# How can genres, languages, and other factors predict the popularity of TV shows?

## Full TMDb TV Shows Dataset 2024 (150K Shows)

## Introduction

In today's rapidly evolving entertainment landscape, understanding what drives the success of TV shows has become increasingly important, especially with the rise of streaming platforms. The Full TMDb TV Shows Dataset 2024 provides a comprehensive collection of data on over 150,000 TV shows, sourced from The Movie Database (TMDb)—a widely recognized platform offering valuable insights into movies and TV shows. This dataset includes detailed information such as genres, networks, release years, vote counts, languages, and more.

By analyzing this dataset, we can uncover which factors influence a show's popularity and, more importantly, predict the success of future TV shows. The primary goal of this project is to develop a predictive model that estimates a TV show's popularity based on key characteristics like genre, language, vote count, and more. Popularity is a crucial metric in the entertainment industry, as it directly impacts decisions related to production budgets, marketing strategies, and content recommendations.

Through this project, we aim to explore the complex relationships between various factors that contribute to a TV show's success, ultimately helping predict which shows are likely to resonate with audiences. Machine learning techniques were employed to build a model capable of accurately predicting TV show popularity. The dataset offers a wealth of information that can influence viewership, including attributes like genres, languages, release dates, ratings, and audience feedback. By uncovering patterns and trends in these features, we can better understand what makes a TV show appealing and why it succeeds in the market.

Predicting TV show popularity holds significant value for multiple stakeholders in the entertainment industry. Content creators can gain insights into the factors that resonate with audiences, enabling them to shape future productions. For distributors and streaming platforms, predictions about which shows will perform well can optimize content offerings, marketing efforts, and distribution decisions. A reliable prediction model can also reduce risk and increase returns on content investments by ensuring that resources are allocated to shows with higher chances of success.

To achieve this, various machine learning models were applied to forecast the success of TV shows based on the dataset. Through data preprocessing, feature engineering, and model evaluation, the project aimed to create a system that provides data-driven recommendations for content development and distribution strategies.

Ultimately, this project demonstrates the power of machine learning in predicting trends within the entertainment industry. The insights derived from this analysis have the potential to guide content creation, marketing strategies, and even offer personalized recommendations for viewers, empowering stakeholders to make more informed, data-driven decisions.

## Dataset Overview

The TMDB TV Shows Dataset 2024 (https://www.kaggle.com/datasets/asaniczka/full-tmdb-tv-shows-dataset-2023-150k-shows) is a comprehensive collection of TV show data sourced from The Movie Database (TMDb), a widely-used platform for information on movies and television shows. This dataset contains valuable details, including ratings, genres, languages, release dates, cast, crew, and more. Featuring information on 150,000 TV shows, the dataset was carefully cleaned and preprocessed to ensure quality data for analysis.

The dataset served as the foundation for building machine learning models to uncover key drivers of TV show popularity. The goal was to extract actionable insights from this data that could benefit industry stakeholders, such as content creators, distributors, and streaming platforms, by helping them make informed decisions about content creation, distribution, and marketing strategies.

## Project Goals

The main objective of this project was to predict the popularity of TV shows based on features such as genres, languages, vote counts, and others. While it was known that factors like genre and language influenced a show's popularity, the exact relationships remained unclear. Through data analysis and machine learning, I aimed to explore and uncover these relationships.

Success was determined by the model's ability to predict popularity accurately, which in turn could help guide content strategy decisions. The key features impacting the model's predictions included the number of seasons, vote count, genre, language, networks, and release dates.

To improve prediction accuracy, several techniques were employed, including Exploratory Data Analysis (EDA), addressing skewness, visualizing histograms, and handling outliers. These efforts ensured the data was suitable for machine learning and helped refine the predictions, enabling streaming platforms to anticipate which shows were likely to resonate with their audiences. This insight could help optimize content offerings and marketing strategies.

## Stages of the Project

1. **Data Preparation**:

   In this stage, the dataset was cleaned and preprocessed to ensure it was suitable for analysis and model training. The data was filtered to include only TV shows released between January 1, 2017, and December 31, 2024. Columns with a high percentage of missing data, such as 'tagline' and 'created_by', were dropped to improve data quality. Redundant or similar columns were also removed to streamline the dataset. Additionally, the 'overview' column was extracted into a separate DataFrame to facilitate feature engineering. Dummy variables were created for categorical columns such as 'languages', 'networks', and 'genres'. Finally, text data was cleaned to ensure consistency and accuracy, ensuring that the dataset was ready for further analysis and model training.

2. **Exploratory Data Analysis (EDA)**:

   In this stage, the dataset was visually analyzed to explore the relationships between continuous and categorical features in relation to the target variable, **popularity**. Various visualization techniques, including histograms, distribution plots, bar charts, scatter plots, and correlation matrices, were employed to identify trends and patterns in key features such as the number of episodes, number of seasons, genres, languages, networks, and others. This step was crucial for gaining insights into the data and understanding how different variables might influence a show’s popularity.

   The skewness values (vote_count: 43.30, popularity: 35.86, networks_count: 24.38, number_of_episodes: 15.40, number_of_seasons: 11.81, and languages_count: 9.26) suggest that most features exhibit a high right skew, indicating a small number of data points with significantly higher values. Since these features did not follow a normal distribution, statistical methods that are robust to skewed data, such as non-parametric tests or transformations, were applied accordingly in the analysis.

**Presentation of Part of the Continuous Features** :

**Figure 1: Log Transformation**: Popularity vs. Log transformed. The log transformation was applied to compress the scale of highly skewed data. This made it easier to visualize the distribution, especially when extreme values were present.

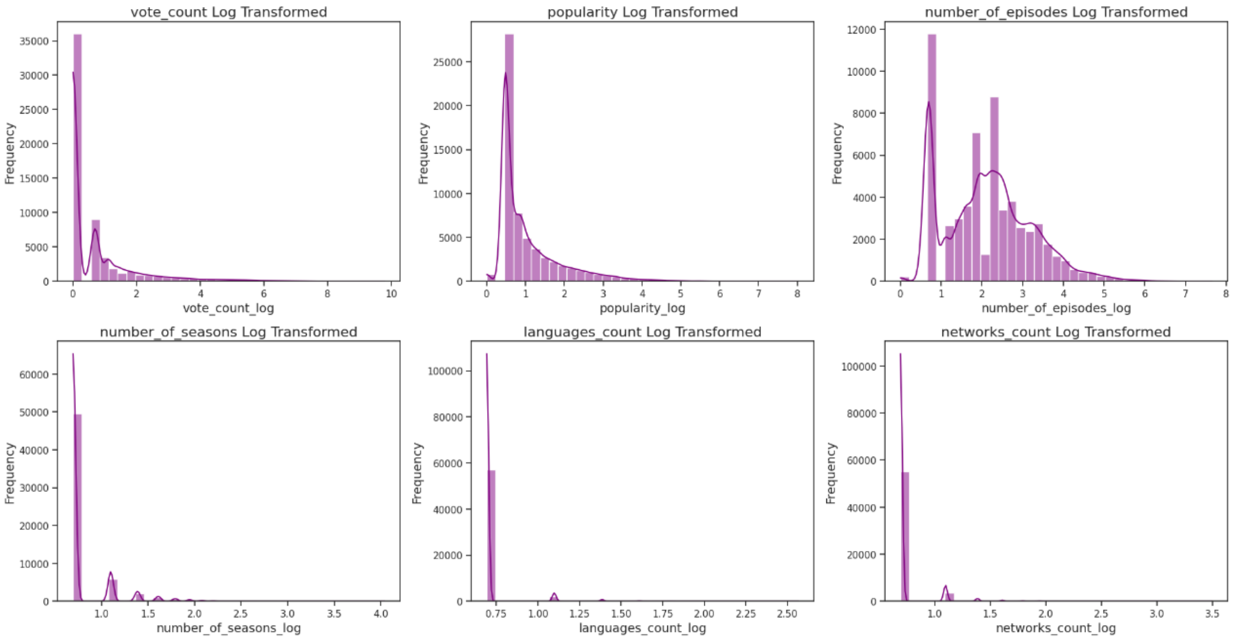


**Figure 2:** Box Plot: A box plot showed distribution and outliers.

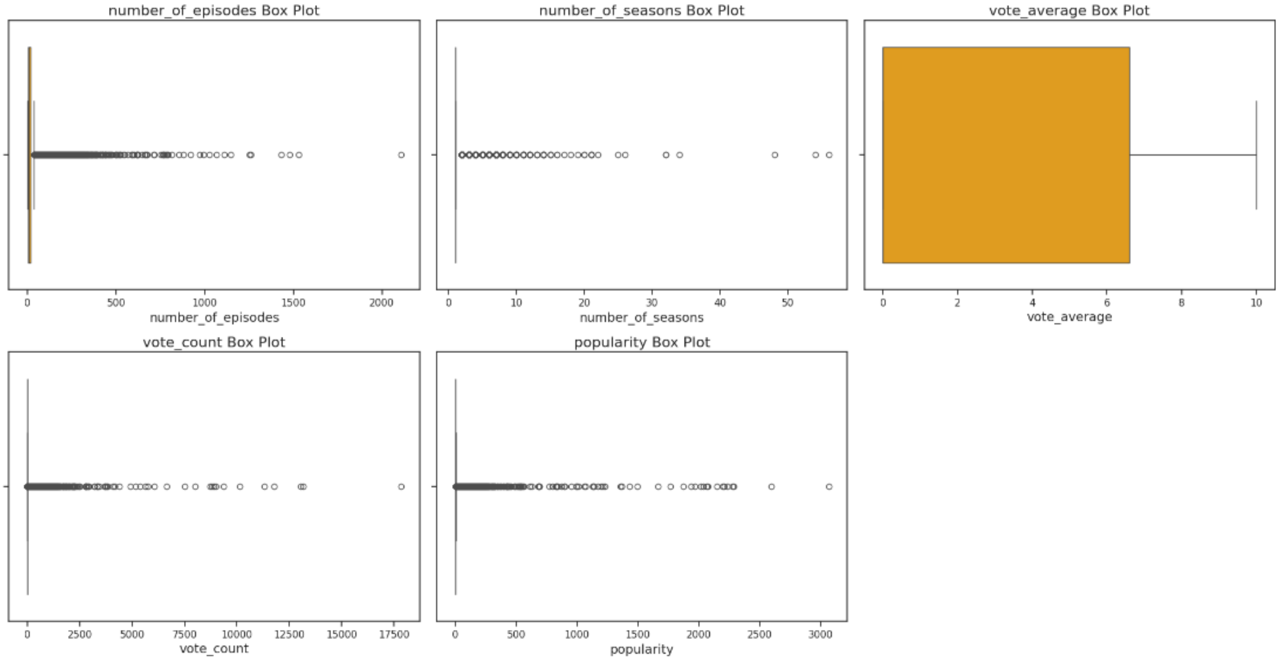
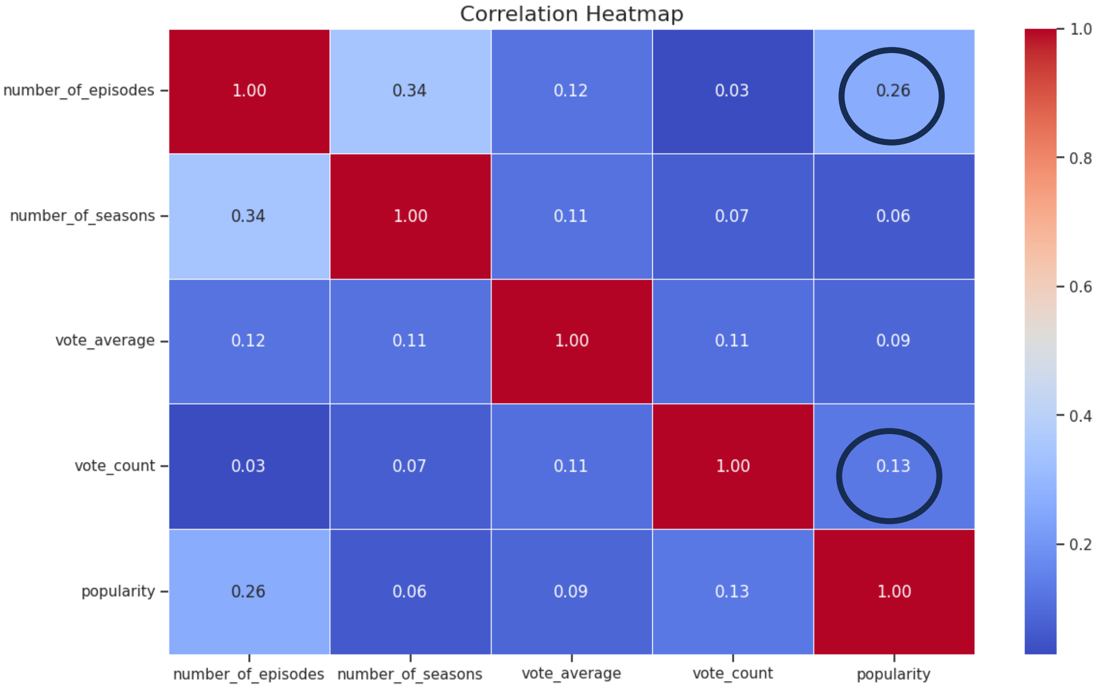


**Figure 3**: Correlation Heatmap of Continuous Variables: This plot visualized correlations among numerical features. 0.2 to 0.4 Weak positive correlation ('popularity' and number_of episodes')=0.26. 0 to 0.2 Very weak or no positive correlation ('popularity' and 'vote_count')= 0.13



| | number_of_episodes | number_of_seasons | vote_average | vote_count | popularity |
|---|---|---|---|---|---|
| number_of_episodes | 1.00 | 0.34 | 0.12 | 0.03 | 0.26 |
| number_of_seasons | 0.34 | 1.00 | 0.11 | 0.07 | 0.06 |
| vote_average | 0.12 | 0.11 | 1.00 | 0.11 | 0.09 |
| vote_count | 0.03 | 0.07 | 0.11 | 1.00 | 0.13 |
| popularity | 0.26 | 0.06 | 0.09 | 0.13 | 1.00 |

**Figure 4**:  Visualizing the Top 10 Most Frequent Categories for Each Categorical Feature: This plot showed the most frequent categories in variables like genre, language, and networks.

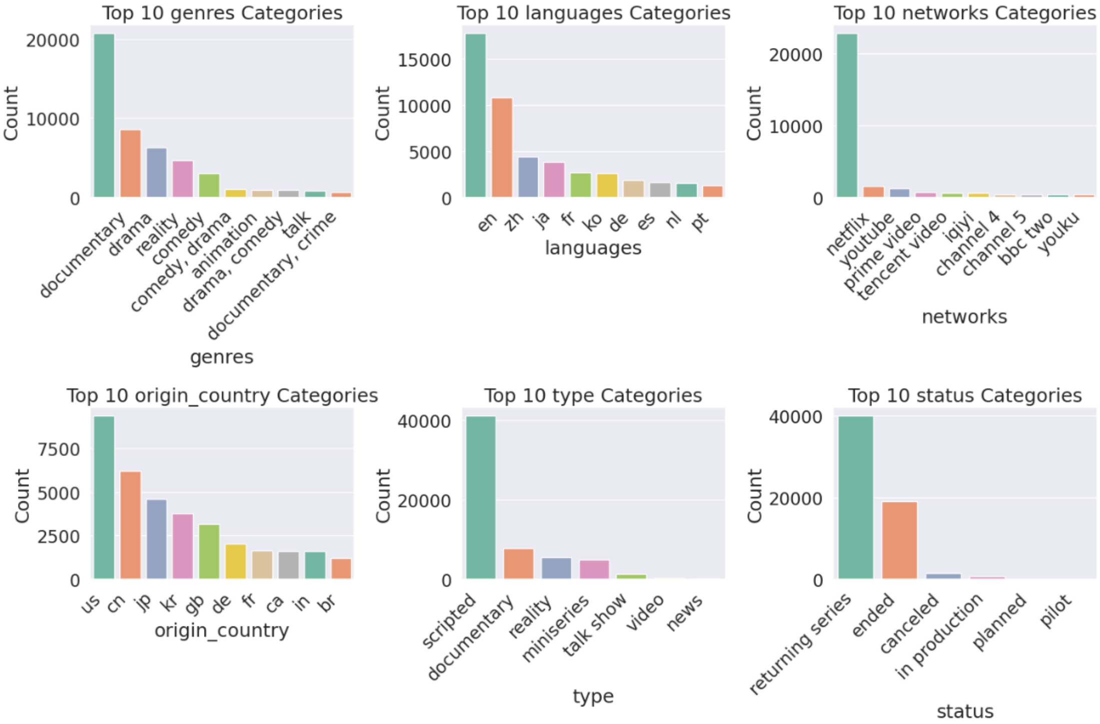
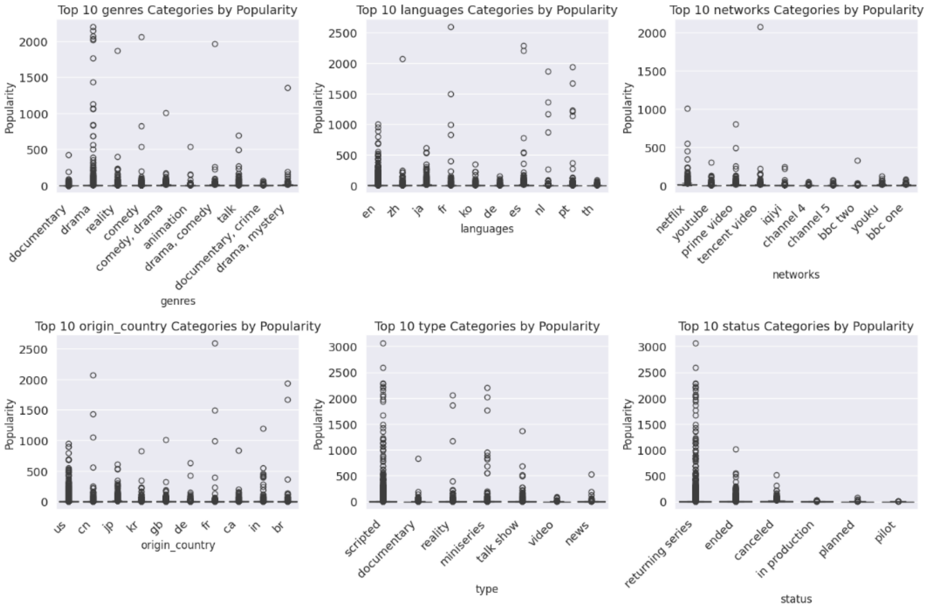
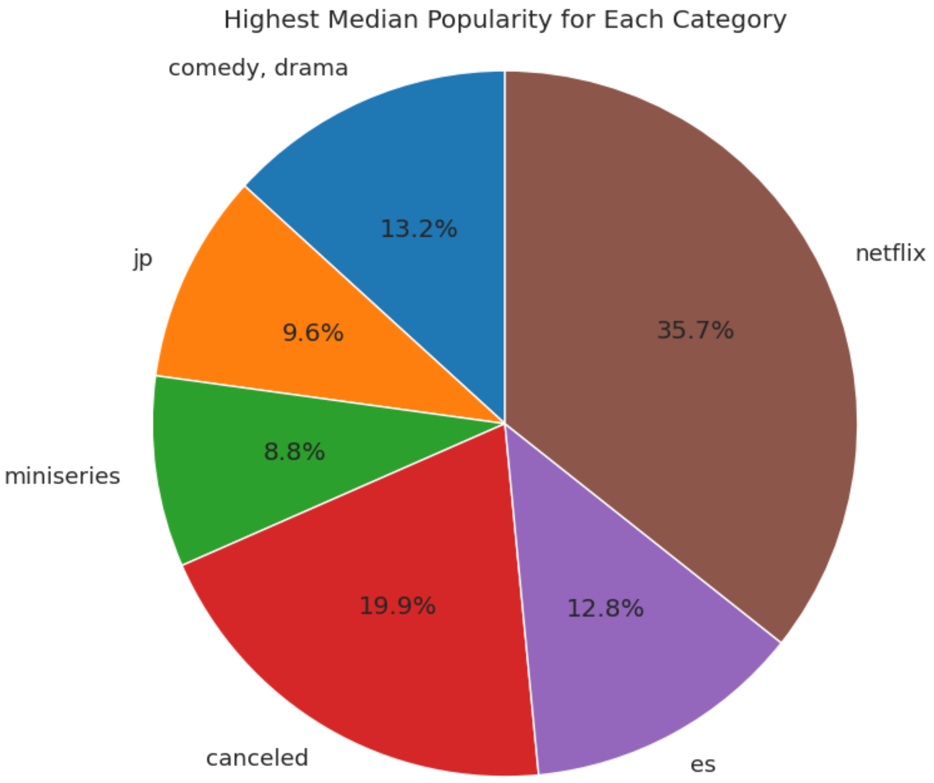


**Figure 5**: A. Analysis of Popularity Across Top 10 Categories for Each Categorical Feature: The Kruskal-Wallis test revealed significant differences in popularity across various categorical variables. Specifically, the results suggested that popularity varied significantly across each of these categories. B. The pie chart below presents the highest median popularity for each category.Each slice represents a category and the size of each slice corresponds to the highest median popularity for that category





### 3. Data Cleansing – Outliers and Missing Values:

Outliers in numerical columns, such as 'vote_count', 'number_of_seasons', 'languages_count', and 'network_count', were identified and handled accordingly. Outliers that were previously imputed with 0 were replaced using MICE (Multiple Imputation by Chained Equations), a more reliable imputation method.

For missing data in the 'last_air_date' column (1.9% missing), the 'in_production' column was leveraged. The approach identifies rows with missing 'last_air_date' values and fills them based on the 'in_production' status. If a show is still in production ('in_production=True'), the 'last_air_date' is set to the current date. For shows not in production, the missing values are left unchanged. This approach ensures that the 'last_air_date' column is appropriately populated while considering ongoing productions.

For missing values in categorical features:

- production_companies (55.9% missing) was dropped as it did not contribute significantly to the model and was found to reduce the $R^2$ score.
- origin_country (13.8% missing) was handled using Random Forest imputation, as earlier attempts using KNN and MICE proved less effective. Any remaining NaN values were imputed using MICE to ensure dataset completeness.

**4. Feature Engineering:**

- Filter the Data (01/01/2022 - 31/12/2023): After evaluating the model's performance, which resulted in a relatively low $R^2$ value, it was decided to narrow the dataset to focus on more recent data. The low $R^2$ suggested that the model wasn't capturing enough variation in the target variable (popularity). Focusing on recent data may provide more insight into current trends and behaviors.

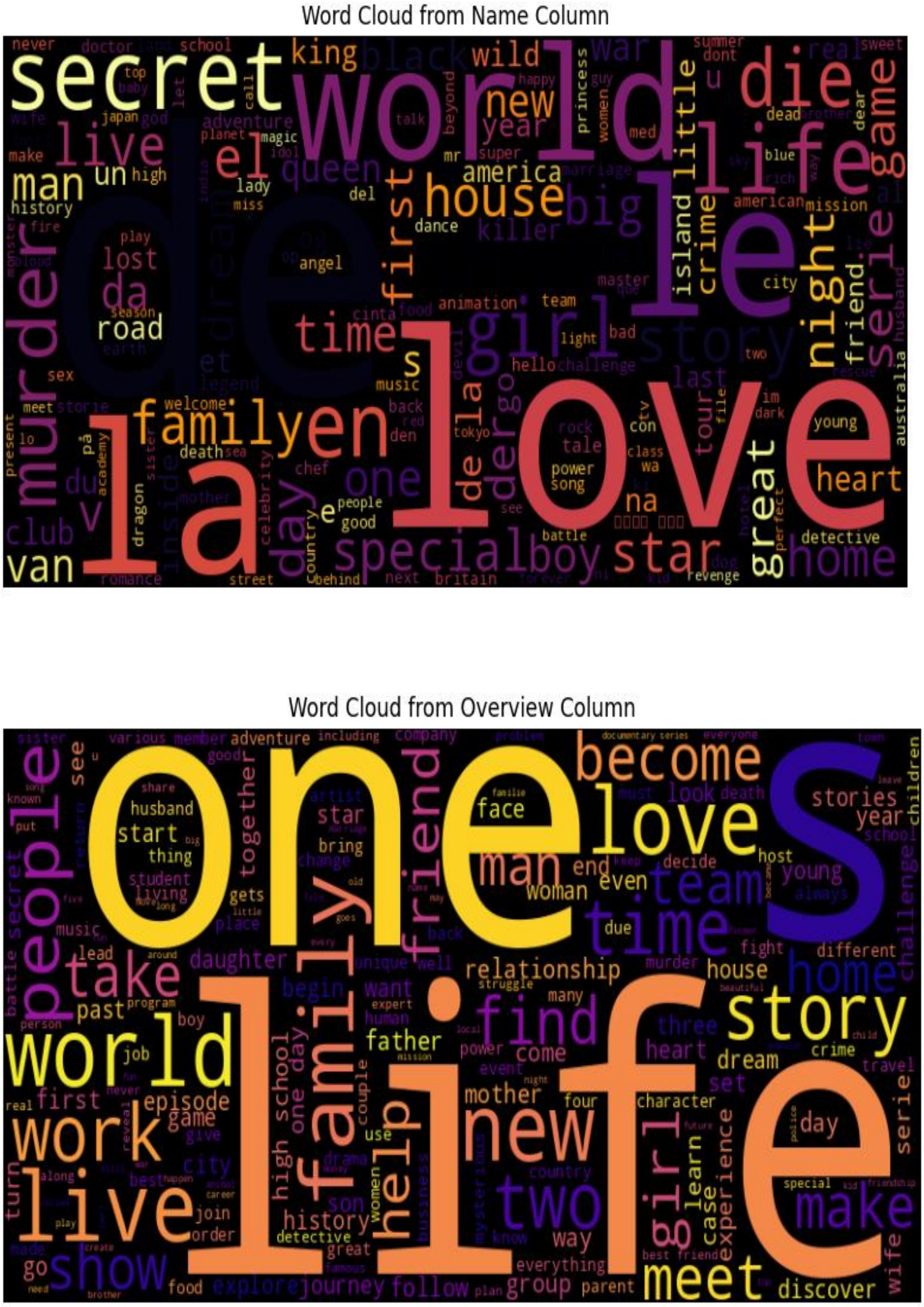
- New Feature Creation: Additional features were created to improve model performance. These included extracting the year from both the 'first_air_date' and 'last_air_date', calculating the duration between these air dates, and extracting year and month information. Additionally, the number of seasons was calculated based on the air dates.

- Handling the 'Overview' Column (Sentiment Analysis): The TextBlob library was used to analyze the sentiment of the 'overview' column. The polarity (sentiment) and subjectivity (personal opinion vs factual) scores were calculated for each overview and stored in a dictionary. These sentiment scores provided valuable insights into the tone and nature of each show’s description.

- Word Clouds for 'Name' and 'Overview' Columns: Word clouds were generated for both the 'name' and 'overview' columns to identify the most frequent words in each. This helped highlight key themes or topics frequently mentioned across the data. The top 5 most frequent words from each column were extracted for further analysis.

- Creating Dummies for Top 5 Words: Dummy variables were created for the top 5 most frequent words in the 'name' and 'overview' columns. These dummy variables were encoded as binary features (1 or 0) and added to the dataset. This encoding helped the model capture the influence of these prominent words, which may have an impact on predicting a show’s popularity.

**Figure 6**: A. Word cloud from 'Name' column; B. Word cloud from 'overview' column



Word Cloud from Name Column

Word Cloud from Overview Column

**Figure 7**A: .The sentiment plot  for 'overview' column**;** B. polarity and subjectivity

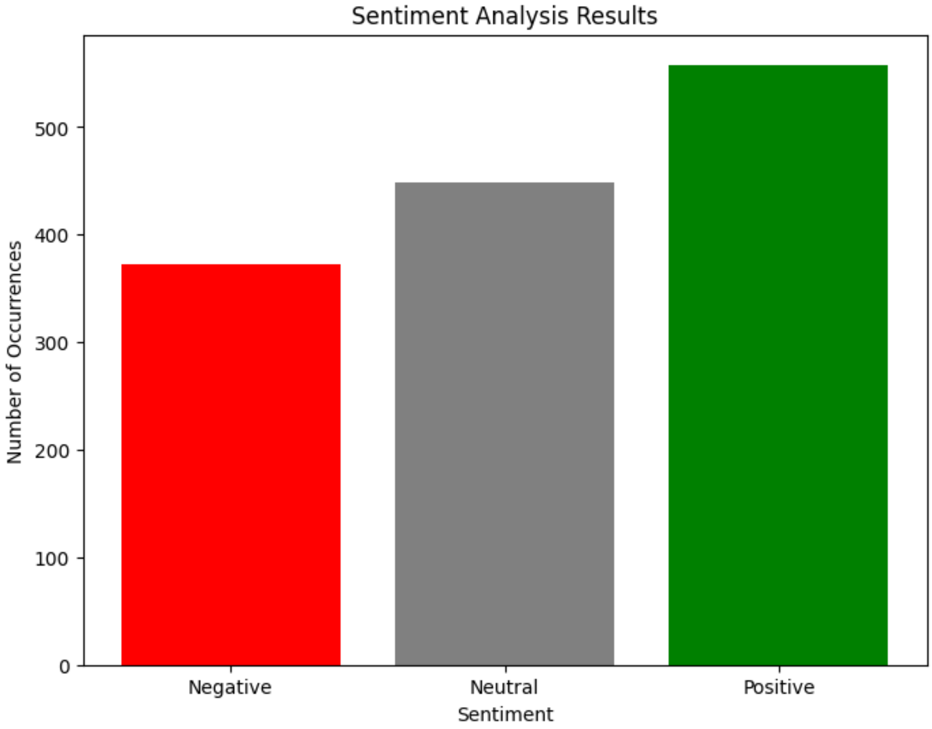


**Figure 7B:** The majority of data points show positive polarity and subjectivity greater than 0, indicating that most of the text reflects subjective opinions or emotions. Additionally, the higher proportion of positive polarity suggests that the content tends to have a more favorable or optimistic tone, with an overall positive sentiment expressed in the majority of the entries

### 5. Model Selection and Fine-tuning:

A variety of machine learning models were tested, including **Lasso**, **Random Forest Regressor**, **XGBoost**, and **Ridge**. These models were fine-tuned through hyperparameter optimization to identify the best-performing model for predicting TV show popularity. During the feature selection process, only the features that were selected by all four models (i.e., those with a selection sum ≥ 4) were retained. As a result, the number of features was reduced from 69 to 24, indicating that only 24 features met the selection criteria for further analysis.

The key features influencing the target variable, "popularity," include factors such as audience reception, platform, genre, timing, and language availability, all of which play significant roles in determining a show's popularity.

### 6. Model Evaluation:

Several regression models were trained to predict TV show popularity, including **Linear Regression**, **Decision Tree**, **Random Forest**, **AdaBoost**, **Gradient Boosting**

**(GBM)**, **SVM**, and **XGBoost**. The models were evaluated using four key metrics: MSE (Mean Squared Error),RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), RMSLE (Root Mean Squared Logarithmic Error), $R^2$ (Coefficient of Determination).

The evaluation results indicated that Random Forest performed the best overall, with the lowest MSE, RMSE, and MAE, as well as the highest $R^2$. On the other hand, SVM and Decision Tree showed poor performance. Both XGBoost and Gradient Boosting offered competitive performances but did not outperform Random Forest by a significant margin. Linear Regression and AdaBoost showed relatively weak performance.

**Hyperparameter Tuning:**

To further enhance the performance of the Random Forest and XGBoost models, RandomizedSearchCV and XGBRegressor were employed to search for the best hyperparameters. After performing hyperparameter tuning:

- **XGBoost** showed improvements in MAE and RMSE compared to the **Random Forest Regressor**, particularly in MAE (14.26% vs. 5.38%).
- The **Random Forest Regressor's** $R^2$ dropped slightly after optimization, whereas **XGBoost** demonstrated a slight improvement in $R^2$ (from 0.318 to 0.323).

Given these results, **XGBoost** appears to be a more promising model, especially after hyperparameter tuning, showing better overall performance. However, the model is not yet ready for production. Further refinement of hyperparameters, the use of cross-validation, and enhancements in feature engineering are expected to lead to additional improvements in model performance, bringing it closer to deployment.

## Project Summary: Deployment and Beneficiaries of Machine Learning

Currently, the model is not ready for full-scale deployment, as further refinement is required to ensure reliable and accurate predictions. Once the necessary adjustments are made, the model can be deployed as a predictive tool for streaming platforms, content creators, and distributors. It will forecast the popularity of upcoming or existing TV shows based on factors such as genre, language, and other relevant features. Once optimized, the model can be seamlessly integrated into content recommendation engines, or assist in production and acquisition decisions to support content strategy.

Who Can Use and Benefit from the Machine Learning Model?

- Streaming Platforms (e.g., Netflix, Hulu, Amazon Prime): Streaming platforms can utilize the model to predict which TV shows are likely to become popular. This will enable more informed content acquisition strategies and improved recommendations that better align with audience preferences.

- Content Creators and Studios: Producers and studios can leverage the model’s predictions to tailor upcoming TV shows in terms of genre, language, and themes, increasing the chances of market success by aligning with audience interests.
- Viewers: The model’s predictions can be used by streaming platforms to offer better, more personalized content recommendations, enriching the user experience and helping viewers discover content they are more likely to enjoy.

This machine learning system, once refined, will help stakeholders in the entertainment industry make data-driven decisions, resulting in more engaging content and an enhanced viewing experience for audiences.