STAT 3355 Project Proposal

Dataset Information:
https://www.kaggle.com/datasets/shivamb/netflix-shows/data
https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

Background and Motivation / Rationale:
We aim to explore Netflix's dataset of shows and movies to gain insights into viewing habits, popular genres, and content distribution across countries. This dataset spans from 1925 to 2021 and includes various entries and variables such as titles, descriptions, imbd ratings, and categories. This will help us understand the relationship between the content features and their popularity on Netflix. The resulting data will be useful for Netflix to make data-driven decisions about content acquisition and recommendations, helping cater better to global user preferences. We will focus on text mining techniques to analyze descriptions and genre tags and conduct a study to identify key factors linked with the popularity of shows and movies. This project can assist in predicting future content success.

Questions:
MAIN QUESTION: How can the key factors identified in our analysis of the Netflix dataset be leveraged to develop personalized show and movie recommendation strategies for customers?

(WHAT)
- What factors correlate the most with the popularity of shows and movies over time?
- What quantitative data can we use to predict whether a customer will like a show?
- What themes or elements in show/movie descriptions are most commonly associated with high viewer ratings?

(HOW)
- How do reviews (out of 10) impact relevance?
- How do social trends (popular shows) affect impact relevance?
- How do factors such as casting, director reputation, or production quality influence customer preferences and ratings?

(TEXT-MINING)
- What unstructured data can we use such as reviews to predict a customer's reaction to a show?
- Can we use text mining based on the comments/reviews left on the shows to find similar shows?
- How can we use the shows/movie descriptions to find similar films to recommend?

(WHEN)
- How does time of year affect movie/show preferences?

(WHERE)
- Are customers likely to watch shows made in a particular country/region?

We decided on these questions because they give us a broad but achievable scope of data we can use to answer our questions. To find out what movies a customer may like, it's important to think of ways we can use easily available data such as ratings and popularity trends, as well as find creative ways to use unstructured data such as movie descriptions to effectively take advantage of all the data available to us.