How to find dataset differences in R, when the pieces of information are changing between datasets it’s a difficult task to identify the same.

Here we are going to discuss the daff package in R, daff package helps us to identify the differences and visualize them in a beautiful way.

**Features**

Its compare rows and columns and provides information on

* Value changes
* Which new row/column added
* Which row/ column removed
* How many rows /columns added/removed?

Let’s see how to execute the daff package in R.

**Load Library**

First load the daff package in R.

If you are not installed the package, you can install the daff package from CRAN

install.packages("daff")

Now we need to create two data frames and let’s name them mydata1 and mydata2.

library(daff)

mydata1<-data.frame(Name=c("P1","P2","P3","P4","P5"),col1=c(1,2,3,4,5),col2=c(11,13,14,15,17))

mydata2<-data.frame(Name=c("P1","P2","P6","P4","P5"),col1=c(1,3,3,6,9))

Let’s see how to find the dataset differences

Daff Comparison: ‘mydata1’ vs. ‘mydata2’

---

@@ Name col1 col2

P1 1 11

-> P2 2->3 13

+++ P6 3

--- P3 3 14

-> P4 4->6 15

-> P5 5->9 17

**->** Indicate the updated value.

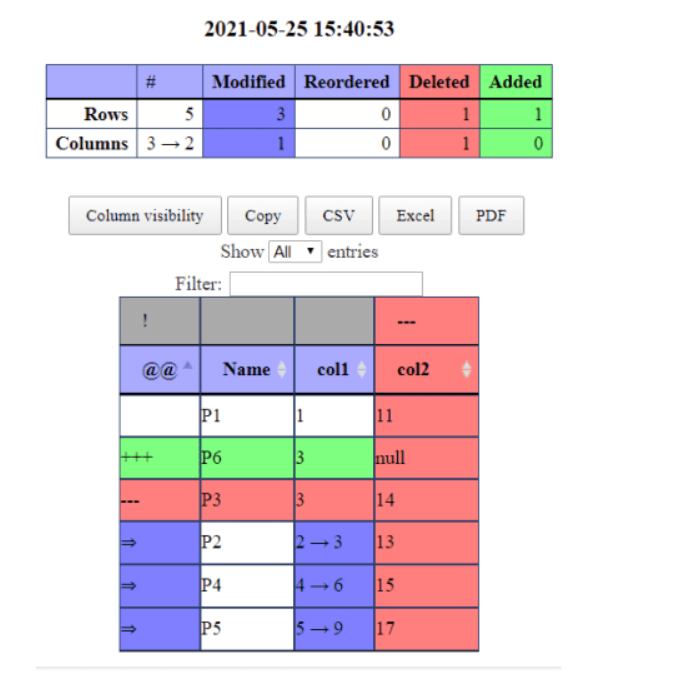
**—**Indicate the row removed from the dataset

**—**If the symbol appears above the column indicate the column removed from the dataset.

**+++**indicate rows which are added

Basically, **->** for updated data, **–** for removed data and **+** for added data.

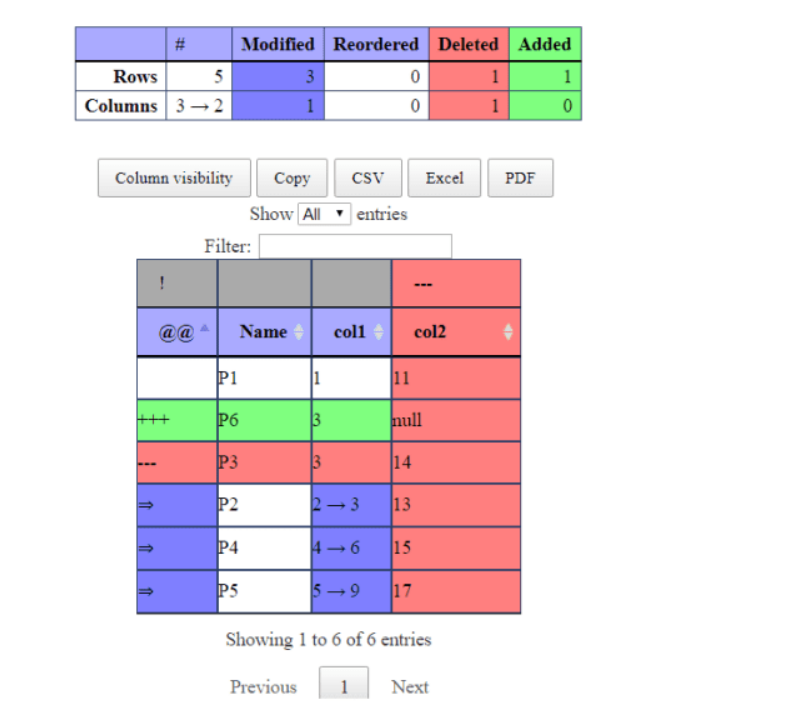
You can visualize the same using render\_diff



The HTML output will be displayed in different colors.

Suppose if you want to compare the datasets based on the primary columns, you can just pass the information like,

render\_diff(diff\_data(mydata1,mydata2,id=c("Name","col1")))



You can provide multiple column names as a primary column. Like c(“Name”,”col1”)

In HTML view, you can download the results into multiple formats like CSV, Excel, and pdf.

Sorting option and filter option also available in HTML view.

In daff package many other arguments are available you can make use of the same.

diff\_data(data\_ref, data, always\_show\_header = TRUE,

  always\_show\_order = FALSE, columns\_to\_ignore = c(),

  count\_like\_a\_spreadsheet = TRUE, ids = c(),

  ignore\_whitespace = FALSE, never\_show\_order = FALSE,

  ordered = TRUE, padding\_strategy = c("auto", "smart", "dense",

  "sparse"), show\_meta = TRUE, show\_unchanged = FALSE,

  show\_unchanged\_columns = FALSE, show\_unchanged\_meta = FALSE,

  unchanged\_column\_context = 1L, unchanged\_context = 1L)

Sometimes we need to ignore some particular columns.

**columns\_to\_ignore** argument will be very useful (character List of columns to ignore in all calculations.

Changes related to these columns should be discounted)

You can export the comparison results into a CSV file, for that you need to store the result into variables.

out<-diff\_data(mydata1,mydata2,id="Name")

write\_diff(out,"D:/RStudio/daff/Result.csv")

Suppose if you want to read the output, you can make use of read\_diff

read\_diff("D:/RStudio/daff/Result.csv")