Notes 4/15/19

Continuation of Regex

* Escape metacharacter
  + \. OR \$ OR \^ For most engines
  + \\. OR \\$ OR \\^ In R language
    - Where \ OR \\ come before any given metacharacter

Character Classes to manipulate strings in R (both work):

* Base R functions
  + Regex
    - Ex: “[0-9]\\.[0-9][0-9]”
* Package “stringr”
  + Str\_verb()
  + Ex: (all are functions)
    - Detect
      * Str\_detect() 🡨 returns a vector
      * Str\_match() 🡨 returns an array
    - Extract a pattern
      * Str\_extract
    - Replace a pattern
      * Str\_replace
    - Split strings
      * Str\_split
    - Str\_sub
      * Extract a subset of the string value
      * Str\_sub(table$variable[1:?], start = ?, end = ?)
* Ex:
  + Read\_html() extracts from an html to get a data frame
    - Str\_detect(table$variable[row], “[0-9]\\.[0-9][0-9]”)
      * Can use str\_locate to find start and end
* Manipulating Strings
  + Always start small
    - One cell, not entire dataframes or variables/columns
* Ex:
  + Long jump record and country
    - Peter <- “Peter O’Connor (IRE)”
    - All countries have “(ABC)” pattern like peter
  + Using stringr
    - Str\_extract(peter, “\\([A-Z][A-Z][A-Z]\\)” ) 🡨 with parentheses
    - Without parentheses: Str\_extract(peter, “[A-Z][A-Z][A-Z]”)
    - Using str\_sub()
      * Without parentheses: Using str\_sub(peter, start = nchar(peter)-3, end = nchar(peter)-1)
* Ex:
  + Getting characters based on position
  + Str\_extract(peter, “\\w”) 🡪 “P”
  + Str\_extract(peter, “\\w\\w”) 🡪 “Pe”
  + Str\_extract(peter, “\\w+”) 🡪 “Peter”
    - Will return all characters not spaced apart
  + Can get last name to (just not for the apostrophe)
  + Str\_extract(peter, “\\w+ \\w+”) 🡪 “Peter O”
    - Works for other names as “John Smith”
  + Can use str\_sub(table$variable, start = 1, nchar(table$variable)-6) 🡪 last and first, even commas and dashes, without the country name