Statistics 405 Project 1:
An Analysis of Baseball Salaries

October 2, 2009

1 Introduction

In 1930 when Babe Ruth was asked by a reporter if he really deserved to be making more money than the president, he famously responded, "I had a better year than he did". In the years since, baseball salaries have become no less contentious. The 1994 strike over a proposed salary cap resulted in more than 900 games being cancelled, lasting until U.S. district court Judge Sonia Sotomayor issued an injunction against the owners. Since that ruling, salaries have continued to skyrocket, particularly for star players. In fact, Alex Rodriguez earned more last year than the entire roster of the Florida Marlins. When asked for his opinion on this, he said, "The Marlins? It's amazing. And they still seem to find a way to be very competitive."[1]

great intro!

So we began to wonder: What does the distribution of baseball salaries look like? How has this distribution changed over time? And are teams like the Florida Marlins with lower payrolls actually competitive i.e. do they stand a chance of winning the World Series?

2 Data cleaning

Since there are only six salary observations for the years before 1985, we chose to focus on the years from 1985 to 2008. Even within these years, there are still a significant number of players for whom no salary information was recorded (see Figure 1). One reason for this may be that each team has not only an "active roster" of 25 men who play throughout the season, but also an "expanded roster" with 15 additional players who are either on the disabled list, or get called up from the minors to play in a few games at the end of the season. These extra players are not included in the teams' official payrolls, but their hitting and pitching stats are recorded in the data set. Comparing the boxplots for the number of games in which players with no salary data batted and the number of games in which players with salary data appeared shows that players with no salary data did generally appear in fewer games (Figure 2).

not a huge difference though

Once we cut the years before 1985, we further scrubbed the data by removing the entries with no salaries recorded or with irregular salary values. There were several salaries entered as \$0 and one entered as \$10,900 which were implausible as yearly salaries within the chosen time frame. The salary of \$10,900 for Dave Silvestri in 1993 is most likely a data entry error since the next two least paid players on the 1993 New York Yankees made \$109,000 each, so it seems possible that the last digit was accidentally dropped.

✓

We assume by neccessity that the missing data is random and would not substantially change the shape of the distribution, but the lack of salary data is a limitation on the strength of the conclusions in our subsequent analysis

good point

[1] http://sports.espn.go.com/mlb/news/story?id=3324199

proportions would probably be better here

Figure 1: Since salary data from before 1985 is mostly unavailable, we narrowed our focus to the years from 1985 on. Even with this restriction, there are still a significant number of players with no salary provided in the data set.

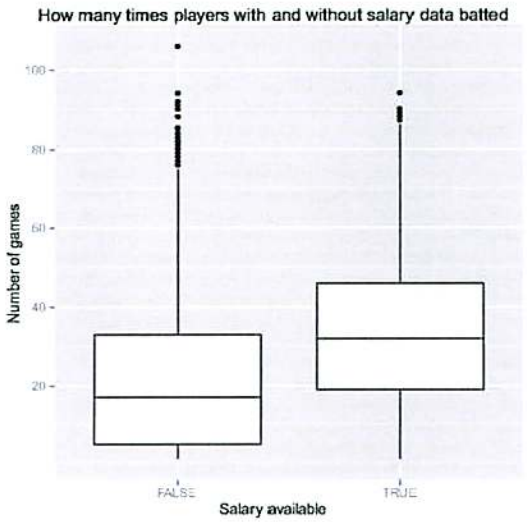


Figure 2: Note that the median and quartiles of the number of games played by players with salary data are higher than the median and quartiles of the number of games played by players with no salary data, but there is significant overlap.

3 Salary Distribution

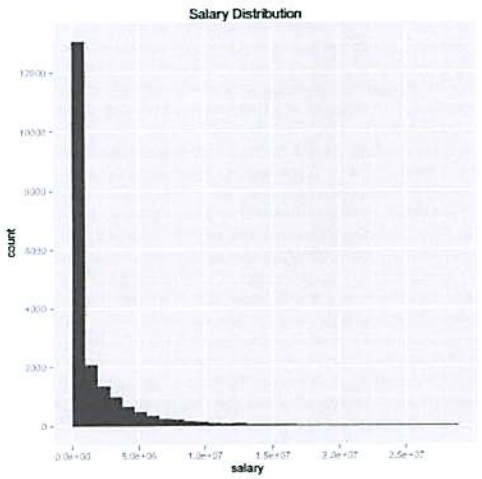
Taking an initial look at the cleaned data, we see that salaries range from a minimum of $50,000 to a maximum of 28 million dollars (Figure 3). The right skew is evident since the median is much less than the mean, which is very close to the seventy-fifth percentile. Taking the log helps get rid of some of this skewness

and makes the distribution easier to deal with.

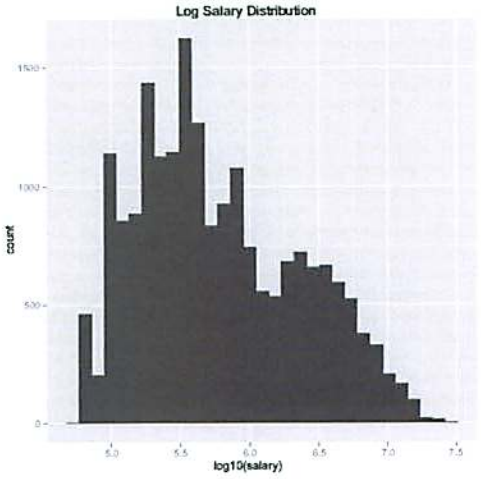
	salary	$\log_{10}$(salary)	career year
1	Min. : 50000	Min. :4.699	Min. : 1.000
2	1st Qu.: 205000	1st Qu.:5.312	1st Qu.: 3.000
3	Median : 450000	Median :5.653	Median : 6.000
4	Mean : 1541033	Mean :5.781	Mean : 6.832
5	3rd Qu.: 1730000	3rd Qu.:6.238	3rd Qu.:10.000
6	Max. :28000000	Max. :7.447	Max. :26.000

Figure 3: Five number summary and mean for the baseball player salaries and length of career

A plot of the cleaned salary data (Figure 4(a)) shows as expected that it is heavily skewed to the right, indicating that few players earn the highest salaries. To make the distribution easier to visualize, we take the log of the salaries (Figure 4(b)). The surprising bimodal pattern of this plot will be discussed further in the next section.



(a) Salary distribution

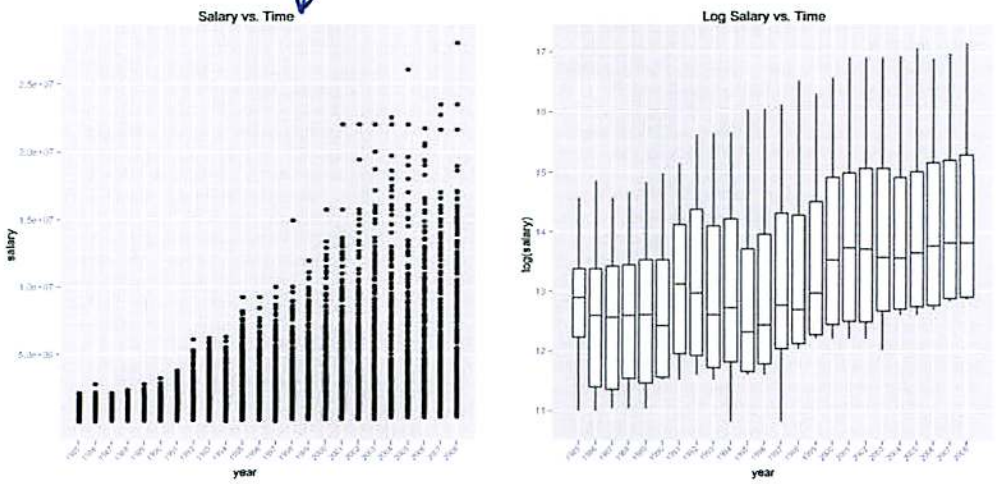


(b) Log salary distribution

To see if the shape of the salary distribution might reflect changes over time, we plotted both the raw salary (Figure 4(c)) and the log of salary (Figure 4(d)) over the selected time interval. Clearly both the median and the range of the salaries increases with time, but the fairly steady upward trend in the quantiles for log salary make it unlikely that the bimodal shape of the log salary plot is due only to salary change over time.

Next, we looked at how the salaries of individual players changed over their career. Players typically start off their major league careers in the low end of the spectrum, leveling off in the middle of their careers until they reach the 20 year mark, where there is an uptick. Salary drops off again beyond this point, perhaps reflecting a decrease in ability with age, although the sparseness of the data makes conclusion difficult.

not necessary



(c) Salary over time

(d) Log salary over time

Figure 4: Salary distributions and salary over time. The fluctuations in median log salary in the early 1990's may reflect the conflicts over salary leading up to the 1994 baseball strike.

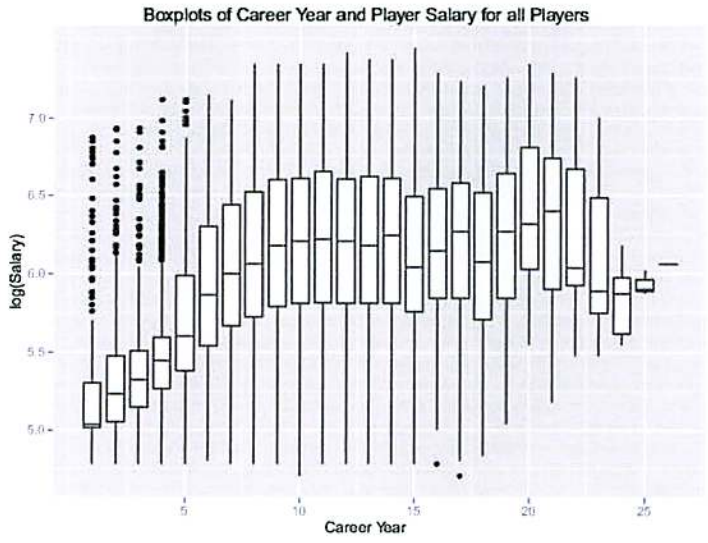


Figure 5: Log salary by career year. Players' salaries generally increase over the first few years and then level off until year 15.

4 Determining factors affecting the distribution of salary

Now we take a closer look at the unusual shape of the log(salary) distribution. With some smoothing, the distribution still shows a large number of players with higher salaries (Figure 6), represented by a peak

around 6.2. When we use less smoothing, the mode of the distribution splits into three peaks, but the second peak remains almost unchanged. To better explain this odd part of the distribution, we explore how it is affected by subsetting the data.

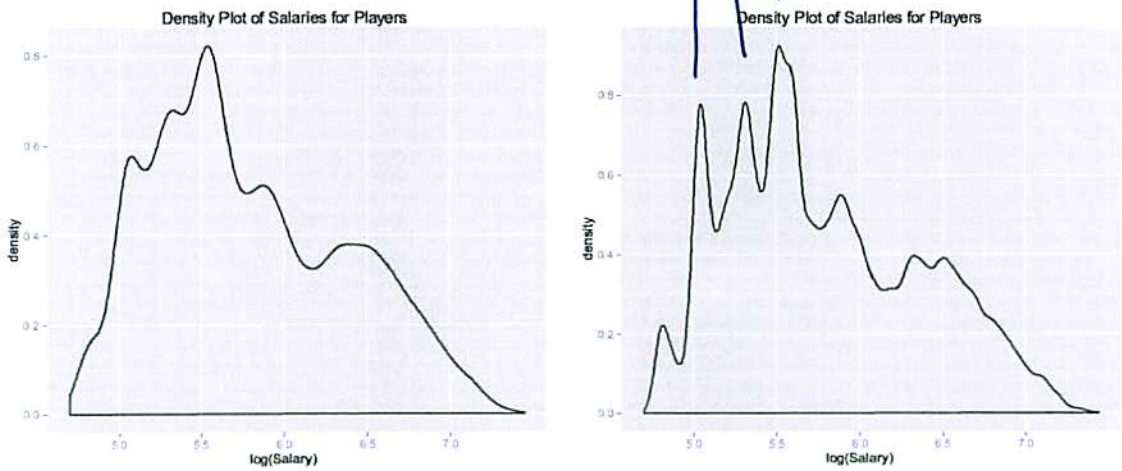


Figure 6: Distribution of salaries of baseball players with different levels of smoothing

Finding factors that lead to the presence or absence of higher end salaries could help explain why there are a larger number of players making 1.6 million dollars (about $10^{6.2}$) after a drop in the number that made a salary lower. In the distribution, as we use less smoothing, we can expect peaks to occur due to rounding and data recording. Ideally, we would hope salaries are reported to the dollar for the distribution to most accurately reflect player salaries, but it seems, for keeping records, such a task is tedious and unnecessary when we are dealing with salaries in the hundred thousands to millions. There may also be common player salaries that a lot of people share, indicated by skill level or importance. This could explain the large peaks with steep drops in between them. However, the wide peak toward the higher end is likely due to other causes, and will be investigated further.

4.1 The effects of position and winning an award

We begin our investigation of the reasons for the second peak's presence by subsetting by various factors to see how the distribution changes in relation to the changes in our sample. Figure 7 shows the distributions of salary for batters (blue) and pitchers (red). Instead of plotting the density, we looked at the counts to compare the two classes to each other without scaling. While it appears that there are fewer pitchers in baseball, the distributions are almost identical in shape. This is interesting because it suggests that the position does not determine a player's salary, as the same proportion of pitchers and field players seem earn each amount. Because the distributions are nearly identical to each other, one can assume little or almost no relationship between position and salary.

Another potential reason we expected to see a change in the distribution was for award winners. Because winning an award and making a high salary are both reserved for better athletes, the assumption that the two are related seems probable. It is not necessarily a causal relationship, but since performance affects money made for the team, it should affect the player's salary. Performance should also affect a player's probability of winning an award. Thus, we expected award winners to be high-salary players. However, while the distribution is skewed left (Figure 8), there are still a large number of low-salary players who win awards as well. Also, the number of award winners seems too small to affect the distribution, as seen by

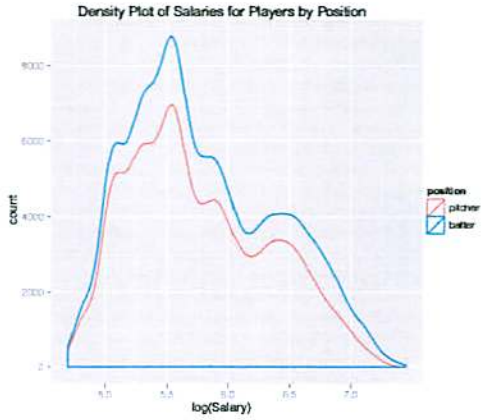


Figure 7: Distribution of salaries of baseball players based on position (pitching or batting)

the similarity of the players who do not win awards with the original distribution of baseball player salaries. This is different from the distributions based on position because there is a relationship illustrated by the change in distribution, but it does not remove the peak completely like we hoped for.

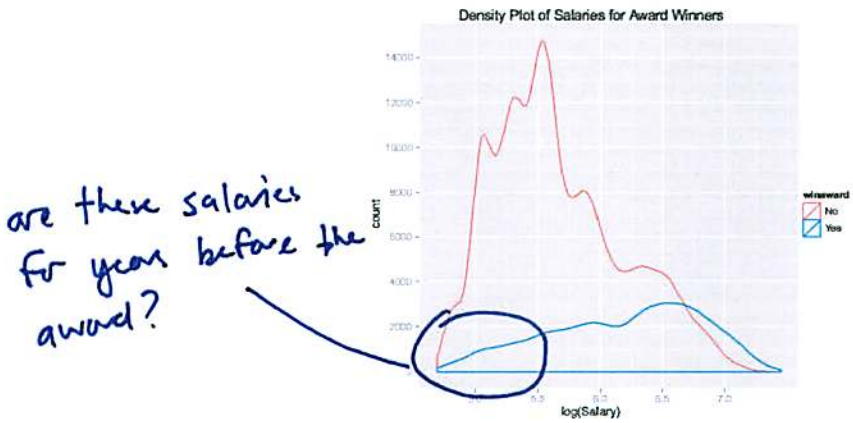


Figure 8: Distribution of salaries of baseball players based on whether the player won an award

4.2 Length of career in the league has an effect on salary

Like most jobs, we should expect to see salary increases with time spent in professional baseball. Assuming time also acts as a filter for worse players, the number of players in the league should decrease as the career length increases, and if career length affects salary, we should expect ~~less~~ fewer players to have higher salaries. Figure 10 shows that as the year increases for the time a person has played professional baseball, his salary increases. Because we are looking at the shape, we do not need to scale the density so that integrates to one, and instead we are able to see how the lengths of the careers compare in shape to each other. It is interesting to note that players in their early years make around the same salary, and they rarely make more

than a million dollars (10^6) each year until they have played a few years. These early years do not show that second peak in the distribution that we saw earlier, suggesting the career length of a player could be responsible for the shape of the distribution.

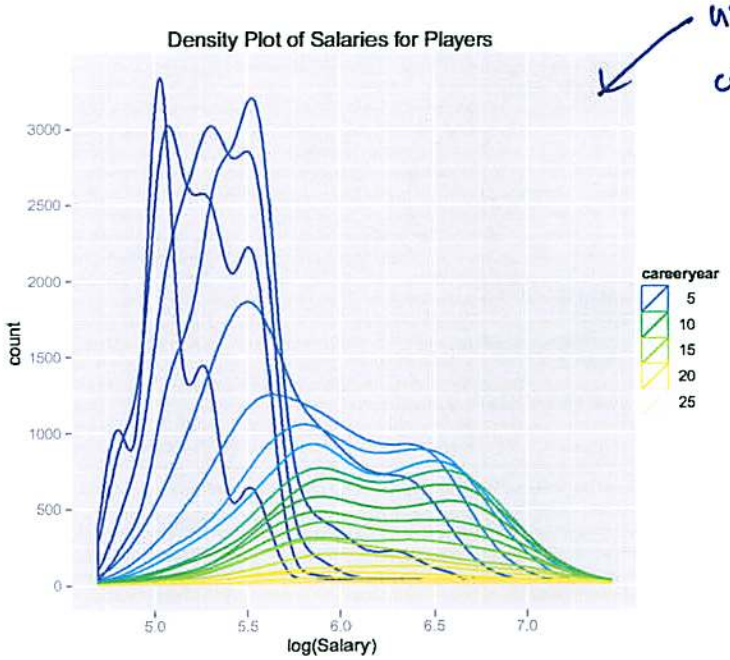
A player's salary tends to increase sharply within the first five years before leveling off as shown by the boxplots in Figure 5. This could help explain the abrupt change in shape of the distribution as players gain experience. However, the distribution in older players is much wider, suggesting there is more to determining a player's salary than the length of time. This seems obvious considering Alex Rodriguez signed a 10-year contract for 252 million dollars in his sixth year while plenty of players do not make that in their whole career (which seems true for the Florida Marlins). Skill and team success should play a key role in determining the player salary, but based on the distribution, the length of career helps explain the shape and the presence of the second peak.

To show the influence of career years on the original distribution, Figure 11 shows the densities of salaries for each career year stacked on each other. The plot shows that the first peaks are most influenced by the early years (years 1 to 4); midrange years and later years influence the part of the distribution after the largest peak. A potential problem with this plot is lack of data for the higher salaries. The distribution can easily be affected by a few salaries since there are not many high-paid players and there are not many players who go beyond 10-12 years. Using the log helps keep highly paid players from becoming outliers, but the little data could still influence the plot dramatically.

career year	1	2	3	4	5	6	7	8	9	10	11	12	13
players	1257	1971	2051	1987	1838	1671	1481	1330	1178	1054	896	743	618
career year	14	15	16	17	18	19	20	21	22	23	24	25	26
players	491	409	299	196	138	84	52	37	20	10	6	3	1

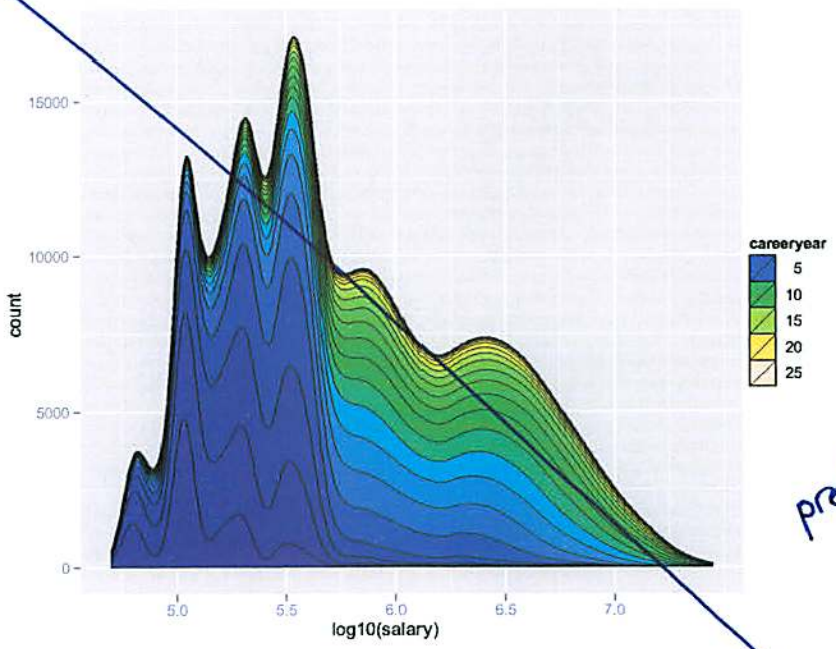
Figure 9: The table shows the drop off in players with reported salaries playing in the league as they get older

Figure 9 also helps show potential reasons for the greater variability in the distribution of older player salaries. There are less older people with reported salaries, so more variability is expected. The change in median over each year also suggests the median could be more influenced by the size of the samples than the earlier years where samples are large enough show the median increases or remains the same. However, the table and boxplots illustrate that there is enough support and large enough numbers in each career to suggest that it is probably the largest reason for the 2nd peak in the distribution. The 1st peak is mainly affected by the first few years, and we still have enough players who play beyond those years to show that the 2nd peak results from the number of older players, who also make more money on average. The boxplot supports this because the median and IQR level off around the same time the distribution changes in terms of career years.



use faceting? combine into 5 year groups?

Figure 10: Distribution of salaries of baseball players based on length of time spent playing in years



pretty, but not informative

Figure 11: Distribution of salaries for each career year stacked to fill the original shape of the distribution for all players

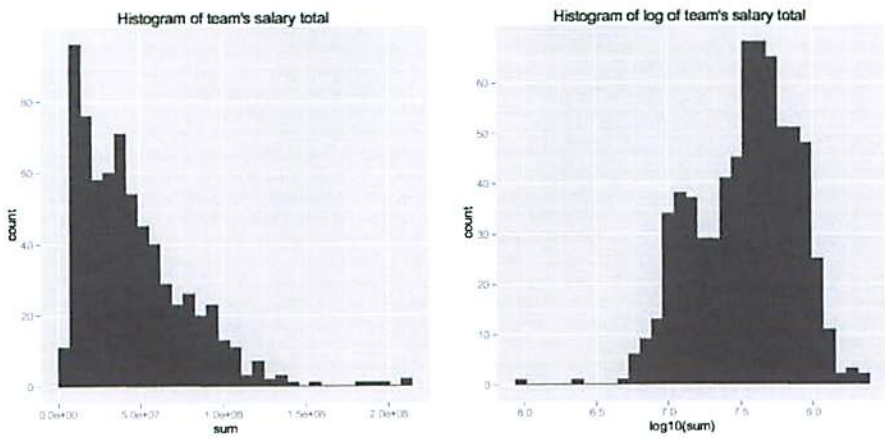


Figure 12: Team salaries and $\log_{10}$ of team salaries

5 Comparison of salaries for World Series winners and losers

Our final concern is the effect of salaries on championships. Can a team essentially buy a championship? Steinbrenner thinks so, but he hasn't had the best luck recently. Winning the World Series is a good way of measuring success of a team. We attempt to determine whether there is a positive relationship between salaries and winning the World Series.

5.1 Total salaries for the team

After incorporating the data for World Series winners and losers for each year, we sum the salaries by team to get aggregate team spending. Figure 12 shows histograms of team salaries and team salaries after a log transform. According to the figures, team salaries are skewed to the right, but the log transformation makes the distribution skewed to the left slightly. For the remainder of the study, we do not use the log transformation and instead use the original salary data. The total salary for each team varies over a wide range, and the number of team players differs from team to team and year to year. We do not have full data on every player either. It is then necessary to plot the average salary for each team against the number of players in each team.

5.2 Average salaries for the team

From Figure 13, we see that most of the teams have more than 25 players, and the average salaries decrease as the number of players increases in that range. It is worth mentioning that some of the teams have very few players and abnormally low average salaries. This data could simply be ignored and excluded because for some years, some teams didn't record the whole salary information in the dataset. For example, Minnesota (MIN) has only 9 players with recorded data in 1987, yet they won the World Series that year. Minnesota had 19 and 17 players other years, which is still lower than a full roster.

5.3 Difference between winners and losers

There are obviously more losers than winners since only one team can be considered the champion, so for comparing the two sets, using counts would be impractical. We avoid this by scaling the plots by density

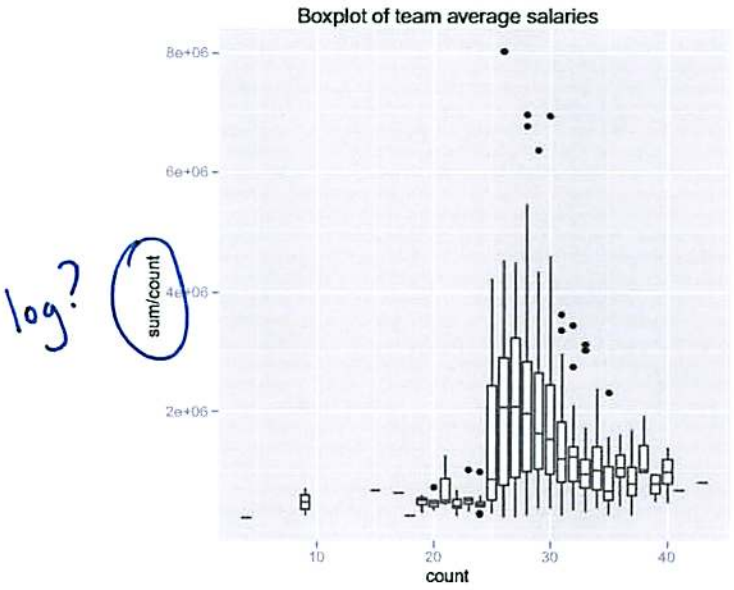


Figure 13: The boxplots show how the amount spent per player changes with the number of people on the team. We only consider the players per team as the ones recorded in the data.

instead, looking at the distribution shape and the probability relative to the individual classes.

We see in Figure 14 it is more probable for winning teams to pay more per player than losers, which means that winners should tend to pay more for each of their players. This may not always be correct, in the case of having star players that greatly increase the average for a particular team and acts as an outlier for the salary distribution of that team. This is quite common in all sports, so we should expect it is possible here. Using average salary does not take this fact into account, and we only worry about the teams overall spending per player without respect to the players.

We are then able to look at whether teams that have spent more tend to win championships. Figure 15 plots the average team salaries for each year, showing that between 1985 and 2008 all winning teams tend to have higher average total salaries than the losing teams for that year. We can assume this means teams tend to pay more to win championships. Although it is interesting to note that the winning team is not usually the team that spent the most money per player. Of course showing the winning team spends the most money would easily support that championships are won and lost based on the team's accountants. Fortunately for baseball, this system has its faults and underdogs can always win.

Looking at which teams payed the most each year (highest paying teams are labeled), we can look at trends in money spent for those teams. Interestingly, though the winning teams have higher than average team salaries, they are not always the highest of the year. This supports that winners pay more than others generally, but paying the most is not necessarily the way to win the championship.

The New York Yankees provide evidence for this statement. It is well known the Yankees have no problem spending money to get big names. They are one of the richest teams, and they are not afraid to show it. Unfortunately, they payed more money per player since 2002 without winning a championship. They managed to win their division or come close each year, but they were not able to win the World Series. It is interesting that they continued this trend of heavy spending without winning, but by coming close each year, there may be the idea that a little more spending would get them past that final obstacle: post-season. They

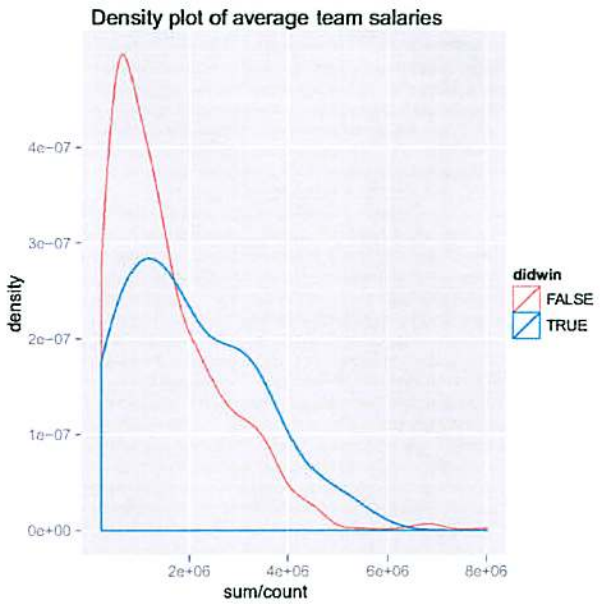


Figure 14: Distribution of average salary for championship winning teams and losing teams

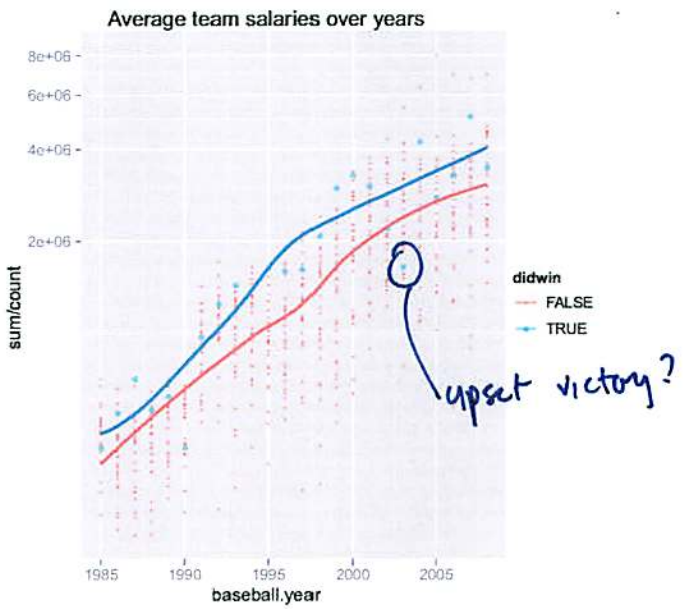


Figure 15: Most World Series winners tend to have higher average player salaries, but they are not usually the highest in the league.

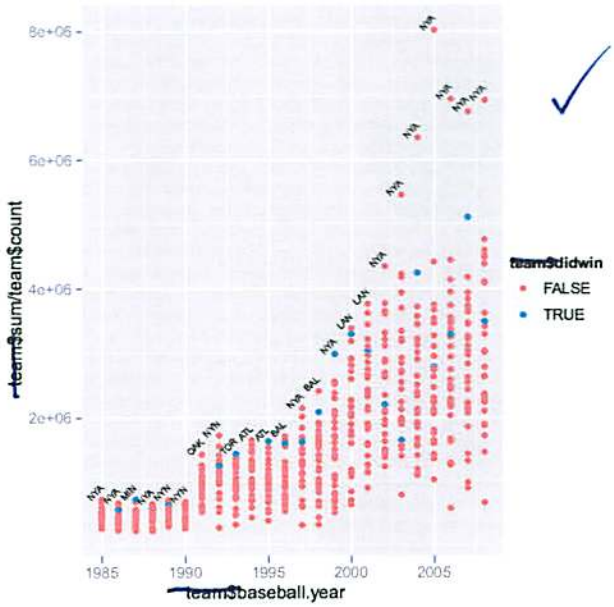


Figure 16: The teams that spend the most per player do not tend to win the World Series, and some will spend the most over multiple years without a championship.

could also potentially have the highest-payed players because they entered contracts with them, meaning players stay on for multiple years.

Another interesting year is the 2003 season, where the Florida Marlins won the championship, yet they were near the bottom of the league in average salary per player. Further inspection of the data showed Florida nearly had a full roster of players in the data and the rest of the league generally reported around twenty-five players per team as well, meaning the results should be accurate. Of course, there is always the chance that a non-reported player was a big name. Florida's biggest player at the time, Ivan Rodriguez, is present in the data. Thus, the Yankee data and 2003 Florida data both support that championships are not bought.

6 Conclusion

Our findings would benefit from having fewer missing salary values, though with wider data collection there could be some issues with consistency. For example, we might need to distinguish players that did not play for the entire season, since here we assume that the salaries reflect pay over the entire year. It would also be interesting to have data further back in time, though this would also complicate the analysis since with a wider time frame we might need to pay more attention to issues such as inflation.

Based on our analysis, we find that salary plays a large role in baseball, yet it does not amount to everything. Alex Rodriguez was not necessary for the Marlins to win in 2003, and he hasn't drastically helped the Yankees in post-season in recent years. Although the salaries of star players receive more attention in the media, a large number of players do not make a million dollars a year. Nonetheless, the upward trend indicates that baseball players will have no trouble making a living in the future, despite Yogi Berra's complaint that "A nickel ain't worth a dime anymore".

Code appendix

```
# Count function - used with ddply to count number of TRUEs for
# players per team and other instances
count <- function(id){
  sum(as.numeric(!is.na(id)))
}

# Setup and data cleaning -----------------------------------------------------
# Load the ggplot library
library(ggplot2)

# Read in the data sets and change the column names to be understandable
pitching <- read.table("Pitching.txt", sep = ",")
names(pitching) <- c("id", "year", "stint", "team", "league", "wins", "losses",
  "games", "starts", "complete_games", "shutouts", "saves", "ipouts",
  "hits_allowed", "earned_runs", "home_runs_allowed", "walks", "strikeouts",
  "baopp", "era", "int_walks", "wild_pitches", "hit_batsmen", "balks",
  "batters_faced", "games_finished", "r")

salary <- read.table("Salaries.txt", sep = ",")
names(salary) <- c("year", "team", "league", "id", "salary")

award <- read.table("AwardsPlayers.txt", sep = ",")
names(award) <- c("id", "award", "year", "league", "tie")

players <- read.csv("players.csv")
batting <- read.csv("batting.csv")
names(batting) <- c("id", "year", "stint", "team", "league", "games", "at_bats",
  "runs", "hits", "doubles", "triples", "home_runs", "runs_batted_in",
  "stolen_bases", "caught_stealing", "walks", "strikeouts",
  "int_walks", "hit_by_pitch", "sacrifice_hit",
  "sacrifice_fly", "grounded_into_double_plays")

# Create a merged table with pitching, batting, and salary data for each player,
# using all.x and all.y so we don't drop entries for players with missing data
baseball <- merge(batting, pitching, by = c("year", "id", "stint", "team",
  "league"), all.x = TRUE, all.y = TRUE)
baseball <- merge(baseball, salary, by = c("year", "id", "team", "league"),
  all.x = TRUE, all.y = TRUE)

# Which players do we have salary data for?
baseball$has_salary <- (!is.na(baseball$salary))
qplot(year, data = baseball, ..count.., geom = "freqpoly",
  xlab = "Year", binwidth = 1,
  colour = has_salary, main = "Availability of salary data by year")
```

```
ggsave("missing.pdf", width = 6, height = 6)

# Cut off all the years before 1985 when salary data was sparse or nonexistent
baseball <- subset(baseball, year >= 1985)

# What entries are still missing or weird?
qplot(year, data = baseball, ..count.., geom = "freqpoly",
  xlab = "Year", binwidth = 1,
  colour = has_salary, main = "Availability of salary data by year after 1985")
ggsave("missing_since_1985.pdf", width = 6, height = 6)

# Take a look at how many games players with and without salaries were in
qplot(!is.na(salary), games.y, data = baseball, geom = "boxplot",
  xlab = "Salary available", ylab = "Number of games",
  main = "How many times players with and without salary data batted")
ggsave("games_by_has_salary.pdf", width = 6, height = 6)

# Remove players with salary 0 and one with salary 10,900
with_salary <- subset(baseball, !is.na(salary) == FALSE & log(salary) > 10)

# General look at salary data ---------------------------------------------
# What does the raw distribution look like?
qplot(salary, data = with_salary) + opts(title = "Salary Distribution")
ggsave("salaryplot1.pdf", width = ..., height = ...)

# Right skew suggests looking at the logged values for salary
qplot(log10(salary), data = with_salary) +
  opts(title = "Log Salary Distribution")
ggsave("salaryplot2.pdf")

# Now see how salaries changed over time
qplot(as.factor(year), salary, data = with_salary) +
  opts(axis.text.x = theme_text(angle = 45, hjust = 1, size = 8,
    colour = "grey50")) +
  opts(title = "Salary vs. Time") +
  labs(x = "year")
ggsave("salaryplot3.pdf")

# Is there a clearer trend for log salaries over time?
qplot(as.factor(year), log(salary), data = with_salary, geom = "boxplot") +
  opts(axis.text.x = theme_text(angle = 45, hjust = 1, size = 8,
    colour = "grey50")) +
  opts(title = "Log Salary vs. Time") +
  labs(x = "year")
ggsave("salaryplot4.pdf")

# Now consider the effect of number of years played on salary
qplot(as.factor(careeryear), log10(salary), data = baseball,
```

```
  geom = "boxplot", ylim = c(4.5, 7.5)) +
  opts(axis.text.x = theme_text(angle = 45, hjust = 1, size = 8,
    colour = "grey50")) +
  opts(title = "Log Salary vs. Career Year") +
  labs(x = "career year")
ggsave("salaryplot5.pdf")

# Investigation of bimodal log salary distribution ----------------------------
# General look at bimodal log salary distribution
qplot(log10(salary), data = with_salary, geom = "density",
  xlab = "log(Salary)", ylab = "Density",
  main = "Density Plot of Salaries for Players")
ggsave("01salarydensity.pdf")

# Create a summary table of salary data and career year data
summarydata <- cbind(with_salary[45], log10(with_salary[45]),
  with_salary[46])
xtable(summary(summarydata))

# Unsmooth some to take a closer look at the distribution
qplot(log10(salary), data = with_salary, geom = "density",
  adjust = 1/2, xlab = "log(Salary)", ylab = "Density",
  main = "Density Plot of Salaries for Players")
ggsave("02lesssmooth.pdf")

# Comparing pitchers and field players/batters
position <- is.na(with_salary$earned_runs)
position <- as.factor(position)
levels(position) <- c("pitcher", "batter")
qplot(log10(salary), ..count.., data = with_salary, geom = "density",
  colour = position,
  xlab = "log(Salary)", ylab = "Density",
  main = "Density Plot of Salaries for Players by Position") +
  scale_colour(name = "position")
ggsave("03pitchorfield.pdf")

# Facet on whether the player ever won any awards
award <- read.csv("award.csv")
winaward <- as.factor(with_salary$id %in% award$id)
levels(winaward) <- c("No", "Yes")
qplot(log10(salary), ..count.., data = with_salary, geom = "density",
  colour = winaward,
  xlab = "log(Salary)",
  main = "Density Plot of Salaries for Award Winners")
ggsave("04awardwinner.pdf")

# Look at effect of career year
qplot(log10(salary), ..count.., data = with_salary,
```

```
  geom = "density",
  colour = careeryear, group = round(careeryear),
  xlab = "log(Salary)",
  main = "Density Plot of Salaries for Players",
  size = I(0.5), adjust = 1.5) +
  scale_color_gradientn(colours = topo.colors(15))
ggsave("05allcareeryears.pdf")

# Stacked density plot
qplot(log10(salary), ..count.., data = with_salary,
  geom = "density", group = round(careeryear),
  fill = careeryear, position = "stack",
  size = I(0.1), adjust = 1.5) +
  scale_fill_gradientn(colours = topo.colors(15))
ggsave("06stackeddensity.pdf")

qplot(log10(salary), ..density.., data = with_salary,
  fill = careeryear, group = round(careeryear))
qplot(log10(salary), ..count.., data = with_salary,
  geom = "density")

# Table of players for each career year
xtable(t(table(with_salary$careeryear)))

# Boxplot of career years
qplot(careeryear, log10(salary), data = with_salary,
  geom = "boxplot", group = round(careeryear),
  xlab = "Career Year", ylab = "log(Salary)",
  main = "Boxplots of Career Year and Player Salary for all Players")
ggsave(file = "07careerbox.pdf")

# Summary of career year data
summary(with_salary$careeryear)

# Early years density
rookies <- subset(with_salary, careeryear < 3)
qplot(log10(salary), ..density.., data = rookies,
  geom = "density", adjust = 2)
ggsave(file = "08earlycareer.pdf")

# Veterans density
vets <- subset(with_salary, careeryear >10)
qplot(log10(salary), ..density.., data = vets,
  geom = "density",
  adjust = 2)
ggsave(file = "09latecareer.pdf")

# Put Veterans and Rookies on same plot
```

```
qplot(geom = "blank") +
  geom_density(aes(log10(vets$salary), ..count..),
  colour = "blue", adjust = 2) +
  geom_density(aes(log10(rookies$salary), ..count..),
  colour = "red", adjust = 2)

# Salaries differences of world series winning teams ---------------------------
# Find the total payroll for each team by year
team <- ddply(baseball$salary, .(baseball$team, baseball$year), "sum")

# Count is number of players on the team that year with recorded data
numplayers <- ddply(baseball$id, .(baseball$team, baseball$year),
  "count")

team <- merge(team, numplayers, by = c("baseball.year",
  "baseball.team"), all = T)
team <- team[team$baseball.year >= 1985,]

# Create logical for if the team won the World Series in that year
worldseries <- read.csv("winners.csv")
names(worldseries) <- c("baseball.year", "winner")
team <- merge(team, worldseries, by = "baseball.year",
  all.x = T, all.y = F)

# Add logical to team table for if that team was a winner
team$didwin <- with(team, as.character(baseball.team) ==
  as.character(winner))

# Histogram of the salary sum
qplot(sum, data = team, main = "Histogram of team's salary total")
ggsave(file = "wssum.pdf")
qplot(log10(sum), data = team, main = "Histogram of log of team's salary total")
ggsave(file = "wslogsum.pdf")

# Investigate average salaries vs. number of team players
qplot(count, sum/count, data = team, geom = "boxplot",
  group = round(count),
  main = "Boxplot of team average salaries")
ggsave(file = "sumcount.pdf")

# Density plot
qplot(sum/count, ..density.., data = team, geom = "density",
  colour = didwin,
  main = "Density plot of average team salaries")
ggsave(file = "dsumcount.pdf")

# Another way to view it
qplot(baseball.year, sum/count, data = team, geom = "point",
```

```
  group = round(baseball.year), colour = didwin, size = didwin,
  alpha = I(1/2), main = "Average team salaries over years") +
  geom_smooth(aes(group = didwin), se = F, size = 1) +
  coord_trans(y = "log10")
ggsave(file = "avesalyear.pdf")

# Can also look at which team spent the most per player
most <- ddply(team, .(baseball.year), subset, sum /
  count == max(sum/count))

qplot(geom = "blank") + geom_point(aes(team$baseball.year,
  team$sum/team$count, group = round(team$baseball.year),
  colour = team$didwin)) +
  geom_text(aes(most$baseball.year, log10(most$sum/most$count)),
  label = most$baseball.team, size = 2, angle = 45, vjust = -1)
ggsave(file = "winningteam.pdf")
```

	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

102/115