Sample Prompt

Cal Poly Pomona Datathon

(Southern California Consortium for Data Science)

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 data 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?"*

First Question

"How exactly is it helping us?"

Load Data and Preliminaries

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

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

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

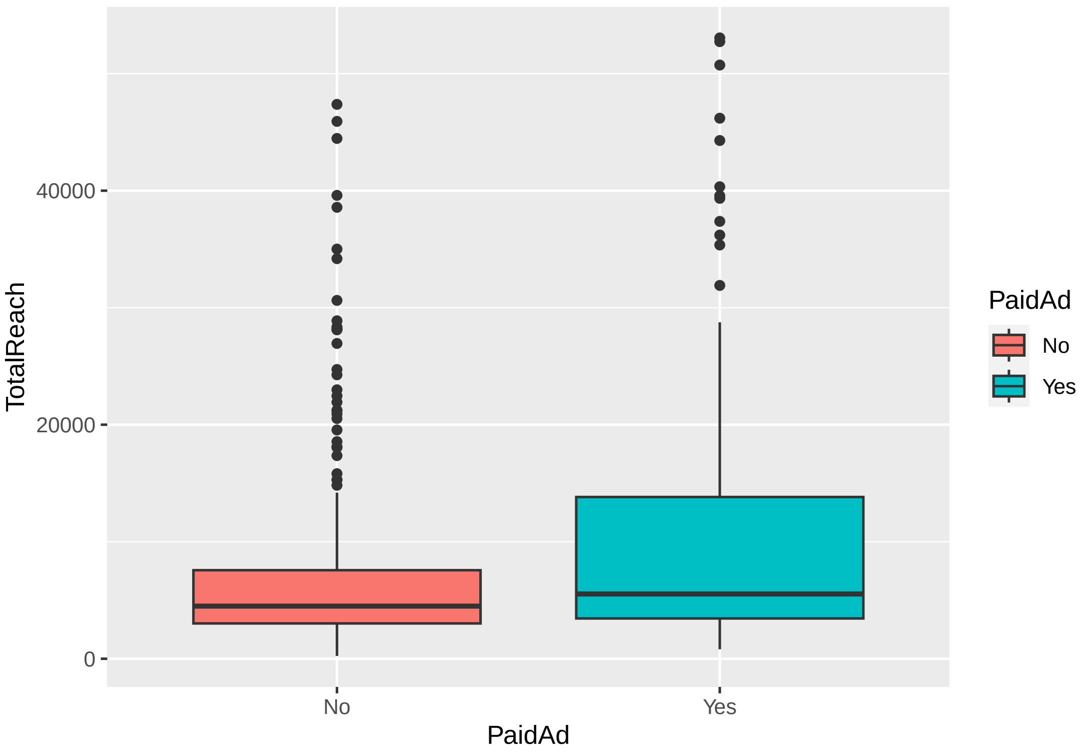
```
##   PageLikes PostType ContentCategory Month Weekday Hour PaidAd TotalReach
## 1    139441    Photo         product    12       4    3     No       2752
## 2    139441   Status         product    12       3   10     No      10460
## 3    139441    Photo     inspiration    12       3    3     No       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
```

Brainstorming

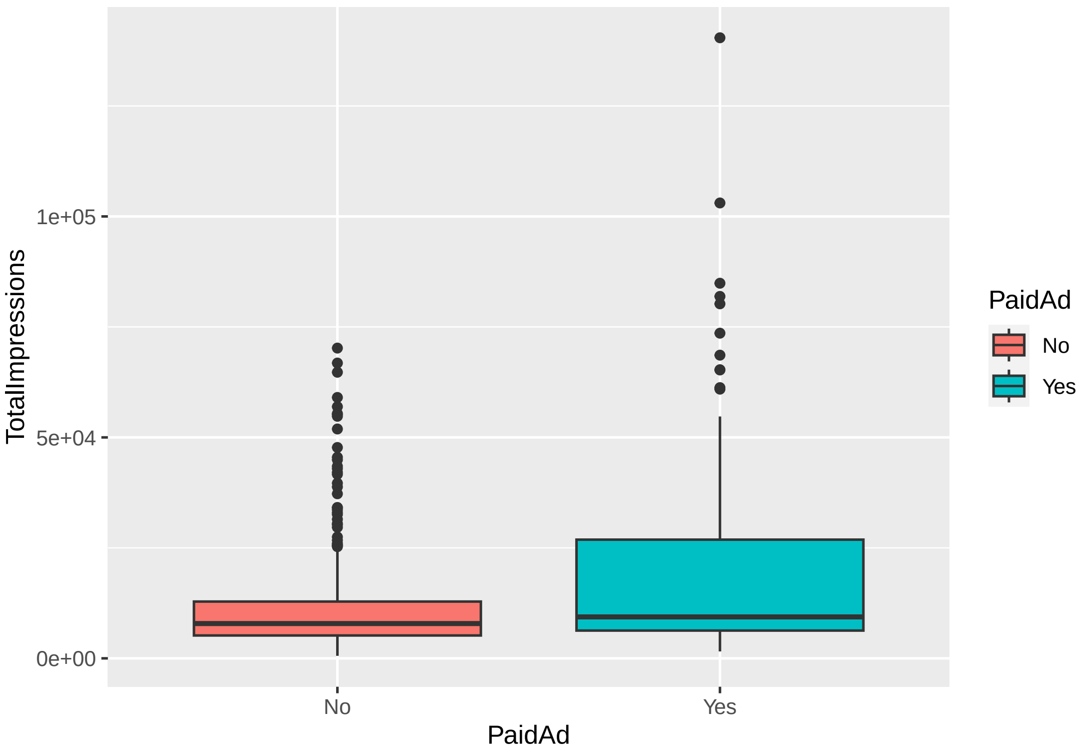
- `PaidAd` has two outcomes and can be considered categorical.
- Desired qualities of a post would be when they have a high `TotalReach`, `TotalImpressions`, `EngagedUsers`, `Comments`, `Likes`, `Shares`, and `Interactions`. These are all **numerical** so we can use boxplots and/or histograms.
    - Since `Interactions` is the total of `Comments`, `Likes`, and `Shares`, we will just look at `Interactions` in this example.

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

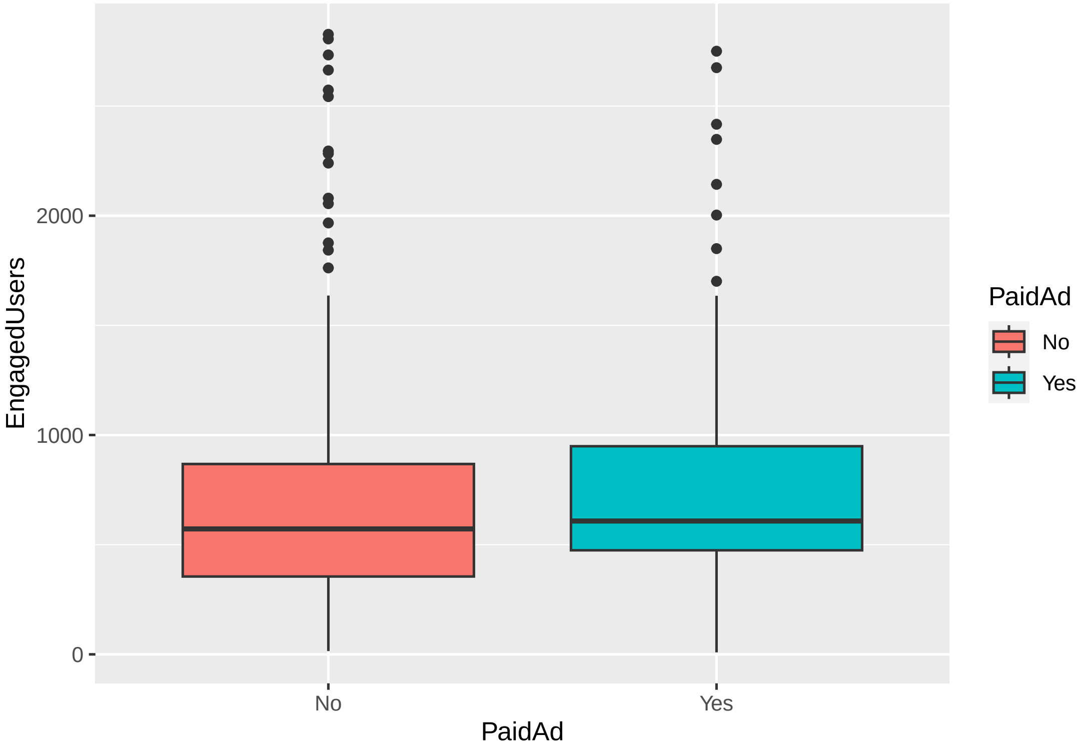
```r
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = TotalReach, fill = PaidAd))
```



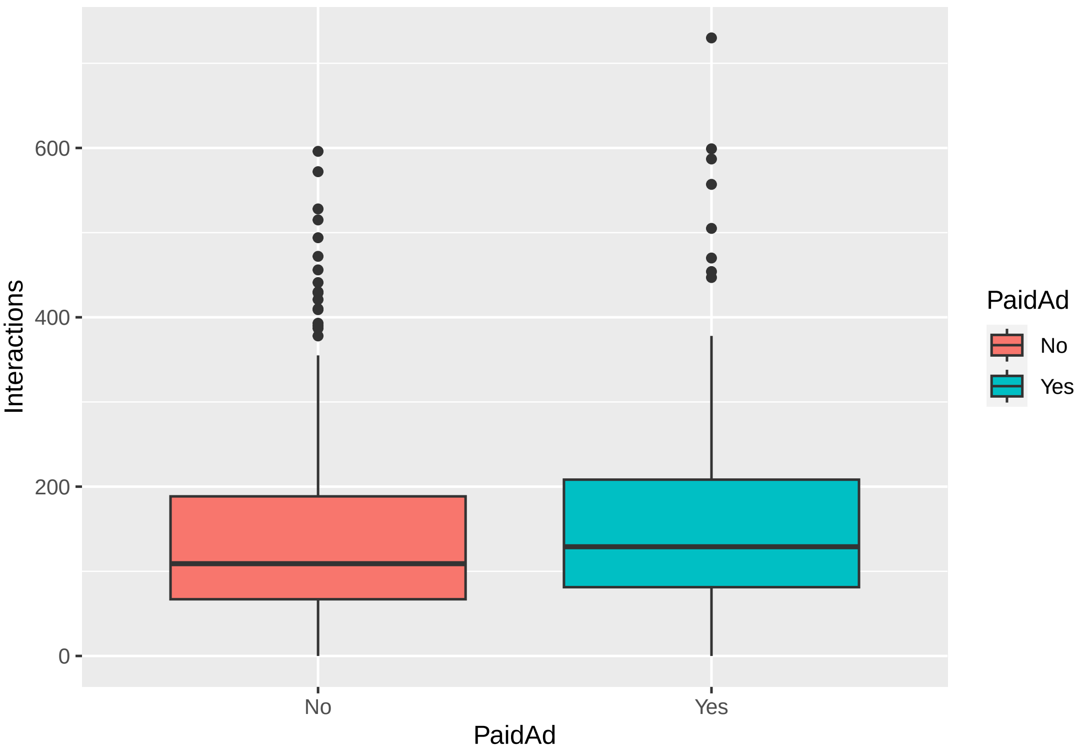
```r
# For efficiency, you should copy+paste the code above to produce this code
ggplot(facebook) + geom_boxplot(aes(x = PaidAd, y = TotalImpressions, fill = PaidAd))
```



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



```
# For efficiency, you should copy+paste the code above to produce this code
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 number of comments, likes, and shares, but not by as much as the reach and impressions.

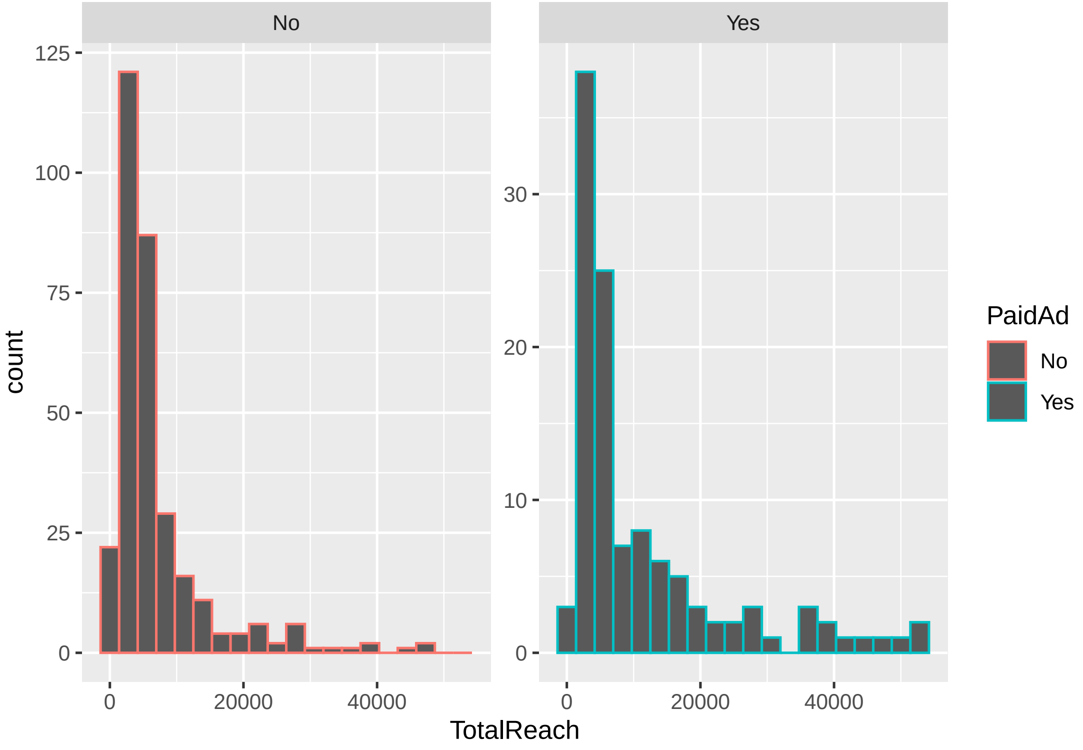
Challenge 2

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?

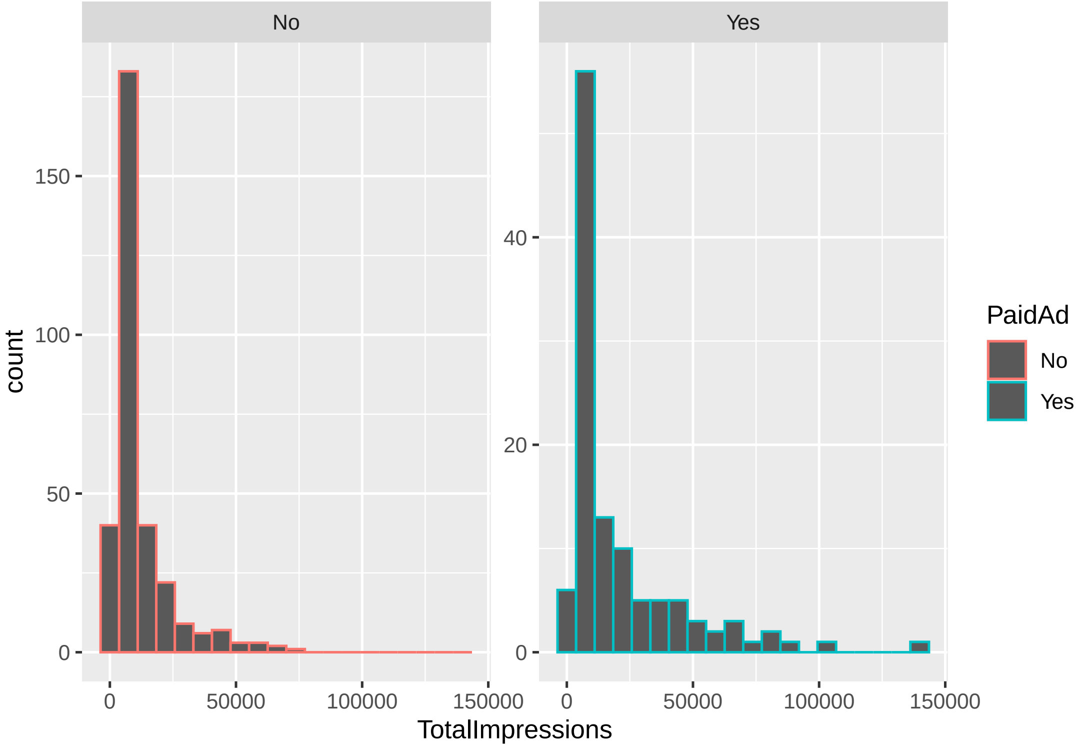
Histograms

Histograms can show similar but more detailed relationships.

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



```
# For efficiency, you should copy+paste the code above to produce this code
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 3

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

Numerical Summary Statistics

```r
# We can also get some numerical summary statistics
facebook_not_paid <- subset(facebook, PaidAd == 0)
facebook_paid <- subset(facebook, PaidAd == 1)


# Summary statistics
summary(facebook_not_paid$TotalReach)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```r
summary(facebook_paid$TotalReach)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```r
# For efficiency, you should copy+paste the code above to produce this code
summary(facebook_not_paid$TotalImpressions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```r
summary(facebook_paid$TotalImpressions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```r
# For efficiency, you should copy+paste the code above to produce this code
summary(facebook_not_paid$EngagedUsers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

```r
summary(facebook_paid$EngagedUsers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

Looking at the means, advertisements

- Increase the average number of users that see a post by 4447 *(going from 7006 without to 11453 with)*
- Increase the average number of times a post was seen by 9194 *(going from 11867 without to 21061 with)*
- Increase the average number of times the post engaged with (clicked on) by 85 *(going from 693.8 without to 778.8 with)*

Math Note!

Even though the number of times the post was engaged with increased only by 85, it increased a decent amount percentage wise. The percent increase is

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

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

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

```r
(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 4

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%.)*

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. **(Use boxplots to help explain it.)**
- 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.

Challenge 5

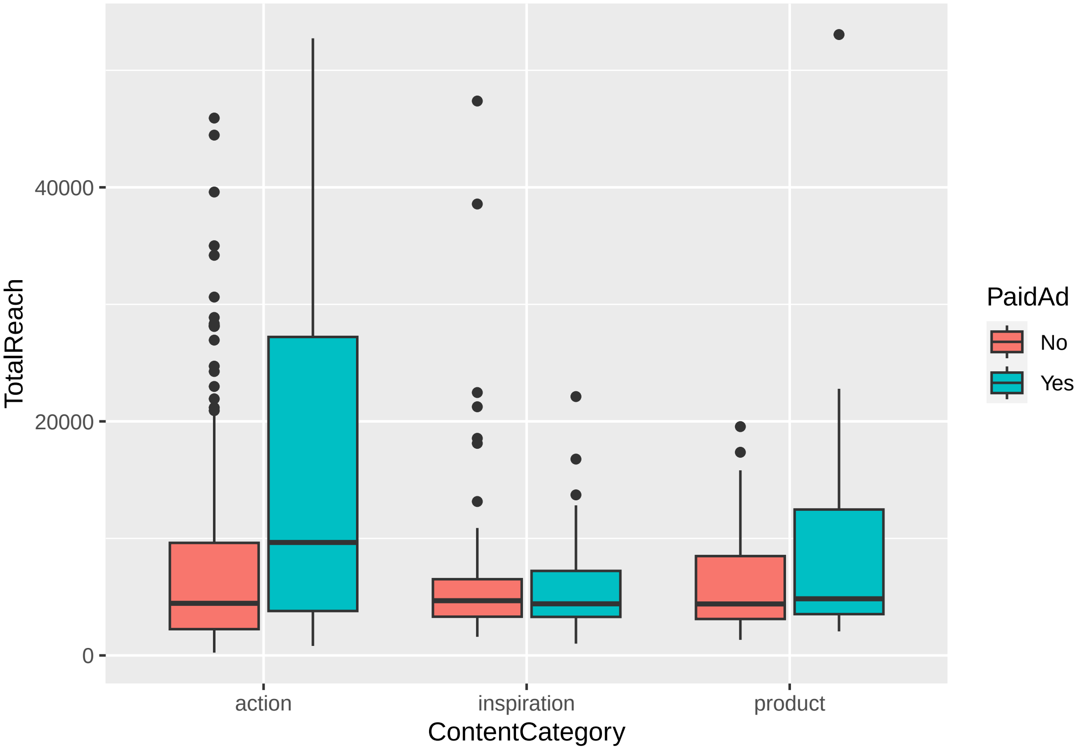
What would you add to this answer based on Challenge 3?

Second Question

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

This is a more challenging question and there are several ways to begin coming up with an answer. Here is a straightforward way to begin:

```r
# 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.

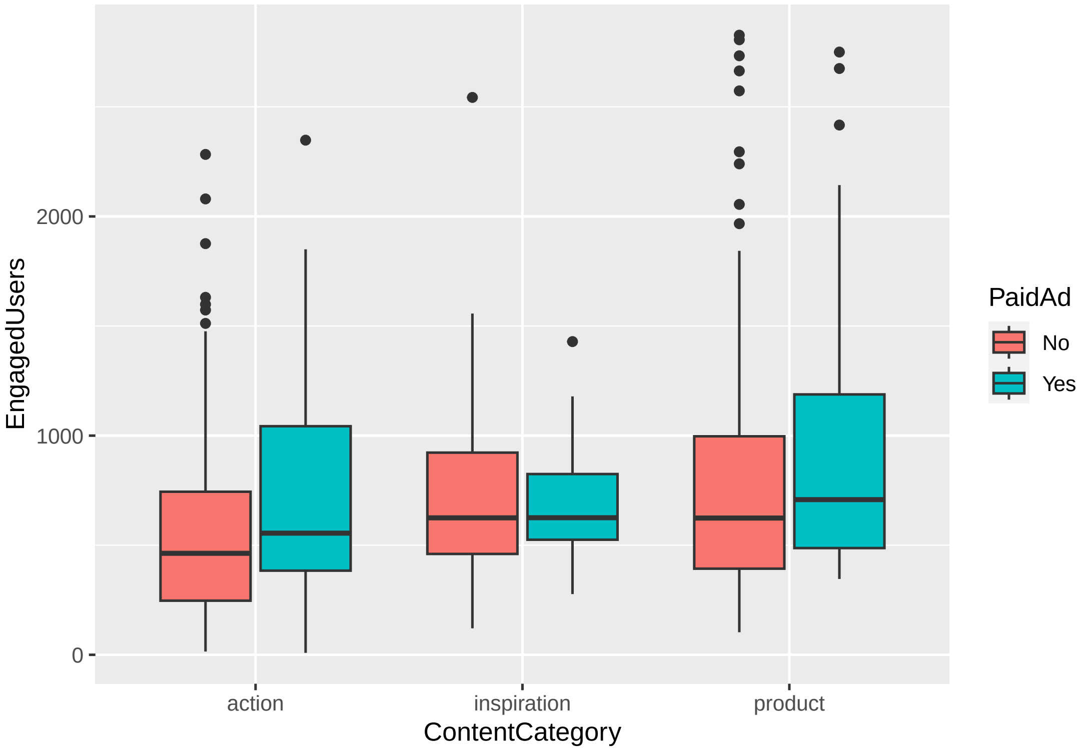
The improvement in **Inspiration** posts *(non-explicit brand related content)* are even less pronounced.

```r
# Frequency table of different categories
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.

```r
# 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 6

Perform the same analysis for `TotalImpressions` and `Interactions`.

Challenge 7

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 8

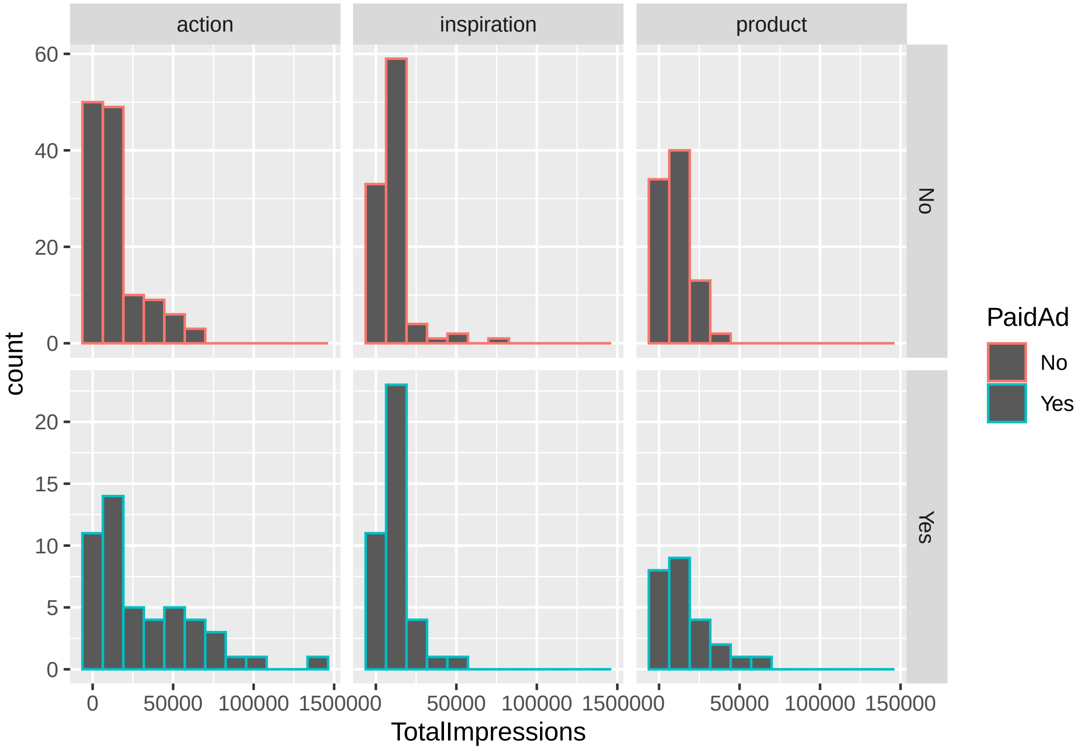
Brainstorm some additional ways to analyze the data to better answer your boss's question. *(This is addressed on the next page)*

Next Steps

There are many ways to perform more analysis for your boss's second question. Here are a few examples:

- Make Histograms like before, but using `facet_grid`, like this:

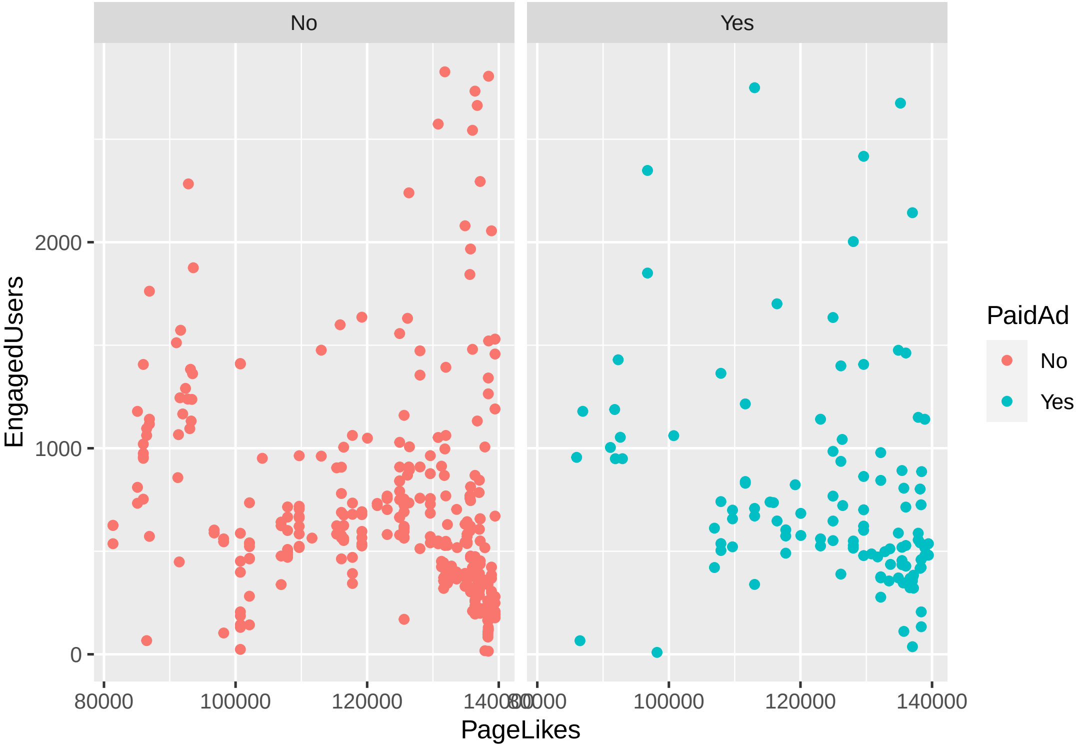
```r
ggplot(facebook) +
  geom_histogram(aes(x = TotalImpressions, col = PaidAd), bins = 12) +
  facet_grid(PaidAd~ContentCategory, scales = "free_y")
```



- 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?
- Perform similar analysis but across other variables like `PostType`, `PageLikes`, and time of posting `Month`, `Weekday`, `Hour`.

For the numeric variable `PageLikes` a visual may look something like

```r
# 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!