

# Introduction to R for Data Management and Analysis

Marcel Ramos, MPH

Session 6

# Announcements

- One more week to go
- Next classes
  - Data Analysis workflow
  - Reporting and reproducibility

# Topics for today

- Merging datasets
  - Identifying duplicate observations
- Review for loops, plotting systems
- Functions / functionals
- Example where we use functions and loops
- Overview of summary tables and statistical tests

# Working in R

*How to actually learn any new programming concept*



*Essential*

Changing Stuff and  
Seeing What Happens

# Merging datasets

- Duplicate observations can be observed with duplicated()
  - Remove these first before merging
- rbind / cbind functions require equal dimensions
  - whether binding by rows or columns
  - row binding requires same column names (colnames)!
- merge function allows binding between unequal dims
  - by argument to tell R what variable to use as the ID
  - no sorting required
- tidyverse: \*\_join type of functions
  - full\_join
  - auto-insertion of NA values

# Functions

- Extend the language
  - Portable
- Group operations for ideally one purpose
- Pure functions - input is the same class as the output
- Well defined inputs and output
- Save you from repeating code
- Increase the flexibility of what R can do

# Structure of a loop

{Pseudocode}

- for loop structure

```
for (variable in vector) {  
  # do something here with  
  variable  
}
```

# Structure of a function

- function structure

```
functionname <-  
  function(argument1 = "default1", argument2 = "default2")  
{  
  anotherfunction(argument1, argument2)  
}
```



# More on functions

argument  
names

```
functionName <- function(argument1, argument2, ...) {
```

```
## body of function ##
```

```
## do something with argument1 and argument2
```

```
return(value)
```

```
}
```

function  
keyword; does  
not change

Good for sending  
additional arguments  
to functions inside the  
body

Curly braces will start and  
end the function (>1 line);  
they indicate *expressions*

```
function( arglist ) expr  
return(value)
```

❖ The body of the function will include the operations to

## Notes on Tuesday's lecture (cont..)

- Functions are powerful tools
- Minimize errors
- Create a set of operations to achieve a goal
- Easy to write
  - Predictable input
  - Predictable output
- Loops are useful but are not easily extensible

# Functionals

- Functional - an argument to a function that it itself is a function
- Many functions can accept other functions as an arguments
  - aggregate, tapply, lapply, sapply, apply, etc.
- Make coding more efficient and customizable
- Increased flexibility but add a layer of complexity
- Why use them?
  - To simplify code and avoid repetition

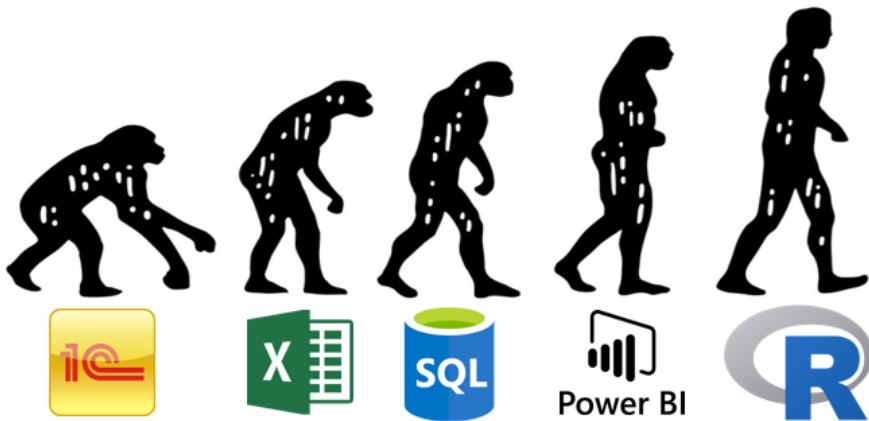
# Air Pollutants Example

- R Programming on Coursera
- Write a function
- Read multiple files at once
- Compute a summary statistic

# Why don't we use Excel?



# Ranking Statistical Software



# Mini Review Session

- Zero-level R Tutorial

# Common Errors and Troubleshooting

- R Basics Chapter



- R is particularly good at statistics
- Packages with new methods get published faster
- Extensibility is an MAJOR advantage compared to other software

# Descriptives

- Frequency tables
  - table function
    - prop.table
- gmodels package
  - CrossTable function
- mean and sd functions

# Statistical Tests

- `chisq.test` function
  - categorical 2x2
- `fisher.test` function
  - categorical with correction for small cells
- `t.test` function
  - categorical (2 levels) & continuous

# Useful functions to apply on model objects

- Functions that work on lm class objects
  - summary
  - fitted
  - resid
  - predict

# Tidy model results with broom

- Use the broom package to clean up results from model functions
  - tidy - model coefficients
  - augment - fitted/residual values and more
  - glance - model level statistics

# Linear Regression

- lm function
- UCLA walk-through

# Logistic Regression and Odds Ratios

- glm function
- Odds Ratio calculation
- [UCLA tutorial](#)

# Community driven development

GitHub, Inc. [US] | <https://github.com/pulls?q=is%3Apr+author%3ALiNK-NY+archived%3Afalse+is%3Aclosed> ☆ 🔔 ⏻ 1

Created Assigned Mentioned

2 Open ✓ 56 Closed

Visibility Organization Sort

**trim trailing ws in versioned deps #366**  
I'm getting issues with a space in the version comparison operator <. I have added a ...  
r-lib/master ← LiNK-NY:master

**travis-ci/travis-build R: update #1707** by LiNK-NY was merged on May 13  
You commented and opened 5

**r-lib/remotes trim trailing ws in versioned deps ✓**  
#366 by LiNK-NY was merged 23 days ago 8

**Bioconductor/BiocManager Informative message, resolves #47 ✗**  
#49 by LiNK-NY was merged on May 10

**Bioconductor/bioconductor.org update install page**  
#26 by LiNK-NY was merged on May 2

**Bioconductor/AnVIL\_rapiclient bug fix: single bracket list subset with vector in get\_message\_body**  
#1 by LiNK-NY was closed on Apr 29 2

**seandavi/BiocPkgTools Make use of `biocPkgList` ✗**  
#32 by LiNK-NY was merged on Apr 23

**seandavi/BiocPkgTools Data pkg ✗**  
#30 by LiNK-NY was merged on Apr 23

**seandavi/BiocPkgTools biocBuildEmail - notify maintainers with email template ✗**  
#29 by LiNK-NY was merged on Apr 22 2

**Bioconductor/AnVIL Document and export Service constructor**  
#10 by LiNK-NY was merged on Apr 28



## GitHub assignment (assigned next week)

- Signup on <https://github.com/>
- Look for the assignment to be posted under <https://github.com/CUNYSPHcode/>
- Fork the repository (will contain an .Rmd file)
- Upload your .Rmd file with the answers
- Create a pull request to submit your .Rmd file