Max Bard, Micah Lee, Arya Patel, Yash Subrahmanyam

Professor: Dr. Orkun Baycik

BA472 Project Report

Section 1. Introduction

Our project focuses on The Walt Disney Company, with a premium placed on the movies produced by Walt Disney Studios. In this project, we utilized two datasets. One of the datasets, “Disney Character Success,” was obtained from Kaggle and includes box office data up until 2016. Though this dataset is titled “Disney Character Success,” we decided to take our project in another direction as the data included was useful for different methods of interpretation. We specifically looked at the file “disney_movies_total_gross.csv” to gauge different factors of movies and how they affected revenue. Table 1 describes the variables included in this dataset.

Table 1: Variable Description of Total Gross Revenue from Disney Movies

Variables	movie_title	release_date	genre	MPAA_rating	total_gross	inflation_adjusted_gross
Description	Movie title	Movie release date	Movie Genre	Film ratings (e.g., G, PG, PG13, R)	Dollar amount of gross revenue	Adjusted dollar amount of gross revenue based on inflation

Another file we studied, described in Table 2, details how different factors affect viewers' preferences of villains or heroes in a given movie. This information could provide Disney with valuable insights into how they should tailor the marketing of their movies to the target audience's age range.

Table 2: Variable Description of Hero vs Villain Dataset

Variables	Date of Survey	Age	Retired	Have Kids Younger than 7	Preference
Description	Date survey was collected	How old the participant in the survey is	Indicates whether the participant is retired or not	Indicates whether the participant has kids younger than the age of 7 or not	Indicates whether the participant prefers the villain or the hero

We decided to focus on the total gross revenue of Disney because Disney is currently performing poorly on the stock market (Sharma, 2022). This decline in shares is attributed to systemic instability in the financial market as well as lackluster releases (e.g., Eternals). Diagnosing this instability would allow Disney to revise their approaches to movie releases, and eventually perform better in future.

The second dataset will allow Disney to judge which characters to develop for a specific movie. For instance, if a movie is being marketed towards an older demographic, the company would benefit from understanding if this demographic is more likely to watch the movie if the hero is advertised to be likable, or if the villain is presented as more evil.

Data Preprocessing

The data in both datasets had to be cleaned. For the data referenced in Table 1, many values were missing. We addressed this issue by using the .dropna() function, then by using the .reset_index() to fix discontinuous indexing. Additionally, the variable we were most interested in, *inflation_adjusted_gross*, contained records of strings, rather than floating point values, which we could work with mathematically. To fix this issue we used the .replace() function to replace any instances of '*$*' or '*,*' with nothing. We then used pd.to_numeric() to finalize the conversion of strings to floats. For the data reference in Table 2, we added a variable called *Bi_Pref*. This variable simplified people's preferences towards heroes or villains. The value "1" to signify a preference for heroes and "0" was used to signify villains. The data in the variable *Kids_younger_than_7* were similarly changed to integer values (henceforth referred to as "ints") as well.

The overarching goal of the project was to examine different factors which impact Disney's revenue. These factors primarily included MPAA ratings and character preferences. We achieved this goal using a range of testing methods, which we detail in Section 3.

Section 2. Research Questions

To achieve the goals listed above, our research questions were as follows:

Is there a difference between the inflation adjusted gross revenue of Disney movies released from 1989-1999 and movies released from 2000-2016?

Throughout the years, Disney has gone through multiple changes in leadership, eras and trends. This has the potential of causing differences in revenues in certain times. To test this difference, we utilized the inflation adjusted gross income of Disney movies. The data was split into two groups, the first of which is marked by the *Disney Renaissance* era. This era spanned the years 1989-1999, where Disney established itself as an animation powerhouse through the release of wildly successful movies such as Aladdin and Lion King, which many today consider classics. The second group included movies released from 2000-2016, the *Post Renaissance* and *Revival* eras. This research question lends insight into which eras Disney should focus their content on in order to maximize their total revenue. Disney can use certain key features which marked these eras in upcoming movies.

Are there differences in inflation adjusted gross revenue by Disney movie's MPAA ratings?

Different movie ratings attract different audiences. To test this in Disney's case, we referred to MPAA ratings and the corresponding inflation adjusted gross income for movies. We used the MPAA ratings G, PG, PG-13 and R. In order to eliminate outliers, we removed movies which fell into the *Not Rated* category, since it is not a classified rating. Understanding whether these ratings cause a difference in revenue can be an important factor in determining whether the company should perform additional tests. These tests could reveal which specific MPAA ratings bring in a higher revenue.

Which MPAA ratings have a greater impact on Disney's inflation adjusted gross revenue?

In general, a movie's MPAA rating is a good measure of the age range of the movie's audience. For instance, a G rated movie will have a greater number of children as their audience, while a R rated movie will have more adults. To test this impact, we used the MPAA rating and inflation adjusted gross income for movies. We used the MPAA ratings G, PG, PG-13 and R.

As a company that has a brand, Disney must recognize which age group their primary audience falls under. This information is vital in understanding which age group is bringing in the most revenue. The resulting data can guide Disney's approach to gearing their content, in the case that they want to increase revenue.

How do Disney's movies' MPAA ratings affect the inflation adjusted gross revenue?

To approach this research question, we tested a specific range of revenue and examined whether the MPAA rating impacted these ranges. We used the MPAA ratings and inflation adjusted gross incomes for movies. We then divided our data into separate categories according to their MPAA ratings. We then further split the rating categories into 2 more categories: an inflation adjusted gross income of above 100 million dollars, and below 100 million dollars.

Disney would benefit from performing this test; it would guide them to study other factors which may impact revenue levels, with a given rating. This additional testing would also allow Disney to study characteristics of their movies that result in higher popularity and increased revenue.

Is being retired a good indicator of whether or not people prefer heroes or villains in Disney movies? If they have a child younger than 7, does that affect their preference?

Disney makes movies for a wide variety of audiences. They produce Pixar movies, but also produce more mature movies such as *Judge Dredd*. Disney should hence develop certain characters for particular audiences.

To answer this question, we used the Hero vs Villain dataset, summarized in Table 2. The preference was our outcome variable, while the person's retired status and whether they had kids younger than 7 were our predictor variables. This test can be beneficial to Disney; it would indicate which age groups prefer heroes and villains. After considering external factors affecting preferences, Disney could use these results to determine whether to advertise hero or villain protagonists of movies to the target audience.

Does having young children cause viewers of Disney movies to prefer heroes or villains over the other?

While Disney focuses its efforts mostly on children's movies, the company should also consider parents in the audience who watch movies with their children. Parents may have different preferences in comparison to their children, such as favoring the villain over the hero. To answer this question, we used the Hero vs Villain dataset as seen in Table 2. The outcome variable was the preference for the hero or villain, and the treatment variable surveyed whether or not the participant had a child under the age of 7.

This experiment could have beneficial implications for Disney. The company could alter its marketing strategy depending on the results. To maximize the application of this experiment's results, Disney producers would not limit themselves to marketing; for instance, the producers could also make the villain more appealing for parents.

Section 3. Experiment and Analysis

Descriptive Statistics for the Total Gross Dataset

This dataset contained 513 records, with a mix of strings and values. The mean of *inflation_adjusted_gross* was around $127,594,100, with a minimum value of $2,984 and a maximum value of around $5,228,953,000. Of the Disney movies accounted for in this dataset, the prominent genre was comedy with 162 entries, followed by adventure and drama with 119 and 103 entries respectively. We reported 185 PG and 141 PG-13 records amongst the movies, along with surprisingly similar numbers of R and G movies, with 98 and 86 records respectively.

Two-Sample Two-Tail Hypothesis Test & Effect Size

We conducted a two-sample t-test to answer the question of whether a difference in the inflation adjusted gross revenue between Renaissance era Disney movies and Post- Renaissance or Revival era Disney movies exists. To perform this test we made use of the *release_date* and *inflation_adjusted_gross* variables indicated in Table 1. Before proceeding with the test, we first had to confirm assumptions on the data.

1. Each observation was independent from all others, as each observation is related to its own movie
2. A simple random sample was obtained with the .sample() function
3. The data was sufficiently large to invoke the central limit theorem, since our datasets for each era held over 200 records.

To conduct the two-sample t-test, we created two new dataframes. The first dataframe contained only movies released from 1989-1999 and the second contained movies released from 2000 onwards. We tested for equality of variances using Levene's Test on the dataframes, and obtained a p-value of 0.006. This p-value was small enough to conclude that the two dataframes did not have equal variances. We were able to conduct our t-test with this knowledge; we set equal_var = False, and the test returned a p-value of 0.002. Since this p-value was less than 0.05, we rejected the null hypothesis and concluded that there exists a significant difference in gross revenue between the Disney Renaissance, and the Post Renaissance and Revival eras.

To examine effect sizes, we set a hypothesized value of $50,000,000, a value we would consider to be a good gross revenue for a "wildly successful" movie. Using this value on the *inflation_adjusted_gross* data from the Renaissance as well as the Post Renaissance and Revival eras, we found effect sizes of 0.315 and 0.483 respectively. This led us to believe that we collected a decent amount of data for the Renaissance era, and could afford to potentially draw a smaller sample size for the Post Renaissance era.

We probed our knowledge that there was a significant difference between the eras further, and found that the Post Renaissance and Revival eras had a higher mean of $113,053,357, compared to the Renaissance era's mean of $81,050,778. This information led us to the conclusion that Disney movies' gross revenues are on a positive path. We attributed this climb in gross revenue to the fact that Disney acquired multiple different studios, which widened the range of movie audiences. This resulted in a higher gross revenue regardless of movie quality, simply because more people gained interest in watching to begin with[1].

[1] We stand to believe that Disney should still look to produce higher quality films in order to not repeat the disaster that was *Eternals*.

One-Way ANOVA & Tukey-Kramer

We sought to answer the question of whether a movie's MPAA rating had an impact on its revenue. With reference to Table 1, we used two variables for our test, *MPAA_rating* and *inflation_adjusted_gross*. We ran assumption checks on our data:

1. Using the Anderson-Darling test, we tested if the sample was drawn from a population that follows a particular distribution. Our results were: [15%, 10%, 5%, 2.5%, 1%]. Since these significance levels indicated a normal distribution, we found the data usable for ANOVA testing.
2. We tested whether our input samples were from a population with equal variances using the Levene test. Our results had a p-value of about 1.00e-06, which violated our assumption. To account for this violation, we carried out a Tukey-Kramer test. This test is appropriate for data sets with unequal variances, unlike a Tukey test.
3. The samples were taken independently, so there was no indication that this assumption was violated.

After running the One-Way ANOVA test, we concluded that a movie's MPAA rating had an impact on its revenue. In order to verify the statistical significance of our test, we obtained the p-value of roughly 2.25e-07. This p-value is much smaller than our significant value of 0.01; we hence confirmed that the results of this test were *highly* statistically significant. We therefore rejected the null hypothesis, and can confidently state that there is a difference in the revenue as the movie rating differs.

We then proceeded with the Tukey-Kramer test to answer the question of which MPAA ratings had a greater impact on Disney's inflation adjusted gross revenue. Figure 1 documents our findings:

	A	B	mean(A)	mean(B)	diff	se	T	p-tukey	hedges
0	G	PG	2.912610e+08	1.026074e+08	1.886536e+08	3.809458e+07	4.952242	5.960801e-06	0.644522
1	G	PG-13	2.912610e+08	1.056656e+08	1.855954e+08	3.993629e+07	4.647286	2.544025e-05	0.633726
2	G	R	2.912610e+08	5.741288e+07	2.338481e+08	4.312809e+07	5.422176	5.460066e-07	0.797856
3	PG	PG-13	1.026074e+08	1.056656e+08	-3.058211e+06	3.263075e+07	-0.093722	9.997049e-01	-0.010453
4	PG	R	1.026074e+08	5.741288e+07	4.519453e+07	3.646766e+07	1.239304	6.021039e-01	0.154422
5	PG-13	R	1.056656e+08	5.741288e+07	4.825274e+07	3.838751e+07	1.256991	5.908133e-01	0.164790

Figure 1: Tukey-Kramer Test

Figure 1 suggests G rated movies bring in the highest revenue in *all* instances. Line 0 indicates G rated movies result in a different revenue at about $291,260,995, compared to PG movies at roughly $102,607,407. Thus, in this case, G rated movies result in a higher revenue. The p-value is about 5.90e-06 and our standard error is about 3.81e-07; these are less than the significant value of 0.01 which we used, so our results are statistically significant. Line 1 suggests G rated movies bring in a different revenue at about $291,260,995 in comparison to PG-13 rated movies at roughly $105,665,618. Therefore, G rated movies result in a higher revenue. The p-value is about 2.54e-05 and our standard error is about 3.99e-07; since these are significantly smaller than our significant value, our results are statistically significant. Finally, Line 2 indicates G rated movies bring in a different revenue at about $291,260,995 compared to R rated movies at roughly $57,412,878. G rated movies once again result in a higher revenue. The p-value is about 5.45e-07 and our standard error is about 4.31e-07. Since these values are notably less than our significant value, our results are statistically significant. In all our cases, our results have a greater than 99% chance of being true. Consequently, we reject the null hypothesis. A difference in the revenue exists as the movie rating differs.

Overall, we proved statistical significance with these 3 cases and came to the conclusion that G rated movies bring in a higher revenue than PG, PG-13 and R ratings. As previously stated, we have an unequal distribution of movies that fall into each category. In this case, G, PG, PG-13 and R rated datasets have 86, 185, 141 and 98 movies respectively. Remarkably, although this dataset of Disney has the lowest number of movies with a G rating, it brings in a higher revenue than the PG, PG-13 and R rated movies.

Of course, further research and testing on other factors, such as movie characters or movie themes, would need to be conducted in order to prove whether the rating has a direct correlation to the revenue. However, looking at Disney as a company, our results make sense. With Disney's amusement parks and whimsical characters, they primarily target young children. Additionally, G-rated movies are made for young kids who need their parents to take them to watch said movies. Taking these two factors into consideration, Disney's G-rated movies would generally attract a large audience where the parents add to the already large pool of children.

Heterogeneity

In order to establish the effect of a movie's MPAA rating on its inflation-adjusted revenue, we ran a heterogeneity test using the *MPAA_Rating* and *inflation_adjusted_gross* variables, indicated in Table 1. To perform this heterogeneity test, we divided the movies by their MPAA ratings, by G, PG-13, PG and R. We then decided to test whether these ratings would have implications on whether the revenue generated, adjusted for inflation, would exceed or fall short of $100,000,000. As mentioned in *"Descriptive Statistics for the Total Gross Dataset"* in Section 3, the mean of the inflation-adjusted gross was roughly $127,000,000, and hence the value of $100,000,000 set an appropriate standard for a successful movie.

We used a *Chi-squared* test to perform this experiment. We operated this test under the assumption that the variables used were categorical. We also assumed that the variables used came from independent random samples, without which the heterogeneity test would not be possible. Using a significance level of 0.1 rather than 0.05, we found the P-value to be 1.221*10^-15, and we therefore rejected our null hypothesis. We hence concluded that the MPAA ratings did have an affect on Disney's revenue and a movie's MPAA rating. This application would prove to be useful if Disney wanted to decide their MPAA rating of a movie with a target revenue in mind; for instance, if a proposed project could be rated either PG or PG-13, Disney could examine which rating would increase their likelihood of achieving a target revenue of $100,000,000 for the movie.

Descriptive Statistics for Hero vs Villain Dataset (Regression & ATE and DoWhy)

The dataset contained 452 instances, and all the data in this dataset was binary after cleaning except for age. The mean age of the dataset was around 45, and the youngest member was 18 years old while the oldest was 70 years old. The data was fairly balanced, with 211 people preferring heroes to villains and 198 people having children under the age of 7. One variable which was particularly imbalanced was the *Retired* column, which only contained 46 instances of people in retirement.

Regression in Experiments & ATE

Using the Hero vs Villain dataset as seen in Table 2, we answered the question of whether or not being retired inclines people to prefer the hero or villain in Disney movies. After running a linear regression analysis on these two variables (*Bi_Pref* and *Retired*), we found retired folks to prefer the villain, apparent in Figure 2 . The coefficients shown in Figure 2, which are the Average Treatment Effect (henceforth referred to as ATE), are particularly important. For one, the row labeled “const,” or constant, indicates the intercept of the equation. Since *Bi_Pref* is a binary variable, the intercept falling close to 0.5 makes sense, as the answer will either be 0 or 1. This demonstrates that the model lies between these two binary values until another variable is added. From the row labeled “Retired,” we notice the ATE is -0.2051. The value of 0 in the column *Bi_Pref* suggests that the participant prefers villains; this signifies that being retired is an important variable for those who prefer the villain. The *Retired* variable is also binary, hence the value of the ATE is significant. This results in the regression equation 0.4877 - 0.2051x. To explain the equation, if the participant identifies as retired, x will equal to 1, meaning the result will be 0.2826. Otherwise, the equation will stay at 0.4877. The standard error of 0.077 suggests that we are confident that the actual value of the ATE for *Retired* lies between -0.1281 and -0.2821. The standard error is rather large, considering the ATE is close to 0 already. Despite this relatively wide range, we can safely assume that retired folks will prefer the villain. The p-value in Figure 2 was lower than our alpha of 0.05; hence, we rejected the null hypothesis of retirement having no effect on the preference between the antagonist or the protagonist.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Bi_Pref   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     7.060
Date:                Wed, 07 Dec 2022   Prob (F-statistic):            0.00816
Time:                        19:55:21   Log-Likelihood:                -323.54
No. Observations:                 452   AIC:                             651.1
Df Residuals:                     450   BIC:                             659.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4877      0.025     19.807      0.000       0.439       0.536
Retired       -0.2051      0.077     -2.657      0.008      -0.357      -0.053
==============================================================================
Omnibus:                     2003.839   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               71.185
Skew:                           0.120   Prob(JB):                     3.49e-16
Kurtosis:                       1.071   Cond. No.                         3.35
==============================================================================
```

Figure 2: Regression Analysis for Character Preference and Retired Status

In the interest of attempting to improve our model, we added a covariate to the regression model. The variable we decided to add was *Kids_younger_than_7*. After running the regression with these two variables, we noticed that the new variable added little to the model. The ATE for *Kids_younger_than_7* was -0.013, suggesting that people with children under 7 years old slightly prefer the villain in movies. We noticed this value is relatively close to 0. The standard error of 0.047 suggested that the variable could lie between -0.060 or 0.034. Hence, this variable could indicate preference towards either the villain or the hero. The *Retired* ATE and the intercept stayed relatively equal, only increasing slightly. This left us with the equation 0.4932 -0.2036x - 0.0130x. The p-value also increased from 0.008 in the first regression to 0.0292. While nonetheless still statistically significant, this increase in p-value suggests that the statistical significance has decreased.

Some takeaways from this model are that retired people prefer villains in Disney movies, while people with children under 7 years old are mostly indifferent to the antagonist or protagonist. There are several implications which subsequently arise: when marketing a new movie, for instance, Disney could attempt to make the villain more appealing if the movie is directed towards older audiences.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Bi_Pref   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     3.560
Date:                Wed, 07 Dec 2022   Prob (F-statistic):             0.0292
Time:                        21:04:55   Log-Likelihood:                -323.50
No. Observations:                 452   AIC:                             653.0
Df Residuals:                     449   BIC:                             665.3
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   0.4932      0.032     15.503      0.000       0.431       0.556
Retired                -0.2036      0.077     -2.628      0.009      -0.356      -0.051
Kids_younger_than_7    -0.0130      0.047     -0.275      0.784      -0.106       0.080
==============================================================================
Omnibus:                     2004.814   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               71.137
Skew:                           0.120   Prob(JB):                     3.57e-16
Kurtosis:                       1.071   Cond. No.                         3.73
==============================================================================
```

Figure 3: Regression Analysis with Covariate

From Figure 3, one notices that the regression is run on “good controls.” The treatment variable, *Retired*, is not obstructed by the covariate, *Kids_younger_than_7*. *Kids_younger_than_7* is a good control because it is not affected by the treatment. Essentially, a participant may be retired without necessarily having children at all - let alone children under the age of 7.

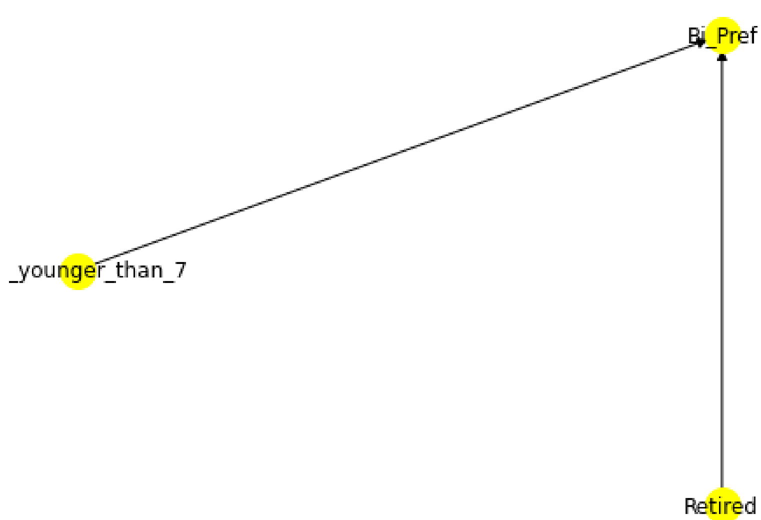

Figure 4: Causal Diagram for Regression Analysis

DoWhy Analysis

Using the Hero vs Villain dataset as seen in Table 2, we answer the question of if having children under 7 causes movie watchers to prefer the hero or the villain. We run DoWhy analysis on our variables to see the effects of this. To start off with, a causal diagram is created. A causal diagram shows the causal relationship between variables. The causal model, Figure 4, is based on prior assumptions. We assume having children under 7 as well as being retired is caused by age. These assumptions are fair to make, as the older someone is, the more likely they will have a child under 7 years old. As for being retired, the older a person is, the more likely they will be retired. We also assume that every predictive variable causes the preference of villain or hero.

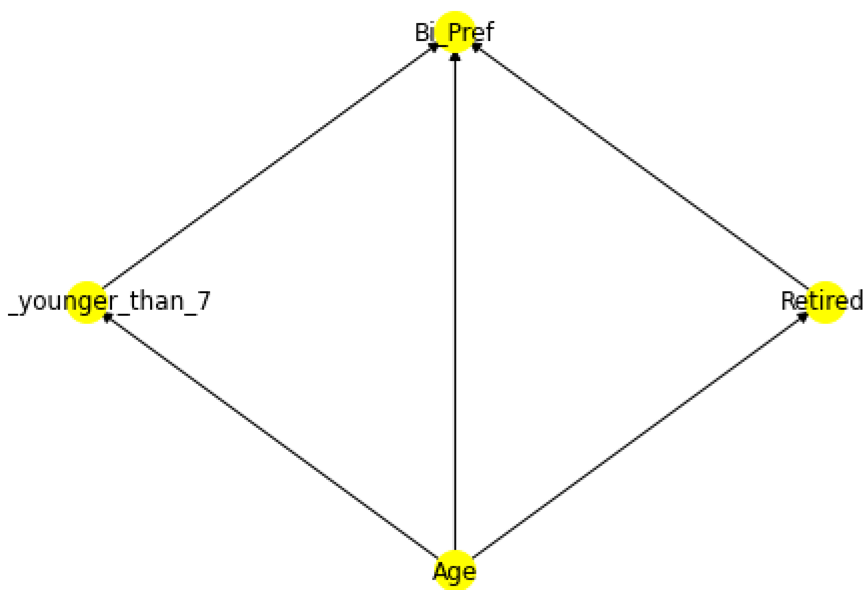

Figure 5: Causal Diagram for DoWhy Analysis

Once the causal diagram was made, we ran a DoWhy analysis. We decided to run a placebo treatment and a data subset validation. The placebo treatment replaces the true treatment variable with a randomly generated variable, while data subset validation checks if the estimated effect changes significantly when we replace the given dataset with a randomly selected subset. Refer to Table 3 for the results of the tests:

Table 3: Validation of Assumptions for DoWhy Analysis

	Estimated Effect	New Effect	p-value
Placebo Treatment	-0.0177	0	2.0
Data Subset Validation	-0.0177	-0.0188	0.4789

For the placebo treatment, we draw our attention to the new effect because if the assumptions are correct, the new effect should be close to 0. From Table 3, we note the new effect is indeed 0, meaning that the assumptions made were extremely accurate. The p-value obtained was higher than the alpha of 0.05. This meant that there were no problems with the estimate, and that our initial assumptions were valid. The new effect was also the important variable for data subset validation. Since we replaced the dataset with a random subset, the new effect would logically be similar to the estimated effect. As Table 3 indicates, the new effect was extremely close to the estimated effect with only a .0011 difference in the two effects. The p-value was also greater than the alpha of 0.05, so the estimates were correct and our initial assumptions were valid.

While our assumptions were correct, this variable is not very significant in determining whether a viewer would be more inclined to prefer the villain or the hero. The response variable, *Bi_Pref,* is binary; this change would mean viewers with children under 7 years old are slightly inclined to prefer villains, as 0 signifies a preference for villain. However, this result is not significant enough to call for any change. The results signify that parents are mostly indifferent to heroes or villains. This makes sense because the movies parents with children under the age of 7 years watch are most likely children's movies that their children want to watch, not what they would necessarily prefer.

Section 4. Conclusions and Discussion

Our research questions explored different variables which led to differences in revenue, with a particular emphasis on movies' MPAA ratings. We additionally looked at supplemental data in order to determine whether adults had a preference for heroes or villains. This supplemental data, foregrounded in the fifth and sixth research questions, tied into the rest of the data we made use of aptly. For instance, if Disney had an idea for a new movie catered to children, they would first determine whether the movie would be catered towards younger children or adolescents - which would then be used to establish the MPAA rating - following which they could evaluate the supplemental data to determine that children would take keener interests in the protagonist. In order to determine whether to cater the movie towards young children or adolescents, Disney could examine whether rating a movie G as opposed to PG or

PG-13 would result in higher revenue, especially over $100,000,000, with a mean of roughly $127,000,000 in mind.

All of our research questions were answered through the experiments we conducted, and some of the results obtained were astounding, suggesting strong implications for future Disney movies. One such finding was that Disney movies with a G rating occurred least frequently, but resulted in the highest revenue in comparison to other movies. We checked assumptions where appropriate and re-tested all results to ensure they were repeatable and reliable, and reflected upon whether our results made sense. In our Two-Sample Two-Tail Hypothesis Test and Heterogeneity tests, for instance, we checked assumptions to ensure we could proceed with these tests without our results being skewed. When we realized our assumptions were not appropriate, we adjusted to match these disparities - as documented in our Tukey-Kramer test.

On balance, the results we obtained do make sense and no irregularities were found when comparing findings with one research question to the next. Some other interesting findings were that retired consumers would generally prefer Disney villains. Combining this knowledge with our newfound findings of PG- or R-rated movie revenue suggests Disney could cater its next project specifically to retired adults, with an apt MPAA rating and emphasis on villains.

There are some limitations to our findings. For instance, the heroes and villains dataset was imbalanced. A more balanced data set would allow our findings to be more conclusive and ensure we do not overlook any information about the dataset, which otherwise may not be accounted for. In order to acquire a more balanced data set, we would have had to gather more samples which would enable us to conclusively state that the regression experiment would hold weight. While our results are important and provoking, their implications would be solidified with more data. Another limitation could be that we had to work with a large amount of missing values in the dataset featuring Total Gross Revenue which we had to eliminate. Due to time constraints, our best solution was to remove these data from the dataset completely, and there were not many other actions we could take to handle this data.

There is scope for further research involving Disney's revenue.. This research may include evaluating data sets pertaining to Disney+, Disney's streaming service. We could examine different factors which impact the revenue generated by the TV shows available on this service and compare them with the findings found from examining the movies as performed through these experiments. Such research would advise Disney on how to approach the content they produce on their streaming service, such as making certain movies available on Disney+, or removing existing movies which are not contributing productively to the revenue being generated.

Bibliography

CC0: Public Domain. (2022, October). *Disney Character Success*. Kaggle. Retrieved December 6, 2022, from https://www.kaggle.com/datasets/thedevastator/disney-character-success-a-comprehensive-analysi?select=disney-characters.csv

Vallat, Raphael. “Pingouin.pairwise_tukey.” *Pingouin.pairwise_tukey - 0.5.2 Documentation*, Pingouin, https://pingouin-stats.org/generated/pingouin.pairwise_tukey.html.

Sharma, Rakesh. “Disney (DIS) Falls after Disappointing Earnings Report.” *Investopedia*, Investopedia, 13 July 2022, https://www.investopedia.com/disney-falls-after-disappointing-earnings-5209283.

Supplemental Data provided by Professor Baycik