NYC Flights Analysis

Import library

```
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 dataset: NYCFLIGHT2013 (CSV FILE)

```
flights <- read.csv("flights.csv")
airlines <- read.csv("airlines.csv")
airports <- read.csv("airports.csv")
planes <- read.csv("planes.csv")
weathers <- read.csv("weather.csv")
```

Data preparation

Overview Data

```
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, 59, 0
```

Check Missing Values

- completed data = 97% missing values = 3%

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

0.971999192341497

Clean data (remove missing value)

```
flights_clean <- drop_na(flights)
flights_clean %>%
    head(5)
```

A data.frame: 5 × 19

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<int>
1	2013	1	1	517	515	2	830	819	11	UA	1545
2	2013	1	1	533	529	4	850	830	20	UA	1714
3	2013	1	1	542	540	2	923	850	33	AA	1141
4	2013	1	1	544	545	-1	1004	1022	-18	B6	725
5	2013	1	1	554	600	-6	812	837	-25	DL	461

Data analysis

Q1: Which top 5 airlines had the highest number of delayed departures?

```
most_delayed <- flights_clean %>%
    group_by(carrier) %>%
    filter(dep_delay > 0) %>%
    summarize(num_delay = n()) %>%
    left_join(airlines ,by = "carrier") %>%
    select(airline_name = name,num_delay) %>%
    arrange(desc(num_delay))

most_delayed %>%
    head(5)
```

A tibble: 5 × 2

airline_name	num_delay
<chr>	<int>
United Air Lines Inc.	27125
ExpressJet Airlines Inc.	22976
JetBlue Airways	21372
Delta Air Lines Inc.	15186
American Airlines Inc.	10105

Q2 : How relative between flights and departure delay flights?

```
total_flights <- flights_clean %>%
    group_by(carrier) %>%
    summarize(num_flights = n()) %>%
    left_join(airlines ,by = "carrier") %>%
    arrange(desc(num_flights)) %>%
    select(airline_name = name,num_flights)

total_flights %>%
    head(5)
```

A tibble: 5 × 2

airline_name	num_flights
<chr>	<int>
United Air Lines Inc.	57782
JetBlue Airways	54049
ExpressJet Airlines Inc.	51108
Delta Air Lines Inc.	47658
American Airlines Inc.	31947

Q3: Top 5 best performance airlines ranked by flight delay ratio

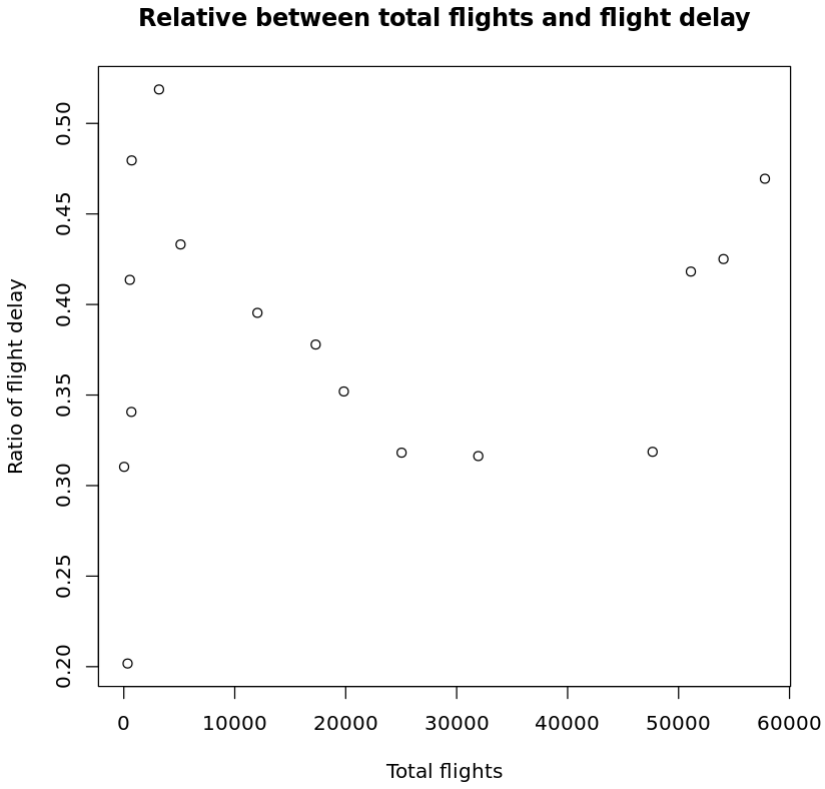
```
relative <- total_flights %>%
    left_join(most_delayed, by = "airline_name") %>%
    mutate(ratio = most_delayed$num_delay / total_flights$num_flights)

relative %>%
    arrange(ratio) %>%
    head(5)
```

A tibble: 5 × 4

airline_name	num_flights	num_delay	ratio
<chr>	<int>	<int>	<dbl>
Hawaiian Airlines Inc.	342	69	0.2017544
SkyWest Airlines Inc.	29	9	0.3103448
American Airlines Inc.	31947	10105	0.3163051
Envoy Air	25037	7966	0.3181691
Delta Air Lines Inc.	47658	15186	0.3186453

```
plot(relative$num_flights,relative$ratio,
    main = "Relative between total flights and flight delay",
    xlab = "Total flights",
    ylab = "Ratio of flight delay")
```

Download



Q4: On average, to which airport do flights arrive most early?

```
flights_clean %>%
    group_by(dest) %>%
    summarize(avg_arr_early = mean(arr_delay)) %>%
    left_join(airports, by = c("dest" = "faa")) %>%
    select(dest,name,avg_arr_early) %>%
    arrange(avg_arr_early) %>%
    head(5)
```

A tibble: 5 × 3

dest	name	avg_arr_early
<chr>	<chr>	<dbl>
LEX	Blue Grass	-22.000000
PSP	Palm Springs Intl	-12.722222
SNA	John Wayne Arpt Orange Co	-7.868227
STT	NA	-3.835907
ANC	Ted Stevens Anchorage Intl	-2.500000

Q5: In which month do flights tend to have the longest delays?

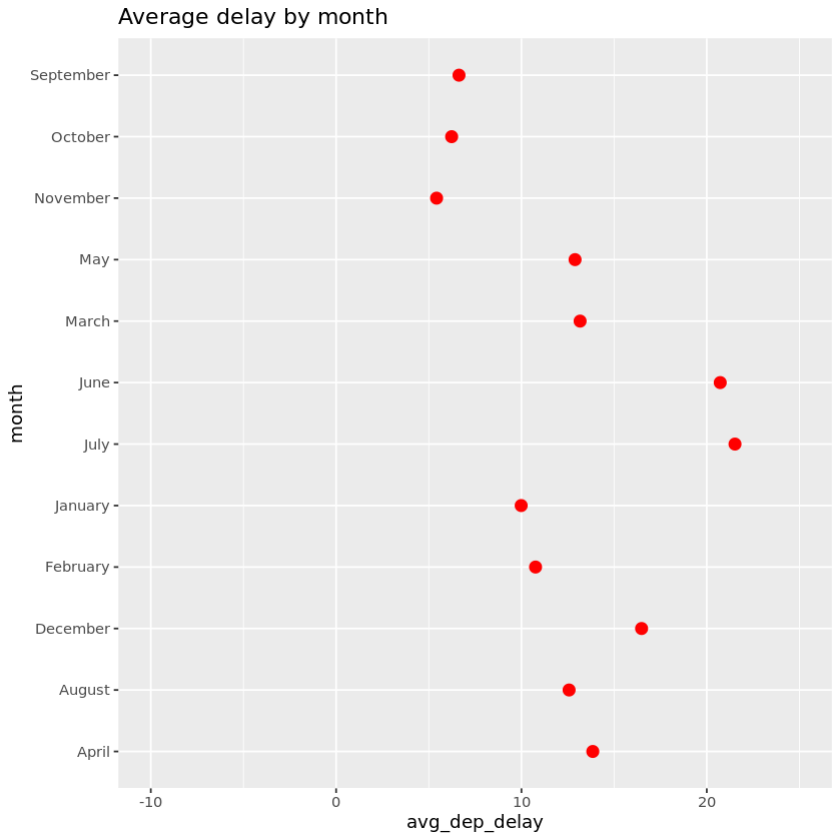
```
delay_by_month <- flights_clean %>%
    group_by(month) %>%
    summarize(avg_dep_delay =  mean(dep_delay)) %>%
    mutate(month = month.name)

delay_by_month %>%
    filter(avg_dep_delay == max(avg_dep_delay))
```

A tibble: 1 × 2

month	avg_dep_delay
<chr>	<dbl>
July	21.52218

```
ggplot(data = delay_by_month) +
    geom_point(aes(x = avg_dep_delay, y = month), color = "red" , size = 3) +
    labs(title = "Average delay by month")+
    xlim(-10,25)
```

Download



Q6: Which Top 5 popular destinations in December 2013

```
flights_clean %>%
    filter(month == 12) %>%
    group_by(dest) %>%
    summarise(flights = n()) %>%
    left_join(airports, by = c("dest" = "faa")) %>%
    select(dest, name, flights) %>%
    arrange(desc(flights)) %>%
    head(5)
```

A tibble: 5 × 3

dest	name	flights
<chr>	<chr>	<int>
ATL	Hartsfield Jackson Atlanta Intl	1429
LAX	Los Angeles Intl	1390
MCO	Orlando Intl	1203
SFO	San Francisco Intl	1159
CLT	Charlotte Douglas Intl	1155