Welcome to the final exam.

First thing you'll need to do is familiarize yourself with the game of baseball if you're not already. I'd suggest a cursory google search on 'rules of baseball.' For this exam we will be examining pitching data from the 2018-2019 major league baseball season. An actual student and I worked on this dataset for the Milwaukee Brewers.

Some basics on the data. When a pitcher throws a pitch there are essentially five outcomes.

1. The batter swings at the ball and puts it in play
2. The batter swings at the ball and puts it out of play resulting in a foul ball
3. The batter swings at a ball and misses it resulting in a swinging strike
4. The batter does not swing at the ball and it's called a strike by the umpire
5. The batter does not swing at the ball and it's called a ball by the umpire

For those avid baseball fans I realize there are a few other possibilities (hit by pitch for example, or a passed ball) but these are pretty rare occurrences so let's just ignore them.

In this truncated dataset you will only be seeing categories 4 and 5, we don't really care about what happened if the batter swings at the ball. We're interested in what factors result in an umpire calling a pitch a ball vs. a strike.

Your response variable in this data is "description." There are only two possible values, called_strike and ball. Each row of data represents a pitch in which the batter did not offer a swing. You will also find the following covariates…

release_speed: speed of the ball when it leaves the pitchers hand
pitch_type: The type of pitch thrown by the pitcher (you'll need to research what these are if you're interested)
plate_x: The horizontal coordinate of the ball as it passes over the plate
plate_z: The vertical coordinate of the ball as it passes over the plate
strikes: The number of strikes the batter has during the current pitch
balls: The number of balls the batter has during the current pitch
home_score: Runs aka score of the away team at the time of pitch
away_score: Runs aka score of the away team at the time of pitch
pitcher: A unique id that represents the pitcher throwing the pitch

Now the great news about baseball data. There are literally hundreds of interesting analyses that we could do with this data. I'm going to give you some general guidelines and then you can pick which analysis you would like to do. ONLY DO ONE! Also, feel free to make up your own.

For each of these analyses you should be focusing on the factors that result in a strike being called instead of a ball. Description should be your response variable.

You will need to focus on the importance/significance of a few of these variables on the probability of a strike being called, for this it's recommended you create a statistical model(s).

BIG HINT! The spatial components (plate_x and plate_z) are going to dominate every single model you create giving your other variables the appearance of having no effect. Plot strikes and balls as different colored points with plate_x and plate_y as your axis. You'll see that there are regions where there are always strikes and regions where there are always balls. I call these 'obvious,' zones. Perhaps we should cut these out of our data and focus on the fringe zone where an umpire actually has to make a judgment call.

Now for the analyses:

1. I've always believed that umpires are more likely to call a fringe pitch as a strike if there are already 3 balls on the batter. Assess whether or not the count (number of strikes and number of balls) impacts the probability of a strike being called if all pitch conditions are comparable.
2. Some pitchers (and catchers by extension) seem to have a wider strike range, meaning that they get more strikes called in similar situations than other pitchers. Mine the data to assess the validity of this 'conventional' baseball wisdom. Are there a few pitchers in particular that stand out? If you choose this analysis you might want to truncate your dataset to only include pitchers that have thrown more than 20 pitches.
3. If I was an umpire I would just start calling more strikes once the game seems out of hand to try to speed up the inevitable end of the game. Create a variable called point differential, which is the difference of scores. Is there a 'shelf,' in score differential in which strikes become more systematic in comparable pitch circumstances?
4. (Maybe a bit challenging). In baseball there is a pitch called 'high heat.' It's a fast-ball (that's the pitch type) that is too high to be a strike. 'Conventional,' baseball wisdom is that 'high heat,' pitches are much more likely to be called strikes than other pitches similarly outside of the strike zone. Assess this conventional baseball wisdom using the data.

Hope you guys have fun with the data! Looking forward to reading these. Lots of office hours next week, let me know if you need me.

Exam is due at 11:59 PM on Sunday May 21st. Legally I cannot grant any extensions.

I will be in zoom virtual office hours on Monday, Wednesday and Thursday next week from 5:30 – 6:30. If you can't make those times I am also available by email appointment.