YOUTUBE SUBSCRIBES AND LIKES

Justin Lee
Data Science at Indiana University

Abstract

YouTube has attracted many users with its voluminous videos and convenience of sharing and watching videos freely (2011). In accordance with its popularity, many streamers have increased their subscribers and likes by sharing quality content videos. In their videos, they use the phrase, “If you liked the video, please press like and subscribe to my channel.” This paper studies if people do really subscribe to the channel if they liked the video. By seasonal ARIMAX model, I failed to find evidence that people subscribe to the channel to the videos they liked. However, the study is incomplete and requires further research to be robust.

Introduction

YouTube is an online video sharing platform where people can upload and share their content with a few clicks (2011). Currently, it is the major platform where over 5 billion videos are watched per day. (2019) The biggest factor that made YouTube successful is users can customize their videos based on their interests and watch the videos they like.

How users can express their interest is by giving likes, dislikes or comment on the videos they watch. When a user liked most of the videos of a certain channel, users can also subscribe and get updates for every video he or she uploads. For YouTube streamers, it is best to get likes and comments on their videos and subscribe to their channels as those counts are proportional to their income from YouTube.

Due to this, streamers ask their viewers to press like and subscribe if they liked the content. My hypothesis of interest arose from this phrase. Will people really press subscribe if they liked the video? Will there be any evidence of avoiding subscribe although the viewer liked the video? I will fit the time series regression model with

YouTube data and try to predict subscribe counts with likes, dislikes, comments of the videos and view counts of the channel.

Methods

Data

The original data was Top 5000 South Korean YouTube channels' videos measured on Dec 15, 2018 ~ Jan 20, 2019. (2019) It contains the number of likes, dislikes and comment for each video and subscribes and view counts for the channel. It is a public dataset that can be accessed in the Kaggle data repository. (2019)

The raw data needed cleaning and wasn't organized. The channel database had a dimension of 1,298,453 rows and 5 columns (index, channel ID, subscriber count, view count and time) and the video database had dimension of 12,788,745 rows and 7 columns (index, video ID, comment count, dislike count, like count, view count and time) For this class project, I randomly chose one channel among 5000 and *Buzzbean11* was selected. He is a famous YouTuber for gaming shows and his initial subscriber count in the data was over 1.9 million.

Using SQL, I filtered the videos that *buzzbean11* has uploaded and summed up likes, dislikes, and comments he got for each time to get the overall counts that *buzzbean11* earned at each time frame. For instance, if *buzzbean11* had 3 videos with 2 likes, 1 like, 4 likes at 5 pm on Jan 1st, it was added up to 7 likes for that time. The final data I used for this analysis was 292 rows and 5 columns with subscribe counts, like counts, dislike counts, comment counts, view counts and time.

There was a time frame that had NA value at 2019-01-04 18:31:25.478978 for subscriber and view counts. I imputed using simple arithmetic average (SAA) methods on time before (t-1) and time after (t+1) and made .csv file. (Moritz, S., 2015) This csv is attached in .zip file.

Time Series Plot

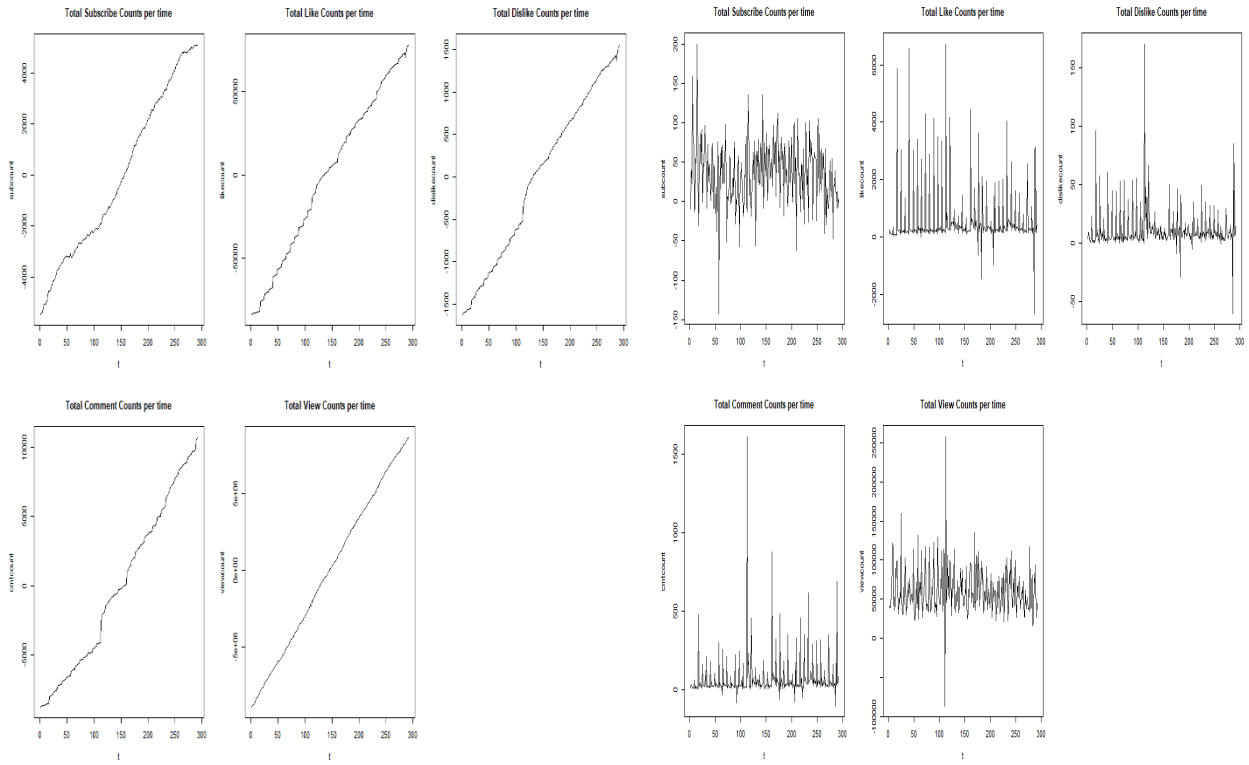

Figure 1 Time Series Plot: Left is the raw time series plot and the right is first-order differentiated.

I plotted the raw data to see if there are any linear or seasonality pattern in any of the variables. Figure 1 shows the total counts (centered at 0) accumulated for each time frame. It decreased at some time points but overall it had a linear trend. For the like, dislike and comments, there seems a sudden increase at time point 100-150. I took a difference to remove the linear trend and see an increase/decrease per time.

By plotting differentiated variables, there were sudden bumps for every variable, which is telling the data is not stable and need further tests. (Figure 1 Right) I decided to see autocorrelation, cross-correlation and spectral analysis.

Autocorrelation & Cross Covariance Plot

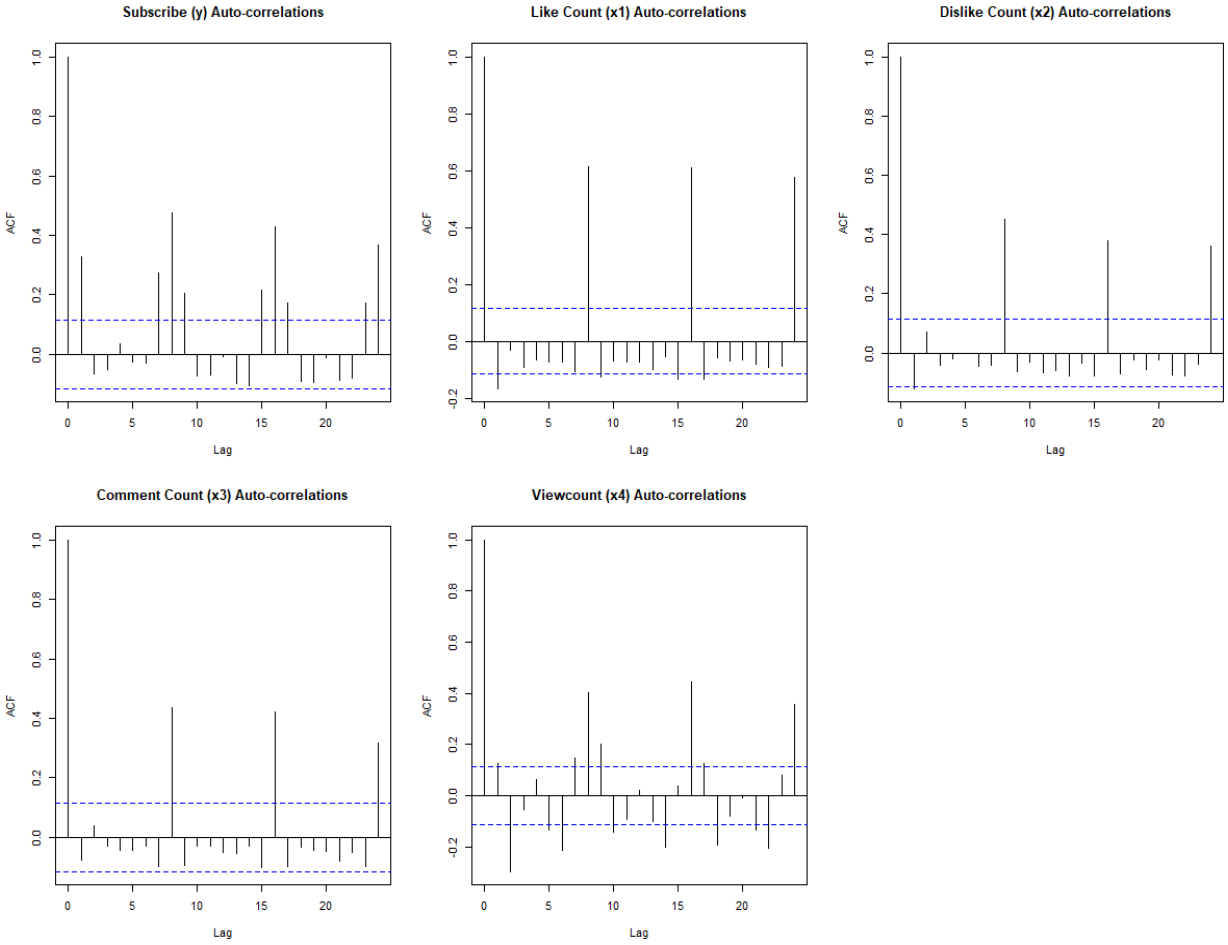

Figure 2 Autocorrelations for Variables

For every 8 lags, there was autocorrelation for each of the variable. (Figure 2, Supp 1) This corresponds to the seasonal effect as every 8 lags * 3 hours equal to one day. However, for subscribes and view count there was also autocorrelation in the 7th and 9th lag.

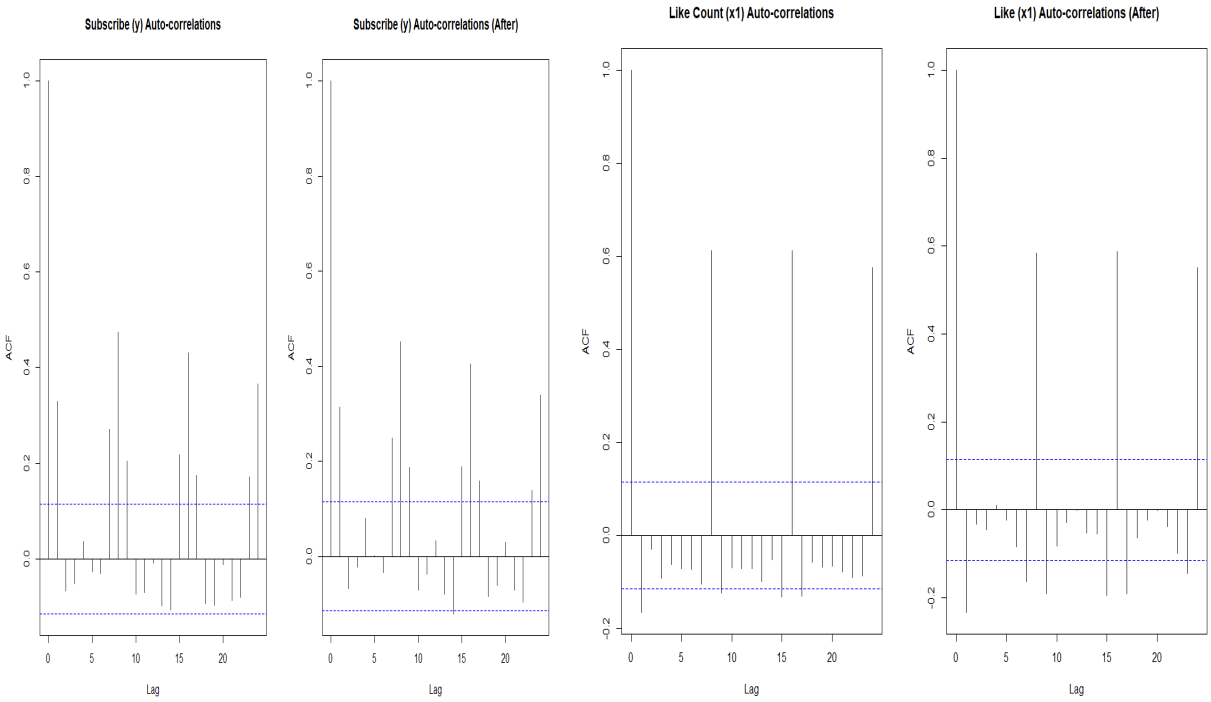

Figure 3 Autocorrelation Plot with Spectral Fit: Left is before the spectral fit while the right is after fitting.

	Before	After
Subscriber	0.474	0.451
Like	0.613	0.585
Dislike	0.453	0.43
Comment	0.436	0.408
View Count	0.402	0.297

Table 1 Comparison of Autocorrelation: The values are autocovariance values at lag term 8 before and after fitting sin and cosine with frequency 8.

To reduce the auto-correlation, I tried fitting for a frequency of 8 with $(sin(2\Pi t/8) + cos(2\Pi t/8)$, it decreased but not much. (Figure 3, Table 1)

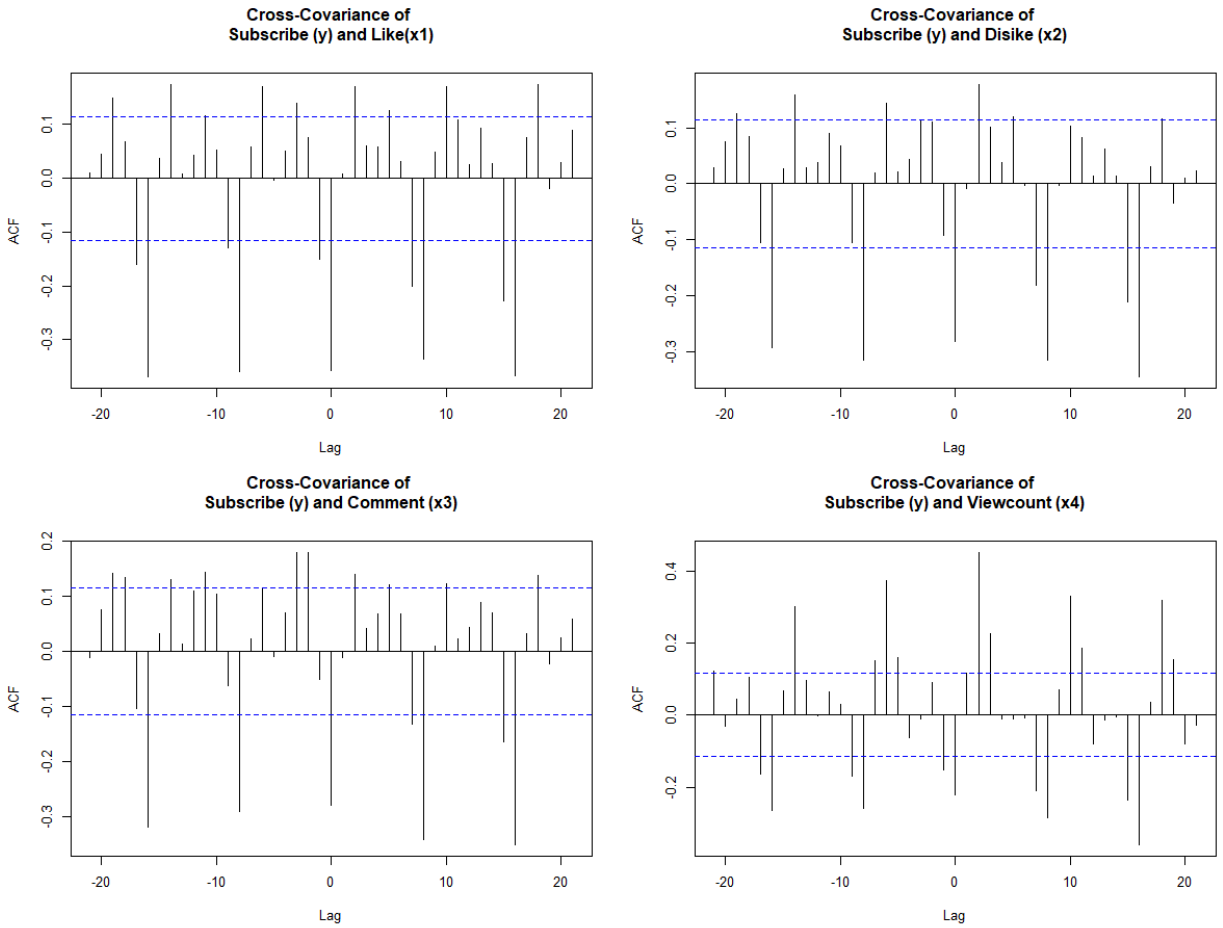

Figure 4 Cross Covariance between Subscribe and X

There are high cross-covariance between subscribe and other variables in lag 0, -1, -8 and -9, 1, 2, 8 and 9. X is both leading and lagging y. (Figure 4, Supp 2)

Spectral Analysis

By plotting autocorrelation plot, it seemed to have a seasonal pattern at lag 8. I performed spectral analysis to validate if it is seasonal and the frequency of its pattern.

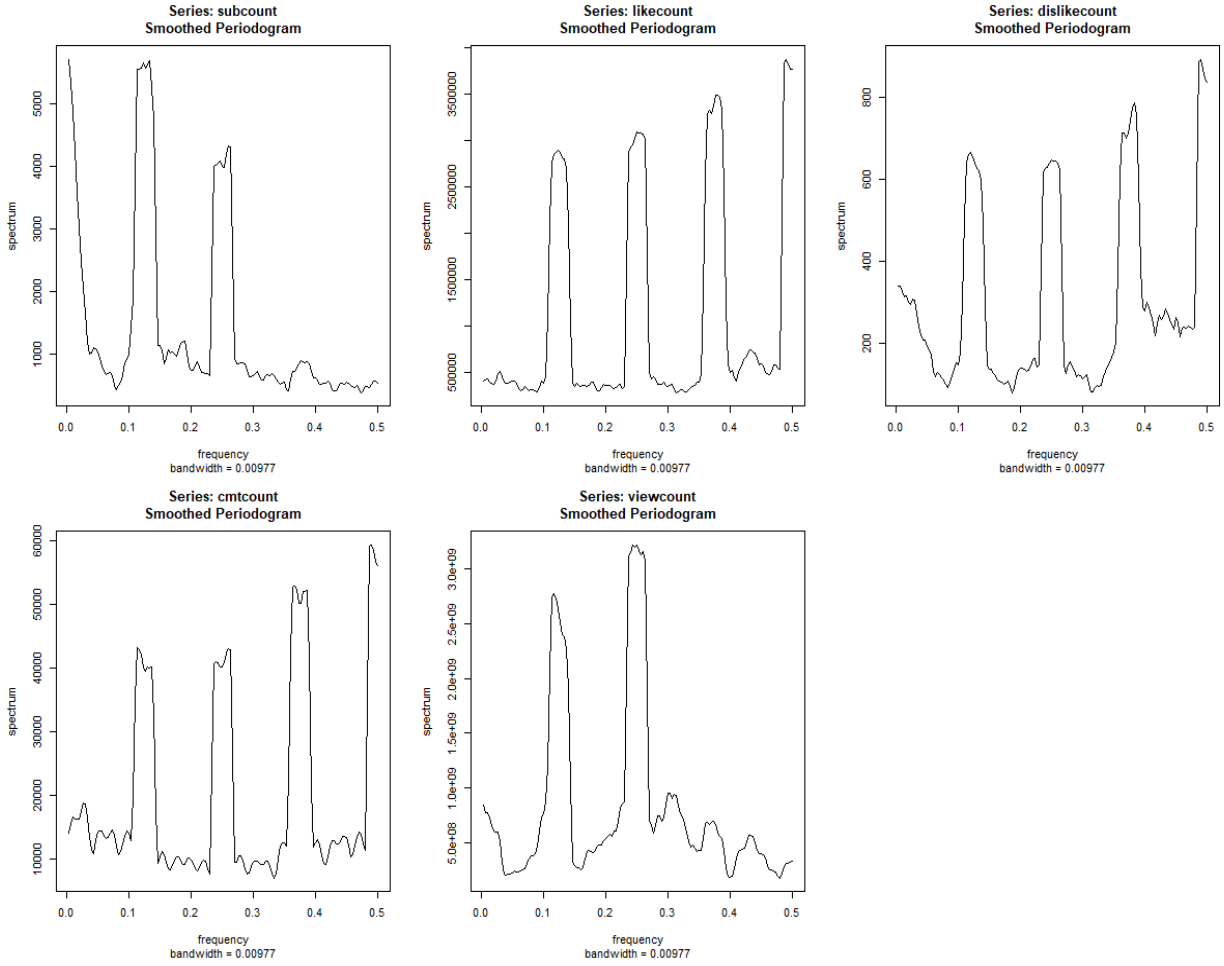

Figure 5 Spectral Analysis with Window Size(= Span) of 10

We can view that there is a seasonal component at frequency 0.123. It is going to be fitted using seasonal ARIMAX model with seasonal period at 8. (Fig 5, Supp 3)

Removal of Seasonal Pattern

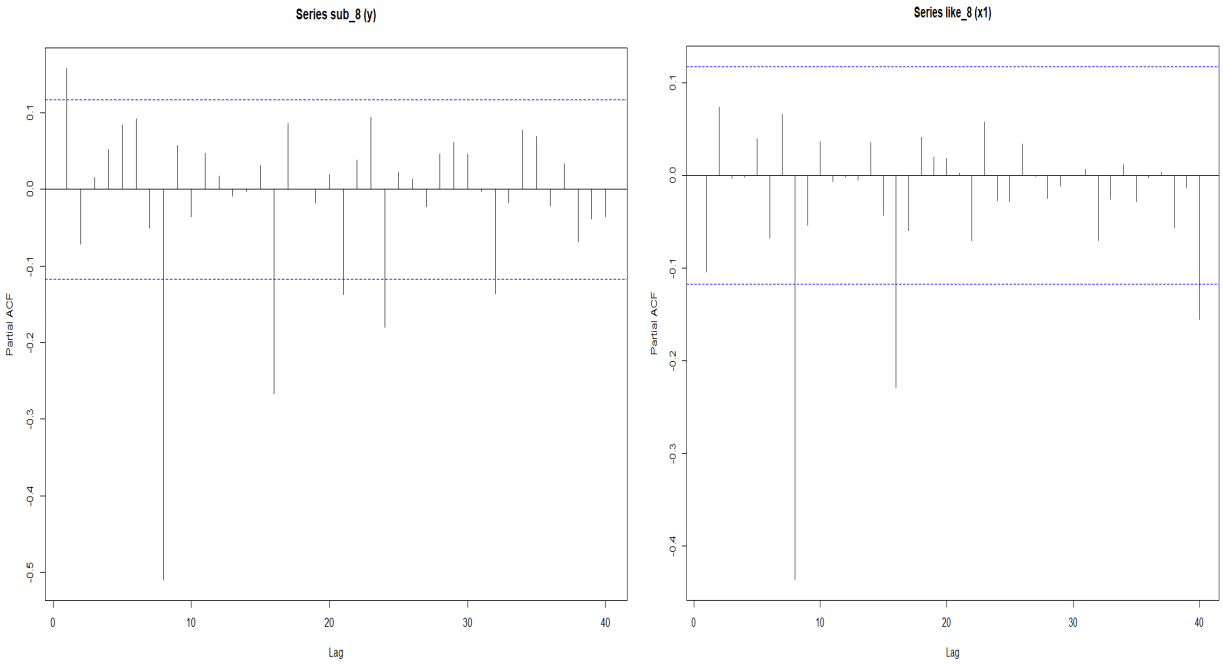

Figure 6 Partial Autocovariance of Subscribe and Like After 8th order differentiation

As taking difference removes the seasonal component, I tried diff(subcount,8) and the same for the like count and saw partial-autocovariance (pacf). For the seasonal term, there need to be 3 seasonal autoregressive terms, SAR (3) and for unseasonal autoregressive, it seems that you only need AR(1).

Seasonal ARIMAX Model

Based on observation of autoregressive and spectral analysis, I decided to fit the seasonal ARIMAX model. There seem to be 3 seasonal autoregressive terms and 1 non-seasonal autoregressive term based on ACF, PACF and Spectral analysis.

Results

	AICval	BICval
(1,0,1,2,1,0)	2268.556	2327.077
(2,0,1,2,1,0)	2268.454	2330.418
(1,0,1,2,1,1)	2252.291	2314.255
(1,0,1,2,1,2)	2254.254	2319.66
(2,1,2,2,2,2)	2225.8	2297.256

Table 2 Model Comparison based on BIC and AIC Values

For model comparison, I changed ARIMA terms and seasonal ARIMA terms and compared AIC, BIC and log-likelihood values to find the best model. Table shows different terms in the order of *(AR, I, MA, SAR, SI, SMA)*. The fifth model had the lowest AIC and BIC values. I didn't compare log-likelihood values as my model is non-nested. My best model by the comparison of these values was the 5^{th} model (2,1,2,2,2,2).

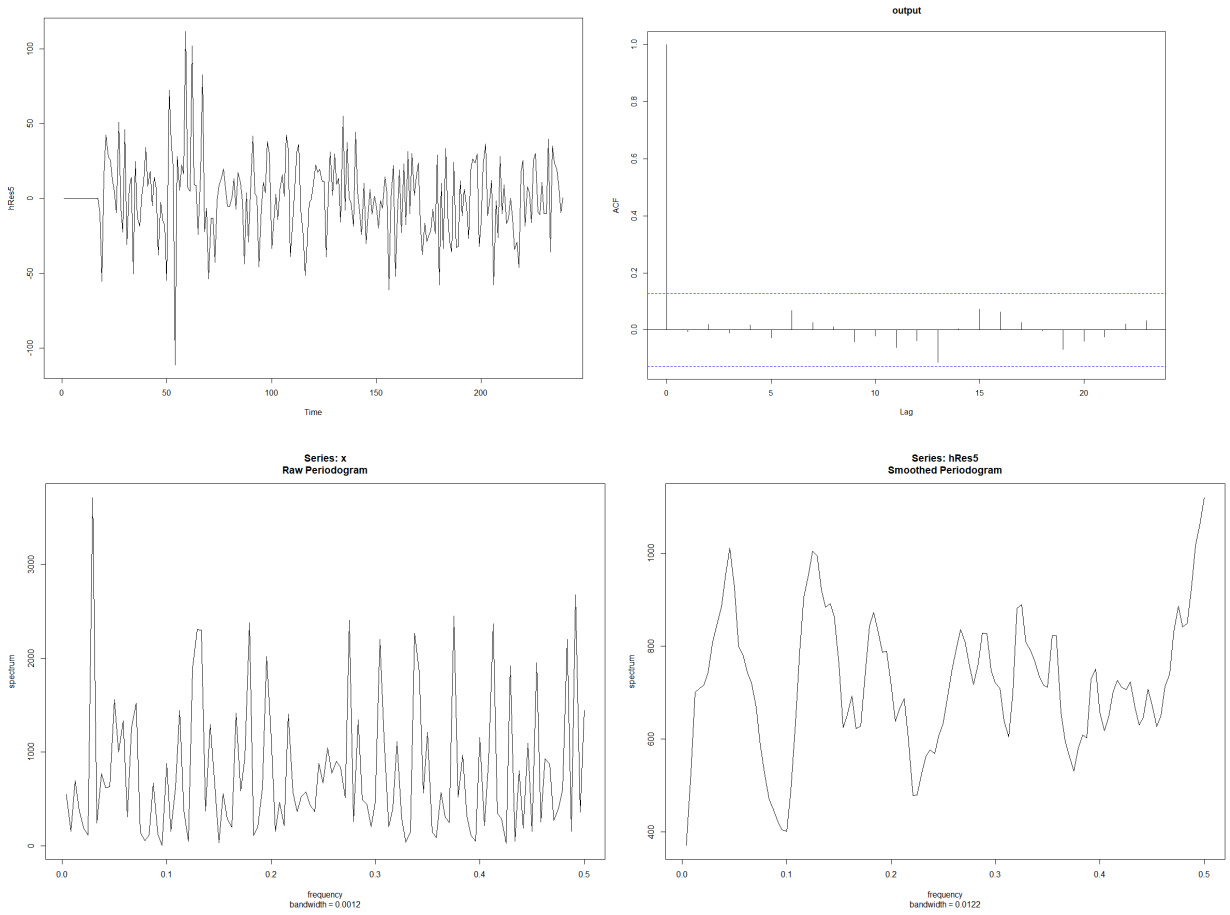

Figure 7 Stationary Test: Plot, Autocovariance, Raw Periodogram, Smoothed Periodogram

I checked the residuals to check if they are white noise. There weren't any spikes on any of the plots, which seems like white noise. Box-Pierce test p-value was >0.05 which is far from being significant. (Supp 4)We found evidence that the residuals are stationary. (Fig 7)

Now, I put aside 50 points to see if my model can forecast the future subscriber counts based on all X's. It was far from the truth, which means the model can't be used to predict the future subscriber counts.

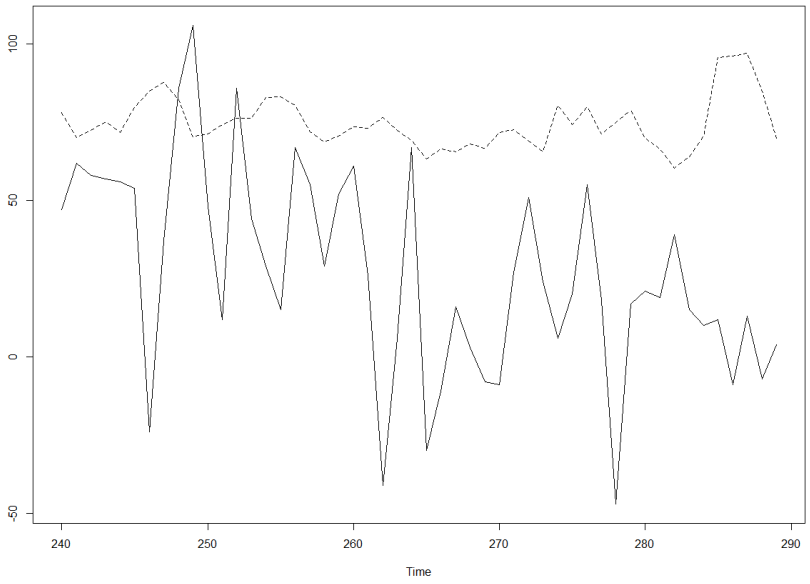

Figure 8 Forecasting Based on the Model : 50 data points were put aside and predicted based on our model.

Conclusion & Discussions

Overall, looking at the coefficients and standard error of different X's, it was not significant that they included zero in 95% C.I., which is b ±1.960* s.e (Supp 5). The model had many insignificant terms and therefore couldn't perform well in forecasting.

This study has failed to reject the hypothesis that *"People who liked the video will not press subscribe"*. Thinking of the possible explanations, it requires more consideration to click subscribe than clicking like. You must be very certain of this channel's content and quality to press subscribe, otherwise; every updated video will keep bothering the user to watch the new video. Comparatively, pressing like are less vulnerable to these stressful situations in the future, thus; people will press like easily than subscribe.

However, we cannot conclude that subscribe and X's have no correlation at all since there are some limitations in my study. As discussed in the data section, I randomly chose one channel to avoid excessive data cleaning. It is possible that *buzzbean11* was the outlier or my study and my study was biased. If I used unbiased

data, I might have found a significant relationship. Also, factor variables such as video category, tags and thumbnails should be included in the model as these may be an important factor for deciding if the user would like to subscribe to the channel or not. For instance, categories are an important factor of chosen to be subscribed or not: news category channels will be more likely to be subscribed than the gaming channels as news are a daily part of life for most of the people while gaming shows don't. Finally, the like counts, dislike counts, and comment counts should be counted for each video. By summing up all the of likes, dislikes, and comments per time frame, I might have lost specifics happening in each video and fitted the model poorly. This was my exploratory analysis for using time series and YouTube data, so future work may be done to make it more solid.

ar1	**ar2**	**ma1**	**ma2**	**sar1**	**sar2**	**sar3**
0.624231	-0.19344	-1.40418	0.491918	-0.03807	0.096511	-0.01953
sma1	**sma2**	**likecount**	**like_1**	**like_2**	**dislikecount**	**dislike_1**
-1.95479	0.987419	-0.00119	-0.00215	-0.00704	0.106135	0.202953
dislike_2	**viewcount**	**view_1**	**view_2**	**cmtcount**	**cmt_1**	**cmt_2**
0.263409	0.000184	6.31E-05	0.000163	0.017594	0.021694	0.032617

Reference

1. YouTube, L. L. C. (2011). YouTube. *Retrieved*, *27*, 2011.
2. 37 Mind Blowing YouTube Facts, Figures and Statistics – 2019. (2019, February 19). Retrieved from https://merchdope.com/youtube-stats/
3. K. (2019, January 20). Youtube 5000 channels' videos Daily count every 3h. Retrieved from https://www.kaggle.com/nnqkfdjq/youtube-5000-channels-videos-daily-count-every-3h
4. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. arXiv preprint arXiv:1510.03924.
5. Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?. Journal of interactive marketing, 26(2), 102-113.