data\_cleaning

May 22, 2016

# Load data

filenames <- list.files("data/", pattern="\*.csv", full.names= TRUE)   
for (i in 1:length(filenames)){  
 # extract file name  
 name <- filenames[i]  
 name <- sub("\*.csv", "", name)  
 name <- sub(".\*//", "", name)  
   
 # read data  
 assign(name, read.csv(filenames[i], stringsAsFactors = F))  
}

## Warning in read.table(file = file, header = header, sep = sep, quote  
## = quote, : incomplete final line found by readTableHeader on 'data//  
## agency.csv'

# 

# check the Operations Dates within dataset  
unique(substr(stop\_times$trip\_id, 1, 12))

## [1] "A20151206SUN" "A20151206WKD" "A20151206SAT" "B20151206SAT"  
## [5] "B20151206SUN" "B20151206WKD" "R20150510SAT" "R20150510SUN"  
## [9] "R20150510WKD"

# data scheme  
str(stop\_times)

## 'data.frame': 533814 obs. of 9 variables:  
## $ trip\_id : chr "A20151206SUN\_069000\_GS.N03R" "A20151206SUN\_069000\_GS.N03R" "A20151206SUN\_069400\_GS.S03R" "A20151206SUN\_069400\_GS.S03R" ...  
## $ arrival\_time : chr "11:30:00" "11:31:30" "11:34:00" "11:35:30" ...  
## $ departure\_time : chr "11:30:00" "11:31:30" "11:34:00" "11:35:30" ...  
## $ stop\_id : chr "901N" "902N" "902S" "901S" ...  
## $ stop\_sequence : int 1 2 1 2 1 2 1 2 1 2 ...  
## $ stop\_headsign : logi NA NA NA NA NA NA ...  
## $ pickup\_type : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ drop\_off\_type : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ shape\_dist\_traveled: logi NA NA NA NA NA NA ...