To find the probability that a pitch was affected by a dewpoint greater than 65 degrees Fahrenheit, the following process was used. Firstly, the data set was checked to see if there were any missing values or duplicate rows. Analysis found that the only column to be incomplete was EVENT\_RESULT\_KEY which will not be relevant for the chosen analysis method, so it is acceptable to continue. After checking for duplicate rows, it was found that there are no duplicated rows and data preprocessing was considered complete.

The statistical method that was used to calculate the probability was to examine certain factors in the data set and to see how many standard deviations away from the mean they were. Based on the degree of variance, a probability was calculated according to a created formula. Four factors were chosen, being RELEASE\_SPEED, SPIN\_RATE\_ABSOLUTE, HORZONTAL\_BREAK, and INDUCED\_VERTICAL\_BREAK. These factors were selected as they are directly linked to the flight path of the pitch which could be affected by the humidity. Other factors are more indicative of the pitcher’s comfortability which could be affected by a multitude of other reasons, such as temperature, leverage, fatigue, general health, etc. Therefore, they were not considered for direct analysis for this problem. Using the four identified factors, they will be compared to the mean of the pitcher’s own statistics, as they are not consistent across the entire league.

After identifying the necessary factors, in order to perform standard deviation analysis, it must first be established that the distributions are Gaussian in nature. This was confirmed by identifying the pitches with the most data points. From the data set, it was found that there were four different pitches with over 1000 samples, being fastballs, sweepers, cutters, and changeups. The five pitchers that threw the most of each of these pitches were individually identified to gather to most samples possible. A histogram for each factor for each pitcher for each pitch type was then created to look at the shape of the distribution. It was found that across the league, the distribution of each of the identified distributions was Gaussian in nature, so it was interpreted that this trend would hold true for the pitches and pitchers with less total samples. Additionally, this type of data is expected to be normally distributed but was also visualized for confirmation purposes.

After confirming that each distribution was normal, it was reasonable to continue with standard deviation analysis. A data frame was created to store the deviation and probability values that were to be calculated also taking relevant identifiers from the original data set. The mean velocity, spin rate, horizontal break, and vertical break was then identified for each pitcher and each pitch that they threw. From there, the entire data set was iterated through and the standard deviation from the mean for each of these factors was recorded. For some pitches, the sample size was too small to gather standard deviation data, so they were skipped over from this portion of the analysis and will be analyzed later.

Using the standard deviations, an overall probability was calculated. Probability can range from 0 to 1, so a base value of 0.5 was selected. Of the four factors, they were weighted slightly differently, with velocity having a maximum value of 0.2, and the other three having a maximum value of 0.1. Therefore, it all factors take on their maximum values, this would bring the probability to its upper or lower limit of 1 and 0 respectively. Taking the base value, if velocity was within one standard deviation of the mean, the overall probability was increased by 0.05 if it fell above the mean and decreased by 0.05 if below the mean. This is because a greater humidity decreases air density, and increases velocity and spin rate, while decreasing break. The probability was increased/decreased by 0.1 if within two standard deviations, and by 0.2 if more than two standard deviations from the mean. This same logic was applied to spin rate (increasing probability if above the mean, decreasing if below), and both break variables (increasing probability if below the mean, decreasing if above). However, these variables were factored in as 0.025, 0.05, and 0.1 based on the three aforementioned standard deviation levels.

After calculating the individual probabilities, the data set was iterated through again to smooth out the values, as well as account for the small sample size pitches. As pitches thrown in the same time frame will likely be affected by the same weather conditions, pitches found to be thrown in the same inning by the same pitcher were average out with the surrounding data points. If there was data before and after a certain pitch, the adjusted probability was factored 10/80/10 for that pitch. If there was only data after than pitch it was adjusted 80/20, and if there was only data before that pitch it was adjusted 20/80. Additionally, for the data points that did not have a standard deviation, the probability was taken to be the average of the previous and next data points. If one of those did not exist, it was taken to just be the one that did. This was then added back into the deviation data frame that was constructed, and then exported into a .csv file for submission.