Data Analysis 2 & Coding 1 - Final Term Project

Analysis of Interactions and Reach on a Facebook Post

Ali Hasnain Khan Sial (2101874)

12/17/2021

Introduction

The goal of this document is to investigate if a correlation exists between interactions on a Facebook post and its reach and also to explore if other factors have an impact. Working in the marketing sector provided me with the opportunity to explore various digital media tools and their mechanics. The idea of how significant each components involved in the success of a social media post has always intrigued me. Therefore, Inspired by the research ***S. Moro, P. Rita and B. Vala.*** (please click for link), I decided to use the same raw data to run an analysis of my own.

Data

The data was obtained from open source at ***Kaggle*** (please click for link). As mentioned earlier, the data being used for the analysis was collected as part of research conducted on Facebook posts of a cosmetic brand for the entire year. The name of the brand is not disclosed due to confidentially purposes. Data has a total of **500 observations** and 17 variables, but only **5 variables will be used for the analysis**.

Data Cleaning & Data Munging

The data obtained was in a completely raw CSV format which required cleaning. Upon loading the data, I realised that it was stored in one column using a delimiter and had to be manually separated into individual columns. To begin the cleaning process, I started by dropping the columns from the table which wont be included in the analysis. The remaining 5 variables are as follows:

- **Y: Total Interactions**: as the name suggests, this variable records the total interactions on a post which means sum of likes, shares and comments (this is the dependent variable).
- **X: Total Reach**: records the total reach per post (this is our independent variable)
- **Z1: If the post was paid?**: a binary variable that records if the post was paid (first confounding variable)
- **Z2: What time of the day was it posted at?**: what hour of the day the post was uploaded between 0-24 (second confounding variable)
- **Z3: Which type of post was it?**:a categorical variable that records if the post was a photo, status ,link or a video (third confounding variable).

The next step was to adjust the type of each columns because originally they were all recorded as a character variable. After that I reviewed and investigated for further filtration and distribution of the variables. The filtration and adjustments made to the data based on my observation are as follows:

- *Total Interactions*: There were few very large values and few very small values and the distribution was also right skewed. Therefore, I filtered the data for values outside the range of 0 to 2,000. The large values were also quite far away from 2,000, thus it made sense to restrict the data within that range. Furthermore, I also decided to take log for the values to make the distribution more normal, an additional column was added to record the values. The graph for the with and without log normal distribution of the Total interactions is provided in the appendix (Exhibit 1).
- *Total Reach*: Again there were few exceptionally large values which had to be removed. I filtered total reach for values below 80,000. Additionally, the distribution had to be be normalized as it was right skewed. A new column was added for log normal values and the graph for distribution is provided in the appendix (Exhibit 2).
- *Paid Post*: 1 value for variable was missing, thus the observation was removed from the data.
- *Additional Variables*: 4 new binary variables were added for the categorical variable "Type of Post". The new variables are **Photo, Status, Link and Video**.

After completing all the transformations and alterations, the data has a total **11 variables with 481 observations**. The final variables are presented in the descriptive statistics Table 1 (Exhibit 3 in appendix). The only variable missing from the table is the categorical variable "Type of Post", the binary variables created for that are included in the table. Apart from the binary variable, Table 1 shows that for total interaction and total reach the mean is greater than median (interactions: mean = 194.19, median = 122 and reach: mean = 11 361.23, median = 5240). This clearly indicated towards a right skewed distribution, while on the other hand, the log value for both variables suggests that mean and median are close to each other, thus the data is more normally distributed. For the numeric confounding variable "time of the day", the distribution seems to be rather normal.

Expectations

Going back to the original question stated in the introduction of this document, "The idea is to establish if there is a linear relationship between the total interaction on a post with total reach". I also believe that other variables such as the time of the day the post was uploaded, what type of post it was and whether it was paid also have an impact on the total interactions. Meanwhile, based on my understanding of the digital marketing platform like Facebook, when and what type of post also has an effect on the total reach of the post and will be considered as controlling variables in this analysis. Therefore, to prove the hunch about this pattern of association between the dependent and independent variable, my hypothesis for this analysis is:

$$H_0 := \beta_1 \neq 0$$
$$H_A := \beta_1 = 0$$

Investigating Patterns of Association

I begin to test this hypothesis by first understanding the key pattern of association which is between interactions and reach based on non parametric regression. For this, I created a Lowess Curve for log of interactions and log of reach. The graph for this has been added in the appendix (Exhibit 4) and shows that there tends to be a relation between both the variables. The slope of the cure is positive and increasing for the first part and tends to slightly decrease in the second part yet staying positive. The confidence interval, as per the graph, is narrow since most of the observations are close to the line of best fit created by this function. The same chart was also used to examine if linear splines are required to be added in the model for the independent variable, which in this case wasn't needed. In the same way, as shown in Exhibit 5, the line of best fit was created for the numeric controlling variable i.e. Time of the day (Post_Hour). As per the graph there tends to be some pattern of association between interaction and the time the post was uploaded but it appears to be slightly on the negative side. Based on what we know at this stage, that the later the post is uploaded the less interaction it will have. The confidence interval appears to be narrow in the first half of the curve and tends to get wider in the later half, meaning that observations for most part are not very close to the Lowess curve. Additionally, I also decided to the examine the linear pattern of association between the dependent and independent variable. In for order to do so, I used the Fit Linear Model as shown in Exhibit 6.

Next, to further understand how all the variables interact with and impact each other I also created a **Correlation Matrix**, this can be reviewed in appendix (Exhibit 7). I agree that these results are not very reliable, but as the name suggests, it can provide the analyst with a basic idea about each variable and its relation to the other variables. Based on the correlations depicted by the matrix we can observe that indeed there tends to be some pattern of association between most of the variables. The log of interactions (dependent variable) has a positive relation with log of reach, paid post, post being a status or video. On the contrary, log of interactions has a negative relation with post being a link and no relation if it is a photo. What's more interesting is the fact that log of reach also has some association with the controlling variables further and thus suggests that interaction terms will need to be added for variables such as paid post and hour of the day when building the regression models.

Regression Models

Now, that a basic pattern of association between the variables has been established, let us examine the regression analysis for the stated hypothesis. There are a total of 4 linear regression models which are explained below in the sequence the variables were added to unveil the pattern of relation. All the models are adjusted for the heteroskedastic robust standard errors. The results of all the regressions are shown in the Table 2, Exhibit 8 of the Appendix.

(1) Log Interaction Vs Log Reach

$$log(Interactions) := \beta_0 + \beta_1 log(Reach)$$

The first linear model tends to present whether there is any relationship between the dependent variable and independent variable i.e. Log of Interaction and Log of Reach respectively. Based on this Log Log Model, ordinary least square estimates that if the total reach on a post is higher by 1%, the total interactions would be higher by approximately 0.66% with a 99.99% level of significance. The R-Square for this model is 0.43, which suggests that the variation is independent variable explains 43% of the variation in dependent variable.

(2) Log Interaction Vs Log Reach + Paid Post

$$log(Interactions) := \beta_0 + \beta_1 log(Reach) + \beta_2(Paid) + \beta_3 log(Reach)(Paid)$$

The Second model estimate the relationship between the two variables in the previous model along with a binary controlling variable i.e. post being paid. As per this model the estimated β coefficient of Log Total Reach suggests that 1% higher reach will result in 0.75% higher total interaction on a posts (at 99.99% level of significance) keeping everything else constant. In the same way, if a post is paid the total interaction will be higher by 203.93%. On the contrary the coefficient of the interaction term Log Reach and post being paid estimates that the total interactions on the post will be 0.23% lower. The β coefficient of paid post and the interaction has 95% level of confidence.

(3) Log Interaction Vs Log Reach + Paid Post + Time of the Day

$$log(Interaction) := \beta_0 + beta_1 log(Reach) + \beta_2(Paid) + \beta_3(Time) + \beta_4 log(Reach)(Paid) + \beta_5 log(Reach)(Time)$$

In this model, an additional confounding variable "Time(hour) of the day" is added to the regression equation in the previous model. The association between interactions and the additional controlling variable has a 99% level of confidence and states that, everything else remaining constant, if a post is uploaded one hour later in the day the total interaction will be lower by 22.42%. On the other hand, interaction term log reach and time of the day estimates if a post is uploaded one hour late, the interactions will be higher by 0.019% (level of confidence is 95%). This is slightly inconsistent with the general trends about the time in the data.

I believe this may be possible that due to a few days for instance, were holidays or weekends and the traffic on social media platforms tends to be higher in the evenings.

(4) Log Interaction Vs Log Reach + Paid Post + Time of the Day + Type of Post

$$log(Interaction) := \beta_0 + beta_1 log(Reach) + \beta_2(Paid) + \beta_3(Time) + \beta_4(Video) +$$

$$\beta_5(Photo) + \beta_6(Status) + \beta_7 log(Reach)(Paid) + \beta_8 log(Reach)(Time)$$

In this final regression, the new added confounder is the type of post (Video, Photo, Status or Link). An interaction term was not included in the model for this variable because according to my understanding the type of post has no direct impact on the reach of the post itself. For instance, the reason why people using social media marketing tend to opt for paid posts is because it provides more reach, but this is usually not consistent with the type of post. Based on this model, Interaction tends to be higher by 89.48% if the post is a video compared to other types of post and keeping everything else constant. This has significance level of 99%. Similarly, in case of the post being a photo or status, the interaction tends to be higher by 120% or 100% respectively, compared to other type of posts. These two β values have a significance level of 99.99%.

Conclusion

To conclude, the regression models created for this analysis clearly suggest that our expectations about the pattern of association between the dependent and independent variable is correct and there appears to be a positive relationship. The same is true for the confounding variables, as except a few, most of the variables tend to have a positive impact on the total interaction of a Facebook post. Moreover, when we compare models with each other, we see a consistent increase in the R Square (Model 1 = 0.43, Model 2 = 0.44, Model 3 = 0.46 and Model 4 = 0.51) and level of significance for each of the coefficients. This further indicates that our models kept on improving as we added more confounding variables. Furthermore, for our above stated hypotheses, we won't be rejecting H_0, since coefficient in all models are significantly different from zero.

We also observe that most coefficient values are also consistent with our with our expectations from that variable's impact on total interactions such as, if the post is paid, in all models, interaction would be higher by approximately 200%. The only coefficient value that was inconsistent with our expeditions form it was the interaction term for the confounding variable "Paid Post". The estimates suggests that the interaction term $\beta log(Reach)(Paid)$ has a negative impact on interactions, which in reality should be the other way around.

Our Preferred Model

$$log(Interaction) := \beta_0 + beta_1 log(Reach) + \beta_2(Paid) + \beta_3(Time) + \beta_4(Video) +$$

$$\beta_5(Photo) + \beta_6(Status) + \beta_7 log(Reach)(Paid) + \beta_8 log(Reach)(Time)$$

This is our preferred model because, alongside being the most consistent with our exceptions, it is also statistically and logically correct. Meaning, all expect one, estimated β coefficient values coincides with how they are intended to perform in the real world for a digital media strategy of a Facebook page. The R Square for this model was the highest compared to all other models i.e. 51% of the variation in interaction is explained by this model. The level of significance for β coefficients was also the highest among other models, most of the coefficients had an 99% or more confidence level and the lowest being 95%.

Overall, I would say that our analysis tends to answer the question we had with regards to this data. The only thing that one can improve in this model is adding more confounding variables that have an impact on the total interactions on a post. I believe there are more variables that would impact interactions such as "Week of the month" or "Month of the year", which were not included in this analysis. Including similar variables to the model would be helpful in explaining the hidden patterns of association that this analysis could not discover.

Appendix

Exhibit 1

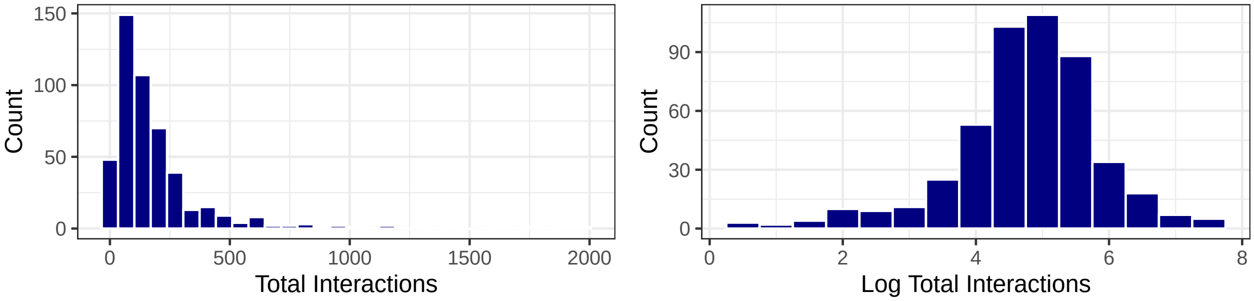

Exhibit 2

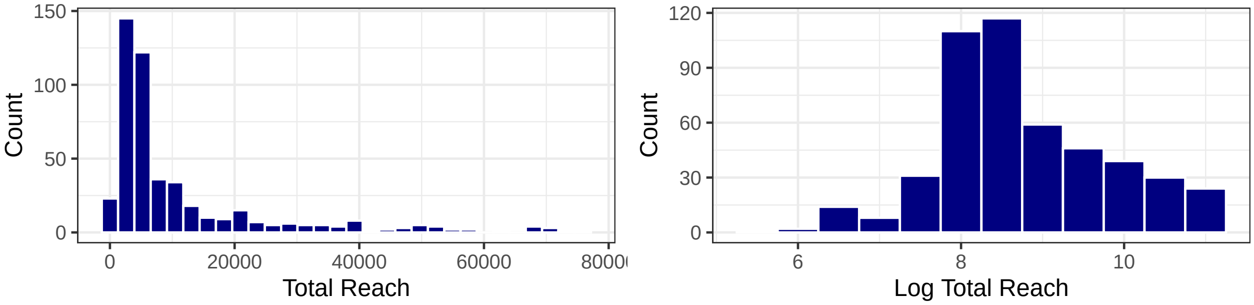

Exhibit 3

Table 1: Descriptive statistics

	mean	Median	SD	Min	Max	P05	P95
Total Interaction	194.19	122.00	240.19	2.00	1974.00	14.00	596.00
Total Reach	11 361.23	5240.00	14 562.90	238.00	76 096.00	1388.00	46 192.00
Log Total Interaction	4.76	4.80	1.10	0.69	7.59	2.64	6.39
Log Total Reach	8.75	8.56	1.06	5.47	11.24	7.24	10.74
Paid Post	0.27	0.00	0.45	0.00	1.00	0.00	1.00
Time of the Day	7.84	9.00	4.37	1.00	23.00	2.00	14.00
Type of Post: Photo	0.85	1.00	0.36	0.00	1.00	0.00	1.00
Type of Post: Status	0.09	0.00	0.29	0.00	1.00	0.00	1.00
Type of Post: Link	0.05	0.00	0.21	0.00	1.00	0.00	0.00
Type of Post: Video	0.01	0.00	0.10	0.00	1.00	0.00	0.00

Exhibit 4

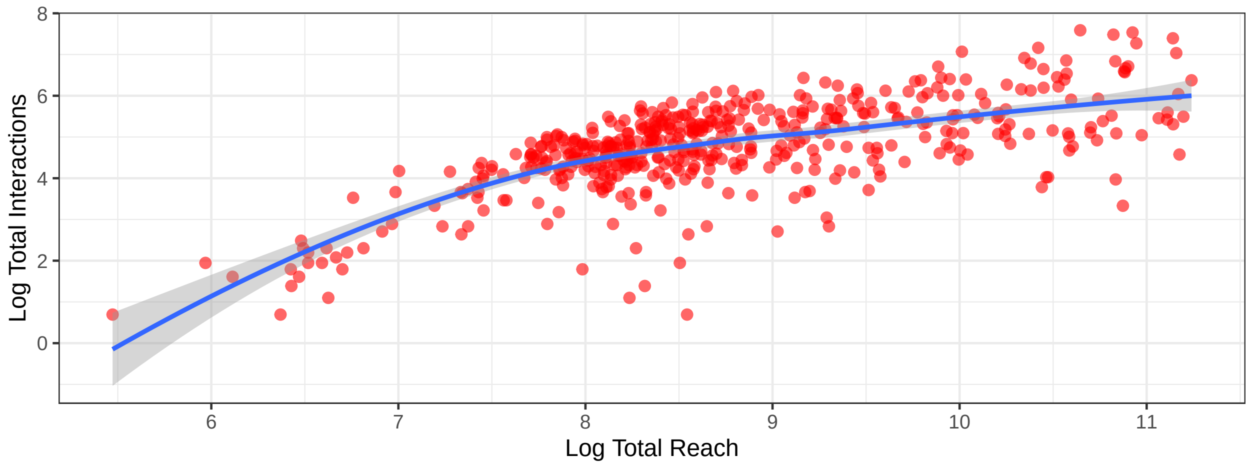

Exhibit 5

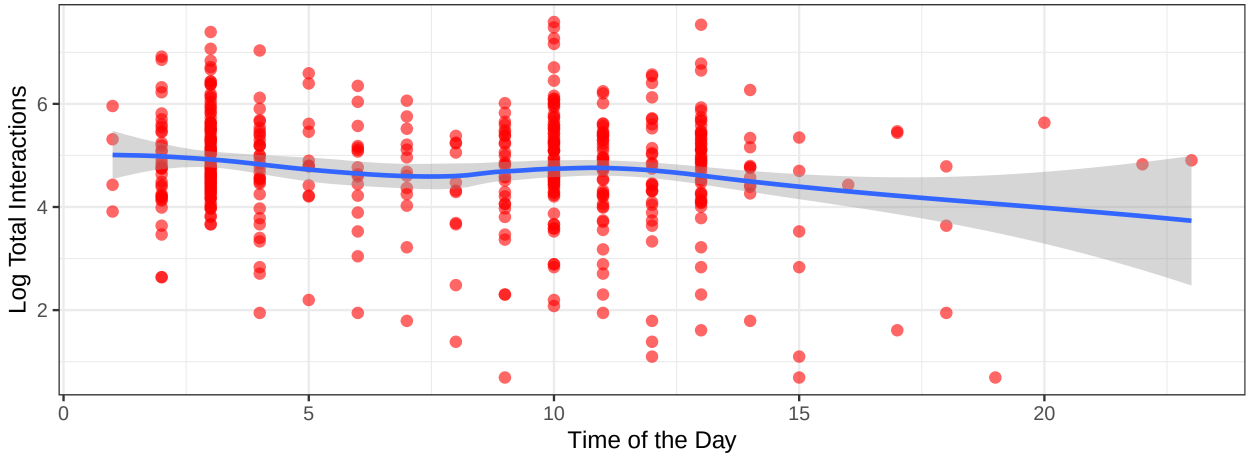

Exhibit 6

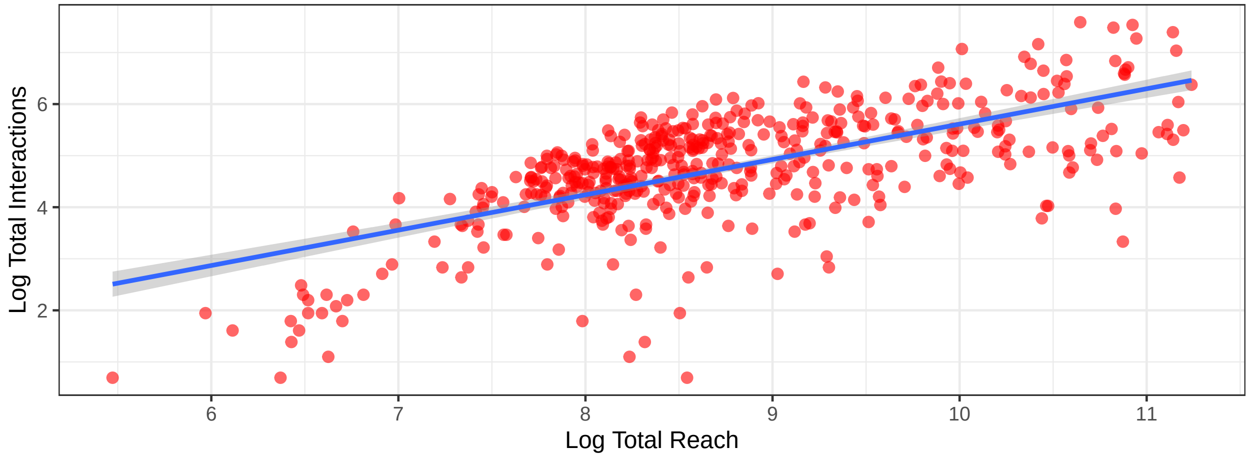

Exhibit 7

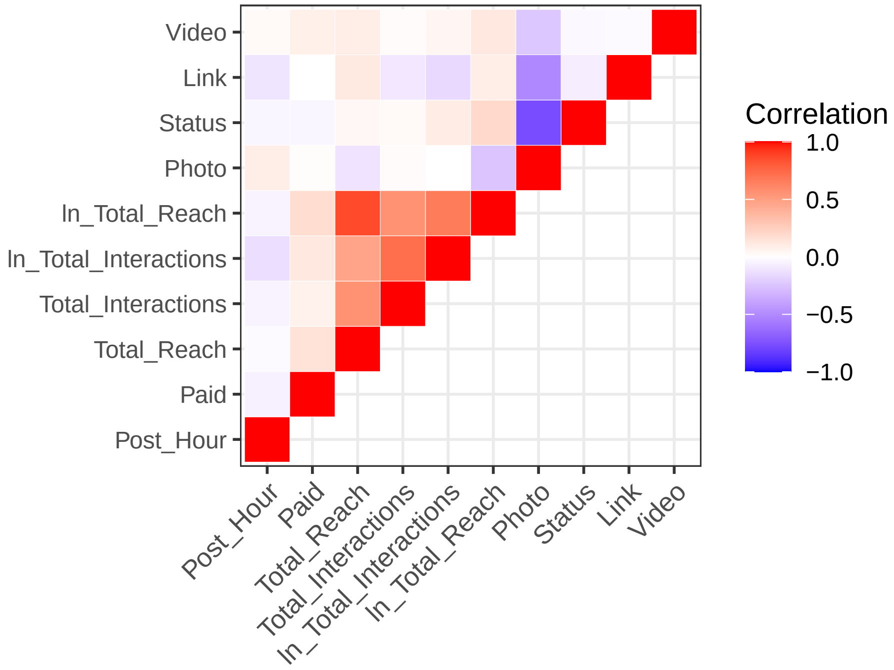

Exhibit 7

[H]

Table 2: Models to uncover relation between interactions and reach

	Model 1	Model 2	Model 3	Model 4
Intercept	-1.244**	-1.772***	0.1687	-1.670*
	(0.3856)	(0.4442)	(0.7238)	(0.7125)
Log Total Reach	0.6855***	0.7465***	0.5492***	0.6319***
	(0.0440)	(0.0512)	(0.0846)	(0.0782)
Paid		2.039*	1.856*	2.219**
		(0.8369)	(0.8102)	(0.7967)
Log Total Reach x Paid Post		-0.2270*	-0.2071*	-0.2501**
		(0.0945)	(0.0921)	(0.0910)
Time of the Day			-0.2242**	-0.1992**
			(0.0691)	(0.0648)
Log Total Reach x Time of the Day			0.0226**	0.0190*
			(0.0080)	(0.0075)
Type: Video				0.8948**
				(0.3092)
Type: Photo				1.260***
				(0.1619)
Type: Status				1.003***
				(0.1866)
Observations	481	481	481	481
R2	0.43113	0.44018	0.46136	0.51871