

# Visualisation for EDA

*April 2017*

Hadley Wickham  
[@hadleywickham](#)  
Chief Scientist, RStudio

# Import



# Tidy

Consistent way of  
storing data



# Transform

Create new variables & new summaries



# Visualise

Surprises, but doesn't scale



# Model

Scales, but doesn't (fundamentally) surprise



*Program*

# Communicate

# Import



## Explore

# Tidy

Consistent way of  
storing data



# Transform

Create new variables & new summaries

# Visualise

Surprises, but doesn't scale



# Model

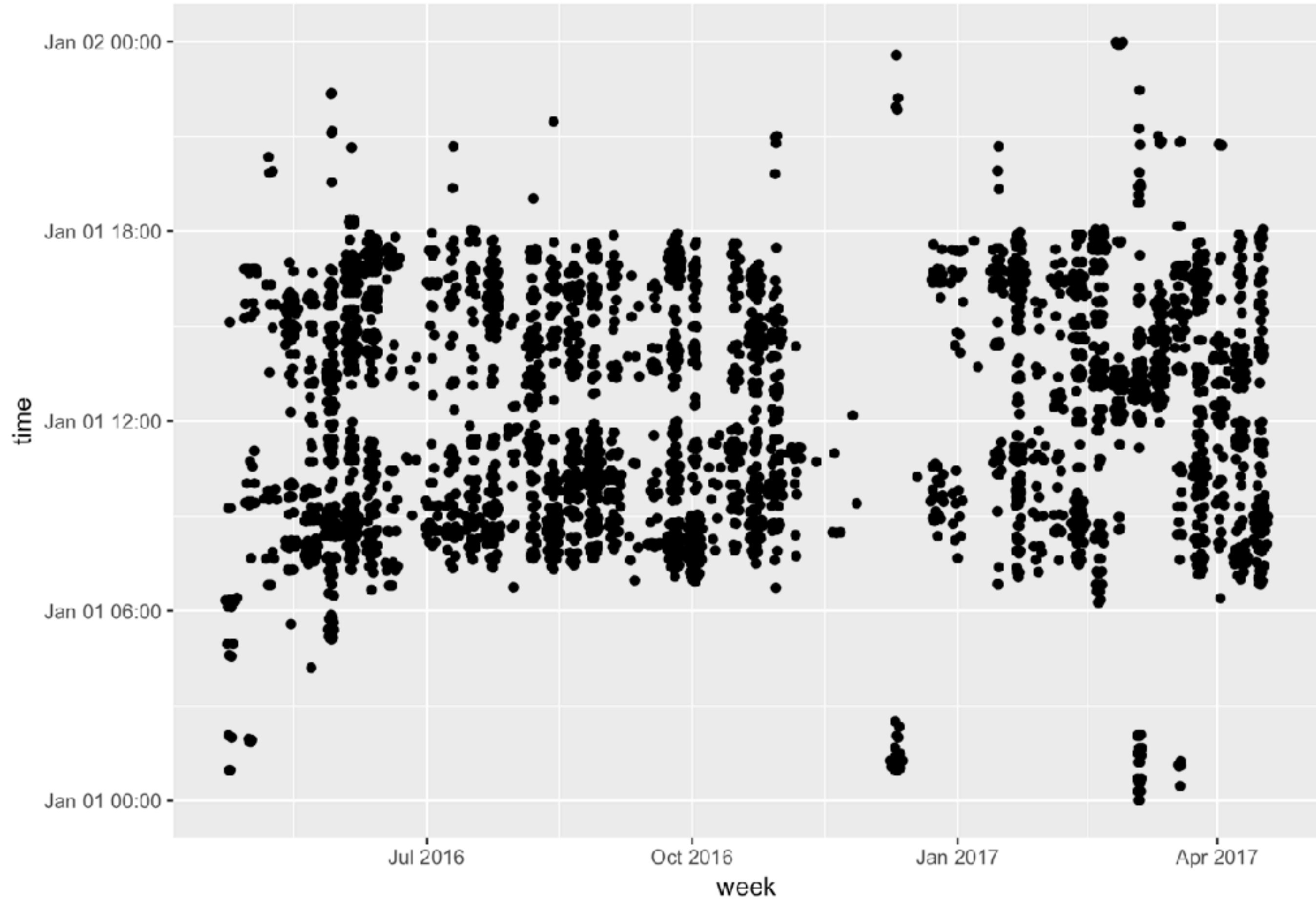
Scales, but doesn't (fundamentally) surprise



# *Program*

# Communicate

# First: my GitHub commits



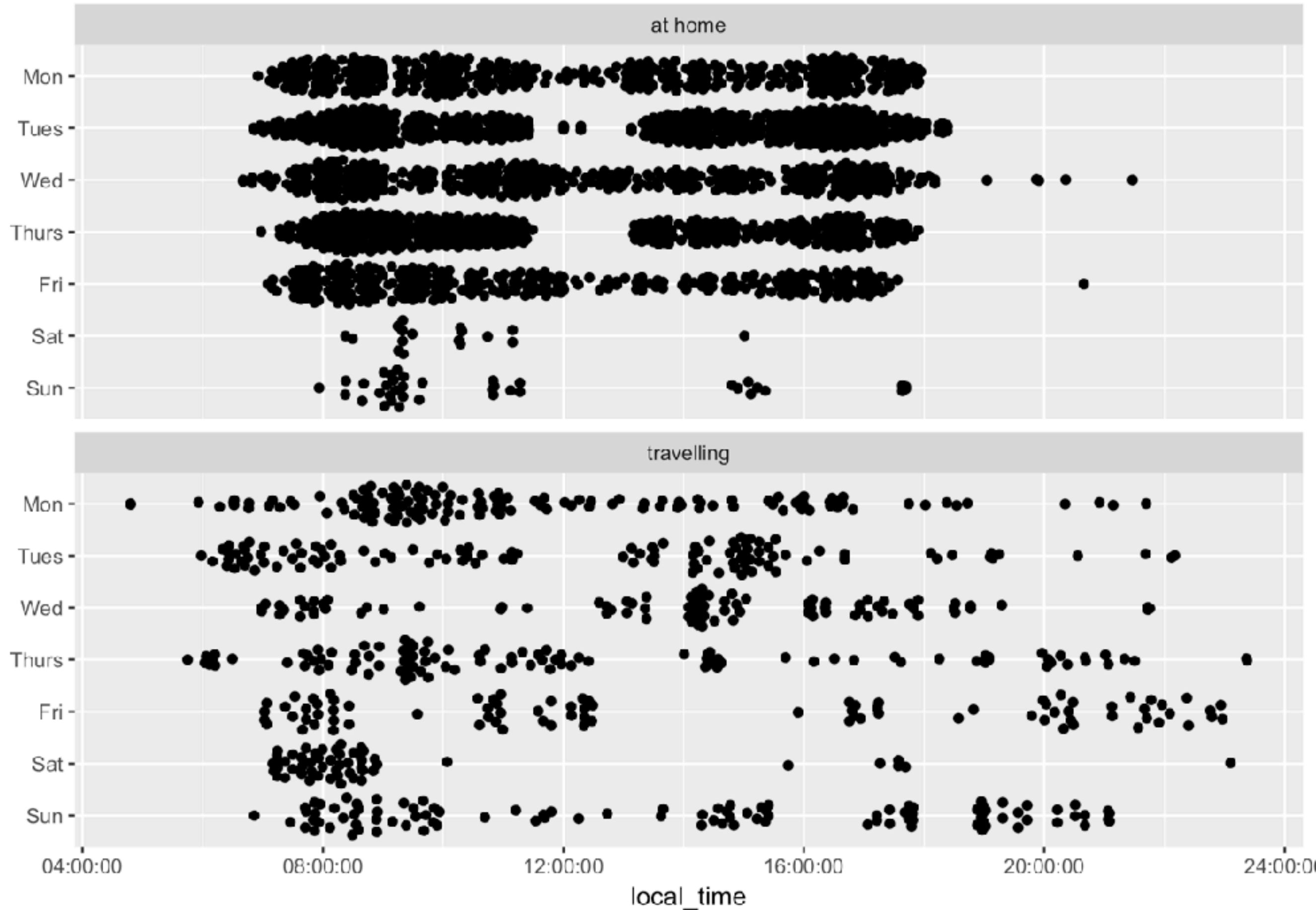
# Second: my travel

## Places I've visited in the US



*As captured by 'primary' address in Triplt*

Which will culminate in this figure

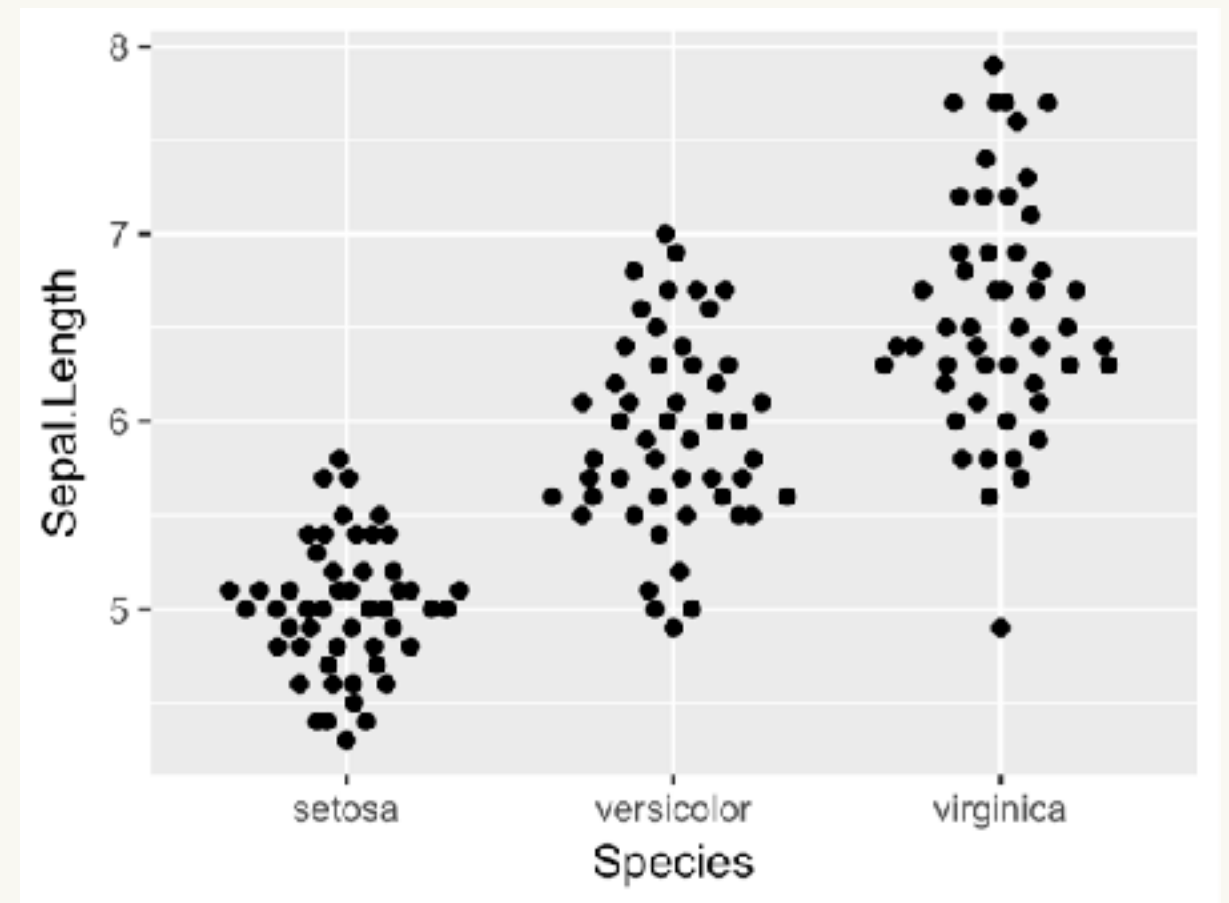
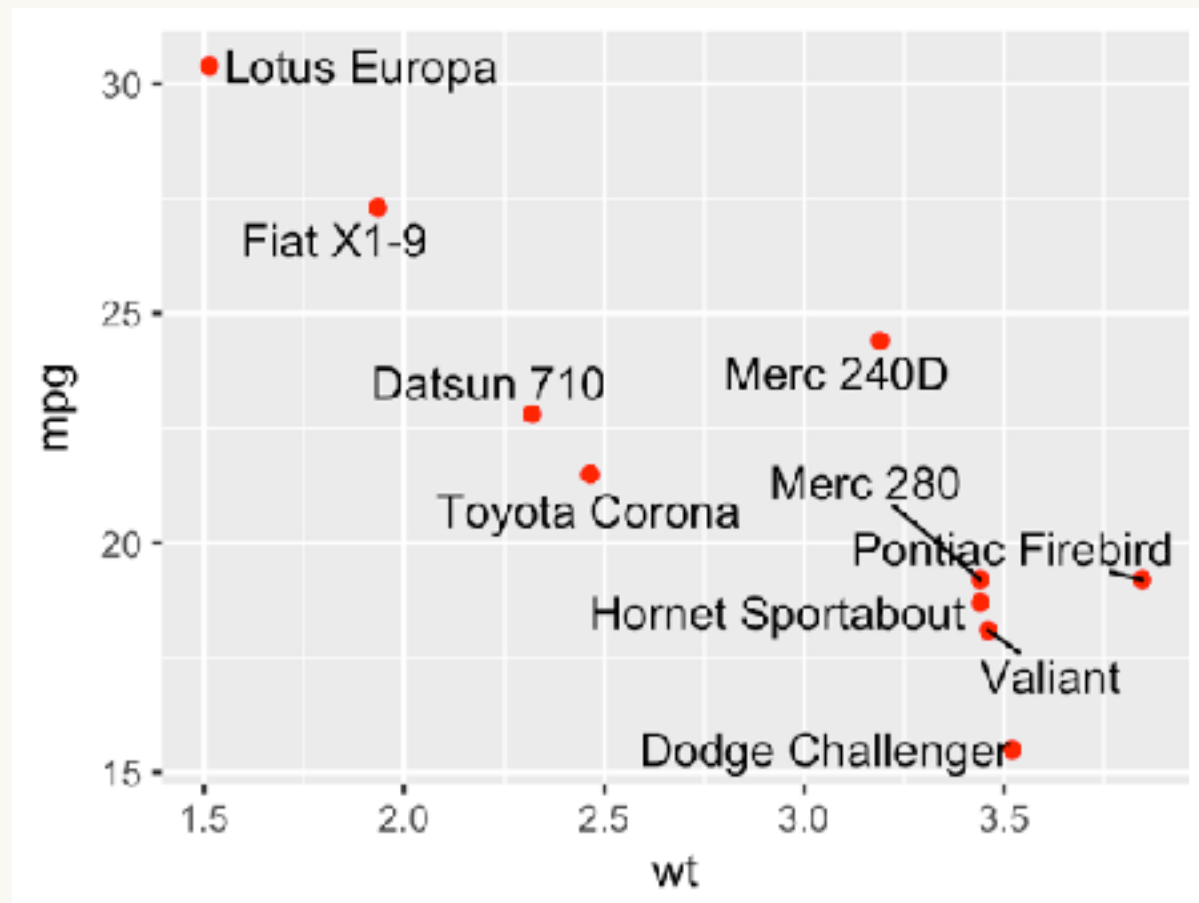


Live demo

# ggrepel

by Kamil Slowikowski

<https://github.com/slowkow/ggrepel>



# ggbeeswarm

by Erik Clark & Scott Sherrill-Max

<https://github.com/eclarke/ggbeeswarm>



# Import

readr  
readxl  
haven  
xml2  
DBI

# Visualise

ggplot2

# Model

broom  
modelr

# Tidy → Transform

tibble  
tidyr

dplyr  
forcats  
hms

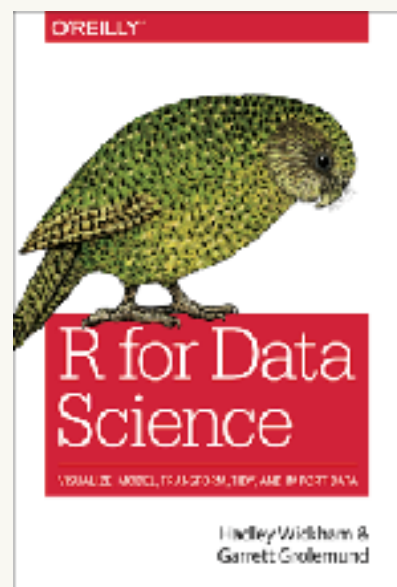
stringr  
lubridate

# Communicate

shiny  
rmarkdown

purrr  
magrittr

# Program



[r4ds.had.co.nz](http://r4ds.had.co.nz)

O'Reilly discount: AUTHD



[tidyverse.org](http://tidyverse.org)

# Console visualisations

# Previously

```
> library(tidyverse)
```

```
Loading ggplot2
```

```
Loading tibble
```

```
Loading tidyr
```

```
Loading readr
```

```
Loading purrr
```

```
Loading dplyr
```

```
Loading stringr
```

```
Loading forcats
```

# Why not make use of the horizontal space?

```
> library(tidyverse)
+ ggplot2 2.2.1.9000      Date: 2017-04-23
+ tibble  1.3.0           R: 3.3.2
+ tidyr   0.6.1           OS: macOS Sierra 10.12.4
+ readr   1.1.0           GUI: RStudio 1.0.141
+ purrr   0.2.2.9000      Locale: en_US.UTF-8
+ dplyr    0.5.0.9004      TZ: America/Chicago
+ stringr 1.2.0
+ forcats 0.2.0
```

Conflicts

```
-----
* filter(),    from dplyr, masks stats::filter()
* is_null(),   from purrr, masks testthat::is_null()
* lag(),       from dplyr, masks stats::lag()
* matches(),   from dplyr, masks testthat::matches()
```

# library(precis)

```
# data.frame [32 x 11]
```












	name	type		precis
	<chr>	<chr>		<chr>
1	mpg	dbl	10.4 [ 15.4 ( 19.2) 22.8]	33.9
2	cyl	dbl	3 levels: 4 (11) 6 (7) 8 (14)	
3	disp	dbl	71.1 [121.0 (196.0) 334.0]	472.0
4	hp	dbl	52 [ 96 ( 123) 180]	335
5	drat	dbl	2.76 [ 3.08 ( 3.70) 3.92]	4.93
6	wt	dbl	1.51 [ 2.54 ( 3.32) 3.65]	5.42
7	qsec	dbl	14.5 [ 16.9 ( 17.7) 18.9]	22.9
8	vs	dbl	2 levels: 0 (18) 1 (14)	
9	am	dbl	2 levels: 0 (19) 1 (13)	
10	gear	dbl	3 levels: 3 (15) 4 (12) 5 (5)	
11	carb	dbl	1 [ 2 ( 2) 4]	8

rescale = TRUE, reorder = TRUE?

```
# data.frame [32 x 11]
```

	name	type						precis	
	<chr>	<chr>						<chr>	
1	mpg	dbl	1.40	[1.54	(1.92)	2.28]	3.39	* 10	
2	disp	dbl	0.71	[1.21	(1.96)	3.34]	4.72	* 100	
3	hp	dbl	5.20	[0.96	(1.23)	1.80]	3.35	* 10	
4	drat	dbl	2.76	[3.08	(3.70)	3.92]	4.93	* 1	
5	wt	dbl	1.51	[2.54	(3.32)	3.65]	5.42	* 1	
6	qsec	dbl	1.45	[1.69	(1.77)	1.89]	2.29	* 10	
7	carb	dbl	1	[2	(2	) 4	] 8	* 1	
8	cyl	dbl	3 levels:	4	(11)	6	(7)	8	(14)
9	vs	dbl	2 levels:	0	(18)	1	(14)		
10	am	dbl	2 levels:	0	(19)	1	(13)		
11	gear	dbl	3 levels:	3	(15)	4	(12)	5	(5)

histogram = TRUE

	name	type			precis
	<chr>	<chr>			<chr>
1	mpg	dbl	10.4		33.9
2	cyl	dbl	4		8
3	disp	dbl	71.1		472.0
4	hp	dbl	52		335
5	drat	dbl	2.76		4.93
6	wt	dbl	1.51		5.42
7	qsec	dbl	14.5		22.9
8	vs	dbl	0		1
9	am	dbl	0		1
10	gear	dbl	3		5
11	carb	dbl	1		8

This work is licensed under the  
Creative Commons Attribution-Noncommercial 3.0  
United States License.

To view a copy of this license, visit  
<http://creativecommons.org/licenses/by-nc/3.0/us/>