

Stat 153 Midterm 2

Samba Njie Jr., Veronika Yang

4/7/2017

Report

Appendix : Code

Establishing working directory:

```
setwd("/Users/sambamamba/Documents/Cal Spring 2017/STAT_153/MT_2/GoogleTimeSeries")  
  
wd <- getwd(); items <- dir()
```

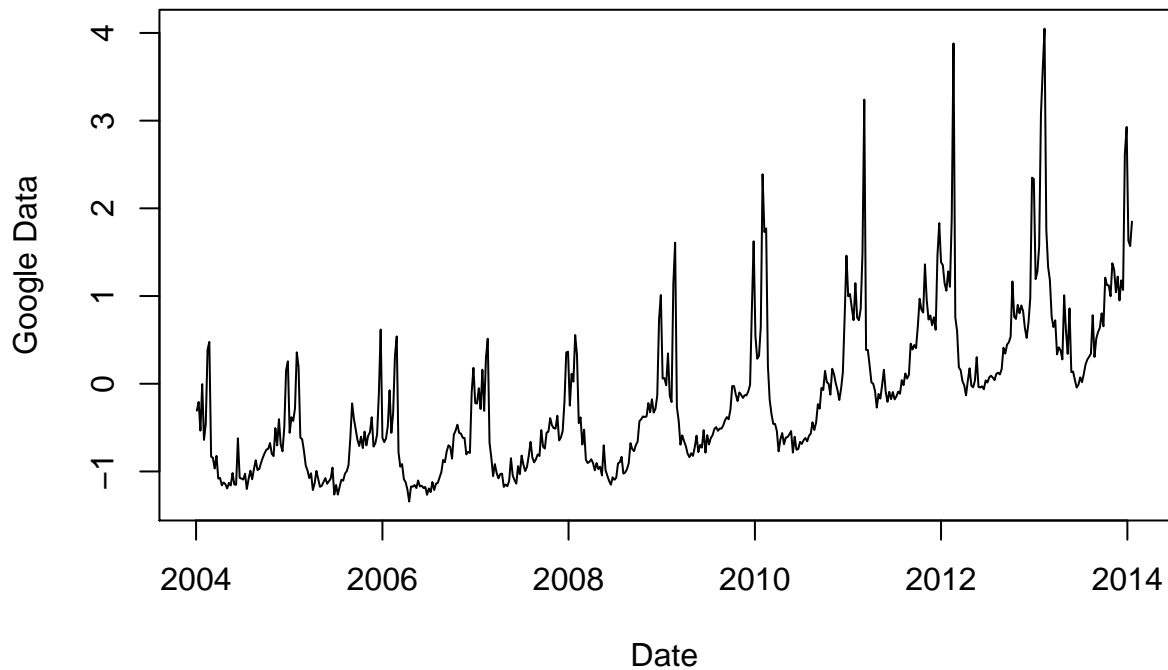
Read in data sets:

```
readData <- function() { # creates a list of the 5 Google data sets  
  dtasets <- items[grepl(".csv", items) == TRUE]  
  dataList <- lapply(dtasets, function(dta) read_csv(file.path(wd, dta)))  
  names(dataList) <- lapply(1:5, function(x) as.vector(paste0("Q",x,"Train")))  
  return(dataList)  
}  
  
data <- readData() # where question i can be found by data[[i]] or data$QiTrain
```

Question 1

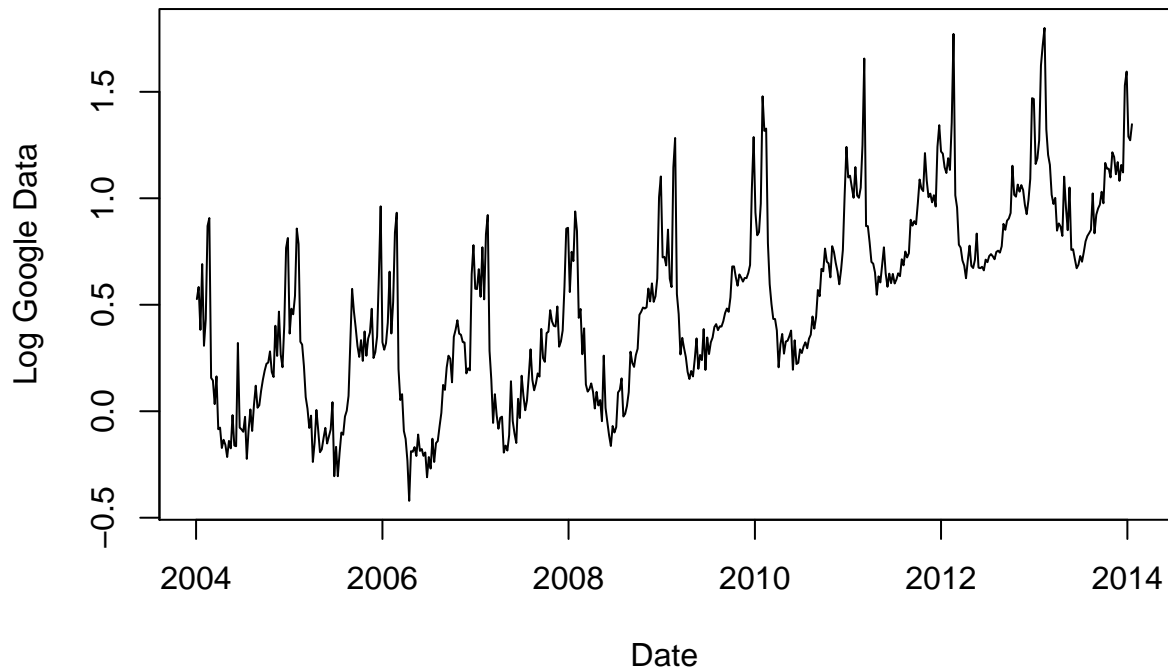
Exploratory Data Analysis

```
Q1Train <- data$Q1Train  
  
plot(Q1Train, type = 'l', xlab = "Date", ylab = "Google Data")
```



There seems to be an increasing linear trend and a clear seasonality in the data set, with a period of around a year. Homoscedasticity in the data set exists. Meaning, as time increases, there seems to be increasing variance in every period. For more convenient analysis and making variance more consistent, we will implement a log transformation of the data. However while log transformation reduces homoskedasticity, logarithms return NaN values with negative data. since the minimum data point in this question is -1.3435, we will shift the data by 2, then perform a log transform, as can be seen in the plot:

```
Q1Train.Log <- data.frame(Date = Q1Train$Date, Activity = log(Q1Train$activity + 2))
plot(Q1Train.Log, type = 'l', xlab = "Date", ylab = "Log Google Data")
```



With the shifted log data at hand, we have reduced homoskedasticity extensively. Now, we must remove the trend by using differencing, aspiring to achieve a stationary data set.

```

qltrain.log <- Q1Train.Log$Activity
#difference <- function(dta, lag.input = 1, order = 1) {
  # Performs differencing for any degree of regular differencing
  #time <- dta[,1]; ts <- dta[,2];
  #ts.out <- diff(ts, lag = lag.input, differences = order)
  #return(data.frame(Date = time[(order + 1):length(time)], Activity = ts.out))
#}

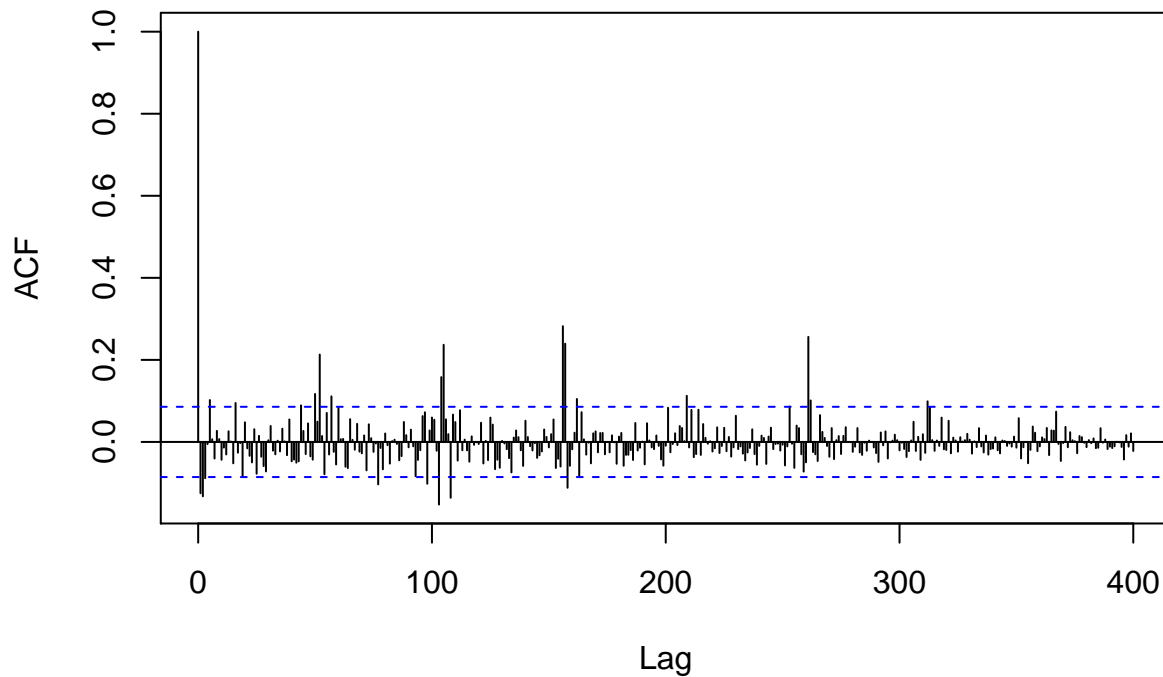
acfIndex <- function(vec, n.max = 1, mod = NA) {
  # input : vector of acf values; output : data frame of the top n.max values and their indices
  stopifnot(n.max <= length(vec));
  val <- rep(NA, n.max); idx <- rep(NA, n.max)
  for (i in 1:n.max) {
    val[i] <- max(vec); idx[i] <- which.max(vec)
    vec <- vec[-idx[i]]
  }
  if (is.na(mod) == FALSE) {
    mod.vec <- idx %% mod
    return(data.frame(index = idx, value = val, remainder = mod.vec))
  }
  return(data.frame(index = idx, value = val))
}

# Observe first and second differenced log data
firstdiff <- diff(qltrain.log)
seconddiff <- diff(qltrain.log, differences = 2)
thirddiff <- diff(qltrain.log, differences = 3)

checkSeas <- acfIndex(acf(firstdiff, lag.max = 400)$acf, n.max = 20, mod = 52)

```

Series firstdiff



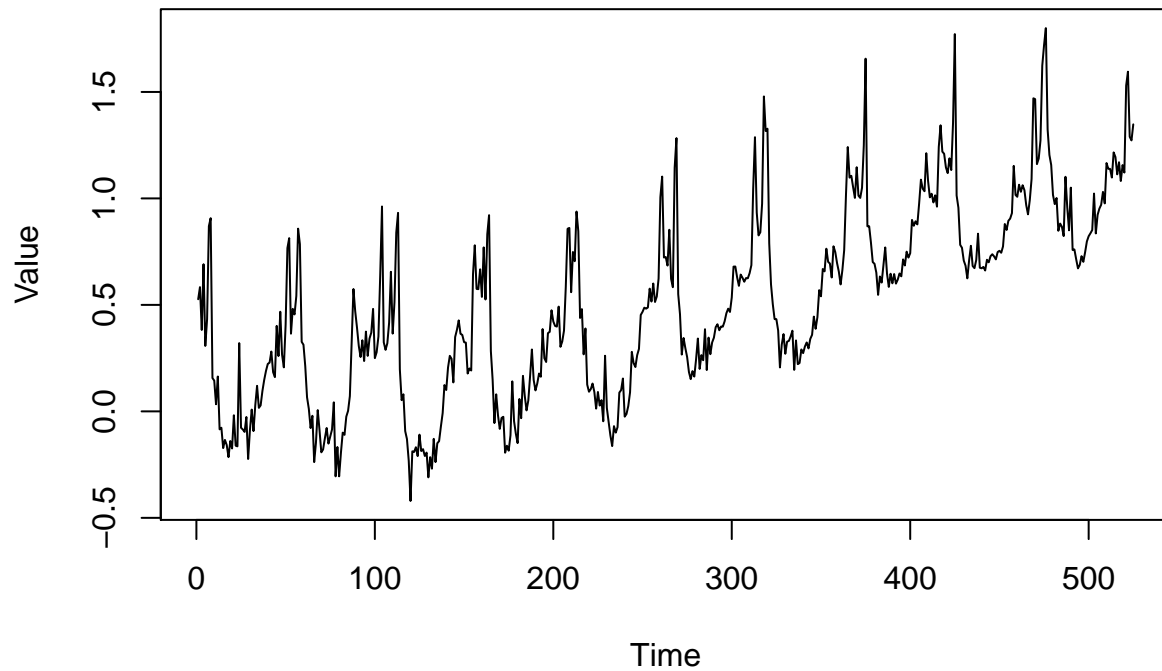
```
# check if yearly
# the acfIndex indicates indices and values of the n highest acf values, and found a lot of them are yearly
# seasonal data differencing
print(checkSeas)
```

##	index	value	remainder
## 1	1	1.00000000	1
## 2	156	0.28215879	0
## 3	260	0.25600402	0
## 4	156	0.23953679	0
## 5	105	0.23686474	1
## 6	52	0.21301361	0
## 7	103	0.15798856	51
## 8	50	0.11698184	50
## 9	203	0.11235940	47
## 10	55	0.11112628	3
## 11	155	0.10472789	51
## 12	5	0.10215912	5
## 13	251	0.10111736	43
## 14	300	0.09900733	40
## 15	15	0.09504322	15
## 16	42	0.08916602	42
## 17	241	0.08642917	33
## 18	297	0.08526497	37
## 19	190	0.08322715	34
## 20	54	0.08272026	2

```
firstdiff.52 <- diff(firstdiff, lag = 52)
```

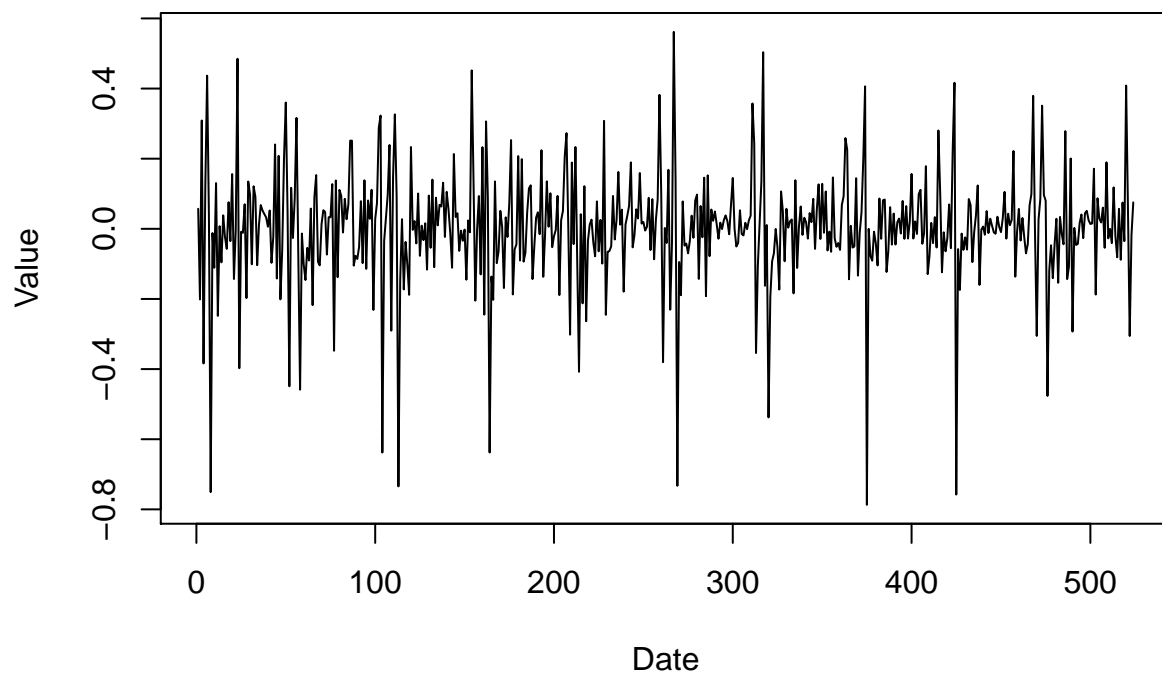
```
# Observe differenced data of orders 1,2
#par(mfrow = c(2,1))
plot(q1train.log, type = 'l', xlab = "Time", ylab = "Value", main = "Undifferenced Google Data")
```

Undifferenced Google Data

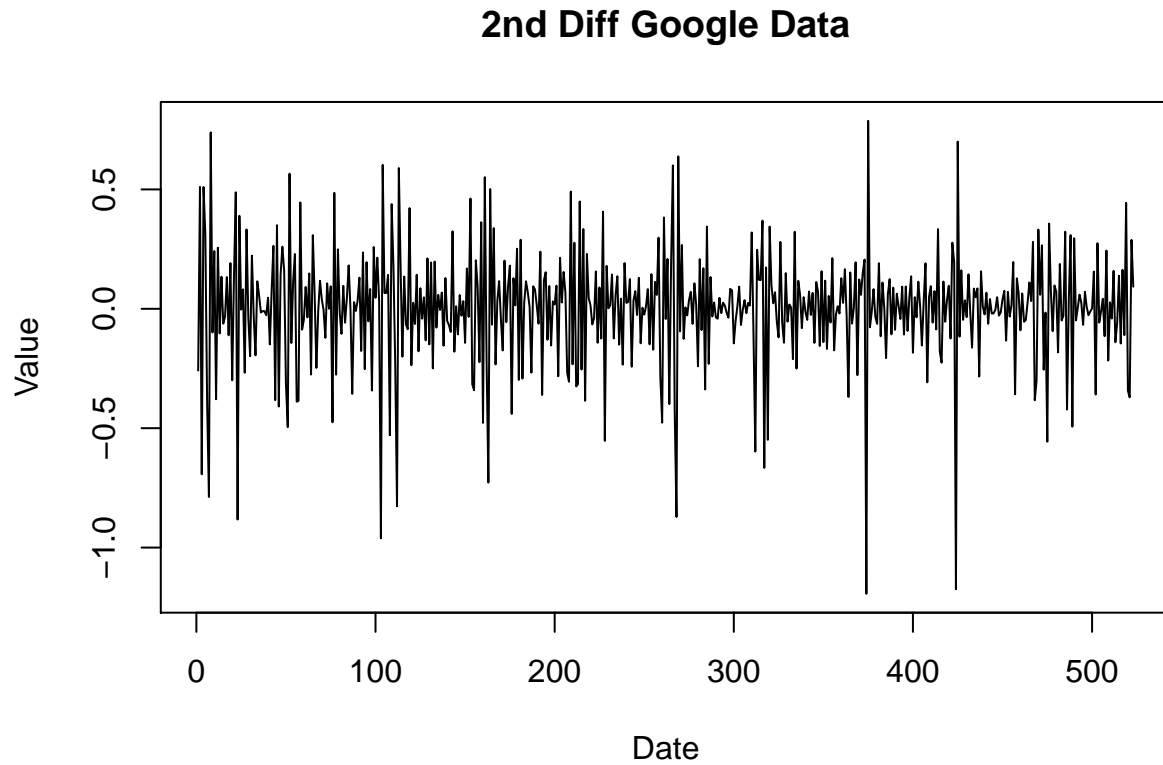


```
plot(firstdiff, type = 'l', xlab = "Date", ylab = "Value", main = "1st Diff Google Data");
```

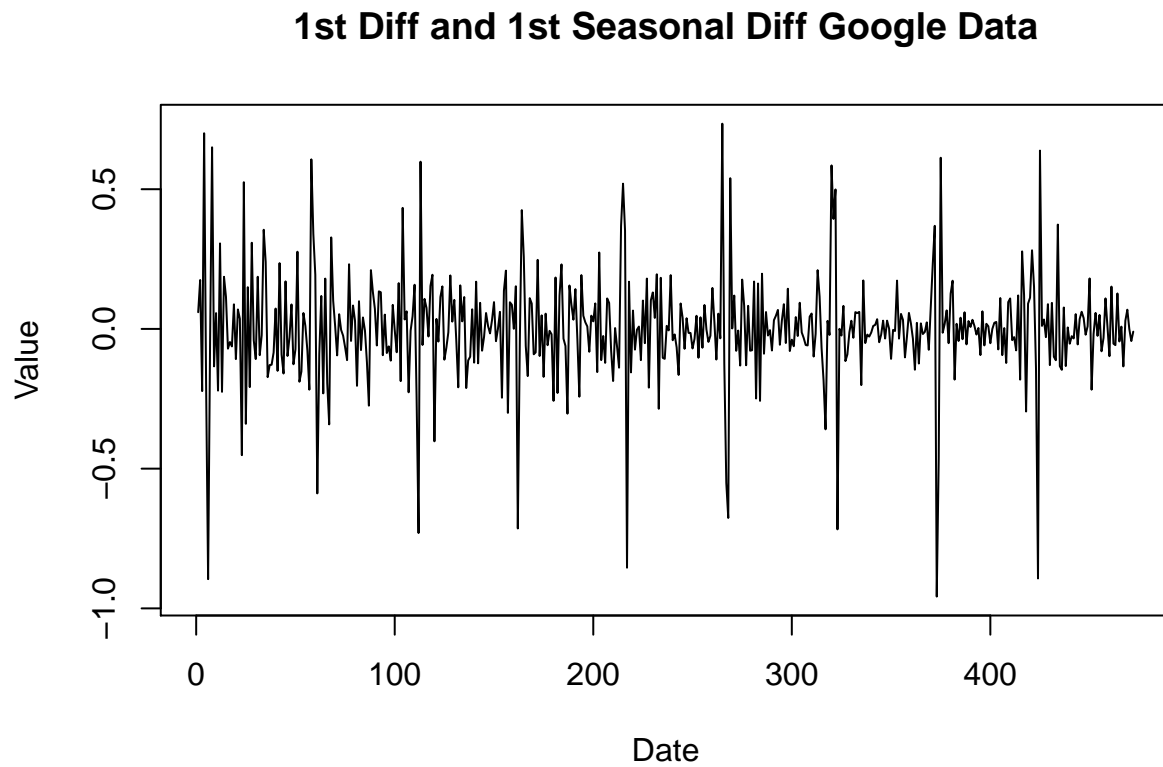
1st Diff Google Data



```
plot(seconddiff, type = 'l', xlab = "Date", ylab = "Value", main = "2nd Diff Google Data")
```

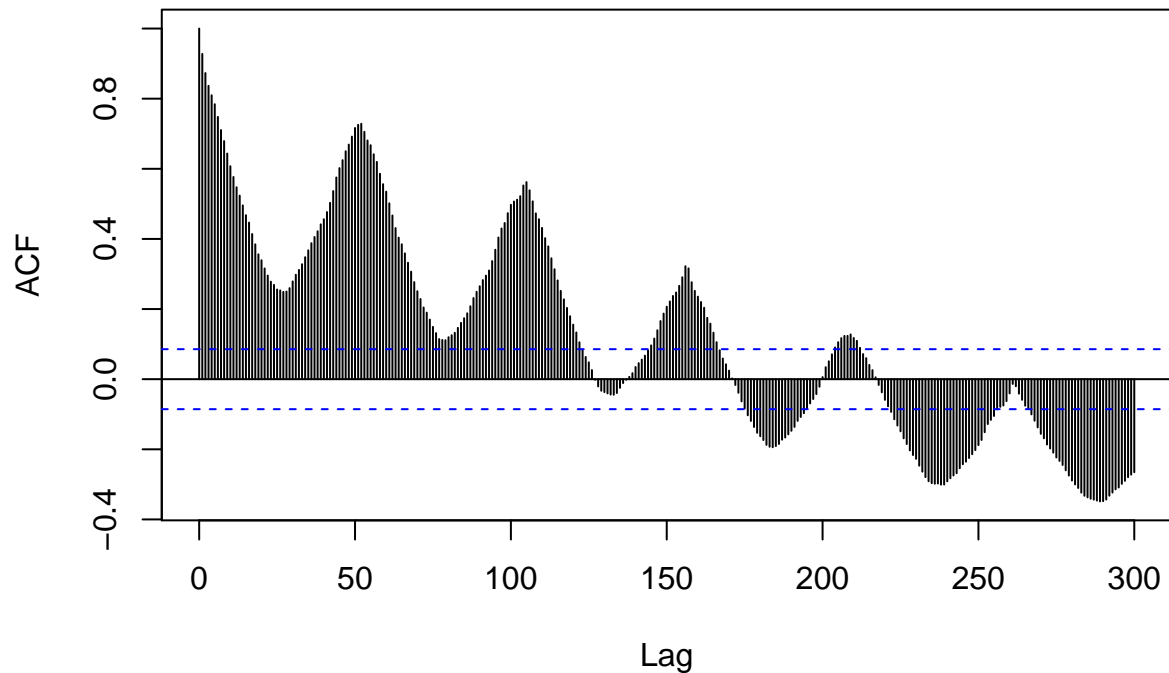


```
plot(firstdiff.52, type = 'l', xlab = "Date", ylab = "Value", main = "1st Diff and 1st Seasonal Diff Google Data")
```

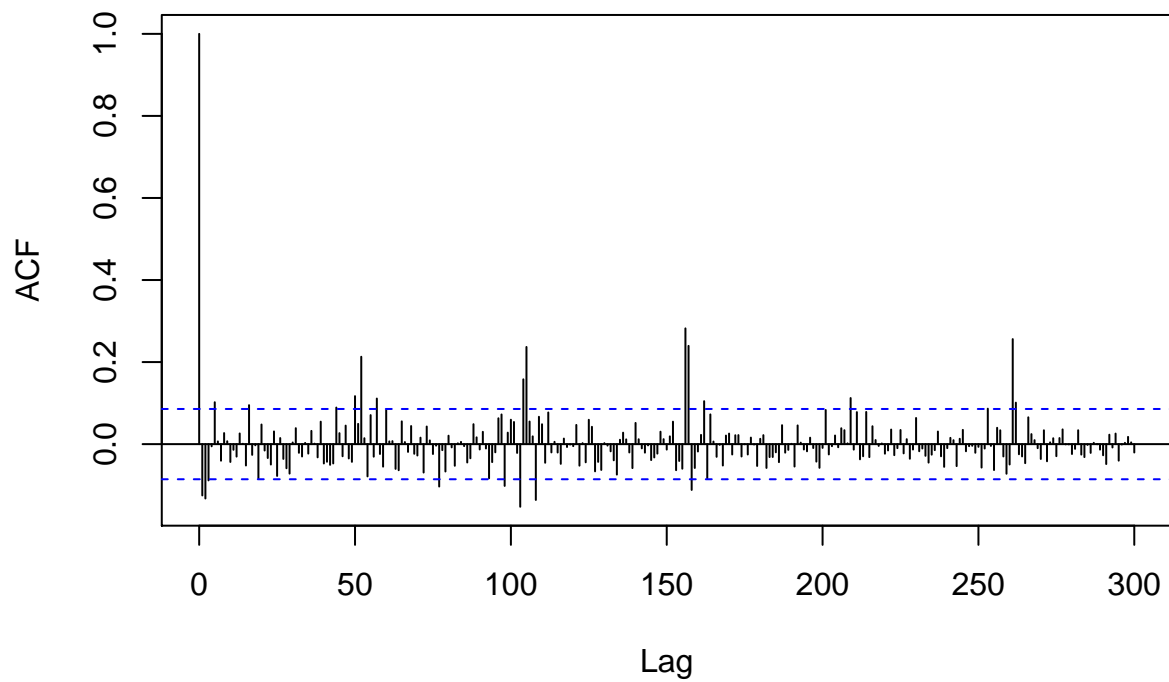


```
# Observe acf of data of orders 1, 2
acf(q1train.log, lag.max = 300); acf(firstdiff, lag.max = 300); acf(seconddiff, lag.max = 300); acf(fir
```

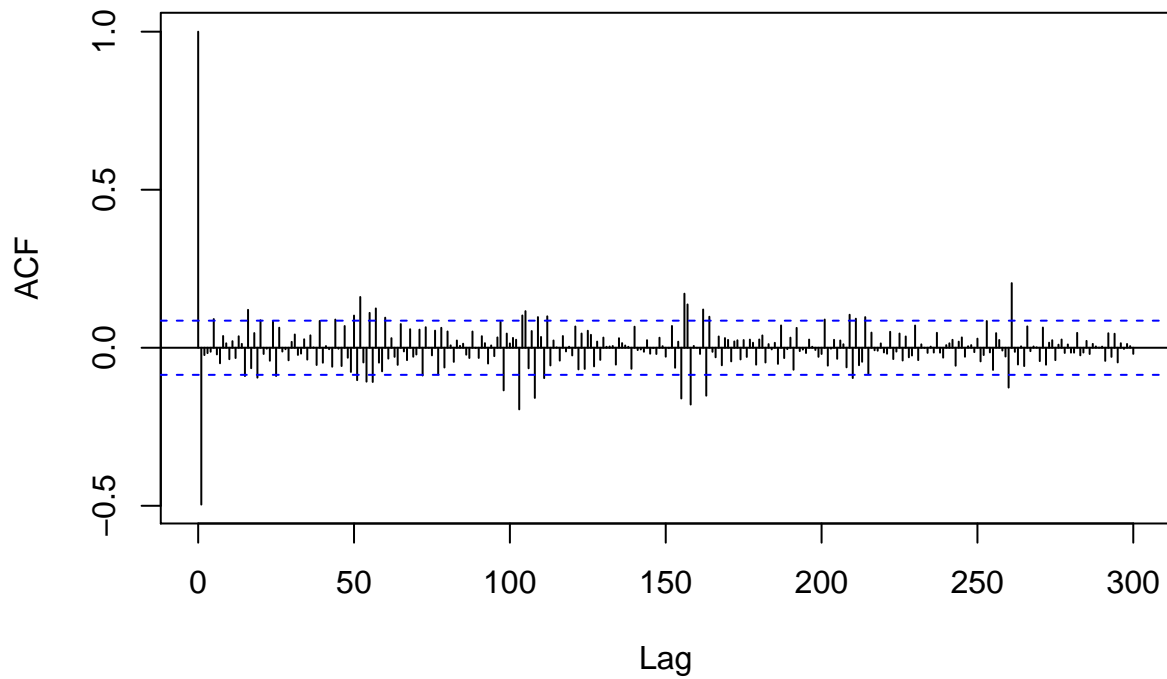
Series q1train.log



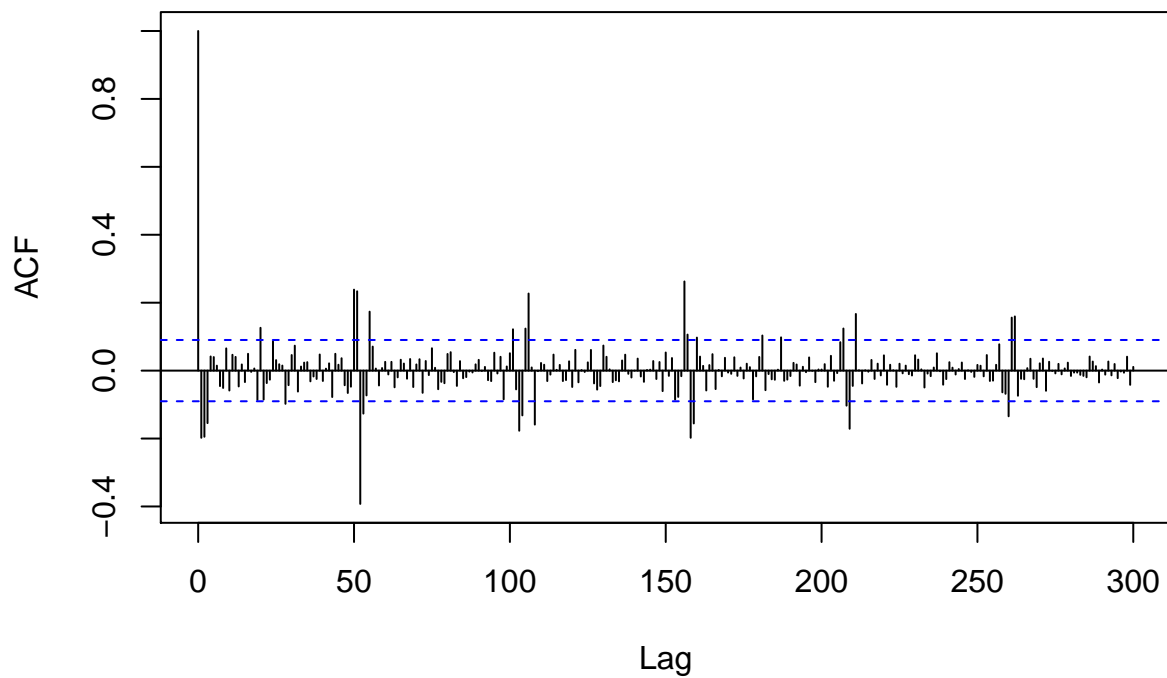
Series firstdiff



Series seconddiff

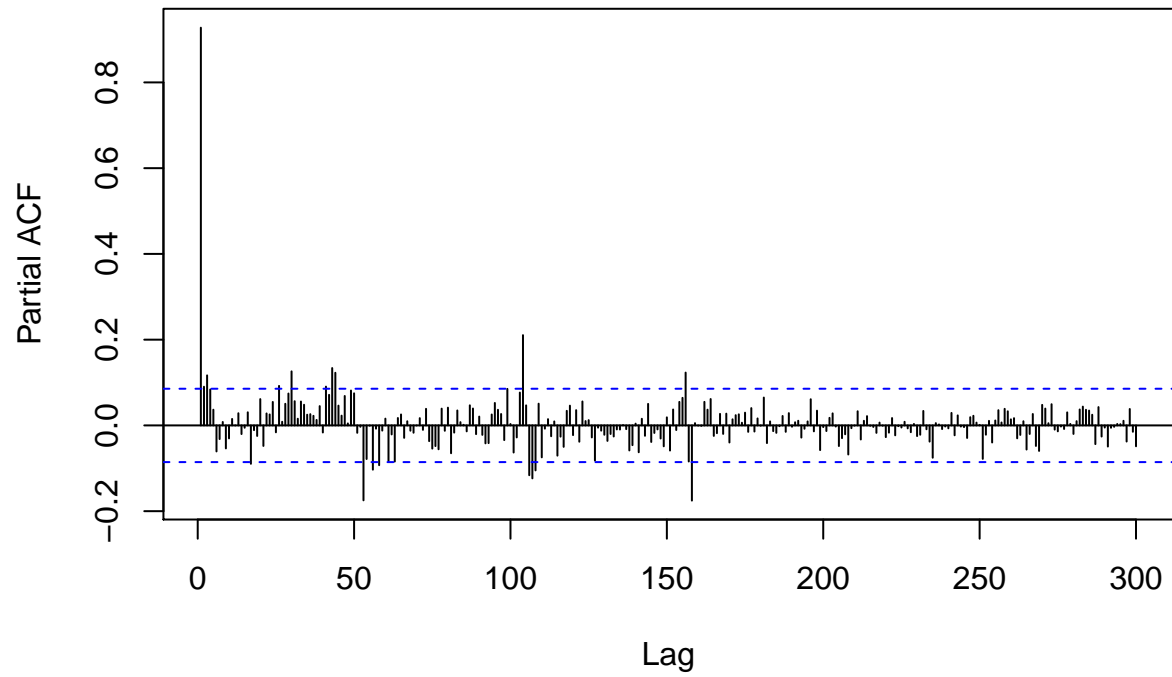


Series firstdiff.52

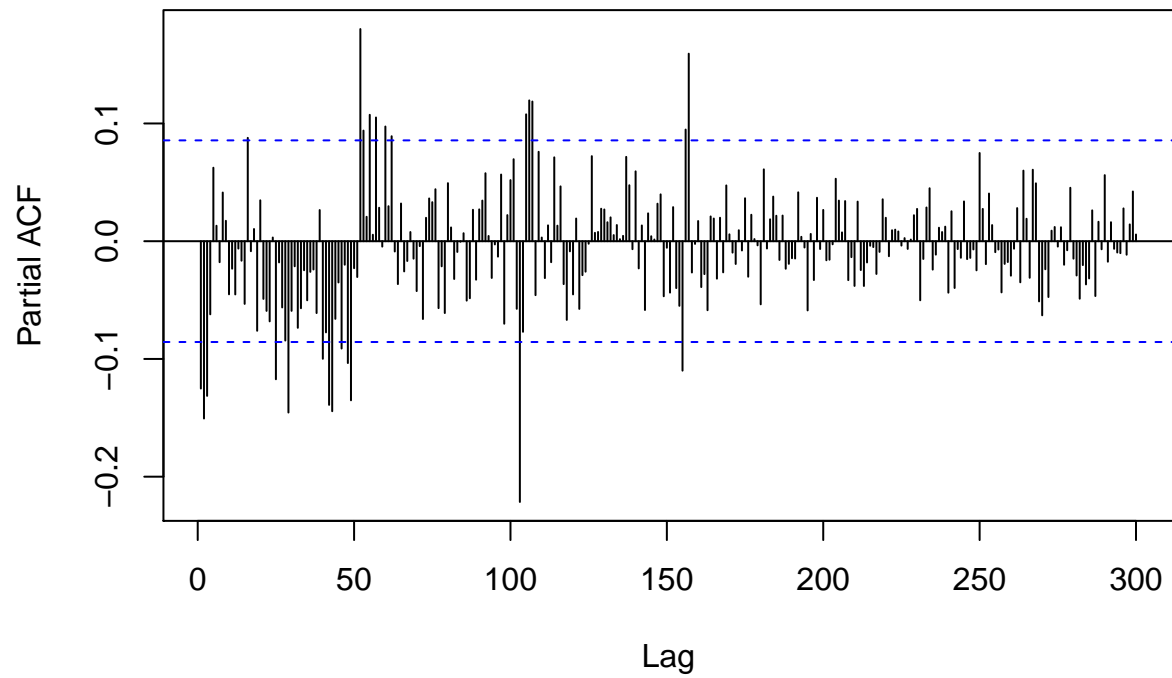


```
pacf(q1train.log, lag.max = 300); pacf(firstdiff, lag.max = 300); pacf(seconddiff, lag.max = 300); pacf
```

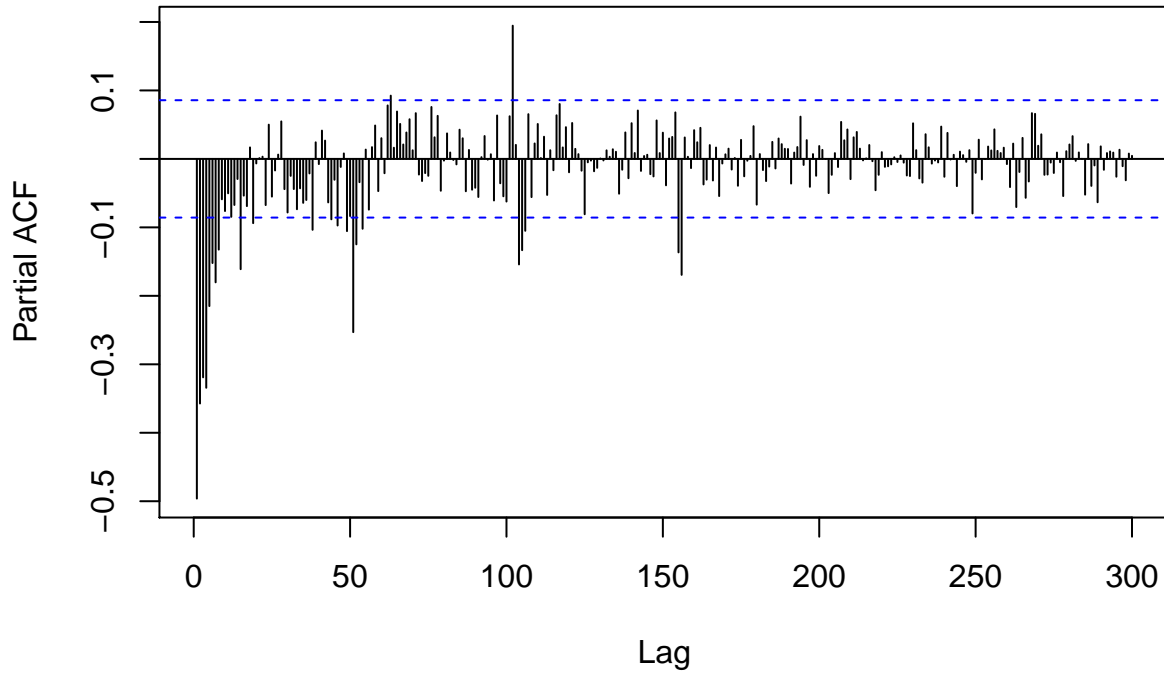

Series q1train.log



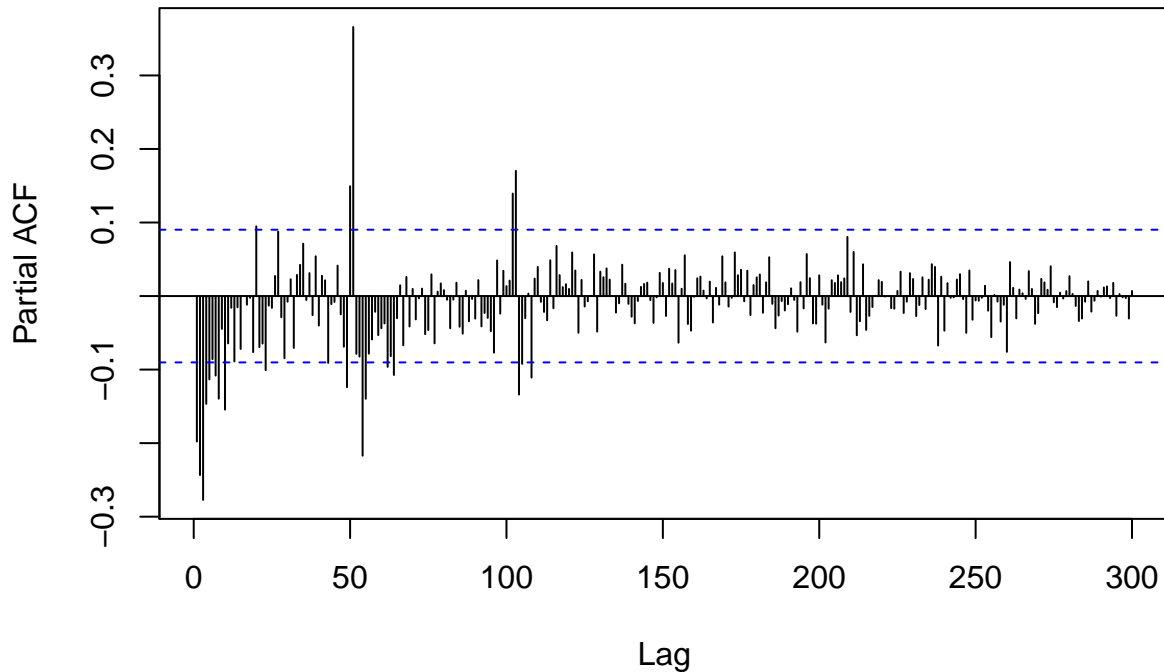
Series firstdiff



Series seconddiff



Series firstdiff.52



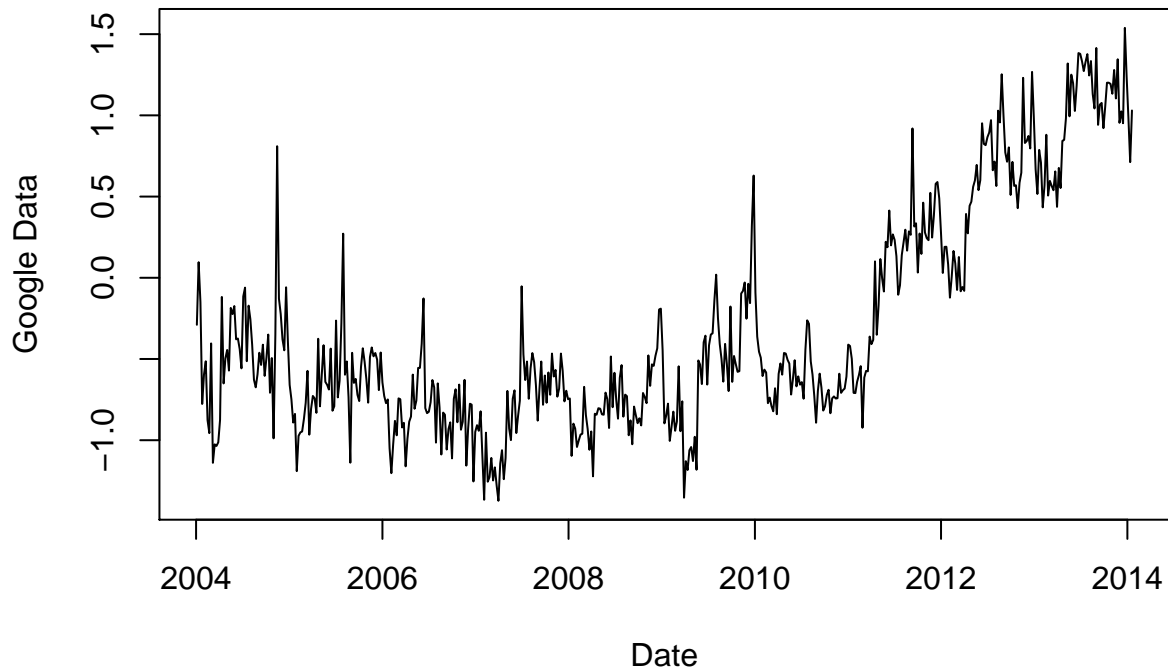
From the plots above, observe that with the second-order differenced data, there is significantly more higher magnitude of lags than the first-order differenced data. As such, we assume that the second-order differencing over-differences the data due to an increase in the breadth of lags with higher magnitudes, and also the existence of lags with higher than -0.5 magnitudes, not seen in the first-order plot. Also, notice that since the lag-1 acf signal is negative (slightly less than -0.5), we have a non-zero degree MA component to our model.

- residuals of plot
- doing adjusted R^2 , AIC, BIC, differencing
- polynomial, exponential fit

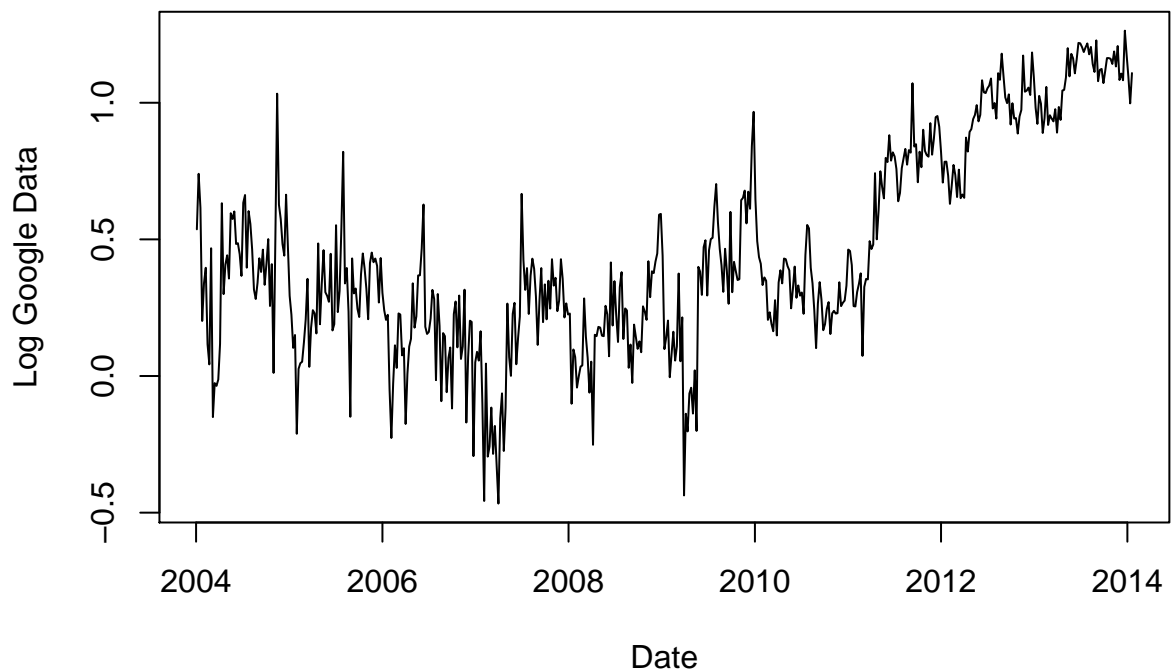
Question 4

```
Q4Train <- data$Q4Train
```

```
plot(Q4Train, type = 'l', xlab = "Date", ylab = "Google Data")
```



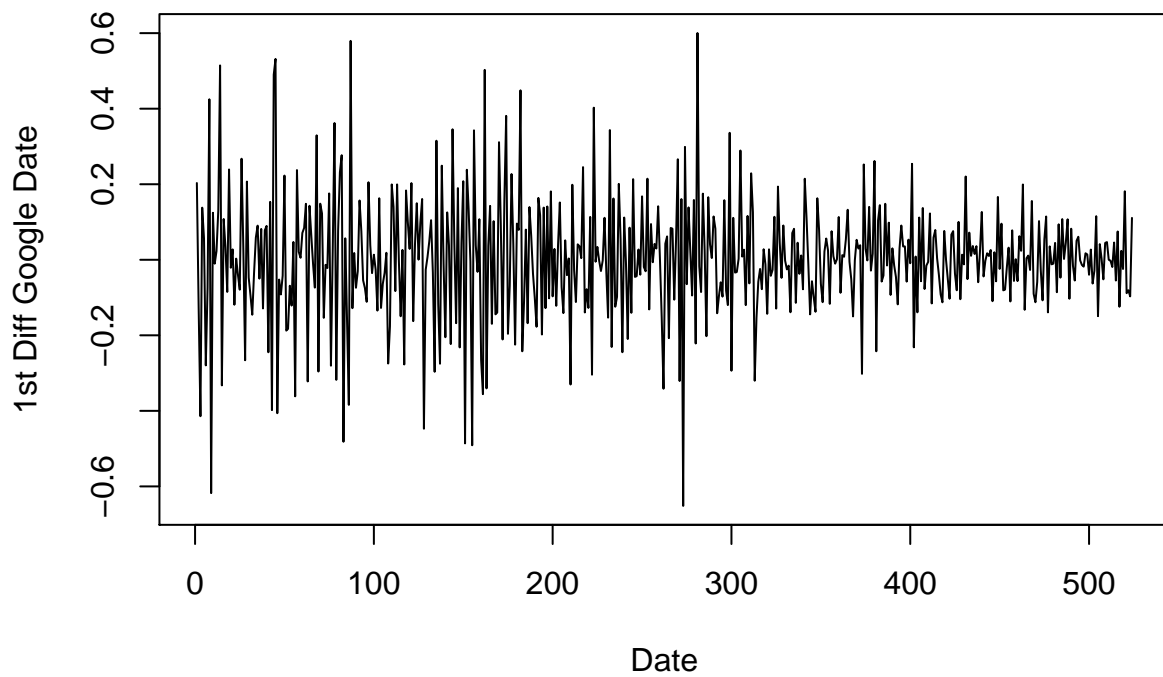
```
Q4Train.Log <- data.frame(Date = Q4Train$Date, Activity = log(Q4Train$activity + 2))  
plot(Q4Train.Log, type = 'l', xlab = "Date", ylab = "Log Google Data")
```



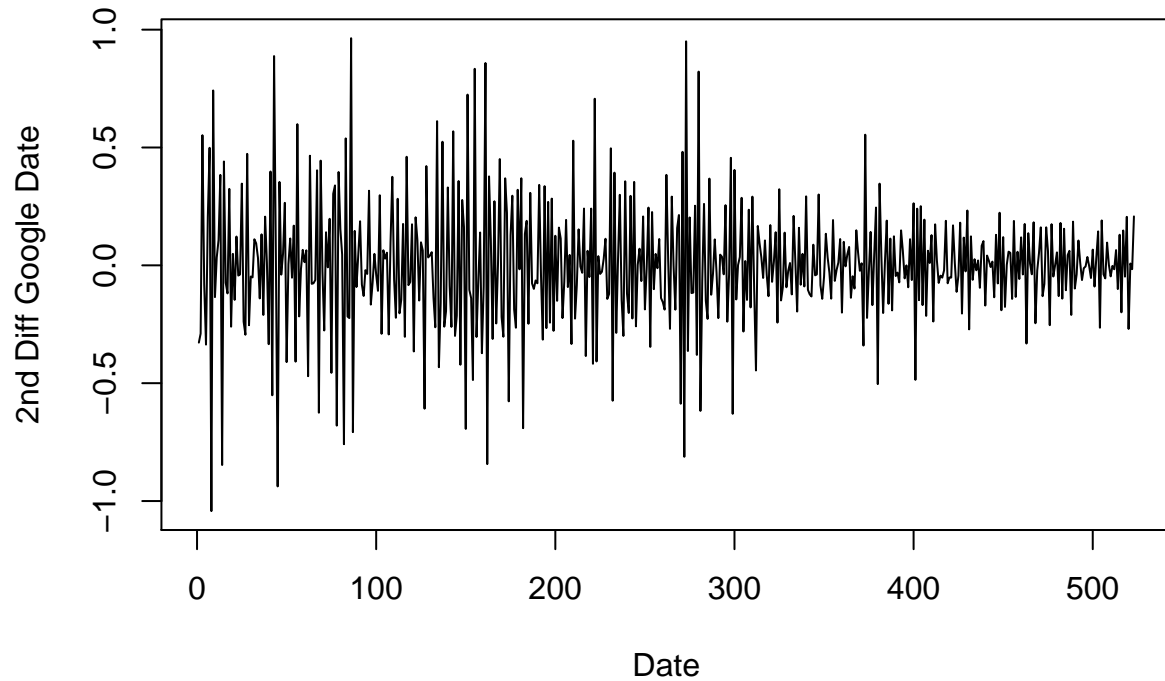
```
q4train.log <- Q4Train.Log$Activity

# Observe first and second differenced log data
q4firstdiff <- diff(q4train.log)
q4seconddiff <- diff(diff(q4train.log))

# Observe differenced data of orders 1,2
plot(q4firstdiff, type = 'l', xlab = "Date", ylab = "1st Diff Google Date");
```

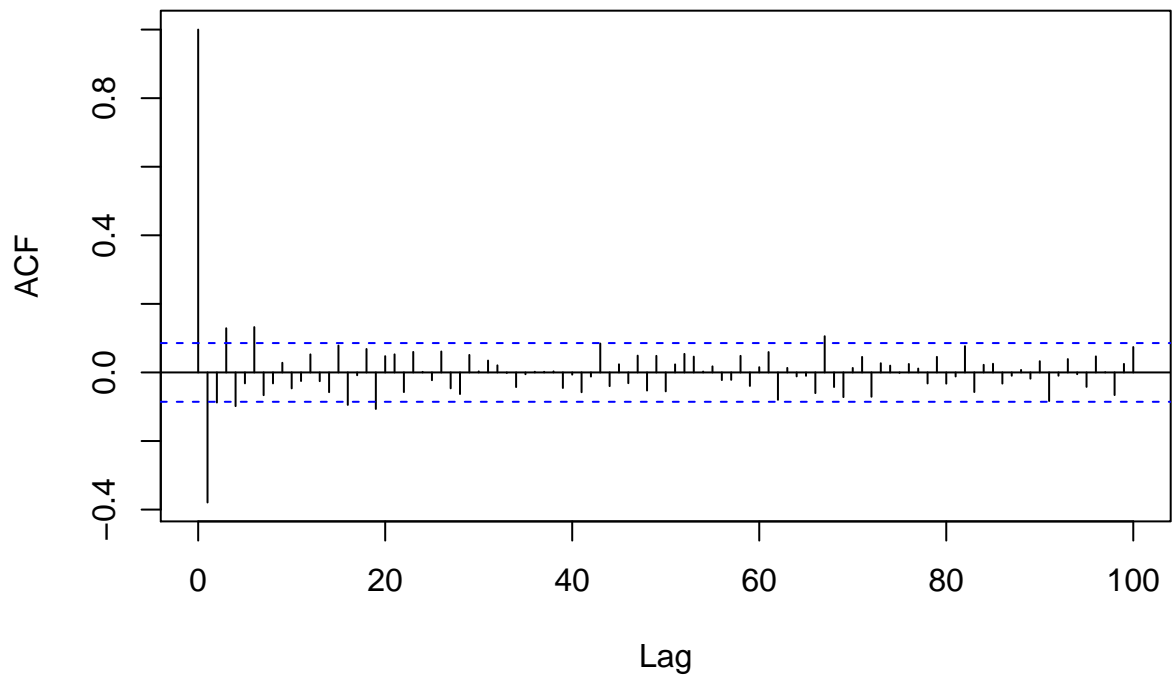


```
plot(q4seconddiff, type = 'l', xlab = "Date", ylab = "2nd Diff Google Date")
```



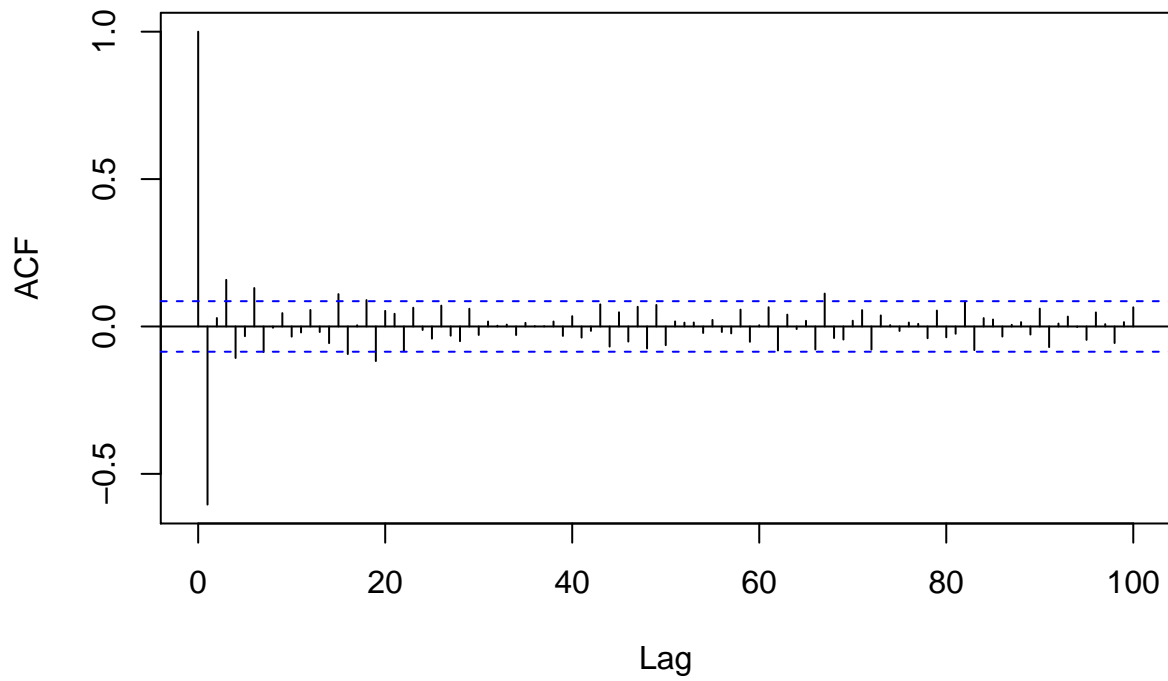
```
# Observe acf of data of orders 1, 2
#par(mfrow = c(2,1))
acf(q4firstdiff, lag.max = 100)
```

Series q4firstdiff



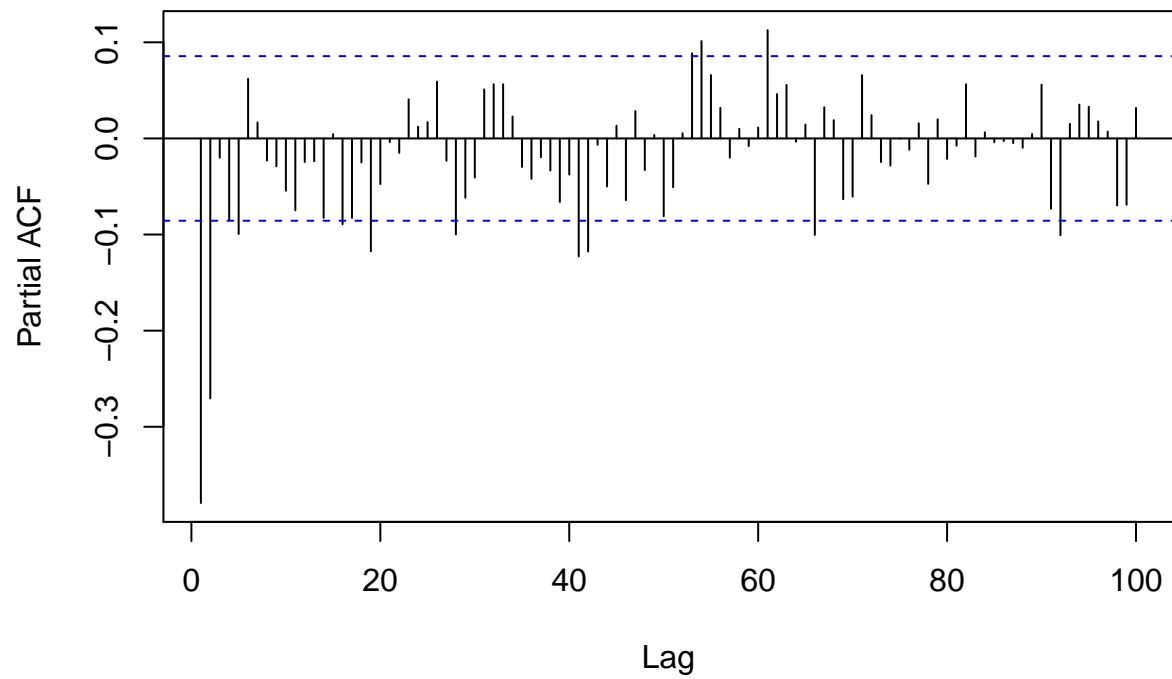
```
acf(q4seconddiff, lag.max = 100)
```

Series q4seconddiff



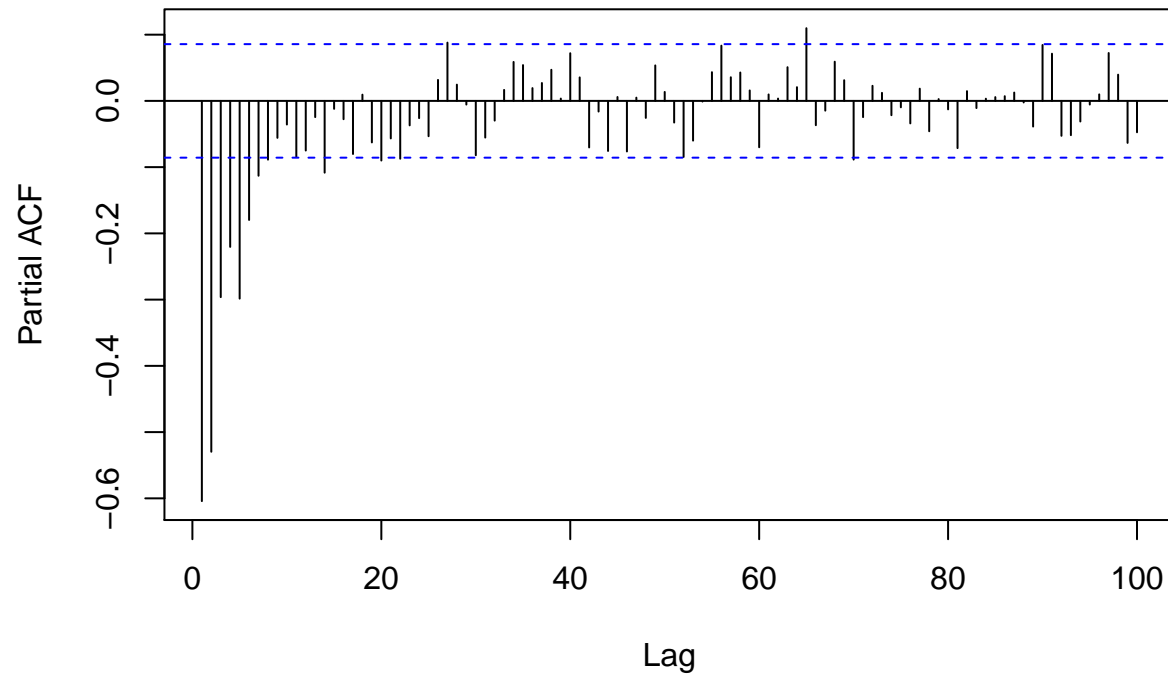
```
#par(mfrow = c(2,1))  
pacf(q4firstdiff, lag.max = 100)
```

Series q4firstdiff



```
pacf(q4seconddiff, lag.max = 100)
```

Series q4seconddiff



Creating the Submission File

```
writeData <- function(dataset, q.num, firstname, lastname, SID) {  
  output <- write.table(dataset,  
    sep = ",",  
    col.names = FALSE,  
    row.names = FALSE,  
    file = paste0("Q",q.num,"_",  
                  firstname,"_",  
                  lastname,"_",SID))  
  return(output)  
}
```