

Project Part 1: NBA Player Performance

Austin Lacey, Nick Orecchia, David Woo

# Executive Summary

After browsing through many different datasets from several sources, we chose the following data for the duration of this project, which effectively covers the average NBA statistics for a select number of players. We extracted our dataset from Kaggle (<https://www.kaggle.com/sachinsharma1123/performance-prediction>), with the hope of exploring the relationship between various statistics and player performance. The raw dataset included 1,340 unique values (which represented the individual number of players) and a total of 20 different statistics / variables per player. The data was structured in a cross-sectional format, as it consisted of various different players across all different era’s with no time factor included. After browsing the variables, the two main variables we determined were the most important were *minutes played* and *target*. The target variable separates the players who have had a career of less than five years (represented by a 0), and the players who have played in the NBA for five years or more (represented by a 1). We believed the minutes played variable to be one of the most important variables because no matter the player’s position or style of play, minutes played is a variable that is equally important and comparable for all of the players. In terms of other variables of interest, we decided to mainly focus on the major statistical categories for basketball which included; Points per Game, Rebounds, Assists, Steals, Blocks, and Turnovers. The overall goal of the first part of the class project is to discover the relationship between the two key variables in the dataset in conjunction with the rest of the variables / statistics per player.

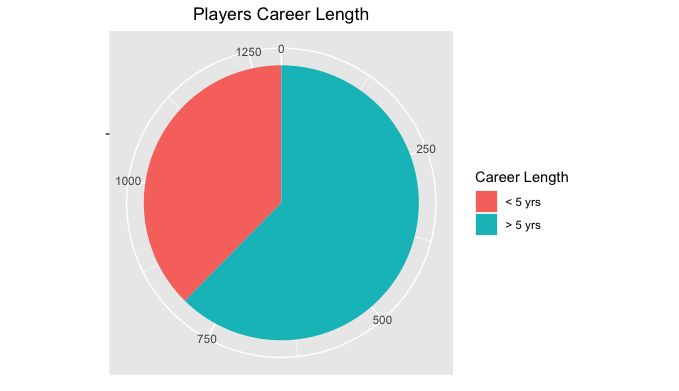
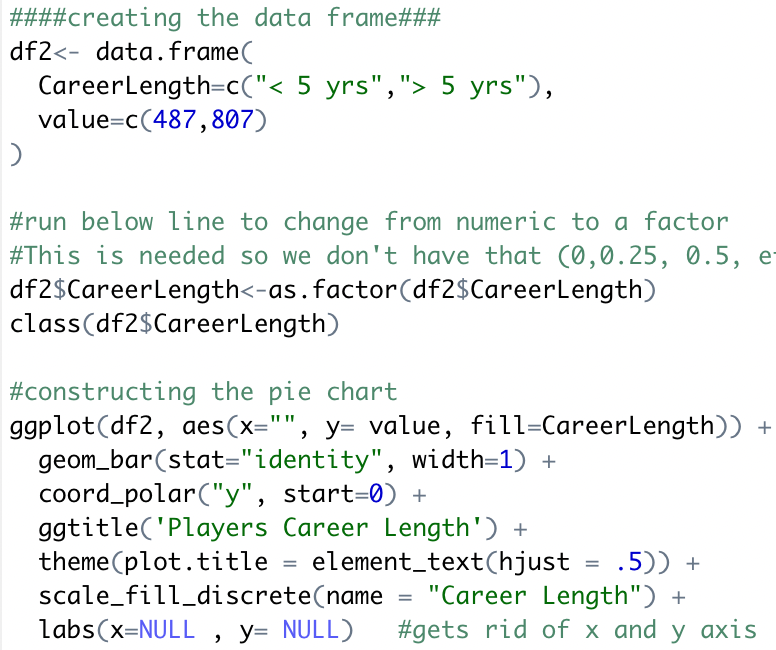
# Preprocessing and Data Cleaning

The data that we collected from Kaggle came to us relatively clean, and as a result didn’t require us to do much in terms of preprocessing and data cleaning. We did not have to deal with any missing or null values, any rescaling or transforming of variables, or even renaming of column headers. However, we did have to deal with various sets of duplicate values, which as a result forced us to delete observations from our data set. As we looked through the dataset we noticed that multiple players had been listed upwards of four times in the original dataset, so in Excel we decided to remove those duplicates. We did this by using conditional formatting, and highlighting the specific observations that were listed more than once. When we identified the specific observations, and realized that they were indeed duplicates (a few father/son combinations were in the dataset, so we had to check via google to see whether or not it was a father/son combo or a duplicate) we deleted them from the data. After removing the 46 duplicate observations, the clean dataset included 1,294 observations. In RStudio we created histograms for all of the variables in our dataset and did this using the code, hist(df$VariableName), as a way to see what the frequencies and distributions of each variable looked like. We also did this to see whether there were any specific outliers that may have been present for specific variables. After running the histograms we noticed that there were no outliers for the variables and at this point our dataset was TIDY.

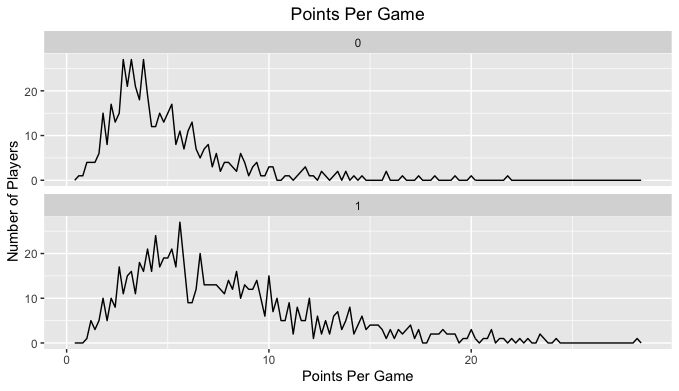
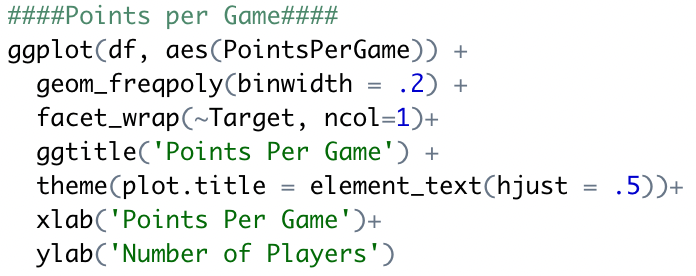
# Exploratory Analysis

Once we had all of our data cleaned, we ran various descriptive analysis to gain a better understanding of the data that we had collected. We did this using the summary function in RStudio, summary(df), and this provided us the mean values, ranges, and interquartile ranges of each of our variables in the dataset. After this we decided to run basic scatter plots using the plot function, plot(df$MinutesPlayed, df$PointsPerGame), and we did this to get a visual understanding of specific correlations or possible relationships that may be present between the variables that we have. After running these we noticed strong positive correlations between variables like minutes played and points per game, and would use this understanding that we gained from the scatter plots for our more in depth analysis and visualizations.

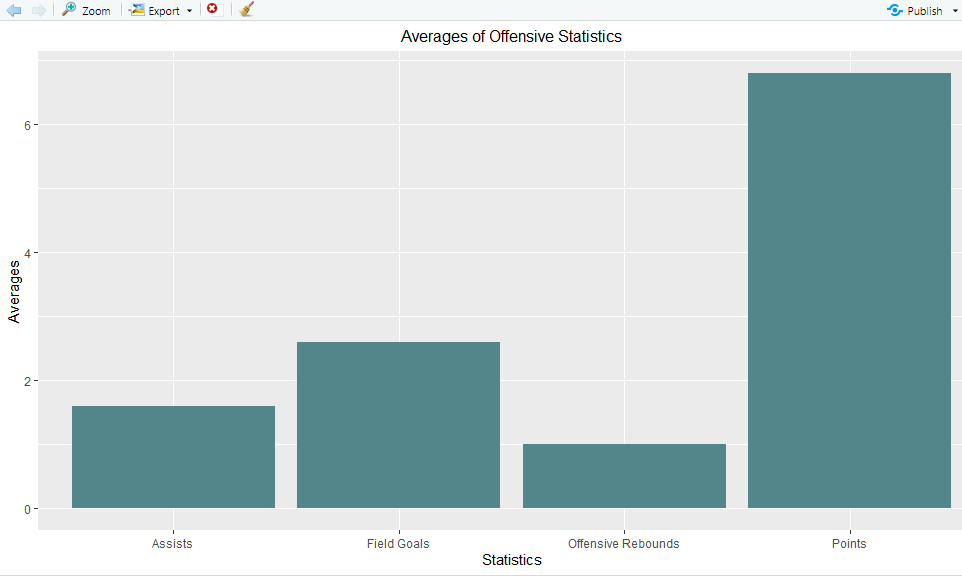
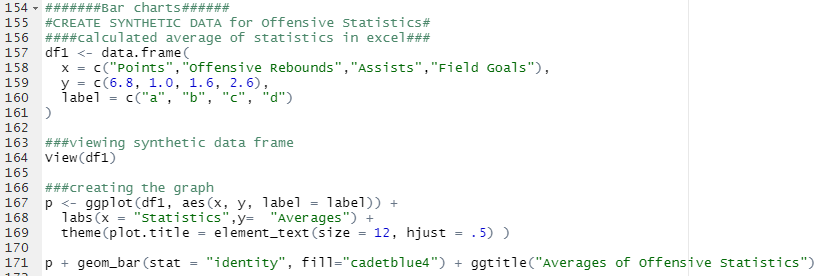
Once we finished with our preliminary exploratory analysis, and gained more knowledge about our data, our next step included looking at the distribution of the “Target” variable which would allow us to understand how many players had a career shorter or longer than 5 years. The visualization that we thought best to explain players' career length was a pie chart, and the first step that we took in making this visualization was creating a new data frame, df2. As the coding in the image below shows, we created df2 and inside this function created two vectors, “Career Length” and “Value” which provides the data for number of players within each of the two Career Length categories(487 for less than 5 years and 807 for more than 5 years). Once we created the df2, the next step we took was changing our Career Length variable from a numeric variable to a factor variable. This allows the variable to be categorized properly and display the pie chart in terms of the two career length categories. Finally we constructed the pie chart leveraging the ggplot2 package. Within the ggplot function we had to set the proper x and y values, and in making a pie chart we left the x value as nothing by only putting quotes, and for the y value referenced the “value” that was created in df2. We then added various chart aesthetic functions like a chart title, centering the chart title, getting rid of the x and y axis labels, and properly labeling the legend. Once the pie chart was completed we were able to have a visual representation that 487 players in our dataset had a career less than 5 years and 807 players had a career that was longer than 5 years.



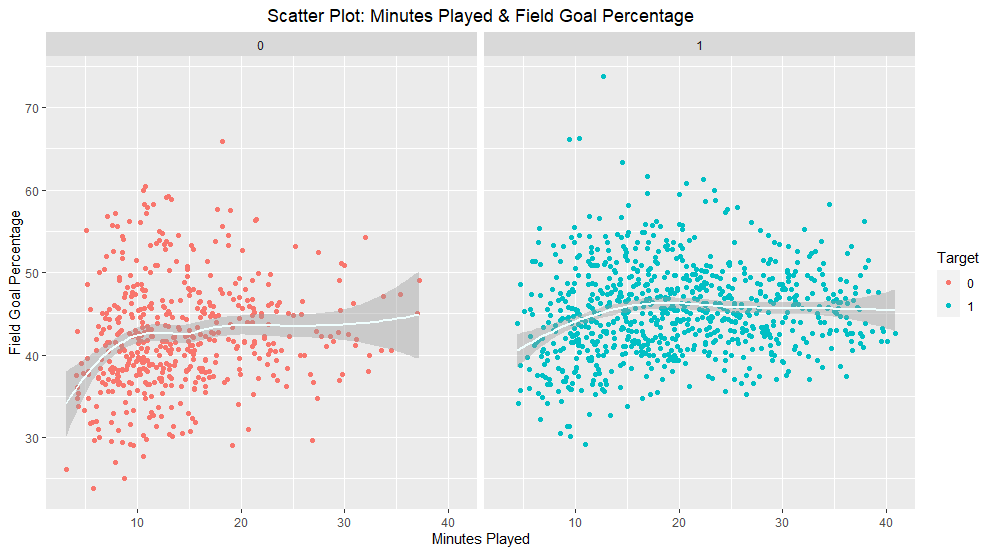
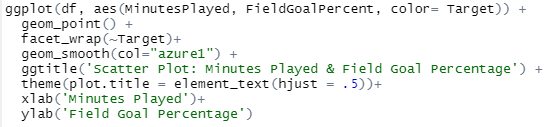
Next, we decided to create frequency poly graphs which allowed us to look at the frequency of various statistical categories across the two “Target'' groups. We decided to graph the variables; Field Goals Made, Points per Game, Rebounds, Assists, Steals, Blocks and Turnovers. We chose to look at these specific variables and their frequencies because they are the major statistical categories that are evaluated in basketball. We created these visualizations using ggplot2 and the geom\_freqpoly function. Within the geom\_freqpoly function we used the bindwith argument to adjust the width of frequency polygraphs. Next, we added the facet\_wrap function with the “Target'' variable as the argument to look at the specific frequencies by the two target variables. We decided to use ncol=1, which positions both the graphs in a vertical manner, to be able to better compare the two separate graphs with each other. We then added various chart aesthetic functions like a chart title, centering the chart title, and x and y lab functions to insert axis labels. For the graph below, we can see that players who had a career longer than five years had more players that averaged higher points per game. The takeaways that we gathered from the frequency poly graphs were that; players who played in the league longer than 5 years had higher frequencies of averaging higher field goals made, points per game, and rebounds. In terms of assists, steals, blocks and turnovers the frequencies seemed to be roughly the same across the two groups.



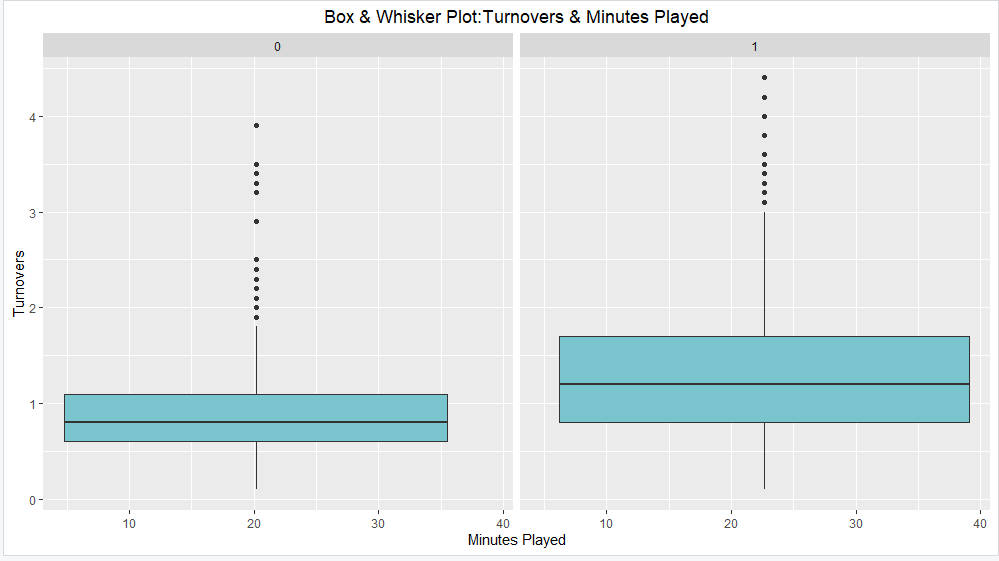
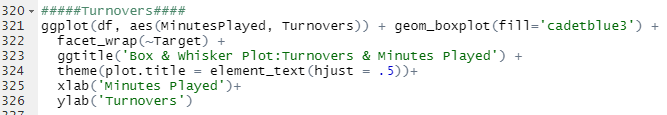
The next visualization we decided to use was a bar chart, which we used to depict the average offensive and defensive statistics across the players. These charts are not facet wrapped, so we aren’t looking at the differences between the 0 and 1 target variables, rather we are trying to just get a better overall understanding of general player statistics. Players who have played in the NBA less than five years and players who have played in the NBA for more than five years are both included in these averages. To create these charts, we first calculated the average for each statistic in excel, and then created a synthetic data frame with these averages in RStudio. The statistics involved in the offensive bar graph include points per game, offensive rebounds, assists, and field goals. Similarly, the statistics involved in the defensive bar graph include steals, defensive rebounds, and blocks.



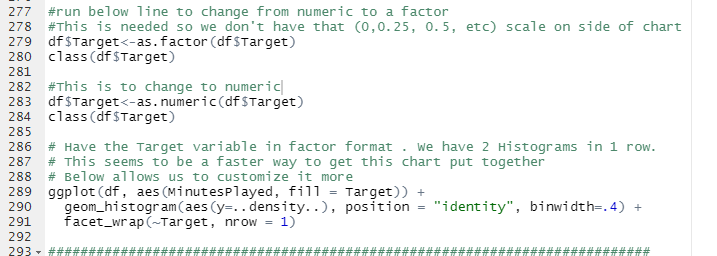
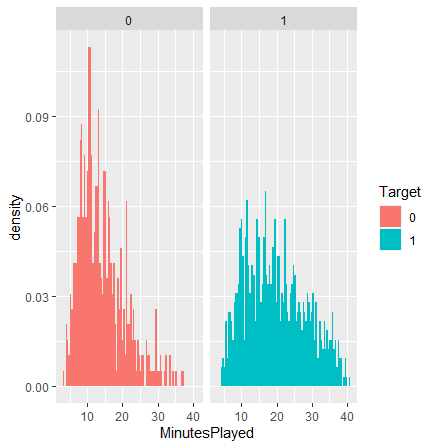
Next, we chose to use scatter plots to visualize the correlations between major statistical variables. On every plot, *minutes played* was depicted on the X axis, with the other statistics plotted on the Y axis (we looked at points per game, field goal percent, free throw percent, rebounds, assists, steals, blocks, and turnovers). We chose to have minutes played represented on every X axis because despite the player’s position or style of play, minutes played in a game is something that can be comparable across all different styles of players. These plots were facet wrapped, so the differences between the 0 and 1 target variables are shown in two separate graphs. We first changed the target variable to a factor so that the legend on the side of the chart is displayed as 0 and 1 rather than a scale from 0, 0.25, 0.5 etc. The target variables are displayed in different colors on each graph to further highlight the difference between the two different groups of players we are looking at. Smoothers are also included to highlight the key trends in each of the variables we looked at. After creating several scatter plots, we came to realize that some sets of variables had obvious correlations (players who had more playing time had more opportunities to get more rebounds and assists) while other sets didn’t. Shown below is an example of one of our scatter plots which shows the relationship between minutes played and field goal percentage. For these two graphs, we were surprised to see that there wasn’t a strong correlation between the two variables. We assumed that the players who had more time on the court were allowed to simply because they were better shooters, and this would go for both groups within our target variables.



The fifth visualization we chose to use for our report is a box and whisker plot which helped visualize the central value and variability across both target groups with the different statistics. Our main goal here is to focus specifically on the distribution of the data across the major statistical categories. Similar to the scatter plot setup, we chose to have minutes played on the X axis of every plot, with points per game, rebounds, assists, blocks, and turnovers on the Y axis. This not only allowed us to keep the visualizations consistent, but also helped us better visualize the distributions. These charts were also facet wrapped, so it was fairly simple to see the differences between the distributions of the players who were in the NBA for less than five years and the players who played in the NBA for more than five years. After looking at the charts, players who had a career in the NBA for longer than five years had more minutes played on average. They also had more points per game, rebounds, assists, blocks, and turnovers than players who played for less than five years. Both sets of players in our target variables had outliers, but the box and whisker plot was a fantastic method which demonstrated the central mean and distribution of data points for our data. Shown below is an example of one of our plots, which shows turnovers and minutes played plotted together.

****

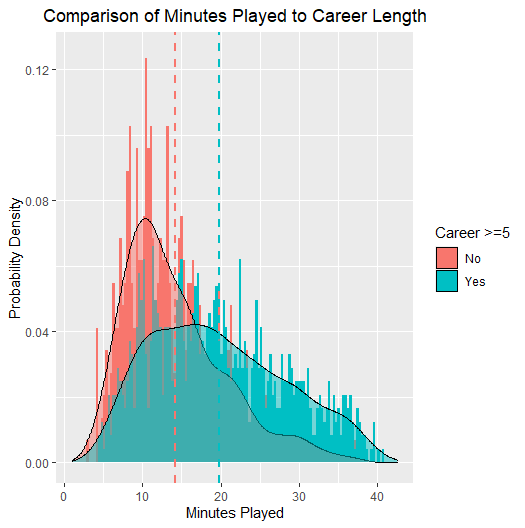
To conclude our project, we looked at the relationship between ”Minutes Played” and “Career Length”. This is also a relevant subject where on ESPN, there has been an emphasis on ensuring that NBA players have limited play time during games and taking rest games. We will see if less minutes played equals a career that spans longer than five years. Since our target variable has been previously converted and stored as a factor this conversion step is no longer needed. Initially we wanted to use the tools that we have been working on most recently in the form of ggplot2, geom\_histogram and facet wrapping. This is a convenient and straightforward method. You will find the image below does show relationships but with the tools we have been provided we can do much better.

First, we must add the plyr package through library(plyr). This package allows us to take complicated problems with many steps and put them into many small pieces that are manageable, and then bring those results all back together. This is important to our analysis because we will be able to stack the charts on top of each other with the many other different analysis tools we have used above and see how they work together.

We start out with nothing but an empty canvas with “Minutes Played” on the X-axis, saving variables along the way using Plot\_a through Plot\_m. It is important for this process to focus on building the essentials early on, adding in complexity throughout the code. This keeps us from having long piecemeal type code, where mistakes can easily be made. So, we started off with putting Minutes Played on the X-axis, then added a histogram that was slightly modified with a binwidth = 0.3 representing our data more efficiently. Next, a vertical line was essential to our dataset which we used geom\_vline to represent this and also added a density curve. We then loaded into one variable (plot\_e). Now that we have a solid backbone for our graph we will start to work on some more aesthetically pleasing features that portray our purpose more clearly. For this to be done color is a must to differentiate the two using the aesthetic fill function. The geom\_density(alpha = .5) function adds the density while using alpha to control the degree of transparency. We then needed the ddply tool to create a vector for the means of each of our groups, giving the mean for “Minutes Played” by “Career Length”. After the dashed vline has been added it was time to then clean up the axis labels with the appropriate names, adding and centering the title and labeling the name of the legend with descriptive labels. Assisting the viewer for clarity purposes. While doing the scaling for the X-axis we found that using the mean minus (-) 2 standard deviations to the left and 3 sd’s to the right showed the information most clearly. With these steps combined, we are left with a very organized and well displayed histogram, showing that there seems to be a correlation between minutes played and career length. On average, players that do not play in the league for at least 5 years are shown to play less time than their counterparts.

* **Please see lines 352 - 432 in file “494Project.R” for work.**



# Conclusion

After analyzing our basketball dataset and taking it through the analytics pipeline our group was able to gain more understanding towards cleaning data sets, gaining specific insights about data, learning to utilize both R and RStudio to create a story in explaining data and finally fully work through the analytics pipeline. Some of our biggest group takeaways were learning how to use the best visualizations to properly explain what questions you are looking to answer. For us, utilizing a pie chart allowed for a great visualization of which players represented which of the two “Target” groups, or layering a histogram to fully understand the comparison between minutes played and a players career length across the two “Target” groups. In terms of our takeaways from our exploratory analysis it appears that the most important statistics/variables that seemed to contribute to playing in the NBA for more than five years included points per game, averaging more playing time, having more rebounds and more assists. Now this makes sense, you’re going to have a more successful career the better you are in these specific categories but it is interesting to see how these statistics cover the different styles of players; the scorers, the facilitators, the big men who rebound, and the iron mans who are on the floor for basically the whole game. It was also interesting that some of the variables weren’t all that different between the two “Target” variable groups. For instance, free throw percentage and field goal percentage were very similar across the two target groups, and this went against what we were initially expecting. We initially thought that the players who had longer careers would have had noticeably higher percentages in these two categories and that there would have been stronger correlations between these variables and the minutes played per game variable. Ultimately we were able to decipher which statistics appear to have the strongest impact on how long a players career was, and more specifically if they were in the NBA for more or less than five years. Now, along with coming up with these interesting and useful understandings for our dataset, our group was also able to learn some important lessons while going through the analytics pipeline. We truly learned the importance of “hacking” and having the patience to stick through a problem in the coding process when maybe something as little as not quoting a variable can be off. We also learned how to properly work as a team, and the importance of explaining the exact steps and processes we were taking throughout our time of coding in R. This allowed for each of our team members to know exactly what the other team member was thinking and to keep our project moving like a well oiled machine. Ultimately, we were able to learn more about what it takes to work with others through the analytics pipeline but also gain specific insights and understandings about our basketball dataset.