PSTAT 174 Time Series Final Project

2024-05-25

Analysis of NYC Shootings dataset

This dataset involves analyzing the number of shootings in NYC over time from January 2006 to December 2022. The data is taken from NYC's public data repository. I had to clean the data because the data before only gave me a table of 27,000 values of the date, time, location, etc. of the shooting. Thus the dataset I'm using is a cleaned and self-modified version of the original one.

```
library(astsa) #acf
library(readr)
library(tseries) #ADF test
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(forecast) #auto.arima function
```

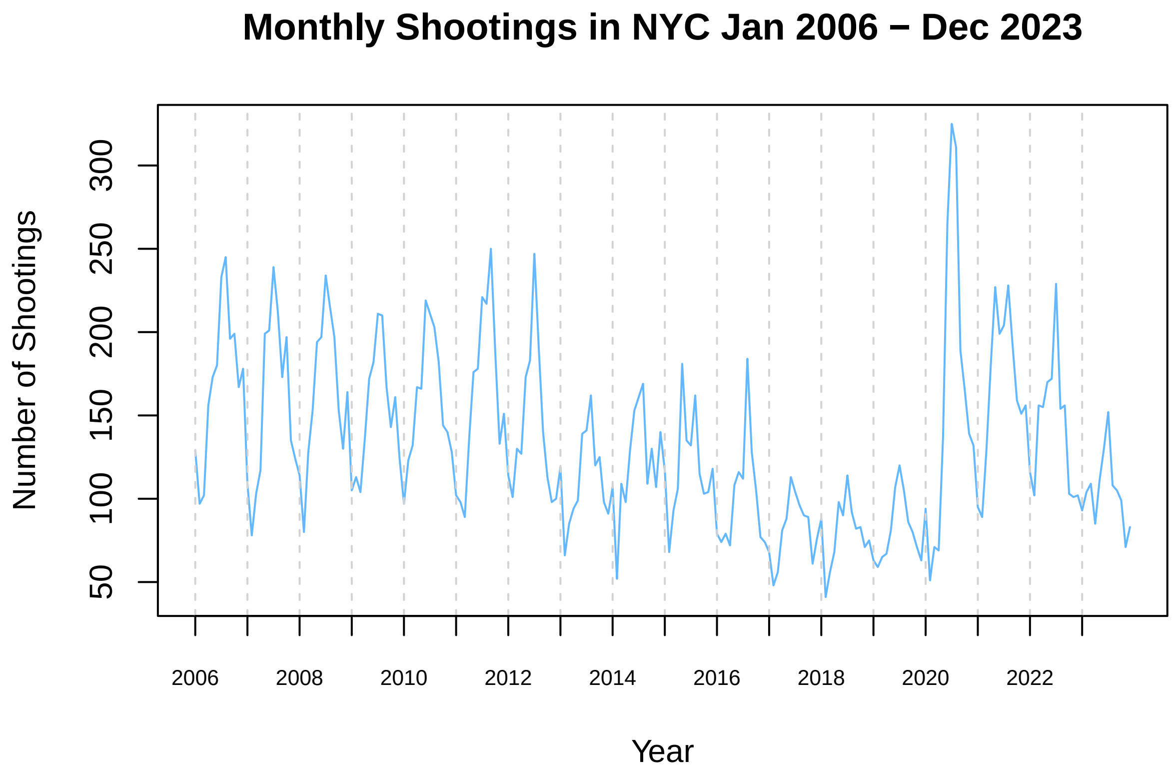
```
##
## Attaching package: 'forecast'

## The following object is masked from 'package:astsa':
##
##     gas
```

```
#Reading the data into R as a time series
dict <- read_csv("dict.csv")$Shootings
```

```
## Rows: 216 Columns: 2

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): Shootings
## date (1): Date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
shootings_ts <- ts((dict), start=c(2006, 1), end=c(2023, 12), frequency=12)

#Plotting our shooting data as a time series
plot(shootings_ts, xlab='Year', ylab='Number of Shootings', main='Monthly Shootings in NYC Jan 2006 - De
years <- seq(2006, 2023, by=1)
```

```r
for (year in years) {
  abline(v=year, col="lightgray", lty=2)
}
axis(1, at=seq(2006, 2023, by=1), labels=seq(2006, 2023, by=1), cex.axis=0.7)
```

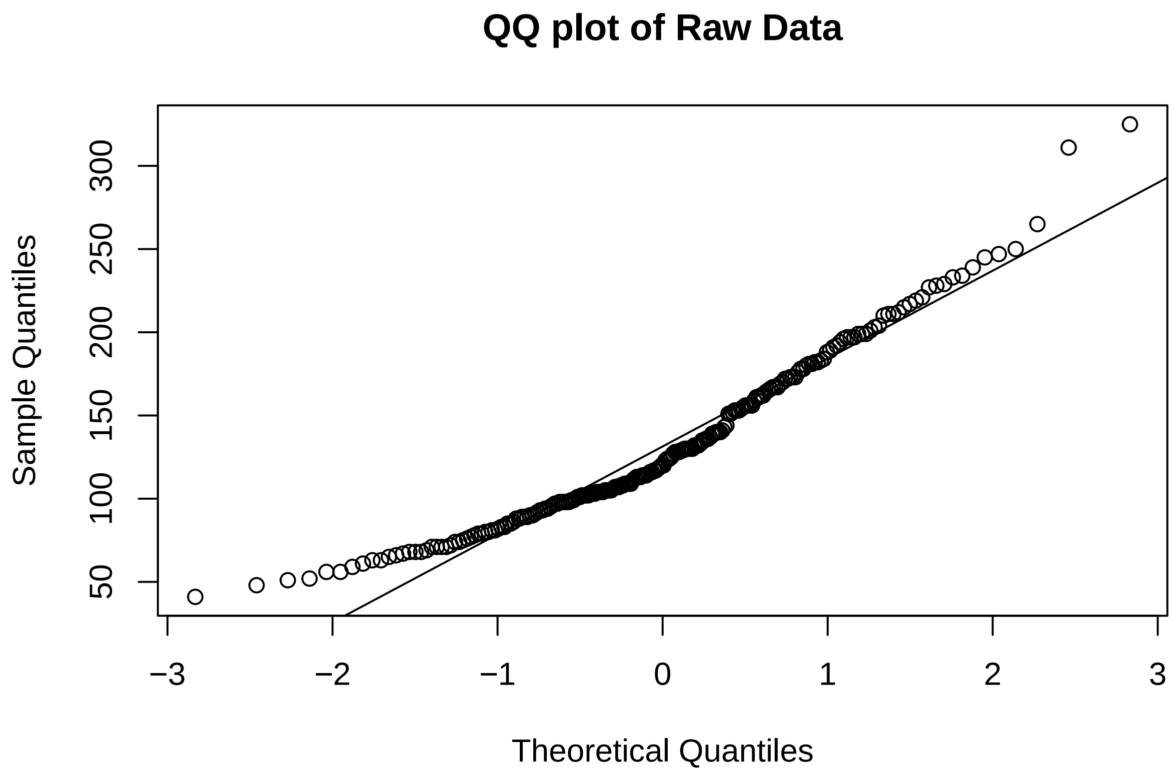



```r
#Analysis of the data
adf.test(shootings_ts) # adf gives result as stationary
```

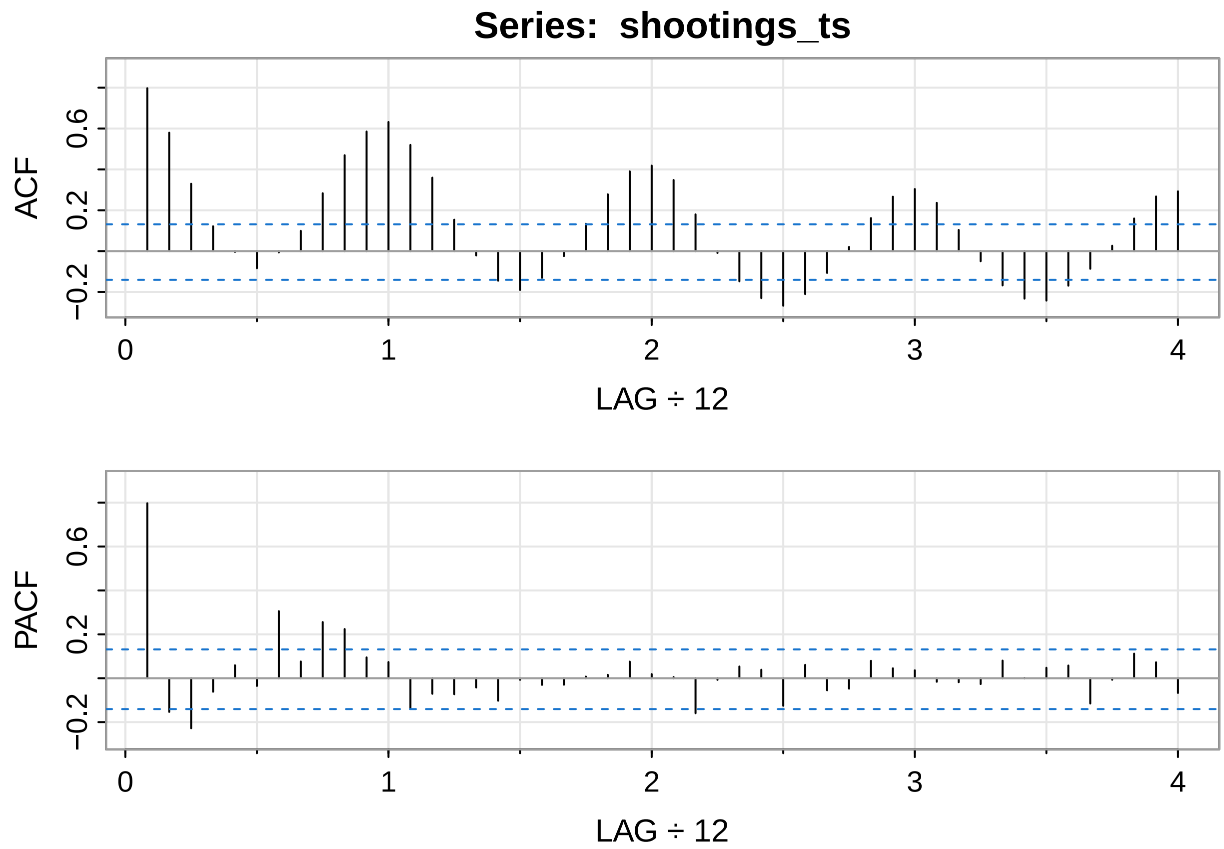
```
## Warning in adf.test(shootings_ts): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  shootings_ts
## Dickey-Fuller = -5.642, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```r
qqnorm(shootings_ts, main='QQ plot of Raw Data')
qqline(shootings_ts) # the curve in the points indicates that data is not normal
```

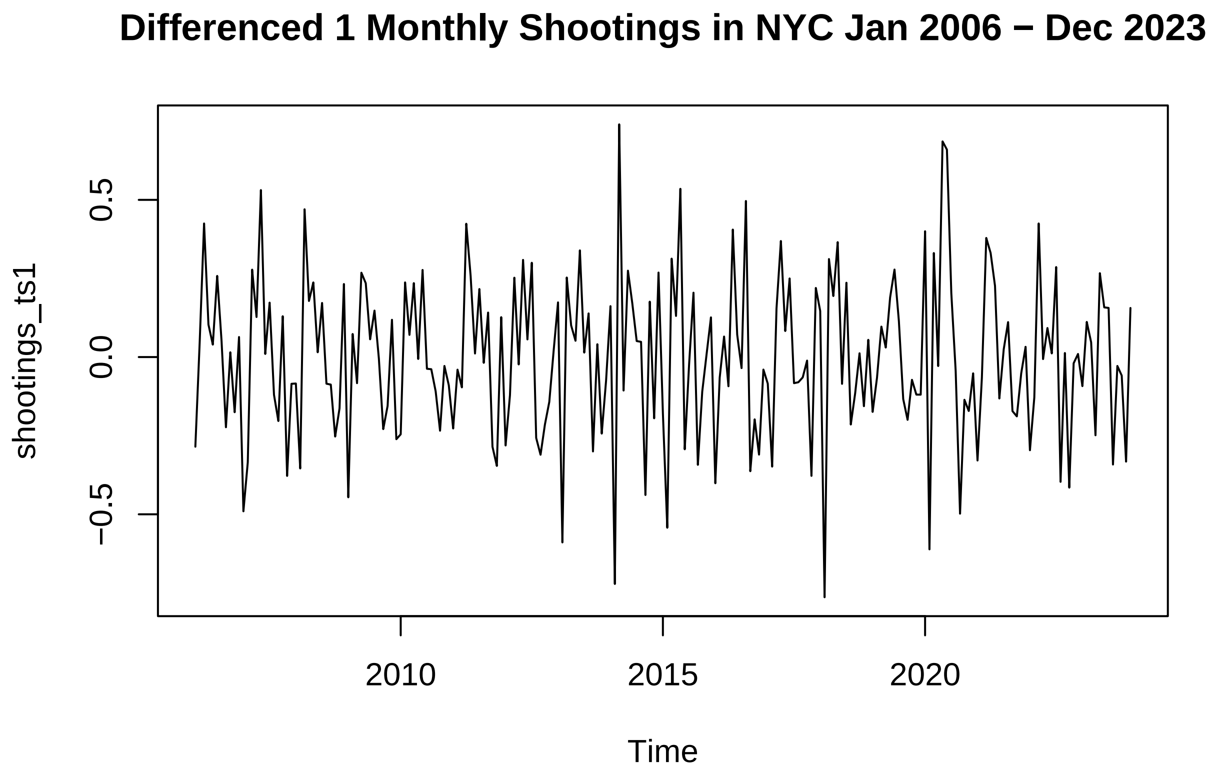



```
acf2(shootings_ts) # acf and pacf show heavy seasonality
```

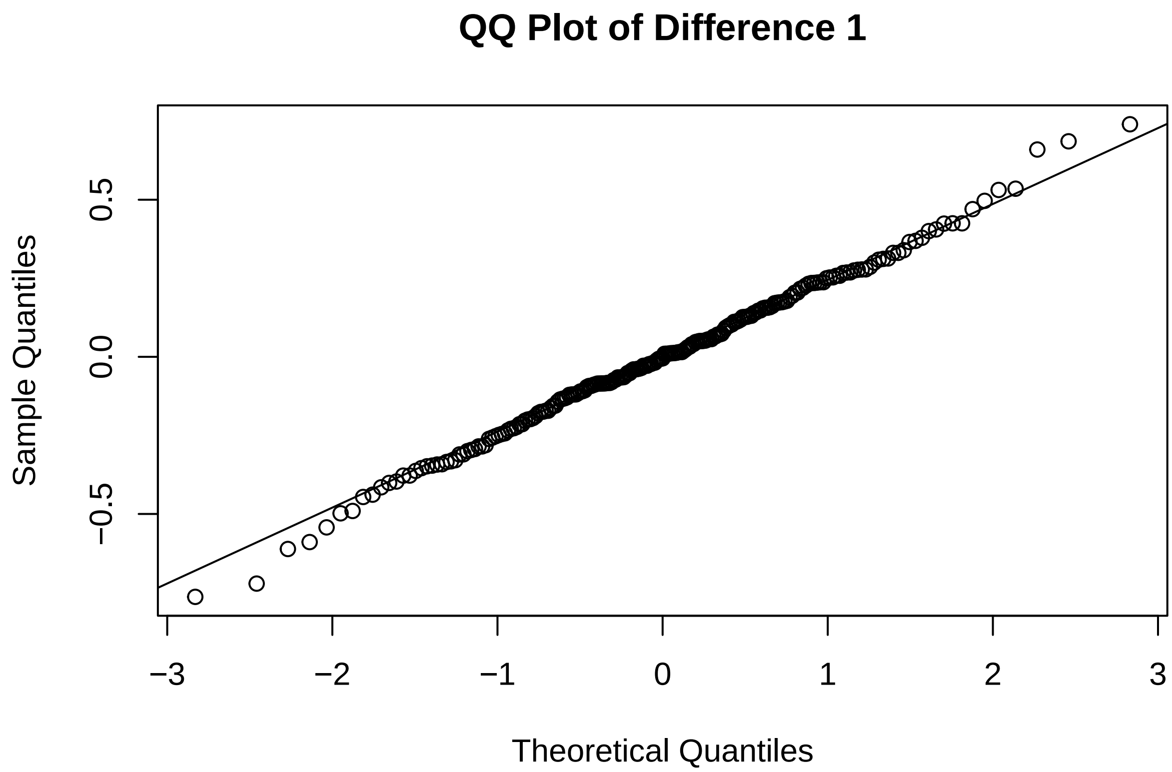



```
##      [,1]  [,2]  [,3]  [,4] [,5]  [,6]  [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## ACF   0.8  0.58  0.33  0.12 0.00 -0.08 -0.01 0.10 0.28  0.47  0.59  0.63  0.52
## PACF  0.8 -0.15 -0.23 -0.06 0.06 -0.04  0.31 0.08 0.26  0.22  0.10  0.07 -0.14
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF   0.36  0.15 -0.02 -0.14 -0.19 -0.13 -0.02  0.13  0.28  0.39  0.42  0.35
## PACF -0.07 -0.07 -0.04 -0.10 -0.01 -0.03 -0.03  0.01  0.02  0.08  0.02  0.01
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF   0.18 -0.01 -0.15 -0.23 -0.27 -0.21 -0.11  0.02  0.16  0.27  0.30  0.24
## PACF -0.16 -0.01  0.05  0.04 -0.13  0.06 -0.06 -0.05  0.08  0.05  0.04 -0.02
##      [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48]
## ACF   0.10 -0.05 -0.17 -0.23 -0.24 -0.17 -0.09  0.03  0.16  0.27  0.29
## PACF -0.02 -0.03  0.08  0.00  0.05  0.06 -0.12 -0.01  0.11  0.07 -0.07
```

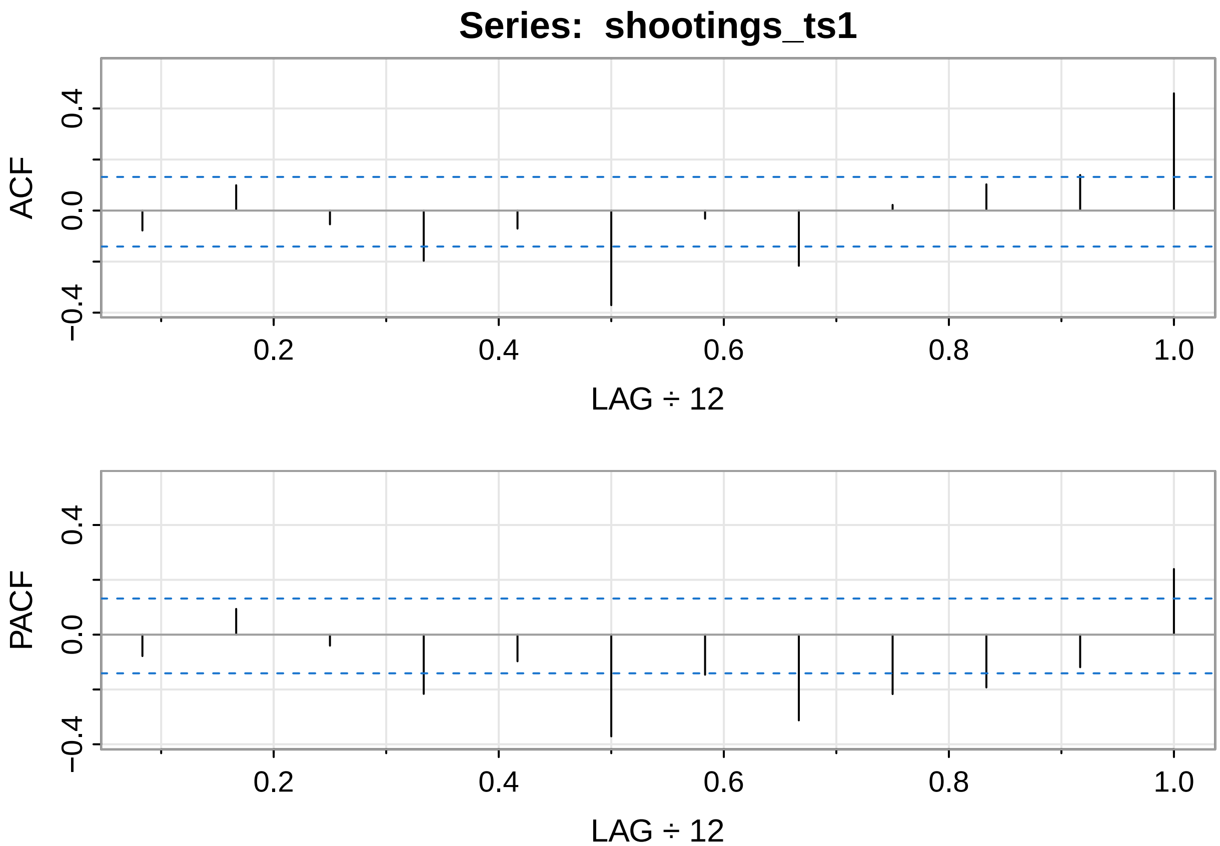
```r
#Since the data is non-normal, we apply Box-Cox transformation
shootings_ts1<-diff(log(shootings_ts))
ts.plot(shootings_ts1, main="Differenced 1 Monthly Shootings in NYC Jan 2006 - Dec 2023")
```




```
qqnorm(shootings_ts1, main='QQ Plot of Difference 1') #This plot looks much better, basically normal/st
qqline(shootings_ts1)
```

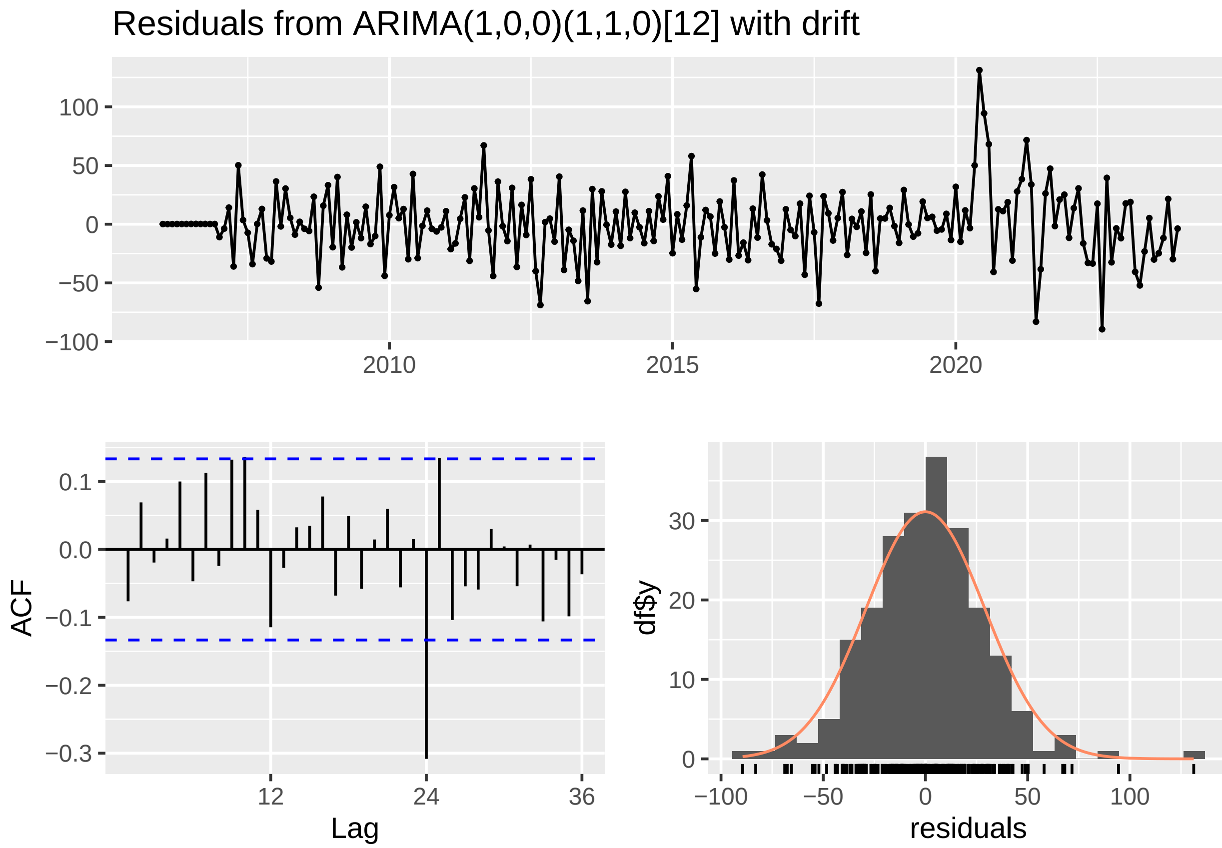



```
acf2(shootings_ts1, 12)
```




```
##        [,1] [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
## ACF  -0.08 0.10 -0.05 -0.20 -0.07 -0.37 -0.03 -0.22  0.02  0.10  0.14  0.46
## PACF -0.08 0.09 -0.04 -0.22 -0.10 -0.37 -0.15 -0.31 -0.22 -0.19 -0.12  0.24
```

```
# Since this doesn't give us a simple model for our SARIMA, we run auto.arima

# Predictions
shootings_modeled<-auto.arima(shootings_ts)
checkresiduals(shootings_modeled) # statistically checking the model
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0)(1,1,0)[12] with drift
## Q* = 49.883, df = 22, p-value = 0.0006081
##
## Model df: 2.   Total lags used: 24
```

```
summary(shootings_modeled) # the residuals look normal
```

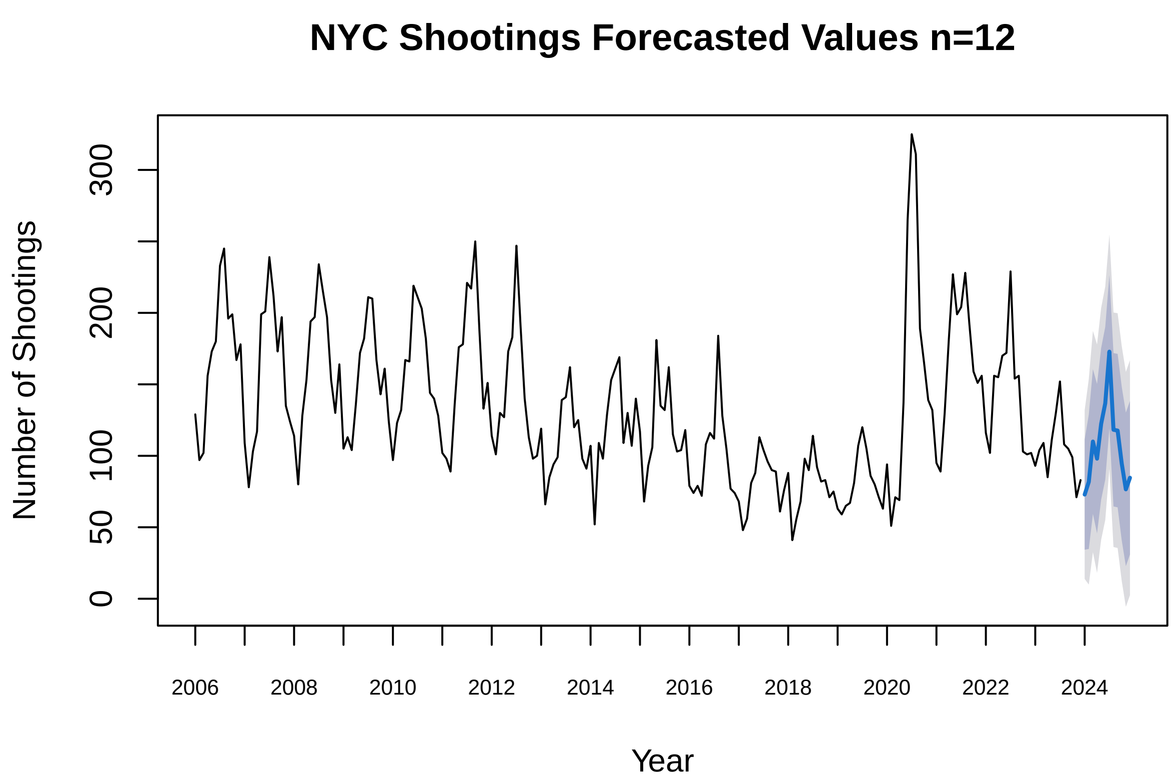
```
## Series: shootings_ts
## ARIMA(1,0,0)(1,1,0)[12] with drift
##
## Coefficients:
##          ar1     sar1    drift
##       0.6957  -0.3703  -0.3007
## s.e.  0.0507   0.0662   0.4204
##
## sigma^2 = 906.9:  log likelihood = -983.79
## AIC=1975.58   AICc=1975.78   BIC=1988.85
##
## Training set error measures:
##                       ME     RMSE      MAE       MPE     MAPE      MASE
## Training set -0.01241812 29.05056 21.49203 -2.150574 17.37663 0.7560569
##                     ACF1
```

```
## Training set -0.07634969
```

```r
# Forecasting future 12 values
future <- forecast(shootings_modeled, h = 12)
print(future)
```

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Jan 2024       72.88312  34.28915 111.4771 13.858744 131.9075
## Feb 2024       81.83495  34.81989 128.8500  9.931618 153.7383
## Mar 2024      109.99223  59.40225 160.5822 32.621539 187.3629
## Apr 2024       97.99690  45.76445 150.2294 18.114259 177.8795
## May 2024      122.35128  69.34213 175.3604 41.280782 203.4218
## Jun 2024      136.74554  83.36453 190.1266 55.106327 218.3848
## Jul 2024      172.87923 119.31916 226.4393 90.966175 254.7923
## Aug 2024      118.21861  64.57210 171.8651 36.173345 200.2639
## Sep 2024      117.63848  63.95017 171.3268 35.529297 199.7477
## Oct 2024       94.63207  40.92354 148.3406 12.491964 176.7722
## Nov 2024       76.53390  22.81559 130.2522 -5.621164 158.6890
## Dec 2024       84.65255  30.92951 138.3756  2.490250 166.8149
```

```r
plot(future, main='NYC Shootings Forecasted Values n=12', xlab = 'Year', ylab='Number of Shootings', xa
axis(1, at=seq(2006, 2024, by=1), labels=seq(2006, 2024, by=1), cex.axis=0.7)
```



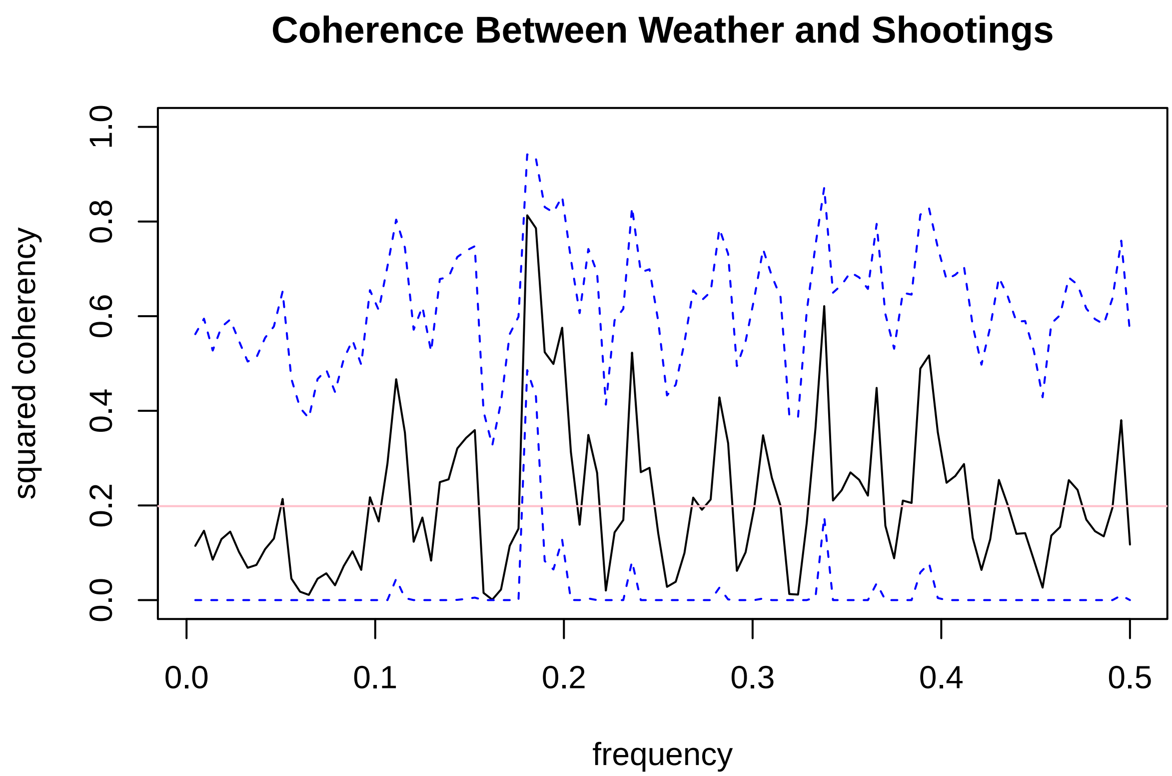
```r
weather_dict <- read_csv("new york.csv")$Value # I pulled in New York Weather Data from Jan 2006 - Dec
```

```
## Rows: 216 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): Date, Value, Anomaly
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

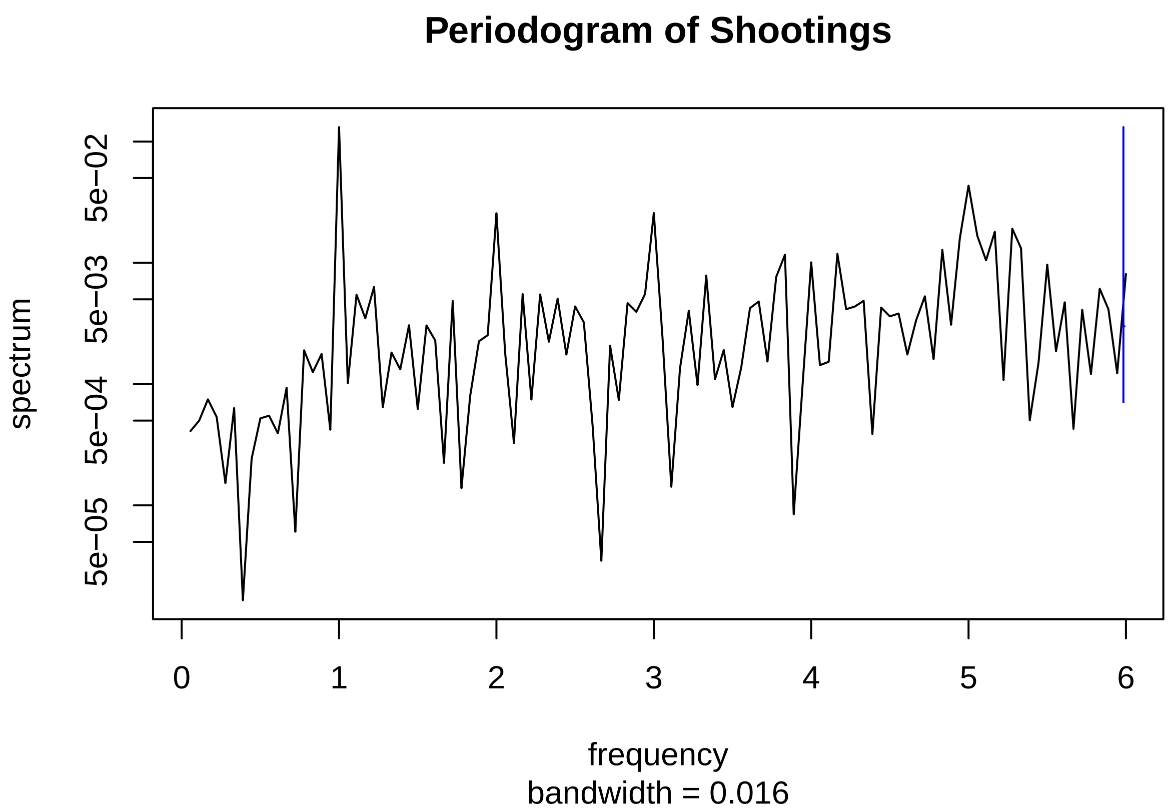
```r
weather_ts <- ts((weather_dict), start=c(2006, 1), end=c(2024, 5), frequency=12)

sr = mvspec(cbind(weather_dict, dict), kernel("daniell",2), plot=FALSE)
sr$df
```

```
## [1] 10
```

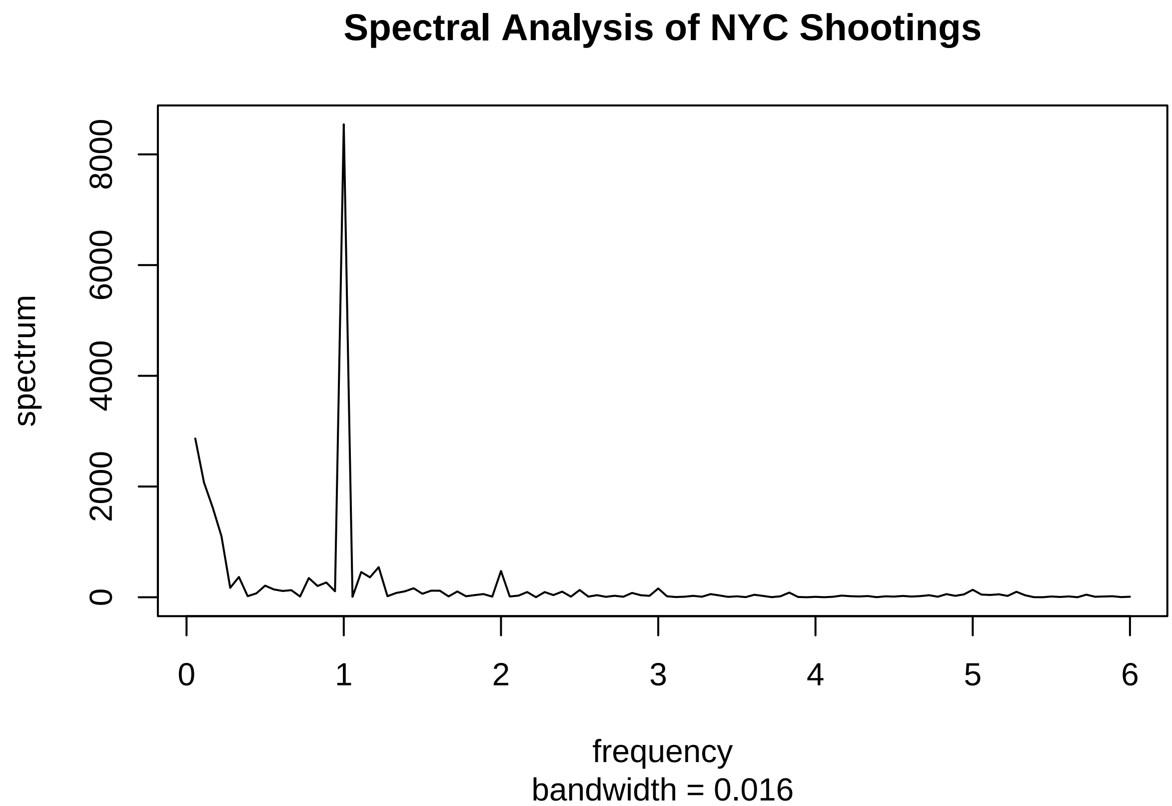
```r
f = qf(.95, 2, sr$df-2)
C = f/(18+f)
plot(sr, plot.type = "coh", ci.lty = 2, main='Coherence Between Weather and Shootings')
abline(h = C, col='pink')
```



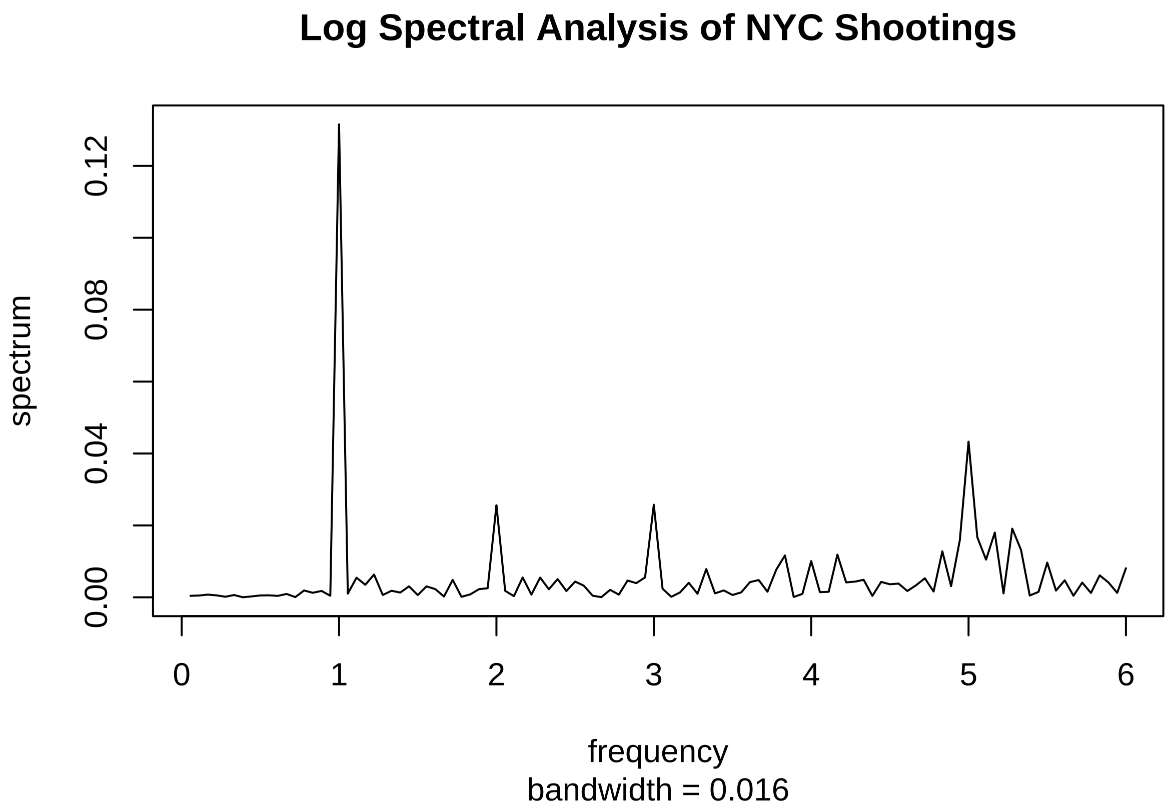
```r
periodogram <- spec.pgram(shootings_ts1, plot=FALSE)
plot(periodogram, log = "y", main = "Periodogram of Shootings") # Frequencies are especially strong eve
```




```r
x.spec <- spectrum(shootings_ts, main= "Spectral Analysis of NYC Shootings", log="no")
```




```
x.logdif1 <- spectrum(shootings_ts1, main= "Log Spectral Analysis of NYC Shootings", log="no") # The tr
```


Log Spectral Analysis of NYC Shootings
spectrum
0.00
0.04
0.08
0.12
0
1
2
3
4
5
6
frequency
bandwidth = 0.016