

# IMDB Data Analysis: Genre and Ratings Between 2005 and 2015

---

Name: Jasmine Hsieh

Course: Math 3210 Data Mining

# Goals

---

- Create a dataset with movies dated from 1873 to 2005, ratings at 2005, ratings at 2015, and genres.
- Look at all the movie genres and see if there's any interesting trend in the change of their ratings from 2005 to 2015


























# Data source 1

---

- Hadley Wickham, assistant professor at Rice University
- 58771 movies
- Title
- Year
- Length
- Ratings
- Votes
- Genres: Action, Animation, Comedy, Drama, Documentary, Romance, Short

# Index of ftp://ftp.fu-berlin.de/pub/misc/movies/database/

 [Up to higher level directory](#)

Name	Size	Last Modified	
 README	2 KB	5/29/2014	12:00:00 AM
 actors.list.gz	282592 KB	11/27/2015	7:48:00 PM
 actresses.list.gz	157865 KB	11/27/2015	7:51:00 PM
 aka-names.list.gz	7873 KB	11/27/2015	8:00:00 PM
 aka-titles.list.gz	8765 KB	11/27/2015	7:58:00 PM
 alternate-versions.list.gz	2441 KB	11/27/2015	8:02:00 PM
 biographies.list.gz	185640 KB	11/27/2015	7:58:00 PM
 business.list.gz	9888 KB	11/27/2015	8:02:00 PM
 certificates.list.gz	3337 KB	11/27/2015	7:59:00 PM
 cinematographers.list.gz	17971 KB	11/27/2015	7:53:00 PM
 color-info.list.gz	16435 KB	11/27/2015	8:00:00 PM
 complete-cast.list.gz	989 KB	3/16/2012	12:00:00 AM
 complete-crew.list.gz	581 KB	3/16/2012	12:00:00 AM
 composers.list.gz	14276 KB	11/27/2015	7:53:00 PM
 contrib		7/6/2005	12:00:00 AM
 costume-designers.list.gz	4773 KB	11/27/2015	7:53:00 PM
 countries.list.gz	16658 KB	11/27/2015	8:01:00 PM
 crazy-credits.list.gz	1283 KB	11/27/2015	7:57:00 PM
 diffs		11/28/2015	7:36:00 AM
 directors.list.gz	32200 KB	11/27/2015	7:52:00 PM
 distributors.list.gz	25768 KB	11/27/2015	8:03:00 PM
 editors.list.gz	22797 KB	11/27/2015	7:53:00 PM
 filesizes	2 KB	11/27/2015	7:48:00 PM
 filesizes.old	2 KB	11/27/2015	7:48:00 PM
 genres.list.gz	16249 KB	11/27/2015	8:00:00 PM

Data source 2- [imdb.com](http://www.imdb.com)

# Index Match In Excel

---

- 2005: 58771 movie titles / 2015: 233401 movie titles
- Formula: `=INDEX(B2:B233402, MATCH(F2, C2:C233402, 0))`
- Issues:
  1. extra space
  2. Name differences, e.g. 100 v.s. A hundred; “The” v.s. “ , the”
  3. identical titles resulting in pulling the wrong value
- Solutions:
  1. trim function
  3. eliminating identical titles on both 2005 and 2015 data

# Initial Match: 27777 titles

eliminated 3815 because votes in 2005 <10

---

1	title	year	length	rating05	rating15	votes05	Action	Animation	Comedy	Drama	Document	Romance	Short
2	90	2005	14	9.1	8.3	10	0	0	0	0	0	0	1
3	37 og et halvt	2005	101	5.6	4.8	27	0	0	1	1	0	0	0
4	500 Years Later	2005	106	9.3	6.9	17	0	0	0	0	1	0	0
5	5th World	2005	75	5.2	6.1	11	0	0	0	1	0	0	0
6	Abel Raises Cain	2005	82	7.8	7.5	13	0	0	0	0	1	0	0
7	Akoibon	2005	95	4.8	4.7	35	0	0	1	0	0	1	0
8	Alien Abduction	2005	90	1.9	2.6	73	0	0	0	0	0	0	0
9	Aliens of the Deep	2005	47	4.4	6.5	88	0	0	0	0	1	0	0
10	Allerzielen	2005	90	6.4	6.9	14	0	0	0	1	0	0	0
11	Alt for Norge	2005	92	6.1	5.7	32	0	0	1	0	1	0	0
12	America 101	2005	86	9.5	6.1	31	0	0	1	0	0	0	0
13	Americano	2005	95	8.1	6.3	31	0	0	1	0	0	1	0
14	Amu	2005	102	6.6	7.4	19	0	0	0	1	0	0	0
15	Anklaget	2005	103	6.7	7.1	33	0	0	0	0	0	0	0
16	Anthony Zimmer	2005	90	6.5	6.5	67	0	0	0	0	0	0	0

# Decide to Reconstruct Movie Genres

---

Genre	# of titles until 2005	Genre	# of titles until 2005
Action	13065	History	3593
Adventure	8901	Horror	5535
Adult	20381	Music	11300
Animation	14947	Musical	5519
Biography	2976	Mystery	4562
Comedy	60273	Romance	14812
Crime	11040	Sci-Fi	4497
Documentary	53583	Short	98699
Drama	69990	Sport	5168
Family	13956	Thriller	9154
Fantasy	5309	War	3971
Film-noir	268	Western	7041

# Index-Match existing 05-15 data with each genre

Genre	Total # of titles	Match
Animation	14947	1035
Biography	2976	363
Comedy	60273	5207
Fantasy	5309	558
Horror	5535	546
Romance	14812	2064
Sci-Fi	4497	734

Excel Macro was used to make the repetition of steps faster



# Big Issue Spotted!

---

- Since I only matched the data with their names, movies sharing the same name but have different genres result in misplacement.
- For example:
  - “Am See” (2001) is a German movie with Drama, Adventure, and Thriller genres. But it was also wrongly labeled as “Short” genre because there is a 11 minute short film with the same name.
- Is about 82% accuracy rate good enough?

# Back to the Beginning!

---

- Import IMDB ratings.list into Excel -> movie15
- Import 2005 ratings.tab -> movie05
- Index and Match function with **double criteria** (name + year)
- Formula: =INDEX(B:B, MATCH(1,(F2=C:C)\*(G2=D:D),0))
- Rating15 match: 37919 data
- Eliminate vote05 <10: 30368
- **Match vote15** <- maybe someone will find it beneficial!

# Example

---

title	year	length	rating05	rating15	vote05	vote15
Star Wars	1977	125	8.8	8.7	134640	805027
Pulp Fiction	1994	168	8.8	8.9	132745	1217498
Fight Club	1999	139	8.5	8.9	112092	1233907
American	1999	121	8.5	8.4	109991	763139
Star Wars:	1980	129	8.8	8.8	103706	734407
Saving Priv	1998	170	8.3	8.6	100267	806448
Schindler's	1993	195	8.8	8.9	97667	795371
Raiders of	1981	115	8.7	8.5	93511	609055
Gladiator	2000	155	8	8.5	92495	902224
Bravehear	1995	177	8.3	8.4	92437	680591
Memento	2000	113	8.7	8.5	90317	783353
Titanic	1997	194	6.9	7.7	90195	734502

Next step: Redo the match with 24 genres again.



# I don't have Hermione's time turner!!!

---

- It takes up to 3 hours for my computer to finish Index-Match double criteria (name and year) for 1 genre.

How can I possibly finish 24 genres?

- Solution: Do the original name match first, eliminate those titles that don't have any match, and then do the double criteria index match with the remaining data.
- **BIG TIME SAVER!**

# Example: Horror genre

title	year	length	rating05	rating15	vote05	vote15	namemat	genre
Frost	1997	270	3.4	7.7	18	53	Horror	#N/A
Giorgino	1994	177	6.4	7.5	118	686	Horror	Horror
Beloved	1998	172	5.6	5.8	1934	5979	Horror	Horror
Canadian Mounties	1953	167	5.3	5.5	13	112	Horror	Horror
Chandramukhi	2005	166	7.1	6.9	28	2848	Horror	Horror
100 Days	1991	161	5.8	6.5	59	592	Horror	Horror
Black Friday	2004	161	8.5	8.6	21	7078	Horror	#N/A
Aliens	1986	154	8.3	8.4	63961	453322	Horror	Horror
Fear of the Dark	2001	150	5.5	4.6	10	65	Horror	Horror
Forever My Love	1962	147	6	7.4	20	98	Horror	#N/A
Dawn of the Dead	1978	139	7.7	8	12621	85559	Horror	Horror
Teito monogatari	1988	135	6.2	6.3	58	236	Horror	#N/A

# Result

---

Genre	Old Match (name)	New Match ( name + year )
Animation	1035	749
Biography	363	396
Comedy	5207	126
Fantasy	558	617
Horror	546	1075
Romance	2064	1141
Sci-Fi	734	752

title	year	length	rating05	rating15	vote05	vote15	genre
8 to 4	1981	77	5.8	6.3	32	147	Adult
800 Fantasia	1979	82	5.1	6.1	30	107	Adult
Aerobisex	1983	85	4.9	6.3	12	26	Adult
Afternoon	1980	80	4.6	6.7	15	63	Adult
All About (	1978	90	4.3	6.9	10	35	Adult
All the Wa	1984	85	3.5	6.1	19	57	Adult
Amanda b	1981	95	5.7	6.3	56	177	Adult
American	1994	84	4.4	6.5	24	30	Adult
American	1981	79	4.5	7.4	11	74	Adult
Army Brat	1987	80	4.8	6.3	18	35	Adult
Aunt Peg	1980	80	5.4	6.5	28	102	Adult
Baby Face	1986	80	7.4	6.7	21	64	Adult
Babylon P	1979	77	5.9	6.3	17	107	Adult
Bad Girls	1981	82	6.4	6.7	59	195	Adult
Bad Girls	1994	99	4.8	5	2150	9243	Adult
Bad Girls I	1986	102	4.6	5.7	13	58	Adult
Barbara B	1977	87	6	6.7	56	367	Adult
Beauty	1981	87	4.1	6.9	11	41	Adult
Behind the	1972	72	5.3	6.1	380	1482	Adult
Bel Ami	1976	104	3.7	5	23	103	Adult
Between t	1985	76	7.2	6.6	31	77	Adult
Beverly Hi	1986	85	6.7	6.5	13	55	Adult
Blond & B	2001	95	8.1	7.8	28	108	Adult
Blonde An	1981	84	3.8	6.6	19	78	Adult
Blonde Fir	1978	86	6.2	7.2	13	66	Adult
Blonde Gc	1982	82	5.8	6.3	14	62	Adult
Blue Jean	1991	87	7.9	7.5	19	28	Adult
Blue Movi	1971	88	4.2	5.1	108	323	Adult
Bobby Sox	1996	90	7	6.8	47	82	Adult
Bodies in I	1983	73	5	7.2	11	39	Adult
Body Talk	1982	81	4.9	6.2	16	52	Adult

[▶ ...](#)
[Action](#)
[Adventure](#)
[Adult](#)
[Animation](#)
[Biography](#)
[Comedy](#)
[Crime](#)
[Documentary](#)
[Drama](#)
[Family](#)
[Fantasy](#)
[Film-Noir](#)
[History.](#)
[Horror](#)
[Music](#)
[Musical](#)
[Myste ...](#)

What the datasets look like.

# Trend On The Ranks Of Genre Ratings

Genre	Average rating 2005	Rank	Genre	Average rating 2015
Film-Noir	6.6	1	Documentary	6.88
Biography	6.59	2	Biography	6.82
Animation	6.56	3	Animation	6.8
Documentary	6.52	4	Film-Noir	6.76
History	6.51	5	History	6.76
War	6.39	6	Short	6.54
Short	6.31	7	War	6.53
Family	6.14	8	Music	6.52
Drama	6.13	9	Family	6.46
Music	6.12	10	Adult	6.44
Romance	6.06	11	Musical	6.39
Musical	6.03	12	Drama	6.32
Mystery	5.92	13	Romance	6.31
Crime	5.88	14	Mystery	6.18
Comedy	5.83	15	Crime	6.18
Sport	5.76	16	Sport	6.17
Western	5.74	17	Comedy	6.1
Fantasy	5.74	18	Western	6.1
Adventure	5.6	19	Fantasy	6.06
Thriller	5.5	20	Adventure	5.94
Adult	5.46	21	Thriller	5.83
Action	5.25	22	Action	5.61
Sci-Fi	5.07	23	Sci-Fi	5.42
Horror	4.75	24	Horror	5.25



# Significance Test: Student's T-test

---

- Evaluate the size of apparent effect you see in your data against the size of the random fluctuations present in your data.
- **Simple example: I throw coins 100 times, getting 45 times heads and 55 times tails. Can I conclude that it is more likely to get tails than heads?**
- T-value:  $\frac{\textit{strength of the "signal"}}{\textit{the surrounding noise}}$
- The bigger the difference, the bigger the T-value
- P-value: likelihood that there is not a reliable difference.
- If  $P=0.05$ , it means there's 5% chance that there is no reliable or statistically significant difference.

# Conclusion

---

- An average movie rating at 2005 is 5.91. An average movie rating at 2015 is 6.23. Average increase is 0.32 point (out of ten).
- Every genre's ratings at 05 and 15 are significantly increased.
- 13 movie genres follows the same trend as overall movies: Action, Adventure, Comedy, Crime, Documentary, Family, Fantasy, Music, Musical, Sci-Fi, Short, Thriller, and Western.
- War genre is the most statistically different. Least average increase of 0.13 point. Rank: 6 -> 7
- Horror genre is the secondly most statistically different. Average increase of 0.50 point (second highest), yet remains the lowest ranked.

Genre	T-value	P-value (two-tailed)	Average increase in rating	Genre	T-value	P-value (two-tailed)	Average increase in rating
Action	-1.73783	0.082406	0.36	History	2.034794	0.042608	0.25
Adventure	-0.62423	0.532595	0.34	Horror	-6.26058	5.42E-10	0.5
Adult	-4.68309	1.45E-05	0.98	Music	-1.88905	0.059503	0.4
Animation	2.898997	0.003847	0.24	Musical	-1.0821	0.279602	0.36
Biography	2.341585	0.019682	0.23	Mystery	2.296458	0.021922	0.26
Comedy	0.70634	0.481291	0.27	Romance	3.539629	0.000415	0.25
Crime	0.889689	0.373795	0.3	Sci-Fi	-0.72674	0.467601	0.35
Documentary	-0.23293	0.816291	0.36	Short	0.992313	0.322431	0.23
Drama	2.700048	0.007286	0.19	Sport	-2.14196	0.033148	0.41
Family	0.243598	0.807591	0.32	Thriller	-0.22352	0.823168	0.33
Fantasy	0.318501	0.750208	0.32	War	6.657406	5.81E-11	0.14
Film-noir	5.34596	2.32E-07	0.16	Western	-1.18006	0.238416	0.36



---

Comments/Questions?