Investigating Engagement in Public Instagram Posts

Introduction

Social media platforms such as Instagram offer users a way to reach wide audiences, and as such, many users and businesses aim to increase engagement on the platform. One key component of engagement is maximizing the number of likes a post receives relative to the number of followers the user has. One aspect of posting users and businesses have control over is the time of day at which they post. Based on Figure 1, the number of average number of likes received by posts vary by time of day and differ depending on whether the post is a video or not. This begs the research question, does the time of day at which an Instagram post is published (Morning, Afternoon, Evening or Night) affect its engagement, and what other factors are correlated with engagement? Other factors investigated include the type of post (carousel with multiple images or video post), if the post was published on a weekend, caption characteristics including length and the presence of hashtags or mentions in the caption, if a location was added, and account characteristics like follower count and following count.

The data used is a sample of over 8,000 public Instagram posts posted between September 2011 and February 2024 posted by 931 randomly sampled users across the world. Each observation is a single post. The original data included the number of likes it received, the number of users following and followed by the post's associated account at the time of posting, the account's username, and the post's caption. Instagram allows users to add a location to a post, but this is not required to publish one. If added, location data was also included, along with the time stamp at which the post was published. One post can include multiple images or videos, known as a carousel, and this was a boolean variable in the orignial data set. If the post included at least one video, this was indicated as well. The data was obtained from Kaggle and originally scraped from Instagram.

Methodology

To ensure independence across observations, one post was randomly sampled per Instagram user, resulting in over 900 unique data points. From post captions and locations, variables were created for the time day of posting, whether the post was published on a weekend, caption length, presence of hashtags or user mentions, and indicators for video posts, carousels, and location tags. Follower count and following count at the time of posting were also included as predictors. The response variable, engagement, was defined as the ratio of likes to followers and then log-transformed to stabilize variance and meet the assumptions of linear regression.

A linear regression model was used to evaluate the relationship between post characteristics and engagement. The initial model (see Appendix Table 3) included all relevant predictors and several motivated interaction terms, such as between post type and time of day, and between follower count and weekend indicator. Correlation analysis and variance inflation factors were performed and indicated that multicollinearity was not a concern. A modified automated backward selection process was applied to reduce the predictors in the model. The final model (see Table 1) was selected based on a combination of AIC, BIC, and relevance to the research question. Model assumptions (linearity, constant variance, normality of residuals, and independence) were checked and found to be satisfied through residual diagnostics and investigation of the data sampling strategy.

Results

Among the 10 variables in the regression model (Table 1), two were statistically significant at the $\alpha = 0.05$ level, the indicator variable for a post being a video, and the indicator for a post caption containing a hashtag. If the post contains a video, the change in the log-engagement ratio is predicted to decrease by 0.761, holding all other variables constant. In other words, holding all covariates constant, posting a video (as opposed to a non-video) is associated with a $e^{-0.761} - 1 = 53.2\%$ decrease in engagement (likes per follower). Holding all covariates constant, publishing a post with at least one hashtag in the caption (as opposed to a caption without a hashtag) is associated with a $e^{-0.257} - 1 = 22.7\%$ decrease in engagement (likes per follower). No particular times of day were statistically significant at the $\alpha = 0.05$ level. The night (9pm–4am) time window had the largest positive estimate, indicating a potential trend toward higher engagement when posting late at night. However, this effect was not statistically significant. It cannot be said with confidence that the time of day is correlated with engagement.

Discussion

Posts that contained videos were predicted to receive many fewer likes per follower compared to non-video posts, after adjusting for other variables. Similarly, posts with hashtags in the caption also were predicted to receive fewer likes per follower compared to non-video posts, holding all other variables constant. For users and businesses aiming to increase engagement, this finding suggests that still images without excessive captions may still be optimal. It is possible that having videos or hashtags in a post distracts the viewing user from liking it. It could also be possible the Instagram feed favors still photos when users view their feed.

The data presents some limitations. It is unknown how the data was scraped, so although it claims to be random, this cannot be confirmed. Additionally, although this study adjusts for follower and following count, other account characteristics such as account type, niche, or post frequency are not accounted for and could impact the engagement of posts.

Detailed Methodology

Many users have multiple posts in the dataset, so the observations are not independent. To account for this, one post was randomly sampled per user. Over 900 data points were still retained. The data cleaning involved making several new variables to use as predictors for engagement. Based on the timestamps, a binary indicator indicating if the post was published on a weekend and a variable representing the time of day the post was published were created (the baselines are not on a weekend and posted in the mornings). A post made between 5 am and 11:59 am is considered posted in the morning, 12 pm to 4:59 pm is considred an afternoon post, 5 pm to 8:59 is considered evening, and 9 pm to 4:59 am is considered night. These were created based on natural breaks during the day in modern society. Based on the caption, the number of characters in the caption and binary indicators if the caption contained hashtags or user mentions were added (the baselines are no hashtags and no mentions). Additionally, an indicator variable representing a location added to the post was created. The variables indicating if a post was a video or if a post was a carousel were converted to factors.

The response variable, engagement, was created by dividing the number of like a post received by the followers the posting account had at the time. Likes and follower count are naturally highly correlated, as users following account see a post in their feed, so more likes are expected. Using likes alone would favor accounts with larger followings, so this standardization allows for a fair comparison across posts from accounts of different sizes. Using the ratio of likes to followers captures the relative success of a post in terms of engagement, rather than absolute popularity. This was then log transformed (a small constant was added to avoid taking the logarithm of zero) to better satisfy the assumptions of linear regression, specifically normality of residuals and homoscedasticity, as it helps stabilize variance and reduce the strong skew present in the engagement distribution.

A linear model was chosen to investigate the relationship between post characteristics and engagement. Linear regression is well-suited for modeling continuous dependent variables, in this case, Instagram post engagement ratio, as a function of multiple continuous and categorical predictors. A beta regression model was also considered as it would be appropriate for modeling a ratio like engagement in this study. However, many posts in the data set had a ratio of 0, meaning the post received no likes. A beta regression cannot handle these values, and transformations to address this issue may make results less interpretable. Next, the correlation between all potential predictors was investigated as seen in Appendix Table 2. The Pearson correlation coefficients all had an absolute value less than 0.6, indicating multicollinearity was not a concern. The indicator and categorical variables were then encoded into factors. These variables included time of day, whether the post was a video, whether it was a carousel, whether it was posted on a weekend, and whether the caption contained a hashtag, mention, or location.

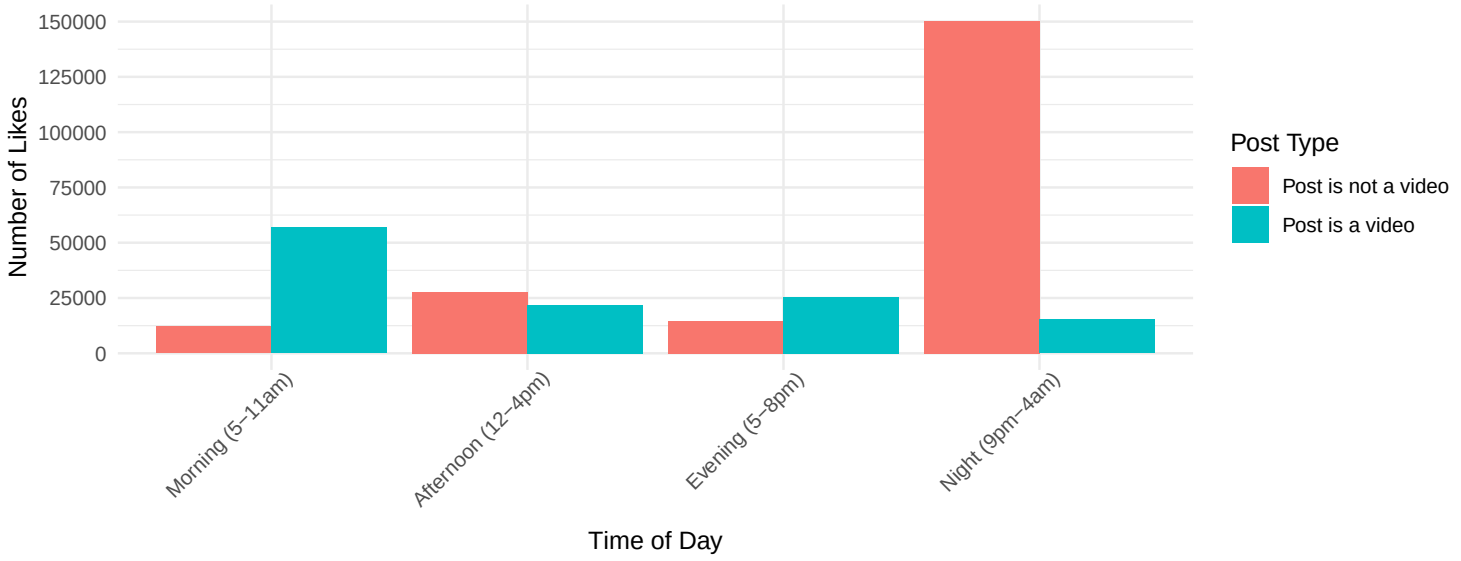
These variables were all included in the initial model, along with the follower and following count, caption length and several interaction terms. This included one between each time of day and video, as the time of day may affect how much engagement videos receive. The interaction between each time of day and if the post was published on the weekend was included to consider if daily patterns of engagement rate differ between weekends and weekdays. An interaction term between the number of followers and weekend was added to investigate if large accounts have more stable engagement on weekends and weekdays. This model is summarized in Appendix Table 3. The variance inflation factor for all predictors was less than 5, so backwards selection was then performed on this model. The backwards selection process was automated due to the high number of predictors. The backwards selection was constrained to retain time of day, the weekend, video and carousel indicators, and number of followers due to their importance in investigating the research question. The resulting model is summarized in Table 1. For the third model, these constraints were removed to further isolate significant predictors (see Appendix Table 4). To compare models, the AIC and BIC were compared for the three candidate models (Appendix Table 5). Although the third model had the lowest AIC and BIC values, the second model (Table 1) was chosen as the final model as it retained variables which are key to investigating the relationship between posting time of day and engagement.

The variance inflation factors of all predictors in the final model were confirmed to be less than 3 (see Appendix Table 6). The assumptions needed for the linear model were then assessed and determined to hold. The first assumption is linearity, in which the response variable has a linear relationship with the explanatory variables in the model. The second assumption is constant variance, in which the variance of the errors is constant regardless of what the predictor values are. It can be seen that the points are randomly scattered around the horizontal axis at 0 in the scatterplot of residuals versus predicted values (Appendix Figure 2), meaning the regression models are linear in the parameters and this assumption is satisfied. After examining plots of the histogram of the residuals (Appendix figure 3) and Normal QQ plots (Appendix Figure 4), we can see that the third Normality assumption is satisfied, as the histograms both appear bell-shaped and symmetric with a center around 0 while in the QQ plots, the points lie approximately along the diagonal line. Independence is also met, as the posts are randomly sampled one post per user, which were randomly sampled. This ensured that each observation came from distinct Instagram accounts, removing within-user correlation may bias standard errors.

Table 1: Regression Model of Instagram Engagement Using Post Characteristics and Time-of-Day Effects

Exposure Variable	Estimate of Slope	95% CI Lower	95% CI Upper	P-Value
Intercept	-4.897	-5.282	-4.513	<0.001
Number of followers	<0.001	<0.001	<0.001	0.692
Afternoon (12–4pm)	0.192	-0.179	0.562	0.310
Evening (5–8pm)	0.104	-0.268	0.476	0.583
Night (9pm–4am)	0.302	-0.088	0.692	0.129
Posted on weekend	0.104	-0.181	0.389	0.472
Caption length (characters)	<0.001	<0.001	<0.001	0.081
Caption contains hashtag	-0.257	-0.513	<0.001	0.049
Post is a video	-0.761	-1.071	-0.451	<0.001
Post is a carousel	-0.211	-0.524	0.102	0.185

Figure 1: Engagement Patterns Across Time-of-Day and Post Type



Appendix

Table 2: Correlation between Model Predictors

	Number of followers	Number of accounts following	Time of day	Posted on weekend	Caption length (characters)	Caption contains hashtag	Caption contains user tag	Post has location tag	Post is a video	Post is a carousel
Number of followers	1.000	-0.005	0.053	-0.011	-0.048	-0.063	-0.025	-0.045	-0.014	0.070
Number of accounts following	-0.005	1.000	-0.023	0.068	0.078	0.042	0.045	0.051	0.056	-0.028
Time of day	0.053	-0.023	1.000	-0.060	-0.048	-0.051	0.006	-0.038	-0.004	-0.025
Posted on weekend	-0.011	0.068	-0.060	1.000	-0.053	-0.012	0.059	0.053	-0.040	0.075
Caption length (characters)	-0.048	0.078	-0.048	-0.053	1.000	0.174	0.178	0.053	0.002	0.040
Caption contains hashtag	-0.063	0.042	-0.051	-0.012	0.174	1.000	0.232	0.117	0.163	-0.113
Caption contains user tag	-0.025	0.045	0.006	0.059	0.178	0.232	1.000	0.076	0.116	-0.031
Post has location tag	-0.045	0.051	-0.038	0.053	0.053	0.117	0.076	1.000	-0.093	0.105
Post is a video	-0.014	0.056	-0.004	-0.040	0.002	0.163	0.116	-0.093	1.000	-0.537
Post is a carousel	0.070	-0.028	-0.025	0.075	0.040	-0.113	-0.031	0.105	-0.537	1.000

Table 3: Model 1: Regression Model of Instagram Engagement Using Post Characteristics and Time-of-Day Effects

Exposure Variable	Estimate of Slope	95% CI Lower	95% CI Upper	P-Value
Intercept	-4.944	-5.430	-4.457	<0.001
Number of followers	<0.001	<0.001	<0.001	0.746
Number of accounts following	<0.001	<0.001	<0.001	0.643
Afternoon (12–4pm)	0.274	-0.256	0.805	0.310
Evening (5–8pm)	0.193	-0.328	0.714	0.467
Night (9pm–4am)	0.528	-0.019	1.074	0.058
Posted on weekend	0.528	-0.154	1.209	0.129
Caption length (characters)	<0.001	<0.001	<0.001	0.117
Caption contains hashtag	-0.228	-0.495	0.039	0.094
Caption contains user tag	-0.135	-0.398	0.128	0.313
Post is a video	-0.893	-1.549	-0.236	0.008
Post is a carousel	-0.185	-0.502	0.132	0.252
Post has location tag	-0.039	-0.299	0.222	0.770
Afternoon × Video	0.256	-0.518	1.031	0.516
Evening × Video	0.177	-0.609	0.963	0.658
Night × Video	-0.121	-0.931	0.689	0.769
Afternoon × Weekend	-0.627	-1.460	0.206	0.140
Evening × Weekend	-0.514	-1.349	0.322	0.228
Night × Weekend	-0.647	-1.559	0.265	0.164
Video × Weekend	0.204	-0.406	0.815	0.511
Followers × Weekend	<0.001	<0.001	<0.001	0.886

Table 4: Model 3: Regression Model of Instagram Engagement Using Post Characteristics and Time-of-Day Effects

Exposure Variable	Estimate of Slope	95% CI Lower	95% CI Upper	P-Value
Intercept	-4.812	-5.014	-4.610	<0.001
Caption length (characters)	<0.001	<0.001	<0.001	0.056
Caption contains hashtag	-0.264	-0.519	-0.009	0.043
Post is a video	-0.646	-0.908	-0.384	<0.001

Table 5: AIC and BIC of Candidate Models

Model	AIC	BIC
Candidate Model 1	2285.703	2382.690
Candidate Model 2 (chosen model)	2270.126	2318.620
Candidate Model 3	2262.942	2284.985

Table 6: VIFs for Chosen Regression Model

	Generalized VIF	Df (Factor Levels - 1)	Scaled GVIF
Number of followers	1.014785	1	1.007366
Time of day	1.021737	3	1.003590
Posted on weekend	1.014499	1	1.007223
Caption length (characters)	1.043284	1	1.021413
Caption contains hashtag	1.067605	1	1.033250
Post is a video	1.435685	1	1.198201
Post is a carousel	1.429504	1	1.195619

Model Assumption Checks for Regression

Figure 2

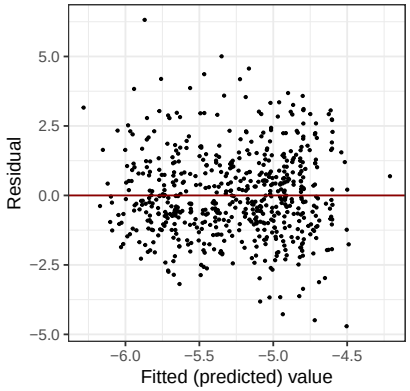


Figure 3

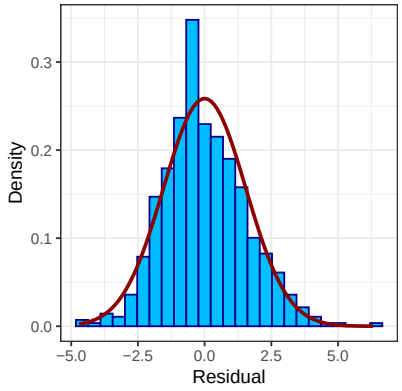


Figure 4