Coding Session 2: Going Through a Sample Prompt

Cal Poly Pomona Datathon – Southern California Consortium for Data Science

Saturday April 27 2024

Sample Prompt: Paid Promotion

Your boss comes up to you and says "we're spending all of this money on promotion. How exactly is it helping us? Are there certain situations where it is less effective than others?"

Challenge 1:

Look at the facebook summary sheet and brainstorm some ways to answer your boss's first question: *"How exactly is it helping us?"*

Also brainstorm ways to answer the second question: *"Are there certain situations where it is less effective than others?"*

In your brainstorming, think about plots and summary statistics you could produce to try to answer the questions.

1. First Question

"How exactly is it helping us?"

Load Data and Preliminaries

```
# Load libraries first if you know which ones you need
library(ggplot2)

# This needs to be done once to load the dataset into the environment.
facebook <- read.csv("facebook_cosmetics.csv", stringsAsFactors = TRUE)

# Look at the first few rows
head(facebook)
```

```
##   PageLikes PostType ContentCategory Month Weekday Hour PaidAd TotalReach
## 1    139441    Photo          product    12       4    3     No       2752
## 2    139441   Status          product    12       3   10     No      10460
## 3    139441    Photo     inspiration    12       3    3     No       2413
## 4    139441    Photo          product    12       2    3     No       7244
## 5    139441   Status          product    12       1    9     No      10472
## 6    139441    Photo     inspiration    12       1    3    Yes      11692
##   TotalImpressions EngagedUsers Comments Likes Shares Interactions
## 1             5091          178        4    79     17          100
## 2            19057         1457        5   130     29          164
## 3             4373          177        0    66     14           80
## 4            13594          671       19   325     49          393
## 5            20849         1191        1   152     33          186
## 6            19479          481        3   249     27          279
```

Challenge 2

Carefully look two of the rows `head(facebook)` and carefully understand what each number means.

Using the *third* row as an example, I would say the following

- at the time of posting, the page had 139441 likes.
- The post was a photo, and it was in the "inspiration" category (*meaning non-explicit brand related content*).
- It was posted in December on Tuesday at 3pm.
- It did not have paid advertising.
- The post was seen by 2413 unique individuals, and it was seen in total 4373 times.
- 177 users engaged with the post *(meaning they clicked on it in some manner)*.
- It had 0 comments, 66 likes, and 14 shares, resulting in a total of 80 interactions.

1.1 Brainstorming and Boxplots

Let's brainstorm some things that will be useful in answering the boss's question.

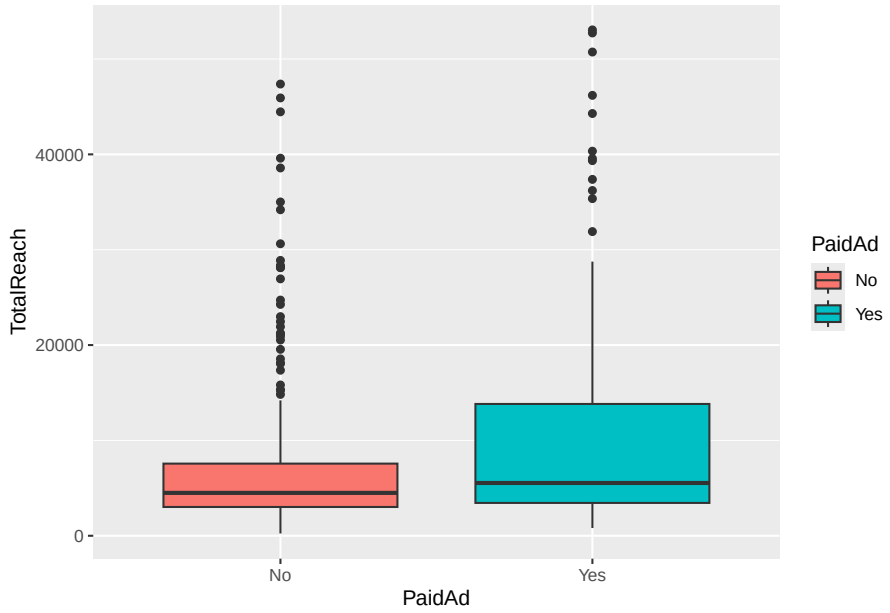
- `PaidAd` has two outcomes and can be considered categorical.
- Desired qualities of a post would be when they have a high `TotalReach`, `TotalImpressions`, `EngagedUsers`, `Comments`, `Likes`, `Shares`, and `Interactions`. These are all **numerical** so we can use boxplots and/or histograms.

– Since `Interactions` is the total of `Comments`, `Likes`, and `Shares`, we will just look at `Interactions` in this example.

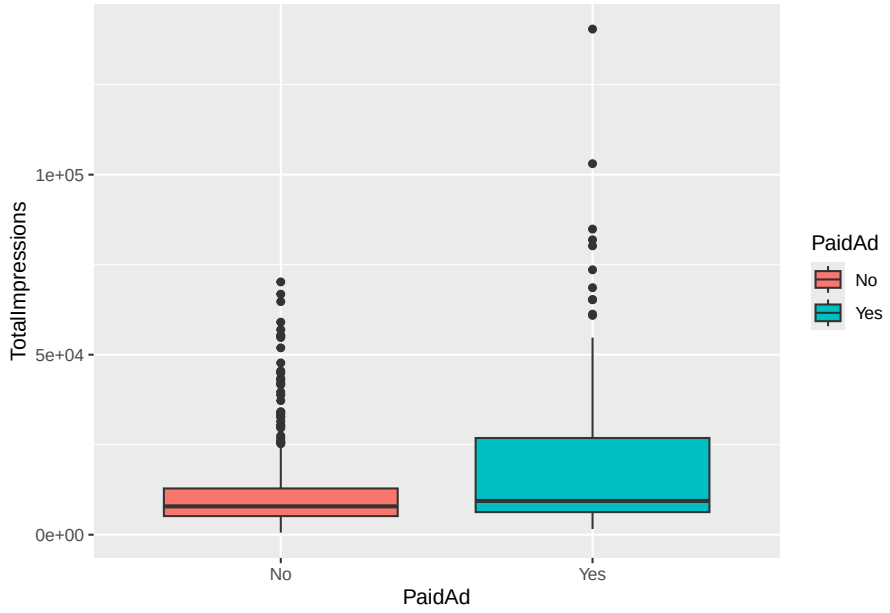
(Make sure you have the "Facebook Dataset" sheet in front of you, so you can quickly reference what the variables mean!)

[Note: `PageLikes` is the number that like the page at the time of posting, not of the post itself.]

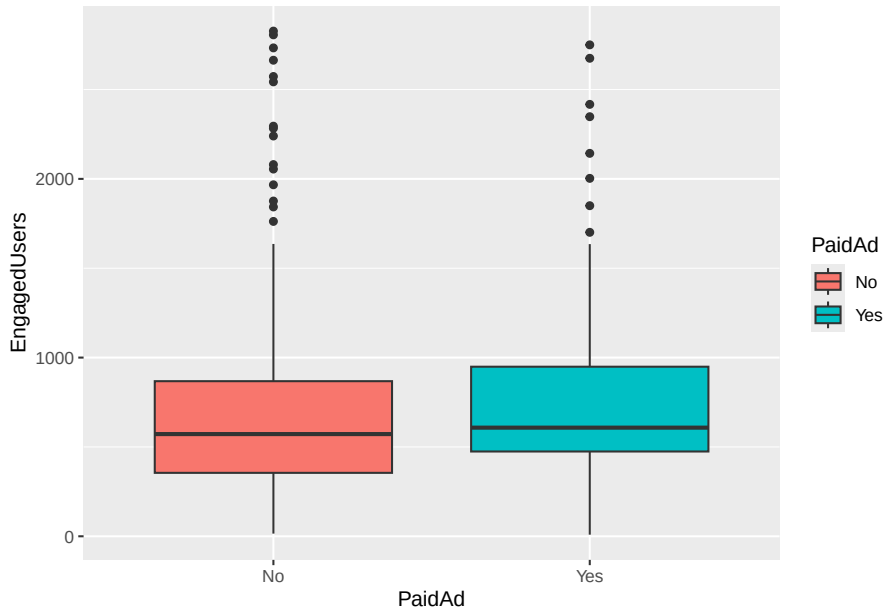
```
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = TotalReach, fill = PaidAd))
```



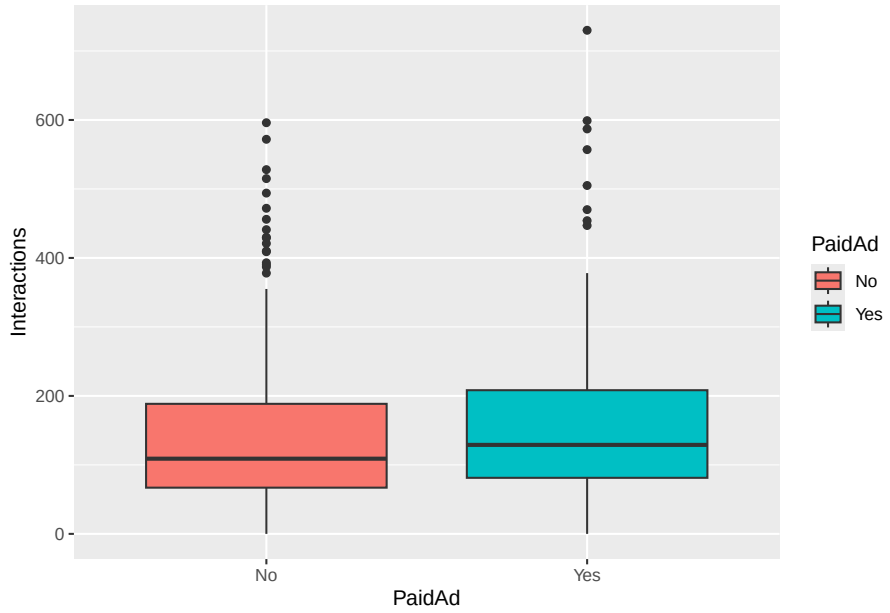

```
# For efficiency, you should copy+paste the other boxplot code to produce this
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = TotalImpressions, fill = PaidAd))
```




```
# For efficiency, you should copy+paste the other boxplot code to produce this
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = EngagedUsers, fill = PaidAd))
```




```
# For efficiency, you should copy+paste the other boxplot code to produce this
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = Interactions, fill = PaidAd))
```



- It looks like paying for advertising increases the reach *(number of unique users to see the post)* and impressions *(number of times page was seen)* by a lot.
- It also increases the number of times the post was engaged with, and the interactions *(total number of comments, likes, and shares)*, but not by as much as the reach and impressions.

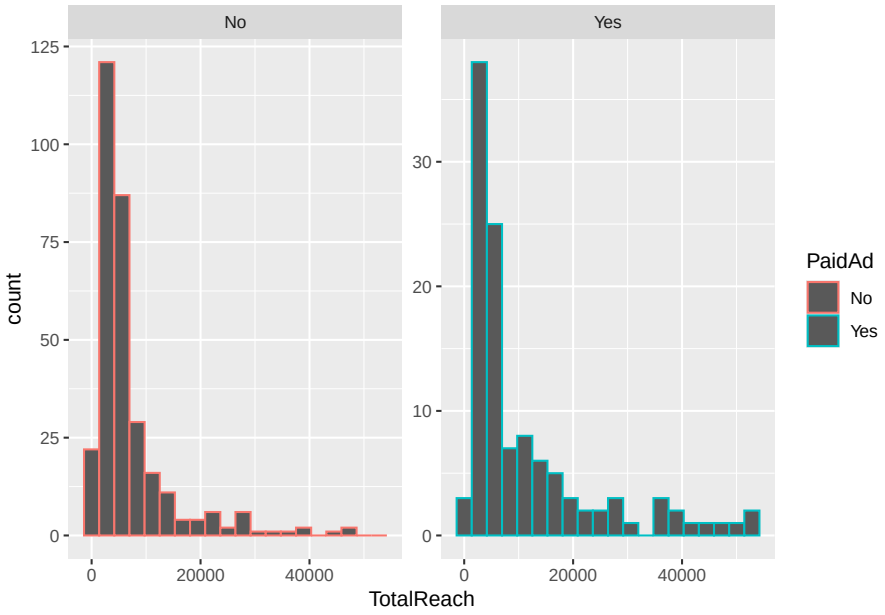
Challenge 3

Think about how advertising works in social media, and connect this to the two bullet points above. Why do you think it didn't increase the engagement as much as the reach?

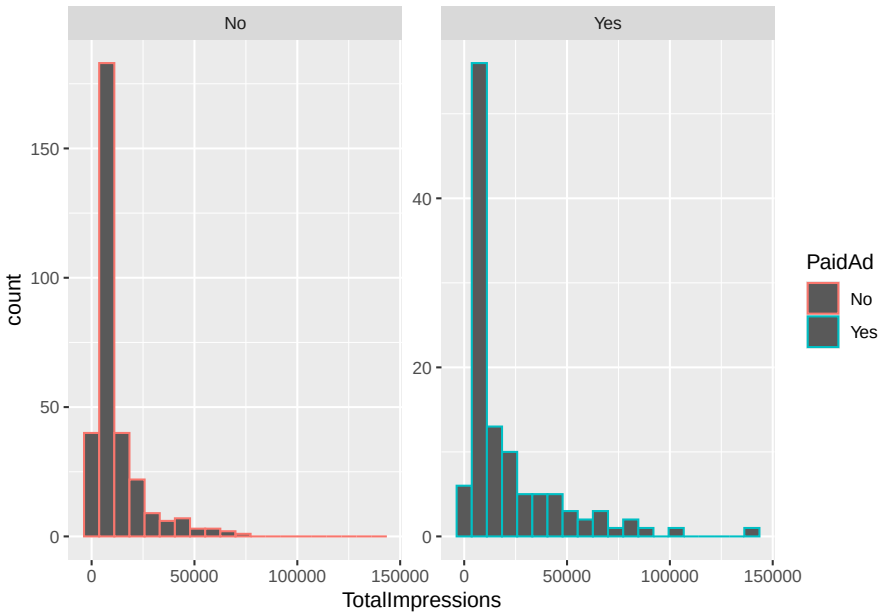
1.2 Histograms

Histograms can show similar but more detailed relationships.

```
ggplot(facebook) +
  geom_histogram(aes(x = TotalReach, col = PaidAd), bins = 20) +
  facet_wrap(~PaidAd, scales = "free_y")
```




```
# For efficiency, you should copy+paste the other histogram code to produce this
ggplot(facebook) +
  geom_histogram(aes(x = TotalImpressions, col = PaidAd), bins = 20) +
  facet_wrap(~PaidAd, scales = "free_y")
```



For example, this shows that advertising generally produces more reach, but more importantly, a larger number of cases where the reach is around 40,000 or larger.

Challenge 4

Produce similar histograms for `EngagedUsers` and `Interactions`. Can you make similar claims?

1.3 Numerical Summary Statistics

```
# We can also get some numerical summary statistics
facebook_not_paid <- subset(facebook, PaidAd == "No")
facebook_paid <- subset(facebook, PaidAd == "Yes")

summary(facebook_not_paid)
```

```
##    PageLikes        PostType      ContentCategory     Month
##  Min.   : 81370   Link  : 14   action     :127    Min.   : 1.000
##  1st Qu.:111132   Photo :272   inspiration:100    1st Qu.: 4.000
##  Median :130791   Status: 28   product    : 89    Median : 7.000
##  Mean   :122992   Video :  2                      Mean   : 7.073
##  3rd Qu.:136642                                   3rd Qu.:10.000
##  Max.   :139441                                   Max.   :12.000
##     Weekday           Hour         PaidAd      TotalReach    TotalImpressions
##  Min.   :1.000   Min.   : 1.000   No :316   Min.   :  238   Min.   :  570
##  1st Qu.:2.000   1st Qu.: 3.000   Yes:  0   1st Qu.: 3018   1st Qu.: 5177
##  Median :4.000   Median : 9.000             Median : 4498   Median : 7878
##  Mean   :4.149   Mean   : 7.949             Mean   : 7006   Mean   :11867
##  3rd Qu.:6.000   3rd Qu.:11.000             3rd Qu.: 7567   3rd Qu.:12853
##  Max.   :7.000   Max.   :23.000             Max.   :47376   Max.   :70212
##   EngagedUsers       Comments         Likes            Shares
##  Min.   :  15.0   Min.   : 0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 355.0   1st Qu.: 1.00   1st Qu.: 52.75   1st Qu.: 10.00
##  Median : 572.0   Median : 2.00   Median : 87.50   Median : 18.00
##  Mean   : 693.8   Mean   : 4.43   Mean   :112.87   Mean   : 21.92
##  3rd Qu.: 868.0   3rd Qu.: 5.00   3rd Qu.:154.00   3rd Qu.: 29.00
##  Max.   :2827.0   Max.   :64.00   Max.   :529.00   Max.   :121.00
##   Interactions
##  Min.   :  0.0
##  1st Qu.: 67.0
##  Median :109.0
##  Mean   :139.2
##  3rd Qu.:188.5
##  Max.   :596.0
```

```
summary(facebook_paid)
```

```
##    PageLikes        PostType      ContentCategory     Month
##  Min.   : 85979   Link  :  6   action     :49     Min.   : 1.000
##  1st Qu.:113613   Photo :100   inspiration:40     1st Qu.: 4.000
##  Median :129600   Status:  6   product    :25     Median : 7.000
##  Mean   :123915   Video :  2                     Mean   : 7.026
##  3rd Qu.:135938                                  3rd Qu.:10.000
##  Max.   :139441                                  Max.   :12.000
##     Weekday           Hour         PaidAd      TotalReach    TotalImpressions
##  Min.   :1.000   Min.   : 2.000   No :  0   Min.   :  813   Min.   :  1568
##  1st Qu.:3.000   1st Qu.: 3.000   Yes:114   1st Qu.: 3442   1st Qu.:  6290
##  Median :4.000   Median : 7.000             Median : 5536   Median :  9378
##  Mean   :4.246   Mean   : 7.219             Mean   :11453   Mean   : 21061
##  3rd Qu.:6.000   3rd Qu.:11.000             3rd Qu.:13822   3rd Qu.: 26864
##  Max.   :7.000   Max.   :22.000             Max.   :53056   Max.   :140432
##   EngagedUsers       Comments          Likes            Shares
##  Min.   :   9.0   Min.   : 0.000   Min.   :  0.0   Min.   :  0.00
```

```
##  1st Qu.: 474.5   1st Qu.: 1.000   1st Qu.: 65.0   1st Qu.: 10.00
##  Median : 608.0   Median : 3.000   Median :106.0   Median : 17.00
##  Mean   : 778.8   Mean   : 5.614   Mean   :141.8   Mean   : 22.18
##  3rd Qu.: 949.0   3rd Qu.: 7.000   3rd Qu.:171.0   3rd Qu.: 29.75
##  Max.   :2750.0   Max.   :47.000   Max.   :696.0   Max.   :123.00
##   Interactions
##  Min.   :  0.00
##  1st Qu.: 81.25
##  Median :129.00
##  Mean   :169.61
##  3rd Qu.:208.25
##  Max.   :730.00
```

Looking at the means, advertisements

- Increase the average number of users that see a post by 4447 *(going from 7006 not paid to 11453 paid)*
- Increase the average number of times a post was seen by 9194 *(going from 11867 not paid to 21061 paid)*
- Increase the average number of times the post engaged with (clicked on) by 85 *(going from 693.8 not paid to 778.8 paid)*

Math Note!

Even though the number of times the post was engaged with increased only by 85, this was an increase of 12.25%.

$$\frac{778.8}{693.8} - 1 = 0.1225 \qquad \text{or} \qquad 12.25\%$$

```
# Compute the percentage in R
778.8 / 693.8 - 1  # as a decimal
```

```
## [1] 0.1225137
```

```
(778.8 / 693.8 - 1) * 100  # as a percentage
```

```
## [1] 12.25137
```

(you can multiply the decimal by 100 to get the %) So, there is a 12.25% increase in the average numbers of users clicking on a post when an advertisement is used.

This is a useful thing to mention to your boss!

Challenge 5

What is the average increase in post `interactions` for when a post has a paid advertisement *(compared to not having one)*?

Also compute the percentage increase in the average post `interactions`. *(Check your answer with the correct one which is 21.85%.)*

1.4 Answer to Boss's Question 1

Here are some things that could be mentioned to the boss for their first question:

- Advertising significantly helps post reach and impressions *(showing the post to more users)*. It helps with post interaction, but not as much as it does for reach and impressions. This is expected, because it is what paying for advertising does. **(Use boxplots to help explain it.)**
  - *[You should add something similar, but for `EngagedUsers` and `Interactions`.]*

- Advertising helps to produce posts that have very large reach. **(Show faceted histograms)**
- Advertisements increase the average number users that see a post was a post by 4447 *(going from 7006 without to 11453 with)*, with a percent increase in the average of 63.47%.

Similar claims would be made about the other average increases and percentage increases.

It is also important to think about what "helping us" means. For example, we could make similar claims about the `likes` or `shares` of a post. But are these as important as things like the `total post reach` and `number of engaged users`? (*The answer is not 100% clear, and it rarely is. This always requires critical thinking!*)

Challenge 6

Fill in the *"You should add something similar..."* bullet point based on your work in Challenge 3.

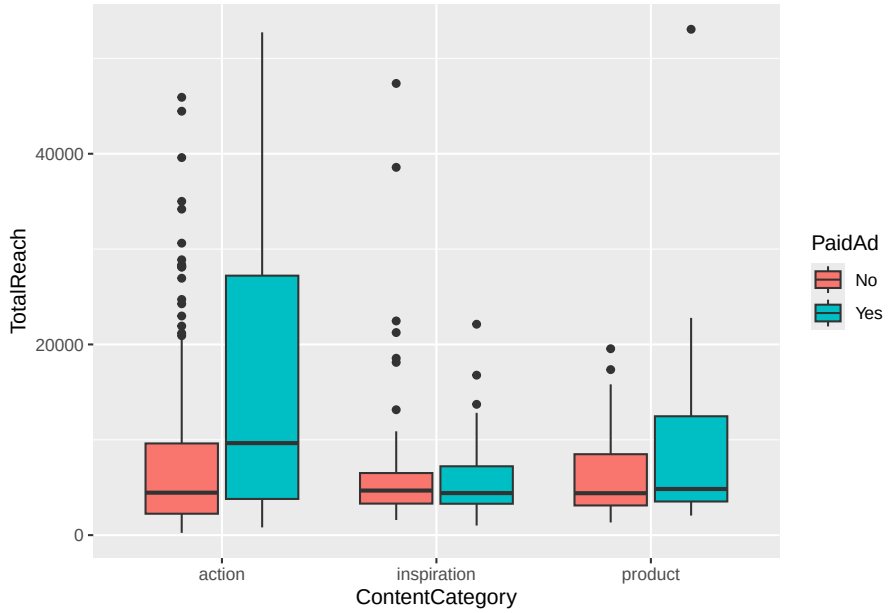
2. Second Question

"Are there certain situations where it is less effective than others?"

This is a much **more challenging question** and there are several ways to begin coming up with an answer.

A straightforward way to begin is to see how `PaidAd` works with *other variables* to affect things like `total reach`, `engaged users`, etc..

```
# Boxplot of TotalReach divided by Content Category and Advertising
ggplot(facebook) +
  geom_boxplot(aes(y=TotalReach, x = ContentCategory, fill = PaidAd))
```



Right away this tells us that paid ads tend to be more effective for **Action** posts *(special offers and contents)* in increasing the number of users that see the post. For **Product** posts *(direct advertisement and explicit brand content)* it offers an increase toward the higher end of posts, but not as much as for action posts.

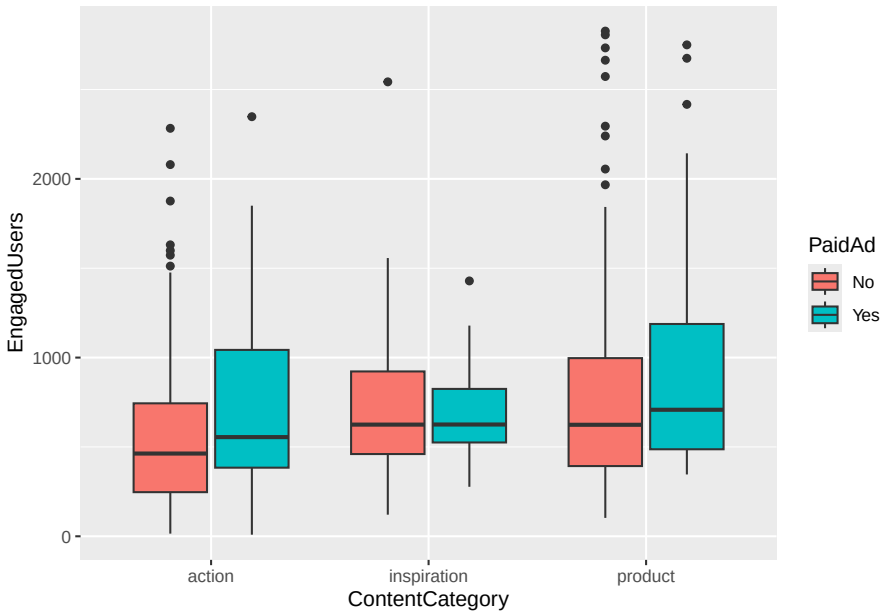
The improvement in **Inspiration** posts *(non-explicit brand related content)* are even less pronounced.

```
# Frequency table of different categories.
#   This checks if the boxplots are reliable.
table(facebook$ContentCategory, facebook$PaidAd)
```

```
##
##                No Yes
##   action      127  49
##   inspiration 100  40
##   product      89  25
```

None of these frequencies are incredibly small. If they were, it might invalidate some of our claims above.

```
# Boxplot of EngagedUsers divided by Content Category and Advertising
ggplot(facebook) +
  geom_boxplot(aes(y=EngagedUsers, x = ContentCategory, fill = PaidAd))
```



The conclusions are mostly the same when working with `EngagedUsers`, just less extreme. The `Inspiration` category is an exception, where it increases the lower end of engaged users, but actually decreases the upper end of engaged users. This decreases the *variability*, but does not tend to offer an overall increase.

Challenge 7

Perform the same analysis for `TotalImpressions` and `Interactions`. This means to make similar plots and analyze them similarly. Are the conclusions mostly the same? What is different?

Challenge 8

Perform similar for `PostType` instead of `ContentCategory`.

2.1 Answer to Boss's Second Question

Challenge 9

Based on the analysis performed so far, what would you report to your boss to answer the question *"Are there certain situations where it is less effective than others?"

Challenge 10

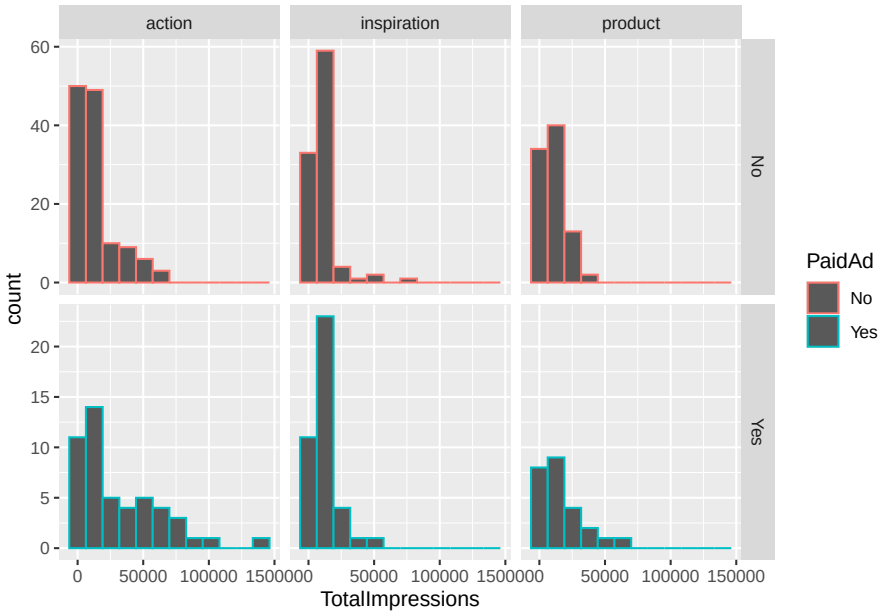
Brainstorm some additional ways to analyze the data to better answer your boss's question. *(some examples are addressed on the next page)*

2.2 Some Next Steps

There are many ways to perform further analysis for your boss's second question. Here are a few things to take into consideration:

1. Histograms like before, but using `facet_grid`, like this:

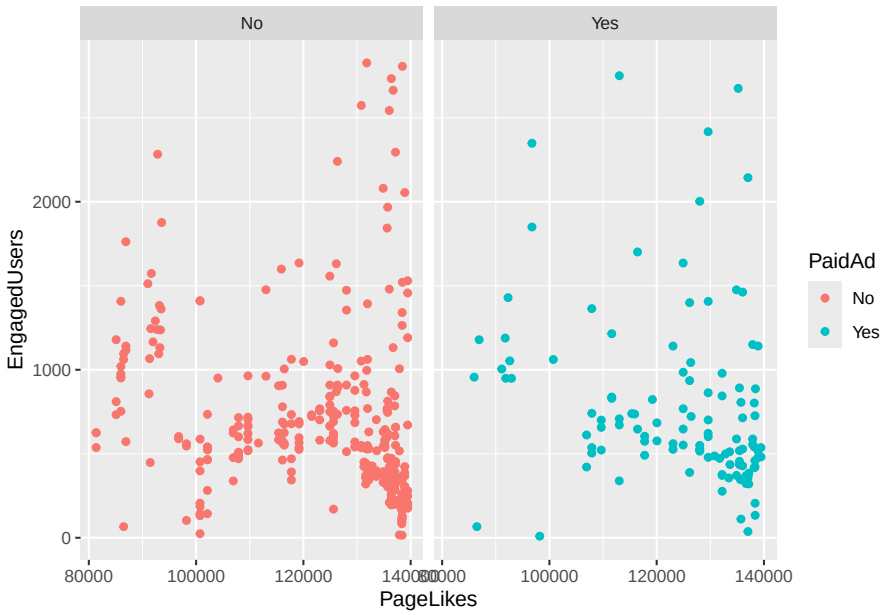
```
ggplot(facebook) +
  geom_histogram(aes(x = TotalImpressions, col = PaidAd), bins = 12) +
  facet_grid(PaidAd~ContentCategory, scales = "free_y")
```



What does this tell us?

2. Critically evaluate what your boss means by "effective." Is it specifically post exposure (through `TotalReach` and `TotalImpressions`) enough? Or is something like `EngagedUsers` more important? Maybe both?
3. Perform similar analysis but across other variables like `PostType`, `PageLikes`, and time of posting `Month`, `Weekday`, `Hour`.
4. For the numeric variable `PageLikes` a visual may look something like

```
# This plot doesn't show much, but it is a proof of concept.
ggplot(facebook) +
  geom_jitter(aes(y = EngagedUsers, x = PageLikes, col=PaidAd)) + facet_wrap(~PaidAd)
```



One of the challenges in data analysis is to know when you are done. Just like with a painting, there is no clearly defined end point. So, try your best to answer the prompts as best you can, and move on when you feel ready!