Data Analysis Flights NYC 2013 

```
library(dplyr)
library(tidyverse)
```

```
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


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 ──
```

Import Data

```
flights <- read.csv("flights.csv")
```

Data preparation

```
glimpse(flights)
```

```
Rows: 336,776
Columns: 19
$ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2…
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, …
$ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, …
$ dep_delay      <int> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1…
$ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,…
$ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,…
$ arr_delay      <int> 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         <int> 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       <int> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1…
$ distance       <int> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, …
$ hour           <int> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6…
$ minute         <int> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0…
```

Finding the percentage of missing values in flight

- complete value = 97.2%
- missing value = 2.8%

```
sum(complete.cases(flights))/nrow(flights)
```

0.971999192341497

Drop missing value

- If the missing value is less than 5%, it's acceptable to drop it.

```
clean_df <- drop_na(flights)
clean_df %>% head(3)
```

A data.frame: 3 × 19

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	ori
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<int>	<chr>	<c
1	2013	1	1	517	515	2	830	819	11	UA	1545	N14228	EW
2	2013	1	1	533	529	4	850	830	20	UA	1714	N24211	LG
3	2013	1	1	542	540	2	923	850	33	AA	1141	N619AA	JFK

Data Analysis 

Q1: What were the top 5 months that had the highest number of delayed arrivals and departure?

```
delay_m_a <- clean_df %>%
group_by(month) %>%
select(month, arr_delay) %>%
filter(arr_delay > 0) %>%
summarize(num_delay = n()) %>%
arrange(desc(num_delay))

delay_m_a %>% head(5)
```

A tibble: 5 × 2

month	num_delay
<int>	<int>
12	14394
7	13304
4	12522
6	12490
8	11629

```
delay_m_d <- clean_df %>%
group_by(month) %>%
select(month, dep_delay) %>%
filter(dep_delay > 0) %>%
summarize(num_delay = n()) %>%
arrange(desc(num_delay))

delay_m_d %>% head(5)
```

A tibble: 5 × 2

month	num_delay
<int>	<int>
7	13773
12	13490
6	12558
8	11665
5	11227

Q2: What were the top 5 destination in December

```
top_d_dec <- clean_df %>%
filter(month == 12, year == 2013) %>%
count(dest) %>%
arrange(desc(n))

top_d_dec %>% head(5)
```

A data.frame: 5 × 2

	dest	n
	<chr>	<int>
1	ATL	1429
2	LAX	1390
3	MCO	1203
4	SFO	1159
5	CLT	1155

Q3: What were the top 5 carrier that had the lowest air time?

```
low_at <- clean_df %>% group_by(carrier) %>%
select(carrier, air_time) %>%
summarize(air_time = n()) %>%
arrange(air_time)

low_at %>% head(5)
```

A tibble: 5 × 2

carrier	air_time
<chr>	<int>
OO	29
HA	342
YV	544
F9	681
AS	709

Q4:Which origin airport had the least amount of arrival "delay" time in October?

```
air_l_delay <- clean_df %>% filter(month == 10, year == 2013,) %>%
group_by(origin)  %>%
select(origin, arr_delay) %>%
summarise(arr_delay =n()) %>%
arrange(arr_delay)

air_l_delay %>% head(3)
```

A tibble: 3 × 2

origin	arr_delay
<chr>	<int>
JFK	9096
LGA	9516
EWR	10006

Q5:What is the average and median of delay departure time for each carrier between June and July

2013?

```
avg_depdelayinJune <- clean_df %>%
filter(between(month,6,7), year == 2013,dep_delay > 0) %>%
group_by(carrier) %>%
summarize(mean_depdealy = mean(dep_delay),
median_depdealy = median(dep_delay)) %>%
arrange(desc(mean_depdealy))
```

```
avg_depdelayinJune %>% head (16)
```

A tibble: 16 × 3

carrier	mean_depdealy	median_depdealy
<chr>	<dbl>	<dbl>
OO	131.00000	131
FL	63.77044	25
YV	62.61404	39
9E	58.58470	34
EV	57.67910	35
VX	56.40036	16
MQ	55.79965	35
B6	52.03178	30
DL	47.80134	22
US	44.90656	24
AA	44.71014	20
WN	44.66869	20
F9	41.06897	25
UA	38.46728	17
HA	33.62500	8
AS	31.37778	8