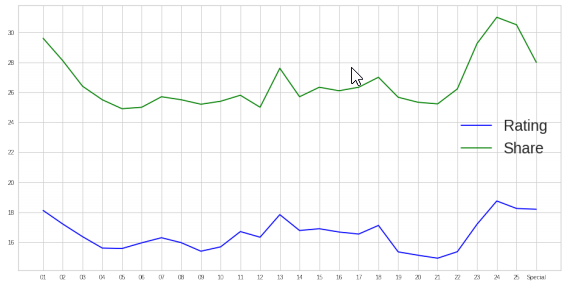
**So no one told you life was gonna be this way...**

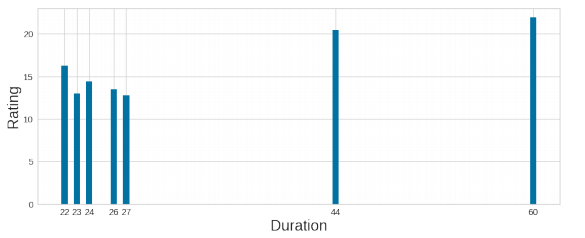
From 1994 to 2004, the show “Friends” was so popular that it spawned fashion trends based off the characters and became a driver behind the culture of “shipping” characters (the desire by followers of a fandom for two or more people to be in a relationship). It is because of this popularity, that ever since the show ended in 2004, fans have cried out for another season, or even just one more episode, as a “Where are they now?” for the characters. However, as people in the show industry well know, reboots and revivals risk ruining the show and its legacy altogether (here’s looking at you, Roseanne). So, if they were to attempt a follow up episode, **how could they guarantee it would be a success with high ratings?**

Once the data from the previous episodes was found, the first step was cleaning and exploring. In this process, I created different plots to identify possible relationships or trends. Some were as expected, some held little insight, while others provided unexpected but valuable results -

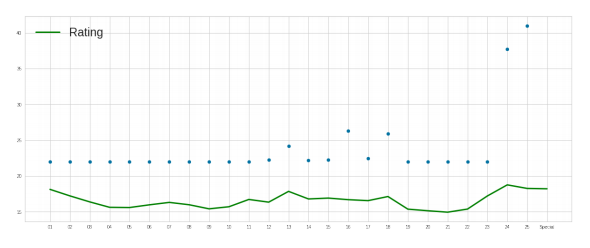
* Plot 1: Share and rating of episodes based on episode number on the same line graph:
  + Nothing unexpected, share and ratings followed similar trend lines



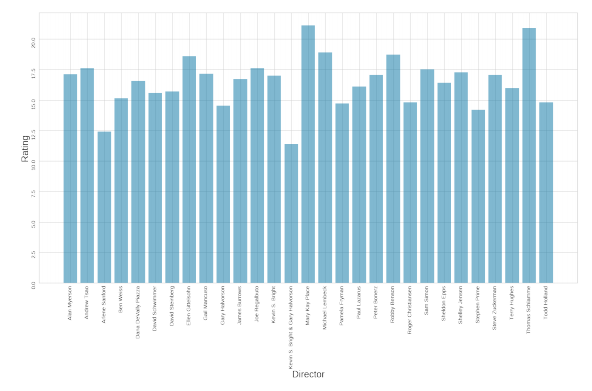
* Plot 2: Ratings by the duration of the episodes based on episode number:
  + The episodes with longer durations (typically premiers, finales, or specials) had higher ratings than shorter episodes. Validated prior expectations.



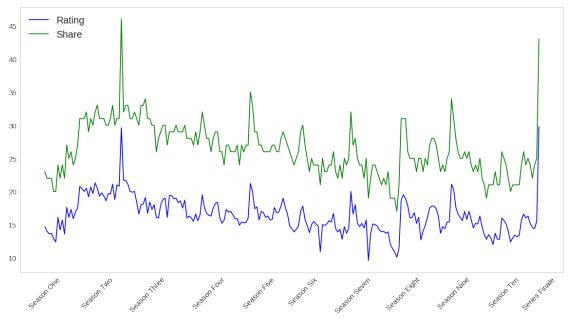
* Plot 3: Average duration vs the average rating:
  + Not particularly helpful, but further confirmed findings in Plot 2.



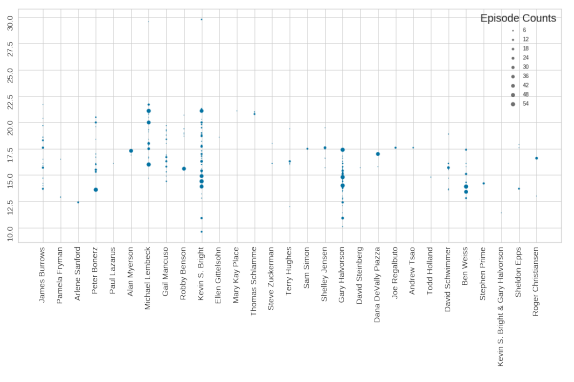
* Plot 4: Average rating of episodes based on the director:
  + Little variation in the averages, but the highest ratings belong to Mary Kay Place and Thomas Schlamme and lowest belongs to Kevin S. Bright/Gary Halvorson.



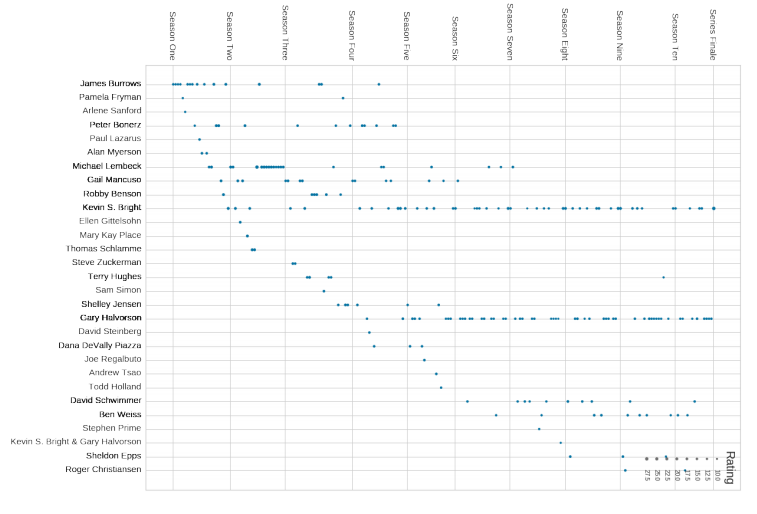
* Plot 5: All of the episode's ratings and shares:
  + More detailed version of Plot 1. Able to see spikes on individual episodes, pointing to the highest rated episodes: S2 E12/13 and S10 E18 (series finale). This was surprising as we saw the expected spikes around premiers and finales, but the middle of Season 2, nothing special happened except that they combined the episodes, giving the “one” episode the cumulative ratings of two.



* Plot 6: Director by Ratings scatter plot with the point size based on number of episodes directed with that rating:
  + Goes further into the ratings per director than Plot 4. This allows us to see that while Mary Kay Place had one of the highest average rating, she directed very few (<6) episodes, so her high rating is a bit of an outlier. It is a similar situation for Kevin S. Bright/Gary Halvorson - they did not direct many episodes together, so their low average rating is an outlier. What is interesting to note here is that Kevin S Bright and Gary Halvorson directed several episodes separately and each had higher average ratings individually.



* Plot 7: Episodes by Directors, with the point size based on the rating:
  + This plot visualizes trends of the directors throughout the seasons, showing that directors often came on board for a season or two, then left. There were few that stayed with the show longer than a few seasons – Kevin S. Bright and Gary Halvorson were two of them.



The plots provided, for the most part, visual confirmation of prior expectations, but some new insights as well - particularly related to the directors, which is where the first few milestones were heavily focused.

Taking that cue, I created dummy variables based on the directors, writers, and characters mentioned in the episode summaries, as well as assigning binary variables to represent episodes with ratings higher/lower than average. Using the director variables as input and ratings as the output, I followed our assignments and calculated the Confusion Matrix, Classification Report and Logistic Regression. These were not quite what I was looking for, though, so I turned to Decision Trees, using both the Gini Index and Entropy as criterion.

Decision Trees had a format closer to what I was looking for, so I began tuning hyper parameters, updating the max\_depth, splitter, and random\_state. This improved accuracy 10%+ to 71%! However, I was still unsatisfied with just one input variable, so I used Decision Trees one last time on new model variables, adding the writers and characters to the input variables.

The accuracy went down with the increased variables. This was not surprising. With a 56% accuracy, though, it was better than a random guess, so I was satisfied.

What does the analysis/model building tell you?

First off, that there are many variables that can affect a show’s ratings and popularity. To narrow it down to just a director, writer, or character is difficult. However, when looking at those specific variables, it was interesting to see what I had had personal bias towards be confirmed with the models (i.e. Kevin Bright directed good episodes, the episode would be a good one if the story had a focus on Chandler, etc).

What are your recommendations?

Create an episode with a highlight on each character’s growth and life updates, with special attention to Chandler and Monica’s relationship (they should NOT divorce), written by Adam Chase and directed by Kevin Bright.

How would you pitch this business problem to a group of stakeholders to gain buy-in to proceed?

With this show appealing to a large demographic of viewers across generations, particularly viewers who affect and oftentimes make the financial decisions in the household, it is vital to create episodes that draw them in and provide more viewership for the advertisements and streaming services: More viewers (higher ratings)-> More sales from ads -> Higher paying ads/revenue from streaming services. To sum up, high ratings lead to more money for the show, and my analysis will provide a guide to creating highly rated episodes, leading to more money for you.

Why should someone in the business care about this solution?

Highly rated shows generate large cash flows, which is something everyone in any business cares about. With the right directors, writers, and by extension, characters, a highly rated episode can be created and generate cash flow. This analysis can be used to identify those right directors, writers, and characters to use.

What are some of the potential challenges or additional opportunities that need to be explored?

Directors, writers, and characters are certainly not the only factors that come into play when working towards higher ratings. The time of day of the show airing, the time of the year, the political climate, world events, even makeup and wardrobe can have big effects (just look at the trend of the “Rachel” hair in the late 90s, early 2000s!). This information was not included in the dataset, but had it been, it could have been included in the input variables to refine the predictions. The amount of time spent tuning and optimizing increases with each new variable, so it would not be a quick feat, but it would certainly provide a wider perspective.

As for the variables already included in the dataset, there are nearly endless visualizations that could have been created and a little more insight gained with each. Had this been my full-time job, I would have explored them further, as well as further tuning for the Decision Trees model to increase accuracy.