DATE: November 6th, 2020
TO: Chris Acme
FROM: Rebecca Redmond, Wendy Merchant, George Tavlarios, John Trotto
SUBJECT: IMDB Project Update

We are writing to update you on our statistical findings from our proposed project on movie data and revenue. Our efforts have not changed much since our initial proposal, but we are focusing our analysis on revenue and how specific variables or the interaction between variables impacts revenue. Since our project proposal, the group has conducted numerous hypothesis tests to analyze the data. We have focused our preliminary analysis on conducting tests related to revenue and genre, runtime, and director. We used two ANOVA tests for genre and runtime and a t-test for director.

Initial Analysis

Our initial analysis consisted of viewing the relationships between revenue and genre, director, and runtime. We also conducted two ANOVA tests with revenue, genre, and runtime. The first was to view whether there are any statistically significant differences in revenue between movie genres, and the second included genre and runtime to view their interaction. The first ANOVA test showed there were significant differences, and a Tukey test showed that the genres with the differences were animation, action, and adventure. The second ANOVA test showed that though genre and runtime were both significantly different on their own, the combined term was not statistically significant. To analyze the relationship between director and revenue, we made a new binary variable for directors to determine if they were a top director or not. This was done by tallying each director name listed and taking the top 15 directors with the most tallies and specified them as a "top director." If the movie was directed by one of the top directors, there is a 1 under the Top_Director column of the data, and a 0 if not. This allowed us to filter the data by top directors and conduct a two-sample t-test to observe if the revenue of movies created by top directors differed (H_A). The p-value was .0004, so we rejected the null and there was statistically significant evidence that there is a difference in average revenue.

Business Relevance:

After making these changes to the data, we believe our team can draw clearer conclusions from the data set about it's variables and effects on revenue. The main question we seek to answer is which variables best predict revenue for popular movies? Do ratings or metascore have a stronger relationship with revenue? Are popular directors making movies that produce the most revenue? Do genre and runtime have a relationship with the movie's revenue? We have refocused our analysis on the main point of seeing which variables impact revenue the most, and slightly reframed our questions to reflect this goal.

Preliminary Analysis Findings November 6th, 2020

Data Source: IMDB data from 2006 to 2016

Data Summary:

Our data set consists of the 1,000 most popular films from 2006-2016 on the movie website IMDB. The data contains the following variables: genre, director, actor, runtime, rating, metascore, and box office revenue. The data set is complete. Slight alterations were made to the categorizing of genre and director for each movie. A single film can be categorized into multiple genres. We thought it would be best to categorize each movie by only one genre, so we categorized each film by the first genre listed. We also made a binary variable for directors to determine if they were a top director or not. This was done by tallying each director name listed and taking the top 15 directors with the most tallies and specified them as a "top director." If the movie was directed by one of the top directors, there is a 1 under the Top_Director column of the data, and a 0 if not. Other main variables we are analyzing are runtime, rating, and metascore.

Summary of Findings:

Based on our data, we wanted to analyze which factors best predict higher revenue for a film. Our preliminary analysis looked at the relationships between revenue, genre, and runtime, specifically how the latter two variables relate to revenue. The One-Way ANOVA test between revenue and genre had a p-value of almost zero, meaning at least one genre has a different average revenue from the others. On viewing the Tukey Test, the only statistically different genres were animation, action, and adventure. On the two-way ANOVA, both genre and runtime were individually significant, but the interaction term had a p-value of 0.674 and therefore was not significantly different for revenue.

Methods and Model Output:

Our team conducted a one-way ANOVA test with our dependent variable as revenue and our grouping variable as the genre of the film. Once this was completed, we also conducted a two-way ANOVA test with runtime as another grouping variable. This allowed us to see each variable's individual effect on revenue as well as the interactions between the two and its combined effect.

```
                    Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(Genre_1)  12 1757510  146459   16.71 <2e-16 ***
Residuals          859 7528477    8764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
128 observations deleted due to missingness
```

```
                                          Df  Sum Sq Mean Sq
as.factor(Genre_1)                        12 1757510  146459
as.factor(Runtime_Minutes)                91 2117162   23266
as.factor(Genre_1):as.factor(Runtime_Minutes) 267 1820982    6820
Residuals                                 501 3590334    7166
                                          F value Pr(>F)
as.factor(Genre_1)                         20.437 <2e-16 ***
as.factor(Runtime_Minutes)                  3.247 <2e-16 ***
as.factor(Genre_1):as.factor(Runtime_Minutes)   0.952  0.674
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
128 observations deleted due to missingness
```

We also conducted a two-sample, two-tailed t-test to determine if the mean revenue for movies with a "Top Director" differed from movies without one, where:

$$H_0: \mu_{\text{Top Director}} = \mu_{\text{Not Top Director}}$$
$$H_A: \mu_{\text{Top Director}} \neq \mu_{\text{Not Top Director}}$$

With a calculated p-value of 0.0003859, we can reject the null hypothesis. The difference in revenue between the data's top directors and the data's average directors is significant.

```
        Welch Two Sample t-test

data:  Revenue_Millions by Yes_Top
t = -3.6932, df = 87.028, p-value = 0.0003859
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -95.24641 -28.59671
sample estimates:
mean in group 0 mean in group 1
       77.20449       139.12605
```