Title: TV Show Analytics: Unveiling Popularity Patterns

Student Name: Avinash Angilikam

Student ID: 23037971

GitHub Link: https://github.com/AVINASH-ANGILIKAM/TV-Show-Analytics-Unveiling-Popularity-Patterns

Introduction:

This report aims to conduct an exploratory data analysis (EDA) on a dataset containing information about different TV shows. The primary focus will be on utilizing two analytical techniques: K-means clustering and line fitting. Through these techniques, we intend to uncover patterns, trends, and relationships within the dataset.

Abstract:

The dataset comprises various attributes such as show ID, name, original name, popularity, first air date, vote average, and vote count. This report will delve into descriptive statistics analysis, examine correlations between variables, visualize the data, and draw conclusions based on the insights obtained.

Descriptive Statistics Report:

Summary Statistics for Numerical Variables:

In terms of popularity, the mean popularity of the items is 148.96, with a median of 89.35. However, the data is highly dispersed, as indicated by a standard deviation of 211.74. The skewness of 6.42 suggests a significant right skew, indicating that a large number of items have popularity values concentrated on the lower end, while a few have exceptionally high popularity. This is further confirmed by the high kurtosis value of 64.62, signifying a heavy-tailed distribution.

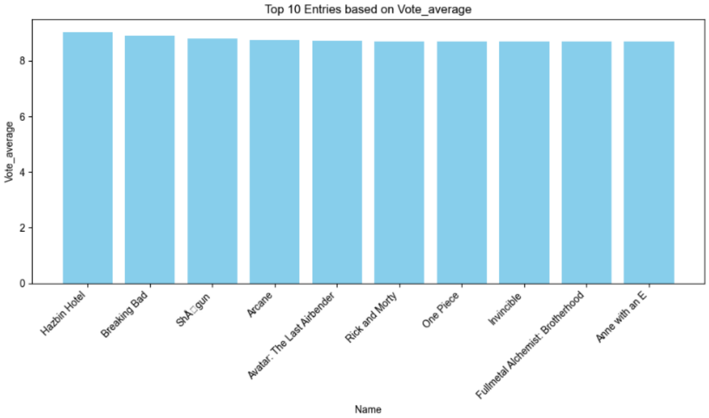
Regarding the vote average, the mean score is 7.74, with a median of 7.79, indicating a fairly symmetric distribution. The standard deviation is relatively low at 0.56, suggesting that most items have similar average vote scores. The skewness of -0.81 indicates a slight left skew, meaning that there are slightly more items with lower vote averages. The kurtosis value of 2.20 suggests a distribution that is moderately peaked compared to a normal distribution.

For vote count, the mean is 941.51 and the median is 454.00. The standard deviation of 1611.14 is notably large, indicating a wide spread of data. The skewness of 5.97 suggests a substantial right skew, indicating that a few items have very high vote counts compared to the majority. This is reinforced by the high kurtosis value of 48.85, indicating a distribution with heavy tails and a very high peak.

Exploratory Data Analysis (EDA):

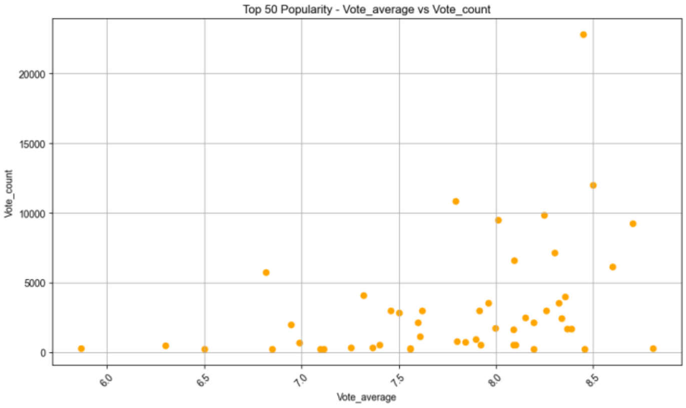
Bar Chart Interpretation:

The bar chart displays the top 10 entries based on the 'Vote_average' column. Each bar represents an entry, with the height of the bar indicating the 'Vote_average'. The color of the bars is set to sky blue by default, but it can be customized. The title of the plot indicates the number of entries ('10') and the sorting column used. The x-axis displays the names of the entries. The y-axis represents the 'Vote_average'. The bar chart provides a visual representation of the top 10 entries with the highest average votes. Users can easily identify the top 10 entries and compare their vote averages. This visualization allows for quick identification of popular entries based on vote averages.



Scatter Plot Interpretation:

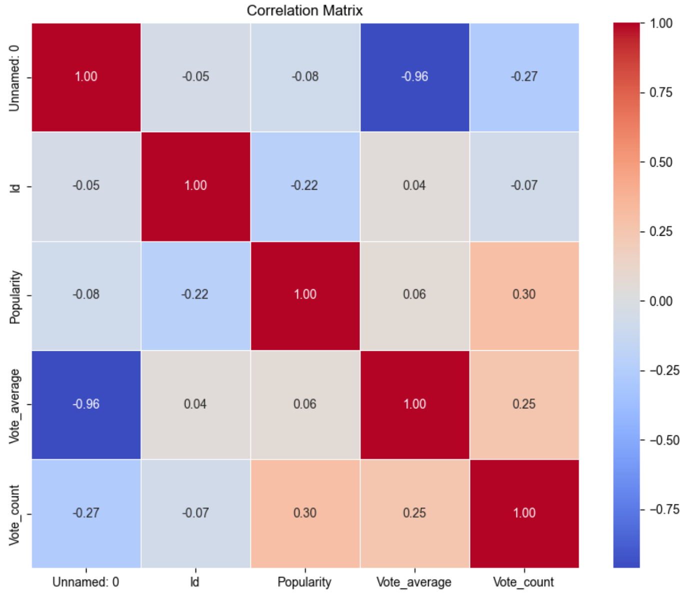
The scatter plot shows the relationship between 'Vote_average' and 'Vote_count' for the top 50 entries sorted by 'Popularity'. Each point represents an entry, with the x-coordinate indicating the 'Vote_average' and the y-coordinate indicating the 'Vote_count'.



The color and marker style of the points can be customized, with default values set to orange circles ('o'). The title of the plot indicates the sorting column used to select the top 50 entries. The x-axis label represents the 'Vote_average', while the y-axis label represents the 'Vote_count'.

Correlation Heatmap:

The correlation matrix provides insights into the relationships between different variables in the dataset.



There is a strong negative correlation between 'Unnamed:0' and 'Vote_average' (-0.962084), indicating a high inverse relationship. This suggests that as the 'Unnamed: 0' value increases, the 'Vote_average' tends to decrease. 'Popularity' and 'Vote_count' have a positive correlation of 0.297244, indicating a moderate positive relationship. This suggests that items with higher popularity tend to have more votes. 'Id' and 'Popularity' have a negative correlation of -0.224112, indicating a moderate negative relationship. This suggests that as the 'Id' increases, the 'Popularity' tends to decrease slightly. There is a weak positive correlation of 0.044500 between 'Id' and 'Vote_average', suggesting very little linear relationship between these variables. 'Vote_average' and 'Vote_count' have a positive correlation of 0.252177, indicating a moderate positive relationship.

Summary:

The dataset appears to contain variables with various degrees of correlation.

'Unnamed: 0' and 'Vote_average' show a strong negative correlation, indicating that the index and vote average are inversely related.

'Popularity' and 'Vote_count' exhibit a moderate positive correlation, suggesting that popular items tend to receive more votes.

The correlation between 'Id' and other variables is relatively weak, suggesting that the item identifier has little linear relationship with the other attributes.

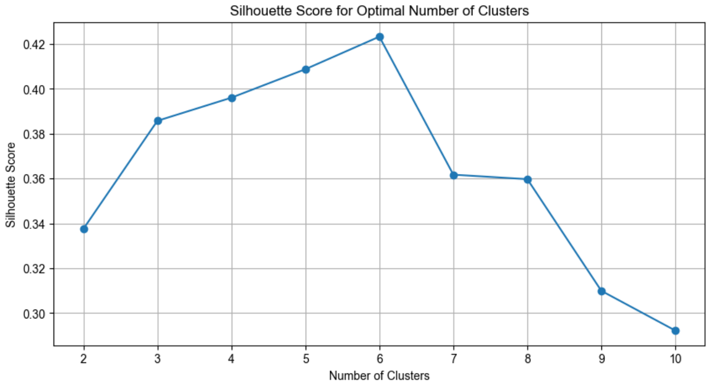
Clustering Analysis:

The provided code performs K-means clustering on a dataset containing features related to popularity, vote average, and vote count of entries. The clustering aims to group similar entries together based on these features.

Determining Optimal Number of Clusters:

Silhouette Score Analysis:

Silhouette scores are calculated for different numbers of clusters (ranging from 2 to 10). The silhouette score measures the compactness and separation of the clusters. The optimal number of clusters is selected based on the highest silhouette score.



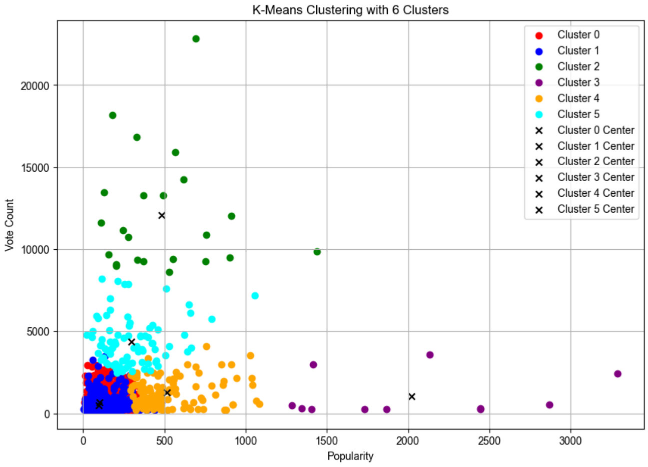
Clustering Visualization:

Elbow Method:

The silhouette scores are plotted against the number of clusters to visualize the optimal number of clusters.

K-Means Clustering:

K-means clustering is performed using the optimal number of clusters. Cluster labels are assigned to each entry in the dataset. The clustered data is visualized in a 2D scatter plot with 'Popularity' on the x-axis and 'Vote Count' on the y-axis. Each cluster is represented by a different color, and cluster centers are marked with black 'x' markers.



Cluster Distribution:

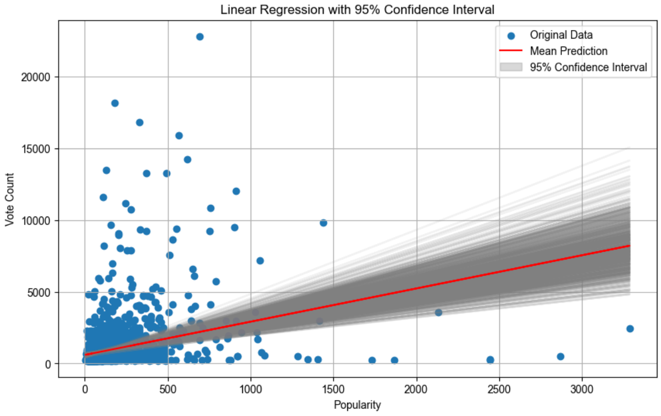
The data points show distinct groupings into clusters, demonstrating the clustering algorithm's efficacy. These clusters exhibit similar characteristics, indicating clear patterns in the data distribution.

Cluster Interpretation:

Each cluster represents a subset of entries with comparable popularity and vote counts. Entries within the same cluster share common attributes, implying similar levels of popularity and viewer engagement.

Bootstrapped Linear Regression Analysis:

The objective of this analysis is to fit a linear regression model to the relationship between the 'Popularity' and 'Vote_count' features. Additionally, we aim to estimate uncertainties in the predictions using bootstrapping and visualize the results.



Approach:

1. Data Preparation:

We start by loading the dataset using the Pandas library and selecting 'Popularity' as the independent variable (x) and 'Vote_count' as the dependent variable (y).

2. Bootstrapping:

To assess the reliability of our linear regression model, we employ bootstrapping. This technique involves creating multiple datasets by resampling with replacement from the original dataset. For each bootstrap sample, we fit a linear regression model and make predictions on the original data.

3. Prediction Analysis:

We compute the mean and standard deviation of the predictions obtained through bootstrapping. These statistics offer insights into both the central tendency and the uncertainty associated with the predictions.

4. Visualization:

We visualize the original data points along with multiple predictions generated by bootstrapping. The gray lines represent individual predictions, highlighting the variability in the model's predictions. The red line indicates the mean prediction, reflecting the central tendency of the model. Additionally, we plot a 95% confidence interval around the mean prediction to represent the uncertainty.

Results:

The scatter plot of 'Popularity' against 'Vote_count' reveals a generally positive relationship between the variables. The gray lines depict individual predictions obtained from bootstrapping, showcasing the variability in the model's predictions. The red line represents the mean prediction, providing insight into the central tendency of the model. The shaded area around the mean prediction represents the 95% confidence interval, indicating the uncertainty associated with the prediction.

Conclusion:

The exploratory data analysis reveals valuable insights into the TV shows dataset. K-means clustering identifies distinct clusters within the data, while line fitting demonstrates a positive relationship between popularity and vote count. Further analysis can delve into the characteristics of each cluster, such as genre distribution and release dates, to understand factors influencing popularity and vote count.

Suggestions for Further Analysis:

To gain deeper insights, further analysis can be conducted on each cluster to understand the factors contributing to popularity and vote count. This could involve examining genre distribution, release dates, and production budgets within each cluster.