Data Analytics Final Project - Data Cleaning

Sai Rakesh Ghanta (sag163@pitt.edu)

2016/11/17

Import the raw dataset and change the stationID to 1~50.

```
library(car)  # for recode
df <- read.csv("~/desktop/dataset.csv", header=FALSE, stringsAsFactors=
FALSE)
colnames(df) <- c("Stime","Syear","Smonth","Sday","week","Shour","Smin"
,"Etime","Eyear","Emonth","Eday","Ehour","Emin",
                  "Fromstation","Tostation")
df$Fromstation=recode(df$Fromstation,"'1000'=0;'1001'=1;'1002'=2;'1003'
=3;'1004'=4;'1005'=5;'1006'=6;'1007'=7;'1008'=8;'1009'=9;'1010'=10;'101
1'=11;'1012'=12;'1013'=13;'1014'=14;'1015'=15;'1016'=16;'1017'=17;'1018
'=18;'1019'=19;'1020'=20;'1021'=21;'1022'=22;'1023'=23;'1024'=24;'1025'
=25;'1026'=26;'1027'=27;'1028'=28;'1029'=29;'1030'=30;'1031'=31;'1032'=
32;'1033'=33;'1034'=34;'1035'=35;'1036'=36;'1037'=37;'1038'=38;'1039'=3
9;'1040'=40;'1041'=41;'1042'=42;'1043'=43;'1044'=44;'1045'=45;'1046'=46
;'1047'=47;'1048'=48;'1049'=49;'1050'=50;else=0")
df$Tostation=recode(df$Tostation,"'1000'=0;'1001'=1;'1002'=2;'1003'=3;'
1004'=4;'1005'=5;'1006'=6;'1007'=7;'1008'=8;'1009'=9;'1010'=10;'1011'=1
1;'1012'=12;'1013'=13;'1014'=14;'1015'=15;'1016'=16;'1017'=17;'1018'=18
;'1019'=19;'1020'=20;'1021'=21;'1022'=22;'1023'=23;'1024'=24;'1025'=25;
'1026'=26;'1027'=27;'1028'=28;'1029'=29;'1030'=30;'1031'=31;'1032'=32;'
1033'=33;'1034'=34;'1035'=35;'1036'=36;'1037'=37;'1038'=38;'1039'=39;'1
040'=40;'1041'=41;'1042'=42;'1043'=43;'1044'=44;'1045'=45;'1046'=46;'10
47'=47;'1048'=48;'1049'=49;'1050'=50;else=0")
df=df[-1,]
rownames(df)= c(1:nrow(df))
df[1:3,]
```

```
##      Stime Syear Smonth Sday week Shour Smin    Etime Eyear Emonth E
day
## 1 7/1/2015  2015      7    1    4     0   44 7/1/2015  2015      7
  1
## 2 7/1/2015  2015      7    1    4     5    4 7/1/2015  2015      7
  1
## 3 7/1/2015  2015      7    1    4     5    4 7/1/2015  2015      7
  1
##   Ehour Emin Fromstation Tostation
## 1     0   58           6         0
## 2     5   23          10        10
## 3     5   24          10        10
```

Count the row numbers base on the station, weekday and hours. Put the result in to the B (rent) and C (return) matrix. AA matrix = B - C (rent-return)

```
B <- matrix(1, nrow = 357, ncol = 24)
C <- matrix(1, nrow = 357, ncol = 24)

for (a in 0:50){ #51 station
  for (b in 0:23){
    for (c in 1:7){
        B[c+(a*7),b+1] <- nrow(df[df$Fromstation==a & df$week==c & df$S
hour==b,])
        C[c+(a*7),b+1] <- nrow(df[df$Tostation==a & df$week==c & df$Eho
ur==b,])
    }
  }
}

#B # from (rent)
#C # to (return)
AA= B-C
AA[1:3,] # from - to
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[,13]
## [1,]  -15   -9   -4    0   -1   -1   -2    0    4    -2    11     4
   -2
## [2,]   -3   -5    0    0   -2    0   -2    0  -15    -5    -8     7
  -30
## [3,]    1    0   -1    0   -1   -1   -2   -3   -3    -7    -4    -8
  -13
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,2
4]
## [1,]     4   -20   -40   -49   -36   -19   -28   -28   -16    -5   -
12
## [2,]    -3   -26   -29   -14   -30    -8   -10   -11   -12    -1
-7
## [3,]   -20    -4   -13    -9    -9   -14   -10   -13    -5    -4
-4
```

Use DD EE FF to record the result and export the result.

```
DD=c(1:8568)#empty row
EE=c(1:8568)
FF=c(1:8568)
#DD

for (d in 0:356) {
  DD[(1+(24*d)):(24+(24*d))] = B[d+1,] #rent
  EE[(1+(24*d)):(24+(24*d))] = C[d+1,] #return
  FF[(1+(24*d)):(24+(24*d))] = AA[d+1,] #rent - return
```

```
}

#DD
#EE
#FF
GG <- cbind(DD,EE,FF)
GG[1:3,]

##      DD EE  FF
## [1,]  0 15 -15
## [2,]  1 10  -9
## [3,]  0  4  -4

GG = as.data.frame(GG) #result matrix
#write.table(GG, "~/desktop/mydata.txt", sep="\t")
```

For pre-processing plot 4 (not specific weekday). We want to campare the difference bwtween weekday and weekend.

```
RR <- matrix(1, nrow = 51, ncol = 24)
TT <- matrix(1, nrow = 51, ncol = 24)

for (q in 0:50){ #51 station
  for (w in 0:23){
    RR[q+1,w+1] <- nrow(df[df$Fromstation==q & df$Shour==w,])
    TT[q+1,w+1] <- nrow(df[df$Tostation==q & df$Ehour==w,])
  }
}

RR[1:3,]

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[,13]
## [1,]   11    4    1    0    0    1    9   28   66    29   102   142
  181
## [2,]   19   21    3    0    1    6    7   16   27    50    91   157
  202
## [3,]   11    3    0    2    0    0    0    4    5    29    33    24
   56
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,2
4]
## [1,]   170   179   170   196   241   175   116    55    26    24
17
## [2,]   160   146   156   131   207   183   162    80    33    61
46
## [3,]    41    39    57    81    79    69    58    26    22    16
10

TT[1:3,]
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
[,13]
## [1,]   42   28   21    9    5   13   25   32   77    71    93   152
  265
## [2,]   26   14   12    2    0   31   37  119  109    73    64   134
  159
## [3,]   15    9    0    0    1    0    4   14   81    37    17    38
   49
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,2
4]
## [1,]   263   278   330   370   445   248   244   199   111    78
54
## [2,]   162   161   207   136   162   183   151   128    60    44
43
## [3,]    41    43    42    39    23    43    34    30    19    27
10

#RR-TT
#write.table(RR, "~/desktop/mydata123.txt", sep="\t")
#write.table(TT, "~/desktop/mydata12344.txt", sep="\t")
```