Noam Greenstein
Methods and Approach write-up

Once I read the data descriptions, I could immediately tell that not all the columns were relevant. I dropped the irrelevant columns and got the info() of the data to ensure everything had the correct data types. Once I fixed those, I checked the other columns for outliers, duplicates, incorrect negative values, as well as logical consistency throughout. Even though I did find outliers, I chose to keep them due to the unknown effect of dew point. I then checked the distribution of each column to check for inconsistencies or skews. I decided to validate the effect of each variable by checking their correlation to the vertical and horizontal break of the pitch. I chose vertical and horizontal break as the baseline because those are the two stats most likely to be affected by dew point due to the nature of their respective statistics. Using the features found by having a significant positive or negative correlation with either vertical or horizontal break, I decided to use k-means clustering to try and see if this would provide any insight as to which pitches were affected. Using the elbow method I found that four clusters would be optimal, so I used that for my model. It did not provide any insights, as seen on the scatter plots. I decided the best way to identify which pitches were affected by dew point, was to take the 95th absolute percentiles of all the variables (the extreme values) and mark them as affected by dew point. Using this new data I trained an xgboost model as a predictor using five-fold cross validation.To optimize the model I tuned three hyperparameters to account for complexity and overfitting. The model performed very well using my data, but I would be interested to see how it performs against new data.