Movie Dataset Analysis Project

2025-11-08

Introduction

This project analyzes movie data to predict US Revenue based on various factors such as budget, opening weekend revenue, number of theaters, ratings, and more. The goal is to build a regression model that explains the variability in US Revenue. Below, we document the step-by-step process of model development, including exploratory data analysis, initial model building, variable selection, transformations, and interactions, leading to the final model.

Data Preparation

```
# Load the dataset
movies <- read.csv('Movie.csv')
head(movies)
```

```
##                       Title USRelease            Genre Rating Sequel Budget
## 1              Man of Steel    16-Jun Action/Adventure  PG-13      0    225
## 2         Monster University    23-Jun        Animation      G      0    200
## 3           Fast & Furious 6    26-May Action/Adventure  PG-13      1    160
## 4 Oz the Great and Powerful    10-Mar Action/Adventure     PG      0    215
## 5  Star Trek: Into Darkness    19-May Action/Adventure  PG-13      1    190
## 6                The Croods    24-Mar        Animation     PG      0    135
##   Opening USRevenue Theaters IntRevenue WorldRevenue Ratings Review Minutes
## 1   116.6     291.0     4207      377.0        668.0     7.1     55     143
## 2    82.4     268.5     4004      475.1        743.6     7.3     65     104
## 3    97.4     238.7     3658      550.0        788.7     7.1     61     130
## 4    79.1     234.9     3912      258.4        493.3     6.3     44     130
## 5    70.2     228.8     3868      238.6        467.4     7.7     72     132
## 6    43.6     187.2     4046      400.0        587.2     7.2     55      98
```

```
summary(movies)
```

```
##     Title            USRelease            Genre              Rating
##  Length:45          Length:45          Length:45          Length:45
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      Sequel           Budget          Opening         USRevenue
```

```
##  Min.   :0.0000   Min.   :  3.00   Min.   :  4.60   Min.   :  8.80
##  1st Qu.:0.0000   1st Qu.: 24.50   1st Qu.: 12.18   1st Qu.: 32.15
##  Median :0.0000   Median : 55.00   Median : 23.10   Median : 71.45
##  Mean   :0.2045   Mean   : 77.08   Mean   : 30.69   Mean   : 91.23
##  3rd Qu.:0.0000   3rd Qu.:122.50   3rd Qu.: 40.75   3rd Qu.:125.03
##  Max.   :1.0000   Max.   :225.00   Max.   :116.60   Max.   :291.00
##  NA's   :1        NA's   :1        NA's   :1        NA's   :1
##     Theaters      IntRevenue      WorldRevenue      Ratings
##  Min.   :2023   Min.   :  0.20   Min.   :  9.3   Min.   :3.500
##  1st Qu.:2832   1st Qu.: 29.75   1st Qu.: 67.3   1st Qu.:6.050
##  Median :3192   Median : 66.20   Median :147.2   Median :6.500
##  Mean   :3169   Mean   :133.13   Mean   :224.4   Mean   :6.382
##  3rd Qu.:3581   3rd Qu.:214.00   3rd Qu.:326.2   3rd Qu.:6.900
##  Max.   :4207   Max.   :550.00   Max.   :788.7   Max.   :7.700
##  NA's   :1      NA's   :1        NA's   :1       NA's   :1
##      Review         Minutes
##  Min.   :11.00   Min.   : 86.00
##  1st Qu.:40.75   1st Qu.: 97.75
##  Median :52.00   Median :108.00
##  Mean   :48.55   Mean   :110.59
##  3rd Qu.:60.00   3rd Qu.:124.25
##  Max.   :75.00   Max.   :143.00
##  NA's   :1       NA's   :1
```

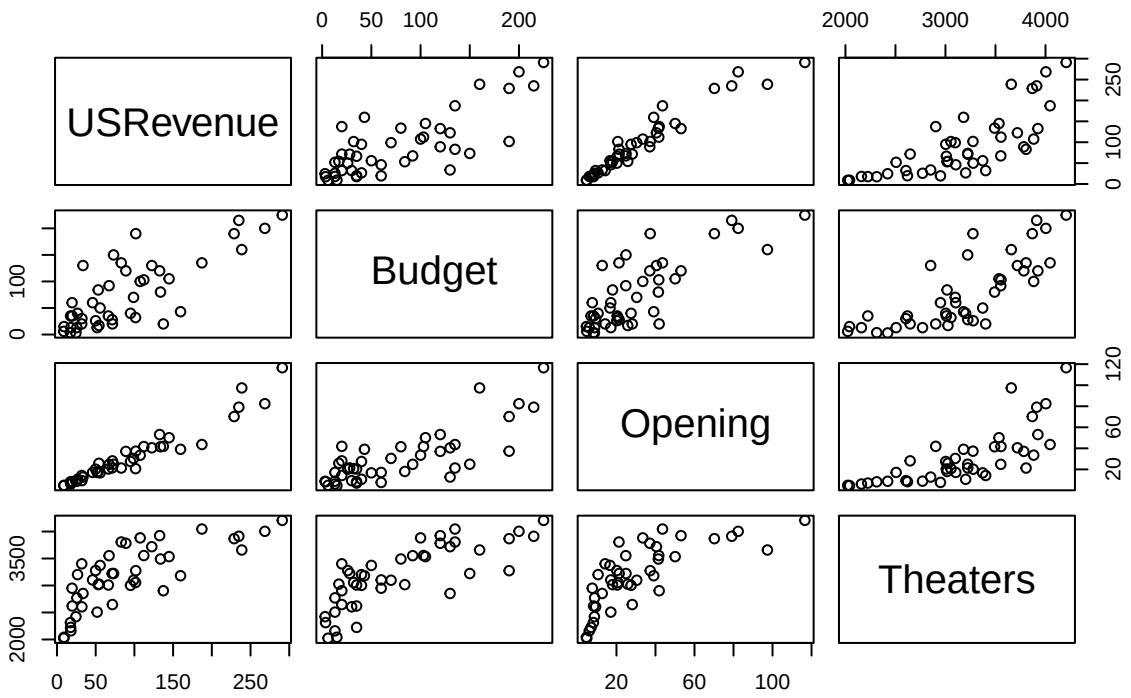
```
# Avoid View() in knitted output to prevent DataTables re-init issues
if (interactive()) View(movies)
movies <- head(movies, -1)  # Remove the last row if it's invalid
tail(movies)
```

```
##                  Title USRelease            Genre Rating Sequel Budget Opening
## 39 Beautiful Creatures    14-Feb            Drama  PG-13      0   60.0     7.6
## 40           Admission    22-Mar           Comedy  PG-13      0   13.0     6.1
## 41              Parker    25-Jan Action/Adventure      R      0   35.0     7.0
## 42          Dark Skies    22-Feb           Horror  PG-13      0    3.5     8.2
## 43             Peeples    10-May           Comedy  PG-13      0   15.0     4.6
## 44            Movie 43    25-Jan           Comedy      R      0    6.0     4.8
##    USRevenue Theaters IntRevenue WorldRevenue Ratings Review Minutes
## 39      19.5     2950       40.5         60.0     6.2     52     124
## 40      18.0     2160        0.7         18.7     5.7     48     107
## 41      17.6     2224       30.9         48.5     6.2     42     118
## 42      17.4     2313       10.3         27.7     6.3     50      97
## 43       9.1     2041        0.2          9.3     5.3     52      95
## 44       8.8     2023       22.3         31.1     4.3     18      94
```

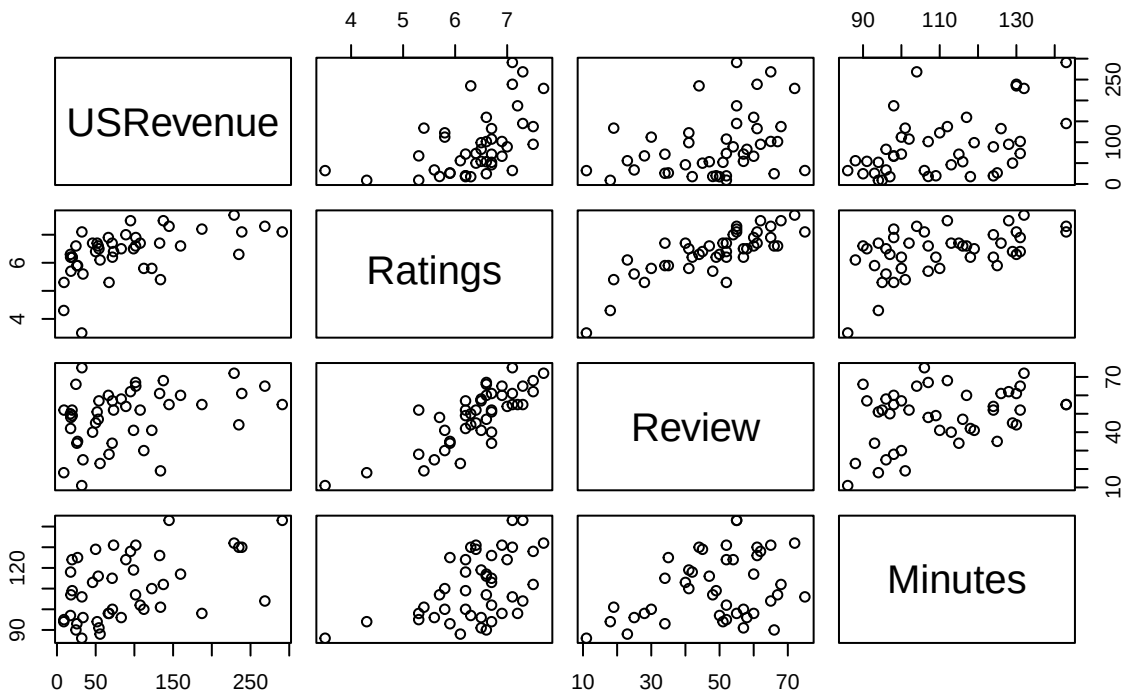
Exploratory Data Analysis

We begin by visualizing the relationships between variables to understand potential correlations and multicollinearity.

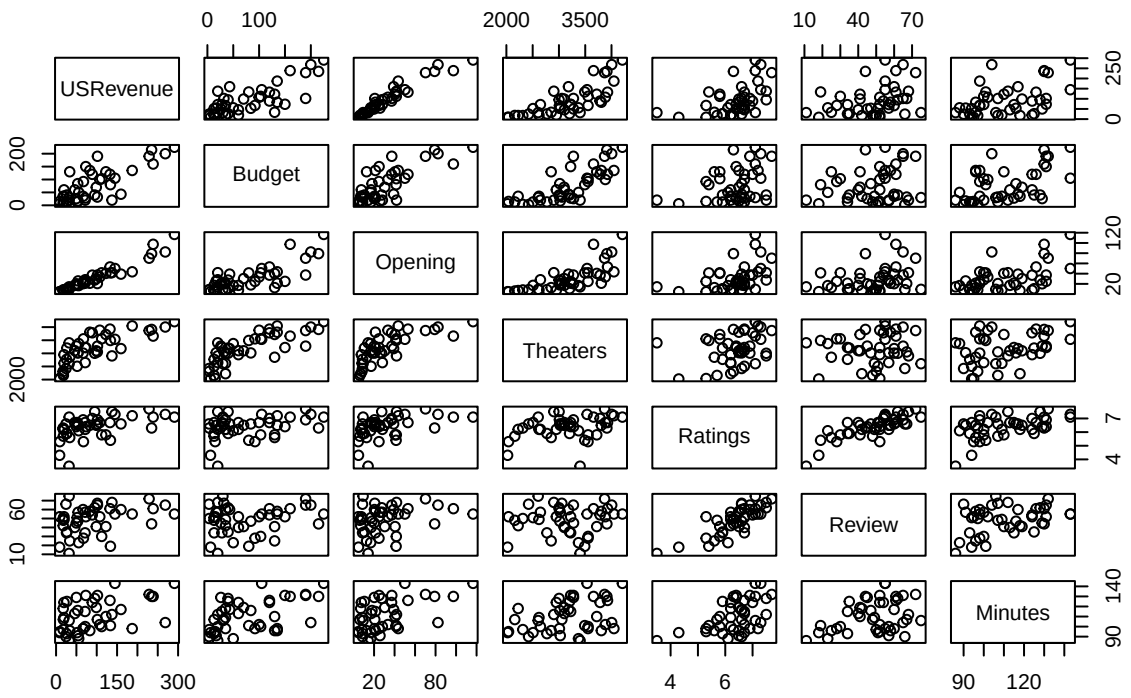
```
pairs(~USRevenue+Budget+Opening+Theaters,data = movies,
      main="Scatterplot Matrix of Key Numerical Variables")
```




```
pairs(~USRevenue+Ratings+Review+Minutes,data = movies,
      main="Scatterplot Matrix Including Ratings and Review")
```




```
pairs(~USRevenue+Budget+Opening+Theaters+Ratings+Review+Minutes,data = movies,
      main="Complete Scatterplot Matrix to Check Multicollinearity")
```




```
library(PerformanceAnalytics)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

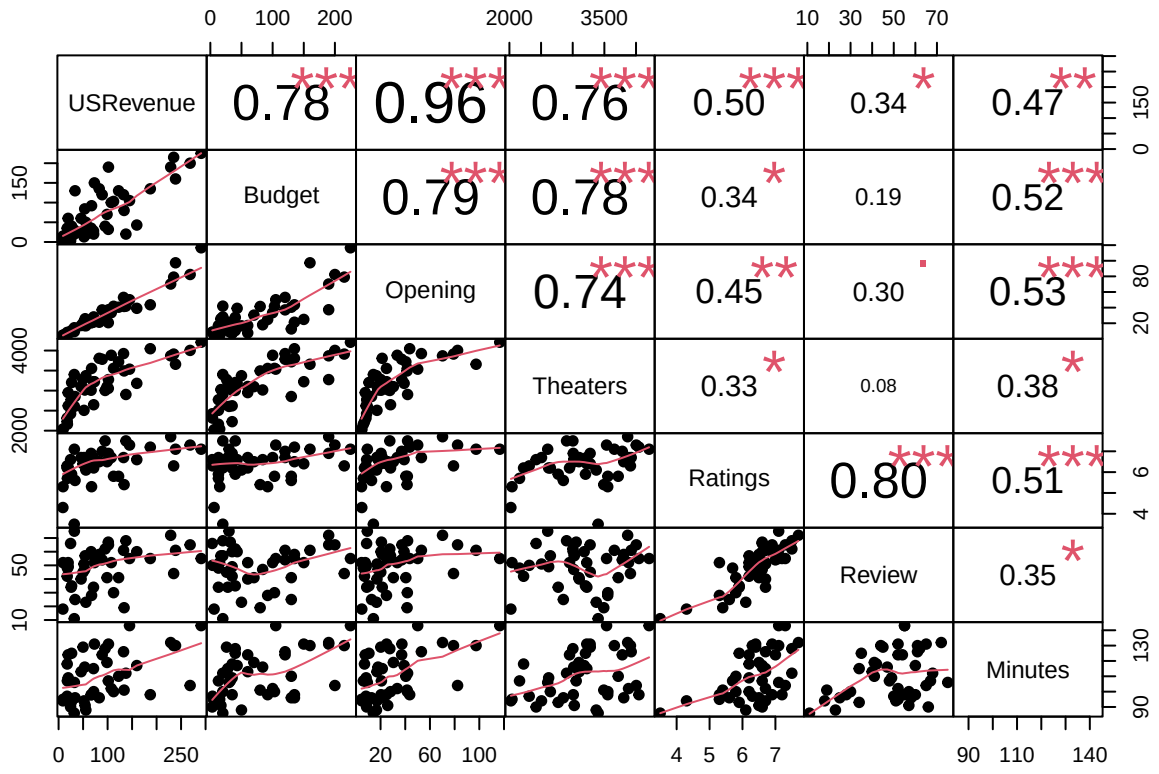
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
chart.Correlation(movies[,c("USRevenue","Budget","Opening","Theaters", "Ratings", "Review", "Minutes")]
```



Initial Model Building

We start with a full model including all potential predictors: numerical variables (Budget, Opening, Theaters, Ratings, Minutes), and categorical variables (Genre, Rating, Sequel).

```
names(movies)
```

```
##  [1] "Title"        "USRelease"    "Genre"        "Rating"       "Sequel"
##  [6] "Budget"       "Opening"      "USRevenue"    "Theaters"     "IntRevenue"
## [11] "WorldRevenue" "Ratings"      "Review"       "Minutes"
```

```
Genre   <- factor(movies$Genre)
Rating <- factor(movies$Rating)
s <- ifelse(movies$Sequel == 1, 1, 0)  # Binary for Sequel

model55 = lm(USRevenue~Budget+Opening+Theaters+Ratings+Minutes+s+factor(Genre)+factor(Rating), data=mov
summary(model55)
```

```
##
## Call:
## lm(formula = USRevenue ~ Budget + Opening + Theaters + Ratings +
##     Minutes + s + factor(Genre) + factor(Rating), data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.02 -10.18   0.00   7.71  32.49
```

```
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -72.381544  49.022387  -1.476   0.1506
## Budget                     0.044302   0.103259   0.429   0.6711
## Opening                    2.278943   0.243749   9.350 2.97e-10 ***
## Theaters                   0.009989   0.009299   1.074   0.2916
## Ratings                   12.368468   4.938513   2.504   0.0181 *
## Minutes                    0.144467   0.334930   0.431   0.6694
## s                          7.143482   8.692724   0.822   0.4179
## factor(Genre)Animation    -1.074878  22.923535  -0.047   0.9629
## factor(Genre)Comedy       23.562711   9.731424   2.421   0.0220 *
## factor(Genre)Crime/Drama   3.473057  15.219870   0.228   0.8211
## factor(Genre)Drama         5.808265   9.937879   0.584   0.5634
## factor(Genre)Horror        7.606842  14.187448   0.536   0.5959
## factor(Rating)PG         -18.287552  20.861010  -0.877   0.3879
## factor(Rating)PG-13      -51.755002  30.174469  -1.715   0.0970 .
## factor(Rating)R          -45.397571  31.190685  -1.455   0.1563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.31 on 29 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9507
## F-statistic: 60.26 on 14 and 29 DF,  p-value: < 2.2e-16
```

Checking Categorical Variables

To understand the impact of categorical variables, we examine their means.

```
aggregate(USRevenue ~ Sequel, data = movies, mean)
```

```
##   Sequel USRevenue
## 1      0  82.65429
## 2      1 124.56667
```

```
aggregate(USRevenue ~ Genre, data = movies, mean)
```

```
##              Genre USRevenue
## 1 Action/Adventure 108.96111
## 2        Animation 161.55000
## 3           Comedy  62.86364
## 4      Crime/Drama  42.05000
## 5            Drama  70.06000
## 6           Horror  70.15000
```

Simplifying the Model: Focusing on Categorical Variables

Given multicollinearity concerns, we first model only the categorical variables to see their individual effects.

```r
model54 = lm(USRevenue~s+factor(Genre)+factor(Rating), data=movies)
summary(model54)
```

```
##
## Call:
## lm(formula = USRevenue ~ s + factor(Genre) + factor(Rating),
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.347 -38.365  -8.965  21.229 201.095
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                377.50      97.20   3.884 0.000451 ***
## s                           48.74      26.66   1.829 0.076253 .
## factor(Genre)Animation    -109.00      73.48  -1.483 0.147161
## factor(Genre)Comedy        -31.75      25.01  -1.270 0.212811
## factor(Genre)Crime/Drama   -32.12      49.98  -0.643 0.524825
## factor(Genre)Drama         -16.70      34.12  -0.489 0.627779
## factor(Genre)Horror        -11.88      36.53  -0.325 0.746931
## factor(Rating)PG          -142.60      73.48  -1.941 0.060615 .
## factor(Rating)PG-13       -287.60      99.37  -2.894 0.006596 **
## factor(Rating)R           -303.33      99.61  -3.045 0.004467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.63 on 34 degrees of freedom
## Multiple R-squared:  0.4067, Adjusted R-squared:  0.2497
## F-statistic:  2.59 on 9 and 34 DF,  p-value: 0.0216
```

```r
model53 = lm(USRevenue~factor(Rating), data=movies)
summary(model53)
```

```
##
## Call:
## lm(formula = USRevenue ~ factor(Rating), data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84.98 -43.36 -11.63  36.90 196.92
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           268.50      65.40   4.105 0.000193 ***
## factor(Rating)PG     -115.35      73.12  -1.577 0.122563
## factor(Rating)PG-13  -174.42      66.87  -2.608 0.012736 *
## factor(Rating)R      -205.96      67.30  -3.060 0.003937 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.4 on 40 degrees of freedom
```

```
## Multiple R-squared:  0.2626, Adjusted R-squared:  0.2073
## F-statistic: 4.749 on 3 and 40 DF,  p-value: 0.006326
```

```
model52 = lm(USRevenue~s, data=movies)
summary(model52)
```

```
##
## Call:
## lm(formula = USRevenue ~ s, data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -92.57 -57.03 -11.86  18.92 208.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    82.65      12.22   6.765 3.16e-08 ***
## s              41.91      27.02   1.551    0.128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.29 on 42 degrees of freedom
## Multiple R-squared:  0.0542, Adjusted R-squared:  0.03168
## F-statistic: 2.407 on 1 and 42 DF,  p-value: 0.1283
```

```
model51 = lm(USRevenue~factor(Genre), data=movies)
summary(model51)
```

```
##
## Call:
## lm(formula = USRevenue ~ factor(Genre), data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91.36 -52.88 -13.01  28.90 182.04
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 108.96      16.62   6.557 9.84e-08 ***
## factor(Genre)Animation       52.59      38.97   1.349   0.1852
## factor(Genre)Comedy         -46.10      26.98  -1.708   0.0957 .
## factor(Genre)Crime/Drama    -66.91      52.55  -1.273   0.2107
## factor(Genre)Drama          -38.90      35.64  -1.091   0.2820
## factor(Genre)Horror         -38.81      38.97  -0.996   0.3256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.5 on 38 degrees of freedom
## Multiple R-squared:  0.1859, Adjusted R-squared:  0.07883
## F-statistic: 1.736 on 5 and 38 DF,  p-value: 0.15
```

Incorporating Numerical Variables with Transformations

We add back numerical variables, starting with Opening and Ratings, and try a quadratic term for Ratings.

```
model50 = lm(USRevenue~Opening+I(Ratings^2)+factor(Rating), data=movies)
summary(model50)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + factor(Rating),
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.199  -9.801  -1.351   4.650  46.316
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           23.4040    21.9797   1.065   0.2937
## Opening                2.5597     0.1220  20.983   <2e-16 ***
## I(Ratings^2)           0.6413     0.2987   2.147   0.0382 *
## factor(Rating)PG     -12.4867    18.8030  -0.664   0.5106
## factor(Rating)PG-13  -42.7897    17.6107  -2.430   0.0199 *
## factor(Rating)R      -38.1403    18.2077  -2.095   0.0429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.39 on 38 degrees of freedom
## Multiple R-squared:  0.956,  Adjusted R-squared:  0.9502
## F-statistic: 165.1 on 5 and 38 DF,  p-value: < 2.2e-16
```

```
model49 = lm(USRevenue~Opening+Ratings+factor(Rating), data=movies)
summary(model49)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + Ratings + factor(Rating),
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.325 -10.006  -2.222   4.518  46.138
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.0260    28.5790   0.211   0.8341
## Opening                2.5750     0.1217  21.160   <2e-16 ***
## Ratings                6.8894     3.5088   1.963   0.0569 .
## factor(Rating)PG     -13.1290    18.9658  -0.692   0.4930
## factor(Rating)PG-13  -43.2854    17.7600  -2.437   0.0196 *
## factor(Rating)R      -38.7172    18.3646  -2.108   0.0417 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.54 on 38 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9493
## F-statistic: 162.1 on 5 and 38 DF,  p-value: < 2.2e-16
```

```
# Model50 with Ratings^2 performs better.
```

Adding Significant Genre Indicator

From the full model, Comedy was the only significant Genre. We add it back.

```
movies$c <- ifelse(movies$Genre == "Comedy", 1, 0)
model48 = lm(USRevenue~Opening+I(Ratings^2)+factor(Rating)+c, data=movies)
summary(model48)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + factor(Rating) +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.218  -9.225  -1.846   8.025  34.202
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.3135    21.4592   0.201  0.84179
## Opening               2.5553     0.1127  22.675  < 2e-16 ***
## I(Ratings^2)          1.0064     0.3062   3.286  0.00223 **
## factor(Rating)PG     -9.5078    17.4015  -0.546  0.58809
## factor(Rating)PG-13 -42.4343    16.2669  -2.609  0.01304 *
## factor(Rating)R     -39.5664    16.8258  -2.352  0.02413 *
## c                    16.9006     6.1546   2.746  0.00926 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.14 on 37 degrees of freedom
## Multiple R-squared:  0.9634, Adjusted R-squared:  0.9575
## F-statistic: 162.5 on 6 and 37 DF,  p-value: < 2.2e-16
```

Verification of Comedy's association:

```
aggregate(USRevenue ~ Genre, data = movies, mean)
```

```
##              Genre USRevenue
## 1 Action/Adventure 108.96111
## 2        Animation 161.55000
## 3           Comedy  62.86364
## 4      Crime/Drama  42.05000
## 5            Drama  70.06000
## 6           Horror  70.15000
```

Experimenting with Other Genres

We try Animation instead of Comedy, but it wasn't significant, so we stick with Comedy.

```
movies$a <- ifelse(movies$Genre == "Animation", 1, 0)
model47 = lm(USRevenue~Opening+I(Ratings^2)+factor(Rating)+a, data=movies)
summary(model47)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + factor(Rating) +
##     a, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.227 -10.255  -1.130   4.556  46.140
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           17.4672    30.0520   0.581   0.5646
## Opening                2.5731     0.1316  19.552   <2e-16 ***
## I(Ratings^2)           0.6189     0.3118   1.985   0.0546 .
## factor(Rating)PG     -10.6644    20.0179  -0.533   0.5974
## factor(Rating)PG-13  -36.3858    28.1541  -1.292   0.2042
## factor(Rating)R      -31.5757    28.9600  -1.090   0.2826
## a                      6.0272    20.5094   0.294   0.7705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.59 on 37 degrees of freedom
## Multiple R-squared:  0.9561, Adjusted R-squared:  0.949
## F-statistic: 134.3 on 6 and 37 DF,  p-value: < 2.2e-16
```

Adding More Variables Back

We add Budget and Theaters back to see if they improve the model.

```
model46 = lm(USRevenue~Budget+Opening+I(Ratings^2)+factor(Rating)+c, data=movies)
summary(model46)
```

```
##
## Call:
## lm(formula = USRevenue ~ Budget + Opening + I(Ratings^2) + factor(Rating) +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.677  -9.247  -1.583   8.485  34.489
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          -1.92386   22.78278  -0.084  0.93317
## Budget                0.06029    0.07162   0.842  0.40547
## Opening               2.45139    0.16742  14.642  < 2e-16 ***
## I(Ratings^2)          1.05781    0.31346   3.375  0.00178 **
## factor(Rating)PG     -9.77411   17.47330  -0.559  0.57937
## factor(Rating)PG-13 -40.57461   16.48009  -2.462  0.01874 *
## factor(Rating)R     -36.25587   17.34422  -2.090  0.04371 *
## c                    18.77029    6.56608   2.859  0.00703 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.2 on 36 degrees of freedom
## Multiple R-squared:  0.9642, Adjusted R-squared:  0.9572
## F-statistic: 138.3 on 7 and 36 DF,  p-value: < 2.2e-16
```

```
model45 = lm(USRevenue~Theaters+Opening+I(Ratings^2)+factor(Rating)+c, data=movies)
summary(model45)
```

```
##
## Call:
## lm(formula = USRevenue ~ Theaters + Opening + I(Ratings^2) +
##     factor(Rating) + c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.140 -11.355   0.023   6.034  32.583
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -31.042913  28.965907  -1.072  0.29098
## Theaters              0.011663   0.006624   1.761  0.08679 .
## Opening               2.379704   0.148194  16.058  < 2e-16 ***
## I(Ratings^2)          1.065083   0.299761   3.553  0.00109 **
## factor(Rating)PG    -14.599292  17.172997  -0.850  0.40087
## factor(Rating)PG-13 -41.412445  15.834702  -2.615  0.01294 *
## factor(Rating)R     -37.641987  16.404195  -2.295  0.02769 *
## c                    19.161577   6.123236   3.129  0.00347 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.73 on 36 degrees of freedom
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9598
## F-statistic: 147.7 on 7 and 36 DF,  p-value: < 2.2e-16
```

Simplifying Rating Categories

We create a binary for high-rated movies (PG-13, R).

```
movies$highrate <- ifelse(movies$Rating %in% c("PG-13", "R"),1, 0)
model44 = lm(USRevenue~Budget+Opening+I(Ratings^2)+highrate+c, data=movies)
summary(model44)
```

```
##
## Call:
## lm(formula = USRevenue ~ Budget + Opening + I(Ratings^2) + highrate +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.297 -10.918  -1.587   8.301  35.997
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.83599   16.14083  -0.547 0.587284
## Budget         0.04096    0.06698   0.611 0.544566
## Opening        2.47261    0.16327  15.145  < 2e-16 ***
## I(Ratings^2)   1.07992    0.30843   3.501 0.001200 **
## highrate     -32.04749    8.04167  -3.985 0.000295 ***
## c             19.00469    6.46837   2.938 0.005587 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15 on 38 degrees of freedom
## Multiple R-squared:  0.9632, Adjusted R-squared:  0.9583
## F-statistic: 198.8 on 5 and 38 DF,  p-value: < 2.2e-16
```

Trying Interactions

We experiment with interactions, like Theaters*Budget.

```
model43= lm(USRevenue~Budget+Theaters+Opening+I(Ratings^2)+highrate+c+Theaters*Budget, data=movies)
summary(model43)
```

```
##
## Call:
## lm(formula = USRevenue ~ Budget + Theaters + Opening + I(Ratings^2) +
##     highrate + c + Theaters * Budget, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1507  -9.3279  -0.2116   8.1713  30.3583
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4.191e+01  2.646e+01  -1.584  0.12193
## Budget           4.147e-01  3.525e-01   1.176  0.24715
## Theaters         1.203e-02  7.089e-03   1.697  0.09838 .
## Opening          2.555e+00  2.249e-01  11.360 1.85e-13 ***
## I(Ratings^2)     1.059e+00  3.058e-01   3.464  0.00139 **
## highrate        -3.469e+01  9.654e+00  -3.593  0.00097 ***
## c                2.056e+01  6.403e+00   3.211  0.00278 **
## Budget:Theaters -1.233e-04  1.050e-04  -1.175  0.24775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14.71 on 36 degrees of freedom
## Multiple R-squared:  0.9664, Adjusted R-squared:  0.9599
## F-statistic:   148 on 7 and 36 DF,  p-value: < 2.2e-16
```

Quadratic for Theaters and Interaction

Based on pairplots showing quadratic relationship with Theaters.

```
model42= lm(USRevenue~I(Theaters^2)+Opening+I(Ratings^2)+highrate+c+I(Theaters^2)*Opening, data=movies)
summary(model42)
```

```
##
## Call:
## lm(formula = USRevenue ~ I(Theaters^2) + Opening + I(Ratings^2) +
##     highrate + c + I(Theaters^2) * Opening, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3630  -8.3427   0.4639   7.0032  29.0647
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.512e+01  1.927e+01  -1.303 0.200469
## I(Theaters^2)          2.023e-06  1.112e-06   1.820 0.076813 .
## Opening                3.259e+00  4.907e-01   6.641 8.57e-08 ***
## I(Ratings^2)           9.680e-01  2.997e-01   3.230 0.002600 **
## highrate              -3.209e+01  8.241e+00  -3.894 0.000398 ***
## c                      1.924e+01  5.885e+00   3.269 0.002335 **
## I(Theaters^2):Opening -5.449e-08  2.953e-08  -1.845 0.072995 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.24 on 37 degrees of freedom
## Multiple R-squared:  0.9676, Adjusted R-squared:  0.9624
## F-statistic: 184.4 on 6 and 37 DF,  p-value: < 2.2e-16
```

```
# This model has the highest Adjusted R-squared so far (.9624).
```

Comparison with a simpler model:

```
model41 = lm(USRevenue~Opening+I(Ratings^2)+highrate+c, data=movies)
summary(model41)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + highrate +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -23.709 -10.695  -1.866   7.773  35.317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.4016    14.3034  -0.308  0.75992
## Opening        2.5488     0.1046  24.371  < 2e-16 ***
## I(Ratings^2)   1.0375     0.2981   3.480  0.00125 **
## highrate     -33.7473     7.4850  -4.509 5.82e-05 ***
## c             17.5773     5.9837   2.938  0.00553 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 39 degrees of freedom
## Multiple R-squared:  0.9628, Adjusted R-squared:  0.959
## F-statistic: 252.4 on 4 and 39 DF,  p-value: < 2.2e-16
```

Adding Release Month

We consider the release month as a factor.

```
movies$USReleaseMonth <- sub("^[0-9]+-", "", movies$USRelease)
model40 = lm(USRevenue~Opening+I(Ratings^2)+highrate+c+USReleaseMonth, data=movies)
summary(model40)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + highrate +
##     c + USReleaseMonth, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.718  -9.804  -1.782   9.506  33.280
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.9768    16.2598  -0.614 0.543693
## Opening             2.4666     0.1238  19.927  < 2e-16 ***
## I(Ratings^2)        0.9909     0.3036   3.264 0.002560 **
## highrate          -30.1712     7.6754  -3.931 0.000409 ***
## c                  14.1342     6.5010   2.174 0.036966 *
## USReleaseMonthFeb   4.6355     9.0163   0.514 0.610590
## USReleaseMonthJan   2.2640     8.7229   0.260 0.796830
## USReleaseMonthJul   8.9704     8.7012   1.031 0.310067
## USReleaseMonthJun  21.3142    10.4595   2.038 0.049653 *
## USReleaseMonthMar  13.0150     8.6302   1.508 0.141049
## USReleaseMonthMay   3.7660     9.6192   0.392 0.697939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.71 on 33 degrees of freedom
## Multiple R-squared:  0.9692, Adjusted R-squared:  0.9599
## F-statistic: 103.9 on 10 and 33 DF,  p-value: < 2.2e-16
```

```r
movies$jun <- ifelse(movies$USReleaseMonth == "Jun", 1, 0)
model39 = lm(USRevenue~Opening+I(Ratings^2)+highrate+c+jun, data=movies)
summary(model39)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + highrate +
##     c + jun, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.654  -8.998  -2.093   7.036  31.443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.1959    14.0026  -0.014   0.9889
## Opening        2.4798     0.1073  23.112  < 2e-16 ***
## I(Ratings^2)   0.9606     0.2910   3.301   0.0021 **
## highrate     -33.4821     7.2392  -4.625 4.25e-05 ***
## c             14.3130     6.0294   2.374   0.0228 *
## jun           14.7640     7.6671   1.926   0.0617 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.38 on 38 degrees of freedom
## Multiple R-squared:  0.9661, Adjusted R-squared:  0.9617
## F-statistic: 216.7 on 5 and 38 DF,  p-value: < 2.2e-16
```

Testing Interactions with Ratings^2

```r
model37   <- lm(USRevenue ~ Opening + I(Ratings^2)*c + highrate + jun, data = movies)
model36 <- lm(USRevenue ~ Opening + I(Ratings^2)*jun + highrate + c, data = movies)
model35_temp  <- lm(USRevenue ~ Opening + I(Ratings^2)*highrate + c + jun, data = movies)

summary(model37)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) * c + highrate +
##     jun, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.747  -9.136  -2.503   6.836  30.906
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.4531    17.7061  -0.139   0.8906
## Opening          2.4719     0.1148  21.526  < 2e-16 ***
## I(Ratings^2)     1.0163     0.3941   2.579   0.0140 *
```

```
## c               18.8545    22.1865   0.850   0.4009
## highrate       -33.4358     7.3351  -4.558 5.47e-05 ***
## jun             15.1812     8.0086   1.896   0.0658 .
## I(Ratings^2):c  -0.1233     0.5789  -0.213   0.8326
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.57 on 37 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9607
## F-statistic: 176.1 on 6 and 37 DF,  p-value: < 2.2e-16
```

```
summary(model36)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) * jun + highrate +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.007  -8.457  -1.467   9.460  29.426
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.2814    13.8631   0.092  0.92685
## Opening            2.5425     0.1149  22.137  < 2e-16 ***
## I(Ratings^2)       0.9327     0.2880   3.239  0.00254 **
## jun              115.5591    71.8295   1.609  0.11616
## highrate         -35.6856     7.3152  -4.878 2.05e-05 ***
## c                 13.9809     5.9569   2.347  0.02439 *
## I(Ratings^2):jun  -2.2141     1.5691  -1.411  0.16657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.2 on 37 degrees of freedom
## Multiple R-squared:  0.9679, Adjusted R-squared:  0.9626
## F-statistic: 185.6 on 6 and 37 DF,  p-value: < 2.2e-16
```

```
summary(model35_temp)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) * highrate +
##     c + jun, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.949  -9.034  -2.802   9.327  31.779
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -75.9649    58.2305  -1.305   0.2001
## Opening                2.4944     0.1067  23.367   <2e-16 ***
```

```
## I(Ratings^2)            2.5890     1.2492   2.073   0.0452 *
## highrate               45.0495    59.0558   0.763   0.4504
## c                      14.0752     5.9700   2.358   0.0238 *
## jun                    12.1018     7.8440   1.543   0.1314
## I(Ratings^2):highrate  -1.6979     1.2674  -1.340   0.1885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.24 on 37 degrees of freedom
## Multiple R-squared:  0.9677, Adjusted R-squared:  0.9624
## F-statistic: 184.7 on 6 and 37 DF,  p-value: < 2.2e-16
```

```
# Compare against additive baseline
model_base <- lm(USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun, data = movies)
anova(model_base, model35_temp)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun
## Model 2: USRevenue ~ Opening + I(Ratings^2) * highrate + c + jun
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 7861.4
## 2     37 7497.7  1    363.69 1.7948 0.1885
```

```
anova(model_base, model36)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun
## Model 2: USRevenue ~ Opening + I(Ratings^2) * jun + highrate + c
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 7861.4
## 2     37 7460.0  1    401.47 1.9912 0.1666
```

```
anova(model_base, model37)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun
## Model 2: USRevenue ~ Opening + I(Ratings^2) * c + highrate + jun
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 7861.4
## 2     37 7851.8  1    9.6207 0.0453 0.8326
```

Model Diagnostics

Checking for multicollinearity and ANOVA.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(model39)
```

```
##      Opening I(Ratings^2)     highrate            c          jun
##     1.551474     1.612272     1.122664     1.449733     1.259287
```

```
anova(model39)
```

```
## Analysis of Variance Table
##
## Response: USRevenue
##              Df Sum Sq Mean Sq   F value    Pr(>F)
## Opening       1 216043  216043 1044.2916 < 2.2e-16 ***
## I(Ratings^2)  1   1614    1614    7.8001 0.0081358 **
## highrate      1   3847    3847   18.5966 0.0001106 ***
## c             1   1909    1909    9.2281 0.0042935 **
## jun           1    767     767    3.7081 0.0616567 .
## Residuals    38   7861     207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model39, model40)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun
## Model 2: USRevenue ~ Opening + I(Ratings^2) + highrate + c + USReleaseMonth
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     38 7861.4
## 2     33 7142.2  5    719.19 0.6646 0.6529
```

```
model38 = lm(USRevenue~Opening+I(Ratings^2)+highrate+c, data=movies)
summary(model38)
```

```
##
## Call:
## lm(formula = USRevenue ~ Opening + I(Ratings^2) + highrate +
##     c, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.709 -10.695  -1.866   7.773  35.317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.4016    14.3034  -0.308  0.75992
## Opening        2.5488     0.1046  24.371  < 2e-16 ***
## I(Ratings^2)   1.0375     0.2981   3.480  0.00125 **
## highrate     -33.7473     7.4850  -4.509 5.82e-05 ***
## c             17.5773     5.9837   2.938  0.00553 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 39 degrees of freedom
## Multiple R-squared:  0.9628, Adjusted R-squared:  0.959
## F-statistic: 252.4 on 4 and 39 DF,  p-value: < 2.2e-16
```

```
anova(model38, model39)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c
## Model 2: USRevenue ~ Opening + I(Ratings^2) + highrate + c + jun
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     39 8628.6
## 2     38 7861.4  1    767.13 3.7081 0.06166 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final Model Development

Incorporating feedback: quadratic for Theaters, interactions.

```
# Best model incorporating Theaters quadratic and interactions
model35 <- lm(USRevenue ~ Theaters + I(Theaters^2) + Opening + Ratings + I(Ratings^2) + highrate + c + I
summary(model35)
```

```
##
## Call:
## lm(formula = USRevenue ~ Theaters + I(Theaters^2) + Opening +
##     Ratings + I(Ratings^2) + highrate + c + I(Theaters^2) * Opening +
##     jun + jun * Opening, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5442  -6.5471   0.1162   4.9479  25.8741
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.107e+02  1.093e+02   1.013 0.318659
## Theaters               -4.254e-02  7.228e-02  -0.589 0.560152
## I(Theaters^2)           9.175e-06  1.289e-05   0.712 0.481652
## Opening                 3.377e+00  7.737e-01   4.364 0.000118 ***
## Ratings                -2.453e+01  2.315e+01  -1.060 0.296874
## I(Ratings^2)            2.914e+00  1.965e+00   1.483 0.147661
## highrate               -2.950e+01  8.972e+00  -3.289 0.002398 **
## c                       1.252e+01  6.034e+00   2.075 0.045878 *
## jun                     3.019e+01  1.461e+01   2.066 0.046781 *
## I(Theaters^2):Opening -6.845e-08  5.918e-08  -1.157 0.255723
## Opening:jun           -2.074e-01  2.917e-01  -0.711 0.482091
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.42 on 33 degrees of freedom
## Multiple R-squared:  0.9744, Adjusted R-squared:  0.9666
## F-statistic: 125.5 on 10 and 33 DF,  p-value: < 2.2e-16
```

```
anova(model38, model35)
```

```
## Analysis of Variance Table
##
## Model 1: USRevenue ~ Opening + I(Ratings^2) + highrate + c
## Model 2: USRevenue ~ Theaters + I(Theaters^2) + Opening + Ratings + I(Ratings^2) +
##     highrate + c + I(Theaters^2) * Opening + jun + jun * Opening
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     39 8628.6
## 2     33 5947.2  6    2681.4 2.4798 0.04322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

Through iterative model building, starting from a full model, we addressed multicollinearity, tested transformations, added significant variables, and explored interactions. The final model (model35) includes quadratic terms for Theaters and Ratings, interactions, and key indicators, achieving a high adjusted R-squared.