Statistical Computing - Exercises 07 - Performance

We return to two problems and add a third: calculating song statistics in the billboard dataset, calculating distances in the flights dataset, and reformatting a hurricane track dataset called hurdat. Below are correct solutions to all three problems. Your task is to make them run faster, while ensuring that your faster code gives the correct result.

Show how long your revised code takes to run, and demonstrate that it gives the right answer.

Song Statistics

```r
hot <- read.csv("../datasets/Hot_100.csv")

weeks_below_k <- function( dat, k ){

    # computes the number of weeks on chart at or below ranking k
    # for each songid in dat

    # get the vector of unique song ids
    unique_songs <- unique( dat$song_id )

    # initialize vector for number of weeks at or below ranking k
    weeks_below_k <- c()

    # loop over all the songs
    for(j in 1:length( unique_songs ) ){

        # find the indices for this song
        inds_song <- which( dat$song_id == unique_songs[j] )

        # initialize a count of number of weeks at or below k
        count <- 0

        # loop over the indices
        for( i in inds_song ){

            # add to the count if the week_position is at or below k
            if( dat$chart_position[i] <= k ){
                count <- count + 1
            }

        }

        # update weeks_below_k with this song's count
        weeks_below_k <- c( weeks_below_k, count )

    }
    # rank the songs
    ord <- order( weeks_below_k, decreasing = TRUE )
    return( data.frame( song_id = unique_songs[ord], weeks_below_k = weeks_below_k[ord] ) )
}

r <- weeks_below_k( hot, 1 )
head(r, n = 20 )
```

```
##                                                                          song_id
## 1                            Old Town RoadLil Nas X Featuring Billy Ray Cyrus
## 2                  DespacitoLuis Fonsi & Daddy Yankee Featuring Justin Bieber
## 3                                     One Sweet DayMariah Carey & Boyz II Men
## 4                                                       As It WasHarry Styles
## 5  Candle In The Wind 1997/Something About The Way You Look TonightElton John
## 6                                          I Gotta FeelingThe Black Eyed Peas
## 7                                       I Will Always Love YouWhitney Houston
## 8                                            I'll Make Love To YouBoyz II Men
## 9                                      Macarena (Bayside Boys Mix)Los Del Rio
## 10                               Uptown Funk!Mark Ronson Featuring Bruno Mars
## 11                                             We Belong TogetherMariah Carey
## 12                              End Of The Road (From "Boomerang")Boyz II Men
## 13                                             The Boy Is MineBrandy & Monica
## 14                                All I Want For Christmas Is YouMariah Carey
## 15                        Blurred LinesRobin Thicke Featuring T.I. + Pharrell
## 16                                           Boom Boom PowThe Black Eyed Peas
## 17                                    CloserThe Chainsmokers Featuring Halsey
## 18                                                        Lose YourselfEminem
## 19                            See You AgainWiz Khalifa Featuring Charlie Puth
## 20                                                     Shape Of YouEd Sheeran
##    weeks_below_k
## 1              19
## 2              16
## 3              16
## 4              15
## 5              14
## 6              14
## 7              14
## 8              14
## 9              14
## 10             14
## 11             14
## 12             13
## 13             13
## 14             12
## 15             12
## 16             12
## 17             12
## 18             12
## 19             12
## 20             12
```

Distance matrix for airline data

```
# distance matrix
dat <- read.csv("../datasets/airline_2019-07-01.csv")

get_distance_matrix <- function( dat ){

    # get the set of unique airports
    ports <- sort( unique( c(dat$Origin, dat$Dest) ) )
    nports <- length(ports)
    # set up a distance matrix
    distmat <- matrix(NA, nports, nports )
    rownames(distmat) <- ports
    colnames(distmat) <- ports

    # loop over all possible origins and destinations
    for(p1 in ports){
        for(p2 in ports){

            # get row and column indices for this pair
            j1 <- which( rownames(distmat) == p1 )
            j2 <- which( colnames(distmat) == p2 )

            # get rows of data frame for this pair, and subset
            ii <- which( dat$Origin == p1 & dat$Dest == p2 )
            subdat <- dat[ii,]

            # find the distance
            distmat[j1,j2] <- subdat$Distance[1]

        }
    }
    return(distmat)
}

distmat <- get_distance_matrix( dat )
v <- c("ATL","ORD","LGA","JFK","DEN","LAX","STL","SEA")
distmat[v,v]
```

```
##      ATL  ORD  LGA  JFK  DEN  LAX  STL  SEA
## ATL   NA  606  762  760 1199 1947  484 2182
## ORD  606   NA  733  740  888 1744  258 1721
## LGA  762  733   NA   NA 1620   NA  888   NA
## JFK  760  740   NA   NA 1626 2475   NA 2422
## DEN 1199  888 1620 1626   NA  862  770 1024
## LAX 1947 1744   NA 2475  862   NA 1592  954
## STL  484  258  888   NA  770 1592   NA 1709
## SEA 2182 1721   NA 2422 1024  954 1709   NA
```

Hurricane data

```
# read in the raw data
dat <- read.csv("../datasets/hurdat2-1851-2021-041922.txt", header= FALSE)

# initialize the processed dataset
hurdat <- data.frame( matrix(NA, 0, ncol(dat)+1) )
colnames(hurdat) <- c("hur_code", colnames(dat))

# counter for the row of hurdat
k <- 0

# loop over rows of raw dataset
for(j in 1:nrow(dat)){
    # extract the current row of raw data
    this_row <- dat[j,]

    # check whether this is a code row
    if( substr( this_row[1,1], 1, 2 ) == "AL" ){

        # if so, update the hurricane code
        hur_code <- this_row[1,1]

    } else {

        # otherwise update the counter and write to the next row
        k <- k + 1
        hurdat[k,] <- cbind( hur_code, this_row )

    }
}
```