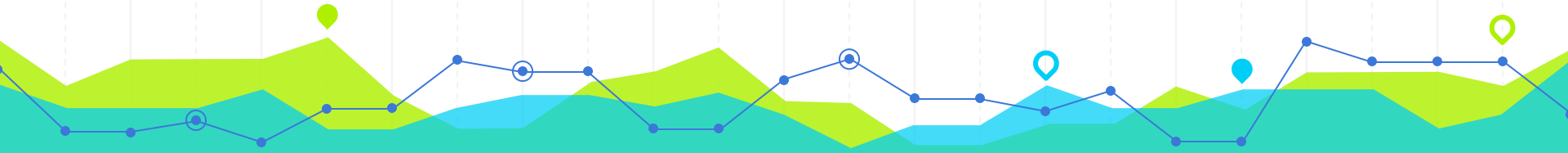
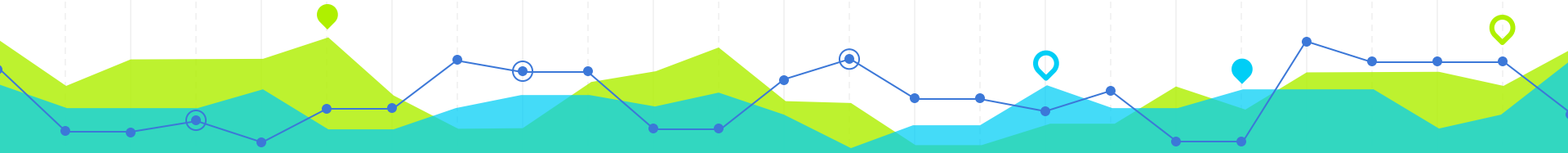


CREATING THE BEST MOVIE

Ketan Jog and Mark Shafran



The Business Problem

1

What is our objective?

Setting the Stage

We work for a major movie studio and the executives are noticing that the most recent movies have not been as successful.

- Low movie ratings on IMDb
- Decrease in profit

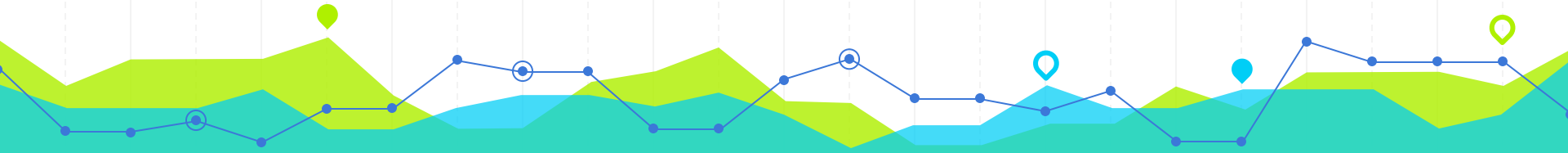
The Tasks

How to increase ratings?

What factors in a movie positively (or negatively) affect the average rating on IMDb?

How to increase profit?

What factors in a movie positively (or negatively) affect the profit? Should the movie target a specific demographic?



Data Collection

Collecting and cleaning data

2

Original Dataset: IMDb movies extensive database - Kaggle

Variables of interest:

title	year	genre	duration	budget	usa_gross_income	worlwide_gross_income
mean_vote	release_date	actors	date_of_birth (of actors)	allgenders_18age_avg_vote	allgenders_18age_votes	allgenders_30age_avg_vote
allgenders_30age_votes	allgenders_45age_avg_vote	allgenders_45age_votes	males_allages_avg_vote	males_allages_votes	males_18age_avg_vote	males_18age_votes
males_30age_avg_vote	males_30age_votes	males_45age_avg_vote	males_45age_votes	females_allages_avg_vote	females_allages_votes	females_18age_avg_vote
females_18age_votes	females_30age_avg_vote	females_30age_votes	females_45age_avg_vote	females_45age_votes	Production company	……(many more)…….

Data Cleaning

- Removed all observations with budget = 0 or N/A
- Dropped rows with non-numeric/corrupt entries
- Removed all observations with with year < 1990
- Converted all budgets/box office revenue from foreign currencies to USD
- Removed data from shutdown/old production companies
- Excluded movies in non-english languages, released in small countries

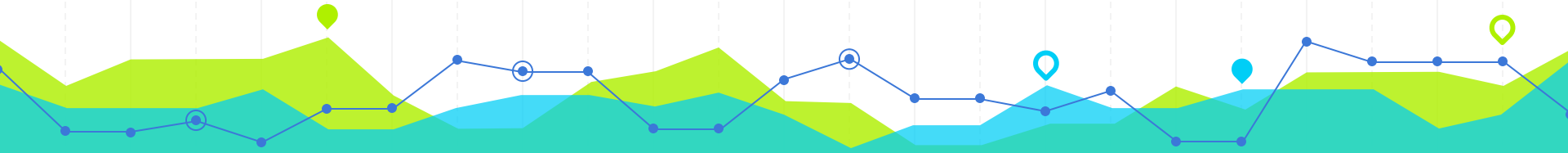
Variable Creation

- Avg_age = average age of actors in a movie (used actor data)
- Title_length = how many words long is the title of a movie
- Genre dummy variables = 1 if movie is of a certain genre, 0 otherwise. (Not mutually exclusive categories)
- US_profit = usa_gross_income - budget
- Month = 1-12 for corresponding month of release (January - December)

Final Dataset

- 17857 observations
  - 3377 data points belong to USA
- Variables of interest:

	month	avg_age	duration	budget	title_length
us_profit	mean_vote	allgenders_18age_avg_vote	allgenders_18age_votes	allgenders_30age_avg_vote	allgenders_30age_votes
allgenders_45age_avg_vote	allgenders_45age_votes	males_allages_avg_vote	males_allages_votes	males_18age_avg_vote	males_18age_votes
males_30age_avg_vote	males_30age_votes	males_45age_avg_vote	males_45age_votes	females_allages_avg_vote	females_allages_votes
females_18age_avg_vote	females_18age_votes	females_30age_avg_vote	females_30age_votes	females_45age_avg_vote	females_45age_votes
action	family	music	crime	fantasy	nonfiction
comedy	drama	horror	romance	sport	western



Increasing Ratings

Running a regression on mean_votes

3

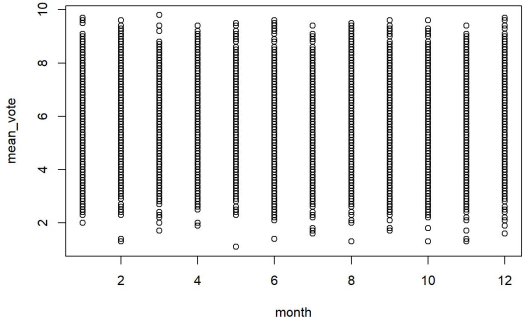
Step 1: Run full regression and check for multicollinearity

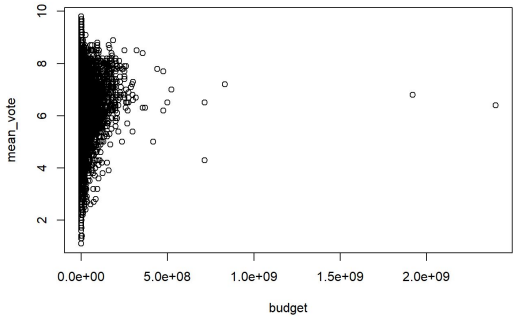
- Correlation matrix shows no values close to 1 or -1, so passes basic test for multicollinearity

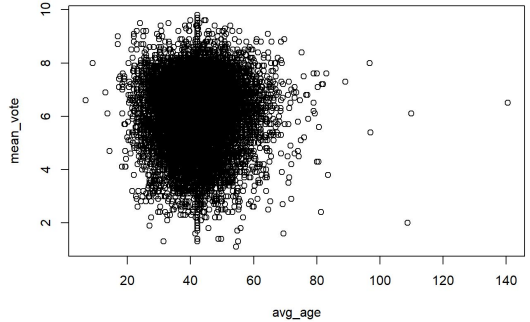
```
##
## Call:
## lm(formula = mean_vote ~ month + avg_age + budget + duration +
##     title_length + action + fam + music + crime + fantasy + nonf
##     comedy + drama + horror + romance + sport + western)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4945 -0.6095  0.0668  0.6823  5.1683
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.524e+00  7.018e-02  64.473  < 2e-16 ***
## month        -1.785e-04  2.327e-03  -0.077 0.938873
## avg_age      -3.692e-03  1.098e-03  -3.363 0.000773 ***
## budget        1.835e-09  2.175e-10   8.439  < 2e-16 ***
## duration      1.513e-02  4.301e-04  35.187  < 2e-16 ***
## title_length  2.859e-03  4.604e-03   0.621 0.534641
## action       -1.483e-01  1.937e-02  -7.654 2.05e-14 ***
## fam           3.150e-01  3.175e-02   9.920  < 2e-16 ***
## music         1.147e-01  4.800e-02   2.389 0.016926 *
## crime         1.090e-01  2.141e-02   5.094 3.54e-07 ***
## fantasy      -1.052e-01  2.696e-02  -3.900 9.65e-05 ***
## nonfiction    2.362e-01  3.203e-02   7.373 1.74e-13 ***
## comedy        4.336e-02  2.025e-02   2.141 0.032280 *
## drama         4.951e-01  1.992e-02  24.859  < 2e-16 ***
## horror       -6.000e-01  2.615e-02 -22.943  < 2e-16 ***
## romance       4.464e-03  2.457e-02   0.182 0.855829
## sport        -4.250e-02  6.307e-02  -0.674 0.500444
## western      -3.973e-01  1.045e-01  -3.802 0.000144 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.082 on 17839 degrees of freedom
## Multiple R-squared:  0.2271, Adjusted R-squared:  0.2263
## F-statistic: 308.3 on 17 and 17839 DF,  p-value: < 2.2e-16
```

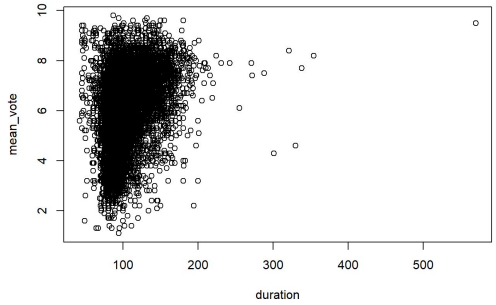
Step 2: Predictors v. Response

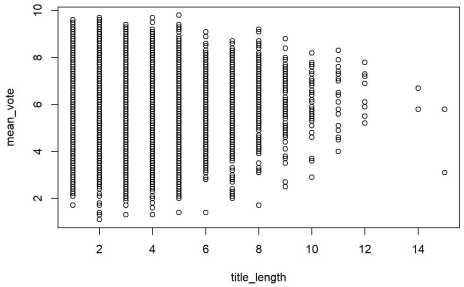
Note: some relationships look non-linear

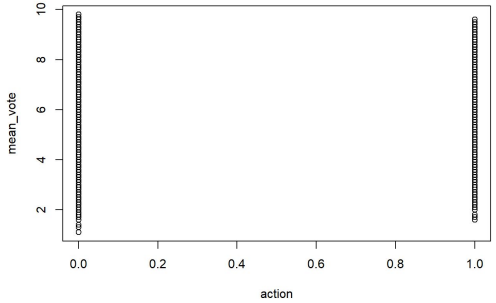










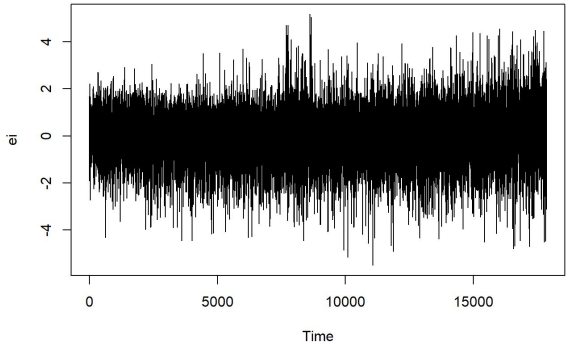


Step 3: Stepwise Variable Selection + Optimized Model

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	duration	addition	0.211	0.211	359.5890	53833.5625	1.0923
2	drama	addition	0.215	0.215	259.6690	53735.3291	1.0892
3	horror	addition	0.218	0.218	194.9790	53671.4377	1.0872
4	fam	addition	0.221	0.221	131.3530	53608.3593	1.0853
5	action	addition	0.223	0.223	84.9270	53562.1838	1.0839
6	budget	addition	0.225	0.224	57.7840	53535.1292	1.0830
7	nonfiction	addition	0.225	0.225	42.5780	53519.9538	1.0825
8	crime	addition	0.226	0.226	27.6130	53505.0017	1.0820
9	fantasy	addition	0.227	0.226	17.5820	53494.9703	1.0817
10	western	addition	0.227	0.226	13.8860	53491.2720	1.0816
11	avg_age	addition	0.227	0.226	10.8940	53488.2761	1.0814

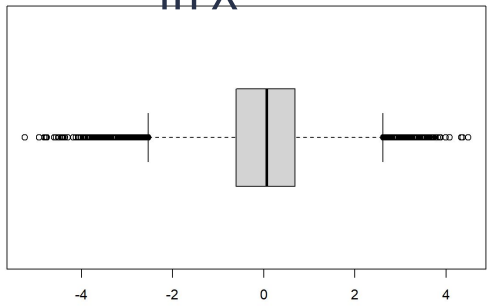
```
## Call:
## lm(formula = mean_vote ~ avg_age + budget + duration + action +
##     fam + music + crime + fantasy + nonfiction + comedy + drama +
##     horror + western)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4972 -0.6089  0.0674  0.6834  5.1664
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.531e+00  6.697e-02  67.652  < 2e-16 ***
## avg_age     -3.693e-03  1.094e-03  -3.374 0.000742 ***
## budget       1.847e-09  2.166e-10   8.526  < 2e-16 ***
## duration     1.513e-02  4.283e-04  35.327  < 2e-16 ***
## action      -1.482e-01  1.896e-02  -7.818 5.67e-15 ***
## fam          3.162e-01  3.135e-02  10.088  < 2e-16 ***
## music        1.165e-01  4.794e-02   2.431 0.015077 *
## crime        1.097e-01  2.119e-02   5.179 2.25e-07 ***
## fantasy     -1.044e-01  2.686e-02  -3.887 0.000102 ***
## nonfiction   2.358e-01  3.181e-02   7.411 1.31e-13 ***
## comedy       4.501e-02  2.014e-02   2.234 0.025466 *
## drama        4.950e-01  1.991e-02  24.867  < 2e-16 ***
## horror      -6.000e-01  2.591e-02 -23.154  < 2e-16 ***
## western     -3.955e-01  1.044e-01  -3.788 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.081 on 17843 degrees of freedom
## Multiple R-squared:  0.227,  Adjusted R-squared:  0.2265
## F-statistic: 403.1 on 13 and 17843 DF,  p-value: < 2.2e-16
```

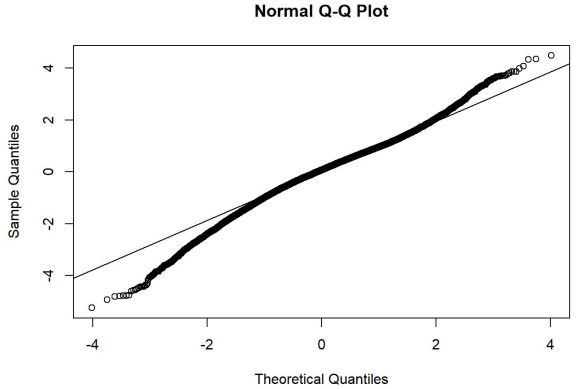
Diagnostics on Residuals: Outliers + Independence + Normality



- Independent

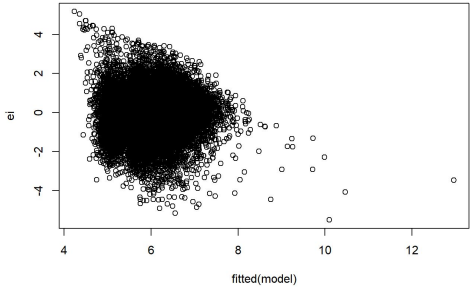
- Many Outliers
- Studentized deleted residuals
  - 4 outliers in Y
- Hat Matrix Leveraged Values
  - 1108 outliers in X

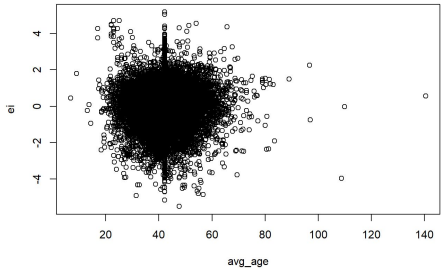


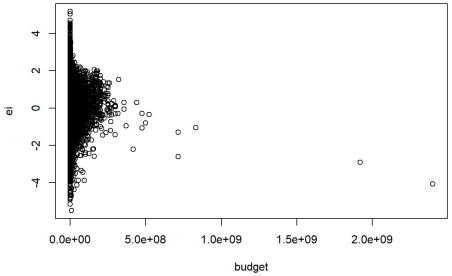


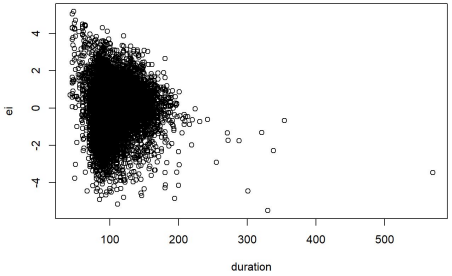
- Not normal
- Also Fails Corr. test for normality
-

Diagnostics on Residuals: Linearity + Constant Variance









- Clearly not linear
- Clearly do not have constant variance
  - Also fail BP Test
  - P-val <2.2e-16

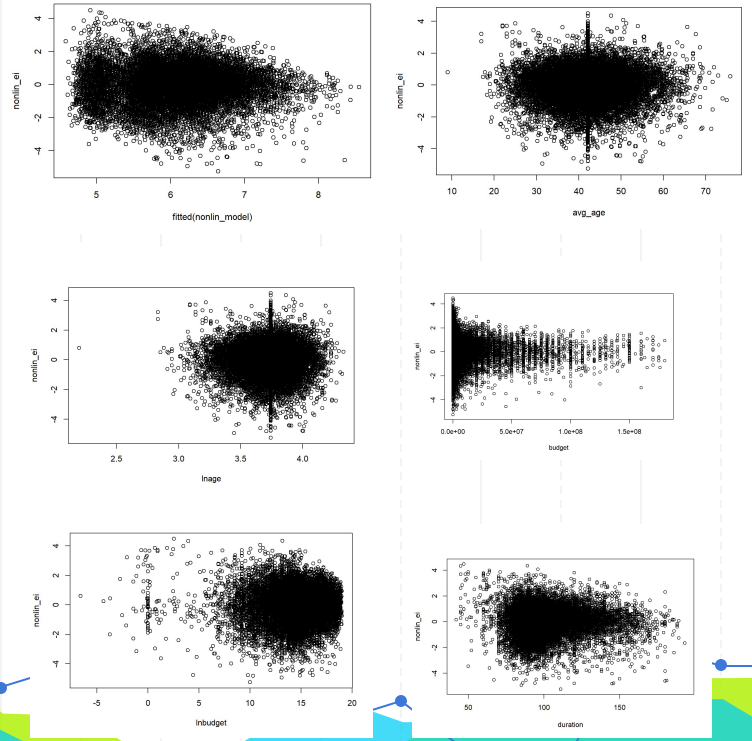
Remediation: Transformation

- Added 4 new variables
  - ln(avg_age)
  - ln(duration)
  - ln(budget)
- Re-ran full regression
- Stepwise Variable Selection

```
## Call:
## lm(formula = mean_vote ~ avg_age + lnage + budget + lnbudget +
##     duration + lnduration + action + fam + crime + fantasy +
##     nonfiction + comedy + drama + horror)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2457 -0.6050  0.0701  0.6854  4.4847
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.557e+01  1.551e+00  10.042  < 2e-16 ***
## avg_age      2.923e-02  7.811e-03   3.742 0.000183 ***
## lnage       -1.373e+00  3.182e-01  -4.314 1.61e-05 ***
## budget       5.644e-09  4.661e-10  12.108  < 2e-16 ***
## lnbudget    -3.693e-02  4.265e-03  -8.660  < 2e-16 ***
## duration     3.475e-02  3.297e-03  10.539  < 2e-16 ***
## lnduration  -1.917e+00  3.573e-01  -5.364 8.25e-08 ***
## action      -1.615e-01  1.942e-02  -8.313  < 2e-16 ***
## fam          2.915e-01  3.427e-02   8.506  < 2e-16 ***
## crime        1.331e-01  2.149e-02   6.192 6.09e-10 ***
## fantasy     -1.237e-01  2.800e-02  -4.416 1.01e-05 ***
## nonfiction   2.502e-01  3.401e-02   7.357 1.97e-13 ***
## comedy       6.540e-02  2.071e-02   3.159 0.001588 **
## drama        5.156e-01  2.044e-02  25.228  < 2e-16 ***
## horror      -6.273e-01  2.640e-02 -23.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 16734 degrees of freedom
## Multiple R-squared:  0.2397, Adjusted R-squared:  0.2391
## F-statistic: 376.9 on 14 and 16734 DF,  p-value: < 2.2e-16
```

Diagnostics: Outliers, Independence, Normality, Linearity, Constant Error Variance

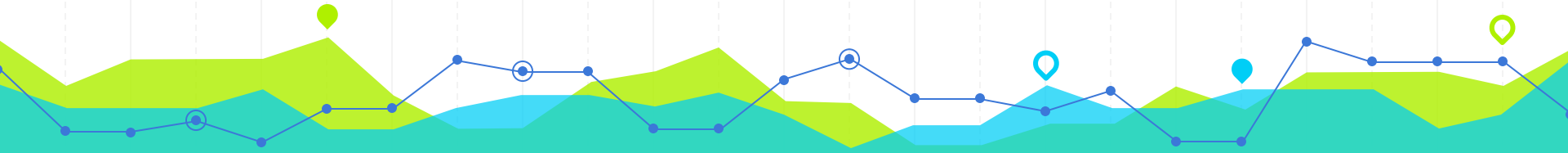
- Far less outliers
- Independent
- Still not normal
- Outliers look more linear
- Error variance still not constant



Regressions on Votes by Age and Gender

	All_18	All_30	All_45	Male_Avg	Male_18	Male_30	Male_45	Female_Avg	Female_18	Female_30	Female_45
intercept	-2.2560	-2.2520	-7.3760	-2.7130	-2.2330	-2.1950	-3.6640	-4.2930	-6.9550	-4.0370	-8.5210
month								0.0056	0.0055		0.0055
avg_age								0.0173		0.0182	
lnage	-0.2839	-0.2592		-0.1718	-0.2504	-0.2496		-0.8591	-0.2397	-0.8925	
budget	3.97E-09	4.13E-09	4.14E-09	4.38E-09	4.13E-09	4.43E-09	3.92E-09	3.23E-09	3.40E-09	3.73E-09	3.50E-09
lnbudget							0.0197			-0.0269	-0.0157
duration			-0.0092					-0.0093	-0.0121	-0.0096	-0.0138
lnduratic	2.0100	1.9400	3.0290	1.9740	1.9760	1.9110	1.9490	2.9690	3.2990	3.0200	3.4990
title_len	-0.0095			-0.0100	-0.0135		-0.0112				
action	-0.1051	-0.1004	-0.0861	-0.0966	-0.1036	-0.0970	-0.0663	-0.0979	-0.0970	-0.1008	-0.0484
fam	0.1688	0.1473	0.1983	0.1110		0.0709	0.1536	0.3660	0.3307	0.3762	0.3338
music		-0.1624	-0.1917	-0.2084	-0.1527	-0.2426	-0.2357				
crime	0.1072	0.0916	0.0545	0.0997	0.1230	0.1059	0.0798			0.0395	
fantasy	-0.1747	-0.1447	-0.1331	-0.1463	-0.1717	-0.1538	-0.1213	-0.1039	-0.1763	-0.0766	-0.0953
nonfictic	0.1375	0.1568	0.2164	0.1772	0.1540	0.1687	0.2218	0.1706	0.0979	0.1529	0.2669
comedy			4.80E-03					-0.0797	-0.0966	-0.0692	-0.0991
drama	0.4169	0.4214	0.4410	0.4045	0.3841	0.4032	0.4494	0.4223	0.3642	0.4130	0.4349
horror	-0.4386	-0.3426	-0.2819	-0.3359	-0.4510	-0.3362	-0.2514	-0.4345	-0.5154	-0.4213	-0.3509
romance	-0.1312	-0.1250	-0.0780	-0.1552	-0.1894	-0.1696	-0.0986	-0.1175	-0.1457	-0.1271	-0.0825
sport											
western			-0.2183					-0.2100		-0.2286	-0.2012
Mult. R^2	0.2198	0.2146	0.2198	0.2189	0.2028	0.204	0.2438	0.2535	0.2185	0.2265	0.2507
Adj. R^2	0.217	0.2119	0.217	0.2159	0.1999	0.2012	0.2409	0.2501	0.2155	0.2229	0.2475

p-val < 0.001
p-val < 0.01
p-val < 0.05
p-val < 0.10



Increasing Profit

4

Running a regression on us_profit

Using no new Variables

- Only explains ~10% variance
- Title length and duration positively impact profit
- Older actors have a negative impact
- Check: sort and split data. Compare mean profit value

```
## Call:
## lm(formula = us_profit ~ month + avg_age + budget + duration +
##     title_length + action + fam + music + crime + fantasy + nonfiction +
##     comedy + drama + horror + romance + sport + western)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -260388077  -17357242   -2936581    9333160  599825782
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.022e+07  8.483e+06  -3.562 0.000373 ***
## month         3.433e+05  2.492e+05   1.378 0.168445
## avg_age      -3.453e+05  1.192e+05  -2.896 0.003801 **
## budget        2.728e-01  2.754e-02   9.907  < 2e-16 ***
## duration      4.249e+05  6.202e+04   6.851 8.69e-12 ***
## title_length  1.106e+06  4.396e+05   2.516 0.011909 *
## action       -8.136e+05  2.201e+06  -0.370 0.711730
## fam           7.257e+06  2.995e+06   2.423 0.015458 *
## music         1.859e+05  4.135e+06   0.045 0.964149
## crime        -5.316e+06  2.150e+06  -2.473 0.013447 *
## fantasy       1.925e+06  2.793e+06   0.689 0.490655
## nonfiction   -7.988e+06  3.559e+06  -2.244 0.024871 *
## comedy        2.429e+06  2.186e+06   1.111 0.266496
## drama        -5.426e+06  2.182e+06  -2.486 0.012963 *
## horror        6.594e+06  3.407e+06   1.935 0.053040 .
## romance      -1.301e+06  2.331e+06  -0.558 0.576927
## sport        -1.393e+06  4.692e+06  -0.297 0.766560
## western      -3.067e+06  9.593e+06  -0.320 0.749220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48230000 on 3377 degrees of freedom
## Multiple R-squared:  0.1057, Adjusted R-squared:  0.1012
## F-statistic: 23.48 on 17 and 3377 DF,  p-value: < 2.2e-16
```

Include data on appeal by age

Anova agrees with addition:

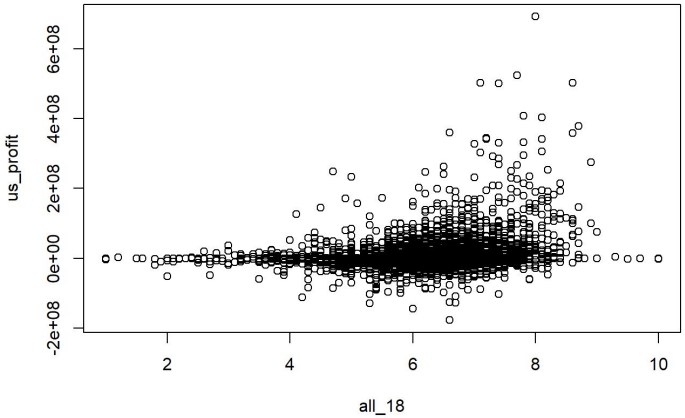
```
> anova(model_old, model_full)
Analysis of Variance Table

Model 1: us_profit ~ month + avg_age + budget + duration + action + fam +
    music + crime + fantasy + nonfiction + comedy + drama + horror +
    romance + sport + western
Model 2: us_profit ~ month + avg_age + budget + duration + action + fam +
    music + crime + fantasy + nonfiction + comedy + drama + horror +
    romance + sport + western + all_18 + all_30 + all_45
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   3378 7.8701e+18
2   3375 7.4266e+18  3 4.4353e+17 67.187 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## lm(formula = us_profit ~ month + avg_age + budget + duration +
##     title_length + action + fam + music + crime + fantasy + nonfiction +
##     comedy + drama + horror + romance + sport + western + all_18 +
##     all_30 + all_45)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -254897082  -19098841   -4148733   12139329  590618372
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.271e+07  9.052e+06  -9.137  < 2e-16 ***
## month         2.934e+05  2.423e+05   1.211 0.226031
## avg_age      -3.174e+05  1.172e+05  -2.708 0.006805 **
## budget        2.203e-01  2.702e-02   8.153 4.95e-16 ***
## duration      1.958e+05  6.244e+04   3.135 0.001731 **
## title_length  1.196e+06  4.272e+05   2.800 0.005140 **
## action        4.071e+05  2.141e+06   0.190 0.849215
## fam           5.006e+06  2.918e+06   1.716 0.086315 .
## music         2.027e+06  4.026e+06   0.503 0.614649
## crime        -6.303e+06  2.091e+06  -3.014 0.002598 **
## fantasy       3.796e+06  2.718e+06   1.397 0.162567
## nonfiction   -1.026e+07  3.466e+06  -2.961 0.003089 **
## comedy        2.764e+06  2.128e+06   1.299 0.193971
## drama        -1.112e+07  2.161e+06  -5.147 2.80e-07 ***
## horror        1.111e+07  3.333e+06   3.332 0.000871 ***
## romance      -9.862e+04  2.270e+06  -0.043 0.965343
## sport        -9.700e+05  4.560e+06  -0.213 0.831566
## western      -7.446e+05  9.326e+06  -0.080 0.936368
## all_18        4.054e+06  1.987e+06   2.040 0.041405 *
## all_30        1.218e+06  2.975e+06   0.409 0.682236
## all_45        7.668e+06  2.620e+06   2.927 0.003444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46860000 on 3374 degrees of freedom
## Multiple R-squared:  0.1565, Adjusted R-squared:  0.1515
## F-statistic: 31.29 on 20 and 3374 DF,  p-value: < 2.2e-16
```

Comparing Regression model using Anova

- Votes variable were correlated, we select age_18 based on F-value, eliminate multicollinearity

- Diagnostics suggested exponential relationship between age_vote
- R^2: 0.15 -> 0.17




```
Residual standard error: 46440000 on 3374 degrees of freedom
Multiple R-squared:  0.1715,    Adjusted R-squared:  0.1666
F-statistic: 34.92 on 20 and 3374 DF,  p-value: < 2.2e-16
```

Including Data on vote by Gender

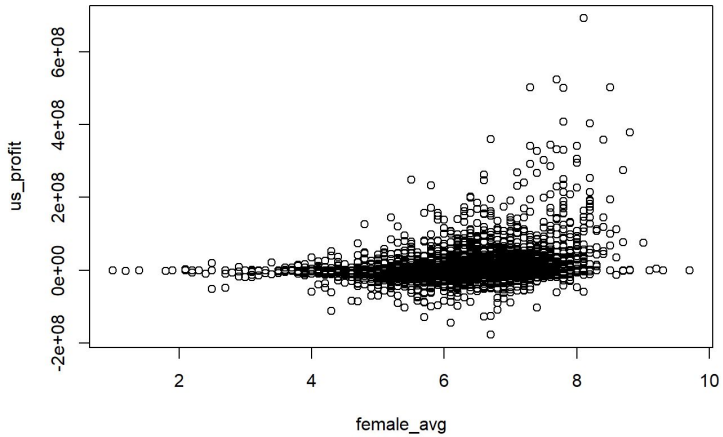
Anova agrees with addition

```
> anova(model_old, model_full)
Analysis of Variance Table

Model 1: us_profit ~ month + avg_age + budget + duration + action + fam +
    music + crime + fantasy + nonfiction + comedy + drama + horror +
    romance + sport + western
Model 2: us_profit ~ month + avg_age + budget + duration + action + fam +
    music + crime + fantasy + nonfiction + comedy + drama + horror +
    romance + sport + western + male_avg + female_avg
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   3378 7.8701e+18
2   3376 7.3600e+18  2 5.1009e+17 116.99 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = us_profit ~ month + avg_age + budget + duration +
    title_length + action + fam + music + crime + fantasy + nonfiction +
    comedy + drama + horror + romance + sport + western + male_avg +
    female_avg)

Residuals:
       Min         1Q     Median         3Q        Max
-254793692  -19094037   -3949942   12248967  587614555

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.229e+07  9.168e+06 -10.067  < 2e-16 ***
month        2.681e+05  2.411e+05   1.112 0.266381
avg_age     -2.849e+05  1.154e+05  -2.469 0.013612 *
budget       2.266e-01  2.689e-02   8.427  < 2e-16 ***
duration     1.866e+05  6.202e+04   3.009 0.002636 **
title_length 1.062e+06  4.260e+05   2.493 0.012712 *
action       5.799e+05  2.132e+06   0.272 0.785583
fam          2.326e+06  2.952e+06   0.788 0.430887
music        4.023e+05  4.016e+06   0.100 0.920207
crime       -5.690e+06  2.085e+06  -2.730 0.006374 **
fantasy      3.370e+06  2.705e+06   1.246 0.212902
nonfiction  -1.030e+07  3.446e+06  -2.988 0.002829 **
comedy       3.477e+06  2.122e+06   1.638 0.101436
drama       -1.154e+07  2.149e+06  -5.370 8.39e-08 ***
horror       1.281e+07  3.326e+06   3.852 0.000119 ***
romance      1.973e+05  2.259e+06   0.087 0.930407
sport       -1.293e+06  4.539e+06  -0.285 0.775801
western     -5.873e+04  9.282e+06  -0.006 0.994952
male_avg     7.801e+05  1.811e+06   0.431 0.666652
female_avg   1.331e+07  1.979e+06   6.725 2.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46660000 on 3375 degrees of freedom
Multiple R-squared:  0.1636, Adjusted R-squared:  0.1589
F-statistic: 34.75 on 19 and 3375 DF,  p-value: < 2.2e-16
```

Comparing Regression model using Anova

- Diagnostics suggested exponential relationship between age_vote




```
Residual standard error: 46100000 on 3375 degrees of freedom
Multiple R-squared:  0.1834,    Adjusted R-squared:  0.1788
F-statistic: 39.89 on 19 and 3375 DF,  p-value: < 2.2e-16
```

By Gender and Age Group

- We observe an exponential relationship across all voting categories with the profit, so we model it as such
- We get a 100% improvement in R square value

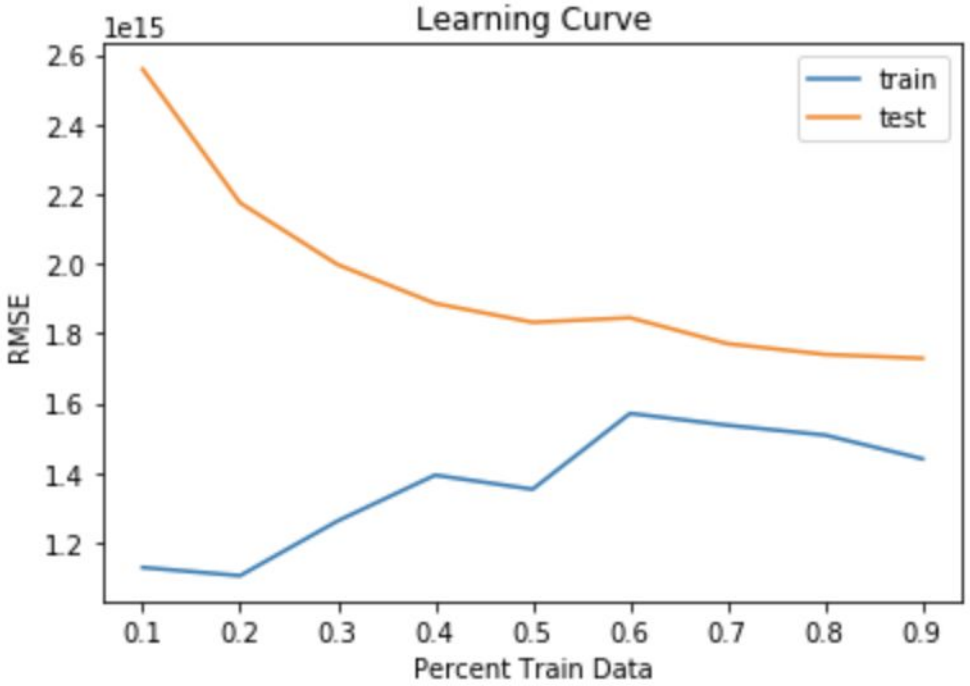
```
Residual standard error: 45520000 on 3366 degrees of freedom
Multiple R-squared:  0.2061,    Adjusted R-squared:  0.1995
F-statistic: 31.21 on 28 and 3366 DF,  p-value: < 2.2e-16
```

We look at number of votes by demographics

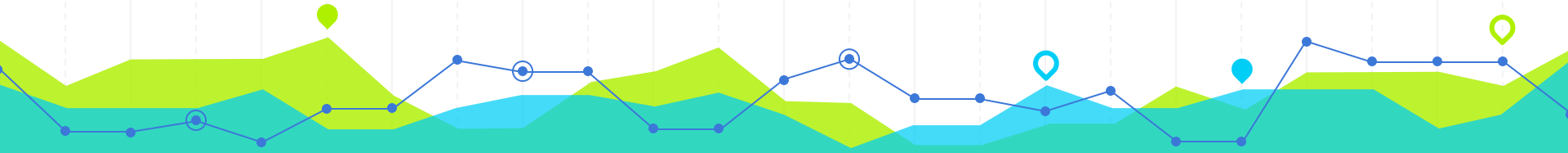
- We look at number of votes per demographic
- Relationships are linear
- R square improves by 250%

```
Residual standard error: 40800000 on 3366 degrees of freedom
  (254 observations deleted due to missingness)
Multiple R-squared:  0.3621,    Adjusted R-squared:  0.3568
F-statistic: 68.23 on 28 and 3366 DF,  p-value: < 2.2e-16
```

Learning Curve



Enough training data to make good inferences



Final Conclusions

Recommendations and findings

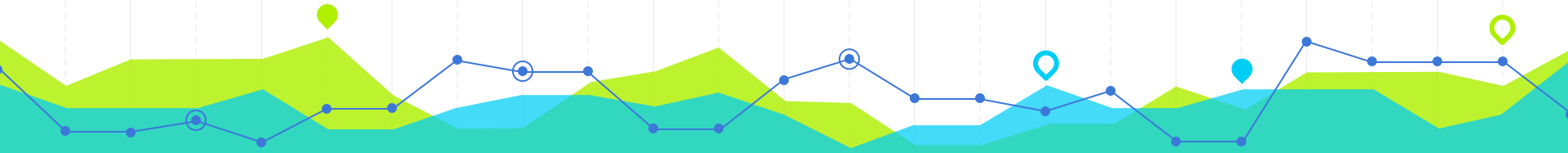
5

How to Improve Ratings (Based on F-Test + Coefficients)

- Overall
  - Lower avg age
  - Higher budget (as long as it is above 10M)
  - Longer duration
  - Genre = family, crime, nonfiction, comedy and/or drama
- Men vs Women
  - Men prefer shorter titles
  - Women don’t like comedies, men indifferent
  - Women like family genre, men indiff.
  - Men don’t like music genre, women indiff.
  - Men like crime genre, women indiff.
  - Women dislike horror more than men
  - Both men and women
    - Do not like action, fantasy, horror, romance
    - Like nonfiction and drama

How to Improve Profit

- Based on Mean Split
  - Choose younger actors
  - Tend to make movies longer
  - Lengthier titles matter
- Based on F-test and Anova
  - Viewers between 18 and 30
  - Female viewers
- Based on coefficient sign:
  - Stick to Action, Fantasy or Comedy genres
  - Avoid War, History and Non-fiction



Potential Improvements

6

How to further improve the model

Model Shortcomings + Potential Improvements

- predictors non-standardized, hard to compare coefficients
  - Improvement: standardize predictors
- Inflation was not accounted for in calculating budget data (Time range: 30 years)
  - Soln: Use conversion rates from when movie was released
- Non-constant error variance
  - Weighted Least Squares
- Non-normality of data
  - Box-Cox Transformation

THANK YOU!

Any questions?