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

Old Geezers

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

The Prediction (Required)

Our prediction of the cumulative domestic box office of "The Marvels" by December 8, 2023 is $112,545,156

Summary of Justification (Required)

Our prediction of 112,545,156 million dollars for the cumulative domestic box office of "The Marvels" by December 8, 2023, was arrived at through the integration of four linear regression models. Initially, we analyzed the historical data of Marvel movies, particularly noting the low opening gross of "The Marvels". We developed three 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 incorporated a fourth model based on daily earnings data since November 12. This model was crucial in capturing the real-time market response and audience trends, providing a contemporary perspective to our analysis. Eventually this model proved most accurate in determining the prediction. We believe this approach, blending historical data patterns with current box office dynamics, makes our 112,545,156 million dollar prediction both robust and reflective of the movie's current trajectory.

Data (Required)

We obtained the dataset from: https://www.boxofficemojo.com/franchise/fr541495045/. ### Preview of dataset

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

```
##        id       perform            title            totalgross
##  Min.   : 1   Length:33          Length:33          Min.   : 65.0
##  1st Qu.: 9   Class :character   Class :character   1st Qu.:215.0
##  Median :17   Mode  :character   Mode  :character   Median :333.0
##  Mean   :17                                         Mean   :356.7
##  3rd Qu.:25                                         3rd Qu.:411.0
##  Max.   :33                                         Max.   :858.0
##  totaltheaters    opengross      opentheaters      date
##  Min.   :3508   Min.   : 46.0   Min.   :3505   Length:33
##  1st Qu.:4080   1st Qu.: 80.0   1st Qu.:4030   Class :character
##  Median :4275   Median :117.0   Median :4253   Mode  :character
##  Mean   :4204   Mean   :132.6   Mean   :4196
```

```
##  3rd Qu.:4349   3rd Qu.:179.0   3rd Qu.:4349
##  Max.   :4662   Max.   :357.0   Max.   :4662
##  distributer           season           production      marketing
##  Length:33          Length:33          Min.   :130.0   Min.   : 65.0
##  Class :character   Class :character   1st Qu.:165.0   1st Qu.: 83.0
##  Mode  :character   Mode  :character   Median :200.0   Median :100.0
##                                        Mean   :204.1   Mean   :102.1
##                                        3rd Qu.:236.0   3rd Qu.:118.0
##                                        Max.   :400.0   Max.   :200.0
##     ratings
##  Min.   :46.00
##  1st Qu.:76.00
##  Median :83.00
##  Mean   :80.88
##  3rd Qu.:91.00
##  Max.   :96.00
```

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

```
##   id perform                             title totalgross totaltheaters
## 1 33      LP                       The Marvels         65          4030
## 2 14      HP    Guardians of the Galaxy Vol. 3        359          4450
## 3 25      HP Ant-Man and the Wasp: Quantumania        215          4345
## 4  7      HP    Black Panther: Wakanda Forever        454          4396
## 5 15      HP            Thor: Love and Thunder        343          4375
##   opengross opentheaters       date                        distributer season
## 1        46         4030 2023-11-10 Walt Disney Studios Motion Pictures   Fall
## 2       118         4450 2023-05-05 Walt Disney Studios Motion Pictures Spring
## 3       106         4345 2023-02-17 Walt Disney Studios Motion Pictures Winter
## 4       181         4396 2022-11-11 Walt Disney Studios Motion Pictures   Fall
## 5       144         4375 2022-07-08 Walt Disney Studios Motion Pictures Summer
##   production marketing ratings
## 1        270       135      62
## 2        250       125      82
## 3        200       100      46
## 4        250       125      83
## 5        250       125      63
```

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

```
##    perform            totalgross    totaltheaters    opengross
##  Length:33          Min.   : 65.0   Min.   :3508   Min.   : 46.0
##  Class :character   1st Qu.:215.0   1st Qu.:4080   1st Qu.: 80.0
##  Mode  :character   Median :333.0   Median :4275   Median :117.0
##                     Mean   :356.7   Mean   :4204   Mean   :132.6
##                     3rd Qu.:411.0   3rd Qu.:4349   3rd Qu.:179.0
##                     Max.   :858.0   Max.   :4662   Max.   :357.0
##   opentheaters     season           production      marketing
##  Min.   :3505   Length:33          Min.   :130.0   Min.   : 65.0
##  1st Qu.:4030   Class :character   1st Qu.:165.0   1st Qu.: 83.0
##  Median :4253   Mode  :character   Median :200.0   Median :100.0
```

```
##  Mean   :4196                     Mean   :204.1   Mean   :102.1
##  3rd Qu.:4349                     3rd Qu.:236.0   3rd Qu.:118.0
##  Max.   :4662                     Max.   :400.0   Max.   :200.0
##     ratings
##  Min.   :46.00
##  1st Qu.:76.00
##  Median :83.00
##  Mean   :80.88
##  3rd Qu.:91.00
##  Max.   :96.00
```

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. ### Distribution of opening gross

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

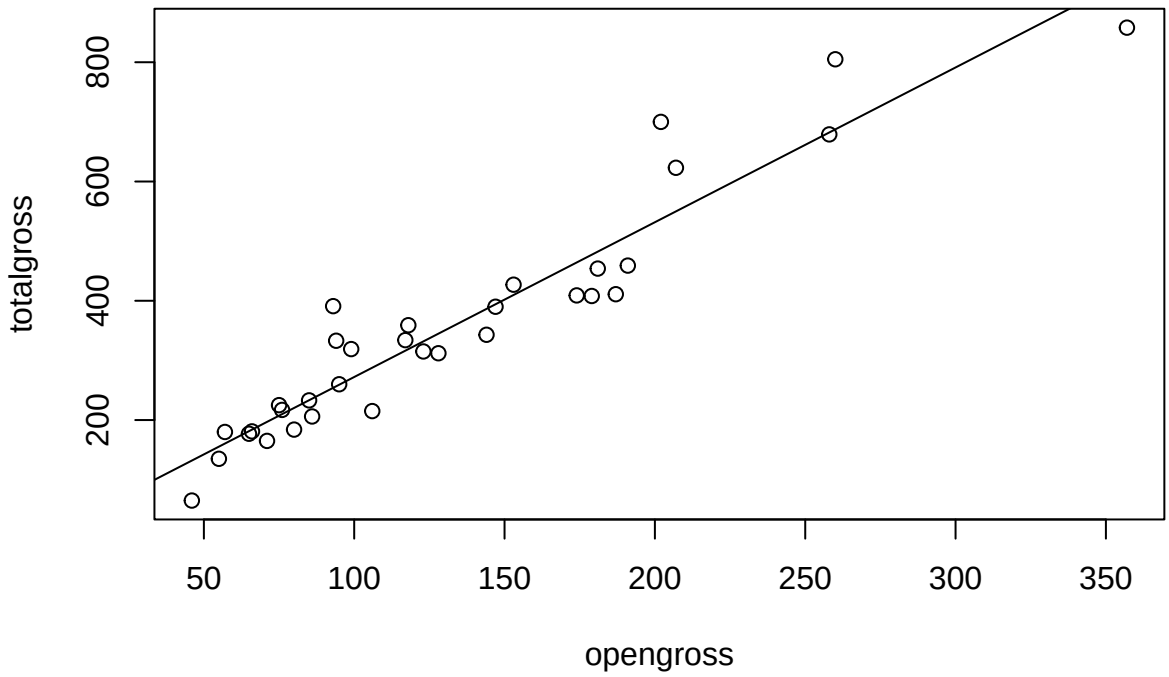
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000460 0.0000800 0.0001170 0.0001326 0.0001790 0.0003570
```

Full model

```
# 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
## -86.980 -36.272  -4.164  17.705 163.089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6400    23.1496   0.546    0.589
## opengross     2.5954     0.1547  16.774   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.63 on 31 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8976
## F-statistic: 281.4 on 1 and 31 DF,  p-value: < 2.2e-16
```

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



Model with low opening gross

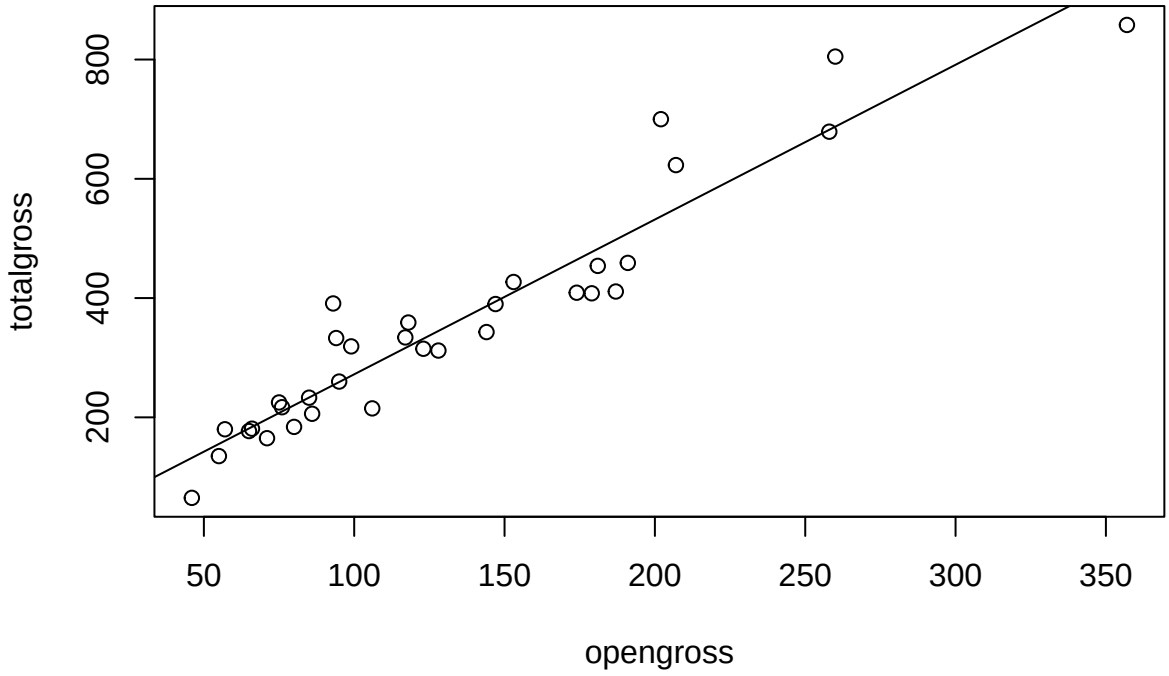
```r
# Get data with opening gross < 100 mil
lowOpeners = subset(modelData, opengross<=100000000)

# 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
## -86.980 -36.272  -4.164  17.705 163.089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  12.6400    23.1496   0.546    0.589
## opengross     2.5954     0.1547  16.774   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.63 on 31 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8976
## F-statistic: 281.4 on 1 and 31 DF,  p-value: < 2.2e-16
```

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



Model with only Fall releases

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

```
##    perform            totalgross    totaltheaters    opengross
##  Length:7           Min.   : 65.0   Min.   :3841   Min.   : 46.00
##  Class :character   1st Qu.:185.5   1st Qu.:3956   1st Qu.: 73.00
##  Mode  :character   Median :225.0   Median :4080   Median : 85.00
```

```
##                     Mean   :237.6   Mean   :4088   Mean   : 95.29
##                     3rd Qu.:274.0   3rd Qu.:4195   3rd Qu.:104.50
##                     Max.   :454.0   Max.   :4396   Max.   :181.00
##   opentheaters     season            production      marketing
##  Min.   :3841   Length:7           Min.   :150.0   Min.   : 75.0
##  1st Qu.:3956   Class :character   1st Qu.:157.5   1st Qu.: 79.0
##  Median :4080   Mode  :character   Median :180.0   Median : 90.0
##  Mean   :4088                      Mean   :200.1   Mean   :100.1
##  3rd Qu.:4195                      3rd Qu.:243.0   3rd Qu.:121.5
##  Max.   :4396                      Max.   :270.0   Max.   :135.0
##     ratings
##  Min.   :47.00
##  1st Qu.:64.50
##  Median :83.00
##  Mean   :76.14
##  3rd Qu.:90.50
##  Max.   :93.00
```

```
head(modelData)
```

```
##   perform totalgross totaltheaters opengross opentheaters season production
## 1      LP         65          4030        46         4030   Fall        270
## 2      HP        359          4450       118         4450 Spring        250
## 3      HP        215          4345       106         4345 Winter        200
## 4      HP        454          4396       181         4396   Fall        250
## 5      HP        343          4375       144         4375 Summer        250
## 6      HP        411          4534       187         4534 Spring        200
##   marketing ratings
## 1       135      62
## 2       125      82
## 3       100      46
## 4       125      83
## 5       125      63
## 6       100      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
## -40.051 -14.042  -7.271  41.973   2.910  23.085  -6.604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.6349    27.7477  -0.672 0.531623
## opengross     2.6888     0.2675  10.050 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 28.99 on 5 degrees of freedom
## Multiple R-squared:  0.9528, Adjusted R-squared:  0.9434
## F-statistic:   101 on 1 and 5 DF,  p-value: 0.0001669
```

```r
# Plot regression line
plot(totalgross~opengross, data=fallData)
abline(fallModel)
```

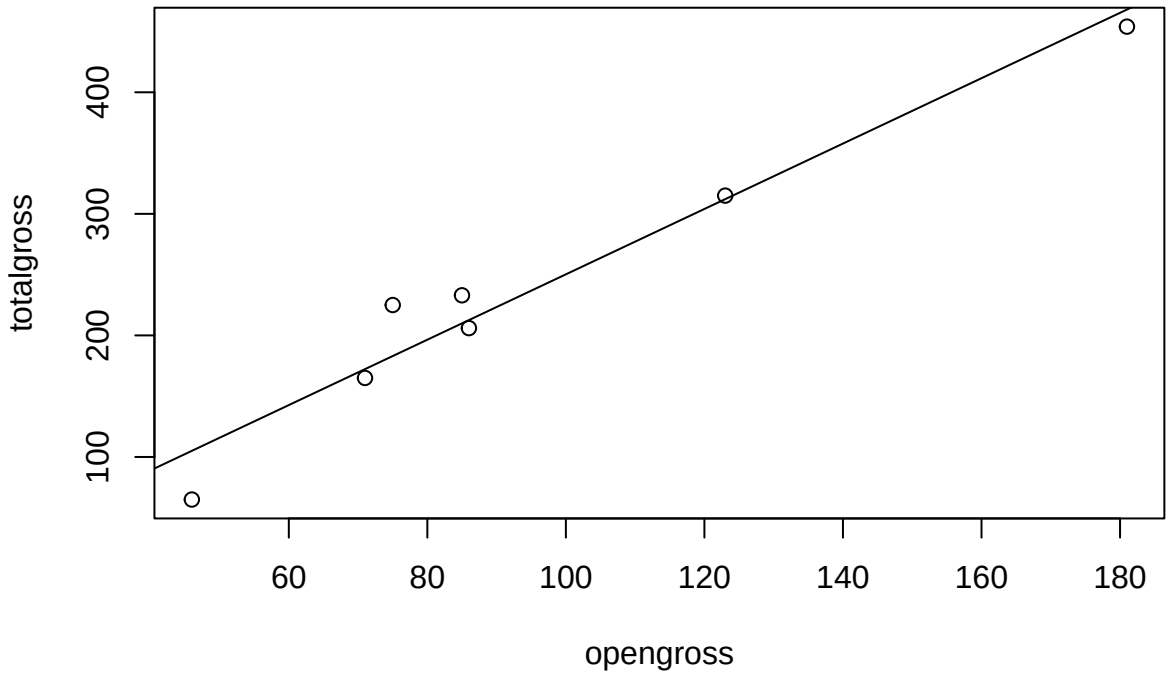


Table with prediction from all models

```r
# Get The Marvels opening gross
theMarvelsOpenGross = data.frame(opengross=(subset(data, title=="The Marvels"))$opengross)

# 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  132.0285     132.0285  105.0509
```

Predict only with daily earnings

```
# Convert table from website into dataframe
dailyEarnings <- data.frame(
  Date = c("Nov 20", "Nov 19", "Nov 18", "Nov 17", "Nov 16", "Nov 15", "Nov 14", "Nov 13", "Nov 12", "No
  Day = c("Monday", "Sunday", "Saturday", "Friday", "Thursday", "Wednesday", "Tuesday", "Monday", "Sunda
  Rank = c(4, 3, 3, 4, 1, 1, 1, 1, 1, 1, 1),
  Revenue = c(1137196, 2910248, 4453682, 2756659, 1251387, 1789239, 3300946, 2372375, 9247703, 15260052
  Change_Daily = c("-60.9%", "-34.7%", "+61.6%", "+120.3%", "-30.1%", "-45.8%", "+39.1%", "-74.3%", "-3
  Change_LastWeek = c("-52.1%", "-68.5%", "-70.8%", "-87.2%", "-", "-", "-", "-", "-", "-", "-"),
  Theaters = c(4030, 4030, 4030, 4030, 4030, 4030, 4030, 4030, 4030, 4030, 4030),
  Avg = c(282, 722, 1105, 684, 310, 443, 819, 588, 2294, 3786, 5360),
  TotalRevenue = c(66082591, 64945395, 62035147, 57581465, 54824806, 53573419, 51784180, 48483234, 4611
  Week = c(11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1)
)

# Preview data
summary(dailyEarnings)
```

```
##      Date               Day                 Rank          Revenue
##  Length:11          Length:11          Min.   :1.000   Min.   : 1137196
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.: 2080807
##  Mode  :character   Mode  :character   Median :1.000   Median : 2910248
##                                        Mean   :1.909   Mean   : 6007508
##                                        3rd Qu.:3.000   3rd Qu.: 6850692
##                                        Max.   :4.000   Max.   :21603104
##  Change_Daily       Change_LastWeek       Theaters         Avg
##  Length:11          Length:11          Min.   :4030   Min.   : 282.0
##  Class :character   Class :character   1st Qu.:4030   1st Qu.: 515.5
##  Mode  :character   Mode  :character   Median :4030   Median : 722.0
##                                        Mean   :4030   Mean   :1490.3
##                                        3rd Qu.:4030   3rd Qu.:1699.5
##                                        Max.   :4030   Max.   :5360.0
##   TotalRevenue          Week
##  Min.   :21603104   Min.   : 1.0
##  1st Qu.:47297046   1st Qu.: 3.5
##  Median :53573419   Median : 6.0
##  Mean   :51262487   Mean   : 6.0
##  3rd Qu.:59808306   3rd Qu.: 8.5
##  Max.   :66082591   Max.   :11.0
```

```r
head(dailyEarnings)
```

```
##      Date       Day Rank Revenue Change_Daily Change_LastWeek Theaters  Avg
## 1 Nov 20    Monday    4 1137196       -60.9%          -52.1%     4030  282
## 2 Nov 19    Sunday    3 2910248       -34.7%          -68.5%     4030  722
## 3 Nov 18  Saturday    3 4453682       +61.6%          -70.8%     4030 1105
## 4 Nov 17    Friday    4 2756659      +120.3%          -87.2%     4030  684
## 5 Nov 16  Thursday    1 1251387       -30.1%               -     4030  310
## 6 Nov 15 Wednesday    1 1789239       -45.8%               -     4030  443
##   TotalRevenue Week
## 1     66082591   11
## 2     64945395   10
## 3     62035147    9
## 4     57581465    8
## 5     54824806    7
## 6     53573419    6
```

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 <- c(8, 7, 6, 5, 4, 3, 2, 1, 0)

# Show data
summary(sinceNov12)
```

```
##      Date               Day                 Rank          Revenue
##  Length:9           Length:9           Min.   :1.000   Min.   :1137196
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:1789239
##  Mode  :character   Mode  :character   Median :1.000   Median :2756659
##                                        Mean   :2.111   Mean   :3246604
##                                        3rd Qu.:3.000   3rd Qu.:3300946
##                                        Max.   :4.000   Max.   :9247703
##  Change_Daily       Change_LastWeek       Theaters         Avg
##  Length:9           Length:9           Min.   :4030   Min.   : 282.0
##  Class :character   Class :character   1st Qu.:4030   1st Qu.: 443.0
##  Mode  :character   Mode  :character   Median :4030   Median : 684.0
##                                        Mean   :4030   Mean   : 805.2
##                                        3rd Qu.:4030   3rd Qu.: 819.0
##                                        Max.   :4030   Max.   :2294.0
##   TotalRevenue          Week    DaysAfterOpeningWknd
##  Min.   :46110859   Min.   : 3   Min.   :0
##  1st Qu.:51784180   1st Qu.: 5   1st Qu.:2
##  Median :54824806   Median : 7   Median :4
##  Mean   :56157900   Mean   : 7   Mean   :4
##  3rd Qu.:62035147   3rd Qu.: 9   3rd Qu.:6
##  Max.   :66082591   Max.   :11   Max.   :8
```

```
sinceNov12
```

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

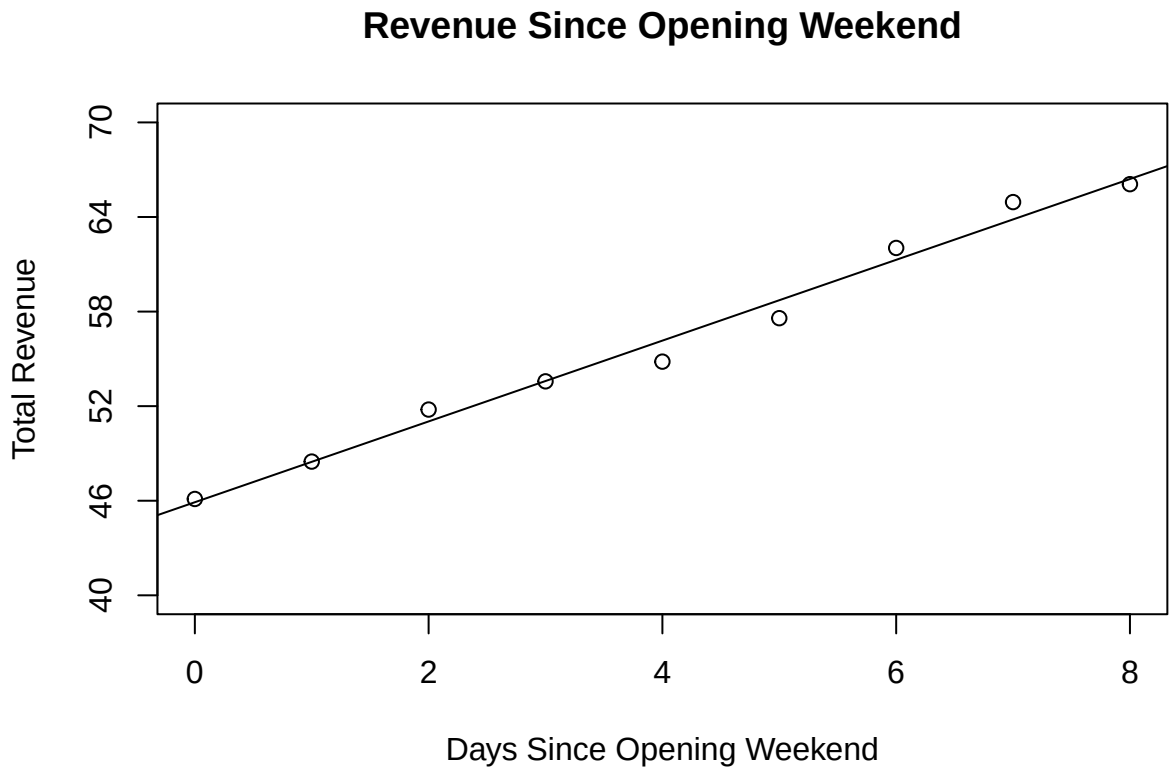
```
# Fit model passed on days since opening weekend
sinceNov12Model = lm(TotalRevenue~DaysAfterOpeningWknd, data=sinceNov12)

# Summarize model
summary(sinceNov12Model)
```

```
##
## Call:
## lm(formula = TotalRevenue ~ DaysAfterOpeningWknd, data = sinceNov12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1333094  -327535    14504   751134  1098326
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          45905674     547891   83.79 9.08e-12 ***
## DaysAfterOpeningWknd  2563057     115080   22.27 9.30e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 891400 on 7 degrees of freedom
## Multiple R-squared:  0.9861, Adjusted R-squared:  0.9841
## F-statistic:   496 on 1 and 7 DF,  p-value: 9.301e-08
```

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

```r
# 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")
abline(milModel)
```




```r
prediction = 45905674+26*2563057
prediction
```

```
## [1] 112545156
```