HiddenBasket: The Most Undervalued NBA Players

By: Xaiver Lim

April 3, 2020

Contents

Introduction

The purpose of this project is to create a model that will determine which basketball players in the NBA are the most undervalued based on their salaries (as of the 2019-2020 season) and performance statistics (e.g. average points, rebounds, assists, and blocks per game).

This concept of finding undervalued athletes comes from the book and movie *Moneyball*, based on a true story about the Oakland Athletics and their general manager Billy Beane who was tasked with assembling a competitive baseball team with a limited salary budget. In 2002, the Oakland Athletics had one of the lowest team payrolls in Major League Baseball which made it difficult to pay high salaries required to attract star baseball players. Thus, Beane had to come up with a creative way to form a team given these salary constraints. He decided to use statistical analysis to find and acquire undervalued baseball players. Ultimately, this method of scouting was intended to help small-market Major League Baseball teams, like Oakland, compete with larger-market teams.

The main objective of this current project is to explore whether the concept of "Moneyball" can be applied to basketball. Statistical analyses performed in this project include multiple logistic regression and discriminant analysis. Prior to the analysis, each player will be assigned to a salary tier based on their salary. Then, logistic regression will be performed to analyze which combination of performance statistics significantly predict player salaries. After determining these variables, they will be incorporated into a discriminant analysis to predict which salary tier each player should belong to based on their performance statistics. Ideally, the most undervalued players will be those who are predicted to be in a high salary tier but are actually in a low salary tier.

Packages

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(readxl)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(conflicted)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```r
options(scipen=10000)
```

```r
conflict_prefer("select", "dplyr")
```

```
## [conflicted] Will prefer dplyr::select over any other package
```

```r
conflict_prefer("filter", "dplyr")
```

```
## [conflicted] Will prefer dplyr::filter over any other package
```

```r
conflict_prefer("mutate", "plyr")
```

```
## [conflicted] Will prefer plyr::mutate over any other package
```

```r
conflict_prefer("count", "dplyr")
```

```
## [conflicted] Will prefer dplyr::count over any other package
```

```r
conflict_prefer("summarize", "plyr")
```

```
## [conflicted] Will prefer plyr::summarize over any other package
```

Data Cleansing

There are three data sets used in this project collected from *https://www.basketball-reference.com/*:

- The **Player Performance Statistics** data set presents information and performance statistics about each player such as their position (Pos), age (Age), team (Tm), games played (G), rebounds per game (REB), assists per game (AST), blocks per game (BLK), and points per game (PTS).
- The **Player Efficiency Rating** data set presents the player efficiency rating (PER) of each player.
  - PER is an advanced metric developed used in basketball to measure the overall rating of a player's per-minute statistical production. The formula for PER is quite complex but, in simple terms, it sums up all the good things a player does and subtracts negative things a player does relative to their team's style of play. To learn more about how PER is calculated: *https://www.basketball-reference.com/about/per.html*
- The **Player Salaries** data set presents the salaries of all NBA players for the 2019-2020 season.

Players who have played less than 10 games in the 2019-2020 season will be removed from the data set due to playing too few games. I have decided to set the cutoff at 10 games because any less than that, players would have played less than 15% of the season which is too small of a sample.

After removing these players from the data set, the three tables will be joined together to show each player's performance statistics and their corresponding salary.

Please note that since there are over 400 players in the NBA, only the first 10 entries of each table will be presented throughout the data cleansing portion of this report. However, a count of entries will be taken after each phase of the data cleansing to monitor changes to the number of entries.

Player Performance Statistics

```
gamePerformance <- read_excel("nba_data.xlsx", sheet = "GamePerformance")
gamePerformance %>% slice(1:10)
```

```
## # A tibble: 10 x 9
##    Player                   Pos     Age Tm        G   REB   AST   BLK   PTS
##    <chr>                    <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Steven Adams             C        26 OKC      58   9.4   2.4   1.1  10.9
##  2 Bam Adebayo              PF       22 MIA      65  10.5   5.1   1.3  16.2
##  3 LaMarcus Aldridge        C        34 SAS      53   7.4   2.4   1.6  18.9
##  4 Nickeil Alexander-Walker SG       21 NOP      41   2     1.8   0.2   5.1
##  5 Grayson Allen            SG       24 MEM      30   2.2   1.4   0     7.4
##  6 Jarrett Allen            C        21 BRK      64   9.5   1.3   1.3  10.6
##  7 Kadeem Allen             SG       27 NYK      10   0.9   2.1   0.2   5
##  8 Al-Farouq Aminu          PF       29 ORL      18   4.8   1.2   0.4   4.3
##  9 Justin Anderson          SF       26 BRK       3   0.7   0     0.3   1
## 10 Kyle Anderson            PF       26 MEM      59   4.4   2.2   0.5   5.7
```

```
count(gamePerformance)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   514
```

There are a total of 514 players in this data set.

Now let's take a look at the summary statistics of the number of games played by the players in the data set.

```
gamePerformance %>% select(G) %>% summary()
```

```
##        G
##  Min.   : 1.00
##  1st Qu.:22.00
##  Median :47.00
##  Mean   :39.89
##  3rd Qu.:58.00
##  Max.   :66.00
```

As you can see, there is quite a lot of variability in the number of games played as the most games played by a player was 66 games and the least number of games played by a player was 1.

Now, I will remove the players who have played less than 10 games.

```
gamePerformance %>% filter(G > 9) -> gamePerformance_tidy
gamePerformance_tidy %>% slice(1:10)
```

```
## # A tibble: 10 x 9
##    Player                   Pos     Age Tm        G   REB   AST   BLK   PTS
##    <chr>                    <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Steven Adams             C        26 OKC      58   9.4   2.4   1.1  10.9
##  2 Bam Adebayo              PF       22 MIA      65  10.5   5.1   1.3  16.2
##  3 LaMarcus Aldridge        C        34 SAS      53   7.4   2.4   1.6  18.9
##  4 Nickeil Alexander-Walker SG       21 NOP      41   2     1.8   0.2   5.1
##  5 Grayson Allen            SG       24 MEM      30   2.2   1.4   0     7.4
```

```
##   6 Jarrett Allen             C         21 BRK       64    9.5    1.3    1.3   10.6
##   7 Kadeem Allen              SG        27 NYK       10    0.9    2.1    0.2    5  
##   8 Al-Farouq Aminu           PF        29 ORL       18    4.8    1.2    0.4    4.3
##   9 Kyle Anderson             PF        26 MEM       59    4.4    2.2    0.5    5.7
##  10 Giannis Antetokounmpo     PF        25 MIL       57   13.7    5.8    1     29.6
```

```r
count(gamePerformance_tidy)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   446
```

After removing 68 players who played less than 10 games, 446 players remained.

Player Efficiency Rating (PER)

```r
efficiency <- read_excel("nba_data.xlsx", sheet = "EfficiencyRating")
efficiency %>% slice(1:10)
```

```
## # A tibble: 10 x 2
##    Player                    PER
##    <chr>                   <dbl>
##  1 Steven Adams             20.8
##  2 Bam Adebayo              20.6
##  3 LaMarcus Aldridge        19.8
##  4 Nickeil Alexander-Walker   7.6
##  5 Grayson Allen            11.4
##  6 Jarrett Allen            20.3
##  7 Kadeem Allen             14  
##  8 Al-Farouq Aminu           7.6
##  9 Justin Anderson          -3.8
## 10 Kyle Anderson            13  
```

Player Salaries

```r
salaries <- read_excel("nba_data.xlsx", sheet = "Salaries")
salaries %>% slice(1:10)
```

```
## # A tibble: 10 x 2
##    Player               Salary
##    <chr>                 <dbl>
##  1 Stephen Curry     40231758
##  2 Chris Paul        38506482
##  3 Russell Westbrook 38178000
##  4 James Harden      37800000
##  5 John Wall         37800000
##  6 LeBron James      37436858
##  7 Kevin Durant      37199000
##  8 Blake Griffin     34234964
##  9 Kyle Lowry        33296296
## 10 Paul George       33005556
```

PER will now be added to the data set containing all the other variables.

```r
all_stats <- inner_join(gamePerformance_tidy, efficiency, by="Player")
all_stats %>% slice(1:10)
```

```
## # A tibble: 10 x 10
##    Player                   Pos     Age Tm        G   REB   AST   BLK   PTS   PER
##    <chr>                    <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Steven Adams             C        26 OKC      58   9.4   2.4   1.1  10.9  20.8
##  2 Bam Adebayo              PF       22 MIA      65  10.5   5.1   1.3  16.2  20.6
##  3 LaMarcus Aldridge        C        34 SAS      53   7.4   2.4   1.6  18.9  19.8
##  4 Nickeil Alexander-Walk~ SG        21 NOP      41   2     1.8   0.2   5.1   7.6
##  5 Grayson Allen            SG       24 MEM      30   2.2   1.4   0     7.4  11.4
##  6 Jarrett Allen            C        21 BRK      64   9.5   1.3   1.3  10.6  20.3
##  7 Kadeem Allen             SG       27 NYK      10   0.9   2.1   0.2   5    14
##  8 Al-Farouq Aminu          PF       29 ORL      18   4.8   1.2   0.4   4.3   7.6
##  9 Kyle Anderson            PF       26 MEM      59   4.4   2.2   0.5   5.7  13
## 10 Giannis Antetokounmpo    PF       25 MIL      57  13.7   5.8   1    29.6  31.6
```

Now that PER has been added to the **Player Performance Statistics** data set, we must attach the player salaries to each corresponding player.

```r
nba <- inner_join(all_stats, salaries, by="Player")
nba %>% slice(1:10)
```

```
## # A tibble: 10 x 11
##    Player          Pos     Age Tm        G   REB   AST   BLK   PTS   PER  Salary
##    <chr>           <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
##  1 Steven Adams    C        26 OKC      58   9.4   2.4   1.1  10.9  20.8  2.58e7
##  2 Bam Adebayo     PF       22 MIA      65  10.5   5.1   1.3  16.2  20.6  3.45e6
##  3 LaMarcus Aldri~ C        34 SAS      53   7.4   2.4   1.6  18.9  19.8  2.60e7
##  4 Nickeil Alexan~ SG       21 NOP      41   2     1.8   0.2   5.1   7.6  2.96e6
##  5 Grayson Allen   SG       24 MEM      30   2.2   1.4   0     7.4  11.4  2.43e6
##  6 Jarrett Allen   C        21 BRK      64   9.5   1.3   1.3  10.6  20.3  2.38e6
##  7 Al-Farouq Aminu PF       29 ORL      18   4.8   1.2   0.4   4.3   7.6  9.26e6
##  8 Kyle Anderson   PF       26 MEM      59   4.4   2.2   0.5   5.7  13    9.07e6
##  9 Giannis Anteto~ PF       25 MIL      57  13.7   5.8   1    29.6  31.6  2.58e7
## 10 Thanasis Antet~ SF       27 MIL      18   1.1   0.5   0.1   2.5  14.2  1.45e6
```

```r
count(nba)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   410
```

After joining the three tables by player name, 410 players remain. Only players who appeared in all three data sets were included into this final compiled data set.

Defining Salary Tiers

In order to perform a logistic regression and discriminant analysis later on, an ordinal response needs to be defined. Thus, salary tiers will be created to group the players.

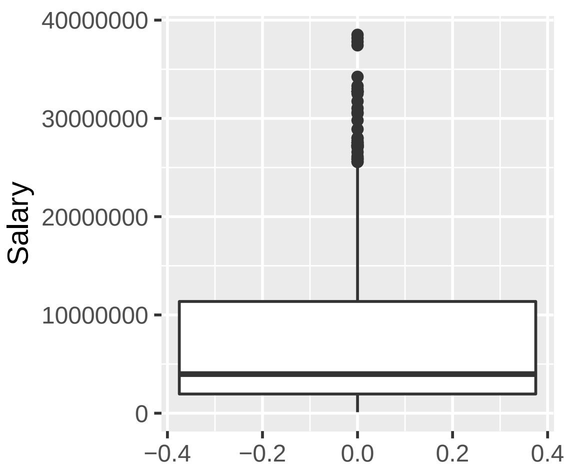
As you can see below from the summary statistics and boxplot of the player salaries:

- The 25th percentile of salaries lie approximately below $2 million
- The middle of the interquartile range (IQR) is at approximately $5 million
- The 75th percentile of salaries appears to be slightly above $10 million
- The upper outliers appear to be salaries over $25 million

```r
nba %>% select(Salary) %>% summary()
```

```
##      Salary
##  Min.   :  101504
##  1st Qu.: 1951650
##  Median : 3976460
##  Mean   : 8117626
##  3rd Qu.:11369948
##  Max.   :38506482
```

```r
ggplot(nba, aes(y = Salary)) + geom_boxplot()
```



Based on the inferences made from the summary statistics and boxplot of the player salaries, I will split up the players into the following salary tiers:

- **Tier 1**: less than $2 million
- **Tier 2**: $2 to 5 million
- **Tier 3**: $5 to 10 million
- **Tier 4**: $10 to 25 million
- **Tier 5**: more than $25 million

```r
nba %>% mutate(Tier=ifelse(Salary < 2000000,"tier1",
                    ifelse(Salary < 5000000,"tier2",
                    ifelse(Salary < 10000000,"tier3",
                    ifelse(Salary < 25000000,"tier4","tier5"))))
               ) -> nba_tiers
nba_tiers <- nba_tiers %>% mutate(Tier = factor(Tier))
nba_tiers %>% slice(1:10)
```

```
## # A tibble: 10 x 12
##    Player     Pos     Age Tm        G   REB   AST   BLK   PTS   PER Salary Tier
##    <chr>      <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <fct>
##  1 Steven Ad~ C        26 OKC      58   9.4   2.4   1.1  10.9  20.8 2.58e7 tier5
##  2 Bam Adeba~ PF       22 MIA      65  10.5   5.1   1.3  16.2  20.6 3.45e6 tier2
##  3 LaMarcus ~ C        34 SAS      53   7.4   2.4   1.6  18.9  19.8 2.60e7 tier5
##  4 Nickeil A~ SG       21 NOP      41   2     1.8   0.2   5.1   7.6 2.96e6 tier2
##  5 Grayson A~ SG       24 MEM      30   2.2   1.4   0     7.4  11.4 2.43e6 tier2
##  6 Jarrett A~ C        21 BRK      64   9.5   1.3   1.3  10.6  20.3 2.38e6 tier2
##  7 Al-Farouq~ PF       29 ORL      18   4.8   1.2   0.4   4.3   7.6 9.26e6 tier3
##  8 Kyle Ande~ PF       26 MEM      59   4.4   2.2   0.5   5.7  13   9.07e6 tier3
##  9 Giannis A~ PF       25 MIL      57  13.7   5.8   1    29.6  31.6 2.58e7 tier5
## 10 Thanasis ~ SF       27 MIL      18   1.1   0.5   0.1   2.5  14.2 1.45e6 tier1
```

Now each row of this data set consists of each player and their performance statistics, salary, and salary tier.

Data Analysis

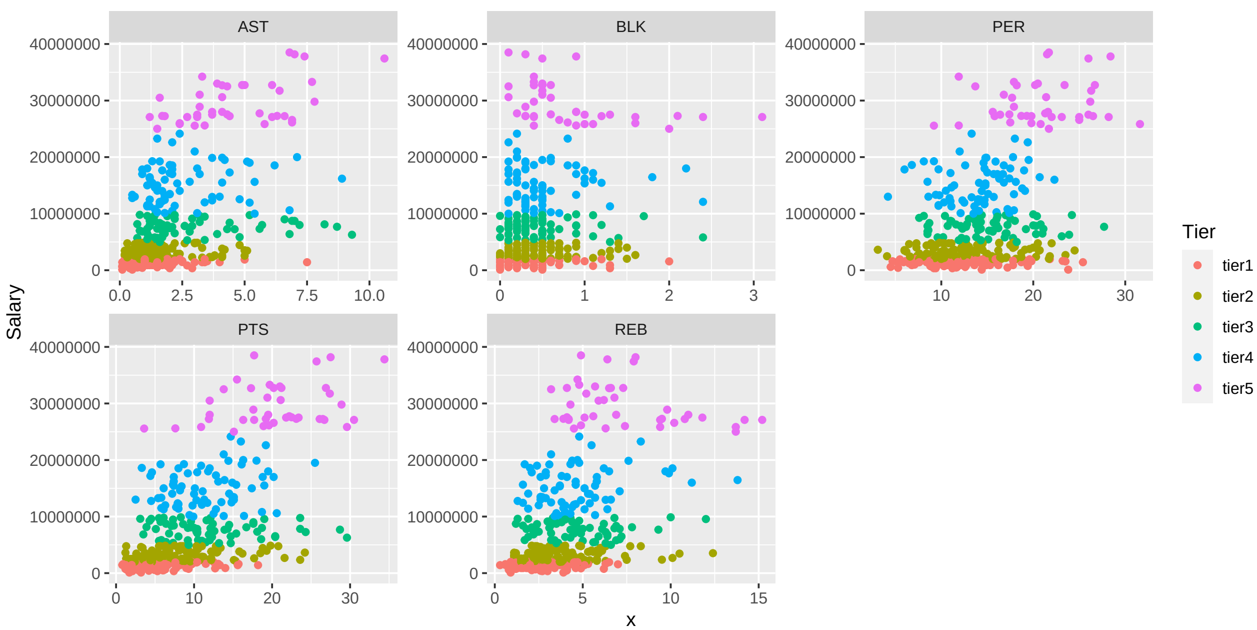
Correlation Between Variables

Before creating a regression model, let's first take a look at the correlation between salary (response variable) and the possible predictor (independent) variables.

```
nba_tiers %>% select(Salary, REB, AST, BLK, PTS, PER) %>% cor() -> COR
COR[1,]
```

```
##    Salary       REB       AST       BLK       PTS       PER
## 1.0000000 0.4844740 0.5328877 0.2606822 0.6312167 0.4667826
```

```
nba_tiers %>%
  pivot_longer((c(REB, AST, BLK, PTS, PER)), names_to = "xname", values_to = "x") %>%
  ggplot(aes(x = x, y = Salary, colour=Tier)) + geom_point() +
  facet_wrap(~xname, scales = "free") -> g1
g1
```



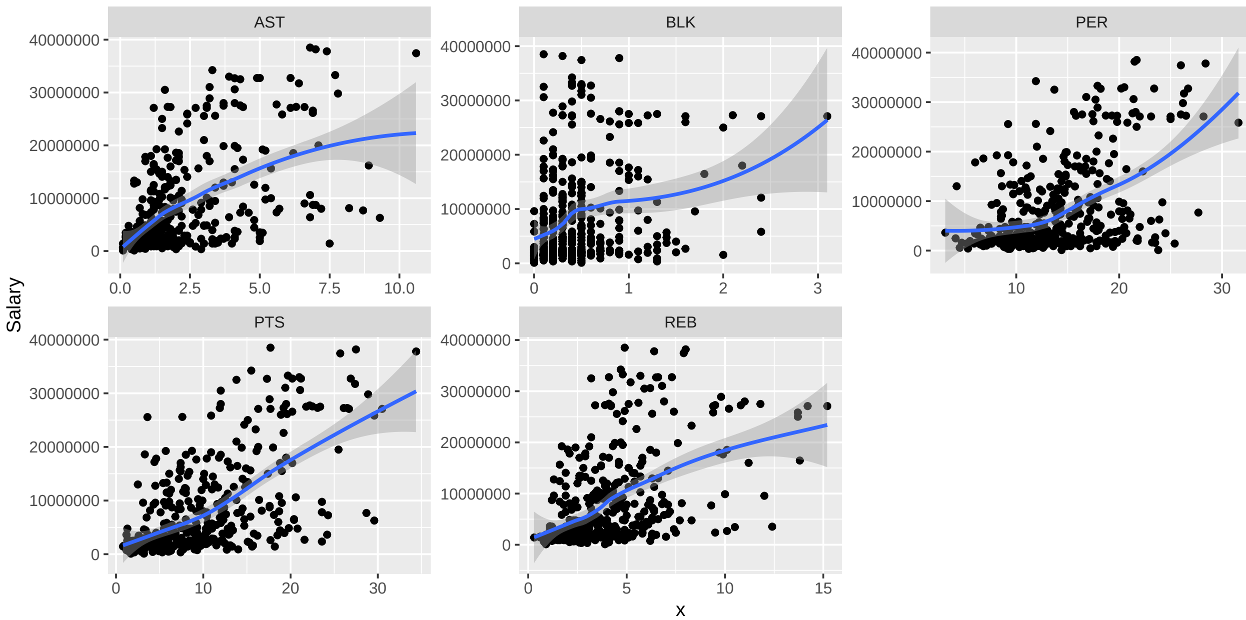
As you can see from the correlations and scatter plots, the direction of all correlations are positive. This means a higher salary is associated with scoring many points and having a high player efficiency rating. In addition, a higher salary is associated with getting a lot of assists, blocks, and rebounds.

However, it is important to note that these relationships vary in terms of correlation strength:

- Salary is strongly correlated with Points Per Game (PTS)
- Salary is moderately correlated with Rebounds Per Game (REB), Assists Per Game (AST), and Player Efficiency Rating (PER)
- Salary is weakly associated with Blocks Per Game (BLK)

```
nba_tiers %>%
  pivot_longer(c(PTS, REB, AST, BLK, PER), names_to = "xname", values_to = "x") %>%
  ggplot(aes(x = x, y = Salary)) + geom_point() + geom_smooth() +
  facet_wrap(~xname, scales = "free") -> g2
g2
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Logistic Regression

Multiple logistic reggression uses predictor variables to model the probability of a certain outcome occurring or a subject belonging to a specific group. The multiple logistic regression model below predicts a player's salary tier based on their average rebounds, assists, blocks, and points per game, along with player efficiency rating (PER). Since the salary tier is an ordered categorical variable (tier 1 to tier 5), an ordered logistic model will be created using *polr*.

```
nba.1 <- polr(Tier ~ REB + AST + BLK + PTS + PER , data = nba_tiers)
drop1(nba.1, test="Chisq")
```

```
## Single term deletions
##
## Model:
## Tier ~ REB + AST + BLK + PTS + PER
##        Df    AIC    LRT       Pr(>Chi)
## <none>    1056.5
## REB     1 1066.6 12.176       0.000484 ***
## AST     1 1070.1 15.683 0.000074875131 ***
## BLK     1 1060.0  5.574       0.018229 *
## PTS     1 1088.3 33.881 0.000000005859 ***
## PER     1 1062.7  8.274       0.004022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the regression model show that REB, AST, BLK, PTS, and PER are all significant as they have p-values of less than 0.05. Thus, all the explanatory variables have some impact on the salary tier, and none of them should be removed from the model.

```
nba.1$coefficients
```

```
##         REB         AST         BLK         PTS         PER
##  0.22406618  0.29930215  0.73864764  0.15879231 -0.08871348
```

With the exception of PER, all of the coefficients are positive which means the model predicts an increase in REB, AST, BLK, or PTS will result in an increase in salary. The model also predicts an increase in PER will result in an decrease in salary.

Now these variables will be incorporated into a discriminant analysis.

Discriminant Analysis

A discriminant analysis predicts group membership based on numeous factors (measured variables), assuming the groups are known. This type of analysis can be performed to predict a player's salary tier based on their performance statistics.

```
salaries.1 <- lda(Tier ~ REB + AST + BLK + PTS + PER, data = nba_tiers)
salaries.1
```

```
## Call:
## lda(Tier ~ REB + AST + BLK + PTS + PER, data = nba_tiers)
##
## Prior probabilities of groups:
##     tier1     tier2     tier3     tier4     tier5
## 0.2634146 0.3048780 0.1585366 0.1731707 0.1000000
##
## Group means:
##            REB      AST       BLK       PTS      PER
## tier1 2.692593 1.275000 0.3240741  5.810185 12.17500
## tier2 3.649600 1.592000 0.4176000  8.046400 12.65520
## tier3 4.550769 2.867692 0.4769231 11.926154 15.10923
## tier4 4.642254 2.594366 0.5197183 11.587324 13.99577
## tier5 7.260976 4.587805 0.8000000 20.297561 20.74634
##
## Coefficients of linear discriminants:
##             LD1         LD2         LD3         LD4
## REB  0.15550187 -0.26570594  0.09212061  0.49804037
## AST  0.25870481  0.07649206 -0.68417575 -0.11827077
## BLK  0.56530753 -0.41081322 -0.06268459 -3.26444824
## PTS  0.12120380 -0.10393985  0.09459629 -0.04657632
## PER -0.03129032  0.31306455  0.07317913  0.04133825
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4
## 0.9482 0.0420 0.0087 0.0011
```

As you can see from looking at the group means, the higher the salary tier, the greater the REB, AST, BLK, PTS, and PER. However, there are a few exceptions to this general trend, especially when looking at tier 4.

The number of linear discriminants is either the number of variables or number of groups - 1, depending on which value is smaller. Since there are 5 variables (REB, AST, BLK, PTS, and PER) and 5 groups (tier 1 to 5), there are 4 linear discrimnants (LD1 to LD4). Each linear discriminant is a linear combination of features (REB, AST, BLK, PTS, and PER) that characterizes a certain group of player.

Now focusing on the proportion of trace, LD1 makes up most of the proportion of trace (0.9482). Thus, we should primarily be focused on LD1.

Moving onto the coefficients of the linear discriminanats, LD1 is positive when REB, AST, BLK, and PTS are high since their coefficients under LD1 are positive.

Finally, since the LD1 coefficient of PER (-0.03129032) is close to 0, PER will have close to no impact on the model.

Now, this model will be used to predict which salary tier each player belongs to.

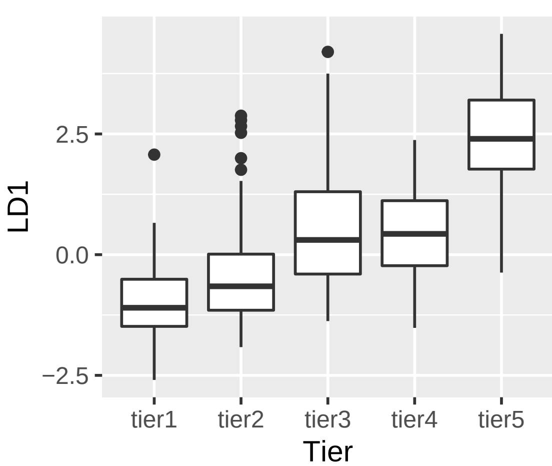
```
salaries.pred <- predict(salaries.1)
pp <- round(salaries.pred$posterior, 3)
predictions <- cbind(nba_tiers, pp)
predictions %>% slice(1:10)
```

```
##                       Player Pos Age  Tm  G  REB AST BLK  PTS  PER   Salary
## 1               Steven Adams   C  26 OKC 58  9.4 2.4 1.1 10.9 20.8 25842697
## 2                Bam Adebayo  PF  22 MIA 65 10.5 5.1 1.3 16.2 20.6  3454080
## 3          LaMarcus Aldridge   C  34 SAS 53  7.4 2.4 1.6 18.9 19.8 26000000
## 4  Nickeil Alexander-Walker  SG  21 NOP 41  2.0 1.8 0.2  5.1  7.6  2964840
## 5              Grayson Allen  SG  24 MEM 30  2.2 1.4 0.0  7.4 11.4  2429400
## 6              Jarrett Allen   C  21 BRK 64  9.5 1.3 1.3 10.6 20.3  2376840
## 7            Al-Farouq Aminu  PF  29 ORL 18  4.8 1.2 0.4  4.3  7.6  9258000
## 8              Kyle Anderson  PF  26 MEM 59  4.4 2.2 0.5  5.7 13.0  9073050
## 9      Giannis Antetokounmpo  PF  25 MIL 57 13.7 5.8 1.0 29.6 31.6 25842697
## 10   Thanasis Antetokounmpo  SF  27 MIL 18  1.1 0.5 0.1  2.5 14.2  1445697
##     Tier tier1 tier2 tier3 tier4 tier5
## 1  tier5 0.073 0.232 0.293 0.261 0.141
## 2  tier2 0.002 0.017 0.109 0.118 0.754
## 3  tier5 0.011 0.070 0.148 0.237 0.534
## 4  tier2 0.341 0.425 0.095 0.139 0.000
## 5  tier2 0.408 0.410 0.091 0.092 0.001
## 6  tier2 0.085 0.293 0.247 0.275 0.099
## 7  tier3 0.221 0.475 0.105 0.198 0.001
## 8  tier3 0.297 0.386 0.155 0.158 0.003
## 9  tier5 0.000 0.000 0.003 0.002 0.996
## 10 tier1 0.716 0.240 0.027 0.017 0.000
```

I have combined player statistics, actual player salary tier, and tier predictions into one table.

Now let's take a look at side-by-side boxplots to see the relationship between LD1 and the salary tiers.

```
tierLD <- cbind(nba_tiers, salaries.pred$x, pp)
ggplot(tierLD, aes(x = Tier, y = LD1)) + geom_boxplot()
```



Since LD1 is positive when REB, AST, BLK, and PTS are high and these stats are associated with a greater salary tier, higher tiers have a greater LD1 score.

Statistical Inferences

Below is a frequency table comparing each player's actual salary tier (obs) to their predicted salary tier (pred).

```
table(obs = nba_tiers$Tier, pred = salaries.pred$class)
```

```
##        pred
## obs     tier1 tier2 tier3 tier4 tier5
##   tier1    60    43     1     3     1
##   tier2    39    69     4     7     6
##   tier3    11    32     9     4     9
##   tier4     2    39    10    15     5
##   tier5     0     3     3     4    31
```

The most undervalued player would be located in the cell with an observed (acutal) tier of tier 1 and a predicted tier of tier 5. This player makes less than $2 million but are predicted to play at the level of someone who makes more than $25 million.

```
data.frame(nba$Player, obs = nba_tiers$Tier, pred = salaries.pred$class) %>%
  filter(obs == "tier1", pred == "tier5") -> pool1
left_join(pool1, nba, by = c("nba.Player" = "Player")) -> Pool1
```

```
## Warning: Column `nba.Player`/`Player` joining factor and character vector,
## coercing into character vector
```

```
Pool1
```

```
##        nba.Player   obs  pred Pos Age  Tm  G REB AST BLK  PTS  PER  Salary
## 1 Devonte' Graham tier1 tier5  PG  24 CHO 63 3.4 7.5 0.2 18.2 15.8 1416852
```

Devonte' Graham is the player in tier 1 who was predicted to be in tier 5. Let's compare his performance statistics to the league average.

League Average

```
nba_tiers %>% summarize(REB = mean(REB), AST = mean(AST), BLK = mean(BLK),
                        PTS = mean(PTS), PER = mean(PER), Salary = mean(Salary)
                        ) -> nba_avg
league_avg <- round(nba_avg, 1) %>% mutate(nba.Player = "NBA Average")
comparison <- rbind.fill(Pool1, league_avg)
comparison
```

```
##        nba.Player   obs  pred  Pos Age   Tm  G REB AST BLK  PTS  PER  Salary
## 1 Devonte' Graham tier1 tier5   PG  24  CHO 63 3.4 7.5 0.2 18.2 15.8 1416852
## 2     NBA Average  <NA>  <NA> <NA>  NA <NA> NA 4.1 2.2 0.5  9.9 14.0 8117626
```

Based on the model, it makes sense that Devonte' Graham was selected since compared to the league average, he scores a lot of points (PTS) and gets a lot of assists (AST). He is also close to the league average in rebounds (REB), blocks (BLK), and player efficiency rating (PER).

Players who are in salary tier 2 but were predicted to be in salary tier 5 are also heavily undervalued and may be considered the next-most undervalued group of players.

```
data.frame(nba$Player, obs = nba_tiers$Tier, pred = salaries.pred$class) %>%
  filter(obs == "tier2", pred == "tier5") -> pool2
left_join(pool2, nba, by = c("nba.Player" = "Player")) -> Pool2
```

```
## Warning: Column `nba.Player`/`Player` joining factor and character vector,
## coercing into character vector
```

```
comparison2 <- rbind.fill(Pool2, league_avg)
comparison2
```

```
##                nba.Player   obs  pred  Pos Age   Tm  G  REB AST BLK  PTS  PER
## 1             Bam Adebayo tier2 tier5   PF  22  MIA 65 10.5 5.1 1.3 16.2 20.6
## 2            John Collins tier2 tier5   PF  22  ATL 41 10.1 1.5 1.6 21.6 23.5
## 3 Shai Gilgeous-Alexander tier2 tier5   SG  21  OKC 63  6.1 3.3 0.7 19.3 17.8
## 4        Donovan Mitchell tier2 tier5   SG  23  UTA 63  4.4 4.2 0.2 24.2 19.1
## 5        Domantas Sabonis tier2 tier5    C  23  IND 62 12.4 5.0 0.5 18.5 20.7
## 6           Pascal Siakam tier2 tier5   PF  25  TOR 53  7.5 3.6 0.9 23.6 18.7
## 7             NBA Average  <NA>  <NA> <NA>  NA <NA> NA  4.1 2.2 0.5  9.9 14.0
##    Salary
## 1 3454080
## 2 2686560
## 3 3952920
## 4 3635760
## 5 3529555
## 6 2351839
## 7 8117626
```

This is the pool of players in tier 2 but who play as if they are in tier 5. Based on the model, it makes sense that these players were selected since compared to the league average, they score a lot of points (PTS), get a lot of rebounds (REB), assists (AST), and blocks (BLK). They also have high player efficiency ratings (PER).

Finally, to further show how undervalued each of these players are, let's take a look at the (posterior) probability of each of these players being in each tier according to the model.

```
tier_predictions <- predictions %>% select(c(Player, tier1, tier2, tier3, tier4, tier5))
left_join(pool2, tier_predictions, by = c("nba.Player" = "Player"))
```

```
## Warning: Column `nba.Player`/`Player` joining factor and character vector,
## coercing into character vector
```

```
##                nba.Player   obs  pred tier1 tier2 tier3 tier4 tier5
## 1             Bam Adebayo tier2 tier5 0.002 0.017 0.109 0.118 0.754
## 2            John Collins tier2 tier5 0.003 0.034 0.076 0.114 0.773
## 3 Shai Gilgeous-Alexander tier2 tier5 0.023 0.122 0.243 0.293 0.319
## 4        Donovan Mitchell tier2 tier5 0.016 0.076 0.226 0.201 0.480
## 5        Domantas Sabonis tier2 tier5 0.001 0.018 0.109 0.103 0.768
## 6           Pascal Siakam tier2 tier5 0.003 0.028 0.099 0.151 0.719
```

According to the model predictions, there was over a 70% chance that Bam Adebayo, John Collins, Domantas Sabonis, and Pascal Siakam would belong to tier 5. This further supports the notion that these players are severely undervalued based on the parameters of the model.

Conclusion

Correlation Between Variables

Salary has a:

- strong positive correlation with Points Per Game
- moderate positive correlation with Rebounds, Assists, and Steals Per Game
- moderate positive correlation with Player Efficiency Rating (PER)
- weak positive correlation with Blocks Per Game.

It intuitively makes sense that points per game is strongly correlated with salary since the primary objective of basketball is to score, so players who score a lot of points should be paid the most. Also, it intuitively makes sense that defensive statistics such as blocks are not as strongly correlated to salary as defense is often overlooked by many teams.

Logistic Regression

The results of the regression model show that REB, AST, BLK, PTS, and PER all have some impact on salary tier. In addition, with the exception of PER, all of the coefficients are positive which means the model predicts an increase in REB, AST, BLK, or PTS will result in an increase in salary.

One may assume that elite players who get paid a lot should be efficient and thus the coefficient for player efficiency rating (PER) should be positive rather than negative. However, since elite players possess the ball more, they have more opportunities to make "mistakes" resulting in inefficiencies. Thus, it is plausible that the coefficient of PER is negative.

Discriminant Analysis

Based on my model, there are numerous players who are undervalued. Specifically, Devonte' Graham, Bam Adebayo, John Collins, Shai Gilgeous-Alexander, Donovan Mitchell, Domantas Sabonis, and Pascal Siakam. The interesting thing is 4 of these 7 players have been named to the 2019-2020 All-Star Team which supports the notion that these players are playing at an elite level but are being underpaid/undervalued.

The model could have also been used to predict the most overvalued players by looking at players in a high salary tier who are predicted to be in a low salary tier. However, I wanted to focus on determining the most undervalued players because I believe it is more important for NBA teams to consider. In conclusion, this model can be used by teams to generate a short list of players to target in free agency.