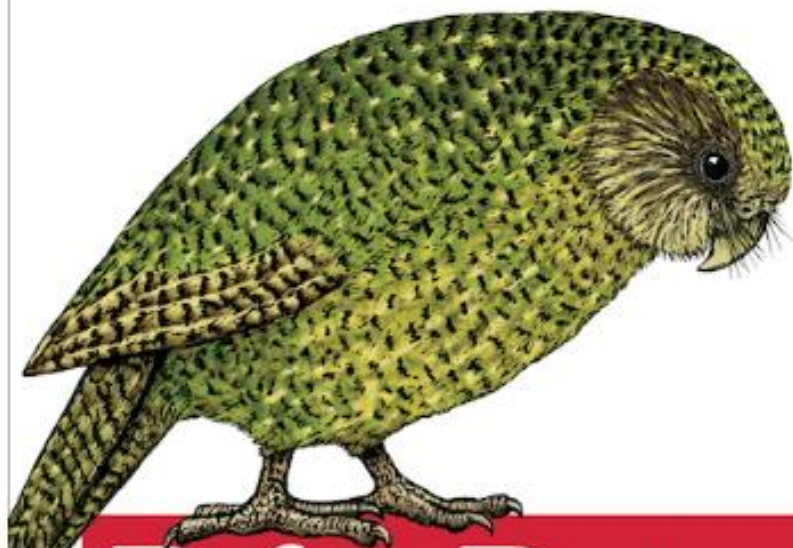


O'REILLY®



R for Data Science

VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

Hadley Wickham &
Garrett Grolemund

Data Wrangling (Part 1)

Quantfish woRkshop 3
November 27, 2018

Getting your data the way you want it

AutoSave Off

OTN_2016.xlsx - Excel

Sign in

File

Home

Insert

Draw

Page Layout

Formulas

Data

Review

View


Add-ins


Help


Foxit PDF


Tell me what you want to do

Share

 Paste

 Cut

 Copy


 Format Painter


Clipboard


Calibri


11

A^A




 B


 I

 U

 A


Font


  


 Wrap Text


Alignment

General


 \$


 %

 ,

 .00

Number

 Conditional Formatting

 Format as Table

Styles

Normal


Bad


Good


Neutral

Calculation


Check Cell


 Insert


 Delete


 Format


Cells

 AutoSum

 Fill


 Clear


 Sort & Filter


 Find & Select

Editing

R8

 X

 ✓

 fx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V			
1	collection	catalog	nur	scientific	common	date	last modified	detected	db	receiver_g	station	receiver	bottom_d	receiver_d	tagname	codespace	sensornam	sensorrow	sensortype	sensorvalu	sensorunit	date collected	timezone	longitude	latitude
2	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS179	250555			A69-9001-	A69-9001			pinger			10/4/2016 4:51	UTC	-59.9838	47.294			
3	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS179	250555			A69-9001-	A69-9001			pinger			10/4/2016 4:52	UTC	-59.9838	47.294			
4	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS178	250535			A69-9001-	A69-9001			pinger			10/4/2016 4:58	UTC	-59.9924	47.290			
5	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS178	250535			A69-9001-	A69-9001			pinger			10/4/2016 4:56	UTC	-59.9924	47.290			
6	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS177	250494			A69-9001-	A69-9001			pinger			10/4/2016 5:05	UTC	-60.0009	47.286			
7	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS178	250535			A69-9001-	A69-9001			pinger			10/4/2016 5:01	UTC	-59.9924	47.290			
8	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:45	UTC	-60.0181	47.278			
9	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:54	UTC	-60.0181	47.278			
10	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:58	UTC	-60.0181	47.278			
11	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:53	UTC	-60.0181	47.278			
12	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:52	UTC	-60.0181	47.278			
13	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:49	UTC	-60.0181	47.278			
14	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:50	UTC	-60.0181	47.278			
15	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:55	UTC	-60.0181	47.278			
16	MMFSRP	MMFSRP-1	Carcharod	white shar	8/11/2017	CBS	CBS	CBS175	250508			A69-9001-	A69-9001			pinger			9/15/2016 0:56	UTC	-60.0181	47.278			
17	MMFSRP	MMFSRP-1	Carcharod	white shar	11/21/2016	CBS	CBS	CBS016	250026	176	171	A69-9001-	A69-9001			pinger			10/20/2016 10:17	UTC	-60.2759	47.106			
18	MMFSRP	MMFSRP-1	Carcharod	white shar	11/21/2016	CBS	CBS	CBS016	250026	176	171	A69-9001-	A69-9001			pinger			10/20/2016 10:15	UTC	-60.2759	47.106			
19	MMFSRP	MMFSRP-1	Carcharod	white shar	6/5/2017	CBS	CBS	CBS263	128522			A69-9001-	A69-9001			pinger			9/15/2016 2:50	UTC	-60.1341	47.290			
20	MMFSRP	MMFSRP-1	Carcharod	white shar	6/5/2017	CBS	CBS	CBS267	128550			A69-9001-	A69-9001			pinger			10/4/2016 7:31	UTC	-60.1612	47.312			
21	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX302	250182	136	128	A69-9001-	A69-9001			pinger			10/25/2016 21:07	UTC	-63.3203	43.819			
22	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX302	250182	136	128	A69-9001-	A69-9001			pinger			10/25/2016 21:08	UTC	-63.3203	43.819			
23	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX302	250182	136	128	A69-9001-	A69-9001			pinger			10/25/2016 21:12	UTC	-63.3203	43.819			
24	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX303	250126	126	118	A69-9001-	A69-9001			pinger			10/25/2016 21:12	UTC	-63.325	43.812			
25	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX303	250126	126	118	A69-9001-	A69-9001			pinger			10/25/2016 21:14	UTC	-63.325	43.812			
26	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX303	250126	126	118	A69-9001-	A69-9001			pinger			10/25/2016 21:15	UTC	-63.325	43.812			
27	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX303	250126	126	118	A69-9001-	A69-9001			pinger			10/25/2016 21:16	UTC	-63.325	43.812			
28	MMFSRP	MMFSRP-2	Carcharod	white shar	5/2/2017	HFX	HFX	HFX304	250173	123	115	A69-9001-	A69-9001			pinger			10/25/2016 21:22	UTC	-63.3295	43.806			
29	MMFSRP	MMFSRP-1	Carcharod	white shar	12/13/2016	HFX	HFX	HFX204	250313	164	159	A69-9001-	A69-9001			pinger			10/26/2016 14:14	UTC	-63.5	43.212			

Things we'll learn today

`filter()` Pick observations by their values

`distinct()` Remove duplicate rows

`arrange()` Reorder rows

`select()` Pick variables by their names

`mutate()` Create new variables with existing variables

`%>%` Combining multiple operations with the pipe

To R!

Filter rows with `filter()`

```
filter(flights, month == 1, day == 1)

#> # A tibble: 842 x 19

#>   year month   day dep_time sched_dep_time dep_delay arr_time
#>   <int> <int> <int>   <int>         <int>       <dbl>   <int>
#> 1  2013     1     1     517           515         2     830
#> 2  2013     1     1     533           529         4     850
#> 3  2013     1     1     542           540         2     923
#> 4  2013     1     1     544           545        -1    1004
#> 5  2013     1     1     554           600        -6     812
#> 6  2013     1     1     554           558        -4     740

#> # ... with 836 more rows, and 12 more variables: sched_arr_time <int>,
#> #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
#> #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#> #   minute <dbl>, time_hour <dtm>
```

Comparisons

> Select values greater than

< Select values less than

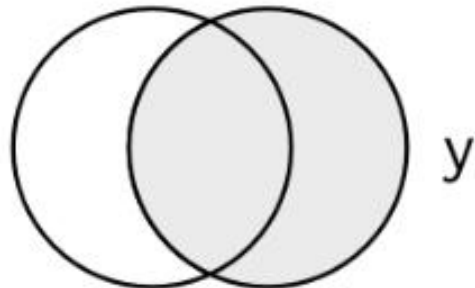
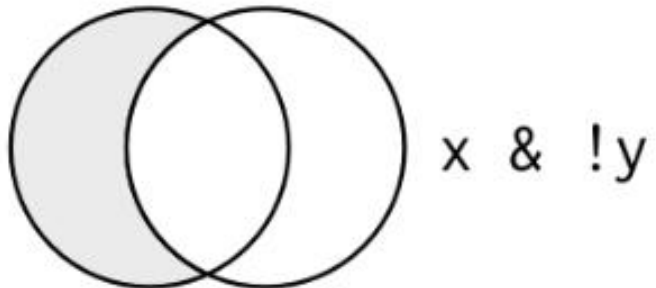
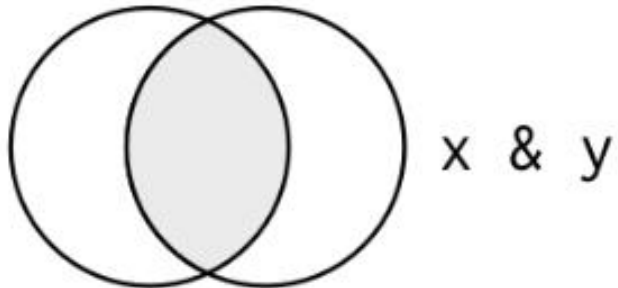
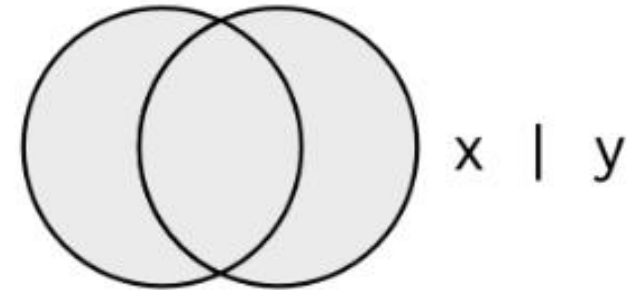
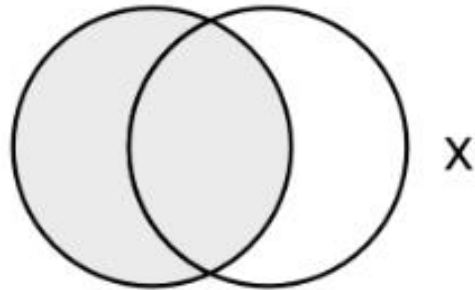
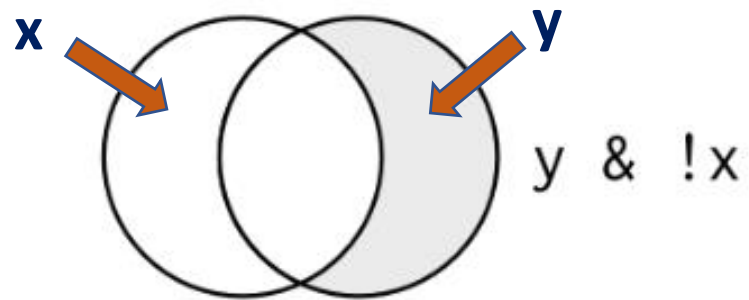
>= Selects values greater than or equal to

<= Selects values less than or equal to

!= Selects values not equal to

== Selects values equal to

Logical operators



Identifying and excluding missing values

Identify whether values are missing or not:

`is.na(x)` Determine if a value is missing

`!is.na(x)` Determine if a value is not missing

Use `filter()` to select missing or not-missing values:

`filter(data, is.na(x))`

`filter(data, !is.na(x))`

Identifying and excluding duplicate rows

Identify and remove rows with duplicate values:

```
distinct(data)
```

Remove duplicate rows based on certain variables:

```
distinct(data, variable)
```

```
distinct(data, variable, .keep_all=TRUE)
```

Sort rows with `arrange()`

```
arrange(flights, year, month, day)

#> # A tibble: 336,776 x 19
#>   year month   day dep_time sched_dep_time dep_delay arr_time
#>   <int> <int> <int>   <int>         <int>         <dbl>   <int>
#> 1  2013     1     1     517           515           2     830
#> 2  2013     1     1     533           529           4     850
#> 3  2013     1     1     542           540           2     923
#> 4  2013     1     1     544           545          -1    1004
#> 5  2013     1     1     554           600          -6     812
#> 6  2013     1     1     554           558          -4     740
#> # ... with 3.368e+05 more rows, and 12 more variables:
#> #   sched_arr_time <int>, arr_delay <dbl>, carrier <chr>, flight <int>,
#> #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#> #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dtm>
```

*** Missing values will always be sorted at the end!**

Select columns with `select()`

```
# Select columns by name
select(flights, year, month, day)

#> # A tibble: 336,776 x 3
#>   year month   day
#>   <int> <int> <int>
#> 1  2013     1     1
#> 2  2013     1     1
#> 3  2013     1     1
#> 4  2013     1     1
#> 5  2013     1     1
#> 6  2013     1     1
#> # ... with 3.368e+05 more rows
```

Select columns with `select()`

There are a number of helper functions you can use within `select()`:

- `starts_with("abc")` : matches names that begin with “abc”.
- `ends_with("xyz")` : matches names that end with “xyz”.
- `contains("ijk")` : matches names that contain “ijk”.
- `matches("(.)\\1")` : selects variables that match a regular expression. This one matches any variables that contain repeated characters. You’ll learn more about regular expressions in [strings](#).
- `num_range("x", 1:3)` : matches `x1` , `x2` and `x3` .

See `?select` for more details.

Add new variables with mutate()

```
mutate(flights_sml,  
  gain = dep_delay - arr_delay,  
  speed = distance / air_time * 60  
)  
  
#> # A tibble: 336,776 x 9  
  
#>   year month   day dep_delay arr_delay distance air_time   gain speed  
#>   <int> <int> <int>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>  
#> 1  2013     1     1         2        11    1400     227    -9   370.  
#> 2  2013     1     1         4        20    1416     227   -16   374.  
#> 3  2013     1     1         2        33    1089     160   -31   408.  
#> 4  2013     1     1        -1       -18    1576     183    17   517.  
#> 5  2013     1     1        -6       -25     762     116    19   394.  
#> 6  2013     1     1        -4        12     719     150   -16   288.  
  
#> # ... with 3.368e+05 more rows
```

Useful creation functions

Logical operators

`>, <, >=, <=, !=, ==`

Arithmetic operators

`+, -, *, /, ^, log, sin, cos`

Modular arithmetic

`%/%, %%`

Offsets

`lead(), lag()`

Cumulative/rolling

`cumsum(), cumprod()`

Ranking

`min_rank()`

Combining multiple operations with the pipe

Instead of this (clunky):

```
flights_sml <- select(flights,
                      year:day,
                      ends_with("delay"),
                      distance,
                      air_time)

flights_sub <- mutate(flights_sml,
                     gain = dep_delay - arr_delay,
                     hours = air_time / 60,
                     gain_per_hour = gain / hours)
```

We can achieve the same thing using %>%:

```
flights_sub <- flights %>%
  select(year:day,
         ends_with("delay"),
         distance,
         air_time) %>%
  mutate(gain = dep_delay - arr_delay,
         hours = air_time / 60,
         gain_per_hour = gain / hours)
```