Group 15 - Anisha Bhuiyan, Shermeen Khan, Sahil Rajapkar
CIS 3920 Data Mining for Business Analytics
Professor Charles Terng

Homework #2

Section II. Univariate Analysis

(1)

    (a) What do the descriptive stats for "share" tell us about the cosmetic company's social media performance?

    The number of shares vary, with the minimum number of posts receiving 0 shares, while the maximum reaching as high as 790. The average number of shares is around 28. 50% of the data is between 10 and 33 shares with the lower quartile of shares being 10 and the upper quartile being approximately 33. The standard deviation is approximately 43. The standard deviation is greater than the mean itself, which is indicative of how spread and dispersed the data is from the mean. Most of the numbers are fitted within one standard deviation of data (from 0 to 42).

    Most of the shares are distributed within the 10-30s range. Based on the distribution of shares seen throughout, it is quite clear that the company can do better in increasing the number of shares coming into the business. They can focus on the posts that are receiving a higher number of shares, and replicate the strategy to increase the average number of shares accordingly.

(2)

    (a) Label the x and y axes using the xlab and ylab parameters. Create a title for the plot using the main parameter.

    The graph above is the histogram of shares. The X axis is the number of Facebook shares and the y axis is the number of total shares.

Histogram of Shares
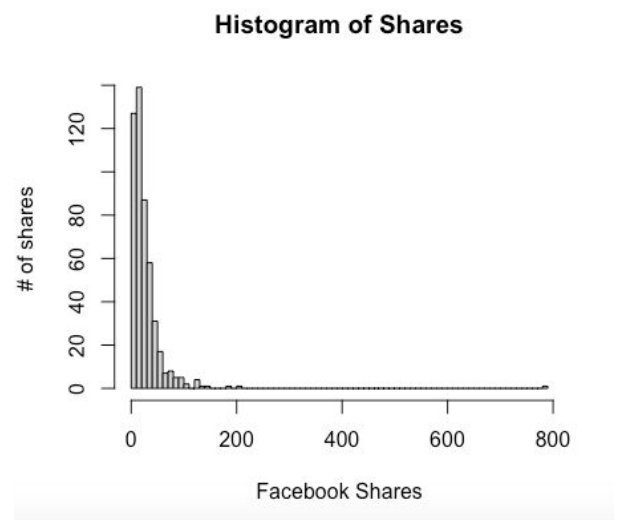
(b) What does the plot tell us about the distribution of shares? Any outliers? Adjust the xlim parameter to examine the 0-200 range on the x-axis. What range of shares do most posts tend to fall in?

The histogram is positively skewed which means that the mean>median>mode. Upon closer analysis, we saw there was an outlier at 790, which is also the maximum value in the histogram. The majority of shares fall in the 0-50 range.


Histogram of Shares
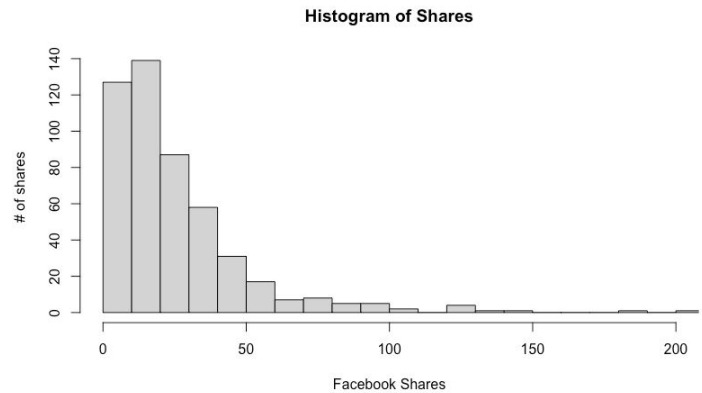
(3)

(a) Do we see any trends in posting frequency by month?

We see a trend in posts based off of the seasons. In spring there is an increase in posting activity in April. We see a brief drop in May, however that is followed by the summer months seeing an increasing number of posts. By the end of summer and beginning of fall, there is a drop once more. However, there is a peak reached in posts during October. After October, there is a decreasing trend again in the winter months.

```
        Category          Paid
action      :211    non_paid:356
inspiration:155     paid    :139
product     :129
```

```
   Post.Month  Post.Weekday   Post.Hour
10      : 57    1:68          3       :105
7       : 52    2:66          10      : 77
4       : 50    3:64          13      : 52
12      : 50    4:71          11      : 44
6       : 49    5:66          2       : 39
11      : 45    6:80          4       : 34
(Other):192     7:80          (Other):144
```


Posts by Month
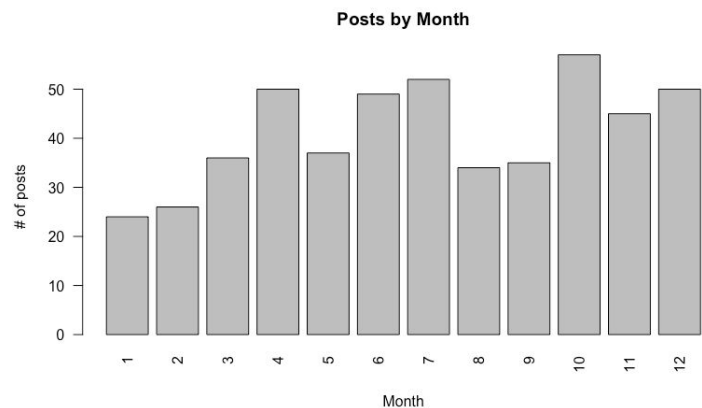
(b) Which "category" of post does the company most frequently make? Least?

As seen in the plot, the cosmetic company is making the most posts (~210) in the action category, and the least posts (~130) in the product category.
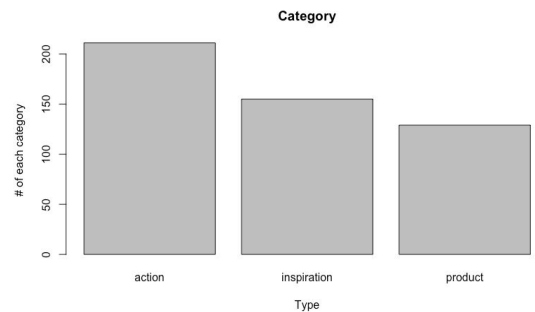
(c) What proportion of posts were paid for Facebook to advertise them?

There was a ratio of approximately 150 to 350 paid posts to non paid posts. The total number of paid posts is less than half of those that are non-paid.
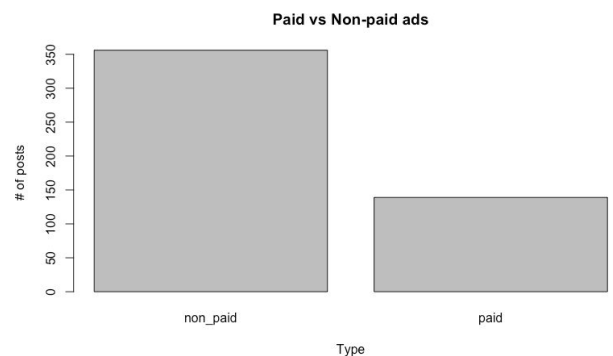

Section III: Bivariate Analysis

## (1) Create a scatterplot of shares against likes. What is the relationship between these two performance metrics?

The relationship between shares and likes is that there is a strong positive correlation. It has a correlation of roughly 0.9 (which is close to a direct positive correlation of 1.0). This means that as the number of shares increases so do the number of likes. The outliers were 5,172 likes with roughly 800 shares and the bulk of the data falls between 0-200 shares and 0-1000 likes.
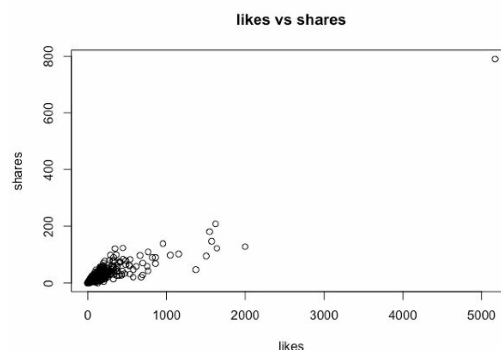
(2) *Using a boxplot, compare shares by month to examine how posts are performing based on when they are published. Consider adjusting the ylim parameter to get a more detailed look. What is a key insight you could give the company from this plot?*

The biggest takeaway from this boxplot is that the averages for the number of shares are roughly in the same range (greater than 0, around the 25 range). Some other insights we found were that September has the highest outlier value occurring (~220 shares), March has the lowest mean average (very close to 0), and June is the only month without an outlier. February, May, and September seem to be the best months because they have the highest averages in shares compared to the other months.


Shares by Month
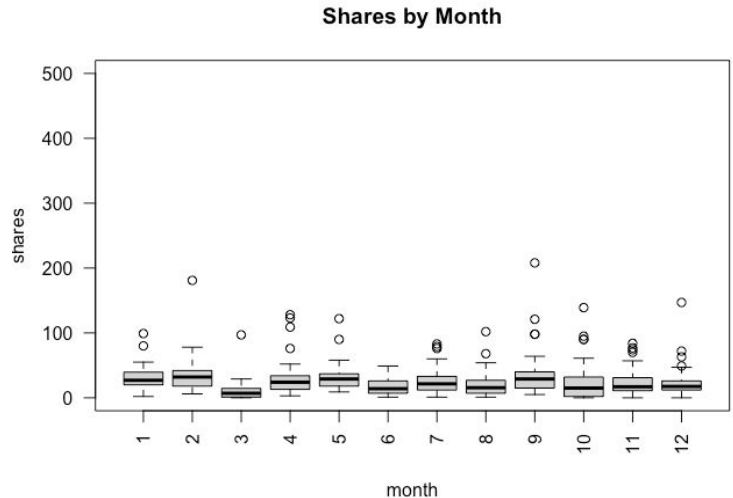
(3)

(a) *Use xtabs( ) and aggregate( ) functions to compare performance (shares) between paid and non-paid posts.*

Using the *xtabs( )* and *aggregate( )* functions, the non-paid ads perform better in terms of shares compared to the paid ad shares. The non-paid shares amount to 8978 while the paid adds total to 4518.

(b) *Compare performance between paid and non-paid posts using boxplots.*

The median of both the paid and non-paid are approximately the same, which means that they both performed equally well, however, there are more outliers when looking at the non-paid ads.


Shares for Non-Paid vs Paid posts
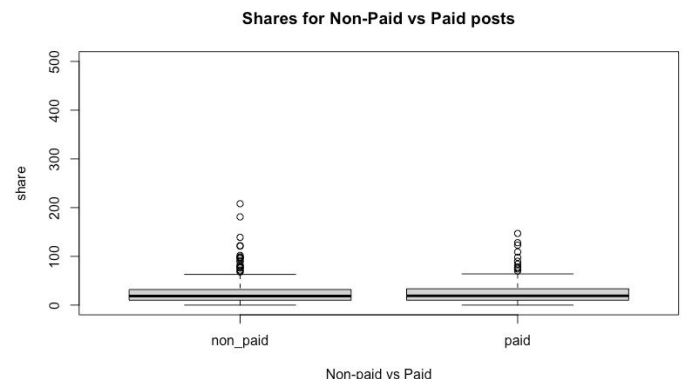
(c) *Based on your interpretation of these results, what would you recommend the company do with paid posts?*

The non-paid and paid have almost the same median and range of numbers. However, the non-paid have higher outliers overall. Since the non-paid are performing as well, we believe it would be a better option for the company to only run non-paid ads.

Section IV: Multivariate Analysis

(1)

(a) Create a 1x2 panel plot comparing shares between paid and non-paid posts for each category.
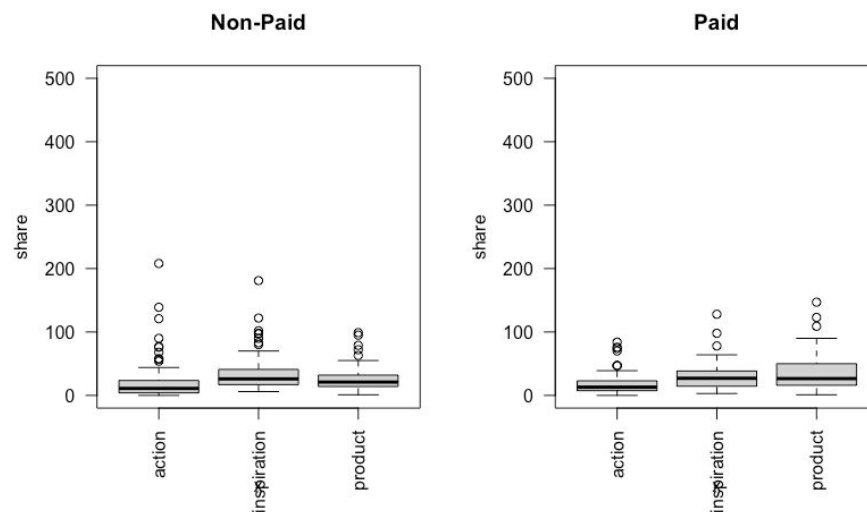
(b) What insight could you provide the marketing team from this?

By comparing the two box plots, we can see that the paid advertising on the brand's product posts result in a slight increase in the average number of shares. Besides that difference the action and inspiration posts have roughly the same number of shares. Some insights that we can provide the marketing team from these results are that they can focus on non-paid advertisements for the action and inspiration categories, and use paid ads for the product category. This decrease in paid ad spend will help them to allocate their budget for other costs.