REVOLUTION ANALYTICS

Introduction to Merging Data with R

Revolution Analytics

Outline

Overview

In this session we'll cover data merging. The objectives of this session are to learn how to:

- Combine two datasets into a single data set that has variables from both of the initial data sets

Outline

Merging data sets

- Many uses
- Most common in relational databases
- Table 1 might have a variable we care about, but Table 2 might have everything else
- How do we put them together?

In R, we use the `merge()` function.

Reading some public data

For this example we'll use a data set detailing average national incomes and alcohol consumption:

```
richest.nations <- read.csv("http://opendata.socrata.com/views/7nh3-7ib4/rows.csv?accessType=DOWN
  header = TRUE)

richest.nations$Countries <- gsub(":", "", richest.nations$Countries)
richest.nations$Amount <- gsub(" per capita", "", richest.nations$Amount)
richest.nations$Amount <- gsub(",", "", richest.nations$Amount)
richest.nations$Amount <- sapply(richest.nations$Amount, function(x) substr(x,
  2, nchar(x)))
richest.nations$Amount <- as.numeric(richest.nations$Amount)
head(richest.nations)
```

```
##   Rank   Countries   Amount
## 1   #1  Luxembourg 89563.63
## 2   #2      Norway 66964.36
## 3   #3     Iceland 53029.30
## 4   #4     Ireland 52892.89
## 5   #5       Qatar 52239.72
## 6   #6 Switzerland 51032.66
```

Inspecting the data

```
alcohol.consumption <- read.csv("http://opendata.socrata.com/views/hj43-2bpj/rows.csv?accessType=
  header = TRUE)
dim(alcohol.consumption)
```

```
## [1] 195   2
```

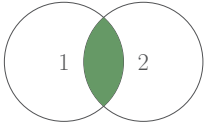
```
head(alcohol.consumption)
```

```
##              Location Liters.per.capita.pure.alcohol.adult.consumption
## 1         Afghanistan                                              0.01
## 2             Albania                                              2.01
## 3             Algeria                                              0.15
## 4             Andorra                                                NA
## 5              Angola                                              3.86
## 6 Antigua and Barbuda                                              5.73
```

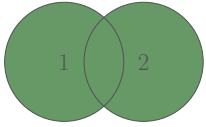
Types of merge

Using the argument `all=...` in `merge()` to change the merge type:

Type = inner
merge(x, y, all=FALSE)

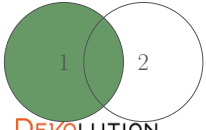


Type = full
merge(x, y, all=TRUE)

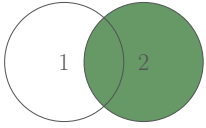


Type = oneToOne
cbind(x, y)

Type = left
merge(x, y, all.x=TRUE)



Type = right
merge(x, y, all.y=TRUE)



Type = union
rbind(x, y)

Merging data sets, all=FALSE

With `merge()` we can join these two datasets using the `Countries` and `Location` columns.

```
new.data.frame <- merge(x = richest.nations, y = alcohol.consumption,
  by.x = "Countries", by.y = "Location", all = FALSE)
dim(new.data.frame)
```

```
## [1] 171   4
```

```
head(new.data.frame)
```

```
##     Countries                                          Rank   Amount
## 1             Historical countries, unions or other regions:       NA
## 2             Historical countries, unions or other regions:       NA
## 3                                          European Union 24217.28
## 4                                          European Union 24217.28
## 5 Afghanistan                                        #197   270.44
## 6     Albania                                        #112  2911.90
```

Merging data sets, all=TRUE

```
new.data.frame.2 <- merge(x = richest.nations, y = alcohol.consumption,
  by.x = "Countries", by.y = "Location", all = TRUE)
dim(new.data.frame.2)
```

```
## [1] 237   4
```

Exercise 1: Practice with `merge()`

- Read in the `Vlookup.csv` file from our course `data` directory
- Split the `value` and `metric` tables into two different objects.
  - Hint: Remember column indexing from prior session.
- Then merge these objects using the key column.
- Write this modified table to file.

Advanced Exercise

- What happens if both names in the component data.frames are the `Key` and `Value`?
- What happens if that is the case, but you specify `by`?

Module review questions

- What is a data merge?
- What command does this in R, and what are the key arguments?

REVOLUTION ANALYTICS

Thank you

Revolution Analytics is the leading commercial provider of software and support for the popular open source R statistics language.

www.revolutionanalytics.com
1.855.GET.REVO
Twitter: @RevolutionR