R basic
基本運算&常用指令2

用DPLYR套件整理結構化的資料

BY R-BASIC GROUP 4 SANA/ANGIE

5個dplyr 套件最常用的函數

filter	自訂條件濾掉column中的資料。
arrange	將資料根據特定欄位來排序。
select	挑選特定column出來 。
mutate	以現有的column資料做運算，形成新的column。
summarise＋group_by	將目前的資料做統計運算，形成統計結論。

邏輯判斷符號

<、>	小於、大於。
<=、>=	小於等於、大於等於。
==、!=	等於、不等於。
A %in% B	A 是否在 B 中。
&&、&	交集，& 適用於向量式的邏輯判斷，&& 適用於單一值的邏輯判斷。
||、|	聯集，| 適用狀況與 & 相同，|| 適用狀況與 && 相同。

以nycflights13套件中的Flights 資料集為例:

- **Flights：所有於2013年在紐約起降的飛機資料**

Untitled1* × flights ×

Filter

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum
1	2013	1	1	517	515	2	830	819	11	UA	1545	N14228
2	2013	1	1	533	529	4	850	830	20	UA	1714	N24211
3	2013	1	1	542	540	2	923	850	33	AA	1141	N619AA
4	2013	1	1	544	545	-1	1004	1022	-18	B6	725	N804JB
5	2013	1	1	554	600	-6	812	837	-25	DL	461	N668DN
6	2013	1	1	554	558	-4	740	728	12	UA	1696	N39463
7	2013	1	1	555	600	-5	913	854	19	B6	507	N516JB
8	2013	1	1	557	600	-3	709	723	-14	EV	5708	N829AS
9	2013	1	1	557	600	-3	838	846	-8	B6	79	N593JB
10	2013	1	1	558	600	-2	753	745	8	AA	301	N3ALAA
11	2013	1	1	558	600	-2	849	851	-2	B6	49	N793JB
12	2013	1	1	558	600	-2	853	856	-3	B6	71	N657JB
13	2013	1	1	558	600	-2	924	917	7	UA	194	N29129

Showing 1 to 14 of 336,776 entries

DPLYR指令用法


```
>filter(flights, arr_delay>0)
```

filter() 根據特定條件篩選資料

#2013年1月2日當天紐約有哪些航

```
> filter(flights, month == 1, day == 2)
# A tibble: 943 × 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl>   <chr>  <int>
1   2013     1     2       42           2359        43      518            442        36      B6    707
2   2013     1     2      126           2250       156      233           2359       154      B6     22
3   2013     1     2      458            500        -2      703            650        13      US   1030
4   2013     1     2      512            515        -3      809            819       -10      UA   1453
5   2013     1     2      535            540        -5      831            850       -19      AA   1141
6   2013     1     2      536            529         7      840            828        12      UA    407
7   2013     1     2      539            545        -6      959           1022       -23      B6    725
8   2013     1     2      554            600        -6      845            901       -16      B6    125
9   2013     1     2      554            600        -6      841            851       -10      B6     49
10  2013     1     2      554            600        -6      909            858        11      B6    371
# ... with 933 more rows, and 8 more variables: tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

filter() 根據特定條件篩選資料

#在2013年的紐約，共有多少班次的飛機起飛延誤呢？

```
> filter(flights,dep_delay>0)
# A tibble: 128,432 × 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier flight
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl>   <chr>  <int>
1   2013     1     1      517            515         2      830            819        11      UA   1545
2   2013     1     1      533            529         4      850            830        20      UA   1714
3   2013     1     1      542            540         2      923            850        33      AA   1141
4   2013     1     1      601            600         1      844            850        -6      B6    343
5   2013     1     1      608            600         8      807            735        32      MQ   3768
6   2013     1     1      611            600        11      945            931        14      UA    303
7   2013     1     1      613            610         3      925            921         4      B6    135
8   2013     1     1      623            610        13      920            915         5      AA   1837
9   2013     1     1      632            608        24      740            728        12      EV   4144
10  2013     1     1      644            636         8      931            940        -9      UA   1701
# ... with 128,422 more rows, and 8 more variables: tailnum <chr>, origin <chr>, dest <chr>,
#   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

arrange() 將資料根據特定欄位來排序

#在2013年的紐約，隨機抽取10個航班

```
> flights2 <- flights[sample(1:336776,10), ]
> arrange(flights2, month, day)
# A tibble: 10 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	1711	1600	71	2005
2	2013	2	3	1801	1810	-9	2155
3	2013	3	12	1027	1030	-3	1302
4	2013	3	27	559	600	-1	715
5	2013	4	17	648	700	-12	951
6	2013	7	8	1812	1759	13	2142
7	2013	8	7	1638	1645	-7	1810

arrange() 將資料根據特定欄位來排序

 # desc():將資料改成遞減排列的函數

```
> arrange(flights2, desc(month), desc(day))
# A tibble: 10 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>
1   2013    12     4     1636           1640        -4     1809
2   2013    11    26     1707           1700         7     2011
3   2013     8    11     2149           2100        49       19
4   2013     8     7     1638           1645        -7     1810
5   2013     7     8     1812           1759        13     2142
6   2013     4    17      648            700       -12      951
7   2013     3    27      559            600        -1      715
8   2013     3    12     1027           1030        -3     1302
```

select() 挑選特定欄位進行分析

```
> 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
7   2013     1     1
8   2013     1     1
9   2013     1     1
10  2013     1     1
```

select() 挑選特定欄位進行分析

```
> select(flights, dep_time:time_hour)
# A tibble: 336,776 × 16
   dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
      <int>          <int>     <dbl>    <int>          <int>     <dbl>
1       517            515         2      830            819        11
2       533            529         4      850            830        20
3       542            540         2      923            850        33
4       544            545        -1     1004           1022       -18
5       554            600        -6      812            837       -25
> select(flights, -(year:day))
# A tibble: 336,776 × 16
   dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
      <int>          <int>     <dbl>    <int>          <int>     <dbl>
1       517            515         2      830            819        11
2       533            529         4      850            830        20
3       542            540         2      923            850        33
4       544            545        -1     1004           1022       -18
5       554            600        -6      812            837       -25
```

mutate() 產生新欄位(變數)

#mutate()產生的新欄位會被排在最末欄,為方便檢視,通常會先縮小資料集#縮小資料集flights為flights_sml (19欄→7欄)

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

	year	month	day	dep_delay	arr_delay	distance	air_time
1	2013	1	1	2	11	1400	227
2	2013	1	1	4	20	1416	227
3	2013	1	1	2	33	1089	160
4	2013	1	1	-1	-18	1576	183
5	2013	1	1	-6	-25	762	116

mutate() 產生新欄位(變數)

#算出飛機飛行速率speed=distance/air_time

```
> mutate(flights_sml,
+         speed_per_hour = distance / air_time * 60
+ )
# A tibble: 336,776 x 8
    year month   day dep_delay arr_delay distance air_time speed_per_hour
   <int> <int> <int>     <dbl>     <dbl>    <dbl>    <dbl>          <dbl>
1   2013     1     1         2        11     1400      227       370.0441
2   2013     1     1         4        20     1416      227       374.2731
3   2013     1     1         2        33     1089      160       408.3750
4   2013     1     1        -1       -18     1576      183       516.7213
5   2013     1     1        -6       -25      762      116       394.1379
6   2013     1     1        -4        12      719      150       287.6000
7   2013     1     1        -5        19     1065      158       404.4304
8   2013     1     1        -3       -14      229       53       259.2453
9   2013     1     1        -3        -8      944      140       404.5714
10  2013     1     1        -2         8      733      138       318.6957
```

summarise()+group_by() 合併匯總

#算出航班起飛的平均延遲時間

```
> summarise(flights, delay = mean(dep_delay, na.rm = TRUE))
# A tibble: 1 x 1
     delay
     <dbl>
1 12.63907
```

summarise()+group_by() 合併匯總

#把year,month,day先group_by組合起來
#再利用summarise()算出”單日”航班起飛的平均延遲時間

```
> by_day <- group_by(flights, year, month, day)
> summarise(by_day, delay = mean(dep_delay, na.rm = TRUE))
Source: local data frame [365 x 4]
Groups: year, month [?]

    year month   day     delay
   <int> <int> <int>     <dbl>
1   2013     1     1 11.548926
2   2013     1     2 13.858824
3   2013     1     3 10.987832
4   2013     1     4  8.951595
5   2013     1     5  5.732218
6   2013     1     6  7.148014
7   2013     1     7  5.417204
8   2013     1     8  2.553073
9   2013     1     9  2.276477
10  2013     1    10  2.844995
```