Question – Does it matter where you play college baseball if your goal is to reach the Major Leagues?

Introduction

The charts from my Midterm Project served as inspiration for the dashboard and charts implemented in JavaScript and D3, and for creating this narrative visualization. Of the 3 allowable hybrid structures, this narrative visualization follows the **interactive slide show** hybrid structure. It is built with scenes, annotations, parameters, and triggers. The **scenes** follow a template for visual consistency and follow an order to best convey the message. Each scene **discusses the layout** and the **visual structure**. The **annotations** also follow a template for visual consistency. Each annotation highlights **specific data trends** and indicates when it is cleared. The **parameters** are identified in each scene and discuss the states of the narrative visualization. The **triggers** are easily identified in terms of the **events** and the **parameter changes** they trigger. User interface events are listed, along with a description of **how possible user events are communicated to the viewer**. The opening scene in the dashboard includes a tab, labeled "**About the Visualization**", that addresses the key points of the rubric required for this assignment.

Dashboard published on publicly available web site:

https://rgjeldum.github.io/uiuc-projects/index.html
The narrative visualization for the year-end project was implemented on a publicly visible website, GitHub. The URL was submitted via Coursera for peer grading.

Data Source

Sean Lahman's Baseball Database
http://www.seanlahman.com

Data Description

This database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2016. It includes data from the two current leagues (American and National), the four other "major" leagues (American Association, Union Association, Players League, and Federal League), and the National Association of 1871-1875.

This database was created by Sean Lahman, who pioneered the effort to make baseball statistics freely available to the general public. What started as a one man effort in 1994 has grown tremendously, and now a team of researchers have collected their efforts to make this the largest and most accurate source for baseball statistics available anywhere.

Software / Tools Used

MAMP PRO 4.1.1 for local web server and IDE

JavaScript language
D3 libraries (https://d3js.org) - Data Driven Documents
MS Excel

Data preparation
The complete set of 27 csv files were easily downloaded from the web site onto my laptop. There was no special data cleaning required on the dataset but one csv file had to be enhanced with an additional column of data. Each of the colleges in the *Schools - corrected.csv* data file had to be identified as a Division I, II, III school, or Community College (CC), as this info was not in the original Lahman database. The division and community college information was easily obtained from the NCAA.org web site and added to Schools.csv file using Excel macros.

Subset of dataset used
I used only 3 of the csv files, *Master.csv*, *CollegePlaying.csv*, and *Schools.csv*, for my data analysis. I utilized an inner join across the *Master.csv* and *CollegePlaying.csv* tables, equating their *playerID* fields. I then created another left join using the CollegPlaying.csv and *Schools.csv* tables, equating their *schoolID* fields. This was how the Data Source, named "MLB Players and Colleges" was created in Tableau. The above mentioned data manipulation was done with a standalone JavaScript process, "".I did not filter out any of the data in the *Master.csv* table, which contains all MLB players from 1871 – 2016. All MLB Players, whomever played, were used in this analysis. Another interesting study would be to look at how, or if, the observed patterns changed over decades.

Data Errors
There were known 4 data errors discovered in two of the csv files I was using from the Lahman dataset. These were found to be confirmed issues after reading baseball data research blogs. For the purpose of this visualization, I corrected the 4 errors, and renamed the *Schools.csv* file as *Schools – corrected.csv*. I also modified the *CollegePlaying.csv* fie but kept the same name.

The **Dashboard** consists of 4 tabs, corresponding to 1 essay about the data visualization and 3 individual charts. Each is described below:

"About the Visualization" essay (PDF)
"Top College States" Chart
"Top 10 Colleges" Chart
"Top NCAA Divisions" Chart

"About the Visualization" essay (PDF)
This tab simply displays a PDF document that gives a complete description of this narrative visualization. This is one of the requirements of the project rubric.

"Top College States" Chart

The data visualization clearly shows that most Major League players got to the big leagues by playing at colleges in largely populated, warm weather states. The top 4 indicated are California (1247), Texas (536), Florida (473), and North Carolina (284). At the other extreme are South Dakota and North Dakota which are at the very bottom of the pack, having only 2 and 1 MLB players, respectively. This is not too surprising as it suggests that for a player to develop his skills to the maximum, it is best to play in a state where the weather allows far more practices and games to sharpen skills over the entire school year. Also worth noting, many of the warm weather schools have done very well in the NCAA tournament over the years, which is a major showcase for prospective players to be seen by MLB scouts before the June amateur draft.

A **treemap** was decided to be the most effective way to visualize the data by having the top states indicated by area, color, and numerical count of players. This design choice was much more intriguing to the viewer than using a simple bar chart. The area of each rectangle, as well as its shade of color were used to indicate the number of MLB players who played college baseball in that state.

The **x-axis** is the U.S. States (using standard U.S. Postal abbreviations). The x data represent a discrete, unordered, categorical dataset.

The **y-axis** is the distinct number of MLB players who played college baseball in each state before becoming a professional. This dataset is discrete, ordered, and quantitative.

The **color choice** was in the cool bluish green range – darker blue representing a higher number of MLB players that played college ball in that state, and light green representing fewer number of MLB players. I also added a simple “count glyph” or numeric label in the box for each state representing the actual number of players who played there. This numeric total reinforces the color and area indicators.

I like the way the **treemap** takes the viewers eyes from the upper left hand corner of the sheet to the lower right hand corner. The viewer’s eyes scan each column, from top to bottom and from left to right, as the player counts go from largest to smallest. The treemap visually reinforces which states (by area and color) had the most MLB players playing college baseball. A simple bar chart would not be as aesthetically pleasing nor would the viewer be able to find the top states as quickly, as his / her eyes would have to scan 50 bars (i.e., states) from left to right, looking for the tallest. With the treemap, the eye flows more naturally rather than having to search back and forth for the top states.

“Top 10 Colleges” Chart

This chart gives a very clear view of the top 10 colleges that produced the most Major League Baseball players over the time periods of 1871- 2016 (all time), 1960-1999 (modern era), and 2000-2016 (current players). The viewer is able to quickly see, by ascending order, what the top 10 producing colleges are. By hovering over any bar, the viewer can see the name of the college, its state location, and the number of MLB players that played at that school. Clearly the top 5 are in warm-weather states: University of Texas at Austin (107 players). USC (105), ASU (101), Stanford (86), and University of Michigan (76). In fact, for all time (1871-2016), 9 of the 10 top colleges were from warm weather sites. Holy Cross was the only exception, which is located in Massachusetts. Clearly coming out of big time baseball programs, increases your odds of becoming a professional baseball player.

"Top NCAA Divisions" Chart

In this chart, we use a simple pie chart with basic colors to show the relative sizes of the colleges that MLB Players played at. There are 4 main categories of colleges assumed: Division I (largest and most competitive), Division II (medium size), and Division III (smallest, often private schools). Community Colleges are a unique category by themselves, in that they are typically schools offering only 2-year degree programs and students often commute to school.

From this simple pie chart it is quite clear that the bigger schools (Division I) produced the most MLB Players and by a far margin; in fact, almost twice as much as the next category (CC). Surprisingly, the Community Colleges (CC) category is a clear second, and the Division II and III schools far behind. This would imply that the best chance for a talented college baseball player is to choose a Division I college or go to a local Community College, if he has aspirations of playing in the major leagues.

Summary

Serious baseball fans have noticed for some time that the annual MLB draft of amateur players results in top picks coming from colleges located in highly populated, warm-weather states. Schools like USC, UCLA and Stanford, located in California, have a rich tradition of providing top talent to Major League Baseball, and many power pitchers have historically come from the state of Texas. Talented high school players should think twice about the state they play their college baseball in if they want to increase their chances of being seen by MLB scouts and ultimately get drafted. Baseball is a highly skilled sport which demands extensive hours of practice to get your skills to the professional level and to keep them sharp. Warm weather states, offering a climate to play baseball year-round, provide the ideal venue for that to happen. Finally, it is clear that playing at a Division I school or Community College (in a warm weather state) clearly increases the odds for success for college baseball players with ambitions of making it the big leagues.

Author / Student: Richard Gjeldum
NetID: gjeldum2
CS 498 - Data Visualization
Instructor: Dr. John Hart
Summer 2017
Final Project: Week #11

Software Acknowledgement
This visualization was made possible by modifying code provided by:

Scott Murray, Choropleth example from "Interactive Data Visualization for the Web"
https://github.com/alignedleft/d3-book/blob/master/chapter_12/05_choropleth.html

Juan Cruz-Benito, Animated Horizontal Bar Chart with Tooltips
Juan Cruz-Benito (juancb)’s Block ab9a30d0e2ace0d2dc8c
http://bl.ocks.org/juan-cb/ab9a30d0e2ace0d2dc8c

Juan Cruz-Benito, Animated Donut Chart with Labels, Legend and Tooltips
Juan Cruz-Benito (juancb)’s Block 1984c7f2b446fffeedde
http://bl.ocks.org/juan-cb/1984c7f2b446fffeedde

Mike Blostock
http://bl.ocks.org/mbostock
great examples of all things D3