Data & Modeling section for Lab2

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.

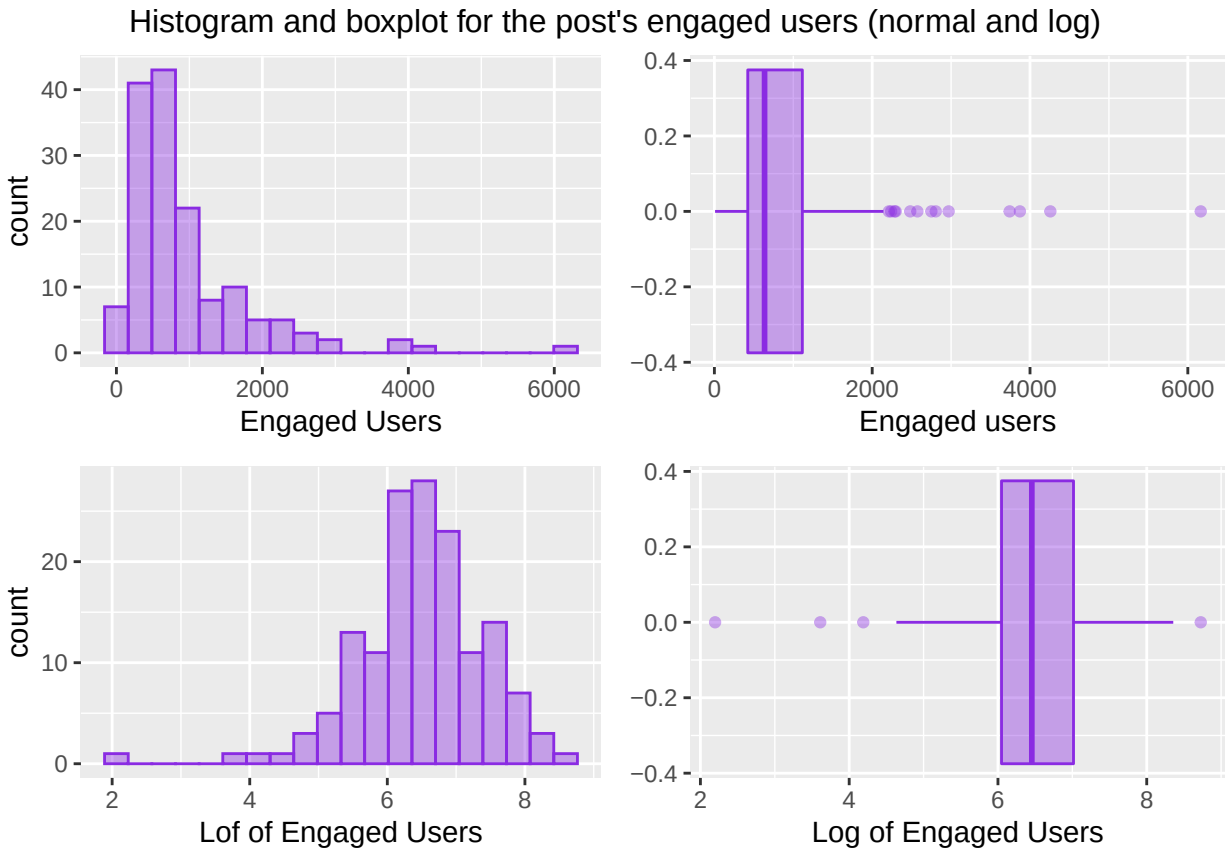
Given the large sample size, we will use a randomized sub-sample of 150 observations for exploration purposes and the remaining 350 for running the models. The only anomaly in the variables of interest is one missing value for `paid` variable, which we will be removing from the analysis, leaving us with a total of 499 observations.

Randomized sub-samples

Split	
Exploration	150
Test	349
Total	499

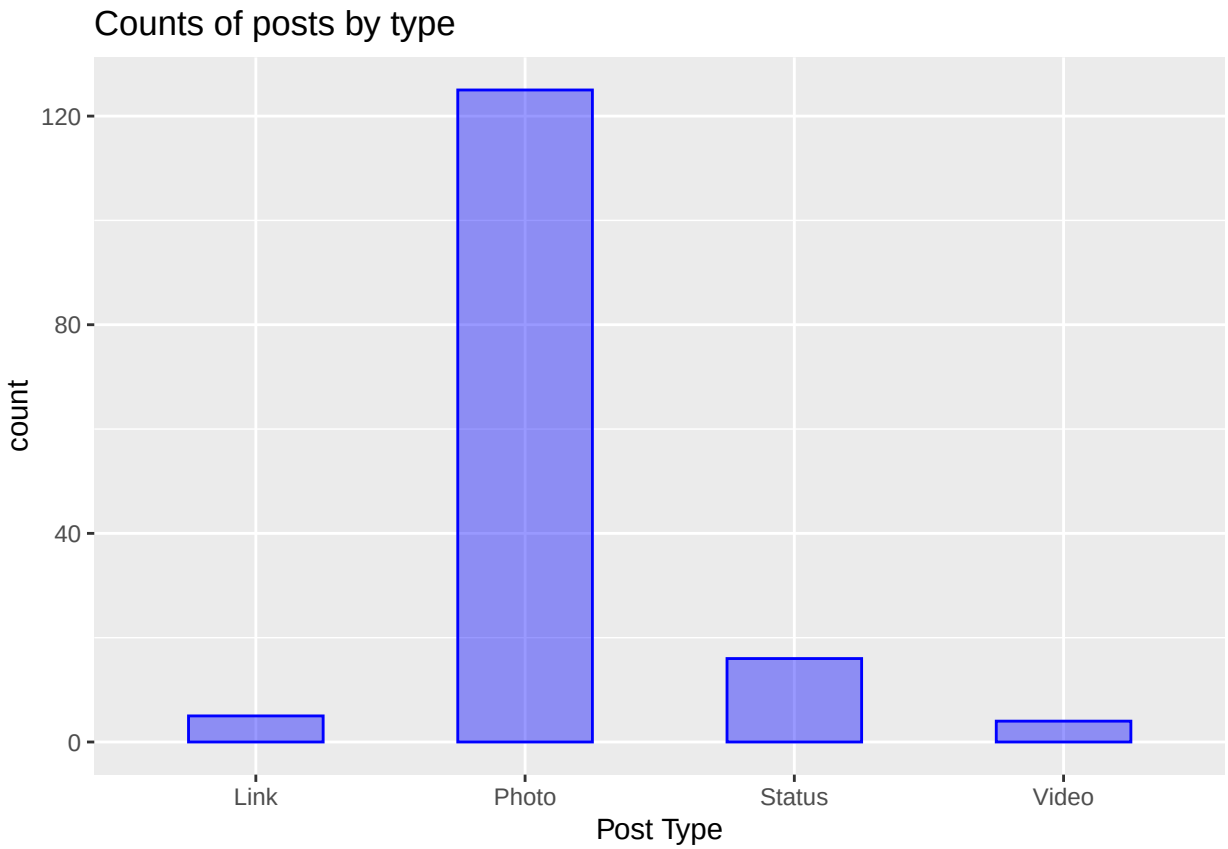
2.1 Engaged users

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:



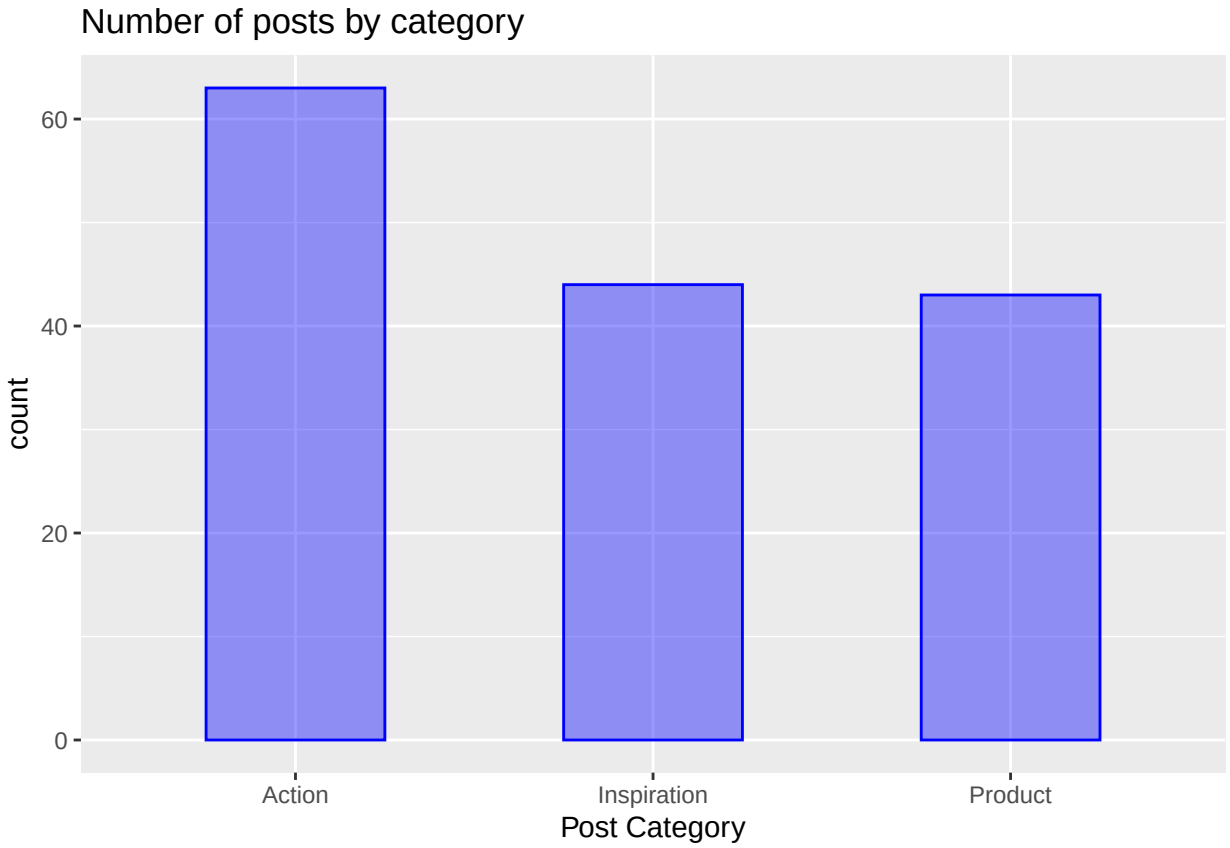
2.2 Category and Type

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:



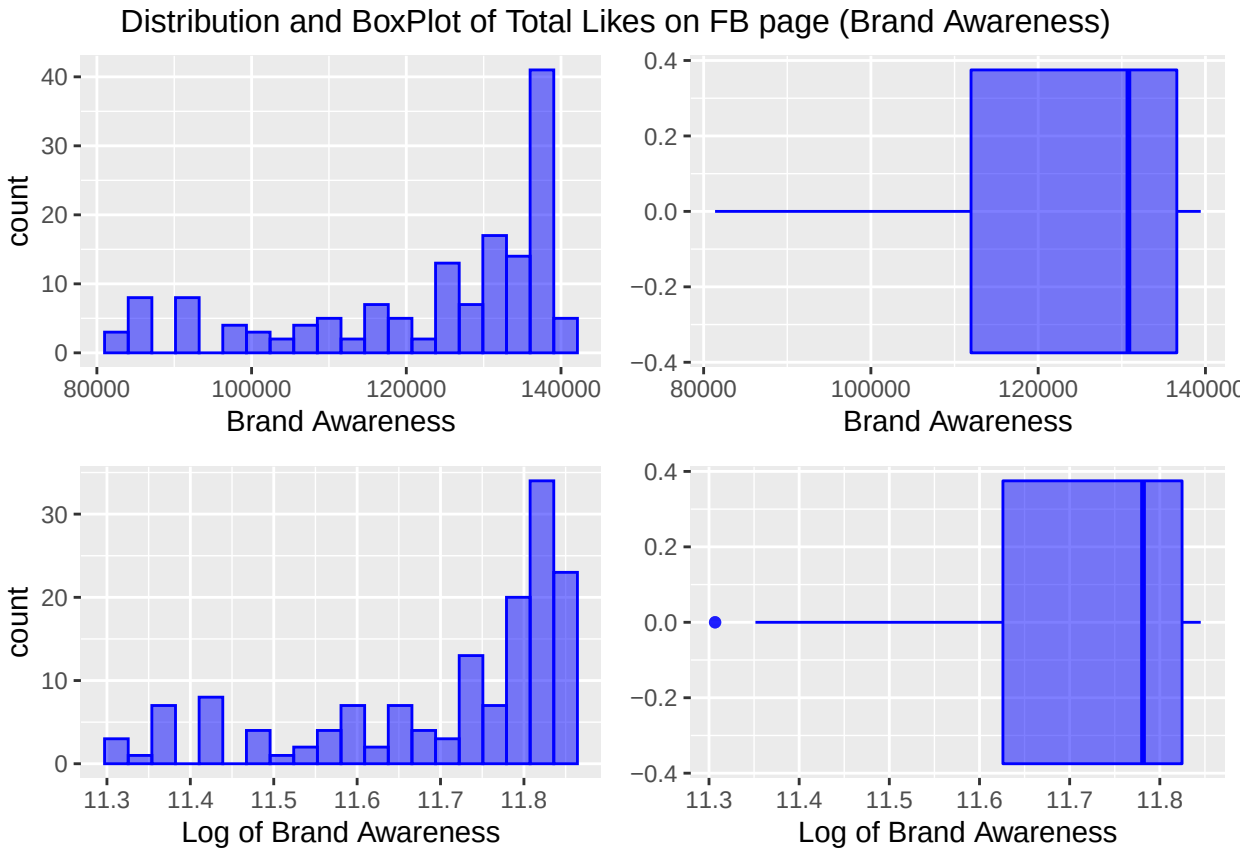
2.3 Covariates

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 in the exploratory dataset, ~32% of all the posts had some kind of paid media support:

Paid media support

Media Support	Number of posts
No Paid support	114
Paid support	36
Total	150

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:

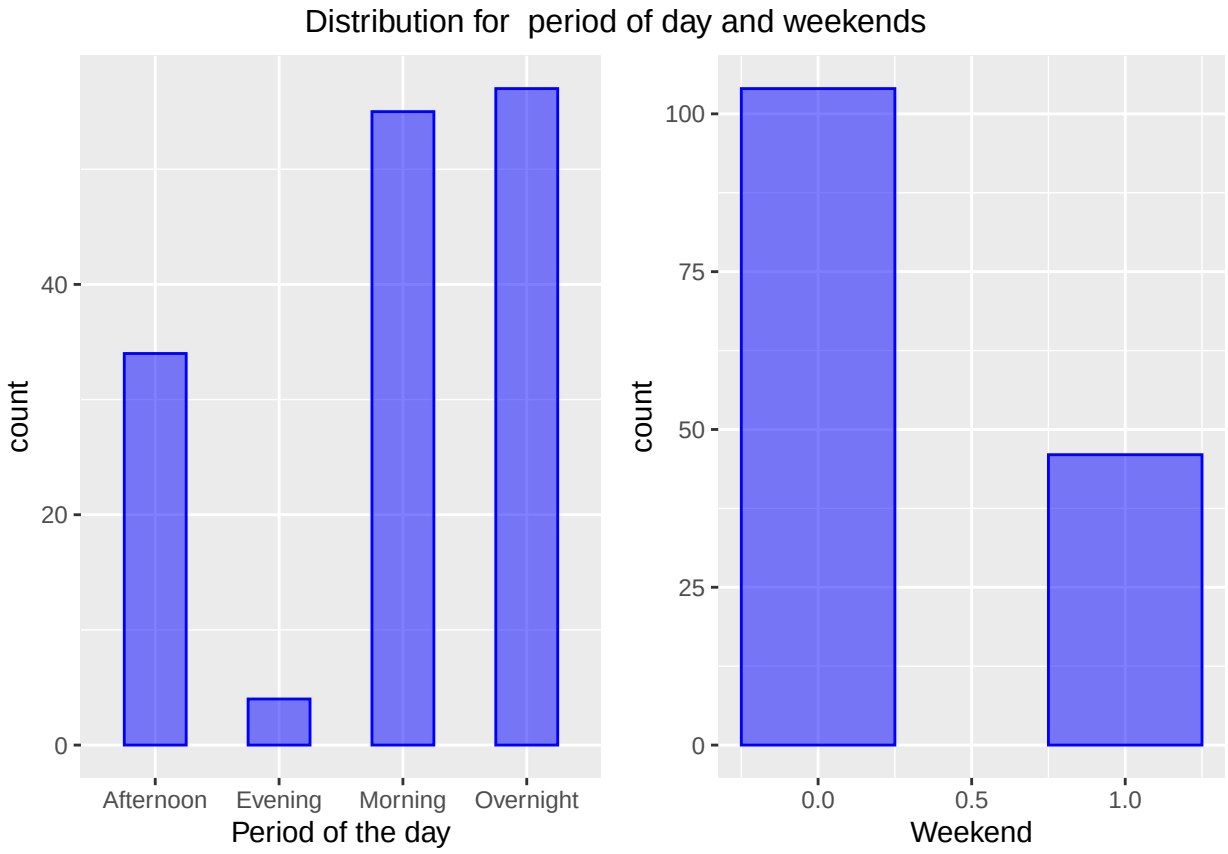


The variable is left skewed and although different transformations were applied to it, none of them helped to reduce the skeweness as it can be seen from the log transformed in the graphs. We will still be applying a log transformation to this variable given that it is easier to interpret and to measure it in a similar scale than the outcome variable.

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 the day of the week the post was published to account for the differences that may exist on user activity at different times and days.

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. This will be encoded in a dummy set to 1 if the period is a weekend.



3. Model

3.1 Base Model

As explained in the data section, we will be applying a log transformation to the outcome variable, engaged users. The base model will only include type and category as main explanatory variables:

$$\log(\widehat{engaged\_users}) = \beta_0 + \beta_1\ type + \beta_2\ category$$

3.2 Adding Covariates

With the base model established, we will be including as control variables paid media efforts, brand awareness, day of the week and period of the day. All these as described on the data section:

$$\begin{aligned}\log(\widehat{engaged\_users}) = \beta_0 + \beta_1\ type + \beta_2\ category \\ + \beta_3\ paid + \beta_4\ \log(brand\_awareness) + \beta_5\ day\_of\_week + \beta_6\ period\_of\_day\end{aligned}$$

3.3 Adding Interaction term

As a next model, we will be including an interaction term to account to see how the different types behave when paired with the different categories:

$$\log(\widehat{engaged\_users}) = \beta_0 + \beta_1\ type + \beta_2\ category + \beta_3\ paid + \beta_4\ \log(brand\_awareness) + \beta_5\ day\_of\_week + \beta_6\ period\_of\_day + \beta_7\ type * category$$

3.4 Standard Errors

We understand that certain dependencies may exist among the posts given that they are all from the same company. This means that the people that know the brand and interact with the social site and posts may be similar on the different posts.

Because of this reason, we will be using *robust clustered standard errors* to adjust the significance for any independence that may exist.

4. Results

Results of the models described are shown in the table below:

Table 1:

	Dependent variable:		
	log(Engaged Users)		
	(1)	(2)	(3)
Type - Photo	0.944***	1.003***	0.997***
	(0.259)	(0.271)	(0.291)
Type - Status	1.822***	2.035***	2.414***
	(0.329)	(0.339)	(0.596)
Type - Video	1.696***	1.852***	1.847***
	(0.441)	(0.370)	(0.382)
Category - Inspiration	0.139	0.024	0.002
	(0.101)	(0.099)	(0.317)
Category - Product	−0.007	0.056	−0.290
	(0.119)	(0.116)	(0.568)
Paid Media		0.240***	0.231***
		(0.088)	(0.088)
Brand Awareness		−1.809***	−1.776***
		(0.340)	(0.343)
Period - Evening		−0.508	−0.503
		(0.535)	(0.532)
Period - Morning		−0.155	−0.153
		(0.117)	(0.118)
Period - Overnight		0.039	0.041
		(0.115)	(0.114)

Weekend		−0.118	−0.125
		(0.096)	(0.096)
Interaction - Photo:Inspiration			0.044
			(0.321)
Interaction - Status:Inspiration			−1.204
			(1.294)
Interaction - Video:Inspiration			
Interaction - Photo:Product			0.335
			(0.576)
Interaction - Status:Product			
Interaction - Video:Product			
Constant	5.431***	26.592***	26.208***
	(0.249)	(4.004)	(4.059)
Observations	349	349	349
R^2	0.147	0.259	0.265
Adjusted R^2	0.135	0.234	0.235
Residual Std. Error	0.815 (df = 343)	0.766 (df = 337)	0.766 (df = 334)
Note:			*p<0.1; **p<0.05; ***p<0.01

All the different types of posts are significant to an α level of 0.05 across all models, with Link being the omitted type. From the magnitude of the coefficients we can see that Video, Status and Photo perform much better than Link, with Status yielding the highest increase on engaged users. All these coefficients are quite stable across models as well. As for the categories, none of them are significant at a 0.05 α level. We also see that the Product coefficient alternates sign in different models, implying the possibility of high correlation among different variables.

The second model includes the covariates paid, period of the day, weekday and brand awareness. We see that the results are similar to the base model. Some of the covariates come out as highly significant such as paid and brand awareness. Surprisingly, the coefficient of Brand Awareness is negative, implying an inverse directionality in the relationship between engaged users and brand awareness, which seems counter intuitive. At the same time, posting on weekends doesn't seem to be significant in any of the models, nor are the day periods.

The Wald Test between the base and the covariate models yields a a very low p-value, meaning that some of the covariates are helping to explain the variability of the engaged users. This phenomenon can also be appreciated in the coefficient of determination (i.e. R^2), as it jumps from just above 14% in the first model, to 24% after the addition of these covariates.

```
## Wald test
##
## Model 1: log(lifetime_engaged_users) ~ 1 + type + category_str
```

```
## Model 2: log(lifetime_engaged_users) ~ 1 + type + category_str + paid +
##     log(page_ttl_likes) + period_of_day + weekend
##   Res.Df Df      F    Pr(>F)
## 1    343
## 2    337  6 9.0198 3.762e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When looking at the model with the interaction term and covariates, we see similar results than the previous one, but with a few modifications.

Examining the interaction term, we notice that none of the terms are significant, which is quite surprising. Some of the coefficients are missing, implying high correlation between some of the interaction terms.

Last, we examine the Wald Test between the model with covariates and the interaction model, and the high p-value indicates that the interaction term doesn't add explanatory power to the model. This can also be appreciated by the fact that the R^2 doesn't increase significantly from one model to another.

```
## Wald test
##
## Model 1: log(lifetime_engaged_users) ~ 1 + type + category_str + paid +
##     log(page_ttl_likes) + period_of_day + weekend
## Model 2: log(lifetime_engaged_users) ~ 1 + type * category_str + paid +
##     log(page_ttl_likes) + period_of_day + weekend
##   Res.Df Df      F Pr(>F)
## 1    337
## 2    334  3 0.3464 0.7918
```