Capstone Project 2: Milestone Report 2

1. The submission demonstrates understanding of how to describe a **problem statement**.

Problem statement: Why it's a useful question to answer and for whom (source this from your proposal)

With so many new options available for customers to watch and stream their favorite movies and TV shows it is key for companies to be able to stand out by providing the pertinent content for all their customers. Being able to accurately recommend new movies and TV shows for individual users keeps the user from venturing to another provider. Being able to predict what combination of movie or TV show features, such as cast, director, genre will give the company insight to what additional content would be best to add to their library.

2. The submission demonstrates an understanding of how to describe a dataset and detail how it was **collected**, **cleaned**, and **wrangled**.

I have acquired datasets from Rotten Tomatoes, IMDb, Amazon, Netflix, and MovieLens. These datasets are vast enough to cover the scope of this project. They also allow me to break up the project into two possible routes; one focusing on a content level prediction and the other to focus on a user level prediction.

The data can be separated to accommodate being applied to either the content or user level prediction. For the content based recommendation system I will utilize the datasets from Rotten Tomatoes, IMDb, Amazon, and Netflix. These datasets are more geared towards the content. They provide information regarding movies available on different platforms, their genres, release dates, directors, cast, and ratings. The ratings given are not associated with individual users, so this doesn't allow for creating a prediction model for recommending a movie or TV show to an individual user. What it

does allow us to do is create models to predict how the content would be rated based on it's features. This helps to solve the problem of determining what currently existing content should be added to the platform. It can also provide a type of blueprint to develop new content around by predicting the combination of features that produce the best ratings.

The IMDb dataset allows for a more complex recommendation system. The dataset contains information on users, broken down by gender, age group, and location (us or non-us). From this dataset we can develop a recommendation model on the demographic level to predict how a certain demographic would rate a movie or TV show that they have not yet watched. This helps to solve relatively the same problem as above, but on the individual level. It allows a company like Netflix to determine what movies or TV shows to recommend that an individual user from a certain demographic should watch next. In doing so creating higher customer satisfaction and increasing customer retention.

3. The submission demonstrates successful application of **exploratory data analysis** (visualization and inferential statistics), for example histograms, scatter plots and hypothesis testing appropriately. And successful application of **machine learning.**

4. The submission demonstrates successful application of **data storytelling** techniques to articulate hypotheses and inferences, appropriate to their target audience.