A little sqldf demonstration

David L. Reiner
XR Trading, Chicago

2013-05-01

```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
…
> require('sqldf')
…
> # Read data file from NOAA
> fname <- "CDO1608826306257.txt"
> classes <- c(rep("NULL",2),"integer",rep(c("numeric","NULL"),6),rep("numeric",2),
+   rep("character",3),"numeric","integer","NULL")
> Weather <- read.csv(fname, colClasses=classes)
> # Clean it up
> Weather$Year <- as.integer(Weather$YEARMODA / 10000)
> Weather$Month <- as.integer((Weather$YEARMODA %% 10000) / 100)
> Weather$Day <- as.integer(Weather$YEARMODA %% 100)
> Weather$DEWP[Weather$DEWP > 999] <- NA
> Weather$SLP[Weather$SLP > 999] <- NA
> Weather$STP[Weather$STP > 999] <- NA
> Weather$MXSPD[Weather$MXSPD > 999] <- NA
> Weather$GUST[Weather$GUST > 999] <- NA
> Weather$PrecipFlag <- substr(Weather$PRCP, 6, 7)
> Weather$PRCP <- as.numeric(substr(Weather$PRCP, 1, 5))
> Weather$PRCP[Weather$PRCP > 99] <- NA # NB: just 99
> Weather$PRCP[is.na(Weather$PRCP)] <- 0
> Weather$SNDP[Weather$SNDP > 999] <- 0
> Weather$MAX <- as.numeric(gsub("\\*","",gsub(" ","",Weather$MAX)))
> Weather$MAX[Weather$MAX > 999] <- NA
> Weather$MIN <- as.numeric(gsub("\\*","",gsub(" ","",Weather$MIN)))
> Weather$MIN[Weather$MIN > 999] <- NA
> head(Weather,2)
  YEARMODA TEMP DEWP SLP   STP VISIB WDSP MXSPD GUST  MAX  MIN PRCP SNDP FRSHTT Year Month Day PrecipFlag
1 19461009 70.5 46.1  NA 988.4   6.6  9.3  13.0   NA 74.3 64.4    0    0      0 1946    10   9          I
2 19461010 64.9 52.4  NA 984.0   4.9 12.1  16.9   NA 80.2 53.4    0    0      0 1946    10  10          I
```

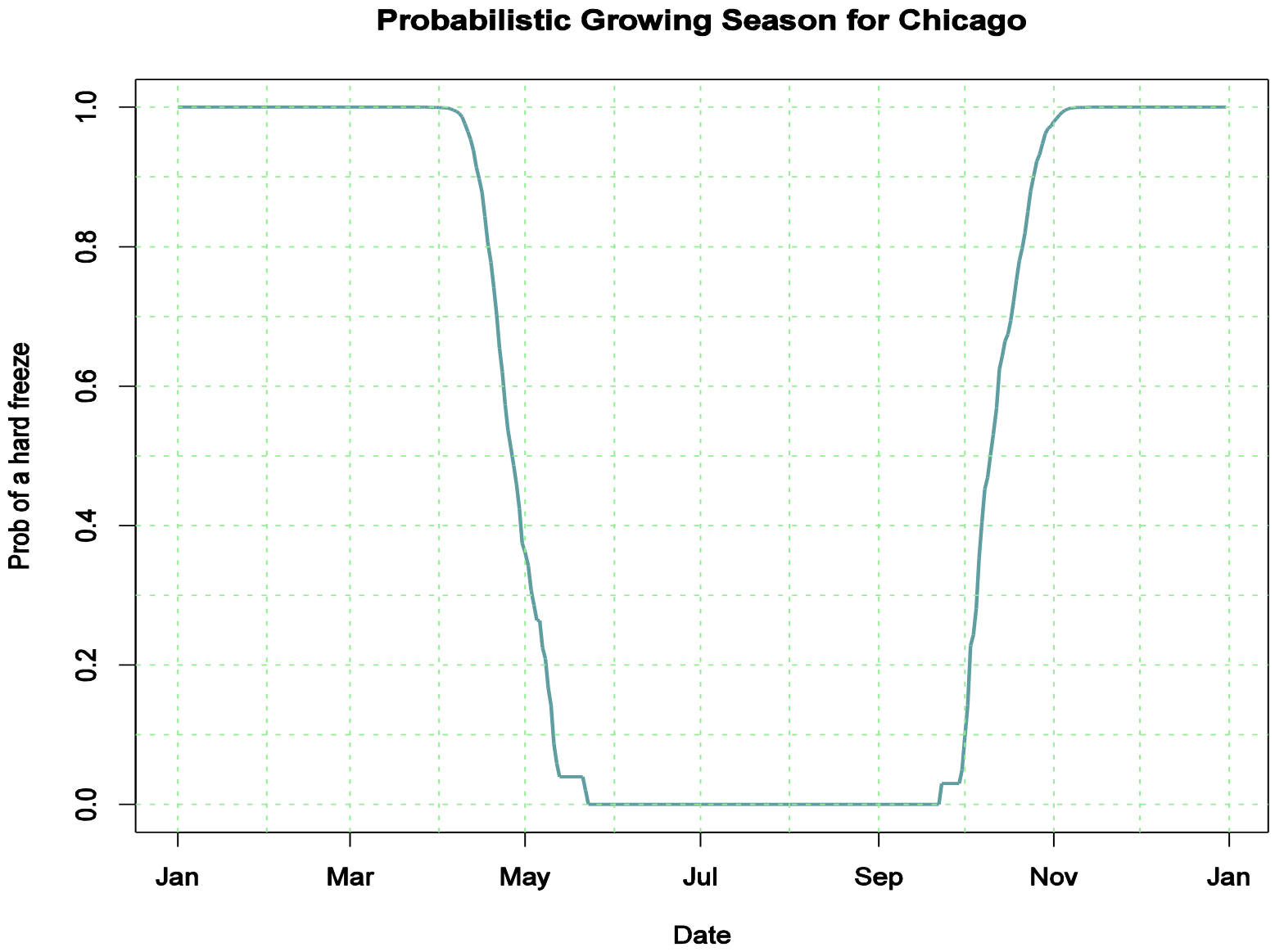
```
> # Look at some weather queries
> sqldf("select * from Weather where MAX > 100")
Loading required package: tcltk
   YEARMODA TEMP DEWP SLP   STP VISIB WDSP MXSPD GUST   MAX  MIN PRCP SNDP FRSHTT Year Month Day PrecipFlag
…
14 19950713 90.7 75.9  NA    NA   7.5  9.7  14.0   NA 102.9 81.0 0.00    0      0 1995     7  13          G
15 20050724 85.9 71.0  NA 989.4   8.7 10.9  17.1 22.9 102.0 64.0 0.06    0      0 2005     7  24          G
16 20120704 89.8 69.0  NA 986.8  10.0  8.4  15.9 22.0 102.0 77.0 0.00    0      0 2012     7   4          G
17 20120705 89.5 67.8  NA 988.4  10.0  6.5  28.9 45.1 102.9 79.0 0.00    0  10010 2012     7   5          G
18 20120706 92.3 72.0  NA 989.3  10.0  6.5  13.0 19.0 102.9 79.0 0.28    0      0 2012     7   6          G
19 20120707 86.9 69.4  NA 990.1  10.0  8.7  15.9 21.0 102.9 75.0 0.00    0      0 2012     7   7          G

> sqldf("select * from Weather where MIN < -20")
   YEARMODA  TEMP  DEWP SLP   STP VISIB WDSP MXSPD GUST   MAX   MIN PRCP SNDP FRSHTT Year Month Day PrecipFlag
1  19510130 -10.7 -23.0  NA    NA  13.0  5.5  11.1   NA   1.4 -21.6 0.00  0.0      0 1951     1  30          I
2  19820110 -19.8 -32.6  NA 997.2   8.8 20.1  24.1 37.9 -14.1 -26.0 0.00  0.8   1000 1982     1  10          D
3  19820111  -2.1 -12.6  NA 990.3   9.7 16.6  30.1 42.0   3.9 -26.0 0.08  1.2   1000 1982     1  11          G
4  19820117 -16.6 -26.1  NA    NA   8.5 10.5  19.0   NA   8.1 -25.1 0.04  4.7 101000 1982     1  17          G
5  19831224 -17.8 -30.6  NA    NA  12.1 18.8  25.1 30.9  -6.0 -25.1 0.00  3.9      0 1983    12  24          G
6  19831225 -10.8 -23.4  NA    NA  14.8 16.1  22.0 26.8  -5.1 -25.1 0.00  3.9      0 1983    12  25          D
7  19840121 -11.1 -23.5  NA    NA  14.9 11.2  21.0 25.8   6.1 -22.0 0.00  3.9      0 1984     1  21          G
8  19850120 -19.9 -33.0  NA 995.9  13.3 16.1  22.0 27.0 -11.0 -27.0 0.00  7.1      0 1985     1  20          G
9  19940118 -13.7 -24.6  NA    NA  11.0 15.2  20.0   NA  10.0 -20.9 0.00  3.9      0 1994     1  18          G
10 19940119 -11.3 -19.2  NA    NA   7.5 10.3  15.0   NA   3.0 -20.9 0.01  3.9   1000 1994     1  19          G

> sqldf("select * from Weather where SNDP > 25")
  YEARMODA TEMP DEWP   SLP   STP VISIB WDSP MXSPD GUST  MAX  MIN PRCP SNDP FRSHTT Year Month Day PrecipFlag
1 19790114 11.6  5.2    NA 986.7   8.7 17.1  25.1 31.9 30.9 -2.9 1.06 28.0 101000 1979     1  14          G
2 19790124 26.9 24.2 996.2 971.5   1.6 13.5  22.0 26.8 35.1 19.9 0.55 26.8 101000 1979     1  24          G
3 19790125 13.3  3.5    NA 986.0  13.3 11.4  18.1 21.0 34.0  5.0 0.16 26.8   1000 1979     1  25          G
4 19820122 21.0 15.0    NA 999.0   8.7 16.9  20.0 22.0 28.0 16.9 0.24 26.8 111000 1982     1  22          G

> sqldf("select min(MIN), max(MAX) from Weather")
  min(MIN) max(MAX)
1      -27      104
```

```
> sqldf("select Month, count(*) from Weather where MAX >  90 group by Month")
  Month count(*)
1     4        1
2     5       42
3     6      249
4     7      395
5     8      283
6     9       82
7    10        1
> sqldf("select Month, count(*) from Weather where MIN < -10 group by Month")
  Month count(*)
1     1       70
2     2       20
3    12       28
> sqldf("select Month, count(*) from Weather where SNDP >  0 group by Month")
  Month count(*)
1     1      922
2     2      706
3     3      318
4     4       38
5    10       20
6    11       93
7    12      542
> sqldf("select Month, count(*) from Weather where PRCP >  1 and SNDP = 0 group by Month")
   Month count(*)
1     1        5
2     2        6
3     3       11
4     4       29
5     5       38
6     6       41
7     7       43
8     8       60
9     9       45
10   10       32
11   11       25
12   12        6
```

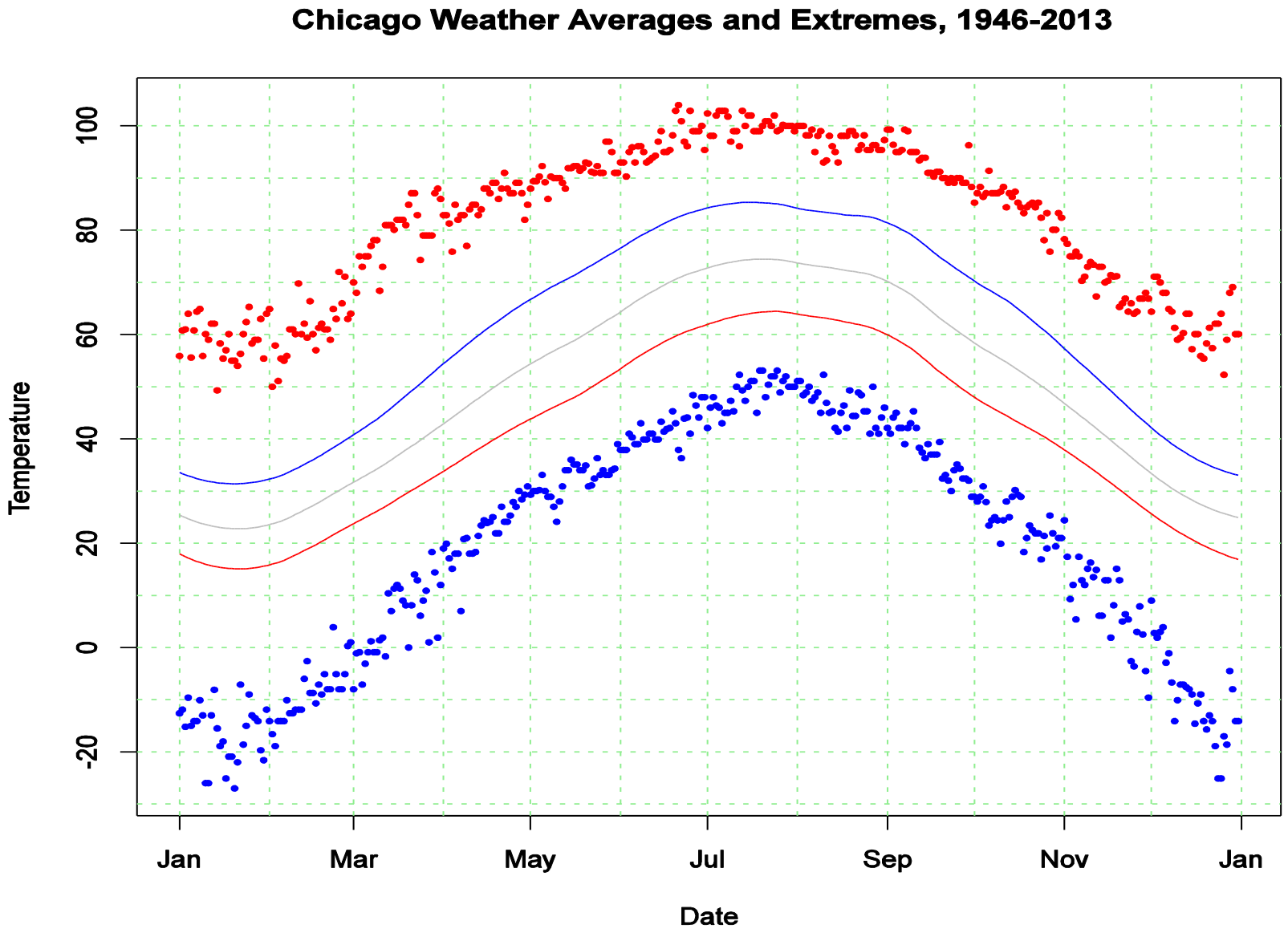
```
> # More complicated queries
> # Form three-day sequences of highs
> DFmax3 <- sqldf("select Weather.YEARMODA, Weather.MAX, BB.lag, BB.lag2 from Weather
+           left join
+           (
+             select A.rowid, A.YEARMODA, A.MAX, B.MAX as lag, C.MAX as lag2
+             from Weather as A join Weather as B join Weather as C
+             where A.rowid-1 = B.rowid
+             and B.rowid-1 = C.rowid
+             order by A.rowid
+           ) as BB
+           on Weather.rowid=BB.rowid ")

> # Were there three days in a row with high >= 99?
> # Note that we must use the fn$ prefix to invoke quasi-perl-style string interpolation functionality.
> tem <- 99
> fn$sqldf("select YEARMODA,lag2,lag,MAX from DFmax3 where MAX>=$tem and lag>=$tem and lag2>=$tem")
  YEARMODA  lag2   lag   MAX
1 19880622 102.9 104.0 100.9
2 19880716  99.0 102.0 102.0
3 19880717 102.0 102.0  99.0
4 19880803 100.0 100.0 100.0
5 20060802  99.0  99.0  99.0
6 20120706 102.0 102.9 102.9
7 20120707 102.9 102.9 102.9
```

```
> # Form three-day sequences of lows
> DFmin3 <- sqldf("select Weather.YEARMODA, Weather.MIN, BB.lag, BB.lag2 from Weather
+           left join
+           (
+             select A.rowid, A.YEARMODA, A.MIN, B.MIN as lag, C.MIN as lag2
+             from Weather as A join Weather as B join Weather as C
+             where A.rowid-1 = B.rowid
+             and B.rowid-1 = C.rowid
+             order by A.rowid
+           ) as BB
+           on Weather.rowid=BB.rowid ")
>
> # What were the 5 coldest stretches of 3 days?
> sqldf("select YEARMODA,lag2,lag,MIN from DFmin3 where MIN+lag+lag2<0 order by MIN+lag+lag2 limit 5")
  YEARMODA  lag2   lag   MIN
1 19831225 -18.9 -25.1 -25.1
2 19831226 -25.1 -25.1 -17.0
3 19820111 -13.0 -26.0 -26.0
4 19940120 -20.9 -20.9 -18.9
5 19940119 -16.1 -20.9 -20.9
```

```
> # What is the probability of freezing weather for each date?
> Frz <- sqldf("select A.Month, A.Day, round((0.0+freezing)/(0.0+total),2) PctFrz
+         from
+         (select Month, Day, count(*) total from Weather
+           group by Month, Day order by Month, Day) as A
+         join
+         (select Month, Day, count(*) freezing from Weather
+           where MIN < 32
+           group by Month, Day order by Month, Day) as B
+         on A.Month=B.Month and A.Day=B.Day
+         group by A.Month, A.Day
+         order by A.Month, A.Day")

> dts <- seq(from=as.Date("2012-01-01"), to=as.Date("2012-12-31"), by="1 day")
> Dts <- data.frame(Date=dts)
> Dts$Month <- 1+as.POSIXlt(dts)$mon
> Dts$Day <- as.POSIXlt(dts)$mday

> frz <- sqldf("select Date,PctFrz from Dts left join Frz on Dts.Month=Frz.Month and Dts.Day=Frz.Day")
> frz[is.na(frz$PctFrz),2] <- 0
> frz$CumFrz[250:366] <- 1-cumprod(1-frz$PctFrz[250:366])
> frz$CumFrz[249:1] <- 1-cumprod(1-frz$PctFrz[249:1])

> plot(frz$Date,frz$CumFrz, type='l', xlab="Date", ylab="Prob of a hard freeze", col="cadetblue", lwd=2)
> title("Probabilistic Growing Season for Chicago")
> abline(h=seq(0,1,.1), col='lightgreen', lty=2)
> abline(v=seq(as.Date("2012-01-01"), as.Date("2013-01-01"), by="1 month"), col='lightgreen', lty=2)
```


Probabilistic Growing Season for Chicago
Prob of a hard freeze
0.0
0.2
0.4
0.6
0.8
1.0
Jan
Mar
May
Jul
Sep
Nov
Jan
Date

```
> # Build a data frame of averages and extremes by date
> # Could be done in one sql statement
> extremes <- sqldf("select Month, Day, min(MIN), max(MAX)
+                   from Weather group by Month, Day order by Month, Day")
> means    <- sqldf("select Month, Day, avg(MIN), avg(MAX) from Weather
+                   group by Month, Day order by Month, Day")
> means[,5] <- sqldf("select avg(TEMP) from Weather group by Month, Day order by Month, Day")
>
> temps <- data.frame(Date=dts, RecLo=extremes[,3], RecHi=extremes[,4],
+                     AvgLo=means[,3], AvgHi=means[,4], Avg=means[,5])
>
> with(temps, plot(Date, RecLo, ylim=c(range(extremes[,3:4])), type="n", ylab="Temperature"))
> abline(h=seq(-30,110,10), col='lightgreen', lty=2)
> abline(v=seq(as.Date("2012-01-01"), as.Date("2013-01-01"), by="1 month"), col='lightgreen', lty=2)
> with(temps, points(Date, RecLo, pch=19, col='blue', cex=0.5))
> with(temps, points(Date, RecHi, pch=19, col='red', cex=0.5))
> loe <- with(temps, loess(AvgLo ~ as.numeric(Date), span=0.2))
> lines(temps$Date, loe$fitted, col='red')
> loe <- with(temps, loess(AvgHi ~ as.numeric(Date), span=0.2))
> lines(temps$Date, loe$fitted, col='blue')
> loe <- with(temps, loess(Avg ~ as.numeric(Date), span=0.2))
> lines(temps$Date, loe$fitted, col='gray')
> title("Chicago Weather Averages and Extremes, 1946-2013")
```


Chicago Weather Averages and Extremes, 1946-2013
Temperature
100
80
60
40
20
0
-20
Jan
Mar
May
Jul
Sep
Nov
Jan
Date