Analisis exploratorio y curacion de datos

Mario Ferreyra, Emiliano Kokic

18 de Mayo 2018

Practico 1

Entregar un Rmd donde se encuentren todos los vuelos que:

- Que arribaron con un retraso de mas de dos horas.
- Volaron hacia Houston (IAH o HOU)
- Fueron operados por United, American o Delta.
- Salieron en Verano (Julio, Agosto y Septiembre)
- Arrivaron mas de dos horas tarde, pero salieron bien.
- Salieron entre medianoche y las 6 am.

Conjunto de datos sobre los vuelos en Nueva York en 2013.

```
library(nycflights13)
flights <- nycflights13::flights
flights
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   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>
```

- 1-Checkeando las dimensiones

```
dim(flights)
```

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

- 2-Estructura

```
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
```

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

- 3-Resumen

```
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
```

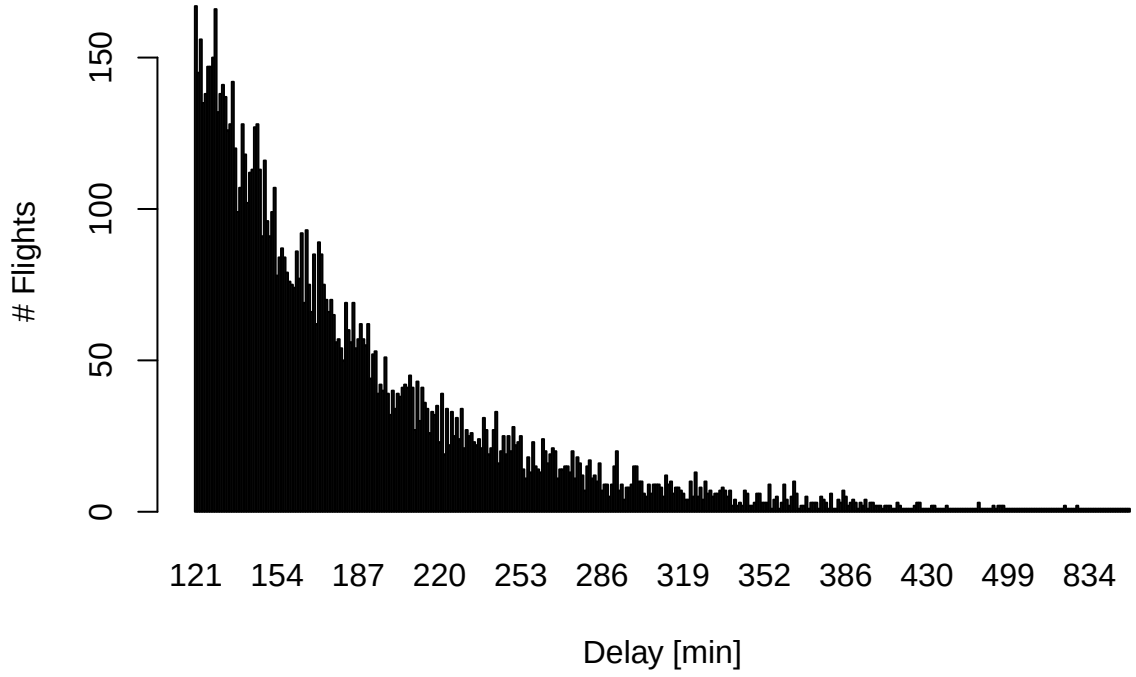
```
##       hour           minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:02:36
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

1. Vuelos que arribaron con un retraso de mas de dos horas.

```
# 2 hs <---> 120 min
flights_delay_more_2hs <- subset(flights, flights$arr_delay > 120)

d = dim(flights_delay_more_2hs)
print(paste0("Cantidad de vuelos que arribaron con retraso mayor a 2hs: ", d[1]))
```

```
## [1] "Cantidad de vuelos que arribaron con retraso mayor a 2hs: 10034"
```

```
counts_flights_delay_more_2hs <- table(flights_delay_more_2hs$arr_delay)
#counts_flights_delay_more_2hs

barplot(counts_flights_delay_more_2hs, xlab="Delay [min]", ylab="# Flights")
```




```
# Distribucion Exponencial
```

Vuelos que volaron hacia Houston (IAH o HOU)

```
flights_2_houston <- subset(flights, flights$dest == "IAH" | flights$dest == "HOU")
```

```r
d = dim(flights_2_houston)
print(paste0("Cantidad de vuelos hacia Houston: ", d[1]))
```

```
## [1] "Cantidad de vuelos hacia Houston: 9313"
```

Vuelos que fueron operados por United, American o Delta.

```r
nycflights13::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.
```

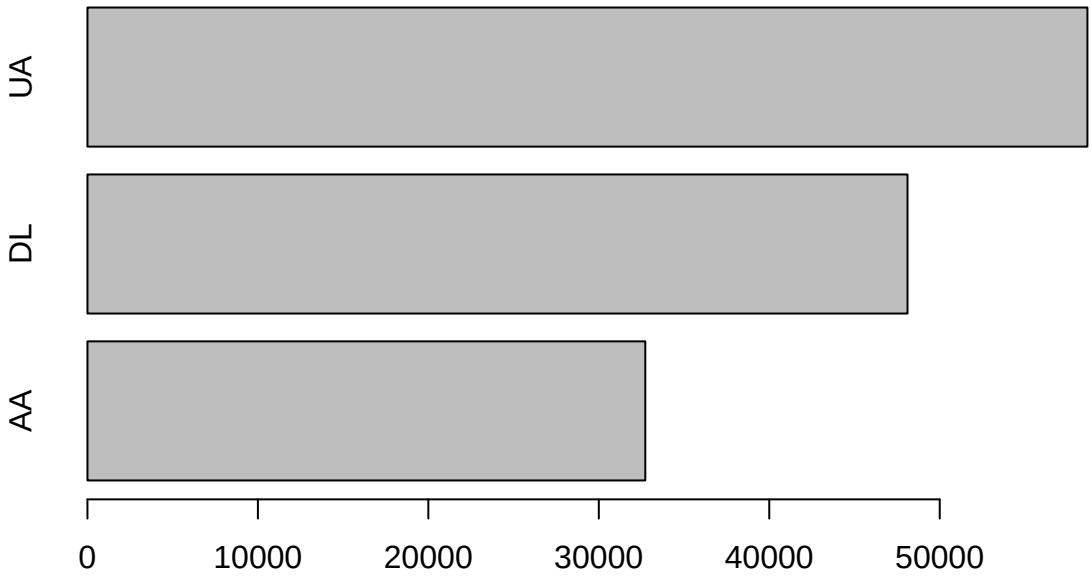
```r
United <- "UA"  # United Air Lines Inc.
American <- "AA"  # American Airlines Inc.
Delta <- "DL"  # Delta Air Lines Inc.

flights_UAD <- subset(flights, (flights$carrier == United) |
                              (flights$carrier == American) |
                              (flights$carrier == Delta))

dim(flights_UAD)
```

```
## [1] 139504     19
```

```r
counts_flights_UAD <- table(flights_UAD$carrier)
counts_flights_UAD
```

```
##
##    AA    DL    UA
## 32729 48110 58665
```

```r
barplot(counts_flights_UAD,  horiz=TRUE)
```



Vuelos que salieron en Verano (Julio, Agosto y Septiembre)

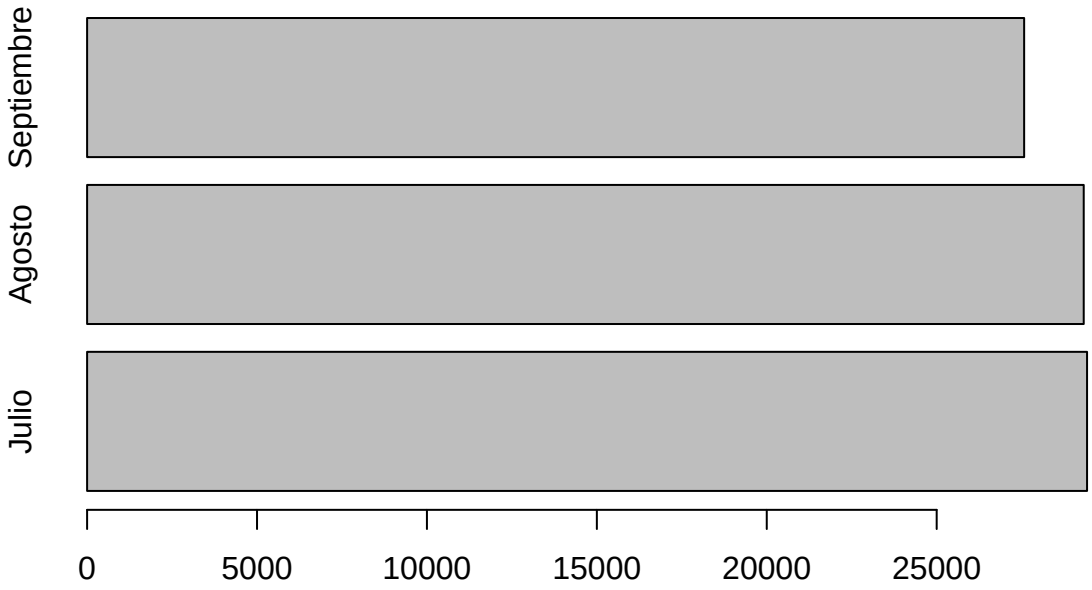
```r
# Julio <---> 7
# Agosto <---> 8
# Septiembre <---> 9
flights_summer <- subset(flights, 7 <= flights$month & flights$month <= 9)

dim(flights_summer)
```

```
## [1] 86326    19
```

```r
counts_flights_summer <- table(flights_summer$month)
counts_flights_summer
```

```
##
##     7     8     9
## 29425 29327 27574
```

```r
barplot(counts_flights_summer, horiz=TRUE, names.arg=c("Julio", "Agosto", "Septiembre"))
```



Vuelos que arrivaron mas de dos horas tarde, pero salieron bien.

```
#subset(flights_delay_more_2hs, is.na(flights_delay_more_2hs))

# Vuelos que salieron bien = No hubo demora en la salida
flights_delay_more_2hs_OK <- subset(flights, flights$arr_delay > 120 & flights$dep_delay <= 0)

d = dim(flights_delay_more_2hs_OK)

print(paste0("Cantidad de vuelos con delay mayor a 2hs que no demoraron en salir: ", d[1]))
```

```
## [1] "Cantidad de vuelos con delay mayor a 2hs que no demoraron en salir: 29"
```

Vuelos que salieron entre medianoche (00:00) y las 6 AM (06:00).

```
time_dep <- function(hour=0, min=0, dep_delay=0){
    return(hour * 60 + min + dep_delay)
}

flights_00_06 <- subset(flights, time_dep(flights$hour, flights$minute, flights$dep_delay) >= time_dep(
                                 time_dep(flights$hour, flights$minute, flights$dep_delay) <= time_dep(

d = dim(flights_00_06)

print(paste0("Cantidad de vuelos que salieron entre medianoche y las 6AM: ", d[1]))
```

```
## [1] "Cantidad de vuelos que salieron entre medianoche y las 6AM: 8165"
```

```
flights_00_06_by_hour <- table(flights_00_06$hour)
flights_00_06_by_hour
```

```
##
##    5    6
## 1829 6336
```

```
barplot(flights_00_06_by_hour, main="Vuelos que salieron entre 00:00 y 06:00",
        xlab="Horas", ylab="Cantidad")
```