Tarea Estadística Computacional Sofia Gerard

2023-10-12

```
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>
```

```
summary(flights)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
```

```
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00.00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00.00
##  Median :29.00   Median :2013-07-03 10:00:00.00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54.64
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00.00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00.00
##
```

```
str(flights)
```

```
## tibble [336,776 x 19] (S3: tbl_df/tbl/data.frame)
##  $ year          : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int [1:336776] 517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int [1:336776] 830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr [1:336776] "UA" "UA" "AA" "B6" ...
##  $ flight        : int [1:336776] 1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr [1:336776] "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr [1:336776] "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr [1:336776] "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num [1:336776] 227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num [1:336776] 1400 1416 1089 1576 762 ...
##  $ hour          : num [1:336776] 5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num [1:336776] 15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct[1:336776], format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
dim(flights)
```

```
## [1] 336776     19
```

Aerolíneas más retrasadas

```
retrasos_aerolineas <- flights |>
  select(dep_delay, arr_delay, carrier) |>
  group_by(carrier) |>
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
             mean_arr_delay = mean(arr_delay, na.rm = TRUE))


# Departure
max_dep_delay <- retrasos_aerolineas %>%
```

```r
  arrange(desc(mean_dep_delay)) |>
  select(carrier, mean_dep_delay) |>
  head(10)

cat("Las 10 aerolíneas que más se retrasan al salir en promedio son:", head(max_dep_delay$carrier, 10))
```

```
## Las 10 aerolíneas que más se retrasan al salir en promedio son: F9 EV YV FL WN 9E B6 VX OO UA
```

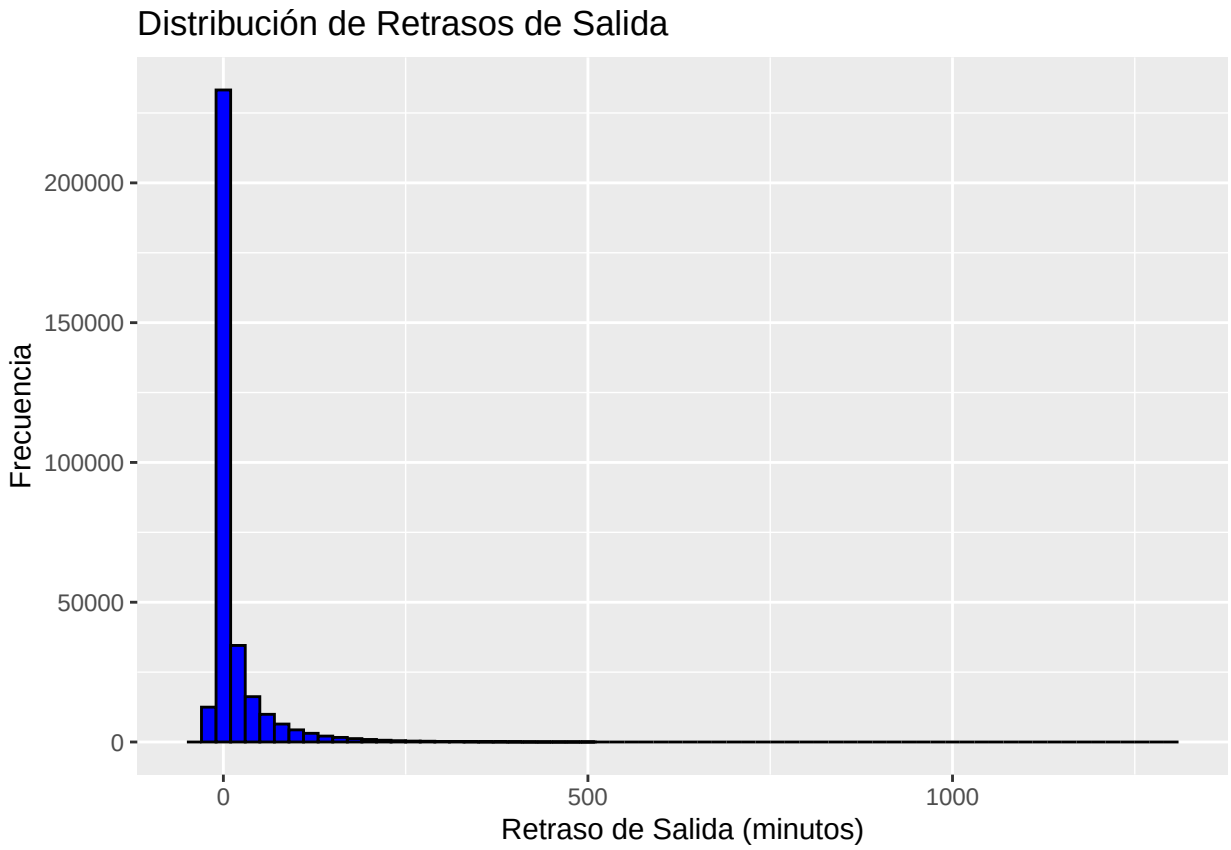
```r
# Arrival
max_arr_delay <- retrasos_aerolineas %>%
  arrange(desc(mean_arr_delay)) |>
  select(carrier, mean_arr_delay) |>
  head(10)

cat("Las 10 aerolíneas que más se retrasan al llegar en promedio son:", head(max_dep_delay$carrier, 10))
```

```
## Las 10 aerolíneas que más se retrasan al llegar en promedio son: F9 EV YV FL WN 9E B6 VX OO UA
```

Distribuciones del retraso

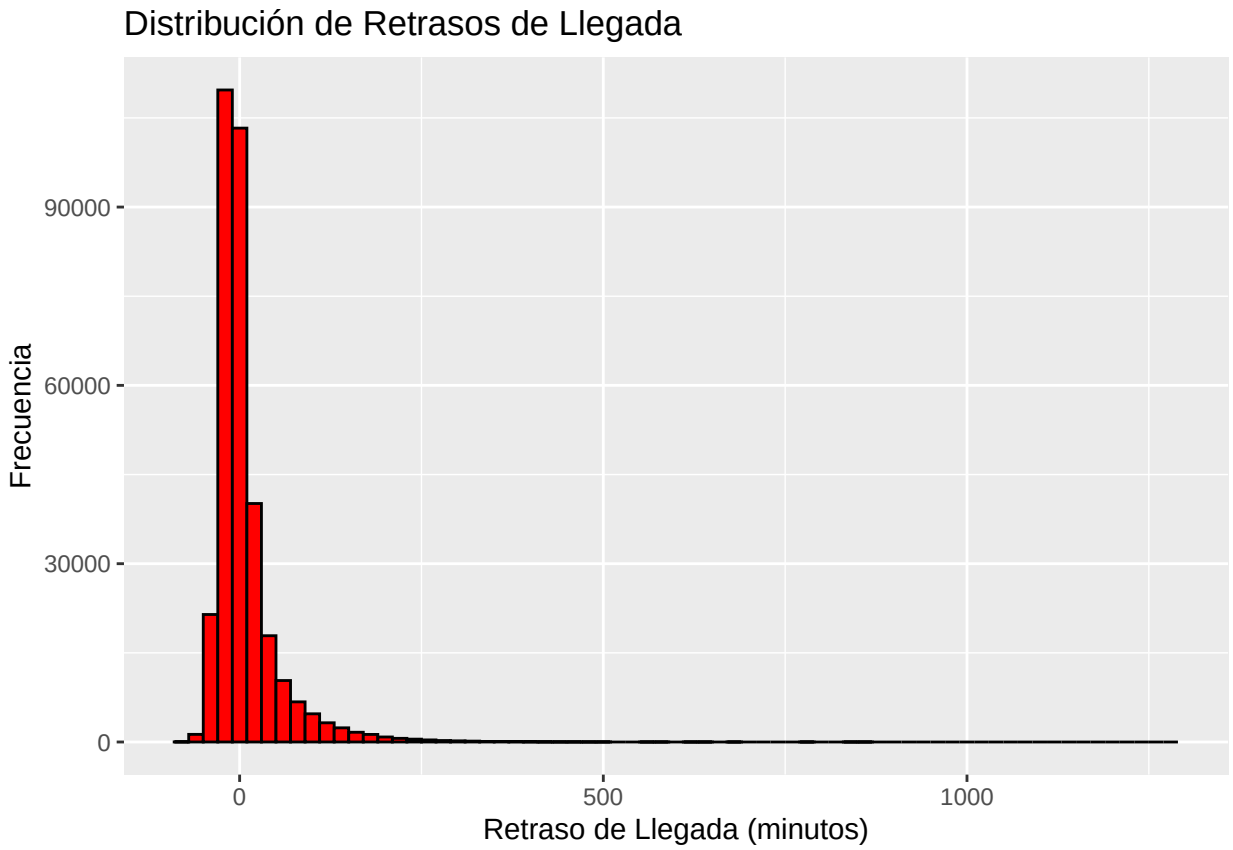
```r
#  dep_delay
ggplot(flights, aes(x = dep_delay)) +
  geom_histogram(binwidth = 20, fill = "blue", color = "black") +
  labs(title = "Distribución de Retrasos de Salida",
       x = "Retraso de Salida (minutos)",
       y = "Frecuencia")
```

```
## Warning: Removed 8255 rows containing non-finite values (`stat_bin()`).
```




```
# arr_delay
ggplot(flights, aes(x = arr_delay)) +
  geom_histogram(binwidth = 20, fill = "red", color = "black") +
  labs(title = "Distribución de Retrasos de Llegada",
       x = "Retraso de Llegada (minutos)",
       y = "Frecuencia")
```

```
## Warning: Removed 9430 rows containing non-finite values (`stat_bin()`).
```



El avion más retrasado

```
#A tail number refers to an identification number painted on an aircraf: tailnum

aviones <- flights %>%
  select(dep_delay,arr_delay,tailnum) %>%
  group_by(tailnum) |>
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
             mean_arr_delay = mean(arr_delay, na.rm = TRUE))
 # Departure
max_dep_delay_avion <- aviones %>%
  arrange(desc(mean_dep_delay)) |>
  select(tailnum, mean_dep_delay) |>
  head(10)

cat("Los 10 aviones que más se retrasan al salir en promedio son:", head(max_dep_delay_avion$tailnum, 1
```

```
## Los 10 aviones que más se retrasan al salir en promedio son: N844MH N922EV N587NW N911DA N851NW N654
```

```
# Arrival
max_arr_delay_avion <- aviones %>%
  arrange(desc(mean_arr_delay)) |>
  select(tailnum, mean_arr_delay) |>
  head(10)

cat("Los 10 aviones que más se retrasan al llegar en promedio son:", head(max_arr_delay_avion$tailnum,
```

```
## Los 10 aviones que más se retrasan al llegar en promedio son: N844MH N911DA N922EV N587NW N851NW N928
```

Avión que más recorrió, el top 10 distancia, viajes

```
distancia <- flights %>%
  select(tailnum, distance) %>%
  group_by(tailnum) %>%
  summarize(media_distancia = mean(distance)) %>%
  arrange(desc(media_distancia))

cat("Los 10 aviones que más distancia han recorrido son:", head(distancia$tailnum, 10))
```

```
## Los 10 aviones que más distancia han recorrido son: N380HA N381HA N382HA N383HA N384HA N385HA N386HA
```

¿Los vuelos están vinculados a la ruta?

```
vuelos_rutas <- flights %>%
  select(flight, origin, dest) %>%
  group_by(flight) %>%
  arrange(flight)

# Parece ser que para un mismo vuelo hay distintas rutas. Pero siempre son
# las mismas 5-6 rutas.
```

Frecuencia de retrasos por hora

```
flights_hora <- flights %>%
  mutate(time_hour = format(time_hour, format = "%H:%M")) %>%
  select(time_hour, arr_delay, dep_delay) %>%
  group_by(time_hour) %>%
  summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
            mean_arr_delay = mean(arr_delay, na.rm = TRUE))
# Departure
max_dep_delay_hora <- flights_hora %>%
  arrange(desc(mean_dep_delay)) |>
  select(time_hour, mean_dep_delay)

gg1 <- ggplot(max_dep_delay_hora, aes(x = time_hour, y = mean_dep_delay)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Máximo Retraso Promedio en la Salida por Hora",
       x = "Hora del Día",
       y = "Máximo Retraso en la Salida (minutos)") +
  theme_minimal()

# Arrival
max_arr_delay_hora <- flights_hora %>%
  arrange(desc(mean_arr_delay)) |>
  select(time_hour, mean_arr_delay)

gg2 <- ggplot(max_arr_delay_hora, aes(x = time_hour, y = mean_arr_delay)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Máximo Retraso Promedio en la Llegada por Hora",
       x = "Hora del Día",
       y = "Máximo Retraso en la Llegada (minutos)") +
  theme_minimal()
```

Hay diferencia por aeropuerto en retrasos

```
# Aeropuerto Origin

aeropuertos_origin <- flights %>%
  select(origin, dep_delay) %>%
  group_by(origin) %>%
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE)) %>%
  arrange(desc(mean_dep_delay))

# Aeropuerto Destination

aeropuertos_destination <- flights %>%
  select(dest, arr_delay) %>%
  group_by(dest) %>%
   summarize(mean_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
   arrange(desc(mean_arr_delay))
```

Para el aeropuerto origen, el que más se atrasa en promedio es EWR (15.1), mientras que el aeropuerto destino con más retrasos en promedio es CAE (41.8).

Hay diferencia por mes

```
por_mes <- flights %>%
  select(month, dep_delay, arr_delay) %>%
  group_by(month) %>%
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
             mean_arr_delay = mean(arr_delay, na.rm = TRUE))

# Departure
max_dep_delay_month <- por_mes %>%
  arrange(desc(mean_dep_delay)) |>
  select(month, mean_dep_delay) |>
  head(10)

cat("El peor mes para salidas, en promedio (más horas de retraso) es:",
    max(max_dep_delay_month$month))
```

```
## El peor mes para salidas, en promedio (más horas de retraso) es: 12
```

```
# Arrival
max_arr_delay_month <- por_mes %>%
  arrange(desc(mean_arr_delay)) |>
  select(month, mean_arr_delay) |>
  head(10)

cat("El peor mes para llegadas, en promedio (más horas de retraso) es:",
    max(max_arr_delay_month$month))
```

```
## El peor mes para llegadas, en promedio (más horas de retraso) es: 12
```

Cuál es el peor vuelo

```
peores_vuelos <- flights %>%
  select(flight, dep_delay, arr_delay) %>%
```

```
  group_by(flight) %>%
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
             mean_arr_delay = mean(arr_delay, na.rm = TRUE))

 # Departure
max_dep_delay_peor_vuelo <- peores_vuelos %>%
  arrange(desc(mean_dep_delay)) |>
  select(flight, mean_dep_delay) |>
  head(10)


# Arrival
max_arr_delay_peor_vuelo <- peores_vuelos %>%
  arrange(desc(mean_arr_delay)) |>
  select(flight, mean_arr_delay) |>
  head(10)

tb_inter <- inner_join(max_dep_delay_peor_vuelo, max_arr_delay_peor_vuelo, by = "flight")



flights |> filter(flight == 1510)
```

```
## # A tibble: 1 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    12     2     1823           1345       278     2123           1640
## # 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>
```

EL peor vuelo es el 1510 de EWR a IAH.

Cuál es el peor día de la semana para volar

```
fechas <- flights %>%
  mutate(time_hour = as.POSIXct(time_hour, format = "%Y-%m-%d %H:%M:%S"),
         dia_semana = format(time_hour, "%A"))

dia_semana <- fechas %>%
  select(dep_delay,arr_delay,dia_semana) %>%
  group_by(dia_semana) %>%
   summarize(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
             mean_arr_delay = mean(arr_delay, na.rm = TRUE))


 # Departure
peor_dia_dep <- dia_semana %>%
  arrange(desc(mean_dep_delay)) |>
  select(dia_semana, mean_dep_delay)


# Arrival
peor_dia_arr <- dia_semana %>%
```

```
  arrange(desc(mean_arr_delay)) |>
  select(dia_semana, mean_arr_delay)
```

El peor día para volar en terminos de horas de atraso promedio es jueves.