Bios 6301: Assignment 6

Yeji Ko

Due Tuesday, 26 October, 1:00 PM

$5^{n=day}$ points taken off for each day late.

40 points total.

Submit a single knitr file (named `homework6.rmd`), along with a valid PDF output file. Inside the file, clearly indicate which parts of your responses go with which problems (you may use the original homework document as a template). Add your name as `author` to the file's metadata section. Raw R code/output or word processor files are not acceptable.

Failure to name file `homework6.rmd` or include author name may result in 5 points taken off.

Question 1

16 points

Obtain a copy of the football-values lecture. Save the five 2021 CSV files in your working directory.

Modify the code to create a function. This function will create dollar values given information (as arguments) about a league setup. It will return a data.frame and write this data.frame to a CSV file. The final data.frame should contain the columns 'PlayerName', 'pos', 'points', 'value' and be orderd by value descendingly. Do not round dollar values.

Note that the returned data.frame should have `sum(posReq)*nTeams` rows.

Define the function as such (10 points):

```
# path: directory path to input files
# file: name of the output file; it should be written to path
# nTeams: number of teams in league
# cap: money available to each team
# posReq: number of starters for each position
# points: point allocation for each category

year <- 2021

ffvalues <- function(path, file='outfile.csv', nTeams=12, cap=200,
                     posReq=c(qb=1, rb=2, wr=3, te=1, k=1),
                     points=c(fg=4, xpt=1, pass_yds=1/25, pass_tds=4, pass_ints=-2,
                     rush_yds=1/10, rush_tds=6, fumbles=-2, rec_yds=1/20, rec_tds=6)){

   ## read in CSV files
  positions <- c('k','qb','rb','te','wr')
  csvfile <- paste('proj_', positions, substr(year, 3, 4), '.csv', sep='')
  files <- file.path(path, csvfile)
  names(files) <- positions
  k <- read.csv(files['k'], header=TRUE, stringsAsFactors=FALSE)
  qb <- read.csv(files['qb'], stringsAsFactors=FALSE)
```

```
rb <- read.csv(files['rb'])
te <- read.csv(files['te'])
wr <- read.csv(files['wr'])

# generate unique list of column names
cols <- unique(c(names(k), names(qb), names(rb), names(te), names(wr)))

# create a new column in each data.frame
k[,'pos'] <- 'k'
qb[,'pos'] <- 'qb'
rb[,'pos'] <- 'rb'
te[,'pos'] <- 'te'
wr[,'pos'] <- 'wr'

# append 'pos' to unique column list
cols <- c(cols, 'pos')

# create common columns in each data.frame
# initialize values to zero
k[,setdiff(cols, names(k))] <- 0
qb[,setdiff(cols, names(qb))] <- 0
rb[,setdiff(cols, names(rb))] <- 0
te[,setdiff(cols, names(te))] <- 0
wr[,setdiff(cols, names(wr))] <- 0

# combine data.frames by row, using consistent column order
x <- rbind(k[,cols], qb[,cols], rb[,cols], te[,cols], wr[,cols])

## calculate dollar values
x[,'p_fg'] <- x[,'fg']*points['fg']
x[,'p_xpt'] <- x[,'xpt']*points['xpt']
x[,'p_pass_yds'] <- x[,'pass_yds']*points['pass_yds']
x[,'p_pass_tds'] <- x[,'pass_tds']*points['pass_tds']
x[,'p_pass_ints'] <- x[,'pass_ints']*points['pass_ints']
x[,'p_rush_yds'] <- x[,'rush_yds']*points['rush_yds']
x[,'p_rush_tds'] <- x[,'rush_tds']*points['rush_tds']
x[,'p_fumbles'] <- x[,'fumbles']*points['fumbles']
x[,'p_rec_yds'] <- x[,'rec_yds']*points['rec_yds']
x[,'p_rec_tds'] <- x[,'rec_tds']*points['rec_tds']
x[,'points'] <- rowSums(x[,grep("^p_", names(x))])

# create new data.frame ordered by points descendingly
x2 <- x[order(x[,'points'], decreasing=TRUE),]

# determine the row indices for each position
k.ix <- which(x2[,'pos']=='k')
qb.ix <- which(x2[,'pos']=='qb')
rb.ix <- which(x2[,'pos']=='rb')
te.ix <- which(x2[,'pos']=='te')
wr.ix <- which(x2[,'pos']=='wr')

# calculate marginal points by subtracting "baseline" player's points
ifelse((x2[k.ix,'points'] - x2[k.ix[nTeams*posReq['k']],'points'] >= 0),
```

```
        x2[k.ix, 'marg'] <- (x2[k.ix,'points'] - x2[k.ix[nTeams*posReq['k']],'points']),
        x2[k.ix, 'marg'] <- 0)
  x2[qb.ix, 'marg'] <- x2[qb.ix,'points'] - x2[qb.ix[nTeams*posReq['qb']],'points']
  x2[rb.ix, 'marg'] <- x2[rb.ix,'points'] - x2[rb.ix[nTeams*posReq['rb']],'points']
  x2[te.ix, 'marg'] <- x2[te.ix,'points'] - x2[te.ix[nTeams*posReq['te']],'points']
  x2[wr.ix, 'marg'] <- x2[wr.ix,'points'] - x2[wr.ix[nTeams*posReq['wr']],'points']

  # create a new data.frame subset by non-negative marginal points
  x2[is.na(x2[,'marg']),'marg'] <- 0
  x3 <- x2[x2[,'marg'] >= 0,]

  # re-order by marginal points
  x3 <- x3[order(x3[,'marg'], decreasing=TRUE),]

  # reset the row names
  rownames(x3) <- NULL

  # calculation for player value
  x3[,'value'] <- (nTeams*cap-nrow(x3)) * x3[,'marg'] / sum(x3[,'marg']) + 1

  # create a data.frame with more interesting columns
  x4 <- x3[,c('PlayerName','pos','points','value')]
  return(x4)

  ## save dollar values as CSV file
  write.csv(x4, file)

}

# 1

x1 <- ffvalues('.')
sum(x1[,'value'] >20)
x1[which(x1[,'pos']=='rb'),][15,]

# 2

x2 <- ffvalues(getwd(), '16team.csv', nTeams=16, cap=150)
sum(x2[,'value'] >20)
sum(x2[1:40,'pos']=='wr')

# 3
x3 <- ffvalues('.', 'qbheavy.csv', posReq=c(qb=2, rb=2, wr=3, te=1, k=0),
            points=c(fg=0, xpt=0, pass_yds=1/25, pass_tds=6, pass_ints=-2,
                     rush_yds=1/10, rush_tds=6, fumbles=-2, rec_yds=1/20, rec_tds=6))
sum(x3[,'value'] >20)
sum(x3[1:30,'pos']=='qb')
```

-2 points, I can't full credit these when I can't see the output in the pdf file.

1. Call:`x1 <- ffvalues('.')`
    1. How many players are worth more than $20? (1 point)

    44 players are worth more than $20.

    1. Who is 15th most valuable running back (rb)? (1 point)

Chris Carson is the 15th most valuable running back.

2. Call:`x2 <- ffvalues(getwd(), '16team.csv', nTeams=16, cap=150)`

    1. How many players are worth more than $20? (1 point)

    There are 44 players who are worth more than $20

    1. How many wide receivers (wr) are in the top 40? (1 point)

    There are 8 wide receivers.

3. Call:`x3 <- ffvalues('.', 'qbheavy.csv', posReq=c(qb=2, rb=2, wr=3, te=1, k=0), points=c(fg=0, xpt=0, pass_yds=1/25, pass_tds=6, pass_ints=-2, rush_yds=1/10, rush_tds=6, fumbles=-2, rec_yds=1/20, rec_tds=6))`

    1. How many players are worth more than $20? (1 point)

    46 players are worth more than $20

    1. How many quarterbacks (qb) are in the top 30? (1 point)

    There are 14 quaterbacks in the top 30.

Question 2

24 points

Import the HAART dataset (`haart.csv`) from the GitHub repository into R, and perform the following manipulations: (4 points each)

1. Convert date columns into a usable (for analysis) format. Use the `table` command to display the counts of the year from `init.date`.

```
library(lubridate)
```

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

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

```
haart <- read.csv("/Users/yejiko/Downloads/Bios6301-main/datasets/haart.csv")
haart$init.date <- mdy(haart$init.date)
haart$last.visit <- mdy(haart$last.visit)
haart$date.death <- mdy(haart$date.death)
table(year(haart$init.date))
```

```
##
## 1998 2000 2001 2002 2003 2004 2005 2006 2007
##    1    5   17   60  270  292  207  104   44
```

2. Create an indicator variable (one which takes the values 0 or 1 only) to represent death within 1 year of the initial visit. How many observations died in year 1?

```
haart$death.1year <- ifelse((haart$date.death - haart$init.date <= 365),1,0)
sum(haart$death.1year, na.rm = TRUE) # 92 observations died within a year of the initial visit.
```

```
## [1] 92
```

3. Use the `init.date`, `last.visit` and `death.date` columns to calculate a followup time (in days), which is the difference between the first and either the last visit or a death event (whichever comes first). If

these times are longer than 1 year, censor them (this means if the value is above 365, set followup to 365). Print the quantile for this new variable.

```
# if last.visit comes first or death.date is null, then plug in last.visit date to last event
haart$last.event <- ifelse((haart$last.visit < haart$date.death) | (is.na(haart$date.death)),
                           as.character(haart$last.visit), as.character(haart$date.death))

# if last.visit is null, always plug in death.date as the last event date
haart[which(is.na(haart$last.event)),"last.event"] <- as.character(haart[which(is.na(haart$last.event))

sum(is.na(haart$last.event))
```

```
## [1] 0
```

```
haart$last.event <- as.Date(haart$last.event)

haart$followup <- haart$last.event - haart$init.date
haart$followup <- ifelse(haart$followup > 365, 365, haart$followup)

summary(haart$followup)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   320.8   365.0   298.4   365.0   365.0
```

4. Create another indicator variable representing loss to followup; this means the observation is not known to be dead but does not have any followup visits after the first year. How many records are lost-to-followup?

```
haart$followup.loss <- ifelse((haart$last.event - haart$init.date > 365) & (haart$death==0),1,0)
table(haart$followup.loss) # 710 records are lost to followup
```

less than?

3/4

```
##
##   0   1
## 290 710
```

5. Recall our work in class, which separated the `init.reg` field into a set of indicator variables, one for each unique drug. Create these fields and append them to the database as new columns. Which drug regimen are found over 100 times?

2/4 credit

```
class(haart$init.reg)
```

```
## [1] "character"
```

```
init.reg <- as.character(haart[,'init.reg'])
(haart[['init.reg_list']] <- strsplit(init.reg, ","))[1:3]
```

```
## [[1]]
## [1] "3TC" "AZT" "EFV"
##
## [[2]]
## [1] "3TC" "AZT" "EFV"
##
## [[3]]
## [1] "3TC" "AZT" "EFV"
```

```
unlist(haart$init.reg_list)[seq(50)]
```

```
##  [1] "3TC" "AZT" "EFV" "3TC" "AZT" "EFV" "3TC" "AZT" "EFV" "3TC" "AZT" "NVP"
## [13] "3TC" "D4T" "EFV" "3TC" "AZT" "NVP" "3TC" "AZT" "NVP" "3TC" "AZT" "EFV"
```

```
## [25] "3TC" "ABC" "AZT" "3TC" "DDI" "NVP" "3TC" "AZT" "NVP" "3TC" "AZT" "IDV"
## [37] "3TC" "AZT" "NVP" "3TC" "AZT" "EFV" "3TC" "AZT" "EFV" "3TC" "D4T" "NVP"
## [49] "3TC" "AZT"
```

```
(all_drugs <- unique(unlist(haart$init.reg_list))) # 18 unique drugs we found
```

```
##  [1] "3TC" "AZT" "EFV" "NVP" "D4T" "ABC" "DDI" "IDV" "LPV" "RTV" "SQV" "FTC"
## [13] "TDF" "DDC" "NFV" "T20" "ATV" "FPV"
```

```
reg_drugs <- matrix(FALSE, nrow=nrow(haart), ncol=length(all_drugs)) # 18 indicator variables

for(i in seq_along(all_drugs)) {
  reg_drugs[,i] <- sapply(haart$init.reg_list, function(x) all_drugs[i] %in% x)
}
reg_drugs <- data.frame(reg_drugs)
names(reg_drugs) <- all_drugs

haart_merged <- cbind(haart, reg_drugs)
head(haart_merged)
```

```
##   male age aids cd4baseline logvl  weight hemoglobin    init.reg  init.date
## 1    1  25    0          NA    NA      NA         NA 3TC,AZT,EFV 2003-07-01
## 2    1  49    0         143    NA 58.0608         11 3TC,AZT,EFV 2004-11-23
## 3    1  42    1         102    NA 48.0816          1 3TC,AZT,EFV 2003-04-30
## 4    0  33    0         107    NA 46.0000         NA 3TC,AZT,NVP 2006-03-25
## 5    1  27    0          52     4      NA         NA 3TC,D4T,EFV 2004-09-01
## 6    0  34    0         157    NA 54.8856         NA 3TC,AZT,NVP 2003-12-02
##   last.visit death date.death death.1year last.event followup followup.loss
## 1 2007-02-26     0       <NA>          NA 2007-02-26      365             1
## 2 2008-02-22     0       <NA>          NA 2008-02-22      365             1
## 3 2005-11-21     1 2006-01-11           0 2005-11-21      365             0
## 4 2006-05-05     1 2006-05-07           1 2006-05-05       41             0
## 5 2007-11-13     0       <NA>          NA 2007-11-13      365             1
## 6 2008-02-28     0       <NA>          NA 2008-02-28      365             1
##   init.reg_list  3TC   AZT   EFV   NVP   D4T   ABC   DDI   IDV   LPV   RTV
## 1 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 3TC, D4T, EFV TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 6 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##     SQV   FTC   TDF   DDC   NFV   T20   ATV   FPV
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 6 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

6. The dataset `haart2.csv` contains a few additional observations for the same study. Import these and append them to your master dataset (if you were smart about how you coded the previous steps, cleaning the additional observations should be easy!). Show the first five records and the last five records of the complete (and clean) data set.

```
haart2 <- read.csv("/Users/yejiko/Downloads/Bios6301-main/datasets/haart2.csv")
```

```
haart2$init.date <- mdy(haart2$init.date)
haart2$last.visit <- mdy(haart2$last.visit)
haart2$date.death <- mdy(haart2$date.death)
haart2$death.1year <- ifelse((haart2$date.death - haart2$init.date <= 365),1,0)

haart2$last.event <- ifelse((haart2$last.visit < haart2$date.death) | (is.na(haart2$date.death)),
haart2[which(is.na(haart2$last.event)),"last.event"] <-
      as.character(haart2[which(is.na(haart2$last.event)),"date.death"])
haart2$last.event <- as.Date(haart2$last.event)
haart2$followup <- haart2$last.event - haart2$init.date
haart2$followup <- ifelse(haart2$followup > 365, 365, haart2$followup)
haart2$followup.loss <- ifelse((haart2$last.event - haart2$init.date > 365) & (haart2$death==0),1,0)

(haart2[['init.reg_list']] <- strsplit(haart2$init.reg, ","))
```

```
## [[1]]
## [1] "3TC" "AZT" "NVP"
##
## [[2]]
## [1] "3TC" "AZT" "NVP"
##
## [[3]]
## [1] "3TC" "DDI" "EFV"
##
## [[4]]
## [1] "3TC" "D4T" "NVP"
```

```
reg_drugs <- matrix(FALSE, nrow=nrow(haart2), ncol=length(all_drugs))
for(i in seq_along(all_drugs)) {
  reg_drugs[,i] <- sapply(haart2$init.reg_list, function(x) all_drugs[i] %in% x)
}
reg_drugs <- data.frame(reg_drugs)
names(reg_drugs) <- all_drugs
haart_merged2 <- cbind(haart2, reg_drugs)

haart_final <- rbind(haart_merged, haart_merged2)
haart_final[1:5,] # first five records
```

```
##   male age aids cd4baseline logvl  weight hemoglobin    init.reg  init.date
## 1    1  25    0          NA    NA      NA         NA 3TC,AZT,EFV 2003-07-01
## 2    1  49    0         143    NA 58.0608         11 3TC,AZT,EFV 2004-11-23
## 3    1  42    1         102    NA 48.0816          1 3TC,AZT,EFV 2003-04-30
## 4    0  33    0         107    NA 46.0000         NA 3TC,AZT,NVP 2006-03-25
## 5    1  27    0          52     4      NA         NA 3TC,D4T,EFV 2004-09-01
##   last.visit death date.death death.1year last.event followup followup.loss
## 1 2007-02-26     0       <NA>          NA 2007-02-26      365             1
## 2 2008-02-22     0       <NA>          NA 2008-02-22      365             1
## 3 2005-11-21     1 2006-01-11           0 2005-11-21      365             0
## 4 2006-05-05     1 2006-05-07           1 2006-05-05       41             0
## 5 2007-11-13     0       <NA>          NA 2007-11-13      365             1
##   init.reg_list  3TC   AZT   EFV   NVP   D4T   ABC   DDI   IDV   LPV   RTV
## 1 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## 4 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 3TC, D4T, EFV TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##     SQV   FTC   TDF   DDC   NFV   T20   ATV   FPV
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
haart_final[1000:1004,] # last five records
```

```
##      male      age aids cd4baseline    logvl  weight hemoglobin    init.reg
## 1000    0 40.00000    1         131       NA 46.2672          8 3TC,D4T,NVP
## 1001    0 27.00000    0         232       NA      NA         NA 3TC,AZT,NVP
## 1002    1 38.72142    0         170       NA 84.0000         NA 3TC,AZT,NVP
## 1003    1 23.00000   NA         154 3.995635 65.5000         14 3TC,DDI,EFV
## 1004    0 31.00000    0         236       NA 45.8136         NA 3TC,D4T,NVP
##       init.date last.visit death date.death death.1year last.event followup
## 1000 2003-07-03 2008-02-29     0       <NA>          NA 2008-02-29      365
## 1001 2003-12-01 2004-01-05     0       <NA>          NA 2004-01-05       35
## 1002 2002-09-26 2004-03-29     0       <NA>          NA 2004-03-29      365
## 1003 2007-01-31 2007-04-16     0       <NA>          NA 2007-04-16       75
## 1004 2003-12-03 2007-10-11     0       <NA>          NA 2007-10-11      365
##      followup.loss init.reg_list  3TC   AZT   EFV   NVP   D4T   ABC   DDI   IDV
## 1000             1 3TC, D4T, NVP TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## 1001             0 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
## 1002             1 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
## 1003             0 3TC, DDI, EFV TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
## 1004             1 3TC, D4T, NVP TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
##        LPV   RTV   SQV   FTC   TDF   DDC   NFV   T20   ATV   FPV
## 1000 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1001 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1002 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1003 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1004 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```