Judson Dunaway-Barlow
HW 6
Data Science
3/13/18

Goose Egg Data:

Link: https://github.com/fivethirtyeight/data/blob/master/goose/README.md

This data contains information about relief pitchers in Major League Baseball starting in 1921. The data was collected using game records stored on RetroSheet and calculated based on play by play information. The data is looking to figure out which relief pitchers have been most effective. Traditionally in baseball, the stat "save" is used to determine good relievers, but it has a number of limitations. First of all, only the last pitcher of the game can earn a save, which eliminates other pitchers who pitch in just as high leverage situations but don't "close" the game from earning recognition for their just-as-valuable work. The new statistic calculated is called a "Goose Egg", which is based on the name of a relief pitcher that leads all baseball history with most goose eggs.

My question would be whether or not pitchers were used correctly. Historically in baseball, the "closer" was supposedly the best relief pitcher. However, were there other pitchers that performed well that didn't get that "closer" status? Or were there closers that actually weren't top performers? The question could be answered in a number of ways, one of which would be to scatterplot goose eggs or goose egg percentage (y) vs saves (x). That way, you could see the pitchers who lived up to their hype (lots of saves and high goose egg percentage), pitchers who could have been overrated (lots of saves, lower goose egg percentage), and pitchers who were underrated (few saves, high goose egg percentage).

Gun Control Data

Link: https://github.com/fivethirtyeight/data/tree/master/poll-quiz-guns

This data contains information about public opinion on different types of gun control measures. The data was collected by a large variety of pollsters in polls of registered voters in the wake of the Stoneman Douglas shooting. It asks questions about raising the age to buy to 21, increasing background checks, or even banning all weapons completely.

A couple questions could be asked here. First of all would be whether party affiliation affects opinion. Clearly it does, and the 538 article shows this, but I think it would be interesting to visualize it differently. More specifically, it would also be interesting to see if the perceived bias of each pollster (i.e. general political lean of its reader base for news organizations) impacted the party line breakdown. For example, a Fox News poll on a specific question might get more conservative leaning answers than one from NPR. This could be visualized by scatterplotting lean vs response tendencies.

New Bechdel Test Data

Link: https://github.com/fivethirtyeight/data/tree/master/next-bechdel

This data contains information about diversity in Hollywood films. The Bechdel Test looks to see if a movie has two women talking to each other about something other than men as a test of diversity. According to the article, only about half of films pass the test. The article looked to find other similar tests for a more modern take on diversity (e.g. race). The other tests look for more than just a single scene towards the full cast, crew, and more granulated data like how often the women appear.

One question that could be asked is if there's a correlation with passing the tests and how much money a movie makes. You could run a machine learning algorithm to figure out which (if any) of the tests are more likely to correlate with more audience interest.