1. For the strike/ball prediction, my basic strategy was to use a classification model in order to try and predict the ball/strike outcome of the pitches. See also the comments left in the python notebook: ‘Strike Prediction’
   1. The first step I took was to do some data exploration and cleaning, first in Excel, and then in Python. I used Excel in order to simply determine the scale of the missing values in the data. I noticed that there were a non-trivial amount of missing values in the dataset, but mostly only one value per line, with the exception of the umpire/catcher combination of '9c6cbb5e', where nearly all of the code was missing. Because the model I ended up using was a random forest regression, I chose to use a placeholder value of 999 for the missing values. This is a sub-optimal approach. A better one would have been to identify pitches that were similar to the ones with missing values using something similar to a K Nearest Neighbor approach and take an average of several of the most similar pitches to come up with an approximate value. With my time constraints (and the fact that I had done that only a few times before) I chose the simpler, but less effective, method.
      1. In the data exploration phase I was sure to determine what was the location of the pitch as it crossed the plate by graphing the data. Graphing the data is a strong reasonableness check, and also helped identify what should be relevant features.
   2. After cleaning the data for input into the model I chose to try two models, a Random Forest Classifier, and a Logistic Regression. Because the task at hand was classification, I chose what I thought to be two relatively simple, but effective classifiers. The Random Forest performed better on the training data set.
      1. I did not spend as much time as would have been ideal on feature selection. Removing some of the features would likely have led to a stronger, simpler model with fewer inputs.
   3. I then imported and similarly cleaned the dataset for the test, and ran the fitted model on that data.
2. See attached report ‘Catcher Report’. Note that I assumed for the purposes of this report that video links would be available for the pitches that I identified as having been well or poorly framed. This would be more useful to a coaching staff than simple plots of the locations or more detail about what % above or below average he was, so I kept my report simple and tried not to get bogged down in what would likely be unhelpful detail.