CAI Check-In

Dylan Blechner (with a lot of help from my dad)

December 11, 2016

So Far

So far, we have successfully read in ten years of NBA data from basketball-reference.com. My dataset now includes player data for each year including Career totals, Playoff data, and Career Playoff totals data. Each player data row also includes the Salary for the year following the data. This is so I can try to understand what factors from this year contribute to the salary the player gets paid in the following year.

Grabbing the data has been one of the more difficult aspects of the project so far and has taken up most of our time.

R Coding Script Examples

NBA Dataset for 2016

Here I show the first nine rows and ten columns of the data table, using the **select** command:

```
NBAData2016 %>% select(1:9) %>% tbl_df
```

```
## # A tibble: 366 × 9
##      No.          Player  Pos.x       Ht    Wt         Birth.Date     X.
##    <int>          <fctr> <fctr>    <dbl> <int>             <fctr> <fctr>
## 1      0    Aaron Brooks     PG 6.000000   161   January 14, 1985     us
## 2      0    Aaron Gordon     SF 6.750000   220 September 16, 1995     us
## 3      9  Aaron Harrison     SG 6.500000   210   October 28, 1994     us
## 4     33   Adreian Payne     PF 6.833333   237  February 19, 1991     us
## 5      9   Alan Anderson     SF 6.500000   220   October 16, 1982     us
## 6     15   Alan Williams     PF 6.666667   260   January 28, 1993     us
## 7     10      Alec Burks     SG 6.500000   214      July 20, 1991     us
## 8     42   Alexis Ajinca      C 7.166667   248        May 6, 1988     fr
## 9     21        Alex Len      C 7.083333   260      June 16, 1993     ua
## 10     8 Al-Farouq Aminu     SF 6.750000   220 September 21, 1990     us
## # ... with 356 more rows, and 2 more variables: Exp <int>, College <fctr>
```

The 2016 data set has 366 rows and 111 columns.

Regression Analysis of the 2016 NBA Data by Player Position

Here is an example script I am working on to calculate a multiple linear regression of Salary vs. many other statistics.

```
Player_Salary16 <-
  read.csv("/home/steve/R/R Projects/NBA Salaries BBRef/Data/2016BBRef.csv", check.names = FALSE)

#Sample Regression for Each Position#
#PG#
subset(Player_Salary16, Pos.x == "PG")
subset(Player_Salary16, Pos.x == "PG") -> PG
PGMod <- lm(data = PG, Salary ~ `Ht` + `Wt` + `Age` + `G` + `GS` + `MP` + `FG` + `FGA` +
              `FG%` + `3P` + `3PA` + `3P%` + `2P` + `2PA` + `2P%` + `eFG%` + `FT` + `FTA` +
              `FT%` + `ORB` + `DRB` + `TRB` + `AST` + `STL` + `BLK` + `TOV` + `PF` + `PTS` +
              `CareerG` + `CareerGS` + `CareerMP` + `CareerFG` + `CareerFGA` + `Career3P` +
              `Career3PA` + `Career2P` + `Career2PA` + `CareerFT` + `CareerFTA` + `CareerORB` +
              `CareerDRB` + `CareerTRB` + `CareerAST` + `CareerSTL` + `CareerBLK` + `CareerTOV` +
              `CareerPF` + `CareerPTS` + `PlayoffG` + `PlayoffGS` + `PlayoffMP` + `PlayoffFG` +
              `PlayoffFGA` + `PlayoffFG%` + `Playoff3P` + `Playoff3PA` + `Playoff3P%` +
              `Playoff2P` + `Playoff2PA` + `Playoff2P%` + `PlayoffeFG%` + `PlayoffFT` +
              `PlayoffFTA` + `PlayoffFT%` + `PlayoffORB` + `PlayoffDRB` + `PlayoffTRB` +
              `PlayoffAST` + `PlayoffSTL` + `PlayoffBLK` + `PlayoffTOV` + `PlayoffPF` +
              `PlayoffPTS` + `CareerPlayoffG` + `CareerPlayoffGS` + `CareerPlayoffMP` +
              `CareerPlayoffFG` + `CareerPlayoffFGA` + `CareerPlayoff3P` + `CareerPlayoff3PA` +
              `CareerPlayoff2P` + `CareerPlayoff2PA` + `CareerPlayoffFT` + `CareerPlayoffFTA` +
              `CareerPlayoffORB` + `CareerPlayoffDRB` + `CareerPlayoffTRB` + `CareerPlayoffAST` +
              `CareerPlayoffSTL` + `CareerPlayoffBLK` + `CareerPlayoffTOV` + `CareerPlayoffPF` +
              `CareerPlayoffPTS`)
summary(PGMod)
anova(PGMod)

#SG#
subset(Player_Salary16, Pos.x == "SG")
subset(Player_Salary16, Pos.x == "SG") -> SG
SGMod <- lm(data = SG, Salary ~ `Ht` + `Wt` + `Age` + `G` + `GS` + `MP` + `FG` + `FGA` +
              `FG%` + `3P` + `3PA` + `3P%` + `2P` + `2PA` + `2P%` + `eFG%` + `FT` + `FTA` +
              `FT%` + `ORB` + `DRB` + `TRB` + `AST` + `STL` + `BLK` + `TOV` + `PF` + `PTS` +
              `CareerG` + `CareerGS` + `CareerMP` + `CareerFG` + `CareerFGA` + `Career3P` +
              `Career3PA` + `Career2P` + `Career2PA` + `CareerFT` + `CareerFTA` + `CareerORB` +
              `CareerDRB` + `CareerTRB` + `CareerAST` + `CareerSTL` + `CareerBLK` + `CareerTOV` +
              `CareerPF` + `CareerPTS` + `PlayoffG` + `PlayoffGS` + `PlayoffMP` + `PlayoffFG` +
              `PlayoffFGA` + `PlayoffFG%` + `Playoff3P` + `Playoff3PA` + `Playoff3P%` +
              `Playoff2P` + `Playoff2PA` + `Playoff2P%` + `PlayoffeFG%` + `PlayoffFT` +
              `PlayoffFTA` + `PlayoffFT%` + `PlayoffORB` + `PlayoffDRB` + `PlayoffTRB` +
              `PlayoffAST` + `PlayoffSTL` + `PlayoffBLK` + `PlayoffTOV` + `PlayoffPF` +
              `PlayoffPTS` + `CareerPlayoffG` + `CareerPlayoffGS` + `CareerPlayoffMP` +
              `CareerPlayoffFG` + `CareerPlayoffFGA` + `CareerPlayoff3P` + `CareerPlayoff3PA` +
              `CareerPlayoff2P` + `CareerPlayoff2PA` + `CareerPlayoffFT` + `CareerPlayoffFTA` +
              `CareerPlayoffORB` + `CareerPlayoffDRB` + `CareerPlayoffTRB` + `CareerPlayoffAST` +
              `CareerPlayoffSTL` + `CareerPlayoffBLK` + `CareerPlayoffTOV` + `CareerPlayoffPF` +
              `CareerPlayoffPTS`)
summary(SGMod)
anova(SGMod)
```

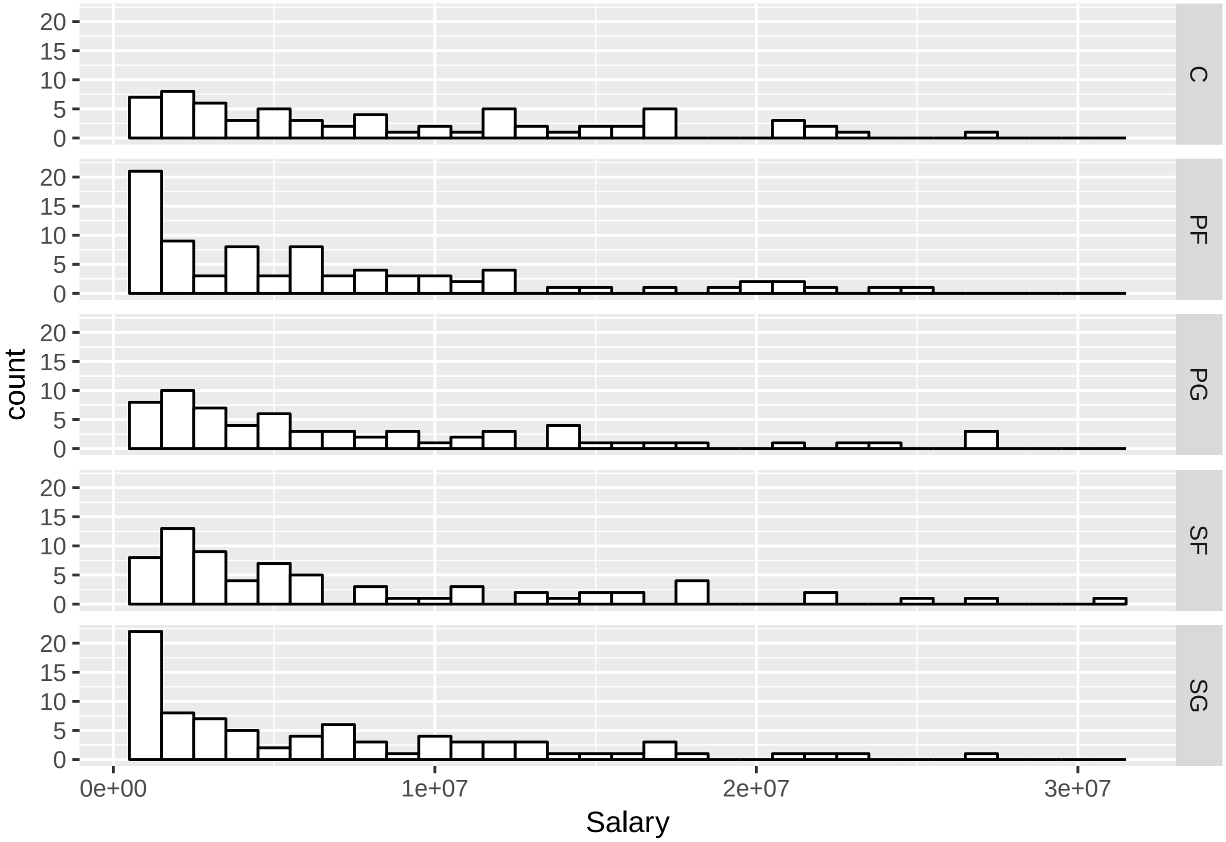
```r
#SF#
subset(Player_Salary16, Pos.x == "SF")
subset(Player_Salary16, Pos.x == "SF") -> SF
SFMod <- lm(data = SF, Salary ~ `Ht` + `Wt` + `Age` + `G` + `GS` + `MP` + `FG` + `FGA` +
              `FG%` + `3P` + `3PA` + `3P%` + `2P` + `2PA` + `2P%` + `eFG%` + `FT` + `FTA` +
              `FT%` + `ORB` + `DRB` + `TRB` + `AST` + `STL` + `BLK` + `TOV` + `PF` + `PTS` +
              `CareerG` + `CareerGS` + `CareerMP` + `CareerFG` + `CareerFGA` + `Career3P` +
              `Career3PA` + `Career2P` + `Career2PA` + `CareerFT` + `CareerFTA` + `CareerORB` +
              `CareerDRB` + `CareerTRB` + `CareerAST` + `CareerSTL` + `CareerBLK` + `CareerTOV` +
              `CareerPF` + `CareerPTS` + `PlayoffG` + `PlayoffGS` + `PlayoffMP` + `PlayoffFG` +
              `PlayoffFGA` + `PlayoffFG%` + `Playoff3P` + `Playoff3PA` + `Playoff3P%` +
              `Playoff2P` + `Playoff2PA` + `Playoff2P%` + `PlayoffeFG%` + `PlayoffFT` +
              `PlayoffFTA` + `PlayoffFT%` + `PlayoffORB` + `PlayoffDRB` + `PlayoffTRB` +
              `PlayoffAST` + `PlayoffSTL` + `PlayoffBLK` + `PlayoffTOV` + `PlayoffPF` +
              `PlayoffPTS` + `CareerPlayoffG` + `CareerPlayoffGS` + `CareerPlayoffMP` +
              `CareerPlayoffFG` + `CareerPlayoffFGA` + `CareerPlayoff3P` + `CareerPlayoff3PA` +
              `CareerPlayoff2P` + `CareerPlayoff2PA` + `CareerPlayoffFT` + `CareerPlayoffFTA` +
              `CareerPlayoffORB` + `CareerPlayoffDRB` + `CareerPlayoffTRB` + `CareerPlayoffAST` +
              `CareerPlayoffSTL` + `CareerPlayoffBLK` + `CareerPlayoffTOV` + `CareerPlayoffPF` +
              `CareerPlayoffPTS`)
summary(SFMod)
anova(SFMod)

#PF#
subset(Player_Salary16, Pos.x == "PF")
subset(Player_Salary16, Pos.x == "PF") -> PF
PFMod <- lm(data = PF, Salary ~ `Ht` + `Wt` + `Age` + `G` + `GS` + `MP` + `FG` + `FGA` +
              `FG%` + `3P` + `3PA` + `3P%` + `2P` + `2PA` + `2P%` + `eFG%` + `FT` + `FTA` +
              `FT%` + `ORB` + `DRB` + `TRB` + `AST` + `STL` + `BLK` + `TOV` + `PF` + `PTS` +
              `CareerG` + `CareerGS` + `CareerMP` + `CareerFG` + `CareerFGA` + `Career3P` +
              `Career3PA` + `Career2P` + `Career2PA` + `CareerFT` + `CareerFTA` + `CareerORB` +
              `CareerDRB` + `CareerTRB` + `CareerAST` + `CareerSTL` + `CareerBLK` + `CareerTOV` +
              `CareerPF` + `CareerPTS` + `PlayoffG` + `PlayoffGS` + `PlayoffMP` + `PlayoffFG` +
              `PlayoffFGA` + `PlayoffFG%` + `Playoff3P` + `Playoff3PA` + `Playoff3P%` +
              `Playoff2P` + `Playoff2PA` + `Playoff2P%` + `PlayoffeFG%` + `PlayoffFT` +
              `PlayoffFTA` + `PlayoffFT%` + `PlayoffORB` + `PlayoffDRB` + `PlayoffTRB` +
              `PlayoffAST` + `PlayoffSTL` + `PlayoffBLK` + `PlayoffTOV` + `PlayoffPF` +
              `PlayoffPTS` + `CareerPlayoffG` + `CareerPlayoffGS` + `CareerPlayoffMP` +
              `CareerPlayoffFG` + `CareerPlayoffFGA` + `CareerPlayoff3P` + `CareerPlayoff3PA` +
              `CareerPlayoff2P` + `CareerPlayoff2PA` + `CareerPlayoffFT` + `CareerPlayoffFTA` +
              `CareerPlayoffORB` + `CareerPlayoffDRB` + `CareerPlayoffTRB` + `CareerPlayoffAST` +
              `CareerPlayoffSTL` + `CareerPlayoffBLK` + `CareerPlayoffTOV` + `CareerPlayoffPF` +
              `CareerPlayoffPTS`)
summary(PFMod)
anova(PFMod)

#C#
subset(Player_Salary16, Pos.x == "C")
subset(Player_Salary16, Pos.x == "C") -> C
CMod <- lm(data = C, Salary ~ `Ht` + `Wt` + `Age` + `G` + `GS` + `MP` + `FG` + `FGA` +
              `FG%` + `3P` + `3PA` + `3P%` + `2P` + `2PA` + `2P%` + `eFG%` + `FT` + `FTA` +
              `FT%` + `ORB` + `DRB` + `TRB` + `AST` + `STL` + `BLK` + `TOV` + `PF` + `PTS` +
              `CareerG` + `CareerGS` + `CareerMP` + `CareerFG` + `CareerFGA` + `Career3P` +
              `Career3PA` + `Career2P` + `Career2PA` + `CareerFT` + `CareerFTA` + `CareerORB` +
              `CareerDRB` + `CareerTRB` + `CareerAST` + `CareerSTL` + `CareerBLK` + `CareerTOV` +
              `CareerPF` + `CareerPTS` + `PlayoffG` + `PlayoffGS` + `PlayoffMP` + `PlayoffFG` +
              `PlayoffFGA` + `PlayoffFG%` + `Playoff3P` + `Playoff3PA` + `Playoff3P%` +
              `Playoff2P` + `Playoff2PA` + `Playoff2P%` + `PlayoffeFG%` + `PlayoffFT` +
              `PlayoffFTA` + `PlayoffFT%` + `PlayoffORB` + `PlayoffDRB` + `PlayoffTRB` +
              `PlayoffAST` + `PlayoffSTL` + `PlayoffBLK` + `PlayoffTOV` + `PlayoffPF` +
              `PlayoffPTS` + `CareerPlayoffG` + `CareerPlayoffGS` + `CareerPlayoffMP` +
              `CareerPlayoffFG` + `CareerPlayoffFGA` + `CareerPlayoff3P` + `CareerPlayoff3PA` +
              `CareerPlayoff2P` + `CareerPlayoff2PA` + `CareerPlayoffFT` + `CareerPlayoffFTA` +
              `CareerPlayoffORB` + `CareerPlayoffDRB` + `CareerPlayoffTRB` + `CareerPlayoffAST` +
              `CareerPlayoffSTL` + `CareerPlayoffBLK` + `CareerPlayoffTOV` + `CareerPlayoffPF` +
              `CareerPlayoffPTS`)
summary(CMod)
anova(CMod)
```

R Plot Examples of 2016 NBA Data

A histogram of salaries for the 2016 NBA Season, for each of the 5 positions:

- C = Center
- PF = Power Forward
- PG = Point Guard
- SF = Small Forward
- SG = Shooting Guard

The width of each salary bin is $1,000,000.



For the 2016-2017 season, the salary distribution for each position is skewed towards lower salaries, with outliers at the higher end. For example, LeBron James, a small forward, is the extreme value for the above **SF** plot at ~ $31 million.

I can see that by taking the 2016 dataset, then filtering it by the player, LeBron James, then selecting only the columns of interest (Player name, Postion, Salary):

```
NBAData2016 %>% filter(Player=="LeBron James") %>% select(Player, Pos.x, Salary)
```

```
##         Player Pos.x   Salary
## 1 LeBron James    SF 30963450
```