Model for Interpretation and Predictive Model for Movies dataset

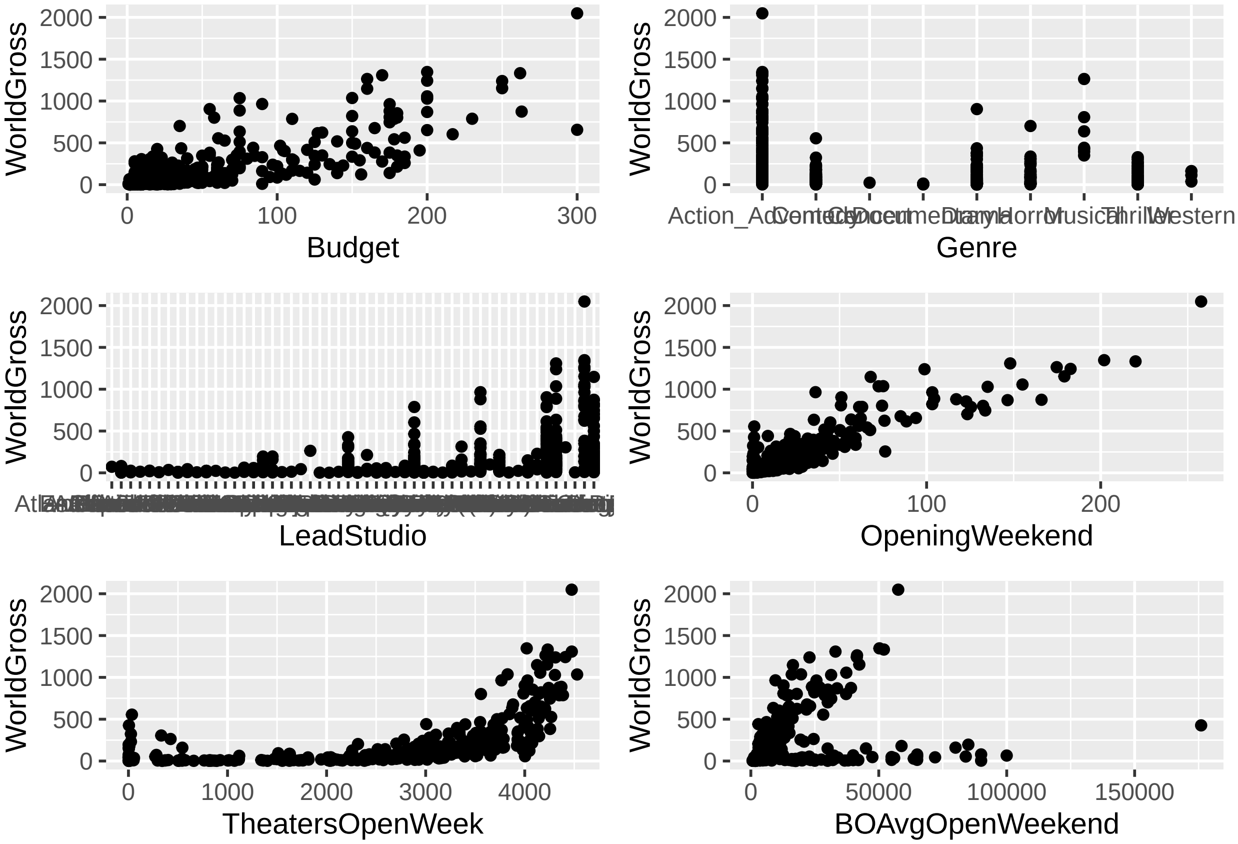
Carol Weng

Because there are many categories for genre, Romance Comedy, Comedy, and Black Comedy are combined into Comedy. Action and Adventure were combined to Action_Adventure.

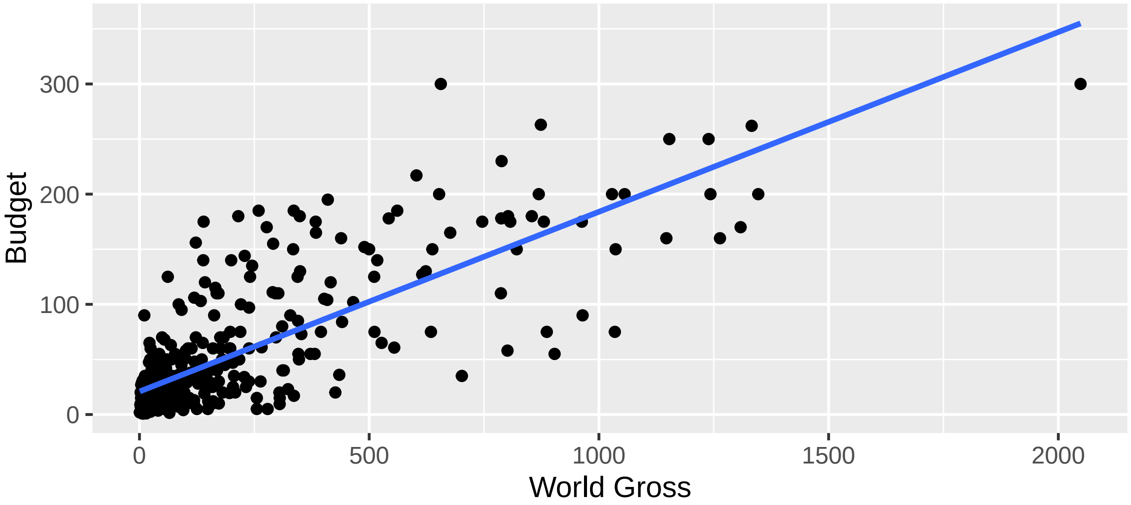
Part 3 - Multivariate Modeling

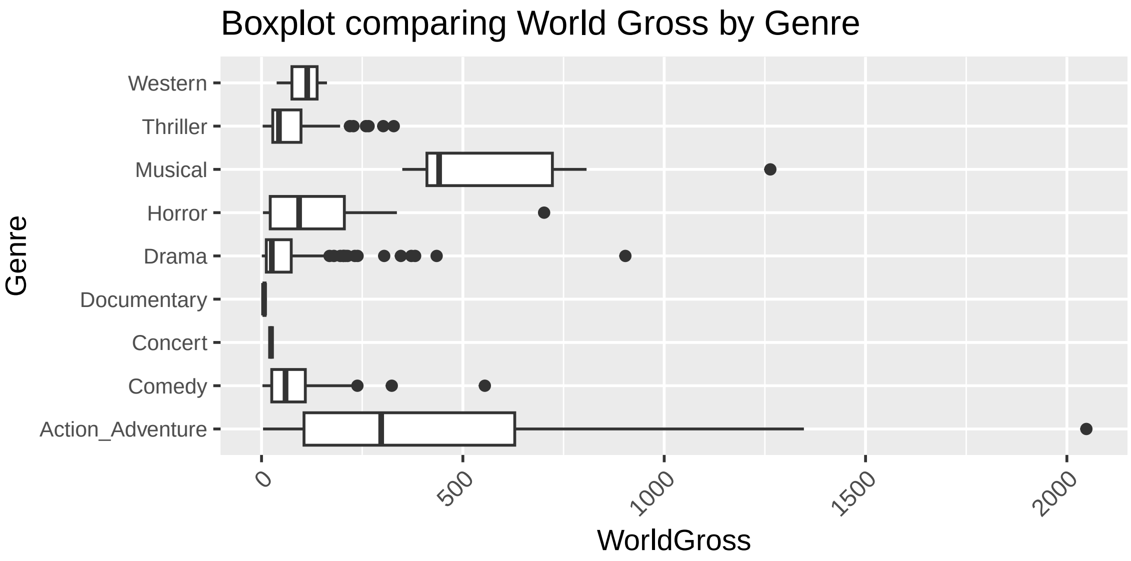
Research Questions: How does production spending impact profit of a movie?

```
##                  RottenTomatoes AudienceScore TheatersOpenWeek OpeningWeekend
## RottenTomatoes             1.00          0.71            -0.22           0.17
## AudienceScore              0.71          1.00            -0.20           0.20
## TheatersOpenWeek          -0.22         -0.20             1.00           0.58
## OpeningWeekend             0.17          0.20             0.58           1.00
## BOAvgOpenWeekend           0.43          0.35            -0.32           0.24
## Budget                     0.05          0.10             0.61           0.74
## DomesticGross              0.26          0.28             0.54           0.95
## WorldGross                 0.21          0.26             0.55           0.91
## ForeignGross               0.18          0.23             0.52           0.84
## Year                       0.05          0.03             0.09           0.09
##                  BOAvgOpenWeekend Budget DomesticGross WorldGross ForeignGross
## RottenTomatoes               0.43   0.05          0.26       0.21         0.18
## AudienceScore                0.35   0.10          0.28       0.26         0.23
## TheatersOpenWeek            -0.32   0.61          0.54       0.55         0.52
## OpeningWeekend               0.24   0.74          0.95       0.91         0.84
## BOAvgOpenWeekend             1.00   0.11          0.30       0.28         0.26
## Budget                       0.11   1.00          0.70       0.79         0.79
## DomesticGross                0.30   0.70          1.00       0.94         0.85
## WorldGross                   0.28   0.79          0.94       1.00         0.98
## ForeignGross                 0.26   0.79          0.85       0.98         1.00
## Year                         0.01   0.01          0.11       0.10         0.09
##                  Year
## RottenTomatoes   0.05
## AudienceScore    0.03
## TheatersOpenWeek 0.09
## OpeningWeekend   0.09
## BOAvgOpenWeekend 0.01
## Budget           0.01
## DomesticGross    0.11
## WorldGross       0.10
## ForeignGross     0.09
## Year             1.00
```



Relationship between Budget and WorldGross





One paragraph discussion (5-7 sentences in length) explaining how you decided which variable(s) you chose your model. Justify why you chose to include the variables you did. Also, if there were any variables you considered adding to the model but didn't, explain why you chose not to. Explain why you did or didn't use any transformations.

Research Questions: How does production spending impact profit of a movie?

WorldGross will be the response variable exploring this research question because it is the total profit earned around the world for the movies.

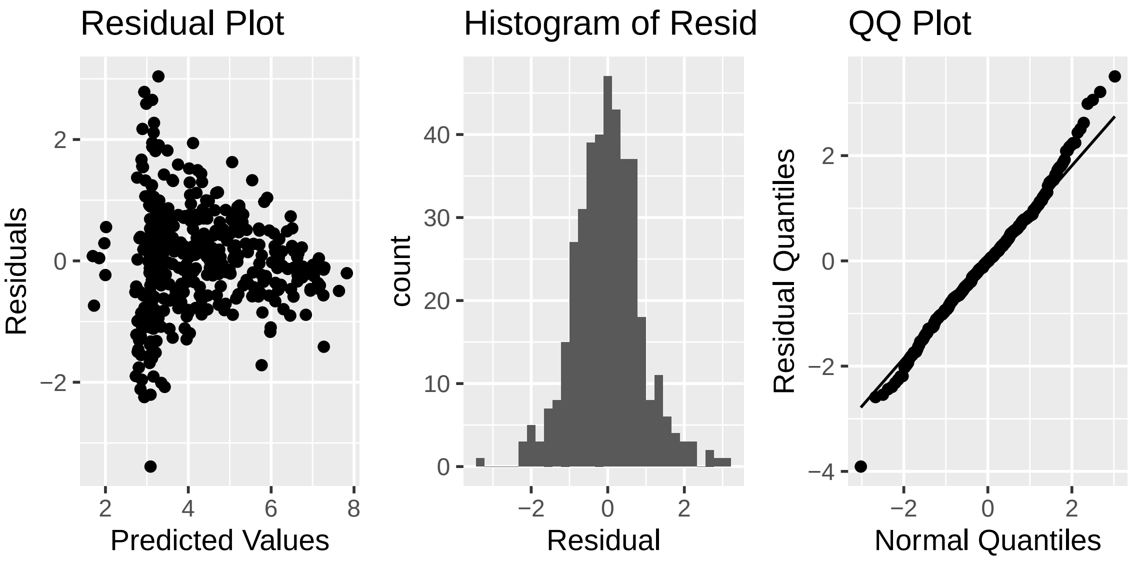
Correlation table: We can see that Foreign Gross, and Domestic Gross are both strongly correlated to World Gross the number was 0.98 and 0.94, this makes sense because world gross include foreign gross and domestic gross. But because Foreign and domestic gross are not production cost, they are not included in the model.

Budget is included in the model because it is a big part of production spending. And we can see in the scatterplot that the dots showed a linear positive trend compared with world gross.

TheatersOpenWeek is the number of screens for opening weekend. We can see in the correlation table that it is highly correlated to World Gross with 0.55, so it was added in the model. This could be related to the production spending because we are not sure whether budget includes advertisement fees or not, usually more numbers of screens for opening weekend means more advertisement fees so it is included in the model. Looking at the residual plot for TheatersOpenWeek, the graph shows a curve, so a quadratic term of TheatersOpenWeek was added to the graph.

Genre is included in the model because the boxplot shows that some type of movies seems to have higher world gross in general. It can be included in production spending because some genre might spend more in production in general.

I was considering including Lead Studio because some studio might be more famous more budget, but decided not to include because there are too many Lead Studio which made it hard to interpret and see the correlation for each of them.



A one paragraph discussion of how well you think the model assumptions are satisfied. If you still have concerns about any model assumptions, explain why

I added log for the response variable WorldGross because the residual graph showed right skewness and severed departure for the qq plot. After adding the log, the histogram seems to be roughly symmetric, QQ plot shows less departure from diagonal line but still some at the end. Some problem with normality assumption but not a lot.

But there is violation of constant variance assumption because there seems to be bigger range on the left than right. This might because of the categorical variable genre because the residual plot showed previously of genre showed different range.

#Confidence intervals for each of your model coefficients.

```
##
## Call:
## lm(formula = log(WorldGross) ~ Budget + Genre + TheatersOpenWeek +
##     I(TheatersOpenWeek^2), data = Movies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3897 -0.5524 -0.0048  0.5175  3.0398
##
## Coefficients:
##                            Estimate     Std. Error t value
## (Intercept)           3.07641326831  0.17171441217  17.916
## Budget                0.00418640014  0.00132907014   3.150
## GenreComedy          -0.03008939894  0.15346874536  -0.196
## GenreConcert         -0.40100167135  0.88953350915  -0.451
## GenreDocumentary     -1.07982506473  0.38899132843  -2.776
## GenreDrama            0.00678875876  0.15227371193   0.045
## GenreHorror           0.25313959578  0.20092962034   1.260
## GenreMusical          0.95702533990  0.34241886539   2.795
## GenreThriller        -0.02206763802  0.17132682751  -0.129
## GenreWestern           0.01986571890  0.51755216238   0.038
## TheatersOpenWeek      -0.00064623237  0.00012734372  -5.075
## I(TheatersOpenWeek^2)  0.00000031906  0.00000003467   9.204
##                                   Pr(>|t|)
## (Intercept)           < 0.0000000000000002 ***
## Budget                             0.00176 **
## GenreComedy                        0.84466
## GenreConcert                       0.65239
## GenreDocumentary                   0.00577 **
## GenreDrama                         0.96446
## GenreHorror                        0.20848
## GenreMusical                       0.00545 **
## GenreThriller                      0.89758
## GenreWestern                       0.96940
## TheatersOpenWeek               0.000000603 ***
## I(TheatersOpenWeek^2) < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8795 on 388 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6794
## F-statistic: 77.88 on 11 and 388 DF,  p-value: < 0.00000000000000022
```

```
##                            2.5 %     97.5 %
## (Intercept)           15.4685062 30.3871687
## Budget                 1.0015746  1.0068226
## GenreComedy            0.7176137  1.3121211
## GenreConcert           0.1164946  3.8493619
## GenreDocumentary       0.1580854  0.7297667
## GenreDrama             0.7463235  1.3582181
## GenreHorror            0.8677015  1.9120705
## GenreMusical           1.3281629  5.1051713
## GenreThriller          0.6984353  1.3699545
## GenreWestern           0.3687295  2.8219367
## TheatersOpenWeek       0.9991038  0.9996042
## I(TheatersOpenWeek^2)  1.0000003  1.0000004
```

A two to three paragraph discussion explaining the most interesting conclusions from your model. In your discussion, include interpretations of at least two model coefficients, their confidence intervals, and their associated p-values in context. Discuss the implications of your findings. Was there anything surprising or unusual? Note any limitations or concerns you might have about your conclusions.

Interpretation:

Budget: On average, for each additional million dollars in budget, the world gross is expected to increase by 0.004 million dollars. We are 95% confident that the world gross of a movie, on average, increases between 0.16% and 0.68% for each additional million dollars in budget. The p-value for Budget 0.00251 represent that there is strong evidence that there is a significant relationship between world gross and budget.

TheatersOpenWeek: On average, for each additional screen for opening weekends, the world gross is expected to decrease by 0.0007 million dollars. We are 95% confident that the world gross of a movie, on average, decreases between 0.09% and 0.04% for each additional screen for opening weekends. The p-value for TheatersOpenWeek is < 0.0001, which means that there is a strong evidence that there is a significant relationship between world gross and number of screens for opening weekends.

Budget, TheatersOpenWeek, I(TheatersOpenWeekˆ2), genre_documentary seems to have significant relationship with world gross. It was surprising that only documentary seems to have an effect on world gross. Genre in general seems not to be a good predictor of world gross/not related to world gross. The limitation and concern is that the model shows a violation of constant variance.

Calculate and interpret a confidence interval for an expected response and also a prediction interval. Choose values/categories of the explanatory variable(s) that are of interest.

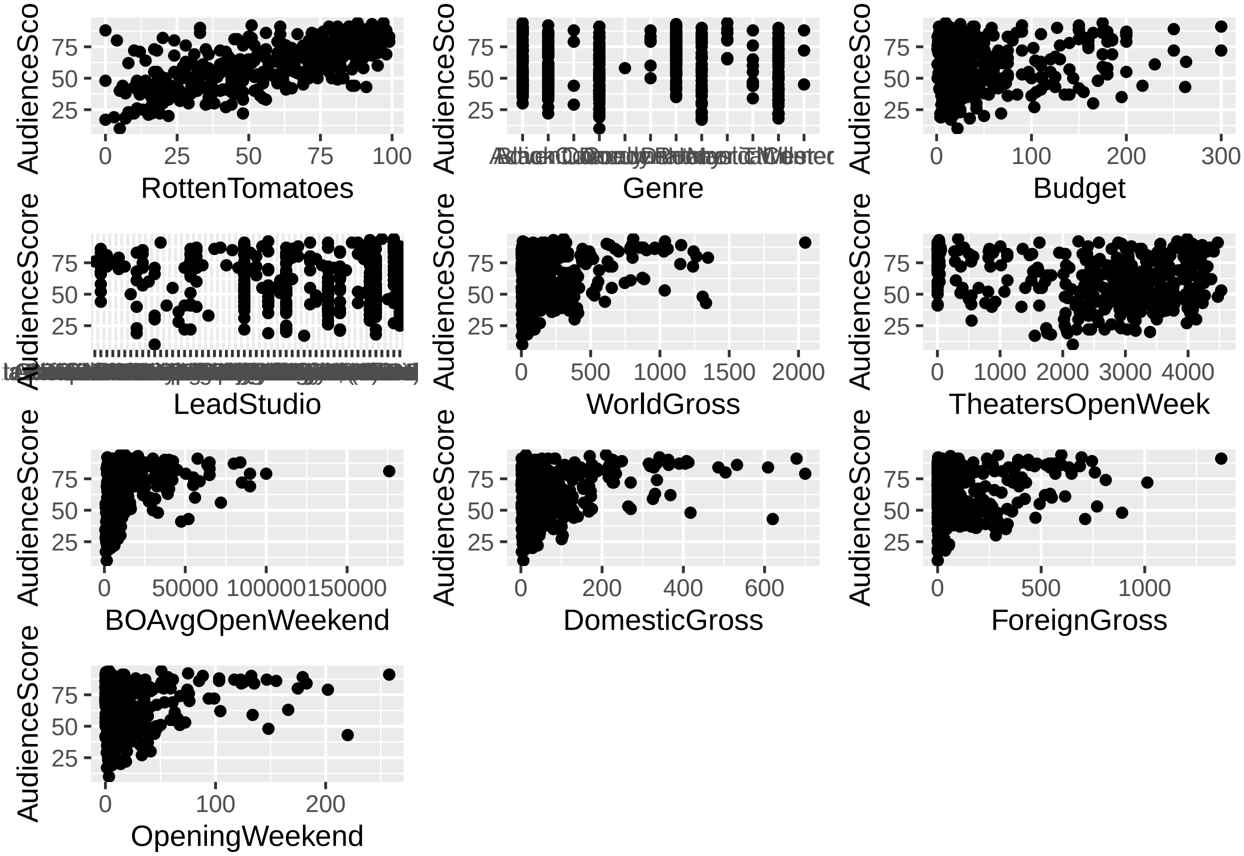
```
##        fit      lwr      upr
## 1 28.44716 13.42793 60.26552
```

```
##        fit      lwr      upr
## 1 28.44716 4.318456 187.3913
```

Interpretation: Confidence Interval: We are 95% confident that the average world gross of all documentary movies with 100 million dollars budget and 3000 screens for opening weekends is between 13.4 and 60.2 million dollars. Prediction Interval: We are 95% confident that the world gross of an individual documentary movie with 100 million dollars budget and 3000 screens for opening weekends is between 4.3 and 187.3 million dollars.

Part 4 - Predictive Modeling

```
##                  RottenTomatoes AudienceScore TheatersOpenWeek OpeningWeekend
## RottenTomatoes             1.00          0.71            -0.22           0.17
## AudienceScore              0.71          1.00            -0.20           0.20
## TheatersOpenWeek          -0.22         -0.20             1.00           0.58
## OpeningWeekend             0.17          0.20             0.58           1.00
## BOAvgOpenWeekend           0.43          0.35            -0.32           0.24
## Budget                     0.05          0.10             0.61           0.74
## DomesticGross              0.26          0.28             0.54           0.95
## WorldGross                 0.21          0.26             0.55           0.91
## ForeignGross               0.18          0.23             0.52           0.84
## Year                       0.05          0.03             0.09           0.09
##                  BOAvgOpenWeekend Budget DomesticGross WorldGross ForeignGross
## RottenTomatoes               0.43   0.05          0.26       0.21         0.18
## AudienceScore                0.35   0.10          0.28       0.26         0.23
## TheatersOpenWeek            -0.32   0.61          0.54       0.55         0.52
## OpeningWeekend               0.24   0.74          0.95       0.91         0.84
## BOAvgOpenWeekend             1.00   0.11          0.30       0.28         0.26
## Budget                       0.11   1.00          0.70       0.79         0.79
## DomesticGross                0.30   0.70          1.00       0.94         0.85
## WorldGross                   0.28   0.79          0.94       1.00         0.98
## ForeignGross                 0.26   0.79          0.85       0.98         1.00
## Year                         0.01   0.01          0.11       0.10         0.09
##                  Year
## RottenTomatoes   0.05
## AudienceScore    0.03
## TheatersOpenWeek 0.09
## OpeningWeekend   0.09
## BOAvgOpenWeekend 0.01
## Budget           0.01
## DomesticGross    0.11
## WorldGross       0.10
## ForeignGross     0.09
## Year             1.00
```



Cross-Validation

```
## Warning: package 'lattice' was built under R version 4.3.2
```

```
control <- trainControl(method="repeatedcv", number=10, repeats=10, savePredictions = "all" )

set.seed(11152023)
model1 <- train(data=Movies1,
                AudienceScore ~ RottenTomatoes ,
                method="lm", trControl=control)

set.seed(11152023)
model2 <- train(data=Movies1,
                AudienceScore ~ RottenTomatoes + BOAvgOpenWeekend,
                method="lm", trControl=control)

set.seed(11152023)
model3 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + Genre + Budget + LeadStudio + WorldGross
                method="lm", trControl=control)
```

```
set.seed(11152023)
model4 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + BOAvgOpenWeekend + I(BOAvgOpenWeekend^2)

set.seed(11152023)
model5 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + BOAvgOpenWeekend + DomesticGross,  metho

set.seed(11152023)
model6 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + BOAvgOpenWeekend + DomesticGross + Openi

set.seed(11152023)
model7 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + BOAvgOpenWeekend + I(BOAvgOpenWeekend^2)

set.seed(11152023)
model8 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + Genre + Budget + LeadStudio + WorldGross

set.seed(11152023)
model9 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + Genre + Budget + LeadStudio + WorldGross

set.seed(11152023)
model10 <- train(data=Movies1, AudienceScore ~ RottenTomatoes + I(RottenTomatoes^2) + Genre + Budget +
                method="lm", trControl=control)


# Calculate RMSPE for each model
RMSPE1 <- sqrt(mean((model1$pred$obs-model1$pred$pred)^2))
RMSPE2 <- sqrt(mean((model2$pred$obs-model2$pred$pred)^2))
RMSPE3 <- sqrt(mean((model3$pred$obs-model3$pred$pred)^2))
RMSPE4 <- sqrt(mean((model4$pred$obs-model4$pred$pred)^2))
RMSPE5 <- sqrt(mean((model5$pred$obs-model5$pred$pred)^2))
RMSPE6 <- sqrt(mean((model6$pred$obs-model6$pred$pred)^2))
RMSPE7 <- sqrt(mean((model7$pred$obs-model7$pred$pred)^2))
RMSPE8 <- sqrt(mean((model8$pred$obs-model8$pred$pred)^2))
RMSPE9 <- sqrt(mean((model9$pred$obs-model9$pred$pred)^2))
RMSPE10 <- sqrt(mean((model10$pred$obs-model10$pred$pred)^2))
```

```
## [1] 13.5343

## [1] 13.52223

## [1] 12.61946

## [1] 13.69228

## [1] 13.4318

## [1] 13.40912

## [1] 13.69228

## [1] 12.6874

## [1] 12.39552

## [1] 12.61001

##   predictions
## 1    84.89408
## 2    76.00513
## 3    37.27391
## 4    23.34207
## 5    80.31602
## 6    71.94609
## 7    66.35857
## 8    41.00000
## 9    43.97618
## 10   42.79211
```

Study your predictions and see if they make sense to you. Write a paragraph summarizing your results. Address the following questions:

```
How complex was the model that did the best in cross-validation? Why do you think this model did the bes
Which predicted value came out the highest? What explanatory variable(s) do you think contributed to thi
Which predicted value came out the lowest? What explanatory variable(s) do you think contributed to this
```

The second most complex model (model9) did the best in the cross-validation. I think this model did the best because it included quadratic terms for the variables that appears to have a curve in the residual plots.

The first movie Doctor Strange had the highest value of 84.89096 which is the highest from all the predicted value. Doctor Strange had the highest rotten tomatoes scores, BOAvgOpenWeekend, OpeningWeekend, WorldGross, and DomesticGross, it had the third highest budget as well. I think all these explanatory variables contributed to this movie getting a high predicted value.

The lowest value is 23.33595 which is for the fourth movie Friend Request. This movie had the lowest rotten tomatoes score and BOAvgOpenWeekend. Budget, OpeningWeekend, WorldGross, and DomesticGross are one of the lowest three movies. I think all these explanatory variables contributed to why this movie gets the lowest predicted value.

Then, refer to the "True_Values.html" file, which contains the true values for each of the cases you're trying to predict. Write a paragraph addressing the following questions.

```
How accurate were your predictions in general?
Which case(s) did the model do a good job of predicting? Why do you think the model did a good job in pr
Which case(s) did the model not predict well? What do you think made these cases hard to predict?
```

I think my predictions in general is pretty accurate, some of them predict exactly what was the audience score, the biggest difference is around 25 lower than the true value.

The model did a good job predicting the 8th movie Phoenix Forgotten because the prediction score is the same as the actual score 41. I think this model did a good job because the rotten tomatoes score is also 41. Rotten tomatoes scores are very correlated to the audience score. So even the other explanatory variables, the 8th movie are fairly low compared to other movies, it is still predicted accurately.

The model did not predict the movie King Arthur: Legend of the Sword well. I think one reason why this movie is hard to predict because it has a low rotten tomatoes of 31 which is 38 lower than the true audience score, and rotten tomatoes are very correlated to the audience score, we can see in the correlation table that the correlation is 0.71. So even the other explanatory variables have higher number, the audience score was still predicted a lot lower.