NYC Flights 2013 Analysis

```
#install packages
install.packages("nycflights13")
```

```
Updating HTML index of packages in '.Library'

Making 'packages.html' ...
 done
```

```
#call Lib
library(tidyverse)
library(dplyr)
library(nycflights13)
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()
```

```
#Read  CSV File
flights <- read.csv("flights.csv",stringsAsFactors = FALSE)
```

```
#display flights
tibble(flights)
```

A tibble: 336776 × 19

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tai
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<int>	<c
2013	1	1	517	515	2	830	819	11	UA	1545	N1
2013	1	1	533	529	4	850	830	20	UA	1714	N2
2013	1	1	542	540	2	923	850	33	AA	1141	N6
2013	1	1	544	545	-1	1004	1022	-18	B6	725	N8
2013	1	1	554	600	-6	812	837	-25	DL	461	N6
2013	1	1	554	558	-4	740	728	12	UA	1696	N3
2013	1	1	555	600	-5	913	854	19	B6	507	N5
2013	1	1	557	600	-3	709	723	-14	EV	5708	N8
2013	1	1	557	600	-3	838	846	-8	B6	79	N5
2013	1	1	558	600	-2	753	745	8	AA	301	N3
2013	1	1	558	600	-2	849	851	-2	B6	49	N7
2013	1	1	558	600	-2	853	856	-3	B6	71	N6
2013	1	1	558	600	-2	924	917	7	UA	194	N2
2013	1	1	558	600	-2	923	937	-14	UA	1124	N5
2013	1	1	559	600	-1	941	910	31	AA	707	N3
2013	1	1	559	559	0	702	706	-4	B6	1806	N7
2013	1	1	559	600	-1	854	902	-8	UA	1187	N7
2013	1	1	600	600	0	851	858	-7	B6	371	N5
2013	1	1	600	600	0	837	825	12	MQ	4650	N5
2013	1	1	601	600	1	844	850	-6	B6	343	N6
2013	1	1	602	610	-8	812	820	-8	DL	1919	N9
2013	1	1	602	605	-3	821	805	16	MQ	4401	N7
2013	1	1	606	610	-4	858	910	-12	AA	1895	N6
2013	1	1	606	610	-4	837	845	-8	DL	1743	N3
2013	1	1	607	607	0	858	915	-17	UA	1077	N5

2013	1	1	607	607	0	858	915	-17	UA	1077	N3
2013	1	1	608	600	8	807	735	32	MQ	3768	N9

```
#display airlines
tibble(airlines)
```

A tibble: 16 × 2

carrier	name
<chr>	<chr>
9E	Endeavor Air Inc.
AA	American Airlines Inc.
AS	Alaska Airlines Inc.
B6	JetBlue Airways
DL	Delta Air Lines Inc.
EV	ExpressJet Airlines Inc.
F9	Frontier Airlines Inc.
FL	AirTran Airways Corporation
HA	Hawaiian Airlines Inc.
MQ	Envoy Air
OO	SkyWest Airlines Inc.
UA	United Air Lines Inc.
US	US Airways Inc.
VX	Virgin America
WN	Southwest Airlines Co.
YV	Mesa Airlines Inc.

2013	1	1	615	615	0	1039	1100	-21	B6	709	N7
				5	0	833	842	-9	DL	575	N3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2013	9	30	2123	2125	-2	2223	2247	-24	EV	5489	N7
2013	9	30	2127	2129	-2	2314	2323	-9	EV	3833	N1
2013	9	30	2128	2130	-2	2328	2359	-31	B6	97	N8
2013	9	30	2129	2059	30	2230	2232	-2	EV	5048	N7
2013	9	30	2131	2140	-9	2225	2255	-30	MQ	3621	N8
2013	9	30	2140	2140	0	10	40	-30	AA	185	N3
2013	9	30	2142	2129	13	2250	2239	11	EV	4509	N1
2013	9	30	2145	2145	0	115	140	-25	B6	1103	N6
2013	9	30	2147	2137	10	30	27	3	B6	1371	N6
2013	9	30	2149	2156	-7	2245	2308	-23	UA	523	N8
2013	9	30	2150	2159	-9	2250	2306	-16	EV	3842	N1

```
## filter NA (missing values)
# write our own function
check_na <- function(col) {
  sum(is.na(col))
}

# validate NA
apply(flights, MARGIN=2, function(col) sum(is.na(col)))
```

```
year:       0 month:      0 day:       0 dep_time:      8255 sched_dep_time:      0 dep_delay:
8255 arr_time:     8713 sched_arr_time:    0 arr_delay:     9430 carrier:      0 flight:    0
tailnum:      2512 origin:      0 dest:       0 air_time:        9430 distance:      0 hour:      0
minute:       0 time_hour:     0
```

2013	9	30	2233	2113	80	112	30	42	UA	471	N5
2013	9	30	2235	2001	154	59	2249	130	B6	1083	N8
2013	9	30	2237	2245	-8	2345	2353	-8	B6	234	N3
2013	9	30	2240	2245	-5	2334	2351	-17	B6	1816	N3

2013	9	30	2240	2250	-10	2347	7	-20	B6	2002	N2

```
# filter NA on  dep_delay ,arr_delay
flights <- flights %>%
filter(!is.na(dep_delay)) %>%
filter(!is.na(arr_delay))
```

2013	9	30	NA	1842	NA	NA	2019	NA	EV	5274	N7

```
# validate NA
apply(flights, MARGIN=2, function(col) sum(is.na(col)))
```

```
year:          0 month:          0 day:          0 dep_time:          0 sched_dep_time:          0 dep_delay:          0
arr_time:          0 sched_arr_time:          0 arr_delay:          0 carrier:          0 flight:          0 tailnum:          0
origin:          0 dest:          0 air_time:          0 distance:          0 hour:          0 minute:          0 time_hour:
0
```

2013	9	30	NA	1159	NA	NA	1344	NA	MQ	3572	N5
2013	9	30	NA	840	NA	NA	1020	NA	MQ	3531	N8

```
# validate NA
apply(airlines, MARGIN=2, function(col) sum(is.na(col)))
```

```
carrier:          0 name:          0
```

"flights" Data frame with columns

Column	Description
year, month, day	Date of departure
dep_time, arr_time	Actual departure and arrival times (format HHMM or HMM), local tz.
sched_dep_time, sched_arr_time	Scheduled departure and arrival times (format HHMM or HMM), local tz.
dep_delay, arr_delay	Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
carrier	Two letter carrier abbreviation. See airlines to get name.
flight	Flight number.
tailnum	Plane tail number. See planes for additional metadata.
origin, dest	Origin and destination. See airports for additional metadata.

Column	Description
air_time	Amount of time spent in the air, in minutes.
distance	Distance between airports, in miles.
hour, minute	Time of scheduled departure broken into hour and minutes.
time_hour	Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

Q1: Number of flights each month

```
resultQ1 <- flights %>%
group_by(month) %>%
summarise(n = n()) %>%
rename(numberOfFlights = n)
```

```
#display resultQ1
resultQ1
```

A tibble: 12 × 2

month	numberOfFlights
<int>	<int>
1	26398
2	23611
3	27902
4	27564
5	28128
6	27075
7	28293
8	28756
9	27010
10	28618
11	26971
12	27020

Q2: Number of flights each carrier

```
resultQ2 <- flights %>%
group_by(carrier) %>%
summarise(n = n()) %>%
arrange(desc(n)) %>%
rename(numberOfFlights = n)  %>%
left_join(airlines,by = "carrier") %>%
select(carrier,name,numberOfFlights)
```

```
#display resultQ12
resultQ2
```

A tibble: 16 × 3

carrier	name	numberOfFlights
<chr>	<chr>	<int>
UA	United Air Lines Inc.	57782
B6	JetBlue Airways	54049
EV	ExpressJet Airlines Inc.	51108
DL	Delta Air Lines Inc.	47658
AA	American Airlines Inc.	31947
MQ	Envoy Air	25037
US	US Airways Inc.	19831
9E	Endeavor Air Inc.	17294
WN	Southwest Airlines Co.	12044
VX	Virgin America	5116
FL	AirTran Airways Corporation	3175
AS	Alaska Airlines Inc.	709
F9	Frontier Airlines Inc.	681
YV	Mesa Airlines Inc.	544
HA	Hawaiian Airlines Inc.	342
OO	SkyWest Airlines Inc.	29

Q3: Number of flights each carrier to arrival delays

```
resultQ3 <- flights %>%
filter(arr_delay > 0) %>%
group_by(carrier) %>%
summarise(n = n()) %>%
arrange(desc(n)) %>%
rename(arrivalDelayNumber = n)  %>%
left_join(airlines,by = "carrier") %>%
select(carrier,name,arrivalDelayNumber)
```

```
#display resultQ3
resultQ3
```

A tibble: 16 × 3

carrier	name	arrivalDelayNumber
<chr>	<chr>	<int>
EV	ExpressJet Airlines Inc.	24484
B6	JetBlue Airways	23609
UA	United Air Lines Inc.	22222
DL	Delta Air Lines Inc.	16413
MQ	Envoy Air	11693
AA	American Airlines Inc.	10706
US	US Airways Inc.	7349
9E	Endeavor Air Inc.	6637
WN	Southwest Airlines Co.	5304
FL	AirTran Airways Corporation	1895
VX	Virgin America	1746
F9	Frontier Airlines Inc.	392
YV	Mesa Airlines Inc.	258
AS	Alaska Airlines Inc.	189
HA	Hawaiian Airlines Inc.	97
OO	SkyWest Airlines Inc.	10

Q4: Number of flights each carrier to departure delays

```
resultQ4 <- flights %>%
filter(dep_delay > 0) %>%
group_by(carrier) %>%
summarise(n = n()) %>%
arrange(desc(n)) %>%
rename(departureDelayNumber = n)  %>%
left_join(airlines,by = "carrier") %>%
select(carrier,name,departureDelayNumber)
```

```
#display resultQ4
resultQ4
```

A tibble: 16 × 3

carrier	name	departureDelayNumber
<chr>	<chr>	<int>
UA	United Air Lines Inc.	27125
EV	ExpressJet Airlines Inc.	22976
B6	JetBlue Airways	21372
DL	Delta Air Lines Inc.	15186
AA	American Airlines Inc.	10105
MQ	Envoy Air	7966
9E	Endeavor Air Inc.	6980
WN	Southwest Airlines Co.	6535
US	US Airways Inc.	4762
VX	Virgin America	2216
FL	AirTran Airways Corporation	1647
F9	Frontier Airlines Inc.	340
YV	Mesa Airlines Inc.	232
AS	Alaska Airlines Inc.	225
HA	Hawaiian Airlines Inc.	69
OO	SkyWest Airlines Inc.	9

Q5: Max distance of each carrier

```
resultQ5 <- flights %>%
filter(arr_delay > 0) %>%
group_by(carrier) %>%
summarise(max = max(distance)) %>%
arrange(desc(max)) %>%
rename(maxDistance = max)  %>%
left_join(airlines,by = "carrier") %>%
select(carrier,name,maxDistance)
```

```
#display resultQ5
resultQ5
```

A tibble: 16 × 3

carrier	name	maxDistance
<chr>	<chr>	<int>
HA	Hawaiian Airlines Inc.	4983
UA	United Air Lines Inc.	4963
AA	American Airlines Inc.	2586
B6	JetBlue Airways	2586
DL	Delta Air Lines Inc.	2586
VX	Virgin America	2586
AS	Alaska Airlines Inc.	2402
US	US Airways Inc.	2153
WN	Southwest Airlines Co.	2133
F9	Frontier Airlines Inc.	1620
9E	Endeavor Air Inc.	1587
EV	ExpressJet Airlines Inc.	1389
MQ	Envoy Air	1147
OO	SkyWest Airlines Inc.	1008
FL	AirTran Airways Corporation	762
YV	Mesa Airlines Inc.	544