The Developmental Cost of Extended Period of Missed Play for MLB Future Prospects

Team: Lulu's Lemons

2/19/2022

Team Members: Abigail Irby, Carolyn Moor, Jonathan Dalsey, Chenfei Li, Aryaman Khandelwal, & Sia Lee

Professor: Dave Zes

Due: February 19, 2022

Abstract

This report sets a milestone for the investigation of the developmental cost of extended periods of missed play for a player's future success. In order to conduct the analysis, we chose and scraped a draft 2005 dataset from multiple websites, including the MLB website and baseball-reference.com, and we studied the relationship of bonus, position, WAR and number of games played. We have also looked at the distribution of positions and recruitment time for future further analysis.

1. Introduction

Our research question is, for MLB prospects, does missing an extended period of play during the developmental time in a player's career affect future success? We will judge a player's success by comparing their game statistics in the Major Leagues to similar players who did not miss a significant period of time. Additionally, we will assume that a successful player plays in the Major Leagues for at least 5 years. We define their development time as during minor league play.

We used web scraping tools to gather all of the data from baseball-reference.com and mlb.com. When gathering our data, we grouped players based on their position on the field and their development time based off of their recruitment time: High School, Junior College, and 4-year College. This grouping allows us to compare the players missing playing time to other players with similar circumstances that do not miss significant playing time which therefore controls for those confounding factors.

2. Background

Our goal is to help managers and recruiters gauge the success of their investment into a player by understanding how extended periods of missed playing time affects the success in a player's career. The time in the minor leagues is seen as development time for the player and there is a consequence to missing extended periods of playing time due to injury, family circumstances, or illness, that cannot be replaced by more training and practicing.

Major League Baseball is the highest level of play and we will be looking at the players in the top 200 signing bonus in the draft picks starting in 2005 to 2010 and who entered the major leagues prior to the 2010-2012 seasons. We are examining only the top 200 signing bonuses because those players are expected to perform the best. We only include players who enter the majors prior to 2010-2012 seasons because there is an accumulation of statistics and data on the player's careers which is beneficial for deriving an accurate answer to the research question. We do not have the ability to get the exact injury or reason for the missed playing time, so we will be looking at whether or not they played each game for that season and determine if there is

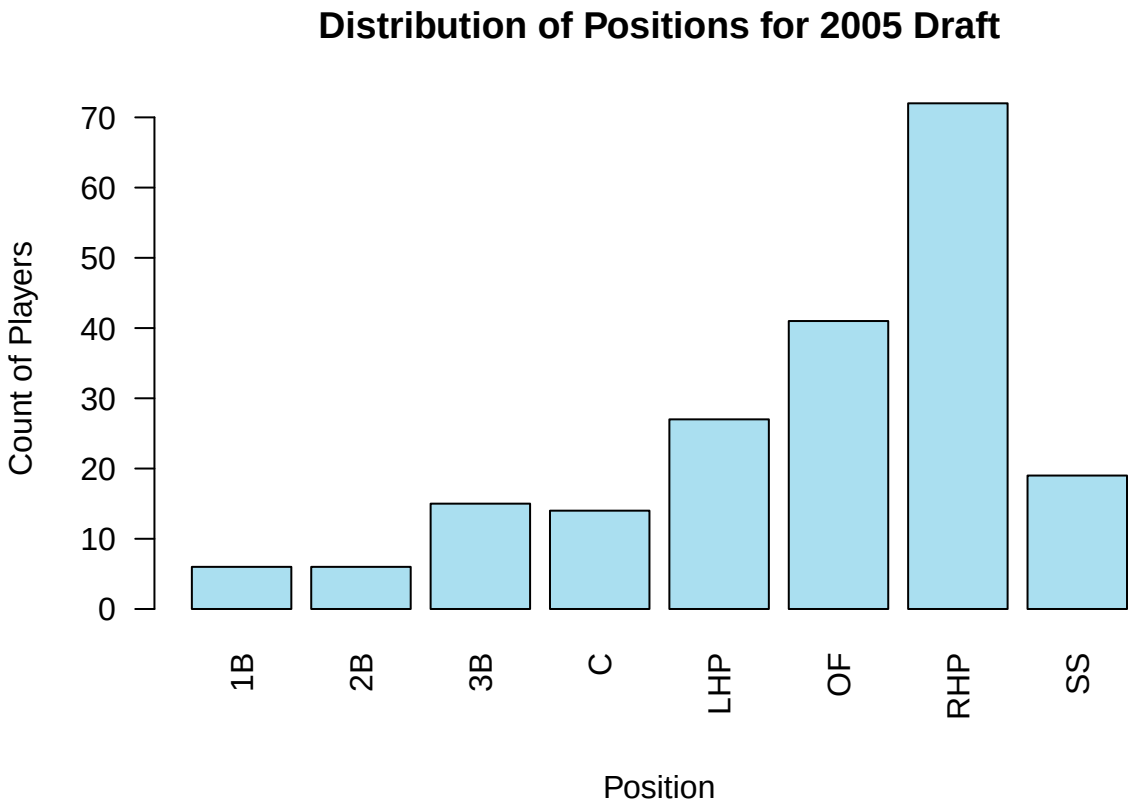
a significant period of time without playing in a game. We defined the extended period of missed playing time as 10 months between games.

3. Results

3.1 Initial Dataset Investigation

We began the analysis by doing a first look at the distribution of important factors in the dataset. Figure 1 shows the distribution of the player positions for the 2005 draft. The definitions of each of the bar labels can be found in Table 1.

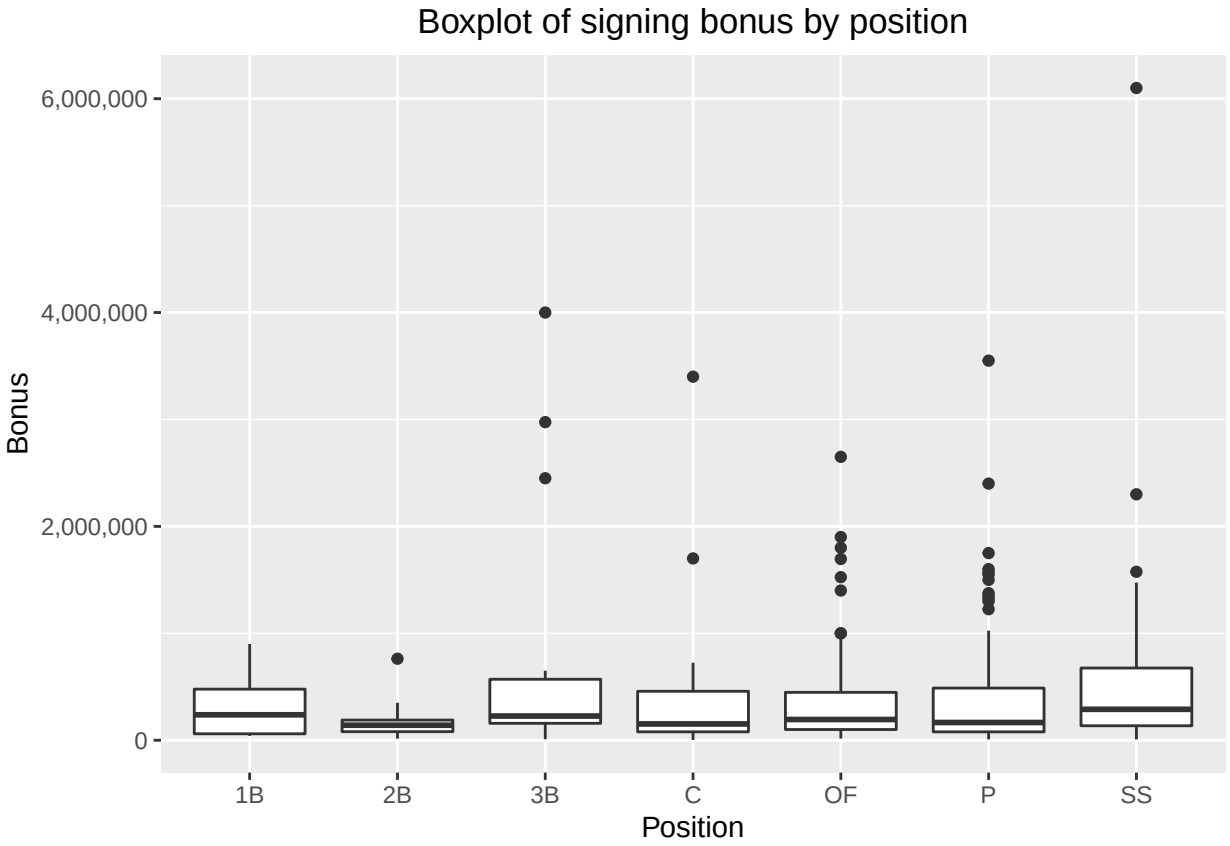
From the figure and table we can see that there are more pitchers than any other position drafted into the MLB. We expected this result because pitchers are only able to physically pitch a certain number of pitches every game. Therefore, they must be rotated in and out frequently so teams will have a larger roster of pitchers.




```
## positions
##  1B  2B  3B   C LHP  OF RHP  SS
##   6   6  15  14  27  41  72  19
```

Another graph that may be worth investigation to answer the primary research question is one that plots the signing bonus of baseball players by position. From the graph provided in Figure 2, we can see that, on average, 1st bases, 3rd bases, and shortstops tend to accrue the highest signing bonuses. Additionally, it is interesting to note that 2nd bases not only earn the lowest signing bonuses on average, but the distribution between quartiles is also the least among positions. Josh Upton, the first over shortstop, earns the highest signing bonus at $6,100,000.

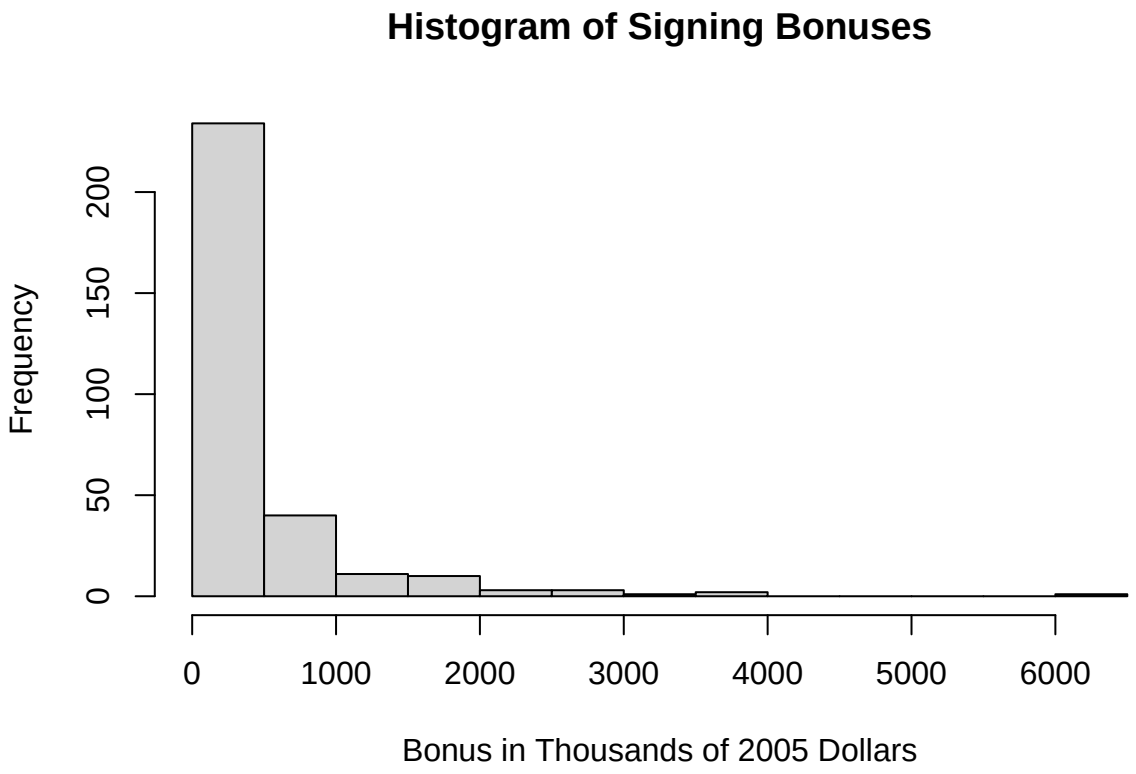
At first, left-hand pitchers and right-hand pitchers were combined into one variable due to the initial lack of depth in data. Moving further into the project, these will be separated once there are more datapoints available.



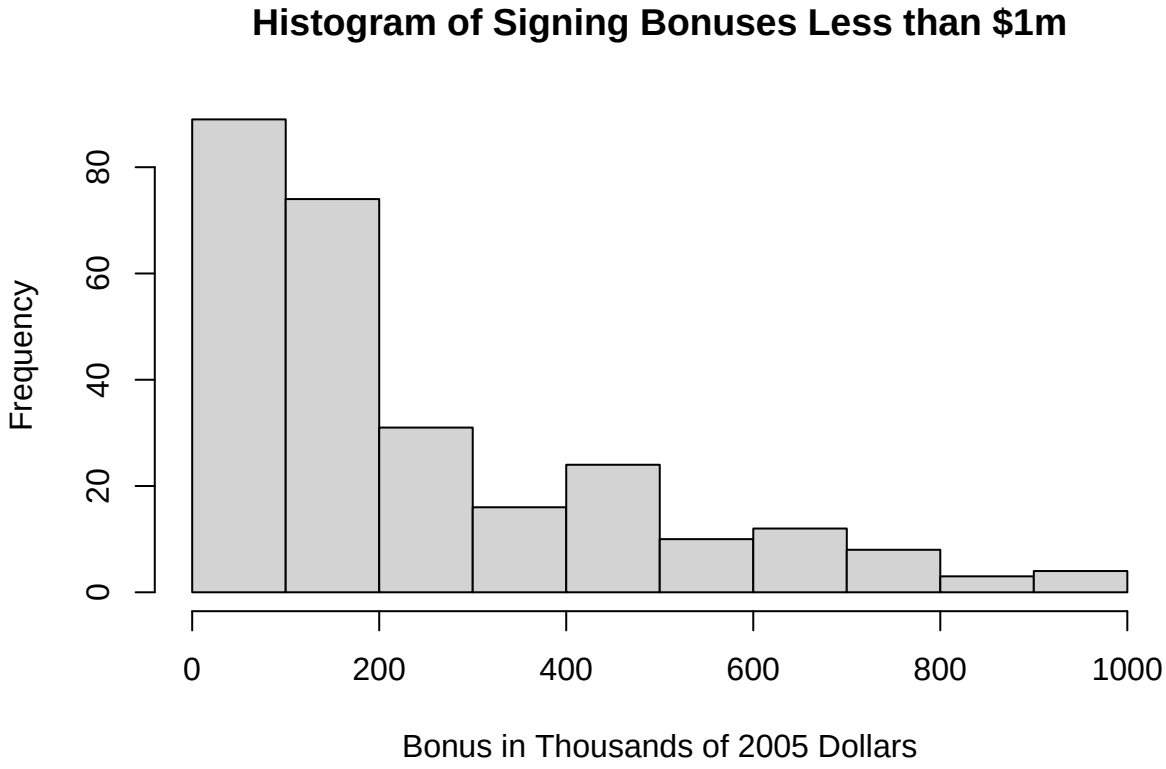
There were few NAs in the data for the type of recruitment time. Thus, all the NAs were removed and the rest of the data were plotted in the decreasing order of the number of players.

Looking at figure 3 below, the most frequent recruitment time is during the 4 years of college. This is as expected as getting recruited in college years is the most common procedure that players take. Having a small portion of the players getting recruited in the Junior College is quite predictable as there are much more players going to the 4 year College rather than Junior College in general, and it is expected that Junior College players are facing a lower level of competition – as such, scouts often do not feel that they can accurately predict the ability of these players at a higher level. Additionally, many players transfer from Junior College Programs to 4 year college programs.

We plan to use the value of a player's signing bonus to determine how highly the team which drafted them values that player, and we believe this should give us a good estimate of how good the team expects this player to be. For the 2005 data we used for this milestone, the figure below (Figure 4) shows a histogram of the signing bonuses, displayed in thousands of 2005 dollars. As expected, this graph is heavily right-skewed: it shows that a few players recieve very large signing bonuses in excess of 2 million dollars, but most players recieve a bonus of less than 1 million dollars.



Given this information, we created an additional histogram (Figure 5) to examine the distribution of the low signing bonus values. The figure below shows the distribution of signing bonuses which are less than 1 million dollars. Here, we see that most players in the draft recieve a bonus of less than 300 thousand dollars.

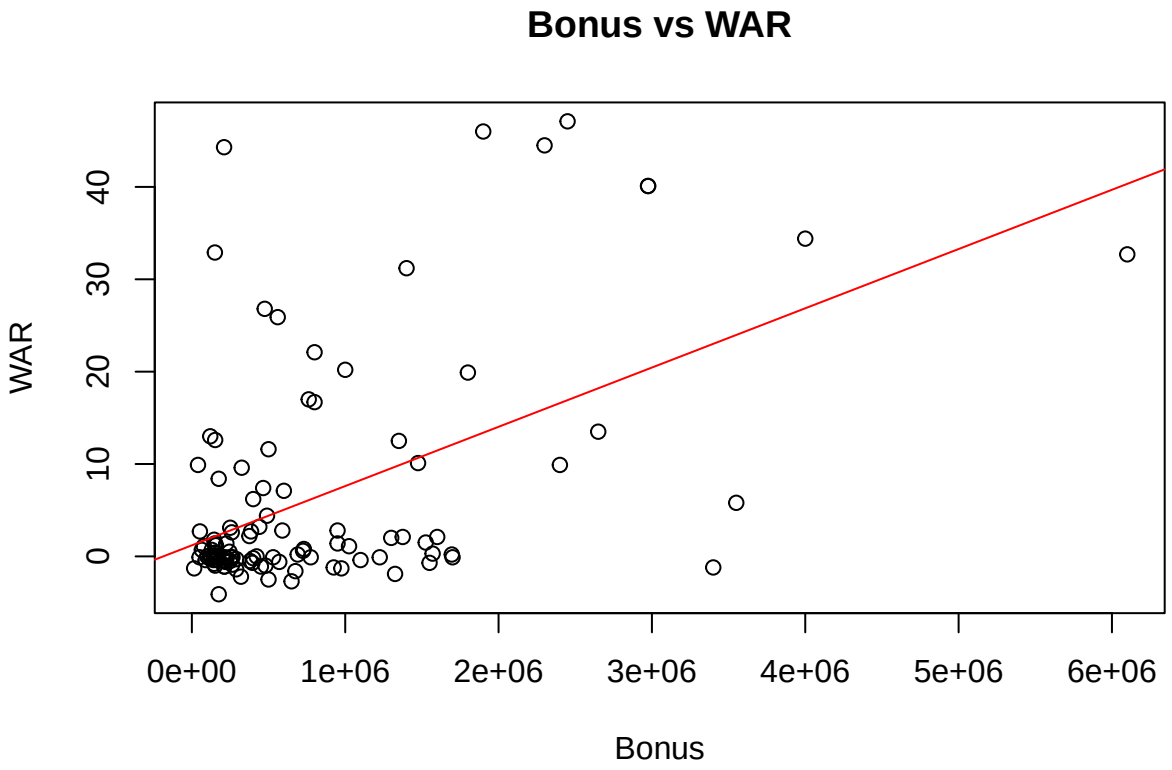


Next, we decided to plot the signing bonuses against draft rank, to confirm the suspicion that players picked

earlier in the draft are generally given a higher signing bonus than players picked later in the draft, because they are valued more by their teams. Figure 6, below, shows that early draft picks recieve the highest signing bonuses, and that bonus value falls off exponentially as the draft rank increases, before leveling off at around pick 150.

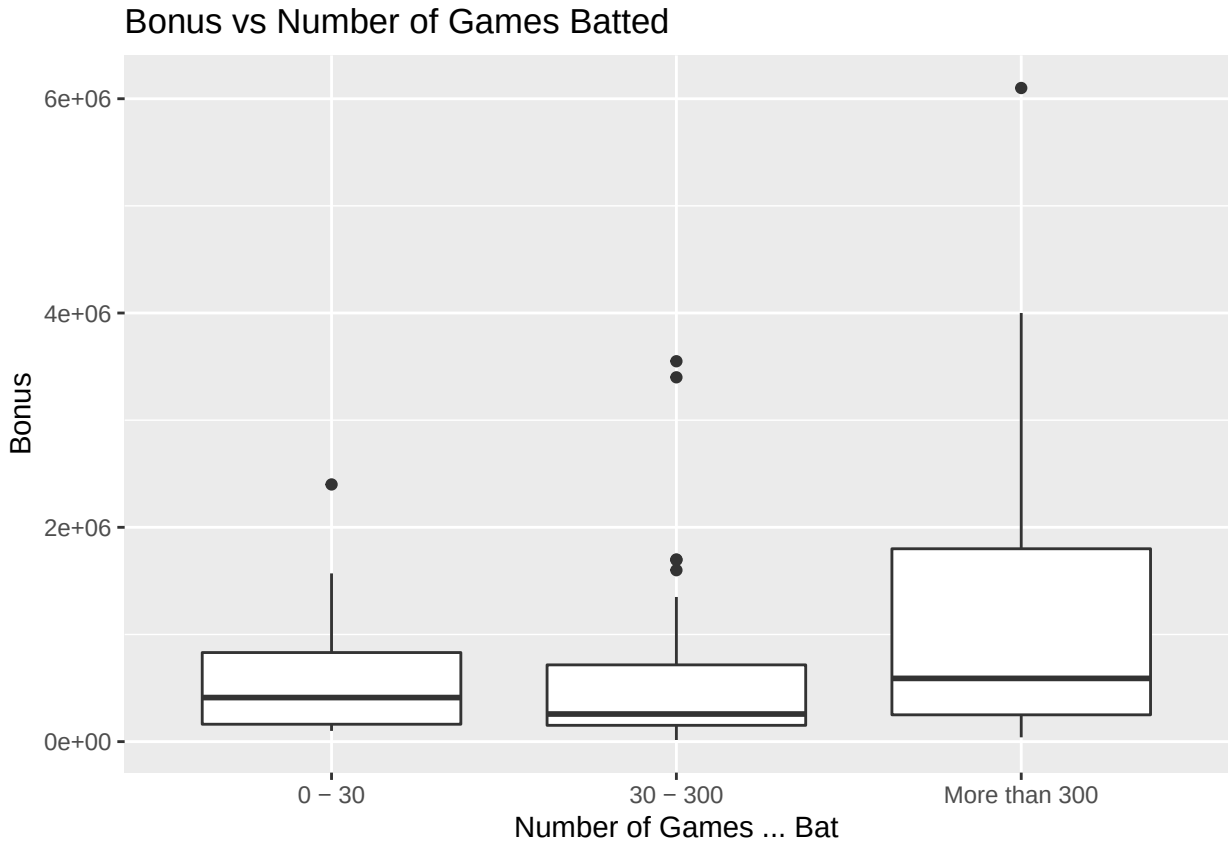
We added the WAR (Wins Above Replacement) metric to color the graph as this is the metric we will use to evaluate player quality. Visually, we did not see a strong relationship between bonus and WAR, but decided to investigate this potential relationship between predicted player value (signing bonus) and actual player contribution (WAR).

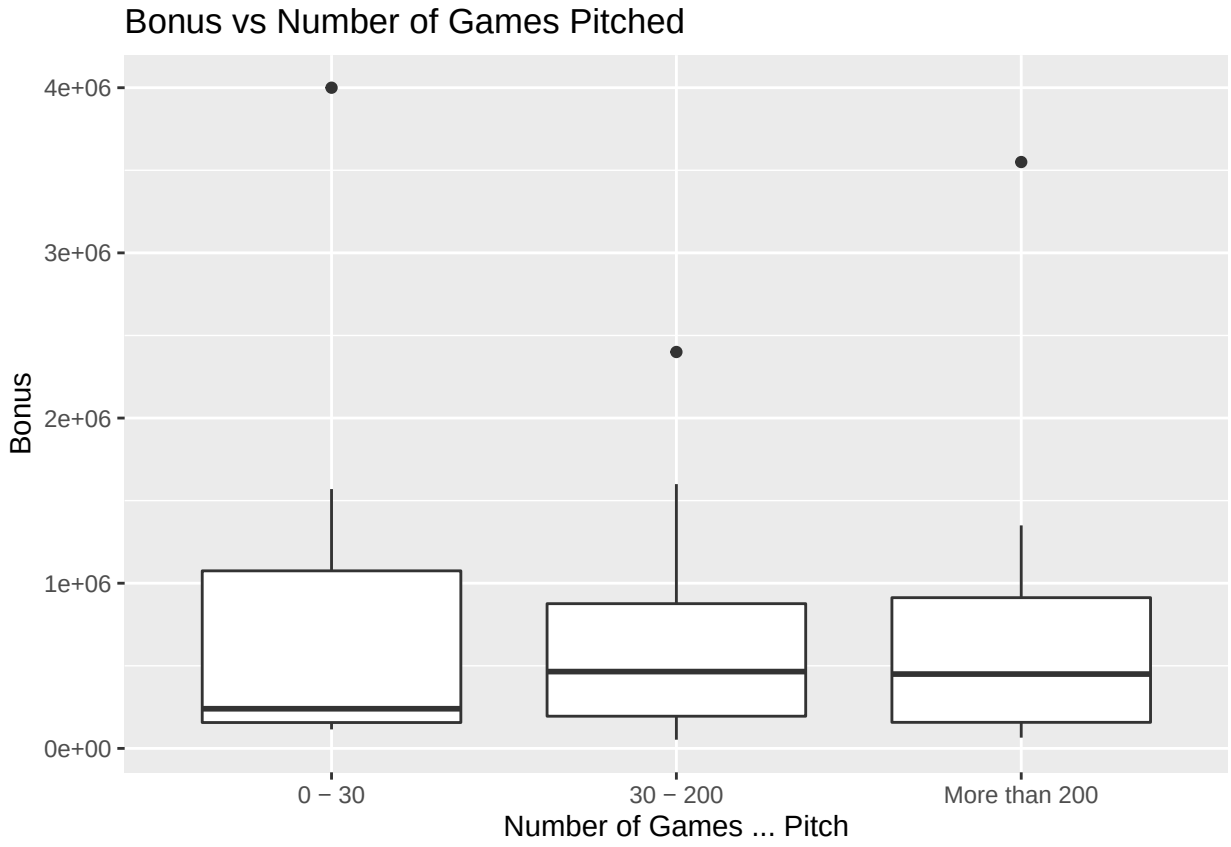
If a player has a larger WAR value, it means that the player is less likely to be replaced by another player. Our graph of Bonus vs WAR shows that there is a weak positive relationship between WAR and Bonus. The linear regression of WAR on Bonus has an R squared of 0.2614, which is not larger. So in general, the players with higher bonuses in fact do not necessarily perform better in the future. This is something we can study in the future final dataset.



We also look into the relationship between the bonus and the number of games played. We divide the players into 3 groups according to the number of games they played as batter and pitcher. As the figure 7 shows, there is no significant difference between the first quantile, medium, and the third quartile for those who played less than 30 games as batter and the players who played 30-300 games. However, the medium and third quartile bonus for players having more than 300 games as pitcher is larger than the medium bonus of the other 2 groups. The figure implies that there could be a weak positive relationship between the bonus and the number of games played as batter.

The figure 8 shows that there is no significant difference between the first quantile, and the third quartile of the 3 groups. However, the medium bonus for players having games as pitcher is smaller than the medium bonus of the other 2 groups. The figure implies that there could be a weak positive relationship between the bonus and the number of games played as pitcher.


Bonus vs Number of Games Batted
Bonus
6e+06
4e+06
2e+06
0e+00
0 − 30
30 − 300
More than 300
Number of Games ... Bat


Bonus vs Number of Games Pitched
Bonus
4e+06
3e+06
2e+06
1e+06
0e+00
0 − 30
30 − 200
More than 200
Number of Games ... Pitch

Conclusions

We have the following conclusions based on the draft 2005 data:

In our draft 2005 dataset, a large proportion of the players are pitchers, so probably in the final dataset, we can group the position of players, such as pitcher vs non pitcher. The most frequent recruitment time is during the 4 years of college in the draft 2005 dataset. As a result, when we study the influence of missing time on bonuses in the future, the recruitment time could be a factor we need to consider. Position, signing bonus, and number of games are some factors related to the WAR. We can analyze these factors in more detailed ways in the final dataset.

The next step in our analysis will be finding data to accurately represent missed time for each players and focus our investigation on how that missed time affects player development.

References

Petti, Bill. et al. "baseballr: The SportsDataverse's R Package for Baseball Data.", 2021, https://billpetti.github.io/baseballr/.

"MLB Stats, Scores, History, & Records." Baseball, https://www.baseball-reference.com/.

"The Official Site of Major League Baseball." MLB.com, https://www.mlb.com/.

Appendix

Code:

```
# Carolyn: distribution of positions
d05_final <- read.csv("Milestone_2.18.22_Data.csv", header=TRUE)
d05_final <- d05_final %>%
  arrange(desc(Bonus))

d05_200 <- d05_final[1:200,]

positions <- d05_200$Pos
positions <- as.factor(positions)
barplot(table(positions), names.arg=levels(positions),
        las=2, xlab = "", ylab="Count of Players",
        main="Distribution of Positions for 2005 Draft", col="#aadff0")
mtext(text = "Position",
      side = 1,
      line = 4)

table(positions)

# Aryaman: Box and whisker plot of signing bonus to position
Raw_05_Draft <- read.csv("Raw_05_Draft.csv")
for (i in 1:nrow(Raw_05_Draft)) {
  if (Raw_05_Draft$Pos[i] == "LHP" || Raw_05_Draft$Pos[i] == "RHP") {
    Raw_05_Draft$Pos[i] <- "P"
  }
}
Raw_05_Draft$Bonus <- parse_number(Raw_05_Draft$Bonus)
ggplot(Raw_05_Draft, aes(x = Pos, y = Bonus)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::comma) +
  ggtitle("Boxplot of signing bonus by position") +
```

```r
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Position") +
  ylab ("Bonus")



#Sia
milestone<-read.csv("Milestone_2.18.22_Data.csv")
type <- as.factor(milestone$Type)
levels(type)[levels(type)==''] <- 'NA'
no.na <- type[-which(type=="NA")]
no.na<-droplevels(no.na)
plot(no.na, main="Distribution of Recruitment Rime",
     xlab="Recruitment Time", ylab="Num of Players",
     col=rgb(0.2,0.4,0.6,0.6))


# Jonny
d05 <- read.csv("Milestone_2.18.22_Data.csv")

hist((d05$Bonus/1000), main = "Histogram of Signing Bonuses",
     xlab = "Bonus in Thousands of 2005 Dollars")

hist((d05$Bonus[d05$Bonus < 1000000] / 1000),
     main = "Histogram of Signing Bonuses Less than $1m",
     xlab = "Bonus in Thousands of 2005 Dollars")

ggplot(d05, aes(x = OvPck, y = (Bonus/1000))) +
  geom_point(aes(color = WAR), na.rm = TRUE) +
  scale_color_gradient(low = "red", high = "green") +
  labs(title = "Signing Bonus vs Pick Rank for 2005 Draftees",
       y = "Bonus in Thousands of 2005 Dollars", x = "Pick Rank in 2005 Draft",
       color = "WAR")

# Chenfei:
data1 = read.csv("./Milestone_2.18.22_Data.csv")
l1 = lm(WAR~Bonus, data = data1)
summary(l1)

plot(WAR~Bonus, data = data1, main = "Bonus vs WAR")
abline(l1, col = "red")

data1 = data1 %>%
  mutate(num_bat = ifelse(data1$G_bat > 300, "More than 300", ifelse(data1$G_bat > 30, "30 - 300", "0 -
data_bat = data1 %>%
  filter(!is.na(G_bat))
ggplot(data_bat, aes(num_bat, Bonus)) +
  geom_boxplot() +
  xlab("Number of Games - Bat") +
  ggtitle("Bonus vs Number of Games Batted")

data1 = data1 %>%
  mutate(num_pitch = ifelse(data1$G_pitch > 200, "More than 200", ifelse(data1$G_pitch > 30, "30 - 200"
```

```r
data_pitch = data1 %>%
  filter(!is.na(G_pitch))
ggplot(data_pitch, aes(num_pitch, Bonus)) +
  geom_boxplot() +
  xlab("Number of Games - Pitch")  +
  ggtitle("Bonus vs Number of Games Pitched")
```