Autocorrelation

Kinga

Wednesday, March 11, 2015

Definition of Autocorrelation

Given a time series $x_1, x_2, \ldots, x_t, \ldots, x_T$, where x_t denotes the observation at time t, the time series is said to show autocorrelation if there is correlation between the lagged values of the time series. The autocorrelation coefficient r_1 is the correlation coefficient of lag 1, i.e. it measures the linear relationship between x_t and x_{t-1} for all time t. In general, the autocorrelation coefficient r_i is the correlation coeffecient of the time series lagged by i, that is it measures the linear relationship between x_t and x_{t-i}.

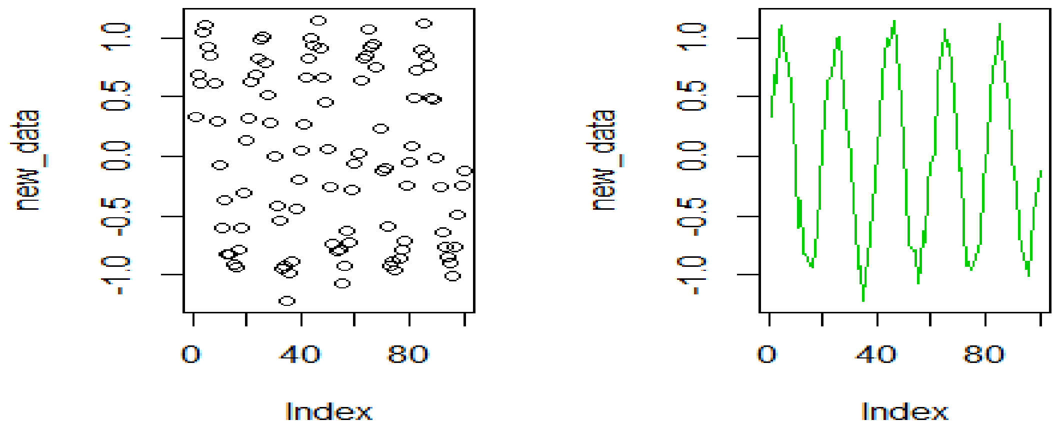
Example of an Auto Correlated Time Series

Let's create a periodic time series data set with some errors thrown in:

```
aSeq <- sin(seq(0.1*pi, 10*pi, 0.1*pi))
# generating a list of 100 random numbers with mean 0 and standard deviation
of 0.1
err <- rnorm(length(aSeq),0, 0.1)  #note: length(aSeq)=100
new_data <- aSeq + err
```

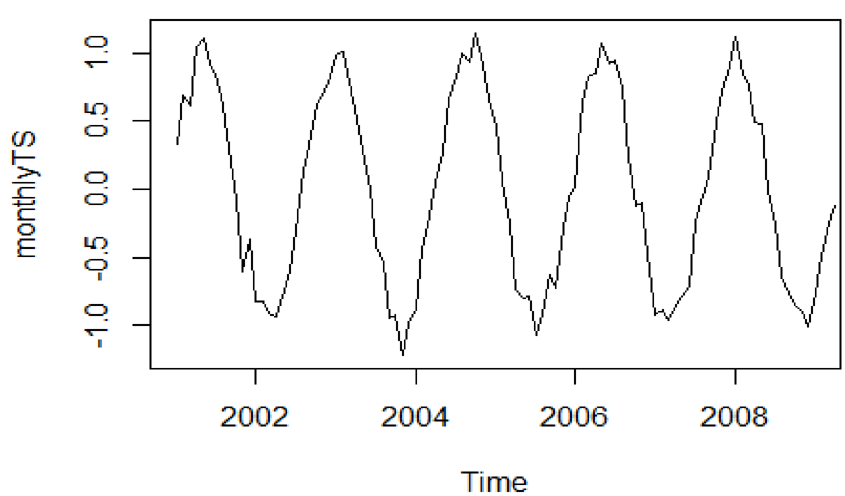
Now let's look at this newly created dataset:

```
par(mfrow=c(1,2))  #arranging the two graphs in one row
plot(new_data)
plot(new_data, type="l", col="3")
```



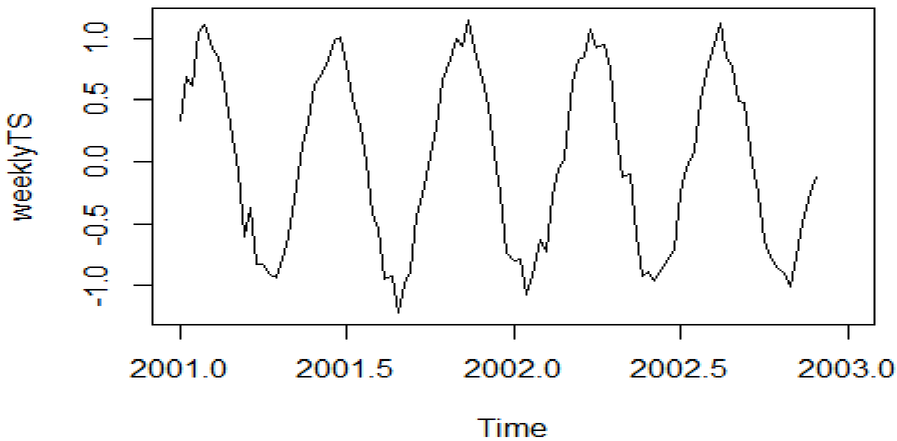
It does look periodic peppered with some errors, so it should work for our purpose. To make it officially a time series, lets apply the *ts*() function to it:

```
#creating a monthly time series, beginning at January 2001
monthlyTS<- ts(new_data, freq = 12, start=c(2001,1))
plot(monthlyTS, xlim = c(2001, 2009))
```




```
monthlyTS

##                Jan          Feb          Mar          Apr          May
## 2001  0.333003457  0.691087298  0.620092363  1.048182189  1.104222087
## 2002 -0.828513613 -0.827707623 -0.912583404 -0.938572141 -0.790060789
## 2003  0.977625001  1.004554778  0.780306821  0.510565569  0.282872307
```

```
## 2004 -0.885536731 -0.448474179 -0.199775208  0.051889840  0.275090947
## 2005  0.452757414  0.064172844 -0.261617384 -0.736935133 -0.804343473
## 2006  0.029150661  0.637858707  0.818125116  0.849564843  1.072620030
## 2007 -0.920299489 -0.891000530 -0.963634897 -0.858792574 -0.784720264
## 2008  1.114549703  0.849097052  0.762521874  0.489908696  0.477813627
## 2009 -0.766388176 -0.494870216 -0.242982554 -0.120230903
##               Jun          Jul          Aug          Sep          Oct
## 2001  0.927117835  0.845204355  0.616786785  0.298442929 -0.075063283
## 2002 -0.596930426 -0.301921703  0.138147801  0.319349627  0.627042608
## 2003 -0.003617255 -0.412733240 -0.546933667 -0.942737390 -0.924877131
## 2004  0.659179838  0.824075796  0.995585361  0.930480566  1.143999138
## 2005 -0.792234686 -1.066599349 -0.925153593 -0.628233800 -0.723352809
## 2006  0.925935099  0.942168253  0.749024375  0.230305359 -0.121229498
## 2007 -0.707183703 -0.244622308 -0.043999159  0.086417967  0.488345366
## 2008 -0.015245254 -0.255219711 -0.636779567 -0.761488077 -0.849237386
## 2009
##               Nov          Dec
## 2001 -0.597720119 -0.373199683
## 2002  0.688668524  0.826475314
## 2003 -1.215934352 -0.985169589
## 2004  0.904199956  0.658081264
## 2005 -0.277870148 -0.058371311
## 2006 -0.103420789 -0.588488590
## 2007  0.725819290  0.897256797
## 2008 -0.901963678 -1.010946936
## 2009

#creating a weekly time series beginning at January 2001
weeklyTS<- ts(new_data, freq = 52, start=c(2001,1))
plot(weeklyTS, xlim = c(2001, 2003))
```

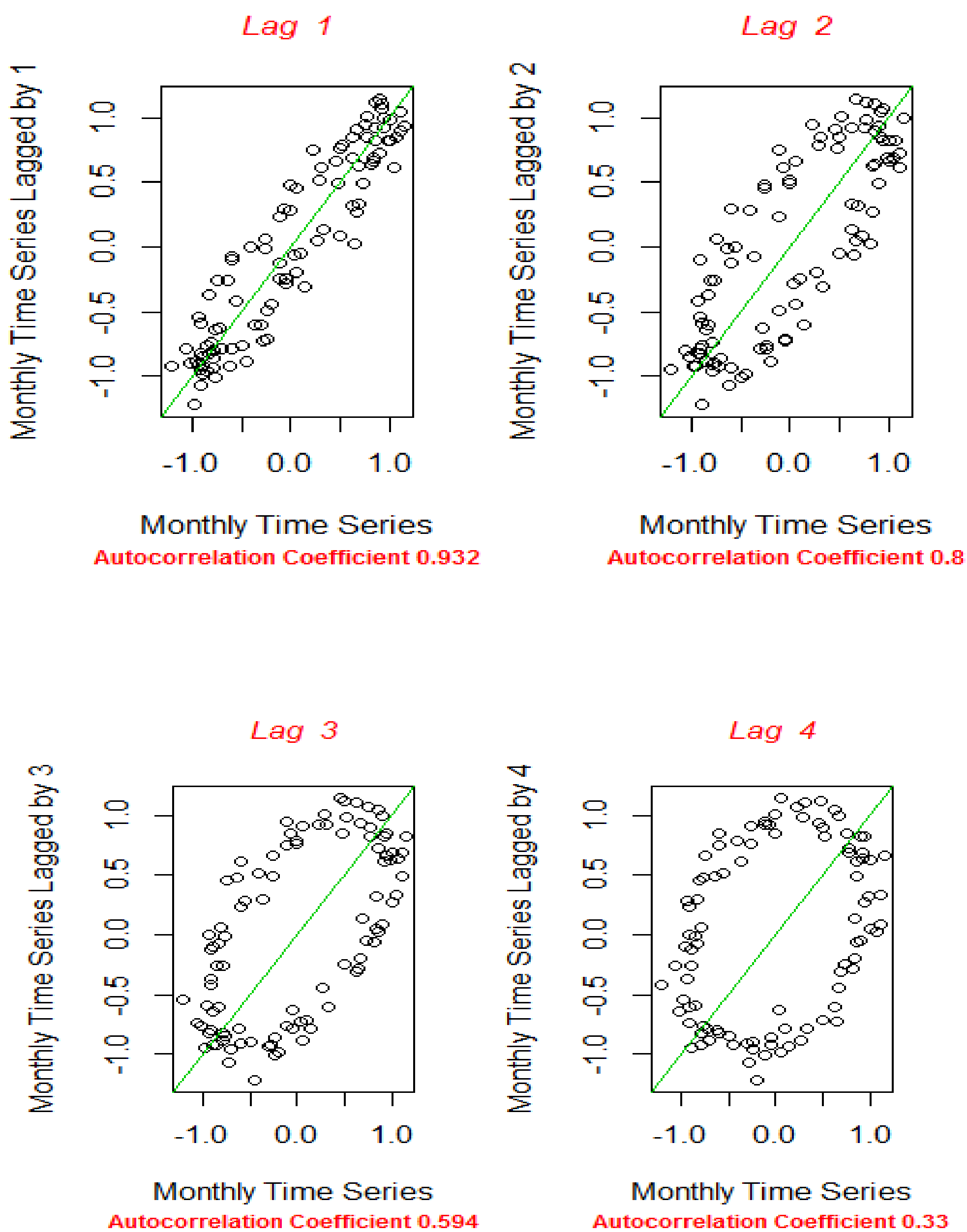


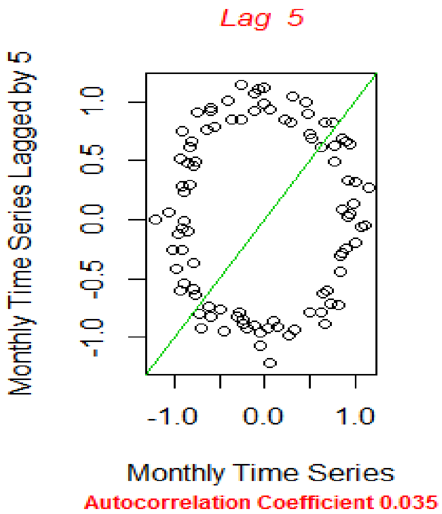
```
weeklyTS

## Time Series:
## Start = c(2001, 1)
## End = c(2002, 48)
## Frequency = 52
##   [1]  0.333003457  0.691087298  0.620092363  1.048182189  1.104222087
##   [6]  0.927117835  0.845204355  0.616786785  0.298442929 -0.075063283
##  [11] -0.597720119 -0.373199683 -0.828513613 -0.827707623 -0.912583404
##  [16] -0.938572141 -0.790060789 -0.596930426 -0.301921703  0.138147801
##  [21]  0.319349627  0.627042608  0.688668524  0.826475314  0.977625001
##  [26]  1.004554778  0.780306821  0.510565569  0.282872307 -0.003617255
##  [31] -0.412733240 -0.546933667 -0.942737390 -0.924877131 -1.215934352
##  [36] -0.985169589 -0.885536731 -0.448474179 -0.199775208  0.051889840
##  [41]  0.275090947  0.659179838  0.824075796  0.995585361  0.930480566
##  [46]  1.143999138  0.904199956  0.658081264  0.452757414  0.064172844
##  [51] -0.261617384 -0.736935133 -0.804343473 -0.792234686 -1.066599349
##  [56] -0.925153593 -0.628233800 -0.723352809 -0.277870148 -0.058371311
##  [61]  0.029150661  0.637858707  0.818125116  0.849564843  1.072620030
##  [66]  0.925935099  0.942168253  0.749024375  0.230305359 -0.121229498
##  [71] -0.103420789 -0.588488590 -0.920299489 -0.891000530 -0.963634897
##  [76] -0.858792574 -0.784720264 -0.707183703 -0.244622308 -0.043999159
##  [81]  0.086417967  0.488345366  0.725819290  0.897256797  1.114549703
##  [86]  0.849097052  0.762521874  0.489908696  0.477813627 -0.015245254
##  [91] -0.255219711 -0.636779567 -0.761488077 -0.849237386 -0.901963678
##  [96] -1.010946936 -0.766388176 -0.494870216 -0.242982554 -0.120230903
```

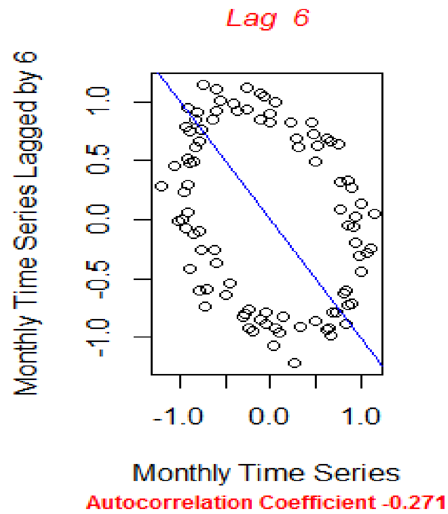
But, we are off track, let's get back to autocorrelation. Let's create various lagged time series data sets and let’s graph them, looking for that linear relationship:

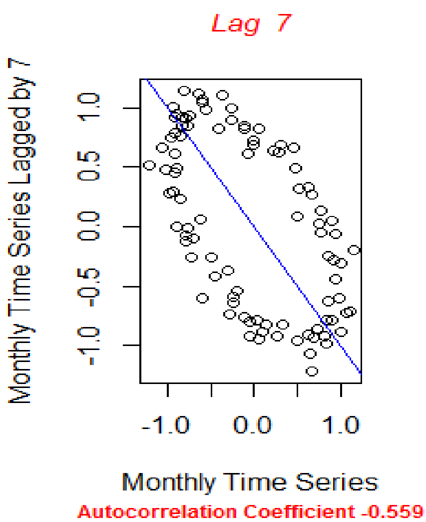
```
l <- length(monthlyTS)
r <-c()
par(mfrow=c(1,2))
for (i in 1:10){
       lagged<- monthlyTS[(1+i): l]
       laggedToo = monthlyTS[1:(l-i)]
       r[i] <- round(cor(lagged, laggedToo),3)

       plot(lagged, laggedToo, xlab ="Monthly Time Series", ylab
=paste("Monthly Time Series Lagged by",i))
       title(main = paste("Lag ",i), sub = paste("Autocorrelation
Coefficient", r[i]),
      cex.main = 1,   font.main= 3, col.main= "red",
      cex.sub = 0.75, font.sub = 2, col.sub = "red")
       if (i <= 5) {
               abline(a=0, b=1, col=3)
```

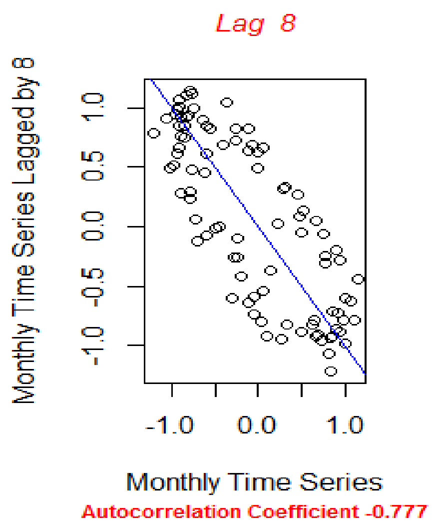
```
        }
if (i > 5){
        abline(a=0, b=-1, col=4)
}

}
```

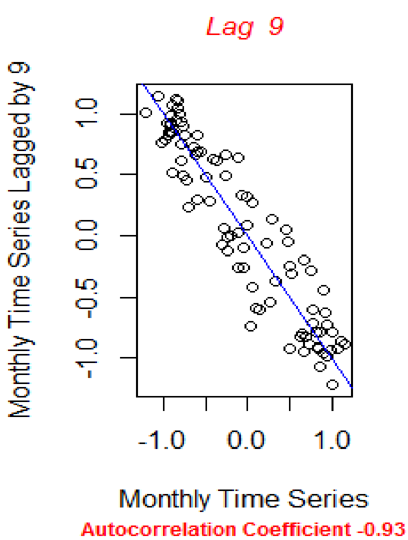


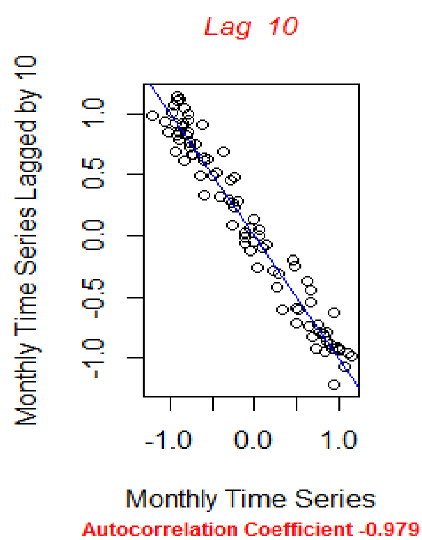

Lag 5
Monthly Time Series Lagged by 5
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient 0.035


Lag 6
Monthly Time Series Lagged by 6
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient -0.271


Lag 7
Monthly Time Series Lagged by 7
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient -0.559

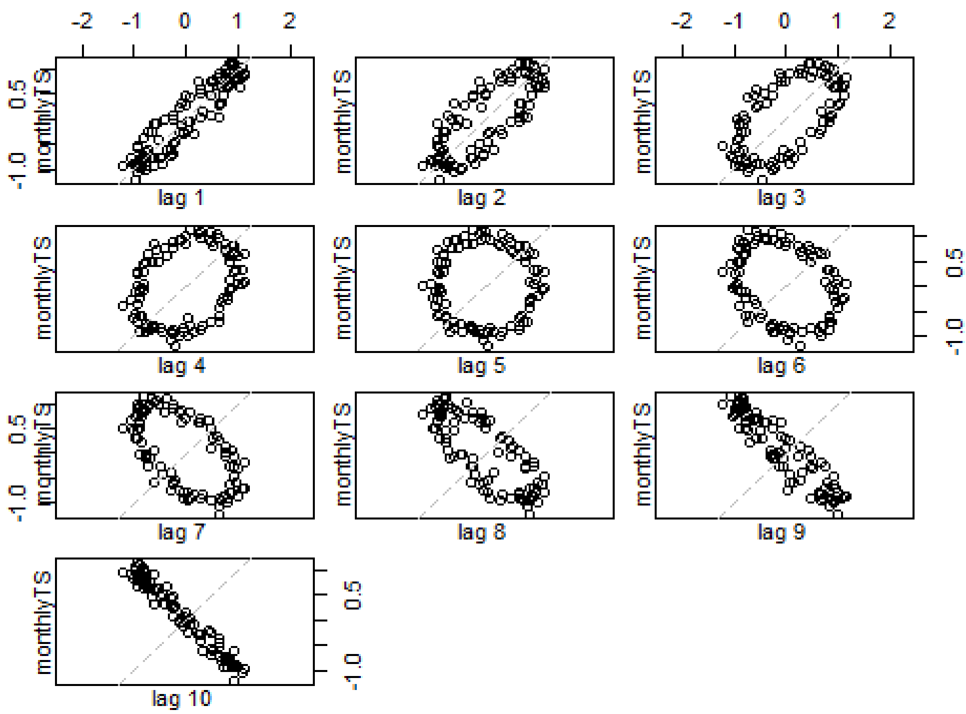

Lag 8
Monthly Time Series Lagged by 8
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient -0.777


Lag 9
Monthly Time Series Lagged by 9
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient -0.93


Lag 10
Monthly Time Series Lagged by 10
1.0
0.5
0.0
-0.5
-1.0
-1.0
0.0
1.0
Monthly Time Series
Autocorrelation Coefficient -0.979

And now I confess that there is an R function that does all this, but I wanted demonstrate what the $lag.plot()$ command is about.

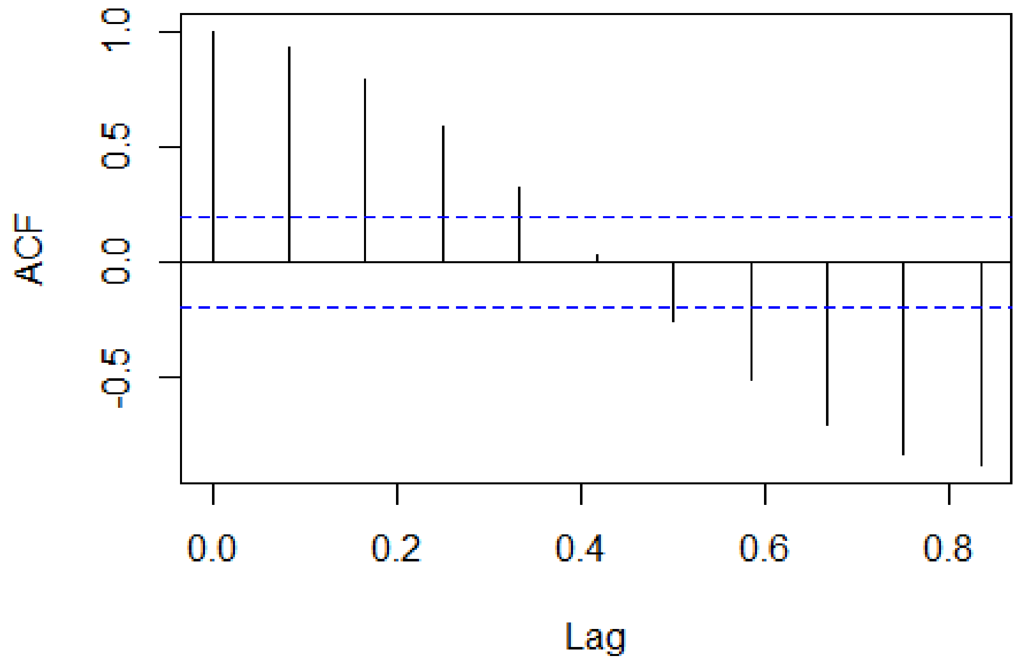
```
lag.plot(monthlyTS, lags=10, do.lines =FALSE)
```



In fact, the $acf()$ command in R will graph the autocorrelation coefficients versus the lag, and this graph is referred to as the auto correlation function or ACF. The data is said to not show auto correlation whenever the auto correlation coefficients are close to zero. The time series that shows no auto correlation is called white noise. But what does it mean for the auto correlation coefficients to be close to zero? Well, it turns out that if 95% of these coefficients are within $\pm\frac{2}{\sqrt{T}}$, where T is the length of the time series, then the time series is a white noise series.

Let's look at the ACF of the $monthlyTS$:

```
acf(monthlyTS, lag.max=10, plot = TRUE)
```



Note that the length of the time series *monthlyTS* is

```
T <- length(monthlyTS)
2/sqrt(T)

## [1] 0.2

-1*2/sqrt(T)

## [1] -0.2
```

And, so the blue horizontal lines on the above graph represent the upper and lower bounds $\pm\frac{2}{\sqrt{T}}$.

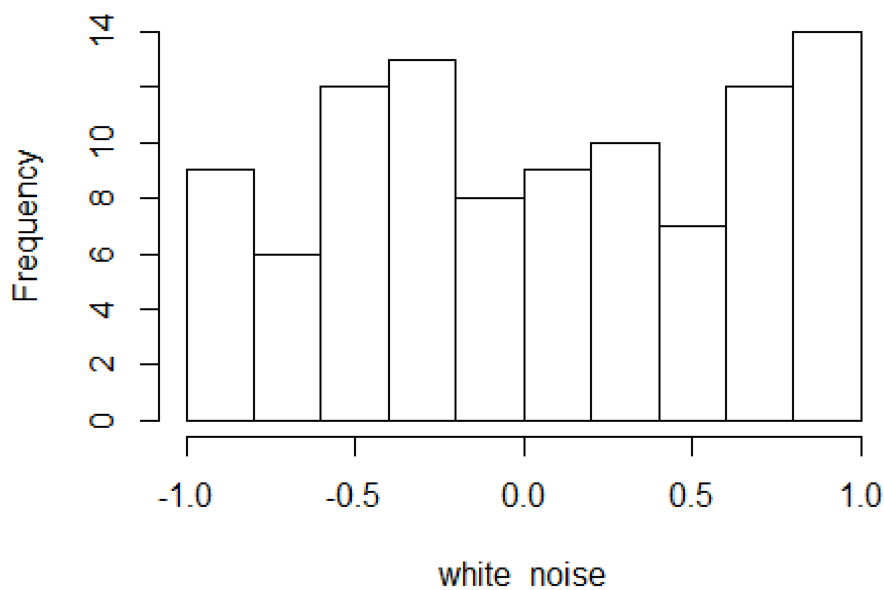
Example of a White Noise Time Series

Let's generate 100 uniformly distributed random numbers between -1 and 1.

```
white_noise <- runif(100, min=-1, max=1)
```

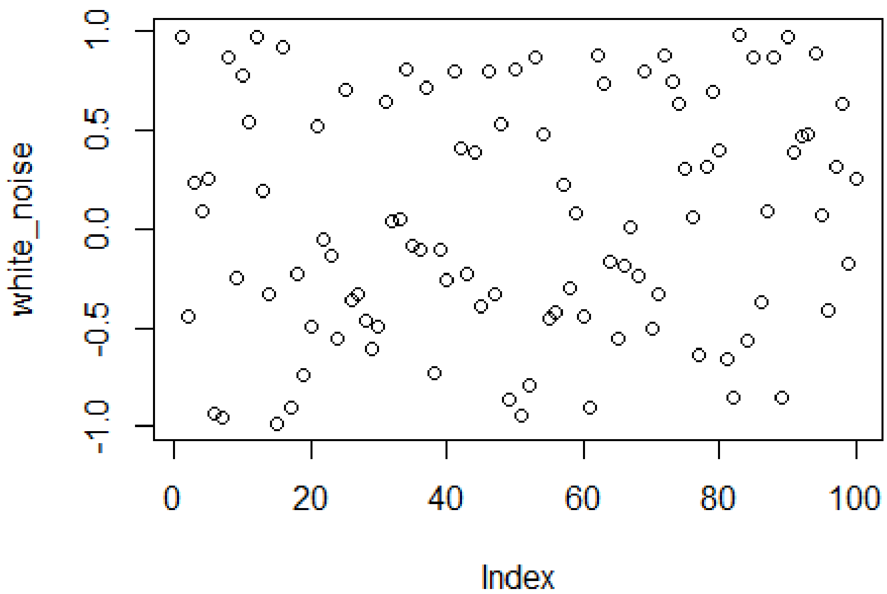
Its histogram should show that it is of uniform distribution:

```
hist(white_noise)
```

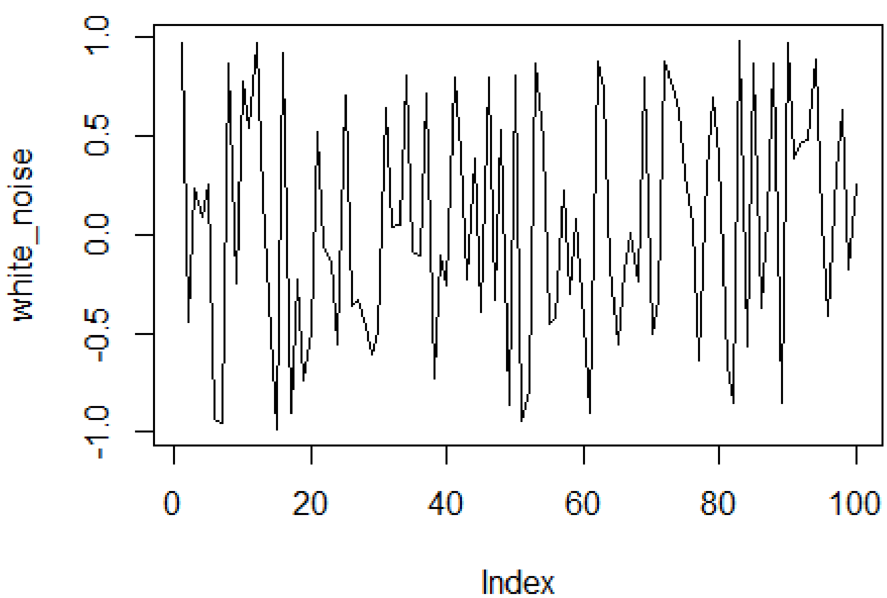


Next, let's plot *white_noise* first as a scatter plot then as a line plot:

```
plot(white_noise)
```

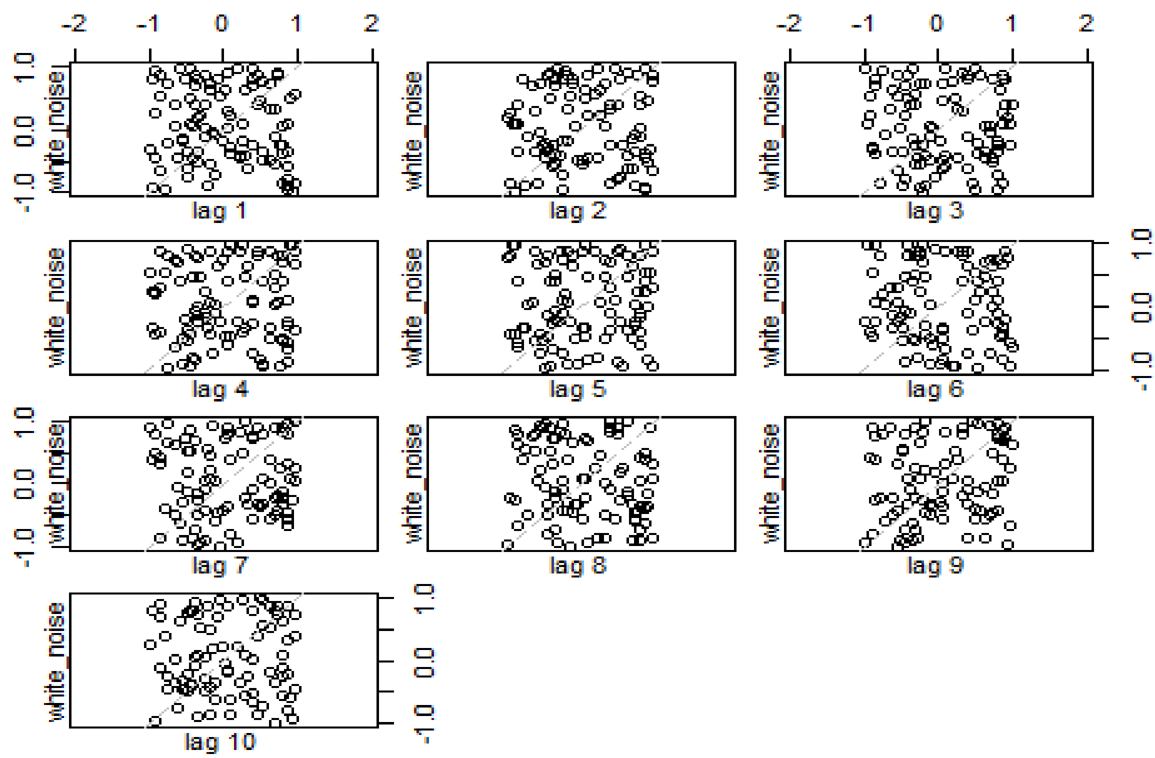



```r
plot(white_noise, type="l")
```



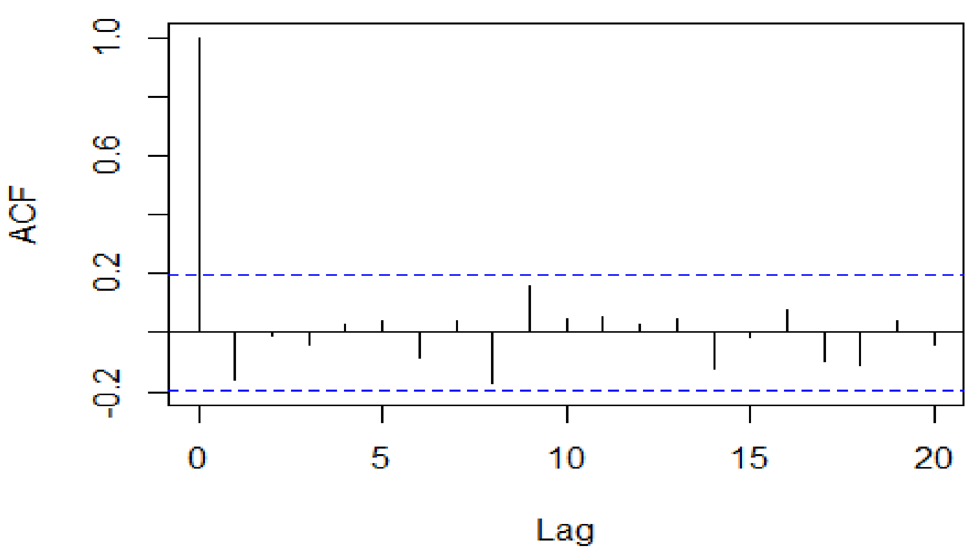
So, the question is, does this data set show auto correlation? To find out, let's graph its lagged scatter plots:

```r
lag.plot(white_noise, lags=10, do.lines =FALSE)
```



Whew, none of these graph show any evidence of a linear relationship, but just to be certain, here is the *ACF* graph of *white_noise*:

```
acf(white_noise, plot = TRUE)
```



Auto Correlation in the Bike Share Demand Data

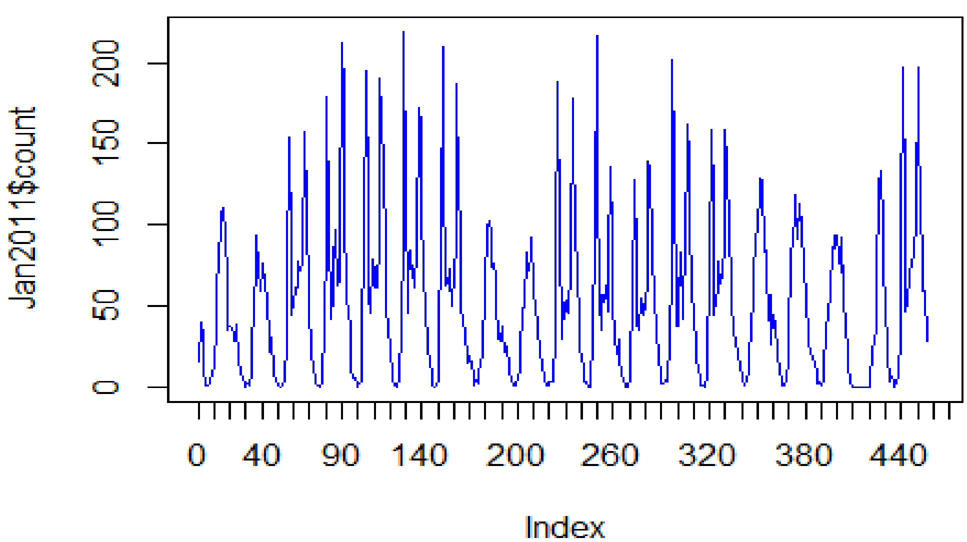
I have imputed the missing values into the original *train.csv*, and I have saved the months of January of 2011 and February of 2011 into separate *.csv* files.

Reading in the Data

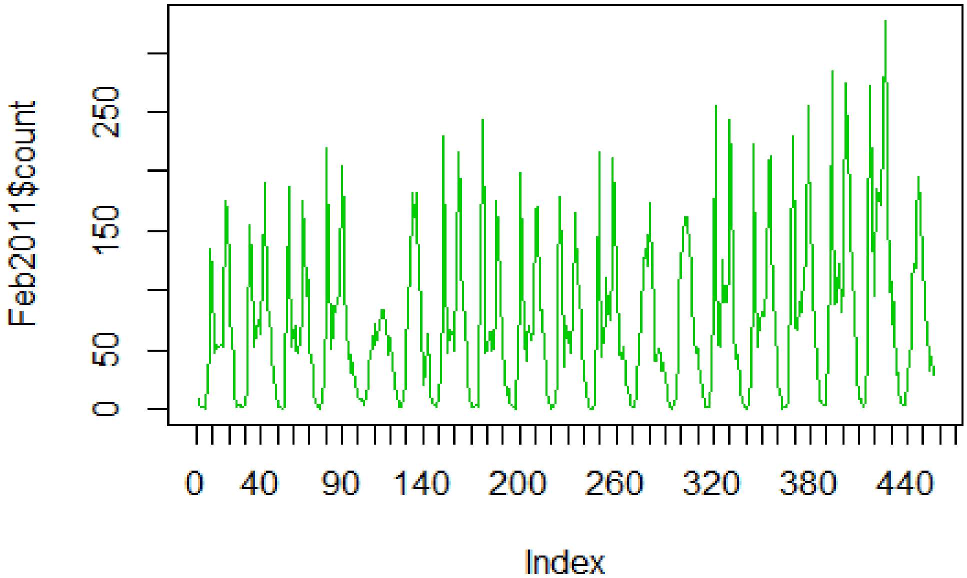
```
Jan2011 <- read.csv("Jan2011.csv")
Feb2011 <- read.csv("Feb2011.csv")
```

Looking at the Data

```
axis_values <- seq(0, 500, by= 10)
#par(mfrow=c(1,2))  #arranging the two graphs in one row
plot(Jan2011$count, type="l", col=4, xaxt="n", xlim=c(0, 460))
axis(1,axis_values )
```

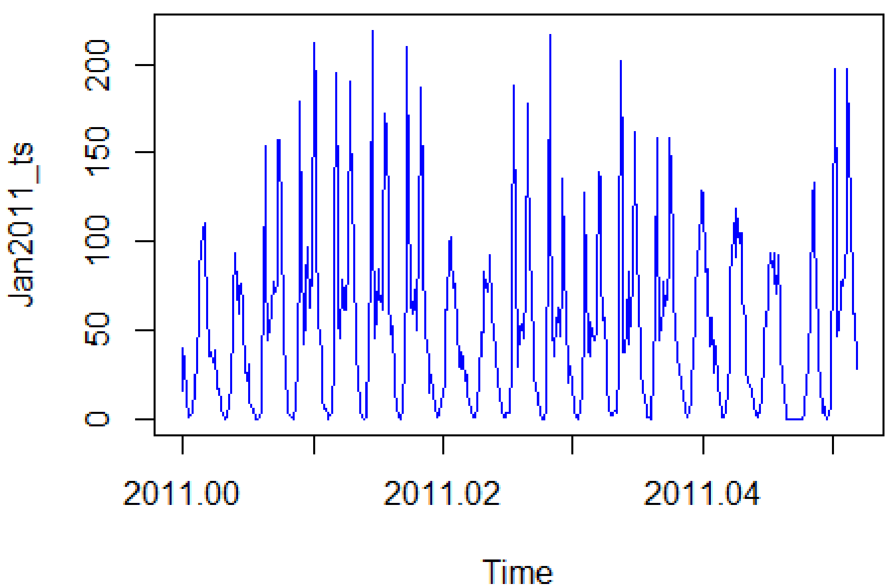



```
plot(Feb2011$count, type="l", col=3, xaxt="n")
#plot(1:100, xaxt = "n")
axis(1,axis_values )
```

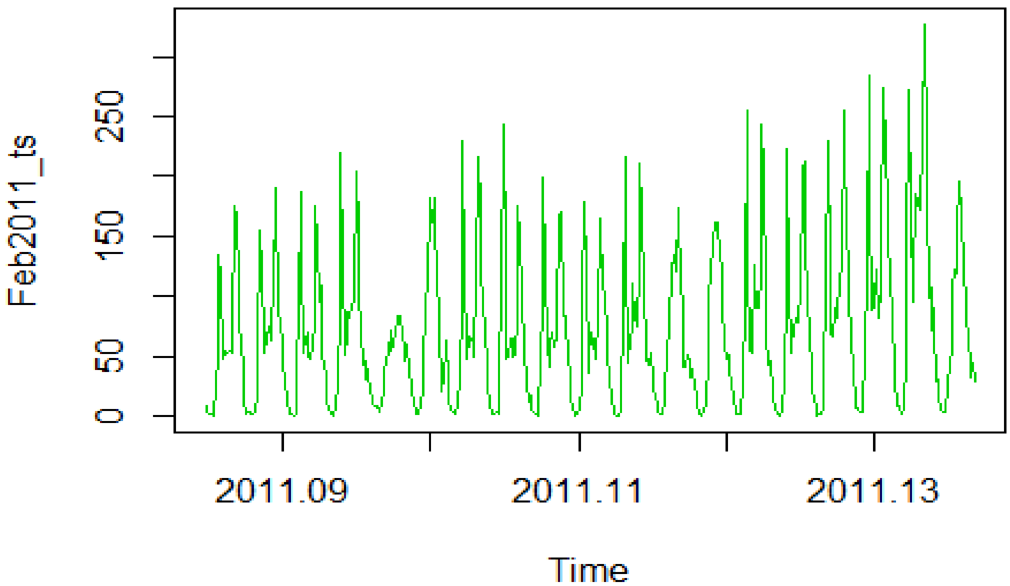


Or, we could look at the data as a time series:

```
Jan2011_ts <- ts(Jan2011$count, freq=365*24, start=c(2011,0) )
Feb2011_ts <- ts(Feb2011$count, freq = 365*24, start=c(2011, 31*24))
plot.ts(Jan2011_ts, col=4)
```



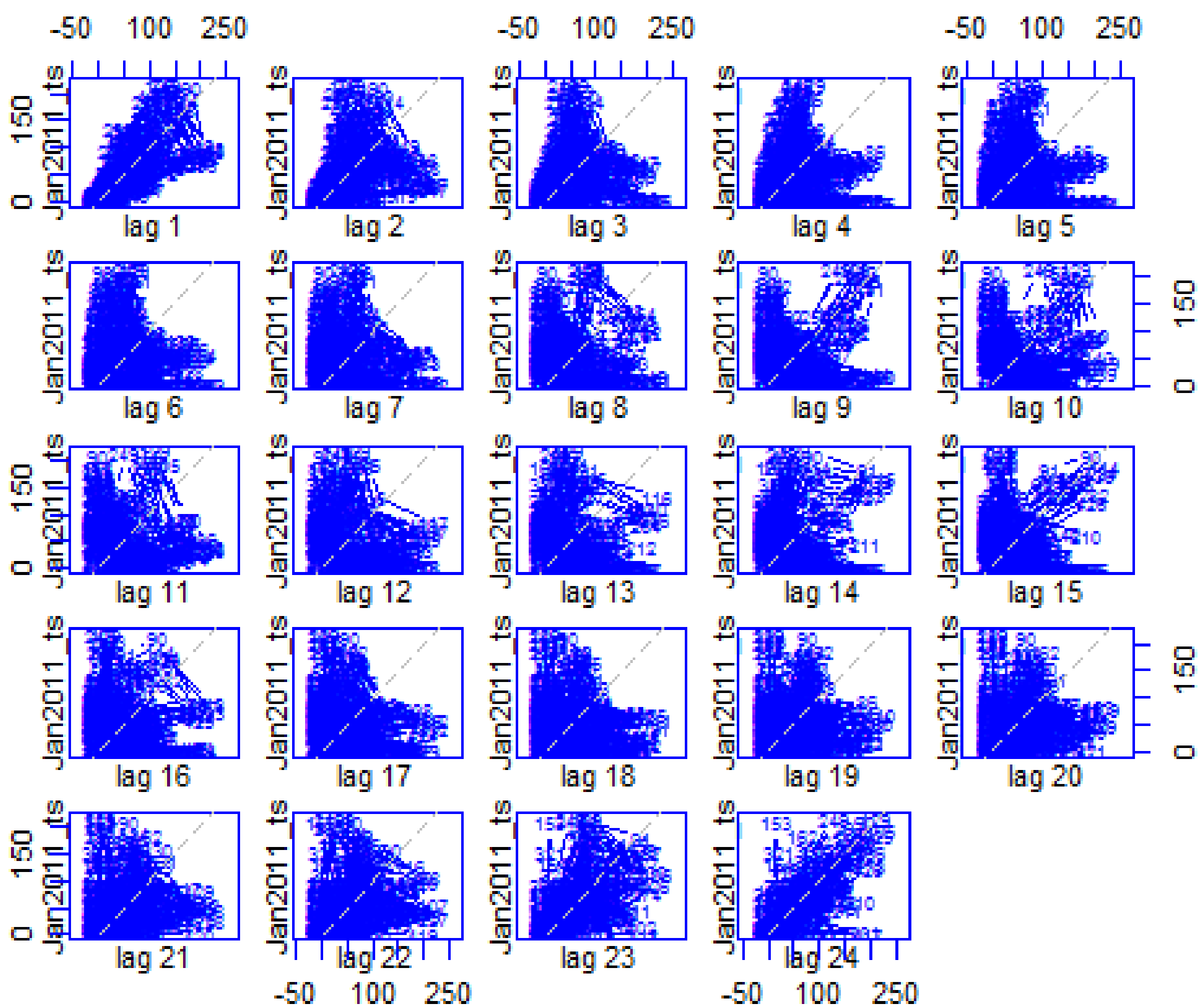
```r
plot.ts(Feb2011_ts, col=3)
```



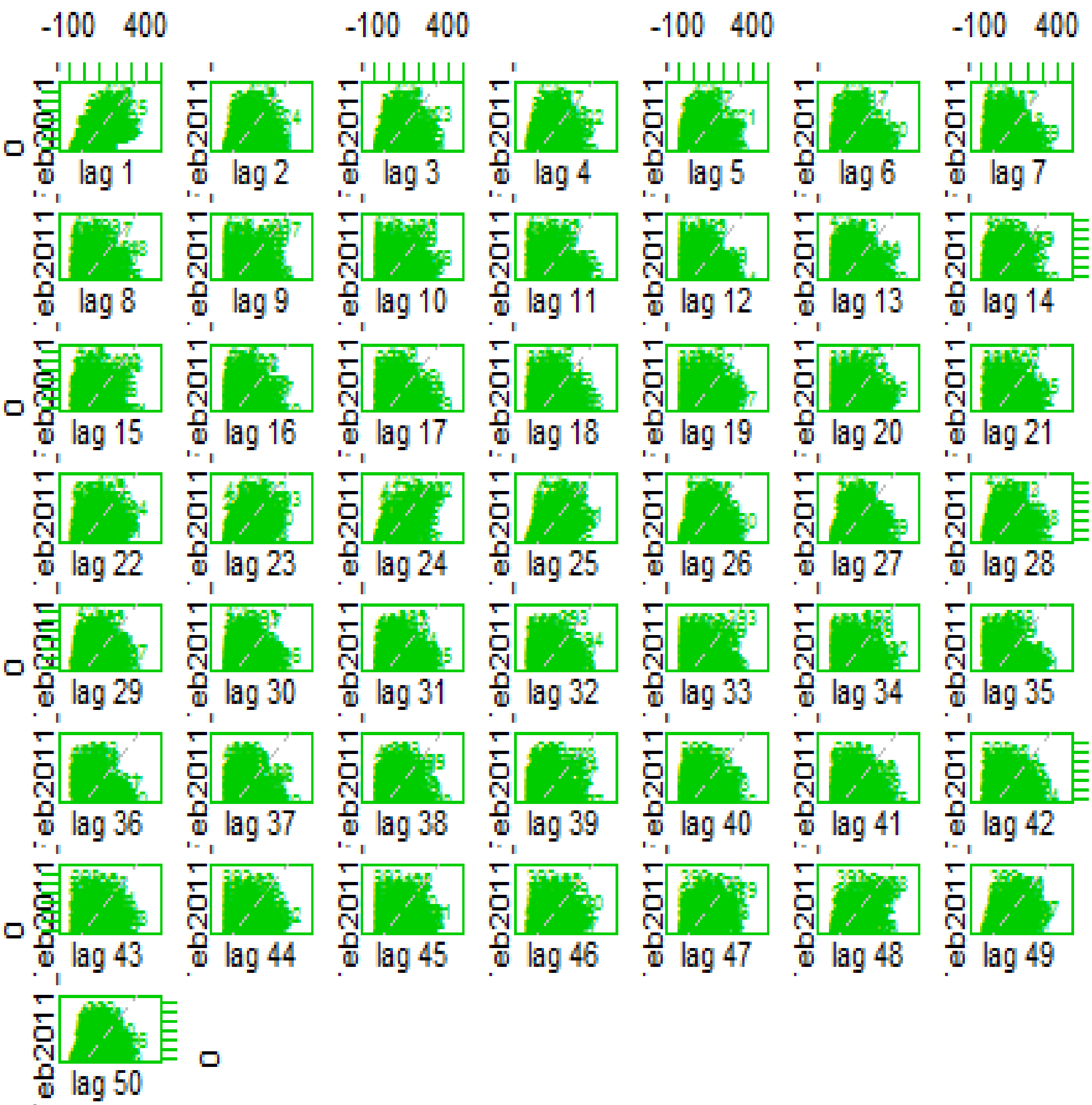
Is There Auto Correlation in the January 2011 and February 2011 Time Series?

Let's look at the lagged scatter plots for each month:

```
lag.plot(Jan2011_ts, lags=24, do.lines =TRUE, col=4)
```

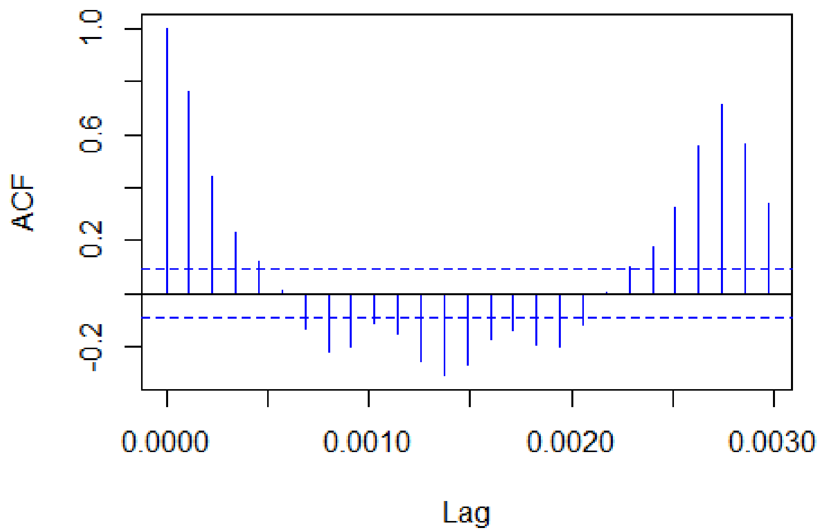


```r
lag.plot(Feb2011_ts, lags=50, do.lines =TRUE, col=3)
```

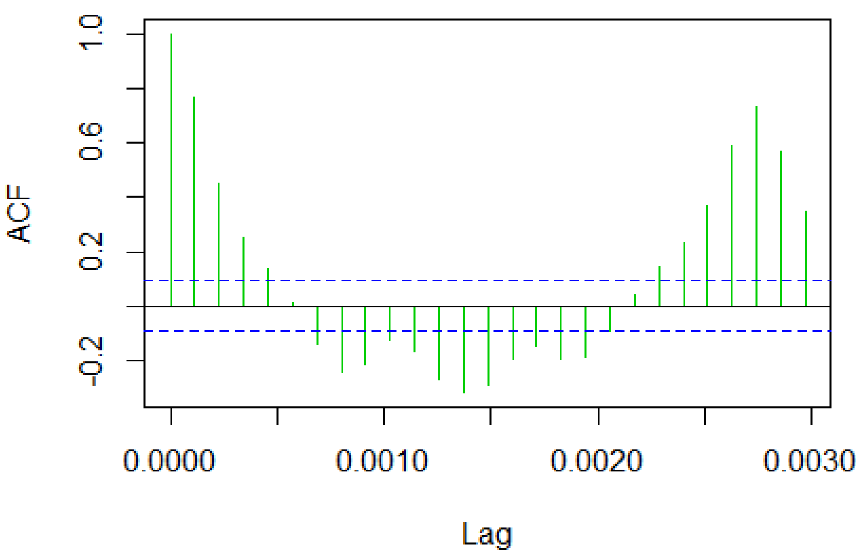


And just to be entirely convincing, here are the ACF graphs of the lags:

```
acf(Jan2011_ts, plot = TRUE, col=4)
```




```
acf(Feb2011_ts, plot = TRUE, col=3)
```



From the graphs it is clear that more than 5% of the lags are outside of the bounds of $\pm\frac{2}{\sqrt{T}}$, where $T = T_{Jan} = T_{Feb} = 456$, the number of observations in each month, and $\pm\frac{2}{\sqrt{T}} = \pm\frac{2}{\sqrt{456}} \approx \pm 0.094$. Therefore the data sets $Jan2011$ and $Feb2011$ show auto correlation.