# The effect of emotional content in episodes of the TV show *Friends* on US viewership

## Elda Pere, Marcus Manos, Casey McGonigle

# 1 - Introduction

The number of fans tuning into a television show is determined by several factors internal and external to the production. Things such as director choice, writer, actors, and emotion tend to affect the number of viewers. Traditionally stakeholders have taken a non-data-driven approach to answer this question. Instead of collecting and analyzing data relating to internal factors, they focus on external factors such as production budget, timing, and trying to understand demographic groups. What these stakeholders are missing is the ability to apply statistical analysis to internal factors like writing, acting, or emotion.

To understand the relationship between these internal factors and television shows we decided to study how emotion affects views on one of the most popular shows in history, Friends. We chose this show because of the availability of data; due to its popularity, there is high-quality data available from reputable sources. We also chose it in part because its popularity made the viewership numbers a goal for many shows to this day. On the series finale in 2004, the show was watched by 52.5 million American viewers, “making it the fifth-most-watched series finale in television”.

Although there are a number of external factors that would be interesting to study, we settled on internal features of television shows since that is what producers are able to control. After taking an initial look into the dataset, we decided to focus on what features of the television program increase viewership. Personally, we measure shows by how much we laugh or cry during an episode, so we chose to specifically study what emotions increase viewership. This leads us to our research question: “Does the frequency of a given emotion affect viewership of Friends in the U.S.?”. We plan on using OLS regression to determine the causal relationship. Answering this question will enable us to better understand viewer preferences and provide a real impact on the process of creating future productions.

# 2 - Description of the Data & Research Design

We used the Friends dataset featured on “Tidy Tuesday” hosted on Github. This dataset included three tables:

1. friends.csv – this table contains the words and the speaker of every line from all 10 seasons of the show
2. friends_emotions.csv – this table identifies the emotional content of most lines from each episode in the first 4 seasons of the show. There are 7 possible emotions. Credit to Emil Hvitfeldt for creating and aggregating this data – we believe he went through transcripts of Friends episodes line-by-line to decide which emotion each line contained.
3. friends_info.csv – unlike the previous two tables, this contains episode-level information. It tells us each episode’s title, director, writer, air date, views in the US, and IMDB rating. For our purposes, only the last two (friends_emotions.csv and friends_info.csv) were relevant.

We began by evaluating the data source and ensuring that it was legitimate and of the highest quality. Since we discovered this data source from a popular Github repository (Tidy Tuesday) and discovered the foundation for it came from several well-respected researchers from institutions such as Emory University, we felt justified in assuming that the data was trustworthy.

Although we had confidence in our data source, the data itself had four important shortcomings. Most notably, the emotions were only tracked through the first 4 seasons of the show. Likewise, not every line in those 4 seasons was analyzed with emotional analysis. Moreover, the data did not account for non-verbal communication like body language or 'laugh tracks'. Finally, there were a few outliers present in the dataset that have significantly more views than other episodes.

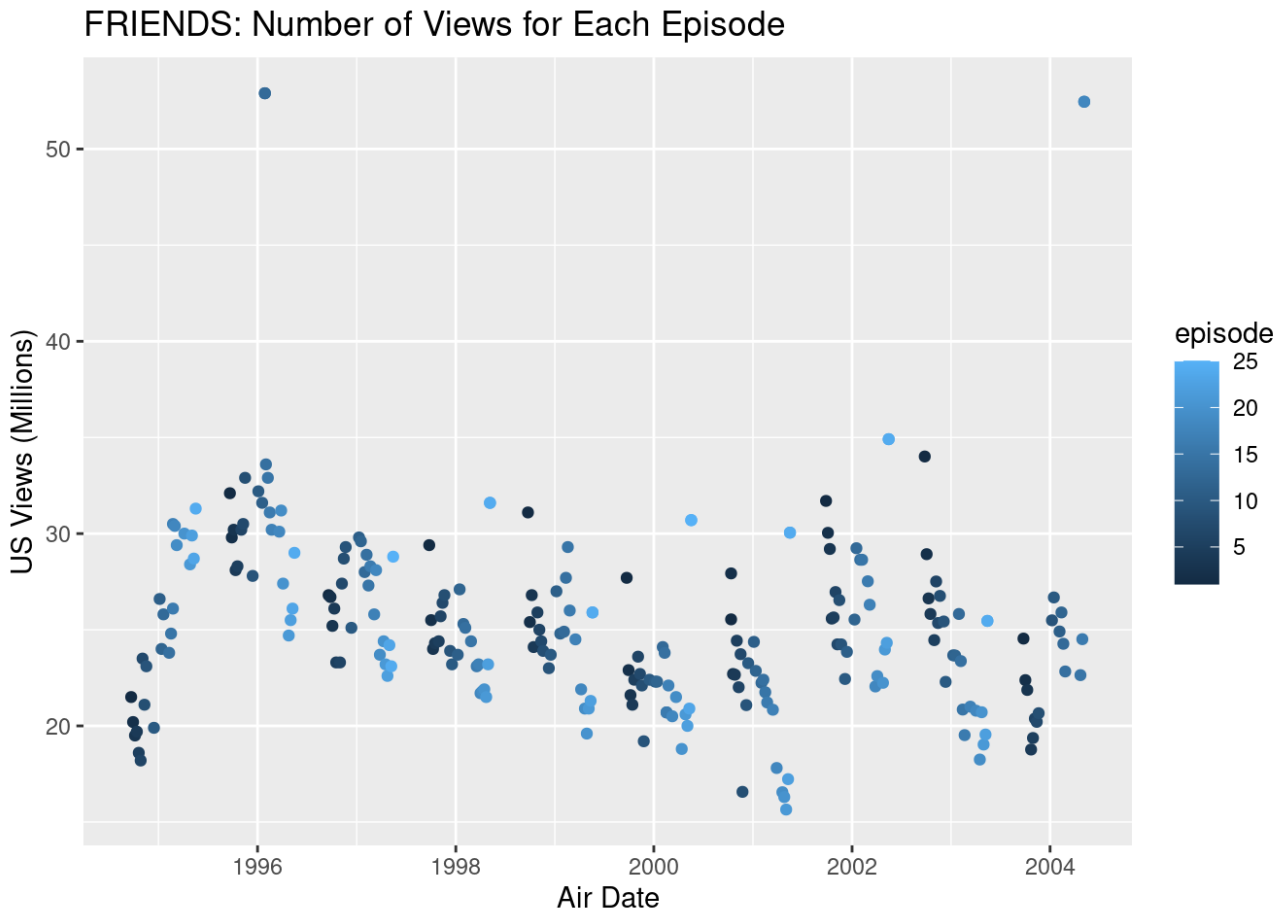
The most pressing issue – that the emotions were only tracked through the first 4 seasons – affected the initial design of our study, causing us to change the scope of the experiment and limit it to the first four seasons. Since this lowered the total number of data points, we had to switch from using the Large Sample Linear Model assumptions to the Classical Linear Model (CLM) assumptions which are more limiting.

The second issue, that not every line was analyzed by the initial researchers does not appear to be a major problem. This is because there were still enough lines analyzed in each episode to understand the general emotion in each episode. Additionally, most lines that weren't analyzed were irrelevant or especially short, which would have likely added more noise to our dataset.

The penultimate issue was that the dataset did not account for body language or other indicators of emotion like laugh tracks. These are a pivotal part of any television show and especially a comedy like Friends. Even though body language and the visual aspect of Friends certainly did have a lot to do with the number of views, they vary greatly from performer to performer. Since the main protagonists stay constant throughout the show, we found it justified to use solely the spoken lines.

The final issue was with the outliers in the dataset. We found two episodes that had roughly twice as many views as the average episode. Upon investigation (see chart 2.0.1 below), we realized they were titled "The One After the SuperBowl", and "The One After the SuperBowl pt. 2". In other words, their viewership was artificially inflated because of the millions of people who watched the Super Bowl and left their TV on afterward. We decided to remove these points from our dataset since they skewed the data by a significant amount (and prevented us from meeting select assumptions for the CLM).

## 2.0.1 – EDA – Number of Views per Episode




After cleaning and wrangling our data, we were left with a dataset covering 95 episodes in seasons 1-4 of Friends. Though our sample size was too small for the Large Sample Linear Model, we were able to use the Classical Linear Model, as long as we met its more stringent requirements.

# 2.1 - Variables

To analyze the impact of emotion on viewership we focused our study on several different variables listed in the data: “Neutral”, “Joyful”, “Mad”, “Peaceful”, “Powerful”, “Sad”, and “Scared”. To get a better understanding of the relationship between emotions, we decided to aggregate the data to an episode level and find the percentage of phrases in each episode by emotion. We chose to use the percent of each emotion (as opposed to the raw count of each emotion) because some episodes had many fewer lines analyzed by emotion than other episodes. We were afraid that using raw counts would encode the total number of lines analyzed in the episode into our model.

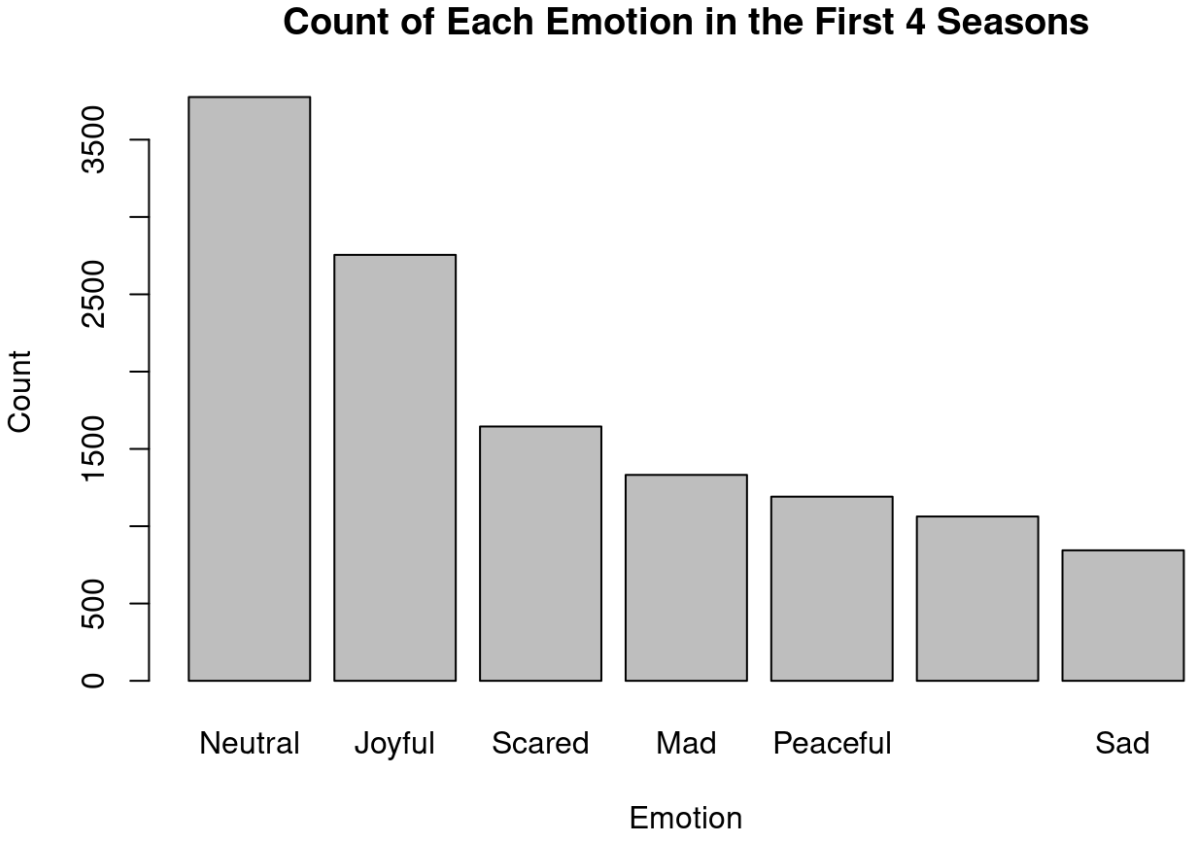
In addition to each of the emotion percentage variables, we chose to include time-related features such as the season of the show and the calendar season to account for any kind of autocorrelation or endogenous influences on viewership.

# 2.2 - Research Design

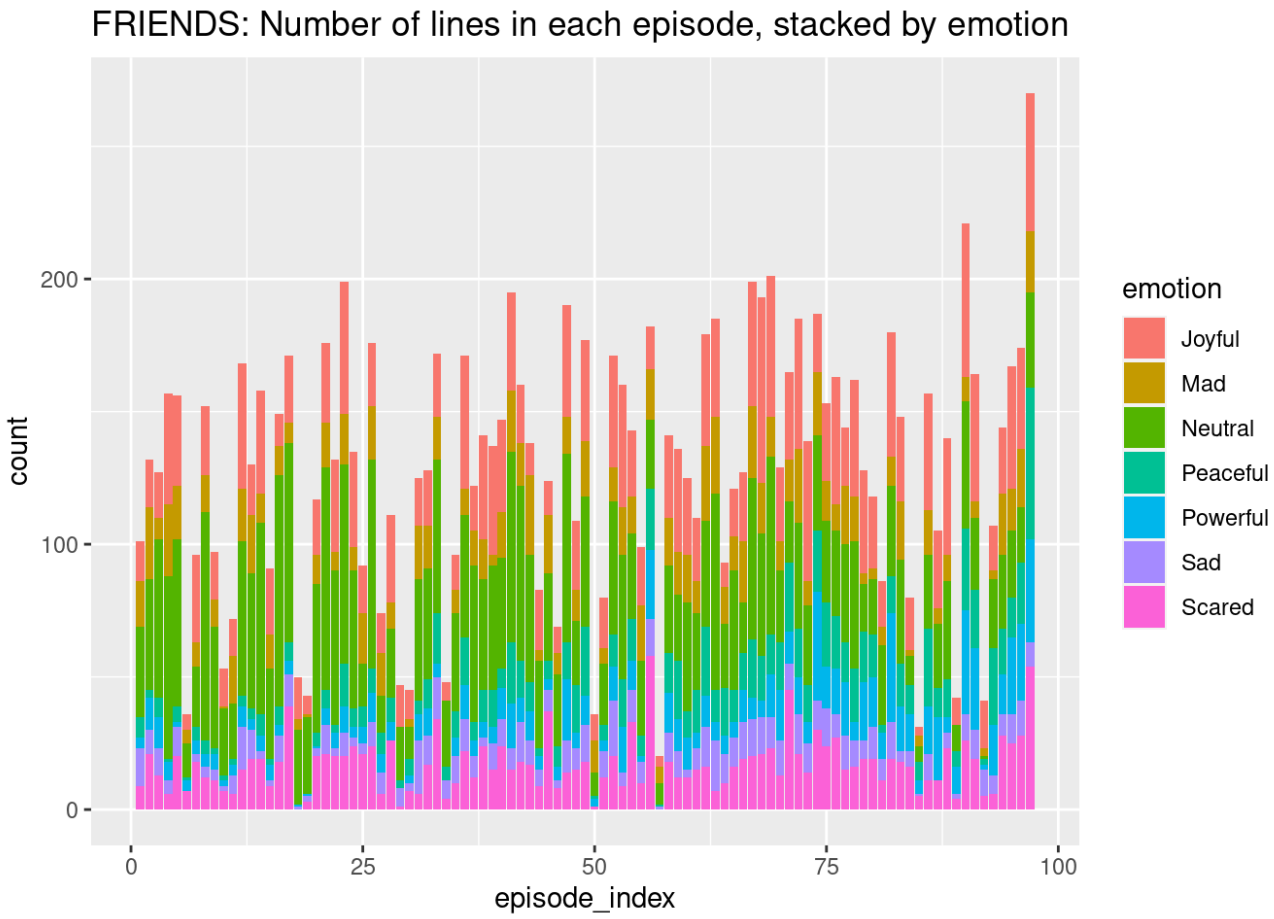
Our research design process included creating a hypothesis, conducting exploratory data analysis, building linear models, evaluating assumptions for them, and analyzing the results in context. Our first step, creating a hypothesis, was completed after taking an initial look at the data. We determined that studying how emotions affect views would be the most practically useful hypothesis to test. We eventually created the null hypothesis: emotions in a Friends episode have no effect on US viewership.

After choosing the topic and developing a research question and null hypothesis, we began to dive into the data and conducted an Exploratory Data Analysis (EDA). During the EDA, we visualized the distribution of emotions (see charts 2.2.1 and 2.2.2 below), looked for outliers in the number of episode views (see chart 2.0.1 above), checked the data types of each column, and visualized the relationship between our emotional variables and the number of views (see chart 2.2.3 below – note: this only shows the Mad emotion, but all 6 graphs for the other emotions looks very similar) This process was crucial in understanding the trends of the data and how the covariates interact with each other.

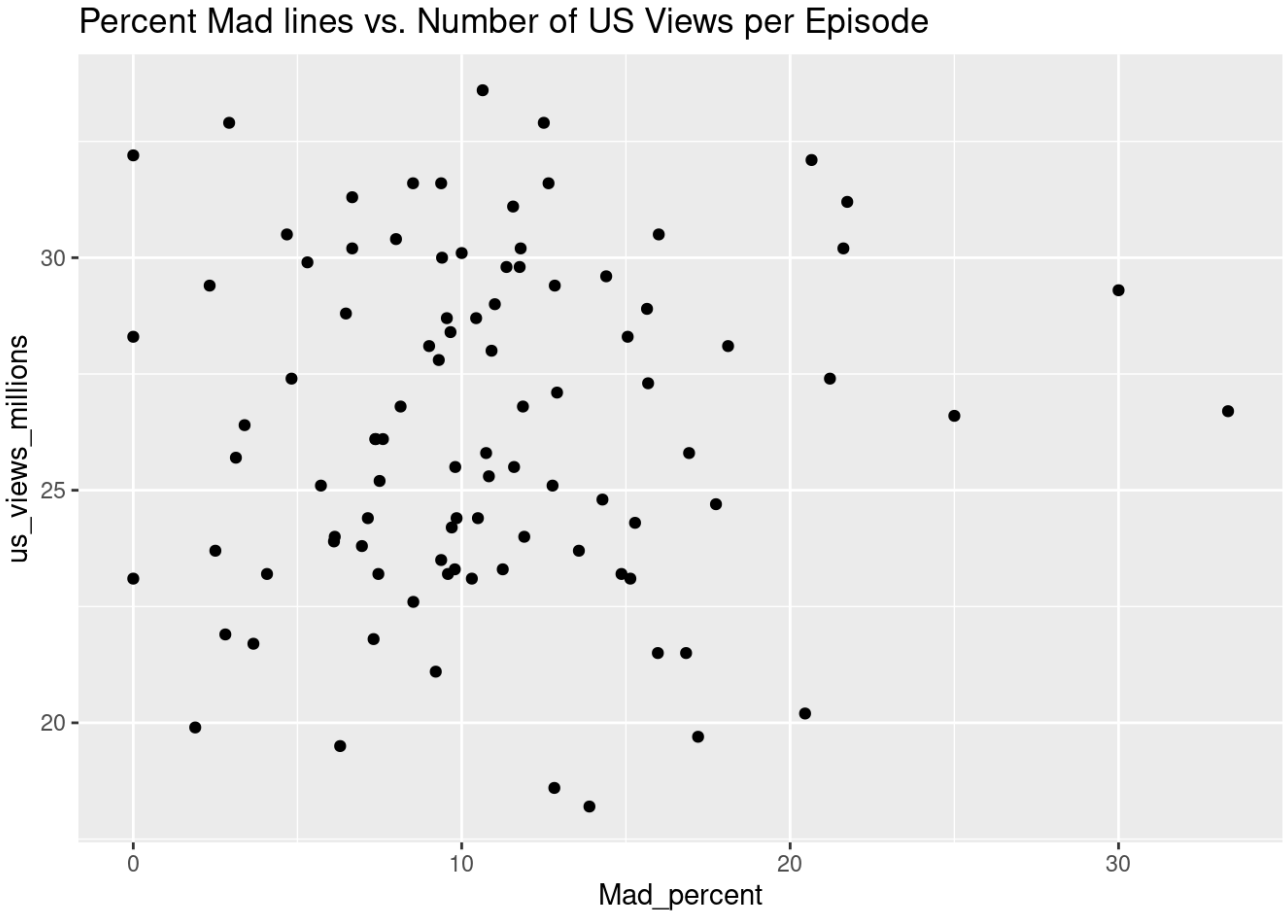
### 2.2.1 – EDA – Distribution of the Counts of Each Emotion in the first 4 Seasons




### 2.2.2 – EDA – Counts of Each Emotion By Episode in the first 4 Seasons




### 2.2.3 – EDA – Relationship between % of 'Mad' lines and US Views



The next step included building three linear models to evaluate our theory that emotional utterances have a causal relationship with views. These models were refined over multiple iterations alongside adjustments to meet the assumptions of the Classical Linear Model – in particular the assumption of heteroskedasticity. After transforming and omitting select data points, we were finally able to meet this assumption and move forward with the analysis.

Next, we had to analyze the results of output linear models. We conducted a t-test with the p-value adjusted using a Bonferonni correction to avoid type 1 error – find more details on this process in the results section.

Finally, we applied our results in the context of the problem. We were successful in rejecting the null hypothesis that emotions have no effect on US viewership, and we thus identified a causal relationship between powerful utterances and views in the United States. We came to this conclusion even after the Bonferroni correction – our p-value was still below 0.05. If the writers and producers were to incorporate our findings by decreasing the presence of powerful emotions in Friends episodes, they'd get more views of their show.

## 2.3 - Brief Introduction to our Models

We will combine our features into three causal models: the first is the simplest. It only uses the percent of neutral emotions to measure the effect of the presence of emotion on views. The second model contains 6 of the emotions ( "Joyful", "Mad", "Peaceful", "Powerful", "Sad", and "Scared") to measure the effect that different emotions have on viewership. The third and final model will include the same emotions as the second model plus control variables relating to time fixed effects such as calendar and television season. These are explored in greater depth in the Model Building Process section below.

To analyze the models we plan on using Ordinary Least Squares (OLS) Regression to find the coefficients for each variable of interest. We will determine significant results by having a p-value of less than or equal to .05. Our goal for these models is to identify a causal relationship between the presence of emotion in Friends and the views for each episode. In addition, we are hoping to identify practical significance we can distill into easy-to-understand advice for producers to take in order to increase viewership.

## 2.4 - Potential Risks

One set of potential risks for our study surround our ability to meet the Classical Linear Model's assumptions. This is a risk since our data does not strictly follow the assumptions, which could potentially invalidate our results. With this in mind, we applied stringent conditions to our model to avert this risk. In the case of the heteroskedasticity assumption, we noticed one model had non-uniform variance which would have biased the results during the analysis stage.

The second set of potential risks surround the data we used to conduct the analysis. The data could be been biased since it only included the first four seasons of the show. The last 6 seasons could have included some important details and given us additional power, both in terms of the data points and the practical significance. If additional seasons were included we could have had more than 200 data points and we are able to apply the Large Sample Linear Model – which has much less stringent assumptions than the CLM. The additional data points would also increase the practical significance since it would include data closer to the current year, giving a better insight into current viewer trends.

# 3 - Model Building Process

We want to investigate if emotion in a Friends episode affects viewership of the episode in the United States. To that end, we built three linear statistical models to estimate the number of episode views from the percentage of lines in the episode that contain a given emotion. Please note that we use the percent of each emotion in the episode (ie. the

number of lines with that emotion / the total number of lines with any emotion) because we do not want to encode the number of analyzed lines in an episode into our model. We've created the variables for the percentages of each emotion to avoid just that. Without further ado, here are our three models:

1. We began with our simplest model, which regresses the number of US views on just one right-hand-side variable: the percentage of lines in an episode that have neutral emotion. In other words, this model looks at how the lack of emotion impacts the number of views an episode receives.

$$views = \beta_0 + \beta_1 PercentNeutral$$

2. Next, we looked at the opposite effect; instead of investigating how the lack of emotion impacts an episode's views we looked into how the presence of each emotion impacts an episode's views. Our dataset explicitly contains six different variables about the presence of certain emotions and we chose to use all six in this model. We used a Bonferroni Correction (covered in part 4: Results) to avoid inflated p-values that come from testing 6 different variables. Note that the neutral variable is still implicitly in this model – we've removed it from this model specification because including it would lead to perfect collinearity between our covariates (in each episode the sum percent of emotions is 100 – if we included all 7 emotions, they'd sum to 100 in every episode)

$$views = \beta_0 + \beta_1 PercentJoyful + \beta_2 PercentMad + \beta_3 PercentPeaceful + \beta_4 PercentPowerful +$$

$$\beta_5 PercentSad + \beta_6 PercentScared$$

3. Finally, we chose to build upon our second model by including control variables for the season of the show (1-4) and the calendar season in which the episode aired (winter, spring, summer, fall). This allows the model to more precisely estimate the effect of each of the emotional variables on the number of views; previously any variation in views that should have been accounted for by one of the season variables was incorrectly absorbed by the emotional variables whereas this architecture clearly separates the season-based effects from the emotion-based effects

$$views = \beta_0 + \beta_1 PercentJoyful + \beta_2 PercentMad + \beta_3 PercentPeaceful + \beta_4 PercentPowerful +$$

$$\beta_5 PercentSad + \beta_6 PercentScared + \beta_7 Season1 + \beta_8 Season2 + \beta_9 Season3 + \beta_{10} Fall + \beta_{11} Winter + \beta_{12} Spring$$

Altogether, these three linear models work to show if a Friends episode's US viewership level is impacted by the emotions present in that episode. The first looks at how lack of emotion impacts viewership while the next two focus on how the presence of specific emotions causes changes in viewership. Finally, the third model adds seasonal control variables to ensure the effects attributed to the emotional variables are not due to the omitted variables.

# 4 - Results

```
##
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                              us_views_millions                us_views_millions
##                                     OLS          coefficient          OLS
##                                                     test
##                                     (1)             (2)              (3)
## --------------------------------------------------------------------------------
## Neutral_percent                    0.024
##                                   (0.031)
##
## Joyful_percent                                     -0.062           -0.035
##                                                    (0.061)          (0.055)
##
## Mad_percent                                         0.003            0.012
##                                                    (0.069)          (0.063)
##
## Peaceful_percent                                    0.160            0.032
##                                                    (0.096)          (0.094)
##
## Powerful_percent                                   -0.207**          -0.102
##                                                    (0.075)          (0.090)
##
## Sad_percent                                        -0.020           -0.073
##                                                    (0.119)          (0.093)
##
## Scared_percent                                     -0.003            0.008
##                                                    (0.079)          (0.060)
##
## factor(season)2                                                     5.384***
##                                                                     (1.064)
##
## factor(season)3                                                      1.953
##                                                                     (1.246)
##
## factor(season)4                                                      1.401
##                                                                     (1.581)
##
## factor(calendar_season)Spring                                        0.505
##                                                                     (0.791)
##
## factor(calendar_season)Winter                                       2.102**
##                                                                     (0.781)
##
## --------------------------------------------------------------------------------
## Observations                         95                               95
## R2                                  0.006                            0.384
## Adjusted R2                        -0.004                            0.303
## Residual Std. Error          3.626 (df = 93)                  3.022 (df = 83)
## F Statistic                 0.599 (df = 1; 93)              4.707*** (df = 11; 83)
## ================================================================================
## Note:                                              *p<0.05; **p<0.01; ***p<0.001
```

```
## [1] "P Values for Model 1"
```

```
##     (Intercept) Neutral_percent
##    4.704801e-43    4.410651e-01
```

```
## [1] "-------------------------------------"
```

```
## [1] "P Values for Model 2"
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      28.1270172  2.8807009  9.7639 1.098e-15 ***
## Joyful_percent   -0.0622985  0.0612341 -1.0174  0.311761
## Mad_percent       0.0034831  0.0688111  0.0506  0.959745
## Peaceful_percent  0.1601934  0.0955736  1.6761  0.097263 .
## Powerful_percent -0.2066851  0.0746706 -2.7680  0.006877 **
## Sad_percent      -0.0204594  0.1193272 -0.1715  0.864259
## Scared_percent   -0.0032018  0.0787965 -0.0406  0.967680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "-------------------------------------"
```

```
## [1] "P Values for Model 3"
```

```
##                   (Intercept)                Joyful_percent
##                  3.361710e-16                  5.270615e-01
##                   Mad_percent              Peaceful_percent
##                  8.538442e-01                  7.344265e-01
##              Powerful_percent                   Sad_percent
##                  2.619724e-01                  4.340906e-01
##                Scared_percent                factor(season)2
##                  8.888650e-01                  2.462836e-06
##               factor(season)3                factor(season)4
##                  1.208541e-01                  3.779095e-01
## factor(calendar_season)Spring factor(calendar_season)Winter
##                  5.249439e-01                  8.600935e-03
```

The motivation behind the model choices above is to provide the writers of Friends with guidance on what the audience responds to. Based on our research and available data, we focused on the statistical significance of the types of emotions within each episode. As seen in the table, the proportion of utterances with powerful emotions has a significant effect on the number of viewers of that episode when regressing on all of the available emotions. With every percent increase in the proportion of powerful utterances in an episode, the number of US viewers decreases by about 200 thousand, all else constant. An example of a powerful utterance is one from Phoebe Buffay in season 2: “But Joey, you’re gonna be fine. You don’t need that show, it was just a dumb soap opera.”

In a practical sense, this provides the show creators with ways to control viewership in the future. The writers can reduce the number of powerful utterances to make the show more popular, and then run additional statistical tests with other variables for the same purpose. There may be a limit to how much they choose to reduce powerful utterances because of the flow of the show, but at least there is a clear direction.

## 4.1 - Bonferroni Correction

We realize that the models run above are regressing views on a number of emotion variables that could be considered part of a family. To keep the family-wise error rate below 5%, we consider a Bonferroni correction for the 7 emotion variables (Neutral_percent, Joyful_percent, Mad_percent, Peaceful_percent, Powerful_percent, Sad_percent, and Scared_percent).

Since the percentage of powerful emotions is the only emotion variable that shows statistical significance in the Stargazer table (**p<0.01), our main question is whether this variable stays significant after applying the correction. When multiplying the current p-value of 0.006877 by the number of family elements - 7 - we get 0.048139. Despite the short distance to 0.05, we conclude that the percent of powerful utterances in a Friends episode has a significant effect on the number of U.S. views (in millions) it receives, and the previous analysis of practical significance stands.

# 5 - Limitations of Our Model

## 5.1 - Statistical Limitations

With less than 100 data points (n = 95) and the limitations analysis that follows, we used the Classical Linear Model and addressed each assumption of the model accordingly. Given that we are only looking at data from one show, the data is not IID if this model is used to generalize the effect of emotions on viewership in other shows. Instead, we can use it to explain the causal relationship between the two variables in Friends and try to generalize to future seasons. In this case, the variables are time-dependent but the autocorrelation it causes will be addressed by controlling for time using the calendar season of the air date. Otherwise given the size and sampling of the data, it is sufficiently IID to create a useful regression model.

An important assumption that is satisfied is that the unique best linear predictor exists. The variables here are not linear combinations of each other; there is no perfect collinearity and no heavy tails in our variable distributions. We have also satisfied the assumptions for linear conditional expectation and normally distributed errors using plots of residuals against fitted values and Q-Q plots.

When testing for heteroscedasticity, we realized we needed to make a correction because the variance of the data in the second model was not constant. The base model that regresses the US views on neutral emotions and the third model with 6 of the 7 emotions and controls both appear to be relatively homoskedastic. This can be seen through Scale-Location plots (the square root of the standardized residuals plotted against the fitted values) as the lines are relatively horizontal.

However, this is not the case for the second model as the line of best fit appears to be almost parabolic in shape. To adjust for the heteroskedasticity in the second model (model_emotions_only) we decided to run a coefficient test on the model with standard robust errors so that heteroskedasticity does not affect the standard errors. Although standard robust errors increase the variance and can potentially bias the results, we did not find issues with it due to the number of data points. This allows us to justify the homoskedastic errors assumption.

## 5.2 - Structural Limitations

Since we did not collect the data and had no say in its creation, we were limited to the provided variables. There are a number of additional variables that are associated with the US views of Friends but that are not part of our dataset. Mainly, we consider the most important omitted variable to be the marketing budget of the show. More concretely,

this could be the dollar amount spent for marketing Friends before each episode. Although it could be measured, the variable was not available for this study.

With an increased amount of marketing, US viewership would intuitively increase - all else constant. Given that the association between powerful utterances and US viewership in the regression model is negative, the bias would be (positive) x (negative) = negative. Since the coefficient of powerful utterances is also negative, the negative bias drives the coefficient value towards zero. This calls into question the significance of powerful utterances against US viewership and is enough cause for future studies, which should attempt to collect the omitted marketing variable by contacting the creators of Friends or any associated parties.

# 6 - Conclusion

Ultimately, we rejected our null hypothesis that emotions have no effect on US viewership. We found a statistically significant relationship (after a Bonferroni Correction) between the number of views and only 1 of our emotional variables. Specifically, the percentage of ‘powerful’ lines has a negative relationship with the number of views an episode receives. This suggests that Friends’ writers and producers could cause an increase in the number of views by including fewer lines with ‘powerful’ emotions. Otherwise, we find no evidence to suggest that the presence of other emotions (sad, joyful, peaceful, mad, scared, and neutral) has any impact on an episode’s viewership. This could pave the way for future studies that consider the same variables with additional data points or in new scenarios; the number of TV show views is a metric that entertainment companies are always trying to optimize.