Bios 6301: Assignment 6

Jonathan Lifferth

Due Tuesday, 24 October, 1:00 PM

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

35/40 - 5/40 for one day late = 30/40

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 2023 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 ordered 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
path <- '~/R_projects/statistical_computing'
# 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 <- params$year
  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'])

# 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 points
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))])

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
if (posReq['k'] != 0) {
  x2[k.ix, 'marg'] <- x2[k.ix,'points'] - x2[k.ix[nTeams*posReq['k']],'points']
```

```
  }
  if (posReq['qb'] != 0) {
    x2[qb.ix, 'marg'] <- x2[qb.ix,'points'] - x2[qb.ix[nTeams*posReq['qb']],'points']
  }
  if (posReq['rb'] != 0) {
    x2[rb.ix, 'marg'] <- x2[rb.ix,'points'] - x2[rb.ix[nTeams*posReq['rb']],'points']
  }
  if (posReq['te'] != 0) {
    x2[te.ix, 'marg'] <- x2[te.ix,'points'] - x2[te.ix[nTeams*posReq['te']],'points']
  }
  if (posReq['wr'] != 0) {
  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 = na.omit(x2)
  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

  ## save dollar values as CSV file
  dollar_values <- x3[c('PlayerName','pos','points','value')]
  write.table(dollar_values, file = "dollar_values.csv", sep = ",", col.names = NA)

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

1. Call ‘x1 <- ffvalues(’.’)‘

x1 <- ffvalues(‘.’)

it is treating the questions as code and your answers as text.

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

print(length(which(x1$value > 20)))

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

rbs <- x1[x1$pos == ‘rb’,] rbs <- rbs[order(rbs[,‘points’], decreasing=TRUE),] print(rbs[15,][‘PlayerName’])

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

x2 <- ffvalues(getwd(), ‘16team.csv’, nTeams=16, cap=150)

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

you dont need to call print it will print anyway

print(length(which(x2$value > 20)))

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

print(length(which(x2[1:40,]$pos==‘wr’)))

1. 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)

print(length(which(x3$value > 20)))

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

print(length(which(x3[1:40,]$pos==‘qb’)))

Question 2

24 points

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

```
haart_df <- read.csv('/Users/jonathanlifferth/R_projects/Bios6301/datasets/haart.csv')
```

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

```
names(haart_df)
```

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

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

library(lubridate)
```

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

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

```
haart_df$init.year <- year(haart_df$init.date)
print(table(haart_df$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?

```
one_year_survival <- list()

for (i in 1:nrow(haart_df)) {
  if (haart_df[i,"death"] == 1) {
    if (difftime(haart_df[i,"date.death"], haart_df[i,"init.date"], units = "days")[[1]] < 365) {
#      print(1)
      one_year_survival <- append(one_year_survival, c(1))
    } else {
#      print(0)
      one_year_survival <- append(one_year_survival, c(0))
    }
  } else {
#    print(0)
    one_year_survival <- append(one_year_survival, c(0))
  }

}
haart_df$one_year_survival <- unlist(one_year_survival)

table(haart_df$one_year_survival)
```

```
##
##   0   1
## 908  92
```

92 died within 1 year of initial visit

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.

```
follow_up_days <- list()
haart_df$follow_up_days <- 365
for (i in 1:nrow(haart_df)) {
#  print(i)
  init <- haart_df[i,'init.date']
  last <- haart_df[i,"last.visit"]
  death <- haart_df[i,"death.date"]
```

```r
#  follow_up <- abs(difftime(haart_df[i,'init.date'],
#                         min(c(haart_df[i,"last.visit"],
#                               haart_df[i,"death.date"]))[[1]]))
  follow_up <-abs(difftime(haart_df[i,'init.date'],
                          min(c(haart_df[i,"last.visit"], haart_df[i,"death.date"]), na.rm = TRUE))[[1
  if (follow_up > 365) {
    follow_up <- 365
  }
#  print(follow_up)
#  follow_up_days <- append(follow_up_days,c(follow_up))
  haart_df$follow_up_days[i] <- follow_up
}
```

```
## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf

## Warning in min.default(structure(NA_real_, class = c("POSIXct", "POSIXt"): no
## non-missing arguments to min; returning Inf
```

```r
quantile(haart_df$follow_up_days)
```

```
##   0%  25%  50%  75% 100%
##    0  342  365  365  365
```

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_df$lost_to_followup <- 0
for (i in 1:nrow(haart_df)) {
  if (haart_df$death[i] != 1) {
    follow_up_delta <- abs(difftime(haart_df$init.date[i], haart_df$last.visit[i])[[1]])
    if (follow_up_delta > 365) {
      haart_df$lost_to_followup[i] <- 1
    }
  }
}
table(haart_df$lost_to_followup)
```

would suggest doing simple subtraction of dates

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

should be 173

710 records are lost to follow-up

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?

```
init.reg <- as.character(haart_df[,'init.reg'])
haart_df[['init.reg_list']] <- strsplit(init.reg, ",")
(all_drugs <- unique(unlist(haart_df$init.reg_list)))
```

```
##  [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_df), ncol=length(all_drugs))
for(i in seq_along(all_drugs)) {
  reg_drugs[,i] <- sapply(haart_df$init.reg_list, function(x) all_drugs[i] %in% x)
}
reg_drugs <- data.frame(reg_drugs)
names(reg_drugs) <- all_drugs
head(reg_drugs)
```

```
##    3TC   AZT   EFV   NVP   D4T   ABC   DDI   IDV   LPV   RTV   SQV   FTC   TDF
## 1 TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 6 TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     DDC   NFV   T20   ATV   FPV
## 1 FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE
## 6 FALSE FALSE FALSE FALSE FALSE
```

```
haart_merged <- cbind(haart_df, 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 init.year one_year_survival follow_up_days
## 1 2007-02-26     0       <NA>      2003                 0      365.00000
## 2 2008-02-22     0       <NA>      2004                 0      365.00000
## 3 2005-11-21     1 2006-01-11      2003                 0      365.00000
## 4 2006-05-05     1 2006-05-07      2006                 1       40.95833
## 5 2007-11-13     0       <NA>      2004                 0      365.00000
## 6 2008-02-28     0       <NA>      2003                 0      365.00000
##   lost_to_followup init.reg_list  3TC   AZT   EFV   NVP   D4T   ABC   DDI   IDV
## 1                1 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 2                1 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 3                0 3TC, AZT, EFV TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 4                0 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
## 5                1 3TC, D4T, EFV TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 6                1 3TC, AZT, NVP TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
##     LPV   RTV   SQV   FTC   TDF   DDC   NFV   T20   ATV   FPV
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 6 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
for (drug in all_drugs) {
  if (table(haart_merged[,drug])[2] >= 100) {
    print(drug)
    print(table(haart_merged[,drug])[2])
  }
}
```

```
## [1] "3TC"
## TRUE
##  973
## [1] "AZT"
## TRUE
##  794
## [1] "EFV"
## TRUE
##  516
## [1] "NVP"
## TRUE
##  358
## [1] "D4T"
## TRUE
```

```
##  146
```

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/jonathanlifferth/R_projects/Bios6301/datasets/haart2.csv')

haart2$init.date <- as.POSIXct(haart2$init.date, format = "%m/%d/%y")
haart2$last.visit <- as.POSIXct(haart2$last.visit, format = "%m/%d/%y")
haart2$date.death <- as.POSIXct(haart2$date.death, format = "%m/%d/%y")

one_year_survival <- list()

for (i in 1:nrow(haart2)) {
  if (haart2[i,"death"] == 1) {
    if (difftime(haart2[i,"date.death"], haart2[i,"init.date"], units = "days")[[1]] < 365) {
#      print(1)
      one_year_survival <- append(one_year_survival, c(1))
    } else {
#      print(0)
      one_year_survival <- append(one_year_survival, c(0))
    }
  } else {
#    print(0)
    one_year_survival <- append(one_year_survival, c(0))
  }

}
haart2$one_year_survival <- unlist(one_year_survival)


follow_up_days <- list()
haart2$follow_up_days <- 365
for (i in 1:nrow(haart2)) {
#  print(i)
  init <- haart2[i,'init.date']
  last <- haart2[i,"last.visit"]
  death <- haart2[i,"death.date"]

#  follow_up <- abs(difftime(haart_df[i,'init.date'],
#                        min(c(haart_df[i,"last.visit"],
#                              haart_df[i,"death.date"]))[[1]]))
  follow_up <-abs(difftime(haart2[i,'init.date'],
                           min(c(haart2[i,"last.visit"], haart2[i,"death.date"]), na.rm = TRUE))[[1]])
  if (follow_up > 365) {
    follow_up <- 365
  }
#  print(follow_up)
#  follow_up_days <- append(follow_up_days,c(follow_up))
  haart2$follow_up_days[i] <- follow_up
}
```

```
haart2$lost_to_followup <- 0
for (i in 1:nrow(haart2)) {
  if (haart2$death[i] != 1) {
    follow_up_delta <- abs(difftime(haart2$init.date[i], haart2$last.visit[i])[[1]])
    if (follow_up_delta > 365) {
      haart2$lost_to_followup[i] <- 1
    }
  }
}

init.reg <- as.character(haart2[,'init.reg'])
haart2[['init.reg_list']] <- strsplit(init.reg, ",")
(all_drugs2 <- unique(unlist(haart2$init.reg_list)))
```

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

```
reg_drugs2 <- matrix(FALSE, nrow=nrow(haart2), ncol=length(all_drugs2))
for(i in seq_along(all_drugs2)) {
  print(i)
  reg_drugs2[,i] <- sapply(haart2$init.reg_list, function(x) all_drugs2[i] %in% x)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
```

```
reg_drugs2 <- data.frame(reg_drugs2)
names(reg_drugs2) <- all_drugs2
head(reg_drugs2)
```

```
##    3TC   AZT   NVP   DDI   EFV   D4T
## 1 TRUE  TRUE  TRUE FALSE FALSE FALSE
## 2 TRUE  TRUE  TRUE FALSE FALSE FALSE
## 3 TRUE FALSE FALSE  TRUE  TRUE FALSE
## 4 TRUE FALSE  TRUE FALSE FALSE  TRUE
```

```
haart2_merged <- cbind(haart2, reg_drugs2)
head(haart2_merged)
```

```
##   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 one_year_survival follow_up_days
## 1 2003-12-01 2004-01-05     0       <NA>                 0       35.00000
## 2 2002-09-26 2004-03-29     0       <NA>                 0      365.00000
## 3 2007-01-31 2007-04-16     0       <NA>                 0       74.95833
```

```
## 4 2003-12-03 2007-10-11     0       <NA>                 0      365.00000
##   lost_to_followup init.reg_list  3TC   AZT   NVP   DDI   EFV   D4T
## 1                0 3TC, AZT, NVP TRUE  TRUE  TRUE FALSE FALSE FALSE
## 2                1 3TC, AZT, NVP TRUE  TRUE  TRUE FALSE FALSE FALSE
## 3                0 3TC, DDI, EFV TRUE FALSE FALSE  TRUE  TRUE FALSE
## 4                1 3TC, D4T, NVP TRUE FALSE  TRUE FALSE FALSE  TRUE
```

```
# need to add columns for reg drugs in the previous dataset but not the new dataset
missing_drugs <- setdiff(all_drugs, all_drugs2)
missing_drugs
```

```
##  [1] "ABC" "IDV" "LPV" "RTV" "SQV" "FTC" "TDF" "DDC" "NFV" "T20" "ATV" "FPV"
```

```
for (drug in missing_drugs) {
  haart2_merged[,drug] <- FALSE
}
haart2_merged
```

```
##   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 one_year_survival follow_up_days
## 1 2003-12-01 2004-01-05     0       <NA>                 0       35.00000
## 2 2002-09-26 2004-03-29     0       <NA>                 0      365.00000
## 3 2007-01-31 2007-04-16     0       <NA>                 0       74.95833
## 4 2003-12-03 2007-10-11     0       <NA>                 0      365.00000
##   lost_to_followup init.reg_list  3TC   AZT   NVP   DDI   EFV   D4T   ABC   IDV
## 1                0 3TC, AZT, NVP TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 2                1 3TC, AZT, NVP TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## 3                0 3TC, DDI, EFV TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## 4                1 3TC, D4T, NVP TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE
##     LPV   RTV   SQV   FTC   TDF   DDC   NFV   T20   ATV   FPV
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
haart1_2_merged <- cbind(haart_df, haart2)
head(haart1_2_merged, 5)
```

```
##   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 init.year one_year_survival follow_up_days
## 1 2007-02-26     0       <NA>      2003                 0      365.00000
## 2 2008-02-22     0       <NA>      2004                 0      365.00000
## 3 2005-11-21     1 2006-01-11      2003                 0      365.00000
```

```
## 4 2006-05-05     1 2006-05-07      2006                1       40.95833
## 5 2007-11-13     0       <NA>      2004                0      365.00000
##   lost_to_followup init.reg_list male      age aids cd4baseline    logvl
## 1                1 3TC, AZT, EFV    0 27.00000    0         232       NA
## 2                1 3TC, AZT, EFV    1 38.72142    0         170       NA
## 3                0 3TC, AZT, EFV    1 23.00000   NA         154 3.995635
## 4                0 3TC, AZT, NVP    0 31.00000    0         236       NA
## 5                1 3TC, D4T, EFV    0 27.00000    0         232       NA
##    weight hemoglobin    init.reg  init.date last.visit death date.death
## 1      NA         NA 3TC,AZT,NVP 2003-12-01 2004-01-05     0       <NA>
## 2 84.0000         NA 3TC,AZT,NVP 2002-09-26 2004-03-29     0       <NA>
## 3 65.5000         14 3TC,DDI,EFV 2007-01-31 2007-04-16     0       <NA>
## 4 45.8136         NA 3TC,D4T,NVP 2003-12-03 2007-10-11     0       <NA>
## 5      NA         NA 3TC,AZT,NVP 2003-12-01 2004-01-05     0       <NA>
##   one_year_survival follow_up_days lost_to_followup init.reg_list
## 1                 0       35.00000                0 3TC, AZT, NVP
## 2                 0      365.00000                1 3TC, AZT, NVP
## 3                 0       74.95833                0 3TC, DDI, EFV
## 4                 0      365.00000                1 3TC, D4T, NVP
## 5                 0       35.00000                0 3TC, AZT, NVP
```

```
tail(haart1_2_merged, 5)
```

```
##      male age aids cd4baseline    logvl  weight hemoglobin    init.reg
## 996     1  42    0         164 5.281029 84.0000         12 3TC,AZT,NVP
## 997     0  39    1         125 4.625312      NA         NA 3TC,AZT,EFV
## 998     0  37    0         122       NA 86.1840         11 3TC,AZT,NVP
## 999     0  31    0         102       NA 61.6896         11 3TC,AZT,NVP
## 1000    0  40    1         131       NA 46.2672          8 3TC,D4T,NVP
##       init.date last.visit death date.death init.year one_year_survival
## 996  2005-04-30 2007-04-13     0       <NA>      2005                 0
## 997  2007-04-24       <NA>     1 2007-08-16      2007                 1
## 998  2005-01-12 2008-03-19     0       <NA>      2005                 0
## 999  2003-05-22 2008-03-07     0       <NA>      2003                 0
## 1000 2003-07-03 2008-02-29     0       <NA>      2003                 0
##      follow_up_days lost_to_followup init.reg_list male      age aids
## 996             365                1 3TC, AZT, NVP    0 31.00000    0
## 997             365                0 3TC, AZT, EFV    0 27.00000    0
## 998             365                1 3TC, AZT, NVP    1 38.72142    0
## 999             365                1 3TC, AZT, NVP    1 23.00000   NA
## 1000            365                1 3TC, D4T, NVP    0 31.00000    0
##      cd4baseline    logvl  weight hemoglobin    init.reg  init.date last.visit
## 996          236       NA 45.8136         NA 3TC,D4T,NVP 2003-12-03 2007-10-11
## 997          232       NA      NA         NA 3TC,AZT,NVP 2003-12-01 2004-01-05
## 998          170       NA 84.0000         NA 3TC,AZT,NVP 2002-09-26 2004-03-29
## 999          154 3.995635 65.5000         14 3TC,DDI,EFV 2007-01-31 2007-04-16
## 1000         236       NA 45.8136         NA 3TC,D4T,NVP 2003-12-03 2007-10-11
##      death date.death one_year_survival follow_up_days lost_to_followup
## 996      0       <NA>                 0      365.00000                1
## 997      0       <NA>                 0       35.00000                0
## 998      0       <NA>                 0      365.00000                1
## 999      0       <NA>                 0       74.95833                0
## 1000     0       <NA>                 0      365.00000                1
##      init.reg_list
```

```
## 996  3TC, D4T, NVP
## 997  3TC, AZT, NVP
## 998  3TC, AZT, NVP
## 999  3TC, DDI, EFV
## 1000 3TC, D4T, NVP
```

there should be records 1001, 1002, 1003, and 1004 as part of your tail incuding 1000