Funky Cats and Their Feisty Stats - Datafest Write-Up

Nichole Feghali, Verna Maullon, Shruthi Vaidyanathan

June 14, 2020

Data Retrieval

Our group wanted to review the question "Do people interact with recipe and DIY YouTube videos more during the physical distancing and social isolation period caused by Covid-19?" While working on this project, we encountered various challenges concerning how to attain our dataset. Initially, we used the 'tuber' package in R to get data for specific videos on YouTube. However, this method only allowed us to see the total number of views for a video instead of the number of views a video gained on a specified date.

We then researched the YouTube API and learned there was an alternate method to getting data from YouTube. Using the YouTube API, we were able to download data for YouTube videos, including the number of comments a video gained on a given date. The first day we used the YouTube API, our group did not realize there was a daily quota on how much data could be downloaded. We reached our quota before we had all of the data we needed and realized we had downloaded extra data we would never need. After realizing this mistake, we planned exactly what data we would need to collect the next day so that we could gather it all before reaching our quota for that day. Using the API, we decided to use the number of comments a video gained everyday from January 1st, 2020 to June 12th, 2020 for the five most popular videos on ten different YouTube channels, as our dataset. This showed us how much people were interacting with the YouTube videos before physical distancing and social distancing started (From January to March) and after it started (from March to the most recent data we collected on June 12th).

Data Analysis

We plotted the number of video comments by the dates specified above, to illustrate significant patterns and relationships between the two variables. As hypothesized, the majority of the plots regarding both cooking and DIY channels showed a significant increase in video comments during March and April. Prior to these months, the plots of cooking channels illustrated a spike in the number of comments during mid January. In contrast, the plots of DIY channels illustrated variations ranging from minimal changes to significant changes in comment numbers. While exploring the data from each channel, we noticed that similar to how there were channels that did not follow the hypothesized trend, there were also channels that did not follow a common trend prior to March. We realized that factors such as time frame, channel popularity, or the video content itself could prevent videos from following a certain pattern regarding its comments. According to our data, there has been a spike in cooking and DIY video comments amid this pandemic, and to further analyze our results, we proceeded with looking for a statistical model that would be a good fit for our data.

We attempted to fit two models to our data, a Gamma generalized linear model (GLM) and a Gamma generalized additive model (GAM). We used a Gamma model since the number of comments per day was a continuous variable with positive values exclusively. After plotting the fitted values for both models, it was determined that the GAM model better accounted for the randomness across time and modeled the data more accurately. While attempting to fit a GAM model for each channel, the model seemed to overfit the data, but then after several attempts of changing the smoothing parameter (k) values, we determined that GAM would be the best model for our data. Answering our research question, there is some evidence of a correlation between increased interaction with cooking and DIY YouTube videos and lockdown and social-distancing measures, implying that YouTube has been a go-to for many of us staying home amid COVID-19.