Memoria TFM

Análisis y predicción del retraso en los vuelos

Ester Llorente San Juan

1. Introducción

El objetivo de este Trabajo Fin de Master no es más que el de aplicar las distintas técnicas aprendidas en el V-Master de Data Science con el fin de analizar y predecir el retraso y el tiempo de retraso en los vuelos comerciales de EE.UU.

Se define como retraso de vuelo cuando un vuelo de una aerolínea despega y aterriza más tarde que su hora programada. La Administración Federal de Aviación (FAA) considera que un vuelo se retrasa cuando aterriza 15 minutos más tarde de su hora programada. Una cancelación ocurre cuando la línea aérea no opera el vuelo por una cierta razón.

En Estados Unidos, cando los vuelos se cancelan o se retrasan, los pasajeros pueden tener derecho a una indemnización debido a la reglas obedecidas por cada compañía de vuelo. Estas reglas generalmente especifican que ante una cancelación, los pasajeros pueden tener derecho a ciertos reembolsos, incluyendo una habitación gratis, si el próximo vuelo es al día siguiente del vuelo cancelado, reembolso, cambio de ruta, llamadas telefónicas y refrigerio.

Cuando un vuelo se retrasa, el Departamento de Transporte impone una multa de hasta $27,500 por pasajero, para los aviones que quedan en pista durante más de 3 horas sin despegar (cuatro horas para vuelos internacionales). En los Estados Unidos, los pasajeros no tienen derecho a una compensación cuando se produce un retraso, ni si quiera un recorte de tarifas. Las compañías deben de pagar los costes de alojamiento de los pasajeros si la demora o cancelación es por su culpa, pero no es así si las causas están fuera de su control, como el clima.

Desde 2003 la Oficina de Estadística de Transporte de los Estados Unidos ha estado haciendo un seguimiento de las causas de los retrasos en los vuelos. El número de retrasos en los vuelos ha aumentado ya que el personal ha sido reducido debido a los problemas financieros que siguieron a los ataques del 11 de Septiembre.

Los retrasos se agrupan el 5 categorías (Aerolínea+Turnaround local, Tiempo Extremo, Aeronaves que llegan tarde o retraso reaccionario, Seguridad y Sistema ATM), siempre teniendo en cuenta que para que se contabilice un retraso en una categoría causal, este retraso ha de haber sido mayor a 15 minutos. Algunas de las causas de los retrasos o cancelaciones de los vuelos incluyen:

- Problemas técnicos de la aerolínea (la causa principal de las demoras en los vuelos)
- Congestión en el tráfico aéreo.
- Terremotos y Tsunamis (por ejemplo, en caso de terremoto y tsunami en el Océano Índico en 2004, terremoto en Chile en 2010 y el terremoto y tsunami en Töhoku en 2011).
- Clima inclemente, tormentas eléctricas, huracanes o ventiscas.
- Acumulación de retraso en las salidas por demoras en las llegadas.
- Problemas de mantenimiento con la aeronave.
- Ataques terroristas (por ejemplo, los bombardeos suicidas de Bruselas en 2016, y el ataque al aeropuerto de Atatürk en 2016).

Para la creación del estudio se utiliza los datos históricos del Departamento de Transporte de los Estados Unidos (US DOT), obtenido de https://www.transtats.bts.gov/DL_SelectFields.asp

2. Descripción de los datos.

Para la realización del análisis, se han descargado datos para los meses de Enero y Febrero de 2014 del US DOT por tener el mayor porcentaje de retrasos 25,43% y 23,56% respectivamente, con un total de 23,02% de vuelos retrasados en ese año.
Se ha realizado un preprocesamiento de los ficheros descargados que incluye: Lectura de los ficheros, y Limpieza, Análisis y tratamiento de los datos.

El dataset descargado contiene en total 902551 objetos y 110 variables, como se puede observar a continuación:

```
> str(flightsAux)
Classes 'data.table' and 'data.frame': 902551 obs. of  110 variables:
 $ Year                  : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ Quarter               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Month                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DayofMonth            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ DayOfWeek             : int  3 4 5 6 7 1 2 3 4 5 ...
 $ FlightDate            : chr  "2014-01-01" "2014-01-02" "2014-01-03" "2014-01-04" ...
 $ UniqueCarrier         : chr  "AA" "AA" "AA" "AA" ...
 $ AirlineID             : int  19805 19805 19805 19805 19805 19805 19805 19805 19805 19805 ...
 $ Carrier               : chr  "AA" "AA" "AA" "AA" ...
 $ TailNum               : chr  "N338AA" "N338AA" "N323AA" "N327AA" ...
 $ FlightNum             : chr  "1" "1" "1" "1" ...
 $ OriginAirportID       : int  12478 12478 12478 12478 12478 12478 12478 12478 12478 12478 ...
 $ OriginAirportSeqID    : int  1247802 1247802 1247802 1247802 1247802 1247802 1247802 1247802 1247802 1247802 ...
 $ OriginCityMarketID    : int  31703 31703 31703 31703 31703 31703 31703 31703 31703 31703 ...
 $ Origin                : chr  "JFK" "JFK" "JFK" "JFK" ...
 $ OriginCityName        : chr  "New York, NY" "New York, NY" "New York, NY" "New York, NY" ...
 $ OriginState           : chr  "NY" "NY" "NY" "NY" ...
 $ OriginStateFips       : chr  "36" "36" "36" "36" ...
 $ OriginStateName       : chr  "New York" "New York" "New York" "New York" ...
 $ OriginWac             : int  22 22 22 22 22 22 22 22 22 22 ...
 $ DestAirportID         : int  12892 12892 12892 12892 12892 12892 12892 12892 12892 12892 ...
 $ DestAirportSeqID      : int  1289203 1289203 1289203 1289203 1289203 1289203 1289203 1289203 1289203 1289203 ...
 $ DestCityMarketID      : int  32575 32575 32575 32575 32575 32575 32575 32575 32575 32575 ...
 $ Dest                  : chr  "LAX" "LAX" "LAX" "LAX" ...
 $ DestCityName          : chr  "Los Angeles, CA" "Los Angeles, CA" "Los Angeles, CA" "Los Angeles, CA" ...
 $ DestState             : chr  "CA" "CA" "CA" "CA" ...
 $ DestStateFips         : chr  "06" "06" "06" "06" ...
 $ DestStateName         : chr  "California" "California" "California" "California" ...
 $ DestWac               : int  91 91 91 91 91 91 91 91 91 91 ...
 $ CRSDepTime            : chr  "0900" "0900" "0900" "0900" ...
 $ DepTime               : chr  "0914" "0857" "" "1005" ...
 $ DepDelay              : num  14 -3 NA 65 110 17 10 23 -1 29 ...
 $ DepDelayMinutes       : num  14 0 NA 65 110 17 10 23 0 29 ...
 $ DepDel15              : num  0 0 NA 1 1 1 0 1 0 1 ...
 $ DepartureDelayGroups  : int  0 -1 NA 4 7 1 0 1 -1 1 ...
 $ DepTimeBlk            : chr  "0900-0959" "0900-0959" "0900-0959" "0900-0959" ...
 $ TaxiOut               : num  20 43 NA 17 32 24 10 12 19 27 ...
 $ WheelsOff             : chr  "0934" "0940" "" "1022" ...
 $ WheelsOn              : chr  "1233" "1220" "" "1308" ...
 $ TaxiIn                : num  5 6 NA 16 6 13 3 6 7 10 ...
 $ CRSArrTime            : chr  "1225" "1225" "1225" "1225" ...
 $ ArrTime               : chr  "1238" "1226" "" "1324" ...
 $ ArrDelay              : num  13 1 NA 59 110 -8 -13 -10 -21 20 ...
 $ ArrDelayMinutes       : num  13 1 NA 59 110 0 0 0 0 20 ...
 $ ArrDel15              : num  0 0 NA 1 1 0 0 0 0 1 ...
 $ ArrivalDelayGroups    : int  0 0 NA 3 7 -1 -1 -1 -2 1 ...
 $ ArrTimeBlk            : chr  "1200-1259" "1200-1259" "1200-1259" "1200-1259" ...
 $ Cancelled             : num  0 0 1 0 0 0 0 0 0 0 ...
 $ CancellationCode      : chr  "" "" "B" "" ...
 $ Diverted              : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CRSElapsedTime        : num  385 385 385 385 385 385 385 385 385 385 ...
 $ ActualElapsedTime     : num  384 389 NA 379 385 360 362 352 365 376 ...
 $ AirTime               : num  359 340 NA 346 347 323 349 334 339 339 ...
 $ Flights               : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Distance              : num  2475 2475 2475 2475 2475 ...
 $ DistanceGroup         : int  10 10 10 10 10 10 10 10 10 10 ...
 $ CarrierDelay          : num  NA NA NA 0 0 NA NA NA NA 0 ...
 $ WeatherDelay          : num  NA NA NA 59 110 NA NA NA NA 20 ...
 $ NASDelay              : num  NA NA NA 0 0 NA NA NA NA 0 ...
 $ SecurityDelay         : num  NA NA NA 0 0 NA NA NA NA 0 ...
 $ LateAircraftDelay     : num  NA NA NA 0 0 NA NA NA NA 0 ...
 $ FirstDepTime          : chr  "" "" "" "" ...
 $ TotalAddGTime         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ LongestAddGTime       : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
$ DivAirportLandings       : int  0 0 0 0 0 0 0 0 0 0 ...
$ DivReachedDest           : num  NA NA NA NA NA NA NA NA NA NA ...
$ DivActualElapsedTime     : num  NA NA NA NA NA NA NA NA NA NA ...
$ DivArrDelay              : num  NA NA NA NA NA NA NA NA NA NA ...
$ DivDistance              : num  NA NA NA NA NA NA NA NA NA NA ...
$ Div1Airport              : chr  "" "" "" "" ...
$ Div1AirportID            : int  NA NA NA NA NA NA NA NA NA NA ...
$ Div1AirportSeqID         : int  NA NA NA NA NA NA NA NA NA NA ...
$ Div1WheelsOn             : chr  "" "" "" "" ...
$ Div1TotalGTime           : num  NA NA NA NA NA NA NA NA NA NA ...
$ Div1LongestGTime         : num  NA NA NA NA NA NA NA NA NA NA ...
$ Div1WheelsOff            : chr  "" "" "" "" ...
$ Div1TailNum              : chr  "" "" "" "" ...
$ Div2Airport              : chr  "" "" "" "" ...
$ Div2AirportID            : int  NA NA NA NA NA NA NA NA NA NA ...
$ Div2AirportSeqID         : int  NA NA NA NA NA NA NA NA NA NA ...
$ Div2WheelsOn             : chr  "" "" "" "" ...
$ Div2TotalGTime           : num  NA NA NA NA NA NA NA NA NA NA ...
$ Div2LongestGTime         : num  NA NA NA NA NA NA NA NA NA NA ...
$ Div2WheelsOff            : chr  "" "" "" "" ...
$ Div2TailNum              : chr  "" "" "" "" ...
$ Div3Airport              : chr  "" "" "" "" ...
$ Div3AirportID            : logi  NA NA NA NA NA NA ...
$ Div3AirportSeqID         : logi  NA NA NA NA NA NA ...
$ Div3WheelsOn             : chr  "" "" "" "" ...
$ Div3TotalGTime           : logi  NA NA NA NA NA NA ...
$ Div3LongestGTime         : logi  NA NA NA NA NA NA ...
$ Div3WheelsOff            : chr  "" "" "" "" ...
$ Div3TailNum              : chr  "" "" "" "" ...
$ Div4Airport              : chr  "" "" "" "" ...
$ Div4AirportID            : logi  NA NA NA NA NA NA ...
$ Div4AirportSeqID         : logi  NA NA NA NA NA NA ...
$ Div4WheelsOn             : chr  "" "" "" "" ...
$ Div4TotalGTime           : logi  NA NA NA NA NA NA ...
$ Div4LongestGTime         : logi  NA NA NA NA NA NA ...
```

Este dataset, está compuesto por datos de tipo int, chr, num y logi.

Se han tratado las variables para poder realizar un análisis del dataset. Ejecutando el comando summary, después de ese tratamiento (las variables que se muestran en rojo, han sido eliminadas) se obtiene:

```
> summary(flightsAux)
```

Year	Quarter	Month	DayofMonth	DayOfWeek
2014 : 845361	Min. : 1	1 : 439620	27 : 31457	1 : 121429
	1st Qu. : 1	2 : 405741	20 : 31288	2 : 111659
	Median : 1		24 : 31282	3 : 128160
	Mean : 1		17 : 31264	4 : 134944
	3rd Qu. : 1		10 : 30904	5 : 138912
	Max. : 1		23 : 30895	6 : 97320
			(Other) : 658271	7 : 112937

FlightDate	UniqueCarrier	AirlineID	Carrier	TailNum
Min. : 2014-01-01	WN : 168404	19393 : 168404	WN : 168404	N477HA : 679
1st Qu. : 2014-01-15	DL : 107389	19790 : 107389	DL : 107389	N485HA : 679
Median : 2014-01-30	EV : 96383	20366 : 96383	EV : 96383	N481HA : 676
Mean : 2014-01-30	OO : 90767	20304 : 90767	OO : 90767	N479HA : 629
3rd Qu. : 2014-02-15	AA : 82675	19805 : 82675	AA : 82675	N486HA : 622
Max. : 2014-02-28	UA : 72257	19977 : 72257	UA : 72257	N488HA : 617
	(Other) : 227486	(Other) : 227486	(Other) : 227486	(Other) : 841459

FlightNum	OriginAirportID	OriginAirportSeqID	OriginCityMarketID	Origin
404 : 681	10397 : 53992	1039705 : 53992	30397 : 53992	ATL : 53992
312 : 627	11298 : 42507	1129803 : 42507	30194 : 49667	DFW : 42507
402 : 621	13930 : 36189	1393003 : 36189	32575 : 48166	ORD : 36189
358 : 608	12892 : 34080	1289203 : 34080	30977 : 47300	LAX : 34080
665 : 575	11292 : 32958	1129202 : 32958	31703 : 45436	DEN : 32958
315 : 562	12266 : 26281	1226603 : 26281	32457 : 37002	IAH : 26281
(Other) : 84168	(Other) : 619354	(Other) : 619354	(Other) : 563798	(Other) : 26281

OriginCityName	OriginState	OriginStateFips	OriginStateName	OriginWac
Atlanta, GA : 53992	CA : 111754	06 : 111754	California : 111754	Min. : 1.00
Chicago, IL : 47300	TX : 108452	48 : 108452	Texas : 108452	1st Qu. : 34.00
Dallas/Fort Worth, TX : 42507	FL : 69853	12 : 69853	Florida : 69853	Median : 54.00
Houston, TX : 35425	GA : 56012	13 : 56012	Georgia : 56012	Mean : 56.18
Los Angeles, CA : 34080	IL : 49234	17 : 49234	Illinois : 49234	3rd Qu. : 82.00
Denver, CO : 32958	CO : 37584	08 : 37584	Colorado : 37584	Max. : 93.00
(Other) : 599099	(Other) : 412472	(Other) : 412472	(Other) : 412472	

DestAirportID	DestAirportSeqID	DestCityMarketID	Dest	DestCityName
10397 : 53846	1039705 : 53846	30397 : 53846	ATL : 53846	Atlanta, GA : 53846
11298 : 42515	1129803 : 42515	30194 : 49636	DFW : 42515	Chicago, IL : 47062
13930 : 36066	1393003 : 36066	32575 : 48163	ORD : 36066	Dallas/Fort Worth, TX : 42515
12892 : 34078	1289203 : 34078	30977 : 47062	LAX : 34078	Houston, TX : 35261
11292 : 32862	1129202 : 32862	31703 : 45135	DEN : 32862	Los Angeles, CA : 34078
12266 : 26112	1226603 : 26112	32457 : 36955	IAH : 26112	Denver, CO : 32862
(Other) : 619882	(Other) : 619882	(Other) : 564564	(Other) : 619882	(Other) : 599737

DestState
CA : 111710
TX : 108318
FL : 69956
GA : 55873
IL : 49013
CO : 37497
(Other) : 412994

DestStateFips
06 : 111710
48 : 108318
12 : 69956
13 : 55873
17 : 49013
08 : 37497
(Other) : 412994

DestStateName
California :111710
Texas :108318
Florida : 69956
Georgia : 55873
Illinois : 49013
Colorado : 37497
(Other) :412994

DestWac
91 : 111710
74 : 108318
33 : 69956
34 : 55873
41 : 49013
82 : 37497
(Other) : 412994

CRSDepTime
Min. : 5
1st Qu. : 930
Median : 1320
Mean : 1324
3rd Qu. : 1715
Max. : 2359

DepTime
Min. : 1
1st Qu. : 939
Median : 1332
Mean : 1340
3rd Qu. : 1729
Max. : 2400

DepDelay
Min. : -112.00
1st Qu. : -4.00
Median : 0.00
Mean : 14.49
3rd Qu. : 15.00
Max. : 1638.00
NA's : 53578

DepDelayMinutes
Min. : 0.00
1st Qu. : 0.00
Median : 0.00
Mean : 16.87
3rd Qu. : 15.00
Max. : 1638.00
NA's : 53578

DepDel15
0 : 630264
1 : 215097

DepartureDelayGroups
-1 : 407285
0 : 219967
1 : 78001
2 : 42064
3 : 25534
4 : 17037
(Other) : 55473

DepTimeBlk
0800-0859 : 61762
1700-1759 : 61731
0700-0759 : 57316
1100-1159 : 56772
1300-1359 : 56157
1000-1059 : 54496
(Other) : 497127

TaxiOut
Min. : 1.00
1st Qu. : 10.00
Median : 14.00
Mean : 16.07
3rd Qu. : 18.00
Max. : 178.00
NA's : 54254

WheelsOff
Min. : 1
1st Qu. : 954
Median :1346
Mean :1363
3rd Qu. :1743
Max. :2400

WheelsOn
Min. : 1
1st Qu. : 1122
Median : 1523
Mean : 1495
3rd Qu. : 1910
Max. : 2400

TaxiIn
Min. : 1.00
1st Qu. : 4.00
Median : 6.00
Mean : 7.22
3rd Qu. : 8.00
Max. : 246.00
NA's : 55583

CRSArrTime
Min. : 1
1st Qu. : 1130
Median : 1527
Mean : 1512
3rd Qu. : 1910
Max. : 2359

ArrTime
Min. : 1
1st Qu. : 1127
Median : 1528
Mean : 1502
3rd Qu. : 1916
Max. : 2400

ArrDelay
Min. :-112.00
1st Qu. :-11.00
Median : -1.00
Mean :11.25
3rd Qu. :16.00
Max. :1628.00
NA's :57190

ArrDelayMinutes
Min. : 0.00
1st Qu. : 0.00
Median : 0.00
Mean : 17.29
3rd Qu. : 16.00
Max. : 1628.00
NA's : 57190

ArrDel15
0 : 623927
1 : 221434

ArrivalDelayGroups
-1 : 316077
0 : 188376
-2 : 119474
1 : 80898
2 : 42968
3 : 25964
(Other) : 71604

ArrTimeBlk
1600-1659 : 62982
1800-1859 : 58957
1400-1459 : 56200
1000-1059 : 55833
2000-2059 : 55129
1200-1259 : 54207
(Other) : 502053

Cancelled
0 : 847980
1 : 54571

CancellationCode
Length : 902551
Class : character
Mode : character

Diverted
0 : 899932
1 : 2619

CRSElapsedTime
Min. : 19.0
1st Qu. : 85.0
Median : 118.0
Mean : 137.3
3rd Qu. : 170.0
Max. : 670.0
NA's : 2

ActualElapsedTime
Min. : 15.0
1st Qu. : 80.0
Median : 116.0
Mean : 134.8
3rd Qu. : 167.0
Max. : 775.0
NA's : 57190

AirTime
Min. :7.0
1st Qu. :59.0
Median :92.0
Mean :111.6
3rd Qu. :142.0
Max. :688.0
NA's :57190

Flights
Min. : 1
1st Qu. : 1
Median : 1
Mean : 1
3rd Qu. : 1
Max. : 1

Distance
Min. : 26.0
1st Qu. : 356.0
Median : 622.0
Mean : 787.1
3rd Qu. : 1020.0
Max. : 4983.0

DistanceGroup
2 : 210225
3 : 165791
4 : 127786
1 : 115122
5 : 86899
7 : 37252
(Other) : 102286

CarrierDelay
Min. : 0.0
1st Qu. : 0.0
Median : 3.0
Mean : 18.7
3rd Qu. : 19.0
Max. : 1620.0
NA's : 681117

WeatherDelay
Min. :0.0
1st Qu. :0.0
Median :0.0
Mean :3.9
3rd Qu. :0.0
Max. :1288.0
NA's :681117

NASDelay
Min. : 0.0
1st Qu. : 0.0
Median : 2.0
Mean : 12.9
3rd Qu. : 17.0
Max. : 799.0
NA's : 681117

SecurityDelay
Min. : 0
1st Qu. : 0
Median : 0
Mean : 0
3rd Qu. : 0
Max. : 263
NA's : 681117

LateAircraftDelay
Min. : 0.0
1st Qu. : 0.0
Median : 8.0
Mean : 25.5
3rd Qu. : 33.0
Max. : 1437.0
NA's : 681117

FirstDepTime
Min. : 8
1st Qu. : 823
Median : 1206
Mean : 1250
3rd Qu. : 1644
Max. : 2357
NA's : 840187

TotalAddGTime
Min. :1.0
1st Qu. :16.0
Median :26.0
Mean :35.3
3rd Qu. :43.0
Max. :327.0
NA's :896647

LongestAddGTime
Min. : 1.0
1st Qu. : 16.0
Median : 26.0
Mean : 34.5
3rd Qu. : 42.0
Max. : 209.0
NA's : 896647

DivAirportLandings
Min. : 0.000000
1st Qu. : 0.000000
Median : 0.000000
Mean : 0.006115
3rd Qu. : 0.000000
Max. : 9.000000

DivReachedDest
Min. : 0.0
1st Qu. : 0.0
Median : 1.0
Mean : 0.6
3rd Qu. : 1.0
Max. : 1.0
NA's : 899932

DivActualElapsedTime
Min. : 90.0
1st Qu. : 254.0
Median : 322.0
Mean : 365.0
3rd Qu. : 429.5
Max. : 1968.0
NA's : 900944

DivArrDelay
Min. : 9.0
1st Qu. : 126.0
Median : 176.0
Mean : 218.2
3rd Qu. : 250.0
Max. : 1651.0
NA's : 900944

DivDistance
Min. : 0.0
1st Qu. : 0.0
Median : 0.0
Mean : 128.4
3rd Qu. : 109.0
Max. : 2556.0
NA's : 899944

Div1Airport
Length : 902551
Class : character
Mode : character

Div1AirportID
Min. : 10135
1st Qu. : 11292
Median : 12892
Mean : 12771
3rd Qu. : 14107
Max. : 16271
NA's : 899615

Div1AirportSeqID
Min. : 1013503
1st Qu. : 1129202
Median : 1289203
Mean : 1277086
3rd Qu. : 1410702
Max. : 1627102
NA's : 899615

Div1WheelsOn
Length : 902551
Class : character
Mode : character

Div1TotalGTime
Min. : 1.0
1st Qu. : 7.0
Median : 16.0
Mean : 23.6
3rd Qu. : 28.0
Max. : 219.0
NA's : 899615

Div1LongestGTime
Min. : 1.0
1st Qu. : 7.0
Median : 11.0
Mean : 19.3
3rd Qu. : 21.0
Max. : 173.0
NA's : 899615

Div1WheelsOff

Div1TailNum

Div2Airport

Div2AirportID

Div2AirportSeqID

```
Length  : 902551
Class   : character
Mode    : character
```

```
Length  : 902551
Class   : character
Mode    : character
```

```
Length  : 902551
Class   : character
Mode    : character
```

```
Min.    : 10397
1st Qu. : 10721
Median : 12302
Mean    : 12348
3rd Qu. : 13930
Max.    : 15096
NA's    : 902483
```

```
Min.    : 1039705
1st Qu. : 1072102
Median : 1230254
Mean    : 1234840
3rd Qu. : 1393003
Max.    : 1509602
NA's    : 902483
```

Div2WheelsOn

```
Length  : 902551
Class   : character
Mode    : character
```

Div2TotalGTime

```
Min.    : 2.0
1st Qu. : 5.0
Median : 8.0
Mean    : 10.5
3rd Qu. : 11.2
Max.    : 50.0
NA's    : 902483
```

Div2LongestGTime

```
Min.    : 2.0
1st Qu. : 5.0
Median : 8.0
Mean    : 9.9
3rd Qu. : 11.0
Max.    : 34.0
NA's    : 902483
```

Div2WheelsOff

```
Length  : 902551
Class   : character
Mode    : character
```

Div2TailNum

```
Length  : 902551
Class   : character
Mode    : character
```

Div3Airport

```
Length  : 902551
Class   : character
Mode    : character
```

Div3AirportID

```
Mode : logical
NA's : 902551
```

Div3AirportSeqID

```
Mode : logical
NA's : 902551
```

Div3WheelsOn

```
Length  : 902551
Class   : character
Mode    : character
```

Div3TotalGTime

```
Mode : logical
NA's : 902551
```

Div3LongestGTime

```
Mode : logical
NA's : 902551
```

Div3WheelsOff

```
Length  : 902551
Class   : character
Mode    : character
```

Div3TailNum

```
Length  : 902551
Class   : character
Mode    : character
```

Div4Airport

```
Length  : 902551
Class   : character
Mode    : character
```

Div4AirportID

```
Mode : logical
NA's : 902551
```

Div4AirportSeqID

```
Mode : logical
NA's : 902551
```

Div4WheelsOn

```
Mode : logical
NA's : 902551
```

Div4TotalGTime

```
Mode : logical
NA's : 902551
```

Div4LongestGTime

```
Mode : logical
NA's : 902551
```

Div4WheelsOff

```
Length  : 902551
Class   : character
Mode    : character
```

Div4TailNum

```
Length  : 902551
Class   : character
Mode    : character
```

Div5Airport

```
Length  : 902551
Class   : character
Mode    : character
```

Div5AirportID

```
Mode : logical
NA's : 902551
```

Div5AirportSeqID

```
Mode : logical
NA's : 902551
```

Div5WheelsOn

```
Length  : 902551
Class   : character
Mode    : character
```

Div5TotalGTime

```
Mode : logical
NA's : 902551
```

Div5LongestGTime

```
Mode : logical
NA's : 902551
```

Div5WheelsOff

```
Length  : 902551
Class   : character
Mode    : character
```

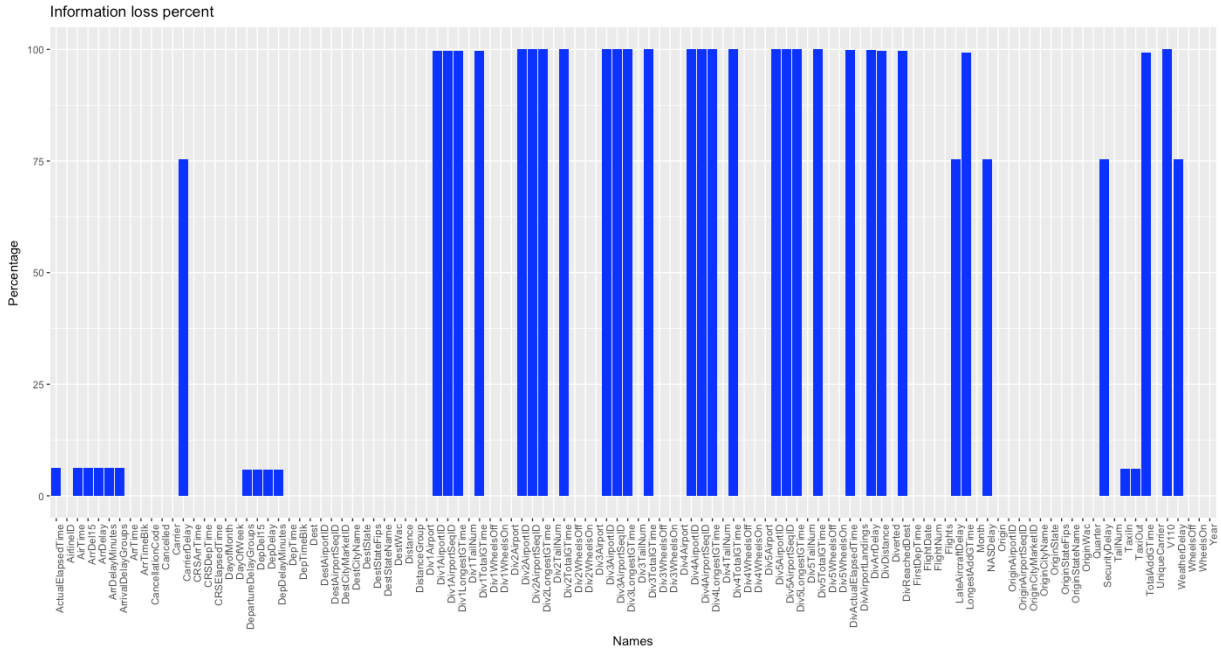
Div5TailNum

```
Length  : 902551
Class   : character
Mode    : character
```

V110

```
Mode : logical
NA's : 902551
```

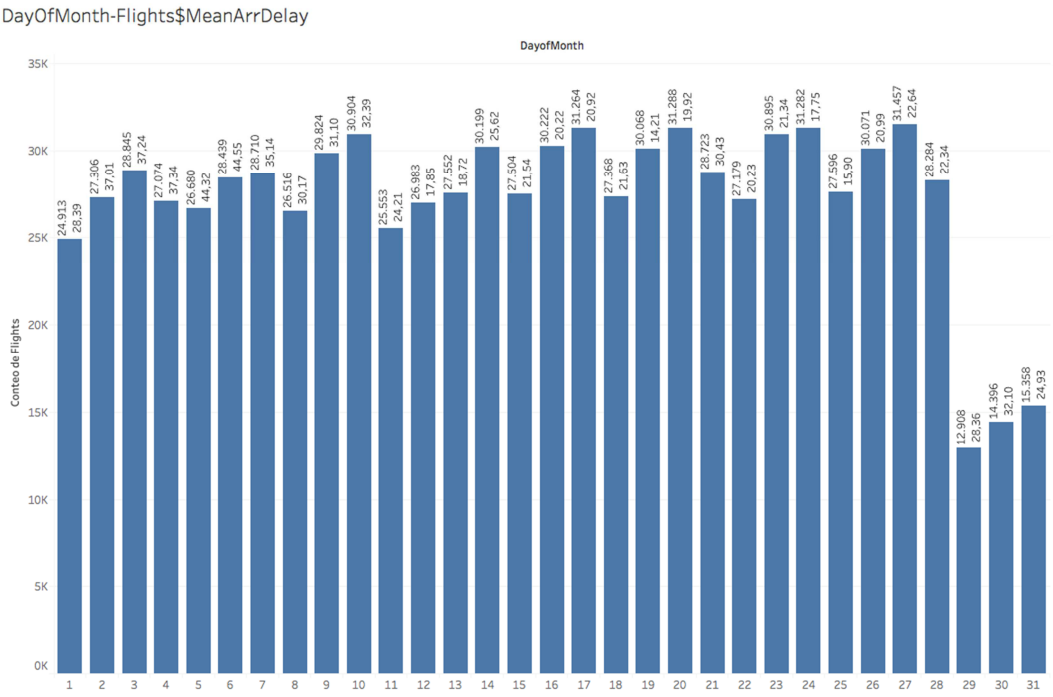
- Muchas de las variables contienen una elevada pérdida de información como se puede observar en la siguiente imagen:



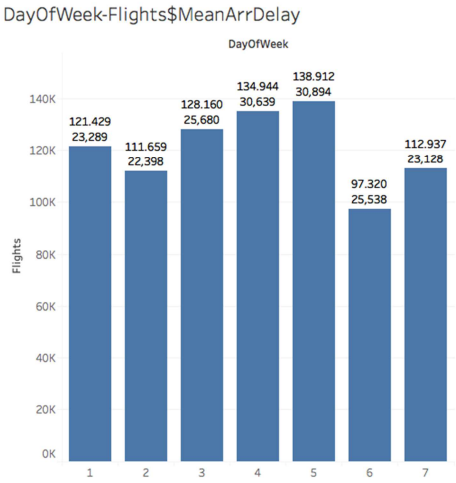
Se han eliminado todas aquellas que contienen >99% de pérdida de información:

	Names	Percent		Names	Percent		Names	Percent
1	TotalAddGTime	99.34585	10	DivArrDelay	99.82195	19	Div4AirportID	100.00000
2	LongestAddGTime	99.34585	11	Div2AirportID	99.99247	20	Div4AirportSeqID	100.00000
3	Div1AirportID	99.67470	12	Div2AirportSeqID	99.99247	21	Div4TotalGTime	100.00000
4	Div1AirportSeqID	99.67470	13	Div2TotalGTime	99.99247	22	Div4LongestGTime	100.00000
5	Div1TotalGTime	99.67470	14	Div2LongestGTime	99.99247	23	Div5AirportID	100.00000
6	Div1LongestGTime	99.67470	15	Div3AirportID	100.00000	24	Div5AirportSeqID	100.00000
7	DivReachedDest	99.70982	16	Div3AirportSeqID	100.00000	25	Div5TotalGTime	100.00000
8	DivDistance	99.71115	17	Div3TotalGTime	100.00000	26	Div5LongestGTime	100.00000
9	DivActualElapsedTime	99.82195	18	Div3LongestGTime	100.00000	27	V110	100.00000

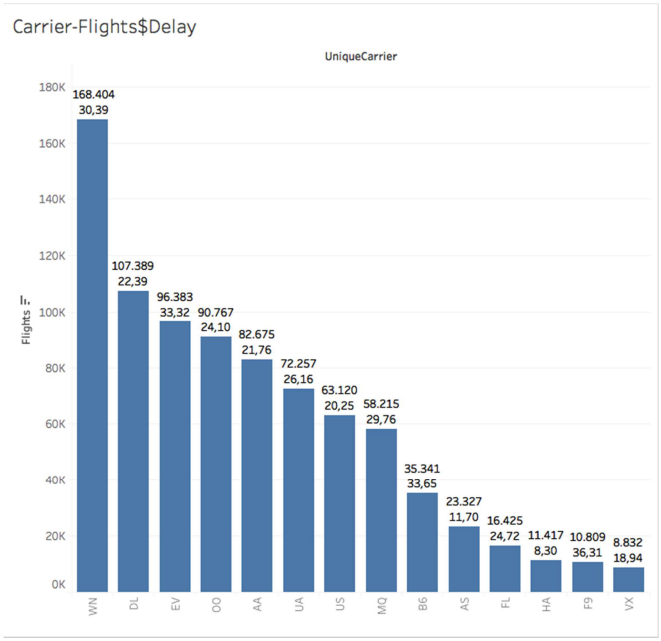
- Existen 54571 vuelos cancelados que han sido eliminados puesto que son operaciones no realizadas. También se han eliminado sus correspondientes variables: Cancelled y CancellationCode por no aportar información para el análisis.
- Existen 2619 vuelos desviados que han sido eliminados. Al tratarse de vuelos no programados (puesto que no aterrizan en el destino programado), no registran el retraso y por tanto no son de utilidad para el análisis. También se han eliminado sus correspondientes variables: Diverted, DivAirportLandings, Div1Airport, Div1WheelsOn, Div1WheelsOff, Div1TailNum, Div2Airport, Div2WheelsOn, Div2WheelsOff, Div2TailNum, Div3Airport, Div3WheelsOn, Div3WheelsOff, Div3TailNum, Div4Airport, Div4WheelsOn, Div4WheelsOff, Div4TailNum, Div5Airport, Div5WheelsOn, Div5WheelsOff y Div5TailNum por no aportar información para el análisis.
- La variable Quarter contiene el cuarto de año al que corresponde el mes, al no ser relevante para el análisis, se ha eliminado.
- Cada UniqueCarrier tiene un único AirlineID asociado, por tanto se ha eliminado esta última.
- La variable Carrier contiene la misma información que UniqueCarrier, por tanto se ha eliminado.
- Cada Origin tiene un único OriginAirportID asociado, por tanto se ha eliminado esta última.
- Cada OriginState tiene un único OriginStateFips asociado, por tanto se ha eliminado esta última.
- Cada OriginState tiene un único OriginWac asociado, por tanto se ha eliminado esta última.
- Cada Dest tienen un único DestAirportID asociado, por tanto se ha eliminado esta última.
- Cada DestState tiene un único DestStateFips asociado, por tanto he eliminado esta última.
- FirstDepTime tienen una pérdida de información del 99.38795% >99%, después de haber eliminado los vuelos cancelados y desviados, por tanto se ha eliminado.
- Cada DestState tiene un único DestWac asociado, por tanto se esta última.
- Representando algunos de estos datos en Tableu se puede ver:
    - La distribución de los vuelos a lo largo de los días del mes es más o menos uniforme exceptuando los últimos días del mes (esto puede ser debido a que la muestra únicamente tiene los meses de Enero y Febrero), se indica también el porcentaje medio de los vuelos retrasados de cada uno de los días:



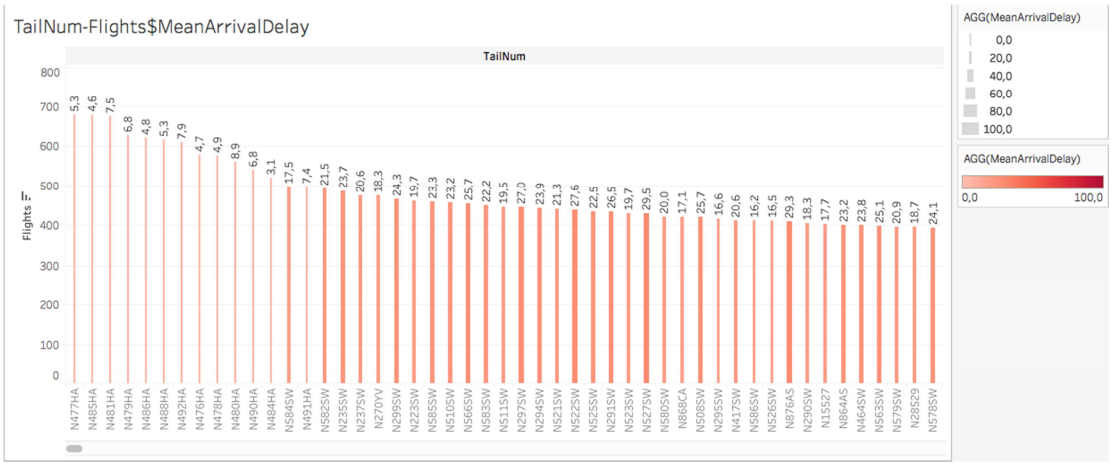
- La distribución de los vuelos a lo largo de la semana es más alta entre semana que el fin de semana, también se indica el porcentaje medio de los vuelos retrasados:



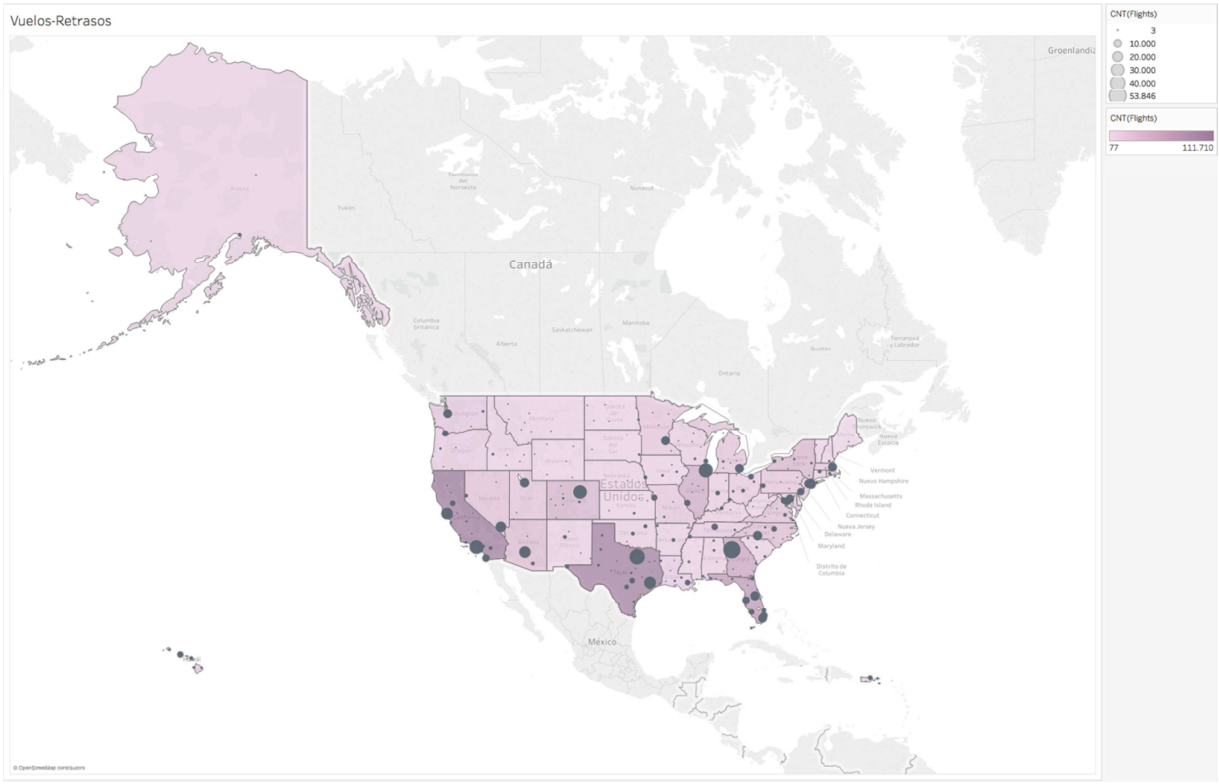
- En la distribución de vuelos por compañía se observa que la compañía WN (Southwest Airlines), DL(Delta Airlines) y EV(Express Jet) son las compañías que más vuelos realizan pero las compañías F9 (Frontier Airlines), B6(JetBlue Airways) y EV son las que mayor porcentaje medio de vuelos retrasados tienen:



- En la siguiente distribución se han ordenado las matrículas por el número de vuelos realizados y el color y tamaño de las barras indican el porcentaje medio del número de vuelos retrasados de esta:

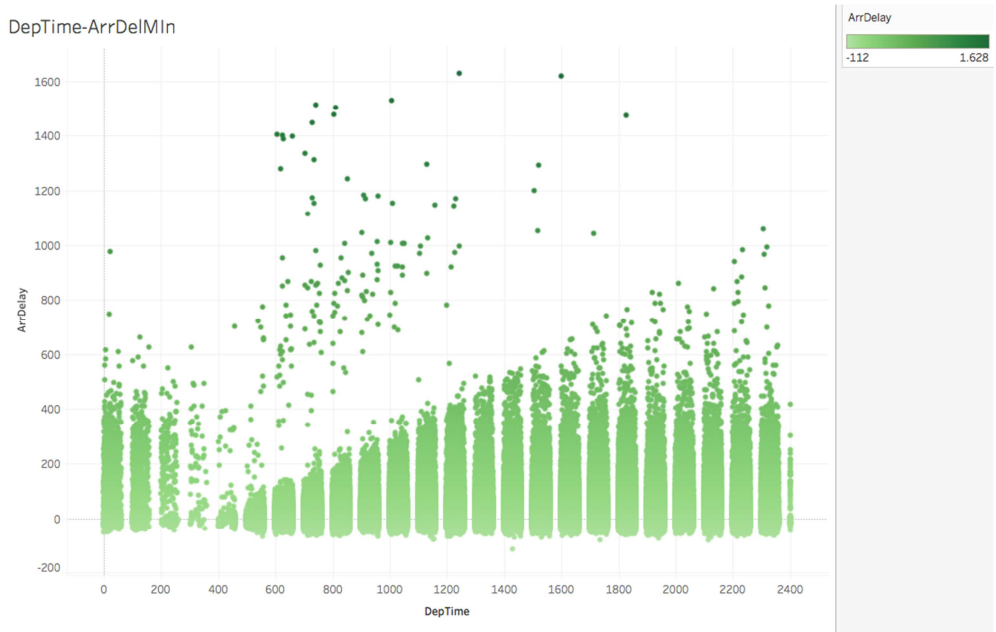


- A través del siguiente mapa, se puede observar que los aeropuertos con más vuelos son: ATL-Aeropuerto Internacional Hartsfield-Jackson de Atlanta, Georgia; DFW-Dallas/Fort Worth Interntional Airport de Dallas y Fort Worth, Texas; y ORD-Aeropuerto Internacional O'Hare de Chicago, Ilinois:

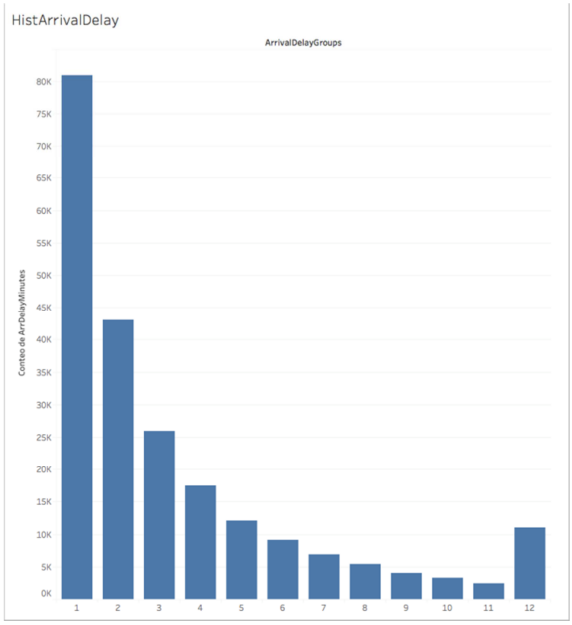


También que los estados con más vuelos son: CA-California, TX-Texas y FL-Florida.

- En la siguiente distribución se puede observar la relación entre la hora de salida el vuelo y el tiempo de retraso en la llegada.



- En este último, se puede observar un histograma de los vuelos que tienen más de 15 minutos de retraso (que son los vuelos considerados como retrasados) clasificado por el grupo de retraso:



Los intervalos de retraso van desde el 1 (15 minutos de retraso) hasta el 12 (>180 minutos de retraso).

Nos interesa conocer cómo es la distribución de los vuelos con más de 180 minutos de retraso, puesto que estos nos pueden modificar la media en el caso de los outliers.

3. Metodologías

Se ha dado respuesta a dos problemas:

1. Predicción de vuelos retrasados en la llegada
2. Tiempo medio de retraso.

Los pasos seguidos en orden para el tratamiento de los datos están incluidos en los siguientes archivos:

- 1_FileDataProcessing.R: En el cual se realiza la lectura de los ficheros, una primera limpieza, análisis y tratamiento de los datos.

  Se eliminan todas aquellas variables que contienen una pérdida de información de más de 99%.

  Se analizan los datos y se eliminan los vuelos cancelados por ser vuelos no operados y los vuelos desviados por no ser vuelos programados y no contener información para el dataset a analizar.

  Se tratan las variables dándolas su correspondiente formato.

  Se hace una copia de dataset (flightsDileDataProcessing) y se guarda el resultado en un fichero flights.csv.

- 2_InputData.R: En el cuál se realiza un segundo análisis del dataset. Se eliminan las variables que no tienen relevancia para el problema a resolver y se toma la decisión de qué hacer con aquellas variables que contienen NA's (pérdida de información).

  Se hace una copia del dataset (finalflights) y se guarda el resultado en un fichero finalFlights.csv.

3.1. Predicción de vuelos retrasados en la llegada:

La solución para este problema de clasificación se encuentra en el archivo: 3_ClassificationModels.R.

Para solucionar este problema de clasificación, me he centrado en la variable ArrDel15 con valor 1, que es la que indica aquellos vuelos que tienen retraso (mayor que 15 minutos) en la llegada.

Para realizar el estudio, he decidido asignar pesos a las variables categóricas y normalizar el resto de variables.

Una vez se han asignado los pesos a las variables categóricas y se han normalizado el resto de variables he guardado el dataset resultante flightsWeights en un fichero flightsWeightsClassi.csv.

He utilizado los modelos de Regresión Logística y Random Forest para determinar cuáles son las variables que más influyen en el retraso de los vuelos. Y la Matriz de Confusión y Precisión/Recall para determinar qué modelo es el que da las mejores predicciones y realiza la mejor clasificación.

3.2. Tiempo medio de retraso:

La solución para este problema de regresión se encuentra en el archivo: 4_RegressionModels.R.

He generado un nuevo dataset flightsDelay que contiene únicamente los vuelos retrasados (ArrDelay = 1).

He analizado los outliers que existen por encima del percentil 95 y he prescindido de ellos para el estudio.

He asignado pesos a las variables categóricas y posteriormente las he normalizado.

Una vez asignados los pesos y normalizadas todas las variables, he guardado el dataset resultante flightsDelayWeights en un fichero flightsWeightsRegression.csv.

He utilizado los modelos de Regresión Lineal y Random Forest con el fin de encontrar el modelo que tenga el error más bajo y prediga mejor el tiempo de retraso. Y las métricas MAPE y RMSE para decidir qué modelo reporta mejor resultado.

4. Conclusiones

Las variables que más influyen en el retraso de los vuelos de llegada (en orden de mayor a menor influencia) son:

- El retraso de salida (DepDelay).
- La hora de salida (DepTime).
- La matrícula (TailNum).
- El número de vuelo (FlightNum).
- El día del mes (DayofMonth).
- El aeropuerto de destino (DestAirportSeqID).
- El estado de origen (OriginState).
- El día de la semana (DayofMonth).
- La distancia (DistanceGroup).
- La compañía (UniqueCarrier).

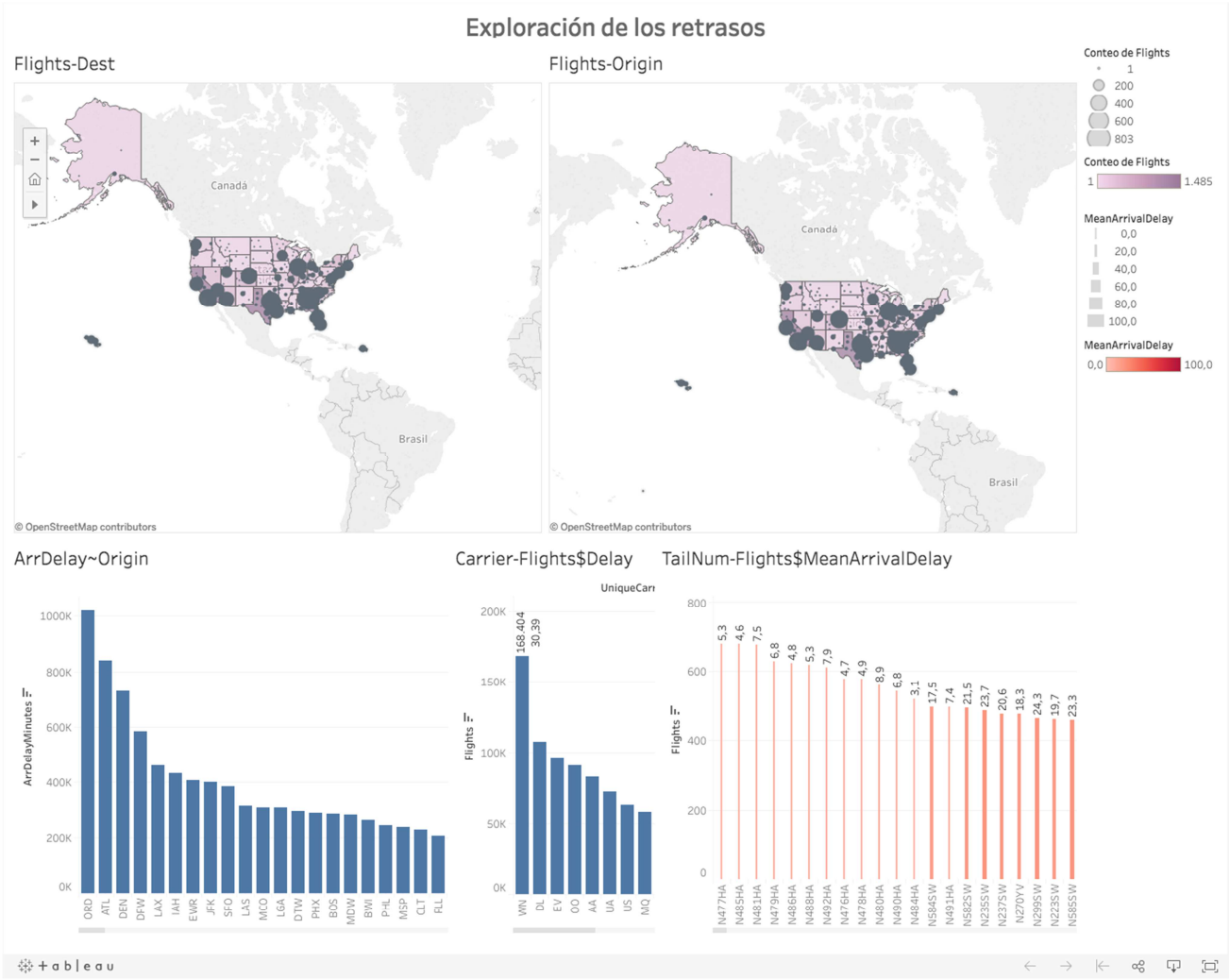
A la hora de predecir el tiempo medio de retraso de un vuelo, la media de porcentaje de error es del 26%.

Futuras mejoras que se pueden realizar sobre el análisis y que no se han llevado a cabo por falta de tiempo:

- Añadir las condiciones meteorológicas y el impacto de estas.
- Analizar cada uno de los grupos de retrasos y la repercusión económica.

5. Dashboard

https://public.tableau.com/views/FlightGraphs/Dashboard2?:embed=y&:display_count=yes



6. Bibliografía:

https://www.transtats.bts.gov/DL_SelectFields.asp

https://srcole.github.io/2017/04/02/flight_delay/

https://es.hortonworks.com/blog/data-science-apacheh-hadoop-predicting-airline-delays/

https://ddd.uab.cat/pub/tfg/2016/169883/MartinezDomenechNerea-TFGAa2015-16.pdf

https://ddd.uab.cat/pub/tfg/2015/146885/MonjeSolaRaul-TFGAa2014-15.pdf

https://www.flyertalk.com/forum/travelbuzz/1202036-how-actual-arrival-time-flight-determined.html

https://github.com/semartin3/TFM

https://stackoverflow.com/questions/25272457/convert-an-integer-column-to-time-hhmm

https://stackoverflow.com/questions/7980622/subset-of-rows-containing-na-missing-values-in-a-chosen-column-of-a-data-frame

http://data.library.virginia.edu/working-with-dates-and-time-in-r-using-the-lubridate-package/

https://www.transtats.bts.gov/HomeDrillChart.asp

http://www.airlineupdate.com/content_public/codes/airlinecodes/iatacodes/iata-f.htm

https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function