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

Introduction
This narrative visualization follows the **interactive slide show** hybrid structure. The **Dashboard** consists 3 charts. The viewer is encouraged to begin the narrative by selecting any of the urls located below the enticing questions on the Dashboard. Once you begin the journey, you can further select the *Next* button (or a preferred *Tab* located at the top of each page). The viewer is quickly enticed to find the answer to - *Does it matter where you play college baseball if your goal is to reach the Major Leagues?*

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

3 Data Visualization Charts
"Top College States"
"Top 10 Colleges"
"Top NCAA Divisions"

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

This narrative visualization 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. The 1st scene is a Choropleth, the 2nd a horizontal bar chart, and the 3rd chart a donut-type, animated pie chart. Each scene has tabs at the top of the page to enable the viewer to quickly go to any of the other charts. It also has a *Next* or *Prev* navigate button to continue on to the next chart in the visualization story (or start over).

The **annotations** also follow a template for visual consistency. Each annotation highlights **specific data trends** and indicates when it is cleared. In each of the 3 charts you will see enticing commentary at the top of the page, along with tooltips (from hovering) that will indicate player counts, college names, state names, NCAA Division, and percentages in the case of the 3rd scene (an animated donut).

The **parameters** are clearly identified in each scene (a 3-way toggle switch at the top of each page). The key parameters consistently used in each scene are the viewer's choice of time period of Major League Baseball that he/she wants to look at. The radio button offers 3 choices: Major League Baseball players over all-time (1871- 2016), those from the modern era (1960-1999), or more current players (2000-2016). After selecting any desired period, each of

the charts will dynamically refresh itself with the appropriate results. The viewer is encouraged via clues on the chart to explore the details of those results.

The **triggers** are easily identified in terms of the **events** and the **parameter changes** they trigger. The underlying data originates from the entire set of MLB Players throughout history and is dynamically refined based on what time period the viewer selects. when that player's debut year occurred.

In addition the dynamic data selection and wealth of tooltips, the most unusual **user interface events** are the Animated Horizontal Bar Chart and the Animated Donut Chart with Labels and Legends.

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 *schooID* 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.

"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 **Choropleth** was decided to be the most effective way to visualize the data by having the top states indicated by a darker reddish color, and numerical count of players. This design choice was much more intriguing to the viewer than using a simple bar chart. The darker the reddish color, was used to indicate the higher number of MLB players who played college baseball in that state.

The **color choice** was in the deep reddish range – darker red representing a higher number of MLB players that played college ball in that state, and beige or crème color 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 indicators.

“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 Acknowledgements
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