Movie Dataset Analysis Project

2025-11-08

Explanation of Objective and Dataset

The objective of this analysis is to explore the Movie.csv dataset and build predictive models for movie success metrics, such as worldwide revenue or ratings, based on various movie attributes. The dataset contains information about movies released in 2013, including financial data, ratings, and other characteristics.

The dataset includes the following variables: - Title: Movie title - USRelease: Release date in US - Genre: Movie genre - Rating: MPAA rating (PG-13, R, etc.) - Sequel: Whether it's a sequel (0/1) - Budget: Production budget in millions - Opening: Opening weekend revenue in millions - USRevenue: US box office revenue in millions - Theaters: Number of theaters - IntRevenue: International revenue in millions – exclude this one - WorldRevenue: Worldwide revenue in millions - Ratings: Average rating (out of 10) - Review: Percentage of positive reviews - Minutes: Movie runtime in minutes

```
# 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
```

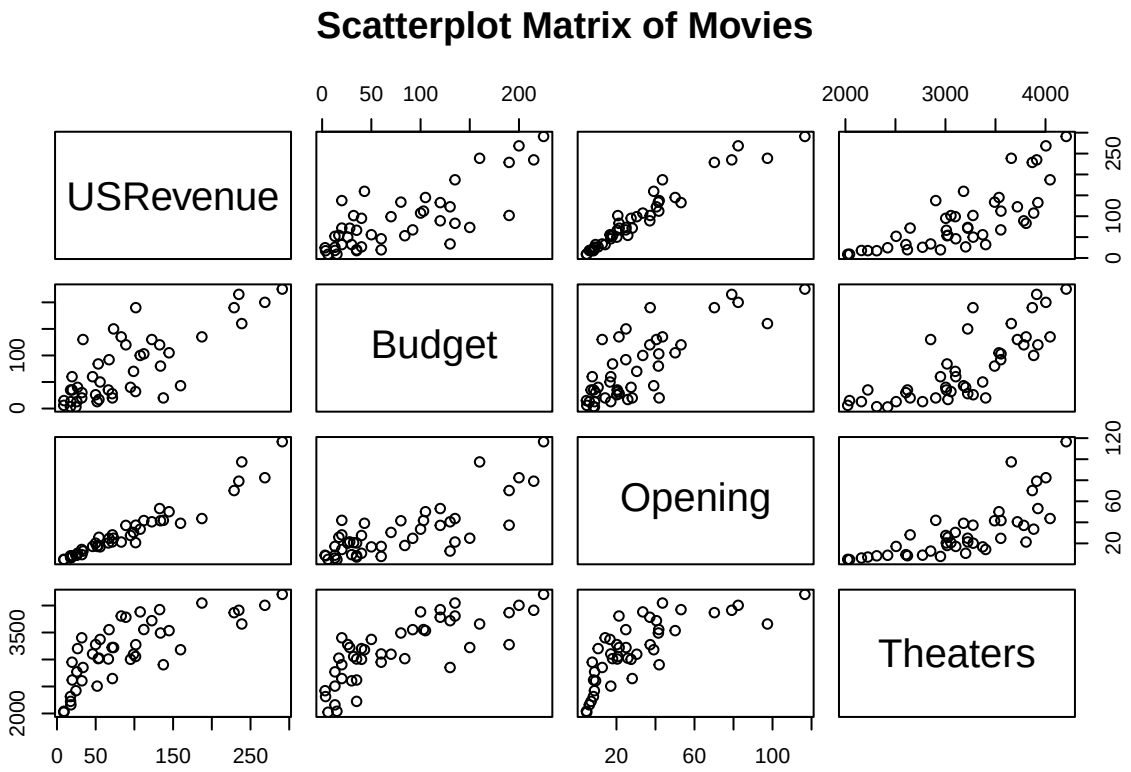
EDA (Exploratory Data Analysis)

Let's start by examining the structure and basic statistics of the dataset.

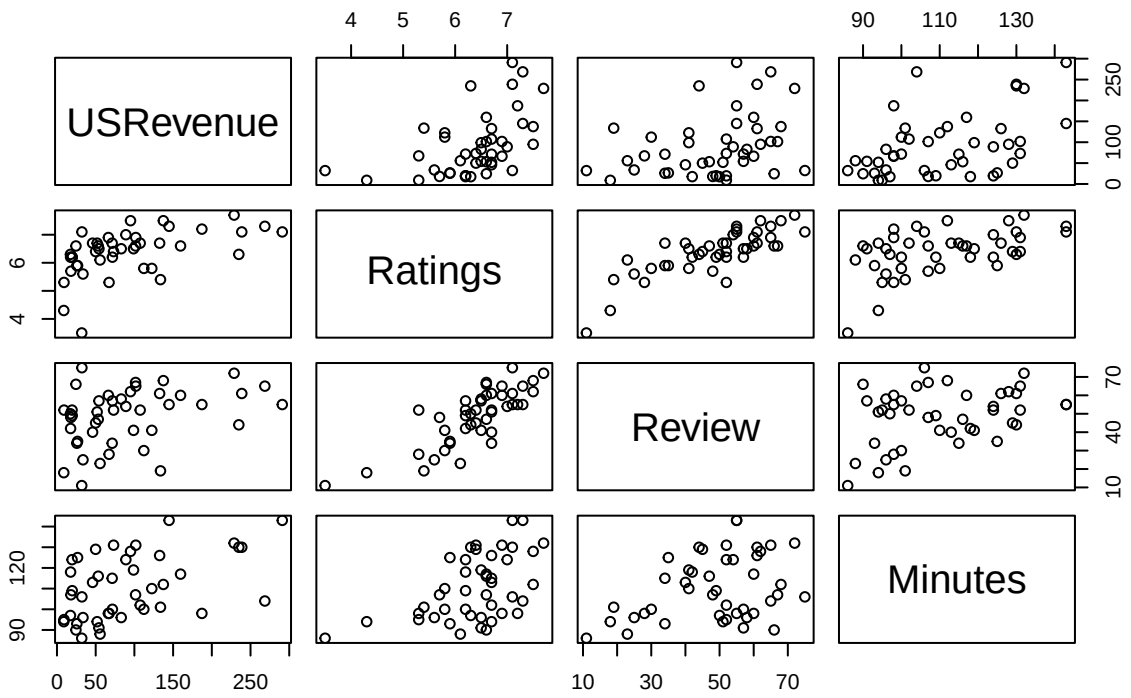
```
str(movies)
```

```
## 'data.frame':    45 obs. of  14 variables:
##  $ Title       : chr  "Man of Steel" "Monster University" "Fast & Furious 6" "Oz the Great and Power
##  $ USRelease   : chr  "16-Jun" "23-Jun" "26-May" "10-Mar" ...
##  $ Genre       : chr  "Action/Adventure" "Animation" "Action/Adventure" "Action/Adventure" ...
##  $ Rating      : chr  "PG-13" "G" "PG-13" "PG" ...
##  $ Sequel      : int  0 0 1 0 1 0 0 0 0 1 ...
##  $ Budget      : num  225 200 160 215 190 135 43 105 20 80 ...
##  $ Opening     : num  116.6 82.4 97.4 79.1 70.2 ...
##  $ USRevenue   : num  291 268 239 235 229 ...
##  $ Theaters    : int  4207 4004 3658 3912 3868 4046 3181 3535 2903 3491 ...
##  $ IntRevenue  : num  377 475 550 258 239 ...
##  $ WorldRevenue: num  668 744 789 493 467 ...
##  $ Ratings     : num  7.1 7.3 7.1 6.3 7.7 7.2 6.6 7.3 7.5 5.4 ...
##  $ Review      : int  55 65 61 44 72 55 60 55 68 19 ...
##  $ Minutes     : int  143 104 130 130 132 98 117 143 112 101 ...
```

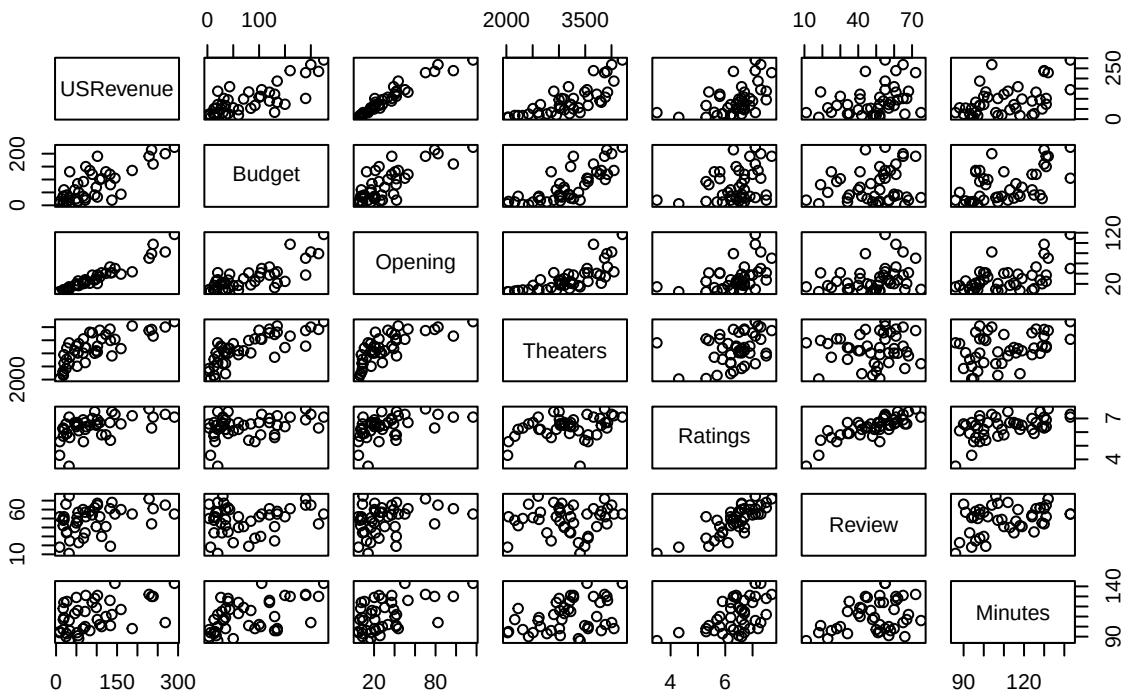
```
pairs(~USRevenue+Budget+Opening+Theaters,data = movies,
      main="Scatterplot Matrix of Movies")
```




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




```
pairs(~USRevenue+Budget+Opening+Theaters+Ratings+Review+Minutes,data = movies,
      main="Scatterplot Matrix of Movies Complete") #to see multi-collinearity
```




```
library(PerformanceAnalytics)
```

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

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

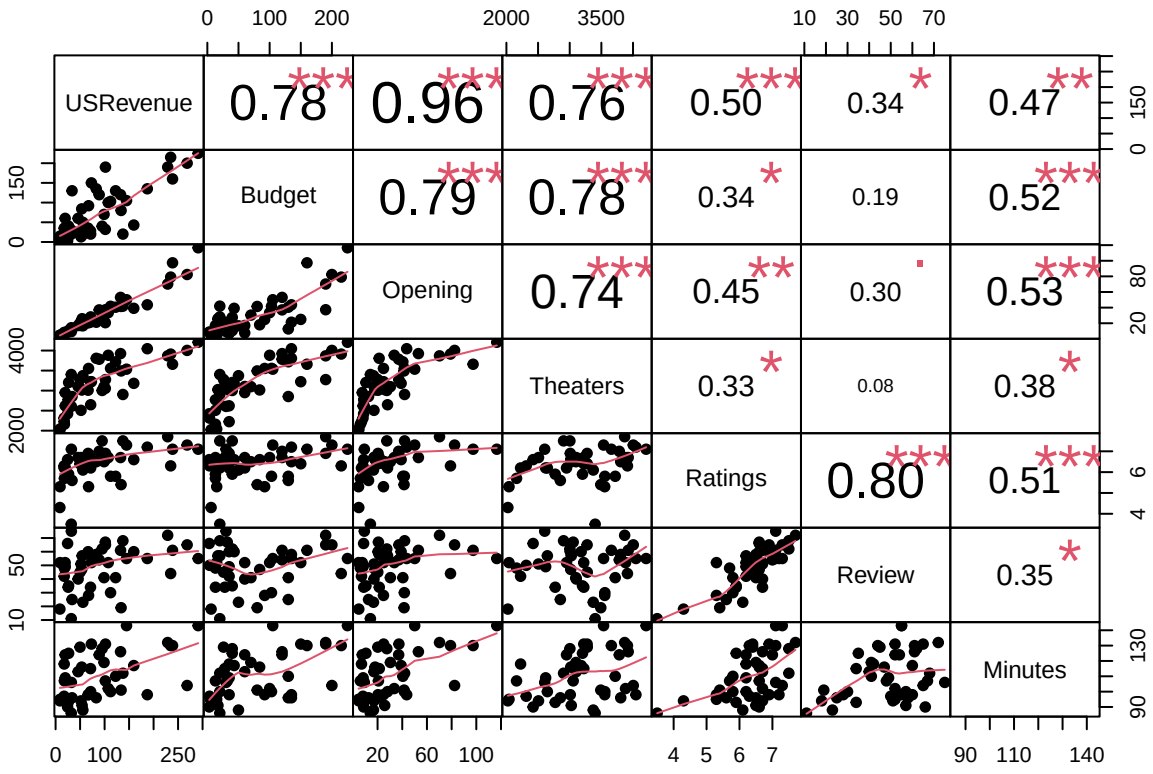
```
##
## Attaching package: 'zoo'
```

```
## 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")]
```

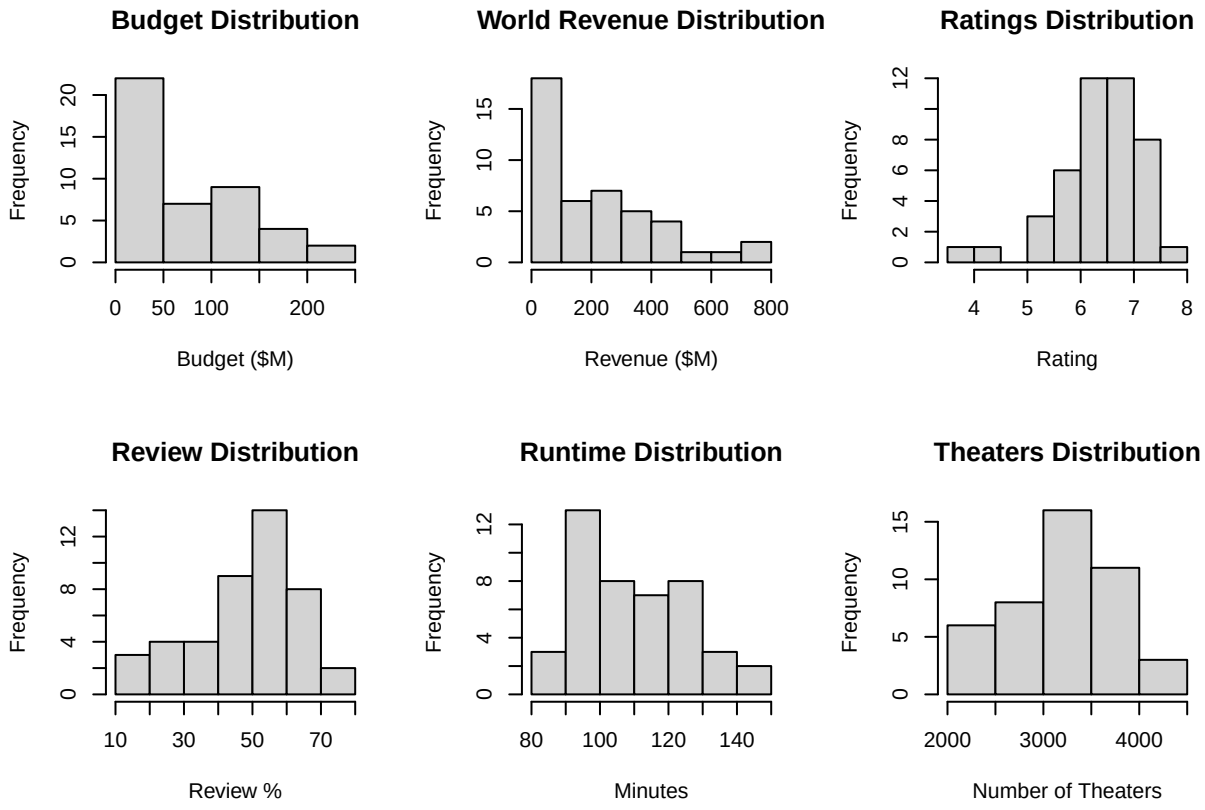


Lets look at variables we think are correlated (multi-collinearity)

```
#Reviews and Ratings (R is 0.80, and reviews and ratings are very similar concepts)
```

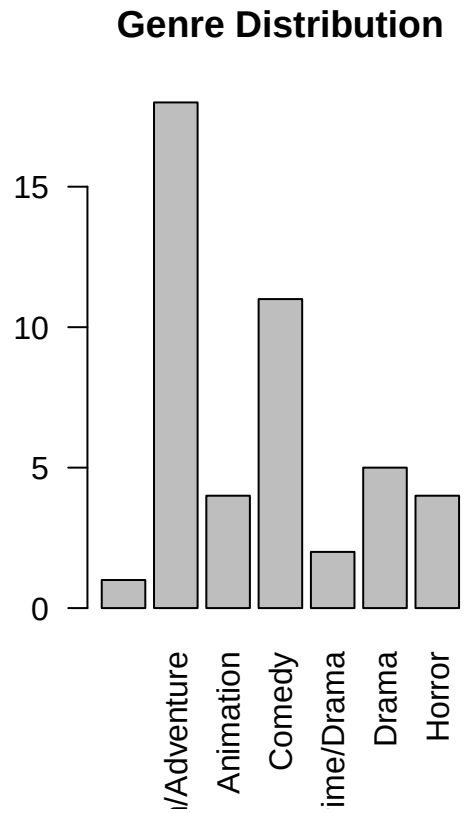
Now, let's look at the distribution of key variables.

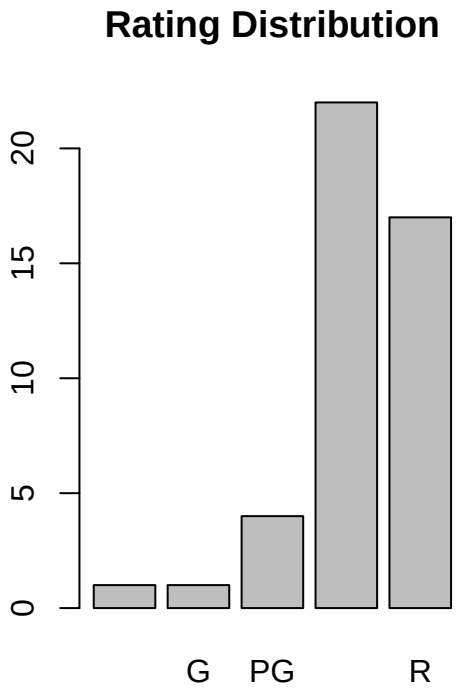
```
# Histograms of numerical variables
par(mfrow=c(2,3))
hist(movies$Budget, main="Budget Distribution", xlab="Budget ($M)")
hist(movies$WorldRevenue, main="World Revenue Distribution", xlab="Revenue ($M)")
hist(movies$Ratings, main="Ratings Distribution", xlab="Rating")
hist(movies$Review, main="Review Distribution", xlab="Review %")
hist(movies$Minutes, main="Runtime Distribution", xlab="Minutes")
hist(movies$Theaters, main="Theaters Distribution", xlab="Number of Theaters")
```



Let's examine the categorical variables.

```
# Bar plots for categorical variables
par(mfrow=c(1,2))
barplot(table(movies$Genre), main="Genre Distribution", las=2)
barplot(table(movies$Rating), main="Rating Distribution")
```



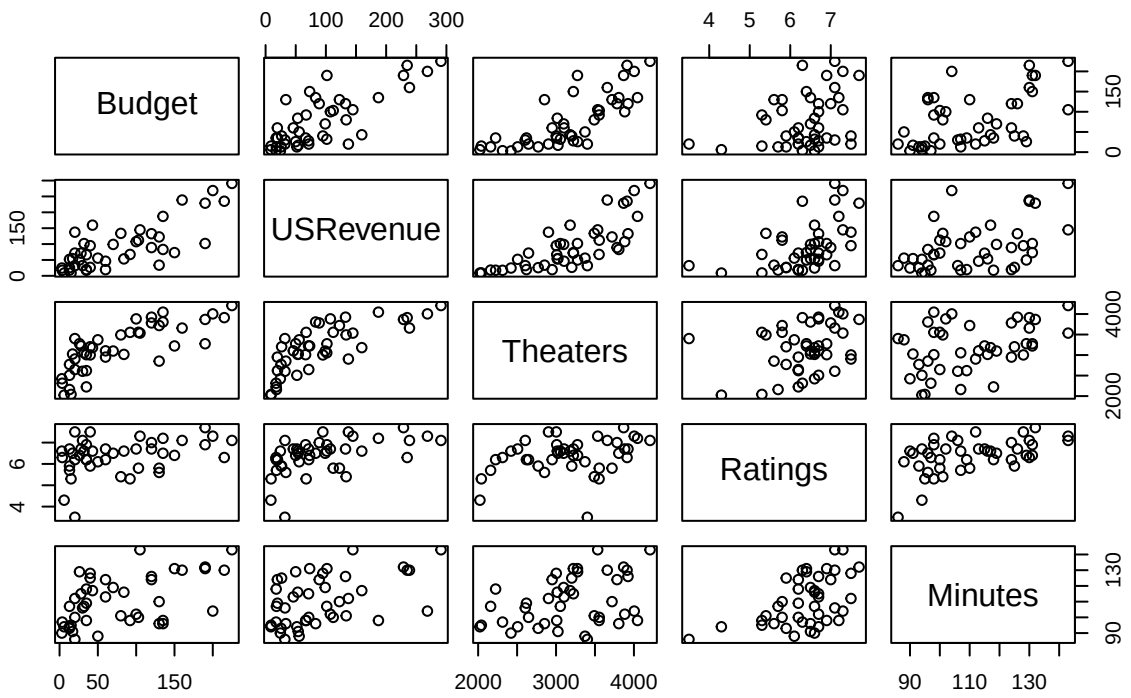


Testing/Models

Scatterplots and Visual Representations of Data

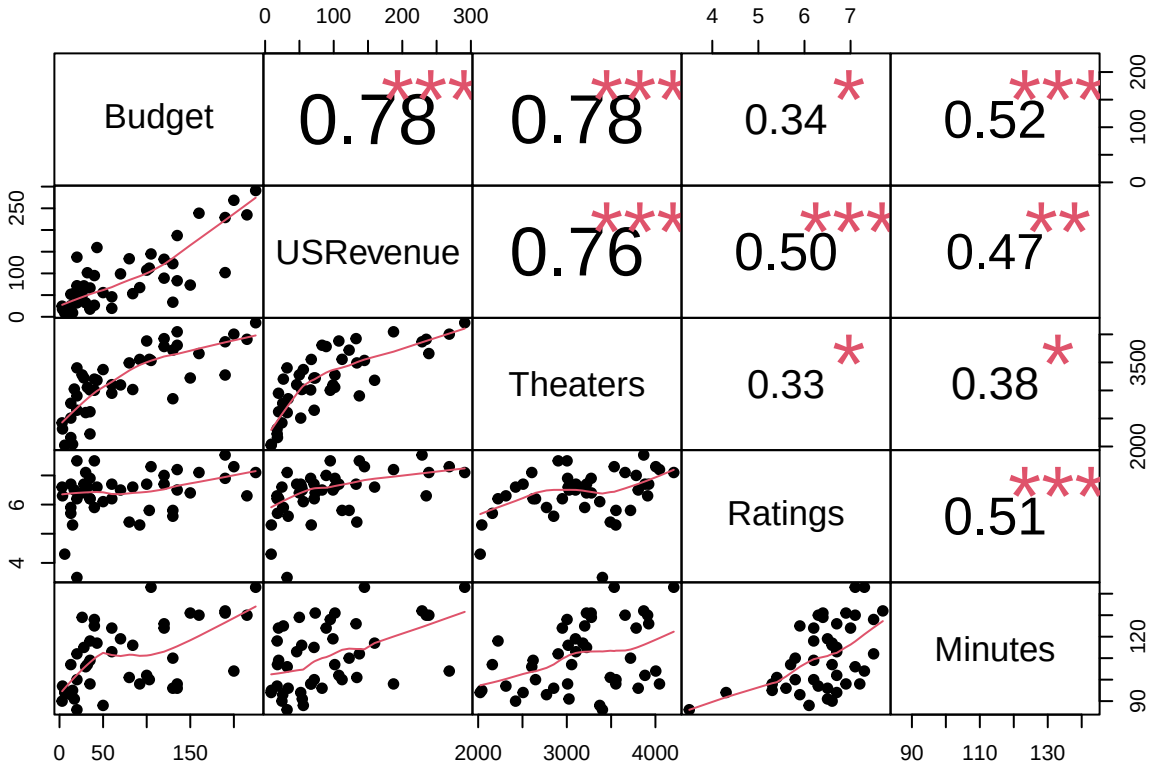
Let's create scatterplot matrices to visualize relationships between variables.

```
# Select numerical variables for correlation analysis
num_vars <- movies[,c("Budget", "USRevenue", "Theaters", "Ratings", "Minutes")]
pairs(num_vars, main="Scatterplot Matrix of Numerical Variables")
```



Let's also look at correlations.

```
library(PerformanceAnalytics)
chart.Correlation(num_vars, histogram=FALSE, pch=19)
```



Deciding Which Variables Should Be Included

Based on the scatterplots and correlations, we can see strong relationships between: - Budget and revenues (Opening, USRevenue, WorldRevenue) - Theaters and revenues - Ratings and Review percentage

For modeling WorldRevenue, we'll consider Budget, Theaters, Ratings, Review, Minutes, and Sequel as predictors.

Let's build several models and compare them.

```
# Model 1: Simple linear model with Budget
model1 <- lm(WorldRevenue ~ Budget, data=movies)
summary(model1)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -286.72  -58.45   -1.73   43.13  343.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4751    26.2562   0.742    0.462
## Budget        2.6580     0.2629  10.109 8.11e-13 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110.7 on 42 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7087, Adjusted R-squared:  0.7018
## F-statistic: 102.2 on 1 and 42 DF,  p-value: 8.111e-13
```

```
# Model 2: Add Theaters
model2 <- lm(WorldRevenue ~ Budget + Theaters, data=movies)
summary(model2)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget + Theaters, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -200.16  -62.98   -9.09   38.78  355.14
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -312.74958  115.19392  -2.715  0.00965 **
## Budget         1.77889    0.38375   4.636 3.59e-05 ***
## Theaters       0.12621    0.04279   2.950  0.00524 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.8 on 41 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7597, Adjusted R-squared:  0.748
## F-statistic: 64.81 on 2 and 41 DF,  p-value: 2.019e-13
```

```
# Model 3: Add Ratings and Review
model3 <- lm(WorldRevenue ~ Budget + Theaters + Ratings + Review, data=movies)
summary(model3)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget + Theaters + Ratings + Review,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.03  -63.67   -4.71   37.60  338.64
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -439.54706  167.89267  -2.618 0.012526 *
## Budget         1.62951    0.38391   4.245 0.000131 ***
## Theaters       0.13325    0.04434   3.005 0.004622 **
## Ratings        4.69775   34.98425   0.134 0.893870
## Review         1.77239    1.77186   1.000 0.323331
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99.67 on 39 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7808, Adjusted R-squared:  0.7583
## F-statistic: 34.73 on 4 and 39 DF,  p-value: 2.27e-12
```

```
# Model 4: Add Minutes and Sequel
model4 <- lm(WorldRevenue ~ Budget + Theaters + Ratings + Review + Minutes + Sequel, data=movies)
summary(model4)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget + Theaters + Ratings + Review +
##     Minutes + Sequel, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.07  -56.49  -10.04   40.04  294.16
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -385.48509  181.19140  -2.128   0.0401 *
## Budget         1.80724    0.40167   4.499 6.54e-05 ***
## Theaters       0.10095    0.04634   2.178   0.0358 *
## Ratings       37.35940   37.81682   0.988   0.3296
## Review         1.35979    1.73885   0.782   0.4392
## Minutes       -1.52285    1.22962  -1.238   0.2233
## Sequel        71.36801   43.04812   1.658   0.1058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.12 on 37 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8026, Adjusted R-squared:  0.7705
## F-statistic: 25.07 on 6 and 37 DF,  p-value: 1.22e-11
```

```
# Model 5: Interaction between Budget and Theaters
model5 <- lm(WorldRevenue ~ Budget * Theaters + Ratings + Review + Minutes + Sequel, data=movies)
summary(model5)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget * Theaters + Ratings + Review +
##     Minutes + Sequel, data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.13  -42.36  -20.01   22.31  293.13
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     -2.334e+02  1.697e+02  -1.375  0.17759
## Budget          -3.441e+00  1.708e+00  -2.015  0.05145 .
## Theaters         2.227e-02  4.856e-02   0.459  0.64925
## Ratings          5.567e+01  3.446e+01   1.616  0.11491
## Review           3.496e-02  1.617e+00   0.022  0.98287
## Minutes         -1.065e+00  1.114e+00  -0.957  0.34512
## Sequel           9.106e+01  3.916e+01   2.325  0.02581 *
## Budget:Theaters  1.449e-03  4.610e-04   3.144  0.00334 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.22 on 36 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.815
## F-statistic: 28.06 on 7 and 36 DF,  p-value: 9.249e-13
```

```
# test model - to teach Brian
```

Let's use stepwise regression to find the best subset of variables.

```
library(MASS)
# Stepwise regression
step_model <- stepAIC(lm(WorldRevenue ~ Budget + Theaters + Ratings + Review + Minutes + Sequel, data=mo
```

```
## Start:  AIC=409.06
## WorldRevenue ~ Budget + Theaters + Ratings + Review + Minutes +
##     Sequel
##
##            Df Sum of Sq    RSS    AIC
## - Review    1      5769 354789 407.78
## - Ratings   1      9206 358226 408.21
## - Minutes   1     14468 363489 408.85
## <none>                  349020 409.06
## - Sequel    1     25927 374947 410.22
## - Theaters  1     44765 393785 412.37
## - Budget    1    190962 539982 426.26
##
## Step:  AIC=407.78
## WorldRevenue ~ Budget + Theaters + Ratings + Minutes + Sequel
##
##            Df Sum of Sq    RSS    AIC
## - Minutes   1     16203 370992 407.75
## <none>                  354789 407.78
## + Review    1      5769 349020 409.06
## - Sequel    1     28296 383085 409.16
## - Theaters  1     39043 393831 410.38
## - Ratings   1     62809 417598 412.96
## - Budget    1    206709 561498 425.98
##
## Step:  AIC=407.75
## WorldRevenue ~ Budget + Theaters + Ratings + Sequel
##
##            Df Sum of Sq    RSS    AIC
```

```
## <none>                  370992 407.75
## + Minutes   1     16203 354789 407.78
## - Sequel    1     26409 397401 408.77
## + Review    1      7503 363489 408.85
## - Theaters  1     44989 415981 410.78
## - Ratings   1     47478 418470 411.05
## - Budget    1    191887 562879 424.09
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = WorldRevenue ~ Budget + Theaters + Ratings + Sequel,
##     data = movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.98  -53.57  -10.80   32.76  287.61
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -528.30119  155.78145  -3.391  0.00161 **
## Budget         1.67152    0.37217   4.491 6.14e-05 ***
## Theaters       0.09505    0.04371   2.175  0.03578 *
## Ratings       48.25010   21.59750   2.234  0.03128 *
## Sequel        71.68224   43.02153   1.666  0.10369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.53 on 39 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7901, Adjusted R-squared:  0.7686
## F-statistic: 36.71 on 4 and 39 DF,  p-value: 9.842e-13
```

Interpretations

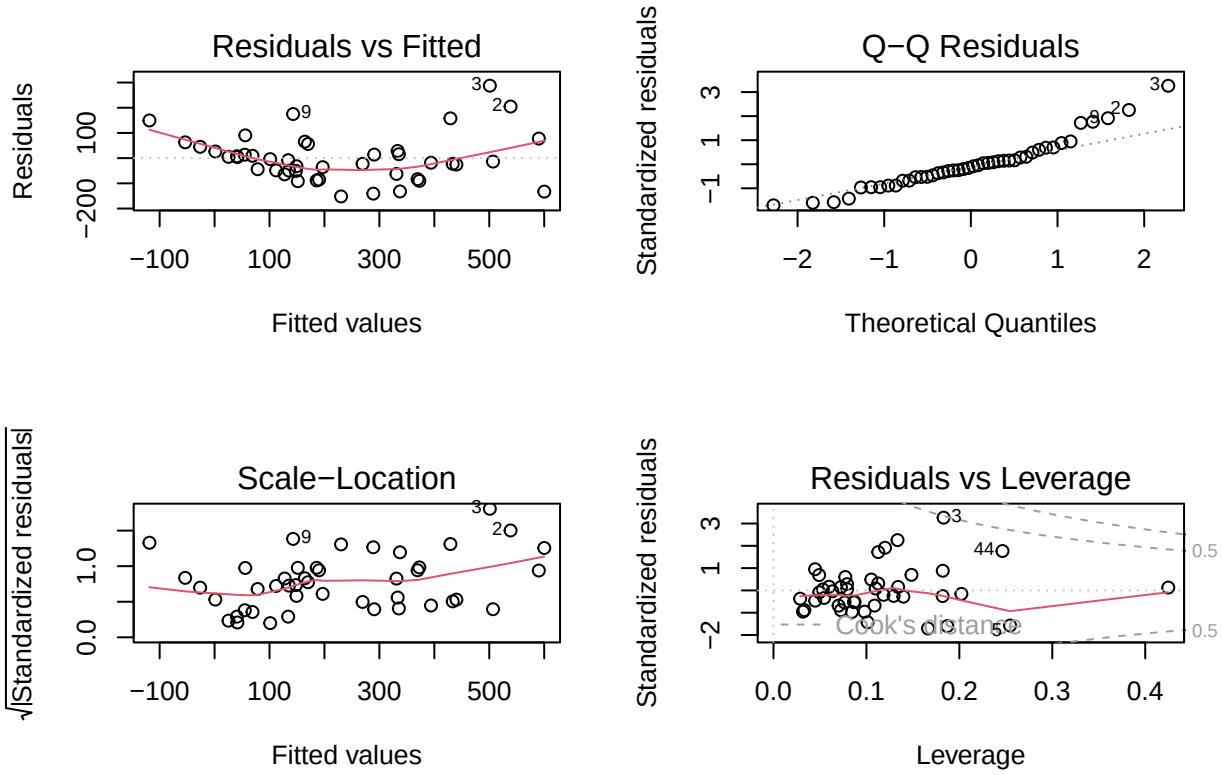
From the models above, we can see that:

- Budget has a strong positive relationship with WorldRevenue
- Number of Theaters also positively impacts revenue
- Higher Ratings and Review percentages tend to correlate with higher revenue
- Sequels may have an advantage in revenue generation

The stepwise model suggests that all variables except Minutes are significant predictors.

Let's check model diagnostics for the best model.

```
# Residual plots for the stepwise model
par(mfrow=c(2,2))
plot(step_model)
```



Conclusion

The analysis shows that movie revenue can be reasonably predicted using budget, number of theaters, ratings, review percentage, and whether it's a sequel. The models explain a significant portion of the variance in worldwide revenue.

The Model We Liked the Best

The stepwise regression model (step_model) is our preferred model as it automatically selected the most significant predictors and achieved good R-squared values with parsimony.

Room and Area for Improvement

Improvements in Our Work:

- Could try polynomial terms or transformations for non-linear relationships
- Consider categorical variables like Genre and Rating in more detail (dummy coding)
- Use cross-validation to assess model stability
- Try other modeling techniques like random forests or neural networks

Improvements for the Dataset:

- More recent data (2013 is quite old for movie analysis)

- Additional variables like director, actors, marketing budget
- Social media metrics, pre-release buzz
- More detailed genre classifications
- International market breakdowns
- Streaming revenue data (though not relevant for 2013)