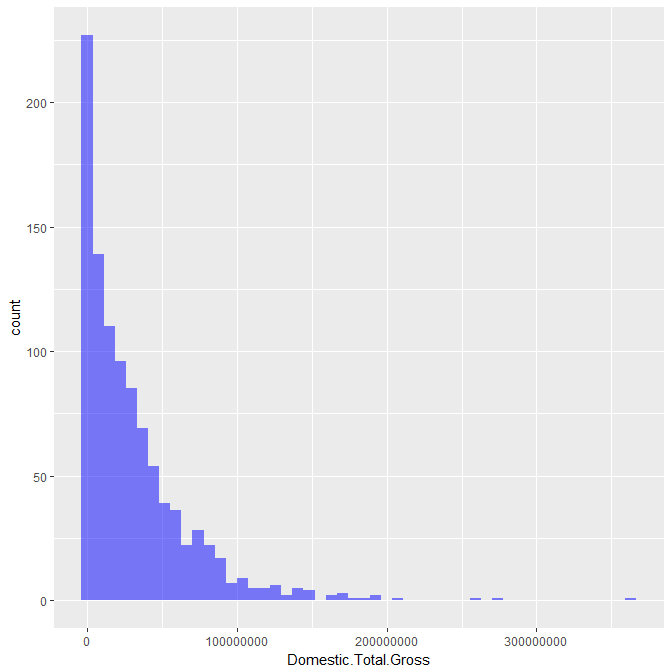
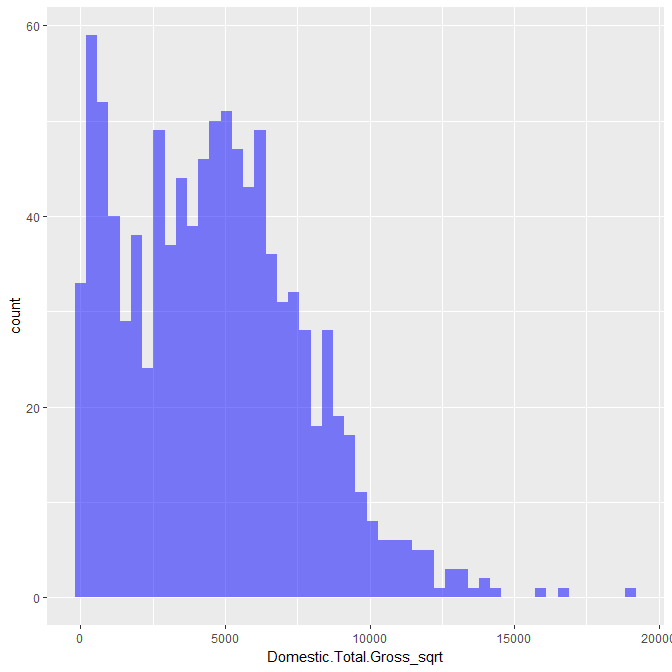
## 6242

Visualization of domestic total gross profit distribution.

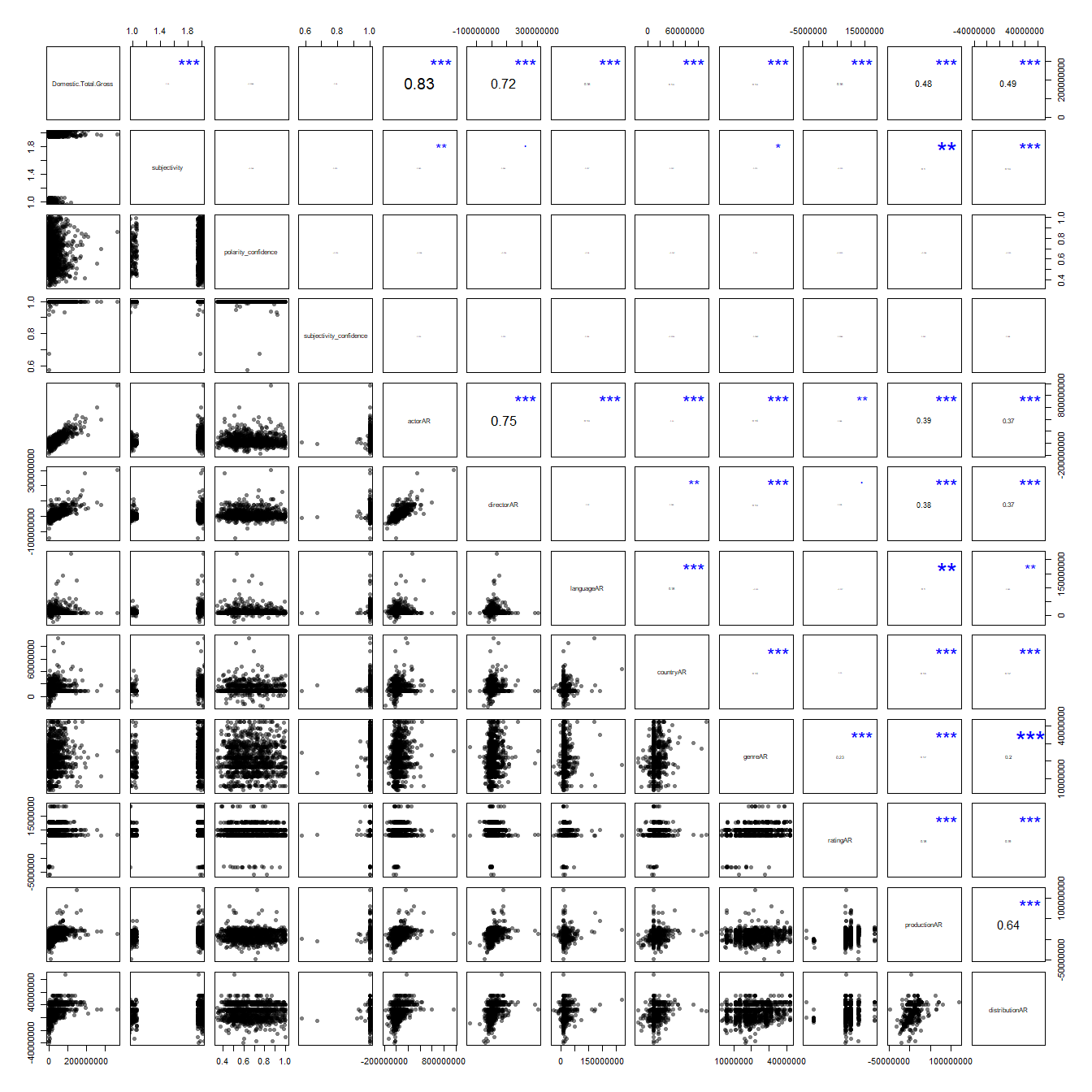


After applying square root transformation to make it less right skewed:

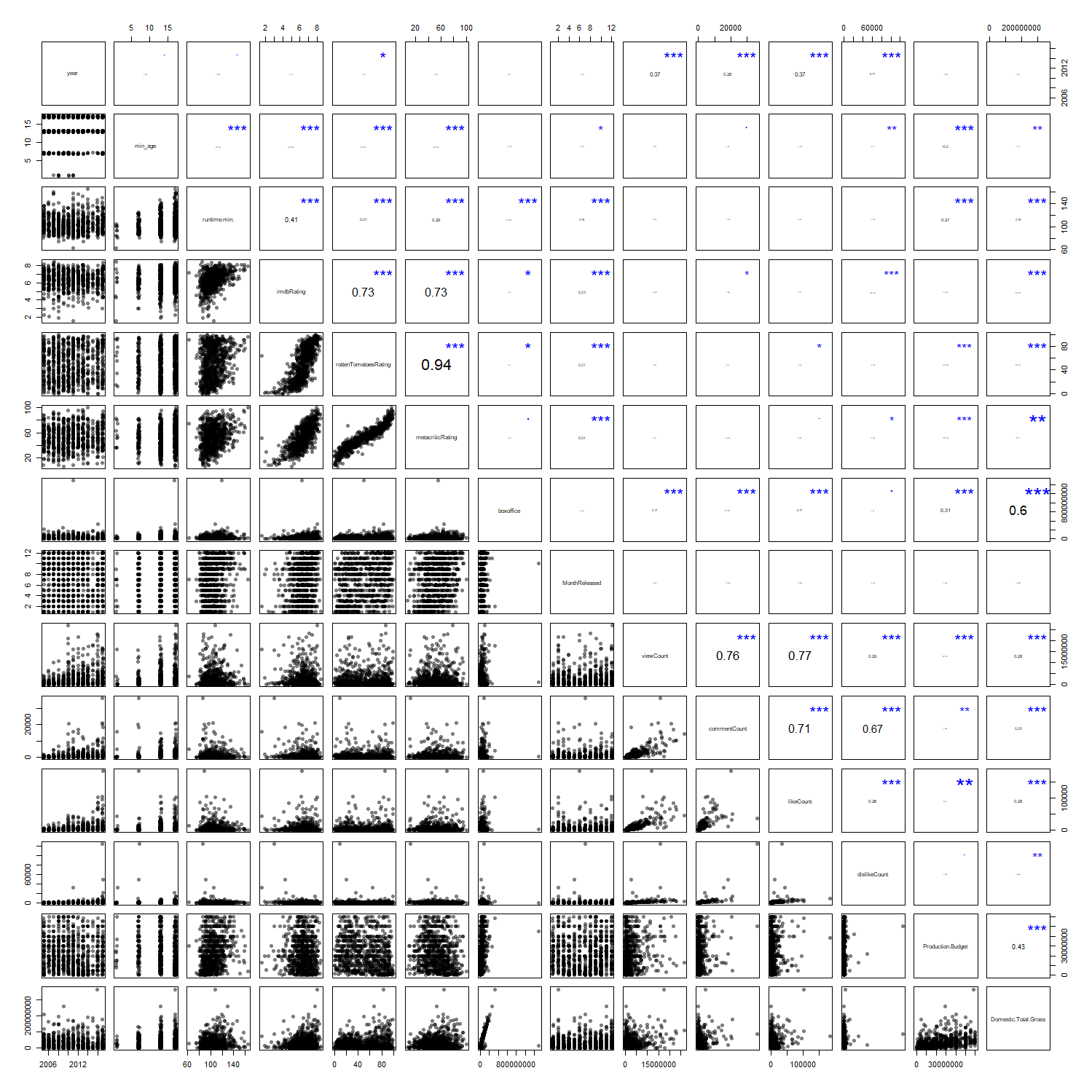


Here is the correlation matrix for all of the variables. Split into two for visibility, domestic total gross revenue is included in both. The variables with the highest correlation are budget and added revenue for actor, director, production, and distribution.

Correlation  
Data : outTrain   
Method : pearson   
Variables: Domestic.Total.Gross, subjectivity, polarity\_confidence, subjectivity\_confidence, actorAR, directorAR, languageAR, countryAR, genreAR, ratingAR, productionAR, distributionAR   
Null hyp.: variables x and y are not correlated  
Alt. hyp.: variables x and y are correlated  
  
Correlation matrix:  
 Domestic.Total.Gross subjectivity polarity\_confidence subjectivity\_confidence actorAR directorAR languageAR countryAR genreAR ratingAR productionAR  
subjectivity 0.11   
polarity\_confidence -0.01 -0.02   
subjectivity\_confidence 0.04 0.00 -0.02   
actorAR 0.83 0.09 -0.04 0.02   
directorAR 0.72 0.06 -0.02 0.03 0.75   
languageAR 0.16 0.04 -0.04 0.01 0.13 0.05   
countryAR 0.13 0.04 -0.01 -0.00 0.11 0.10 0.18   
genreAR 0.13 0.07 0.01 0.01 0.15 0.13 -0.01 0.15   
ratingAR 0.16 -0.02 -0.00 0.01 0.09 0.06 -0.03 0.03 0.23   
productionAR 0.48 0.10 -0.02 0.04 0.39 0.38 0.10 0.14 0.12 0.18   
distributionAR 0.49 0.13 -0.02 0.05 0.37 0.37 0.08 0.12 0.20 0.19 0.64   
  
p.values:  
 Domestic.Total.Gross subjectivity polarity\_confidence subjectivity\_confidence actorAR directorAR languageAR countryAR genreAR ratingAR productionAR  
subjectivity 0.00   
polarity\_confidence 0.79 0.56   
subjectivity\_confidence 0.21 0.95 0.51   
actorAR 0.00 0.00 0.26 0.51   
directorAR 0.00 0.05 0.46 0.31 0.00   
languageAR 0.00 0.24 0.21 0.66 0.00 0.11   
countryAR 0.00 0.26 0.65 0.94 0.00 0.00 0.00   
genreAR 0.00 0.02 0.72 0.79 0.00 0.00 0.69 0.00   
ratingAR 0.00 0.50 0.94 0.79 0.01 0.08 0.39 0.35 0.00   
productionAR 0.00 0.00 0.56 0.18 0.00 0.00 0.00 0.00 0.00 0.00   
distributionAR 0.00 0.00 0.46 0.15 0.00 0.00 0.01 0.00 0.00 0.00 0.00

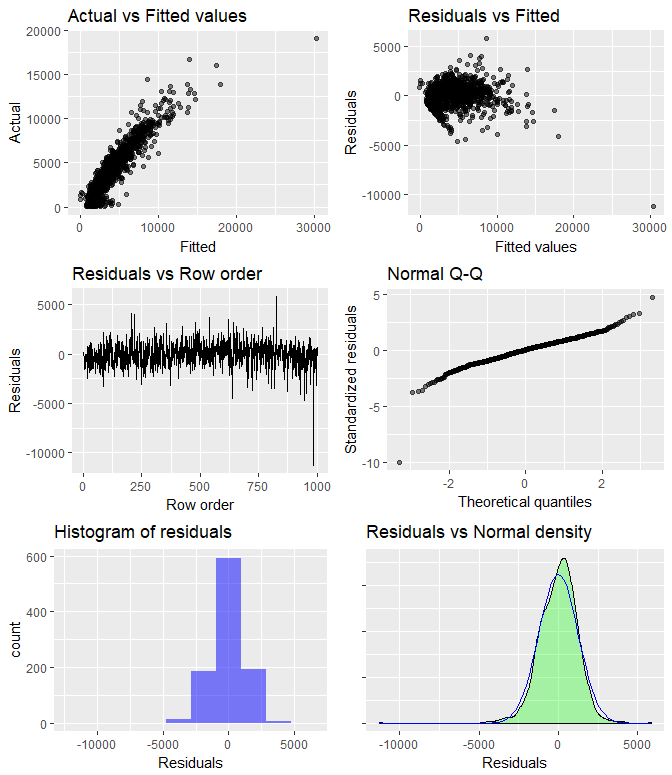


Correlation  
Data : outTrain   
Method : pearson   
Variables: year, min\_age, runtime.min., imdbRating, rottenTomatoesRating, metacriticRating, boxoffice, MonthReleased, viewCount, commentCount, likeCount, dislikeCount, Production.Budget, Domestic.Total.Gross   
Null hyp.: variables x and y are not correlated  
Alt. hyp.: variables x and y are correlated  
  
Correlation matrix:  
 year min\_age runtime.min. imdbRating rottenTomatoesRating metacriticRating boxoffice MonthReleased viewCount commentCount likeCount dislikeCount Production.Budget  
min\_age 0.06   
runtime.min. 0.06 0.14   
imdbRating -0.03 0.14 0.41   
rottenTomatoesRating 0.08 0.13 0.24 0.73   
metacriticRating 0.05 0.14 0.29 0.73 0.94   
boxoffice 0.02 -0.05 0.14 0.09 0.07 0.06   
MonthReleased -0.02 0.08 0.18 0.23 0.21 0.24 0.03   
viewCount 0.37 0.02 0.05 -0.01 0.04 0.03 0.17 0.03   
commentCount 0.28 -0.06 -0.02 -0.08 0.02 -0.00 0.15 0.00 0.76   
likeCount 0.37 -0.02 0.01 0.03 0.08 0.06 0.17 0.01 0.77 0.71   
dislikeCount 0.17 -0.11 -0.05 -0.19 -0.05 -0.08 0.06 -0.02 0.29 0.67 0.26   
Production.Budget -0.05 -0.20 0.27 -0.00 -0.16 -0.14 0.31 0.02 0.12 0.10 0.10 0.07   
Domestic.Total.Gross 0.03 -0.11 0.18 0.16 0.13 0.10 0.60 0.01 0.28 0.25 0.28 0.10 0.43   
  
p.values:  
 year min\_age runtime.min. imdbRating rottenTomatoesRating metacriticRating boxoffice MonthReleased viewCount commentCount likeCount dislikeCount Production.Budget  
min\_age 0.06   
runtime.min. 0.07 0.00   
imdbRating 0.40 0.00 0.00   
rottenTomatoesRating 0.02 0.00 0.00 0.00   
metacriticRating 0.15 0.00 0.00 0.00 0.00   
boxoffice 0.63 0.14 0.00 0.01 0.05 0.09   
MonthReleased 0.52 0.03 0.00 0.00 0.00 0.00 0.41   
viewCount 0.00 0.50 0.17 0.89 0.23 0.44 0.00 0.43   
commentCount 0.00 0.09 0.49 0.02 0.63 0.91 0.00 0.96 0.00   
likeCount 0.00 0.52 0.68 0.45 0.03 0.10 0.00 0.86 0.00 0.00   
dislikeCount 0.00 0.00 0.12 0.00 0.12 0.02 0.09 0.55 0.00 0.00 0.00   
Production.Budget 0.13 0.00 0.00 0.93 0.00 0.00 0.00 0.66 0.00 0.01 0.00 0.06   
Domestic.Total.Gross 0.38 0.00 0.00 0.00 0.00 0.00 0.00 0.82 0.00 0.00 0.00 0.01 0.00



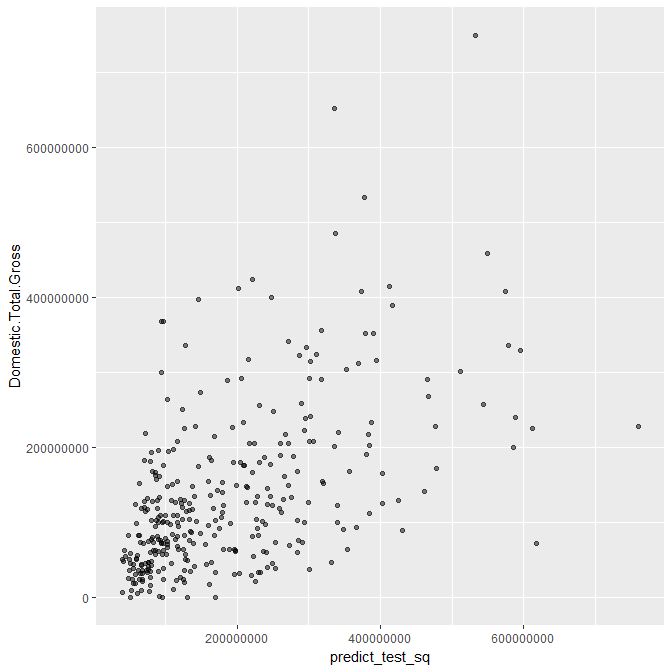
First model: Budget and added revenue for actor, director, production, and distribution. These variables had the highest correlation with domestic total gross revenue. I used the square root transformation, so when analyzing the test results I needed to square the results to get the actual prediction. Not that while the coefficients appear to be 0, they are not actually zero - it is just that the values for budget and added revenue are so large that a one unit change has a small effect.

Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: Production.Budget, actorAR, directorAR, productionAR, distributionAR   
Null hyp.: the effect of x on Domestic.Total.Gross\_sqrt is zero  
Alt. hyp.: the effect of x on Domestic.Total.Gross\_sqrt is not zero  
  
 coefficient std.error t.value p.value   
 (Intercept) 1472.040 69.355 21.225 < .001 \*\*\*  
 Production.Budget 0.000 0.000 30.715 < .001 \*\*\*  
 actorAR 0.000 0.000 27.516 < .001 \*\*\*  
 directorAR 0.000 0.000 7.773 < .001 \*\*\*  
 productionAR 0.000 0.000 3.699 < .001 \*\*\*  
 distributionAR 0.000 0.000 9.035 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.841, Adjusted R-squared: 0.841   
F-statistic: 1054.272 df(5,994), p.value < .001  
Nr obs: 1,000   
  
Sum of squares:  
 df SS  
Regression 5 8014114474  
Error 994 1511190408  
Total 999 9525304882  
  
Variance Inflation Factors  
 actorAR directorAR distributionAR productionAR Production.Budget  
VIF 2.368 2.329 1.902 1.829 1.190  
Rsq 0.578 0.571 0.474 0.453 0.159



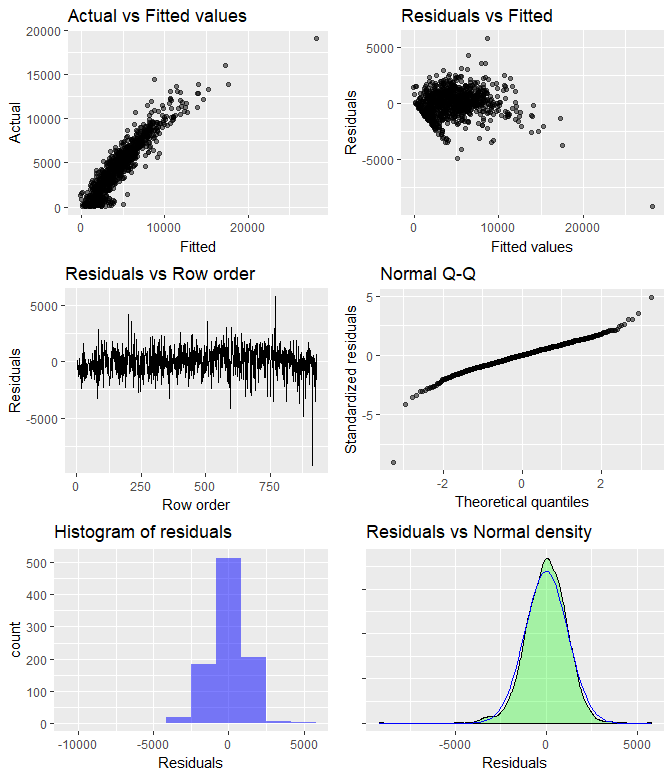
Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: Production.Budget, actorAR, directorAR, productionAR, distributionAR   
Prediction dataset : outTest   
Rows shown : 10 of 332   
  
 Production.Budget actorAR directorAR productionAR distributionAR Prediction 2.5% 97.5% +/-  
 60000000 -15261663 30522105 -2697372 12943408 7112.322 4677.043 9547.600 2435.278  
 60000000 0 0 21745046 20762074 7606.972 5178.518 10035.426 2428.454  
 60000000 -4866478 -27365284 22280237 20132369 7070.807 4639.265 9502.349 2431.542  
 60000000 -12636329 0 3463844 -395282 6245.271 3813.489 8677.054 2431.782  
 60000000 -29182980 -14724554 17163082 10358090 6333.348 3903.537 8763.159 2429.811  
 60000000 0 0 3463844 -395282 6487.643 4055.851 8919.435 2431.792  
 60000000 -5875187 0 3463844 -395282 6374.954 3943.181 8806.726 2431.772  
 60000000 20958976 16535902 -2697372 12943408 7589.556 5157.421 10021.690 2432.135  
 60000000 -19997146 -3355187 22280237 20132369 7153.970 4725.014 9582.927 2428.957  
 60000000 -8030088 0 21745046 20441134 7440.094 5011.517 9868.672 2428.577

First model results - y axis is actual revenue and x axis is predicted revenue.



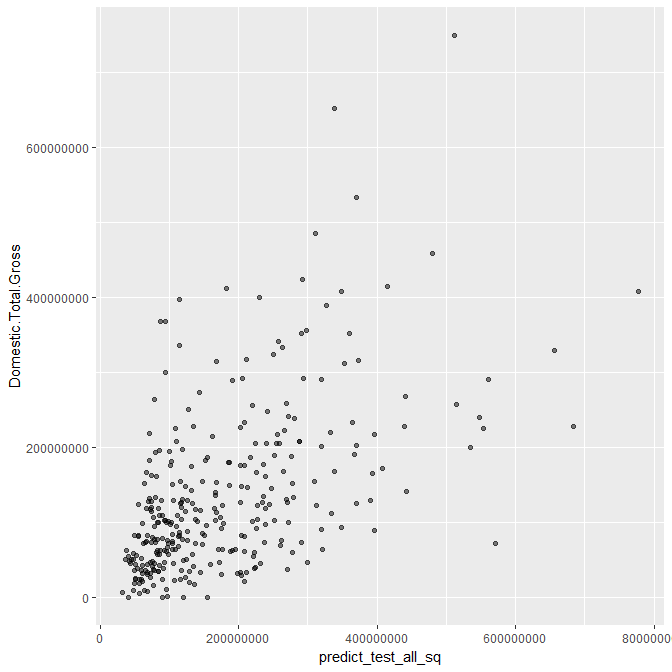
Second model: With all variables. Has a lower sum of squared error than the first model that had only 5 variables.

Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: rated, runtime.min., metacriticRating, YearReleased, MonthReleased, viewCount, commentCount, likeCount, dislikeCount, Production.Budget, polarity, subjectivity, polarity\_confidence, subjectivity\_confidence, actorAR, directorAR, languageAR, countryAR, genreAR, ratingAR, productionAR, distributionAR   
Null hyp.: the effect of x on Domestic.Total.Gross\_sqrt is zero  
Alt. hyp.: the effect of x on Domestic.Total.Gross\_sqrt is not zero  
  
 coefficient std.error t.value p.value   
 (Intercept) 76438.639 24211.219 3.157 0.002 \*\*   
 rated|NC-17 -2051.053 937.980 -2.187 0.029 \*   
 rated|NOT RATED -1456.918 486.119 -2.997 0.003 \*\*   
 rated|PG -324.013 402.191 -0.806 0.421   
 rated|PG-13 -476.434 388.424 -1.227 0.220   
 rated|R -755.464 387.827 -1.948 0.052 .   
 runtime.min. 7.847 3.058 2.566 0.010 \*   
 metacriticRating 8.798 2.478 3.551 < .001 \*\*\*  
 YearReleased -38.842 11.944 -3.252 0.001 \*\*   
 MonthReleased -3.041 12.062 -0.252 0.801   
 viewCount 0.000 0.000 0.293 0.770   
 commentCount 0.100 0.035 2.837 0.005 \*\*   
 likeCount -0.015 0.005 -2.959 0.003 \*\*   
 dislikeCount -0.021 0.013 -1.620 0.105   
 Production.Budget 0.000 0.000 25.974 < .001 \*\*\*  
 polarity|neutral 3.313 96.776 0.034 0.973   
 polarity|positive -75.798 120.904 -0.627 0.531   
 subjectivity|subjective -119.455 121.475 -0.983 0.326   
 polarity|\_confidence 4.769 255.116 0.019 0.985   
 subjectivity|\_confidence 2403.933 2162.718 1.112 0.267   
 actorAR 0.000 0.000 25.181 < .001 \*\*\*  
 directorAR 0.000 0.000 8.233 < .001 \*\*\*  
 languageAR 0.000 0.000 1.503 0.133   
 countryAR 0.000 0.000 1.347 0.178   
 genreAR 0.000 0.000 1.766 0.078 .   
 productionAR 0.000 0.000 2.526 0.012 \*   
 distributionAR 0.000 0.000 9.526 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.854, Adjusted R-squared: 0.85   
F-statistic: 203.838 df(26,904), p.value < .001  
Nr obs: 931   
  
The set of explanatory variables exhibit perfect multicollinearity.  
One or more variables were dropped from the estimation.  
Sum of squares:  
 df SS  
Regression 26 7509453106  
Error 904 1280909710  
Total 930 8790362816  
  
Multicollinearity diagnostics were not calculated.



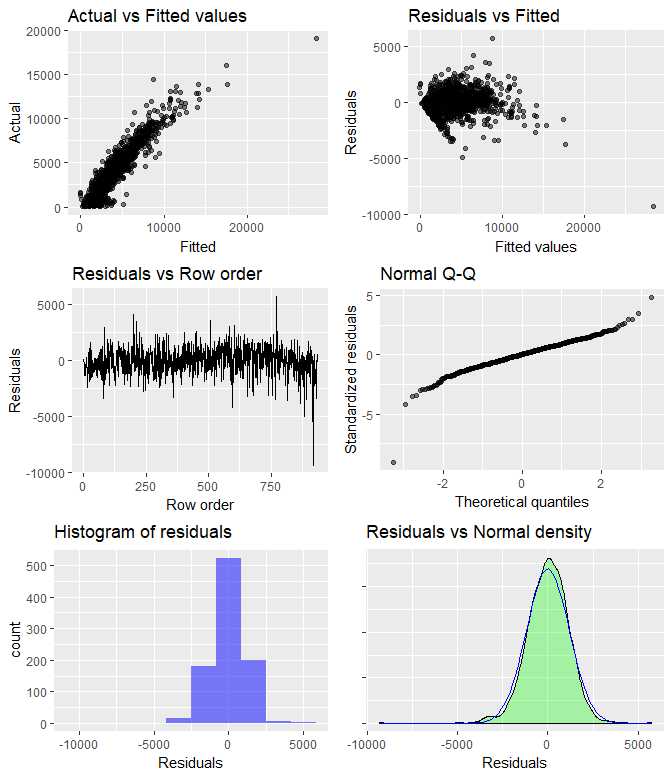
Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: rated, runtime.min., metacriticRating, YearReleased, MonthReleased, viewCount, commentCount, likeCount, dislikeCount, Production.Budget, polarity, subjectivity, polarity\_confidence, subjectivity\_confidence, actorAR, directorAR, languageAR, countryAR, genreAR, ratingAR, productionAR, distributionAR   
Prediction dataset : outTest   
Rows shown : 10 of 317   
  
 rated runtime.min. metacriticRating YearReleased MonthReleased viewCount commentCount likeCount dislikeCount Production.Budget polarity subjectivity polarity\_confidence subjectivity\_confidence actorAR directorAR languageAR countryAR genreAR  
 R 113 40 2013 1 8778.000 4.000 23.000 1.000 60000000 neutral subjective 0.576 1.000 -15261663 30522105 9159624 9575722 13464425  
 PG-13 130 33 2013 8 6507433.000 11440.000 29867.000 1042.000 60000000 neutral subjective 0.610 1.000 0 0 9159624 26515417 23008630  
 PG-13 105 57 2014 1 150574.000 145.000 382.000 63.000 60000000 neutral subjective 0.874 1.000 -4866478 -27365284 15344220 -7396445 21913317  
 PG 96 61 2014 9 274261.000 169.000 1159.000 42.000 60000000 positive subjective 0.857 1.000 -12636329 0 9159624 9575722 22293110  
 R 107 27 2015 1 4486055.000 885.000 9363.000 527.000 60000000 negative objective 0.588 1.000 -29182980 -14724554 9159624 12012154 18782711  
 R 99 28 2016 3 1714519.000 1084.000 3003.000 347.000 60000000 negative subjective 0.480 1.000 0 0 25371079 57480737 21913317  
 PG 101 84 2016 8 6319206.000 3426.000 28815.000 1522.000 60000000 negative objective 0.614 1.000 -5875187 0 9159624 9575722 29429955  
 PG-13 96 74 2016 9 14765703.000 8703.000 54735.000 2057.000 60000000 positive objective 0.477 1.000 20958976 16535902 9159624 9575722 10930973  
 PG-13 118 47 2016 10 125561.000 9.000 172.000 7.000 60000000 neutral subjective 0.762 1.000 -19997146 -3355187 16312415 16850658 16722852  
 PG 90 40 2017 4 867614.000 392.000 1808.000 327.000 60000000 positive objective 0.933 1.000 -8030088 0 9159624 20979198 22293110  
 ratingAR productionAR distributionAR Prediction 2.5% 97.5% +/-  
 8133748 -2697372 12943408 6635.208 4272.145 8998.272 2363.063  
 9899276 21745046 20762074 8250.936 5827.047 10674.824 2423.889  
 9899276 22280237 20132369 6753.973 4380.987 9126.959 2372.986  
 12706625 3463844 -395282 6096.868 3724.595 8469.142 2372.274  
 8133748 17163082 10358090 5636.303 3263.281 8009.326 2373.023  
 8133748 3463844 -395282 6127.905 3722.119 8533.692 2405.786  
 12706625 3463844 -395282 6568.314 4179.568 8957.059 2388.746  
 9899276 -2697372 12943408 7495.798 5097.871 9893.724 2397.926  
 9899276 22280237 20132369 6903.025 4545.253 9260.797 2357.772  
 12706625 21745046 20441134 7133.263 4749.182 9517.344 2384.081

Here is a scatterplot comparing on the y axis the actual revenue and the x axis the predicted revenue for the second model.



This third model has the variables from the second model with the insignificant variables removed. The sum of squared error is slightly higher than the second model, however now all of the variables are at least significant at alpha = 0.1

Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: rated, min\_age, runtime.min., metacriticRating, YearReleased, commentCount, likeCount, dislikeCount, Production.Budget, actorAR, directorAR, genreAR, ratingAR, productionAR, distributionAR   
Null hyp.: the effect of x on Domestic.Total.Gross\_sqrt is zero  
Alt. hyp.: the effect of x on Domestic.Total.Gross\_sqrt is not zero  
  
 coefficient std.error t.value p.value   
 (Intercept) 77404.388 23095.053 3.352 < .001 \*\*\*  
 rated|NC-17 -2028.808 930.314 -2.181 0.029 \*   
 rated|NOT RATED -1431.888 483.564 -2.961 0.003 \*\*   
 rated|PG -332.156 400.240 -0.830 0.407   
 rated|PG-13 -449.088 386.663 -1.161 0.246   
 rated|R -725.869 385.482 -1.883 0.060 .   
 runtime.min. 8.092 3.024 2.676 0.008 \*\*   
 metacriticRating 9.157 2.392 3.827 < .001 \*\*\*  
 YearReleased -38.177 11.495 -3.321 < .001 \*\*\*  
 commentCount 0.108 0.030 3.602 < .001 \*\*\*  
 likeCount -0.014 0.005 -3.168 0.002 \*\*   
 dislikeCount -0.023 0.012 -1.850 0.065 .   
 Production.Budget 0.000 0.000 26.641 < .001 \*\*\*  
 actorAR 0.000 0.000 26.080 < .001 \*\*\*  
 directorAR 0.000 0.000 8.144 < .001 \*\*\*  
 genreAR 0.000 0.000 1.991 0.047 \*   
 productionAR 0.000 0.000 2.674 0.008 \*\*   
 distributionAR 0.000 0.000 9.510 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.853, Adjusted R-squared: 0.85   
F-statistic: 312.512 df(17,915), p.value < .001  
Nr obs: 933   
  
The set of explanatory variables exhibit perfect multicollinearity.  
One or more variables were dropped from the estimation.  
Sum of squares:  
 df SS  
Regression 17 7507084779  
Error 915 1292934278  
Total 932 8800019056  
  
Multicollinearity diagnostics were not calculated.



Linear regression (OLS)  
Data : outTrain   
Response variable : Domestic.Total.Gross\_sqrt   
Explanatory variables: rated, min\_age, runtime.min., metacriticRating, YearReleased, commentCount, likeCount, dislikeCount, Production.Budget, actorAR, directorAR, genreAR, ratingAR, productionAR, distributionAR   
Prediction dataset : outTest   
Rows shown : 10 of 318   
  
 rated min\_age runtime.min. metacriticRating YearReleased commentCount likeCount dislikeCount Production.Budget actorAR directorAR genreAR ratingAR productionAR distributionAR Prediction 2.5% 97.5% +/-  
 R 17 113 40 2013 4.000 23.000 1.000 60000000 -15261663 30522105 13464425 8133748 -2697372 12943408 6658.621 4305.033 9012.208 2353.588  
 PG-13 13 130 33 2013 11440.000 29867.000 1042.000 60000000 0 0 23008630 9899276 21745046 20762074 8271.905 5876.801 10667.010 2395.105  
 PG-13 13 105 57 2014 145.000 382.000 63.000 60000000 -4866478 -27365284 21913317 9899276 22280237 20132369 6912.604 4560.846 9264.362 2351.758  
 PG 7 96 61 2014 169.000 1159.000 42.000 60000000 -12636329 0 22293110 12706625 3463844 -395282 6215.541 3853.367 8577.715 2362.174  
 R 17 107 27 2015 885.000 9363.000 527.000 60000000 -29182980 -14724554 18782711 8133748 17163082 10358090 5533.265 3181.768 7884.762 2351.497  
 R 17 99 28 2016 1084.000 3003.000 347.000 60000000 0 0 21913317 8133748 3463844 -395282 5763.059 3407.324 8118.794 2355.735  
 PG 7 101 84 2016 3426.000 28815.000 1522.000 60000000 -5875187 0 29429955 12706625 3463844 -395282 6505.576 4134.689 8876.464 2370.887  
 PG-13 13 96 74 2016 8703.000 54735.000 2057.000 60000000 20958976 16535902 10930973 9899276 -2697372 12943408 7519.612 5142.694 9896.530 2376.918  
 PG-13 13 118 47 2016 9.000 172.000 7.000 60000000 -19997146 -3355187 16722852 9899276 22280237 20132369 6894.507 4545.474 9243.539 2349.033  
 PG 7 90 40 2017 392.000 1808.000 327.000 60000000 -8030088 0 22293110 12706625 21745046 20441134 7030.963 4668.103 9393.822 2362.859

Finally, a scatterplot of the predictions for this third model with the actual revenue for the testing data.

