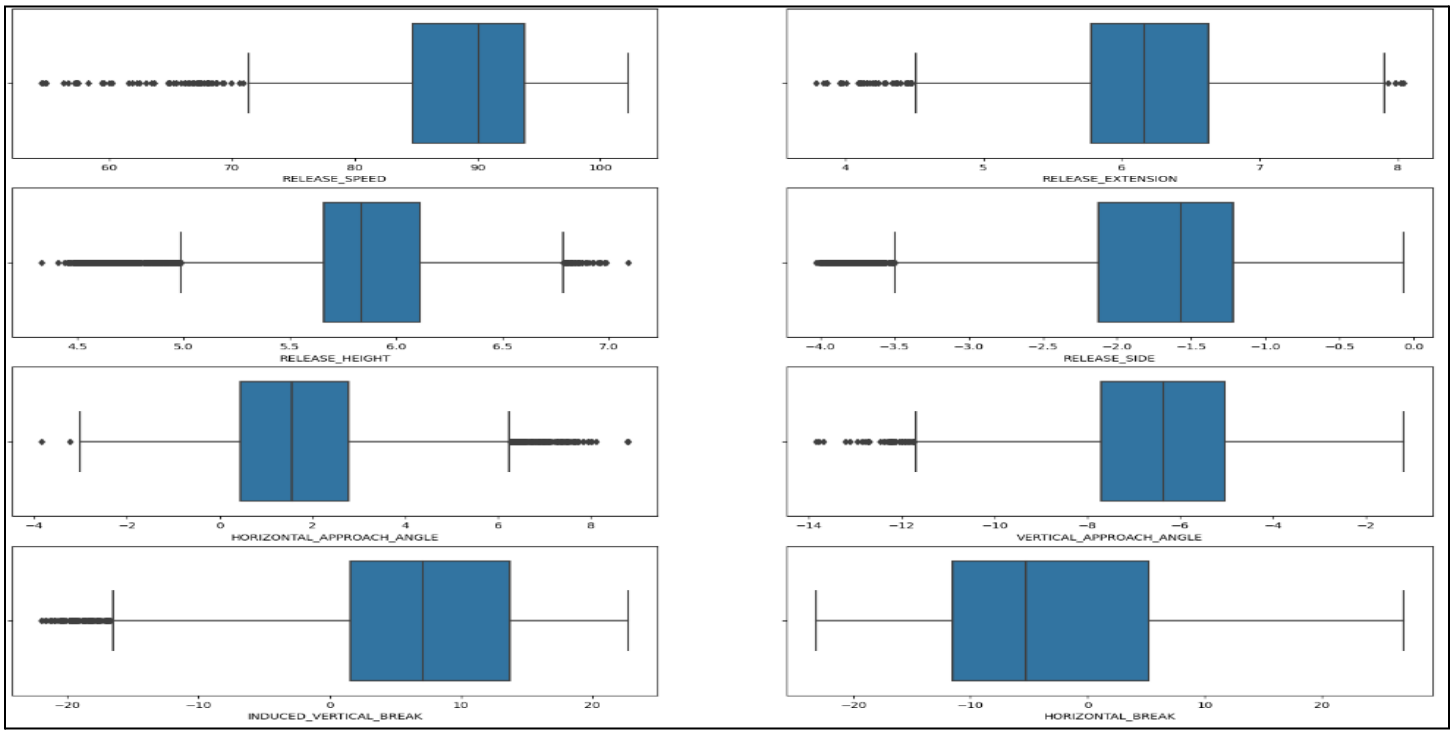
This write up consists of my findings into how weather, in this case dew point, can alter pitching in a baseball game. I used the pitch data provided in data.csv to account for and identify the probability of a pitch being affected by dew point greater than 65 degrees F.

Data Exploration & Preprocessing

The data provided consists of 9889 pitches thrown by 37 pitchers at the Great American Ballpark with 26 columns comprising game events and pitch characteristics. The first thing I noticed was that some of the pitch characteristics varied based on the handedness of the pitcher. Therefore, I wanted to standardize them by considering *THROW_SIDE_KEY* as 'R', and for that I flipped the sign for *HORIZONTAL_BREAK*, *RELEASE_SIDE* & *HORIZONTAL_APPROACH_ANGLE*. After this I checked for NaN values where only the column *EVENT_KEY* had 7258, which makes sense since it's pitch-by-pitch data for at-bats in multiple games.

We then check for outliers present in the data using boxplots and describe functions in Python. Even though there are a number of data points outside the boxplots min & max region, these can be easily due to the complex nature of baseball and it wouldn't make sense to just remove these data points. For example, some pitchers (like position players) may throw the ball exceptionally slow and other pitch types might require a slower release speed. Similarly, the arm slot and the pitcher's height may vary a lot, resulting in multiple outliers in *RELEASE_HEIGHT*. However, it is good to know that these outliers are not anything extraordinary, like throwing a pitch at 130 mph. From the KDE plot (in the notebook), we can also see that the distribution for most of the pitch characteristics is definitely not normal.

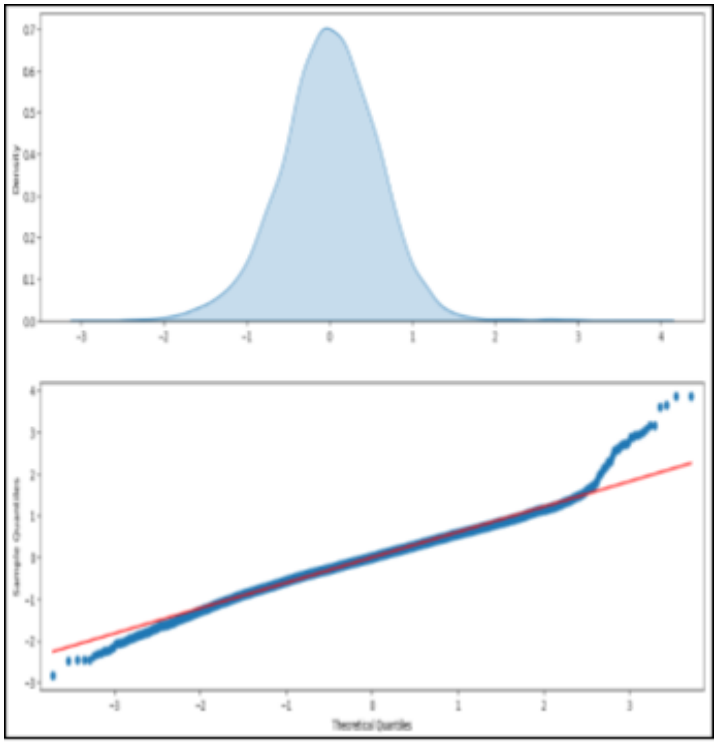


Methodology

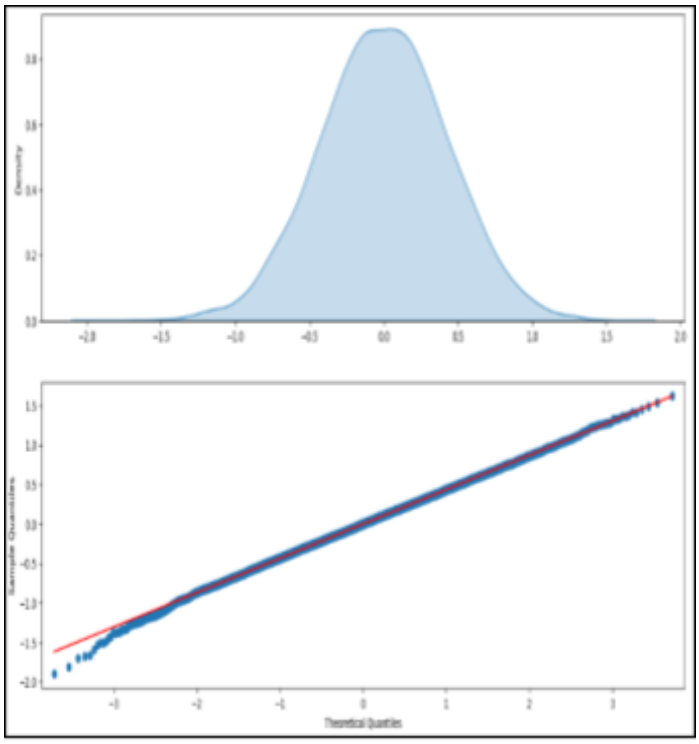
When I received the assessment, I was a little confused over the fact that there was no response variable to compare against and build a machine learning model. But then, it may also not be a machine learning question. So, how do we find the pitches affected by dew point without an actual target variable? This is when I saw the hidden hint mentioned in the documentation. The statement "pitch movement due to environmental factors" made me realize that along with gravity and air resistance, we may also have dew point affecting Induced Vertical Break and Horizontal Break. Therefore, we would be able to discern this effect of dew point over pitch movement by accounting for other factors already provided in the data, such as, the pitcher, release characteristics, approach angle, etc.

We can set up a mixed effect model with the release characteristics, approach angle, spin rate being the fixed effects and pitchers & their pitch types being the random effects. We will model the fixed and random effects against Induced Vertical Break as well as Horizontal Break. We can then interpret the residuals from these models, where a high residual may suggest that the pitch was affected by something not included in the model, possibly dew point. After finding the residual, we can use that to calculate the probability of dew point affecting each pitch, assuming that the residual is present likely due to dew point. We calculate the mean and standard deviation of the residuals and use that to calculate z-score for each pitch. Since the probability of a pitch not getting affected by dew point would be given by the cumulative distribution function (CDF) of the normal distribution at that z-score, we take the complement to find the probability of a pitch being affected by dew point > 65 degrees F.

To implement this model, I built a class Model because I needed to implement it for both Induced Vertical Break and Horizontal Break. The class contains functions for scaling data, building models, getting residuals and probabilities for each pitch. You can pass the formula and the groups you want as parameters for your mixed effect model.



Induced Vertical Break



Horizontal Break

I also wanted to check for normality of the residuals graphically to ensure the assumption made by the mixed effect model. From the above Kernel Density and QQ plot, we can see that the residuals for the Induced Vertical Break model exhibit slightly heavier tails than a perfect normal distribution, also evident from the QQ plot with the right tail having a higher deviation. However, for the Horizontal Break model, it appears to follow the normal distribution more closely, with only any deviation of the left tail.

<brain not braining but i will be back>

References,

1. https://www.baseballvmi.com/baseball-humidity#:~:text=Humidity%20actually%20has%20little%20effect,through%20than%20dry%20air%20is.

2. https://www.daytondailynews.com/weather/does-humidity-lead-more-home-runs-baseball/oP4c9mGNHtBAo1gYDpFODJ/

3. https://www.pythonfordatascience.org/mixed-effects-regression-python/