STA 141A Project: Does NBA Draft Position Matter?

6/6/2022

Name:	**Email Address**	**Contributions**
Anthony Nguyen	akvnguyen@ucdavis.edu	Introduction, Descriptive Statistics, Scoring Rubric, Methodology, Results & Discussion, Conclusion
Arya Gupta	aaggupta@ucdavis.edu	Introduction, Scoring Rubric, Descriptive Statistics, Methodology, Results & Discussion
Diego Alatorre	dialatorre@ucdavis.edu	Descriptive Statistics, Linear Regression, Results and Discussion, Diagnostic Plots, Response
Jared Choy	jchoy@ucdavis.edu	Dataset Description and Construction, All Plots, Descriptive Statistics, Linear Regression Part I, Response
Jasper Liu	jetliu@ucdavis.edu	Plotting, Research Questions, Descriptive Statistics, Conclusion

Contents

Introduction

Each year, an NBA draft takes place where all NBA teams have the opportunity to select from the best prospects around the world. Sixty players are drafted in a year, with 30 players drafted in each round. The purpose of this project is to determine whether draft position matters in the long run. In other words, does a high draft pick (for example, the first overall player to be selected) consistently have a more successful career than a lower draft pick (for example, the last player selected in the first round)? To compare which players are "more successful" we will consider numerous data points and create a grading scale which we will denote as their "career score". The scoring rubric to create the career score is listed on page 2. We will be using two datasets of NBA players drafted anywhere from 1963 to 2020. After computing an overall career score for each player and determining if, on average, higher draft picks had more successful careers than later draft picks, we will reorganize players by their respective draft era to test if draft position mattered more or less during specific eras of the NBA.

Dataset Description

We pulled two datasets from *Kaggle.* The first dataset supplied the information regarding player statistics by season, and the second assigned career accolades won per player. Additionally, a player can have a draft year lower than 1996, as long as they were an active player in 1996. We merged the two frames using a personally made function to create our final dataset. Our final dataset contains player information between the 1996 season and 2020 season, with 1640 observations and 13 attributes. Listed below are the attributes of the data frame.

Attribute	Description
player_name	The name of each player
draft_year	Year the player was drafted. Values include 1963-2020.
draft_round	Round the player was drafted in. Binary variable containing values one or two
draft_pick	The number the player was selected in the corresponding year's draft
pts	Career average for points scored
reb	Career average for rebounds accrued
ast	Career average for assists performed
dpoys	Number of Defensive Player of the Year awards a player won
mips	Number of Most Improved Player awards a player won
mvps	Number of Most Valuable Player awards a player won
rotys	Number of Rookie of the Year awards a player won (the maximum is 1)
sixth.mans	Number of Sixth Man of the Year awards a player won.
career.score	The overall career score assigned by our scoring rubric below. Created by us

Research Questions

1. Is draft position a good indicator of overall NBA career success?
    a. Can a model be used to depict if career scores correlate alongside the draft order?
    b. How do the career scores differ between top-5 picks, lottery picks, and the rest of the first round picks?
2. Did draft position matter more or less during different era ranges of the NBA?
    a. If so, is there a certain era where the draft pick was the most important?

Scoring Rubric

Accolade	Points
Most Valuable Player	20
Defensive Player of the Year	15
Rookie of the Year	10
Most Improved Player	10
Sixth Man of the Year	10

Player Statistics	Points
Points Per Game	1 PPG = 1 points
Rebounds Per Game	1 RPG = 1.2 points
Assists Per Game	1 APG = 1.5 points

* We assigned the points in the accolades section by prestige associated with the award, and the player statistics by descending overall averages. For example, PPG is only worth 1 point because it is the highest personal average of the 3.

Descriptive Statistics

To better understand our dataset and quantify the average NBA player, we calculated some descriptive statistics. This gave us a better visual representation of the league-wide averages as well as any outliers or trends within our data. For an average player in the NBA, the performance statistics are as follows:

$Mode_{\text{year}}$	$Mode_{round}$	$Mode_{pick}$	μ_{points}	μ_{reb}	$\mu_{assists}$	SD_{pts}	SD_{reb}	$SD_{assists}$	$\mu_{career.\ score}$	$SD_{career.\ score}$
2019	1	4, 8, 9	~6.90	~3.09	~1.50	~4.9	~2.06	~1.47	13.71	11.20

Max/Outliers

The draft years were similar with roughly 59 players drafted each year from 1996 to 2020. There were 960 first-round picks and 660 second-round picks total, and the earliest draft years were 1963, 1976, and 1978. Picks 4, 8, and 9 were the most common in this dataset, but the top ten draft picks were all fairly represented in the data. There were 39 outliers for player points, 26 outliers for player rebounds, 105 outliers for assists, and 72 outliers for career score. Players who exceeded 16.69 career points, 7.22 rebounds, and 4.43 assists respectively are considered to be in the top 5% for their corresponding category. We must note that if we look at these statistics together, incredibly only 5 out of 1640 players averaged more than 16.69 points, 7.22 rebounds, and 4.43 assists together over their careers.

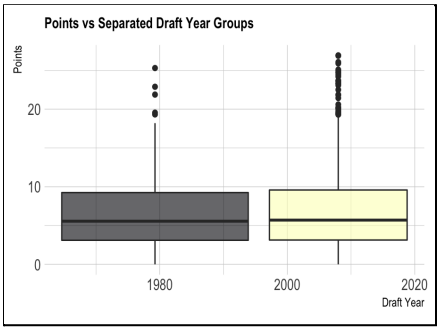

Figure 1

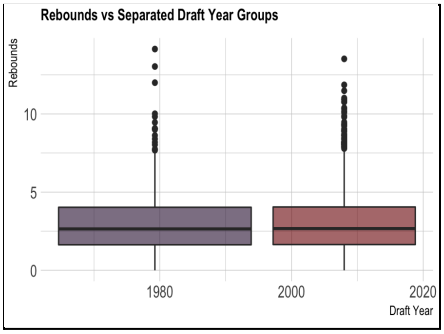

Figure 2

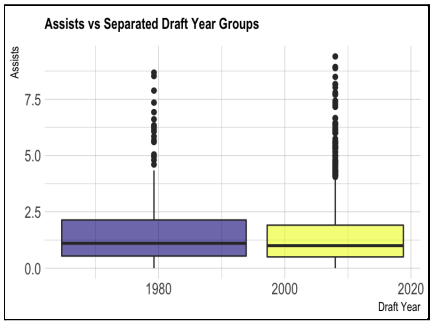

Figure 3

Listed above are three plots separating overall career PPG, RPG, APG by respective draft year. The boxplots are separated by pre-1996 and post-1996 draft classes. We can see that there are consistently more outliers for players drafted after 1996 across all plots. As per the boxplots, 3 out of the 5 players with the highest rebound averages were drafted pre 96'.

Personal Statistics

PPG

Name	Total
Lebron James	26.93
Kevin Durant	26.88
Allen Iverson	26.06
Luka Dončić	25.90
Michael Jordan	25.30

RPG

Name	Total
Dennis Rodman	14.15
Andre Drummond	13.52
Jayson Williams	13.03
Charles Barkley	12.00
Dwight Howard	11.85

APG

Name	Total
Chris Paul	9.40
Trae Young	8.93
John Wall	8.89
Jason Kidd	8.68
John Stockton	8.53

PPG: The average amongst these five players is 26.21645 PPG, which is more than five standard deviations above the average amount of points scored. Evidently, these players are heralded as some of the most elite and gifted scorers in NBA history.

RPG: The top five total rebounders of all time are all not included in this list, due to their era not being within 1996 to 2020. Players such as Kevin Garnett, who has more rebounds than all of the other players on this list, is not included because we are looking at averages and not total. The average mean among the five is 12.91287, which is more than six SDs above the average.

APG: The top five for this category are all renowned for their assisting capabilities. It is important to note that John Stockton, the all-time total assist leader, is fifth on this list because the data frame only presents the last seven years of his career as, once again, we are only considering data after the year 1996. The average among the five players is 8.8849 assists, which is six standard deviations above the average.

Accolades and Career Score

In the tables displayed below, we see the top five players ranked by our career score as well as the top five who have won each award the most times since the 1996 season. Since only one player can win each award every season, even the most coveted players win just a handful of awards. On the other hand, the average NBA player will most likely not win any individual award in their career.

Career Score

Name	Total
LeBron James	137.03667
Giannis Antetokounmpo	104.46750
Tim Duncan	86.32737
Karl Malone	80.06250
Stephen Curry	78.79417

MVP

Name	Total
Lebron James	4
Giannis Antetokounmpo	2
Karl Malone	2
Michael Jordan	2
Stephen Curry	2

DPOY

Name	Total
Dikembe Mutombo	3
Dwight Howard	3
Rudy Gobert	3
Alonzo Mourning	2
Kawhi Leonard	2

SMOY

Name	Total
Jamal Crawford	3
Lou Williams	3
Aaron McKie	1
Antawn Jaminson	1
Ben Gordon	1

Methodology

Is a player's draft position a good indicator of overall NBA career success? To answer this question we split our project into two stages. First, we created an outcome variable we described as “career score" to quantify the career success of each player. To determine a player’s career score, we factored in each player's average points, rebounds, and assists from 1996 to 2020. In addition to this, we filtered the dataset by accolades won and grouped each player by the number of times they won each award. Next, we created a function called ‘accolades.finder’ which merged the number of accolades won by a player onto our dataset. Finally, we used the scoring rubric described above to multiply the players’ personal statistics and accolades by the rubric's respective score to create their career scores. By using career scores, we were able to quantify the success of each player’s career which allowed us to effectively compare the careers of each player.

After we computed the career scores of each player, we moved on to the second stage of our project. To test if there was a correlation between draft position and overall career success, we created a linear regression model that used the original draft selection of the player as the independent variable, and our “career scores” as the dependent variable. Finally, we created plots to find the difference between the career scores of players picked in different tiers incrementing by tens.

We also tested to see whether draft position mattered more or less during different era ranges of the NBA. In other words, were NBA teams able to evaluate and assess players’ potential better during different periods of the NBA? To answer this question, we grouped every player by their respective draft years and reorganized the players by three different eras (1963 to 1995 = Era 1 | 1996 to 2010 = Era 2 | 2011 to 2020 = Era 3). We then regressed each respective model to compare the coefficients of draft pick and round to see if these values changed depending on the era the players participated in.

Results and Discussion

Is draft position a good indicator of overall NBA career success?

After computing the career scores of every NBA player and grouping them by tiers (picks 1-10 were denoted as tier 1, 11-20 were tier 2, etc.) we found that draft position did have some correlation with overall career success. As seen in figure 4, *Career Score by Tier*, on average, the boxplots of higher tier players were greater than those of the lower-tier players. However, this difference falls off drastically when comparing middle-tier players to lower-tier players. In other words, the overall career score of players was more evenly distributed the later into the draft they were selected. Looking at Figure 5, *Career Score by Draft Pick*, we can see that these two points are made clear, as the regression lines become more horizontal and the career score decreases as the pick number increases. Based on this, it is suggested that draft position can be a more accurate predictor of overall NBA career success the earlier the players were selected.

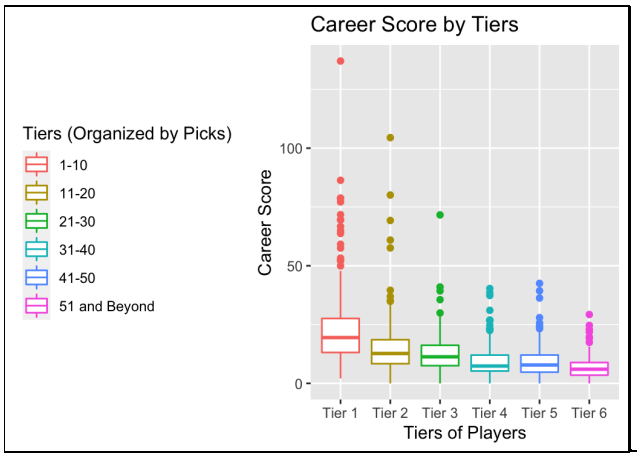

Figure 4

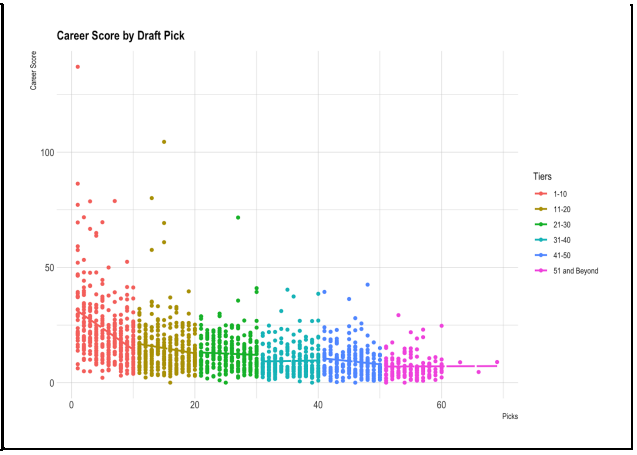

Figure 5

To further investigate if draft position is a good indicator of overall career success, a linear regression model was used. Different models were tested, but adding new regressors and transforming the dependent variable did not seem to significantly change the results nor improve the model. Therefore, it was concluded that it was best to use a model with only two regressors. The model is of the following form: player.score ~ draft_round + draft_pick. The summary of the regression and diagnostic plots are shown below:

	Estimate	Standard Error	t	P-value
Intercept	21.6297	0.8370	25.841	<2e-16
Draft Round	-0.8608	0.8617	-0.999	0.318
Draft Pick	-0.2513	0.0237	-10.603	<2e-16

As seen from the table above, draft pick is the only significant variable and has an estimated coefficient value of -0.2513. This indicates that when draft picks increase by one position, career score decreases by .2513, holding other factors fixed. This supports our claim, as we can see from the previous box plots and regression model that as the tiers increase from 1 to 5, the career score consequently decreases. This means that career score and pick position do correlate highly with one another, and the top-rated players will typically be from the first tier while players with lower ratings will be in the subsequent tiers. Moreover, it is worth noting that this model has a few issues. First, the R-squared value is very low, so our model is not a great fit. Additionally, we see from the diagnostic plots that our model has issues regarding heteroskedasticity, non-normality, outliers, and high leverage points. Therefore, the relationship between our predictors and the response variable cannot be reliably determined from this model.

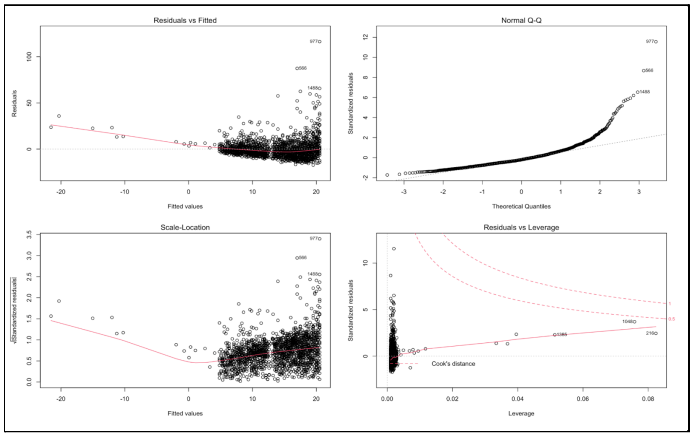

Figure 6

Did draft pick position matter more / less during different era ranges of the NBA?

After reorganizing the players into different eras and running our regressions again without considering the draft round, we were able to compare the differences between the coefficients for draft pick by era. Based on these summaries, we observed that draft pick position had the largest impact in Era 2, which is supported by the fact that the draft pick estimate coefficient was the smallest (-0.3442). This was very closely followed by the coefficient from Era 3 (-0.3274). The coefficient from Era 1 (-0.1525) was by far the lowest, indicating that draft pick position was least impactful in Era 1. It is important to note that since all of our draft pick estimators are negative, a higher draft position was negatively correlated with overall career success, especially during era 2.

We should also note that the number of observations are different across the three eras. For Era 1, there are 349 observations, for Era 2 there are 740 observations, and for Era 3 there are 539 observations. Since Era 1 had fewer observations and given that the majority of the players had to have successful careers to continue playing in 1996, it is evident that the coefficient for draft pick would be more positive than the other two eras. Era 2 and Era 3 are quite similar but Era 2 may have a smaller coefficient because just like for Era 1, Era 3 has fewer observations. Likewise, players in Era 3 were drafted in recent years, giving them less experience which could potentially lead to inflated statistics. Compared to Era 1, Eras 2 and 3 had a significantly larger t-value, supporting the idea that there exists significant differences between Era 1 and Eras 2 and 3. In Figure 7, we can see that the tier 1 regression line has a notably steep slope, with tiers 2 and 3 being slightly more constant. Although tier 4 had an increasing slope, the overall trend of the model remained consistent: the later into the draft players were taken, the lower their average career score was. In Figure 8 and 9, we see a similar trend of tier 1 having the steepest slope, but the remaining data points vary from decreasing to slightly increasing. In all, the three graphs showed us that, on average, a higher draft pick does affect a player's career score.

Era 1

	Intercept	Draft Pick
Estimate	17.3189	-0.1525
Standard Error	0.7794	0.0233
t	22.220	-6.555
P-value	<2e-16	1.94e-10

Era 2

	Intercept	Draft Pick
Estimate	22.9455	-0.3442
Standard Error	0.7505	0.0241
t	30.57	-14.27
P-value	<2e-16	<2e-16

Era 3

	Intercept	Draft Pick
Estimate	22.9081	-0.3274
Standard Error	0.7689	0.0236
t	29.79	-13.85
P-value	<2e-16	1.57e-15

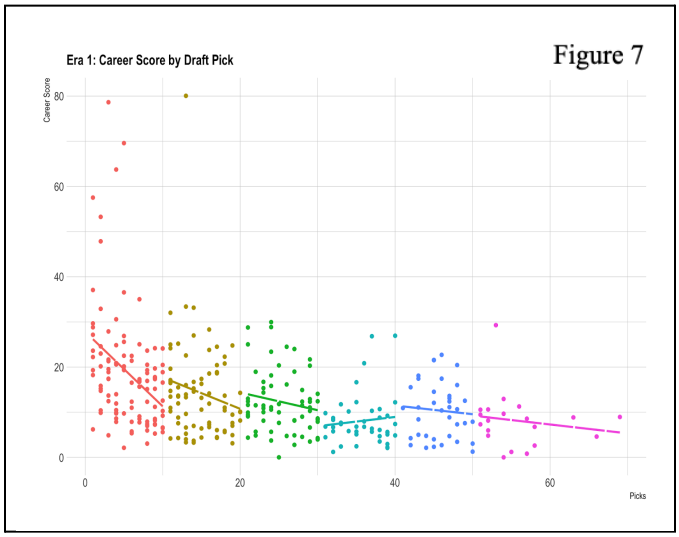

Figure 7

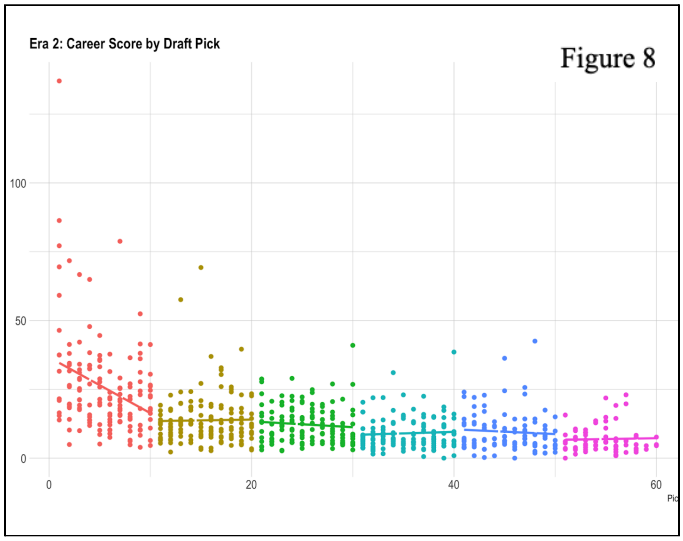

Figure 8

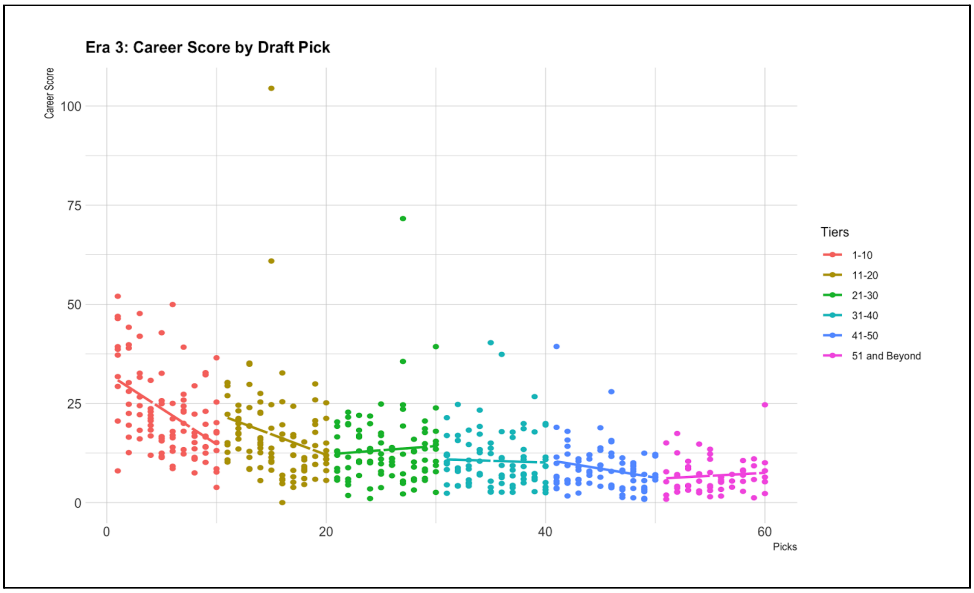

Figure 9

Conclusion

For our project, there were two main stages. First, we quantified the overall career success of NBA players by creating an outcome variable, 'career score', which included personal statistics and individual accolades earned through the course of the respective careers of each player. We used our created grading rubric to calculate the outcome variable for our players, both active and inactive. Next, we ran multiple linear regressions models and plotted graphs of our data with players grouped by their respective draft pick positions and draft years. Based on the results of our models, we found that draft pick position was a good predictor of overall NBA career success as players who were selected closer to the beginning of their draft classes were more likely to have better careers. Likewise, the average predicted career score value slightly decreased as each draft pick went by. On the other hand, we found that the later into the draft a player was selected, the less influential the draft position was. In addition to this, based on our coefficients, we also found that draft position mattered the most during Eras 2 and 3 of the NBA and had the least impact in Era 1.

Despite our findings, the diagnostic plots of our model suggest that there are issues regarding heteroskedasticity, non-linearity, outliers, and high leverage points. Additionally, we used log transformation in an attempt to normalize the data and alleviate heteroskedasticity. We decided to not move forward with the log transformation because even after the transformation of the dependent variable, issues such as heteroskedasticity still remained. Therefore, the relationship between our predictors and the response variable may not be reliably determined from this model. From a part of our summary statistics, we determined that draft pick was the only significant variable with a p-value of less than 2e-16 compared to that of draft round. Based on our career score calculations, we determined that Lebron James, Giannis Antetokounpo, Tim Duncan, Karl Malone, and Stephen Curry had the top five highest career scores based on our rubric.

Overall, drafting players, especially later on in the draft, proves to be a difficult task for NBA scouts and teams. Since the draft pick position was more influential in the first half of the draft, perhaps NBA teams should devote more resources to finding what makes first round picks more gifted than the second round picks. In addition, tier 1 players are expected to perform significantly better than the rest of their draft class, and we found that these Tier 1 players do end up performing better than the rest. Perhaps NBA scouts could determine if these players gain confidence and team expectations from being picked early on, and if that influences their gameplay, if any.

Response

"*It sounds that the career score is a variable that you are building, is that correct?*"

- Yes, this is correct. We created a career score using pre-existing variables within the dataset and separate variables from our other dataset involving personal accolades. We assigned each variable with a certain value to determine our career score.

"*Before you build such a variable, please show some descriptive statistics of the variables you will be putting together. Please, also show descriptive statistics of the variable you create.*"

- This is included in the descriptive statistics section of the report. We found the career score average among the entire data set to be 13.71. We also headed the top five for each category to analyze other properties within our data.

"*I am a bit confused on the use of logistic regression for question 2. What is the outcome variable of question 2? Could run hierarchical clustering to create subgroups of players with similar performance/characteristics?*"

We determined that answering our second question by running logistic regression or hierarchical clustering to not be as meaningful given the structure of our variables and data. Instead, we used linear regression once more, as it was more effective in answering our query.

Appendix:

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
library(dplyr)
library(tidyverse)
library(ggplot2)
library(GGally)
library(ggrepel)
library(viridis)
library(hrbrthemes)
library(ISLR)
# Dataframe Creation
all_seasons <- read.csv("all_seasons.csv")
all_seasons <- all_seasons[-1]
all_seasons <- all_seasons[, -c(2:5)]
all_seasons <- all_seasons[, -c(2:3)]
all_seasons <- all_seasons[-5]
all_seasons <- all_seasons[, -c(8:14)]
all_seasons <- all_seasons %>% filter(draft_year != "Undrafted", draft_round != "Undrafted",
draft_number != "Undrafted")

# convert characters into numeric vectors
draft.yr.num <- as.numeric(all_seasons$draft_year)
draft.round.num <- as.numeric(all_seasons$draft_round)
draft.pick.num <- as.numeric(all_seasons$draft_number)

update.frame <- data.frame(player_name = all_seasons$player_name, draft_year = draft.yr.num,
draft_round = draft.round.num, draft_pick = draft.pick.num, pts = all_seasons$pts, reb =
all_seasons$reb, ast = all_seasons$ast) # this allows us to aggregate everything because they are
now all numeric

final.frame <- update.frame %>% group_by(player_name) %>% summarise(draft_year = mean(draft_year),
draft_round = mean(draft_round), draft_pick = mean(draft_pick), pts = mean(pts), reb =  mean(reb),
ast = mean(ast))

final.frame[final.frame$player_name == "Nikola Jokic",]

# Descriptive Stats Part 1
```
```

```
mean(final.frame$draft_year)
mean(final.frame$draft_round)
mean(final.frame$draft_pick)
mean(final.frame$pts)
mean(final.frame$reb)
mean(final.frame$ast)

sd(final.frame$pts)
sd(final.frame$reb)
sd(final.frame$ast)

# Number of outliers for PPG
ppg_out <- boxplot.stats(final.frame$pts)$out
ppg_ind <- which(final.frame$pts %in% c(ppg_out))
length(ppg_ind)

# Number of outliers for REB
reb_out <- boxplot.stats(final.frame$reb)$out
reb_ind <- which(final.frame$pts %in% c(reb_out))
length(reb_ind)

# Number of outliers for AST
ast_out <- boxplot.stats(final.frame$ast)$out
ast_ind <- which(final.frame$ast %in% c(ast_out))
length(ast_ind)

head(sort(table(final.frame$draft_year), decreasing = TRUE), 3) # most players in draft year
head(sort(table(final.frame$draft_year), increasing = TRUE), 3) # least per draft year
head(sort(table(final.frame$draft_round), decreasing = TRUE), 1) # max round pick
head(sort(table(final.frame$draft_pick), decreasing = TRUE), 10) # draft pick counts

head(sort(unlist(split(final.frame$pts, final.frame$player_name)), decreasing = TRUE), 5)  # top 5
points
(26.93333 + 26.88462 + 26.06429 + 25.90000 + 25.30000) / 5

head(sort(unlist(split(final.frame$reb, final.frame$player_name)), decreasing = TRUE), 5) #top5
rebounds
(14.15000 + 13.52222 + 13.03333 + 12.00000 + 11.85882) / 5

head(sort(unlist(split(final.frame$ast, final.frame$player_name)), decreasing = TRUE), 5) #top 5
assist
(9.400000+8.933333+8.880000+8.682353+8.528571) / 5
```

```
# Now we must create the accolades portion of the dataframe.
Player.Award.Shares <- read.csv("Player Award Shares.csv")
award.frame <- Player.Award.Shares %>% filter(winner == "TRUE") %>% filter(season >= 1996)
award.frame$winner = 1

dpoys <- award.frame %>% filter(award == "dpoy") %>% dplyr::select(player, winner)
dpoy.num <- dpoys %>% group_by(player) %>% summarise(winner = sum(winner))

mips <- award.frame %>% filter(award == "mip") %>% dplyr::select(player, winner)
mip.num <- mips %>% group_by(player) %>% summarise(winner = sum(winner))

mvps <- award.frame %>% filter(award == "nba mvp") %>% dplyr::select(player, winner)
mvp.num <- mvps %>% group_by(player) %>% summarise(winner = sum(winner))

roty <- award.frame %>% filter(award == "nba roy") %>% dplyr::select(player, winner)
roty.num <- roty %>% group_by(player) %>% summarise(winner = sum(winner))

sixth.man <- award.frame %>% filter(award == "smoy") %>% dplyr::select(player, winner)
sixth.man.num <- sixth.man %>% group_by(player) %>% summarise(winner = sum(winner))

accolades.finder <- function(x){
  dpoy.wins  <- merge(x, dpoy.num, all.x = TRUE, by.x = "player_name", by.y = "player")
  dpoy.wins <- dpoy.wins %>% replace(is.na(.), 0)
  mip.wins <- merge(dpoy.wins, mip.num, all.x = TRUE, by.x = "player_name", by.y = "player")
  mip.wins <- mip.wins %>% replace(is.na(.), 0)
  mvp.wins <- merge(mip.wins, mvp.num, all.x = TRUE, by.x = "player_name", by.y = "player")
  mvp.wins <- mvp.wins %>% replace(is.na(.), 0)
  roty.wins <- merge(mvp.wins, roty.num, all.x = TRUE, by.x = "player_name", by.y = "player")
  roty.wins <- roty.wins %>% replace(is.na(.), 0)
  sixth.man.wins <- merge(roty.wins, sixth.man.num, all.x = TRUE, by.x = "player_name", by.y =
"player")
  sixth.man.wins <- sixth.man.wins %>% replace(is.na(.), 0)
}
new.frame <- accolades.finder(final.frame)
colnames(new.frame)[8] <- "dpoys"
colnames(new.frame)[9] <- "mips"
final.frame2 <- new.frame %>% dplyr::rename(mvps = winner.x, rotys = winner.y, sixth.mans = winner)
final.frame2 <- mutate(final.frame2, career.score = pts + (reb*1.2) + (ast*1.5) + (dpoys*15) +
(mips*10) + (mvps*20) + (rotys * 10) + (sixth.mans * 5))

final.frame2$draft_round[final.frame2$draft_pick>30] <- 2
final.frame2 <- final.frame2 %>% filter(draft_round == 1 | draft_round == 2)
```

```
# Descriptive Stats part 2 | MIP and ROTY are redundant as you can only win them once.

head(sort(unlist(split(final.frame2$dpoys, final.frame2$player_name)), decreasing = TRUE), 5) #
dpoys top 5

head(sort(unlist(split(final.frame2$mvps, final.frame2$player_name)), decreasing = TRUE), 5)
# mvps top 5

head(sort(unlist(split(final.frame2$sixth.mans, final.frame2$player_name)), decreasing = TRUE), 5)
#sixth mans top 5

head(sort(unlist(split(final.frame2$career.score, final.frame2$player_name)), decreasing = TRUE),
5) # overall top 5 for career score

out <- boxplot.stats(final.frame2$career.score)$out
out_ind <- which(final.frame2$career.score %in% c(out))
out_ind

length(out_ind) # tells us how many outliers there are total for career score
sd(final.frame2$career.score)

# Linear Regression Q1
test <- lm(career.score~draft_round + draft_pick + draft_year, data = final.frame2)
summary(test)
plot(test)

test.2 <- lm(career.score~draft_round + draft_pick, data = final.frame2)
summary(test.2)

test.3 <- lm(career.score ~ draft_round + draft_pick + as.factor(draft_year), data = final.frame2)
summary(test.3)

test.4 <- lm(career.score ~ draft_pick, data = final.frame2)
summary(test.4)
```

```
# Plots
final.frame2 %>% #i think we can use this
    ggplot(aes(x=draft_year, y=pts, fill = (draft_year>=1996))) +
    geom_violin() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    theme_classic() +
    theme_ipsum() + theme(
        legend.position="none",
        plot.title = element_text(size=11)) +
    ggtitle("Violin Plot of Points vs Draft Year") +
    xlab("Draft Year") + ylab("Points")

# These are for desc stats
final.frame2 %>%
    ggplot(aes(x=draft_year, y=pts, fill = (draft_year>=1996))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="A") +
    theme_classic() +
    theme_ipsum() +
    theme(legend.position="none",
    plot.title = element_text(size=11)) +
    ggtitle("Points vs Separated Draft Year Groups") +
    xlab("Draft Year") +
    ylab("Points")

final.frame2 %>%
    ggplot(aes(x=draft_year, y=reb, fill = (draft_year>=1996))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=.6, option="H") +
    theme_classic() +
    theme_ipsum() +
    theme(legend.position="none",
    plot.title = element_text(size=11)) +
    ggtitle("Rebounds vs Separated Draft Year Groups") +
    xlab("Draft Year") +
    ylab("Rebounds")

final.frame2 %>%
    ggplot(aes(x=draft_year, y=ast, fill = (draft_year>=1996))) +
    geom_boxplot() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6, option="C") +
    theme_classic() +
    theme_ipsum() +
```

```
    theme(legend.position="none",
    plot.title = element_text(size=11)) +
    ggtitle("Assists vs Separated Draft Year Groups") +
    xlab("Draft Year") +
    ylab("Assists")


regression.class <- mutate(final.frame2, tier = ifelse(draft_pick %in% 1:10, "Tier 1",
                                    ifelse(draft_pick %in% 11:20, "Tier 2",
                                    ifelse(draft_pick %in% 21:30, "Tier 3",
                                    ifelse(draft_pick %in% 31:40, "Tier 4",
                                    ifelse(draft_pick %in% 41:50, "Tier 5", "Tier 6"))))))

regression.2 <- regression.class %>% filter(draft_pick < 70)
regression.2 <- regression.2[-c(219),]

ggplot(data = regression.2, mapping = aes(x = draft_pick, y = career.score, color = tier)) +
    geom_point() +
    geom_smooth(se = FALSE, method='lm', linetype = "F1") +
    ggtitle("Career Score by Draft Pick") +
    xlab("Picks") +
    ylab("Career Score") +
    theme_ipsum() + theme(
    legend.position="right",
    plot.title = element_text(size=14)
    ) +
    scale_colour_discrete(
    labels =
      c("Tier 1" = "1-10",
        "Tier 2" = "11-20",
        "Tier 3" = "21-30",
        "Tier 4" = "31-40",
        "Tier 5" = "41-50",
        "Tier 6" = "51 and Beyond")) +
     labs(color='Tiers') # i like this plot

ggplot(data = regression.class, mapping = aes(x = tier, y = career.score)) +
    geom_boxplot(aes(color = tier)) +
    ggtitle("Career Score by Tiers") +
    xlab("Tiers of Players") +
    ylab("Career Score") +
    theme(legend.position = "none") +
    scale_colour_discrete(
```

```
    labels = c("Tier 1" = "1-10",
               "Tier 2" = "11-20",
               "Tier 3" = "21-30",
               "Tier 4" = "31-40",
               "Tier 5" = "41-50",
               "Tier 6" = "51 and Beyond")) + labs(color='Tiers (Organized by Picks)')
  #ggrepel::geom_label_repel(aes(label = tier), data = regression.class)#this is a good plot

# Linear Regression Q2

era1 <- regression.2 %>% filter(draft_year %in% 1963:1995)
era2 <- regression.2 %>% filter(draft_year %in% 1996:2010)
era3 <- regression.2 %>% filter(draft_year %in% 2011:2020)

model_era1 <- lm(career.score~draft_pick, data = era1)
summary(model_era1)
par(mfrow=c(2,2))
plot(model_era1, 1)
plot(model_era1, 2)
plot(model_era1, 5)

model_era2 <- lm(career.score~draft_pick, data = era2)
summary(model_era2)
par(mfrow=c(2,2))
plot(model_era2, 1)
plot(model_era2, 2)
plot(model_era2, 5)

model_era3 <- lm(career.score~draft_pick, data = era3)
summary(model_era3)
par(mfrow=c(2,2))
plot(model_era3, 1)
plot(model_era3, 2)
plot(model_era3, 5)
```

```{r}
# Plots for question 2

ggplot(data = era1, mapping = aes(x = draft_pick, y = career.score, color = tier)) +
    geom_point() +
    geom_smooth(se = FALSE, method='lm', linetype = "F1") +
    ggtitle("Era 1: Career Score by Draft Pick") +
```

```
    xlab("Picks") +
    ylab("Career Score") +
    theme_ipsum() + theme(
    legend.position="none",
    plot.title = element_text(size=14)
    ) +
    scale_colour_discrete(
    labels =
      c("Tier 1" = "1-10",
        "Tier 2" = "11-20",
        "Tier 3" = "21-30",
        "Tier 4" = "31-40",
        "Tier 5" = "41-50",
        "Tier 6" = "51 and Beyond")) +
     labs(color='Tiers')

ggplot(data = era2, mapping = aes(x = draft_pick, y = career.score, color = tier)) +
    geom_point() +
    geom_smooth(se = FALSE, method='lm', linetype = "F1") +
    ggtitle("Era 2: Career Score by Draft Pick") +
    xlab("Picks") +
    ylab("Career Score") +
    theme_ipsum() + theme(
    legend.position="right",
    plot.title = element_text(size=14)
    ) +
    scale_colour_discrete(
    labels =
      c("Tier 1" = "1-10",
        "Tier 2" = "11-20",
        "Tier 3" = "21-30",
        "Tier 4" = "31-40",
        "Tier 5" = "41-50",
        "Tier 6" = "51 and Beyond")) +
     labs(color='Tiers')

ggplot(data = era3, mapping = aes(x = draft_pick, y = career.score, color = tier)) +
    geom_point() +
    geom_smooth(se = FALSE, method='lm', linetype = "F1") +
    ggtitle("Era 3: Career Score by Draft Pick") +
    xlab("Picks") +
    ylab("Career Score") +
    theme_ipsum() + theme(
```

```
    legend.position="none",
    plot.title = element_text(size=14)
    ) +
    scale_colour_discrete(
    labels =
      c("Tier 1" = "1-10",
        "Tier 2" = "11-20",
        "Tier 3" = "21-30",
        "Tier 4" = "31-40",
        "Tier 5" = "41-50",
        "Tier 6" = "51 and Beyond")) +
     labs(color='Tiers')
```