For assignment 2, I loaded up Forrest Gump to look at the social networks based on a dataset from Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/T4HBA3

I have two ideas for categorical variables.

The dataverse dataset is great as it provides the social network for the movie, but it doesn't list any other information about the characters. So I would like to use a different dataset to try to match the character names to a dataset. For assignment 1, I pulled in the genres for a movie dataset I got from Kaggle: <https://www.kaggle.com/rounakbanik/the-movies-dataset>. This dataset also has characters mapped to actors, which is mapped to gender, So I believe I will be able to take the network data and then map it to gender.

The big struggle here will be that likely the character names will not match 1 to 1. I will have to either use some fuzzy text algorithm and/or manually match things up.

My second idea was to take series (Star Wars, Lord of the Rings) that exist in the database. Then, for a categorical variable I can check to see which characters exist in which movie. I could then answer the question: does centrality in movie 1, predict whether the character appears in movie 2 or movie 3? For this, the idea would be to load the movie data, then create a categorical variable based on matching the movie character names. This will likely be a smaller challenge as (I would hope!) the character would be named the same in the same dataset from one movie to the next.