Report for DM Mini project

Title:

Movie Data Analysis

Presented by:

Kavya Purushothaman 1514001

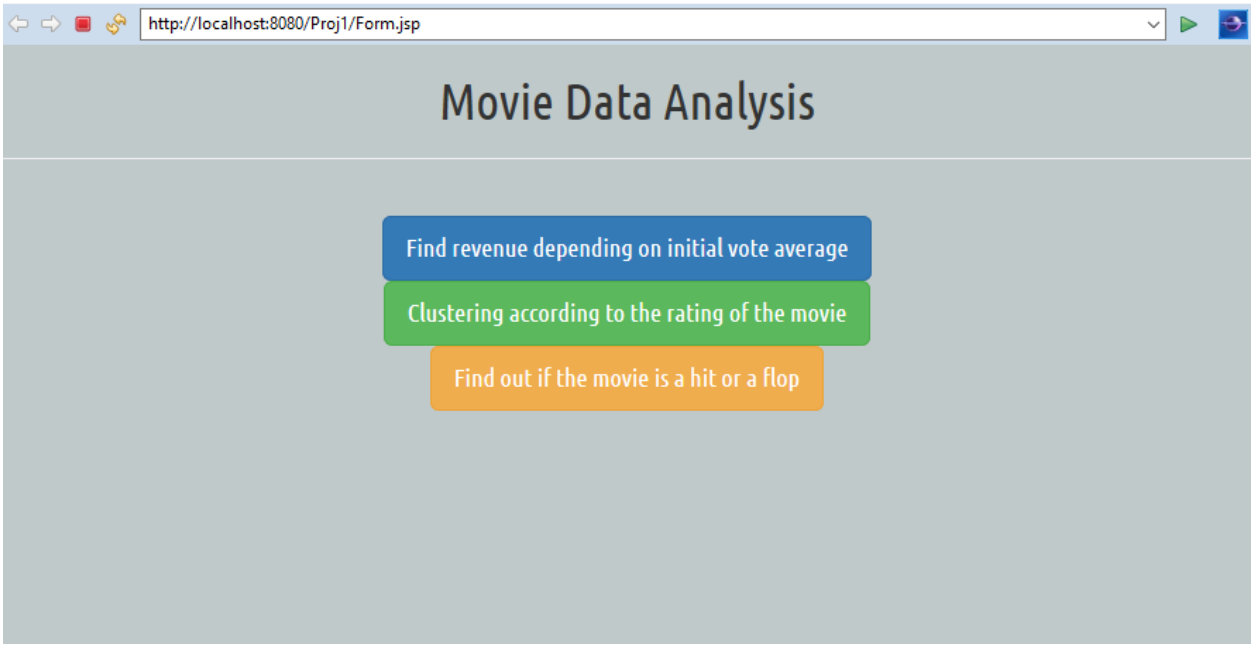
Introduction

Through this mini project, we try to analyze the revenue as well as whether the movie is a hit or a flop. We achieve the revenue through linear regression and using naïve Bayesian classification we can know whether the movie is a hit or a flob.

The various attributes use in the project are as follows:

1. Budget
2. ID
3. Language
4. Title
5. Popularity
6. Revenue
7. Runtime
8. Status
9. Vote_average
10. Vote_count
11. Rating
12. Result
13. Actor

We have prepared the following GUI for an end user to perform various action as displayed on the screen, the GUI is bootstrap enabled.



We have applied various algorithms to the above mentioned attributes, as shown below:

1. Linear Regression

- Functionalities:
  This method is used to predict the revenue of the movie depending on the initial vote average received, depending on the votes of the audience in the first week of the release, the end user would be able to predict the overall revenue of the movie.

- Algorithms Used:

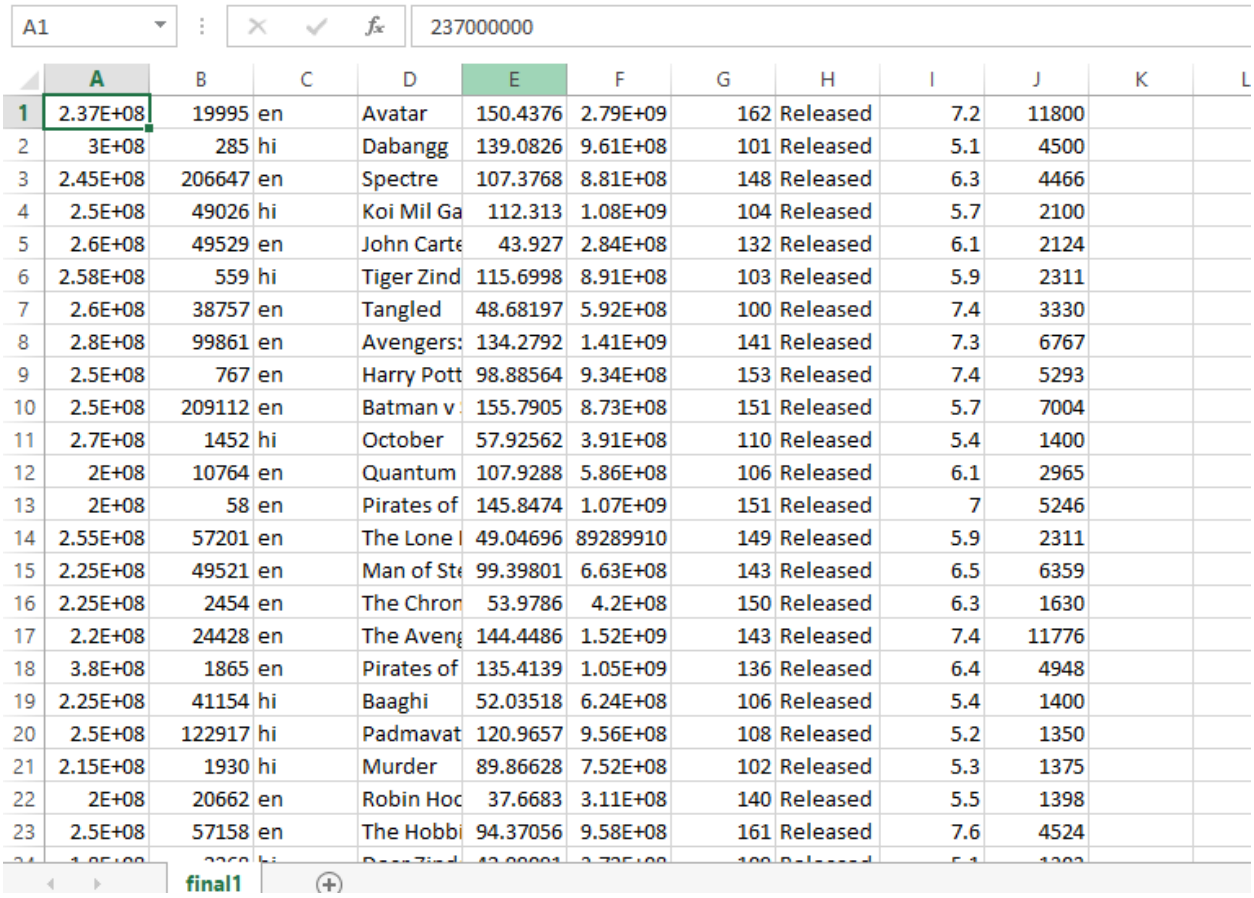
1. Linear Regression

- Resources Used:

1. JDK
2. Notepad
3. Xampp

- Result:
  In this experiment we have applied linear regression on the Vote_average field of the movie. We predicted the total revenue of the movie based on the votes of the audience in the initial week.

- Sheet Used:

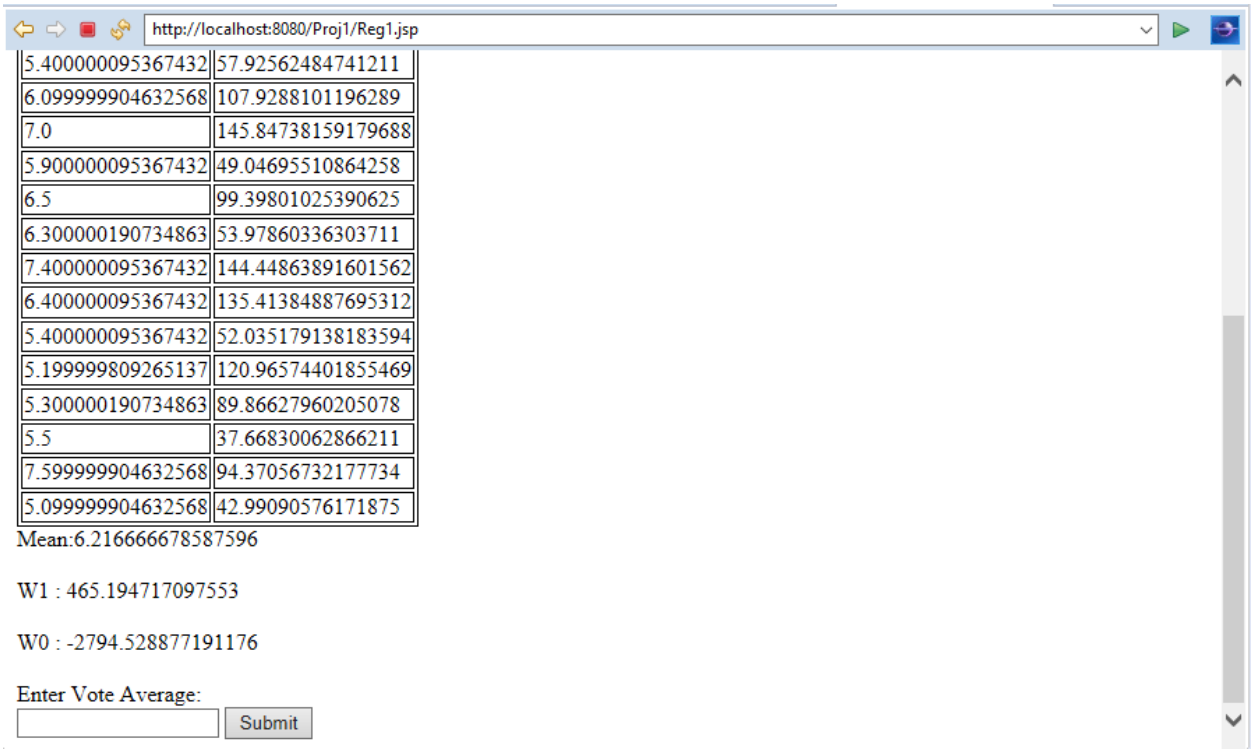


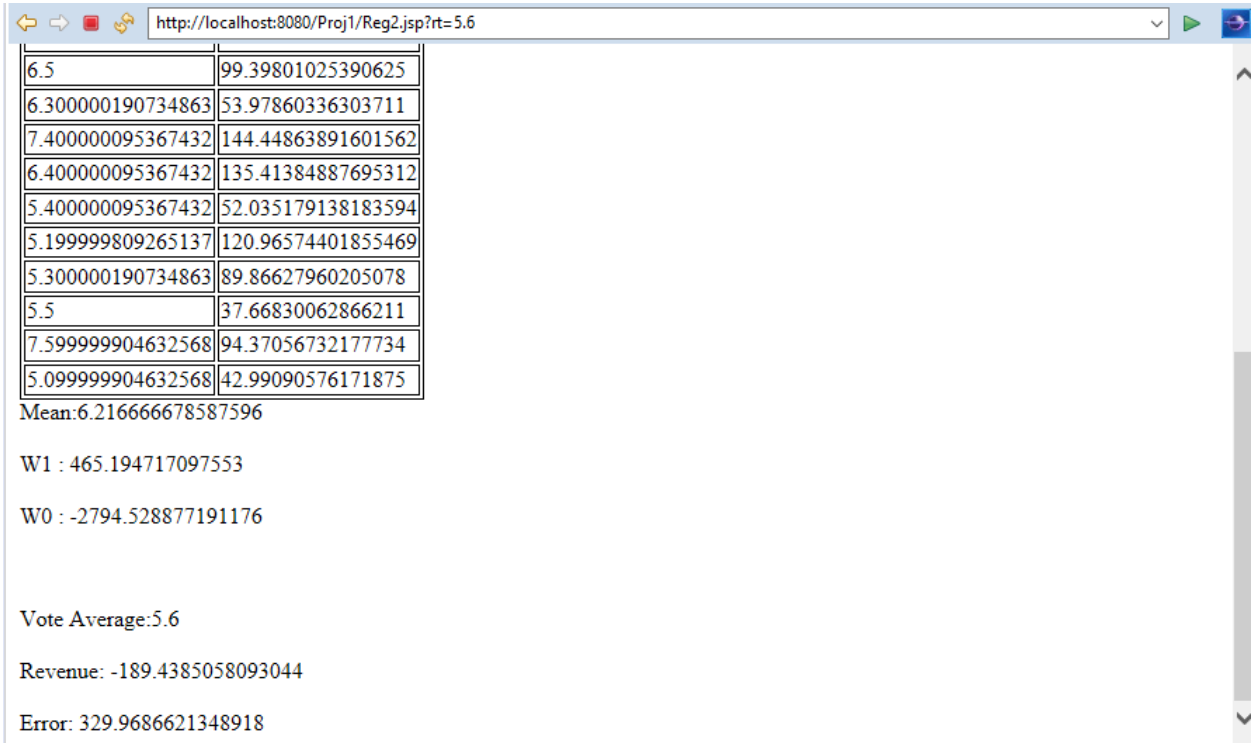
A1 | 237000000

	A	B	C	D	E	F	G	H	I	J	K	L
1	2.37E+08	19995	en	Avatar	150.4376	2.79E+09	162	Released	7.2	11800		
2	3E+08	285	hi	Dabangg	139.0826	9.61E+08	101	Released	5.1	4500		
3	2.45E+08	206647	en	Spectre	107.3768	8.81E+08	148	Released	6.3	4466		
4	2.5E+08	49026	hi	Koi Mil Ga	112.313	1.08E+09	104	Released	5.7	2100		
5	2.6E+08	49529	en	John Carte	43.927	2.84E+08	132	Released	6.1	2124		
6	2.58E+08	559	hi	Tiger Zind	115.6998	8.91E+08	103	Released	5.9	2311		
7	2.6E+08	38757	en	Tangled	48.68197	5.92E+08	100	Released	7.4	3330		
8	2.8E+08	99861	en	Avengers:	134.2792	1.41E+09	141	Released	7.3	6767		
9	2.5E+08	767	en	Harry Pott	98.88564	9.34E+08	153	Released	7.4	5293		
10	2.5E+08	209112	en	Batman v	155.7905	8.73E+08	151	Released	5.7	7004		
11	2.7E+08	1452	hi	October	57.92562	3.91E+08	110	Released	5.4	1400		
12	2E+08	10764	en	Quantum	107.9288	5.86E+08	106	Released	6.1	2965		
13	2E+08	58	en	Pirates of	145.8474	1.07E+09	151	Released	7	5246		
14	2.55E+08	57201	en	The Lone I	49.04696	89289910	149	Released	5.9	2311		
15	2.25E+08	49521	en	Man of Ste	99.39801	6.63E+08	143	Released	6.5	6359		
16	2.25E+08	2454	en	The Chron	53.9786	4.2E+08	150	Released	6.3	1630		
17	2.2E+08	24428	en	The Aveng	144.4486	1.52E+09	143	Released	7.4	11776		
18	3.8E+08	1865	en	Pirates of	135.4139	1.05E+09	136	Released	6.4	4948		
19	2.25E+08	41154	hi	Baaghi	52.03518	6.24E+08	106	Released	5.4	1400		
20	2.5E+08	122917	hi	Padmavat	120.9657	9.56E+08	108	Released	5.2	1350		
21	2.15E+08	1930	hi	Murder	89.86628	7.52E+08	102	Released	5.3	1375		
22	2E+08	20662	en	Robin Hoc	37.6683	3.11E+08	140	Released	5.5	1398		
23	2.5E+08	57158	en	The Hobbi	94.37056	9.58E+08	161	Released	7.6	4524		

final1

- Output:


http://localhost:8080/Proj1/Reg1.jsp
5.400000095367432 57.92562484741211
6.099999904632568 107.9288101196289
7.0 145.84738159179688
5.900000095367432 49.04695510864258
6.5 99.39801025390625
6.300000190734863 53.97860336303711
7.400000095367432 144.44863891601562
6.400000095367432 135.41384887695312
5.400000095367432 52.035179138183594
5.199999809265137 120.96574401855469
5.300000190734863 89.86627960205078
5.5 37.66830062866211
7.599999904632568 94.37056732177734
5.099999904632568 42.99090576171875
Mean:6.216666678587596
W1 : 465.194717097553
W0 : -2794.528877191176
Enter Vote Average:
Submit


http://localhost:8080/Proj1/Reg2.jsp?rt=5.6
6.5 99.39801025390625
6.300000190734863 53.97860336303711
7.400000095367432 144.44863891601562
6.400000095367432 135.41384887695312
5.400000095367432 52.035179138183594
5.199999809265137 120.96574401855469
5.300000190734863 89.86627960205078
5.5 37.66830062866211
7.599999904632568 94.37056732177734
5.099999904632568 42.99090576171875
Mean:6.216666678587596
W1 : 465.194717097553
W0 : -2794.528877191176
Vote Average:5.6
Revenue: -189.4385058093044
Error: 329.9686621348918

2. Data classification

- Functionalities:

  Bayesian classifier can predict membership probabilities such as the probabilities that a sample belongs to a particular class or groupings.

- Algorithms Used:

  Naïve Bayesian Classification Algorithm

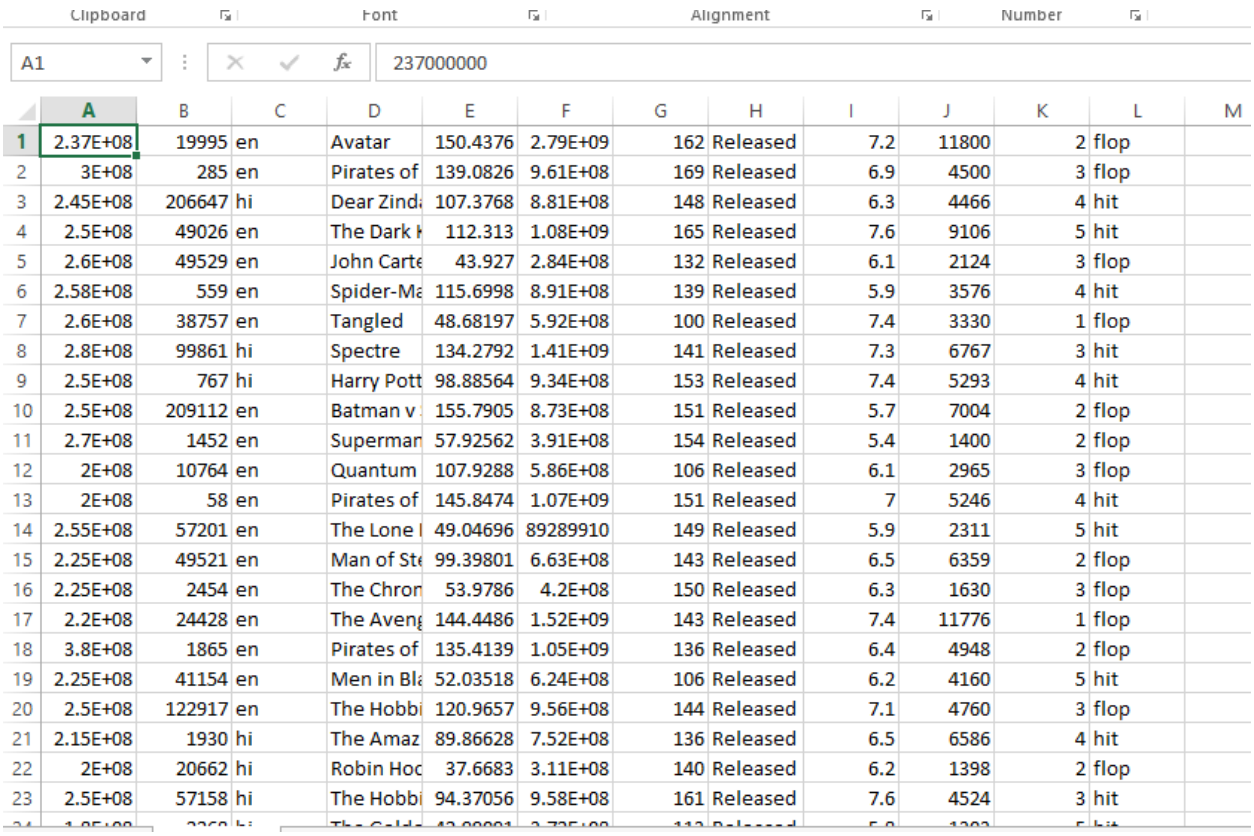
- Resources Used:

  1. JDK
  2. Notepad
  3. Xampp
  4. Microsoft Excel

- Result

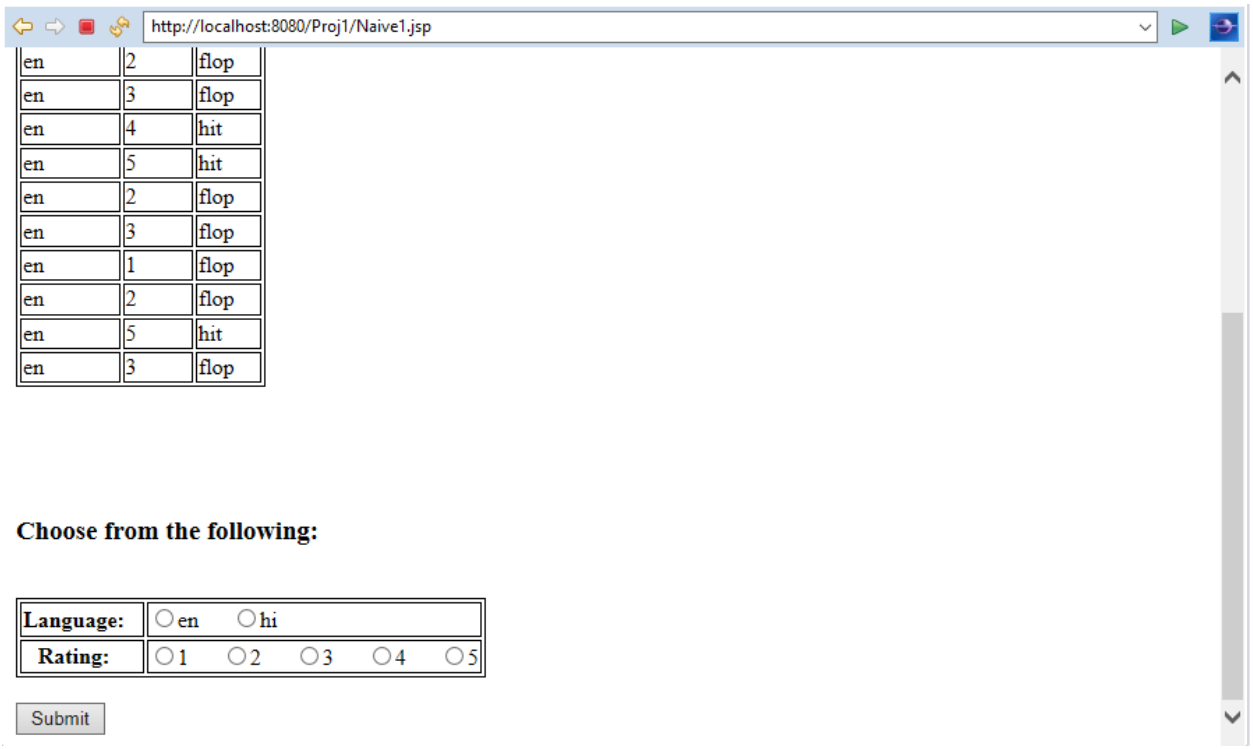
  In this experiment, we have calculated the probabilities of whether the movie released is a flop or a hit depending upon the language of the movie and also the rating of the particular movie. When a user fills out the form with the movie details, we use this algorithm to calculate the probability of this movie being a hit or a flop. If the probability of the movie being a hit is greater than the movie being a flop, then we assume that the movie is a hit and vice versa.

- Sheet Used:

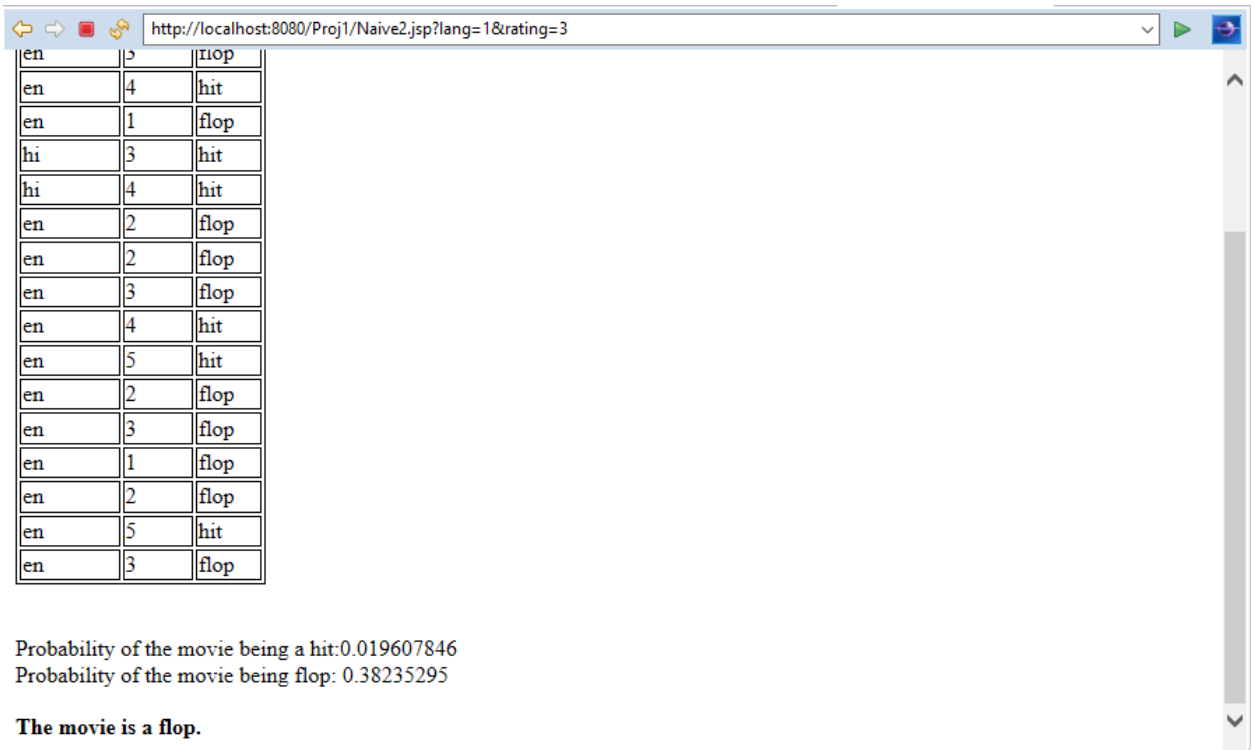


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	2.37E+08	19995	en	Avatar	150.4376	2.79E+09	162	Released	7.2	11800	2	flop	
2	3E+08	285	en	Pirates of	139.0826	9.61E+08	169	Released	6.9	4500	3	flop	
3	2.45E+08	206647	hi	Dear Zind	107.3768	8.81E+08	148	Released	6.3	4466	4	hit	
4	2.5E+08	49026	en	The Dark K	112.313	1.08E+09	165	Released	7.6	9106	5	hit	
5	2.6E+08	49529	en	John Carte	43.927	2.84E+08	132	Released	6.1	2124	3	flop	
6	2.58E+08	559	en	Spider-Ma	115.6998	8.91E+08	139	Released	5.9	3576	4	hit	
7	2.6E+08	38757	en	Tangled	48.68197	5.92E+08	100	Released	7.4	3330	1	flop	
8	2.8E+08	99861	hi	Spectre	134.2792	1.41E+09	141	Released	7.3	6767	3	hit	
9	2.5E+08	767	hi	Harry Pott	98.88564	9.34E+08	153	Released	7.4	5293	4	hit	
10	2.5E+08	209112	en	Batman v	155.7905	8.73E+08	151	Released	5.7	7004	2	flop	
11	2.7E+08	1452	en	Superman	57.92562	3.91E+08	154	Released	5.4	1400	2	flop	
12	2E+08	10764	en	Quantum	107.9288	5.86E+08	106	Released	6.1	2965	3	flop	
13	2E+08	58	en	Pirates of	145.8474	1.07E+09	151	Released	7	5246	4	hit	
14	2.55E+08	57201	en	The Lone I	49.04696	89289910	149	Released	5.9	2311	5	hit	
15	2.25E+08	49521	en	Man of Ste	99.39801	6.63E+08	143	Released	6.5	6359	2	flop	
16	2.25E+08	2454	en	The Chron	53.9786	4.2E+08	150	Released	6.3	1630	3	flop	
17	2.2E+08	24428	en	The Aveng	144.4486	1.52E+09	143	Released	7.4	11776	1	flop	
18	3.8E+08	1865	en	Pirates of	135.4139	1.05E+09	136	Released	6.4	4948	2	flop	
19	2.25E+08	41154	en	Men in Bla	52.03518	6.24E+08	106	Released	6.2	4160	5	hit	
20	2.5E+08	122917	en	The Hobbi	120.9657	9.56E+08	144	Released	7.1	4760	3	flop	
21	2.15E+08	1930	hi	The Amaz	89.86628	7.52E+08	136	Released	6.5	6586	4	hit	
22	2E+08	20662	hi	Robin Hoo	37.6683	3.11E+08	140	Released	6.2	1398	2	flop	
23	2.5E+08	57158	hi	The Hobbi	94.37056	9.58E+08	161	Released	7.6	4524	3	hit	

- Output:


http://localhost:8080/Proj1/Naive1.jsp
en 2 flop
en 3 flop
en 4 hit
en 5 hit
en 2 flop
en 3 flop
en 1 flop
en 2 flop
en 5 hit
en 3 flop
Choose from the following:
Language: en hi
Rating: 1 2 3 4 5
Submit

Case 1:


http://localhost:8080/Proj1/Naive2.jsp?lang=1&rating=3
en 4 hit
en 1 flop
hi 3 hit
hi 4 hit
en 2 flop
en 2 flop
en 3 flop
en 4 hit
en 5 hit
en 2 flop
en 3 flop
en 1 flop
en 2 flop
en 5 hit
en 3 flop
Probability of the movie being a hit:0.019607846
Probability of the movie being flop: 0.38235295
The movie is a flop.

Case 2:

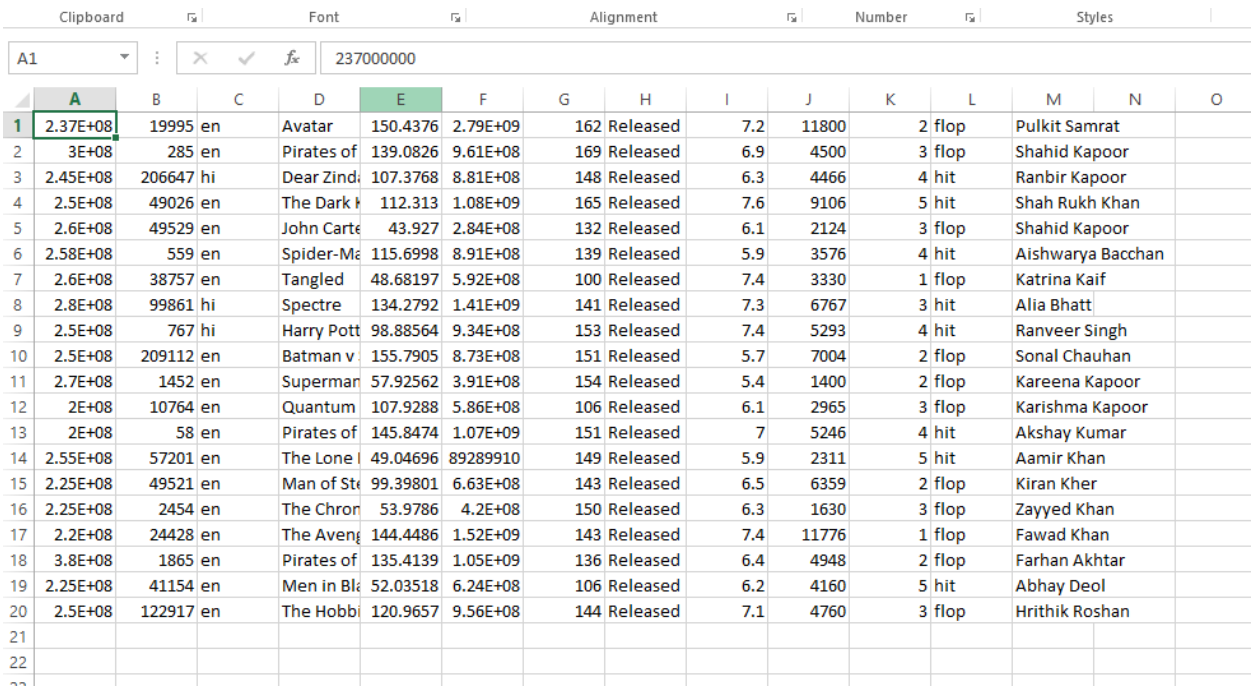
3. Clustering:

- Functionalities:

  We form various clusters using this algorithm, clusters basically have high intragroup similarity and very low intergroup similarity. In this project, we make clusters depending upon the rating of the movie and we also display the actors with the rating, the clusters so obtained will be different with each click as we take a random variable in order to calculate the mean.

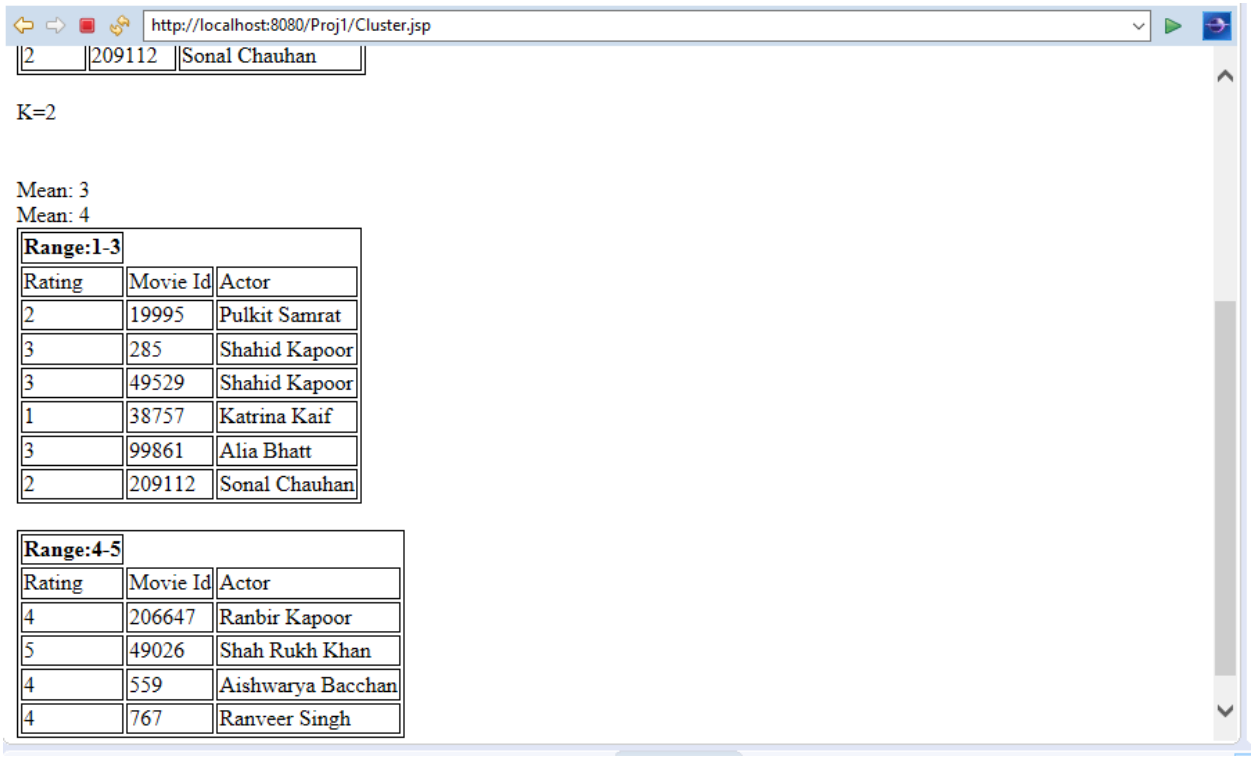
- Resources Used:
  1. JDK
  2. Notepad
  3. Xampp

- Sheet Used:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	2.37E+08	19995	en	Avatar	150.4376	2.79E+09	162	Released	7.2	11800	2	flop	Pulkit Samrat		
2	3E+08	285	en	Pirates of	139.0826	9.61E+08	169	Released	6.9	4500	3	flop	Shahid Kapoor		
3	2.45E+08	206647	hi	Dear Zind	107.3768	8.81E+08	148	Released	6.3	4466	4	hit	Ranbir Kapoor		
4	2.5E+08	49026	en	The Dark	112.313	1.08E+09	165	Released	7.6	9106	5	hit	Shah Rukh Khan		
5	2.6E+08	49529	en	John Carte	43.927	2.84E+08	132	Released	6.1	2124	3	flop	Shahid Kapoor		
6	2.58E+08	559	en	Spider-Ma	115.6998	8.91E+08	139	Released	5.9	3576	4	hit	Aishwarya Bacchan		
7	2.6E+08	38757	en	Tangled	48.68197	5.92E+08	100	Released	7.4	3330	1	flop	Katrina Kaif		
8	2.8E+08	99861	hi	Spectre	134.2792	1.41E+09	141	Released	7.3	6767	3	hit	Alia Bhatt		
9	2.5E+08	767	hi	Harry Pott	98.88564	9.34E+08	153	Released	7.4	5293	4	hit	Ranveer Singh		
10	2.5E+08	209112	en	Batman v	155.7905	8.73E+08	151	Released	5.7	7004	2	flop	Sonal Chauhan		
11	2.7E+08	1452	en	Superman	57.92562	3.91E+08	154	Released	5.4	1400	2	flop	Kareena Kapoor		
12	2E+08	10764	en	Quantum	107.9288	5.86E+08	106	Released	6.1	2965	3	flop	Karishma Kapoor		
13	2E+08	58	en	Pirates of	145.8474	1.07E+09	151	Released	7	5246	4	hit	Akshay Kumar		
14	2.55E+08	57201	en	The Lone	49.04696	89289910	149	Released	5.9	2311	5	hit	Aamir Khan		
15	2.25E+08	49521	en	Man of St	99.39801	6.63E+08	143	Released	6.5	6359	2	flop	Kiran Kher		
16	2.25E+08	2454	en	The Chron	53.9786	4.2E+08	150	Released	6.3	1630	3	flop	Zayyed Khan		
17	2.2E+08	24428	en	The Aveng	144.4486	1.52E+09	143	Released	7.4	11776	1	flop	Fawad Khan		
18	3.8E+08	1865	en	Pirates of	135.4139	1.05E+09	136	Released	6.4	4948	2	flop	Farhan Akhtar		
19	2.25E+08	41154	en	Men in Bla	52.03518	6.24E+08	106	Released	6.2	4160	5	hit	Abhay Deol		
20	2.5E+08	122917	en	The Hobbi	120.9657	9.56E+08	144	Released	7.1	4760	3	flop	Hrithik Roshan		
21															
22															

- Output:



Conclusion:

Hence, we have performed various actions on the movie dataset, we have predicted the total revenue (numeric attribute) and also whether the movie is a hit or a flop (categorical attribute) depending on other fields.