TV SHOW POPULARITY

BY ANDREW DELIS
DECEMBER 15, 2024

OVERVIEW

The top streaming companies make and release hundreds of their own titles per year. This means that the business model is reliant on generating as many good ideas for movies and shows as quickly as possible. Not every good idea is good enough to get made, yet they still make and release a baffling number of shows and movies per year. Sifting through thousands of ideas a year obviously takes a lot of time and money. I do not work in the TV industry, but I am doing this analysis in hopes of identifying what types of titles are the most popular with viewers this year.

GOALS

1. Identify the most influential variables on 2024 releases' TMDB popularity score
2. Make a recommendation based on the results of my analysis
3. Explore next steps for further analysis

DATA ANALYSIS TECHNIQUES

- Penalized Regression
- Partition Clustering
- Hierarchical Clustering
- PCA
- PCA + Partition Clustering
- PCA + Hierarchical Clustering

NOTES ON DATA ANALYSIS TECHNIQUES

Penalized Regression

I started off my analysis by running a penalized regression on my data to predict the popularity of a TV show. I did this for a number of reasons. First, my data is highly dimensional, so the dimensionality reduction features of lasso were very appealing. Second, just by reading the names of the variables it seemed like there were a lot of confounding variables. Both lasso and ridge would help me identify which variables were confounding.

I ran three iterations of both a lasso and ridge regression model.

1. Initial run contained all the variables
2. The second iteration did not contain obviously confounding variables
3. The third iteration did not contain outliers

I gained two key things from running these models.

1. Confounding Variable Identification
2. Influential Variable Overview.

As expected, there were many confounding variables. These first models allowed me to identify those and remove them from the data I used in my subsequent clustering analysis. Additionally, these models showed that country of origin and genre (specifically soap operas) were clearly important to the popularity of a show.

Removing outliers was also useful to my analysis. When removing outliers the most influential variables changed, but genre and country of origin were still at the top. Additionally, when soap operas and other outlier shows are removed from the data, adult content actually reduces the popularity of a show fairly significantly.

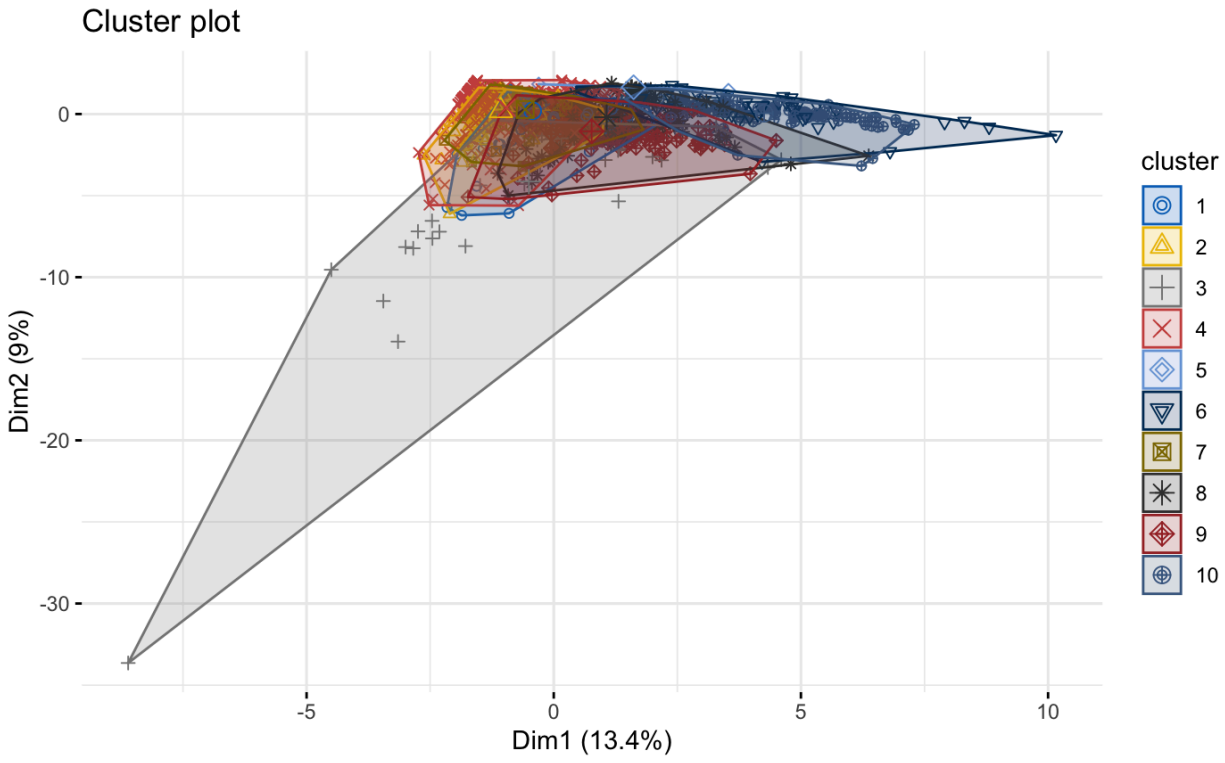
Partition Clustering & Hierarchical Clustering

For both my partition clustering (k-means) and hierarchical clustering (ward) models, I calculated the optimal number of clusters using both the silhouette and elbow methods. The optimal number of clusters was four and two respectively.

Initially, I just ran the models with two, four, and six clusters. Eventually I also ran both my k-means and my hierarchical clustering models with three and ten clusters too due to the results of my hierarchical clustering analysis and the silhouette graph of the PCA data.

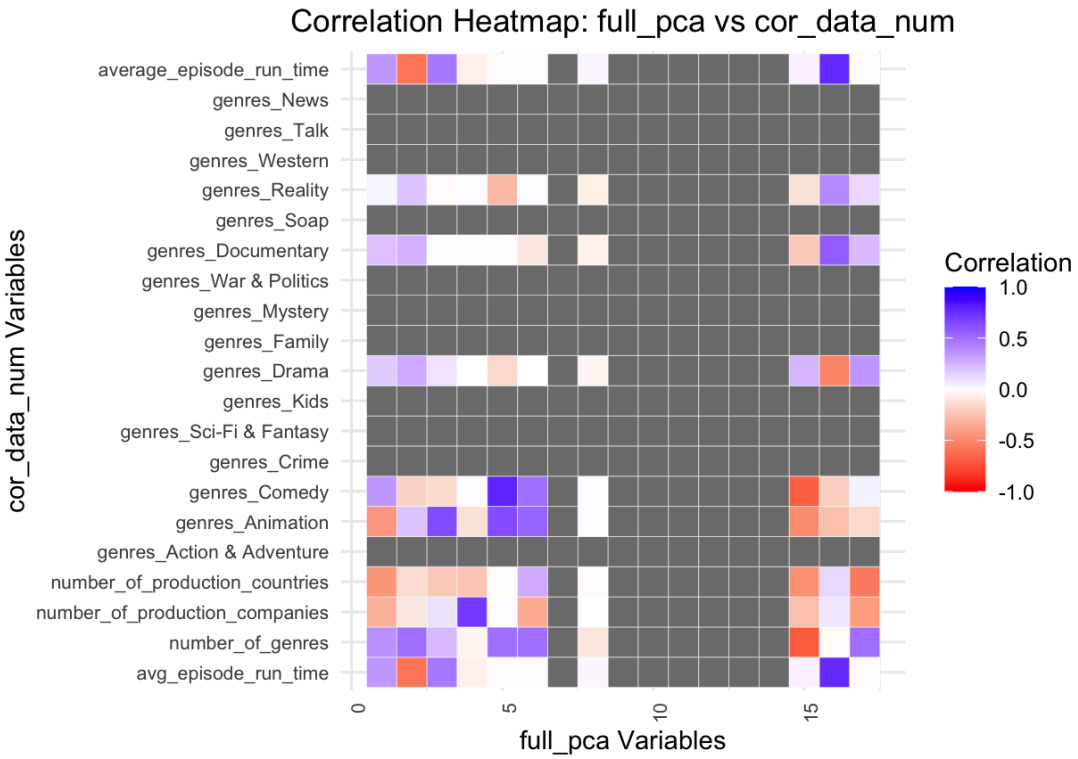
After running the four partition clustering models I created a table that contained the cluster that had the highest popularity coefficient. Interestingly the models with two, three, four, and six clusters had very similar coefficient values. However, the model with 10 clusters actually had the highest popularity coefficient and had totally different coefficient values. Despite this genre was still very important.

The clustering models did not visually show any clear clusters. Here is an example:



PCA

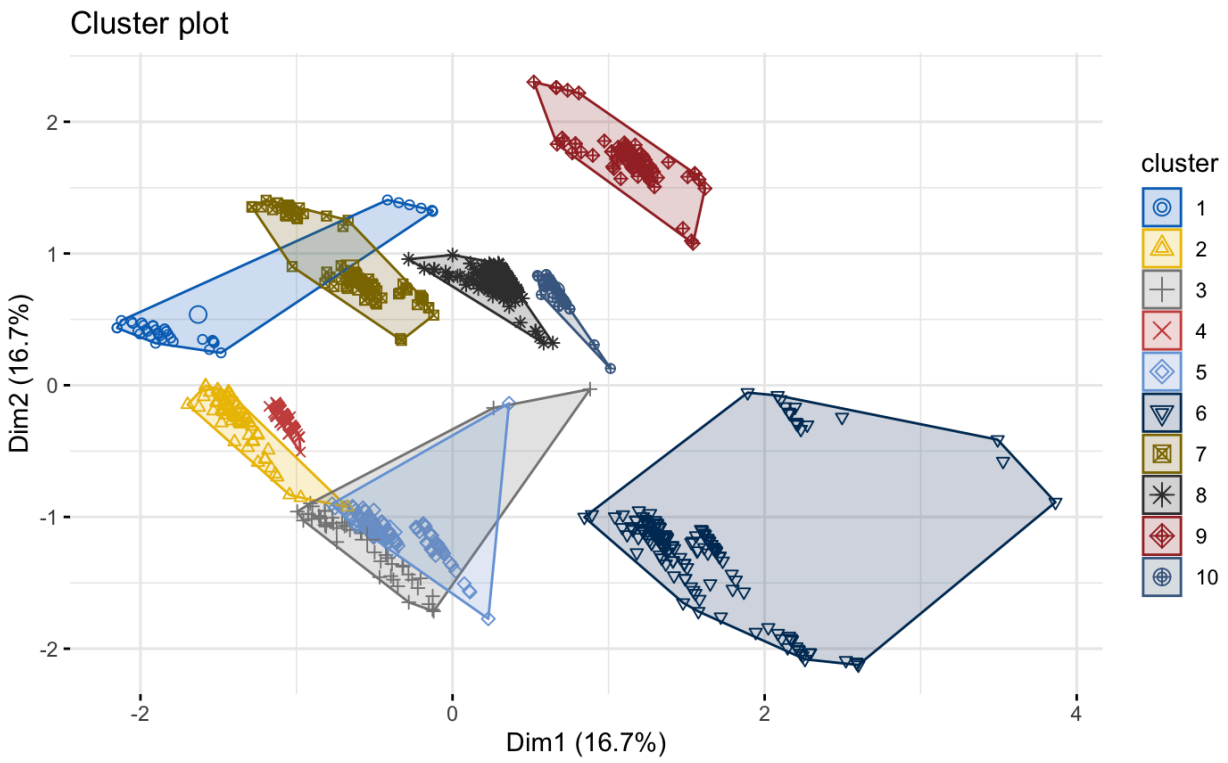
After finishing my clustering models I did PCA on my data and created a heat map correlation matrix of the variables that went into the PCA calculation and the principal components themselves.

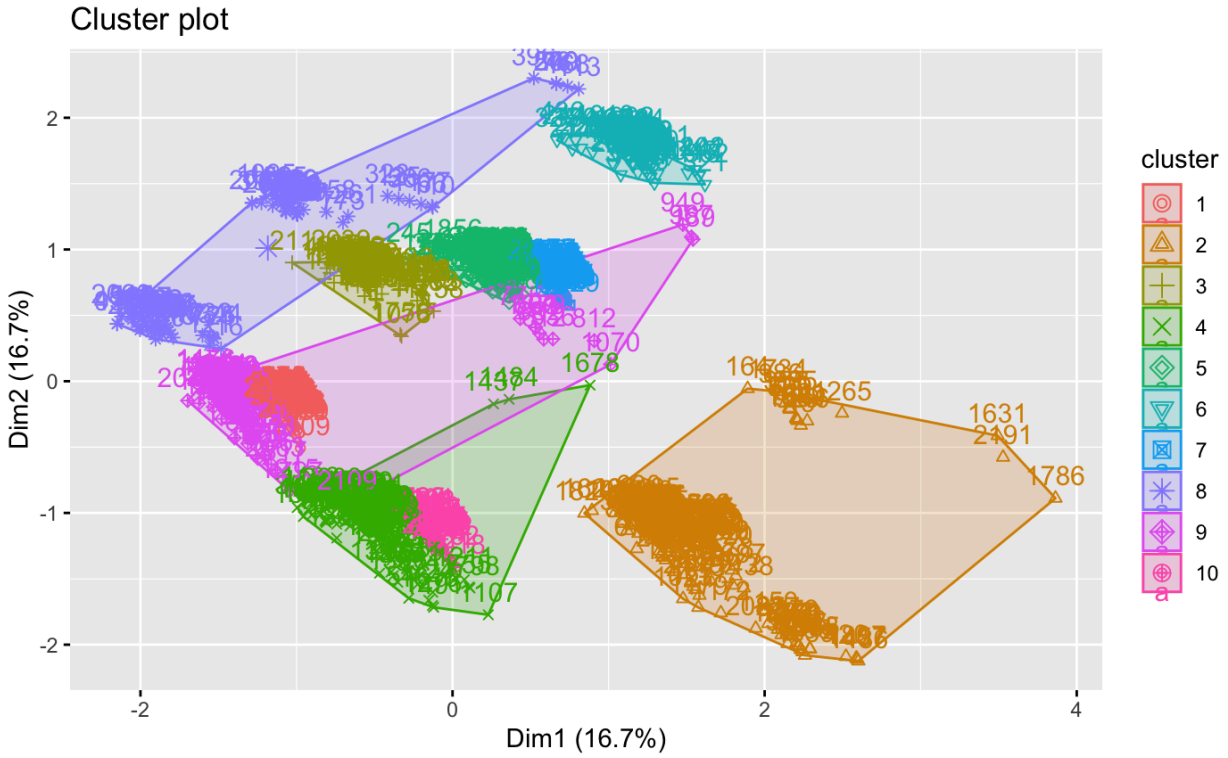


The first five principal components account for 81.8% of the variation in the data, but the first two principal components only account for 44% of that. There is no strong correlation between the first two components and a single variable, but notably many of the correlated variables are genre variables.

PCA + Partition Clustering & PCA + Hierarchical Clustering

Finally, I ran the clustering models on my PCA data. As I alluded to earlier, the results of the silhouette graphs on the PCA data showed that a much higher number of clusters was possibly optimal. The graphs of the models visually support this too. The possible clusters are much more distinct.


Cluster plot
Dim2 (16.7%)
Dim1 (16.7%)
cluster
1
2
3
4
5
6
7
8
9
10


Cluster plot
Dim2 (16.7%)
Dim1 (16.7%)
cluster
1
2
3
4
5
6
7
8
9
10

This suggests that there are ways to segment the data even further than I already have. These finer clusters may even tend to have higher popularity scores than any cluster I have seen so far. Unfortunately I did not think to include popularity back into the PCA data, so that will have to be one of this analysis next steps.

CONCLUSION

Based on my full analysis, I recommend that streaming service be very mindful of genre. Based on my penalized regression models, I recommend that they start picking up any soap operas not made in the US. If they do start shooting something, make sure that children can go see it. The ridge model showed that adult shows do not do well, and the clusters with the highest popularity showed tended to have higher coefficients for the ‘kids’ genre or genres that they stereotypically like.

NEXT STEPS

My clustering analysis did not include some of the influential variables from the penalized regression analysis due to them being factor variables (one-hot-encoding would have given my data even higher dimensionality and skyrocketed run time). Seeing if I can find a way to add those back in is something that I plan on doing next.

I only used ward linkage for my hierarchical clustering models. Testing other linkages as a robustness check is necessary for those models.

Adding popularity back into the PCA analysis to see how things change is something I want to do in the future.