TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Análisis del mercado aéreo doméstico de Estados Unidos

Autor: Augusto Cortejana Retamozo

Directores: Lluís Marco y Montserrat Termes

Departamento: Departamento de Estadística e Investigación Operativa/ Departamento de Econometría, Estadística y Economía Aplicada

Convocatoria: Junio 2018

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Análisis del mercado aéreo doméstico de Estados Unidos

Augusto Cortejana Retamozo

29 de junio de 2018

Resumen

El presente trabajo analiza el mercado aéreo doméstico de los Estados Unidos. Primero de todo, se investiga la evolución de su estructura y grado de competencia de las aerolíneas en los últimos años a través de los indicadores de concentración, volatilidad y variables claves del propio mercado como el factor de carga o precio de los billetes.

Después se procede a explotar la base de datos de los 5,8 millones de vuelos domésticos en el año 2015 a partir de gráficos, tests y el Análisis de Correspondencias Múltiples. Debido al gran coste que suponen los retrasos para toda la economía, se realiza un modelo logístico para clasificar el estado de un vuelo y predecir la probabilidad que un vuelo sufra retraso.

Por último, se crea una aplicación dashboard en Shiny orientada a los pasajeros donde pueden consultar información de las aerolíneas, aeropuertos y rutas.

Palabras clave*: Indicadores de concentración, Mercado aéreo, Competencia, regresión logística, Mineria de datos, aplicación Shiny*

Abstract

The following project analyzes the USA domestic air industry. First, the evolution of the structure and level of competence between airlines in the last fifteen years is analyzed through concentration, volatility indexes and key factors of the market such as the passenger load factor and the airfares.

After that, using graphical tools, statistical tests and the Correspondence Multiple Analysis the 5.8 million flights database of 2015 is examined. Due to the big costs that flight delays imply for all the economy, a logistic regression is built to classify the flight status and predict the probability of delays.

In the last chapter, a Shiny dashboard app is implemented for the passengers to collect more information about airlines, airports and flight routes.

Key words: *concentration and volatility indexes, air industry, competence, logistic regression, data mining, Shiny app*

Clasificación AMS

62-07 Data analysis

62-09 Graphical methods

62H15 Hypothesis testing

91B82 Statistical methods; economic indices and measures

97K40 Descriptive statistics

62J12 Generalized linear models

62H25 Factor analysis and principal components; correspondence analysis

ÍNDICE

INTRODUCCIÓN

El mercado aéreo está formado por las aerolíneas que operan rutas con el fin de transportar pasajeros. A partir de la liberalización del mercado en la década de los 70, el transporte aéreo ha tenido una evolución espectacular y cada vez más viajar en avión es más común en la sociedad actual.

El presente trabajo se centra en el mercado aéreo doméstico de los Estados Unidos, que consta de las rutas cuyo origen y destino son dentro del mismo país.

Según la Federal Aviation Administration (en adelante FAA), un vuelo se considera oficialmente retrasado cuando llega al menos 15 minutos después de la hora prevista y se considera cancelado cuando no logra iniciar el despegue. Es por eso por lo que, en el presente trabajo, cuando se hable de un vuelo en retraso o cancelado se hace referencia a la definición oficial de la FAA.

Existen 5 causas oficiales de retraso y de cancelación:

-Sistema aéreo: Retrasos y cancelaciones atribuidas al sistema de aviación nacional referentes a un amplio conjunto de condiciones como: condiciones no extremas de mal tiempo, operaciones del aeropuerto, control del tráfico aéreo y volumen denso del tráfico.

-Seguridad: Retrasos y cancelaciones causados por la evacuación de una terminal, violación de la seguridad, equipos de control inoperativos y colas de espera con más de 29 minutos de espera en las áreas de detección (screening areas).

-Aerolínea: circunstancias que la aerolínea tiene control como los problemas con la tropa, mantenimiento del avión, problemas con el equipaje, etc.

-Tiempo: Condiciones meteorológicas extremas (actuales o previstas) que, según la aerolínea, retrasan o evitan la operación de un vuelo como tornados, tormentas de nieves.

-Avión en retraso: Se refiere a que el avión anterior llegó tarde y produjo retrasos a los posteriores vuelos.

Es necesario aclarar que el retraso en la llegada de un avión puede ser debido a una sola de estas causas o puede tener más de una causa asociada.

Justificación

A la hora de elegir el tema para el trabajo final de carrera, me centré en dos requisitos claves: temáticas que me resultasen interesantes poder analizar y sobre todo poder demostrar los conocimientos que he adquirido a lo largo de mi formación tanto de economía como de estadística.

El análisis del mercado aéreo doméstico de los Estados Unidos cumplía con estos dos requisitos dada la inmensa información que se puede extraer.

Estructura del trabajo

A continuación, se explica la estructura del presente trabajo:

En el **capítulo 1**, se determinará el grado de competencia que hay en el caso de las aerolíneas de los Estados Unidos a partir de la evolución que han tenido los indicadores de concentración, volatilidad, así como otras variables propias del mercado doméstico aéreo en los últimos 15 años.

Después de haber presentado la estructura y competencia del mercado, en el **segundo capítulo** se procede al análisis exploratorio del producto que ofrecen las aerolíneas: los vuelos. Mediante, técnicas descriptivas, inferenciales y gráficas se analiza que factores determinan el estado de un vuelo, donde y cuando se concentran los retrasos más importantes, que aerolíneas son las mejores para según qué ruta, etc.

Después de este análisis exploratorio de los vuelos y debido al gran coste económico e ineficiencia que provocan los retrasos, en **el capítulo 3** se procede a la construcción de una regresión logística que prediga si un vuelo sufrió o no retraso utilizando solo la información que se dispone antes de volar.

En el **último capítulo**, se implementa una aplicación web que muestra de forma interactiva los resultados más interesantes del análisis. Esta aplicación está orientada a pasajeros que quieran informarse de cualquier ruta, aerolínea, aeropuerto y probabilidad de que su vuelo sufra retraso.

Metodología

Para el presente trabajo se analizarán los datos principalmente con el software Rstudio que utiliza el lenguaje de programación R: los paquetes ggplot2 (para la creación de gráficos), tidyverse (orientando al data science y manipulación de bases de datos), leaflet (para la creación de mapas) y shiny (para la construcción del dashboard) han sido los más utilizados. También se ha utilizado el programa Microsoft Excel.

En el capítulo 2 se utilizan tests estadísticos a partir de una muestra de 100000 registros. Al tratarse de una base de datos de tamaño enorme, la potencia del test será bastante alta y por lo tanto el test será muy sensible en detectar diferencias significativas. Como estadísticamente significativo no implica importante, siempre que sea posible se adjuntan también intervalos de confianza.

En casos de comparaciones múltiples 2 a 2 se ha aplicado la corrección conservadora de Bonferroni.

Objetivo

El objetivo principal de este trabajo es caracterizar la estructura y grado de competencia del sector aéreo doméstico de las aerolíneas de los Estados Unidos e intentar explicar el comportamiento de los vuelos: ver si hay comportamientos diferenciadores entre aerolíneas, donde se concentran los retrasos y que factores afectan más a los retrasos y predecir la probabilidad de retraso con la información a priori.

Agradecimientos

Por último, agradecer a mis tutores Lluís Marco y Montserrat Termes por el tiempo que me han dedicado cuando tenía dudas, a mi familia, pareja y amigos por apoyarme y darme ánimos en la realización de este trabajo.

CAPÍTULO I.
ANÁLISIS ECONÓMICO DEL MERCADO AÉREO DOMÉSTICO DE ESTADOS UNIDOS

La economía industrial es la rama de la economía que estudia la forma en la que actúan las fuerzas del mercado, el comportamiento de sus agentes y los resultados concretos que estos obtienen de un determinado sector económico. (Tirole,1988). En esta rama, medir el grado de competencia de un sector es clave. Se entiende por competencia a la existencia de un número de empresas que ofrecen productos en un mercado donde los consumidores demandan estos productos a los ofertantes y que no pueden incidir en los precios.

Para Schumpeter (1940), economista austro estadounidense del siglo pasado, la competencia de las empresas toma una dimensión dinámica, es decir, además de la estructura del mercado, que se suele medir con el grado de concentración, la movilidad de las empresas también es un factor clave para explicar el grado de competencia. Mejor dicho, un mercado en donde entren y salgan empresas es un mercado en donde la competencia funciona bien. Es lo que él denominó "la destrucción creativa".

Pero, en realidad, la concentración y la movilidad, aunque sean condiciones necesarias, no son suficientes para determinar el grado de competencia en un sector y si la competencia entre empresas funciona correctamente. Para ello, también hace falta analizar otras variables como la evolución de los costes y precios a lo largo del tiempo o ver en qué fase se encuentra el mercado, es decir, si está en crecimiento o se ha estancado. (Goreki,1989). Este último paso, muchas veces no se realiza y por lo tanto se construyen conclusiones en base solo a ver la concentración y movilidad.

Es por eso por lo que, en este primer capítulo, se analizará el grado de competencia de las aerolíneas que configuran el mercado aéreo doméstico de los Estados Unidos desde una perspectiva dinámica. Por ello, para no acabar en conclusiones erróneas, se calcularán no solo los índices de concentración, sino también los de volatilidad para el conjunto de aerolíneas y también se analizarán otras variables asociadas al propio mercado aéreo.

El siguiente análisis se ha llevado a cabo para el período 2002-2016. Los datos utilizados en este capítulo provienen de 4 fuentes:

- El Bureau of Transportation Statistics (la agencia federal de estadísticas de transporte de los Estados Unidos, en adelante, BTS)
- La International Civil Aviation Organization (organización de la ONU que entre otras funciones, publica indicadores del sector aéreo; en adelante ICAO)
- Airlines for America (A4A, organización comercial de las principales aerolíneas de Estados Unidos,en adelante A4A)
- International Air Transport Association (Es el instrumento para la cooperación entre aerolíneas, promoviendo la seguridad, fiabilidad, confianza y economía en el transporte aéreo, en adelante IATA)

1.1 Índices de concentración

La concentración se define como el grado en que un número de empresas ejerce el control o dirección de la actividad económica de un sector (Sunkel, Geoffroy,2001). Por lo tanto, conocer el grado de concentración de un mercado es vital para saber cómo está estructurado: en el caso de concentración máxima estaríamos hablando de un monopolio (una sola empresa domina el mercado), si pocas empresas dominan el mercado entonces estamos ante un mercado oligopolizado, y, por último, la situación de concentración mínima corresponde al caso de competencia perfecta (Sunkel, Geoffroy,2001).

Para determinar la estructura de mercado, es muy común en los estudios utilizar los índices de concentración. Éstos índices sirven para determinar si la distribución en el reparto total del mercado entre empresas es igualitaria o no mediante el cálculo de la cuota de mercado que cada empresa tiene:

$$Cuota\ de\ la\ empresa_i = S_i = \frac{X_i}{\sum_i^n X_i}$$

donde X_i representa la variable a partir de la cual se calcula la cuota y el sumatorio es el valor agregado de X de todas las n empresas del mercado. A continuación, se explica brevemente los indicadores de concentración para el análisis del presente trabajo

1.1.1 Ratios de concentración CR_k.

El indicador de concentración más sencillo es el denominado CR_k. Esta ratio acumula las cuotas de mercado de las k mayores empresas de un sector. (Martínez, 2008) Por ejemplo CR_4 calcula la suma de las cuotas de mercado de las 4 empresas mayores que operan en este sector.

$$CR_k = \sum_{i=1}^{k} S_i$$

Éste índice pondera equitativamente las cuotas de las empresas mayores y el valor del índice depende del valor k que elijamos a tener en cuenta, dando lugar a que este índice puede llevarnos a conclusiones diferentes según el número de empresas que se tengan en cuenta. Es por esta principal razón por la que se recomienda calcularlo para distintos valores de k.

Aun así, es de los más utilizados ya que no se necesita la información de todas las empresas del mercado (ya se ignora la información de las $n - k$ empresas mayores). El índice varía entre $k/\ n$ (mínima concentración) y 100 (concentración máxima, caso de un monopolio absoluto). Para decidir el grado de concentración se suele utilizar el CR_4: Si este es mayor que 50, estamos ante un mercado con alta concentración.

La representación del CR_k se plasma en la curva de concentración que une los valores de CR para todo valor de k. Si todas las cuotas fueran idénticas, la representación gráfica del CR seria la bisectriz de 45º. Cuanto más concentrado esté un mercado, la curva del indicador está mas alejada por encima de esta bisectriz. En el caso de la figura 1.1, el mercado D es el menos concentrado y el A; el que más.

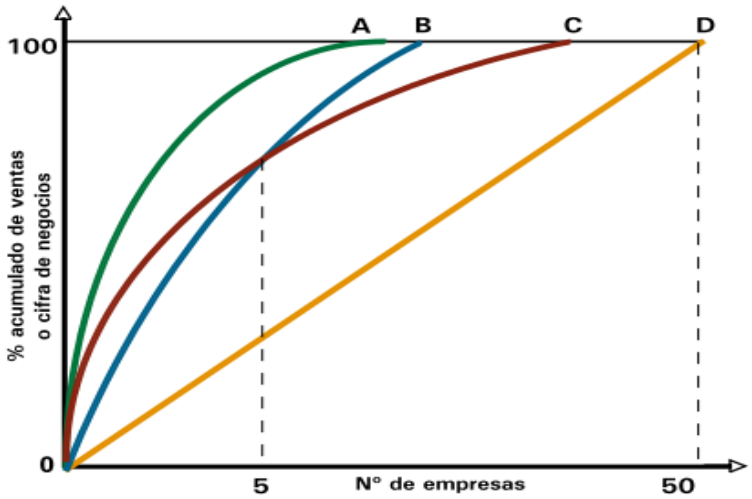


Figura 1.1: Ejemplo de curva de concentración

1.1.2 Índice Herfindahl- Hirschman

El segundo índice que se utilizará para el análisis es el índice de Herfindahl- Hirschman. A diferencia del CR_k, en este caso se tiene en cuenta todas las n empresas que componen el mercado (Martínez,2008).

$$HH = \sum_{i=1}^{n} S_i^2 = \frac{\sigma^2 + 1}{n}$$

El índice se calcula con la suma de todas las cuotas de mercado elevadas al cuadrado. Para decidir el grado de concentración se suele utilizar valores límites: Si HH es superior a 1000 y menor a 1800, estamos ante un mercado con grado medio de concentración y si es superior a 1800, grado alto.

La gran ventaja es que en el cálculo se tiene en cuenta el número de empresas (a mayor numero, menor concentración) y también el grado de dispersión que existe entre las cuotas de mercado (σ^2). Pero puede presentar problemas de ponderaciones excesivas para empresas de mayor cuota.

1.2 Medidas de volatilidad

La gran limitación de todos los indicadores de concentración es que son medidas estáticas, es decir, no tienen en cuenta la evolución de la concentración a lo largo del tiempo de forma implícita, imposibilitando así, el análisis dinámico que defiende Schumpeter (Hannah y Kay,1977).

El análisis puramente estático de un mercado puede conducir a conclusiones erróneas y es por eso que se necesitan también medidas dinámicas para saber cómo varia la participación de las empresas a lo largo del tiempo y también poder tener en cuenta la posible entrada y salida de empresas en el mercado. (Mazzucato, Semmler, 2002)

El índice de volatilidad que se utilizará es el siguiente:

$$I = \frac{1}{2}\sum_{i=1}^{n} |S_{i2} - S_{i1}|$$

Donde S_{ik} representa la cuota de mercado de la empresa i en el periodo k

Se trata de las semisumas de los valores absolutos de las diferencias entre las cuotas de mercado de las n empresas entre dos años.

Si I toma valor 0, estamos ante un caso de máxima estabilidad en términos dinámicos (las cuotas de todas las empresas no varían respecto el periodo 1) y si I vale 1 se trata de un mercado en donde en el periodo 2 las empresas que operaban en el periodo 1 abandonan el mercado (máxima inestabilidad).

1.3 Análisis de la estructura del mercado aéreo

1.3.1 Cálculo de los índices de concentración en el mercado aéreo

Para el cálculo de las cuotas de mercado de las aerolíneas se ha decidido optar por la variable "Número de pasajeros" a diferencia de otros estudios que utilizan los ingresos de facturación o el volumen de las ventas de billetes.

El uso del volumen de las ventas puede provocar un sesgo en la información cuando los productos (en este caso los vuelos) no son homogéneos en cuanto a características (Goreki,1989). En un mercado como el aéreo, en que los vuelos que ofrecen las distintas compañías varían según precio, clase, trayecto, estamos hablando de productos heterogéneos.

Todos los datos de este apartado provienen de un fichero del BTS donde cada fila representa los datos de una aerolínea para un determinado año. En la página web puedes seleccionar el tipo de información que se desea: en este caso, es suficiente con el número de pasajeros para cada año. El número de total de aerolíneas en el fichero es de 211.

Con esta información se ha podido calcular las cuotas de mercado de todas las aerolíneas de la siguiente manera:

$$S_i = \frac{Nº\ de\ pasajeros\ transportados\ por\ la\ aerolinea_i}{Total\ de\ pasajeros\ transportados}\ x\ 100$$

En este mercado operan más de 200 aerolíneas, pero para este análisis se ha decidido reducir este número para no sesgar los resultados dado que:

-Algunas aerolíneas representaban una cuota casi nula del total (son, por ejemplo, aerolíneas que tienen menos de 100 pasajeros al año, o que se centran solo en una zona muy reducida del mercado).

-Existen aerolíneas regionales que operan en nombre de aerolíneas nacionales. Lo correcto en estos casos es tener en cuenta la cuota de mercado global de la aerolínea decisora (Bergensen,2002).

También es importante remarcar que el número de aerolíneas no es constante a lo largo del período como se estudiará más adelante. A medida que han pasado los años, el número de empresas que configura el mercado aéreo doméstico ha ido reduciéndose. Es decir, en 2002 había 53 aerolíneas, pero en el año 2016, el número se redujo a 28.

Con el software R, se realizó una función que calculase el valor de **la ratio de concentración** para cualquier año y valor de k. Los resultados, para 6 valores diferentes de k se muestran a continuación:

Tabla 1.1 Valores del CR

k	2002	2005	2010	2016
1	12,12	13,39	16,81	20,62
2	23,10	25,13	31,07	37,13
4	49,47	52,25	55,47	68,85
10	61,61	71,05	73,46	82,48
20	79,81	88,95	89,60	96,05
53	100	100	100	100

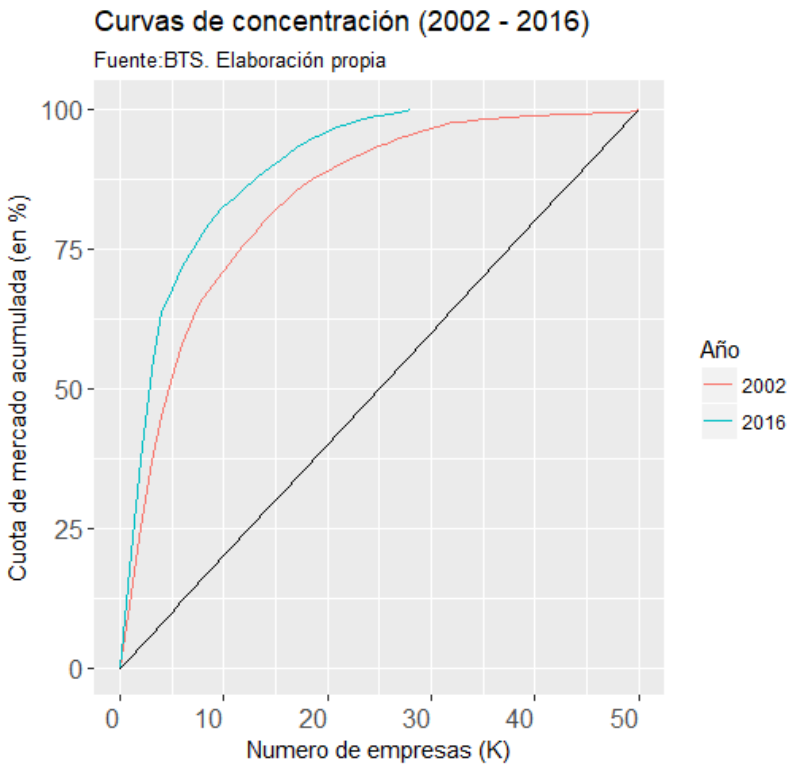


Figura 1.02: Curvas de concentración

Se observa que a medida que pasan los años, ha habido un aumento en la concentración del mercado dado que, en el año 2002, las 4 aerolíneas más grandes (Southwest, United Air, Delta y American Airlines) dominaban casi el 50% del mercado, pero en el año 2016, este porcentaje fue casi de un 70%. De igual forma, para cada año se llega a la misma conclusión con el indicador CR: el mercado aéreo doméstico presenta una concentración elevada. La mayoría

de las aerolíneas presentes obtienen unas cuotas muy inferiores ya que son aerolíneas que no pueden hacer frente a la superioridad de las cuatro más grandes y que, por lo general, no operan vuelos por todo el territorio estadounidense.

De manera gráfica, esto quiere decir que la curva de concentración (las ratios de concentración para cada valor posible de k) está significativamente por encima de la línea de concentración igual a cero (bisectriz 45º) y que, con el paso de los años, la curva ha ido alejándose cada vez más, como se observa en la figura 1.2.

Debido a las limitaciones de las ratios de concentración explicadas en el anterior apartado, se calcula a continuación el **índice de Herfindahl** para cada año, el cual se obtiene con la información de todas las aerolíneas que configuran el mercado:

Tabla 1.2: Valores del Índice de Herfindahl

	2002	**2005**	**2010**	**2015**	**2016**
HHI	652,21	693,66	805,8	1061,4	1148,3
N	53	37	34	30	28
σ	2,08	2,19	2,58	3,02	3,11

En la tabla 1.2 se adjunta el valor del índice de Herfindahl, así como el número de aerolíneas y una medida de la disparidad entre las cuotas de mercado (la desviación típica).

El valor del índice ha aumentado considerablemente a lo largo del tiempo (casi el doble respecto al año 2002) dado que la disparidad entre cuotas ha aumentado y el número de aerolíneas en el mercado ha disminuido. A la vista de los resultados, se puede determinar que, a partir del año 2015, el mercado aéreo ya se encuentra en un grado medio de concentración dado que el índice HH es superior a 1000.

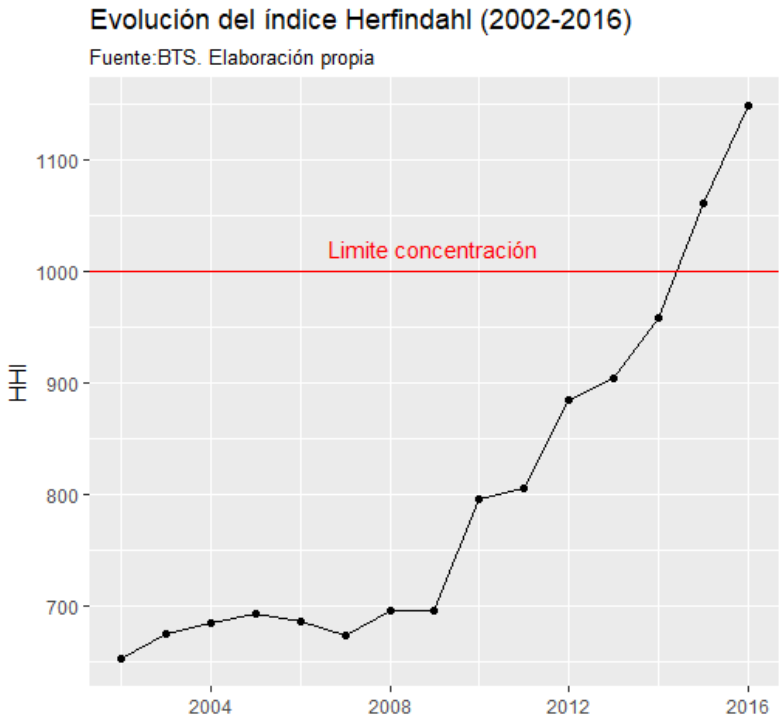


Figura 1.3: Evolución HHI

Existen dos motivos que explican la reducción del número de aerolíneas y se explicarán con más detalle en el siguiente apartado (Williams,2017): el primero es que algunas dejaron de operar por problemas financieros y quiebra (es el caso de Aloha Airlines, Colgan Air, Independence Air,Comair Inc… entre otras) y el segundo motivo son las fusiones entre aerolíneas nacionales y adquisiciones de aerolíneas grandes con aerolíneas regionales como las de Delta con Northwest en el año 2009, Southwest con AirTran Airways y US Airways con American Airlines en 2014. A finales de 2016, ha tenido lugar la última de las fusiones: Alaska y Jetblue.

Para incorporar el efecto de la salida de las aerolíneas al análisis y llevar a cabo un análisis más dinámico, a continuación, se calculará el índice de volatilidad I. Para ello, se elige como período 1, el año 2002 y como segundo período, el año 2016.

Para cada aerolínea, se calcula la diferencia absoluta entre la cuota de mercado que obtuvo en 2002 y la que obtuvo en 2016. Las aerolíneas que han desaparecido ya sea por quiebra o por adquisición tienen una cuota en 2016 de 0 y las aerolíneas que no existían en 2002 pero sí en 2016 tienen asignadas unas cuotas en 2002 iguales a 0.

Para el mercado aéreo doméstico de EE. UU**, el índice I de volatilidad toma valor igual a 0,35**. Esto indica que el mercado presenta una volatilidad media baja. A pesar de las fusiones y adquisiciones que provocaron una reducción considerable del número de empresas, la distribución de las cuotas de mercado es bastante parecida entre 2002 y 2016, hecho que explica esta poca volatilidad en el sector.

Así y para concluir este apartado, se destaca que las ratios de concentración y el índice HHI del mercado aéreo doméstico señalan que está bastante oligopolizado ya que un reducido número de aerolíneas controlan la mayor parte del sector. Además, esta concentración no ha hecho más que aumentar a lo largo de los últimos años.

Según una perspectiva industrial dinámica, el sector presenta poca inestabilidad en las cuotas de mercado de las aerolíneas a lo largo del tiempo. Lo que Schumpeter denomina “destrucción creativa” no se ha dado en este sector ya que, en diez años, las aerolíneas que dominaban el sector en 2002 (Delta, Southwest, American Airlines y United Air) lo siguen haciendo en 2016 y con mayor fuerza.

1.3.2 Análisis de variables claves

Por último, se analiza la evolución de diversas variables para así entender mejor la estructura del mercado y como se ha llegado a tal punto.

En general, como se puede ver en la figura 1.4, el mercado aéreo doméstico ha tenido una evolución bastante positiva en cuanto a número de pasajeros: en el año 2002 volaron 550 millones de pasajeros y en el 2016, este dato se incrementó hasta los 740 millones (esto se traduce en una tasa anual de crecimiento medio del 2,2%). Es de destacar el descenso de los años 2007 y 2008 debido a la crisis financiera del 2007 pero en 2009 continuó la tendencia positiva y ya en 2015 superó los niveles anteriores a la crisis.

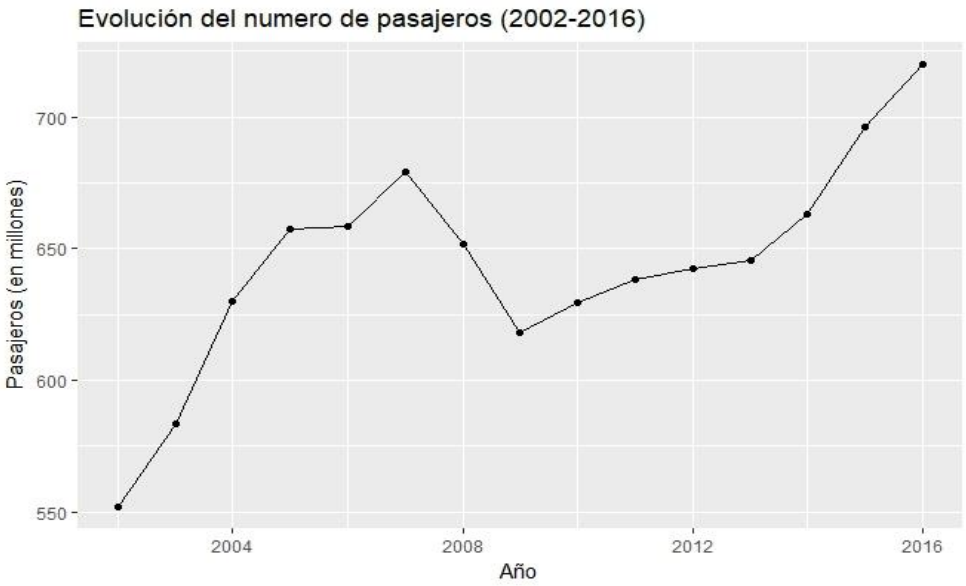


Figura 1.4: Evolución número de pasajeros

A pesar de esta evolución positiva en cuanto al número de pasajeros, el contexto para las aerolíneas era diferente: los ataques terroristas del 2001, la crisis de las .com y la crisis de 2007 produjo un ambiente de pérdidas financieras generales durante toda la primera década de los 2000. Nada comparable con la época de los 90 en que los costes eran más bajos debido al precio bajo del jet fuel y del dólar fuerte (Yglesias,2017). Es desde el año 2013 cuando las aerolíneas han vuelto a tener grandes beneficios como se verá a continuación. ¿Cómo se explica esto?

La estructura de costes de una aerolínea viene muy marcada por sus altos costes fijos: el coste de mantener un avión es elevado y lo sigue siendo independientemente del nivel de demanda que tenga (Doganis,2001). Es por eso que, en una época de caída de la demanda generalizada, es normal que haya habido grandes pérdidas financieras ya que los costes seguían siendo los mismos, mientras que los ingresos disminuían.

Estas pérdidas explican en gran parte, la gran reducción de aerolíneas desde el año 2002 (como se explicó en el apartado de los indicadores) ya que muchas se declararon en bancarrota (Prince, Simon,2014). Por otra parte, esto ha conducido a una época de fusiones y adquisiciones (algunas de las aerolíneas en bancarrota fueron rescatadas por aerolíneas existentes) como se puede ver a continuación en la figura 1.5.

El gráfico de la A4A corresponde a la evolución que han tenido las 4 aerolíneas dominantes: los puntos son períodos de bancarrota que por lo general vemos que inician fusiones y adquisiciones como es el caso de Nortwest con Delta y United con Continental. Por último, el año pasado se anunció la fusión de Virgin America con Alaska Airlines para el 2018.

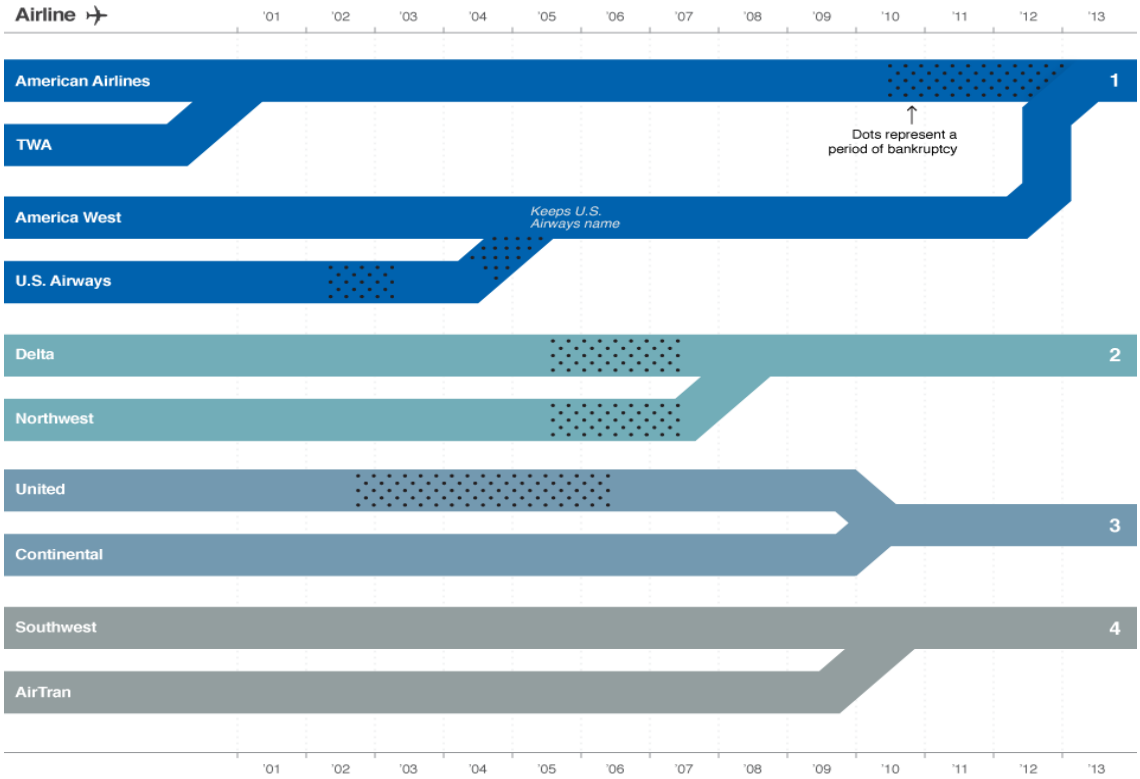


Figura 1.5: Evolución de las fusiones y adquisiciones de las principales aerolíneas (2001-2013)

Esta concentración de las aerolíneas ha tenido dos consecuencias (Moss,2014): la **primera es que el mercado aéreo esté dominado en un 70% por las 4 aerolíneas más grandes** como ya se vio anteriormente (de las 11 grandes aerolíneas de 2002, solo 5 siguen operando actualmente), y la **segunda consecuencia es el aumento de los beneficios totales de las aerolíneas** como veremos a continuación. A excepción del 2008 y 2009 donde existen grandes pérdidas por la crisis financiera, la evolución estos últimos años ha sido también positiva.

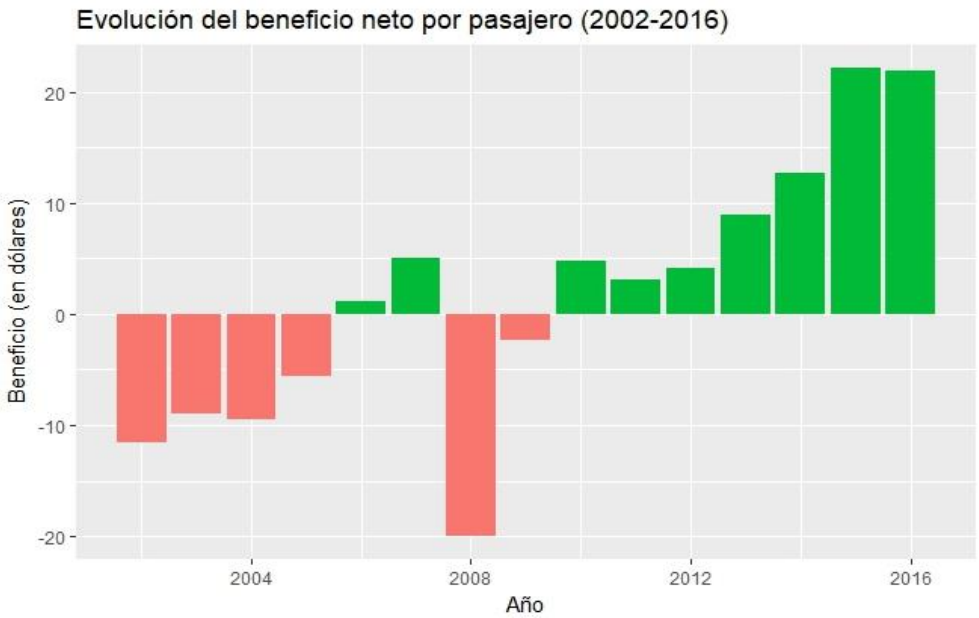


Figura 1.06: Evolución del beneficio neto por pasajero

Parece no ser de todo correcto decir que solo la concentración explica esta evolución de los beneficios ya que se ha de tener en cuenta otros factores (Helleloid,2015):

Para empezar, los costes, y en especial, los costes fijos tienen gran importancia para las aerolíneas: el 54% de los costes de una aerolínea grande vienen determinados por el mantenimiento y depreciación de los aviones. De estos costes, destaca sobre todo el combustible de aviación (el llamado jet fuel) que, según la ICAO, representa hasta una tercera parte de los costes totales. Su evolución es la siguiente:

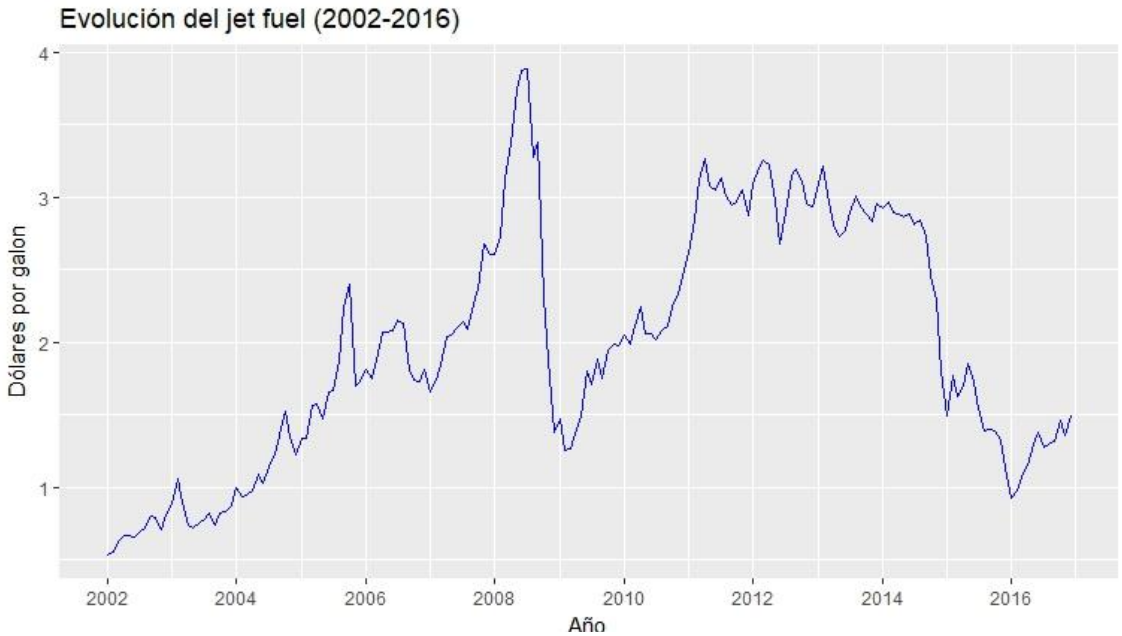


Figura 1.07: Evolución del jet fuel

Se observa un encarecimiento del jet fuel hasta el 2008 y después el precio cae a más de la mitad debido a la crisis financiera (pasa de 4 dólares a 1,5 dólares el galón). Esta gran caída no tiene un impacto positivo en los beneficios del mercado aéreo ya que las consecuencias de la

crisis financiera fueron mayores (el mercado aéreo tuvo unas pérdidas globales de más de 20 billones de dólares). Sin embargo, sí se observa que, a partir del 2012, el jet fuel ha ido abaratándose considerablemente: de los casi 3,5 dólares por galón en 2012, baja hasta menos de 2 dólares al inicio del 2016. Esta reducción de costes ha favorecido a la evolución positiva de los beneficios de las aerolíneas.

Por otra parte, el factor de carga de las aerolíneas ha aumentado. El factor de carga es una medida muy utilizada en los sectores relacionados con el transporte (en inglés, *passenger load factor* y en adelante, PLF) que sirve como una medida de eficiencia en la utilización del servicio aéreo para las aerolíneas debido a los altos costes fijos. El PLF mide el porcentaje de capacidad bajo el que vuelan los aviones: Un PLF alto indica que la aerolínea utiliza eficientemente sus aviones ya que estos vuelan bastante llenos, en cambio un PLF bajo indica que la aerolínea no está siendo eficiente en el uso de sus aviones ya que no utiliza toda la capacidad posible (Jadhav,2016).

La evolución del PLF para las 4 aerolíneas más importantes, junto con la media del sector aéreo, es la siguiente:

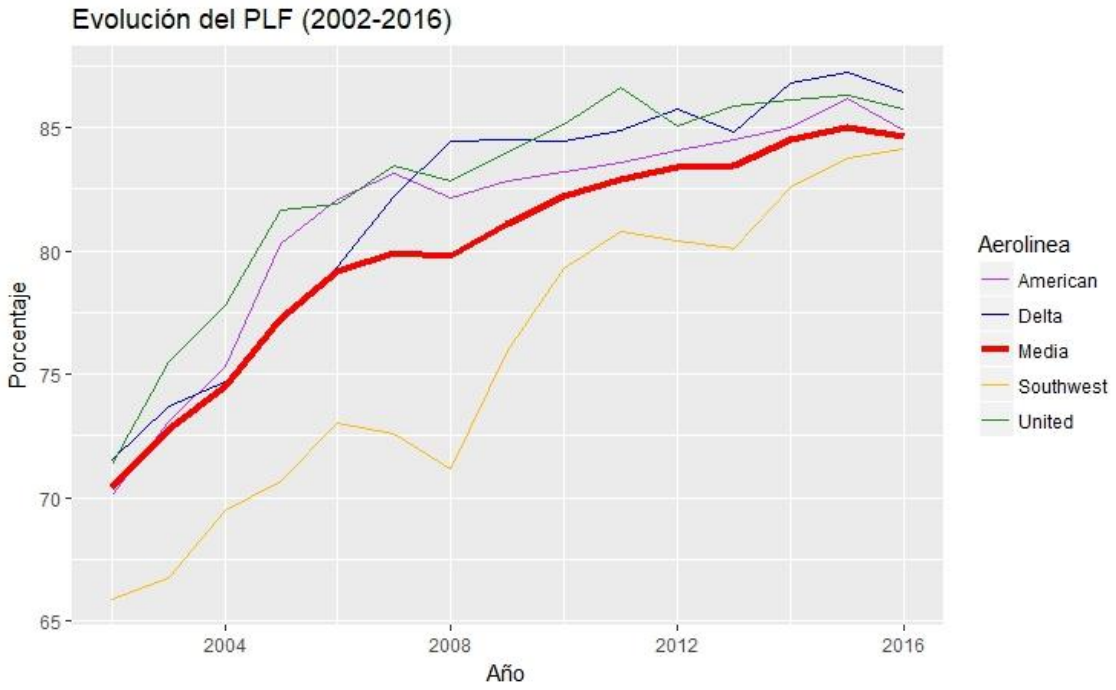


Figura 1.08: Evolución PLF

Se puede ver claramente como la evolución del PLF nos indica que las aerolíneas vuelan cada vez con aviones más llenos ya que aprovechan en un 85% su capacidad en el año 2016 a comparación del 2002 que la media era de un 70%. Las 4 aerolíneas han visto una gran mejora en la utilización de la capacidad, siendo la lowcost Southwest la que ha tenido un crecimiento mayor estos últimos años, pero por debajo de la media.

Como cita Helleloid en su artículo de 2015: "*la combinación de un entorno económico favorable con menores precios de jet fuel y sobre todo una reducción en el número de competidores junto a un aumento de los ingresos extras de los vuelos han contribuido a los beneficios récord del sector aéreo de los últimos años*".

Estos beneficios, no se han trasladado a los pasajeros ya que como se observa en este gráfico de evolución del precio medio de los billetes (en valor real para tener en cuenta la inflación

de cada año) desde el 2009, los precios han ido aumentando considerablemente coincidiendo con los años en que los costes de las aerolíneas bajaban.

A partir del 2015 vemos una leve bajada de los precios, pero el precio final se sitúa muy por encima del que había en 2002 cuando el mercado no estaba tan concentrado.

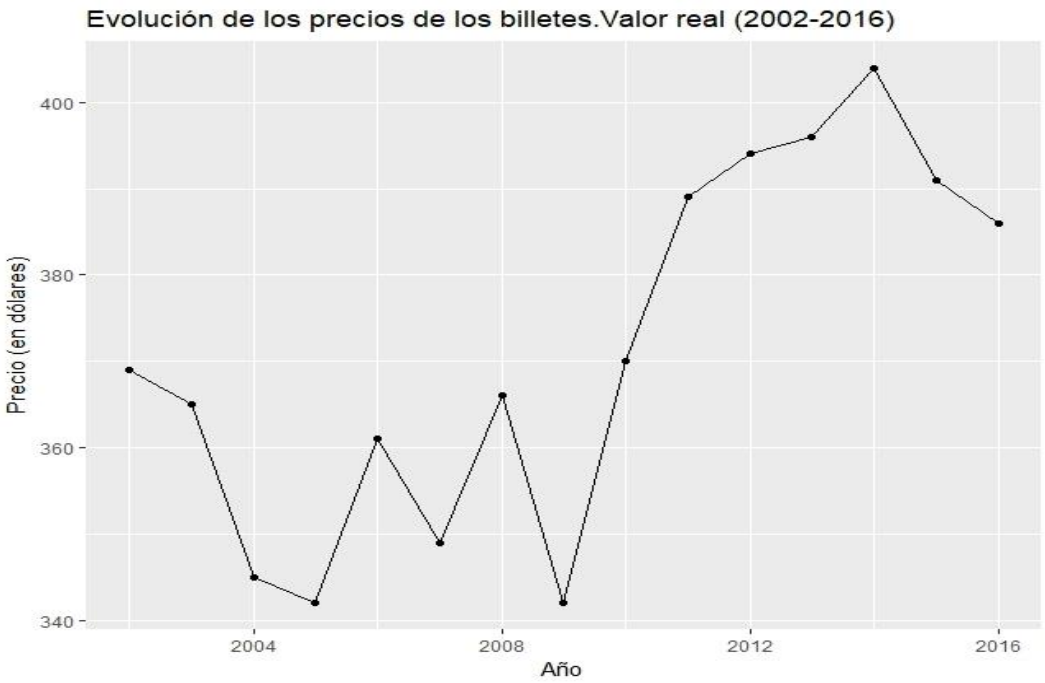


Figura 1.9: Evolución precios billetes

En resumen, se ha podido comprobar que estos últimos quince años el sector aéreo doméstico ha ido concentrándose más y más, en lo que podemos llamar un oligopolio donde las aerolíneas han sabido aprovechar esta situación para ganar más beneficios y ser más productivas (tienen un mayor PLF y menores costes que en 2002). Esto no ha supuesto algo igual de positivo para los consumidores ya que los precios de los billetes no han ido a la baja a diferencia del margen de beneficios de las aerolíneas, por lo que el excedente del consumidor no ha aumentado.

Además antes había más oferta de aerolíneas que elegir para las diferentes rutas y los aviones no iban tan llenos como en la actualidad y eso significa más tiempo de espera y más posibilidad de retrasos. (Ball, Barnhart,2011 y Harlan,2014)

Es ahora que, a partir de ver la estructura oligopolista del mercado, la poca inestabilidad que existe y la evolución de variables clave podemos determinar que **el grado de competencia** para el sector aéreo doméstico en el periodo 2002-2016 es **bastante bajo**.

CAPÍTULO II.
ANÁLISIS DE LOS VUELOS DOMÉSTICOS

Una vez realizado la presentación del mercado aéreo doméstico y visto su estructura y evolución de las aerolíneas en estos últimos años, en el capítulo 2 se procede a analizar los vuelos para el año 2015 que se consideran los productos que ofrecen las aerolíneas a los pasajeros, consumidores del mercado aéreo.

2.1 Presentación de los datos

A continuación, se presenta los ficheros utilizados para el análisis:

2.1.1 Ficheros del análisis

2.1.1.1 Vuelos

El fichero de vuelos contiene más de 5.816.708 registros y 29 variables procedentes del BTS.

Cada fila hace referencia a un determinado vuelo doméstico operado en el año 2015 con información de los tiempos de vuelo, la ruta, el día, mes y hora que inició y acabó el recorrido y, por último, el estado final del vuelo.

Las variables y su significado se presentan a continuación:

Tabla 2.1: Significado de las variables del fichero VUELOS

INFORMACIÓN BÁSICA DEL VUELO	
month	Mes del vuelo
day	Dia del mes del vuelo (1,…,31)
day_of_week	Dia de la semana que tuvo lugar el vuelo(Lunes,…Viernes)
airline	Código IATA de la aerolínea que operó el vuelo
flight_number	ID del vuelo
tail_number	ID del avión
origin_airport	Código IATA del aeropuerto de origen
destination_airport	Código IATA del aeropuerto de destino
INFORMACIÓN DE TIEMPOS DE VUELO	
scheduled_departure	Hora prevista de salida del vuelo
scheduled_arrival	Hora prevista de llegada del vuelo
departure_time	Hora real de salida del vuelo
arrival_time	Hora real de llegada del vuelo
wheels_off	Hora en que el avión despega del origen
scheduled_time	Tiempo previsto del vuelo (en minutos)
air_time	Duración en minutos desde que despega hasta que aterriza
distance	Distancia (en millas) entre los aeropuertos
wheels_on	Hora en que el avión toca tierra en la llegada

departure_delay	Diferencia en minutos entre la hora prevista y la hora real de partida. Si positiva, indica retraso.
arrival_delay	Diferencia en minutos entre la hora prevista y la hora real de llegada. Si positiva, indica retraso.
INFORMACIÓN SOBRE EL ESTADO FINAL DEL VUELO	
diverted	Si el vuelo se desvió o no (1=sí)
cancelled	Si el vuelo fue cancelado o no (1=sí)
cancellation reason	Razón de la cancelación
air_system delay	Retraso en minutos debido al sistema aéreo
security_delay	Retraso en minutos debido por Seguridad
airline_delay	Retraso en minutos debido por la aerolínea
late_aircraft_delay	Retraso en minutos debido a retrasos anteriores
weather_delay	Retraso en minutos debido al tiempo

Se decidió ampliar la información de los vuelos buscando nuevos datos en las diferentes administraciones de los Estados Unidos relacionadas con el mercado aéreo. La razón de esta decisión fue que el fichero de vuelos solo proporcionaba información relacionada con el propio vuelo y para este análisis interesa tener en consideración más factores.

A continuación, se describe los ficheros con información ampliada:

2.1.1.2 **Aviones**

La información relacionada con los aviones se obtuvo del registro de aviones del 2015 de la página web de la FAA a partir de dos ficheros:

-El primero (*actref.txt*) contenía información sobre el año de manufacturación del avión y de tipo legal (a quien estaba registrado el avión, dirección legal, etc).

-El segundo fichero (*master.txt)* proporcionaba información sobre las características técnicas de los aviones.

El segundo fichero era el más importante para el análisis, pero, a diferencia del fichero *actref.txt* no tenía ninguna variable válida para enlazar con el fichero de vuelos. Es por eso que se unió estos dos ficheros a partir de un segundo código compartido (código de serie del avión: *Serial.number*)

El fichero final de aviones se enlaza con el de vuelos a partir de la variable *N-number*. La información a considerar es la siguiente. En negrita se señala la clave primaria:

Tabla 2.2: Significado de las variables del fichero AVIONES

INFORMACIÓN SOBRE AVIONES	
Identificador n-number	Identificador del avión (N_number): N+ 5 dígitos
no.seats	Número de asientos del avión
no.eng	Número de motores del avión
type.aircraft	Tipo de avión
mfr_date	Fecha de manufacturación del avión
model	Modelo del avión
type.eng	Tipo de motor/motores que tiene el avión

2.1.1.3 **Aerolíneas**

Contiene el nombre de 14 aerolíneas que operan en el mercado junto al código asignado por la International Air Transport Association (en adelante, IATA). Cada aerolínea del mundo tiene asignado un único código IATA. Solo sirve para enlazar el código de la aerolínea con su nombre real, por ello no se adjunta ninguna tabla.

2.1.1.4 **Aeropuertos**

Este fichero contiene el código IATA de los 322 aeropuertos más importantes que configuran el mercado aéreo y la siguiente información. Al igual que el fichero de aerolíneas, la fuente procede de la IATA.

Este fichero, además de servir para saber información de los aeropuertos también sirve para poder posicionar en el mapa las distintas rutas del fichero de vuelos gracias a las coordenadas geográficas.

Tabla 2.3 Significado de las variables del fichero AEROPUERTOS

INFORMACIÓN SOBRE AEROPUERTOS	
iata_code	Código IATA del aeropuerto
airport	Nombre completo
city	Ciudad
state	Estado
latitude	Latitud del aeropuerto
longitude	Longitud del aeropuerto

A continuación, se muestra las relaciones entre estas 4 tablas:

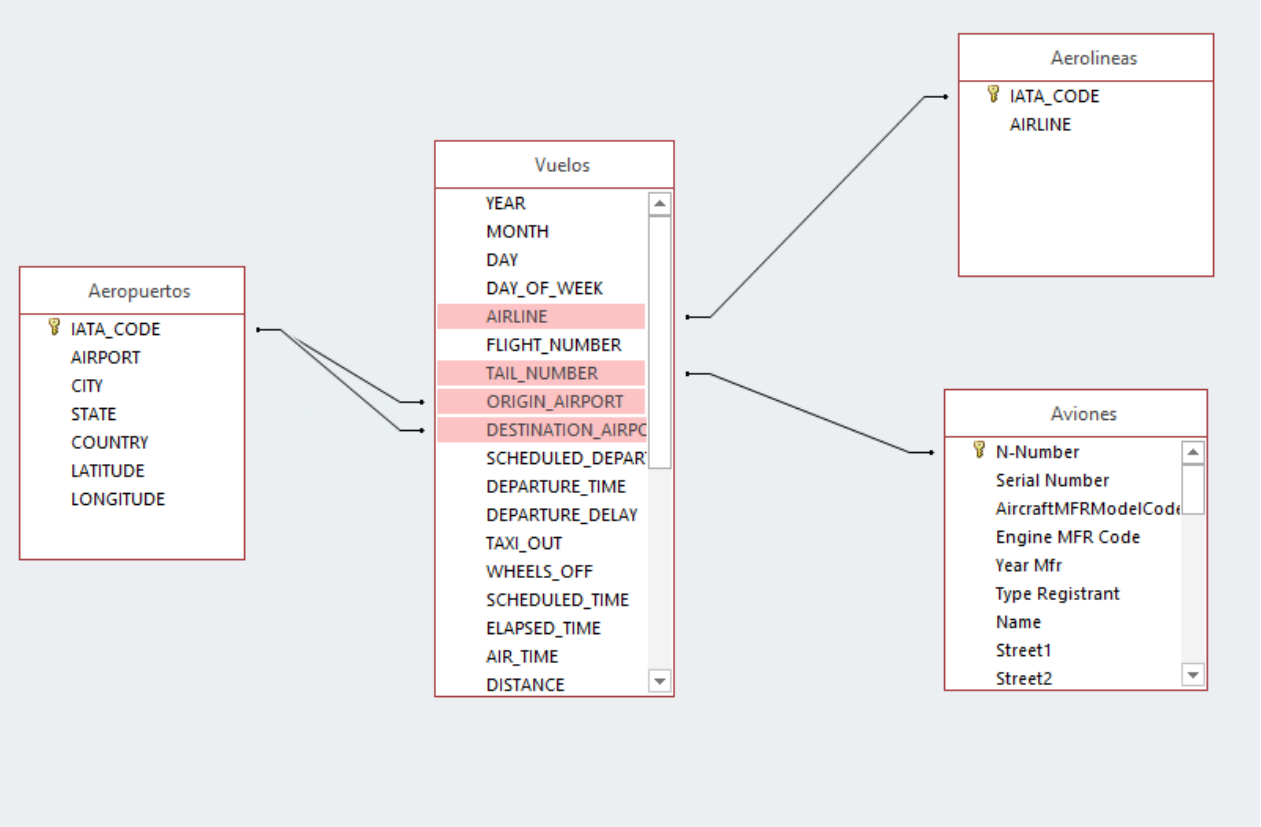


Figura 2.1: Relaciones de la base de datos

Como se observa en la Figura 2.1, el fichero de aerolíneas se relaciona con el de vuelos mediante el código IATA de cada aerolínea. El fichero de aviones se une al de vuelos con el identificador (*N-number*) del avión.

Y, por último, el fichero de aeropuertos se relaciona con el de vuelos a partir del código IATA de cada aeropuerto (y se relaciona dos veces; una para el aeropuerto de origen y otra para el de destino).

2.1.1.5 **Clima**

Para el apartado del clasificador del estado de un vuelo, se creyó conveniente añadir información meteorológica ya que juega un papel importante en el estado de un vuelo.

Se consiguió esta información, mediante una petición en la página web de la NOAA (National Oceanic and Atmospheric Administracion) donde se dispone de información meteorológica histórica para cada una de las estaciones repartidas por todo el país. Se decidió escoger los mismos aeropuertos como estaciones para el análisis.

Este fichero solo se incluyó para las 10 rutas más populares debido la dificultad de conseguir la información (la petición en la web de la NOAA solo incluía una estación por petición y se les mandó un mensaje para ver si podían pasarme directamente las estaciones de los 322 aeropuertos, pero no se obtuvo respuesta) y relacionarla con el fichero de vuelos.

Se logró relacionar este fichero con el de vuelos mediante la creación de una variable que juntara la fecha, hora de salida y el aeropuerto de origen.

Tabla 2.4: Significado de las variables del fichero TIEMPO

INFORMACIÓN METEOROLÓGICA	
station	Nombre de la estación
date	Fecha y hora en que se recoge la información del tiempo
hourly_visibility	Grado de visibilidad en que se identifica un objeto (unidad:millas)
hourly_drybulbtemp	Temperatura estándar en Cº
hourly_humidity	Humedad relativa en porcentaje
hourly_pressure	Presión atmosférica observada (unidad: Hg)
hourly_windspeed	Velocidad del viento (unidad: millas por hora)
hourly_precip	Cantidad total de lluvia en pulgadas en la última hora

2.2 Creación, modificación y eliminación de variables

2.2.1 Creación de variables

Para sacar el máximo de potencial a los datos, se han creado las siguientes 6 variables para el análisis:

- **Estado vuelo**: Indica el estado final de un vuelo que puede ser a tiempo (si *arrival_delay*<15), retrasado (si *arrival_delay*>=15) y cancelado (si *cancelled*=1).

- **Retraso_recode**: Se categorizó la variable *arrival_delay* en 4 categorías (<15 minutos, entre 15 y 30 minutos, entre 31 y 60 minutos y > 1 hora) para ver el grado de retraso.

- **Velocidad**: Dividiendo distancia entre tiempo de vuelo y luego multiplicar por 60 y 1,60934 para pasarla a km/h.

- **Antigüedad**: A partir de la fecha de fabricación del avión (*Year MFR)*. Se calcula de la siguiente manera: Antigüedad= 2015 - *Year MFR.*

- **Es_festivo**: Variable categórica que toma valor 1 cuando es época de festivos y 0 en caso contrario. Para la industria de transporte aéreo, la FAA define 7 épocas oficiales de festivo (*holiday travel seasons*):

Tabla 2.5: Fechas de los festivos oficiales en 2015

ÉPOCA DE FESTIVO	**FECHAS EN 2015**
PRESIDENT'S DAY	12 febrero - 17 febrero
EASTER	29 marzo - 12 abril
MEMORIAL DAY	18 mayo - 27 mayo
INDEPENDENCE DAY	26 junio - 5 julio
LABOR DAY	2 septiembre – 9 septiembre
THANKSGIVING	20 noviembre – 1 diciembre
WINTER HOLIDAY SEASON	16 diciembre - 5 enero

- **Tamaño_avión**: A partir de la variable *no.seats* se creó 3 categorías: Avión pequeño, avión mediano y avión grande. La clasificación se hizo a partir de los saltos que había en la variable número de asientos.

2.2.2 Modificación y eliminación de variables

También se cambió el formato de las variables de hora *wheels off, wheels on, scheduled_arrival_time* y *scheduled_departure_tim*e ya que el formato anterior no permitía diferenciar la hora de los minutos (por ejemplo, 30 quería decir las 00h con 30min, 1237; las 12h con 37 minutos, etc). El nuevo formato utilizado era hora: minutos.

Se decidió eliminar las variables *departure_time* y *arrival_time* (formato hora) ya que pueden llevar a confusión dado que no se indica la fecha: si para un vuelo con una hora prevista de salida a las 00:10, la variable *departure_time* vale 23:49, no se sabe si sufrió un retraso de muchas horas o que la salida se adelantó unos minutos.

La información de cuantificación del retraso ya nos la proporcionan las variables *departure_delay* y *arrival_delay*.

2.3 Problemas en la base de datos

2.3.1 Problema con los códigos de los aeropuertos

Como se ha mencionado anteriormente, para cada vuelo, se especifica el código IATA de los dos aeropuertos (el de origen y el de llegada) de manera que el código sirve de conexión con la tabla de aeropuertos.

El problema surgió en que había códigos de aeropuertos en el fichero de vuelos que no se correspondían con el código IATA (3 letras) sino que seguían un formato numérico de 5 dígitos. Este problema afectaba a 486.165 vuelos (el 8,3% del total de registros) los cuales tenían tanto el código de origen como el código de llegada con el formato incorrecto.

En la página del BTS donde se extrajeron los datos, se encontró 2 ficheros: el primero *T1.txt* contenía el código de 5 dígitos con el nombre del aeropuerto, el segundo *T2.txt*; proporcionaba el código IATA y el nombre del aeropuerto.

Enlazando las dos tablas (mediante left join para conservar todos los códigos erróneos) se concluye que estos códigos eran provenientes de aeropuertos que ya existía información en el fichero AIRPORTS y no eran unos nuevos como se creía.

Finalmente se recodificó las variables *origin_airport* y *destination_airport* uniendo mediante left join la base de vuelos con la de aeropuertos actualizada dos veces (una para el de origen y la segunda para el aeropuerto de destino).

2.3.2 Problema con la agrupación de las aerolíneas

Los vuelos de la base de datos son operados por 14 aerolíneas diferentes, pero algunas son de tipo regional (solo vuelan en regiones con una demanda insuficiente para atraer a aerolíneas nacionales).

Gracias al análisis económico del primer apartado del presente trabajo, se sabe que la mayoría de las aerolíneas regionales son controladas por las aerolíneas nacionales más importantes,

es decir, operan vuelos en distintas regiones de Estados Unidos, pero en nombre de las aerolíneas grandes.

Se decidió asignar los vuelos de estas aerolíneas regionales a las nacionales correspondientes ya que las aerolíneas presentan siempre sus resultados teniendo en cuenta los vuelos de las regionales a las que contratan.

A continuación, se muestra las relaciones de las aerolíneas regionales con las nacionales:

Tabla 2.6: Relación de las aerolíneas regionales con las nacionales

AEROLINEA REGIONAL	AEROLÍNEA(S) NACIONAL
AMERICAN EAGLE	American
SKYWEST	American, United, Delta y Alaska
EXPRESSJET	American, United y Delta

En el caso de American Eagle, sus vuelos fueron fáciles de asignar ya que solo operan para una sola aerolínea nacional (American).

Para los casos de Skywest y ExpressJet, que operan para más de una compañía, se realizó la asignación a partir de dos fuentes de información: los números de vuelos (variable *flight_number*) y el aeropuerto de origen y de destino (muchas de las rutas aéreas que vuelan las aerolíneas regionales son exclusivas de cada aerolínea nacional). Esta información se obtuvo en las webs de las compañías regionales.

Por último, se decidió asignar los vuelos de US Airways a American Airlines dado que a finales de 2014 esta última adquirió a la primera. Con esta agrupación, se pasó de 13 aerolíneas a un final de 9.

2.4 Tratamiento de valores missings y anómalos

El último paso del preprocessing fue examinar los valores faltantes y valores anómalos en las bases de datos para ver si era necesario o no recurrir a un método de imputación.

2.4.1 Valores anómalos

En el summary (en anexo Salida R.1) se observó que la variable *distance* tenía un mínimo de 1. Esto significa que el vuelo tuvo una distancia de solo una milla. No tenía sentido dado el origen y destino así que se substituyó este valor por el valor correcto de la ruta. También se eliminó 16 registros con antigüedad igual a 2015 años (en el fichero de aviones, el año de fabricación tenía valor 0).

Se comprobó también que cada ruta tuviera la misma distancia, que la suma de los tiempos de vuelo coincidiera con la hora de llegada y con el retraso en minutos, para así evitar cualquier sesgo en el análisis.

2.4.2 Valores missings

De los tres ficheros, solo el fichero de vuelos tenía valores faltantes, algunos codificados como *NA* y otros no tenían ningún dato, pero R no los asignaba como *NA*. Se recodificó entonces los datos sin valores por *NA's*.

La distribución del número de *NA's* para el total de 5.816.708 vuelos es la siguiente:

Tabla 2.7: Distribución de los NA's

Nº DE NA'S	**1**	**2**	**6**	**7**	**11**	**12**	**>12**
FRECUENCIA FILAS	1.062.750	330	4.646.718	1.856	12.549	833	91.672
PORCENTAJE	18.3 %	0,005%	80%	0,03%	0,2%	0,01%	1.6%

La conclusión es que no hay ningún vuelo que tenga todas las variables con algún valor. Esto tiene una explicación lógica.

La gran cantidad de missings corresponde a 3 situaciones:

- Vuelos no cancelados que tienen en la variable *cancelation_reason* un NA como es lógico.
- Vuelos que no han sufrido retraso, y por consiguiente tienen NA en las variables de causa de retraso. (*air_system_delay, security_delay, airline_delay, late_ aircraft_delay* y *weather_delay*)
- Vuelos cancelados que tienen NA's en las variables de tiempo de vuelo.

Se decidió recodificar estos valores missing por un 0 y no aplicar un método de imputación ya que obviamente se trata de valores faltantes estructurales. Después de esta recodificación, solo un 1,8% de los vuelos presentaban algún valor restante que no se correspondía a una de estas tres situaciones.

Estos vuelos tenían valores missing en las variables de tiempo del vuelo. Había algunos vuelos que les faltaba el tiempo de vuelo (*air_time*) que se podía construir a partir de las variables *wheels off* y *wheels_on* ya que el tiempo de vuelo, según el BTS, es la diferencia entre estas dos horas.

Pero no se decidió obtener los valores faltantes de *air_time* a partir de estas dos variables, ya que la hora de cuando aterriza (*wheels on*) no tiene porqué compartir el mismo huso horario que la hora de cuando despega el avión y, por lo tanto, el tiempo de vuelo no es exactamente la diferencia.

En resumen, en esta primera etapa previa al análisis se han preparado definitivamente los ficheros mediante la creación de nuevas variables, la eliminación de variables confusas o poco informativas y la eliminación de registros con información faltante o incorrecta para así empezar el análisis sin sesgo alguno.

2.5 Análisis exploratorio de los datos

El objetivo de esta fase del análisis es a partir del análisis descriptivo-gráfico y tests de inferencia entender mejor los datos e identificar relaciones entre las variables. Por último, se realiza un análisis multivariante con el Análisis de Correspondencia Múltiples para analizar las posibles relaciones de forma simultánea.

Este capítulo no trata solo de una descriptiva univariante o bivariante sino de generar preguntar sobre los datos y mediante la visualización, transformar estos datos en respuesta a las preguntas.

En el mundo actual del big data, los gráficos cada vez más son una herramienta muy potente para exponer conclusiones claras sobre datos. Rstudio dispone de un paquete llamado *ggplot2* con el que se logran gráficos atractivos y de diferentes tipos. Los gráficos de este análisis están hechos a partir del mencionado paquete.

2.5.1 Vuelos

2.5.1.1 **¿Como se distribuye el estado de los vuelos?**

En la Tabla 2.8: Estado vuelo se constata que en 2015 la inmensa mayoría de vuelos, el 80,8%, llegó a tiempo a su destino, el 17,7% sufrió algún retraso y sólo el 1,5% no pudo llegar a destino debido a que fue cancelado.

De los vuelos con retraso en la llegada, las causas más comunes son las de aerolínea, sistema aéreo y acumulación de retrasos anteriores (cada una representa entorno el 30%), después debidas al tiempo extremo (solo un 3,7%). Los retrasos debido a seguridad son muy escasos ya que representan un mínimo 0,2% del total de retrasos.

Tabla 2.8: Estado vuelo

Estado	Frecuencia	%
No retrasado	4.592.481	80,8 %
Retrasado	1.004.540	17,7 %
Cancelado	87.860	1,5 %
TOTAL	**5.684.882**	**100%**

Tabla 2.9: Motivos de retraso

Motivo	Frecuencia	%
Aerolínea	561.157	32,5%
Sistema	554.208	32,1%
Avión retrasado	545.626	31,6%
Tiempo	64371	3,7%
Seguridad	3424	0,2%

Tabla 2.10: Motivos de cancelación

Motivo	Frecuencia	%
Tiempo	48.408	55,1 %
Aerolínea	23.824	27,1 %
Sistema	15.606	17,8 %
Seguridad	22	0,025%

La mitad de los vuelos cancelados son debido a condiciones meteorológicas extremas (el 55%), a diferencia de los vuelos con retraso donde esta causa solo representaba un escaso 3,7% (es lógico ya que cuando hay tormentas fuertes o tornados, suelen cancelarse vuelos por precaución). Al igual que en los retrasos, las cancelaciones por seguridad son muy escasas. (solo un 0,02%)

2.5.1.2 **¿Qué causas provocan más retraso?**

Lo verdaderamente importante no es ver la frecuencia de cada causa de retraso en los vuelos, sino que interesa el impacto que tienen sobre el retraso final. Por ello se ha representado el boxplot del retraso para cada causa.

Los retrasos por avión anterior retrasado, que se encontraban en la tercera posición en cuanto a frecuencia, son la clase de retrasos con más impacto: una media de 48 minutos y un retraso mediano de 26 (muy alejado del resto de causas).

Los retrasos por mal tiempo que representan solo un 4% del total de retrasos, son los que ocupan la segunda posición en cuanto a impacto con un retraso medio de 14 minutos y mediana igual a 6.

Para todas las causas, se observa que la distribución del retraso es asimétrica a la derecha, donde la mayoría de los retrasos están acumulados debajo del retraso mediano.

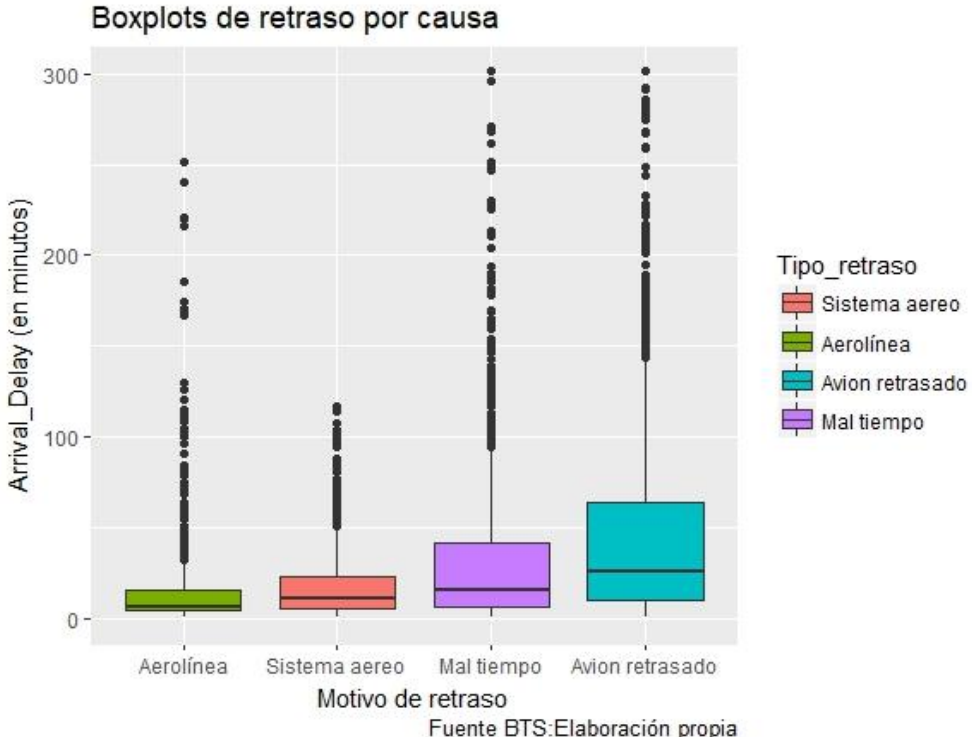


Figura 2.2: Retraso por causa

A continuación, se presenta la distribución de numero de causas de retraso por cada vuelo:

Tabla 2.11: Número de causas de retraso

Nº DE CAUSAS	**1**	**2**	**3**	**4**
Nº VUELOS	481417	561402	654	623
%	46,1%	53,8%	0,1%	0,1%

La mayoría de los vuelos retrasados tienen una o dos causas de retraso asociadas siendo dos el número de causas más frecuente (53,8% de los vuelos presentan dos causas).

La pareja de causas de retraso más frecuente es la de Aerolínea- Avión retrasado (más del 40% de vuelos retrasados tienen como causas estas dos a la vez).

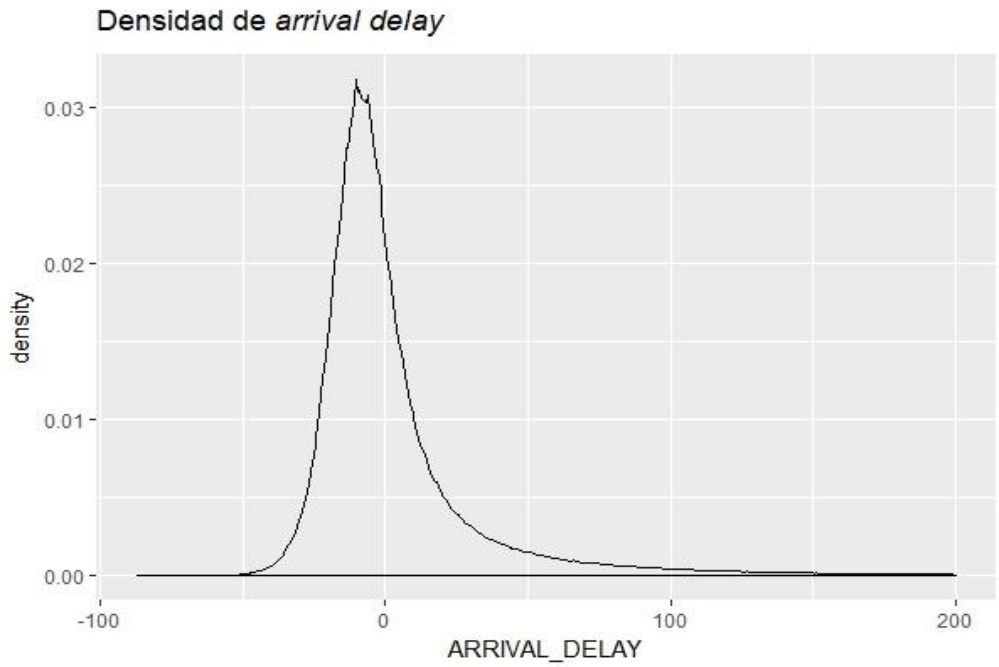


Figura 2.3: Densidad de la variable de retraso

De la figura 2.3 se puede deducir que la distribución de la variable *arrival_delay* no es normal. Si se aplica el test de Shapiro Wilk en una muestra de 5000 vuelos (es lo máximo que permite R en este test) para comprobar esta hipótesis se obtiene un p-valor inferior a 0.05 rechazando la hipótesis nula de normalidad. La distribución es no normal, asimétrica a la derecha (skewness=6,5) y muy leptocúrtica (kurtosis=98,4). Esto quiere decir que la mayoría de los vuelos presentan retrasos pequeños y cada vez menos vuelos presentan retrasos mayores.

Salida R 2.1: Resultado del test de Shapiro

```
> shapiro.test(mostra$ARRIVAL_DELAY)

        Shapiro-Wilk normality test

data:  mostra$ARRIVAL_DELAY
W = 0.5484, p-value < 2.2e-16
```

2.5.1.3 **¿Qué fechas son en general las mejores para viajar?**

Para todos los meses existen casos outliers y la distribución del retraso sigue siendo asimétrica a la derecha (la mayoría de los vuelos no presentan ningún retraso). Por ello, en lugar de la mediana (valen todas 0) se comparan los percentiles 75 (informa de cuanto de malo puede llegar a ser el retraso para cada mes).

Los meses de febrero, junio, julio y diciembre son los peores meses para viajar con retrasos medios y percentil75 mayor a los 10 minutos. En cambio, en los meses de setiembre, octubre, noviembre hay menos grado de retraso (sus percentiles75 no superan los 5 minutos).

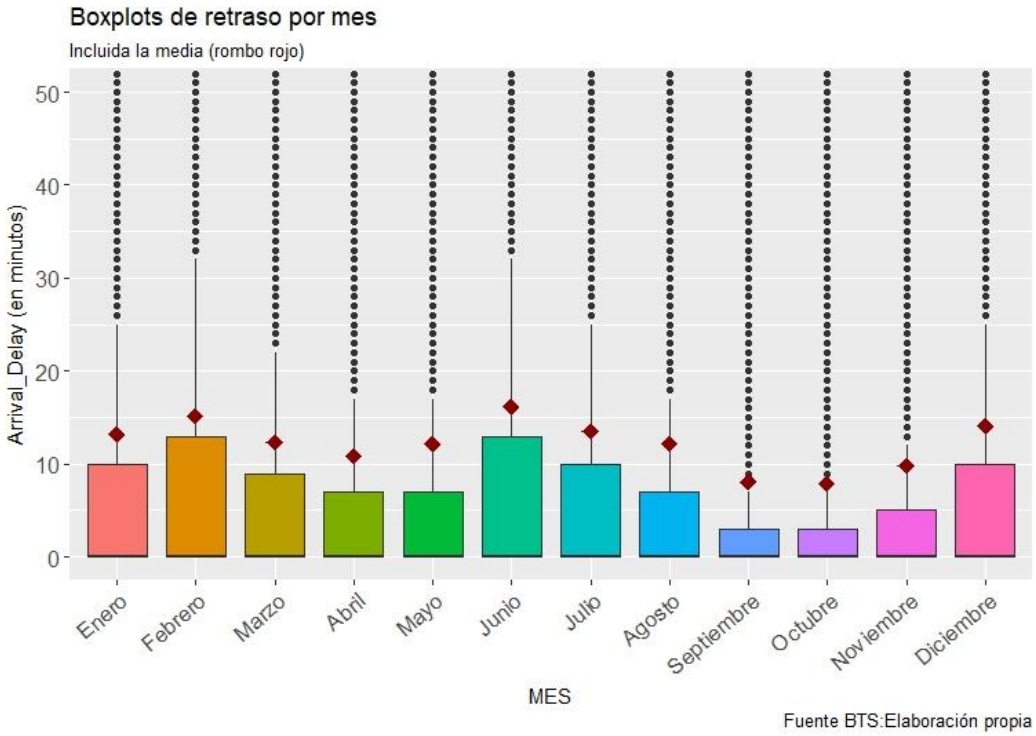


Figura 2.4: Retraso por mes del año

Las diferencias de retraso entre los doce meses del año son estadísticamente significativas (mediante el test de Kruskal Wallis con un p-valor<0.05). En el anexo se adjunta la tabla de comparaciones múltiples mediante la instrucción pairwise.wilcox.test (en Anexo Salida R.2).

A excepción de los pares febrero -junio, agosto-enero, agosto-marzo y abril-mayo, entre todos los demás pares de meses existen diferencias significativas en el retraso.

Salida R 2.2: Resultado del test de Kruskal Wallis

```
> kruskal.test(flights.2$ARRIVAL_DELAY,flights.2$MONTH)

        Kruskal-Wallis rank sum test

data:  flights.2$ARRIVAL_DELAY and flights.2$MONTH
Kruskal-Wallis chi-squared = 1072.5, df = 11, p-value < 2.2e-16
```

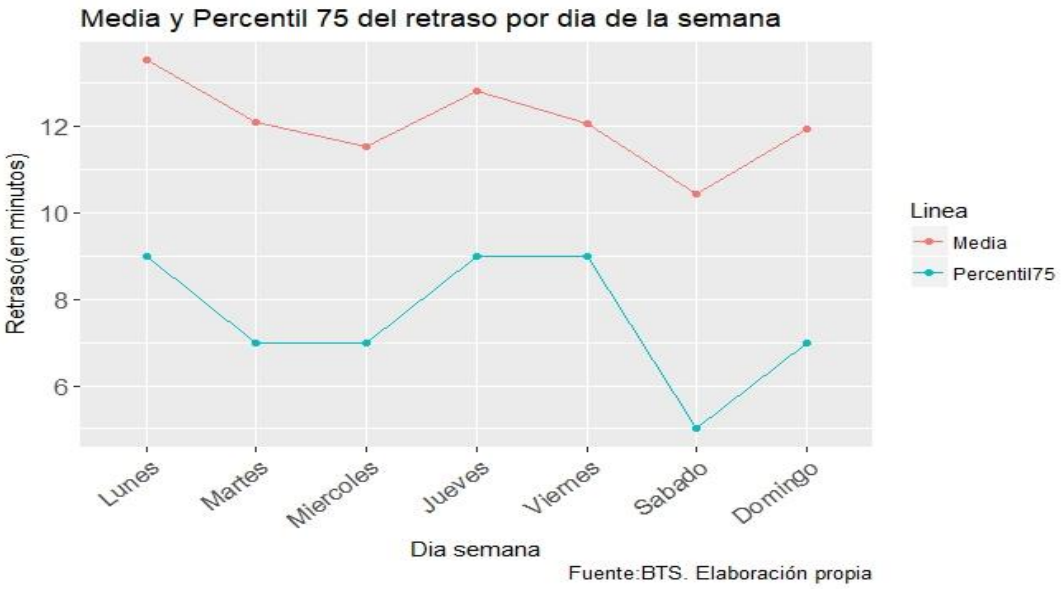


Figura 2.5: Retraso por día de la semana

En días de semana, los miércoles, sábados y domingos son los mejores días para coger un vuelo a diferencia de los lunes, jueves y viernes. El sábado presenta un retraso medio de 10 minutos y un percentil 75 de 5 minutos, valor que vale el doble si se viaja un lunes.

Cabe destacar que los sábados también es el día que vuelan menos aviones (872 vuelos diarios es la media ese día), mientras que los lunes, es el día con más demanda, presenta una media diaria de 2300 vuelos. En el anexo se adjunta la tabla de frecuencias por día de la semana (en Anexo Tabla 1).

Las diferencias de retraso también son significativas entre días de la semana (test de Kruskal Wallis con p-valor <0.05). Si se observa en detalle, todas las comparaciones 2 a 2 salen como significativas (en Anexo Salida R.3)

Salida R 2.3: Resultado del test de Kruskal Wallis

```
> kruskal.test(flights.2$ARRIVAL_DELAY,flights.2$DAY_OF_WEEK)

        Kruskal-Wallis rank sum test

data:  flights.2$ARRIVAL_DELAY and flights.2$DAY_OF_WEEK
Kruskal-Wallis chi-squared = 207.46, df = 6, p-value < 2.2e-16
```

Para ver interacciones entre el mes y el día de la semana, se ha construido el gráfico de retrasos para cada día de la semana de cada mes del año.

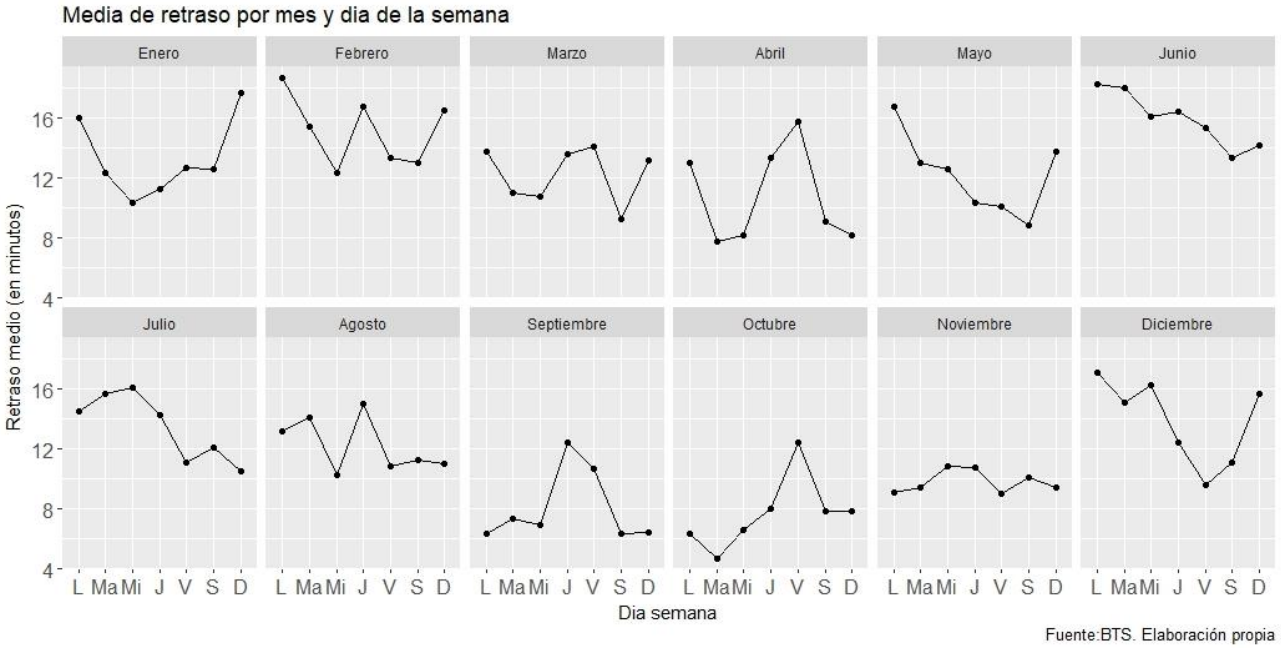


Figura 2.6: Media de retraso por mes y día de la semana

Se observa claramente que la tendencia del retraso durante la semana es distinta según el mes en que se vuela, es decir, el retraso en cada día de la semana depende del mes (existe interacción entre estos dos factores). Por ejemplo, en setiembre y octubre el mejor día de la semana para viajar son los lunes, pero en junio y diciembre pasa todo lo contrario.

También se observa que la variabilidad entre días de la semana es diferente: por ejemplo, en noviembre vemos que apenas existe diferencias de medias de retraso, a diferencia de diciembre en que viajar según qué día presenta niveles bastante diferentes de retraso.

2.5.1.4 **¿Qué hora es la mejor para coger un vuelo?**

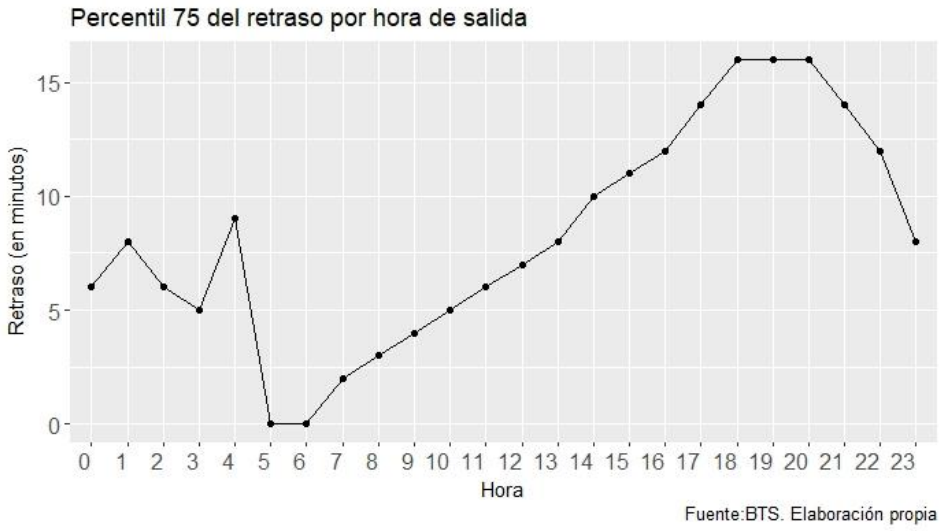


Figura 2.7: Percentil 75 del retraso por hora de salida

Se observa una tendencia creciente en el retraso en los vuelos que la hora de salida es entre la medianoche y las tres de la mañana y entre las 6 de la mañana y las 6 de la tarde. La mejor hora para coger un vuelo es a las 5 o 6 de la mañana con un retraso prácticamente nulo (el

percentil 75 es de 0 minutos, quiere decir que un 75% de las veces el vuelo no sufrirá ningún retraso).

La peor hora es entre las 6 de la tarde y 9 de la noche. Es normal ya que los retrasos se han acumulado durante el día. Para este intervalo de horas, el percentil75 del retraso es superior a 15 minutos.

Esta tendencia del retraso para las diferentes horas es igual para todos los meses, es decir, no hay interacción entre estas dos variables y por lo tanto no se adjunta el gráfico como en el caso del mes y el día de la semana.

2.5.1.5 **Evolución temporal del estado de los vuelos**

En el siguiente gráfico temporal se presentan los porcentajes de vuelos cancelados y retrasados para cada día del año 2015. Se observa como la evolución de vuelos cancelados tiene una evolución similar a la de vuelos retrasados, solo que, desplazada más hacia abajo. Esta tendencia se hace más visible en las épocas de más retraso.

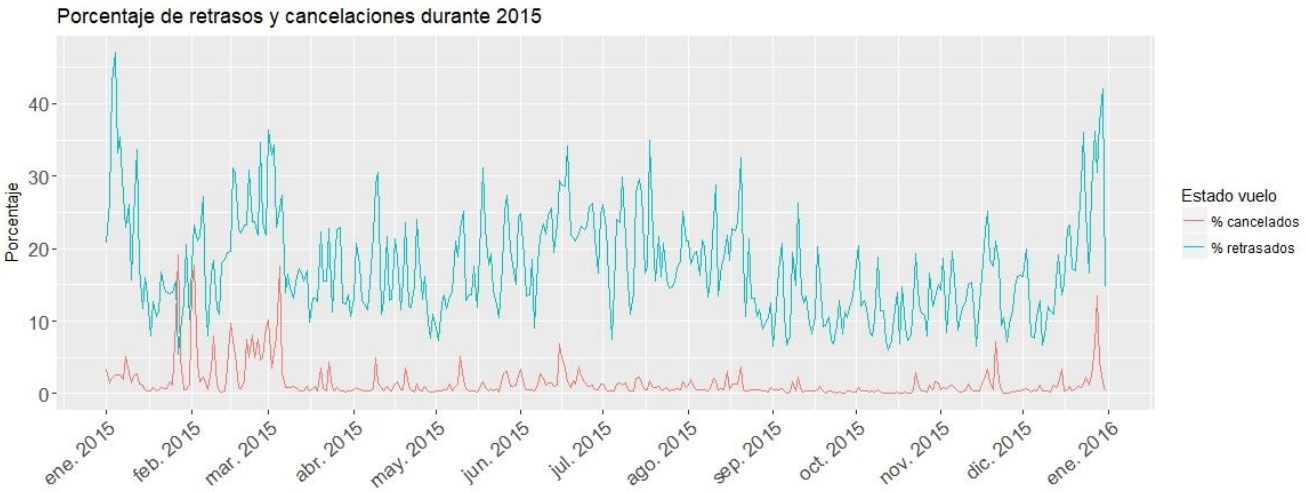


Figura 2.8: Evolución temporal: % retrasos y cancelaciones

Para observar mejor la evolución de los vuelos cancelados, se adjunta a continuación solo la evolución de las cancelaciones.

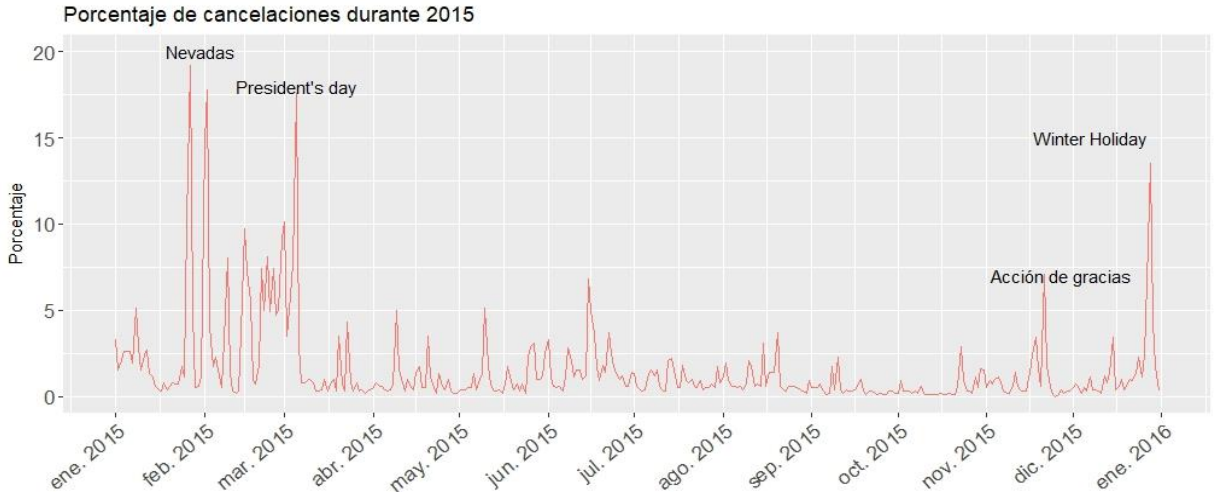


Figura 2.9: Evolución temporal: % cancelaciones

Es notable que, desde inicios de diciembre, el porcentaje de vuelos retrasados, que desde setiembre se mantiene entre el 10 y 20%, va aumentando y llega hasta su pico máximo en las épocas festivas navideñas (supera el 40%). Lo mismo pasa para los vuelos cancelados, ya que durante todo el año se mantienen debajo del 5%, pero para estas fechas logra rozar el 15%.

Los meses de septiembre a noviembre son los más estables (octubre es el mes con un retraso medio menor como se observa en la figura 2.8).

El único momento del año en que el porcentaje de cancelados supera el de retrasados es a finales de enero y principios de febrero, ya que uno de cada cinco vuelos fue cancelado. Este hecho es algo puntual que se explica a partir del mal temporal que hizo para estas fechas debido a una gran nevada que azotó el norte del país y obligó a las aerolíneas a cancelar masivamente sus vuelos.

Además, se observa que en los días festivos como el President's Day y Acción de Gracias, aumenta el porcentaje de vuelos cancelados y retrasados debido al aumento de demanda por coger vuelos. El test de la Chi cuadrado concluye que existe relación significativa entre el porcentaje de retrasos y las épocas de festivo.

El intervalo de confianza resultante del test de Wilcoxon es [0.9, 2] (en Anexo Salida R.4 se encuentra el intervalo de este test). El retraso mediano en épocas de festivo está entre 1 y 2 minutos por encima del de época no festiva.

Salida R 2.4: Resultado del test de la Chi cuadrado

```
> chisq.test(table(flights.2$festivo,flights.2$DELAYED))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(flights.2$festivo, flights.2$DELAYED)
X-squared = 168.25, df = 1, p-value < 2.2e-16
```

2.5.2 Aerolíneas

2.5.2.1 ¿Qué aerolíneas dominan el mercado según el número de vuelos?

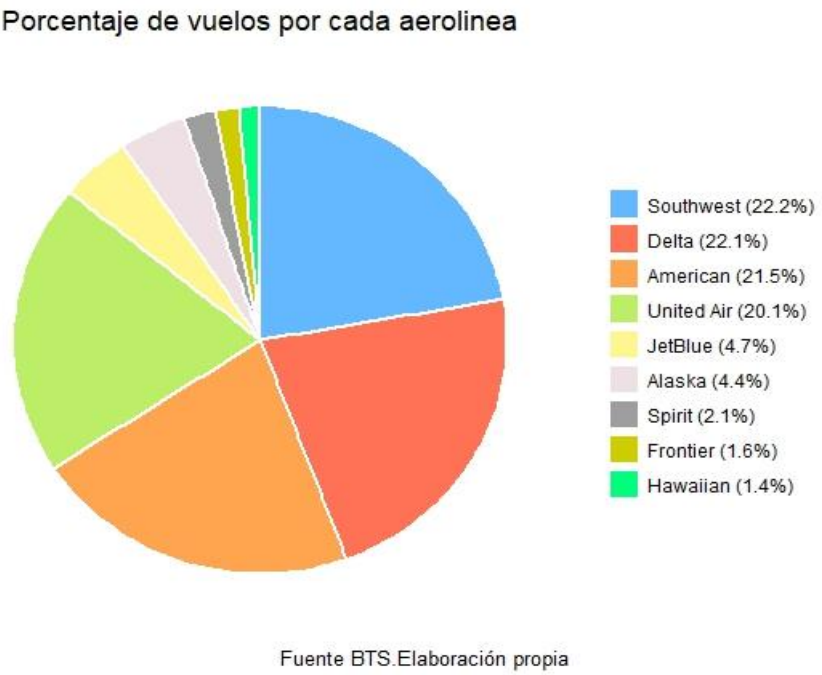


Figura 2.10: Distribución de los vuelos por aerolínea

El gráfico muestra claramente que más del 80% del total de vuelos domésticos de nuestra base fueron operados por solo 4 aerolíneas: la lowcost Southwest y las aerolíneas Delta, American y United Airlines. Estas 4 aerolíneas presentan unos porcentajes bastante similares, rondando el 20% que en términos absolutos representa más de un millón de vuelos por aerolínea.

Esta distribución tan desigual se sigue viendo en el número de rutas aéreas que cada aerolínea realiza y en el tamaño de la flota que cada una posee (Tabla 2.12): está claro que a más rutas y más aviones para volar, más vuelos hace la compañía. Por ejemplo, Southwest realiza más de 1000 rutas diferentes mientras que la regional Hawaiian solo 56.

Tabla 2.12: Rutas y flota para cada aerolínea

Aerolínea	Número de rutas	Número aviones
United	1460	1339
Southwest	1329	705
Delta	1309	1093
American	1109	1323
Alaska	344	230
Jetblue	341	215
Frontier	331	79
Spirit	314	63
Hawaiian	56	51

2.5.2.2 **¿Qué aerolíneas presentan más retrasos y cancelaciones?**

En el siguiente gráfico se representa el porcentaje oficial de vuelos cancelados, retrasados y a tiempo para cada aerolínea.

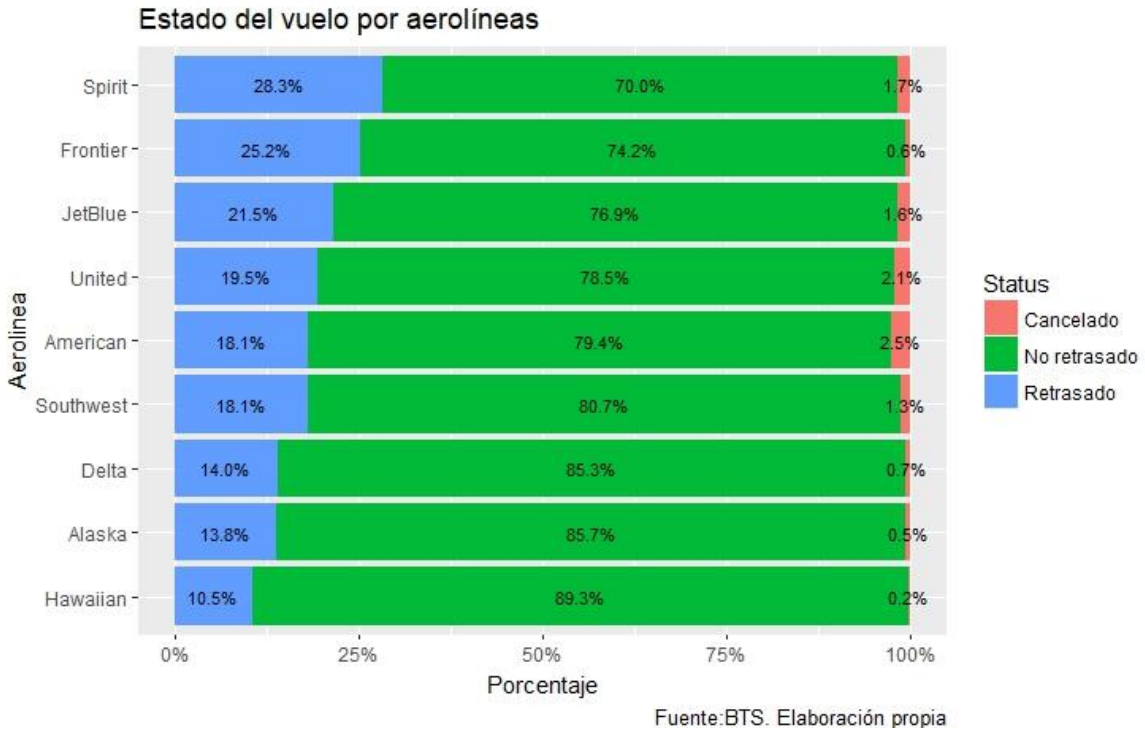


Figura 2.11: Estados del vuelo por aerolínea

Spirit Airlines es la aerolínea que presenta una tasa de retrasos mayor (más del 28% de sus vuelos sufre retrasos) seguida de cerca de Frontier Air y Jetblue. La pequeña Hawaian, que era la que menos vuelos opera de todas las aerolíneas, es la que presenta una tasa de retraso menor (sólo un 10% de sus vuelos llega con algún retraso).

El porcentaje de vuelos cancelados es mínimo para todas las compañías, pero destaca American Airlines con un 2,5% de tasa de cancelación (el porcentaje medio de vuelos cancelados para las demás aerolíneas se sitúa en el 1%).

Para dar conclusiones más detalladas sobre el nivel de retraso en los vuelos de cada compañía es necesario trabajar con la variable *arrival_delay* (retraso en minutos en destino) ya que está claro que es más interesante saber el grado de retraso que no el porcentaje de vuelos retrasados para cada aerolínea (aunque dos aerolíneas presenten porcentajes similares, preferiremos la que tenga menor grado de retraso en sus vuelos).

A continuación, se muestra la variable *arrival-delay* categorizada para cada aerolínea:

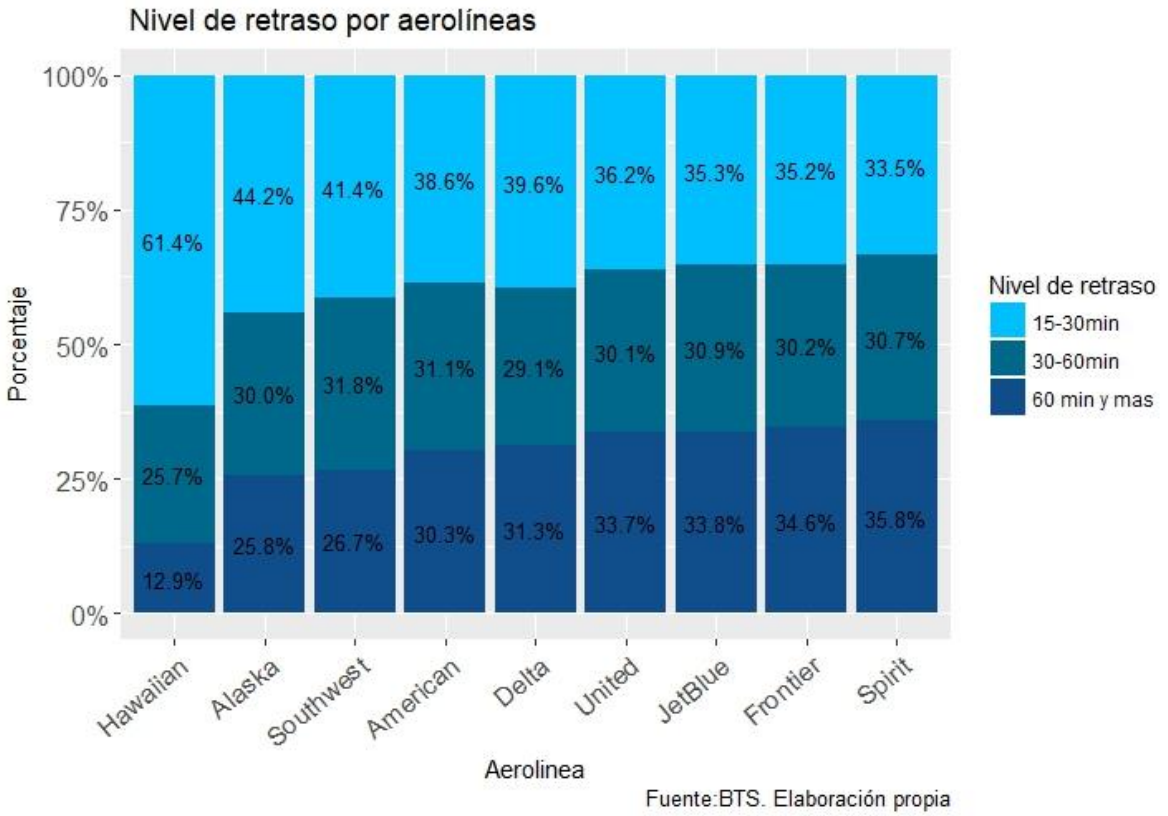


Figura 2.12: Nivel de retraso por aerolíneas

La aerolínea Hawaiian, que era la que presentaba un porcentaje inferior de retrasos, también es la compañía con un menor grado (es la única compañía en la que si un vuelo suyo sufre retraso más del 50% de estos será menor a media hora y solo el 13% de sus retrasos son mayores a una hora).

En el otro extremo, se encuentra otra vez la aerolínea Spirit donde el 36% de los retrasos son superiores a una hora (triplica a Hawaian en este porcentaje). Por lo tanto, si se puede evitar viajar con esta aerolínea, mejor.

Si no se tiene en cuenta Hawaian, las otras aerolíneas presentan porcentajes similares en los tres grados de retraso: los retrasos menores rondan el 40% y los retrasos medios; el 30%. En cuanto a los retrasos de más de una hora, sí que se observa mayor variación entre aerolíneas ya que algunas presentan porcentajes alrededor del 25% (casos de Alaska, Southwest) y otras que rondan el 35% (United Air, Spirit y Jetblue).

Para ver si existen diferencias estadísticamente significativas de los retrasos entre las 9 aerolíneas se utiliza el test de Kruskal Wallis. Se concluye que sí hay diferencias significativas en el retraso de las aerolíneas (p-valor<0.05).

Salida R 2.5: Resultado del test de Kruskal Wallis

```
> kruskal.test(ARRIVAL_DELAY~AIRLINE,data=flights.2)

        Kruskal-Wallis rank sum test

data:  ARRIVAL_DELAY by AIRLINE
Kruskal-Wallis chi-squared = 734.21, df = 8, p-value < 2.2e-16
```

Si se observa en detalle las comparaciones múltiples 2 a 2 (en anexo Salida R.5) existe diferencias significativas entre todos los pares de aerolíneas a excepción de Frontier con Spirit (las dos aerolíneas peor situadas en el ranking)

En el siguiente gráfico se adjunta los gráficos de densidad de la variable *arrival_delay* para las aerolíneas Hawaian y Spirit (que representan los dos extremos de nivel de retraso)

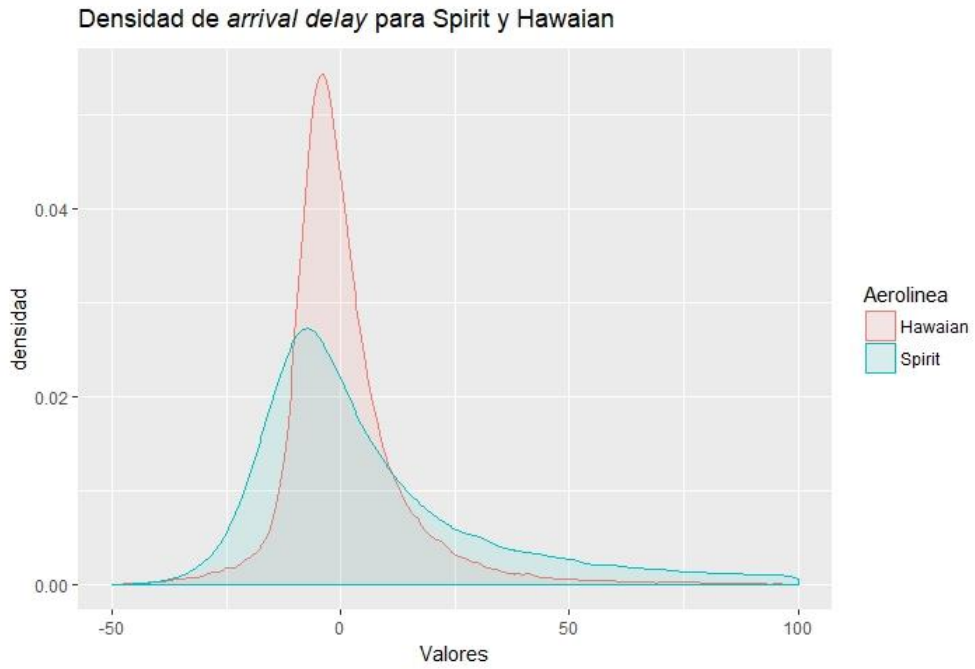


Figura 2.13: Densidad del retraso de Spirit y Hawaian

Se observa claramente que las dos distribuciones presentan formas bastante diferentes: por una parte, la densidad de Hawaian es bastante simétrica, mientras que la de Spirit, presenta una asimetría positiva, reflejando así el mayor porcentaje de retrasos.

También se concluye que la densidad del retraso en el caso de Hawaian presenta menos variación que la de Spirit (la densidad de Hawaiian es la más apuntada de todas las aerolíneas) y por lo tanto se concluye que un pasajero que vuele con Hawaian tiene menos riesgo de sufrir retrasos que si viajara con Spirit.

El test de Mann Whitney Wilcoxon para muestras independientes concluye que hay diferencias estadísticamente significativas entre los retrasos de estas dos aerolíneas. El intervalo de confianza confirma esto: el retraso mediano de Hawaiian está por debajo del de Spirit 2 minutos menos como mínimo. (si se calculase el intervalo con las medias, las diferencias son mayores que el de medianas debido a la asimetría de la distribución)

Salida R 2.6: Resultado del test de Wilcoxon

```
> wilcox.test(ARRIVAL_DELAY ~ AIRLINE, data = spirit_hawaian,conf.int = T)

        Wilcoxon rank sum test with continuity correction

data:  ARRIVAL_DELAY by AIRLINE
W = 4076100000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -2.000010 -1.999941
sample estimates:
difference in location
             -1.999982
```

2.5.2.3 **¿Qué aerolíneas tuvieron más cancelaciones por culpa suya?**

Lo lógico sería que también en los vuelos cancelados se mantenga la proporción de vuelos vista anteriormente (es decir, si una aerolínea opera el 40% de los vuelos, se espera que más o menos presente el mismo % de cancelado respecto el total global).

Tabla 2.13: Vuelos cancelados por aerolínea

Aerolinea	Frecuencia	Porcentaje
American	30899	35,2%
United Air	23789	27,1%
Southwest	16043	18,3%
Delta	8788	10%
Jetblue	4276	4,8%
Spirit	2004	2,3%
Alaska	1302	1,5%
Frontier	588	0,6%
Hawaiian	171	0,2%

Las aerolíneas que superan el porcentaje esperado de vuelos cancelados son American (opera el 21% de vuelos, pero suyos son el 35% del total de cancelados) y United Air. En cambio, las aerolíneas Delta, Southwest, Spirit, Frontier y Alaska presentan porcentajes inferiores a lo esperado.

Está claro que es más grave un vuelo cancelado que un vuelo con retraso y por ello se profundiza en ver que cancelaciones son culpa de la aerolínea.

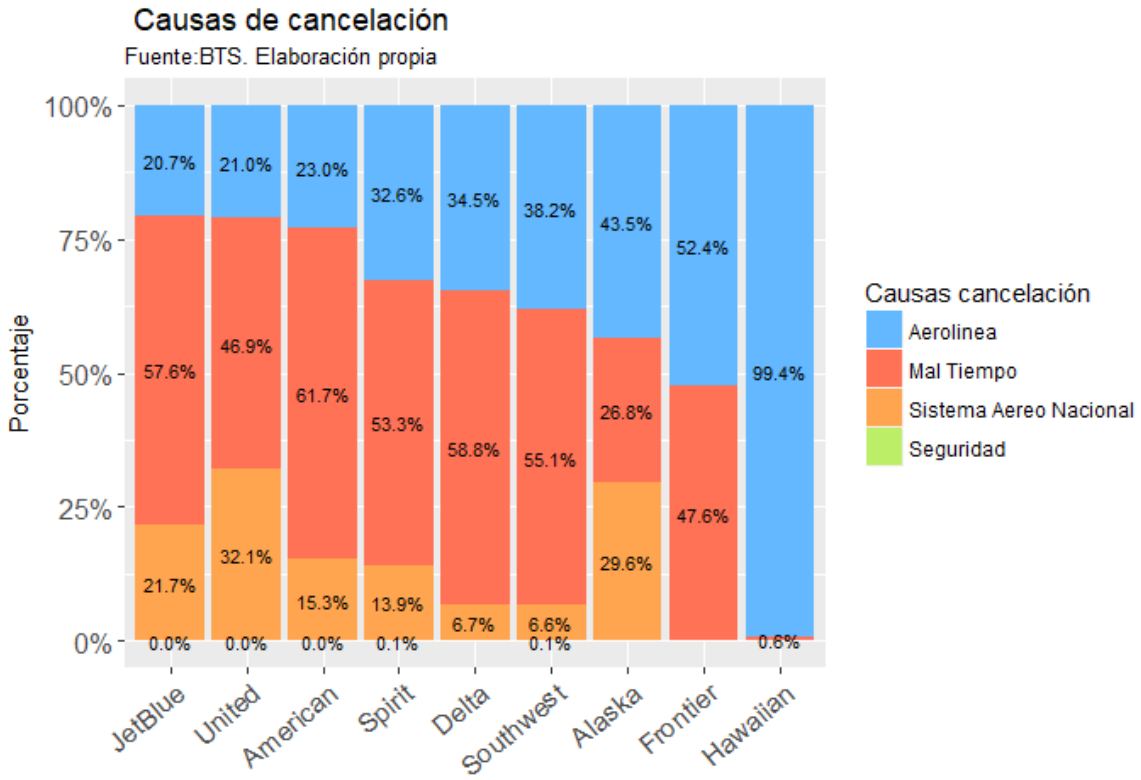


Figura 2.14: Causas cancelación por aerolínea

Sorprende ver como el 99,4% de las cancelaciones de Hawaian son a causa suya y no por otros factores. American y United Air que eran las dos aerolíneas con el porcentaje de cancelados más alto, son ahora las que menos culpa tienen de sus vuelos cancelados.

2.5.2.4 **Efecto relleno**

Se descubrió un comportamiento curioso en las aerolíneas respecto la duración programada de cada ruta: los tiempos previstos de vuelos para una misma ruta varían según cada aerolínea: Por ejemplo, la ruta de Tucson a Los Angeles tiene una duración programada de 92 minutos para Southwest pero en cambio, para Delta el tiempo programado es de 106 minutos.

Se buscó en internet por qué los tiempos programados para una misma ruta varían. Este hecho se le denomina "padding effect" según el BTS (en castellano; efecto relleno): algunas aerolíneas modifican sus tiempos al alza para así reducir la probabilidad de que el vuelo se considere oficialmente en retraso.

Esto es debido a dos factores: el retraso implica costes económicos para todos, como se verá en el capítulo siguiente y, por otra parte, los resultados de las tasas de retraso son de carácter público (el BTS publica estos datos abiertamente) y se utiliza como medida de calidad del servicio de las aerolíneas.

Se calculó para cada aerolínea la media de su efecto relleno (excluyendo las rutas que son exclusivas de cada aerolínea para no llegar a conclusiones erróneas) a partir de la media de las diferencias entre la duración programada media para cada ruta y la duración programada de la ruta según cada aerolínea.

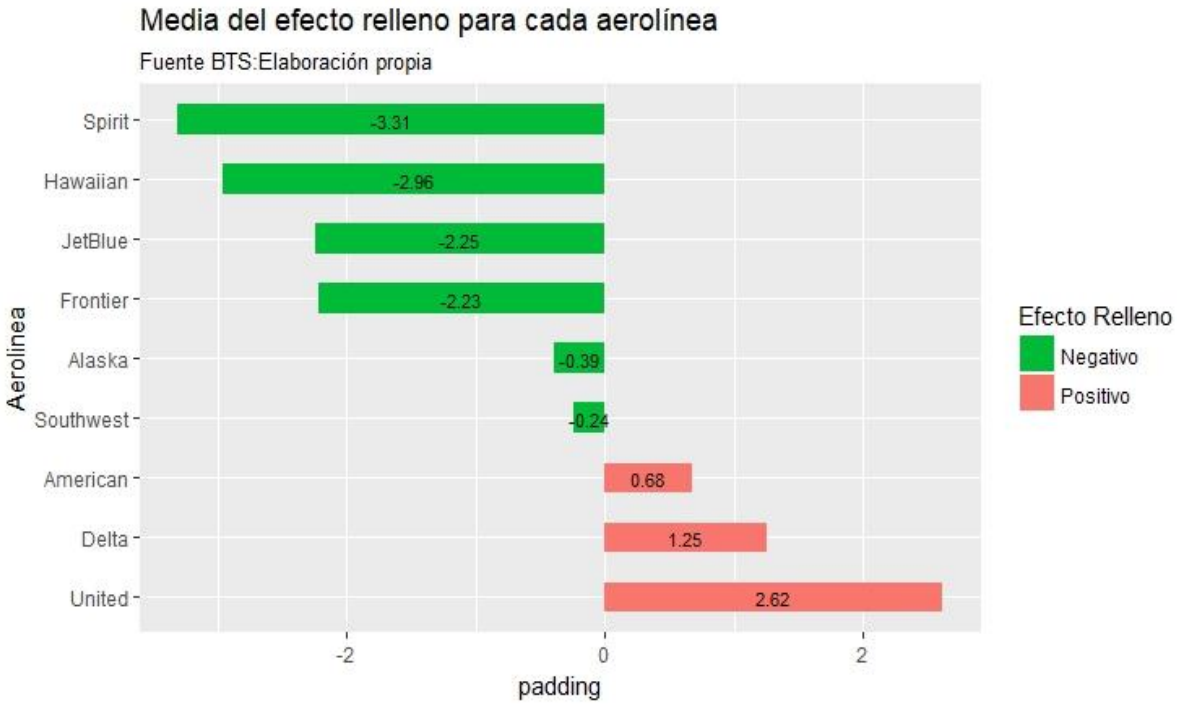


Figura 2.15: Efecto relleno

Se observa un comportamiento diferenciador según el tipo de aerolínea (pequeña o grande) a la hora de rellenar la duración prevista de cada vuelo:

Las aerolíneas regionales independientes Spirit, Hawaian Jetblue y Frontier vuelan con unos tiempos previstos más ajustados a la baja: Spirit y Hawaian restan 3 minutos de media en su tiempo previsto para cada vuelo comparado con las demás aerolíneas. Les siguen de cerca Jetblue y Frontier que restan 2,2 minutos de media.

En cambio, las aerolíneas grandes sí rellenan los tiempos previstos de vuelo al alza, siendo United Airlines la que más rellena: más de 2,6 minutos por vuelo, seguida de Delta que añade 1,2 minutos más.

El efecto relleno, como se ha explicado anteriormente, modifica los porcentajes oficiales de vuelos que llegan a tiempo a su destino. Para calcular el efecto que tiene sobre este porcentaje el hecho de rellenar los tiempos previstos de vuelo se calcula el % sin efecto relleno (se resta al retraso de cada vuelo de las aerolíneas su correspondiente efecto medio relleno)

A continuación, se adjunta la tabla con el efecto sobre los porcentajes de vuelos a tiempo.

Tabla 2.14: Efecto relleno

Aerolinea	% sin relleno	% oficial (con relleno)	Efecto
Spirit	72,4	70	-2,4
Hawaian	90,7	89,3	-1,4
Jetblue	78,2	76,9	-1,3
Frontier	75,7	74,2	-1,5
Alaska	85,7	85,7	0
Southwest	80,7	80,7	0
American	78,7	79,4	+0,7
Delta	84	85,3	+1,3
United	76,2	78,5	+2,3

Se constata claramente, que el efecto relleno sí logra cambiar los porcentajes de vuelos a tiempo para casi todas las aerolíneas: United Airlines ve aumentado su porcentaje oficial de vuelos que llegan a tiempo en 2,3 puntos porcentuales, mientras que Spirit, que vuela con unos tiempos previstos más ajustados, pierde 2,4 puntos en su porcentaje.

Por último, dado que su efecto relleno era casi nulo, Alaska y Southwest presentan los mismos porcentajes para las dos situaciones. Además, aunque se acortan diferencias, el orden de porcentajes de vuelos a tiempo entre aerolíneas no se ve alterado ya que Spirit sigue siendo la aerolínea con una peor tasa de vuelos a tiempo y Hawaian la mejor.

2.5.3 Aeropuertos

2.5.3.1 Los aeropuertos con más trafico

Figura 2.16: Aeropuertos con más tráfico

Los aeropuertos del siguiente gráfico son los 20 aeropuertos con más pasajeros embarcados según la FAA. A todos estos aeropuertos se les denomina hub airports (son aeropuertos con un gran tráfico y varias conexiones). El aeropuerto con más conexiones es el de Hartsfied Jackson Atlanta situado en el estado de Georgia con más de 49 millones de pasajeros en el año 2015. Después se sitúan los aeropuertos de Los Ángeles, Chicago, Dallas y New York.

Para ver una comparativa de aeropuertos en cuanto al retraso, se calcularon las medias de retrasos para cada aeropuerto teniendo en cuenta los retrasos en la salida y llegada.

En este gráfico se muestra el nivel de retraso normalizado para una mejor interpretación.

Figura 2.17: Nivel de retraso por aeropuerto

En este gráfico de barras divergentes, se observa como los tres aeropuertos de Nueva York (La Guardia, John Kennedy y Newark Liberty) presentan unos retrasos medios más por encima de la media global (se sitúan a más de una desviación típica de ésta).

En cambio, los aeropuertos de Seattle, Charlotte, Phoenix son los que menos retraso presentan de los 20 aeropuertos más populares.

2.5.3.2 ¿Qué estados presentan más cancelaciones y retrasos?

Se calcula a continuación para cada uno de los 50 estados su tasa de cancelación de la siguiente manera:

$$Tasa\ de\ cancelación(o\ retraso)\ del\ estado_i =$$

$$\sum \frac{Vuelos\ cancelados(retrasados)\ de\ los\ aeropuertos\ del\ estado_i}{Vuelos\ totales\ aeropuertos\ del\ estado\ _i}$$

Como en un estado puede haber más de un aeropuerto, la tasa de cancelación o de retraso de cada estado es la suma de todos los vuelos cancelados o retrasados de todos los aeropuertos pertenecientes a un mismo estado dividido entre los vuelos totales de estos aeropuertos. La tasa de retraso se ha calculado a partir del retraso en la salida.

Al comparar tasas y no valores absolutos se logra expresar los resultados en términos de probabilidad y se logra eliminar las diferencias de número de conexiones de cada estado para poder comparar entre estados.

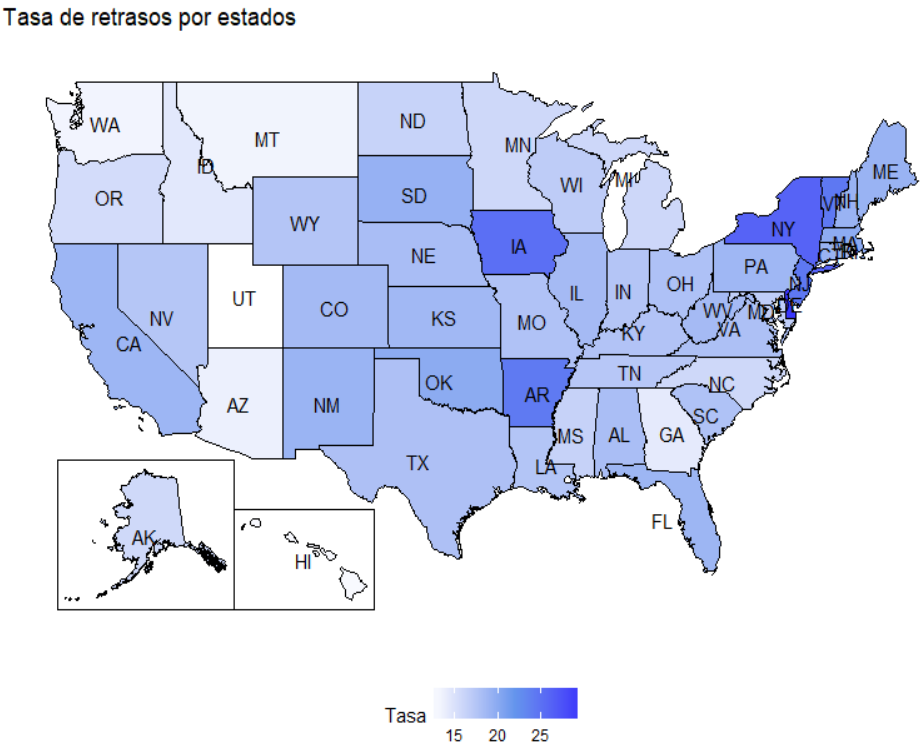


Figura 02.18: Tasa de retraso por estados

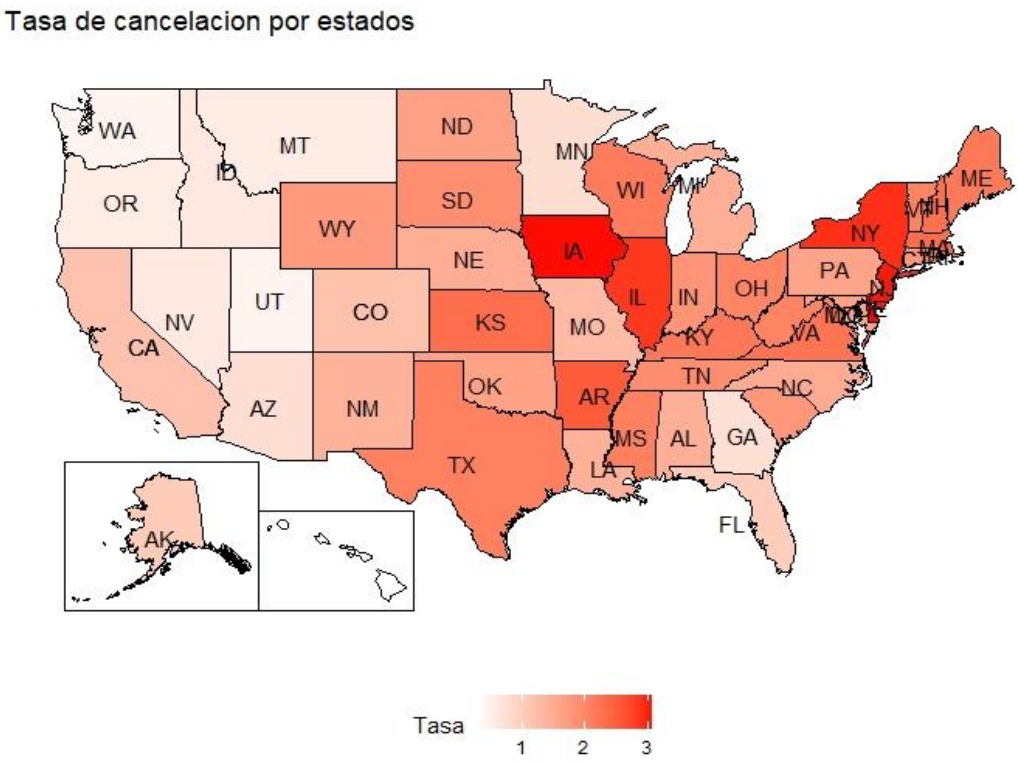


Figura 2.19: Tasa de cancelación por estados

Se observa que los 3 estados con más tasas de cancelación son los de Nueva York, un 3,2 % de los vuelos operados en este estado acaban siendo cancelados, Nueva Jersey con una tasa de cancelación del 3,1% y por último Delaware (2.9%). En el caso de Nueva York y Iowa también ocupan el primer puesto en cuando a tasas de retraso, por encima de 22 %. Estos tres estados se encuentran en la costa este de los Estados Unidos

En cambio, en la parte más este del mapa, se encuentra los estados con menos cancelación y retrasos como es el caso de Hawái con solo un 0,3% de vuelos cancelados y 12% de tasa de retraso, seguido de Washington con un 0,4% de cancelación y un 13,4% de retrasos.

2.5.3.3 **¿Cuántas rutas hay?**

Los vuelos hacen referencia a 4597 diferentes rutas por todo el territorio estadounidense. Graficar todas las rutas en un mapa no tendría sentido ya que se vería una nube indistinguible de líneas, es por eso, que en el dashboard se pueden visualizar las rutas de forma interactiva (ver las rutas con más retraso o con más tráfico).

De las 4597 rutas, un 70% son operadas por una única aerolínea, un 20% por dos, un 7% por tres y solo un escaso 3% por más de tres. Esto indica la ya justificada poca competencia que existe en este mercado.

Tabla 2.15: Distribución del número de aerolíneas por ruta

AEROLINEAS/RUTA	**1**	**2**	**3**	**4**	**5**	**6**
Nº RUTAS	3215	928	325	89	38	2
PORCENTAJE	70%	20%	7%	2%	1%	0,04%

En cuanto a las distancias, la ruta con una distancia mayor es la de Honolulu a Nueva York-JFK (una distancia de 4983 millas) y la que menos distancia presenta es la de Newark a Nueva York-JFK con solo 21 millas.

2.5.3.4 **Efecto jet-stream**

Las corrientes en chorros (en inglés jet streams) son un flujo de aire rápido y fuerte que se encuentra en la atmosfera de la Tierra. Están en la misma altitud que cuando los aviones vuelan y afectan a la velocidad de su ruta.

En la Figura 2.20 se observa la dirección de estas corrientes que van de oeste a este:

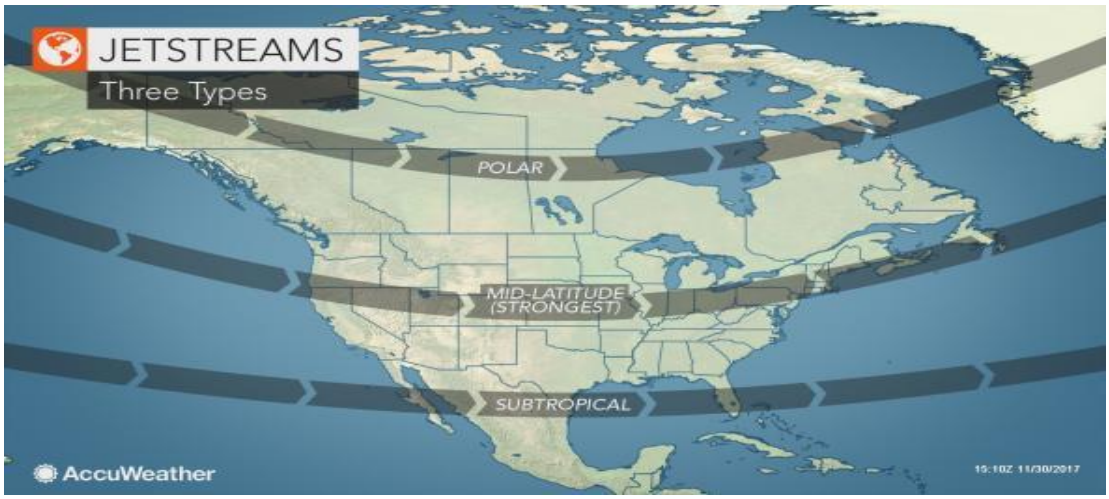


Figura 2.20: Ejemplo de corrientes en chorro. Fuente: AccuWeather

Se calcula para cada ruta su dirección (se compara las longitudes del origen y destino: una ruta va dirección este si la coordenada longitud de origen es menor que la de destino). A continuación, se adjunta el gráfico que confirma el efecto jet stream en los vuelos.

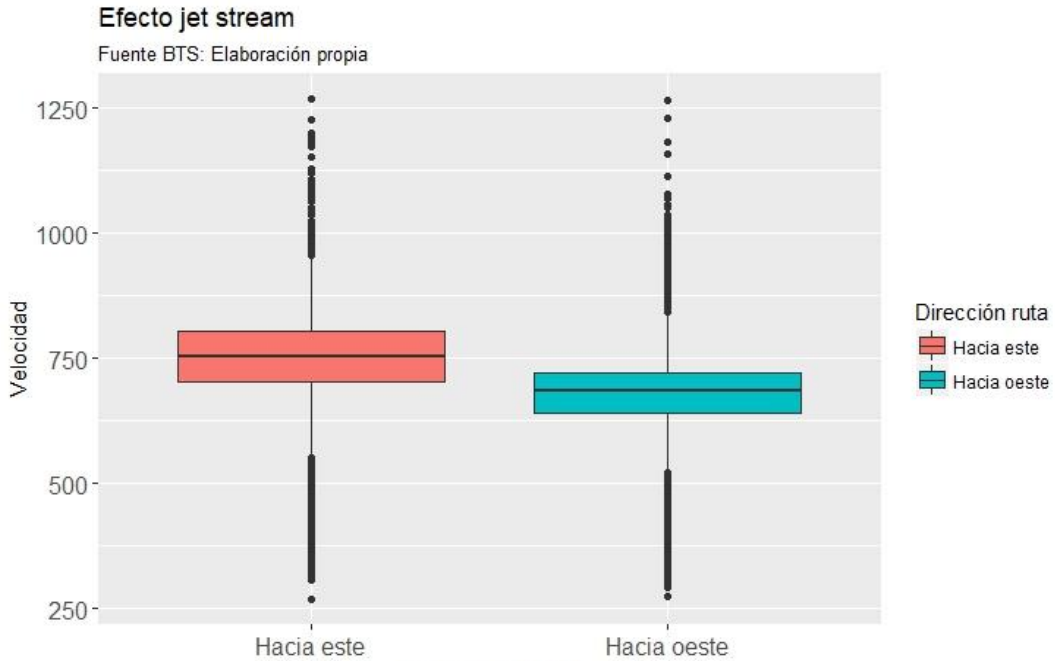


Figura 2.21: Efecto jetstream (Velocidad según dirección ruta)

Se grafica los boxplots de velocidad según la dirección de la ruta. El efecto jet stream se observa claramente: la velocidad mediana de las rutas hacia el este, es decir las que aprovechan este efecto es de 754km/h mientras que las otras rutas tienen una velocidad mediana de 684 km/h. Estadísticamente existen diferencias significativas en la velocidad según la ruta: si se mira el intervalo de confianza, la velocidad mediana de las rutas hacia el este está entre 59,7 y 62,3 km/h por encima de la velocidad mediana de las rutas hacia el oeste.

Salida R 2.7: Resultado test de Wilcoxon

```
> wilcox.test(speed ~ direccion2,data=flights.2 ,alternative="two.sided",conf.int=T)

        Wilcoxon rank sum test with continuity correction

data:  speed by direccion2
W = 1602400000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 59.75554 62.35040
sample estimates:
difference in location
              61.05231
```

¿Entonces, esto quiere decir que las rutas que vuelan hacia el este presentan tasas menores de retraso? La respuesta es no.

El efecto jetstream no solo es conocido por las aerolíneas, por lo que el tiempo previsto de una misma ruta sí varia si va en dirección este u oeste (por ejemplo: ir de Nueva York a Honolulu tiene una duración prevista de 11 horas, pero el regreso solo de 9 horas y media)

Es por eso por lo que no hay una relación estadísticamente significativa entre el porcentaje de retrasos y la dirección de la ruta (p-valor =0,4058)

Salida R 2.8: Resultado test de independencia Chi

```
> chisq.test(table(flights.2$DELAYED,flights.2$direccion2))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(flights.2$DELAYED, flights.2$direccion2)
X-squared = 0.69093, df = 1, p-value = 0.4058
```

2.5.4 Aviones

2.5.4.1 Distribución del tipo de avión y número de motores

La tabla de porcentajes fila del tipo de avión con el número de motores es la siguiente:

Tabla 2.16: Tabla de contingencia de tipo de avión con número de motores

	1 motor	2 motores	3 motores	4 motores	% total
Ala fija >1 motor	16389 (0,3%)	4917480 (99,6%)	4077 (0,08%)	554 (0,02%)	**99,9%**
Rotorcraft	2513 (95,3%)	123 (4,7%)	0	0	**0,1%**
% total	**0,38%**	**99,52%**	**0.08%**	**0.01%**	**100%**

La mayoría de los vuelos utilizan el avión tipo ala fija (representa el 99,9% de los vuelos). La mínima parte restante son vuelos realizados por aviones rotocraft (helicóptero). Casi todos los aviones multimotor son de 2 motores (representa un 99,9% de este tipo de avión) y solo una mínima parte tienen más de 2.

Los aviones de 1 motor y el rotorcraft pertenecen solo a las aerolíneas American y Jetblue y los de 4 motores a las aerolíneas grandes Delta y United. Las demás, solo operan sus vuelos con aviones de más de 1 motor.

2.5.4.2 Antigüedad de la flota

En el fichero de aviones también se dispone de la antigüedad de cada avión. La distribución de antigüedad de la flota de cada aerolínea es la siguiente:

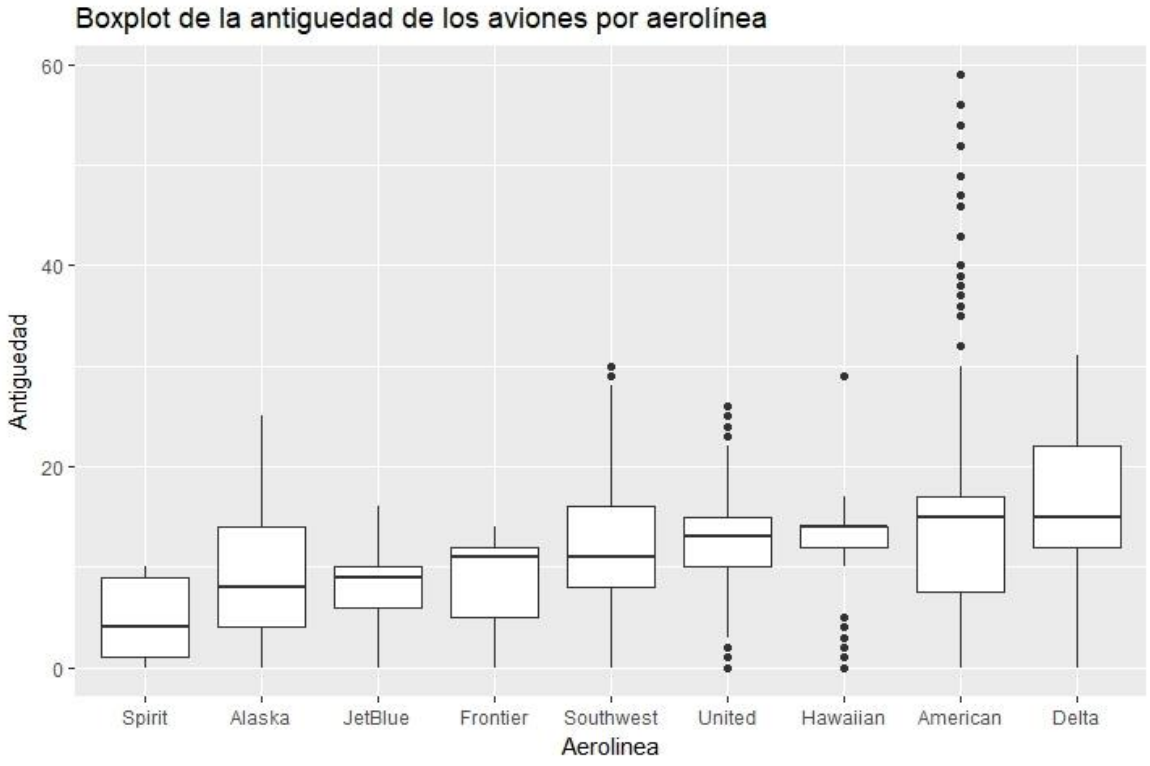


Figura 2.22: Antigüedad flota por aerolínea

La aerolínea pequeña Spirit tiene una flota de aviones con una antigüedad mediana de 4 años, la más baja de todas las aerolíneas. En cambio, Delta tiene una antigüedad de casi 4 veces más la de Spirit, con una mediana de 15 años.

Se observa también que para cada aerolínea la distribución de antigüedad es diferente: America, Frontier, Jetblue y Hawaiian tienen una distribución asimétrica a la izquierda (la mayoría de su flota son aviones con más años de antigüedad). En cambio, Southwest y Delta presenta asimetría a la derecha (aunque tenga una mediana superior que Frontier, la mayoría de su flota contiene aviones con menos años en relación con su misma flota).

American es la aerolínea que posee los aviones más antiguos (14 aviones superan los 30 años) y tiene la antigüedad mediana más alta de todas, solo superada por Delta.

2.5.5 Relación entre las variables numéricas

Para todos los pares de variables numéricas se calcula el coeficiente de correlación de Spearman (ninguna variable satisface el supuesto de normalidad). El corrplot resultante es el siguiente:

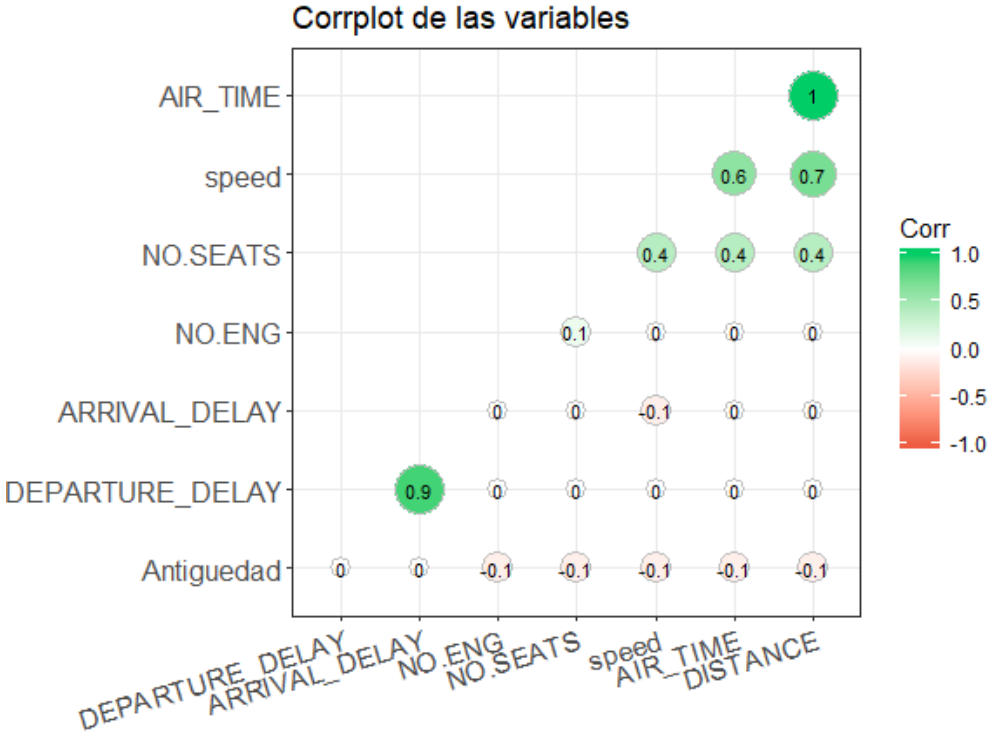


Figura 2.23: Corrplot

Existe una fuerte correlación positiva de 0,9 entre las variables de retraso en la salida y retraso en la llegada (*dep_delay* y *arrival_delay*). Si un vuelo ya empieza a volar más tarde de la hora prevista, a no ser que aumente la velocidad (no es el caso ya que no existe correlación entre velocidad y retraso en la salida) el vuelo casi seguro llegará tarde.

Este dato es bastante revelador: gran parte de los retrasos se originan ya en la salida. Investigando esto, se descubrió que el 74,4% de los vuelos con retraso en la llegada, presentaron retraso en la salida.

También es de destacar la fuerte correlación positiva (0,7) entre velocidad y distancia: más distancia entre las dos ciudades, se asocia a vuelos con una velocidad mayor.

Para el clasificador del capítulo 3 se tendrá en cuenta estas relaciones para evitar problemas graves de multicolinealidad.

2.5.6 Análisis multivariante de los datos: ACM

Por último, se realiza el Análisis de Correspondencias Múltiples para ver relaciones de todo tipo entre las variables categóricas de manera simultánea.

Se decide categorizar las variables numéricas Antigüedad y distancia del vuelo de acuerdo con sus cuartiles.

Las 7 variables activas para realizar el análisis son: la aerolínea, la antigüedad del avión, distancia del vuelo, si fue festivo o no, la hora prevista de salida, el tamaño del avión y el nivel de retraso a la llegada.

Se utiliza la regla del último codo (*last elbow rule*) para determinar el número de dimensiones significativas. Se resta a cada valor propio la inercia trivial (la inversa del número de variables activas) para saber el porcentaje real de variabilidad explicada en cada dimensión.

Con esta regla, el número de dimensiones significativas es 5 (el último codo se encuentra en la dimensión número 6). En el anexo se adjunta la tabla de los valores propios, inercia e inercia acumulada para cada una de estas 5 dimensiones.

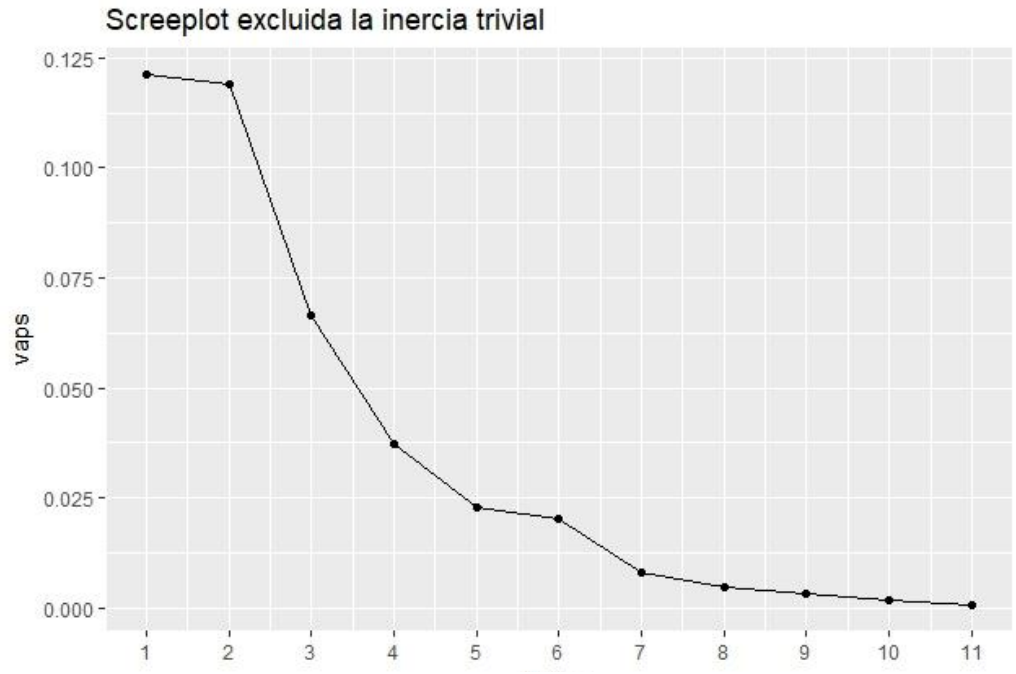


Figura 2.24: Screeplot

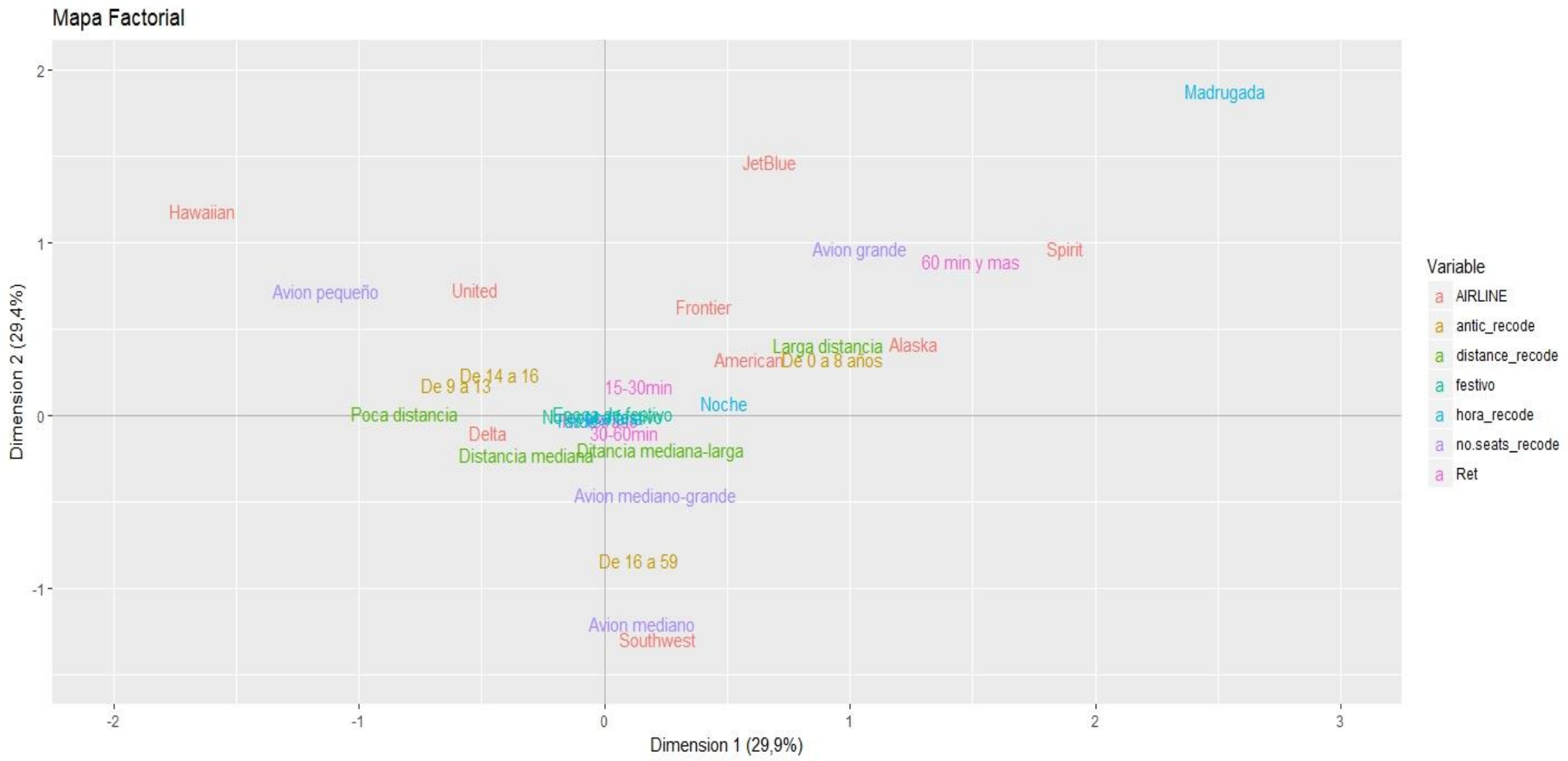
Con 5 dimensiones, se explica el 90.5% de la inercia total y las dos primeras dimensiones explican ya más de un 50% de la variabilidad total.

Tabla 2.17: Valores propios de las dimensiones significativas

	Vap	Variancia (%)	Var. acumulada (%)
dim 1	0,1212	29,9%	29,9%
dim 2	0,119	29,4%	59,3%
dim 3	0,066	16,4%	75,7%
dim 4	0,037	9,2%	84,9%
dim 5	0,022	5,6%	90,5%

La proyección de los individuos, así como el de las variables suplementarias cuantitativas se adjunta en el anexo. (ver Figura 1 y 2). Ninguna de las variables cuantitativas está bien representada en el plano.

A continuación, se adjunta el plano factorial con las proyecciones de las categorías.


Mapa Factorial
Madrugada
JetBlue
Hawaiian
Avion grande
60 min y mas
Spirit
Avion pequeño
United
Frontier
Larga distancia
Alaska
American
De 0 a 8 años
De 14 a 16
De 9 a 13
15-30min
Noche
Poca distancia
Delta
30-60min
Distancia mediana
Distancia mediana-larga
Avion mediano-grande
De 16 a 59
Avion mediano
Southwest
Variable
AIRLINE
antic_recode
distance_recode
festivo
hora_recode
no.seats_recode
Ret
Dimension 2 (29,4%)
Dimension 1 (29,9%)
2
1
0
-1
-2
-1
0
1
2
3

Las 9 aerolíneas están posicionadas en diferentes puntos del mapa: a la izquierda del eje 1 se sitúa Hawaian, United, Delta y Southwest, aerolíneas que presentaban unos porcentajes de retraso inferiores. En cambio, en la parte derecha nos encontramos con Spirit y Frontier que eran las dos aerolíneas con más retrasos (la categoría de retraso superior a una hora está cerca de este grupo de aerolíneas).

Lo mismo pasa con la distancia de los vuelos. Los vuelos de poca distancia se sitúan en la parte izquierda, junto a Hawaian (la mayoría de sus rutas son viajes entre las islas de Hawaii que están cerca entre ellas), United y Delta. En cambio, los vuelos de larga distancia se sitúan a la derecha muy cerca de la aerolínea Alaska (y tiene sentido ya que opera rutas del estado de Alaska, situado al noreste de Canadá, hacia los demás estados).

Como se puede observar en el mapa factorial, las categorías de distancia del vuelo y tamaño del avión están muy relacionadas (y es una relación no lineal), ya que las líneas que unen las categorías ordenadas son paralelas.

La categoría de tamaño medio del avión y aviones con más años de antigüedad están posicionadas muy cerca de Southwest que es una aerolínea que tiene una flota compuesta por aviones de un mismo tipo (Boeing) y de este tamaño.

Los aviones grandes se sitúan alrededor de la categoría de vuelos de larga distancia y también cerca de los aviones con menos de 8 años de antigüedad. Todas estas categorías están alrededor del grupo de aerolíneas con más retrasos ya comentado anteriormente.

En el centro de gravedad se sitúan las categorías no festivo y hora prevista de salida a la mañana o tarde y niveles de retraso inferiores a la hora. Todas estas categorías hacen referencia a vuelos normales que no explican la variabilidad de la nube de puntos (son vuelos considerados normales en media).

En cambio, la categoría hora de salida a la madrugada se encuentra alejada de este centro dado que hay pocos vuelos que salgan a esa hora y cerca de las aerolíneas Spirit y Jetblue que son las aerolíneas que operan más vuelos a esas horas.

CAPÍTULO III.
CONSTRUCCIÓN DE UN CLASIFICADOR DEL ESTADO DE UN VUELO

3.1 Motivación de predecir el estado de un vuelo

Los retrasos aéreos son un problema para el sector ya que genera costes económicos tanto para las aerolíneas como para los pasajeros: según el último informe publicado por el BTS, el coste de los retrasos se cuantificó en 32,9 mil millones de dólares el año 2016.

Se estima que 8,3 mil millones de dólares son costes a la aerolínea como gastos en la tripulación, fuel y mantenimiento, 16,7 mil millones son coste a los pasajeros basado en el tiempo que pierden por el retraso y 3.9 mil millones son el coste de la demanda perdida a causa de que algunos pasajeros evitan viajar por culpa de los retrasos.

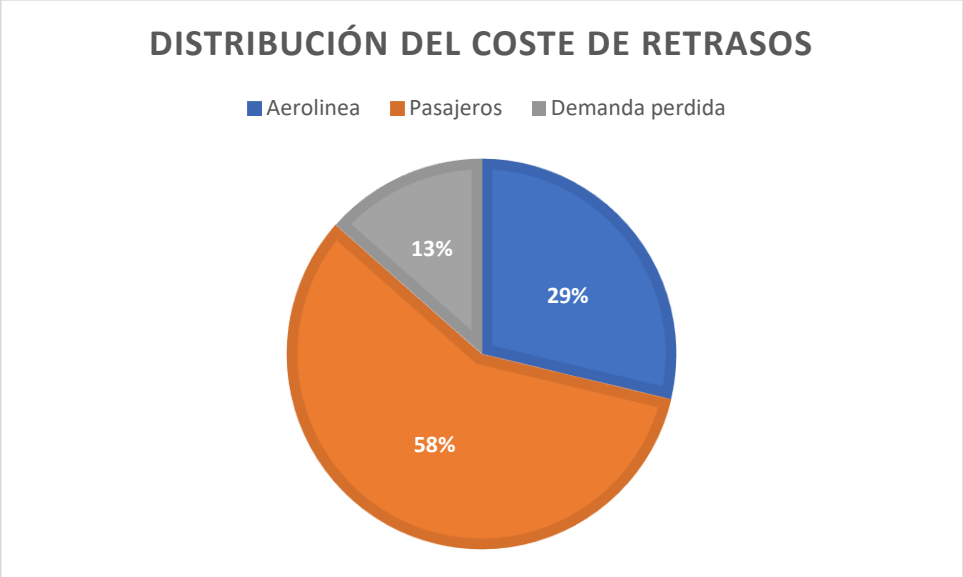


Figura 03.01: Coste de los retrasos

Además de estos costes directos, existe un coste indirecto para la economía en general: la ineficiencia del sector genera menos productividad. Este coste según el BTS es de 4 mil millones de dólares.

El BTS también publica el coste medio del retraso por minuto para las aerolíneas de Estados Unidos: 62,3 dólares por minuto. La base de datos de los vuelos del 2015 que se ha analizado en el capítulo 2 presenta un **retraso total de 67.722.079 minutos** (más de 1 millón de horas) que se traduce en un **coste de 41,8 mil millones de dólares**.

Es por eso por lo que predecir el estado de un vuelo sirve tanto a pasajeros como aerolíneas a anticiparse antes y reorganizar sus tiempos y por último evitar estos grandes costes para todos. Si un pasajero sabe de antemano la probabilidad de retraso de su vuelo, puede anticiparse y cambiar sus planes.

A continuación, se explica el proceso de creación de una regresión logística que prediga si un vuelo sufre o no retraso.

3.2 Reducción de la base de datos e incorporación de variables meteorológicas

Se decidió reducir la base de datos ya que, con la potencia de un solo ordenador, realizar un modelo con millones de datos no es posible (R no puede almacenar tanta memoria).

En lugar de escoger una muestra aleatoria de la base de datos, se escogió todos los vuelos de las 10 primeras rutas más populares (en total 87294 registros haciendo así factible la modelización).

El motivo de este método de reducción fue que se decidió incorporar variables meteorológicas al análisis para así poder aumentar la capacidad predictiva del modelo y entonces buscar estos datos para los 322 aeropuertos requería mucho tiempo de espera.

Se divide la base de datos reducida en el training set (2/3 de los registros totales) y el test set (el 1/3 restante).

3.3 Selección de variables predictoras

La variable a predecir es DELAYED, variable binaria que toma valor 0 si el vuelo no sufrió retraso menor a 15 minutos, y 1 si sufrió retraso mayor o igual a 15 minutos.

Se decidió descartar del modelo las variables referentes a los tiempos de vuelo (departure_delay, *taxi_out, wheels,off, wheels_on, taxi_in*). También se eliminó las variables de causa de retraso (*security_delay, airline_delay, late_aircraft_delay, weather_delay, air_system_delay*) y de cancelación (*cancelled, cancellation_reason*).

La decisión de eliminarlas fue tomada ya que son variables que se recogen durante o después del vuelo e interesa saber el resultado de la predicción antes de esto.

Dado que la creación de esta regresión tiene un objetivo principalmente predictivo, la posible multicolinealidad no es un problema grave si no es excesiva. De toda forma, se revisa que las variables que se incluyen no estén muy correlacionadas entre ellas (se elimina también *distance* por la correlación alta con *air_time*)

Finalmente, para la construcción de la regresión logística se utilizará la siguiente información: el mes, día de la semana, la hora de salida prevista, si es época de festivos o no, la aerolínea, el tiempo de vuelo, el origen y destino, la velocidad, el número de asientos y de motores del avión, su antigüedad, el tipo de motor, el tipo de avión y todas las variables meteorológicas.

3.4 Elección de la función de enlace

Se estima un modelo con todas las variables predictoras sin interacciones para cada función de enlace: logit (modelo m1) y probit (modelo m2).

La regla de decisión se basa en ver qué modelo tiene el criterio de información más pequeño (aunque al ser modelos con los mismos grados de libertad, la devianza residual también serviría como criterio de elección).

Al tener los dos criterios de información menores, la función logit es la elegida.

Tabla 3.1: Resultados del AIC y BIC para las dos link functions

	AIC	BIC
Modelo logit	55693,75	56151,55
Modelo probit	55812,73	56270,53

3.5 Estudio de interacciones de orden 2 y selección final con stepwise

Después de decidir la función de enlace, se realiza el test de la devianza para ver la significación de las variables en el modelo con la función Anova que nos proporciona los efectos netos.

Salida R 3.1: Tabla Anova

```
> Anova(m1,type=2)
Analysis of Deviance Table (Type II tests)

Response: DELAYED
                     LR Chisq Df Pr(>Chisq)
MONTH                  617.45 11  < 2.2e-16 ***
DAY_OF_WEEK            335.15  6  < 2.2e-16 ***
AIRLINE                122.18  6  < 2.2e-16 ***
AIR_TIME               448.14  1  < 2.2e-16 ***
ORIGIN                 396.61  3  < 2.2e-16 ***
DESTINATION            481.58  3  < 2.2e-16 ***
speed                   31.21  1  2.311e-08 ***
Antiguedad               2.23  1     0.1353
TYPE.ENG                 2.54  3     0.4674
hora_prev_salida      1334.85 18  < 2.2e-16 ***
HOURLYVISIBILITY        23.93  1  9.995e-07 ***
HOURLYDRYBULBTEMPC       0.91  1     0.3393
HOURLYRelativeHumidity  86.56  1  < 2.2e-16 ***
HOURLYWindSpeed        101.51  1  < 2.2e-16 ***
HOURLYStationPressure   33.64  1  6.643e-09 ***
HOURLYPrecip            99.30  1  < 2.2e-16 ***
DAILYSnowDepth           0.62  1     0.4319
seats_recode            31.93  3  5.404e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

De las 18 variables del modelo, 14 son significativas: 7 variables categóricas y 7 covariables. La antigüedad del avión, el tipo de motor y la temperatura no son factores significativos según el test Anova de efectos netos (p-valor>0.05).

Por lo tanto, existen un total de 63 interacciones posibles de orden 2 (21 interacciones de 2 factores y 42 de un factor con una covariable). Para intentar sacar el máximo potencial a los datos se decide evaluar todas estas interacciones de la siguiente manera:

A continuación, se construye un modelo con las 14 variables significativas del test Anova (modelo m1dep).

Para ver la significación de cada una de estas 63 interacciones, se ha comparado dos modelos: el modelo con las 14 variables significativas (modelo m1dep) con el modelo que resulta de incluir la interacción a estudiar (m3 = m1dep + interacción). Esta comparación se ha realizado para cada una de las 63 posibles interacciones mediante un bucle en R.

Como los dos modelos son anidados se utiliza el test de la Chi para ver si la reducción de la devianza residual al incorporar la interacción es realmente significativa (que querrá decir que la interacción en si es significativa). De acuerdo con el resultado de los tests, 46 interacciones de un total de 63 son significativas.

El modelo que resulta de incluir todas estas 46 interacciones junto con las variables significativas (modelo m4) consigue reducir en 4052 unidades la deviancia residual y esta reducción, según el test es significativa (el test de la Chi nos da un pvalor menor a 0.05).

Salida R 3.2: Test Chi de los dos modelos

```
> anova(m1dep,m4,test="Chi")
Analysis of Deviance Table

Model 1: DELAYED ~ MONTH + DAY_OF_WEEK + AIRLINE + DESTINATION + TYPE.ENG +
    festivo + AIR_TIME + speed + NO.SEATS + hora_prev_salida +
    HOURLYVISIBILITY + HOURLYRelativeHumidity + HOURLYWindSpeed +
    HOURLYStationPressure + HOURLYPrecip + DAILYSnowDepth
Model 2: DELAYED ~ MONTH + DAY_OF_WEEK + AIRLINE + DESTINATION + TYPE.ENG +
    festivo + AIR_TIME + speed + NO.SEATS + hora_prev_salida +
    HOURLYVISIBILITY + HOURLYRelativeHumidity + HOURLYWindSpeed +
    HOURLYStationPressure + HOURLYPrecip + DAILYSnowDepth + MONTH *
    AIR_TIME + DAY_OF_WEEK * AIR_TIME + AIRLINE * AIR_TIME +
    DESTINATION * AIR_TIME + TYPE.ENG * AIR_TIME + MONTH * speed +
    DAY_OF_WEEK * speed + AIRLINE * speed + DESTINATION * speed +
    TYPE.ENG * speed + MONTH * NO.SEATS + AIRLINE * NO.SEATS +
    DESTINATION * NO.SEATS + MONTH * hora_prev_salida + DAY_OF_WEEK *
    hora_prev_salida + AIRLINE * hora_prev_salida + DESTINATION *
    hora_prev_salida + TYPE.ENG * hora_prev_salida + festivo *
    hora_prev_salida + MONTH * HOURLYVISIBILITY + AIRLINE * HOURLYVISIBILITY +
    festivo * HOURLYVISIBILITY + MONTH * HOURLYRelativeHumidity +
    DAY_OF_WEEK * HOURLYRelativeHumidity + AIRLINE * HOURLYRelativeHumidity +
    DESTINATION * HOURLYRelativeHumidity + MONTH * HOURLYWindSpeed +
    DAY_OF_WEEK * HOURLYWindSpeed + DESTINATION * HOURLYWindSpeed +
    MONTH * HOURLYStationPressure + AIRLINE * HOURLYStationPressure +
    DESTINATION * HOURLYStationPressure + MONTH * HOURLYPrecip +
    DAY_OF_WEEK * HOURLYPrecip + DESTINATION * HOURLYPrecip +
    DAY_OF_WEEK * DAILYSnowDepth + DESTINATION * DAILYSnowDepth +
    MONTH * DAY_OF_WEEK + MONTH * AIRLINE + MONTH * DESTINATION +
    MONTH * TYPE.ENG + MONTH * festivo + DAY_OF_WEEK * AIRLINE +
    DAY_OF_WEEK * DESTINATION + DAY_OF_WEEK * festivo + AIRLINE *
    DESTINATION
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1     58444      56023
2     57889      51971 555   4052.5 < 2.2e-16 ***
```

Por último, para escoger el modelo final se realiza una selección automática con el método stepwise partiendo del modelo nulo, con direction="both" (el método permite

eliminar e incluir variables durante el procesos de selección) y como límite (upper scope) se fija el modelo con las 46 interacciones significativas (modelo m4). Así, la convergencia es más rápida y no se considera la inclusión de una interacción hasta que las dos variables están ya en el modelo.

El modelo final resultante contiene todas las variables significativas de la tabla Anova y 31 interacciones de las 46 que eran significativas. Al final se consigue una devianza residual de 50716, frente a la devianza del modelo sin ninguna interacción que vale 56025 y la devianza del modelo con todas las interacciones significativas de 51971. Aunque se gastan muchos más grados de libertad, la reducción es significativa.

3.6 Validación y capacidad predictiva del modelo

Para una validación completa del modelo logístico es necesario tener datos agregados. En este caso debido al gran número de predictores no es factible hacerlo. En un nivel máximo de desagregación, interpretar los residuos al 100% es complicado.

Si se observa el gráfico de los residuos de Pearson contra las variables del modelo (Figuras 3.2 y 3.3) vemos que la hipótesis de linealidad parece cumplirse para todas estas. Junto a estos gráficos R adjunta el resultado del test de lack of fit para ver si una variable está relacionada con los residuos. El test añade un término cuadrático al modelo para examinar si este es estadísticamente significativo.

Salida R 3.3: Test lack of fit

	Test stat	Pr(>|Test stat|)	
hora_prev_salida			
speed	1.8061	0.1789826	
DESTINATION			
MONTH			
AIR_TIME	207.7014	< 2.2e-16	***
DAY_OF_WEEK			
HOURLYWindSpeed	13.6336	0.0002222	***
AIRLINE			
HOURLYRelativeHumidity	0.0632	0.8015230	
HOURLYPrecip	4.2617	0.0389811	*
festivo			
HOURLYVISIBILITY	2.4617	0.1166555	
HOURLYStationPressure	0.0007	0.9796303	

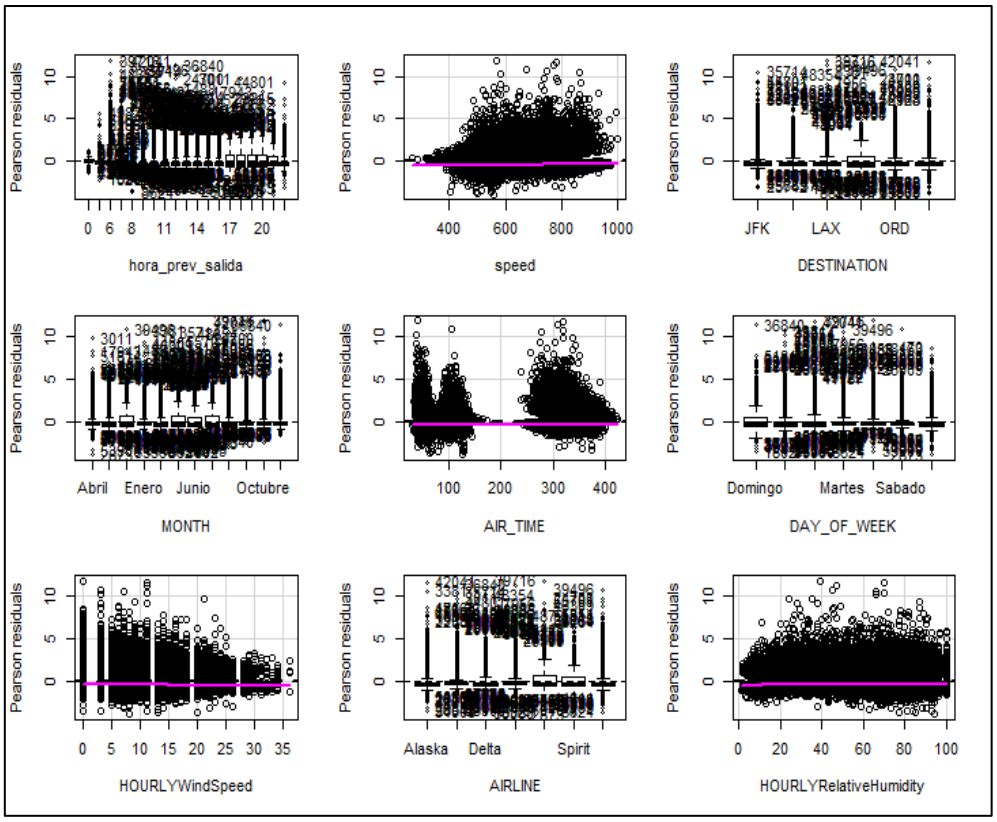


Figura 3.2: Gráficos de validación I

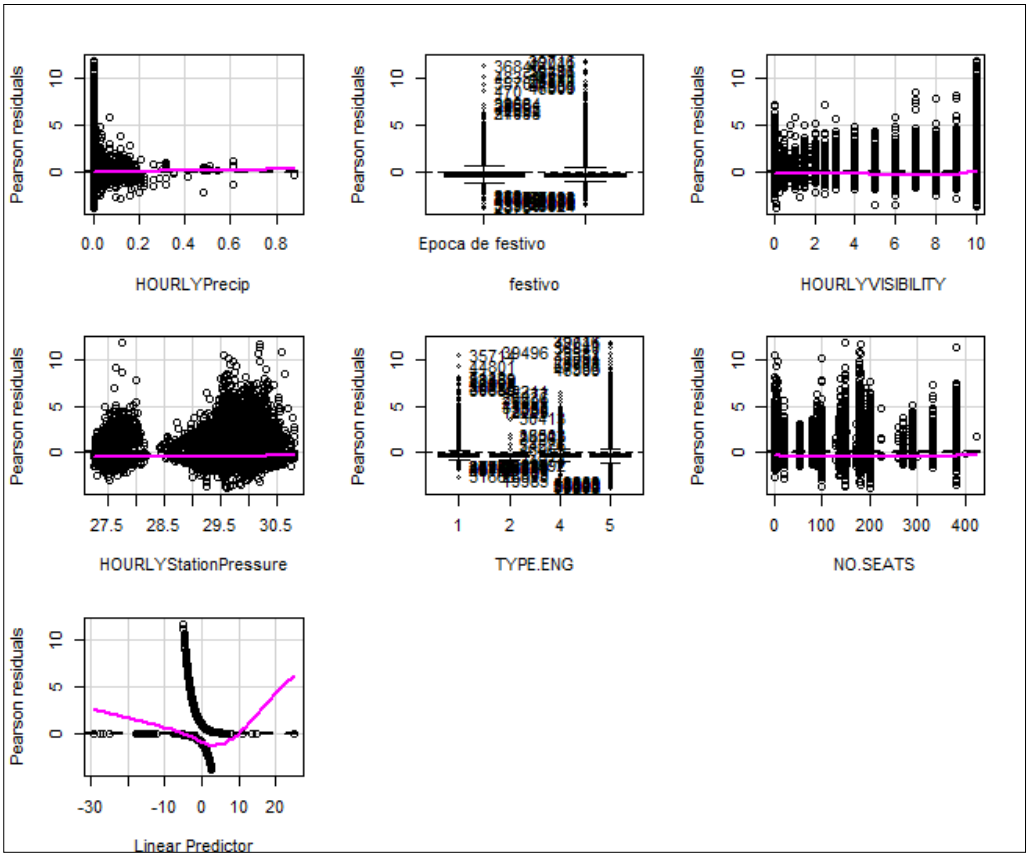


Figura 3.3: Gráficos de validación II

Las variables *air_time, hourlywindspeed y hourlyPrecip* salen significativas según el test, por lo que se añade el término cuadrático de cada una al modelo (con la instrucción I(variable^2)). Los resultados del test se adjuntan en Anexo Salida R.6. Todos los términos salen como no significativos por lo que se concluye que el modelo esta correctamente especificado.

Se analiza también la presencia de outliers y datos influyentes.

Con la instrucción outlierTest se analizan los residuos estudentizados para ver si existen outliers. La observación 7649 es la que posee un resido estudentizado mayor (superior a 3).

Salida R 3.4: Test de outliers

```
> outlierTest(mdef2)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferonni p
7649 3.382849         0.00071738           NA
```

Para los datos influyentes y outliers se visualiza los gráficos de diagnóstico. Se utiliza la distancia de Cook como regla de decisión ya que es una medida que sintetiza la información del apalancamiento y de outlier:

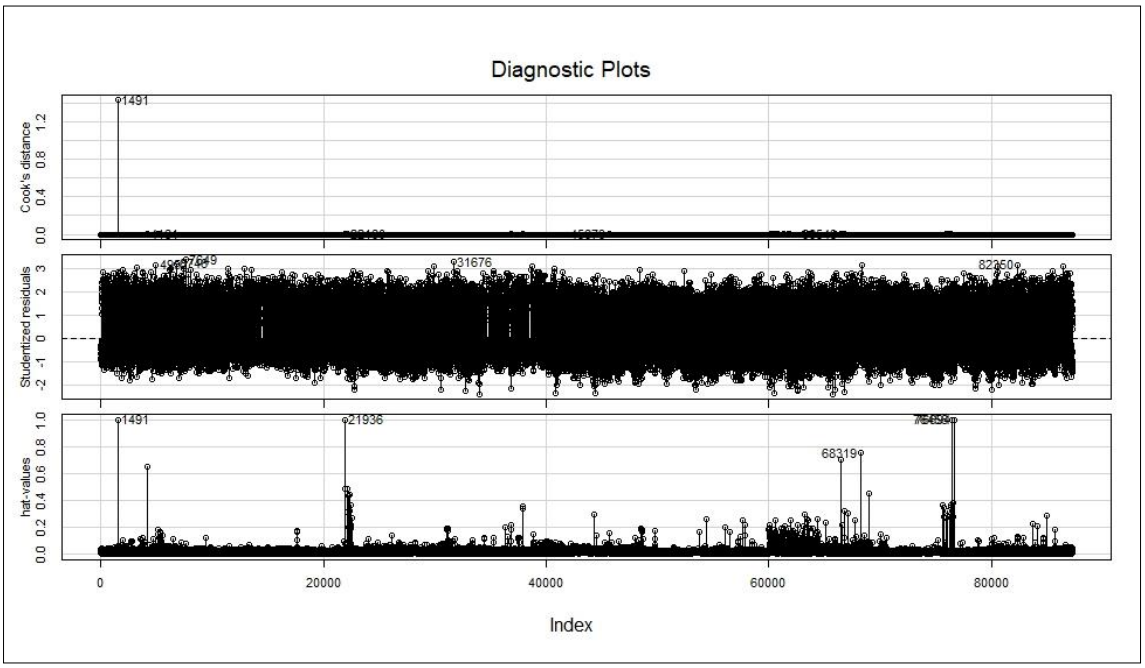


Figura 3.4: Gráficos de diagnóstico

Destaca la observación 1491, con una distancia de Cook muy elevada, superior a 1 y muy alejada del resto de valores. Esta observación hace referencia a un vuelo con hora prevista a las 12 de la noche con retraso de San Francisco a Nueva York y que no llovió ni nevó, operado por Jetblue con un avión estrenado ese mismo año.

Para verificar si esta observación tiene un gran impacto en los valores de los coeficientes del modelo, se vuelve a estimar el modelo sin esta observación y se compara los coeficientes de ambos modelos con la instrucción compareCoefs() del paquete car. Se

observa que la exclusión de esta observación no afecta a los coeficientes y por lo tanto no es un dato verdaderamente influyente.

Después de la validación, se calcula la capacidad predictiva del modelo a partir de los datos de test. Con la función predict de R se obtiene la predicción del modelo. El cutoff elegido para clasificar el estado del vuelo es de 0,5 (es el valor predeterminado)

La matriz de confusión e indicadores de calidad del modelo para los datos de test con un valor límite de 0,5 son los siguientes:

		Predicción	
		No Retraso	Retraso
Observado	No Retraso	21357	987
	Retraso	4817	1646

Tabla 3.2: Matriz de confusión

	Test set
Sensibilidad	25,5%
Precisión	62,5%
Ratio de error	20,1%

Tabla 3.3: Indicadores

Con estos datos se concluye que el modelo realiza un 80% de las predicciones de forma correcta y que un 62,5% de los vuelos sin retraso son correctamente predichos (el modelo tiene una notable capacidad de detectar los vuelos sin retrasos que es la clase mayoritaria).

Pero, el indicador más importante para el objetivo es la sensibilidad ya que informa de la capacidad que tiene el modelo de detectar los vuelos con retrasos y por desgracia, la sensibilidad es bastante baja ya que solo un 25,3% de las predicciones de vuelos con retraso son correctas.

La regresión logística finalmente utilizada es mejor para predecir que la que resultaría de no tener en cuenta ninguna de las interacciones (modelo m1dep con las 14 variables significativas). Aunque la tasa de error sea prácticamente la misma, la sensibilidad del modelo m1dep es de un 10,7%, frente al 25,5% finalmente conseguido). La matriz de confusión y los indicadores se encuentran en el Anexo Tabla 2.

Estos tres indicadores dependen del valor límite que se escoja: por ejemplo, una manera de aumentar la sensibilidad sería reducir el valor límite por el cual se decide clasificar un vuelo como retrasado, pero esto hace aumentar bastante la tasa de error global (si el cutoff fuera de 0.1, la sensibilidad aumenta hasta un 50% pero la tasa de error seria casi del 90%).

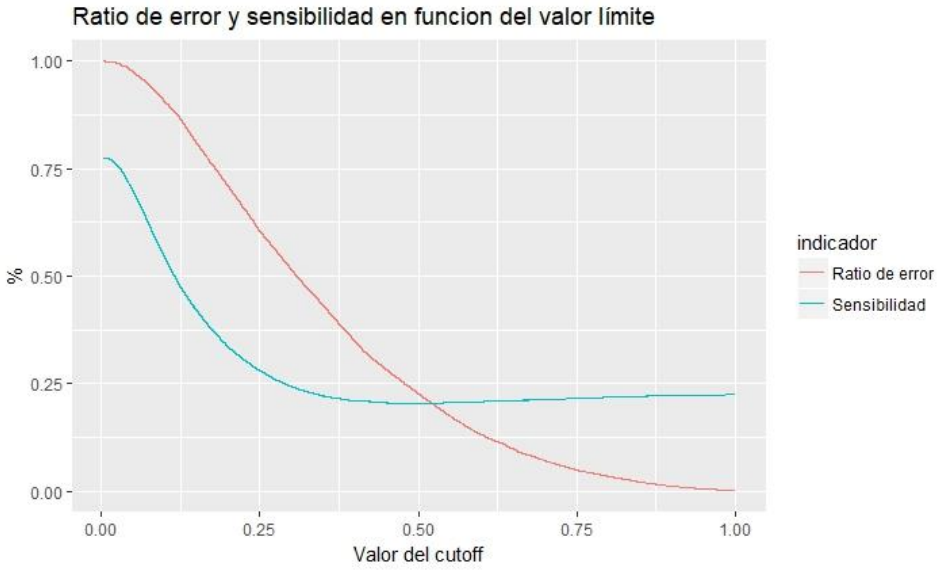


Figura 3.5: Ratio de error y sensibilidad en función del valor límite

La medida por excelencia para validar la capacidad predictiva de los modelos es la curva ROC y su estadístico AUC ya que no dependen del valor límite escogido, a diferencia de los indicadores ya mencionados. Si se calcula el área debajo de esta curva, el estadístico **AUC vale 0,76** que significa que el modelo tiene una buena capacidad predictiva, sin llegar a ser extraordinaria (per mejora el AUC del modelo m1dep que vale 0,7)

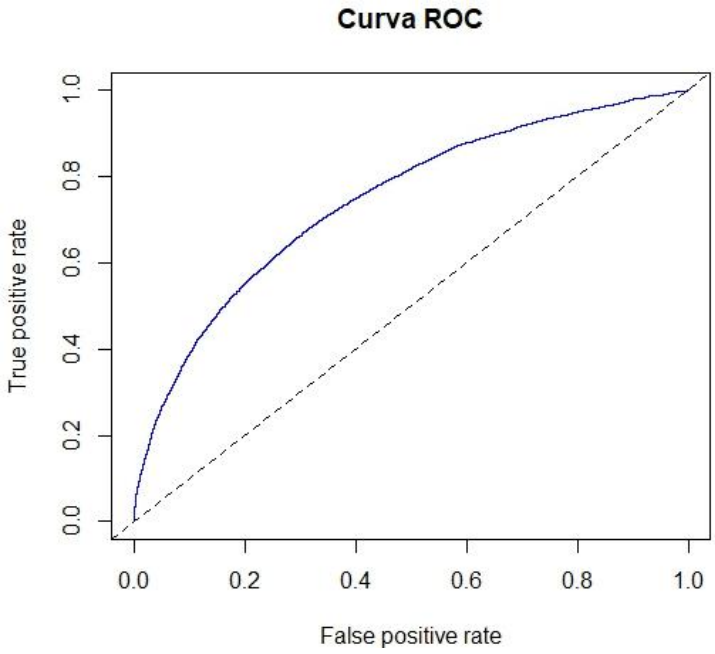


Figura 3.6: Curva ROC

La conclusión es que con las variables disponibles es difícil predecir bien los retrasos. Una de las causas puede ser el hecho de que se haya filtrado la base de datos (de más de 5 millones de registros a solo 87 mil) o que falten más variables claves que expliquen el retraso. Desde febrero de este mismo año, Google Flights nos permite saber con antelación el estado de un vuelo. Para ello, Google utiliza datos **históricos** sobre el

estado de un vuelo de diferentes años. Seguro que, teniendo en cuenta, no solo un año sino un periodo más extenso, el modelo es capaz de mejorar su capacidad predictiva.

El problema de las clases descompensadas (un 80% son vuelos sin retraso) también agrava el problema de encontrar un buen clasificador que prediga bien la clase minoritaria.

CAPÍTULO IV. CREACIÓN DEL DASHBOARD

En este capítulo se explica cómo se ha implementado una aplicación interactiva para visualizar la información que proporcionan los vuelos del análisis.

4.1 Justificación de implementar una aplicación

En la descriptiva del capítulo 2 se analizó aerolíneas, aviones, aeropuertos y rutas, pero es tanta la cantidad de información que es por eso que se construye un dashboard donde el usuario pueda recopilar aún más información como por ejemplo saber que aerolínea es la mejor para según qué ruta, ver donde se concentran los aeropuertos con más retraso y un largo etcétera.

Por otra parte, en el ámbito del data science, es importante saber comunicar los resultados de un análisis a un público amplio y es por ello que Shiny es una gran opción ya que permite presentar los resultados de una forma interactiva mediante tablas, gráficos y otras opciones de visualización.

4.2 Shiny

Shiny es un paquete de R creado en el año 2012 por Rstudio. Su función principal es crear aplicaciones interactivas mediante código R que luego el mismo paquete traduce a lenguaje HTML,CSS y javascript logrando así una aplicación web. Es por ello por lo que no es necesario saber programación de páginas webs.

La aplicación siempre se compone de dos partes:

- **User-interface script** (UI): contiene el diseño y el aspecto de la aplicación como por ejemplo cuantas pestañas tendrá la aplicación, donde se ubicará cada elemento a visualizar, etc.
- **Server script** (SERVER): A partir de los objetos definidos en el UI se construye los resultados a mostrar.

Estas dos partes se guardan o en dos archivos separados (ui.R y server.R) o en un único archivo llamado app.R

Los objetos de una aplicación son de dos tipos:

- **Inputs**: conjuntos de valores que el usuario puede introducir y cambiar mediante los widgets (cajas de selección, botones, inputs numéricos, etc). Gracias a ellos, el usuario interacciona con la aplicación. Los inputs se construyen en el UI.

- **Outputs**: Son los objetos que cambian en función de los inputs mediante reactividad. Pueden ser tablas, gráficos, textos, imágenes y un largo etc. La creación de estos se realiza en el SERVER.

La aplicación se ha realizado mediante el paquete *Shinydashboard* el cual proporciona un formato predeterminado de presentación. El formato del dashboard siempre se divide en 3 partes:

- Header: el título principal de la aplicación. Se sitúa en la parte superior de la aplicación

- Sidebar: Hace referencia a las pestañas en las cuales se divide la aplicación. Se sitúa al lado izquierdo.
- Body: Panel principal donde se visualizan todos los inputs y outputs. Por cada pestaña, el cuerpo principal cambia, lo que nos permite visualizar los elementos de manera más fácil.

Un ejemplo de dahsboard con 4 pestañas y de título “MyDashboard”:

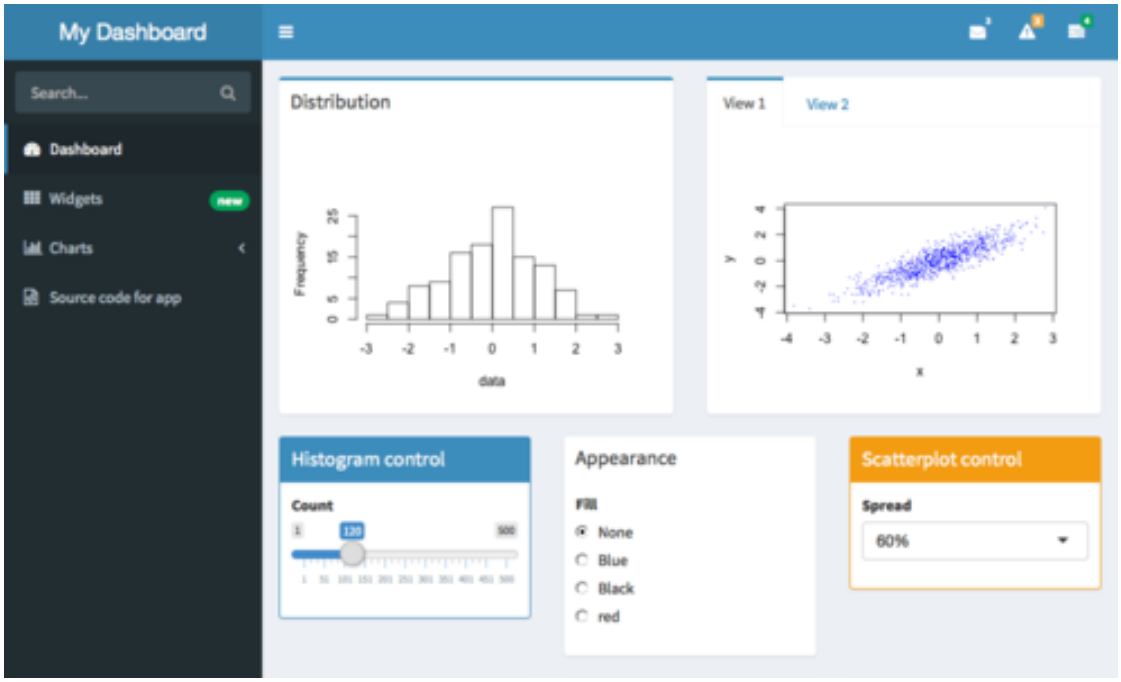


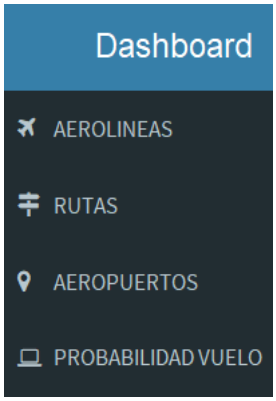
Figura 4.1 Ejemplo aplicación en Shiny

4.3 Estructura de la aplicación

En este apartado se explica la estructura de la aplicación Shiny. Debido a la inmensa cantidad de información que podemos extraer de la base de datos de vuelos se ha estructurado la aplicación en 4 pestañas: AEROLINEAS, RUTAS, AEROPUERTOS, PROBABILIDAD DEL VUELO.

La aplicación está dirigida a usuarios que quieran saber sobre que aerolíneas son las mejores para según qué rutas, los aeropuertos y rutas con más tráfico o con más retrasos, cuanto de probable es que un vuelo llegue con retraso a su destino, etc.

El sidebar de la aplicación, junto al código de R que lo define se muestra a continuación:




```
sidebar<- dashboardSidebar(
  sidebarMenu(
    menuItem("AEROLINEAS", tabName = "AEROLINEAS", icon = icon("plane"),
             menuSubItem(text = "Estado de los vuelos", tabName = "Estado"),
             menuSubItem(text = "Rutas", tabName = "rutasaero"),
             menuSubItem(text="Flota",tabName="flotas"),
             menuSubItem(text = "Comparativa retrasos", tabName = "comparativa")
    ),
    menuItem("RUTAS", tabName = "RUTAS", icon = icon("map-signs"),
             menuSubItem(text="Mejor aerolinea",tabName="mejoraer"),
             menuSubItem(text="Ranking de rutas",tabName = "ranking"),
             menuSubItem(text="Historico de rutas",tabName = "historicorutas")

    ),
    menuItem("AEROPUERTOS", tabName = "AEROPUERTOS", icon = icon("map-marker")),
    menuItem("PROBABILIDAD VUELO", tabName = "clasificador", icon = icon("laptop"))
```

Figura 4.2: Sidebar de la aplicación

Las pestañas Aerolineas y Rutas constan de submenús como se ve en el código de R. A continuación, se explica el contenido de cada pestaña.

4.3.1 Pestaña AEROLINEAS

Esta pestaña consta de 4 submenús: Estado de los vuelos, Rutas, Flota y Comparativa de retrasos.

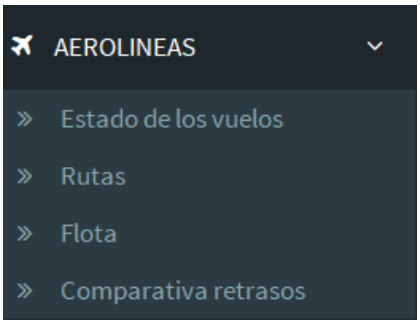


Figura 4.3: Submenús pestaña AEROLINEAS

4.3.1.1 **Estado de los vuelos**

En el submenú de Estado de los vuelos se visualiza lo siguiente:

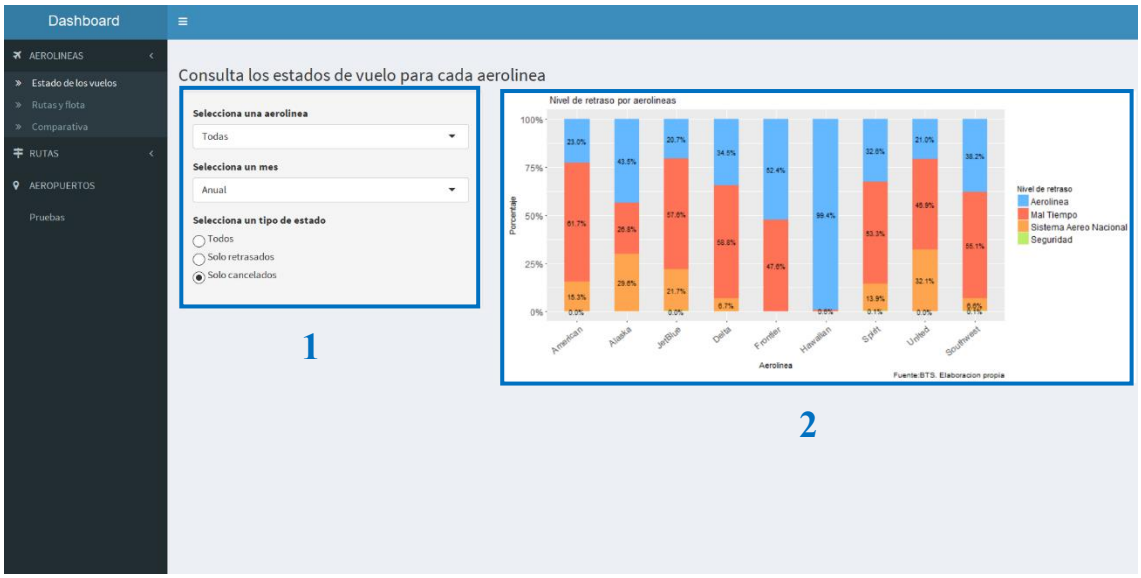


Figura 4.4: Estado de los vuelos (si se decide ver todas las aerolíneas)

Esta pestaña consta de un sidebarPanel (1) donde se selecciona la aerolínea, el mes y el tipo de estado de vuelo a visualizar. También cabe la posibilidad de seleccionar todas las aerolíneas y/o el dato anual.

El mainPanel (2) consta de un gráfico que va cambiando según la opción elegida. Además, según si se ha seleccionado una aerolínea o todas, el gráfico cambia de tipo (gráfico de barras acumuladas si todas o un gráfico circular si es una aerolínea determinada).

Esto se realizó gracias al conditionalPanel (es un mainPanel pero que cambia según una condición). En la imagen anterior, es el caso de mainPanel si se desea visualizar todas las aerolíneas. La imagen de a continuación es para el caso de una sola aerolínea donde además del gráfico se presenta las diferentes tasas.

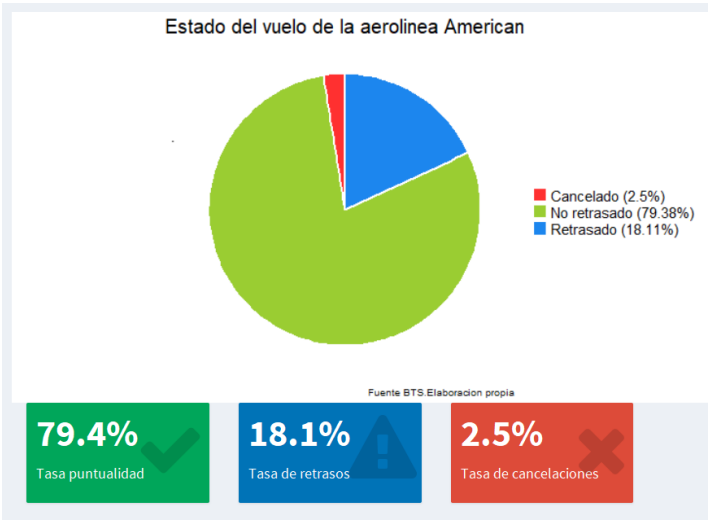


Figura 4.5: Estado de los vuelos (si solo una aerolínea)

4.3.1.2 **Rutas**

En el sidebarPanel (1), según la ciudad de origen seleccionada, en el segundo desglosador se ha de elegir las aerolíneas que vuelan desde esta ciudad. En el mainPanel se visualiza un mapa con las posibles rutas que existen desde esta ciudad para la aerolínea seleccionada (2), junto con una tabla con la misma información (3). Además, la aplicación nos muestra el número de rutas totales para aquella aerolínea.

Como la selección de aerolínea depende de la ciudad de origen que se escoja (no todas operan vuelos en la misma ciudad), el input "selector de aerolínea" se creó en el server mediante la función renderUI.

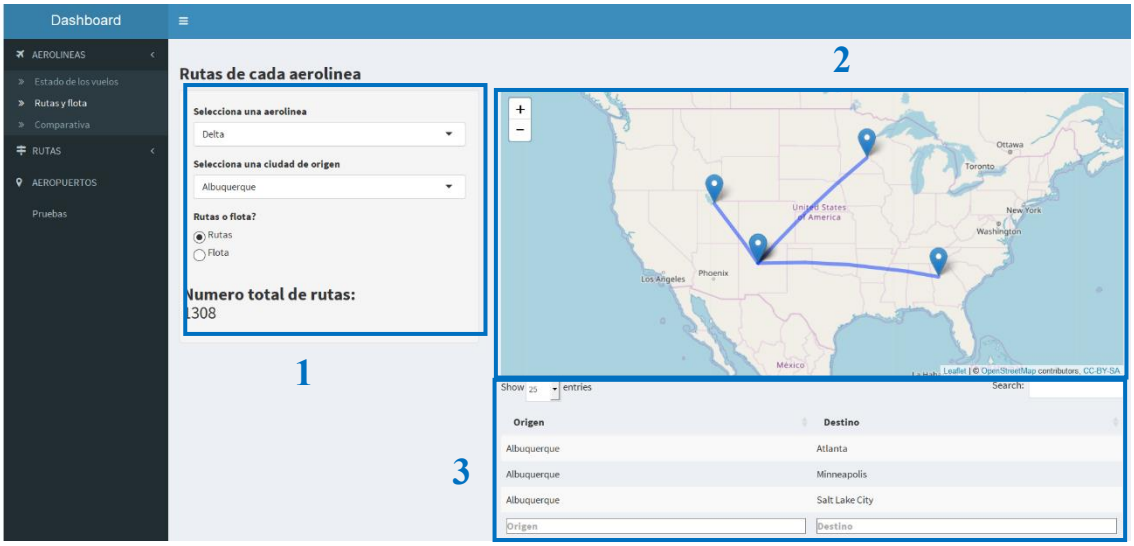


Figura 4.6: Rutas de cada aerolínea

4.3.1.3 **Comparativa de retrasos**

Por último, el submenú de Comparativa nos permite visualizar las diferencias de retrasos que existe entre aerolíneas.

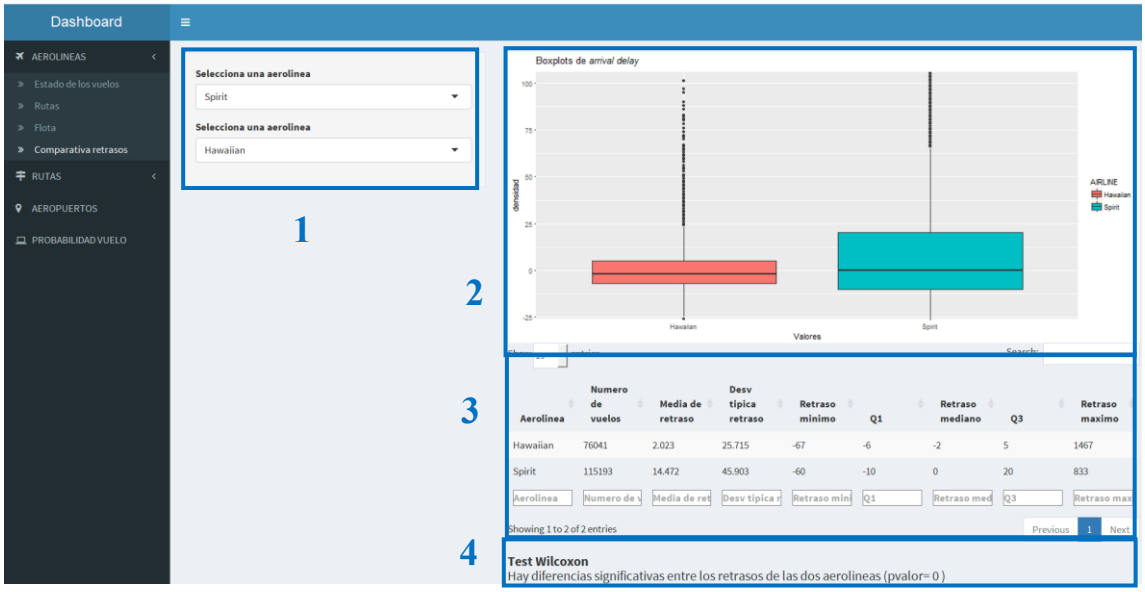


Figura 4.7: Comparativa de retrasos

En el sidebarPanel (1) se elige las dos aerolíneas a comparar.

La aplicación nos permite visualizar la distribución del retraso para las dos aerolíneas (2), así como una tabla descriptiva comparativa del retraso (3). Por último, nos da la conclusión del test de Wilcoxon para ver si hay diferencias estadísticamente significativas en el nivel de retraso entre las dos aerolíneas seleccionadas (al tratarse de comparaciones múltiples 2 a 2, los pvalores están ajustados por la corrección conservadora de Bonferroni) (4).

4.3.2 RUTAS

Esta pestaña consta de dos submenús:

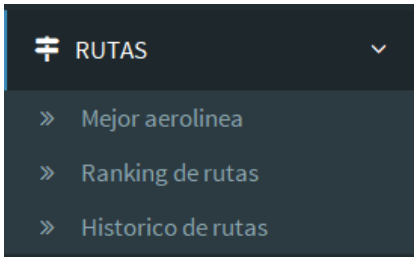


Figura 4.8: Submenús pestaña RUTAS

4.3.2.1 **Mejor aerolínea**

El primer menú nos permite seleccionar la ruta deseada (1) a partir de dos seleccionadores. El seleccionador de destino depende de la elección de la ciudad de origen ya que no todas las combinaciones posibles entre las 322 ciudades son rutas factibles.

A continuación, se visualiza en el mapa la ruta (2) y en forma de tabla, se adjunta el tiempo medio y la variabilidad del tiempo de vuelo de todas las aerolíneas que operan esta ruta (3). Por último, teniendo en cuenta la variabilidad del tiempo de vuelo, la media del tiempo de vuelo y el % de retrasos la aplicación nos dice la mejor aerolínea (4).

La mejor aerolínea para volar cada ruta es la que minimiza la combinación entre media, variabilidad y % de retrasos.

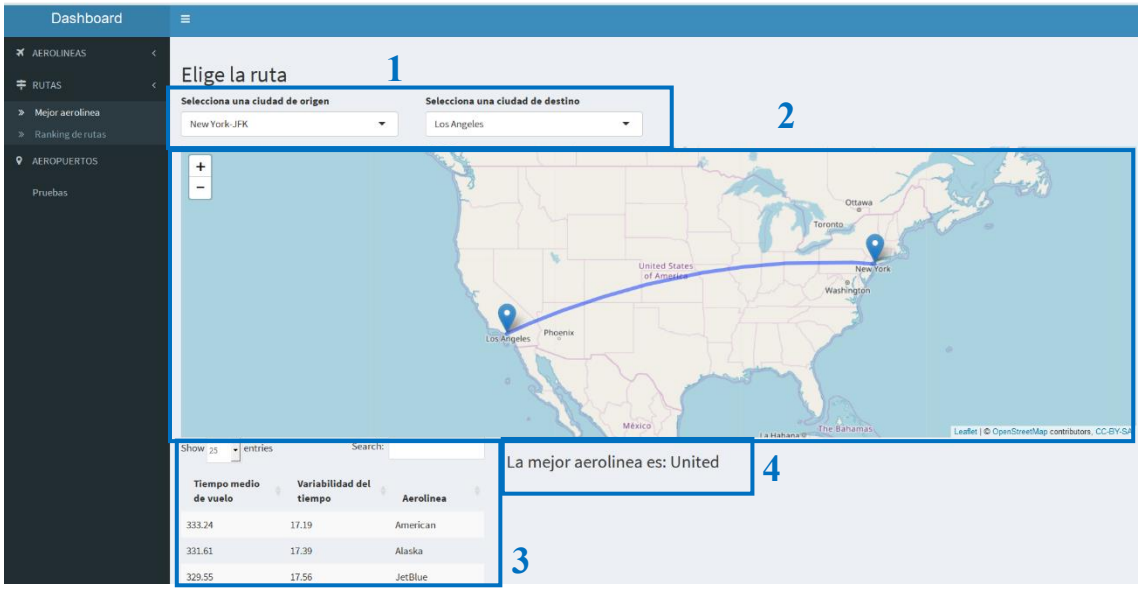


Figura 4.9: Mejor aerolínea para cada ruta

4.3.2.2 **Ranking de rutas**

En el submenú de Ranking de rutas, la aplicación permite elegir en el sidebarPanel (1) el número de rutas a visualizar en orden descendente según dos criterios: por tráfico o por tasa de retraso (calculada como el número de veces que la ruta presenta un retraso >15 entre el total de veces que se voló esa ruta).

Los outputs resultantes son un mapa con las rutas (2) y una tabla resumen de aquella ruta y el valor de tráfico o tasa de retraso correspondiente (3). En este ejemplo, se visualizan 5 rutas por el criterio de tasa de retraso.

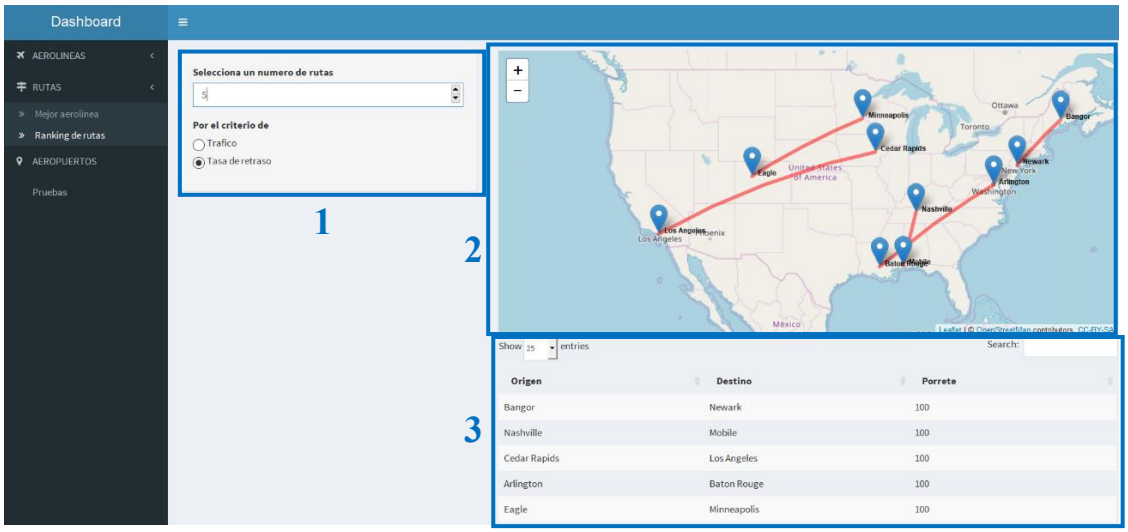


Figura 4.10: Ranking de rutas

4.3.2.3 **Histórico de rutas**

En esta pestaña, el usuario puede elegir 2 opciones en el selector de botones (1):

1) Visualizar en el mapa la ruta elegida para un día determinado

2) Visualizar en el mapa todas las rutas de ese día y de ese estado (2) (Figura)

Para las 2 opciones, la aplicación muestra el mapa (3) donde las rutas se clasifican en 3 grupos: verde (Rutas que en su mayoría llegaron a tiempo), amarillo (rutas que presentaron en su mayoría retrasos) y rojo (rutas que ese día fueron mayormente canceladas).

Para la segunda opción, además se informa del porcentaje de rutas canceladas, retrasadas para ese día.

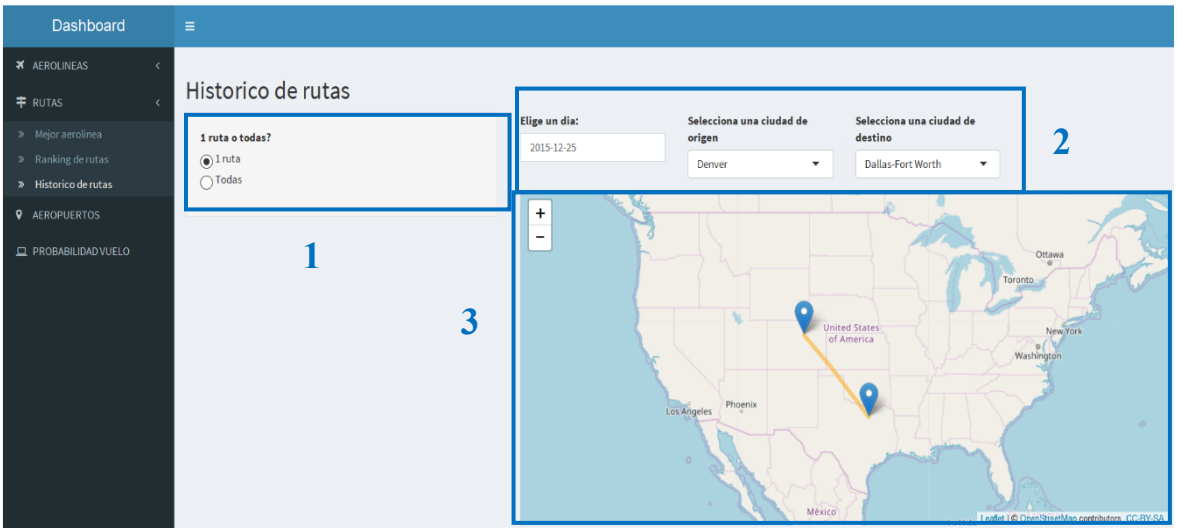


Figura 4.11: Histórico de rutas. Opcion1

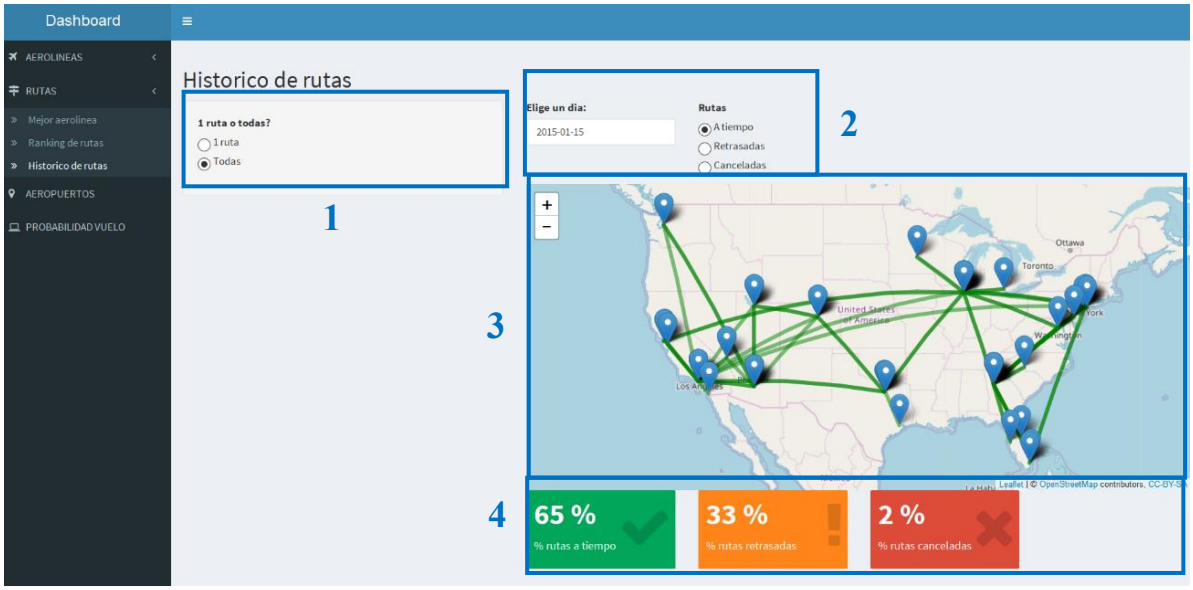


Figura 4.12: Histórico de rutas. Opción 2

4.3.3 Aeropuertos

La aplicación muestra el ranking de aeropuertos según el tráfico o nivel de retraso.

Los aeropuertos con un marcador rojo son los que presentan un retraso mediano por encima del cuartil 75 del retraso global, los de color naranja se sitúan dentro del rango intercuartílico y los de color verde son los que presentan un retraso por debajo del cuartil 25.

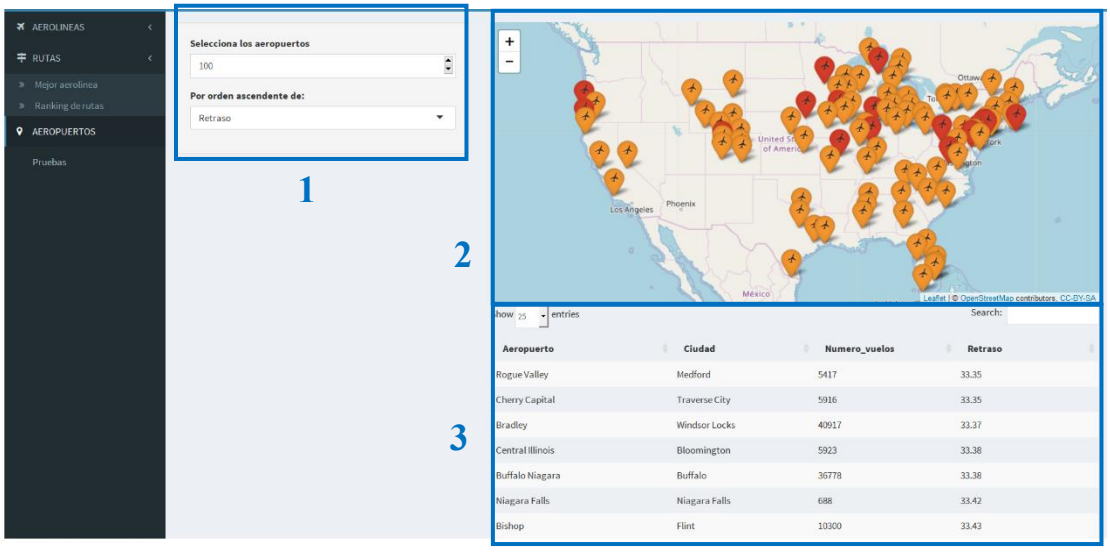


Figura 4.13: Aeropuertos

4.3.4 Probabilidad de retraso de un vuelo

La aplicación ofrece la posibilidad de calcular la probabilidad de que un vuelo presente retraso mayor a 15 minutos en su llegada.

Para ello, el usuario debe introducir en los inputs (1) de la pestaña los datos siguientes: el mes, día de la semana, hora de salida, aerolínea, ruta, características del avión y las condiciones meteorológicas de esa hora.

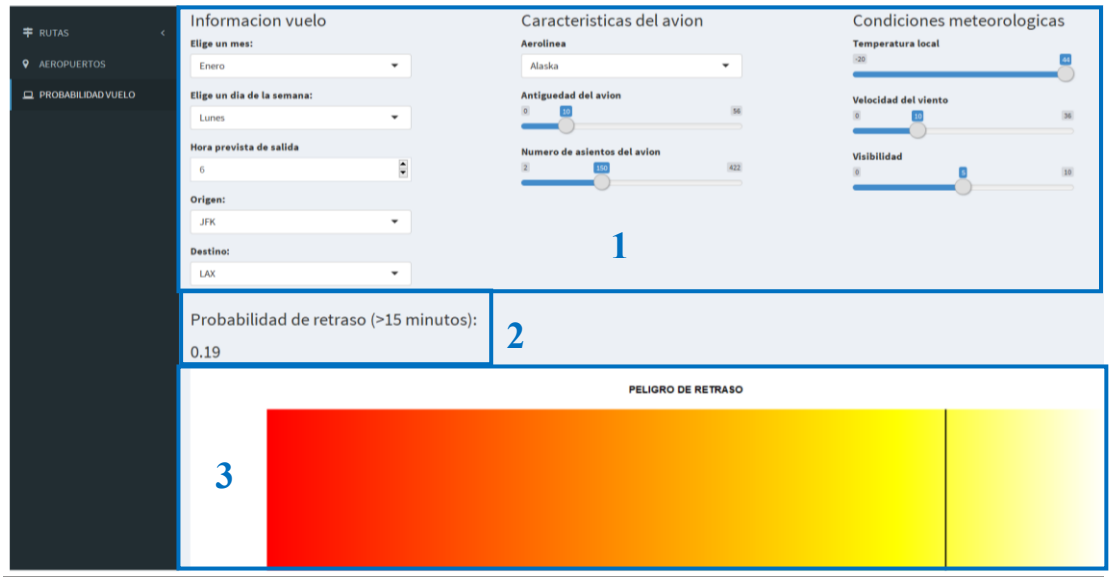


Figura 4.14: Probabilidad de retraso del vuelo

A continuación, la aplicación calcula la probabilidad de retraso a partir del modelo del capítulo 3 (2) y de manera visual se adjunta un gráfico de barras degradadas donde el extremo izquierdo (rojo) simboliza probabilidad máxima de sufrir retraso y el extremo derecho pintado de amarillo probabilidad mínima (como es el caso del ejemplo).

Conclusiones

En este capítulo se resumen las conclusiones más importantes obtenidas a lo largo del presente trabajo. Después se comentan las posibles extensiones del análisis y por último se expone la valoración personal.

Resultados Obtenidos

El presente trabajo tenía como objetivo principal utilizar los conocimientos de estadística y de economía adquiridos para analizar el mercado aéreo doméstico de los Estados Unidos.

En el capítulo 1 se analizó la evolución de la estructura del mercado aéreo de las aerolíneas en estos últimos 15 años a partir del cálculo de indicadores de concentración, de volatilidad y de variables propias de este mercado. La conclusión, una vez analizadas estas 3 fuentes es bastante clara: el mercado ha ido concentrándose cada vez más a favor de las aerolíneas debido en parte a las fusiones y adquisiciones de las aerolíneas regionales o pequeñas por parte de las más grandes.

Esta concentración asociada a un margen de beneficio más amplio para las aerolíneas (favorecido también por el contexto económico de los últimos años: bajada del precio del jet fuel y crecimiento del número de pasajeros) no se ha traducido en un aumento del excedente del consumidor ya que los pasajeros han visto como el precio de los billetes ha ido encareciéndose estos últimos años.

Una vez presentado el mercado aéreo, en el capítulo 2 se analizó todos los vuelos domésticos operados por las aerolíneas para el año 2015.

Las aerolíneas nacionales, a diferencia de las pequeñas, tienden a rellenar más sus tiempos previstos para así mejorar las tasas de puntualidad. A esto se le conoce como efecto relleno y se ha de tener en cuenta para calcular las tasas reales (las oficiales están sesgadas por este efecto).

Existen diferencias significativas del retraso entre aerolíneas siendo la pequeña Spirit la peor y Hawaian la mejor (incluso excluyendo el efecto relleno). Las aerolíneas nacionales se encuentran en la mitad del ranking en cuanto a retrasos.

Los retrasos por culpa de la aerolínea son los más frecuentes pero los más largos son los que se originan por acumulación de retrasos anteriores (el 74,4% de los retrasos pasan antes del despegue). Es por eso por lo que las mejores horas para volar son por la mañana a partir de las 6. Después hay una tendencia creciente del retraso debido a esta acumulación de retrasos.

Además, la tendencia y variabilidad del retraso varía en la semana según el mes en que se vuele: en diciembre existe mucha más variabilidad entre días a diferencia de noviembre que da igual el día a volar.

Del análisis temporal de los vuelos, se concluye por una parte que en los meses de verano y diciembre hay más retrasos y también en las épocas de festivo debido al aumento de demanda de vuelos.

Geográficamente, los aeropuertos del este son los que más retrasos presentan (sobre todo los tres de Nueva York que son los que más vuelos operan).

También quedó confirmado el efecto corriente en chorro (jet stream) que afecta positivamente a la velocidad de las rutas que vuelan en dirección este, pero no influye en las tasas de retraso ya que, en este caso, las aerolíneas tienen en cuenta este efecto a la hora de programar los tiempos de vuelo.

La concentración del mercado explicada en el capítulo 1 analizando los indicadores de concentración también se constata en la distribución de las rutas ya que un 70% de las rutas son operadas por una sola aerolínea.

En el capítulo 3 se construyó un modelo logístico para predecir la probabilidad que un vuelo sufriera retraso. Se realizó un primer modelo con todas las variables recogidas antes del vuelo y con el test Anova se seleccionó las significativas. Se concluyó que el día, mes, aerolínea, ruta, velocidad, duración y las variables meteorológicas, a excepción de la temperatura, son significativas a la hora de predecir el retraso.

Después se estudió todas las interacciones dobles posibles a partir del test de la devianza residual para mejorar la predicción del modelo.

El modelo resultante tenía una tasa de error del 20% y una buena precisión, pero la sensibilidad del modelo era baja concluyendo así que es bastante difícil predecir los vuelos con retraso con la información a priori.

Posibles extensiones del trabajo realizado

Visto el resultado final del trabajo cabe señalar que existen posibles extensiones a partir del presente trabajo:

- Utilizar todos los millones de vuelos para predecir la probabilidad que un vuelo sufra retraso. Con un solo ordenador la limitación de memoria obligó a reducir bastante el número de registros a utilizar en la construcción de la regresión.
- Un análisis más detallado de los pasajeros para analizar en profundidad todas las partes que configuran el mercado aéreo. Sería interesante analizar las opiniones de estos a través de técnicas de sentimiento.

Valoración personal

El mayor reto ha sido compaginar la jornada laboral de 8 horas con la realización del Trabajo de Fin de Grado. Estoy bastante satisfecho con el resultado final por dos motivos:

El primero y que más valoro ha sido que he adquirido nuevos conocimientos: nuevos paquetes de R (tidyverse, leaflet, ggplot) que han mejorado mi destreza a la hora de explotar una base de datos y también he aprendido a realizar aplicaciones interactivas con Shiny.

El segundo motivo es que he podido aplicar los conocimientos adquiridos en mi formación de estadística y economía y extraer conclusiones de un caso real.

Bibliografía

Artículos

Baldwing, J. Y Goreki. O. Measuring the Dynamics of market structure. A: *Annales d'Economie et de Statistique No. 15-16*, 1989. n 5/16.

Ball, Michael; Barnhart, Cinthya: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States. A: *National Center of Excellence for Aviation Operations Research,* 2011.

Gonzalez Martínez, Rolando: Una Aproximación de Indicadores de Concentración y Movilidad Intra-industrial. A: *Munich Personal RePEc Archive, 2008.*

Harlan, Chico: Airlines have never been better at making certain your flight is full. A: *The Washington Post*, 2014.

Helleloid, D; Nam: The U.S. airline industry in 2015. *Journal of the International Academy for Case Studies*,2015 pag 21-25.

Mazzucato, M; Semmler, W: Market share instability and stock price volatility during the industry life-cycle: the US automobile industry. A: *Journal of Evolutionary Economics Volumen 9*, 2002.

Mehmet Yaşar, Kasım Kiracı: Market Share, the Number of Competitors and Concentration: An Empirical Application on the Airline Industry. A: *Anadolu International Conference in Economics.* 2017.

Moss, Diana:Delivering the Benefits? Efficiencies and Airline Mergers. A: *American Antitrust Institute.* 2013.

Phelan, John J (2015): Does industry concentration matter? *A: Journal of Economics and Economic Education Research*. 2015.

Prince, Jeffrey; Simon, Daniel. The impact of Mergers on Quality Provision: Evidence from the Airline Industry. A: *Kelley School of Business Research,* 2015 No.2014-03

Libros

Doganis, Rigas. *The airline bussines in the 21st century,* ed. Routledge, 2000. ISBN 0415208831 pag 20-35.

Grolemund, Garret; Wickham Hadley: *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, ed. O'Rilley, 2007. ISBN 978-1491910399. Versión online: http://r4ds.had.co.nz/

L Hannah, JA Kay: *Concentration in modern industry: Theory, measurement and the UK experience*, ed Macmillan, 1977. ISBN 978-1-349-02773-6, pag 41-63.

Williams, George: *The airline industry and The Impact of Deregulation*, ed. Routledge,1993. ISBN 78-0291398246, Capítulo 1.

Páginas Webs (Última consulta: 28/06/2018)

https://shiny.rstudio.com/tutorial/

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

https://www.faa.gov/data_research/

https://www4.icao.int/newdataplus/

https://www.ncdc.noaa.gov/cdo-web/datasets#LCD

http://airlines.org/data/

Anexo

Salidas de R

Salida R. 1: Summary del fichero antes del preprocessing

```
> summary(flights2)
     MONTH              DAY            DAY_OF_WEEK          AIRLINE        FLIGHT_NUMBER  TAIL_NUMBER
 Julio  : 506588   2      : 191623   Lunes    :845396   Southwest:1258446   Min.   :   1   Length:5684882
 Agosto : 496630   16     : 191445   Martes   :825331   Delta    :1245545   1st Qu.: 715   Class :character
 Marzo  : 494752   20     : 191166   Miercoles:836542   American :1233609   Median :1642   Mode  :character
 Junio  : 488018   13     : 190639   Jueves   :852356   United   :1150156   Mean   :2134
 Mayo   : 482834   9      : 189572   Viernes  :842743   JetBlue  : 266318   3rd Qu.:3217
 Octubre: 475189   8      : 189400   Sabado   :683370   Alaska   : 246720   Max.   :6948
 (Other):2740871   (Other):4541037   Domingo  :799144   (Other)  : 284088
 SCHEDULED_DEPARTURE DEPARTURE_DELAY      TAXI_OUT       WHEELS_OFF    SCHEDULED_TIME   ELAPSED_TIME
 Min.   :   1        Min.   : -82.00   Min.   :  1.00   Min.   :   1   Min.   : 18.0   Min.   : 14.0
 1st Qu.: 916        1st Qu.:  -5.00   1st Qu.: 11.00   1st Qu.: 935   1st Qu.: 87.0   1st Qu.: 83.0
 Median :1325        Median :  -2.00   Median : 14.00   Median :1343   Median :124.0   Median :120.0
 Mean   :1330        Mean   :   9.39   Mean   : 16.06   Mean   :1357   Mean   :142.8   Mean   :138.1
 3rd Qu.:1730        3rd Qu.:   7.00   3rd Qu.: 19.00   3rd Qu.:1754   3rd Qu.:175.0   3rd Qu.:170.0
 Max.   :2359        Max.   :1988.00   Max.   :225.00   Max.   :2400   Max.   :718.0   Max.   :766.0
                     NA's   :84291     NA's   :87109    NA's   :87109  NA's   :6       NA's   :87861
    AIR_TIME        DISTANCE        WHEELS_ON        TAXI_IN       SCHEDULED_ARRIVAL ARRIVAL_DELAY
 Min.   :  7.0   Min.   :   1.0   Min.   :   1   Min.   :  1.00   Min.   :   1      Min.   : -87.0
 1st Qu.: 61.0   1st Qu.: 386.0   1st Qu.:1054   1st Qu.:  4.00   1st Qu.:1110      1st Qu.: -13.0
 Median : 95.0   Median : 666.0   Median :1509   Median :  6.00   Median :1520      Median :  -5.0
 Mean   :114.6   Mean   : 831.4   Mean   :1472   Mean   :  7.42   Mean   :1495      Mean   :   4.4
 3rd Qu.:145.0   3rd Qu.:1069.0   3rd Qu.:1912   3rd Qu.:  9.00   3rd Qu.:1919      3rd Qu.:   8.0
 Max.   :690.0   Max.   :4983.0   Max.   :2400   Max.   :248.00   Max.   :2400      Max.   :1971.0
 NA's   :87861                    NA's   :87861  NA's   :87861                      NA's   :87861
    DIVERTED     CANCELLED              CANCELLATION_REASON  AIR_SYSTEM_DELAY   SECURITY_DELAY
 Min.   :0e+00   0:5597022   No                    :5597022   Min.   :   0.000   Min.   :  0.000
 1st Qu.:0e+00   1:  87860   Aerolinea             :  23824   1st Qu.:   0.000   1st Qu.:  0.000
 Median :0e+00               Mal Tiempo            :  48408   Median :   0.000   Median :  0.000
 Mean   :2e-07               Sistema Aereo Nacional:  15606   Mean   :   2.488   Mean   :  0.014
 3rd Qu.:0e+00               Seguridad             :     22   3rd Qu.:   0.000   3rd Qu.:  0.000
 Max.   :1e+00                                              Max.   :1134.000   Max.   :573.000

 AIRLINE_DELAY      LATE_AIRCRAFT_DELAY WEATHER_DELAY          ORIGIN          DESTINATION          speed
 Min.   :   0.000   Min.   :   0.000   Min.   :   0.000   ATL     : 378539   ATL     : 378366   Min.   :  52.05
 1st Qu.:   0.000   1st Qu.:   0.000   1st Qu.:   0.000   ORD     : 310379   ORD     : 310297   1st Qu.: 592.73
 Median :   0.000   Median :   0.000   Median :   0.000   DFW     : 239436   DFW     : 238908   Median : 670.56
 Mean   :   3.465   Mean   :   4.295   Mean   :   0.541   DEN     : 213551   DEN     : 213327   Mean   : 659.78
 3rd Qu.:   0.000   3rd Qu.:   0.000   3rd Qu.:   0.000   LAX     : 203144   LAX     : 203376   3rd Qu.: 734.95
 Max.   :1971.000   Max.   :1331.000   Max.   :1211.000   SFO     : 161666   SFO     : 161753   Max.   :1267.95
                                                          (Other):4178167   (Other):4178855   NA's   :87861
```

Salida R.2: Comparaciones múltiples entre meses

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  flights.2$ARRIVAL_DELAY and flights.2$MONTH

           Enero   Febrero Marzo   Abril   Mayo    Junio   Julio   Agosto  Septiembre Octubre Noviembre
Febrero    0.00536 -       -       -       -       -       -       -       -          -       -
Marzo      1.00000 7.5e-05 -       -       -       -       -       -       -          -       -
Abril      0.00142 1.6e-14 0.02727 -       -       -       -       -       -          -       -
Mayo       9.8e-07 < 2e-16 2.9e-05 1.00000 -       -       -       -       -          -       -
Junio      2.4e-07 1.00000 3.3e-10 < 2e-16 < 2e-16 -       -       -       -          -       -
Julio      1.00000 0.03760 1.00000 3.6e-05 4.5e-09 3.0e-06 -       -       -          -       -
Agosto     0.06151 2.3e-11 0.61018 1.00000 0.86624 < 2e-16 0.00279 -       -          -       -
Septiembre < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 -          -       -
Octubre    < 2e-16 < 2e-16 < 2e-16 < 2e-16 5.8e-15 < 2e-16 < 2e-16 < 2e-16 1.00000    -       -
Noviembre  < 2e-16 < 2e-16 < 2e-16 < 2e-16 5.9e-12 < 2e-16 < 2e-16 < 2e-16 1.00000    1.00000 -
Diciembre  0.00062 2.0e-14 0.01101 1.00000 1.00000 < 2e-16 7.9e-06 1.00000 < 2e-16    < 2e-16 1.5e-14

P value adjustment method: bonferroni
```

Salida R.3: Comparaciones múltiples entre días de la semana

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  flights.2$ARRIVAL_DELAY and flights.2$DAY_OF_WEEK

          Lunes   Martes  Miercoles Jueves  Viernes Sabado
Martes    0.0257  -       -         -       -       -
Miercoles 0.1123  1.0000  -         -       -       -
Jueves    0.0053  8.4e-11 1.4e-09   -       -       -
Viernes   0.2073  1.0e-07 1.2e-06   1.0000  -       -
Sabado    < 2e-16 2.5e-07 1.2e-08   < 2e-16 < 2e-16 -
Domingo   0.0138  1.0000  1.0000    5.3e-11 6.1e-08 1.6e-06

P value adjustment method: bonferroni
```

Salida R.4: Resultado test Wilcoxon e intervalo de confianza para FESTIVO

```
> wilcox.test(ARRIVAL_DELAY ~ festivo,data=flights3 ,conf.int=T)

        Wilcoxon rank sum test with continuity correction

data:  ARRIVAL_DELAY by festivo
W = 8781400, p-value = 0.001088
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.9999533 2.0000334
sample estimates:
difference in location
             1.000038
```

Salida R.5: Comparaciones múltiples entre aerolíneas

```
        Pairwise comparisons using Wilcoxon rank sum test

data:  flights.2$ARRIVAL_DELAY and flights.2$AIRLINE

          American Alaska  JetBlue Delta   Frontier Hawaiian Spirit  United
Alaska    1.00000  -       -       -       -        -        -       -
JetBlue   1.00000  1.00000 -       -       -        -        -       -
Delta     2.1e-11  0.00089 2.0e-06 -       -        -        -       -
Frontier  < 2e-16  < 2e-16 3.3e-11 < 2e-16 -        -        -       -
Hawaiian  1.0e-15  3.4e-14 4.6e-09 < 2e-16 1.00000  -        -       -
Spirit    < 2e-16  < 2e-16 < 2e-16 < 2e-16 0.72215  0.00256  -       -
United    3.1e-05  0.01493 1.00000 < 2e-16 1.3e-11  4.0e-09  < 2e-16 -
Southwest < 2e-16  7.1e-12 1.9e-06 < 2e-16 2.3e-06  4.6e-05  < 2e-16 7.0e-09

P value adjustment method: bonferroni
```

Salida R.6: Resultado test lack of fit

	Test stat	Pr(>|t|)
hora_prev_salida	NA	NA
speed	311.190	0.000
DESTINATION	NA	NA
MONTH	NA	NA
AIR_TIME	0.000	1.000
DAY_OF_WEEK	NA	NA
HOURLYWindSpeed	0.000	1.000
AIRLINE	NA	NA
HOURLYRelativeHumidity	0.699	0.403
HOURLYPrecip	0.000	1.000
festivo	NA	NA
HOURLYVISIBILITY	0.894	0.345
HOURLYStationPressure	0.234	0.629
TYPE.ENG	NA	NA
NO.SEATS	1.744	0.187
I(HOURLYWindSpeed^2)	0.895	0.344

Tablas

Tabla 1: Frecuencia de vuelos por día de la semana

Dia Semana	**Lu**	**Ma**	**Mi**	**Ju**	**Vi**	**Sa**	**Do**
Frecuencia	845396	825331	836542	852356	842743	683370	799144

Tabla 2: Matriz de confusión e indicadores del modelo m1dep

		Predicción	
		No Retraso	Retraso
Observado	No Retraso	21955	389
	Retraso	5771	692

	Test set
Sensibilidad	10,7%
Precisión	64,5%
Ratio de error	20,1%

Gráficos

Figura 1: Proyección individuos

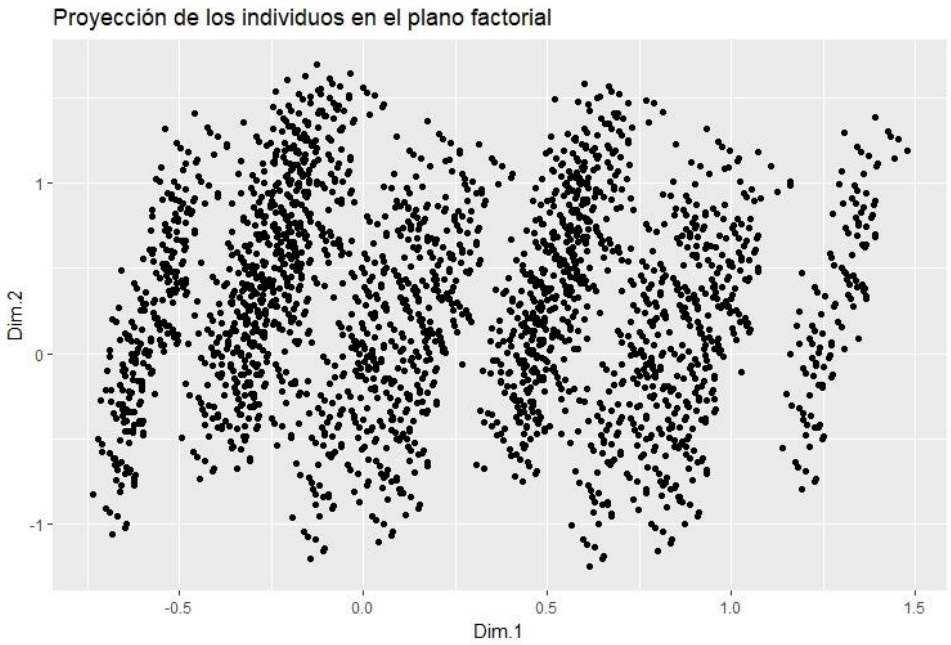


Figura 2: Proyección variables quantitativas suplementarias