<Final Project: IMDB Data Analysis >

MATH 3210 Jasmine Hsieh 4120903 12/14/2015

Executive Summary

As a response to the lack of movie rating data available on the internet, I constructed a movie data set with movie from 1983 to 2005, their ratings/votes at both 2005 and 2015, and their genres. The program I chose is Excel and my primary matching method is the Index-Match function with Multiple Criteria. My dataset consist of 25 worksheets. The first worksheet contains 30368 movie titles with 7 attributes: title, year, length, rating at 2005, rating at 2015, votes at 2005, and votes at 2015. Each of the following 24 worksheets represents a genre and contains data of movie titles labelled as such genre.

Besides building the dataset, I also conducted some analysis to the ratings of movies at 2005 and 2015. I concluded that there was no major change in the rankings of genres based on their average ratings. I did note, however, that drama genre had the most significant drop in its rank (from 9^{th} to 12^{th}) and that the adult genre had the most significant improvement (from 21th to 10^{th}).

I also did the Significant Test – Student's T-test to verify that there are significant differences in every movie genre's ratings from 2005 and 2015. Moreover, War genre is the most statistically different with the least average increase of 0.13 point and Horror genre is the secondly most statistically different, with the average increase of 0.50 point. Interestingly enough, with such a notable average increase, the Horror genre still remains the lowest ranked.

Problem Description

In my final project, I am to achieve two goals. First, I am to create a dataset with movies dated from 1873 to 2005, their ratings at 2005, their ratings at 2015, and genres. Secondly, I am to look at all the movie genres and see if there's any interesting trend in the change of their ratings from 2005 to 2015.

I have two data sources. The first one is a data set named "Movie Data Set" made by Hadley Wickham, a statistics professor at Rice University. The data set contains 58771 data

entries with the following list: title, year, budget, length, rating, votes, r1-10 (distribution of votes for each rating), MPAA rating, and genres (action, animation, comedy, drama, documentary, romance, and short). (Wickham, 2006)

The other data source is IMDB's "Alternative Interfaces" page which provides plain text data files of all the attributes you can think of that a movie would have. What's difficult is that since every attribute is a separate .list files, it takes quite some time to process to get what one would like.

Analysis Technique

<Part A>

In my initial steps, I import both data sources into Excel, then I use Index-Match function to match the movies titles in both data. Wherever there's a match, the rating at 2015 would be pulled and put besides the rating at 2005 column. Some underlying issues here include the name differences (e.g. "100" v.s. "a hundred"; and "The" v.s. ", the") and identical titles resulting in pulling the wrong values. I do the best I can to fix the issues and I end up eliminating all the identical movies titles on both data sets.

I also eliminate those data with less than 10 votes in 2005. In the end, I get 27777 matches with my data set looking like this:

	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

Figure 1 - Initial Match

My next step is to reconstruct the genre data. The reason I decided on rebuilding the genres is because there are only 7 genres in the data professor Wickham provided. It would have been alright if I just went along with what he had, but I also felt like it would be nice if I can build a complete data set since there's nothing like this available on the internet.

After I import the genre files from imdb.com into excel, this is the amount of data I get from each genre. And then I do Index-Match functions with the 05-15 data I just created and each genre.

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

However, a big issue is spotted after I finish the steps. There are 2388 films that are recognized as Short, but 424 actually don't belong in this genre. For example, there is a 90 minute film called "Dead People", and it was recognized as Short because there are two other short films with the same name. So that means my genre matches can possibly suffer from films that share the same name but are in different genres.

Before I decide how to deal with this, I try to find out the magnitude of this issue: the Short genre contains almost 99 thousand films so that it presumably has the most repeating names. And 1963 out of 2388 films that were recognized as Short do belong in this genre (this

can be confirmed by their lengths as Short genre contains films less than 45 minutes long). The accuracy rate here is 82% and presumably other genres should have better accuracy rate than this.

The question underlining here is: Is an 82% accuracy rate of a dataset good enough to provide reliable analysis? If not, how can I improve the accuracy? My answer is that even though 82% accurate doesn't sound bad, I want to make my dataset as accurate as possible. After some research, I learned how to match data with both the name and the year in Excel.

The second time matching my data sources, I get 30368 data entries (as opposed to the 27777 last time). Why do I get more data matches this time even though my criteria is more strict? It's because since I'm also taking the "year" into consideration, I don't have to eliminate those identical titles as I did before.

title	year	length	rating05	rating15	vote05	vote15
Star Wars	1977	125	8.8	8.7	134640	805027
Pulp Fictio	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

Figure 2 - what the data look like after my second attempt

Another huge time-consuming issue happens when I'm repeating this procedure with each of the 24 genres. 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? My solution is to 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. Excel Macros are also used to make the repetition of steps faster.

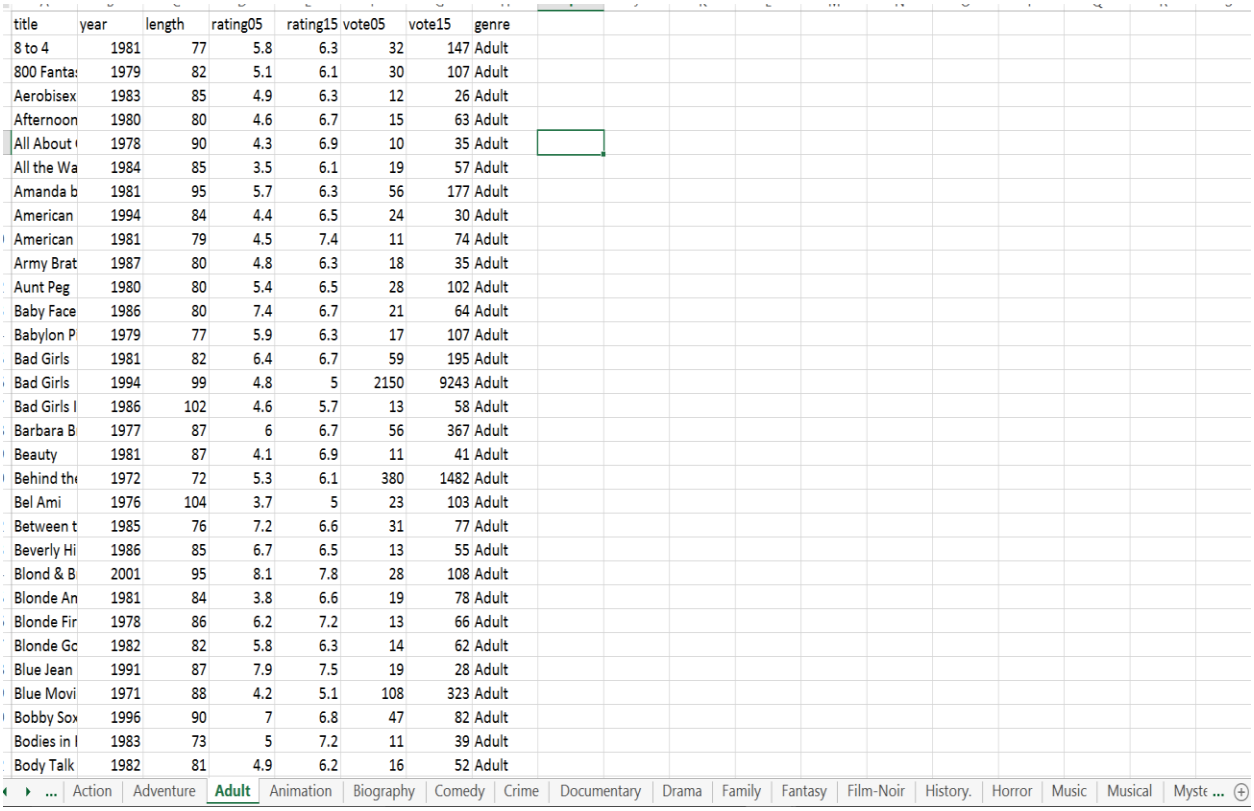
title	year	length	rating05	rating15	vote05	vote15	namemat	genre
Frost	1997	270	3.4	7.7	18	53	Horror	#N/A
Giorgino	1994	177	6.4	7.5	118	686	Horror	Horror
Beloved	1998	172	5.6	5.8	1934	5979	Horror	Horror
Canadian Mounties	1953	167	5.3	5.5	13	112	Horror	Horror
Chandramukhi	2005	166	7.1	6.9	28	2848	Horror	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

Figure 3 - A sample of the difference before and after double criteria match

The chart below is a sample of the differences between the old match method and the new match method. Most genres suffer a loss in the number of titles; however, there are some genres that benefit from the fact that there are movies with identical titles existing in the dataset. The extreme of such case is the horror genre. Its number of titles almost doubled and that makes me wonder if it implies that the movie producers aren't very creative when it comes to naming horror movies.

Genre	Old Match (only name)	New Match (name + year)
Comedy	5207	126
Crime	1574	1172
Documentary	1029	102
Drama	6030	330
Family	1218	942
Fantasy	558	617
Film-noir	141	203
History	307	349
Horror	546	1075
Music	472	456
Musical	505	638
Mystery	622	712
Romance	2064	1141

Up until this point, my data construction is completed. I have created a workbook with 25 worksheets. The first worksheet contains 30368 movie titles with 7 attributes: title, year, length, rating at 2005, rating at 2015, votes at 2005, and votes and 2015. Each of the following 24 worksheets represents a genre and contains data of movie titles labelled as such genre.



title	year	length	rating05	rating15	vote05	vote15	genre
8 to 4	1981	77	5.8	6.3	32	147	Adult
800 Fanta:	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 th	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

Figure 4 - What my final workbook looks like

<Part B>

The second part of my analysis, I'm going to do two things. First, I will do a simple analysis on the ranks of movie genres based on their average ratings in both 2005 and 2015, and report on any interesting finding. Second, I'm going to do a significance test (Student's T-test) within each genre and between genres, to see if there are significant differences in the change of ratings throughout these 10 years.

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

Figure 5 - Trend On The Ranks Of Genre Ratings

Above is the chart that I created to show the trend on the ranks of genre ratings. On the left side are the ranks of genres based on their average ratings in 2005 and on the right side is that of 2015. After reading both ranks, I notice that there really isn't much change of the genre ranks. The top five highly-rated genres (film-noir, biography, animation, documentary, history) in 2005 are still the top five currently; and the bottom three lowest rated genres (action, sci-fi, horror) also remains the bottom three lowest today.

Three biggest differences in the ranks of 2005 and 2015 are: 1. Film-Noir and Documentary genres switched places; 2. Drama's rank dropped the most (from the 9th to the 12th); 3. Adult's rank increased the most (from the 21st to the 10th).

Now we're getting into the significant test.

A T-test is a statistic test that checks if two means (averages) are reliably different from each other. Looking at the means, we can tell the difference. But we can't be sure if that's a reliable difference. Simple example: If I throw a coin 100 times, and I get 45 times heads and 55 times tails. Can I conclude that it is more likely to get tails than heads? No. It's just random fluctuations.

We normally would get two values after running the T-test: the T-value and the P-value. The T-value can be described as T = $\frac{Variance\ between\ groups}{Variance\ within\ groups}$ or $\frac{strength\ of\ the\ "signal"}{the\ surrounding\ noise}$. The bigger the T-value is, the bigger the difference. But how do we know if the T-value is big enough to show a difference? That's when we need to look at the P-value. The P-value tells us the likelihood that there is not really a difference.

Specifically, the P-value is the probability that the pattern of data in our sample could be produced by random data. If P=0.05, it means there's 5% chance that there is no real difference. The P value only depends on the size of the sample. Bigger sample makes it easier to detect differences. A good guideline is to have at least 30+ data points in each group.

I first run T-test for each genre's rating at 2005 and rating at 2015. The result is more than clear that there are significant difference within each genre between its rating at 2005 and 2015. Then I run each genre's difference in rating (2015 rating − 2005 rating) against the overall movies' difference in rating. Below is the result I get.

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

Figure 6 - T-test: each genre's difference in ratings against the overall's difference in ratings.

Action, Adventure, Comedy, Crime, Documentary, Family, Fantasy, Music, Musical, Sci-Fi, Short, Thriller, Western genres are not significantly different from the total movies. In other words, 13 out of 24 genres have the same trend in ratings in 10 years (averagely 0.32 point increase).

War genre is the most statistically different from the average movies with an average increase only 0.13 point. It is also has the least average increase. On the other hand, Horror genre is the secondly most statistically different from the average movies with an average increase of 0.50 point.

I think it's worth noting that the Adult genre actually has the seemingly biggest difference from the overall movies than any other genre (it has a 0.98 average increase). But the

T-test helps us determine that some of its difference is due to random fluctuation, so that it is actually the secondly statistically most different.

Assumptions

I assume that all the data I acquired online are accurate, And, I also assume that there are no two movies sharing the same name made in the same year. If such case existed, then there would be errors in my data.

Results

Based on the analysis and research carried out in the Analysis Technique section, I have the following observations.

- An average movie rating at 2005 is 5.91. An average movie rating at 2015 is 6.23. The average increase is 0.32 point (out of ten).
- There isn't much change of the rating ranks based on genres from 2005 to 2015. The top five highly-rated (film-noir, biography, animation, documentary, and history) and the bottom three lowest rated (action, sci-fi, and horror) remains the same.
- Two biggest differences in the ranks of movie genres' ratings between 2005 and 2015 are: 1. Drama's rank dropped the most (from the 9th to the 12th); 2. Adult's rank increased the most (from the 21st to the 10th).
- According to the T-test, 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, with the least average increase of 0.13 point. (Its rank dropped from 6^{th} to 7^{th})
- Horror genre is the secondly most statistically different, with the average increase of 0.50 point (second highest), yet it remains the lowest ranked.

Issues

I've fairly discussed this part in the Analysis Technique section.

References

The Internet Movie Database (IMDB). (2015, Nov 27). *Alternative Interfaces*. Retrieved from The Internet Movie Database (IMDB): http://www.imdb.com/interfaces

Wickham, H. (2006, June 5). *Movies dataset.* Retrieved from the website of Hadley Wickham: http://had.co.nz/data/movies/description.pdf