**Peter Fontana**

**SlideRule / Springboard Capstone Project**

**Objective**

The objective of this capstone project is to arrive at a model that will best reveal variables for posting that drive social interactions on various social platforms for brands, organizations, or other pages with large community sizes. Armed with this knowledge, strategies can be implemented to impact these independent variables in order to help drive the dependent variable of interactions.

**Initial Data Sourcing (Tumblr)**

Tumblr was initially chosen for this analysis, due to data access. Seven tumblrs were researched for active post frequency (in a day) over the past 12 months. Tumblr post data sourcing was conducted via the Apigee Tumblr API interface, harvesting date, time, URL, and total notes for 1,000 posts. The raw data output was cleaned through a variety of manipulations in Excel. All posts on dates which had at least one post receiving unusually high notes were removed; a standard rule was applied across Tumblrs, removing any posts that received a number of notes that exceeded more than one standard deviation from the mean. These outliers were likely caused by paid investment and were not included in the analysis. Calculated fields were added to create variables for hour (ET), post frequency (in a day), and a notes index (number of notes per post relative to the mean notes received for the individual Tumblrs). Since followers are unavailable through the API, this index was used to provide a more even metric across the seven Tumblrs.

http://calvinklein.tumblr.com/archive

http://sendthemasignal.tumblr.com/archive

http://disney.tumblr.com/archive

http://gq.tumblr.com/archive

http://adidasoriginals.tumblr.com/archive

http://comedycentral.tumblr.com/archive

http://glamour.tumblr.com/archive

**Initial Data Visualizations & Observations**

Prior to running a regression model to find relationships between engagement and different variables, the variables were visualized to identify observational relationships to guide the model.

Notes received by posts were graphed by the hour (ET) in which they were posted. The chart below is a scatter plot output showing that posts published between 12pm and 2:59am appear to receive more notes than posts published between 3am and 11:59am.

*R syntax for reference*

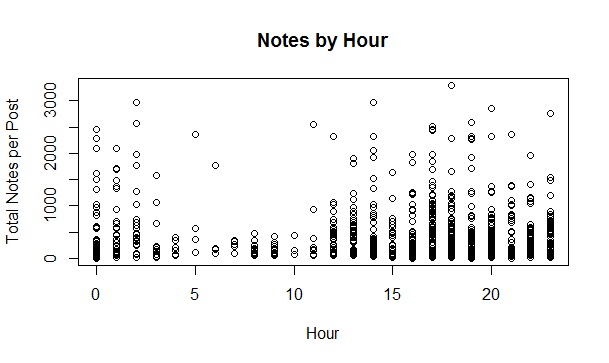
*tumblr2=read.csv("tumblrData2.csv")*

*str(tumblr2)*

*hour2 <- tumblr2$hour*

*notes2 <- tumblr2$notes*

*plot(hour2, notes2, main="Notes by Hour", xlab="Hour", ylab="Total Notes per Post")*

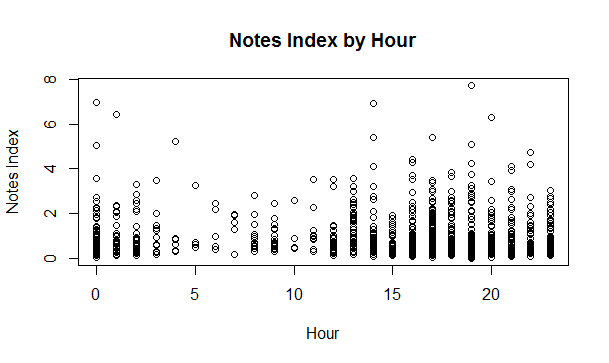


To account for the varying community sizes of the seven Tumblrs, this same chart was produced using the Notes Index instead of raw number of Notes. This index divided the number of notes received per post by the mean notes for the individual Tumblr. The observation remains largely the same with more engagement being received between 12pm and 2:59am between 2am and 11:59am.

*R syntax for reference*

*notes\_index2 <- tumblr2$notes\_index*

*plot(hour2, notes\_index2, main="Notes Index by Hour", xlab="Hour", ylab="Notes Index")*



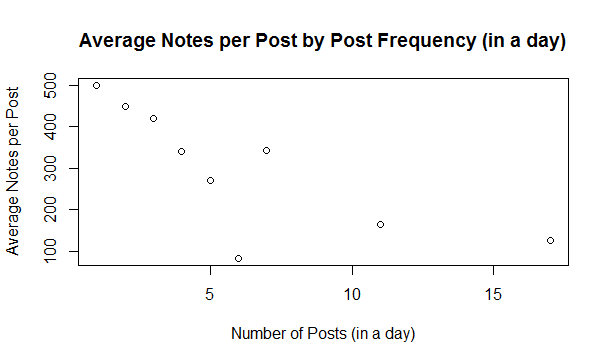
To observe the impact of post frequency (in a day) on the average number of notes per post, an aggregate data set was created and plotted. The chart reveals that as more posts are published on a single day, engagement is spread across more posts and that doubling or tripling the number of posts does not double or triple the engagement received. The observations for post frequencies > 5 appear erratic due to the sample size of days with posting frequencies > 5 (further demonstrated by the table proceeding).

*R syntax for reference*

*AverageNotesByPostFrequencyTable2 <- aggregate(tumblr2$notes, list(tumblr2$post\_frequency), FUN=mean)*

*AverageNotesByPostFrequencyTable2*

*plot(AverageNotesByPostFrequencyTable2, main="Average Notes per Post by Post Frequency (in a day)",xlab="Number of Posts (in a day)",ylab="Average Notes per Post")*

**

In the supplementary table below, the variables in the chart are presented alongside the average total notes received (notes per post \* frequency (in a day)). The multiplier variable calculates the incremental notes received above 1 post. This illustrates that the doubling of the post frequency (in a day) does not equate to a doubling of engagement. The figures in red denote the frequency (in a day) observations with a low sample size (less than 5) across the 1,000 posts on the seven Tumblrs.

**Frequency/Day** **Avg Notes per Post Avg Total Notes Multiplier Sample Obs#**

1 499 500 1 225

2 450 900 1.8 108

3 420 1263 2.5 79

4 340 1361 2.7 38

5 271 1355 2.7 15

6 82 497 1.0 2

7 343 2402 4.8 3

11 163 1798 3.6 4

17 125 2138 4.3 1

**Tumblr Linear Regression Model**

Despite these observations, a linear regression showed no significant impact on the variability in notes received and the hour (ET) published or post frequency (in a day). The low R-squared indicates the variability is not explained by these relationships.

In the two linear regressions below, a subset of the Tumblr data was created with a post frequency (in a day) less than 6, to exclude the observations with a low sample size. Notes\_index as well as Notes were used as the dependent variables in the two linear regressions, with neither producing significant results:

Notes\_index linear regression:

*R syntax for reference*

*#linear regression with notes\_index*

*tumblr3 = subset(tumblr2, post\_frequency<6)*

*notes\_indexReg=lm(notes\_index ~ hour + post\_frequency,data=tumblr3)*

*summary(notes\_indexReg)*

*notes\_indexReg$residuals*

*SSE = sum(notes\_indexReg$residuals^2)*

*SSE*

*#linear regression with notes\_index*

*Call:*

*lm(formula = notes\_index ~ hour + post\_frequency, data = tumblr3)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-1.0224 -0.5961 -0.3279 0.2612 6.7026*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 1.145723 0.100462 11.404 <2e-16 \*\*\**

*hour -0.005579 0.004611 -1.210 0.227*

*post\_frequency -0.024412 0.025471 -0.958 0.338*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 0.9599 on 902 degrees of freedom*

*Multiple R-squared: 0.002561, Adjusted R-squared: 0.000349*

*F-statistic: 1.158 on 2 and 902 DF, p-value: 0.3146*

*SSE = [1] 831.153*

*#correlation with notes\_index*

*> cor(tumblr3$notes\_index, tumblr3$hour)*

*[1] -0.03930385*

*> cor(tumblr3$notes\_index, tumblr3$post\_frequency)*

*[1] -0.03069046*

Notes linear regression:

*R syntax for reference*

*tumblr3 = subset(tumblr2, post\_frequency<6)*

*notesReg=lm(notes ~ hour + post\_frequency,data=tumblr3)*

*summary(notesReg)*

*notesReg$residuals*

*SSE = sum(notesReg$residuals^2)*

*SSE*

*#linear regression with notes*

*Call:*

*lm(formula = notes ~ hour + post\_frequency, data = tumblr3)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-508.9 -320.7 -182.2 116.2 2947.8*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 569.2834 54.4768 10.450 < 2e-16 \*\*\**

*hour -0.5473 2.5004 -0.219 0.826782*

*post\_frequency -53.8187 13.8117 -3.897 0.000105 \*\*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 520.5 on 902 degrees of freedom*

*Multiple R-squared: 0.01657, Adjusted R-squared: 0.01439*

*F-statistic: 7.597 on 2 and 902 DF, p-value: 0.0005347*

*SSE = [1] 244397665*

*#correlation with notes*

*> cor(tumblr3$notes, tumblr3$hour)*

*[1] -0.003409674*

*> cor(tumblr3$notes, tumblr3$post\_frequency)*

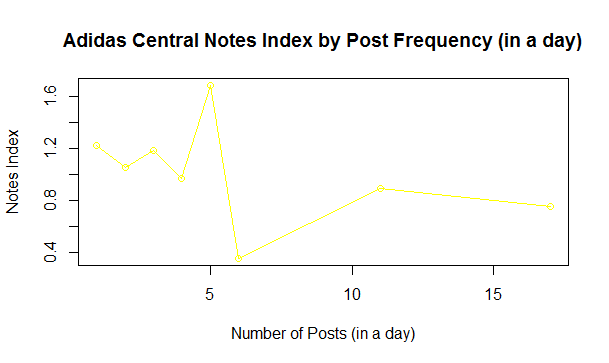
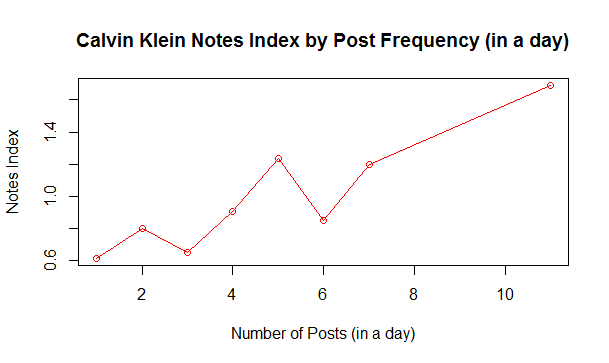
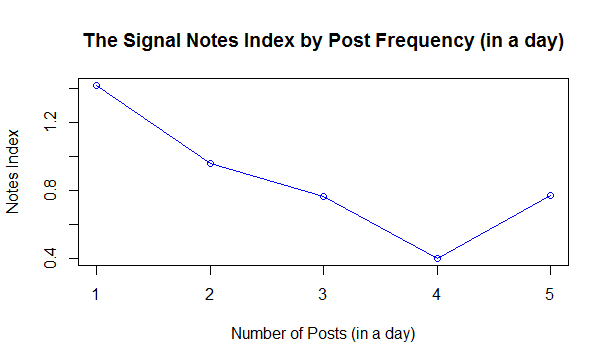
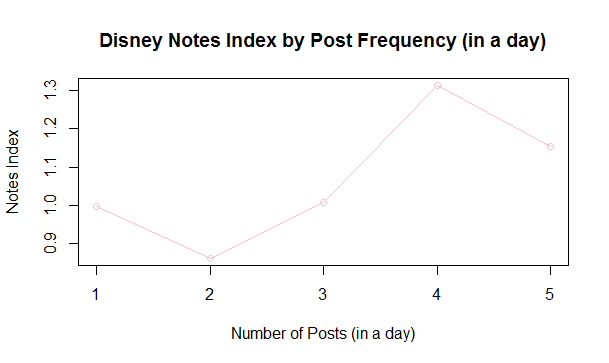
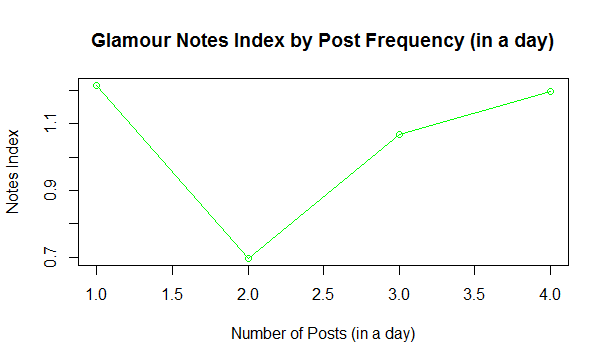
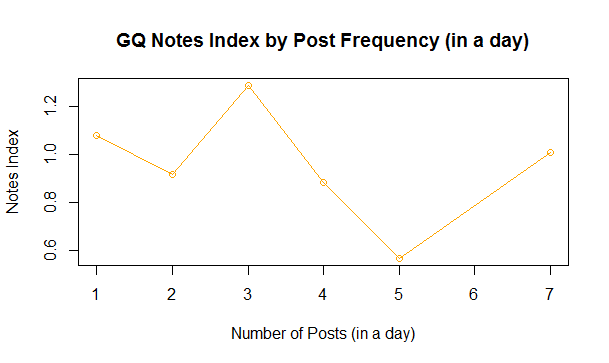
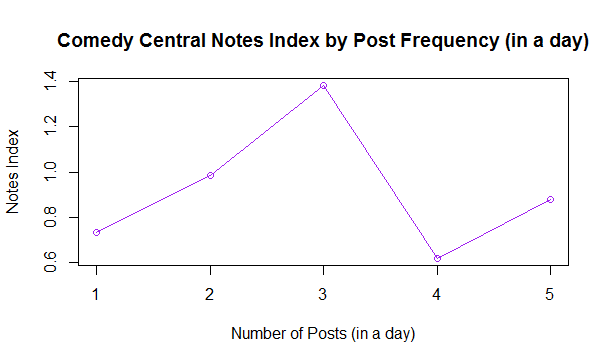
*[1] -0.1285052*

**Tumblr Challenges**

While the model appears to be a good methodology for revealing variables impacting engagement, no independent variables were significantly impacting the number of notes or the notes index on Tumblr. This could be due to the nature of Tumblr post shelf-life compared to other social networks. For example, a Facebook post receives between 60% and 80% of its lifetime engagement within the first 48 hours of going live; this long tail drops off quickly thereafter. However, on Tumblr a post will continue to receive steady engagement through reblogs months or even years later. As a result of this test and user behavior, further analysis was conducted on a limited data set on Facebook.

**Observing Bias Among the Seven Tumblr Sample**

Before moving onto the Facebook test, each of the seven Tumblrs sampled for this analysis were compared for any observable trends in notes relative to post frequency (in a day). Notes index (number of notes per post relative to the mean notes received for the individual Tumblrs) was graphed against the number of posts (in a day), or post frequency (in a day). Each Tumblr is shown using a different color in each line graph below:



Based on observations from these seven graphs, notes per post does not appear to show a consistent trend across Tumblrs. In fact, each tumblr’s post frequency (in a day) shows a different pattern in notes index as number of posts (by day) increases. With these observations and with consideration for the aforementioned challenges in analyzing Tumblr patterns (see **Tumblr Challenges**), this analysis approach may be more suited to a platform like Facebook, Twitter, or Instagram.

**Revised Data Sourcing (Facebook)**

Over 950 posts were sourced from Banana Republic’s Facebook page from over two years of activity on the page. This data was acquired for academic purposes via the page’s Insights post export. All posts on dates which had at least one post receiving paid impressions were removed. Calculated fields were added to create variables for hour (ET), post frequency (in a day), and engagement rate (number of engaged users divided by the reach of the post).

https://www.facebook.com/BananaRepublic/

**Initial Data Visualizations & Observations**

Charting the interactions by hour (ET) reveals that Banana Republic is mostly posting from 6am to 6pm, and is seeing the most interactions (likes, comments, and shares) on posts published between 9am and 4pm.

*R syntax for reference*

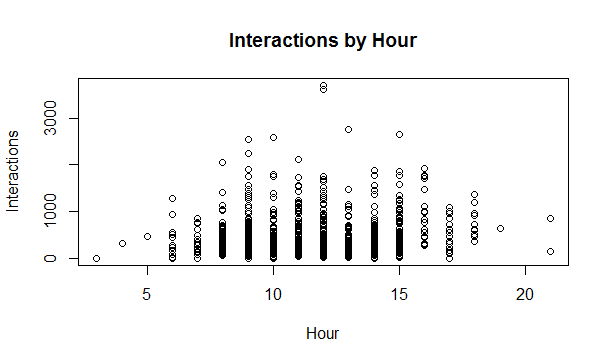
*BRFB=read.csv("BRFB.csv")*

*str(BRFB)*

*FBhour <- BRFB$hour*

*FBinteractions <- BRFB$interactions*

*plot(FBhour, FBinteractions, main="Interactions by Hour", xlab="Hour", ylab="Interactions")*



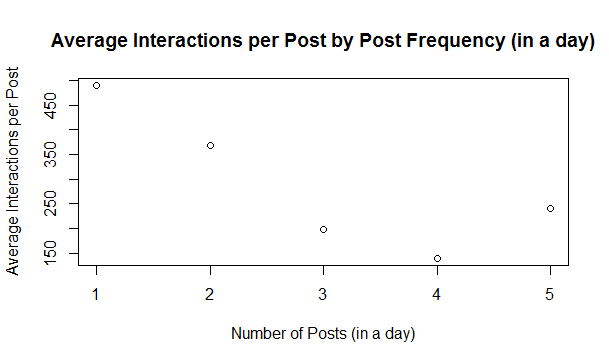
To observe the impact of post frequency (in a day) on the average number of interactions per post, an aggregate data set was created and plotted. Similar to the Tumblr analysis, the chart reveals that as more posts are published on a single day, engagement is spread across more posts and that doubling or tripling the number of posts does not double or triple the engagement received. The observations for post frequencies > 4 appear erratic due to the sample size of days with posting frequencies > 4 (further demonstrated by the table proceeding).

*R syntax for reference*

*FBInteractionsByPostFrequencyTable <- aggregate(BRFB$interactions, list(BRFB$post\_frequency), FUN=mean)*

*FBInteractionsByPostFrequencyTable*

*plot(FBInteractionsByPostFrequencyTable, main="Average Interactions per Post by Post Frequency (in a day)",xlab="Number of Posts (in a day)",ylab="Average Interactions per Post")*



In the supplementary table below, the variables in the chart are presented alongside the average total interactions received (interactions per post \* frequency (in a day)). The multiplier variable calculates the incremental notes received above 1 post. This illustrates that the doubling of the post frequency (in a day) does not equate to a doubling of engagement. The figures in red denote the frequency (in a day) observations with a low sample size (less than 4) across the 1,000 posts on the Facebook page.

**Frequency/Day** **Avg Interactions/Post Avg Total Interactions Multiplier Sample Obs#**

1 489 489 1 625

2 368 737 1.5 260

3 198 595 1.2 33

4 140 561 1.1 8

5 242 1208 2.5 5

This pattern is similar to the metric reach, which is controlled by Facebook’s algorithm (further explanation in the next section).

**Frequency/Day** **Avg Reach/Post Avg Total Reach Multiplier Sample Obs#**

1 30992 30992 1 625

2 28024 56048 1.8 260

3 18542 55627 1.8 33

4 10187 40748 1.3 8

5 14189 70947 2.3 5

**Facebook Linear Regression Models**

Linear regression models were initially run with three independent variables:

1. hour (ET) of the publishing
2. post frequency on the day
3. day of the week for the post

In two models, these were tested against interactions on the posts as well as reach. Significant results were observed from the Banana Republic Facebook data set for both hour (ET) and post frequency (in a day)in both models. Day of the week did not yield significant results. Hour (ET) of the day yielded a positive coefficient, suggesting that later posts have a better chance of garnering reach and interactions than earlier posts—in line with the observations. Post frequency (in a day)yielded a negative relationship in the model, suggesting that a lower post frequency (in a day) results in a higher amount of reach and number of interactions—also in line with observations.

Much like Tumblr, Facebook has unique challenges. It’s algorithm for brands, organizations, and other pages is designed to prevent spamming of users. The more messages sent out, the less reach they will receive and therefore fewer interactions per post. Although these variables are designed to avoid marketers from “gaming the system,” these regression findings are consistent with this algorithm behavior. Furthermore, since posting to Facebook is essentially free (in terms of media placement, and discounting production and labor), maximizing the amount of total reach and interactions is the goal.

Interactions linear regression:

*R syntax for reference*

*#linear regression with interactions*

*FBinteractionsReg=lm(interactions ~ hour + post\_frequency + dowNum,data=BRFB)*

*summary(FBinteractionsReg)*

*FBinteractionsReg$residuals*

*SSE = sum(FBinteractionsReg$residuals^2)*

*SSE*

*Call:*

*lm(formula = interactions ~ hour + post\_frequency + dowNum, data = BRFB)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-525.54 -238.78 -127.10 97.72 3155.74*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 332.409 67.225 4.945 9.01e-07 \*\*\**

*hour 29.125 5.106 5.704 1.56e-08 \*\*\**

*post\_frequency -131.705 20.756 -6.345 3.43e-10 \*\*\**

*dowNum -7.946 6.802 -1.168 0.243*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 411.3 on 953 degrees of freedom*

*(1 observation deleted due to missingness)*

*Multiple R-squared: 0.06479, Adjusted R-squared: 0.06185*

*F-statistic: 22.01 on 3 and 953 DF, p-value: 8.733e-14*

*R syntax for reference*

*#linear regression with reach*

*FBreachReg=lm(total\_reach ~ hour + post\_frequency + dowNum,data=BRFB)*

*summary(FBreachReg)*

*Call:*

*lm(formula = total\_reach ~ hour + post\_frequency + dowNum, data = BRFB)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-41412 -16865 -9592 6284 256764*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 2107.9 4927.3 0.428 0.669*

*hour 3429.0 374.2 9.162 < 2e-16 \*\*\**

*post\_frequency -6574.5 1521.4 -4.322 1.71e-05 \*\*\**

*dowNum -426.7 498.5 -0.856 0.392*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 30150 on 953 degrees of freedom*

*(1 observation deleted due to missingness)*

*Multiple R-squared: 0.09025, Adjusted R-squared: 0.08738*

*F-statistic: 31.51 on 3 and 953 DF, p-value: < 2.2e-16*

**Expanded Facebook Linear Regression Models (Interactions)**

To expand beyond hypotheses to see what else might be unexpectedly impacting interactions, a linear regression was run for each dependent variable with these independent variables:

1. hour (ET)
2. post\_frequency (in a day)
3. dowNum (day of the week)
4. year
5. month
6. Wkend (1 for weekend, 0 for weekday)
7. total\_reach
8. viral\_reach
9. impressions
10. engaged\_users
11. engagement\_rate

As it turns out, several of these returned significant results for interactions:

*R syntax for reference*

*#linear regression with interactions*

*FBinteractionsReg=lm(interactions ~ hour + post\_frequency + dowNum + year + month + total\_reach + viral\_reach + impressions+ Wkend + engaged\_users + engagement\_rate,data=BRFB)*

*summary(FBinteractionsReg)*

*Call:*

*lm(formula = interactions ~ hour + post\_frequency + dowNum +*

*year + month + hour + total\_reach + viral\_reach + impressions +*

*Wkend + engaged\_users + engagement\_rate, data = BRFB)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-1235.67 -204.17 -27.26 154.89 2016.06*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 6.168e+05 1.270e+05 4.857 1.64e-06 \*\*\**

*hour 2.755e+00 6.355e+00 0.433 0.664861*

*post\_frequency -4.441e+01 2.549e+01 -1.742 0.082145 .*

*dowNum -1.576e+01 8.596e+00 -1.834 0.067356 .*

*year -3.060e+02 6.304e+01 -4.854 1.66e-06 \*\*\**

*month -3.391e+01 8.821e+00 -3.844 0.000138 \*\*\**

*total\_reach 9.918e-03 2.286e-03 4.339 1.76e-05 \*\*\**

*viral\_reach 5.063e-02 4.663e-03 10.857 < 2e-16 \*\*\**

*impressions -3.607e-03 1.072e-03 -3.366 0.000828 \*\*\**

*Wkend 1.927e+02 3.974e+01 4.849 1.70e-06 \*\*\**

*engaged\_users 2.690e-02 8.619e-03 3.121 0.001916 \*\**

*engagement\_rate -3.937e+01 1.810e+02 -0.218 0.827872*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 355.4 on 457 degrees of freedom*

*(489 observations deleted due to missingness)*

*Multiple R-squared: 0.5175, Adjusted R-squared: 0.5058*

*F-statistic: 44.55 on 11 and 457 DF, p-value: < 2.2e-16*

While post frequency (in a day)and remained significant, several others also showed significant results:

Year & Month – As time moved forward (years and months), interactions decreased. This is consistent with the decrease in organic exposure Facebook’s algorithm has permitted in the years spanning this analysis (2013 to 2015).

total\_reach, viral\_reach, impressions – Reaching more people is associated with higher interactions, which makes sense, and confirms that reaching more people will result in more interactions. However, higher impressions resulted in lower interactions. While this seems inconsistent with the reach finding, higher impressions can be due to the same people seeing posts more than once. Therefore, this is also closely associated with post frequency (in a day), which also results in a lower interaction number.

Wkend – Taking a deeper look at the day of week and separating weekdays from weekends, the regression revealed a significant, positive co-efficient for the weekend variable, meaning interactions increase on weekends.

engaged\_users – Engaged users being significant, and positively associated is not surprising as interactions are a component of engaged users. It is interesting that this finding is only significant at 99% and not 99.9%---like other variables.

**Expanded Facebook Linear Regression Models (Reach)**

The same expanded regression analysis was repeated for reach, resulting in the following:

*R syntax for reference*

*#linear regression with reach*

*FBreachReg=lm(total\_reach ~ hour + post\_frequency + dowNum + year + month + Wkend + interactions + viral\_reach + impressions + engaged\_users + engagement\_rate,data=BRFB)*

*summary(FBreachReg)*

*Call:*

*lm(formula = total\_reach ~ hour + post\_frequency + dowNum + year +*

*month + hour + Wkend + interactions + viral\_reach + impressions +*

*engaged\_users + engagement\_rate, data = BRFB)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-64909 -2739 -880 1364 55201*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -5.643e+06 2.599e+06 -2.172 0.03039 \**

*hour 9.938e+01 1.274e+02 0.780 0.43573*

*post\_frequency 3.045e+01 5.130e+02 0.059 0.95269*

*dowNum 2.134e+02 1.728e+02 1.235 0.21732*

*year 2.803e+03 1.290e+03 2.173 0.03029 \**

*month 1.925e+02 1.795e+02 1.072 0.28424*

*Wkend -5.228e+02 8.168e+02 -0.640 0.52247*

*interactions 3.989e+00 9.194e-01 4.339 1.76e-05 \*\*\**

*viral\_reach -1.545e-01 1.047e-01 -1.476 0.14060*

*impressions 4.406e-01 6.962e-03 63.296 < 2e-16 \*\*\**

*engaged\_users 1.010e+00 1.682e-01 6.002 3.97e-09 \*\*\**

*engagement\_rate -1.128e+04 3.591e+03 -3.142 0.00178 \*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 7128 on 457 degrees of freedom*

*(489 observations deleted due to missingness)*

*Multiple R-squared: 0.9644, Adjusted R-squared: 0.9635*

*F-statistic: 1124 on 11 and 457 DF, p-value: < 2.2e-16*

interactions – This is the significant reciprocal of the significant findings above for interactions as the dependent variable

impressions – It makes sense that reach would have a significant, positive relationship with impressions—as there are more people reached by a post, there will be more impressions on the post

engaged\_users – A significant, positive relationship between reach and engaged users makes sense as there is a higher group of reached people with the opportunity to engage with posts

engagement\_rate – As reach is the denominator to this calculated field, it makes sense that a significant, negative relationship would exist between these two

**Decision Trees**

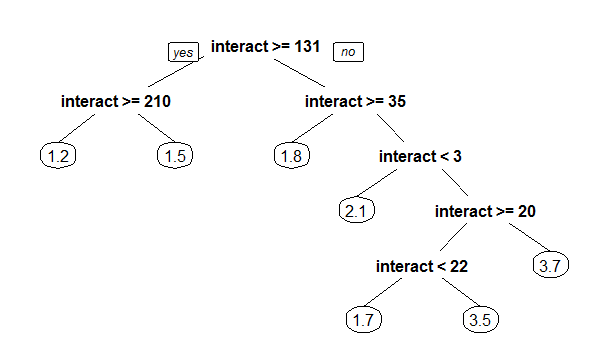
To better understand the thresholds within post frequency (in a day) and the number of interactions or reach received on a post, regression tress were run for post frequency (in a day) among both variables: interactions and reach.

Both trees showed the highest interactions and reach with behavior of posting just over 1 post on average, or 1 to 2 posts in reality. This is consistent with the observation that posting more often does not garner more interactions or reach.

*R syntax for reference*

*FBTree\_Interactions=rpart(post\_frequency ~ interactions,data=BRFB,minbucket=2)*

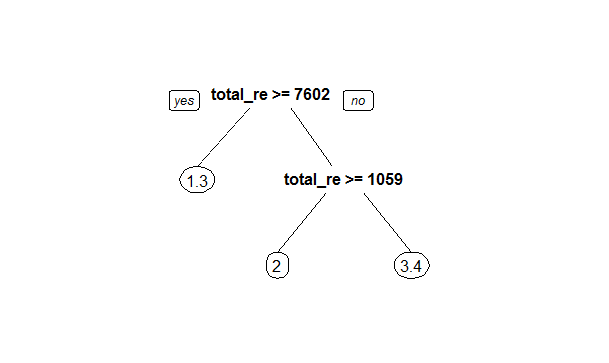
*prp(FBTree\_Interactions)*

****

*R syntax for reference*

*FBTree\_Reach=rpart(post\_frequency ~ total\_reach,data=BRFB,minbucket=5)*

*prp(FBTree\_Reach)*



**Facebook Challenges and Conclusion**

With these observations and the confirmation of the regression model, Banana Republic should be posting 1 to 2 times per day to maximize reach and interactions, as it receives no more reach and only minimal incremental interactions for a 3rd post. Posting 3 times will do no harm, but over 3 times may result in lower reach and punishment from the algorithm based on observations. A later hour (ET) in the day will produce a stronger result, and Banana Republic should continue with this strategy. Posting more on the weekends should expect more engagement per post than during the week.

As time moves forward, it seems the number of social interactions on these organic posts (posts without paid support and subject to Facebook’s algorithm) should decrease as they have year-over-year and across the months of a year.

Using a similar approach, additional Facebook pages can be analyzed. Similarly, as demonstrated by the failure to capture significant results on Tumblr, this model can be replicated across social networks with a content shelf life more like Facebook (e.g. Instagram, Twitter).