TV Show Recommendations

Grace Bero & Abby Maahs

Researchers: Abby Maahs and Grace Bero

Research Focus: Statistics

School: Drake University

Presentation Type: PowerPoint

Abstract

The goal of our project was to find which country produces the most popular television shows. We judged popularity based on average ratings and number of views. The methods we used to research this problem were R studio and SAS studio using a CSV file.

Based on our data, we found that the United States produced the most popular television shows. The most popular show produced by the United States was ‘House of the Dragon’, followed by other well-known television shows like ‘Game of Thrones’ and ‘American Horror Story’. These U.S. produced shows had popularity ratings above 1,000 and vote counts close to 200,000. On the other hand, we found that Mexico had the lowest average popularity scores on produced television shows. Most shows produced in Mexico had popularity ratings and vote counts under ten.

Acknowledgements

We would like to thank our professors and colleges for helping us achieve the skills needed to fulfill the project requirements. We appreciate Professor Herath for his guidance and supervision which has provided us with the resources needed.

Abby Maahs did the SAS coding, and Grace Bero did the R coding.

Introduction

Over the past semester, we have been learning how to code with R and SAS. In this project, we were able to use our skills to solve real statistical questions. We wanted to find which country produced the most popular television shows based on our data. Our goal was to identify the country and the most popular shows from that country.

We used a dataset called ‘data_TV.csv’ found on Kaggle. This data has a list of 2,617 shows from different countries around the world. We decided to write our R codes first to clean our data before analyzing with SAS. This way we could clean our data. We chose to clean our data by removing duplicates and blank columns. We created a new data set from the cleaned data, we used the new dataset to analyze our data with R and SAS.

The data includes first air date, original country and language, name, popularity, vote average, vote count, and overview. Vote average is based on a scale 1-10. Most of the shows are produced in the United States along with other countries like Japan and Korea. Most of the popular shows are in English because the United States and Great Britain are two of the highest producing countries and are both English dominated. Similarly, a great number of television shows are produced in Spanish because Spanish-speaking countries are also top producers, for example; Spain and Mexico.

Results

R Results

Our first step in R was importing the csv files and merging them. We had one csv file that contained the numerical ratings and one with the origin countries. We merged them by name of the television show. We then cleaned the data by removing duplicates and blank lines of data. For the purpose of our research question, we will not be recommending shows with an average vote of below 5.0 and will be recommending shows with an average rating above 8.0. Best rating and worst rating were found by using 'slice_min' and 'slice_max' to find the top three best and worst rated shows. This was determined to be too broad due to the number of shows that tied with a rating of 8.7. Another way that we recommended television shows was based on a voter/watcher ratio. This was then determined to be unreliable due to a low correlation between the voter/watcher ratio; this is shown in figure 4. To filter our data, we took out shows with less than 500 votes. The United States has the highest rated show of our data, 'The D'Amelio Show'. Japan had the most highly rated television shows in the top ten rated shows. 28 of Japan's shows all tied for their top rated shows at a rating of 8.7. We used conditionals in R to put vote averages into categories. The distribution is outlined in figure 5. Television shows with ratings less than 6 were put in the lowest category, 'Bad'. Ratings between 6 and 8.5 were 'Good'. Every other show with ratings above 8.5 were in the 'Great' category. We used these categories to make a distribution of how many shows were in the 'Great category per country; this is shown in figure 6. It was determined that Japan and the US produced the highest quality shows.

SAS Results

Our first step in SAS was using ‘proc means’ to display the statistical properties of our data. We were able to find numerical data on popularity, vote count, and vote average including minimum, maximum, and standard deviation. We filtered the data by dropping the overview column. It was not necessary for the data we needed to answer our research question. We also filtered the popularity column to only print data that has popularity ratings above 7. From this filtered data, we formatted the numerical data to have commas. We created a smaller dataset using dataline that showcased the country, language, and number of shows produced. Conditionals were used to separate the vote counts into categories. If the vote count was less than 20, the show is considered to be a ‘Flop. Vote counts above 20 and less than 100, the show falls under the category of ‘Minor Success’. Vote counts between 100 and 300 are considered ‘Blockbuster’, and any vote count higher than 300 falls under the category of ‘Super Hit’. We made a frequency bar graph to visualize the conditions shown in Figure 1. The other two visual plots we made were based on the origin country. We made one sgplot with the horizontal axis as vote average and group as origin country shown in Figure 2. The last bar graph we made was the frequency of shows in different languages based on the origin language shown in Figure 3. Most countries produced shows in their dominant language, the only exceptions were Canada and the United States.

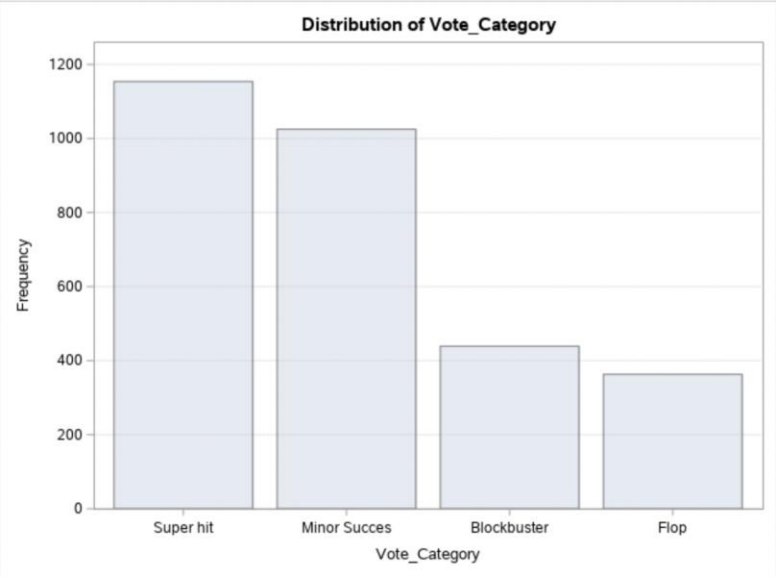
References

Statistics in APA - Purdue OWL® - Purdue University. (n.d.). https://owl.purdue.edu/owl/research_and_citation/apa6_style/apa_formatting_and_style_guide/statistics_in_apa.html

Find open datasets and Machine Learning Projects. Kaggle. (n.d.). Retrieved December 6, 2022, from https://www.kaggle.com/datasets

Figures

Figure 1.



Super Hit = Vote Count >200, Blockbuster = Vote Count 100-199, Minor Success =Vote Count 20-199, Flop = Vote Count < 20

Figure 2.

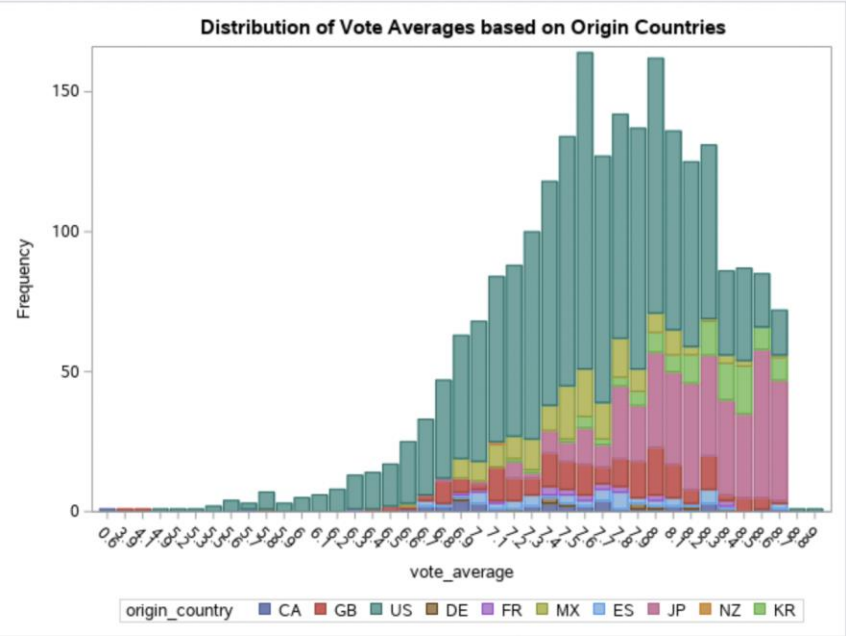


Figure 3.

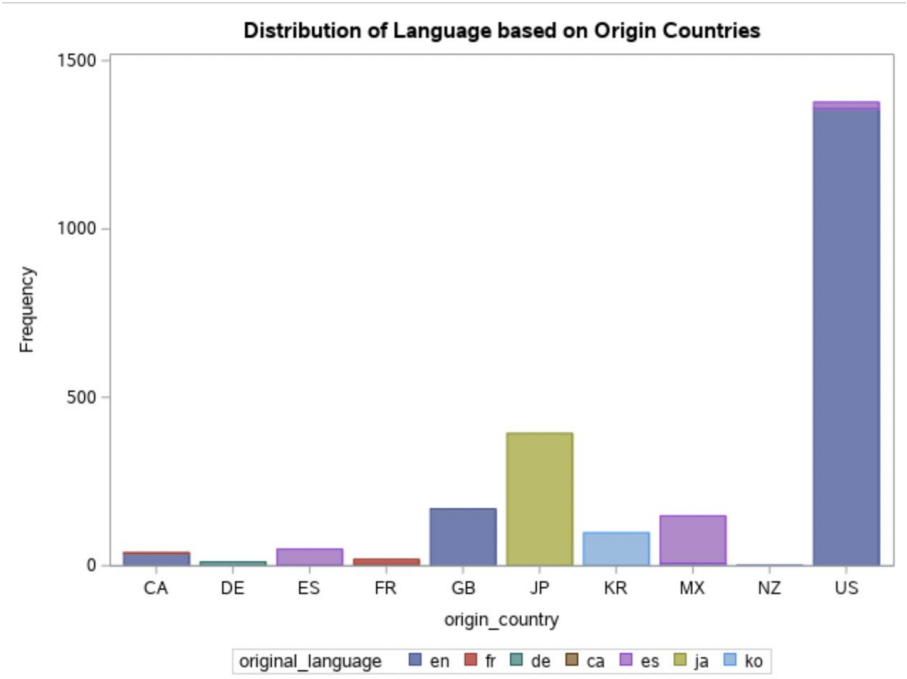
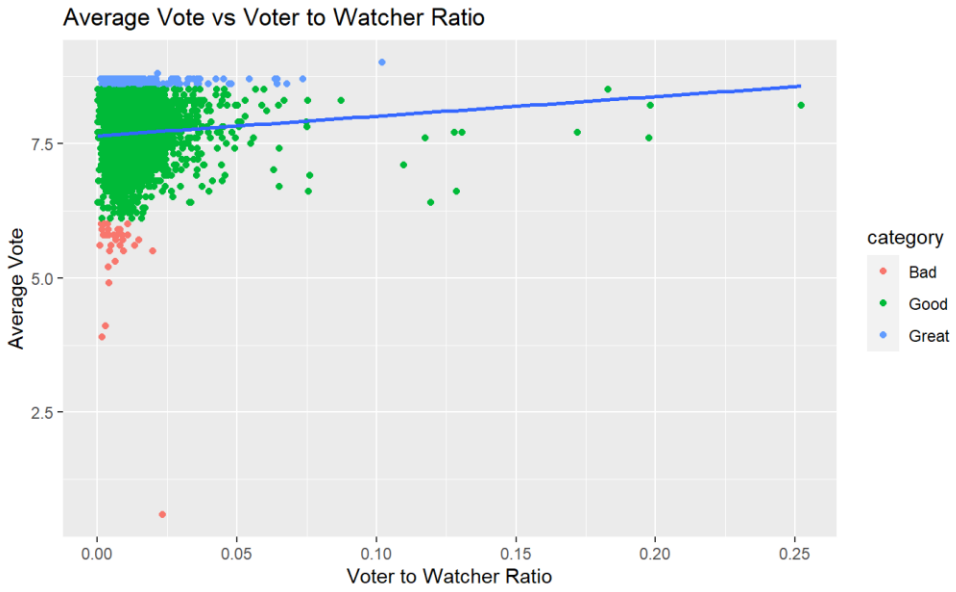
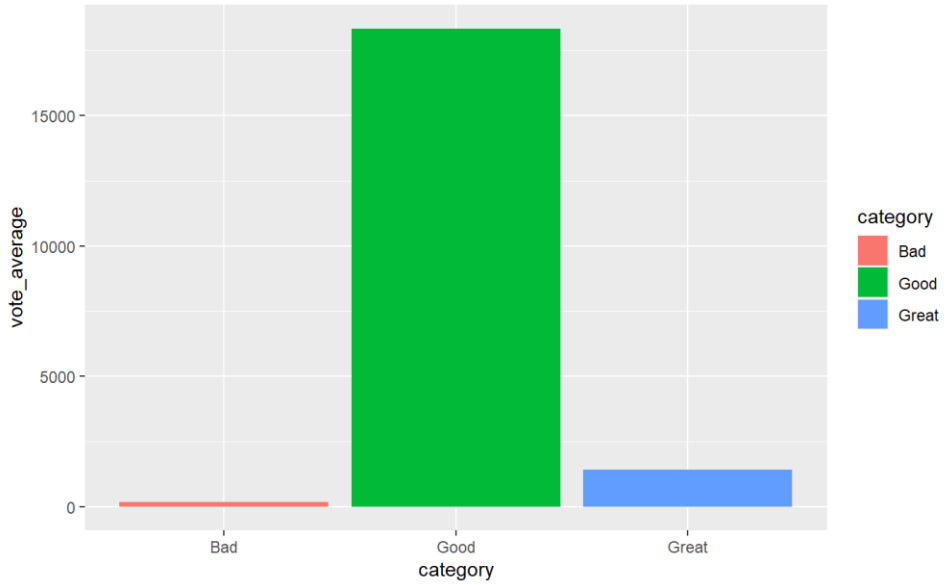


Figure 4



Correlation coefficient = 0.08587

Figure 5



Great = Avg Rating 8.5-10, Good = Avg Rating 6-8.499, Bad = Avg Rating 0-6

Figure 6