# Final Exam - Version B

### What You have Learned about Exploring Variation

Let’s start with a data set called **twitter\_presidents** from [FiveThirtyEight](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/). If you want to know more information about these variables at any time, google: **twitter\_presidents r documentation**. Also remember that you need require(fivethirtyeight) to see this data in our DataCamp Sandbox.

There is a theory (put out by pop culture journalists) that in the twitterverse, **retweets** and **favorites** are signs of “good” engagement where people are agreeing with your tweet. However, **replies** are considered “bad” engagement. “The lengthier the conversation” sparked by a post, “the surer it is that someone royally messed up,” Luke O’Neil wrote recently [in Esquire](http://www.esquire.com/news-politics/news/a54440/twitter-ratio-reply/). But some people believe that any engagement with tweets is good engagement, so more of all of these (**retweets**, **favorites**, and **replies**) are good!

1. Go ahead and use R to look at the first six lines of this data frame. What are the cases in this data frame?
2. Make a jitter plot to look at the variation in **replies** (how many times a post was retweeted) by **user** (either @BarackObama or @realDonaldTrump). Write R code here:
3. By eyeballing this visualization, what is the relationship between **replies** and **user**? *[Not graded for correct/incorrect response but response required.]*
4. Make a visualization to look at the variation in **replies** by **favorites**. Write R code here:
5. By eyeballing this visualization, what is the relationship between **replies** and **favorites**? *[Not graded for correct/incorrect response but response required.]*
6. Just out of curiosity, who had the most favorited tweet in this data set?

As statisticians (you are a statistician now!), we want to **explain** the variation in engagement with tweets. Let’s say we have two hypotheses.

**User Hypothesis:** Knowing the **user** would help us explain more of the variation in **replies**.

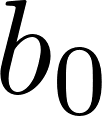
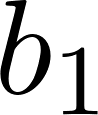
**Favorites Hypothesis:** Knowing the number of **favorites** would help us explain more of the variation in **replies**.

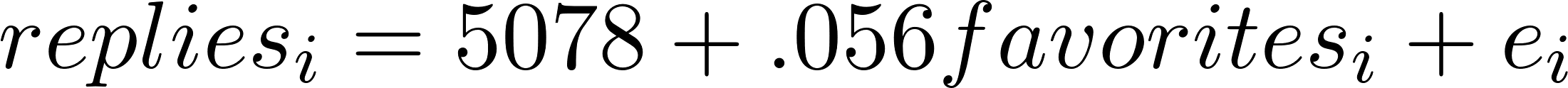
1. Write a word equation to represent the **user** hypothesis:
2. Write a word equation to represent the **favorites** hypothesis:
3. In your word equations (above), label which is the outcome and which is the explanatory variable.

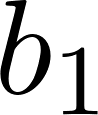
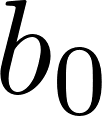
### What You have Learned about Modeling Variation

1. Here is a jitter plot of the user hypothesis here. Highlight the empty model in blue and the predictions of the **user** model in another color.

### 

1. In the jitter plot above, pick an Obama tweet and a Trump tweet and draw residuals from the empty model (in blue) as well as residuals from the user model (in the other color).
2. Now let’s create a statistical model of the **user** hypothesis. Write the best fitting model in General Linear Model (GLM) format.
3. What does your [](https://www.codecogs.com/eqnedit.php?latex=b_0%0) mean?
4. What does your [](https://www.codecogs.com/eqnedit.php?latex=b_1%0) mean?
5. Dr. Ji found that the best fitting model to explain variation in **replies** using **favorites** is this:

[****](https://www.codecogs.com/eqnedit.php?latex=replies_i%20%3D%205078%20%2B%20.056favorites_i%20%2B%20e_i%0)

She argues that .056 is a pretty small [](https://www.codecogs.com/eqnedit.php?latex=b_1%0) compared to the [](https://www.codecogs.com/eqnedit.php?latex=b_0%0) so this model explains .056 of the variation in **replies**. You suspect that her argument is wrong. What statistic(s) might tell you whether this **favorites** model explained more variation or the **user** model? [Circle one answer below.]

1. PRE
2. F
3. SS Model
4. All of the above
5. Run R code to figure out which is the better model, **user** model or **favorites** model. Which model is better? Support your answer with statistics from your R code (but you do not have to write the R code here).

### What You have Learned about Evaluating Models

The **user** model tells us about the relationship between **user** and **replies** *in the sample data.* But we want to know if this relationship exists in the Data Generating Process.

1. The sample F for this model is 2405. Is it *possible* that we could have gotten this F value if there was no relationship between **user** and **replies** in the DGP?
2. Let’s use randomization (**shuffle**) to replicate what we found in the supernova table.
   1. To start, let’s create a sampling distribution with 10,000 randomized Fs. Write that code below.
   2. Also save the sample F as **sampleF**. Write that code as well.
3. Draw a picture of the distribution triad here. Include a drawing of the resulting sampling distribution of Fs and label where the p-value is.
4. What does your p-value mean?
5. Go to this website to submit the results of your analysis (from the Evaluating Models section) in APA style.

**http://bit.ly/finalexam21**

=====================================================================

## Ask Dr. Ji Coupon

Use this coupon to ask Dr. Ji about whether you have the right or wrong answer for any 1 question. (Props to Asha for thinking of this coupon!)

If this has been signed, coupon is VOID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_