Dew Point and Pitching

Weather can alter the way a pitcher throws their pitches. Pitchers tend to not throw in rainy weather for very long -- usually, rainy games result in rain delays or rainouts -- but pitchers do still throw in games where humidity is a factor. Dew point is a measure of how much water vapor is in the air. As dew point increases, the "mugginess" feeling increases. This can affect the flight of a pitch and it can also affect the pitcher's comfortability.

Cincinnati in the summer can be one of the most humid-feeling parks to play in. At a dew point of 65 degrees F, the feeling of humidity becomes uncomfortable. For the purposes of this project and using the pitch data provided, please identify the probability a pitch was affected by a dew point greater than 65 degrees F.

Data

Two files are provided for this problem: `data.csv` and `sample_submission.csv`. Please do not use any public data as supplement; all analysis should use only the data provided.

`data.csv`

This file contains all the data available for your analysis.

PID: Unique ID for a pitch in the dataset. One row of data represents one pitch. PID is also ordered by game date, inning, batter in inning, and pitch number vs the current batter.

INNING_KEY: The inning a pitch is thrown in

BATTER_IN_INNING_KEY: The number of batters faced in a given inning

PITCH_NUMBER: The number of pitches thrown during the plate appearance

OUT_KEY: Number of outs during the pitch

BALLS: Amount of balls in the count before the pitch

STRIKES: Amount of strikes in the count before the pitch

IS_RUNNER_ON_*: Whether a baserunner is on 1B, 2B, and/or 3B

PITCHER_KEY: MLB Advanced Media ID for the pitcher

THROW_SIDE_KEY: Whether the pitcher is throwing left-handed or right-handed

EVENT_RESULT_KEY: The result after a pitch is thrown and the pitch ends in a plate-appearance ending event (strikeout, walk, etc)

PITCH_RESULT_KEY: The result after the pitch is thrown (ball, strike, etc)

INDUCED_VERTICAL_BREAK: Vertical pitch movement due to environmental factors (gravity, air resistance), where 0 represents 0 inches different from a pitch thrown in a vacuum

HORIZONTAL_BREAK: Horizontal pitch movement due to environmental factors (gravity, air resistance), where 0 represents 0 inches different from a pitch thrown in a vacuum. Pitcher handedness can affect horizontal break (i.e. a slider from a right-handed pitcher may break 12" while the same pitch from a left-handed pitcher may break -12")

SPIN_RATE_ABSOLUTE: Spin rate of the pitch in rotations per minute

RELEASE_SPEED: Speed of the pitch in miles per hour

RELEASE_SIDE: Horizontal location at pitch release. Pitcher handedness can affect release side (i.e. a right-handed pitcher who release 2', left-handed pitchers may release at -2')

RELEASE_HEIGHT: Height of the pitch at release

RELEASE_EXTENSION: Distance from the pitcher rubber at pitch release

HORIZONTAL_APPROACH_ANGLE: The horizontal angle of the pitch as it approaches home plate. Pitcher handedness can affect horizontal approach angle (i.e. a right-handed pitcher whose pitch approaches at 3 degrees, a left-handed pitcher may have an angle of -3 degrees)

VERTICAL_APPROACH_ANGLE: The vertical angle of the pitch as it approaches home plate

PLATE_X: The horizontal location of the pitch as it crosses the front of home plate

PLATE_Z: The vertical location of the pitch as it crosses the front of home plate

`sample_submission.csv`

This file is what the output from your analysis should look like. There should be a value representing all unique pitch IDs.

PID: Unique ID for a pitch in the dataset. One row of data represents one pitch. These are the same pitch IDs as in the `data.csv` file.

DEWPOINT_AFFECTED: the probability a pitch was affected by dewpoint.

Your submission

For a complete project, please submit the following three files:

A write-up of your methods and approach. Please make this 2 pages max. This should contain any and all explanations of your discovery work and how you set up your data to return dewpoint-affected probability.

The notebook that you used to produce the submission file. This should be able to run off of the `data.csv` file provided. Please annotate and comment your notebook where appropriate.

a CSV called `submission.csv` that contains your dewpoint-affected probabilities. This file should look nearly identical to `sample_submission.csv` except the DEWPOINT_AFFECTED column should have your probability results rather than the sample submission results.