|  |  |
| --- | --- |
| Hulu Review 2022: Is Hulu On-Demand Worth It? | TV Guide - TV Guide  HULU  Show Popularity | Abstract  This project focuses on what aspects of our HULU data set affect show popularity.  Ceara Stewart, Kawshif Muhammed, Caitlin Rookey and Randy Garcia  IST687 Fall 2022 |

**Table of Content**

1. Introduction-----------------------------------------------------------------------------2

2. Business Questions--------------------------------------------------------------------2

3. Data Cleanse/Munge/Preparation ---------------------------------------------------2

4. Descriptive Statistics and Visualization -------------------------------------------3

5. Data Modeling ------------------------------------------------------------------------7

7. Appendix – Code ---------------------------------------------------------------------14

**Introduction**

This report presents the findings of the Top 1,000 Most Popular HULU Shows dataset consisting of 1000 rows and 181 columns consisting of various aspects of the HULU shows, such as episode count, network, and genre. The objective of this report is to look at the variables that affect a shows popularity to provide insight to HULU on what changes they can make to less popular shows to make them popular, and thus increase business. Our team developed various business questions to steer our data analysis direction. After reviewing and reading in the data, our team cleaned, munged, and prepared the data set to create data visualization plots to better illustrate our findings. Our team also developed models to analyze the data to identify attributes of various variables that have the most impact on show popularity.

**Business Questions:**

1. What makes a show “popular”? What factors make a show “popular”?
2. How does show popularity compare to other values within the dataset?
3. Are there correlations between variables that result in a show being “popular”?

**Data Cleanse/Munge/Preparation**

1. We began by reading in the data and linking it to a data frame name.
2. Due to the original dataset having 181 columns, we selected a few to keep in our analysis process to keep this report concise. We “deleted” all the columns except for #1, #2, #3, #6, #9, #10, #15, #16, #17, #21, #39, #40, and #41
   1. #1 Show ID – the unique id for each show.
   2. #2 Show Time – the time the show airs.
   3. #3 Show Conical Name – the full name of the show.
   4. #6 Show Episode Count – the number of episodes for each show.
   5. #9 Show Genre – the genre of each show.
   6. #10 Show Sub Genre – each possible genre of each show.
   7. #15 Show Name – the name of each show as it appears on air.
   8. #16 Show Rating – the rating from 1-5 of each show.
   9. #17 Show Season Count- the number of seasons for each show.
   10. #21 Show Type – the type of show.
   11. #39 Show Company ID – the unique id of the company where the show is aired.
   12. #40 Show Company Channel ID- the unique id of the channel the show is aired on.
   13. #41 Show Company Name – the full name of the company that owns the show.
3. The original dataset had longer and more confusing column names, so our team renamed them to be simplified and easier to read.
4. Next, our team removed any weird symbols from the showType column and removed the “~” symbol from the showSubGenre column, as well as converted the showTime into a readable time of show.

**Descriptive Statistics and Visualization**

**Histogram 1**

Show season count: This histogram shows that as season numbers increase, the overall episodes per season tends to decrease.

Chart, histogram

Description automatically generated

**Histogram 2**

Show Episode count: This histogram shows the general count of episodes. We can see that most shows remain in the lower episode count range.

Chart, histogram

Description automatically generated

**Histogram 3**

Show rating: Since our overall goal is to assess show rating, we created a histogram of show rating. Since the dataset we originally pulled from is the 1000 most popular shows on Hulu, we can see that show ratings remain at about a 3 or higher.

Chart, bar chart

Description automatically generated

**Plot 1**

Season Episode Count: This histogram plots show episode count based on show season count and show genre. For each genre, we can see what shows had the most seasons and most episodes. We can see a trend where, as the number of seasons increases, the number of episodes increases as well.

Chart, scatter chart

Description automatically generated

**Plot 2**

Season episode count 2: Since most shows that are popular have a lot of episodes, we also decided to condense our view and plot shows that had 250 episodes or more. With this condensed view, we can see that there is less of a pattern once we look at shows with more than 250 episodes.

**Chart, scatter chart

Description automatically generated**

**Plot 3**

Show Genre vs Show Rating: This plot shows us what shows within their respective genres received the highest show ratings. From this plot we can see that “Drama” and “Comedy” genres contained the most shows with the highest ratings.

**Scatter chart

Description automatically generated with low confidence**

**Plot 4**

Show Episode Count vs Show Rating: This plot shows us what episode counts have the highest show ratings. From this plot we can see a that as episode count increases, show ratings tend to also be higher.

**Chart, scatter chart

Description automatically generated**

**Plot 5**

Show Season Count vs Show Rating: This plot shows us what shows have the highest show ratings based on season count. From this plot we can see a that as season count increases, show ratings tend to also be higher.

Chart, scatter chart

Description automatically generated

**Plot 6**

Show Company vs Show Rating: This plot shows us what shows have the highest show ratings based on the company they belong to. From this plot we can see a that companies such as Nickelodeon, FOX and ABC, have more shows with higher ratings.

**Chart

Description automatically generated with medium confidence**

**Data Modeling**

**Linear Regression Model**

**Introduction:**

In total we ran seven linear regression models to rest the relationships between independent and dependent variables. Based on these models, we discovered what independent variables affected our dependent variable of show rating.

**Model 1**

In our first model, we put show rating as y, and show episode count as our x variable. We got an R2 of 0.0004533 and adjusted R2 of -0.0005482, showing us that there is not a strong, or significant relationship between show rating and show episode count.

Text, letter

Description automatically generated

**Model 2**

Our second model looks at show ratings as our y variable and show genre as our x variable. We got an R2 of 0.4843 and adjusted R2 of 0.4665, which, compared to our model 1, is statistically significant. We can take a closer look at our intercepts and see that certain columns are also more statistically significant than other columns. From that perspective, showGenreClassics and showGenreScience Fiction are less statistically significant than other aspects within show genre.

**Table

Description automatically generated**

**Model 3**

Model 3 looks at show rating versus show season count, being our y and x variables respectively. We get an R2 of 4.283e-07 and adjusted R2 of -0.001002, which, similar to model 1, is not statistically significant when we compare our x variable to our y variable.

**Text, letter

Description automatically generated**

**Model 4**

Our fourth model looks at show rating as our y variable and show company id as our x variable. This model produces an R2 of 0.00948 and adjusted R2 of 0.008488, which is not statistically significant, showing us that show company id does not affect the outcome of show ratings as much as other models.

Text, letter

Description automatically generated

**Model 5**

This model looks at show rating versus show company channel id, as our y and x variables. The R2 and adjusted R2 values of 0.01081 and 0.009818 are more statistically significant than other models. This shows, if we company model 5 to model 4, that show channel id has more of an effect on the outcome of show ratings than show company id does.

**Text, letter

Description automatically generated**

**Model 6**

Model six looks at show ratings as the y variable and show company name as our x variable. We got an R2 of 0.3755 and adjusted R2 of 0.3541, which is statistically significant. We can also look at the intercepts of the columns belonging to show company id and see that showCompanyNameMTV and showCompanyNameABC are among some of our statistically significant columns compared to showCompanyNameDisney Junior which is not statistically significant.

**Table

Description automatically generated**

**Graphical user interface, text, application, table

Description automatically generated**

**Table

Description automatically generated**

**Model 7**

Our final model is our multiple regression model looking at show rating, as the y variable, versus show company id, show company channel id and show company name as our x variables. This multiple regression model produces an R2 of 0.3755 and adjusted R2 of 0.3541. This happens to be the same values from our model 6, which gives an idea that show company name has more of an effect on show rating compared to the other variables. Looking at model 7 as a whole, it is still statistically significant, meaning our x variables together result in a large effect on the outcome of show rating. We can also look at the combined aspects of the variables to see that showCompanyNameHulu Original Series and showCompanyNameAdult Swim are among some of the more statistically significant columns within our variables.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Appendix – Code**

Our R code can be found in an attached pdf file.