*Digital Marketing Analyst Case-Study – Using Data Analysis Techniques to Determine Best Advertising Practices on Facebook*

In this case study, I demonstrate my proficiency in programs like R and Excel as well as my general thought process on how I go about approaching the problem. Likewise, I utilize techniques like regression analysis to gain insight with respect to what variables drive traffic to a particular Facebook page. The data used in this exercise is publicly available, and the data is related to posts published during the 2014 calendar year on the Facebook page of a renowned cosmetics brand. The dataset contains 500 of the original 790 rows; some information was omitted due to confidentiality issues. Ultimately, I prepare the dataset by making the necessary transformations, perform a regression analysis, and interpret the results.

To begin, I start with downloading the CSV-file of the dataset, which can be accessed here: <https://archive.ics.uci.edu/ml/datasets/Facebook+metrics>. I then import the dataset into R with the following code:

---------------------------------------------------------------------------------------------------------------------

fb\_data <- read.csv("facebook data.csv")

---------------------------------------------------------------------------------------------------------------------

Upon importing the dataset, I immediately notice something unusual about the dataset from the Environment window in R. As an illustration, my dataset has 500 rows of observations; however, only 1 variable. Surely, something must be amiss here. So, I go back to examining the original dataset. I notice from text file as well as in Excel that the data is not separated by commas, rather it is separated by a semi-colon. So, I must make adjustments to my R code to import the data correctly.

---------------------------------------------------------------------------------------------------------------------

fb\_data <- read.csv("facebook data.csv", sep = ";")

---------------------------------------------------------------------------------------------------------------------

There, now my dataset appears to be imported correctly. This is confirmed by the Environment window in R, stating that I now have a data table with 500 rows of observations and 19 variables. I now run the following code to ensure that I have the proper packages activated that will be necessary for analysis.

---------------------------------------------------------------------------------------------------------------------

library(tidyverse)

library(dplyr)

---------------------------------------------------------------------------------------------------------------------

Upon loading in my data, the first thing I want to do is to examine the data table structure. It is important for me to check because I need to understand whether or not the variables within my dataset are associated with their expected data type.

---------------------------------------------------------------------------------------------------------------------

str(fb\_data)

---------------------------------------------------------------------------------------------------------------------

Upon doing so, I notice some discrepancies in the dataset. For example, the *Type* variable is defined as a character or string variable, whereas it would make far more sense for it to be a factor or as a categorical variable. Similarly, the *Post Weekday* variable is defined as an integer variable, but it doesn’t make sense to me to have a integer data type here. I determine that I will need to transform this variable into a factor data type as that is more appropriate. To address these issues, I run the code in R.

---------------------------------------------------------------------------------------------------------------

fb\_data$Type <- as.factor(fb\_data$Type)

fb\_data$Post.Weekday <- as.factor(fb\_data$Post.Weekday)

str(fb\_data)

---------------------------------------------------------------------------------------------------------------------

Using the structure function again, I can confirm that the transformations to the *Type* & *Post Weekday* variables were successful. Circling back to the business problem at hand, ultimately, I want to determine the effect certain variables have on Facebook interactions. Upon examining my newly transformed dataset in the Environment window in R, I notice that the numbers associated with *Page Total Likes* & *Lifetime Post Total* Impressions are exceptionally large.

Similarly, with respect to the business task, it is far more beneficial to determine the expected percentage increases with respect to a marginal increase in an independent variable. So, for instance, I want to share my findings in such a manner that would determine the percentage increase in traffic for a single additional comment. Given this context, I decide that log-level or log-log regression analysis technique would be the most beneficial way to tackle the problem at hand.

Moreover, I handpick particular variables from the dataset that I think would be most beneficial in providing insight with respect to the business task at hand. Ultimately, I determine that *Lifetime Post Total Impressions*, *Page Total Likes*, *Type*, *Post Weekday*, *Paid*, *Comment*, *Like*, and *Share* are the most relevant to my analysis.

To me, it is apparent that these variables would have the biggest impact on determining the effect of Facebook traffic. As an illustration, I suspect that the *Paid* variable would influence Facebook traffic because a paid post is likely to have far more outreach and interactions. Also, it is important to include the *Paid* variable in my regression model because it serves as a great control variable.

Now, given the context of business task as well as outline of the regression model I want to employ, I need to transform my data a little more. To do so, I run the following code in R:

---------------------------------------------------------------------------------------------------------------------

fb\_data$Lifetime.Post.Total.Impressions <- log(fb\_data$Lifetime.Post.Total.Impressions)

fb\_data$Page.total.likes <- log(fb\_data$Page.total.likes)

View(fb\_data)

---------------------------------------------------------------------------------------------------------------------

Here, I use the View function to make sure the transformations to *Lifetime Post Total Impressions* & *Page Total Likes* were performed correctly. Upon viewing the data table, I confirm that these transformations were carried out correctly. Now, I need to switch focus towards building my model. As aforementioned, I already determined the variables most relevant to the analysis task I want to perform include *Lifetime Post Total Impressions*, *Page Total Likes*, *Type*, *Post Weekday*, *Paid*, *Comment*, *Like*, and *Share*. So, I want to get rid of unnecessary information irrelevant to the business task at hand. Similarly, I want to be working in a data table that just includes the information I need. To do so, I run the following code in R:

---------------------------------------------------------------------------------------------------------------------

fb\_data1 <- fb\_data %>%

select(Lifetime.Post.Total.Impressions,

Page.total.likes,

Type,

Post.Weekday,

Paid,

comment,

like,

share)

---------------------------------------------------------------------------------------------------------------------

Using this code, I created a new data table titled fb\_data1 including only the information relevant to my regression model. Whereas the initial data table titled fb\_data has 500 observations & 19 variables, the newly created data table titled fb\_data1 has 500 observations & 8 variables. Moreover, the variables in fb\_data1 have already been transformed to the correct data type from previous code, so that is no longer a concern for me.

Now, I can finally focus on running my regression model. To do so, I execute the following code in R:

---------------------------------------------------------------------------------------------------------------------

model <- lm(Lifetime.Post.Total.Impressions ~ .,

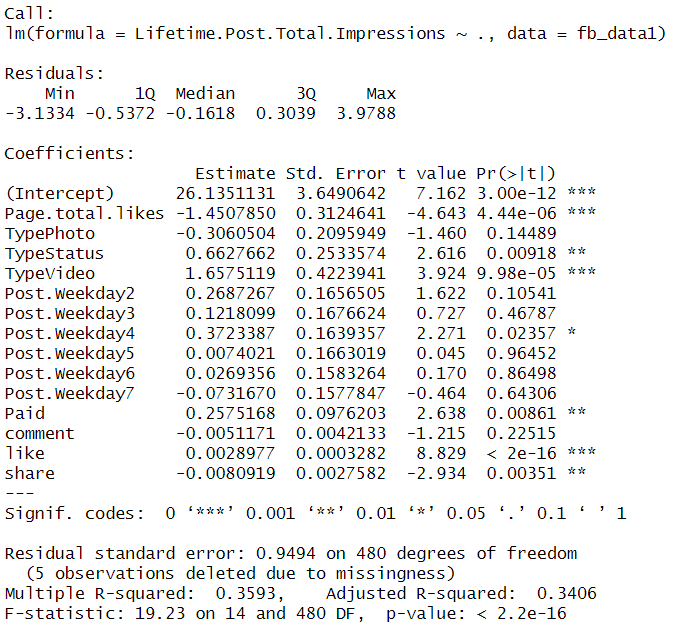
data = fb\_data1)

summary(model)

---------------------------------------------------------------------------------------------------------------------

In this first part of the code, I create a data frame titled model. This model will include all of the attributes associated with my regression technique. Similarly, the lm function indicates that this is linear regression model, and the other part of the code indicates that *Lifetime Post Total Impressions* is my dependent variable in the model, and all other variables from the data frame titled fb\_data1 serve as independent variables. Lastly, the summary function gives me a summary of the regression model I just executed, including information like the coefficient estimates, standard errors, and t-values associated with all the independent variables.

---------------------------------------------------------------------------------------------------------------------



Asdfasdfasdf

When interpreting the table above, the first thing I want to do is to think about the sign of the coefficients. While this relies on experience and intuition, I ultimately want to ask myself if the sign makes sense. As an illustration, in a hypothetical linear model related to *Income*, I wouldn’t expect to see a negative sign on the coefficient of *Years of Education*. Logically, it makes little sense because years of education should increase your human capital and skills; therefore, education should have a positive effect when determining an individual’s income.

Now, back to our data, I want to view the coefficient signs in a similar framework. With respect to *Page Total Likes*, I clearly notice that this has a negative coefficient. However, intuitively, this does not make much sense. Logically speaking, an increase in page total likes should help increase Facebook traffic. However, here, our data and regression model is telling us the opposite essentially. Typically, I would expect that this coefficient should display a positive relationship.

Moving on, let us interpret the *Type* coefficients. Here, I can clearly see that the *Type: Photo* has a negative coefficient, which is odd. However, when looking at the table, I can immediately tell this isn’t statistically significant. The **Pr(>|t|)** column represents the p-value associated with the value in the t-value column. If the p-value is less than a certain significance level, typically .05 in a linear regression model, then the predictor variable is said to have a statistically significant relationship with the response variable in the model.

So, back to our original example in the summary table, the **Pr(>|t|)** associated with *Type:Photo* = 0.14489. And, 0.14489 > .05. Therefore, I conclude that the unusual negative relationship between *Type:Photo* & *Lifetime Post Total Impressions* can be ignored.

Let us review some other variables in our model further. Whereas *Type:Photo* had a negative coefficient, *Type:Status* & *Type:Video* have the expected positive relationship. Likewise, I can tell from the summary statistics table of my data frame titled model that this relationship is statistically significant. So, I can interpret this as follows: a Facebook post made including a status update or video is more likely to receive impressions and drive traffic to the Facebook page of the renowned cosmetics brand. In fact, a post of a video is the most likely to generate traffic, and I know this to be true considering the positive coefficient associated with *Type:Video* of 1.65 is greater than the positive coefficient associated with *Type Status* of .66.

Now, let’s shift our focus towards the *Post Weekday* variables. Here, I notice something interesting. I notice that *Post Weekday 4* is the only statistically significant relationship. Similarly, it has the largest positive coefficient. Next, I must ask myself, once again, does this make sense.

My model & the data are telling me that Thursday is the best day to post in order to drive traffic towards the Facebook page. Now, I must digest this information from an intuitive perspective. Upon reflection, I confirm that this relationship does make sense. Thursday is the middle of the week, and people are often browsing on social media to distract from work or figure out their weekend plans.

Exploring the *Paid* variable, the coefficient is positive as expected. And, the *Paid* variable is statistically significant. However, I realize that coefficient is pretty darn small. This seems to suggest that it doesn’t matter much whether a post is paid or unpaid in nature with respect to generating traffic to a Facebook page.

Lastly, let’s take a look at our final three variables: *Comment*, *Like*, and *Share*. An interesting finding from the summary statistics is that a comment is not statistically significant in determining traffic to a Facebook page. Conversely, a like is statistically significant and does drive traffic to a Facebook page in a meaningful way. Finally, the summary statistics associated with the variable *Share* appear a bit strange.

As an illustration, the coefficient on the share variable is negative. Moreover, it is statistically significant. How can this be? Logically speaking, someone sharing a post on Facebook should make it visible to their network, which should organically drive traffic to the original Facebook page. This is counter-intuitive. Is the story the data is communicating to me a lie?

Well, no. I just needed to consider how Facebook works further at an individual level, which helps me rationalize why this value is negative. As an illustration, when you share a post, all the engagement from the post is directed towards you as the individual user and not the original page you are promoting. So, although counter-intuitive, the relationship is perfectly rationale. Indeed, the relationship is statistically significant, but the coefficient is so small that the relationship between sharing a page and lifetime post impressions is practically negligible. Perhaps, this relationship would become definitely positive given more observations; however, it is uncertain, and more research should be done on this topic.

---------------------------------------------------------------------------------------------------------------------

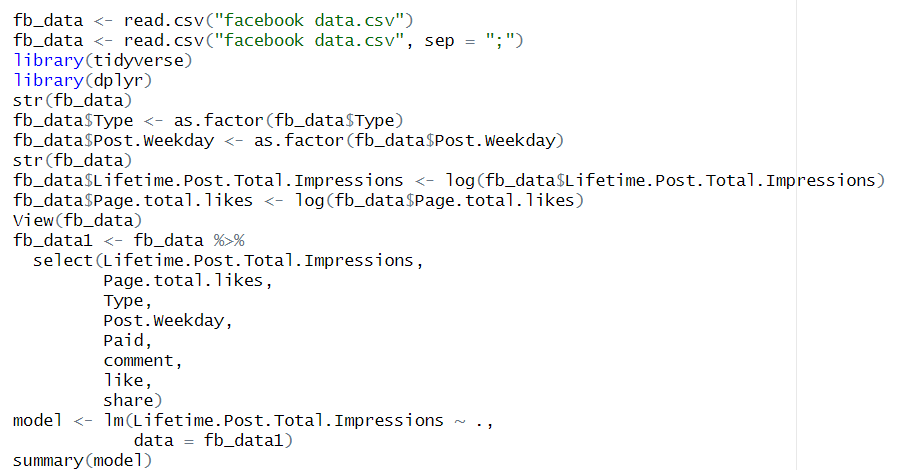
Going back full circle, the original business task was to figure out what drives traffic to the Facebook page of renowned cosmetics brand. Using page traffic data from Facebook, I imported this data in R. Likewise, I manipulated and transformed the data pertinent to my analysis. I then ran a linear regression model to gain insight on which factors are most important to increase traffic. From this exercise, I present the findings of my analysis to the key stakeholders of the company.

**Key Findings:**

* *There is a negligible difference between paid & unpaid content with respect to traffic generated to the Facebook page. Therefore, advertising dollars are best spent elsewhere on other platforms or through entirely different mediums.*
* *Advertising efforts on Facebook should focus on generating likes through a video format on Thursday.*
* *Additional market research with a larger sample size would be helpful to inform future advertisement strategies through Facebook.*

---------------------------------------------------------------------------------------------------------------------

**Full Code in R:**



**Information About the Data:**

The dataset can be accessed through the following link: <https://archive.ics.uci.edu/ml/datasets/Facebook+metrics>. The dataset is publicly available for research purposes. The details are described in (Moro et al., 2016).  
Please include this citation if you plan to use this database:  
(Moro et al., 2016) Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.