Purpose

The aim of this project is to utilize scripts from the popular show “The Simpsons” to predict episode ratings. These findings can be used by busy networks bombarded with pilot scripts to save time and money on the selection process, filtering through less promising show ideas.

Data

The datasets for this project were obtained from the website *data.world* via the link: https://data.world/data-society/the-simpsons-by-the-data. In order to access the datasets a free account was created in order to join the website as a member. … The *simpsons_script_lines.csv* file containing raw text scripts per episode throughout “The Simpsons” series, as well as normalized text and and spoken word text was downloaded. Additionally, the *simpsons_episodes.csv* file was downloaded, containing imdb ratings for each episode in the series.

These csv files were imported into python where they were joined to create one dataset in the format of a dataframe. All irrelevant columns were dropped leaving *imdb_rating* column as well as the *raw_text* column. At this time the extent of data cleaning involved removing rows with missing values and transforming the *raw_text* column of the dataframe into an array.

Exploratory Data Analysis

Next steps involve exploratory data analysis and will include exploring the frequency and length of lines per character over the course of the series, and the distribution of ratings of the series as well as the number of views. Using inferential statistics next steps will involve exploring whether the gender of a character has an impact on the number and length of lines that they have.