Final Project Research Proposal

W203 Statistics for Data Science

By: Alejandro Pelcastre, Elias Saravia, and Shanie Hsieh

Research Question: How would changes to features like *comments, likes, shares, Month, Day, Hour, paid* (did someone pay for post to reach more people) and other variables affect our outcome variable, *total interactions*?

Dataset: The dataset we will be exploring to answer our research question comes from the UCI Machine Learning Repository. The dataset explores posts that were published on Facebook by a cosmetics brand in 2014. There are 19 variables and 500 observations contained in this dataset. Essentially, 7 variables are features prior to the post and the other 12 variables evaluate the post. The outcome variable we are using is *Total Interactions* and this seems to be a good metric to use for our model. The list of all variables within the dataset are:

- Page total likes
- Post info: Type, Category, Paid
- Post time: Post Month, Post Weekday, Post Hour
- Lifetime post effects: Lifetime Post Total Reach, Lifetime Post Total Impressions, Lifetime Engaged Users, Lifetime Post Consumers, Lifetime Post Consumptions, Lifetime Post Impressions by people who have liked your Page, Lifetime Post reach by people who like your Page, Lifetime People who have liked your Page and engaged with your post
- Post engagement: Comment, Like, Share, Total Interactions

Link to UCI ML Repository - https://archive.ics.uci.edu/ml/datasets/Facebook+metrics

Plan of Action: Our plan is to create a regression analysis to make sense of how changes in our independent variables alter our outcome variable, *total interactions*. We will address the linear model assumptions and convince the reader that our dataset satisfies the requirements. We will begin by preprocessing our data to make sure there are no complications with null values and erroneous data as well as one-hot encode any necessary categorical variables. Then we plan to create different lm models and compare them with an F-test to see whether some variables are statistically significant in our model. The key independent variables we hope to look into include the post comments, likes, and shares. We would also like to create additional models that look at some of the other variables including page total likes, type, paid, lifetime post total reach, and lifetime post total impressions. We can further explore any potential additional variables and omitted variables along with their bias in relation to our model.

Statistical Methods and Visualization Techniques: Some examples of potential visualizations that can supplement our analysis are histograms and scatterplots for variables with continuous values during our Exploratory Data Analysis (EDA) Phase. We can further analyze the statistics in our model such as the residuals vs. fitted values to evaluate the large-model assumptions. Lastly, we will create a regression table and discuss the statistical significance of the coefficients in our model in relation to our product.