Stat535_HW5

Jie Wang

October 29, 2017

1.

```
rm(list=ls())
library(RSQLite)
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
```

```
library(DBI)
```

2.

a.

```
con <- dbConnect(RSQLite::SQLite(), dbname='baseball.db')
dbListTables(con)
```

```
##  [1] "AllstarFull"          "Appearances"          "AwardsManagers"
##  [4] "AwardsPlayers"        "AwardsShareManagers" "AwardsSharePlayers"
##  [7] "Batting"              "BattingPost"          "Fielding"
## [10] "FieldingOF"           "FieldingPost"         "HallOfFame"
## [13] "Managers"             "ManagersHalf"         "Master"
## [16] "Pitching"             "PitchingPost"         "Salaries"
## [19] "Schools"              "SchoolsPlayers"       "SeriesPost"
## [22] "Teams"                "TeamsFranchises"      "TeamsHalf"
## [25] "sqlite_sequence"      "xref_stats"
```

b.

```
payroll <- dbReadTable(con, "Salaries")
payroll %>%
  dplyr::filter(yearID == 2010) %>%
  group_by(teamID) %>%
  summarise(payroll = sum(salary)) %>%
  arrange(desc(payroll)) %>%
  head(n = 5)
```

```
## # A tibble: 5 x 2
##   teamID   payroll
##    <chr>     <dbl>
## 1    NYA 206333389
## 2    BOS 162447333
## 3    CHN 146609000
## 4    PHI 141928379
## 5    NYN 134422942
```

```
#payroll2010 <- subset(payroll, yearID == 2010)
#payroll2010S <- cbind(aggregate(salary ~ teamID, payroll2010, sum))
#payroll2010S$teamID[payroll2010S$salary == max(payroll2010S$salary)]
```

c.

```
dbGetQuery(con, "
                 SELECT teamID, sum(salary) AS payroll
                 FROM Salaries
                 WHERE yearID == 2010
                 GROUP BY teamID
                 ORDER BY payroll DESC
                 LIMIT 5;
                 ")
```

```
##   teamID   payroll
## 1    NYA 206333389
## 2    BOS 162447333
## 3    CHN 146609000
## 4    PHI 141928379
## 5    NYN 134422942
```

d.

```
team_payroll <- dbGetQuery(con, "
                 SELECT yearID, teamID, sum(salary) AS payroll
                 FROM Salaries
                 WHERE yearID BETWEEN 1985 and 2010
                 GROUP BY teamID, yearID
                 ORDER BY payroll DESC;")
head(team_payroll, n = 10)
```

```
##    yearID teamID   payroll
## 1    2005    NYA 208306817
## 2    2008    NYA 207896789
## 3    2010    NYA 206333389
## 4    2009    NYA 201449189
## 5    2006    NYA 194663079
## 6    2007    NYA 189259045
## 7    2004    NYA 184193950
## 8    2010    BOS 162447333
## 9    2003    NYA 152749814
## 10   2009    NYN 149373987
```

3.

a.

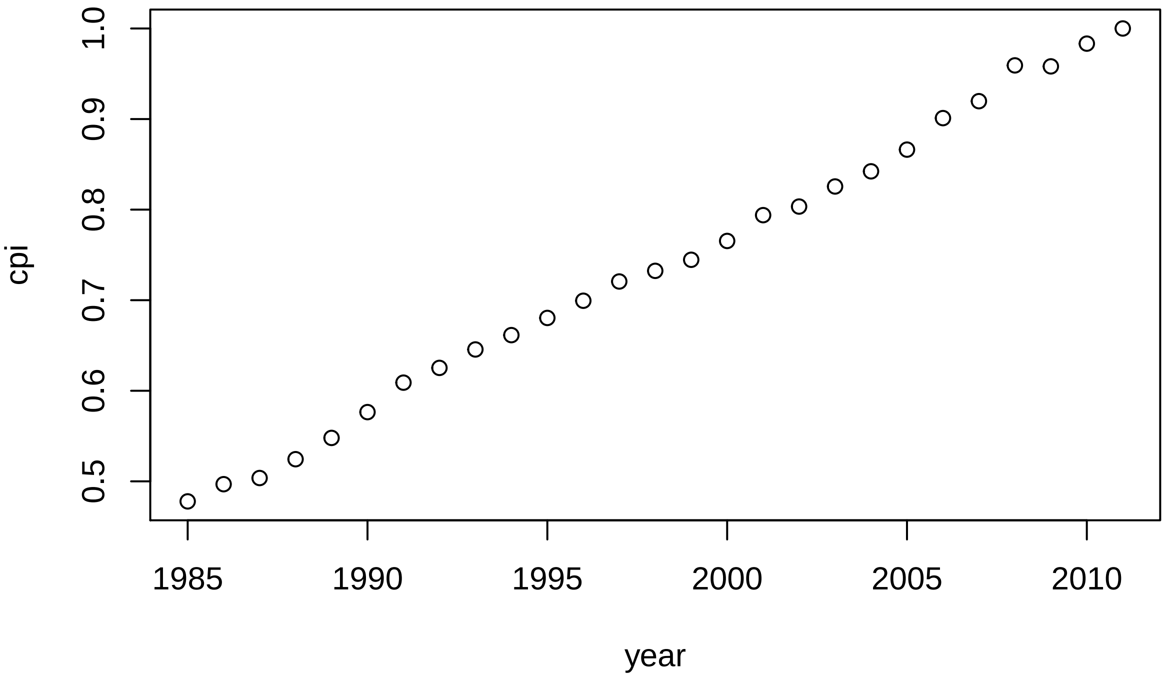
```
library(fImport)
```

```
## Loading required package: timeDate

## Loading required package: timeSeries
```

```
library(timeDate)
library(timeSeries)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
cpi <- read.table("CPIAUCSL.txt", head=T, skip=54)
cpi$DATE <- ymd(cpi$DATE)
cpi <- subset(cpi,months(DATE)=="January" & DATE <= ymd("2011-01-01") & DATE >= ymd("1985-01-01"))
cpi <- cpi[,2]
cpi <- cpi/cpi[length(cpi)]
year <- 1985:2011
plot(year,cpi)
```



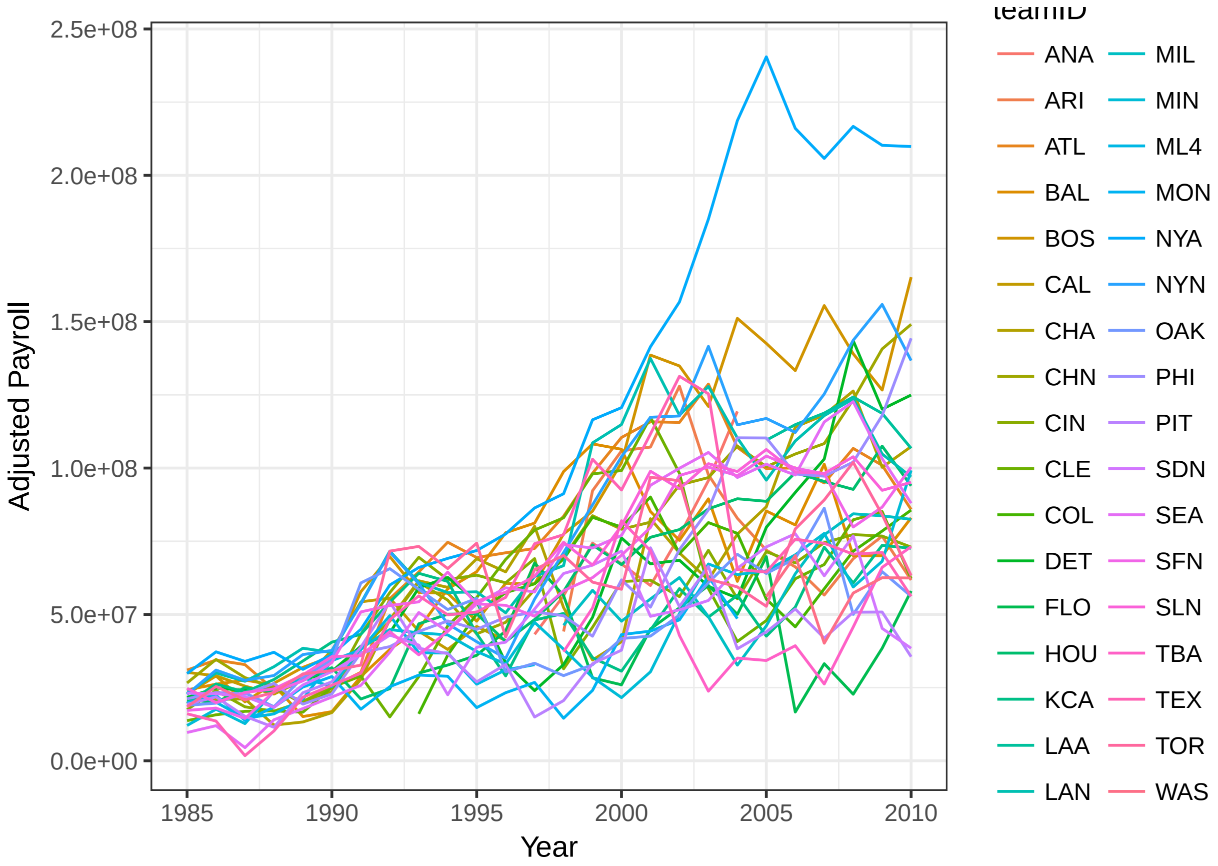
b. Calculate the inflation-adjusted payroll of each baseball team over time. (Hint: You may find plyr helpful here.)

```
#convert_to_2011 <- function(amt, yr) {
#  y <- amt / cpi[yr - 1985 + 1]
#}
#convert_to_2011(payroll, yearID))
team_payroll <- team_payroll %>%
  mutate(adj_payroll = payroll/cpi[yearID-1985+1])
head(team_payroll, n=10)
```

```
##    yearID teamID   payroll adj_payroll
## 1    2005    NYA 208306817   240473695
## 2    2008    NYA 207896789   216728096
## 3    2010    NYA 206333389   209842673
## 4    2009    NYA 201449189   210245416
## 5    2006    NYA 194663079   216040855
## 6    2007    NYA 189259045   205772010
## 7    2004    NYA 184193950   218686566
## 8    2010    BOS 162447333   165210210
## 9    2003    NYA 152749814   185028878
## 10   2009    NYN 149373987   155896364
```

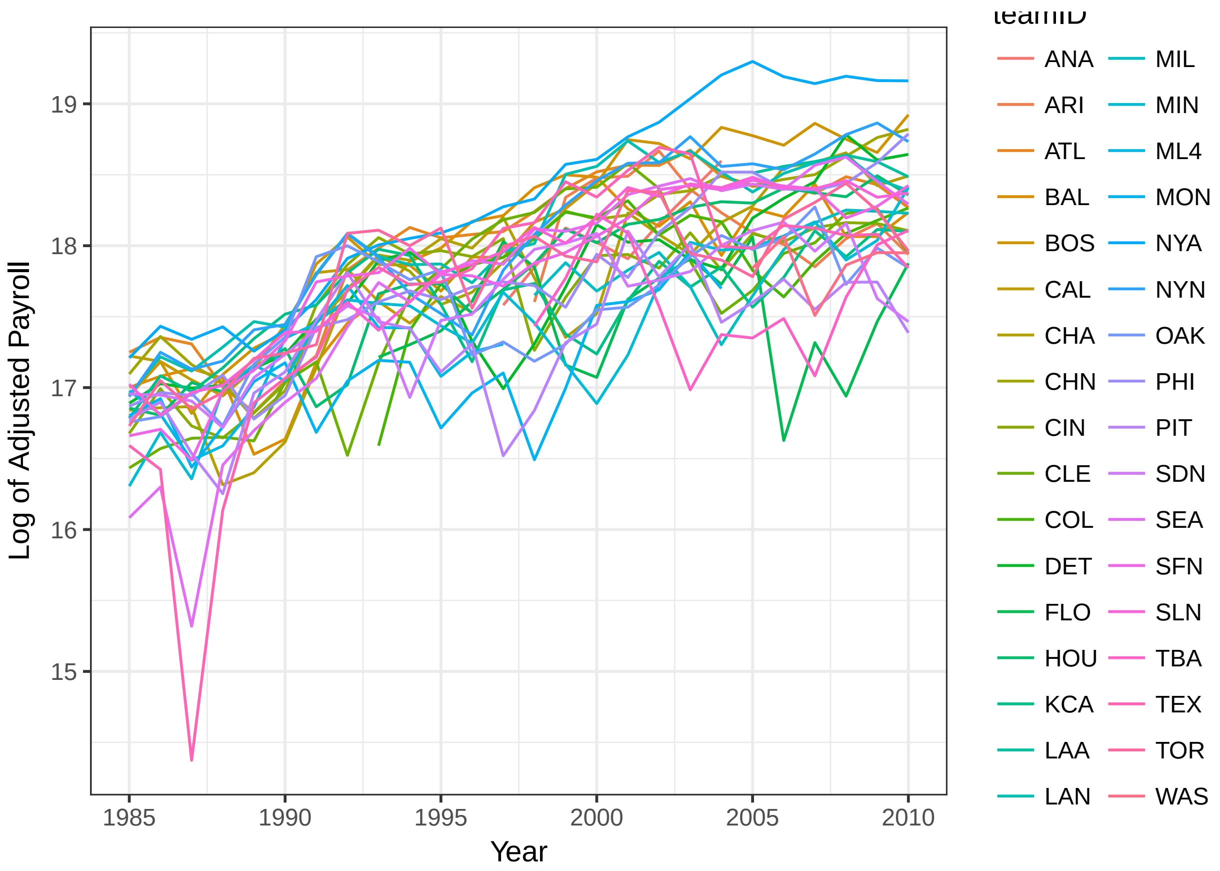
c.

```
library(ggplot2)
ggplot(team_payroll,
       aes(x = yearID,
           y = adj_payroll,
           color = teamID)) +
  geom_line() +
  theme_bw() +
  labs(x = "Year",
       y = "Adjusted Payroll")
```



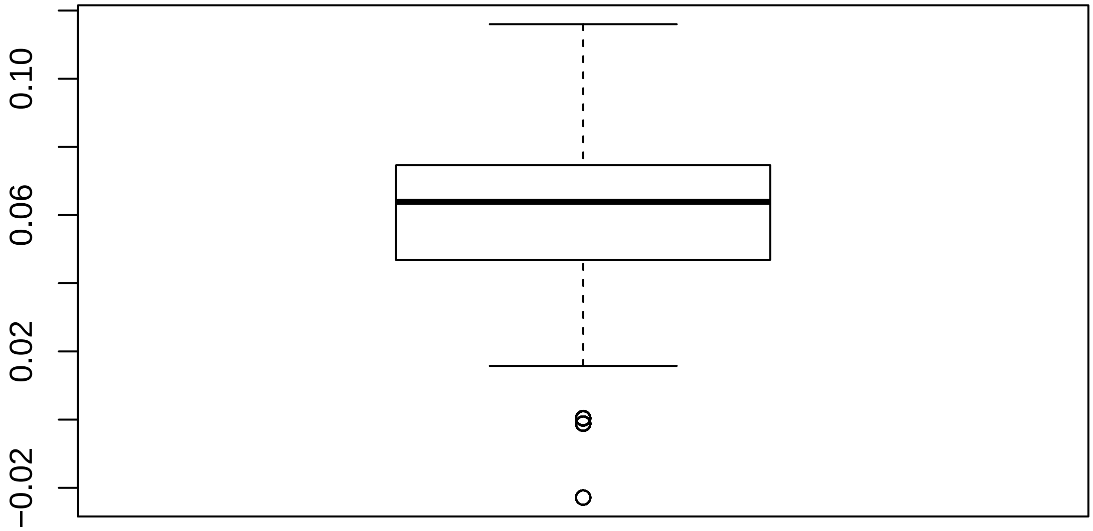
d.

```
library(ggplot2)
ggplot(team_payroll,
       aes(x = yearID,
           y = log(adj_payroll),
           color = teamID)) +
  geom_line() +
  theme_bw() +
  labs(x = "Year",
       y = "Log of Adjusted Payroll")
```



e.

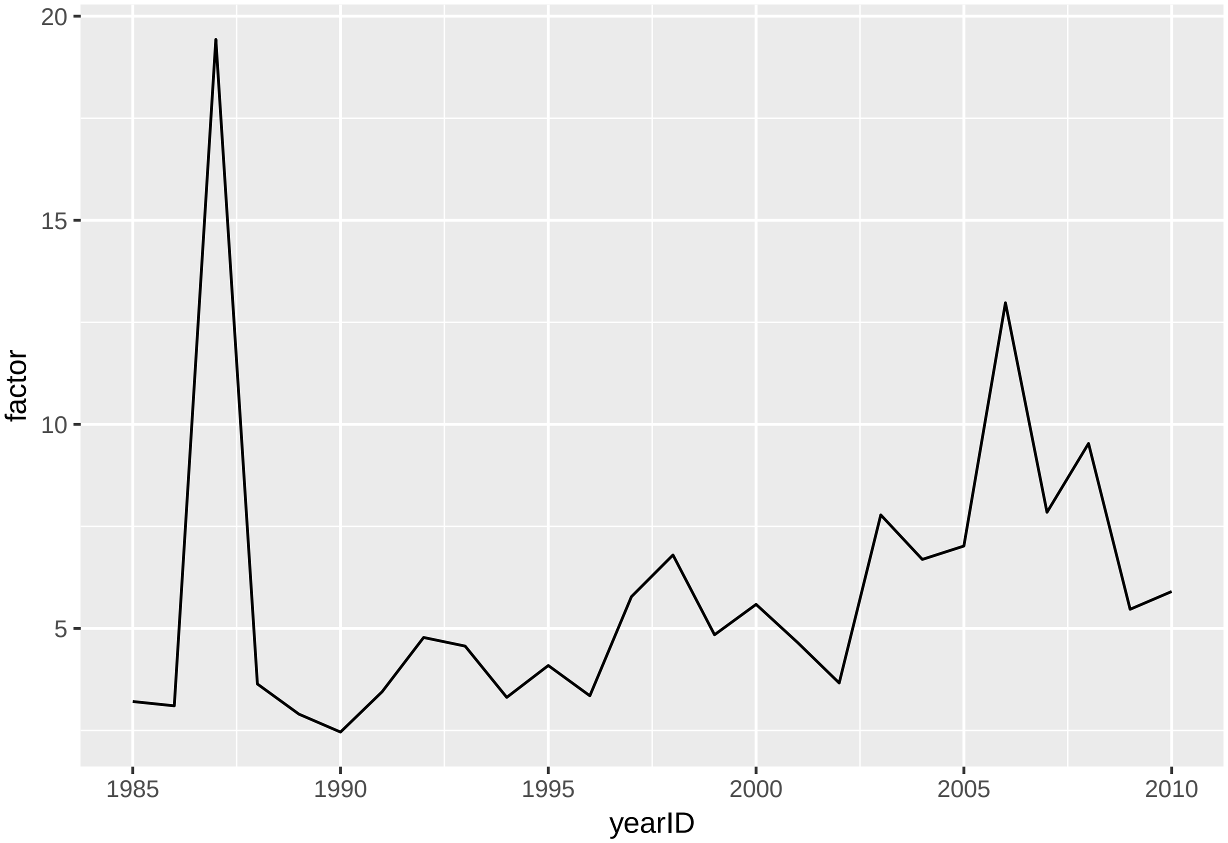
```
slopes <- team_payroll %>% group_by(teamID) %>%   mutate(slopes=coef(lm(log(adj_payroll)~yearID))[2])

boxplot(slopes$slopes)
```



```
team_payroll %>% group_by(yearID)  %>%  top_n(n = 5, wt = adj_payroll) %>% arrange(desc(yearID)) %>% gr
```

```
## # A tibble: 22 x 2
## # Groups:   teamID [22]
##    teamID     n
##     <chr> <int>
##  1    NYA    24
##  2    NYN    16
##  3    BOS    14
##  4    ATL    11
##  5    LAN    11
##  6    BAL     7
##  7    CHA     6
##  8    CHN     5
##  9    CLE     4
## 10    KCA     4
## # ... with 12 more rows
```

```
team_payroll %>% group_by(yearID) %>% summarise(max=max(adj_payroll), min=min(adj_payroll)) %>%
  mutate(factor=max/min) %>% ggplot(aes(x=yearID, y=factor)) +
  geom_line()
```

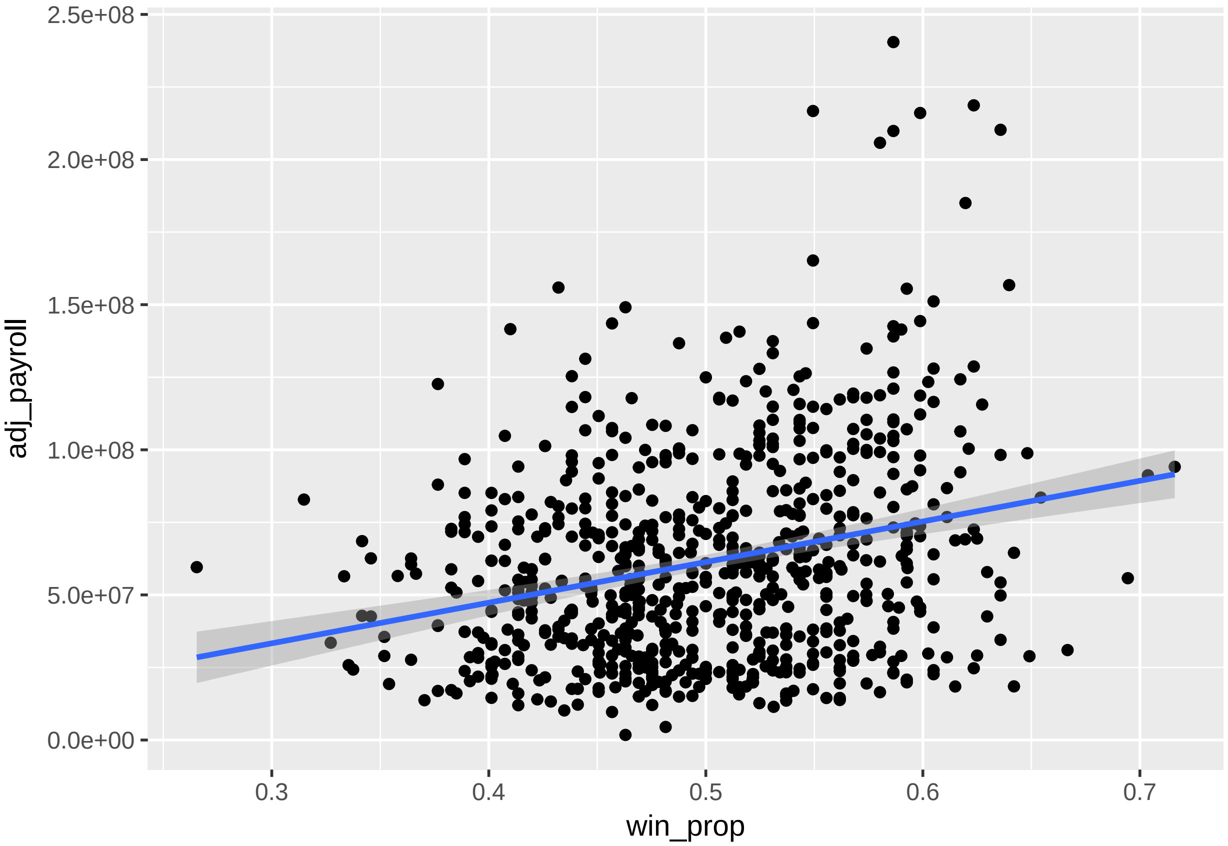


From the result, we can see that in general, the payrolls fell behind the inflation. Also we can see from the pictures in Q3, there is no team whose payrolls have consistently been higher than the others. The gap between the highest and the lowest payrolls has grown.

4.

```
teams <- dbGetQuery(con, "select yearID, teamID, W, G from Teams")
team_payroll %>% inner_join(teams) %>% mutate(win_prop=W/G)%>%
  ggplot(aes(x=win_prop, y=adj_payroll))+geom_point()+geom_smooth(method="lm")
```

```
## Joining, by = c("yearID", "teamID")
```



5.

```
salary_class_df <- payroll %>% dplyr::group_by(playerID) %>%
  summarise(mean.salary = mean(salary, na.rm=T)) %>%
  mutate(salary_class = cut(mean.salary, breaks=c(0, 50000, 100000, 150000, 200000, max(mean.salary)),
head(salary_class_df, n = 15)
```

```
## # A tibble: 15 x 3
##     playerID mean.salary salary_class
##        <chr>       <dbl>       <fctr>
##  1 aardsda01    851950.0            5
##  2  aasedo01    575000.0            5
##  3  abadan01    327000.0            5
##  4 abbotje01    246250.0            5
##  5 abbotji01   1440055.6            5
##  6 abbotku01    470777.8            5
##  7 abbotky01    129500.0            3
##  8 abbotpa01    924428.6            5
##  9 abercre01    327000.0            5
## 10 abernbr01    257500.0            5
## 11 abnersh01    144700.0            3
## 12 abreubo01   7598547.6            5
## 13 abreuto01    400000.0            5
## 14 accarje01    548600.0            5
## 15 aceveal01    435650.0            5
```

```r
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Librar
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found
```

```
## Could not load tcltk.  Will use slower R code instead.
```

```r
salary_class_df1 <- dbGetQuery(con,"select playerID,avg(salary) as mean_salary from Salaries group by pl
salary_class_df2 <- sqldf::sqldf("select playerID,mean_salary,min(5,floor(mean_salary/5000)) as salary_
head(salary_class_df2, n = 15)
```

```
##     playerID mean_salary salary_class
## 1  aardsda01    851950.0            5
## 2   aasedo01    575000.0            5
## 3   abadan01    327000.0            5
## 4  abbotje01    246250.0            5
## 5  abbotji01   1440055.6            5
## 6  abbotku01    470777.8            5
## 7  abbotky01    129500.0            5
## 8  abbotpa01    924428.6            5
## 9  abercre01    327000.0            5
## 10 abernbr01    257500.0            5
## 11 abnersh01    144700.0            5
## 12 abreubo01   7598547.6            5
## 13 abreuto01    400000.0            5
## 14 accarje01    548600.0            5
## 15 aceveal01    435650.0            5
```