Introducción

En la era moderna, los vuelos comerciales se han convertido en un pilar fundamental de la sociedad, permitiendo a las personas explorar destinos lejanos y conectando a empresas en todo el mundo. Estos vuelos desempeñan un papel crucial en la economía global y son indispensables para el crecimiento y desarrollo de diversas industrias. Se tiene registros de que, durante el año 2019, hubo más de 4,500 millones de vuelos comerciales en todo el mundo, y se espera que esta cifra siga aumentando en los próximos años. Un ejemplo destacado es el Aeropuerto Internacional Benito Juárez de la Ciudad de México, que en 2017 atendió a un total de 44.7 millones de pasajeros, según datos del Grupo Aeroportuario de la Ciudad de México (GACM).

Sin embargo, a pesar de la importancia de los vuelos comerciales, los retrasos son una realidad frecuente en la industria. Según la Oficina de Estadísticas de Transporte de Estados Unidos, más del 20 % de los vuelos comerciales en ese país experimentaron retrasos en 2017. Esta situación tiene un impacto significativo tanto para los pasajeros como para las aerolíneas, los aeropuertos y las empresas relacionadas, además ed que es innegable que representa un costo sustancial para la economía. De acuerdo con un estudio de la Universidad de California, los retrasos en los vuelos domésticos en los Estados Unidos generaron pérdidas de 32.9 mil millones de dólares en 2007, dentro de los cuales están incluídos los gastos por la productividad perdida, el aumento en el consumo de combustible y los gastos adicionales tanto para las aerolíneas como para los pasajeros.

Además de los costos directos, los vuelos retrasados también tienen consecuencias indirectas. Por ejemplo, las cadenas de suministro pueden verse interrumpidas, ocasionando pérdidas en ventas y productividad. Asimismo, los retrasos pueden dificultar que las empresas cumplan con sus plazos y objetivos, lo que se traduce en pérdidas financieras. Ante esta problemática, es crucial implementar estrategias que permitan predecir y prevenir los retrasos en los vuelos comerciales. Esto no solo mejoraría la confiabilidad y eficiencia de los viajes aéreos, sino que también brindaría beneficios tanto a los pasajeros como a las aerolíneas y los aeropuertos.

Descripción de los datos

Los datos iniciales para desarrollar este proyecto fueron proporcionados por una aerolínea en el año 2021 a uno de los autores de este trabajo. Los datos consisten de un conjunto de registros de vuelos nacionales e internacionales durante los años 2017 y 2018 de un aeropuerto latinoamericano, destacando una presencia mayoritaria de vuelos de la aerolínea que proporcionó la información. Por motivos de confidencialidad, se omitieron todos los registros asociados con dicha aerolínea, por lo que los datos que se usan a lo largo de este proyecto asumen la presencia de registros de vuelos de otras aerolíneas durante los mismos años de operación para el mismo aeropuerto.

A continuación se presenta una muestra de los datos con los que se cuenta:

Table 1: Muestra de los datos 1

Fecha_I	Hora_I	Vlo-I	Des-I	Emp-I	Fecha_O	Hora_O	Vlo-O	Des-O	Emp-O
2017-07-16	6:35:00	501	SAEZ	SKU	2017-07-16	6:34:00	501	SAEZ	SKU
2017-11-02	5:00:00	26	SCFA	JAT	2017-11-02	4:55:00	26	SCFA	JAT
2017-02-22	15:10:00	800	SPJC	SKU	2017-02-22	15:40:00	800	SPJC	SKU
2017-10-08	13:00:00	507	SAEZ	SKU	2017-10-08	13:11:00	507	SAEZ	SKU
2017-06-08	2:12:00	174	MPTO	CMP	2017-06-08	2:33:00	174	MPTO	CMP
2017-08-11	6:00:00	31	SCTE	LAW	2017-08-11	6:02:00	31	SCTE	JMR
2017-04-19	10:45:00	73	SCIE	SKU	2017-04-19	10:55:00	73	SCIE	SKU
2017-01-23	1:33:00	240	SKBO	AVA	2017-01-23	1:19:00	240	SKBO	AVA

Table 2: Muestra de los datos 2

DIA	MES	AÑO	DIANOM	TIPOVUELO	OPERA	SIGLADES
16	7	2017	Domingo	I	Sky Airline	Buenos Aires
2	11	2017	Jueves	N	JetSmart SPA	Antofagasta
22	2	2017	Miercoles	I	Sky Airline	Lima
8	10	2017	Domingo	I	Sky Airline	Buenos Aires
8	6	2017	Jueves	I	Copa Air	Ciudad de Panama
11	8	2017	Viernes	N	Latin American Wings	Puerto Montt
19	4	2017	Miercoles	N	Sky Airline	Concepcion
23	1	2017	Lunes	I	Avianca	Bogota

Como puede observarse, la información con la que se cuenta puede agruparse en tres tipos: información sobre la programación del vuelo, información sobre la operación del vuelo, e información adicional. La principal diferencia entre las primeras dos es que la información sobre la programación del vuelo corresponde a los datos con los que el vuelo es anunciado ante los clientes, mientras que la segunda son los datos con los que el vuelo fue registrado para su funcionamiento y seguimiento antes del despegue.

Respecto a la información sobre la programación del vuelo, se cuenta con la fecha y hora en la que cada vuelo fue programado (**Fecha_I** y **Hora_I**), el número de vuelo (**Vlo-I**), las siglas de la ciudad de destino (**Des-I**), y con las siglas de la aerolínea (**Emp-I**). Por otro lado, respecto a la información sobre la operación del vuelo, se cuenta con la fecha y hora en la que cada vuelo realizó el despegue (**Fecha-O** y **Hora_O**), el número de vuelo con el que operó (**Vlo-O**), las siglas de la ciudad destino (**Des-O**), y las siglas de la aerolínea que llevó a cabo el vuelo (**Emp-O**). Por último, se cuenta con información adicional sobre el **Día**, **Mes** y **Año** (en formato numérico) en los que el vuelo se llevó a cabo, el día de la semana nominal (**DIANOM**), el tipo de vuelo que sirve para diferenciar vuelos nacionales de los internacionales (**TIPOVUELO**), y el nombre de la ciudad de destino (**SIGLADES**). Es importante recordar que todos los vuelos de los que se tiene registro corresponden al mismo aeropuerto de origen.

Considerando la descripción anterior, se procede a presentar un análisis exploratorio de la información para realizar su respectivo tratamiento y limpieza. Gracias al resumen presentado en la siguiente tabla, podemos observar el número de vuelos con los que se cuenta después de realizar la anonimización de la aerolínea emisora de la información.

Table 3: Número de vuelos por año

Año	Número de vuelos	Porcentaje
2017	27300	100

Como se describió al inicio de esta sección, contamos con información sobre el número identificador del vuelo con el que se anunció comercialmente y con el que se llevó a cabo la operación, por lo tanto resulta interesante conocer cuántos cambios de este número existieron durante todo el año.

Table 4: Cantidad de cambios de número de vuelo

¿Cambió?	Número de vuelos	Porcentaje
Cambió	33	0.1209
No cambió	27267	99.8791

Gracias a la tabla anterior se puede observar que la proporción de vuelos que registraron un cambio en su número de vuelo corresponde a una mínima parte del total de todos los datos, por lo que podemos prescindir de elos y trabajar únicamente con los vuelos que no presentaron cambios en sus números. Esta decisión nos ayudará posteriormente en el modelado de la solución, ya que, al asumir que todos los vuelos tuvieron el mismo número de programación y de operación, se están eliminando dos columnas de información, lo que ayuda a reducir la complejidad del problema.

También se pueden proporcionar elementos visuales que ayuden a entender el problema con el que se cuenta, además de comprender la distribución de los datos para ciertas variables de interés.

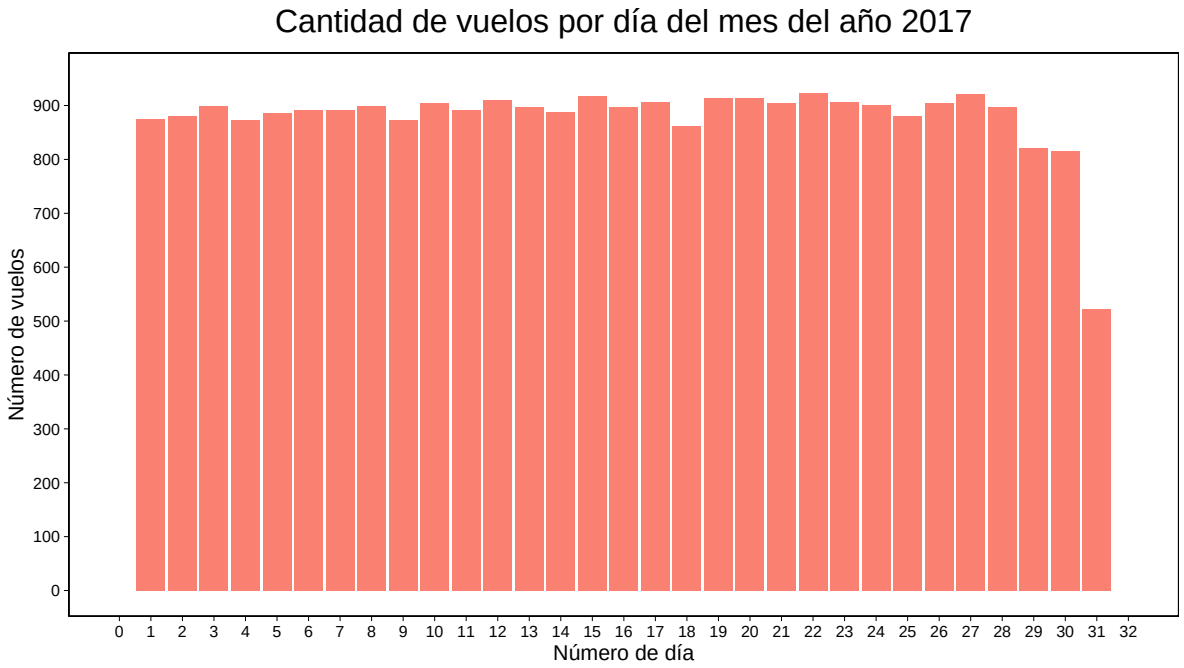

Figure 1: Distribución de la cantidad de vuelos internacionales por día ordinal

De la gráfica anterior se puede observar que la distribución del número de vuelos por día del mes es bastante similar para todos, con excepción de los últimos dos días. Este comportamiento es el esperado si consideramos que no todos los meses cuentan con 31 días, por lo que se entiende que es normal que exista un desbalance para el último día de la gráfica. Por otro lado, el comportamiento para los demás días es bastante similar, ya que cantidad de vuelos para cada uno de ellos se encuentra en el intervalo de 800 a 900, sin embargo no existe una varianza en la cantidad de vuelos que sugiera que pudiera ser una variable de interés para poder predecir un posible retraso.

De forma análoga, podemos proporcionar un análisis sobre la distribución del número de vuelos por día de la semana con la intención de encontrar algún patrón en la información, o en su defecto, algún comportamiento para ciertos días específicos que nos permitan considerar los días como una variable importante a la hora de predecir retrasos para vuelos.

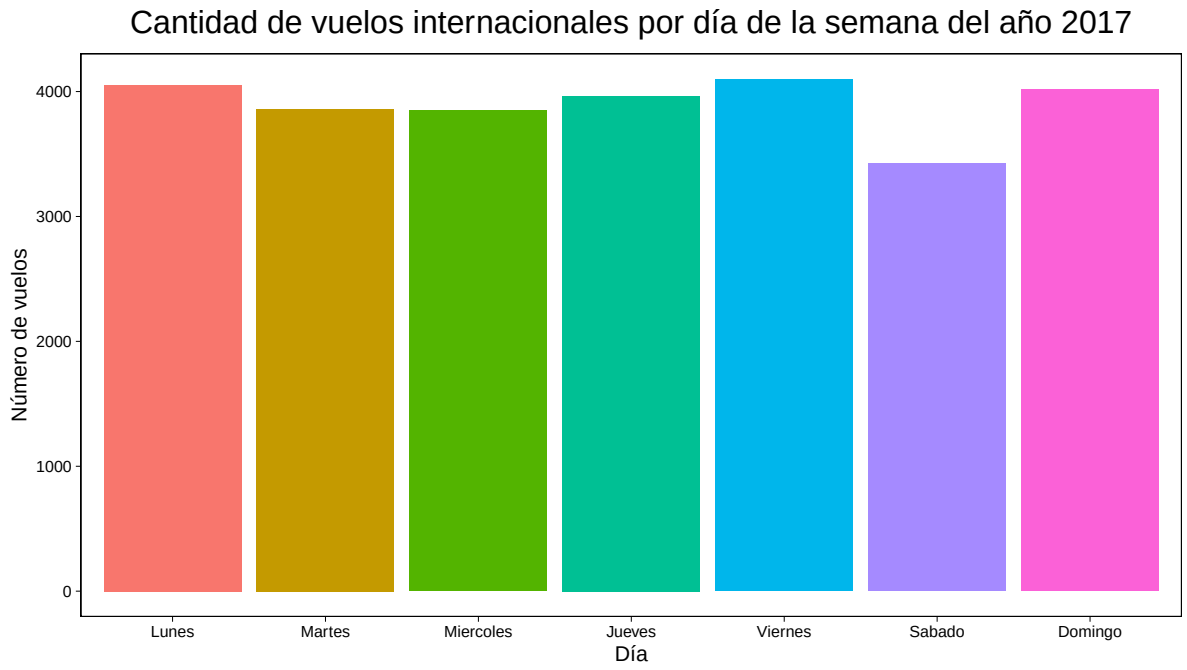

Figure 2: Distribución de la cantidad de vuelos por día de la semana

Como puede observarse de la gráfica anterior, la distribución para todos los días es bastante similar con excepción del día sábado. En general se puede deducir que existe un ligero aumento de afluencia de vuelos para los días lunes y viernes, mientras que los días martes y miércoles son los días que presentan un ligero decremento. Esperaríamos tener una mayor claridad sobre qué días hay más vuelos en el año, ya que pudiera ser un indicio de que ante mayor afluencia de vuelos la probabilidad de retraso aumentara, sin embargo este comportamiento no es claro por sí solo en la gráfica anterior. A pesar de ello, debido al comportamiento de la afluencia de vuelos para el día sábado y el contexto del problema, esta variable pudiera ser de ayuda para predecir el atraso de vuelos, por lo que será considerada.

Otra gráfica que resulta pertinente visualizar es la distribución de vuelos por mes, buscando capturar temporadas altas y bajas, vacaciones, o algún otro patrón de interés. Estos comportamientos se muestran satisfactoriamente en la siguiente gráfica, donde la distribución de vuelos es la esperada considerando el contexto del problema. En general puede observarse que existe una mayor afluencia de vuelos durante los meses de octubre a diciembre, lo que es consistente con las festividades mundiales y los fenómenos denominados “temporadas altas”. Por otro lado, se puede observar que los meses donde hay menor afluencia de vuelos es durante los meses de febrero a junio, lo cual también es consistente con los meses donde se les denomina de “temporadas bajas”. Podemos concluir que esta variable sí es de interés para nuestro análisis, ya que ofrece bastante información que puede ser utilizada durante nuestro modelado.

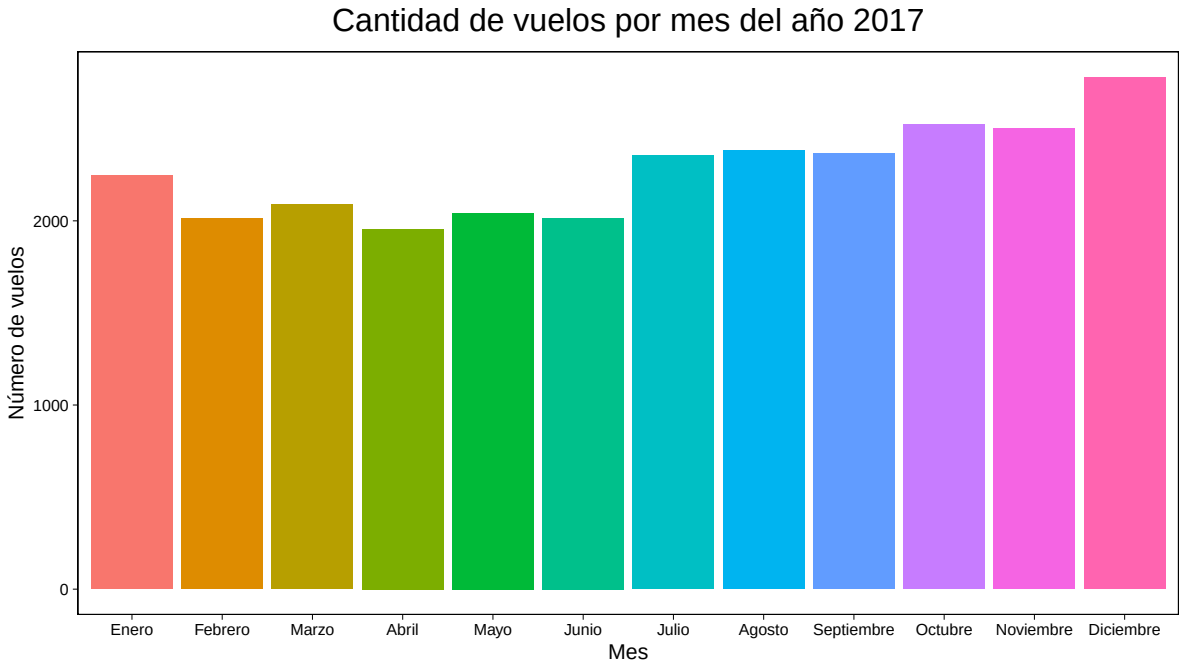

Figure 3: Distribución de la cantidad de vuelos por mes

También resulta pertinente analizar la proporción de vuelos nacionales e internacionales que conforman los datos disponibles. Gracias a la siguiente tabla que detalla dicho análisis, podemos concluir que la proporción de vuelos nacionales e internacionales es bastante similar, lo cual no nos proporciona suficiente información para poder considerarla como una variable de interés al momento de predecir los retrasos en vuelos. Sin embargo, un problema particular a resolver sería la predicción de restrasos en vuelos internacionales, ya que, de acuerdo con la experiencia, este tipo de vuelos son los que pudieran ser más caóticos y podrían incurrir en una mayor cantidad de pérdidas monetarias para las aerolíneas. De esta manera, los siguientes análisis y procedimientos subsecuentes consideran únicamente la presencia de vuelos internacionales, para los cuales también resulta pertinente obtener un nuevo resumen estadístico y visual.

Table 5: Número de vuelos internacionales

TIPOVUELO	Número de vuelos	Proporción
Internacional	14054	0.5154
Nacional	13213	0.4846

Análisis de vuelos internacionales

Table 6: Número de vuelos internacionales

TIPOVUELO	Número de vuelos
Internacional	14054

A pesar de haber realizado un recorte del total de información disponible, aún se cuenta con bastantes registros para realizar la implemetación de un modelo para predicción de retrasos para vuelos internacionales.

Sin embargo, es necesario conocer las distribuciones de estos vuelos para verificar si los patrones anteriormente vistos siguen siendo los mismos, o en su defecto identificar nuevos patrones de interés.

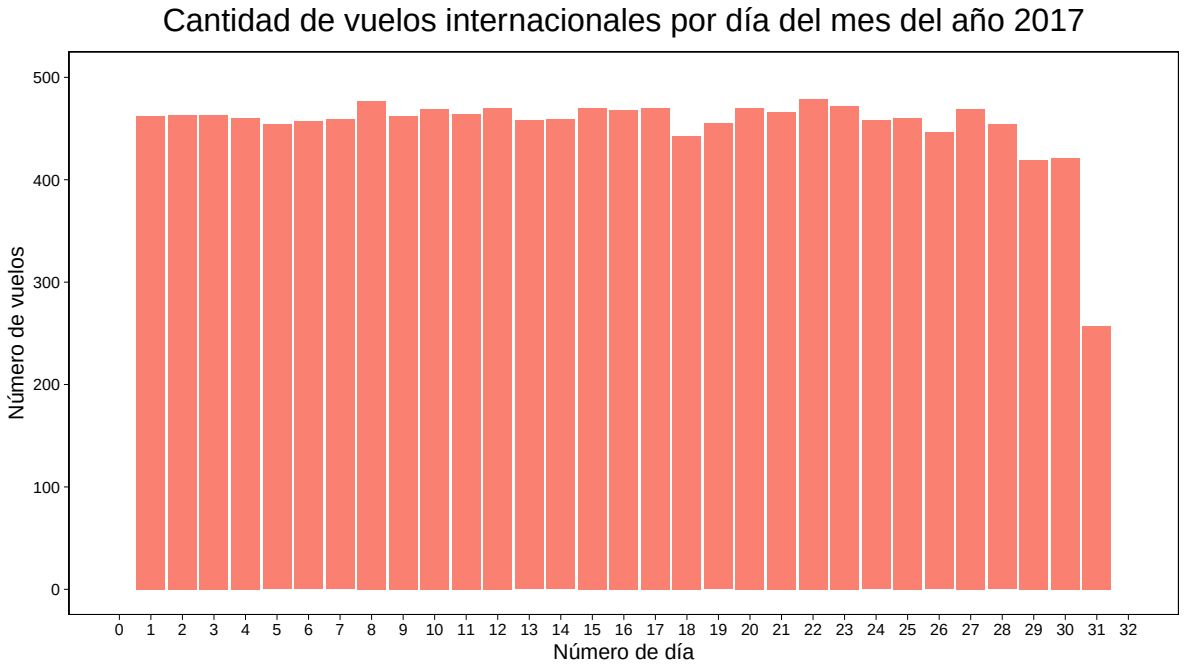

Figure 4: Distribución de la cantidad de vuelos internacionales por día ordinal

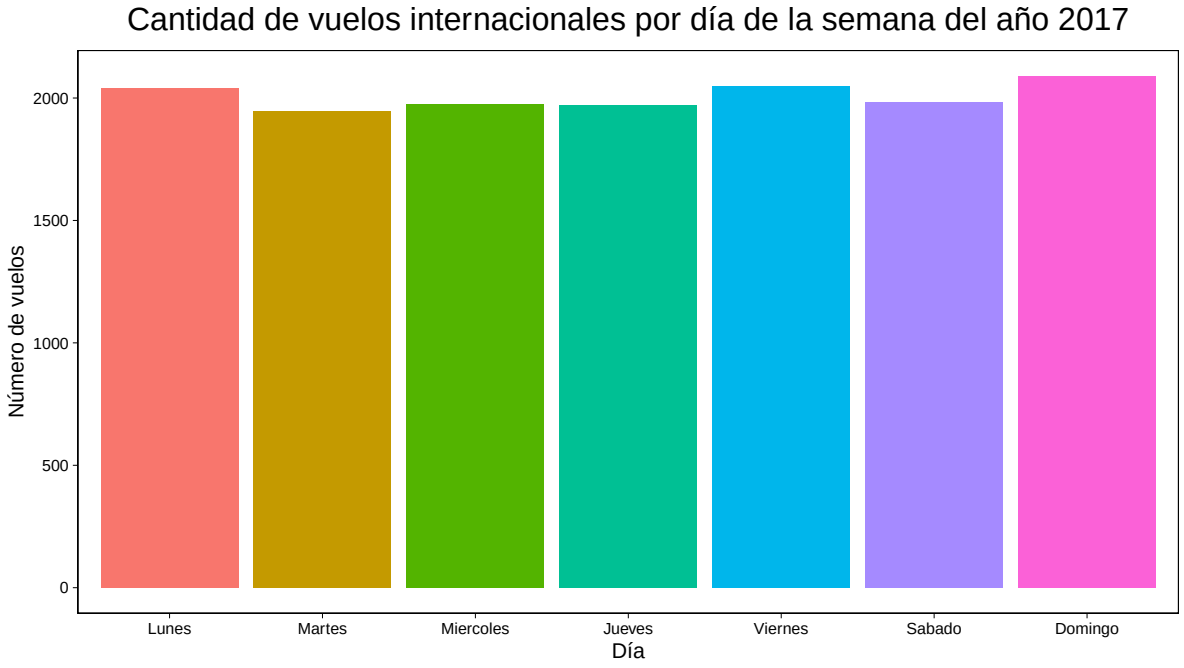

Figure 5: Distribución de la cantidad de vuelos internacionales por día de la semana

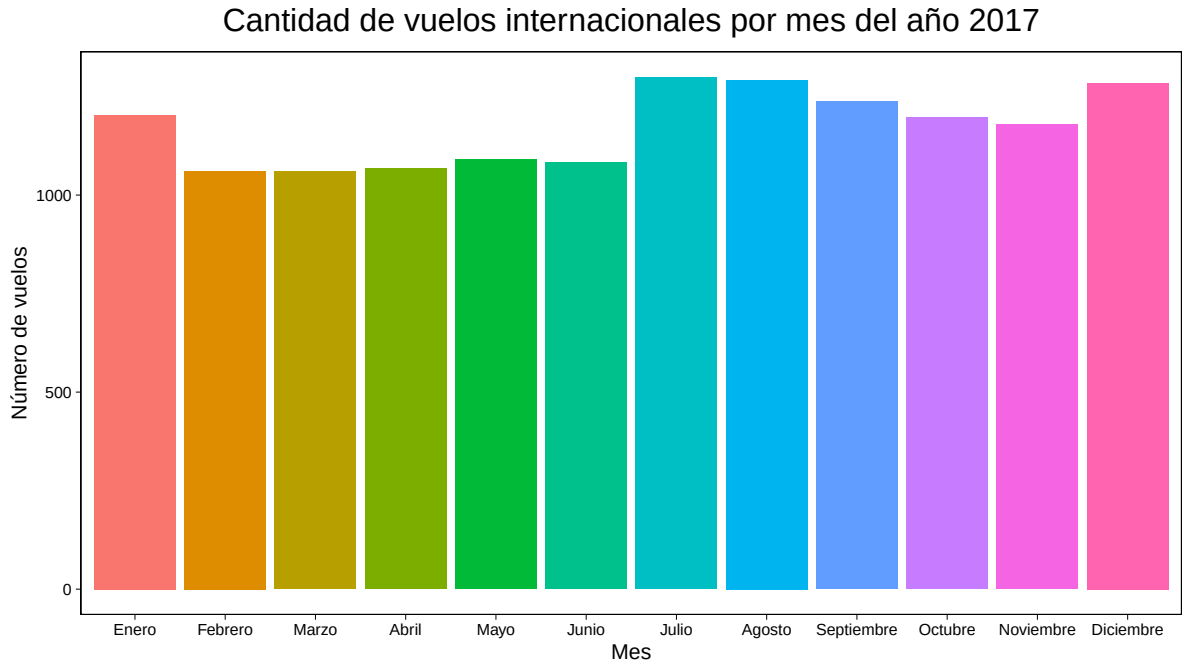

Figure 6: Distribución de la cantidad de vuelos internacionales por mes

Afortunadamente se puede concluir que un análisis para vuelos internacionales resulta más conveniente, ya que en las gráficas anteriores puede observarse que los patrones iniciales de la información se mantuvieron, además de que se logró capturar nuevos patrones que son de mayor relevancia. Respecto a la distribución de vuelos por día del mes, se mantiene el mismo comportamiento en la distribución de los datos, denotando que la cantidad de vuelos para los últimos dias de cada mes es menor en comparación con el resto. Respecto a la distribución de vuelos por día de la semana, la distribución se mantuvo bastante similar, además de que el desbalance que había para el día sabado se corrigió. Por último, respecto a la distribución de vuelos por mes, se puede observar que se captura de mejor manera las temporadas altas y bajas para los meses anteriormente analizados, además de que en esta ocasión los meses de junio y julio presentan una alta afluencia, lo cual es completamente entendible considerando las épocas de vacaciones de verano.

Distribución de vuelos por aerolínea

Un último análisis que resulta importante considerar es la distribución de vuelos por aerolínea, de modo que se pueda entender qué aerolíneas son las que tienen mayor presencia de vuelos internacionales y cuáles son susceptibles a algún retraso en sus operaciones. Esta información puede observarse en la siguiente gráfica, donde se muestra el número total de registros de múltiples aerolíneas sin embargo se puede destacar que la mayor cantidad de registros se concentra entre las aerolíneas *Aerolíneas Argentinas*, *Copa Air* y *Sky Airline*, lo cual podría contaminar el modelo a implementar.

Teniendo en cuenta esta distribución, se tomó la decisión de eliminar aquellos registros de vuelos que fueron operados por las aerolíneas con mayor presencia, de modo que se mantuvieran todas aquellas aerolíneas con una proporción menor de registros en la base de datos.

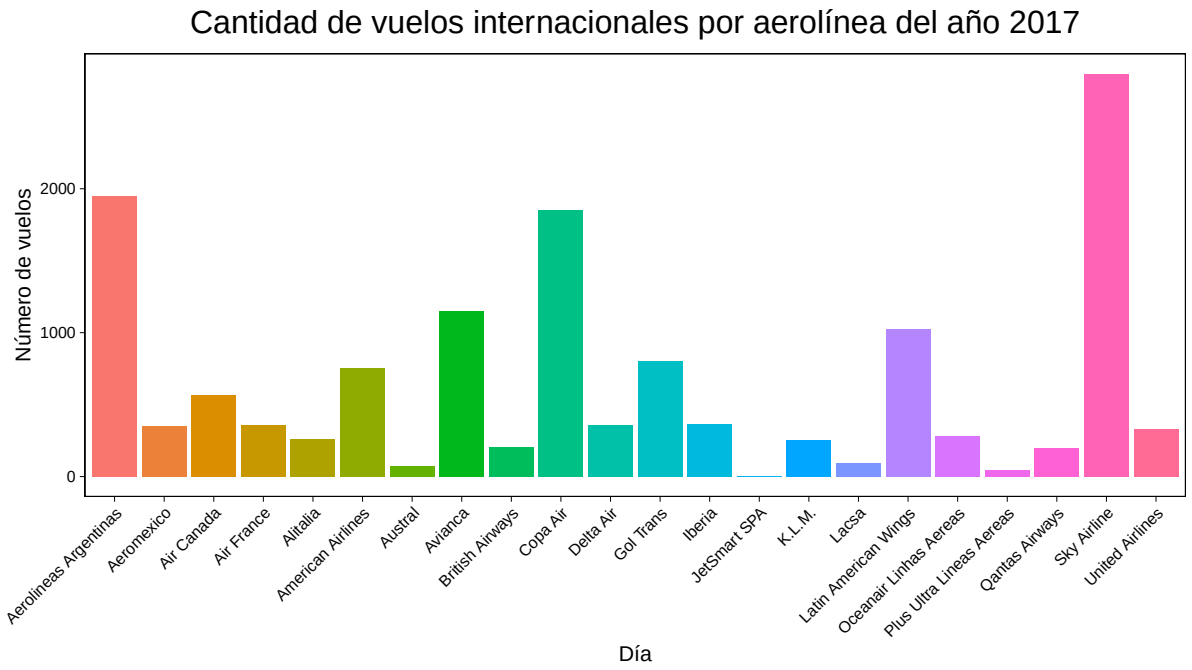

Figure 7: Distribución de la cantidad de vuelos internacionales por aerolínea

Ingeniería de variables

Después del análisis exploratorio de los datos y de las variables con las que se cuenta, un paso adicional para favorecer el modelado de la solución es la creación de variables auxiliares y de nuestra variable objetivo. Para este procedimiento es necesario considerar el contexto del problema a resolver y la información con la que se cuenta, por lo tanto en las subsecciones siguiente se incluye una justificación de cada una de las variables creadas y el método para crearlas.

Temporada alta

Como pudo observarse en el análisis exploratorio, existe una tendencia muy clara sobre los meses donde la afluencia de vuelos es mayor y donde no lo es. Esto puede traducirse en temporadas altas y bajas para las aerolíneas, y la experiencia sugiere que ante una mayor afluencia de vuelos y de personas, la probabilidad de un retraso incrementa. Por lo tanto, la primer variable auxiliar creada fue la variable **temporada_alta**, la cual fue creada a partir de algunos casos particulares que la aerolínea original proporcionó de acuerdo con la experiencia.

- Si la fecha del vuelo se encuentra entre el 1 de enero y el 3 de marzo, es **temporada alta**.
- Si la fecha del vuelo se encuentra entre el 15 de julio y el 30 de julio, es **temporada alta**.
- Si la fecha del vuelo se encuentra entre el 11 de septiembre y el 30 de septiembre, es **temporada alta**.
- Si la fecha del vuelo se encuentra entre el 15 de diciembre y el 31 de diciembre, es **temporada alta**.
- Cualquier otro caso no considerado, es **temporada baja**.

Un resumen de la distribución de vuelos después de la creación de esta variable se muestra en la siguiente tabla.

Table 7: Resumen del tipo de temporada de los vuelos

Temporada	Número de vuelos
Baja	5026
Alta	2432

Diferencia en minutos

Para definir si un vuelo se atrasó o no se debe considerar la fecha y hora con las que se anunció a los clientes y la fecha y hora en los que efectivamente el vuelo operó. Por lo tanto, la variable **diff_min** fue creada considerando la diferencia absoluta en minutos entre la hora de partida y la hora anunciada, lo cual nos permite tener una mayor idea sobre la gravedad del retraso que sufrió un vuelo.

Periodo del día

Otra variable que puede ayudar en el modelado de la solución es conocer para qué momento del día se había anunciado el vuelo. De acuerdo con la experiencia, se sabe que durante la noche existe una menor afluencia de vuelos en comparación con los que se operan durante la mañana o la tarde, sin embargo es necesario tener definidos los horarios en los que se consideran dichos periodos del día. De acuerdo con la aerolínea original, los periodos del día se definen de acuerdo con las siguientes condiciones:

- **Mañana:** Si el vuelo fue anunciado entre las **5:00 am** y las **11:59 am**.
- **Tarde:** Si el vuelo fue anunciado entre las **12:00 pm** y las **18:59 pm**.
- **Noche:** Si el vuelo fue anunciado entre las **19:00 pm** y las **4:59 am**.

Por lo tanto, se procedió a generar la variable auxiliar **periodo_dia** considerando las reglas anteriores. La siguiente tabla muestra un resumen de la distribución de vuelos para dicha variable.

Table 8: Resumen del número de vuelos por periodo del día

Periodo del día	Número de vuelos	Porcentaje
mañana	1791	0.2401
noche	3076	0.4124
tarde	2591	0.3474

Variable objetivo: Atraso 15 minutos

Considerando que el resultado esperado con este proyecto es predecir si un vuelo se retrasará o no con base en ciertas variables que lo describen, el último paso en la ingeniería de variables fue definir la varible de referencia o la variable objetivo. De acuerdo con la aerolínea original y su experiencia, un vuelo puede considerarse como retrasado si la diferencia en minutos entre su hora de anuncio y su hora de salida es igual o mayor a 15 minutos, por lo tanto se tomó de apoyo la variable **diff_min** también creada en este proceso de ingeniería de variables para crear nuestra variable objetivo llamada **atraso_15**. Esta variable toma únicamente el valor de 1 si el vuelo se retrasó con base en la descripción anterior, en caso contrario toma el valor de 0.

Un resumen de la distribución de vuelos restrasados y no retrasados se puede observar en la siguiente tabla, así como la proporción correspondiente que representa del total de vuelos.

Table 9: Resumen del número de vuelos retrasados y no retrasados

atraso	Número de vuelos	Porcentaje
No atraso	5674	0.7608
Atraso	1784	0.2392

Base para modelo

1. Analisis de la base de datos

GrafiCas eda

2. Modelo

Analisis modelo1