Assignment 3

Tushar Ponkshe, tvp2110

October 11, 2018

ii. (a)

```
setwd("C:/Users/tusha/OneDrive/Documents/Columbia Fall 2018 Courses/STAT COMP & INTRO TO DATA SC
IENCE/Datasets")
nets1819 = readLines("NetsSchedule1819.html", warn = FALSE)
length(nets1819)
```

```
## [1] 106
```

ii. (b)

```
sum(nchar(nets1819))
```

```
## [1] 463423
```

ii. (c)

```
max(nchar(nets1819))
```

```
## [1] 249693
```

iii Who and when are they playing first? Who and when are they playing last? Nets first play Detroit Pistons on Wed, Oct 17, 2018 at 7:00 PM and last play Miami Heat on Wed, Apr 10, 2019 at 8:00 PM

iv Line 64 contains information about the games of the regular season

v Regular expression to extract the line that contains the time, location, and opponent of all games

```
#time format is "time":"2018-10-17T23:00Z"
#location format is "location":"Brooklyn","links":"......"
#opponent format is "opponent":{"id":"8","abbrev":"DET","displayName":"Detroit...
line64 <- nets1819[64]
time.expr <- "[0-9]{1}[:][0-9]{2}\\s[P][M]"
grep(time.expr, nets1819)
```

```
## [1] 64 65
```

```
#time <- regmatches(nets1819, regexpr(time.expr, nets1819))
opponent.expr <- '/[a-z]*[-]*[a-z]+\"[>][A-Z]{1,2}\\s*[A-Z]*[a-z]*'
grep(opponent.expr, nets1819)
```

```
## [1] 64
```

```
#regmatches(nets1819, regexpr(opponent.expr, nets1819))
location.expr <- "\"location\":\"[A-Z]+[a-z]*\""
grep(location.expr, nets1819)
```

```
## [1] 65
```

vii

```
date.expr <- "[A-Z]{1}[a-z]{2}[[:punct:]]\\s[A-Z]{1}[a-z]{2}\\s[0-9]{1,2}"
lines = gregexpr(date.expr, nets1819)
line64_82 = lines[[64]]
line64_82
```

```
##  [1] 144299 145499 146703 147896 149114 150360 151592 152790 153996 155207
## [11] 156428 157607 158789 160026 161261 162453 163667 164860 166022 167227
## [21] 168465 169696 170879 172100 173311 174540 175774 176985 178174 179414
## [31] 180636 181842 183118 184307 185510 186712 187929 189160 190356 191586
## [41] 192786 193970 195152 196353 197561 198768 199960 201147 202359 203563
## [51] 204766 205958 207155 208382 209559 210823 212020 213221 214429 216277
## [61] 217511 218739 219953 221180 222392 223556 224814 226041 227232 228444
## [71] 229692 230855 232034 233241 234456 235767 236998 238205 239409 240618
## [81] 241862 243061
## attr(,"match.length")
##  [1] 11 11 11 11 11 11 11 11 10 10 10 10 11 11 11 11 11 11 11 11 11 11 11
## [24] 10 10 10 10 10 11 11 11 11 11 11 11 11 11 11 10 10 10 10 10 11 11 11
## [47] 11 11 11 11 11 11 11 10 10 10 10 11 11 11 11 11 11 10 10 10 10 10 11
## [70] 11 11 11 11 11 11 11 11 10 10 10 10 11
```

It contains 82 lines and as seen later in (vi), every line contains information of a single game. Confirmed also that the first and the last games match the ones found in (ii) (see vi output).

viii

```
date <- regmatches(line64, gregexpr(date.expr, line64))
date
```

```
## [[1]]
##  [1] "Wed, Oct 17" "Fri, Oct 19" "Sat, Oct 20" "Wed, Oct 24" "Fri, Oct 26"
##  [6] "Sun, Oct 28" "Mon, Oct 29" "Wed, Oct 31" "Fri, Nov 2"  "Sun, Nov 4"
## [11] "Tue, Nov 6"  "Fri, Nov 9"  "Sat, Nov 10" "Mon, Nov 12" "Wed, Nov 14"
## [16] "Fri, Nov 16" "Sat, Nov 17" "Tue, Nov 20" "Wed, Nov 21" "Fri, Nov 23"
## [21] "Sun, Nov 25" "Wed, Nov 28" "Fri, Nov 30" "Sat, Dec 1"  "Mon, Dec 3"
## [26] "Wed, Dec 5"  "Fri, Dec 7"  "Sat, Dec 8"  "Wed, Dec 12" "Fri, Dec 14"
## [31] "Sun, Dec 16" "Tue, Dec 18" "Wed, Dec 19" "Fri, Dec 21" "Sun, Dec 23"
## [36] "Wed, Dec 26" "Fri, Dec 28" "Sat, Dec 29" "Wed, Jan 2"  "Fri, Jan 4"
## [41] "Sun, Jan 6"  "Mon, Jan 7"  "Wed, Jan 9"  "Fri, Jan 11" "Mon, Jan 14"
## [46] "Wed, Jan 16" "Fri, Jan 18" "Mon, Jan 21" "Wed, Jan 23" "Fri, Jan 25"
## [51] "Mon, Jan 28" "Tue, Jan 29" "Thu, Jan 31" "Sat, Feb 2"  "Mon, Feb 4"
## [56] "Wed, Feb 6"  "Fri, Feb 8"  "Mon, Feb 11" "Wed, Feb 13" "Thu, Feb 21"
## [61] "Sat, Feb 23" "Mon, Feb 25" "Wed, Feb 27" "Fri, Mar 1"  "Sat, Mar 2"
## [66] "Mon, Mar 4"  "Wed, Mar 6"  "Sat, Mar 9"  "Mon, Mar 11" "Wed, Mar 13"
## [71] "Sat, Mar 16" "Sun, Mar 17" "Tue, Mar 19" "Fri, Mar 22" "Mon, Mar 25"
## [76] "Thu, Mar 28" "Sat, Mar 30" "Mon, Apr 1"  "Wed, Apr 3"  "Sat, Apr 6"
## [81] "Sun, Apr 7"  "Wed, Apr 10"
```

vi Write a regular expression to split the whole line into 82 lines, with each line displaying the information of one game. (You may obtain some hint from problem (vii). Information is date, time, opponent.

```
loc <- gregexpr(date.expr,nets1819)
vec = c()
for (i in 1:81) {
  vec[i] = substr(line64, loc[[64]][i], loc[[64]][i+1])
}
last = substr(line64, 243061, 243061+1210)
vec[82] = last
head(vec,3)
```

```
## [1] "Wed, Oct 17</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-
logo\"><span class=\"pr2\">@</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a hr
ef=\"/nba/team/_/name/det/detroit-pistons\"><img alt=\"Detroit\" title=\"Detroit\" src=\"http://
a.espncdn.com/combiner/i?img=/i/teamlogos/nba/500/det.png&amp;w=50&amp;h=50\"/> </a></span><span
><a href=\"/nba/team/_/name/det/detroit-pistons\">Detroit<!-- --> </a></span></div></td><td clas
s=\"Table2__td\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401070694\" to=
\"http://www.espn.com/nba/game?gameId=401070694\">7:00 PM<!-- --> </a></span></td><td class=\"Ta
ble2__td\"></td><td colSpan=\"3\" class=\"Table2__td\"><a class=\"Schedule__ticket\" href=\"https
://www.vividseats.com/nba-basketball/detroit-pistons-tickets/pistons-10-17-2805257.html?wsUser=
717\" target=\"_blank\" to=\"https://www.vividseats.com/nba-basketball/detroit-pistons-tickets/p
istons-10-17-2805257.html?wsUser=717\">2,234 tickets as low as $11<!-- --> <svg class=\"external
ml2 icon__svg\" viewBox=\"0 0 24 24\"><use xlink:href=\"#icon__external\"></use></svg></a></td>
</tr><tr class=\"filled Table2__tr Table2__tr--sm Table2__even\" data-idx=\"3\"><td class=\"Tabl
e2__td\"><span>F"
## [2] "Fri, Oct 19</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-
logo\"><span class=\"pr2\">vs</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a h
ref=\"/nba/team/_/name/ny/new-york-knicks\"><img alt=\"New York\" title=\"New York\" src=\"htt
p://a.espncdn.com/combiner/i?img=/i/teamlogos/nba/500/ny.png&amp;w=50&amp;h=50\"/> </a></span><s
pan><a href=\"/nba/team/_/name/ny/new-york-knicks\">New York<!-- --> </a></span></div></td><td c
lass=\"Table2__td\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401070704\" t
o=\"http://www.espn.com/nba/game?gameId=401070704\">7:30 PM<!-- --> </a></span></td><td class=
\"Table2__td\"></td><td colSpan=\"3\" class=\"Table2__td\"><a class=\"Schedule__ticket\" href=
\"https://www.vividseats.com/nba-basketball/brooklyn-nets-tickets/nets-vs-knicks-10-19-2805409.h
tml?wsUser=717\" target=\"_blank\" to=\"https://www.vividseats.com/nba-basketball/brooklyn-nets-
tickets/nets-vs-knicks-10-19-2805409.html?wsUser=717\">2,617 tickets as low as $54<!-- --> <svg
class=\"external ml2 icon__svg\" viewBox=\"0 0 24 24\"><use xlink:href=\"#icon__external\"></use
></svg></a></td></tr><tr class=\"Table2__tr Table2__tr--sm Table2__even\" data-idx=\"4\"><td cla
ss=\"Table2__td\"><span>S"
## [3] "Sat, Oct 20</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-
logo\"><span class=\"pr2\">@</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a hr
ef=\"/nba/team/_/name/ind/indiana-pacers\"><img alt=\"Indiana\" title=\"Indiana\" src=\"http://
a.espncdn.com/combiner/i?img=/i/teamlogos/nba/500/ind.png&amp;w=50&amp;h=50\"/> </a></span><span
><a href=\"/nba/team/_/name/ind/indiana-pacers\">Indiana<!-- --> </a></span></div></td><td class
=\"Table2__td\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401070710\" to=
\"http://www.espn.com/nba/game?gameId=401070710\">7:00 PM<!-- --> </a></span></td><td class=\"Ta
ble2__td\"></td><td colSpan=\"3\" class=\"Table2__td\"><a class=\"Schedule__ticket\" href=\"https
://www.vividseats.com/nba-basketball/indiana-pacers-tickets/pacers-10-20-2805553.html?wsUser=71
7\" target=\"_blank\" to=\"https://www.vividseats.com/nba-basketball/indiana-pacers-tickets/pace
rs-10-20-2805553.html?wsUser=717\">2,457 tickets as low as $8<!-- --> <svg class=\"external ml2
icon__svg\" viewBox=\"0 0 24 24\"><use xlink:href=\"#icon__external\"></use></svg></a></td></tr>
<tr class=\"filled Table2__tr Table2__tr--sm Table2__even\" data-idx=\"5\"><td class=\"Table2__t
d\"><span>W"
```

```
tail(vec, 3)
```

```
## [1] "Sat, Apr 6</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-l
ogo\"><span class=\"pr2\">@</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a hre
f=\"/nba/team/_/name/mil/milwaukee-bucks\"><img alt=\"Milwaukee\" title=\"Milwaukee\" src=\"htt
p://a.espncdn.com/combiner/i?img=/i/teamlogos/nba/500/mil.png&amp;w=50&amp;h=50\"/> </a></span><
span><a href=\"/nba/team/_/name/mil/milwaukee-bucks\">Milwaukee<!-- --> </a></span></div></td><t
d class=\"Table2__td\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401071866
\" to=\"http://www.espn.com/nba/game?gameId=401071866\">5:00 PM<!-- --> </a></span></td><td clas
s=\"Table2__td\"><div class=\"network-container\"><div>NBATV</div></div></td><td colSpan=\"3\" c
lass=\"Table2__td\"><a class=\"Schedule__ticket\" href=\"https://www.vividseats.com/nba-basketba
ll/milwaukee-bucks-tickets/bucks-4-6-2805033.html?wsUser=717\" target=\"_blank\" to=\"https://ww
w.vividseats.com/nba-basketball/milwaukee-bucks-tickets/bucks-4-6-2805033.html?wsUser=717\">1,68
4 tickets as low as $15<!-- --> <svg class=\"external ml2 icon__svg\" viewBox=\"0 0 24 24\"><use
xlink:href=\"#icon__external\"></use></svg></a></td></tr><tr class=\"Table2__tr Table2__tr--sm T
able2__even\" data-idx=\"82\"><td class=\"Table2__td\"><span>S"
## [2] "Sun, Apr 7</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-l
ogo\"><span class=\"pr2\">@</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a hre
f=\"/nba/team/_/name/ind/indiana-pacers\"><img alt=\"Indiana\" title=\"Indiana\" src=\"http://a.
espncdn.com/combiner/i?img=/i/teamlogos/nba/500/ind.png&amp;w=50&amp;h=50\"/> </a></span><span><
a href=\"/nba/team/_/name/ind/indiana-pacers\">Indiana<!-- --> </a></span></div></td><td class=
\"Table2__td\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401071872\" to=\"h
ttp://www.espn.com/nba/game?gameId=401071872\">5:00 PM<!-- --> </a></span></td><td class=\"Table
2__td\"></td><td colSpan=\"3\" class=\"Table2__td\"><a class=\"Schedule__ticket\" href=\"http
s://www.vividseats.com/nba-basketball/indiana-pacers-tickets/pacers-4-7-2805754.html?wsUser=717
\" target=\"_blank\" to=\"https://www.vividseats.com/nba-basketball/indiana-pacers-tickets/pacer
s-4-7-2805754.html?wsUser=717\">2,396 tickets as low as $12<!-- --> <svg class=\"external ml2 ic
on__svg\" viewBox=\"0 0 24 24\"><use xlink:href=\"#icon__external\"></use></svg></a></td></tr><t
r class=\"filled bb--none Table2__tr Table2__tr--sm Table2__even\" data-idx=\"83\"><td class=\"T
able2__td\"><span>W"
## [3] "Wed, Apr 10</span></td><td class=\"Table2__td\"><div class=\"flex items-center opponent-
logo\"><span class=\"pr2\">vs</span><span class=\"tc pr2\" style=\"width:20px;height:20px\"><a h
ref=\"/nba/team/_/name/mia/miami-heat\"><img alt=\"Miami\" title=\"Miami\" src=\"http://a.espncd
n.com/combiner/i?img=/i/teamlogos/nba/500/mia.png&amp;w=50&amp;h=50\"/> </a></span><span><a href
=\"/nba/team/_/name/mia/miami-heat\">Miami<!-- --> </a></span></div></td><td class=\"Table2__td
\"><span class=\"\"><a href=\"http://www.espn.com/nba/game?gameId=401071894\" to=\"http://www.es
pn.com/nba/game?gameId=401071894\">8:00 PM<!-- --> </a></span></td><td class=\"Table2__td\"></td
><td colSpan=\"3\" class=\"Table2__td\"><a class=\"Schedule__ticket\" href=\"https://www.vividse
ats.com/nba-basketball/brooklyn-nets-tickets/nets-vs-heat-4-10-2805394.html?wsUser=717\" target=
\"_blank\" to=\"https://www.vividseats.com/nba-basketball/brooklyn-nets-tickets/nets-vs-heat-4-1
0-2805394.html?wsUser=717\">2,441 tickets as low as $39<!-- --> <svg class=\"external ml2 icon__
svg\" viewBox=\"0 0 24 24\"><use xlink:href=\"#icon__external\"></use></svg></a></td></tr></tbod
y></table></td></tr></tbody></table></div><div class=\"Table2__shadow--right\" style=\"opacity:0
\"></div></div></div></t"
```

```
length(vec)
```

```
## [1] 82
```

The 64th line containing information on all the games is split into 82 lines each containing information on single games. As from code output, the first game is on Wed, Oct 17 and the last game is on Wed, Apr 10.

ix

```
time <- regmatches(line64, gregexpr(time.expr, line64))
time
```

```
## [[1]]
##  [1] "7:00 PM" "7:30 PM" "7:00 PM" "7:00 PM" "8:00 PM" "5:00 PM" "7:30 PM"
##  [8] "7:30 PM" "7:30 PM" "6:00 PM" "9:00 PM" "9:00 PM" "8:30 PM" "8:00 PM"
## [15] "7:30 PM" "7:00 PM" "6:00 PM" "7:30 PM" "8:30 PM" "2:00 PM" "6:00 PM"
## [22] "7:30 PM" "7:30 PM" "7:00 PM" "7:30 PM" "7:30 PM" "7:30 PM" "7:30 PM"
## [29] "7:00 PM" "7:30 PM" "3:00 PM" "7:30 PM" "8:00 PM" "7:30 PM" "6:00 PM"
## [36] "7:30 PM" "7:00 PM" "5:00 PM" "7:30 PM" "8:00 PM" "3:30 PM" "7:30 PM"
## [43] "7:30 PM" "7:30 PM" "7:30 PM" "8:00 PM" "7:00 PM" "3:30 PM" "7:30 PM"
## [50] "7:30 PM" "7:30 PM" "7:30 PM" "8:30 PM" "7:00 PM" "7:30 PM" "7:30 PM"
## [57] "7:30 PM" "7:30 PM" "7:00 PM" "7:30 PM" "7:00 PM" "7:30 PM" "7:30 PM"
## [64] "7:30 PM" "7:30 PM" "7:30 PM" "7:30 PM" "7:00 PM" "7:30 PM" "8:00 PM"
## [71] "9:00 PM" "9:00 PM" "0:00 PM" "0:30 PM" "0:00 PM" "7:00 PM" "6:00 PM"
## [78] "7:30 PM" "7:30 PM" "5:00 PM" "5:00 PM" "8:00 PM"
```

x

```
# [<][i-l]{2}\\s[a-z]{5}[=]\\"[a-z]{4}[-][a-z]{6}\\"[>][@|v]


home.expr <- '[<]div class=\"flex items-center opponent-logo\"><span class=\"pr2\">[@|vs]'
homeaway <- regmatches(line64, gregexpr(home.expr, line64))
homeaway <- homeaway[[1]]
homeboolean <- (substr(homeaway, nchar(homeaway[1]), nchar(homeaway[1])) == "v")
home = c()
for (i in 1:82) {
  home[i] = as.numeric(homeboolean[i])
}
home
```

```
##  [1] 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 1 1 1 1 0 1 1 1 0 0 1 1 1 0 1 1
## [36] 1 0 0 1 0 0 0 1 0 1 0 0 1 1 1 0 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1 1 0 1 0
## [71] 0 0 0 0 0 0 1 1 1 0 0 1
```

xi

```
opponent.expr <- '/[a-z]*[-]*[a-z]+[-][6,7]*[a-z]+\">[A-Z]{1,2}\\s*[A-Z]*[a-z]*\\s*[A-Z]*[a-z]*'
opponent.info <- (regmatches(nets1819, gregexpr(opponent.expr, nets1819)))[[64]]

want <- '[A-Z]{1,2}[a-z]*[-]*[6-7]*[a-z]*\\s*[A-Z]*[a-z]*'
opponent.info1 <- regmatches(opponent.info, gregexpr(want, opponent.info))

opponent = c()

for (i in 1:82) {
  opponent[i] = opponent.info1[[i]]
}

head(opponent)
```

```
## [1] "Detroit"      "New York"     "Indiana"      "Cleveland"
## [5] "New Orleans"  "Golden State"
```

```
tail(opponent)
```

```
## [1] "Boston"    "Milwaukee" "Toronto"   "Milwaukee" "Indiana"   "Miami"
```

```
length(opponent)
```

```
## [1] 82
```

xii

```
data.frame(date, time, opponent, home)[1:10,]
```

```
##    c..Wed..Oct.17....Fri..Oct.19....Sat..Oct.20....Wed..Oct.24...
## 1                                                    Wed, Oct 17
## 2                                                    Fri, Oct 19
## 3                                                    Sat, Oct 20
## 4                                                    Wed, Oct 24
## 5                                                    Fri, Oct 26
## 6                                                    Sun, Oct 28
## 7                                                    Mon, Oct 29
## 8                                                    Wed, Oct 31
## 9                                                     Fri, Nov 2
## 10                                                    Sun, Nov 4
##    c..7.00.PM....7.30.PM....7.00.PM....7.00.PM....8.00.PM....5.00.PM...
## 1                                                                7:00 PM
## 2                                                                7:30 PM
## 3                                                                7:00 PM
## 4                                                                7:00 PM
## 5                                                                8:00 PM
## 6                                                                5:00 PM
## 7                                                                7:30 PM
## 8                                                                7:30 PM
## 9                                                                7:30 PM
## 10                                                               6:00 PM
##        opponent home
## 1       Detroit    0
## 2      New York    1
## 3       Indiana    0
## 4     Cleveland    0
## 5   New Orleans    0
## 6  Golden State    1
## 7      New York    0
## 8       Detroit    1
## 9       Houston    1
## 10 Philadelphia    1
```

The display format is a bit weird, but you can see that the first 10 games in the output match the first 10 games as seen on the website.