Lotto_Data_Cleaning

Chris Richardson

February 17, 2019

```
library(stringi)
library(stringr)
```

```
substrRight <- function(x, n){
  substr(x, nchar(x)-n+1, nchar(x))
}
```

```
# Data comes in TXT form from https://www.calottery.com/play/draw-games/daily-3/winning-numbers
df <-  read.table('C:/Users/crich/Desktop/Googz/lottoResearch/Data/DownloadAllNumbers.txt', sep
 = '\t')
```

```
# REMOVE FIRST 3 UNNECESSARY ROWS
# which are unnecessary text which makes R unable to render the table
lotto <-  data.frame(df[-(1:3),])
head(lotto, n=5)
```

```
##                                                    df...1.3....
## 1 15752     Fri. Feb 15, 2019         9         5          8
## 2 15751     Thu. Feb 14, 2019         3         5          7
## 3 15750     Thu. Feb 14, 2019         7         6          6
## 4 15749     Wed. Feb 13, 2019         5         5          1
## 5 15748     Wed. Feb 13, 2019         0         6          7
```

```
# Rename column
colnames(lotto) <- 'Record'

# remove white space
lotto$strip <-  gsub("\\s", "", lotto$Record)

lotto$digitsOnly <-  gsub('\\D','', lotto$strip)

# this is built on the constant variable of digits always ending in order 'Serial #, date, and w
inning #s', wtih a char length of 9, thus we clip anythin unnecessary
lotto$Serial <-  as.integer(substr(lotto$digitsOnly, 1, nchar(lotto$digitsOnly)-9))

head(lotto, n = 5)
```

```
##                                                                 Record
## 1 15752     Fri. Feb 15, 2019          9          5          8
## 2 15751     Thu. Feb 14, 2019          3          5          7
## 3 15750     Thu. Feb 14, 2019          7          6          6
## 4 15749     Wed. Feb 13, 2019          5          5          1
## 5 15748     Wed. Feb 13, 2019          0          6          7
##                     strip     digitsOnly Serial
## 1 15752Fri.Feb15,2019958 15752152019958  15752
## 2 15751Thu.Feb14,2019357 15751142019357  15751
## 3 15750Thu.Feb14,2019766 15750142019766  15750
## 4 15749Wed.Feb13,2019551 15749132019551  15749
## 5 15748Wed.Feb13,2019067 15748132019067  15748
```

```
# Extract DaY as %b%d%Y
lotto$stringDate <- str_extract(lotto$Record, '(\\D{9}).+(\\d{2})')
lotto$stringDate <- substring(lotto$stringDate, 10)

# Remove whitespace from Date
lotto$stringDate <- gsub("\\s", "", lotto$stringDate)

# insert , between month & day for regex purposes
lotto$stringDate <-  sub('^(.{3})(.*)', "\\1,\\2", lotto$stringDate)

lotto$stringDate <-  gsub(",", "/", lotto$stringDate)

# convert stringDate into international date standard
lotto$Date <-  as.Date(lotto$stringDate, format = '%b/%d/%Y' )

# grab string containing date
lotto$Day <-  gsub('\\d','', lotto$strip)
# grab exact day
lotto$Day <-  stri_sub(lotto$Day, 1, 3)

# label rows according to draw time
lotto$Time[lotto$Serial %% 2 == 1] <-  'Evening'
lotto$Time[lotto$Serial %% 2 == 0] <- 'Midday'

# grab winning #s
lotto$Winning <-  as.character(substrRight(lotto$strip,3))
head(lotto, n=5)
```

```
##                                                                 Record
## 1 15752     Fri. Feb 15, 2019          9          5          8
## 2 15751     Thu. Feb 14, 2019          3          5          7
## 3 15750     Thu. Feb 14, 2019          7          6          6
## 4 15749     Wed. Feb 13, 2019          5          5          1
## 5 15748     Wed. Feb 13, 2019          0          6          7
##                     strip     digitsOnly Serial  stringDate       Date Day
## 1 15752Fri.Feb15,2019958 15752152019958  15752 Feb/15/2019 2019-02-15 Fri
## 2 15751Thu.Feb14,2019357 15751142019357  15751 Feb/14/2019 2019-02-14 Thu
## 3 15750Thu.Feb14,2019766 15750142019766  15750 Feb/14/2019 2019-02-14 Thu
## 4 15749Wed.Feb13,2019551 15749132019551  15749 Feb/13/2019 2019-02-13 Wed
## 5 15748Wed.Feb13,2019067 15748132019067  15748 Feb/13/2019 2019-02-13 Wed
##      Time Winning
## 1  Midday     958
## 2 Evening     357
## 3  Midday     766
## 4 Evening     551
## 5  Midday     067
```

```
# Remove
lotto <-  lotto[,c(1,4,6:9)]

lotto$One <- as.integer(stri_sub(lotto$Winning, 1, -3))
lotto$Two <-  as.integer(stri_sub(lotto$Winning, 2, -2))
lotto$Three <-  as.integer(stri_sub(lotto$Winning, 3, -1))

# column for the total sum of winning numbers
lotto$WinningSum <- rowSums(lotto[,7:9])

#display columns built from Record column (column number one/lotto[1])
head(lotto[,c(2:10)], n = 5)
```

```
##   Serial       Date Day    Time Winning One Two Three WinningSum
## 1  15752 2019-02-15 Fri  Midday     958   9   5     8         22
## 2  15751 2019-02-14 Thu Evening     357   3   5     7         15
## 3  15750 2019-02-14 Thu  Midday     766   7   6     6         19
## 4  15749 2019-02-13 Wed Evening     551   5   5     1         11
## 5  15748 2019-02-13 Wed  Midday     067   0   6     7         13
```

```
# copy original claned DF to new DF for future experimentation
lottoExtended = lotto

# split winning numbers into two columns consisting of 1&2 and 2&3
lottoExtended$FrontTwo <- stri_sub(lotto$Winning, 1,2)
lottoExtended$LastTwo <-  stri_sub(lotto$Winning,2,3)

lottoExtended$FrontTwoSum <-  lotto$One + lotto$Two
lottoExtended$LastTwoSum <- lotto$Two + lotto$Three
head(lottoExtended[,c(5:12)])
```

```
##      Time Winning One Two Three WinningSum FrontTwo LastTwo
## 1  Midday     958   9   5     8         22       95      58
## 2 Evening     357   3   5     7         15       35      57
## 3  Midday     766   7   6     6         19       76      66
## 4 Evening     551   5   5     1         11       55      51
## 5  Midday     067   0   6     7         13       06      67
## 6 Evening     962   9   6     2         17       96      62
```

```
# save basic Lotto Table
write.table(lotto, file = 'C:/Users/crich/Desktop/Googz/lottoResearch/Data/cleanedLotto.csv', se
p = '\t', col.names = T)

# contains extra 3 columns (refer to chunk/cell above)
write.table(lottoExtended, file = 'C:/Users/crich/Desktop/Googz/lottoResearch/Data/cleanedLottoE
xtended.csv', sep = '\t', col.names = T)

# TABLEAU FRIENDLY
#remove first column, which is unnecessary for tableau
lottoExp = lotto[,c(2:10)]
write.table(lottoExp, file = 'C:/Users/crich/Desktop/Googz/lottoResearch/Data/cleanedLottoTable.
csv', sep = '\t', col.names = T)
```