Brandon J Knight

Data mining

Project

Project Report

For my project I will be analyzing the correlation between known steroid users and non-steroid users of the MLB. Alongside this data it is important to note that I will be analyzing a few of the greatest baseball players of all time. I will be analyzing this data in the following specified areas of interests: hits, rbi's, runs, and home runs. All of my major data on my spreadsheets were provided by Sean Lahman. An award winning database journalist, author, and creator of the Lahman database. The Lahman database is the largest, free to use collection of statistics for major league baseball teams, players, and seasons from 1871 to present. I will also be using stats provided by Baseballreferece.com.

To conduct this analyses I chose to search for the most well-known steroid users in baseball history. I chose to use the data provided by the Mitchell Report. The **Mitchell Report** is the result of former US Senator George J. **Mitchell**'s 20-month investigation into performance-enhancing drug use in Major League Baseball (MLB). After taking all of the players that were recognized as steroid users in the MLB, I took a list of the highest ranking players within the MLB. In order to get this data set, I visited ESPN's hall of fame for the top 100 players of all time. This is the list of the top 20 players. Some of the names mentioned in this top 20 were also considered to be steroid users and some were mentioned in the Mitchell Report, so those are the names that I will not be comparing to other steroid users.

There are a few things that should be mentioned that are important in analyzing these few data sets. The first thing to note is the various change within the start of the players career compared to later on in the players career. There will most likely always be a noticeable increase in the stats after the first two or so year of the player's career. This is due to the increase of games that the player was actually active in. Due to the increase of games per year after the first 2 years, there will most likely be an increase in there stats which is unrelated to the steroid usage during that time. There are also heavy decreases and sometime radical slope changes throughout certain graphs that also implicate a change within the amount of games that player played that year due to either injury, suspension or any other personal venture. To account for these radical changes, I analyzed 3 similar stats (in comparison to number of games) of each player so that I could obtain a closer margin and more accurate analysis throughout my data. The 3 similar stats I chose to analyze where in chronological order of the players first year (To which I took the stats of each players first year of play and compared them), similar number of games (To which I took the stats of a year both players had a similar number of games and highest stat per game year), and finally the players 10th year (To which I took the players 10th year of play and compared those two sets of data). The highest stats of these players will also be recorded however it will only include the highest stats within the 10 year period.

There were a couple of noticeable trends that I saw throughout my data. Although I predicted very high data sets for the players that were known steroid users, there were many times throughout the sets where that was not the case. As said before in the previous slides this could be due to a multitude of many different variables such as the amount of games played and when certain players were actually on steroids and when they were off them. Age is also another factor that I did not add into my data that could prove to vary in graphs when looking at that particular portion of data. Another valid variable is the recent change in supplement nutrition and food production within the 21st century's current market. There are needless to say, a lot more options when choosing what workout and what foods a player would like to choose while in season. There have also been many technological advances in recent years that no only help the player become a better athlete, but also help analyze specific areas of improvement needed for the player to have his best season.

There are many similarities between both players that are being analyzed in each graph however there are also many differences as I have stated and as you can probably suspect. The main differences is in the slope variables where the men that were on steroids seemed to always have a higher slope value than the men without. Meaning that the performance of these men changed slightly more than those who did not take steroids weather that meant for the better or for the worse. It also seemed that the steroid users had lower stats that some other recent MLB stars mainly because of the modern era we are in with the advances in the world of MLB and also the overall natural skill that possibly came from genetics rather than steroids that seemed to turn regular players into better players. Thank you for your time and attendance.