### What You have Learned about Exploring Variation

Let’s start with a data set called **candy\_rankings** 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: **candy\_rankings r documentation**. Go ahead and use R to look at the first six lines of this data frame. Write code here:

What are the cases in this data frame?

Take a look at the variation in **winpercent** (what percentage of time a particular candy won when matched up against other candies in a head to head contest). Write R code here:

Just out of curiosity, you might be interested in the candy that had the lowest **winpercent**, the *worst* candy according to this data. How would you find that out? Write the R code here:

What is the worst candy in this data set?

As statisticians (you are a statistician now!), we want to do is **explain** the variation we see here. Why are some candies considered better than others? Why are some candies disdained? Take a look at the **categorical variables** in this data set and come up with one possible story. Describe your story and also write it as a word equation.

In your word equation (above), label which is the outcome and which is the explanatory variable.

Let’s explore the relationships between the outcome and explanatory variable in the data. Make a visualization that would help you explore this relationship and write the R code here:

In this situation, what would it mean to “explain variation”?

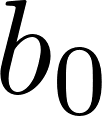
Just from eyeballing your visualization, do you get the sense that your explanatory variable “explains” variation in the outcome variable well? Why or why not?

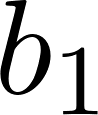
### What You have Learned about Modeling Variation

Sketch a faceted histogram of your hypothesis here. Draw in the empty model in blue. Draw in your complex model in another color.

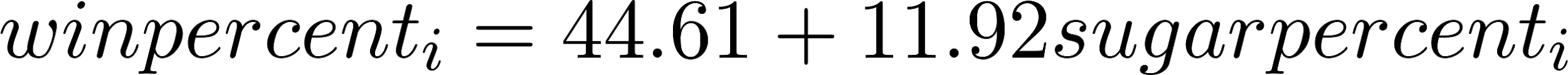
Now let’s create a statistical model of your story. Write the R code below to find the best fitting model.

Write the best fitting model in General Linear Model (GLM) format.

What does your [](about:blank) mean?

What does your [](about:blank) mean?

Dr. Ji surmises that sugar content might explain a lot of the variation in **winpercent** and reports that the following is the best fitting **sugarpercent** model:

[](about:blank)

She argues that 11.92 is a pretty big coefficient so this model must explain more variation than your model. You suspect that she is wrong. What statistic(s) might tell you whether this **sugarpercent** model explained more variation or your model?

Write the R code to get the statistics that tell you how much variation has been explained by your story.

Which is the better model, your story or Dr. Ji’s **sugarpercent** story? Support your answer with statistics that you got from running the code above.

### What You have Learned about Evaluating Models

Your model tells us about the relationship between **winpercent** and your explanatory variable *in the sample data.* But we want to know if this relationship exists in the Data Generating Process.

Could we have gotten the PRE you got if there was no relationship between the explanatory and outcome variable in the DGP?

Let’s use randomization (**shuffle**) to replicate what we found in the supernova table. To start, let’s create a sampling distribution with 10,000 randomized PREs. Write that code below. Also save the sample PRE as **samplePRE**.

Draw a picture of the distribution triad here. Include a drawing of the resulting sampling distribution of PREs and color in the p-value in a different color.

What is p-value? (Describe the picture above.)

Write up the results of your analysis here in APA style.