CSP 571 Course Project

Basketball Salaries Team

Load, Clean, and Link Data

Load NBA 2K Data

Note: Primary dataset is directly downloaded from Kaggle. This video-game rankings dataset is scraped from http://mtdb.com/20

```
library(stringr)
library(rvest)
library(tidyr)
if (!file.exists('data/raw/nba2k/nba2k_16.csv')){ # only run if data is not already scraped
# constants
root <- 'data/raw/nba2k'
years <- c(16,17,18,19,20)
pages = c(84,68,72,68,46)
url_f <- 'http://mtdb.com/%d?page=%d&sortedBy=overall&sortOrder=Descending&'
for (i in 1:length(years)){
  year_df <- vector('list',12)
  names(year_df) <- c('name','position','ovr','out','ins','pla','ath','def','reb','xbox','ps4','pc')
  year <- years[i]
  page <- pages[i]
  for (page in 1:page){
    # load webpage
    url <- sprintf(url_f,year,page)
    webpage <- read_html(url)
    # load salary table
    player_tables <- html_nodes(webpage, css = 'table')
    player_df_page <- html_table(player_tables[[1]])#[-(1),]
    names(player_df_page) <- c('name','position','ovr','out','ins','pla','ath','def','reb','xbox','ps4','pc')
    year_df <- rbind(year_df,player_df_page)}
  write.csv(year_df,sprintf('%s/nba2k_%d.csv',root,year))
  cat(sprintf('%d nrows: %d\n',year,nrow(year_df)))}}
```

Clean Primary Dataset

```
library("readxl")
df_primary <- read_excel('data/raw/primary_dataset_raw.xlsx')
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in D24626 / R24626C4: got 'z'
```

```
df_primary <- df_primary[,!(names(df_primary)%in%c('#','blanl','blank2'))] # drop empty/non-stat columns
colnames(df_primary)[1:3] <- c('year','name_p','salary')
df_primary <- df_primary[!is.na(df_primary[['salary']]),] # drop rows with no salaryes
df_primary[is.na(df_primary)] <- 0
df_primary <- df_primary[df_primary$year%in%c(2016:2020),]   # take 2016-2017 player data
head(df_primary)
```

```
## # A tibble: 6 x 51
##    year name_p salary Pos     Age Tm        G    GS    MP   PER `TS%` `3PAr`
##   <dbl> <chr>   <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
```

```
## 1  2017 A.J. ~ 1.31e6 C        24 DAL      22     0   163   8.4 0.472  0.238
## 2  2016 Aaron~ 2.70e6 PG       31 CHI      69     0  1108  11.8 0.494  0.394
## 3  2017 Aaron~ 2.12e6 PG       32 IND      65     0   894   9.5 0.507  0.427
## 4  2016 Aaron~ 4.35e6 PF       20 ORL      78    37  1863  17   0.541  0.245
## 5  2017 Aaron~ 5.50e6 SF       21 ORL      80    72  2298  14.4 0.53   0.309
## 6  2016 Aaron~ 3.76e5 SG       21 CHO      21     0    93   4.3 0.371  0.526
## # ... with 39 more variables: FTr <dbl>, `ORB%` <dbl>, `DRB%` <dbl>,
## #   `TRB%` <dbl>, `AST%` <dbl>, `STL%` <dbl>, `BLK%` <dbl>, `TOV%` <dbl>,
## #   `USG%` <dbl>, OWS <dbl>, DWS <dbl>, WS <dbl>, `WS/48` <dbl>, OBPM <dbl>,
## #   DBPM <dbl>, BPM <dbl>, VORP <dbl>, FG <dbl>, FGA <dbl>, `FG%` <dbl>,
## #   `3P` <dbl>, `3PA` <dbl>, `3P%` <dbl>, `2P` <dbl>, `2PA` <dbl>, `2P%` <dbl>,
## #   `eFG%` <dbl>, FT <dbl>, FTA <dbl>, `FT%` <dbl>, ORB <dbl>, DRB <dbl>,
## #   TRB <dbl>, AST <dbl>, STL <dbl>, BLK <dbl>, TOV <dbl>, PF <dbl>, PTS <dbl>
```

```
summary(df_primary)
```

```
##       year         name_p              salary             Pos
##  Min.   :2016   Length:965         Min.   :   11534   Length:965
##  1st Qu.:2016   Class :character   1st Qu.: 1551659   Class :character
##  Median :2017   Mode  :character   Median : 4000000   Mode  :character
##  Mean   :2017                      Mean   : 6789399
##  3rd Qu.:2017                      3rd Qu.:10500000
##  Max.   :2017                      Max.   :34682550
##       Age             Tm                  G               GS
##  Min.   :19.00   Length:965         Min.   : 1.00   Min.   : 0.00
##  1st Qu.:23.00   Class :character   1st Qu.:32.00   1st Qu.: 1.00
##  Median :26.00   Mode  :character   Median :61.00   Median :12.00
##  Mean   :26.48                      Mean   :53.41   Mean   :25.99
##  3rd Qu.:29.00                      3rd Qu.:75.00   3rd Qu.:52.00
##  Max.   :40.00                      Max.   :82.00   Max.   :82.00
##        MP            PER              TS%              3PAr
##  Min.   :   1   Min.   :-35.30   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 496   1st Qu.: 10.50   1st Qu.:0.5040   1st Qu.:0.1360
##  Median :1197   Median : 13.30   Median :0.5380   Median :0.3110
##  Mean   :1247   Mean   : 13.61   Mean   :0.5324   Mean   :0.3045
##  3rd Qu.:1954   3rd Qu.: 16.30   3rd Qu.:0.5710   3rd Qu.:0.4470
##  Max.   :3125   Max.   : 39.30   Max.   :1.0000   Max.   :1.0000
##       FTr              ORB%             DRB%            TRB%
##  Min.   :0.0000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:0.1670   1st Qu.: 1.900   1st Qu.:10.30   1st Qu.: 6.200
##  Median :0.2400   Median : 3.300   Median :14.00   Median : 8.800
##  Mean   :0.2682   Mean   : 4.868   Mean   :15.13   Mean   : 9.992
##  3rd Qu.:0.3380   3rd Qu.: 7.100   3rd Qu.:19.20   3rd Qu.:13.100
##  Max.   :2.0000   Max.   :27.300   Max.   :39.20   Max.   :30.300
##       AST%             STL%             BLK%             TOV%
##  Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 7.00   1st Qu.: 1.100   1st Qu.: 0.500   1st Qu.: 9.90
##  Median :10.40   Median : 1.500   Median : 1.200   Median :12.50
##  Mean   :13.38   Mean   : 1.583   Mean   : 1.652   Mean   :12.82
##  3rd Qu.:17.80   3rd Qu.: 1.900   3rd Qu.: 2.300   3rd Qu.:15.20
##  Max.   :72.30   Max.   :11.100   Max.   :15.100   Max.   :43.60
##       USG%             OWS              DWS              WS
##  Min.   : 0.00   Min.   :-3.300   Min.   :0.000   Min.   :-2.10
##  1st Qu.:15.30   1st Qu.: 0.100   1st Qu.:0.400   1st Qu.: 0.50
##  Median :18.40   Median : 0.800   Median :1.000   Median : 1.80
##  Mean   :18.85   Mean   : 1.387   Mean   :1.272   Mean   : 2.66
##  3rd Qu.:21.80   3rd Qu.: 2.100   3rd Qu.:1.900   3rd Qu.: 3.80
##  Max.   :41.70   Max.   :13.800   Max.   :6.000   Max.   :17.90
##      WS/48               OBPM               DBPM              BPM
##  Min.   :-0.28300   Min.   :-17.3000   Min.   :-8.5000   Min.   :-24.100
##  1st Qu.: 0.05000   1st Qu.: -2.4000   1st Qu.:-1.5000   1st Qu.: -3.100
##  Median : 0.08700   Median : -0.9000   Median :-0.3000   Median : -1.200
```

```
##  Mean   : 0.08683   Mean   : -0.9566   Mean   :-0.2671   Mean   : -1.225
##  3rd Qu.: 0.12100   3rd Qu.:  0.4000   3rd Qu.: 1.0000   3rd Qu.:  0.700
##  Max.   : 0.63400   Max.   : 15.3000   Max.   :12.0000   Max.   : 15.600
##       VORP               FG              FGA               FG%
##  Min.   :-1.4000   Min.   :  0.0   Min.   :   0.0   Min.   :0.0000
##  1st Qu.:-0.1000   1st Qu.: 62.0   1st Qu.: 146.0   1st Qu.:0.4050
##  Median : 0.2000   Median :166.0   Median : 368.0   Median :0.4410
##  Mean   : 0.6493   Mean   :200.8   Mean   : 441.5   Mean   :0.4463
##  3rd Qu.: 1.0000   3rd Qu.:294.0   3rd Qu.: 644.0   3rd Qu.:0.4810
##  Max.   :12.4000   Max.   :824.0   Max.   :1941.0   Max.   :1.0000
##        3P               3PA              3P%               2P
##  Min.   :  0.00   Min.   :  0.0   Min.   :0.0000   Min.   :  0
##  1st Qu.:  3.00   1st Qu.: 12.0   1st Qu.:0.2450   1st Qu.: 43
##  Median : 30.00   Median : 92.0   Median :0.3330   Median :113
##  Mean   : 47.83   Mean   :133.8   Mean   :0.2846   Mean   :153
##  3rd Qu.: 77.00   3rd Qu.:215.0   3rd Qu.:0.3750   3rd Qu.:219
##  Max.   :402.00   Max.   :886.0   Max.   :1.0000   Max.   :730
##       2PA              2P%              eFG%              FT
##  Min.   :   0.0   Min.   :0.0000   Min.   :0.0000   Min.   :  0.00
##  1st Qu.:  93.0   1st Qu.:0.4460   1st Qu.:0.4670   1st Qu.: 23.00
##  Median : 235.0   Median :0.4830   Median :0.5010   Median : 59.00
##  Mean   : 307.8   Mean   :0.4837   Mean   :0.4986   Mean   : 92.23
##  3rd Qu.: 444.0   3rd Qu.:0.5290   3rd Qu.:0.5360   3rd Qu.:120.00
##  Max.   :1421.0   Max.   :1.0000   Max.   :1.0000   Max.   :746.00
##       FTA              FT%              ORB               DRB
##  Min.   :  0.0   Min.   :0.0000   Min.   :  0.00   Min.   :  0
##  1st Qu.: 33.0   1st Qu.:0.6740   1st Qu.: 13.00   1st Qu.: 62
##  Median : 78.0   Median :0.7640   Median : 33.00   Median :143
##  Mean   :120.3   Mean   :0.7305   Mean   : 52.69   Mean   :173
##  3rd Qu.:161.0   3rd Qu.:0.8310   3rd Qu.: 70.00   3rd Qu.:243
##  Max.   :881.0   Max.   :1.0000   Max.   :395.00   Max.   :817
##       TRB              AST              STL               BLK
##  Min.   :   0.0   Min.   :  0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  79.0   1st Qu.: 30.0   1st Qu.: 14.00   1st Qu.:  5.00
##  Median : 178.0   Median : 74.0   Median : 33.00   Median : 15.00
##  Mean   : 225.7   Mean   :115.5   Mean   : 40.02   Mean   : 25.03
##  3rd Qu.: 307.0   3rd Qu.:151.0   3rd Qu.: 58.00   3rd Qu.: 33.00
##  Max.   :1198.0   Max.   :906.0   Max.   :169.00   Max.   :269.00
##       TOV               PF              PTS
##  Min.   :  0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.: 25.00   1st Qu.: 47.0   1st Qu.: 166.0
##  Median : 57.00   Median :102.0   Median : 437.0
##  Mean   : 70.13   Mean   :103.4   Mean   : 541.8
##  3rd Qu.: 99.00   3rd Qu.:152.0   3rd Qu.: 780.0
##  Max.   :464.00   Max.   :278.0   Max.   :2558.0
```

Pool Together and Clean NBA 2K Data (Secondary Dataset)

```
secondary_attriutes <- c('name_s','position_s','ovr','out','ins','pla','ath','def','reb')
df_secondary <- vector('list',9)
names(df_secondary) <- secondary_attriutes
path_f = 'data/raw/nba2k/nba2k_%d.csv'
for (year in c(16:20)){
  df_year <- read.csv(sprintf(path_f,year))
  headers <- names(df_year)
  names(df_year) <- c('drop1',headers[1:length(headers)-1])
  df_year <- df_year[,c('name','position','ovr','out','ins','pla','ath','def','reb')]
  names(df_year) <- secondary_attriutes
  df_year[,'year'] <- 2000+year
  df_secondary <- rbind(df_secondary,df_year)}
```

```
df_secondary[is.na(df_secondary)] <- 0
df_secondary <- df_secondary[df_secondary$year%in%c(2016,2017),]   # take 2016-2017 2K ratings data
head(df_secondary)
```

```
##                    name_s position_s ovr out ins pla ath def reb year
## 1      '96 Michael Jordan         SG  99  95  88  91  93  92  75 2016
## 2         '15 Kobe Bryant         SG  99  97  79  95  84  88  65 2016
## 3           Stephen Curry         PG  99  98  66  98  89  78  54 2016
## 4            LeBron James         SF  99  94  89  91  92  91  91 2016
## 5 '71 Kareem Adbul-Jabbar          C  99  75  93  56  89  86  98 2016
## 6            Kyrie Irving         PG  98  98  70  95  91  74  49 2016
```

```
summary(df_secondary)
```

```
##                name_s       position_s       ovr             out
##  Jimmy Butler     :  10   PG     :812   Min.   :40.00   Min.   :25.0
##  Kyrie Irving     :  10   SF     :782   1st Qu.:71.00   1st Qu.:62.0
##  Russell Westbrook:  10   SG     :749   Median :78.00   Median :73.0
##  Damian Lillard   :   9   PF     :710   Mean   :78.89   Mean   :71.3
##  Demar Derozan    :   9   C      :708   3rd Qu.:86.00   3rd Qu.:82.0
##  James Harden     :   9   C/PF   :  0   Max.   :99.00   Max.   :99.0
##  (Other)          :3704   (Other):  0
##       ins             pla             ath             def
##  Min.   :25.00   Min.   :25.00   Min.   :25.00   Min.   :25.00
##  1st Qu.:58.00   1st Qu.:48.00   1st Qu.:68.00   1st Qu.:58.00
##  Median :64.00   Median :61.00   Median :74.00   Median :65.00
##  Mean   :65.43   Mean   :62.04   Mean   :73.68   Mean   :66.28
##  3rd Qu.:72.00   3rd Qu.:76.00   3rd Qu.:80.00   3rd Qu.:73.00
##  Max.   :98.00   Max.   :99.00   Max.   :98.00   Max.   :98.00
##
##       reb             year
##  Min.   :25.00   Min.   :2016
##  1st Qu.:43.00   1st Qu.:2016
##  Median :57.00   Median :2016
##  Mean   :59.62   Mean   :2016
##  3rd Qu.:75.00   3rd Qu.:2017
##  Max.   :99.00   Max.   :2017
##
```

Merge Primary and Secondary Datasets

```
library(stringr)
clean_names <- function(names){
  names <- tolower(names)
  names <- str_squish(names)
  names <- gsub('\\.','',names)
  names <- gsub('-',' ',names)
  return (names)}
df_primary$name <- clean_names(df_primary[['name_p']])
df_secondary$name <- clean_names(df_secondary[['name_s']])
# if multiple versions of a player, take the one with the max overall
df_secondary_max <- aggregate(df_secondary['ovr'],df_secondary[c('name','year')],max)
df_secondary_max <- merge(df_secondary_max,df_secondary,by=c('name','year','ovr'),all=F)
df_secondary_max_2 <- aggregate(df_secondary_max['out'],df_secondary_max[c('name','year')],max)
df_full_s <- merge(df_secondary_max,df_secondary_max_2,by=c('name','year','out'),all=F)
# only take totals from players who changed teams mid-year
df_p_tot <- df_primary[df_primary$Tm=='TOT',]
traded_player_years <- interaction(df_primary[,c('year','name')]) %in%
                       interaction(df_p_tot[,c('year','name')])
df_p_wo_tot <- df_primary[!traded_player_years,]
```

```
df_full_p <- rbind(df_p_wo_tot,df_p_tot)
# join datasets
df_full <- merge(df_full_p,df_full_s,by=c('name','year'),all=F)
df_full <- df_full[order(df_full$name,df_full$year),]
df_full <- unique(df_full)
head(df_full[,1:5])
```

```
##            name year        name_p  salary Pos
## 1  aaron brooks 2016  Aaron Brooks 2700000  PG
## 2  aaron brooks 2017  Aaron Brooks 2116955  PG
## 3  aaron gordon 2016  Aaron Gordon 4351320  PF
## 4  aaron gordon 2017  Aaron Gordon 5504420  SF
## 5 adreian payne 2016 Adreian Payne 2022240  PF
## 6    aj hammons 2017  A.J. Hammons 1312611   C
```

Clean Up Merged Data and

```
drop_cols <- c('name','name_s','position_s')
df_final <- df_full[,!(names(df_full)%in%drop_cols)]
names(df_final)[names(df_final)=='position_p'] <- 'position'
names(df_final)[names(df_final)=='name_p'] <- 'name'
s_columns <- c('ovr','out','ins','pla','ath','def','reb')
df_p_final <- df_final[,!(names(df_final)%in%s_columns)] # final primary dataset
df_s_final <- df_final[,c('name',s_columns)] # final secondary dataset
head(df_final) # final complete (combined primary and secondary) datasets
```

```
##   year          name  salary Pos Age  Tm  G GS   MP  PER   TS%  3PAr   FTr ORB%
## 1 2016  Aaron Brooks 2700000  PG  31 CHI 69  0 1108 11.8 0.494 0.394 0.136  2.0
## 2 2017  Aaron Brooks 2116955  PG  32 IND 65  0  894  9.5 0.507 0.427 0.133  2.3
## 3 2016  Aaron Gordon 4351320  PF  20 ORL 78 37 1863 17.0 0.541 0.245 0.333  9.0
## 4 2017  Aaron Gordon 5504420  SF  21 ORL 80 72 2298 14.4 0.530 0.309 0.251  5.3
## 5 2016 Adreian Payne 2022240  PF  24 MIN 52  2  486  5.6 0.422 0.221 0.179  4.8
## 6 2017  A.J. Hammons 1312611   C  24 DAL 22  0  163  8.4 0.472 0.238 0.476  5.4
##   DRB% TRB% AST% STL% BLK% TOV% USG%  OWS DWS   WS  WS/48 OBPM DBPM  BPM VORP
## 1  7.5  4.8 26.0  1.4  0.7 14.2 22.9  0.2 0.7  0.9  0.040 -0.5 -2.8 -3.3 -0.4
## 2  6.3  4.3 20.7  1.4  0.9 17.2 19.2 -0.2 0.5  0.3  0.016 -2.1 -2.6 -4.6 -0.6
## 3 21.3 15.1 10.3  1.6  2.4  9.0 17.3  3.2 2.2  5.4  0.139  0.6  1.2  1.8  1.8
## 4 14.1  9.6 10.5  1.4  1.4  8.5 20.1  2.0 1.7  3.7  0.076 -0.2 -0.4 -0.7  0.8
## 5 21.5 13.3  8.9  1.7  1.8 18.7 17.7 -0.9 0.4 -0.5 -0.047 -5.9 -0.2 -6.1 -0.5
## 6 20.9 12.8  3.8  0.3  7.2 16.4 17.6 -0.2 0.2  0.0 -0.001 -7.5  1.9 -5.6 -0.1
##    FG FGA   FG% 3P 3PA   3P%  2P 2PA   2P%  eFG%  FT FTA   FT% ORB DRB TRB AST
## 1 188 469 0.401 66 185 0.357 122 284 0.430 0.471  49  64 0.766  21  80 101 180
## 2 121 300 0.403 48 128 0.375  73 172 0.424 0.483  32  40 0.800  18  51  69 125
## 3 274 579 0.473 42 142 0.296 232 437 0.531 0.509 129 193 0.668 154 353 507 128
## 4 393 865 0.454 77 267 0.288 316 598 0.528 0.499 156 217 0.719 116 289 405 150
## 5  53 145 0.366  9  32 0.281  44 113 0.389 0.397  17  26 0.654  20  91 111  29
## 6  17  42 0.405  5  10 0.500  12  32 0.375 0.464   9  20 0.450   8  28  36   4
##   STL BLK TOV  PF  PTS out ovr ins pla ath def reb
## 1  30  10  82 132  491  79  75  52  74  77  52  36
## 2  25   9  66  93  322  87  85  51  81  82  57  37
## 3  59  55  66 153  719  87  90  91  69  86  69  87
## 4  64  40  89 172 1019  86  92  91  49  86  75  94
## 5  16  11  36  77  132  56  69  65  43  66  64  68
## 6   1  13  10  21   48  47  66  64  40  58  57  71
```

```
summary(df_final)
```

```
##       year          name               salary            Pos
##  Min.   :2016   Length:729         Min.   :   11534   Length:729
##  1st Qu.:2016   Class :character   1st Qu.: 2116955   Class :character
##  Median :2016   Mode  :character   Median : 5200000   Mode  :character
```

```
##  Mean   :2016                     Mean   : 7858289
##  3rd Qu.:2017                     3rd Qu.:12078652
##  Max.   :2017                     Max.   :34682550
##       Age             Tm                   G                GS
##  Min.   :19.00   Length:729         Min.   : 1.00   Min.   : 0.00
##  1st Qu.:23.00   Class :character   1st Qu.:52.00   1st Qu.: 3.00
##  Median :26.00   Mode  :character   Median :68.00   Median :21.00
##  Mean   :26.53                      Mean   :61.19   Mean   :31.81
##  3rd Qu.:29.00                      3rd Qu.:77.00   3rd Qu.:63.00
##  Max.   :40.00                      Max.   :82.00   Max.   :82.00
##        MP            PER              TS%              3PAr
##  Min.   :   6   Min.   :-7.70   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 854   1st Qu.:10.90   1st Qu.:0.5090   1st Qu.:0.1060
##  Median :1508   Median :13.70   Median :0.5410   Median :0.3050
##  Mean   :1476   Mean   :14.17   Mean   :0.5382   Mean   :0.2928
##  3rd Qu.:2125   3rd Qu.:16.90   3rd Qu.:0.5720   3rd Qu.:0.4420
##  Max.   :3125   Max.   :32.00   Max.   :1.0000   Max.   :0.9000
##       FTr              ORB%             DRB%            TRB%
##  Min.   :0.0000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:0.1760   1st Qu.: 2.000   1st Qu.:10.60   1st Qu.: 6.30
##  Median :0.2470   Median : 3.600   Median :14.60   Median : 9.30
##  Mean   :0.2708   Mean   : 5.076   Mean   :15.56   Mean   :10.32
##  3rd Qu.:0.3380   3rd Qu.: 7.500   3rd Qu.:19.60   3rd Qu.:13.30
##  Max.   :1.2190   Max.   :21.800   Max.   :36.30   Max.   :25.60
##       AST%            STL%             BLK%            TOV%             USG%
##  Min.   : 0.0   Min.   : 0.000   Min.   :0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 7.1   1st Qu.: 1.100   1st Qu.:0.60   1st Qu.:10.00   1st Qu.:15.40
##  Median :10.3   Median : 1.500   Median :1.20   Median :12.50   Median :18.50
##  Mean   :13.4   Mean   : 1.586   Mean   :1.74   Mean   :12.74   Mean   :19.17
##  3rd Qu.:17.7   3rd Qu.: 1.900   3rd Qu.:2.50   3rd Qu.:15.10   3rd Qu.:22.10
##  Max.   :57.3   Max.   :11.100   Max.   :9.70   Max.   :43.60   Max.   :41.70
##       OWS              DWS              WS              WS/48
##  Min.   :-3.300   Min.   :0.000   Min.   :-2.100   Min.   :-0.28300
##  1st Qu.: 0.200   1st Qu.:0.700   1st Qu.: 1.100   1st Qu.: 0.05600
##  Median : 1.100   Median :1.300   Median : 2.500   Median : 0.09100
##  Mean   : 1.717   Mean   :1.526   Mean   : 3.243   Mean   : 0.09283
##  3rd Qu.: 2.500   3rd Qu.:2.200   3rd Qu.: 4.400   3rd Qu.: 0.12700
##  Max.   :13.800   Max.   :6.000   Max.   :17.900   Max.   : 0.34300
##       OBPM              DBPM               BPM               VORP
##  Min.   :-17.300   Min.   :-8.20000   Min.   :-24.1000   Min.   :-1.400
##  1st Qu.: -2.100   1st Qu.:-1.30000   1st Qu.: -2.7000   1st Qu.:-0.100
##  Median : -0.700   Median :-0.10000   Median : -0.7000   Median : 0.400
##  Mean   : -0.673   Mean   :-0.08217   Mean   : -0.7543   Mean   : 0.837
##  3rd Qu.:  0.500   3rd Qu.: 1.10000   3rd Qu.:  1.0000   3rd Qu.: 1.300
##  Max.   : 12.400   Max.   :12.00000   Max.   : 15.6000   Max.   :12.400
##        FG              FGA              FG%               3P
##  Min.   :  0.0   Min.   :   0.0   Min.   :0.0000   Min.   :  0.00
##  1st Qu.:116.0   1st Qu.: 259.0   1st Qu.:0.4110   1st Qu.:  4.00
##  Median :208.0   Median : 462.0   Median :0.4460   Median : 42.00
##  Mean   :240.5   Mean   : 526.5   Mean   :0.4529   Mean   : 56.23
##  3rd Qu.:338.0   3rd Qu.: 731.0   3rd Qu.:0.4880   3rd Qu.: 90.00
##  Max.   :824.0   Max.   :1941.0   Max.   :1.0000   Max.   :402.00
##       3PA              3P%               2P              2PA
##  Min.   :  0.0   Min.   :0.0000   Min.   :  0.0   Min.   :   0.0
##  1st Qu.: 16.0   1st Qu.:0.2500   1st Qu.: 75.0   1st Qu.: 157.0
##  Median :120.0   Median :0.3330   Median :153.0   Median : 308.0
##  Mean   :156.9   Mean   :0.2846   Mean   :184.3   Mean   : 369.6
##  3rd Qu.:256.0   3rd Qu.:0.3730   3rd Qu.:258.0   3rd Qu.: 513.0
##  Max.   :886.0   Max.   :1.0000   Max.   :730.0   Max.   :1421.0
##       2P%              eFG%              FT              FTA
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0   Min.   :  0.0
```

```
##  1st Qu.:0.4530   1st Qu.:0.4730   1st Qu.: 38   1st Qu.: 51.0
##  Median :0.4860   Median :0.5060   Median : 80   Median :110.0
##  Mean   :0.4885   Mean   :0.5038   Mean   :111   Mean   :144.7
##  3rd Qu.:0.5310   3rd Qu.:0.5370   3rd Qu.:145   3rd Qu.:194.0
##  Max.   :1.0000   Max.   :1.0000   Max.   :746   Max.   :881.0
##       FT%              ORB              DRB              TRB
##  Min.   :0.0000   Min.   :  0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.:0.6940   1st Qu.: 21.00   1st Qu.:103.0   1st Qu.: 128.0
##  Median :0.7690   Median : 44.00   Median :181.0   Median : 231.0
##  Mean   :0.7438   Mean   : 63.55   Mean   :207.3   Mean   : 270.9
##  3rd Qu.:0.8310   3rd Qu.: 87.00   3rd Qu.:279.0   3rd Qu.: 365.0
##  Max.   :1.0000   Max.   :395.00   Max.   :817.0   Max.   :1198.0
##       AST              STL              BLK              TOV
##  Min.   :  0.0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 47.0   1st Qu.: 22.00   1st Qu.:  9.00   1st Qu.: 39.00
##  Median : 97.0   Median : 42.00   Median : 20.00   Median : 69.00
##  Mean   :137.3   Mean   : 47.42   Mean   : 30.31   Mean   : 83.28
##  3rd Qu.:176.0   3rd Qu.: 66.00   3rd Qu.: 39.00   3rd Qu.:114.00
##  Max.   :906.0   Max.   :169.00   Max.   :269.00   Max.   :464.00
##        PF               PTS              out              ovr
##  Min.   :  0.0   Min.   :   0.0   Min.   :30.00   Min.   :61.00
##  1st Qu.: 79.0   1st Qu.: 307.0   1st Qu.:60.00   1st Qu.:71.00
##  Median :125.0   Median : 544.0   Median :72.00   Median :76.00
##  Mean   :121.8   Mean   : 648.3   Mean   :71.12   Mean   :78.45
##  3rd Qu.:165.0   3rd Qu.: 894.0   3rd Qu.:82.00   3rd Qu.:85.00
##  Max.   :278.0   Max.   :2558.0   Max.   :99.00   Max.   :99.00
##       ins              pla              ath              def              reb
##  Min.   :44.00   Min.   :28.00   Min.   :49.00   Min.   :43.00   Min.   :27.0
##  1st Qu.:58.00   1st Qu.:47.00   1st Qu.:68.00   1st Qu.:58.00   1st Qu.:44.0
##  Median :64.00   Median :59.00   Median :73.00   Median :64.00   Median :59.0
##  Mean   :65.32   Mean   :61.21   Mean   :73.45   Mean   :65.46   Mean   :60.9
##  3rd Qu.:71.00   3rd Qu.:75.00   3rd Qu.:79.00   3rd Qu.:72.00   3rd Qu.:74.0
##  Max.   :97.00   Max.   :98.00   Max.   :98.00   Max.   :98.00   Max.   :98.0
```

```
# Output final complete, primary, and seconday datasets
write.csv(df_final,'data/pooled/complete.csv')
write.csv(df_p_final,'data/pooled/primary.csv')
write.csv(df_s_final,'data/pooled/secondary.csv')
```

Explore Data

Summarize Datasets

```
# primary dataset
str(df_p_final)
```

```
## 'data.frame':    729 obs. of  51 variables:
##  $ year  : num  2016 2017 2016 2017 2016 ...
##  $ name  : chr  "Aaron Brooks" "Aaron Brooks" "Aaron Gordon" "Aaron Gordon" ...
##  $ salary: num  2700000 2116955 4351320 5504420 2022240 ...
##  $ Pos   : chr  "PG" "PG" "PF" "SF" ...
##  $ Age   : num  31 32 20 21 24 24 25 26 29 30 ...
##  $ Tm    : chr  "CHI" "IND" "ORL" "ORL" ...
##  $ G     : num  69 65 78 80 52 22 82 61 82 68 ...
##  $ GS    : num  0 0 37 72 2 0 82 25 82 68 ...
##  $ MP    : num  1108 894 1863 2298 486 ...
##  $ PER   : num  11.8 9.5 17 14.4 5.6 8.4 12.7 11.3 19.4 17.7 ...
##  $ TS%   : num  0.494 0.507 0.541 0.53 0.422 0.472 0.533 0.506 0.565 0.553 ...
##  $ 3PAr  : num  0.394 0.427 0.245 0.309 0.221 0.238 0.485 0.455 0.244 0.302 ...
##  $ FTr   : num  0.136 0.133 0.333 0.251 0.179 0.476 0.217 0.292 0.123 0.169 ...
```

```
##  $ ORB%  : num  2 2.3 9 5.3 4.8 5.4 4.5 4.8 6.3 4.9 ...
##  $ DRB%  : num  7.5 6.3 21.3 14.1 21.5 20.9 18.6 23.5 18.2 18.6 ...
##  $ TRB%  : num  4.8 4.3 15.1 9.6 13.3 12.8 11.5 14.1 12.4 11.8 ...
##  $ AST%  : num  26 20.7 10.3 10.5 8.9 3.8 8.8 7.9 16.7 24.4 ...
##  $ STL%  : num  1.4 1.4 1.6 1.4 1.7 0.3 1.5 1.7 1.3 1.2 ...
##  $ BLK%  : num  0.7 0.9 2.4 1.4 1.8 7.2 1.8 2 3.6 3.3 ...
##  $ TOV%  : num  14.2 17.2 9 8.5 18.7 16.4 13.2 15.2 8.8 11.9 ...
##  $ USG%  : num  22.9 19.2 17.3 20.1 17.7 17.6 16.9 15.4 20.6 19.8 ...
##  $ OWS   : num  0.2 -0.2 3.2 2 -0.9 -0.2 1.7 -0.1 4.9 3.6 ...
##  $ DWS   : num  0.7 0.5 2.2 1.7 0.4 0.2 2.3 2 4.5 2.7 ...
##  $ WS    : num  0.9 0.3 5.4 3.7 -0.5 0 4 1.9 9.4 6.3 ...
##  $ WS/48 : num  0.04 0.016 0.139 0.076 -0.047 -0.001 0.082 0.051 0.172 0.137 ...
##  $ OBPM  : num  -0.5 -2.1 0.6 -0.2 -5.9 -7.5 -0.4 -2.3 1.5 1 ...
##  $ DBPM  : num  -2.8 -2.6 1.2 -0.4 -0.2 1.9 0.7 1.2 2.6 2.1 ...
##  $ BPM   : num  -3.3 -4.6 1.8 -0.7 -6.1 -5.6 0.2 -1.1 4.1 3.1 ...
##  $ VORP  : num  -0.4 -0.6 1.8 0.8 -0.5 -0.1 1.3 0.4 4.1 2.8 ...
##  $ FG    : num  188 121 274 393 53 17 299 183 529 379 ...
##  $ FGA   : num  469 300 579 865 145 ...
##  $ FG%   : num  0.401 0.403 0.473 0.454 0.366 0.405 0.416 0.393 0.505 0.473 ...
##  $ 3P    : num  66 48 42 77 9 5 126 70 88 86 ...
##  $ 3PA   : num  185 128 142 267 32 10 349 212 256 242 ...
##  $ 3P%   : num  0.357 0.375 0.296 0.288 0.281 0.5 0.361 0.33 0.344 0.355 ...
##  $ 2P    : num  122 73 232 316 44 12 173 113 441 293 ...
##  $ 2PA   : num  284 172 437 598 113 32 370 254 792 559 ...
##  $ 2P%   : num  0.43 0.424 0.531 0.528 0.389 0.375 0.468 0.445 0.557 0.524 ...
##  $ eFG%  : num  0.471 0.483 0.509 0.499 0.397 0.464 0.503 0.468 0.547 0.527 ...
##  $ FT    : num  49 32 129 156 17 9 115 96 103 108 ...
##  $ FTA   : num  64 40 193 217 26 20 156 136 129 135 ...
##  $ FT%   : num  0.766 0.8 0.668 0.719 0.654 0.45 0.737 0.706 0.798 0.8 ...
##  $ ORB   : num  21 18 154 116 20 8 98 77 148 95 ...
##  $ DRB   : num  80 51 353 289 91 28 401 374 448 369 ...
##  $ TRB   : num  101 69 507 405 111 36 499 451 596 464 ...
##  $ AST   : num  180 125 128 150 29 4 138 99 263 337 ...
##  $ STL   : num  30 25 59 64 16 1 72 60 68 52 ...
##  $ BLK   : num  10 9 55 40 11 13 53 44 121 87 ...
##  $ TOV   : num  82 66 66 89 36 10 120 94 107 116 ...
##  $ PF    : num  132 93 153 172 77 21 171 102 163 138 ...
##  $ PTS   : num  491 322 719 1019 132 ...
```

```
# secondary dataset
str(df_s_final)
```

```
## 'data.frame':    729 obs. of  8 variables:
##  $ name: chr  "Aaron Brooks" "Aaron Brooks" "Aaron Gordon" "Aaron Gordon" ...
##  $ ovr : int  75 85 90 92 69 66 91 83 83 91 ...
##  $ out : int  79 87 87 86 56 47 90 75 81 80 ...
##  $ ins : int  52 51 91 91 65 64 77 72 76 82 ...
##  $ pla : int  74 81 69 49 43 40 60 59 58 82 ...
##  $ ath : int  77 82 86 86 66 58 81 75 75 77 ...
##  $ def : int  52 57 69 75 64 57 76 66 70 80 ...
##  $ reb : int  36 37 87 94 68 71 94 65 73 87 ...
```

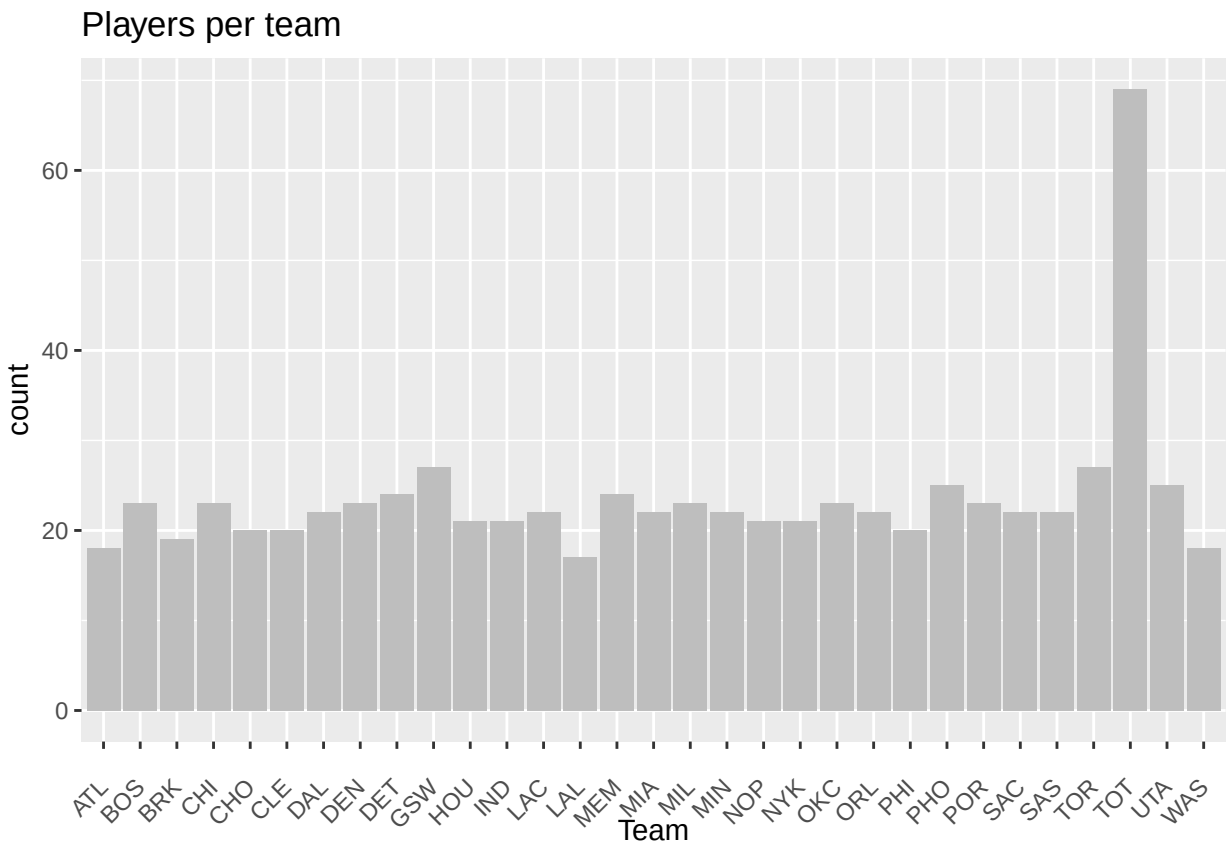
Complete Dataset Histograms

```
library(purrr)
library(tidyr)
library(ggplot2)
df_final %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
```

```
      facet_wrap(~ key, scales = "free") +
      geom_histogram(aes(y=..density..), fill = "grey") +
      geom_density()
ggsave("figures/hist_complete_vars.png", width=15, height=13)
```

Bar Chart of Player by Team from Complete Dataset

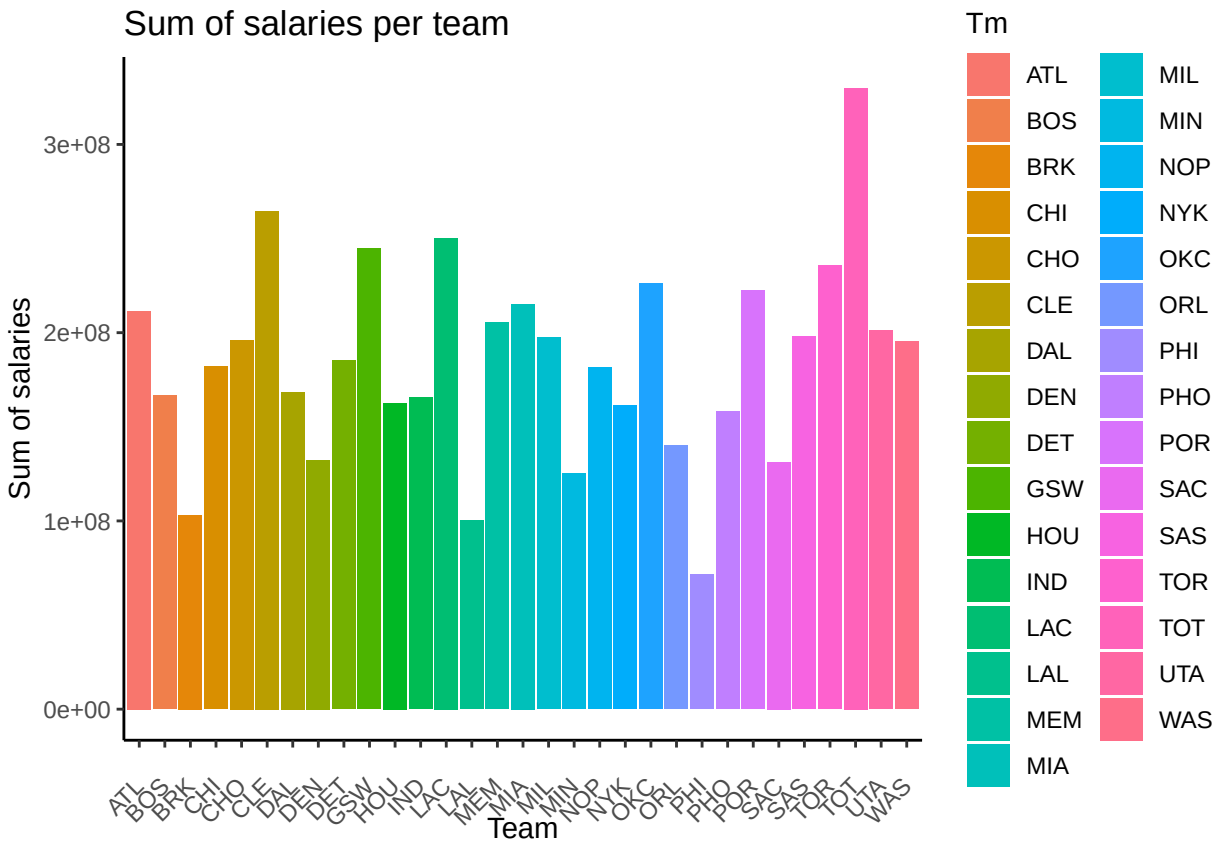
```
library(ggplot2)
ggplot(df_final, aes(x = Tm)) +
  geom_bar(fill = "grey") +
  labs(x = "Team", title = "Players per team") +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```




```
ggsave("figures/bar_complete_player_per_team.png", width=10, height=7)
```

Sum of Salaries per Team for Complete Dataset

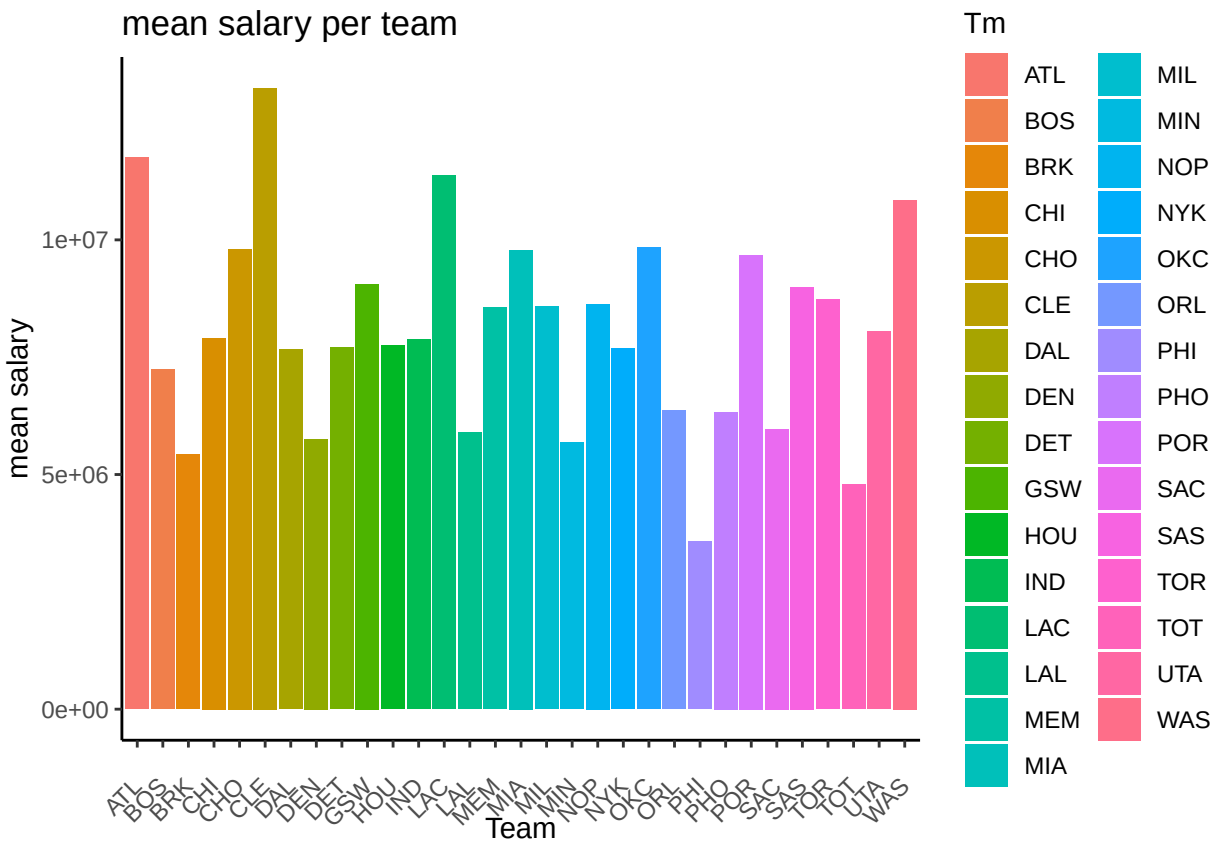
```
library(ggplot2)
library(tidyr)
library(dplyr)
df_final %>%
  group_by(Tm) %>%
  summarise(sum_salary = sum(salary)) %>%
  ggplot(aes(x = Tm, y = sum_salary, fill = Tm)) +
    geom_bar(stat = "identity") +
    theme_classic() +
    labs(
        x = "Team",
        y = "Sum of salaries",
        title = paste("Sum of salaries per team")) +
    theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```




```
ggsave("figures/bar_complete_sum_salaries_per_team.png", width=10, height=7)
```

Mean Salaries per Team for Complete Dataset

```
library(ggplot2)
library(tidyr)
library(dplyr)
df_final %>%
  group_by(Tm) %>%
  summarise(mean_salary = mean(salary)) %>%
  ggplot(aes(x = Tm, y = mean_salary, fill = Tm)) +
    geom_bar(stat = "identity") +
    theme_classic() +
    labs(
        x = "Team",
        y = "mean salary",
        title = paste(
            "mean salary per team")) +
    theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

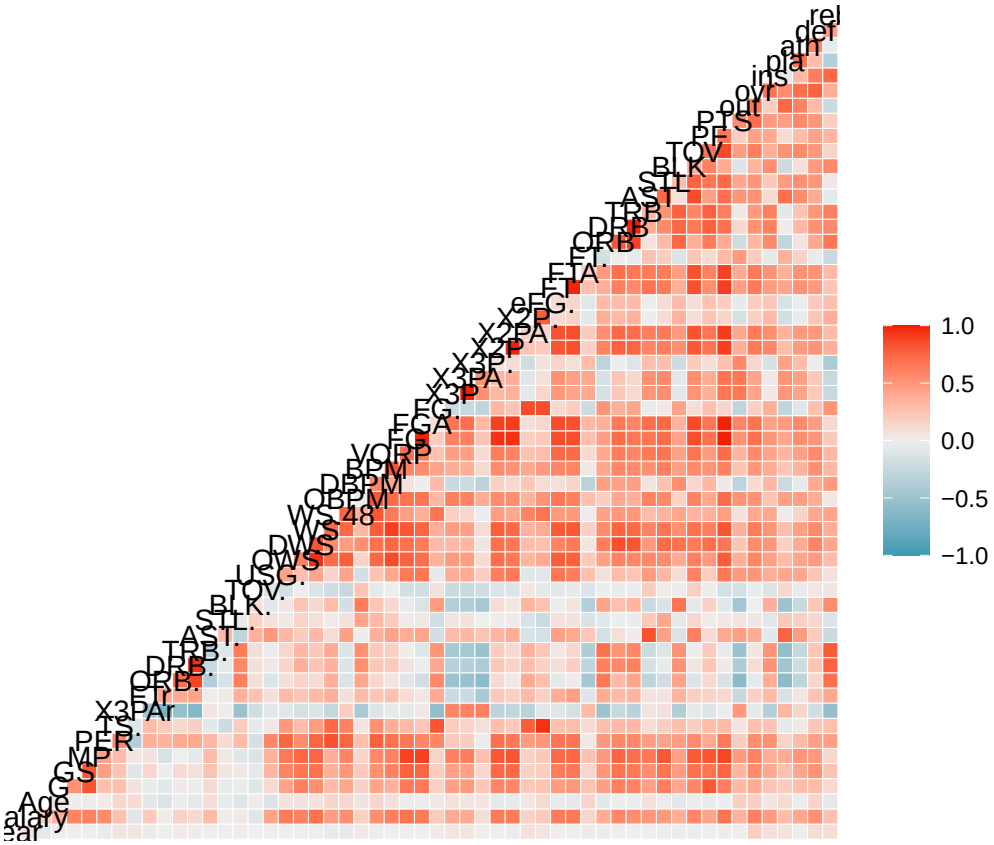



```
ggsave("figures/bar_complete_mean_salaries_per_team.png", width=10, height=7)
```

Correlation Matrix for Complete Dataset

```
library(GGally)
ggcorr(df_final)
```

```
## Warning in ggcorr(df_final): data in column(s) 'name', 'Pos', 'Tm' are not
## numeric and were ignored
```




```
ggsave("figures/complete_correlation_matrix.png", width=10, height=10)
```

Top Salary Correlations from Complete Dataset

```
corr_matrix <- cor(Filter(is.numeric,df_final),method = "pearson")
correlation_salary <- sort(corr_matrix[,'salary'],decreasing = TRUE)
correlation_salary
```

```
##      salary          WS         PTS          FG         FTA         FGA
##  1.00000000  0.69694007  0.68277318  0.67887345  0.65272914  0.64998761
##         OWS          2P          FT         2PA        VORP         ovr
##  0.64938687  0.64906100  0.64179967  0.63494348  0.62381023  0.60924644
##          MP         DWS          GS         TOV         DRB         PER
##  0.60343107  0.60088760  0.59389658  0.58462787  0.58028532  0.55123246
##         TRB         BPM        OBPM         def         AST         STL
##  0.54086403  0.53954352  0.53817279  0.52781018  0.49523840  0.48903287
##         ins       WS/48        USG%          PF         ath          3P
##  0.47107915  0.45339858  0.42959581  0.42180306  0.41425254  0.39254104
##         3PA         ORB         BLK         out           G        AST%
##  0.39105939  0.37101526  0.36750253  0.34764250  0.34594849  0.29618555
##         pla         TS%         reb         FTr        eFG%         FG%
##  0.28681732  0.26814690  0.25776743  0.20931740  0.20513457  0.20096204
##         Age        DBPM          2P%         DRB%         FT%        TRB%
##  0.17217057  0.17133671  0.16770795  0.16645373  0.14290182  0.12041378
##          3P%        year         BLK%        STL%        ORB%        TOV%
##  0.09282571  0.06562855  0.03505738  0.01660475  0.01620331 -0.08835855
##        3PAr
## -0.08875199
```

VARIABLE SELECTION

Helper Functions

```
get_salary_formula <- function(x_vars){
  return(as.formula(sprintf('salary ~ `%s`',paste(x_vars,collapse='` + `'))))}
```

Primary Dataset Variable Selection Using Automated F-Test-Based Backward Selection

```
library(rms)
```

```
## Loading required package: Hmisc
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
## The following object is masked from 'package:rvest':
##
##     html
## The following objects are masked from 'package:base':
##
##     format.pval, units
## Loading required package: SparseM
##
## Attaching package: 'SparseM'
## The following object is masked from 'package:base':
##
##     backsolve
```

```
p_x_vars <- names(df_p_final)[!(names(df_p_final)%in%c('salary','name','2P','2PA','PTS','TRB'))]
# 2P, 2PA, PTS, and TRB were causing singularity in predictor matrix, so they were dropped
p_formula <- get_salary_formula(p_x_vars)
p_formula
```

```
## salary ~ year + Pos + Age + Tm + G + GS + MP + PER + `TS%` +
##     `3PAr` + FTr + `ORB%` + `DRB%` + `TRB%` + `AST%` + `STL%` +
##     `BLK%` + `TOV%` + `USG%` + OWS + DWS + WS + `WS/48` + OBPM +
##     DBPM + BPM + VORP + FG + FGA + `FG%` + `3P` + `3PA` + `3P%` +
##     `2P%` + `eFG%` + FT + FTA + `FT%` + ORB + DRB + AST + STL +
##     BLK + TOV + PF
## <environment: 0x557b7550df88>
```

```
p_selection_model <- ols(p_formula, data = df_p_final)
p_selection_model
```

```
## Linear Regression Model
##
##  ols(formula = p_formula, data = df_p_final)
##
##                        Model Likelihood     Discrimination
##                           Ratio Test           Indexes
```

```
##  Obs           729    LR chi2    767.95    R2       0.651
##  sigma4562607.7641    d.f.           78    R2 adj   0.609
##  d.f.          650    Pr(> chi2) 0.0000    g  6434028.617
##
##  Residuals
##
##        Min         1Q    Median         3Q        Max
##  -15413435  -2767864   -214787   2588116  14947940
##
##
##             Coef           S.E.            t     Pr(>|t|)
##  Intercept -2.904331e+09 941365783.6821 -3.09 0.0021
##  year       1.445455e+06    467526.4743  3.09 0.0021
##  Pos=PF    -3.993265e+05    684912.6640 -0.58 0.5601
##  Pos=PF-C   5.581257e+05   3376233.9080  0.17 0.8688
##  Pos=PG    -4.226804e+06   1146245.1595 -3.69 0.0002
##  Pos=SF    -9.651593e+05    919027.3941 -1.05 0.2940
##  Pos=SG    -2.300728e+06    999349.1031 -2.30 0.0216
##  Age        2.208793e+05     46707.6258  4.73 <0.0001
##  Tm=BOS    -1.799787e+06   1534506.4422 -1.17 0.2413
##  Tm=BRK    -1.766209e+06   1834085.4806 -0.96 0.3359
##  Tm=CHI    -1.467240e+06   1603873.0340 -0.91 0.3606
##  Tm=CHO    -1.199611e+06   1631865.3181 -0.74 0.4625
##  Tm=CLE     1.975514e+06   1662176.5118  1.19 0.2351
##  Tm=DAL    -4.952237e+05   1670872.4331 -0.30 0.7670
##  Tm=DEN    -2.398784e+06   1804535.0031 -1.33 0.1842
##  Tm=DET    -1.300965e+06   1681378.5475 -0.77 0.4394
##  Tm=GSW    -1.194590e+06   1557774.1134 -0.77 0.4434
##  Tm=HOU    -1.865145e+06   1747377.5045 -1.07 0.2862
##  Tm=IND    -2.004682e+06   1589674.6873 -1.26 0.2077
##  Tm=LAC     9.150092e+05   1599531.0996  0.57 0.5675
##  Tm=LAL    -2.456010e+05   2071065.2206 -0.12 0.9056
##  Tm=MEM     4.618436e+05   1698808.8655  0.27 0.7858
##  Tm=MIA    -1.283444e+06   1589069.3816 -0.81 0.4196
##  Tm=MIL     4.310121e+05   1731244.9022  0.25 0.8035
##  Tm=MIN    -2.293160e+06   1884380.4829 -1.22 0.2241
##  Tm=NOP     6.705877e+05   1741684.4359  0.39 0.7003
##  Tm=NYK    -1.324540e+06   1783971.1901 -0.74 0.4581
##  Tm=OKC     1.000927e+06   1675005.1823  0.60 0.5503
##  Tm=ORL    -8.523000e+05   1715902.4517 -0.50 0.6196
##  Tm=PHI    -3.657564e+06   1776553.9927 -2.06 0.0399
##  Tm=PHO    -5.155295e+04   1835285.7651 -0.03 0.9776
##  Tm=POR     2.281753e+06   1756641.0396  1.30 0.1944
##  Tm=SAC    -1.307577e+06   1748756.4154 -0.75 0.4549
##  Tm=SAS    -2.719821e+06   1599463.6795 -1.70 0.0895
##  Tm=TOR     1.961734e+05   1666683.7593  0.12 0.9063
##  Tm=TOT    -1.997084e+06   1434042.2459 -1.39 0.1642
##  Tm=UTA    -1.307648e+06   1562344.4909 -0.84 0.4029
##  Tm=WAS     7.641096e+05   1687983.8983  0.45 0.6509
##  G         -8.373489e+04     22785.5807 -3.67 0.0003
##  GS         2.140579e+04     11773.4286  1.82 0.0695
##  MP         2.577508e+03      1785.4615  1.44 0.1493
##  PER       -5.479893e+04    436861.2968 -0.13 0.9002
##  TS%       -6.247633e+06  21715218.9945 -0.29 0.7737
##  3PAr      -9.829886e+06   6653229.8444 -1.48 0.1400
##  FTr       -8.596919e+05   2976447.9042 -0.29 0.7728
##  ORB%       5.112267e+04    990699.6612  0.05 0.9589
##  DRB%       1.574234e+05    957434.1370  0.16 0.8694
##  TRB%      -2.568829e+05   1937176.4474 -0.13 0.8945
##  AST%       2.367601e+04     88246.8813  0.27 0.7886
##  STL%      -3.104940e+05    550829.7511 -0.56 0.5732
```

```
##  BLK%      -1.059258e+05    448473.7762 -0.24 0.8134
##  TOV%       1.698296e+05     94088.6584  1.80 0.0715
##  USG%       8.220136e+04    193283.4390  0.43 0.6708
##  OWS        4.149648e+06   3707384.5304  1.12 0.2634
##  DWS        6.011305e+06   3727937.6153  1.61 0.1073
##  WS        -3.114980e+06   3687674.0038 -0.84 0.3986
##  WS/48     -1.225996e+07  23475406.8763 -0.52 0.6017
##  OBPM      -3.425475e+06   3777728.1377 -0.91 0.3649
##  DBPM      -4.539602e+06   3742456.6474 -1.21 0.2256
##  BPM        4.509642e+06   3743953.7363  1.20 0.2288
##  VORP      -1.548741e+06    582281.1506 -2.66 0.0080
##  FG         7.859587e+03     30036.7000  0.26 0.7937
##  FGA       -3.268416e+03     14894.3876 -0.22 0.8264
##  FG%       -2.173782e+07  37877504.5070 -0.57 0.5662
##  3P        -2.859535e+04     38445.2554 -0.74 0.4573
##  3PA        1.826901e+04     15758.8612  1.16 0.2468
##  3P%       -1.505102e+06   2069856.9309 -0.73 0.4674
##  2P%       -2.560485e+06   6781959.0108 -0.38 0.7059
##  eFG%       1.538210e+07  36728214.4975  0.42 0.6755
##  FT        -2.850325e+04     24822.9593 -1.15 0.2513
##  FTA        2.806539e+04     15307.3174  1.83 0.0672
##  FT%        9.313072e+05   2283628.3141  0.41 0.6835
##  ORB       -9.305379e+03     13676.4629 -0.68 0.4965
##  DRB        1.020433e+03      6447.4653  0.16 0.8743
##  AST        5.053867e+03      9055.2430  0.56 0.5770
##  STL       -1.689472e+04     18607.6337 -0.91 0.3642
##  BLK        1.281055e+04     18216.7337  0.70 0.4822
##  TOV       -1.006664e+04     22805.1935 -0.44 0.6591
##  PF        -2.424438e+04      9089.1998 -2.67 0.0078
##
```

```
p_seleced <- fastbw(p_selection_model, rule = "p", sls = 0.1)
p_seleced
```

```
##
##  Deleted Chi-Sq d.f. P      Residual d.f. P      AIC    R2
##  ORB%    0.00   1    0.9588  0.00     1   0.9588  -2.00 0.651
##  PER     0.01   1    0.9031  0.02     2   0.9913  -3.98 0.651
##  DRB     0.02   1    0.8916  0.04     3   0.9982  -5.96 0.651
##  FGA     0.04   1    0.8480  0.07     4   0.9994  -7.93 0.651
##  FG      0.06   1    0.8094  0.13     5   0.9997  -9.87 0.651
##  AST%    0.04   1    0.8376  0.17     6   0.9999 -11.83 0.651
##  TS%     0.11   1    0.7383  0.28     7   0.9999 -13.72 0.651
##  eFG%    0.06   1    0.8129  0.34     8   1.0000 -15.66 0.651
##  FT%     0.06   1    0.8060  0.40     9   1.0000 -17.60 0.651
##  BLK%    0.10   1    0.7476  0.50    10   1.0000 -19.50 0.651
##  TOV     0.13   1    0.7208  0.63    11   1.0000 -21.37 0.651
##  FTr     0.31   1    0.5806  0.94    12   1.0000 -23.06 0.651
##  3P%     0.39   1    0.5299  1.33    13   1.0000 -24.67 0.651
##  2P%     0.43   1    0.5127  1.76    14   1.0000 -26.24 0.650
##  BLK     0.46   1    0.4959  2.22    15   0.9999 -27.78 0.650
##  AST     0.58   1    0.4453  2.81    16   0.9999 -29.19 0.650
##  OBPM    0.71   1    0.3992  3.52    17   0.9998 -30.48 0.649
##  WS      0.72   1    0.3958  4.24    18   0.9996 -31.76 0.649
##  STL%    0.89   1    0.3445  5.13    19   0.9993 -32.87 0.649
##  USG%    0.51   1    0.4734  5.65    20   0.9993 -34.35 0.648
##  TRB%    0.76   1    0.3837  6.41    21   0.9990 -35.59 0.648
##  DRB%    0.82   1    0.3639  7.23    22   0.9987 -36.77 0.647
##  3P      0.83   1    0.3632  8.06    23   0.9983 -37.94 0.647
##  3PA     1.60   1    0.2053  9.66    24   0.9958 -38.34 0.646
##  GS      2.97   1    0.0848 12.63    25   0.9807 -37.37 0.644
##  FT      4.85   1    0.0277 17.48    26   0.8938 -34.52 0.642
```

```
##  ORB     3.98   1    0.0460 21.46    27   0.7642 -32.54 0.640
##  STL     4.39   1    0.0362 25.85    28   0.5814 -30.15 0.637
##  TOV%    5.08   1    0.0242 30.93    29   0.3689 -27.07 0.635
##  FG%     3.03   1    0.0818 33.96    30   0.2826 -26.04 0.633
##
## Approximate Estimates after Deleting Factors
##
##                 Coef      S.E.  Wald Z         P
## Intercept -2.726e+09 6.958e+08 -3.9173 8.956e-05
## year       1.355e+06 3.451e+05  3.9266 8.615e-05
## Pos=PF    -8.137e+05 6.104e+05 -1.3331 1.825e-01
## Pos=PF-C   5.494e+05 3.332e+06  0.1649 8.690e-01
## Pos=PG    -3.414e+06 7.520e+05 -4.5397 5.635e-06
## Pos=SF    -1.688e+06 7.264e+05 -2.3233 2.016e-02
## Pos=SG    -2.703e+06 8.025e+05 -3.3683 7.564e-04
## Age        2.515e+05 4.346e+04  5.7867 7.177e-09
## Tm=BOS    -1.697e+06 1.474e+06 -1.1518 2.494e-01
## Tm=BRK    -2.484e+06 1.588e+06 -1.5639 1.179e-01
## Tm=CHI    -1.586e+06 1.463e+06 -1.0842 2.783e-01
## Tm=CHO    -1.025e+06 1.503e+06 -0.6817 4.954e-01
## Tm=CLE     2.039e+06 1.535e+06  1.3279 1.842e-01
## Tm=DAL    -8.993e+05 1.501e+06 -0.5991 5.491e-01
## Tm=DEN    -2.779e+06 1.550e+06 -1.7928 7.300e-02
## Tm=DET    -1.373e+06 1.451e+06 -0.9463 3.440e-01
## Tm=GSW    -4.378e+05 1.432e+06 -0.3057 7.598e-01
## Tm=HOU    -1.836e+06 1.552e+06 -1.1828 2.369e-01
## Tm=IND    -2.682e+06 1.495e+06 -1.7946 7.272e-02
## Tm=LAC     8.928e+05 1.500e+06  0.5951 5.518e-01
## Tm=LAL    -9.671e+05 1.684e+06 -0.5743 5.658e-01
## Tm=MEM    -2.319e+05 1.484e+06 -0.1563 8.758e-01
## Tm=MIA    -1.144e+06 1.477e+06 -0.7750 4.383e-01
## Tm=MIL    -2.407e+05 1.507e+06 -0.1597 8.731e-01
## Tm=MIN    -3.065e+06 1.578e+06 -1.9417 5.217e-02
## Tm=NOP     2.756e+05 1.529e+06  0.1802 8.570e-01
## Tm=NYK    -1.531e+06 1.545e+06 -0.9912 3.216e-01
## Tm=OKC     6.711e+05 1.489e+06  0.4508 6.521e-01
## Tm=ORL    -1.076e+06 1.512e+06 -0.7115 4.768e-01
## Tm=PHI    -3.969e+06 1.569e+06 -2.5307 1.138e-02
## Tm=PHO    -9.995e+05 1.535e+06 -0.6512 5.149e-01
## Tm=POR     2.151e+06 1.543e+06  1.3945 1.632e-01
## Tm=SAC    -2.341e+06 1.540e+06 -1.5205 1.284e-01
## Tm=SAS    -2.116e+06 1.485e+06 -1.4254 1.540e-01
## Tm=TOR    -3.270e+05 1.445e+06 -0.2264 8.209e-01
## Tm=TOT    -2.112e+06 1.274e+06 -1.6582 9.728e-02
## Tm=UTA    -1.105e+06 1.430e+06 -0.7726 4.398e-01
## Tm=WAS     2.629e+05 1.570e+06  0.1674 8.670e-01
## G         -1.029e+05 1.909e+04 -5.3898 7.053e-08
## MP         3.665e+03 8.188e+02  4.4758 7.611e-06
## 3PAr      -4.222e+06 1.449e+06 -2.9137 3.572e-03
## OWS        9.085e+05 2.610e+05  3.4805 5.004e-04
## DWS        2.967e+06 4.982e+05  5.9560 2.585e-09
## WS/48     -3.670e+07 8.905e+06 -4.1217 3.760e-05
## DBPM      -1.195e+06 2.477e+05 -4.8229 1.415e-06
## BPM        1.152e+06 2.393e+05  4.8137 1.482e-06
## VORP      -1.229e+06 4.120e+05 -2.9828 2.856e-03
## FTA        1.027e+04 2.631e+03  3.9047 9.436e-05
## PF        -2.448e+04 6.843e+03 -3.5775 3.469e-04
##
## Factors in Final Model
##
##  [1] year  Pos   Age   Tm    G     MP    3PAr  OWS   DWS   WS/48 DBPM  BPM
```

```
## [13] VORP  FTA   PF
```

Checking for Multicollinearity Among Optimal Subset of Primary Variables.

```
p_subset_formula <- get_salary_formula(p_seleced[['names.kept']])
p_subset_formula
```

```
## salary ~ year + Pos + Age + Tm + G + MP + `3PAr` + OWS + DWS +
##     `WS/48` + DBPM + BPM + VORP + FTA + PF
## <environment: 0x557b78c050f8>
```

```
p_subset_lm <- lm(p_subset_formula , data=df_p_final)
summary(p_subset_lm)
```

```
##
## Call:
## lm(formula = p_subset_formula, data = df_p_final)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -15502276  -3036236   -166445   2773803  16858913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.726e+09  6.978e+08  -3.906 0.000103 ***
## year         1.355e+06  3.461e+05   3.915 9.94e-05 ***
## PosPF       -8.137e+05  6.122e+05  -1.329 0.184220
## PosPF-C      5.494e+05  3.342e+06   0.164 0.869461
## PosPG       -3.414e+06  7.542e+05  -4.527 7.08e-06 ***
## PosSF       -1.688e+06  7.285e+05  -2.317 0.020825 *
## PosSG       -2.703e+06  8.048e+05  -3.359 0.000827 ***
## Age          2.515e+05  4.358e+04   5.770 1.20e-08 ***
## TmBOS       -1.697e+06  1.478e+06  -1.148 0.251174
## TmBRK       -2.484e+06  1.593e+06  -1.559 0.119384
## TmCHI       -1.586e+06  1.467e+06  -1.081 0.280065
## TmCHO       -1.025e+06  1.508e+06  -0.680 0.496896
## TmCLE        2.039e+06  1.540e+06   1.324 0.185938
## TmDAL       -8.993e+05  1.505e+06  -0.597 0.550456
## TmDEN       -2.779e+06  1.555e+06  -1.788 0.074277 .
## TmDET       -1.373e+06  1.455e+06  -0.944 0.345725
## TmGSW       -4.378e+05  1.436e+06  -0.305 0.760602
## TmHOU       -1.836e+06  1.557e+06  -1.179 0.238645
## TmIND       -2.682e+06  1.499e+06  -1.789 0.074001 .
## TmLAC        8.928e+05  1.505e+06   0.593 0.553115
## TmLAL       -9.671e+05  1.689e+06  -0.573 0.567090
## TmMEM       -2.319e+05  1.488e+06  -0.156 0.876193
## TmMIA       -1.144e+06  1.481e+06  -0.773 0.439939
## TmMIL       -2.407e+05  1.512e+06  -0.159 0.873551
## TmMIN       -3.065e+06  1.583e+06  -1.936 0.053270 .
## TmNOP        2.756e+05  1.534e+06   0.180 0.857444
## TmNYK       -1.531e+06  1.549e+06  -0.988 0.323356
## TmOKC        6.711e+05  1.493e+06   0.450 0.653209
## TmORL       -1.076e+06  1.516e+06  -0.709 0.478324
## TmPHI       -3.969e+06  1.573e+06  -2.523 0.011849 *
## TmPHO       -9.995e+05  1.539e+06  -0.649 0.516349
## TmPOR        2.151e+06  1.547e+06   1.390 0.164848
## TmSAC       -2.341e+06  1.544e+06  -1.516 0.129959
## TmSAS       -2.116e+06  1.489e+06  -1.421 0.155687
## TmTOR       -3.270e+05  1.449e+06  -0.226 0.821492
## TmTOT       -2.112e+06  1.277e+06  -1.653 0.098715 .
## TmUTA       -1.105e+06  1.434e+06  -0.770 0.441352
```

```
## TmWAS        2.629e+05  1.575e+06   0.167 0.867450
## G           -1.029e+05  1.915e+04  -5.374 1.06e-07 ***
## MP           3.665e+03  8.212e+02   4.463 9.46e-06 ***
## `3PAr`      -4.222e+06  1.453e+06  -2.905 0.003789 **
## OWS          9.085e+05  2.618e+05   3.470 0.000552 ***
## DWS          2.967e+06  4.996e+05   5.939 4.58e-09 ***
## `WS/48`     -3.670e+07  8.930e+06  -4.110 4.44e-05 ***
## DBPM        -1.195e+06  2.484e+05  -4.809 1.87e-06 ***
## BPM          1.152e+06  2.400e+05   4.800 1.95e-06 ***
## VORP        -1.229e+06  4.132e+05  -2.974 0.003042 **
## FTA          1.027e+04  2.639e+03   3.893 0.000109 ***
## PF          -2.448e+04  6.863e+03  -3.567 0.000386 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4576000 on 680 degrees of freedom
## Multiple R-squared:  0.633,  Adjusted R-squared:  0.6071
## F-statistic: 24.44 on 48 and 680 DF,  p-value: < 2.2e-16
```

```
vif(p_subset_lm) # All variables have low VIF values. So no multicollinearity.
```

```
##      year     PosPF   PosPF-C     PosPG     PosSF     PosSG       Age     TmBOS
##  1.042488  2.068084  1.063938  3.039324  2.898172  3.612086  1.235888  2.323207
##     TmBRK     TmCHI     TmCHO     TmCLE     TmDAL     TmDEN     TmDET     TmGSW
##  2.242634  2.289195  2.112056  2.202889  2.309227  2.571270  2.345968  2.561503
##     TmHOU     TmIND     TmLAC     TmLAL     TmMEM     TmMIA     TmMIL     TmMIN
##  2.361245  2.188260  2.306629  2.261795  2.454998  2.234695  2.431149  2.553628
##     TmNOP     TmNYK     TmOKC     TmORL     TmPHI     TmPHO     TmPOR     TmSAC
##  2.291435  2.337216  2.371049  2.342636  2.298772  2.732122  2.546828  2.430115
##     TmSAS     TmTOR     TmTOT     TmUTA     TmWAS         G        MP    `3PAr`
##  2.258760  2.606798  4.867458  2.371320  2.079046  5.303619 14.425244  3.133272
##       OWS       DWS   `WS/48`      DBPM       BPM      VORP       FTA        PF
## 11.305787 10.973791 12.008925  8.080378 21.840249 14.072543  4.481791  6.096161
```

```
p_vars_final <- p_seleced[['names.kept']]
```

Complete Dataset Variable Selection Using Automated F-Test-Based Backward Selection

```
library(rms)
c_x_vars <- names(df_final)[!(names(df_final)%in%c('salary','name','2P','2PA','PTS','TRB'))]
# 2P, 2PA, PTS, and TRB were causing singularity in predictor matrix, so they were dropped
c_formula <- get_salary_formula(c_x_vars)
c_formula
```

```
## salary ~ year + Pos + Age + Tm + G + GS + MP + PER + `TS%` +
##     `3PAr` + FTr + `ORB%` + `DRB%` + `TRB%` + `AST%` + `STL%` +
##     `BLK%` + `TOV%` + `USG%` + OWS + DWS + WS + `WS/48` + OBPM +
##     DBPM + BPM + VORP + FG + FGA + `FG%` + `3P` + `3PA` + `3P%` +
##     `2P%` + `eFG%` + FT + FTA + `FT%` + ORB + DRB + AST + STL +
##     BLK + TOV + PF + out + ovr + ins + pla + ath + def + reb
## <environment: 0x557b79ce7d90>
```

```
c_selection_model <- ols(c_formula, data = df_final)
c_selection_model
```

```
## Linear Regression Model
##
##  ols(formula = c_formula, data = df_final)
##
##                       Model Likelihood     Discrimination
##                          Ratio Test           Indexes
##  Obs           729    LR chi2    784.36    R2       0.659
```

```
##  sigma4536019.7526    d.f.           85    R2 adj   0.614
##  d.f.          643    Pr(> chi2) 0.0000    g  6466358.479
##
##  Residuals
##
##        Min         1Q    Median         3Q        Max
##  -15228458  -2574926   -164661   2407168  14763079
##
##
##             Coef           S.E.          t     Pr(>|t|)
##  Intercept -1.862157e+09 1.089539e+09 -1.71 0.0879
##  year       9.264194e+05 5.411141e+05  1.71 0.0874
##  Pos=PF    -1.245596e+05 7.253857e+05 -0.17 0.8637
##  Pos=PF-C   1.368401e+06 3.385775e+06  0.40 0.6862
##  Pos=PG    -4.539750e+06 1.379005e+06 -3.29 0.0010
##  Pos=SF    -1.250292e+06 1.003014e+06 -1.25 0.2130
##  Pos=SG    -2.566197e+06 1.171328e+06 -2.19 0.0288
##  Age        1.784575e+05 5.067069e+04  3.52 0.0005
##  Tm=BOS    -1.893085e+06 1.530779e+06 -1.24 0.2167
##  Tm=BRK    -1.045437e+06 1.839828e+06 -0.57 0.5701
##  Tm=CHI    -1.233178e+06 1.601371e+06 -0.77 0.4415
##  Tm=CHO    -8.359099e+05 1.627861e+06 -0.51 0.6078
##  Tm=CLE     1.778669e+06 1.662906e+06  1.07 0.2852
##  Tm=DAL    -2.277301e+05 1.664375e+06 -0.14 0.8912
##  Tm=DEN    -2.116142e+06 1.803191e+06 -1.17 0.2410
##  Tm=DET    -8.663116e+05 1.682168e+06 -0.51 0.6067
##  Tm=GSW    -1.399756e+06 1.561722e+06 -0.90 0.3704
##  Tm=HOU    -1.875317e+06 1.741230e+06 -1.08 0.2819
##  Tm=IND    -1.734918e+06 1.588570e+06 -1.09 0.2752
##  Tm=LAC     8.090358e+05 1.592459e+06  0.51 0.6116
##  Tm=LAL    -8.012139e+04 2.063844e+06 -0.04 0.9690
##  Tm=MEM     4.396382e+05 1.705404e+06  0.26 0.7967
##  Tm=MIA    -1.075528e+06 1.583534e+06 -0.68 0.4973
##  Tm=MIL     7.321702e+05 1.739865e+06  0.42 0.6740
##  Tm=MIN    -1.939443e+06 1.888965e+06 -1.03 0.3049
##  Tm=NOP     9.163479e+05 1.737309e+06  0.53 0.5981
##  Tm=NYK    -1.104200e+06 1.788813e+06 -0.62 0.5373
##  Tm=OKC     9.963334e+05 1.673803e+06  0.60 0.5519
##  Tm=ORL    -4.275795e+05 1.711287e+06 -0.25 0.8028
##  Tm=PHI    -3.146405e+06 1.774363e+06 -1.77 0.0767
##  Tm=PHO     3.165694e+05 1.832103e+06  0.17 0.8629
##  Tm=POR     2.454171e+06 1.752860e+06  1.40 0.1620
##  Tm=SAC    -9.984937e+05 1.746793e+06 -0.57 0.5678
##  Tm=SAS    -2.843545e+06 1.604203e+06 -1.77 0.0768
##  Tm=TOR     1.527904e+05 1.664926e+06  0.09 0.9269
##  Tm=TOT    -1.880008e+06 1.429814e+06 -1.31 0.1890
##  Tm=UTA    -1.370369e+06 1.559036e+06 -0.88 0.3797
##  Tm=WAS     1.074738e+06 1.685293e+06  0.64 0.5239
##  G         -7.213246e+04 2.303575e+04 -3.13 0.0018
##  GS         1.934477e+04 1.183759e+04  1.63 0.1027
##  MP         2.043614e+03 1.800936e+03  1.13 0.2569
##  PER       -7.232608e+04 4.357120e+05 -0.17 0.8682
##  TS%        2.804291e+06 2.187451e+07  0.13 0.8980
##  3PAr      -9.360439e+06 6.623436e+06 -1.41 0.1581
##  FTr       -1.847125e+06 2.995828e+06 -0.62 0.5377
##  ORB%       1.908199e+05 9.905421e+05  0.19 0.8473
##  DRB%       2.248505e+05 9.559642e+05  0.24 0.8141
##  TRB%      -4.680003e+05 1.935267e+06 -0.24 0.8090
##  AST%       8.723315e+03 8.936894e+04  0.10 0.9223
##  STL%      -2.532261e+05 5.530271e+05 -0.46 0.6472
##  BLK%      -2.309969e+05 4.524496e+05 -0.51 0.6098
```

```
##  TOV%       1.754818e+05 9.490810e+04  1.85 0.0649
##  USG%       1.181789e+05 1.946107e+05  0.61 0.5439
##  OWS        3.889040e+06 3.709470e+06  1.05 0.2948
##  DWS        5.555895e+06 3.720148e+06  1.49 0.1358
##  WS        -2.865891e+06 3.683318e+06 -0.78 0.4368
##  WS/48     -6.579129e+06 2.349473e+07 -0.28 0.7795
##  OBPM      -3.194178e+06 3.761883e+06 -0.85 0.3961
##  DBPM      -4.084174e+06 3.729023e+06 -1.10 0.2738
##  BPM        4.107698e+06 3.728656e+06  1.10 0.2710
##  VORP      -1.377750e+06 5.909769e+05 -2.33 0.0200
##  FG         7.955394e+03 3.036633e+04  0.26 0.7934
##  FGA       -2.775365e+03 1.500940e+04 -0.18 0.8534
##  FG%       -2.816772e+07 3.808067e+07 -0.74 0.4598
##  3P        -3.707326e+04 3.886119e+04 -0.95 0.3404
##  3PA        2.117817e+04 1.584865e+04  1.34 0.1819
##  3P%       -7.082239e+05 2.120252e+06 -0.33 0.7385
##  2P%       -2.057427e+06 6.866814e+06 -0.30 0.7646
##  eFG%       1.275957e+07 3.669827e+07  0.35 0.7282
##  FT        -2.548486e+04 2.499030e+04 -1.02 0.3082
##  FTA        2.449323e+04 1.537310e+04  1.59 0.1116
##  FT%        1.058487e+06 2.273883e+06  0.47 0.6417
##  ORB       -1.061636e+04 1.376503e+04 -0.77 0.4408
##  DRB        4.820145e+03 6.571272e+03  0.73 0.4635
##  AST        6.632647e+03 9.202001e+03  0.72 0.4713
##  STL       -2.662004e+04 1.880369e+04 -1.42 0.1574
##  BLK        9.317229e+03 1.839855e+04  0.51 0.6127
##  TOV       -1.330621e+04 2.317883e+04 -0.57 0.5661
##  PF        -2.325120e+04 9.054637e+03 -2.57 0.0105
##  out       -7.741260e+04 3.674630e+04 -2.11 0.0355
##  ovr        1.262315e+05 8.249180e+04  1.53 0.1265
##  ins        1.299027e+04 4.412704e+04  0.29 0.7686
##  pla       -2.278761e+02 2.830248e+04 -0.01 0.9936
##  ath       -1.306647e+04 4.425939e+04 -0.30 0.7679
##  def        4.946639e+04 3.538236e+04  1.40 0.1626
##  reb       -5.595674e+04 2.478000e+04 -2.26 0.0243
##
```

```
c_seleced <- fastbw(c_selection_model, rule = "p", sls = 0.1)
c_seleced
```

```
##
##  Deleted Chi-Sq d.f. P      Residual d.f. P      AIC    R2
##  pla       0.00   1   0.9936  0.00     1   0.9936  -2.00 0.659
##  AST%      0.01   1   0.9224  0.01     2   0.9952  -3.99 0.659
##  TS%       0.02   1   0.9015  0.02     3   0.9990  -5.98 0.659
##  PER       0.02   1   0.8866  0.05     4   0.9997  -7.95 0.659
##  FGA       0.02   1   0.8884  0.06     5   0.9999  -9.94 0.659
##  ORB%      0.03   1   0.8619  0.10     6   1.0000 -11.90 0.659
##  DRB%      0.05   1   0.8213  0.15     7   1.0000 -13.85 0.659
##  2P%       0.07   1   0.7959  0.21     8   1.0000 -15.79 0.659
##  ins       0.07   1   0.7843  0.29     9   1.0000 -17.71 0.659
##  ath       0.08   1   0.7777  0.37    10   1.0000 -19.63 0.659
##  3P%       0.12   1   0.7289  0.49    11   1.0000 -21.51 0.659
##  FG        0.14   1   0.7042  0.63    12   1.0000 -23.37 0.659
##  WS/48     0.26   1   0.6093  0.89    13   1.0000 -25.11 0.659
##  FT%       0.25   1   0.6205  1.14    14   1.0000 -26.86 0.658
##  eFG%      0.31   1   0.5776  1.45    15   1.0000 -28.55 0.658
##  BLK       0.29   1   0.5895  1.74    16   1.0000 -30.26 0.658
##  TOV       0.35   1   0.5526  2.09    17   1.0000 -31.91 0.658
##  AST       0.34   1   0.5618  2.43    18   1.0000 -33.57 0.658
##  BLK%      0.60   1   0.4373  3.03    19   1.0000 -34.97 0.657
##  STL%      0.56   1   0.4553  3.59    20   1.0000 -36.41 0.657
```

```
##  OBPM     0.66   1   0.4181  4.25    21   1.0000 -37.75 0.657
##  WS       0.78   1   0.3770  5.03    22   0.9999 -38.97 0.656
##  FTr      0.90   1   0.3438  5.92    23   0.9999 -40.08 0.656
##  3P       1.08   1   0.2978  7.01    24   0.9997 -40.99 0.655
##  3PA      0.78   1   0.3764  7.79    25   0.9996 -42.21 0.655
##  TRB%     0.72   1   0.3954  8.51    26   0.9995 -43.49 0.655
##  DRB      1.09   1   0.2971  9.60    27   0.9992 -44.40 0.654
##  def      1.71   1   0.1912 11.31    28   0.9978 -44.69 0.653
##  GS       2.43   1   0.1193 13.73    29   0.9926 -44.27 0.652
##  VORP     2.37   1   0.1236 16.10    30   0.9818 -43.90 0.650
##  FT       3.55   1   0.0597 19.65    31   0.9430 -42.35 0.649
##  FTA      0.73   1   0.3928 20.38    32   0.9443 -43.62 0.648
##  ORB      2.33   1   0.1266 22.72    33   0.9105 -43.28 0.647
##  year     5.42   1   0.0199 28.13    34   0.7501 -39.87 0.644
##  3PAr     2.89   1   0.0890 31.03    35   0.6605 -38.97 0.643
##  Tm      46.83  30   0.0258 77.86    65   0.1316 -52.14 0.618
##
## Approximate Estimates after Deleting Factors
##
##                Coef      S.E.  Wald Z         P
## Intercept  -7443996 2890995.2 -2.5749 1.003e-02
## Pos=PF       305142  616656.4  0.4948 6.207e-01
## Pos=PF-C     534967 3277297.6  0.1632 8.703e-01
## Pos=PG     -2950590  973572.7 -3.0307 2.440e-03
## Pos=SF      -649584  796423.8 -0.8156 4.147e-01
## Pos=SG     -1620820  929173.4 -1.7444 8.110e-02
## Age          194132   43375.1  4.4757 7.617e-06
## G            -86619   17274.8 -5.0142 5.326e-07
## MP             5636     705.4  7.9891 1.332e-15
## TOV%         108590   48513.3  2.2384 2.520e-02
## USG%         160857   43909.6  3.6634 2.489e-04
## OWS          628726  156302.7  4.0225 5.759e-05
## DWS         1845071  329440.5  5.6006 2.136e-08
## DBPM        -540674  201878.4 -2.6782 7.402e-03
## BPM          383920  138070.0  2.7806 5.426e-03
## FG%       -10577878 3567092.0 -2.9654 3.023e-03
## STL          -36992   10782.3 -3.4308 6.019e-04
## PF           -20043    6409.1 -3.1273 1.764e-03
## out         -131999   25879.2 -5.1006 3.386e-07
## ovr          276397   43182.1  6.4007 1.546e-10
## reb          -54947   20397.0 -2.6939 7.063e-03
##
## Factors in Final Model
##
##  [1] Pos  Age  G    MP   TOV% USG% OWS  DWS  DBPM BPM  FG%  STL  PF   out  ovr
## [16] reb
```

Checking for Multicollinearity Among Optimal Subset of Complete Variables.

```
c_subset_formula <- get_salary_formula(c_seleced[['names.kept']])
c_subset_formula
```

```
## salary ~ Pos + Age + G + MP + `TOV%` + `USG%` + OWS + DWS + DBPM +
##     BPM + `FG%` + STL + PF + out + ovr + reb
## <environment: 0x557b7b5234f8>
```

```
c_subset_lm <- lm(c_subset_formula , data=df_final)
summary(c_subset_lm)
```

```
##
## Call:
```

```
## lm(formula = c_subset_formula, data = df_final)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -16885929  -2941962   -364289   2588753  16723179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.444e+06  2.917e+06  -2.552 0.010925 *
## PosPF        3.051e+05  6.222e+05   0.490 0.624003
## PosPF-C      5.350e+05  3.307e+06   0.162 0.871532
## PosPG       -2.951e+06  9.824e+05  -3.004 0.002763 **
## PosSF       -6.496e+05  8.036e+05  -0.808 0.419179
## PosSG       -1.621e+06  9.376e+05  -1.729 0.084292 .
## Age          1.941e+05  4.377e+04   4.436 1.06e-05 ***
## G           -8.662e+04  1.743e+04  -4.969 8.44e-07 ***
## MP           5.636e+03  7.118e+02   7.918 9.37e-15 ***
## `TOV%`       1.086e+05  4.895e+04   2.218 0.026851 *
## `USG%`       1.609e+05  4.431e+04   3.631 0.000303 ***
## OWS          6.287e+05  1.577e+05   3.986 7.40e-05 ***
## DWS          1.845e+06  3.324e+05   5.550 4.03e-08 ***
## DBPM        -5.407e+05  2.037e+05  -2.654 0.008128 **
## BPM          3.839e+05  1.393e+05   2.756 0.006007 **
## `FG%`       -1.058e+07  3.599e+06  -2.939 0.003402 **
## STL         -3.699e+04  1.088e+04  -3.400 0.000712 ***
## PF          -2.004e+04  6.467e+03  -3.099 0.002017 **
## out         -1.320e+05  2.611e+04  -5.055 5.49e-07 ***
## ovr          2.764e+05  4.357e+04   6.343 4.01e-10 ***
## reb         -5.495e+04  2.058e+04  -2.670 0.007765 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4577000 on 708 degrees of freedom
## Multiple R-squared:  0.6177, Adjusted R-squared:  0.6069
## F-statistic: 57.21 on 20 and 708 DF,  p-value: < 2.2e-16
```

```
vif(c_subset_lm) # All variables have low VIF values. So no multicollinearity.
```

```
##     PosPF   PosPF-C     PosPG     PosSF     PosSG       Age         G        MP
##  2.135637  1.041163  5.153781  3.524824  4.899340  1.245715  4.394009 10.832663
##    `TOV%`    `USG%`       OWS       DWS      DBPM       BPM     `FG%`       STL
##  1.571512  1.922782  4.101722  4.854849  5.430365  7.358386  2.670350  4.677699
##        PF       out       ovr       reb
##  5.410165  5.588898  5.706429  4.767951
```

```
c_vars_final <- c_seleced[['names.kept']]
```

Subset Primary and Complete Dataframes to Include Only Name, Salary, and Selected Variables

```
p_vars_subset <- c('name','salary',p_vars_final)
df_p_subset_final <- df_p_final[,p_vars_subset]
c_vars_subset <- c('name','salary',c_vars_final)
df_c_subset_final <- df_final[,c_vars_subset]
```

Split Train-Test

```
library(caret)
set.seed(7)
```

Primary Dataset

```
train_rows <- createDataPartition(y=df_p_subset_final[,'salary'], list=FALSE, p=.8)
p_train_df <- df_p_subset_final[train_rows,]
p_test_df <- df_p_subset_final[-train_rows,]
stopifnot(nrow(p_train_df) + nrow(p_test_df) == nrow(df_p_subset_final))
nrow(p_train_df)
```

```
## [1] 585
```

```
nrow(p_test_df)
```

```
## [1] 144
```

```
names(p_train_df)
```

```
##  [1] "name"   "salary" "year"   "Pos"    "Age"    "Tm"     "G"      "MP"
##  [9] "3PAr"   "OWS"    "DWS"    "WS/48"  "DBPM"   "BPM"    "VORP"   "FTA"
## [17] "PF"
```

```
head(p_train_df)
```

```
##            name  salary year Pos Age  Tm  G   MP  3PAr  OWS DWS  WS/48 DBPM
## 1  Aaron Brooks 2700000 2016  PG  31 CHI 69 1108 0.394  0.2 0.7  0.040 -2.8
## 2  Aaron Brooks 2116955 2017  PG  32 IND 65  894 0.427 -0.2 0.5  0.016 -2.6
## 3  Aaron Gordon 4351320 2016  PF  20 ORL 78 1863 0.245  3.2 2.2  0.139  1.2
## 4  Aaron Gordon 5504420 2017  SF  21 ORL 80 2298 0.309  2.0 1.7  0.076 -0.4
## 5 Adreian Payne 2022240 2016  PF  24 MIN 52  486 0.221 -0.9 0.4 -0.047 -0.2
## 6  A.J. Hammons 1312611 2017   C  24 DAL 22  163 0.238 -0.2 0.2 -0.001  1.9
##    BPM VORP FTA  PF
## 1 -3.3 -0.4  64 132
## 2 -4.6 -0.6  40  93
## 3  1.8  1.8 193 153
## 4 -0.7  0.8 217 172
## 5 -6.1 -0.5  26  77
## 6 -5.6 -0.1  20  21
```

```
write.csv(p_train_df,'data/train_test/primary/train.csv')
write.csv(p_test_df,'data/train_test/primary/test.csv')
```

Complete Dataset

```
library(caret)
set.seed(7)
train_rows <- createDataPartition(y=df_c_subset_final[,'salary'], list=FALSE, p=.8)
c_train_df <- df_c_subset_final[train_rows,]
c_test_df <- df_c_subset_final[-train_rows,]
stopifnot(nrow(c_train_df) + nrow(c_test_df) == nrow(df_c_subset_final))
nrow(c_train_df)
```

```
## [1] 585
```

```
nrow(c_test_df)
```

```
## [1] 144
```

```
names(c_train_df)
```

```
##  [1] "name"   "salary" "Pos"    "Age"    "G"      "MP"     "TOV%"   "USG%"
##  [9] "OWS"    "DWS"    "DBPM"   "BPM"    "FG%"    "STL"    "PF"     "out"
## [17] "ovr"    "reb"
```

```
head(c_train_df)
```

```
##            name  salary Pos Age  G   MP TOV% USG%  OWS DWS DBPM  BPM   FG% STL
```

```
## 1  Aaron Brooks 2700000  PG  31 69 1108 14.2 22.9  0.2 0.7 -2.8 -3.3 0.401  30
## 2  Aaron Brooks 2116955  PG  32 65  894 17.2 19.2 -0.2 0.5 -2.6 -4.6 0.403  25
## 3  Aaron Gordon 4351320  PF  20 78 1863  9.0 17.3  3.2 2.2  1.2  1.8 0.473  59
## 4  Aaron Gordon 5504420  SF  21 80 2298  8.5 20.1  2.0 1.7 -0.4 -0.7 0.454  64
## 5 Adreian Payne 2022240  PF  24 52  486 18.7 17.7 -0.9 0.4 -0.2 -6.1 0.366  16
## 6  A.J. Hammons 1312611   C  24 22  163 16.4 17.6 -0.2 0.2  1.9 -5.6 0.405   1
##    PF out ovr reb
## 1 132  79  75  36
## 2  93  87  85  37
## 3 153  87  90  87
## 4 172  86  92  94
## 5  77  56  69  68
## 6  21  47  66  71
```

```
write.csv(c_train_df,'data/train_test/complete/train.csv')
write.csv(c_test_df,'data/train_test/complete/test.csv')
```