Joseph Crockett  
ES 207: Environmental Data Analysis  
April 1st, 2016

**Homework Assignment 7: Time & Space**

**Objective Statement:**  
Purposeful flooding of the Cosumnes River floodplain could stimulate localized recharge of the groundwater aquifer and spur fish populations. However, the expense may not be worth the investment if flooding does not occure often enough. We will examine the discharge levels at the USGS gaging station at Michigan Bar to determine the percentage of years that discharges >= 800 cubic feet per second occur 100 days or more. Groundwater aquifers will benefit if floods occur in 50% of years.

**Methods:**  
Our first goal was to gather data from the California Data Exchange Center with a unique function that extracted streamflow data from a query URL. We next removed leap days and examined monthly autocorrelation, and total, mean, and maximum discharge by month. Finally, the number of flood events (days during which average discharge equals or exceeds 800 cfs) per water year were calculated.

**Data:**  
USGS maintains streamflow records from water year 1908 to the present at approximately 1.5 million sites in the United States and associated commonwealths. We used historical river conditions at Michigan Bar from the California Data Exchange Center for the period 1908 to 2015. Data were converted from CFS to acre-foot / day for time-series analysis.

**Code:**

# The provided function getCDEC encounters problems when the date format "m/d/Y" is used: csv files will not be written with the / symbol. The fixed function, also provided, corrects this with the addition of "as.character(strptime(...))" on lines 53 and 54. An alternative to this would be to use the function "format" to rearrange the start and end times to "m-d-Y" regardless of the input format.  
  
#Extracting data from URL  
get.dailyUSGS <- function(site, sensor, start, end){  
 #site = site number (11335000 for Michigan Bar)  
 #sensor = sensor parameter (00060 for discharge)  
 #start and end values can be in mm-dd-yyyy or mm/dd/yyyy forms  
 start <- format(as.Date(start, "%m/%d/%Y"),"%Y-%m-%d") #format date in form m - d - y for URL  
 print(start)  
 end <- format(as.Date(end,"%m/%d/%Y"), "%Y-%m-%d")  
 print(end)  
 print(site)  
 print(sensor)  
 print(paste0("http://waterdata.usgs.gov/nwis/dv?cb\_",sensor,"=on&format=rdb&site\_no=",site,"&referred\_module=sw&period=&begin\_date=",start,"&end\_date=",end))  
 data <- read.table(paste0("http://waterdata.usgs.gov/nwis/dv?cb\_",sensor,"=on&format=rdb&site\_no=",site,"&referred\_module=sw&period=&begin\_date=",start,"&end\_date=",end), header = T) #header = T removes one of the two unnecessary rows  
   
 assign("daily\_sf", data, envir = .GlobalEnv) #assign to the global environment  
   
}

get.dailyUSGS("11335000", "00060", "10/01/1907", "09/30/2015")

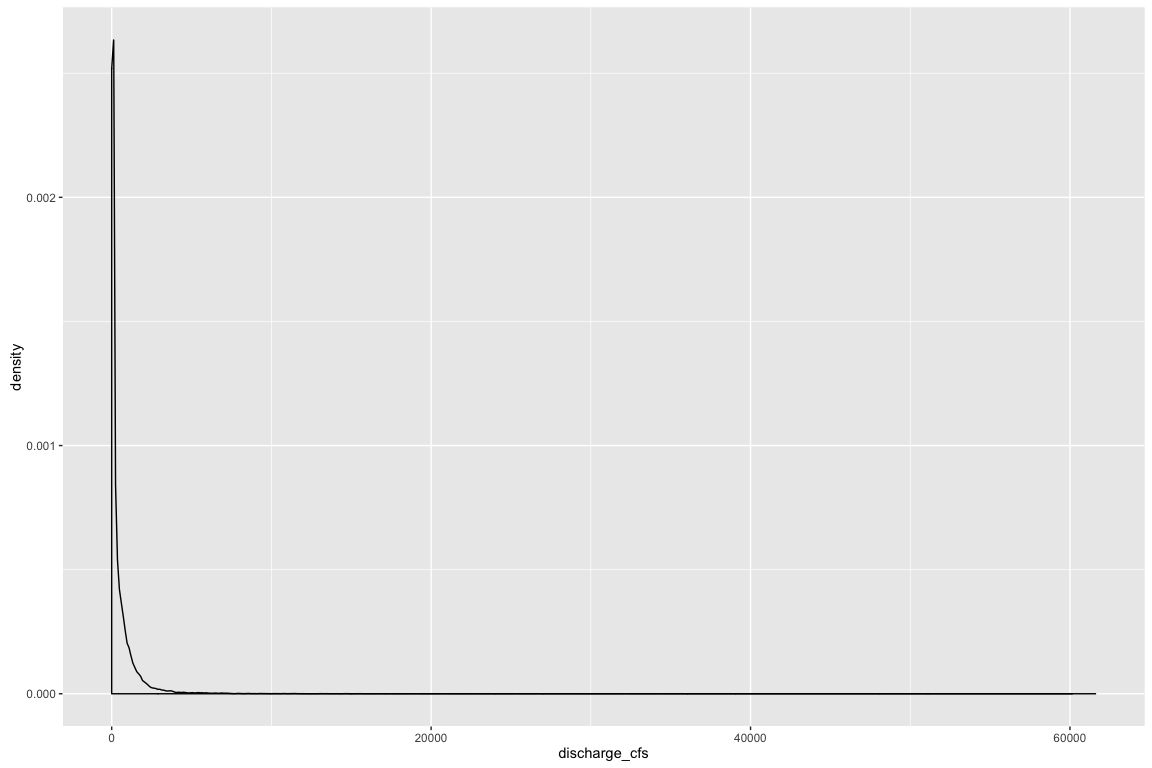
## [1] "1907-10-01"  
## [1] "2015-09-30"  
## [1] "11335000"  
## [1] "00060"  
## [1] "http://waterdata.usgs.gov/nwis/dv?cb\_00060=on&format=rdb&site\_no=11335000&referred\_module=sw&period=&begin\_date=1907-10-01&end\_date=2015-09-30"

water\_dat <- daily\_sf  
water\_dat <- droplevels(water\_dat[-1,]) #remove the first row  
colnames(water\_dat) [4:5] <- c("discharge\_cfs","dat\_qvalue") #rename the last two columns  
  
water\_dat[,"datetime"] <- as.POSIXct(water\_dat[,"datetime"], tz = "America/Los\_Angeles") #convert datetime into POSIXct format, with timezone PST.  
  
water\_dat[,"discharge\_cfs"] <- as.double(as.character(water\_dat[,"discharge\_cfs"])) #convert discharge cfs to numeric(double format)  
  
#From the Nevada division of water resources, 1 cfs for 24 hours = 1.9835 Acre-Feet  
water\_dat[,"discharge\_afd"] <- 1.98347109902 \* water\_dat[,"discharge\_cfs"]

#Add year, month, and day columns, including water year and water month columns  
  
#Extracting year, month, and day values from datetime column  
water\_dat[,"year"]<-as.numeric(format(water\_dat[,"datetime"], format = "%Y"))   
water\_dat[,"month"] <-as.numeric(format(water\_dat[,"datetime"], format = "%m"))  
water\_dat[,"day"] <- as.numeric(format(water\_dat[,"datetime"], format = "%d"))  
  
#water years start October 1 and end september 30, thus for calender months >9, we can add 1 to the year because it will be in the next year. We also reorder months so that october = 1 and september =12  
water\_dat[,"water\_year"] <- ifelse(water\_dat[,"month"] > 9, water\_dat[,"year"] + 1, water\_dat[,"year"])  
water\_dat[,"water\_month"] <- ifelse(water\_dat[,"month"] > 9, water\_dat[,"month"] - 9, water\_dat[,"month"] + 3)  
  
#removing leap days  
water\_dat <- water\_dat[!(water\_dat[,"month"] == 2 & water\_dat[,"day"] == 29),]  
summary(water\_dat)

## agency\_cd site\_no datetime   
## USGS:39420 11335000:39420 Min. :1907-10-01 00:00:00   
## 1st Qu.:1934-09-30 18:00:00   
## Median :1961-09-30 12:00:00   
## Mean :1961-09-30 09:08:24   
## 3rd Qu.:1988-09-30 06:00:00   
## Max. :2015-09-30 00:00:00   
## discharge\_cfs dat\_qvalue discharge\_afd year   
## Min. : 0.0 A :39246 Min. : 0.00 Min. :1907   
## 1st Qu.: 26.0 A:e: 14 1st Qu.: 51.57 1st Qu.:1934   
## Median : 98.0 P : 160 Median : 194.38 Median :1961   
## Mean : 484.3 Mean : 960.66 Mean :1961   
## 3rd Qu.: 525.0 3rd Qu.: 1041.32 3rd Qu.:1988   
## Max. :61600.0 Max. :122181.82 Max. :2015   
## month day water\_year water\_month   
## Min. : 1.000 Min. : 1.00 Min. :1908 Min. : 1.000   
## 1st Qu.: 4.000 1st Qu.: 8.00 1st Qu.:1935 1st Qu.: 3.000   
## Median : 7.000 Median :16.00 Median :1962 Median : 7.000   
## Mean : 6.526 Mean :15.72 Mean :1962 Mean : 6.501   
## 3rd Qu.:10.000 3rd Qu.:23.00 3rd Qu.:1988 3rd Qu.:10.000   
## Max. :12.000 Max. :31.00 Max. :2015 Max. :12.000

# No negative values for discharge  
  
ggplot(water\_dat, aes(discharge\_cfs)) + geom\_density()



#The function ts will create a time series object, ts, with attributes that make time series analysis easier, such as aligning time axes. Yearly totals will have seasonal variation, so the following functions aggregate by month/year  
  
  
water\_dat\_mon\_sum <- ts(as.vector(aggregate(discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = sum)), frequency = 12)  
  
water\_dat\_mon\_mean <- ts(as.vector(aggregate(discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = mean)), frequency = 12)  
  
water\_dat\_mon\_max <- ts(as.vector(aggregate(discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = max)), frequency = 12)  
  
#Determining driest year, wettest year, and year with highest daily average discharge  
wat\_yr\_s <- aggregate(discharge\_afd~water\_year, data = water\_dat, FUN = sum)  
  
which.min(wat\_yr\_s$discharge\_afd) #will show the row number of the minimum year sum of discharge\_afd

## [1] 70

#[1] 70  
wat\_yr\_s[70,] #1977

## water\_year discharge\_afd  
## 70 1977 15764.11

which.max(wat\_yr\_s$discharge\_afd) #similarly, the row number of the max value

## [1] 76

#[1] 76  
wat\_yr\_s[76,] #1983

## water\_year discharge\_afd  
## 76 1983 1221247

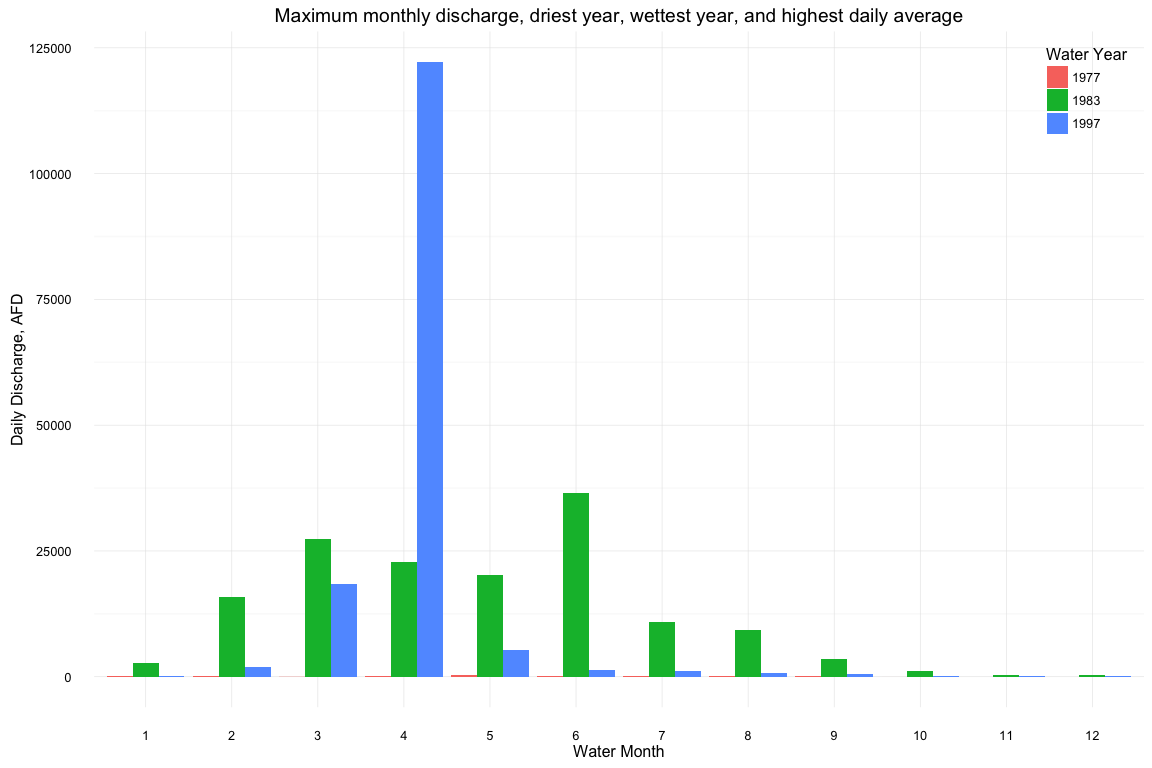
#Interesting that the minimum and the maximum discharge years occur within 10 years of each other  
  
#The year with the highest daily average discharge: Is this the year in which the highest flood occured?  
  
which.max(water\_dat$discharge\_afd)

## [1] 32579

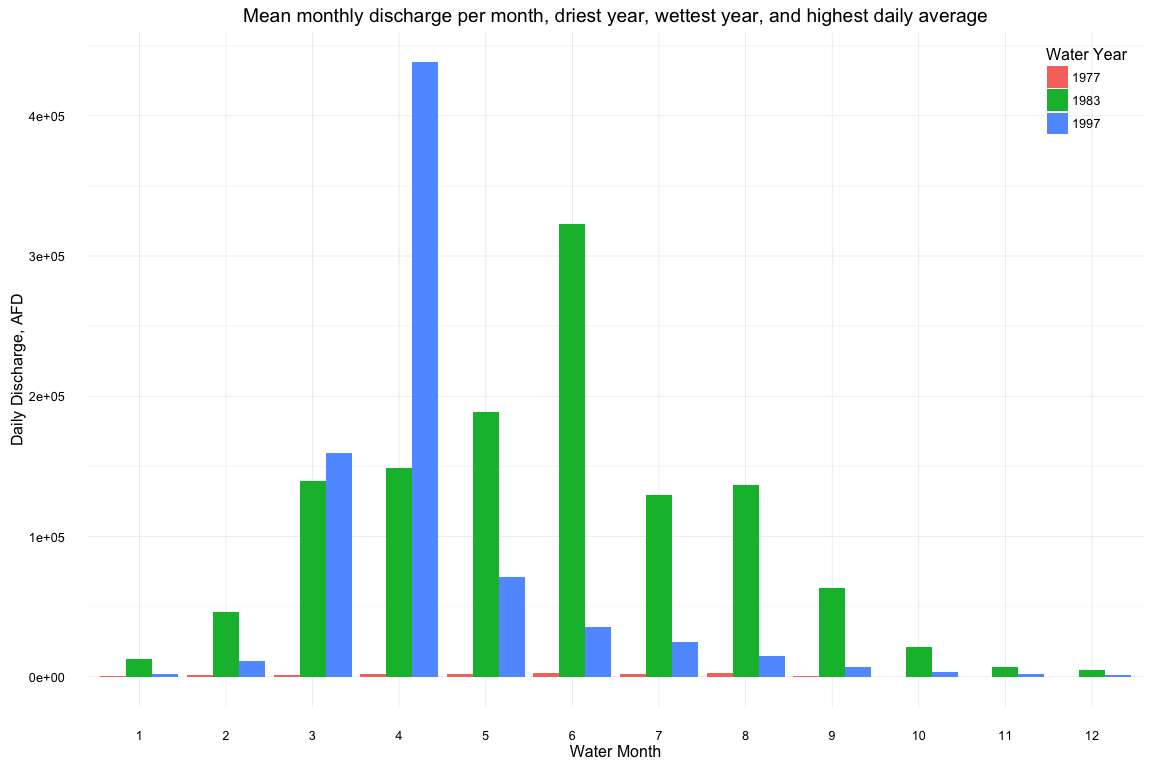
#[1] 32579  
  
water\_dat[32579,"datetime"]

## [1] "1997-01-02 PST"

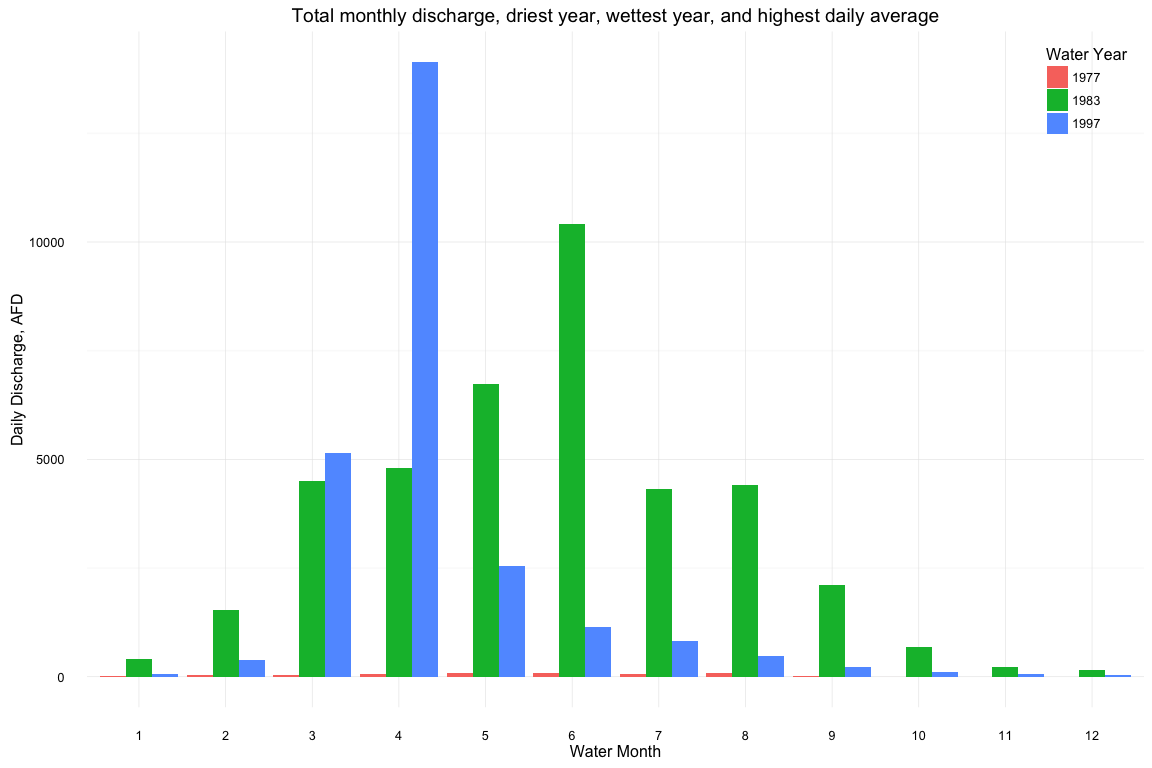
#[1] "1997-01-02 PST"  
  
#Plotting monthly hydrograph results for the above years: 1977, 1983, 1997  
  
d1 <- as.data.frame(water\_dat\_mon\_max)  
d2 <- as.data.frame(water\_dat\_mon\_sum)  
d3 <- as.data.frame(water\_dat\_mon\_mean)  
  
#Max monthly discharge  
df\_max <- melt(list(max\_77 = filter(d1, water\_year == 1977), max\_83 = filter(d1, water\_year == 1983), max\_97 = filter(d1, water\_year == 1997)), id.vars = c("water\_year", "water\_month"))  
  
ggplot(df\_max, aes(x = as.factor(water\_month), y = value, fill = as.factor(water\_year))) + geom\_bar(stat = "identity", position = "dodge") + theme\_minimal() + labs(title = "Maximum monthly discharge, driest year, wettest year, and highest daily average", x = "Water Month", y = "Daily Discharge, AFD") + theme(legend.position = c(1,1), legend.justification = c(1,1)) + scale\_fill\_discrete(name = "Water Year")



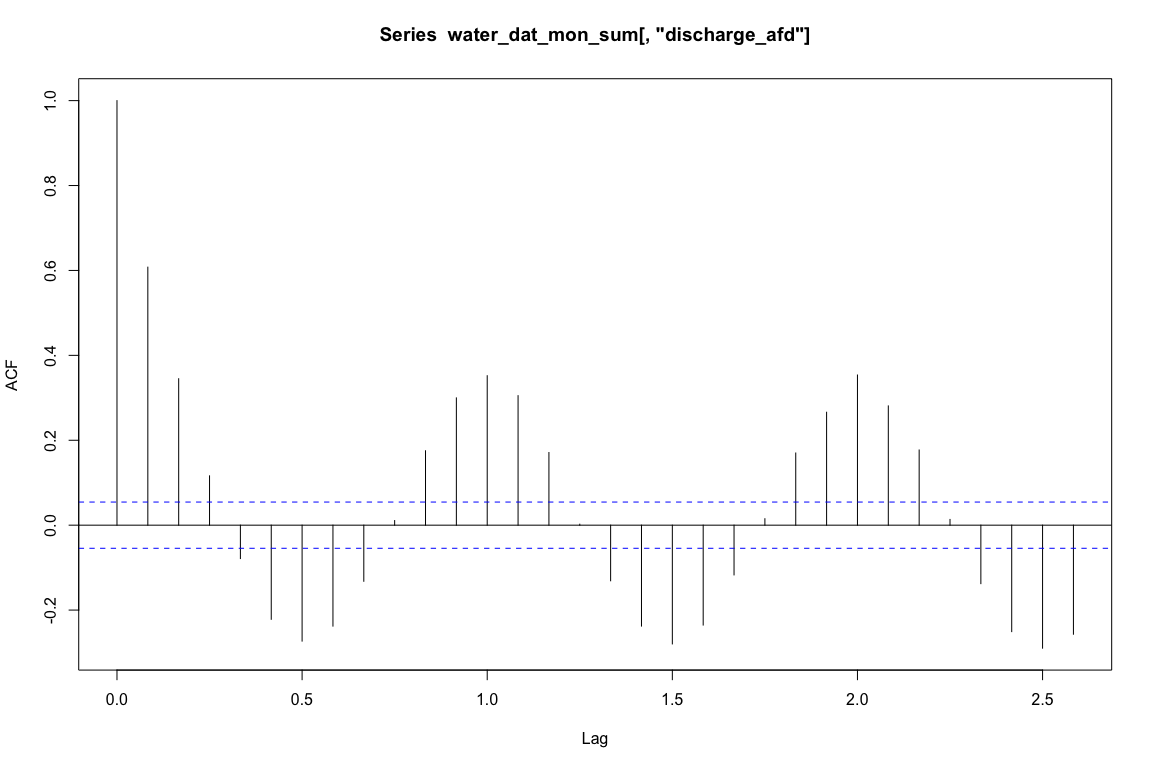
#Mean monthly discharge  
df\_mean <- melt(list(mean\_77 = filter(d2, water\_year == 1977), mean\_83 = filter(d2, water\_year == 1983), mean\_97 = filter(d2, water\_year == 1997)), id.vars = c("water\_year", "water\_month"))  
  
ggplot(df\_mean, aes(x = as.factor(water\_month), y = value, fill = as.factor(water\_year))) + geom\_bar(stat = "identity", position = "dodge") + theme\_minimal() + labs(title = "Mean monthly discharge per month, driest year, wettest year, and highest daily average", x = "Water Month", y = "Daily Discharge, AFD") + theme(legend.position = c(1,1), legend.justification = c(1,1)) + scale\_fill\_discrete(name = "Water Year")



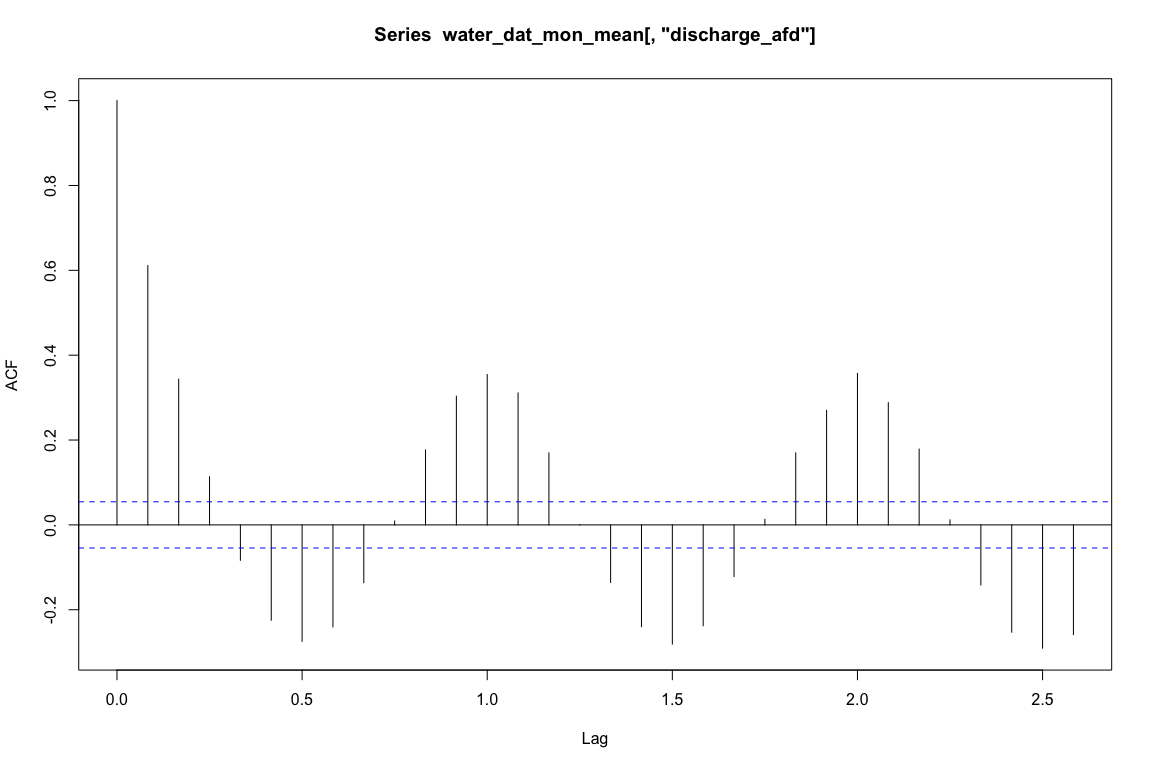
#total monthly discharge  
df\_sum <- melt(list(sum\_77 = filter(d3, water\_year == 1977), sum\_83 = filter(d3, water\_year == 1983), sum\_97 = filter(d3, water\_year == 1997)), id.vars = c("water\_year", "water\_month"))  
  
ggplot(df\_sum, aes(x = as.factor(water\_month), y = value, fill = as.factor(water\_year))) + geom\_bar(stat = "identity", position = "dodge") + theme\_minimal() + labs(title = "Total monthly discharge, driest year, wettest year, and highest daily average", x = "Water Month", y = "Daily Discharge, AFD") + theme(legend.position = c(1,1), legend.justification = c(1,1)) + scale\_fill\_discrete(name = "Water Year")



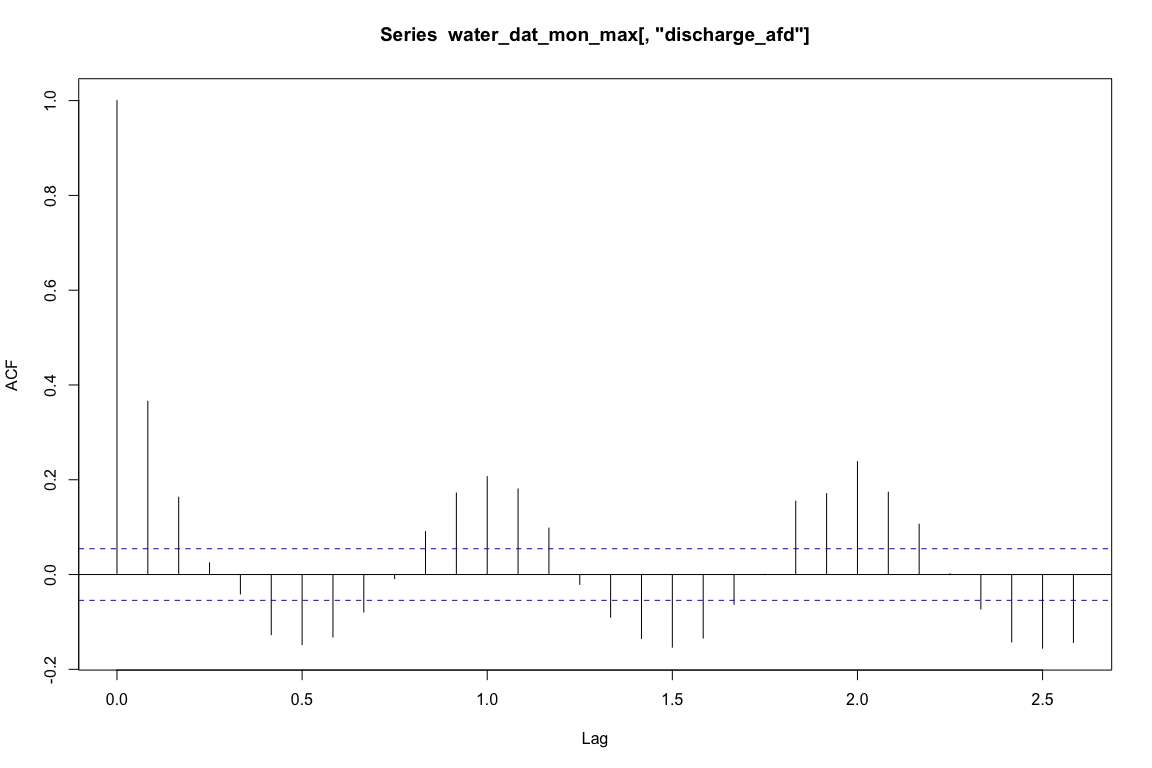
#Autocorrelation functions of sum, mean, and max  
acf(water\_dat\_mon\_sum[,"discharge\_afd"])



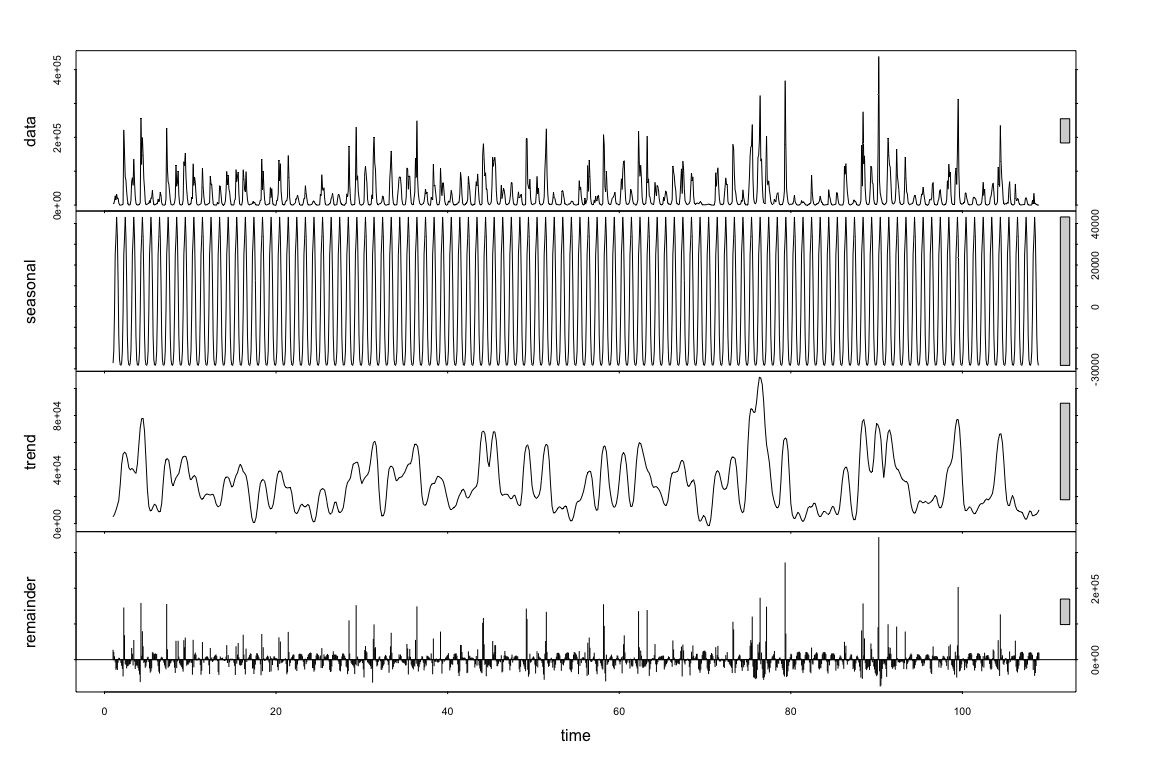
acf(water\_dat\_mon\_mean[,"discharge\_afd"])



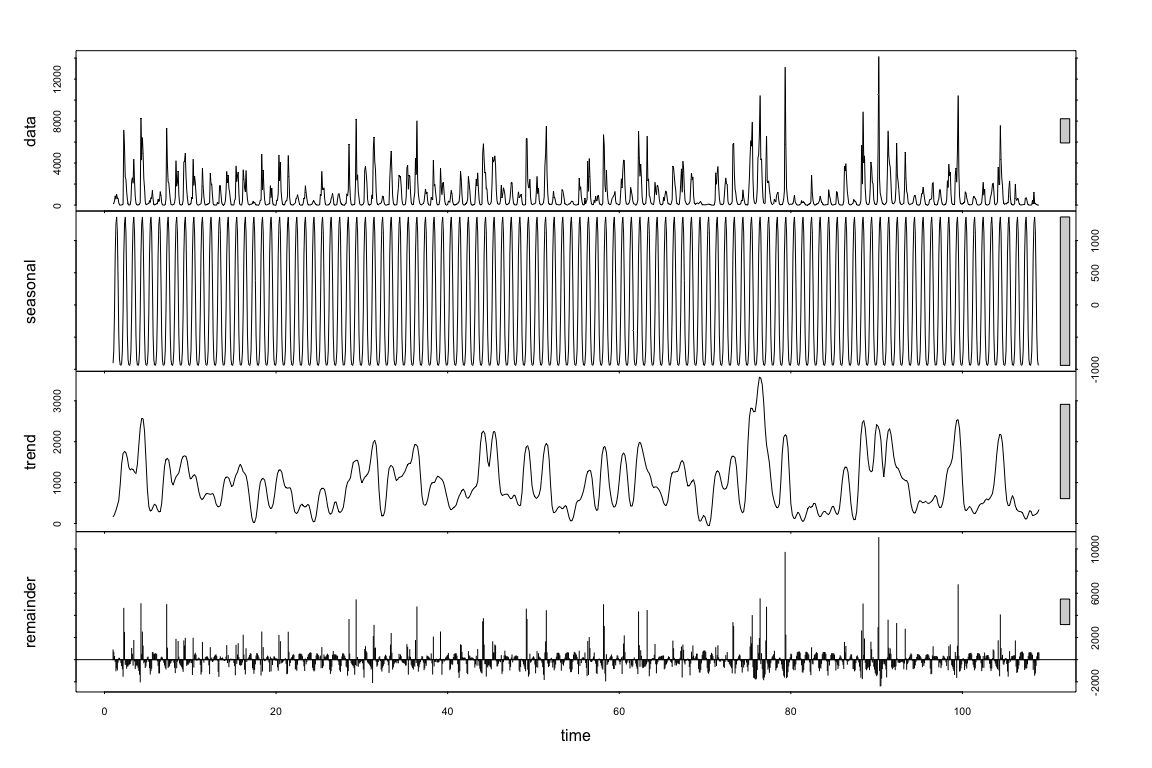
acf(water\_dat\_mon\_max[,"discharge\_afd"])



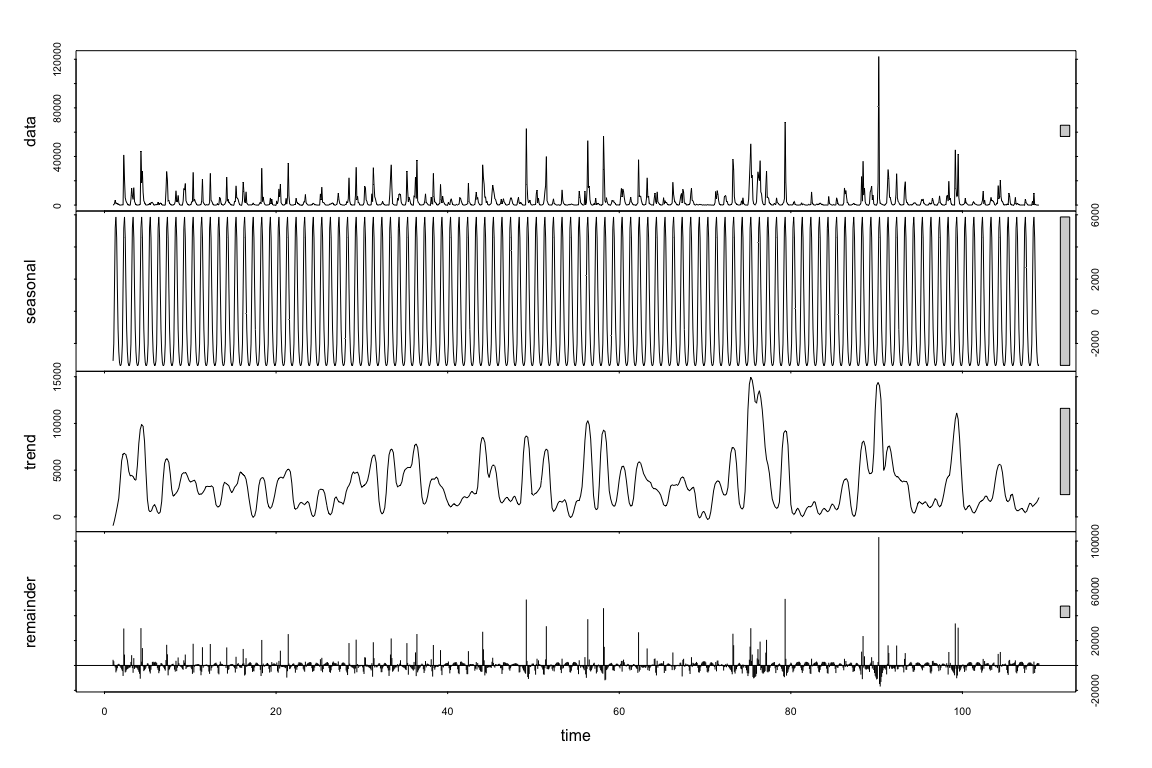
#The pattern in the ACF indicates seasonality, especially considering that water year is being used.. Lags at approximately 6 months are negative (summer dry season to winter wet) and 1 year are positive (summer to summer or winter to winter). Interestingly, the lags two years out are similar to one year away.  
  
plot(stl(water\_dat\_mon\_sum[,"discharge\_afd"], s.window = "per"))



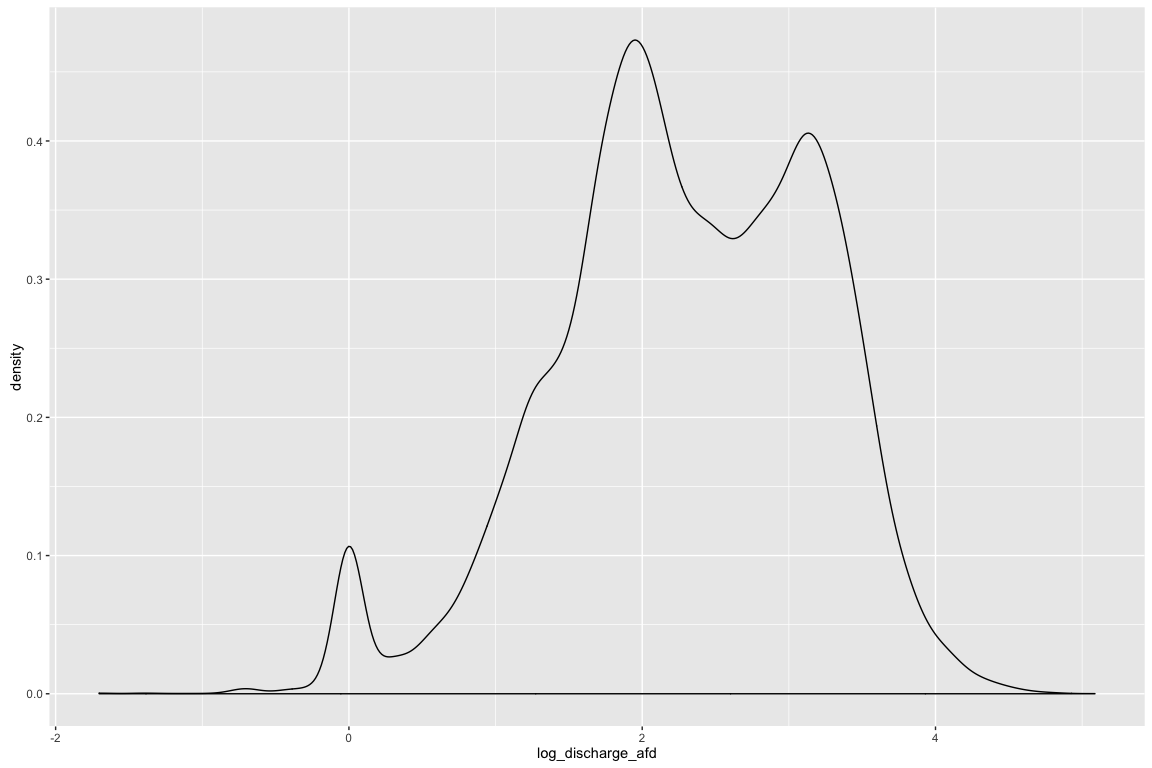
plot(stl(water\_dat\_mon\_mean[,"discharge\_afd"], s.window = "per"))



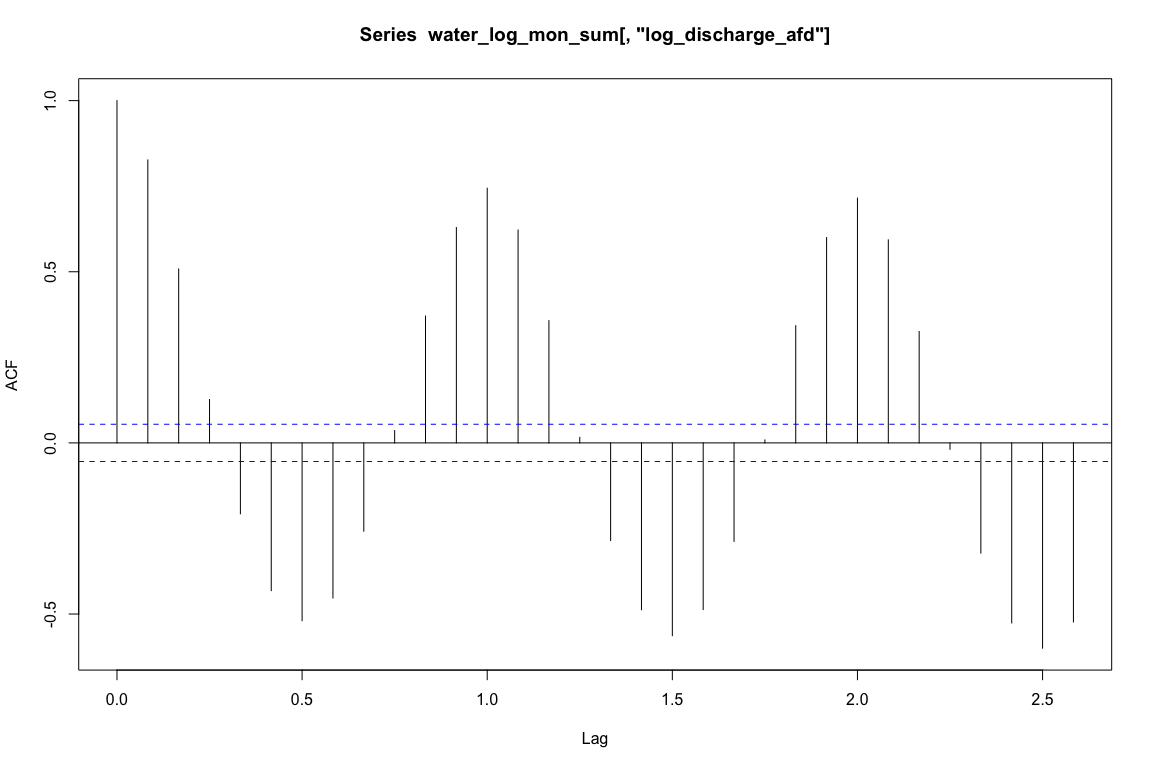
plot(stl(water\_dat\_mon\_max[,"discharge\_afd"], s.window = "per"))



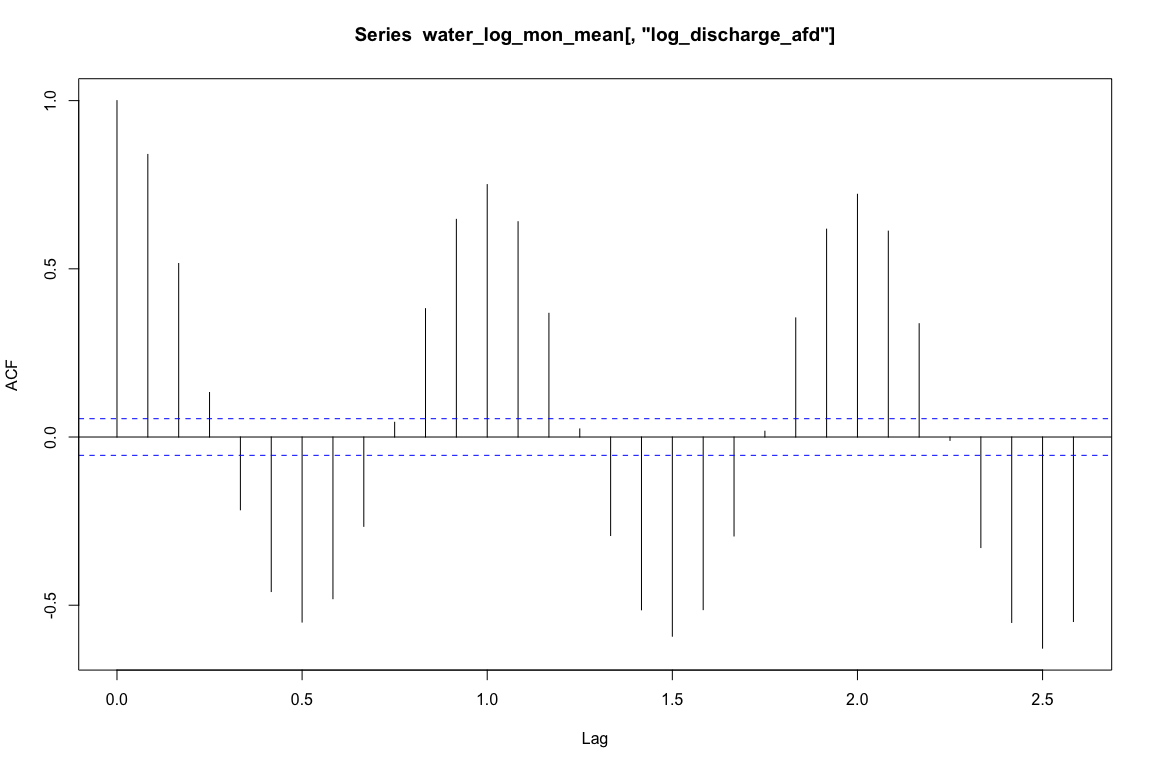
#Detrending the raw data reveals the seasonal structure swings by winter/summer. We also observe that the magnitude of flood months is greater than any individual dry month, but the dry months outnumber the total wet months.  
  
  
#log10 values to normalize  
water\_dat$log\_discharge\_afd <- ifelse(water\_dat$discharge\_afd <= 0, log10(1), log10(water\_dat$discharge\_afd)) #replacing 0 and negative values with 1 to remove inf values  
  
ggplot(water\_dat, aes(log\_discharge\_afd)) + geom\_density()



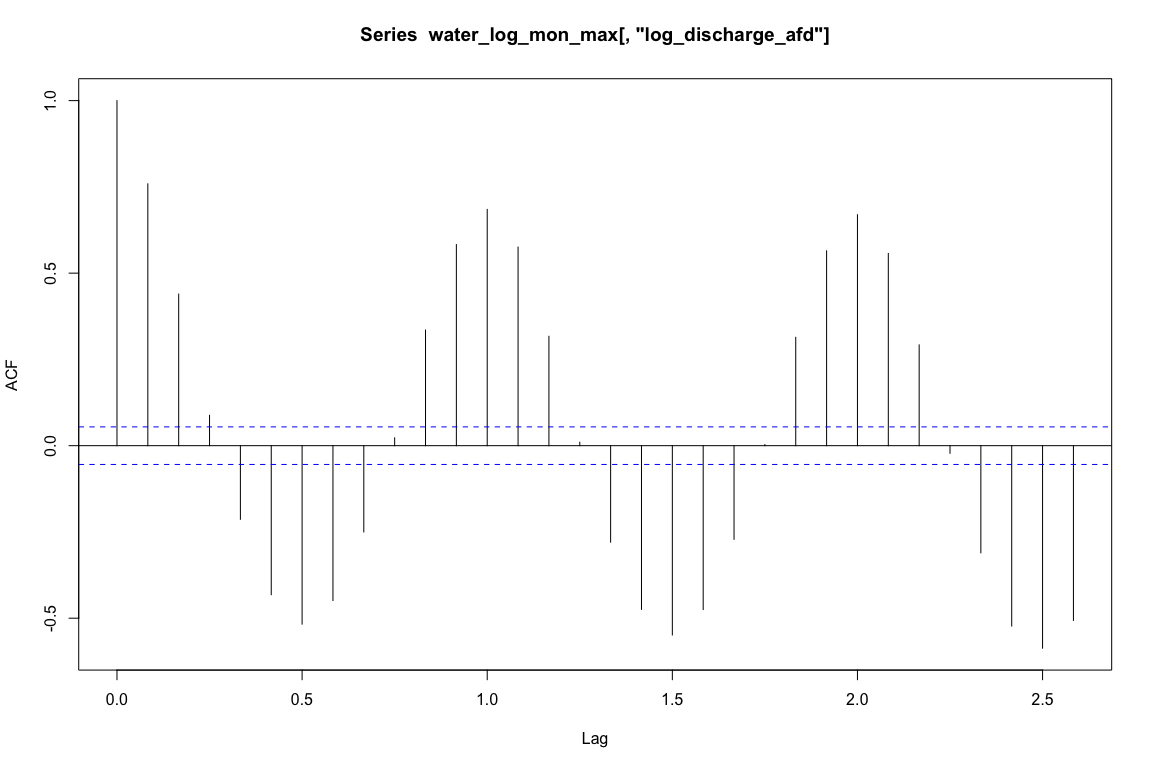
#Normalish distribution achieved  
  
water\_log\_mon\_sum <- ts(as.vector(aggregate(log\_discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = sum)), frequency = 12)  
  
water\_log\_mon\_mean <- ts(as.vector(aggregate(log\_discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = mean)), frequency = 12)  
  
water\_log\_mon\_max <- ts(as.vector(aggregate(log\_discharge\_afd ~ water\_month + water\_year, data = water\_dat, FUN = max)), frequency = 12)  
  
acf(water\_log\_mon\_sum[,"log\_discharge\_afd"])



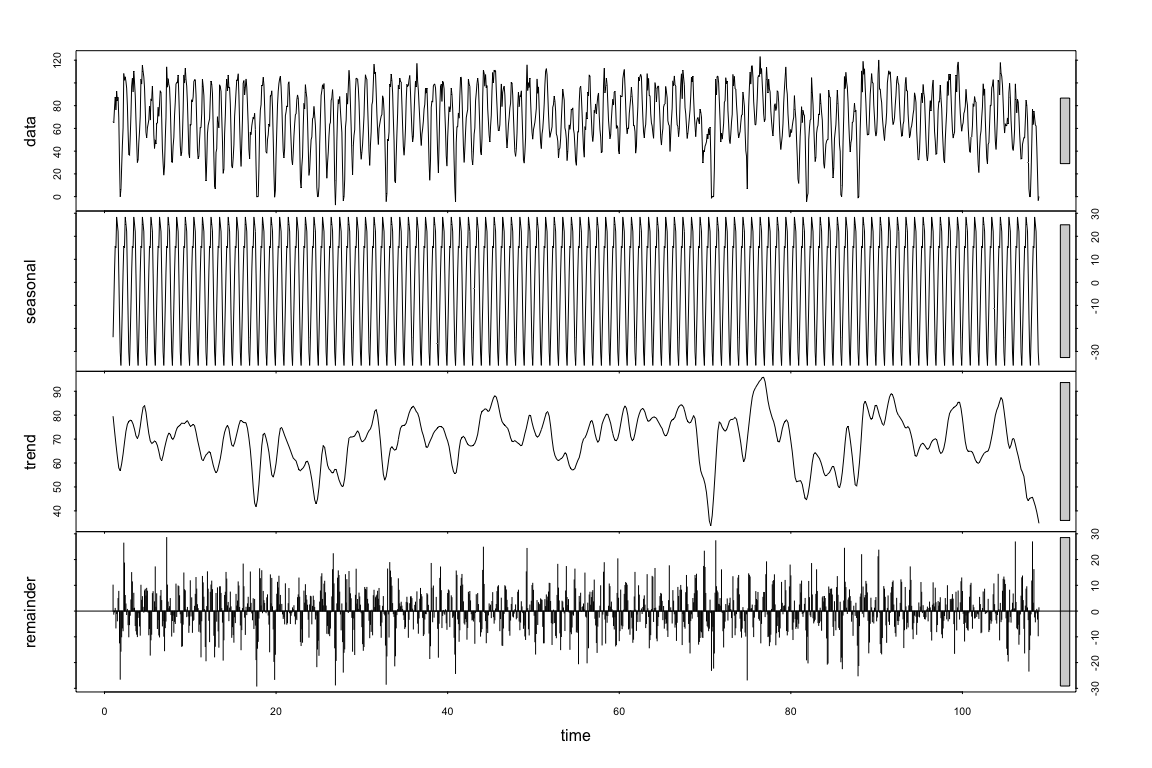
acf(water\_log\_mon\_mean[,"log\_discharge\_afd"])



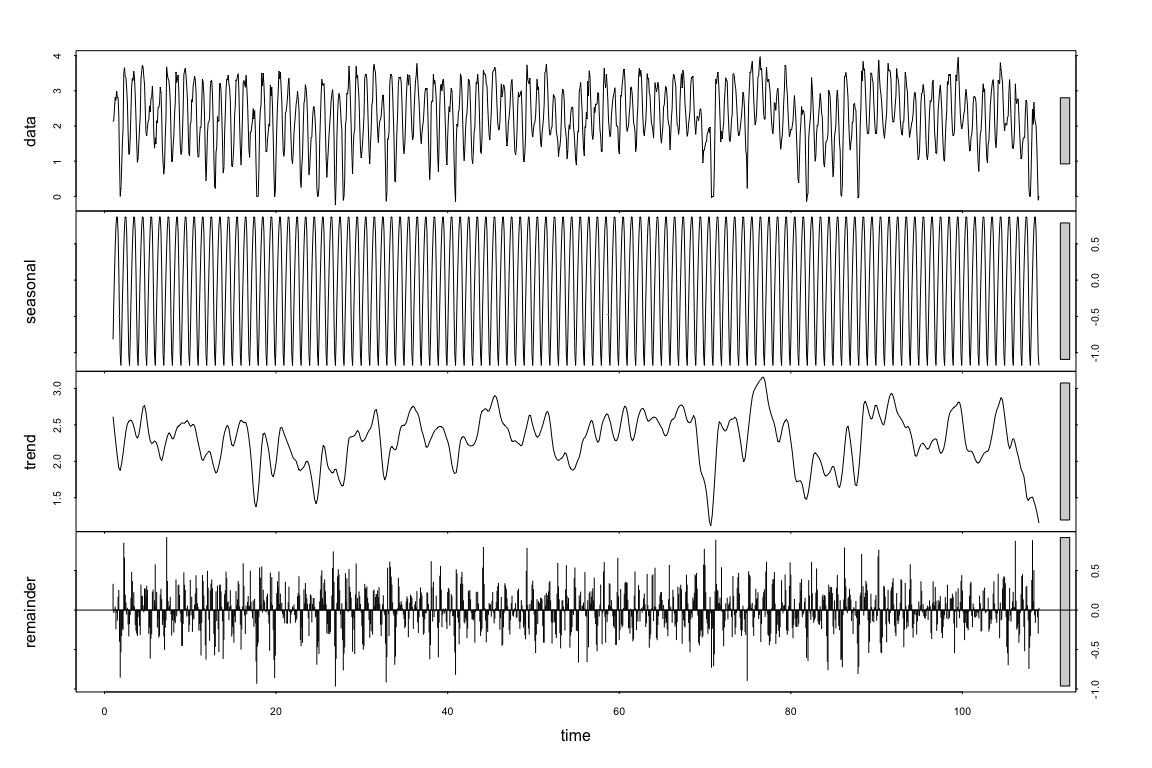
acf(water\_log\_mon\_max[,"log\_discharge\_afd"])



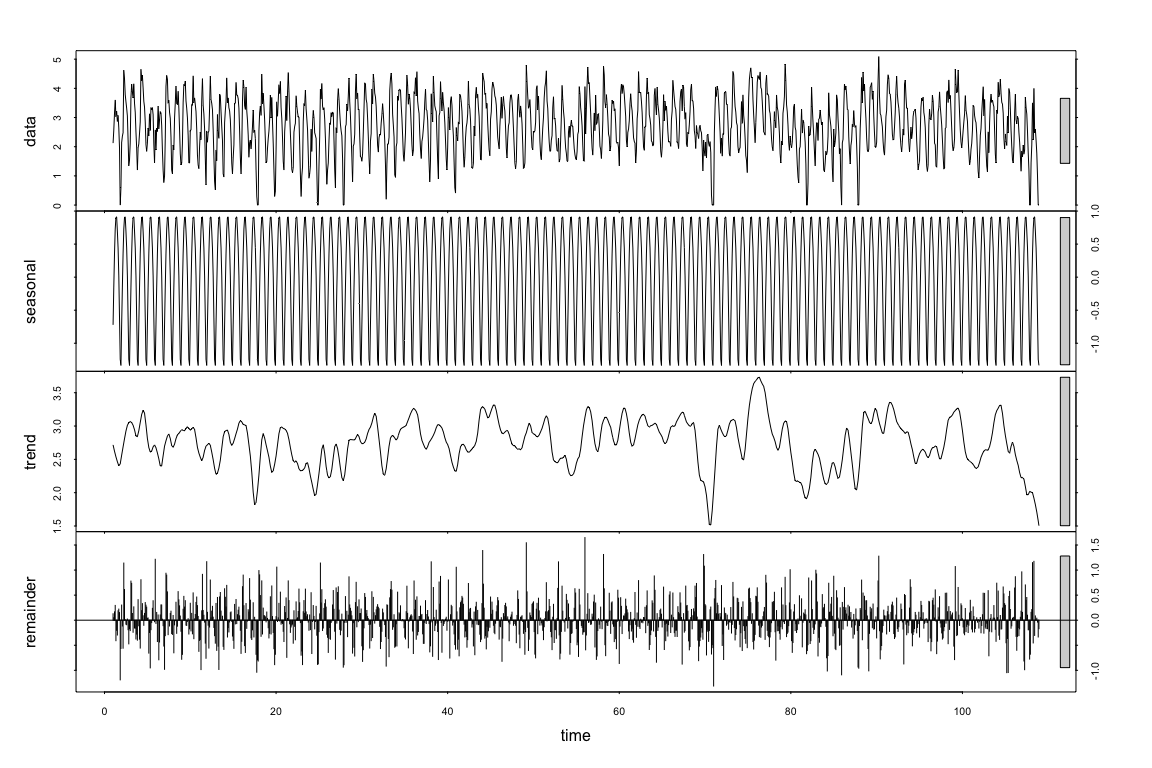
plot(stl(water\_log\_mon\_sum[,"log\_discharge\_afd"], s.window = "per"))



plot(stl(water\_log\_mon\_mean[,"log\_discharge\_afd"], s.window = "per"))

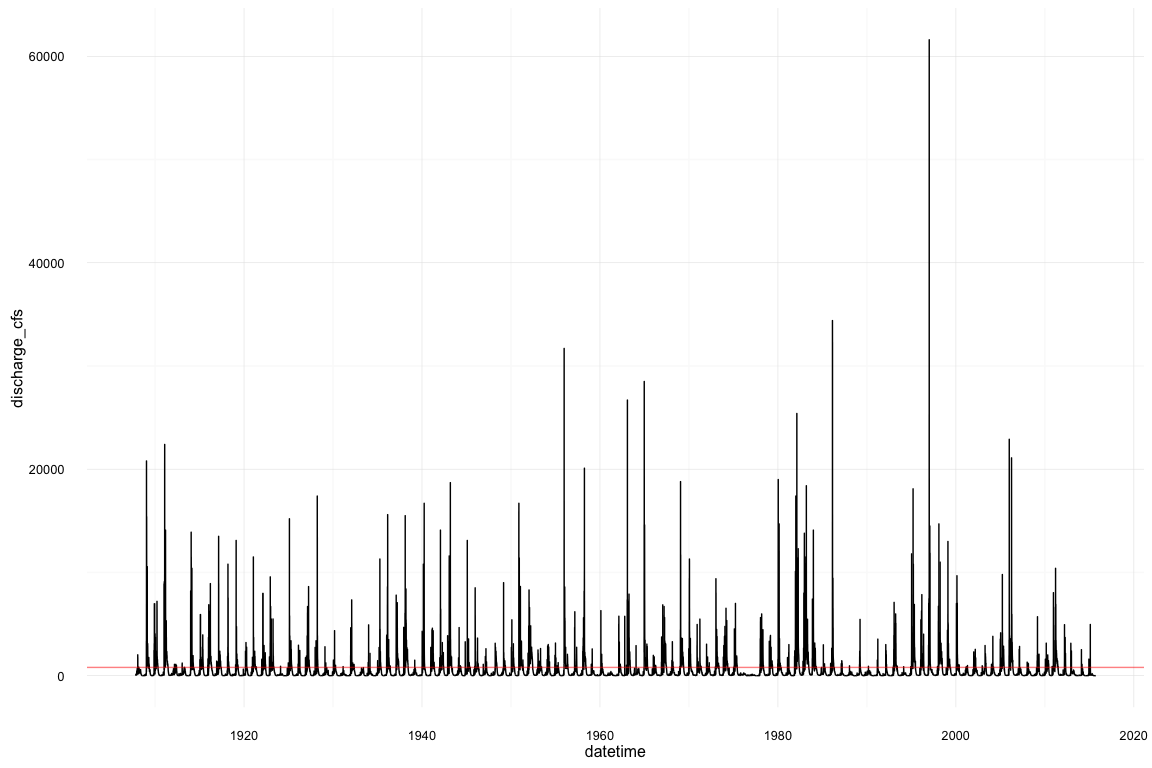


plot(stl(water\_log\_mon\_max[,"log\_discharge\_afd"], s.window = "per"))



#The seasonal trend is still apparent, but what is more interesting is that the trend takes a steep dive in the last few years of the time series. The difference between the log transformed data and the raw is that the last few years of the raw data appear to be in the midst of a depression, but about to upswing, while the transformed data appears to be continuing a downward trend.

#Calculating flood events  
#Flood days are defined as days in which discharges in excess of 800 cfs occur.  
  
water\_dat[,"flood"] <- ifelse(water\_dat[,"discharge\_cfs"] >= 800, 1, 0)  
  
ggplot(water\_dat, aes(x = datetime, y = discharge\_cfs)) + geom\_line() + geom\_hline(yintercept = 800, col = "red", alpha = 1/2) + theme\_minimal()



water\_dat\_flood <- aggregate(flood ~ water\_year, data = water\_dat, FUN = sum)  
  
water\_dat\_flood$ab <- ifelse(water\_dat\_flood$flood >= 100, "Flood Year" , "Not Flood Year")  
  
sum(ifelse(water\_dat\_flood$ab == "Flood Year", 1, 0)/nrow(water\_dat\_flood))

## [1] 0.3055556

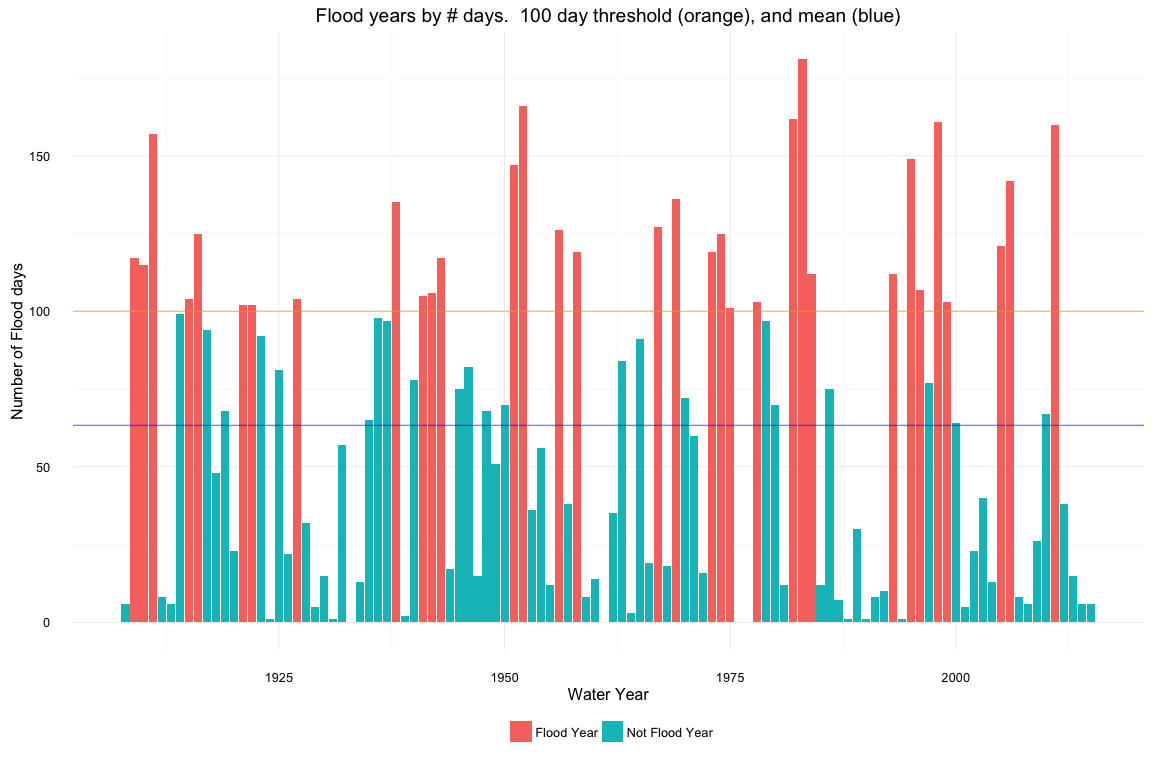
#[1] 0.3055556  
  
  
water\_dat\_flood$ab3 <- ifelse(water\_dat\_flood$flood >= 60, "Flood Year" , "Not Flood Year")  
  
sum(ifelse(water\_dat\_flood$ab3 == "Flood Year", 1, 0)/nrow(water\_dat\_flood))

## [1] 0.5185185

#[1] 0.5185185  
  
water\_dat[,"flood2"] <- ifelse(water\_dat[,"discharge\_cfs"] >= 550, 1, 0)  
  
water\_dat\_flood2 <- aggregate(flood2 ~ water\_year, data = water\_dat, FUN = sum)  
  
water\_dat\_flood2$ab <- ifelse(water\_dat\_flood2$flood2 >= 100, "Flood Year" , "Not Flood Year")  
  
sum(ifelse(water\_dat\_flood2$ab == "Flood Year", 1, 0)/nrow(water\_dat\_flood))

## [1] 0.4907407

#[1] 0.4907407  
  
ggplot(water\_dat\_flood, aes(x = water\_year, y = flood, fill = ab)) + geom\_bar(stat = "identity") + geom\_hline(yintercept = 100, color = "orange", alpha = 1/2) + theme\_minimal() + theme(legend.title = element\_blank(), legend.position = "bottom") + geom\_hline(yintercept = mean(water\_dat\_flood$flood), color = "blue", alpha = 1/2) + labs(title = "Flood years by # days. 100 day threshold (orange), and mean (blue)", x = "Water Year", y = "Number of Flood days")



**Results:**

As expected, streamflow exhibited yearly auto-correlation, in which a one year lag had positive autocorrelation and a 6 month lag had negative correlation. Lags were similar 1 year, 2 year, and expected, 3 years away. When seasonal trends were removed, the results reveal multi-year wet/dry cycles. Beginning in approximately the 1990s, the wet/dry pattern changes relative to the previous century to longer periods of low stream flow with peaks of intense flooding. Wet/dry cycles in the previous years followed a ragged sine curve with fewer extremes.  
36% of water years had more than 100 days in which discharge equaled or exceeded 800 cfs. Visual inspection of flood years reveals that flood years cluster more starting in the 1970s compared to previous years. Since 1975, ~ 3 clusters of multiple flood years occurred (2 or greater consecutive years of 100 days greater than 800 cfs). These number of days within flood years also seems to be increasing.  
Monthly decomposition of flood events show that discharge occurs primarily in the 4 – 6th months of the water year (January through March). 1997, the year with the largest daily average discharge, had its largest total, mean, and max discharge in January.

**Discussion:**

The final analysis shows that the threshold of years needed to justify the creation of a floodplain has not been met. To meet the 50% threshold, we would need to reduce the number of days a year to 60 or the discharge to 550 cfs.  
The number of flood days within a flood year seem to be increasing, as does the time between wet years. This could be due to extreme precipitation events within a longer drought period. Interestingly, the wettest year, 1983 occurred just 6 years after the driest year. Finally, extreme weather events are captured. The 97 -98 El Nino is clearly noticeable as a massive flood spike. In the monthly graphs, we see that 1997 also laid claim to the largest daily discharge, in January of that year. The 2012 – 2015 drought appears as a period of low discharge rates, comparable to the 1930s. Future analysis may include trend analysis of high discharge days: whether the trend of isolated extreme flood years is in fact increasing may change criteria for floodplain construction.

**Limitations:**

There is an unanswered question of whether our chosen data period accurately depicts future flooding. Earlier droughts, such as the 1930s Dustbowl, drew our final statistic down, but this might not be representable of current climate conditions. A more in-depth analysis would include discharge forecasts and removal of outliers.