Predicting Piracy Rates of Movies using ML Algorithms

Alisha Baral

Background

- Piracy has significant economic impact on the entertainment industry. It is the unauthorized use or reproduction of copyrighted works, and it has become increasingly prevalent with the rise of the internet and digital media.
- ML algorithms are a powerful tool that can be used to predict piracy rates of movies by analyzing various factors that can influence piracy.

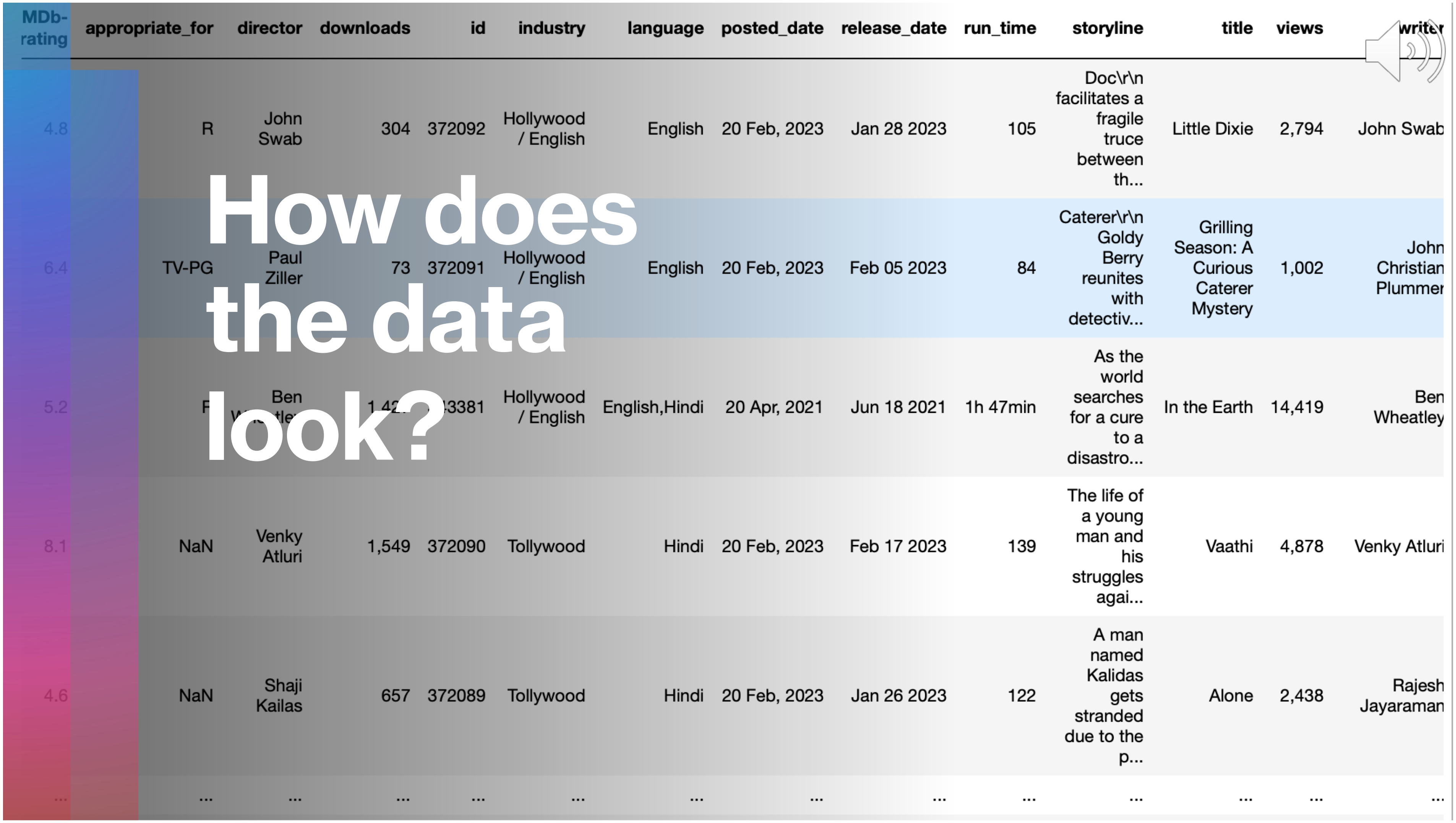
About this data

- The movie dataset used for this projects was taken from Kaggle where it was gathered from a pirated website that has a user base of around 2M visitors per month.
- It contains information about movies, including details such as movie tittle, genre, director, release date, runtime, storyline, language, writer, and user ratings.

Research question

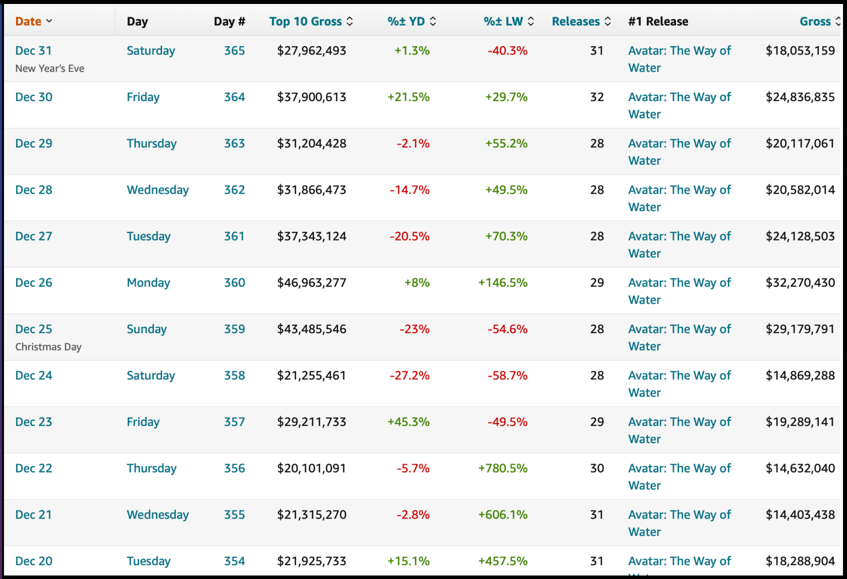
1) What factors influence the number of downloads (Also meaning mostly pirated) a movie receives? How much of an impact does the industry, language, or release date have on the number of downloads?
2) Does the IMDb rate of any movie affect the download number of the movie? What is the highest pirated movie together and from each industry?
3) Which industry is mostly affected by piracy?

How does the data look?

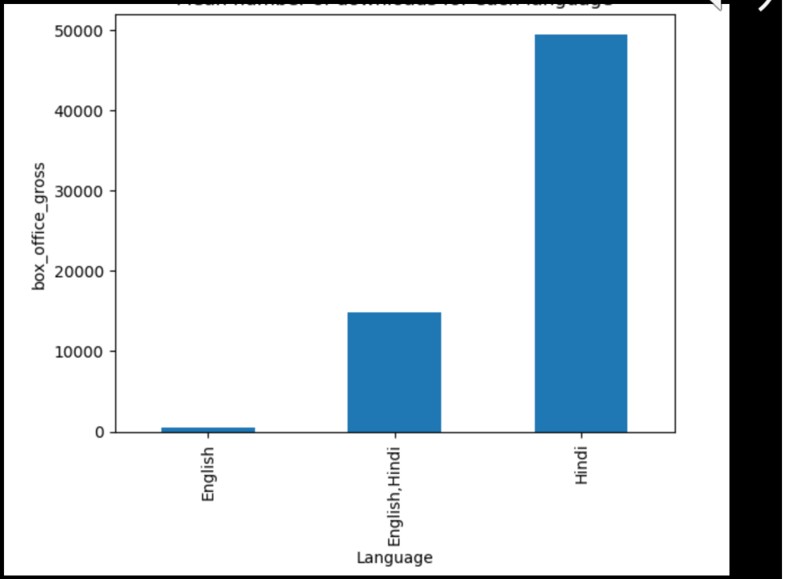


MDb-rating	appropriate_for	director	downloads	id	industry	language	posted_date	release_date	run_time	storyline	title	views	writer
4.8	R	John Swab	304	372092	Hollywood / English	English	20 Feb, 2023	Jan 28 2023	105	Doc\r\n facilitates a fragile truce between th...	Little Dixie	2,794	John Swab
6.4	TV-PG	Paul Ziller	73	372091	Hollywood / English	English	20 Feb, 2023	Feb 05 2023	84	Caterer\r\n Goldy Berry reunites with detectiv...	Grilling Season: A Curious Caterer Mystery	1,002	John Christian Plummer
5.2	R	Ben Wheatley	1,42[illegible]	[illegible]43381	Hollywood / English	English,Hindi	20 Apr, 2021	Jun 18 2021	1h 47min	As the world searches for a cure to a disastro...	In the Earth	14,419	Ben Wheatley
8.1	NaN	Venky Atluri	1,549	372090	Tollywood	Hindi	20 Feb, 2023	Feb 17 2023	139	The life of a young man and his struggles agai...	Vaathi	4,878	Venky Atluri
4.6	NaN	Shaji Kailas	657	372089	Tollywood	Hindi	20 Feb, 2023	Jan 26 2023	122	A man named Kalidas gets stranded due to the p...	Alone	2,438	Rajesh Jayaraman
...	...	...	...	...	...	...	...	...	...	...	...	...	...

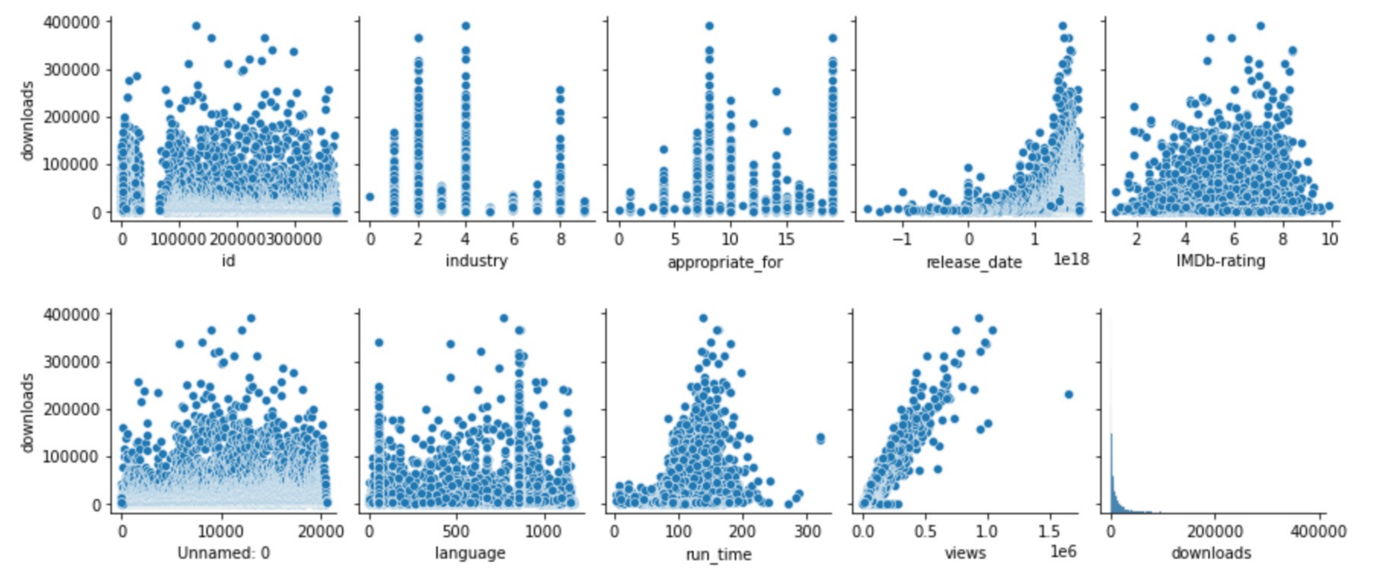
Data preprocessing and EDA



Date	Day	Day #	Top 10 Gross	%± YD	%± LW	Releases	#1 Release	Gross
Dec 31 New Year's Eve	Saturday	365	$27,962,493	+1.3%	-40.3%	31	Avatar: The Way of Water	$18,053,159
Dec 30	Friday	364	$37,900,613	+21.5%	+29.7%	32	Avatar: The Way of Water	$24,836,835
Dec 29	Thursday	363	$31,204,428	-2.1%	+55.2%	28	Avatar: The Way of Water	$20,117,061
Dec 28	Wednesday	362	$31,866,473	-14.7%	+49.5%	28	Avatar: The Way of Water	$20,582,014
Dec 27	Tuesday	361	$37,343,124	-20.5%	+70.3%	28	Avatar: The Way of Water	$24,128,503
Dec 26	Monday	360	$46,963,277	+8%	+146.5%	29	Avatar: The Way of Water	$32,270,430
Dec 25 Christmas Day	Sunday	359	$43,485,546	-23%	-54.6%	28	Avatar: The Way of Water	$29,179,791
Dec 24	Saturday	358	$21,255,461	-27.2%	-58.7%	28	Avatar: The Way of Water	$14,869,288
Dec 23	Friday	357	$29,211,733	+45.3%	-49.5%	29	Avatar: The Way of Water	$19,289,141
Dec 22	Thursday	356	$20,101,091	-5.7%	+780.5%	30	Avatar: The Way of Water	$14,632,040
Dec 21	Wednesday	355	$21,315,270	-2.8%	+606.1%	31	Avatar: The Way of Water	$14,403,438
Dec 20	Tuesday	354	$21,925,733	+15.1%	+457.5%	31	Avatar: The Way of	$18,288,904

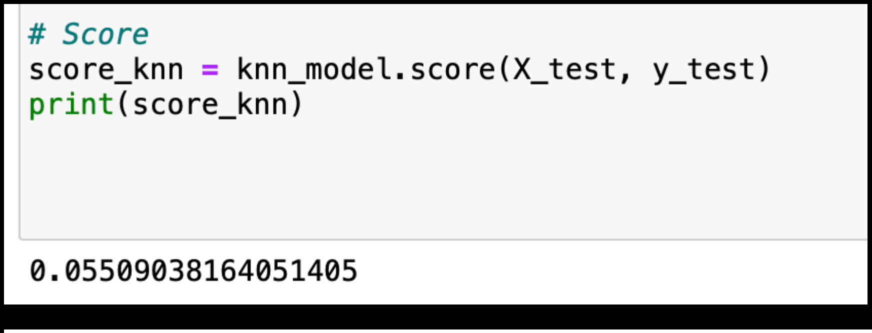


Limitations of Integrating Box Office Data



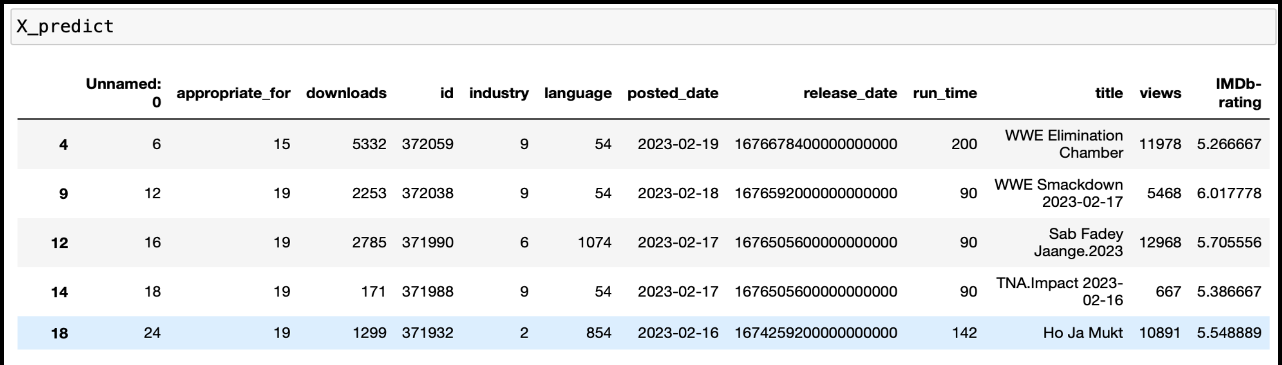
Creating a scatter plot with 'downloads' on the y-axis and each variable in the group on the x-axis.

Regression with kNN



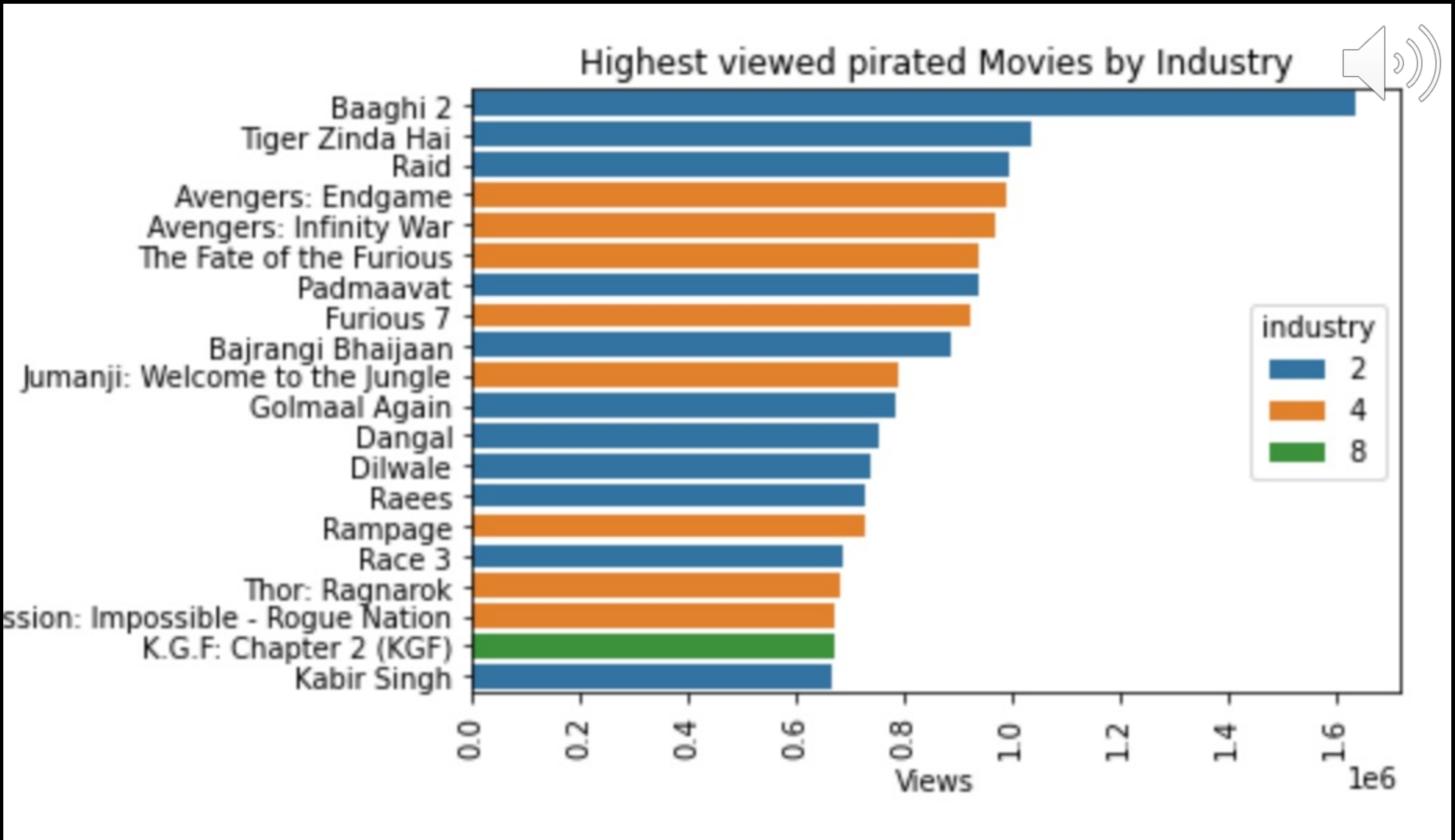
```
# Score
score_knn = knn_model.score(X_test, y_test)
print(score_knn)
```

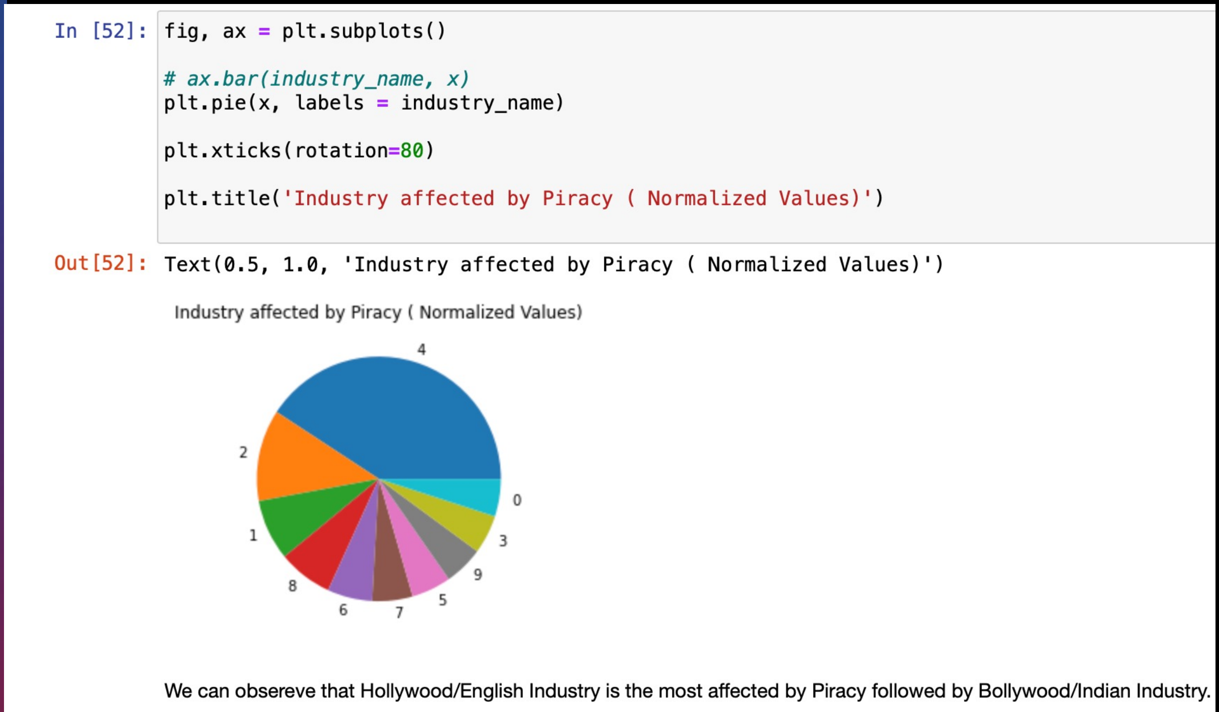
```
0.05509038164051405
```



```
X_predict
```

	Unnamed: 0	appropriate_for	downloads	id	industry	language	posted_date	release_date	run_time	title	views	IMDb-rating
4	6	15	5332	372059	9	54	2023-02-19	1676678400000000000	200	WWE Elimination Chamber	11978	5.266667
9	12	19	2253	372038	9	54	2023-02-18	1676592000000000000	90	WWE Smackdown 2023-02-17	5468	6.017778
12	16	19	2785	371990	6	1074	2023-02-17	1676505600000000000	90	Sab Fadey Jaange.2023	12968	5.705556
14	18	19	171	371988	9	54	2023-02-17	1676505600000000000	90	TNA.Impact 2023-02-16	667	5.386667
18	24	19	1299	371932	2	854	2023-02-16	1674259200000000000	142	Ho Ja Mukt	10891	5.548889


Highest viewed pirated Movies by Industry
Baaghi 2
Tiger Zinda Hai
Raid
Avengers: Endgame
Avengers: Infinity War
The Fate of the Furious
Padmaavat
Furious 7
Bajrangi Bhaijaan
Jumanji: Welcome to the Jungle
Golmaal Again
Dangal
Dilwale
Raees
Rampage
Race 3
Thor: Ragnarok
ssion: Impossible - Rogue Nation
K.G.F: Chapter 2 (KGF)
Kabir Singh
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e6
Views
industry
2
4
8

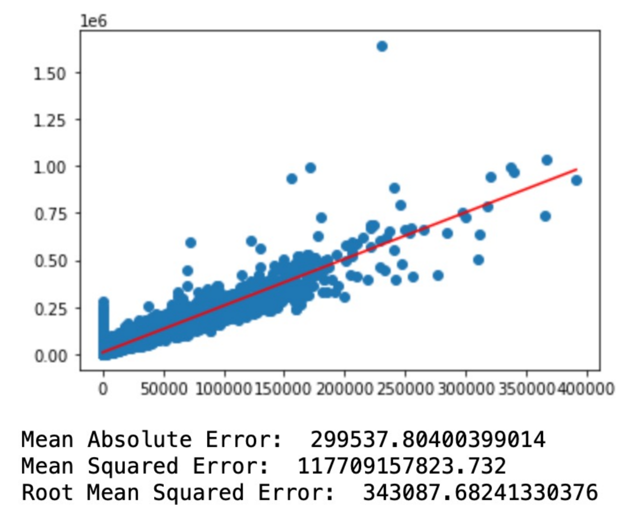


```
In [52]: fig, ax = plt.subplots()

         # ax.bar(industry_name, x)
         plt.pie(x, labels = industry_name)

         plt.xticks(rotation=80)

         plt.title('Industry affected by Piracy ( Normalized Values)')

Out[52]: Text(0.5, 1.0, 'Industry affected by Piracy ( Normalized Values)')
```

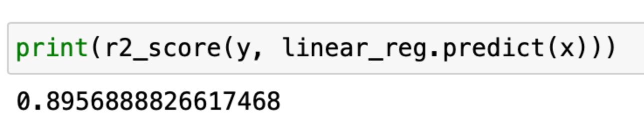
Industry affected by Piracy (Normalized Values)

We can obsereve that Hollywood/English Industry is the most affected by Piracy followed by Bollywood/Indian Industry.

Machine Learning Model



```
Mean Absolute Error:  299537.80400399014
Mean Squared Error:  117709157823.732
Root Mean Squared Error:  343087.68241330376
```



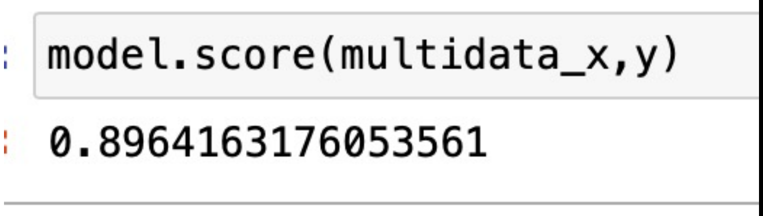
```
print(r2_score(y, linear_reg.predict(x)))
```

```
0.8956888826617468
```

Linear regression

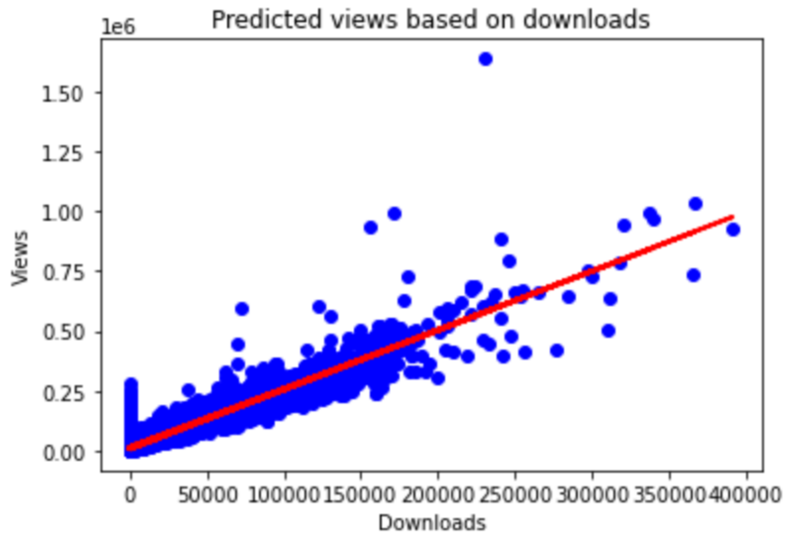
- Predicting the number of views based on the number of downloads.

Multi Linear Regression



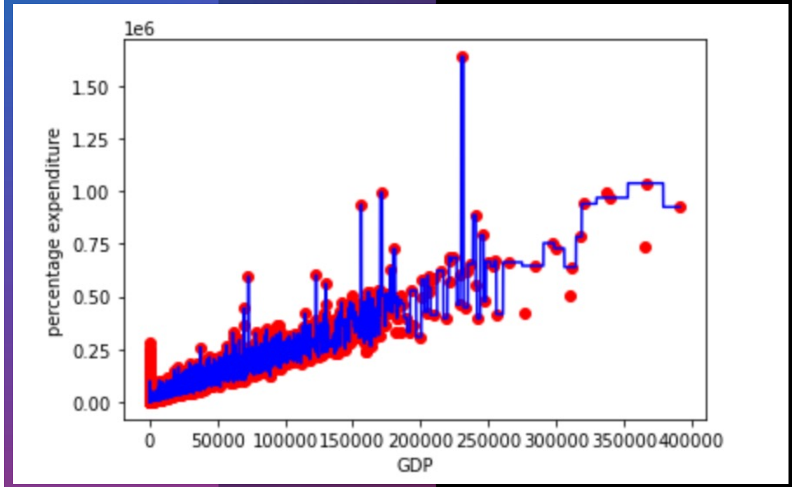
```
model.score(multidata_x,y)
```

```
0.8964163176053561
```

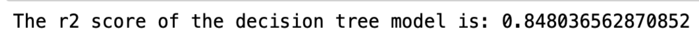


- Predicting the number of views based on several variables like number of downloads, run_time, language and IMDb-rating of a movie.

Decision Tree Regression



Predicting the number of views based on the number of downloads of the movie.



```
The r2 score of the decision tree model is: 0.848036562870852
```

Conclusion

- Piracy is a major issue in the movie industry, causing significant losses each year.
- Bollywood (Hindi) movies dominate the top 20 pirated movies.
- Variables like runtime, language, and views have a strong positive correlation with the number of downloads.
- Models such as K-Nearest Neighbors, Linear Regression, and Decision Trees can be used to predict downloads and views based on various input variables.
- It is crucial to find solutions to combat piracy while still providing audiences with access to quality films in a legal and ethical way.