—— MEMORIA ——

Minería de Datos sobre un Data Mart
PRÁCTICA 5

CURSO 2023/2024

Simón Alonso (821038)
Óscar Brizuela (820773)
Jan Carlos Franco (806955)

Índice

1. Predictores influyentes en el retraso de los vuelos

En primer lugar, es necesario tener en una sola tabla la información correspondiente a todas las tablas de nuestra base de datos, para posteriormente conformar el *dataset* apropiado en *R*. Para ello, se combinaron las distintas columnas situadas en hojas diferentes de *Excel* (del mismo libro) utilizando la función *VLOOKUP,* de *Excel*.

Dado que contábamos con más de medio millón de vuelos en nuestra hoja *Excel*, se han elegido 50000 filas aleatorias, sin valores nulos, del conjunto completo. De no ser así, los tiempos de computación con nuestras máquinas serían demasiado altos a la hora de computar todos los modelos posibles con los predictores seleccionados del *dataset*.

Se ha optado por elegir los siguientes 8 predictores:

1. *TIEMPO_TOTAL_VUELO*
2. *CIUDAD_ORIGEN*
3. *CIUDAD_DESTINO*
4. *MARCA*
5. *DIA_SEMANA*
6. *IATA_AEROLINEA*
7. *FRANJA_HORARIA_LLEGADA*
8. *FRANJA_HORARIA_SALIDA*

Para poder evaluar los distintos modelos y encontrar cuáles son los factores que más afectan al retraso de los vuelos lo primero que se ha realizado ha sido convertir los predictores cualitativos en cuantitativos. Esto se ha llevado a cabo mediante la función *as.factor()*, que asigna un número a cada variable cualitativa, de modo que las que tengan el mismo valor, tendrán el mismo número.
Lo siguiente que se ha realizado es iterar por cada número de predictores, evaluando el modelo con todas las combinaciones posibles con dicho número de predictores. Una vez se ha creado el modelo, se obtiene la métrica R^2 y se comparan los modelos que tengan el mismo número de predictores para quedarnos con el mejor modelo para ese número de predictores. El código correspondiente a dichas iteraciones se puede encontrar en el Anexo.

Una vez se cuenta con los mejores modelos para cada número de predictores es necesario el R^2 ajustado, el *BIC* o el *AIC*. Se han utilizado los tres para estar más seguros de los resultados al tener más métricas que analizar.

Al comparar los modelos se ha buscado el que mayor R^2 ajustado tenga (cuanto más cercano a 1 mejor). Además, se han considerado las métricas *BIC* y *AIC*, para las cuales se han buscado los valores más bajos en los modelos.

Resultados:

Número de predictores	Predictores del mejor modelo (mayor R^2)	R^2 ajustado	*BIC*	*AIC*
1	*CIUDAD_ORIGEN*	0.0094	617567	614921
2	*CIUDAD_ORIGEN* *FRANJA_HORARIA_LLEGADA*	0.024	616818	614154
3	*CIUDAD_ORIGEN* *DIA_SEMANA* *FRANJA_HORARIA_LLEGADA*	0.035	616334	613617
4	*CIUDAD_ORIGEN* *CIUDAD_DESTINO* *DIA_SEMANA* *FRANJA_HORARIA_LLEGADA*	0.037	619162	613826
5	*CIUDAD_ORIGEN* *CIUDAD_DESTINO* *DIA_SEMANA* *MARCA* *FRANJA_HORARIA_LLEGADA*	0.039	619163	613712
6	*TIEMPO_TOTAL_VUELO* *CIUDAD_ORIGEN* *CIUDAD_DESTINO* *DIA_SEMANA* *IATA_AEROLINEA* *FRANJA_HORARIA_LLEGADA*	0.04	619102	613643
7	*TIEMPO_TOTAL_VUELO* *CIUDAD_ORIGEN* *CIUDAD_DESTINO* *DIA_SEMANA* *IATA_AEROLINEA* *FRANJA_HORARIA_LLEGADA* *MARCA*	0.041	619176	613602
8	*TIEMPO_TOTAL_VUELO* *CIUDAD_ORIGEN* *CIUDAD_DESTINO* *DIA_SEMANA* *IATA_AEROLINEA* *FRANJA_HORARIA_LLEGADA* *MARCA* *FRANJA_HORARIA_SALIDA*	0.042	619171	613579

Tabla 1: mejores modelos en función del número de predictores

Se puede observar que teniendo en cuenta el R^2 ajustado y el *AIC,* el mejor modelo es el de 8 predictores (todos). Sin embargo, la métrica *BIC* nos indica que el mejor es el de 3 predictores (los presentes en la tabla). Esto se puede deber a que el *BIC* penaliza de forma mayor a los modelos más complejos. No obstante, el hecho de que el mejor modelo coincida en tener las mejores métricas tanto para el R^2 ajustado como para el *AIC* nos da una confianza extra a la hora de determinar cuál es el mejor modelo.

De todas formas, estos resultados no son suficientemente relevantes, ya que indican que utilizando una regresión lineal, independientemente del número de predictores que se seleccione (de los que hemos preseleccionado), el modelo no se consigue ajustar bien a los datos, pues los valores del R^2 ajustado son demasiado pequeños, y los valores de *AIC* y *BIC,* quizá demasiado grandes. Es posible que estos valores se deban a la gran cantidad de datos heterogéneos con la que se trabaja.

Tras el descontento de los resultados, y para confirmar que la metodología llevada a cabo era la correcta, se ha utilizado el predictor *RETRASO_LLEGADA*, el cual sumado a *RETRASO_SALIDA* proporciona el *RETRASO_TOTAL* del vuelo. Como se esperaba, solamente con uno de estos predictores se obtiene un R^2 ajustado de 0.98, puesto que evidentemente el retraso de llegada (o de salida) influye de forma directa en el retraso total.

2. Clusterización de los retrasos en 3 grupos en función de una variable

Para la realización de estos agrupamientos se ha utilizado el algoritmo de *k-medias*, que inicia *k* centroides aleatoriamente y asigna a cada punto el clúster que tiene el centroide más cercano a dicho punto. Una vez se han asignado todos los puntos a un clúster, el centroide se coloca en el lugar de las medias de dichos puntos y se vuelve a repetir el proceso hasta que converja o hasta que se ejecute un número máximo de iteraciones.

2.1. Clasificación en función del aeropuerto origen

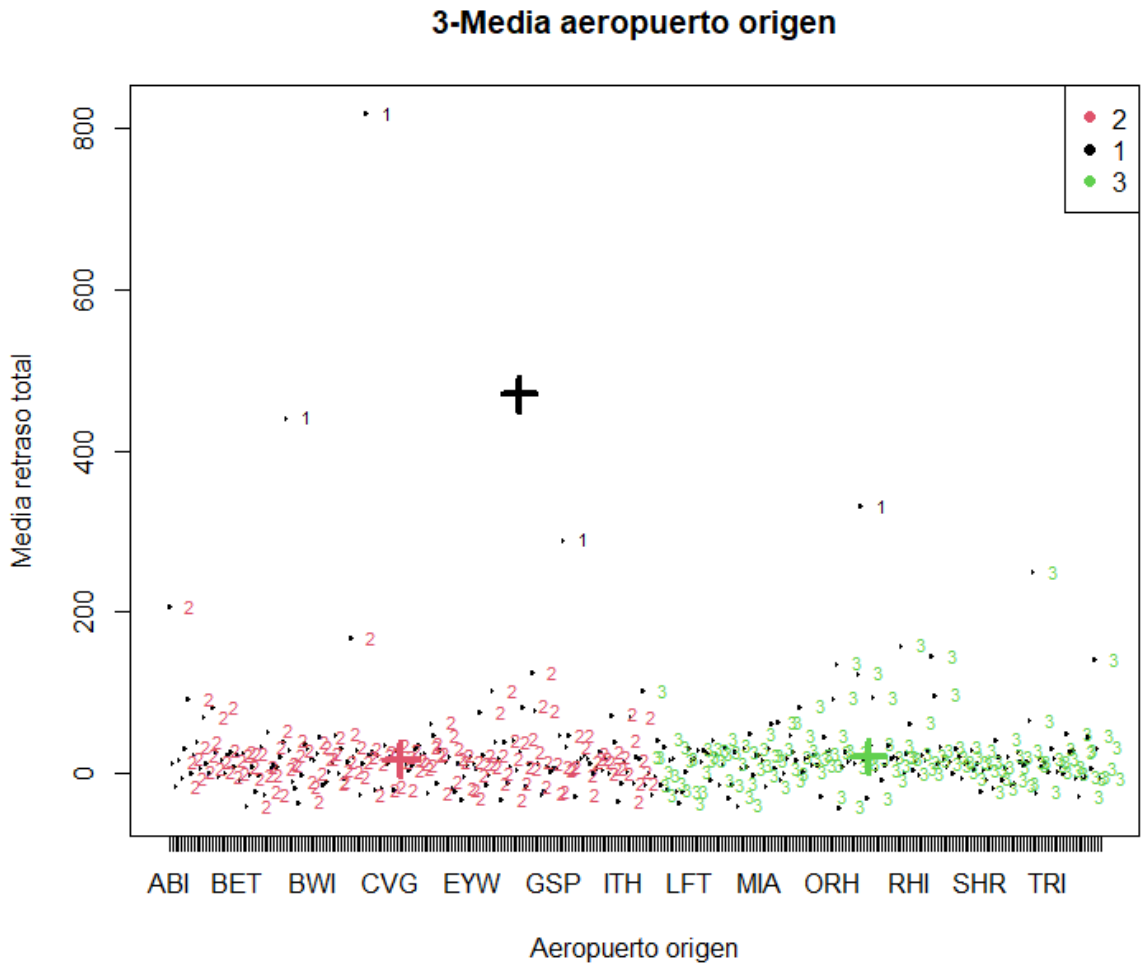

Figura 1: gráfico de los aeropuertos origen agrupados en función del retraso medio total, utilizando k-medias (k = 3)

Se puede observar cómo la media del retraso de los vuelos dado el aeropuerto de origen se ha podido agrupar en 3 grupos. Se observa que el grupo 1 tiene una media de retraso de aproximadamente 450 minutos. De esta forma podemos distinguir cuales son los aeropuertos cuya media de retraso más se acerca a 450 minutos de media.

Lo mismo ocurre para los grupos 2 y 3, los cuales tienen el centroide a una altura muy similar en cuanto al retraso medio total, lo que indica que ha conseguido separar en grupos los aeropuertos para que la media del retraso de dichos aeropuertos sean muy similares, aunque pertenezcan a diferentes grupos.

2.2. Clasificación en función del aeropuerto destino

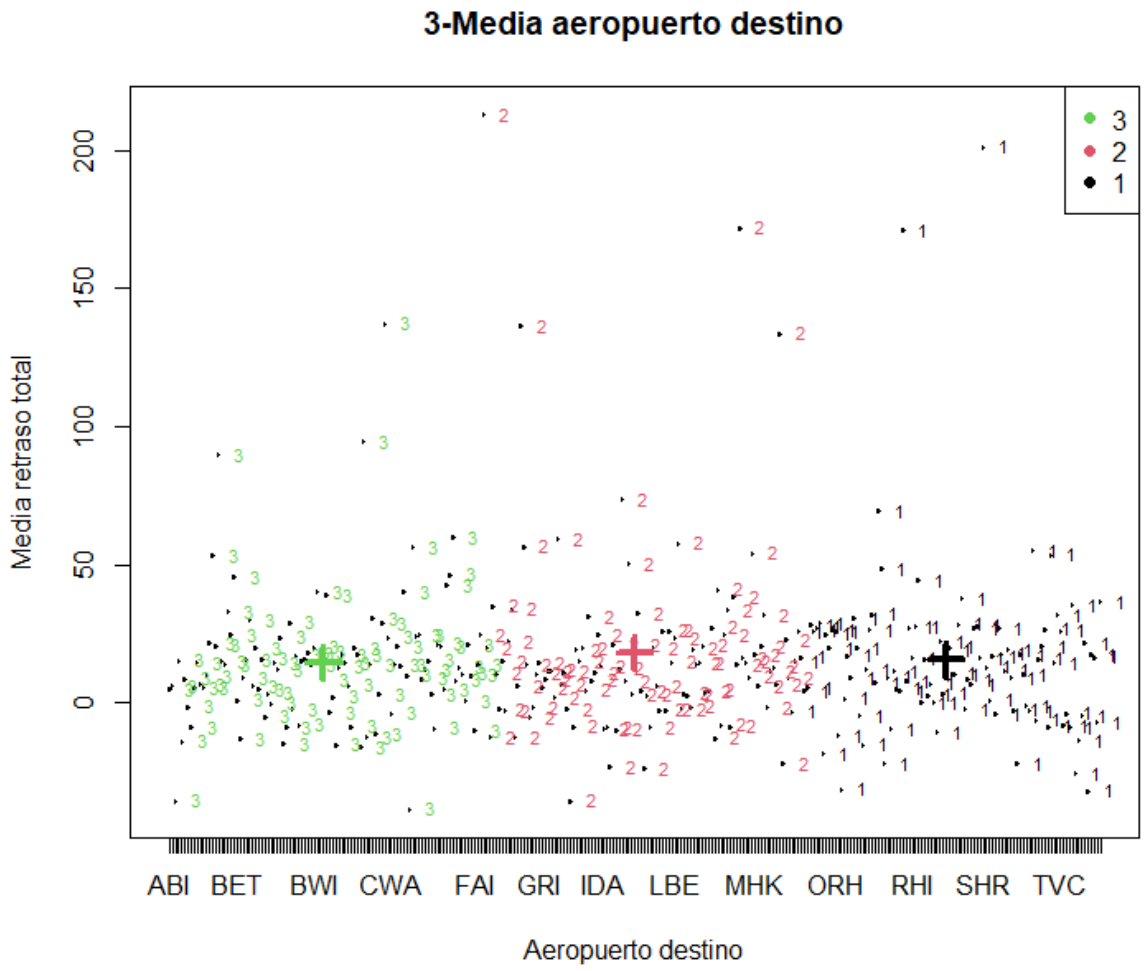

Figura 2: gráfico de los aeropuertos destino agrupados en función del retraso medio total, utilizando k-medias (k = 3)

Se puede observar cómo la convergencia del *k-medias* ha dado con 3 centroides casi en la misma posición en el eje *y*, lo que indica que la media de todos los aeropuertos de destino de cada grupo es muy similar con la de los otros grupos. Esto podría indicar que los grupos son homogéneos en términos del retraso medio total por aeropuerto destino, ya que se observan patrones muy similares entre los distintos clústers.

2.3. Clasificación en función de la aerolínea

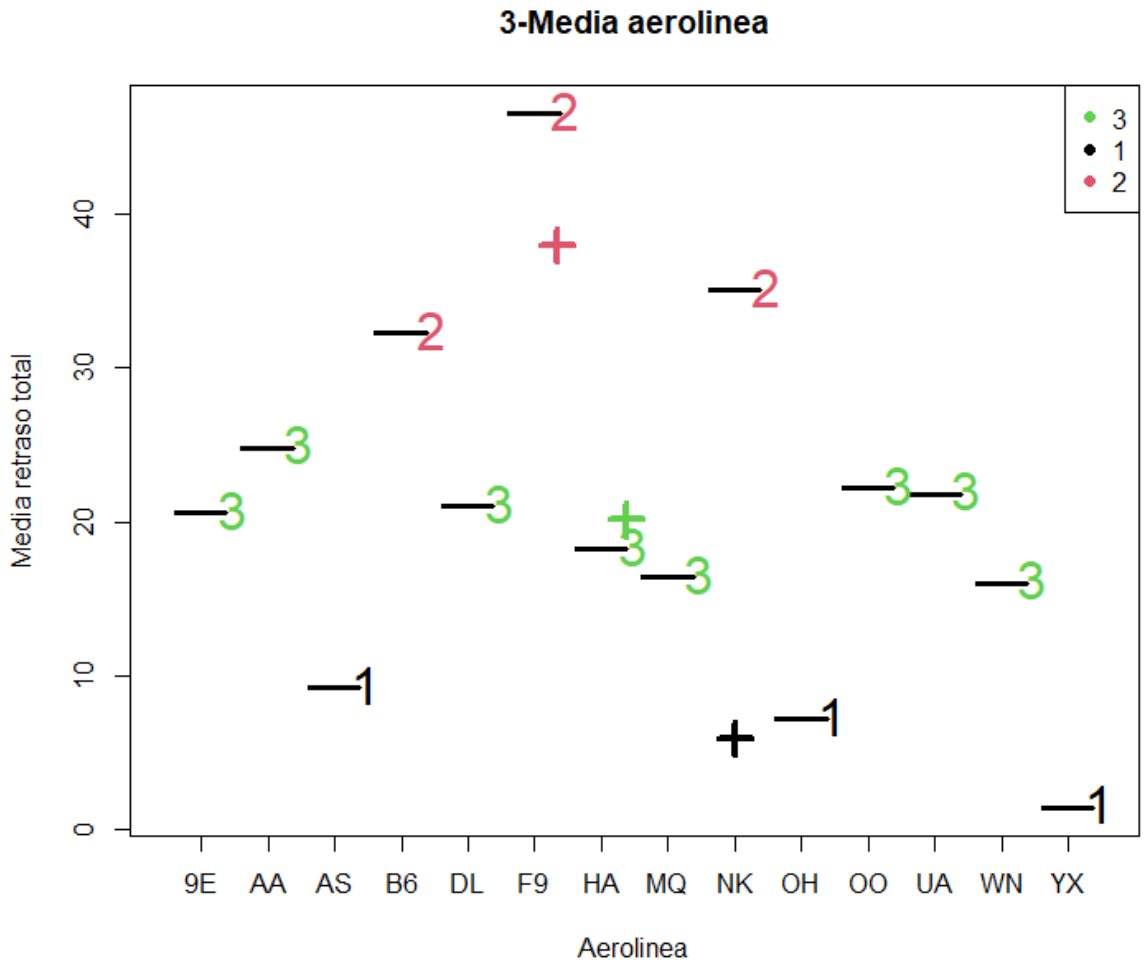

Figura 3: gráfico de las aerolíneas agrupadas en función del retraso medio total, utilizando k-medias (k = 3)

En este caso, los 3 grupos generados tienen una media de retraso total bastante diferente entre sí. El grupo con mayor retraso de media entre todas las aerolíneas es el 2, con cerca de 38 minutos, seguido por el grupo 3, con una media de retraso de 20 minutos. Las aerolíneas del grupo 1 cuenta con un retraso medio de alrededor de 7 minutos, por lo que se puede concluir que las aerolíneas “AS”, “OH” y “YX” son las más eficientes en cuanto a retrasos.

2.4. Clasificación en función del modelo de avión

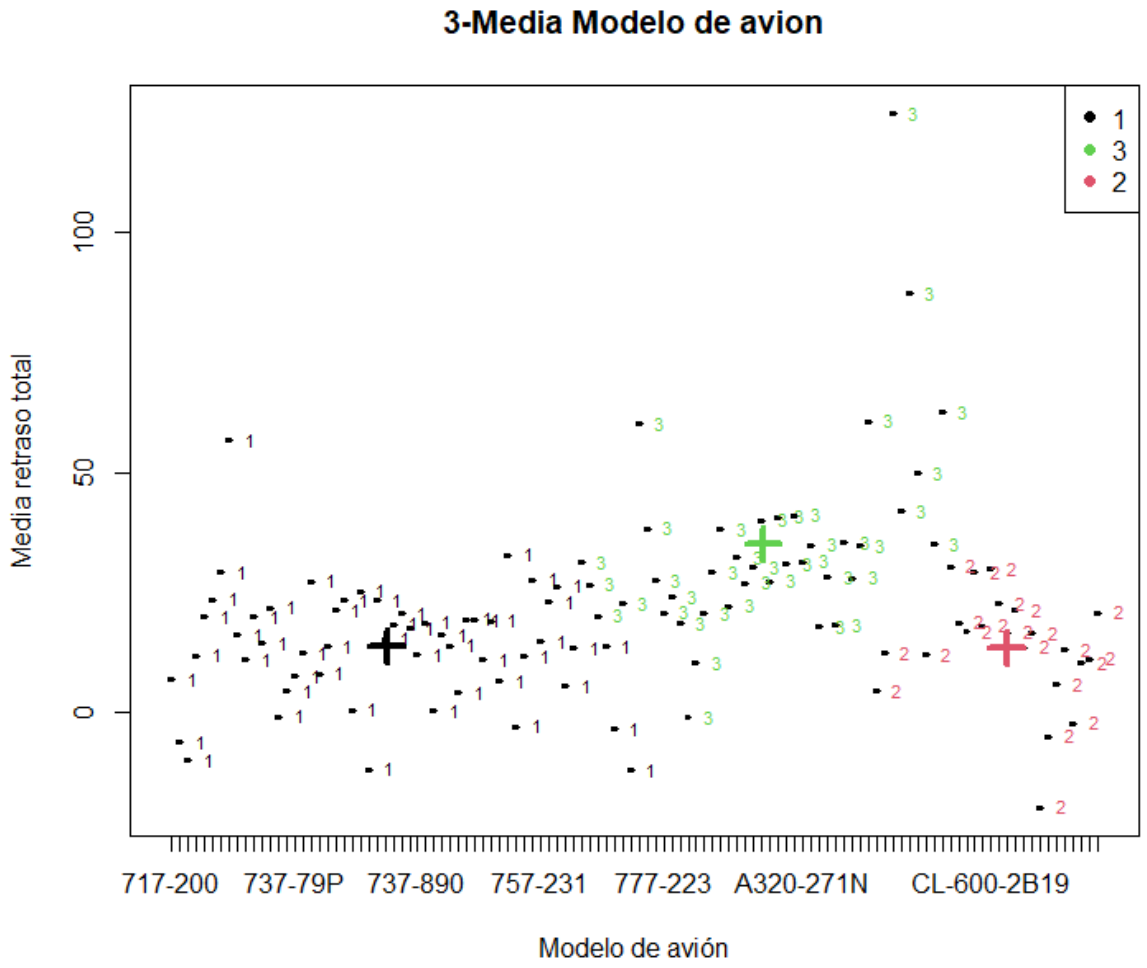

Figura 4: gráfico de los modelos de avión agrupados en función del retraso medio total, utilizando k-medias (k = 3)

Se puede observar cómo se han agrupado los modelos de avión de forma que la media de los clústers es muy similar entre los distintos grupos. Estas agrupaciones han tenido lugar debido a que los modelos de avión dentro de cada cluster tienen retrasos medios similares, diferenciándose en alrededor de 10 minutos.

3. Influencia de la franja horaria en los retrasos

Para medir la influencia de las franjas horarias en los distintos retrasos con los que se cuentan se ha realizado un análisis de la varianza (*ANOVA*), puesto que se desea investigar la influencia de una variable cualitativa sobre una variable cuantitativa. Una vez realizado el análisis de la varianza se lleva a cabo la prueba *Tukey*. Estos análisis y pruebas se han realizado con todas las posibles combinaciones del retraso total, retraso de salida y retraso de llegada con la franja horaria de llegada y la franja horaria de salida. Un ejemplo del código necesario para realizar estos análisis se encuentra en el anexo.

Retraso total con franja horaria de llegada

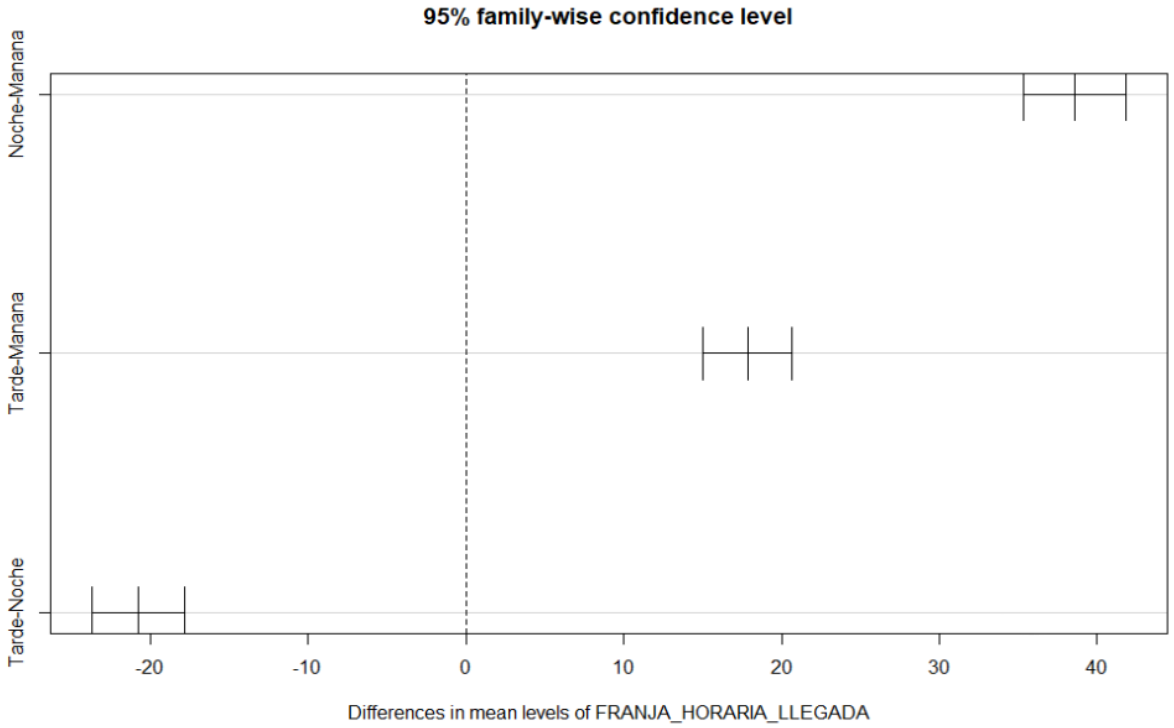

Figura 5: gráfico de caja de test de Tukey representando el retraso total frente a la franja horaria de llegada

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = RETRASO_TOTAL ~ FRANJA_HORARIA_LLEGADA, data = datos)

$FRANJA_HORARIA_LLEGADA
                    diff        lwr        upr p adj
Noche-Manana   38.60538  35.36302  41.84775     0
Tarde-Manana   17.83518  15.01603  20.65433     0
Tarde-Noche   -20.77021 -23.68989 -17.85053     0
```

Figura 6: valores estadísticos del test de Tukey del retraso total frente a la franja horaria de llegada

Retraso total con franja horaria de salida

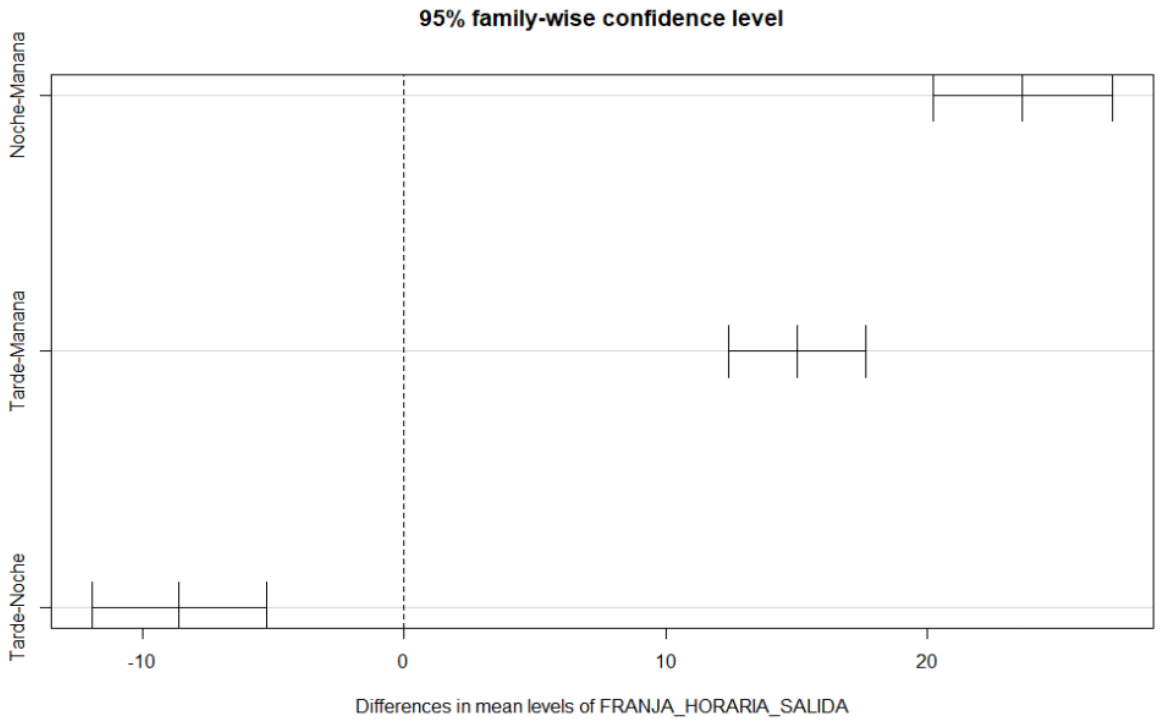

Figura 7: gráfico de caja de test de Tukey representando el retraso total frente a la franja horaria de salida

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = RETRASO_TOTAL ~ FRANJA_HORARIA_SALIDA, data = datos)

$FRANJA_HORARIA_SALIDA
                    diff       lwr      upr p adj
Noche-Manana  23.665601  20.21906 27.11215     0
Tarde-Manana  15.050410  12.41935 17.68147     0
Tarde-Noche   -8.615192 -11.94220 -5.28818     0
```

Figura 8: valores estadísticos del test de Tukey del retraso total frente a la franja horaria de salida

Retraso de llegada con franja horaria de llegada

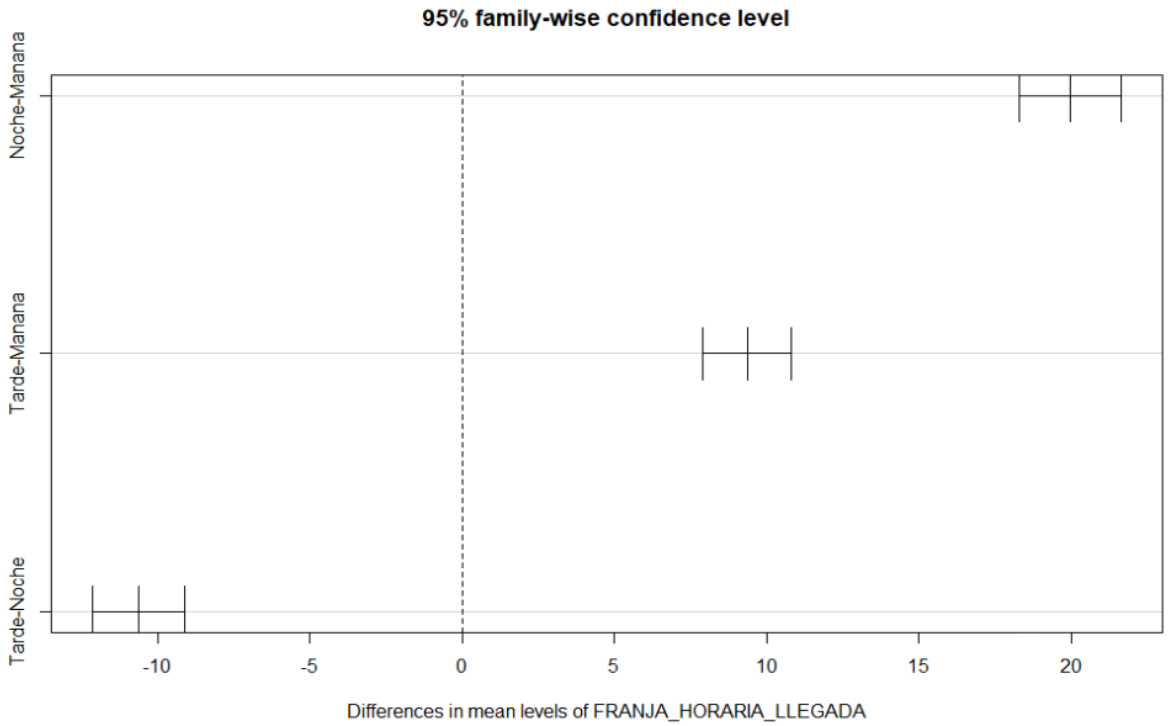

Figura 9: gráfico de caja de test de Tukey representando el retraso de llegada frente a la franja horaria de llegada

Retraso llegada y franja horaria salida

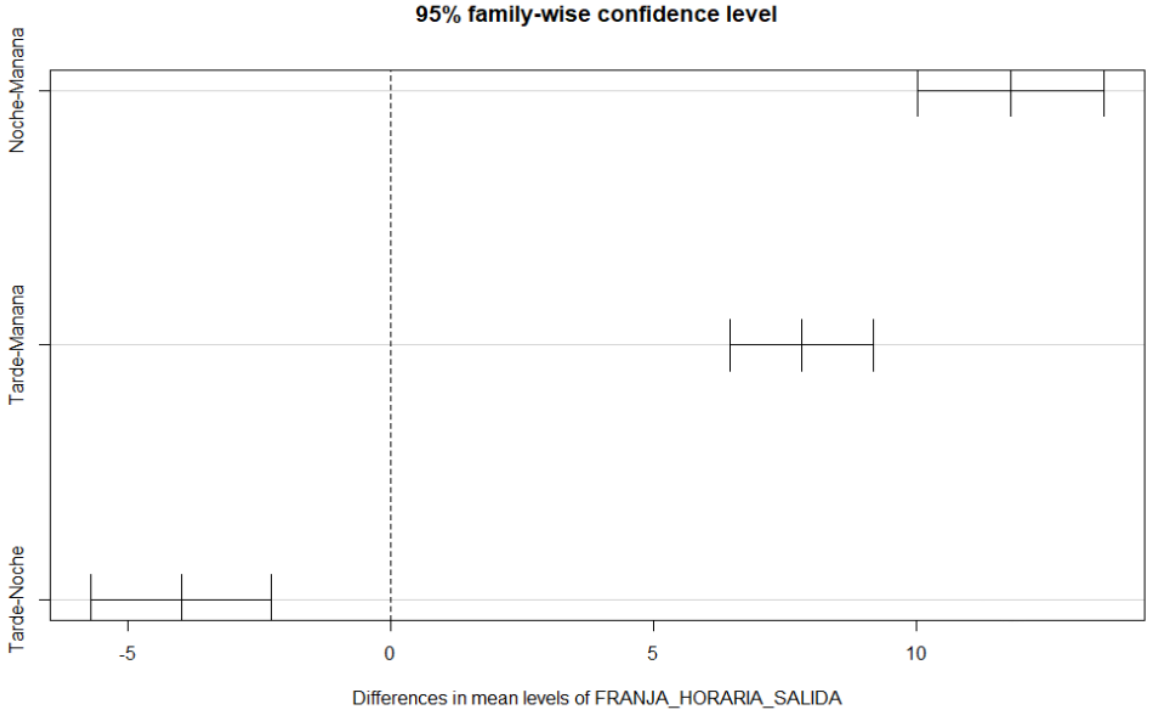

Figura 10: gráfico de caja de test de Tukey representando el retraso de llegada frente a la franja horaria de salida

Retraso salida y franja horaria salida

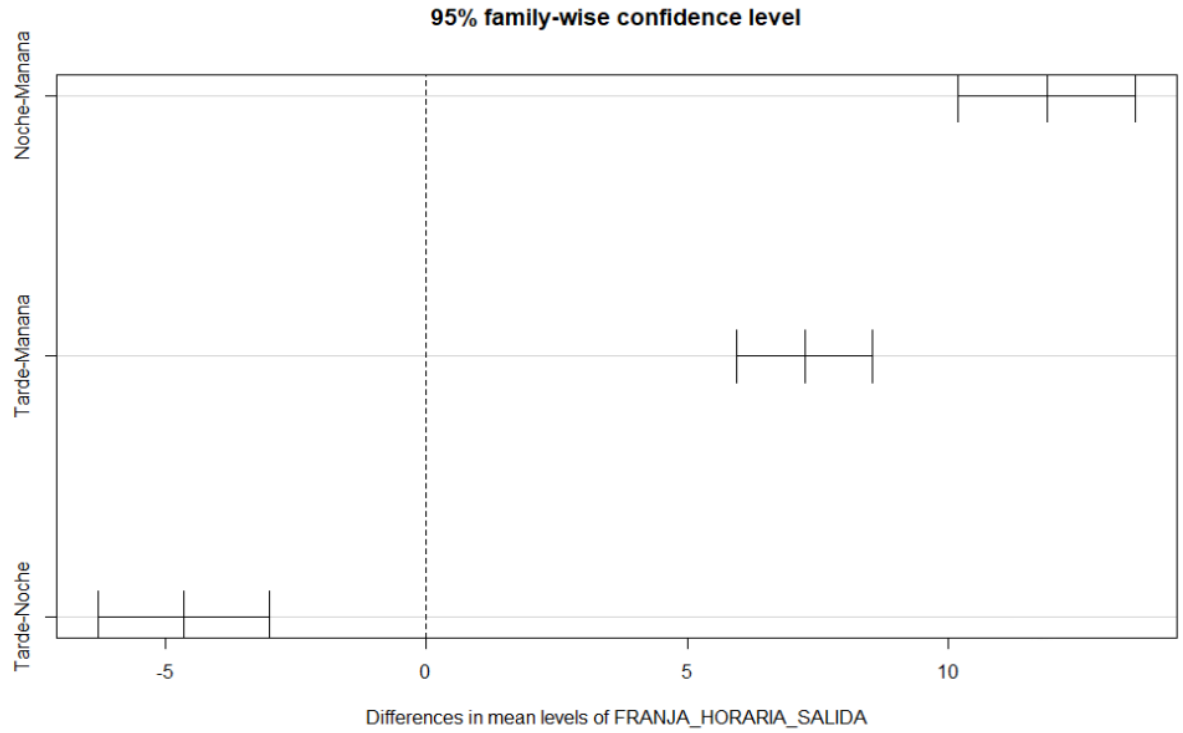

Figura 11: gráfico de caja de test de Tukey representando el retraso de salida frente a la franja horaria de salida

Retraso salida y franja horaria llegada

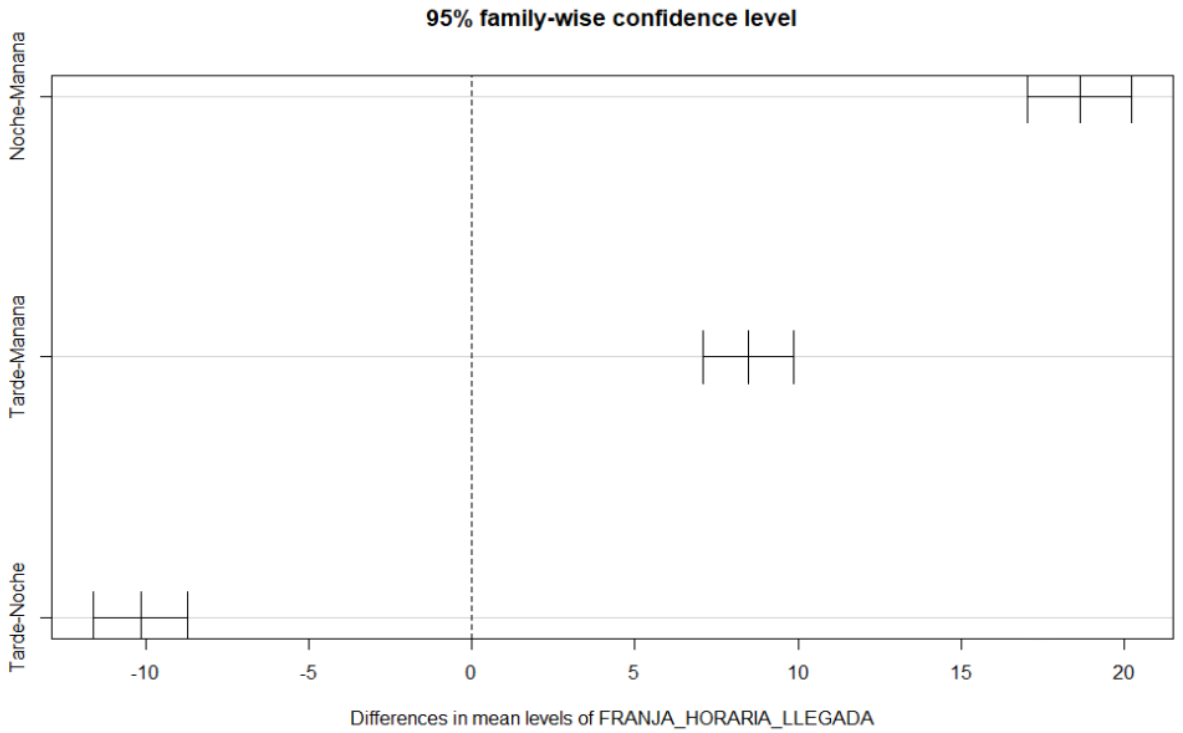

Figura 12: gráfico de caja de test de Tukey representando el retraso de salida frente a la franja horaria de llegada

Como podemos observar en los p-valores obtenidos en los test de Tukey, los cuales son muy cercanos a 0 (redondeados directamente a 0 en las figuras 6 y 7), tenemos evidencia de que existen medias distintas entre los distintos grupos con los intervalos de confianza utilizados. Esta teoría, además, se comprueba gráficamente gracias a las figuras 5, 7, 9, 10, 11 y 12, donde se ve que el 0 no pertenece a ningún intervalo en ninguno de los casos. Por tanto, podemos concluir que las franjas horarias sí que influyen en los distintos retrasos de los vuelos.

4. Estudio de reglas de asociación en los vuelos

Para analizar si existían patrones en los datos mediante reglas de asociación primero fue necesario pasar a "factor" aquellas variables cualitativas de las que se querían extraer posibles patrones. De esta forma, transformamos las variables cualitativas en ítems, teniendo un ítem por cada nivel de cada variable cualitativa. No obstante, previamente se reemplazaron los valores con tilde, "Sábado" y "Miércoles", de la variable *DIA_SEMANA*, por sus valores sin tilde, y los valores con "ñ", "Mañana", de las variables correspondientes a *FRANJA_HORARIA_SALIDA* por "manana". Este reemplazo de "carácteres extraños para *R*" evita que aparezcan posteriores errores en la creación y filtrado de reglas de asociación.

Se decidió tener en cuenta varias de las variables cualitativas del *data mart* presentes en la hoja de cálculo *Excel* descrita en el punto 1. No se tuvieron en cuenta identificadores (de fecha y vuelo), las variables correspondientes a las horas, tanto estimadas como reales, de salida y llegada ni el periodo vacacional.

En un principio estas variables fueron las únicas en no ser tenidas en cuenta en el *dataset*. Sin embargo, tras la generación de multitud de reglas, nos dimos cuenta de que la gran mayoría de ellas presentaban obviedades: no aportaban información debido a que algunas de las variables presentes en el antecedente y en el consecuente se encuentran relacionadas entre sí por sentido común. Por ejemplo, en este momento existía un gran número de reglas en las que en el antecedente aparecía un aeropuerto (como *DEN*, aeropuerto de *Denver*), y en el consecuente aparecía el estado o la ciudad de dicho aeropuerto (*Denver, CO*). Esta regla, por tanto, no aportaba información sobre ningún patrón, ya que evidentemente todos los vuelos que salen desde un aeropuerto salen desde el estado o la ciudad correspondiente a dicho aeropuerto. Esto no solo sucedía con las variables pertenecientes a "lugares" (aeropuertos, ciudades y estados tanto origen como destino), sino también a los retrasos (si un vuelo se había retrasado en la salida era normal que también se retrasase en la llegada), las franjas horarias (reglas con horas pertenecientes a la noche, por ejemplo, tenían en el antecedente la franja horaria correspondiente a "Noche") o las marcas (un modelo solo es de una marca, por lo que las reglas con una marca en el antecedente y uno de sus modelos en el consecuente, o viceversa, no aportaban información).

Para solventar este problema y disminuir considerablemente el número de reglas generadas, se decidió tener en cuenta en el *dataframe* solo el retraso total (retraso de salida + retraso de llegada) y la franja horaria de llegada (puesto que existían también muchas reglas con la misma franja horaria en la salida, que aparecía en el consecuente, y en la llegada, en el antecedente, o viceversa).

Por otro lado, debido a que sabiendo el aeropuerto se puede conocer la ciudad a la que pertenece, su población y el estado, se decidió tener en cuenta para el *dataframe* solamente el aeropuerto (tanto de origen como de destino) de los vuelos. Esta decisión redujo el número de reglas notablemente. Siguiendo la misma idea, solo se tuvo en cuenta el modelo de avión y no la marca, puesto que dado un modelo de avión concreto podemos conocer la marca a la que corresponde. Para implementar estas ideas simplemente se asignaron las variables que no se deben considerar en el *dataframe* a *NULL*.

Además, también fue necesario discretizar las variables continuas, puesto que las reglas de asociación no se pueden aplicar a este tipo de variables.
Las variables “continuas” discretizadas fueron las siguientes:

- *TIEMPO_TOTAL_VUELO*: debido a que el tiempo total de vuelo está descrito en minutos, se han considerado vuelos “cortos” aquellos con una duración inferior a una hora y media (90 minutos); vuelos “medios” aquellos con una duración de entre 91 minutos y 3 horas; y vuelos “largos” aquellos que duran más de 3 horas.
- *RETRASO_TOTAL*: aquellos vuelos con un retraso mayor a 5 minutos se han considerado como “retrasados”. Aquellos vuelos con un “adelanto” mayor a 5 minutos se han considerado como “adelantados”. Por tanto, se ha decidido dar un margen de +- 5 minutos para considerar al resto de vuelos como “puntuales”, ya que prácticamente ningún vuelo sale o llega exactamente a la hora estimada.

Una vez discretizadas todas las variables, se creó la matriz de transacciones. De esta forma, cada vuelo se conforma como una transacción, que contendrá una serie de ítems.

4.1. Generación de reglas

Tras tener la matriz de transacciones definitiva, habiendo eliminado las obviedades antes comentadas, se generaron diversas reglas de asociación mediante los algoritmos *apriori (76 reglas)* y Eclat (76 reglas). Para ambos algoritmos se utilizaron los mismos parámetros:

- Soporte mínimo (*minsup*, correspondiente al parámetro *support* en R): 0.01
- Confianza mínima (*minconf*, correspondiente al parámetro *confidence* en R): 0.8
- Número máximo de ítems que contiene la regla (parámetro *maxLen* en R): 10

El código correspondiente a la generación de algunas de estas reglas se encuentra en el Anexo.

4.2. Filtrado de reglas

A pesar de que el número de reglas no era tan elevado como en un principio (donde superaba el millón), fue necesario filtrar reglas relevantes que pudieran aportar información.

Las reglas que se describen a continuación son simplemente subconjuntos de las reglas anteriormente generadas.

4.2.1. Reglas con *DIA_SEMANA*

Para investigar si existe algún patrón en el dataset que tenga relación con los días de la semana de los vuelos se filtraron, en primer lugar, aquellas reglas en las que apareciera algún día de la semana en el consecuente (right hand side). Debido a que ninguna regla cumplía con estas características, se realizó el mismo filtrado, pero ahora para ver si existían reglas en las que algún dia de la semana aparecía en el antecedente. Estas fueron ordenadas según los criterios "lift" y "confidence". Sin embargo, a pesar de todos los esfuerzos realizados para eliminar las obviedades anteriormente descritas, todavía obtenemos reglas poco relevantes, como las presentes en la figura 13. Estas muestran patrones en los que aparece la aerolínea "WN", siempre con el mismo modelo de avión en el antecedente ("737-7H4") además de los días de la semana, por lo que las reglas pueden simplemente deberse a que esta aerolínea solo dispone de ese modelo de avión, el cual hace volar todos los días.

Por otro lado, se filtraron aquellas reglas en las que en el consecuente aparecieran las demás variables (y, como sucederá de ahora en adelante, que tuvieran un *lift* mayor que 2). Por ejemplo, se filtraron aquellas en las que en el consecuente aparecieran los aeropuertos origen.
En la gran mayoría de casos, simplemente, no salía ninguna regla en las que determinadas variables aparecieran en el consecuente. Solamente salían reglas cuando la variable *IATA_AEROLINEA* se encontraba en el consecuente. Estas reglas coinciden con las reglas obtenidas al iniciar este apartado, utilizando tanto el algoritmo *apriori* como el algoritmo Eclat. Un ejemplo del código para filtrar este tipo de reglas se encuentra en el Anexo.

```
[1]  {DIA_SEMANA=Jueves,
      MODELO=737-7H4}              => {IATA_AEROLINEA=WN} 0.01136
1  0.01136 4.779201   568
[2]  {DIA_SEMANA=Viernes,
      MODELO=737-7H4}              => {IATA_AEROLINEA=WN} 0.01130
1  0.01130 4.779201   565
[3]  {DIA_SEMANA=Martes,
      MODELO=737-7H4}              => {IATA_AEROLINEA=WN} 0.01318
1  0.01318 4.779201   659
[4]  {DIA_SEMANA=Domingo,
      MODELO=737-7H4}              => {IATA_AEROLINEA=WN} 0.01442
1  0.01442 4.779201   721
```

Figura 13: algunas reglas generadas con DIA_SEMANA en el antecedente, ordenadas por "lift"

4.2.2. Reglas con *IATA_AEROLINEA*

Utilizando ambos algoritmos de generación de reglas obtenemos los modelos más utilizados por las aerolíneas cuando forzamos a que la la aerolínea aparezca como consecuente. Esta información se puede visualizar en la figura 14.

```
[1]  {MODELO=A320-251N}          => {IATA_AEROLINEA=F9} 0.01880
0.01880 39.215686   940
[2]  {MODELO=A320-251N,
      TIEMPO_TOTAL_VUELO=medio} => {IATA_AEROLINEA=F9} 0.01070
0.01070 39.215686   535
[3]  {MODELO=A320-271N}          => {IATA_AEROLINEA=NK} 0.01276
0.01276 24.850895   638
[4]  {MODELO=ERJ 190-100 IGW}    => {IATA_AEROLINEA=B6} 0.01260
0.01260 23.169601   630
[5]  {MODELO=ERJ 170-200 LL}     => {IATA_AEROLINEA=OO} 0.01158
0.01158 10.570825   579
[6]  {MODELO=A319-131}           => {IATA_AEROLINEA=UA} 0.01078
0.01078  9.150805   539
[7]  {MODELO=737-924ER}          => {IATA_AEROLINEA=UA} 0.01896
0.01896  9.150805   948
[8]  {MODELO=737-824}            => {IATA_AEROLINEA=UA} 0.02048
0.02048  9.150805  1024
```

Figura 14: algunas reglas generadas con IATA_AEROLINEA en el consecuente, ordenadas por "lift"

Cuando, por otro lado, hacemos que la aerolínea aparezca en el antecedente, obtenemos también 5 aparentes patrones que revelan información sobre los modelos utilizados por ciertas aerolíneas para tipos de vuelo (en duración o en franja horaria, por ejemplo) concretos. De la misma manera, también podemos ver que la aerolínea "YX", cuando opera vuelos de menos de hora y media, suele tener un retraso total negativo, haciendo que se adelante el vuelo respecto a la hora estimada. En la siguiente figura apreciamos estos 5 patrones.

```
    lhs                                rhs                              support
fidence coverage      lift count
[1] {IATA_AEROLINEA=9E,
     TIEMPO_TOTAL_VUELO=medio}     => {MODELO=CL-600-2D24}        0.01526
9159664  0.01666 16.653934   763
[2] {IATA_AEROLINEA=YX,
     TIEMPO_TOTAL_VUELO=corto}     => {RETRASO_TOTAL=adelantado} 0.01168
8295455  0.01408  1.535001   584
[3] {IATA_AEROLINEA=YX,
     TIEMPO_TOTAL_VUELO=corto}     => {MODELO=ERJ 170-200 LR}     0.01146
8139205  0.01408  7.908283   573
[4] {IATA_AEROLINEA=9E,
     FRANJA_HORARIA_SALIDA=Tarde} => {MODELO=CL-600-2D24}        0.01450
8136925  0.01782 14.794409   725
[5] {IATA_AEROLINEA=9E}           => {MODELO=CL-600-2D24}        0.02732
8063754  0.03388 14.661372  1366
```

Figura 15: todas las reglas generadas con IATA_AEROLINEA en el antecedente, ordenadas por "lift"

A pesar de que casi todas las reglas en las que las variables a cruzar con la aerolínea aparecen en el antecedente no presentan información nueva relevante, sí existen dos tipos de reglas que aportan patrones interesantes. Estos se obtienen haciendo que el aeropuerto origen (y también el destino) aparezca en el consecuente de la regla, mostrando que, de nuevo, la aerolínea "WN" opera en dos aeropuertos principalmente (*DAL* y *MDW*). Gracias a esta información, y junto a la obtenida en el punto 4.2.1, podemos deducir que la aerolínea

“WN”, conocida como “Southwest Airlines”, es alguna aerolínea low-cost que solo utiliza un mismo modelo de avión y que opera principalmente los vuelos entre Dallas y Chicago. Así lo confirma una rápida búsqueda en internet, que muestra que la sede de la aerolínea está ubicada en Texas y que solo utiliza aviones Boeing 737 (modelo de negocio similar al de Ryanair en Europa).

```
    lhs                                rhs                 support confidence
coverage     lift count
[1] {IATA_AEROPUERTO_ORIGEN=DAL} => {IATA_AEROLINEA=WN} 0.01090  0.9749553
0.01118 4.659507   545
[2] {IATA_AEROPUERTO_ORIGEN=MDW} => {IATA_AEROLINEA=WN} 0.01064  0.9156627
0.01162 4.376136   532
```

Figura 16: todas las reglas generadas con IATA_AEROLINEA en el consecuente y IATA_AEROPUERTO_ORIGEN en el antecedente, ordenadas por “lift”

4.2.3. Reglas con *MATRICULA*

Cuando forzamos a que la variable *MATRICULA* aparezca en el antecedente no obtenemos ninguna regla de asociación. Lo mismo sucede cuando hacemos que aparezca en el consecuente. Esto da a entender que no existe ningún patrón relacionado con un avión concreto (no podríamos realizar afirmaciones similares a, por ejemplo, “el avión *N543852* solo es utilizado para vuelos cortos”).

4.2.4. Reglas con *MODELO*

En las reglas en las cuales la variable *MODELO* aparece en el antecedente encontramos algunas de las reglas presentes en el apartado 4.2.2, en el cual se muestran los modelos más utilizados por las aerolíneas. Por otro lado, cuando investigamos las reglas en las que la variable *MODELO* aparece en el consecuente, también obtenemos algunas de las reglas presentes en el punto 4.2.2, en este caso mostrando los modelos utilizados por las aerolíneas para tipos de vuelos determinados. Esta información se puede comprobar en las similitudes entre la figura 17 y la figura 15.

```
    lhs                                rhs                          support
e coverage      lift count
[1] {IATA_AEROLINEA=9E,
     TIEMPO_TOTAL_VUELO=medio} => {MODELO=CL-600-2D24}    0.01526
4  0.01666 16.653934   763
[2] {IATA_AEROLINEA=YX,
     TIEMPO_TOTAL_VUELO=corto} => {MODELO=ERJ 170-200 LR} 0.01146
5  0.01408  7.908283   573
```

Figura 17: todas las reglas generadas con MODELO en el consecuente y IATA_AEROPUERTO_ORIGEN en el antecedente, ordenadas por “lift”

4.2.5. Reglas con *IATA_AEROPUERTO_ORIGEN* y *IATA_AEROPUERTO_DESTINO*

En el caso de intentar que el aeropuerto origen aparezca en el consecuente, no obtenemos ninguna regla con ninguno de los algoritmos de generación de reglas. Sin embargo, cuando hacemos que aparezcan en el antecedente, obtenemos la misma información que en la figura 13, confirmando que la aerolínea "WN" opera principalmente entre el aeropuerto de Dallas y el de Midway, en Chicago. Lo mismo sucede cuando generamos las reglas con los aeropuertos destino.

4.2.6. Reglas con *FRANJA_HORARIA_SALIDA*

A pesar de que no sale ninguna regla cuando *FRANJA_HORARIA_SALIDA* aparece en el consecuente, sí aparecen ciertas reglas cuando dicha variable aparece en el antecedente. Sin embargo, estas reglas se deben, de nuevo, a los modelos utilizados por dos aerolíneas en este caso no solo "WN", sino también "AA". Por tanto, quizá podríamos deducir que la aerolínea "AA" (*American Airlines*, una de las más importantes en *EEUU*) utiliza sus modelos *737-823* cuando necesita realizar vuelos de entre 1 hora y media y 3 horas entre las 12 y las 20 horas. esta información se puede confirmar en las reglas 3 y 4 de la siguiente figura.

```
[1]  {MODELO=737-8H4,
      FRANJA_HORARIA_SALIDA=Tarde}    => {IATA_AEROLINEA=WN} 0.01586
1 4.779201     162
[2]  {MODELO=737-8H4,
      FRANJA_HORARIA_SALIDA=Manana}   => {IATA_AEROLINEA=WN} 0.01260
1 4.779201     163
[3]  {MODELO=737-823,
      FRANJA_HORARIA_SALIDA=Tarde,
      TIEMPO_TOTAL_VUELO=medio}       => {IATA_AEROLINEA=AA} 0.01102
1 7.122507     223
[4]  {MODELO=737-823,
      FRANJA_HORARIA_SALIDA=Tarde}    => {IATA_AEROLINEA=AA} 0.01802
1 7.122507     227
```

Figura 18: algunas reglas generadas con FRANJA_HORARIA_SALIDA en el antecedente, ordenadas por "lift"

4.2.7. Reglas con *TIEMPO_TOTAL_VUELO*

La información obtenida con las reglas en las que el tiempo total de vuelo aparece en el antecedente coincide con algunas de las presentes en el punto 4.2.6, por lo que no aportan información relevante. Por ello, no podemos afirmar, por ejemplo, que los vuelos de mayor duración tienden a tener retrasos frente a aquellos de duración menor.

4.2.8. Reglas con itemsets frecuentes maximales

También se han realizado, para los dos algoritmos de generación de reglas, reglas para las que se han utilizado itemsets frecuentes maximales. Obtenemos, de nuevo, información sobre los modelos utilizados por las aerolíneas en determinados tipos de vuelo. Podemos decir, por ejemplo, que la aerolínea "F9" utiliza aviones *Airbus A320-251N* para sus vuelos entre una hora y media y 3 horas de duración. En este punto se muestran las reglas generadas tanto por el algoritmo *apriori* (figura 19) como por el algoritmo Eclat (figura 20).

```
     lhs                                rhs                    support confiden
ce      lift itemset
[1]  {MODELO=A320-251N,
      TIEMPO_TOTAL_VUELO=medio}     => {IATA_AEROLINEA=F9}    0.01070  1.00000
00 39.215686     223
[2]  {MODELO=A320-271N}             => {IATA_AEROLINEA=NK}    0.01276  1.00000
00 24.850895      31
[3]  {MODELO=ERJ 190-100 IGW}       => {IATA_AEROLINEA=B6}    0.01260  1.00000
00 23.169601      30
[4]  {IATA_AEROLINEA=9E,
      TIEMPO_TOTAL_VUELO=medio}     => {MODELO=CL-600-2D24}   0.01526  0.91596
64 16.653934     236
[5]  {IATA_AEROLINEA=9E,
      FRANJA_HORARIA_SALIDA=Tarde} => {MODELO=CL-600-2D24}   0.01450  0.81369
25 14.794409     235
[6]  {MODELO=ERJ 170-200 LL}        => {IATA_AEROLINEA=OO}    0.01158  1.00000
00 10.570825      26
```

Figura 19: algunas reglas generadas con itemsets frecuentes maximales siguiendo el algoritmo a priori, ordenadas por "lift"

```
     lhs                                rhs                    support confiden
ce      lift itemset
[1]  {MODELO=A320-251N,
      TIEMPO_TOTAL_VUELO=medio}     => {IATA_AEROLINEA=F9}    0.01070  1.00000
00 39.215686      22
[2]  {MODELO=A320-271N}             => {IATA_AEROLINEA=NK}    0.01276  1.00000
00 24.850895      11
[3]  {MODELO=ERJ 190-100 IGW}       => {IATA_AEROLINEA=B6}    0.01260  1.00000
00 23.169601      10
[4]  {IATA_AEROLINEA=9E,
      TIEMPO_TOTAL_VUELO=medio}     => {MODELO=CL-600-2D24}   0.01526  0.91596
64 16.653934     129
[5]  {IATA_AEROLINEA=9E,
      FRANJA_HORARIA_SALIDA=Tarde} => {MODELO=CL-600-2D24}   0.01450  0.81369
25 14.794409     130
[6]  {MODELO=ERJ 170-200 LL}        => {IATA_AEROLINEA=OO}    0.01158  1.00000
00 10.570825       6
```

Figura 20: algunas reglas generadas con itemsets frecuentes maximales siguiendo el algoritmo Eclat, ordenadas por "lift"

4.2.9. Reglas con itemsets frecuentes cerrados

Además de las reglas generadas con los itemsets frecuentes maximales, también se han generado aquellas para el resto de itemsets frecuentes cerrados. De nuevo, encontramos patrones que muestran información sobre los modelos más utilizados por las aerolíneas.

```
     lhs                                  rhs                  support confiden
ce      lift itemset
[1]  {MODELO=A320-251N}                => {IATA_AEROLINEA=F9}  0.01880  1.00000
00 39.215686     124
[2]  {MODELO=A320-251N,
      TIEMPO_TOTAL_VUELO=medio}        => {IATA_AEROLINEA=F9}  0.01070  1.00000
00 39.215686     549
[3]  {MODELO=A320-271N}                => {IATA_AEROLINEA=NK}  0.01276  1.00000
00 24.850895     113
[4]  {MODELO=ERJ 190-100 IGW}          => {IATA_AEROLINEA=B6}  0.01260  1.00000
00 23.169601     112
[5]  {IATA_AEROLINEA=9E,
      TIEMPO_TOTAL_VUELO=medio}        => {MODELO=CL-600-2D24} 0.01526  0.91596
64 16.653934     562
[6]  {IATA_AEROLINEA=9E,
      FRANJA_HORARIA_SALIDA=Tarde}     => {MODELO=CL-600-2D24} 0.01450  0.81369
```

Figura 21: algunas reglas generadas con itemsets frecuentes cerrados siguiendo el algoritmo a priori, ordenadas por "lift"

```
     lhs                                  rhs                  support confiden
ce      lift itemset
[1]  {MODELO=A320-251N,
      TIEMPO_TOTAL_VUELO=medio}        => {IATA_AEROLINEA=F9}  0.01070  1.00000
00 39.215686      22
[2]  {MODELO=A320-251N}                => {IATA_AEROLINEA=F9}  0.01880  1.00000
00 39.215686      23
[3]  {MODELO=A320-271N}                => {IATA_AEROLINEA=NK}  0.01276  1.00000
00 24.850895      11
[4]  {MODELO=ERJ 190-100 IGW}          => {IATA_AEROLINEA=B6}  0.01260  1.00000
00 23.169601      10
[5]  {IATA_AEROLINEA=9E,
      TIEMPO_TOTAL_VUELO=medio}        => {MODELO=CL-600-2D24} 0.01526  0.91596
64 16.653934     144
[6]  {IATA_AEROLINEA=9E,
      FRANJA_HORARIA_SALIDA=Tarde}     => {MODELO=CL-600-2D24} 0.01450  0.81369
```

Figura 22: algunas reglas generadas con itemsets frecuentes cerrados siguiendo el algoritmo Eclat, ordenadas por "lift"

Conclusiones

En el apartado 1 se ha demostrado que aunque salgan unos resultados que apuntan a que el modelo no se ajuste bien a los datos, no es indicativo de que el proceso esté mal, sino que el modelo utilizado no es el adecuado para dichos datos, o que simplemente los datos no tienen ningún patrón a analizar según una regresión lineal.

En el apartado 2 se ha conseguido separar por grupos los aeropuertos origen y destino, aerolínea y modelo dado el retraso medio total. Esto puede servir para conseguir separar unos aeropuertos de otros, lo que facilita la implementación de enfoques y estrategias específicas para cada grupo.

En el apartado 3 se comprueba cómo las variables cualitativas pueden tener influencia en ciertas variables cuantitativas, determinando en nuestro caso que la franja horaria en la que salen o llegan los vuelos influye en los distintos tiempos de retraso.

En el apartado 4 se observa cómo el uso de reglas de asociación muestra información sobre patrones presentes en los datos. En nuestro caso, principalmente se ha obtenido información sobre aquellos modelos más utilizados por las aerolíneas, así como ciertas aerolíneas que operan de manera muy definida, con sólo un itinerario y haciendo uso de un único modelo de avión, siguiendo así el modelo de negocio de las aerolíneas *low-cost*.

Tarea	**Jan**	**Simón**	**Óscar**
Problema 1	4 horas		
Problema 2	4 horas		
Problema 3	4 horas		
Problema 4	6 horas		
Realización de la memoria	2 horas		
Total: horas	20 horas		

Tabla 2: esfuerzos realizados durante esta práctica

Anexo

Código relevante apartado 1

```
# Iterar sobre diferentes números de predictores
for (num_predictores in 1:num_total_predictores) {
  # Obtener combinaciones de predictores
  combinaciones <- combn(predictores, num_predictores)

  # Iterar sobre cada combinación
  for (i in 1:ncol(combinaciones)) {
    counter <- counter + 1
    predictors_subset <- combinaciones[, i]
    numero_predictores <- length(predictors_subset)

    cat("Iteración:", counter, "Predictores:", predictors_subset, "\n")

    # Crea la fórmula de regresión
    formula <- as.formula(paste("RETRASO_TOTAL ~", paste(predictors_subset, collapse =
"+")))

    # Guardamos la formula para saber luego cual es el mejor modelo
    formula_predictores <- as.character(paste(predictors_subset, collapse = "+"))

    model <- lm(formula, data = datos)
    summary(model)
    # Obtener R^2
    r2 <- summary(model)$r.squared

    # Comprobamos para el mismo numero de predictores, si el r2 es mayor, en caso
afirmativo este es el nuevo
    # mejor modelo para ese número de predictores
    if (r2 > r2_modelos[numero_predictores]) {
      mejores_modelos[[numero_predictores]] <- formula_predictores
      r2_modelos[numero_predictores] <- r2
      r2Adj_modelos[numero_predictores] <- summary(model)$adj.r.squared
      BIC_modelos[numero_predictores] <- BIC(model)
      AIC_modelos[numero_predictores] <- AIC(model)
    }
  }
}
```

Código relevante apartado 2

```
# a) aeropuerto origen
datos$IATA_AEROPUERTO_ORIGEN <- as.factor(datos$IATA_AEROPUERTO_ORIGEN)
# Obtenemos la media
medias <- aggregate(RETRASO_TOTAL ~ IATA_AEROPUERTO_ORIGEN, data = datos, FUN = mean)
combined_data <- cbind(as.numeric(medias$IATA_AEROPUERTO_ORIGEN), medias$RETRASO_TOTAL)

# Realizamos el kmeans
kmeans_result <- kmeans(combined_data, centers = 3, nstart = 50)
```

```
# Ploteamos los puntos
plot(medias$RETRASO_TOTAL ~ medias$IATA_AEROPUERTO_ORIGEN,
     main = "3-Media aeropuerto origen",
     xlab = "Aeropuerto origen", ylab = "Media retraso total",
     col = kmeans_result$cluster, pch = 19)
```

Código relevante apartado 3

```
print("Analisis ANOVA modeloAnovaTL")
modeloAnovaTL=aov(RETRASO_TOTAL~FRANJA_HORARIA_LLEGADA,data=datos)
print(summary(modeloAnovaTL))
print("Prueba de Tukey modeloAnovaTL")
print(TukeyHSD(modeloAnovaTL))
plot(TukeyHSD(modeloAnovaTL))
```

Código relevante apartado 4

4.1.

```
# Generar reglas mediante apriori
# minsup = 0.01, minconf = 0.8 y reglas contienen a lo sumo 10 items
reglas_vuelos_a_priori = apriori(transacciones_vuelos,
                       parameter = list(support = 0.01, confidence = 0.8, maxlen = 10))
```

4.2.1

```
reglas_dia_semana_cons_a_priori = subset(reglas_vuelos_a_priori, rhs %pin% c("DIA_SEMANA"))
# Ninguna regla, nos descartamos que existan reglas con dia de la semana como consecuente
inspect(head(sort(reglas_dia_semana_cons_a_priori, by = "lift"), 10))
reglas_dia_semana_cons_eclat = subset(reglas_vuelos_eclat, rhs %pin% c("DIA_SEMANA")) # Ninguna
regla, nos descartamos que existan reglas con dia de la semana como consecuente
inspect(head(sort(reglas_dia_semana_cons_eclat, by = "lift"), 10))
```

```
reglas_dia_semana_aeropuerto_origen_a_priori =
subset(reglas_vuelos_a_priori, lhs %pin% c("DIA_SEMANA") & rhs %pin%
c("IATA_AEROPUERTO_ORIGEN") & lift > 2)
inspect(head(sort(reglas_dia_semana_aeropuerto_origen_a_priori, by =
"confidence"), 10))
```