

dplyr

a grammar of
data manipulation

Group	dose 1	dose 2
A	3	3
A	4	5
B	3	1
B	1	3
C	1	3
C	2	2

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

n	min	mean	max
6	4	5.2	9

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9

Group	Total
A	15

B	3	1	4
B	1	3	4

B	8
---	---

C	1	3	4
C	2	2	4

C	8
---	---

n	min	mean	max
6	4	5.2	9

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

Group	Sum
A	6
A	9
B	4
B	4
C	4
C	4

Group	Sum
A	6
C	4
C	4

tbl
%>%



database



American Airlines



American Eagle



FRONTIER
AIRLINES

jetBlue
AIRWAYS®

SkyWest
AIRLINES*

 **DELTA**





RStudio

Go to File/Function

Console - /

```
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
```

```
Natural language support but running in an English locale
```

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

```
[Workspace loaded from ~/.RData]
```

```
> library(hflights)
> hflights
```

Environment History

Import Dataset Clear

Global Environment

Name Type Len

Envir

Files Plots Packages Help Viewer

Zoom Export Clear All



RStudio

File Edit View Insert Cell Help

Go to File/Function

74968	TPA	787	6	17	0	0
74971	MSP	1034	11	25	0	0
74975	MSY	305	3	26	0	0
74978	BWI	1235	4	11	0	0
74981	LGA	1416	5	14	0	0
74985	CLE	1091	5	16	0	0
74990	DCA	1208	9	16	0	0
74993	EGE	935	4	16	0	0
74994	SEA	1874	6	11	0	0
75001	BOS	1597	9	13	0	0
75004	MCO	853	8	21	0	0
75010	LAS	1222	7	32	0	0
75014	TPA	787	3	20	0	0
75015	PDX	1825	7	11	0	0
75018	MSY	305	2	17	0	0
75025	LGA	1416	5	27	0	0
75029	SAN	1303	3	13	0	0
75030	SAN	1303	4	35	0	0
75031	AUS	148	5	20	0	0
75032	CLE	1091	5	28	0	0
75033	ORD	925	11	24	0	0
75034	SLC	1195	5	18	0	0
75036	ATL	689	9	16	0	0
75039	MSY	305	3	10	0	0
75040	SEA	1874	6	30	0	0
75043	SFO	1635	5	6	0	0
75044	SFO	1635	7	29	0	0
75045	PHL	1324	9	19	0	0
75050	MCO	853	6	16	0	0
75052	DTW	1076	8	19	0	0
75054	LAX	1379	13	15	0	0
75057	DEN	862	17	15	0	0
75062	EWR	1400	14	18	0	0
75067	PHX	1009	6	22	0	0
75073	TPA	787	3	13	0	0

[reached getOption("max.print") -- omitted 227020 rows]

>

Environment History

Import Dataset Clear Global Environment

Name	Type	Len
Envir		

Files Plots Packages Help Viewer

Zoom Export Clear All

tbl

Console ~/

```
> hflights <- tbl_df(hflights)
> hflights
Source: local data frame [227,496 x 21]
```

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier
5424	2011	1	1	6	1400	1500	AA
5425	2011	1	2	7	1401	1501	AA
5426	2011	1	3	1	1352	1502	AA
5427	2011	1	4	2	1403	1513	AA
5428	2011	1	5	3	1405	1507	AA
5429	2011	1	6	4	1359	1503	AA
5430	2011	1	7	5	1359	1509	AA
5431	2011	1	8	6	1355	1454	AA
5432	2011	1	9	7	1443	1554	AA
5433	2011	1	10	1	1443	1553	AA
...
Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)							

> |

Environment History

Import Dataset Clear

Global Environment

Name	Type	Len
hflights	tbl_df	21

Files Plots Packages Help Viewer

Zoom Export Clear

Console ~/ Go to file/function

Project

Console ~/
#useairline
(int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)
> hflights
Source: local data frame [227,496 x 21]

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime
5424	2011	1	1	6	1400	1500
5425	2011	1	2	7	1401	1501
5426	2011	1	3	1	1352	1502
5427	2011	1	4	2	1403	1513
5428	2011	1	5	3	1405	1507
5429	2011	1	6	4	1359	1503
5430	2011	1	7	5	1359	1509
5431	2011	1	8	6	1355	1454
5432	2011	1	9	7	1443	1554
5433	2011	1	10	1	1443	1553
..

Variables not shown: UniqueCarrier (chr), FlightNum (int), TailNum (chr), ActualElapsedTime (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)

> |

Environment History Import Dataset Global Environment

Name	Type	Length	Size	Value
hflights	tbl_df	21	23.6 .. 227496 obs. of	

Files Plots Packages Help Viewer

> glimpse(hflights)

Variables:

```
$ Year          (int) 2011, 2011, 2011, 2011, 2011, 2011, 2011, ...
$ Month        (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ DayofMonth   (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ...
$ DayOfWeek    (int) 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, ...
$ DepTime      (int) 1400, 1401, 1352, 1403, 1405, 1359, 1359, ...
$ ArrTime      (int) 1500, 1501, 1502, 1513, 1507, 1503, 1509, ...
$ UniqueCarrier (chr) "AA", "AA", "AA", "AA", "AA", "AA", "AA", ...
$ FlightNum    (int) 428, 428, 428, 428, 428, 428, 428, 428, 42...
$ TailNum      (chr) "N576AA", "N557AA", "N541AA", "N403AA", "N...
$ ActualElapsedTime (int) 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70...
$ AirTime       (int) 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42...
$ ArrDelay      (int) -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 2...
$ DepDelay      (int) 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19...
$ Origin        (chr) "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
$ Dest          (chr) "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", ...
$ Distance      (int) 224, 224, 224, 224, 224, 224, 224, 224, 22...
$ TaxiIn        (int) 7, 6, 5, 9, 9, 6, 12, 7, 8, 6, 8, 4, 6, 5, ...
$ TaxiOut       (int) 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, ...
$ Cancelled     (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ CancellationCode (chr) "", "", "", "", "", "", "", "", "", "", ...
$ Diverted      (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

> |

> glimpse(hflights)

Variables:

```
$ Year          (int) 2011, 2011, 2011, 2011, 2011, 2011, 2011, ...
$ Month        (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ DayofMonth   (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ...
$ DayOfWeek    (int) 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, ...
$ DepTime      (int) 1400, 1401, 1352, 1403, 1405, 1359, 1359, ...
$ ArrTime      (int) 1500, 1501, 1502, 1513, 1507, 1503, 1509, ...
$ UniqueCarrier (chr) "AA", "AA", "AA", "AA", "AA", "AA", "AA", ...
$ FlightNum    (int) 428, 428, 428, 428, 428, 428, 428, 428, 42...
$ TailNum      (chr) "N576AA", "N557AA", "N541AA", "N403AA", "N...
$ ActualElapsedTime (int) 60, 60, 70, 70, 62, 64, 70, 59, 71, 70, 70...
$ AirTime       (int) 40, 45, 48, 39, 44, 45, 43, 40, 41, 45, 42...
$ ArrDelay      (int) -10, -9, -8, 3, -3, -7, -1, -16, 44, 43, 2...
$ DepDelay      (int) 0, 1, -8, 3, 5, -1, -1, -5, 43, 43, 29, 19...
$ Origin        (chr) "IAH", "IAH", "IAH", "IAH", "IAH", "IAH", ...
$ Dest          (chr) "DFW", "DFW", "DFW", "DFW", "DFW", "DFW", ...
$ Distance      (int) 224, 224, 224, 224, 224, 224, 224, 224, 22...
$ TaxiIn        (int) 7, 6, 5, 9, 9, 6, 12, 7, 8, 6, 8, 4, 6, 5, ...
$ TaxiOut       (int) 13, 9, 17, 22, 9, 13, 15, 12, 22, 19, 20, ...
$ Cancelled     (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ CancellationCode (chr) "", "", "", "", "", "", "", "", "", "", ...
$ Diverted      (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

> as.data.frame(hflights)|

select
filter
arrange
mutate
summarize

Group	count	dose	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

select

Group	Sum
A	6
A	9
B	4
B	4
C	4
C	4

select

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

filter

Group	dose 1	dose 2	Sum
A	3	3	6
C	1	3	4
C	2	2	4

filter

Group	dose 1	dose 2	Sum
C	1	3	4
A	3	3	6
A	4	5	9
B	3	1	4
B	2	3	4
C	5	2	4

arrange

Group	dose 1	dose 2	Sum
C	1	3	4
B	2	3	4
B	3	1	4
A	3	3	6
A	4	5	9
C	5	2	4

arrange

Group	dose 1	dose 2
A	3	3
A	4	5
B	3	1
B	1	3
C	1	3
C	2	2

mutate

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

mutate

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

summarise

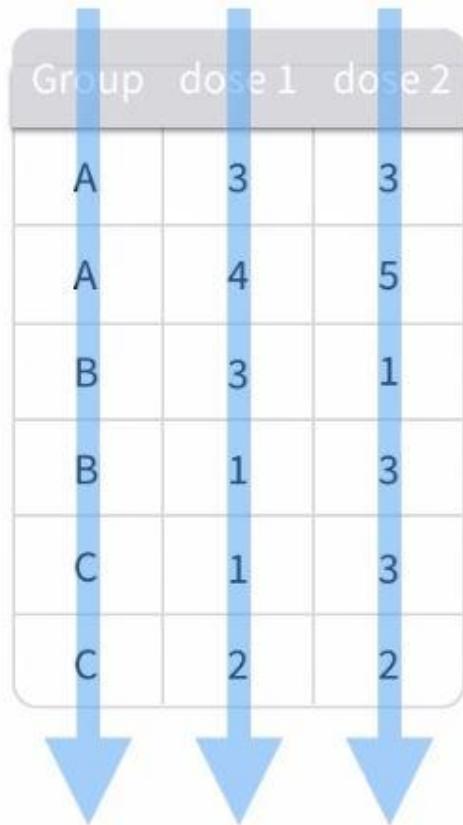
Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

n	min	mean	max
6	4	5.2	9

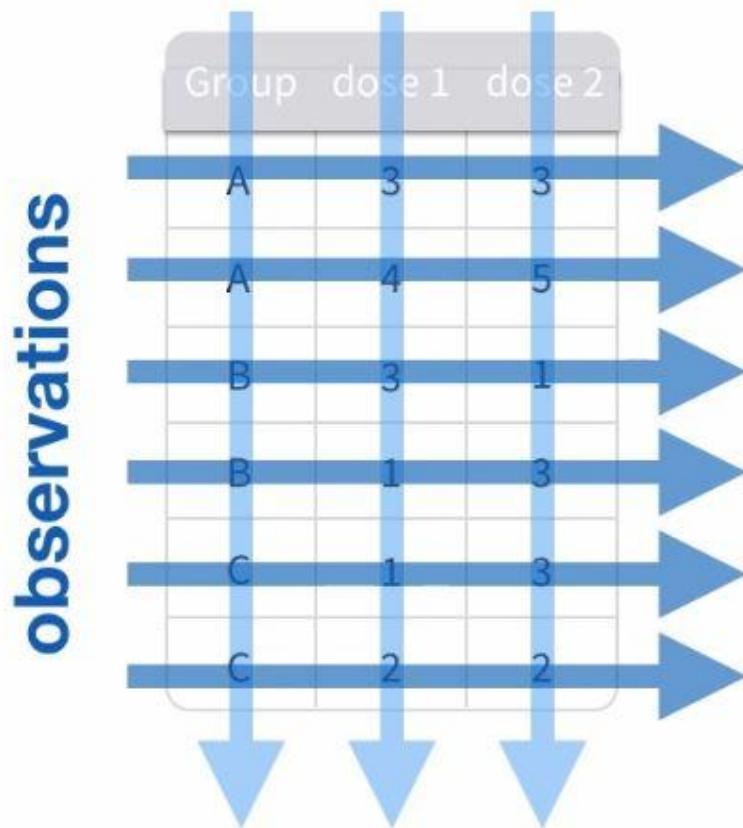
summarise

variables

Group	dose 1	dose 2
A	3	3
A	4	5
B	3	1
B	1	3
C	1	3
C	2	2



variables



select

mutate

filter

arrange

summarize

select

mutate

variables

filter

arrange

summarize

select

mutate

variables

filter

arrange

observations

summarize

select

mutate

variables

filter

arrange

observations

summarize

groups

Year	ActualElapsedTime
Month	AirTime
DayofMonth	ArrDelay
DayofWeek	DepDelay
DepTime	Origin
ArrTime	Dest
UniqueCarrier	Distance
FlightNum	TaxiIn
TailNum	TaxiOut
	Cancelled
	CancellationCode
	Diverted

Year	ActualElapsedTime
Month	AirTime
DayofMonth	ArrDelay
DayofWeek	DepDelay
DepTime	Origin
ArrTime	Dest
UniqueCarrier	Distance
FlightNum	TaxiIn
TailNum	TaxiOut
	Cancelled
	CancellationCode
	Diverted

tbl

columns to
select

```
select(df, Group, Sum)
```

tbl

columns to
select

select(df, Group, Sum)

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

tbl

columns to
select

select(df, Group, Sum)

Group	Sum
A	6
A	9
B	4
B	4
C	4
C	4

Console ~/ ↵

> select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)

Source: local data frame [227,496 x 4]

	ActualElapsedTime	AirTime	ArrDelay	DepDelay
5424	60	40	-10	0
5425	60	45	-9	1
5426	70	48	-8	-8
5427	70	39	3	3
5428	62	44	-3	5
5429	64	45	-7	-1
5430	70	43	-1	-1
5431	59	40	-16	-5
5432	71	41	44	43
5433	70	45	43	43
...

> |

Console ~/ ↗

> `hflights`

Source: local data frame [227,496 x 21]

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier
5424	2011	1	1	6	1400	1500	AA
5425	2011	1	2	7	1401	1501	AA
5426	2011	1	3	1	1352	1502	AA
5427	2011	1	4	2	1403	1513	AA
5428	2011	1	5	3	1405	1507	AA
5429	2011	1	6	4	1359	1503	AA
5430	2011	1	7	5	1359	1509	AA
5431	2011	1	8	6	1355	1454	AA
5432	2011	1	9	7	1443	1554	AA
5433	2011	1	10	1	1443	1553	AA
...

Variables not shown: FlightNum (int), TailNum (chr), ActualElapsedTime
(int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr),
Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled
(int), CancellationCode (chr), Diverted (int)

|

Console ~/ ↗

```
> select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)
```

Console ~/ ↗

> select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)

Source: local data frame [227,496 x 4]

	ActualElapsedTime	AirTime	ArrDelay	DepDelay
5424	60	40	-10	0
5425	60	45	-9	1
5426	70	48	-8	-8
5427	70	39	3	3
5428	62	44	-3	5
5429	64	45	-7	-1
5430	70	43	-1	-1
5431	59	40	-16	-5
5432	71	41	44	43
5433	70	45	43	43
...

> |

Console ~/ ↗

```
> select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)
Source: local data frame [227,496 x 4]
```

	ActualElapsedTime	AirTime	ArrDelay	DepDelay
5424	60	40	-10	0
5425	60	45	-9	1
5426	70	48	-8	-8
5427	70	39	3	3
5428	62	44	-3	5
5429	64	45	-7	-1
5430	70	43	-1	-1
5431	59	40	-16	-5
5432	71	41	44	43
5433	70	45	43	43
...
>	>	>	>	>

```
> h1 <- select(hflights, ActualElapsedTime, AirTime, ArrDelay, DepDelay)
```

Group	dose 1	dose 2
A	3	3
A	4	5
B	3	1
B	1	3
C	1	3
C	2	2

mutate

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

mutate

length x width x height =

length	width	height
2	3	3
2	4	5
3	3	1
1	1	3

length x width x height = volume

length	width	height	volume
2	3	3	18
2	4	5	40
3	3	1	9
1	1	3	3

length x width x height = volume

mass / volume =

mass	volume
50	10
45	15
35	10

length x width x height = volume

mass / volume = density

mass	volume	density
50	10	5
45	15	3
35	10	3.5

```
mutate(h1, loss = ArrDelay - DepDelay)
```

tbl

```
mutate(h1, loss = ArrDelay - DepDelay)
```

tbl

new column
name

```
mutate(h1, loss = ArrDelay - DepDelay)
```

tbl new column name = expression

```
mutate(h1, loss = ArrDelay - DepDelay)
```

tbl

new column
name

=

expression

```
mutate(h1, loss = ArrDelay - DepDelay)
```

Year	ArrDelay	DepDelay	loss
2011	-10	0	-10
2011	-9	1	-10
2011	-8	-8	0
2011	3	3	0
2011	-3	5	-8
2011	-7	-1	-6

Console ~/ ↗

```
> mutate(h1, loss = ArrDelay - DepDelay)
```

Source: local data frame [227,496 x 5]

	ActualElapsedTime	AirTime	ArrDelay	DepDelay	loss
1	60	40	-10	0	-10
2	60	45	-9	1	-10
3	70	48	-8	-8	0
4	70	39	3	3	0
5	62	44	-3	5	-8
6	64	45	-7	-1	-6
7	70	43	-1	-1	0
8	59	40	-16	-5	-11
9	71	41	44	43	1
10	70	45	43	43	0
..
>					

Console ~/ ↗

```
> mutate(h1, loss = ArrDelay - DepDelay)
```

Source: local data frame [227,496 x 5]

	ActualElapsedTime	AirTime	ArrDelay	DepDelay	loss
1	60	40	-10	0	-10
2	60	45	-9	1	-10
3	70	48	-8	-8	0
4	70	39	3	3	0
5	62	44	-3	5	-8
6	64	45	-7	-1	-6
7	70	43	-1	-1	0
8	59	40	-16	-5	-11
9	71	41	44	43	1
10	70	45	43	43	0
..
>	h2 <- mutate(h1, loss = ArrDelay - DepDelay)				
~					

```
> mean(h2$loss, na.rm = TRUE)  
[1] -2.32
```

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

filter

Group	dose 1	dose 2	Sum
A	3	3	6
C	1	3	4
C	2	2	4

filter

Console ~/ ↗

```
> f1 <- select(hflights, starts_with("Cancel"), DepDelay)
```

```
> f1
```

```
Source: local data frame [227,496 x 3]
```

	Cancelled	CancellationCode	DepDelay
5424	0		0
5425	0		1
5426	0		-8
5427	0		3
5428	0		5
5429	0		-1
5430	0		-1
5431	0		-5
5432	0		43
5433	0		43
...

```
> |
```

```
filter(hflights, Cancelled == 1)
```

tbl

logical test

```
filter(hflights, Cancelled == 1)
```

tbl

logical test

```
filter(hflights, Cancelled == 1)
```

Year	Cancelled	Dest
2011	0	DFW
2011	1	DFW
2011	0	ELP
2011	1	ELP

tbl

logical test

```
filter(hflights, Cancelled == 1)
```

Year	Cancelled	Dest
2011	1	DFW
2011	1	ELP

Console ~/ ↵

> filter(f1, Cancelled == 1)

Source: local data frame [2,973 x 3]

	Cancelled	CancellationCode	DepDelay
1	1	A	NA
2	1	B	NA
3	1	B	NA
4	1	A	NA
5	1	B	NA
6	1	B	NA
7	1	B	NA
8	1	B	NA
9	1	A	NA
10	1	B	NA
...

> |

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	2	3	4
C	1	3	4
C	5	2	4

arrange

Group	dose 1	dose 2	Sum
C	1	3	4
B	2	3	4
B	3	1	4
A	3	3	6
A	4	5	9
C	5	2	4

arrange

tbl

column
name

```
arrange(hflights, DepDelay)
```

Year	DepDelay	Dest
2011	-2	DFW
2011	3	DFW
2011	0	ELP
2011	10	ELP

tbl

column
name

```
arrange(hflights, DepDelay)
```

Year	DepDelay	Dest
2011	-2	DFW
2011	0	ELP
2011	3	DFW
2011	10	ELP

Console ~/ ↗

```
> a1 <- select(hflights, TailNum, contains("Delay"))
> a1
```

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
5424	N576AA	-10	0
5425	N557AA	-9	1
5426	N541AA	-8	-8
5427	N403AA	3	3
5428	N492AA	-3	5
5429	N262AA	-7	-1
5430	N493AA	-1	-1
5431	N477AA	-16	-5
5432	N476AA	44	43
5433	N504AA	43	43
..
>			

Console ~/ ↗

> arrange(a1, DepDelay)

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
1	N728SK	-25	-33
2	N648MQ	-23	-23
3	N13908	-12	-19
4	N11107	-40	-19
5	N27610	-17	-18
6	N134EV	-31	-18
7	N14960	-15	-17
8	N14604	-9	-17
9	N13995	-17	-17
10	N502MQ	-23	-17
..
>			

Console ~/ ↗

> arrange(a1, DepDelay)

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
1	N728SK	-25	-33
2	N648MQ	-23	-23
3	N13908	-12	-19
4	N11107	-40	-19
5	N27610	-17	-18
6	N134EV	-31	-18
7	N14960	-15	-17
8	N14604	-9	-17
9	N13995	-17	-17
10	N502MQ	-23	-17
..

> |

Console ~/ ↗

> arrange(a1, DepDelay, ArrDelay)

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
1	N728SK	-25	-33
2	N648MQ	-23	-23
3	N11107	-40	-19
4	N13908	-12	-19
5	N134EV	-31	-18
6	N27610	-17	-18
7	N14943	-33	-17
8	N879AS	-32	-17
9	N368NB	-31	-17
10	N26215	-28	-17
..
>			

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

summarise

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

n	min	mean	max
6	4	5.2	9

summarise

tbl

new column
name

=

expression

```
summarise(df, sum = sum(A),  
          avg = mean(B),  
          var = var(B))
```

tbl

new column
name

=

expression

```
summarise(df, sum = sum(A),  
          avg = mean(B),  
          var = var(B))
```

A	B	C
105	6	20
108	3	18
144	3	7
132	5	8

tbl

new column
name

=

expression

```
summarise(df, sum = sum(A),  
          avg = mean(B),  
          var = var(B))
```

A	B	C
105	6	20
108	3	18
144	3	7
132	5	8

sum avg var

489	4.25	44.92
-----	------	-------

```
summarise(df, sum = sum(A),  
          avg = mean(B),  
          var = var(B))
```

```
summarise(df, sum = sum(A),  
          avg = mean(B),  
          var = var(B))
```

Console ~/ ↗

> a1

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
5424	N576AA	-10	0
5425	N557AA	-9	1
5426	N541AA	-8	-8
5427	N403AA	3	3
5428	N492AA	-3	5
5429	N262AA	-7	-1
5430	N493AA	-1	-1
5431	N477AA	-16	-5
5432	N476AA	44	43
5433	N504AA	43	43
..

> |

Console ~/ ↗

> a1

Source: local data frame [227,496 x 3]

	TailNum	ArrDelay	DepDelay
5424	N576AA	-10	0
5425	N557AA	-9	1
5426	N541AA	-8	-8
5427	N403AA	3	3
5428	N492AA	-3	5
5429	N262AA	-7	-1
5430	N493AA	-1	-1
5431	N477AA	-16	-5
5432	N476AA	44	43
5433	N504AA	43	43
..

> a1 <- filter(a1, !is.na(DepDelay))

>

Console ~/ ↗

```
> summarise(a1, min = min(DepDelay), max = max(DepDelay), avg = mean(Dep  
y), med = median(DepDelay))
```

Source: local data frame [1 x 4]

	min	max	avg	med
1	-33	981	9.444951	0

> |

A	B	C
105	6	20
108	3	18
144	3	7
132	5	8



summarise

sum	avg	var
489	4.25	44.92

```
a1 <- select(a, X, Y, Z)
a2 <- filter(a1, X > Y)
a3 <- mutate(a2, Q = X + Y + Z)
a4 <- summarise(a3, all = sum(Q))
```

```
a1 <- select(a, X, Y, Z)
a2 <- filter(a1, X > Y)
a3 <- mutate(a2, Q = X + Y + Z)
a4 <- summarise(a3, all = sum(Q))
```

```
summarise(  
  mutate(  
    filter(  
      select(a, X, Y, Z),  
      X > Y),  
      Q = X + Y + Z),  
    all = sum(Q))  
)
```

%>%

magrittr

some
object

pipe

some
function

arguments

```
object %>% function(____, arg2, arg3, ...)
```

Console ~/ ↗

```
> c(1,2,3) %>% sum()
[1] 6
> c(1,2,3, NA) %>% sum(na.rm = TRUE)
[1] 6
> hflights$ArrDelay %>% hist(co|
```

Console ~/ ↗

```
> c(1,2,3) %>% sum()
[1] 6
> c(1,2,3, NA) %>% sum(na.rm = TRUE)
[1] 6
> hflights$ArrDelay %>% hist(col = "steelblue", border = "white", xlim =
50,400))
> |
```

```
summarise(  
  mutate(  
    filter(  
      select(a, X, Y, Z),  
      X > Y),  
      Q = X + Y + Z),  
  all = sum(Q))  
)
```

```
a %>%  
  select(X, Y, Z) %>%  
  filter(X > Y) %>%  
  mutate(Q = X + Y + Z) %>%  
  summarise(all = sum(Q))
```

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

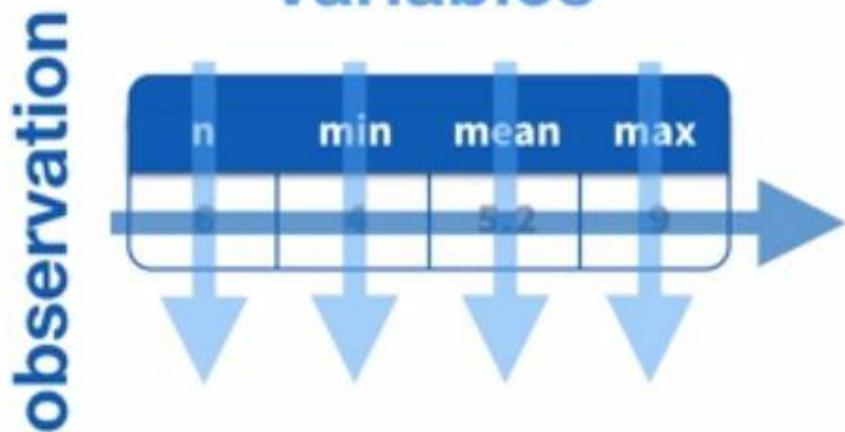


n	min	mean	max
6	4	5.2	9

summarise

Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9
B	3	1	4
B	1	3	4
C	1	3	4
C	2	2	4

variables



Group	dose 1	dose 2	Sum
A	3	3	6
A	4	5	9

Group	Total
A	15

B	3	1	4
B	1	3	4

B	8
---	---

C	1	3	4
C	2	2	4

C	8
---	---

n	min	mean	max
6	4	5.2	9

group_by

tbl

column to
group by

group_by(df, Group)

tbl

column to
group by

```
group_by(df, Group)
```

```
df %>%
```

```
group_by(Group)
```

```
Untitled1* 
Source on Save | Run | Source | 
1 hflights %>%
2   group_by(UniqueCarrier) %>%
3   summarise(avgDep = mean(DepDelay, na.rm = T),
4             avgArr = mean(ArrDelay, na.rm = T)) %>%
5   arrange(avgArr, avgDep)

1:1 (Top Level) : R Script : 

Console ~/ ~
> hflights %>%
+   group_by(UniqueCarrier)
Source: local data frame [227,496 x 21]
Groups: UniqueCarrier

  Year Month DayofMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum TailNum
1 5424 2011         1          1        6    1400    1500        AA       428  N576AA
2 5425 2011         1          2        7    1401    1501        AA       428  N557AA
3 5426 2011         1          3        1    1352    1502        AA       428  N541AA
4 5427 2011         1          4        2    1403    1513        AA       428  N403AA
5 5428 2011         1          5        3    1405    1507        AA       428  N492AA
6 5429 2011         1          6        4    1359    1503        AA       428  N262AA
7 5430 2011         1          7        5    1359    1509        AA       428  N493AA
8 5431 2011         1          8        6    1355    1454        AA       428  N477AA
9 5432 2011         1          9        7    1443    1554        AA       428  N476AA
10 5433 2011         1         10        1    1443    1553        AA       428  N504AA
...
Variables not shown: ActualElapsedTime (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)
> |
```

The screenshot shows the RStudio interface. The top panel is a code editor titled "Untitled1*". It contains the following R code:

```
1 hflights %>%
2   group_by(UniqueCarrier) %>%
3   summarise(avgDep = mean(DepDelay, na.rm = T),
4             avgArr = mean(ArrDelay, na.rm = T)) %>%
5   arrange(avgArr, avgDep)
```

The code editor has a toolbar with icons for file operations, search, and source control. The status bar at the bottom left shows "1:1" and "(Top Level)". The status bar at the bottom right shows "R Script".

The bottom panel is a "Console" window. It displays the same R code and its output:

```
> hflights %>%
+   group_by(UniqueCarrier) %>%
+   summarise(avgDep = mean(DepDelay, na.rm = T),
+             avgArr = mean(ArrDelay, na.rm = T))
Source: local data frame [15 x 3]
```

Below the console, the data frame is displayed as a table:

	UniqueCarrier	avgDep	avgArr
1	AA	6.390144	0.8917558
2	AS	3.712329	3.1923077
3	B6	13.320532	9.8588410
4	CO	9.261313	6.0986983
5	DL	9.370627	6.0841374
6	EV	12.482193	7.2569543
7	F9	5.093637	7.6682692
8	FL	4.716376	1.8536239
9	MQ	11.071745	7.1529751
10	OO	8.885482	8.6934922
11	UA	12.918707	10.4628628
12	US	1.622926	-0.6307692
13	WN	13.488241	7.5871430
14	XE	7.713728	8.1865242
15	YV	1.538462	4.0128205

Untitled1* Source on Save

Run Source

```
1 hflights %>%
2   group_by(UniqueCarrier) %>%
3   summarise(avgDep = mean(DepDelay, na.rm = T),
4             avgArr = mean(ArrDelay, na.rm = T)) %>%
5   arrange(avgArr, avgDep)
```

1:1 (Top Level) R Script

Console ~ / ↻

```
> hflights %>%
+   group_by(UniqueCarrier) %>%
+   summarise(avgDep = mean(DepDelay, na.rm = T),
+             avgArr = mean(ArrDelay, na.rm = T)) %>%
+   arrange(avgArr, avgDep)
Source: local data frame [15 x 3]
```

	UniqueCarrier	avgDep	avgArr
1	US	1.622926	-0.6307692
2	AA	6.390144	0.8917558
3	FL	4.716376	1.8536239
4	AS	3.712329	3.1923077
5	YV	1.538462	4.0128205
6	DL	9.370627	6.0841374
7	CO	9.261313	6.0986983
8	MQ	11.071745	7.1529751
9	EV	12.482193	7.2569543
10	WN	13.488241	7.5871430
11	F9	5.093637	7.6682692
12	XE	7.713728	8.1865242
13	OO	8.885482	8.6934922
14	B6	13.320532	9.8588410
15	UA	12.918707	10.4628628

> |

The screenshot shows the RStudio interface. The top panel is a script editor titled "Untitled1" containing R code. The bottom panel is a console window showing the execution of the code and its output.

```
1 hflights %>%
2   group_by(UniqueCarrier) %>%
3   summarise(avgDep = mean(DepDelay, na.rm = T),
4             avgArr = mean(ArrDelay, na.rm = T)) %>%
5   arrange(avgArr, avgDep)
```

Console output:

```
> hflights %>%
+   group_by(UniqueCarrier) %>%
+   summarise(avgDep = mean(DepDelay, na.rm = T),
+             avgArr = mean(ArrDelay, na.rm = T)) %>%
+   arrange(avgArr, avgDep)
Source: local data frame [15 x 3]
```

	UniqueCarrier	avgDep	avgArr
1	US	1.622926	-0.6307692
2	AA	6.390144	0.8917558
3	FL	4.716376	1.8536239
4	AS	3.712329	3.1923077
5	YV	1.538462	4.0128205
6	DL	9.370627	6.0841374
7	CO	9.261313	6.0986983
8	MQ	11.071745	7.1529751
9	EV	12.482193	7.2569543
10	WN	13.488241	7.5871430
11	F9	5.093637	7.6682692
12	XE	7.713728	8.1865242
13	OO	8.885482	8.6934922
14	B6	13.320532	9.8588410
15	UA	12.918707	10.4628628

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Includes file navigation icons (File, Open, Save, etc.) and a "Go to file/function" search bar.
- Script Editor:** Displays an R script with four lines of code:

```
1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())
```
- Console:** Shows the prompt ">" followed by a blank line where output would appear.
- Run Button:** Located in the top right of the script editor area.
- Source Tab:** Located in the top right of the script editor area.
- Environment Pane:** On the right side, it shows the "Global Envir" with two entries: "Name" and "hfligh...".
- Files and Plots Tabs:** Located at the bottom right of the interface.

Untitled1* Untitled2*

Source on Save Run Source

```

1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())

```

1:1 (Top Level) R Script

Console

```

> hflights %>%
+   group_by(UniqueCarrier, Dest)
Source: local data frame [227,496 x 21]
Groups: UniqueCarrier, Dest

```

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum
5424	2011	1	1	6	1400	1500	AA	428	N576AA
5425	2011	1	2	7	1401	1501	AA	428	N557AA
5426	2011	1	3	1	1352	1502	AA	428	N541AA
5427	2011	1	4	2	1403	1513	AA	428	N403AA
5428	2011	1	5	3	1405	1507	AA	428	N492AA
5429	2011	1	6	4	1359	1503	AA	428	N262AA
5430	2011	1	7	5	1359	1509	AA	428	N493AA
5431	2011	1	8	6	1355	1454	AA	428	N477AA
5432	2011	1	9	7	1443	1554	AA	428	N476AA
5433	2011	1	10	1	1443	1553	AA	428	N504AA
...

Variables not shown: ActualElapsedTime (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)

> |

Environment

Global Envir.

Name

hfligh...

Files Plots

Untitled1* Untitled2*

Source on Save Run Source

```

1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())

```

1:1 (Top Level) R Script

Console

```

> hflights %>%
+   group_by(UniqueCarrier, Dest)
Source: local data frame [227,496 x 21]
Groups: UniqueCarrier, Dest

```

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum	TailNum
5424	2011	1	1	6	1400	1500	AA	428	N576AA
5425	2011	1	2	7	1401	1501	AA	428	N557AA
5426	2011	1	3	1	1352	1502	AA	428	N541AA
5427	2011	1	4	2	1403	1513	AA	428	N403AA
5428	2011	1	5	3	1405	1507	AA	428	N492AA
5429	2011	1	6	4	1359	1503	AA	428	N262AA
5430	2011	1	7	5	1359	1509	AA	428	N493AA
5431	2011	1	8	6	1355	1454	AA	428	N477AA
5432	2011	1	9	7	1443	1554	AA	428	N476AA
5433	2011	1	10	1	1443	1553	AA	428	N504AA
...

Variables not shown: ActualElapsedTime (int), AirTime (int), ArrDelay (int), DepDelay (int), Origin (chr), Dest (chr), Distance (int), TaxiIn (int), TaxiOut (int), Cancelled (int), CancellationCode (chr), Diverted (int)

>

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Includes icons for file operations (New, Open, Save, etc.) and a "Go to file/function" search bar.
- Code Editor:** Displays R code in the "Untitled2" script pane:

```
1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())
```
- Console:** Shows the prompt ">" followed by a blank line where output would appear.
- Environment:** A sidebar pane listing objects in the global environment, including "hflights".
- Plots:** A small preview pane showing a histogram or similar plot.

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Contains icons for file operations (New, Open, Save, etc.) and a "Go to file/function" search bar.
- Script Pane:** Displays R code in a syntax-highlighted editor. The code uses the dplyr package to group flights by carrier and destination, then summarise the count of flights and destinations. The code is as follows:

```
1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())
```
- Console Pane:** Shows the execution of the R code. It displays the source code followed by the resulting data frame information and its contents.

```
> hflights %>%
+   group_by(UniqueCarrier, Dest) %>%
+   summarise(nflights = n())
Source: local data frame [241 x 3]
Groups: UniqueCarrier [241]

  UniqueCarrier Dest nflights
1          AA  DFW     2105
2          AA  MIA     1139
3          AS  SEA      365
4          B6  JFK      695
5          CO  ABQ      101
6          CO  AGS       1
7          CO  ANC      125
8          CO  ATL      927
9          CO  AUS     2645
10         CO  BHM       5
...
> |
```
- Environment Sidebar:** Shows the global environment with objects like "hflights" listed under "Global Envir".
- Bottom Right:** Buttons for "Files" and "Plots".

Untitled1* Untitled2* |

Source on Save Run Source

```
1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())
```

1:1 (Top Level) R Script

Console ~/

```
> hflights %>%
+   group_by(UniqueCarrier, Dest) %>%
+   summarise(nflights = n())
Source: local data frame [241 x 3]
Groups: UniqueCarrier
```

	UniqueCarrier	Dest	nflights
1	AA	DFW	2105
2	AA	MIA	1139
3	AS	SEA	365
4	B6	JFK	695
5	CO	ABQ	101
6	CO	AGS	1
7	CO	ANC	125
8	CO	ATL	927
9	CO	AUS	2645
10	CO	BHM	5
..

Files Plots

Go to file/function

Untitled1* Untitled2*

Source on Save Run Source

```
1 hflights %>%
2   group_by(UniqueCarrier, Dest) %>%
3   summarise(nflights = n()) %>%
4   summarise(ndests = n())
```

1:1 (Top Level) R Script

Console ~/

```
> hflights %>%
+   group_by(UniqueCarrier, Dest) %>%
+   summarise(nflights = n())
Source: local data frame [241 x 3]
Groups: UniqueCarrier
```

	UniqueCarrier	Dest	nflights
1	AA	DFW	2105
2	AA	MIA	1139
3	AS	SEA	365
4	B6	JFK	695
5	CO	ABQ	101
6	CO	AGS	1
7	CO	ANC	125
8	CO	ATL	927
9	CO	AUS	2645
10	CO	BHM	5
..

Files Plots

Untitled1* Untitled2* | Source on Save | Run | Source | Environment
Global Envir... Name hfligh...
1 hflights %>%
2 group_by(UniqueCarrier, Dest) %>%
3 summarise(nflights = n()) %>%
4 summarise(ndests = n())

1:1 (Top Level) R Script

Console ~/

	CU	ANL	TZC
8	CO	ATL	927
9	CO	AUS	2645
10	CO	BHM	5
..

> hflights %>%
+ group_by(UniqueCarrier, Dest) %>%
+ summarise(nflights = n()) %>%
+ summarise(ndests = n())
Source: local data frame [15 x 2]

	UniqueCarrier	ndests
1	AA	2
2	AS	1
3	B6	1
4	CO	62
5	DL	3
6	EV	5
7	F9	2
8	FL	2
9	MQ	3
10	OO	34
11	UA	6
12	US	3
13	WN	33
14	XE	81
15	YV	3

>

Source

Console -/ Groups: UniqueCarrier

	UniqueCarrier	Dest	nflights
1	AA	DFW	2105
2	AA	MIA	1139
3	AS	SEA	365
4	B6	JFK	695
5	CO	ABQ	101
6	CO	AGS	1
7	CO	ANC	125
8	CO	ATL	927
9	CO	AUS	2645
10	CO	BHM	5
..

```
> hflights %>%
+   group_by(UniqueCarrier, Dest) %>%
+   summarise(nflights = n()) %>%
+   summarise(ndests = n())
```

Source: local data frame [15 x 2]

	UniqueCarrier	ndests
1	AA	2
2	AS	1
3	B6	1
4	CO	62
5	DL	3
6	EV	5
7	F9	2
8	FL	2
9	MQ	3
10	OO	34
11	UA	6
12	US	3
13	WN	33
14	XE	81
15	YV	2

Environment

Global Envir

Name

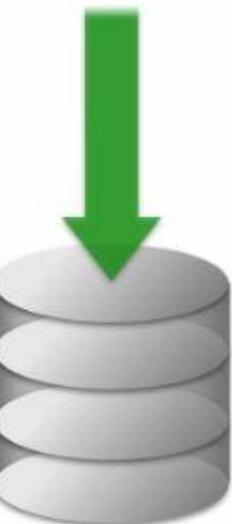
hfligh...

Files Plots

data frame

data table

database



database