Weather Data

> *It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet. We provide you with the dataset from NOAA of the weather conditions of 2007 to 2014, during the months of the tests. (From https://www.kaggle.com/c/predict-west-nile-virus/data)*

Read the weather data

```
weather.data <- read.csv("../data/weather.csv", na.strings=c("M", "-", "", " "))
dim(weather.data)
```

```
## [1] 2944   22
```

```
str(weather.data)
```

```
## 'data.frame':    2944 obs. of  22 variables:
##  $ Station    : int  1 2 1 2 1 2 1 2 1 2 ...
##  $ Date       : Factor w/ 1472 levels "2007-05-01","2007-05-02",..: 1 1 2 2 3 3 4 4 5 5 ...
##  $ Tmax       : int  83 84 59 60 66 67 66 78 66 66 ...
##  $ Tmin       : int  50 52 42 43 46 48 49 51 53 54 ...
##  $ Tavg       : int  67 68 51 52 56 58 58 NA 60 60 ...
##  $ Depart     : int  14 NA -3 NA 2 NA 4 NA 5 NA ...
##  $ DewPoint   : int  51 51 42 42 40 40 41 42 38 39 ...
##  $ WetBulb    : int  56 57 47 47 48 50 50 50 49 50 ...
##  $ Heat       : int  0 0 14 13 9 7 7 NA 5 5 ...
##  $ Cool       : int  2 3 0 0 0 0 0 NA 0 0 ...
##  $ Sunrise    : int  448 NA 447 NA 446 NA 444 NA 443 NA ...
##  $ Sunset     : int  1849 NA 1850 NA 1851 NA 1852 NA 1853 NA ...
##  $ CodeSum    : Factor w/ 97 levels "BCFG BR","BR",..: NA NA 2 3 NA 19 23 NA NA NA ...
##  $ Depth      : int  0 NA 0 NA 0 NA 0 NA 0 NA ...
##  $ Water1     : logi  NA NA NA NA NA NA ...
##  $ SnowFall   : Factor w/ 3 levels "0.0","0.1","  T": 1 NA 1 NA 1 NA 1 NA 1 NA ...
##  $ PrecipTotal: Factor w/ 167 levels "0.00","0.01",..: 1 1 1 1 1 1 167 1 167 167 ...
##  $ StnPressure: num  29.1 29.2 29.4 29.4 29.4 ...
##  $ SeaLevel   : num  29.8 29.8 30.1 30.1 30.1 ...
##  $ ResultSpeed: num  1.7 2.7 13 13.3 11.7 12.9 10.4 10.1 11.7 11.2 ...
##  $ ResultDir  : int  27 25 4 2 7 6 8 7 7 7 ...
##  $ AvgSpeed   : num  9.2 9.6 13.4 13.4 11.9 13.2 10.8 10.4 12 11.5 ...
```

```
head(weather.data)
```

```
##   Station       Date Tmax Tmin Tavg Depart DewPoint WetBulb Heat Cool
## 1       1 2007-05-01   83   50   67     14       51      56    0    2
## 2       2 2007-05-01   84   52   68     NA       51      57    0    3
## 3       1 2007-05-02   59   42   51     -3       42      47   14    0
## 4       2 2007-05-02   60   43   52     NA       42      47   13    0
## 5       1 2007-05-03   66   46   56      2       40      48    9    0
## 6       2 2007-05-03   67   48   58     NA       40      50    7    0
##   Sunrise Sunset CodeSum Depth Water1 SnowFall PrecipTotal StnPressure
## 1     448   1849    <NA>     0     NA      0.0        0.00       29.10
```

```
## 2      NA     NA    <NA>    NA     NA     <NA>        0.00       29.18
## 3     447   1850      BR     0     NA      0.0        0.00       29.38
## 4      NA     NA   BR HZ    NA     NA     <NA>        0.00       29.44
## 5     446   1851    <NA>     0     NA      0.0        0.00       29.39
## 6      NA     NA      HZ    NA     NA     <NA>        0.00       29.46
##   SeaLevel ResultSpeed ResultDir AvgSpeed
## 1    29.82         1.7        27      9.2
## 2    29.82         2.7        25      9.6
## 3    30.09        13.0         4     13.4
## 4    30.08        13.3         2     13.4
## 5    30.12        11.7         7     11.9
## 6    30.12        12.9         6     13.2
```

Notice of the weather data

From the file: QUALITY CONTROLLED LOCAL CLIMATOLOGICAL DATA

1. The dry bulb, dew point and wet bulb temperatures were originally reported to the nearest tenth of a degree Fahrenheit. The **Automated Surface Observing System (ASOS)** records temperatures and dew points in whole degrees Fahrenheit and converts these values to the nearest tenth of a degree Celsius for observation transmission. Until this date, these values online have incorrectly been converted back to the nearest tenth of a degree Fahrenheit, implying a level of precision that is not present at the instrument level.
2. Two stations.

- Whole degree Celsius temperature values for **AWOS** stations;
- Tenths degrees Celsius temperature values for **ASOS** stations.

Their location:

- Station 1: CHICAGO O'HARE INTERNATIONAL AIRPORT Lat: 41.995 Lon: -87.933 Elev: 662 ft. above sea level
- Station 2: CHICAGO MIDWAY INTL ARPT Lat: 41.786 Lon: -87.752 Elev: 612 ft. above sea level

From https://www.kaggle.com/c/predict-west-nile-virus/data

Note of some features

- **WetBulb**: Wet-bulb temperature is largely determined by both actual air temperature (dry-bulb temperature) and the amount of moisture in the air (humidity)

Data engineering

Separate the data set by station

```
weather.data.split <- split(weather.data, weather.data$Station)
weather.stn1 <- weather.data.split[[1]]
weather.stn2 <- weather.data.split[[2]]
dim(weather.stn1)
```

```
## [1] 1472   22
```

```
dim(weather.stn2)
```

```
## [1] 1472   22
```

Sunrise and Sunset

Only station 1 has such data, but they should be the same for the two stations

```
weather.stn2$Sunrise <- weather.stn1$Sunrise
weather.stn2$Sunset <- weather.stn1$Sunset
```

Depart

Depart means "DEPARTURE FROM NORMAL".

```
summary(weather.stn1$Depart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -17.000  -3.000   2.000   1.954   7.000  23.000
```

```
summary(weather.stn2$Depart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA    1472
```

```
normal.tmp <- weather.stn1$Tavg - weather.stn1$Depart
weather.stn2$Depart <- weather.stn2$Tavg - normal.tmp
summary(weather.stn1$Depart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -17.000  -3.000   2.000   1.954   7.000  23.000
```

```
summary(weather.stn2$Depart)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -17.000  -2.000   3.000   3.207   8.000  25.000      11
```

CodeSum

First find out all the unique code for different weather.

```
code.A <- unique(unlist(strsplit(unique(as.character(weather.stn1$CodeSum)), " ")))
code.B <- unique(unlist(strsplit(unique(as.character(weather.stn2$CodeSum)), " ")))
code <- union(code.A, code.B)[-1]
code
```

```
##  [1] "BR"   "RA"   "HZ"   "VCTS" "TSRA" "FU"   "DZ"   "TS"   "FG+"  "BCFG"
## [11] "MIFG" "FG"   "SQ"   "SN"   "VCFG" "GR"
```

```
code.name <- paste(rep("Code.", length(code)), code, sep="")
code.name
```

```
##  [1] "Code.BR"   "Code.RA"   "Code.HZ"   "Code.VCTS" "Code.TSRA"
##  [6] "Code.FU"   "Code.DZ"   "Code.TS"   "Code.FG+"  "Code.BCFG"
## [11] "Code.MIFG" "Code.FG"   "Code.SQ"   "Code.SN"   "Code.VCFG"
## [16] "Code.GR"
```

Then add new columns indicating distinct weather code in the data frames. Observe all levels in CodeSum, "FG", "TS" and "RA" are the tricky ones, since these letter pairs appear in more than one code.

```
code
```

```
##  [1] "BR"   "RA"   "HZ"   "VCTS" "TSRA" "FU"   "DZ"   "TS"   "FG+"  "BCFG"
## [11] "MIFG" "FG"   "SQ"   "SN"   "VCFG" "GR"
```

```
## rewrite the reguar expression of "FG", "TS" and "RA" to avoid wrong matching
code[2] <- "^RA | RA$| RA |^RA$"
code[8] <- "^TS | TS$| TS |^TS$"
code[12] <- "^FG | FG$| FG |^FG$"
```

```
for(i in 1:length(code)) {
     new.code <- code[i]
    new.code.name <- code.name[i]
    weather.stn1[, new.code.name] <- grepl(new.code, weather.stn1$CodeSum)
    weather.stn2[, new.code.name] <- grepl(new.code, weather.stn2$CodeSum)
}
## show the resultant features
head(weather.stn1[,c(13, 23:38)])
```

```
##    CodeSum Code.BR Code.RA Code.HZ Code.VCTS Code.TSRA Code.FU Code.DZ
## 1     <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 3       BR    TRUE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 5     <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 7       RA   FALSE    TRUE   FALSE     FALSE     FALSE   FALSE   FALSE
## 9     <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 11    <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
##    Code.TS Code.FG+ Code.BCFG Code.MIFG Code.FG Code.SQ Code.SN Code.VCFG
## 1    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 3    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 5    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 7    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 9    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 11   FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
##    Code.GR
## 1    FALSE
## 3    FALSE
## 5    FALSE
## 7    FALSE
## 9    FALSE
## 11   FALSE
```

```
head(weather.stn2[,c(13, 23:38)])
```

```
##    CodeSum Code.BR Code.RA Code.HZ Code.VCTS Code.TSRA Code.FU Code.DZ
## 2     <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 4    BR HZ    TRUE   FALSE    TRUE     FALSE     FALSE   FALSE   FALSE
## 6       HZ   FALSE   FALSE    TRUE     FALSE     FALSE   FALSE   FALSE
## 8     <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 10    <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
## 12    <NA>   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE
##    Code.TS Code.FG+ Code.BCFG Code.MIFG Code.FG Code.SQ Code.SN Code.VCFG
## 2    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 4    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 6    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 8    FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 10   FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
## 12   FALSE    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE
##    Code.GR
## 2    FALSE
## 4    FALSE
## 6    FALSE
## 8    FALSE
## 10   FALSE
## 12   FALSE
```

Water1

```
summary(weather.stn1$Water1)
```

```
##    Mode    NA's
## logical    1472
```

```
summary(weather.stn2$Water1)
```

```
##    Mode    NA's
## logical    1472
```

This Water1 is useless, remove it from data frames.

```
weather.stn1 <- weather.stn1[,-15]
weather.stn2 <- weather.stn2[,-15]
```

SnowFall

SnowFall is not a good predictor, since the time is from May to October.

```
summary(weather.stn1["SnowFall"])
```

```
##  SnowFall
##  0.0:1459
##  0.1:   1
##    T:  12
```

```
summary(weather.stn2["SnowFall"])
```

```
##  SnowFall
##  0.0 :   0
##  0.1 :   0
##    T :   0
##  NA's:1472
```

```
-which(names(weather.stn1)=="Depth")
```

```
## [1] -14
```

```
weather.stn1 <- weather.stn1[,-which(names(weather.stn1)=="SnowFall")]
weather.stn2 <- weather.stn2[,-which(names(weather.stn2)=="SnowFall")]
```

Date

```
weather.stn1$Date <- as.Date(weather.stn1$Date)
weather.stn2$Date <- as.Date(weather.stn2$Date)
```

PrecipTotal

```
weather.stn1$PrecipTotal <- as.numeric(weather.stn1$PrecipTotal)
weather.stn2$PrecipTotal <- as.numeric(weather.stn2$PrecipTotal)
```

Depth

Remove this feature, since there are all NAs.

```
weather.stn1 <- weather.stn1[,-which(names(weather.stn1)=="Depth")]
weather.stn2 <- weather.stn2[,-which(names(weather.stn2)=="Depth")]
```

Missing values

Fill all the missing values by the means of surrounding values.

```
library("zoo")
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Tavg
weather.stn1$Tavg <- (na.locf(weather.stn1$Tavg) + rev(na.locf(rev(weather.stn1$Tavg))))/2
weather.stn2$Tavg <- (na.locf(weather.stn2$Tavg) + rev(na.locf(rev(weather.stn2$Tavg))))/2
## Depart
weather.stn1$Depart <- (na.locf(weather.stn1$Depart) + rev(na.locf(rev(weather.stn1$Depart))))/2
weather.stn2$Depart <- (na.locf(weather.stn2$Depart) + rev(na.locf(rev(weather.stn2$Depart))))/2
## WetBulb
weather.stn1$WetBulb <- (na.locf(weather.stn1$WetBulb) + rev(na.locf(rev(weather.stn1$WetBulb))))/2
weather.stn2$WetBulb <- (na.locf(weather.stn2$WetBulb) + rev(na.locf(rev(weather.stn2$WetBulb))))/2
## Heat
weather.stn1$Heat <- (na.locf(weather.stn1$Heat) + rev(na.locf(rev(weather.stn1$Heat))))/2
weather.stn2$Heat <- (na.locf(weather.stn2$Heat) + rev(na.locf(rev(weather.stn2$Heat))))/2
## Cool
weather.stn1$Cool <- (na.locf(weather.stn1$Cool) + rev(na.locf(rev(weather.stn1$Cool))))/2
weather.stn2$Cool <- (na.locf(weather.stn2$Cool) + rev(na.locf(rev(weather.stn2$Cool))))/2
## PrecipTotal
weather.stn1$PrecipTotal <- (na.locf(weather.stn1$PrecipTotal) + rev(na.locf(rev(weather.stn1$PrecipTota
weather.stn2$PrecipTotal <- (na.locf(weather.stn2$PrecipTotal) + rev(na.locf(rev(weather.stn2$PrecipTota
## StnPressure
weather.stn1$StnPressure <- (na.locf(weather.stn1$StnPressure) + rev(na.locf(rev(weather.stn1$StnPressu
weather.stn2$StnPressure <- (na.locf(weather.stn2$StnPressure) + rev(na.locf(rev(weather.stn2$StnPressu
## SeaLevel
weather.stn1$SeaLevel <- (na.locf(weather.stn1$SeaLevel) + rev(na.locf(rev(weather.stn1$SeaLevel))))/2
weather.stn2$SeaLevel <- (na.locf(weather.stn2$SeaLevel) + rev(na.locf(rev(weather.stn2$SeaLevel))))/2
## AvgSpeed
weather.stn1$AvgSpeed <- (na.locf(weather.stn1$AvgSpeed) + rev(na.locf(rev(weather.stn1$AvgSpeed))))/2
weather.stn2$AvgSpeed <- (na.locf(weather.stn2$AvgSpeed) + rev(na.locf(rev(weather.stn2$AvgSpeed))))/2
```

Combined Main and Weather Data Set

Merge train/test and weather.stn1/weather.stn2

Each row in main data set is merged to the weather record by the closer station.

```
## load the train and test sets first
train <- read.csv("../data/train2B.csv")
test <- read.csv("../data/test2B.csv")
## transform the data format
train$Date <- as.Date(train$Date)
train$Month <- factor(train$Month,
                      levels=c("May", "June", "July", "August", "September", "October"))
train$Weekday <- factor(train$Weekday,
                        levels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
test$Date <- as.Date(test$Date)
test$Month <- factor(test$Month,
                      levels=c("June", "July", "August", "September", "October"))
test$Weekday <- factor(test$Weekday,
                        levels=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday"))
```

```
## combine the weather.stn1 and weather.stn2 to one data frame first
weather.stn <- rbind(weather.stn1, weather.stn2)
```

```
## merge
train <- merge(train, weather.stn, by.x=c("Date", "ClosestStn"), by.y=c("Date", "Station"))
test <- merge(test, weather.stn, by.x=c("Date", "ClosestStn"), by.y=c("Date", "Station"))
test <- test[order(test$Id), ]
## show the new data sets
str(train, strict.width="cut")
```

```
## 'data.frame':    8475 obs. of  54 variables:
##  $ Date                   : Date, format: "2007-05-29" "2007-05-29" ...
##  $ ClosestStn             : int  1 1 1 1 1 1 1 2 2 2 ...
##  $ Address                : Factor w/ 138 levels "1000 East 67th Street, "..
##  $ Species                : Factor w/ 7 levels "CULEX ERRATICUS",..: 3 4 4..
##  $ Block                  : int  41 41 62 79 79 75 65 25 11 15 ...
##  $ Street                 : Factor w/ 128 levels " E 105TH ST",..: 33 33 2..
##  $ Trap                   : Factor w/ 136 levels "T001","T002",..: 2 2 7 1..
##  $ AddressNumberAndStreet: Factor w/ 138 levels "1000  E 67TH ST, Chicag"..
##  $ Latitude               : num  42 42 42 42 42 ...
##  $ Longitude              : num  -87.8 -87.8 -87.8 -87.8 -87.8 ...
##  $ AddressAccuracy        : int  9 9 9 8 8 8 8 8 8 8 ...
##  $ Year                   : int  2007 2007 2007 2007 2007 2007 2007 2007 2..
##  $ Month                  : Factor w/ 6 levels "May","June","July",..: 1 1..
##  $ Week                   : int  22 22 22 22 22 22 22 22 22 22 ...
##  $ Weekday                : Factor w/ 5 levels "Monday","Tuesday",..: 2 2 ..
##  $ NumMosquitos           : int  1 1 1 1 4 1 1 1 1 2 ...
##  $ WnvPresent             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ TrapNumber             : int  2 2 7 15 15 148 143 46 48 45 ...
##  $ TrapMS                 : Factor w/ 3 levels "B","C","M": 3 3 3 3 3 3 3 ..
##  $ DisStn1                : num  11.8 11.8 13.53 9.24 9.24 ...
##  $ DisStn2                : num  19.2 19.2 23.3 21.8 21.8 ...
##  $ Tmax                   : int  88 88 88 88 88 88 88 88 88 88 ...
##  $ Tmin                   : int  60 60 60 60 60 60 60 65 65 65 ...
##  $ Tavg                   : num  74 74 74 74 74 74 74 77 77 77 ...
##  $ Depart                 : num  10 10 10 10 10 10 10 13 13 13 ...
##  $ DewPoint               : int  58 58 58 58 58 58 58 59 59 59 ...
##  $ WetBulb                : num  65 65 65 65 65 65 65 66 66 66 ...
##  $ Heat                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Cool                   : num  9 9 9 9 9 9 9 12 12 12 ...
##  $ Sunrise                : int  421 421 421 421 421 421 421 421 421 421 ...
##  $ Sunset                 : int  1917 1917 1917 1917 1917 1917 1917 1917 1..
##  $ CodeSum                : Factor w/ 97 levels "BCFG BR","BR",..: 3 3 3 3..
##  $ PrecipTotal            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ StnPressure            : num  29.4 29.4 29.4 29.4 29.4 ...
##  $ SeaLevel               : num  30.1 30.1 30.1 30.1 30.1 ...
##  $ ResultSpeed            : num  5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.8 5.8 ...
##  $ ResultDir              : int  18 18 18 18 18 18 18 16 16 16 ...
##  $ AvgSpeed               : num  6.5 6.5 6.5 6.5 6.5 6.5 6.5 7.4 7.4 7.4 ...
##  $ Code.BR                : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Code.RA                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.HZ                : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Code.VCTS              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.TSRA              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.FU                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.DZ                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
##  $ Code.TS               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.FG+              : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.BCFG             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.MIFG             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.FG               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.SQ               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.SN               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.VCFG             : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Code.GR               : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
head(test)
```

```
##         Date ClosestStn Id
## 1 2008-06-11          1  1
## 2 2008-06-11          1  2
## 3 2008-06-11          1  3
## 4 2008-06-11          1  4
## 5 2008-06-11          1  5
## 6 2008-06-11          1  6
##                                          Address
## 1 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 2 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 3 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 4 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 5 4100 North Oak Park Avenue, Chicago, IL 60634, USA
## 6 4100 North Oak Park Avenue, Chicago, IL 60634, USA
##                  Species Block          Street Trap
## 1 CULEX PIPIENS/RESTUANS    41  N OAK PARK AVE T002
## 2         CULEX RESTUANS    41  N OAK PARK AVE T002
## 3          CULEX PIPIENS    41  N OAK PARK AVE T002
## 4       CULEX SALINARIUS    41  N OAK PARK AVE T002
## 5        CULEX TERRITANS    41  N OAK PARK AVE T002
## 6          CULEX TARSALIS    41  N OAK PARK AVE T002
##              AddressNumberAndStreet Latitude Longitude AddressAccuracy
## 1 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
## 2 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
## 3 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
## 4 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
## 5 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
## 6 4100  N OAK PARK AVE, Chicago, IL 41.95469 -87.80099               9
##   Year Month Week   Weekday NumMosquitos TrapNumber TrapMS  DisStn1
## 1 2008  June   24 Wednesday      7.36107          2      M 11.79739
## 2 2008  June   24 Wednesday      7.36107          2      M 11.79739
## 3 2008  June   24 Wednesday      7.36107          2      M 11.79739
## 4 2008  June   24 Wednesday      7.36107          2      M 11.79739
## 5 2008  June   24 Wednesday      7.36107          2      M 11.79739
## 6 2008  June   24 Wednesday      7.36107          2      M 11.79739
##   DisStn2 Tmax Tmin Tavg Depart DewPoint WetBulb Heat Cool Sunrise Sunset
## 1 19.1911   86   61   74      7       56      64    0    9     416   1926
## 2 19.1911   86   61   74      7       56      64    0    9     416   1926
## 3 19.1911   86   61   74      7       56      64    0    9     416   1926
## 4 19.1911   86   61   74      7       56      64    0    9     416   1926
## 5 19.1911   86   61   74      7       56      64    0    9     416   1926
## 6 19.1911   86   61   74      7       56      64    0    9     416   1926
```

```
##   CodeSum PrecipTotal StnPressure SeaLevel ResultSpeed ResultDir AvgSpeed
## 1    <NA>           1       29.28    29.99         8.9        18       10
## 2    <NA>           1       29.28    29.99         8.9        18       10
## 3    <NA>           1       29.28    29.99         8.9        18       10
## 4    <NA>           1       29.28    29.99         8.9        18       10
## 5    <NA>           1       29.28    29.99         8.9        18       10
## 6    <NA>           1       29.28    29.99         8.9        18       10
##   Code.BR Code.RA Code.HZ Code.VCTS Code.TSRA Code.FU Code.DZ Code.TS
## 1   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
## 2   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
## 3   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
## 4   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
## 5   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
## 6   FALSE   FALSE   FALSE     FALSE     FALSE   FALSE   FALSE   FALSE
##   Code.FG+ Code.BCFG Code.MIFG Code.FG Code.SQ Code.SN Code.VCFG Code.GR
## 1    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
## 2    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
## 3    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
## 4    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
## 5    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
## 6    FALSE     FALSE     FALSE   FALSE   FALSE   FALSE     FALSE   FALSE
```

Save this data set

```
write.csv(train, "../data/train4B.csv", row.names=F)
write.csv(test, "../data/test4B.csv", row.names=F)
```