Simran Pabla

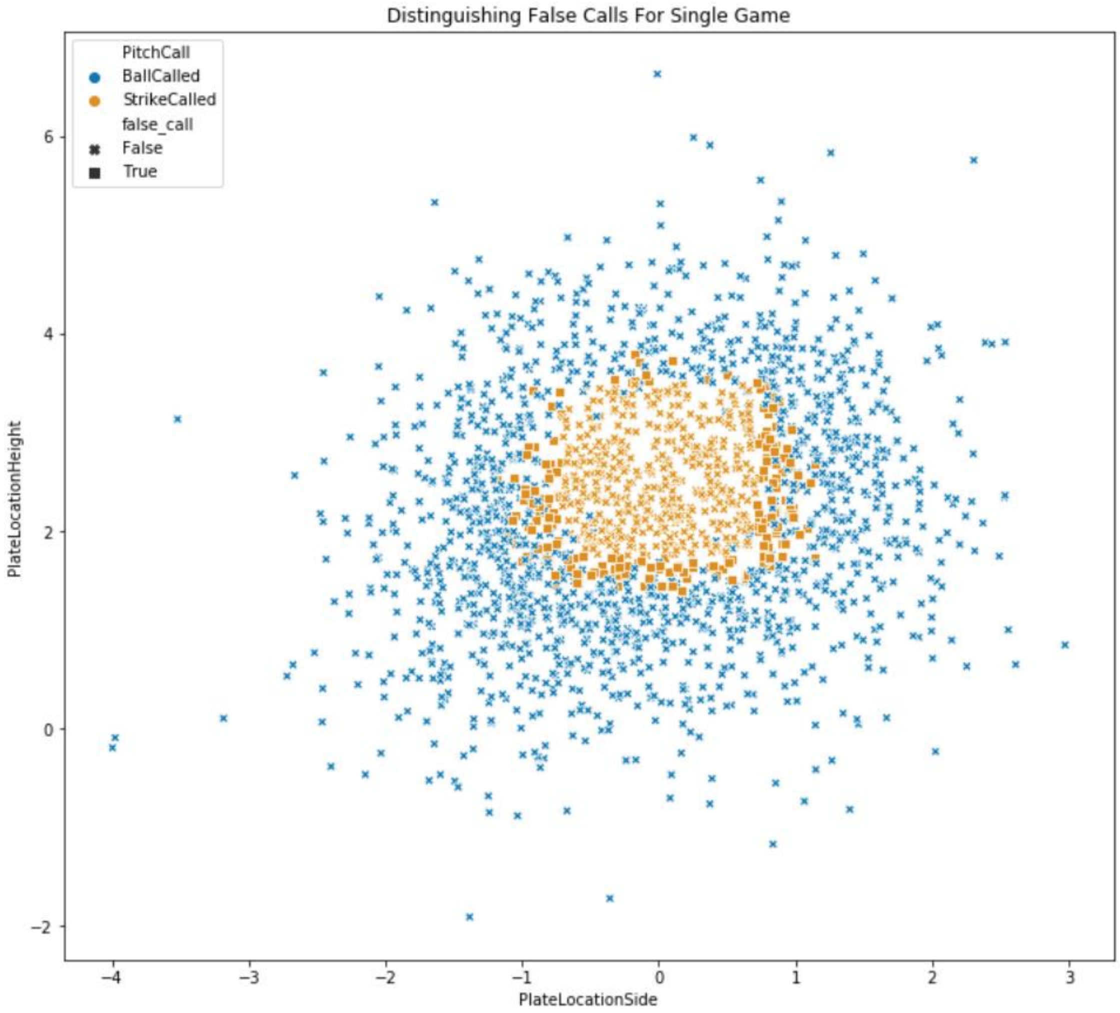
Individual Deliverable #1

The intent of my initial probe into the data was to 1) identify potential gaps in the data and 2) specifically pay closer attention to the time features in the dataset. To achieve these goals, I broke down the visualizations into iterative steps.

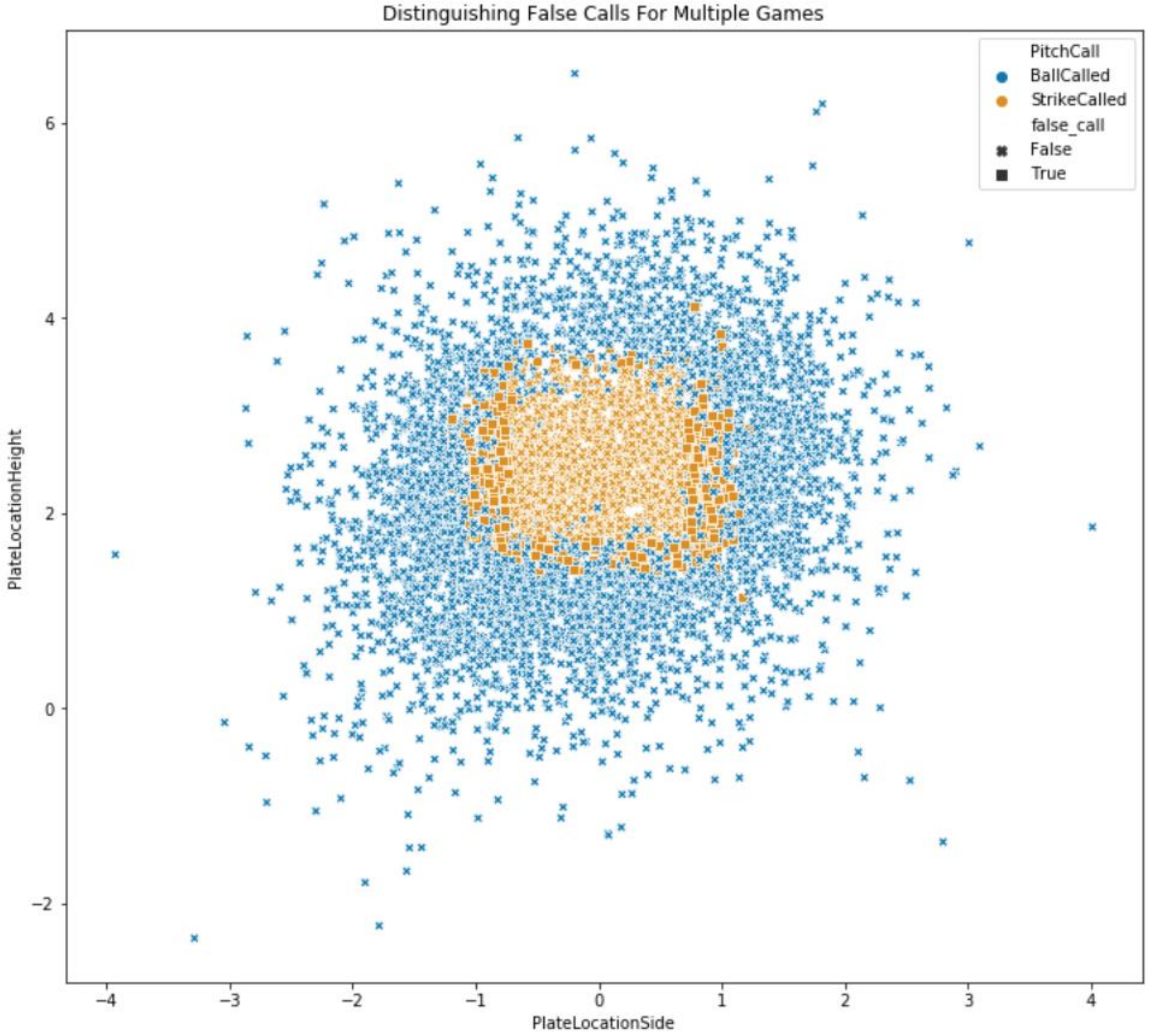
Some assumptions made in each of these charts are defined as follows:

- The strike zone is estimated to be 8.5 inches in either direction from the center of the plate in width and ranging from 1.75ft to 3.5ft. Already, this is a source of inaccuracy, as the values aren't consistent among all players.

First, I plotted the calls for a single game and distinguished calls visually. The false calls are represented as squares and the correct calls are crosses. As expected, a box forms along the border between the strike and ball calls.

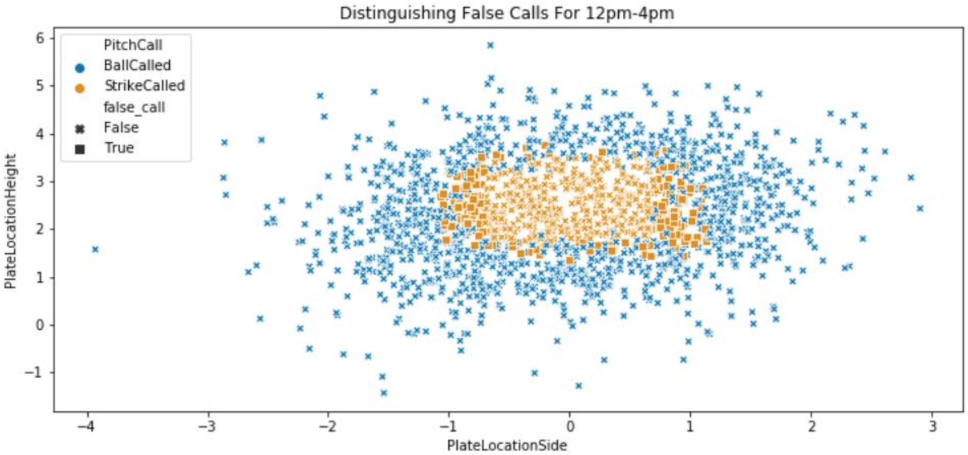


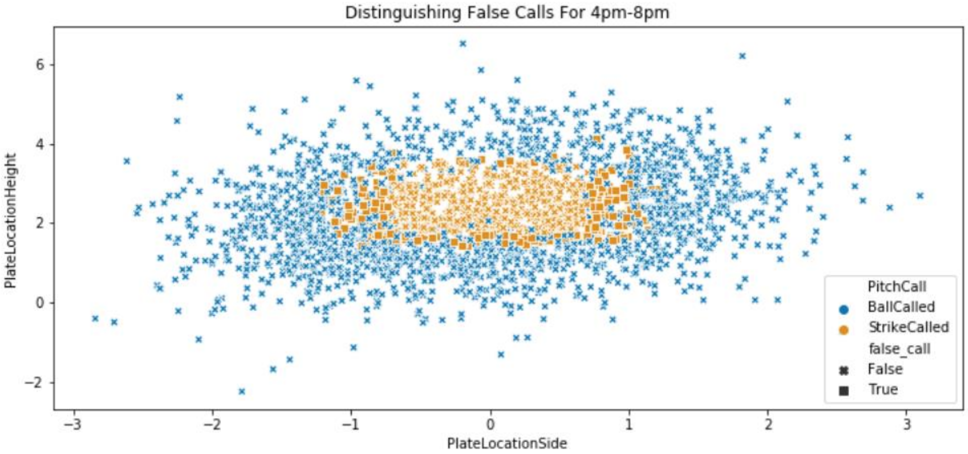
To expand on this, I extended to games over three separate days: 2016-10-02, 2-16-10-01, and 2016-09-30.

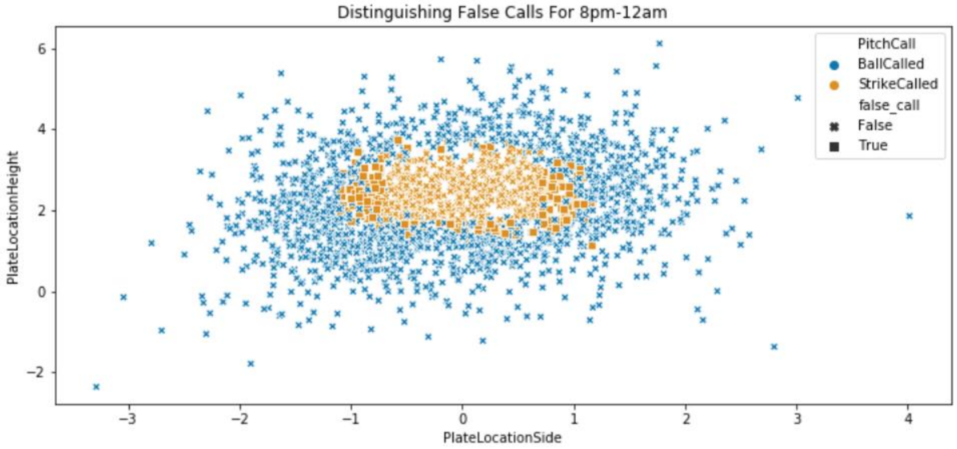


The dataset in its current form lacks a field for game identification, making it more difficult to assess the number of games represented within any given subset of the data. To mitigate this, I chose three days and estimated game boundaries such that each day had three games: 12-4pm, 4-8pm, and 8pm-12am were the three bins I generated for the games. This also forced an assumption, as time was represented as a 12-hour clock rather than 24-hour. I gathered additional statistics regarding these bins.

I began by plotting the calls for each of these three “times of day” separately.

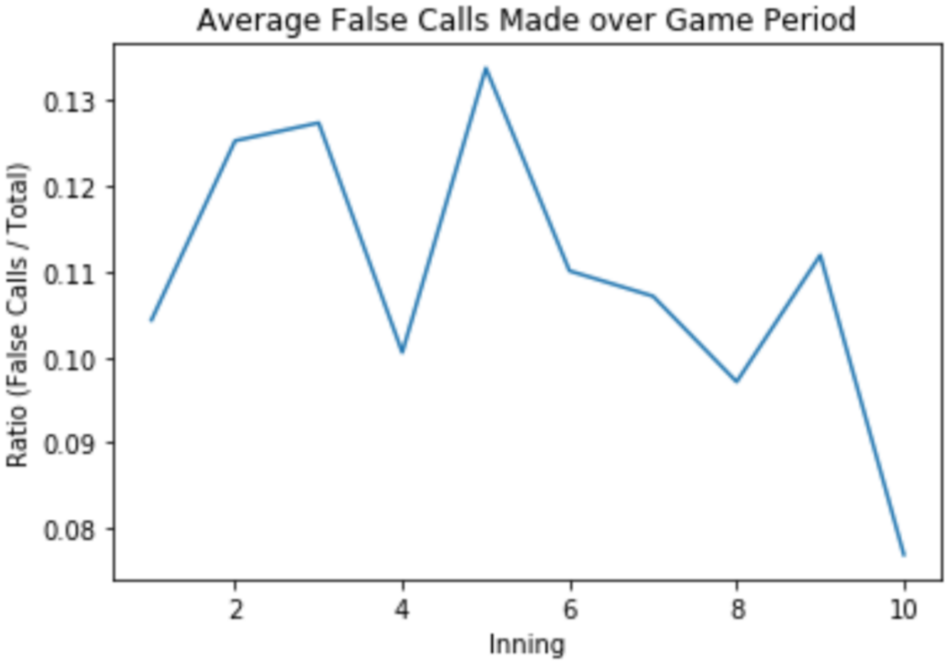

Distinguishing False Calls For 12pm-4pm


Distinguishing False Calls For 4pm-8pm


Distinguishing False Calls For 8pm-12am

Visually, these three graphs proved to be relatively similar, though I’d like to conduct further analysis regarding the significance of the widths of the average strike zones established in the estimated three games. Without further analysis, it’s difficult to gauge whether the tighter nature of the strike zone in 8pm-12am is significant or not.

Both time and time elapsed within a game are interesting factors that we as a team are curious about analyzing further. Using the data that I sifted through above, I generated a preliminary plot of the average ratio of false calls to total calls for each inning.



With this chart, including more samples will be beneficial.

Through this initial exploration of the data, I identified possible directions that we may be able to explore over the next few weeks, as well as certain shortcomings in the data, and hopefully our team will begin to explore the ideas and shortcomings we've collectively found.