Unsupervised Machine Learning Final Project

The grading will center around 5 main points:

* ~~Does the report include a section describing the data?~~
* ~~Does the report include a paragraph detailing the main objective(s) of this analysis?~~
* Does the report include a section with variations of Unsupervised Learning models and specifies which one is the model that best suits the main objective(s) of this analysis?
* Does the report include a clear and well presented section with key findings related to the main objective(s) of the analysis?
* Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different modeling techniques?

Sections required in your report:

* ~~Main objective of the analysis that also specifies whether your model will be focused on clustering or dimensionality reduction and the benefits that your analysis brings to the business or stakeholders of this data.~~
* ~~Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.~~
* ~~Brief summary of data exploration and actions taken for data cleaning or feature engineering.~~
* Summary of training at least three variations of the unsupervised model you selected. For example, you can use different clustering techniques or different hyperparameters.
* A paragraph explaining which of your Unsupervised Learning models you recommend as a final model that best fits your needs in terms.
* Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise.
* Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.
* A strong business problem statement might read as follows:

**Clustering Analysis of Highly Ranked IMDB Television Shows**

**Business Problem Statement:**

In an increasingly competitive market, identifying the driving factors behind a successful television show is crucial for content creators, streaming platforms, and advertisers. Analysis of the top 250 IMDB-rated television shows will help identify the key features that distinguish high-performing content. This insight can guide future investment, content development, and marketing decisions to promote enhanced viewer engagement.

**Objectives:**

* Uncover underlying data structure through linear and nonlinear dimensionality reduction of the IMDB dataset to reveal key interdependencies among critical features (year, episodes, rating and description keywords) that drive content performance.
* Apply clustering analysis to these significant dimensions to segment the data into distinct groups, enabling stakeholders to pinpoint target content segments, optimize recommendation systems, and support strategic decision-making.

**Dataset Description:**

The dataset consists of the top 250 IMDB-rated television shows sourced from Kaggle (<https://www.kaggle.com/datasets/khushipitroda/imdb-top-250-tv-shows>). It includes a mix of categorical and numerical attributes such as:

* **Name:** Title of the television show.
* **Year:** Production year (or range) indicating when the show aired.
* **Episodes:** Number of episodes, which is encoded as a numeric value after removing extraneous text.
* **Type:** Content rating/category (e.g., TV-MA, PG-13) that has been re-mapped to consolidate similar groups.
* **Rating:** Viewer rating from IMDB, serving as a proxy for audience reception.
* **Description:** Textual synopsis of the show, used for additional textual analysis.
* **Additional Attributes:** Such as image source and hyperlink data, which have been dropped for our focused analysis.

**Analysis Outline:**

1. **Data Preprocessing:**

* Clean and transform the data (handle missing values in ‘Type,’ convert year ranges to the first year the show aired, convert episode list to integers).
* Scale numerical features using MinMax scaling prior to dimensionality reduction and clustering.

1. **Dimensionality Reduction:**

* Apply both linear (PCA) and non-linear (Kernel PCA) techniques to uncover latent relationships among the key features.
* Explore how these methods reveal interdependencies in the dataset, especially focusing on features like Year, Episodes, and Rating.

1. **Clustering Analysis:**

* Use clustering algorithms (KMeans and DBSCAN) on the transformed data to identify distinct segments within the top-rated shows.
* Evaluate cluster quality using metrics such as silhouette scores and analyze the resulting cluster centroids to determine key differentiating features.

1. **Interpretation & Business Implications:**

* Compare the clusters by examining original feature distributions (using box plots and group summaries).
* Extract actionable insights that can inform content development, marketing, and investment strategies.

Exploratory Data Analysis:

Below is an example summary for the EDA section that highlights both the data cleaning/feature engineering steps and the keyword analysis via a word cloud:

**Exploratory Data Analysis (EDA)**

During the EDA phase, the following steps were taken:

* **Data Cleaning and Feature Engineering:**
* **Missing Values & Data Consistency:**  
  Addressed missing values in the 'Type' column by imputing or creating a new category, and standardized 'Year' values by extracting the starting year from range entries.
* **Data Transformation:**  
  Cleaned the 'Episodes' column by removing extraneous text (e.g., "eps") and converting it into a numeric format.
* **Scaling:**  
  Applied Min/Max scaling to numerical features (Year, Episodes, and Rating) to ensure uniformity prior to dimensionality reduction and clustering analyses.
* **Keyword Analysis via Text Mining:**
  + **TF-IDF Vectorization:**  
    Processed the 'Description' text using TF-IDF to quantify the importance of terms across the dataset.
  + **Word Cloud Generation:**  
    Aggregated the TF-IDF scores to create a word cloud that visually emphasizes the key terms present in the television show descriptions, thereby providing insights into thematic patterns in the data.

A close up of words

AI-generated content may be incorrect.

These steps not only helped to prepare the dataset for advanced modeling (dimensionality reduction and clustering) but also provided a deeper understanding of the content's underlying themes and structures, which are critical for deriving actionable business insights.

**Model Training Summary**

To uncover the natural groupings in the top 250 IMDB-rated television shows dataset, we experimented with several unsupervised clustering variations:

1. **KMeans on PCA-Reduced Data:**
   * *Approach:*  
     We first applied linear PCA to reduce the dimensionality of the scaled numerical features (Year, Episodes, Rating), retaining two principal components that captured approximately 58% of the variance.
   * *Clustering:*  
     KMeans clustering was then performed on the PCA-transformed data. Using the elbow method and silhouette scores, we determined that three clusters provided a good balance between compactness and separation.
   * *Insights:*  
     This approach allowed us to interpret the primary variance directions in the data and relate the cluster centers back to the original features, offering straightforward interpretability.
2. **DBSCAN on Min/Max Scaled Data:**
   * *Approach:*  
     The dataset was scaled using MinMaxScaler, and DBSCAN was applied directly on the scaled features.
   * *Parameter Tuning:*  
     We experimented with a range of eps values (guided by k-distance plots and silhouette scores) and settled on eps = 0.07 with min\_samples = 5. Initial results yielded four clusters; however, by excluding noise (points labeled as -1), the effective number of clusters was reduced to three.
   * *Insights:*  
     The DBSCAN method’s ability to detect outliers provided additional insights into irregular data points and helped ensure that our clusters represented dense regions of similar shows.
3. **Kernel PCA Followed by KMeans Clustering:**
   * *Approach:*  
     To capture non-linear relationships, we applied Kernel PCA using an RBF kernel with gamma = 5 and set n\_components = 5. This transformation mapped the original features into a new space where complex interactions could be more clearly discerned.
   * *Clustering:*  
     KMeans was then used on the Kernel PCA-transformed data, and we again identified three distinct clusters.
   * *Insights:*  
     This combination provided a more nuanced view of the data’s structure, capturing subtle patterns that were not apparent with linear techniques. Hyperparameter tuning (such as varying gamma) further refined the clustering outcomes.

**Evaluation and Selection:**  
Each clustering approach was evaluated using metrics like the silhouette score and visual inspection of cluster distributions. The consistency across methods—particularly the emergence of three clusters—strengthened the validity of our segmentation. This multi-model strategy not only cross-validates our findings but also offers different perspectives on the underlying factors driving television show success.

This summary addresses the training of different model variations and explains how each approach contributes to understanding the data and guiding strategic decisions.

**Model Evaluation:**

Based on our experiments with multiple unsupervised models, we recommend the Kernel PCA followed by KMeans clustering approach as our final model. This combination best meets our needs by capturing the non-linear relationships among key features—Year, Episodes, and Rating—that drive show performance, while robustly segmenting the dataset into three distinct clusters. The non-linear transformation of Kernel PCA provides a richer representation of the underlying structure, enabling KMeans to identify meaningful groups that align with our business objective of uncovering actionable insights for content development, targeted marketing, and strategic investments. This approach not only reveals the complex interdependencies in the data but also offers a clear, interpretable framework for guiding decision-making in a competitive media landscape.

Summary of Key Findings:

Next Steps:

A chart of a number of blue boxes

AI-generated content may be incorrect.

A chart with many colored dots

AI-generated content may be incorrect.

Cluster Year Episodes keyword\_score

0 1993.833333 151.555556 0.000000

1 2012.425806 38.522581 0.004326

2 2011.608696 63.000000 0.172317

Introduction of Rating into PCA resulted in nonlinearity. Kernel-PCA (n\_components = 5, gamma = 5) was used and got 3 clusters again:

A diagram of a data

AI-generated content may be incorrect.

Year: 0.117

keyword\_score: 0.253

Rating: 0.630