Assignment 2

Vaibhav Girish - 2019121
Shashank Chaurasia - 2019105

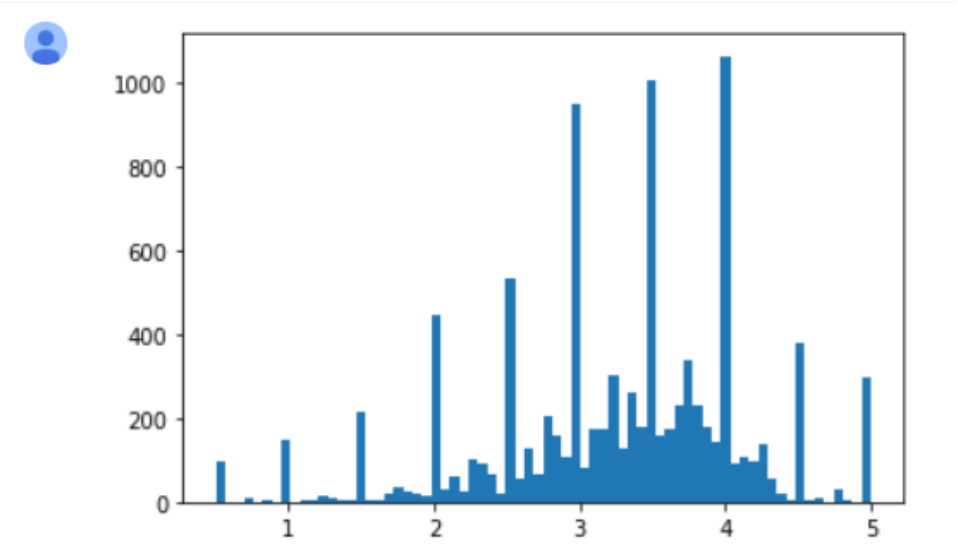
Question 1:
Exploratory Data Analysis:
We are only exploring the data in movies.csv, ratings.csv and tags.csv. We excluded links.csv because it only contains the movie ids in different websites like movielens, imdb and tmdb.

1. Checking for nan values: There were no nan values found in the dataset.
2. Checking for the most rated movies: We found the top 20 rated movies to be the following:

```
title
Forrest Gump (1994)                                                    329
Shawshank Redemption, The (1994)                                       317
Pulp Fiction (1994)                                                    307
Silence of the Lambs, The (1991)                                       279
Matrix, The (1999)                                                     278
Star Wars: Episode IV - A New Hope (1977)                              251
Jurassic Park (1993)                                                   238
Braveheart (1995)                                                      237
Terminator 2: Judgment Day (1991)                                      224
Schindler's List (1993)                                                220
Fight Club (1999)                                                      218
Toy Story (1995)                                                       215
Star Wars: Episode V - The Empire Strikes Back (1980)                  211
Usual Suspects, The (1995)                                             204
American Beauty (1999)                                                 204
Seven (a.k.a. Se7en) (1995)                                            203
Independence Day (a.k.a. ID4) (1996)                                   202
Apollo 13 (1995)                                                       201
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)    200
Lord of the Rings: The Fellowship of the Ring, The (2001)              198
Name: rating, dtype: int64
```

   Even though there are over 9000 movies in the data. More than 75% of them are not even rated at least 10 times. This makes sense as not all the movies are popular.

3. Checking for the mean of ratings for each movie: We calculated the mean of every rating for each movie and found the following result:



Most of the movies were rated between 3 and 4 and there were many movies which were rated in increments of 0.5 like at 2, 2.5, 3, 3.5, 4 because it is possible that these movies were only rated a few times, for example many unpopular movies could be rated 3 or 4 and only rated once or twice. That's why the histogram peaks at values of 3, 3.5, 4.

4. Checking for the most common tags: These were the most common tags that we found:

```
tag
In Netflix queue      131
atmospheric            36
thought-provoking      24
superhero              24
surreal                23
funny                  23
Disney                 23
religion               22
quirky                 21
psychology             21
dark comedy            21
sci-fi                 21
suspense               20
twist ending           19
crime                  19
visually appealing     19
politics               18
mental illness         16
music                  16
time travel            16
Name: tag, dtype: int64
```

5. Checking for the most reviewed movies: These were the most reviewed movies that we found:

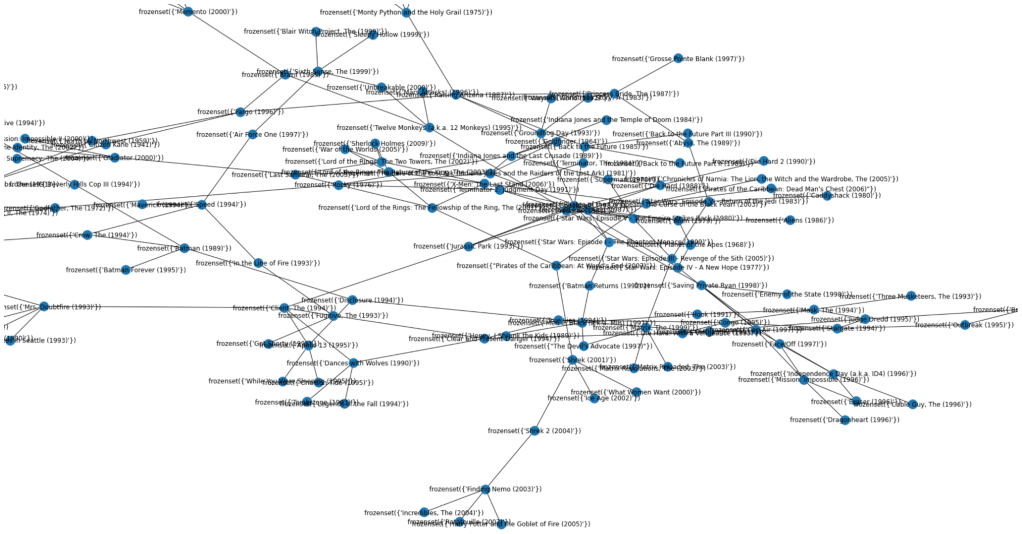
```
title
Pulp Fiction (1994)                                                181
Fight Club (1999)                                                   54
2001: A Space Odyssey (1968)                                        41
Léon: The Professional (a.k.a. The Professional) (Léon) (1994)      35
Eternal Sunshine of the Spotless Mind (2004)                        34
Big Lebowski, The (1998)                                            32
Donnie Darko (2001)                                                 29
Star Wars: Episode IV - A New Hope (1977)                           26
Inception (2010)                                                    26
Suicide Squad (2016)                                                19
Avatar (2009)                                                       18
In the Mood For Love (Fa yeung nin wa) (2000)                       18
Eraserhead (1977)                                                   17
Pi (1998)                                                           17
Avengers: Infinity War - Part I (2018)                              15
Memento (2000)                                                      13
Mary and Max (2009)                                                 13
Blade Runner (1982)                                                 13
Punch-Drunk Love (2002)                                             13
Up (2009)                                                           13
Name: tag, dtype: int64
```

Question 2:

We built a recommendation system using association rule mining techniques on the given dataset.
We first extracted the user information, by merging the movie dataset with the ratings to see which users watched which movies.
We then created a sparse matrix containing the movies watched and the user ID. Using this sparse matrix and the fpgrowth method we generate frequent itemsets with min_support 0.1.
Using these generated frequent itemsets we created association rules to determine the recommendations.
This would then recommend 4 movies based on the title of the movie.
In case the movie entered did not exist as an antecedent, we picked the best rated movies in the genre of the entered movie and returned this as a recommendation.

Question 3:

Using the frequent itemsets found above we found the maximal frequent itemsets.
We then plot all the antecedents with their corresponding consequents as edges in the graph. We then went through the list of maximal frequent all rules and changed the color of the nodes that were part of the maximal frequent itemsets.



Learnings:

1. We learned about Exploratory Data Analysis techniques and how to extensively use pandas and matplotlib to analyse huge chunks of data.
2. We researched different association rule mining techniques like apriori and fpgrowth and how to use them using libraries in python.
3. We learned how to make use of pandas to manipulate and gain insights on data.
4. We learned different association mining terms like confidence, support, lift, leverage and conviction.
5. We learned how to build a recommendation system after using association mining techniques to acquire rules.

References:

1. https://www.analyticsvidhya.com/blog/2021/04/mastering-exploratory-data-analysiseda-for-data-science-enthusiasts/
2. https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining
3.