Assignment 3: Physical Properties of Rivers

Felipe Raby Amadori

OVERVIEW

This exercise accompanies the lessons in Hydrologic Data Analysis on the physical properties of rivers.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_RiversPhysical.Rmd") prior to submission.

The completed exercise is due on 18 September 2019 at 9:00 am.

Setup

1. Verify your working directory is set to the R project file,
2. Load the tidyverse, dataRetrieval, and cowplot packages
3. Set your ggplot theme (can be theme_classic or something else)
4. Import a data frame called "MysterySiteDischarge" from USGS gage site 03431700. Upload all discharge data for the entire period of record. Rename columns 4 and 5 as "Discharge" and "Approval.Code". DO NOT LOOK UP WHERE THIS SITE IS LOCATED.
5. Build a ggplot of discharge over the entire period of record.

```
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

#Verify your working directory is set to the R project file
getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/Ramos 3 Semestre/Hydrologic_Data_Analysis"
```

```
#Load the tidyverse, dataRetrieval, and cowplot packages
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cowplot)
```

```
##
## ************************************************************
```

```
## Note: As of version 1.0.0, cowplot does not change the
##   default ggplot2 theme anymore. To recover the previous
##   behavior, execute:
##   theme_set(theme_cowplot())
## ************************************************************
```

```
library(dataRetrieval)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:cowplot':
##
##     stamp
## The following object is masked from 'package:base':
##
##     date
```

```
#Set your ggplot theme (can be theme_classic or something else)
felipe_theme <- theme_light(base_size = 12) +
  theme(axis.text = element_text(color = "grey8"),
        legend.position = "right", plot.title = element_text(hjust = 0.5))
theme_set(felipe_theme)

#Import a data frame called "MysterySiteDischarge" from USGS gage site 03431700.
#Upload all discharge data for the entire period of record. Rename columns 4 and
#5 as "Discharge" and "Approval.Code".

MysterySiteDischarge <- readNWISdv(siteNumbers = "03431700",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "",
                     endDate = "")

names(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")

#Checking for missing data
sum(is.na(MysterySiteDischarge$Discharge))
```

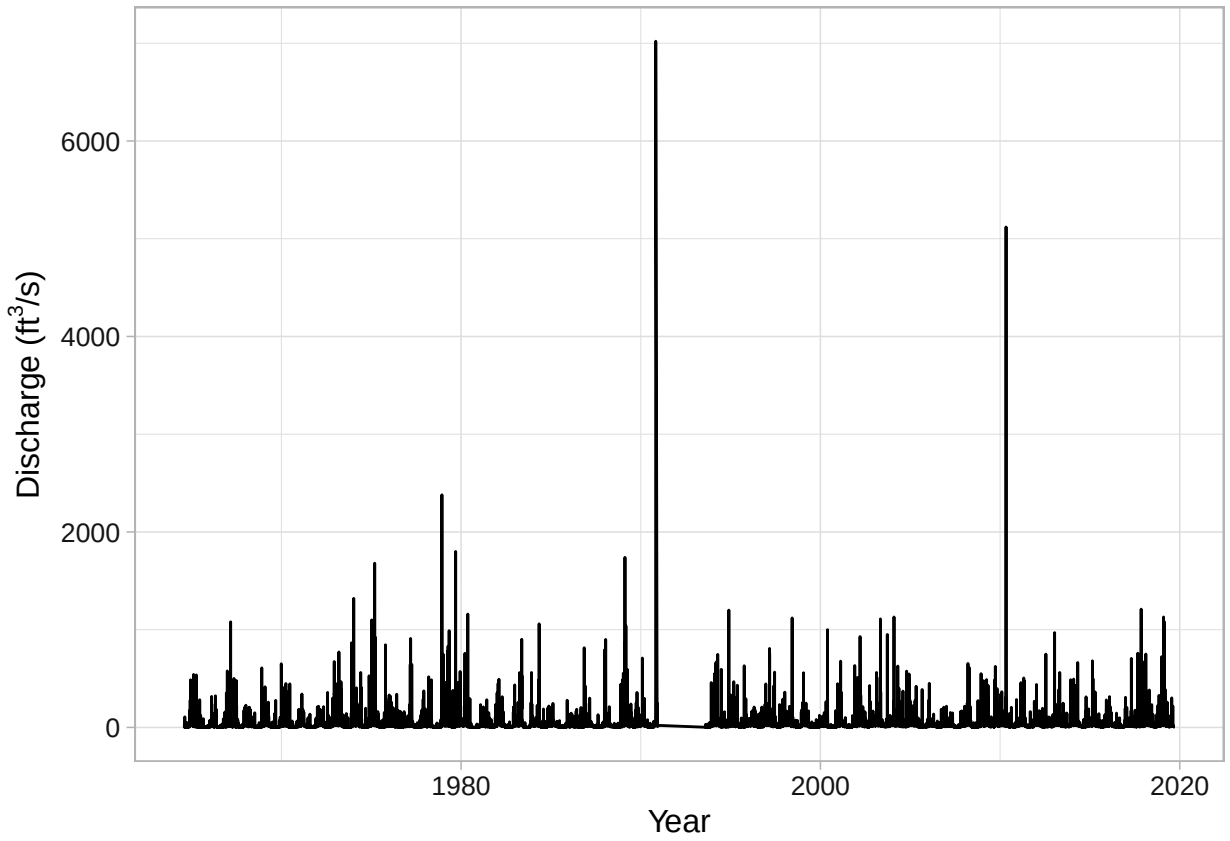
```
## [1] 0
```

```
summary(MysterySiteDischarge)
```

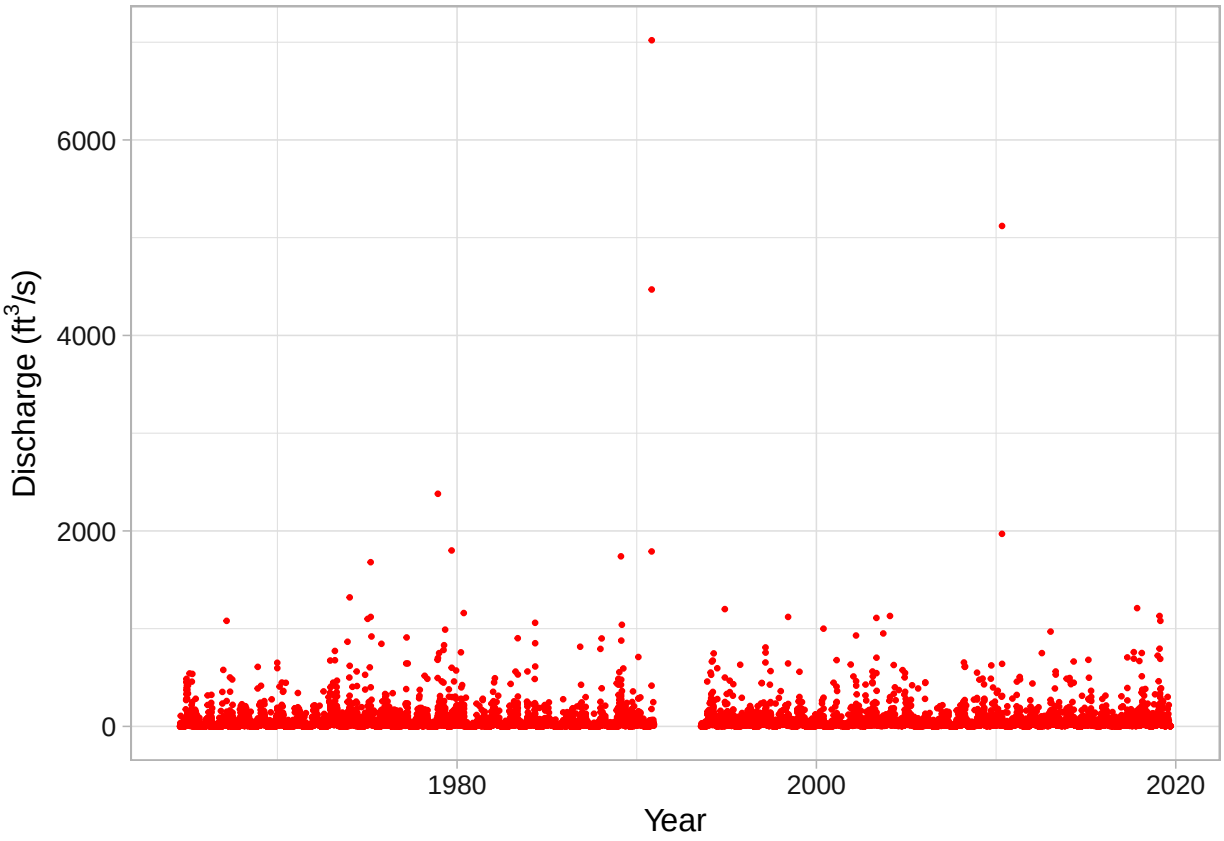
```
##   agency_cd           site_no              Date
##  Length:19127       Length:19127       Min.   :1964-08-01
##  Class :character   Class :character   1st Qu.:1977-09-03
##  Mode  :character   Mode  :character   Median :1990-10-30
##                                        Mean   :1992-02-21
##                                        3rd Qu.:2006-08-13
##                                        Max.   :2019-09-17
##    Discharge       Approval.Code
##  Min.   :   0.05   Length:19127
##  1st Qu.:   4.70   Class :character
##  Median :  13.00   Mode  :character
##  Mean   :  35.28
```

```
##  3rd Qu.:  32.40
##  Max.   :7020.00
```

```
#Build a ggplot of discharge over the entire period of record.

MisteryPlot <-
  ggplot(MysterySiteDischarge, aes(x = Date, y = Discharge)) +
         geom_line() +
         labs(x = "Year", y = expression("Discharge (ft"^3*"/s)"))
  print(MisteryPlot)
```




```
#Build a ggplot of discharge over the entire period of record using
#geom_point to look for missing data.

MisteryPlot <-
  ggplot(MysterySiteDischarge, aes(x = Date, y = Discharge)) +
         geom_point(color = "red", size = 0.5) +
         labs(x = "Year", y = expression("Discharge (ft"^3*"/s)"))
print(MisteryPlot)
```



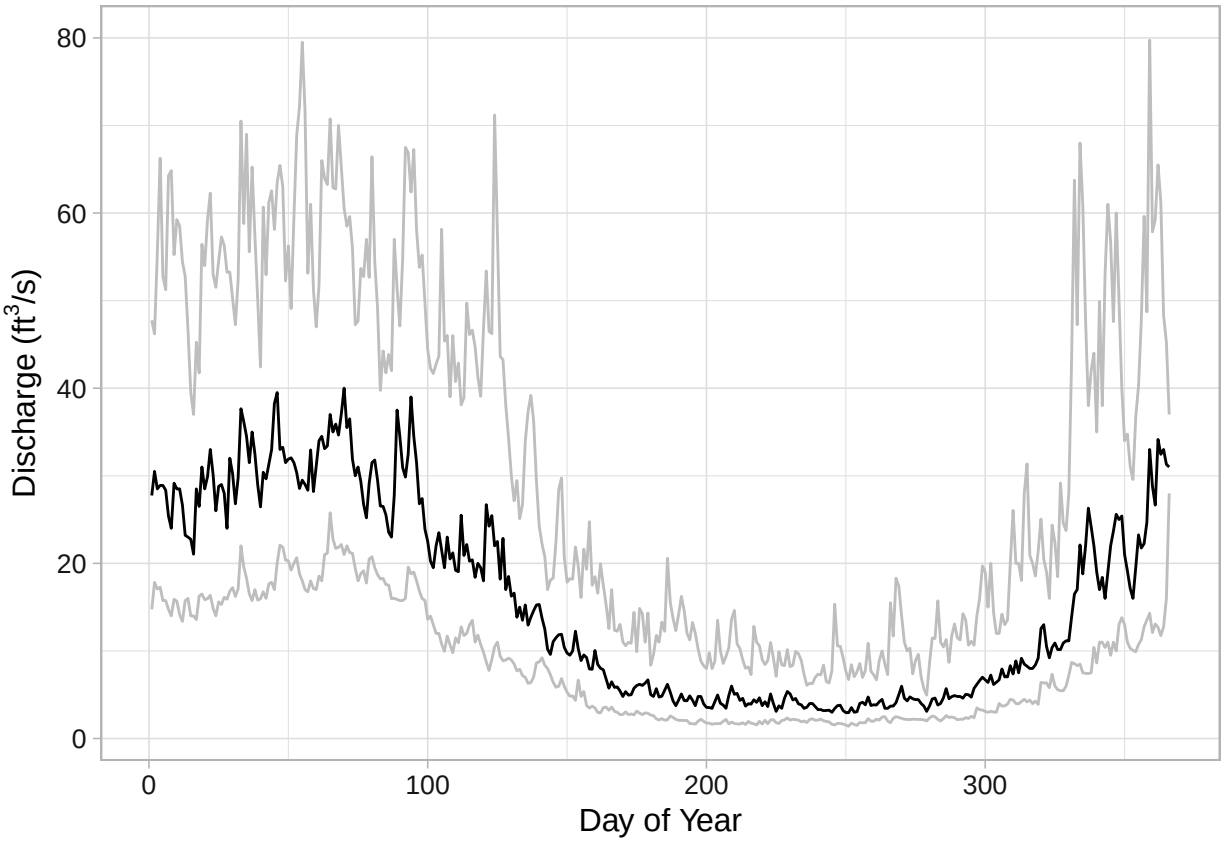

```
#There is a period of approx. 2 years and 7 months with no data between
#1990-12-16 and 1993-08-01
```

Analyze seasonal patterns in discharge

5. Add a "Year" and "Day.of.Year" column to the data frame.
6. Create a new data frame called "MysterySiteDischarge.Pattern" that has columns for Day.of.Year, median discharge for a given day of year, 75th percentile discharge for a given day of year, and 25th percentile discharge for a given day of year. Hint: the summarise function includes `quantile`, wherein you must specify `probs` as a value between 0 and 1.
7. Create a plot of median, 75th quantile, and 25th quantile discharges against day of year. Median should be black, other lines should be gray.

```
#Add a "Year" and "Day.of.Year" column to the data frame.
MysterySiteDischarge <-
  MysterySiteDischarge %>%
  mutate(Year = year(Date)) %>%
  mutate(Day.of.Year = yday(Date))

#Create a new data frame called "MysterySiteDischarge.Pattern" that has columns
#for Day.of.Year, median discharge for a given day of year, 75th percentile
#discharge for a given day of year, and 25th percentile discharge for a given
#day of year.

MysterySiteDischarge.Pattern <-
  MysterySiteDischarge %>%
```

```
  group_by(Day.of.Year) %>%
  summarise(MedianDischarge = quantile(Discharge, probs=0.5),
            Percent75Discharge = quantile(Discharge, probs=0.75),
            Percent25Discharge = quantile(Discharge, probs=0.25))

#Create a plot of median, 75th quantile, and 25th quantile discharges against
#day of year. Median should be black, other lines should be gray.

MysterySiteDischargePatternPlot <-
  ggplot(MysterySiteDischarge.Pattern, aes(x = Day.of.Year)) +
  geom_line(aes(y = MedianDischarge)) +
  geom_line(aes(y = Percent75Discharge), color = "gray") +
  geom_line(aes(y = Percent25Discharge), color = "gray") +
  labs(x = "Day of Year", y = expression("Discharge (ft"^3*"/s)"))
print(MysterySiteDischargePatternPlot)
```



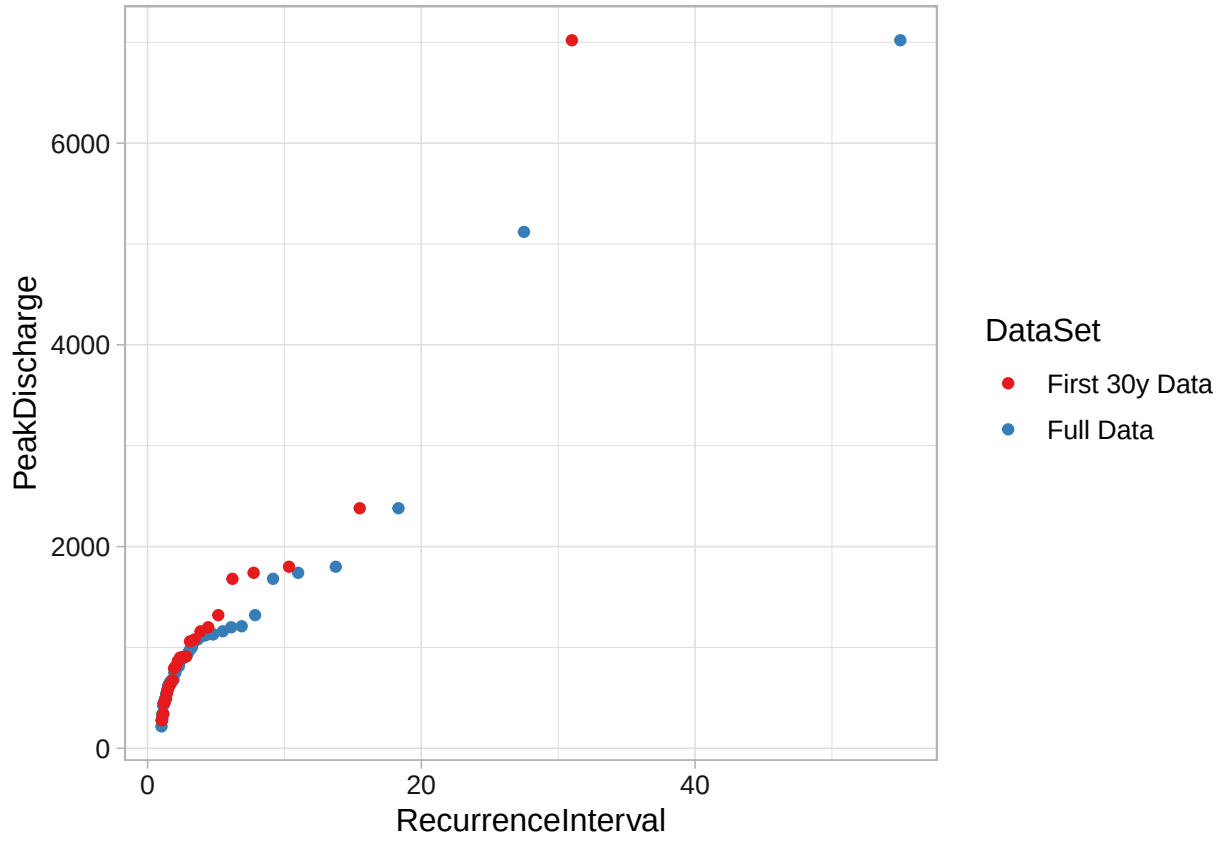
8. What seasonal patterns do you see? What does this tell you about precipitation patterns and climate in the watershed?

   The period of lower discharge values is approx. between day #150 and day #300 which corresponds to the end of Spring and the Summer season. Discharge values start to increase importantly during the Fall season (probably due to seasonal rains). Discharge values stay at their highest levels during the whole winter and start to decline towards the beginning of Spring. Mystery Site probably is located in a place with low spring/summer and high fall/winter precipitations. There is no clear sign of an increase of flow when temperatures start to increase (Spring) meaning that there is probably no snow pack melting going to Mystery Site.

Create and analyze recurrence intervals

9. Create two separate data frames for MysterySite.Annual.30yr (first 30 years of record) and MysterySite.Annual.Full (all years of record). Use a pipe to create your new data frame(s) that includes the year, the peak discharge observed in that year, a ranking of peak discharges, the recurrence interval, and the exceedende probability.
10. Create a plot that displays the discharge vs. recurrence interval relationship for the two separate data frames (one set of points includes the values computed from the first 30 years of the record and the other set of points includes the values computed for all years of the record.
11. Create a model to predict the discharge for a 100-year flood for both sets of recurrence intervals.

```
#Create two separate data frames for MysterySite.Annual.30yr (first 30 years of
#record) and MysterySite.Annual.Full (all years of record). Use a pipe to create
#your new data frame(s) that includes the year, the peak discharge observed in
#that year, a ranking of peak discharges, the recurrence interval, and the
#exceedende probability.

MysterySite.Annual.30yr <-
  MysterySiteDischarge %>%
  filter(Year < 1996) %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval, DataSet = "First 30y Data")

MysterySite.Annual.Full <-
  MysterySiteDischarge %>%
  group_by(Year) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(Year) + 1)/Rank,
         Probability = 1/RecurrenceInterval, DataSet = "Full Data")


#Analize incomplete years

#Create a plot that displays the discharge vs. recurrence interval relationship
#for the two separate data frames (one set of points includes the values
#computed from the first 30 years of the record and the other set of points
#includes the values computed for all years of the record.

MisterySite.Annual.Both <- rbind(MysterySite.Annual.Full, MysterySite.Annual.30yr)

MysterySiteRecurrencePlot.Both <-
  ggplot(MisterySite.Annual.Both,
         aes(x = RecurrenceInterval, y = PeakDischarge, color = DataSet)) +
  geom_point() +
  scale_color_brewer(palette = "Set1")
print(MysterySiteRecurrencePlot.Both)
```




```
#Create a model to predict the discharge for a 100-year flood for both sets of
#recurrence intervals.

#30year
MysterySite.Annual.30yr.Model <- lm(data = MysterySite.Annual.30yr,
                                    PeakDischarge ~ log(RecurrenceInterval))
summary(MysterySite.Annual.30yr.Model)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log(RecurrenceInterval), data = MysterySite.Annual.30yr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -974.12 -337.65   34.84  232.57 2908.00
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -69.87     185.73  -0.376     0.71
## log(RecurrenceInterval)  1217.79     147.16   8.275 5.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 673.9 on 28 degrees of freedom
## Multiple R-squared:  0.7098, Adjusted R-squared:  0.6994
## F-statistic: 68.48 on 1 and 28 DF,  p-value: 5.261e-09
```

```
MysterySite.Annual.30yr.Model$coefficients[1] +
  MysterySite.Annual.30yr.Model$coefficients[2]*log(100)
```

```
## (Intercept)
##    5538.257
```

```
#Full Record
MysterySite.Annual.Full.Model <- lm(data = MysterySite.Annual.Full,
                                    PeakDischarge ~ log(RecurrenceInterval))
summary(MysterySite.Annual.Full.Model)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log(RecurrenceInterval), data = MysterySite.Annual.Full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -955.95 -236.29   41.91  210.67 2805.35
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.001    116.322  -0.017    0.986
## log(RecurrenceInterval) 1052.234     88.834  11.845   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.3 on 52 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7244
## F-statistic: 140.3 on 1 and 52 DF,  p-value: < 2.2e-16
```

```
MysterySite.Annual.Full.Model$coefficients[1] +
  MysterySite.Annual.Full.Model$coefficients[2]*log(100)
```

```
## (Intercept)
##    4843.717
```

12. How did the recurrence interval plots and predictions of a 100-year flood differ among the two data frames? What does this tell you about the stationarity of discharge in this river?

    Looking at the recurrence interval plots, it can be seen that below 1000 ft3/s the curves are very similar. Above that value, data differs, having the data of the first 30 years on record (30yearmodel) higher peak discharge values for the same recurrence interval than the full data (fullmodel). More over, for the prediction of the discharge for a 100-year flood for the 30yearmodel we got a discharge value of 5538.257 ft3/s and for the prediction of the discharge for a 100-year flood for the fullmodel we got a discharge value of 4843.717 ft3/s which is approx. a 13% decrease. That difference between predictions is relevant considering that those values could be use for designing flood protection infrastructure or other type of important water infrastructure. This tells me that the assumption of stationary flow regimes in Mystery River is not perfectly correct. Recurrence values vary importantly depending on the data selected to do the calculations specially in data sets that include extreme outliers and periods of missing data.

Reflection

13. What are 2-3 conclusions or summary points about river discharge you learned through your analysis?

1. River discharge is highly seasonal. It can be learned a lot about the climate of a site by just looking at the discharge information.
2. Predicting the recurrence of discharge values in rivers is essential for infrastructure design. However, non-stationary discharge regimes can alter this calculations importantly. A designer or decision maker has to know this facts when selecting the data used for calculations. Every piece of extra information that can be obtained in recurrence calculations should be used or at least considered when making decisions. If a more conservative or risky approach is followed depends on the nature of the study or the design that is been performed.

14. What data, visualizations, and/or models supported your conclusions from 13?

    I supported conclusion #1 with the plot of median, 75th quantile, and 25th quantile discharges against day of year and all the data that was used in creating that plot. I supported conclusion #2 with the plot that displays the discharge vs. recurrence interval relationship for the two separate data frames and the model to predict the discharge for a 100-year flood for both sets of recurrence intervals.

15. Did hands-on data analysis impact your learning about discharge relative to a theory-based lesson? If so, how?

    It definitely impacted my learning. It allowed me to work with the data, to try to solve problems that happen while doing it. Also using real data for learning allows us to immediately gain experience working with "real life" data and understand the gaps between theory and practice.

16. How did the real-world data compare with your expectations from theory?

    Real world data is rarely "perfect" and I think that is one the most important things that we need to learn how to deal with. Also important is how relevant are certain assumptions that are accepted in some fields that sometimes can lead to important mistakes. Expectations from theory would be to observe perfect seasons, perfect stationary data, no missing values, no calibration mistakes in sensors. I think it could be said that those expectations will be met almost never.