A few weeks ago, my friend and I were talking about our favourite TV show, Star Trek: The Next Generation (TNG). The discussion soon moved to who our characters were. I was rooting for **Data**, while my friend preferred **Picard**. Later that day, I wanted to find out what the broader community thinks of this. I found several discussion forums and articles, each with varying opinions. Not satisfied, the data scientist within me began to wonder if I can quantitatively answer this question. I needed to turn to the most extensive TV show data source out there: IMDb.

Luckily IMDb is very willing to share their data and I downloaded all the relevant files. Exploring and cleaning through the database, I settled on a neat tabular dataset with variables like episode number, season number, writer, director, and the IMDb rating.

Checkout report.pdf for the full analysis.

But wait, what is Star Trek anyway? Star Trek is a Science Fiction media franchise spanning several TV shows, movies, books, and video games. It was created back in the 60s with its first TV show. Many more Star Trek TV shows have come out since, including The Next Generation that ran in 1987-1994. The TV shows typically follow an episodic formula, with each episode having a self contained plot. As with TV shows of this nature, there is a core set of characters, along some recurring characters and one-off characters. There were also many directors involved, each directing a varying number of episodes. In this analysi, my focus was on the characters, and how characters influence the IMDb rating of episodes.

Thus the assumption is that we deem a character to be more favourably perceived by the audience if their presence improves IMDb ratings on average. The IMDb data source did not have information on chracter screentimes. Luckily, I was able to find script data online. This script data contained scripts from evey star trek episode. I was able to construct a proxy for a character's screentime in a given episode by calculating the percentage of words spoken by that chacter within that episode.

While cleaning my data, I realized that I can easily extend the analysis to other TV shows, and I did so by including three other Star Trek TV shows that followed TNG's run. The entire work was done using R v4.1.1.

Data

- IMDb Ratings and Episode information from https://datasets.imdbws.com (collected on Nov 23, 2021). This formed the source for the tabular data used in the analysis.
- Script Data from www.chakoteya.net/StarTrek/. This was the source for the script of each episode in text form. The text was used to calculate a proxy for each character's relative screentime in each episode, the proxy being percentage of words spoken by a chraracter in an episode. This proxy variable was then merged with the tabular data.
- Four TV Shows:
    - Star Trek: The Next Generation (1987–1994)
    - Star Trek: Deep Space 9 (1993–1999)
    - Star Trek: Voyager (1995–2001)
    - Star Trek: Enterprise (2001–2005)
- 614 Episodes in all
- Relevant Variables:
    - Name of Director(s)
    - Name of writer(s)
    - Episode Rating

    - Proxy Character Screentimes for main characters (33 in all)

Methodology 🛠

Data Extraction and Cleaning

This step involved cleaning up the script data, creating the proxy variable, and accurately matching it with the IMDb source data. This ste also involved cleaning up inconcsistencies between data from different shows and extracting relevant items from different datasets to create our final tabular dataset.

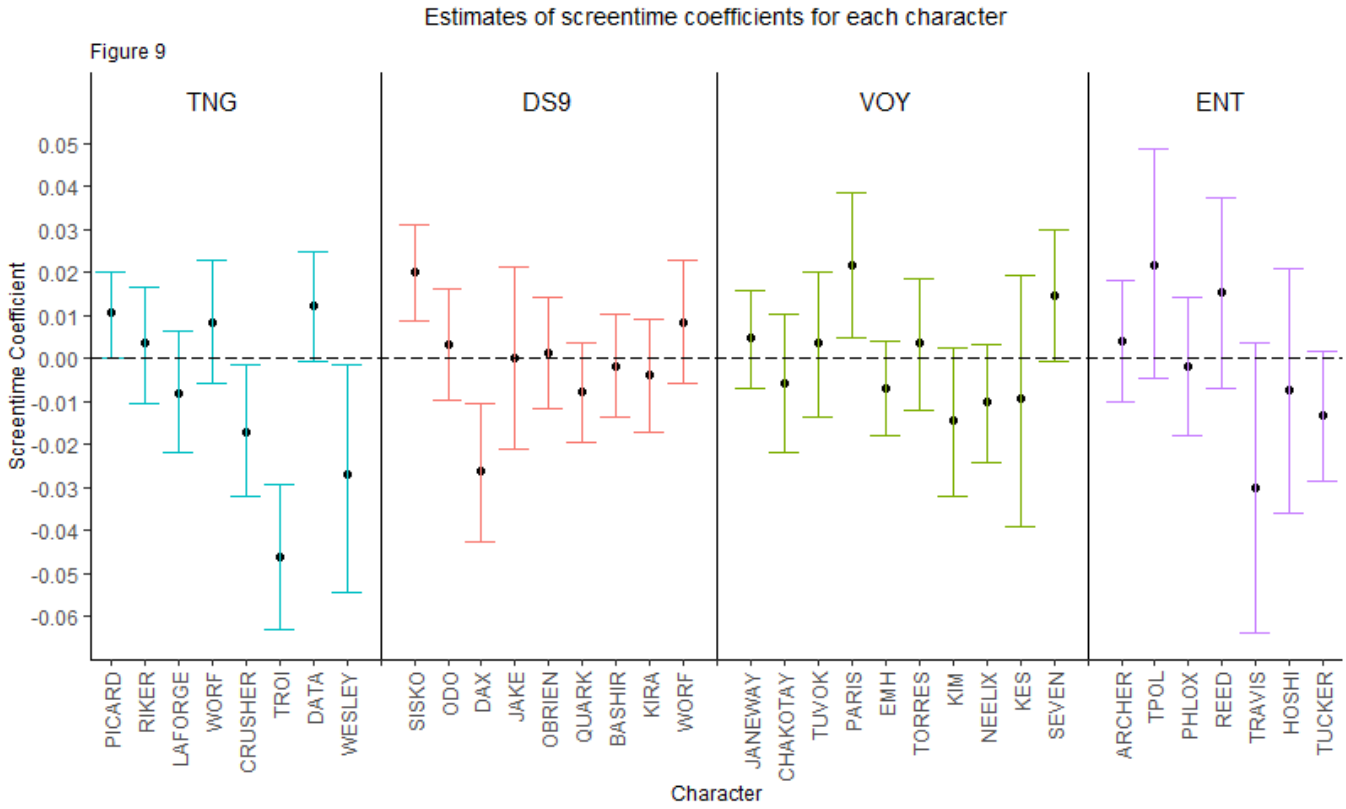
Statistical Modeling

The main idea behind our statistical methodology is that we create a proxy screentime variable for each character and quantify the character's likeability as the regression coefficient of their screentime on IMDb ratings. We also want to adjust the model to take into account the fact that not all seasons and not all directors are created equally. We use an heirarchical regression methodology as this allows us to adjust for general quality of shows and seasons while also allowing for the "director" variable. We cannot treat the "director" variable as a regular one hot encoded variable because of the large number of directors involved, with many of them directing only a single episode. We formulated our regression as:

- Target variable: IMDB rating
- Unit of observation: Episode
- Fixed Effects: Character Screentimes
- Random Effects: Show:Season combination, Name of director(s)

Results 📊

Influence of Characters



The coefficient of the fixed effects for each character will tells us how well a character is perceived by the audience. The figure illustrates the coefficient, which is the estimated average change in episode rating when a character's screentime is increased by 1 percentage point. This increase is at the expense of the screentime of "other" characters i.e. any character not included in the main character list. Corresponding 95% confidence intervals are also illustrated.

- Most characters were not observed to have a significant impact on episode ratings.
- Screentimes of Sisko from Star Trek: Deep Space 9 and Paris from Star Trek: Voyager have the strongest positive impact on episode ratings.
- Screentimes of Dr. Crusher, Troi, and Wesley from Star Trek: The Next Generation, and Dax from Star Trek: Deep Space 9 have the strongest negative impact on the ratings.
- There is not sufficient data to conclusively answer whether **Data** or **Picard** were more favoured in TNG. The jury is still out.

Conclusions

While we were not able to settle the original question, this proved to be a very interesting and fun analysis, and I learned much about applying statistical methods to answer questions by indirect framing of the question. I also realise that this analysis methodology can easily be applied to many other TV shows. One of the more prospective ones is the popular sitcom **Friends**.

Limitations

There are a few limitations to this study that need to be taken into consideration.

- IMDb calculates ratings as a weighted mean from individual user ratings, with the weight calculation algorithm being a secret. While the weighting scheme is in place to prevent manipulation, it leaves an uncertainty as to exactly what our target variable is.

- Characters change and evolve through a show's natural progression, and thus their effect on episode ratings is likely to be variable between seasons. This effect can be modeled using a random slopes model however the multifold increase in number of variables makes the modeling difficult. Thus we are forced to assume that each character has a uniform linear effect on episode ratings.
- We did not perform post-hoc analysis on our regression estimates. Due to the large number of statistical tests done, we are likelier to commit Type I errors and misrepresent the confidence bands.