Analysis of Movies (2005 – 2014)

By Arti Annaswamy

For:

Foundations of Data Science Workshop by MySlideRule

Data & Sources

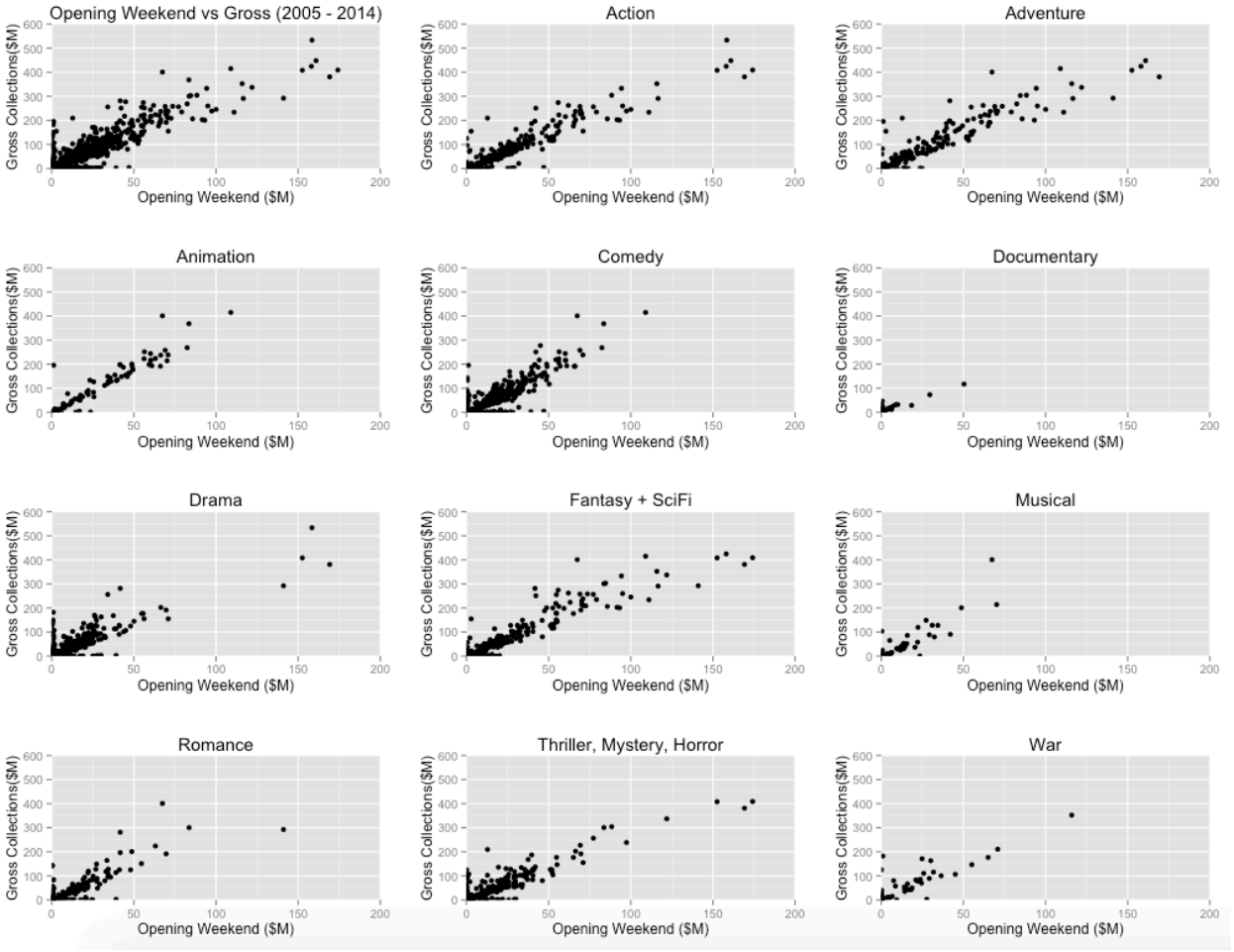
The following data, from their respective sources identified below, are used in this analysis.

Column Name	Example Data	Source
Year	2012	OMDb Data API
imdbId	tt1345836	MovieLens dataset
opening_weekend	160887295	IMDb Data Scraper
budget_est	250000000	IMDb Data Scraper
gross	448130642	IMDb Data Scraper
title	The Dark Knight Rises	OMDb Data API
Rated	PG-13	OMDb Data API
Released	7/20/12	OMDb Data API
Director	Christopher Nolan	OMDb Data API
Metascore	78	OMDb Data API
imdbRating	8.5	OMDb Data API
tomatoMeter	87	OMDb Data API
tomatoRating	8	OMDb Data API
tomatoUserRating	4.3	OMDb Data API
movieId	91529	MovieLens dataset
searchString	the+dark+knight+rises+2012+trailer	Manually created using movie Title and Year
ratingMean	3.995676293	Summarise() on Ratings.csv in MovieLens dataset
nRat	6129	
ratingMedian	4	
viewCount	66211508	YouTube Data API v3
likeCount	197311	YouTube Data API v3
dislikeCount	6112	YouTube Data API v3
favCount	0	YouTube Data API v3
commentCount	150827	YouTube Data API v3
Actors	Christian Bale, Gary Oldman, Tom Hardy, Joseph Gordon-Levitt	OMDb Data API
Plot	Eight years after the Joker's reign of anarchy, the Dark Knight is forced to return from his imposed exile to save Gotham City from the brutal guerrilla terrorist Bane with the help of the enigmatic Catwoman.	OMDb Data API
tomatoConsensus	The Dark Knight Rises is an ambitious, thoughtful, and potent action film that concludes Christopher Nolan's franchise in spectacular fashion.	OMDb Data API
gAct	1	List of genres in MovieLens dataset, manually parsed
gAdv	1	
gAnim	0	
gChi	0	
gCom	0	
gCri	1	
gDocu	0	
gDra	0	
gFant	0	
gFno	0	

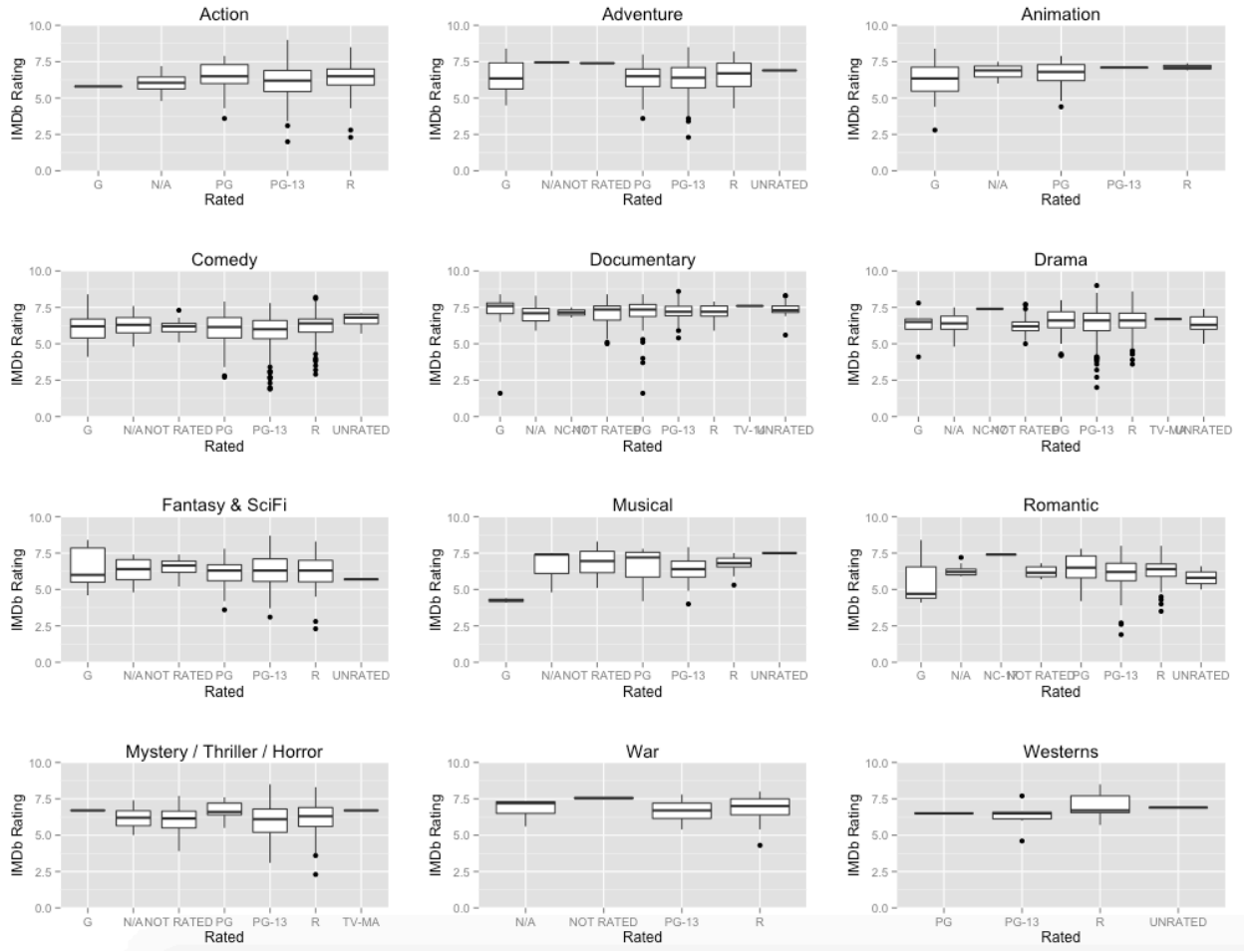
gHor	0	
gMus	0	
gMys	0	
gRom	0	
gSci	0	
gThr	0	
gWar	0	
gWes	0	
gNA	0	
gTotal	3	
runtime	165	OMDb Data API
ratedG	0	OMDb Data API, manually parsed
ratedNA	0	
ratedNC17	0	
ratedNOTRATED	0	
ratedPG	0	
ratedPG13	1	
ratedR	0	
ratedTV14	0	
ratedUNRATED	0	
ratedTVMA	0	
grp	4	Result of clustering analysis
Inflation.Factor	2.439129696	Factor of increase in Internet users from 2005 to 2014
viewCount_adj	27145546.26	Adjusted YouTube Views based on “inflation” factor above

Exploring the Data

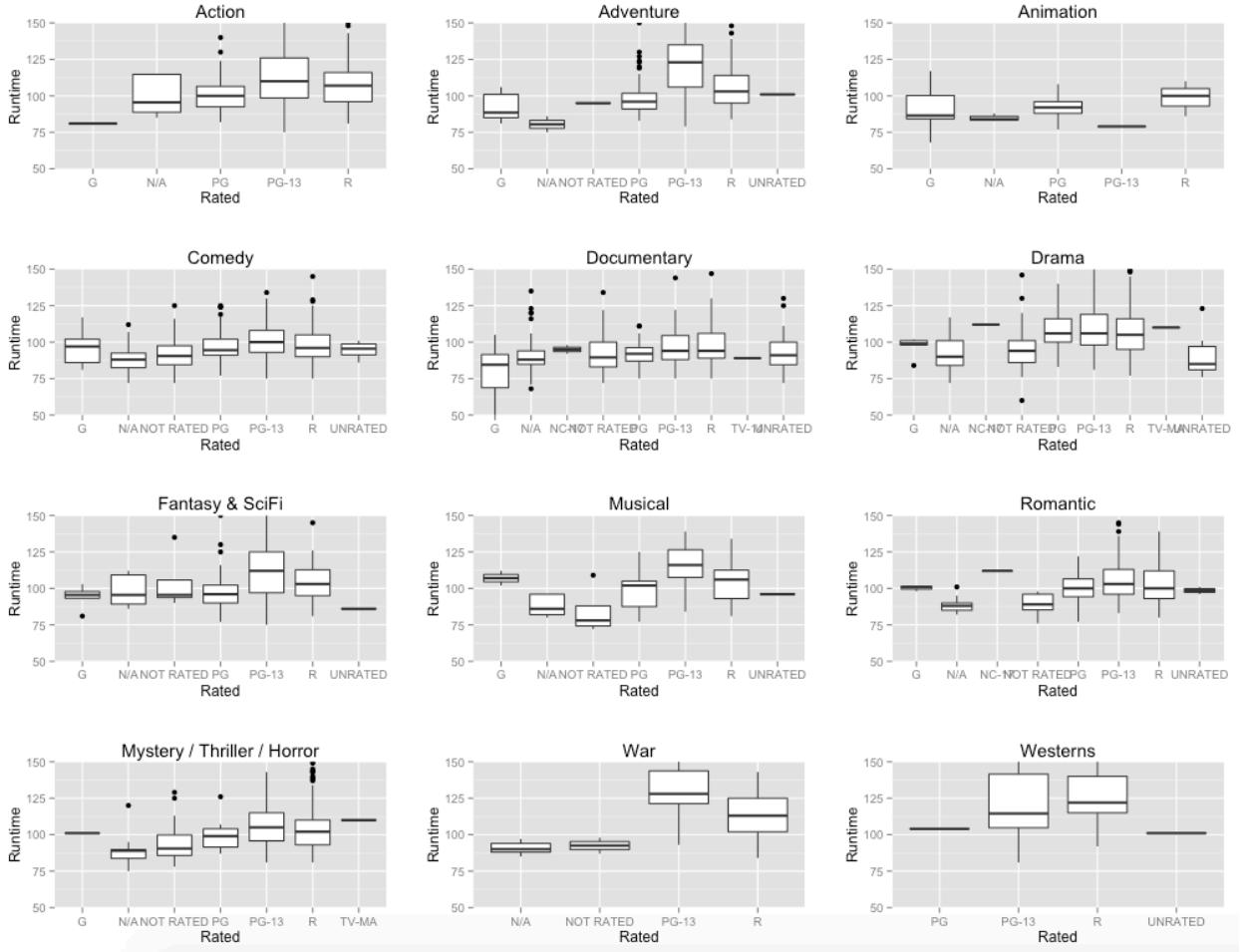
Plotting the Opening Weekend Collections vs the Gross (Total) collections shows an expected correlation. General wisdom in the movie industry assumes that the Opening Weekend collections are ~ 25% of the total Box Office collections for a film. This is a general thumb-rule, and the data supports a 3x-4x multiple.



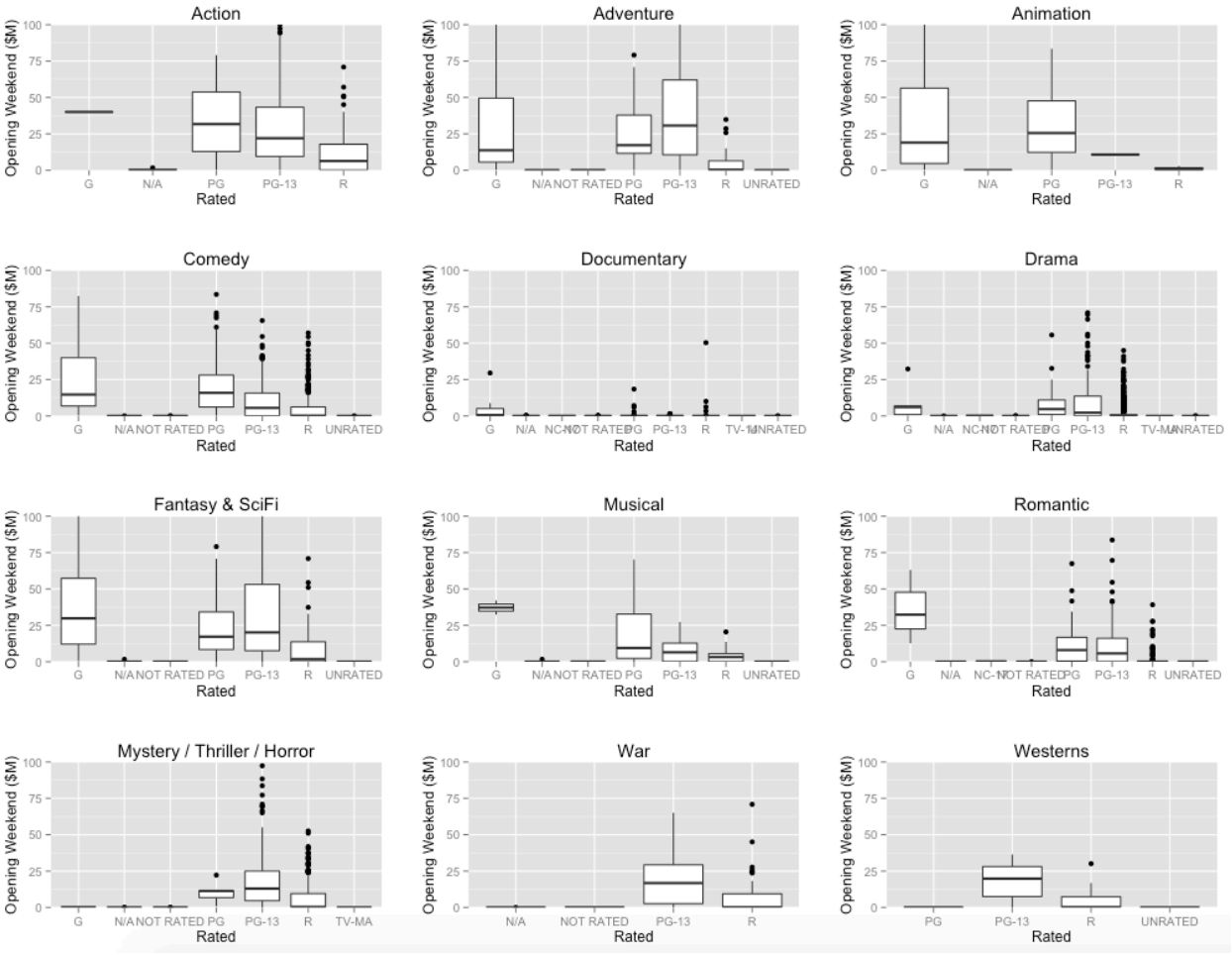
The boxplots below of IMDb Ratings by MPAA Rating, faceted by Genre show an average IMDb Rating that hovers around the 5.0 – 7.5 region. Documentaries on average are rating higher, while there is more variation in the Fantasy, SciFi, Musical and Romantic genres. Action and Adventure movies are rated consistently high as well.



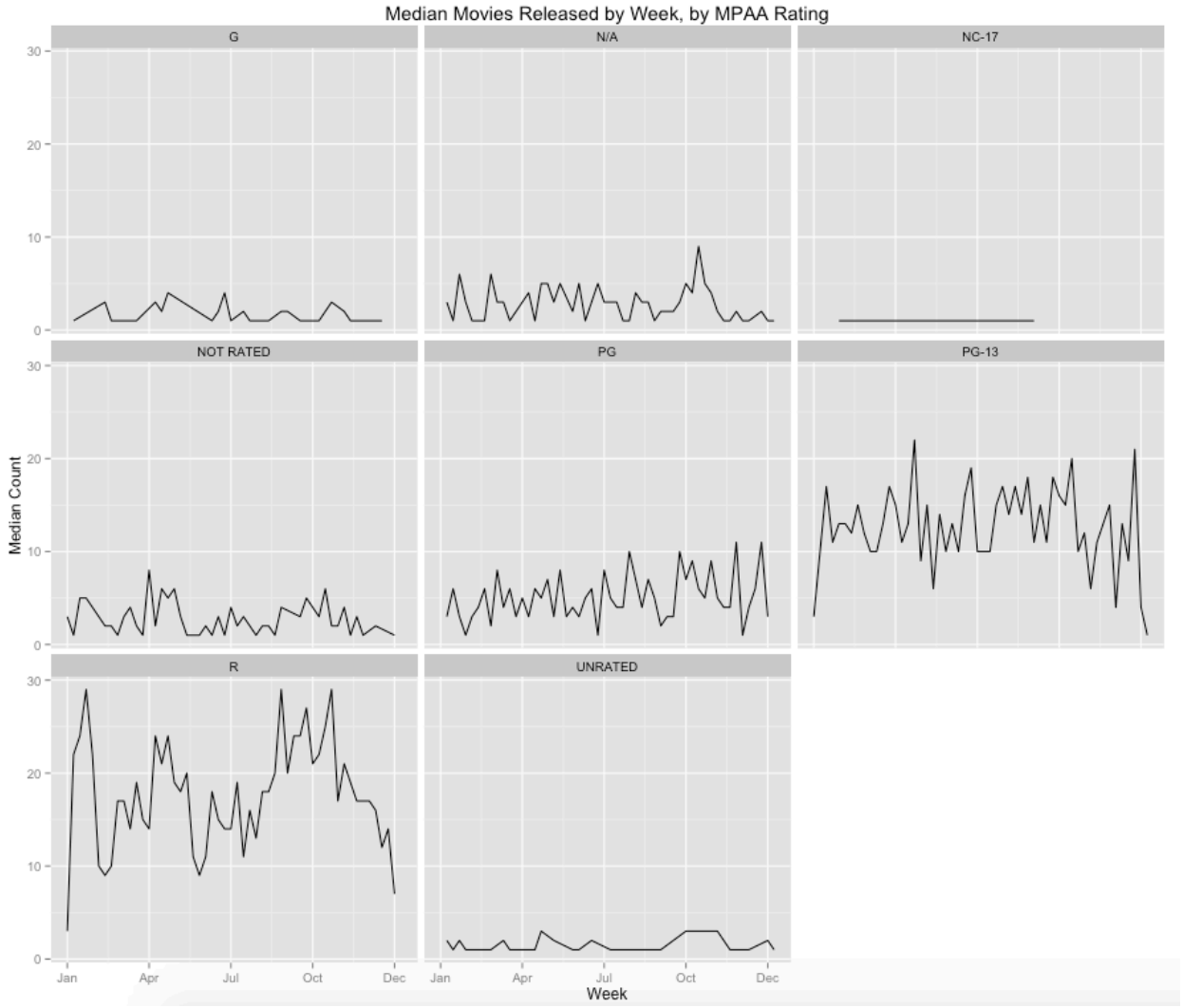
The boxplots below plot the average runtimes (in mins) of movies in various genres. Westerns, War films, and Action seem to have wide variation in times, and have some of the longer runtimes. Animation and Musicals are among the shorter movie lengths, the former definitely due to production costs directly associated with the length of a film.



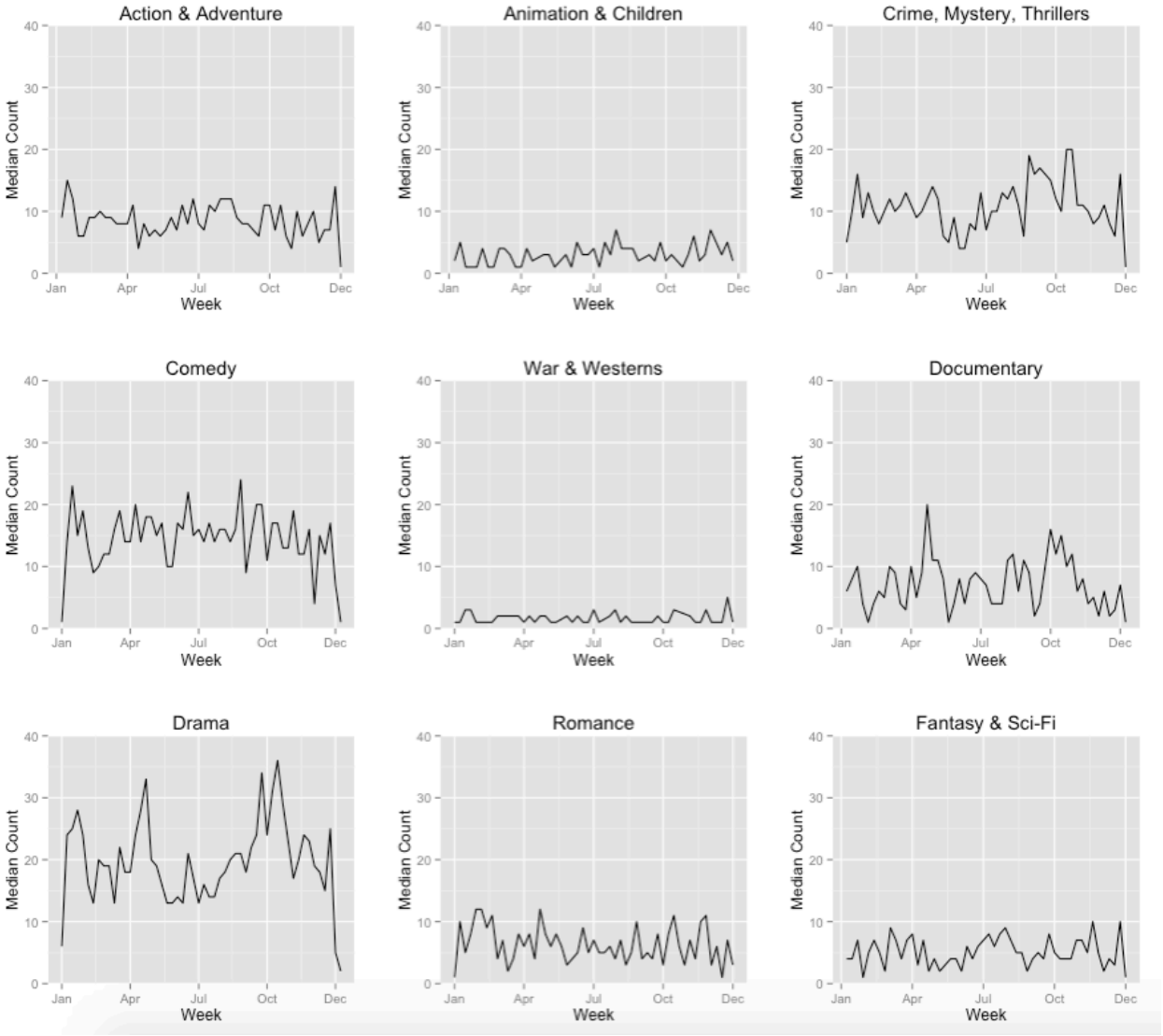
It is fairly unusual for movies to make more than $50M in its opening weekend, as shown in the boxplots below of Opening Weekend collections by Genre and MPAA rating. Occasionally, a movie becomes a breakaway hit (typically in the Action, Comedy, Drama or Thriller/Mystery categories) and appears in the 4th quartile in the boxplots below.



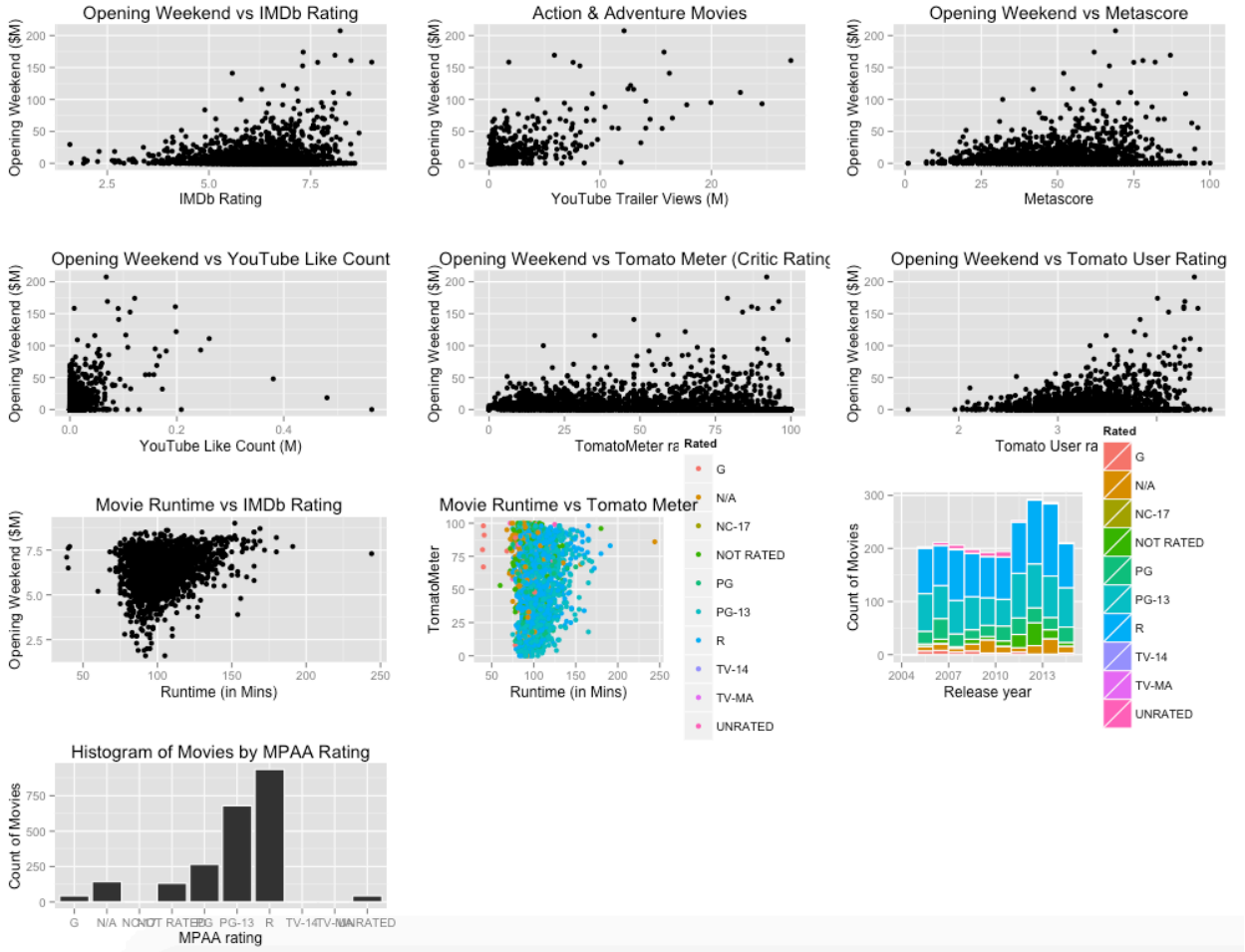
Plotting the median number of movies released each week from 2005 – 2014, across MPAA ratings shows PG and PG-13 movies released all year round and R-rated movies released mostly between the end of summer (Aug-Sep) and January every year (colloquially known as Awards season)



The chart below show the median number of movies released, across various genres. Action movies appear to spike in Dec-Jan, in addition to the traditional summer season. Crime/Mystery/Thriller films spike close to Halloween. Documentary and Drama films show spikes just before and after the summer season.



The charts below do not indicate any particularly strong correlations, but are interesting regardless. For example, the Metascore and Tomato Meter (Critic Ratings) seem to span a wide range without being particularly predictive of Opening Weekend grosses. User Ratings on IMDb and Tomato seem to show a slight positive correlation for highly rated films with high Opening Weekend grosses. Movie runtimes don't particularly correlate with ratings.



Analyses

1. Linear Regression

Performing a general linear regression of all numerical independent variables against the dependent variable, Opening Weekend, yields the following results:

```
Call:
lm(formula = opening_weekend ~ Year + Rated + Metascore + imdbRating +
    tomatoMeter + tomatoRating + tomatoUserRating + ratingMean +
    nRat + ratingMedian + viewCount + likeCount + dislikeCount +
    favCount + commentCount + gAct + gAdv + gAnim + gChi + gCom +
    gCri + gDocu + gDra + gFant + gFno + gHor + gMus + gMys +
    gRom + gSci + gThr + gWar + gWes + gNA + gTotal + runtime,
    data = movies2)

Residuals:
      Min        1Q    Median        3Q       Max
-68206293  -4569312   -145665   3609051 130312290

Coefficients: (2 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.485e+08  1.970e+08  -1.769  0.07697 .
Year             1.694e+05  9.825e+04   1.724  0.08490 .
RatedN/A        -6.912e+06  2.151e+06  -3.213  0.00133 **
RatedNC-17      -8.377e+06  6.944e+06  -1.206  0.22777
RatedNOT RATED  -6.780e+06  2.183e+06  -3.106  0.00192 **
RatedPG         -3.726e+06  1.948e+06  -1.913  0.05591 .
RatedPG-13      -3.193e+06  2.026e+06  -1.576  0.11518
RatedR          -8.470e+06  2.027e+06  -4.178 3.05e-05 ***
RatedTV-14      -1.564e+07  1.172e+07  -1.334  0.18218
RatedTV-MA      -8.033e+06  1.175e+07  -0.684  0.49416
RatedUNRATED    -6.629e+06  2.613e+06  -2.537  0.01125 *
Metascore        2.040e+04  4.923e+04   0.414  0.67862
imdbRating      -1.308e+06  5.698e+05  -2.296  0.02174 *
tomatoMeter      5.527e+04  3.857e+04   1.433  0.15201
tomatoRating    -1.598e+06  9.398e+05  -1.700  0.08928 .
tomatoUserRating 6.711e+06  9.373e+05   7.160 1.10e-12 ***
ratingMean      -2.684e+06  1.583e+06  -1.695  0.09025 .
nRat             2.804e+03  2.154e+02  13.017  < 2e-16 ***
ratingMedian     8.875e+05  1.324e+06   0.670  0.50284
viewCount        1.092e+00  8.460e-02  12.902  < 2e-16 ***
likeCount       -1.034e+01  1.717e+01  -0.602  0.54718
dislikeCount    -2.030e+03  2.867e+02  -7.079 1.94e-12 ***
favCount                NA         NA      NA       NA
commentCount     3.773e+02  7.064e+01   5.341 1.02e-07 ***
gAct             5.241e+06  8.398e+05   6.241 5.21e-10 ***
gAdv             6.157e+06  1.006e+06   6.120 1.11e-09 ***
gAnim            8.220e+06  1.637e+06   5.020 5.58e-07 ***
gChi             2.176e+06  1.531e+06   1.422  0.15529
gCom             7.206e+05  6.733e+05   1.070  0.28468
gCri             7.150e+05  9.123e+05   0.784  0.43330
gDocu           -4.540e+06  1.117e+06  -4.064 5.00e-05 ***
gDra            -3.563e+06  6.882e+05  -5.177 2.46e-07 ***
```

```
gFant             3.712e+06  1.127e+06   3.295  0.00100 **
gFno             -4.629e+06  4.471e+06  -1.035  0.30068
gHor              1.331e+06  1.090e+06   1.222  0.22203
gMus             -7.375e+05  1.436e+06  -0.514  0.60761
gMys             -7.471e+05  1.215e+06  -0.615  0.53855
gRom             -7.409e+04  7.472e+05  -0.099  0.92102
gSci              3.299e+06  1.097e+06   3.008  0.00266 **
gThr              3.156e+05  7.870e+05   0.401  0.68847
gWar              2.223e+05  1.597e+06   0.139  0.88932
gWes             -3.906e+06  2.699e+06  -1.447  0.14794
gNA              -1.001e+07  6.815e+06  -1.469  0.14200
gTotal                   NA         NA      NA       NA
runtime           1.382e+05  1.850e+04   7.472 1.14e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11530000 on 2203 degrees of freedom
Multiple R-squared:  0.5885,  Adjusted R-squared:  0.5807
F-statistic: 75.02 on 42 and 2203 DF,  p-value: < 2.2e-16
```

To improve this and to create a test and train set, we split the dataset into a 0.7 / 0.3 split, and remove the insignificant variables.

```
Call:
lm(formula = opening_weekend ~ imdbRating + tomatoUserRating +
    ratingMean + nRat + viewCount + dislikeCount + commentCount +
    gAct + gAdv + gAnim + gDocu + gDra + gFant + gSci + runtime +
    ratedG + ratedPG + ratedPG13, data = moviesTrain)

Residuals:
      Min        1Q    Median        3Q       Max
-69075871  -4734578    -93721   3615251 104064147

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.788e+07  2.856e+06  -6.262 4.91e-10 ***
imdbRating       -2.250e+06  6.336e+05  -3.551 0.000395 ***
tomatoUserRating  7.651e+06  1.043e+06   7.334 3.57e-13 ***
ratingMean       -1.656e+06  1.003e+06  -1.652 0.098837 .
nRat              2.757e+03  2.235e+02  12.336  < 2e-16 ***
viewCount         1.002e+00  9.924e-02  10.098  < 2e-16 ***
dislikeCount     -2.269e+03  3.334e+02  -6.806 1.43e-11 ***
commentCount      5.284e+02  7.419e+01   7.123 1.61e-12 ***
gAct              5.196e+06  9.126e+05   5.694 1.48e-08 ***
gAdv              5.754e+06  1.168e+06   4.925 9.35e-07 ***
gAnim             7.425e+06  1.809e+06   4.105 4.25e-05 ***
gDocu            -5.180e+06  1.069e+06  -4.846 1.39e-06 ***
gDra             -4.209e+06  7.143e+05  -5.893 4.65e-09 ***
gFant             2.430e+06  1.307e+06   1.859 0.063196 .
gSci              3.780e+06  1.267e+06   2.983 0.002903 **
runtime           1.390e+05  2.025e+04   6.865 9.58e-12 ***
ratedG            1.047e+07  2.380e+06   4.400 1.16e-05 ***
ratedPG           4.645e+06  1.041e+06   4.461 8.73e-06 ***
ratedPG13         5.054e+06  6.842e+05   7.387 2.44e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 11350000 on 1553 degrees of freedom
Multiple R-squared:  0.6004,  Adjusted R-squared:  0.5958
F-statistic: 129.6 on 18 and 1553 DF,  p-value: < 2.2e-16
```

This improves the R-squared on the model. Using this to predict the test set, we get the following results.

```
predLM <- predict(moviesLM3, newdata = moviesTest)
SSE <- sum((predLM - moviesTest$opening_weekend)^2)
SST <- sum((mean(moviesTrain$opening_weekend) -
moviesTest$opening_weekend)^2)
R2_lm = 1 - SSE/SST
R2_lm
```

```
0.5362176
```

2. Bag of Words – "Tomato Consensus"

For the next analysis, we select the 'Tomato Consensus' and 'Plot Summary' fields from the dataset. The 'Tomato Consensus' field contains a gist of Critics Reviews, while the Plot Summary is a snippet of the full summary available on IMDb.

After performing the data processing, removing the stopwords, calculating the frequencies and removing sparse words (0.99), we perform linear regression on the resulting data set.

```
Call:
lm(formula = opening_weekend ~ ., data = trainTC)

Residuals:
      Min        1Q    Median        3Q       Max
-46685951  -4898533    137767   4540907 115127640

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.594e+07  3.346e+06  -4.764 2.10e-06 ***
act             2.711e+06  1.816e+06   1.492 0.135810
action          1.751e+06  1.753e+06   0.999 0.318201
adapt          -2.652e+06  2.598e+06  -1.021 0.307649
add             1.914e+06  2.994e+06   0.639 0.522790
also            1.735e+06  2.793e+06   0.621 0.534540
ambiti          4.093e+06  2.901e+06   1.411 0.158484
american       -4.652e+06  2.570e+06  -1.810 0.070467 .
anim            1.497e+06  3.113e+06   0.481 0.630687
anoth          -1.377e+06  2.387e+06  -0.577 0.564124
appeal          1.571e+06  2.478e+06   0.634 0.526097
audienc        -3.734e+06  2.431e+06  -1.536 0.124820
beauti         -8.380e+05  2.145e+06  -0.391 0.696143
benefit        -1.634e+06  2.893e+06  -0.565 0.572406
best            1.862e+06  2.056e+06   0.905 0.365479
better          2.396e+06  3.056e+06   0.784 0.433107
big             6.555e+05  2.457e+06   0.267 0.789657
boast          -1.917e+05  2.391e+06  -0.080 0.936109
```

```
cant            -1.090e+06  2.381e+06  -0.458 0.647161
cast            -3.332e+05  1.368e+06  -0.244 0.807633
charact          1.322e+06  1.429e+06   0.925 0.355047
charm           -1.734e+06  1.678e+06  -1.034 0.301398
chemistri        5.067e+05  3.334e+06   0.152 0.879233
classic         -2.089e+06  3.075e+06  -0.679 0.497166
clever           4.499e+05  2.471e+06   0.182 0.855537
clich           -8.260e+05  2.334e+06  -0.354 0.723449
cliché           2.760e+05  2.671e+06   0.103 0.917709
come            -3.684e+06  2.613e+06  -1.410 0.158795
comedi           1.496e+06  1.406e+06   1.064 0.287498
comic            4.616e+06  2.735e+06   1.688 0.091607 .
compel           6.590e+06  2.288e+06   2.881 0.004030 **
complex          2.228e+06  3.103e+06   0.718 0.472937
dark             2.459e+06  2.227e+06   1.104 0.269668
debut           -2.206e+06  2.627e+06  -0.839 0.401370
deliv           -1.592e+06  2.061e+06  -0.773 0.439915
despit           7.902e+05  1.734e+06   0.456 0.648748
dialogu         -2.986e+06  2.399e+06  -1.245 0.213294
direct          -1.824e+05  1.735e+06  -0.105 0.916261
director        -1.242e+06  1.538e+06  -0.807 0.419563
documentari     -5.928e+05  1.930e+06  -0.307 0.758732
doesnt           1.626e+06  1.975e+06   0.823 0.410530
drama           -1.930e+06  1.610e+06  -1.198 0.230939
dull            -5.042e+06  2.871e+06  -1.756 0.079298 .
effect          -1.439e+06  2.079e+06  -0.692 0.488879
effort          -5.068e+05  2.955e+06  -0.171 0.863872
emot            -1.124e+06  2.052e+06  -0.548 0.583806
end              2.654e+06  2.662e+06   0.997 0.318984
enough           2.427e+06  1.426e+06   1.702 0.089061 .
entertain        5.050e+06  1.440e+06   3.508 0.000466 ***
even             2.118e+06  1.888e+06   1.122 0.261925
explor           4.357e+06  2.840e+06   1.534 0.125316
fail             2.637e+05  1.772e+06   0.149 0.881730
famili           3.594e+06  2.078e+06   1.729 0.083972 .
familiar         2.971e+05  2.379e+06   0.125 0.900645
fan              3.666e+05  2.106e+06   0.174 0.861866
fascin          -5.435e+05  2.495e+06  -0.218 0.827626
featur           2.545e+06  1.777e+06   1.432 0.152431
feel             4.788e+06  2.675e+06   1.790 0.073701 .
film            -4.540e+05  1.121e+06  -0.405 0.685532
filmmak          3.060e+05  2.713e+06   0.113 0.910203
find            -1.653e+06  2.174e+06  -0.760 0.447212
fine             2.045e+06  2.525e+06   0.810 0.418186
formula          1.744e+06  1.870e+06   0.933 0.351013
franchis         1.998e+07  3.048e+06   6.556 7.80e-11 ***
full            -2.000e+05  3.068e+06  -0.065 0.948024
fun             -7.758e+06  2.615e+06  -2.967 0.003059 **
funni           -8.315e+05  1.561e+06  -0.532 0.594469
gag              1.209e+06  2.672e+06   0.453 0.650890
genr            -8.127e+05  2.189e+06  -0.371 0.710429
get              3.781e+06  2.434e+06   1.554 0.120521
give            -9.784e+05  3.040e+06  -0.322 0.747647
good             8.848e+04  2.345e+06   0.038 0.969904
great           -6.410e+05  2.980e+06  -0.215 0.829731
heart           -3.987e+06  2.675e+06  -1.490 0.136387
horror          -2.493e+06  2.630e+06  -0.948 0.343375
```

```
human            5.013e+06  2.651e+06   1.891 0.058863 .
humor            1.431e+06  1.823e+06   0.785 0.432653
impress          1.124e+07  2.898e+06   3.879 0.000110 ***
insight         -3.333e+06  2.948e+06  -1.131 0.258411
inspir           3.229e+05  2.218e+06   0.146 0.884262
interest         2.072e+06  2.983e+06   0.695 0.487405
isnt            -8.957e+05  2.734e+06  -0.328 0.743279
just            -1.022e+06  2.346e+06  -0.436 0.663200
keep            -4.952e+05  2.882e+06  -0.172 0.863614
lack             2.770e+05  1.806e+06   0.153 0.878155
larg             5.100e+05  3.015e+06   0.169 0.865710
laugh            2.863e+06  1.938e+06   1.477 0.139889
lead            -9.725e+05  1.909e+06  -0.509 0.610563
less            -3.174e+06  2.648e+06  -1.199 0.230844
life            -1.699e+05  1.934e+06  -0.088 0.930009
likabl          -3.131e+06  3.136e+06  -0.999 0.318129
like            -1.878e+06  2.004e+06  -0.937 0.348752
littl            1.261e+05  1.744e+06   0.072 0.942390
live             5.738e+06  2.240e+06   2.561 0.010532 *
look            -1.283e+06  1.997e+06  -0.643 0.520569
love             5.139e+05  1.883e+06   0.273 0.784895
make             1.392e+06  1.381e+06   1.008 0.313410
man              6.567e+06  2.442e+06   2.689 0.007248 **
mani             1.538e+06  2.467e+06   0.623 0.533145
materi          -7.341e+06  3.284e+06  -2.235 0.025563 *
may             -9.968e+05  1.654e+06  -0.603 0.546804
messag           1.919e+06  2.708e+06   0.709 0.478676
might            4.455e+06  2.764e+06   1.612 0.107282
moment           4.215e+05  2.676e+06   0.157 0.874875
move             1.481e+06  3.150e+06   0.470 0.638326
movi            -2.437e+05  1.362e+06  -0.179 0.857978
much             1.660e+06  2.542e+06   0.653 0.513765
music           -7.838e+05  2.597e+06  -0.302 0.762875
narrat           1.717e+06  2.707e+06   0.635 0.525832
never            3.522e+05  2.662e+06   0.132 0.894767
new              2.458e+06  2.147e+06   1.145 0.252431
occasion        -9.092e+05  2.654e+06  -0.343 0.732000
offer           -2.153e+06  1.390e+06  -1.549 0.121579
often           -7.655e+05  3.214e+06  -0.238 0.811771
one             -3.145e+06  1.847e+06  -1.703 0.088848 .
origin           2.691e+06  2.238e+06   1.202 0.229511
over             1.963e+06  2.922e+06   0.672 0.501727
pace            -2.101e+06  2.707e+06  -0.776 0.437832
part             6.131e+06  2.668e+06   2.299 0.021681 *
perform          2.471e+05  1.037e+06   0.238 0.811727
play             4.524e+06  2.862e+06   1.580 0.114229
plot            -1.020e+06  1.407e+06  -0.725 0.468627
polit            2.550e+06  2.264e+06   1.126 0.260294
poor            -2.311e+06  2.886e+06  -0.801 0.423430
power           -1.612e+06  1.773e+06  -0.909 0.363436
predecessor      4.163e+06  2.782e+06   1.497 0.134747
predict         -1.635e+06  1.872e+06  -0.873 0.382673
premis          -2.767e+06  2.066e+06  -1.339 0.180744
prove            2.393e+05  2.293e+06   0.104 0.916920
quit            -3.091e+06  2.896e+06  -1.068 0.285881
remak            4.293e+06  3.027e+06   1.418 0.156349
role            -2.554e+06  3.362e+06  -0.760 0.447538
```

```
romant            7.030e+05  2.935e+06   0.239 0.810763
satir            -2.779e+06  2.835e+06  -0.980 0.327116
satisfi          -5.496e+06  2.893e+06  -1.900 0.057670 .
script            1.332e+06  1.594e+06   0.835 0.403648
seem             -7.256e+06  2.830e+06  -2.564 0.010450 *
set               4.843e+06  2.557e+06   1.894 0.058370 .
sharp            -2.592e+06  2.455e+06  -1.056 0.291193
short             5.131e+05  2.653e+06   0.193 0.846673
smart             7.233e+06  1.970e+06   3.672 0.000250 ***
solid            -1.978e+06  2.023e+06  -0.977 0.328557
sourc             1.311e+07  3.499e+06   3.747 0.000186 ***
sport            -1.522e+06  2.928e+06  -0.520 0.603301
star              4.247e+05  1.929e+06   0.220 0.825759
still             2.279e+06  2.023e+06   1.126 0.260216
stori            -1.421e+06  1.368e+06  -1.039 0.298889
strong            1.971e+06  1.761e+06   1.119 0.263166
subject           5.404e+05  1.923e+06   0.281 0.778767
success           6.587e+06  3.155e+06   2.088 0.037018 *
suffer            3.734e+06  2.435e+06   1.533 0.125399
surpris           6.811e+06  2.095e+06   3.251 0.001180 **
sweet            -2.491e+06  2.763e+06  -0.901 0.367536
take              2.169e+06  2.368e+06   0.916 0.359700
tale             -3.001e+05  2.420e+06  -0.124 0.901315
talent            3.244e+06  1.790e+06   1.812 0.070231 .
thank             5.695e+06  2.610e+06   2.182 0.029301 *
that              1.197e+06  2.409e+06   0.497 0.619339
thin              2.273e+06  2.901e+06   0.783 0.433500
though            1.457e+06  1.651e+06   0.882 0.377774
thrill            1.431e+06  2.013e+06   0.711 0.477155
thriller         -1.913e+06  1.842e+06  -1.038 0.299359
time              1.568e+06  2.004e+06   0.782 0.434248
tone             -1.612e+05  3.048e+06  -0.053 0.957825
true              1.745e+06  2.901e+06   0.602 0.547606
turn             -7.596e+06  3.007e+06  -2.526 0.011663 *
twist             4.058e+06  2.859e+06   1.419 0.155987
ultim            -1.706e+06  1.868e+06  -0.913 0.361330
uneven            2.290e+06  2.311e+06   0.991 0.321925
viewer           -9.901e+05  1.925e+06  -0.514 0.607047
visual           -2.853e+06  1.702e+06  -1.676 0.094039 .
wast              2.080e+06  3.072e+06   0.677 0.498464
way              -1.402e+06  3.324e+06  -0.422 0.673348
well             -3.882e+06  2.180e+06  -1.781 0.075189 .
wellact          -3.492e+06  2.950e+06  -1.184 0.236689
will              1.894e+06  1.881e+06   1.007 0.314218
wit              -2.823e+06  2.973e+06  -0.949 0.342566
work             -2.872e+06  1.889e+06  -1.520 0.128651
world            -4.861e+06  3.458e+06  -1.406 0.160047
writerdirector   -2.081e+06  2.955e+06  -0.704 0.481312
young            -1.579e+06  3.320e+06  -0.476 0.634370
imdbRating       -2.189e+06  6.510e+05  -3.362 0.000795 ***
tomatoUserRating  5.328e+06  1.103e+06   4.831 1.51e-06 ***
ratingMean       -9.912e+05  1.048e+06  -0.946 0.344308
nRat              2.866e+03  2.606e+02  10.998  < 2e-16 ***
viewCount         1.073e+00  9.755e-02  10.997  < 2e-16 ***
dislikeCount     -2.005e+03  3.285e+02  -6.104 1.34e-09 ***
commentCount      2.190e+02  7.317e+01   2.994 0.002805 **
gAct              4.881e+06  1.029e+06   4.743 2.32e-06 ***
```

```
gAdv              4.583e+06  1.232e+06   3.721 0.000207 ***
gAnim             7.609e+06  2.148e+06   3.543 0.000409 ***
gDocu            -2.183e+06  1.253e+06  -1.742 0.081676 .
gDra             -3.145e+06  8.103e+05  -3.881 0.000109 ***
gFant             3.690e+06  1.382e+06   2.669 0.007694 **
gSci              3.954e+06  1.341e+06   2.948 0.003247 **
runtime           1.468e+05  2.318e+04   6.334 3.23e-10 ***
ratedG            1.141e+07  2.396e+06   4.761 2.13e-06 ***
ratedPG           5.632e+06  1.124e+06   5.012 6.08e-07 ***
ratedPG13         4.394e+06  7.357e+05   5.972 2.97e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11270000 on 1373 degrees of freedom
Multiple R-squared:  0.6639,  Adjusted R-squared:  0.6155
F-statistic:  13.7 on 198 and 1373 DF,  p-value: < 2.2e-16
```

This yields a better R-squared, but with several insignificant variables. Removing several insignificant variables reduces the overall R-squared, but improves the adjusted R-squared.

```
Call:
lm(formula = opening_weekend ~ compel + entertain + franchis +
    fun + human + impress + live + man + satisfi + seem + set +
    smart + sourc + success + surpris + turn + imdbRating + tomatoUserRating
+
    nRat + viewCount + dislikeCount + commentCount + gAct + gAdv +
    gAnim + gDocu + gDra + gFant + gSci + runtime + ratedG +
    ratedPG + ratedPG13, data = trainTC)

Residuals:
      Min        1Q    Median        3Q       Max
-52023177  -4770722     13095   4211196 120552164

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.446e+07  2.933e+06  -4.930 9.11e-07 ***
compel            4.661e+06  2.060e+06   2.263 0.023790 *
entertain         4.860e+06  1.320e+06   3.683 0.000238 ***
franchis          2.155e+07  2.777e+06   7.762 1.51e-14 ***
fun              -7.680e+06  2.396e+06  -3.206 0.001375 **
human             5.782e+06  2.498e+06   2.315 0.020765 *
impress           1.144e+07  2.707e+06   4.225 2.53e-05 ***
live              4.968e+06  2.013e+06   2.468 0.013685 *
man               6.036e+06  2.284e+06   2.643 0.008311 **
satisfi          -6.123e+06  2.673e+06  -2.291 0.022125 *
seem             -6.670e+06  2.655e+06  -2.512 0.012093 *
set               5.626e+06  2.405e+06   2.340 0.019427 *
smart             6.902e+06  1.809e+06   3.816 0.000141 ***
sourc             7.187e+06  2.181e+06   3.296 0.001004 **
success           6.656e+06  2.924e+06   2.276 0.022961 *
surpris           6.270e+06  1.968e+06   3.186 0.001474 **
turn             -6.787e+06  2.771e+06  -2.449 0.014432 *
imdbRating       -2.865e+06  4.792e+05  -5.979 2.79e-09 ***
tomatoUserRating  5.855e+06  1.020e+06   5.740 1.14e-08 ***
nRat              2.955e+03  2.316e+02  12.758  < 2e-16 ***
viewCount         1.048e+00  9.115e-02  11.495  < 2e-16 ***
```

```
dislikeCount     -1.957e+03  3.141e+02  -6.232 5.95e-10 ***
commentCount      2.366e+02  6.830e+01   3.465 0.000545 ***
gAct              5.654e+06  9.174e+05   6.163 9.11e-10 ***
gAdv              4.408e+06  1.137e+06   3.877 0.000110 ***
gAnim             7.301e+06  1.828e+06   3.994 6.80e-05 ***
gDocu            -3.672e+06  1.054e+06  -3.483 0.000509 ***
gDra             -3.961e+06  7.118e+05  -5.564 3.10e-08 ***
gFant             3.228e+06  1.275e+06   2.531 0.011463 *
gSci              2.808e+06  1.245e+06   2.255 0.024261 *
runtime           1.371e+05  2.137e+04   6.417 1.85e-10 ***
ratedG            1.035e+07  2.181e+06   4.746 2.26e-06 ***
ratedPG           5.342e+06  1.026e+06   5.205 2.20e-07 ***
ratedPG13         4.341e+06  6.782e+05   6.401 2.05e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11230000 on 1538 degrees of freedom
Multiple R-squared:  0.6266,  Adjusted R-squared:  0.6186
F-statistic: 78.21 on 33 and 1538 DF,  p-value: < 2.2e-16
```

The R-squared of the prediction calculation is

```
[1] 0.5491726
```

which is a slight improvement over the Linear Regression model prior to the Bag of Words method.

3. Bag of Words – “Plot Summary”

The same analysis performed on the Plot Summary variable produces as good or worse results.

```
Call:
lm(formula = opening_weekend ~ ., data = trainPS)

Residuals:
      Min        1Q    Median        3Q       Max
-44509939  -5134177   -216566   4072024 106090650

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.555e+07  3.315e+06  -4.690 3.00e-06 ***
agent             2.427e+06  2.055e+06   1.181 0.237649
america           6.693e+05  1.980e+06   0.338 0.735455
american         -2.584e+05  1.864e+06  -0.139 0.889756
around           -1.091e+05  2.345e+06  -0.047 0.962901
attempt          -5.719e+05  1.937e+06  -0.295 0.767886
away              2.590e+05  2.539e+06   0.102 0.918777
back             -2.304e+06  1.812e+06  -1.271 0.203836
base              1.243e+06  2.790e+06   0.445 0.656059
battl             6.493e+05  2.774e+06   0.234 0.814988
becom            -5.250e+05  1.373e+06  -0.382 0.702296
begin            -1.553e+06  2.130e+06  -0.729 0.466165
best             -1.514e+06  2.559e+06  -0.592 0.554109
boy               9.247e+04  1.912e+06   0.048 0.961439
bring            -2.886e+05  2.548e+06  -0.113 0.909867
brother           3.793e+05  1.836e+06   0.207 0.836359
call              1.868e+06  2.636e+06   0.709 0.478591
```

```
can             -4.339e+05  2.251e+06  -0.193 0.847176
career          -5.769e+05  2.538e+06  -0.227 0.820248
center           1.411e+06  2.224e+06   0.635 0.525852
chang            3.842e+05  2.336e+06   0.164 0.869378
children         9.394e+04  2.310e+06   0.041 0.967568
citi             1.082e+06  1.810e+06   0.598 0.550007
colleg          -6.853e+05  1.997e+06  -0.343 0.731534
come             2.753e+06  1.622e+06   1.697 0.089865 .
comedi          -3.150e+06  2.337e+06  -1.348 0.178006
countri         -1.806e+06  2.305e+06  -0.783 0.433477
coupl            1.960e+06  1.979e+06   0.991 0.322085
daughter        -5.058e+05  1.966e+06  -0.257 0.797008
day             -9.009e+05  1.862e+06  -0.484 0.628586
dead            -2.212e+06  2.647e+06  -0.836 0.403408
death           -1.381e+06  1.913e+06  -0.722 0.470456
decid           -1.939e+06  2.468e+06  -0.786 0.432238
discov          -1.867e+06  1.652e+06  -1.130 0.258660
documentari     -1.041e+06  1.615e+06  -0.645 0.519146
drama           -1.291e+06  2.326e+06  -0.555 0.578886
dream            8.894e+05  2.358e+06   0.377 0.706135
drug             5.634e+05  1.854e+06   0.304 0.761240
end             -2.143e+06  2.407e+06  -0.890 0.373487
escap           -3.638e+04  2.662e+06  -0.014 0.989099
event            1.103e+06  2.299e+06   0.480 0.631384
experi           1.006e+06  2.602e+06   0.387 0.698982
explor          -1.659e+06  2.458e+06  -0.675 0.499930
face            -4.198e+06  2.110e+06  -1.989 0.046868 *
fall             6.533e+05  2.117e+06   0.309 0.757611
famili           3.037e+05  1.053e+06   0.289 0.772997
father          -2.427e+06  1.491e+06  -1.629 0.103638
fight            4.369e+06  2.092e+06   2.088 0.036989 *
film            -2.323e+04  1.575e+06  -0.015 0.988234
find            -4.428e+05  1.062e+06  -0.417 0.676855
first            4.526e+05  2.159e+06   0.210 0.833989
follow           3.661e+06  1.946e+06   1.882 0.060099 .
forc            -1.313e+06  1.491e+06  -0.881 0.378630
form            -1.269e+06  2.230e+06  -0.569 0.569499
former          -2.868e+06  1.914e+06  -1.499 0.134133
four             1.163e+06  2.349e+06   0.495 0.620724
friend           2.221e+06  1.429e+06   1.554 0.120309
futur           -3.839e+06  2.571e+06  -1.493 0.135627
get              2.210e+06  1.506e+06   1.467 0.142562
girl             4.694e+05  1.540e+06   0.305 0.760572
goe             -4.102e+06  2.798e+06  -1.466 0.142780
group           -9.137e+05  1.546e+06  -0.591 0.554502
guy             -2.303e+06  2.139e+06  -1.077 0.281789
help            -1.711e+06  1.692e+06  -1.011 0.312038
hes              1.157e+06  2.424e+06   0.477 0.633109
high             1.292e+06  2.952e+06   0.438 0.661784
home             2.002e+06  1.711e+06   1.170 0.242240
hous             6.036e+05  2.172e+06   0.278 0.781127
human            4.077e+06  2.355e+06   1.731 0.083600 .
husband         -7.426e+05  2.593e+06  -0.286 0.774615
investig        -3.652e+05  2.120e+06  -0.172 0.863238
job             -2.657e+06  2.153e+06  -1.234 0.217409
journey         -3.412e+06  2.594e+06  -1.315 0.188641
kill             3.034e+06  2.174e+06   1.396 0.162998
```

```
last            -2.772e+06  2.758e+06  -1.005 0.315060
lead            -9.693e+04  2.154e+06  -0.045 0.964117
learn            4.644e+06  1.922e+06   2.416 0.015826 *
left             5.207e+05  2.359e+06   0.221 0.825360
life            -8.627e+05  9.462e+05  -0.912 0.362077
live            -7.742e+05  1.230e+06  -0.629 0.529275
look            -1.235e+06  1.508e+06  -0.819 0.412861
love             2.526e+05  1.411e+06   0.179 0.857977
make            -5.080e+05  1.945e+06  -0.261 0.793977
man             -3.509e+05  1.143e+06  -0.307 0.758985
marri            1.010e+06  2.363e+06   0.427 0.669194
meet            -7.649e+05  1.757e+06  -0.435 0.663402
men              1.210e+05  2.275e+06   0.053 0.957601
mission         -8.777e+05  2.802e+06  -0.313 0.754161
mother          -3.415e+05  1.618e+06  -0.211 0.832907
move             7.679e+05  2.218e+06   0.346 0.729179
murder           6.758e+05  1.947e+06   0.347 0.728584
music            4.577e+04  2.214e+06   0.021 0.983514
must             2.128e+06  1.662e+06   1.280 0.200649
mysteri         -7.388e+05  1.832e+06  -0.403 0.686786
new             -6.373e+05  1.348e+06  -0.473 0.636480
night           -7.615e+05  2.416e+06  -0.315 0.752630
old             -1.833e+06  2.115e+06  -0.867 0.386175
one             -3.736e+05  1.204e+06  -0.310 0.756395
order            1.533e+06  2.210e+06   0.694 0.487976
parent           1.229e+06  2.339e+06   0.525 0.599444
past             2.745e+06  2.547e+06   1.078 0.281381
peopl           -2.614e+05  2.059e+06  -0.127 0.898975
person          -5.733e+05  2.341e+06  -0.245 0.806566
plan            -4.037e+06  2.422e+06  -1.667 0.095676 .
play            -3.619e+06  2.258e+06  -1.603 0.109143
protect          3.297e+06  2.578e+06   1.279 0.201142
put              1.743e+06  2.112e+06   0.825 0.409417
relationship     7.153e+05  1.647e+06   0.434 0.664072
return          -2.389e+05  1.859e+06  -0.129 0.897755
save            -3.064e+06  2.112e+06  -1.450 0.147222
school          -1.621e+06  2.404e+06  -0.674 0.500362
search           1.421e+06  2.189e+06   0.649 0.516252
secret           4.560e+04  2.243e+06   0.020 0.983784
seri             2.022e+05  2.325e+06   0.087 0.930684
set             -6.098e+05  1.647e+06  -0.370 0.711224
sister           3.377e+06  2.004e+06   1.685 0.092175 .
son             -2.064e+06  1.677e+06  -1.231 0.218465
start            3.259e+06  2.232e+06   1.460 0.144485
stori           -2.033e+06  1.263e+06  -1.611 0.107512
struggl          6.058e+05  1.568e+06   0.386 0.699220
student         -4.586e+05  2.084e+06  -0.220 0.825878
take             1.662e+06  1.264e+06   1.316 0.188476
team             3.558e+06  1.583e+06   2.248 0.024742 *
teenag           4.672e+04  1.848e+06   0.025 0.979833
three           -1.056e+06  1.533e+06  -0.689 0.491045
time            -6.760e+05  2.066e+06  -0.327 0.743526
togeth           2.508e+06  1.941e+06   1.293 0.196372
town            -1.144e+06  1.830e+06  -0.625 0.532173
travel          -8.514e+05  2.004e+06  -0.425 0.670976
tri             -9.205e+05  1.484e+06  -0.620 0.535323
trip            -4.096e+06  2.512e+06  -1.631 0.103188
```

```
troubl            1.207e+06  2.479e+06   0.487 0.626312
true             -3.248e+06  2.558e+06  -1.269 0.204519
turn              9.898e+04  1.836e+06   0.054 0.957018
two               2.095e+05  1.177e+06   0.178 0.858799
use              -2.152e+06  2.553e+06  -0.843 0.399467
war              -3.509e+06  1.645e+06  -2.134 0.033055 *
way              -1.440e+06  1.916e+06  -0.751 0.452678
whose            -1.065e+06  2.209e+06  -0.482 0.629777
wife              1.577e+05  1.887e+06   0.084 0.933408
will             -3.102e+06  2.252e+06  -1.377 0.168592
woman             1.172e+06  1.305e+06   0.899 0.369021
work              2.047e+06  1.551e+06   1.320 0.187085
world            -1.476e+06  1.118e+06  -1.320 0.186954
year             -1.396e+06  1.420e+06  -0.984 0.325507
york              1.944e+06  2.299e+06   0.846 0.397818
young            -3.034e+06  1.042e+06  -2.911 0.003661 **
imdbRating       -1.505e+06  6.948e+05  -2.166 0.030483 *
tomatoUserRating  6.834e+06  1.155e+06   5.919 4.07e-09 ***
ratingMean       -2.223e+06  1.136e+06  -1.956 0.050645 .
nRat              2.538e+03  2.970e+02   8.547  < 2e-16 ***
viewCount         9.793e-01  1.072e-01   9.132  < 2e-16 ***
dislikeCount     -1.991e+03  3.477e+02  -5.725 1.27e-08 ***
commentCount      5.205e+02  8.289e+01   6.279 4.53e-10 ***
gAct              4.768e+06  1.084e+06   4.400 1.16e-05 ***
gAdv              6.447e+06  1.290e+06   4.998 6.53e-07 ***
gAnim             9.495e+06  2.018e+06   4.705 2.79e-06 ***
gDocu            -4.172e+06  1.406e+06  -2.967 0.003059 **
gDra             -3.328e+06  8.142e+05  -4.087 4.62e-05 ***
gFant             4.510e+06  1.408e+06   3.204 0.001387 **
gSci              3.171e+06  1.431e+06   2.216 0.026850 *
runtime           1.199e+05  2.245e+04   5.340 1.08e-07 ***
ratedG            9.795e+06  2.556e+06   3.832 0.000132 ***
ratedPG           4.951e+06  1.177e+06   4.206 2.77e-05 ***
ratedPG13         5.313e+06  7.568e+05   7.021 3.41e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11850000 on 1407 degrees of freedom
Multiple R-squared:  0.629,   Adjusted R-squared:  0.5858
F-statistic: 14.55 on 164 and 1407 DF,  p-value: < 2.2e-16
```

Once again, cleaning up the independent variables produces a worse R-squared, but a better Adjusted R-squared

```
Call:
lm(formula = opening_weekend ~ fight + human + learn + sister +
    team + war + young + imdbRating + tomatoUserRating + ratingMean +
    nRat + viewCount + dislikeCount + commentCount + gAct + gAdv +
    gAnim + gDocu + gDra + gFant + gSci + runtime + ratedG +
    ratedPG + ratedPG13, data = trainPS)

Residuals:
      Min        1Q    Median        3Q       Max
-45886950  -4956367   -306624   3798189 115531709

Coefficients:
```

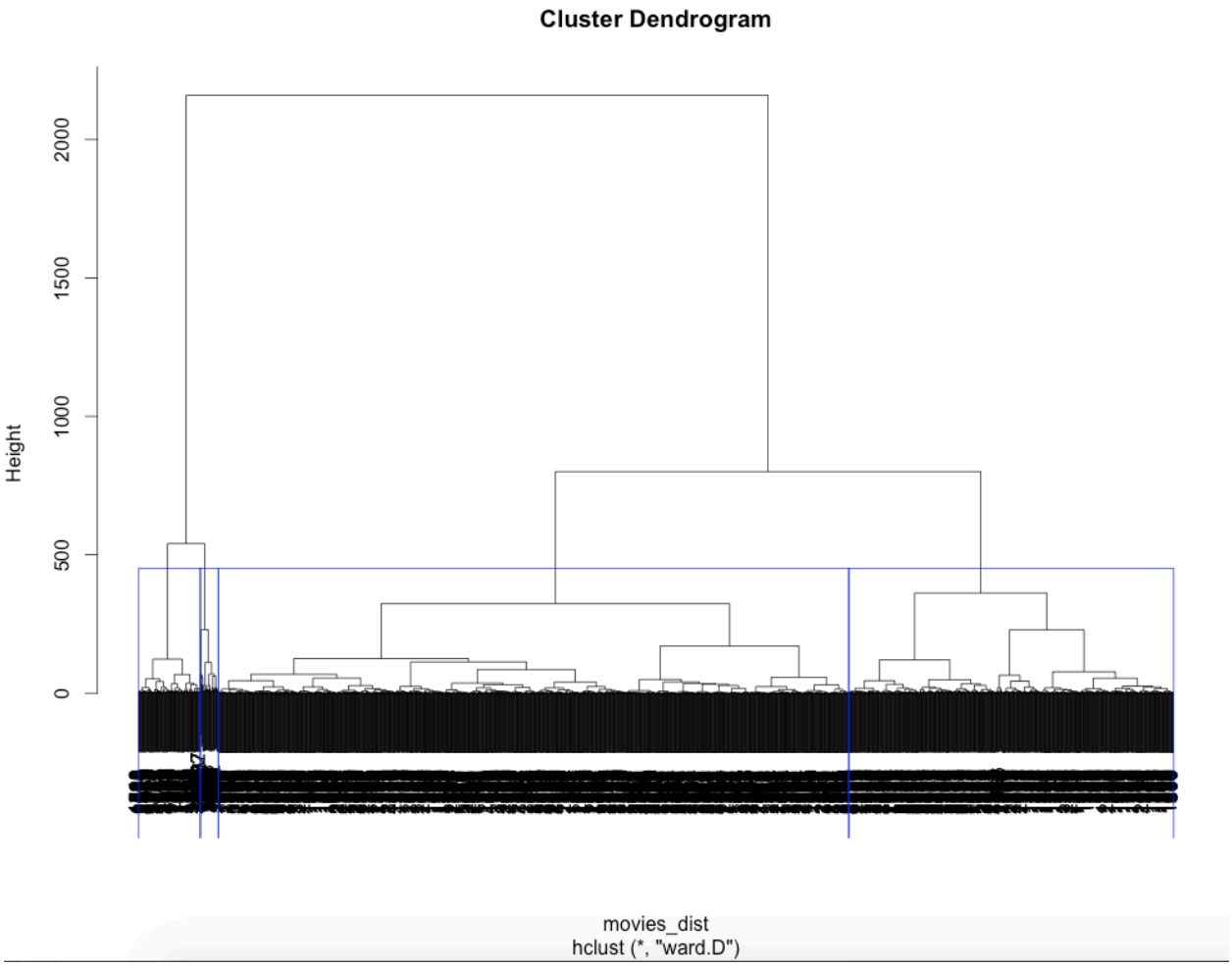
```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.652e+07  2.977e+06  -5.550 3.35e-08 ***
fight              4.325e+06  1.963e+06   2.203 0.027730 *
human              4.185e+06  2.250e+06   1.860 0.063133 .
learn              4.276e+06  1.802e+06   2.374 0.017741 *
sister             3.160e+06  1.839e+06   1.718 0.085953 .
team               2.632e+06  1.458e+06   1.806 0.071191 .
war               -3.890e+06  1.533e+06  -2.537 0.011293 *
young             -2.383e+06  9.590e+05  -2.485 0.013061 *
imdbRating        -1.554e+06  6.574e+05  -2.364 0.018223 *
tomatoUserRating   7.111e+06  1.073e+06   6.625 4.77e-11 ***
ratingMean        -2.284e+06  1.072e+06  -2.132 0.033191 *
nRat               2.470e+03  2.806e+02   8.801  < 2e-16 ***
viewCount          1.028e+00  1.009e-01  10.190  < 2e-16 ***
dislikeCount      -2.109e+03  3.295e+02  -6.402 2.03e-10 ***
commentCount       4.943e+02  7.790e+01   6.346 2.90e-10 ***
gAct               5.472e+06  9.770e+05   5.601 2.52e-08 ***
gAdv               6.135e+06  1.207e+06   5.083 4.16e-07 ***
gAnim              9.031e+06  1.899e+06   4.757 2.15e-06 ***
gDocu             -4.597e+06  1.131e+06  -4.065 5.04e-05 ***
gDra              -3.422e+06  7.464e+05  -4.585 4.91e-06 ***
gFant              4.697e+06  1.330e+06   3.530 0.000427 ***
gSci               3.312e+06  1.285e+06   2.578 0.010040 *
runtime            1.159e+05  2.113e+04   5.485 4.83e-08 ***
ratedG             8.891e+06  2.412e+06   3.685 0.000236 ***
ratedPG            4.459e+06  1.107e+06   4.027 5.91e-05 ***
ratedPG13          4.997e+06  7.076e+05   7.063 2.46e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11700000 on 1546 degrees of freedom
Multiple R-squared:  0.603,   Adjusted R-squared:  0.5965
F-statistic: 93.91 on 25 and 1546 DF,  p-value: < 2.2e-16
```

The R-squared of the prediction calculation is
```
[1] 0.5540814
```
which is a slight improvement over both the general Linear Regression model, and the LR model run after Bag of Words method on the 'TomatoConsensus' variable.

4. Clustering

After transforming the variables to make the data suitable for the clustering method, the 'ward.D' method is used to produce a dendrogram of the clustering.

The ideal number of clusters seems to be 3 – 4. The data is divided up into k = 4 clusters as outlined in the dendrogram below, and the dataset assigned to these groups.



Performing the analysis for each of the training and test set groups does not yield very promising results.

Group 1 Linear Regression & R-squared

```
Residual standard error: 13230000 on 483 degrees of freedom
Multiple R-squared:  0.4234,  Adjusted R-squared:  0.4043
F-statistic: 22.16 on 16 and 483 DF,  p-value: < 2.2e-16
```

Prediction R-squared:
```
[1] 0.3837773
```

Group 2 Linear Regression & R-squared

```
Residual standard error: 2694000 on 923 degrees of freedom
Multiple R-squared:  0.1922,  Adjusted R-squared:  0.1782
F-statistic: 13.73 on 16 and 923 DF,  p-value: < 2.2e-16
```

Prediction R-squared:
```
[1] 0.2168196
```

Group 3 Linear Regression & R-squared

```
Residual standard error: 17730000 on 85 degrees of freedom
Multiple R-squared:  0.5618,  Adjusted R-squared:  0.4845
F-statistic: 7.266 on 15 and 85 DF,  p-value: 5.54e-10
```

Prediction R-squared:

```
[1] 0.5108711
```

Group 4 Linear Regression & R-squared

```
Residual standard error: 43820000 on 15 degrees of freedom
Multiple R-squared:  0.7527,  Adjusted R-squared:  0.5054
F-statistic: 3.044 on 15 and 15 DF,  p-value: 0.01925
```

Prediction R-squared:

```
[1] -0.4705494
```

Table summarizing R^2 of all analyses methods (second run, with set.seed = 5249)

Test Details	Var Name in R Script	# of Dep. Variables in Regression	R2 Measure	R2 Value
Linear Regression	moviesLM	37	Multiple R2	0.588505312
Linear Regression	moviesLM2	19	Multiple R2	0.584631884
Linear Regression on Training set	moviesLM3	19	Calculated R2	0.561908492
Tomato Consensus - Bag of Words, Lin Regr, on Training Set	TC_lm	199	Multiple R2	0.678003983
Tomato Consensus - Bag of Words, Lin Regr, Pred. on Test Set	TC_lm	199	Calculated R2	0.520915301
Tomato Consensus - Bag of Words, Lin Regr, on Training Set	TC_lm2	120	Multiple R2	0.666388717
Tomato Consensus - Bag of Words, Lin Regr, Pred. on Test Set	TC_lm2	120	Calculated R2	0.550142969
Tomato Consensus - Bag of Words, Lin Regr, on Training Set	TC_lm3	34	Multiple R2	0.629268687
Tomato Consensus - Bag of Words, Lin Regr, Pred. on Test Set	TC_lm3	34	Calculated R2	0.559980184
Plot Summary - Bag of Words, Lin Regr, on Training Set	PS_lm	165	Multiple R2	0.629954347

Plot Summary - Bag of Words, Lin Regr, Pred. on Test Set	PS_lm	165	Calculated R2	0.544810698
Plot Summary - Bag of Words, Lin Regr, on Training Set	PS_lm2	26	Multiple R2	0.595727472
Plot Summary - Bag of Words, Lin Regr, Pred. on Test Set	PS_lm2	26	Calculated R2	0.544810698
Clustering, Lin Regr, on Training Set - Group 1	cLM1	17	Multiple R2	0.559356774
Clustering, Lin Regr, Pred. on Test Set - Group 1	cLM1	17	Calculated R2	0.635766894
Clustering, Lin Regr, on Training Set - Group 2	cLM2	17	Multiple R2	0.553757972
Clustering, Lin Regr, Pred. on Test Set - Group 2	cLM2	17	Calculated R2	0.554591159
Clustering, Lin Regr, on Training Set - Group 3	cLM3	17	Multiple R2	0.515555359
Clustering, Lin Regr, Pred. on Test Set - Group 3	cLM3	17	Calculated R2	0.592075031
Clustering, Lin Regr, on Training Set - Group 4	cLM4	17	Multiple R2	0.680273013
Clustering, Lin Regr, Pred. on Test Set - Group 4	cLM4	17	Calculated R2	0.669983117

Conclusions

Further analysis and data are required to improve the predictive abilities of the linear regression model. While it appears that critic's ratings do not predict Opening Weekend grosses very well, User Ratings do exhibit a correlation with Opening Weekend grosses, especially in specific genres. Due to the fluid nature of the data, and the fact that the data was collected well after the movie has released, it is difficult to determine when the user ratings were posted. Consequently, user ratings may well be after the fact of the movie's success.