Project #1 - NBA Stats and Salaries

Gabriella Cerrato (gac2625)

5/9/2021

R Markdown

```
#setting up mirror
options(repos =c(CRAN ="http://cran.rstudio.com"))
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/v6/xdp535n500b2ks1gszsfl5880000gn/T//RtmpE1TxPo/downloaded_packages
```

```
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/v6/xdp535n500b2ks1gszsfl5880000gn/T//RtmpE1TxPo/downloaded_packages
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/v6/xdp535n500b2ks1gszsfl5880000gn/T//RtmpE1TxPo/downloaded_packages
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.1.0     ✓ stringr 1.4.0
## ✓ tidyr   1.1.3     ✓ forcats 0.5.1
## ✓ readr   1.4.0
```

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

```
install.packages("factoextra")
```

```
##
## The downloaded binary packages are in
##  /var/folders/v6/xdp535n500b2ks1gszsfl5880000gn/T//RtmpE1TxPo/downloaded_packages
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#I have chosen to explore data regarding the salaries and game statistics of NBA players during the 2018-2019 season. I chose a data set of the salaries of NBA players for the 2018-2019 season (https://hoopshype.com/salaries/2018-2019/ (https://hoopshype.com/salaries/2018-2019/)) as well as a data set of game statistics (https://www.nbastuffer.com/2018-2019-nba-player-stats/ (https://www.nbastuffer.com/2018-2019-nba-player-stats/)) that included data of minutes played per game, ppints scored per game, age of player, etc.. I chose these data sets because I thought it would be interestign to see how the salaries earned by the players related to their performance in games.

```
library(readxl)
NBAplayerstats <- read_excel("~/Downloads/2018-2019 NBA Player Stats.project2.xlsx")
```

```
## New names:
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * `` -> ...6
## * ...
```

```
View(NBAplayerstats)

#data set of the salaries of NBA players during the 2018-2019 season
library(readxl)
NBA_salaries <- read_excel("~/Downloads/NBA salaries.project2.xlsx")
```

```
## New names:
## * `` -> ...4
```

```
View(NBA_salaries)
```

```
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/v6/xdp535n500b2ks1gszsfl5880000gn/T//RtmpE1TxPo/downloaded_packages
```

```
library(dplyr)
#to tidy the data set of the salaries earned by players, I removed a redundant column th
at specified row number and removed a column of the salaries adjusted for inflation
NBAsalaries <- subset(NBA_salaries, select = -c(1, 4)) %>%
#I renamed the columns in the salary data set
  rename("NAME" = 1, "Salary ($)" = 2)

#In the second data set, I also renamed some columns for clarity and brevity. MPG stands
for minutes played per game, etc.
NBAplayerstats <- subset(NBAplayerstats, select = -c(1)) %>%
  rename(
   "NAME" = 1, "TEAM" = 2, "Position" =  3, "Age" = 4, "GP" = 5, "MPG" = 6,  "Percentage
of Team Minutes Used" = 7, "Usage Rate" = 8, "Turnover Rate" = 9, "FTA" = 10, "FT%" = 11
, "2PA" = 12, "2P%" = 13, "3PA" = 14, "3P%" = 15, "Effective Shooting %" = 16, "True Sho
oting %" = 17, "Points Per Game" = 18, "Rebounds Per Game"= 19, "Total Rebound Percentag
e" = 20, "Assists Per Game" = 21, "Assists %" = 22, "Steals Per Game" = 23, "Blocks Per
 Game" = 24, "Turnovers Per Game" = 25, "Versatility Index" = 26, "Offensive Rating" = 2
7, "Defensive Rating" = 28)

#I used full join to join the data sets. The common variable used to join the data set w
as the name of the player.
NBA <- NBAsalaries %>%
  full_join(NBAplayerstats, by = c("NAME"))

# I removed NAs from the joined data set. 424 cases were dropped because the salary data
set included every single player in the league, but there were not game stats for every
 single player in the league in the other data set.
options(repos = c(CRAN = "http://cran.rstudio.com"))
NBAclean <- NBA %>% drop_na()
```

*Review: My cleaned, joined data set had 29 columns and 169 rows and a total of 593 observations.

```r
#use group by and summarise to determine mean, standard deviation, and counts of salary
 by NBA team
NBAclean %>% group_by(TEAM) %>% summarise(mean(`Salary ($)`), n = n(), sd('Salary'))
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion

## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm =
## na.rm): NAs introduced by coercion
```

```
## # A tibble: 16 x 4
##    TEAM  `mean(\`Salary ($)\`)`     n `sd("Salary")`
##    <chr>                  <dbl> <int>          <dbl>
##  1 Bos                10040634.    12             NA
##  2 Bro                 4246577.    14             NA
##  3 Den                11620785      9             NA
##  4 Det                 9519020.    12             NA
##  5 Gol                10927084     13             NA
##  6 Hou                15933234.     9             NA
##  7 Ind                 8325404.    11             NA
##  8 Lac                 6591601.    11             NA
##  9 Mil                 9860350.    10             NA
## 10 Okc                14269973.     8             NA
## 11 Orl                 8301316.     9             NA
## 12 Phi                 8903224.    12             NA
## 13 Por                13146703.     9             NA
## 14 San                10133440.     8             NA
## 15 Tor                13590404.    10             NA
## 16 Uta                 8169345.    12             NA
```

```
#the team with the highest mean salary is the Houston Rockets with a mean salary of $15, 933, 234.00. The Brooklyn Nets have the most players included in the dataset but have th e lowest mean salary of $4,246,577.00.

#The team with a highest standard deviationis Oklahoma (OKC) with a standard deviaiton o f $13,770, 410.00.

#use select to mutate and keep only numeric values and drop some stat variables

NBA_numeric <- NBAclean %>%
  #drop name,, team, position and other stat columns
  select(-c(1,3, 4, 9, 10, 11, 12,13, 24, 25,26 )) %>%
  #keep only numeric variables and use mutate if to make numeric
  mutate_if(is.character, as.numeric)

#calculate summary statistics for numeric variables Points PEr Game, Game points, age, a nd salary

mean(NBA_numeric$`Salary ($)`)
```

```
## [1] 9916752
```

```
#The mean salary of the players in the data set is $9,916,752.00.

sd(NBA_numeric$`Points Per Game`)
```

```
## [1] 7.166074
```

```
#The standard deviation of points per game is 7.166 points.
sd(NBA_numeric$`Salary ($)`)
```

```
## [1] 9276552
```

```
#The standard deviation of the salaries is $9,276,552.00

n_distinct(NBA_numeric$`Points Per Game`)
```

```
## [1] 113
```

```
#The number of distinct values in points per game is 113. This means many players had th
e same number of points per game during this season
n_distinct(NBA_numeric$`Salary ($)`)
```

```
## [1] 143
```

```
#The number of distinct values in the salary column is 143.

cor(NBA_numeric$`Points Per Game`, NBA_numeric$`Salary ($)`)
```

```
## [1] 0.6681942
```

```
#the correlation coefficient between the salary earned and points per game is 0.6682.
cor(NBA_numeric$MPG, NBA_numeric$`Salary ($)`)
```

```
## [1] 0.6278735
```

```
# The corerlation coefficient between the salary earned and minutes played per game is
 0.6279.

#Next, I created a new variable, points per minute using the existing variable points pe
r game and minutes played per game
NBA_numeric2 <- NBA_numeric %>%
mutate(
Points_Per_Minute = `Points Per Game`/`MPG`)

NBA_numeric2
```

```
## # A tibble: 169 x 19
##    `Salary ($)`   Age    GP   MPG `Percentage of Team Minutes… `2P%` `3PA` `3P%`
##           <dbl> <dbl> <dbl> <dbl>                         <dbl> <dbl> <dbl> <dbl>
##  1     37457154  31.2    22  38.4                          80.1 0.524   244 0.377
##  2     35665000  30.6     5  39.4                          82.2 0.377    34 0.324
##  3     35654150  34.1    11  36.1                          75.3 0.576    63 0.27
##  4     32700000  33.2    24  37.5                          78.2 0.526   145 0.359
##  5     31873932  30.2     2  29.1                          60.5 0.462    13 0.462
##  6     31214295  29.2     9  29.6                          61.7 0.435    24 0.375
##  7     30570000  29.8    11  38.6                          80.3 0.48    137 0.35
##  8     30560700  29.1     5  40.8                          85   0.537    47 0.319
##  9     30000000  30.7    12  36.8                          76.7 0.552    80 0.438
## 10     29230769  34.3    14  33.5                          69.8 0.513    38 0.316
## # … with 159 more rows, and 11 more variables: Effective Shooting % <dbl>,
## #   True Shooting % <dbl>, Points Per Game <dbl>, Rebounds Per Game <dbl>,
## #   Total Rebound Percentage <dbl>, Assists Per Game <dbl>, Assists % <dbl>,
## #   Versatility Index <dbl>, Offensive Rating <dbl>, Defensive Rating <dbl>,
## #   Points_Per_Minute <dbl>
```

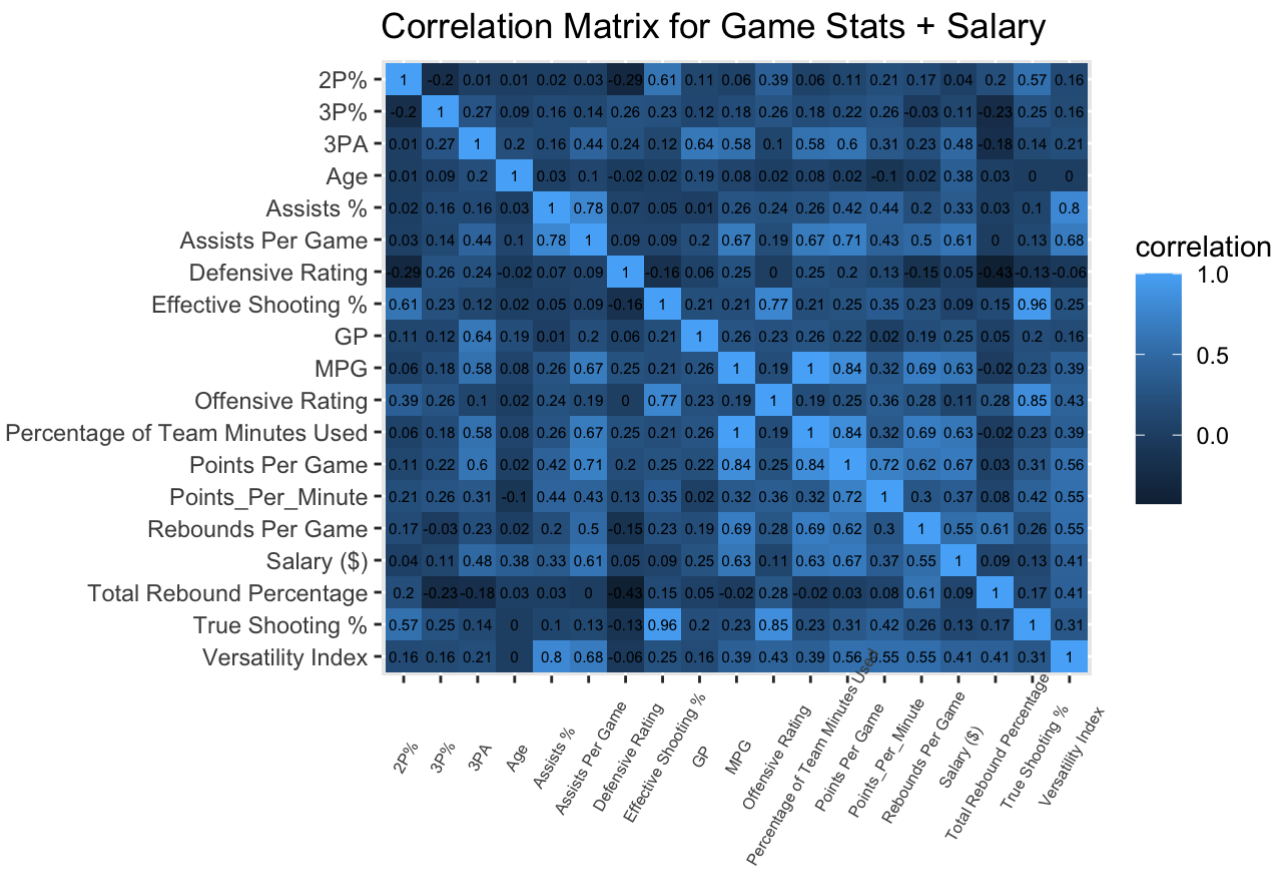
```
#use filter and arrange to view the number and stats of players who score greater than a verage points per game and are older than the average age
mean(NBA_numeric$Age)
```

```
## [1] 27.80704
```

```
#The average age of the players in the data set is 27.8 years.
NBA_ppg_by_age <- NBA_numeric2 %>%
filter(`Points Per Game` > 9.97 & Age > 27.8) %>% arrange(desc(Age))

#Seventy players scored over 9.97 points per game and are over the average
```

Review:I removed the structure statistics from the project.

```
#no scientific notation
options(scipen=999)

#make a correlation heat map
cor(NBA_numeric2) %>%
# Save as a data frame
as.data.frame %>%
# Convert row names to an explicit variable
rownames_to_column %>%
# Pivot so that all correlations appear in the same column
pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>% # Specify variab
les are displayed alphabetically from top to bottom
ggplot(aes(rowname, factor(other_var, levels = rev(levels(factor(other_var)))), fill=cor
relation))  +
# Heatmap with geom_tile
geom_tile() +
# Change the scale to make the middle appear neutral scale_fill_gradient2(low="red",mid
="white",high="blue")           + # Overlay values
geom_text(aes(label = round(correlation,2)), color = "black", size = 2) +
# Give title and labels
labs(title = "Correlation Matrix for Game Stats + Salary", x = "", y = "") + theme(axis.
text.x  = element_text(angle=60,vjust=0.7, size=6))
```

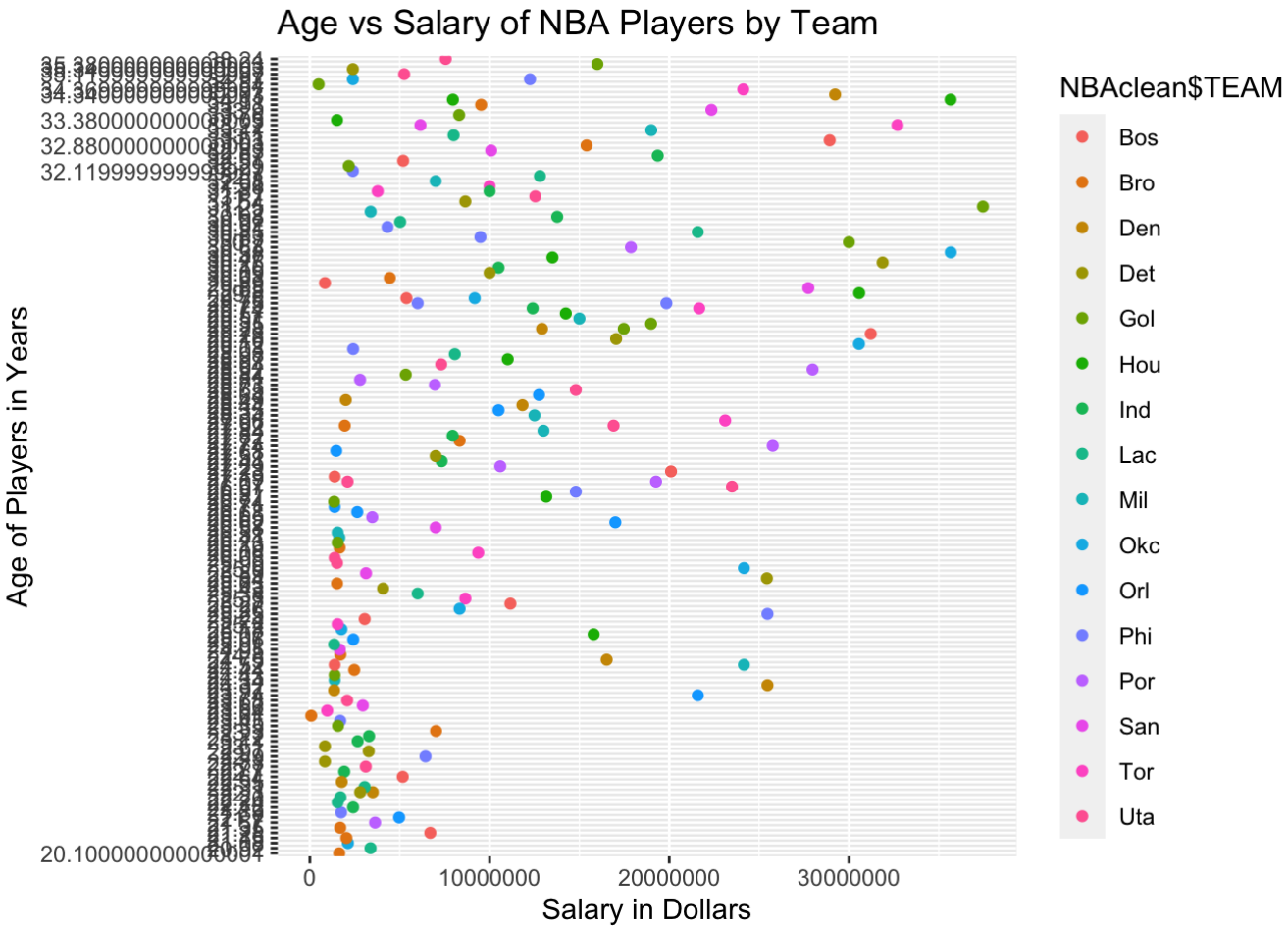


```
#additional plot
ggplot() + geom_point(data = NBAclean, aes(x =NBAclean$`Salary ($)`, y = NBAclean$Age, c
olor = NBAclean$TEAM)) +    ylab("Age of Players in Years") + xlab("Salary in Dollars")
 + ggtitle("Age vs Salary of NBA Players by Team")
```

```
## Warning: Use of `NBAclean$`Salary ($)`` is discouraged. Use `Salary ($)`
## instead.
```

```
## Warning: Use of `NBAclean$Age` is discouraged. Use `Age` instead.
```

```
## Warning: Use of `NBAclean$TEAM` is discouraged. Use `TEAM` instead.
```

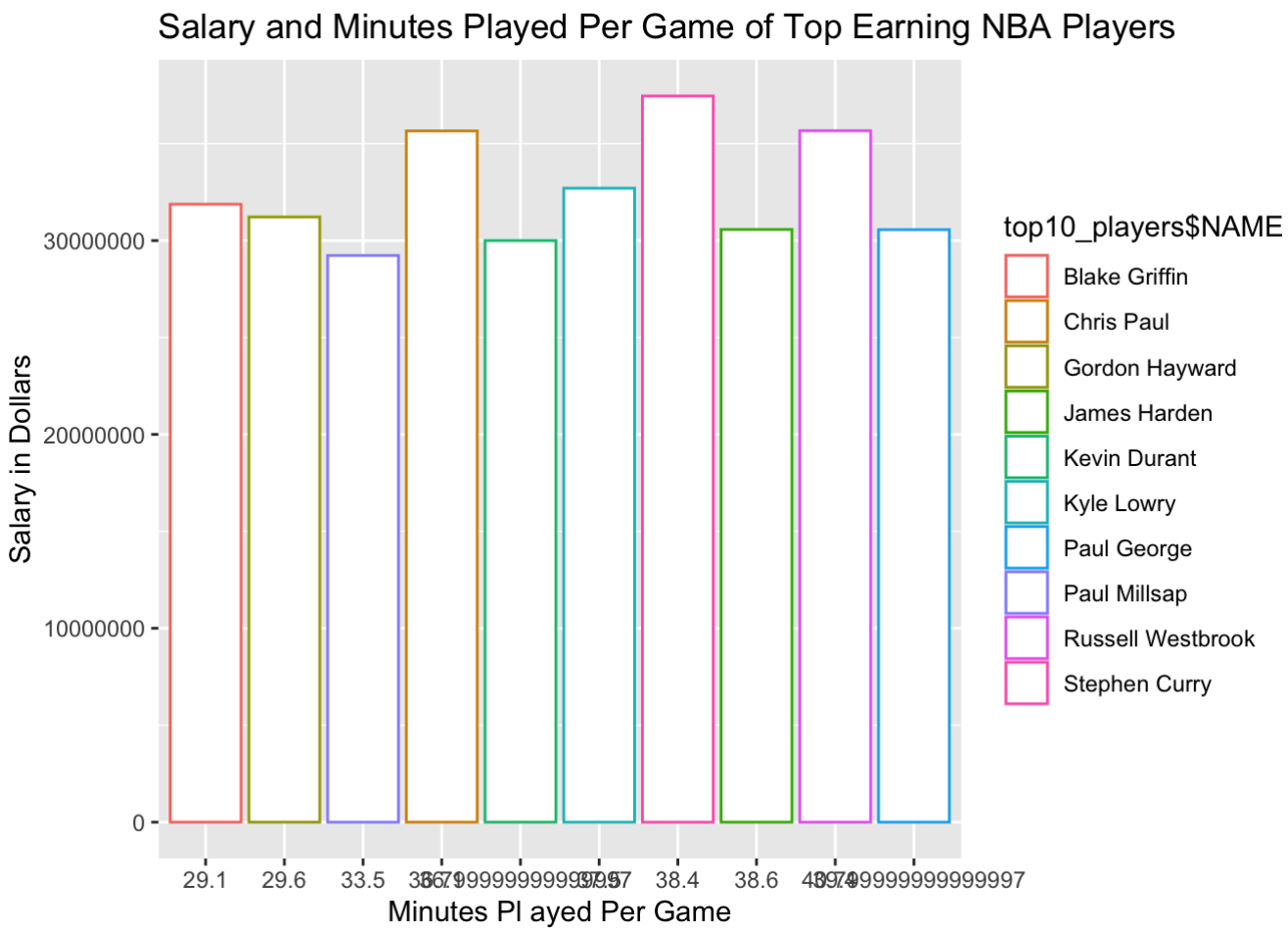



```
#this plot shows how salary and age are correlated by team
#additional plot
#I created a subset of the dataset with the top 10 earners in the league.
library(dplyr)
top10_players <- NBAclean[c(1:10),]

ggplot(data=top10_players, aes(x= top10_players$'MPG', y=top10_players$`Salary ($)`, col
or = top10_players$NAME)) +
geom_bar(stat="identity", fill="white") + ylab("Salary in Dollars") + xlab("Minutes Pl a
yed Per Game") +
ggtitle("Salary and Minutes Played Per Game of Top Earning NBA Players")
```

```
## Warning: Use of `top10_players$MPG` is discouraged. Use `MPG` instead.
```

```
## Warning: Use of `top10_players$`Salary ($)`` is discouraged. Use `Salary ($)`
## instead.
```

```
## Warning: Use of `top10_players$NAME` is discouraged. Use `NAME` instead.
```




```
#This plot showed the relationship between salary and minutes played per game of the top
earning players in the league.


#make a covariance matrix
NBA_numeric2 %>%
select_if(is.numeric) %>% cov
```

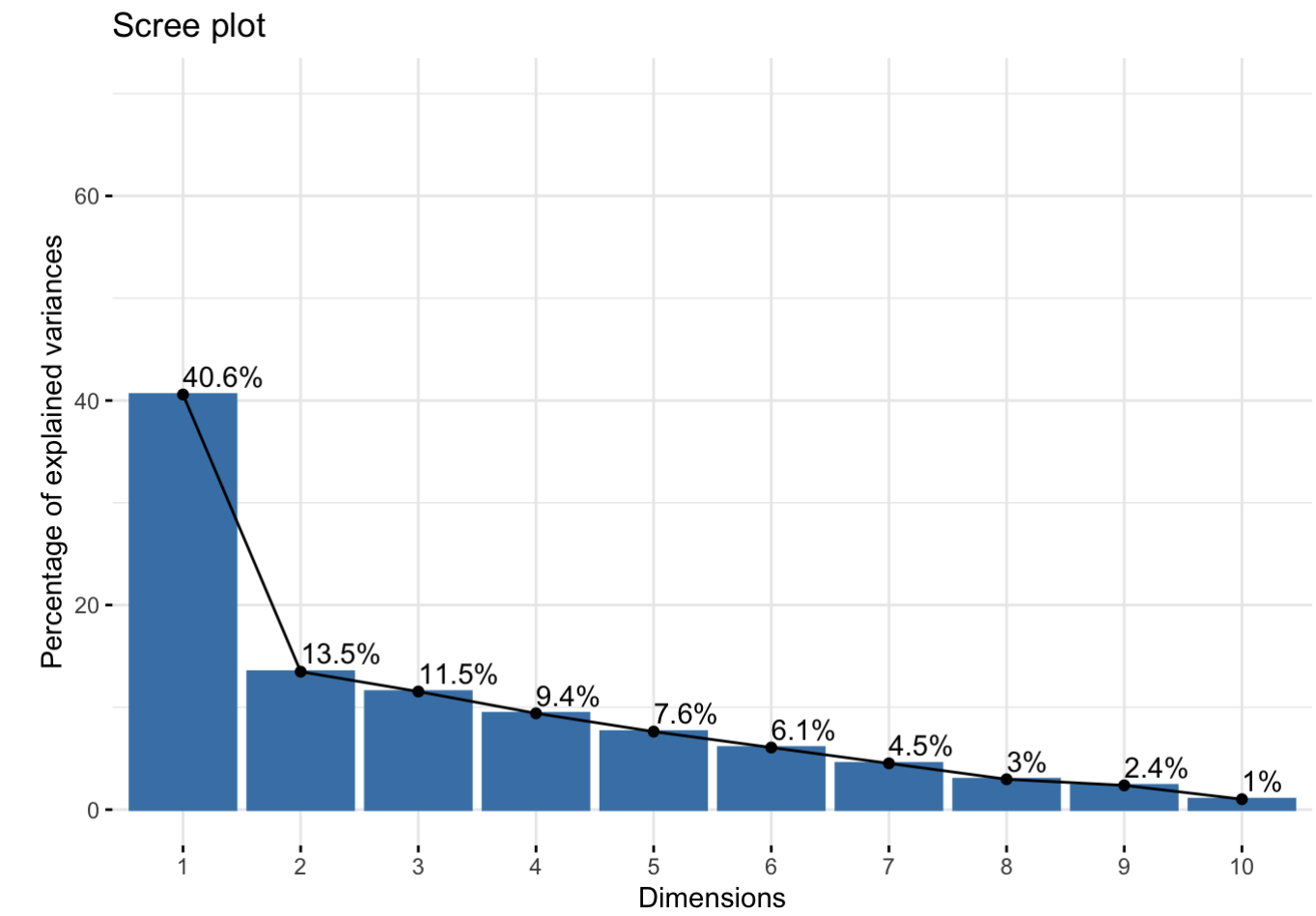
```
##                                    Salary ($)                 Age
## Salary ($)                    86054417910637.19 13697114.8570681885
## Age                               13697114.86       14.9728876233
## GP                                14161504.18        4.4682294308
## MPG                               63029432.78        3.3675425824
## Percentage of Team Minutes Used  131308612.61        6.9951723373
## 2P%                                  61487.62        0.0034122327
## 3PA                              167471114.84       28.4509411102
## 3P%                                 206989.53        0.0649701895
## Effective Shooting %                100061.97        0.0108392716
## True Shooting %                     131697.02        0.0001759062
## Points Per Game                   44419181.79        0.6091703297
## Rebounds Per Game                 13902730.97        0.1995200056
## Total Rebound Percentage           4165096.34        0.4936995985
## Assists Per Game                  11190672.97        0.7292620456
## Assists %                         33723616.18        1.1978914835
## Versatility Index                 11193960.00       -0.0251455340
## Offensive Rating                  19415303.84        1.4763485841
## Defensive Rating                   2688438.58       -0.4854183221
## Points_Per_Minute                   563966.36       -0.0637486161
##                                            GP              MPG
## Salary ($)                    14161504.18283319 63029432.7816392
## Age                                  4.46822943        3.3675426
## GP                                  36.77550014       17.2993132
## MPG                                 17.29931319      117.1033333
## Percentage of Team Minutes Used     36.10593477      244.0106731
## 2P%                                  0.10552719        0.1061468
## 3PA                                144.38620034      232.2906136
## 3P%                                  0.13862345        0.3834734
## Effective Shooting %                 0.15550507        0.2761825
## True Shooting %                      0.13540113        0.2750307
## Points Per Game                      9.49249084       65.1020833
## Rebounds Per Game                    3.18041702       20.5516163
## Total Rebound Percentage             1.57326360       -1.0573489
## Assists Per Game                     2.37530642       14.4260623
## Assists %                            0.65132784       30.7533333
## Versatility Index                    2.94001127       12.3350549
## Offensive Rating                    27.49700972       40.3742170
## Defensive Rating                     1.93583404       14.3638278
## Points_Per_Minute                    0.02098279        0.5735451
##                                Percentage of Team Minutes Used             2P%
## Salary ($)                                  131308612.6131129 61487.619532333
## Age                                                 6.9951723     0.003412233
## GP                                                 36.1059348     0.105527191
## MPG                                               244.0106731     0.106146841
## Percentage of Team Minutes Used                   508.4552966     0.221187465
## 2P%                                                 0.2211875     0.023858292
## 3PA                                               484.3225768     0.064534305
## 3P%                                                 0.7996372    -0.006096545
## Effective Shooting %                                0.5759648     0.011621457
## True Shooting %                                     0.5736207     0.009648338
## Points Per Game                                   135.6606090     0.123103434
## Rebounds Per Game                                  42.8041226     0.071015596
```

```
## Total Rebound Percentage                          -2.2561042     0.158388789
## Assists Per Game                                  30.0593160     0.009190494
## Assists %                                         64.1259844     0.033028846
## Versatility Index                                 25.7043054     0.071847284
## Offensive Rating                                  84.2056301     1.181135091
## Defensive Rating                                  29.9724119    -0.238097380
## Points_Per_Minute                                  1.1954594     0.005371794
##                                              3PA             3P%
## Salary ($)                       167471114.83999014 206989.532781312
## Age                                     28.45094111      0.064970189
## GP                                     144.38620034      0.138623450
## MPG                                    232.29061355      0.383473397
## Percentage of Team Minutes Used        484.32257678      0.799637169
## 2P%                                      0.06453431     -0.006096545
## 3PA                                   1391.03402367      1.959826043
## 3P%                                      1.95982604      0.038073190
## Effective Shooting %                     0.56816455      0.005644942
## True Shooting %                          0.59766272      0.005493727
## Points Per Game                        161.40709707      0.308392399
## Rebounds Per Game                       23.26133418     -0.016277684
## Total Rebound Percentage               -34.60666737     -0.228295978
## Assists Per Game                        32.27799732      0.053684094
## Assists %                               65.45961538      0.332414011
## Versatility Index                       22.91686038      0.089798640
## Offensive Rating                        72.05556143      1.004053783
## Defensive Rating                        49.04290645      0.272701180
## Points_Per_Minute                        1.90868991      0.008370588
##                                 Effective Shooting %   True Shooting %
## Salary ($)                      100061.974363976 131697.0156048183
## Age                                     0.010839272      0.0001759062
## GP                                      0.155505072      0.1354011341
## MPG                                     0.276182463      0.2750306777
## Percentage of Team Minutes Used         0.575964828      0.5736207347
## 2P%                                     0.011621457      0.0096483380
## 3PA                                     0.568164553      0.5976627219
## 3P%                                     0.005644942      0.0054937270
## Effective Shooting %                    0.015409991      0.0131213463
## True Shooting %                         0.013121346      0.0122212487
## Points Per Game                         0.223881044      0.2461054487
## Rebounds Per Game                       0.078275775      0.0780022401
## Total Rebound Percentage                0.095039434      0.0964469428
## Assists Per Game                        0.021865205      0.0286125493
## Assists %                               0.063542674      0.1203004121
## Versatility Index                       0.090855748      0.0997996584
## Offensive Rating                        1.861348172      1.8217925402
## Defensive Rating                       -0.110222855     -0.0775871865
## Points_Per_Minute                       0.007211344      0.0076640239
##                                  Points Per Game Rebounds Per Game
## Salary ($)                      44419181.7878663 13902730.96541631
## Age                                    0.6091703        0.19952001
## GP                                     9.4924908        3.18041702
## MPG                                   65.1020833       20.55161630
## Percentage of Team Minutes Used      135.6606090       42.80412264
## 2P%                                    0.1231034        0.07101560
```

```
## 3PA                                161.4070971       23.26133418
## 3P%                                  0.3083924       -0.01627768
## Effective Shooting %                 0.2238810        0.07827577
## True Shooting %                      0.2461054        0.07800224
## Points Per Game                     51.3526190       12.17860806
## Rebounds Per Game                   12.1786081        7.55163779
## Total Rebound Percentage             0.9099496        8.43839708
## Assists Per Game                    10.1067491        2.69683749
## Assists %                           32.4092262        6.08712912
## Versatility Index                   11.8429579        4.44776627
## Offensive Rating                    35.0154991       15.14764335
## Defensive Rating                     7.7681685       -2.18178395
## Points_Per_Minute                    0.8492652        0.13367927
##                              Total Rebound Percentage   Assists Per Game
## Salary ($)                        4165096.337450691 11190672.974098338
## Age                                     0.493699598        0.729262046
## GP                                      1.573263595        2.375306424
## MPG                                    -1.057348901       14.426062271
## Percentage of Team Minutes Used        -2.256104184       30.059316005
## 2P%                                     0.158388789        0.009190494
## 3PA                                   -34.606667371       32.277997323
## 3P%                                    -0.228295978        0.053684094
## Effective Shooting %                    0.095039434        0.021865205
## True Shooting %                         0.096446943        0.028612549
## Points Per Game                         0.909949634       10.106749084
## Rebounds Per Game                       8.438397084        2.696837489
## Total Rebound Percentage               25.677020287        0.006891378
## Assists Per Game                        0.006891378        3.920758664
## Assists %                               1.402898352       16.841858974
## Versatility Index                       6.056266202        3.951478233
## Offensive Rating                       27.731319034        7.168345661
## Defensive Rating                      -11.637193928        1.001353550
## Points_Per_Minute                       0.067288600        0.139766043
##                                    Assists % Versatility Index
## Salary ($)                   33723616.17783883 11193959.99748873
## Age                                 1.19789148       -0.02514553
## GP                                  0.65132784        2.94001127
## MPG                                30.75333333       12.33505495
## Percentage of Team Minutes Used    64.12598443       25.70430544
## 2P%                                 0.03302885        0.07184728
## 3PA                                65.45961538       22.91686038
## 3P%                                 0.33241401        0.08979864
## Effective Shooting %                0.06354267        0.09085575
## True Shooting %                     0.12030041        0.09979966
## Points Per Game                    32.40922619       11.84295788
## Rebounds Per Game                   6.08712912        4.44776627
## Total Rebound Percentage            1.40289835        6.05626620
## Assists Per Game                   16.84185897        3.95147823
## Assists %                         118.44404762       25.68932234
## Versatility Index                  25.68932234        8.70161595
## Offensive Rating                   51.09501374       24.74732530
## Defensive Rating                    4.17366300       -0.95661031
## Points_Per_Minute                   0.77848726        0.26697594
##                              Offensive Rating Defensive Rating
```

```
## Salary ($)                      19415303.8357284 2688438.57904692
## Age                                    1.4763486      -0.48541832
## GP                                    27.4970097       1.93583404
## MPG                                   40.3742170      14.36382784
## Percentage of Team Minutes Used       84.2056301      29.97241195
## 2P%                                    1.1811351      -0.23809738
## 3PA                                   72.0555614      49.04290645
## 3P%                                    1.0040538       0.27270118
## Effective Shooting %                   1.8613482      -0.11022286
## True Shooting %                        1.8217925      -0.07758719
## Points Per Game                       35.0154991       7.76816850
## Rebounds Per Game                     15.1476434      -2.18178395
## Total Rebound Percentage              27.7313190     -11.63719393
## Assists Per Game                       7.1683457       1.00135355
## Assists %                             51.0950137       4.17366300
## Versatility Index                     24.7473253      -0.95661031
## Offensive Rating                     378.2615279       0.36052480
## Defensive Rating                       0.3605248      29.09381305
## Points_Per_Minute                      1.1640395       0.11652427
##                                 Points_Per_Minute
## Salary ($)                       563966.357075407
## Age                                -0.063748616
## GP                                  0.020982788
## MPG                                 0.573545076
## Percentage of Team Minutes Used     1.195459365
## 2P%                                 0.005371794
## 3PA                                 1.908689907
## 3P%                                 0.008370588
## Effective Shooting %                0.007211344
## True Shooting %                     0.007664024
## Points Per Game                     0.849265243
## Rebounds Per Game                   0.133679268
## Total Rebound Percentage            0.067288600
## Assists Per Game                    0.139766043
## Assists %                           0.778487263
## Versatility Index                   0.266975945
## Offensive Rating                    1.164039474
## Defensive Rating                    0.116524272
## Points_Per_Minute                   0.026915845
```

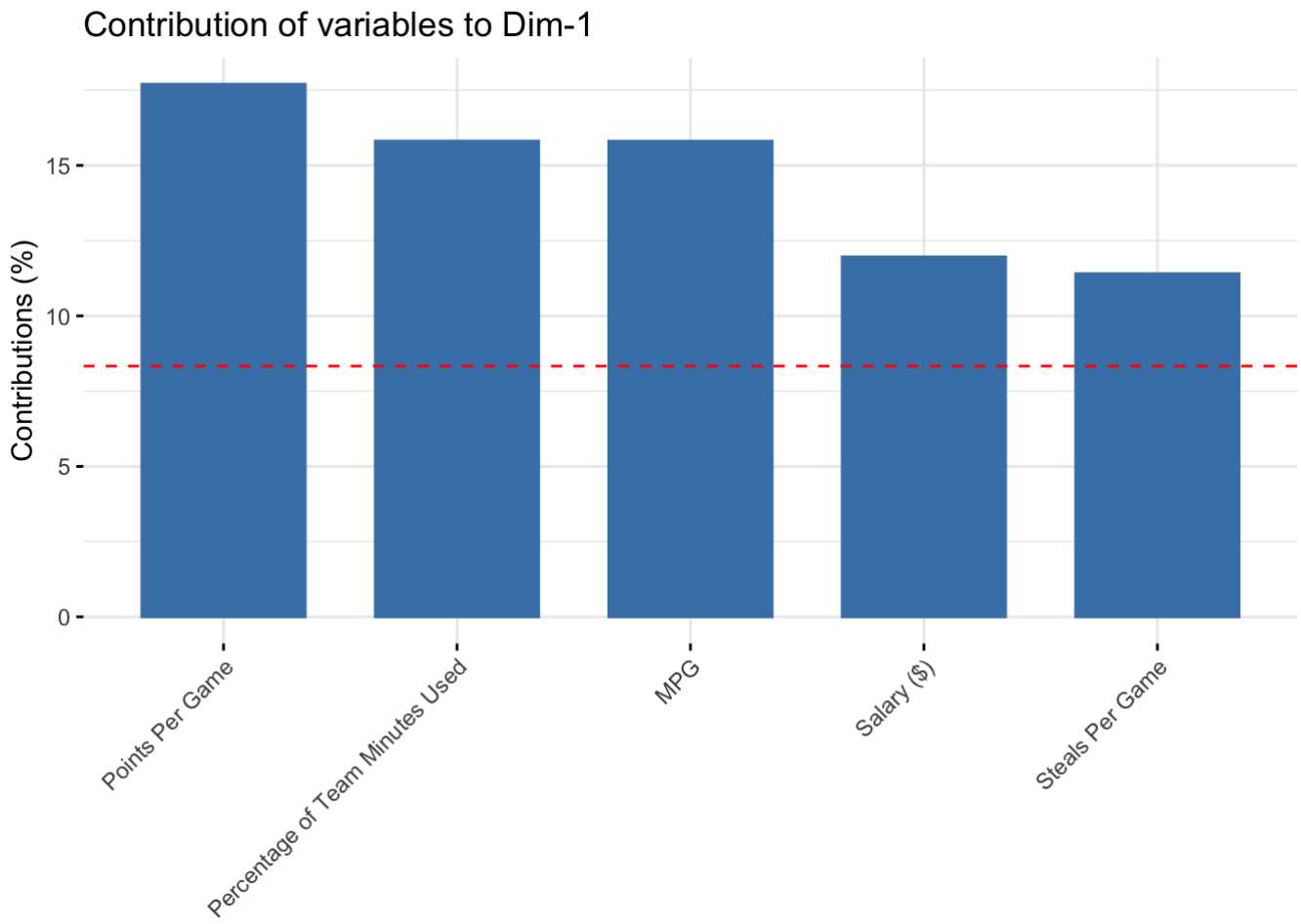
```
#Review: I am using PCA because I have many different variables, many of which are simil
ar. Because there are so many variables that are also  similar, PCA will reduce dimensio
nality.

# Visualize the eigenvalues and variances of the PCS in a table


NBA_numeric3 <- NBAclean %>%
#drop NAME, TEAM, Position and other stat columns
select(-c(1,3, 4,10, 11, 12,13,14, 15, 16, 17, 18, 20, 21, 22, 25,26, )) %>%
#keep only numeric variables
mutate_if(is.character, as.numeric)
# Prepare data for PCA and run PCA
pca <- NBA_numeric3 %>%
#  Scale to 0 mean and unit variance (standardize)
scale() %>%
prcomp()


# Visualize percentage of variances for each PC in a scree plot
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 70))
```



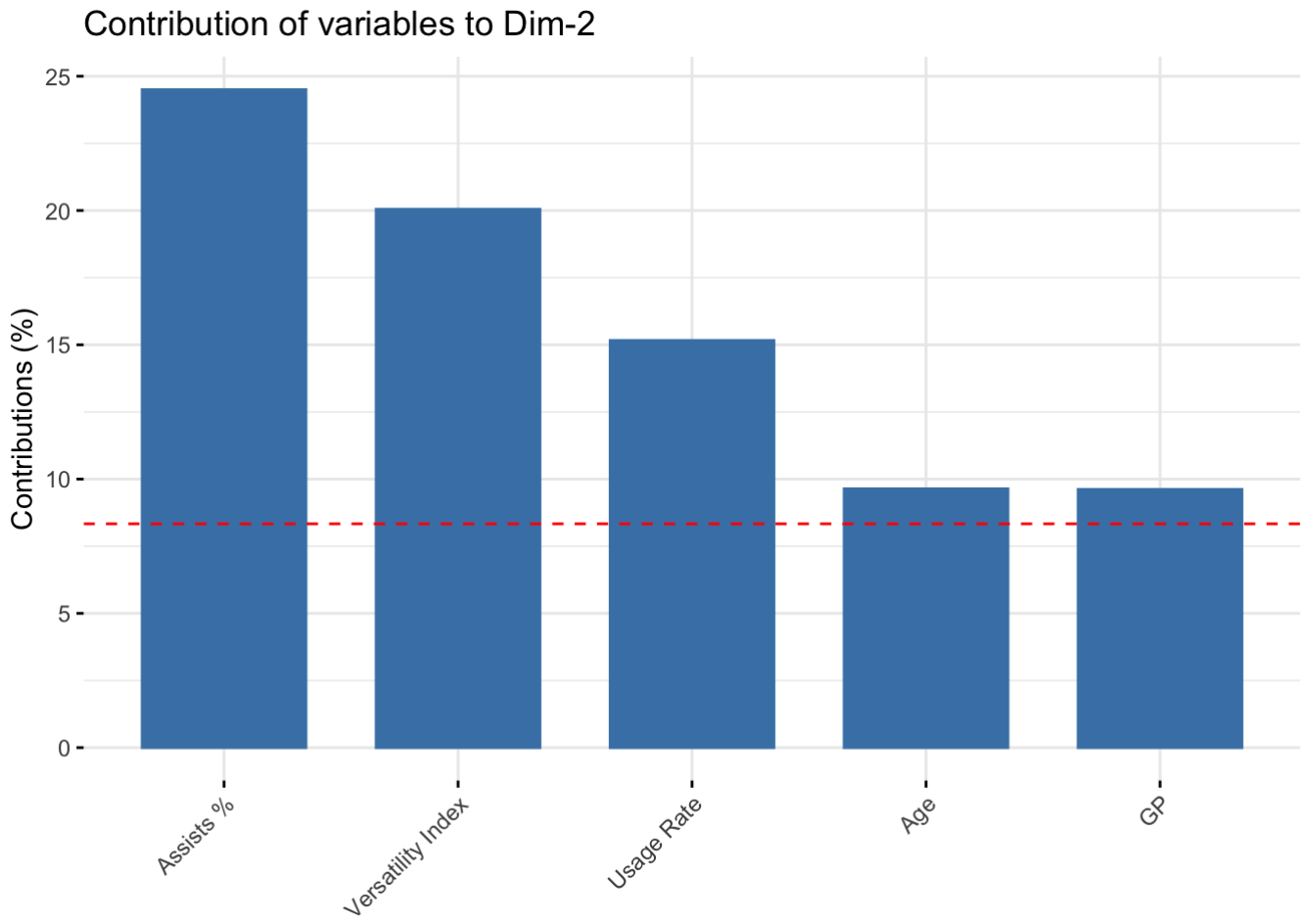

```
# Visualize the contributions of the variables to the PCs in a table
get_pca_var(pca)$contrib
```

```
##                                    Dim.1         Dim.2      Dim.3       Dim.4
## Salary ($)                     11.9621624  2.2707561101  0.2243728 12.93001496
## Age                             0.4055832  9.6382295612 10.3085120 38.07538525
## GP                              1.8085189  9.6154293563 17.7339495  1.56959334
## MPG                            15.8078391  7.0436783820  1.2198870  2.25491588
## Percentage of Team Minutes Used 15.8099454  7.0445191902  1.2176200  2.26758874
## Usage Rate                      5.8682892 15.1622638633 12.3922704  3.94404542
## Points Per Game                17.6986103  0.0003981622  1.6704535  0.44612424
## Assists %                       6.9561877 24.5024639019  2.1560473  0.67007007
## Steals Per Game                11.4037918  1.2760409096  1.5894049  0.89266090
## Versatility Index               9.6290474 20.0489465002  7.8788850  0.00946579
## Offensive Rating                1.8016179  0.8949222497 34.3806032 19.46368527
## Defensive Rating                0.8484067  2.5023518134  9.2279944 17.47645015
##                                     Dim.5       Dim.6        Dim.7      Dim.8
## Salary ($)                     0.33766351  0.01420129  7.149360932  6.6860615
## Age                           14.86123019  8.26282345  0.374436524  2.3088523
## GP                             2.77917865 59.33695547  6.129922285  0.3820796
## MPG                            2.61671841  1.09038969  0.003816022  6.1642652
## Percentage of Team Minutes Used  2.60885369  1.08416761  0.004493782  6.1507545
## Usage Rate                     3.25742868 13.04500383 12.034124350 13.4089545
## Points Per Game                0.23538450  1.04445048 10.145392972  0.5963011
## Assists %                      4.37340592  2.22844177 21.388002024  3.5995638
## Steals Per Game                8.00164330  2.32961702 25.498696197 41.5146715
## Versatility Index              0.10046960  0.08276100  0.736544219  5.6477154
## Offensive Rating               0.05382421  8.61796672 15.445572861 13.2574431
## Defensive Rating              60.77419934  2.86322167  1.089637832  0.2833376
##                                       Dim.9      Dim.10       Dim.11
## Salary ($)                    58.049674213  0.21030397  0.165427256
## Age                           15.744358053  0.01731108  0.003272549
## GP                             0.001127179  0.56038414  0.082847217
## MPG                            5.533302354  0.27694512  8.030190869
## Percentage of Team Minutes Used  5.517152915  0.28369252  7.969364601
## Usage Rate                     2.018410779  0.19734794 18.671846627
## Points Per Game                3.259760215  4.21607125 60.687020989
## Assists %                      0.240081336 33.86654793  0.019173033
## Steals Per Game                3.952732647  3.34373894  0.196997860
## Versatility Index              2.480438386 52.60209420  0.783624894
## Offensive Rating               1.426751351  1.50283175  3.154781335
## Defensive Rating               1.776210573  2.92273116  0.235452769
##                                      Dim.12
## Salary ($)                    0.0000009761751
## Age                           0.0000058409912
## GP                            0.0000144499581
## MPG                           49.9580519778487
## Percentage of Team Minutes Used 50.0418471506354
## Usage Rate                    0.0000143590161
## Points Per Game               0.0000323679536
## Assists %                     0.0000151540586
## Steals Per Game               0.0000040392877
## Versatility Index             0.0000076155649
## Offensive Rating              0.0000000301928
## Defensive Rating              0.0000060383178
```

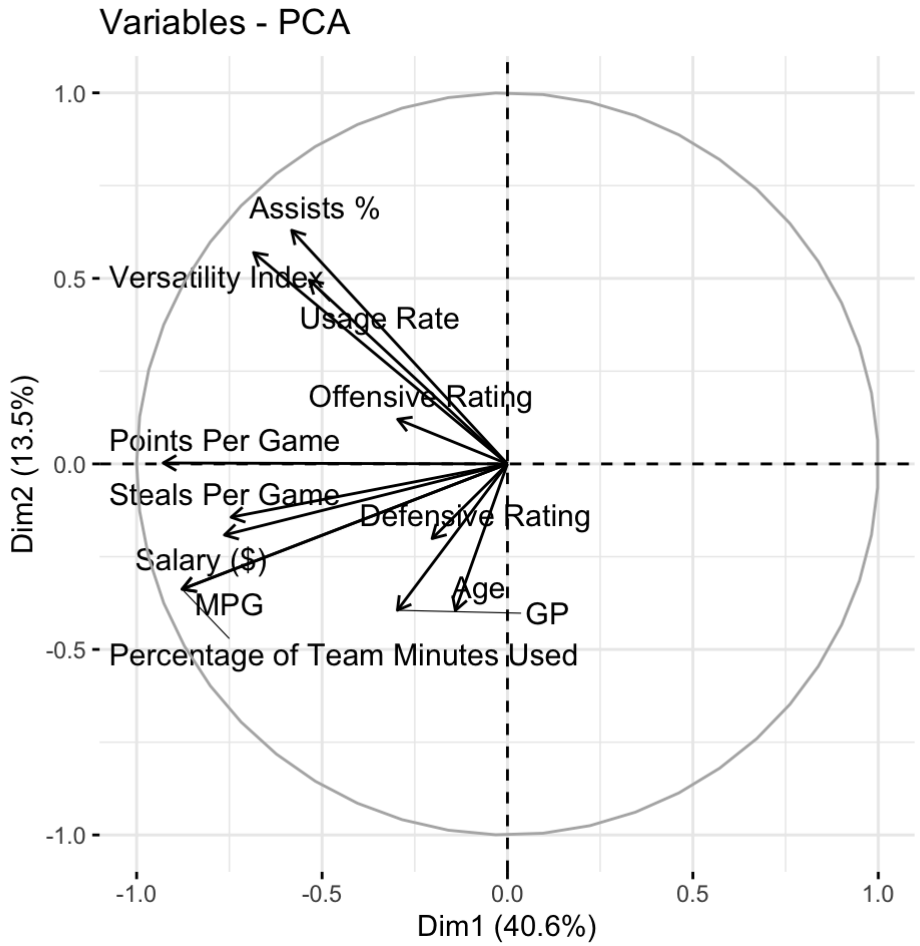
```
# Visualize the 5 top contributions of the variables to the PCs in a bar graph
  # Note the red dash line indicates the average contribution
fviz_contrib(pca, choice = "var", axes = 1, top = 5) # on PC1
```




```
fviz_contrib(pca, choice = "var", axes = 2, top = 5) # on PC2
```




```
# Visualize the contributions of the variables to the PCs in a correlation circle
fviz_pca_var(pca, col.var = "black",
             repel = TRUE) # Avoid text overlapping
```




```
# view results from PCA
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
#  Visualize the results
pca
```

```
## Standard deviations (1, .., p=12):
##  [1] 2.206923962 1.272120035 1.176942611 1.062863941 0.956841127 0.853380057
##  [7] 0.736739832 0.596600014 0.533666326 0.349538120 0.216385273 0.002173729
##
## Rotation (n x k) = (12 x 12):
##                                         PC1         PC2         PC3
## Salary ($)                      -0.34586359 -0.15069028  0.04736801
## Age                             -0.06368541 -0.31045498  0.32106872
## GP                              -0.13448118 -0.31008756  0.42111696
## MPG                             -0.39759073 -0.26539929 -0.11044850
## Percentage of Team Minutes Used -0.39761722 -0.26541513 -0.11034582
## Usage Rate                      -0.24224552  0.38938752 -0.35202657
## Points Per Game                 -0.42069716  0.00199540 -0.12924602
## Assists %                       -0.26374586  0.49499964  0.14683485
## Steals Per Game                 -0.33769501 -0.11296198 -0.12607160
## Versatility Index               -0.31030706  0.44776050  0.28069352
## Offensive Rating                -0.13422436  0.09460033  0.58634975
## Defensive Rating                -0.09210899 -0.15818824 -0.30377614
##                                         PC4         PC5         PC6
## Salary ($)                       0.359583300 -0.05810882 -0.01191692
## Age                              0.617052552 -0.38550266  0.28745127
## GP                              -0.125283412 -0.16670869 -0.77030485
## MPG                             -0.150163773  0.16176274  0.10442173
## Percentage of Team Minutes Used -0.150585150  0.16151946  0.10412337
## Usage Rate                       0.198596209 -0.18048348 -0.36117868
## Points Per Game                 -0.066792532  0.04851644 -0.10219836
## Assists %                        0.081857808 -0.20912690  0.14927966
## Steals Per Game                  0.094480733  0.28287176  0.15263083
## Versatility Index               -0.009729229 -0.03169694 -0.02876821
## Offensive Rating                -0.441176668  0.02320004  0.29356374
## Defensive Rating                -0.418048444 -0.77957809  0.16921057
##                                         PC7         PC8          PC9
## Salary ($)                       0.267382889 -0.25857420  0.761903368
## Age                              0.061191219  0.15194908 -0.396791608
## GP                              -0.247586799  0.06181259 -0.003357349
## MPG                             -0.006177396 -0.24827938 -0.235229725
## Percentage of Team Minutes Used -0.006703568 -0.24800715 -0.234886205
## Usage Rate                       0.346902354  0.36618239 -0.142070784
## Points Per Game                  0.318518335  0.07722053 -0.180548060
## Assists %                       -0.462471643 -0.18972516  0.048998096
## Steals Per Game                 -0.504962337  0.64431880  0.198814804
## Versatility Index               -0.085822154 -0.23764922 -0.157494076
## Offensive Rating                 0.393008560  0.36410772  0.119446697
## Defensive Rating                -0.104385719  0.05322947  0.133274550
##                                        PC10        PC11           PC12
## Salary ($)                       0.04585891 -0.04067275  0.00009880158
## Age                              0.01315716  0.00572062  0.00024168143
## GP                              -0.07485881 -0.02878319 -0.00038013101
## MPG                             -0.05262558 -0.28337591 -0.70681010164
## Percentage of Team Minutes Used -0.05326279 -0.28230063  0.70740262334
## Usage Rate                      -0.04442386 -0.43210932  0.00037893292
## Points Per Game                 -0.20533074  0.77901875 -0.00056892841
## Assists %                       -0.58194972 -0.01384667 -0.00038928214
```

```
## Steals Per Game                    0.18285893  0.04438444 -0.00020097979
## Versatility Index                  0.72527301  0.08852259  0.00027596313
## Offensive Rating                  -0.12259004 -0.17761704  0.00001737607
## Defensive Rating                   0.17095997  0.04852348 -0.00024572989
```

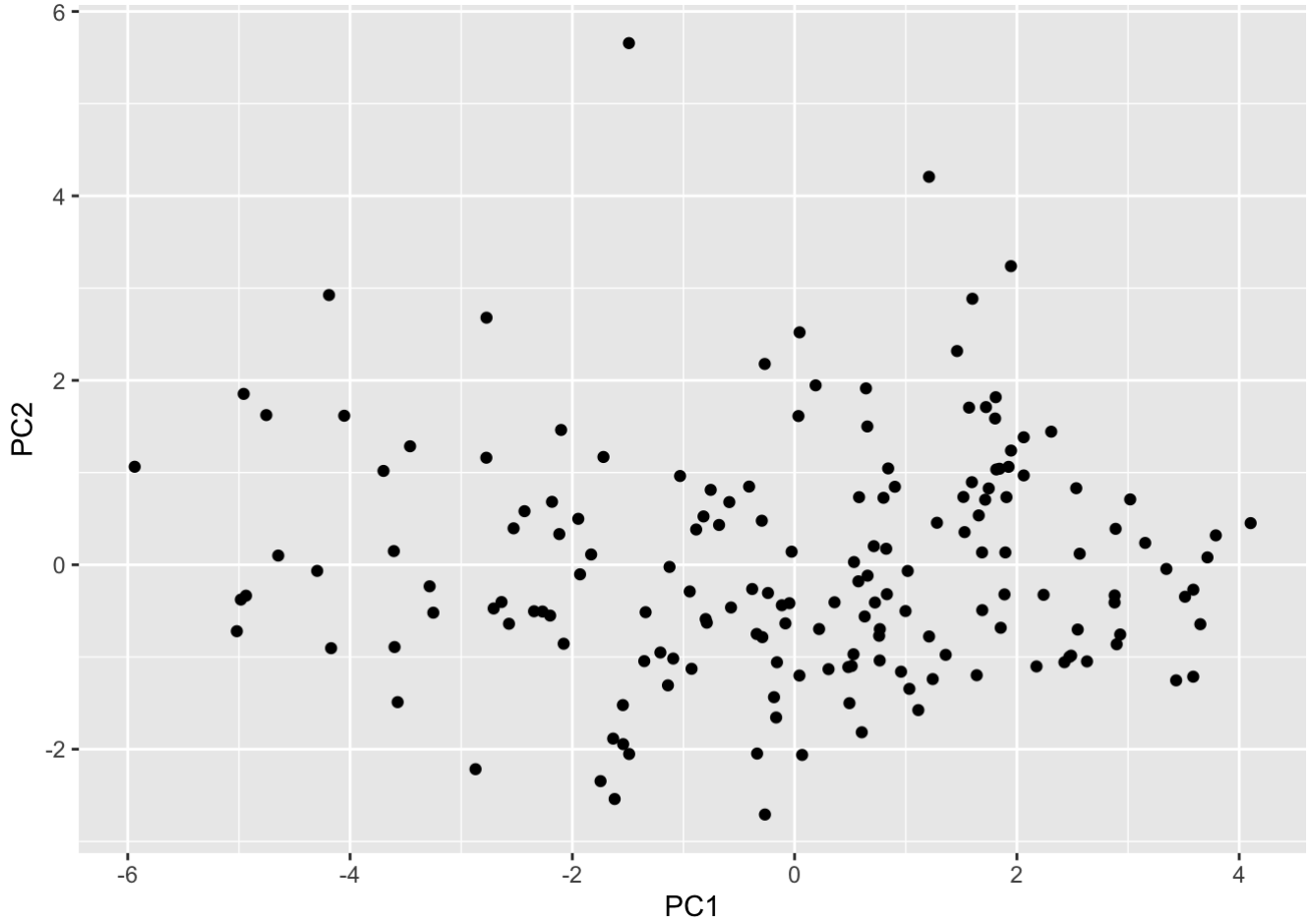
```
# Visualize the rotated data
head(pca$x)
```

```
##            PC1        PC2        PC3       PC4         PC5        PC6
## [1,] -5.018674 -0.7205645  0.6566582 0.5465537 -1.18665992 -1.3593977
## [2,] -4.957130  1.8532591 -0.3152085 1.8358828 -0.41844192  0.4937030
## [3,] -4.170430 -0.9045077  0.2576625 1.8613509 -0.05170314  0.7856883
## [4,] -3.571787 -1.4896253  1.8558630 1.0795503 -0.26713624 -0.8177767
## [5,] -4.189173  2.9250806  0.1679974 2.0496718  0.07411503  0.5189288
## [6,] -1.139847 -1.3075263 -0.1244741 0.4512135 -0.61611786  0.5424844
##             PC7        PC8       PC9       PC10        PC11         PC12
## [1,]  0.8521410 -0.1499074 0.7489948 -0.2092819  0.39353013  0.002600396
## [2,]  0.1644823 -1.3972890 0.3762989 -0.4671300 -0.15209484  0.004491845
## [3,] -0.7419946  0.6693132 1.1459702  0.3317596 -0.09172399  0.003215630
## [4,] -0.7035050 -0.3944264 0.7774890 -0.5653247 -0.35615897  0.002061286
## [5,]  0.9370303 -0.3772936 0.4055050 -0.5272902  0.20946439 -0.003030879
## [6,]  0.2902759 -1.0228472 1.5267919  0.1370699 -0.17808561  0.001574204
```

```
# Add the information about the different groups back into PCA data
pca_data <- data.frame(pca$x)
head(pca_data)
```

```
##         PC1        PC2        PC3       PC4         PC5        PC6        PC7
## 1 -5.018674 -0.7205645  0.6566582 0.5465537 -1.18665992 -1.3593977  0.8521410
## 2 -4.957130  1.8532591 -0.3152085 1.8358828 -0.41844192  0.4937030  0.1644823
## 3 -4.170430 -0.9045077  0.2576625 1.8613509 -0.05170314  0.7856883 -0.7419946
## 4 -3.571787 -1.4896253  1.8558630 1.0795503 -0.26713624 -0.8177767 -0.7035050
## 5 -4.189173  2.9250806  0.1679974 2.0496718  0.07411503  0.5189288  0.9370303
## 6 -1.139847 -1.3075263 -0.1244741 0.4512135 -0.61611786  0.5424844  0.2902759
##          PC8       PC9       PC10        PC11         PC12
## 1 -0.1499074 0.7489948 -0.2092819  0.39353013  0.002600396
## 2 -1.3972890 0.3762989 -0.4671300 -0.15209484  0.004491845
## 3  0.6693132 1.1459702  0.3317596 -0.09172399  0.003215630
## 4 -0.3944264 0.7774890 -0.5653247 -0.35615897  0.002061286
## 5 -0.3772936 0.4055050 -0.5272902  0.20946439 -0.003030879
## 6 -1.0228472 1.5267919  0.1370699 -0.17808561  0.001574204
```

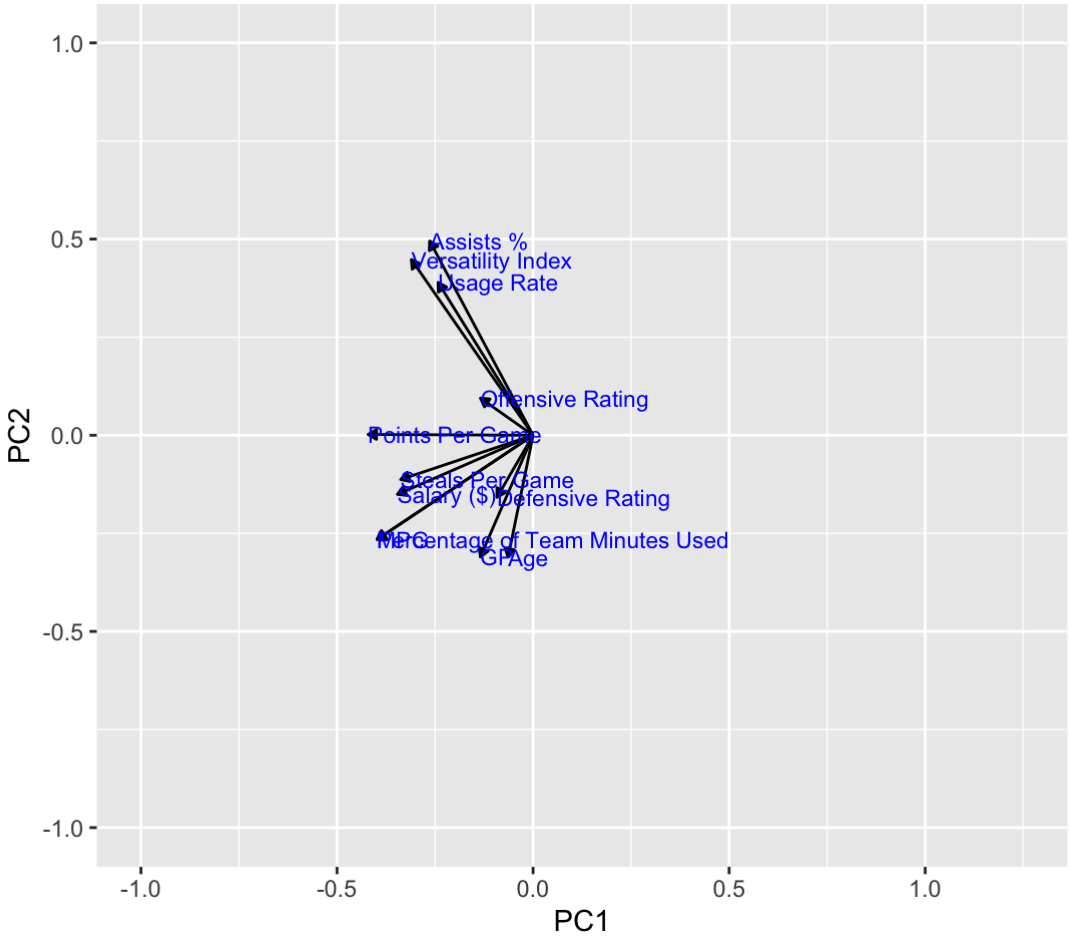
```
# Plot the data according to the new coordinate system: PC1 and PC2
ggplot(pca_data, aes(x = PC1, y = PC2)) +
geom_point()
```




```
# Take a look at the rotation matrix
pca$rotation
```

```
##                                        PC1         PC2         PC3
## Salary ($)                      -0.34586359 -0.15069028  0.04736801
## Age                             -0.06368541 -0.31045498  0.32106872
## GP                              -0.13448118 -0.31008756  0.42111696
## MPG                             -0.39759073 -0.26539929 -0.11044850
## Percentage of Team Minutes Used -0.39761722 -0.26541513 -0.11034582
## Usage Rate                      -0.24224552  0.38938752 -0.35202657
## Points Per Game                 -0.42069716  0.00199540 -0.12924602
## Assists %                       -0.26374586  0.49499964  0.14683485
## Steals Per Game                 -0.33769501 -0.11296198 -0.12607160
## Versatility Index               -0.31030706  0.44776050  0.28069352
## Offensive Rating                -0.13422436  0.09460033  0.58634975
## Defensive Rating                -0.09210899 -0.15818824 -0.30377614
##                                         PC4         PC5         PC6
## Salary ($)                       0.359583300 -0.05810882 -0.01191692
## Age                              0.617052552 -0.38550266  0.28745127
## GP                              -0.125283412 -0.16670869 -0.77030485
## MPG                             -0.150163773  0.16176274  0.10442173
## Percentage of Team Minutes Used -0.150585150  0.16151946  0.10412337
## Usage Rate                       0.198596209 -0.18048348 -0.36117868
## Points Per Game                 -0.066792532  0.04851644 -0.10219836
## Assists %                        0.081857808 -0.20912690  0.14927966
## Steals Per Game                  0.094480733  0.28287176  0.15263083
## Versatility Index               -0.009729229 -0.03169694 -0.02876821
## Offensive Rating                -0.441176668  0.02320004  0.29356374
## Defensive Rating                -0.418048444 -0.77957809  0.16921057
##                                         PC7         PC8          PC9
## Salary ($)                       0.267382889 -0.25857420  0.761903368
## Age                              0.061191219  0.15194908 -0.396791608
## GP                              -0.247586799  0.06181259 -0.003357349
## MPG                             -0.006177396 -0.24827938 -0.235229725
## Percentage of Team Minutes Used -0.006703568 -0.24800715 -0.234886205
## Usage Rate                       0.346902354  0.36618239 -0.142070784
## Points Per Game                  0.318518335  0.07722053 -0.180548060
## Assists %                       -0.462471643 -0.18972516  0.048998096
## Steals Per Game                 -0.504962337  0.64431880  0.198814804
## Versatility Index               -0.085822154 -0.23764922 -0.157494076
## Offensive Rating                 0.393008560  0.36410772  0.119446697
## Defensive Rating                -0.104385719  0.05322947  0.133274550
##                                        PC10        PC11           PC12
## Salary ($)                       0.04585891 -0.04067275  0.00009880158
## Age                              0.01315716  0.00572062  0.00024168143
## GP                              -0.07485881 -0.02878319 -0.00038013101
## MPG                             -0.05262558 -0.28337591 -0.70681010164
## Percentage of Team Minutes Used -0.05326279 -0.28230063  0.70740262334
## Usage Rate                      -0.04442386 -0.43210932  0.00037893292
## Points Per Game                 -0.20533074  0.77901875 -0.00056892841
## Assists %                       -0.58194972 -0.01384667 -0.00038928214
## Steals Per Game                  0.18285893  0.04438444 -0.00020097979
## Versatility Index                0.72527301  0.08852259  0.00027596313
## Offensive Rating                -0.12259004 -0.17761704  0.00001737607
## Defensive Rating                 0.17095997  0.04852348 -0.00024572989
```

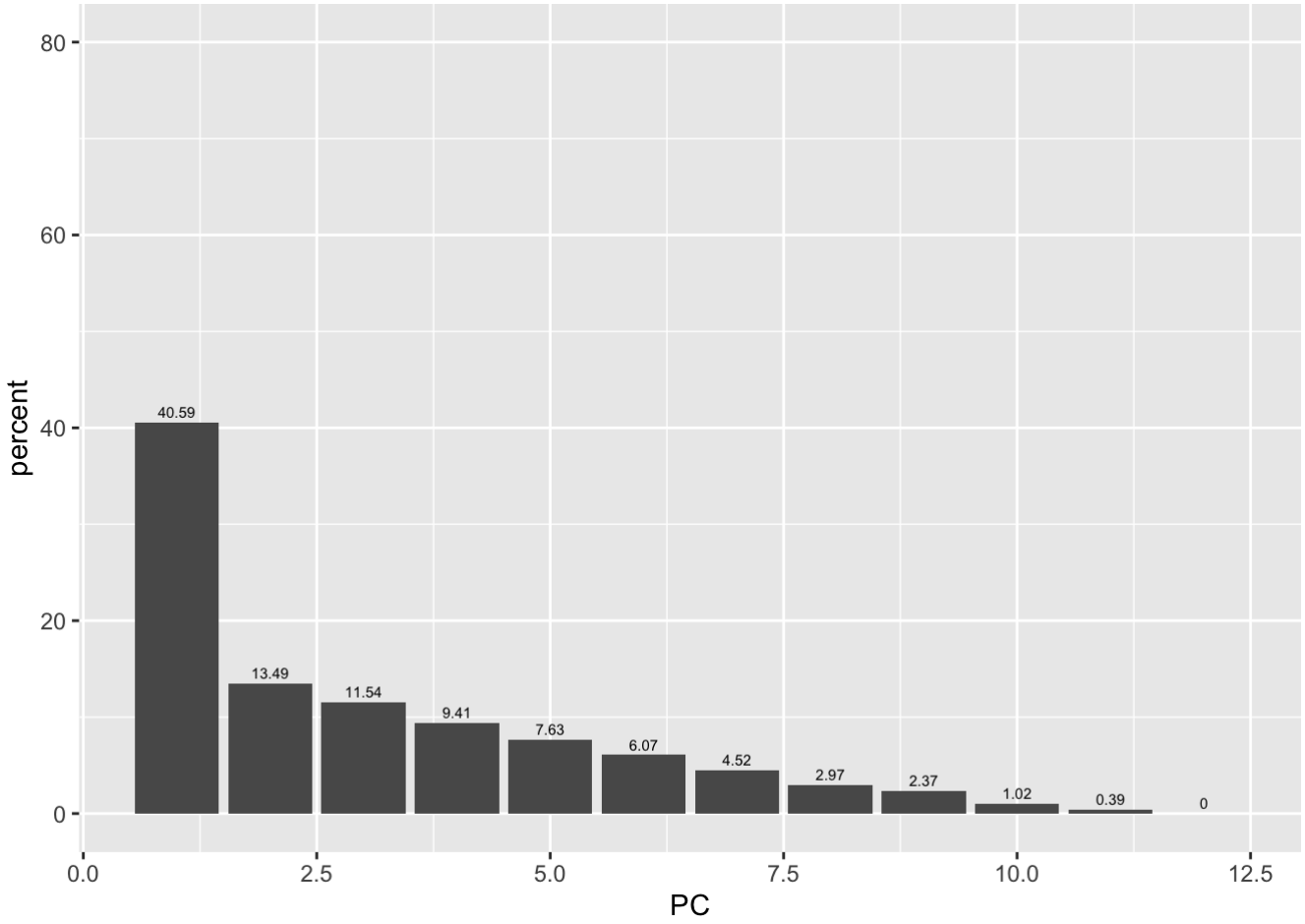
```
# Save the rotation matrix in a data frame
rotation_data <- data.frame(
pca$rotation,
variable = row.names(pca$rotation))

# Define an arrow style
arrow_style <- arrow(length = unit(0.05, "inches"), type = "closed")

# Plot the contribution of variables to PCs using geom_segment() for arrows and geom_tex
t() for labels
ggplot(rotation_data)  +
geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style) + geom_text
(aes(x = PC1, y = PC2, label = variable), hjust = 0, size = 3, color = "blue"
) +
xlim(-1., 1.25) +
ylim(-1., 1.) + coord_fixed()
```




```
# Determine the percentage of variance explained by each component with sdev
percent <- 100* (pca$sdev^2 / sum(pca$sdev^2))
percent
```

```
##  [1] 40.58761145511 13.48574486189 11.54328257193  9.41399797541  7.62954118269
##  [6]  6.06881268557  4.52321316368  2.96609646930  2.37333122964  1.01814081004
## [11]  0.39018821891  0.00003937583
```

```
# Visualize the percentage of variance explained by each component
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
ggplot(perc_data, aes(x = PC, y = percent)) + geom_col() +
geom_text(aes(label = round(percent, 2)), size = 2, vjust = -0.5) + ylim(0, 80)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.