Proposal

Using data from YouTube, can we predict the number of views a video will receive using criteria such as location, category, number of likes, number of dislikes, and/or number of comments?

Introduction

YouTube videos are a billion dollar industry. The top 10 YouTube channels for half this year alone amassed over $162 million. The current top channel “Vlad and Nikita” earn around $312 thousand per video (<https://www.cashlady.com/youtube-league/>). Each video invariably contains at least one advertisement as well as most of the videographers have a separate sponsor who pays them directly. Understanding the underpinnings of what gets the most views is beneficial to anyone entering or already in the YouTube world.

Source of Data

The source of the data is kaggle (https://www.kaggle.com/datasnaek/youtube-new) and contains data on the daily trending videos from 2017 with up to 200 listed trending videos per day. We will be utilizing, at minimum, the json’s and CSV’s for Canada, United States, Great Britain, Germany and France, totaling 30581 unique values. The site has data for Japan, Mexico, Korea, India, and Russia; we will attempt to pull them in, if required, for a larger picture. However, the five already chosen datasets should be sufficient to cover the project statement.

Methodology

Utilizing R, we will

* Clean and prep the data for eventual linear regression models
  + Subsetting the datasets into the train/test structure.
  + Dealing with heteroscedasticity, which we will need to take into account (see graph below).
* Once the data is ready for modeling, we will perform at least three and up to five separate linear models on the subset datasets
* Chose the best performing model

