Analysis of the effect of high dew points on baseball pitches
By: Jameel Kaba

Objective
Determine how much influence high dew points (high humidity) has on pitches thrown at Great American Ballpark in Cincinnati.

Process
Data Exploration
- Check the dataset for null/missing values
- Visualize the distribution of pitching factors which may be affected by humidity.
    - These are factors which are related to movement, and would potentially be impacted by a high dew point
        - Factors used: Vertical/Horizontal break, spin rate, release speed, and approach angles
    - Factors with varying distributions include vertical/horizontal break, spin rate, and release speed
    - The approach angles had a more normal distribution to it

Interaction
- Look at interactions between factors in order to potentially capture the more complex relationships by factors affected by humidity
- Interactions derived from data:
    - Spin-speed, spin-vertical/horizontal break, speed-vertical/horizontal break.

Model choice
- Because we are looking for the impact high dew points have on pitches, an anomaly detection model to determine whether a pitch was impacted will be used.
    - Choices considered: K-Nearest Neighbors, Local Outlier Factor, and Isolation Forest.
- We will use an isolation forest as we have a large dataset, and the anomaly we are looking at impacts the characteristics of pitches.
- Anomaly score produced from the model can be converted into the probability a pitch was affected by a high dew point
- We can use the isolation forest to predict whether the pitches in the dataset have been affected

Evaluation
- Visualize the distribution of probabilities to determine the likelihood of each pitch being affected by high dew points
- Many pitches were classified as normal, but a subset were classified as anomalies (affected by high dew points)
    - As a result, we will visualize the characteristics of these pitches that were classified as anomalies to try and draw our conclusions.

Conclusion
The results, based on patterns which are observed by the data and assumptions from the Isolation Forest model offers us insights into the influence humidity has on pitch behavior in Cincinnati.

Induced Vertical Break: leans more towards -10 and 0, with a spike occurring as we get closer to 20

Horizontal Break: peaks between -15 and 0 and again from 10 to 20

Spin Rate: most impacted occurs with around the 3000 range

Release Speed: between 80 and 83 is where most pitches are affected

Vertical and Horizontal Approach Angle: the distribution seems to retain it's shape when compared to against the visualization of the whole database.