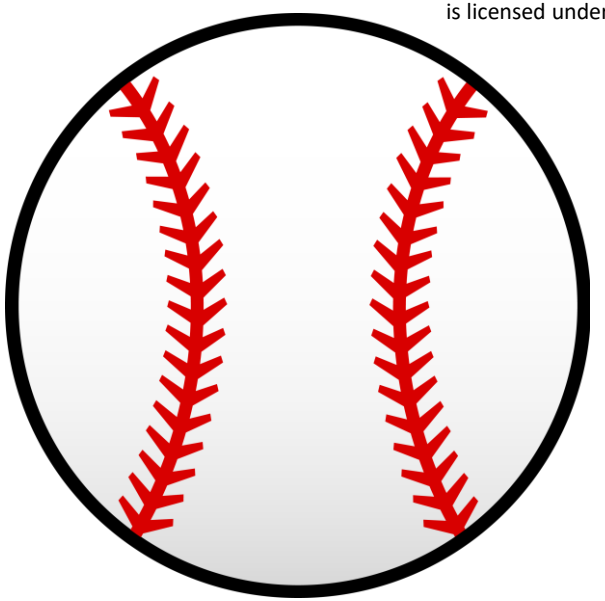
Motivation



My interest in the baseball dataset was twofold.

i. I had not worked with sports data before.

ii. While I enjoy watching the Nationals, I never pay close attention to baseball stats

After working with the data in class, I wanted to do some further exploration to see if I could discover anything interesting in the scope of this assignment.

A glimpse at an amateur's feature engineering…

In class, we discussed adjusting statistics to adjust a per game basis.

Here, I do that for several stats.

```
df = df.withColumn('HpG', df.H / df.G)
df = df.withColumn('HRpG', df.HR / df.G)
df = df.withColumn('HBPpG', df.HBP / df.G)
df = df.withColumn('BBpG', df.BB / df.G)
df = df.withColumn('SLG', df.H/df.AB)
```

SLG, or slugging average, I discovered on Wikipedia, but decided not to delve into.

(https://en.wikipedia.org/wiki/Baseball_statistics)

Instead, I look at the relationship between Hit/Game vs. Ball on Base/Game

Visualization

While the visual depicts a loosely positive relationship between H/G and BB/G, they share a Pearson's r value of .75

While I initially wanted to increase point transparency to better visualize the linear relationship in the highly dense portion of the graph, I did not want to lose the outliers (some of which I've put in circles).