CS210 project Osman Enes Erdoğdu

Osman Enes Erdoğdu-29344

January 2024

1) Data Part

The data of this project is the movies I have watched for the last 3 years.It consists the name of the movies with: the year I've watched the film, genre of the film, IMDB rank/score, studio/distributor, director of the film and the release date of the film.

	Movie	Year Watched	Genre	IMDB rank	Studio	Director	Relase year
0	Split	2021	Horror/Mystery	7.3	Blinding Edge Pictures	M. Night Shyamalan	2016
1	Focus	2021	Comedy/Crime	6.6	Warner Bros	Glenn Ficarra	2015
2	Dark Knight Rises	2021	Action/Thriller	9.0	Warner Bros	Christopher Nolan	2012
3	Rocketman	2021	Musical/Drama	7.3	Paramount Pictures	Dexter Fletcher	2019
4	The Shawshank Redemption	2021	Thriller/Crime	9.3	Warner Bros	Frank Darabont	1994
5	Star Wars 4	2021	Sci-fi/Fantasy	8.6	20th Century Studios	George Lucas	1977
6	Star Wars 5	2021	Sci-fi/Fantasy	8.7	20th Century Studios	Irvin Kershner	1980
7	Here Comes the Boom	2021	Comedy/Action	6.4	Sony Pictures	Frank Coraci	2012
8	Star Wars 6	2021	Sci-fi/Fantasy	8.3	20th Century Studios	Richard Marquand	1983
9	Sonic	2021	Comedy/Action	6.5	Paramount Pictures	Jeff Fowler	2020

Here is the example look of the first 10 rows of the data and columns.

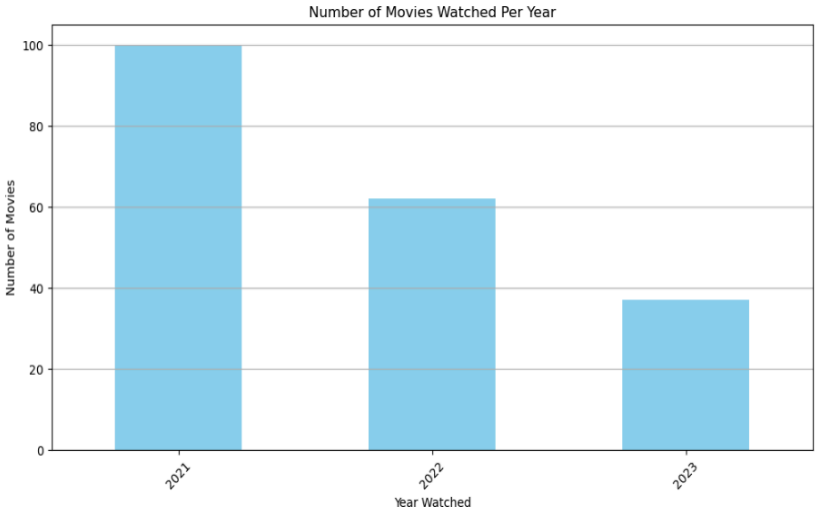
2) Motivation of project

Motivation of this project is to see if there is a relationship between my data. Such as how my preferences changed over the years, which genre I switched to, how are the quality of these films(according to IMDB scores) etc.

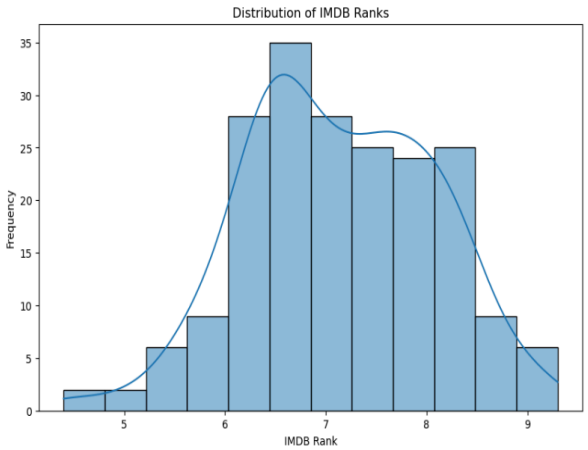
3) EDA Part

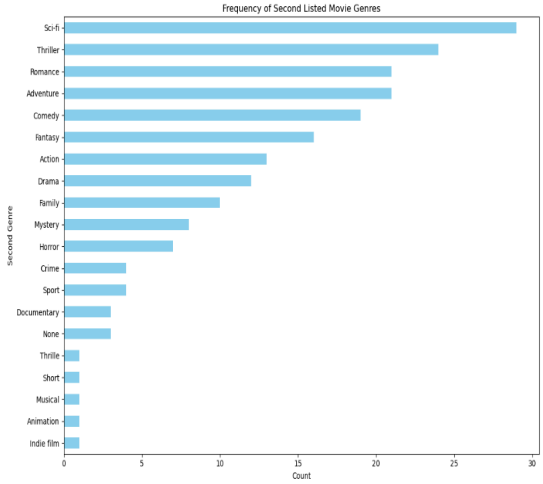
This is the data analysis and exploration part. In this part I have shown the various graphs to visualize the data.

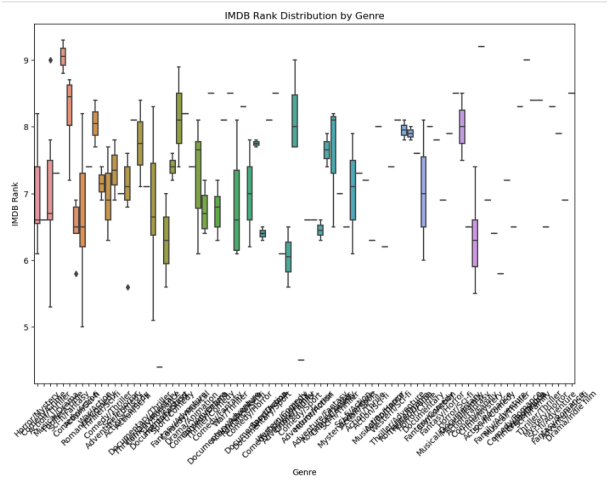
```
Missing values in each column:
 Movie           0
Year Watched     0
Genre            0
IMDB rank        0
Studio           0
Director         0
Relase year      0
dtype: int64

Data types of each column:
 Movie           object
Year Watched      int64
Genre            object
IMDB rank       float64
Studio           object
Director         object
Relase year       int64
dtype: object
```

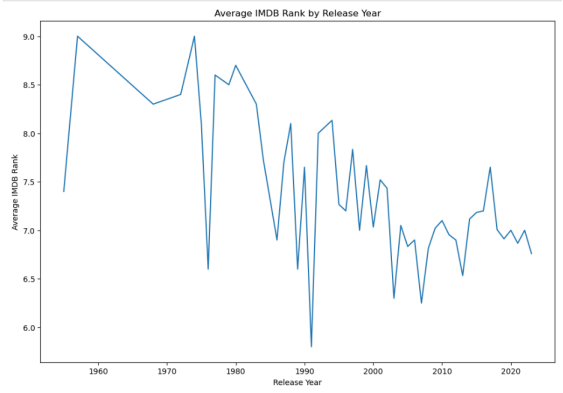


First I checked for any missing values and analyzed the data types of columns. Then at the right I looked for the movies I've watched each year. Since there was a lockdown in 2021 it can be seen that I watched more movies that year.

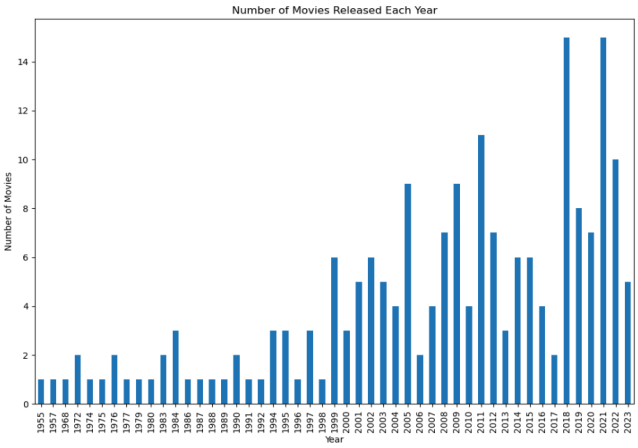


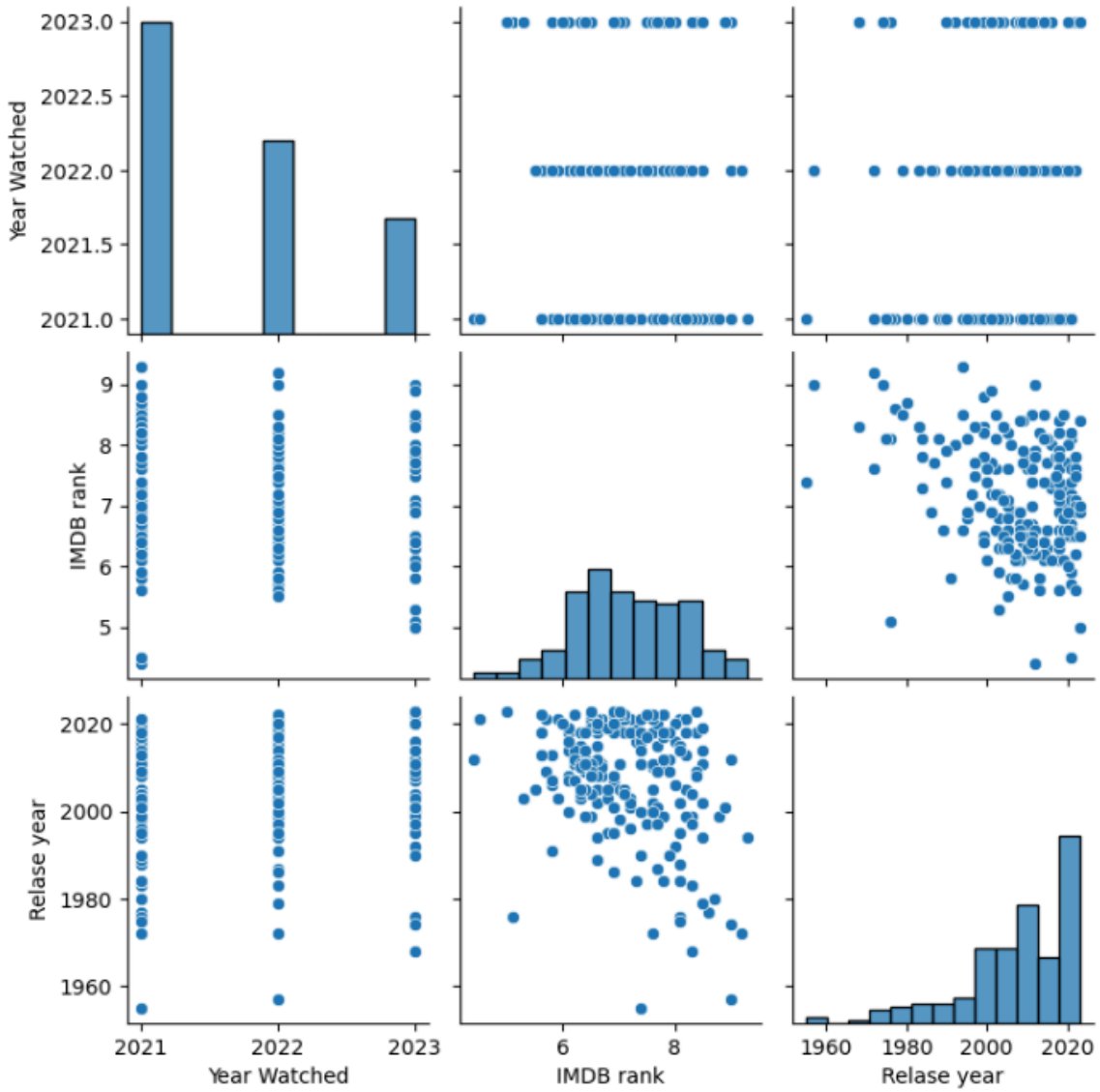




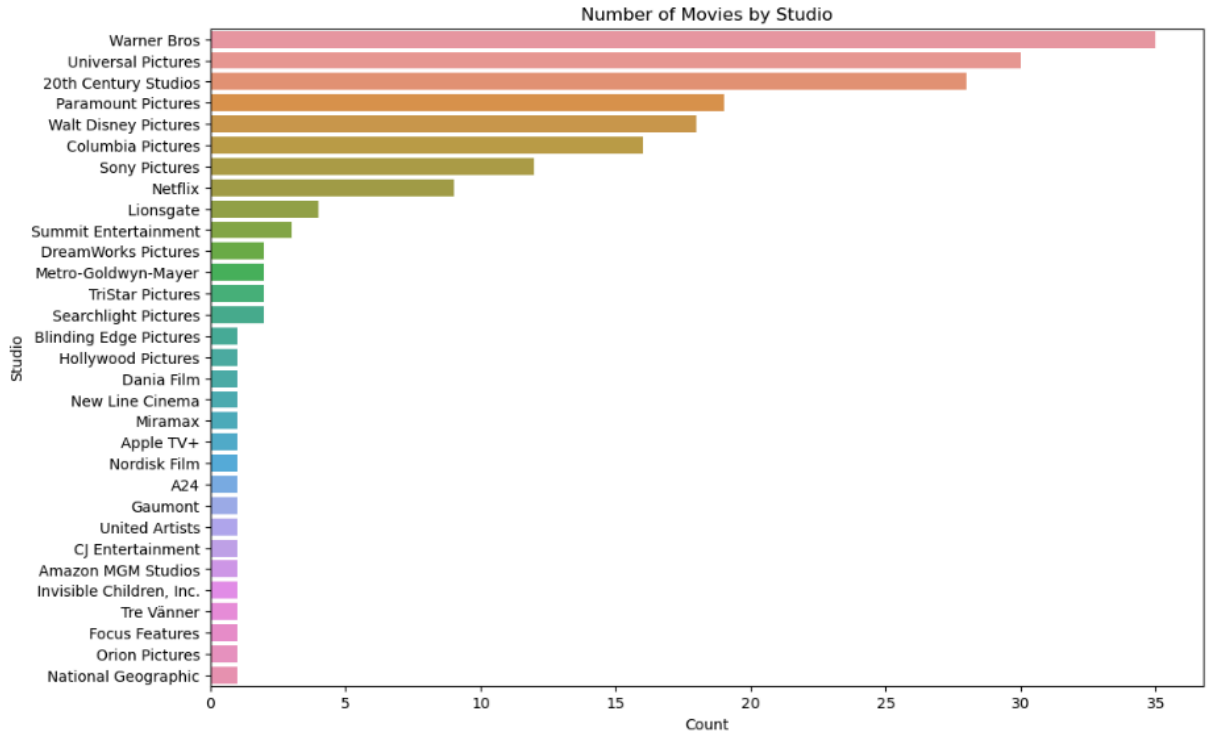


```
Director
Frank Darabont           9.3
Francis Ford Coppola     9.1
Sidney Lumet             9.0
Peter Jackson            8.9
Irvin Kershner           8.7
                         ...
Mark Steven Johnson      5.3
Mariano Laurenti         5.1
Ben Wheatley             5.0
Malcolm D. Lee           4.5
Jason Russell            4.4
Name: IMDB rank, Length: 162, dtype: float64
```

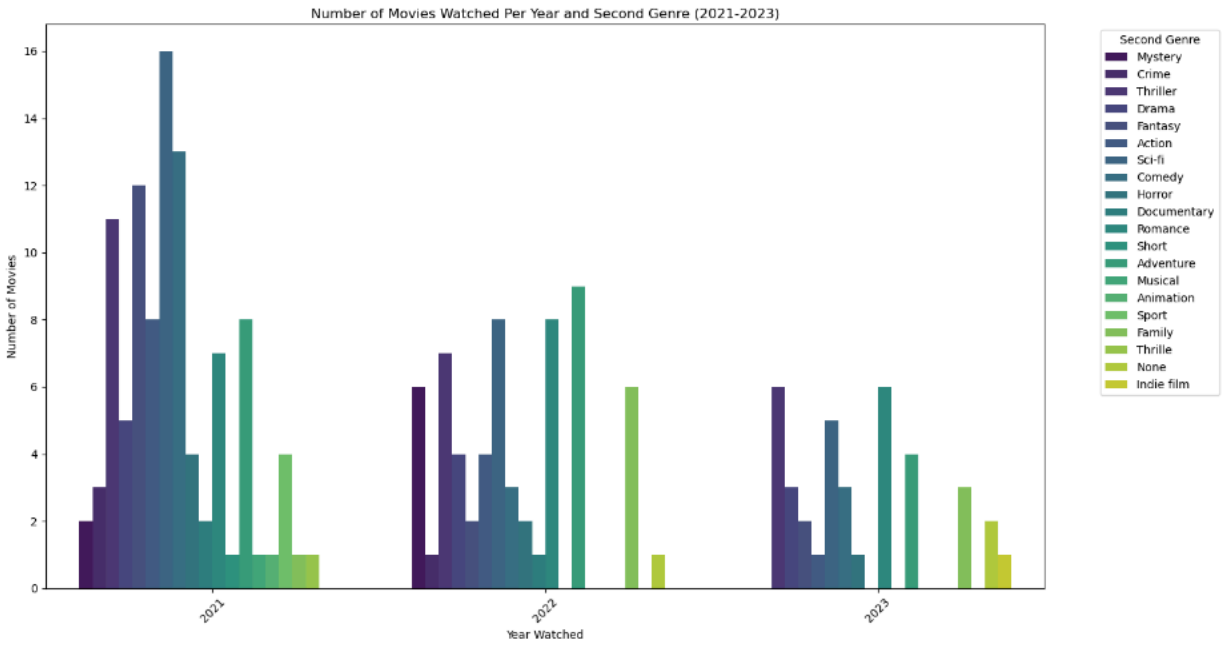




Then above there are different analysis for different features of the movies.



In this part I checked if there are any specific studios I preferred and I realized Warner Bros. is my number one choice, then comes Universal Pictures and 20th Century Studios.



In here I have seen the change of my movie preference over the years. In 2021 when there was covid-19 my favorite film genre was sci-fi but over the years it can be seen that genres like mystery, thriller increased when genres like sci-fi and fantasy decreased. I think because of the earthquake my preference might change to more realistic movies.

4) Prediction and machine learning
   Below there is the output of my code for how accurate my program would predict my future movie preferences. The lower the number is the more accurate it gets.

```
0.7010433877391228
```

And here is the visualization of the prediction accuracy