Ronin Chasan (rcc263), Jacob Wise (jmw555)
INFO3300 - Project 2 Written Description

Data Description

In our research, we had our eyes set on finding the quantity of professional sports players in the four major American sports leagues from each state in the US. Despite there being few professional athletes in the US compared to the rest of the population, we decided to only collect data on players from the current or most recent season for each league, as this still amounts to thousands of data points. Plus, limiting data to only current athletes will allow the user to draw more complete conclusions about the proportions of athletes to the rest of the population. In order to collect this data, we needed to find reputable sources for the origins of players from each league. Fortunately for us, a company called Sports Reference compiles significant amounts of data for the four leagues we sought to analyze, operating a website for each of them. We were able to locate data about the hometowns and home states for NBA, NFL, and NHL players from Basketball Reference[1], Pro Football Reference[2], and Hockey Reference[3], respectively. While adequate data for MLB players existed on Baseball Reference, further research led us to a dataset from Baseball Almanac[4] that was already organized into the quantity of MLB players from each state. Lastly, we decided to include population data for each state to set up the percentage visualizations we would make later on. This data was gathered using the most recent estimates from the US Census Bureau[5]. Unfortunately, the data from all of these sources was only available as either HTML tables or in a bulleted list format. As a result, we decided the most efficient way to gather the data would be to manually copy and enter the data into spreadsheets. From here, we imported our datasets and were able to begin filtering.

Each dataset represented its own set of challenges in order to extract the data we wanted and format it into a usable form. First, each of the datasets needed to have their formatting cleaned up. This included casting number values to Number type objects, trimming spaces from state names, and creating a uniform name for Washington, D.C. across all datasets. Additionally, the NBA and NHL datasets, which still consisted of individual player data, needed to be filtered to only active players. After this cleaning step, the NFL, MLB, and population datasets were ready to use, but the other two needed more work. Thus, we created a function that would count the number of players from each state and return it as one usable object.

To handle all of the data in an effective manner, we elected to attach data values to the properties of each state object. This meant that every state would have an accompanying total count of all athletes from that state, a count of the number of athletes from each league in that state, the population of that state, and finally each of these values (minus the population, naturally) divided by the population of that state. This created a set of statewide percentages representing the actual portion of that state's population that is a professional athlete in each of these leagues. First, however, we needed to import map data from D3.topojson for US states, state borders, and the nation as a whole. Finally, with our data sorted, filtered, cleaned, and mapped, and our map data prepared, we were ready to proceed.

Ronin Chasan (rcc263), Jacob Wise (jmw555)
INFO3300 - Project 2 Written Description

Data Sources:

1. Basketball Reference
   - https://www.basketball-reference.com/friv/birthplaces.fcgi
2. Pro Football Reference
   - https://www.pro-football-reference.com/friv/birthplaces.htm
3. Hockey Reference
   - https://www.hockey-reference.com/friv/birthplaces.cgi
4. Baseball Almanac
   - https://www.baseball-almanac.com/players/birthplace.php?y=2021
5. US Census Bureau
   - https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/national/totals/nst-est2019-alldata.csv

Design Rationale

When working with quantitative data such as athlete counts and population percentages, it's important to map it to visual elements in a way that allows them to be compared to one another. We want our users to be able to look at the data in two ways. For each league, they can compare the counts or percentages of each state. For each state, they can compare the athlete count of each league. These affordances ultimately fueled our design decision for the two visualizations.

It was clear from the beginning that our first visualization would be a map. The marks on the map are the states, whose sizes correspond to their actual geographical areas. Even though varying size of an element (e.g. radius of a circle, length of a line) is often useful for visualizing and comparing quantitative data, the option in this case would be the area of the states, which isn't very intuitive and would lead to a mess of a map. We chose to use color hue instead, which let us display the data on a linear scale in an intuitive and appealing way. The color legend maps the color scale to the data itself, so the user can quickly discover where each state generally lies on the scale. There are some tradeoffs in using color hue - the user can't view individual observations, and can't pinpoint the exact count or percentage just from looking at a state's colors - but we decided it was ultimately the best way to allow the user to gain a complete overview, or get the "bigger picture," from one quick glance at our map. The hovering and on click functionalities let them explore individual states further. For each league, and the total across all leagues, a representative color was picked. Our linear color scales vary states from lightgrey to this color, so the more saturated the color, the more athletes or a higher proportion of athletes from that state.

When it came to choosing visual channels, the graph was almost the opposite of the map. Rather than showing all states and varying color for one league, the graph shows only one state, and that state's information for each league. In terms of scale, the difference in athlete count between two leagues is proportionate to the difference in population proportion. This allowed us to put both scales on the graph, one on each side, and use rectangles in a bar graph to show the count/proportion for each league.

Although each bar has a different color, this is only to indicate league (these correspond to the colors in the map). Bar height was the visual channel used to represent the data. The taller the bar, the more athletes from that state in that league, and the higher proportion of the population in that league.

Interactivity Rationale

Our interactivity design decisions were based on the affordances we aimed to grant our users. The two main pieces of data conveyed in the interactive map are athlete count and population percentage, so buttons to toggle between these features were immediately included. The map should also allow users to explore not only the totals for each state across all leagues, but also each league individually. Similar to the count/percent toggle, buttons for each league and the total were added for users to easily navigate between. Changing the button color and the mouse to a hand when the buttons are hovered, as well as the color and border when selected, indicates to the users that they can switch amongst them, and which is currently displayed. This mouse change is used to indicate clickability throughout the project. Sidebar text also aids in guiding the user towards toggling the map setup and league.

The sidebar buttons are only a small portion of the map's interactive features. The states' color only tells the user so much; when they hover the map, they can explore each state further. The hovered state gets a border, and a text box appears with the relevant info for that display setting (which can be seen in the sidebar). The hover info guide and border make it very clear to the user what the mouseover does. The mouse style indicates states can be clicked, which is how we connected the two visualizations. We wanted our users to be able to adjust the bar graph to look at an individual state. The best way to do this, avoiding adding any unnecessary visual elements, was to connect the graph with the map. Clicking on a state adds a thick grey border, changes the title of the graph to the state's name, and adjusts the bars and axes scales using animated transitions.

Our last major interactive feature is the brush functionality on the map's color scale. Hovering the scale turns the mouse into a brush tool, letting the user discover that they can drag and move the brush element to select a certain count or percent range to examine. One of our goals is to allow users to see where states fall on the scale for each league; the brush functionality lets them select a region and examine, for instance, which states fall at the low and high ends of each league count or population proportion. When deciding how to best convey which states are selected by the brush on the map, we ran through a few ideas. First we thought about adjusting their fill, but that would interfere with the color scale's ability to convey differences in the data. We tried changing opacity, fading out the unselected states, but it generally just made them look lighter and could confuse users into thinking they fall lower on the scale. We settled on borders, using a different color and thickness than the click border to distinguish between the two. We decided this was the best way to indicate the selected states without interfering with the rest of the information conveyed by the map. We specifically chose border colors and

patterns for different interactive features so that users can see both at once, for example if a state is both brushed and clicked. This final piece of interactivity allows the user to complete a full-coverage, in-depth analysis of how states differ by athlete count and population proportion across the four major sports leagues.

Story

Throughout history, some of the world's most famous celebrities have been athletes, with the likes of Muhammad Ali, Michael Jordan, and Tiger Woods, to name a few. Most sports fans often know plenty about athletes' professional careers, and even sometimes college, but rarely get glimpses of them at the high school level, with the exception of a few infrequent extraordinary prospects. Our visualizations aim to tell the story of athlete origins, particularly those of the Big Four American sports leagues (NBA, MLB, NFL, and NHL). It allows users to do a full comparison of athlete home states, showing the athlete count for each state by league, as well as the percentage of the population that are professional athletes in each league. Users can compare states to each other using the map, and leagues within each state with the bar graph below it. By doing this, one can gain a complete understanding of the regionality of sports, the importance of population, and the athletic powerhouses of this country.

Our visualizations shed light on several significant and surprising insights about athlete origins. For example, California, Texas, Florida, and New York are the four most populous states - by a solid margin. Yet, the Total Count map tells us New York is not even near the top four in professional athletes, falling below much smaller states like Ohio, Pennsylvania, and Louisiana. Toggling to percentages shows us one reason why this is the case; athlete density is heavily concentrated in the south. In fact, Louisiana, Mississippi, and Alabama all have athlete density percentages that are several times as much as that of New York. Even further, they do this almost solely with NFL athletes. The NFL is the biggest of the four leagues by far, and most states, as can be seen by the bar graph, have many more athletes in the NFL than each of the other leagues. MLB athletes are a little more spread out, with higher percentages in Georgia, Tennessee, and Oklahoma. The NHL, however, is the most unique. Over 40% of the league is made up of athletes from Minnesota, Michigan, and New York. Hockey is extremely regional, especially due to the cold conditions in which it is desirable to play. NHL athlete origins are concentrated in the Upper Midwest and Northeast, straying from every other sport. Minnesota and Alaska (which has an extremely small sample size) are the only states with more NHL players than NFL players. The NFL is simply too big for its impact on the distribution of total athlete origins to be significantly diminished by the other leagues.

With this data, athletic scouts can scale down their search to the regions that produce the most athletes for their sport, and that probably have highly-funded, highly-regarded athletic programs for their high schools and colleges. Aspiring athletes could gain a better understanding of the sports culture within each state. Fans could learn more about where their favorite athletes come from, and how geography may have contributed to their progression in their sport. Even aside from these uses, it's still an interesting, fun

way to learn more about professional sports for fans across the country. Many big sports fans, even if they don't plan to put this knowledge to a particular use, would find it interesting enough to be worthwhile.

Team Contributions

Ronin

- Data collection and manipulation for NFL, NHL
- Sidebar display
- Choropleth and scales setup
- Mapping data to states
- Mouseover interactions for map and for graph
- Connecting graph to map on click

Jacob

- Data collection and manipulation for NBA, MLB
- Sidebar button interactivity
- Map color legend
- Brush functionality
- Graph setup
- Reset button

Time Spent

Over the course of this project, we would estimate that we combined spent a total of roughly thirty hours researching, collecting data, coding, and writing between the two of us. Most of our time was spent building out the interactive features of our visualizations, as well as connecting different components (e.g. making buttons that would perform multiple tasks) and fixing bugs. Towards the end of the project, we also focused a lot of effort on styling the user interface to enhance interaction.