Atlanta Braves R&D Questionnaire

**Analysis Questions**

Please limit each answer to a maximum of 500 words.

1. On 8/24/2021, the Cardinals trailed the Tigers 4-3 going into the top of the 9th. To begin this inning, Daz Cameron doubled, Akil Baddoo struck out, and Jonathan Schoop grounded out, moving Cameron to 3rd. The batter is now Robbie Grossman. Assume Luis Garcia will pitch through the 5th spot in the batting order, Jeimer Candelario. Should the Cardinals intentionally walk Grossman? Describe what your process would be to determine whether to pitch to him. The following link contains the box score information for this game: <https://www.mlb.com/gameday/tigers-vs-cardinals/2021/08/24/632781#game_state=final,lock_state=final,game_tab=box,game=632781>

I would not intentionally walk Grossman. Although the hitter after Robbie Grossman, Zack Short, was statistically a much worse hitter than Grossman in 2021, walking Grossman would give the Tigers a free base runner and allow them a greater opportunity to extend their lead against the Cardinals and put the game out of reach. This can be backed up by run expectancy research, as situations with two outs and runners on first and third have higher expected run values than situations with two outs and a runner only on third. This difference in expected run value has a greater importance than the difference between Grossman and Short’s probabilities of hitting into outs. We must also consider how the remainder of the inning would play out in the event that the Cardinals intentionally walk Grossman. In a situation where Zack Short does not hit into an out, but instead gets on base, it brings to the plate Jeimer Candelario, who is arguably the Tigers’ best hitter, in a situation where he may have multiple runners in scoring position. A hit from Candelario, which is more likely than a hit from both Grossman and Short according to their batting averages, would blow the game open and most likely ruin the Cardinals chances of coming back and winning. Thus, why give the Tigers’ best hitter an opportunity to extend a lead when you could pitch to Grossman, who has a higher probability of hitting into an out, in a situation where you are less vulnerable to allowing a significant number of runs? Despite the fact that I would personally pitch to Grossman and avoid the intentional walk, there are other factors to consider when making the decision. Luis Garcia is a right-handed pitcher, while Robbie Grossman and Jeimer Candelario are both switch hitters and Zack Short is a right-handed hitter. Thus, out of the three hitters, pitching to Short would provide the greatest platoon advantage for the Cardinals, giving incentive to walk Grossman. However, I would like to know Grossman and Candelario’s splits from both sides of the plate, as it could be possible that Grossman struggles as a left-handed hitter, which would incentivize pitching to him rather than walking him. I would also want to know recent trends for each hitter. If Robbie Grossman is currently on a hot streak, then I would possibly prefer to walk him, but if Zack Short is the one who is on fire, then I would be even more sure about pitching to Grossman. Overall, I wouldn’t be able to make a sure decision without information on recent trends, hot streaks, and more for Grossman, Short, Candelario, and even Luis Garcia. However, I believe that intentional walks should be reserved for elite hitters, as they allow for the opposing team to have free base runners, and even the best hitters in the league get out more than half the time. Thus, without further information, I would choose not to intentionally walk Robbie Grossman in this situation.

1. You are running a generic mid-market team and are exploring the idea of signing Cody Bellinger this offseason. What contract would you be willing to offer him? Please explain your thought process and discuss any important considerations.

Cody Bellinger had a resurgence in 2023, recording his best season since 2019 when he won MVP. However, when digging into his peripheral numbers, via Baseball Savant, I have to be weary of his success. Bellinger’s xwOBA in 2023 was .331, 39 points lower than his actual figure. This implies that he may have benefited from luck this season, as his expected statistics place this season closer to his 2018 season, in which he recorded a 120 wRC+ (from FanGraphs), rather than his Rookie-of-the-Year season in 2017 or MVP campaign in 2019. However, that 120 wRC+ is still well above average and, if we expect this to be his true offensive talent, he is certainly worth a multi-year contract. This is backed up by the fact that he is able to consistently perform at an elite level defensively at a premium position in center field. Thus, it is important to value Bellinger at the proper figure. Although he is most likely never going to have an MVP-caliber season again, he has still proven that he can post good all-around seasons year in and year out. Some comparable players to Bellinger offensively are Brandon Nimmo, Ketel Marte, and his current teammate Seiya Suzuki. Each of these players are currently tenured under multi-year deals worth $20.25 million, $15.2 million, and $17 million per year, respectively. A contract I would be comfortable with giving Bellinger would be somewhere within this same range. Another important factor to consider here is Bellinger’s age. At 28 years old, I would be cautious about how his skills may deteriorate over time. Players often lose athleticism after the age of 30, and this might result in Bellinger losing the elite defense that makes him so valuable. Thus, I wouldn’t be willing to give him a contract longer than 5 years, with 3 or 4 years being my preferred length. Without consideration of any team constraints, my offer to Bellinger would either be $60 million over three years or $75 million over five years. However, I must also take into account the current budget of my team. As a mid-market team, it would most likely not be a financially smart decision for me to give Bellinger a contract similar to Nimmo’s with the Mets. The Marte and Suzuki contracts would seem more reasonable. I also have to keep in mind how much the players on my current roster are making. If my team is young and has a lot of pre-arbitration players, then I may be more inclined to overpay for Bellinger. However, if I already have multiple players locked up on big deals, then I might have to lowball Bellinger or not give him an offer at all. I don’t know this information, but, given that I have at least some restriction on my budget as a mid-market team, I would formulate an offer worth $15-17 million per year. Therefore, my final offer to Bellinger would be a four year deal worth $64 million in total.

1. Pitcher A walks half the batters he faces and strikes out the other half. Pitcher B doesn’t walk or strike out any of the batters he faces. Which pitcher would you prefer? What ratio of strikeouts to walks would make you indifferent between the two pitchers?

Since Pitcher A strikes out half the batters he faces, I know he has a 50% chance of recording an out for any single batter. The probability of Pitcher A getting through an inning is the sum of the probabilities that he: walks 3 batters and records 3 strikeouts, walks 2 batters and records 3 strikeouts, walks 1 batter and records 3 strikeouts, and walks 0 batters and records 3 strikeouts. This calculated probability is 5C3(.5)6 + 4C2(.5)5 + 3C1(.5)4 + 2C0(.5)3 = 0.65625 or 65.625%. The probability of Pitcher A allowing at least one run in an inning is one minus this number, or 0.34375. Multiplying this by 9 to convert it to an ERA gives us 3.09. However, this is the minimum possible ERA this pitcher could have, as it is built off the assumption that Pitcher A can only allow one run, which isn’t true, as he could allow multiple runs by walking more than 4 hitters before reaching his third strikeout. To show how this ERA inflates as I factor in those situations where he walks in one or more runs I can multiply the probabilities of him allowing four, three, two, and one runs by their run values, then multiply the sum by 9 to get a more accurate ERA: 9 \* [ 4(9C7(.5)10) + 3(8C6(.5)9)+ 2(7C5(.5)8)+ 1(6C4(.5)7) ]= 5.2835625. An ERA of 5.28 is not good, and it would only grow as I add in probabilities of Pitcher A allowing 5 or more runs. In comparison, I know that Pitcher B does not walk or strike out any batters. Thus, basically every at-bat he pitches in ends with a ball in play. There is no mention of Player B’s performance on these balls in play so I will assume that he performs at a league average level. The league average BABIP is 0.295, according to FanGraphs. This means that Pitcher B’s probability of recording an out is one minus his BABIP, or 0.705. This is much higher than the 0.5 probability of Pitcher A recording an out and implies that Player B is preferred as a pitcher. Additionally, since I am assuming that Pitcher B is a league average pitcher, his ERA would be 4.33, once again according to FanGraphs. This is also a better figure than the one I calculated for Pitcher A. Thus, assuming that Pitcher B is at least a league average pitcher, I would be confident in choosing him over Pitcher A. A strikeout to walk ratio that would make me indifferent towards which pitcher to choose would be either: 1) a 70.5% strikeout rate, meaning that Pitcher A and Pitcher B are equally likely to record an out, or 2) a strikeout rate that reduces Pitcher A’s calculated ERA so that it matches the league average ERA.

1. Briefly explain how you would go about estimating the effect of catcher framing at the major league level? Assume you only have access to the identities of the people involved, information about the pitch (location, characteristics, etc.), and information about the game (count, inning, score, etc.).

There are two different ways I would want to quantify a catcher’s framing ability. The first is by a pure percentage, determining how often a catcher can turn a ball into a strike. The second is by influence on the game. There may be catchers who aren’t as good at framing as others, but do it well in important or high leverage situations, and should get credit for that. To begin, I would group the data by catcher identity to ensure that each player is looked at independently from the others. I would then compare pitch location data with pitch call data and count how many times a pitch was thrown outside of the strike zone but labeled a called strike. Adding to this, I would weigh each of these pitches by their distance from the strike zone, as a pitch that was further away from the zone but still framed as a strike should be viewed as more impressive for the catcher. Adding these weighted counts together and dividing them by the total number of pitches outside the strike zone a catcher received would give me a “framing score” that I could use to compare catchers. Turning to the second method of quantifying framing, I would combine these weighted counts with game information like the inning, score, and count, to give a greater perspective on how influential they were. For example, a pitch that was framed as a strike on a 3-0 count with the bases empty in a blowout will have much less significance than a pitch that was framed as a strike on a 3-2 count with the bases loaded in a tie game. This second method will account for that by measuring the difference in probabilities of recording an out or allowing a run between the situation where the pitch was called a strike and where the pitch was called a ball. This is very similar to what umpire scoring algorithms do when determining how an umpire’s misclassified pitch influences a team’s chances of scoring or preventing runs. Both of these methods will prove useful, as the first gives an overall impression as to which catchers around the league are the best at framing in general, while the second reveals which catchers have saved or cost their team the most runs via their framing capabilities. There are some external variables to consider while implementing this project. For example, some catchers may be unfortunate to play with pitching staffs that don’t allow them many opportunities to frame pitches, and may have low scores on these metrics despite great framing talent. Additionally, umpires have a strong influence on whether a pitch is framed or not given that they call the pitches, and some umpires are more vulnerable to falling for a frame job than others, which should be taken into account. Despite this, I believe that I would be able to adequately quantify catchers’ framing abilities around the league.

**Modeling Questions**

For these exercises, you will be using your knowledge of R or Python to answer a few baseball-related questions.

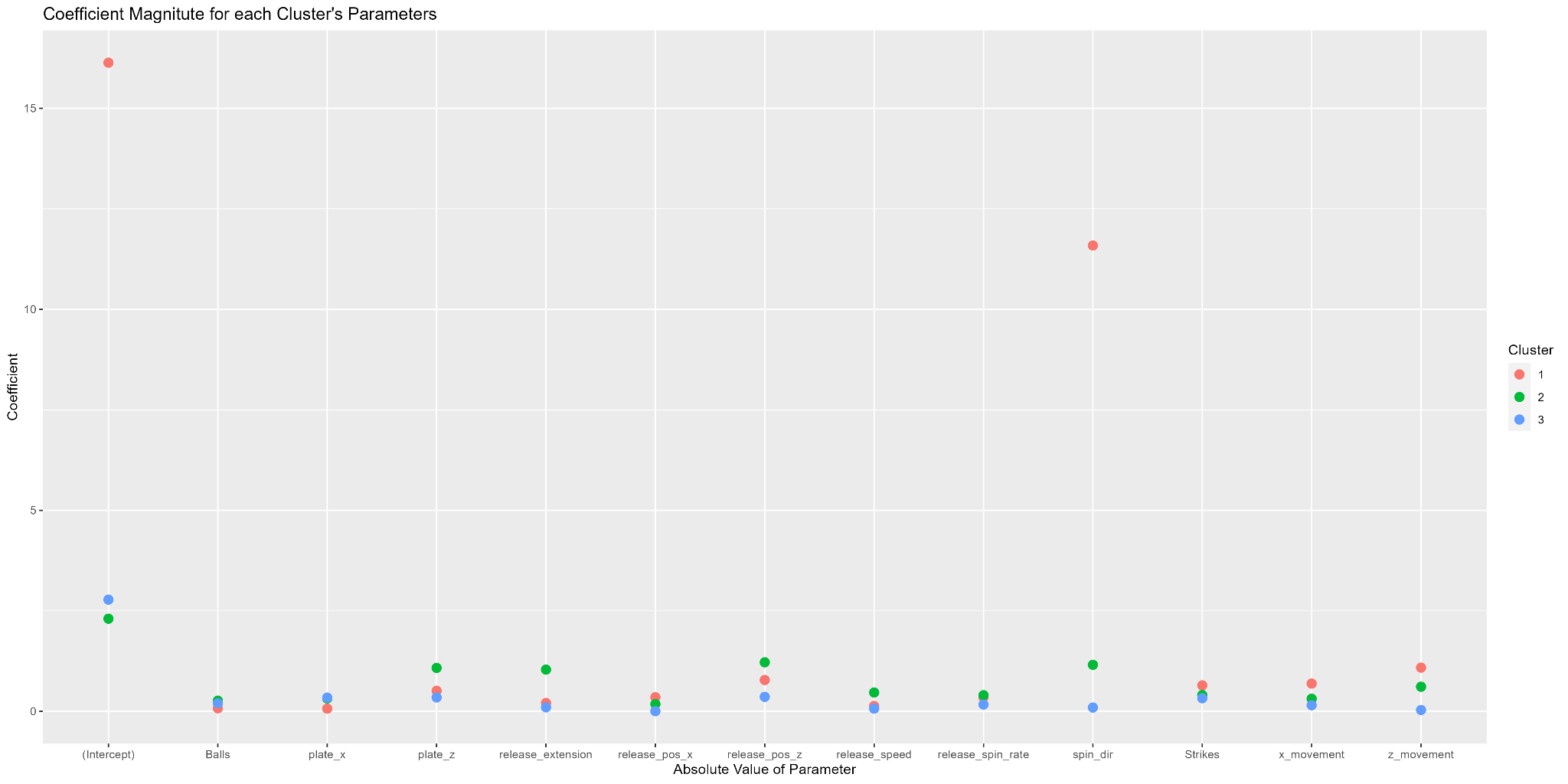
Use the attached Trackman pitch by pitch data of Braves pitchers from the 2018 season to answer the following prompts. It is not necessary to utilize every column in the attached file; only use those you feel are necessary. There is a GLOSSARY defining the columns in the PitchData.csv file on the second page of this document.

This exercise should not take more than a few hours. Please include all of your code with your responses. If you do not know how to complete one or more of the questions, feel free to leave them blank.

* 1. **Create TWO models to predict the likelihood of a swing and miss based on the characteristics of a curveball. Evaluate and compare the performances of your models using any method(s) you’d prefer. Explain your results in 500 words or less.**

The two models I chose to predict the likelihood of a swing and miss based on the characteristics of a curveball were a clustered logistic regression and a random forest. Before implementing the models, I converted X movement and Release Position X to their absolute values in order to eliminate the influence of handedness. Next, I created a binomial variable named “Whiff” as the response variable for the models. I then standardized all the pitch metrics, including Release Speed, X Movement, Z Movement, Spin Rate, Spin Direction, Release Position X, Release Position Z, Release Extension, Plate X, and Plate Z. From there, I turned to the models. For the logistic regression, I used K-means clustering with three centers to create three different clusters to perform the model on. Building three independent logistic regression models with a 70% training split on the individual clusters gave me test accuracies of 0.908046, 0.8712121, and 0.891253. For the random forest, I used a 70% training split once again and the model resulted with 500 trees, trying 4 variables at each split. The corresponding accuracy on the testing set was 0.8939198, which demonstrates consistency with the three logistic regression models. There are some concerns I had with the models and their results. Due to the fact that there is a disproportionate amount of non-whiffs to whiffs in the dataset, both models were more trained to correctly predict non-swing and misses than they were to predict swings and misses. This resulted in both models frequency misclassifying swings and misses. However, if we are operating under the assumption that we have a full set of 2018 data, we can be more confident that these results are accurate.

* 1. **Using your preferred model from Question #3, create a visualization to display the most important characteristics of a curveball in recording a swing-and-miss. Explain your visualization in 500 words or less.**



This plot illustrates the strength of the coefficients for each parameter in the three logistic regression models for the three clusters. This strength is defined as the absolute value of the coefficients, as the sign of the coefficients don’t matter to this plot. From the plot, spin direction has the most variance between the three models, which makes sense, as spin directions are measured in terms of degrees. Outside of spin direction and the intercept, Plate Z, Release Extension, Release Position Z, and Z movement seem to have the most influence on the model. This also makes sense intuitively, as a curveball with good vertical movement and location is expected to generate more whiffs. Overall, for most of the parameters, there isn’t much differentiation between the coefficients for each cluster, meaning that a single logistic regression model may have worked almost as well as three on different clusters.

**Note:** Models in this exercise will be less accurate due to small samples of pitches and pitchers, so proceed with your evaluations and conclusions as if there were a complete set of 2018 data.

**PitchData.csv Glossary**

| **Variable** | **Definition** |
| --- | --- |
| Pitcher\_ID | The pitcher’s MLBAM ID |
| Pitcher | The pitcher’s full name |
| Pitcher\_Throws | The pitcher’s handedness |
| Batter\_ID | The batter’s MLBAM ID |
| Batter | The batter’s full name |
| Batter\_Hits | The batter’s handedness |
| Game\_Date | The date the game occurred |
| Top\_Bot | Whether it is the top or bottom of the inning (1 signifies the top and 2 signifies the bottom) |
| Inning | The inning the pitch was thrown |
| Balls | The number of balls when the pitch was thrown |
| Strikes | The number of strikes when the pitch was thrown |
| Outs | The number of outs when the pitch was thrown |
| Pitch\_Outcome | The outcome after the pitch was thrown |
| Pitch\_Type | The pitch type (4-Seam and 2-Seam are grouped as fastballs) |
| release\_speed | The pitch’s velocity (mph) |
| x\_movement | The pitch’s horizontal movement (inches) |
| z\_movement | The pitch’s vertical movement (inches) |
| release\_spin\_rate | The pitch’s spin rate (rpm) |
| spin\_dir | The pitch’s spin axis (degrees) |
| release\_pos\_z | The horizontal release point for that pitch (ft) |
| release\_pos\_z | The vertical release point for that pitch (ft) |
| release\_extension | The release extension for that pitch (ft) |
| plate\_x | The horizontal location of the ball when it crosses home plate (ft) |
| plate\_z | The vertical location of the ball when it crosses home plate (ft) |