Analyzing Efficient Strategies for Economic Success in the National Basketball Association

Author: Deon Ababio

Discussants: Kaggle (Website)

Introduction

The average salary of an NBA basketball player is 7.7 million dollars. With such a high salary, many kids grow up and aspire to be a professional athlete. However, when I first heard of the average salary, I was shocked. I pondered as to how the average salary is 7.7 million when players such as James Harden, Russell Westbrook, Lebron James, and Draymond Green are making over 25 million dollars per season. At first I believed it was just because they were superstars, but when looking at Draymond Green, who averaged 7.7 points per game last season, land a 100 million dollar contract with the Golden State Warriors, I started to wonder what is the best way a potential professional athlete should take in order to achieve financial success in the National Basketball Association.

My data was collect from Here, a website named Kaggle. The Data set contains Salary and Statistics (advanced/basic) for NBA players during the 17-18 regular season. I chose this data set because I could not find data on the 2018-2019 or 2019-2020 season. However, I believe this data set will still be relevant because the data was collected not along ago.

This topic is important to me because my friend, Miye Oni, was drafted in the second round of the 2019 NBA draft. He was not as big of a star such as Zion Williamson and Ja Morant, but I do not believe that Miye's inability to be drafted in the first round will affect his potential Salary in the NBA.

My overall plan with this data is to analyze the common ways professional athletes receive contracts in the NBA and their respective salaries. Here are the 4 most common ways:

- *Cap Space*
  - A Salary Cap is the total amount of money NBA teams are allowed to use to pay its players. The salary cap for the 2017-18 season was set at $99.093 million. Jun 30, 2018. Thus, normally if a player received a contract due to Cap Space, then
- *Bird Rights*
  - Players that signed using Bird rights, formally known as Bird Exception, allows teams to exceed the salary cap to sign their own free agents at an amount up to the maximum salary.
- *Minimum Salary*
  - The minimum salary is in place in order to protect all players (especially veterans, who are rewarded with higher values on their minimum deals depending on their years of service in the league), providing them with an absolute floor amount for what they will earn from teams interested in signing them.
- *First Round pick*
  - The four-year minimum base salaries for players in this year's draft are as follows: $480,000 (Year 1), $555,000 (Year 2), $630,000 (Year 3) and $705,000 (Year 4). After four years, players are allowed to sign maximum salaries.

Jason Huang, a Professor at the University of Pennsylvania, conducted a similar study which helped me create my analysis. His work can be found here. In his paper he analyzes she association between NBA statistics and winning percentage and determine if that same relationship holds for player compensation and individual player statistics. The only difference in his investigation and my own is that I am focusing more on the way a player signed their contract and their respective salaries. Also, he focuses on the 2013-2014 and 2014-2015 NBA seasons while I only focus on 2017-2018 seasons.

Data wrangling: Cleaning up the 2017-2018 NBA Players Statistics.

The original data set contained an NBA player's name, country, salary, draft number, position, age, and their respective statistics per game. The statistics column names are abbreviated but a glossary of them can be found on this website.

I filtered my data set to only contain players from the United States because it is the only sample size I want to focus on. I also realized the column with a player's salary was a string data type, so I used the as.numeric function to make it integers. Also, Since the Salaries were very high, I used a logarithmic transformation on the salaries.

```
USA_players <- filter(mydata, NBA_Country == "USA")

USA_players <- mutate(USA_players, SALARY = sub("\\$", "", Salary),
    SALARY_change = sub("\\,", "", SALARY), SALARY_final = sub("\\,",
        "", SALARY_change), SALARY_numeric = as.numeric(SALARY_final))
```

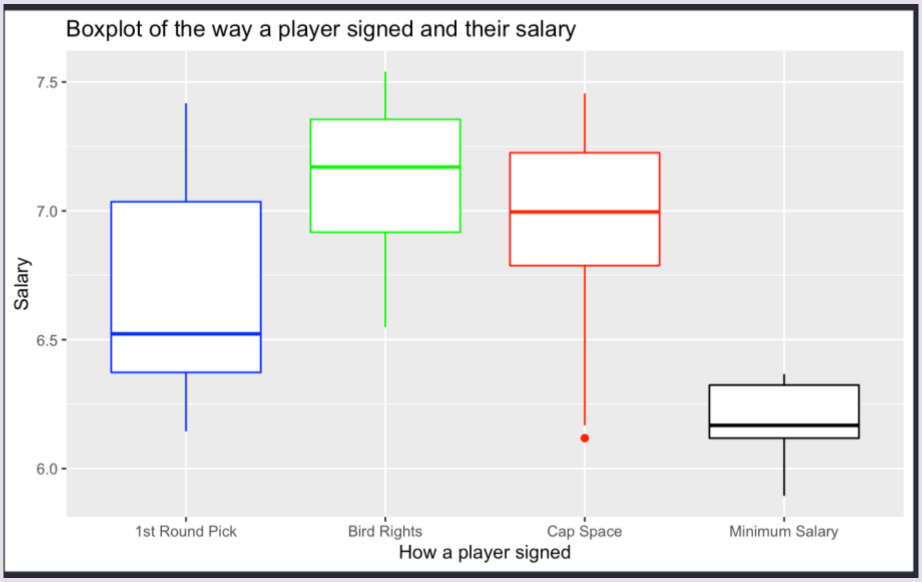
I also noticed my data set has duplicated rows. This is because throughout the season, many players were traded to other teams, so the data accounted for that. However, this is not important for my analyses. Thus, to make the data more simplified I decided to remove any duplicated players or players that are missing data in their respective salary statistics.

```
USA_players <- USA_players[!duplicated(USA_players$Player), ]  #Remove's Duplicates
USA_players <- USA_players[!(is.na(USA_players$Salary) | USA_players$Salary ==
    ""), ]  #Removes ones without salary
USA_players <- USA_players[!(is.na(USA_players$Signed.Using) |
    USA_players$Signed.Using == ""), ]
```

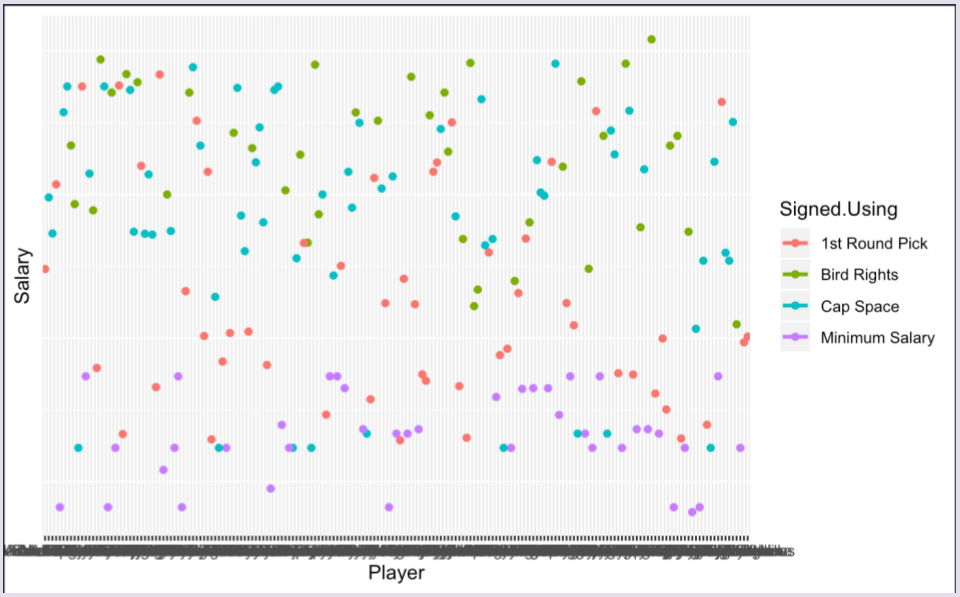
Visualize the data: Boxplots of Players' Salaries based on how they got their respective contracts

Using ggplot, I created a boxplot which shows the relationship between a player's salary and how they signed their contract (Cap Space, 1st round pick, minimum salary, bird rights). This is of interest to me because I want to see the way a basketball player signs their contract and their salary. I also wanted to see whether there is a difference in the salary ranges. I also created a scatterplot to see if there is a difference among salary versus the way they signed.

```
library(ggplot2)
USA_players_salaries_reason_to_sign <- USA_players[USA_players$Signed.Using %in%
    c("Cap Space", "1st Round Pick", "Minimum Salary", "Bird Rights"),
    ]
col = c("blue", "green", "red", "black")

ggplot(USA_players_salaries_reason_to_sign, aes(x = Signed.Using,
    y = log_Salary, )) + geom_boxplot(col = col) + ggtitle("Boxplot of the way a player s
igned and their salary") +
    xlab("How a player signed") + ylab("Salary")
```



```
ggplot(USA_players_salaries_reason_to_sign, aes(SALARY_numeric,
Player, col = Signed.Using)) + geom_point(apha = 0.5) + geom_smooth(method = "lm",
   se = FALSE) + xlab(" Player") + ylab("Salary") +
theme(axis.text.y = element_blank(),
   axis.ticks.y = element_blank())
```



The boxplot of the data shows that there is difference in the range of salaries between how a player signed. It seems as though players who signed using "bird rights" have a longer range and median of salaries. Therefore, I believe there is a difference of salaries. But to prove this I will run a hypothesis tests for more than two means.

The scatterplot of the data shows that there is variation among a player's salary and how they signed their contract among players who signed by. However, it seems as though players who signed using minimum salary relatively have the same salary, which makes sense because it is the minimum salary.

With these Visualizations, it allows me to discover any differences between the mean salaries of the groups.

Analyses: Analyzing Differences between groups, Significant Predictors, and Correlation

Hypothesis test for differences in Salaries between groups

I am borrowing code that was used to get the MAD statistic from homework 5, for which the code can for get_group_means and get_MAD_stat. My Null, Alternative, and Significance Level are below:

Null: There is no difference in the mean salaries for each way to sign a contract. Mean(salary of 1st Round Picks) = Mean(salary of Bird Rights players) = Mean(salary of Cap Space players) = Mean(salary of Minimum Salary players)

Alternative: At least 1 pair of the mean's salaries are different.

In Symbols: $\mu_i = \mu_j$

NULL: $H_0: \mu_{1stRoundPick} = \mu_{BirdRights} = \mu_{CapSpace} = \mu_{MinimumSalary}$

ALTERNATIVE: $H_A: \mu_{1stRoundPick} = \mu_{BirdRights} = \mu_{CapSpace} = \mu_{MinimumSalary}$ must be false. At least one of these equatilites are not true.

The Significance Level

Significance level is $\alpha = 0.05$. If the significance level falls below this, I will REJECT my null hypothesis.

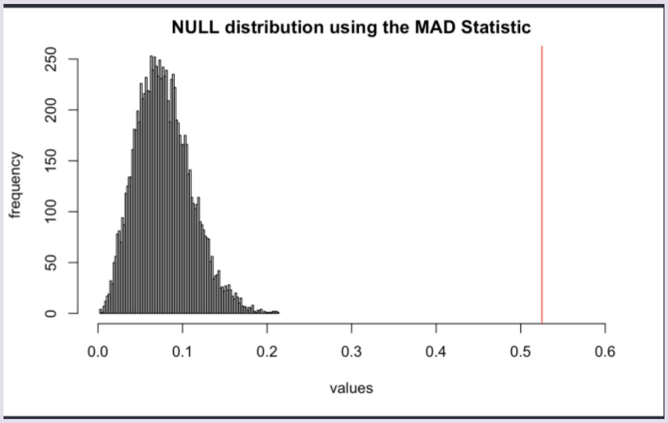
For this analysis, I use the MAD statistic to compare the mean Salaries between the different methods players signed as the observed statistic.

```
# store the Salary and How they signed in objects
player_salary_table <- USA_players_salaries_reason_to_sign$SALARY_numeric
# get the mean salary for each way a player got signed (For table below)

player_salary <- USA_players_salaries_reason_to_sign$log_Salary
player_signing <- USA_players_salaries_reason_to_sign$Signed.Using

# get the mean salary for each way a player got signed
group_means <- get_group_means(player_salary, player_signing)
group_means
## [1] 6.653110 7.124224 6.915170 6.161104
# Calculate the MAD statistic
obs_stat <- get_MAD_stat(group_means)
obs_stat
## [1] 0.525237
```

How the player signed	Mean Salary
1st Round Pick	2910078
Bird Rights	2909117
Cap Space	2911725
Minimum Salary	2911725

I then simulated steps 2-5 of the hypothesis test. I plotted the null distribution along with a red vertical line at the real MAD statistic value. I will also report the p-value below.

```
null_dist <- NULL  #Which will create the new MAD Statistics.

for (i in 1:10000) {
    new_signings <- sample(player_signing)
    new_means <- get_group_means(player_salary, new_signings)
    null_dist[i] <- get_MAD_stat(new_means)
}


# plot the null distribution with a red vertical line for the
# statistic value
hist(null_dist, nclass = 100, main = "NULL distribution using the MAD stat",
    xlab = "values", ylab = "frequency", xlim = c(0, 20000000))

abline(v = obs_stat, col = "red")
```



```
(p_value <- sum(null_dist >= obs_stat)/length(null_dist))
## [1] 0
```

Based on this analysis comparing group means using the MAD statistic, there does appear to be a difference between the player's salary depending on the way they signed. With a P-value of 0, I can reject the null hypothesis, which would make my Alternative Hypothesis (that there is a difference between the salaries of the way players signed) to be statistically significant. Overall, it is statistically safe to believe that players who sign using bird rights tend to have higher salaries.

Backwards Stepwise Regression

After discovering that, I wanted to see any other variables within the data set were statistically significant of predicting a player's salary. I used backwards stepwise regression to do this. It is a stepwise regression approach that begins with a full (saturated) model and at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data. I fitted a linear model that shows the relationship between a player's salary and their other numeric respective statistics.

The Descriptive Statistics below show that Age, Games, and Minutes Played were statistically significant predictors of a player's salary. (My entire process can be found in my appendix) However, when using Backwards Stepwise regression, I would repeat this process again but remove any predictors that are not statistically significant one by one.

```
lm_fit <- lm(log_salary ~ NBA_DraftNumber + Age + G + MP +
    WS, data = USA_players)  #Runs a linear model to determine the relationship between S
alary Numeric and the possible predictors: Age, DraftNumber, Games, Minutes Played, and w
ins
summary(lm_fit)
##
##
## Call:
## lm(formula = log_Salary ~ NBA_DraftNumber + Age + G + MP + WS,
##     data = USA_players)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63357 -0.20667  0.00854  0.22275  0.74607
##
## Coefficients:
##                   Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)      5.67624125  0.12874801  44.088 < 0.0000000000000002 ***
## NBA_DraftNumber -0.00583833  0.00102251  -5.710      0.0000000285600 ***
## Age              0.04350409  0.00458934   9.479 < 0.0000000000000002 ***
## G               -0.01149903  0.00169884  -6.769      0.0000000000747 ***
## MP               0.00041771  0.00006025   6.933      0.0000000000279 ***
## WS               0.02462928  0.01094377   2.251               0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3185 on 283 degrees of freedom
## Multiple R-squared:  0.5475, Adjusted R-squared:  0.5395
## F-statistic: 68.49 on 5 and 283 DF,  p-value: < 0.00000000000000022
```

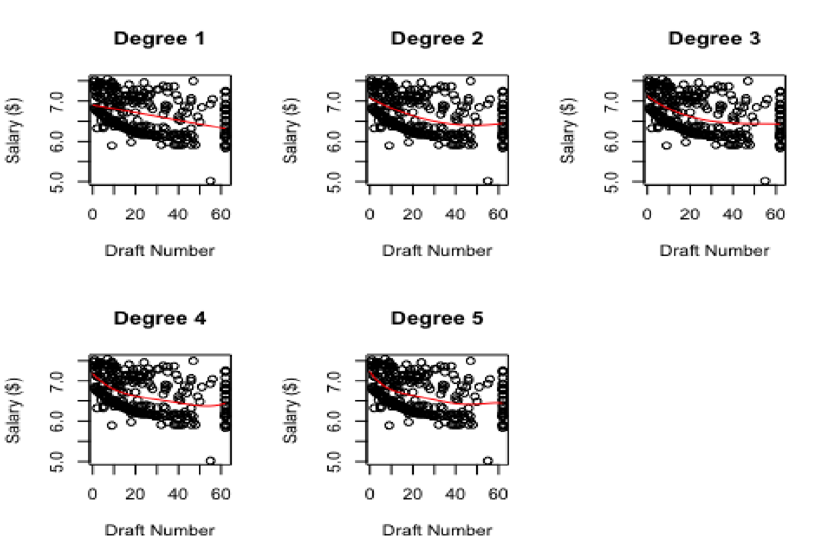
Building Linear Models

To show that the linear relationship between Salaries and the predictor variables are statistically significant, I ran a polynomial model to determine if there is a linear relationship that is statistically significant. Below I will display only display the code for NBA_DraftNumber I use to show the polynomial models but the linear models for all of the significant predictors (The rest will be in the Appendix).

```
par(mfrow = c(2, 3))
x_vals_df <- data.frame(NBA_DraftNumber = 0:62)
for (i in 1:5) {
curr_model <- lm(log_Salary ~ poly(NBA_DraftNumber, degree = i),
       data = USA_players)
model_summary <- summary(curr_model)
y_vals_predicted <- predict(curr_model, newdata = x_vals_df)


plot(SALARY_numeric ~ NBA_DraftNumber, data = USA_players,
       xlab = "Draft Number", ylab = "Salary ($)", main = paste("Degree",
           i))

points(x_vals_df$NBA_DraftNumber, y_vals_predicted, type = "l",
       col = "red")
}

(model_summary)
```

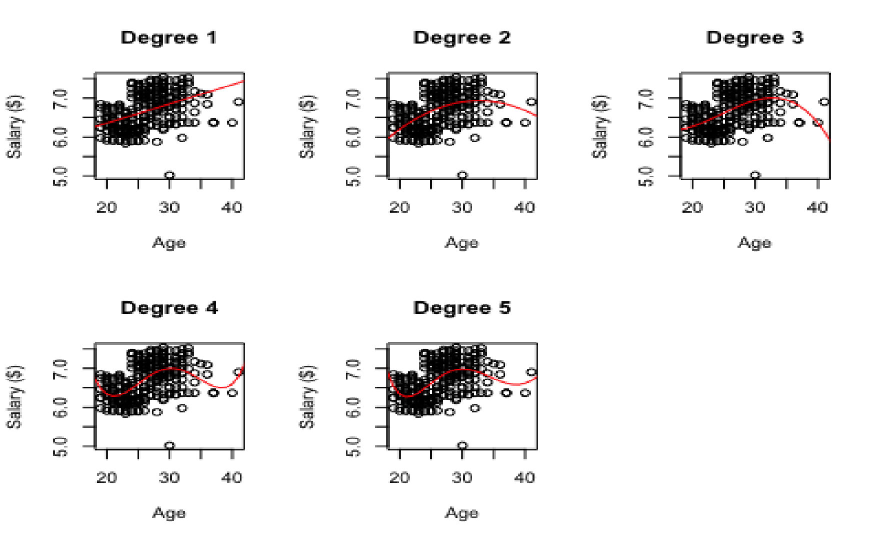
Linear Model of Draft Number and Salary

```
## Coefficients:
##                                                 Pr(>|t|)
## (Intercept)                        < 0.0000000000000002 ***
## poly(NBA_DraftNumber, degree = i)1     0.00000000000164 ***
## poly(NBA_DraftNumber, degree = i)2             0.000176 ***
## poly(NBA_DraftNumber, degree = i)3             0.393564
## poly(NBA_DraftNumber, degree = i)4             0.373934
## poly(NBA_DraftNumber, degree = i)5             0.579304
```



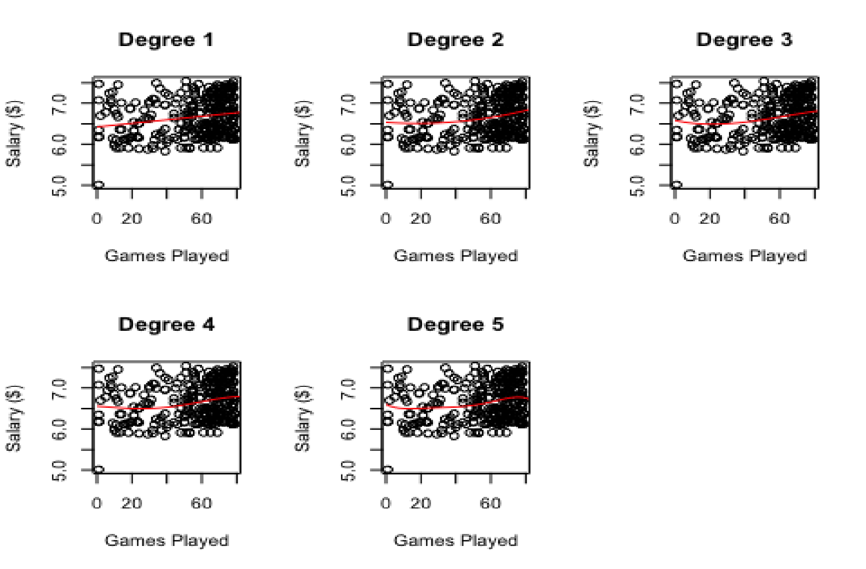
Linear Model of Age and Salary

```
(model_summary)
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)               6.6619     0.0233 285.885 < 0.0000000000000002
## poly(Age, degree = i)1    3.4877     0.3962   8.804 < 0.0000000000000002
## poly(Age, degree = i)2   -1.8262     0.3962  -4.610            0.0000061
## poly(Age, degree = i)3   -0.9738     0.3962  -2.458               0.0146
## poly(Age, degree = i)4    1.5815     0.3962   3.992            0.0000835
## poly(Age, degree = i)5   -0.3155     0.3962  -0.797               0.4264
##
## (Intercept)             ***
## poly(Age, degree = i)1 ***
## poly(Age, degree = i)2 ***
## poly(Age, degree = i)3 *
## poly(Age, degree = i)4 ***
## poly(Age, degree = i)5
```



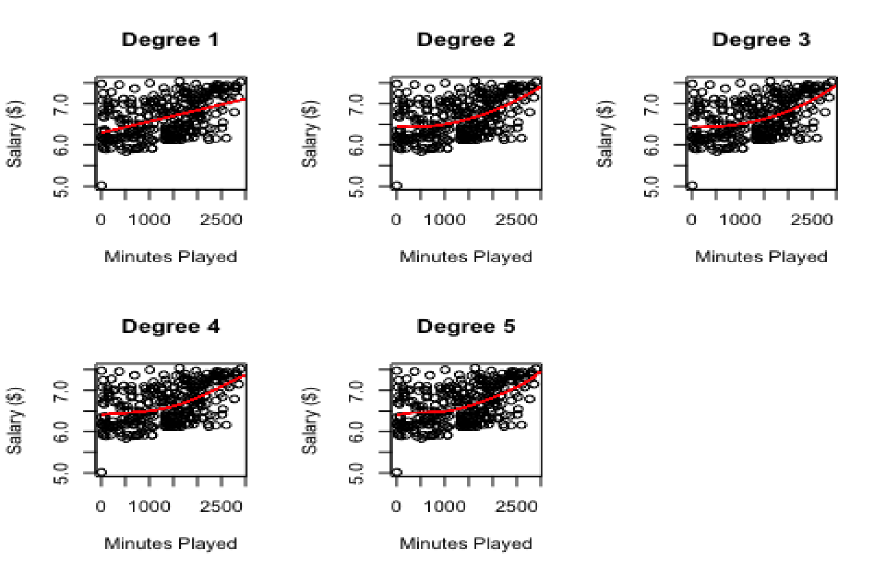
Linear Model of Games Played and Salary

```
(model_summary)
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)             6.6619     0.0272 244.941 < 0.0000000000000002 ***
## poly(G, degree = i)1    1.5824     0.4624   3.422             0.000712 ***
## poly(G, degree = i)2    0.6024     0.4624   1.303             0.193652
## poly(G, degree = i)3   -0.2045     0.4624  -0.442             0.658653
## poly(G, degree = i)4   -0.1150     0.4624  -0.249             0.803774
## poly(G, degree = i)5   -0.1940     0.4624  -0.419             0.675170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



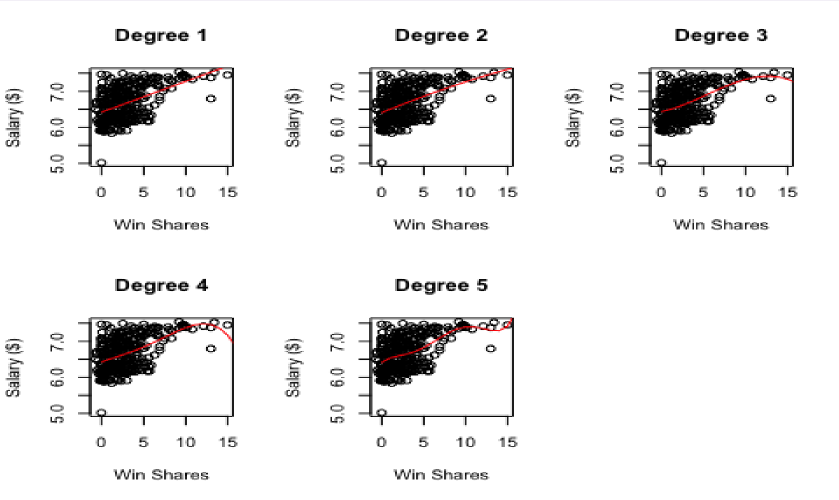
Linear Model of Minutes Played and Salary

```
(model_summary)
## Coefficients:
##                          Estimate Std. Error t value              Pr(>|t|)
## (Intercept)               6.66191    0.02435 273.551 < 0.0000000000000002 ***
## poly(MP, degree = i)1  3.61449    0.41401   8.730 0.000000000000000224 ***
## poly(MP, degree = i)2  1.35220    0.41401   3.266              0.00122 **
## poly(MP, degree = i)3  0.10381    0.41401   0.251              0.80219
## poly(MP, degree = i)4 -0.16186    0.41401  -0.391              0.69613
## poly(MP, degree = i)5  0.15134    0.41401   0.366              0.71497
```



Linear Model of Win Shares and Salary

```
(model_summary)
## Coefficients:
##                         Estimate Std. Error t value            Pr(>|t|)
## (Intercept)              6.66191    0.02383 279.541 <0.0000000000000002 ***
## poly(WS, degree = i)1  4.02541    0.40514   9.936 <0.0000000000000002 ***
## poly(WS, degree = i)2 -0.20680    0.40514  -0.510               0.610
## poly(WS, degree = i)3 -0.54578    0.40514  -1.347               0.179
## poly(WS, degree = i)4 -0.33487    0.40514  -0.827               0.409
## poly(WS, degree = i)5  0.59386    0.40514   1.466               0.144
## ---
```



The last column in each of the images above show that there is a statistically significance in the linear relationship between Salary and each of the predictor variables respectively.

This makes sense this may be because players who play more games (assuming they take some games off to rest) are less prone to injury, thus if they stay healthy throughout the season more teams are willing to sign them. Also, players that tend to be drafted late in the round have less expectations. Thus, when they surpass them, many teams may be more likely to pay them more.

Correlation between variables in groups.

After discovering which variables were statistically significant predictors of a player's salary, I tested the correlation between the Salary and the significant predictors to test if the correlation is statistically significant.

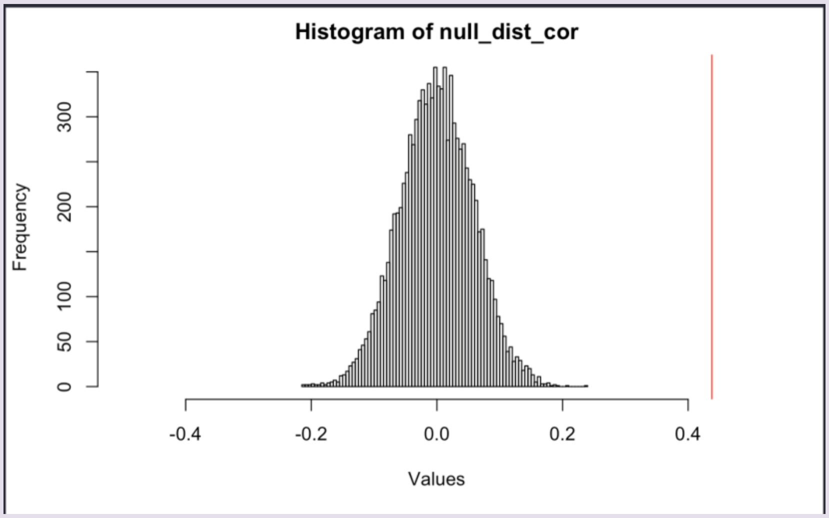
```
(obs_stat_Age <- cor(USA_players$Age, USA_players$SALARY_numeric))
## [1] 0.3863592
(obs_stat_G <- cor(USA_players$G, USA_players$SALARY_numeric))
## [1] 0.150394
(obs_stat_MP <- cor(USA_players$MP, USA_players$SALARY_numeric))
## [1] 0.429037
(obs_stat_WS <- cor(USA_players$WS, USA_players$SALARY_numeric))
```

```
## [1] 0.5816724
```

I did not include the scatterplots of these graphs because it would go above the page limit (But it will be in the Appendix). However, when I ran the correlation between a player's salary and the statistically significant predictors, it shows that there could be a positive correlation between salaries and each of the predictor variables.

I then ran a one tailed permutation test to create a null distribution of the correlation between salary and the statistically significant predictor variables. I will only display the null distribution of the correlation between WS (win shares) and salary due to page limit restrictions but the other three null distributions will be in the appendix.

```r
null_dist_cor_WS <- NULL

for (i in 1:10000) {
    shuff_salary <- sample(USA_players$log_Salary)
    null_dist_cor_WS[i] <- cor(USA_players$WS, shuff_salary)
}

hist(null_dist_cor_WS, nclass = 100, xlim = c(-0.6, 0.6), xlab = "Values",
    ylab = "Frequency")
abline(v = obs_stat_WS, col = "red")
```



```r
(p_value_WS <- sum(null_dist_cor_WS >= obs_stat_WS)/length(null_dist_cor_WS))
## [1] 0
```

With a P-Value of less than 0.05 (which was common among all of the null distributions I ran) shows that the correlation between Age and Salary, Games and Salary, Minutes Played and Salary, and Win Shares and Salary are statically significant because I can reject my null hypothesis. This means that there tends to be a linear relationship. Overall, the more games and minutes a player plays, and the older one gets, have a higher salary. This makes sense because if any player were to sign a contract using Bird rights, they would have had to be in the league for a reasonable amount of time.

Conclusion

Every professional athlete's path to getting an NBA contract varies. After running hypothesis test, I noticed that highest range of salaries for professional athletes are the ones that are signed using Bird Rights. Bird Rights, once again, is when a team is allowed to go beyond its salary cap to sign a player to its max contract.

Some of my Analyses, especially where I tested for the difference in the means of multiple groups, it showed that the difference in the salaries per the way a player signed were statistically significant. Thus, it helped me receive insight on the what is the best way to achieve economic success. My Analyses showed that you don't have to be a star coming out of high school to get paid. But a player's loyalty to a team and vice versa play a factor. This does seem a bit unusual especially since an NBA team's mentality is to win a championship, loyalty should not be much of a factor. However, my analysis shows that a player's salary tends to be higher if they sign a contract with bird rights (meaning that they have been on the same team for a while).

With this information Miye Oni should feel more comfortable about his future in the NBA. Just because he was not considered an NBA star coming out of college, his future with the Utah Jazz can be economically successful if he plays more games/minutes (by avoiding injury) or is within reasonable age to sign a contract using Bird Rights

Reflection

One Issue I encountered with this project was converting string values to integers. For example, in the salary column of my data set, each data point contained a $ beforehand, making it a string variable. This cause me to do research that can remove the dollar sign. However, even after I removed the dollar sign, the value was still a string variable, so I had to discover a way to convert the data type into integers.

I was very glad about the visualization of my data. The boxplots showed a clear difference between the salaries between groups. This made me more comfortable when I ran steps 2-5 of my hypothesis test to show that the difference in the means between the two groups were statistically significant.

In the future, I wish I could test whether categorical variables are good predictors of a player's Salary (Such as what team they are on). From there, I would probability guess that players in Los Angelo's or Texas get paid the most because the LA market is very big and there is no income tax in Texas.

Appendix:

Link to my R Code and PDF (PDF will look slightly different from this Document)