Case Study I - US 2019 Film Releases

Emma Schmidt

STA 610 - Fall 2022

1 Introduction

The purpose of this study is to determine what leads a film to see success in terms of overall net profit. To best capture film success, two separate analyses are run in parallel: one examining net profits, and one examining the log ratio of box office earnings and film budget. This study makes use of data captured on 2019 United States film releases. The data includes an abundance of covariates for each film, specifically: title, release date, production company, three leading cast members, director, box office earnings, budget, run time, critic score, and genre. In addition to defining what makes a successful movie, it is of specific interest to identify the relationships between net income and budget, as well as net income and critic score. The first steps in approaching this problem include pre-processing the data and conducting exploratory data analysis. From there I proceed with selecting covariates and running a variety of model types. Finally, I assess my models, noting key outputs, model fit, and overall takeaways.

2 Data Pre-Processing

The first major issue with the film data is that it contains a high degree of missing values. The missingness occurs in the box office earnings and budget covariates. Specifically, roughly 19.8 percent of the box office earnings and 31.0 percent of the budget values are N/A. This is particularly troubling because both box office and budget are used to calculate a film's net profit. Because net profit is the value of interest in this model it does not make sense to impute this value, so I proceed with a refined data set that removes all films with N/A values. After removing missing values, net profit is calculated for each film in the new data set by subtracting the budget from box office earnings. Finally, net profit is scaled by dividing by 1e5.

3 EDA

Exploratory data analysis is the next critical step in covariate selection and the model refinement process. Two separate analyses are justified by the histograms in Figure 1. Notice that net profit does not exhibit signs of normality. To combat this a second analysis is conducted on the log ratio of box office earnings and budget, as this dependent variable behaves much closer to a normal distribution. For this reason two analyses are carried out with each of these dependent variables.

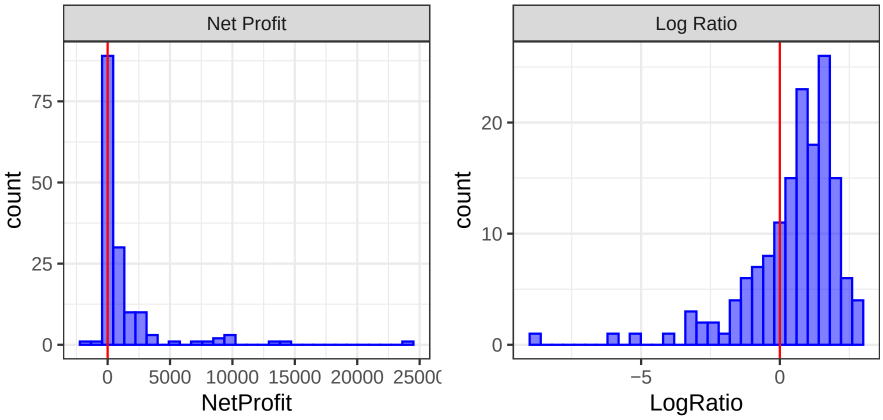


Figure 1: Histogram of the Dependent Variables

The next step in the process is eliminating covariates that will make the model prone to over fitting. These covariates include: title, director, and the three leading cast members. These covariates are not ideal for the model because they are unique for nearly every movie. Any given director or leading cast member occurs in a max of two films, with the majority only occurring in one. Title and cast member covariates are removed from further exploration, but instead of throwing out director completely, a deeper look into this variable reveals that a gender analysis could be of high interest. More specifically, the relationship between net profits for films that are directed by males vs. females. To achieve this I create an additional covariate for director gender by manually entering the data. Found in the appendix is a figure that shows that female directed films perform better than male directed films. Director gender looks like it could be informative as a fixed effect in a future model.

Another interesting covariate to consider is film run time. Instead of approaching this as a quantitative covariate, I categorize films as $<$ 90min, 90min-2h, 2h-2.5h, and $>$ 2.5h. Recall that budget and critic score each in relation to net profit are covariates of high interest in this case study. Figure 2 offers scatter plots for net income plotted against budget and critic score respectively. While analyses of both net income and the log ratio of net income exhibit slightly positive correlation between the dependent variables and both budget and critic score, this can best be seen in the analysis of net profit. For this reason Figure 2 only includes the plots for net profit, but the log ratio plots can be located in the appendix. The correlation with net profit seen in the plots indicates that budget and critic score are promising potential fixed effects in the modeling process. Additionally, it is clear that longer run times tend to yield higher success. Run time as a categorical covariate is a strong candidate for a random effect, as there is variation across groups, and the groups are of varying sizes.

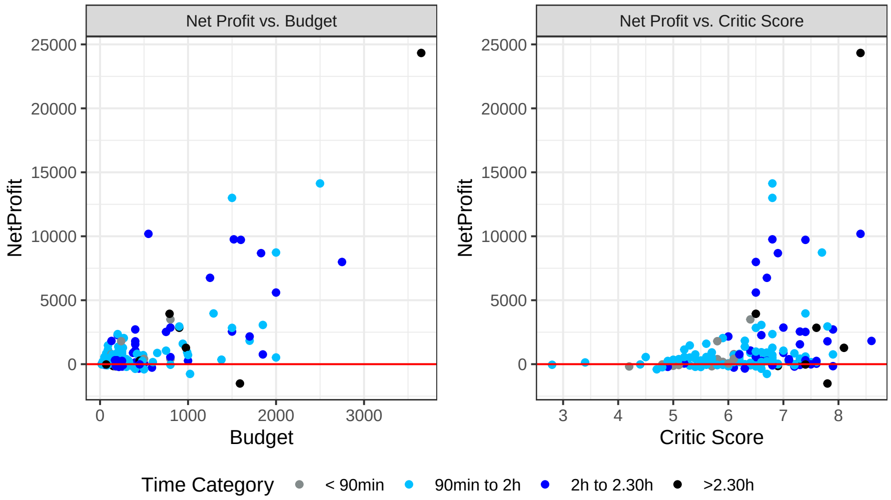


Figure 2: Run Time Scatterplots

For the remaining categorical covariates I proceed with exploring a variety groupings. Figure 3 highlights the grouping of release date by season, with the seasons defined as: Winter Jan.-March, Spring April-June, Summer July-Sept., Fall Oct.-Dec.. While most of the seasons show only moderate variation from one another, summer stands out as particularly interesting. Figure 3 reveals that films appear to be more successful when released in the summer. Like run time, season also possesses many of the qualities of a potential random effect.

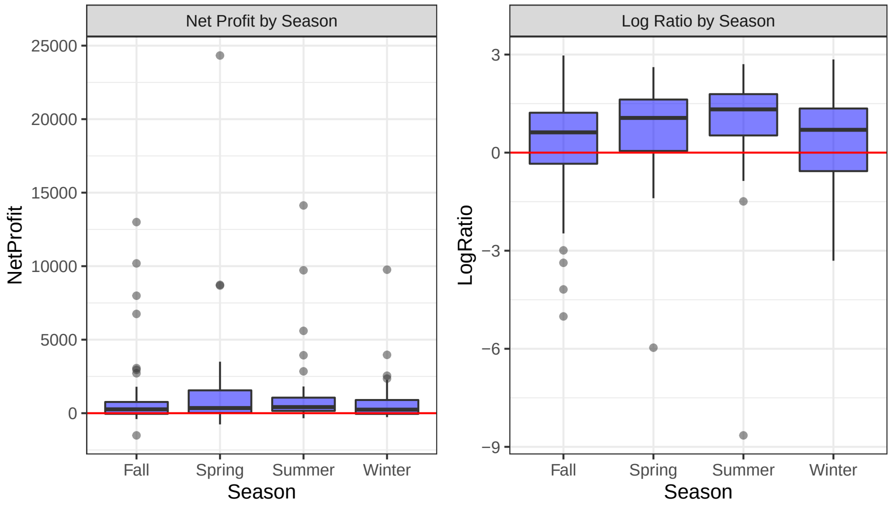


Figure 3: Box plots by Season

The next categorical covariate to hone in on is film genre. This covariate is particularly challenging to deal with because there are 33 distinct genres and many films are categorized with more than one genre. Because many of the genres have a small number of films in the data, I am only using the top five genres as distinct genre groups, and all remaining films are categorized in an ’other’ category. The top 5 genres are comedy, drama, thriller, horror, and action. Each of these categories have a range of 21-43 films. Films that are categorized with multiple genres are double counted to account for the appearance of each genre. For example, 'The Upside' belongs to both the drama and comedy genres, and thus is included in both groupings. Located in the appendix is a figure that includes the box plots for each film genre grouping. Differences across genre groups are fairly minimal, it is unlikely that this covariate will hold significance in the final models.

Lastly, production company is considered as a final potential covariate. Production company is handled much like genre; films are bucketed into one of the top five production companies or into an 'other bucket. Double counting is handled the same as above. The top five production companies are: Universal, Warner Bros, 20th Century Fox, Lionsgate, and Columbia pictures, each ranging between 10-20 total films. Differences in success across production company are present, but not strongly so. It may be worth including the covariate in preliminary models, but will likely not make the final cut. A figure with box plots for each producion company grouping can be found in the appendix.

4 Model Selection

The first preliminary model is a simple linear regression that includes all of the covariates from the EDA section of this case study. This model is using net income as the dependent variable. The model can be specified as:

$$\begin{aligned} y_i =& \mu + \beta_1(Budget_i) + \beta_2(Critical\ Score_i) + \sum_{g\in G}\gamma_g \mathbb{1}(g \in genre_i) + \sum_{p\in P}\gamma_p \mathbb{1}(p \in Production\ Company_i) + \\ & \sum_{s\in S}\gamma_s \mathbb{1}(sSeason_i = s) + \sum_{t\in T}\gamma_t \mathbb{1}(Time_i = t) + \delta_i + \epsilon_i \end{aligned}$$

$$\delta_i = Director\ Gender\ Male, \quad \epsilon_i \sim N(0, \sigma^2)$$

Table 1 can be referenced on page five to view the simple linear regression summary. Notable covariates from the all inclusive linear regression are: budget, critic score, 20th century fox, director gender, summer, and the intercept. Budget, critic score, and summer season have positive coefficients. This means films with higher budgets and or critic scores, or films with summer release dates, are expected to increase the net profit. Conversely the other covariates have negative coefficients and negative correlation to net profit. These covariates could be potentially useful considerations for a mixed effects linear model, but further exploration is necessary. Playing with a number of additional covariate combinations for simple linear regression models, budget, critic score, director gender, season, and run time are most consistently significant. This finding aligns with the initial exploratory data analysis.

Advancing the modeling process, I consider three variations of a mixed effect model. Now layering the assumption of normality on the selected random effects, I pivot to the second analysis where log ratio of net income is the dependent variables. The three models are as follows:

$$\begin{aligned} Model\ 1: y_{ij} &= \mu + \beta_1(Log(Budget_i)) + \beta_2(Critic\ Score_i) + \alpha_j + \epsilon_i \\ \alpha_j &= Run\ Time\ Category, \quad \alpha_j \sim N(0, \tau_1^2), \quad \epsilon_i \sim N(0, \sigma^2) \end{aligned}$$

$$Model\ 2 : y_{ijs} = \mu + \beta_1(Log(Budget_i)) + \beta_2(Critic\ Score_i) + \alpha_j + \gamma_s + \epsilon_i$$

$$\alpha_j = Run\ Time\ Category, \quad \alpha_j \sim N(0, \tau_1^2), \quad \delta_i = Director\ Gender\ Male, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$Model\ 3 : y_{ij} = \mu + \beta_1(Log(Budget_i)) + \beta_2(Critic\ Score_i) + \alpha_j + \gamma_s + \delta_i + \epsilon_i$$

$$\alpha_j = Run\ Time\ Category, \quad \alpha_j \sim N(0, \tau_1^2), \quad \delta_i = Director\ Gender\ Male, \quad \epsilon_i \sim N(0, \sigma^2),$$

$$\gamma_s \sim N(0, \tau_2^2)$$

Table 1:

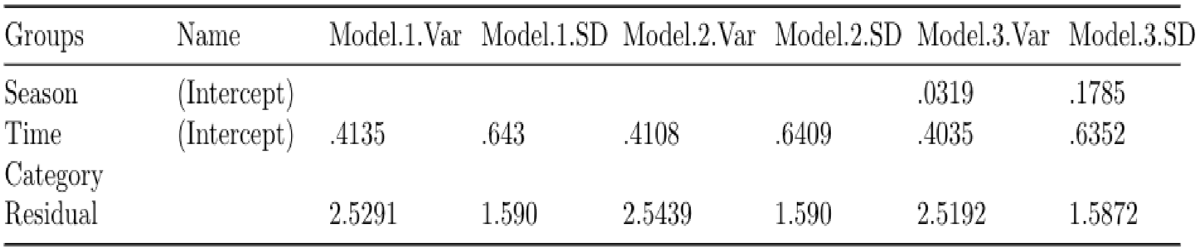
	Dependent variable:	
	NetProfit	
Budget	3.957***	(0.332)
‘Critic Score‘	702.696***	(209.014)
Horror	512.037	(694.107)
Thriller	205.788	(602.079)
Drama	−650.784	(636.981)
Action	−1,160.979*	(643.601)
Comedy	−173.367	(544.245)
‘Other Genre‘	−551.703	(728.350)
‘20th Century Fox‘	−2,195.708***	(711.997)
Columbia	406.936	(675.163)
Universal	103.396	(553.365)
‘Warner Bros.‘	−377.363	(505.816)
Lionsgate	−8.173	(586.857)
‘Director Gender‘Male	−1,002.956**	(469.506)
‘Time Category‘90min to 2h	132.127	(649.882)
‘Time Category‘2h to 2.30h	116.441	(765.129)
‘Time Category‘>2.30h	−129.529	(1,031.784)
SeasonSpring	480.622	(431.912)
SeasonSummer	804.097*	(472.014)
SeasonWinter	509.527	(462.736)
Intercept	−4,250.493***	(1,414.408)
Observations	155	
R^2	0.644	
Adjusted R^2	0.591	
Residual Std. Error	1,968.755 (df = 134)	
F Statistic	12.106*** (df = 20; 134)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2 located below, illustrates the summary output data from the three specified mixed effects model. As expected critic score and log(budget) are significant in all three models. Interestingly, director gender is not significant in either case, suggesting that maybe the relationship between film success of male and female directed movies is not that significant. Finally, the random effects placed on run time and season are significant. The coefficients for critic score, budget, director gender, and intercept are consistent with the findings from the simple linear regression model. That is, critic score and log(budget) have positive coefficients and director gender and the intercept have negative coefficients. Figure 4, also below, offers the variance and standard deviation for the random effects in each model.

Table 2:

	Dependent variable:		
	LogRatio		
	(1)	(2)	(3)
‘Critic Score‘	0.461***	0.466***	0.470***
	(0.150)	(0.151)	(0.150)
log(Budget)	0.221*	0.225*	0.231**
	(0.115)	(0.117)	(0.116)
‘Director Gender‘Male		−0.136	−0.138
		(0.355)	(0.355)
Intercept	−4.109***	−4.050***	−4.088***
	(1.179)	(1.190)	(1.188)
Observations	155	155	155
Log Likelihood	−296.023	−296.066	−295.905
Akaike Inf. Crit.	602.045	604.131	605.811
Bayesian Inf. Crit.	617.262	622.392	627.115

Note: *p<0.1; **p<0.05; ***p<0.01



Groups	Name	Model.1.Var	Model.1.SD	Model.2.Var	Model.2.SD	Model.3.Var	Model.3.SD
Season	(Intercept)					.0319	.1785
Time	(Intercept)	.4135	.643	.4108	.6409	.4035	.6352
Category							
Residual		2.5291	1.590	2.5439	1.590	2.5192	1.5872

Figure 4: Random Effects

5 Model Fit

Using mixed effects Model Three as the final model, I now explore model fit. Figure 5 and Figure 6 shown below, examine the adequacy of modeling assumptions. Figure 5 does not raise any red flags, as the residuals do not appear to demonstrate any type of pattern. Figure 6 is however a bit concerning. Even after log transforming the data a heavy tail still looks to be present. This heavy tail indicates that this mixed effects model may not be the best fit. However, the mixed effects model does do substantially better in meeting assumptions than the original simple linear regression. Fitted vs. residual and normal qq-plots for the simple linear regression model can be found in the appendix.

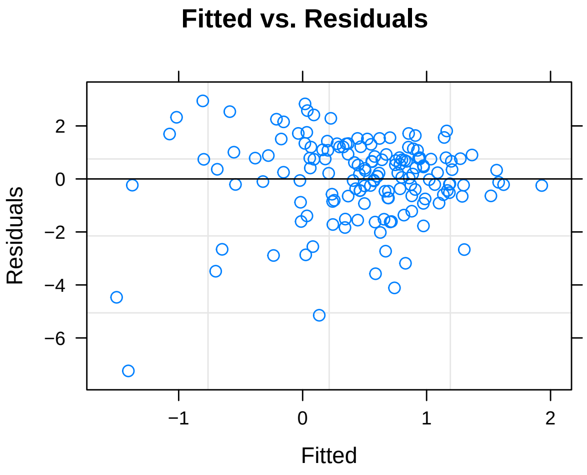


Figure 5: Fitted vs. Residuals

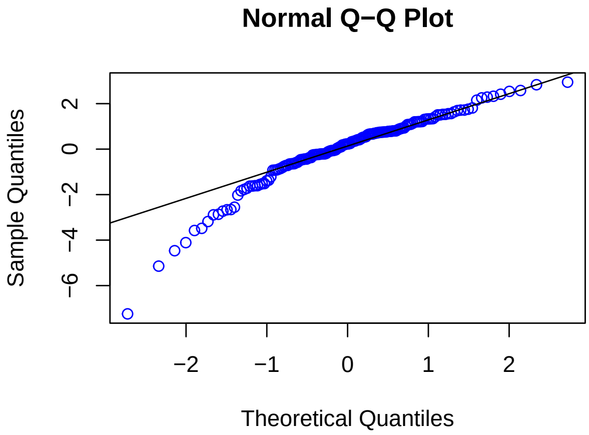


Figure 6: QQ Plot

6 Conclusion

The final model indicates that budget and critic score are important to determining a film's success. In fact, these covariates were significant in every stage of the modeling process. In short the bigger the budget and higher the critic score, the better chance a film has to be profitable. Additionally, the season a film is released during can also be a potential explanation for a film's success of lack there of. Films that were released in the summer season outperformed films from other seasons. This could be because people may have more time to attend the theatre in the summer, or perhaps the most highly anticipated movies are released in the summer. Finally, run time was also a significant factor to a film's net profit. The EDA and model indicate that movies that run over 2 hours tend to make more money than those that don't. Overall, the models were successful in helping explain what makes profitable film.

Appendix

Emma Schmidt

2022-10-06

```r
knitr::opts_chunk$set(echo = FALSE)
library(stringr)
library(ggplot2)
library(tidyr)
library(lubridate)
library(dplyr)
library(data.table)
library(lme4)
library(ggpubr)
library(stargazer)
# Read in the Data
films <- read.csv("~/Desktop/Duke/610/United_States_Film_Releases_2019.csv")
# Data Preprocessing
f_miss <- films[!(films$Box.Office == "N/A" | films$Budget == "N/A"), ]

count(films[films$Budget == "N/A", ])

f_miss$Box.Office <- as.numeric(gsub(",", "", gsub("$", "", f_miss$Box.Office,

                                        fixed = TRUE)))/1e5

f_miss$Budget <- as.numeric(gsub(",", "", gsub("$", "", f_miss$Budget,

                                        fixed = TRUE)))/1e5

unique_genre <- (unlist(str_split(f_miss$Genre, "/")))

tot_genre <- table(gsub(" ", "", unique_genre, fixed = TRUE))

unique_prod <- (unlist(str_split(f_miss$Production.Company, "/")))

tot_prod <- table(gsub(" ", "", unique_prod, fixed = TRUE))

max_dir <- max(table(f_miss['Director']))

max_1 <- max(table(f_miss['Lead.Cast.1']))
```

```r
max_2 <- max(table(f_miss['Lead.Cast.2']))

max_3 <- max(table(f_miss['Lead.Cast.3']))

f_miss <- f_miss %>%
  mutate(NetProfit = Box.Office - Budget,
         LogBudget = log(Budget),
         LogBoxOffice = log(Box.Office),
         LogRatio = log(Box.Office) - log(Budget),
         Month = month(mdy(f_miss$Release.Date..mmddyyyy)),
         Horror = grepl("Horror", Genre),
         Thriller = grepl("Thriller", Genre),
         Drama = grepl("Drama", Genre),
         Action = grepl("Action", Genre),
         Comedy = grepl("Comedy", Genre),
         `Other Genre` = !grepl("Horror|Thriller|Drama|Comedy|Action", Genre, fixed = FALSE),
         `20th Century Fox` = grepl("20th Century Fox", Production.Company),
         `Columbia` = grepl("Columbia Pictures", Production.Company),
         Lionsgate = grepl("Lionsgate", Production.Company),
         `Universal` = grepl("Universal Pictures", Production.Company),
         `Warner Bros.` = grepl("Warner Bros. Pictures", Production.Company),
         `Other Production` = !grepl("20th Century Fox|Columbia Pictures|Lionsgate|Universal Pi
         `Time Category` = case_when(Run.Time..minutes. < 90 ~ "< 90min",
                                     Run.Time..minutes. < 120 & Run.Time..minutes. >= 90 ~ "90mi
                                     Run.Time..minutes. < 150 & Run.Time..minutes. >= 120 ~ "2h t
                                     Run.Time..minutes. >= 150 ~ ">2.30h"),
         `Time Category` = factor(`Time Category`,
                                  levels = c("< 90min", "90min to 2h", "2h to 2.30h", ">2.30h")),
         Season = case_when(Month = 1 & Month <=3 ~ "Winter",
                            Month >= 4 & Month <=6 ~ "Spring",
                            Month >= 7 & Month <=9 ~ "Summer",
                            Month >= 10 & Month <=12 ~ "Fall"))



f_miss$Month[f_miss$Month == '5/'] = '05'

f_miss$`Director Gender` <- "Male"
ID_female <- c(8, 19, 23, 36, 38, 40,
               41, 45, 48, 52, 59, 65, 79,
               84, 90, 100, 105, 124, 131,
               133, 136, 137,140,145, 153)
f_miss$`Director Gender`[ID_female] <- "Female"
f_miss$`Director Gender` <- as.factor(f_miss$`Director Gender`)


names(f_miss)[names(f_miss) == 'Critic.Score..IMDB.x.10.'] <- 'Critic Score'
```

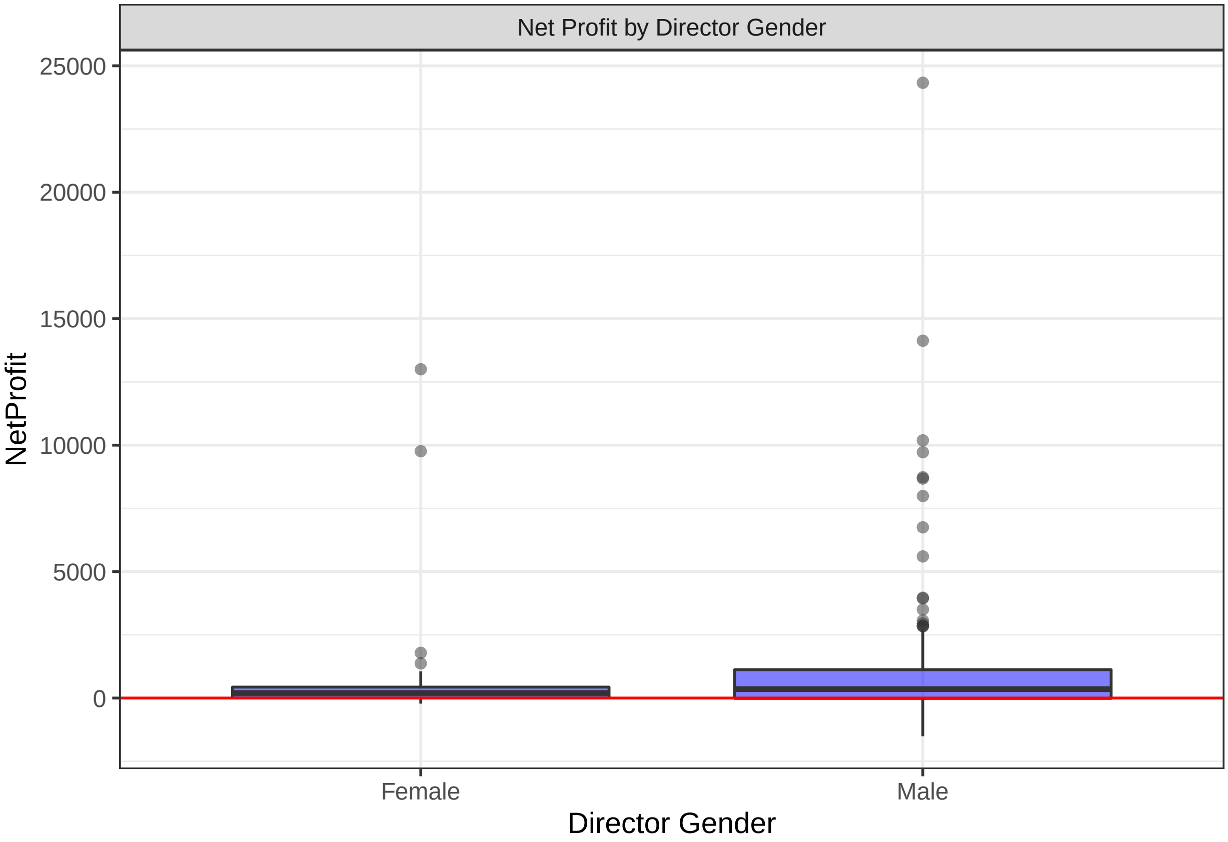
```r
# Director Gender Plots
d <- ggplot(data = f_miss, aes(x=`Director Gender`, y =NetProfit)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Net Profit by Director Gender")

logd <- ggplot(data = f_miss, aes(x=`Director Gender`, y =LogRatio)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Log Ratio by Director Gender")

d

logd
# Season Plots
season <- ggplot(data = f_miss, aes(x=Season, y =NetProfit)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Net Profit by Season")

logseason <- ggplot(data = f_miss, aes(x=Season, y =LogRatio)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Log Ratio by Season")

season

logseason
# Time Category Plots

t <- ggplot(data = f_miss, aes(x=`Time Category`, y =NetProfit)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Net Profit by Time Category")

logt <- ggplot(data = f_miss, aes(x=`Time Category`, y =LogRatio)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Log Ratio by Time Category")

t
```

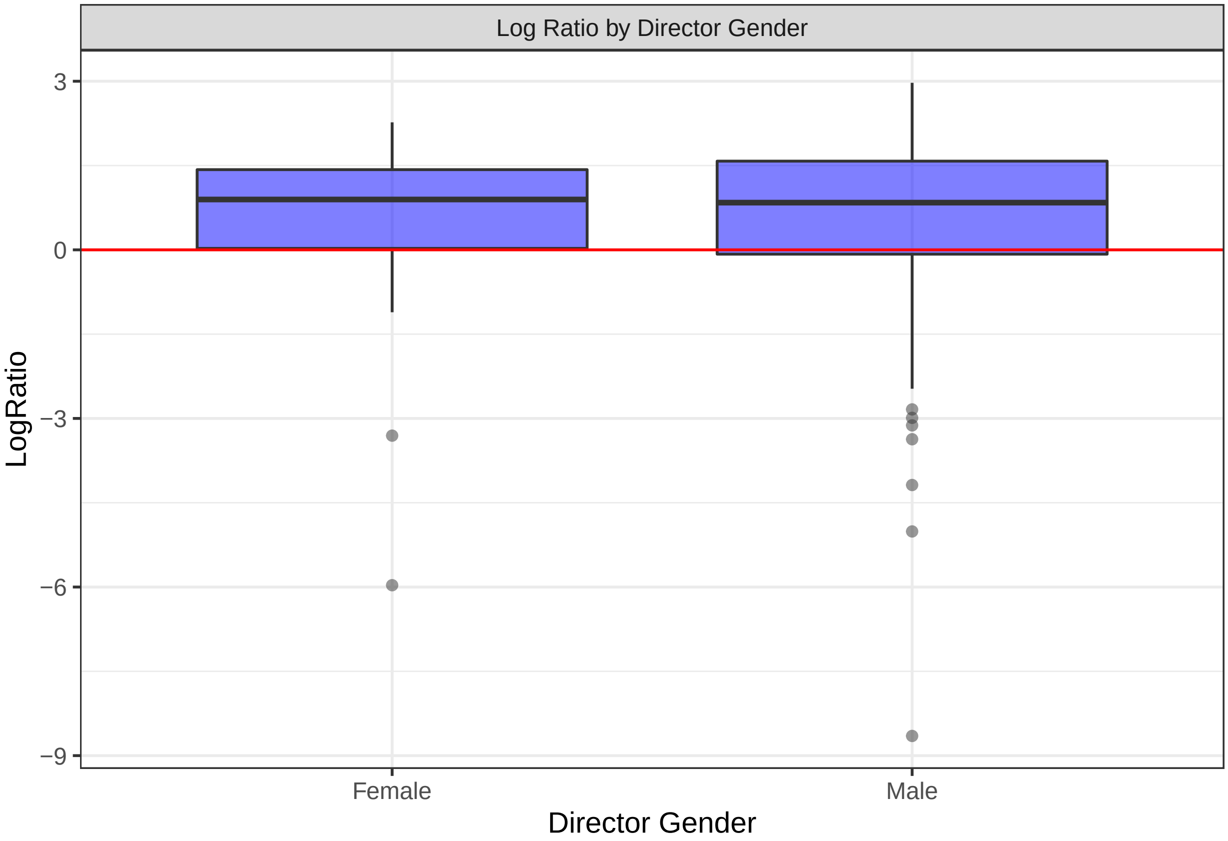
```
logt
# Run time ungrouped plot

rt <- ggplot(data = f_miss, aes(x= Run.Time..minutes., y = NetProfit)) +
  geom_point(colour = "blue", alpha = .5) +
  geom_hline(yintercept = 0, colour = "red") +
  theme_bw() +
  facet_wrap(~"Run Time vs. Net Profit")

rt
# Budget Plot
b <- ggplot(data = f_miss, aes(x= Budget, y = NetProfit)) +
  geom_point(colour = "blue", alpha = .5) +
  geom_hline(yintercept = 0, colour = "red") +
  theme_bw() +
  facet_wrap(~"Budget vs. Net Profit")

b
# Critic Score Plot
cs <- ggplot(data = f_miss, aes(x= `Critic Score`, y = NetProfit)) +
        geom_point(colour = "blue", alpha = .5) +
        geom_hline(yintercept = 0, colour = "red") +
        theme_bw() +
        facet_wrap(~"Critic Score vs. Net Profit")

cs
# Month Plots
m <- ggplot(data = f_miss, aes(x=Month, y =NetProfit)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Net Profit by Month")

logm <- ggplot(data = f_miss, aes(x=Month, y =LogRatio)) +
      geom_boxplot(fill = "blue", alpha = 0.5) +
      geom_hline(yintercept = 0, colour = "red") +
      theme_bw() +
      facet_wrap(~"Log Ratio by Month")

m

logm
# Genre Plots
rbind(f_miss %>% filter(Horror == TRUE) %>% mutate(Type = "Horror"),
      f_miss %>% filter(Thriller == TRUE) %>% mutate(Type = "Thriller"),
      f_miss %>% filter(Drama == TRUE) %>% mutate(Type = "Drama"),
      f_miss %>% filter(Action == TRUE) %>% mutate(Type = "Action"),
      f_miss %>% filter(Comedy == TRUE) %>% mutate(Type = "Comedy"),
      f_miss %>% filter(`Other Genre` == TRUE) %>% mutate(Type = "Other")) %>%
      ggplot(aes(x=Type, y =NetProfit)) +
        geom_boxplot(fill = "blue", alpha = 0.5) +
        geom_hline(yintercept = 0, colour = "red") +
        theme_bw() +
        facet_wrap(~"Net Profit by Genre Type")

rbind(f_miss %>% filter(Horror == TRUE) %>% mutate(Type = "Horror"),
      f_miss %>% filter(Thriller == TRUE) %>% mutate(Type = "Thriller"),
      f_miss %>% filter(Drama == TRUE) %>% mutate(Type = "Drama"),
      f_miss %>% filter(Action == TRUE) %>% mutate(Type = "Action"),
      f_miss %>% filter(Comedy == TRUE) %>% mutate(Type = "Comedy"),
      f_miss %>% filter(`Other Genre` == TRUE) %>% mutate(Type = "Other")) %>%
      ggplot(aes(x=Type, y =LogRatio)) +
        geom_boxplot(fill = "blue", alpha = 0.5) +
        geom_hline(yintercept = 0, colour = "red") +
        theme_bw() +
        facet_wrap(~"Log Ratio by Genre Type")

# Histogram and Production Company Plots
phist1 <- ggplot(data=f_miss, aes(x=NetProfit)) +
  geom_histogram(fill = "blue",colour = "blue", alpha = .5, bins = 30) +
  theme_bw() +
  geom_vline(xintercept = 0, colour = "red") +
  facet_wrap(~"Net Profit")

phist2 <- ggplot(data=f_miss, aes(x=LogRatio)) +
  geom_histogram(fill = "blue", , colour = "blue", alpha = .5, bins = 30) +
  theme_bw() +
  geom_vline(xintercept = 0, colour = "red") +
  facet_wrap(~"Log Ratio")

ggarrange(phist1, phist2)

rbind(f_miss %>% filter(`20th Century Fox` == TRUE) %>% mutate(Company = "20th Century Fox"),
      f_miss %>% filter(`Columbia` == TRUE) %>% mutate(Company = "Columbia"),
      f_miss %>% filter(Lionsgate == TRUE) %>% mutate(Company = "Lionsgate"),
      f_miss %>% filter(`Universal` == TRUE) %>% mutate(Company = "Universal"),
      f_miss %>% filter(`Warner Bros.` == TRUE) %>% mutate(Company = "Warner Bros."),
      f_miss %>% filter(`Other Production` == TRUE) %>% mutate(Company = "Other")) %>%
      ggplot(aes(x=Company, y =NetProfit)) +
        geom_boxplot(fill = "blue", alpha = 0.5) +
        geom_hline(yintercept = 0, colour = "red") +
        theme_bw() +
        facet_wrap(~"Net Profit by Production Company")
```

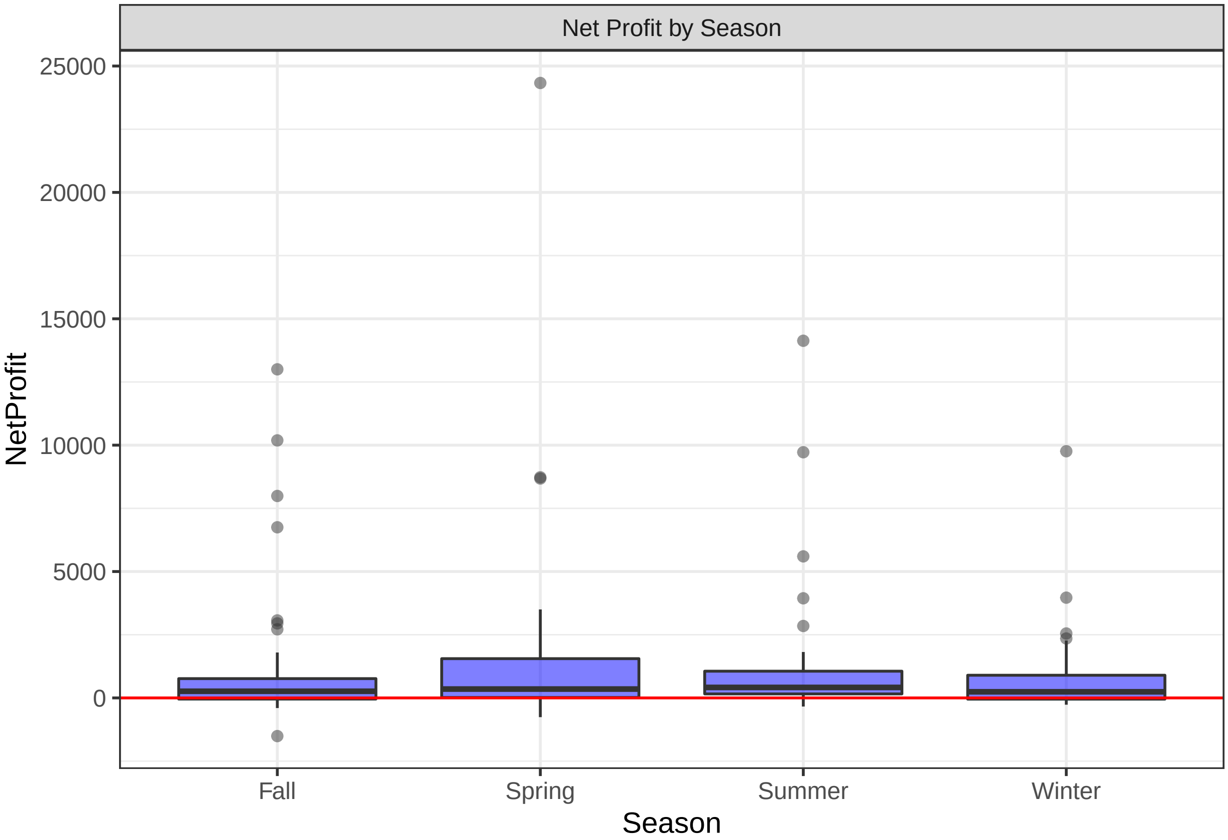
```r
rbind(f_miss %>% filter(`20th Century Fox` == TRUE) %>% mutate(Company = "20th Century Fox"),
      f_miss %>% filter(`Columbia` == TRUE) %>% mutate(Company = "Columbia"),
      f_miss %>% filter(Lionsgate == TRUE) %>% mutate(Company = "Lionsgate"),
      f_miss %>% filter(`Universal` == TRUE) %>% mutate(Company = "Universal"),
      f_miss %>% filter(`Warner Bros.` == TRUE) %>% mutate(Company = "Warner Bros."),
      f_miss %>% filter(`Other Production` == TRUE) %>% mutate(Company = "Other")) %>%
      ggplot(aes(x=Company, y =LogRatio)) +
        geom_boxplot(fill = "blue", alpha = 0.5) +
        geom_hline(yintercept = 0, colour = "red") +
        theme_bw() +
        facet_wrap(~"Log Ratio by Production Company")
# Scatterplots Budget and Critic Score Colored by Director Gender and Run Time
ggplot(data = f_miss) +
  geom_point(aes(x = `Critic Score` , y = NetProfit, colour = `Director Gender`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Net Profit vs. Critic Score") +
  scale_color_manual(values = c("black", "blue"))

ggplot(data = f_miss) +
  geom_point(aes(x = `Critic Score` , y = LogRatio, colour = `Director Gender`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Log Ratio vs. Critic Score") +
  scale_color_manual(values = c("black", "blue"))

ggplot(data = f_miss) +
  geom_point(aes(x = Budget, y = NetProfit, colour = `Director Gender`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Net Profit vs. Budget") +
  scale_color_manual(values = c("black", "blue"))

ggplot(data = f_miss) +
  geom_point(aes(x = LogBudget , y = LogRatio, colour = `Director Gender`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Log Ratio vs. Log Budget") +
  scale_color_manual(values = c("black", "blue"))

p2 <- ggplot(data = f_miss) +
  geom_point(aes(x = `Critic Score` , y = NetProfit, colour = `Time Category`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Net Profit vs. Critic Score") +
  scale_color_manual(values = c("azure4", "deepskyblue", "blue", "black"))
```

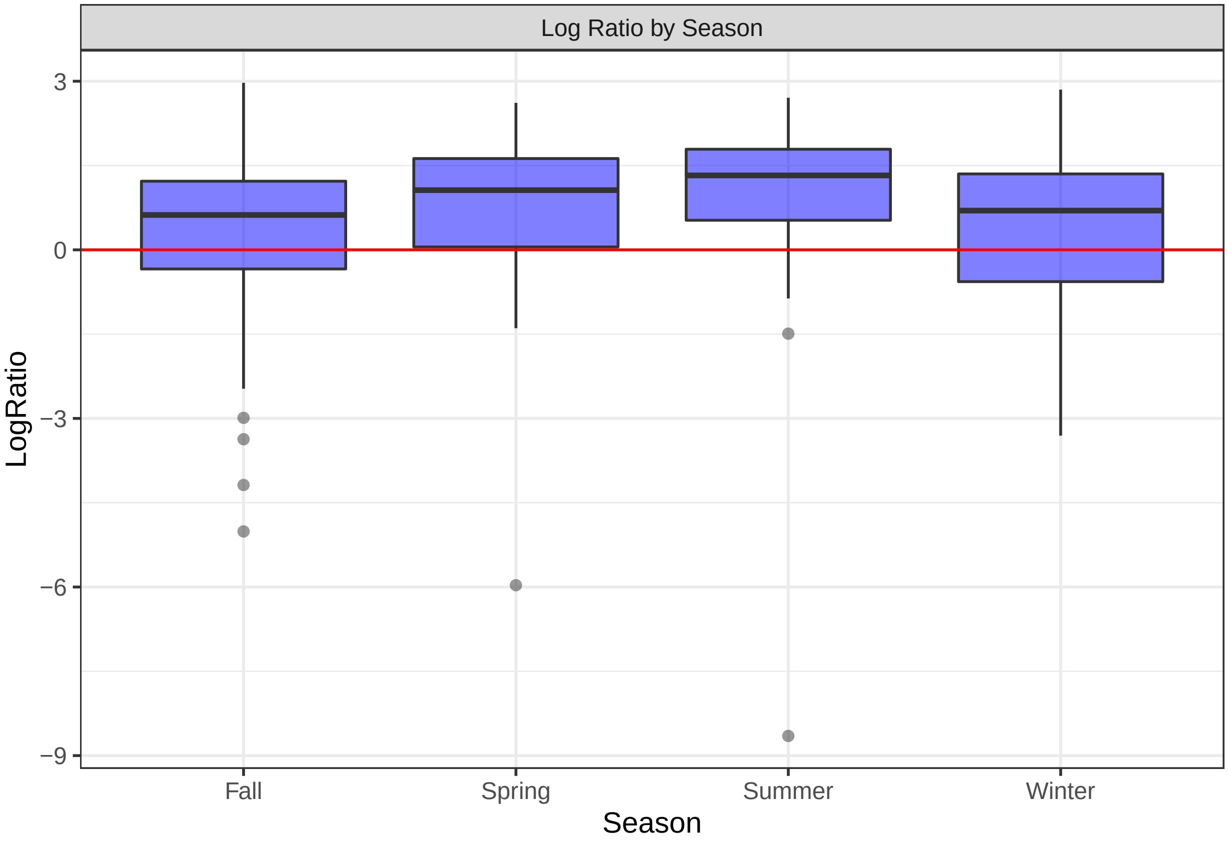
```
p4<-ggplot(data = f_miss) +
  geom_point(aes(x = `Critic Score` , y = LogRatio, colour = `Time Category`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Log Ratio vs. Critic Score") +
  scale_color_manual(values = c("black", "blue", "azure4", "deepskyblue"))

p1 <- ggplot(data = f_miss) +
  geom_point(aes(x = Budget, y = NetProfit, colour = `Time Category`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Net Profit vs. Budget") +
  scale_color_manual(values = c("azure4", "deepskyblue", "blue", "black"))

p3 <-ggplot(data = f_miss) +
  geom_point(aes(x = LogBudget , y = LogRatio, colour = `Time Category`)) +
  geom_hline(yintercept = 0, col = "red") +
  theme_bw() +
  facet_wrap(~"Log Ratio vs. Log Budget") +
  scale_color_manual(values = c("black", "blue", "azure4", "deepskyblue"))
# GG Arrange
ggarrange(p1, p2, p3, p4, legend = "bottom", common.legend = T, nrow = 2, ncol = 2)
# Modeling
data_reg <- f_miss %>%
  select(NetProfit,
         Budget, `Critic Score`, Horror, Thriller, Drama, Action, Comedy, `Other Genre`,
         `20th Century Fox`, Columbia, Universal, `Warner Bros.`, Lionsgate,
         `Director Gender`, `Time Category`, Season)

lreg <- lm(NetProfit ~ ., data= data_reg)

summary(lreg)

data_reg2 <- f_miss %>%
  select(LogRatio,
         LogBudget, `Critic Score`, `Director Gender`, `Time Category`, Season)

lfit <- lmer(LogRatio ~ `Critic Score` + LogBudget + (1|`Time Category`), data = data_reg2)
lfit2 <- lmer(LogRatio ~ `Critic Score` + LogBudget + `Director Gender` +  (1|`Time Category`)
              data = data_reg2)
lfit3 <- lmer(LogRatio ~ `Critic Score` + LogBudget + `Director Gender` + (1|Season) + (1|`Tim

summary(lfit)
summary(lfit2)
summary(lfit3)
```

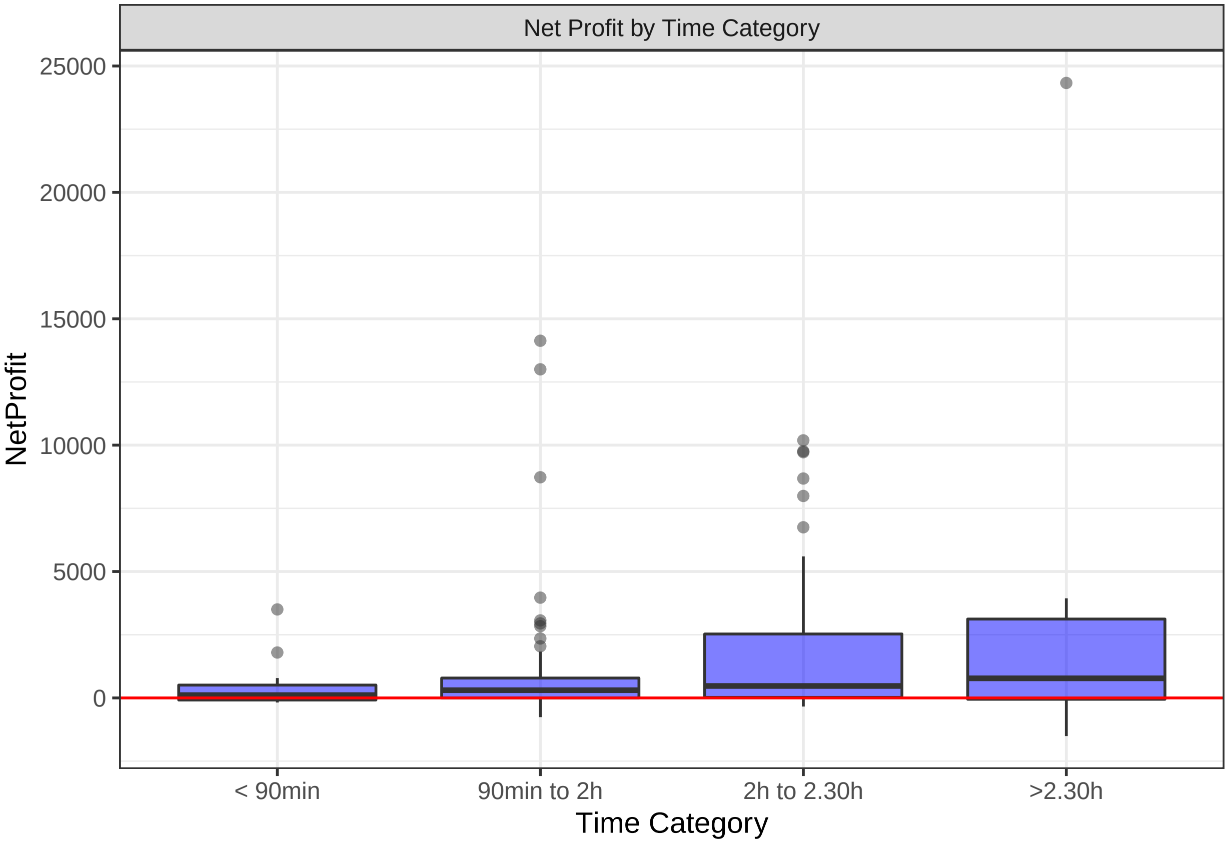
```
random <- data.frame(Groups = c("Season", "Time Category", "Residual"),
                     Name = c("(Intercept)", "(Intercept)", ""),
                     `Model 1 Var` = c("", ".4135", "2.5291"),
                     `Model 1 SD` = c("",".643", "1.590"),
                     `Model 2 Var` = c("",".4108", "2.5439"),
                     `Model 2 SD` = c("",".6409", "1.590"),
                     `Model 3 Var` = c(".0319",".4035", "2.5192"),
                     `Model 3 SD` = c(".1785",".6352", "1.5872"))

knitr::kable(random)
# Model Fit
plot(lfit3, xlab = "Fitted", ylab = "Residuals", main = "Fitted vs. Residuals")
qqnorm(residuals(lfit3), col = "blue")
qqline(residuals(lfit3))

plot(lreg)
```

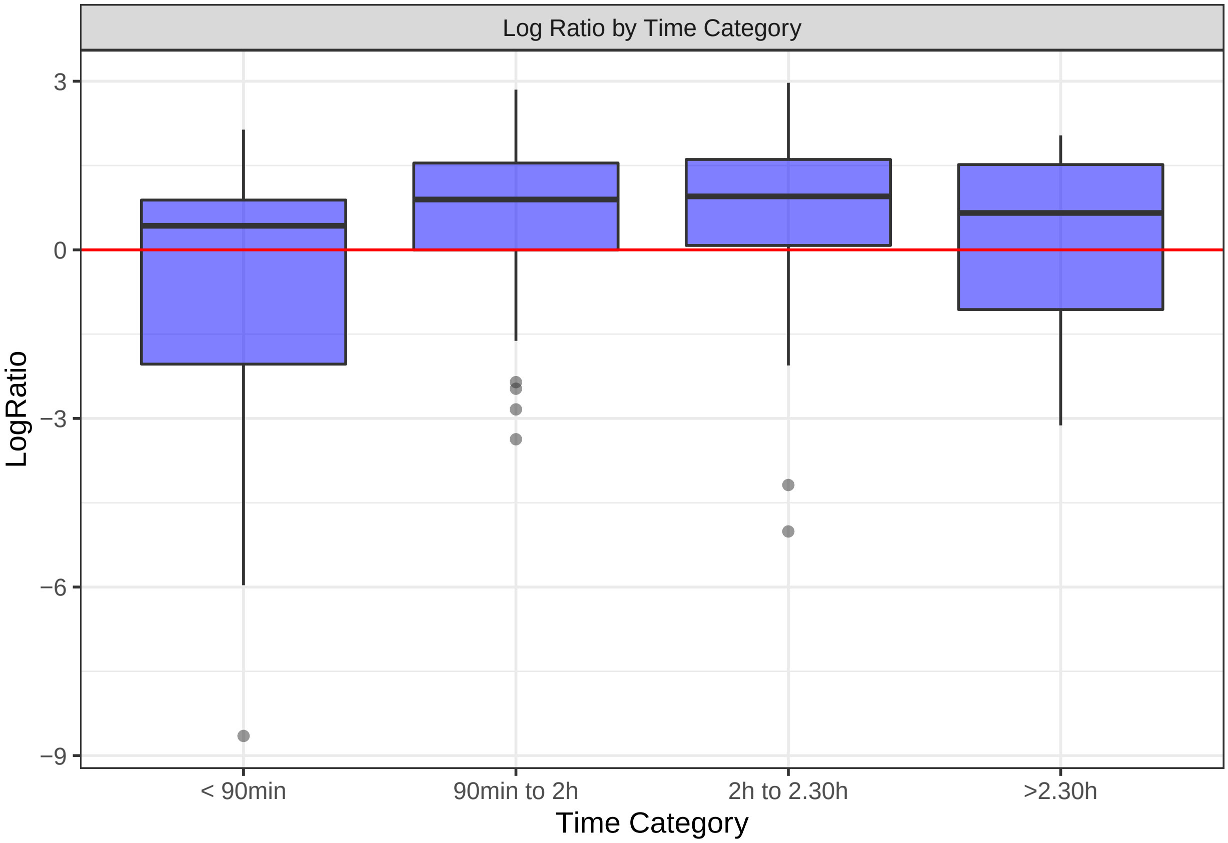


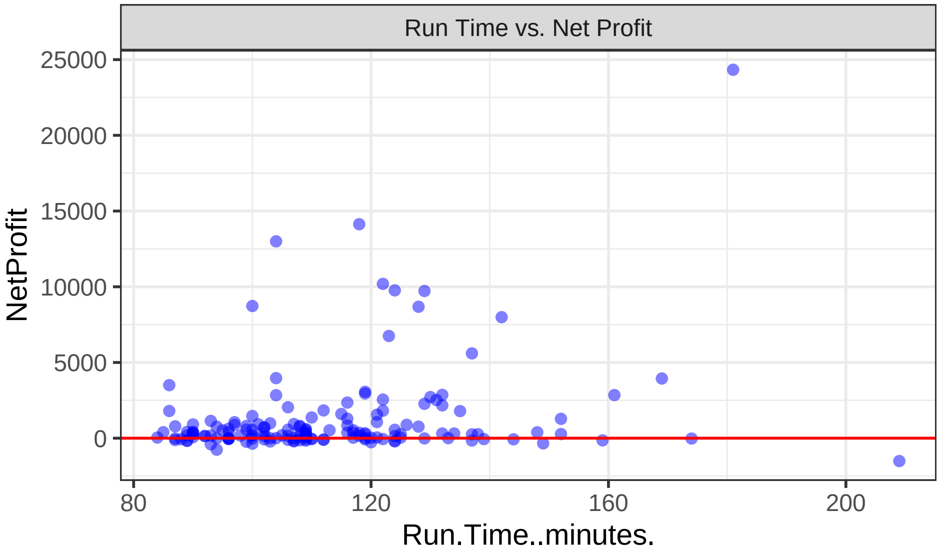

Log Ratio by Director Gender
3
0
−3
−6
−9
LogRatio
Female
Male
Director Gender

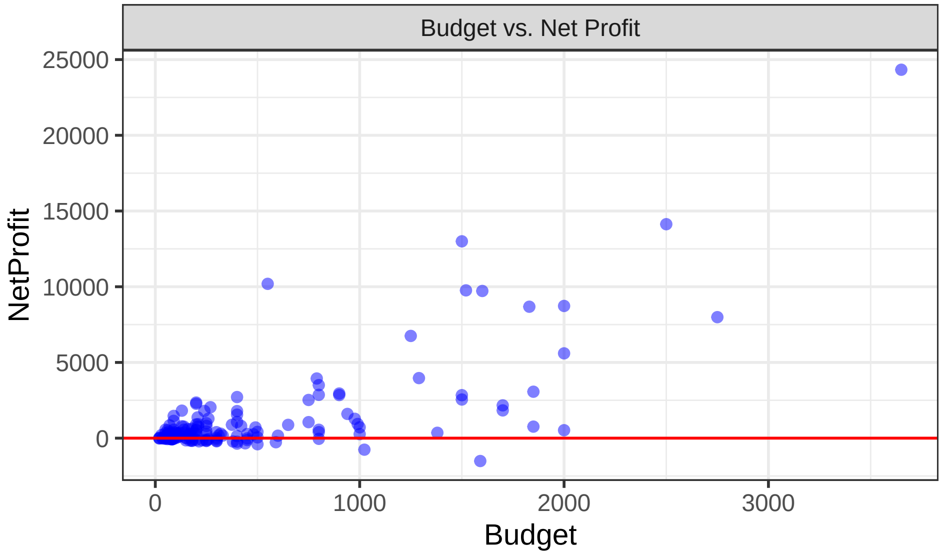

Net Profit by Season
25000
20000
15000
10000
5000
0
NetProfit
Fall
Spring
Summer
Winter
Season

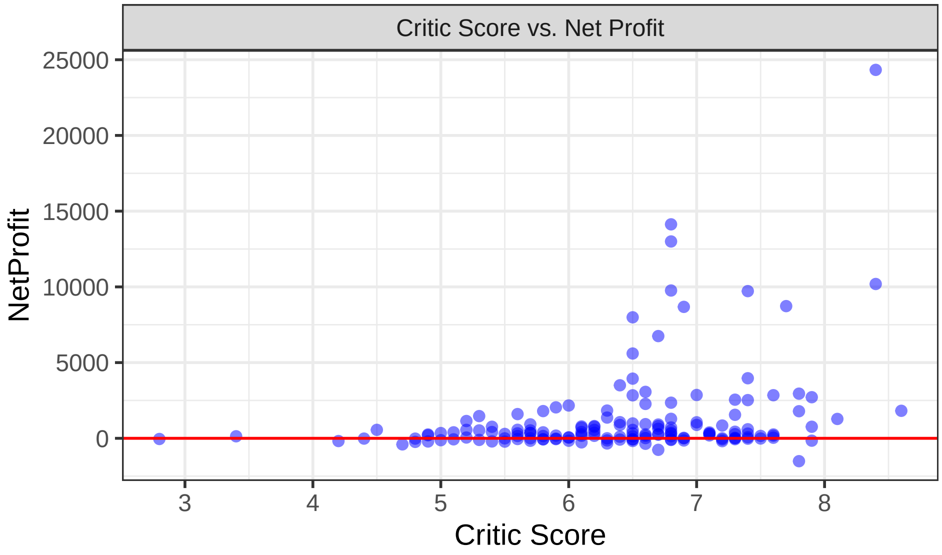

Log Ratio by Season
LogRatio
3
0
−3
−6
−9
Fall
Spring
Summer
Winter
Season

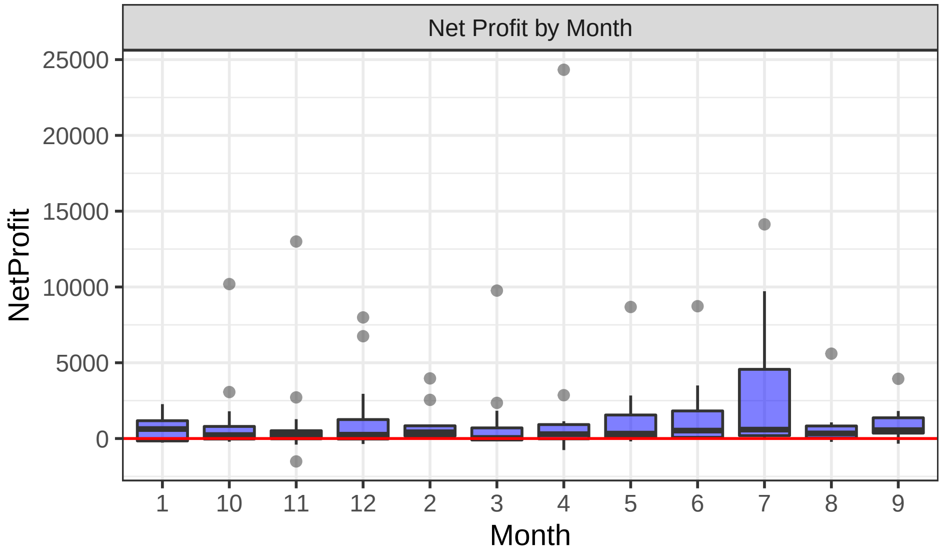

Net Profit by Time Category
NetProfit
25000
20000
15000
10000
5000
0
< 90min
90min to 2h
2h to 2.30h
>2.30h
Time Category

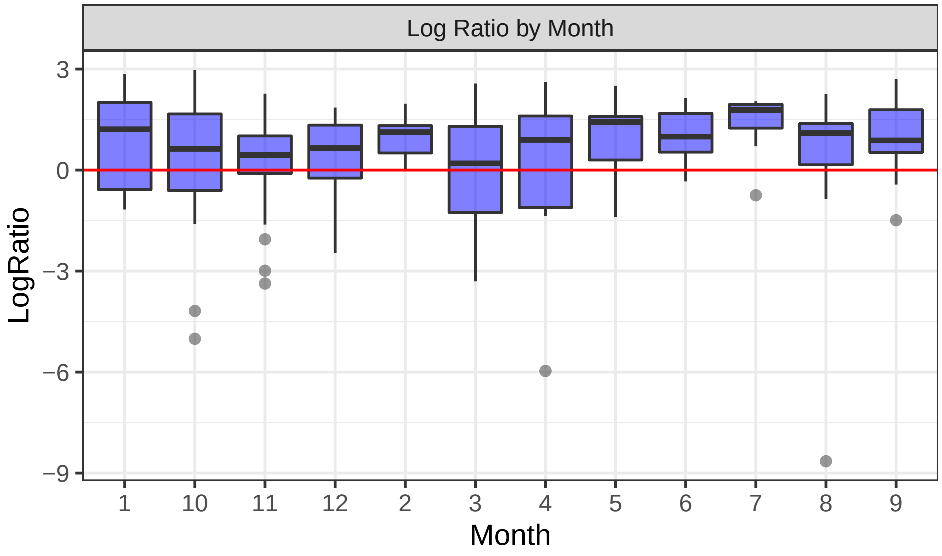

Log Ratio by Time Category
3
0
−3
−6
−9
LogRatio
< 90min
90min to 2h
2h to 2.30h
>2.30h
Time Category

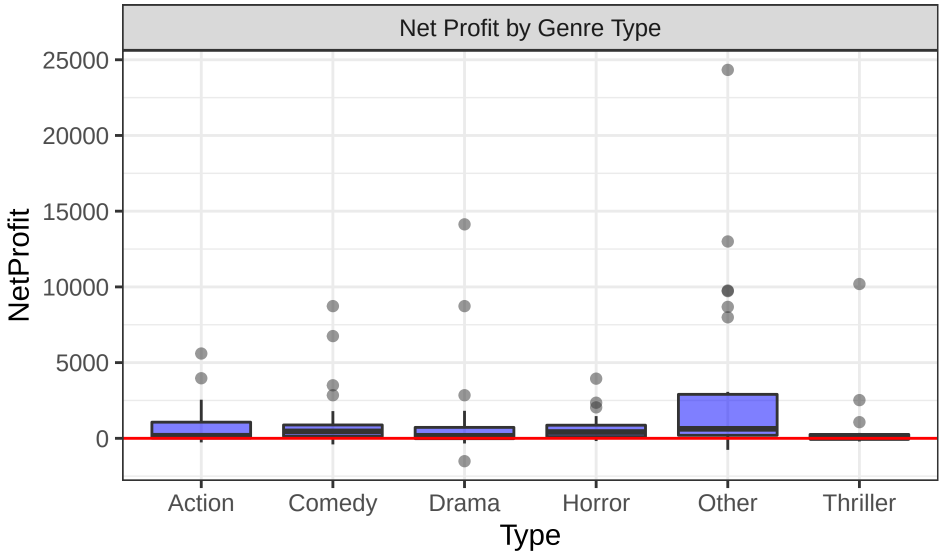

Run Time vs. Net Profit
25000
20000
15000
10000
5000
0
NetProfit
80
120
160
200
Run.Time..minutes.

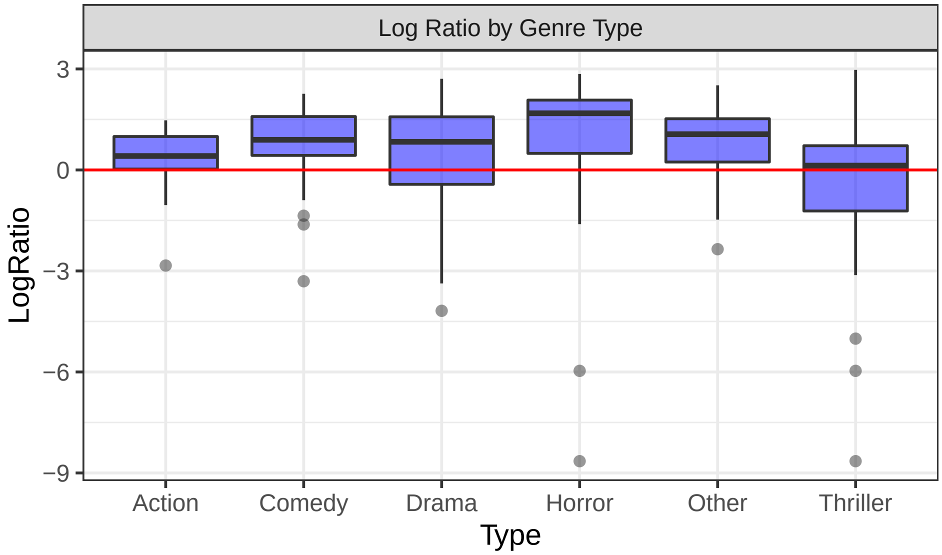

Budget vs. Net Profit
NetProfit
25000
20000
15000
10000
5000
0
0
1000
2000
3000
Budget

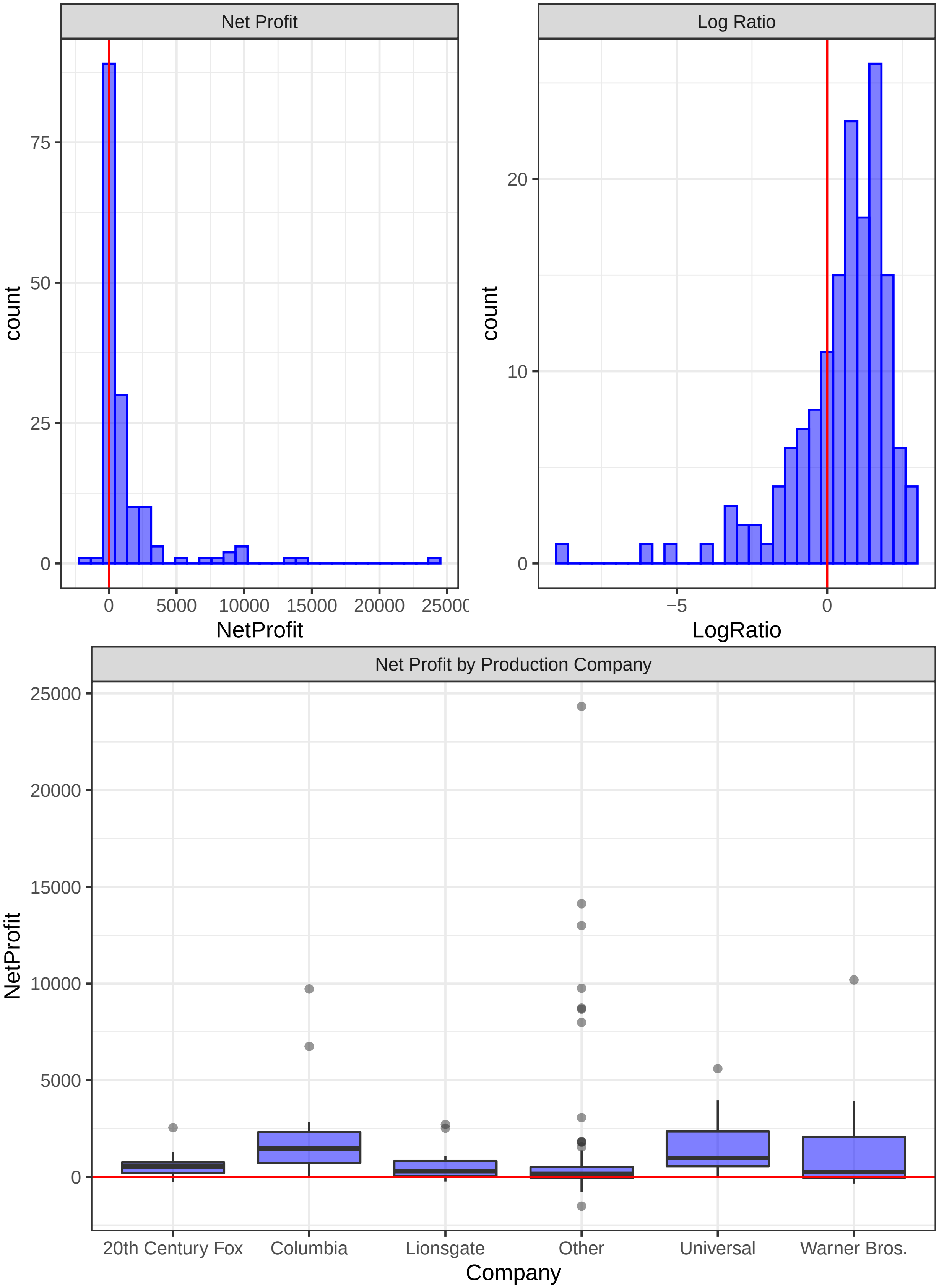

Critic Score vs. Net Profit
NetProfit
25000
20000
15000
10000
5000
0
3
4
5
6
7
8
Critic Score

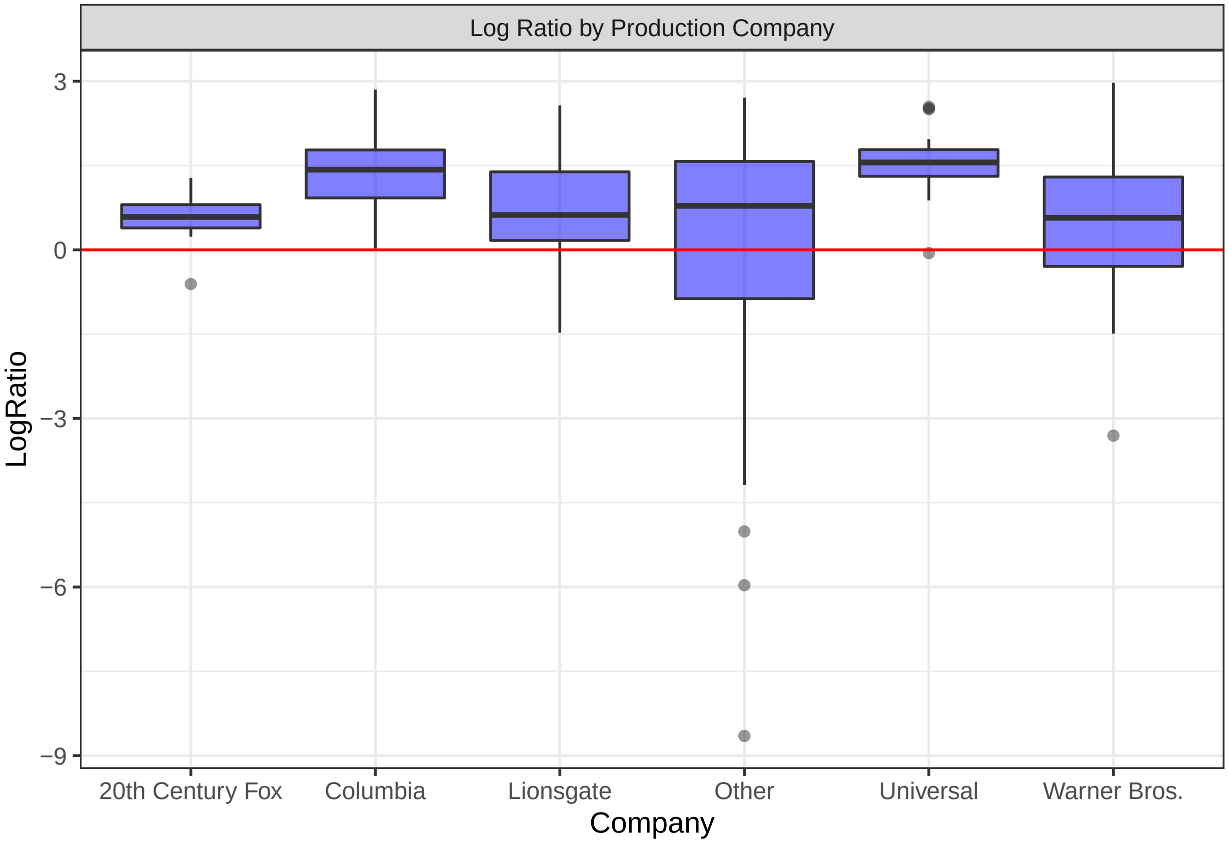

Net Profit by Month
NetProfit
25000
20000
15000
10000
5000
0
1
10
11
12
2
3
4
5
6
7
8
9
Month

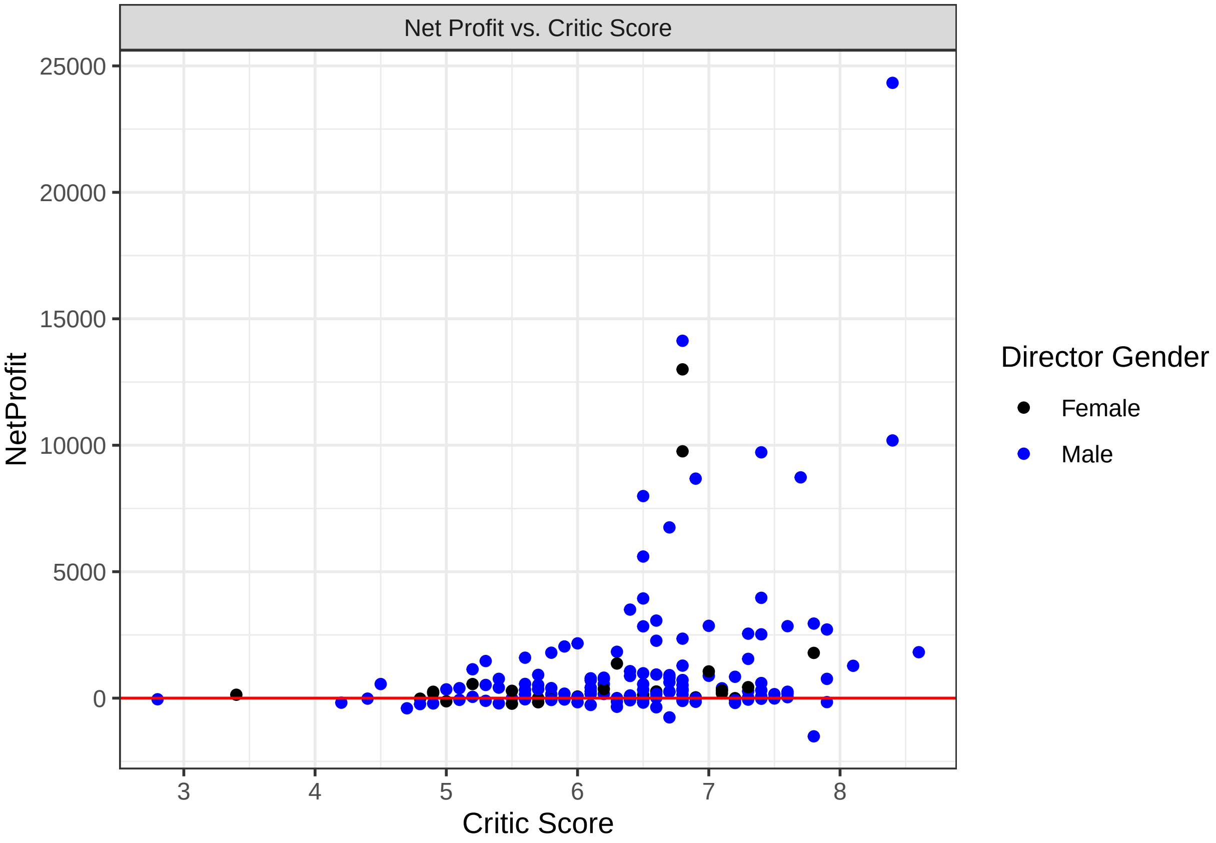

Log Ratio by Month
LogRatio
3
0
−3
−6
−9
1
10
11
12
2
3
4
5
6
7
8
9
Month

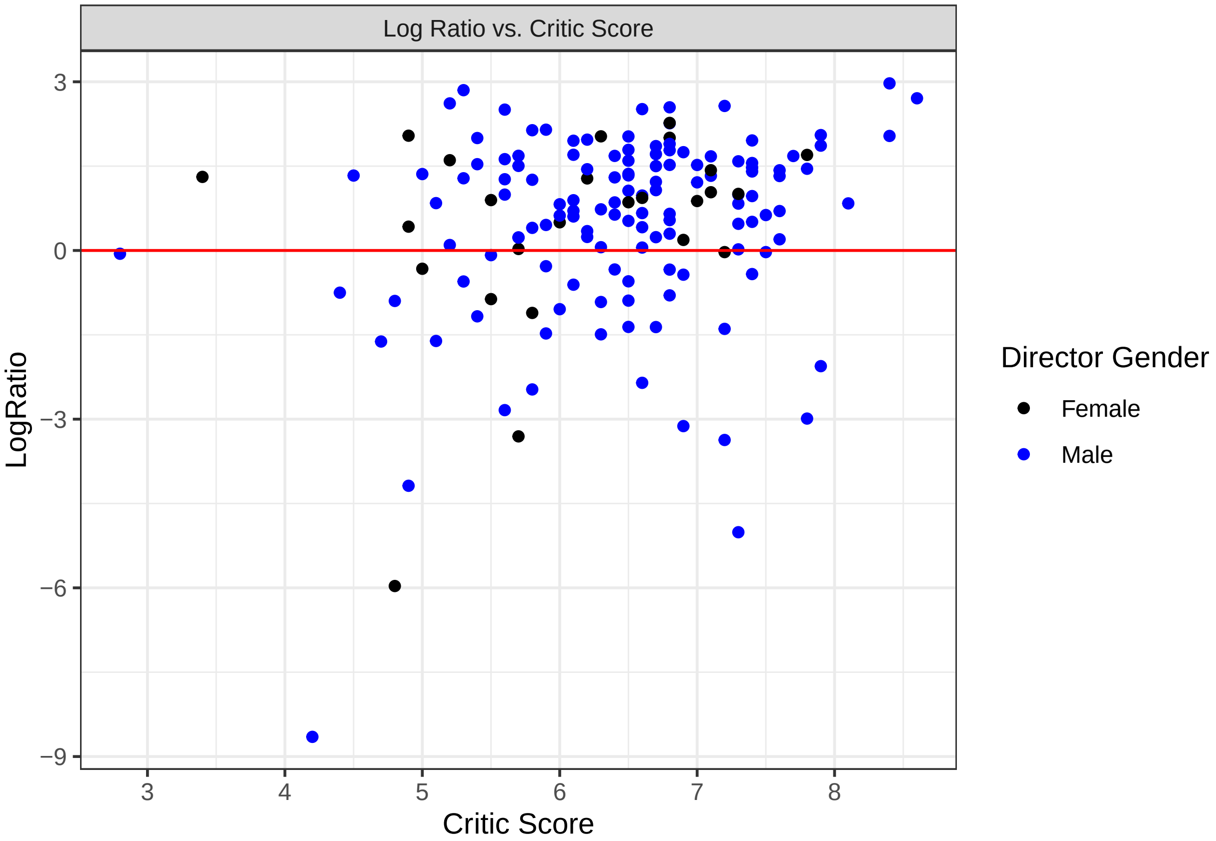

Net Profit by Genre Type
NetProfit
25000
20000
15000
10000
5000
0
Action
Comedy
Drama
Horror
Other
Thriller
Type

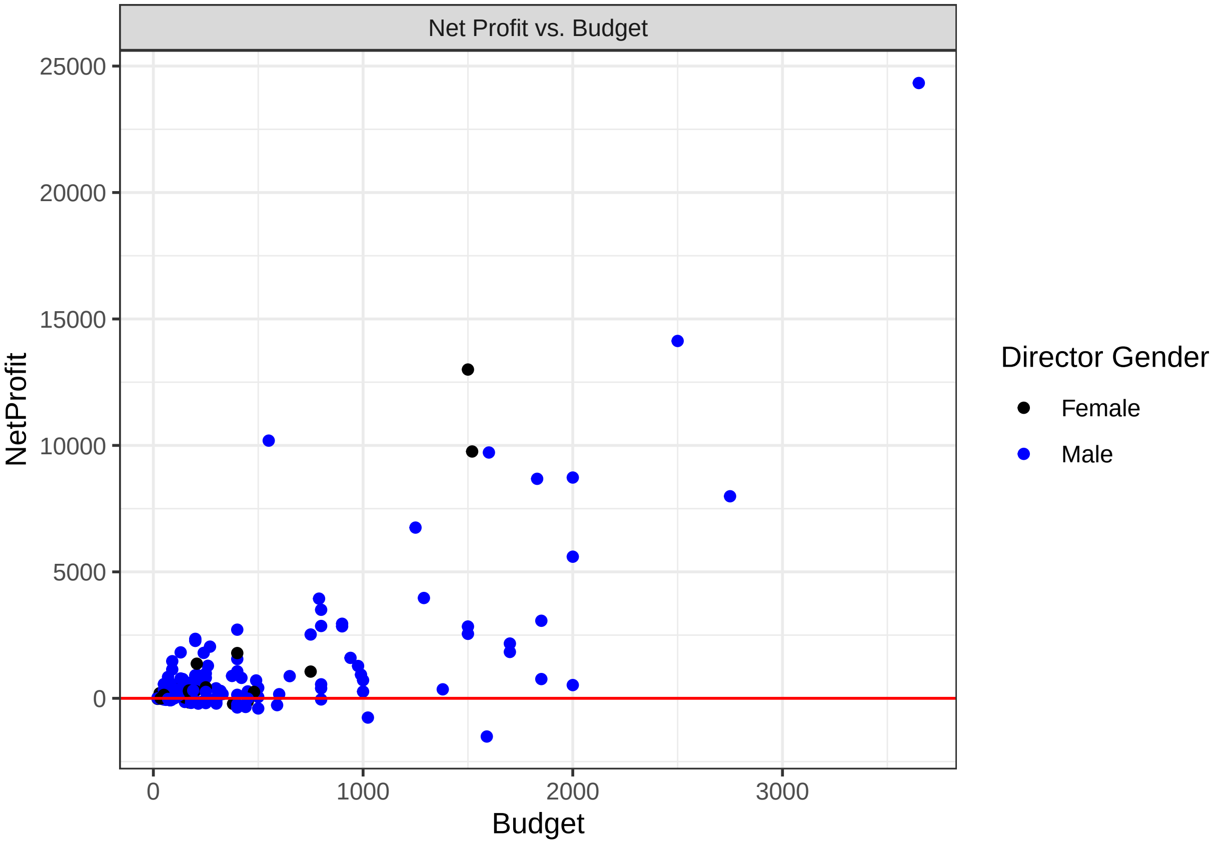

Log Ratio by Genre Type
LogRatio
3
0
−3
−6
−9
Action
Comedy
Drama
Horror
Other
Thriller
Type

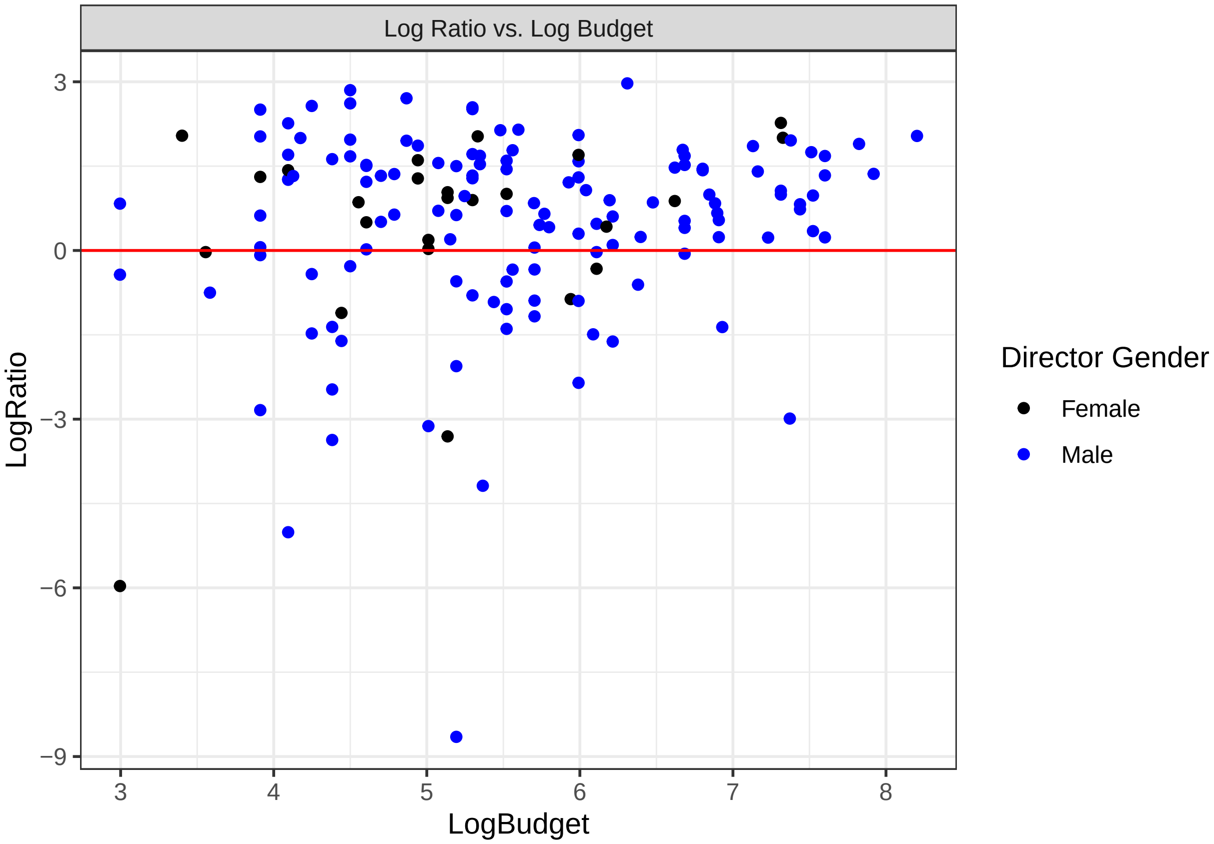

Net Profit
count
75
50
25
0
0
5000
10000
15000
20000
25000
NetProfit
Log Ratio
count
20
10
0
−5
0
LogRatio
Net Profit by Production Company
NetProfit
25000
20000
15000
10000
5000
0
20th Century Fox
Columbia
Lionsgate
Other
Universal
Warner Bros.
Company

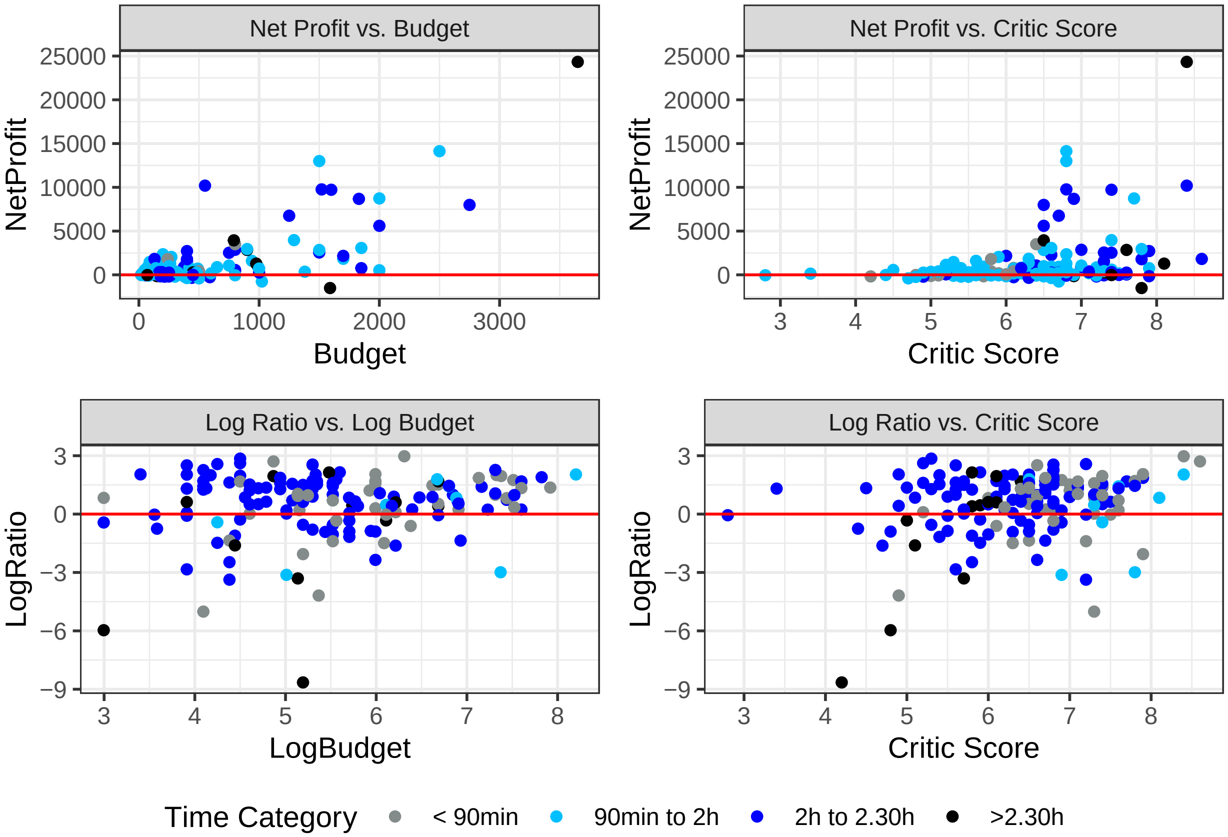

Log Ratio by Production Company
LogRatio
3
0
−3
−6
−9
20th Century Fox
Columbia
Lionsgate
Other
Universal
Warner Bros.
Company

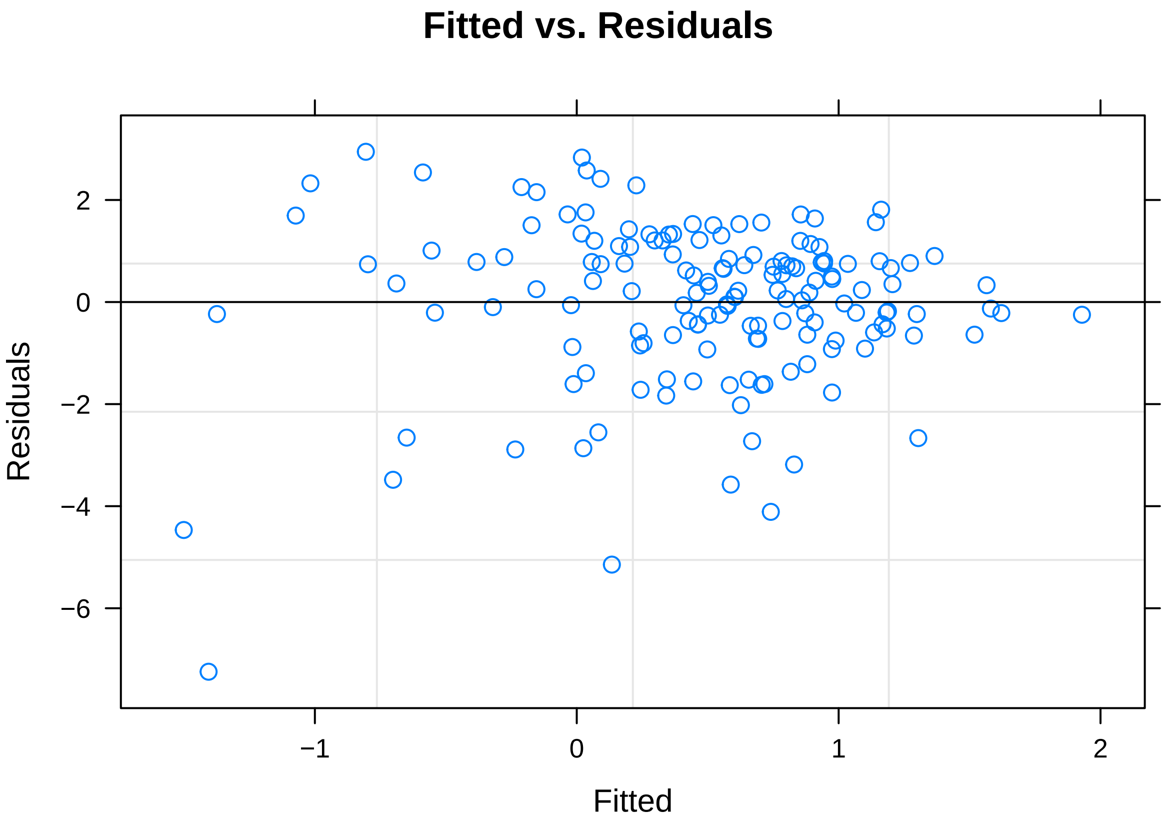

Net Profit vs. Critic Score
NetProfit
25000
20000
15000
10000
5000
0
3
4
5
6
7
8
Critic Score
Director Gender
● Female
● Male

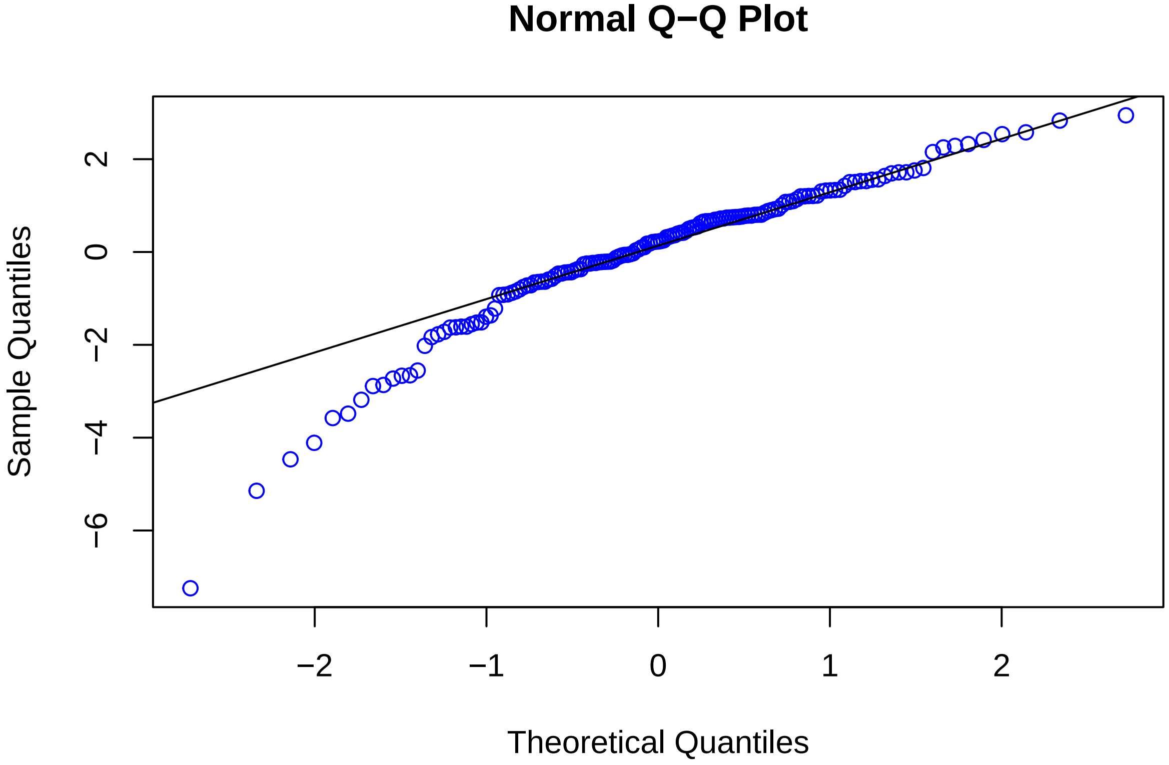

Log Ratio vs. Critic Score
LogRatio
3
0
−3
−6
−9
Critic Score
3
4
5
6
7
8
Director Gender
● Female
● Male

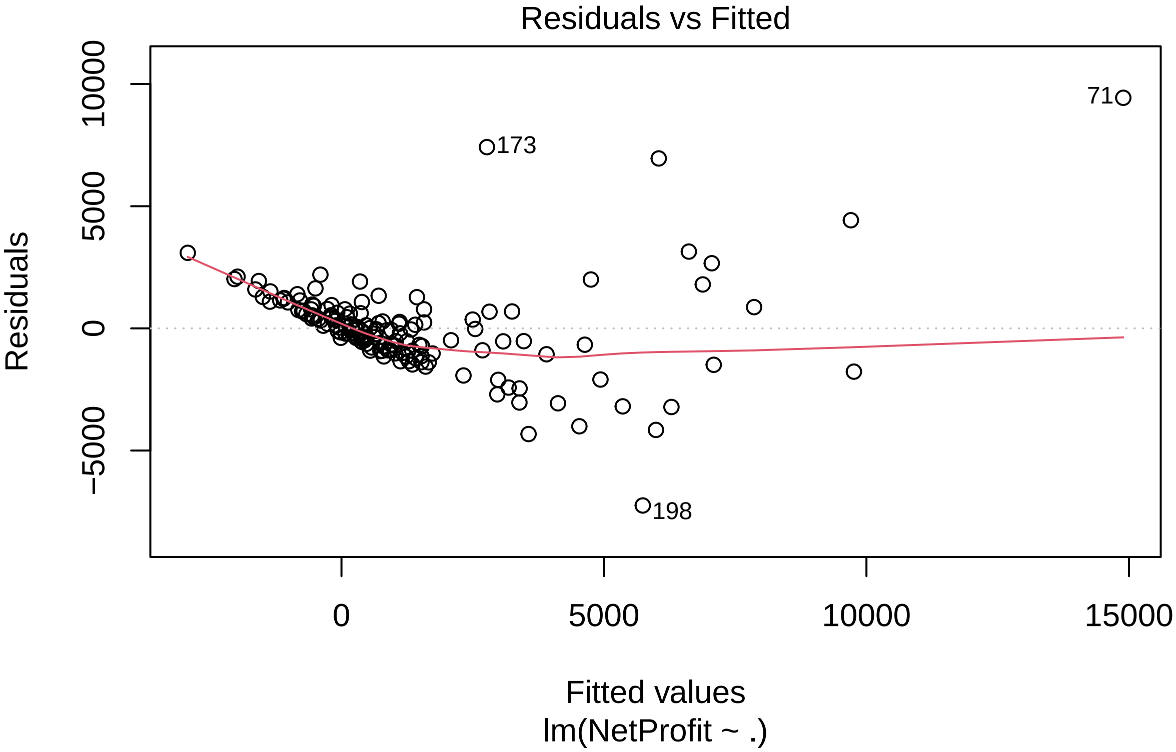

Net Profit vs. Budget
NetProfit
25000
20000
15000
10000
5000
0
Budget
0
1000
2000
3000
Director Gender
● Female
● Male

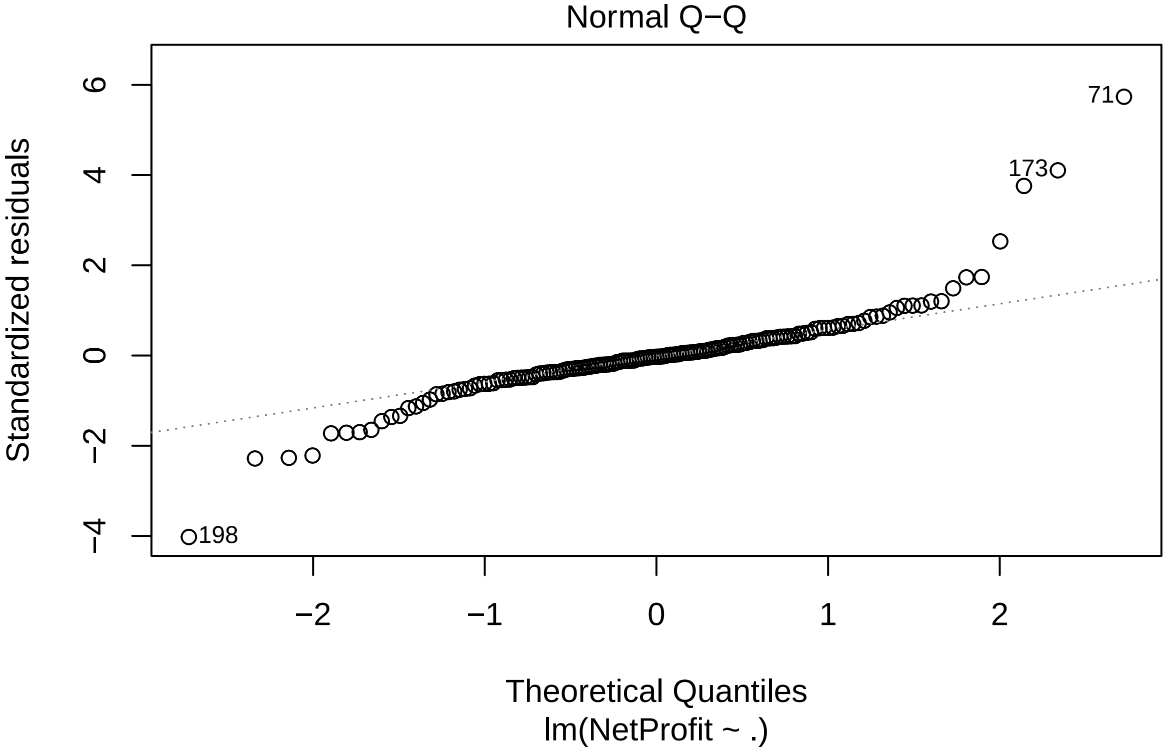

Log Ratio vs. Log Budget
LogRatio
3
0
−3
−6
−9
3
4
5
6
7
8
LogBudget
Director Gender
● Female
● Male

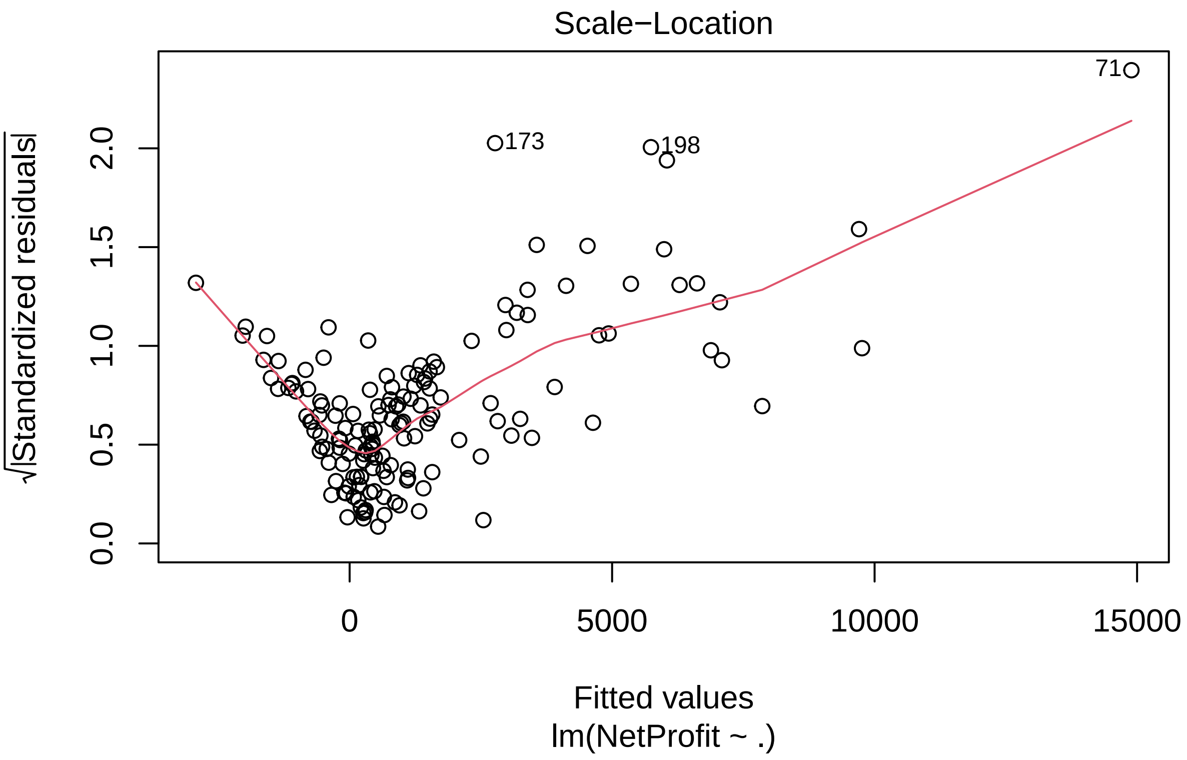

Net Profit vs. Budget
NetProfit
25000
20000
15000
10000
5000
0
0
1000
2000
3000
Budget
Net Profit vs. Critic Score
NetProfit
25000
20000
15000
10000
5000
0
3
4
5
6
7
8
Critic Score
Log Ratio vs. Log Budget
LogRatio
3
0
−3
−6
−9
3
4
5
6
7
8
LogBudget
Log Ratio vs. Critic Score
LogRatio
3
0
−3
−6
−9
3
4
5
6
7
8
Critic Score
Time Category
● < 90min
● 90min to 2h
● 2h to 2.30h
● >2.30h

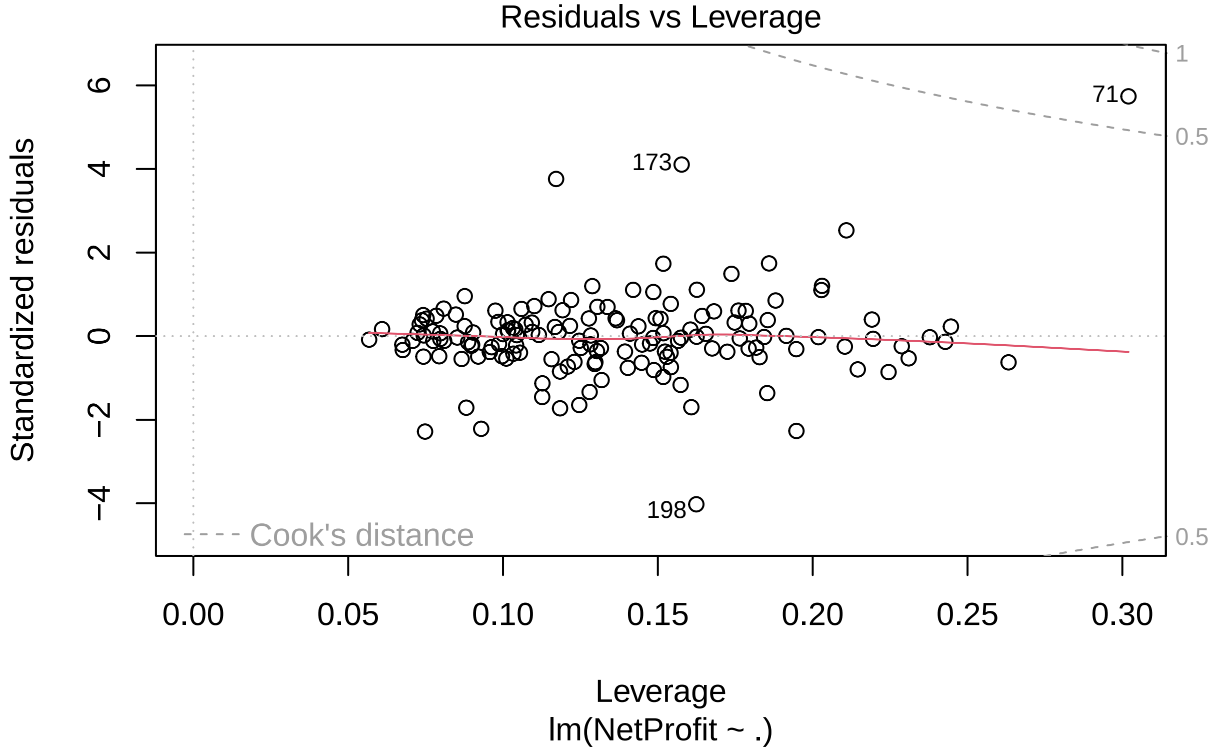

Fitted vs. Residuals
Residuals
2
0
−2
−4
−6
−1
0
1
2
Fitted


Normal Q−Q Plot
Sample Quantiles
2
0
−2
−4
−6
−2
−1
0
1
2
Theoretical Quantiles


Residuals vs Fitted
Residuals
10000
5000
0
−5000
0
5000
10000
15000
Fitted values
lm(NetProfit ~ .)
71
173
198


Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−2
−1
0
1
2
Theoretical Quantiles
lm(NetProfit ~ .)
71
173
198


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
5000
10000
15000
71
173
198
Fitted values
lm(NetProfit ~ .)


Residuals vs Leverage
Standardized residuals
−4
−2
0
2
4
6
0.00
0.05
0.10
0.15
0.20
0.25
0.30
71
173
198
1
0.5
0.5
Cook's distance
Leverage
lm(NetProfit ~ .)