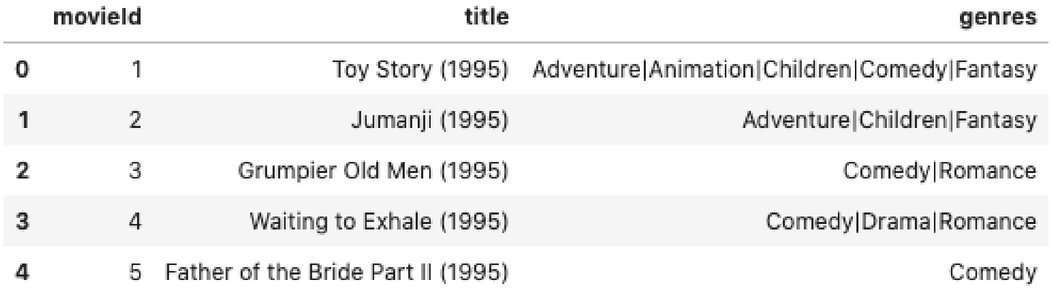
Joel James, Abira Hossain, Iqra Qumar
CSc 46000
Preliminary Analysis
Professor Conolly

- **Data Cleaning Code**
    - Code for cleaning and processing your data. Include a data dictionary for your transformed dataset.

```
import re

import pandas as pd


# https://files.grouplens.org/datasets/movielens/ml-25m.zip

movies = pd.read_csv("movies.csv")


def clean_title(title):

    title = re.sub("[^a-zA-Z0-9 ]", "", title)

    return title

movies["clean_title"] = movies["title"].apply(clean_title)
```

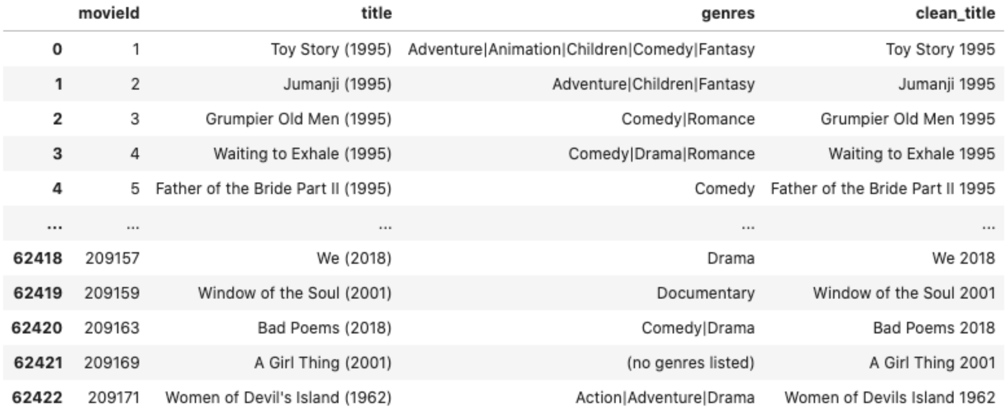
Field Name	Data Type	Data Format	Description	Example
movieNumber (not going to be explicitly declared)	Integer		Movie counter	10203
movieId	Integer		Unique number for movie	590
title	String		Title of movie (includes year released)	4.5
genre	String		Genre of the movie	Romance
clean_title	String		Cleaned movie title after removing any special characters	Toy Story 1995

- **Exploratory Analysis**
    - Describe what work you have done so far and include the code. This may include descriptive statistics, graphs and charts, and preliminary models.
        - This is a screenshot of the output before the cleaning.



	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

        -
        - This is a screenshot of the output after the cleaning.



	movieId	title	genres	clean_title
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	Toy Story 1995
1	2	Jumanji (1995)	Adventure|Children|Fantasy	Jumanji 1995
2	3	Grumpier Old Men (1995)	Comedy|Romance	Grumpier Old Men 1995
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance	Waiting to Exhale 1995
4	5	Father of the Bride Part II (1995)	Comedy	Father of the Bride Part II 1995
...	...	...	...	...
62418	209157	We (2018)	Drama	We 2018
62419	209159	Window of the Soul (2001)	Documentary	Window of the Soul 2001
62420	209163	Bad Poems (2018)	Comedy|Drama	Bad Poems 2018
62421	209169	A Girl Thing (2001)	(no genres listed)	A Girl Thing 2001
62422	209171	Women of Devil's Island (1962)	Action|Adventure|Drama	Women of Devils Island 1962

        - As you can see, it is cleaning out the special characters in the movie titles.

- **Challenges**
    - Describe any challenges you've encountered so far. **Let me know if there's anything you need help with!**
        - Understanding the various uses of the sklearn and numpy libraries
        - Cleaning data
            - Using regex to clean the movie titles in the dataset
            - Addressing the question of whether we need to clean more than what we’ve done
        - Learning how to commit to the Github repository

- **Future Work**
    - Describe what work you are planning to complete for the final analysis.
        - Next, we are looking into TF-IDF matrices (Term Frequency - Inverse Document Frequency), which are handy algorithms that uses the frequency of words to determine how relevant those words are to a given document.
        - Then, we want to create a search box to use to obtain user input for our recommendation system.
        - Ultimately, we want to build sort of a recommendation score based off of similar users to us who liked the same movie, and general users who watched the movie.
        - In the end, we want to make sure that the higher the recommendation score, the more similar the movie should be to the input given by the user.

- **Contributions as of right now:**
    - **Joel** - finding resources that contained implementations of similar models, writing the data dictionary, cleaning the data
    - **Iqra** - finding data, refining data dictionary, coming up with the analysis plan and identifying challenges
    - **Abira** - identifying anticipated challenges for the project and ethical considerations, cleaning the data