

# 126 Data project, step 1

Sam Ream, Valeria Lopez, Skyler Yee

## Data Description

“The History of Baseball” is a record of all the stats of baseball players who have played in the MLB up until the year 2015. While the data in this source is extensive, we elected to narrow the range of the years down to 2000-2015 to more accurately represent the changing landscape of Baseball strategy.

## Data Source

Source: “The History of Baseball” by Sean Lahman,

URL: <https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball> (<https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball>)

## Population Description

Since we sampled from games played between 2000-2015, our population is all of the MLB players from 2000-2015. We did not want to generalize the whole history of the MLB because the rules of the game and the strategies used by teams and players are always changing.

## Variables

**BMI** - categorical (underweight, healthy, overweight, obese); The formula is  $\text{weight}/(\text{height})^2 \cdot 703$ . Weight is in pounds and height is in inches.

**Batting Hand** - categorical (left, right, both); whether the player hits with his left or right hand

**Singles** - quantitative; The number of times the player gets a hit and lands safely on first without stopping

**Doubles** - quantitative; The number of times the player gets a hit and lands safely on second without stopping

**Triples** - quantitative; The number of times the player gets a hit and makes it safely to third without stopping

**Home Runs** - quantitative; The number of times player hits the ball and makes it all the way around the base path to home plate safely without stopping

**Walks** - quantitative; The number of times the player refrains from swinging at 4 bad pitches (balls) during their at bat and is allowed to advance to first base

**Stolen Bases** - quantitative; The number of times the player safely makes it from one base to another when there is not a ball in play

**Hit by Pitch** - quantitative; The number of times the player is hit by a pitch and is allowed to advance to first base

**Intentional Walks** - quantitative; The number of times the opposing team elected to walk the player on purpose, allowing them to advance to first base

## Summary Statistics

Data summary

---











Name	batting
------	---------

Number of rows	500
Number of columns	13
Column type frequency:	
character	3
numeric	10
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
player_id	0	1	7	9	0	500	0
BMI	0	1	1	1	0	3	0
HAND	0	1	1	1	0	3	0

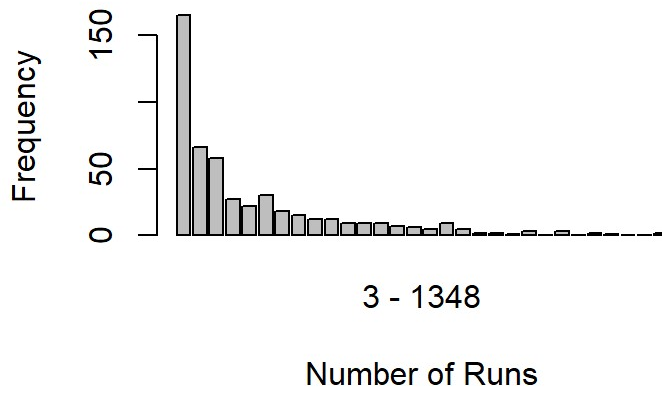
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
AT_BAT	0	1	1547.25	1688.83	101	303.50	886.0	2197.25	9362	
RUNS	0	1	206.61	248.71	3	29.00	104.5	288.00	1348	
HOME_RUNS	0	1	48.68	70.75	0	5.00	17.5	63.00	409	
TRIPLE	0	1	7.79	10.93	0	1.00	3.5	11.00	91	
DOUBLE	0	1	81.73	97.56	0	11.00	41.5	111.50	478	
SINGLES	0	1	270.59	317.67	6	45.00	145.0	378.00	2390	
WALKS	0	1	145.13	184.50	0	20.75	65.5	213.00	1160	
INT_WALKS	0	1	11.45	20.92	0	0.00	4.0	14.00	179	
STOLEN_BASES	0	1	25.20	50.25	0	1.00	7.0	27.00	498	
HIT_BY_PITCH	0	1	15.71	23.71	0	2.00	6.0	20.00	179	

These tables are derived from a randomly selected group of 500 players who played in the MLB (Major League Baseball) during the years between 2000 and 2015 (inclusive). From these tables, we can observe that all variables have no missing values, which indicates that the dataset is complete. The mean number of runs that a player will have is 206.61 and standard deviation is 248.71 with a minimum number of runs equal to 3 and a maximum of 1348.

# Individual Distributions

## Frequency of Number of Runs

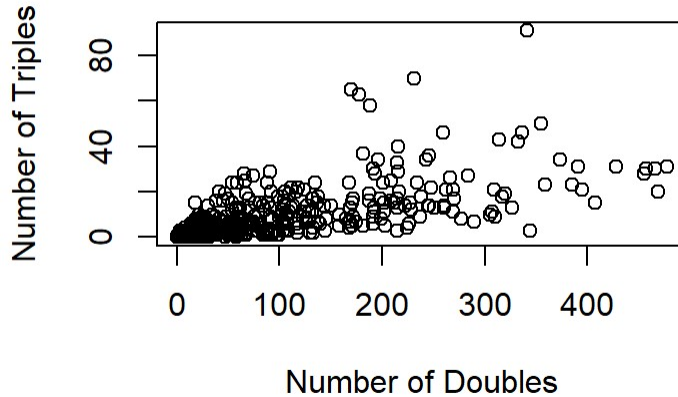


Frequency of Number of Runs

This Histogram shows the distribution of the frequency of the number of runs obtained by a random sample of 500 players. This distribution takes a form similar to the exponential distribution, implying that a large number of players obtain very few runs while a small number of player obtain many runs. The range of the number of runs is 3-1348 as previously noted in the summary tables.

## Numerical Relationships

### Comparison of Doubles and Triples

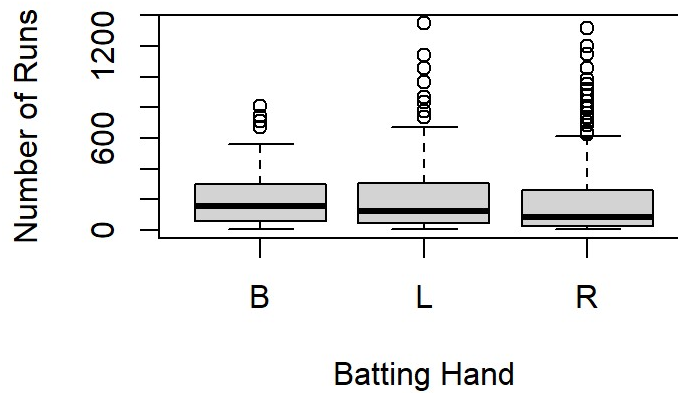


Comparison of Doubles and Triples

While we were initially concerned that Doubles and Triples may be correlated as they both represent a players ability to get to a different base. However, we can see that the correlation between the two is not as strong as we worried in this graph.

## Categorical Relationships

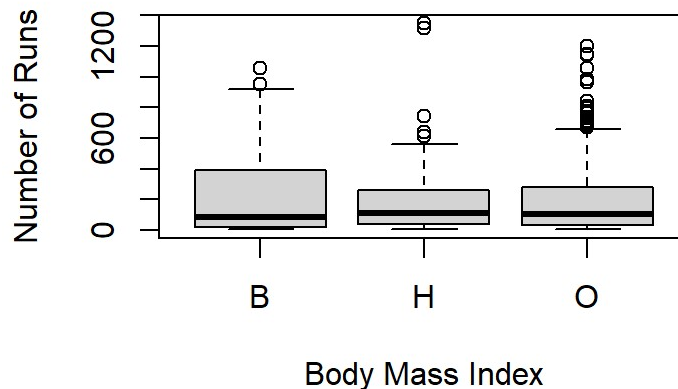
## Number of Runs and Handedness



Number of Runs and Handedness

This graph displays the batting hand of players in relation to the number of runs. The “L” stands for “left”, “R” stands for right, and the “B” stands for both. There was not much difference between the three categories as they have similar minimums, means, and third quartile values, while maximums vary a bit more. While the 3 different categories displayed a similar range for the number of runs, those with a right hand more regularly placed above the fourth quartile of their group.

## Number of Runs and BMI



Number of Runs and BMI

This graph displays the relationship between players' BMI and the number of runs. In the graph “B” stands for obese, “H” stands for healthy, and “O” stands for overweight. A player falls under the underweight category if their BMI falls below 18.5, healthy if between 18.5 and 24.9, overweight if between 24.9-29.9, and obese if 30 or above. In our sample, there were no underweight players. Most of the players that fell under the obese category scored the most runs and had the fewest persons outside of the first and fourth quartile. The smallest range of data was for the players that had a healthy category with a few outliers who had the highest runs in the population.

## Conclusion

The data was approximately what we had expected which is shown in our calculated batting average (the number of a player's hits divided by their total number of at-bats) being around 0.246 which is close to MLB's 0.250 calculated value for the league. We sampled our data randomly to get an accurate representation of the population. We originally had hits as an independent variable, but realized that we would be double counting and would make our estimators partially unidentifiable, so we removed it from our data.