Lafayette College Senior Project 2020
User Manual

Cameron Zurmuhl **Joseph Teddick** **Matthew Gerber** **Xingwen Wei**

Project Background and Goal

Soccer Hotspots is a data visualization project intended to locate high schools of Division-One collegiate soccer players who have a proven record of success. Literature from authors such as Daniel Coyle suggests that talent isn't born, rather developed. In the football world, for example, Super Bowl MVP quarterbacks Nick Foles and Drew Brees both hail from Westlake High School in Austin, Texas, an acclaimed football powerhouse. This project applies similar logic to soccer.

Using roster data from 34 colleges and universities with rigorous admission standards across 11 Division-One conferences from 2006 onward, we compiled a web application that allows Lafayette College's Men's Soccer recruiters to strategically locate the players they need. By filtering players based on their league, college, position, number of years they were a starter (playing at least 50% of season games), and number of years they received an all-conference award, recruiters have great versatility in finding talent. We hope our software will help streamline the recruiting process for future seasons and supplement current practices.

Our visualization is made up of two components: a heat map and markers. Both components can be toggled by the user. Each marker represents a high school that had at least one player that met the query criteria. Each marker has a label for the number of players who attended that high school under the given criteria. When the map is zoomed out, markers will cluster. The heatmap is a good method for visualizing the concentration of players. The more players a high school has per query, the more red the area.

In addition, our project is scalable and future proof. After proper admin authentication is obtained, the user can choose to upload data in one of two ways. To upload many players at a time, the user can choose 3 files consisting of lists of players in what is called a “scraped data upload”. In order to upload one player at a time, the user can choose “manual data upload” and fill out the required fields by hand.

Disclaimer

Due to the nature of this project and application, there are inherent risks in making assumptions from the data visualizations. Those risks are outlined below.

Data Limitations

Due to the nature of the scraped data, our application currently includes approximately 25% of the colleges' roster data between the years 2006 and 2018. In order to display the correct location of the high school on the map, the name of the high school on the roster data must be exactly matched to the location data provided by the College Board and the government. In order to maintain 100% locations accuracy, the data is limited in this way. However, as described below, the user has the option to manually upload single players one at a time in the case that the location is known to be correct. This process bypasses our matching algorithm.

Unintended Consequences

In addition to our data limitation, we recognize that our data presents its own set of biases. Given that the 33 colleges we focus on are mostly private, expensive schools, our data favors soccer players who come from wealthy areas and can therefore afford to attend such prestigious colleges. This no doubt favors high schools in wealthier areas of the country and of the world. Recruiting solely based on the information gained from the use of this application has the potential to leave out talented high schools and players that are located in less fortunate areas. We urge the user to take this into account when recruiting using our software.

Home Page

Our home page has 3 main components designed to guide the workflow of the user: the selectors, the map, and the data table. Combined, these 3 elements help to support both the visualization and statistical gathering of the data. After choosing specific filters to query the data and subsequently pressing the submit button, the appropriate data will be displayed on the map. This data is also shown on the table below the map and can be sorted, searched, and exported as the user sees fit.

Selectors

There are 5 main selectors the user has to create a search:

- College League: Choose 1 or more college leagues from the dropdown list. The colleges that are part of the selected leagues will automatically be checked in the drop down list of colleges.
- College: Select the desired colleges from the drop down menu. This includes any colleges already selected through the college league option as well as any additional colleges selected.
- Position: Select any or all desired positions. This position correlates to the player's most recent position played.
- Number of Years Starter: Select a value ranging from 0 to '4+' in order to query players based on the number of years they started, or played in more than 50% of the games in a given season.
- Number of Years All Conference: Select a value ranging from 0 to '4+' in order to query players based on the number of years in which they received an all-conference award.

If the user wishes to only display high schools that produced 2 or more soccer players, they can choose to exclude schools with only one player using the checkbox next to the submit button. Once the user chooses the appropriate values from the selectors, the submit button can be pressed to display the results.

Note: Not selecting a value within a filter will default to all of the values being "selected". For example, selecting nothing and just hitting submit will return all players.

Map

After the submit button is pressed, the page will load the appropriate students onto the map. The map is made up of two primary icons: high schools and clusters. A high school is represented by a yellow schoolhouse (), with the number on each corresponding to the number of soccer

players from that high school. The cluster icons () represent a geographic grouping of high schools, with the number again representing the number of students in that region. The color of the cluster is representative of the relative amount of students in the cluster. As the user zooms in and out of the map, the clusters will expand and condense respectively. The user can automatically expand a cluster by clicking on the specific cluster. Additionally, clicking on a high school will display an info window with the high school name, location, and number of students. This will also update the data table to reflect only students from the high school. Other features of the map are described below:

Feature	Description
Map / Satellite	Toggle between a traditional map view and a satellite view.
Toggle Heat Map	Remove the markers and turn on the heat map to show approximate geographic hotspots. Note: this feature auto-toggles satellite view to on. The redness of an area is proportional to the concentration of players from that area.
Toggle Markers	Turn on and off the map markers.
Reset Table	Reset the table to display all of the players in the original query (i.e. before clusters or schools were clicked) and zoom out to fit all the data.
	View the map in full screen.

Data Table

After the data is loaded, the user can scroll down to view the information for the queried players in a table format. This table contains the following values for each player: years on roster (presented as a list), first name, last name, position, height in inches (most recent), weight in pounds (most recent), number of years starter, number of years all conference, league, college, hometown, state or country, and high school. As described above, the data in the table defaults to that returned from the initial query, but is modified on each click of a high school marker. This can be reset by clicking the “reset table” button on the map. Additionally, the user has the ability to search the data, sort by each column, and change the number of players being displayed. Also, clicking on a player’s row will take you to his most recent bio site. Some biolinks may be broken. Finally, by clicking “Export Table (CSV)” the user can download a CSV file of the currently displayed data to be opened in Excel or any preferred spreadsheet editing software.

Data Upload

To promote scalability as well as to future-proof the application, two methods of uploading are implemented: scraped data upload and individual upload. While scraped upload lets the user upload groups of students at a time via multiple spreadsheets of data, individual upload allows the user to quickly and automatically add data for one individual player per an individual roster year. Details for each upload process are provided below, as well as a method of restoring previous versions of the data in the case of an error. Note: only site superusers can access the data upload pages.

Upload Scraped Data

The scraped data upload process, available on the page titled "Upload Scraped Data", is intended to upload large amounts of data at a time. For example, after the newest year of player data becomes available annually, the user will want to upload the new data to keep the database current and correct.

The scraped data upload process consists of 3 files: Roster Data, Starter Data, and Accolade Data. This upload also takes in a unique "Description", which is used in the data restoration process. Each scraped data upload creates a new version of the database using the description as the name of the version. As described below, these versions can be viewed and restored. In order for the data uploaded to work properly, the files uploaded must follow the exact format (same column names) of the roster_data.csv, accolades.csv, and starter_data.csv files. As a note the id column in all files must exist, but can be left blank as this is just a Django requirement. The original files can be seen in the data folder of the repository and templates for these files can be found in the templates folder of the repository.

Upload Manual Data

In order to upload data for one player at a time, the user can choose "Upload Manual Data". The user can fill out the required attributes and submit the new player data.

Note: The manual upload process bipasses the high school matching schema implemented in the back end. This means that, in the case of our algorithm being unable to match a player's high school information to the roster data, manual upload can be used instead to display this player in the application. Since this is the case, it is important to double check before adding in manual data and ensure all the information you want is included. Before submitting a manual upload, you will be presented with a confirmation alter back which has your inputted information. You can click "OK" to add the player to the database, or "Cancel" to make changes. **Once data is confirmed, it cannot be changed.**

If you wish to see the unmatched players and the list of all high schools (with location info) to assist with manual data upload, you can find them in files called `unmatched_players.csv`

in the `templates` directory and `high_school_data.csv` in the `data` directory, respectively.

Restore Data

In the case of an error in the data upload process, the user can choose to restore previous versions of the database by choosing "Restore Data". By clicking on the drop down menu, the user can display the names of the previous versions of the database created in scraped data uploads. By selecting one and clicking "Restore", the user can revert the application to a previous version.

Note: The current version of the database in use is displayed in the top corner of the main and restore data pages, and made an initial backup to start. Restoring takes a decent amount of time, so be sure to refresh the restore page to see if the restore is finished loading. You will see a "loading…" status or a "loaded" status in the upper left hand corner of the screen. Only site superusers can access the restore page.

Admin Login

The Admin Login page is used to add super users to the site so they can access password-protected pages, such as the data uploads. The admin login site can also be used to change data in the database, but this is highly discouraged.

Acknowledgements

We would like to thank Coach Bohn for providing the project idea. We also want to thank Professor Joann Ordille of Lafayette College's computer science department for providing insightful advice for our project. We hope our application proves useful and can supplement current recruiting practices.