Assignment 1: Introduction

Sophia Bryson (sab159)

OVERVIEW

This exercise accompanies the lessons in Water Data Analytics on introductory material.

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 (marked with >).
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After completing your assignment, fill out the assignment completion survey in Sakai.

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: 2022-01-18

Course Setup

1. Post the link to your forked GitHub repository below. Your repo should include one or more commits and an edited README file.

   Link: https://github.com/sab159/Water_Data_Analytics_2022

Data Visualization Exercises

2. Set up your work session. Check your working directory, load packages `tidyverse`, `dataRetrieval`, and `zoo`. Set your ggplot theme as theme_classic (you may need to look up how to set your theme).

```
# Check working directory
   getwd()
```

```
## [1] "C:/Users/lenovo/Desktop/Duke/Spring 2022/ENV790/Water_Data_Analytics_2022/Assignments"
```

```
# Load necessary packages
   #install.packages("zoo") #only need to run once. Hence commented out.
   library(tidyverse)
   library(dataRetrieval)
```

```
## Warning: package 'dataRetrieval' was built under R version 4.1.2
```

```
   library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.1.2
```

```
# Set ggplot theme to classic
   theme_set(theme_classic()) #see https://ggplot2.tidyverse.org/reference/theme_get.html
```

3. Upload discharge data for the Eno River at site 02096500 for the same dates as we studied in class (2012-01-01 through 2021-12-31). Obtain data for discharge. Rename the columns with informative titles, as we did in class.

```
# Import data using dataRetrieval package
   # Parameters
   site = "02096500"
   startDate = "2012-01-01"
   endDate = "2021-12-31"
   parameter = "00060" #corresponds to discharge (cfs)

   # Get daily values (dv) from NWIS
   EnoDischarge <- readNWISdv(siteNumbers = site,
                              startDate = startDate,
                              endDate = endDate,
                              parameterCd = parameter)

# Rename columns - useful things & standard capitalization
   EnoDischarge <- EnoDischarge %>% rename(date = Date,
                                           discharge_cfs = X_00060_00003,
                                           approvalCode = X_00060_00003_cd)

   #examine attribute information
   attr(EnoDischarge, "variableInfo")
```

```
##   variableCode           variableName              variableDescription
## 1        00060 Streamflow, ft&#179;/s Discharge, cubic feet per second
##       valueType  unit options noDataValue
## 1 Derived Value ft3/s    Mean          NA
```

```
   attr(EnoDischarge, "siteInfo")
```

```
##                   station_nm  site_no agency_cd timeZoneOffset
## 1 HAW RIVER AT HAW RIVER, NC 02096500      USGS         -05:00
##   timeZoneAbbreviation dec_lat_va dec_lon_va       srs siteTypeCd    hucCd
## 1                  EST   36.08722  -79.36611 EPSG:4326         ST 03030002
##   stateCd countyCd network
## 1      37    37001    NWIS
```

4. Build a plot called EnoPlot2. Use the base plot we made in class and make the following changes:

- Add a column to your data frame for discharge in meters cubed per second. hint: package dplyr in tidyverse includes a `mutate` function

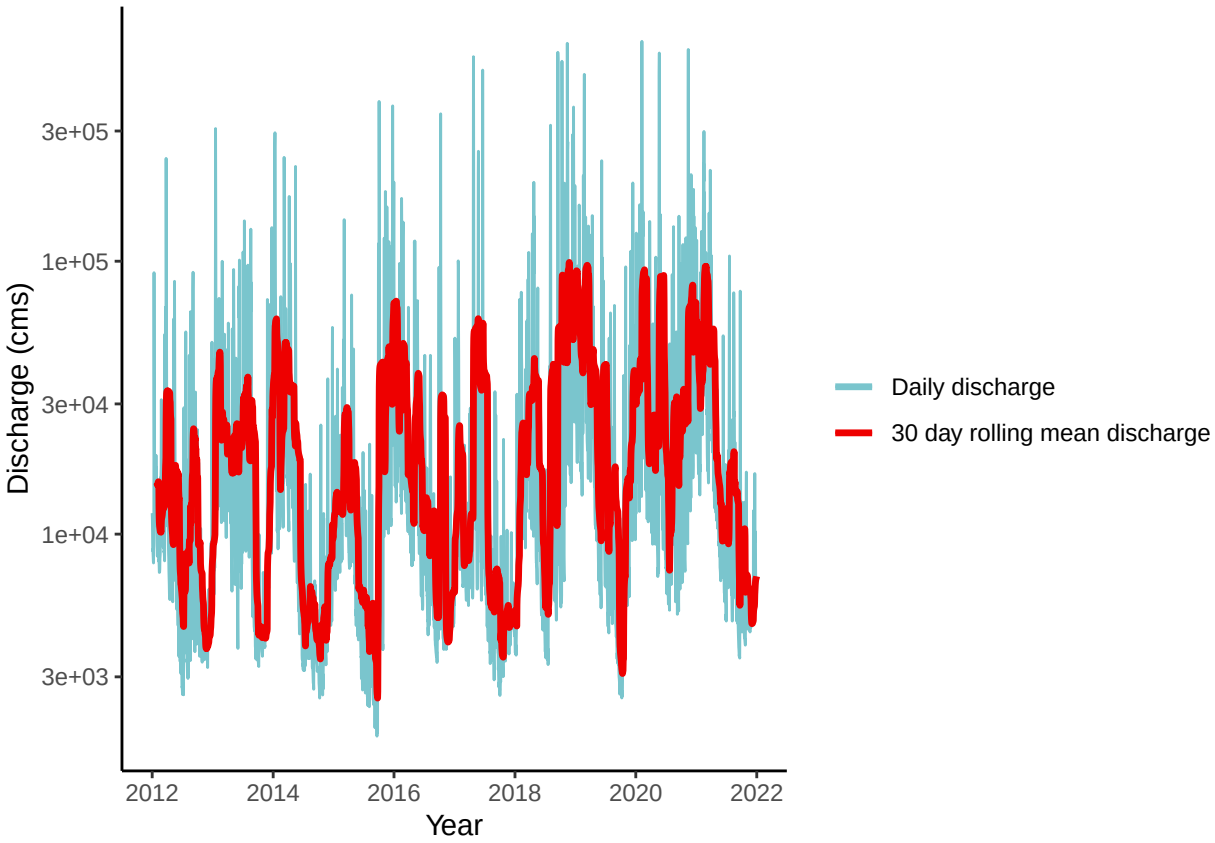
- Add a column in your data frame for a 30-day rolling mean of the metric discharge. (hint: package dplyr in tidyverse includes a `mutate` function. hint: package zoo includes a `rollmean` function)
- Create two geom_line aesthetics, one for daily discharge (meters cubed per second) and one for rolling mean of discharge. Color these differently.
- Update your ggplot theme. I suggest "classic." (hint: https://ggplot2.tidyverse.org/reference/ggtheme.html)
- Update axis names
- Change the y axis from a linear to a log10 axis (hint: google "ggplot logged axis")
- Add a legend. (hint: Google "add legend two geom layers ggplot")

```
# Add a column for discharge in cubic meters per second & 30 day rolling mean of metric discharge
   EnoDischarge <- EnoDischarge %>% mutate(discharge_cms = discharge_cfs * 35.315,
                                           rolling30MeanDischarge_cms = rollmean(discharge_cms, 30,
                                                                                 fill = NA,
                                                                                 align = "right")) #fir

# Plot daily and rolling mean discharge values
   EnoPlot2 <- ggplot(EnoDischarge, aes(x = date)) +
               geom_line(aes(y = discharge_cms, color = "Daily discharge")) +
               geom_line(aes(y = rolling30MeanDischarge_cms, color = "30 day rolling mean discharge"), s
               labs(x = "Year", y = "Discharge (cms)") +
               theme_classic() +
               scale_y_log10() +
               scale_color_manual(name = "", values = c("Daily discharge" = "cadetblue3",
                                                        "30 day rolling mean discharge" = "red2")) #see

   EnoPlot2
```

```
## Warning: Removed 29 row(s) containing missing values (geom_path).
```



5. In what ways was the second plot a more effective visualization than the first?

   ANSWER: The second plot smooths out some of the noise from the daily discharge values alone, allowing for both inter- and intra- year trends to be observed.

6. What portions of the coding were challenging for you?

   ANSWER: I'm not yet familiar with the grammar of data visualization used by ggplot, and it took me a while to figure out how to add a legend that showed what I wanted. Even now it feels like it was a somewhat roundabout and non-intuitive way of achieving that.

7. Interpret the graph you made. What are the things you notice about within- and across-year variability, as well as the differences between daily values and 30-day rolling mean?

   ANSWER: There are annual fluctuations between high and low flows, with low flows near the end of each year and in the summer, and higher flows in the spring. The precise magnitude of these highs and lows depend on the year, and there can be substantial variation beween years (though the range is, over the whole time period, fairly consistent). There is more fluctuation in daily values than in 30 day rolling means, both overalland within and between years.