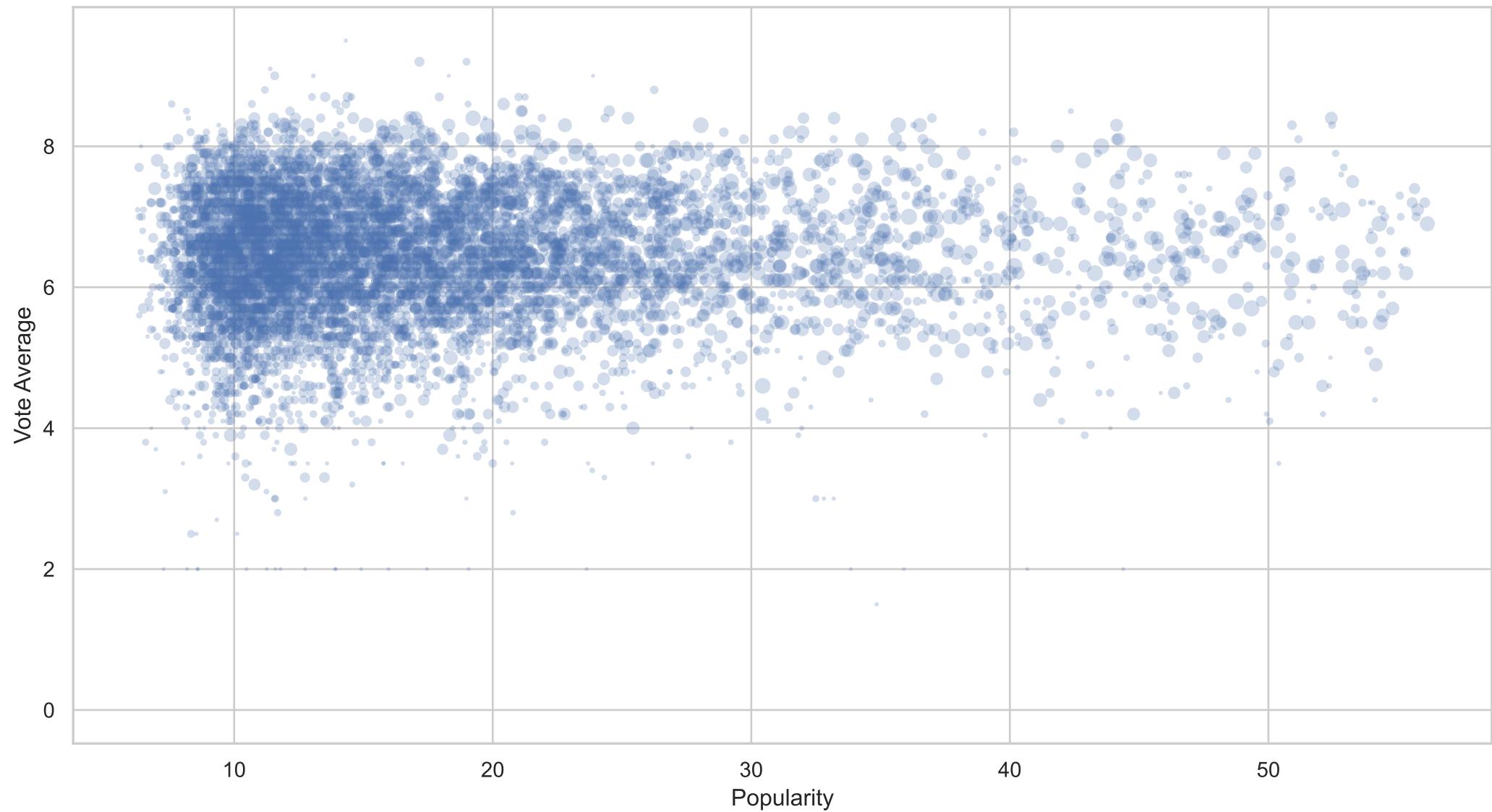
# Box Plot: Revenue by Language (Top 6)



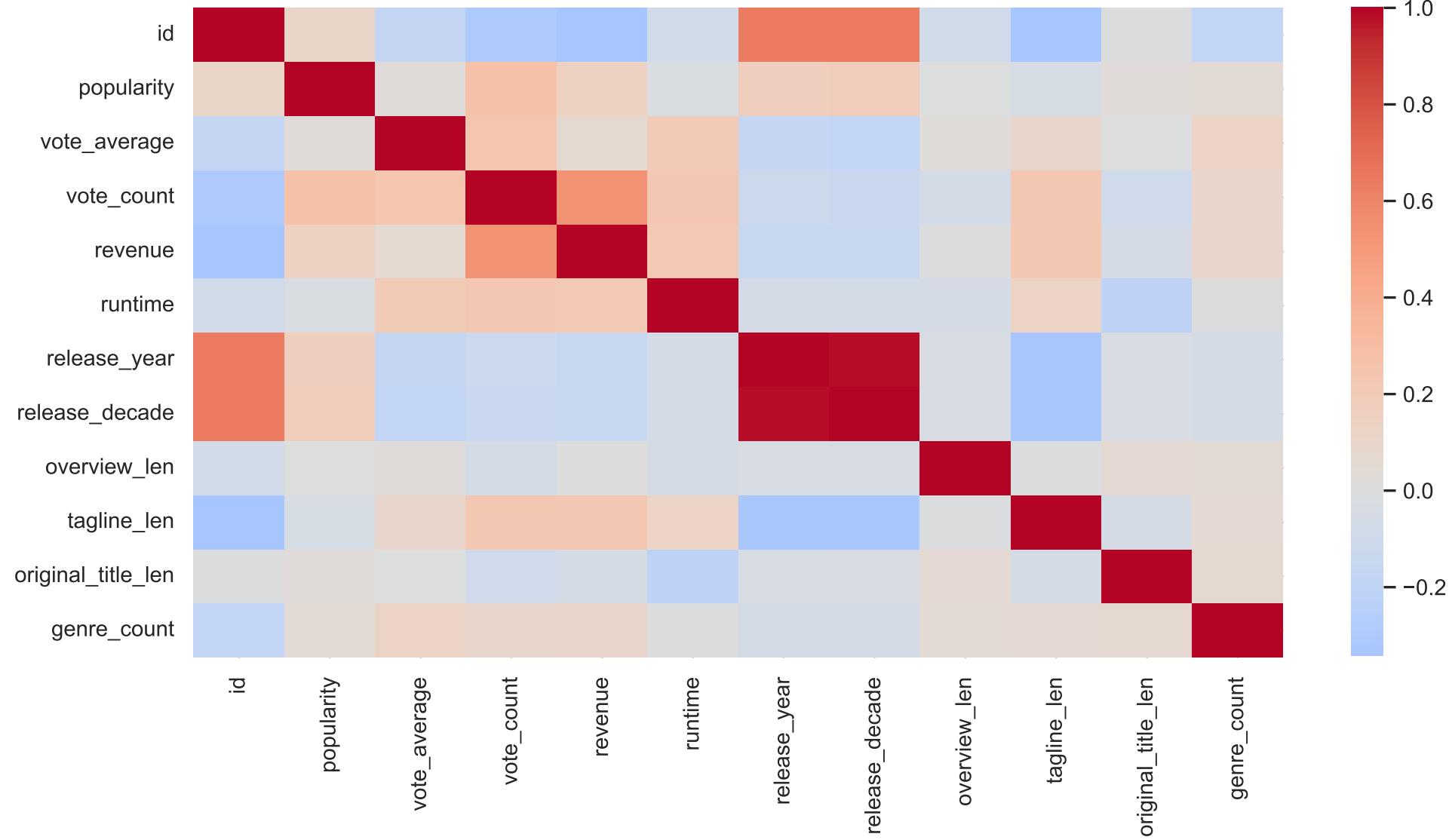
# Scatter Plot: Popularity vs Revenue



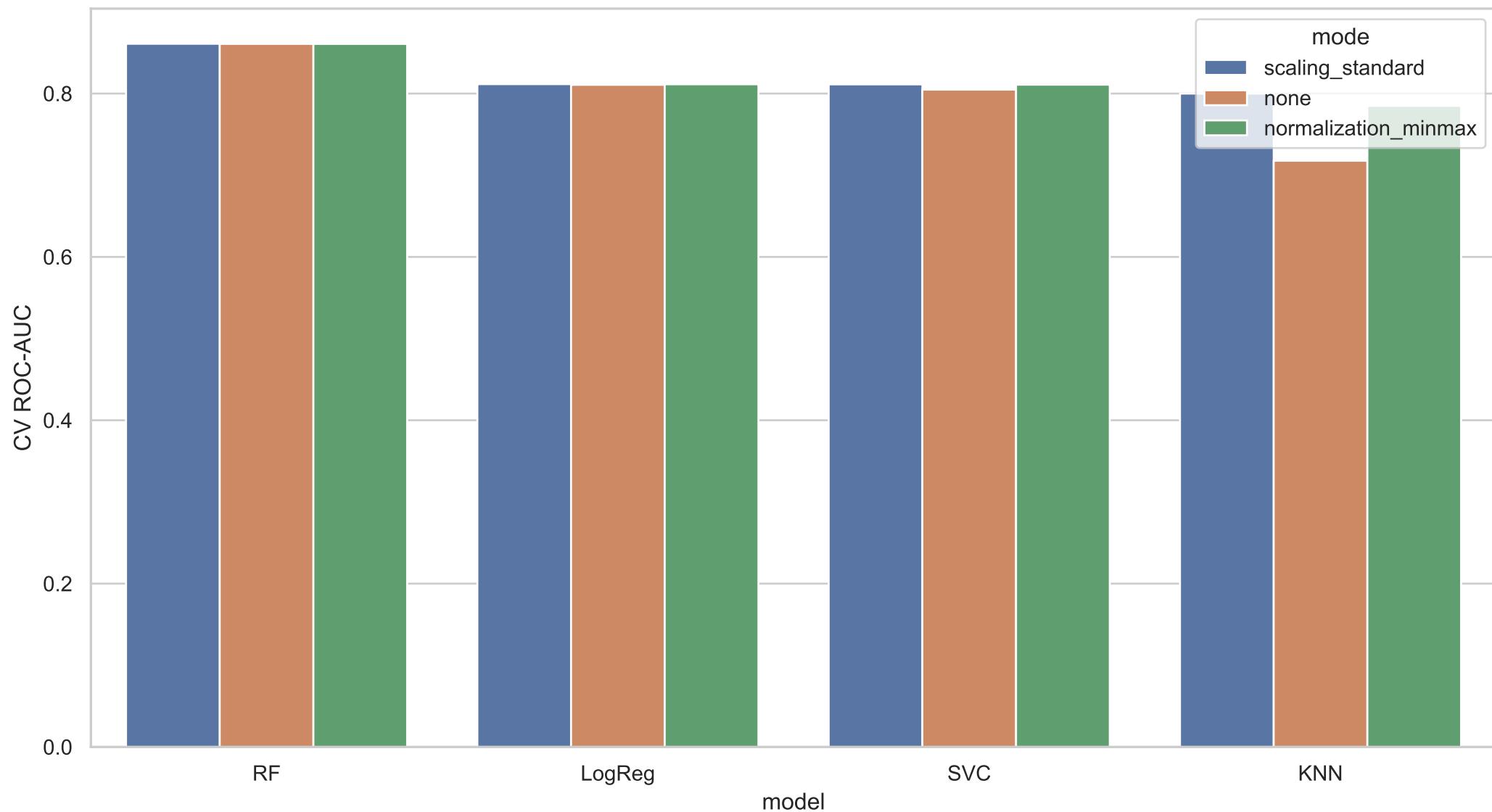
### Bubble Plot: Popularity vs Vote Average (size = vote\_count)



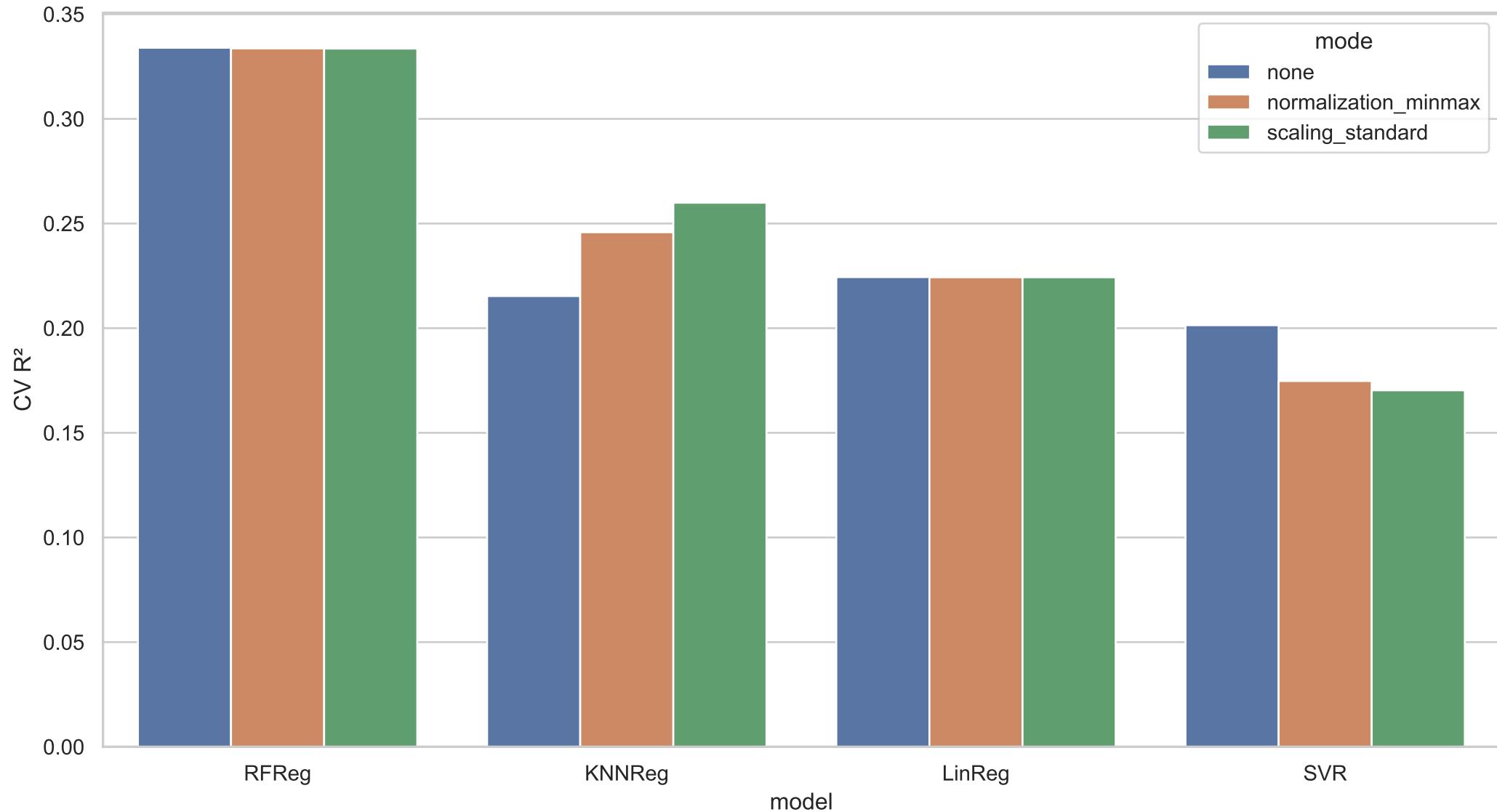
## Correlation Heatmap (Numeric Features)



## Classification Model Comparison (CV ROC-AUC)



## Regression Model Comparison (CV R<sup>2</sup>)



## ROC Curve (Hold-out Set)

