ASSIGNMENT 2

Submitted by:
Mohammad Osama Ataullah
MT21127
Sehban Fazili
MT21143

1. Exploratory Data Analysis

Frequently occurring values in categorical features:

Frequent ‘title’ in movies.csv -

- Confessions of a Dangerous Mind (2002) 2
- Saturn 3 (1980) 2
- Eros (2004) 2

Frequently ‘genre’ in movies.csv -

- Drama
- Comedy
- Romance

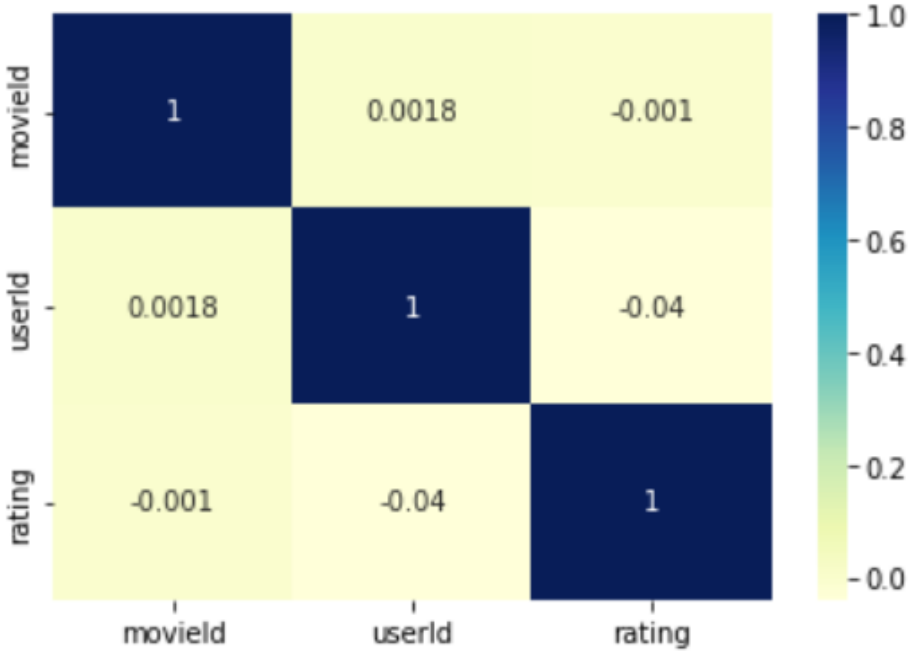
Frequent tags given by users to movies (tags.csv) -

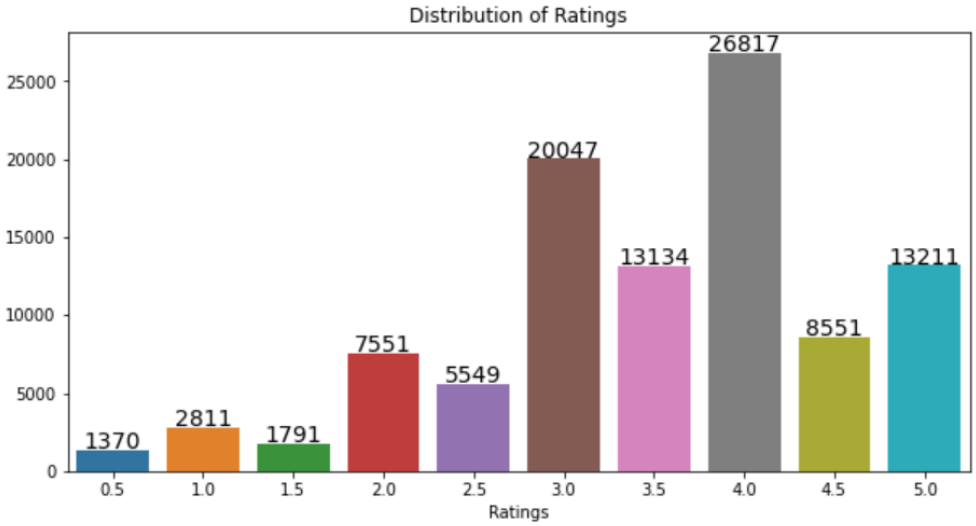
- In Netflix queue 131
- atmospheric 36
- thought-provoking 24

Missing Values

In our analysis, we found that no missing values were found in links, movies, tags csv file. However 8 values were missing (null) in ‘tmbdId’ attribute of links.csv

Correlation between features





Insights

As seen from the rating distribution, there are more movies that are highly rated than movies that have low ratings. This can be because of the fact that people who didn't like a movie, didn't bother to rate it either. Thus it is better to recommend movies that have high ratings.

2. Association Rule Mining

Steps:

a) We made the transaction file using the rating csv and movie csv files because the user who has rated the movie would have definitely watched the movie.

b) We converted the transaction data in appropriate numerical format using one hot encoding so that it can be given as input for apriori algorithm to generate frequent itemsets.

c) We then used these frequent itemsets with support > 0.1 to generate association rules.

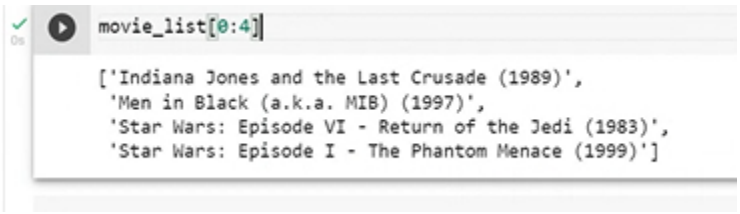
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1907651	(Kill Bill: Vol. 1 (2003), Fight Club (1999), ...	(Kill Bill: Vol. 2 (2004), Matrix, The (1999),...	0.106557	0.111475	0.100000	0.938462	8.418552	0.088121	14.438525
1907662	(Kill Bill: Vol. 2 (2004), Matrix, The (1999),...	(Kill Bill: Vol. 1 (2003), Fight Club (1999), ...	0.111475	0.106557	0.100000	0.897059	8.418552	0.088121	8.679157
1907681	(Kill Bill: Vol. 2 (2004), Star Wars: Episode ...	(Matrix, The (1999), Kill Bill: Vol. 1 (2003),...	0.114754	0.106557	0.100000	0.871429	8.178022	0.087772	6.948998
1051321	(Kill Bill: Vol. 1 (2003), Fight Club (1999), ...	(Kill Bill: Vol. 2 (2004), Star Wars: Episode ...	0.106557	0.114754	0.100000	0.938462	8.178022	0.087772	14.385246
1051340	(Kill Bill: Vol. 2 (2004), Star Wars: Episode ...	(Kill Bill: Vol. 1 (2003), Fight Club (1999), ...	0.114754	0.106557	0.100000	0.871429	8.178022	0.087772	6.948998

d) Then using these association rules we recommended top four movies to the user based on the movies that the user had previously watched.

Example1 Input:

```
movies_watched =  ['Star Wars: Episode IV - A New Hope (1977)','Indiana Jones and the Temple of Doom (1984)']
```

Example1 Output:



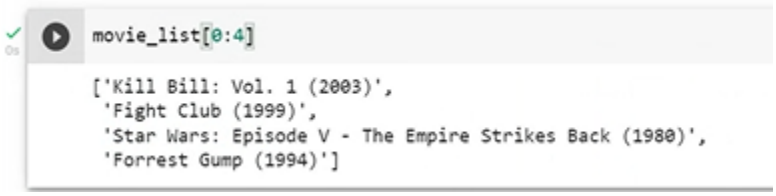
```
movie_list[0:4]

['Indiana Jones and the Last Crusade (1989)',
 'Men in Black (a.k.a. MIB) (1997)',
 'Star Wars: Episode VI - Return of the Jedi (1983)',
 'Star Wars: Episode I - The Phantom Menace (1999)']
```

Example2 Input:

```
movies_watched =  ['Matrix, The(1999)']
```

Example2 Output:



```
movie_list[0:4]

['Kill Bill: Vol. 1 (2003)',
 'Fight Club (1999)',
 'Star Wars: Episode V - The Empire Strikes Back (1980)',
 'Forrest Gump (1994)']
```

3. Using the above generated frequent itemsets we then find maximal frequent item sets. Some of the maximal frequent itemsets generated are as following:

```
[frozenset({'40-Year-Old Virgin, The (2005)'}),
 frozenset({'Amadeus (1984)'}),
 frozenset({'American President, The (1995)'}),
 frozenset({'Animal House (1978)'}),
 frozenset({'Austin Powers in Goldmember (2002)'}),
 frozenset({'Avengers, The (2012)'}),
 frozenset({'Big Fish (2003)'}),
 frozenset({'Blair Witch Project, The (1999)'}),
 frozenset({'Blazing Saddles (1974)'}),
 frozenset({'Borat: Cultural Learnings of America for Make Benefit
Glorious Nation of Kazakhstan (2006)'}),
 frozenset({"Bridget Jones's Diary (2001)"}),
 frozenset({'Butch Cassidy and the Sundance Kid (1969)'}),
 frozenset({"Charlie's Angels (2000)"}),
 frozenset({'Chicken Run (2000)'}),
 frozenset({'Children of Men (2006)'}),
 frozenset({'Chronicles of Narnia: The Lion, the Witch and the Wardrobe,
The (2005)'}),
 frozenset({'Citizen Kane (1941)'})]
```

Learnings

1. Learned how to analyze data and prepare it into suitable format for analysis.
2. Learned how recommendation systems work using association rule mining.
3. Learned how association rule mining generates frequent itemset which in turn generate association rules that can have very high business value.
4. Learned how to convert transaction data into numerical format such as one hot encoding.
5. Learned how to evaluate association rules using different metrics such as support, confidence, lift and select best rules.

References

1. On Scalability of Association-rule-based Recommendation: A Unified Distributed-computing Framework
2. https://towardsdatascience.com/how-to-find-closed-and-maximal-frequent-itemsets-from-fp-growth-861a1ef13e21
3. https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184