NYC Flights Analysis 2013

```
library(dplyr)
library(readr)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

```
flights <- read_csv("result_flights.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 messag
```

```
glimpse(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, 5, 6, 6, 6
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0
```

Q1: Which top 3 of destination is the most flights in March 2013 ?

```
flights %>%
  filter(month==3, year==2013) %>%
  count(dest) %>%
  arrange(desc(n)) %>%
  head(3)
```

A tibble: 3 × 2

dest	n
<chr>	<int>
ATL	1448
ORD	1343
BOS	1324

Q2: Which carrier is the most distance in Jan 2013 ?

```
flights %>%
    group_by(carrier) %>%
    filter(month==1, year==2013) %>%
    summarise(sum_distance = sum(distance)) %>%
    arrange(desc(sum_distance)) %>%
    head(10)
```

A tibble: 10 × 2

carrier	sum_distance
<chr>	<dbl>
UA	6777189
B6	4699834
DL	4503241
AA	3773186
EV	2178833
MQ	1284653
WN	938403
US	858820
VX	788439
9E	749305

Q3: Which month had the most departure delay (hr) in 2013 ?

```
flights %>%
    group_by(month) %>%
    filter(!is.na(dep_delay),dep_delay > 0) %>%
    summarise(sum_dep_mins = sum(dep_delay)) %>%
    mutate(sum_delay_hr = round(sum_dep_mins/60,2)) %>%
    arrange(desc(sum_dep_mins)) %>%
    select(month, sum_delay_hr) %>%
    head(5)
```

A tibble: 5 × 2

month	sum_delay_hr
<dbl>	<dbl>
7	11314.47
6	10501.73
12	8401.78
4	7764.08
3	7401.00

Q4 : Which top 3 of the destination are the eariest departure in June 2013 ?

```
flights %>%
    filter(!is.na(origin),!is.na(dest),!is.na(dep_delay),dep_delay < 0,month==6,
    group_by(dest) %>%
    summarise(sum_early_hr = round(sum((dep_delay)*(-1))/60,2)) %>%
    arrange(desc(sum_early_hr)) %>%
    head(3)
```

A tibble: 3 × 2

dest	sum_early_hr
<chr>	<dbl>
BOS	66.68
CLT	56.95
ORD	55.27

Q5: Which carrier had the lowest avg amount of time spent in the air ?

```
flights %>%
    filter(!is.na(carrier), !is.na(air_time)) %>%
    group_by(carrier) %>%
    summarise(avg_air_time = round(mean(air_time),2))%>%
    arrange(avg_air_time) %>%
    head(5)
```

A tibble: 5 × 2

carrier	avg_air_time
<chr>	<dbl>
YV	65.74
OO	83.48
9E	86.78
US	88.57
EV	90.08