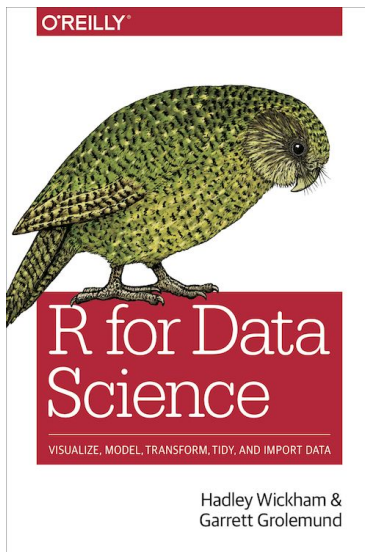


Introduction to data science with R and the tidyverse



Daniel Lundin

<http://r4ds.had.co.nz/>



The “Excel view” of data, a.k.a. “wide” format

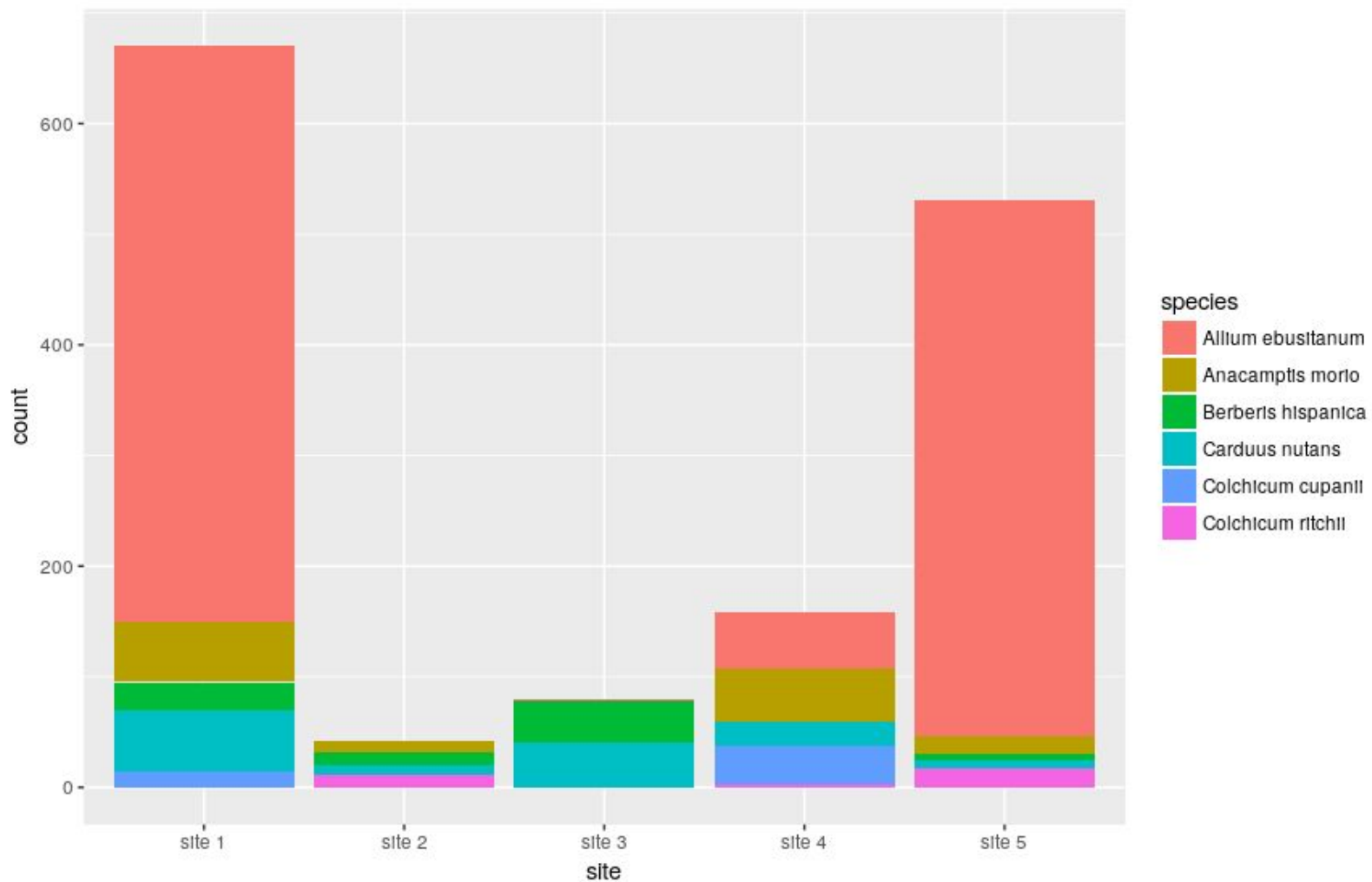
species	order	site 1	site 2	site 3	site 4	site 5
<i>Berberis hispanica</i>	<i>Ranunculales</i>	26	12	37	0	5
<i>Allium ebusitanum</i>	<i>Asparagales</i>	521	0	2	51	485
<i>Anacamptis morio</i>	<i>Asparagales</i>	54	10	0	48	16
<i>Carduus nutans</i>	<i>Asterales</i>	55	8	41	23	8
<i>Colchicum cupanii</i>	<i>Liliales</i>	14	0	0	34	0
<i>Colchicum ritchii</i>	<i>Liliales</i>	0	12	0	3	17



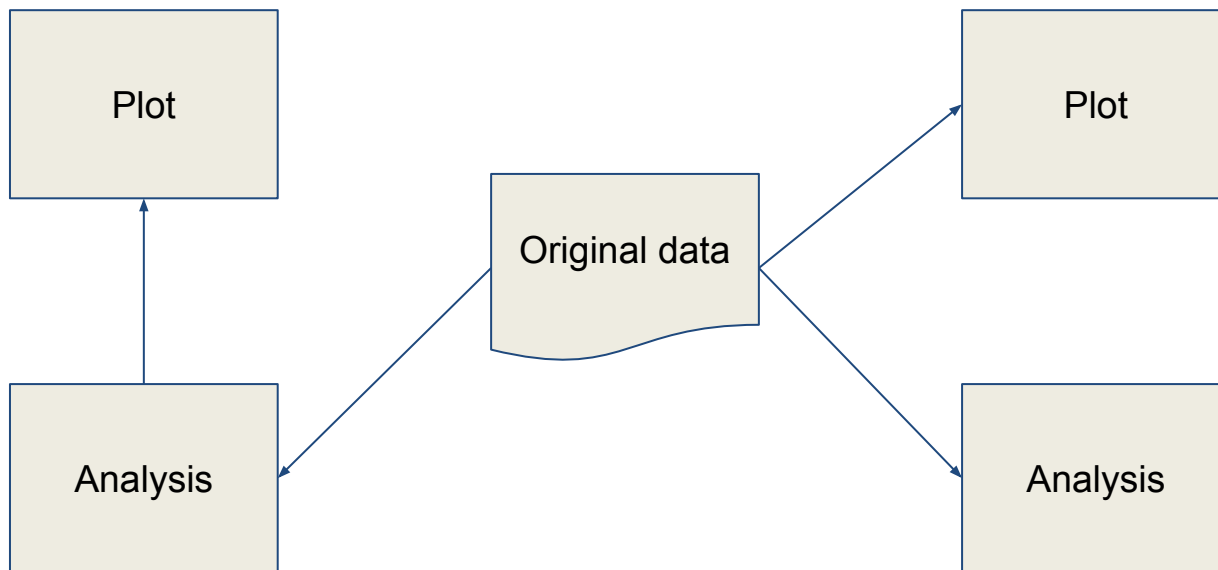
The “data view” of data, a.k.a. “long” format

species	order	site	count
Berberis hispanica	Ranunculales	site 1	26
Allium ebusitanum	Asparagales	site 1	521
Anacamptis morio	Asparagales	site 1	54
Carduus nutans	Asterales	site 1	55
Colchicum cupanii	Liliales	site 1	14
Colchicum ritchii	Liliales	site 1	0
Berberis hispanica	Ranunculales	site 2	12
Allium ebusitanum	Asparagales	site 2	0
Anacamptis morio	Asparagales	site 2	10
Carduus nutans	Asterales	site 2	8
Colchicum cupanii	Liliales	site 2	0
Colchicum ritchii	Liliales	site 2	12
Berberis hispanica	Ranunculales	site 3	37

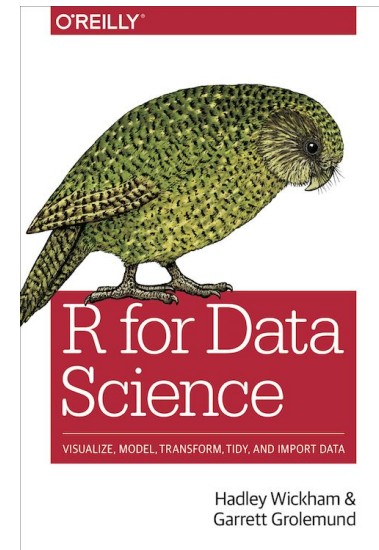
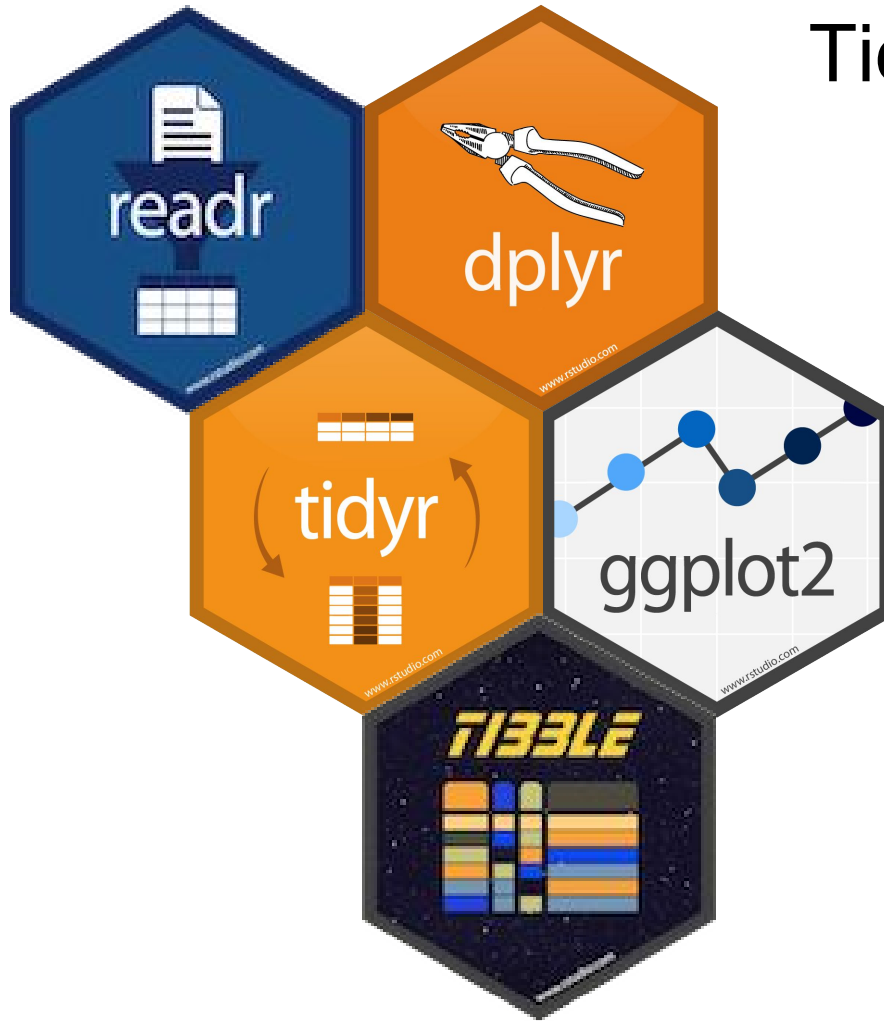




You need to be good at format conversion...



Tidyverse



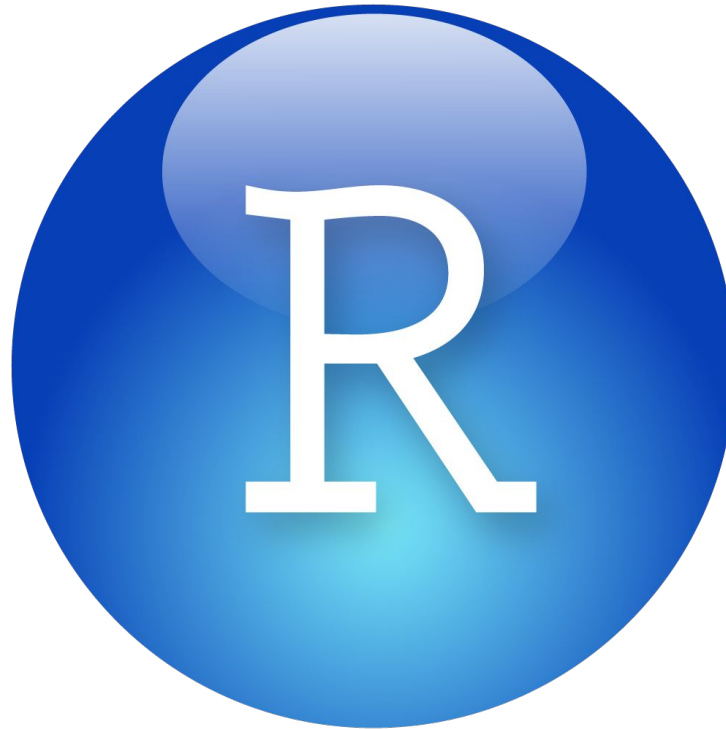
<http://r4ds.had.co.nz/>

```
> install.packages('tidyverse')  
> library(tidyverse)
```

<https://www.tidyverse.org/>



RStudio



<https://www.rstudio.com/>

RStudio and Shiny are trademarks of RStudio, Inc.



RStudio

File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

analysis.R * prep.R *

Source on Save Run Source

Run the current line or selection (Ctrl+Enter)

```
1 # User Analysis
2
3 setwd("~/analysis")
4 source("prep.R")
5
6 library(plyr)
7 library(lattice)
8 library(ggplot2)
9
10 # Import data set
11 rawdata <- read.csv("stats.csv")
12 dim(rawdata)
13
14 # Clean data set
15 clean <- prepareStats(rawdata)
16
17 # Subset of active users
18 active <- subset(clean, active == 1)
19 count(active, "daysSinceAccountCreated < 30")[2,2]
20 mean(active$age)
21
22
23
24
```

16:1 (Top Level) R Script

Workspace History

Load Save Import Dataset Clear All

Data

clean	360404 obs. of 35 variables
rawdata	530750 obs. of 35 variables

Functions

- evalPercentage(expression, index)
- formatInteger(object, ...)
- prepareStats(data, sampleSize = NA)

Files Plots Packages Help

Install Packages Check for Updates

<input type="checkbox"/>	compiler	The R Compiler Package
<input checked="" type="checkbox"/>	datasets	The R Datasets Package
<input type="checkbox"/>	dichromat	Color schemes for dichromats
<input type="checkbox"/>	digest	Create cryptographic hash digests of R objects
<input type="checkbox"/>	evaluate	Parsing and evaluation tools that provide more details than the default.
<input type="checkbox"/>	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...
<input type="checkbox"/>	formatR	Format R Code Automatically
<input checked="" type="checkbox"/>	ggplot2	An implementation of the Grammar of Graphics
<input checked="" type="checkbox"/>	graphics	The R Graphics Package
<input checked="" type="checkbox"/>	grDevices	The R Graphics Devices and Support for Colours and Fonts

Console ~/analysis/

```
> library(plyr)
> library(lattice)
> library(ggplot2)
> # Import data set
> rawdata <- read.csv("stats.csv")
> dim(rawdata)
[1] 530750    35
>
> # Clean data set
> clean <- prepareStats(rawdata)
>
>
```


What I expect from you

1. No knowledge of R
2. Knowledge of “base” R
3. Knowledge of tidyverse R
4. You prefer Python, Perl, Ruby, ...

“Daniel, I’m not following!”



What to show (notes to myself)

- readr: `read_tsv()/write_tsv()`
- tidyr: `gather()/spread()`
- tidyr: `separate()/unite()`
- dplyr: `filter()/select()`
- dplyr: `group_by()` and `summarise()`
- dplyr: `union()/inner_join()` *et al.*
- ggplot2: `ggplot(): geom_col(), geom_point(), facet_wrap()`

