Assignment 3: Physical Properties of Rivers

Jack Carpenter

OVERVIEW

This exercise accompanies the lessons in Water Data Analytics 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, check your PDF against the key and then submit your assignment completion survey at https://forms.gle/ydeD5axzCnaNzgss9

Having trouble? See the assignment's answer key if you need a hint. Please try to complete the assignment without the key as much as possible - this is where the learning happens!

Target due date: 2022-02-08

Setup

1. Verify your working directory is set to the R project file. Load the tidyverse, dataRetrieval, lubridate, and lfstat packages. Set your ggplot theme (can be theme_classic or something else).
2. Import a data frame called "MysterySiteDischarge" from USGS gage site 03431700. Import discharge data starting on 1964-10-01 and ending on 2021-09-30. Rename columns 4 and 5 as "Discharge" and "Approval.Code". DO NOT LOOK UP WHERE THIS SITE IS LOCATED.
3. Build a ggplot of discharge over the entire period of record.

```
#1: Setup
#verify working directory
getwd()
```

```
## [1] "/Users/Jack/Documents/Duke/Spring 2022/Water Data Analytics/Water_Data_Analytics_2022/Assignment
```

```
#load packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

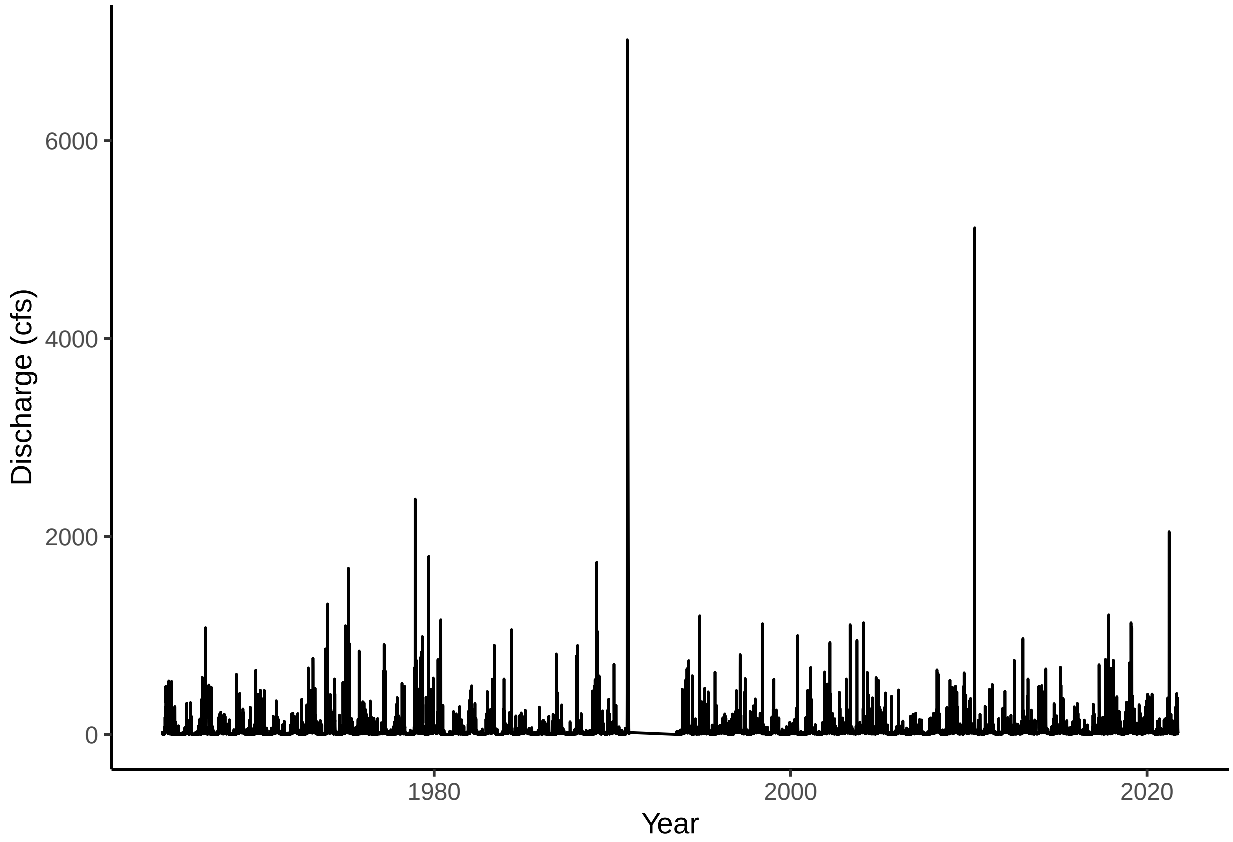
```
library(dataRetrieval)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(lfstat)
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## Loading required package: lmom

## Loading required package: lattice
```

```r
#set ggplot theme
theme_set(theme_classic())

#2: Import dataset
MysterySiteDischarge <- readNWISdv(siteNumbers = "03431700",
                                  parameterCd = "00060",
                                  startDate = "1964-10-01",
                                  endDate = "2021-09-30")
#rename columns
names(MysterySiteDischarge)[4:5] <- c("Discharge", "Approval.Code")

#3: GGplot
ggplot(MysterySiteDischarge, aes(x = Date, y = Discharge)) +
  geom_line() +
  labs(x = "Year", y = "Discharge (cfs)")
```



Analyze seasonal patterns in discharge

4. Add a "WaterYear" and "DayOfYear" column to the data frame. Hint: Use a pipe, and you will need both the lubridate and lfstat packages. Set WaterYear to numeric.
5. 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.
6. Create a plot of median, 75th quantile, and 25th quantile discharges against day of year. Median should be black, other lines should be gray.

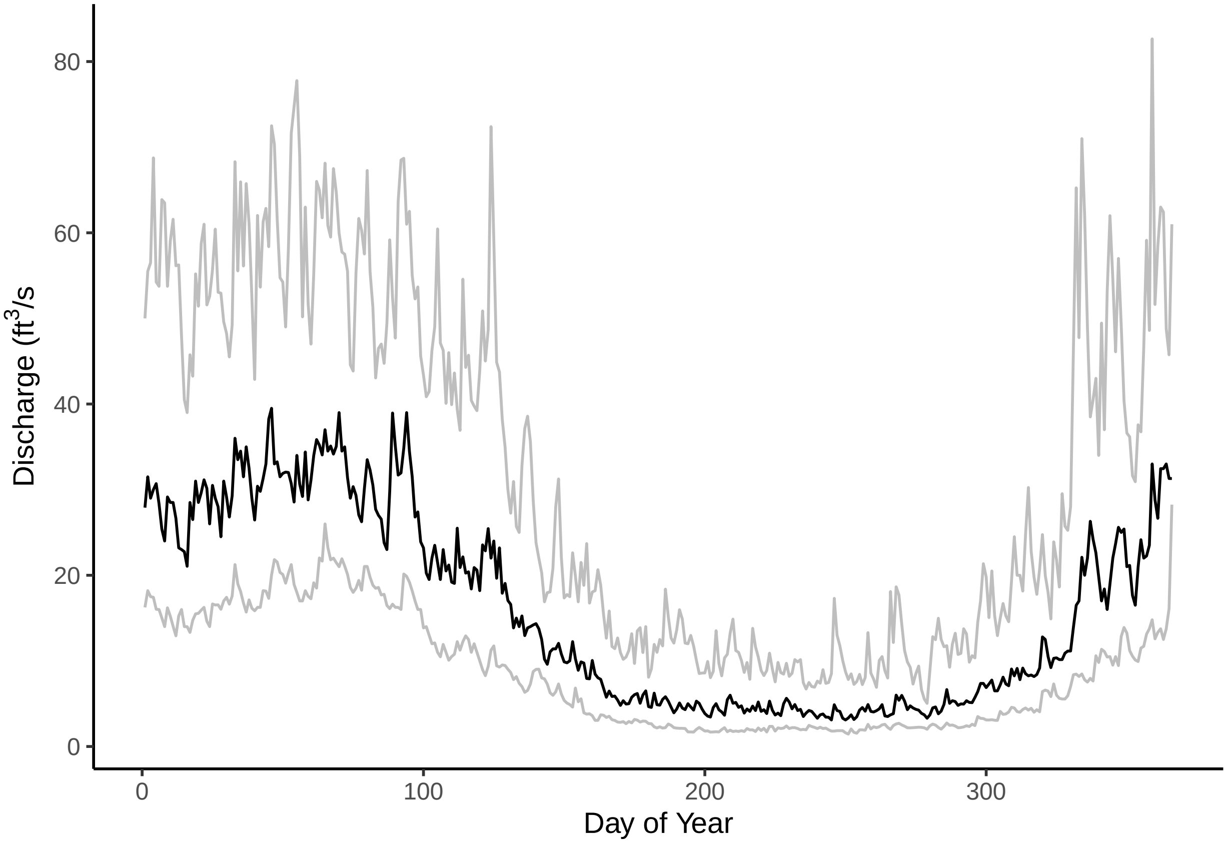
```
#4 Add columns
MysterySiteDischarge <- MysterySiteDischarge %>%
  mutate(DayOfYear = yday(Date),
         WaterYear = water_year(Date, origin = "usgs"))
#set WaterYear to numeric
class(MysterySiteDischarge$WaterYear)
```

```
## [1] "factor"
```

```
MysterySiteDischarge$WaterYear <- as.numeric(as.character(MysterySiteDischarge$WaterYear))
class(MysterySiteDischarge$WaterYear)
```

```
## [1] "numeric"
```

```
#5 Create new Dataframe
MysterySiteDischarge.Pattern <- MysterySiteDischarge %>%
  group_by(DayOfYear) %>%
  summarise(Median.Discharge = median(Discharge, na.rm = TRUE),
            p75.Discharge = quantile(Discharge, 0.75, na.rm = TRUE),
            p25.Discharge = quantile(Discharge, 0.25, na.rm = TRUE))
# I think I might have the 75th and 25th percentiles backwards, but I'm not sure
# this can also be done with quantile(data, probs = percentile)

#6 GGplot!
ggplot(MysterySiteDischarge.Pattern, aes(x = DayOfYear)) +
  geom_line(aes(y = Median.Discharge), color = "black") +
  geom_line(aes(y = p75.Discharge), color = "grey") +
  geom_line(aes(y = p25.Discharge), color = "grey") +
  labs(x = "Day of Year", y = expression("Discharge (ft"^3*"/s"))
```



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

   There is a very strong seasonal pattern where discharge is elevated in the Winter and Spring months and very depressed during the summer and Fall. There is more variation in the high-discharge months as well, indicating that the increase is probably due to precipitation patterns. This looks like a relatively wet winter and dry summer. This looks a lot like a West Coast climate with relatively dry summers and wet winters - California's meditteranean climate and the PNW are both candidates for me. It could also be somewhere in the Mountain West, but I would expect more variability in the summer in that case due to summer thunderstorms in the Rockies occassionally dumping significant amounts of water. Missed that there is no strong snowmelt pulse in the spring, so probably somewhere warm enough that snow is not part of the equation - rules out CA and probably the PNW as well.

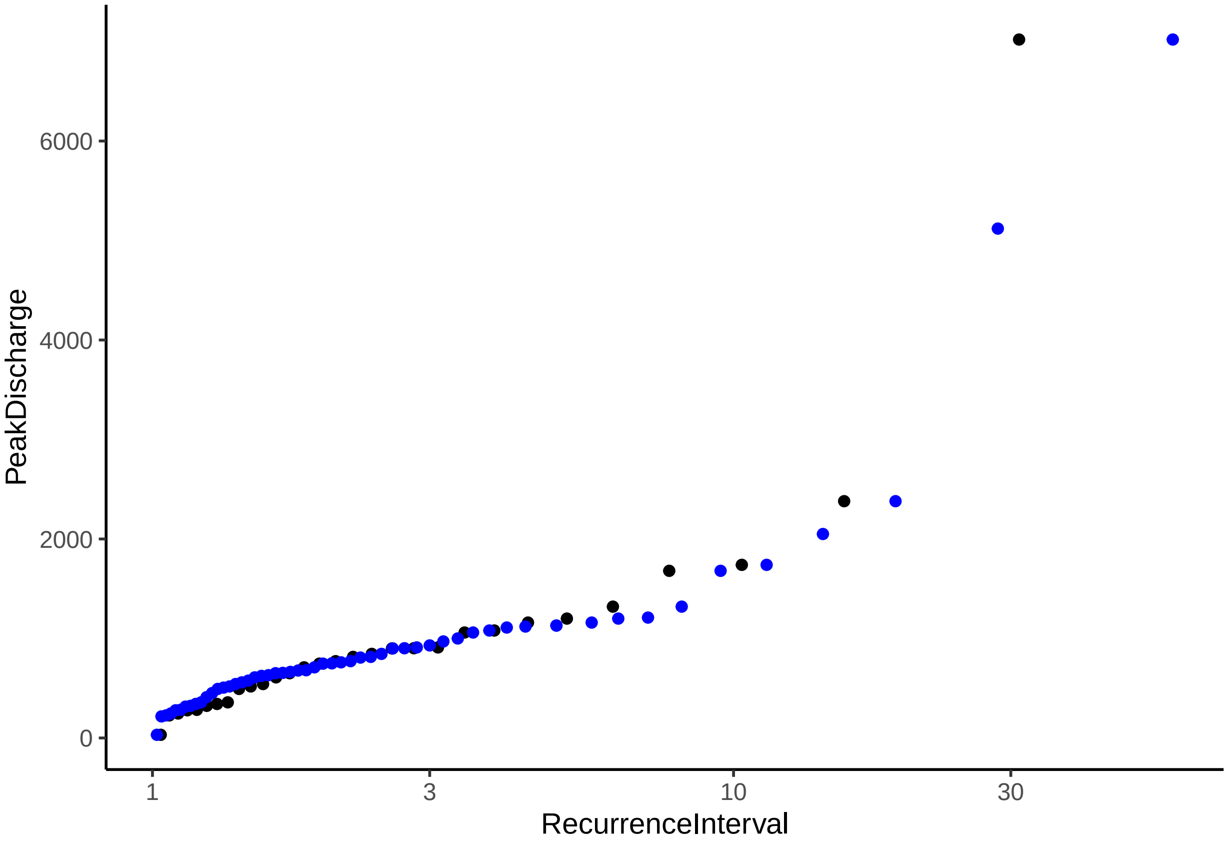
Create and analyze recurrence intervals

8. 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 water year, the peak discharge observed in that year, a ranking of peak discharges, the recurrence interval, and the exceedende probability.

9. 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.

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

```r
#8 Create dataframes
# first the first 30 years
MysterySite.Annual.30yr <- MysterySiteDischarge %>%
  filter(WaterYear <= 1995) %>%
  group_by(WaterYear) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(WaterYear)+1)/Rank,
         Probability = 1/RecurrenceInterval)
# now the whole timeframe
MysterySite.Annual.Full <- MysterySiteDischarge %>%
  group_by(WaterYear) %>%
  summarise(PeakDischarge = max(Discharge)) %>%
  mutate(Rank = rank(-PeakDischarge),
         RecurrenceInterval = (length(WaterYear)+1)/Rank,
         Probability = 1/RecurrenceInterval)

#9 Create a plot comparing the two different Recurrence Intervals
ggplot(MysterySite.Annual.30yr, aes(x = RecurrenceInterval, y = PeakDischarge)) +
  geom_point() +
  geom_point(data = MysterySite.Annual.Full, color = "blue",
             aes(x = RecurrenceInterval, y = PeakDischarge)) +
  scale_x_log10()
```



```
# I hate reading log scales, but the plot does look better with a log scale

#10 Model both time periods
MysterySite.RIModel.30yr <- lm(data = MysterySite.Annual.30yr, PeakDischarge ~ log10(RecurrenceInterval
summary(MysterySite.RIModel.30yr)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log10(RecurrenceInterval), data = MysterySite.Annual.30yr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -978.0 -319.0  111.1  195.6 2947.9
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -160.4      189.0  -0.849    0.403
## log10(RecurrenceInterval)   2838.0      344.8   8.230 5.88e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 685.9 on 28 degrees of freedom
## Multiple R-squared:  0.7075, Adjusted R-squared:  0.6971
## F-statistic: 67.73 on 1 and 28 DF,  p-value: 5.876e-09
```

```
#now the whole timeframe
MysterySite.RIModel.Full <- lm(data = MysterySite.Annual.Full, PeakDischarge ~ log10(RecurrenceInterval
summary(MysterySite.RIModel.Full)
```

```
##
## Call:
## lm(formula = PeakDischarge ~ log10(RecurrenceInterval), data = MysterySite.Annual.Full)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -862.8 -245.3   67.0  206.3 2779.1
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -35.31     110.67  -0.319    0.751
## log10(RecurrenceInterval)  2435.40     194.25  12.537   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 560.9 on 54 degrees of freedom
## Multiple R-squared:  0.7443, Adjusted R-squared:  0.7396
## F-statistic: 157.2 on 1 and 54 DF,  p-value: < 2.2e-16
```

```
#now lets make some predictions for 100-yr recurrence intervals
MysterySite.RIModel.30yr$coefficients[1] + MysterySite.RIModel.30yr$coefficients[2]*log10(100)
```

```
## (Intercept)
##    5515.593
```

```
MysterySite.RIModel.Full$coefficients[1] + MysterySite.RIModel.Full$coefficients[2]*log10(100)
```

```
## (Intercept)
##    4835.488
```

```
#still not exactly sure what the coefficients stand for, but I'll go with it
```

11. 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?

    The recurrence interval plots are very similar out until about 20 years, then they start to diverge. This indicates a degree of stationarity, as anything smaller than a 20-year event is likely the same size now as it was 30-years ago. However, past 20 years is where there is some change. The 30-year event for the first 30 years is the same size as the 50-year event for the full timeline, indicating potentially that large events are becoming more rare. This is backed up by the predictions for the 100-yr floods. The models predicted a 100-year flood about 1000cfs higher for the first 30 years than it was for the full timeframe, indicating that potentially this river is seeing slightly reduced flows and thus a slightly smaller 100-year flood event now than was expected 30 years ago. Overall, this river is experiencing stationarity with maybe some minor shifts.