R Notebook

```
library(ggthemes)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

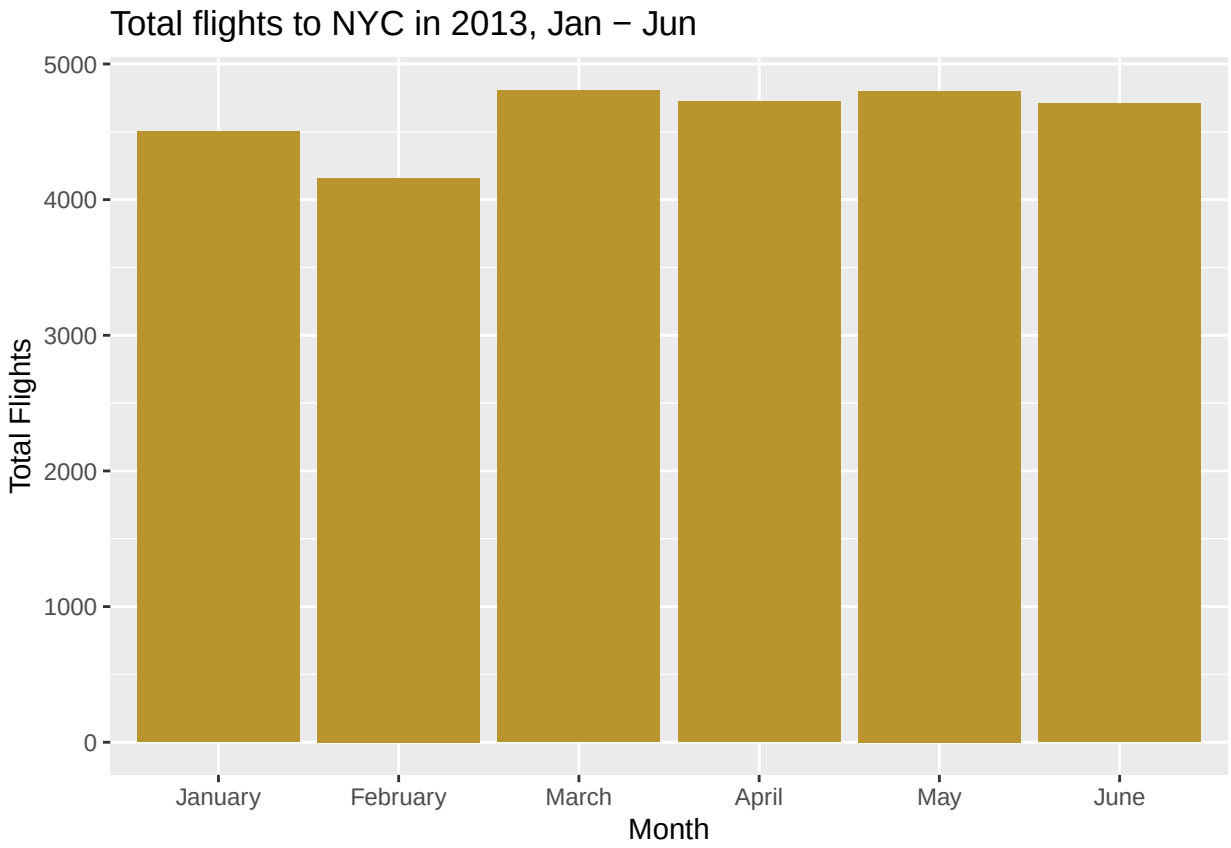
```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(nycflights13)
library(dplyr)
library(ggplot2)
data("flights")
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, 1~
## $ day            <int> 1, 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      <dbl> 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      <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         <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       <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, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
```

```
flights %>%
  filter(month == c(1:6)) %>%
  mutate(full_month = month(month, label = T, abbr = F)) %>%
  ggplot(aes(full_month)) +
  geom_bar(fill = "#ba952f") +
  labs(title = "Total flights to NYC in 2013, Jan - Jun",
       x = "Month",
       y = "Total Flights")
```

```
## Warning in month == c(1:6): longer object length is not a multiple of shorter
## object length
```

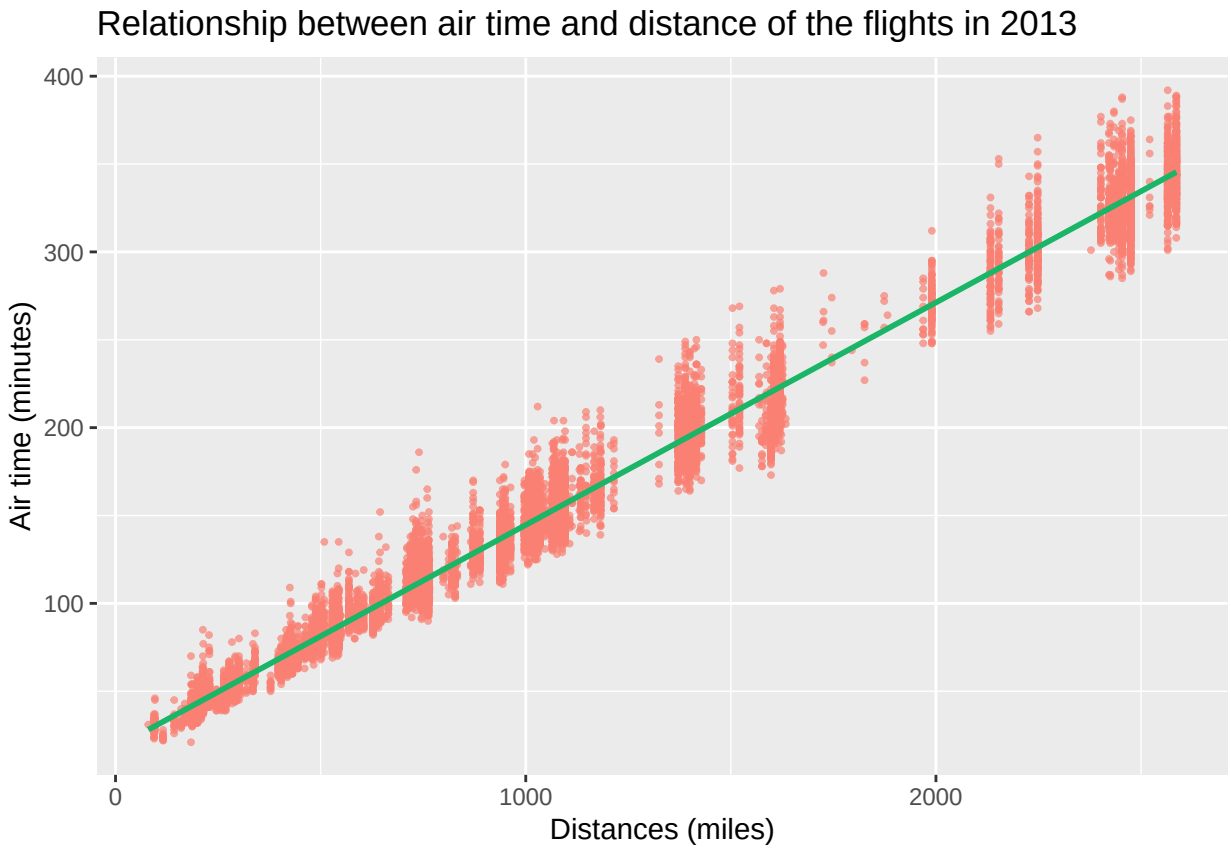


Total flights in March to June were slightly different, but in February was drop from January.

```
flights <- drop_na(flights)
f_clean <- flights %>%
  filter(distance <= 3000)

set.seed(11)
ggplot(sample_n(f_clean, size = 10000), aes(distance,
                                            air_time)) +
  geom_point(color = "Salmon",
             size = 0.7,
             alpha = 0.7) +
  geom_smooth(method="lm",
              color = "#1bb567") +
  labs(title = "Relationship between air time and distance of the flights in 2013",
       x = "Distances (miles)",
```

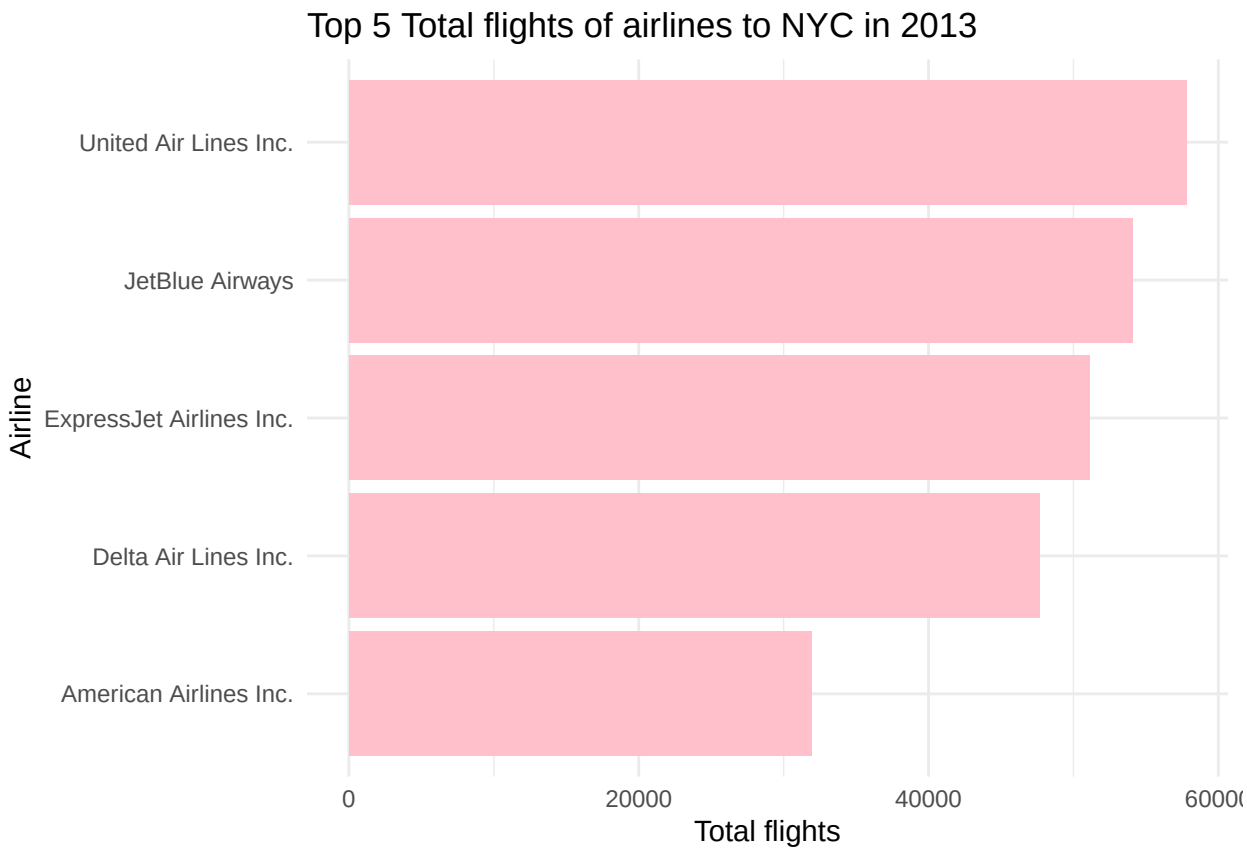
```
       y = "Air time (minutes)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



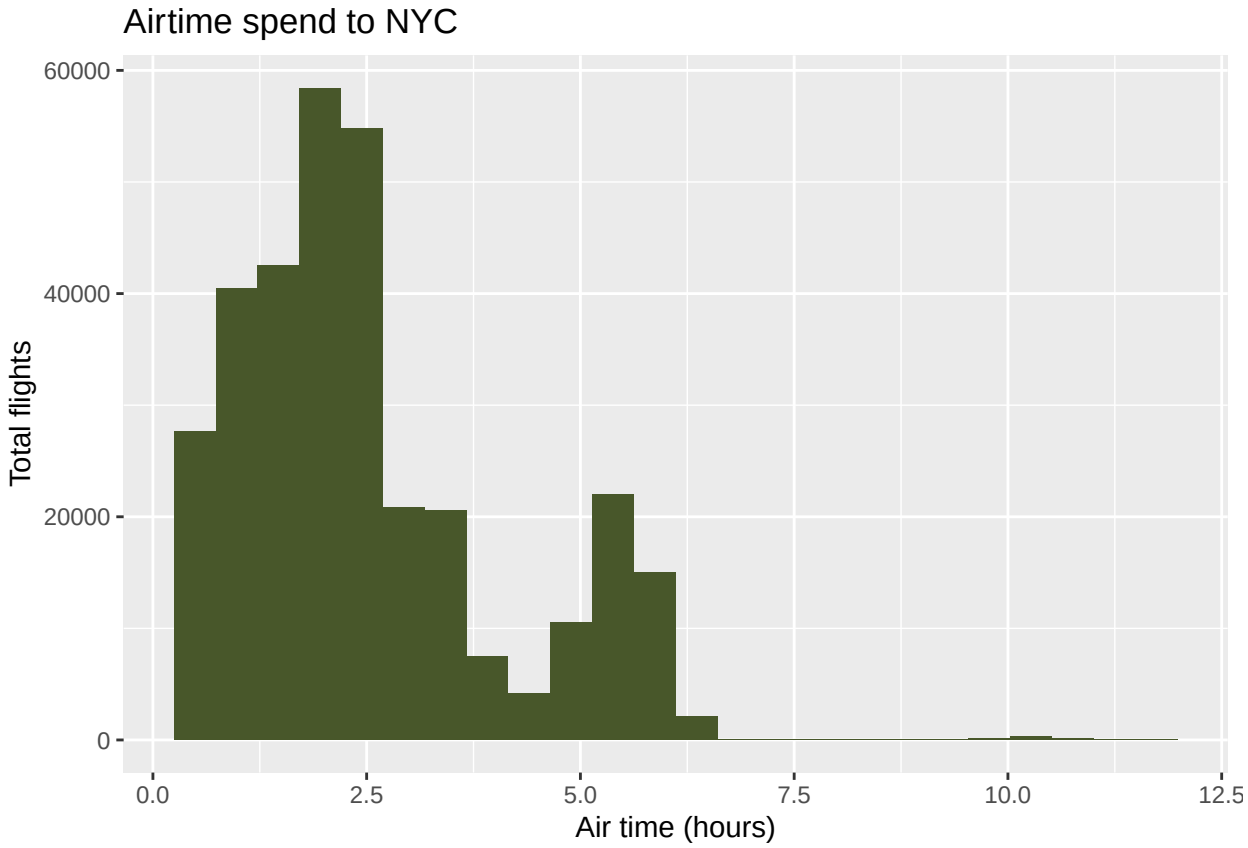
At the trend for the graph, it can be seen that air time were increased when distances were increased.

```
f <- flights %>%
    group_by(carrier) %>%
    count(carrier) %>%
    arrange(desc(n)) %>%
    left_join(airlines, by = "carrier") %>%
    rename(airline = carrier,
           total_flights = n) %>%
    head(5)

ggplot(f, aes(name, total_flights)) +
  geom_col(fill = "pink") +
  coord_flip() +
  labs(title = "Top 5 Total flights of airlines to NYC in 2013",
       x = "Airline",
       y = "Total flights") +
  theme_minimal()
```



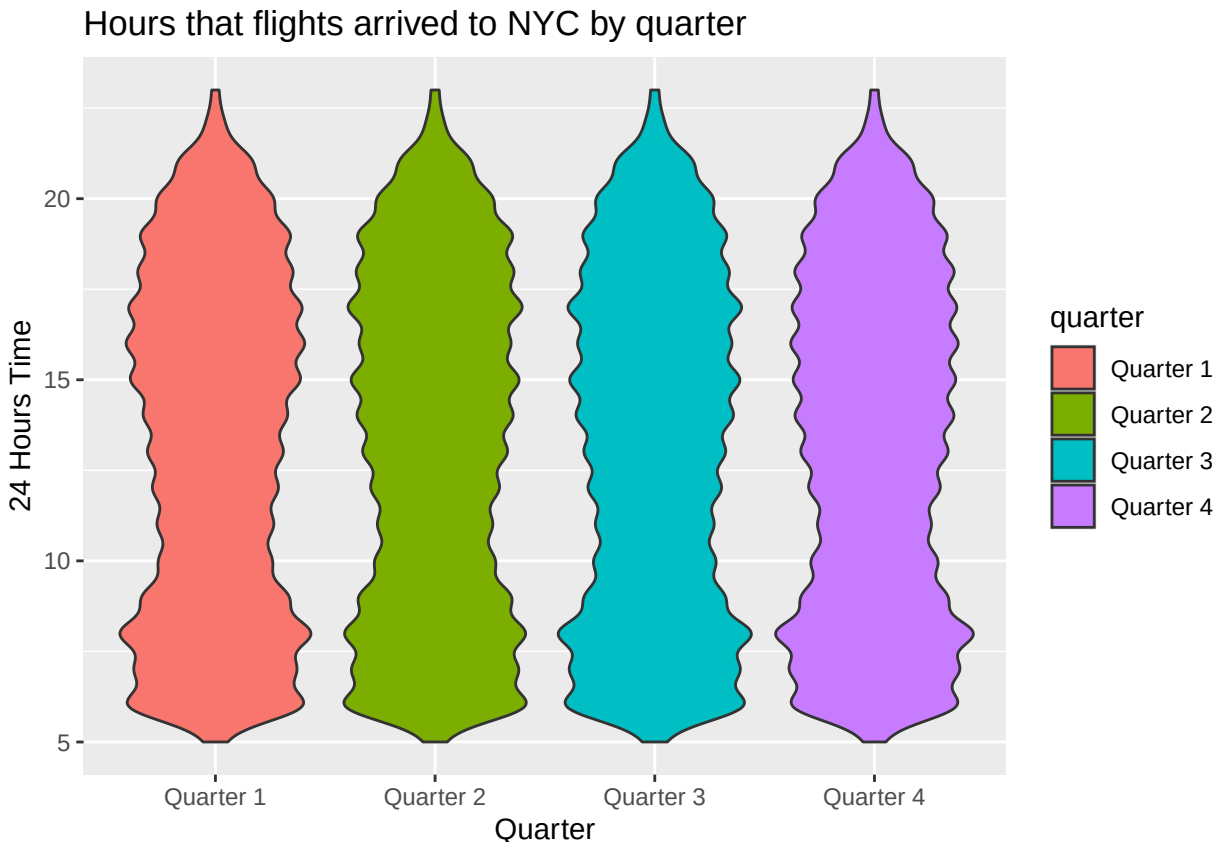
United Air Lines Inc had the most flights to NYC in 2013.

```
flights <- flights %>%
  mutate(air_time_hour = air_time/60)

ggplot(flights, aes(air_time_hour), color = air_time_hour) +
  geom_histogram(fill = "#48572a",
                 bins=24) +
  labs(title = "Airtime spend to NYC",
       x = "Air time (hours)",
       y = "Total flights")
```



The most airlines had air time to NYC less than 2.5 hours.

```
flights <- flights %>%
    mutate(quarter = case_when(
                month %in% c(1,2,3) ~ "Quarter 1",
                month %in% c(4,5,6) ~ "Quarter 2",
                month %in% c(7,8,9) ~ "Quarter 3",
                month %in% c(10,11,12) ~ "Quarter 4"
    ))


ggplot(flights, aes(quarter, hour, fill = quarter)) +
  geom_violin() +
  labs(title = "Hours that flights arrived to NYC by quarter",
       x = "Quarter",
       y = "24 Hours Time")
```



The flights's time arrived to NYC in each quarter were slightly different.