# Flights Data Visualization

2024-07-28

### Show data set

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(nycflights13)
library(glue)
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # i 336,766 more rows
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
airlines
```

```
## # A tibble: 16 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
```

```
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
## 11 OO      SkyWest Airlines Inc.
## 12 UA      United Air Lines Inc.
## 13 US      US Airways Inc.
## 14 VX      Virgin America
## 15 WN      Southwest Airlines Co.
## 16 YV      Mesa Airlines Inc.
```

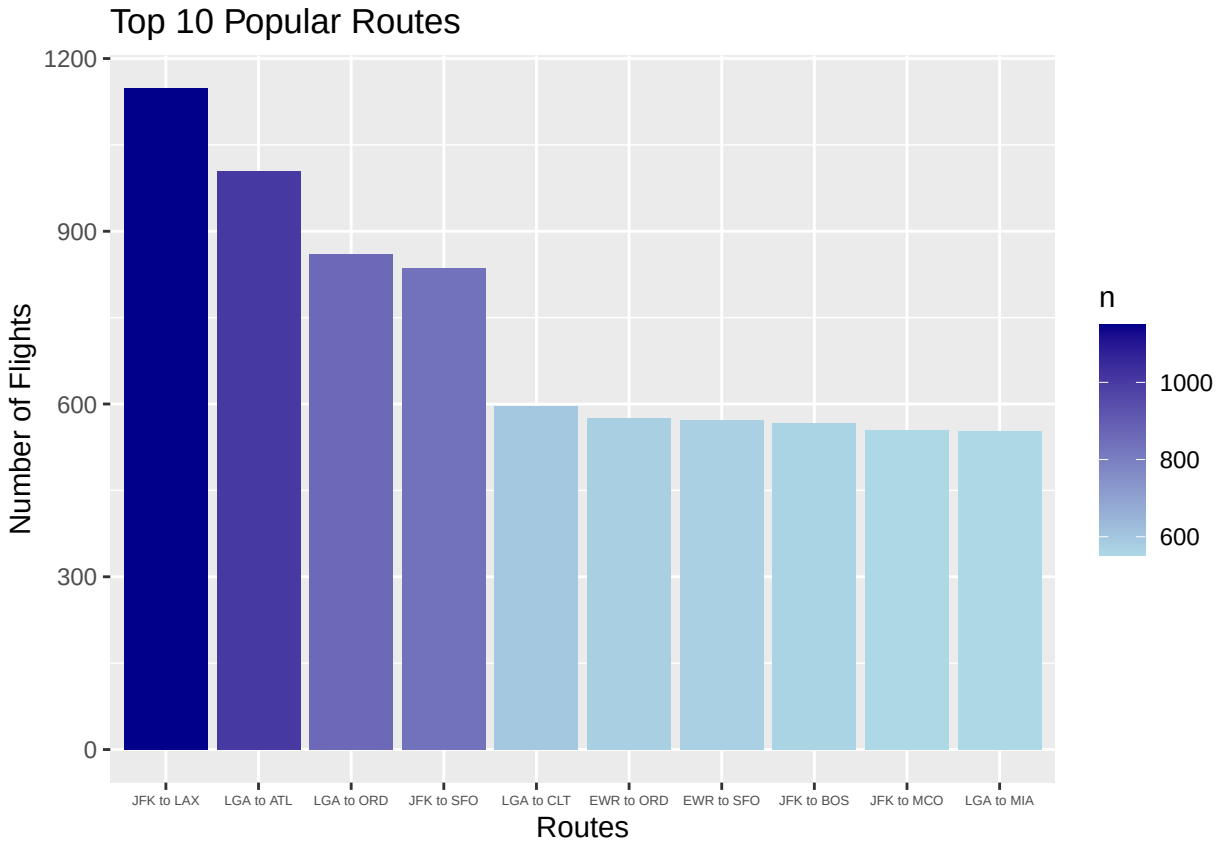
### Data Preparation

```
flights = flights %>%
  sample_frac(0.1) %>%
  filter(!rowSums(is.na(.)))
```

## Top 10 Popular Routes (show orgin and destination)

```
# Create a new data frame with only the necessary columns
flights_routes = flights %>%
  select(origin, dest)

# Count the frequency of each route and Select the top 10 routes
route_counts = flights_routes %>%
  count(origin, dest) %>%
  arrange(desc(n)) %>%
  head(10)

# Create new column "route" showing origin and destination
route = route_counts %>%
  mutate(route = glue("{origin} to {dest}")) %>%
  arrange(desc(n))

# Convert route to factor for correct ordering
route$route = factor(route$route, levels = route$route)
```

```
ggplot(route, aes(route, n, fill = n)) +
  geom_col() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  theme(axis.text.x = element_text(size = 5)) +
  labs(title = "Top 10 Popular Routes", x = "Routes", y = "Number of Flights")
```



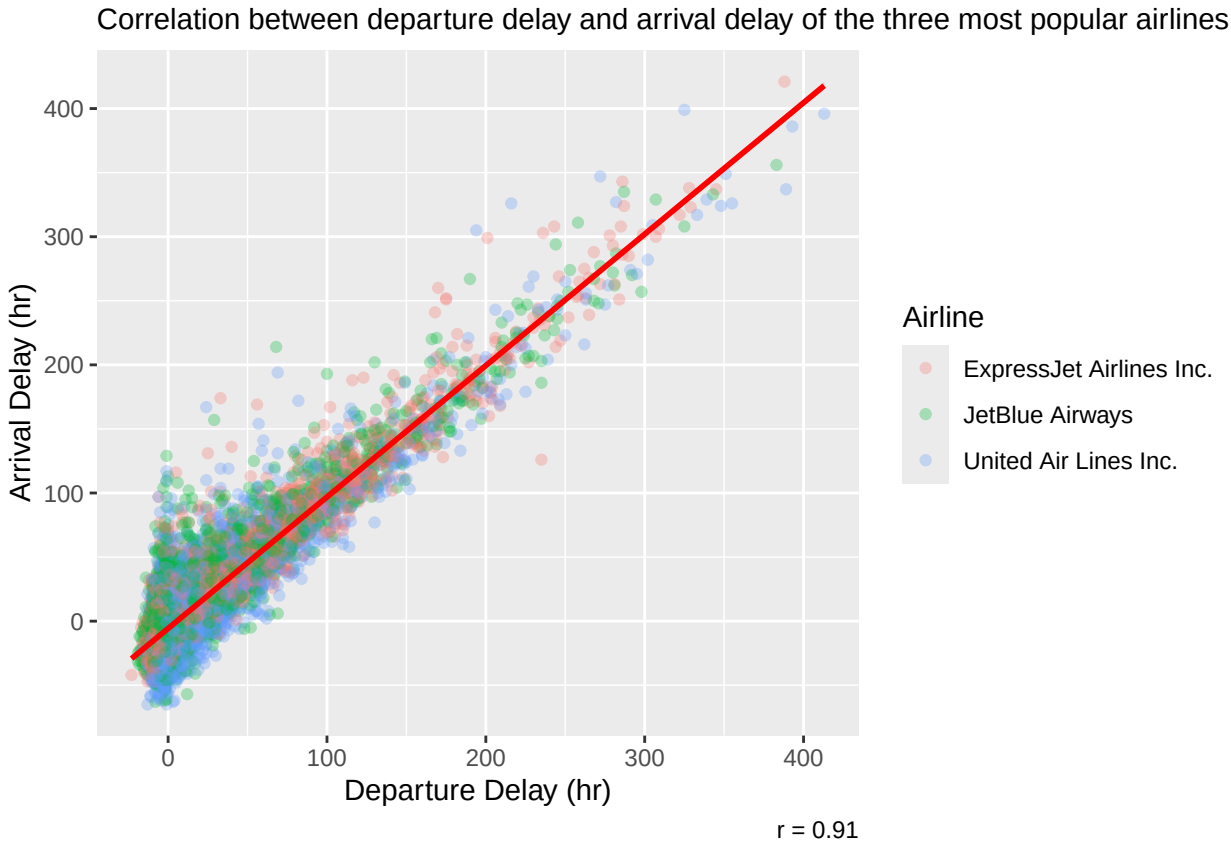

## Correlation between departure delay time and arrival delay time for the 3 airlines with highest number of flights

```
# Form a table with information from 'flights' and 'airlines' tables
flights_airlines = flights %>%
  left_join(airlines, by = 'carrier')

# Find 3 airlines with highest number of flight with the help of table 'airlines'
pop_airlines = flights_airlines %>%
  group_by(name) %>%
  count() %>%
  arrange(-n) %>%
  head(3)

# Select only rows of our three popular airlines and convert time to hours
delay_airlines = flights_airlines %>%
  filter(name %in% pop_airlines$name) %>%
  mutate(dep_delay == dep_delay/60 & arr_delay == arr_delay/60) %>%
  rename(Airline = name)
```

```
# Calculate correlation coefficient
correlation = cor(flights_airlines$dep_delay, flights_airlines$arr_delay)

ggplot(delay_airlines, aes(dep_delay,arr_delay,col=Airline)) +
  geom_point(alpha = 0.3) +
```

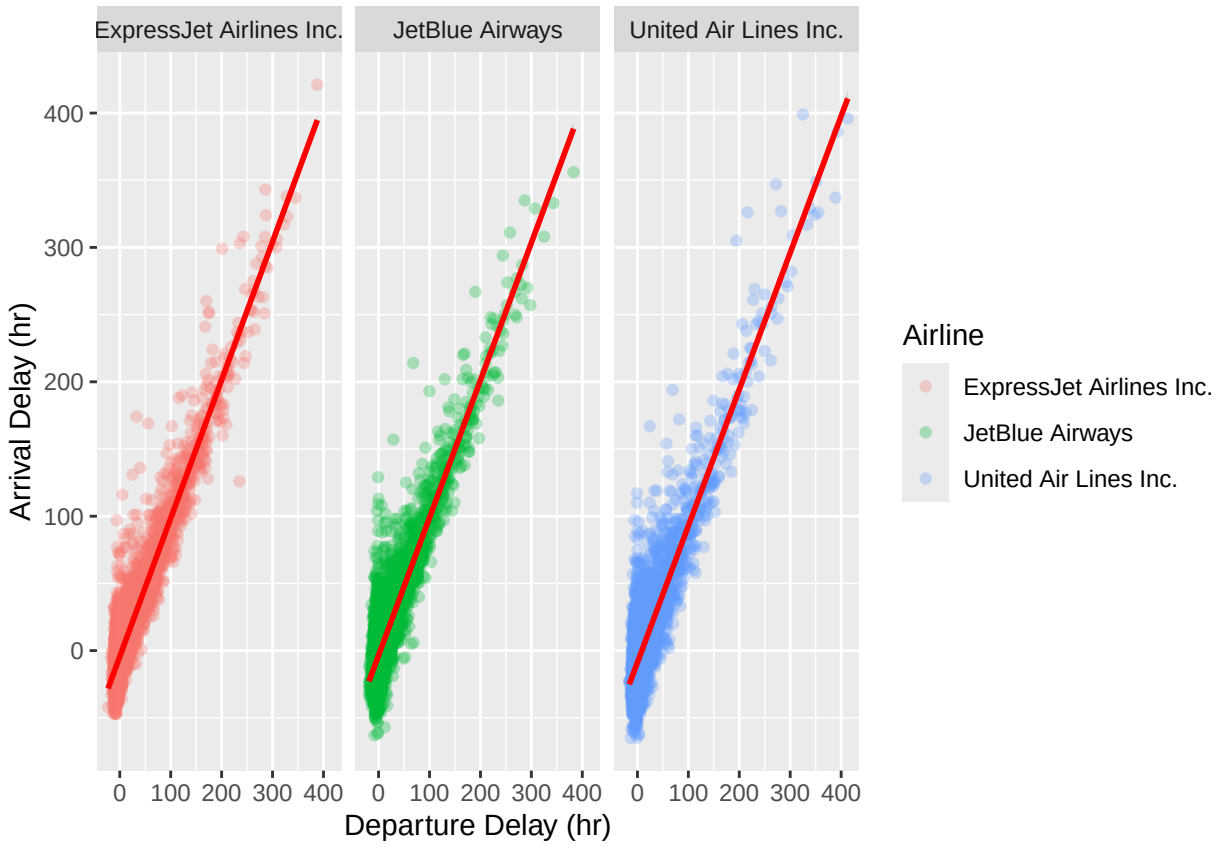
```
  geom_smooth(method = "lm", col = "red") +
  labs(subtitle = "Correlation between departure delay and arrival delay of the three most popular airli
```

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




```
ggplot(delay_airlines, aes(dep_delay,arr_delay,col=Airline)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", col = "red") +
  facet_wrap(~Airline) +
  labs(x = "Departure Delay (hr)", y = "Arrival Delay (hr)")
```

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


ExpressJet Airlines Inc.
JetBlue Airways
United Air Lines Inc.
Arrival Delay (hr)
Departure Delay (hr)
0
100
200
300
400
Airline
ExpressJet Airlines Inc.
JetBlue Airways
United Air Lines Inc.