12-2018

Analytics Improving Professional Sports Today

Logan C. McDavid
lmcdavid@vols.utk.edu

Analytics Improving Professional Sports Today

CHANCELLOR HONORS THESIS

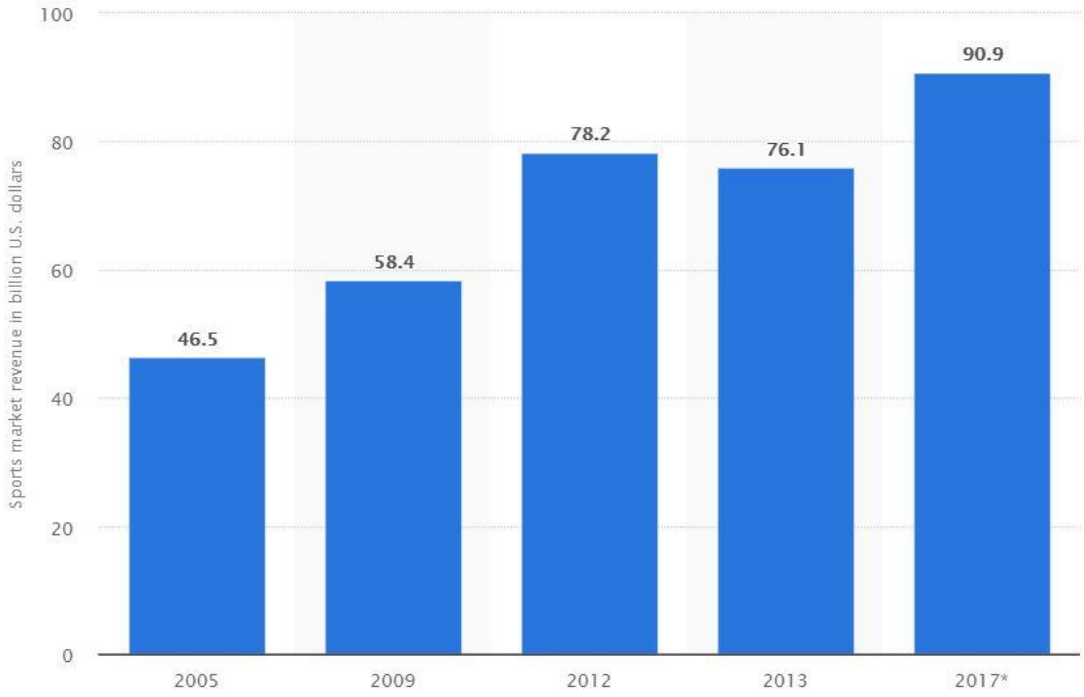
Logan McDavid
UNIVERSITY OF TENNESSEE |

Table of Contents

Introduction:

Around the world, professional sports continue to be one of the main forms of entertainment in many ways. Basketball, football, baseball, soccer, and many other sports helps connect players, coaches, and fans to one another while creating an inviting environment for everyone. The joy of playing, coaching, and watching has been an outlet that will continue to be there for people and provide entertainment. The sports market has grown over $12 billion dollars over the last 5 years, and this growth will continue as the professional games become more and more competitive. Figure 1 shows the growth of the sports market over the years, which helps give a better visualization for how this market is expanding (Global Sports Market Revenue 2005-2017 | Statistic).

Figure 1



Some may think only players and coaches can improve the quality of the game by making extravagant plays, developing entertaining sports media, and drawing up game-winning plays. But unknowingly, some one behind a computer screen messing with a keyboard can make more than a change to any sport. Looking at historical sports data allows analysts to analyze previous games, focus on current players performances, and make predictions for future games. Using statistics rather than reputation of players can create a competitive advantage for professional teams. This includes looking at minor and major details that could make remarkable changes to the game. Using tableau graphics, predictive analytics, and descriptive analytics gives analyst statistical evidence that will help achieve a better franchise for any sports team. Players are constantly practicing and striving for greatness, but practice can now be even more effective since statistics can show the necessary workouts and techniques to practice to fully develop a player's game. Sports analytics has helped improve professional sports by giving statistical evidence in many areas.

Current Process:

Pretty much every professional team has an analytical department or something relating to statistical analysis. The process starts with team staff gathering data from each individual player during their games; this data is sent to data experts to manipulate the data and run analyses to create valuable insight. Next, the experts will discuss the results with coaches, general managers, or other people on the staff to see what current players need to do to improve and what potential players will be the best addition to the team. The following information will go through this process more in-depth and will show examples of how sports analytics will be the catch to the future (Steinberg).

Data Manipulation:

I've been provided with statistics from the National Basketball Association and the Major League baseball organization that will be analyzed to create much needed value in the professional sports world. When it comes to using data in analytical software, there must be some proper tweaking and adjustments. In order to better understand the analysis, knowing the information that is being used for statistical graphics and insight is an important factor that will help get the most out of the results. Manipulating the data with specific techniques will ultimately make the analysis on the data run smoother. (1) First, the large set of data covering the National Basketball Association's players and their stats over the 2016-2017 season came from Adam Milliken, a University of Tennessee graduate student. Using the 2016-2017 data seemed more appropriate than the 2017-2018 data because that season is not yet over, and percentages could change during finals time. Given a player's name, some examples of their statistics are games played, minutes per game, field goal percentage, three-point percentage, turnovers, and much more for a total of 67 variables. This data is extremely large and messy, so cleaning the data is extremely necessary before performing any type of analysis. The columns consisted of individual stats and their specific team's stats. The analysis will be done on individual statistics, so we can determine a player's skill. The team columns were removed since they wouldn't be used. Next, there were some columns that contained NA's instead of 0's, so I had to make each NA = 0 in order to run the appropriate functions and not result in an NA error. Lastly, when running code in R Studio, having percentages and other symbols in the column names creates many errors for the functions, so I had to rename all the columns to make sure the appropriate output is given. Then, the NBA data is prepared for analysis. (2)

Next, the Major League Baseball data set contains over 100,000 baseball players from 1870-2017. This data comes from SeanLahman.com, who has been collecting and updating this data for many years (Lahman). This data set contains the players' names and their individual statistics, such as: amount of at bats, hits, stolen bases, RBI's, and many other valuable information. When manipulating this data, there were columns that needed to be removed and columns that needed to be added. The dataset was missing one of the most important factors to baseball, batting average, which I calculated by doing the amount of hits divided by the amount of at bats for each player. This value gives a relative look at how good someone is hitting based on the amount of times they've been given a chance to actually hit. Next, there were four columns with no values for any players, so they were removed from the data set. Lastly, there were some random NA values in my calculated batting average column because some players had 0 at bats; after performing some cleaning steps, these NA's were replaced with 0's, so that we could continue our analysis. All the adjustments have been completed, and now the statistical analysis shouldn't output any errors or corrupt data.

Analytics in Basketball:

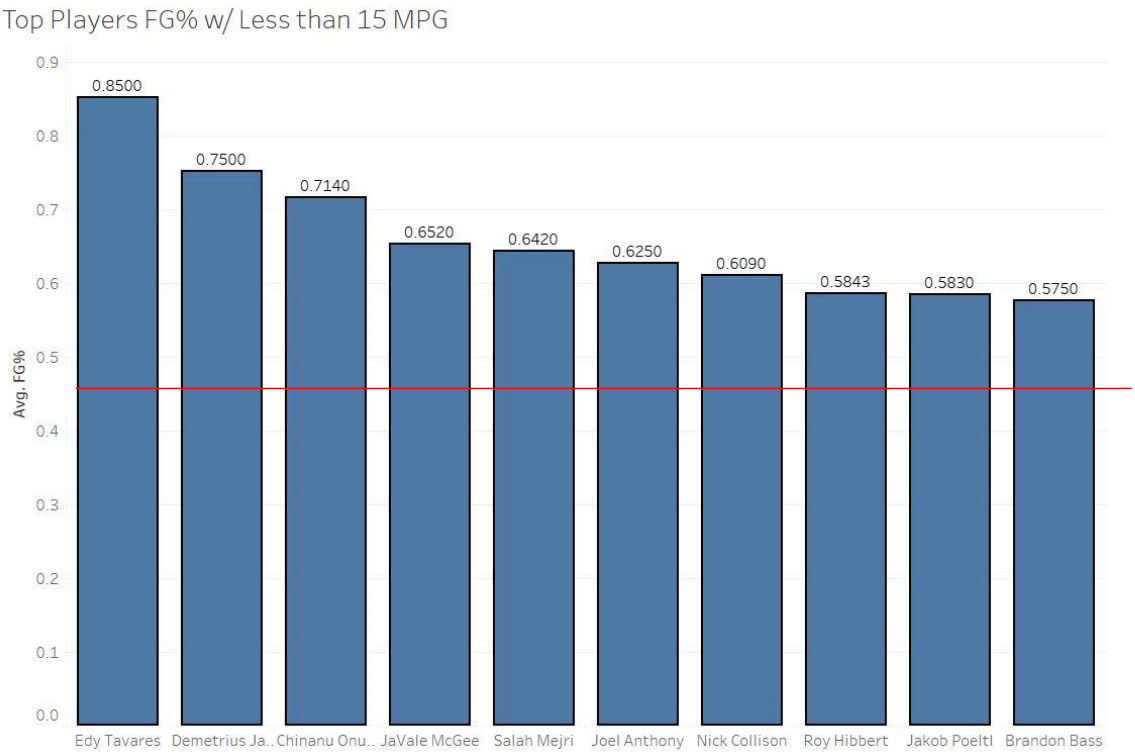
The National Basketball Association attracts millions of viewers for every game with its constant entertainment and popularity around the world. Expanding its horizons and recruiting players from other countries continues to intrigue new viewers and will continue to help the National Basketball Association grow. The most important part that the National Basketball Association needs to focus on is keeping all viewers by making the game more competitive. When it comes down to performing analytics, the question is what can be improved to help a specific team win more games?

Tableau:

Looking at the analytical side of making the sport more competitive, Tableau brings a major and simple help to analyst in the basketball world today. I've provided three Tableau graphics that will continue to help coaches and players keep the sport as competitive as it can be:

(1) Field goal percentage is one of the most crucial aspects in basketball and being able to look at the statistics gives us a better look at specific players. Coaches and players may think the amount of points someone scores in a game means they are a better player, but if someone shoots 50 times making only 15 shots (30%) and another player shoots 20 times and makes 10 shots (50%), the most effective player would be the player with the higher FG%. Shown in figure 2, each of these players play less than 15 minutes per game, but they have an extremely high field goal percentage. Seeing this information allows team analysts to make a case to coaches that these players may deserve even more playing time since they are shooting well above the average for NBA players.

Figure 2

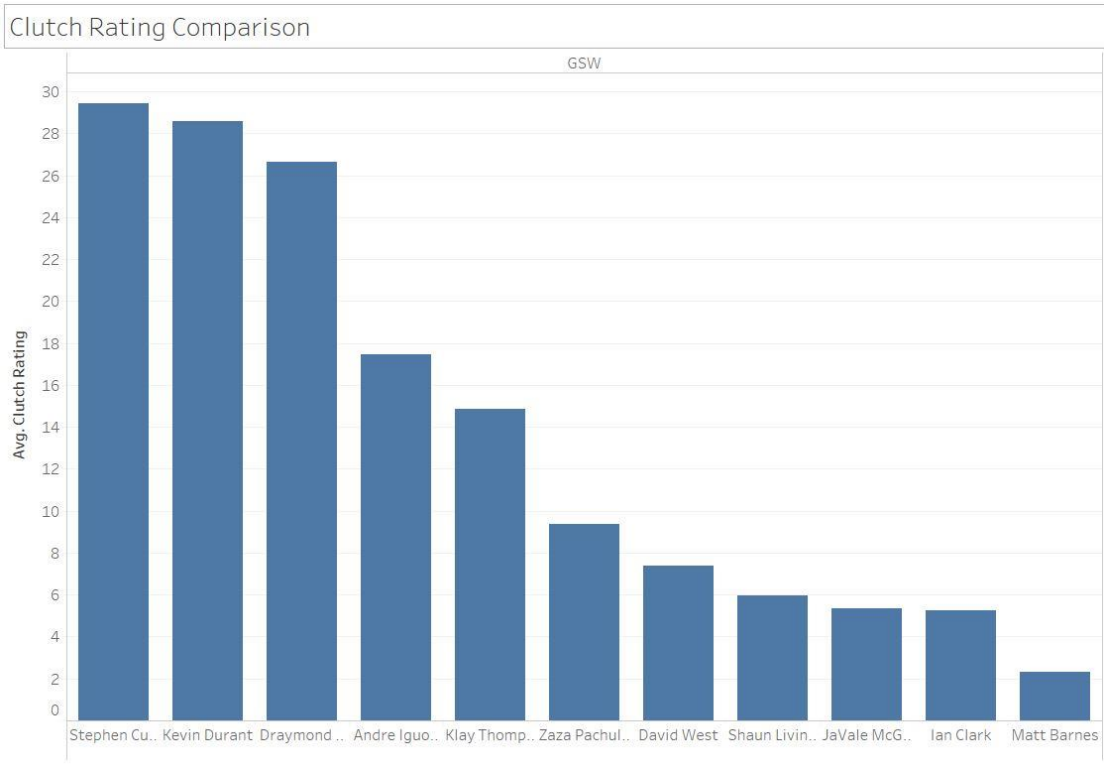


The red line depicts the average field goal percentage of NBA players

(2) Critical times cause for critical measures. There are plenty of teams that have phenomenal players looking to take that final shot to win the game. But statistically which player is more clutch in last minute games? A player's clutch rating has been calculated by multiplying the player's points produced per shot and the player impact estimate. The player's impact estimate takes into consideration assists, turnovers, points, rebounds and other factors of a player's games. Figure 3 shows a comparison of clutch ratings between the players on the Golden State Warriors (arguably the most dominant team in the NBA). Kevin Durant, Klay Thompson, Draymond Green, and Steph Curry all have the reputation to be given that last shot, but with statistics, a team should go into that moment having the highest confidence that the last second shot will be made. This analysis allows professional basketball teams the ability to

know who should be taking the shot in critical moments. In this case, Stephen Curry should be the one with the ball when a basket is needed and not the other players with relatively high clutch ratings. This could cause conflict with the coach and players, but having this statistical evidence gives the coach some support that is able to be seen by the players.

Figure 3



(3) In any basketball game, free throws play a major role in winning a basketball game. Shooting 50% from the free throw line rather than 100% makes a drastic difference when it comes down to the last minutes. We are able to identify which players are struggling in this area in order to benefit the team as a whole. In figure 4, I have organized the field goal percentages of multiple players on the Golden State Warriors in ascending order, so that we can see which players need the most practice in this area to make the team more competitive.

We can present this information to coaches and players to determine a practice plan for players lacking in this area. This tool gives players an outline on what they need to work on and how they can help the team more.

Figure 4

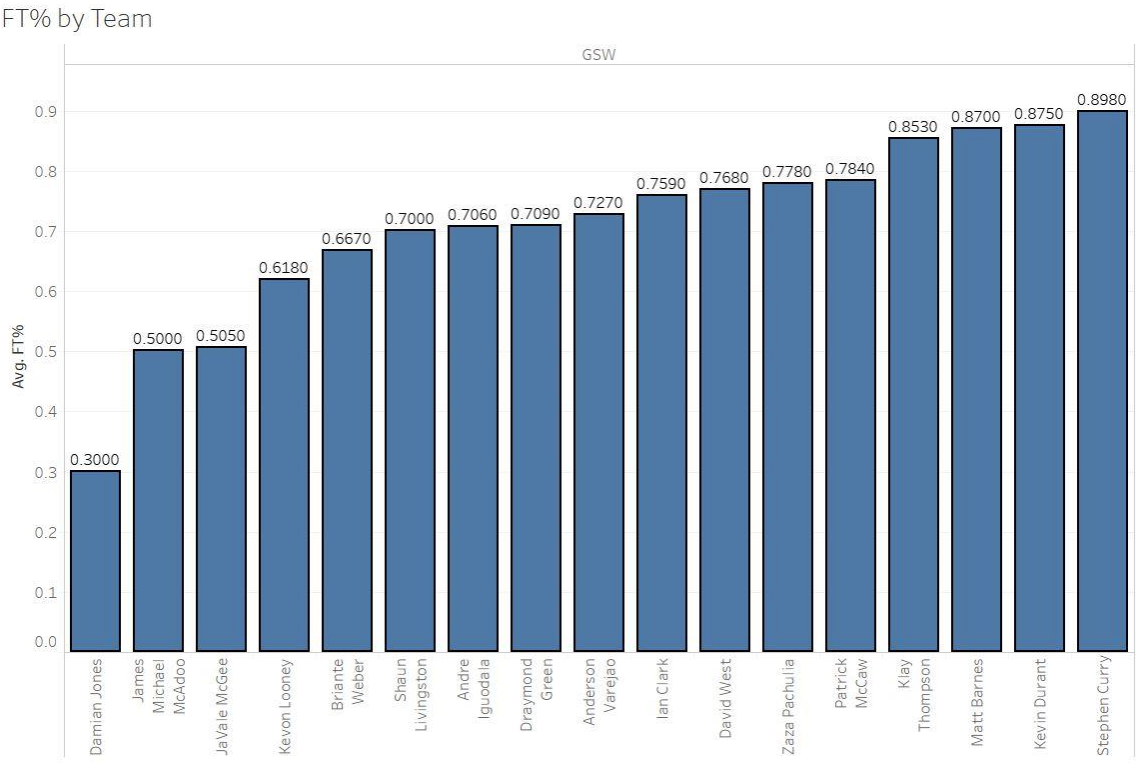


Tableau gives easy-to-read visualizations that provide evidence that will help coaches understand where they can make improvements, and to motivate players to work on their individual skills. Using this tool in the National Basketball Association would make the league more competitive and would overall make the games more interesting. Viewers love seeing high scoring games and clutch shots, so being able to find the players that make the best percentage of shots while having players making more clutch shots would definitely intrigue the majority of viewers.

Predictive Analytics:

Although Tableau was a very simple graphical tool, there are other way to use analytics in order to predict which players to play or even choose to trade. Using R Studio software, we are able to use predictive analytics, in order to look at certain characteristics to determine a specific outcome for professional basketball players. After the cleaning steps, we can see which players are most likely to be a superstar and which are not by running certain lines of code. Before running the code, a judgement call must be made to determine what a superstar is based on the given predictors. I've decided to classify a superstar as a player that has a field goal percentage over 40%, points per game greater than 24, and a clutch rating over 25. Once this was declared, I was able to run some code in R to output two important results that will help a team grow. We ran a decision tree model to figure out which variables are the most important when determining if someone is a superstar, and to create a tree that gives the probabilities that a player will be a superstar or not. The model resulted in an accuracy of 97.6%, which is an extremely strong model, and these two outputs are a major key to a team's success:

(1) Figure 5 shows the importance of the variables in the model, so now we are able to focus on these when making our conclusions. To read these results, start at the top with the most important predictor having a 100, and then the following predictors are ranked by strength between 1-100. Free throws attempted, free throws made, and field goals attempted are the three most important variables when making our predictions. When looking coaches are looking at players to recruit or trade, they can look at these factors to get a better idea of whether or not a person is a superstar. When doing this, an

analyst should contact the owner or coach in order to determine their exact definition of a superstar.

Figure 5

rpart variable importance

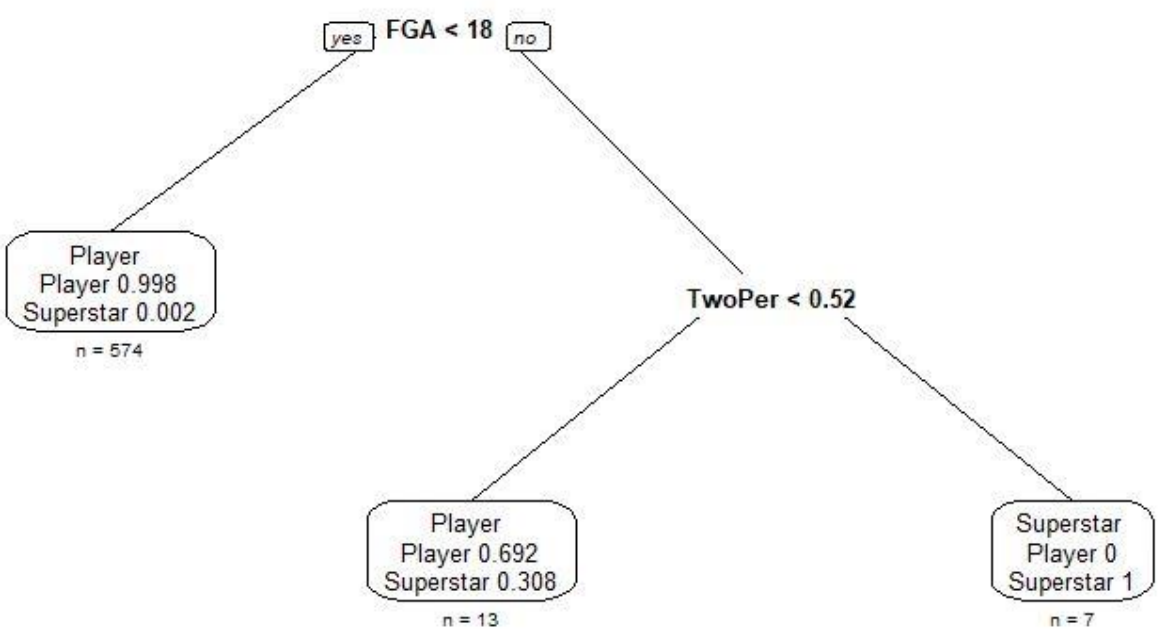
only 20 most important variables shown (out of 24)

	Overall
FTA	100.00
FT	95.08
FGA	83.02
TwoP	50.73
MP	45.53
TwoPer	31.16
FGPer	26.86
AST	20.13

(2) Figure 6 shows the decision tree that shows the probability a player will be a superstar given the specific factors. The way to read this tree is from the top to the bottom. The first portion of the tree looks at the field goals attempted per game factor being less than 18; if a player's field goal attempts are less than 18 then go left down the tree. This shows us the probability that a player is a superstar or not. The probability that a player is a superstar given this factor is 0.2%, and the probability that a player is not a superstar is 99.8%. Now, if the field goal attempts are greater than 18, go down the right side of the tree where we see two-point percentage. If the player's two-point percentage is greater than 52%, then 100% of the players are superstars. If the player's two-point percentage is less than 52%, then there is a 30.8% chance of the player being a superstar, and a 69.2% chance that the player is a not a superstar. This tree gives us a set of criteria that we follow that gives us statistical probabilities to determine what we should characterize the player as. Giving this tree to a coach or owner allows them to evaluate every player and even potential players they plan to recruit or trade for. This

will not only help the coaches make better, faster decisions, but it will make each team more competitive by getting the best possible players they can. These coding tactics will ultimately give teams an advantage and gives a different perspective when looking at the factors.

Figure 6



Basketball Conclusions:

There is an endless amount of techniques that will provide important information for these players and coaches. This upcoming analytical wave gives even more people the opportunity to help the game with a mix of their creativity and data skills. Overall, these predictive analytics and tableau graphical techniques have given insight into the basketball world and will continue to add value to each players' game, while making this sport today even more competitive.

Analytics in Baseball:

Major League Baseball has been a revolutionary movement for sports around the world; bringing together cultures, eliminating the racial barrier, and uniting fans in different countries. This industry continues to be a popular market that has gotten up to 18 million viewers a game, and something this massive need major analytical attention that will develop the game even more (Average US TV Viewership Selected MLB Games 2017 | Statistic). Through descriptive analytics and tableau visualizations, baseball can be dwindled down to simple numbers and names to understand what needs to be done to improve a professional team.

Descriptive Analytics:

Descriptive analytics can make a major impact in the baseball world today. Looking at historical data, a lot of valuable information can be generated to better the players and ultimately the team. Unlike being a fast-paced game like basketball, baseball shows more of a slow strategic side that can be monitored by looking at several statistical analyses. Some descriptive analytical techniques are described below to show how this strategy could work for the Major League teams.

(1) Season by season, coaches are trying to figure out what components need working on, and which ones are significant enough to make a priority. Figure 7 shows a connecting letters report, which helps us determine the statistical significance between a variable given a sorted group. In this example, I chose batting average as the main component to look at in order to see if there is a statistically significant difference over the years. To read this output, you must look at each group's letters and see which ones have different letters in it. For example, "1930 to 1950" has a

“c” and “1950 to 1970” has a “d”; since they have no similar letters, there is a statistically significant difference in batting average between the groups. When two groups have two similar letters, there is no statistical significance in batting average between those decades. For example, the group “1870 to 1890” has the letters “ab” and the group “1890 to 1910” has the letter “a”; since they have the same letter “a”, there is no statistically significant difference between the two batting averages. Furthermore, the farther the letters are apart, the bigger the difference in batting averages between the groups. The “1890 to 1910” group and the “2010 to 2017” group have the biggest difference in batting average because the letters “a” and “f” are the farthest letters apart from one another. Whereas, the two closest letters have the smallest significant difference. In this case, I did the connecting letters report grouped by every two decades since the data was set up that way. There isn’t much value in this specific case because the players between two decades would be a lot different, but this is just an example to show the analysis. This technique would be very valuable to a coach if it was done season by season. Looking to see if their batting average has significantly changed over the past 5 seasons allows the coach to see if some serious changes need to be made or not. This technique could also be done with pitching data. Given the amount of strikes per game a pitcher throws grouped by month could show if there is a significant difference, and if there is, coaches could look at the factors affecting their pitchers and make adjustments. This technique gives coaches a good idea for what they need to focus on and leads them to make adjustments if there is significant output.

Figure 7

```
1890to1910 1870to1890 1910to1930 1930to1950 1950to1970 1970to1990 1990to2010 2010to2017
       "a"        "ab"        "bc"         "c"         "d"         "e"         "f"         "g"
```

(2) One of the most known uses of sports analytics comes in 2002 from the Major League team the Oakland Athletics. Although this example does not have a valuable visualization, this moment was a major turning point for the analytical side of sports and inspired many others to follow the same steps. Before general manager Billy Beane introduced this new strategy, scouts were focusing their time on players that were powerful, fast, smart, and overall just had phenomenal baseball technique. Which in a way, this makes sense because the majority of the time these players have a lot of skill. But, how can teams with a low budget compete with other high-end teams that will most likely get those obvious, skillful players? Billy Beane was motivated by an analytical guru that helped form an organization around statistical data and introduced the idea of sabermetrics, which is using statistical analytics to analyze baseball records. Instead of focusing on those remarkable, high school standouts, the key component in Beane's eyes was on base percentage. He didn't care if they walked, got hit, or hit a weak ball over the second baseman's head, as long as they got on base. This concept focused on people who got on base and scored runs, and with Oakland's low budget, this gave them the best competitive advantage they could have. Going from a team that wouldn't make the playoffs to winning 103 games, making the playoffs, and winning the American League West Title shows how statistical evidence can give any team an advantage if it is properly analyzed (An Examination of the Moneyball Theory: A Baseball Statistical Analysis).

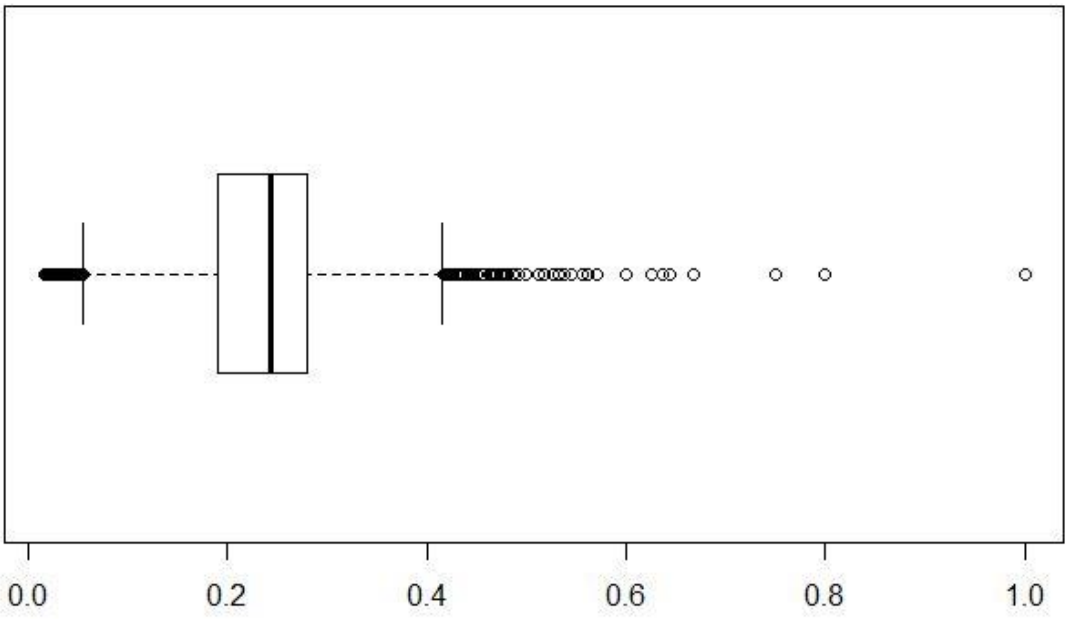
(3) A very simple way to look at historical data is to generate tables in R studio to look at individual players statistics and compare variables. Running one command can produce loads of valuable information that coaches, and owners can use to focus on in practice. Figure 8 shows the layout of the simple table along with all the variables. The table contains over 100,000 baseball players but provided is just an example of 11 of the players. The outcome variable depicts whether or not the batting average is over or under 0.300. If it is below, the outcome variable will say “Bad”, and if it is above, the variable will say “Good”. This 0.300 line was chosen by me because growing up in baseball, having a 0.300 or above has always been a respected batting average. The coaches will able to look at games played, at bats, total runs, total hits, number of singles, doubles, triples and homeruns, strike outs, and batting average. These characteristics are extremely useful in the preseason when developing a practice plan for each individual player. Being a good coach requires the ability to know what to look for. Instead of focusing on a player’s strength, coaches are able to focus on their weaknesses and eliminate those in the future. By the click of a button, all of this information is displayed and used to better each and every team.

Figure 8

playerID	yearID	stint	teamID	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	GIDP	RunsPerAB	Year	Outcome	BatAvg
madlobi01	1981	1	PIT	82	279	35	95	23	1	6	45	18	6	34	17	5	0.125	1970to1990	Good	0.341
bailobo01	1977	1	TOR	122	496	62	154	21	5	5	32	15	6	17	26	7	0.125	1970to1990	Good	0.310
poqueto01	1976	1	KCA	104	344	43	104	18	10	2	34	6	5	29	31	9	0.125	1970to1990	Good	0.302
jonescl01	1971	1	NYN	136	505	63	161	24	6	14	69	6	5	53	87	9	0.125	1970to1990	Good	0.319
rayjo01	1988	1	CAL	153	602	75	184	42	7	6	83	4	1	36	38	10	0.125	1970to1990	Good	0.306
randlle01	1974	1	TEX	151	520	65	157	17	4	1	49	26	17	29	43	10	0.125	1970to1990	Good	0.302
crawfwi01	1976	1	SLN	120	392	49	119	17	5	9	50	2	1	37	53	10	0.125	1970to1990	Good	0.304
sanguma01	1975	1	PIT	133	481	60	158	24	4	9	58	5	4	48	31	12	0.125	1970to1990	Good	0.328
mattido01	1989	1	NYA	158	631	79	191	37	2	23	113	3	0	51	30	15	0.125	1970to1990	Good	0.303
simmote01	1971	1	SLN	133	510	64	155	32	4	7	77	1	3	36	50	20	0.125	1970to1990	Good	0.304
johnsla02	1979	1	CHA	133	479	60	148	29	1	12	74	8	2	41	56	23	0.125	1970to1990	Good	0.309

(4) Sometimes a coach may want to see the overall distribution of their teams batting averages in order to depict any outliers that may be exceeding their expectations or not meeting their standards. Figure 9 shows a box plot of the distribution of batting averages and we are easily able to see the mean batting average and how the majority compares to it. Since one of the components we are analyzing is the mean, I removed all batting averages that were 0 because these were most likely people who had 0 at bats. The dots are on the right and left of the “whiskers” are considered outliers since their batting averages are extremely high and low. When looking at these, we need to determine why these are so different and determine if the data is correct and not a human input error. The dark black line shows the batting average mean, which gives use a solid baseline to compare specific players to the mean. If players are performing above the average, coaches can give them the credit they deserve, but if the players are below the average, coaches can confront these players to motivate them and help them improve.

Figure 9



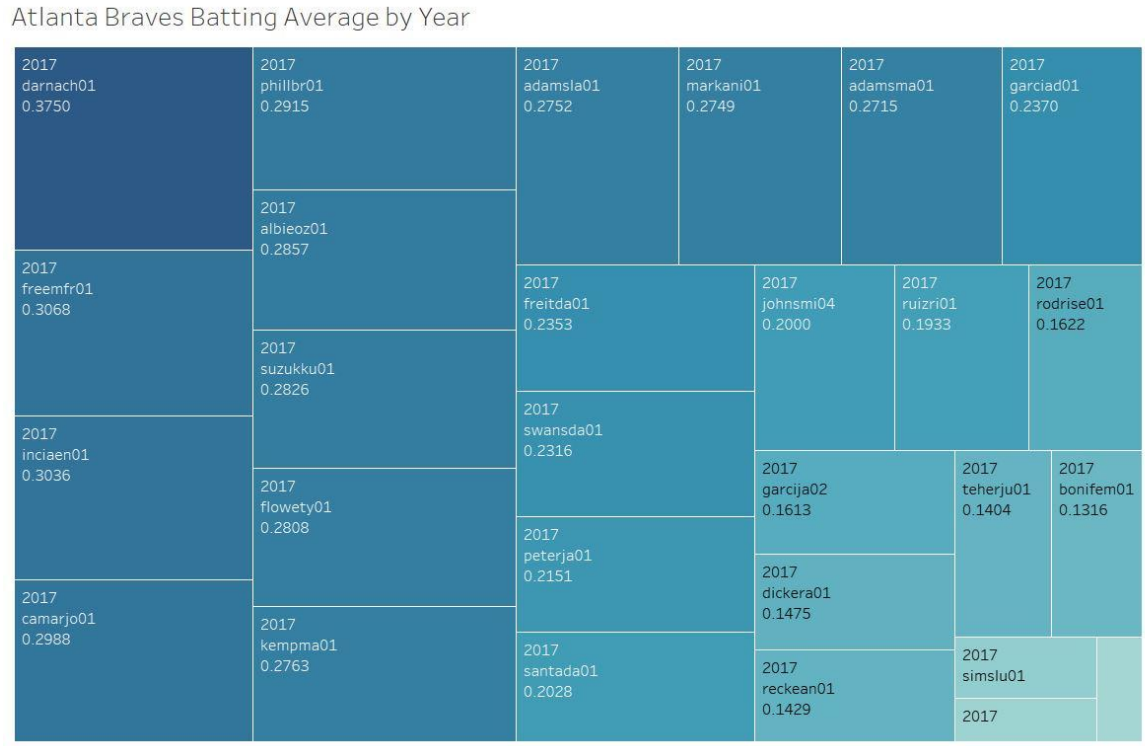
These descriptive analytics techniques give a historical look at the player's skills and weaknesses and can benefit each team by developing their game plan around this data. The only difficult task when using this analytical process is gathering the historical data. This process takes time and patience, and sometimes this could take two to three baseball seasons to see significance results. Making sure to be consistent and having a person focus on the data process will keep the data accurate and reliable.

Tableau:

Hitting back on the tableau graphics, this software does a lot more than just bar charts. When looking at the baseball data, tableau gives a variety of different visualizations that can easily be filtered when given messy data. The batting average column must be calculated, and then inputted with the other variables in order to begin the analysis. When doing an analysis, sometimes there needs to be a lot of information in the visualization but it just will be too messy and hard to understand.

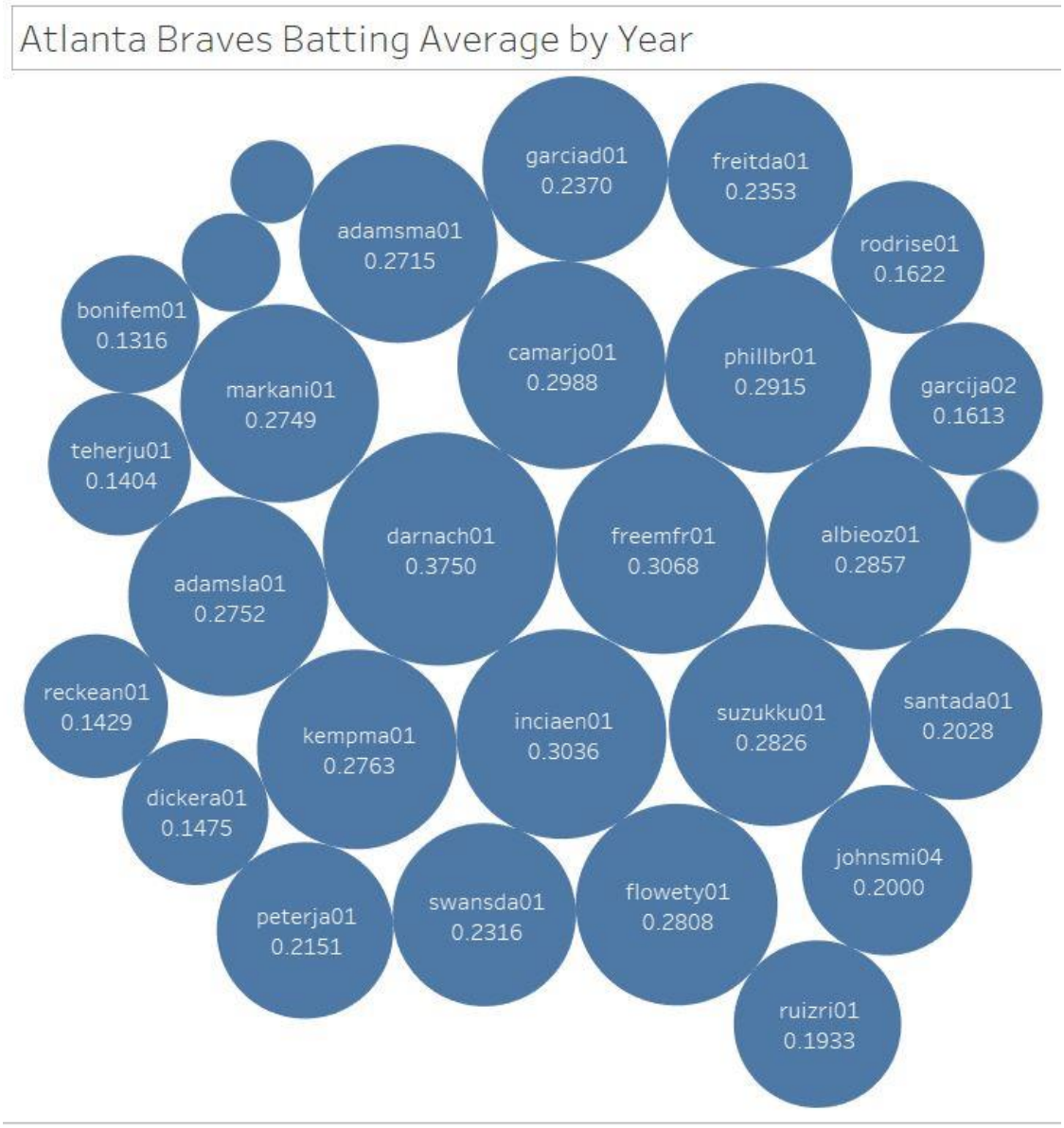
(1) Figure 9 shows a mosaic plot that was a good alternative for wanting a team, the individual players, and the batting averages all in the same graphic. This chart shows us the batting averages of the players on the Atlanta Braves. The darker squares are the players with the best batting average, so this graphic makes it easy to distinguish which players are doing good and which players are doing poorly. The actual batting average number and the player's ID are shown in the top left corner of each box, so the coach can determine which box matches with which person and to know the set number for each player's batting average. The player ID's are set by the coach, so they will be able to distinguish the appropriate person. This visualization benefits coaches and owners because they are able to distinguish where everyone stands and whether or not someone should be pushed to practice harder or be traded. This graphic relates to players on a specific team, but another graphic like this can be created for potential draft picks. Being able to depict the best statistical players rather than the most athletic players helps make smarter draft picks to build a future team.

Figure 10



(2) Figure 10 shows another tableau visualization that may be appealing to different coaches. This dot plot shows the best batting averages for each player on the Atlanta Braves based on the size of the circles. The bigger circles depict players with large batting averages, whereas the smaller circles depict players with smaller batting averages. Having a different variety of graphs gives the recipient many options to choose from and allows them to pick their favorite for viewing. With this chart, there is the ability to remove the bigger circles in case you only want to focus on the smaller ones. When showing these graphs to coaches or owners, finding the graph that is easiest to read is the advantage of this tool.

Figure 11



Predictive Analytics:

Lastly, predictive analytics can perform in a way to understand if a player will be successful or not in the future. General managers or coaches should be able to look at specific variables to determine how a player will perform, and with the amount of players that could potentially be added to the team, the coaching staff will want to determine all of their predictions and compare them to each other. With R Studio, the data can be split into a "train"

dataset and a “holdout” data set. The “train” dataset is used to create a predictive model that looks at all the variables to determine if someone is successful or not. In this case, the term successful is to have a batting average over 0.300. The train model created outputted an accuracy of 98.37%, which is a very strong model. Overfitting the model on the training set is the only issue that we need to be cautious about. This means that the model molds to closely to the training set rather than reflecting the overall model. After creating the predictive model, we are able to run a test on the remaining “holdout” data to determine if the model predicts players success correctly, and to determine if the model is overfit or not. The accuracy on the “holdout” data was a 98.53%, which means the model was not overfit and made really good predictions on the remaining data. Figure 12 shows an example of the code to illustrate what was needed to do to get these outputs. This specific technique is called a “Random Forest”; the model creates many decision trees using all the players in the training set, like the basketball one on page 12 above, and chooses the most common output in all of the trees as the final output model. There are many other techniques to use, but this is just an example of how one predictive model is created and used.

Figure 12

```
train.rows <- sample(1:nrow(BattingForest),0.7*nrow(BattingForest))
TRAIN <- BattingForest[train.rows,] #training model data
HOLDOUT <- BattingForest[-train.rows,]  #Test data

fitControl <- trainControl(method="cv",number=4, allowParallel = TRUE)

forestGrid <- expand.grid(mtry=c(4,7))  #put in values

FOREST <- train(Outcome~.-BatAvg,data=TRAIN,method='rf',tuneGrid=forestGrid,
                trControl=fitControl, preProc = c("center", "scale"))


FOREST$bestTune #Gives best parameters
FOREST$results[rownames(FOREST$bestTune),]  #Shows optimal results
varImp(FOREST)

postResample(predict(FOREST,newdata=HOLDOUT),HOLDOUT$Outcome)
```

Making of train and holdout data set

Coding parameters, can be adjusted

Making of the train model

Testing of outputs

Baseball Conclusions:

Overall, using the tableau visualizations and descriptive analytics allows coaches to focus on their weaknesses and develop their team's abilities as best as they can. Besides learning a few key coding methods and graphical displays, these techniques are fairly simple and add an enormous amount of value and competitive advantage to each team. The predictive analytical techniques give recruiters the opportunity to predict player's success based on variables such as: at bats, hits, games played, RBI's, and other factors. This allows for teams to choose the players that will most likely be the best fit for their team. Implementing these factors will ultimately improve professional baseball teams chance of winning more games and possibly winning a world series.

Overall Analysis:

Using analytical software, the development of professional sports has improved throughout the years, and it will continue to change as teams take advantage of these powerful tools. Using connecting letter reports, box plots, aggregate tables, decision trees, and all the other examples above allows a professional team to better understand their players and where they should move forward from here. Simple statistical evidence can make a major difference in the amount of wins a team can obtain over each season. Basketball and baseball have always been a passion of mine, so being able to analyze this data and make suggestions to coaches and owners would be an amazing opportunity. These techniques can be done in many other professional sports, and analyst that have that passion for other sports would use tools similar to these in order to make valuable outputs. Other tactics will definitely originate that will help make better decisions, and as software advances, analyst will be able to make better

predictions, better decisions on historical data, and more in-depth graphics that may add more value. Sport analytics can make a major impact on professional sports and keeping up to date with these tools will continue to help teams have a competitive advantage.

Works Cited

"An Examination of the Moneyball Theory: A Baseball Statistical Analysis." The Sport Journal, 20 Mar. 2015, thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/.

"Average US TV Viewership Selected MLB Games 2017 | Statistic." Statista, www.statista.com/statistics/251536/average-tv-viewership-of-selected-major-league-baseball-games/.

"Global Sports Market Revenue 2005-2017 | Statistic." Statista, www.statista.com/statistics/370560/worldwide-sports-market-revenue/.

Lahman, Sean. "Download Lahman's Baseball Database." SeanLahmancom, www.seanlahman.com/baseball-archive/statistics/.

Steinberg, Leigh. "CHANGING THE GAME: The Rise of Sports Analytics." Forbes, Forbes Magazine, 18 Aug. 2015, www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#32ccb05a4c1f.