Relationship of Likes/Dislikes on Comments of a Trending Youtube Video

Sanchit Singhal
December 17,2018
SDS358

Introduction

Objectives: Investigate if, in the US region, after controlling for category type and number of views, the number of likes and number of dislikes account for a significantly greater proportion of variance in the number of comments on a trending YouTube video?

Motivation: By understanding the relationship between likes/dislikes on the number of comments on a Youtube video, we can create content that is more engaging with its audience.

Hypotheses: The number of likes and number of dislikes on a trending Youtube video will account for a significantly greater proportion of variance in the number of comments over and above its category type and the number of views.

Methods

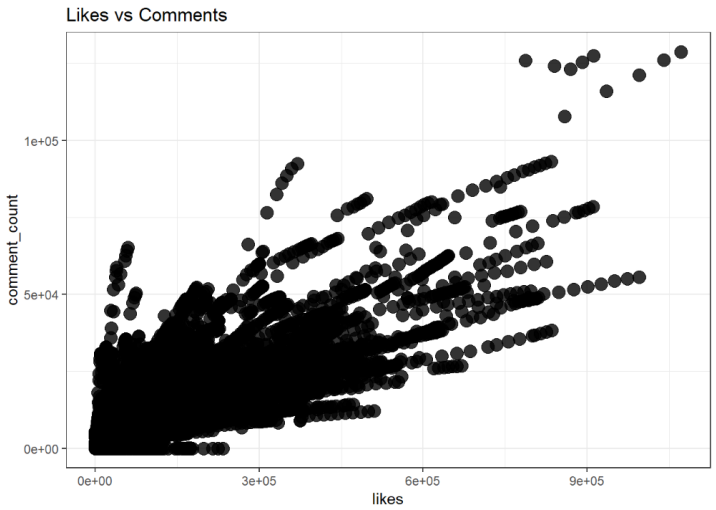
Sample: Initial dataset was subset to just US region: 40949 trending Youtube videos with information about Number of Comments (quantitative), number of Likes (quantitative), number of Dislikes (quantitative),Category Type (categorical), Number of Views (quantitative). 1235 outliers, with high Cook's Distance scores (above 0.03) were removed leaving us with a total of 39714 of sample records that were fitted to each model.

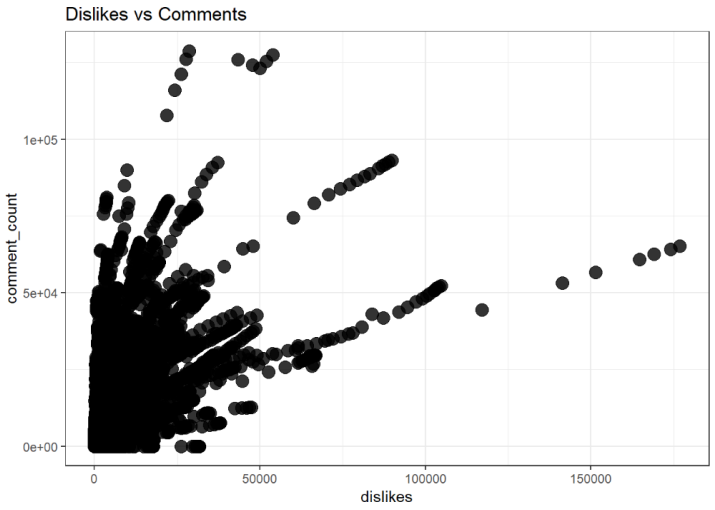
Analysis Method: Sequential Multiple Regression Analysis with several models runs and use of anova function in R

Descriptives

Variables in Models:

	Type	Mean	SD
# of Comments	Response	5067.37	9004.09
# of Likes	Independent	50655.91	93052.86
# of Dislikes	Independent	2154.57	5633.08
# of Views	Nuisance	50655.91	93052.86
Category type	Nuisance	Total of 15 category types	





Results

Overall Sequential Results:

Model	R2	RSS	Df	F-value	P-value
Cat, View	0.4368	1.81e12			
Cat, View, Likes, Dislikes	0.7792	7.11e11	2	30794	<0.05

Full Model – Examining variables of interest:

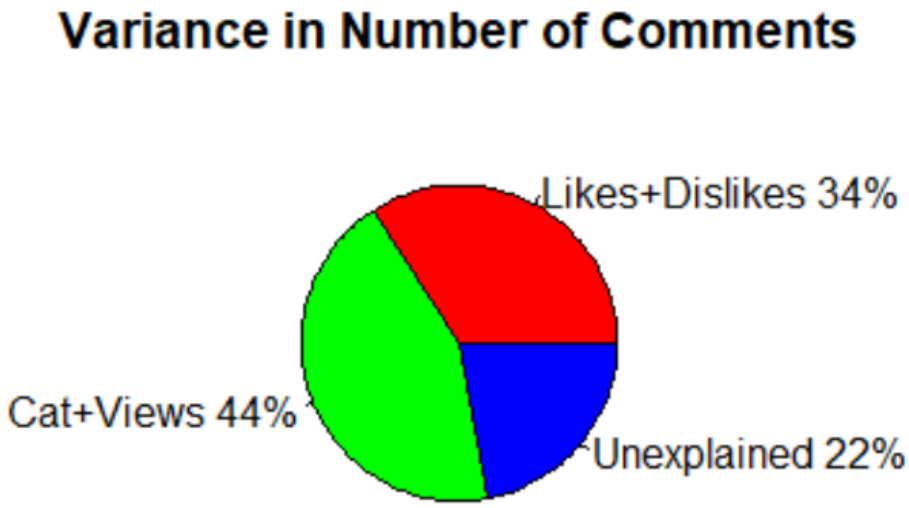
Predictor	Estimate	SE	t-value	P-value
Likes	0.0923	4.18e-4	220.998	<0.05
Dislikes	0.5054	4.88e-3	103.535	<0.05

Examining Variable Impact – Standardized Betas:

Predictor	Estimate
Likes	0.9535
Dislikes	0.3162

Examining Variable Impact – Part Correlation Squared:

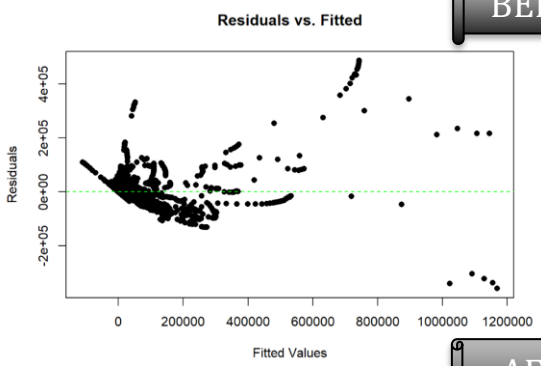
Predictor	Estimate
Likes	0.2716
Dislikes	0.0596

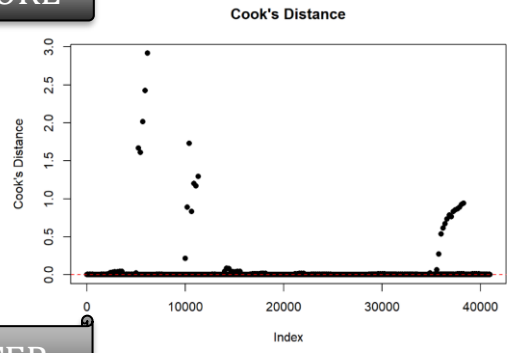


Assumptions

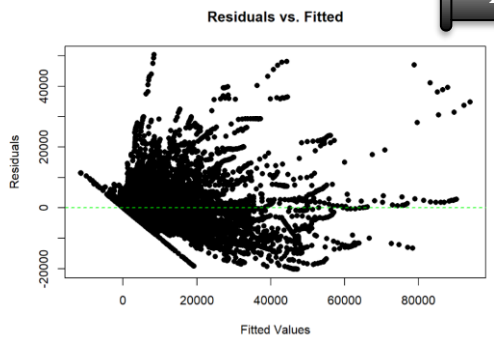
Assumptions: Homoscedasticity was confirmed visually via a Residual vs.Fitted Plot. Cook's Distance was used to remove outliers. VIF was used to make there is no multi-collinearity.

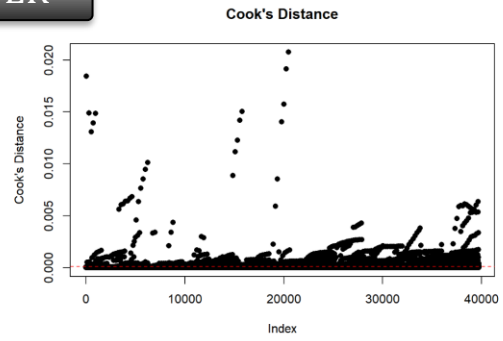
BEFORE





AFTER





VIF

	GVIF	Df	GVIF^(1/(2*Df))
category_id	1.169008	14	1.005593
views	3.762690	1	1.939766
likes	3.347894	1	1.829725
dislikes	1.677295	1	1.295104

Discussion

Interpretation: The initial model, containing only the nuisance variables of category type and views accounted for 43.68% of variance in the number of comments, $F(15,39698)=2053$, $p<0.05$. The second model showed that the addition of Likes and Dislikes in the model accounts for an additional 34.25% of variance for a total of 77.92%. This change in proportion was confirmed to be significant via ANOVA, $F(2,39696) = 30794$, $p<0.05$. Although both variables of interest were significant, Number of Likes ($b=0.0923$, standard beta = 0.95, $t(39696)= 220.998$, $p<0.05$) has a bigger impact on the number of comments than Number of Dislikes ($b=0.5054$,standard beta = 0.32, $t(39696)= 103.535$, $p<0.05$) and can uniquely account for around 27% of variance explained while Dislikes accounts for only 6% of variance in the number of comments on a video.

Limitations: Although the model did not fail any assumptions, it does not take into account other factors that may impact the number of comments such as time since posted, etc. The category type may have been a confounding variable that could have impacted the results as well.

Implications: The method used in this study was sufficient to answer the research question. This data was specifically for the US region which could lead to biased results that do not generalize to everyone. Even though the research question was targeted at the US region, I would like to continue additional research into how these results compare to other regions and what that means for the internet as a whole – this might give us a better insight into comments on videos and how to enable content creators to upload videos that are more engaging with their audience.

References: Data sourced from Kaggle platform:

J, M. (2018, November 20). Trending YouTube Video Statistics. Retrieved from https://www.kaggle.com/datasnaek/youtube-new/data.