

STAT 405 Project 1: August Born

October 2, 2009

1 Introduction

This report is based on the data collected during the so-called live ball era, which dates back to the 1920's. Its counterparty is dead ball era, which has very different rules with current baseball games. Especially, the dead ball era is characterized by extremely low-scoring games[6].

Our report focuses mainly on the birth distribution of players and possible effects of it on their performance on the field. We start with the "player" dataset, and outside data is introduced from Centers for Disease Control And Prevention (CDC) when the birth distribution of the original dataset is compared with nation-wide population. Two major factors are found responsible for the unusual accumulation of August-born players. The first factor is the natural birth trend by month in US, which normally contributes a peak in between August and September. Besides the demographic explanation, the so-called "little league effect" is also responsible. It seems that the long-last enrollment cut-off age rule of little league teams favors especially August players. Our explanation is that the two factors together contribute to the unusual accumulation of August-born players. ✓

The relative performance of August-born player is the next concern in this report. We begin analyzing this by searching and comparing August born players in Hall of Fame with players of different birth month. The comparison shows hardly any advantage of August born. Since the small sample size of Hall of Fame players, two statistics are introduced into the report: RC for evaluation of players' batting skills; strike-out to walk ratio for measuring players' pitching skills. The two statistics require "batting" and "pitching" datasets, which are obtained from baseball databank. The two statistics provide more precise access to further comparison of August born and others. The result shows no significant difference between August-born and others.

Based on previous analysis, we find the enrollment age cutoff of little league tends to favor mediocre baseball players born in August systematically, although it is not the only culprit to blame, since natural birth distribution also contributes some degree to the result.

Some suggestions are proposed to correct the little league effect on major league. Problems and shortcomings of reasoning and conclusion are also raised, mostly due to lack of more relevant data.

This report is written centering on the following questions:

- Why do US born baseball players show such special birth month distribution?
- Is nationwide birth seasonality the only reason for the phenomenon?
- How does August-born perform relative to players born in other months?
- What can be done to correct the situation? — Does it need correcting?

2 Keyword

birth month, little league effect, age cutoff date

3 Data Cleaning

Besides batting and players dataset, we also used pitching dataset which was obtained from the Baseball Databank. Since our goal is to explore the birth seasonality of US baseball players, the first thing we did was to separate the player dataset into US-born and Non-US born players. For both datasets, date was converted into number and month and year information of players' birthday were extracted and merged into the original datasets. In order to explore the players' monthly birth amount change with time, we decided to look at the monthly birth pattern for every twenty years. Because the data size for 80s is small, we grouped them with the period of 1960-1980. We have also looked at population change for every ten years and found it appropriate to set the period as twenty years since the patterns are similar. The number of people who were born in different months was summed up. The overall US population data from 1989-1997 (Curtin SC, Park MM, 1999) was merged into the US players dataset so we can make a comparison of these two and find out if there is any unique birth pattern for US baseball players regardless of the US population influence.

We created two major new variables rc and sown as two indicators to measure the players' batting and pitching abilities respectively. RC stands for runs created and has the following formula: $((h + bb) * (tb)) / (ab + bb)$. SOWN stands for strikeout to walk ratio which can be calculated as so / bb . Because ab and bb are denominators for rc and sown, we replaced those ab and bb that are equal to zero with NA.

After we merged the batting and pitching dataset with players, we encountered the problem that some ids have played for multiple years. Since we are supposed to know the average rc value for every player, we consolidated the data by id to obtain mean of rc or sown for each player and every year.

what are these abbreviations?

4 The Distribution of Birth Month of U.S. Players vs non-U.S. Players

We are interested in the influence of individual players' date of birth on their performance and career as well as the sport itself.

To begin with, we first look at the distribution of birth month of U.S. players and non-U.S. players. Referring to Figure-01 U.S. born players have prominently distinct distribution of birth trend with respect to non-U.S. players. An accumulation of August born players can be observed in the figure and a birth trough in June. The birth peak of U.S. born players' in August contributes more than 10.25% births in the whole year, while the trough in June ~~only~~ consists of no more than 7.25% births of annual total.

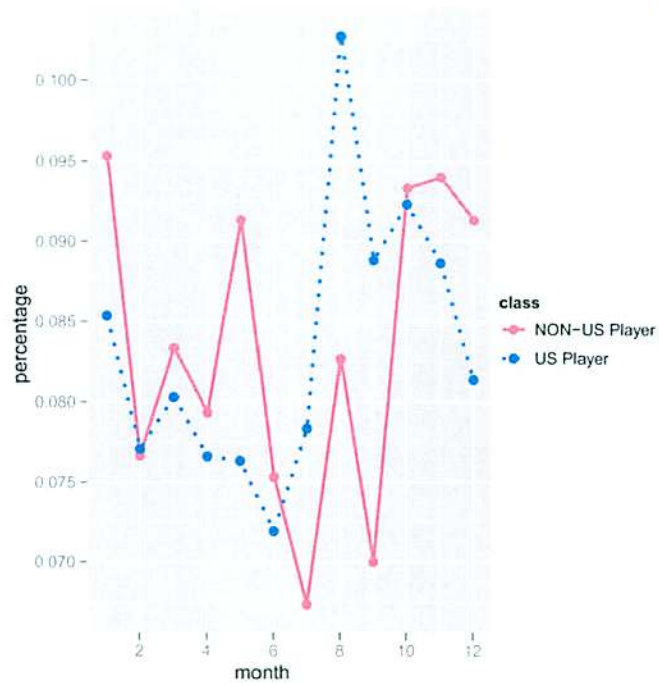
~~Yet~~ Non-U.S. born players illustrate a totally different trend with U.S. born. Birth dates of foreign born players tend to cluster at around December and January, with a peak at December of approximately 9.5% of total live births. In contrast to the U.S. curve, the non-U.S. curve shows a trough in July with 6.5%, and the bounce-back in August is also considerably small, only reaching 8.25% of total births. The comparison is sharp. We think that there must be something special in U.S. Two reasons are most convincing in this case: the natural seasonality of U.S. births or some structural factor in major league's selection process.

plausible

5 What Is The Reasoning Behind U.S. Players' Particular Birth Month Distribution?

It is reasonable to first consider the natural seasonality of birth by month, since it is logically the most direct response. Researches in the field of social biology and eugenics also confirm the guess. Daniel Seiver (1985)[7] points out that seasonal variation in American births is one of the great demographic regularities, which births in September exceed births in April by 10-20 percent in every year of the postwar period, while the pattern for the other months of the year is also surprisingly stable. His research basically focuses on seasonality of birth from 1947 to 1976. To better illustrate his idea, we extract data from Vital Statistics, CDC (1999)[2] and set up Table 1.

are you using IRFS?



I'd make this plot wider & shorter

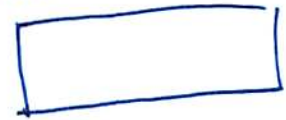


Figure 1: Birth month of U.S. born baseball players compared to foreign born players. Notice the obvious difference in between at August. U.S. trend is special in this case.

Month	Births	Percentage	Month	Births	Percentage
January	2916198	8.10%	July	3177368	8.83%
February	2731961	7.59%	August	3194077	8.87%
March	3028342	8.41%	September	3115876	8.66%
April	2904538	8.07%	October	3047931	8.47%
May	3044568	8.46%	November	2867539	7.97%
June	2995540	8.32%	December	2976137	8.27%

you can probably omit this & just show figure 3.

Source: *Trends in the Attendant, Place, and Timing of Births, and in the Use of Obstetric Interventions: United States, 1989 - 97*, Division of Vital Statistics, CDC

We sum up the original data from the report year by year, and calculate the percent of births in particular months within the 9 years. During the period of 1989 - 1997, U.S. birth seasonality shows a similar trend as Daniel Seiver concludes. However, the birth peak is in August, which contains 8.87% of total births per year. The two obvious troughs in the trend are February (7.59%) and November (7.97%). There is trivial difference with the experience of postwar period, yet it still catches most of the particular trend of U.S. birth seasonality.

The U.S. birth seasonality is unique, according to Ursula Cowgill (1966)[1], the European countries that she examines have contrasted pattern of seasonality (refer to Figure-02).

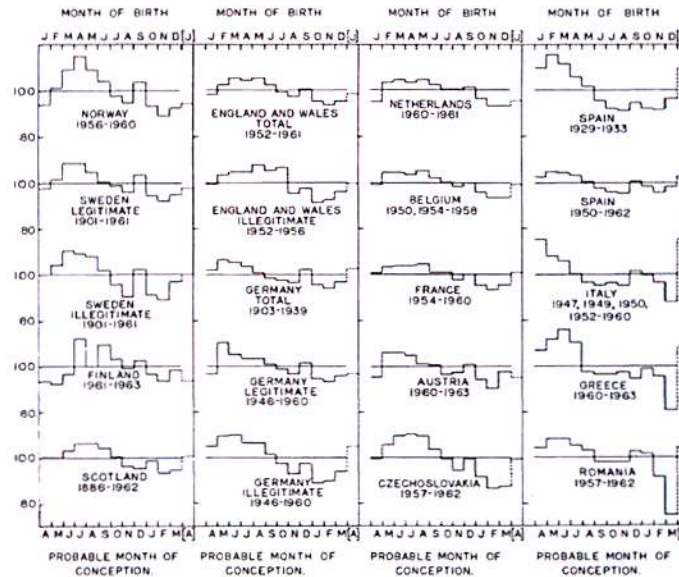


Figure 2: Seasonality of Birth in European Countries. Source: *Season of Birth in Man. Contemporary Situation With Special Reference To Europe And Southern Hemisphere*, Ursula Cowgill

Almost all European countries in the graphic show a concentration of births in the first half of year during the period studied. This can explain the distinct patterns of U.S. born players' birth date and of non-U.S. born players'.

Thus, we can say that the particular trend of U.S. natural seasonality is an important reason why there are so many August born baseball players in the data we have. However, can we draw the conclusion that this is the only reason?

great investigation

To answer the question, we input the data in the table to R, and compare the trend with U.S. born baseball players' birth dates. The result is Figure-03.

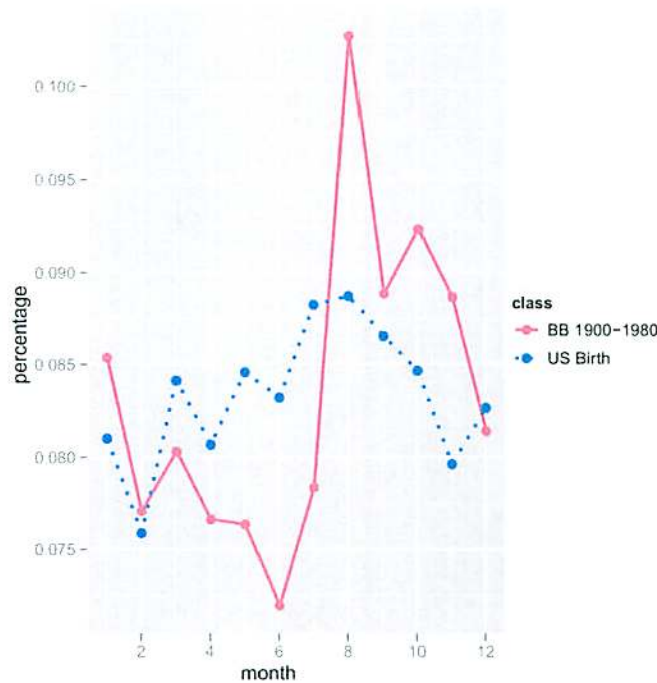


Figure 3: Birth Date of Baseball Players vs Seasonality of U.S. Birth Trend

The birth date distribution of baseball players in our data set has some similarity with respect to the seasonality of U.S. nation-wide birth trend. First of all, the births in spring and winter are at relatively low level in both cases. And the peak seasons of both kick in at August. However, the distinction is even more prominent than similarity. The nation-wide seasonality shows a gradual rising trend from January to August, after reaching the peak of nearly 9%, it begins to slowly decline, until ending at a similar level in December. Yet the baseball player curve shows a sharp decrease in June and an even sharper rise of births in August. The percentage of Baseball players born in August is higher than the percentage of total American babies born in August by almost 1.25%. This is a clear contrast, which means natural seasonality alone may not be able to explain the abnormal fall and rise of percentage of August born players'.

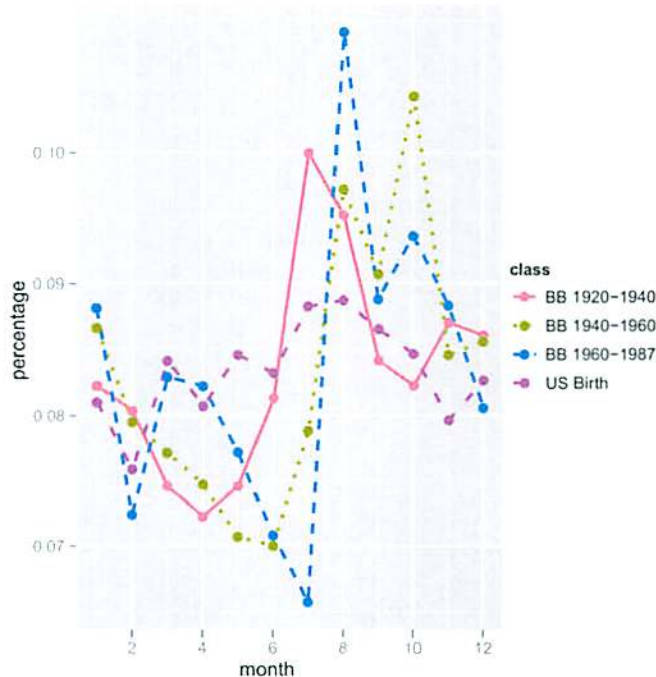
After searching and study, we find a so-called "little league age cutoff effect" responsible for the phenomenon.

6 Age Cutoff Effect

In the bestselling book "Outliers", Malcolm Gladwell describes a situation where children born on January 2nd has the most advantage applying for hockey team since Canadian hockey team set their age cutoff date at January 1st, which makes January born children almost one year stronger than December born. Thus, the kids born at the "right time" end up with higher possibilities to be chosen by a coach. In the sport of baseball in U.S., similar phenomenon also exists. It is called the little league effect.

Little League Baseball and Softball is founded by Carl Stotz in 1939 as a three-team league in Williamsport, Pennsylvania. Several specific divisions of Little League baseball and softball are available to children ages

5 to 18. When the July 31 league age determination date is settled upon in the mid 1940s[8], Little League is confined to Pennsylvania and a few other states, and it is the only youth baseball organization of any significant size. The function of the age cutoff date, July 31, is that it made the league enroll more children born on August. This can back to the population of children born between 1930 and 1940. The age cutoff date of July 31 just starts to show impact on the children of this period, so the August peak effect is not so obvious, and even move forward to July (see Figure-04).



Again -shape would make it easier to compare trends

Figure 4: U.S. birth seasonality compared to different periods of baseball players birth months

Later on, the age limit of Little League changes into 12-year old below. In the 1950s, other youth baseball organizations are formed, primarily for young teens. These organizations and Little League get together and agree to use one date (July 31). Thus, the greater August peak formed in 1940-1960 can be explained by the broad adoption of age cutoff date. At the same time, we notice there is also an obvious peak in October. This may result from the large change from the US birth seasonality during this period of time. From the perspective of little league effect, the greater peak in October can result from the inadequate spread of little league in the U.S. in the first decade of this period.

When it comes to the era of 1960-1987, with the age cutoff rule becoming a regulation recognized by the baseball circle all around the world, the remarkable August peak is then expressed perfectly by age cutoff date effect. And the trough in July also indicates children born in this month are disadvantaged worst by the rule.

good points

However, we cannot figure out the reason of the August peak during 1900-1920, when there should be no little league effect at all (see Figure-05). The August birth peak may originate from an unusual seasonality trend in the period, where births in August climb to rocket high. Unfortunately, we cannot verify this assumption, since no relevant monthly data is found concerning this era.

Other baseball leagues, like Minor League, Major League, just follow Little League's lead on age cutoff date. There could be seasonal variation on births, which helps represent the extra fluctuation of the every

I also wonder if you could look at the age a player debuts - do players with birthdays in August tend to be older?

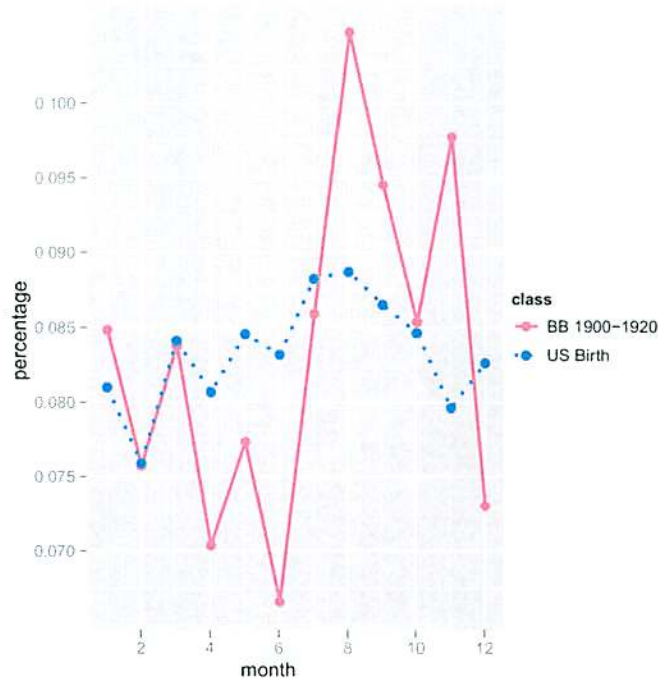


Figure 5: U.S. birth seasonality compared to 1900 - 1920

20-year data. Still, we believe that August-born peak among U.S. players can be explained by the little league age cutoff date for a large part.

That is to say, the age cutoff date provides August-born children with significant advantage when they apply for Little League. Thus, we couldn't help asking one question:

Are players born in August superior to others?

7 How does August-born perform relative to players born in other months?

7.1 Intuitive Evaluation

While the All Star example seems to support the argument above (see Figure-06), we think the criterion is biased.

7.1.1 The All Star Approach

The Major League Baseball All-Star[4] Game is an annual baseball game between players from the National League and the American League, currently selected by a combination of fans, players, coaches, and managers. Each league's All-Star team consisted of 32 players, including fan voting 8 players, player voting 16 players, manager selection 8 players, final vote 1 player. The selection of players in All Star based more on the popularity of the player, instead of totally focusing on the ability of the player. It is not based on objective data collected during the game process. As a result, we cannot judge whether a player is successful or not simply on this.

→
 don't need
 a heading
 for such
 a small
 section

what bin width did you use?

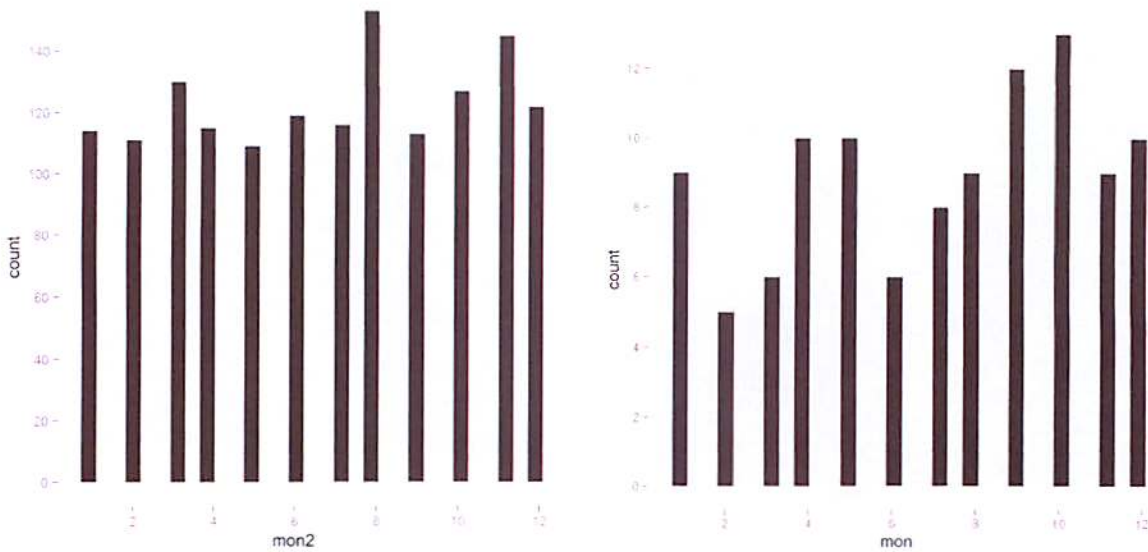


Figure 6: (Left) Birth month distribution of All Star players; (Right) Birth month distribution of Hall of Fame players. Notice that although the All Star case shows some kind of a peak in August, yet we have reasons to believe that the measure itself cannot reflect a player's ability, and the slight margin compared to others is just a reflection of natural seasonality.

7.1.2 The Hall of Fame Approach

Instead, we prefer the Hall of Fame approach[5]. Among baseball fans, "Hall of Fame" means the pantheon of players who have been enshrined in the Hall. The voting team would consider the scoring, defending, and other abilities of the players. It has strict criterion during the selection based on collected data of different players. It can objectively show the ability of a player. That is why we choose it as a factor of showing the successful players' birth concentration. Referring to Figure-07, August does not contain the most Hall of Fame players. This is to say, from a Hall of Fame perspective, August born are not superior.

7.2 Indices Evaluation

To better evaluate relative performance of August-borns, we introduce the following two statistics[3].

7.2.1 Runs Created (RC)

In the 1970, a night watchman, Bill James, studied baseball box scores long into night and self published books filled with his findings. These yearly books, entitled The Baseball Abstract, firstly put forward the new notion of measuring offensive behavior in baseball, which called runs created.

According to James, the philosophy of offense could be explained as followed two axioms:

- A ballplayer's purpose in playing baseball is to do those things which create win for his team, while avoiding those things which create losses for his team.
- Wins result form runs scored. Losses result from runs allowed.

Deduction based on the two axioms, we could form a conclusion: an offensive player's job is to create runs for his team.

this sentence doesn't para very well

In order to get a feel for how many runs a player can be credited with actually having produced, Bill James devised a model called runs created (RC). The basic version of the runs created formula is

$$\frac{(H + BB) \times (TB)}{(AB + BB)}$$

There are two essential elements to an offense: its ability to get on base, and its ability to advance runners. In the RC formula, the first part of the numerator, the H + BB, is the "on base" portion of the formula, while the second part of the numerator, the TB, represents the "advancement" part. The denominator is roughly the number of opportunities.

Runs created are believed to be an accurate measure of an individual's offensive contribution, because when used on whole teams, the formula normally closely approximates how many runs the team actually scores. Even the basic version of runs created usually predicts a team's run total within a 50% - otherwise just

Therefore, now we use run created to measure the average offensive abilities of players in terms of different birth month to speculate its pattern. *a comment*

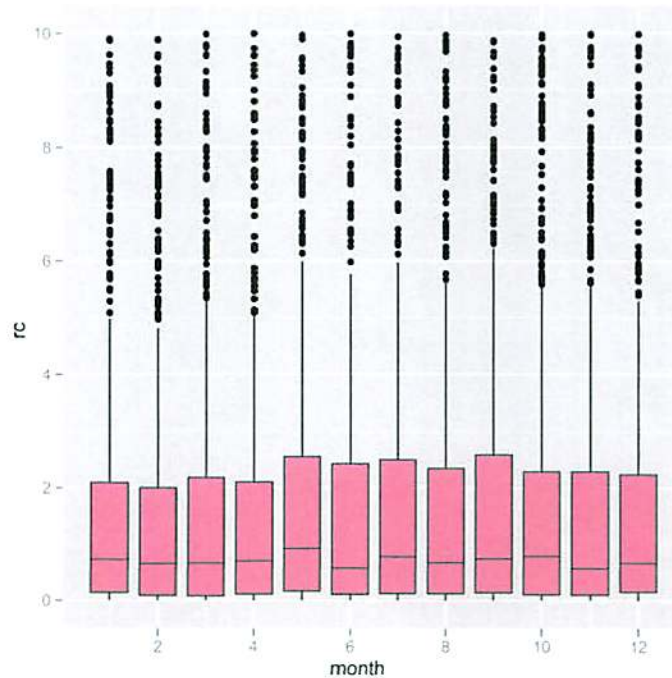


Figure 7: RC index for U.S. born players by birth month. Notice that medians of the boxplots show randomness; no sign of high index for August.

At a first glance of the above picture, there seems no significant discrepancy of offensive performance between month and month. But in order to reach more convincing conclusion, we perform eight times of permutation on the RC variable to see if any changes in the pattern. The following nine pictures show the original picture with comparison to the other permutation pictures. From these, we could obtain that the RC hold steady with the change in month, which exhibits no significant correlation between the offensive abilities of a player and his birth month.

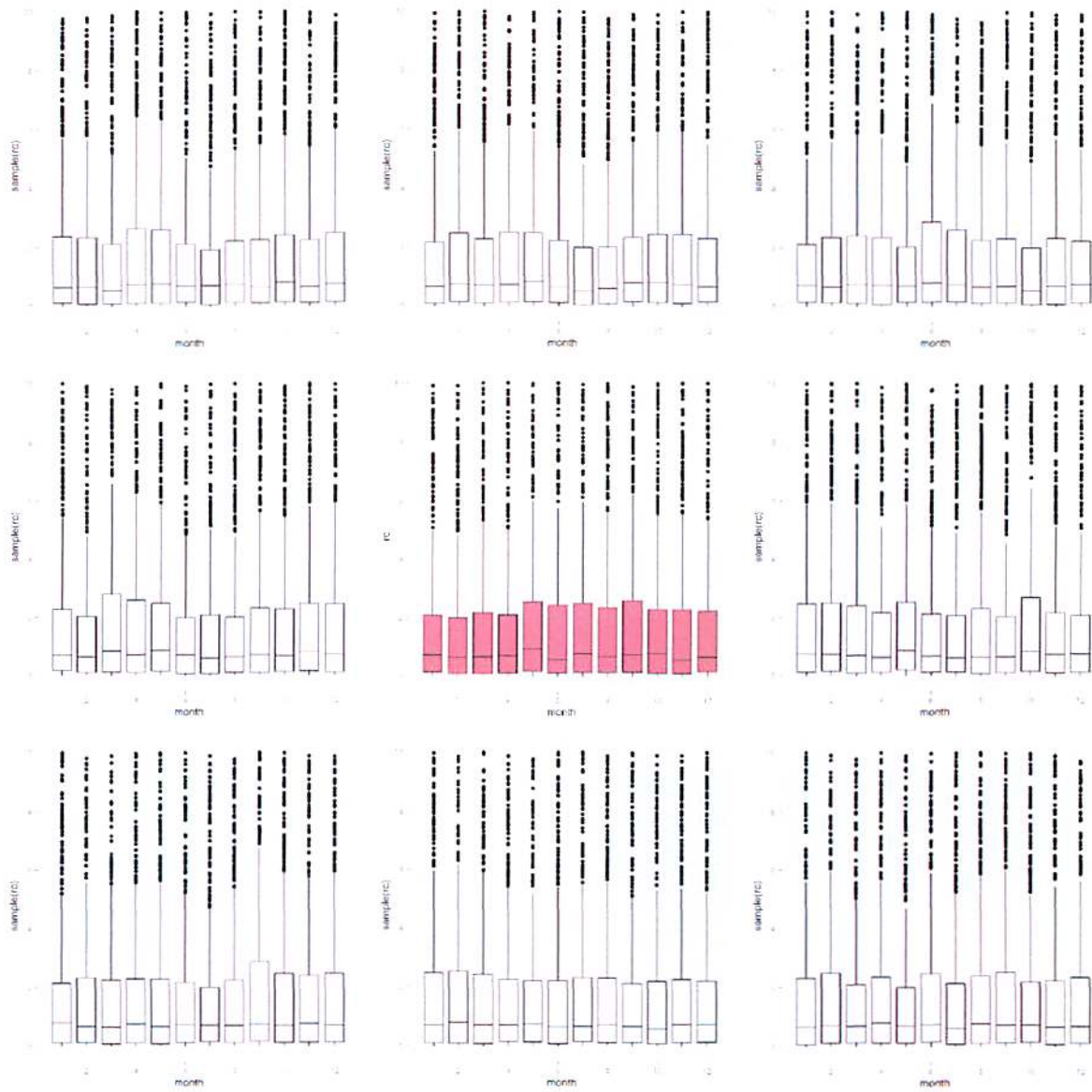


Figure 8: Permutations of RC. The graphic in the middle is the original RC boxplot. Notice that the permutations show no significant difference with respect to the original one, which means it is highly possible that August born shows no advantage in batting skills.

Good - but would be even better if true result wasn't highlighted - then you really have to look hard

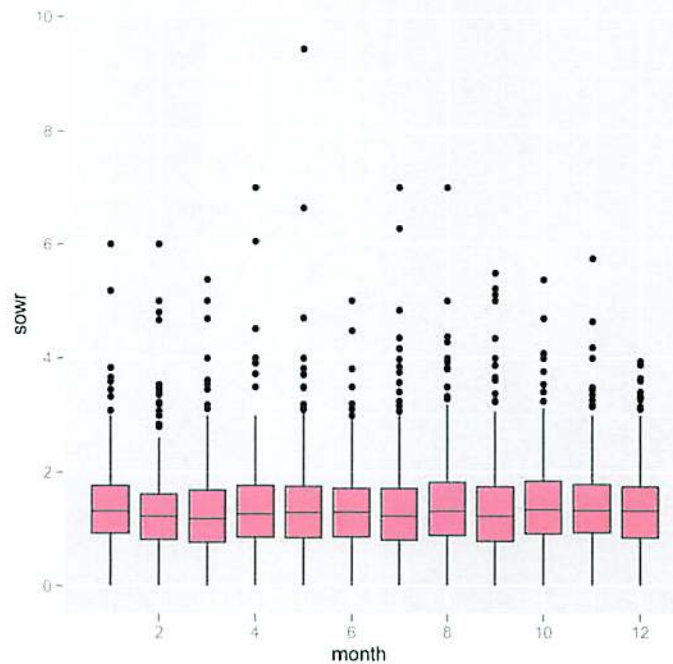


Figure 9: SOWR index for U.S. born players by birth month. Notice that, again, medians of the boxplots show randomness; still no sign of high index for August.

7.2.2 Strikeout to Walk Ratio (SOWR)

Since the measuring aspects of defensive abilities are much more complicated due to we should take into account every position of a team when in defense turn, we pick up a very important position (maybe the most important person in defense), as the position of pitching. ✓

Strikeout-to-walk ratio is a measure of a pitcher's ability to control pitches; calculated as strikeouts divided by bases on balls.

$$SOWR = \frac{SO}{BB}$$

A pitcher that possesses a great SO/BB ratio is usually a dominant power pitcher, such as Randy Johnson, Pedro Martnez, Curt Schilling, or Ben Sheets.

Therefore, we use strikeout-to-walk ratio of a player to observe the average pitching performance in different months.

To do the same work as above, we perform eight permutations to compare the original picture with its permutations. All these show the pitching performance of players also does not correlate with month change.

8 Conclusion

To sum up our analysis, both seasonal trend of birth in U.S. and Little League age cutoff date contribute to the sharp jump of number of baseball players born in July to that of players born in August. According to our comparison of August born with other players on number of players who enter Hall of Fame, individual RC and SOWR indices, August born players show no sign of superior to others. This is to say, the current

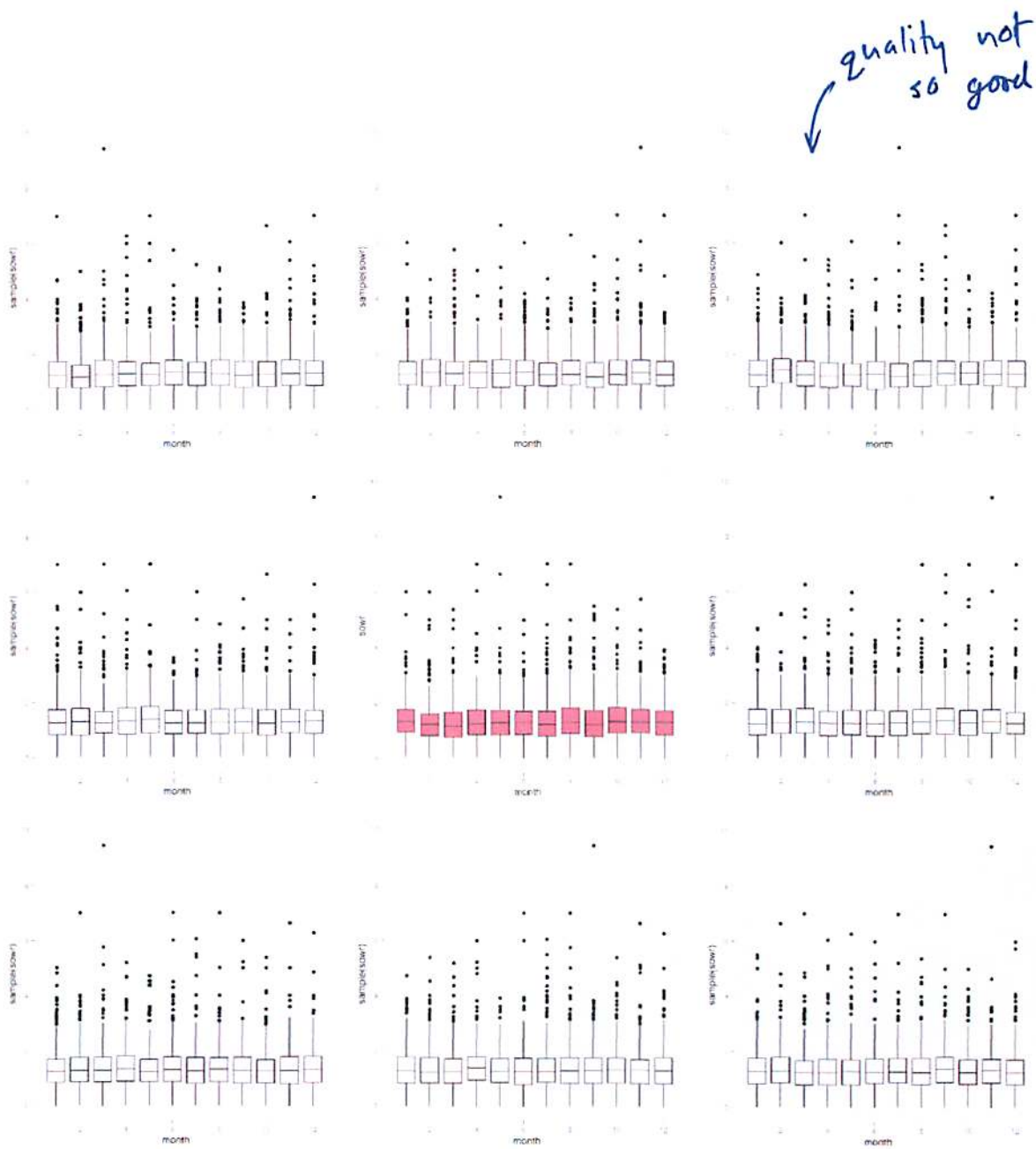



Figure 10: Permutations of SOWR. The graphic in the middle is the original SOWR boxplot. Notice that the permutations show no significant difference with respect to the original one, which means it is highly possible that August born shows no advantage in pitching skills.

Little League selection system favors mediocre players born right after the age cutoff date, which impedes the efficiency of cultivating baseball players.

9 Suggestion

In order to avoid the long-lasting effects from one fixed cutoff date, we suggest that the Little League use for cutoff ages for one year. These four can averagely distribute over a year. For example, we can make them, March 1, June 1, September 1, and December 1. Under this condition, the children born in different month basically have a fair chance to be selected into the league. They would not waste their time to wait for a cutoff date once a year passing the age limitation. Multi-time selection improves the selection system. It minimizes the gap of physical status between different birth months, which allows eligible players to enter into the league on a more fair begin. At the same time, we hope that the new selection system bring efficiency into the league by providing more opportunity to children born just before mid-year.



References

- [1] Ursula M. Cowgill. Season of birth in man. contemporary situation with special reference to europe and the southern hemisphere. *Ecology*, 47(4):614–623, 1966.
- [2] S. C. Curtin and M. M. Park. Trends in the attendant, place, and timing of births, and in the use of obstetric interventions: United states, 1989-97. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System.*, 47(27):1–12, 1999.
- [3] Michael R. Huber Gabriel B. Costa and John T. Saccoman. *Understanding Sabermetrics*. McFarland Company, 2008.
- [4] <http://en.wikipedia.org>. Major League Baseball All-Star Game.
- [5] <http://en.wikipedia.org>. National Baseball Hall of Fame and Museum.
- [6] James Lincoln Ray, 2007. What Was the Dead Ball Era?
- [7] Daniel A. Seiver. Trend and variation in the seasonality of u.s. fertility. *Demography*, 22(1):89–100, 1985.
- [8] www.littleleague.org, 2005. League Age Determination Date Talking Points and QA.

A Code

```
library(ggplot2)

# Load data to R.
f <- read.csv("Fielding.csv")
b <- read.csv("batting.csv")
p <- read.csv("Pitching.csv")
s <- read.csv("Salaries.csv")
pl <- read.csv("players.csv")

# Convert dates to numbers.
parse_date <- function(p) as.Date(strptime(p, "%m/%d/%Y"))
date_vars <- c("birth", "debut", "final", "death")
pl[date_vars] <- lapply(pl[date_vars], parse_date)

# Extract month and year from birthday.
pl$mon <- as.POSIXlt(pl$birth)$mon + 1
pl$year <- 1900 + as.POSIXlt(pl$birth)$year

# Calculate US-born players' monthly birth count for every 20 years.
# Subset data with US-born players.
plusa <- subset(pl, country = "USA")

# Count birth for every month from 1900 to 1920.
pl0020 <- subset(plusa, byear >= 1900 & byear < 1921)
count <- rep(0, 12)
for (i in 1 : 12) {
  for (j in 1 : length(pl0020$mon)){
    if (pl0020$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1900 -1920.
pl0020mon <- data.frame(1:12, count)
names(pl0020mon) <- c("month", "count")
pl0020mon$class <- "BB 1900-1920"
pl0020mon <- transform(pl0020mon, percentage = count / sum(count))

# Count birth for every month from 1920 to 1940.
pl2040 <- subset(plusa, byear >= 1920 & byear < 1941)
count <- rep(0, 12)
for (i in 1 : 12) {
```

← hopefully you could now
do this more easily with plyr
< dplyr

indenting

```

    for ( j in 1 : length(pl2040$mon)){
      if (pl2040$mon[j] == i) {
        count[i] <- count[i] + 1
      }
    }
  }

# Create a new data frame containing percentage of birth for month from 1920 -1940.
pl2040mon <- data.frame(1:12, count)
names(pl2040mon) <- c("month", "count")
pl2040mon$class <- "BB 1920-1940"
pl2040mon <- transform(pl2040mon, percentage = count / sum(count))

# Count birth for every month from 1940 to 1960.
pl4060 <- subset(plusa, byear >= 1940 & byear < 1961)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(pl4060$mon)){
    if (pl4060$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1940 -1960.
pl4060mon <- data.frame(1:12, count)
names(pl4060mon) <- c("month", "count")
pl4060mon$class <- "BB 1940-1960"
pl4060mon <- transform(pl4060mon, percentage = count / sum(count))

# Count birth for every month from 1960 to 1987.
pl6080 <- subset(plusa, byear >= 1960 & byear < 1987)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(pl6080$mon)){
    if (pl6080$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1960 -1987.
pl6080mon <- data.frame(1:12, count)
names(pl6080mon) <- c("month", "count")
pl6080mon$class <- "BB 1960-1987"
pl6080mon <- transform(pl6080mon, percentage = count / sum(count))

# Upload the US 1989-1997 population dataset.

```



```

usp8997 <- read.csv("US1989-1997.csv")

# Create a dataset containing US-born players' birth count for every twenty years and US population bir
usp <- rbind(pl0020mon, pl2040mon)
usp <- rbind(usp, pl4060mon)
usp <- rbind(usp, pl6080mon)
usp <- rbind(usp8997, usp)

# Calculate non-us-born monthly birth count for every 20 years.
# Subset data with non-USA born players.
plnousa <- subset(plusa, country != "USA" )

# Count birth for every month from 1900 to 1920.
plnous0020 <- subset(plnousa, byear >= 1900 & byear < 1921)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(plnous0020$mon)){
    if (plnous0020$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1900 -1920.
plnous0020mon <- data.frame(1:12, count)
names(plnous0020mon) <- c("month", "count")
plnous0020mon$class <- "BB 1900-1920"
plnous0020mon <- transform(plnous0020mon, percentage = count / sum(count))

# Count birth for every month from 1920 to 1940.
plnous2040 <- subset(plnousa, byear >= 1920 & byear < 1941)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(plnous2040$mon)){
    if (plnous2040$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1920 -1940.
plnous2040mon <- data.frame(1:12, count)
names(plnous2040mon) <- c("month", "count")
plnous2040mon$class <- "BB 1920-1940"
plnous2040mon <- transform(plnous2040mon, percentage = count / sum(count))

# Count birth for every month from 1940 to 1960.

```

line wrapping

```

plnous4060 <- subset(plnousa, byear >= 1940 & byear < 1961)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(plnous4060$mon)){
    if (plnous4060$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1940 -1960.
plnous4060mon <- data.frame(1:12, count)
names(plnous4060mon) <- c("month", "count")
plnous4060mon$class <- "BB 1940-1960"
plnous4060mon<- transform(plnous4060mon, percentage = count / sum(count))

# Count birth for every month from 1960 to 1987.
plnous6080 <- subset(plnousa, byear >= 1960 & byear < 1987)
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(plnous6080$mon)){
    if (plnous6080$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Create a new data frame containing percentage of birth for month from 1960 -1987.
plnous6080mon <- data.frame(1:12, count)
names(plnous6080mon) <- c("month", "count")
plnous6080mon$class <- "BB 1960-1987"
plnous6080mon <- transform(plnous6080mon, percentage = count / sum(count))

# Create a data frame containing non-US-born players' birth count for every twenty years and US populat.
usp8997$class <- "US Birth"
nosp <- rbind(plnous0020mon, plnous2040mon)
nosp <- rbind(nosp, plnous4060mon)
nosp <- rbind(nosp, plnous6080mon)
nosp <- rbind(usp8997, nosp)

# Create a plot to compare US population monthly birth seasonality with
# US-born players' monthly seasonality for every 20 years.
ggplot()+
geom_line(data = usp, aes(month, percentage, colour = class), size = 1) +
geom_point(data = usp, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

```

```

ggsave("uspop.pdf", height = 6, width = 6)

# Create a plot to compare US population monthly birth seasonality with
# non-US-born players' monthly seasonality for every 20 years.
ggplot()+
geom_line(data = nousp, aes(month, percentage, colour = class), size = 1) +
geom_point(data = nousp, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

ggsave("nouspop.pdf", height = 6, width = 6)

# Count monthly birth count for all US-born players.
uspl <- plusa
countus <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(uspl$mon)){
    if (uspl$mon[j] == i) {
      countus[i] <- countus[i] + 1
    }
  }
}

# Count monthly birth count for all non-US-born players.
nouspl <- plnosa
countnous <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(nouspl$mon)){
    if (nouspl$mon[j] == i) {
      countnous[i] <- countnous[i] + 1
    }
  }
}

# Create a data frame with monthly birth counts and percentage of all non-US-born players.
uspls <- data.frame(1 : 12, countus)
uspls$class <- "US Player"
names(uspls) <- c("month", "count", "class")
uspls <- transform(uspls, percentage = count / sum(count))

# Create a data frame with monthly birth counts and percentage of all non-US-born players.
nouspls <- data.frame(1 : 12, countnous)
nouspls$class <- "NON-US Player"
names(nouspls) <- c("month", "count", "class")
nouspls <- transform(nouspls, percentage = count / sum(count))

# Combine data with both US-born and non-US-born players.
nousanduspl <- rbind(uspls, nouspls)

```

```

# Compare monthly birth amount change for US and Non-US players.
ggplot()+
geom_line(data = nousanduspl, aes(month, percentage, colour = class, linetype = class), size = 1) +
geom_point(data = nousanduspl, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

ggsave("USvsNonUS.pdf", height = 6, width = 6)

# Combine US population birth count dataset with players born from 1900-1920.
usand20s <- rbind(usp8997, pl0020mon)

# Compare monthly birth amount change of US population with US baseball players from 1900 - 1920.
ggplot()+
geom_line(data = usand20s, aes(month, percentage, colour = class, linetype = class), size = 1) +
geom_point(data = usand20s, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

ggsave("usand20s.pdf", height = 6, width = 6)

# Create a data frame with US population and all US players from 1930 - 1987.
usand30to80 <- rbind(usp8997, pl2040mon)
usand30to80 <- rbind(usand30to80, pl4060mon)
usand30to80 <- rbind(usand30to80, pl6080mon)

# Compare monthly birth amount change of US population with US baseball players from 1930 - 1987.
ggplot()+
geom_line(data = usand30to80, aes(month, percentage, colour = class, linetype = class), size = 1) +
geom_point(data = usand30to80, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

ggsave("usand30to80.pdf", height = 6, width = 6)

# Count the birth for all the US-born players.
pl0080 <- plusa
count <- rep(0, 12)
for (i in 1 : 12) {
  for ( j in 1 : length(pl0080$mon)){
    if (pl0080$mon[j] == i) {
      count[i] <- count[i] + 1
    }
  }
}

# Combine the US population data with all US-born players monthly birth count data.
pl0080mon <- data.frame(1:12, count)
names(pl0080mon) <- c("month", "count")
pl0080mon$class <- "BB 1900-1980"
pl0080mon <- transform(pl0080mon, percentage = count / sum(count))

```

```

us0080 <- rbind(usp8997, pl0080mon)

# Compare US population birth seasonality vs. US-born baseball player birth seasonality.
ggplot()+
geom_line(data = us0080, aes(month, percentage, colour = class, linetype = class), size = 1) +
geom_point(data = us0080, aes(month, percentage, colour = class), size = 3) +
xlab("month") + ylab("percentage")

ggsave("usand00to80.pdf", height = 6, width = 6)

# Load data to R before analyzing technical variables.
batting <- read.csv("batting.csv")
pitching <- read.csv("Pitching.csv")
players <- read.csv("players.csv")

# Create a new variable abba and set abbb that are equal to zero to NA.
batting$abbb <- with(batting, ab + bb)
batting$abbb[batting$abbb == 0] <- NA

# Create a new variable rc(Runs Created)
batting$tb <- with(batting, (h - X2b - X3b - hr) + 2 * X2b + 3 * X3b + 4 * hr)
batting$rc <- with(batting, ((h + bb) * (tb)) / abbb)

# Consolidate rc by id to get mean of rc of each player for every year.
my.mean.rc <- function(df) {
  columns <- df["rc"]
  total <- apply(columns, 2, FUN = "mean", na.rm = T)
}

rc.mean.year <- ddply(batting, .(id), "my.mean.rc")

# Merge the consolidated data set rc.mean.year with player to get information of birthday.
names(rc.mean.year) <- c("id", "rc")
players_rc <- merge(rc.mean.year, players, by = "id")

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

# Extract year, month information from birth.
players_rc$year <- 1900 + as.POSIXlt(players_rc$birth)$year
players_rc$month <- as.POSIXlt(players_rc$birth)$mon + 1

# Explore the relationship of born month and rc.
qplot(month, rc, data = players_rc, geom = "boxplot", group = month, ylim=c(0,10),
colour = "red") + opts(legend.position = "none")

```

I'm not completely sure what you're doing here.

```

ggsave("rc~month.pdf", height = 6, width = 6)

# Permutation for rc to see if the pattern of the plot is random or not.
perm <- function() {
  sample <- qplot(month, sample(rc), data = players_rc, geom = "boxplot",
    group = month, ylim=c(0,10))
}
perm()
ggsave("sample1.png")
perm()
ggsave("sample2.png")
perm()
ggsave("sample3.png")
perm()
ggsave("sample4.png")
perm()
ggsave("sample5.png")
perm()
ggsave("sample6.png")
perm()
ggsave("sample7.png")

#####

# Set bb that are equal to zero to NA in pitching dataset.
pitching$bb[pitching$bb == 0] <- NA

# Create a new variable sowr(Strikeout Walk Ratio)
pitching$sowr <- with(pitching, so / bb)

# Consolidate sowr by id to get mean of sowr of each player for every year.
my.mean.sowr <- function(df) {
  columns <- df["sowr"]
  total <- apply(columns, 2, FUN = "mean", na.rm = T)
}

sowr.mean.year <- ddply(pitching, .(id), "my.mean.sowr")

# Merge the consolidated data set sowr.mean.year with player to get information of birthday.
names(sowr.mean.year) <- c("id", "sowr")
players_sowr <- merge(sowr.mean.year, players, by = "id")

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

```

good use of a function

✓

```

# Extract year, month information from birth.
players_sowr$year <- 1900 + as.POSIXlt(players_sowr$birth)$year
players_sowr$month <- as.POSIXlt(players_sowr$birth)$mon + 1

# Explore the relationship of sowr with born month of the players.
qplot(month, sowr, data = players_sowr, geom = "boxplot", group = month, ylim=c(0,10),
colour = "red") + opts(legend.position = "none")

ggsave("sowr~month.pdf", height = 6, width = 6)

# Permutation of sowr to check if the pattern of the plot is random or not.
perm2 <- function() {
  qplot(month, sample(sowr), data = players_sowr, geom = "boxplot",
    group = month, ylim=c(0,10))
}

perm2()
ggsave("sowrperm1.png")
perm2()
ggsave("sowrperm2.png")
perm2()
ggsave("sowrperm3.png")
perm2()
ggsave("sowrperm4.png")
perm2()
ggsave("sowrperm5.png")
perm2()
ggsave("sowrperm6.png")
perm2()
ggsave("sowrperm7.png")

```

	Outstanding (A+)	Good (A)	Acceptable (B)	Needs work (C)	Inadequate (F)
Introduction	10	8	6	4	2
	Clearly and concisely describes the data, and why it is of interest. Sets up a clear roadmap for the rest of the paper.	Good introduction to data, but roadmap for rest of paper lacking.	Introduction and roadmap unclear and missing important details.	Rote description of data. No context provided for data or questions.	Fails to introduce data and questions of interest.
Questions and findings (see homework rubric)					
Curiosity	20	16	12	8	4
Scepticism	20	16	12	8	4
Organisation	20	16	12	8	4
Conclusion	10	8	6	4	2
	Conclusions follows logically from results and findings. Includes interesting further questions and ideas for future research.	Good summary, but doesn't pull pieces together into cohesive whole. Interesting ideas for future research	Summary patchy, but some attempt at synthesis and development of ideas for future work.	Repeats findings with no synthesis. No proposals for future work.	Fails to summarise findings or ask more questions.
Presentation					
Text	5	4	3	2	1
	English is polished, concise and clear. No grammar or spelling mistakes.	Clear and concise, but not elegant. A few spelling and grammatical errors.	Readable, but excessively verbose, or lacking in detail. A number of errors in text.	Marginally readable. Many errors.	Barely readable. Many spelling and grammar errors. No evidence of proof reading.
Graphs	5	4	3	2	1
	Graphs carefully tuned for desired purpose. Evidence that many graphs were created before choosing one for presentation. Each graph illustrates one point.	Graphs well chosen, but a few have minor problems: inappropriate aspect ratios, poor labels, poor quality when printed.	Most graphs appropriate. Many graphs have minor problems.	Graphs poorly chosen to support questions. Some redundant or fundamentally flawed.	Graphs do not support questions and findings. Major presentation problems.
Tables	5	4	3	2	1
	All tables carefully constructed to make it easy to perform important comparisons. Careful styling highlights important features.	Tables generally well constructed, but some have minor flaws: too many d.p, tables too large.	Most tables appropriate. Many tables have minor problems.	Tables badly arranged to support comparisons of interest. Too many, or inconsistent, decimal places.	Tables do no support questions and findings. Major display problems.
Code	25	20	15	10	5
See code rubric					

Comments

Fantastic project - good use of external data sets and I think you've ~~found~~ come up with some ~~good~~ plausible explanations for some of the birth patterns.

101/115