INTRODUCTION

In the movie *Moneyball*, Oakland Athletics general manager Billy Beane attempts to build a baseball team with a budget of less than $40 million, less than 1/3 than the budget of the New York Yankees. In just one season (2002), he built a team nearly from the ground up that won 103 out of 162 games, winning their division by 4 games. Their season ended when they were defeated by the Minnesota Twins in the first round of the playoffs, but Beane changed the game forever by introducing robust statistical analysis to the game of baseball.

If someone wanted to do this today, how could a new team owner or GM achieve this? What is the minimum amount of money an MLB team needs to be successful? If a new team is going to be formed, how much money does the owner need to spend on players for the team to win at least half (81) their games?

While in reality some team owners may have goals other than winning championships (such as having flashy players or minimizing costs above all), our target client will be a brand-new team owner or general manager who wants to be successful on the field. His or her main goal will be to maximize the number of wins and chance of winning the World Series, while spending the least possible amount of money on player salaries.

A WORD ABOUT MAJOR LEAGUE BASEBALL

In baseball, there are two different kinds of players: pitchers and batters. Among batters, there are also other positions, such as infielders and outfielders. The infield positions are first base, second base, shortstop, and third base. The outfield positions are left field, center field, and right field. There is also a designated hitter (a non-fielding batter) in the American League. Each of these positions may have different salaries.

One factor that may contribute to differences in stats is that some teams are in the American League while others are in the National League. The American League, because they have a designated hitter and the pitchers do not hit, tend to have bigger hitters, who may have different salaries.

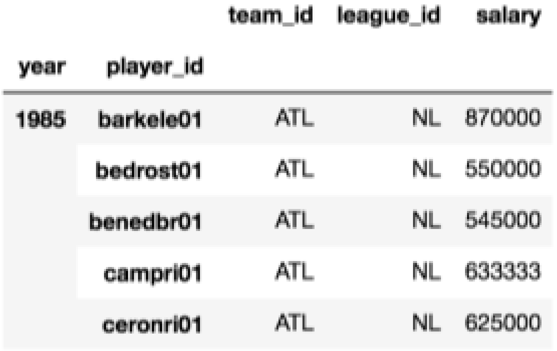
Another complicating factor for both batters and pitchers (though more of a problem for pitchers) is that some of the players are starters and some are not. For batters, most of the time only the starters actually play in the game. However, in a single game, as many as 4 or 5 pitchers may play, but only one appears as the starting pitcher.

DATA

For this project, I will use the Lahman Baseball Database. I acquired the data through Kaggle. The dataset includes a wide variety of data from every Major League Baseball (MLB) league, team, and player. Although much of the data is available from the 1871 to the 2015 seasons, the `salary` dataset only includes data on players from 1985 to 2015, so we will limit our analysis to these years.

Most of the datasets are organized by season and player. See below for how the `salary` dataset is organized. Players often play for multiple seasons in their career. The mean number of seasons played for all players 1985-2015 is about 5 seasons. Modern-day seasons include 162 games, running from March to September.

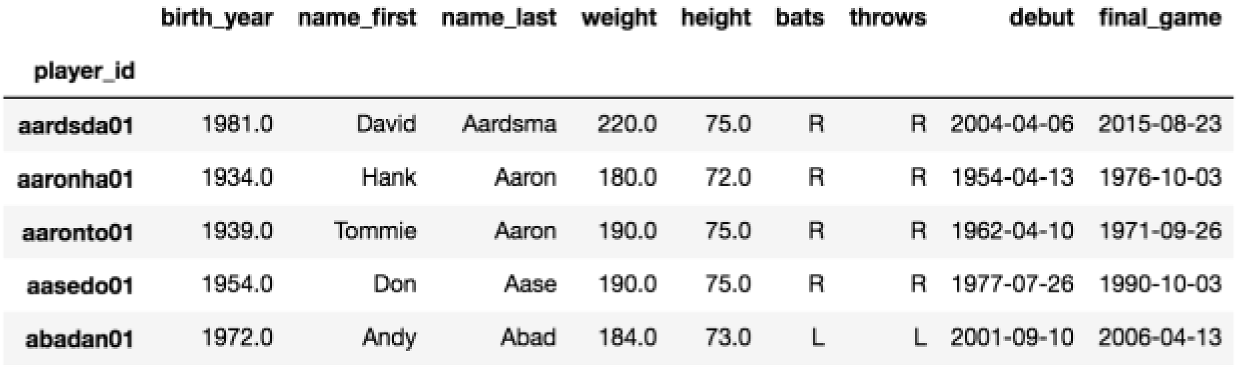
salary.csv



		team_id	league_id	salary
year	player_id			
1985	barkele01	ATL	NL	870000
	bedrost01	ATL	NL	550000
	benedbr01	ATL	NL	545000
	campri01	ATL	NL	633333
	ceronri01	ATL	NL	625000

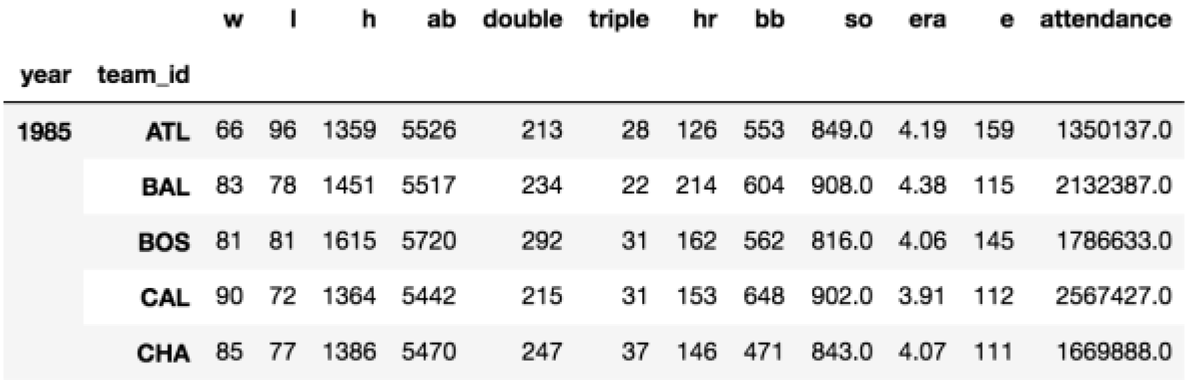
The `player` dataset is organized by player alone, as these data do not change from year to year. The `team` dataset, obviously, is organized by season and team, not player. See below for select variables from the first five rows of each of these datasets. For the `team` dataset, the variables are wins and losses, team total hits, at-bats, doubles, triples, homeruns, walks, strikeouts, earned run average, errors, and total home attendance for that year. For the `player` dataset, the birth_year and debut columns can be used to determine a player's age and their time in the MLB for a given year.

player.csv



	birth_year	name_first	name_last	weight	height	bats	throws	debut	final_game
player_id									
aardsda01	1981.0	David	Aardsma	220.0	75.0	R	R	2004-04-06	2015-08-23
aaronha01	1934.0	Hank	Aaron	180.0	72.0	R	R	1954-04-13	1976-10-03
aaronto01	1939.0	Tommie	Aaron	190.0	75.0	R	R	1962-04-10	1971-09-26
aasedo01	1954.0	Don	Aase	190.0	75.0	R	R	1977-07-26	1990-10-03
abadan01	1972.0	Andy	Abad	184.0	73.0	L	L	2001-09-10	2006-04-13

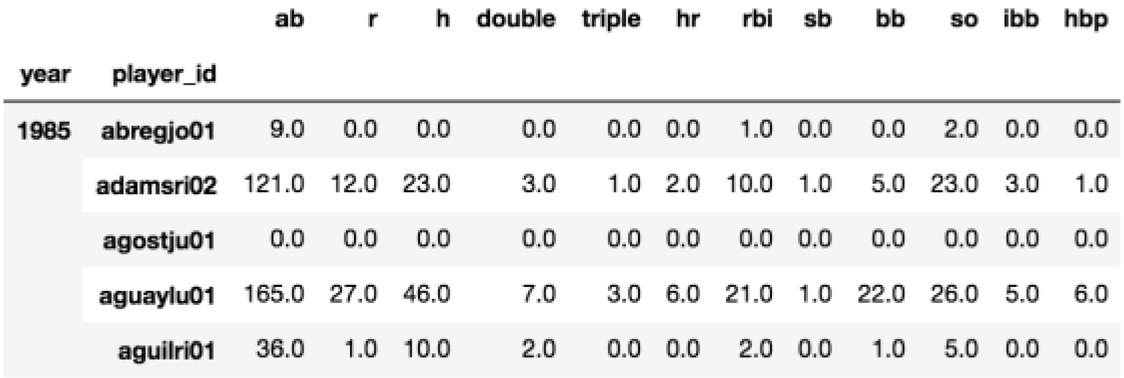
team.csv



		w	l	h	ab	double	triple	hr	bb	so	era	e	attendance
year	team_id												
1985	ATL	66	96	1359	5526	213	28	126	553	849.0	4.19	159	1350137.0
	BAL	83	78	1451	5517	234	22	214	604	908.0	4.38	115	2132387.0
	BOS	81	81	1615	5720	292	31	162	562	816.0	4.06	145	1786633.0
	CAL	90	72	1364	5442	215	31	153	648	902.0	3.91	112	2567427.0
	CHA	85	77	1386	5470	247	37	146	471	843.0	4.07	111	1669888.0

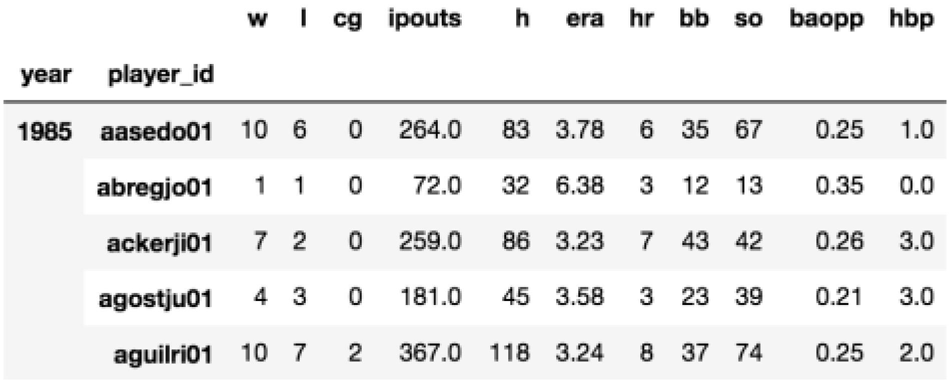
For player statistics, we will use the `batting` and `pitching` datasets. Below are the first five rows of these datasets, only including the variables we will use later. For batting, these variables are at-bats, runs, hits, doubles, triples, homeruns, runs-batted-in, stolen bases, walks, strikeouts, intentional walks, and hits-by-pitch. For pitching, the variables are wins, loses, complete games, number of outs pitched, hits, earned run average, homeruns against, walks, strikeouts, opponent batting average, and hits-by-pitch.

batting.csv



		ab	r	h	double	triple	hr	rbi	sb	bb	so	ibb	hbp
year	player_id												
1985	abregjo01	9.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0
	adamsri02	121.0	12.0	23.0	3.0	1.0	2.0	10.0	1.0	5.0	23.0	3.0	1.0
	agostju01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	aguaylu01	165.0	27.0	46.0	7.0	3.0	6.0	21.0	1.0	22.0	26.0	5.0	6.0
	aguilri01	36.0	1.0	10.0	2.0	0.0	0.0	2.0	0.0	1.0	5.0	0.0	0.0

pitching.csv

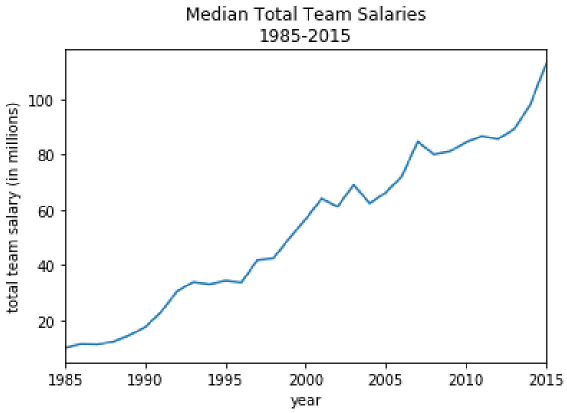


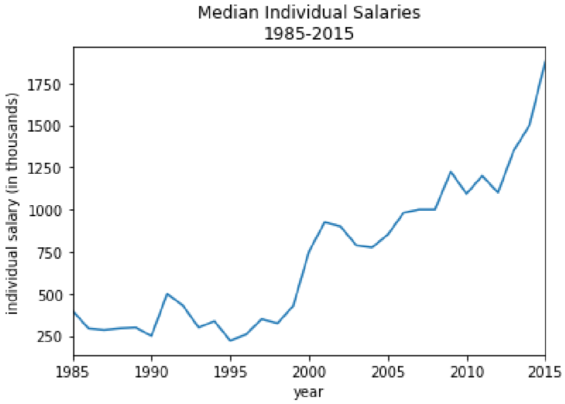
		w	l	cg	ipouts	h	era	hr	bb	so	baopp	hbp
year	player_id											
1985	aasedo01	10	6	0	264.0	83	3.78	6	35	67	0.25	1.0
	abregjo01	1	1	0	72.0	32	6.38	3	12	13	0.35	0.0
	ackerji01	7	2	0	259.0	86	3.23	7	43	42	0.26	3.0
	agostju01	4	3	0	181.0	45	3.58	3	23	39	0.21	3.0
	aguilri01	10	7	2	367.0	118	3.24	8	37	74	0.25	2.0

EXPLORATORY DATA ANALYSIS: TRENDS AND DISTRIBUTIONS

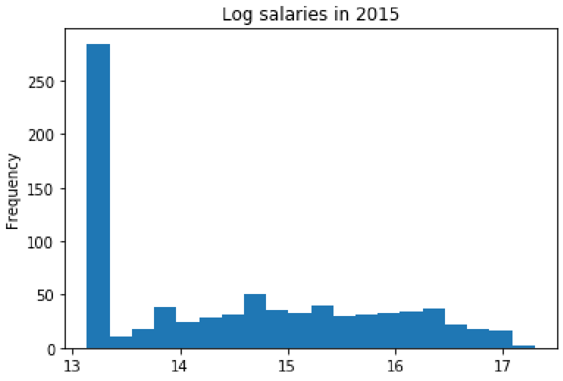
SALARIES

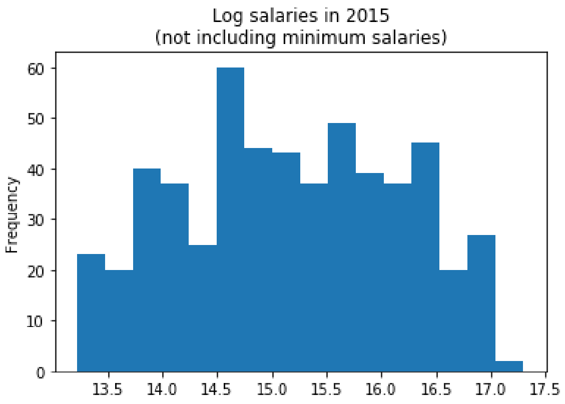
Salaries have increased significantly since 1985. From 1985 to 2015, total team salaries increased by over a factor of 11, though about half of the increase is due to inflation. Adjusting for inflation, total team salaries have still increased by a factor of 5. The rise in salaries has been fairly steady over time, though there is a sharp increase in median salary beginning in 2013. Below are line charts showing the increase in both median team salaries and individual salaries over time.



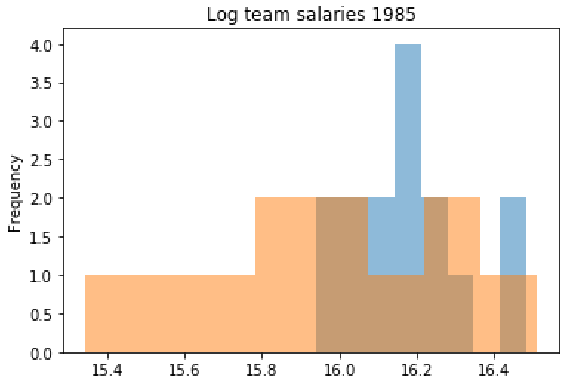


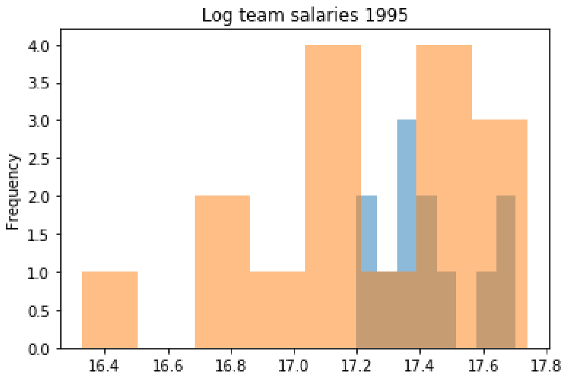
The distributions of log individual player salaries are consistently bi-modal over time. Below is a histogram of log player salaries in 2015 (most other years looks similar) and another histogram of log salaries without the minimum salaries. As we can see, most players receive salaries at the low-end of the distribution (around the minimum of $507,500), while the distribution of those who receive higher salaries is actually fairly normal.

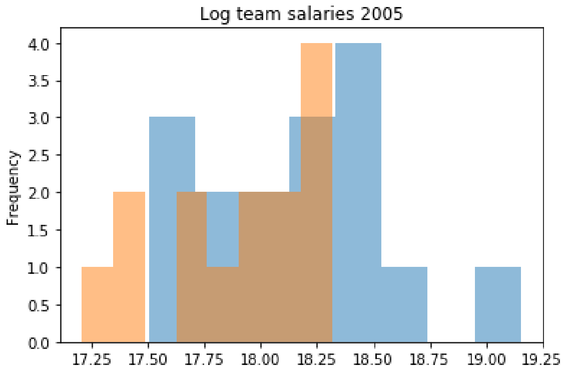


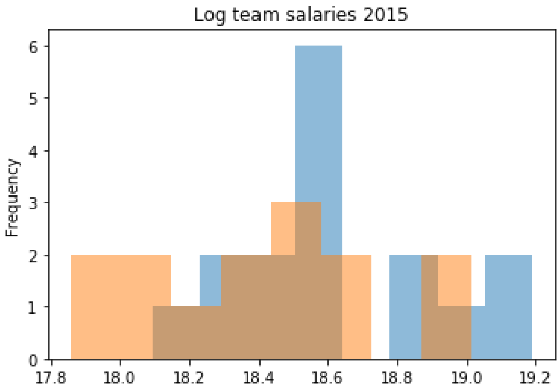


Total team salaries are fairly normally distributed, especially when logged. Below are histograms of log total team salaries from 1985 to 2015, every ten years. The blue bars represent teams with a win-loss record at or above .500 and the orange bars represent teams with a record below .500.

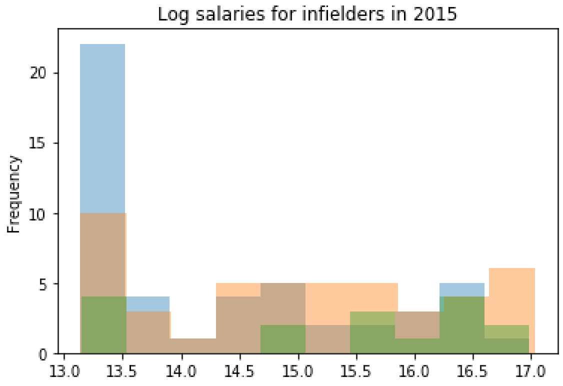


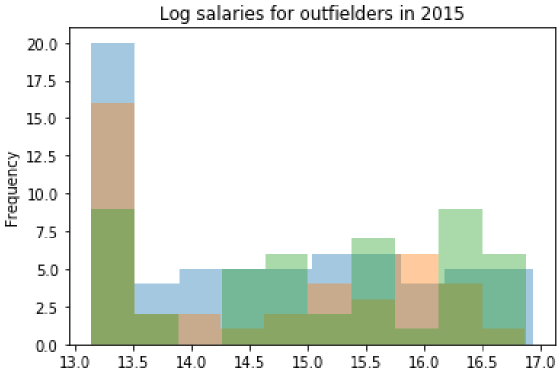




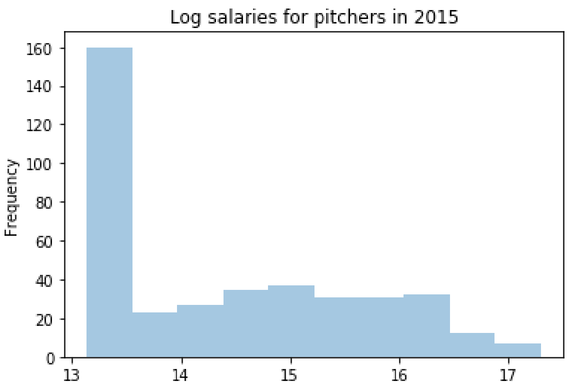


As mentioned above, each player is usually responsible for a position on the field, such as first base or left field. We can examine how player salaries differ by starting position. Below are histograms showing salary in 2015 by starting position. For infielders, blue bars are second basemen, orange bars are first basemen, and green bars are designated hitters. For outfielders, blue bars are left, orange bars are center, and green bars are right fielders.



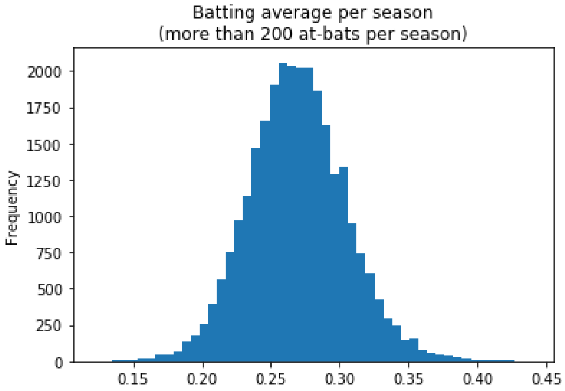


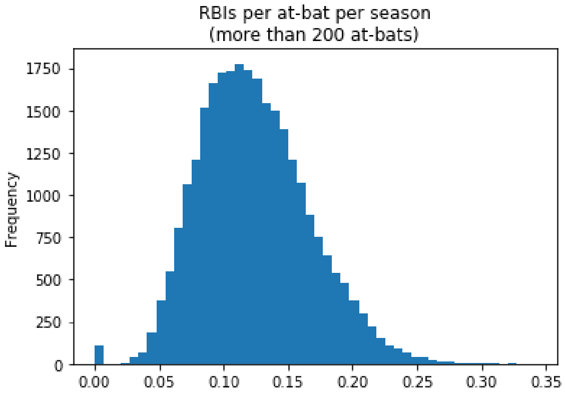
We can see a difference in salary by position here. Designated hitters and right fielders tend to be paid more than other positions, while second basemen and left fielders are paid less.

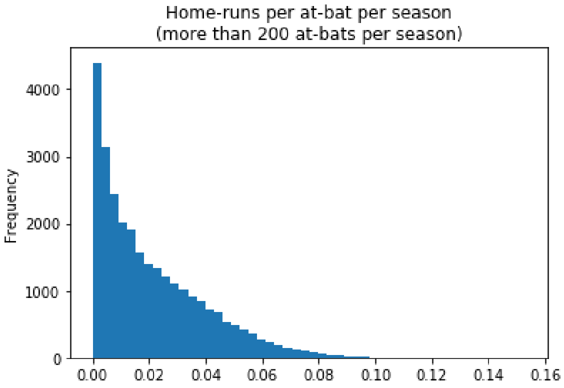


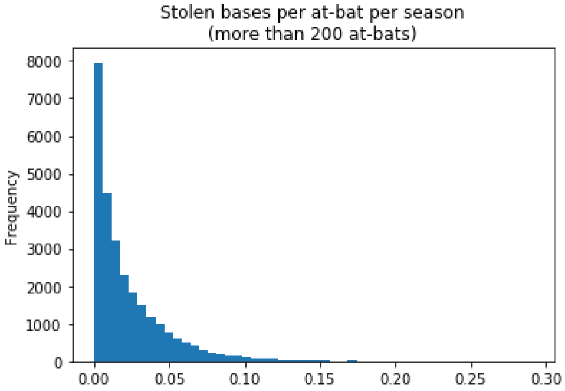
Player stats

Below are histograms for select player stats. On their own, each stat tends to be right-tailed. By simply averaging some of the stats by at-bats or innings-pitched, we can produce some normally distributed variables, but some remain quite right-tailed. First, I show some histograms for batting stats, averaged by at-bats (and only including players with at least 200 at-bats):

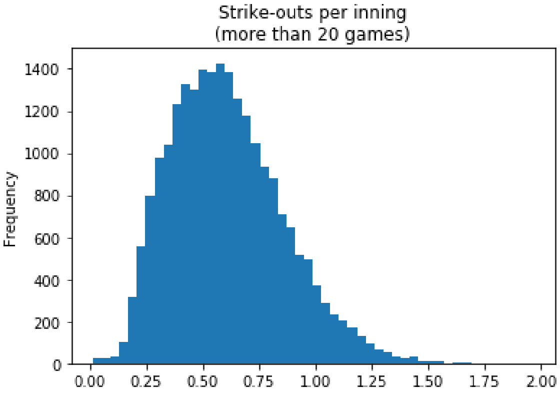


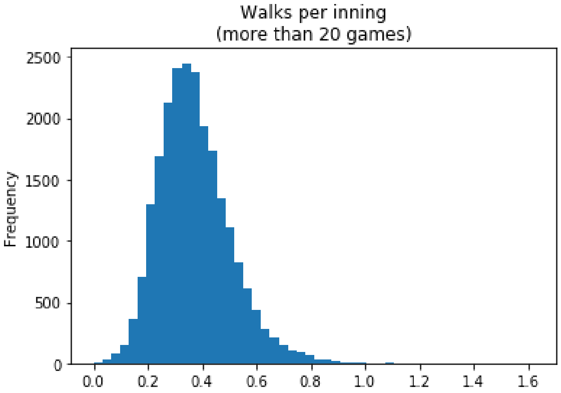






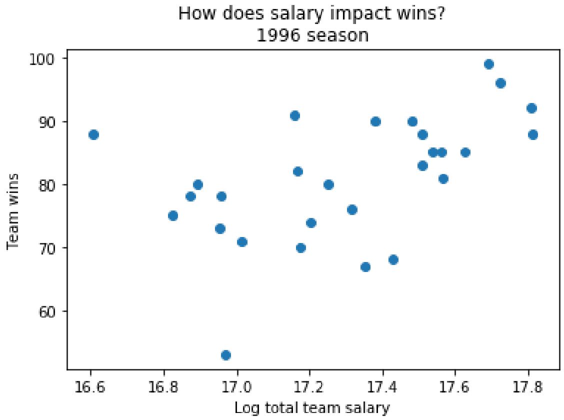
Next, some pitching stats, averaged by season total innings-pitched (and only including pitchers who played in at least 20 games):

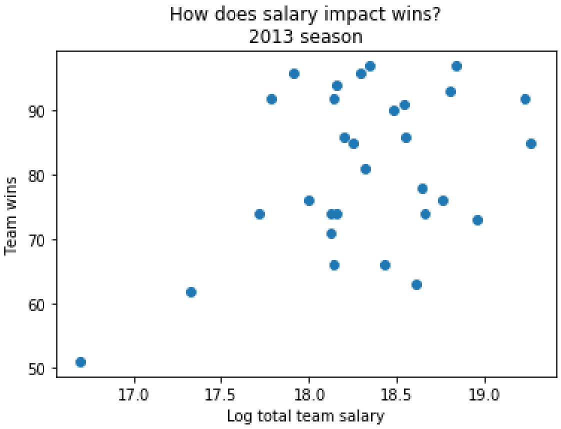




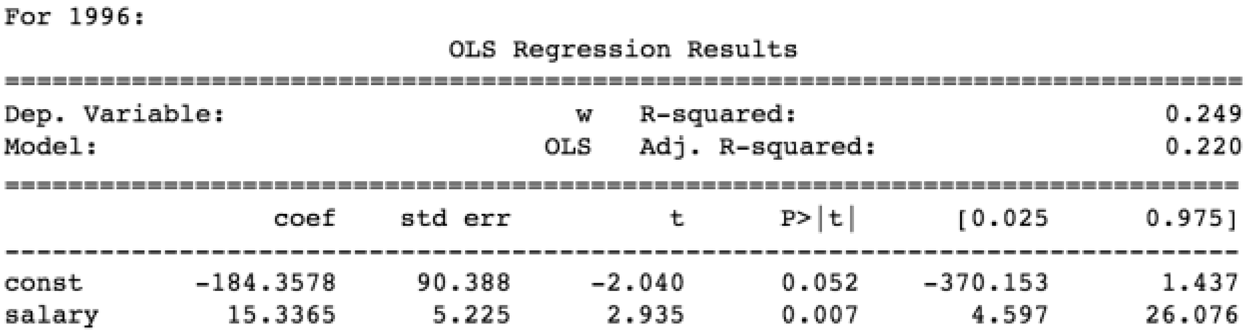
What is the impact of salary on wins?

In our proposed problem, we assume a relationship between total team salary and the number of wins a team will achieve. Below are scatterplots of log total team salary and number of wins for the 1996 and 2013 seasons (the least- and most-recent statistically significant years).

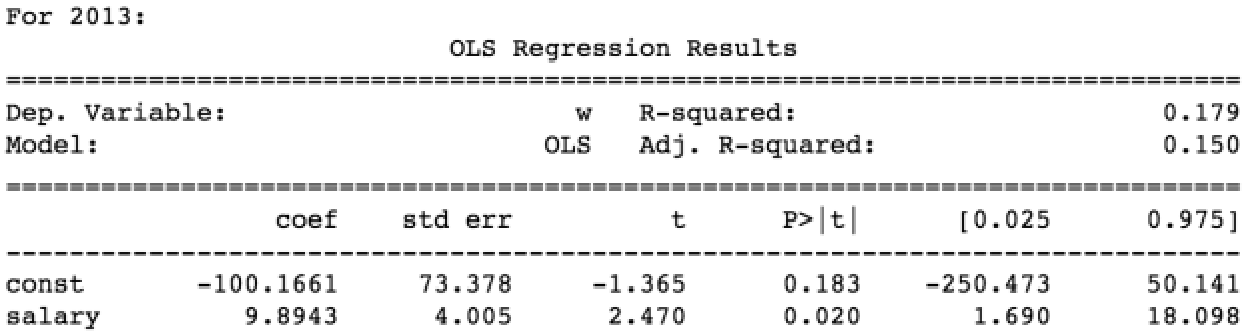




In order to test this relationship, I set up a set of linear regressions between log of total team salary for a given year and total number of wins for that year. Results vary from year to year. A number of years (1996, 1998, 1999, 2002-2007, 2009, 2010, and 2013) show a statistically significant (alpha=0.05) positive relationship between log of total team salary and number of wins. For years with significant results, the coefficients tend to be around 10 to 15, meaning that an increase of total team salary by 10% predicts an increase of 1 to 1.5 wins, on average. Below are the regression results for 1996 and 2013.

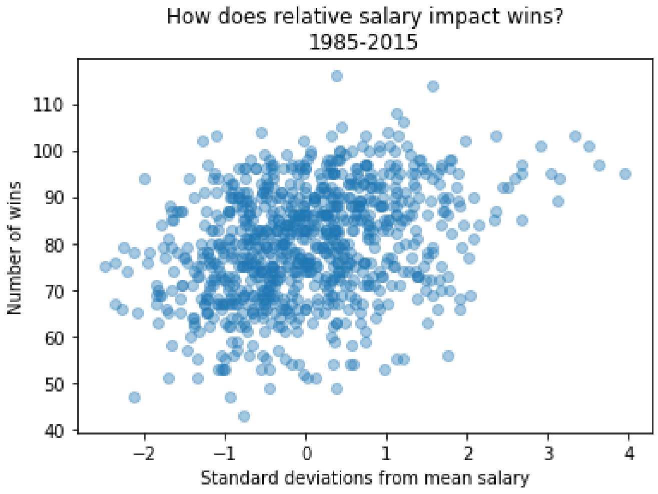


```
For 1996:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      w   R-squared:                       0.249
Model:                            OLS   Adj. R-squared:                  0.220
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -184.3578     90.388     -2.040      0.052    -370.153       1.437
salary        15.3365      5.225      2.935      0.007       4.597      26.076
```

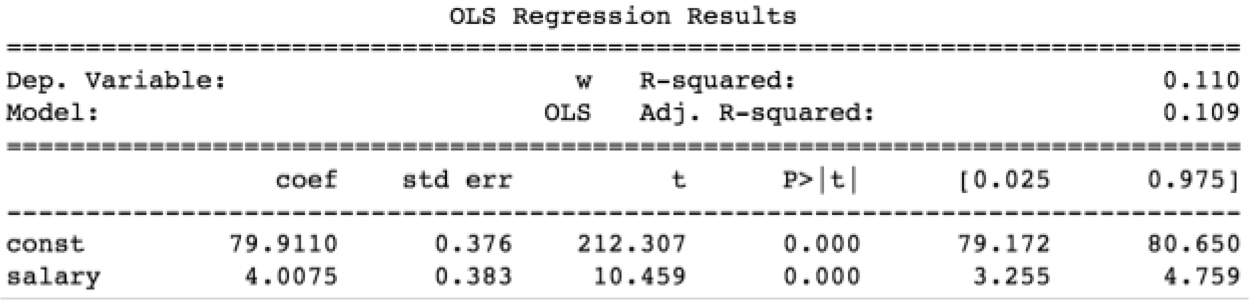


```
For 2013:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      w   R-squared:                       0.179
Model:                            OLS   Adj. R-squared:                  0.150
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -100.1661     73.378     -1.365      0.183    -250.473      50.141
salary         9.8943      4.005      2.470      0.020       1.690      18.098
```

In order to include data for all years, I normalized the data (mean=0, standard deviation=1). Normalized, I can compare the data year-to-year without worrying about yearly trends and inflation, and run a linear regression including all the data. The scatterplot of this data is below.



Below are the results of the regression. We get a statistically significant result: an increase of total team salary by one standard deviation from the mean predicts an increase of four wins, on average.



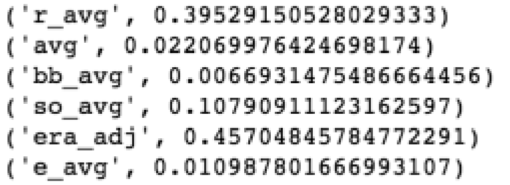
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      w   R-squared:                       0.110
Model:                            OLS   Adj. R-squared:                  0.109
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         79.9110      0.376    212.307      0.000      79.172      80.650
salary         4.0075      0.383     10.459      0.000       3.255       4.759
```

It seems that, overall, a higher total team salary is correlated with more wins.

DECISION TREES: WHAT STATS ARE IMPORTANT FOR WINNING?

In order to determine what stats in players we should look for when constructing a team, I set up several decision tree regression models. [1] In the first model, I set the dependent variable as number of wins and the independent variables are a variety of stats, aggregated at the team level. In order to account for variation in length of games and other non-performance factors that would cause biased absolute stats, we average batting stats by number of total at-bats. In the second model, I set the dependent variable as average number of runs per at-bat and the independent variables are three commonly used averages, at the individual player level. In the third model, the dependent variable is the same as in the second model, but the independent variables are averages of several batting stats.

In our first decision tree model, we use team batting average, earned run average, and averages of runs, strikeouts, walks, and errors to predict winning percentage. We limit the data to seasons since 1984 in accordance with our salary data. The model produces the following feature importances:

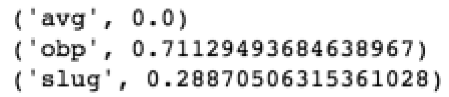


```
('r_avg', 0.39529150528029333)
('avg', 0.022069976424698174)
('bb_avg', 0.0066931475486664456)
('so_avg', 0.10790911123162597)
('era_adj', 0.45704845784772291)
('e_avg', 0.010987801666993107)
```

The most important predictors of winning percentage are, unsurprising to most baseball fans, the number of runs the team scores and the number of runs scored against the team.

But simply the number of runs is not a helpful statistic for picking players to build a team. What other statistics best predict the scoring of runs? For the next set of models, we will look at individual batting statistics, limited to seasons since 1984 and players with at least 100 at-bats in the season. In our second model, we use batting average (hits per at-bat), on-base percentage (number of times a batter gets on base per at-bat), and slugging average (a batting average weighted toward bigger hits) to predict the number of runs per at-bat. The model produces the following feature importances:

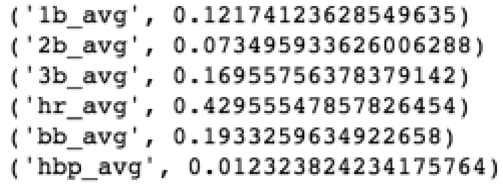
[1] I choose to only use batting stats because of some of the complications of using the pitching stats: adjusting for starters versus non-starters, lack of meaningful stats, etc.



```
('avg', 0.0)
('obp', 0.71129493684638967)
('slug', 0.28870506315361028)
```

On-base percentage is by far the most predictive. Normal batting average contributes nothing! So the most important metric for scoring runs is the ability to get simply on base.

A batter can get on base by getting a hit, a walk, or getting hit by a pitch. Additionally, there are four kinds of hits: singles, doubles, triples, and home-runs. In our third model, we use at-bat averages of each of these statistics to predict runs. The feature importances of this model are:



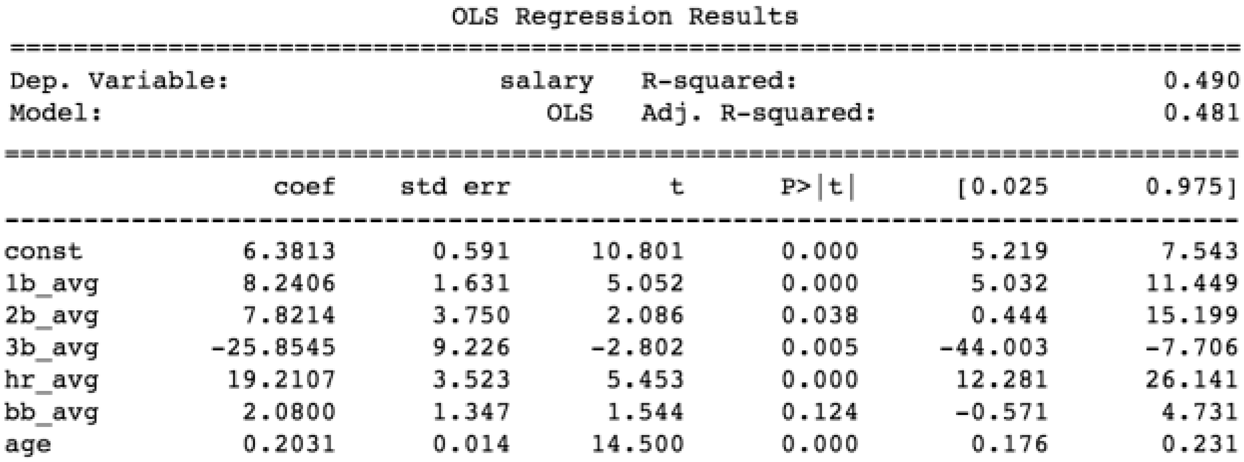
```
('1b_avg', 0.12174123628549635)
('2b_avg', 0.073495933626006288)
('3b_avg', 0.16955756378379142)
('hr_avg', 0.42955547857826454)
('bb_avg', 0.1933259634922658)
('hbp_avg', 0.012323824234175764)
```

Home runs are by far the most predictive variable for scoring runs, which makes sense. A home run is scores a run every time, while any other method of getting on base leaves the possibility of getting out or being left on base when the inning ends. However, the other variables are less obvious. Doubles and hit-by-pitches are by far the least predictive, while walks and triples are more predictive than I originally expected.

PLAYER STATS AND SALARY

Now we know that scoring runs is the most important metric for winning, and that the ability to get on base is the way score runs. Further, there are different ways to get on base, and we determined that home runs, triples, and walks are the most predictive of scoring runs. What does this mean for our team owner who wants to create a winning team on a budget?

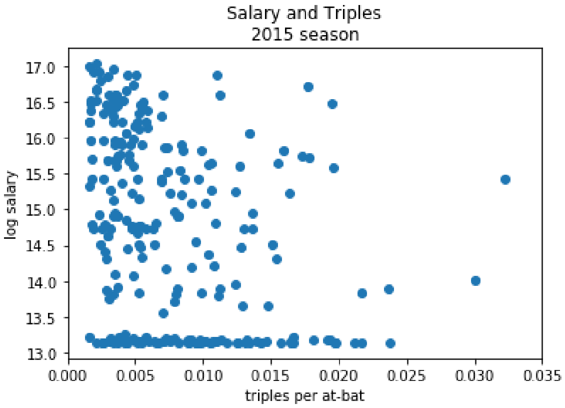
In order to find out how each of these batting statistics is correlated with salary, I set up a linear regression with the same stats as the third decision tree from above as the independent variables (as well as player age to control for other salary differences). Similar to above, I also limit the data to the 2015 season and to players with at least 100 at-bats in the season. Below are the results:

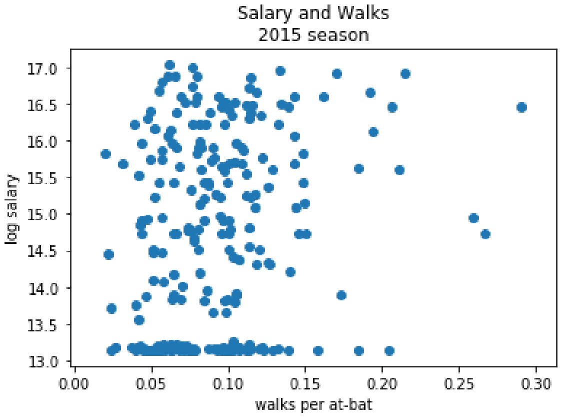


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.490
Model:                            OLS   Adj. R-squared:                  0.481
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.3813      0.591     10.801      0.000       5.219       7.543
1b_avg         8.2406      1.631      5.052      0.000       5.032      11.449
2b_avg         7.8214      3.750      2.086      0.038       0.444      15.199
3b_avg       -25.8545      9.226     -2.802      0.005     -44.003      -7.706
hr_avg        19.2107      3.523      5.453      0.000      12.281      26.141
bb_avg         2.0800      1.347      1.544      0.124      -0.571       4.731
age            0.2031      0.014     14.500      0.000       0.176       0.231
```

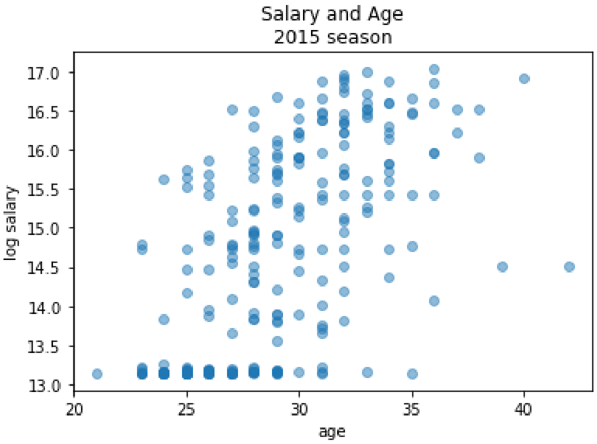
We see statistically significant results from nearly all the variables. Unfortunately, there are likely to be some serious multicollinearity issues with these variables. Since we are using the numbers as qualitative measures of what statistics most impact salary however, not as precise predicators of salary, they are still useful for our analysis here. We are also making the assumption that a player's past statistics are fairly predictive of his performance in the future. While this may not be accurate for all players, this assumption should be true on average.

Run individually, we find a negative relationship between triples and salary, and only a very weak positive relationship between walks and salary. Both triples and walks are somewhat predictive of scoring runs, so these are good statistics to look for in players. Below are scatterplots showing the relationships between triples and walks and log salaries for individual players for the 2015 season.





Age also plays a major factor in player salaries. There is a fairly strong positive relationship between age and salary, so younger players can be hired for lower salaries. Below is a scatterplot showing the relationship between age and log salaries for individual players for the 2015 season.



CONCLUSION

In order to win Major League Baseball games, there is no doubt a team needs to spend a lot of money constructing a team. In order to win a majority of their games, a new team owner should expect to spend about $80 million on player salaries. But money can be spent in an intelligent way. Trying to hire the "very best" players is not necessarily the optimal way to win games, especially on a small budget.

If someone wants to start a brand-new MLB team, they should hire players who know how to get on base, reflected in the ability to hit singles and triples and draw walks. A team should particularly look for a few inexpensive players who are known for hitting triples. The team owner should also resist the urge to hire big homerun hitters because, while homeruns are a sure way to score runs, homerun hitters are also more expensive. Homerun hitters are flashier and may make a team more exciting to watch, but they also demand higher salaries. Finally, the team owner should hire younger but promising players, because while older players may be better and more predictable, they also tend to be paid more money.