IST 707 Tending Youtube Videos Project

Maryse Khoury, Brian Taylor

11/13/2019

Goal:

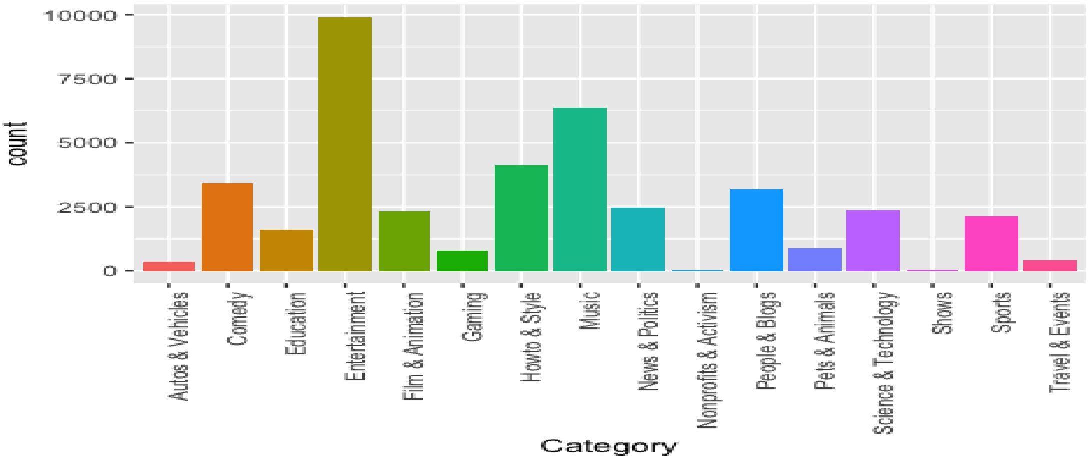
Identifying Social Media trends is big business. Multiple multi-million dollar careers have started and been built off of “going viral” with a video posted online. Advertising is a growing business for online video platforms, with about 86% of marketers using video to get their messages out. By the year 2023, money spent on video marketing is expected to surpass $100 billion. We want to identify where to best advertise on YouTube to get our message out. The two key questions we are looking to answer with this dataset are:

1. What key factors effect popular videos on YouTube?

2. How can we identify popular videos early on to maximize views of our ad in a popular video?

By using data collected from Kaggle about YouTube (found at: https://www.kaggle.com/datasnaek/youtube-new#USvideos.csv), we attempt to answer these questions.

From the 8 months of data, we identified how many of what category of videos were listed in YouTube’s trending videos data.

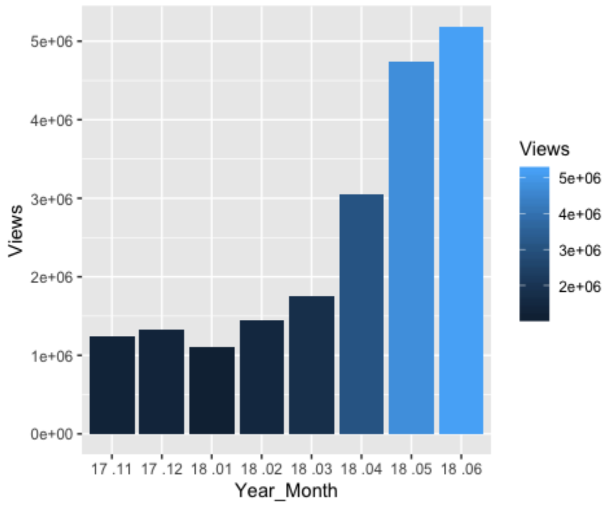


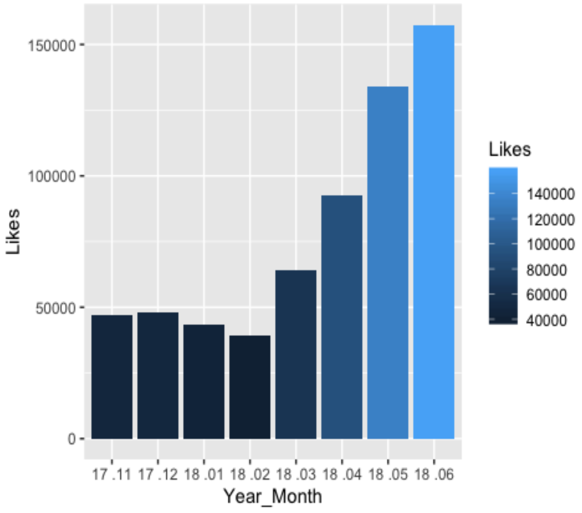
It is obvious that Entertainment, Music and How-to & Style are the top three categories of the top trending videos.

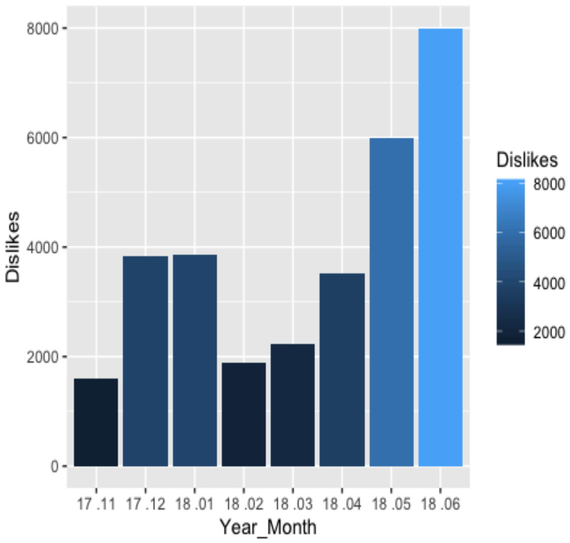
YouTube uses a proprietary algorithm to identify trending videos. Some trending videos are viewed more but not interacted with via likes, dislikes and comments.

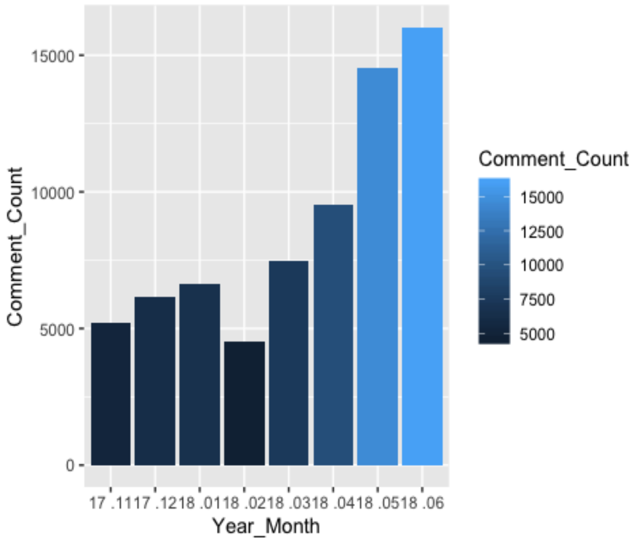
Our popularity index (“PI”) is calculated by summing likes, dislikes and comment counts, and then dividing that sum by the number of views for the video, representing populary, or engaging content that viewers spend time on.

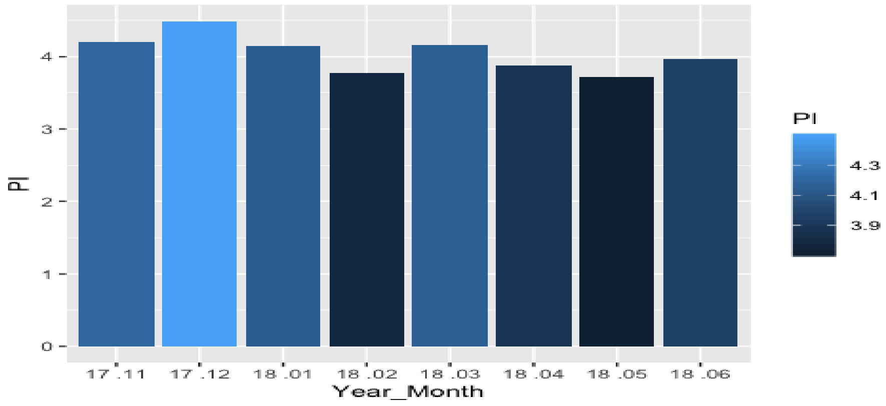
Below are aggregated average of views, likes, Dislikes, Comments and our variable "PI", meaning Popularity index, by month.







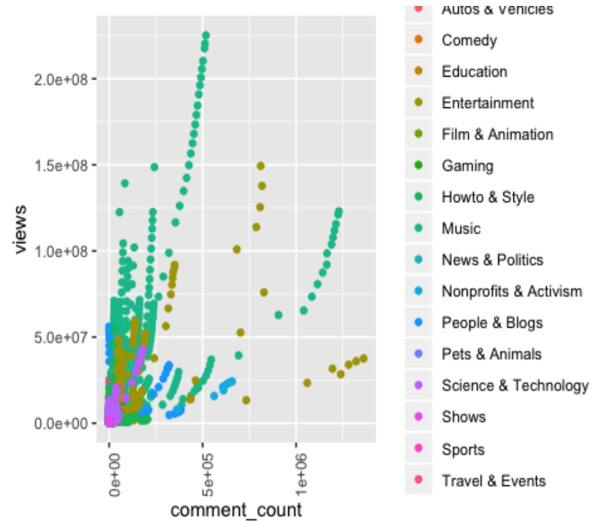


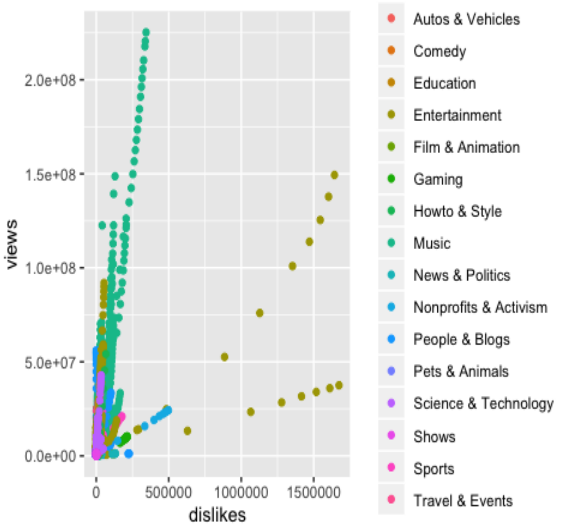


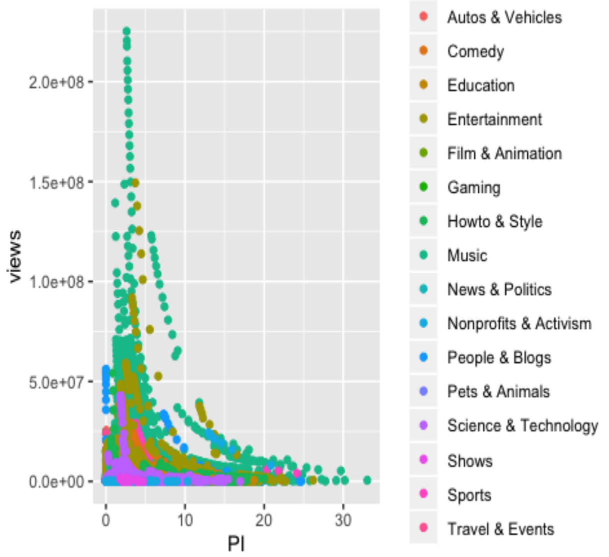
Notice that even though views and interactions vary by month, the PI remains relatively constant. This is partly by design, but also shows that the propensity for interacing with the videos does not appear seasonal, but consistent over time. Therefore, a large PI could indicate a true "viral" video.

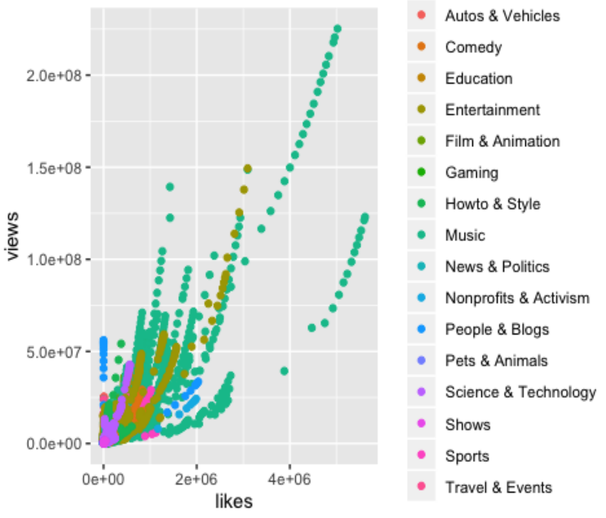
For reference, a larger PI is better.

The following charts display a variety of interactions by category.

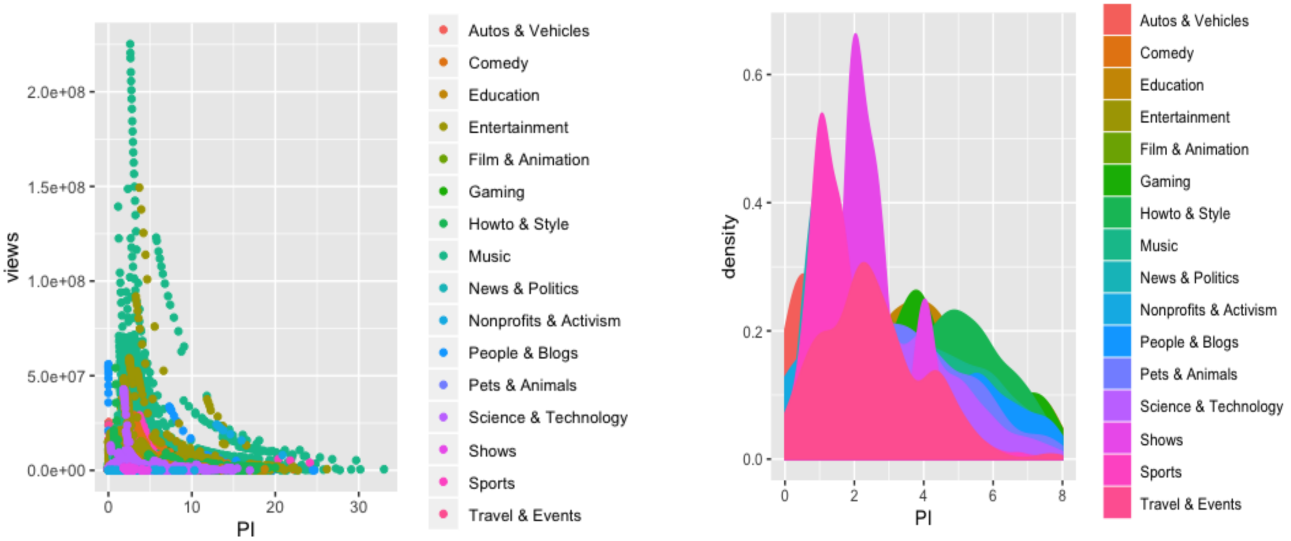




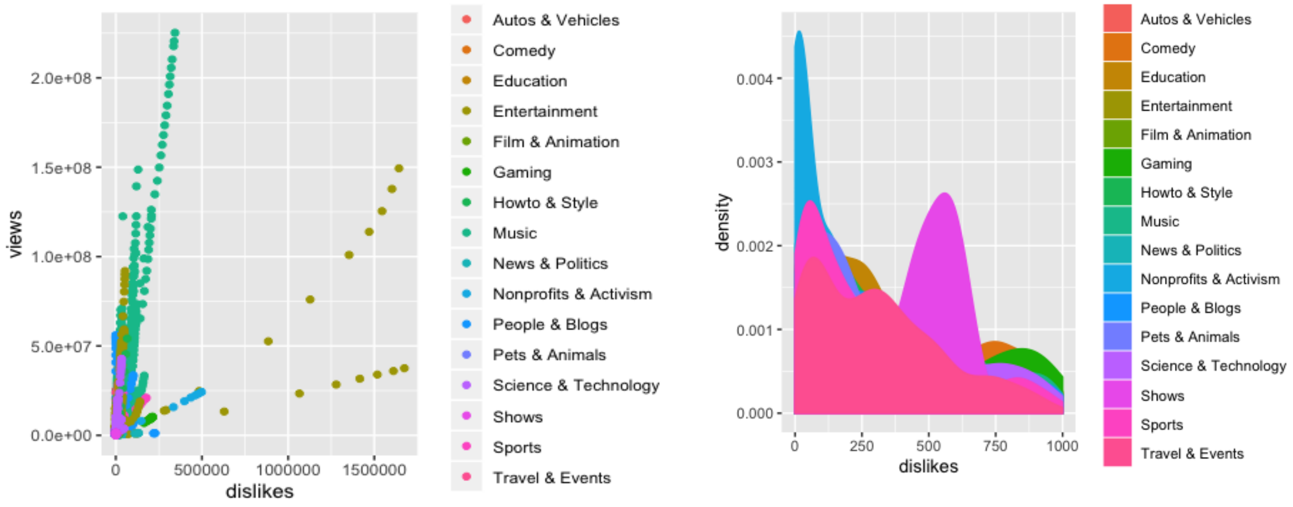




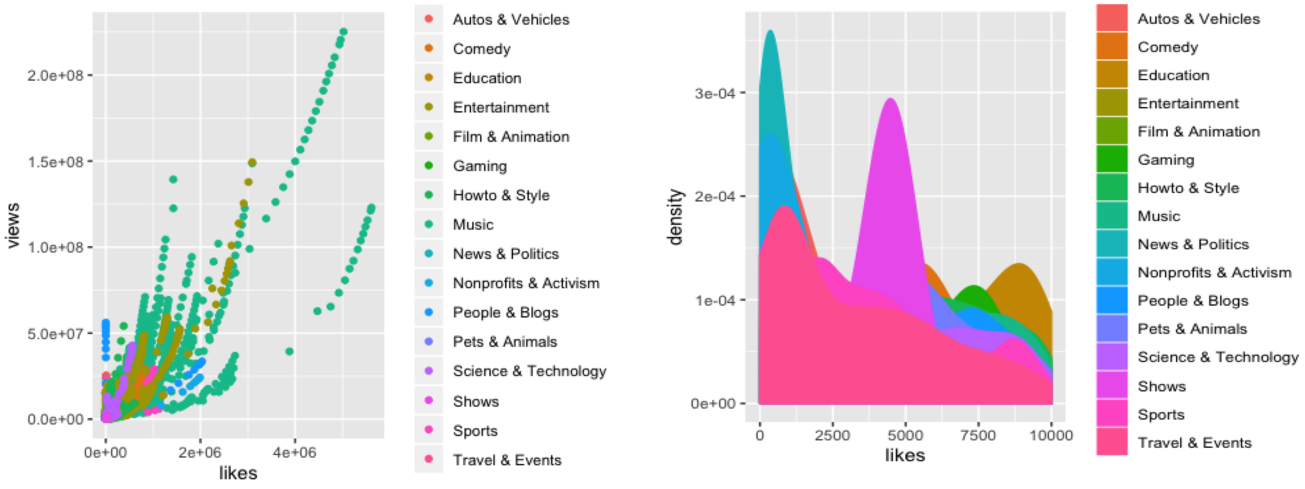
PI with respect to views by category.

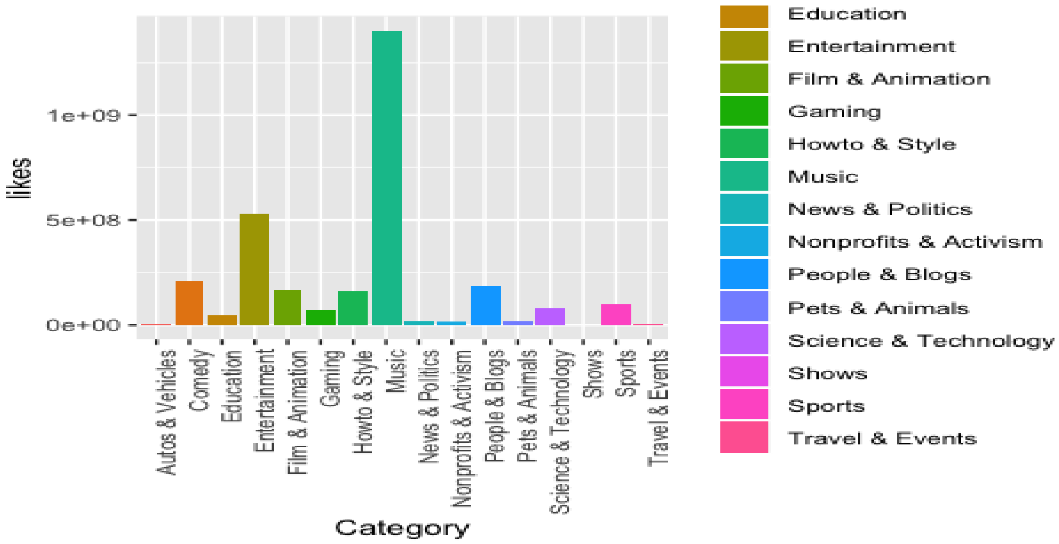


Dislikes with respect to views by category.



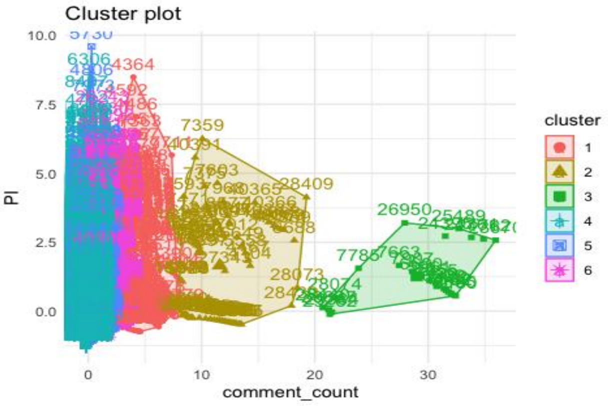
Likes with respect to views by category.



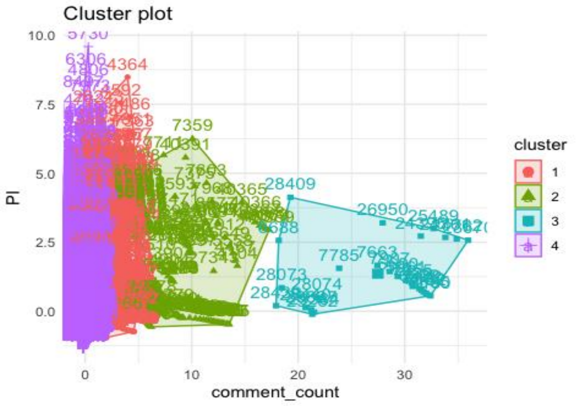


Utilizing Kmeans clustering, we compared K clusters at 6, 4, and 2 respectively. The elbow method noted below indicate that these k values should provide meaningful results. The number reported below each plot are the centers for each model.

K = 6 `## [1] 2.813572e+1`

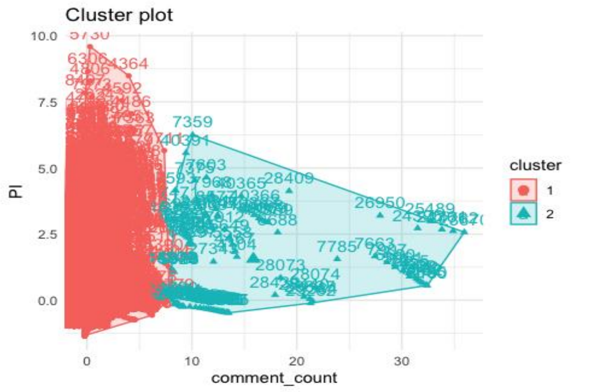


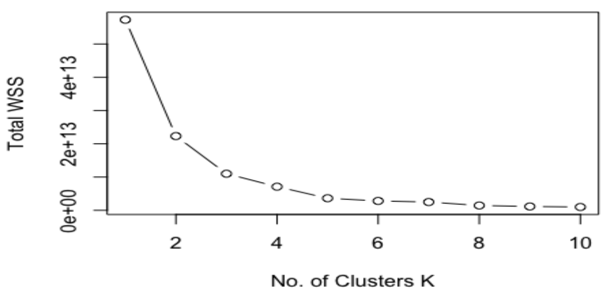
K = 4 `## [1] 7.10207e+12`



K=2 `## [1] 2.231194e+13`

Here is the elbow method mentioned above.





A Decision Tree model was evaluated using the C4.5 Algorithm.

```
## Confusion Matrix and Statistics
##            Reference
## Prediction   0   1   2   3
##          0 203  33   9   3
##          1  23 221  30  24
##          2   0  10  69  11
##          3   0   6  13 154
## Overall Statistics
##                Accuracy : 0.7998
##                  95% CI : (0.7705, 0.8268)
##     No Information Rate : 0.3337
##     P-Value [Acc > NIR] : < 2.2e-16
##                   Kappa : 0.7234
##  Mcnemar's Test P-Value : 4.813e-06
## Statistics by Class:
##                      Class: 0 Class: 1 Class: 2 Class: 3
## Sensitivity            0.8982   0.8185  0.57025   0.8021
## Specificity            0.9228   0.8571  0.96948   0.9692
## Pos Pred Value         0.8185   0.7416  0.76667   0.8902
## Neg Pred Value         0.9590   0.9041  0.92768   0.9403
## Prevalence             0.2794   0.3337  0.14957   0.2373
## Detection Rate         0.2509   0.2732  0.08529   0.1904
## Detection Prevalence   0.3066   0.3684  0.11125   0.2138
## Balanced Accuracy      0.9105   0.8378  0.76986   0.8856
```

PI values were binned into classes of popularity.

Classes of popularity correlate to quartiles of the PI values, in that a PI = 0 is the 1st quartile, or Class 0 of popularity. Further, PI = 1 is second quartile or Class 1 of populariy, PI = 2 is the third quartile or Class 2 of popularity and PI = 3 is the fourth quartile, Class 3. We are interested in Class 3, or the fourth quartile, for advertizing purposes.

The accuracy of this model is surprising given the simplicity. At 79.98% accuracy, and a Kappa of .72, this appears to be a fairly accurate model, with minimal potential for over-fitting the data.

A Random Forest model was also evaluated.

```
## Call:
##  randomForest(formula = pop ~ ., data = P2_train, ntree = 100)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 1
##         OOB estimate of  error rate: 56.97%
## Confusion matrix:
##     0   1  2   3 class.error
## 0 386 330  6 124   0.5437352
## 1 258 595 26 247   0.4715808
## 2  76 214 26 153   0.9445629
## 3  74 310 26 386   0.5150754
```

The accuracy of this model is surprisingly low given. The more trees that were added, the error rate grew. This was in part due to a lack of use of channel titles in developing this particlaur model due to limitations within the random forrest itself. With an error rate greater than 50%, you are better off flipping a coin.

So a Naive Beyes model was developed and tested.

```
## Confusion Matrix and Statistics
##           Reference
## Prediction   0   1   2   3
##          0 109  24   6   5
##          1  20 123  17  18
##          2   1  10  55   9
##          3   9  16   6  93
## Overall Statistics
##                Accuracy : 0.7294
##                  95% CI : (0.689, 0.7671)
##     No Information Rate : 0.3321
##     P-Value [Acc > NIR] : <2e-16
##                   Kappa : 0.6305
##  Mcnemar's Test P-Value : 0.2681
## Statistics by Class:
##                      Class: 0 Class: 1 Class: 2 Class: 3
## Sensitivity            0.7842   0.7110   0.6548   0.7440
## Specificity            0.9084   0.8420   0.9542   0.9217
## Pos Pred Value         0.7569   0.6910   0.7333   0.7500
## Neg Pred Value         0.9204   0.8542   0.9350   0.9194
## Prevalence             0.2668   0.3321   0.1612   0.2399
## Detection Rate         0.2092   0.2361   0.1056   0.1785
## Detection Prevalence   0.2764   0.3417   0.1440   0.2380
## Balanced Accuracy      0.8463   0.7765   0.8045   0.8329
```

The Naive Beyes model performed similarly to the decision tree model, with slightly less accuracy as well as a lower Kappa value. Like the decision tree model, this model accounted for the YouTube channel name and reinforces the idea that the channel the video is posted on has a large effect on the popularity of a video.

Next a k-Nearest Neighbor model was utilized.

```
##     test_pop          ##     test_pop          ##      test_pop
## knn1   0   1   2   3  ## knn5   0   1   2   3  ## knn10   0   1   2   3
##    0 279  68   0   0  ##    0 250  98   0   0  ##     0 199 133   0   0
##    1  49 296  29   1  ##    1  78 270  30  19  ##     1 129 239  49  18
##    2   0  28  46  18  ##    2   0  23  31  13  ##     2   0  19  22  32
##    3   0   1  22 163  ##    3   0   2  36 150  ##     3   0   2  26 132
## [1] 0.784             ## [1] 0.701             ## [1] 0.592
```

K nearest neighbors proved as accurate as the decision tree and Niave Beyes models, however, as the nearest neighbors values increased, the accuracy decreased. A 78.4% accuracy for 1 neighbor is on par with the previous models.

Finally, an SVM model was utilized

First: A Polynomial Kernel

```
Reference
Prediction   0   1   2   3
         0   0   0   0   0
         1 193 293 122 200
         2   0   0   0   0
         3   0   0   0   0
Overall Statistics
               Accuracy : 0.3626
                 95% CI : (0.3294, 0.3968)
    No Information Rate : 0.3626
    P-Value [Acc > NIR] : 0.5133
                  Kappa : 0
 Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.0000   1.0000    0.000   0.0000
Specificity            1.0000   0.0000    1.000   1.0000
Pos Pred Value            NaN   0.3626      NaN      NaN
Neg Pred Value         0.7611      NaN    0.849   0.7525
Prevalence             0.2389   0.3626    0.151   0.2475
Detection Rate         0.0000   0.3626    0.000   0.0000
Detection Prevalence   0.0000   1.0000    0.000   0.0000
Balanced Accuracy      0.5000   0.5000    0.500   0.5000
```

Second: A Radial Kernel

```
Call:
svm(formula=pop~.,data=svm_train,type="C-classification", kernel = "radial")
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
Number of Support Vectors:  3125
Confusion Matrix and Statistics
          Reference
Prediction   0   1   2   3
         0   0   0   0   0
         1 193 293 122  76
         2   0   0   0   0
         3   0   0   0 124
Overall Statistics
               Accuracy : 0.5161
                 95% CI : (0.481, 0.5511)
    No Information Rate : 0.3626
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.2612
 Mcnemar's Test P-Value : NA
Statistics by Class:
                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.0000   1.0000    0.000   0.6200
Specificity            1.0000   0.2408    1.000   1.0000
Pos Pred Value            NaN   0.4284      NaN   1.0000
Neg Pred Value         0.7611   1.0000    0.849   0.8889
Prevalence             0.2389   0.3626    0.151   0.2475
Detection Rate         0.0000   0.3626    0.000   0.1535
Detection Prevalence   0.0000   0.8465    0.000   0.1535
Balanced Accuracy      0.5000   0.6204    0.500   0.8100
```

SVM radial kernel model resulted in higher accuracy at 51.61% than the polynomial kernel, which is at 36.26%. Nevertheless, noting these results, the svm model is not the best model predicting results for the trending Youtube videos dataset, as the models implemented before gave higher accuracy.

Conclusion

Multiple key factors that effect video popularity on YouTube were identified. A combination of the channel sponsoring the video, as well as the category of video play a large role in determining the viralness of a video. Using these attributes, models with up to almost 80% accuracy were developed to predict videos with a high populairty index (PI). Areas for further exploration would be into if certain “tags” can be used to improve accuracy of the predictions as well as applications in markets outside of the United States.