Historical research can be challenging, especially if one has to pick a beginning and end to the topic in question. Is

Some research historical research questions are at first glance simple. If you are writing about the American Civil War, a logical place to begin the analysis is with the shelling of fort Sumpter in April 1861 <http://www.smithsonianmag.com/history/fort-sumter-the-civil-war-begins-1018791/> and a good place to end it is with the surrender at Appamottox court house in April 1865. <https://www.nps.gov/apco/learn/historyculture/the-surrender-meeting.htm> . Sounds easy, until one has to analyze the causes and/or effects of the civil, then the point of analysis becomes more difficult. Do we end the analysis with the so-called reconstruction era? Do we begin the analysis with the formation of the union, or do we go even earlier to the beginnings of slavery?

This is where machine learning can help us parse through history, and divide the centuries into specific periods. It’s also where baseball can serve as a laboratory for this exploration into historical classification.

K-Means Cclustering is an area of unsupervised machine learning that essentially separates samples into n groups of equal variance. It does so by taking centroids, and assigning samples to the nearest centroid. It then creates new centroids based on the mean value of the samples. It continues this process until the centroids stop moving.

Sound confusing? Imagine empting the contents of a fruit loop cereal box, and then trying to make sense of the mess on the floor. We can organize it by color, we can organize it by size. Or, we can create centers in the floor and compute the distance of the cereal bits on the floor from those centroids and assigne them accordingly. http://scikit-learn.org/stable/modules/clustering.html#k-means

In basic terms, the algorithm has three steps. The first step chooses the initial centroids, with the most basic method being to choose  samples from the dataset . After initialization, K-means consists of looping between the two other steps. The first step assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly.

Baseball historians break down time into several eras as well, such as the Dead-ball era 1901-1919 or 1920), and the Free Agency Era (1977-1993). We recently saw the advent of the Steroid Era.

When Examining baseball statistics, or any other data analysis, where one begins and ends an analysis can influence the end result.

Let's look at home runs for example