CS 542 Descriptives

Netflix full movie data

100+ million movie-user pairs over 5+ years. This is with 17k movies and 480k users

Most ratings are in 2005 and most users have at least 1 rating that year as well

|  |  |  |  |
| --- | --- | --- | --- |
| Rating Year | Ratings | Distinct Users | Distinct Movies |
| 1999 | 2,178 | 77 | 784 |
| 2000 | 924,443 | 8,227 | 4,470 |
| 2001 | 1,769,031 | 19,801 | 6,538 |
| 2002 | 4,342,871 | 51,732 | 8,418 |
| 2003 | 9,985,337 | 117,500 | 11,949 |
| 2004 | 30,206,574 | 259,407 | 15,506 |
| 2005 | 53,250,073 | 451,435 | 17,768 |

Most users joined in 04-05 but earlier users have had more time to rate more titles. Raters from each year 2000+ have had a chance to rate about every movie available

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| New User Rating Year | Ratings | Distinct Users | Distinct Movies | Ratings per User |
| 1999 | 48,494 | 77 | 7,653 | 630 |
| 2000 | 3,683,055 | 8,151 | 17,553 | 452 |
| 2001 | 4,959,402 | 13,096 | 17,677 | 379 |
| 2002 | 11,671,048 | 34,308 | 17,767 | 340 |
| 2003 | 19,253,412 | 69,779 | 17,767 | 276 |
| 2004 | 31,431,147 | 141,955 | 17,763 | 221 |
| 2005 | 29,433,949 | 212,823 | 17,750 | 138 |

Netflix movies have been rated in a more even distribution. Movies join Netflix or get their first rating in a year pretty evenly post 2000. The number of ratings is per film decrease with less years to rate as expected. But the quality of rating stays the same.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Netflix Rating Start Year | Movies | Ratings | Average Rating | Ratings per Movie |
| 1999 | 784 | 20,973,288 | 3.62 | 26,752 |
| 2000 | 3,686 | 22,243,461 | 3.60 | 6,035 |
| 2001 | 2,112 | 14,508,356 | 3.60 | 6,869 |
| 2002 | 2,148 | 12,126,635 | 3.59 | 5,646 |
| 2003 | 3,440 | 14,641,874 | 3.60 | 4,256 |
| 2004 | 3,424 | 13,498,144 | 3.62 | 3,942 |
| 2005 | 2,176 | 2,488,749 | 3.61 | 1,144 |

Since Netflix has a long list of movies I decided to check how people enjoyed older movies vs new ones. Seems like people are more likely to rate newer movies but still quality stayed constant

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Movies | Ratings | Average Rating | Ratings per Movie |
| Pre 1980 | 3,507 | 10,556,815 | 3.57 | 3,010 |
| 1980s | 2,136 | 13,490,153 | 3.60 | 6,316 |
| Early 90s | 1,789 | 11,583,022 | 3.61 | 6,475 |
| Late 90s | 3,384 | 20,350,332 | 3.61 | 6,014 |
| 2000 | 1,234 | 6,599,655 | 3.60 | 5,348 |
| 2001 | 1,184 | 7,241,888 | 3.61 | 6,116 |
| 2002 | 1,310 | 8,640,932 | 3.60 | 6,596 |
| 2003 | 1,271 | 9,576,604 | 3.61 | 7,535 |
| 2004 | 1,436 | 10,456,339 | 3.62 | 7,282 |
| 2005 | 512 | 1,983,802 | 3.62 | 3,875 |

Data on our 100k movie-user prediction sample. This includes 68k users and 6.8k movies

User Info

Most Users we are testing have rated a significant number of movies. The range will be useful in determining how many movies someone will be likely to rate in 06

|  |  |  |
| --- | --- | --- |
| **%** | **Ratings** | **Ratings in 05** |
| **100% Max** | 9,822 | 8,415 |
| **99%** | 1,786 | 1,206 |
| **95%** | 1,064 | 661 |
| **90%** | 787 | 474 |
| **75% Q3** | 443 | 254 |
| **50% Median** | 200 | 117 |
| **25% Q1** | 73 | 47 |
| **10%** | 24 | 12 |
| **5%** | 9 | 4 |
| **1%** | 2 | 0 |
| **0% Min** | 1 | 0 |

Some Users we are checking for multiple ratings which we will watch in how it pertains to model accuracy. But the vast majority we will only check for 1 movie rating

|  |  |
| --- | --- |
| **Entries** | **Frequency** |
| **1** | 49,183 |
| **2** | 12,369 |
| **3** | 3,972 |
| **4** | 1,539 |
| **5** | 637 |
| 6-10 | 614 |
| 10+ | 43 |

Most of the users we are checking for ratings entered their first rating in 2005 which is proportional to the main dataset. Also, almost all users had at least one rating in 2005

Movies

Most movies in our dataset to predict on are well reviewed (in # of times). Some near the top have been reviewed over 100k times (even in just 2005). Also while the prediction set is fairly disperse there are a few movies that will account for a large number of predictions

|  |  |  |  |
| --- | --- | --- | --- |
| **Level** | **Ratings** | **Ratings in 05** | **Entries in Data** |
| **100% Max** | 200,832 | 145,099 | 632 |
| **99%** | 104,031 | 59,300 | 126 |
| **95%** | 37,310 | 19,119 | 68 |
| **90%** | 18,801 | 9,042 | 41 |
| **75% Q3** | 4,218 | 2,079 | 14 |
| **50% Median** | 1,055 | 541 | 4 |
| **25% Q1** | 322 | 173 | 2 |
| **10%** | 154 | 84 | 1 |
| **5%** | 115 | 60 | 1 |
| **1%** | 73 | 29 | 1 |
| **0% Min** | 3 | 3 | 1 |

All movie info is proportional to the aggregate dataset