Summary

Baseball dataset is consisted of physical characteristic of 1157 baseball players and their performance scores. The handedness (whether players right or left handed), weight, height, and names. Then their home run and average batting scores.

In my story, I wanted to examine relationships between players physical characteristics and their performance. Then verify if there is correlation and what is it.

Feel free to view my story by clicking or copy/paste the link below:

https://public.tableau.com/profile/ali6684#!/vizhome/BaseballVariables_AliHasan/Story1?publish=yes

Initial Design

From an initial data exploration dropping measures and dimensions to columns and rows, I have noticed the followings:

- There are a lot of players with average batting zero. Most likely these players are the catchers and it is irrelevant to include them in our calculations to examine home run and batting average. Way to get more accurate results. You can't be a professional baseball player with zero average batting score. Yet this doesn't apply to home runs as you might be able to hit while batting but never made a home run. I have created a new group for player with Non-Zero Average Batting to help me filter this value out
- I must be careful when using SUM, COUNT, or AVERAGE in my calculations. Like sum or count of left or right handed players scores show different results when using average. Which later helped me find an interesting fact that will be explained later in my project. According to the BMI scale, if you fall between 18.5 and 24.9 you're considered to be a normal, healthy weight. If you clock in below 18.5, you are considered underweight. On the other side of the scale, if you're between 25–29.9 this that you're overweight. And, finally, if your BMI is higher any higher than 30 means that you are obese.
- I noticed some players have duplicated records. This came when I saw some players with BMI values over 46.07 which is not realistic for a baseball player and a very over weight measure

Data Cleaning

- Taking catchers off resulted in reducing records numbers from Total = 1157, right = 737, left = 316, and both 104 to Total = 891, right = 546, left = 256, and both 89. So that you will see different numbers from original dataset.
- For players have duplicated records, filter them out either using BMI summation and apply filters then filter out whatever BMI over 46.07:

name	handedness	height	weight	avg	HR
Dave Roberts	R	75	215	0.239	49
Dave Roberts	L	75	195	0.194	7
Mike Brown	R	74	195	0.265	23

Mike Brown	R	74	195	0	0
Mel Stottlemyre	R	73	178	0.16	7
Mel Stottlemyre	R	72	190	0	0
Dave Stapleton	L	73	185	0	0
Dave Stapleton	R	73	178	0.271	41
Bobby Mitchell	L	70	170	0.243	3
Bobby Mitchell	R	75	185	0.235	21
Jim Wright	R	77	205	0	0
Jim Wright	R	73	165	0	0

Updated design

I have decided to use bar, histogram, scattered, table and line plots that can illustrate my explanations slides.

In the first slide, I used bar charts and tables to explore handedness with graded blue color. It wasn't very clear to distinguish between right and both hands so that I added HR scores and handedness labels to be clearer. I added it as a table too.

In the second slide, I used bar chart for players names vs BMI. Clearly shows up duplicated names when I sum their BMIs (to be filtered out later). I have added it as a table too.

In the third slide, it is the most interesting slide with all features and performance scores. I used color labels for handiness. They I used text labels and bar charts for BMI, Home Run counts and Average Batting.

In the fourth slide, I used histogram plot to show players BMI distribution then BMI distribution vs home run.

In the fifth slide, I used scattered plot and line chart to plot home run vs average batting scores.

Finally, I used histogram again in the sixth slide to plot Home Run distribution and average batting distribution.

Findings

I have noticed some new facts and confirmed others I have already knew so let's go:

- The number of right-handed player is more than double left-handed yet the average is opposite. Which means the rate of left-handed players score home runs and average batting is higher than both or right handed. You can see that in slide 1 and 3. In Slide 3, it is more obvious that the first best 4 players in average batting are left-handed (green color). And the 3 out 4 best players in home runs are left-handed too.
- Most of the baseball players are in a healthy/athletic shape which and specifically between BMI values 22 and 26 as you can see in slide 4. I expect that won't be the case with American football and some other sport. The highest home runs were scored at BMI bin 24^{th}.

- The third and logical finding was for comparison between Home Run and Average Batting. The players with good batting average scoring more home runs. It is a positive correlation and logical at same time. That's illustrated in slide 5.
- The home run distribution it right skewed which is logical. So that the number of players scored less than 20 home runs is higher. Slide 6
- The average batting is normally distributed between 0.22 and 0.28. Slide 6

Feedback

I'm not very knowledge able about baseball so that I have asked my friend who is a baseball fan and very knowledgeable about the game. He suggested adding slugging percentage to the dataset. Slugging percentage reflects the percentage of plyers doubles, triples and homes. Which is in this case can provide another performance measurement and better variable to rank player accordingly

Resources

I have used baseball dataset provided to us by Udacity