Assignment 1: Introduction

Felipe Raby Amadori

OVERVIEW

This exercise accompanies the lessons in Hydrologic Data Analysis 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 Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FILENAME") prior to submission.

The completed exercise is due on 2019-09-04 before class begins.

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/fr55/Hydrologic_Data_Analysis.git

2. Complete the Consent Form in Sakai. You must choose to either opt in or out of the research study being conducted in our course.

Did you complete the form? (yes/no)

yes

Course Project

3. What are some topics in aquatic science that are particularly interesting to you?

   ANSWER: Water movement, the use of water by humans, groundwater study, watershed water balance.

4. Are there specific people in class who you would specifically like to have on your team?

   ANSWER: Tristen Townsend, Keith Bollt, Simon Warren

5. Are there specific people in class who you would specifically *not* like to have on your team?

   ANSWER: No

Data Visualization Exercises

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

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

getwd()
```

```
## [1] "C:/Users/Felipe/OneDrive - Duke University/1. DUKE/Ramos 3 Semestre/722 Hydro Data/Hydrologic_Da
```

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

theme_set(theme_classic())
```

7. Upload discharge data for the Eno River at site 02096500 for the same dates as we studied in class (2009-08-01 through 2019-07-31). Obtain data for discharge and gage height (you will need to look up these parameter codes). Rename the columns with informative titles. Imperial units can be retained (no need to change to metric).

```
EnoDischarge <- readNWISdv(siteNumbers = "02096500",
                     parameterCd = "00060", # discharge (ft3/s)
                     startDate = "2009-08-01",
                     endDate = "2019-07-31")

EnoGageHeight <- readNWISdv(siteNumbers = "02096500",
                     parameterCd = "00065", # gage height (ft)
                     startDate = "2009-08-01",
                     endDate = "2019-07-31")



# Renaming columns

names(EnoDischarge)[1:2] <- c("Agency.Discharge","Site_No.Discharge")
names(EnoDischarge)[4:5] <- c("Discharge_ft3.s","Discharge.Approval.Code")
names(EnoGageHeight)[1:2] <- c("Agency.GageHeight","Site_No.GageHeight")
names(EnoGageHeight)[4:5] <- c("GageHeight_ft","GageHeight.Approval.Code")

#Data exploring

str(EnoDischarge)
```

```
## 'data.frame':    3647 obs. of  5 variables:
##  $ Agency.Discharge       : chr  "USGS" "USGS" "USGS" "USGS" ...
##  $ Site_No.Discharge      : chr  "02096500" "02096500" "02096500" "02096500" ...
##  $ Date                   : Date, format: "2009-08-01" "2009-08-02" ...
##  $ Discharge_ft3.s        : num  186 129 123 118 84 112 161 88.6 74.8 71.6 ...
##  $ Discharge.Approval.Code: chr  "A" "A" "A" "A" ...
##  - attr(*, "url")= chr "https://waterservices.usgs.gov/nwis/dv/?site=02096500&format=waterml,1.1&Para
##  - attr(*, "siteInfo")='data.frame': 1 obs. of  13 variables:
##   ..$ station_nm          : chr "HAW RIVER AT HAW RIVER, NC"
##   ..$ site_no             : chr "02096500"
##   ..$ agency_cd           : chr "USGS"
##   ..$ timeZoneOffset      : chr "-05:00"
##   ..$ timeZoneAbbreviation: chr "EST"
##   ..$ dec_lat_va          : num 36.1
##   ..$ dec_lon_va          : num -79.4
```

```
##   ..$ srs                 : chr "EPSG:4326"
##   ..$ siteTypeCd          : chr "ST"
##   ..$ hucCd               : chr "03030002"
##   ..$ stateCd             : chr "37"
##   ..$ countyCd            : chr "37001"
##   ..$ network             : chr "NWIS"
##  - attr(*, "variableInfo")='data.frame': 1 obs. of  7 variables:
##   ..$ variableCode       : chr "00060"
##   ..$ variableName       : chr "Streamflow, ft&#179;/s"
##   ..$ variableDescription: chr "Discharge, cubic feet per second"
##   ..$ valueType          : chr "Derived Value"
##   ..$ unit               : chr "ft3/s"
##   ..$ options            : chr "Mean"
##   ..$ noDataValue        : logi NA
##  - attr(*, "disclaimer")= chr "Provisional data are subject to revision. Go to http://waterdata.usgs
##  - attr(*, "statisticInfo")='data.frame':    1 obs. of  2 variables:
##   ..$ statisticCd  : chr "00003"
##   ..$ statisticName: chr "Mean"
##  - attr(*, "queryTime")= POSIXct, format: "2019-09-03 19:23:53"
```

```
str(EnoGageHeight)
```

```
## 'data.frame':    3640 obs. of  5 variables:
##  $ Agency.GageHeight       : chr  "USGS" "USGS" "USGS" "USGS" ...
##  $ Site_No.GageHeight      : chr  "02096500" "02096500" "02096500" "02096500" ...
##  $ Date                    : Date, format: "2009-08-01" "2009-08-02" ...
##  $ GageHeight_ft           : num  2.13 1.89 1.85 1.84 1.66 1.77 2.02 1.69 1.61 1.59 ...
##  $ GageHeight.Approval.Code: chr  "A" "A" "A" "A" ...
##  - attr(*, "url")= chr "https://waterservices.usgs.gov/nwis/dv/?site=02096500&format=waterml,1.1&Para
##  - attr(*, "siteInfo")='data.frame': 1 obs. of  13 variables:
##   ..$ station_nm          : chr "HAW RIVER AT HAW RIVER, NC"
##   ..$ site_no             : chr "02096500"
##   ..$ agency_cd           : chr "USGS"
##   ..$ timeZoneOffset      : chr "-05:00"
##   ..$ timeZoneAbbreviation: chr "EST"
##   ..$ dec_lat_va          : num 36.1
##   ..$ dec_lon_va          : num -79.4
##   ..$ srs                 : chr "EPSG:4326"
##   ..$ siteTypeCd          : chr "ST"
##   ..$ hucCd               : chr "03030002"
##   ..$ stateCd             : chr "37"
##   ..$ countyCd            : chr "37001"
##   ..$ network             : chr "NWIS"
##  - attr(*, "variableInfo")='data.frame': 1 obs. of  7 variables:
##   ..$ variableCode       : chr "00065"
##   ..$ variableName       : chr "Gage height, ft"
##   ..$ variableDescription: chr "Gage height, feet"
##   ..$ valueType          : chr "Derived Value"
##   ..$ unit               : chr "ft"
##   ..$ options            : chr "Mean"
##   ..$ noDataValue        : logi NA
##  - attr(*, "disclaimer")= chr "Provisional data are subject to revision. Go to http://waterdata.usgs
##  - attr(*, "statisticInfo")='data.frame':    1 obs. of  2 variables:
##   ..$ statisticCd  : chr "00003"
##   ..$ statisticName: chr "Mean"
```

```
##  - attr(*, "queryTime")= POSIXct, format: "2019-09-03 19:23:54"
```

```
#Dates Correctly Formated. Discharge and Height data correctly formated as numeric.

summary(EnoDischarge$Discharge_ft3.s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    51.5   140.0   265.0   627.4   530.0 17800.0
```

```
summary(EnoGageHeight$GageHeight_ft)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.460   1.950   2.440   3.116   3.272  23.110
```

```
any(is.na(EnoDischarge$Discharge_ft3.s))
```

```
## [1] FALSE
```

```
any(is.na(EnoGageHeight$GageHeight_ft))
```

```
## [1] FALSE
```

```
#No NAs in the data

EnoDischarge_GageHeight <- EnoGageHeight %>%
left_join(EnoDischarge, by = "Date")
```

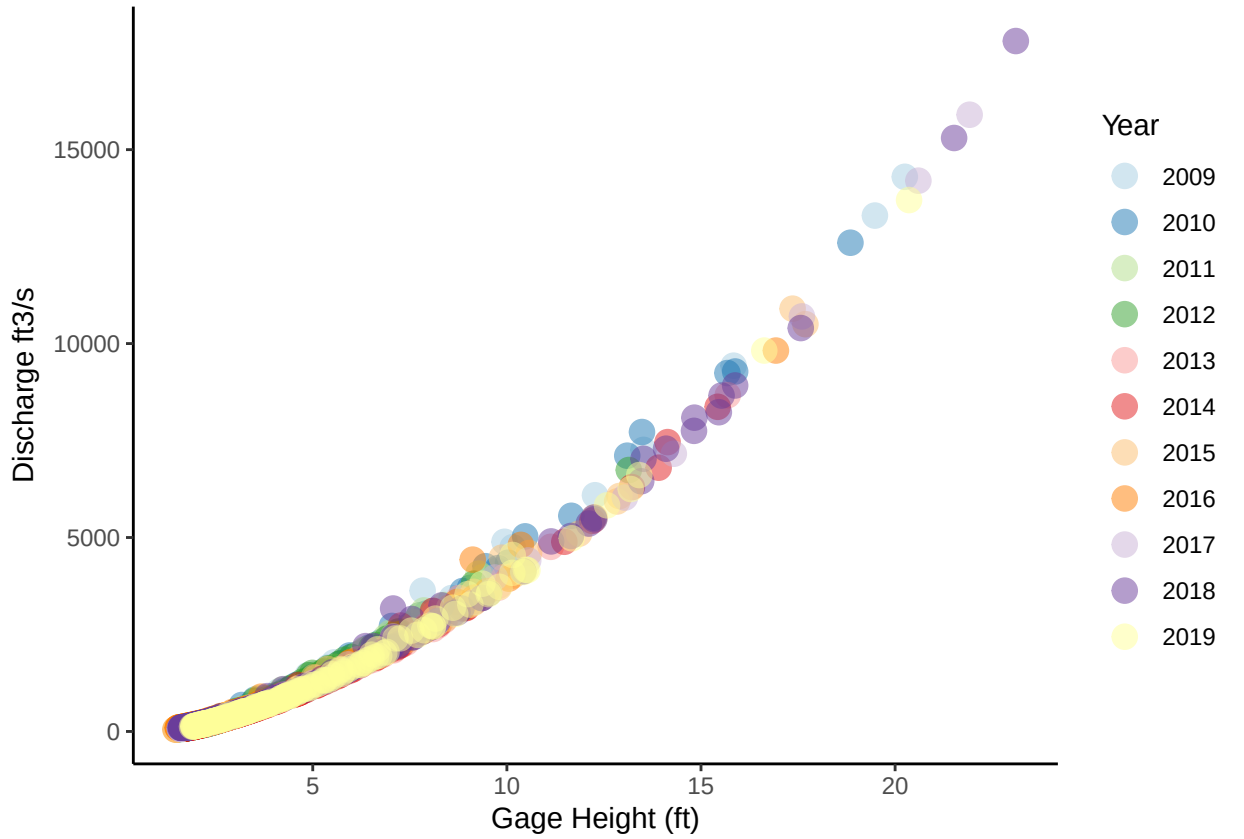
8. Add a "year" column to your data frame (hint: lubridate has a `year` function).

```
EnoDischarge_GageHeight <- mutate(EnoDischarge_GageHeight, year = year(Date))
```

9. Create a ggplot of discharge vs. gage height, with gage height as the x axis. Color each point by year. Make the following edits to follow good data visualization practices:

- Edit axes with units
- Change color palette from ggplot default
- Make points 50 % transparent

```
ggplot(EnoDischarge_GageHeight, aes(x = GageHeight_ft, y = Discharge_ft3.s)) +
  geom_point(aes(color = factor(year)), size = 4, alpha = 0.5) +
  scale_color_brewer(palette = "Paired") +
  xlab(expression("Gage Height (ft)")) +
  ylab(expression("Discharge ft3/s")) +
  labs(color = 'Year')
```



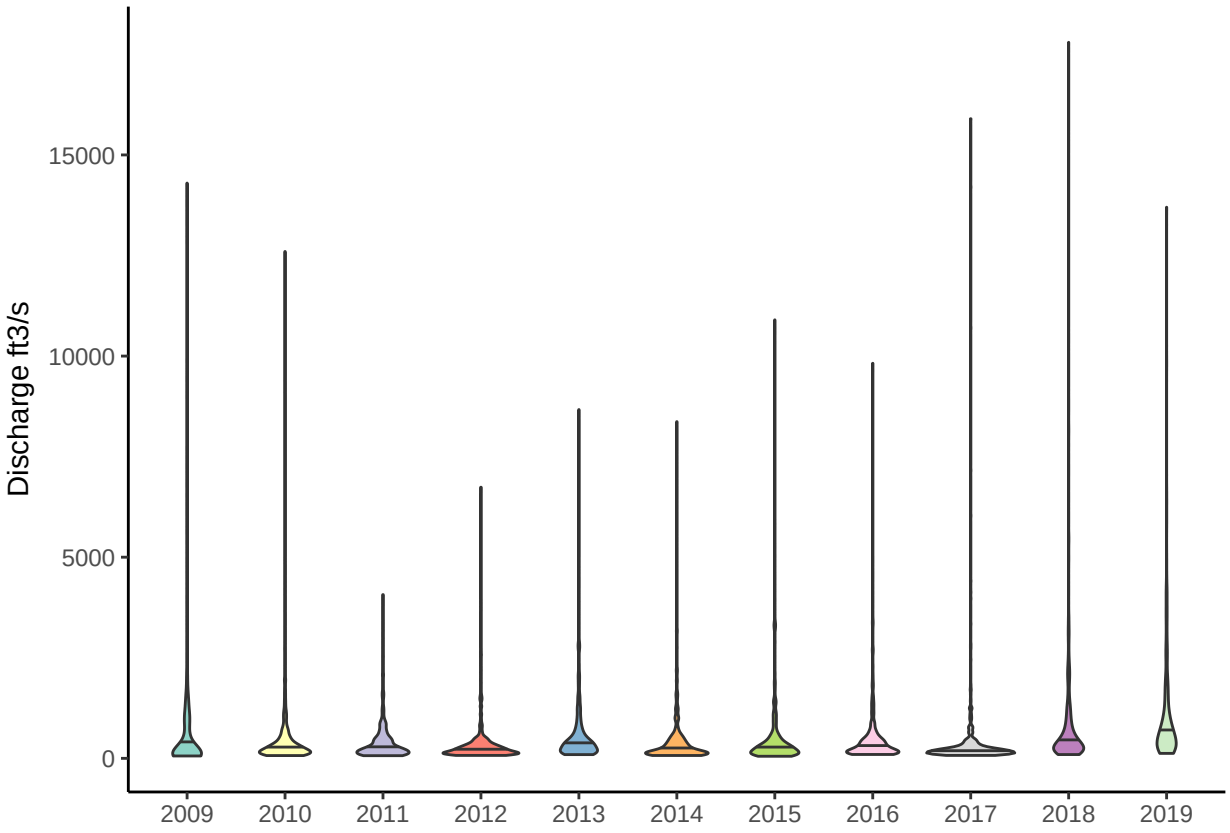
10. Interpret the graph you made. Write 2-3 sentences communicating the main takeaway points.

    ANSWER: The Gage Height vs Discharge function appears to be parabolic although there are years that some of the values fall slightly outside the curve such as 2010, 2016, 2018. This phenomenon could be produced by a decalibration in the sensors that are measuring gage height and flow. It also could be produced by a bad designed gage that under certain conditions the height in it is influenced by downstream flow. 2018 and 2017 have the highest Discharge and Gage Height values recorded.

11. Create a ggplot violin plot of discharge, divided by year. (Hint: in your aesthetics, specify year as a factor rather than a continuous variable). Make the following edits to follow good data visualization practices:

- Remove x axis label
- Add a horizontal line at the 0.5 quantile within each violin (hint: draw_quantiles)

```
ggplot(EnoDischarge_GageHeight, aes(x = factor(year), y = Discharge_ft3.s)) +
  geom_violin(aes(fill = factor(year)), draw_quantiles = 0.5) +
  scale_fill_brewer(palette = "Set3") +
  xlab(expression("")) +
  theme(legend.position = "none") +
  ylab(expression("Discharge ft3/s"))
```



12. Interpret the graph you made. Write 2-3 sentences communicating the main takeaway points.

    ANSWER: In the graph it can be observed that the data is positively skewed as it is expected for flow data. The tail on the right side of the distribution is definitely longer. The year with highest discharge recorded is 2018. Years 2017 and 2019 were also high discharge years. The year with lowest discharge recorded was 2011, which started a period of relatively low discharges up to year 2016. Even though year 2017 has high discharge values recorded, it has a relatively low median compared to the others years considered in the study.