R: Web Scraping

Web scrapping in R is accomplished through the **rvest** package.

 library(rvest)

Example #1

Website: <http://course1.winona.edu/cmalone/dsci210/Exams/USQuickFacts_Webscrape1.html>



 HTML Snipit



The data to be scraped sits inside a HTML table, thus the function html\_table() should be used to put data into a useable R format.

url <- "http://course1.winona.edu/cmalone/dsci210/Exams/USQuickFacts\_Webscrape1.html"

page <- read\_html(url)

temptable <- html\_table(page, header=TRUE)

df<-as.data.frame(temptable)

Data put into a data.frame in R



Getting summaries for Avg. Travel Time to Work

summary(df$Avg.Travel.Timeto.Work)
hist(df$Avg.Travel.Timeto.Work)





Example #2

Website: <http://course1.winona.edu/cmalone/dsci210/Exams/USQuickFacts_Webscrape2.html>



HTML Snipit



R code to read in this version of webpage. Note the [[2]] is pulling off the second table instance on this webpage.

url <- "http://course1.winona.edu/cmalone/dsci210/Exams/USQuickFacts\_Webscrape2.html"

page <- read\_html(url)

temptable <- html\_table(page, header=TRUE)[[2]]

df<-as.data.frame(temptable)

This data as read in my R and put into a data.frame in R



Example #3

Website: <https://www.ssa.gov/oact/babynames/numberUSbirths.html>



HTML Snipit



Chrome Extension: SelectorGadget

The SelectorGradget extension for Chrome will allow us to identify the location of content to be scraped.



Using the SelectorGadget inside Chrome.



R code to scrape this data

url<-"https://www.ssa.gov/oact/babynames/numberUSbirths.html"

page <- read\_html(url)

temp<-html\_text(html\_nodes(page, ".border td"))

df<-as.data.frame(temp)

Structure of data.frame is not correct, no rows…



Data processing steps. Note: gsub() function is necessary to remove comma in counts before converting to numeric.

rowid<-rep(c(1,2,3,4),135)

Year<-as.numeric(as.character(df$temp[rowid==1]))

Males<-as.numeric(gsub(",","",df$temp[rowid==2]))

Females<-as.numeric(gsub(",","",df$temp[rowid==3]))

Getting a simple plot to compare trends over time between Males and Females.

plot(Year,Males,type='n')

lines(Year,Males,col="blue")

lines(Year,Females,col='pink')

