Bios 6301: Assignment 6

Yiqing Pan

Due Tuesday, 22 October, 1:00 PM

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

40 points total.

Submit a single quarto file (named `homework6.qmd`), 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.qmd` or include author name may result in 5 points taken off.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.4.2     v purrr   1.0.2
## v tibble  3.2.1     v dplyr   1.1.2
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(lubridate)
```

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

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

```
library(dplyr)
```

Question 1

16 points

Obtain a copy of the football-values lecture. Save the five 2024 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):

```r
# 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
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
  year <- 2024
  positions <- c('k','qb','rb','te','wr')
  csvfile <- paste('proj_', positions, substr(year, 3, 4), '.csv', sep='')
  files <- file.path(year, 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'])

  ## append 5 df to 1
  k[,'pos'] <- 'k'
  qb[,'pos'] <- 'qb'
  rb[,'pos'] <- 'rb'
  te[,'pos'] <- 'te'
  wr[,'pos'] <- 'wr'

  cols <- unique(c(names(k), names(qb), names(rb), names(te), names(wr)))
  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


  x <- rbind(k[,cols], qb[,cols], rb[,cols], te[,cols], wr[,cols])

  ##calculate points based on point allocation
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 indeces 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 value
ix_group = c("k", "qb", "rb", "te", "wr")

for (i in (1: length(ix_group))) {
  if (posReq[[ix_group[i]]] == 0){
    next
  } else {
    x2[which(x2[,'pos']==ix_group[i]), 'marg'] <- x2[which(x2[,'pos']==ix_group[i]),'points'] - x2[whic
  }

}

# x2[k.ix, 'marg'] <- x2[k.ix,'points'] - x2[k.ix[posReq[["k"]]*nTeams], 'points']
# x2[qb.ix, 'marg'] <- x2[qb.ix,'points'] - x2[qb.ix[posReq[["qb"]]*nTeams],'points']
# x2[rb.ix, 'marg'] <- x2[rb.ix,'points'] - x2[rb.ix[posReq[["rb"]]*nTeams],'points']
# x2[te.ix, 'marg'] <- x2[te.ix,'points'] - x2[te.ix[posReq[["te"]]*nTeams],'points']
# x2[wr.ix, 'marg'] <- x2[wr.ix,'points'] - x2[wr.ix[posReq[["wr"]]*nTeams],'points']

# create a new data.frame subset by non-negative marginal points
x3 <- x2[x2[,'marg'] >= 0 & !is.na(x2[,'marg']),]
# re-order by marginal points
x3 <- x3[order(x3[,'marg'], decreasing=TRUE),]
# reset the row names
rownames(x3) <- NULL

  ## calculate dollar values
x3[,'value'] <- (nTeams*cap-nrow(x3)) * x3[,'marg'] / sum(x3[,'marg']) + 1

x4 <- x3[,c('PlayerName','pos','points',"marg",'value')]
x4 <- x4[order(x4[,'value'], decreasing=TRUE),]

  ## save dollar values as CSV file
write_csv(x4, file)
```

```r
  ## return data.frame with dollar values
return(x4)
}
```

1. Call `x1 <- ffvalues('.')`

```r
x1 <- ffvalues('.')
```

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

```r
length(which(x1[,'value']>20))
```

```
## [1] 44
```

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

```r
rb.ix <- which(x1[,'pos']=='rb')
x1[rb.ix[15], ]
```

```
##           PlayerName pos points   marg    value
## 31 De'Von Achane MIA  rb  166.8 26.975 29.59927
```

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

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

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

```r
length(which(x2[,'value']>20))
```

```
## [1] 42
```

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

```r
wr.ix <- which(x2[,'pos']=='wr')
sum((wr.ix <= 40))
```

```
## [1] 12
```

1. Call:

   ```r
   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)

```r
length(which(x3[,'value']>20))
```

```
## [1] 48
```

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

```r
wr.ix <- which(x3[,'pos']=='qb')
sum((wr.ix <= 30))
```

```
## [1] 14
```

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`.

```r
haart <- read.csv("haart.csv")
colnames(haart)
```

```
##  [1] "male"        "age"         "aids"        "cd4baseline" "logvl"
##  [6] "weight"      "hemoglobin"  "init.reg"    "init.date"   "last.visit"
## [11] "death"       "date.death"
```

```r
haart$init.date <- as.Date(haart$init.date, format="%m/%d/%y")
haart$last.visit <- as.Date(haart$last.visit, format="%m/%d/%y")
haart$date.death <- as.Date(haart$date.death, format="%m/%d/%y")

haart$init.year <- year(haart$init.date)

table(haart$init.year)
```

```
##
## 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?

```r
# Create indicator
haart$death.within.1yr <- ifelse(!is.na(haart$date.death) &
                                  difftime(haart$date.death, haart$init.date, units = "days") <= 365, 1

head(haart)
```

```
##   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 init.year death.within.1yr
## 1 2007-02-26     0       <NA>      2003               0
## 2 2008-02-22     0       <NA>      2004               0
## 3 2005-11-21     1 2006-01-11      2003               0
## 4 2006-05-05     1 2006-05-07      2006               1
## 5 2007-11-13     0       <NA>      2004               0
## 6 2008-02-28     0       <NA>      2003               0
```

```
# Count how many observations died within 1 year
sum(haart$death.within.1yr)
```

```
## [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.

```
haart$followup.time <- pmin(difftime(haart$date.death, haart$init.date, units = "days"),
                            difftime(haart$last.visit, haart$init.date, units = "days"),
                            na.rm = TRUE)

haart$followup.time <- ifelse(haart$followup.time > 365, 365, haart$followup.time)

quantile(haart$followup.time, na.rm = TRUE)
```

```
##     0%    25%    50%    75%   100%
##   0.00 320.75 365.00 365.00 365.00
```

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$lost.to.followup <- ifelse(is.na(haart$date.death) & haart$followup.time < 365, 1, 0)
sum(haart$lost.to.followup)
```

```
## [1] 173
```

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?

```r
# Split init.reg into separate columns for each drug regimen
haart <- haart %>%
  mutate(across(starts_with("init.reg"), ~ as.character(.))) %>%
  separate(init.reg, into = paste0("drug_", 1:3), sep = ",")  # Adjust separator and number of drugs as
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 76 rows [20, 62, 69,
## 86, 94, 102, 112, 122, 137, 147, 149, 153, 162, 176, 212, 216, 218, 219, 236,
## 242, ...].
```

```r
# Convert each unique drug to its own indicator variable
drugs <- unique(c(haart$drug_1, haart$drug_2, haart$drug_3))
for (drug in drugs) {
  haart[paste0("ndrug_", drug)] <- ifelse(grepl(drug, paste(haart$drug_1, haart$drug_2, haart$drug_3, se
}

# Find drug regimens found more than 100 times
colSums(haart[, grepl("ndrug_", colnames(haart))]) > 100
```

```
## ndrug_3TC ndrug_D4T ndrug_DDI ndrug_ABC ndrug_FTC ndrug_AZT ndrug_EFV ndrug_LPV
##      TRUE      TRUE     FALSE     FALSE     FALSE      TRUE      TRUE     FALSE
## ndrug_ATV ndrug_NVP ndrug_DDC ndrug_RTV ndrug_FPV ndrug_IDV ndrug_TDF ndrug_SQV
##     FALSE      TRUE     FALSE     FALSE     FALSE     FALSE     FALSE     FALSE
## ndrug_NFV
##     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.

```r
haart2 <- read.csv("haart2.csv")
haart2
```

```
##   male      age aids cd4baseline    logvl  weight hemoglobin    init.reg
## 1    0 27.00000    0         232       NA      NA         NA 3TC,AZT,NVP
## 2    1 38.72142    0         170       NA 84.0000         NA 3TC,AZT,NVP
## 3    1 23.00000   NA         154 3.995635 65.5000         14 3TC,DDI,EFV
## 4    0 31.00000    0         236       NA 45.8136         NA 3TC,D4T,NVP
##   init.date last.visit death date.death
## 1   12/1/03     1/5/04     0         NA
## 2   9/26/02    3/29/04     0         NA
## 3   1/31/07    4/16/07     0         NA
## 4   12/3/03   10/11/07     0         NA
```

```r
# Convert date columns
haart2$init.date <- as.Date(haart2$init.date, format="%m/%d/%y")
haart2$last.visit <- as.Date(haart2$last.visit, format="%m/%d/%y")
haart2$date.death <- as.Date(haart2$date.death, format="%m/%d/%y")

# Append to master dataset
haart_combined <- bind_rows(haart, haart2)
```

```r
# Show first 5 and last 5 records of the complete dataset
head(haart_combined, 5)
```

```
##   male age aids cd4baseline logvl  weight hemoglobin drug_1 drug_2 drug_3
## 1    1  25    0          NA    NA      NA         NA    3TC    AZT    EFV
## 2    1  49    0         143    NA 58.0608         11    3TC    AZT    EFV
## 3    1  42    1         102    NA 48.0816          1    3TC    AZT    EFV
## 4    0  33    0         107    NA 46.0000         NA    3TC    AZT    NVP
## 5    1  27    0          52     4      NA         NA    3TC    D4T    EFV
##    init.date last.visit death date.death init.year death.within.1yr
## 1 2003-07-01 2007-02-26     0       <NA>      2003                0
## 2 2004-11-23 2008-02-22     0       <NA>      2004                0
## 3 2003-04-30 2005-11-21     1 2006-01-11      2003                0
## 4 2006-03-25 2006-05-05     1 2006-05-07      2006                1
## 5 2004-09-01 2007-11-13     0       <NA>      2004                0
##   followup.time lost.to.followup ndrug_3TC ndrug_D4T ndrug_DDI ndrug_ABC
## 1           365                0         1         0         0         0
## 2           365                0         1         0         0         0
## 3           365                0         1         0         0         0
## 4            41                0         1         0         0         0
## 5           365                0         1         1         0         0
##   ndrug_FTC ndrug_AZT ndrug_EFV ndrug_LPV ndrug_ATV ndrug_NVP ndrug_DDC
## 1         0         1         1         0         0         0         0
## 2         0         1         1         0         0         0         0
## 3         0         1         1         0         0         0         0
## 4         0         1         0         0         0         1         0
## 5         0         0         1         0         0         0         0
##   ndrug_RTV ndrug_FPV ndrug_IDV ndrug_TDF ndrug_SQV ndrug_NFV init.reg
## 1         0         0         0         0         0         0     <NA>
## 2         0         0         0         0         0         0     <NA>
## 3         0         0         0         0         0         0     <NA>
## 4         0         0         0         0         0         0     <NA>
## 5         0         0         0         0         0         0     <NA>
```

```r
tail(haart_combined, 5)
```

```
##      male      age aids cd4baseline    logvl  weight hemoglobin drug_1 drug_2
## 1000    0 40.00000    1         131       NA 46.2672          8    3TC    D4T
## 1001    0 27.00000    0         232       NA      NA         NA   <NA>   <NA>
## 1002    1 38.72142    0         170       NA 84.0000         NA   <NA>   <NA>
## 1003    1 23.00000   NA         154 3.995635 65.5000         14   <NA>   <NA>
## 1004    0 31.00000    0         236       NA 45.8136         NA   <NA>   <NA>
##      drug_3  init.date last.visit death date.death init.year death.within.1yr
## 1000    NVP 2003-07-03 2008-02-29     0       <NA>      2003                0
## 1001   <NA> 2003-12-01 2004-01-05     0       <NA>        NA               NA
## 1002   <NA> 2002-09-26 2004-03-29     0       <NA>        NA               NA
## 1003   <NA> 2007-01-31 2007-04-16     0       <NA>        NA               NA
## 1004   <NA> 2003-12-03 2007-10-11     0       <NA>        NA               NA
##      followup.time lost.to.followup ndrug_3TC ndrug_D4T ndrug_DDI ndrug_ABC
## 1000           365                0         1         1         0         0
## 1001            NA               NA        NA        NA        NA        NA
## 1002            NA               NA        NA        NA        NA        NA
## 1003            NA               NA        NA        NA        NA        NA
## 1004            NA               NA        NA        NA        NA        NA
##       ndrug_FTC ndrug_AZT ndrug_EFV ndrug_LPV ndrug_ATV ndrug_NVP ndrug_DDC
## 1000          0         0         0         0         0         1         0
## 1001         NA        NA        NA        NA        NA        NA        NA
## 1002         NA        NA        NA        NA        NA        NA        NA
## 1003         NA        NA        NA        NA        NA        NA        NA
## 1004         NA        NA        NA        NA        NA        NA        NA
##       ndrug_RTV ndrug_FPV ndrug_IDV ndrug_TDF ndrug_SQV ndrug_NFV    init.reg
## 1000          0         0         0         0         0         0        <NA>
## 1001         NA        NA        NA        NA        NA        NA 3TC,AZT,NVP
## 1002         NA        NA        NA        NA        NA        NA 3TC,AZT,NVP
## 1003         NA        NA        NA        NA        NA        NA 3TC,DDI,EFV
## 1004         NA        NA        NA        NA        NA        NA 3TC,D4T,NVP
```