Convert_NA_in_DF

Kurt Stoneburner

6/27/2020

Scenario:

Convert every NA in a data frame to something else. In this example, NA is converted to 0.

Code:

“‘ { r convert_na_in_df, echo=TRUE }

any_df = data.frame()

for (colCount in 1:length(any_df)) {

thisColumn <- unlist(any_df[colCount])

new_column <- sapply(thisColumn, function(x){ if (is.na(x)) { return(0) } else { return(x) } },simplify=“array”)

any_df[colCount] <- new_column

}

head(any_df)

```
## Initial Loop ##

**for (colCount in 1:length(any_df ) ) **
Generates a counter from 1 to length of data frame (This loop works well in python too).

## Get each Column ##

**any_df[colCount]**
We want the column as a vector, but this returns a list.

**thisColumn <- unlist(any_df[colCount])**
unlist() converts a list to a vector.


**new_column <- sapply(thisColumn, function(x){ ... },simplify="array")**
sapply() returns a vector when using simplify="array". sapply() (along with lapply(), mapply()) is the

**any_df[colCount] <- new_column**
Replace the old column vector with the new one.
```

~~~~
Getting a handle on loops is a big deal. Here is one more scenario. How about calculating the differenc
Code

‘‘‘ { r vector_differences, echo=TRUE }

some_vector <- c(1,4,5,2,12,35,2,5,6,10)
i <- 0
diff_vector <- sapply(some_vector, function(x){
    i <<- i+1

  if (i == 1) {
    return(0)
  }

  return(some_vector[i]-some_vector[i-1])

},simplify="array")
~~~~

diff_vector = 0 3 1 -3 10 23 -33 3 1 4

One issue with sapply() is that it returns value of the vector. To get the difference between two values the index must be referenced.

i < - 0

Initializes a standard counter outside of the sapply() loop.

i < < - i+1

The < <- references i from outside of the loop. A single <- would reference an i that was internal to the loop and would never increment.

return(some_vector[i]-some_vector[i-1])

Gets a value at index (i) and subtracts the value from the previous index (i-1).

This technique opens a lot of possibilities for generating derived data. Using this technique it would be straightforward to generate an average of the previous 7 numbers. When applied to COVID numbers this would be the now popular 7-day rolling average metric.