

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()
```

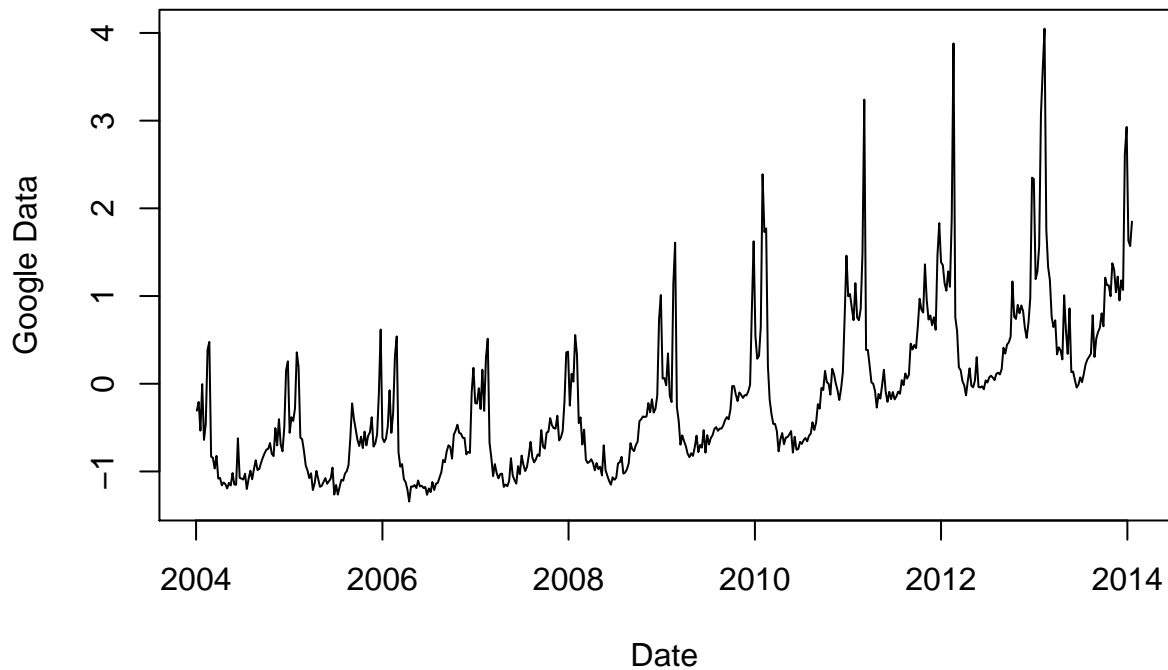
Read in data sets:

```
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

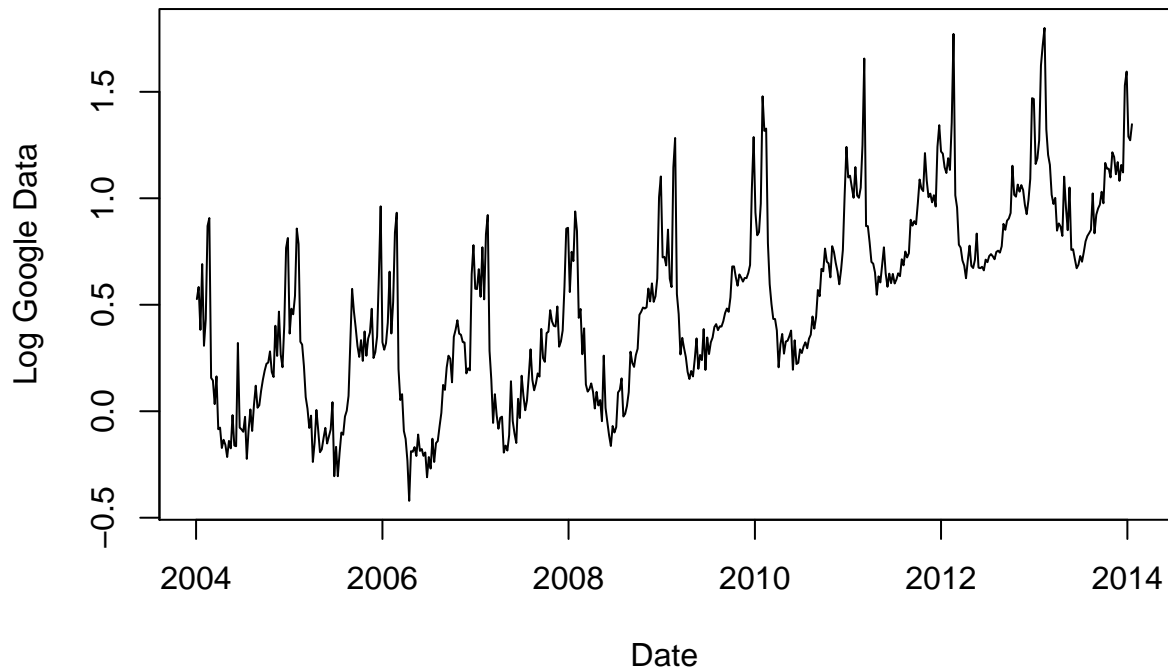
Exploratory Data Analysis

```
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. Homoskedasticity 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.

```

difference <- function(dta, order = 1) {
  # Performs differencing for any degree of regular differencing
  time <- dta[,1]; ts <- dta[,2];
  for (i in 1:order) {
    ts <- diff(ts)
  }
  return(data.frame(Date = time[(order + 1):length(time)], Activity = ts))
}

```

```

# Observe first and second differenced log data

```

```

firstdiff <- difference(q1train.log, 1)
seconddiff <- difference(q1train.log, 2)

```

```

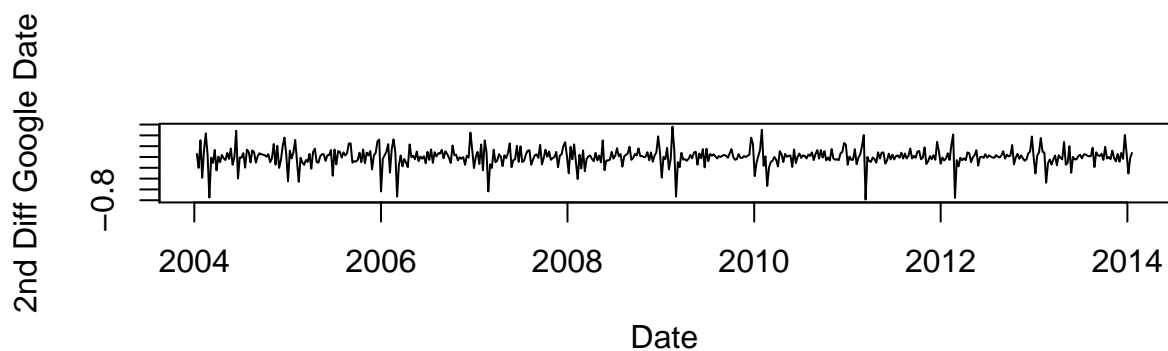
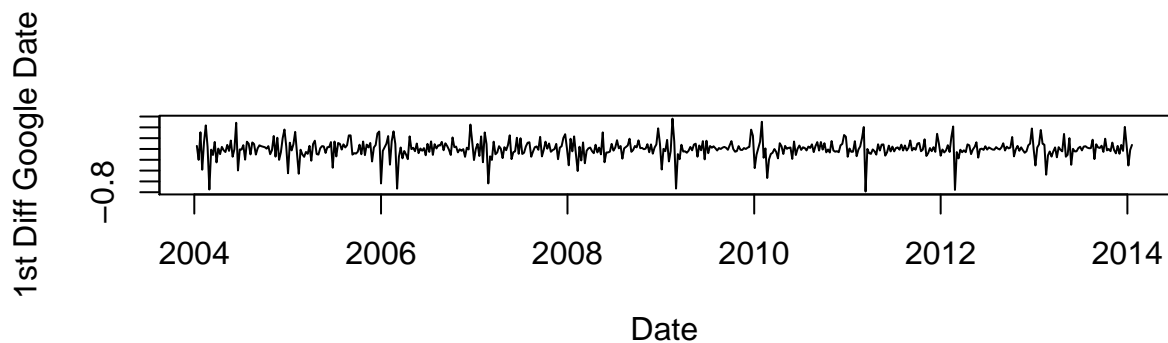
# Observe differenced data of orders 1,2

```

```

par(mfrow = c(2,1))
plot(firstdiff, type = 'l', xlab = "Date", ylab = "1st Diff Google Date");
plot(seconddiff, type = 'l', xlab = "Date", ylab = "2nd Diff Google Date")

```

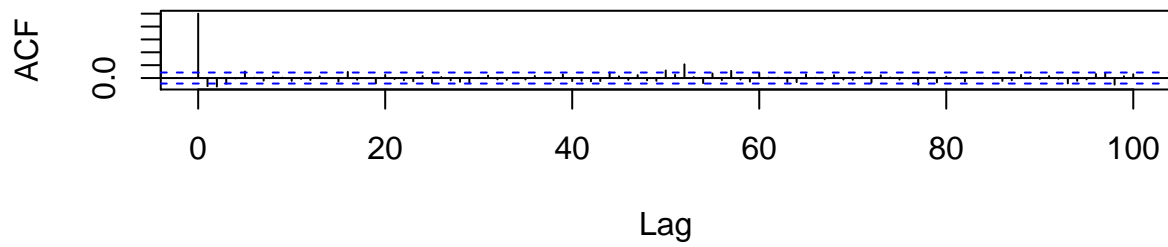


```

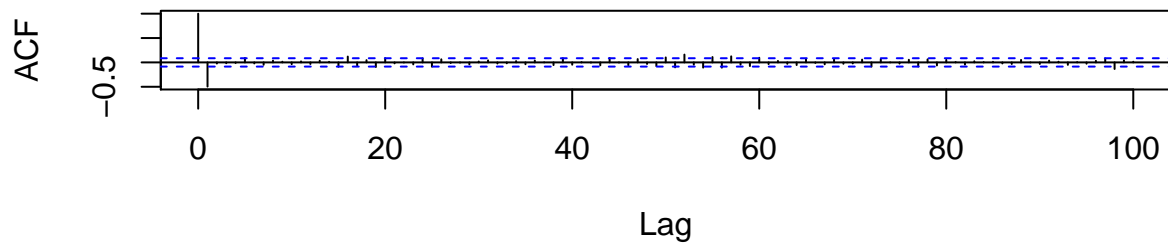
# Observe acf of data of orders 1, 2
par(mfrow = c(2,1))
acf(firstdiff$Activity, lag.max = 100)
acf(seconddiff$Activity, lag.max = 100)

```

Series firstdiff\$Activity



Series seconddiff\$Activity



Next step : see if there is seasonality in differencing

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)  
}
```