DaigleInClassLab_Wk5D1RevB.R

2011home

Wed Feb 14 09:50:50 2018

```
# 1 which league is the best in terms of wage (eur_wage) and overall?

getwd()
```

```
## [1] "/Users/2011home/Library/Mobile Documents/com~apple~CloudDocs/Education/UConn/Spring 2018/R/Week5
```

```
setwd("/Users/2011home/Library/Mobile Documents/com~apple~CloudDocs/Education/UConn/Spring 2018/R/Week5/
dir()
```

```
## [1] "complete.csv"                  "district_rev_exp_readtable.txt"
## [3] "district_rev_exp.csv"          "district_rev_exp.txt"
## [5] "district_rev_exp.xlsx"         "school15doc.pdf"
```

```
sc <- read.csv("complete.csv", stringsAsFactors = FALSE, na.strings = "")
# na.strings = "" means that any empty strings become "NA", we can change that from "" to " " or "." or
sc <- sc[, c("name", "club", "age", "league", "eur_value", "eur_wage", "overall")]
head(sc)
```

```
##                name               club age                    league
## 1 Cristiano Ronaldo      Real Madrid CF  32 Spanish Primera División
## 2          L. Messi        FC Barcelona  30 Spanish Primera División
## 3            Neymar Paris Saint-Germain  25           French Ligue 1
## 4         L. Suárez        FC Barcelona  30 Spanish Primera División
## 5          M. Neuer    FC Bayern Munich  31        German Bundesliga
## 6    R. Lewandowski    FC Bayern Munich  28        German Bundesliga
##   eur_value eur_wage overall
## 1  9.55e+07   565000      94
## 2  1.05e+08   565000      93
## 3  1.23e+08   280000      92
## 4  9.70e+07   510000      92
## 5  6.10e+07   230000      92
## 6  9.20e+07   355000      91
```

```
str(sc)
```

```
## 'data.frame':    17994 obs. of  7 variables:
##  $ name     : chr  "Cristiano Ronaldo" "L. Messi" "Neymar" "L. Suárez" ...
##  $ club     : chr  "Real Madrid CF" "FC Barcelona" "Paris Saint-Germain" "FC Barcelona" ...
##  $ age      : int  32 30 25 30 31 28 26 26 27 29 ...
##  $ league   : chr  "Spanish Primera División" "Spanish Primera División" "French Ligue 1" "Spanish Pr
##  $ eur_value: num  9.55e+07 1.05e+08 1.23e+08 9.70e+07 6.10e+07 9.20e+07 6.45e+07 9.05e+07 7.90e+07 
##  $ eur_wage : num  565000 565000 280000 510000 230000 355000 215000 295000 340000 275000 ...
##  $ overall  : int  94 93 92 92 92 91 90 90 90 90 ...
```

```
# str tells us about what each of these columns are in the data.frame
summary(sc)
```

```
##      name               club                age           league
##  Length:17994       Length:17994       Min.   :16.00   Length:17994
##  Class :character   Class :character   1st Qu.:21.00   Class :character
```

```
##  Mode  :character   Mode  :character   Median :25.00   Mode  :character
##                                        Mean   :25.12
##                                        3rd Qu.:28.00
##                                        Max.   :47.00
##    eur_value            eur_wage         overall
##  Min.   :        0   Min.   :     0   Min.   :46.00
##  1st Qu.:   300000   1st Qu.:  2000   1st Qu.:62.00
##  Median :   700000   Median :  4000   Median :66.00
##  Mean   :  2370511   Mean   : 11504   Mean   :66.25
##  3rd Qu.:  2000000   3rd Qu.: 12000   3rd Qu.:71.00
##  Max.   :123000000   Max.   :565000   Max.   :94.00
```

```r
lWage <- aggregate(sc[, "eur_wage"], list(sc$league), mean)
# this gives us the mean of the eur_wage by each league (the list)
names(lWage) <- c("league", "avgWage")
head(lWage)
```

```
##                          league   avgWage
## 1         Argentinian Superliga  6837.179
## 2           Australian A-League  3046.610
## 3           Austrian Bundesliga  6046.332
## 4      Belgian First Division A  8912.844
## 5 Campeonato Brasileiro Série A 15515.625
## 6      Chilian Primera División  4800.000
```

```r
tail(lWage)
```

```
##                       league   avgWage
## 36 Spanish Segunda División  5846.154
## 37      Swedish Allsvenskan  2213.368
## 38       Swiss Super League  6897.338
## 39        Turkish Süper Lig 16563.745
## 40 Ukrainian Premier League  1000.000
## 41  USA Major League Soccer  4086.400
```

```r
rankWage <- order(lWage[, "avgWage"], decreasing = TRUE)
lWage <- lWage[rankWage, ]

lOverall <- aggregate(sc[,"overall"], list(sc$league), mean)
names(lOverall) <- c("league", "avgOverall")

rankOverall <- order(lOverall[, "avgOverall"], decreasing = TRUE)
lOverall <- lOverall[rankOverall, ]

# 2

rank_value <- order(sc[,"eur_value"], decreasing = TRUE)
lValue <- sc[rank_value, ]
lValue100 <- lValue[1:100,]
lValue100$rank_value <- 1:100

str(lValue100)
```

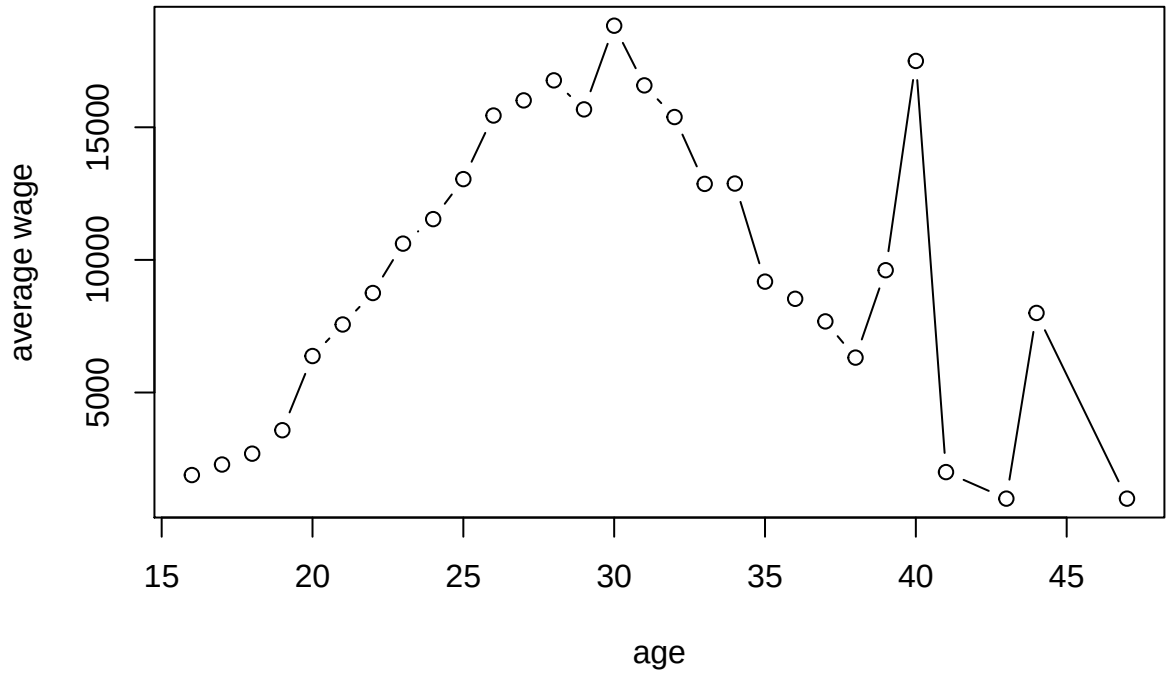
```
## 'data.frame':    100 obs. of  8 variables:
##  $ name      : chr  "Neymar" "L. Messi" "L. Suárez" "Cristiano Ronaldo" ...
##  $ club      : chr  "Paris Saint-Germain" "FC Barcelona" "FC Barcelona" "Real Madrid CF" ...
```

```
##  $ age       : int  25 30 30 32 28 26 26 27 23 29 ...
##  $ league    : chr  "French Ligue 1" "Spanish Primera División" "Spanish Primera División" "Spanish
##  $ eur_value : num  1.23e+08 1.05e+08 9.70e+07 9.55e+07 9.20e+07 9.05e+07 8.30e+07 7.90e+07 7.90e+07
##  $ eur_wage  : num  280000 565000 510000 565000 355000 295000 285000 340000 215000 275000 ...
##  $ overall   : int  92 93 92 94 91 90 89 90 88 90 ...
##  $ rank_value: int  1 2 3 4 5 6 7 8 9 10 ...
```

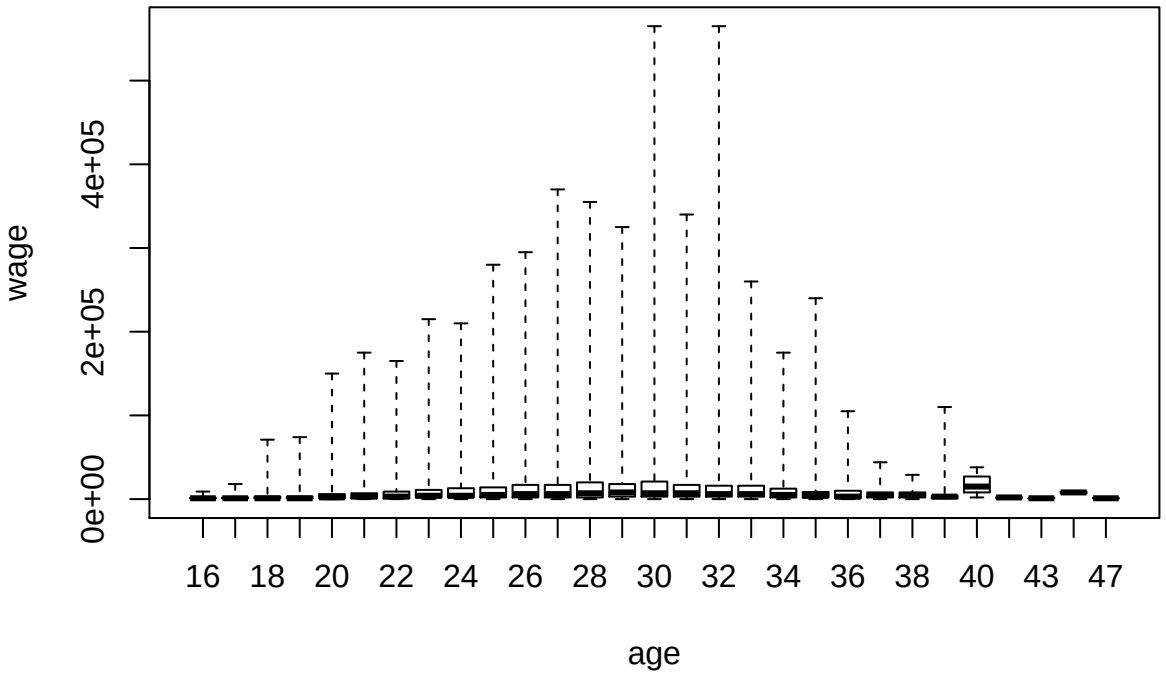
```r
teamStat <- summary(factor(lValue100$club))
teamStat <- sort(teamStat, decreasing = TRUE)
teamStat
```

```
##          Real Madrid CF        FC Bayern Munich                Chelsea
##                      12                     11                      8
##           FC Barcelona       Manchester United     Paris Saint-Germain
##                       7                      7                      7
##               Juventus         Manchester City         Atlético Madrid
##                       6                      6                      5
##      Tottenham Hotspur               Liverpool                 Napoli
##                       5                      4                      4
##                Arsenal       Borussia Dortmund                AS Monaco
##                       3                      3                      2
##                  Inter                   Roma Athletic Club de Bilbao
##                       2                      2                      1
##    Bayer 04 Leverkusen           FC Schalke 04                  Milan
##                       1                      1                      1
##             RB Leipzig                 Torino
##                       1                      1
```

```r
# 3 Present the distribution of avg wage based on age
ageWage <- aggregate(sc[, "eur_wage"], list(sc$age), mean)
names(ageWage) <- c("age", "avgWage")
plot(ageWage$age, ageWage$avgWage, xlab = "age", ylab = "average wage", type = "b")
```



```r
boxplot(eur_wage~age, data = sc, xlab = "age", ylab = "wage", range = 0)
```




```r
# 4 Which team has the most players under 23?

team23 <- factor(sc[sc$age<23, "club"])
teamAge <- table(team23)
teamAgeRank <- sort(teamAge, decreasing = TRUE)
teamAgeRank[1:10]
```

```
## team23
##   Sevilla Atlético     FC Barcelona B         LOSC Lille
##                 28                 22                 22
##   Werder Bremen II Olympique Lyonnais           Banfield
##                 22                 20                 19
##        Envigado FC    FC Nordsjælland               Ajax
##                 19                 19                 18
##          AS Monaco
##                 18
```