NYC flights 2013 Analysis

```
##Ctrl + M -> convert to markdown
```

```
library(dplyr)
library(readr)
#NYC flights 2013 Analysis
```

```
flights <- read_csv("flights.csv")
airlines <- read_csv("airlines.csv")
```

```
Rows: 336776 Columns: 19

── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr   (4): carrier, tailnum, origin, dest
dbl  (14): year, month, day, dep_time, sched_dep_time, dep_delay, arr_time, ..
dttm  (1): time_hour


ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this messa

Rows: 16 Columns: 2

── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (2): carrier, name


ℹ Use `spec()` to retrieve the full column specification for this data
```

```
glimpse(flights)
#tibble(flights)
```

```
Rows: 336,776
Columns: 19
$ year           <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2
$ month          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ day            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ dep_time       <dbl> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558,
$ sched_dep_time <dbl> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600,
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1
$ arr_time       <dbl> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,
$ sched_arr_time <dbl> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "
$ flight         <dbl> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733,
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0
```

Q1 : How many flights in 2013 ?

```
flights %>%
    select(carrier,flight) %>%
    summarise(n=n())
```

A tibble: 1 × 1

n
<int>
336776

A1 : There are all 336,776 flights.

Q2: How many flights of each carrier and sort data from maximum to minimum?

```
flights %>%
    group_by(carrier)%>%
    summarise(Total_of_carrier = n()) %>%
    arrange(desc(Total_of_carrier)) %>%


    inner_join (airlines,by = c("carrier" = "carrier"))%>%
    select(carrier ,name, Total_of_carrier)#%>%
   # head(10) #%>%
   # tail(5)
```

A tibble: 16 × 3

carrier	name	Total_of_carrier
<chr>	<chr>	<int>
UA	United Air Lines Inc.	58665
B6	JetBlue Airways	54635
EV	ExpressJet Airlines Inc.	54173
DL	Delta Air Lines Inc.	48110
AA	American Airlines Inc.	32729
MQ	Envoy Air	26397
US	US Airways Inc.	20536
9E	Endeavor Air Inc.	18460
WN	Southwest Airlines Co.	12275
VX	Virgin America	5162
FL	AirTran Airways Corporation	3260
AS	Alaska Airlines Inc.	714
F9	Frontier Airlines Inc.	685
YV	Mesa Airlines Inc.	601
HA	Hawaiian Airlines Inc.	342
OO	SkyWest Airlines Inc.	32

A2 : The top 5 of flights are UA, B6, EV, DL, AA and the last 5 of fights are MQ,US,9E,WN and VX.

Q3: Find avg,min,max,sd for dep_delay and arr_delay?

```
flights %>%
       filter(dep_delay > 0,arr_delay >0) %>%
       summarise(avg_dep_delay = mean(dep_delay),
                 min_dep_delay = min(dep_delay),
                 max_dep_delay = max(dep_delay),
                 sd_dep_delay = sd(dep_delay),
                 avg_arr_delay = mean(arr_delay),
                 min_arr_delay = min(arr_delay),
                 max_arr_delay = max(arr_delay),
                 sd_arr_delay = sd(arr_delay))
```

A tibble: 1 × 8

avg_dep_delay	min_dep_delay	max_dep_delay	sd_dep_delay	avg_arr_delay	min_arr_delay	max_arr_dela
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
51.43601	1	1301	59.17872	52.43045	1	1272

A3 :

The result for dep_delay : average = 51.43601 ,min =1, max = 1301,sd = 59.17872 and

The result for arr_delay : average = 52.43045 ,min =1, max = 1272,sd = 59.4824.

Q4.How many times is each airline that dep_delay higher than the average?

```
flights %>%
       select(dep_delay,carrier) %>%
       mutate(dep_delay_Check = if_else(dep_delay >= 51.43601 , "NG", "Good")) %
       filter(dep_delay_Check == "NG") %>%
       group_by(carrier) %>%
       summarise(Total = n())%>%
       arrange(desc(Total))
```

A tibble: 16 × 2

carrier	Total
<chr>	<int>
EV	8022
B6	5403
UA	4654
DL	3133
MQ	2411
AA	2339
9E	2269
WN	1268
US	933
VX	408
FL	363
YV	87
F9	84
AS	46
HA	13
OO	4

Q5 : Which month has the lowest number of flights in the top 5?

```
flights %>%
    group_by(month)%>%
    summarise(Total_of_flight_each_month = n()) %>%
    arrange(Total_of_flight_each_month)%>%
    head(5)
```

A tibble: 5 × 2

month	Total_of_flight_each_month
<dbl>	<int>
2	24951
1	27004
11	27268
9	27574
12	28135

A5 : the lowest number of flights are Feb, Jan, Nov, Sep, Dec.

Q6 : Where is the destination that people are popular to go?

```
flights %>%
    group_by(dest)%>%
    summarise(Total_of_orginal = n()) %>%
    arrange(desc(Total_of_orginal)) %>%
    head(10)
```

A tibble: 10 × 2

dest	Total_of_orginal
<chr>	<int>
ORD	17283
ATL	17215
LAX	16174
BOS	15508
MCO	14082
CLT	14064
SFO	13331
FLL	12055
MIA	11728
DCA	9705

A6 : The popular destination top 5 is ORD, ATL ,LAX, BOS ,MCO.