STAT 336 Final Project Meta Description

By Curtis Leaver

I chose the topic that I did because I wanted an interesting dataset to explore, that likely had not been explored too much. I figured that topic like health, politics, etc. would be widely studied and would have lots of visualizations. After searching for interesting datasets, I found the repository for tidy Tuesday, which basically adds an interesting dataset every Tuesday. After searching through the tidy Tuesday repository, one dataset about Star Wars caught my eye because I am a Star Wars fan. After looking through, I realized that there were more than enough observations and columns, so I decided to go with that dataset. The dataset includes survey data. The types of columns include questions like "Have you seen any 6 of the films in the Star Wars franchise," "Rank the Star Wars movies in order," or demographic questions like gender, age, or household income. Although answering a question about the data relating to demographics and its relation to Star Wars would have been useful, I find it more useful and interesting to focus purely on Star Wars data, being a Star Wars fan. Therefore, I knew that evaluating the opinion of a character would be the most variable result that is less easy to predict and more interesting to visualize. I also did not want to compare the popularity of two or more characters, as that is obvious to predict if you know the characters well enough intuitively. So, I knew that evaluating a controversial character like Darth Vader would be the best option because, honestly, I did now know how well he would be liked by the survey respondents.

My intended audience would be a newspaper article for a company like FiveThirtyEight or possibly executives in the film industry so I can portray to them that the way characters are portrayed heavily influence audience perception.

In order to clean the data, I used R Studio. I used the dplyr select function, which allowed me to select only the columns that I needed, which were “Have you seen any of the 6 Star Wars movies,” a column that had no meaningful name that represented if the respondent had seen Episode 3, and the column representing the respondent’s opinion of Darth Vader. To clean, I had to remove rows of respondents who had never seen a Star Wars movie. Then, mutated a column to say “Yes” or “No” depending on if they had seen Episode 3. Then, I removed all rows with N/A data for their opinion of Darth Vader. Finally, I got the counts of people for each opinion for seeing Episode 3 vs not and I converted them to percentages.

I wrote my write-up using information from Numbers in the Newsroom by Sarah Cohen.

I made the background black because Darth Vader’s main color is black, and so is Star Wars in general (space). I chose to make all the text white because it is easy to read in front of the black background. Percentages are the easiest thing to compare because you can line up stacked bar charts and they are on the same scale, so that’s what I did. Finally, for the legend and colors of the bars, I chose to make the colors be in order from most favorable to least favorable. I also color coded most favorable to the greenest and least favorable to the reddest, because this is a common color scheme for good vs bad. These changes were made based on our class readings of Design Elements by Dennis Puhalla and Show me the numbers by Stephen Few.

Citations

R for Data Science. "TidyTuesday." GitHub, 2021, https://github.com/rfordatascience/tidytuesday.

Cohen, Sarah. Numbers in the Newsroom: Using Math and Statistics in News, Second Edition, Investigative Reporters & Editors, Inc., 2014.

Few, Stephen. Show Me the Numbers: Designing Tables & Graphs to Enlighten. Perceptual Edge, 2004.

Puhalla, Dennis. "Chapter 3: Color Structure." Design Elements, Form & Space: A Graphic Style Manual for Understanding Structure and Design, Rockport Publishers, 2011, pp. 1 online resource (169 p.).