Dplyr

Dplyr library is used for data manipulation

Installation

In [1]:
```
install.packages("dplyr")
```

In [3]:
```
options(warn=-1)
library("dplyr")
```

Installation of nyc flights data

In [4]:
```
install.packages('nycflights13')
```

In [5]:
```
library(nycflights13)
```

In [6]:
```
head(flights)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	1	517	515	2	830	819	11	
2013	1	1	533	529	4	850	830	20	
2013	1	1	542	540	2	923	850	33	
2013	1	1	544	545	-1	1004	1022	-18	
2013	1	1	554	600	-6	812	837	-25	
2013	1	1	554	558	-4	740	728	12	

In [8]:
```
summary(flights)
```

```
      year          month             day           dep_time    sched_dep_time
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
 Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
 Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
 Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
                                                 NA's   :8255
   dep_delay          arr_time    sched_arr_time   arr_delay
 Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
 1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
 Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
 Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
 3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
 Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
 NA's   :8255      NA's   :8713                  NA's   :9430
   carrier             flight       tailnum             origin
 Length:336776      Min.   :   1   Length:336776      Length:336776
 Class :character   1st Qu.: 553   Class :character   Class :character
 Mode  :character   Median :1496   Mode  :character   Mode  :character
                    Mean   :1972
                    3rd Qu.:3465
                    Max.   :8500

    dest              air_time        distance         hour
 Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
 Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
 Mode  :character   Median :129.0   Median : 872   Median :13.00
                    Mean   :150.7   Mean   :1040   Mean   :13.18
                    3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
                    Max.   :695.0   Max.   :4983   Max.   :23.00
                    NA's   :9430
    minute        time_hour
 Min.   : 0.00   Min.   :2013-01-01 05:00:00
 1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
 Median :29.00   Median :2013-07-03 10:00:00
 Mean   :26.23   Mean   :2013-07-03 05:22:54
 3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
 Max.   :59.00   Max.   :2013-12-31 23:00:00
```

In [9]:
```
dim(flights)
```

```
336776  19
```

1) filter

filter() allows you to select a subset of rows in a data frame

In [10]:
```
head(filter(flights,month==11,day==3,carrier=='AA'))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	11	3	538	545	-7	824	855	-31	
2013	11	3	556	600	-4	900	905	-5	
2013	11	3	604	610	-6	844	855	-11	
2013	11	3	624	629	-5	907	929	-22	
2013	11	3	625	630	-5	736	805	-29	
2013	11	3	653	655	-2	925	920	5	

In [11]:
```
head(flights[flights$month == 11 & flights$day == 3 & flights$carrier == 'AA',])
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	11	3	538	545	-7	824	855	-31	
2013	11	3	556	600	-4	900	905	-5	
2013	11	3	604	610	-6	844	855	-11	
2013	11	3	624	629	-5	907	929	-22	
2013	11	3	625	630	-5	736	805	-29	
2013	11	3	653	655	-2	925	920	5	

2) slice

Select row with position

In [12]:
```
slice(flights,4:8)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	1	544	545	-1	1004	1022	-18	
2013	1	1	554	600	-6	812	837	-25	
2013	1	1	554	558	-4	740	728	12	
2013	1	1	555	600	-5	913	854	19	
2013	1	1	557	600	-3	709	723	-14	

3) arrange

orders data frame as per specific columns

In [13]:
```
head(arrange(flights,year,month,day,air_time))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	1	2302	2200	62	2342	2253	49	
2013	1	1	1318	1322	-4	1358	1416	-18	
2013	1	1	2116	2110	6	2202	2212	-10	
2013	1	1	2000	2000	0	2054	2110	-16	
2013	1	1	2056	2004	52	2156	2112	44	
2013	1	1	908	915	-7	1004	1033	-29	

desc()

To arrange in descending order

In [14]:
```
head(arrange(flights,desc(arr_delay)))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	9	641	900	1301	1242	1530	1272	
2013	6	15	1432	1935	1137	1607	2120	1127	
2013	1	10	1121	1635	1126	1239	1810	1109	
2013	9	20	1139	1845	1014	1457	2210	1007	
2013	7	22	845	1600	1005	1044	1815	989	
2013	4	10	1100	1900	960	1342	2211	931	

4) select()

select() allows you to select a subset of columns in a data frame

In [15]:
```
head(select(flights,year,month))
```

year	month
2013	1
2013	1
2013	1
2013	1
2013	1
2013	1

5) rename()

rename columns

In [16]:
```
head(rename(flights,tail_num=tailnum))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	1	517	515	2	830	819	11	
2013	1	1	533	529	4	850	830	20	
2013	1	1	542	540	2	923	850	33	
2013	1	1	544	545	-1	1004	1022	-18	
2013	1	1	554	600	-6	812	837	-25	
2013	1	1	554	558	-4	740	728	12	

6) distinct

returns unique values

In [17]:
```
distinct(flights,carrier)
```

carrier
UA
AA
B6
DL
EV
MQ
US
WN
VX
FL
AS
9E
F9
HA
YV
OO

7) mutate()

adding new columns using feature engineering

In [18]:
```
head(mutate(flights,total_delay = arr_delay + dep_delay))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	1	1	517	515	2	830	819	11	
2013	1	1	533	529	4	850	830	20	
2013	1	1	542	540	2	923	850	33	
2013	1	1	544	545	-1	1004	1022	-18	
2013	1	1	554	600	-6	812	837	-25	
2013	1	1	554	558	-4	740	728	12	

transmute()

returns only new columns

In [19]:
```
head(transmute(flights, total_delay = arr_delay+dep_delay))
```

total_delay
13
24
35
-19
-31
8

8) summarize()

summarize dataframe with single value using aggregate function

In [20]:
```
# na.rm=TRUE for removing nan

summarise(flights,avg_air_time=mean(air_time,na.rm=TRUE))
```

avg_air_time
150.6865

9) Sampling methods

select sample from dataframe

sample_n()

select random number of rows

In [21]:
```
sample_n(flights,5)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carr
2013	6	27	NA	2000	NA	NA	2225	NA	M
2013	10	15	1055	1055	0	1334	1353	-19	
2013	3	25	NA	1001	NA	NA	1129	NA	
2013	7	14	732	735	-3	932	1005	-33	A
2013	1	2	637	640	-3	832	809	23	

sample_frac()

select random fraction of data

In [22]:
```
sample_frac(flights,0.3) #30%
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	c
2013	4	23	101	2130	211	325	16	189	
2013	2	3	1736	1745	-9	1924	1951	-27	
2013	4	24	1702	1715	-13	1847	1900	-13	
2013	5	18	1156	1200	-4	1304	1313	-9	
2013	10	6	830	830	0	1123	1124	-1	

Pipe operator %>%

In [23]:
```
df <- mtcars
```

In [24]:
```
df
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

In [25]:
```
a <- filter(df,mpg>20)
print(a)
```

```
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
5  24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
6  22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
7  32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
8  30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
9  33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
10 21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
11 27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
12 26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
13 30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
14 21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

In [26]:
```
b <- sample_n(a,10)
print(b)
```

```
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
2  33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
4  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
5  26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
6  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
7  21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
8  32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
9  30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
10 21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
```

In [27]:
```
c <- arrange(b,desc(hp))
```

In [28]:
```
head(c)
```

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1

Using pipe operator

In [29]:
```
df %>% filter(mpg > 20) %>% sample_n(5) %>% arrange(desc(mpg))
```

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2

group_by()

group by specific category and perform aggregate operations

In [30]:
```
df %>% group_by(gear) %>% summarize(mean_hp=mean(hp))
```

gear	mean_hp
3	176.1333
4	89.5000
5	195.6000

join()

join operation is used to merge/join multiple dataframes

In [34]:
```
c1 <- c('Tendulkar','Kohli','Dhoni','Bumrah','Chahal')
c2 <- c('Tendulkar','Shami','Umesh','Bumrah','Chahal')
c3 <- c(10000,7100,5800,890,870)
c4 <- c(11,200,220,370,420)
```

In [35]:
```
runs <- data.frame(players=c1,runs=c3)
wickets <- data.frame(players=c2,wickets=c4)
```

In [36]:
```
runs
```

players	runs
Tendulkar	10000
Kohli	7100
Dhoni	5800
Bumrah	890
Chahal	870

In [37]:
```
wickets
```

players	wickets
Tendulkar	11
Shami	200
Umesh	220
Bumrah	370
Chahal	420

In [38]:
```
inner_join(runs,wickets,by="players")
```

players	runs	wickets
Tendulkar	10000	11
Bumrah	890	370
Chahal	870	420

In [39]:
```
full_join(runs,wickets,by="players")
```

players	runs	wickets
Tendulkar	10000	11
Kohli	7100	NA
Dhoni	5800	NA
Bumrah	890	370
Chahal	870	420
Shami	NA	200
Umesh	NA	220