Scan to know paper details and
author's profile

Association between Cyberbullying Crude Comments and the Number of user Subscribers and uploads on YouTube

Srimayee Dam

Columbia University

ABSTRACT

This article entails a quantitative analysis of the association between cyberbullying crude and number of user subscribers and uploads on YouTube. Of the several numbers of variables listed in the said dataset from Kaggle.com, the variables of comments, subscribers and uploads were found to be continuous. The data included 3464 YouTube user ids, and were analyzed using Descriptives, Normality Tests and Spearman's Rho on SPSS. The findings included non-linear, positively skewed, non-normal distribution. The correlation analysis depicted a slight positive association between User subscribers and rude comments, whereas no connection between User uploads and crude comments. Future work to focus on other factors influencing hurtful commenting on YouTube.

Keywords: NA

Classification: LCC Code: QA76.9.H85

Language: English

Great Britain
Journals Press

LJP Copyright ID: 975815
Print ISSN: 2514-863X
Online ISSN: 2514-8648

London Journal of Research in Computer Science & Technology

Volume 25 | Issue 5 | Compilation 1.0

Association between Cyberbullying Crude Comments and the Number of user Subscribers and uploads on YouTube

Srimayee Dam

ABSTRACT

This article entails a quantitative analysis of the association between cyberbullying crude and number of user subscribers and uploads on YouTube. Of the several numbers of variables listed in the said dataset from Kaggle.com, the variables of comments, subscribers and uploads were found to be continuous. The data included 3464 YouTube user ids, and were analyzed using Descriptives, Normality Tests and Spearman's Rho on SPSS. The findings included non-linear, positively skewed, non-normal distribution. The correlation analysis depicted a slight positive association between User subscribers and rude comments, whereas no connection between User uploads and crude comments. Future work to focus on other factors influencing hurtful commenting on YouTube.

Author: Doctoral student in Health Education, Teachers College, Columbia University.

I. BACKGROUND/INTRODUCTION

A 2024 US national survey found that 79% of youth experienced victimization on YouTube, with the common forms of cyberbullying in YouTube content and comments being offensive interactions, online harassment, and cyberstalking (Muminovic, 2025). YouTube is not only one of the platforms where such behavior persists, but through advanced machine learning techniques instances of cyberbullying within YouTube comments can be identified

(Thamaraiselvi et al., 2024). Hence it is important to assess which factors contribute to offensive commenting and cyberbullying on YouTube.

The data for this paper has been used from Kaggle.com, a 2022 dataset on cyberbullying detection consisting of datasets from several sources projecting different types of cyberbullying like hate speech, aggressions, insults and toxicity. The data includes sources and social media platforms such as Twitter and YouTube. Hence, it becomes even more relevant to evaluate the associations between toxic comments on YouTube with other variables/factors.

The sample size/population sample includes a number of 3464 YouTube ids (members of different ages, having a certain number of subscribers to their channels, and those who received derogatory comments on their uploads). The measures used for the project include descriptive statistics and tests of normality for each of the variables, as well as Spearman's correlation between the variables. In order to run Spearman's correlation, the variables include a) to determine an association between the *number of subscribers* and the *number of derogatory comments* on YouTube; b) to assess the relationship between the *number of YouTube uploads* (by user/s) and the *number of toxic comments* on YouTube.

The rationale for analyzing the association between the above variables is because prevalence of toxicity in YouTube comments is still high (despite its anti-bullying regulations and policies) as per the data in recent studies. So, it seems justified to try and understand if there are variables that influence these cyberbullying behaviors and tendencies through derogatory commenting. Similarly the rationale behind using the above dataset is because it was the most readily available, relevant to the topic and easy to download dataset.

As mentioned above, for the purpose of this paper, the correlations of two independent variables (like the number of user subscribers, and number of user uploads) with the number of toxic comments online would be assessed using Spearman's Rho, a non-parametric test in SPSS. Since the dataset is positively skewed with significant outliers, hence the use of non-parametric tests (such as Spearman's Rho) to assess the relationship. This large dataset of 3464 cases is unique in a way because it compiles harmful YouTube comments and provides a numeric value to it for each YouTube user or member.

1.1 Research Questions

Based on the variables in the dataset, a correlation analysis seems most appropriate to inquire about the association between the variables. These associations can be run in SPSS using the *Spearman's correlation* and reason needs to be provided for choosing non-parametric tests.

This could be justified by demonstrating the *descriptive statistics* and *tests of normality* results.

Thus all these above tests would be required to be run on SPSS.

The variables that I chose to focus on are the number of user subscribers and number of user uploads (as the independent variables), and the number of toxic/hurtful comments (as the dependent variable). The two research questions that have been developed are:

1. Is there a relationship between the number of user subscribers and the number of toxic comments on YouTube?
2. Is there an association between the number of user uploads and the number of hurtful comments on YouTube?

To be able to determine an association between the variables as stated, it would provide some insight into which factors may or may not influence offensive commenting and consequently cyber-harassment on YouTube.

Based on the above research questions for the project, the null and alternative hypotheses would be as follows:

Null Hypothesis for Research Question 1: There is no relationship between the number of user subscribers and the number of mean or toxic comments on YouTube

Alternative Hypothesis for Research Question 1: There is a relationship between the number of user subscribers and the number of mean or toxic comments on YouTube

Null Hypothesis for Research Question 2: There is no association between the number of YouTuber uploads and the number of mean and toxic comments online

Alternative Hypothesis for Research Question 2: There is an association between the number of YouTuber uploads and the number of hurtful comments online

Each of the research questions could be adequately answered using a non-parametric test

(such a Spearman's Rho) in SPSS to assess the association between two continuous variables.

II. METHODS

Nature of the Data: The study participants include 3464 YouTube user ids (members of different ages, having a certain number of subscribers, and those who received derogatory comments on their YouTube uploads). However, it's hard to say how the data was collected, as this was found as a public dataset to use from Kaggle.com (Kaggle might have compiled all the raw data from YouTube).

The dataset appears to be a collection of data on YouTube users, analyzing their activity and toxic content of the comments. Each row with the columns represents a specific user profile, a collection of the raw text of comments received by the user, the total number hurtful comments, the user's subscriber count, his/her YouTube membership duration, total number of videos uploaded, profanity in user id, age and oh label of the users.

In short, the dataset links user activity metrics (like subscribers, uploads, comments) with the text content of the mean and hurtful comments.

This is in a way to determine an association between hurtful/toxic/bullying comments with the membership duration, number of uploads, subscriptions, age, profanity in username and oh labels of the users.

Measures: Pearson's correlation coefficient would have been most appropriate to analyze the data as well as address each of the research questions. However, the *descriptive statistics* to analyze the skewness and kurtosis values and *normality tests such as KS*, as well as looking into relevant graphs such as *histograms, qq-plots and box plots*, a positively skewed (with a longer tail extending to the right of the distribution), non-normal distribution would prompt a non-parametric test such as *Spearman's Rho*.

Spearman's Rho would be used for this dataset to measure the strength and direction of the monotonic relationship between the "continuous" variables, as stated in each of the above research questions. Because the data did not meet the assumptions for a Pearson's correlation and has significant outliers (as determined by the *skewness values of descriptive statistics, significance values of KS tests and graphical representations of normality such as histograms, qq-plots and box plots*); hence *Spearman's correlation* would be used to assess the non-linear relationship. Besides, with each of the variables having a ratio level of measurement, a correlation test to determine the association between the variables in each of the research questions would be most relevant.

Through *Spearman's Rho*, the correlation between the variables would be measured, whether one increases or decreases in relation to the other; however the relationship is not necessarily a straight line (*Spearman's rank-order correlation using SPSS statistics*, n.d.).

III. RESULTS

The results from the related tests in SPSS are as follows:

Descriptive Statistics and Tests of Normality for the Number of Comments variable:

Mean	15.45
Median	14.00
Variance	117.994
Standard Deviation	10.863
Range	49
Skewness	.579
Kurtosis	-.548
KS Test Significance	<.001

With the mean of 15.45 slightly higher than the median of 14.00, this indicates that the distribution is not perfectly symmetrical (there are outliers) and that there could be a slight pull in the positive direction (right side). A positive skewness value of .579 suggests that the tail of the distribution is longer on the right side. A KS test significance value of <.001 implies that the distribution was not random and statistically significant, thereby providing strong evidence to reject the null hypothesis. The other normality tests depict a positively skewed histogram, indicating few high-value outliers (Image in Appendix 1c). The QQ Plot also demonstrates non-normal distribution, with points deviating from the straight line (Image in Appendix 1d).

Descriptive Statistics and Tests of Normality for the Number of Subscribers variable:

Mean	304.32
Median	2.00
Variance	240886923.46
Standard Deviation	15520.532
Range	912377
Skewness	58.642
Kurtosis	3446.793
KS Test Significance	<.001

An extreme low value of 2.00 as median does inflate the mean and the variance, and there is the presence of extreme outliers in this dataset. The data is also extremely skewed (58.642), with the long tail extending towards the right. Hence it is a highly, positively skewed dataset. Similarly, a KS test significance value of <.001 implies that there is very strong evidence to reject the null hypothesis. The other normality tests (histogram and QQ plot images in Appendix 2c and 2d) project a highly positively skewed, non-normal distribution.

Descriptive Statistics and Tests of Normality for the Number of Uploads variable:

Mean	10.29
Median	5.00
Variance	820.623
Standard Deviation	28.647
Range	819
Skewness	13.416
Kurtosis	274.457
KS Test Significance	<.001

With a mean value of 10.29 higher than the median of 5.00, this shows that there are extreme outliers that are influencing the overall statistics. The data is also heavily right-skewed as 13.416 is a large positive value. With the KS test significance value of <.001 which is lower than the standard p-value, one would reject the null hypothesis. Other tests of normality (images of histogram and QQ plot in Appendix 3c and 3d), depict a non-normal distribution.

Because the data in all of the above variables are heavily right-skewed and non-normal, a Spearman's correlation would be used to assess the relationship based on the research questions. The findings are as follows:

Spearman's Rho	Findings
Correlation Coefficient: *Number of Comments* and *Number of Subscribers*	.079
Correlation Coefficient: *Number of Comments* and *Number of Uploads*	-.009
Two-tailed ignifiScance: *Number of Comments* and *Number of Subscribers*	<.001
Two-tailed Significance: *Number of Comments* and *Number of Uploads*	.597

With a two-tailed significance of <.001, it is a highly statistically significant correlation; thereby the null hypothesis can be rejected. However, a Correlation Coefficient between the *Number of Comments* and the *Number of Subscribers* being 0.079, it is a very weak- almost negligible positive relationship between the number of rude comments received and the number of user subscribers. This means that as the value of one variable increases, the other slightly tends to go up. The connection is minimal, random, close to zero with non-linear relationship between the two variables. Hence subscriber count is not a good predictor for comment count (image in Appendix 4a).

The Correlation Coefficient of -.009 between the *Number of Comments* and *Number of Uploads* indicate an extremely weak, almost non-existent linear relationship. Hence there is no connection between the number of hurtful comments received and the number of user uploads.

Besides, the two-tailed significance value of 0.597 shows that there is no statistically significant relationship between the two variables. Hence, one would fail to reject the null hypothesis. It is random, and unlikely to predict if the number of user uploads influence the number of toxic comments received (image in Appendix 4b).

Discussion/Conclusion:

Based on the findings from the statistical interpretation, it is evident that there could be a slight/very weak, almost negligible positive correlation between the number of rude comments and the number of user subscribers; thereby *we reject the null hypothesis in Research Question 1*. On the contrary, based on the results from the statistical procedures, there is no correlation between the number of crude comments received and the number of user uploads; thereby *we fail to reject the null hypothesis in Research Question 2*. In order to arrive at the above conclusions, appropriate measures were used such as non-parametric tests (Spearman's Rho) as *the data was heavily right-skewed and non-normal*. The Descriptive Statistics, Normality Tests, Graphs for statistical interpretation (histograms, QQ plots) for each variable helped to assess the skewness and non-normality in the data.

To further interpret the above results, future research could focus on the data and findings between the *number of comments* and *number of user subscribers* correlation. The number of crude comments and number of user uploads' association can be excluded from future studies and analysis as there is no correlation as such. Future work could also look into the possibility of other variables in datasets and their association with cyberbullying comments on YouTube. *Notable Limitations and Recommendations:* While analyzing the data for the *Subscribers* and the *Uploads* variables, there were significant extreme outliers that heavily influenced the output. A future recommendation would be to use more authentic, normally distributed data for statistical analysis.

Similarly, there was hardly any mention about how and/why the data was collected on YouTube user activity metrics. Going forward, another recommendation would be to provide enough background information for context regarding data collection and that might help to present more authentic data (with less or no extreme outliers).

There was also a lack of mention of timeline, as in when the data was collected. As a recommendation, this needs to be looked into and incorporated in future datasets.

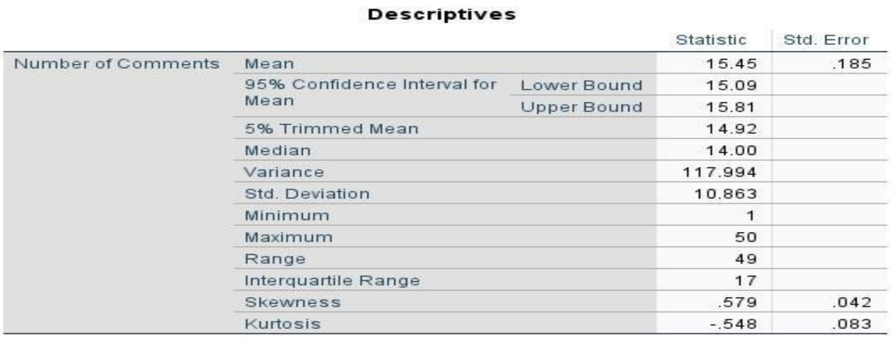
Uniqueness of the Dataset and Findings and Implications: In all, the dataset was unique as it helped address the inquiry based on the research questions. The research questions developed were unique too as such inquiry was missing in the current literature. The findings presented would help advance conversations around the topic based on quantitative statistical analysis. It would encourage future research and develop quantitative data as well as the use of appropriate statistical tools to guide intervention/prevention research on cyberbullying.

REFERENCES

1. *Cyberbullying tweets*. (n.d.). Kaggle.com. https://www.kaggle.com/datasets/pradeepjswl/cyberbullying-tweets?utm_source=chatgpt.com
2. Muminovic, A. (2025). Moderating harm: Benchmarking large language models for cyberbullying detection in YouTube comments. *ArXiv*. https://arxiv.org/html/2505.18927v2
3. *Spearman's rank-order correlation using SPSS statistics*. (n.d.). Laerd Statistics. https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php#:~:text=This%20is%20why%20we%20dedicate,another%20measure%20would%20be%20better
4. Thamaraiselvi, A., Sinduja, S., Devadharshini, S., Gnanadharshini, S., Kaviyasri, V., & Rama Jeevitha, R. (2024). Detection of cyberbullying on YouTube using machine learning.
5. *International Journal of Engineering Research & Technology, 13*(10). IJERTV13IS100110. https://www.ijert.org/research/detection-of-cyber-bullying-on-youtube-using-machine-lear ning-IJERTV13IS100110.pdf

APPENDIX

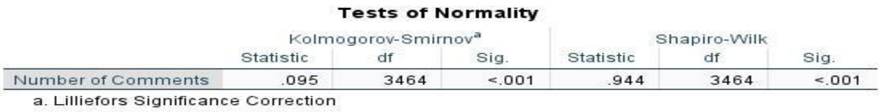
Appendix 1a.



Descriptives

			Statistic	Std. Error
Number of Comments	Mean		15.45	.185
	95% Confidence Interval for Mean	Lower Bound	15.09	
		Upper Bound	15.81	
	5% Trimmed Mean		14.92	
	Median		14.00	
	Variance		117.994	
	Std. Deviation		10.863	
	Minimum		1	
	Maximum		50	
	Range		49	
	Interquartile Range		17	
	Skewness		.579	.042
	Kurtosis		-.548	.083

Appendix 1b

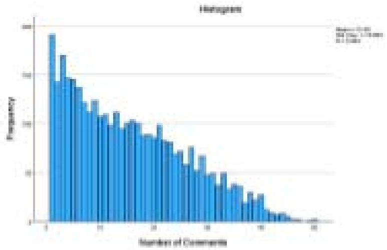


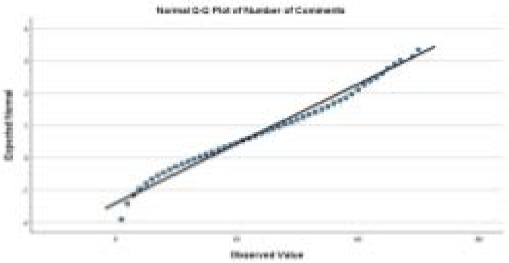
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Comments	.095	3464	<.001	.944	3464	<.001

a. Lilliefors Significance Correction

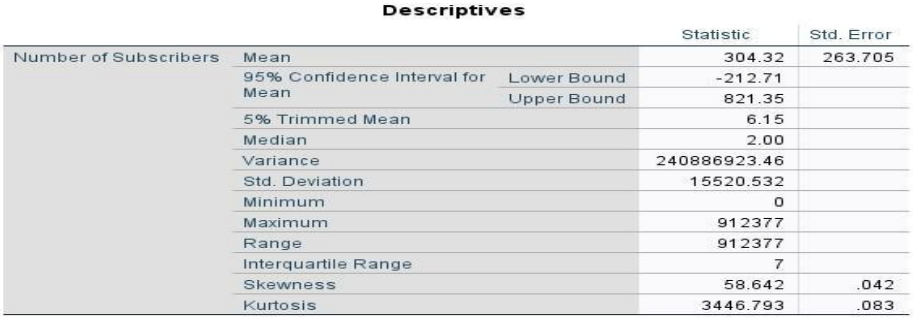
Appendix 1d





Appendix 1c.

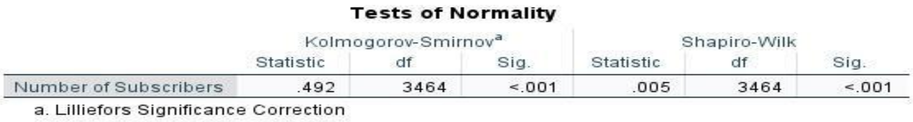
Appendix 2a



Descriptives

			Statistic	Std. Error
Number of Subscribers	Mean		304.32	263.705
	95% Confidence Interval for Mean	Lower Bound	-212.71	
		Upper Bound	821.35	
	5% Trimmed Mean		6.15	
	Median		2.00	
	Variance		240886923.46	
	Std. Deviation		15520.532	
	Minimum		0	
	Maximum		912377	
	Range		912377	
	Interquartile Range		7	
	Skewness		58.642	.042
	Kurtosis		3446.793	.083

Appendix 2b

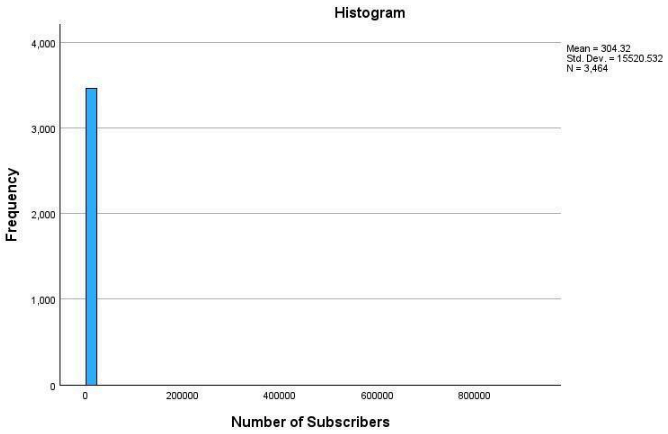


Tests of Normality

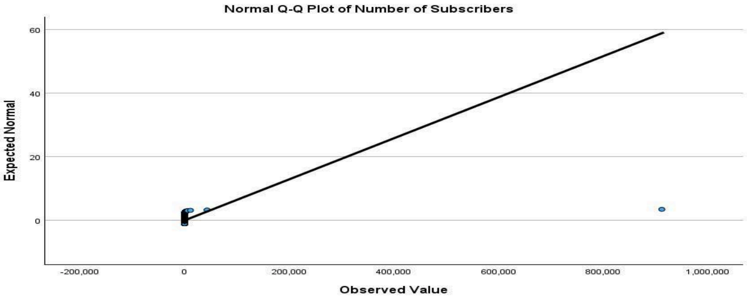
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Subscribers	.492	3464	<.001	.005	3464	<.001

a. Lilliefors Significance Correction

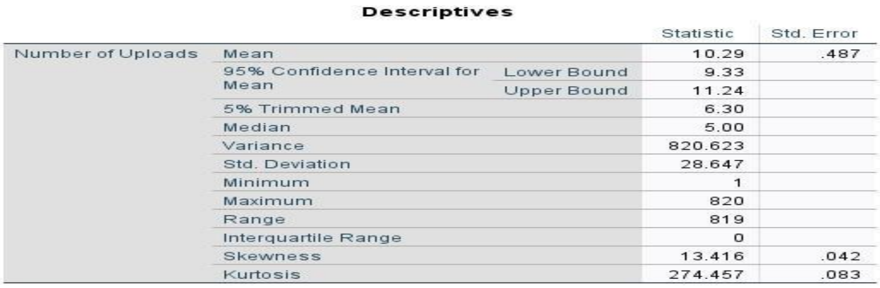
Appendix 2c



Appendix 2d



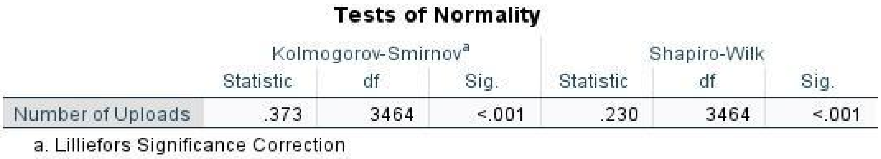
Appendix 3a



Descriptives

			Statistic	Std. Error
Number of Uploads	Mean		10.29	.487
	95% Confidence Interval for Mean	Lower Bound	9.33	
		Upper Bound	11.24	
	5% Trimmed Mean		6.30	
	Median		5.00	
	Variance		820.623	
	Std. Deviation		28.647	
	Minimum		1	
	Maximum		820	
	Range		819	
	Interquartile Range		0	
	Skewness		13.416	.042
	Kurtosis		274.457	.083

Appendix 3b

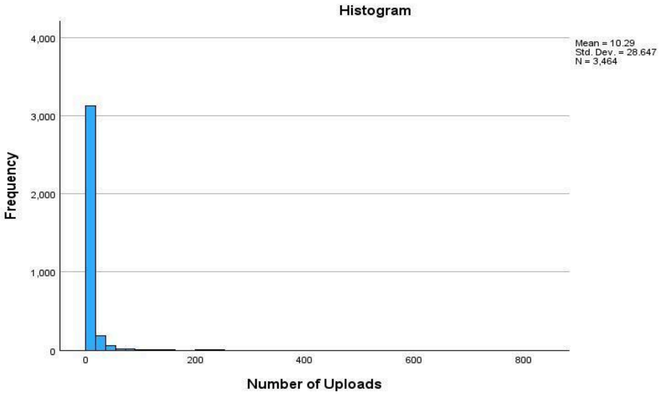


Tests of Normality

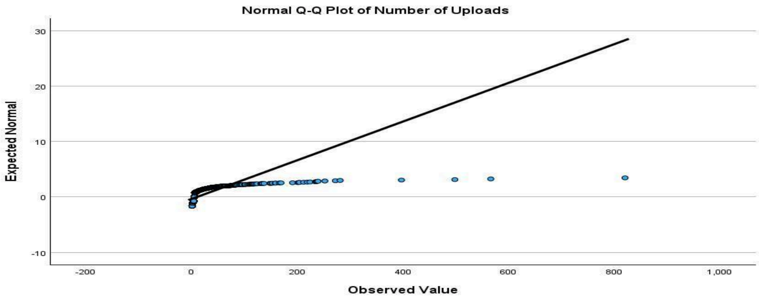
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Uploads	.373	3464	<.001	.230	3464	<.001

a. Lilliefors Significance Correction

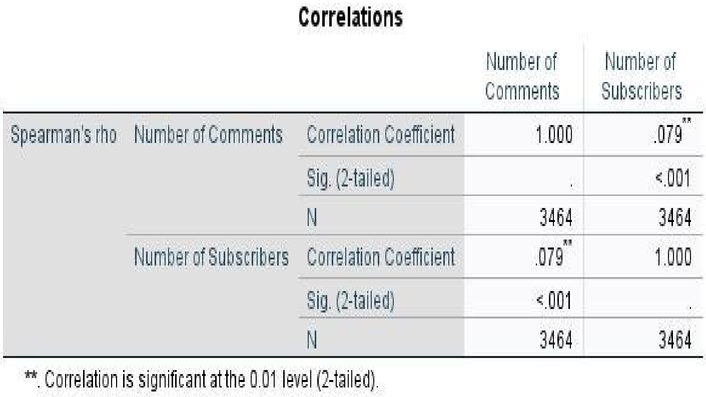
Appendix 3c



Appendix 3d



Appendix 4a

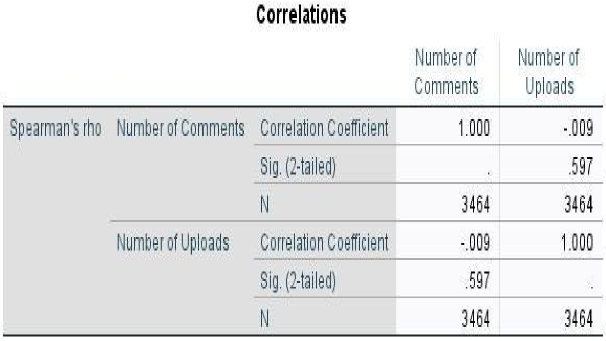


Correlations

			Number of Comments	Number of Subscribers
Spearman's rho	Number of Comments	Correlation Coefficient	1.000	.079**
		Sig. (2-tailed)	.	<.001
		N	3464	3464
	Number of Subscribers	Correlation Coefficient	.079**	1.000
		Sig. (2-tailed)	<.001	.
		N	3464	3464

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix 4b



Correlations

			Number of Comments	Number of Uploads
Spearman's rho	Number of Comments	Correlation Coefficient	1.000	-.009
		Sig. (2-tailed)	.	.597
		N	3464	3464
	Number of Uploads	Correlation Coefficient	-.009	1.000
		Sig. (2-tailed)	.597	.
		N	3464	3464