Lab_deliverable_2

Yihang Duanmu, Minjun Kim, Annelise Schreiber

2025-10-02

Part 1

The three questions of interest remain the same for us, they are:

1. Does the date and time in which a post is made impact the engagement that said post receives?
2. Does the fact of a post being paid/unpaid impact the average reach of the post significantly?
3. Does the cosmetics brand gain significant exposure over the course of the year, from first to latest post analyzed in 2014?

Part 2

First, we standardize and clean the data.

All the post_weekday, post_hour, and post_month data are turned to integers with non-numeric text being turned into NA. Then, wday_fac maps weekday codes from 1-7 to "Mon"–"Sun"; hour_fac turns hours into an ordered 0–23 factor; and month_fac maps 1–12 to "Jan"–"Dec" (or falls back to whatever strings exist).

Then we build the analysis tibble dat—it picks the response variables (post consumers) and the cleaned timing predictors (wday, hour, month), adds the promotion flag (paid or unpaid), and creates a simple time-order index (idx). The summary(dat$consumers) line quickly checks the outcome's distribution; the output shows a right-skewed variable with large upper outliers, which motivates options like log1p(consumers) or robust SEs in later models.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0   332.5   551.5   798.8   955.5 11328.0
```

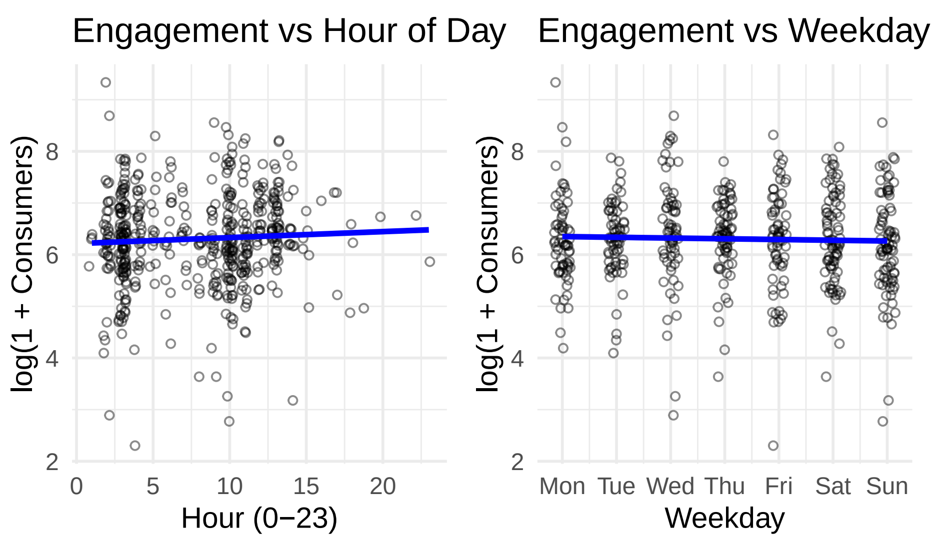
Question 1: Does the date and time in which a post is made impact the engagement that said post receives?

We used the posted hours and posted weekdates as two seperate covariates. And we used log1p(consumers) as the response variable because engagement counts are extremely right-skewed with a few huge outliers, which compresses most points near zero on the raw scale and can overly influence the fit. The following plots show the distribution of the data points.

```r
# Hour plot
plot_df <- dat |> mutate(hour_num = as.numeric(as.character(hour)))
p1 <- ggplot(plot_df, aes(x = hour_num, y = log1p(consumers))) +
  geom_point(shape = 1, alpha = 0.45, size = 1.2,
             position = position_jitter(width = 0.25, height = 0)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Engagement vs Hour of Day",
       x = "Hour (0-23)", y = "log(1 + Consumers)") +
  theme_minimal()

# Weekday plot
wk_map <- setNames(1:7, c("Mon","Tue","Wed","Thu","Fri","Sat","Sun"))
plot_df2 <- dat |> mutate(wday_num = unname(wk_map[as.character(wday)]))
p2 <- ggplot(plot_df2, aes(x = wday_num, y = log1p(consumers))) +
  geom_point(shape = 1, alpha = 0.45, size = 1.2,
             position = position_jitter(width = 0.15, height = 0)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_x_continuous(breaks = 1:7, labels = names(wk_map)) +
  labs(title = "Engagement vs Weekday",
       x = "Weekday", y = "log(1 + Consumers)") +
  theme_minimal()

p1 + p2
```



The point plot of log1p(consumers) vs hour shows a very slight upward slope, indicating a weak tendency for engagement to be higher at later hours.

The point plot of log1p(consumers) by weekday shows an almost flat best-fit line, with similar clouds of points for Mon–Sun.

The two OLS outputs are similar so we take the OLS output for fitting post consumers to post hour as an example:

```r
# OLS: engagement vs hour
model_hour <- lm(log1p(consumers) ~ hour_num, data = plot_df)
summary(model_hour)
```

```
##
## Call:
## lm(formula = log1p(consumers) ~ hour_num, data = plot_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9584 -0.4873  0.0022  0.5826  3.0973
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.214732   0.082564  75.271   <2e-16 ***
## hour_num    0.011562   0.009202   1.256     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.898 on 498 degrees of freedom
## Multiple R-squared:  0.00316,    Adjusted R-squared:  0.001158
## F-statistic: 1.579 on 1 and 498 DF,  p-value: 0.2095
```

The result shows slight linear relationship between post consumers and post date and time. However, the p-value shows that the results are not significant, meaning the model can be improved.

Question 2: Does the fact of a post being paid/unpaid impact the average reach of the post significantly?

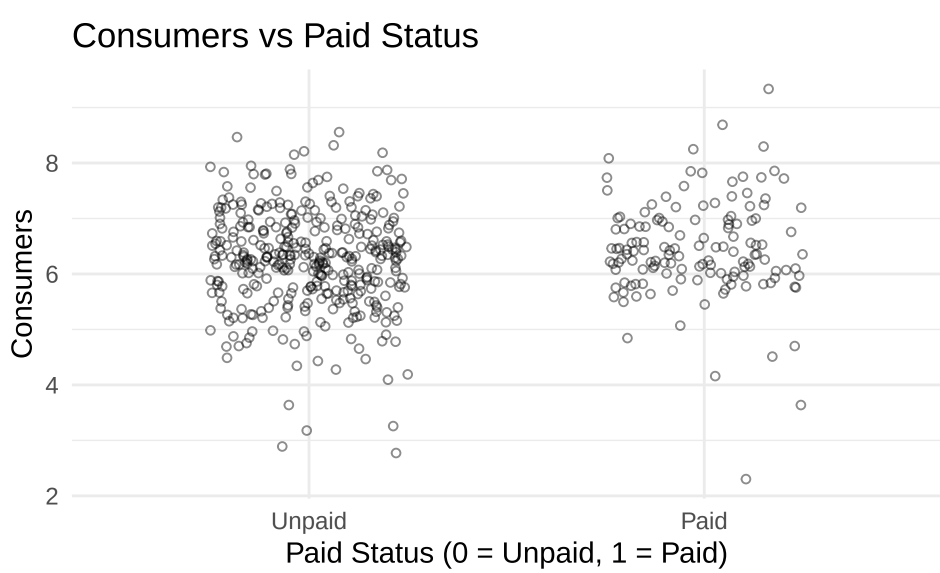
To analyze whether a post being paid or unpaid impacts reach, we have:

Response variable (Y): "consumers" - This represents lifetime post consumers, a numerical value. Covariate (X): "paid" - This is a binary variable (e.g., 1 for paid posts, 0 for unpaid). It allows us to compare the average reach between the two groups.

Since we want to see whether paid status impacts average reach, a linear model is appropriate with paid status as the independent variable, and lifetime consumers as teh response. This variable was used to represent reach as we felt the number of consumers of a post best represented overall reach.

Once again, using log1p(consumers) to negate the impact of outliers.

```r
ggplot(dat, aes(x = paid, y = log1p(consumers))) +
  geom_point(shape = 1, alpha = 0.45, size = 1.2, position = position_jitter(width = 0.25,   height = 0)) +
  labs(title = "Consumers vs Paid Status",
       x = "Paid Status (0 = Unpaid, 1 = Paid)",
       y = "Consumers") +
  theme_minimal()
```



There seems to be a very slight upward trend in reach as a result of posts being paid.

```r
# Basic OLS model
model <- lm(log1p(consumers) ~ paid, data = dat)
summary(model)
```

```
##
## Call:
## lm(formula = log1p(consumers) ~ paid, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1463 -0.4842  0.0152  0.5583  2.8862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.25012    0.04710 132.693   <2e-16 ***
## paidPaid     0.19875    0.08933   2.225   0.0265 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8949 on 498 degrees of freedom
## Multiple R-squared:  0.009841,   Adjusted R-squared:  0.007853
## F-statistic:  4.95 on 1 and 498 DF,  p-value: 0.02654
```

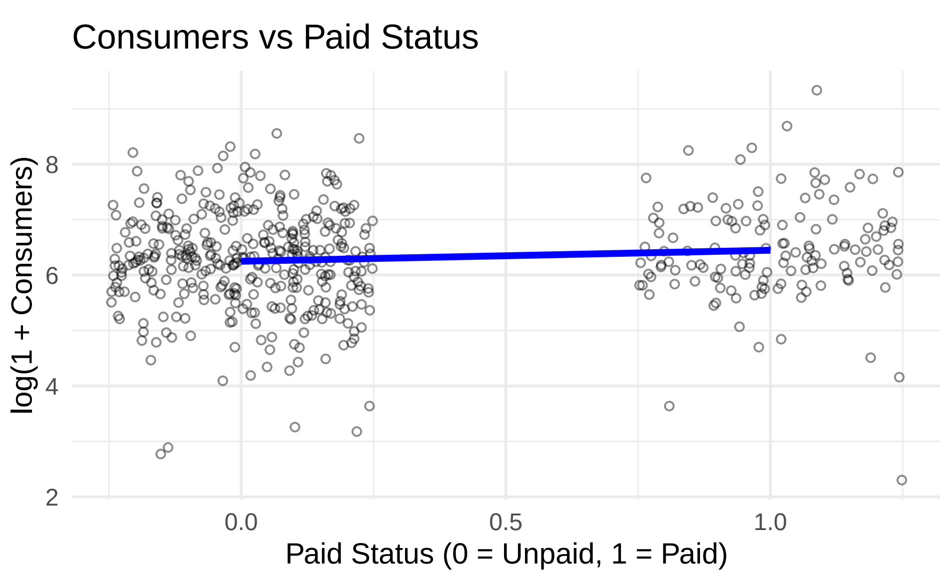
We can interpret these results from the OLS model…

The intercept is the average number of log1p(consumers) for unpaid posts. The paid coefficient is the average additional reach for paid posts as compared to unpaid.

Intercept: ~6.25, which would be ~517 consumers after back transforming. Paid Coefficient: ~114 more consumers for paid posts (after back transforming) Since the p-value of 0.0265 is less than 0.05, the difference between paid and unpaid is statistically significant. As the t-value of 2.225 is greater than 2, it is also likely that the difference is statistically significant.

```
dat$paid_numeric <- as.numeric(dat$paid) - 1

ggplot(dat, aes(x = paid_numeric, y = log1p(consumers))) +
  geom_point(shape = 1, alpha = 0.45, size = 1.2,
    position = position_jitter(width = 0.25, height = 0)
  ) + geom_smooth(method = "lm", se = FALSE, color = "blue", linewidth = 1.2) +
  labs(
    title = "Consumers vs Paid Status",
    x = "Paid Status (0 = Unpaid, 1 = Paid)",
    y = "log(1 + Consumers)"
  ) +
  theme_minimal()
```

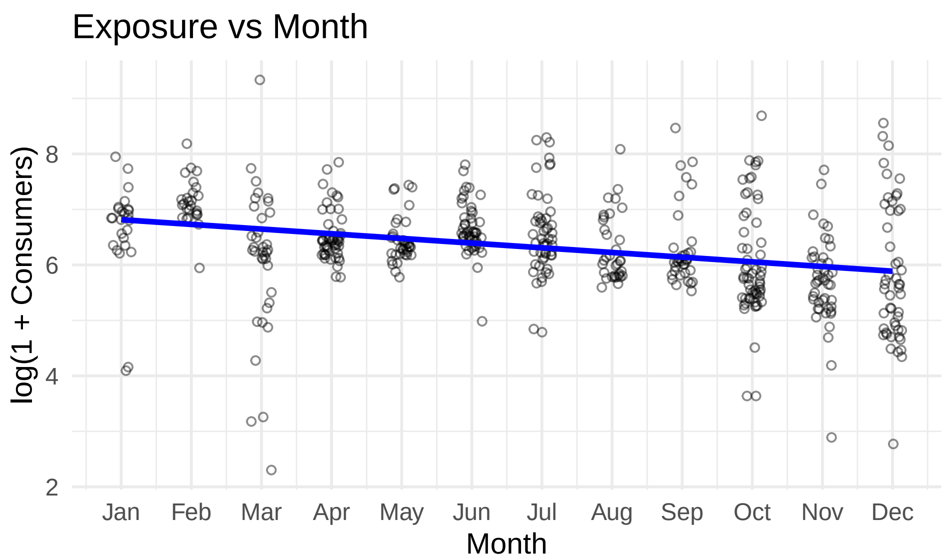


This model determines that there is an impact on reach based on paid status; however, it is not necessarily the strongest factor as the R^2 value is so small. This likely indicates that the variation in the response variable is due more largely to other factors than whether or not a post was paid.

Question 3: Does the cosmetics brand gain significant exposure over the course of the year, from first to latest post analyzed in 2014?

The response variable is still the post consumers and the covariate chosen is the post month. An OLS analysis on these two variables can show whether the posts reach more people as time progresses in the year. log1p(consumers) is chosen as the dependent variable to negate the impact of outliers.

```
plot_df2 <- dat |> dplyr::mutate(mon_num = match(as.character(month), month.abb))
ggplot(plot_df2, aes(x = mon_num, y = log1p(consumers))) +
  geom_point(shape = 1, alpha = 0.45, size = 1.2,
             position = position_jitter(width = 0.15, height = 0)) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +
  labs(title = "Exposure vs Month", x = "Month", y = "log(1 + Consumers)") +
  theme_minimal()
```



The scatter plot and the fitted regression line shows a slight downward trend. This contradicts our assumption and suggests the brand gradually loses exposure throughout 2014.

```
model_month <- lm(log1p(consumers) ~ mon_num, data = plot_df2)
summary(model_month)
```

```
##
## Call:
## lm(formula = log1p(consumers) ~ mon_num, data = plot_df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3436 -0.4192 -0.0832  0.4078  2.6889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.89940    0.08995  76.704  < 2e-16 ***
## mon_num     -0.08440    0.01157  -7.296 1.18e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8548 on 498 degrees of freedom
## Multiple R-squared:  0.09657,    Adjusted R-squared:  0.09475
## F-statistic: 53.23 on 1 and 498 DF,  p-value: 1.178e-12
```

The negative coefficient confirms the negative relationship between exposure and months. The low p-value suggests that the linear model is significant but the low R^2 value indicates high variation in the data.

Author Contribution Statement

1. **First Deliverable**
   1. **Yihang Duanmu** — insight for the motivation section
   2. **Minjun Kim** — introduction/summary
   3. **Annelise Schreiber** — questions/impact; editing
2. **Second Deliverable**
   1. **Yihang Duanmu** — OLS analysis for Question 3
   2. **Minjun Kim** — OLS analysis for Question 1
   3. **Annelise Schreiber** — OLS analysis for Question 2