Baseball data analysis

October 15, 2008

This paper uses two data sets that contain information on baseball players from as early as 1871. One set contained offensive statistics such as games played, at bats, and home runs while the other set contained more personal information such as the player's country of origin, birth date, and height. Baseball is currently the easiest sport of the four major American sports (baseball, basketball, football, and hockey) to analyze using statistics as baseball is the most individualized of these team sports. When a player is batting, it is just the batter versus the pitcher. The other players on the field have a much smaller effect on the batter than, for example, the effect an offensive line might have on the quarterback in football. I have chosen to look at how certain offensive categories have changed over the years and over the course of a player's career.

The first thing I needed to do was clean up the two data sets given to us. The first problem I noticed was that intentional bases on balls, hit by pitches, and sacrifice flies were missing for players in earlier years. Intentional bases on balls and hit by pitches were probably not separated from bases on balls and sacrifice flies were likely recorded as sacrifice hits until a certain year. Therefore I set these missing values to 0 in order to preserve the new variables I want to introduce. which year?

which years?

Next, I separated data from the dead-ball and live-ball eras. In 1920, Major League Baseball changed some of its rules to give more of an advantage to the hitter, creating the two eras. For most of my analysis, I did not use the data from before 1920 because any variation between the dead-ball era and the live-ball era could be due to rule changes instead of the hitters themselves. I will also look at the difference between the two eras in on-base plus slugging (ops) which I will talk more about shortly.

how does that affect the variables above?

I was also worried about the number of plate appearances (pa) affecting the data. I chose to look at plate appearances instead of at-bats becasue plate appearances accounts for walks and sacrifices. Some players had very few plate appearances that led to misleading statistics in the variables I wanted to use. After trying many different minimums, I decided to make the requirement 100 plate appearances for the year. Figure 1 looks at plate appearances for each year. Interestingly, there were dips at 1981 and 1994. After searching Wikipedia for these seasons, I realized that these dips were caused by strikes that shortened the seasons and consequently the number of plate appearances. These seasons should not affect my analysis as the two main statistics I will look at are rate statistics instead of counting statistics. — ~~what about~~ should you have used 50 pa's for the strike year?

good investigation

The last thing I did to clean the data was convert the dates in the players data set to numbers that represented the year and merged the birth year, debut year, and final year to the batters data. I will use these years to see how time affects certain offensive statistics to see if there is any relationship between these years and ops or stolen base success rate (sbsucc).

Next, I created new variables that I will use to analyze the data. I will use on-base plus slugging (ops) as a measure of a player's hitting ability. This is a much better measure than the traditional measures such as batting average, runs batted in, or home runs (hr). Studies have shown that a team's ops has a higher correlation with the runs scored by that team in the following season than any of the statistics given to us in the batting data set. Since scoring runs is the goal of baseball from an offensive standpoint it makes since that we use ops as our offensive measure. Another statistic I want to look at is a player's success rate when stealing bases. I calculated this as a player's number of stolen bases divided by the number of times he attempted to steal a base (sa). I plan on looking at how these variables have changed over time. Also, ✓

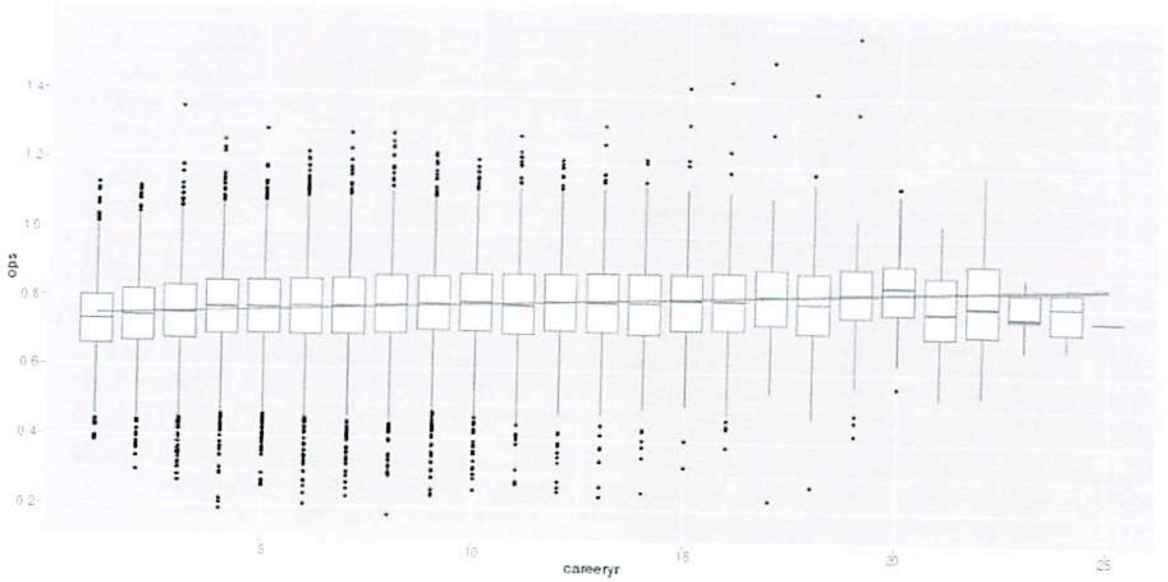

Figure 6: A boxplot of ops vs the player's career year with a linear model. There is one box for each possible career year. Again, the model shows a positive trend.

3 Stolen base success rate

The other statistic examined was stolen base success rate. I plotted this value against year, age, and number of steal attempts. Figure 8 suggests that base stealing has become more successful over the years. I believe this could be due to the decreased frequency of steal attempts in today's game. Next, I looked at the success rate versus the number of steal attempts. Figure 9 shows a positive trend as expected as we would assume that a player with a higher success rate would be given more attempts by his manager to steal a base than a player with a lower success rate. I finished by observing stolen base success rate vs age. There is a negative relationship between age and stolen base success rate for the top graph in Figure 10, but there is a positive relationship in the bottom graph of the same figure. The difference between the two graphs is that the lower graph has set the minimum number of stolen base attempts at 20. The negative relationship is expected as players tend to get slower as they get older, but why does the relationship become positve when we set the minimum number of attempts? Figure 11 seems to provide an answer. Again, the top graph is without restrictions and the bottom graph has the minimum number of attempts set at 20. The counts decrease much less in the 25 and 30 age bins then they do in the other bins. Therefore, by limiting the stolen base attempts, we got rid of more players in the extremes of the age range (35+ years old), many of whom probably had lower stolen base success rates since they were older.

nice exploration.

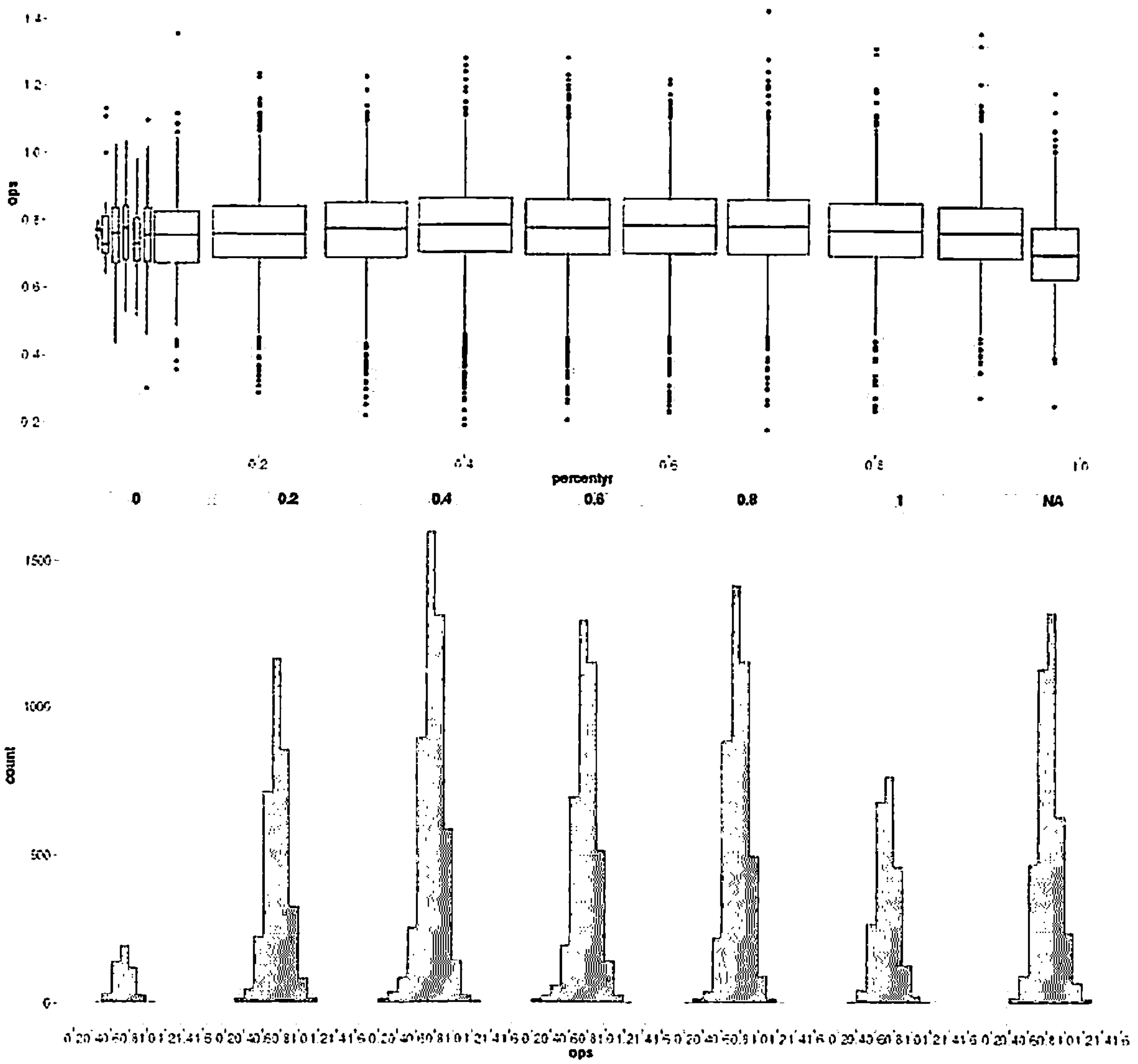

Figure 7: (Top) A boxplot of ops vs how far a player is into his career. There is one box for every .1 in percentyr. (Bottom) A histogram of ops counts for bins of size .2 in percentyr. The top graph shows the peak I was wanting to find when observing ops. I would have like to fit a smooth curve here but I got a memory error message. The bottom graph shows that these boxplots are normal.

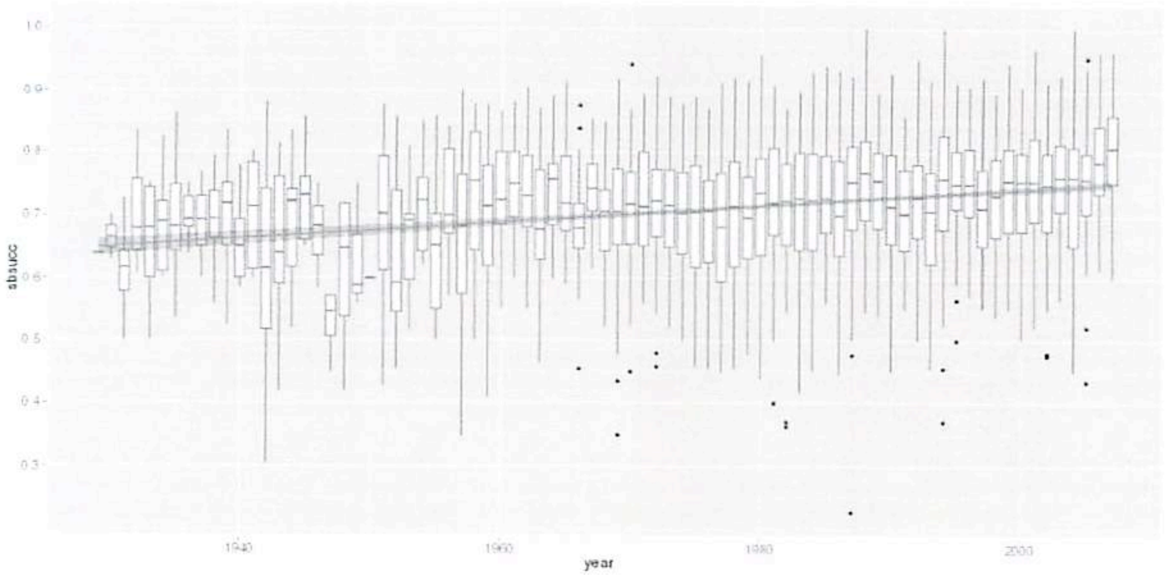

Figure 8: A boxplot of stolen base success rate vs year with a linear model. There is one box for each year. There is a positive trend.

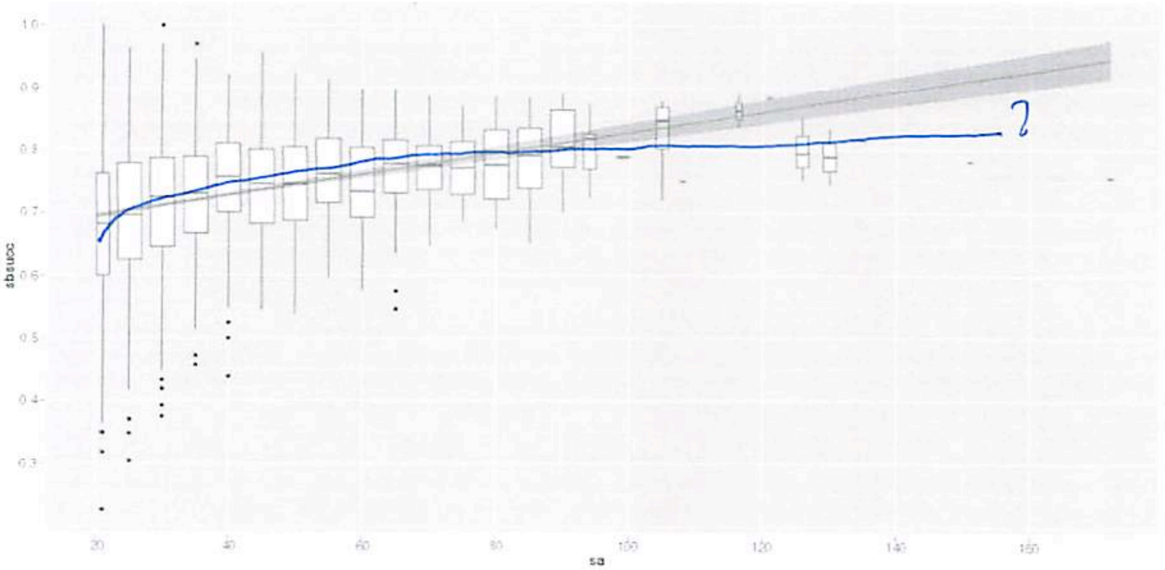

Figure 9: A boxplot of stolen base success rate vs number of steal attempts with a linear model. The boxes have a binwidth of 5 attempts. There is a positive trend between the two variables.

but over estimated by liner model.

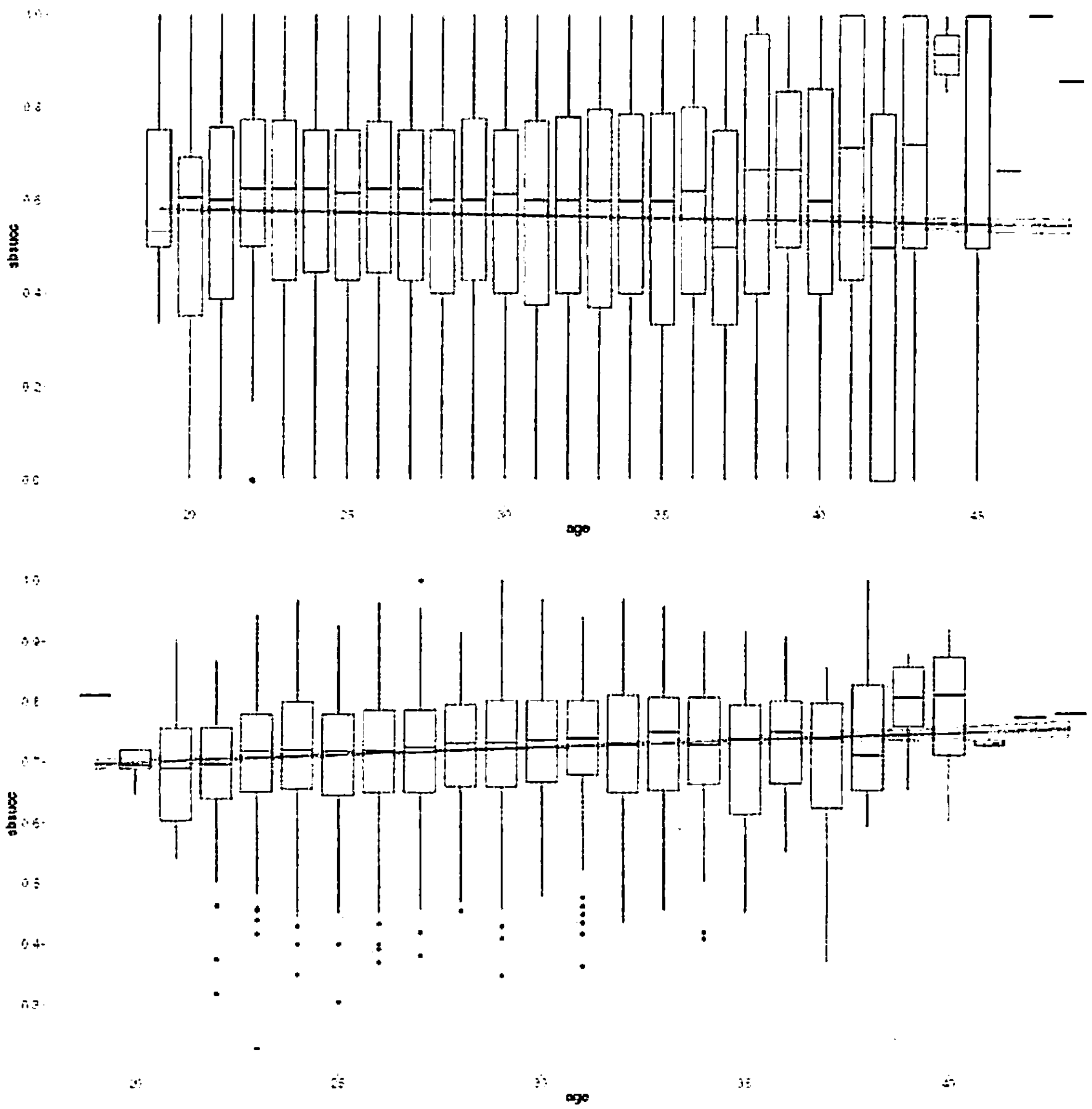

Figure 10: (Top) A boxplot of stolen base success rate vs age with a linear model. There is one box for each age. (Bottom) The same plot with the minimum number of stolen base attempts set to 20. The top graph shows a negative trend between the two variables until age 35. The bottom shows that there is a positive trend.

+ these guys disappeared

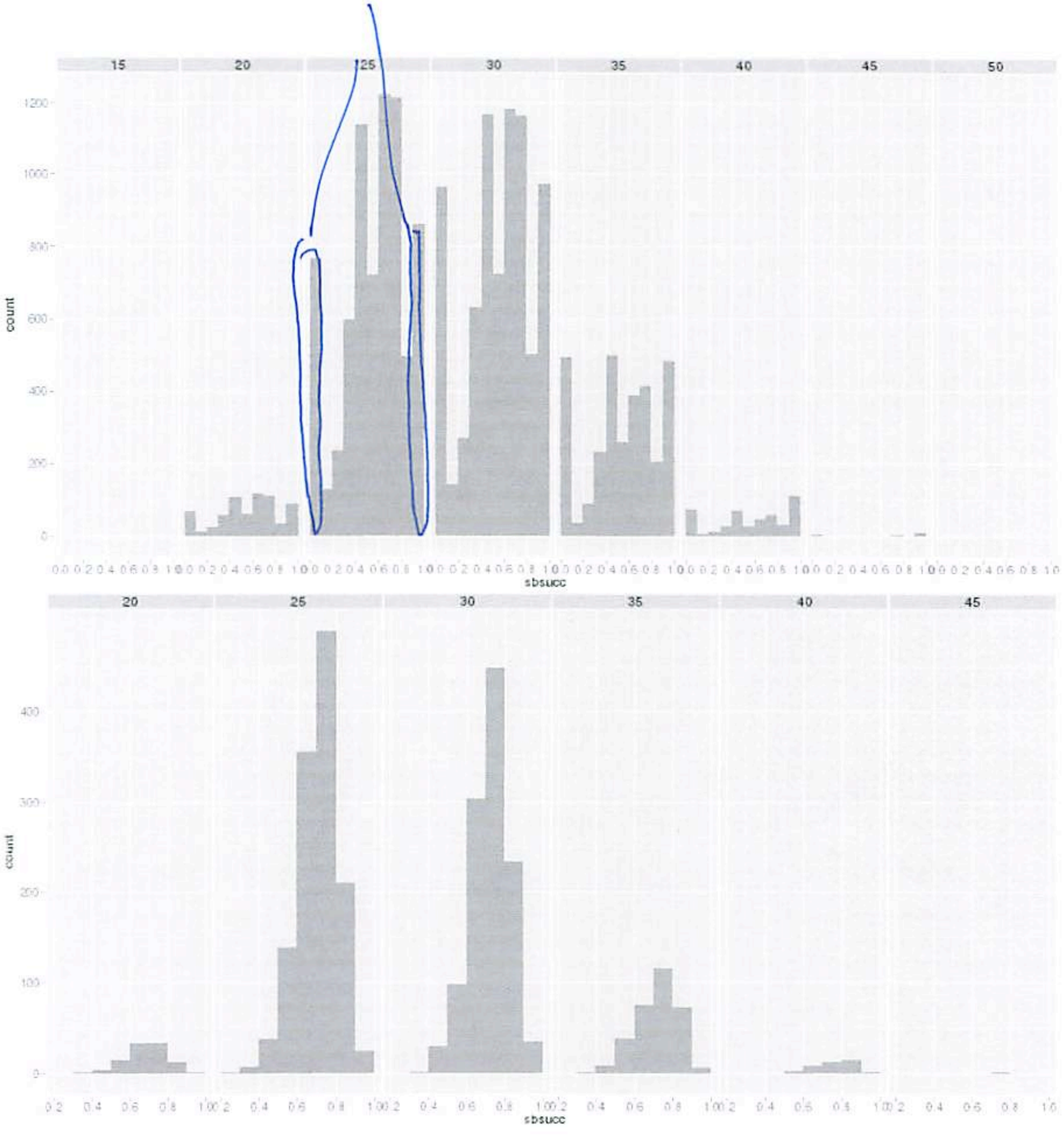

Figure 11: (Top) A histogram of stolen base success rate counts for bins of size 5 in year. (Bottom) The same plot with the minimum number of stolen base attempts set to 20. The counts decrease less in the bins for ages 25 and 30 than in other bins.

4 Conclusion

The main conclusion of my analysis is that baseball is becoming an increasingly offensive game. Home runs, ops, and stolen base success rates have increased over the years and show no sign of decreasing. This is partially by the design of Major League Baseball to increase offense in order to attract more fans. The first evidence to support this design is the rule changes in 1920 that created the live-ball era. Another was the addition of the designated hitter in the American League in 1973. When I analyzed this using the abliveball ? data set, I didn't see much difference between the leagues, probably because most pitchers will not have 100 at bats in a season. Since pitchers are generally the worst batters on the team, taking them out of the calculation minimizes the ops difference between the leagues. More research could be done to observe this effect on ops and home runs. I also could have looked into the years following the two strikes to see if the strike had any effect on player performance the following year. One thing I would have liked to analyze was the difference between steal attempts and stolen base success rate between the live-ball and dead-ball eras. I would expect that there would be a higher number of steal attempts in the dead-ball era as it was rarer for a batter to get into scoring position (second or third base) without stealing a base as ops was lower for this time period than in the live-ball era. Unfortunately, I was unable to conduct such an analysis since the number of times a player was caught stealing was not recorded for most players in the dead-ball era. ✓

There might have been an error in my calculation of ops. I was unsure if the data includes intentional bases on balls and hit by pitches in bases on balls or if sacrifice flies were also counted as sacrifice hits. This is potentially harmful to the actual ops values but it should not change the trends seen in the figures. An analysis could be done to determine if these statistics were double-counted.

A Code

```
library(ggplot2)

# saves data sets as b and p

b <- read.csv("batting.csv")
p <- read.csv("players.csv")

# sets NA entries in ibb to 0, ibb probably not always recorded, these would be
# in bb

b$ibb[is.na(b$ibb)] <- 0

# sets NA entries in sf to 0, sf probably not always recorded, these would be
# in sh

with(b, table(year, is.na(sf)))
b$sf[is.na(b$sf)] <- 0

# sets NA entries in hbp to 0, hbp probably not always recorded, these would be
# in bb

b$hbp[is.na(b$hbp)] <- 0

# create slugging percentage (formulas from wikipedia)

b$slg <- with(b, ((h - X2b - X3b - hr) + 2*X2b + 3*X3b + 4*hr + bb + ibb + hbp)  /
/ (ab + bb + ibb + hbp))

# create on-base percentage, (do you want to count ibb? - usually more due to
# situation than player)

b$obp <- with(b, (h + bb + ibb + hbp) / (ab + bb + ibb + hbp + sh + sf))

# creates on-base plus slugging, this is a much better measure of hitting than
# any of the given stats)

b$ops <- with(b, slg + obp)




# want to know plate appearances so that we can get rid of data with
# insufficient number of plate appearances

b$pa <- with(b, ab + bb + ibb + hbp + sh + sf)
```

```
# used plots to look at what would be a good minimum number of ab's

qplot(year, pa, data=b)
# two years were low, 1981 and 1994, both had strikes during the season-
# wikipedia)
qplot(year, pa, data=b, geom="boxplot", group=year)



# create stolen base success percentage (sa = steal attempts)

b$sa <- with(b, sb + cs)

b$sbsucc <- with(b, sb / sa)

# changes years from factors to numbers

parse_date <- function(x) as.Date(strptime(x, "%m/%d/%Y"))
date_vars <- c("birth", "debut", "final", "death")
p[date_vars] <- lapply(p[date_vars], parse_date)

# merge important years into b


year <- function(x) 1900 + as.POSIXlt(x)$year


# birth year

years <- with(p, data.frame(
  id = id,
  byear = year(birth),
  dyear = year(debut),
  fyear = year(final)
))


b4 <- merge(b, years, by ="id")

# list the decade of the given player

b4$decade <- round_any(b4$year, 10, floor)

# age

b4$age <- with(b4, year - byear)

# merge height and weight
```

```
size <- with(p, data.frame(
id = id,
weight = weight,
height = height
))

b5 <- merge(b4, size, by ="id")

# create weight classes and height class to use for sbsucc

b5$wtclass <- round_any(b5$weight, 5)
b5$wtclass2 <- round_any(b5$weight, 15)
b5$htclass <- round_any(b5$height, 5)

# want to only look at live-ball era (1920 and on, wikipedia)

liveball <- subset(b5, year >= 1920)

# look at deadball era to compare with liveball era

deadball <- subset(b, year < 1920)

# sets minimum plate appearances to 100, looks at a plot to see if this was a
# good number, also looked at 500, 250, 200, 50 and 20

abliveball <- subset(liveball, pa >= 100)
#qplot(year, pa, data=abliveball, geom="boxplot", group=year)

abdeadball <- subset(deadball, pa >= 100)

# want to look at what how many years a player has been in the league, assumes
# they are in the league every year from their debut year until the final year

abliveball$careeryr <- with(abliveball, year - dyear + 1)

# look at the distribution of the lengths of careers

qplot(careeryr, data=abliveball, geom="histogram", binwidth=1)

# also want to look at what part of their career the given player is in for
# that year, will be missing for players that have not retired

abliveball$percentyr <- with(abliveball, careeryr / (fyear - dyear + 1))

# look at the trend of ops based on year

qplot(year, ops, data=abliveball, geom="boxplot", group=year)
```

```
# lost 4705 rows, all before 1955, why?

earlyabliveball <- subset(abliveball, year <=1950)
head(earlyabliveball)

# did not change NA entries in ibb and sf before calculating slg and obp, also
# found a mistake in careeryr and percentyr, have now corrected the order, will
# not have this problem agian

# set a minimum number of sb attempts (sb + cs), don't want to remove these
# players from ops graphs becasue they could still have a significant number of
# at bats without a significant number of sb attempts

sbabliveball <- subset(abliveball, sa >= 20)
qplot(sbsucc, data=sbabliveball, geom="histogram", binwidth=.01)

# tried min number of attempts at 20, 30, and 50
# tried binwidth of .1 and .01

# place percentyr into bins so you can use the bins as facets

abliveball$pybin <- round_any(abliveball$percentyr, .2)

# place ages into bins of size 5 so you can use the bins as facets

sbabliveball$agebin <- round_any(sbabliveball$age, 5)

abliveball$agebin <- round_any(abliveball$age, 5)

# look at only players from 1973 on to see if the DH affects the ops of the AL

dh <- subset(abliveball, year >= 1973)


#######################################################################

# end of data cleaning, plots used

#######################################################################

# look at minimum number of plate appearances

qplot(year, pa, data=abliveball, geom="boxplot", group=year) +
geom_smooth(aes(group = 1), method="lm")

# ggsave(file = "pavyear.png", width=10, height=5)
# dips in 1981 and 1994 due to strikes - wikipedia

# hr vs year (live-ball era)

qplot(year, hr, data=abliveball, geom="boxplot", group=year)
# ggsave(file = "hrvyear.png", width = 10, height = 5)

# hr vs year (all years)

qplot(year, hr, data=b, geom="boxplot", group=year)
# ggsave(file = "hrvallyears.png", width=10, height=5)


#######

# ht and wt vs year

#######

qplot(year, weight, data=abliveball, geom="boxplot", group=year)
# ggsave(file = "weightvyear.png", width = 10, height = 5)

qplot(year, height, data=abliveball, geom="boxplot", group=year)
# ggsave(file = "heightvyear.png", width = 10, height = 5)


#######

# ops graphs

#######

# look at the distribution of ops

qplot(ops, data=b, geom="histogram", binwidth=.01)
# also tried binwidth of .05 and .001
# ggsave(file = "opscount.pdf", width=10, height=10)

# look at the trend of ops based on year (with and without y = .75)

qplot(year, ops, data=abliveball, geom="boxplot", group=year) +
geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "opsvyear2.png", width=10, height=5)

qplot(year, ops, data=abliveball, geom="boxplot", group=year) +
geom_smooth(aes(group = 1), method="lm") + geom_hline(intercept = 0.75,
colour = "red")
# ggsave(file = "opsvyear.png", width=10, height=5)
# look at this trend for deadball as well

qplot(year, ops, data=abdeadball, geom="boxplot", group = year) +
geom_smooth(aes(group = 1), method="lm") + geom_hline(intercept = 0.75,
colour = "red")
# ggsave(file = "opsvsdeadyear.png", width=10, height=5)

# look at ops versus careeryr

qplot(careeryr, ops, data=abliveball, geom="boxplot", group=careeryr) +
geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "careeryrvops.png", width=10, height=5)

# look at ops versus percentyr

qplot(percentyr, ops, data=abliveball, geom="boxplot", group=signif(percentyr,
digits = 1))
# also tried digits = 2, too many boxes
# ggsave(file = "percentyrvops.png", width=10, height=5)

# make sure the histograms are normal for each boxplot

qplot(ops, data=abliveball, geom="histogram", binwidth = .1, facets = .~ pybin)
# ggsave(file = "opsctscheck.png", width=10, height=5)

# look at ops versus age

qplot(age, ops, data=abliveball, geom="boxplot", group=age) +
geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "agevops.png", width=10, height=5)

# look at the difference between leagues since the designated hitter was first
# used in the AL

qplot(year, ops, data=dh, geom="boxplot", colour=lg,
group=interaction(lg, year))
# ggsave(file="alvnl.png", height = 5, width = 10)



#######

# sbsucc graphs

#######

# look at the distribution of sbsucc when we set min attempts
```

```
qplot(sbsucc, data=sbabliveball, geom="histogram", binwidth=.01)
# ggsave(file = "sbdist.pdf", width=10, height=10)

# look at the trend of sbsucc based on year

qplot(year, sbsucc, data=sbabliveball, geom="boxplot", group=year) +
geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "sbsuccvyear.png", width=10, height=5)

# look at sbsucc vs weight class

qplot(wtclass, sbsucc, data=sbabliveball, geom="boxplot", group=wtclass,
xlab = "Weight", ylab = "Sb success")
# ggsave(file = "sbsuccvweight.png", width = 10, height = 10)

# look at sbsucc vs height

qplot(height, sbsucc, data=sbabliveball, geom="boxplot", group=height,
xlab = "Height", ylab = "Sb success")
# ggsave(file = "sbsuccvheight.png", width = 10, height = 10)

# sbsucc vs age (with and without the limit on attempts)

qplot(age, sbsucc, data=sbabliveball, geom="boxplot", group=age) +
  geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "sbsuccvage.png", width = 10, height = 5)

qplot(age, sbsucc, data=abliveball, geom="boxplot", group=age) +
  geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "agevsbsucc.png", width = 10, height = 5)

# why are the two previous graphs so different?

qplot(sbsucc, data=sbabliveball, geom="histogram", binwidth = .1, facets =
.~ agebin)
# ggsave(file = "sbsucc_check.png", width = 10, height = 5)

qplot(sbsucc, data=abliveball, geom="histogram", binwidth = .1, facets =
.~ agebin)
# ggsave(file = "sbsucc_origcheck.png", width = 10, height = 5)


# sbsucc vs sa

qplot(sa, sbsucc, data=sbabliveball, geom="boxplot", group=round_any(sa,5)) +
geom_smooth(aes(group = 1), method="lm")
# ggsave(file = "sbsuccvsa.png", width = 10, height = 5)
# sbsucc vs percentyr

qplot(percentyr, sbsucc, data=sbabliveball, geom="boxplot",
  group=round_any(percentyr,.05))
# ggsave(file = "sbsuccvpercentyr.png", width = 10, height = 10)



########################################################

# end of plots, calculations

########################################################

#######

# hr calculations

#######

# look at average home runs of liveball and deadball eras

mean(abdeadball[,12])
# [1] 1.668974

mean(abliveball[,12])
# [1] 8.655226



#######

# ops calculations

#######

# look at average ops of liveball and deadball eras

mean(abliveball$ops)
# [1] 0.7762875

mean(abdeadball$ops)
# [1] 0.6904932



#######
```

```
# sbsucc calculations

#######

# how well does height predict sbsucc?

h = lm(sbsucc ~ height, data=sbabliveball)
# summary(h)

# how well does weight predict sbsucc?

w = lm(sbsucc ~ weight, data=sbabliveball)
# summary(w)

# what if we put both height and weight together?

hw = lm(sbsucc ~ weight + height, data=sbabliveball)
# summary(hw)

# add age?

a = lm(sbsucc~ weight + height + age, data=sbabliveball)
# summary(a)

# none of these have a decent R-squared, models probably aren't linear




write.table(abliveball, "baseball-clean.csv", sep=",", row=F)

write.table(abdeadball, "baseball-clean-dead.csv", sep=",", row=F)
```