project Description

IMDb

Delivered by:

Ibrahim alhammad

Waad alduhami

Abstract:

The Internet Movie Database (IMDB) is an online database of information related movies, TV shows, celebrities, genre, reviews, etc. The IMDB website enables registered users to rate different movies,TV shows and actors on scale of 1 to 10.

As we all now that there are a lot of Production Companies in the World like (Universal Pictures, Warner Bros., Columbia Pictures and Walt Disney Pictures)

so in our project we will talk about how to help them to get a higher **rating** based on iMDb top 250 movies scarped data from iMDb website

data:

The data scraping was one of the most time consuming parts of this project because it took quite a bit of time to define the business problem enough to come up with an idea of what I needed and it was also difficult at times to navigate through the nested tables (the HTML). However, BeautifulSoup made it quite simple to grab the HTML and parse through it.

With this problem in mind, it was time to start the scraping process. I decided to start by scraping as many movies as possible so that I could build out my training data set. When I looked at the a sample page on boxofficemojo, I noticed that a lot of interesting information was captured in the table at the top of each movie page. As a first step, I realized it would be quick to just take down this information.

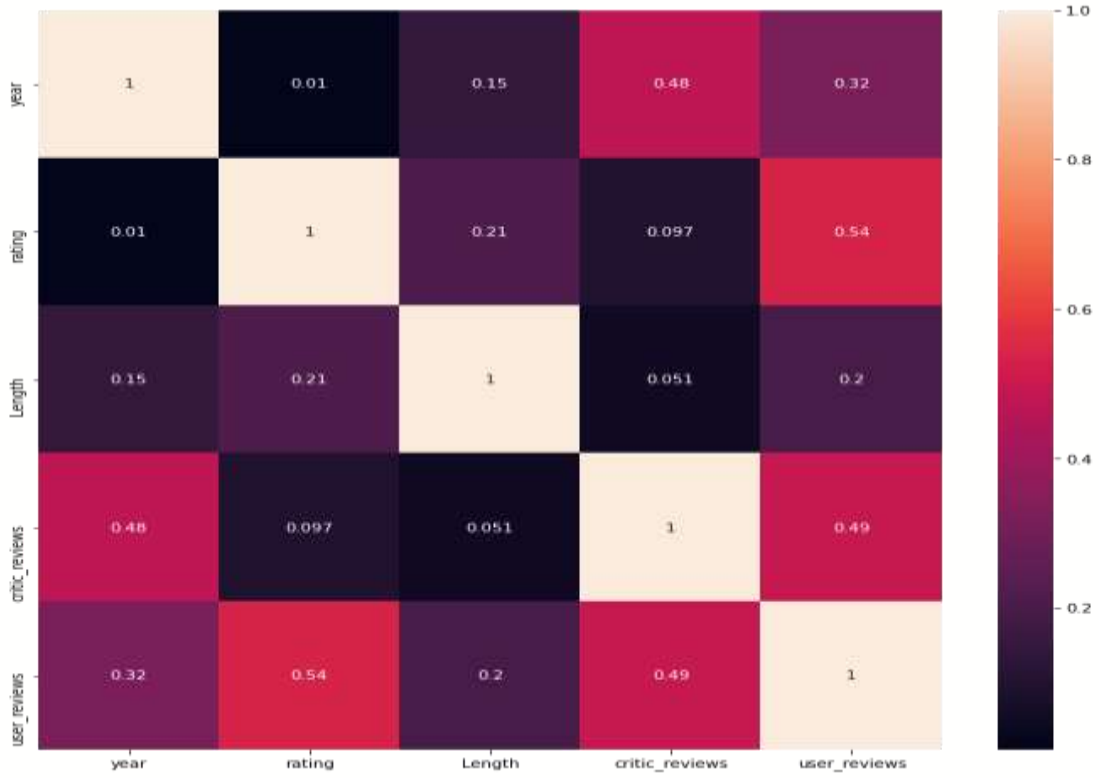
we used the top 250 movies rated based on iMDb website and 5 Columns

	movie_title	year	director	rating	Motion_Rating	Length	critic_reviews	user_reviews
0	The Shawshank Redemption	1994	Frank Daraboni (dir.)	9.220472	R12	142.0	192.0	9500.0
1	The Godfather	1972	Francis Ford Coppola (dir.)	9.147190	PG12	175.0	267.0	4600.0
2	The Godfather: Part II	1974	Francis Ford Coppola (dir.)	8.980494	R12	202.0	189.0	1200.0
3	The Dark Knight	2008	Christopher Nolan (dir.)	8.972840	PG12	152.0	434.0	7600.0
4	12 Angry Men	1957	Sidney Lumet (dir.)	8.938812	Approved	96.0	159.0	1800.0

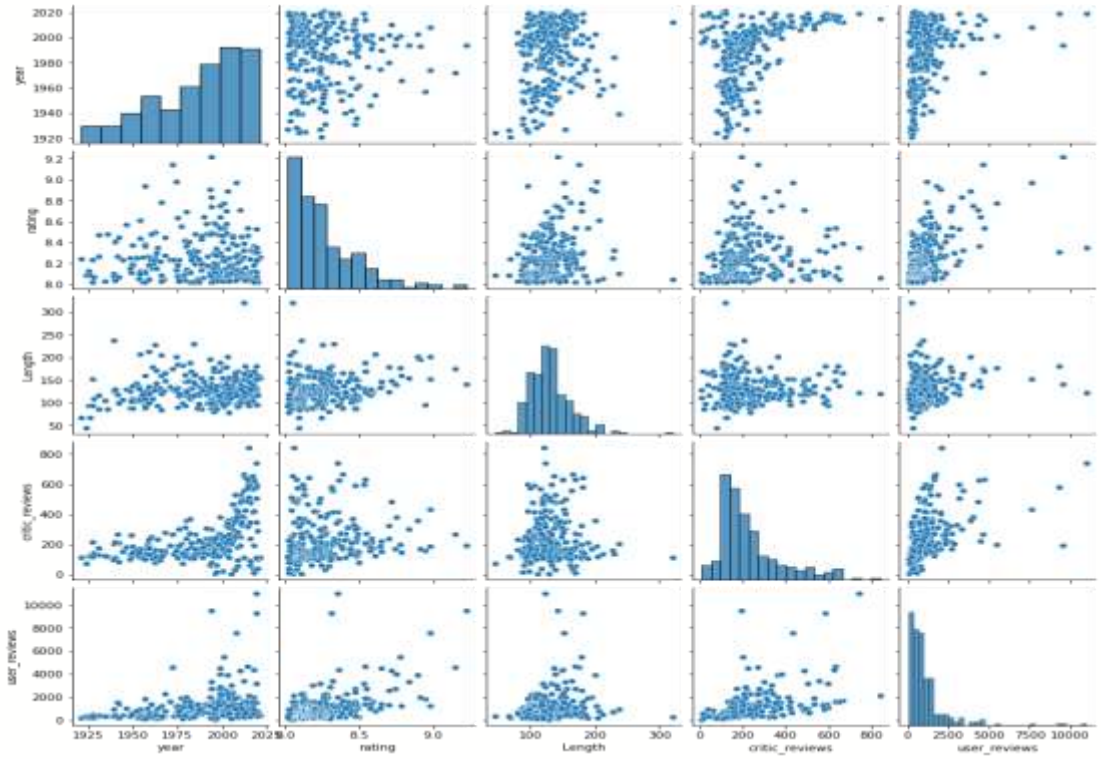
Tools:

- Numpy and Pandas
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Seaborn

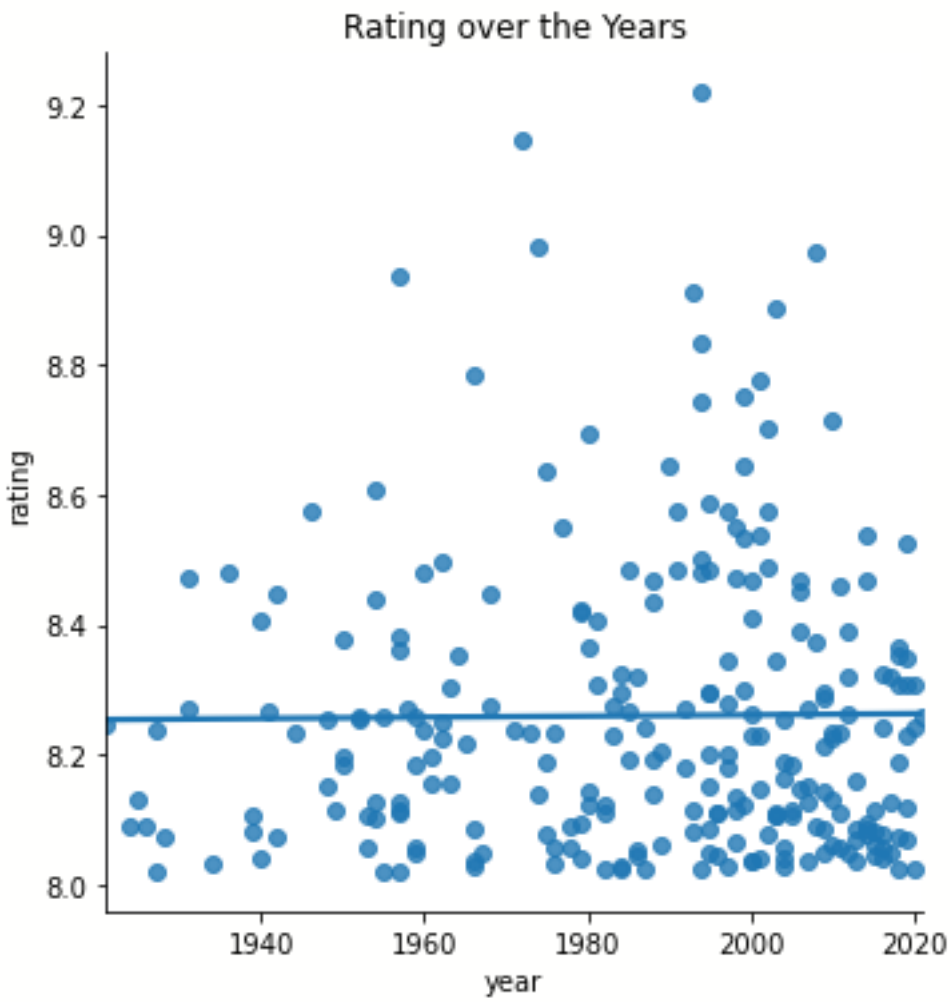
Communication:



Pairplot :



The difference between year , rating


Rating over the Years
rating
9.2
9.0
8.8
8.6
8.4
8.2
8.0
1940
1960
1980
2000
2020
year

Dep. Variable:	rating	**R-squared:**	0.453
Model:	OLS	**Adj. R-squared:**	0.437
Method:	Least Squares	**F-statistic:**	28.02
Date:	Sun, 26 Sep 2021	**Prob (F-statistic):**	1.29e-20
Time:	07:30:08	**Log-Likelihood:**	60.964
No. Observations:	175	**AIC:**	-109.9
Df Residuals:	169	**BIC:**	-90.94
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.1893	1.140	10.694	0.000	9.939	14.439
year	-0.0021	0.001	-3.632	0.000	-0.003	-0.001
Length	0.0011	0.000	2.375	0.019	0.000	0.002
user_reviews	7.798e-05	1.23e-05	6.365	0.000	5.38e-05	0.000
PG12	0.3448	0.114	3.017	0.003	0.119	0.570
R12	0.1952	0.053	3.685	0.000	0.091	0.300

Omnibus:	16.734	**Durbin-Watson:**	2.055
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.933
Skew:	0.660	**Prob(JB):**	4.69e-05
Kurtosis:	3.995	**Cond. No.**	2.12e+05