**Plots for Shots: Using R’s Shiny App and Culturally Relevant Data to Enhance Statistics and Data Science Learning**

Immanuel Williams

California Polytechnic State University

GATO365 Learning Center

Gracie Briggs

California Polytechnic State University

Joshua Havstad

California Polytechnic State University

## Abstract

It is known that students often struggle with statistics and data science for a variety of reasons. This paper presents a method to reduce students’ apprehension of statistics and enhance their engagement with data science by integrating a Shiny package in R to visualize a culturally relevant dataset that compares Michael Jordan, Kobe Bryant, and LeBron James. A study was conducted to determine if a classroom activity that implemented the Shiny app and culturally relevant data had an impact on students' self-reported understandings of statistics and data science concepts, as well as graphics. The study’s results showed that students, regardless of sports preference, were positively impacted by the proposed Shiny package in R. Students' perception of their understanding and feelings towards statistics, data science, and graphics also increased. This study implies that more research needs to be done to clearly determine if and how using a culturally relevant dataset increases students' understanding of statistics and data science through introductory statistics and data science courses.

**Keywords**

Visualizations, Shiny app, R, classroom activity, CRD, teaching relevant data in the classroom, learning statistics, data science education

## 1. Introduction

Some students can face an uphill battle when they are first introduced to statistics and data science. Statistical concepts can be unfamiliar, daunting, and unrelatable, which can reduce students’ likelihood to succeed and comprehend the material. Knowing this, the job of an instructor is to break down possible barriers that students may encounter during their acquisition of skills. These skills can, in turn, teach them how to become both excellent consumers and producers of statistical analyses as well as proficient in data science methodologies. By removing these unnecessary obstacles, instructors can go beyond merely presenting statistics to students, and instead, they can guide students to appreciate and utilize statistics and data science. In other words, it is better to encourage students to learn statistics and data science by incorporating familiar or interesting material rather than using examples that are designed only to teach a specific statistical methodology. Studies have shown that the use of real and relevant data sets have an association with increased understanding of material, greater memory retention of material, and deeper interest, motivation, and engagement levels amongst students (Neumann et al., 2013).

Reaching students from their unique perspectives is especially imperative when teaching students of marginalized populations. An instructor may need to adopt a culturally responsive framework to ensure that all students in the classroom can understand and relate to the material being taught. Incorporating culturally diverse content can lead to students having higher interest in the material and greater understanding of the content which can result in students performing better in the course (Byrd 2016; Brown et al. 2018). This is an essential component of culturally responsive teaching (Gay 2000; Gay 2002; Ladson-Billings 1995) – teaching through the lens of the students’ backgrounds and understanding of the world (Gay 2010). Through familiarized content, students can more readily learn new concepts (Morell & Duncan-Andrade 2002). This framework needs to be adopted within statistics and data science courses to increase engagement within STEM-related fields.

Culturally Relevant Datasets (CRD) are the combination of statistics and data within culture (Weiland & Williams, 2023). A CRD is defined as data that either is relevant and popular among students or relates to a social justice issue. These datasets may come from current events, pop culture, or any topic that is relevant to or interests students. There are three major benefits of using CRD in the classroom. First, students are more likely to feel connected to the material and therefore can learn more efficiently because of increased brain stimulation. Second, as students associate statistical concepts with their own experiences, their retention may increase. Third, students gain the ability to observe how statistics and data science can be applied outside the classroom. The benefits of using CRD, especially in an introductory statistics course, increases the likelihood of all students gaining statistical literacy skills and understanding data science principles due to a new understanding of the relevancy behind the statistical concepts. This idea is supported by a study in which real-life data was used in a statistics classroom. The students reported on many themes such as relevance, interest, learning, motivation, engagement, and understanding. Relevance was the most cited theme, and one student commented that real-life data “helps you realize how important [statistics] is and how relevant it is to everything that you do” (Neumann et al., 2013).

When looking into data at the university level, an example of a CRD would be a dataset that involves college admission. More specifically, variables such as SAT/ACT scores, admission, race, and gender would intrigue students in introductory statistics courses because of their recent experience during their application process. Another example of a CRD would be to analyze a dataset that investigates the birth weight of babies, that may or may not participate in a prenatal program which is meant to help low-income women ensure a certain birth weight for their children. Though this dataset would not directly relate with most students in high school or college, the impacts of certain variables on the baby’s birth weight - such as whether the mother smokes/drinks, the mother’s educational level, or the birth parent’s race - will draw awareness of the ramifications of these concepts and causes. To be more specific, students may draw logical conclusions before visualizing or implementing a statistical technique, which could allow them to understand the statistical technique and the data science insights that will eventually be taught and implemented.

## 2. An Overview: The Example CRD

### 2.1 The Players

Michael Jordan (MJ) was raised in North Carolina, where he played three years of college basketball for the University of North Carolina at Chapel Hill before entering the National Basketball Association (NBA) draft (Morrissey, 2009). He was drafted in the first round and was picked third by the Chicago Bulls. Jordan is mostly known for his extraordinary athleticism and elite ability to score the ball, as he owns the highest career scoring average by any player who has played in the NBA. Perhaps the most iconic name and brand in all of basketball’s history, Jordan’s accolades speak for themselves, as he has 6 championships, 6 finals MVPs, 5 regular season MVPs, a defensive player of the year award, and many other accomplishments. He also inspired the generation of basketball players in the modern era.

Kobe Bryant (KB) was born in Philadelphia, Pennsylvania and decided to enter the NBA draft right after high school (ESPN, 2010). He was drafted 13th to the Charlotte Hornets but was traded on draft night to the Los Angeles Lakers where he would go on to spend the rest of his career. Bryant entered the NBA when Jordan had won his fourth championship and, at that time, was the youngest player to enter the NBA. He is often referred to as perhaps the most skilled player the NBA has ever seen, and he is known for his “killer instinct” on the court. This ultra-competitive nature is why he was nicknamed the “Black Mamba” and led the most popular franchise in the NBA to 5 championships. A superb player both offensively and defensively, Bryant’s accolades also include 2 finals MVPs, 1 regular season MVP, and 18 all-star appearances.

LeBron James (LJ) was another young superstar that began his career in the NBA right after high school. Born in Akron, Ohio, James was the first overall draft pick for the 2003-2004 season by the Cleveland Cavaliers (NBA Draft, 2003). Deemed “The Chosen One” by Sports Illustrated when he was still in high school, James was, and still is, compared to Michael Jordan regularly. Known for his ability to excel at many areas of basketball, James’ name is all over the record books, including being the all-time leading scorer in NBA history and 4th all-time in assists. Other accomplishments of his include 4 championships, 4 finals MVPs, 4 regular season MVPs, and 20 all-star appearances.

### 2.2 The Question at Hand

The CRD that is discussed in this paper was designed to answer one of the age-old questions in modern sports history, “Who is the Greatest Basketball Player of All Time?”. This question may resonate with students who follow sports, play basketball, or have a significant sports influence in their lives. Initially focusing on MJ and LJ, this debate often polarizes fans into two camps. However, the question then arises: should Kobe Bryant also be included in this discussion? By expanding our dataset to include the modern NBA legends of Michael Jordan, Kobe Bryant, and LeBron James, we aim to provide a more comprehensive analysis that could sway the debate or at least broaden the conversation. There will be parts of the paper that reference this case study (The GOAT Case Study), and it will be explained more in-depth later on, offering a unique perspective that challenges readers to consider whether the inclusion of Kobe Bryant could redefine the criteria for what makes the Greatest Basketball Player of All Time.

## 3. Implementation in the Classroom

### 3.1 General Comparison

Table 1 highlights the physical attributes, accolades, and a few statistics for Jordan, Bryant, and James. A difference that many highlight in comparing these players is the number of championships won, in which Jordan currently leads with six. Equally important, every time Jordan appeared in the finals, not only did he win, but he also won the finals MVP. However, there are some statistics in which Jordan trails both Bryant and James, such as all-star appearances and points scored. It should be noted that some accolades, including MVPs, are voted on by members of the media or fans, and should not be taken as a perfectly accurate representation of the players’ abilities. Another point which gets brought up in debates about the players is peak performance versus sustained excellence, as Kobe and LeBron played at a superstar level longer than Jordan did.

#### Table 1 NBA Accolades

|  |  |  |  |
| --- | --- | --- | --- |
|  | Jordan | Bryant | James |
| Year entered the League | 21 | 18 | 18 |
| Height | 6 ft 6 in. (1.98 m) | 6 ft 6 in. (1.98 m) | 6 ft 8 in. (2.03 m) |
| Teams | Chicago Bulls  Washington Wizards | Los Angeles Lakers | Cleveland Cavaliers  Miami Heat  Los Angeles Lakers |
| First Year in NBA | 1984 | 1996 | 2003 |
| Main Positions | Shooting Guard | Shooting Guard | Small Forward  Power Forward |
| NBA Finals | 6 | 5 | 4 |
| NBA Finals MVPs | 6 | 2 | 4 |
| Final Appearances | 6 | 7 | 10 |
| Regular Season MVPs | 5 | 1 | 4 |
| NBA All-Star Game Selections | 14 | 18 | 20 |
| All-Star MVPs | 3 | 4 | 3 |
| Rookie of the Year | Yes | No | Yes |
| Number of Seasons | 15 | 20 | 21 |
| All-Defensive First Teams | 9 | 9 | 5 |
| Scoring Championships | 10 | 2 | 1 |
| Number of Points | 32292 | 33643 | 39771 |
| Rebounds | 6672 | 7047 | 11000 |
| Assists | 5633 | 6306 | 10764 |

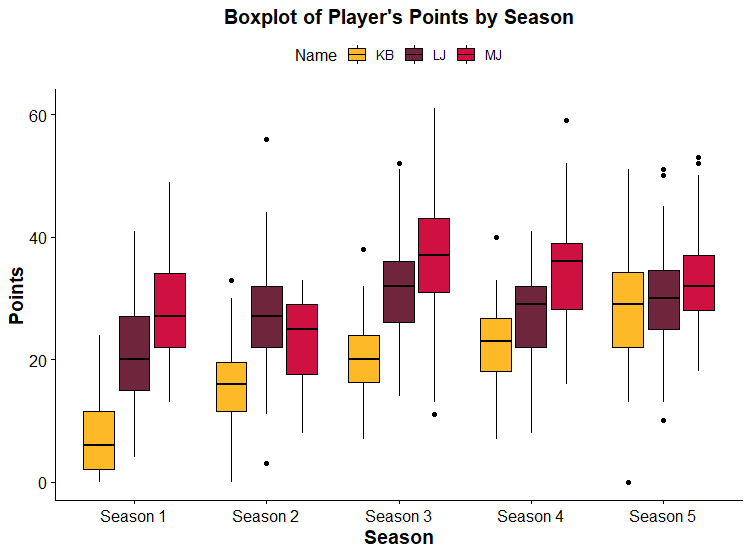
When comparing these players, it is imperative to highlight that James played a different position than Jordan and Bryant, which implies different tasks on the court. James’ main responsibilities as a small forward and power forward were to shoot, defend, and rebound. Jordan and Bryant were shooting guards so their main responsibilities were to shoot and to steal the ball (NBA Terms, 2018). This can explain how LeBron James has exceedingly more rebounds than the other two players. Though Jordan and Bryant played the same position, Bryant was also known to play small forward at times (NBA Terms, 2018). It is important to note that Kobe Bryant’s responsibility on the court would change depending on the position he played. This could have been the case because of the dynamics of the team rather than because of ability.

Though Bryant and James seem to outperform Jordan in the statistics shown in Table 2, it is important to remember that James and Bryant both played longer than Jordan. Analyzing overall attributes as well as accolades only provides a small fragment of their success which is why it is important to investigate their statistics for a more detailed comparison of the players.

### 3.2 Descriptive Statistics & Graphics

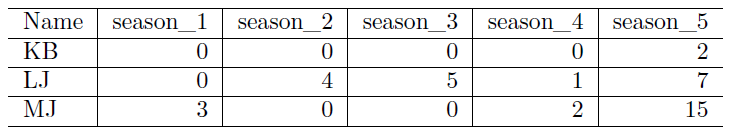
One way of introducing this content to an introductory statistics course is to discuss descriptive statistics (mean, median, variation, distribution, etc.) as well as graphics. A boxplot from The GOAT (Greatest of All Time) Case Study of the three players' points during their first five years can be seen in Figure 1. This figure shows Kobe’s rise to stardom, as he becomes more acclimated to the league his scoring ability becomes comparable to Jordan and James. This graphic also shows how when Jordan was injured in his second season, causing him to miss 64 games, it heavily affected his performance in terms of points. Year 2 is the only year when James' median points was larger than Jordan’s. Figure 1 allows for an educational opportunity to discuss boxplots and what they can tell us about the data in context of the CRD being used.

#### **Figure 1** Players during their First Season - Points in the NBA



Tables 2 and 3 below show the number of double-doubles[[1]](#footnote-1) and triple-doubles[[2]](#footnote-2), respectively, each player was able to achieve during their first five seasons. It should be apparent that Jordan started off well, though he was affected by an injury which consequently affected his performance. Since Bryant did not start once he entered the NBA, his ability to obtain either accolade was a tough task, but once he started, he was able to do well. The tables show how James’ game may be more well-rounded rather than focusing on one aspect of the game as compared to Jordan and Bryant. This could also be due to positional responsibility or team needs. Over 60% of his double-doubles were caused by him scoring more than ten points and obtaining more than ten defensive rebounds. Take into consideration that he is slightly taller than both Jordan and Bryant which could explain why he has more rebounds.

#### **Table 2** Double-Double of Players during their First Five Seasons in NBA **Table 3** Triple-Double of Players during their First Five Seasons in NBA

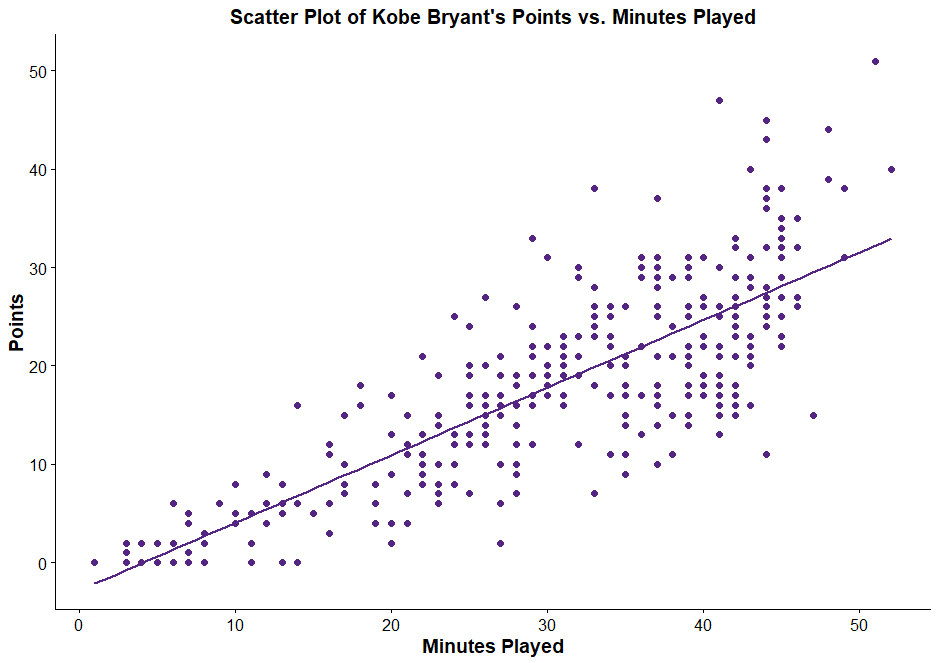


### 3.3 Other Teaching Methods

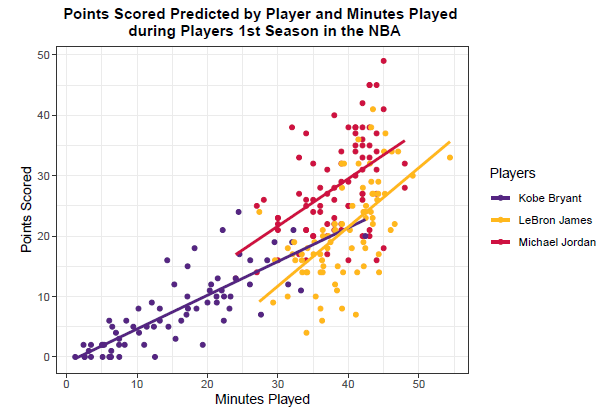
It would be best to describe the players in terms of their history and accolades, then show data (Table 1, then Tables 2 and 3) which validates facts that are known about the players. This could lead to questions such as *“What is the most likely reason Bryant did not achieve a double-double during his first two seasons in the league?”*. The GAISE College Report from 2016 advocates for these “engaging examples” in which students can begin with a research question and explore relevant data in order to come to a conclusion (Wood et al., 2017). Allowing students the freedom to think critically and investigate their own research questions will foster problem-solving skills and allow educators to teach statistical concepts in the context of the students’ curiosity.

An instructor may want to discuss simple linear regression, in which they may use *minutes played* as the explanatory variable and *points* as the response variable. This can be seen in Figure 2 that displays Kobe Bryant’s points during his first season. Though this analysis is logical in that more minutes played implies a higher likelihood of scoring more points, it provides an opportunity for students to reaffirm this line of reasoning visually and implement the simple linear regression analysis to verify it. However, being in the game longer has a different impact across the different players. In other words, implementing the multiple regression model with an interaction term in which the response variable is still *points* but the explanatory variables are now *player*, *minutes played*, and *player\*minutes played*, one can see that as James’ minutes increased, his scoring increased at a higher rate than Jordan and Bryant comparatively (see Figure 3). This fact could also be realized by evaluating Table 4 which is the multiple regression output.

#### **Figure 2** Kobe Bryant’s First Season - Predicting Points



#### **Figure 3** Players in their First Seasons and Minutes Played - Predicting Points



#### **Table 4** Multiple Regression on the players in their First Season and Minutes Played - Predicting Points

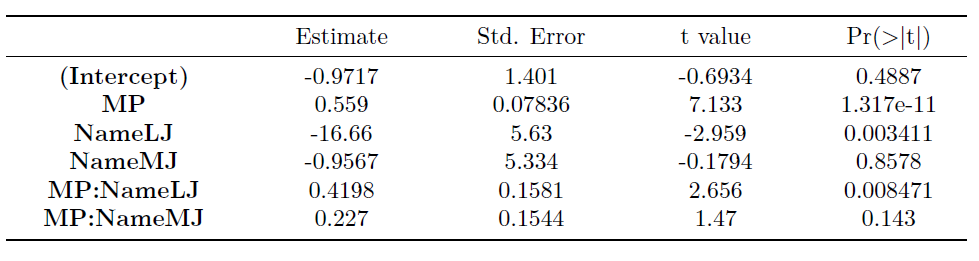
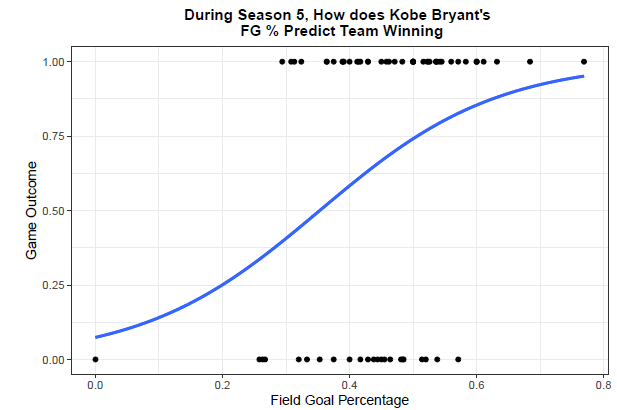


Table 4 shows that LeBron James has a different y-intercept than Michael Jordan and Kobe Bryant. This is a great opportunity to instruct students about when y-intercepts do not have interpretable or useful meanings. In this instance, the intercept is -0.9717, but a player cannot score negative points. Another conclusion we can come to from the interaction effects is that there is evidence that LeBron James scores points at a higher rate than Kobe Bryant when adjusting for minutes played, but there is no evidence that Kobe and Michael score at a different rate when adjusting for minutes played. This also provides the instructor an opportunity to discuss other regression models that are not linear, as it could be the case that the longer a player stays in the game, the rate of scoring points is potentially exponential.

One example of a non-linear model that could be introduced with this data is logistic regression. Logistic regression is often used when working with a categorical response variable. This could provide a valuable educational moment for the students, since they can learn when and how to use a fundamental statistical modeling technique. One of the scenarios that would be of interest is to determine if field goal percentage (shots made/shots taken) is associated with the game outcome. In other words, the response variable would be *game outcome* (win or loss) and the explanatory variable would be Bryant's *field goal percentage*. This logistic regression model can be seen in Figure 4. Students can also learn how to interpret the results of this logistic regression model in the context of the question. Figure 4 shows the students that Kobe Bryant’s field-goal percentage appears to be associated with the log-odds of Bryant’s team winning the game. This non-traditional example could spark further interest from the students on what additional factors impact the probability of a player winning a game.

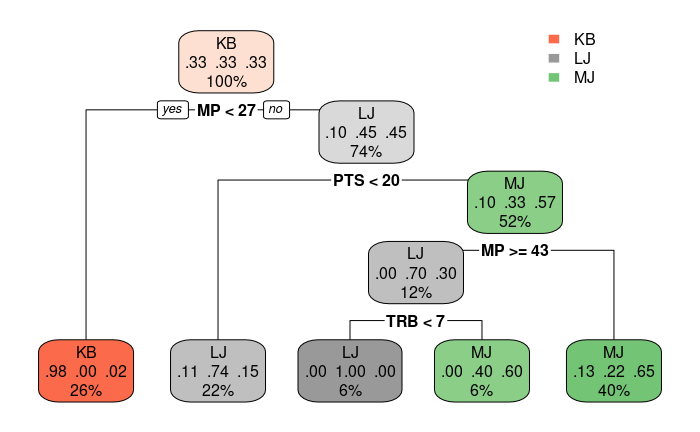
#### **Figure 4** Logistic Regression Prediction on Kobe Bryant’s Game Outcome Predicted by Field Goal Percentage



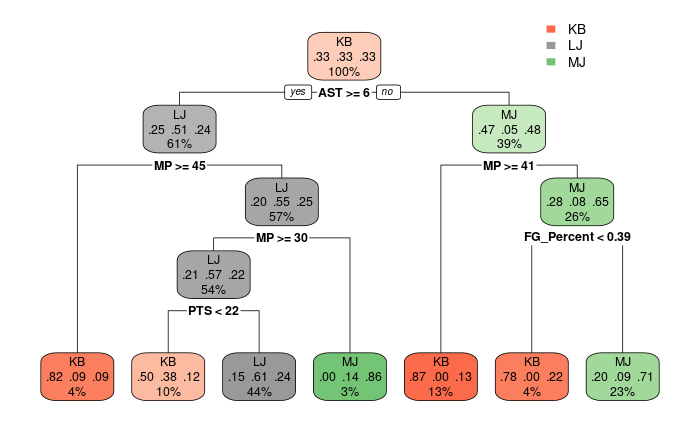
This data provides many useful teaching opportunities that will keep students engaged while ensuring students comprehension of the statistical concept being taught. This paper highlights examples using descriptive statistics and graphics, analysis of variance, and regression. However, an instructor of a more advanced statistics class could also incorporate this data in their classroom to introduce or implement techniques such as time series (points through all five seasons), principal component analysis (to evaluate the differences in the players), or classification and regression trees (to determine which type of performance contributes to a team winning).

This dataset is multifaceted in that it could also be used to introduce non-traditional graphic techniques such as the analysis of decision trees, making it an excellent tool for both statistics and data science education. As you can see in Figures 5 and 6 below, the players’ names are given on the tops of each data box. The top of each of the trees starts with 100% of our data and is then broken down by each decision (MP, PTS, TRB) below the trees. The categories are broken down by the splits and percentages as the trees branch further down. This approach not only enhances understanding of statistical analysis but also introduces students to fundamental data science concepts such as predictive modeling and classification. Notice that each of the numbers inside the nodes (boxes), the 3 numbers below the names, represent the proportion of instances in our dataset which satisfy the conditions. For example, in Figure 5, we can see that Kobe played less than 27 minutes in 98% of his games in his first season whereas LeBron played less than 27 minutes in 0% of his games in his first season. The percentage at the bottom of the nodes indicates the classification accuracy of the decision tree based on the conditions specified at that node. This practical application of decision trees in analyzing sports data serves as a bridge between theoretical knowledge and real-world data science applications, further enriching the educational experience.

#### **Figure 5** Season 1 Analysis using Decision Trees



#### **Figure 6** Season 2 Analysis using Decision Trees



## 4. Extracting and Transforming the Data

### 4.1 Cleaning and Transformation

Cleaning the data is not only valuable for the R Shiny app to run more efficiently, but it also provides a practical application of data science methodologies, allowing students to actively learn more about the data while it is right in front of them. While extracting and transforming the data, the student can learn an immense amount of knowledge about the history of the numbers, the variables, and where they came from. This information can then be further used to enhance their analytical skills and deepen their understanding of data science principles in addition to strengthening their analysis and understanding of the project.

### 4.2 Extracting the Data

R (R Core Team 2015), a statistical software, was used to scrape the example data from www.basketball-reference.com using packages in R called XML (Lang & CRAN Team. 2017) and httr (Wickham, 2016). Once the data was extracted from the website, a process of transforming and organizing the data was done in order to create graphics and perform statistical analyses. This process included removing unnecessary characters from the variables, creating new variables, and ensuring the variables from all three players were consistent (the type of data collected changed in the early 2000s). The code for this process is provided within the supplementary folder to this article. The data set includes Michael Jordan, Kobe Bryant, and LeBron James' first fifteen years in the NBA. The data only includes 15 seasons for each player to attempt to accurately illustrate the comparison and attempt to reduce the number of confounding factors. Though this data set only includes these three players' first 15 years in the NBA, more years and other players in the NBA (past or present) can be added by using the R code provided. This practice of data extraction and transformation is foundational in data science education, as it teaches students the critical steps of cleaning and preparing data for analysis.

There are 34 different variables in the data set which are further classified as being: descriptive, outcome, offense, defense, or both offense and defense which can be seen in Table A.1. This table provides a teacher who is not familiar with basketball further insight into how these variables relate to one another. The descriptive type of variable provides context of the actual game, the outcome type of variable describes the results of the game, whereas the three other classes signify three aspects of the game of basketball: offense, defense, and both offense and defense. The offense's main objective is to score points whereas the defense's main objective is to prevent points. However, the classification of a variable being both offense and defense is where the variable can occur on either side. This distinction between variable types is important because it provides students and teachers the ability to not only understand the data in a basketball context but also to apply data science techniques to extract insights, showcasing the versatility of data science in sports analytics.

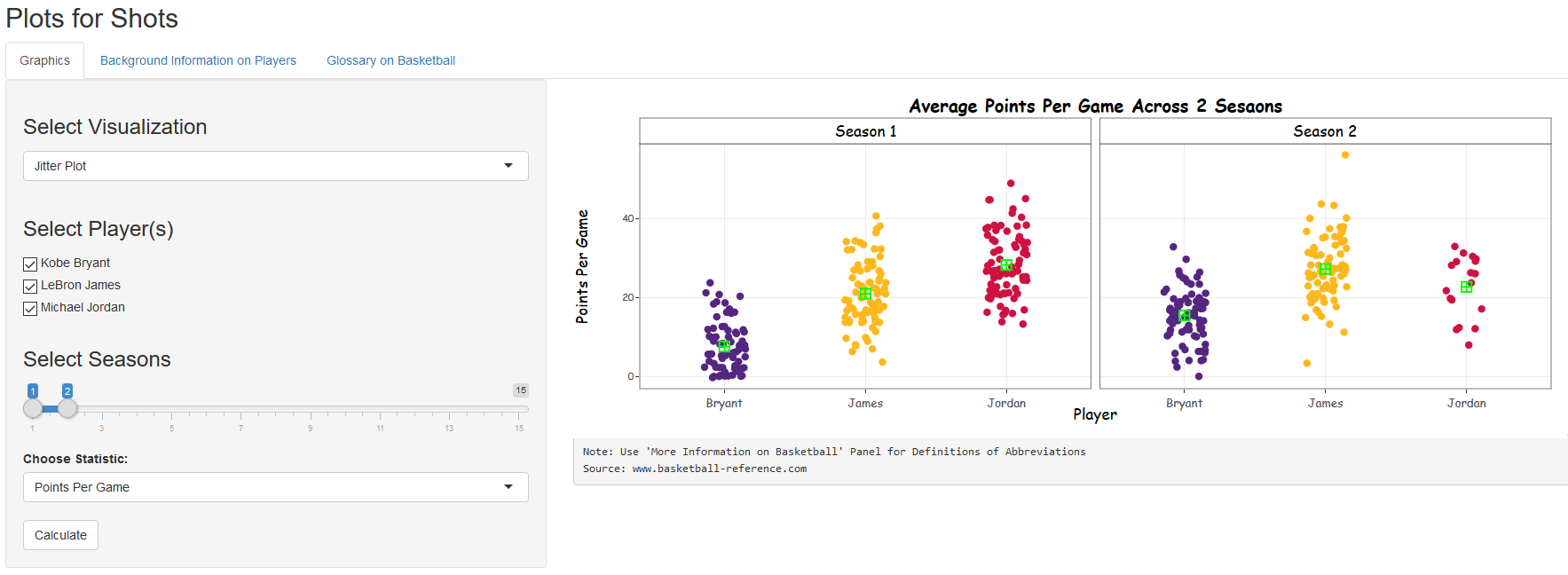
## 5. Plots For Shots

An R Shiny named Plots For Shots (PFS) was created as an application of the data which consists of 3 panels: graphics, background information, and glossary. The main functionality of this app action is located within the graphics panel which allows the user to manipulate widgets to view different types of plots. The background information and glossary serve as reference points for those unfamiliar with the NBA about various components in basketball. The proper usage of this R Shiny is to allow students to experience the data in such a way that they can strengthen their opinion or form a new one regarding these players based on data.

### 5.1 Graphics Panel

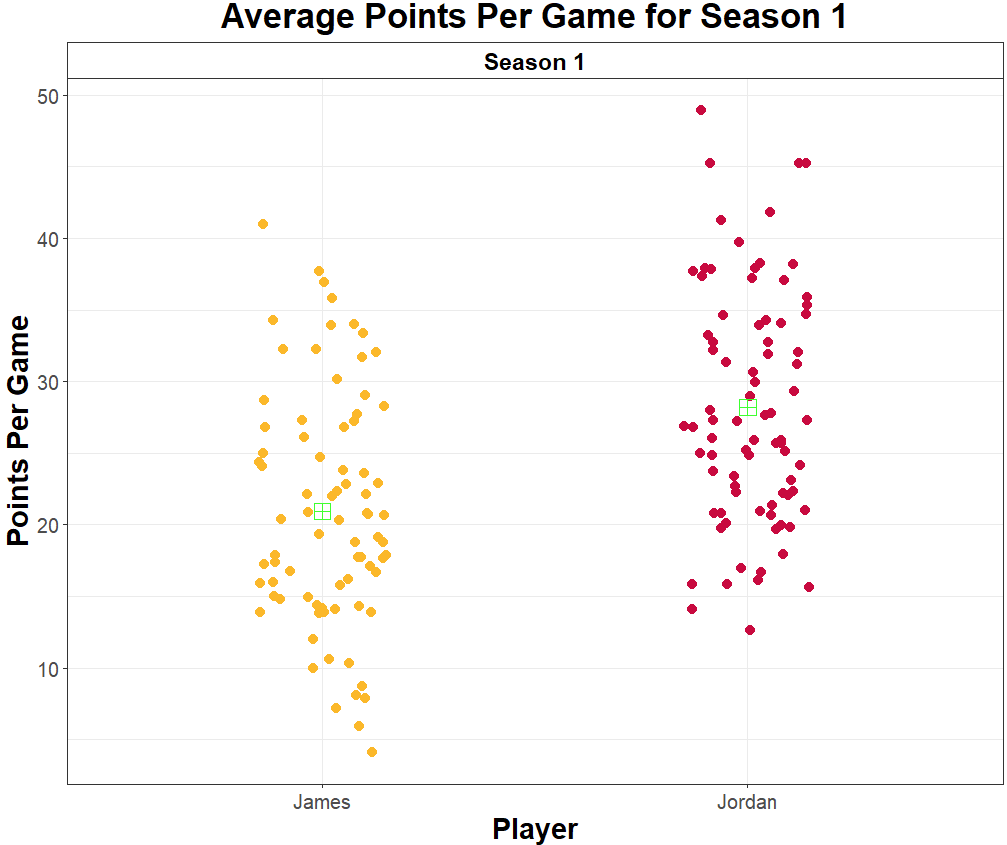
Once the user interacts with the PFS R Shiny, they will have the ability to select the type of graphics, which players, seasons, and statistics to compare, which can be seen in Figure 7. It is best if the user first selects a graphic and then a player. From here they can select the season and statistics, it is advised for the user to work with at most four seasons at a time because computer screens will distort the figures. The graphics that are created are based on only one statistic in that the relationships between NBA stats (i.e. points versus minutes played) is not permitted in the PFS R Shiny. Though it would have been advantageous to compare multiple statistics, the explanation required to comprehend the graphics would take away from the experience with the R Shiny.

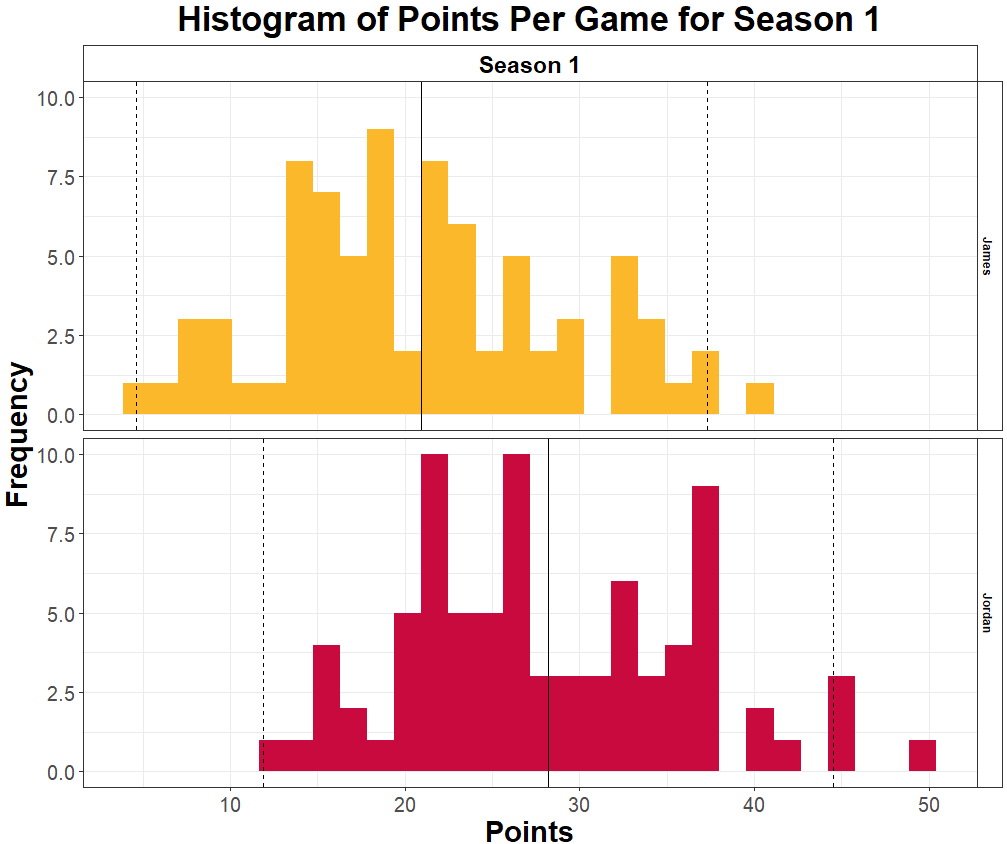
#### **Figure 7** Graphics Panel within PFS R Shiny

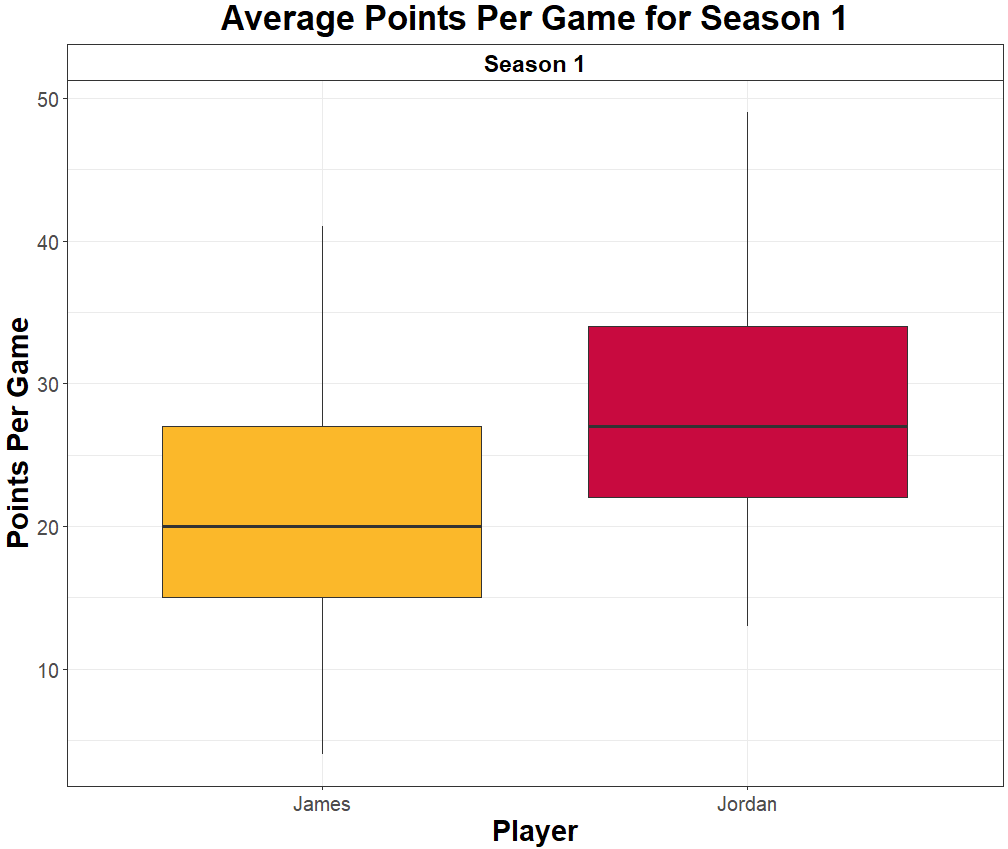


There were 3 different types of plots that the R Shiny allowed the user to select: the jitter plot, the histogram, and the boxplot. An example of each type of plot can be seen in Figure 8. Each graphic can be used to represent the same information, each one having a specific benefit. A jitter plot is a way to visualize data in terms of its distribution, then add random jitter to be able to recognize dense areas of values. The other graphics do not have the capacity to emphasize certain regions being more (or less) dense than others. This is an important feature to see in a graphic because it is a sign of consistency. If a number of data points are clustered in one area, it shows that the player does not deviate much from a particular performance. The boxplot provides a graphical representation of the five number summary while allowing a user to see a clear difference between players. A histogram is typically used to visualize the distribution of the data. Unlike the boxplot, the histogram has the ability to display the skewness within a graph.

#### **Figure 8** Jitter Plot, Boxplot, and Histogram within PFS R Shiny







Users have the ability to select offensive statistics such as: field goals, three-point field goals, points, offensive rebounds, and assists. Users also have the capacity to select from the following defensive statistics: turnovers, defensive rebounds, steals, and blocks. These were selected from the list of possible statistics from the website due to the fact they are easily explained to individuals that are not familiar with the sport. This is a great opportunity for students to visualize the many discrepancies between players’ statistics across various seasons.

### 5.2 Background Information

When comparing these players, it is important to note each player's background information, which can be illustrated in Table 1. One component of this panel is a table that holds the physical characteristics, accolades, and career statistics. The physical characteristics are crucial to allow the user to investigate because it allows the person to know how their physiques differ. Reviewing accolades shows how these players are comparable and can convince someone that these players should be analyzed in the same discussion. Career statistics such as rebounds, points, and assists are essential so that users can get a general sense on how these players compare in terms of statistics that are critical in the NBA. A timeline is also included so that the users can easily know when these stars played in the NBA, the overlap between these players, and which teams they played for. This is important because it shows how they played in different time periods of the NBA. There are other aspects that one could use to make an argument regarding comparing players, such as their teammates, opponents, coaches, or changes to rules. However these factors will add more variability that may further explain the difference between these players.

### 5.3 Glossary Panel

All the concepts discussed in the R Shiny are contained here to help those that might forget and not be familiar with this (Table 5). The user has the capacity to select the following from a dropdown menu to gather information that they may not be well-acquainted with: Offensive Statistics, Defensive Statistics, Other Statistics, Positions, and Graphics. The offensive and defensive statistics are based on the ones that are provided in the graphics panel, whereas, the other statistics option includes statistics such as free-throws and fouls so that the user is comfortable with other statistics that are important in the game of basketball. A definition is also provided for each player’s position within the position option so that the student can recollect what the main responsibilities are of each player while on the court. The background and glossary panels are paramount to utilizing the graphics panel in that it allows users to see which graphics are more beneficial to the data they are trying to interpret.

#### **Table 5** Glossary Panel within PFS R Shiny

### 5.4 Implementing PFS in the Classroom

This activity was centered around asking probing questions to get students to think about a problem from a statistical framework perspective. There were 142 students that made up three back-to-back classes at a California State University, where most were business majors. About 20% of the class took AP Statistics before coming to college. Each class was a 50 minute class where the PFS R Shiny was put on the schools server to be able to allow students to use it from their personal laptops. Before this classroom activity, the class had recently discussed graphics and they were familiar with center and spread. When implementing this activity, it is imperative to note that students were able to use the internet, giving them the ability to look up more information outside the R Shiny and understand the data set in a better way.

A 10 minute presentation was given to the students before they interacted with the R Shiny so that they could get excited about the material at hand. In order to drive the conversation, a series of questions were asked during the presentation. The first question asked was *“Who is the GOAT, Michael Jordan or LeBron James?”*. Though this question is subjective, it allows students to articulate their opinions or hypotheses regarding these players before interacting with the data. Then the students are asked *“Should Kobe Bryant be added to the discussion?”*. This became a yelling match in two out of the three courses because some of the students were extreme Laker fans and believe that Bryant should automatically be added to the conversation without considering the data. We found that this question made students that were not engaged initially start to perk up because everyone was excited to state their opinion. The third question was *“Can you back up your opinions and claims with data?”*. This question is probably the most important question in that it shows the importance of not only statistics but also data. We used this moment to further explain that drawing conclusions without data is very dangerous and we should take a look at the data before having such strong opinions. From here we started to look at the facts of a condensed version of the accolade table in the General Comparison Table 1. Then the last question was asked: *“What type of information would help compare these players in a more effective way?”.* Students started to say things like ‘players’ performance’ or ‘statistics like points and assists’ which allowed me to guide them to thinking about analyzing and visualizing the players in terms of game-by-game performance. We introduced the first season statistics of Jordan and James via the R Shiny so that they can not only get used to the web application and how it works, but also for them to compare Jordan and James without Bryant in the discussion. This opened the discussion about statistical concepts such as median in the boxplot and spread within the jitter plot. We also discussed how you can use the other tabs to get more information about the players and the NBA in general.

### 5.5 Group Activity & Questions

We then let the students break up into groups of two to four and answer the second question, *“Should Kobe Bryant be added to this discussion of who is the GOAT?”,* using the PFS R Shiny. To answer this question, the groups had to first decide on a claim (Is Kobe taken into account or not?) and then find graphics that support their claim. A Google Form was created for them to submit their answers. Since Plotly allows users to save graphics, the students were able to save their work on their laptops and submit their graphics within Google Forms. Along with submitting a graphic, they had to also provide two sentences that explain why they choose their graphic and how it answers the question at hand. This is a very critical component of this activity because it allows students to gain a further understanding of graphics and what information is being displayed within it. Not only do they start with a better understanding of the topic because of CRD, but this research showed that they were more intrigued after further analysis of the second question.

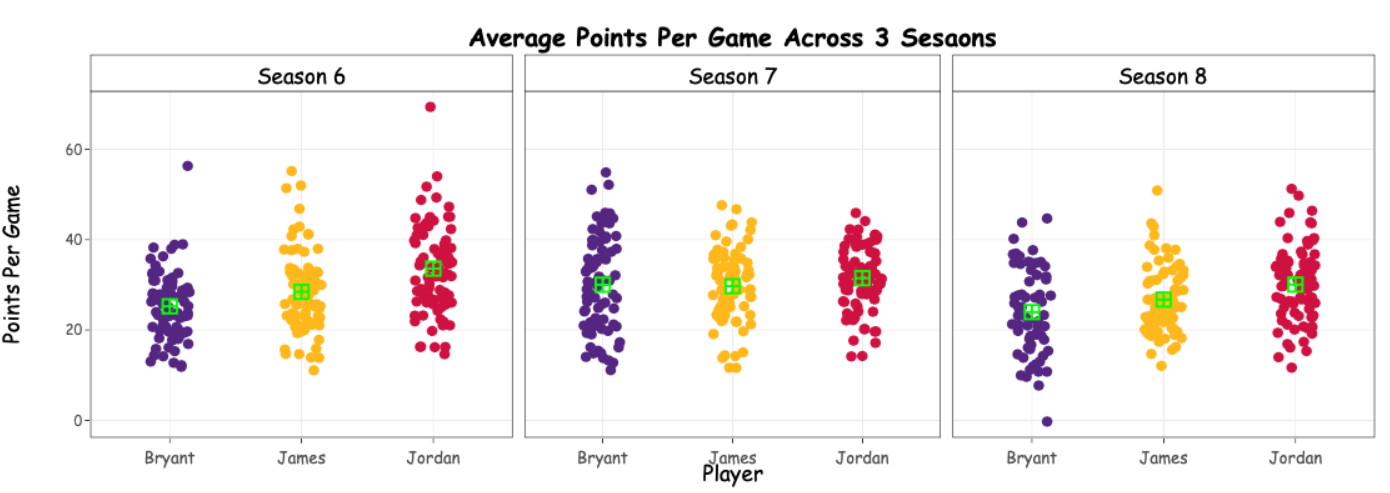
In validating their claims with visuals, students were able to acquire a deeper understanding of the data rather than just performing inferential tests. The capacity for the students to be able to manipulate up to three variables at one point, while providing explanations for their selection of the graphic, provided great information for the students and gave the instructor the ability to confirm that the students understood the learning objective. The GAISE College Report includes recommendations of strategies for teaching statistics which include emphasizing statistical literacy and developing statistical thinking, using real data, stressing conceptual understanding rather than mere knowledge of procedures, fostering active learning in the classroom, and using technology for developing conceptual understanding and analyzing data (Wood et al., 2017). The PFS activity was able to implement all of those strategies mentioned in the duration of one class, showing the wide range of educational benefits that this activity provides.

Figure 9 displays a graphic that was submitted via the Google Form by one group of students. Their claim was that Kobe Bryant should not be added to the discussion and they provided ample reasoning for this. The students first explained why they were selecting certain seasons:

“We chose these seasons because they are usually known as their ‘prime’.”

They then went on to say that Kobe Bryant was not performing at the same rate as the other two players during their prime years. Although the figure is not provided, their explanation for analyzing the assists during the same time period was eye opening. These types of observations within an introductory statistics course are incredibly beneficial to leveraging data within the classroom so that students get as much out of the course as possible.

#### **Figure 9** Selected Student Figure



"We chose to use assists [in seasons six through eight where these are usually their prime seasons] because they are just as important as making the points yourself and exhibit significant skill and teamwork in a player. In terms of assists, James has the greatest median in seasons six to seven. Bryant and Jordan are more comparable, with their medians being similar or the same [as in season seven]. However Bryant never has a greater median of assists than Jordan. Additionally, Bryant has greater variability than Jordan in seasons six and seven, showing inconsistency. Bryant was almost there in terms of comparison to James and Jordan, but never comes out on top and never beats Jordan in assists per game."

## 6. Results of Classroom Activity

### 6.1 Description of Survey

A pre-activity and a post-activity survey were administered to the three classrooms via Google Forms. Within this study there were two particular areas that are of interest: 1) Could PFS change student’s point of view on these legends? and 2) Could PFS have a positive impact on students' understanding and feelings towards statistics and graphics? The survey questions can be seen in Table 6. Question one on both surveys was used to create identifiers for each student so that their responses could be linked without directly knowing names. Questions nine and ten within both surveys were used to determine if visualizing these legends’ data changed their opinion about these players.

#### **Table 6** Survey Questions

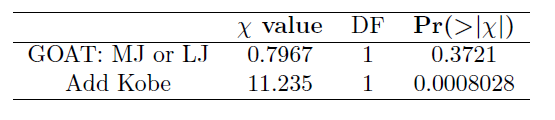
|  |  |  |
| --- | --- | --- |
| Questions | Pre Survey | Post Survey |
| Question 1 | What numeric month were you born in? Think of a number that is between 100 and 999 that you can remember. Attach these two numbers in the following manner "342". For example, Jamear was born in March and he chose 342 so his identifier number would be "03-342". (Identifier Question) Remember this number for the second survey. | Enter the month and three digit code you created before this activity. |
| Question 2 | Do you follow sports? | Your understanding of statistics increased during the Plots for Shots activity. |
| Question 3 | Do you follow basketball? | Your understanding of graphical representations of data increased during the Plots for Shots activity. |
| Question 4 | State gender. | Your feelings towards statistics are positive. |
| Question 5 | Your feelings towards statistics are positive. | Your understanding of statistics is strong. |
| Question 6 | Your understanding of statistics is strong. | Your feelings towards graphics of data (i.e. histograms, boxplots or jitter plots) is positive. |
| Question 7 | Your feelings towards graphics of data (i.e. histograms, boxplots or jitter plots) is positive. | Your understanding of graphics of data (i.e. histograms, boxplots or jitter plots) is strong. |
| Question 8 | Your understanding of graphics of data (i.e. histograms, boxplots or jitter plots) is strong. | Who do you think is the greatest of all time (GOAT) after this activity? |
| Question 9 | Who do you think is the greatest of all time (GOAT? | Do you think Kobe Bryant should be added to this discussion? |
| Question 10 | Do you think Kobe Bryant should be added to this discussion? | Did this activity change your perception about these players? |
|  |  | Comments for Plots for Shots |

Questions five to eight within the pre-activity survey and four to seven within the post-activity survey were meant to determine if PFS has an impact on students' understanding and feelings towards statistics and graphics. This concept was directly confirmed within the post-activity survey. It is important to gauge the students’ feelings or attitudes towards statistics, as research has shown that positive attitudes towards statistics can lead to an increase in both student performance and achievement (Gopal et al., 2018). Though a comparison has not been investigated regarding the difference between using a CRD and a non-CRD in the classroom, it is still important to measure the impact in using a certain dataset within a classroom. The students were also asked questions about whether they are interested in sports, if they followed basketball, as well as their gender. This information is important to know because students' understanding of sports, more specifically basketball, can have an impact on their experience with PFS. Questions regarding students' feelings and understanding were on a five point Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree) in which these values were converted to a numeric scale with (-2,-1,0,1,2). This was done so that the proper statistical techniques could be used to evaluate the data.

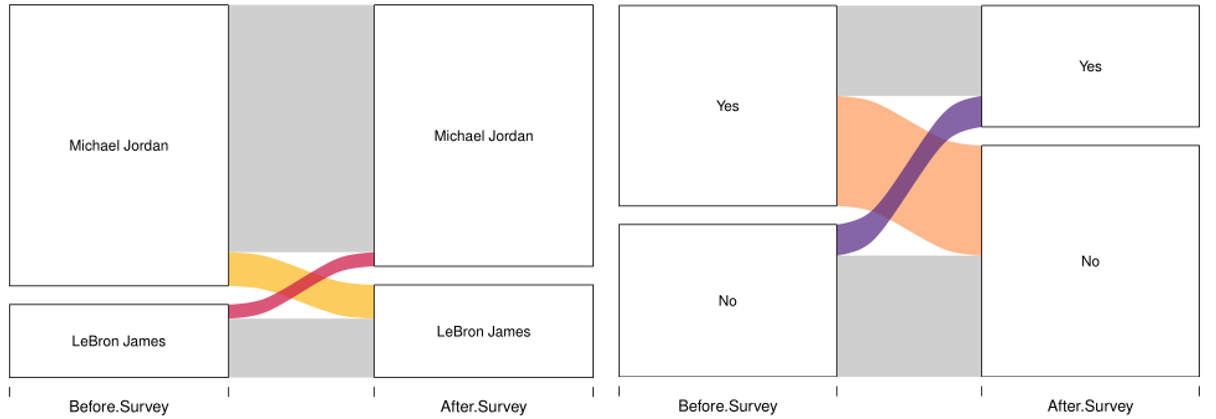
### 6.2 Change of Opinion

Figure 10 displays the results of who is GOAT; showing how opinions change regarding who is the GOAT: Jordan or James. Though a majority of students felt as though Jordan is the GOAT, more students changed their opinion after PFS than those that changed their opinion about James. However a Chi-Square test of independence seen in Table 7 indicated that PFS did not have an impact on students' opinion on who is the GOAT. One way of interpreting this result is that PFS confirmed students' opinion regarding this question.

#### **Table 7** Chi-Square Test of Independence on Students’ Opinions Changing



#### **Figure 10** Alluvial Plot on Students’ Opinions Changing

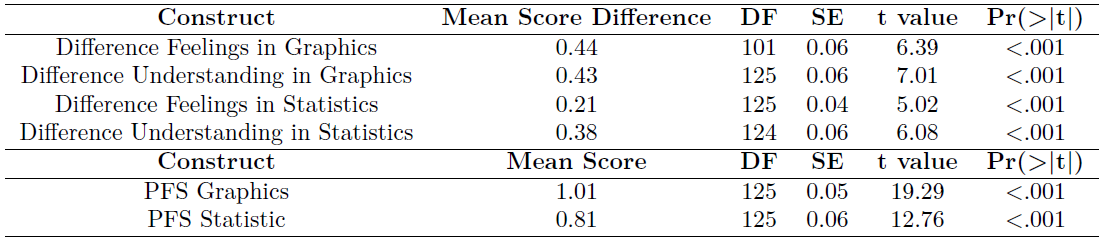


Though PFS could not change students’ opinions about who is the GOAT, Figure 10 illustrates that PFS influenced students to change their opinion about Kobe. More students changed their mind about Kobe being added to the discussion compared to those who originally believed that Kobe should be added after interacting with PFS. The Chi-Square test of independence indicates that PFS did have a significant impact on students' opinion about Kobe which can be seen in Table 7. This result provides an opportunity to appreciate how using real data to understand a statistical concept can have an impact on students' perception on concepts that are not traditionally discussed in the statistical classroom. In other words, visualizing data as a statistical concept must be taught and can also change students' beliefs on preconceived notions.

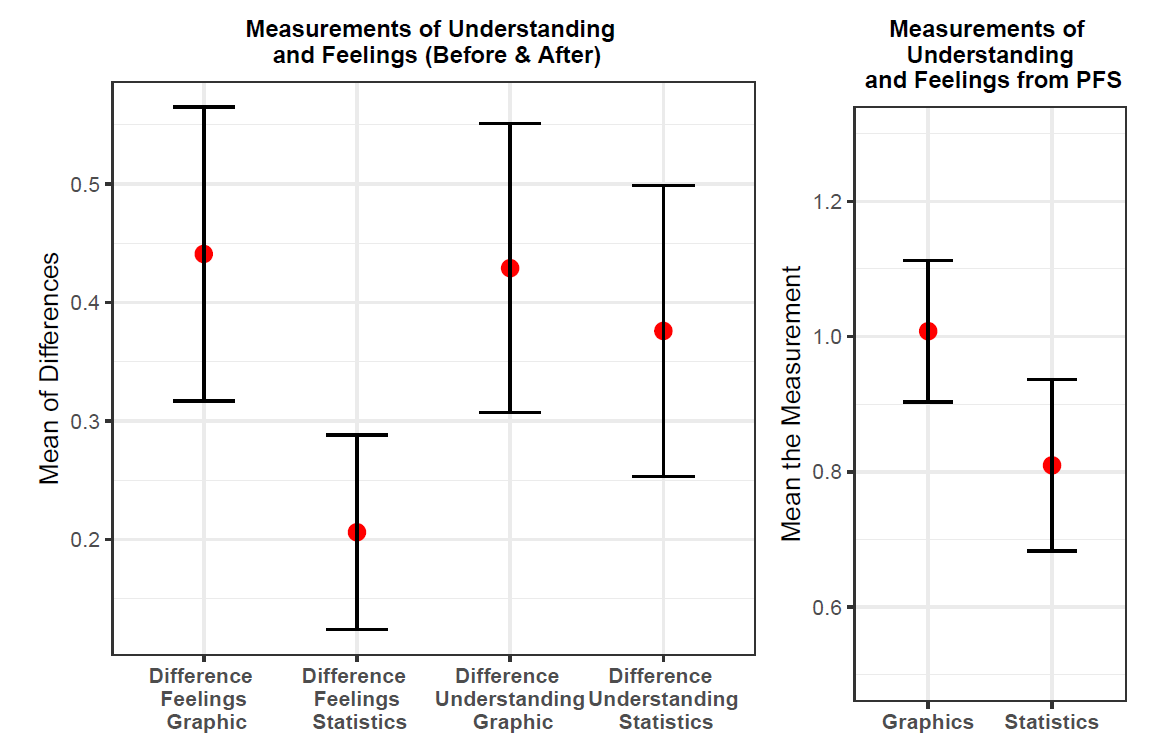
### 6.3 Perception and Understanding of Statistics and Graphics

This method of learning allowed students to see statistics in a more applicable manner. After interacting with PFS, students were positively impacted in terms of their feelings and understanding of graphics as well as their feelings and understanding of statistics. PFS seemed to impact the students' feelings towards graphics the most because the average difference between before and after PFS was 0.44 (p-value of <.001), as shown in Table 8 and Figure 11. Students' understanding of graphics’ average difference was 0.429 (post minus pre) where a paired t-test had a p-value of <0.001. These results imply that the PFS activity provided students with the information and ability to not only change their opinion about the players, but it also provided students with an improved understanding of graphics. Though the p-values for the students’ understanding and feeling towards statistics were significant after PFS, these differences between before and after were not as large. However, this activity was able to show that both the students’ perception and understanding of statistics increased almost similarly to their perception and understanding of graphics.

#### **Table 8** Understanding and Feelings Paired t-tests



#### **Figure 11** Plot of Understanding and Feelings



Questions two and three on the post-activity survey were meant to validate the difference in feelings and understanding of statistics and graphics which it was able to do. The tests from Table 8 confirm the idea that the PFS activity provides an educational benefit for students as well as sparks more student interest in the subject of statistics when working with culturally relevant data.

### 6.4 Student Feedback

Incorporating Culturally Relevant Data (CRD) into educational activities has proven to be an innovative approach to teaching students about data visualization. This method not only engages students but also enhances their understanding of statistical concepts through the lens of topics they find relatable and interesting. The feedback collected from students post-activity underscores the effectiveness of using CRD in the classroom. The overwhelming positive feedback from students highlights a significant increase in engagement and enjoyment during the learning process. Students expressed that the activity was "very fun," "engaging," and "a great visual for organizing data." Such responses indicate that CRD not only makes learning more enjoyable but also aids in the comprehension of visual representations of data.

The choice of data set—focusing on basketball, a subject of wide interest among students—played a crucial role in capturing their attention. As one student mentioned, "I enjoyed this activity as I am a big basketball fan. It helped me to understand visualizations of data more, and I had fun while doing it." This sentiment was echoed by many, demonstrating that the relevancy of the data set contributed significantly to the activity's success. Beyond engagement, students reported a deeper understanding of data visualization concepts. Comments such as "it helped me understand visual representations of data more" and "This activity helped me become more comfortable with reading stats graphs" suggest that CRD facilitates a more profound comprehension of statistical analysis and its applications.

The feedback from students serves as a powerful testament to the value of incorporating CRD into educational activities. Professors may not always grasp what is relevant to their students, but the positive reactions to this activity underscore the importance of understanding and utilizing CRD. By choosing data sets that resonate with students' interests and cultural backgrounds, educators can create a more inclusive, engaging, and effective learning environment. The use of CRD sets in teaching data visualization offers a promising avenue for enhancing student engagement and understanding. As demonstrated through student testimonials, such an approach not only makes learning more enjoyable but also facilitates a deeper understanding of complex concepts. Moving forward, it is crucial for educators to continue exploring and integrating CRD into their teaching strategies to cater to the diverse interests and cultural backgrounds of their students.

## 7. Conclusion

The modern NBA Legend dataset addresses an age-old question about who is the GOAT when comparing three basketball icons – Michael Jordan, LeBron James, and Kobe Bryant. This dataset is culturally relevant within the context of pop culture, and therefore students are better able to connect to the material. Additionally, this paper outlines a process for instructors to obtain similar sports data or other types of relevant data from the web to use in a classroom-based activity. Moreover, it emphasizes the integration of data science education by illustrating the use of data analysis and visualization techniques that are pivotal in the field. Overall, using a successful classroom activity as an example, we propose strategies to excite students about learning statistical concepts and data science methodologies through CRDs in their courses.

The results showed that neither the students’ gender nor sports preference mattered in terms of PFS impact on their experience. This is a critical finding because it implies that developing course content that speaks to marginalized populations can be beneficial for all. It is worth noting that not only did students’ opinions change regarding Kobe Bryant but also their self-reported understanding of statistics and data science increased, since the students were able to see statistics used in a relatable and engaging manner. This approach highlights the effectiveness of applying data science in educational settings, where real-world data makes the abstract concepts more tangible. Though the students’ feelings and perceived understanding do not directly represent change in ability, this outcome suggests that this is possible with PFS as a deeper understanding and more positive outlook of statistics and data science should naturally lead to proficiency in statistical concepts and methods. However, since this study only examines one CRD within a limited sample of students, these results are not representative of all student learning.

It is conceivable that using CRDs as course material can reduce students’ fear towards statistics and foster a more intuitive grasp of data science principles. Consequently, this may allow instructors to teach advanced material more efficiently, as students will be able to grasp the foundational concepts with less difficulty. With an easy implementation of the R Shiny and access to the retrieval of the original dataset, other statistics instructors could use this material as a foundation for creating their own CRD. This method also applies seamlessly to data science education, where students can be encouraged to analyze and interpret complex datasets. This method can be applied to other sports and popular athletes. Though web scraping can be a daunting task, there are many R functions and online forums that assist in this process.

In regards to future research, there are many questions and possible avenues to explore based on the findings of this study. First, what is and what is not a CRD needs to be more clearly defined. Many factors affect whether a CRD is interesting to a general student population, such as region, student demographics, and political climate. What is culturally relevant to one group of students may be completely removed from another. How can instructors acquire and create CRDs that will target their specific students? Additionally, integrating data science education into the development and analysis of CRDs could further personalize learning experiences and enhance engagement. It is worth studying which factors are most important when creating a CRD in order to help instructors take an intentional approach when developing course material. Another important line of inquiry is to examine how students react to datasets that are not culturally relevant, or non-CRDs. A future study could test not only if students prefer CRDs or non-CRDs but how that preference impacts performance. Furthermore, investigations to determine if CRDs increase student learning compared to non-CRDs and how data science education can be leveraged to enhance this effect would be beneficial to instructors designing their courses and developing textbooks.

## 8. References

Acosta, M. M., & Denham, A. R. (2018). Simulating Oppression: Digital Gaming, Race and the Education of African American Children. *The Urban Review*, 50(3), 345-362.

Arnholt, A. T. (2018). Using a Shiny app to teach the concept of power. *Teaching Statistics.*

Bartell, T., Wager, A., Edwards, A., Battey, D., Foote, M., & Spencer, J. (2017). Toward a framework for research linking equitable teaching with the standards for mathematical practice. *Journal for Research in Mathematics Education*, 48(1), 7-21.

Basketball Reference, (2019), Sports Direct Inc. Retrieved from <http://www.basketballreference.com/>

Brown, B. A., Boda, P., Lemmi, C., & Monroe, X. (2018). Moving Culturally Relevant Pedagogy From Theory to Practice: Exploring Teachers’ Application of Culturally Relevant Education in Science and Mathematics. *Urban Education*, 0042085918794802.

Brand, B. R., Glasson, G. E., & Green, A. M. (2006). Sociocultural factors influencing students' learning in science and mathematics: An analysis of the perspectives of African American students. *School Science and Mathematics*, *106*(5), 228-236.

Byrd, C. M. (2016). Does culturally relevant teaching work? An examination from student perspectives. *Sage Open*, 6(3), 2158244016660744.

Chance, B., & Rossman, A. (2006, July). Using simulation to teach and learn statistics. In *Proceedings of the Seventh International Conference on Teaching Statistics* (pp. 1-6). Voorburg, The Netherlands: International Statistical Institute.

Conners, F. A., McCown, S. M., & Roskos-Ewoldson, B. (1998). Unique challenges in teaching undergraduates statistics. *Teaching of Psychology*, 25(1), 40-42.

ESPN, (2010), Before they were stars: Kobe Bryant, Retrieved from <https://www.espn.com/nba/playoffs/2010/columns/story?page=beforetheywerestars-kobe-100601> (accessed 20 June 2019)

Ford, D. Y., & Kea, C. D. (2009). Creating culturally responsive instruction: For students' and teachers' sakes. *Focus on Exceptional Children*, 41(9), 1.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of teacher education*, *53*(2), 106-116.

Gay, G. (2010). *Culturally Responsive Teaching: Theory, Research, and Practice*. New York, NY: Teachers College Press.

Gopal, K., Salim, N. R., & Ayub, A. F. M. (2018). The influence of attitudes towards statistics on statistics engagement among undergraduate students in a Malaysian public university. *AIP Conference Proceedings*, *1974*(1). https://doi.org/10.1063/1.5041704

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3), 465-491.

Morrissey R., (2009), Chapter 1: Brooklyn, <https://www.chicagotribune.com/sports/bulls/chi-michael-jordan-chicago-bulls-chapter-1-story.html> (accessed 25 June 2019)

National Center for Education Statistics. (1986). IPEDS : Integrated Postsecondary Education Data System : less than two-year institutions. [Washington, D.C.?] :[National Center for Education Statistics]

NBA Draft, (2003), Prospect Profile: LeBron James, Retrieved from <http://archive.nba.com/draft2003/profiles/JamesLeBron.html> (accessed 23 June 2019)

NBA History, (2017), NBA Rules History, Retrieved from https://cdn.nba.net/nba-drupal-prod/nba-rules-changes-history.pdf, (accessed 16 June 2019)

NBA Terms, (2018), 250+ Basketball Terms all Coaches and Players Must Know, Retrieved from <https://www.basketballforcoaches.com/basketball-terms/> (accessed 16 June 2019)

Neumann, D. L., Hood, M., & Neumann, M. M. (2013). Using real-life data when teaching statistics : student perceptions of this strategy in an introductory statistics course. *Statistics Education Research Journal*, *12*(2), 59–70. https://doi.org/10.52041/serj.v12i2.304

Plotly Technologies Inc. (2015). Collaborative data science. Montreal, QC. <https://plot.ly>

Potter, G., Wong, J., Alcaraz, I., & Chi, P. (2016). Web application teaching tools for statistics using R and shiny. *Technology Innovations in Statistics Education*, *9*(1).

Rodríguez, A. J., & Kitchen, R. S. (Eds.). (2004). *Preparing mathematics and science teachers for diverse classrooms: Promising strategies for transformative pedagogy*. Routledge.

Sciutto M. J. (1995). Student-centered methods for decreasing anxiety and increasing interest level in undergraduate statistics courses. *Journal of Instructional Psychology*, 22, 277–280.

Villegas, A. M., & Lucas, T. (2002). Preparing culturally responsive teachers: Rethinking the curriculum. *Journal of Teacher Education*, 53(1), 20-32.

Weiland, T., & Williams, I. (2023). Culturally Relevant Data in Teaching Statistics and Data Science Courses. *Journal of Statistics and Data Science Education*, 1-14.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Williams, I. J., & Williams, K. K. (2016). Understanding summary statistics and graphical techniques to compare Michael Jordan versus LeBron James. *Teaching Statistics*, 38(3), 108111.

Williams, I. J., & Williams, K. K. (2018). Using an R shiny to enhance the learning experience of confidence intervals. *Teaching Statistics*, 40(1), 2428.

Wilson, S. G. (2013). The flipped class: A method to address the challenges of an undergraduate statistics course. *Teaching of Psychology*, 40(3), 193-199.

Wood, B. L., Mocko, M., Everson, M., Horton, N. J., & Velleman, P. (2017). Updated guidelines, updated curriculum: The GAISE College Report and introductory statistics for the modern student. *arXiv.Org*. https://doi.org/10.48550/arxiv.1705.09530

# Appendix

## Table A.1

|  |  |  |
| --- | --- | --- |
| Name of Variable | Description | Variable Type |
| DD | Double-Double | Both |
| GmSc | Game Score | Both |
| PF | Personal Fouls | Both |
| TRB | Total Rebounds | Both |
| TD | Triple-Double | Both |
| AST | Assists | Defensive |
| BLK | Blocks | Defensive |
| DRB | Defensive Rebounds | Defensive |
| STL | Steals | Defensive |
| TOV | Turnovers | Defensive |
| Age | Age of Player | Descriptive |
| Date | Date Game Played | Descriptive |
| G | Game Number | Descriptive |
| Game\_Location | Home or Away | Descriptive |
| GS | Games Started | Descriptive |
| MP | Minutes Played | Descriptive |
| Name | Name of Player | Descriptive |
| Opp | Opposing Team | Descriptive |
| Season | Season Number | Descriptive |
| Tm | Player's Team | Descriptive |
| Rk | Rank | Descriptive |
| 3P | Three Pointers | Offensive |
| 3P\_percent | Three Pointers Percentage | Offensive |
| 3PA | Three Pointers Attempted | Offensive |
| FG | Field Goals | Offensive |
| FG\_percent | Field Goals Percentage | Offensive |
| FGA | Field Goals Attempted | Offensive |
| FT | Free Throws | Offensive |
| FT\_percent | Free Throws Percentage | Offensive |
| FTA | Free Throws Attempted | Offensive |
| ORB | Offensive Rebounds | Offensive |
| PTS | Points | Offensive |
| Game\_Outcome | Win or Loss | Outcome |
| Point\_Margin | Point Difference | Outcome |

Author’s Information:

**First author's name:** Immanuel James Williams Ph.D.

**First author's address:** 1 Grand Avenue San Luis Obispo CA, 93401

**First author's E-mail address:** imwillia@calpoly.edu

**First author's phone number and fax number:** ​t -  ​​805.756.2709 & f - 805.756.2700

**Second author's name:** Gracie Marie Briggs

**Second author's E-mail address:** [graciembriggs@gmail.com](mailto:graciembriggs@gmail.com)

1. When a player reaches double figures (10 or more) in two of the five main statistical categories – points, rebounds, assists, steals and blocks. [↑](#footnote-ref-1)
2. When a player reaches double figures (10 or more) in three of the five main statistical categories – points, rebounds, assists, steals and blocks. [↑](#footnote-ref-2)