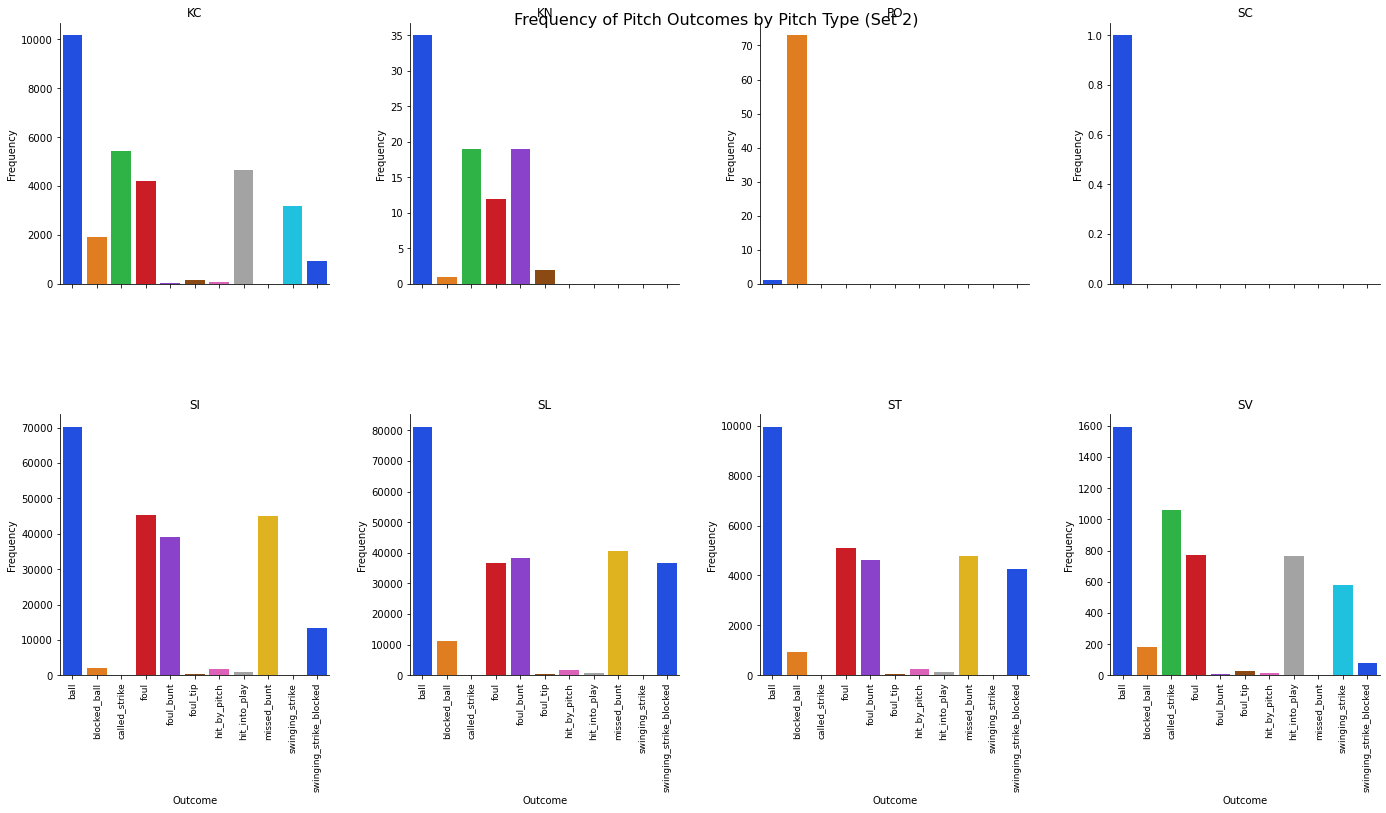
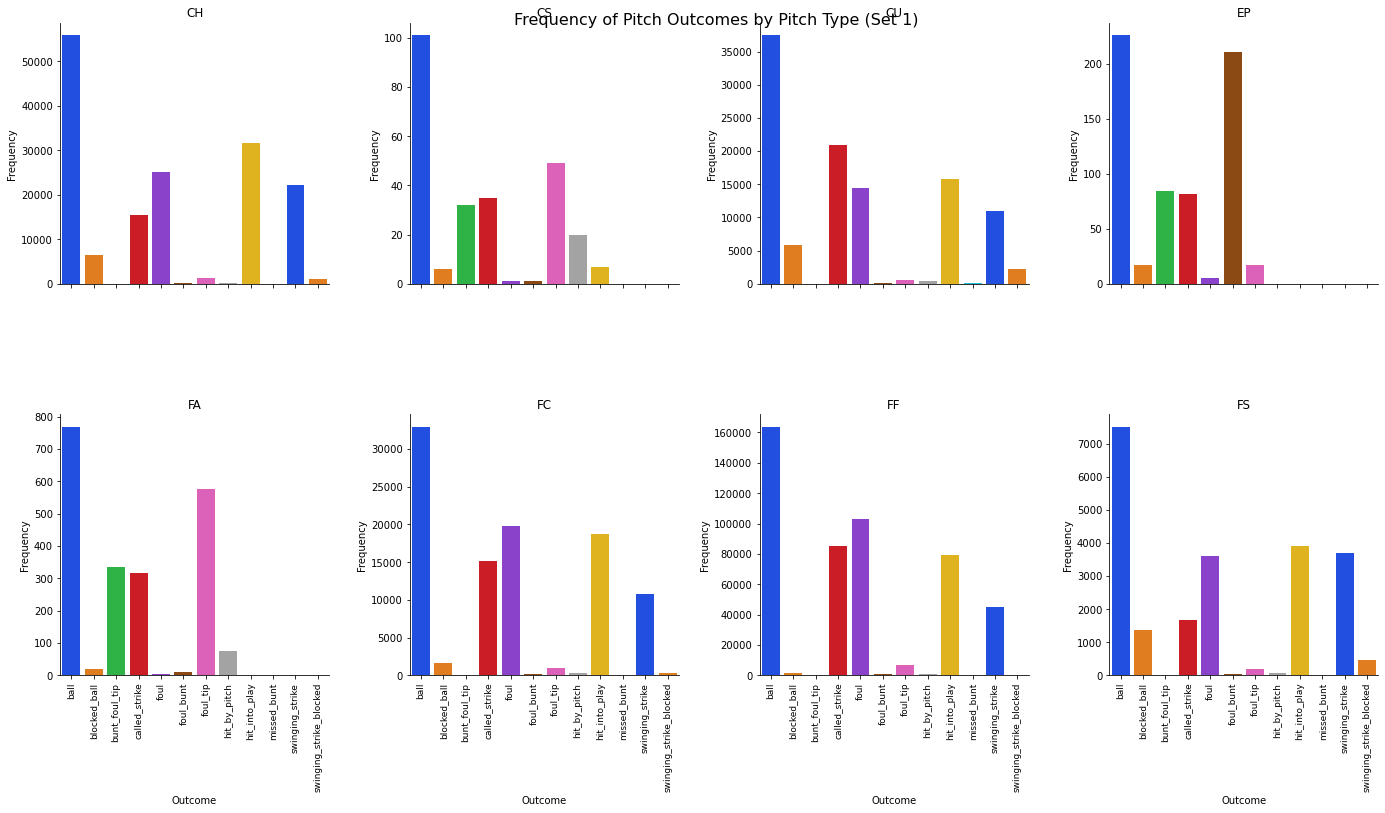
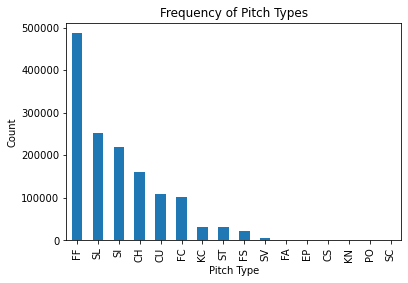
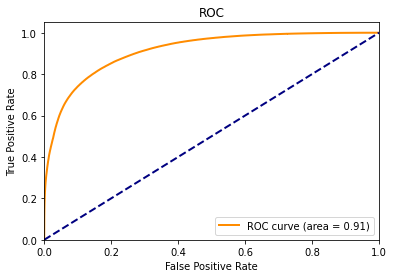
I started with data consolidation work, merging datasets from the first two seasons to construct a combined training set. Simultaneously, the third season's data was set as well for validation purposes and I applied the same preprocessing steps from the combined dataset to this year 3 dataset. Initial data exploration revealed missing entries, which were addressed by designating “Unknown” to absent pitch types and generating unique identifiers for missing pitch IDs. I then dropped all rows which contained the “Unknown” tag for pitch ID since those rows were missing all tracking data as well as dropping the batter, and pitcher columns. This meant that my model was going to focus on the pitch characteristics. This decision was based on my hypothesis that a pitch's outcome is more influenced by its intrinsic attributes rather than the players involved. For the missing under release speeds, I calculated the median speed by pitch type and filled in the missing cells with the corresponding median. Once both datasets were filled in, I conducted some exploratory work to see any relationships between the variables given. For example, I looked at how often each pitch type was used and the frequency of the various outcomes for each pitch type as shown in the visuals below.

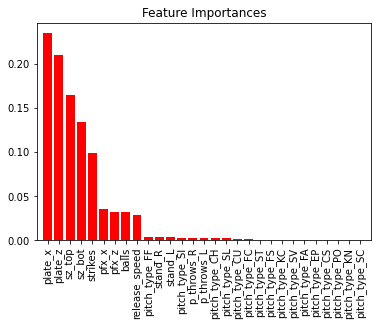


For feature engineering, I created a binary target from descriptive outcomes—swing or no swing- based on the outcomes given in the description. I set it to where a swing is when the outcome was 'hit\_into\_play','ball', 'called\_strike','swinging\_strike' ,'foul', 'foul\_bunt','foul\_tip', and 'in\_play'. I didn't include bunts here because I don’t think of bunts as a swing, but more of a deliberate way for a batter to just create contact with the ball. I used a random forest classifier to model the data due to its resistance to overfitting and ability to handle possible multicollinearity between any feature I included. Hyperparameter tuning was conducted through randomized search due to the computational limitations of my laptop rather than utilizing gridsearch.

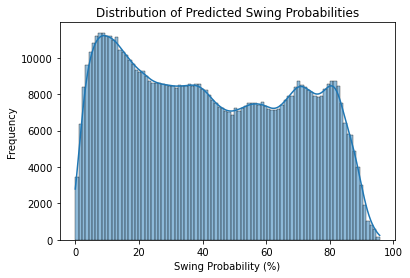
The model produced an ROC with a 0.91 area under the curve, which validates its strong discriminative power between swings and non-swings. This curve showcases the model's high true positive rate across varying thresholds, emphasizing its effectiveness in identifying swings.



The feature importance plot revealed to me that the pitch's location relative to the strike zone (captured by plate\_x and plate\_z) and the size of the strike zone (sz\_top) are important in predicting a batter's decision to swing. These features, along with the count of balls and strikes, hold the most weight in the model, underscoring that the spatial parameters of a pitch influence a batter's decision-making process. The stance of the batter and the throwing arm of the pitcher, while still relevant, exert less influence on the model compared to the pitch's location and game context. The same applied to the various pitch types which all vary in their influence on the model, but to a lesser extent than I had previously thought.



These plots show the model's reliance on the situational and spatial context of each pitch rather than the players' identities, aligning with my initial assumptions. By focusing on the immediate circumstances of each pitch, I created a model which aimed to capture the game's dynamic nature. With this, I was able to apply the model to the year 3 dataset, providing swing probability predictions. Below is a distribution of the predicted swing probabilities in my year 3 dataset.



The results illustrate to me that the predicted probabilities are somewhat bimodal with peaks around what might be seen as low and high probabilities of swing. This suggests to me that my model tends to predict either a high likelihood of a swing or a low one, with fewer predictions in the middle range. This could be a positive sign in the model’s ability to discern the characteristics that lead to a possible swing.