How Post Length Affects Engagement on Facebook: An Explanatory Model

Arisa Nguyen, Ayman Bari, Emanuel Mejía, Jorge Bonilla

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Research Question . . . 1

2 Data and Methodology **1**
2.1 About the Data . . . 1
2.2 Data exploration . . . 2
2.2.1 Skew of the data . . . 3
2.2.2 Outliers in the Data . . . 4
2.3 Checking our intuitions . . . 5

3 Research Design **6**

4 Modeling **8**
4.1 Base Model . . . 8
4.2 Second Model . . . 9
4.3 Third Model . . . 11
4.4 Fourth model . . . 12

5 Results **14**

6 Model Limitations **17**
6.1 Omitted Variables . . . 17
6.2 Feedback Loops . . . 18
6.3 Statistical Limitations . . . 18
6.3.1 Independent and Identically Distributed Data . . . 18
6.3.2 No Collinearity/ Unique Best Linear Predictor . . . 19
6.3.3 Linear Conditional Expectation . . . 19
6.3.4 Homoskedastic Errors . . . 19
6.3.5 Normally Distributed Errors . . . 20

7 Conclusion and Impacts **20**
7.1 Overall Effect . . . 20

1 Introduction

1.1 Motivation

Professional sports teams often create Facebook pages as a way to build community, advertise, and keep their fans in the loop. Facebook is the largest social network in the world with an annual revenue of approximately $117 billion, with 97.5% generated from the advertisement sales[1]. Posts made by these organizations are the primary way of communicating to their audience. Our aim is to investigate how specific attributes of the post, Facebook page, and organization affect engagement on each post. The data set to be analyzed comes from UC Irvine's Machine Learning Repository, and has scraped information from 2,770 pages, 57,000 posts, and 4,120,532 comments.

1.2 Research Question

For this project we'll refer to **engagement** as the number of comments made on a post in the 24 hour period after it was published on Facebook. Our research question is as follows:

how is user engagement on the Facebook platform affected by post length while controlling for other variables such as post date, post share counts, page category, and page popularity?

The approach to operationlize this study is discussed in detail in the **Research Design** section of this report. Our results and processes are explained explicitly throughout our paper, and we provide a recommendation to changes to the Facebook platform that can lead to greater user engagement. Keeping in mind that user engagement is complex in nature in the world's largest social media platform, we also discuss the limitations of our approach and the steps we took to mitigate these effects.

2 Data and Methodology

2.1 About the Data

The data set to be analyzed is the "Facebook Comment Volume Dataset Data Set". The initial data set includes 55 features and 602813 records. We narrow down the feature set to the following variables of interest[2]:

Table 1. Data Description

Feature	Description[3]
CC4 (Outcome)	Number of comments left on a post 24 hours after it has been published
Post Length (Intervention)	Number of characters that a post has
Post Share Count	Number of times the post has been shared
Page Likes	Number of likes on a post
Page Talking About	Number of people who return to the page after liking (includes comments, likes, shares, etc. by visitors of the page)
Page Category	Description of the page category
Base time	Value between 0 and 72 indicating the number of hours between a post being published, and the when the data was collected
Published Day	Seven binary features indicating the day of the week the post was published
Base Day	Seven binary features indicating the day of the week the post data was scraped

[1] Meta's (formerly Facebook Inc.) annual revenue from 2009 to 2021: https://www.tinyurl.com/3sruyj2f
[2] Full features list avaialable at https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset#
[3] Feature definitions taken from data source. Available at https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset#

There are five variants of the data set, each scraped at a different time. For convenience in analysis and data exploration, we merge the five data sets and add a new variable to indicate the source variant for each record. We wrangle the data to reduce the 14 binary features indicate the days of the week the posts were published and scraped into two features, "Day Published" and "Base Day". We then created two binary features indicating where the new features were weekends (Saturday or Sunday) or not. The added variables include:

Table 2. Additional Features

Additional Features	Description
Day Published	Day of the week the post was published
Post Length (Intervention)	Day of the week the post data was scraped
Published Weekend	Binary value indicating if post was published on the weekend
Base Weekend	Binary value indicating if post data was scraped on a weekend

We are looking to create a model that explains levels of engagement 24 hours after a post is published. Therefore, we filter our dataset to remove all posts that are less than 24 hours old. Keeping records for posts published less than 24 before data was collected would skew the CC4 variable (number of comments in the last 24 hours since post was published), as we would be potentially comparing the number of comments for some posts that had 24 hours to accrue comments, with others that had only 2 for example.

We then subset our data to include only those where the category variable includes the term "Professional Sports Team". This reduces the total volume of our data set to 71875 records. We further subset this, taking a 30% sub sample to perform our explanatory data analysis (EDA).

2.2 Data exploration

We start our data exploration by looking at the distribution of our variables of interest as shown in Image 1.

Image 1. Histograms of Variables

We have two findings in our initial exploration:

2.2.1 Skew of the data

We notice that the distributions are highly skewed to the left for Comments (CC4, our response variable), Post Length, Post Shares, Page likes, and Page talking about. This is especially the case for the Comments (CC4), Post Length, and Post Shares features where it appears that there is a large cluster around the 0 value. We count the number of zeros in our sub exploration data sample for each feature to confirm this.

Table 3. Proportion of Zeroes

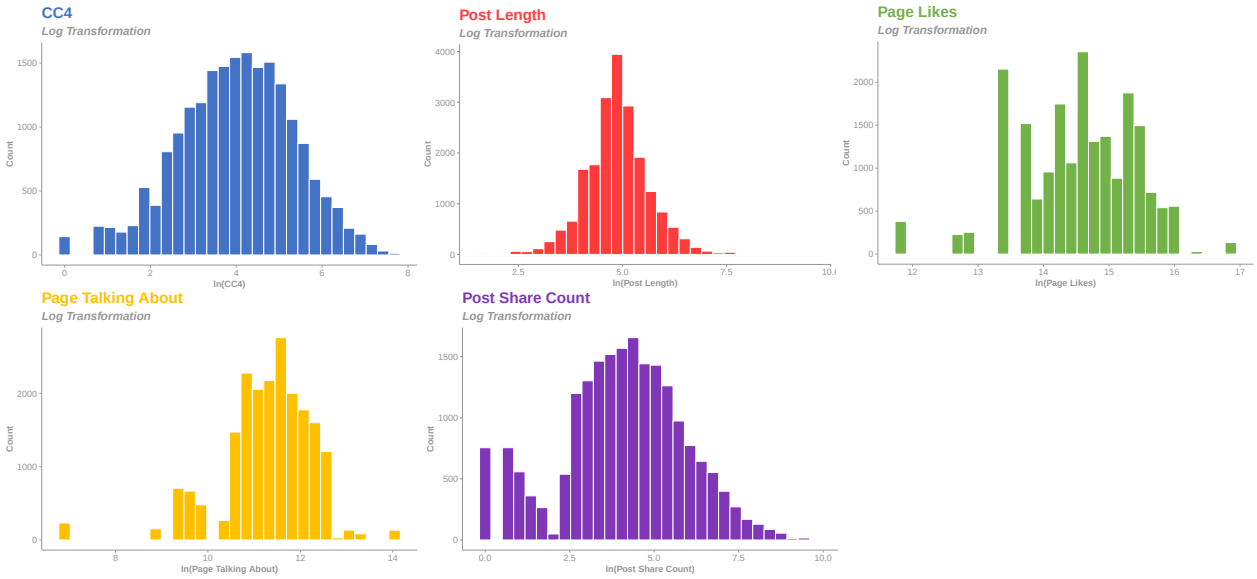
Feature	Proportion of zero values in sample
Comments (CC4)	1%
Post Length	6%
Page Likes	0%
Page Talking About	0%
Post Share Count	0%

We find that 1% of the Facebook posts in our sample have no comments. The skew in our feature distributions suggests that variable transformations may be useful in modeling and interpretation of our data. A natural log transformation seems appropriate given the shape of our distributions. We choose to remove the zero values from our sample, as we would not be able to apply the same transformation to them.

This further reduces our sample size down to 67327 comments.

We apply natural log transformations to our comments, page likes, page likes, page talking about, post length, and post shares features. The resulting distributions can be found below in Image 2.

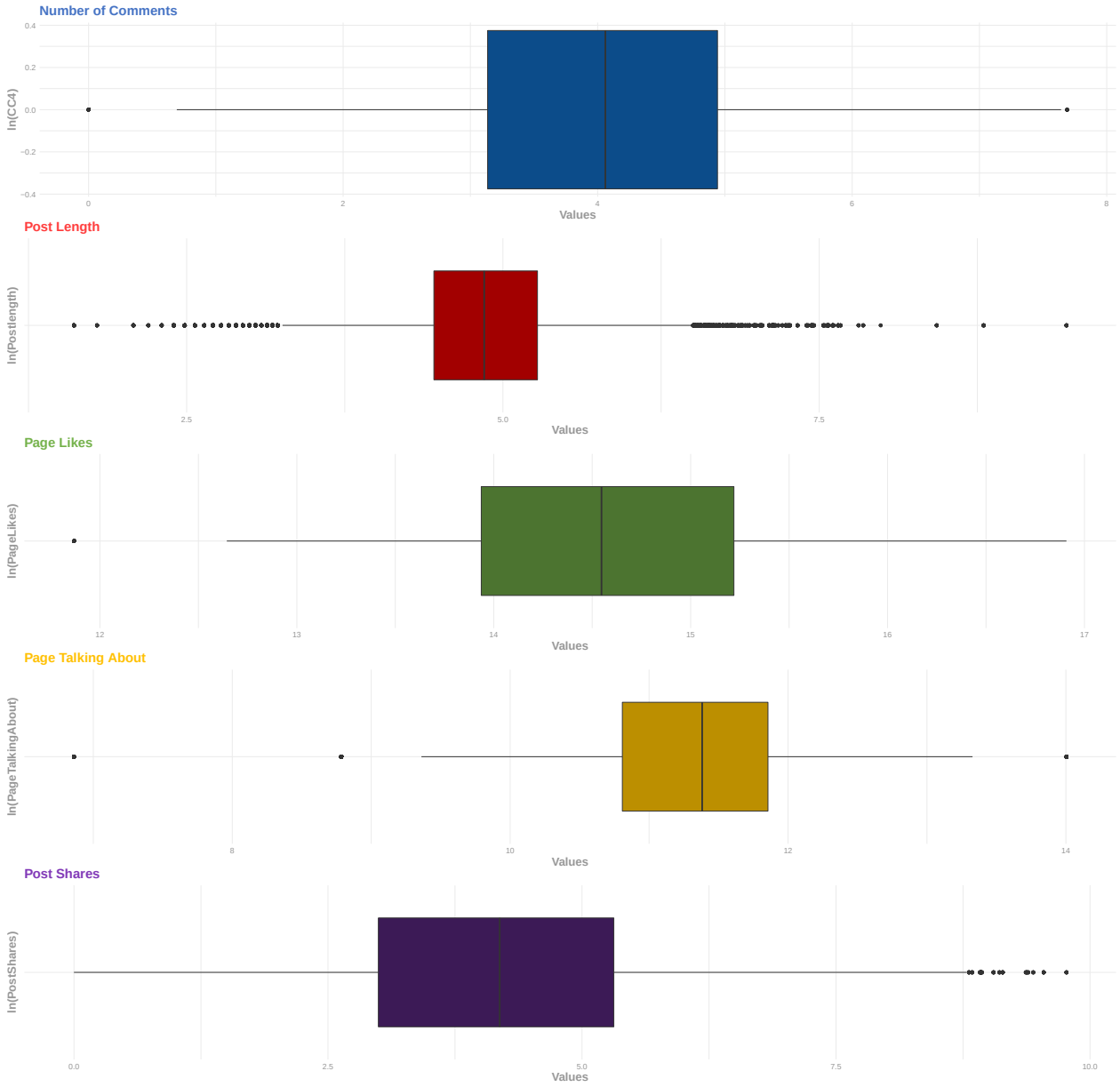
Image 2. Histograms of Transformed Variables



2.2.2 Outliers in the Data

The histograms above for the untransformed variables confirms that our data set contains several outliers that heavily skew our data. The features with outliers include CC4, Page Likes, Page Talking About, Post Length, and Post Shares which can make interpreting the coefficient of the variables more difficult once we have the results from the linear regressions. The boxplots below provide a representation on the quantile distributions of these outliers.

Image 3. Boxplots of Variables



The presence of these outliers in our outcome and explanatory variables suggests that we will perform additional statistical procedures such as Cook's Distance and Statistical Leverage during the modeling phase in order to take into account for the effect of outliers in our data. Performing these tests will also allow us to troubleshoot some of the potential issues that we will face in estimating the results from the Variance-Covariance matrix between the explanatory variables in each model in this study.

2.3 Checking our intuitions

Ahead of preparing our model, we have some assumptions regarding the relationships between our features and our outcome variable (post engagement, as measured by number of comments). Namely, we assume page likes, page talking about, and post shares to be positively correlated with the engagement. Here, we produce joint distribution plots between our outcome variable, and features of interest in Image 4.

Image 4. Joint Distributions of Variables

Adding a regression line through our joint distributions provides some support for our intuitions. We see a positive gradient for page likes, page talking about, and post shares feature.

3 Research Design

This will be a quantitative, retrospective analysis of engagement of Facebook posts. *We operationalize the concept of engagement as the number of comments contributed to a Facebook post. The aim of our design will be to produce an explanatory model for engagement.* Our target will be to produce a result that can be shared with the product team at Facebook in order to make an update to, or around the functionality of the posts feature that will enhance engagement levels on the platform. For example, the finding that shorter posts drive more engagement may provide the rational for introducing character restrictions. Alternatively, different recommendations may be pushed to owners of different page categories nudging them on the direction of a more optimal post length. In our case, we focus on posts published on professional sports team pages.

We first conceptualize a causal theory for the engagement level of a post during the last 24 hours relative to when the post was initially published. We consider the number of comments that a post receives as our metric for product success. Our causal theory states that the length of a Facebook post affects the number of comments within the next 24 hours while controlling for other variables such as post share count, page popularity and the base time at which the post was published relative to when the data was collected. Considering the variables that we believe to impact engagement as represented by the number of comments that a post receives, we form the following structural model:

To evaluate our hypothesis we established the following regression model:

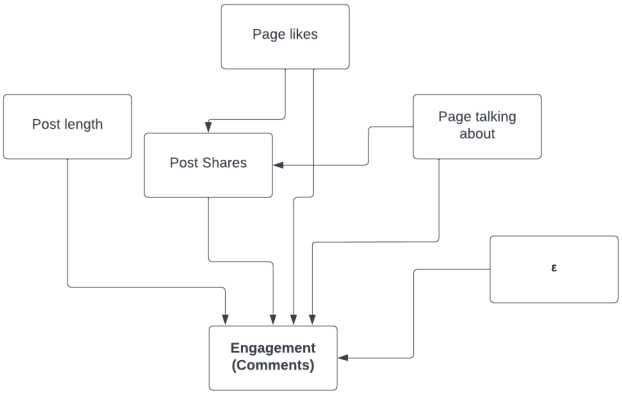


Figure 1: Structural Model for Facebook Post Engagement.

$$\begin{aligned}\text{Comments} = \beta_0 &+ \beta_1 \text{PostLength} + \beta_2 \text{PostShares} + \beta_3 \text{PostPopularity} + \beta_4 \text{PageTalkingAbout} \\ &+ \beta_5 \text{DayPublished} + \beta_6 \text{PostAge} + \varepsilon\end{aligned}$$

We start with an exploration of the data, checking the joint distributions of each causal variable with our outcome variable. Following an initial assessment to validate our assumptions, we perform transformations for our variables that may better model the impact on engagement. The source of the data set stems from a program written in Java and Facebook Query Language to scrape the pages, posts, and comments[4]. We use the Ordinary Least Squares Regression to produce our model of the data, and validate the assumptions required. We plan to include in our model a number of control features, which we do not believe to have a direct causal effect on our outcome variable, but may be associated with levels of engagement. During the modeling stage of the project, ANOVA F-Tests will be conducted to assess the efficacy of adding additional variables to the model to determine how successful the Facebook platform is in terms of number of comments which is the engagement metric we wish to analyze. We will do this process to develop at least three models:

1. The base model where we regress the number of comments on the post length alone.
2. A model where we regress the number of comments on all of the desired covariates (control as well as causal variables).
3. A third model where we add interaction terms for our covariates.

[4]Singh, Kamaljot. "Comment Volume Prediction using Neural Networks and Decision Trees." International Conference on Modelling and Simulation, 2015, https://ijssst.info/Vol-16/No-5/paper16.pdf

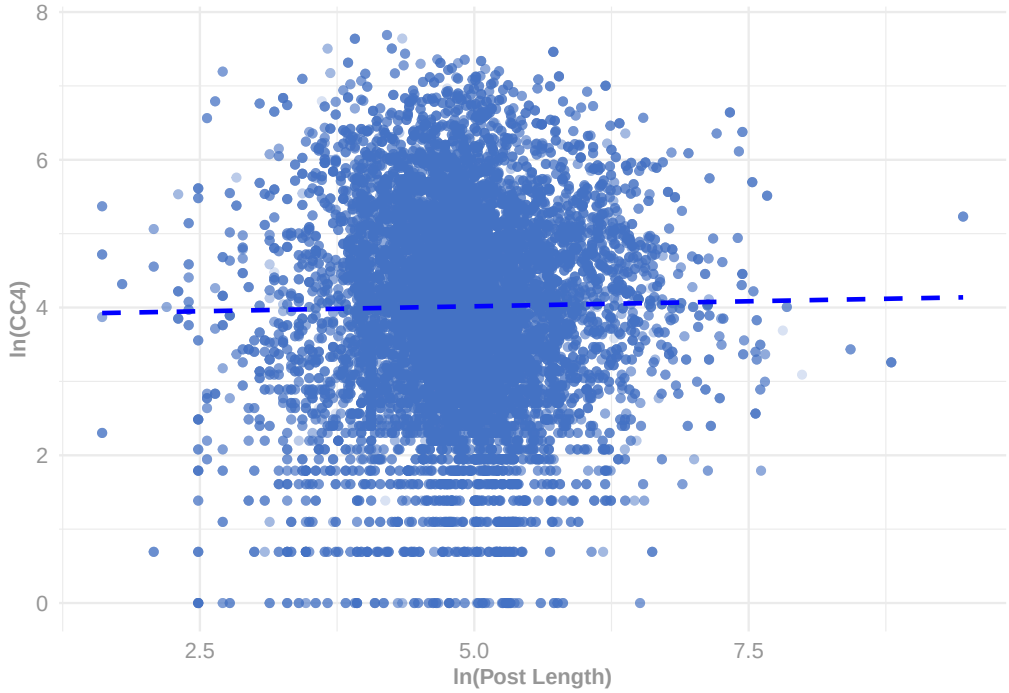
4 Modeling

4.1 Base Model

Our first model will regress our outcome variable on our treatment variable alone.

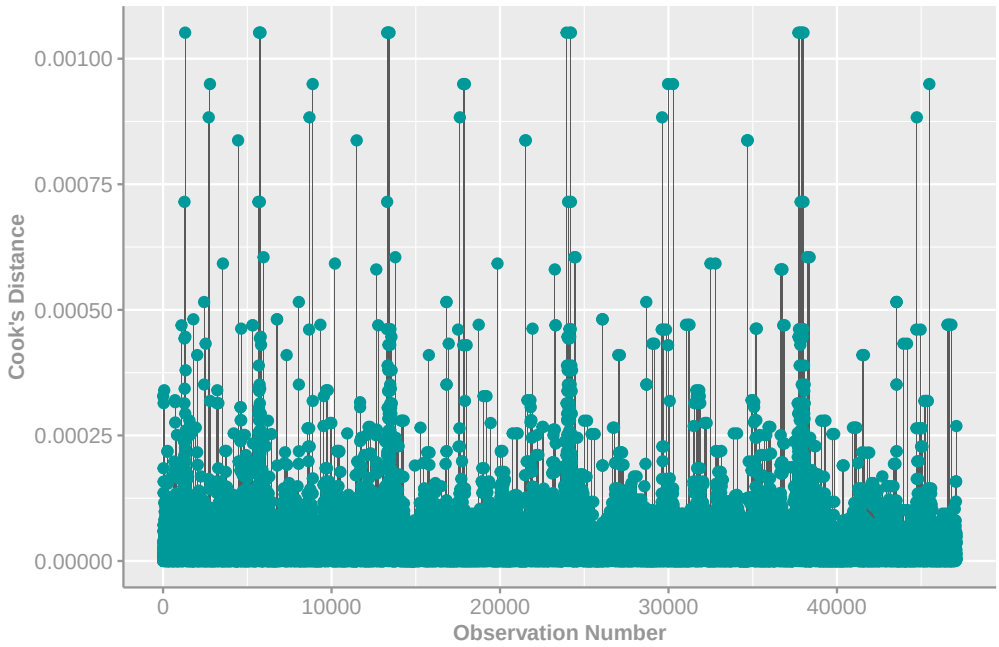
$$\text{Model 1} : ln(\text{Comments}) = \beta_0 + \beta_1 ln(\text{PostLength}) + \varepsilon$$

Image 5. Joint Distributions of Variables



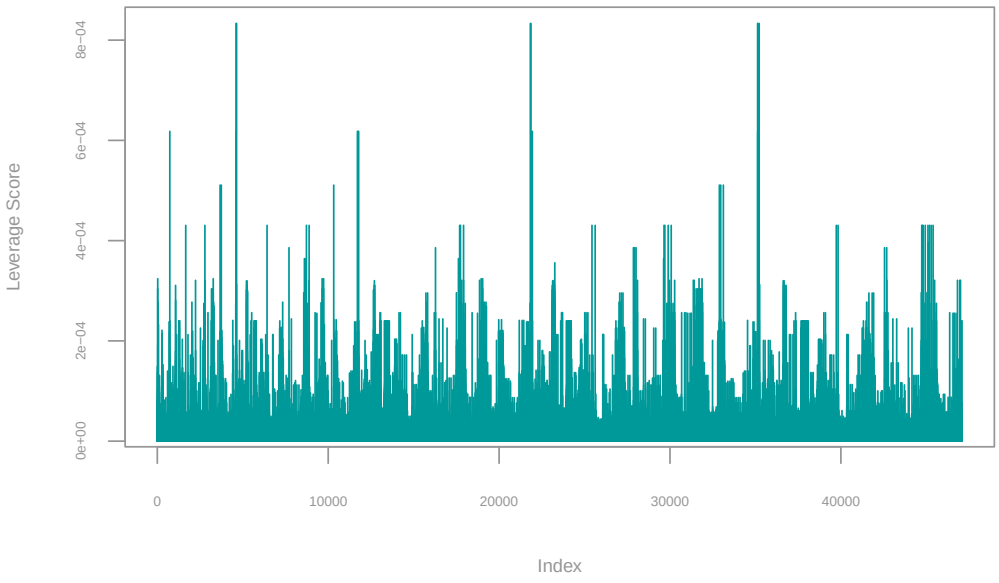
The plots below show the results with Cook's distance to remove the outliers on the base model. The results from the Cook's distance plot indicate the values above a Cook's Distance of 0.00050 are highly influential outliers. Using these influential values, we will remove the outliers from the data set.

Image 6: Cook's Distance for Base Model



The plots for the Statistical Leverage on the base model indicate that none of the values are greater than 2. Since the threshold for Statistical Leverage has been set at 2, and none of the observations are greater than 2, then none of the observations will be removed using the Statistical Leverage approach.

Image 7: Statistical Leverage for Base Model



4.2 Second Model

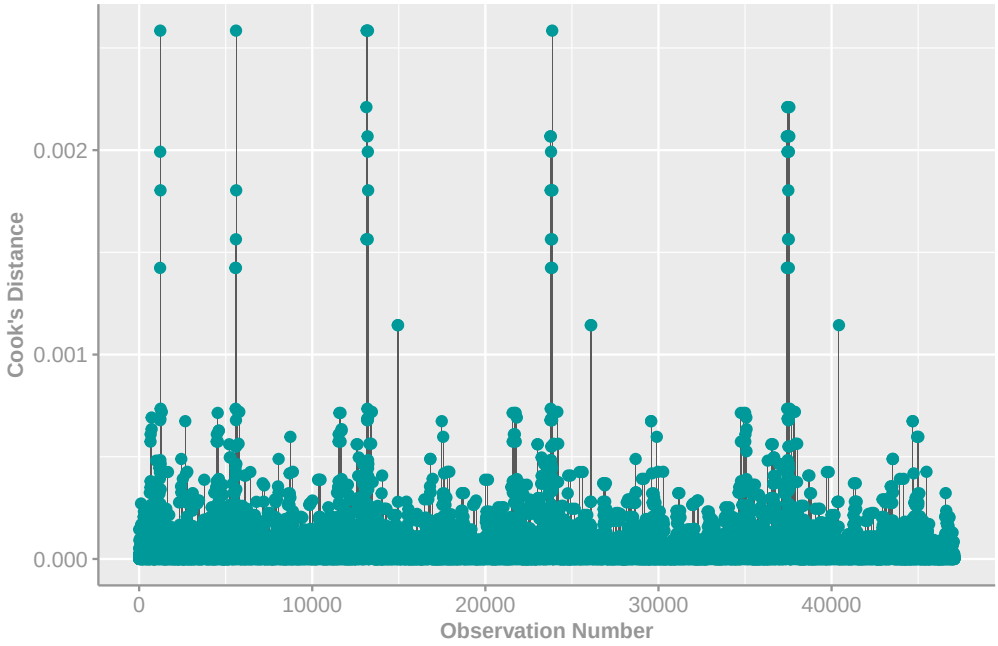
In the second model, we include a number of control variables that we believe to have a causal link with the number of comments generated (our response): 1. Page Likes: We expect that the more likes a page has, the more comments posts on that page will receive 2. Page Talking About: We expect a positive relationship between the number of users that interact with a page, and the engagement on the posts on that page 3. Post Shares: We expect that the more times a post has been shared, the more comments it will receive

The second model is evaluated as follows:

$$\text{Model 2} : ln(\text{Comments}) = \beta_0 + \beta_1 ln(\text{PostLength}) + \beta_2 ln(\text{PageTalkingAbout}) + \beta_3 ln(\text{PageLikes}) + \beta_4 ln(\text{PostShares}) + \varepsilon$$

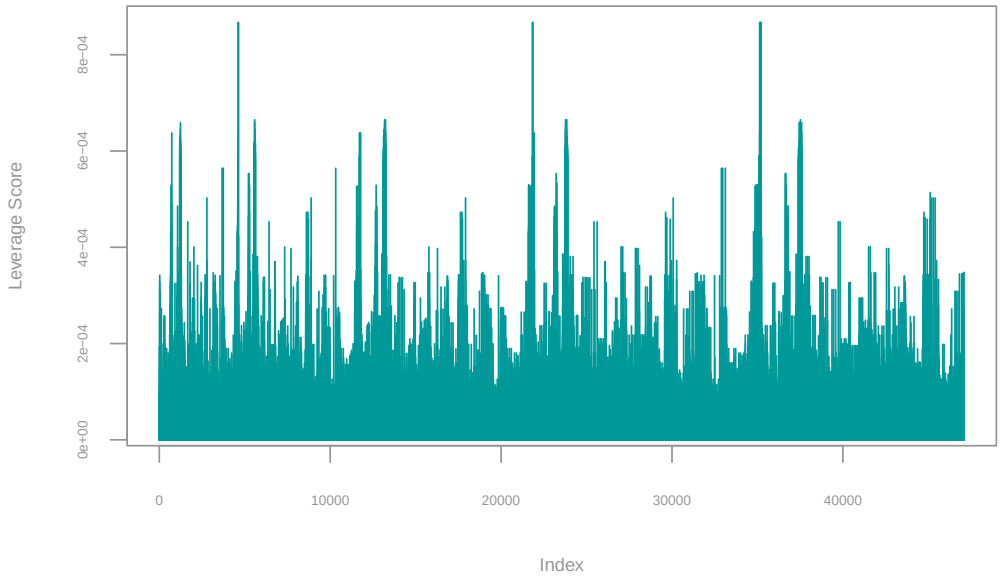
For Model 2, the plots below show the results with Cook's distance to remove the outliers. This plot indicates that values above a Cook's Distance of 0.0020 are highly influential outliers. Using these influential values, we will remove the outliers from the data set based on the results from the second model.

Image 8: Cook's Distance for Model 2



The plots for the Statistical Leverage on Model 2 indicate that none of the values are greater than 2. Since the threshold for Statistical Leverage has been set at 2, and none of the observations are greater than 2, then none of the observations will be removed using the Statistical Leverage approach.

Image 9: Statistical Leverage for Model 2



4.3 Third Model

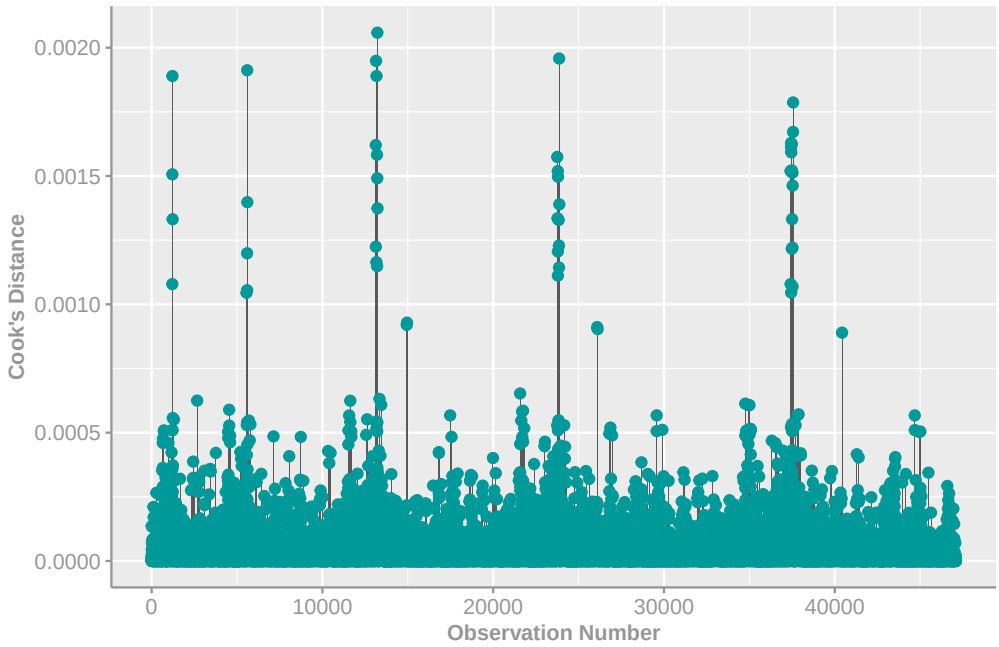
For our third model, we include an additional variables of when the post was published. We do not draw a direct causal link between the day a post was published, but expect that engagement levels on Facebook may vary by day of the week. The additional variables are: 1. Post Age (Basetime): We expect the age of a post to impact the number of comments it generates 2. Weekend: Whether the post was published on a weekend or not

Our third model is evaluated as follows:

$$\begin{aligned}\text{Model 3} : ln(\text{Comments}) = \beta_0 + \beta_1 ln(\text{PostLength}) + \beta_2 ln(\text{PageTalkingAbout}) + \beta_3 ln(\text{PageLikes}) \\ + \beta_4 ln(\text{PostShares}) + \beta_5 \text{Weekend} + \beta_6 \text{PostAge} + \varepsilon\end{aligned}$$

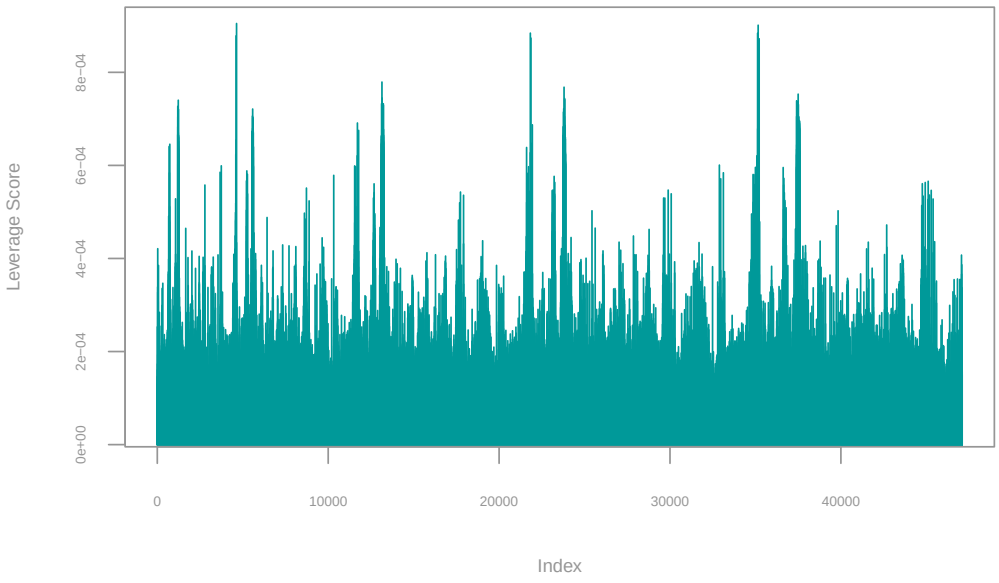
For Model 3, the plots below show the results with Cook's distance to remove the outliers. The results from this Cook's distance plot indicate that values above a Cook's Distance of 0.00025 are highly influential outliers. Using these influential values, we will remove the outliers from the data set based on the results from the third model.

Image 10: Cook's Distance for Model 3



The plots for the Statistical Leverage on Model 3 indicate that none of the values are greater than 2. Since the threshold for Statistical Leverage has been set at 2, and none of the observations are greater than 2, then none of the observations will be removed using the Statistical Leverage approach.

Image 11: Statistical Leverage for Model 3



4.4 Fourth model

Our final model includes interaction terms for features that we believe to be related. We include interaction terms for:

1. Post share count and the post age (Basetime variable): We expect that the duration of a post's presence on a page will impact the number of comments it generates.

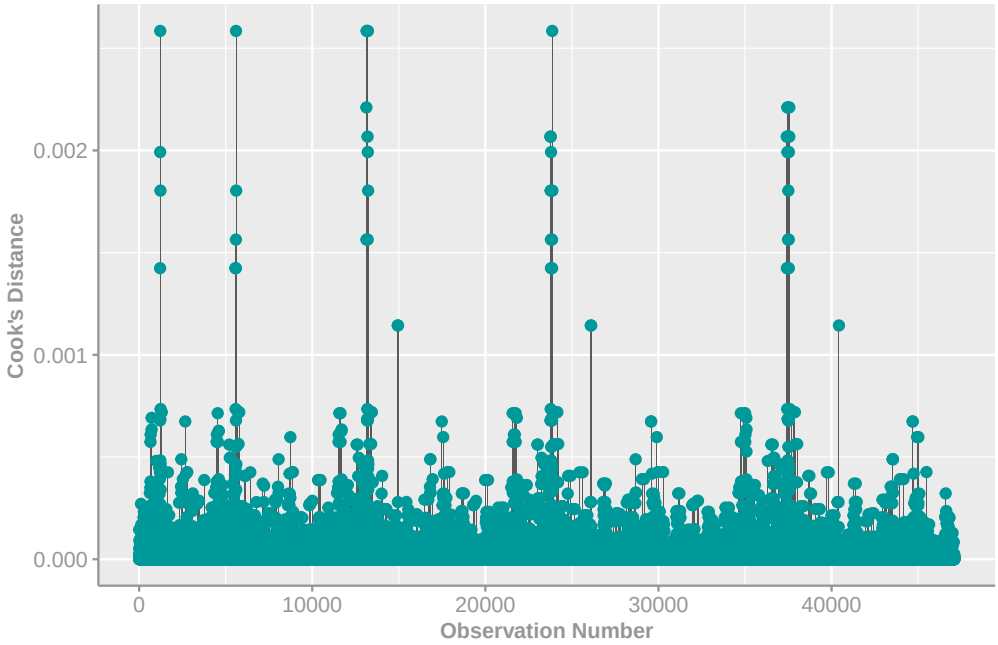
2. We expect that there is some interaction between the number of likes a page has (a proxy for page popularity) and the number of people that interact with that page (page talking about). It would be reasonable to assume that the more popular a page is, the more frequently it is visited.
3. Post length and Weekend, as we may assume that posting behavior varies on the weekend as compared with weekdays.

So our fourth and final model is:

$$\begin{aligned}\text{Model 4}: ln(\text{Comments}) = \beta_0 &+ \beta_1 ln(\text{PostLength}) + \beta_2 ln(\text{PageTalkingAbout}) + \beta_3 ln(\text{PageLikes}) \\ &+ \beta_4 ln(\text{PostShares}) + \beta_5 \text{Weekend} + \beta_6 \text{PostAge} \\ &+ \beta_7 ln(\text{PostLength}) * \text{Weekend} + \beta_8 ln(\text{PageTalkingAbout}) * \text{PageLikes} \\ &+ \beta_9 ln(\text{PostShares}) * \text{PostAge} + \varepsilon\end{aligned}$$

Lastly, the plot below show the results with Cook's distance to remove the outliers based on the regression of model 4. The results from the Cook's distance plot indicate the values above a Cook's Distance of 0.0015 are highly influential outliers. Using these influential values, we will remove the outliers from the data set based on the results from the fourth model.

Image 12: Cook's Distance for Model 4

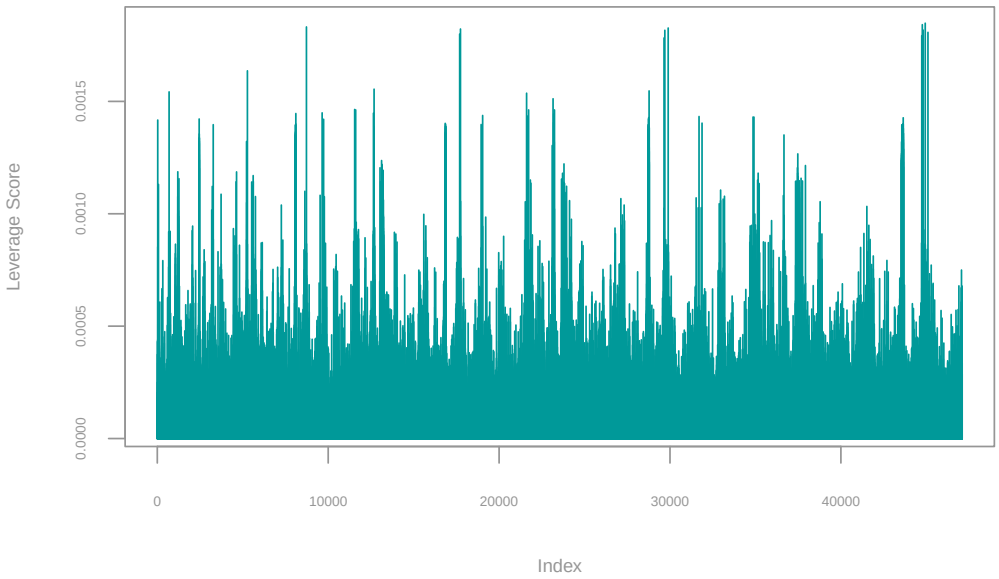


The plots for the Statistical Leverage on Model 4 indicate that none of the values are greater than 2. Since the threshold for Statistical Leverage has been set at 2, and none of the observations are greater than 2, then none of the observations will be removed using the Statistical Leverage approach.

Table 4: Variance-Covariance Matrix for Model 4

	(Intercept)	ln(Post Length)	ln(Page Talking About)	ln(Page Likes)	ln(Post Share Count)	Weekend	Post Age (hours)	Post Length * Weekend	Page Talking about * Page Likes	Post Age * ln(Post Share Count)
(Intercept)	0.453	0	-0.039	-0.032	0	0.000	0	0.000	0.003	0
ln(Post Length)	0.000	0	0.000	0.000	0	0.000	0	0.000	0.000	0
ln(Page Talking About)	-0.039	0	0.004	0.003	0	0.000	0	0.000	0.000	0
ln(Page Likes)	-0.032	0	0.003	0.002	0	0.000	0	0.000	0.000	0
ln(Post Share Count)	0.000	0	0.000	0.000	0	0.000	0	0.000	0.000	0
Weekend	0.000	0	0.000	0.000	0	0.004	0	-0.001	0.000	0
Post Age (hours)	0.000	0	0.000	0.000	0	0.000	0	0.000	0.000	0
Post Length * Weekend	0.000	0	0.000	0.000	0	-0.001	0	0.000	0.000	0
Page Talking about * Page Likes	0.003	0	0.000	0.000	0	0.000	0	0.000	0.000	0
Post Age * ln(Post Share Count)	0.000	0	0.000	0.000	0	0.000	0	0.000	0.000	0

Image 13: Statistical Leverage for Model 4



Taking the variance covariance matrix of our final model in Table 4, we note that the variance of the coefficients is close zero (also see standard errors in results section). This makes us skeptical of the results of the model. We experiment with a smaller sample size to check the effects of sample size on the variances, however we find similar results even when reducing the sample size to 30% of the total. We also attempt to remove zero values and outliers from our models, but the results in the variances remain near zero. We note that this result requires further investigation.

5 Results

The results of our models can be found in Table 5 on the next page of this report.

The base model confirms a positive relationship between the length of a post and the number of comments received 24 hours after a post was published. Holding everything else constant, this model states that a 1% increase in the post length leads to an average increase of 0.027% in the number of comments. The R^2 is quite low at 0.0002 suggesting that 0.0002 percent of the variation in the number of commnets is attributed to the single explanatory variable. Although the explanatory variable is statistically significant, this model alone does not produce practically significant results due to the low fit of the model and the omitted variable bias.

Next, we add three control variables as explained in section 4.2 of the report and we see improvements in our results. The addition of these three explanatory variables reflects user behavior deemed necessary to assess user engagement. This is reflected with an R^2 of 0.538 and statistically significant explanatory varibles. Model 2 infers that, while holding everything else constant, a 1% increase in the post length leads to an average increase of 0.029% in the number of comments. Moreover, there is a negative coefficient on the number of likes which intuitively does not make sense as one would expect that higher likes on Facebook posts would lead to higher comments. For this model, the results may not be practically significant due to the contradictory results reported on the coefficient for the number of likes. When performing the ANOVA

Table 5: Regression Results

	Dependent variable:			
	Comments Count (log_CC4)			
	(1)	(2)	(3)	(4)
(Intercept)	3.882^{***}	-0.233^{***}	-0.236^{***}	-1.482^{**}
	(0.041)	(0.075)	(0.076)	(0.673)
ln(Post Length)	0.027^{***}	0.029^{***}	0.030^{***}	0.036^{***}
	(0.008)	(0.006)	(0.006)	(0.007)
ln(Page Talking About)		0.256^{***}	0.255^{***}	0.370^{***}
		(0.006)	(0.006)	(0.059)
ln(Page Likes)		-0.050^{***}	-0.049^{***}	0.044
		(0.007)	(0.007)	(0.048)
ln(Post Share Count)		0.476^{***}	0.477^{***}	0.454^{***}
		(0.002)	(0.002)	(0.008)
Weekend			0.076^{***}	0.201^{***}
			(0.009)	(0.065)
Post Age (hours)			-0.001^{**}	-0.003^{***}
			(0.0003)	(0.001)
Interaction: Post Length | Weekend				-0.026^{*}
				(0.013)
Interaction: Page Talking about | Page Likes				-0.008^{*}
				(0.004)
Interaction: Post Age | ln(Post Share Count)				0.0005^{***}
				(0.0002)
Observations	47,095	47,095	47,095	47,095
R^2	0.0002	0.538	0.539	0.539
Adjusted R^2	0.0002	0.538	0.539	0.539
Residual Std. Error	1.334 (df = 47093)	0.907 (df = 47090)	0.906 (df = 47088)	0.906 (df = 47085)
F Statistic	10.671^{***} (df = 1; 47093)	$13{,}717.350^{***}$ (df = 4; 47090)	$9{,}170.037^{***}$ (df = 6; 47088)	$6{,}117.015^{***}$ (df = 9; 47085)

Note: *p<0.1; **p<0.05; ***p<0.01

F-Test between the base model and model 2, we obtain the results below. Given that the p-value is below 0.05, we reject the null hypothesis that there is no difference between both models and we choose model 2 over the base model.

```
anova(model_2, base_model, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: log_CC4 ~ log_Postlength + log_PageTalkingAbout + log_PageLikes +
##     log_PostShareCount
## Model 2: log_CC4 ~ log_Postlength
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1  47090 38730
## 2  47093 83840 -3    -45110 18282 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In model three, we control for the weekend binary variable and age of the post. As expected, the model presents a positive relationship between the Weekend control binary variable and the number of comments. If the post was published on a weekend, this would lead to an increase of 7.6% in the number of comments while holding everything else constant. This is intuitive as one would expect that people on weekends would have more time to comment on a Facebook post as opposed to a weekday when they may have other obligations such as school or work. The covariates in model 3 are statistically significant, and we find the results could be practically significant for Facebook as they serve billions of users daily on their platform. This company should incentivize individuals to create posts on weekends to increase user engagement. Additionally, the R^2 marginally increased to 0.539 and the explanatory variable of interest, ln(Post Length), remains with a positive coefficient. However, the negative coefficient on Page Likes remains negative which remains counterintuitive. The results for the ANOVA F-Test between model 2 and model 3 are shown below. Given that the p-value is below 0.05, we reject the null hypothesis that there is no difference between both models and we choose model 3 over model 2.

```
anova(model_3, model_2, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: log_CC4 ~ log_Postlength + log_PageTalkingAbout + log_PageLikes +
##     log_PostShareCount + PublishedWeekend + Basetime
## Model 2: log_CC4 ~ log_Postlength + log_PageTalkingAbout + log_PageLikes +
##     log_PostShareCount
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  47088 38672
## 2  47090 38730 -2   -58.085 35.363 4.504e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lastly, model 4 contains the interactions discussed in section 4.4 along with all previous covariates from model 3. The regression indicates that there is still a positive relationship between the number of comments and the length of a post. This tells us that we are capturing the necessary omitted variables to further explain the variation in the number of comments. The fourth model states that a 1% increase in the post length leads to an average increase of 0.036% in the number of comments received in a Facebook post. It is also worth noting that the coefficient on Page Likes is now positive, although this variable is no longer statistically significant. Moreover, the R^2 remained at 0.539 and the explanatory variable of interest, ln(Post Length), remains with a positive coefficient. The results for the ANOVA F-Test between model 3 and model 4 are shown below. Given that the p-value is below 0.05, we reject the null hypothesis that there is no difference between both models and we choose model 4 over model 3.

```
anova(model_4, model_3, test = 'F')
```

```
## Analysis of Variance Table
##
## Model 1: log_CC4 ~ log_Postlength + log_PageTalkingAbout + log_PageLikes +
##     log_PostShareCount + PublishedWeekend + (log_Postlength *
##     PublishedWeekend) + (log_PageLikes * log_PageTalkingAbout) +
##     (log_PostShareCount * Basetime)
## Model 2: log_CC4 ~ log_Postlength + log_PageTalkingAbout + log_PageLikes +
##     log_PostShareCount + PublishedWeekend + Basetime
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  47085 38658
## 2  47088 38672 -3   -13.791 5.5991 0.0007789 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Despite the statistical significance of features added in models 3 and 4, we notice great consistency in the coefficients of the features in our initial causal mode (model 2). We suspect that the addition of new features and interactions in models 3 and 4 does not add any further explanatory power. Therefore, we assessed the improvements to the model by conducting F-test between each version as previously shown. Table 6 below shows the Mean Squared Residuals (MSR) for each of the model indicating that the MSR decreases with each model iteration.

Table 6. Mean Squared Residuals

Model	MSR
Base Model	1.7802287
Model 2	0.8223857
Model 3	0.8211523
Model 4	0.8208595

6 Model Limitations

Our model faces the common limitation of one equation structural models. The three key limitations we introduce here are omitted variable bias, a likely causal feedback loop skewing our results, and statistical limitations.

6.1 Omitted Variables

When it comes to the number of comments that are left on a post, there are a number of other variables that may be influencing our outcomes that are not captured in this data set. In our analysis, we consider some page features (page likes and page talking about) and post features (post length and post shares). A post feature that was not included in the data set but could influence outcomes is the number of likes the post received. If many people are willing to interact with the post through liking it, it follows that many people may also be willing to interact with the post through commenting. Additionally, users may be more willing to comment on popular posts validated with many likes. This omitted variable bias is away from 0.

Another post feature that was not included in the model is the promoted status. When a user pays to promote a post, the post is boosted so that it is seen by more people thereby possibly receiving more comments. The omitted variable bias in this case would be away from 0. This variable was included in the original data set in binary form (0 for not promoted, 1 for promoted). However, the EDA showed that all posts in the data set were not promoted (0), and it was therefore was excluded from the analysis since it would not have added any new insight into the model.

While there are other page and post features that could be included, we suggest that there is another category of feature that has been excluded from this analysis (due to availability of data). We do not consider any user features. For example, we can reasonably assume that larger numbers of Facebook friends or followers a user has will increase the number of comments one of his or her posts receive. This omitted variable bias would be away from 0.

Furthermore, we do not include any qualitative features of the posts. We imagine the substantive contents of a post to be a significant driving factor in the number of engagements it generates. Further research may include a factor for some qualitative measures of the post, such as post sentiment. The same applies for posts with images (data that is not included in our data set) and their qualitative aspects. Posts with highly emotional sentiment may capture user's attentions and increase the engagement compared with posts with less emotion sentiment. This again would be a bias away from 0.

Consideration must also be given to events that occur outside of the Facebook platform that will impact engagement levels. In our case of professional sports teams, the build up to and/or post game effect may be a significant factor when modeling the engagement levels of posts on the team's Facebook pages and would be a bias away from 0. For example, we can imagine that anticipation of an upcoming "big game" can spark more engagement.

6.2 Feedback Loops

The rules of a structural model require that the effect on the measured variable to be uni-directional, flowing from cause to effect, and never back from the response variable to one of the modeled causal variables. In our case, it is plausible that we have such an effect. Our causal model assumes that the more frequently a post is shared, the more comments it will generate, simply because shares make the post available to more users on the Facebook platform. However, it is plausible that the more comments a post has generated, the more likely a user will be to share it as the number of comments validate the post as being "worth sharing" in the view of some users.

6.3 Statistical Limitations

Linear regression as performed in this analysis requires normally distributed data, which the log-transformed variable CC4 meets (see Image 2 under Data and Methodology section) after removing all zero values from the data set. All other log-transformed variables also had zero values removed from the data set. We did this because log-transforming a zero value is not possible, and also because our data contained many zeros for some of these variables which is zero-inflating the data. To include these zero values in future analyses, zero-inflated poisson regression should be used.

The five assumptions of Classical Linear Modeling include independent and identically distributed (IID) data, no perfect collinearity (unique best linear predictor), linear conditional expectation, homoskedastic errors, and normally distributed errors. Since the data meets the requirements to be considered a large sample, only the assumptions of IID and no perfect collinearity (also known as having a unique best linear predictor) need to be evaluated. The remaining 3 assumptions should be met through the Central Limit Theorem, but will be assessed anyway for validity.

6.3.1 Independent and Identically Distributed Data

The first large sample assumption that needs to be met is the the data is independent and identically distributed (IID).

The data set is the result of combining 5 variants of Facebook post data. Each variant contains the same set of columns, but scraped the posts at different times. We were unable to verify whether the posts were truly randomly scraped across the entire platform. Since multiple variants were combined into a single data set, it is possible that the same post was scraped during different variants and appear more than one time. In this case, it would make the data not IID.

Additionally, it seems that there were many posts coming from the same Facebook pages. Although there was no Facebook page identifier in the data set, many lines listed the same exact number of "page likes", leading us to deduce that they are likely from the same page. The potential clustering of posts from the same Facebook page violates the IID assumption. Another indication that this data set is not IID is that the promoted status was set to 0 (for not promoted) for all posts. Facebook gives posters the ability to pay to boost their post so that a wider audience views it. An IID data set would include a mix of posts that were promoted (1) and not promoted (0).

The lack of IID data means that the sample is not representative of the population and that results from this analysis may not be able to explain the variation on the number of comments. To remedy this problem, future analysis should scrape data in a way that randomly scrapes posts across the entire platform, only scraping each post once.

6.3.2 No Collinearity/ Unique Best Linear Predictor

The second large sample assumption that needs to be met is that there is a unique best linear predictor exists (unique BLP).

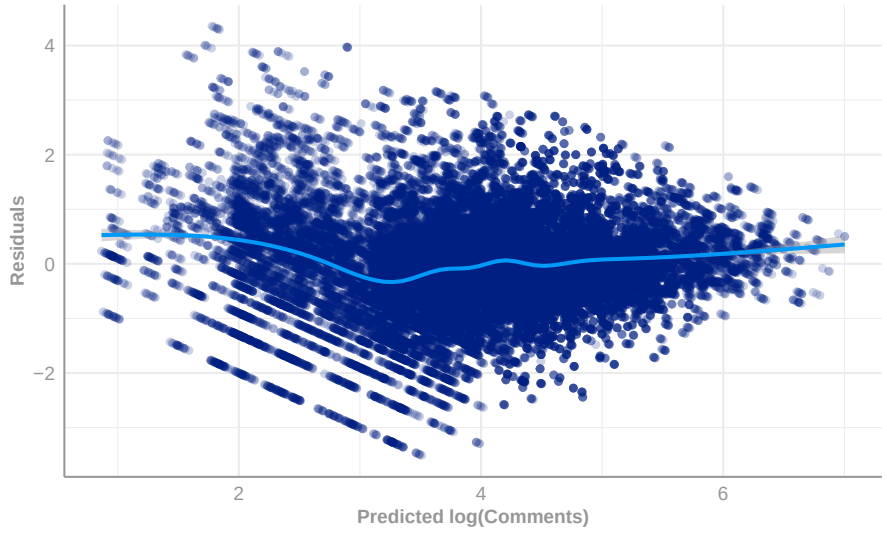
In order to prove that a unique BLP exists, there can be no perfect collinearity between the explanatory variables. In other words, each variable in the model cannot be linearly predicted by the other variables. Since no variables were dropped while creating the linear model using the R function lm(), this means that there was no perfect collinearity. To verify this, we confirmed that, in each of our models, none of the standard errors was much larger than the others.

6.3.3 Linear Conditional Expectation

The graph between predicted values and residuals show that as the predicted values increase, the residuals stay the same at approximately 0. This indicates that the predicted value does not deviate from the actual value and thus the linear conditional expectation assumption is met.

Image 14: Predicted Values VS Residuals

Scatterplot



6.3.4 Homoskedastic Errors

When assessing the assumption of homoskedastic errors, we are looking for evidence that heteroskedacicity exists among the errors. The p-value of the Breusch-Pagan test is $< 2.2e\text{-}16$, which is smaller than our alpha of 0.05. Thus we reject the null hypothesis that there is no evidence of heteroskedasticity. The assumption

of homoskedastic error is not met, and the variance of our parameters and OLS standard errors are biased. Robust standard errors should be used to obtain unbiased standard errors.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_4
## BP = 2703, df = 9, p-value < 2.2e-16
```

6.3.5 Normally Distributed Errors

When looking at the distribution of our residuals we find indeed a bell-shaped density, which is confirmed when comparing it with a normal distribution using a Q-Q Plot. This assumption is satisfied in this case.

Image 15: Model Residuals

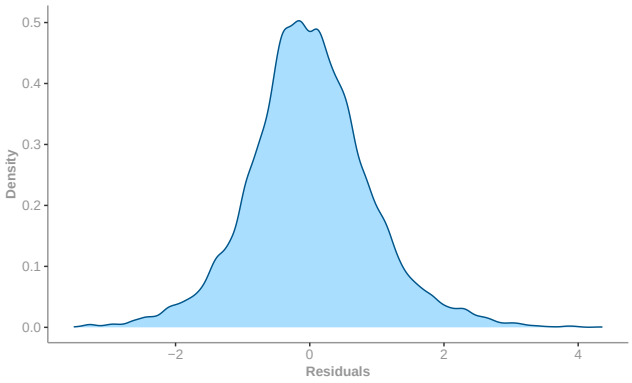
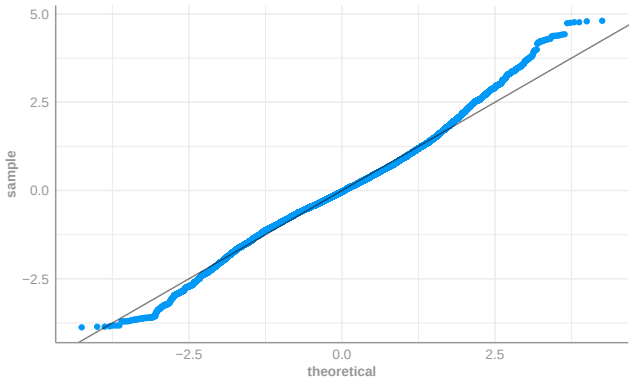


Image 16: Model Residuals



7 Conclusion and Impacts

Our research paper investigated changes to a product we have defined as the Facebook platform where users can engage with Sports Team pages. Our main research question aimed to establish a causal relationship between the number of user comments on a post and the length of a Facebook post while controlling for other variables such as post date, post share counts, page category, and page popularity. Based on the models we developed and the limitations we have outlined, we are confident that increases to post length increases user engagement if we are able to obtain a random sample of the population. Though some effects could be practically significant, it is worth noting that Omitted Variables such as Post Promoted status can be an additional explanatory variable to increase engagement.

7.1 Overall Effect

- We find the length of a post to be statistically significant across the 4 different models.
- Our model 4 was chosen as the preferred model after running ANOVA F-Tests. This model provides evidence that a 1% increase in the post length leads to an average increase of 0.036% in the number of comments that this post receives 24 hours after it has been published while holding everything else constant.
- Due to Facebook's global reach engaging billions of daily users, we find that the find can find a practically significant effect because marginal increases in the length of a Facebook post can lead to higher engagement even if the statistical values are small percentage values.
- Due to the limitations of the model in terms of promoting content and controlling for user features such as location and age, we suffer from Omitted Variable Bias that limit the interpretations of our model.
- Future research in this area needs to control for zero inflated Poisson regressions which can lead to zero covariances between the coefficients in our model. Moreover, the data must be collected across a random sample of the population to ensure that it meets the IID assumption for classical linear models.