# RECOMMENDATION SYSTEM

## USING MULTIPLE METHODS TO RETURN USER RECOMMENDATIONS

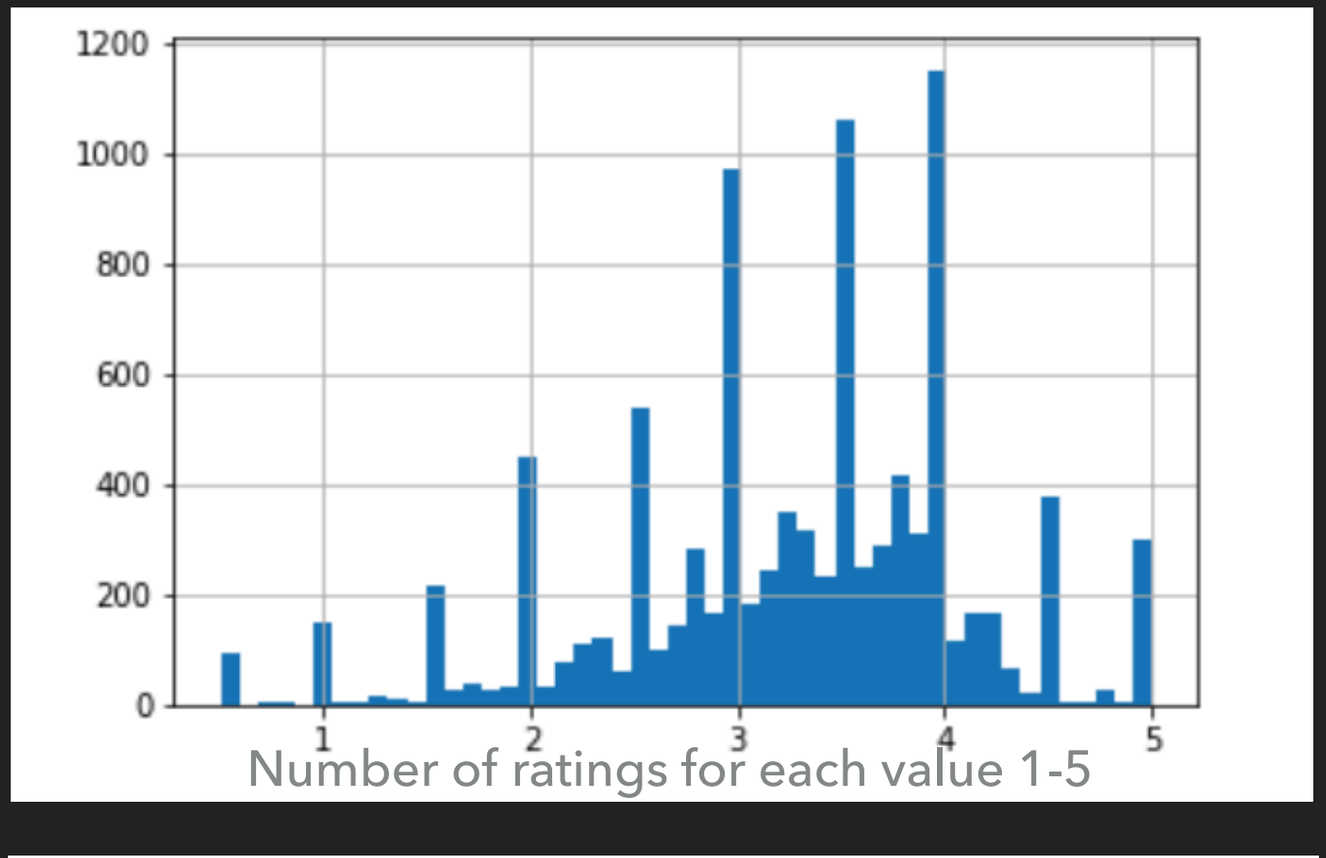
# OVERVIEW

- Detailed exploratory analysis of data
- Development of a simple recommendation system based on correlation between rating and rating quantity
- Development of model based recommendation system utilizing a learning model system

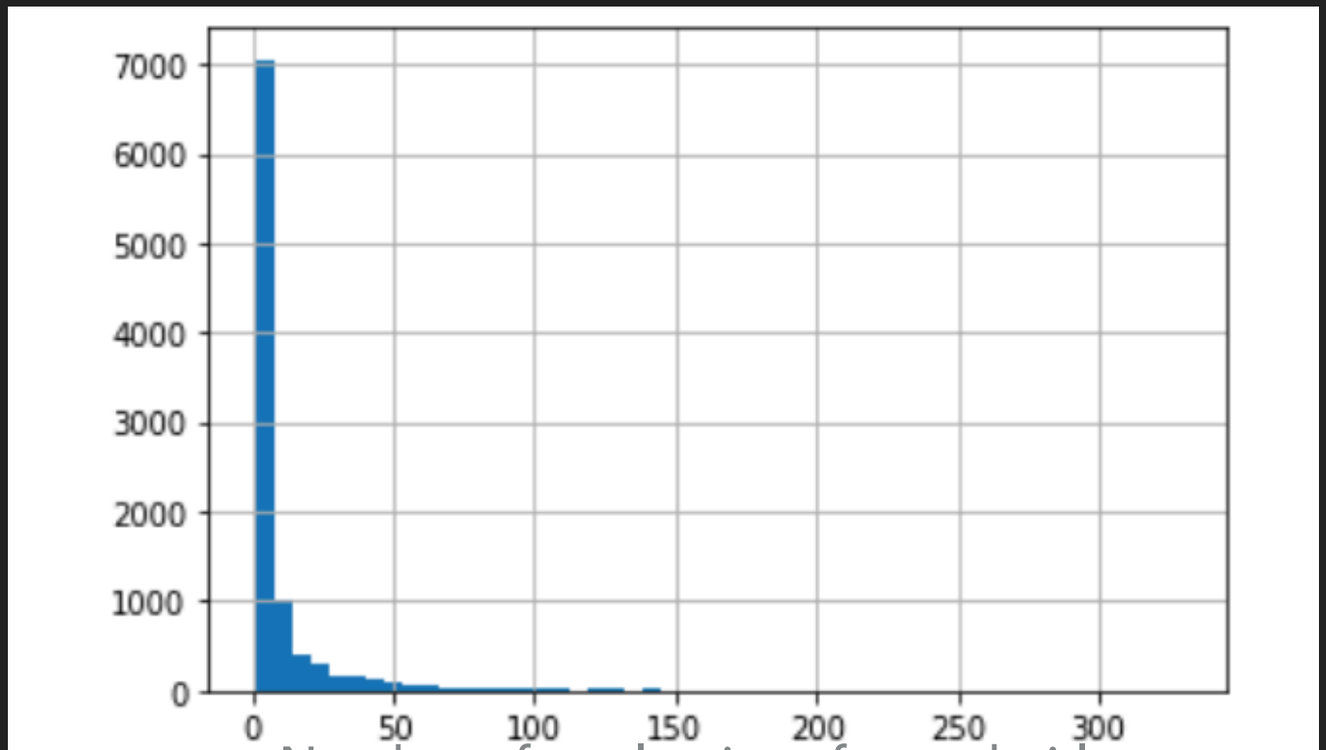
# ANALYSIS OF DATA

- Key metrics identified: Number of Ratings, Rating Value, User ID, Movie Title
- Ratings clustered with most titles having few ratings but better titles having more.
- Cross reference of both values confirm a strong correlation between rating value and number of ratings.

Visualizations showing user rating and rating value




Number of ratings for each value 1-5

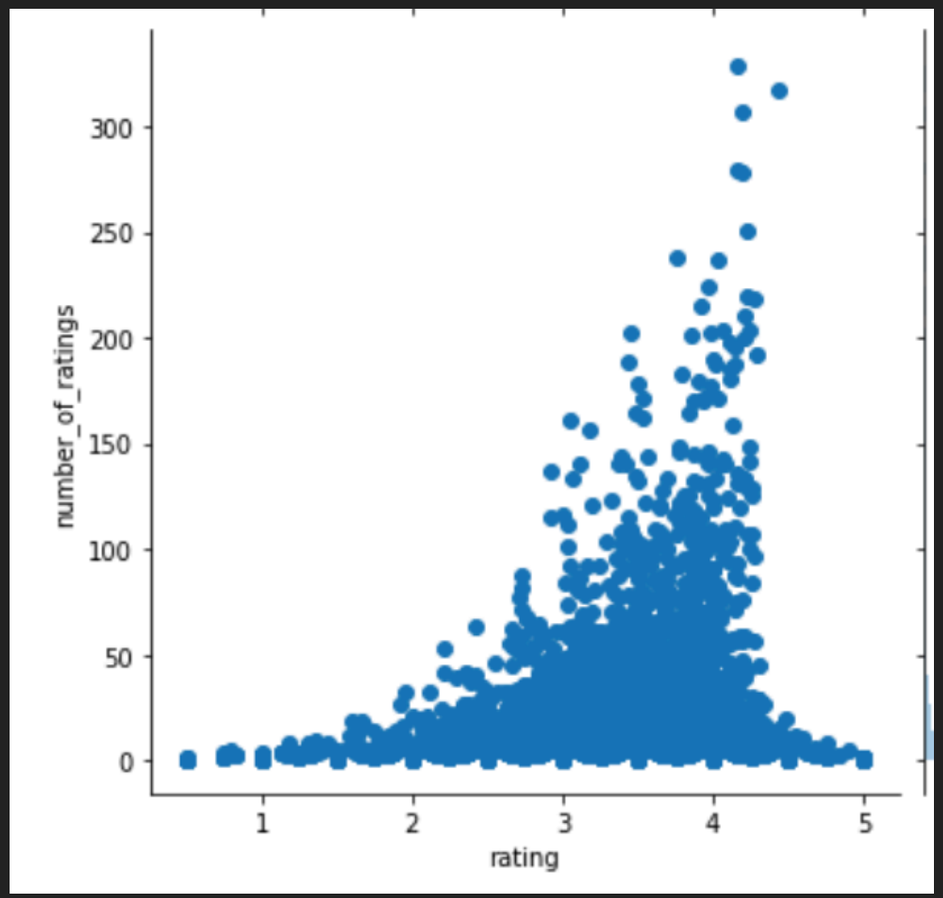



Number of total ratings for each title

# SIMPLE RECOMMENDATION SYSTEM BASED ON CORRELATED DATA

- Data cleaned into simplest form examining only rating, number of ratings, and title
- Correlation patters between highly rated and titles with over 100 ratings were developed
- Highest correlation columns were included in the final data

# Recommendation system results based on correlation of rating data




| title | rating | number_of_ratings |
|---|---|---|
| Forrest Gump (1994) | 4.164134 | 329 |
| Shawshank Redemption, The (1994) | 4.429022 | 317 |
| Pulp Fiction (1994) | 4.197068 | 307 |
| Silence of the Lambs, The (1991) | 4.161290 | 279 |
| Matrix, The (1999) | 4.192446 | 278 |
| Star Wars: Episode IV - A New Hope (1977) | 4.231076 | 251 |
| Jurassic Park (1993) | 3.750000 | 238 |
| Braveheart (1995) | 4.031646 | 237 |
| Terminator 2: Judgment Day (1991) | 3.970982 | 224 |
| Schindler's List (1993) | 4.225000 | 220 |

| title | Correlation | number_of_ratings |
|---|---|---|
| Forrest Gump (1994) | 1.000000 | 329 |
| Good Will Hunting (1997) | 0.484042 | 141 |
| Aladdin (1992) | 0.464268 | 183 |
| American History X (1998) | 0.457287 | 129 |
| Truman Show, The (1998) | 0.432556 | 125 |
| Braveheart (1995) | 0.416976 | 237 |
| Ferris Bueller's Day Off (1986) | 0.405830 | 109 |
| Mrs. Doubtfire (1993) | 0.401408 | 144 |
| Full Metal Jacket (1987) | 0.397241 | 102 |
| Saving Private Ryan (1998) | 0.390074 | 188 |

| title | Correlation | number_of_ratings |
|---|---|---|
| Pulp Fiction (1994) | 1.000000 | 307 |
| Fight Club (1999) | 0.543465 | 218 |
| Kill Bill: Vol. 1 (2003) | 0.504147 | 131 |
| Trainspotting (1996) | 0.437714 | 102 |
| Kill Bill: Vol. 2 (2004) | 0.421685 | 110 |
| Usual Suspects, The (1995) | 0.411700 | 204 |
| Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001) | 0.402193 | 120 |
| Eternal Sunshine of the Spotless Mind (2004) | 0.401534 | 131 |
| Reservoir Dogs (1992) | 0.394687 | 131 |
| Twelve Monkeys (a.k.a. 12 Monkeys) (1995) | 0.391141 | 177 |

# BUILDING OF RECOMMENDATION SYSTEM USING MODELING

- Surprise recommendation modeling system was used
- Users asked to rate 5 titles and model was trained using that user input to return model results.
- KNN and Grid Search used to train model
- Top 10 results are returned

Screenshots below show input and return of model data

```
Out[143]: [{'userId': 1000, 'movieId': 7099, 'rating': '5'},
           {'userId': 1000, 'movieId': 1722, 'rating': '3'},
           {'userId': 1000, 'movieId': 8607, 'rating': '2'},
           {'userId': 1000, 'movieId': 2384, 'rating': '5'},
           {'userId': 1000, 'movieId': 941, 'rating': '4'}]
```

```
Recommendation #  1 :  841    Streetcar Named Desire, A (1951)
Name: title, dtype: object

Recommendation #  2 :  602    Dr. Strangelove or: How I Learned to Stop Worr...
Name: title, dtype: object

Recommendation #  3 :  2462    Boondock Saints, The (2000)
Name: title, dtype: object

Recommendation #  4 :  933    Boot, Das (Boat, The) (1981)
Name: title, dtype: object

Recommendation #  5 :  694    Casablanca (1942)
Name: title, dtype: object
```

# CONCLUSIONS

- User recommendation systems are useful in a number of ways, from this example predicting similar movie titles a user could be interested in to sales and marketing.
- Models can be continuously improved by adding additional variables, expanding data collection on individual users
- Even a simplified system based entirely on data correlation can be effective in predicting user choices.

# THANK YOU.

Will VanDerKloot