STOR 455 Group Project (Due 5pm on November 24th)

Old Geezers

Hunter Buresh (730481420), Alex Georgiev (730483404), Zach Richardson (730408740)

The Prediction (Required)

Our prediction of the cumulative domestic box office of “The Marvels” by December 8, 2023 is $104,112,752.

Summary of Justification (Required)

Our prediction of $104,112,752 million dollars for the cumulative domestic box office of The Marvels by December 8, 2023, was derived by investigating several linear regression models. Initially, we analyzed the historical data of Marvel movies, particularly noting the low opening gross of The Marvels. We developed three linear regression models based on this dataset: one considering all movies, another focusing specifically on those with lower opening grosses (≤ $100 Million), and a third for films released in the fall season. Alongside this, we also built a fourth model based on daily earnings data since November 12. This model was trained on live daily data from the Box Office Mojo website. The reason why we chose Nov 12 is because the movie was released on Nov 10, so the increase in earnings between Nov 10 and Nov 12 was nonlinear. Therefore, we chose Nov 12 because it was the earnings after the opening weekend, so the earnings started to follow a more linear trend after that. This model was crucial for capturing the real-time market response and audience trends, providing a contemporary perspective to our analysis. We decided to solely rely on this model for our final prediction because The Marvels had the lowest opening gross earnings out of all of the Marvel movies in the previous dataset we used, which put it around 1.23 standard deviations below the mean. Not only that, but the standard deviation of opening gross earnings was around $70 million, which is almost double the opening gross of The Marvels, so The Marvels was pretty far from the average opening gross earnings. This means that it doesn’t make sense to build a model based on Marvel movies that were so much more successful than The Marvels. We also needed to make sure that our model predicted exactly for December 8 (26 days after the movie was released) and we can’t guarantee that the total gross earnings from the other dataset was exactly 26 days after the movie was released. We believe that only using past earnings for The Marvels is the most reflective of the movie’s current trajectory.

Data (Required)

We created the marveldata.csv dataset in Excel by using the table from: (https://www.boxofficemojo.com/franchise/fr541495045/). The data has the opening gross and total gross earnings for 33 Marvel movies.

Preview of dataset

```
data <- read.csv("marveldata.csv", quote="'")
summary(data)
```

```
##        id        title              totalgross        totaltheaters
##  Min.   : 1   Length:33          Min.   : 48483234   Min.   :3508
##  1st Qu.: 9   Class :character   1st Qu.:214504909   1st Qu.:4080
##  Median :17   Mode  :character   Median :333176600   Median :4275
##  Mean   :17                      Mean   :356173708   Mean   :4204
##  3rd Qu.:25                      3rd Qu.:411331607   3rd Qu.:4349
##  Max.   :33                      Max.   :858373000   Max.   :4662
##    opengross          opentheaters      date             distributer
##  Min.   : 46110859   Min.   :3505   Length:33          Length:33
##  1st Qu.: 80366312   1st Qu.:4030   Class :character   Class :character
##  Median :117027503   Median :4253   Mode  :character   Mode  :character
##  Mean   :132665838   Mean   :4196
##  3rd Qu.:179139142   3rd Qu.:4349
##  Max.   :357115007   Max.   :4662
##     season             production          marketing            ratings
##  Length:33          Min.   :130000000   Min.   : 65000000   Min.   :46.00
##  Class :character   1st Qu.:165000000   1st Qu.: 82500000   1st Qu.:76.00
##  Mode  :character   Median :200000000   Median :100000000   Median :83.00
##                     Mean   :204081818   Mean   :102040909   Mean   :80.88
##                     3rd Qu.:236200000   3rd Qu.:118100000   3rd Qu.:91.00
##                     Max.   :400000000   Max.   :200000000   Max.   :96.00
```

```r
head(data, 5)
```

```
##   id                             title totalgross totaltheaters opengross
## 1 33                       The Marvels   48483234          4030  46110859
## 2 14    Guardians of the Galaxy Vol. 3  358995815          4450 118414021
## 3 25 Ant-Man and the Wasp: Quantumania  214504909          4345 106109650
## 4  7    Black Panther: Wakanda Forever  453829060          4396 181339761
## 5 15            Thor: Love and Thunder  343256830          4375 144165107
##   opentheaters       date                         distributer season production
## 1         4030 2023-11-10 Walt Disney Studios Motion Pictures   Fall  270000000
## 2         4450 2023-05-05 Walt Disney Studios Motion Pictures Spring  250000000
## 3         4345 2023-02-17 Walt Disney Studios Motion Pictures Winter  200000000
## 4         4396 2022-11-11 Walt Disney Studios Motion Pictures   Fall  250000000
## 5         4375 2022-07-08 Walt Disney Studios Motion Pictures Summer  250000000
##   marketing ratings
## 1 135000000      62
## 2 125000000      82
## 3 100000000      46
## 4 125000000      83
## 5 125000000      63
```

```r
# We don't need the id, distributor, title, or date columns
modelData <- subset(data, select = -c(id, distributer, title, date))
summary(modelData)
```

```
##    totalgross        totaltheaters    opengross          opentheaters
##  Min.   : 48483234   Min.   :3508   Min.   : 46110859   Min.   :3505
##  1st Qu.:214504909   1st Qu.:4080   1st Qu.: 80366312   1st Qu.:4030
##  Median :333176600   Median :4275   Median :117027503   Median :4253
##  Mean   :356173708   Mean   :4204   Mean   :132665838   Mean   :4196
##  3rd Qu.:411331607   3rd Qu.:4349   3rd Qu.:179139142   3rd Qu.:4349
##  Max.   :858373000   Max.   :4662   Max.   :357115007   Max.   :4662
##     season             production          marketing            ratings
##  Length:33          Min.   :130000000   Min.   : 65000000   Min.   :46.00
##  Class :character   1st Qu.:165000000   1st Qu.: 82500000   1st Qu.:76.00
##  Mode  :character   Median :200000000   Median :100000000   Median :83.00
##                     Mean   :204081818   Mean   :102040909   Mean   :80.88
##                      3rd Qu.:236200000   3rd Qu.:118100000   3rd Qu.:91.00
##                      Max.   :400000000   Max.   :200000000   Max.   :96.00
```

Daily earnings dataset

This dataset was put together from the table found at (https://www.boxofficemojo.com/release/rl247366145/?ref_=bo_tt_gr_1). It has the to-date revenue of The Marvels from Nov 10 (release) to Nov 22. We believe that it's more accurate to build a model around past data of The Marvels rather than combining data from many movies that did not follow a similar earnings trend.

```r
# Convert table from website into dataframe
dailyEarnings <- data.frame(
  Date = c("Nov 22", "Nov 21", "Nov 20", "Nov 19",
           "Nov 18", "Nov 17", "Nov 16", "Nov 15", "Nov 14",
           "Nov 13", "Nov 12", "Nov 11", "Nov 10"),
  DayOfWeek = c("Wednesday", "Tuesday", "Monday", "Sunday",
                "Saturday", "Friday", "Thursday", "Wednesday",
                "Tuesday", "Monday", "Sunday", "Saturday", "Friday"),
  Rank = c(5, 3, 4, 3, 3, 4, 1, 1, 1, 1, 1, 1, 1),
  Revenue = c(1500000, 1570855, 1137196, 2910248, 4453682,
              2756659, 1251387, 1789239, 3300946, 2372375,
              9247703, 15260052, 21603104),
  Change_Daily = c("-4.5%", "+38.1%", "-60.9%", "-34.7%",
                   "+61.6%", "+120.3%", "-30.1%", "-45.8%",
                   "+39.1%", "-74.3%", "-39.4%", "-29.4%", "-"),
  Change_LastWeek = c("-16.2%", "-52.4%", "-52.1%",
                      "-68.5%", "-70.8%", "-87.2%",
                      "-", "-", "-", "-", "-", "-", "-"),
  Theaters = c(3070, 4030, 4030, 4030, 4030, 4030, 4030,
               4030, 4030, 4030, 4030, 4030, 4030),
  Avg = c(488, 389, 282, 722, 1105, 684, 310, 443, 819,
          588, 2294, 3786, 5360),
  TotalRevenue = c(69153446, 67653446, 66082591, 64945395,
                   62035147, 57581465, 54824806, 53573419,
                   51784180, 48483234, 46110859, 36863156, 21603104),
  Day = c(13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1)
)

# Preview data
summary(dailyEarnings)
```

```
##      Date            DayOfWeek              Rank          Revenue
##  Length:13          Length:13          Min.   :1.000   Min.   : 1137196
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.: 1570855
##  Mode  :character   Mode  :character   Median :1.000   Median : 2756659
##                                        Mean   :2.231   Mean   : 5319496
##                                        3rd Qu.:3.000   3rd Qu.: 4453682
##                                        Max.   :5.000   Max.   :21603104
##  Change_Daily       Change_LastWeek       Theaters         Avg
##  Length:13          Length:13          Min.   :3070   Min.   : 282
##  Class :character   Class :character   1st Qu.:4030   1st Qu.: 443
##  Mode  :character   Mode  :character   Median :4030   Median : 684
##                                        Mean   :3956   Mean   :1328
##                                        3rd Qu.:4030   3rd Qu.:1105
##                                        Max.   :4030   Max.   :5360
##   TotalRevenue           Day
##  Min.   :21603104   Min.   : 1
##  1st Qu.:48483234   1st Qu.: 4
##  Median :54824806   Median : 7
##  Mean   :53899558   Mean   : 7
##  3rd Qu.:64945395   3rd Qu.:10
##  Max.   :69153446   Max.   :13
```

```
head(dailyEarnings)
```

```
##     Date DayOfWeek Rank Revenue Change_Daily Change_LastWeek Theaters  Avg
## 1 Nov 22 Wednesday    5 1500000        -4.5%          -16.2%     3070  488
## 2 Nov 21   Tuesday    3 1570855       +38.1%          -52.4%     4030  389
## 3 Nov 20    Monday    4 1137196       -60.9%          -52.1%     4030  282
## 4 Nov 19    Sunday    3 2910248       -34.7%          -68.5%     4030  722
## 5 Nov 18  Saturday    3 4453682       +61.6%          -70.8%     4030 1105
## 6 Nov 17    Friday    4 2756659      +120.3%          -87.2%     4030  684
##   TotalRevenue Day
## 1     69153446  13
## 2     67653446  12
## 3     66082591  11
## 4     64945395  10
## 5     62035147   9
## 6     57581465   8
```

Analysis (Required)

We noticed that the opening gross for The Marvels was only $46 million. Looking at the distribution of the opening gross of other Marvel movies shows that The Marvels has the lowest opening gross out of all the movies in the dataset. We found that The Marvels was around 1.23 standard deviations below the mean. The standard deviation was around $70 million, which is almost double the opening gross of The Marvels. This means that The Marvels was pretty far from the average opening gross of other Marvel movies. This is why the prediction from the model with all Marvel movies is too optimistic. We decided to train a model based on Movies that had low opening weekend earnings to better reflect the performance of The Marvels, and a model that only contained movies with fall releases. The problem with these models is that we can't guarrantee the period of time between the movie's release and the gross earnings that we're predicting for is the same 26 days after release that we're predicting for The Marvels. However, these models still gave us ballpark figures to compare our final daily earnings model against. Our final prediction was pretty close to the model with only fall releases. All models showed that it makes sense to predict future gross earnings based on the opening gross earnings because of the clear linear relationship between the two showcased by the plots with the regression lines.

Distribution of opening gross

```
openGrossMil <- data$opengross/1000000
summary(openGrossMil)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.11   80.37  117.03  132.67  179.14  357.12
```

```
# Find std dev and mean of opening gross
oneStd = sd(data$opengross)
cat("One standard deviation:",oneStd, "\n")
```

```
## One standard deviation: 70424430
```

```r
meanOpenGross = mean(data$opengross)
cat("Mean of opening gross: ", meanOpenGross, "\n")
```

```
## Mean of opening gross:  132665838
```

```r
# Get The Marvels opening gross
theMarvelsOpenGross = data.frame(opengross=(subset(data, title=="The Marvels"))$opengross)
theMarvelsOpenGrossNum = theMarvelsOpenGross$opengross
# Find standard deviations below the mean The Marvels was
numBelowMean = (meanOpenGross - theMarvelsOpenGrossNum) / oneStd
cat("Number of Std Devs The Marvels is below the mean: ", numBelowMean, "\n")
```

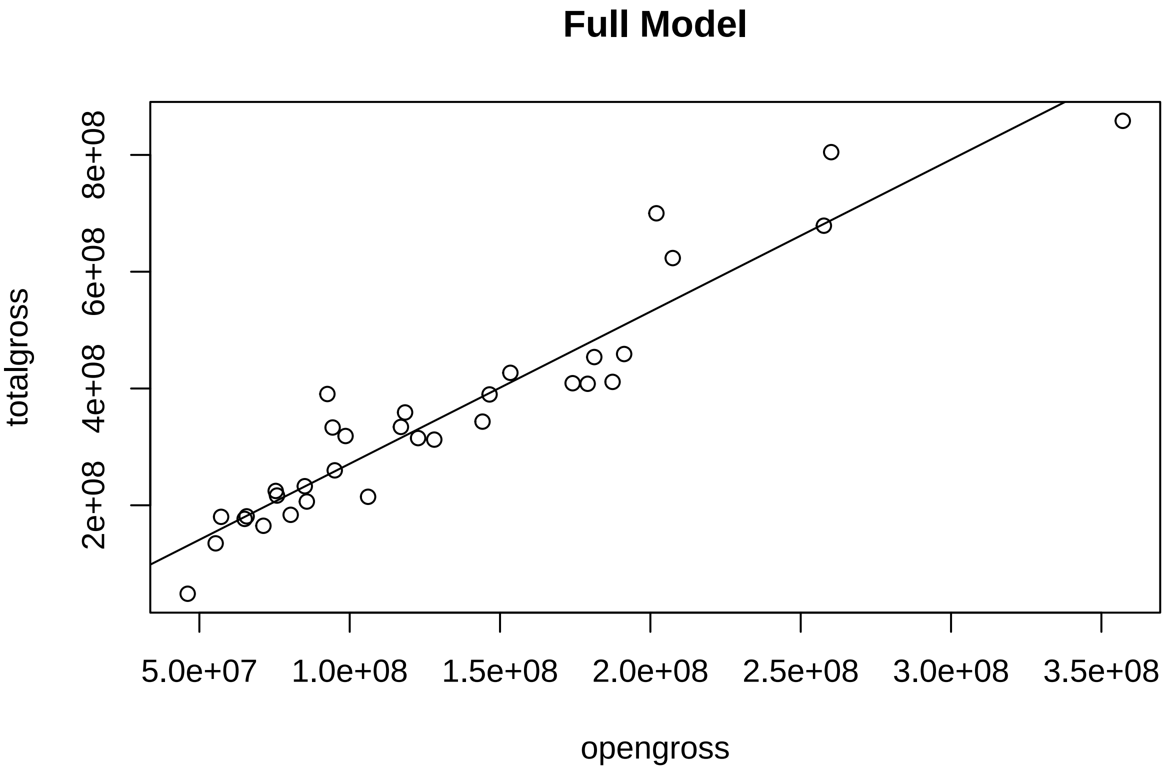
```
## Number of Std Devs The Marvels is below the mean:  1.229048
```

Full model

```r
# Fit a model with all data
fullModel = lm(totalgross~opengross, data=modelData)
summary(fullModel)
```

```
##
## Call:
## lm(formula = totalgross ~ opengross, data = modelData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -87453553 -36306375  -3008707  18757942 163292669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.064e+07  2.343e+07   0.454    0.653
## opengross   2.604e+00  1.565e-01  16.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62350000 on 31 degrees of freedom
## Multiple R-squared:  0.8993, Adjusted R-squared:  0.8961
## F-statistic: 276.9 on 1 and 31 DF,  p-value: < 2.2e-16
```

```r
# Plot regression line
plot(totalgross~opengross, data=modelData,
     main="Full Model")
abline(fullModel)
```

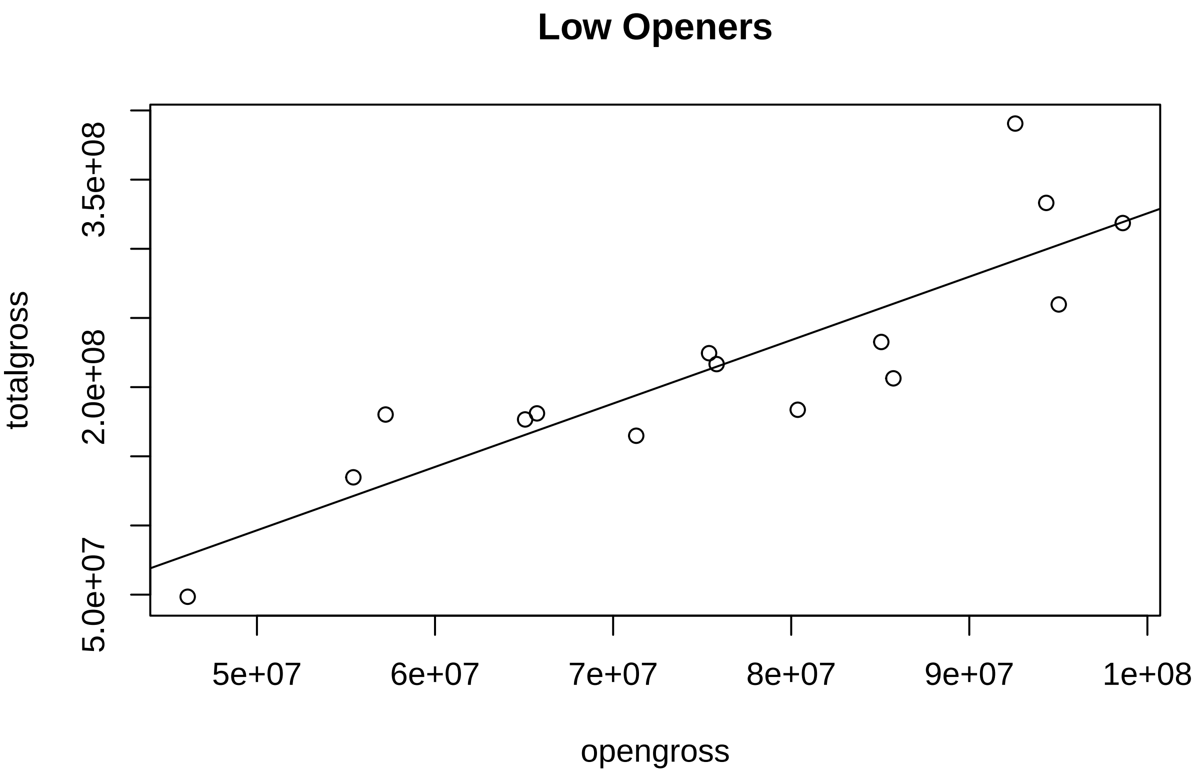


Model with low opening gross

```r
# Get data with opening gross < 100 mil
lowOpeners = subset(modelData, opengross<=100000000)

# Fit model with low openers
lowOpenModel = lm(totalgross~opengross, data=lowOpeners)
summary(lowOpenModel)
```

```
##
## Call:
## lm(formula = totalgross ~ opengross, data = lowOpeners)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -53920824 -29717738   1852240  12979095  98896996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.326e+08  5.563e+07  -2.384   0.0331 *
## opengross    4.583e+00  7.150e-01   6.409 2.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42880000 on 13 degrees of freedom
## Multiple R-squared:  0.7596, Adjusted R-squared:  0.7411
## F-statistic: 41.08 on 1 and 13 DF,  p-value: 2.31e-05
```

```
# Plot regression line
plot(totalgross~opengross, data=lowOpeners,
     main="Low Openers")
abline(lowOpenModel)
```



Model with only fall releases

```
# Only take movies that had fall releases
fallData = subset(modelData, season=="Fall")
summary(fallData)
```

```
##    totalgross        totaltheaters    opengross          opentheaters
##  Min.   : 48483234   Min.   :3841   Min.   : 46110859   Min.   :3841
##  1st Qu.:185616187   1st Qu.:3956   1st Qu.: 73342954   1st Qu.:3956
##  Median :224543292   Median :4080   Median : 85058311   Median :4080
##  Mean   :235112596   Mean   :4088   Mean   : 95382524   Mean   :4088
##  3rd Qu.:273850104   3rd Qu.:4195   3rd Qu.:104241415   3rd Qu.:4195
##  Max.   :453829060   Max.   :4396   Max.   :181339761   Max.   :4396
##     season            production          marketing            ratings
##  Length:7           Min.   :150000000   Min.   : 75000000   Min.   :47.00
##  Class :character   1st Qu.:157500000   1st Qu.: 78750000   1st Qu.:64.50
##  Mode  :character   Median :180000000   Median : 90000000   Median :83.00
##                     Mean   :200171429   Mean   :100085714   Mean   :76.14
##                     3rd Qu.:243100000   3rd Qu.:121550000   3rd Qu.:90.50
##                     Max.   :270000000   Max.   :135000000   Max.   :93.00
```

```
head(modelData)
```

```
##   totalgross totaltheaters opengross opentheaters season production marketing
## 1   48483234          4030  46110859         4030   Fall  270000000 135000000
## 2  358995815          4450 118414021         4450 Spring  250000000 125000000
## 3  214504909          4345 106109650         4345 Winter  200000000 100000000
## 4  453829060          4396 181339761         4396   Fall  250000000 125000000
## 5  343256830          4375 144165107         4375 Summer  250000000 125000000
## 6  411331607          4534 187420998         4534 Spring  200000000 100000000
##   ratings
## 1      62
## 2      82
## 3      46
## 4      83
## 5      63
## 6      73
```

```
# Fit model with fall releases
fallModel = lm(totalgross~opengross, data=fallData)
summary(fallModel)
```

```
##
## Call:
## lm(formula = totalgross ~ opengross, data = fallData)
##
## Residuals:
##         1         4         8         9        17        20        26
## -50747029 -18338034  -3819451  44570077   4484968  26001635  -2152166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.794e+07  3.209e+07   -0.87 0.423893
## opengross    2.758e+00  3.092e-01    8.92 0.000295 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33510000 on 5 degrees of freedom
## Multiple R-squared:  0.9409, Adjusted R-squared:  0.929
## F-statistic: 79.56 on 1 and 5 DF,  p-value: 0.000295
```

```
# Plot regression line
plot(totalgross~opengross, data=fallData,
     main="Fall Releases")
abline(fallModel)
```

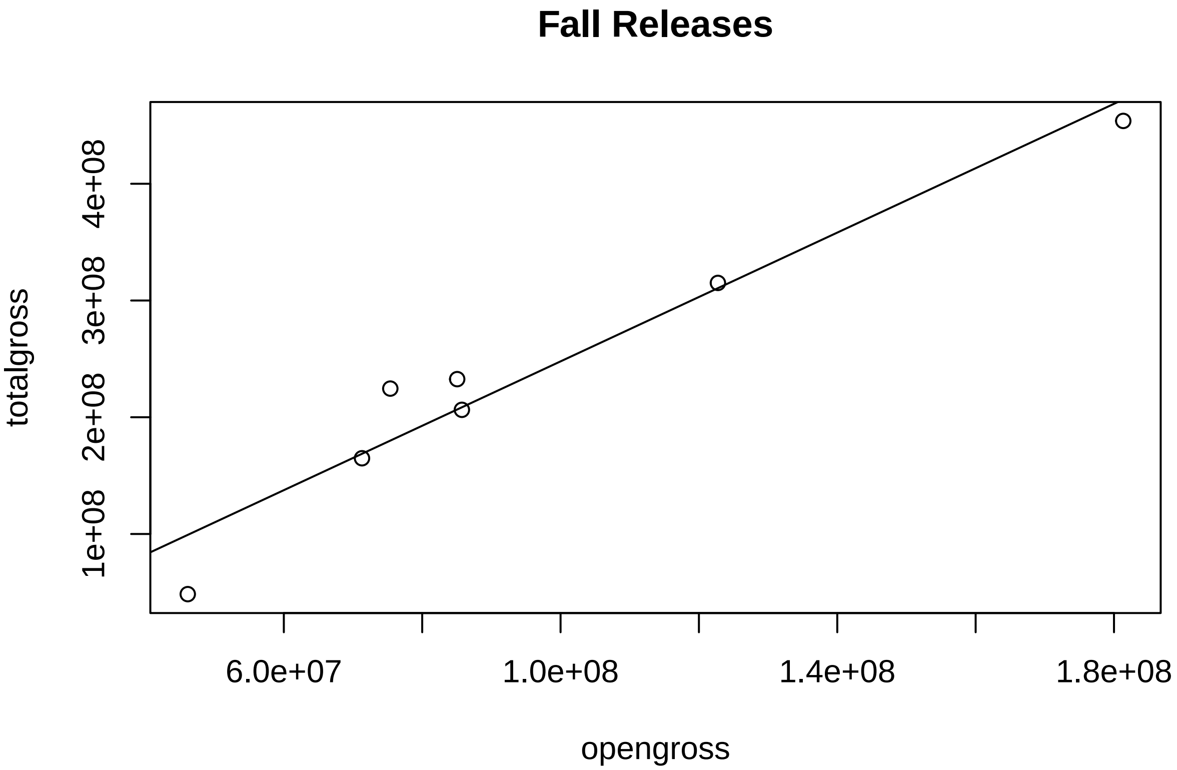


Table with prediction from all models

```r
# Predictions for each model
fullPred = predict(fullModel, newdata = theMarvelsOpenGross)
lowOpenPred = predict(lowOpenModel, newdata = theMarvelsOpenGross)
fallPred = predict(fallModel, newdata = theMarvelsOpenGross)

# Create a data frame for the table
predictionTable <- data.frame(
  FullModel = fullPred,
  LowOpenModel = lowOpenPred,
  FallModel = fallPred
)

# Display the table
print(predictionTable)
```

```
##   FullModel LowOpenModel FallModel
## 1 130738678     78683385  99230263
```

SLR based on earnings since Nov 12

```r
# Data that only contains Nov 12 and later
sinceNov12 = subset(dailyEarnings, (Date != "Nov 10") & (Date != "Nov 11"))

# Add column that represents days since Nov 12
sinceNov12$DaysAfterOpeningWknd = sinceNov12$Day - 3

# Show data
```

```
summary(sinceNov12)
```

```
##      Date            DayOfWeek              Rank          Revenue
##  Length:11          Length:11          Min.   :1.000   Min.   :1137196
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:1535428
##  Mode  :character   Mode  :character   Median :3.000   Median :2372375
##                                        Mean   :2.455   Mean   :2935481
##                                        3rd Qu.:3.500   3rd Qu.:3105597
##                                        Max.   :5.000   Max.   :9247703
##  Change_Daily       Change_LastWeek       Theaters         Avg
##  Length:11          Length:11          Min.   :3070   Min.   : 282.0
##  Class :character   Class :character   1st Qu.:4030   1st Qu.: 416.0
##  Mode  :character   Mode  :character   Median :4030   Median : 588.0
##                                        Mean   :3943   Mean   : 738.5
##                                        3rd Qu.:4030   3rd Qu.: 770.5
##                                        Max.   :4030   Max.   :2294.0
##   TotalRevenue            Day       DaysAfterOpeningWknd
##  Min.   :46110859   Min.   : 3.0   Min.   : 0.0
##  1st Qu.:52678800   1st Qu.: 5.5   1st Qu.: 2.5
##  Median :57581465   Median : 8.0   Median : 5.0
##  Mean   :58384363   Mean   : 8.0   Mean   : 5.0
##  3rd Qu.:65513993   3rd Qu.:10.5   3rd Qu.: 7.5
##  Max.   :69153446   Max.   :13.0   Max.   :10.0
```

```
sinceNov12
```

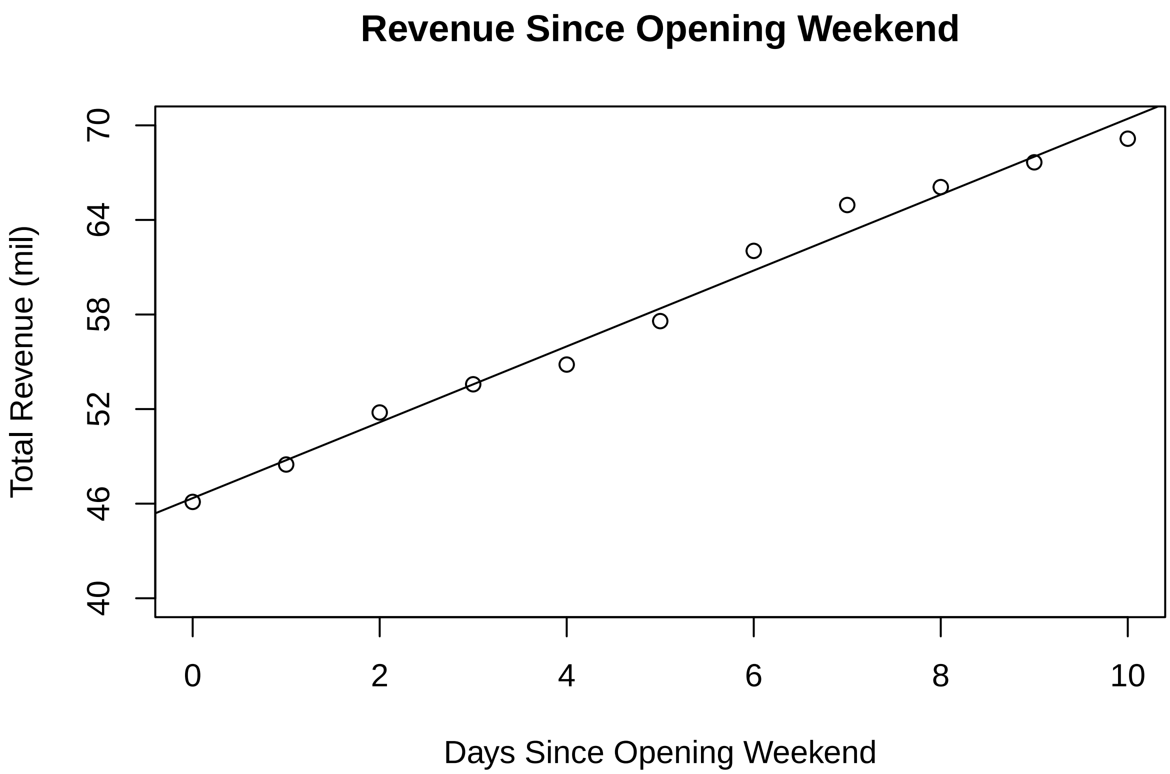
```
##        Date DayOfWeek Rank Revenue Change_Daily Change_LastWeek Theaters  Avg
## 1  Nov 22 Wednesday    5 1500000        -4.5%          -16.2%     3070  488
## 2  Nov 21   Tuesday    3 1570855       +38.1%          -52.4%     4030  389
## 3  Nov 20    Monday    4 1137196       -60.9%          -52.1%     4030  282
## 4  Nov 19    Sunday    3 2910248       -34.7%          -68.5%     4030  722
## 5  Nov 18  Saturday    3 4453682       +61.6%          -70.8%     4030 1105
## 6  Nov 17    Friday    4 2756659      +120.3%          -87.2%     4030  684
## 7  Nov 16  Thursday    1 1251387       -30.1%               -     4030  310
## 8  Nov 15 Wednesday    1 1789239       -45.8%               -     4030  443
## 9  Nov 14   Tuesday    1 3300946       +39.1%               -     4030  819
## 10 Nov 13    Monday    1 2372375       -74.3%               -     4030  588
## 11 Nov 12    Sunday    1 9247703       -39.4%               -     4030 2294
##    TotalRevenue Day DaysAfterOpeningWknd
## 1      69153446  13                   10
## 2      67653446  12                    9
## 3      66082591  11                    8
## 4      64945395  10                    7
## 5      62035147   9                    6
## 6      57581465   8                    5
## 7      54824806   7                    4
## 8      53573419   6                    3
## 9      51784180   5                    2
## 10     48483234   4                    1
## 11     46110859   3                    0
```

```
# Fit model passed on days since opening weekend
sinceNov12Model = lm(TotalRevenue~DaysAfterOpeningWknd, data=sinceNov12)

# Summarize model
```

```r
summary(sinceNov12Model)
```

```
##
## Call:
## lm(formula = TotalRevenue ~ DaysAfterOpeningWknd, data = sinceNov12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1264703  -580422  -239717   549023  1747518
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          46350576     565584   81.95 3.04e-14 ***
## DaysAfterOpeningWknd  2406757      95601   25.18 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1003000 on 9 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.9844
## F-statistic: 633.8 on 1 and 9 DF,  p-value: 1.183e-09
```

```r
# Convert to millions for easier viewing
plotData = sinceNov12
plotData$TotalRevenue = plotData$TotalRevenue / 1000000

# Plot regression line
milModel = lm(TotalRevenue~DaysAfterOpeningWknd, data=plotData)
plot(TotalRevenue~DaysAfterOpeningWknd, data=plotData,
     yaxp=c(40,70,5), ylim=c(40, 70),
     main="Revenue Since Opening Weekend",
     xlab="Days Since Opening Weekend", ylab="Total Revenue (mil)")
abline(milModel)
```

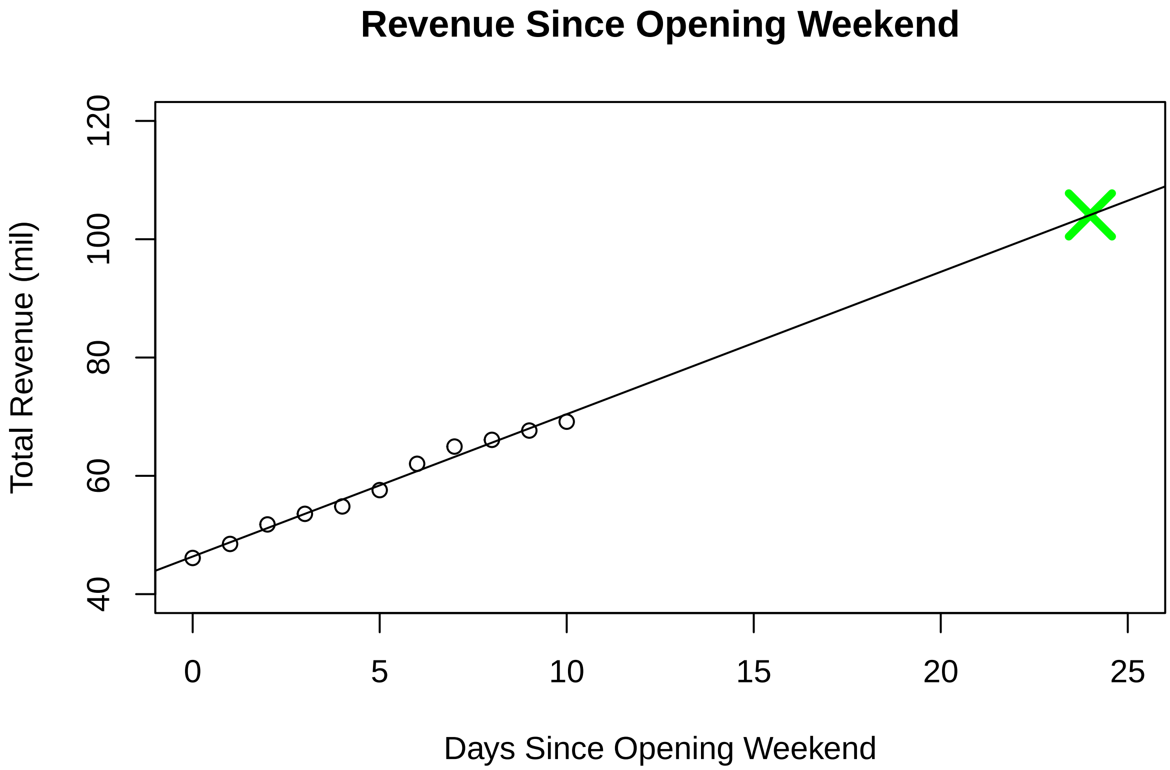


```r
finalPrediction = predict(sinceNov12Model,
                          newdata=data.frame(DaysAfterOpeningWknd=c(24)))
cat("Our prediction based on daily earnings since Nov 12: ",
    finalPrediction, "\n")
```

```
## Our prediction based on daily earnings since Nov 12:  104112752
```

Plot with final prediction

```r
plot(TotalRevenue~DaysAfterOpeningWknd, data=plotData,
     ylim=c(40, 120), xlim=c(0, 25),
     main="Revenue Since Opening Weekend",
     xlab="Days Since Opening Weekend", ylab="Total Revenue (mil)")
points(x=24,y=finalPrediction/1000000,pch=4, cex=3, lw=4, col="green")
abline(milModel)
```


Revenue Since Opening Weekend
Total Revenue (mil)
40
60
80
100
120
0
5
10
15
20
25
Days Since Opening Weekend