WHAT MAKES A MOVIE A BOX OFFICE SUCCESS?

July 29, 2018

Team D
STAT 444 Final Project
Yuanjing Cai, Haonan Duan, Haidi Shui
Department of Statistics and Actuarial Science, University of Waterloo

Contents

0.1 INTRODUCTION

Movies play an important role in the modern entertainment industry. They provide a perfect bonding opportunity for friends and family. Film industry also creates thousands of employment opportunities, such as actors, actresses, cameramen, producers and company representatives. More importantly, movie industry is crucial for supporting the modern economy. In 2009, across major territories, there were over 6.8 billion cinema admissions creating global box office revenues of over US$30 billion.

Therefore, predicting box office of movies is of extreme importance. It can help producers to see what the market is like for the films, and what to develop or not to develop. Box office prediction will also help directors in casting, to see how that effected the data if it was showing the success rate as borderline. Company directors can also use the prediction results as a benchmark to validate their casting decisions, and also as a tool to help the investors see the value of the project.

Actually, there are already a lot of people working on box office prediction. According to *Waterloo Stories*, Jack Zhang, a graduate of Waterloo's mathematical economics program, founded a startup recently, called *BoxOfficePrediction*, to predict the minimum amount of revenue a movie will generate in the first week it's released ("Movies, math and money: Waterloo grad predicts box office revenues"). Zhang's model has caught the eye of the film industry. After attending the Toronto International Film Festival, Zhang has connected with the Canadian Film Centre accelerator program, ideaBOOST, which acts as a bootcamp for companies looking to bring new techniques and technologies into the entertainment ecosystem.

Our report aims to predict box office for movies based on their attributes via various predictive models. In general, our analysis will use 4 modeling techniques, smoothing spline, local regression, k-nearest neighbor, random forest and boosting method. To evaluate each model's predictive accuracy, we use 5-fold cross validation to compare their Average Prediction Squared Error (APSE).

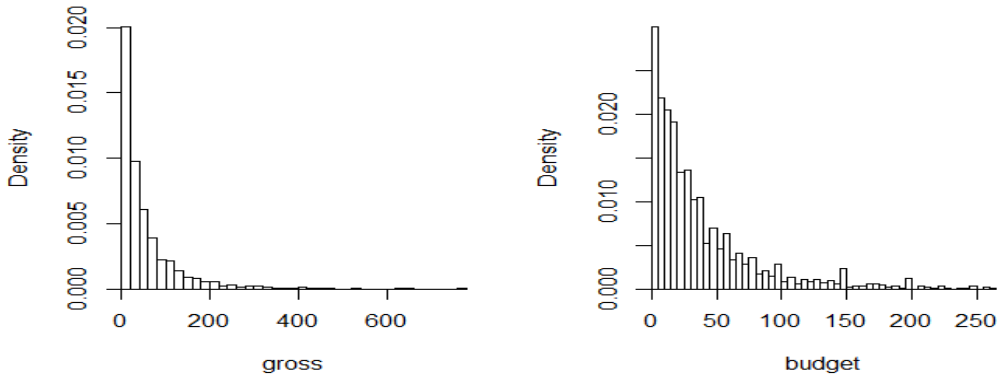
0.2 DATA VISUALIZATION

After removing observations containing missing values (≈ 13%), our dataset consists of 3756 observations, 1 response variable ("gross" - the box office a movie grossed (million USD)), and 23 covariates:

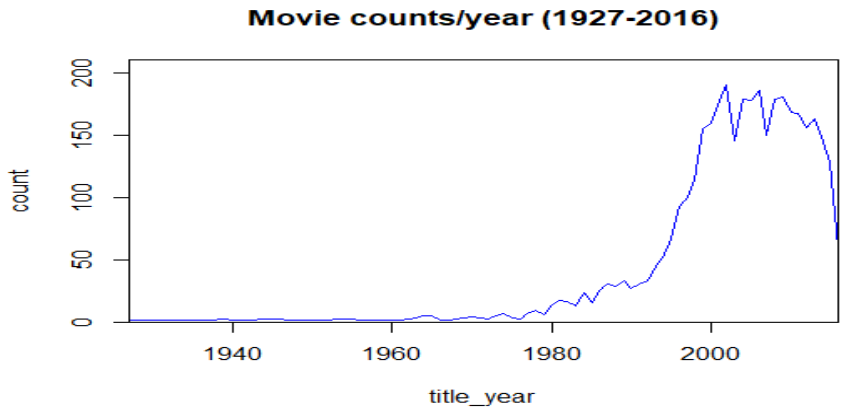
```
 [1] "color"                    "num_critic_for_reviews"   "duration"
 [4] "director_facebook_likes"  "actor_3_facebook_likes"   "actor_1_facebook_likes"
 [7] "gross"                    "genres"                   "num_voted_users"
[10] "cast_total_facebook_likes" "facenumber_in_poster"    "num_user_for_reviews"
[13] "language"                 "country"                  "content_rating"
[16] "budget"                   "title_year"               "actor_2_facebook_likes"
[19] "imdb_score"               "aspect_ratio"             "movie_facebook_likes"
[22] "total_facebook_likes"     "return_on_investment"     "profit_loss"
```

Below are some visualizations to give the audience an overview of some important variables in our dataset.

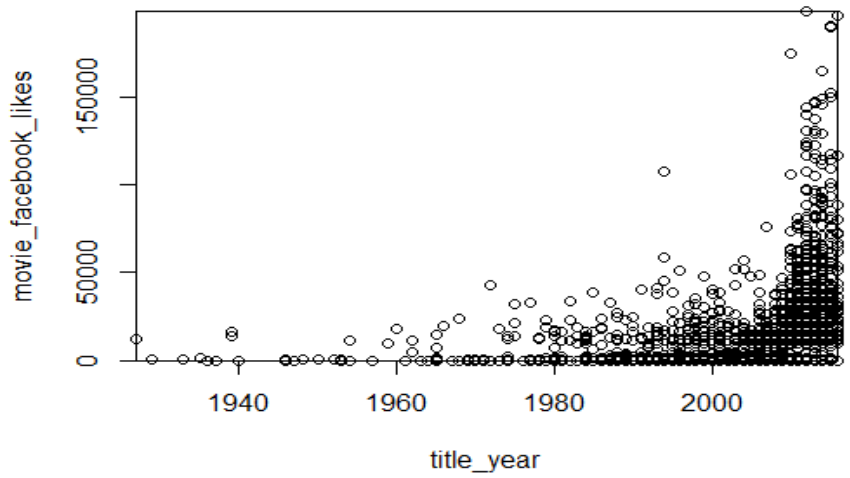
- The box office revenue and budget of the movies are both right-skewed.



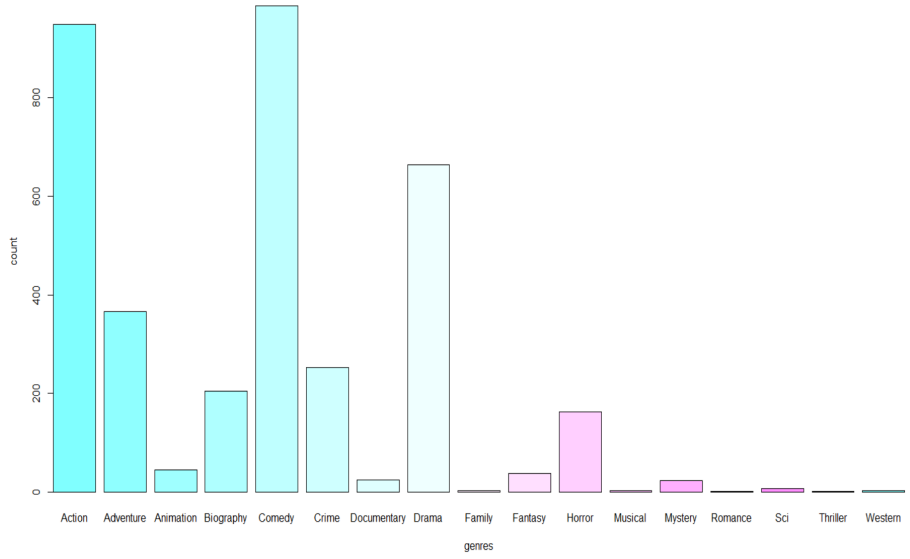
- The number of movies produced every year grows sharply since 1990, and remained at a high level until 2010.



- The number of facebook likes of the movies produced after 2010 almost tripled that of those produced before 2010, thanks to the rapid growth of Facebook around that time.



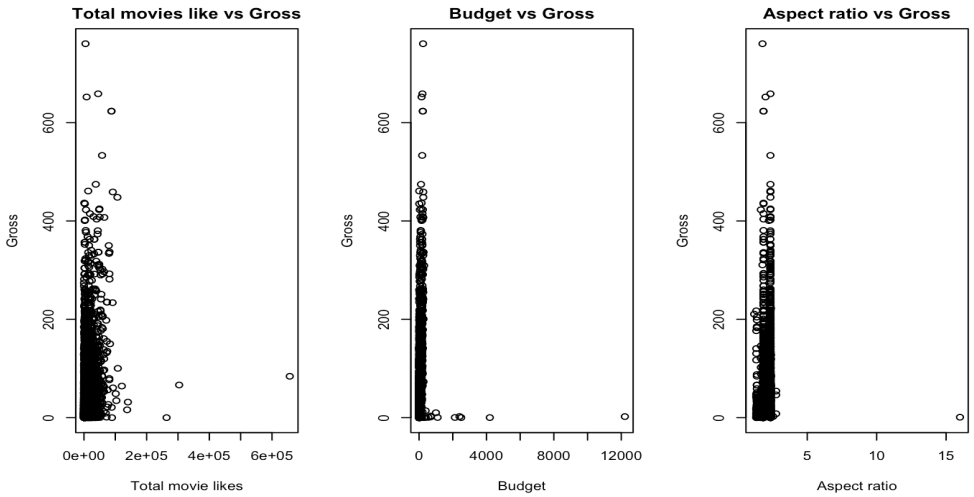
- Action, Comedy and Drama movies dominate the movie market in terms of quantity.



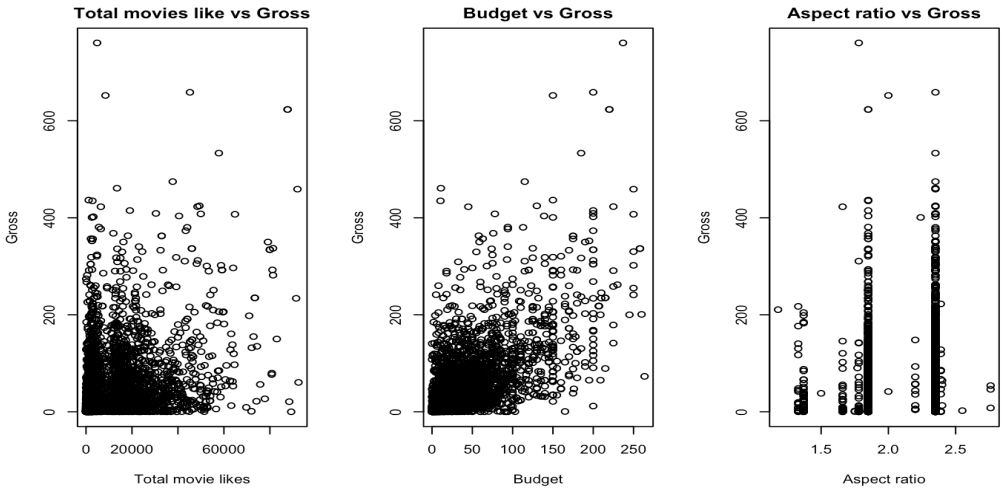
0.3 DATA PREPROCESSING

0.3.1 Remove outliers

We plan to remove some obvious outliers by looking at the scatter plots of each predictor against the response. Below are three scatter plots which has several obvious outliers.



The outliers make the above three plots look very strange. It may change the total shape of the fitting model, so we decide to remove them. Below are the 3 corresponding scatter plots after we remove them. We can see that the trend looks much more clear without the outliers.



0.3.2 Reduce predictors and level of categorical variables

The original dataset contains 23 predictors. Since large predictor numbers may increase the computational cost and model complexity, we decide to throw away some variates manually. Considering that $actor1FacebookLike$, $actor2FacebookLike$, $actor3Facebooklike$ and $totalFacebookLike$ represents similar things, we decide to throw away $actor1FacebookLike$, $actor2FacebookLike$ and $actor3FacebookLike$.

Also, for the language predictor, more than 90% movies use English in our dataset. So we decide to encode all other languages into "others" in the dataset. Similarly, we decide to encode all countries besides USA and UK into "others". This will reduce the difficulty in fitting the categorical variables.

0.4 MODEL FITTING

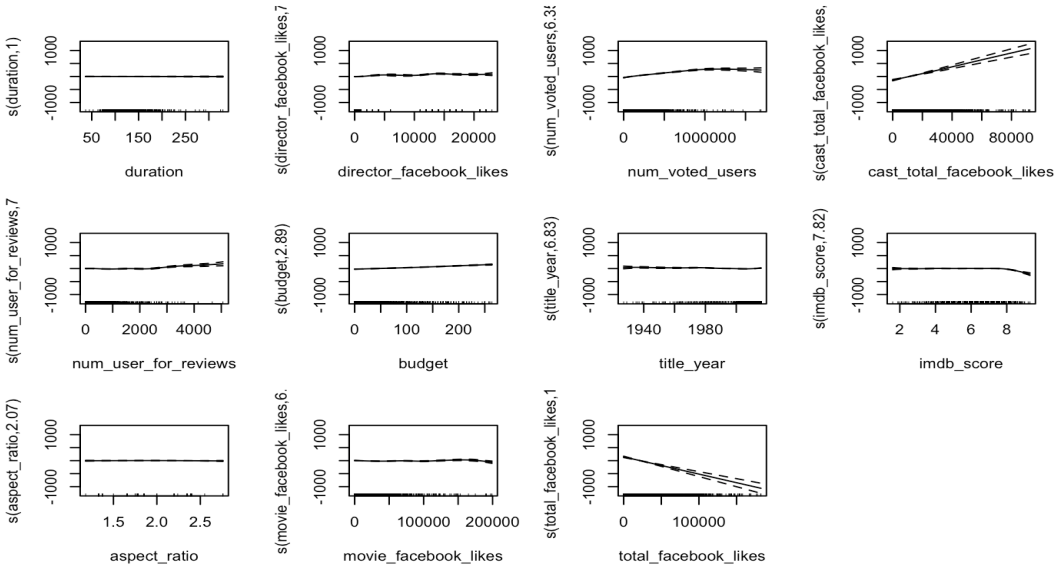
0.4.1 Spline method

We use bam function in the mgcv package to fit the spline method.

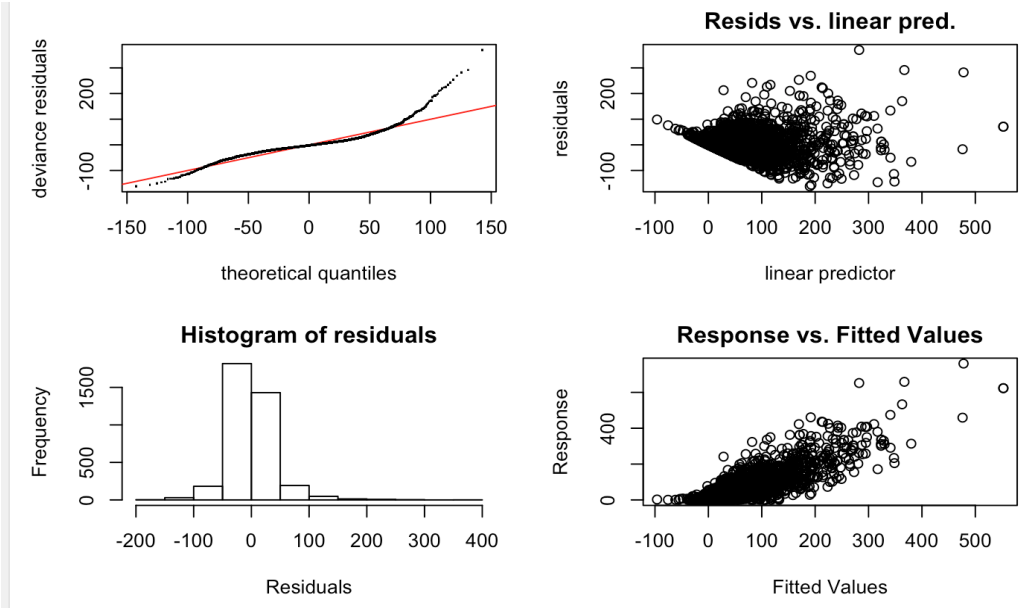
```
Parametric coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.110e+03  1.702e+02   6.523 7.81e-11 ***
colorColor                         1.659e+01  4.002e+00   4.145 3.47e-05 ***
duration                          -9.198e-02  3.817e-02  -2.410 0.016017 *
director_facebook_likes            6.957e-03  6.569e-04  10.592  < 2e-16 ***
num_voted_users                    1.779e-04  8.915e-06  19.953  < 2e-16 ***
cast_total_facebook_likes          1.530e-02  1.165e-03  13.129  < 2e-16 ***
num_user_for_reviews               9.028e-03  2.875e-03   3.140 0.001704 **
languageother                     -1.015e+01  3.943e+00  -2.573 0.010113 *
I(country == "USA")TRUE            1.194e+01  1.907e+00   6.261 4.25e-10 ***
I(content_rating == "Passed")TRUE -7.608e+01  2.742e+01  -2.775 0.005551 **
I(genres == "Adventure")TRUE       1.370e+01  2.517e+00   5.443 5.58e-08 ***
I(genres == "Animation")TRUE       2.743e+01  6.576e+00   4.172 3.10e-05 ***
I(genres == "Comedy")TRUE          6.782e+00  1.813e+00   3.740 0.000187 ***
I(genres == "Crime")TRUE          -1.194e+01  2.894e+00  -4.126 3.77e-05 ***
I(genres == "Family")TRUE          1.100e+02  2.468e+01   4.459 8.49e-06 ***
I(genres == "Horror")TRUE          8.100e+00  3.635e+00   2.228 0.025927 *
I(genres == "Musical")TRUE         8.145e+01  3.310e+01   2.461 0.013908 *
budget                             7.611e-01  2.133e-02  35.673  < 2e-16 ***
title_year                        -5.655e-01  8.513e-02  -6.643 3.52e-11 ***
imdb_score                         2.839e+00  8.592e-01   3.304 0.000964 ***
aspect_ratio                      -8.168e+00  2.879e+00  -2.837 0.004578 **
movie_facebook_likes               8.287e-05  4.316e-05   1.920 0.054969 .
total_facebook_likes              -7.793e-03  5.968e-04 -13.059  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


R-sq.(adj) =  0.629   Deviance explained = 63.2%
-REML =  19301  Scale est. = 1817.2    n = 3730
```

First, we fit a multiple linear regression model to do variable selection. According to the summary, we delete all variates that are not statistics significant. Above is the plot of the fitting spline on each predictor after deleting all irrelevant variables.

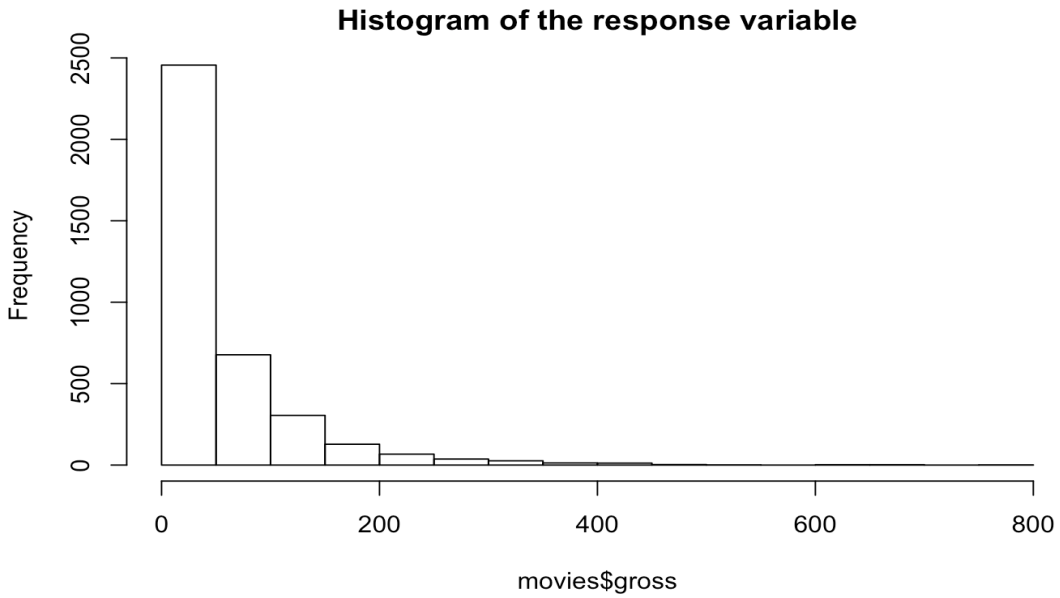
Then we use the smoothing spline to fit all remaining covariates. Below is the plot of the model.



Below is the diagnostic plot for the above model.



Below is the histogram of the response variable, total gross:



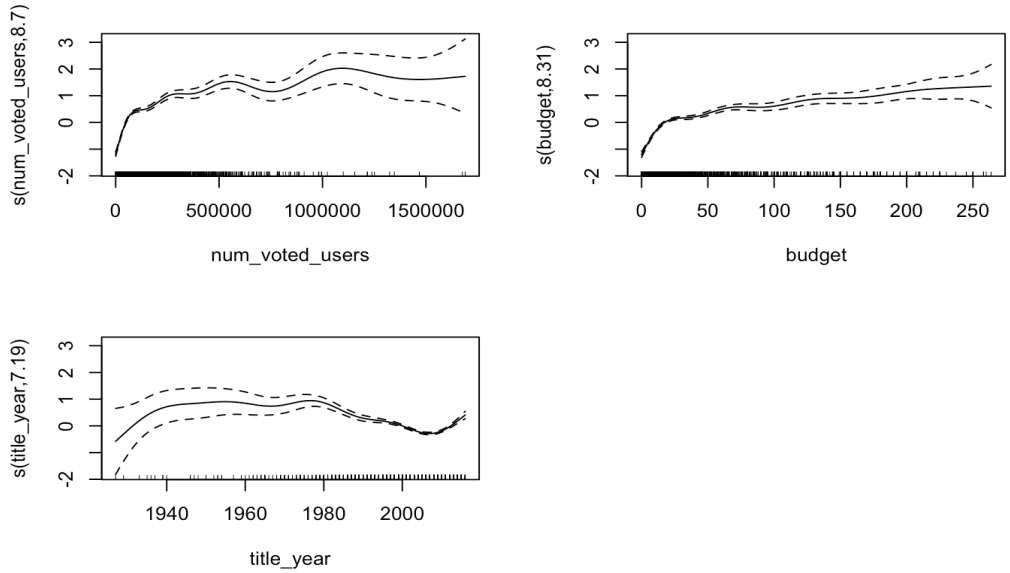
Here are 3 adjustments we decide to take after looking at the above three plots:

- We decide to use use gamma distribution with log link function instead of the default normal distribution to build the model for 3 reasons listed below. The reasons are as follow. First, the second diagonostic plot shows an obvious heteroscedasticity pattern, which implies that the normal distribution assumption may be violated. Second, the histogram of the response shows that distribution of gross is highly skewed. Third, the response, the box-office of a movie should always be positive.

- As we can see, there is an obvious decreasing linear relationship between total facebook likes and the total gross. If we think about it, this does not make sense in real life. How can movies with fewer facebook likes tend to gain more in gross? We investigate this problem and conclude that total facebook like is not a very good predictor by design. Many movies don't have a facebook page, especially those that were made before people started to use facebook. Hence this predictor is highly biased towards the recent movies. Therefore, we decide to drop this predictor.

- Looking at the first plot, we do not need to fit a smoothing spline matrix on every predictor. The linear term is fine.

After changing the underlying distribution to gamma distribution, dropping the said predictor and changing some into linear terms, below is the plot of the new model we have.



Now let's compare the model using the normal distribution and gamma distribution. Below is the result(mo3 uses normal distribution and mo7 uses gamma distribution).

```{r}
AIC(mo3, mo7)
```

	df <dbl>	AIC <dbl>
mo3	70.27858	38015.23
mo7	35.90785	33113.90

2 rows

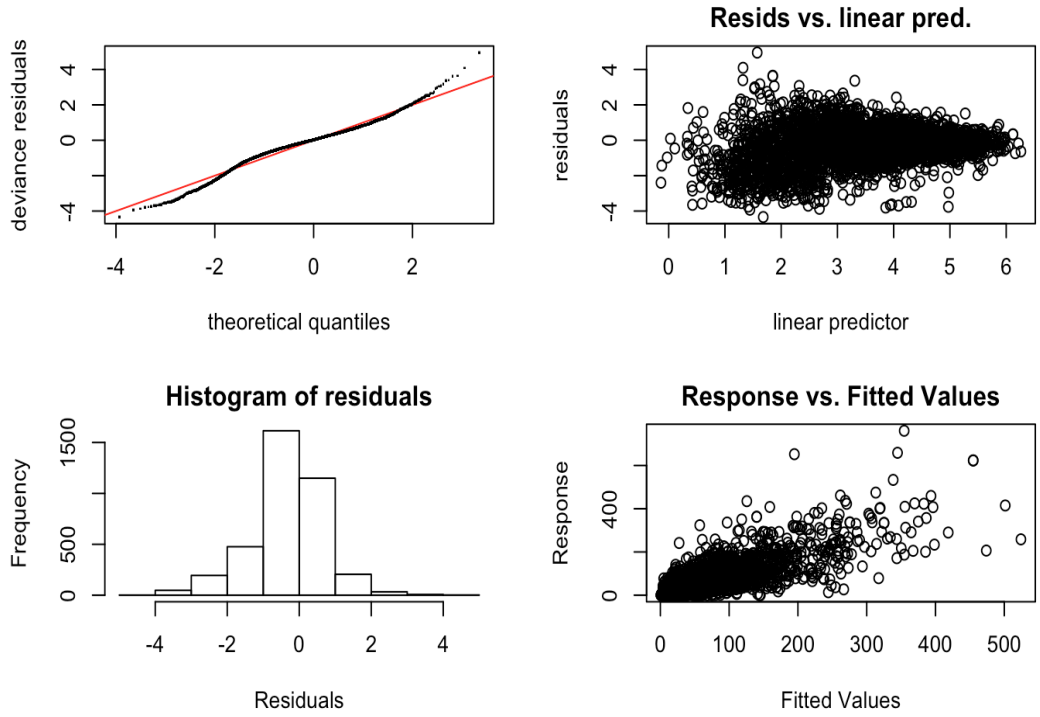
The above result shows that using gamma distribution decreases in both the degree freedom and AIC. Hence, our choice of using gamma distribution seem to be legit.

Next, we will investigate the interaction effect between predictors. Below is some comparison between model with interaction terms and model without interaction terms. Mo7 is the model without the interaction term, and mo8, mo9 and mo10 are the models without the interaction terms.

	df <dbl>	AIC <dbl>
mo7	35.90785	33113.90
mo8	44.48905	33265.95
mo9	43.56659	33168.67
mo10	45.32842	33136.88

4 rows

We can see that adding the interaction terms do not decrease AIC. Hence we decide not to include the interaction terms in our model.

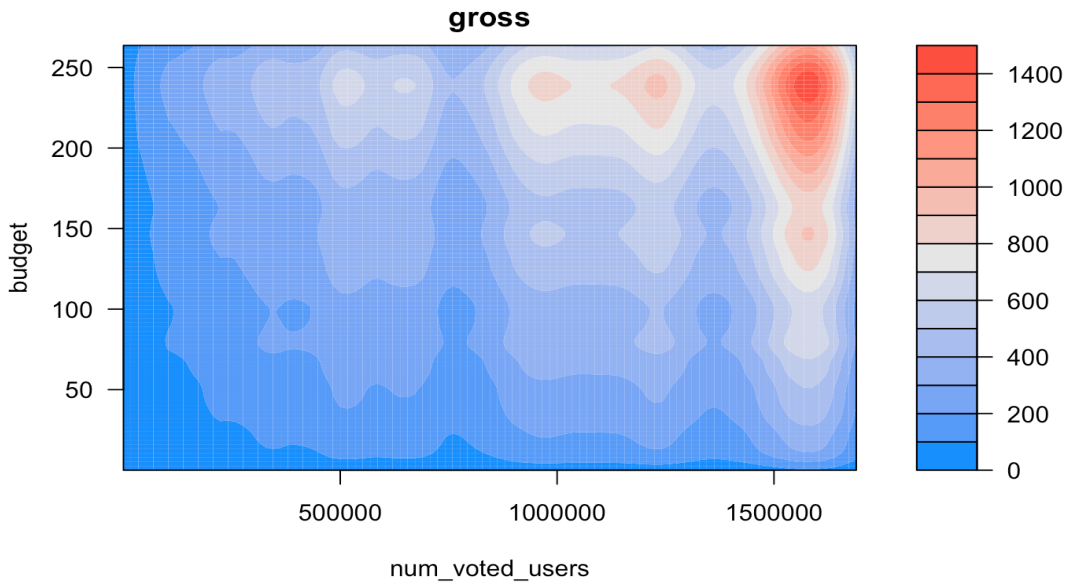


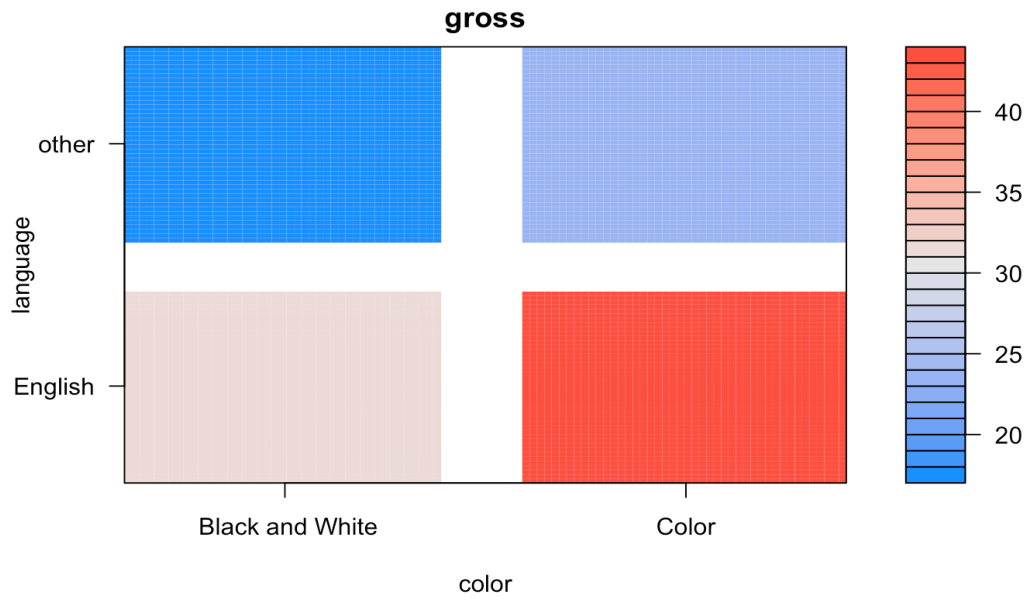
Compared with the previous diagnostic plot, the above plots show some promising improvements in that:

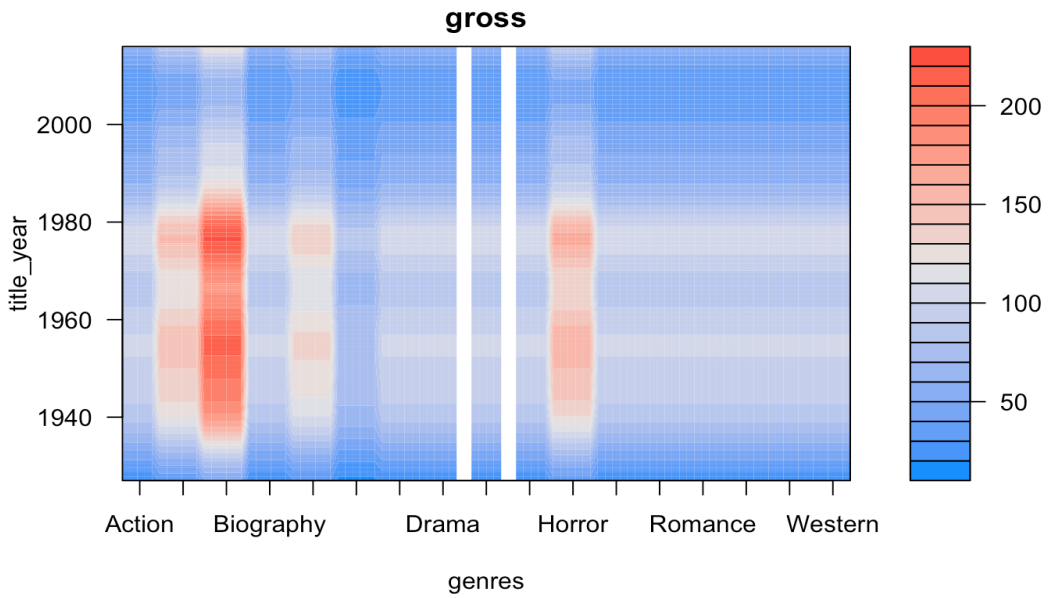
- The residual plot looks more random and shows less heteroscedasticity pattern.
- The range of residuals are much narrower.
- The deviance residuals fit better with the theoretical quantity.

Finally, since there are almost 4000 points, we decide to increase number of knots in the spline. We can see that AIC decreases by more than 100 after increasing the number of knots.

Below is some visualization of the final model.


gross
budget
250
200
150
100
50
500000
1000000
1500000
num_voted_users
1400
1200
1000
800
600
400
200
0


gross
language
other
English
Black and White
Color
color
40
35
30
25
20



Here are some simple and direct conclusions from those plots:

- In general, the more budget and more number of voted users, the higher gross a move can gain.
- Color movie in English gain the highest gross, while black-and-white movie in laguages other than English gain the lowest gross.
- Movie gross has declined since 1980. It is consistent with recent research that "movie industry is declining"("Is the Cinema Industry Set For A Clear Downfall?", Dina Zipin).
- Action, Biography and Horror gain more favour among audience compared with other genres.

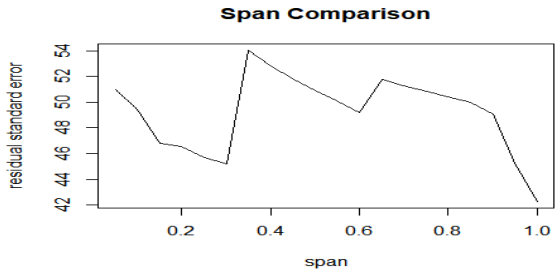
Using 5-fold cross validation, the average prediction squared error for the final model in the spline method is **1885.511**.

0.4.2 Local linear regression method

We use the simple loess to fit a locally weighted Sum of squares estimate to our data. Since we have five categorical predictors(color, genres, language, country and content rating), and each of them has 2, 17,2, 3 and 12 levels, we want to use the indicator method to fit the data aiming to select significant predictors. However, an error is thrown that for simple loess only 1-4 predictors are allowed, so we choose to use the most significant 4 predictors from

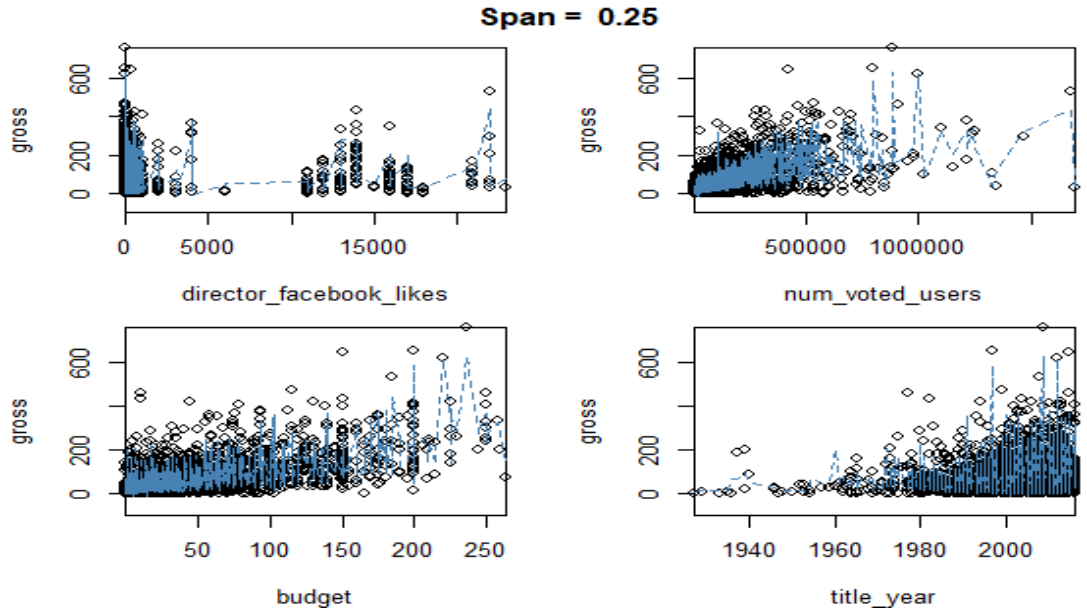
spline method (3.1) based on their pvalue, which are title_year, budget, num_vote_users and director_facebook_likes.

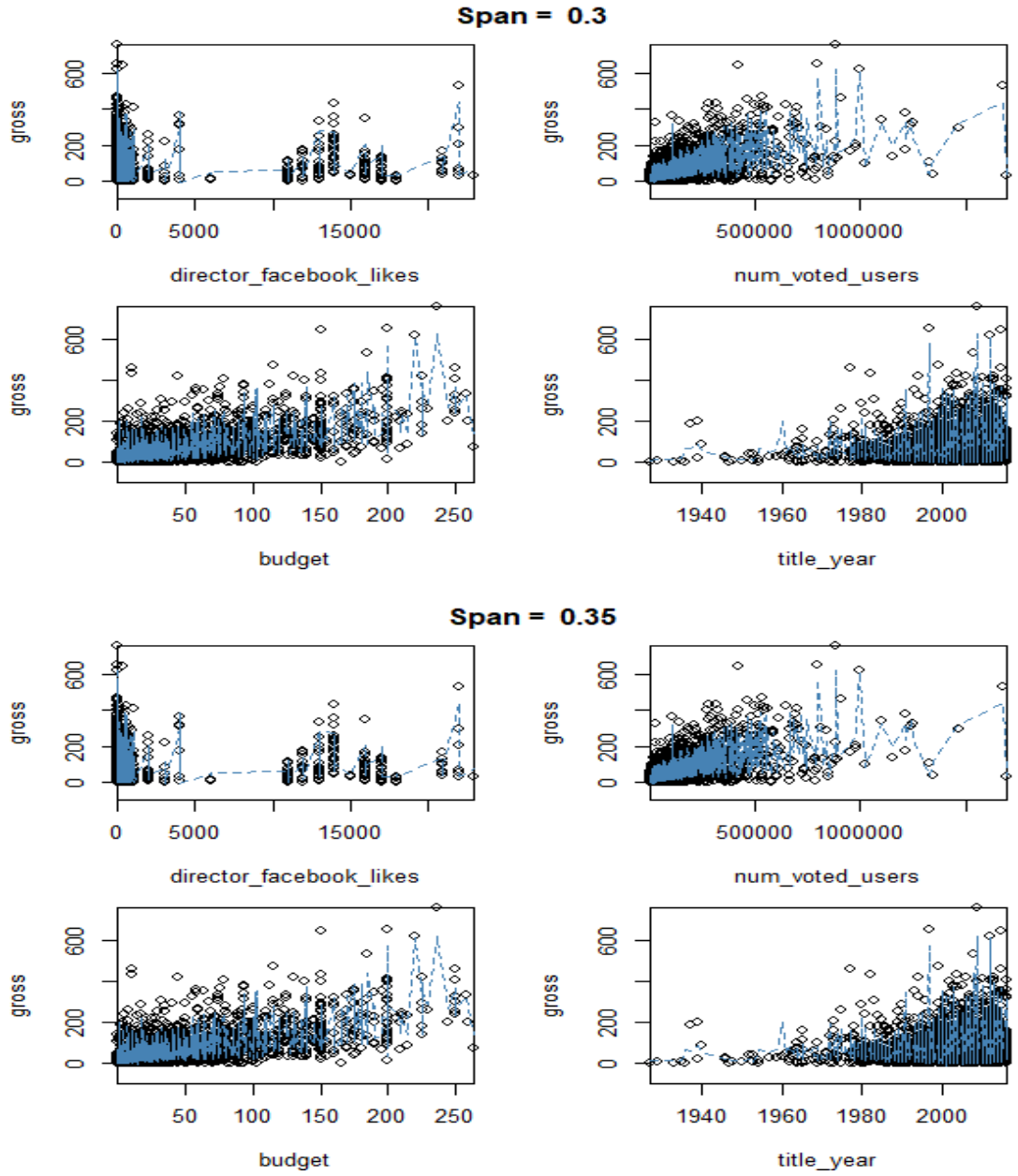
We then adjust the span to get models with different complexities. We select 20 different spans from 0.05 to 1, and get the RSS of each of these models.

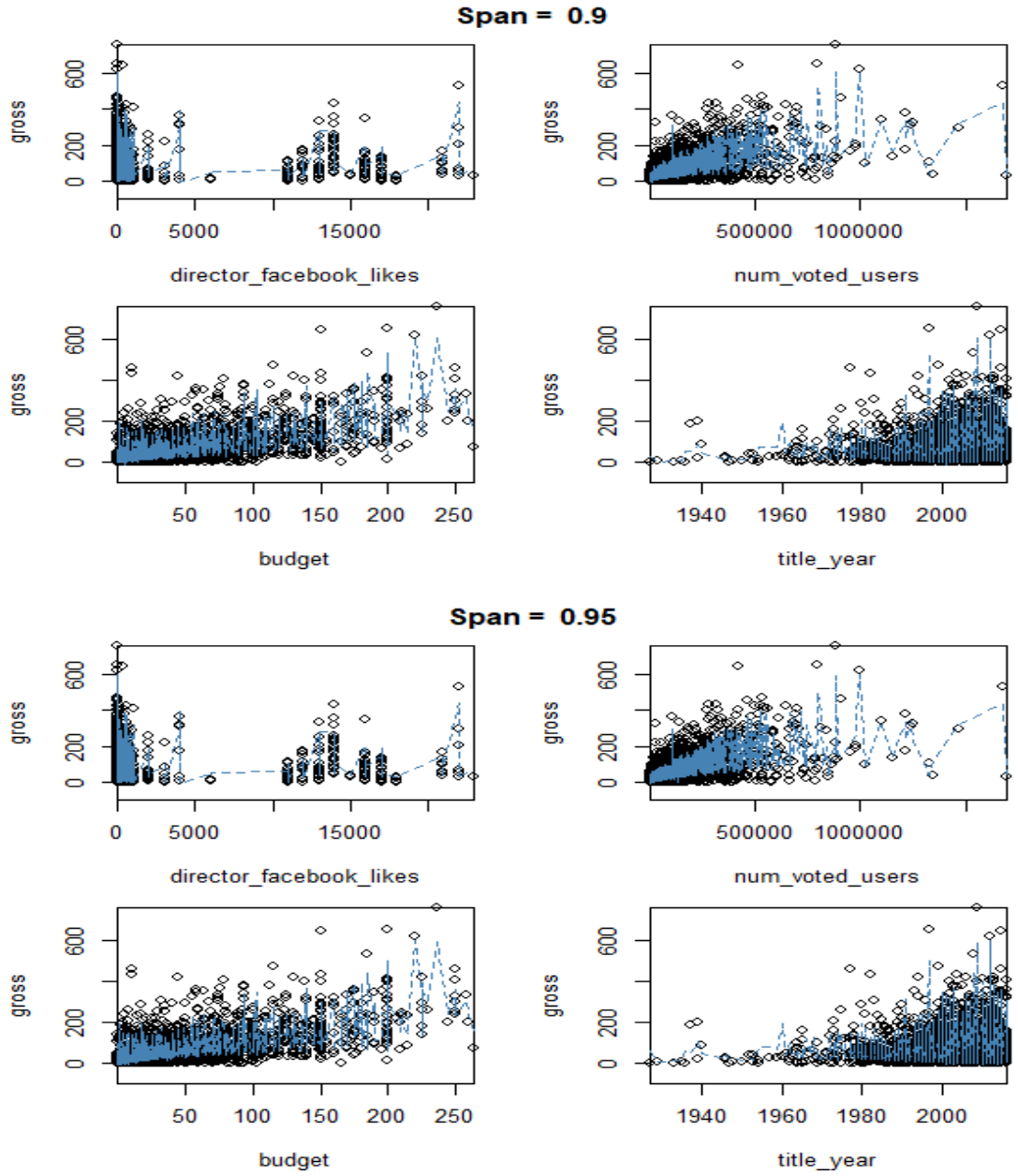


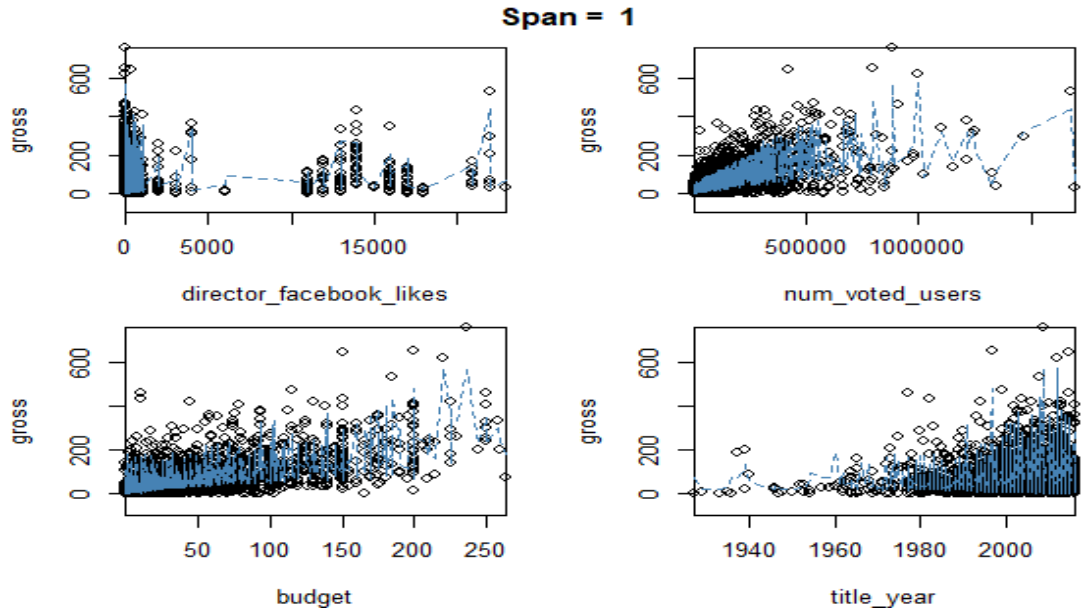
From the graph above, the residual standard error is fluctuating with the change of span. Our goal is to find the span that gives smallest predictive error, and the spans of 0.2-0.4 and 0.9-0.1 seem to have the lowest RSS, so we select six different spans 0.25, 0.3, 0.35, 0.9, 0.95, and 1.0 to see which model fits the data best.

For each span, we plot the fitted value of $gross$ versus each predictor.




Span = 0.3
gross
director_facebook_likes
num_voted_users
budget
title_year
Span = 0.35
gross
director_facebook_likes
num_voted_users
budget
title_year


Span = 0.9
gross
director_facebook_likes
num_voted_users
budget
title_year
Span = 0.95
gross
director_facebook_likes
num_voted_users
budget
title_year



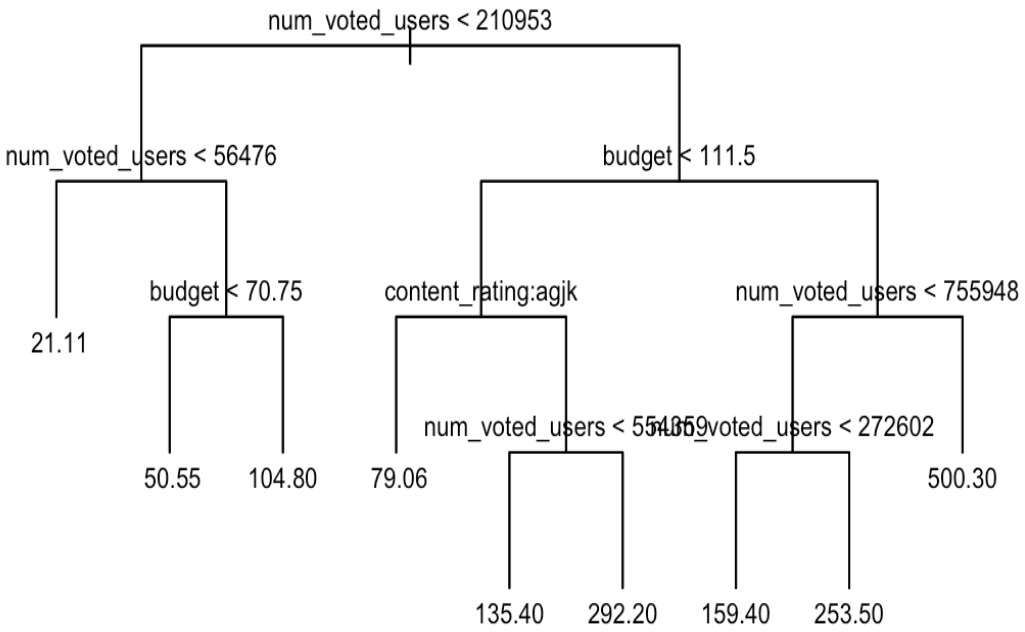
We can hardly tell any difference among the fitted lines across different spans. Therefore, we use 5-fold cross-validation to compare the average prediction squared error, and whichever gives the best prediction will be our final model for LOESS.

span	apse
0.25	1813.455
0.30	1802.372
0.35	1794.488
0.90	1783.010
0.95	1795.318
1.00	1794.566

In this case, we will choose span = 0.9, whose apse is **1783.010**.

0.4.3 Random forest method

First, let's build a single regression tree on the dataset to get a feel of what the tree model might look like. Below is the visualization of the model.



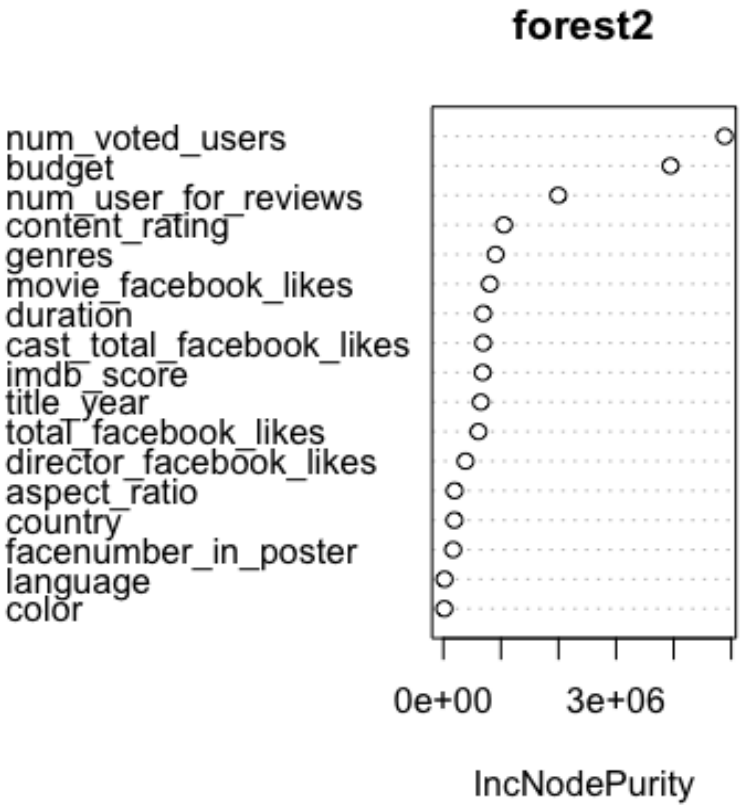
Here are some interesting insights we can gain from the above tree:

1. Number of voted users is the most important factor in determining the box office. The more voted users, the larger the box office.

2. When number of voted users is greater than 56476, budget begins to to play an important role. The greater the budget, the larger the box office. However, when number of voted users is below 56476, budget doesn't affect the box office at all.

3. The content rating will influence the box office only when number of content users is greater than 210953 and budget is smaller than 111.5. In this case, the higher the content rating, the greater the box office.

Then, let's build a random forest model on all variables with $m = 3$ using $randomforest()$. The summary of this model shows that 71.82% of variance is explained by this model. Below is the summary of the importance of each covariate.



From the above plot, we can see that number of voted users, budget, and number of users for reviews are three most important factors on influencing the total gross.

Using 5-fold cross validation, the average prediction squared error for the regression tree method is **1403.481**.

0.4.4 Boosting method

Boosting method "boosts" the performance of random forest by taking a linear combination of the predictions from all the trees in a forest. Starting from the random forest we have obtained in section 3.2, where 5 variates are randomly selected in each split, num_voted_users, $budget$, $num\_user\_for\_reviews$, $content\_rating$, $genres$ turn out to account for largest drop in RSS. To make the ASPE's comparable with and without boosting, we also use those 5 variables in boosting method, i.e. the formula we pass to the $gbm()$ function will be

$$gross \sim num_voted\_users + budget + num\_user\_for\_reviews + content\_rating + genres$$

Using $n.tree = 1000$ in the random forest and the default shrinkage parameter, the new model we obtained decreases the ASPE (5-fold cross-validation) by $\approx 13\%$. The relative influence of each variable is displayed below

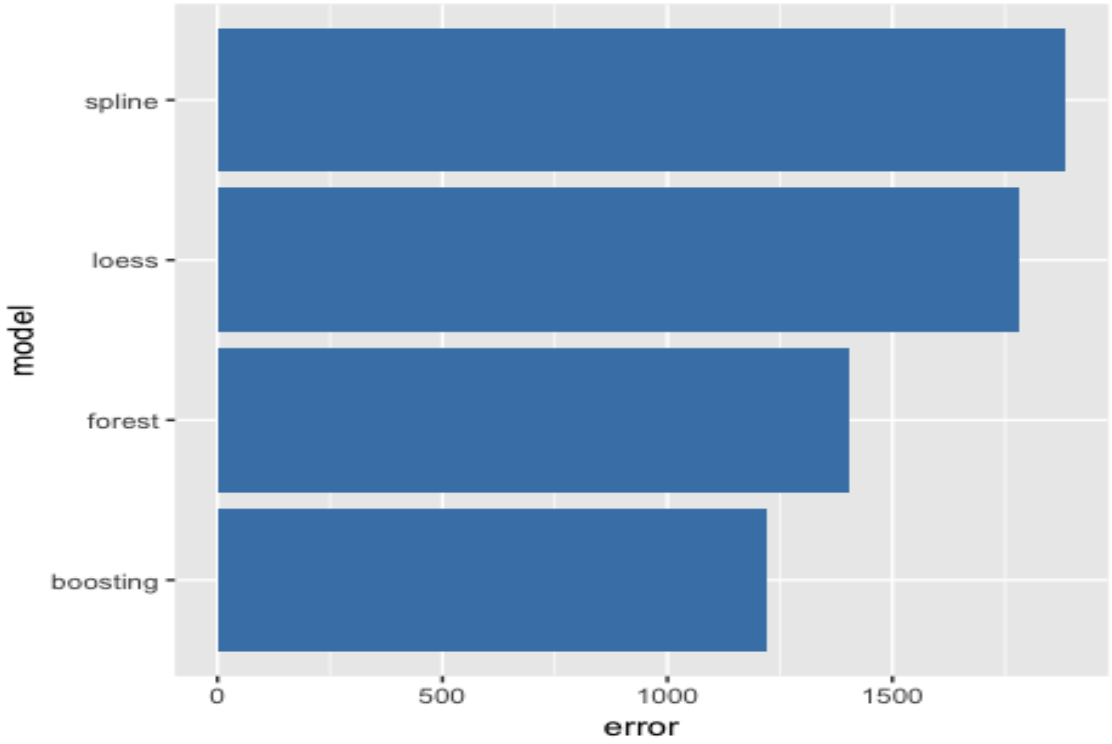
var	rel.inf
num_voted_users	47.984896
budget	31.836612
num_user_for_reviews	15.084165
content_rating	3.266412
genres	1.827915

Using 5-fold cross validation, the average prediction squared error for the final model in the boosting method is **1221.85**.

0.5 STATISTICAL COMPARISON AMONG 4 MODELS

0.5.1 Prediction accuracy

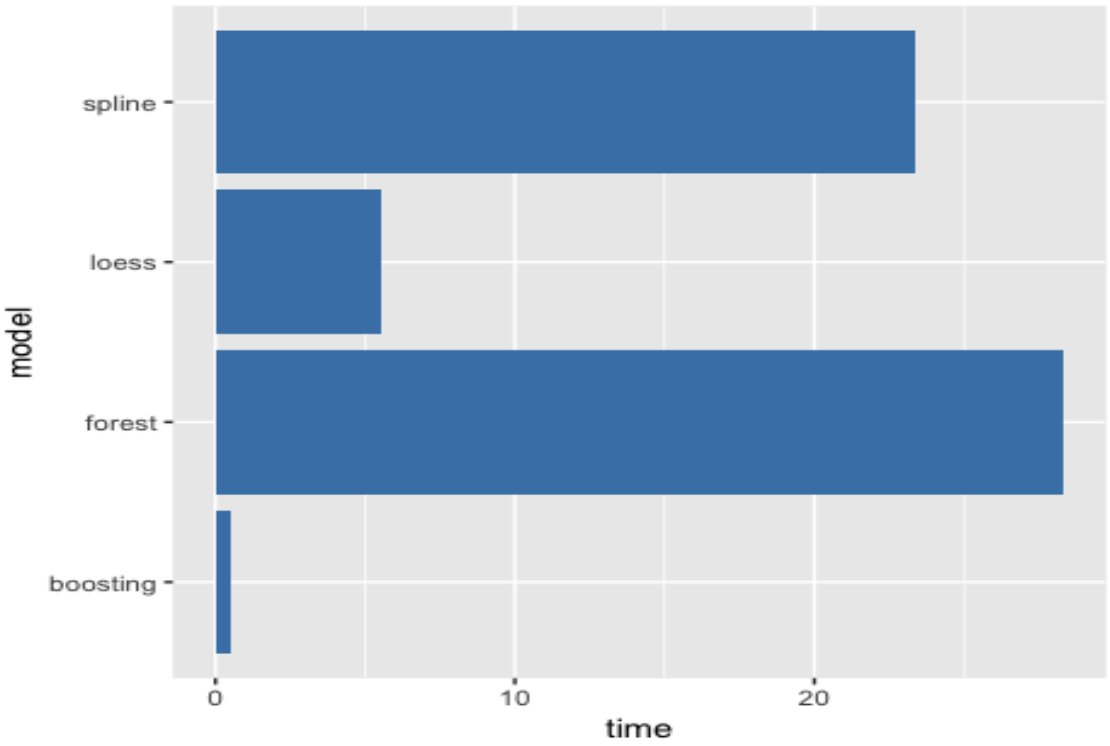
Below is the plot of prediction errors of 4 models using 5-fold cross validation:



Hence, in terms of prediction accuracy, **boosting method** is the best.

0.5.2 Computation time

We use $proc.time()$ in R to track the fitting time of each model. Below is the result.



Hence, in terms of computation time, **boosting method** is the best.

0.5.3 Easiness of interpretation

Spline method is the easiest to interpret. We can interpret this model using its coefficients and p-value. Also, spline method can be easily visualized, which helps us understand the model in a greater depth.

0.5.4 Final model

Since our goal is to choose the best predictive model, **boosting model** is selected as the final model for its low prediction error and computational time.

0.6 CONCLUSION AND INSIGHT

Even though spline and decision tree haven't been chosen as the final model, visualization of both models provide some valuable and interesting insights into the movie market.

- Number of voted users on *IMDb* is the best predictor on the box office. More people voted on the movie, the more money it will gain.
- In general, the highest grossing movies have the highest budget as well.
- Most popular movies use English as the language.
- Black and white movies make less money compared to color movies.
- Action, Biography and Horror gain more favour among audience compared with other genres.
- Movie industry is declining since 1980.

0.7 FUTURE WORK

- We wish that we could do more analysis on how genres influence the box office. It will be a very interesting topic. However, in the current data set, the number of movies in each genre is pretty uneven. This makes such analysis difficult to conduct.
- We wish to try a different method to deal with missing data. In the original analysis, we simply deleted all missing values. But we would like to see how the models will change if we deal with them in a different way.
- We think that this problem can also be investigated in a classification manner, that is, what movie is profitable and what movie will lose money. This is also an interesting and crucial question for the film makers. However, classification technique is beyond the scope of this course.

0.8 INDIVIDUAL CONTRIBUTION

Below is each group member's contribution:

Yuanjing Cai: Data visualization, boosting method, 5-fold cross-validation prediction error.

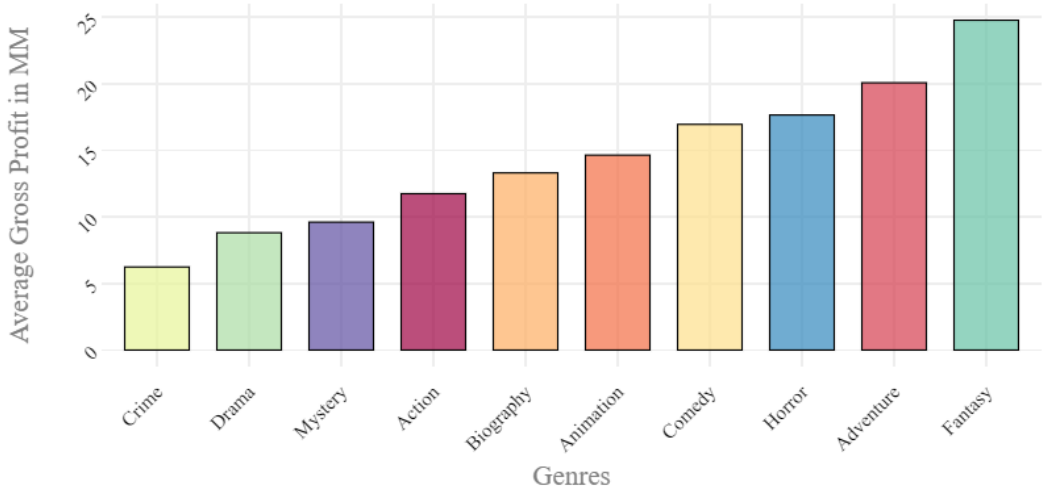
Haonan Duan: Spline method, random forest method, writing introduction and conclusion.

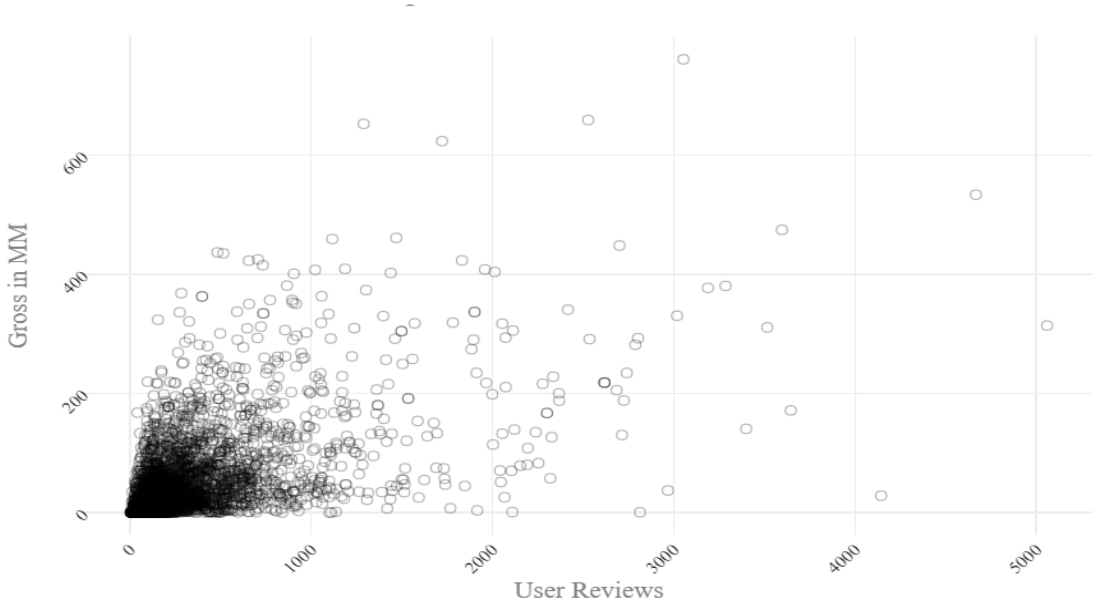
Haidi Shui: Loess method, KNN method.

0.9 APPENDIX

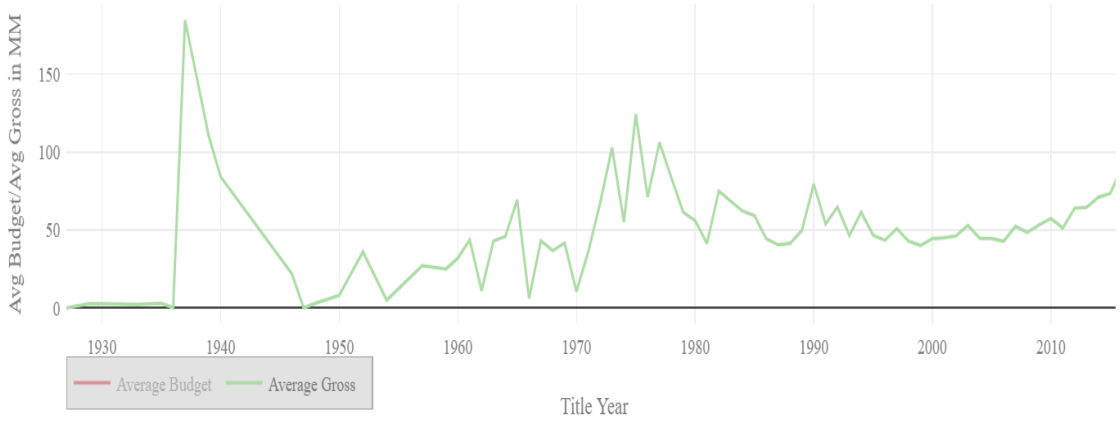
0.9.1 Data and Literature

The dataset is downloaded from a Kaggle project "Understand thenanding Movies through Data". It was originally scraped from www.imdb.com by Kaggle user chuansun76, and then processed by the author of this project, Gautam Joshi. Joshi removed all the duplicates, blank items and NA values which reduced the data to 3739 observations. He also did some rough analysis and data visualization of the movie data. His work helped us identify some major factors that could contribute to the box office revenue of a movie, which proved to be the case in our analysis. The factors are genres, number of users voted and reviews on IMDB, and year. From the following plots (source: https://www.kaggle.com/karrrimba/understanding-movies-through-data), the relationship seems straightforward:





Line Graph for Average Budget vs Average Gross



0.9.2 k nearest neighbor method

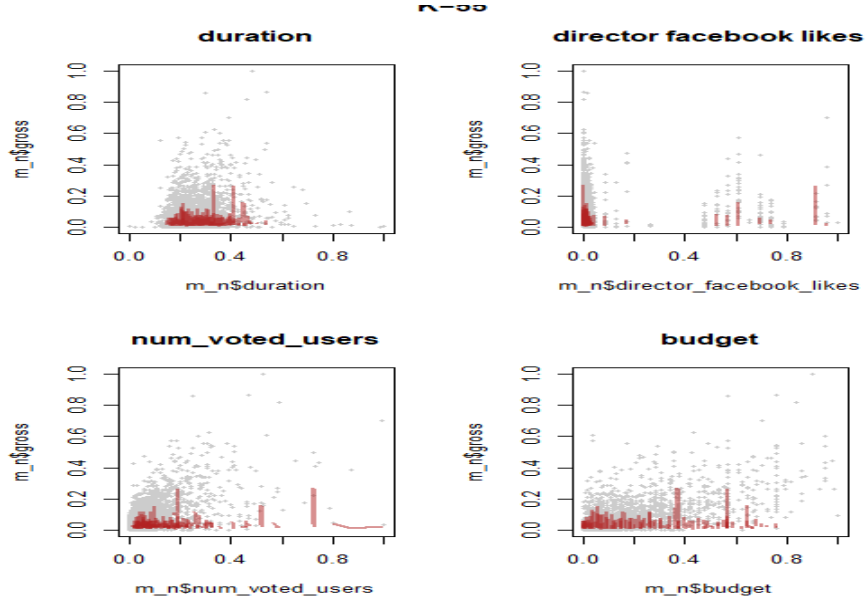
Another model we use is the KNN nearest neighbor.

First, we convert all category variables to numeric variables to calculate the distance. However, if we treat all values numerically, KNN doesn't make sense. For example, the distance between "English" and "other language" is not very interpretable. This is the main reason why we don't include KNN in the main part.

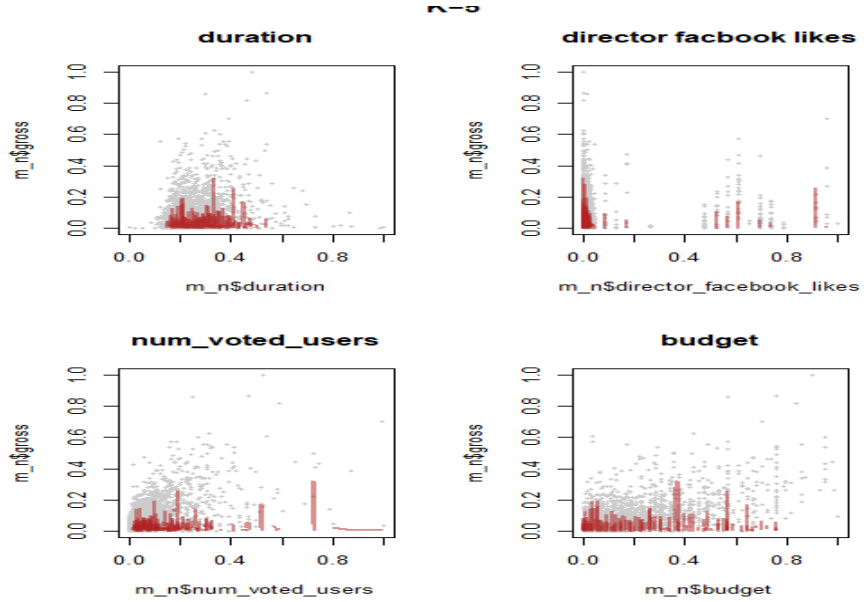
We find out that the ranges of all predictors are wide. In this case, we consider normalizing

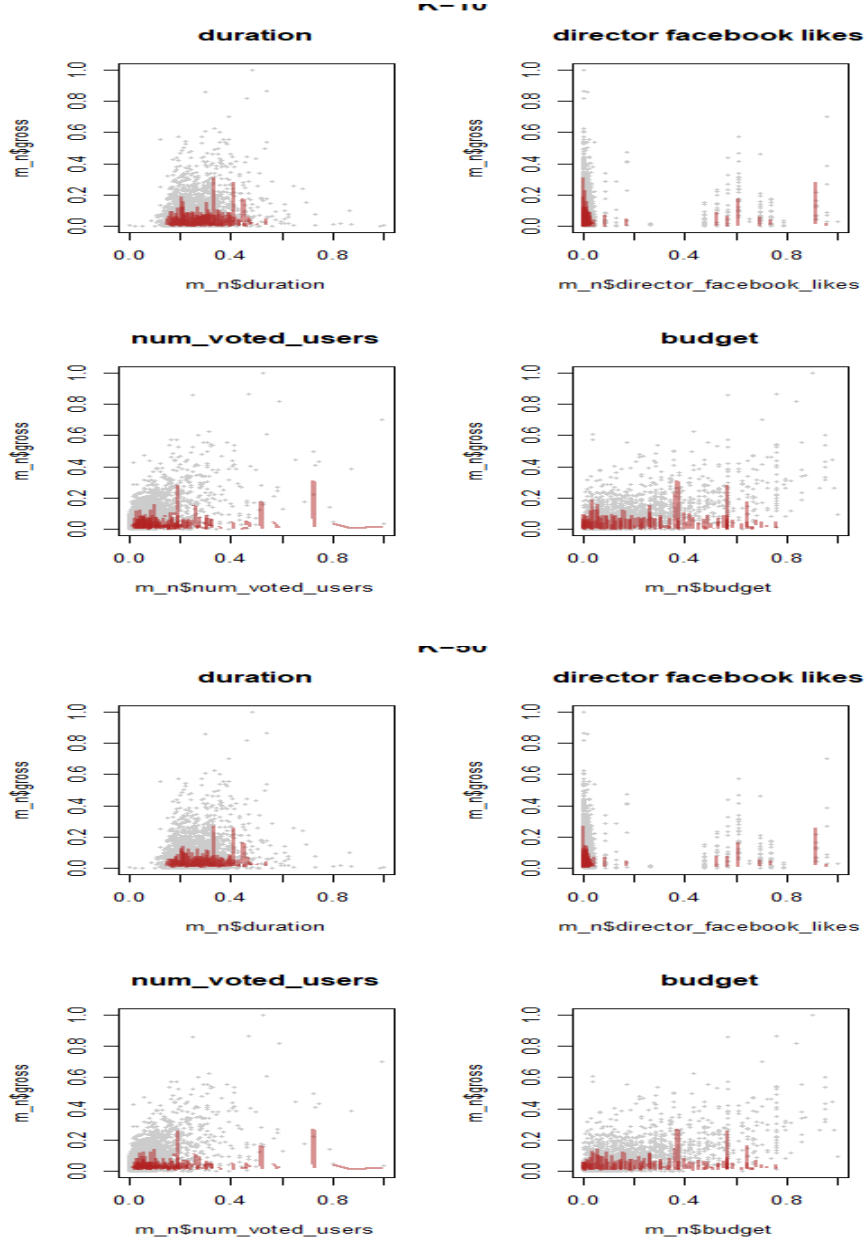
all data. For KNN, we initially take k=55 since it is approximately square root of number of data points.

However, from the graph below we can see that when k=55 the predicted model doesn't fit well.



Hence, we need to take more values of k and find the optimal k. We take k=5, 10, 50, and 100. We visualize the graphs below from the models with different k.





It is clear that when the k increases from 5 to 50 the fitting is going to be better.

However, we still decide not to use KNN because of the ambiguity in categorical variable.