Exploring Factors Influencing User Engagement with Academic Institution-Related YouTube Videos: A Case Study of "AAA College"

Abstract

In recent years, the use of social media by higher-education institutions has become increasingly prevalent as a means of connecting with prospective students. However, there remains a gap in understanding the factors influencing user engagement with such content. In this study, we investigate three determinants of user engagement – video duration, category, and whether the video was uploaded by the college's official channel – on videos retrieved by the search query "AAA College," a small, private college in the United States. Our findings highlight the significance of video category and official channel designation in predicting user engagement, as measured by the proportion of likes to views. We find that video category and video duration significantly impact user engagement, as measured by the proportion of likes to views. Additionally, we uncover interaction effects between video duration and both category and official channel designation, suggesting nuanced dynamics in user engagement patterns.

Introduction

In recent years, higher education institutions have increasingly leveraged social media technology to forge connections with prospective students, reflecting the shifting communication landscape and the preferences of young adults (Faculak, 2012). YouTube has emerged as a platform offering a diverse array of content, ranging from personal accounts to official promotional material, that provides insights into student life and campus culture. Despite the prominence of YouTube in the higher education sphere, there remains a dearth of research exploring the factors influencing user engagement with academic institution-related videos. Previous research, such as Oliphant's (2023) analysis of mental health-related videos, revealed higher engagement levels for personal narratives compared to commercial, governmental, or organization-produced content. Tutan (2022) also underscores the significance of video length and comment counts in discerning between official and unofficial sources, a finding further supported by Yang (2022), who identifies shorter videos as garnering more engagement, with likes, comments, and views serving as indicators of social endorsement.

In this study, we aim to investigate certain determinants of user engagement in YouTube videos retrieved through the search query "AAA college," representing a small, private college in the United States. To our knowledge, a similar dataset – namely, the top YouTube videos related to an academic institution – has not been studied for factors related to user engagement. We examine the characteristics associated with each video – duration, category, and whether the video was uploaded by the college's official channel – and seek to understand their significance in predicting a video's engagement as measured by the proportion of likes to views.

Data

We used the Python requests library and YouTube API to create a novel dataset of 300 videos returned by the search query "AAA College." The videos were listed in order of relevance as determined by the YouTube API. For each video, we collected details including channel, duration, view count, like count, and video category. We use like proportion (like count divided by view count) as the response variable capturing engagement to create an interpretable measure that accounts for the large variance in the number of views (SD = 76789.13). To ensure parameter estimability, we also chose to collapse video categories from 14 into 4 levels: Education, Entertainment, People and Blogs, and Others, a level that encompasses the 10 categories with the least videos each (Appendix, Figure 1). We also created a binary variable based on channel name – "AAA college" or not "AAA college" – that indicates whether the video was uploaded by the college's official channel.

Methods

We then calculated summary statistics for all variables, and fit an additive multiple linear regression model as we have a continuous response variable (Model 1). We checked the linearity assumption by accessing the residual plots for all predictors(Appendix, Figure 2-10). Using the additive model, we removed potentially influential points by setting a threshold of 0.015 for Cook's Distance (Appendix, Figure 5). Given the large variance in video duration (SD = 789.265), we chose to omit high leverage points for video duration. Our interaction plots suggest an interaction between video category and duration, as well as between official channel designation and duration (Appendix, Figures 2, 3). Therefore we fit two interaction models (Model 2, Model 3), and compare them to the additive model to determine whether these interactions are significant. We list hypothesized models in Appendix, Table 2. We interpreted the variables for which we found significant results, and justified our choice of a multiple linear regression model by assessing the validity conditions using the best model. We fix α = 0.05 for this investigation and use R for all analyses and visualizations (version 2023.12.1+402).

Results

There were n= 300 videos collected. After screening the variable using Cook's Distance and leverage value we used 281 videos for analysis. This set has an average like proportion of 1.659 (SD = 1.436). The average duration of each video was 571.6 seconds (SD = 789.265 sec). The majority of videos were categorized as either Education (25.62%) or Other (37.37%), and 11.03% of videos were uploaded by the college's official channel. A full summary of statistics can be found in Appendix, Table 1.

Model 1: Additive Model with Duration, Category and OfficialChannel

The additive model with category, duration, and official account explained approximately 9.3% of the variation in like proportion (Adj. $R^2 = 0.093$). We prefer this model to a model with no predictors (F-statistic = 6.723; $F_{5, 275}$; p-value < 0.0001).

$$\begin{aligned}\widehat{LikeProportion}_i &= 1.298 + 0.0002\, Duration_i - 0.378\, Cat(Education)_i \\ &\quad + 0.678\, Cat(Entertainment)_i + 0.893\, Cat(People\ and\ Blogs)_i + 0.340\, Official\ Channel_i\end{aligned}$$

From this model, we expect a video categorized as "Other" and from a channel other than the college's official channel with a duration of 0 seconds to be 1.298%, a value that should be considered for model-fitting only. For each 1-second increase in duration, we expect the like proportion to increase by 0.0002%, given all other variables remain constant (p-value = 0.046). Compared to videos categorized as "Other", we expect like proportion to increase by 0.678% (p-value = 0.009) for an educational video, and by 0.893% for a video categorized as "People and Blogs" (p-value < 0.0001). Official channel designation was not significant at the α = 0.05 level.

Model 2: Interaction Model with Duration x Category

$$\begin{aligned}\widehat{LikeProportion}_i &= 1.267 + 0.0003\, Duration_i - 0.197\, Cat(Education)_i + 0.161\, Cat(Entertainment)_i \\ &\quad + 0.688\, Cat(People\ and\ Blogs)_i + 0.373\, Official\ Channel_i - 0.0003\, (Duration_i \times Cat(Education)_i) \\ &\quad + 0.0017\, (Duration_i \times Cat(Entertainment)_i) + 0.0004 (Duration_i \times Cat(People\ and\ Blogs)_i)\end{aligned}$$

There was a significant interaction between video duration and category, adjusting for official channel (F-statistic = 2.667; $F_{3, 272}$; p-value = 0.048). However, individual interaction terms did not significantly differ from "Other" category. This model was an improvement over the model with no predictors (F-statistic = 5.28; $F_{8, 272}$; p-value < 0.001) and additive model (Adj. $R^2 = 0.109$). Video category was the sole significant variable at α = 0.05 (F-statistic = 56.78; $F_{3, 272}$; p-value < 0.0001).

Model 3: Interaction Model with Duration x OfficialChannel

$$\begin{aligned}\widehat{LikeProportion}_i &= 1.215 + 0.0003\, Duration_i - 0.376\, Cat(Education)_i + 0.721\, Cat(Entertainment)_i \\ &\quad + 0.906\, Cat(People\ and\ Blogs)_i + 0.750\, Official\ Channel_i - 0.0006\, (Duration_i \times Official\ Channel_i)\end{aligned}$$

There was also evidence of a statistically significant interaction between video duration and official channel designation after adjusting for category (t-statistic = -2.30; t_{274}; p-value = 0.022). This model was an improvement over the model with no predictors (F-statistic = 5.28; $F_{6, 274}$; p-value < 0.022), and approximately 10.7 % of the variation in like proportion could be explained by the model (Adj. $R^2 = 0.107$). Given the highest Adj. R^2, and that all variables are significant at the α = 0.05 level, we determine Model 3 to be our best model.

As before, the intercept value of 1.215 is only useful for model fitting. For each 1-second increase in duration for an official channel categorized as "Other", we expect the like proportion to decrease by 0.0006%, as compared to a video from a non-official channel (p-value = 0.046). While video category as a variable was significant (p-value < 0.0001), in this model the p-value for the variable cat(Education) indicates the predicted like proportion for educational videos is not significantly different from that of an “Other” video of the same duration (p-value = 0.133). By comparison, the predicted like proportion for “Entertainment” and “People and Blogs” videos were significantly greater than that of an “Other” video of the same duration (p-value = 0.005 and p-value < 0.0001 respectively). On average, for videos with the same duration “Entertainment” and “People and Blogs” videos had 0.721% and 0.906% more engagement than an “Other” video, respectively.

Model Adequacy

The plots of the residuals regressed on duration, category, and official channel, only the category shows relatively small deviations from the zero-line (Appendix, Figures 6-14). This indicates that the linearity assumption for duration and official channel variables may not be appropriate. Both duration and the official account variable display a discernible pattern in the residual plot suggesting that the equal variance assumption might not hold (Appendix, Figures 6-14). Our Levene test confirms this conclusion for video category (F-statistic = 8.99; $F_{3, 277}$; p-value < 0.001) and the official channel variable (F-statistic = 16.771; $F_{1, 279}$; p-value < 0.001). The normal Q-Q plots show residuals that mostly follow a 45◦ line with some slight deviations in the tails, suggesting that the normality assumption is mostly appropriate (Appendix, Figures 15-17). We also used Variance Inflation Factors (VIF) to check for multicollinearity between our variables, and observed the overall VIF values remain relatively low (all below 1.3), suggesting minimal multicollinearity concerns.

Discussion

We found several insights into factors influencing user engagement with academic institution-related videos on YouTube, including interaction effects between duration and both category and official channel designation. While the interaction between duration and category did not significantly alter the predicted engagement levels, for videos compared to those categorized as "Other," videos uploaded by the college’s official channel demonstrated higher levels of engagement as video duration increased. This suggests that longer videos from the official channel were more likely to receive likes relative to views compared to videos from unofficial sources. Furthermore, videos uploaded by the college’s official channel demonstrated higher levels of engagement as video duration increased. Institutions may benefit from investing in longer-form content, particularly if it aligns with educational or informative themes, to enhance user engagement and promote positive perceptions among prospective students. From our best model, the interaction model with official channel designation (Model 3), our results demonstrate that the video category plays a significant role as videos categorized as "Education" or "People and Blogs" exhibited higher levels of engagement compared to videos categorized as "Other." This finding aligns with previous research highlighting the appeal of educational content and personal narratives on social media platforms.

These results underscore the importance of considering both content characteristics and channel sources in understanding user engagement with academic institution-related videos on YouTube. However, it's essential to acknowledge several limitations of our study. Firstly, our analysis focused solely on videos related to a single institution, limiting the generalizability of our findings. Furthermore, the violations of our model assumptions indicate that a multiple linear regression model may not be appropriate. The limited utility of our best model for explaining the change in like-proportion, as determined by the magnitude of individual variable coefficients and Adj. R^2, suggests that we can explore that other metrics such as comments, shares, and subscriber growth. These variables could provide further insights into user interactions and perceptions of academic institution-related videos on YouTube.

References

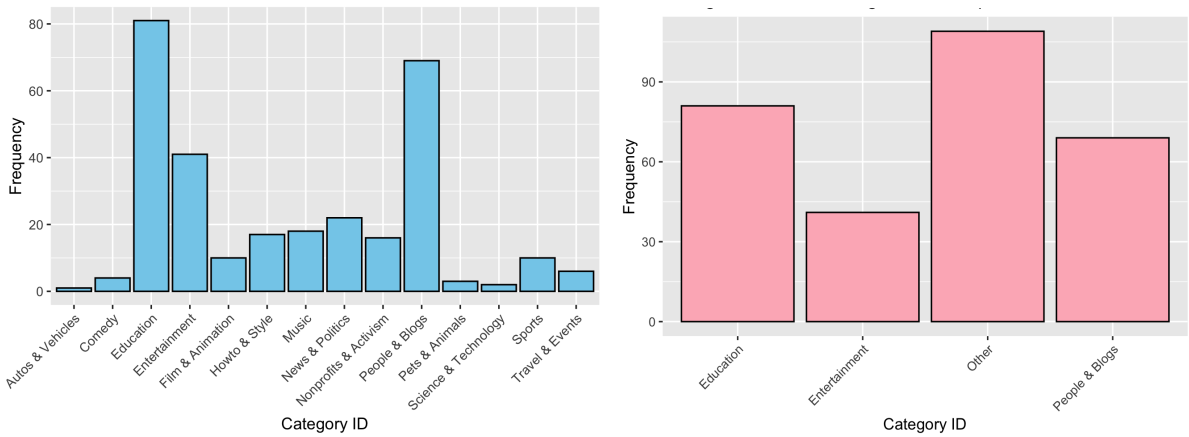
Faculak, Natalie, "College Admissions Use of Social Media in Recruitment Marketing: A Literature Review & Strategic Plan for Western Michigan University's Office of Admissions" (2012). Honors Theses. 1773. https://scholarworks.wmich.edu/honors_theses/1773

Oliphant, T. (2013). User Engagement with Mental Health Videos on YouTube. Journal of the Canadian Health Libraries Association Journal De l'Association Des bibliothèques De La Santé Du Canada, 34(3), 153–158. https://doi.org/10.5596/c13-057

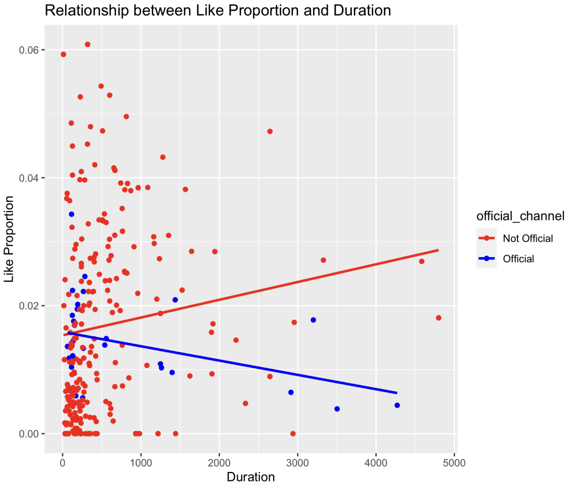
Tutan D, Kaya M. Evaluation of YouTube Videos as a Source of Information on Hepatosteatosis. Cureus. 2023 Oct 11;15(10):e46843. doi: 10.7759/cureus.46843. PMID: 37829652; PMCID: PMC10566639.

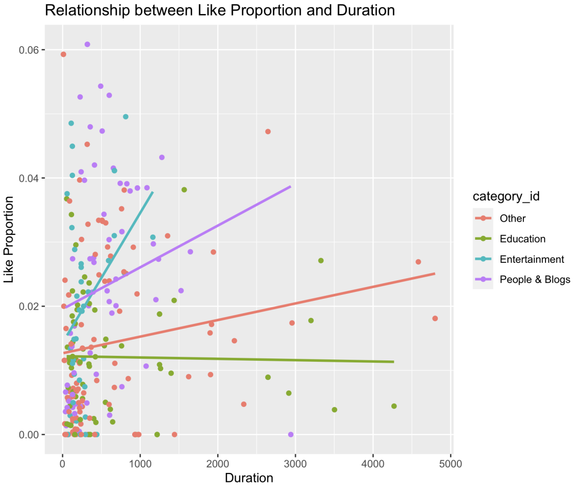
Yang, Shiyu et al. "The science of YouTube: What factors influence user engagement with online science videos?." PloS one vol. 17,5 e0267697. 25 May. 2022, doi:10.1371/journal.pone.0267697

Appendix



Figures 1 and 2: Barplot of video categories before collapsing levels (left) and after (right). The category Other contains all categories with less than 40 videos each.





Figures 3 and 4 : Interaction plots for Duration with Official Account (left) and Duration with Category

Variable Type	Variable Name	Mean (SD)
Response	**Like proportion (%)**	1.7 (1.44)
Explanatory	**Duration (seconds)**	571.6 (789.27)
		n (%)
	Video Category	
	Education	72 (25.62%)
	Entertainment	39 (13.88%)
	People and Blogs	65 (23.13%)
	Other	105 (37.37%)
		n (%)
	Official Channel	
	Official College Channel	31 (11.03%)
	Not College Official Channel	250 (88.97%)

Table 1: Summary of Video Characteristics (n=281)

Model 1: Additive Model

$$LikeProportion_i = \beta_0 + \beta_1 Duration_i + \beta_2 Cat(Education)_i + \beta_3 Cat(Entertainment)_i + \beta_4 Cat(People\ and\ Blogs)_i + \beta_5 Official\ Channel_i + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0,\ \sigma)$$

Model 2: Interaction Model for Video Category and Duration

$$LikeProportion_i = \gamma_0 + \gamma_1 Duration_i + \gamma_2 Cat(Education)_i + \gamma_3 Cat(Entertainment)_i + \gamma_4 Cat(People\ and\ Blogs)_i + \gamma_5 Official\ Channel_i + \gamma_6 (Duration_i \times Cat(Education)_i) + \gamma_7 (Duration_i \times Cat(Entertainment)_i) + \gamma_8 (Duration_i \times Cat(People\ and\ Blogs)_i) + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0,\ \varrho)$$

Model 3: Interaction Model for Official Channel and Duration

$$LikeProportion_i = \delta_0 + \delta_1 Duration_i + \delta_2 Cat(Education)_i + \delta_3 Cat(Entertainment)_i + \delta_4 Cat(People\ and\ Blogs)_i + \delta_5 Official\ Channel_i + \delta_6 (Duration_i \times Official\ Channel_i) + \varepsilon_i$$

$$\varepsilon_i \sim Normal(0,\ \phi)$$

$Cat(Other)_i$ and $NotOfficial\ Channel_i$ are considered the baselines.

Table 2: Hypothesized Models, including the additive model and two interaction models

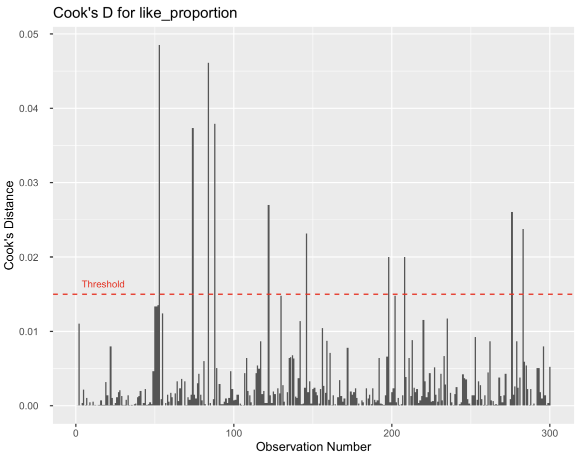
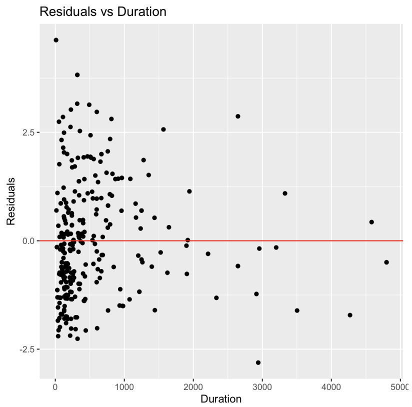
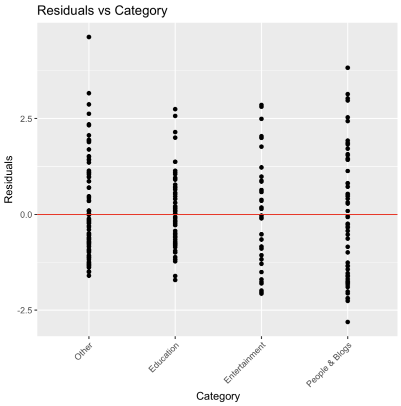


Figure 5: Plot of Cook's Distance using additive model to determine potentially influential points. We set a threshold of 0.015 for omitting points.





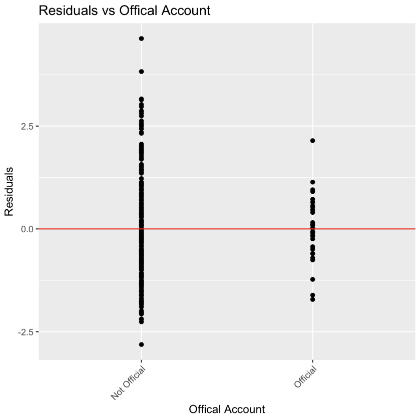
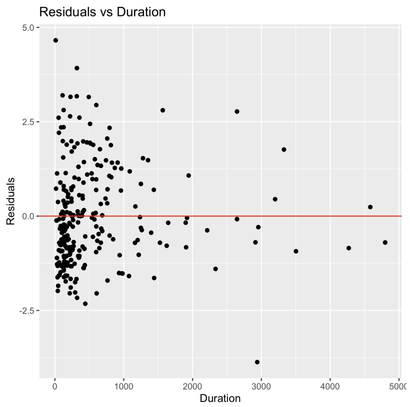
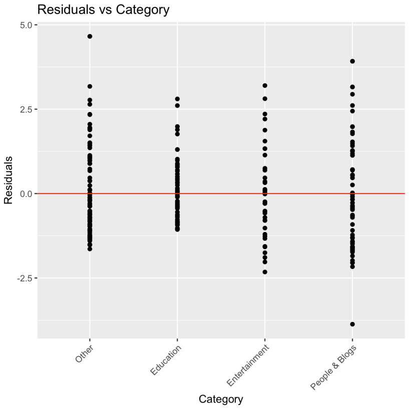


Figure 6 (left): Model 1 residual plot for Duration
Figure 7 (middle): Model 1residual plot for Category
Figure 8 (right): Model 1 residual plot for Offical Account





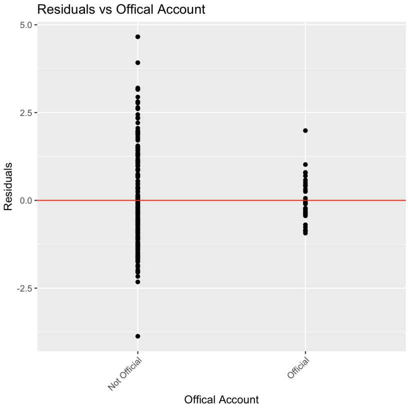
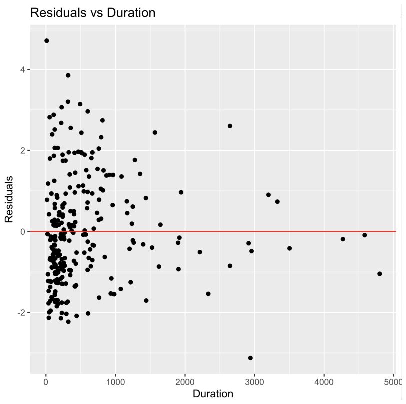
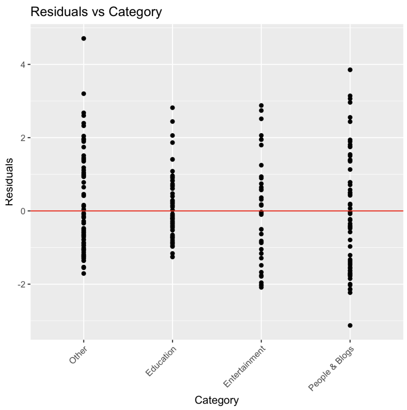


Figure 9 (left): Model 2 residual plot for Duration
Figure 10 (middle): Model 2 residual plot for Category
Figure 11 (right): Model 2 residual plot for Offical Account





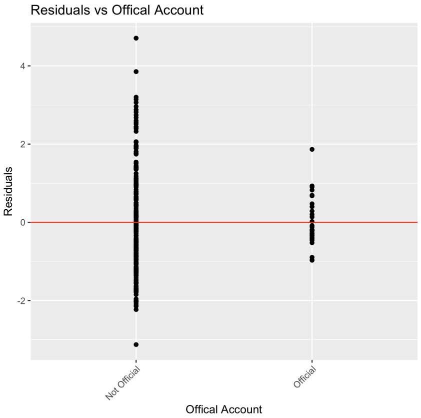
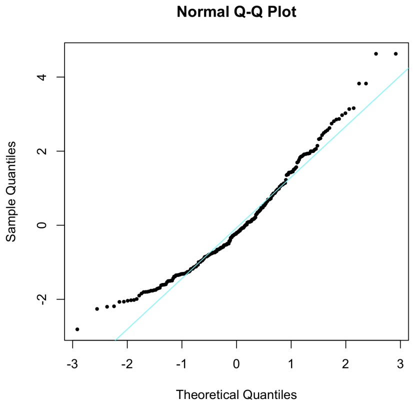
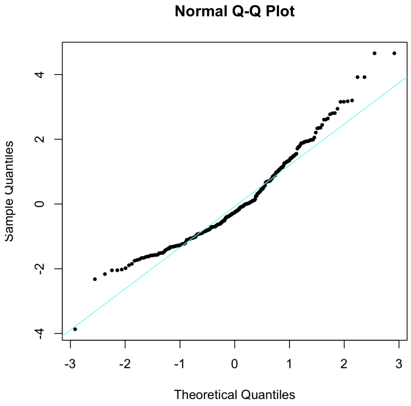


Figure 12 (left): Model 3 residual plot for Duration
Figure 13 (middle): Model 3 residual plot for Category
Figure 14 (right): Model 3 residual plot for Offical Account





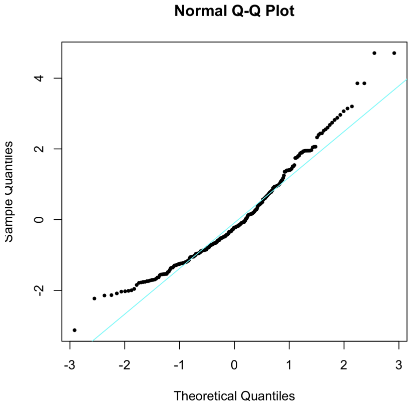


Figure 15 (left): Normal Quantile-Quantile Plot for Model 1
Figure 16 (middle): Normal Quantile-Quantile Plot for Model 2
Figure 17 (right): Normal Quantile-Quantile Plot for Model 3