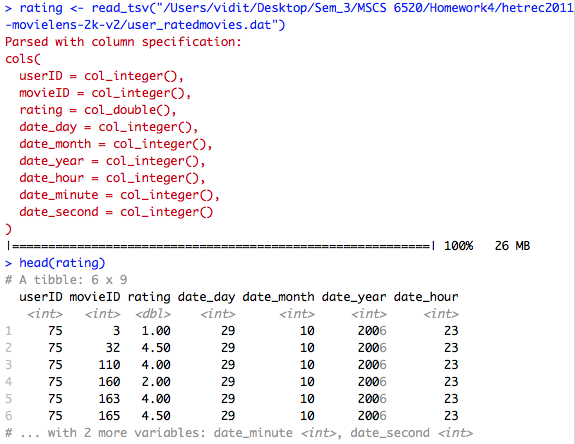
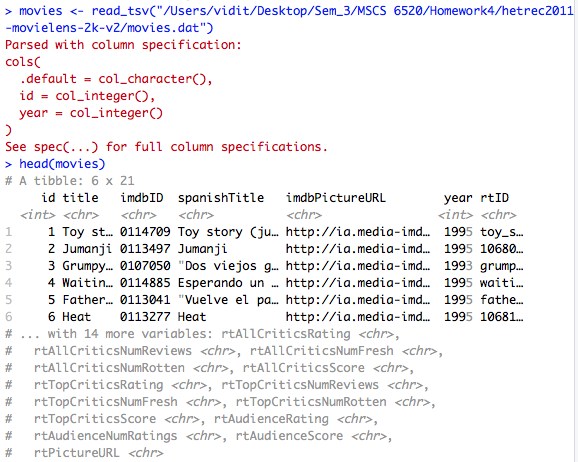
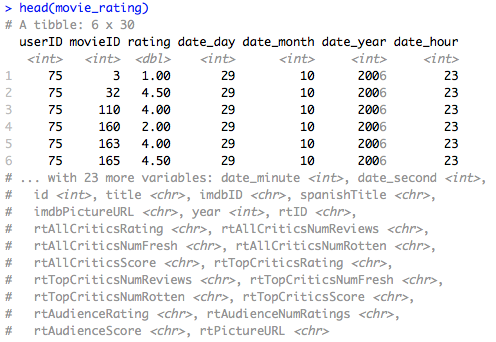
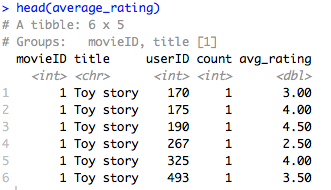
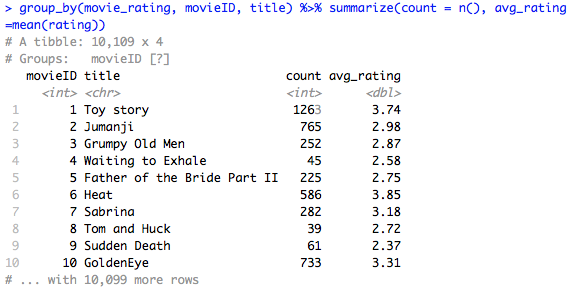
1. Read in the user\_ratedmovies.dat and movies.dat files. These are tab-separated value (TSV) files so you will need to use read\_tsv() function.

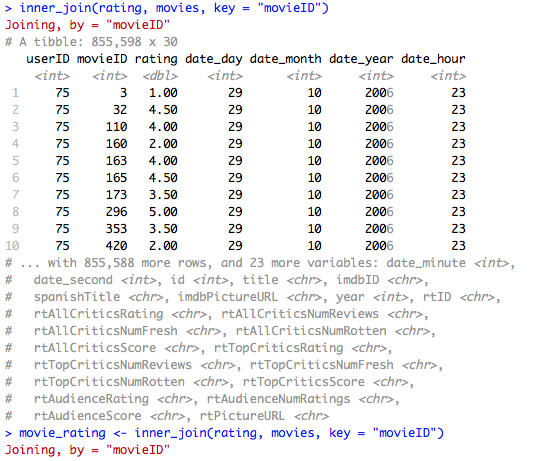


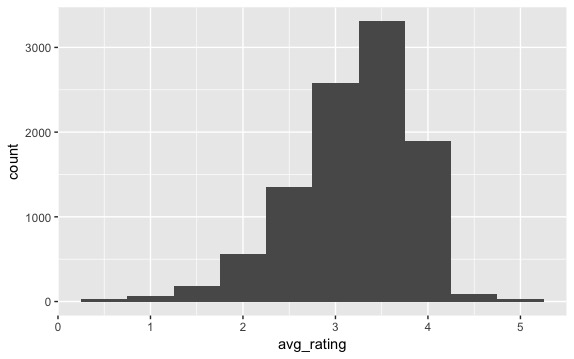


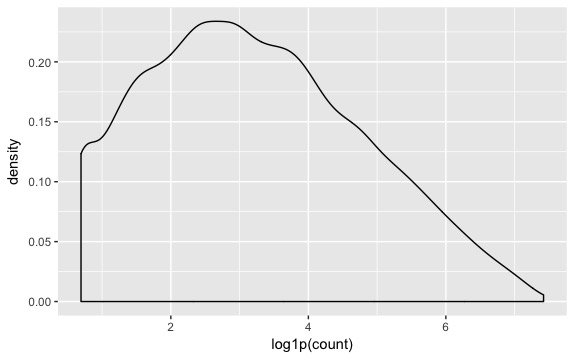
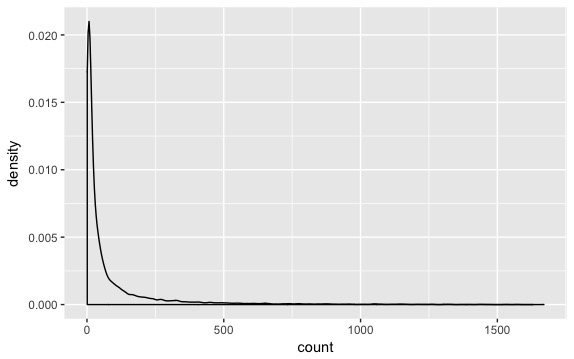
1. To join the two tibbles, we need to have a common column name. Use the mutate() and select() functions to rename the id column of the movies tibble to movieId so that it matches the movieId column of the user\_ratedmovies tibble. Join the user\_ratingsmovies and movies tibbles. 
2. Compute the average and number of ratings by the MovieLens users.



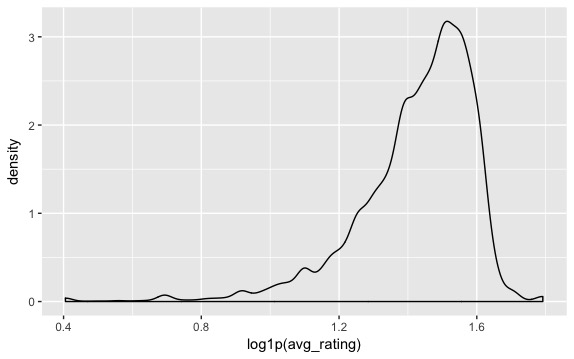


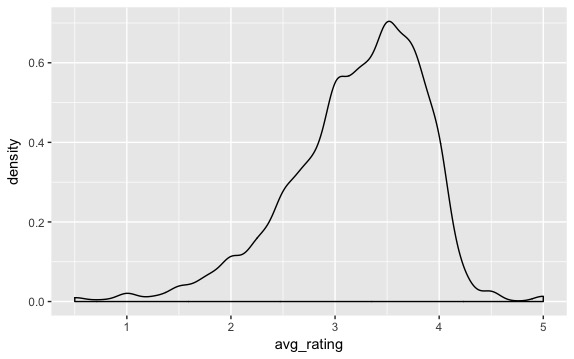


1. Plot the distributions of the average MovieLens ratings as a histogram. What seem to be the most common ratings given to movies? 
2. Plot the number of MovieLens ratings per movie (log scale) as densities. Does the number of ratings per movie seem uniformly distributed (each movie gets the same number of ratings) or skewed (some movies are rated more times than others)?



1. Plot the number of ratings (log scale) per movie vs the average rating per movie as a scatter plot. Is there a relationship between the two variables like we saw in the lecture? How do you interpret this relationship?





1. Build 4 models and evaluate their predictions using RMSE like we did in class: Baseline: use the overall average rating across all movies to predict the ratings by each user

Model 1: use the average rating per movie to predict the MovieLens ratings by each user Models 2 – 4: use the average rating per movie, filtered by the number of ratings, to predict the MovieLens ratings by each user Which model was best?

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **Rsquared** | **MAE** |
| Baseline | 1.0036010 | NA | 0.7875257 |
| 1 | 0.8841864 | 0.2243652 | 0.6820687 |
| 2 | 0.8795120 | 0.2282744 | 0.6785775 |
| 3 | 0.8781777 | 0.2265803 | 0.6772564 |
| 4 | 0.8767028 | 0.2243053 | 0.6759678 |

* Baseline: 1.0036
* Model 1 (Avg rating): 0.8841
* Model 2 (Avg rating and count >= 10): 0.8795
* Model 3 (Avg rating and count >= 20): 0.8781
* Model 4 (Avg rating and count >= 30): 0.8767

Model 1 is best.