Reds Dew Point

My overall approach was to evaluate how the expected (predicted) horizontal and vertical movement related to the actual. Since high humidity results in less dense air, I made the assumption that high dew point would mean higher horizontal and vertical break. I would construct a p-value to get a probability and subtract that p-value by 1 to obtain the probability that there were indeed outside factors.

I started by organizing and manipulating the data. The first thing I wanted to do was make handedness irrelevant by multiplying the horizontal features for lefties by -1. Next, I wanted to add some features I thought could be important. I believed that how much a pitcher has thrown would be important. I created columns for the number of pitches in that inning for the pitcher and the number of pitches in the game for the pitcher. Additionally, I created a column for the number of innings that pitcher had thrown in up to that point in the game.

Two big factors in expected horizontal and vertical break would be the pitcher and the pitch type. These were both categorical data. Starting with the pitch types, there were few enough of them that it made sense to one-hot encode that variable. After investigating, there were a few pitches marked 'UN' for unknown. I decided to push these to the closest pitch type they resembled. This was based on the specific pitcher's pitch types and their average movement, spin, and speed profiles. After changing these, I could one-hot encode the pitch types. The pitchers also did not have that many distinct people, so for this dataset it could make sense to one-hot encode, but in the world outside of this dataset, this dimensionality would get too extreme. I instead decided to add a variable that was the pitcher's mean movement profiles for that pitch type.

An important part of comparing the predicted and actual movement profiles will be the standard deviation. I created a column for the standard deviation of every pitcher-pitch type group. However, after investigation of these groups, I found some had very small counts. I decided to take the average standard deviation of each pitch type across all pitcher-pitch type group that had a count of at least 10. If the pitcher-pitch type group had less than 10, I gave it the average standard deviation. This excludes the knuckleball, which I left alone, as it only had 3 examples.

Finally, I created independent linear regression models with XGBoost for horizontal and vertical break. I calculated the z-score of the prediction from the actual using the prediction as the "mean" and the actual standard deviation as the standard deviation. Using the predicted value as the mean and the actual standard deviation, I thought I could shift the distribution to match the features that were involved in the prediction. This way I can take into account features other than dew point that would be affecting the movement, such as release point metrics. Combining these into a chi-square statistic, I could find a joint p-value. If the either z-score was less than 0, I pushed the p-value to 1. My probability that the dew point affected the pitch was 1-p.

The other route I thought about for this project was looking at the release point metrics, as the dew point's effect on comfortability would likely show up there. I concluded that either way, the effect would show up in the movement profiles.