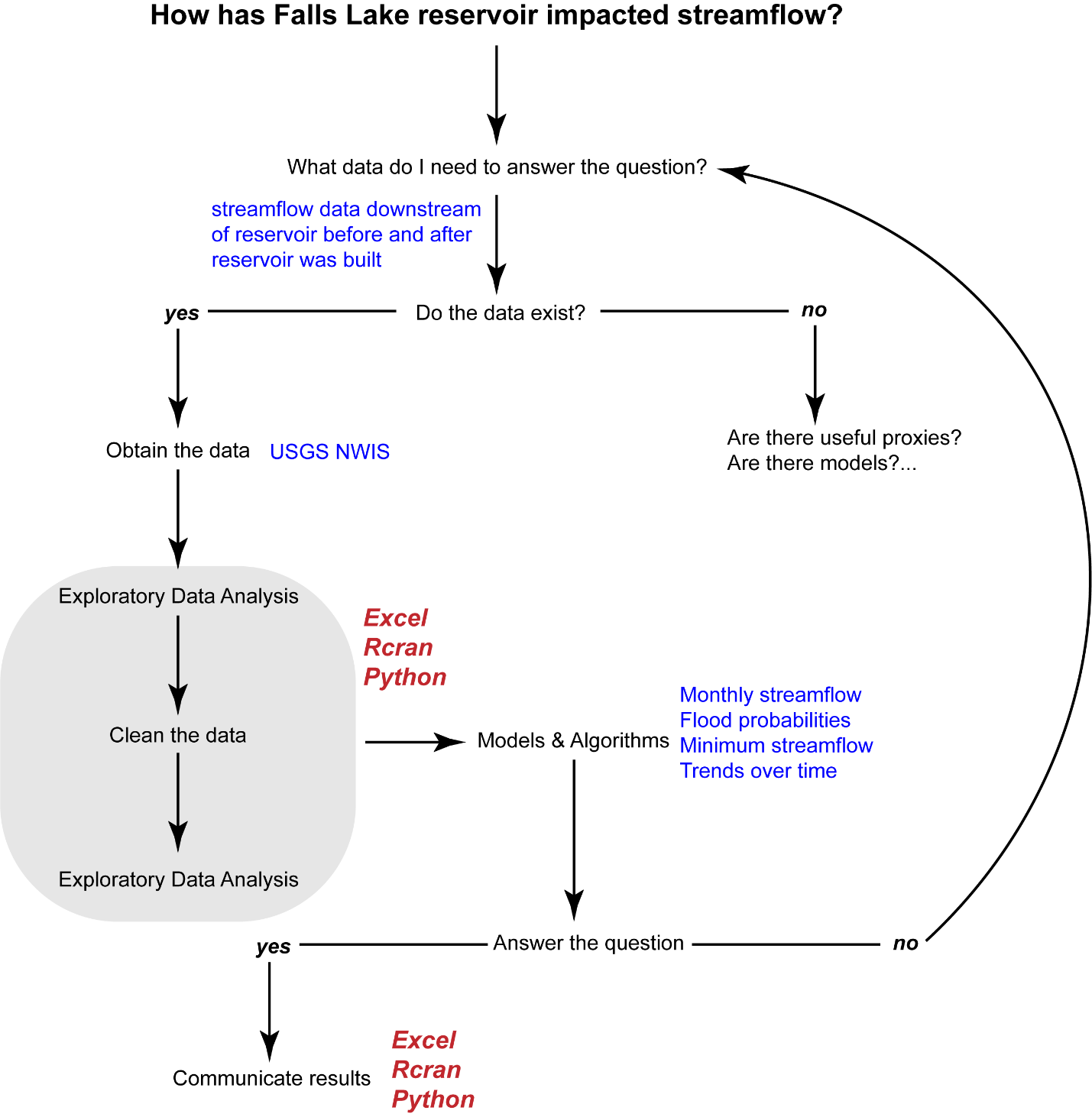
Unit 1: Streamflow Data

# Rcran Exercise:

USGS actively supports R: <https://owi.usgs.gov/R/packages.html>

Information on how to use data Retrieval: <https://owi.usgs.gov/R/dataRetrieval.html#2>

Need to **install** packages (one time only) and **load** libraries each session.



# Obtain the data

1. Use daily discharge data for Neuse River near Clayton and Goldsboro
   1. Save usgs file into a .csv file

OR

* 1. Install and load the following packages into Rstudio
     1. dataRetrieval
     2. Identify site No, parameter code, statistical code, start and end dates
  2. Use readNWIS… function to load in data of interest

Library: dataRetrieval and EGRET

If ever you need help: type in the function with a question mark in front of it. For example, ?plot

Some packages will have vignettes that will walk you through examples. You can find them online or load them directly into R

# Explore Data: Visualize with plots

1. Plot using basic package: plot()
2. Plot using ggplots: ggplot() 🡪 library ggplot2

## ED: Confidence Level and Summary Statistics

1. How confident are we in the data?
   1. How many observations are there?
      1. Use dim() to answer
   2. How many observations were in each qualification code?
      1. Use table() to answer
      2. Transform into a dataframe to manipulate: as.data.frame()
   3. What percentage of observations fall into each category
   4. How confident are we in the quality of the data at this location?
      1. Create a pie chart to quickly visualize information: pie()

**There are many ways to arrive at the same answer**

1. Summary Statistics: FUNCTIONS
   1. Calculating with a function 🡪 create a process and call that process to run with different variables.
      1. Here we create the function “gen\_stats”
      2. There are two parameters we want to read into the function:
         1. Data
         2. Column number of the table we want to fill. In this instance, we will create the table outside of the function.
      3. Within the function we want to do the following:
         1. Calculate statistics:
            1. Min()
            2. Quantile(data, percentile (0.1, 0.25, 0.5, …)
            3. Median()
            4. Mean()
            5. Max()
         2. Fill the table based on row and column number
   2. Create a dataframe:
      1. As.data.frame(matrix(nrow, ncol))
         1. A dataframe reads [row,col] and you can point to a cell by listing the row and column number you wish to fill.
      2. Colnames(dataframe) 🡪 allows you to provide headings to the columns in your data frame
      3. Fill in the descriptor column by calling the column of interesting and filling the row with a vectors.
   3. To divide the data between two periods, we create 2 new variables and use subset()
      1. Subset(data, column…)
   4. Call the function with the data of interest and fill the appropriate column
2. Summary Statistics: DPLYR and PIPES
   1. Library: dplyr and magrittr
   2. Create the dataframe to fill
      1. Pipes allow you to continue a train of manipulation without recalling the data over and over
         1. For example, we are pointing to the data once, filtering the data once (not creating a new dataset), and calling statistics as we go.
      2. Lots of useful functions with dplyr and magrittr
   3. The data don’t necessarily format into something usable – but that is easy to fix with t()
   4. Transpose into a data frame and add column headers
   5. Did the summary statistics change with the construction of Falls Lake Reservoir (1981-1984)?

## ED: Seasonal variation in streamflow

1. Parse out the Date to include Year and Month: lubridate library
   1. Can calculate multiple ways
      1. Dplyr
      2. Function
      3. For Loop
         1. For loop allows you to loop through a data frame, subsetting for specific conditions. In this case we will subset by month and perform the same calculation over and over again, filling a data frame with results as we go.
   2. Plot Results
      1. What do you observe?

## EXTRA 🡪 LOOK FOR AND LOAD MULTIPLE SITES

1. Load in sites for NC
   1. Filter by parameter, statistics, and period
   2. Map site locations in R
   3. Opportunities to loop through multiple sites and perform statistics…
      1. Simply plotted flow on a graph to demonstrate

# Has the reservoir impacted the 100 year flood frequency?

1. Calculate the Water Year
   1. Ifelse()
2. Calculate the maximum annual streamflow
   1. Use dplyr to group by year and grab max flow
   2. Use arrange() to sort flows in descending: desc() order
   3. Use rank() to create a rank number from 1 to n
   4. Calculate the Return Interval, RI = (n+1)/m; where n is the number of years of data and m is the rank
   5. Calculate the annual probability of this flow occurring in any given year
3. Plot your recurrence interval on the x-axis and your max discharge on the y-axis
   1. Place the x-axis on a log scale
   2. Add a regression and find the one with the best fit: lm()
   3. Using the best-fit regression – estimate the discharge for the 100-year, 500-year and 1000 year events: predict()
4. Turn this into a function and recalculate using only data from 1930-1979 and then for 1984-2017
   1. Calculate the discharge for different return periods and exceedance probabilities
5. Add data to the plot and compare
6. Create a function that prints out the probability of experiencing a 100, 500, or 1000 year flood during x years.
   1. Use the print() function to print results
   2. Use the paste0() function to intersperse text with variables

# How has Falls Lake impacted minimum flows?

1. Load (EGRET) package and read in daily data. Notice that Q7 is already calculated

<https://cran.r-project.org/web/packages/EGRET/vignettes/EGRET.pdf>

Notice the plotFour() 🡪 you can see a big impact on minimum flows in summer months following the installation of Falls Lake

1. Load in data
   1. Use TTR package to calculate rolling 7 day average
      1. SMA()
   2. Use dplyr package to calculate minimum Q7 each year
   3. Calculate the Rank, RI and probability
2. Plot the 7Q flow with the Probability factor
   1. Fit with best fitting regression
   2. Estimate the 7Q10 using the linear regression.
3. How many days were below 7Q10 in the POR?
4. When did the low flow days occur relative to Falls Lake?
5. How does the answer change if we only use data post Falls Lake?
   1. How big of a difference is it?
6. Did Falls Lake increase or decrease the minimum flows?

# How has streamflow changed over time and by season?

In excel, the only regression options

1. Mann Kendall tests (non-parametric assumptions) are commonly used for streamflow analyses.
   1. This requires installing and loading the trend package
   2. We will aggregate total streamflow by water year for the annual trend analysis
      1. Use dplyr to do this
   3. Use mk.test() to extract the pvalue and sen() to get the slope and confidence intervals
      1. How does this compare with the linear regression?
2. Repeat for 1930-1979 and for 1984-2017
   1. What do you observe?
3. If there are not annual trends, are there seasonal ones?
   1. Use a for loop to answer and fill in a data frame with results

# Grab all streams in North Carolina with 30 years of data and determine which, if any have significant changes in annual streamflow?

1. Use readNWISdata to find all NC streamgauge sites
2. Filter those sites for 30 years of data and with beginning dates prior to 1960 and ending dates after 2010
3. Set up a for loop to calculate the annual change in streamflow
   1. Constrain time period to 1970-2017
   2. Fill those results into a data frame
4. Map the direction (sen slope) and significance (tau pval) of trends
   1. How many trends were positive? Negative?
   2. How many trends were significant?
   3. Are there any spatial patterns?