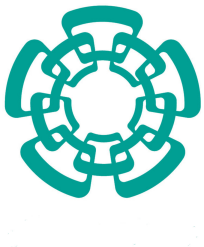
Ejercicio 5: Aplicación de técnicas de agrupación a dataset del mundo real



Análisis de datos

Profesor: Iván López Arévalo,

Rafael Pérez Torres

LTI CINVESTAV Tamaulipas

1 Introducción

Las técnicas de aprendizaje no supervisado permiten descubrir información en los datos sin tener información *a priori* de los mismos. Dichas técnicas han permitido la obtención de conocimiento a simple vista oculto en enormes datasets de información proveniente de la industria y de los ámbitos empresarial y científico.

Dos de las grandes familias de algoritmos para generar grupos (clústers) a partir de datos son las técnicas jerárquicas y particionales. Las técnicas jerárquicas buscan hacer divisiones incrementales en los datos, creando nuevos niveles con grupos específicos en cada iteración. Por otro lado, las técnicas particionales buscan dividir el espacio de objetos en áreas que se encentran representadas por un punto central (prototipo) denominado centroide.

En este documento se muestran los resultados del análisis de un dataset de vuelos comprendidos durante el año 2002 en Estados Unidos, así como las observaciones encontradas al utilizar técnicas de clústering jerárquico (*híbrido*) y particional (*kmeans*) utilizando la suite ***R***.

2 Preprocesamiento

2.1 Reducción de dimensionalidad

A partir de una primera inspección de los datos se eliminaron algunas de las columnas cuyo valor se mantenía constante, se encontraba ausente en todas las instancias, o que no aportaba valor alguno para el descubrimiento de información. A través de instrucciones como las que se muestran en el Listing 1, fue posible realizar la remoción de dichas columnas.

```
dsCrudo$CarrierDelay <- NULL
dsCrudo$WeatherDelay <- NULL
dsCrudo$NASDelay <- NULL
dsCrudo$SecurityDelay <- NULL
dsCrudo$LateAircraftDelay <- NULL
dsCrudo$Year <- NULL
```

Listing 1: Eliminación de columnas con valores constantes

Después de realizar una segunda vista rápida a los datos y a las descripciones de cada uno de los atributos, se determinó realizar la separación de los datos en tres familias: vuelos cancelados, vuelos desviados y vuelos con la información completa.

La determinación de los vuelos cancelados se realizó a través del código mostrado en el Listing 2.

```
dsVuelosCancelados <- ( subset(dsCrudo, dsCrudo$Cancelled == 1) )
```

Listing 2: Detección de vuelos cancelados

La determinación de los vuelos desviados se realizó a través del código mostrado en el Listing 3:

```
dsVuelosDesviados <- ( subset(dsCrudo, dsCrudo$Diverted == 1) )
```

Listing 3: Detección de vuelos desviados

Finalmente, la determinación de los vuelos con información de horarios completa fue determinada de forma manual, omitiendo aquellos registros que carecieran de algún valor en las columnas con componentes de fechas, tal como se muestra en el Listing 4.

```
dsTiemposCompletos <- (subset(dsCrudo, !is.na(dsCrudo$DepTime)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    ArrTime)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    CRSArrTime)))
```

```
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    CRSDepTime)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    ActualElapsedTime)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    AirTime)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    DepDelay)))
dsTiemposCompletos <- (subset(dsTiemposCompletos, !is.na(dsTiemposCompletos$
    ArrDelay)))
```

Listing 4: Detección de vuelos con información de horarios completa

2.2 Tratamiento de los registros con información de horarios completa

Los registros con la información de horarios completa fueron los únicos considerados para ingresar en la etapa de descubrimiento de información (clústering). Su preparación fue realizada a través de las actividades descritas a continuación.

2.2.1 Separación por mes

Con la intención de realizar un análisis y muestreo estratificado que permitiera reducir la cantidad de instancias a analizar, los datos se separaron en base al mes de la fecha de partida del viaje. Instrucciones similares a la mostrada en el Listing 5 permitieron realizar la separación.

```
dsEnero <- subset(dsTiemposCompletos, dsTiemposCompletos$Month == 1)
```

Listing 5: Separación de registros de vuelos por mes

Un caso especial fue el de los registros pertenecientes al mes de abril, ya que algunos de los valores de las fechas no podían ser transformados a su representación en milisegundos (más información de esta transformación es encontrada en las secciones posteriores del documento). Por ello, se omitieron algunas de estas instancias (44 en total) si fueron registradas con una fecha del domingo 7 de abril y en el rango de las 02:00 horas. Dicha discriminación fue realizada a través del código mostrado en el Listing 6.

```
dsAbril <- subset(dsTiemposCompletos, dsTiemposCompletos$Month == 4)
dsAbrilProbablesConflictivos <- subset( dsAbril, dsAbril$DayofMonth == 7)
dsAbrilConflictivosHraSalida <-
  subset( dsAbrilProbablesConflictivos, dsAbrilProbablesConflictivos$DepTime
    >= 200 & dsAbrilProbablesConflictivos$DepTime < 300)
```

```
dsAbrilConflictivosHraEstimadaSalida <-
  subset( dsAbrilProbablesConflictivos, dsAbrilProbablesConflictivos$
    CRSDepTime >= 200 & dsAbrilProbablesConflictivos$CRSDepTime < 300)
dsAbrilConflictivosHraLlegada <-
  subset( dsAbrilProbablesConflictivos, dsAbrilProbablesConflictivos$ArrTime
    >= 200 & dsAbrilProbablesConflictivos$ArrTime < 300)
dsAbrilConflictivosHraEstimadaLlegada <-
  subset( dsAbrilProbablesConflictivos, dsAbrilProbablesConflictivos$
    CRSArrTime >= 200 & dsAbrilProbablesConflictivos$CRSArrTime < 300)
dsAbrilConflictivos <- rbind(dsAbrilConflictivosHraEstimadaLlegada,
    dsAbrilConflictivosHraEstimadaSalida, dsAbrilConflictivosHraSalida,
    dsAbrilConflictivosHraLlegada)
diferencia <- dsAbril[!duplicated(rbind(dsAbrilConflictivos, dsAbril))[-seq_
    len(nrow(dsAbrilConflictivos))], ]
dsAbril <- diferencia
```

Listing 6: Separación de registros del mes de abril

2.2.2 Muestreo aleatorio

Se realizó un submuestro aleatorio de las instancias de cada uno de los meses. Para el caso del clústering particional (*skmeans*) se optó por un submuestro del 5% del total de los registros. Sin embargo, esta cantidad resultaba demasiado alta para el algoritmo jerárquico (el equipo de cómputo utilizado agotaba su memoria) por lo que para este caso se utilizó únicamente el 1%. A través de instrucciones similares a las mostradas en el Listing 7 se realizó este muestreo.

```
dsEnero <- dsEnero[sample(1:nrow(dsEnero), as.integer(length(dsEnero$Month) *
    tamanyoMuestra), replace=FALSE),]
```

Listing 7: Muestreo aleatorio de registros por mes

Cálculo de las fechas en formato de milisegundos Debido a que la información de fechas estimadas y reales tanto de llegada como de salida de los vuelos se encontraba dividida, se procedió a realizar su unificación en un solo atributo *sintético* que representa la cantidad de milisegundos transcurridos entre cada una de las fechas en cuestión y el 1 de enero de 1970 (conocido como ***the epoch***).

Al final, la intención es reducir la cantidad de columnas y mantener una variable continua que permitiera ser utilizada por los algoritmos. La función mostrada en el Listing 8. Permite realizar el cálculo de las representaciones de las fechas en milisegundos.

```r
calcularMsHoras <- function (ds){
  parteFecha <- sprintf("2002-%02d-%02d", as.numeric(ds$Month), as.numeric(ds$
    DayofMonth))

  horaSalida <- sprintf("%04d", ds$DepTime)
  horaLlegada <- sprintf("%04d", ds$ArrTime)
  horaEstimadaSalida <- sprintf("%04d", ds$CRSDepTime)
  horaEstimadaLlegada <- sprintf("%04d", ds$CRSArrTime)

  # Hora real salida
  fechaSalida <- sprintf("%s %s:%s:00",parteFecha, substr(horaSalida, 1, 2),
    substr(horaSalida, 3, 4))
  msHoraSalida <- as.integer( ymd_hms(fechaSalida, tz = "America/Mexico_City")
    )

  # Hora real llegada
  fechaLlegada <- sprintf("%s %s:%s:00",parteFecha, substr(horaLlegada, 1, 2),
     substr(horaLlegada, 3, 4))
  msHoraLlegada <- as.integer( ymd_hms(fechaLlegada, tz = "America/Mexico_City
    "))

  # Hora estimada salida
  fechaEstimadaSalida <- sprintf("%s %s:%s:00",parteFecha, substr(
    horaEstimadaSalida, 1, 2), substr(horaEstimadaSalida, 3, 4))
  msHoraEstimadaSalida <- as.integer( ymd_hms(fechaEstimadaSalida, tz = "
    America/Mexico_City"))

  # Hora estimada llegada
  fechaEstimadaLlegada <- sprintf("%s %s:%s:00",parteFecha, substr(
    horaEstimadaLlegada, 1, 2), substr(horaEstimadaLlegada, 3, 4))
  msHoraEstimadaLlegada <- as.integer( ymd_hms(fechaEstimadaLlegada, tz = "
    America/Mexico_City"))

  mss <- list("msHoraEstimadaSalida" = msHoraEstimadaSalida,
              "msHoraSalida" = msHoraSalida,
              "msHoraEstimadaLlegada" = msHoraEstimadaLlegada,
              "msHoraLlegada" = msHoraLlegada)
  mss
}
```

Listing 8: Transformación de las horas (timestamps) a milisegundos

2.2.3 Eliminación de información superflua y/o redundante

Debido a los cambios introducidos por el cálculo de milisegundos, algunas de las columnas pierden su razón de ser y se procedió a eliminarlas de los datasets. Por ejemplo, los datos señalando al mes, día del mes y a las fechas en particular son sustituidos por las columnas creadas.

Adicionalmente, datos como los que indican si el vuelo ha sido cancelado o desviado fueron también eliminados, ya que los registros de estos datasets cuentan con la información de horarios completa (recordando que los registros de vuelos desviados o cancelados ya fueron separados al inicio del pre-procesamiento).

Es importante destacar que otras columnas como el número de vuelo, el identificador del carrier y del avión también fueron eliminadas dado que representan identificadores que, según la perspectiva seguida, no aportan mayor información al proceso.

Finalmente, se realizó la detección de información redundante a través del cálculo de correlaciones. La Figura 2.1 muestra las correlaciones de uno de los datasets (se observó comportamiento similar en el resto de datasets). Puede observarse que existe gran correlación entre los datos que indican información temporal, lo cual es una situación esperada ya que, por ejemplo, el crecimiento en la hora de partida de un avión incrementa también la hora de llegada. Otro ejemplo evidente es lo que sucede entre las variables de tiempo total de vuelo y distancia recorrida, que tienen una correlación amplia debido a que mientras más tiempo tenga el avión en vuelo, mayor distancia habrá recorrido.

Sin embargo, no se detectó alguna correlación entre el resto de los atributos. Las relaciones entre cuestiones temporales dominan este dataset.

2.2.4 Conversión de valores nominales a numéricos

En una de las experimentaciones se consideró que la información referente al origen y al destino podría tener algún grado de importancia, por lo que fueron transformadas de un tipo factor a uno numérico para su inclusión en el proceso de clasificación. Esto fue realizado a través de instrucciones similares a las mostradas en el Listing 9:

```
1 dsEnero$Origin <- as.integer(factor(dsEnero$Origin))
2 dsEnero$Dest  <- as.integer(factor(dsEnero$Dest))
```

Listing 9: Conversión de valores nominales (factor) a numérico

La conversión a valor numérico de los atributos nominales permitió realizar la ejecución del clústering particional. Sin embargo, los resultados arrojados por el clústering contenían alta dimensionalidad y estos atributos al ser identificadores fueron candidatos

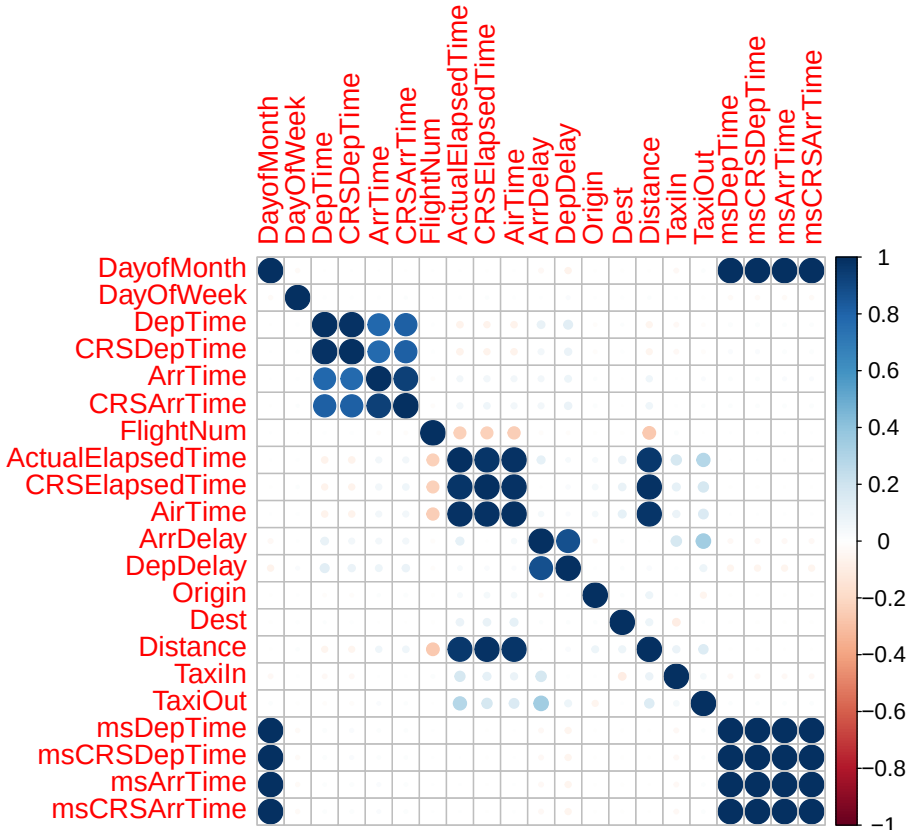


Figure 2.1: Correlaciones existentes en el dataset (enero)

a ser eliminados para la ejecución final.

Al final de estas tareas, se obtuvo la lista definitiva de atributos a considerar para el proceso de análisis de componentes principales (*PCA*), indicada en la Tabla 2.1.

Columna	**Significado**
DayOfWeek	Día de la semana
AirTime	Tiempo de vuelo
ArrDelay	Atraso en llegada
DepDelay	Atraso en salida
TaxiIn	Tiempo en hangar entrada
TaxiOut	Tiempo en hangar salida

Table 2.1: Conjunto final de columnas consideradas

2.2.5 Análisis de componentes principales (*PCA*)

Adicionalmente a los pasos anteriores, se implementó como etapa previa al clústering el análisis de componentes principales (*PCA*, por sus siglas en inglés). El proceso de *PCA* fue alimentado con los registros de los meses con sus datos escalados. De los componentes principales generados, originalmente se habían elegido cinco para realizar el proceso de clústering. Sin embargo esta dimensionalidad complicaba el análisis de los resultados por lo que se eligió un valor final de 3 componentes.

Al final de estas tareas, se obtuvieron doce datasets con un promedio de 21,656 registros y 6 columnas para el clústering particional y X registros y también 6 columnas para el clústering jerárquico; originalmente, dichos datasets tenían 5,271,359 registros y 23 columnas.

3 Análisis de datos

3.1 En base a frecuencias

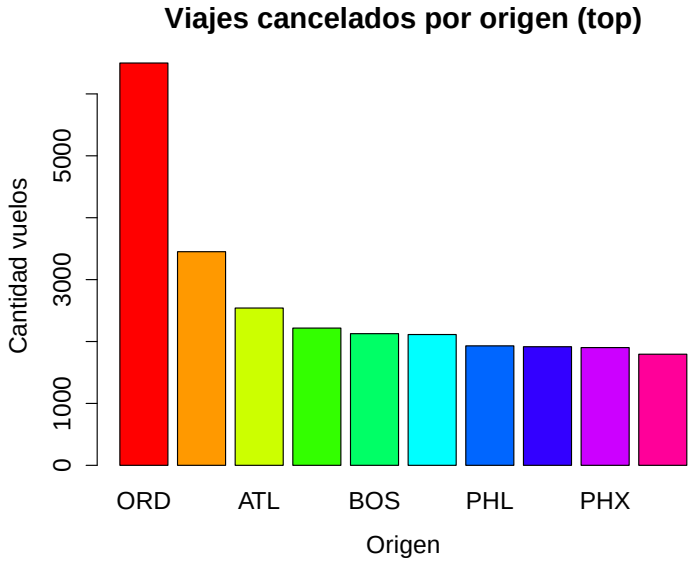
En esta sección se describe el análisis realizado sobre los datos utilizando únicamente información relacionada a la frecuencia de los mismos. El análisis fue desarrollado de forma individual por cada uno de los tres grandes grupos de datos: vuelos cancelados, vuelos desviados y vuelos con información de horarios completa.

3.1.1 Análisis de vuelos cancelados

Orígenes con la mayor cantidad de vuelos cancelados Representa a aquellos aeropuertos que, como punto de partida, cancelaron la mayor cantidad de vuelos. Se muestran en la Figura 3.1.

En este apartado y en el de destinos con la mayor cantidad de vuelos cancelados llama la atención que el aeropuerto de la ciudad de Chicago (ORD) canceló la mayor cantidad de vuelos. Haciendo un análisis a un nivel inferior se detectó que en el mes de junio se llevó a cabo la mayor cantidad de cancelaciones. Sin embargo, después de una búsqueda rápida en la web, no se detectó algún evento histórico (fenómeno meteorológico, social, etc.) como causa aparente.

Destinos con la mayor cantidad de vuelos cancelados Representa a aquellos aeropuertos que, como punto de llegada, cancelaron la mayor cantidad de vuelos. Se muestran en la Figura 3.2

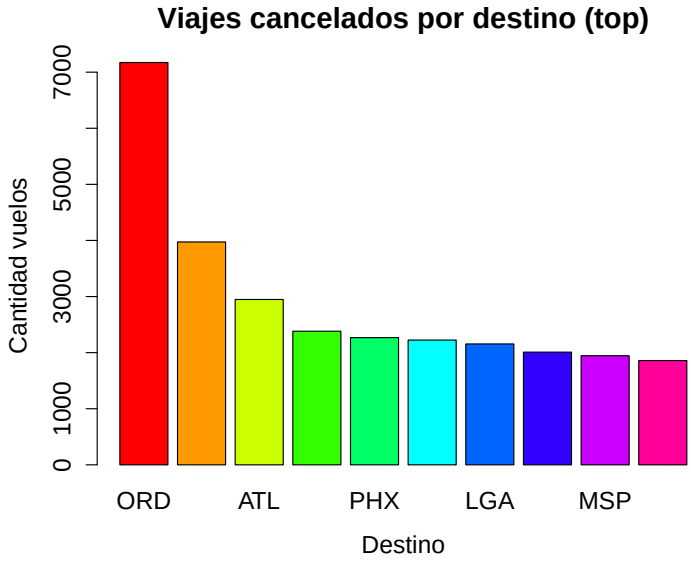


Aeropuerto	Cantidad
ORD	6499
DFW	3451
ATL	2541
DTW	2217
BOX	2127
LGA	2113
PHL	1930
LAX	1916
PHX	1902
MSP	1795

(a) Vuelos cancelados por origen

(b) Datos asociados

Figure 3.1: Resumen de vuelos cancelados por origen

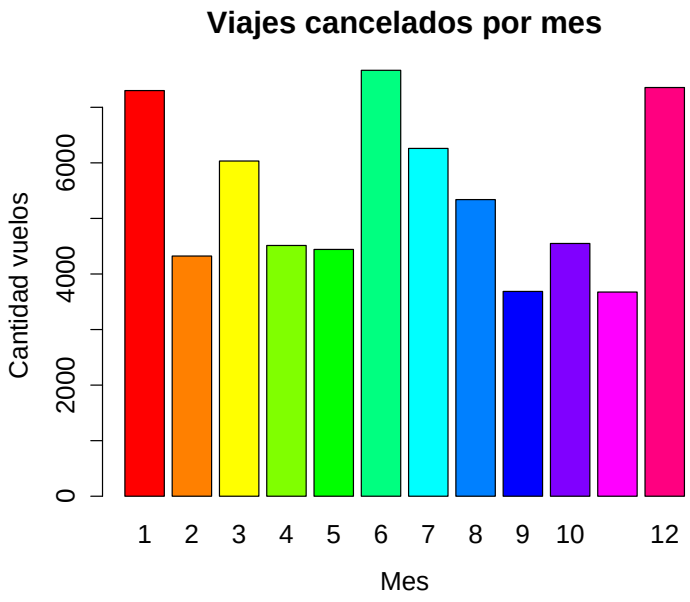


Aeropuerto	Cantidad
ORD	7172
DFW	3973
ATL	2948
DTW	2382
BOX	2268
LGA	2224
PHL	2155
LAX	2009
PHX	1945
MSP	1859

(a) Vuelos cancelados por destino

(b) Datos asociados

Figure 3.2: Resumen de vuelos cancelados por destino



(a) Vuelos cancelados por mes

Mes	Cantidad
Junio	7666
Diciembre	7356
Enero	7301
Julio	6260
Marzo	6033
Agosto	5339
Octubre	4549
Abril	4513
Mayo	4442
Febrero	4323
Septiembre	3686
Noviembre	3675

(b) Datos asociados

Figure 3.3: Resumen de vuelos cancelados por mes

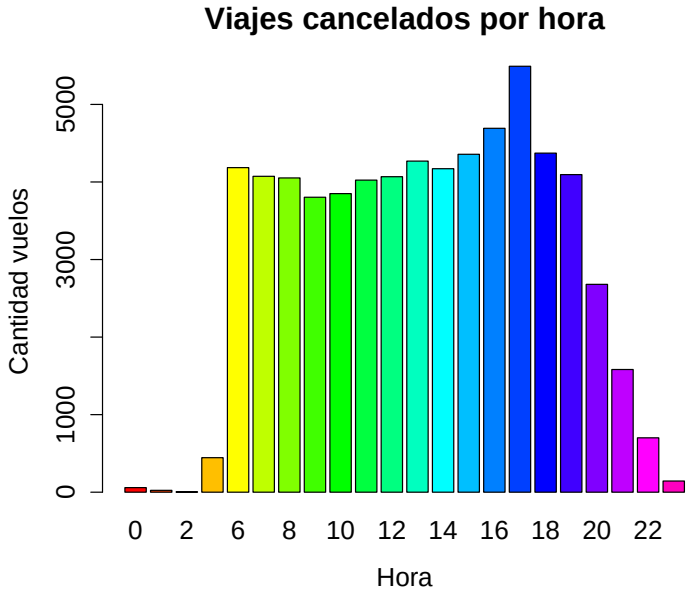
Vuelos cancelados por mes Cantidad de vuelos cancelados en cada uno de los meses. Se aprecia que en el mes de junio, diciembre y enero ocurre la mayor cantidad de vuelos cancelados. Diciembre y enero pueden comprenderse por cuestiones climatológicas, sin embargo el mes de junio representa una incógnita, siendo Chicago la ciudad que más contribuyó con las cancelaciones de vuelos. Se muestran en la Figura 3.3.

Vuelos cancelados por hora Cantidad de vuelos cancelados por hora del día. Se aprecia que alrededor de las 17:00 horas es cuando más vuelos se cancelan. Se muestran en la Figura 3.4.

Vuelos cancelados por aerolínea Indica aquellas aerolíneas que más vuelos cancelan. Se muestran en la Figura 3.5.

3.1.2 Análisis de vuelos desviados

Orígenes con la mayor cantidad de vuelos desviados Representa a aquellos aeropuertos que, como punto de partida, desviaron la mayor cantidad de vuelos. Se muestran en la Figura 3.6.



(a) Vuelos cancelados por hora

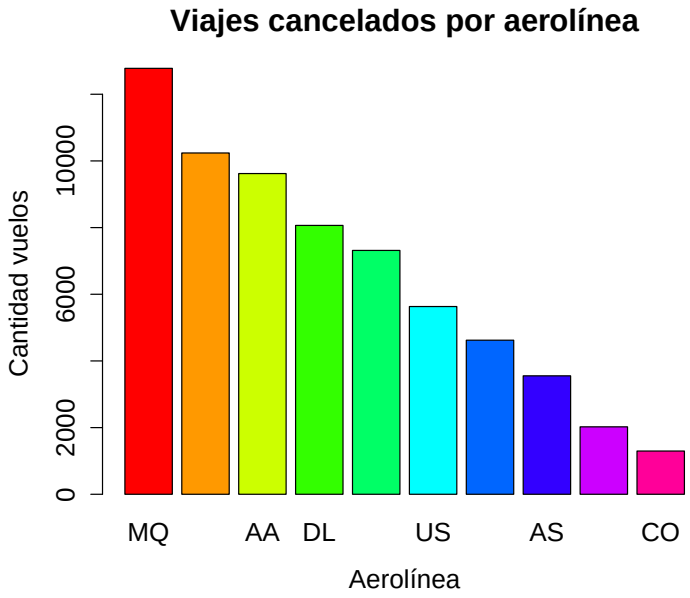
Hora	Cantidad
0	59
1	24
2	5
5	445
6	4185
7	4073
8	4052
9	3803
10	3850
11	4024
12	4068
13	4269
14	4171
15	4357
16	4692
17	5492
18	4372
19	4095
20	2680
21	1582
22	701
23	144

(b) Datos asociados

Figure 3.4: Resumen de vuelos cancelados por hora

Destinos con la mayor cantidad de vuelos desviados Representa a aquellos aeropuertos que, como punto de llegada, desviaron la mayor cantidad de vuelos. Se muestran en la Figura 3.7.

Vuelos desviados por mes Cantidad de vuelos desviados en cada uno de los meses. Se muestran en la Figura 3.8.

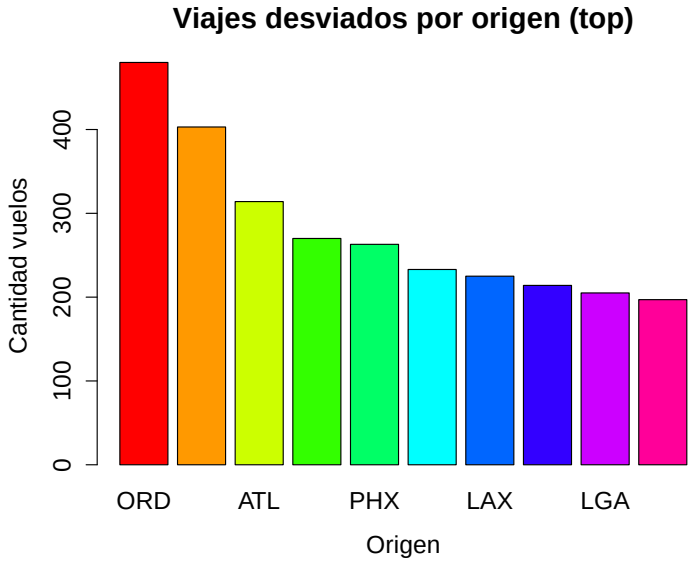


Aerolínea	Cantidad
MQ	12777
WN	10238
AA	9621
DL	8065
NW	7315
US	5632
UA	4622
AS	3553
HP	2023
CO	1297

(a) Vuelos cancelados por aerolínea

(b) Datos asociados

Figure 3.5: Resumen de vuelos cancelados por aerolínea

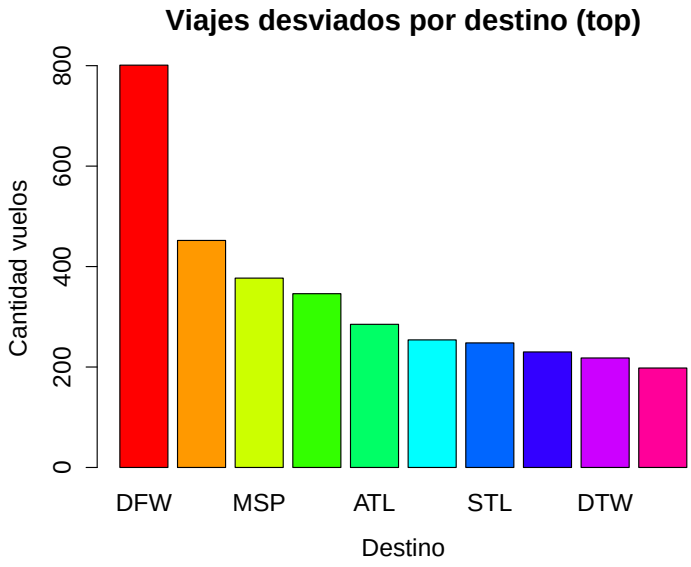


Aeropuerto	Cantidad
ORD	480
DFW	403
ATL	314
DTW	270
PHX	263
MSP	233
LAX	225
LAS	214
LGA	205
SEA	197

(a) Vuelos desviados por origen

(b) Datos asociados

Figure 3.6: Resumen de vuelos desviados por origen

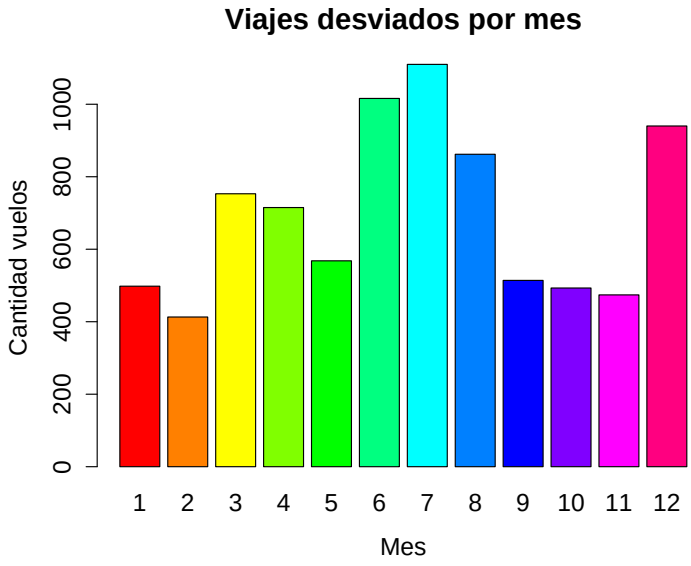


(a) Vuelos desviados por destino

Aeropuerto	Cantidad
DFW	801
ORD	452
MSP	377
IAH	346
ATL	285
LGA	254
STL	248
PHL	230
DTW	218
SFO	198

(b) Datos asociados

Figure 3.7: Resumen de vuelos desviados por destino

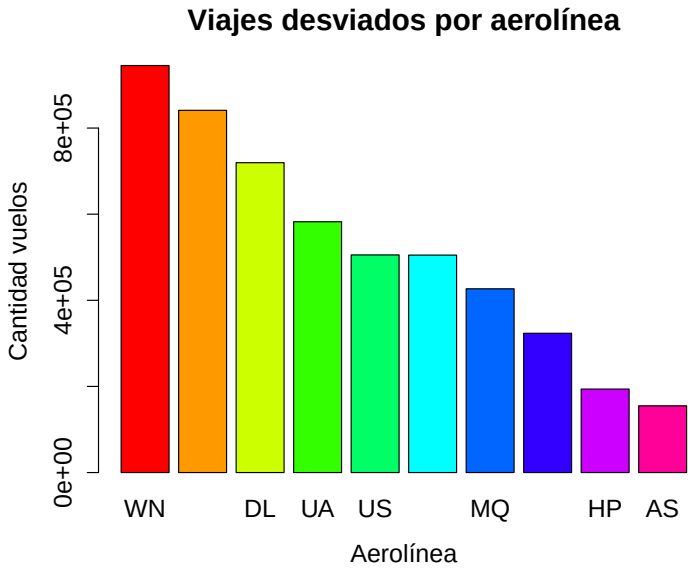


(a) Vuelos desviados por mes

Mes	Cantidad
Enero	498
Febrero	413
marzo	753
Abril	715
Mayo	568
Junio	1016
Julio	1110
Agosto	862
Septiembre	514
Octubre	493
Noviembre	474
Diciembre	940

(b) Datos asociados

Figure 3.8: Resumen de vuelos desviados por mes



(a) Vuelos desviados por aerolínea

Aerolínea	Cantidad
WN	945219
AA	841164
DL	719553
UA	582437
US	505369
NW	504976
MQ	426720
CO	323425
HP	193886
AS	154990

(b) Datos asociados

Figure 3.9: Resumen de vuelos desviados por aerolínea

Vuelos desviados por aerolínea Indica aquellas aerolíneas que más vuelos desviaron en el año. Se muestran en la Figura 3.9.

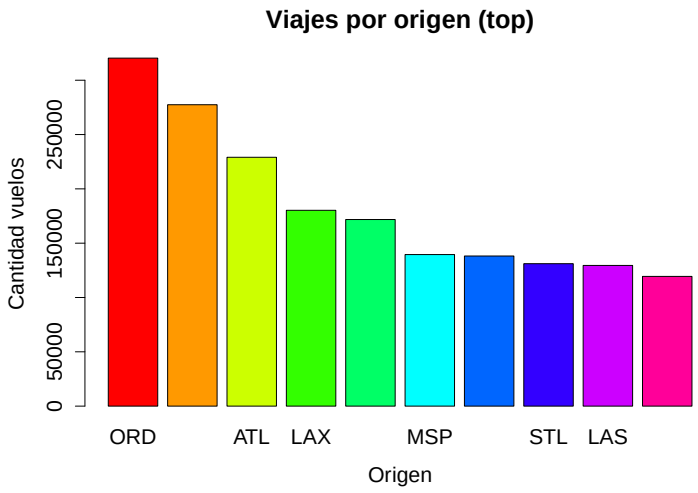
3.1.3 Análisis de vuelos con información de horarios completa

Vuelos por origen Representa a aquellos aeropuertos que, como punto de partida, realizaron la mayor cantidad de vuelos. El aeropuerto de Chicago (ORD) es del que más vuelos parten. Se muestran en la Figura 3.10.

Vuelos por destino Representa a aquellos aeropuertos que que tenían la mayor cantidad de vuelos como punto de destino. Nuevamente, al aeropuerto de Chicago (ORD) es al que más vuelos arriban. Se muestran en la Figura 3.11.

Vuelos por mes Indica la cantidad de vuelos realizados por mes. El mes de agosto es en el que más vuelos fueron realizados, seguido de cerca por julio y mayo. Se muestran en la Figura 3.12.

Vuelos por día del mes Indica la cantidad de vuelos realizados según cada día del mes. No existe alguna tendencia detectable en los datos. Se muestran en la Figura 3.13.

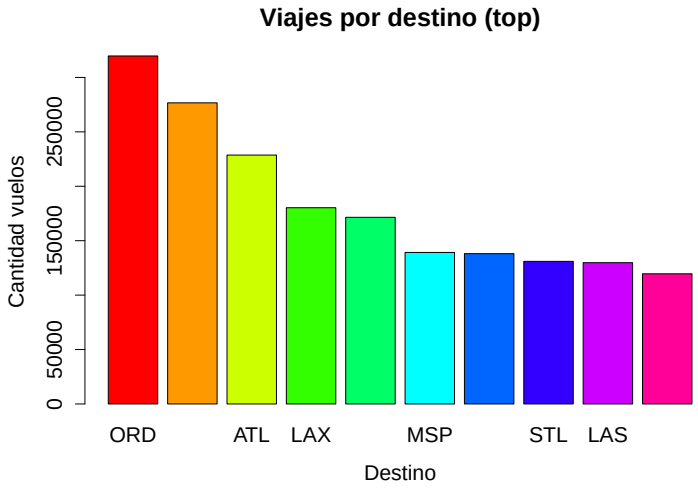


(a) Vuelos concretados por origen

Aeropuerto	Cantidad
ORD	320349
DFW	277445
ATL	229073
LAX	180260
PHX	171777
MSP	139512
DTW	138194
STL	131110
LAS	129509
DEN	119408

(b) Datos asociados

Figure 3.10: Resumen de vuelos concretados por origen



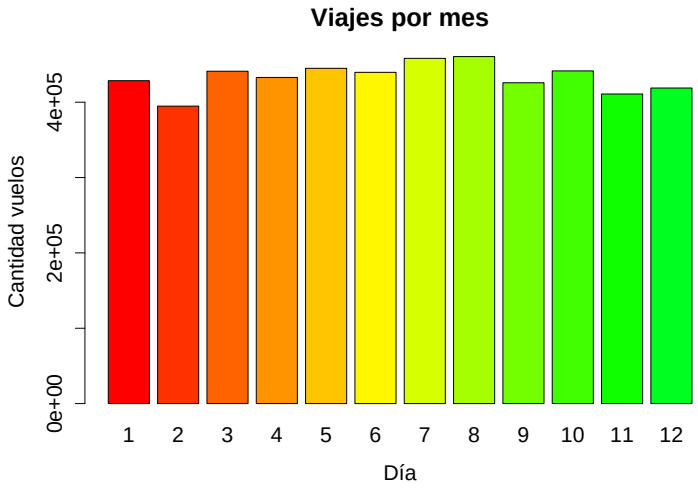
(a) Vuelos concretados por destino

Aeropuerto	Cantidad
ORD	319709
DFW	276636
ATL	228665
LAX	180290
PHX	171481
MSP	139212
DTW	138076
STL	131008
LAS	129773
DEN	119580

(b) Datos asociados

Figure 3.11: Resumen de vuelos concretados por destino

Vuelos por día de la semana Indica la cantidad de vuelos realizados según cada día de la semana. La intención de este análisis era detectar algún patrón de viaje por parte de los usuarios. Se encontró que el día sábado es el menos utilizado para viajar. Se muestran en la Figura 3.14.



(a) Vuelos concretados por mes

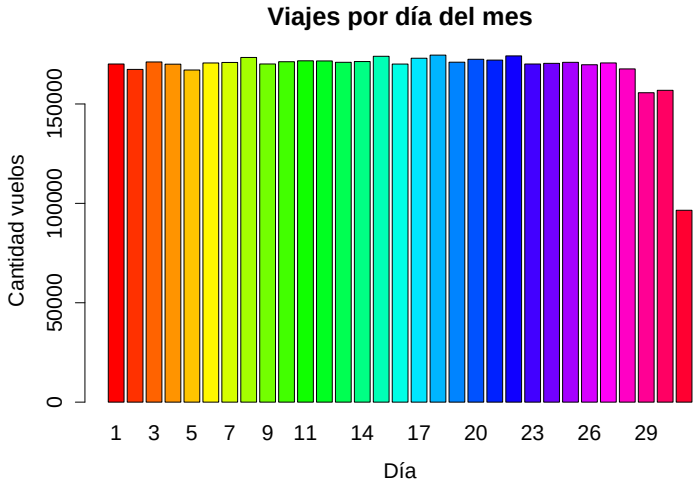
Mes	Vuelos
Enero	428537
Febrero	394799
Marzo	441110
Abril	432854
Mayo	444974
Junio	439651
Julio	458203
Agosto	460563
Septiembre	425796
Octubre	441548
Noviembre	410875
Diciembre	418829

(b) Datos asociados

Figure 3.12: Resumen de vuelos completos por mes

Vuelos por hora Indica la cantidad de vuelos realizados por hora en el día. 17:00 es la hora más común en la que se realizaron los vuelos. Se muestran en la Figura 3.15.

Vuelos por aerolínea Indica aquellas aerolíneas que más vuelos concretaron en el año. Southwest Airlines (WN) es la aerolínea más utilizada. Se muestran en la Figura 3.16.

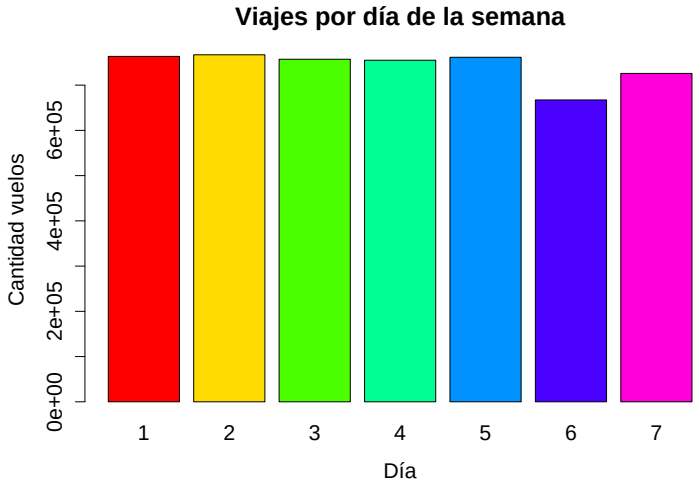


Dia	Cantidad
18	176518
22	176085
15	175853
8	175248
17	174909
3	174506
19	174505
20	174185
11	174146
21	173882

(a) Vuelos concretados por día del mes

(b) Datos asociados

Figure 3.13: Resumen de vuelos completos por día del mes



Día	Cantidad
Lunes	763551
Martes	767106
Miércoles	757236
Jueves	755090
Viernes	761604
Sábado	667250
Domingo	725902

(a) Vuelos concretados por día de la semana

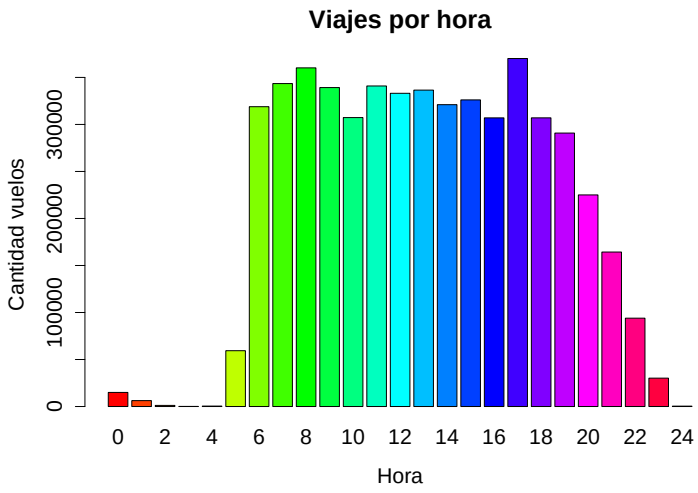
(b) Datos asociados

Figure 3.14: Resumen de vuelos completos por día de la semana

3.2 Con herramientas de clústering

3.2.1 Ejecución del proceso de clústering

Las técnicas de clústering fueron implementadas sobre las versiones reducidas de los datasets. El equipo de cómputo donde se ejecutaron los procesos es una iMac de 21.5



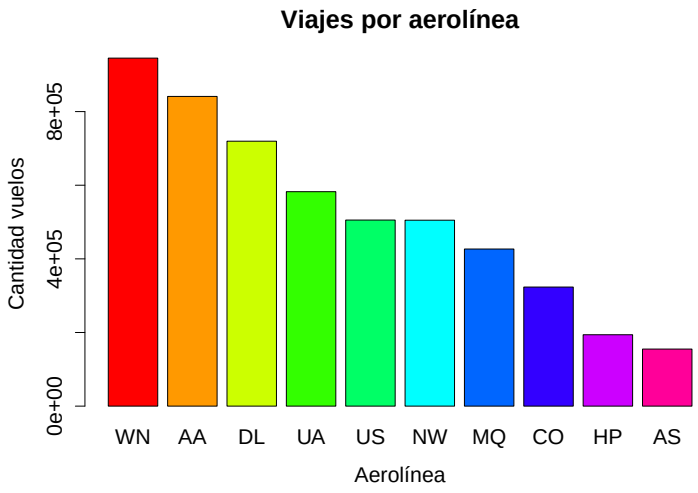
(a) Vuelos concretados por hora

Hora	Cantidad
0	14987
1	6287
2	1189
3	124
4	471
5	59368
6	318859
7	343458
8	360145
9	339205
10	307290
11	340906
12	333104
13	336507
14	321020
15	326087
16	306969
17	370101
18	306993
19	290809
20	225021
21	164291
22	93961
23	30181
24	406

(b) Datos asociados

Figure 3.15: Resumen de vuelos completos por hora

pulgadas, con 4GB de memoria RAM y un procesador Intel i5 con 4 núcleos con una velocidad de 2.5 GHz.



(a) Vuelos concretados por aerolínea

Codigo	Cantidad
WN	945219
AA	841164
DL	719553
UA	582437
US	505369
NW	504976
MQ	426720
CO	323425
HP	193886
AS	154990

(b) Datos asociados

Figure 3.16: Resumen de vuelos concretados por aerolínea

3.2.2 Clústering particional

El algoritmo indicado para la realización del clustering particional fue el ***K-Means***, en particular la función implementada dentro del paquete *skmeans* de ***R***, llamada también *skmeans*. La ejecución de la función *skmeans* fue realizada con tamaños de clúster desde 2 hasta 8 por cada uno de los datasets de los meses. Algunos de los resultados técnicos de este experimento se encuentran descritos en la Tabla 3.1.

k	Tiempo(seg)	Valor (criterion)
2	9.979287	8778.09
3	16.98486	5639.61
4	16.06621	4101.42
5	19.70341	3151.56
6	28.34242	2670.34
7	36.63261	2356.24
8	44.77737	2107.43

Table 3.1: Resultados técnicos de la ejecución de la técnica *skmeans* (promedios)

Clústering jerárquico La técnica indicada para la ejecución del clustering jerárquico fue el *hybridHclust* (Hybrid hierarchical clustering using mutual clusters), el cual sigue el enfoque jerárquico top-down. Este algoritmo no requiere parámetros adicionales al dataset a procesar. La ejecución de este algoritmo fue realizada en base a los datos escalados de cada uno de los meses. Algunos de los resultados técnicos de este experimento se encuentran descritos en la Tabla F2.

4 Interpretación de los resultados

Los procesos de análisis en base a frecuencias y agrupamiento permitieron obtener algunos detalles de la información contenida en los datos.

En cuanto a las técnicas de agrupamiento utilizadas, es importante mencionar los siguientes puntos. Se ejecutaron varias rondas debido a modificaciones en los parámetros, sobre todo las columnas seleccionadas en cada uno de los datasets.

En este sentido, inicialmente se habían considerado datos que tenían relación con las componentes temporales (*timestamp*) de los registros, como por ejemplo las fechas reales y estimadas de llegada y partida; sin embargo, éstos dificultaban la generación y la representación de la información, ya que al reflejar eventos que suceden a lo largo del tiempo, su intersección con otros elementos era mínima.

Debido a dichas modificaciones, hubo constante retroalimentación entre las etapas de preprocesamiento y análisis de los datos hasta alcanzar diseños que pudieran ser ejecutados respetando restricciones de memoria, cómputo y tiempo.

5 Conclusiones

La minería de datos se auxilia de gran variedad de herramientas que tienen su origen en otras disciplinas, como la teoría de la información, las matemáticas o la estadística. Sin embargo, es un proceso que para ejecutarse de forma satisfactoria requiere de conocimiento por parte del analista, en base a su experiencia, para la selección de los atributos y técnicas de análisis apropiadas según el conjunto de datos a procesar. Este documento ha presentado la labor de análisis de datos sobre un dataset utilizando técnicas basadas netamente en frecuencias, así como utilizando un par de técnicas de agrupamiento jerárquico y particional. Se realizó un preprocesamiento para mejorar la calidad de los datos así como la ejecución de *PCA* para elegir las columnas que representaron la mayor varianza de los datos. Adicionalmente, se realizó un submuestreo aleatorio para reducir la cantidad

de instancias a introducir en los algoritmos y hacer factible su ejecución en términos de memoria. Después de rechazar los atributos relacionados con instantes de tiempo, se sospecha que los clústers formados tienen relación con XXXXXXXXX-YYYYYYYY. Adicionalmente se han mostrado estadísticas técnicas de los tiempos de ejecución de dichos algoritmos.