1 Introduction

The data used accounted for batted ball data tracked by two different systems. Additionally, it was provided that system A was suspected to be much more accurate. First, the data from the two systems was explored to identify the quirks associated with them, and the difference between the two systems. Afterward, batter profiles were created to assess the average speed off bat as well as average launch angle for each player.

2 Data Exploration

Simple data exploration was conducted to identify the type and distribution of data, as well as plot the data by column shown in Figure 1. System A had a higher average batted ball speed, lower variance in batted ball speed, and lower average vertical launch angle. Additionally, System A was much more likely to have missing data for a batted ball. System A had launch angle and speed data for about 89.7% of batted balls, while System B had angle and speed data for about 98.1% of data – a significant difference.

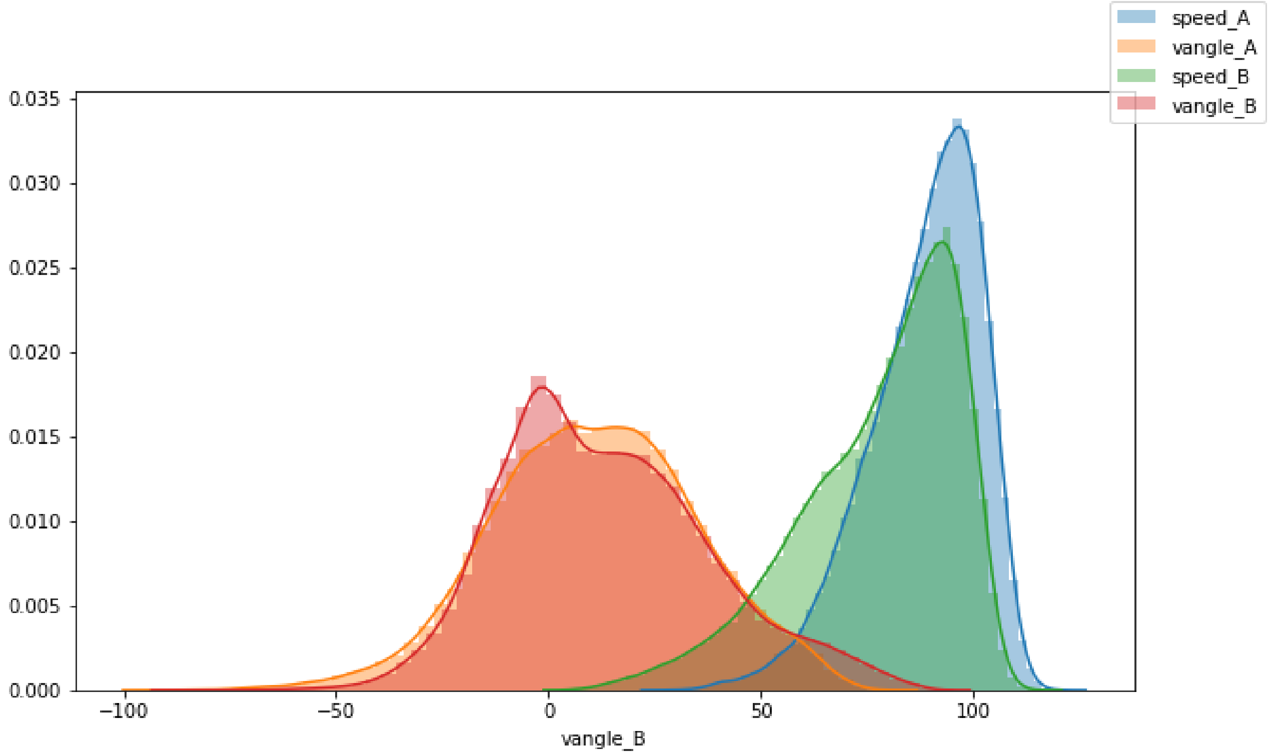
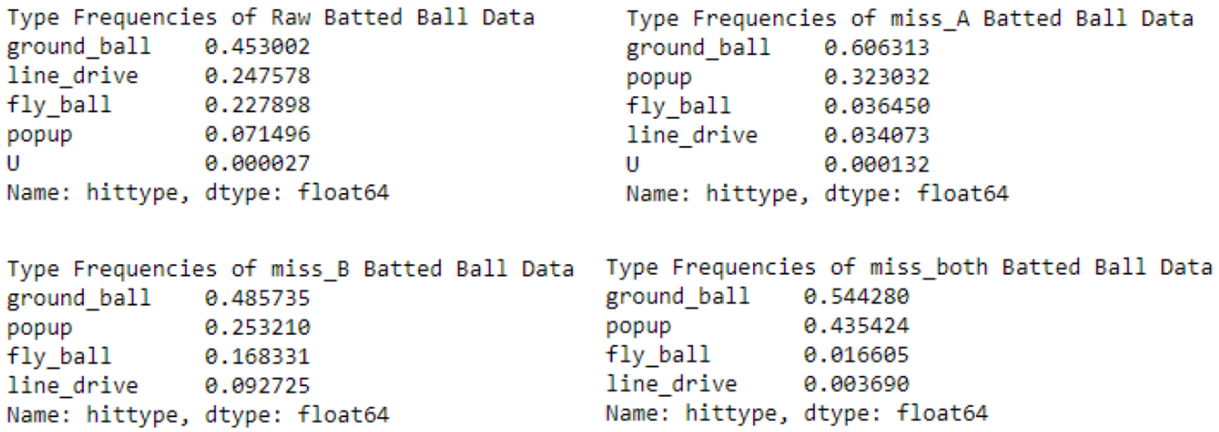


Figure 1: Plot of Categorical Data Distributions

To investigate why there was a significant difference in percentage of batted balls tracked by each system and check if balls were missing at random or not at random, the data was divided into subsets based on what data was missing. Batted ball instances with data from only system B were grouped together such that this group contained nan values for system A measurements. This data will be referenced as miss_A for this report. The same was done for

groups with data from only system A (miss_B), and with data missing from both systems (miss_both). Data instances containing measurements from both systems was left out for now.
The data distribution in miss_B was relatively similar to the original data set. However, in this data set it was much more likely to be missing data from system A as well than in the original data set. The data distribution in miss_A was significantly different from the original data set, with a mean of 58 for batted ball speed measured by system B and a mean of 19 for launch angle. The batted balls missing data from system A have a much lower average exit velocity. Additionally, the first and third quartiles of average launch angle in miss_A are much further from the mean than in the raw data, suggesting that there is a higher probability of a ball having a low launch angle or a high launch angle in this subset.

To further highlight this finding, the relative frequencies of batted ball hit types were recorded for the raw data, miss_A, miss_B and miss_both. These findings are displayed in Figure 2. The relative frequencies for ground balls and popups were significantly higher in miss_A. Ground balls and popups are even more common in miss_both, representing 98% of the data of miss_both. Ground ball rates remained fairly constant in miss_B, however popups were much more common, with a noticeable decrease in line drive rates. This indicates that system A was much more likely to fail to produce measurements tracking popups and groundballs, while system B mainly had difficulty in tracking popups.



```
Type Frequencies of Raw Batted Ball Data        Type Frequencies of miss_A Batted Ball Data
ground_ball    0.453002                         ground_ball    0.606313
line_drive     0.247578                         popup          0.323032
fly_ball       0.227898                         fly_ball       0.036450
popup          0.071496                         line_drive     0.034073
U              0.000027                         U              0.000132
Name: hittype, dtype: float64                   Name: hittype, dtype: float64

Type Frequencies of miss_B Batted Ball Data   Type Frequencies of miss_both Batted Ball Data
ground_ball    0.485735                       ground_ball    0.544280
popup          0.253210                       popup          0.435424
fly_ball       0.168331                       fly_ball       0.016605
line_drive     0.092725                       line_drive     0.003690
Name: hittype, dtype: float64                 Name: hittype, dtype: float64
```

Figure 2: Hit Type Frequencies for Raw Data, miss_A, miss_B, and miss_both

3 Methods

Player profiles containing number of batted balls, average exit velocity, and average launch angle were then created for each batter. Based on the data exploration, the missing data was not categorized as missing at random. This is because balls with tracking data missing on at least one system have a disproportionately large probability of having negative or large positive launch angles – a strong indication of data being missing not at random. Imputation methods for such data sets tend to lead to bias in results. Due to the data being labeled as missing not at random, and given the information that system A is suspected to be much more accurate, a simple decision tree served as the method of imputation when calculating that the speed off bat was for each instance. If the batted ball has data tracked by system A, then system A data is used for the batted ball speed and vertical launch angle. If the batted ball does not have measurements from system A, then system B data is used for the speed and launch angle of the batted ball. If data was missing from both systems, the data was dropped. This was done because only 542 (< 1% of the data) of 73,375 batted balls were missing data from both systems, so the impact was assumed to be small.

4 Discussion

More complex statistical and data science methods may be used in this project in an attempt to do a better job of finding the “true” speed off bat average for each player. The imputation process was chosen to be very simple in this project due to a number of constraints such as limited sample size, prior belief that one system was more accurate, a lack of other features for batted balls, and the trend that balls were not missing randomly. An example of a possible simple imputation method would be to replace missing data with the average exit velocity and launch angle of the given hit type for that system. However, correlation analysis showed that the exit velocity and launch angle of a batted ball are not particularly correlated as shown in Figure 3, which would lead to a biases in results. More advanced methods of imputation such as multiple imputation and maximum-likelihood estimation techniques may produce better estimates of missing data may capture ways to use the system that does have data to fill in missing data in the other system effectively, which is a route to consider when attempting imputation for this type of data in the future. If possible, however, expanding the range of trackable launch angles in the more accurate system without reducing accuracy would be ideal, as this would lessen the burden of needing advanced imputation methods to handle missing data. Lastly, given the relatively small sample size of the data for each batter, their “true” averages for speed off bat and launch angles are most likely closer to the average distribution across all players, and a prior distribution of player averages for exit velocity and launch angle may be used to infer a better estimate of an individual player’s performance.

5 Conclusion

Differences in System A and B were explored showing that System A seems to have difficulty in tracking balls with extreme launch angles – ground balls or popups, which resulted in a much higher rate of missing data. A simple imputation method used for reasons discussed in Section 4 was used to handle missing data when creating profiles containing player averages for exit velocity and launch angle. A practical use of this data would be to use profile data and the original in conjunction with average outcomes given a batted ball’s exit velocity and launch angle to assess player performance by creating metrics. With a prior distribution of batted ball data, it would be possible to produce more complex imputation methods that would do a better job of calculating a player’s true speed off bat average.