STAT 430: Final Project

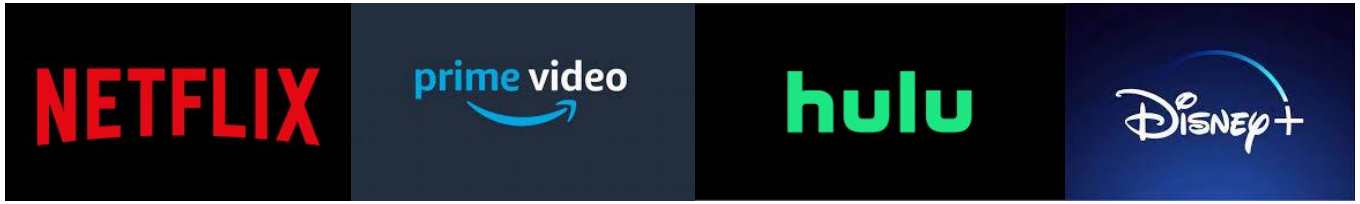
By: Atharv Pathak and W. Jonas Reger

Introduction

- Motivation
    - We wanted to work with a clusterable dataset that is associated with a well-known company or product, and something that we use ourselves in our personal lives.
    - Initial interests in analyzing Netflix data.

Dataset Research

- Dataset Information
  - We found a dataset on Kaggle called “Movies on Netflix, Prime Video, Hulu and Disney+”.
  - We chose this one since it included data across multiple streaming platforms.



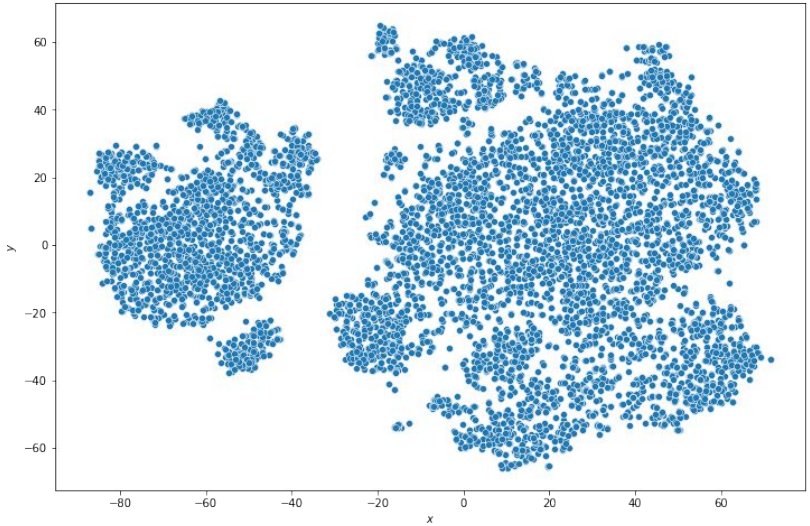
- 16744 observations
- 17 attributes
  - Row index, movie ID, Type
  - Title, Year, Age rating, IMDb rating, Rotten Tomatoes rating
  - Netflix, Hulu, Prime Video, Disney+, Directors, Genres, Country, Language, Runtime

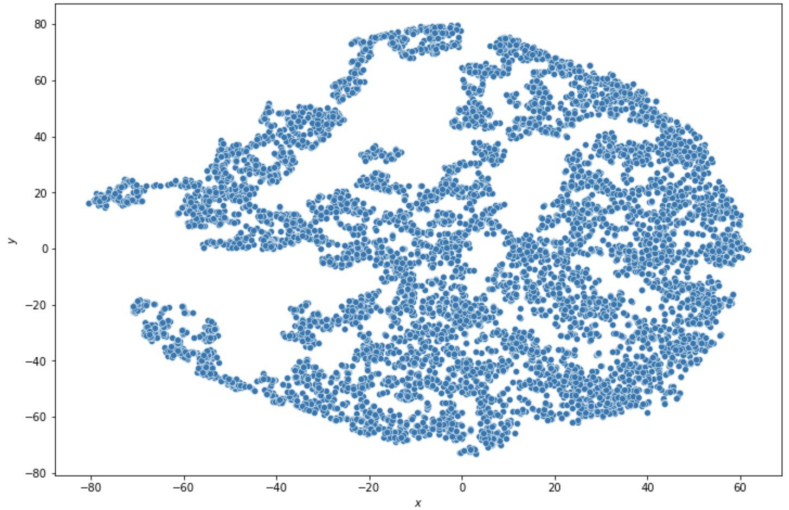
Exploratory Data Analysis

- There is a lot of missing information.
  - Removed missing information
  - Text processing to separate categorical levels of attributes
  - Binned Language and Country into Continental/Regional groups
  - After cleaning the data we found that:
    - Prime Video accounts for 67.6% of the data (Larger sample)
    - Netflix accounts for 22.8% of the data
    - Hulu accounts for 8.0% of the data
    - Disney+ accounts for 6.9% of the data
  - We found a few strong relationships that were expected
    - E.g. Disney+ is positively correlated with Family genre movies.

Pre-Analysis Questions

- Is the data clusterable?
  - Yes, all versions of the modified dataset has a Hopkins statistics between 0.04 and 0.16.
  - T-SNE plots also show evidence of clustering structure
    - Separation and cohesion is not very ideal



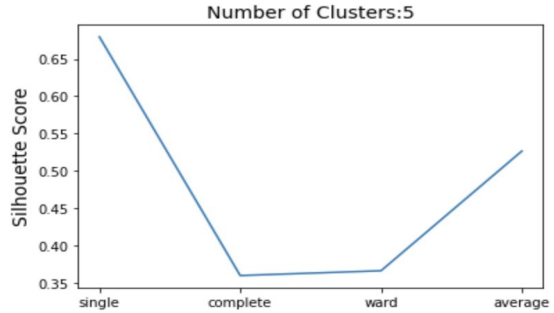


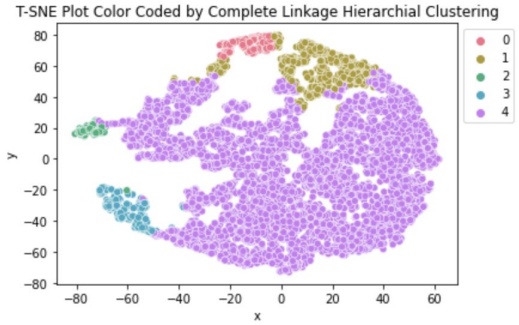
Algorithm Selection Motivation

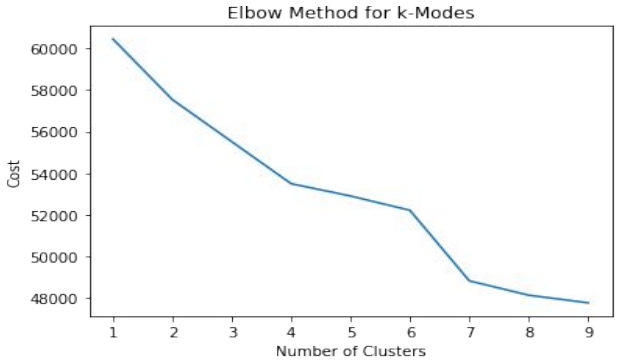
- Since the modified dataset is mostly categorical, we binned numerical attributes and used categorical clustering analysis.
- Algorithms:
  - Agglomerative Hierarchical Clustering Algorithm
    - Works with categorical dummy variables (especially when standardized)
  - KModes Clustering Algorithm
    - Works well with categorical data

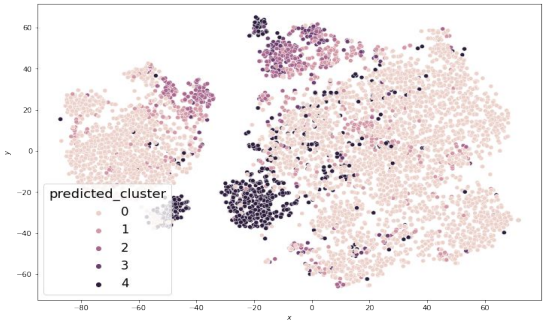
Algorithm Results

- Selected 5 clusters for both algorithms



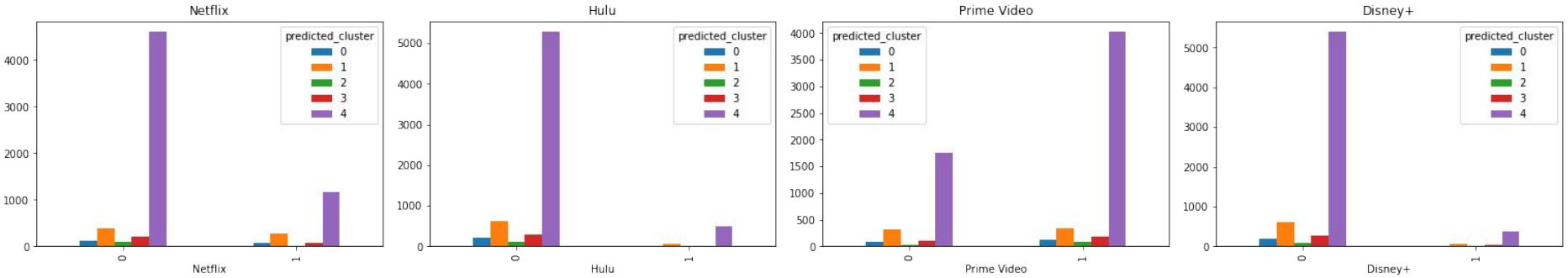






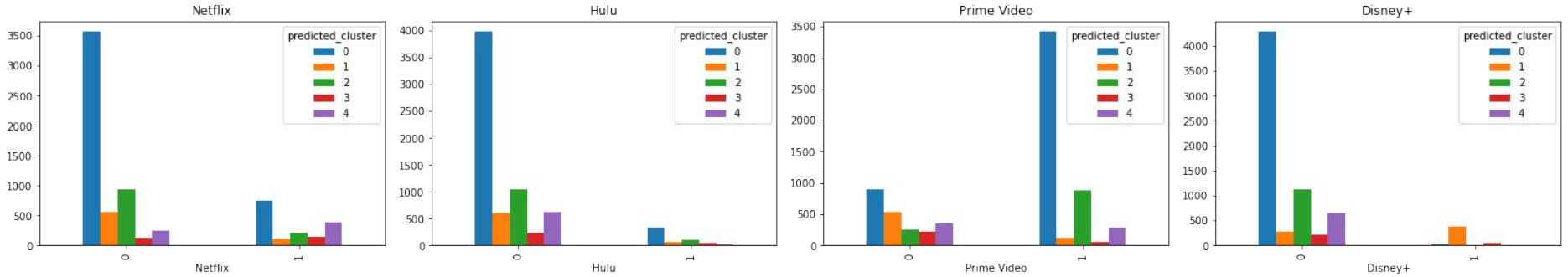
Post Analysis Questions

- Agglomerative Hierarchical predicted cluster (some) characteristics
  - Cluster 0 - Netflix & Prime Video, Young/Adolescent, Action, Drama & Romance, Shorter, Asia
  - Cluster 1 - Netflix, Hulu & Prime Video, Adolescent/Adult, Drama, Shorter, N. America & Asia
  - Cluster 2 - Prime Video & Disney+, Young, Western & Comedy, Shorter, N. America, Older movies
  - **Cluster 3** - Prime Video, Adult & All ages, Comedy & Documentary, Shorter, N. America, Newer movies
  - Cluster 4 - Hulu & Prime Video, Adult, Thriller & Horror, Longer, N. America & Europe



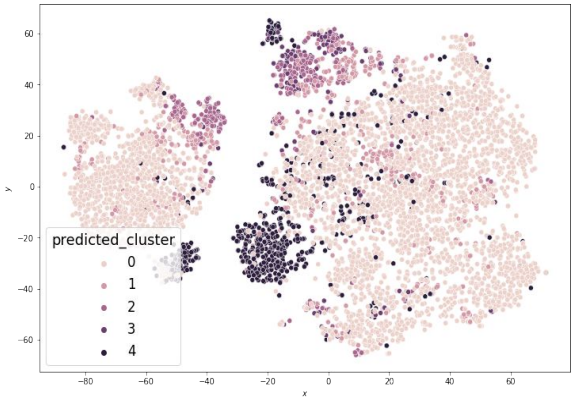
Post Analysis Questions

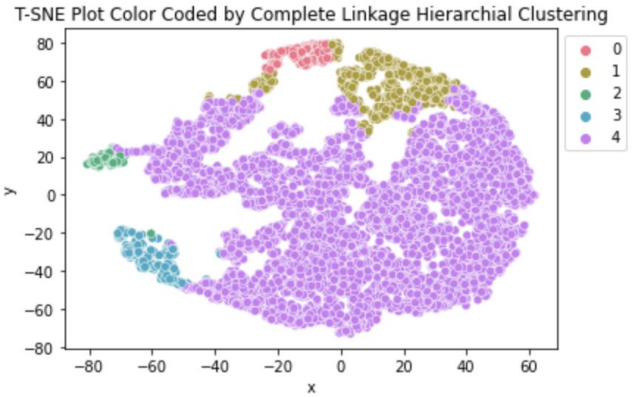
- KModes predicted cluster (some) characteristics
  - Cluster 0 - Prime Video, Adult, Horror, Longer, N. America
  - Cluster 1 - Netflix, Hulu & Disney+, Young, Comedy, Shorter, N. America
  - **Cluster 2** - Disney+, Young, Adventure, Animation & Family, Shorter, N. America
  - **Cluster 3** - Netflix, Adolescent, Action & Drama, Shorter, Newer Movies, Asia
  - Cluster 4 - Hulu & Prime Video, Adolescent/Adult, Thriller & Drama, Longer, N. America & Europe



Analysis Summary

- KModes clustered the dataset better than the Agglomerative Hierarchical Clustering Algorithm
- NMF or Spectral Clustering might work better on this dataset, or the data could be simplified a bit more (i.e. some attribute levels had few observations and could be binned into an “Other” level).





Conclusion

- While KModes Clustering Algorithm didn't cluster the data as well as hoped according to the T-SNE plots, the clustering characteristics did match with previous assumptions and expectations of what viewers might be interested in.
    - E.g. Comedies would appeal more to younger audiences while Horrors would appeal to older audiences.
    - E.g. Cluster 2 is the "Disney+" cluster where characteristics are as expected for Disney films and viewers.
- Agglomerative Hierarchical Clustering struggled with clustering the processed dataset.
    - E.g. Cluster 3 had "conflicting" results with both adult and younger audiences
    - E.g. Other clusters were muddled in terms of streaming platforms, but had good insights in terms of genre and other categories

References

1. Aggarwal C.C. (2016) An Introduction to Recommender Systems. In: Recommender Systems. Springer, Cham. https://doi.org/10.1007/978-3-319-29659-3_1

2. Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction. Journal of Marketing, 74(1), 108–121. https://doi.org/10.1509/jmkg.74.1.108

3. Technophilo. (2012, October). Recommender Systems: Pros and Cons. Technophilo. http://www.technophilo.in/2012/10/recommender-systems-pros-and-cons.html