Data Science Capstone - Movielense project

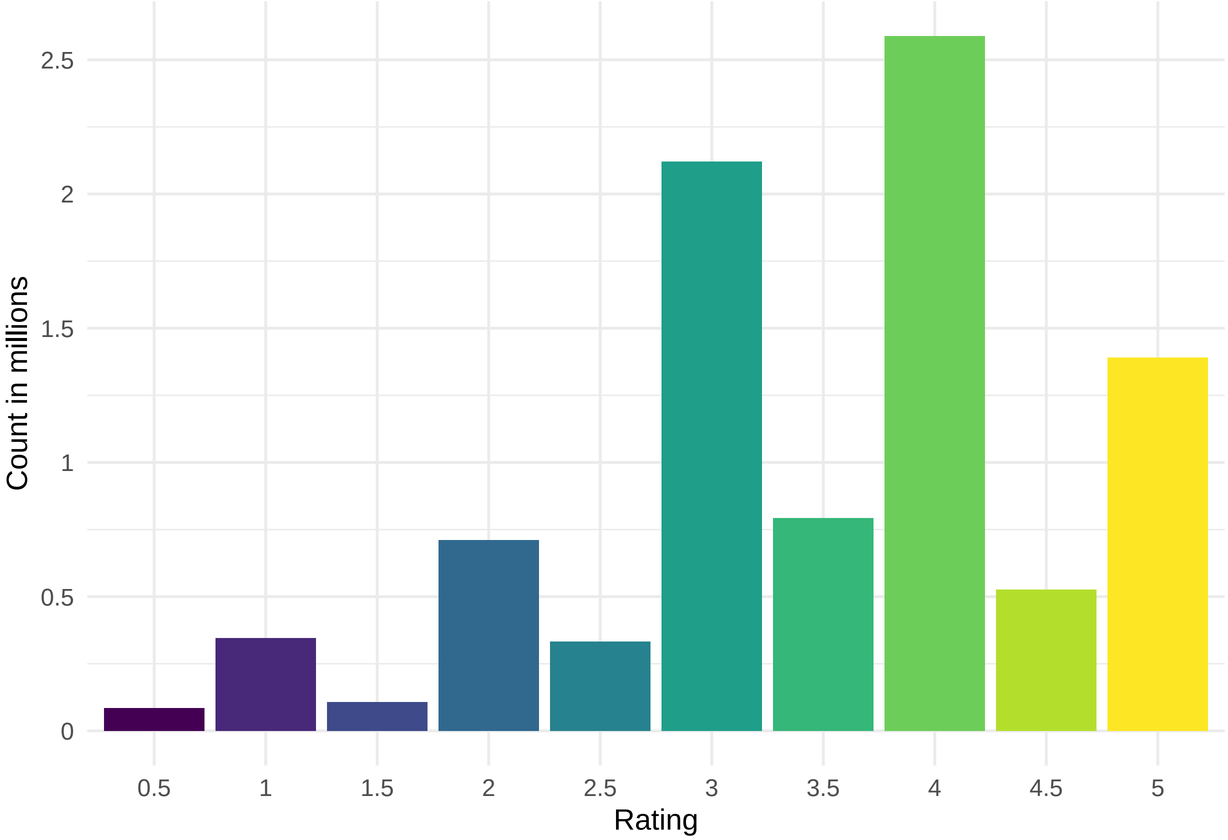
Julia Schröder

11/8/2020

Introduction

Description of dataset

The dataset used in this analysis is called "movielens" and comprises 9000055 ratings of 10677 given by 69878 users. These are the occurences per rating:

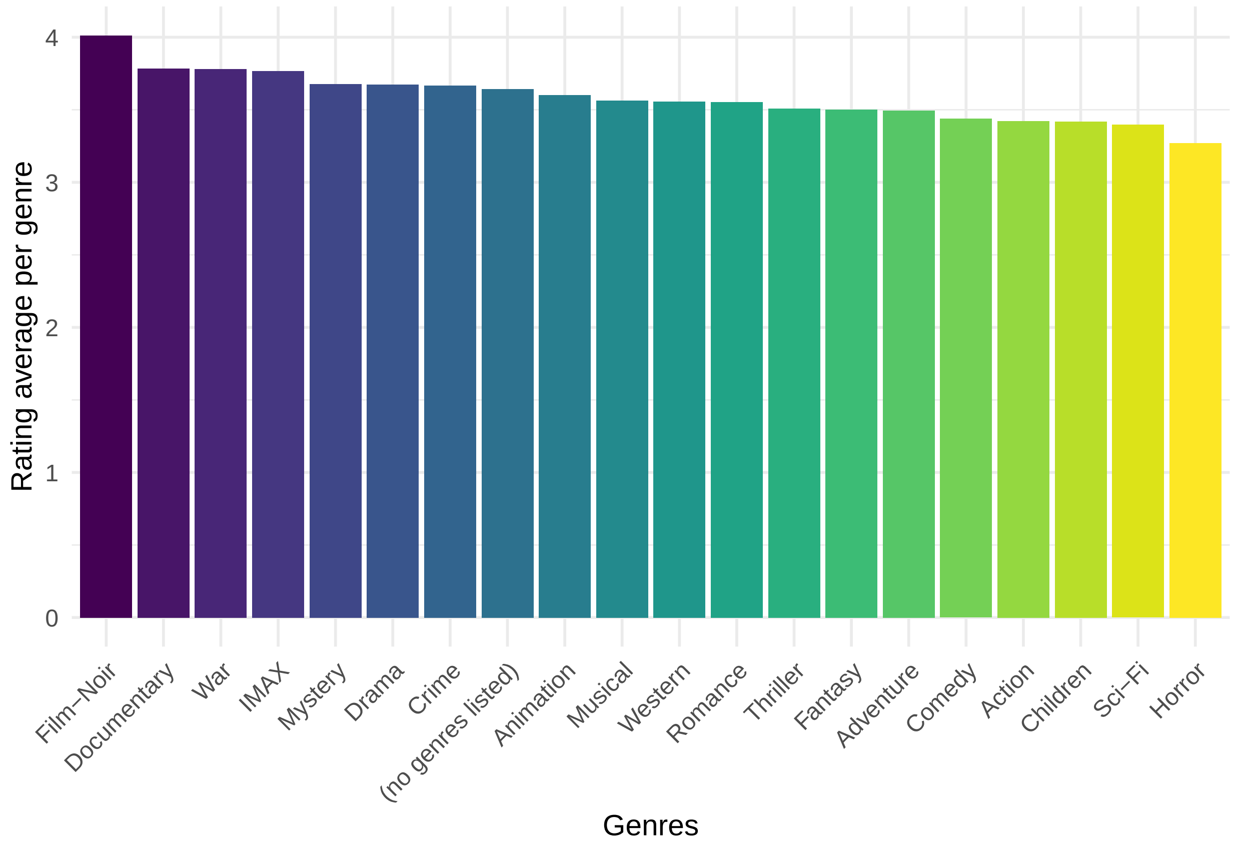


These are the 10 highest rated movies:

rating_avg	title
5.00	Hellhounds on My Trail (1999)
5.00	Satan's Tango (Sátántangó) (1994)
5.00	Shadows of Forgotten Ancestors (1964)
5.00	Fighting Elegy (Kenka erejii) (1966)
5.00	Sun Alley (Sonnenallee) (1999)
5.00	Blue Light, The (Das Blaue Licht) (1932)
4.75	Who's Singin' Over There? (a.k.a. Who Sings Over There) (Ko to tamo peva) (1980)
4.75	Human Condition II, The (Ningen no joken II) (1959)
4.75	Human Condition III, The (Ningen no joken III) (1961)
4.75	Constantine's Sword (2007)

These are the average ratings per genre:

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



Summary of the project's goal

Goal of this project was to build a movie rating predicition. This is possible by training a machine learning (ML) algorithm using the available ratings in the training dataset and evaluation of different ML-methods and parameters in cross-validation.

Key analysis steps

Data cleaning
Data cleaning involved the following steps:
* changing the timestamp to a date object
* extracting the year of the movie release from the title * extracting the genres from the nested column

Model selection

Model evaluation

Methods

explaination of process + techniques used: incl. data cleaning, data exploration + visualization insights gained modeling approach

Data cleaning

The timestamp was converted to a date object using the as_datetime funciton from the lubridate package.
The release year and title of the movie was extracted from the title column using the str_match function from the stringr package.
The genres were unnested by splitting the genres column by the "|" string and joining the data back using the movieId column.
This is the head of the final dataframe used for training the ML algorithms:

userId	rating	rating_date	movieId	movie_title	movie_year	movie_genre
1	5	1996-08-02 11:24:06	122	Boomerang	1992	Comedy
1	5	1996-08-02 11:24:06	122	Boomerang	1992	Romance
1	5	1996-08-02 10:58:45	185	Net, The	1995	Action
1	5	1996-08-02 10:58:45	185	Net, The	1995	Crime
1	5	1996-08-02 10:58:45	185	Net, The	1995	Thriller
1	5	1996-08-02 10:57:01	292	Outbreak	1995	Action
1	5	1996-08-02 10:57:01	292	Outbreak	1995	Drama
1	5	1996-08-02 10:57:01	292	Outbreak	1995	Sci-Fi
1	5	1996-08-02 10:57:01	292	Outbreak	1995	Thriller
1	5	1996-08-02 10:56:32	316	Stargate	1994	Action

Results

presentation of modeling results
discussion of model performance

Conclusion

brief summary of report
limitations
future work