In-Class Exercise: Multivariate Statistics and Plots

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: October 10, 2017

PDF file location: http://www.murraylax.org/rtutorials/facebook-plotsstats.pdf

HTML file location: http://www.murraylax.org/rtutorials/facebook-plotsstats.html

---

Note on required packages: The following code requires the `psych` package and the packages in the `tidyverse`. The `tidyverse` actually contains many packages that allow you to organize, summarize, and plot data. The package `psych` is used to perform statistics related to the median. If you have not already done so, download and install the libraries (needed only once per computer), and load the libraries (need to do every time you start R) with the following code:

```
install.packages("psych") # This only needs to be executed once for your machine
install.packages("tidyverse") # This only needs to be executed once for your machine
library("psych") # This needs to be executed every time you load R
library("tidyverse") # This needs to be executed every time you load R
```

Data Set

The following data set comes from the following study on Facebook marketing and performance metrics:

Moro, S., Rita, P. and Vala, B., (2016) "Predicting Social Media Performance Metrics and Evaluation of the Impact on Brand Building: A Data Mining Approach" *Journal of Business Research*, Vol. 68, pp. 3341-3351. Available at http://www.sciencedirect.com/science/article/pii/S0148296316000813

Download and load into memory the data set (see other formats at end of document):

```
load(url("http://murraylax.org/datasets/facebook.RData"))
```

The data set includes statistics from 500 Facebook posts in 2014 related to the marketing of a globally known cosmetic brand. Facebook marketing is an important part of many businesses marketing strategy. Facebook interaction can help businesses build their brand and market new products. Marketing executives such statistics to better understand the effectiveness of their Facebook marketing.

The data set includes the following variables:

1. **Type**: Scale / Class: Nominal / Factor. Type of post. Possible outcomes are "Link", "Photo", "Status", and "Video"
2. **Month**: Scale / Class: Ordinal / Ordered factor. Month of the year for the post.
3. **Weekday**: Scale / Class: Ordinal / Ordered factor. Day of the week for the post.
4. **Hour**: Scale / Class: Ratio / Integer. Hour of the day - between 0 (12:00AM) and 23 (11:00PM)
5. **Paid**: Scale / Class: Binary / Integer. Dummy variable equal to 1 if a paid post, 0 if a free or unsolicited post.

6. **Reach**: Scale / Class: Ratio / Integer. Number of unique individuals who saw the post appear on their news feeds.
7. **Impressions**: Scale / Class: Ratio / Integer. Number of times the post appeared on people's news feeds (some individuals may have had the post appear more than once)
8. **EngagedUsers**: Scale / Class: Ratio / Integer. Number of unique individuals that clicked anywhere in the post.
9. **Comments**: Scale / Class: Ratio / Integer. Number of comments on the post.
10. **Likes**: Scale / Class: Ratio / Integer. Number of likes for the post
11. **Shares**: Scale / Class: Ratio / Integer. Number of shares for the post
12. **Interactions**: Scale / Class: Ratio / Integer. The sum, Comments + Likes + Shares.
13. **Weekday.Int**: Scale / Class: Ordinal / Integer: Number associated with day of the week in **Weekday**
14. **Month.Int**: Scale / Class: Ordinal / Integer: Number associated with month in **Month**

Exercises

1. Create a bar graph with error bars (using normal distribution to compute confidence intervals) to illustrate the average number of engaged users by month. Based on your plot, identify the month that results in the most number of engaged users per post, and identify the months where there is statistical evidence that they result in a lower number of engaged users.
2. Suppose your audience is a group of marketing professionals that recently engaged in a Facebook marketing campaign a new product. They did this work in August. Highlight the bar for August with a different color than the others. Make the plot look pretty with colors of your choosing, no labels on the axes, a title, and no legend.
3. Is there evidence that mean number of engaged users from posts made in August are different than other months? Test the appropriate hypothesis.
4. Report the interpolated median and the median number of engaged users by month. Is there evidence that the median number of engaged users from posts made in August is different than other months?
5. Your marketing team is considering whether to create photo or video Facebook posts. Is there statistical evidence that one results in a higher mean number of interactions than the other?

To answer this question, create a sub-sample of just video and photo posts, as follows:

```
df.sub <- filter(df, Type=="Photo" | Type=="Video")
df.sub <- droplevels(df.sub)
```

6. Are the number of comments on a post correlated with the number of impressions? Test the appropriate hypothesis.
7. Illustrate the relationship between comments and impressions with a scatter plot and a best fit straight line illustrating the relationship. To make the best use of space, zoom in on your plot so that it just shows data with comments between 0 and 30 and impressions between 0 and 60,000. Give your plot a descriptive title.