homework_data_viz_dsb10_pornnapat

Pornnapat K.

2024-08-05

Install libraries

```
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(lubridate)
library(ggplot2)
```

List column names of flights

```
colnames(flights)
```

```
##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"
```

1. Which carries have the best on-time performance, and how does this vary by airport (origin)?

necessary: carrier, dep_delay, arr_delay, origin

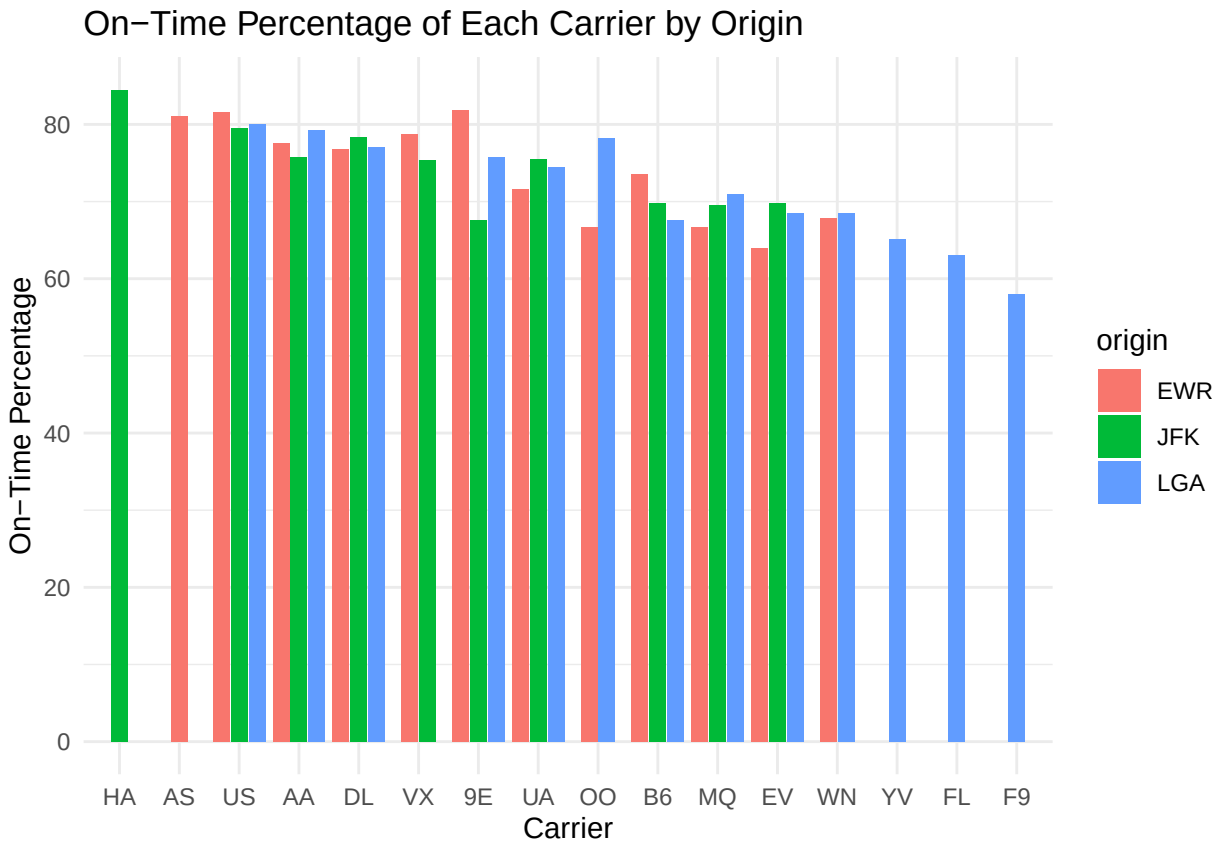
*Within 15 minutes after its schedule is considered on-time

```
on_time_perf_all <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay)) %>%
  mutate(on_time = (dep_delay <= 15 & arr_delay <= 15)) %>%
  group_by(carrier, origin) %>%
  summarize(on_time_percentage = mean(on_time) * 100) %>%
  arrange(desc(on_time_percentage))
```

```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```

```r
on_time_perf_carrier <- on_time_perf_all %>%
  group_by(carrier) %>%
  summarize(avg_on_time_percentage = mean(on_time_percentage))
```

```r
ggplot(on_time_perf_all, aes(fct_rev(fct_reorder(carrier, on_time_percentage)), on_time_percentage, fil
  geom_col(position = position_dodge2(preserve = "single")) +
  theme_minimal() +
  labs(title = "On-Time Percentage of Each Carrier by Origin",
       x = "Carrier",
       y = "On-Time Percentage")
```



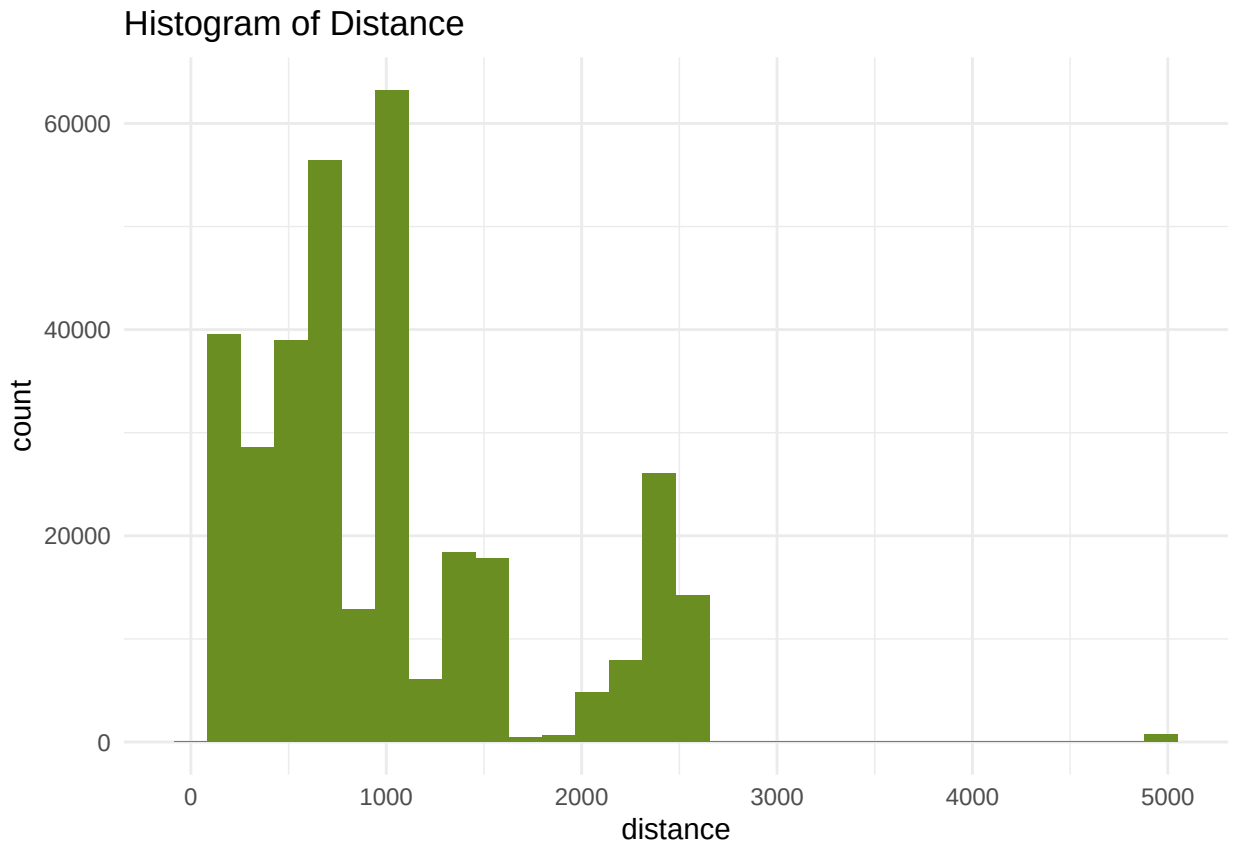
2. What is the distribution of flight distances, and how does this relate to flight delays?

necessary: distance, dep_delay, arr_delay

```r
summary(flights$distance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      17     502     872    1040    1389    4983
```

```r
ggplot(flights, aes(distance)) +
  geom_histogram(bins = 30, fill = "olivedrab") +
  theme_minimal() +
  labs(title = "Histogram of Distance")
```

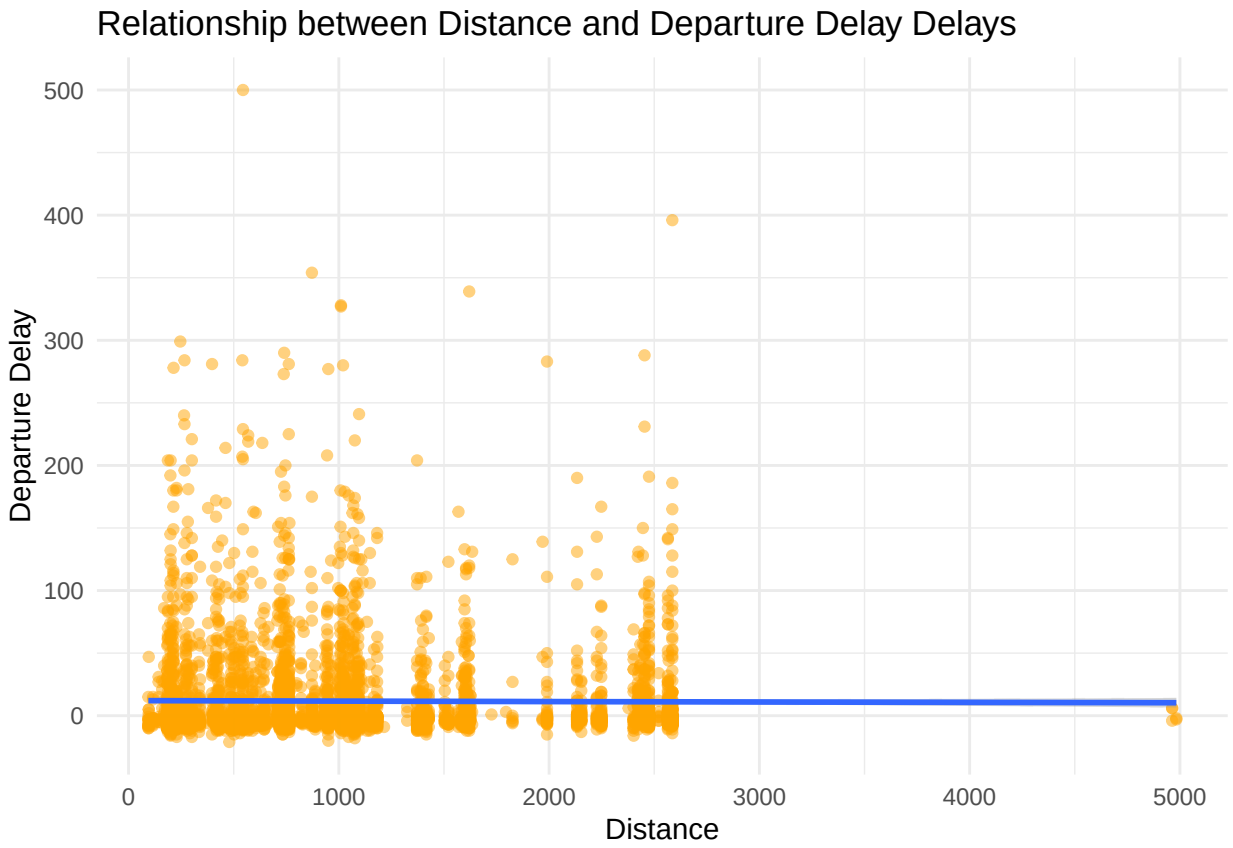



```
set.seed(38)
small_flights <- sample_n(flights, 5000)

ggplot(small_flights, aes(distance, dep_delay)) +
  geom_point(color = "orange", na.rm = TRUE, alpha=0.5) +
  geom_smooth() +
  theme_minimal() +
  labs(title = "Relationship between Distance and Departure Delay Delays",
       x = "Distance",
       y = "Departure Delay")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

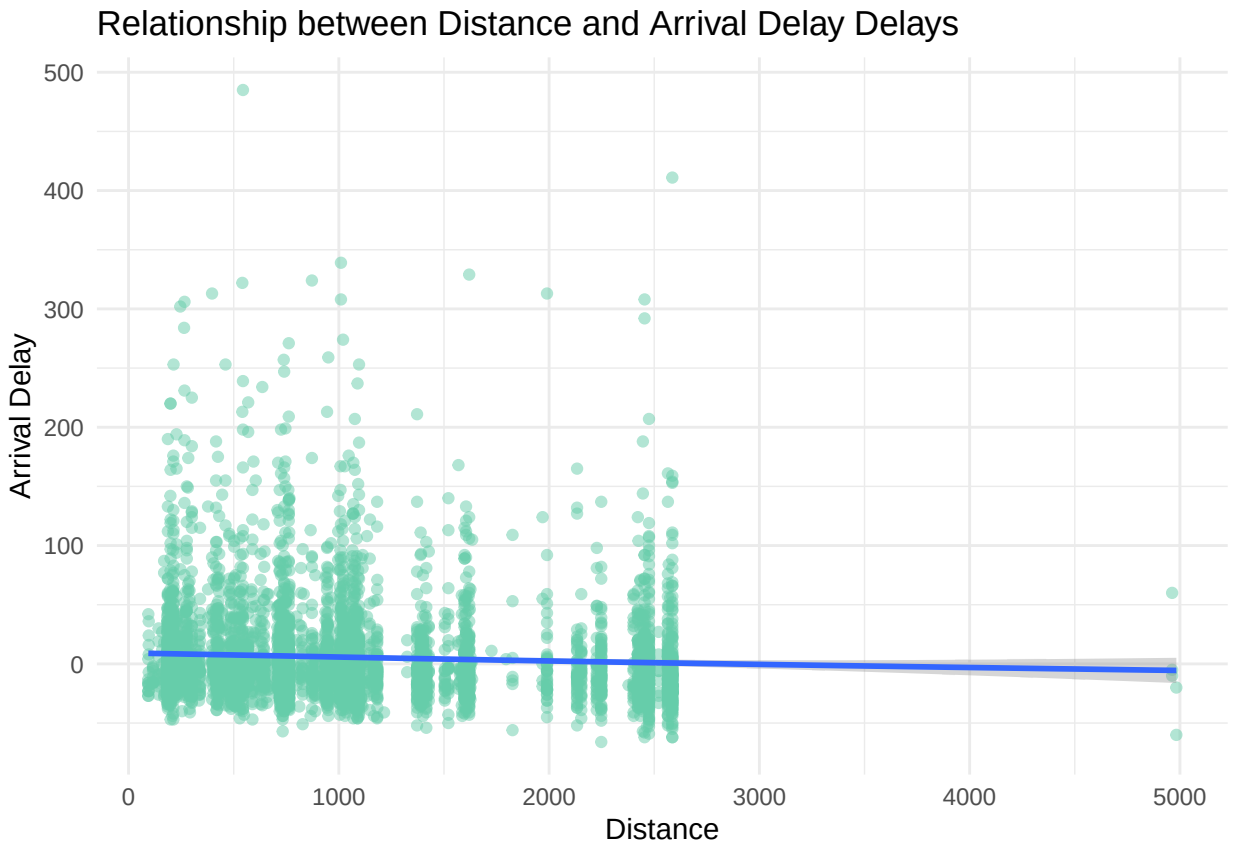
```
## Warning: Removed 112 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```




```
ggplot(small_flights, aes(distance, arr_delay)) +
  geom_point(color = "aquamarine3", na.rm = TRUE, alpha=0.5) +
  geom_smooth() +
  theme_minimal() +
  labs(title = "Relationship between Distance and Arrival Delay Delays",
       x = "Distance",
       y = "Arrival Delay")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 127 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



3. How do departure delays vary by time of day and day of week?

necessary: year, month, day, sched_dep_time, dep_delay

```
departure_delay <- flights %>%
  mutate(date = make_date(year, month, day),
         day_of_week = wday(date, label = TRUE),
         sched_dep_time_bin = cut(
          sched_dep_time,
          breaks = seq(0, 2400, by = 100),
          labels = seq(0, 23))
         ) %>%
  group_by(day_of_week, sched_dep_time_bin) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'day_of_week'. You can override using the
## `.groups` argument.
```

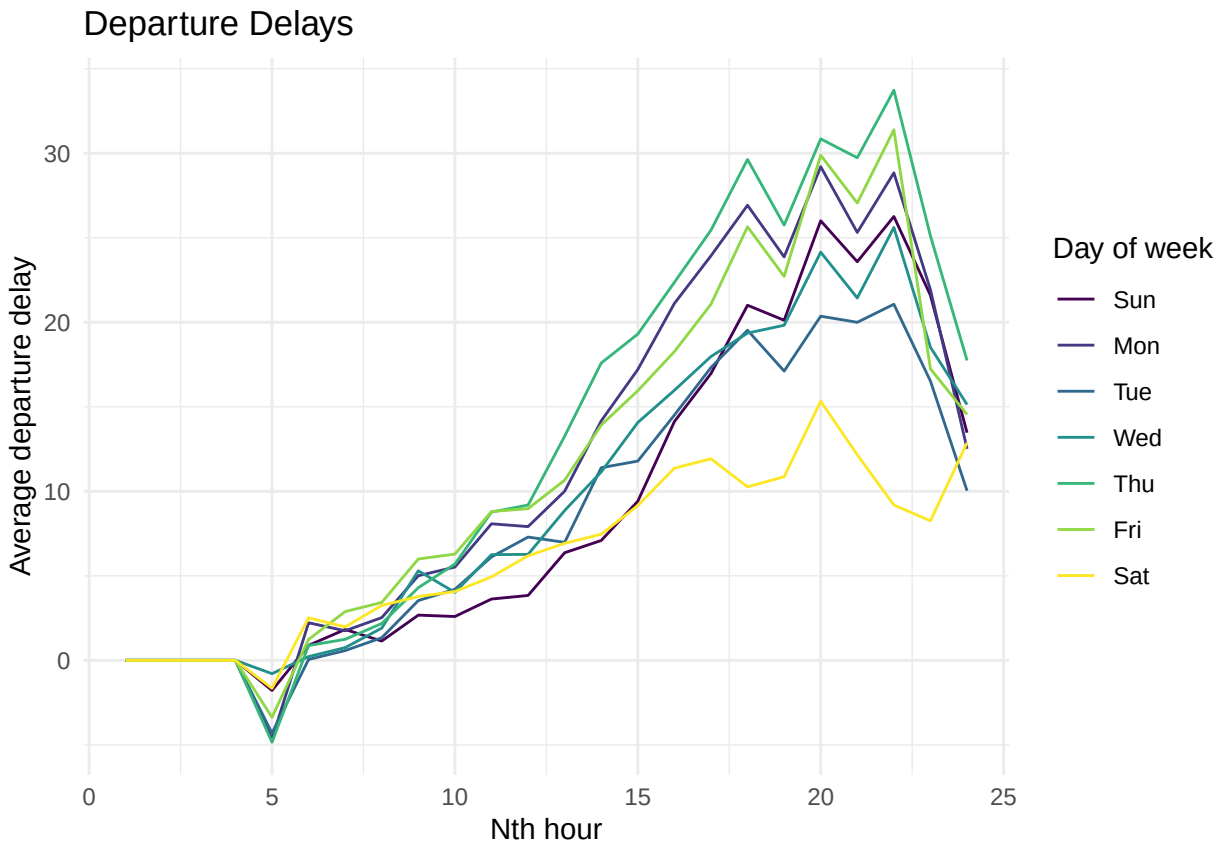
```
all_days <- factor(
  c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),
  levels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),
  ordered = TRUE
)
all_bins <- factor(0:23, levels = 0:23)
complete_grid <- expand.grid(day_of_week = all_days, sched_dep_time_bin = all_bins)

complete_departure_delay <- complete_grid %>%
```

```
  left_join(departure_delay, by = c("day_of_week", "sched_dep_time_bin")) %>%
  mutate(mean_dep_delay = ifelse(is.na(mean_dep_delay), 0, mean_dep_delay),
         sched_dep_time_bin = as.numeric(sched_dep_time_bin)) %>%
  group_by(day_of_week, sched_dep_time_bin) %>%
  summarize(mean_dep_delay = mean(mean_dep_delay, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'day_of_week'. You can override using the
## `.groups` argument.
```

```
ggplot(complete_departure_delay, aes(sched_dep_time_bin, mean_dep_delay, col=day_of_week)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Departure Delays",
       x = "Nth hour",
       y = "Average departure delay",
       color = "Day of week")
```



4. Which destination are most popular from each New York airport, and how have these changed over time?

necessary: dest, origin, month

```
max_dest_df <- flights %>%
  group_by(origin, dest) %>%
  summarize(count = n()) %>%
  group_by(origin) %>%
```

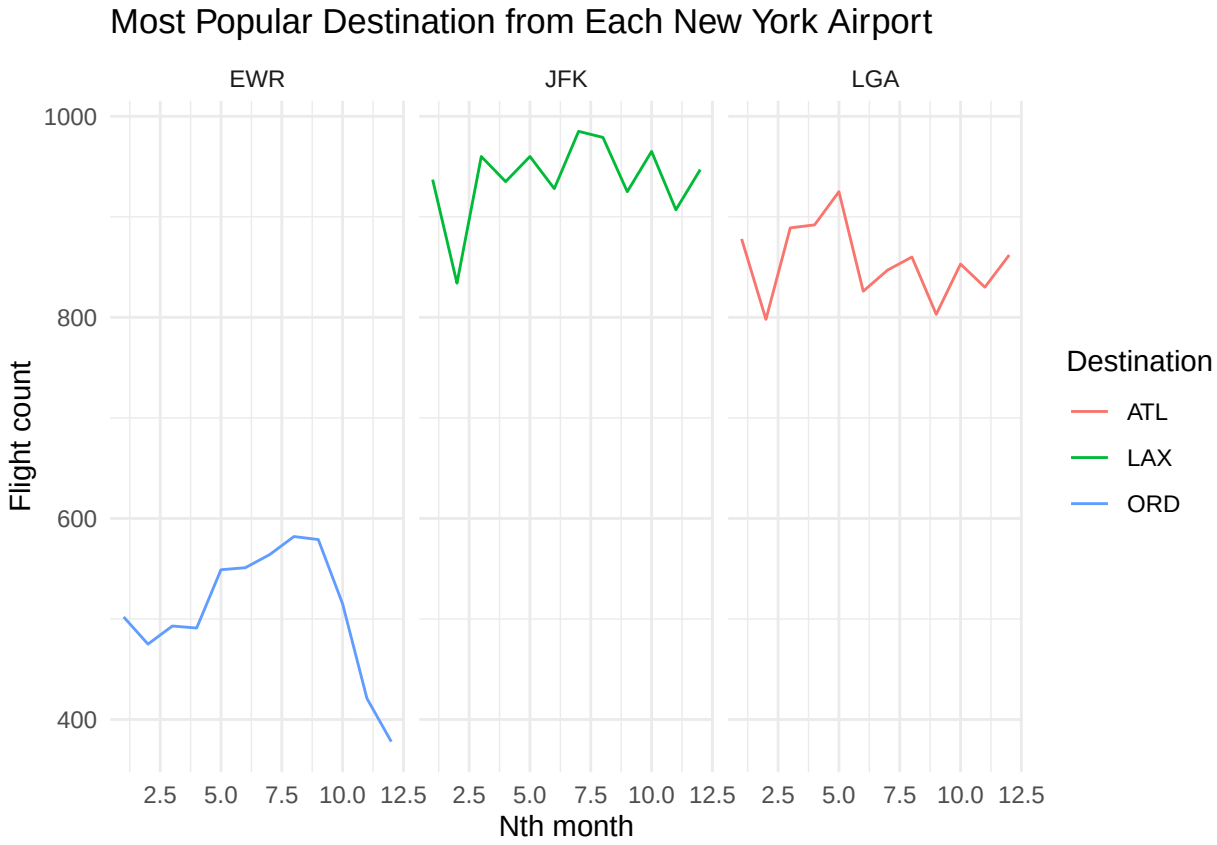
```
  filter(count == max(count))
```

```
## `summarise()` has grouped output by 'origin'. You can override using the
## `.groups` argument.
```

```
popular_dest <- flights %>%
  inner_join(max_dest_df, by = c("origin", "dest")) %>%
  group_by(origin, dest, month) %>%
  summarize(flight_count = n())
```

```
## `summarise()` has grouped output by 'origin', 'dest'. You can override using
## the `.groups` argument.
```

```
ggplot(popular_dest, aes(month, flight_count, col=dest)) +
  geom_line() +
  theme_minimal() +
  facet_wrap(~origin) +
  labs(title = "Most Popular Destination from Each New York Airport",
       x = "Nth month",
       y = "Flight count",
       color = "Destination")
```



5. What is the relationship between air time and distance for different carriers?

necessary: air_time, distance, carrier

```
flights_sampled <- flights %>%
  group_by(carrier) %>%
  sample_n(min(n(), 100)) %>%
  ungroup()
```

```
ggplot(flights_sampled, aes(distance, air_time)) +
  geom_point(na.rm = TRUE, alpha=0.5, col="salmon") +
  geom_smooth(method = "lm") +
  theme_minimal() +
  facet_wrap(~carrier) +
  labs(title = "Relationship between Distance and Air Time of Each Carry",
       x = "Distance",
       y = "Air time")
```

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

```
## Warning: Removed 47 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```