NYC flights 2013 Analysis

```
# import
library(tidyverse)
flights <- read_csv("flights.csv")
df <- flights
```

```
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()

Rows: 336776 Columns: 19

── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
```

```
glimpse(df)
```

```
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,
$ 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,
```

Q1: Top 10 airlines in May 2013

```
df %>%
    filter(month == 5, year == 2013) %>%
    count(carrier) %>%
    arrange(desc(n)) %>%
    head(10)
```

A tibble: 10 × 2

carrier	n
<chr>	<int>
UA	4960
EV	4817
B6	4576
DL	4082
AA	2803
MQ	2284
US	1785
9E	1462
WN	1006
VX	496

Q2: Average distance of each airline in Q1 2013

```
df %>%
    group_by(carrier) %>%
    filter(between(month, 1, 3), year == 2013) %>%
    summarise(mean(distance))
```

A tibble: 16 × 2

carrier	mean(distance)
<chr>	<dbl>
9E	473.7514
AA	1349.6699
AS	2402.0000
B6	1060.7052
DL	1232.8168
EV	528.8364
F9	1620.0000
FL	684.7755
HA	4983.0000
MQ	565.0895
OO	733.0000
UA	1451.3840
US	538.7930
VX	2493.2820
WN	956.9239
YV	229.0000

Q3: Average flight delay of each carrier in September 2013

```
df %>%
    filter(month == 9, year == 2013) %>%
    group_by(carrier) %>%
    summarise(avg_dep_delay = mean(dep_delay, na.rm = TRUE),
              avg_arr_delay = mean(arr_delay, na.rm = TRUE))
```

A tibble: 16 × 3

carrier	avg_dep_delay	avg_arr_delay
<chr>	<dbl>	<dbl>
9E	7.754232	-7.143052
AA	5.694272	-8.573988
AS	-4.516667	-34.900000
B6	6.634260	-2.514520
DL	5.526071	-4.414533
EV	8.237970	-1.870712
F9	8.263158	2.285714
FL	16.948819	15.071146
HA	-5.440000	-19.400000
MQ	5.350545	1.562560
OO	-4.941176	-8.882353
UA	6.890823	-7.668464
US	1.962583	-4.721113
VX	6.988962	-6.577434
WN	14.166832	2.243756
YV	8.880952	-2.785714

Q4: Which carrier had longest distance in Q2 2013 ?

```
df %>%
    filter(between(month, 4, 6), year == 2013) %>%
    group_by(carrier) %>%
    summarise(max_distance = max(distance)) %>%
    arrange(desc(max_distance))
```

A tibble: 16 × 2

carrier	max_distance
<chr>	<dbl>
HA	4983
UA	4963
AA	2586
B6	2586
DL	2586
VX	2586
AS	2402
US	2153
WN	2133
F9	1620
9E	1391
EV	1389
MQ	1147
FL	762
YV	544
OO	488

Q5: Which Destination had most flights in May 2013 ?

```
df %>%
    filter(month == 5, year == 2013) %>%
    count(dest) %>%
    arrange(desc(n)) %>%
    head(10)
```

A tibble: 10 × 2

dest	n
<chr>	<int>
ORD	1582
ATL	1499
LAX	1453
BOS	1327
SFO	1198
CLT	1167
MCO	1125
MIA	991
FLL	919
DCA	882

Q6: What is maximum arrived delay in 2013, which carrier and when ?

```
df %>%
    select(day, month, year, carrier, arr_delay ) %>%
    group_by(carrier) %>%
    arrange(desc(arr_delay)) %>%
    head(10)
```

A grouped_df: 10 × 5

day	month	year	carrier	arr_delay
<dbl>	<dbl>	<dbl>	<chr>	<dbl>
9	1	2013	HA	1272
15	6	2013	MQ	1127
10	1	2013	MQ	1109
20	9	2013	AA	1007
22	7	2013	MQ	989
10	4	2013	DL	931
17	3	2013	DL	915
22	7	2013	DL	895
5	12	2013	AA	878
3	5	2013	MQ	875

Q7: Top 3 of the month with the most of flights

```
df %>%
    count(month) %>%
    arrange(desc(n)) %>%
    head(3)
```

A tibble: 3 × 2

month	n
<dbl>	<int>
7	29425
8	29327
10	28889

Proportion between delay and early arrival flight in 2013.

```
flights %>%
  mutate( arrival_type = case_when(arr_delay > 0 ~ "delay",
                                   arr_delay < 0 ~ "early",
                                   arr_delay == 0 ~ "on time")) %>%
  count(arrival_type) %>%
  mutate(percent = n*100/sum(n)) %>%
  arrange(desc(percent))
```

A spec_tbl_df: 4 × 3

arrival_type	n	percent
<chr>	<int>	<dbl>
early	188933	56.100494
delay	133004	39.493313
NA	9430	2.800081
on time	5409	1.606112