HW week 11

w203: Statistics for Data Science

Group 2: Jeremy Lan, Taehun Kim, Nicolas Loffreda

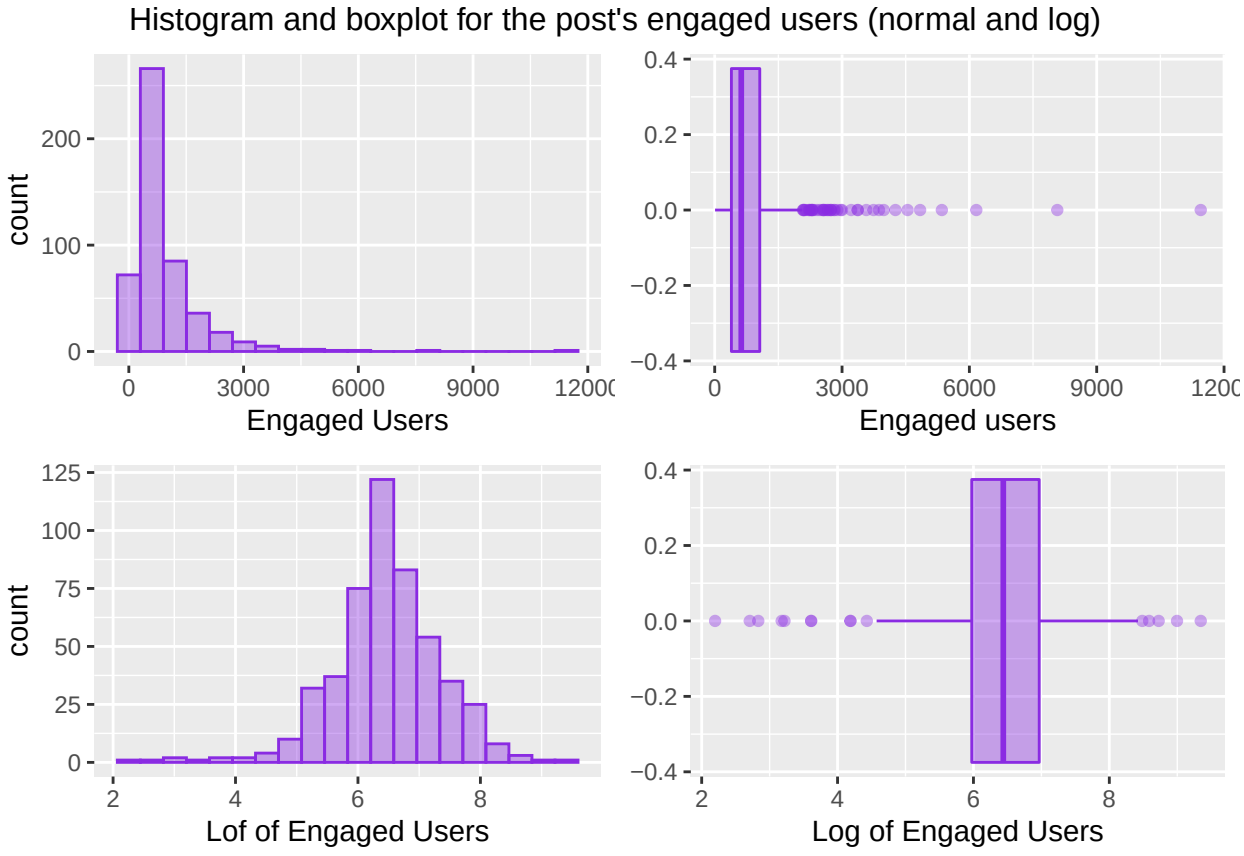
2. Data

The dataset we will be using for this analysis is a subset of that collected by Moro et al. (2016). The dataset contains a representative sample of 500 Facebook posts from a worldwide renowned cosmetic brand, collected between January 1st and December 31st of 2014. By the time the data was collected, Facebook was the most used social website, with roughly 1.28 billion monthly active users (Insights 2014).

Each observation from the dataset represents a post from this company, for which a variety of features have been collected.

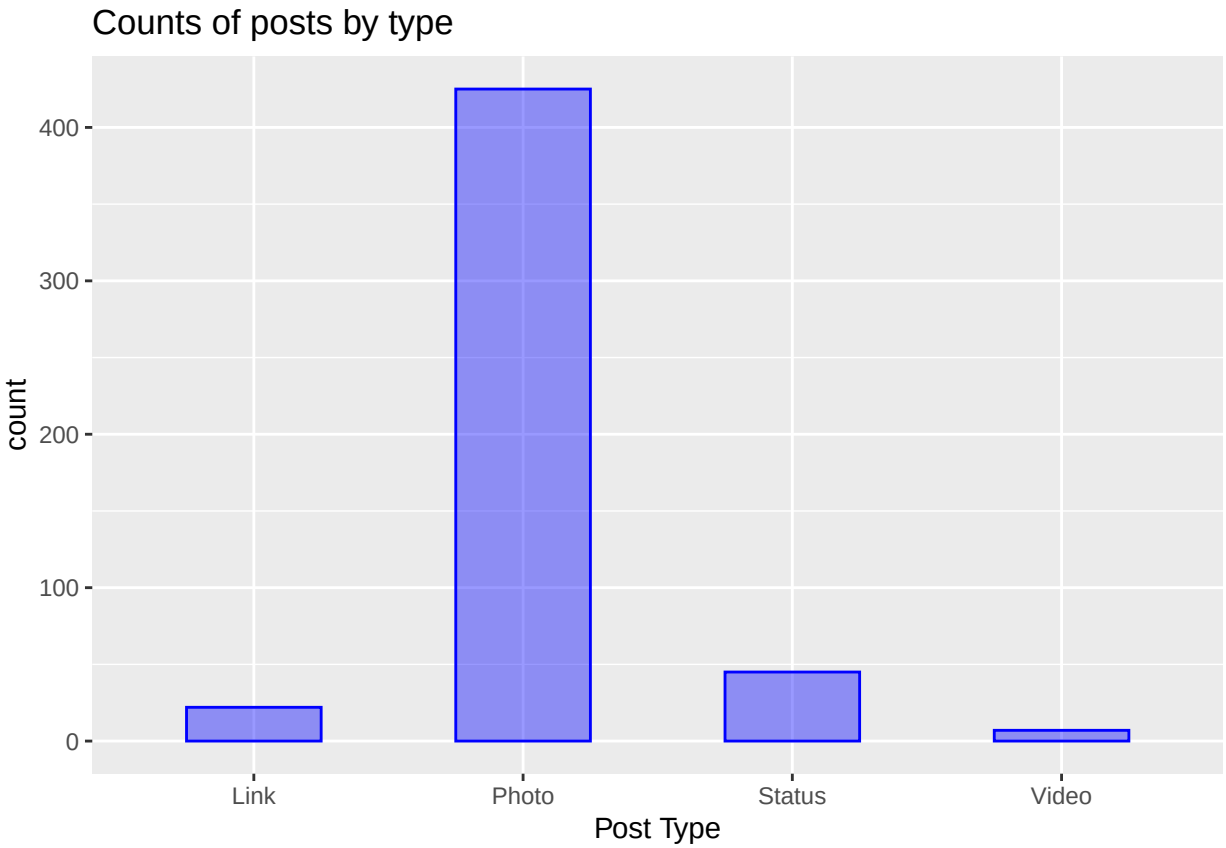
2.1 Engaged users

The outcome variable will be the number of unique *engaged users* the post had through its lifetime. An engaged user is defined as someone who clicked in the post. Looking into this variable, we can see that it is fairly skewed to the right. To make the variable easier to work with, we will be applying a log transformation:



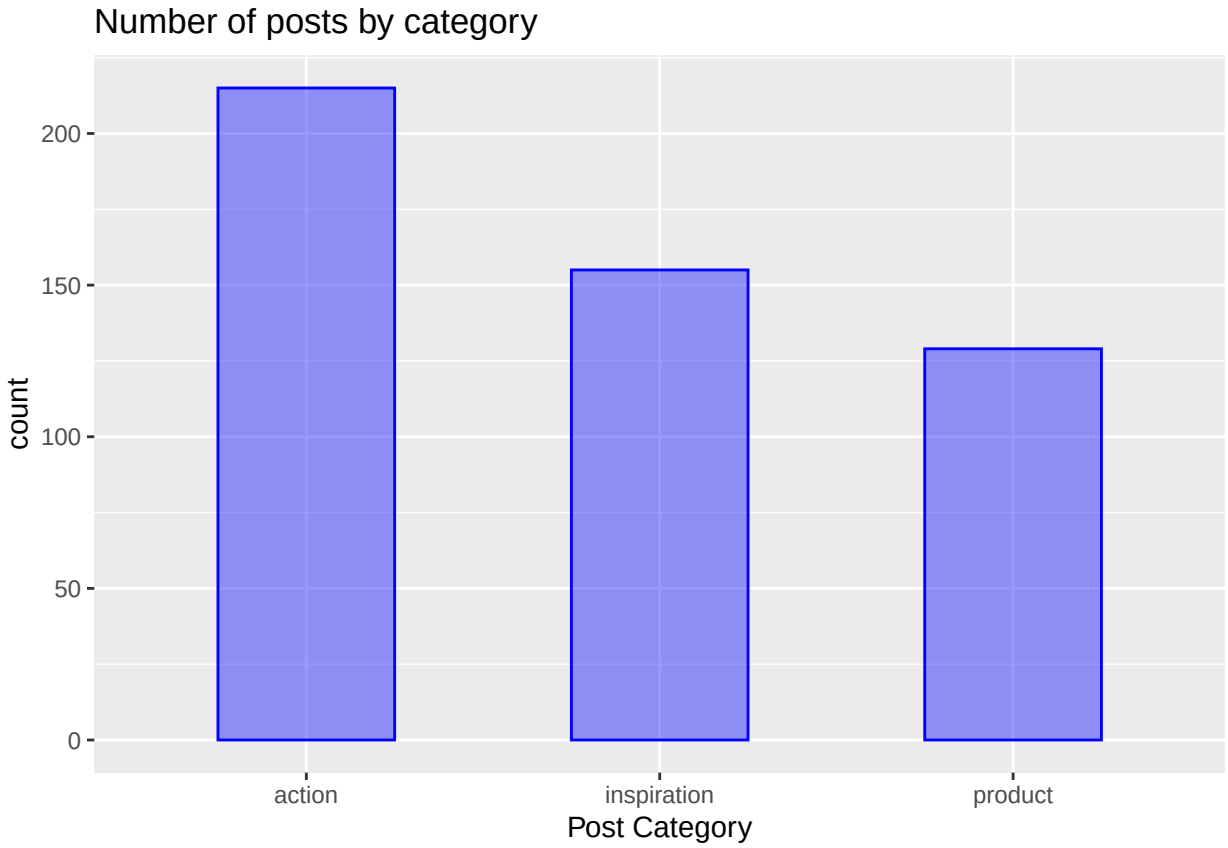
2.2 Category and Type

The main variables we want to measure the impact on engaged users are the `type` and `category` of the post. The `type` is categorized in Photo, Video, Link or Status, and it represents what kind of content the post contained. We can see that most of the posts published were photos:



On the other hand, the category describes how the content of the post was displayed to the user. There were 3 distinct categories the dataset differentiates: - Action: Special offers and contests - Product: Direct advertisement or explicit brand content - Inspiration: Non-explicit brand related content

The number of posts published of each category are as follows:



2.3 Covariates

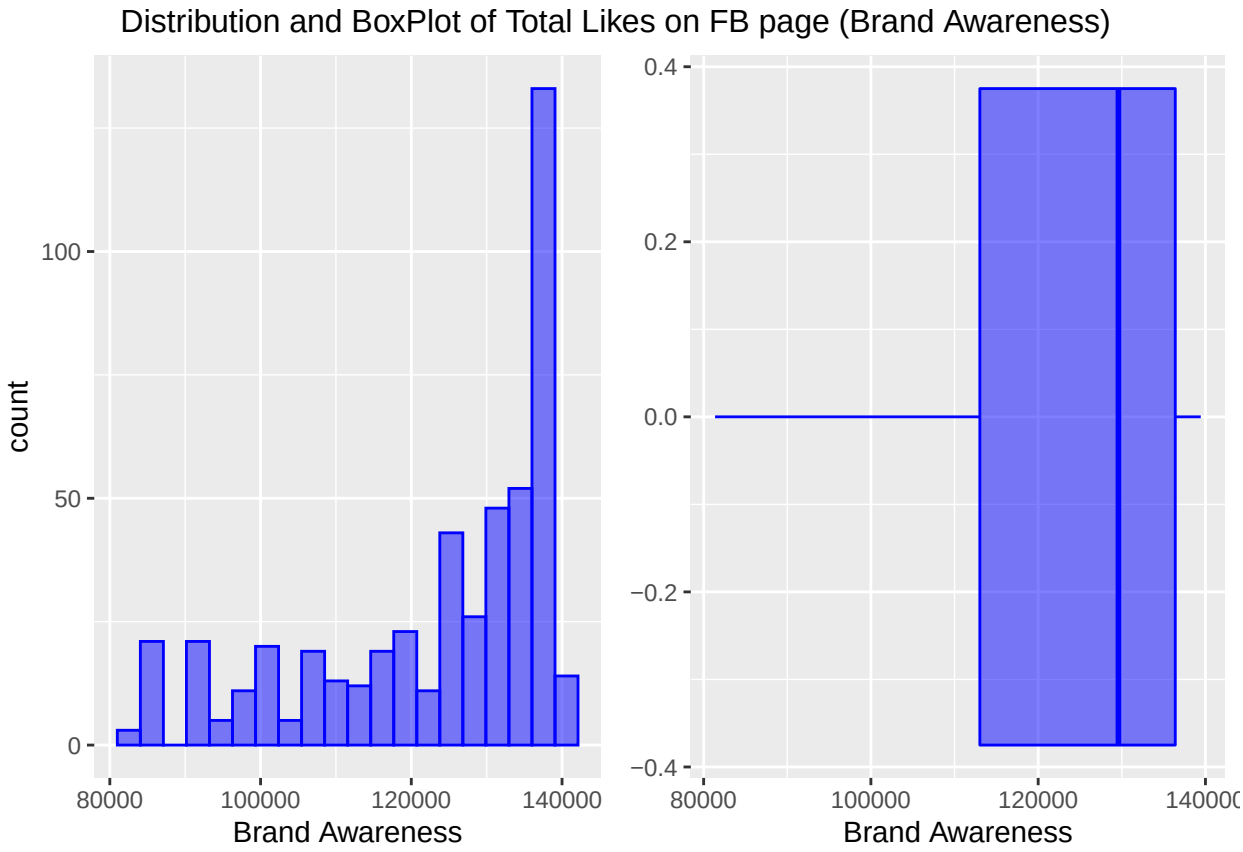
2.3.1 Paid Among the covariates we will be including in the model is paid advertising. The variable `paid` will be encoded as a dummy variable to indicate whether the post had any paid media associated with it or not. We can see that 28% of all the posts had some kind of paid media support:

Paid media support

Media Support	Number of posts
No Paid support	360
Paid support	139
Total	499

Something to notice as well is that `paid` has a missing value. Given the large number of samples still remaining we removed that missing value leaving a total of 499 observations.

2.3.2 Brand Awareness Another important control variable will be Brand Awareness. This variable represents how much users are aware or the brand. As it is a difficult concept to measure, we will be accounting for this as the number of likes the Facebook site of the company had at the time that the post was published:



The variable is left skewed and although different transformations were applied to it, none of them helped to reduce the skeweness. For this, the variable will be included as is.

2.3.3 Period of day and Day of the week The last variables we will be including as control are the period of the day and day of the week. In particular, we will distinguish 4 periods of the day, overnight, morning, afternoon and evening. The first going from 12am to 6am, the second one from 6am to 12pm, then 12pm to 6pm, and 6pm to 12am.

On the other hand, the days of the week will be divided into weekdays and weekends.