Movie Buzz Breakdown🐝

Jennifer Warren

Presentation Date: 12/7/2021

Summary of Data

I found this data online because I don’t have any of my own data. So as a preface: this whole script is just me playing around with R and R Markdown!

Since I had free range on this project, I thought it would be interesting to look at the relationship between how much money a movie makes in its first U.S. box office run, and several different factors such as if the movie is a:

- sequel
- action movie
- comedy movie
- animated movie
- horror movie

as well how what the film’s rating and budget were, how many views it received on http://traileraddict.com (http://traileraddict.com), number of message board comments on https://https://www.comingsoon.net/ (https://https://www.comingsoon.net/), and the percentage of votes on Fandango that “can’t wait to see” the film.

However, this script focuses on a just few of these data. These are the codes for each data type that are analyzed:

- BOX = Gross ($) from U.S. box office first run
- MPRATING = MPAA Rating code, where 1=G, 2=PG, 3=PG13, AND 4=R
- BUDGET = Production Budget in Millions ($)
- ADDICT = Number of Trailer Views on traileraddict.com
- CMNGSOON = Number of message board comments at comingsoon.net
- CNTWAIT3 = Percentage of Fandango votes that can’t wait to see the film

Importing Data

```
df_movies <- read_csv('C:/Users/091wa/Documents/5th year/EconS 523 (Data)/Final Presentation/movie_data.csv')

df <- df_movies %>%
  select(BOX, MPRATING, BUDGET, SEQUEL, ACTION, COMEDY, ANIMATED, HORROR, ADDICT, CMNGSOON, CNTWAIT3)
```

Raw Data

This is what the data from the Excel file I used looks like:

```
print(df)
```

```
## # A tibble: 62 x 11
##         BOX MPRATING BUDGET SEQUEL ACTION COMEDY ANIMATED HORROR ADDICT CMNGSOON
##       <dbl>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>    <dbl>  <dbl>  <dbl>    <dbl>
##  1 19167085        4   28        0      0      1        0      0  7860.       10
##  2 63106589        2  150        1      0      0        1      0  5737        59
##  3  5401605        4   37.4      0      0      1        0      0   850        24
##  4 67528882        3  200        1      1      0        0      0 15326        93
##  5 26223128        2  150        0      0      0        1      0  4574.       30
##  6 69637740        3   37        0      0      0        0      0 33324       533
##  7 14800723        3  130        0      0      0        0      0  3890.       20
##  8 31069826        3   80        0      0      1        0      0  2340.        6
##  9 12063452        3   40        1      1      0        0      0  3678        75
## 10  4271451        4   35        1      1      0        0      0  3586.      419
## # ... with 52 more rows, and 1 more variable: CNTWAIT3 <dbl>
```

Shortened Data for Summary

I didn’t want to use that much information though, so I shortened it:

```
df1 <- subset(df, select = -c(SEQUEL, ACTION, COMEDY, ANIMATED, HORROR))
print(df1)
```

```
## # A tibble: 62 x 6
##         BOX MPRATING BUDGET ADDICT CMNGSOON CNTWAIT3
##       <dbl>    <dbl>  <dbl>  <dbl>    <dbl>    <dbl>
##  1 19167085        4   28    7860.       10     0.49
##  2 63106589        2  150    5737        59     0.79
##  3  5401605        4   37.4   850        24     0.36
##  4 67528882        3  200   15326        93     0.76
##  5 26223128        2  150    4574.       30     0.53
##  6 69637740        3   37   33324       533     0.77
##  7 14800723        3  130    3890.       20     0.49
##  8 31069826        3   80    2340.        6     0.63
##  9 12063452        3   40    3678        75     0.59
## 10  4271451        4   35    3586.      419     0.62
## # ... with 52 more rows
```

Using the “subset” function, I created a new data frame with only the columns that I wanted to use later on. I used the “-c” to delete the data columns I didn’t want and left everything else alone.

Summary Stats for Shortened Data

I didn’t really want to see an average for all the ones and zeros in the five columns that I deleted (whether the movie was action, comedy, animated, horror, or a sequel). This table is much shorter and easier to digest that it otherwise would have been!

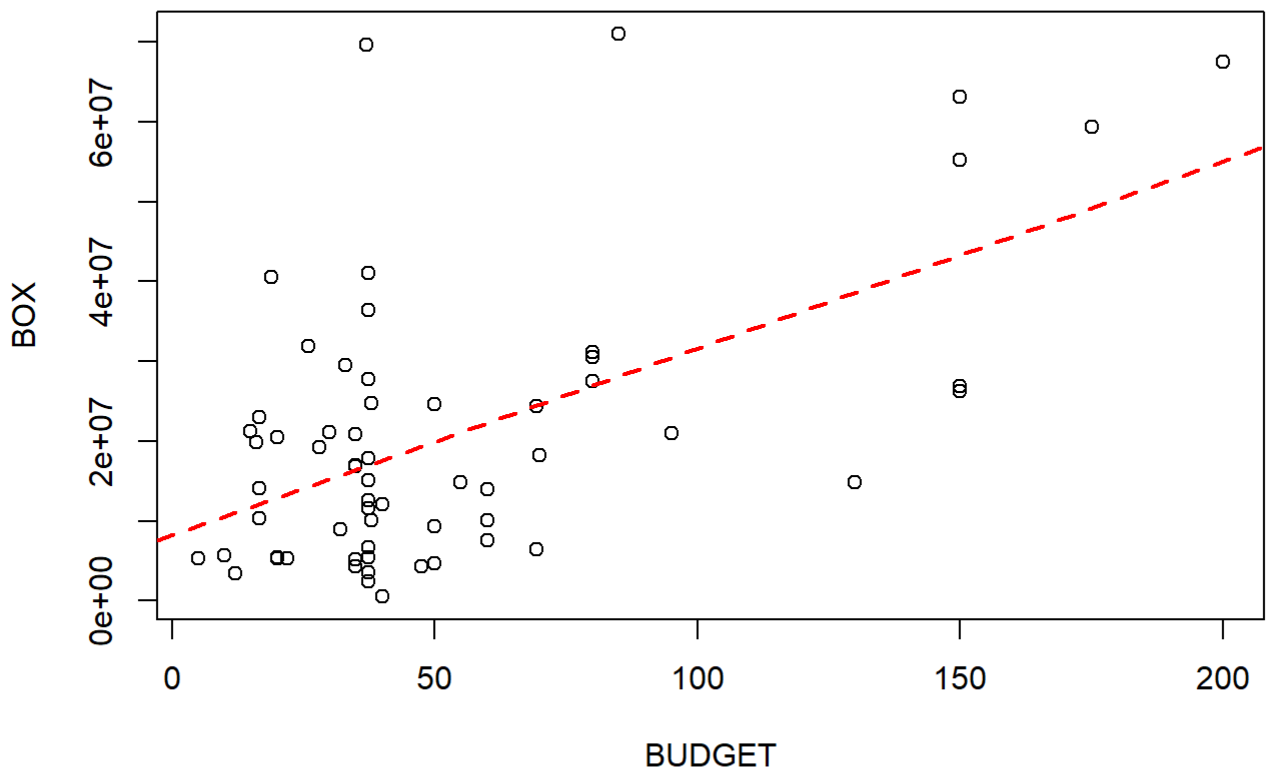
```
ss_movies <- sapply(df1,
                    function(i) c(mean(i), min(i), max(i), sd(i))) %>%
  data.frame() %>%
  round(digits = 2)

row.names(ss_movies) <- c('mean', 'min', 'max', 'sd')
ss_movies %>% kable(caption = 'Summary Statistics')
```

Summary Statistics

	BOX	MPRATING	BUDGET	ADDICT	CMNGSOON	CNTWAIT3
mean	20720651	2.97	53.29	5933.81	78.21	0.48
min	511920	1.00	5.00	568.00	2.00	0.15
max	70950500	4.00	200.00	45865.69	594.00	0.79
sd	17492443	0.81	42.87	7674.61	124.55	0.16

Correlation between Budget and Box Office $$

```
attach(df)
movies <- lm(BOX~BUDGET)
plot(BUDGET,BOX)
abline(coefficients(movies), lwd=2, lty=2, col='red')
```



Wow! Imagine that! Movies that have higher budgets make more money! The theory has been confirmed.

And who doesn't love a good summary stats table? Here it is:

```
summary(movies)
```

```
## 
## Call:
## lm(formula = BOX ~ BUDGET)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -23844872 -10060273  -3032775   7565032  52722776 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8269445    2944005   2.809   0.0067 ** 
## BUDGET        233663      43183   5.411 1.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14460000 on 60 degrees of freedom
## Multiple R-squared:  0.3279, Adjusted R-squared:  0.3167 
## F-statistic: 29.28 on 1 and 60 DF,  p-value: 1.153e-06
```

We can see that a movie’s budget is not the main reason for its box office revenue, but it can take some credit for why a movie makes the money it does.

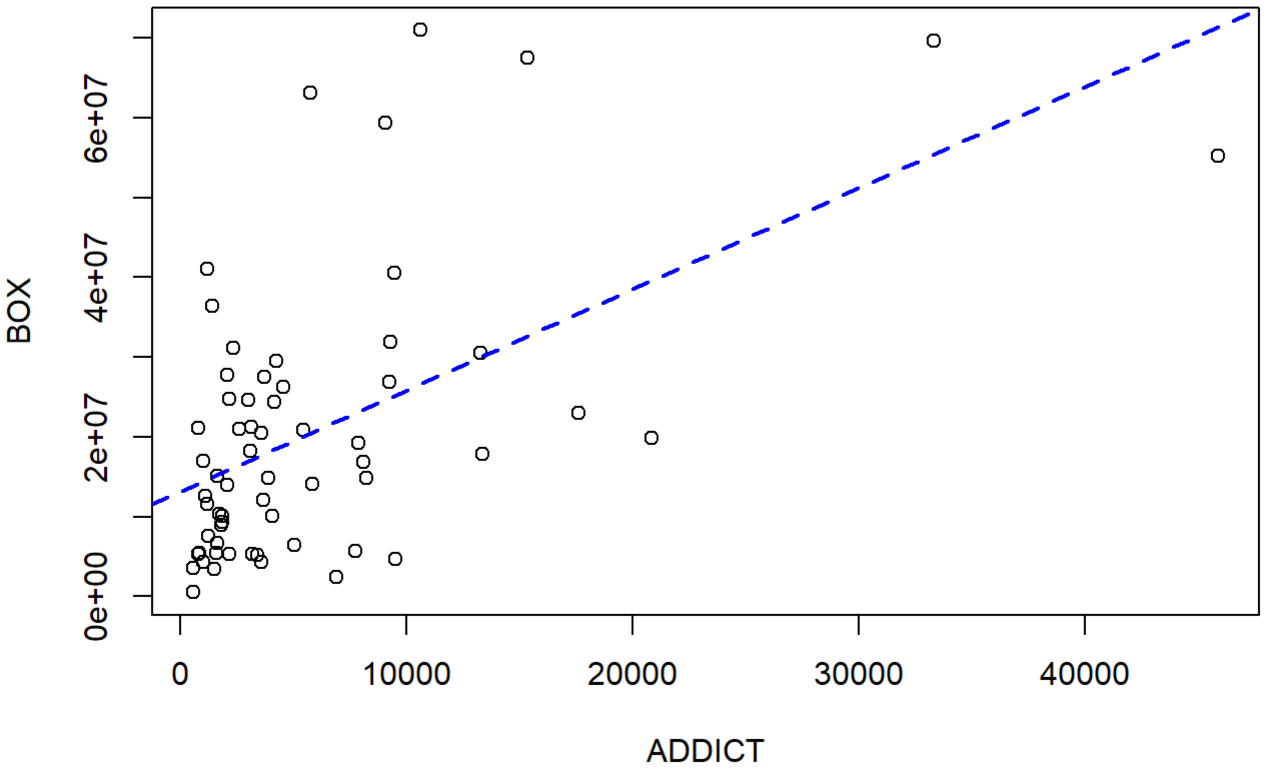
Correlation between Trailer Views and Box Office $$

Now let’s look at how the amount of views a movie trailer recieved corresponds to how much money it made:

```
attach(df)
```

```
## The following objects are masked from df (pos = 3):
## 
##     ACTION, ADDICT, ANIMATED, BOX, BUDGET, CMNGSOON, CNTWAIT3, COMEDY,
##     HORROR, MPRATING, SEQUEL
```

```
movies2 <- lm(BOX~ADDICT)
plot(ADDICT,BOX)
abline(coefficients(movies2), lwd=2, lty=2, col='blue')
```

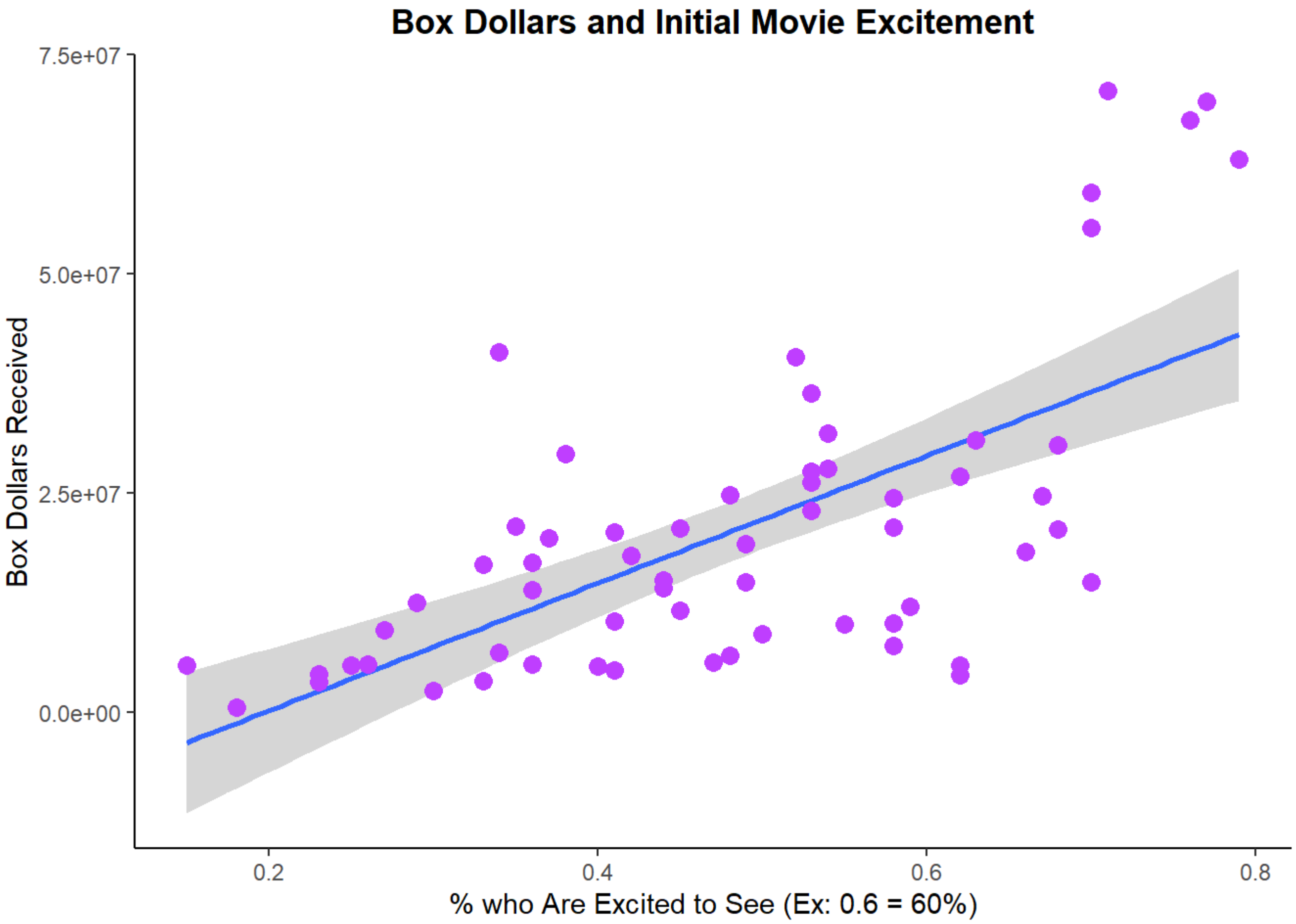


This involves the same steps as the last graph, I just wanted to switch up one of the variables.

Correlation between % Who “Can’t Wait to See” and Box Office $$

```
ggplot(df, aes(CNTWAIT3, BOX)) +
  theme_classic() +
  geom_smooth(method = 'lm') +
  geom_point(color = 'darkorchid1', size = 3) +
  labs(title = 'Box Dollars and Initial Movie Excitement',
       x = '% who Are Excited to See (Ex: 0.6 = 60%)',
       y = 'Box Dollars Received') +
  theme(plot.title = element_text(face = 'bold', hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



That’s all Folks!

"That's all Folks!"

```
```