UNIVERSIDAD DE
BUENOS AIRES

FACULTAD DE INGENIERÍA

[75.06/95.58] ORGANIZACIÓN DE
DATOS

Curso : Martinelli

1ER CUATRIMESTRE DE 2024

TRABAJO PRÁCTICO N° 1: Análisis exploratorio de datos

GRUPO: 6	
APELLIDO, Nombres	N° PADRÓN
Dominguez,Gonzalo Alejo	109759
Hsieh,Cindy Teresa	108051

Link al Colab:
https://colab.research.google.com/drive/17888PziHmesU3ODikUjjUE4Sml3VxaM1?usp=sharing

Introducción:

En este informe, se presenta un análisis exploratorio del registro histórico de consumos de la línea aérea. El objetivo principal es comprender mejor el comportamiento de los clientes y las tendencias de los mismos realizando un detallado análisis de los datos proporcionados. El mismo tiene como finalidad descubrir patrones, relaciones y tendencias que puedan resultar de interés para la empresa.

Datos Utilizados:

Se utilizan dos conjuntos de datos:

1. Customer_airways_data: Contiene información sobre reservas, ventas de pasajes y detalles de vuelo.
2. Cleaned-reviews: Contiene opiniones de clientes sobre su experiencia de vuelo.

Limpieza de datos:

Inspeccionando brevemente los datos está claro que los mismos no están listos para ser analizados.

Limpieza de cleaned_reviews:

Al examinar el conjunto de datos 'cleaned_reviews', identificamos la presencia de reviews y comentarios repetidos. Utilizando el método 'value_counts()' en la columna de reviews, observamos que hay instancias donde el mismo review está asociado con diferentes calificaciones ('rates'). Para abordar este problema, decidimos eliminar las filas duplicadas. Eliminamos ambas filas y no solo una de las dos, ya que no podemos determinar cuál de los registros duplicados es el válido.

Además, haciendo uso de la columna ya provista originalmente en el set de datos llamada 'date' (fecha) agregamos una columna que indica el día de la semana en el que se realizó el vuelo, a esta columna la llamamos 'flight_day'. Esto lo hicimos con el objetivo de facilitar operaciones de fusión con otros conjuntos de datos, lo cual será explicado en detalle más adelante. Como aclaración le cambiamos el nombre al dataframe a reviews_limpias.

Limpieza de customer_airways_data:

```
reviews_limpias['date'] = pd.to_datetime(reviews_limpias['date'])
reviews_limpias['flight_day'] = reviews_limpias['date'].dt.day_name().str.slice(stop=3)
```

En cuanto al conjunto de datos `customer_airways_data`, comenzamos eliminando filas duplicadas al igual que como hicimos con el set anterior. La diferencia en este caso estuvo en que tan solo se eliminó una de las dos columnas repetidas. Esto se debe a que al tener

exactamente la misma información podría tratarse de un error en la carga de la información y no queremos desechar datos potencialmente válidos.

Luego para exprimir al máximo la información ya presente en el set de datos y aprovechando que la columna route está conformada por el código IATA del aeropuerto de origen más el código del aeropuerto de destino del vuelo , usamos esa información para agregar nuevos datos. Primero creamos dos nuevas columnas llamadas flight_origin y flight_destiny, en la primera pusimos el código del aeropuerto de salida y en la segunda el del aeropuerto de llegada. Acto seguido con la ayuda de la librería de python 'airportsdata'(cuya documentación se encuentra en el sig. link: airportsdata · PyPI) a partir del código IATA obtuvimos el código **ISO 3166-1** del país en el que se encuentra el aeropuerto. Por último usando una segunda librería 'pycountry' (cuya documentación se encuentra en el sig. link: pycountry · PyPI) mapeamos esos código de país al nombre correspondiente del mismo. Finalmente dejamos las columnas flight_origin y flight_destiny con los nombres del país de origen y destino del vuelo respectivamente. Como toque final, acomodamos estas columnas nuevas para que las que son afines queden juntas.

Con estas acciones, hemos completado la limpieza de los datos y los hemos preparado para su análisis.

Merge de Ambos Dataframes:

Para obtener un conjunto de datos combinado, realizamos un merge (o unión) entre los dataframes customer_airways_data y reviews_limpias. Para realizar esta tarea realizamos un par de modificaciones y consideraciones. Primero para buscar una columna en común se consultó si en el set cleaned_reviews la columna de país representaba al país desde donde se hizo la reserva, ante la respuesta afirmativa ya teníamos la primera columna compartida entre ambos conjuntos de datos. Aun así resultaba insuficiente la información para realizar algún tipo de unión. Para abordar esto, transformamos la columna 'date' del dataframe reviews_limpias (Como se mencionó más arriba) en días de la semana (flight_day) para poder fusionar los conjuntos de datos en función del país de origen de la reserva y el día de la semana del vuelo. Esto lo hicimos con el objetivo de buscar relaciones entre los datos y no limitarnos a analizar cada archivo por separado.

El tipo de merge realizado fue el interno (how='inner') para conservar sólo las filas que tienen coincidencias en ambos dataframes y no quedarnos con información incompleta.

```
final_df = pd.merge(customer_airways_data, reviews_limpias, left_on=['booking_origin', 'flight_day'],
right_on=['country', 'flight_day'], how='inner')
```

Es importante tener en cuenta que al no tener un identificador único del cliente en ninguno de los conjuntos de datos, esta operación puede no ser precisa, ya que pasajeros del mismo país y en la misma fecha pueden tener calificaciones y opiniones diferentes resultando así en una explocion de las cantidades de datos y en la sobrerrepresentación de algunos clientes. A pesar de este problema, decidimos continuar con el merge para explorar posibles relaciones entre los datos, como fue expresamente pedido.

Análisis Exploratorio

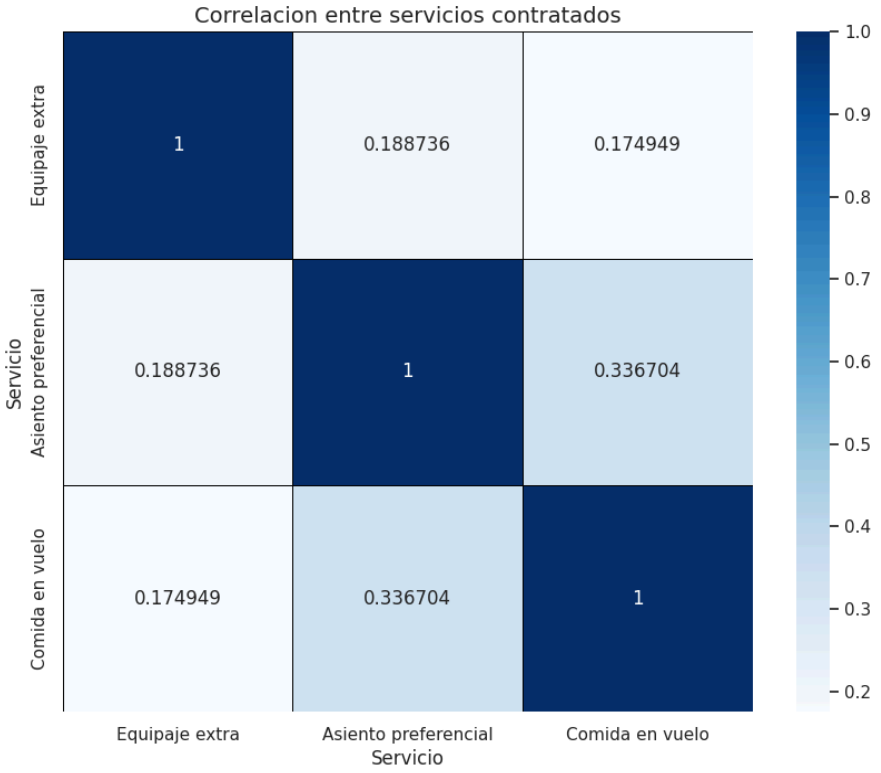
Obs 1: Todos los análisis que van a ser mostrados a continuación son considerando sólamente a aquellos datos de vuelos cuya reserva efectivamente se haya finalizado. Esto lo hacemos ya que solo queríamos considerar los datos de aquellos pasajeros que efectivamente hayan formalizado la compra.

Obs 2: Teniendo en cuenta la situación detallada en la sección anterior respecto al merge, cada plot fue realizado usando el dataframe correspondiente a la información de la que era necesaria disponer para crear el mismo.

Relación entre los servicios contratados

La primera hipótesis que consideramos a la hora de analizar los datos fue que debía haber una fuerte relación entre los distintos servicios ofrecidos a bordo en los vuelos. En otras palabras queríamos explorar la idea de que por ejemplo aquellos pasajeros que contratan un servicio como equipaje extra, por decir algo, tenderían (Quizá al tener un mayor presupuesto) a contratar también otros servicios como comida durante el vuelo o asiento preferencial y viceversa.
Para ello armamos una matriz de correlación entre las columnas de los distintos registros de vuelos. Esta matriz para cada par de servicios ofrecidos a bordo nos indica el coeficiente de correlación entre ambos. Este coeficiente es un valor que puede ir de -1 a 1 con 0 indicando que no hay correlación. Un valor entre 0 y 1 indica que hay una correlación positiva, lo que quiere decir que si una de las variables crece la otra también lo hace. Mientras más cerca de 1 más fuerte la correlación positiva.

Para representar esta matriz de correlación hicimos un uso de un heatmap como puede verse a continuación:



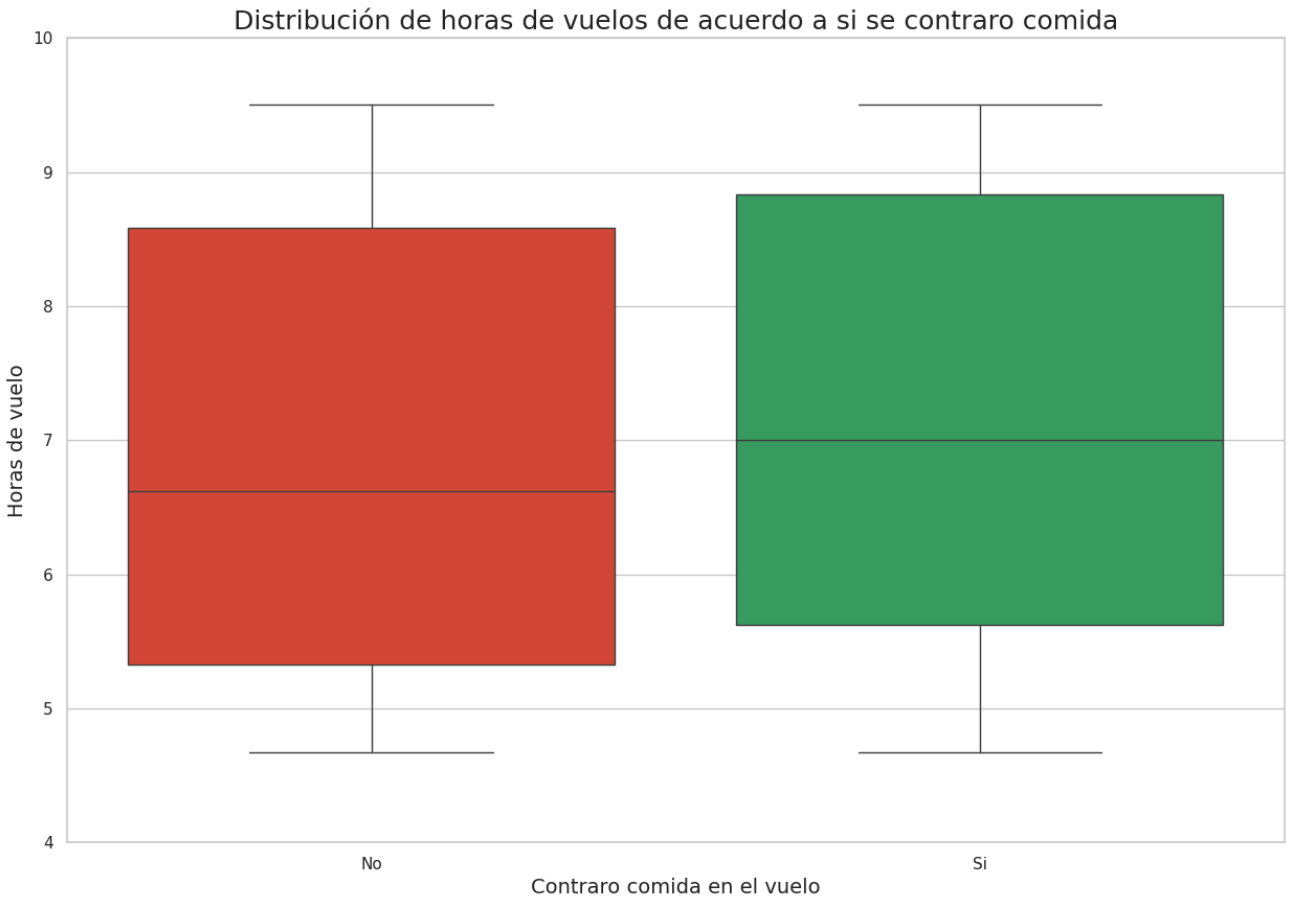
Un color más claro marca menor correlación y uno más fuerte una mayor correlación. Como puede verse en el gráfico, nuestra hipótesis no fue del todo correcta ya que a pesar de que los servicios a bordo guardan una correlación positiva entre todos esta no es lo suficientemente fuerte o alta como pensábamos que iba a ser.
Otro aspecto que se puede observar es que tanto comida en vuelo como equipaje extra tiene una correlación muy baja con equipaje extra, siendo las mismas de 0.17 y 0.18 respectivamente. Por otro lado, en el caso de asiento preferencial y comida en vuelo la correlación es bastante más alta, de 0.33. Esto podría llegar a indicar que los clientes que contratan alguno de estos dos servicios priorizan principalmente la comodidad y confort dentro de la cabina por sobre poder llevar equipaje de más. Esto es algo que podría resultar conveniente que la empresa tenga en cuenta.

Distribución de horas de vuelos de acuerdo a los diferentes servicios contratados.
La hipótesis en este caso era ver si aquellos pasajeros que van a realizar vuelos más largos tienden a contratar servicios. Ya que al tener más horas de viaje, los pasajeros tienden a querer servicios extra para así aumentar su comodidad durante el vuelo.

Para ello armamos una serie de boxplots donde se puede apreciar las diferentes distribuciones de las horas de vuelo para cada servicio y de acuerdo a si se contrató o no algo.

Comida en vuelo

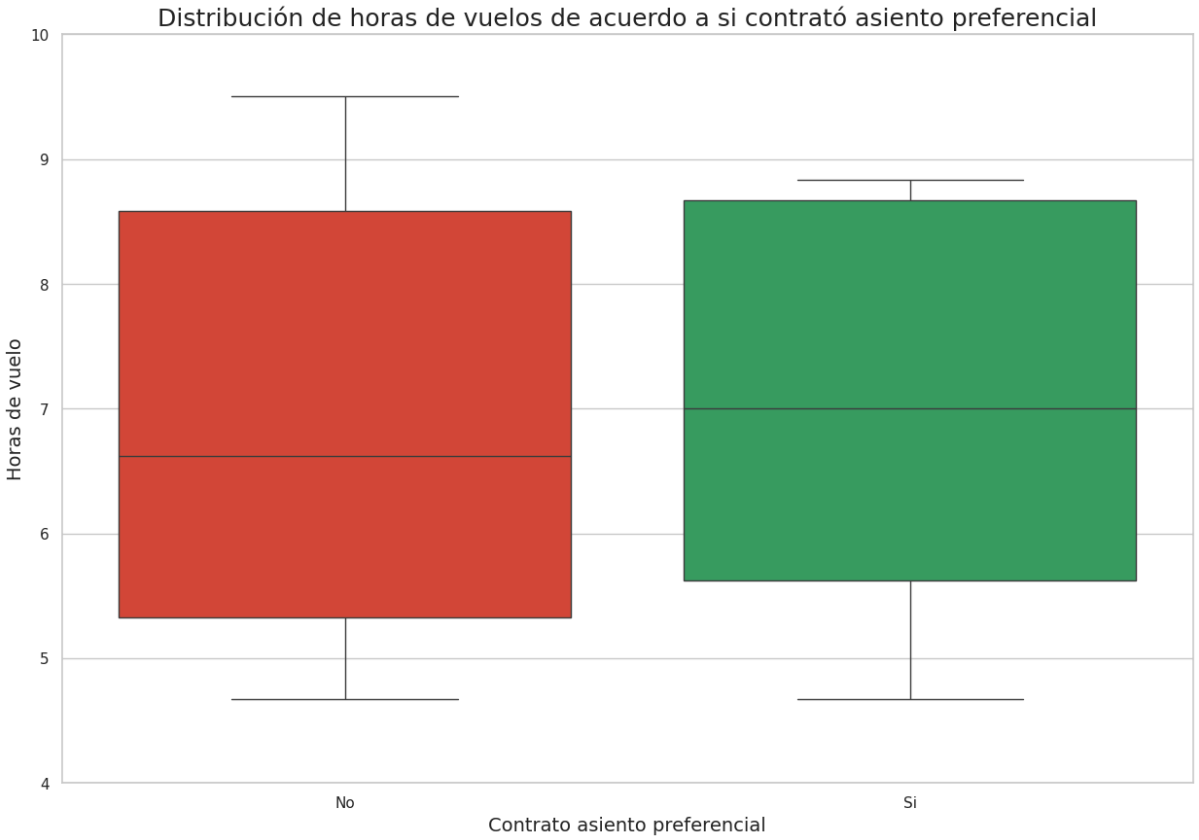
Consideramos la idea de que aquellos pasajeros que tienen vuelos más largos al pasar tantas horas sin comer algo van a tender a contratar comida durante el vuelo.



Se puede apreciar de forma clara como en aquellos vuelos en los que se contrató comida la media de duración de los vuelos es más alta que en aquellos en los que no. El boxplot de la derecha está más arriba, lo que marca que tanto el máximo como el mínimo de duración de los vuelos con comida extra son más altos que los de aquel vuelo en los que no se contrata.

Asiento preferencial

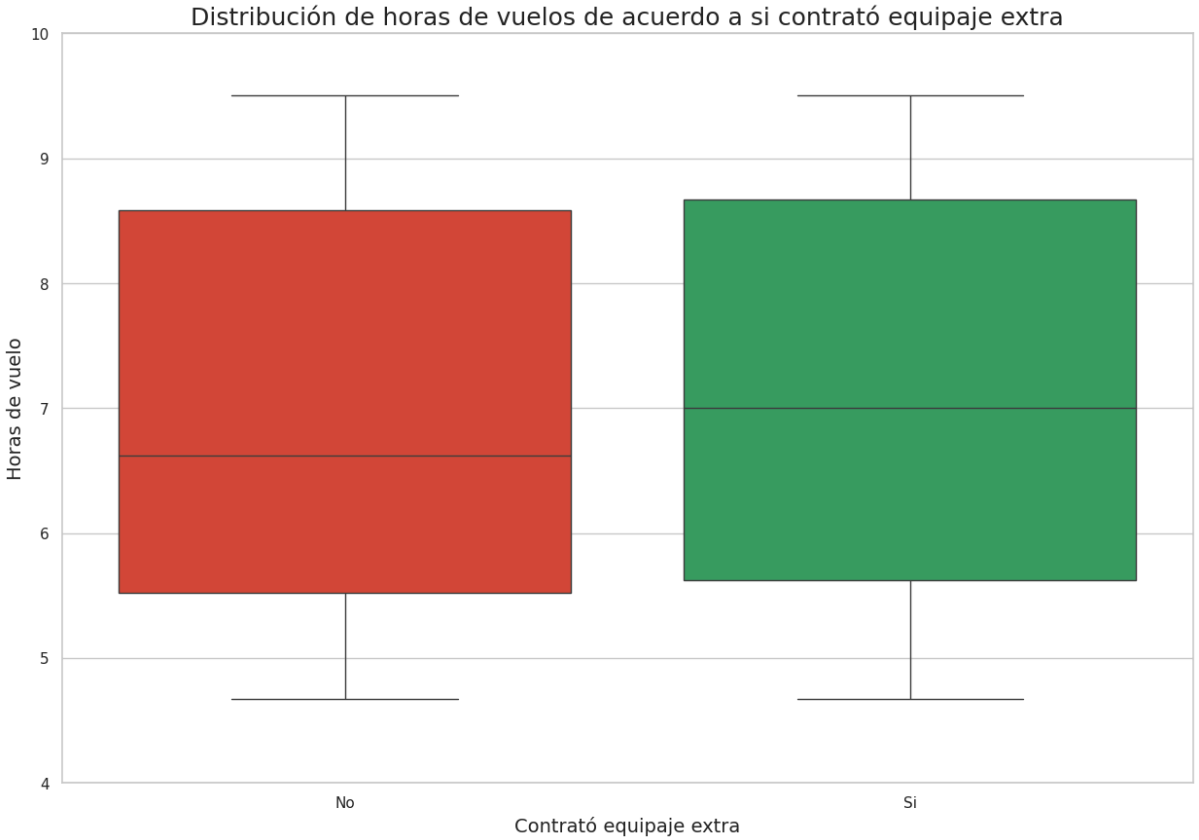
Siguiendo con la idea planteada antes, era probable que en vuelos de mayor duración los pasajeros al tener que afrontar tantas horas de viaje por delante prefieran elegir su asiento para de esa forma puedan transitar el momento de la manera más cómoda.



De igual manera que en el gráfico anterior se puede apreciar de forma clara como en aquellos vuelos en los que se contrató asiento preferencial la media de duración de los vuelos es más alta que en aquellos en los que no, aunque en menor medida en comparación con la comida durante el vuelo.

Equipaje extra

Acá queríamos ver si aquellos pasajeros que realizaban vuelos de mayor duración, al probablemente realizar viajes a destinos más lejanos, tendían a contratar equipaje extra.

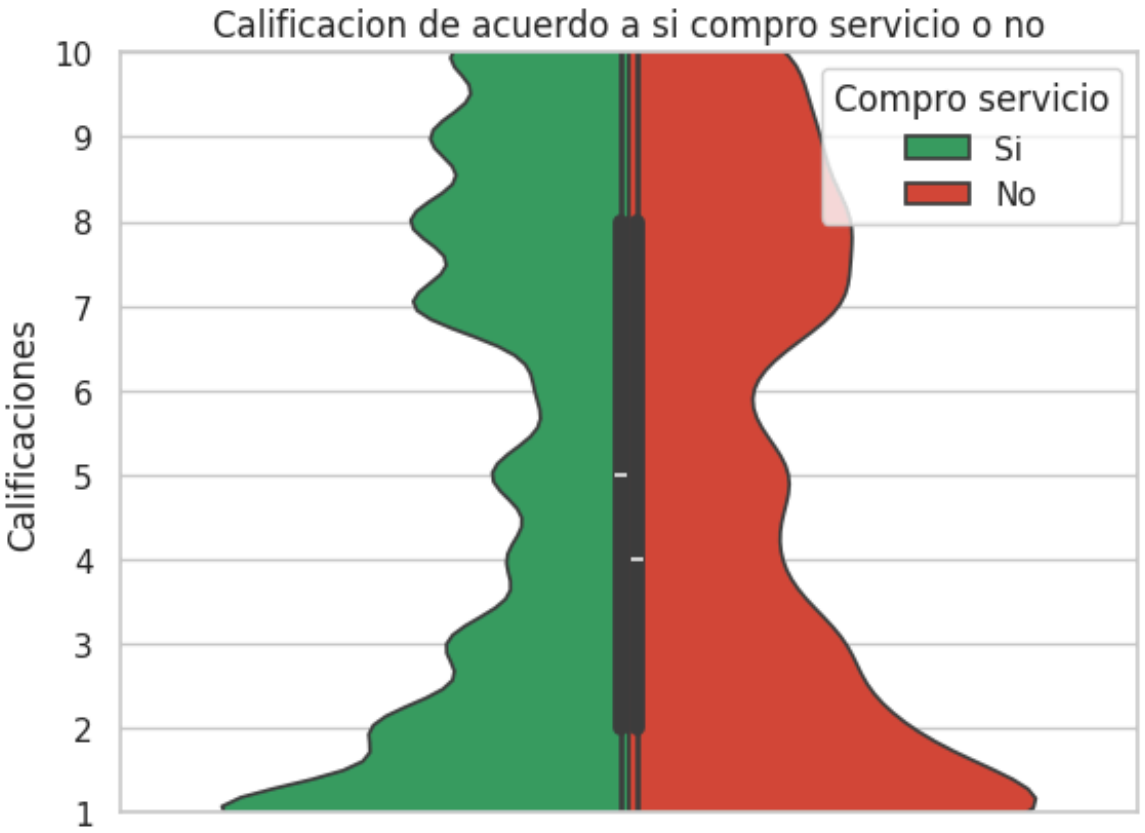


Como en los gráficos anteriores, en aquellos vuelos en los que se contrató equipaje extra la media de duración de los vuelos es más alta que en aquellos en los que no.

Como conclusión puede apreciarse que en todos los boxplot confeccionados se sigue la misma tendencia, evidentemente la duración de los vuelos tiene una gran influencia en la contratación o no de los servicios extra ofrecidos por la aerolínea.
Por ende resultaría muy conveniente para la compañía ofrecer una mayor cantidad de vuelos de larga duración y promocionarlos para así incentivar un mayor consumo y gasto por parte de los clientes.

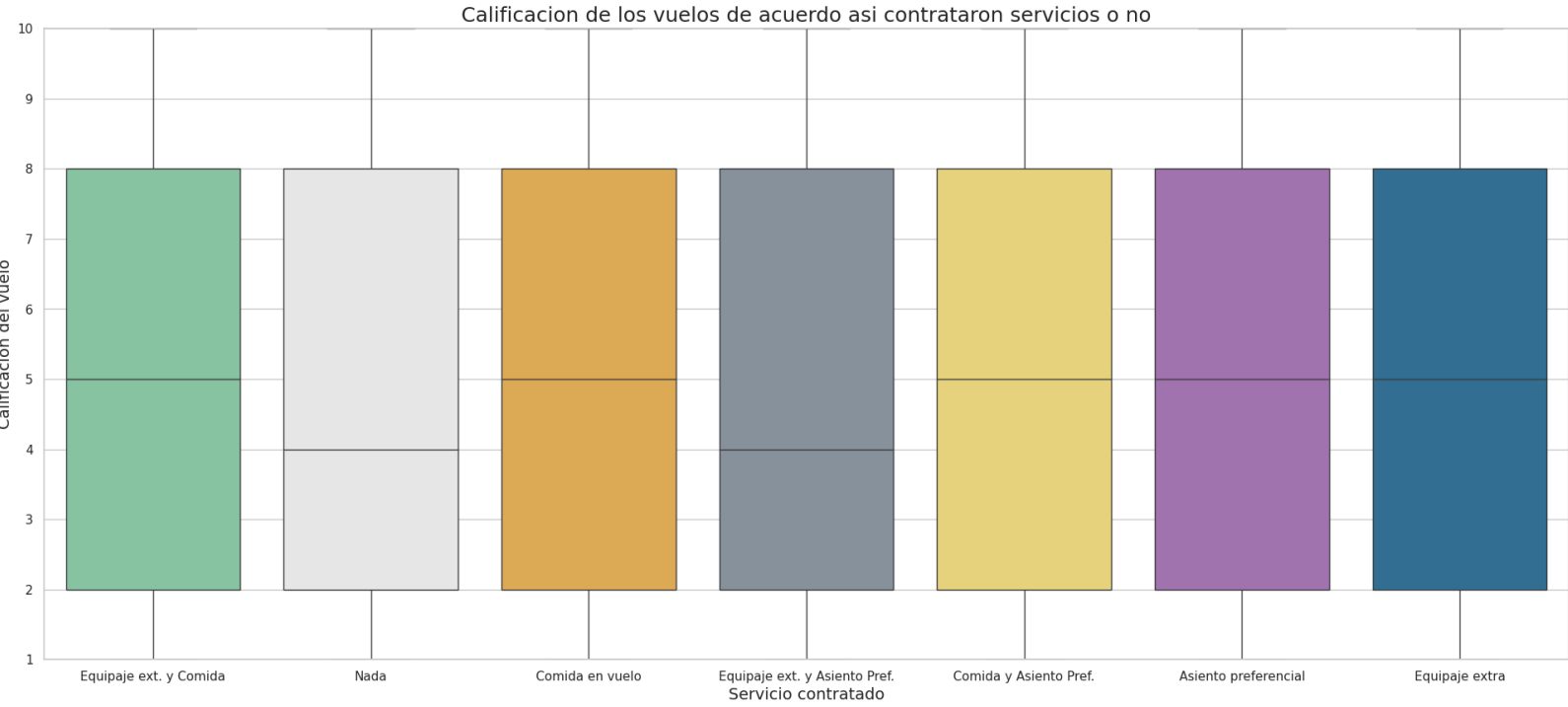
Calificación de los vuelos de acuerdo a si contrataron servicios o no

Estamos interesados en saber si los servicios adicionales ofrecidos tienen un impacto positivo en la calidad general del servicio ofrecido por la aerolínea, lo que se puede traducir en calificaciones más altas.
Para ello quisimos analizar y comparar las calificaciones de aquellas personas que contratan algún servicio con las notas de aquellos que no.



Como se puede ver en este gráfico parecería que el contratar un servicio extra se traduce en una mejor calificación por parte del pasajero, aunque no necesariamente tiene un gran impacto ya que no se da una gran diferencia entre una media de notas y la otra . Se puede apreciar que de media aquellas personas que contratan servicio ponen una nota alrededor de 5 (cinco) mientras que aquellos que no lo hacen ponen nota 4 (cuatro), tan solo un punto por debajo.
La diferencia que se puede observar claramente está en el número de calificaciones dadas. Del lado de los que contrataron algún tipo de servicio los picos en las distintas notas son más marcados lo que indicaría que hay una mayor cantidad de notas dadas por parte de los clientes. Por otro parte, del lado de aquellos que no contrataron nada no hay picos y el gráfico está más suavizado, lo cual es un resultado de que el muestreo de notas sea mucho menor. Se puede llegar a concluir con esto que en la amplia mayoría de los vuelos que se reservaron y se calificaron se contrató algún tipo de servicio extra. De igual manera se analizará con más detalle.

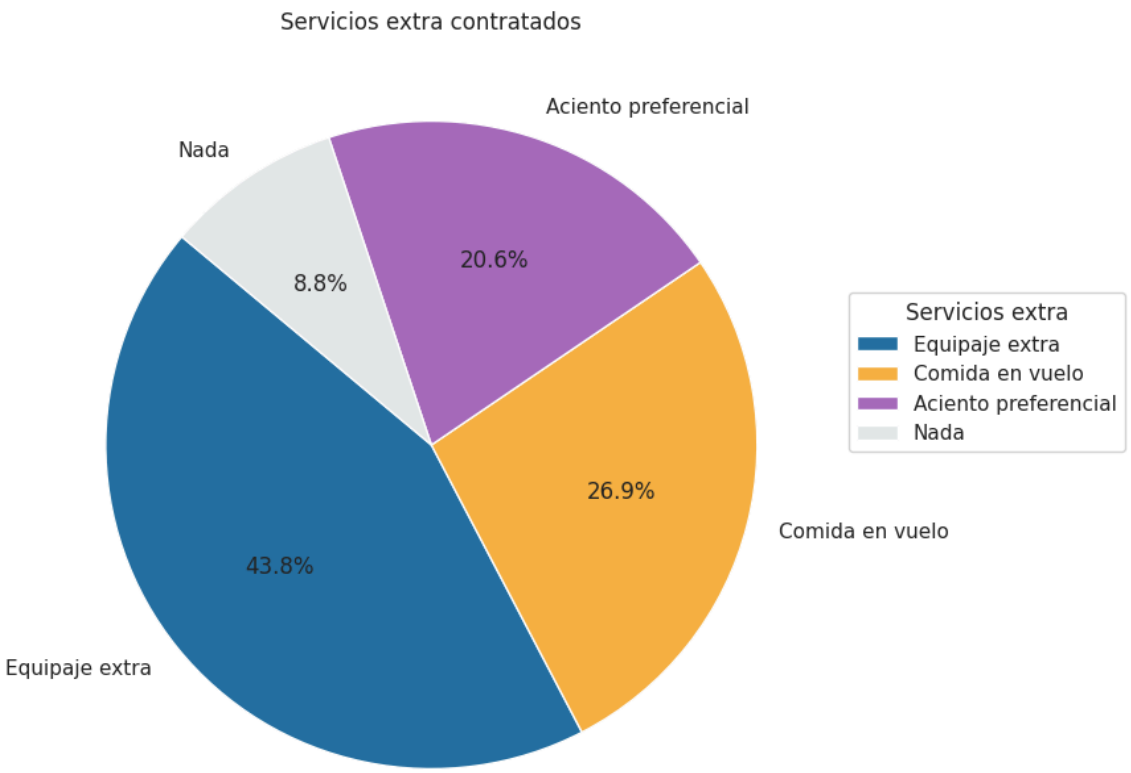
Para seguir analizando el impacto de la contratación de servicios en la calificación del vuelo y ver si podíamos descubrir algo más, armamos un boxplot en el que se pueda apreciar las distintas distribuciones de las notas dadas por los clientes de acuerdo al servicio que se contrató. Cabe aclarar que también tuvimos en cuenta aquellos casos en los que no se contrató nada y las distintas combinaciones posibles de servicios a la vez.



Podemos apreciar que de igual manera que en el plot anterior la media de notas de aquellos pasajeros que contratan algún servicio es 5 (cinco) y de aquellos que no es 4 (cuatro). También todos los boxplot están a la misma altura lo que indica que no hay grandes diferencias entre las distribuciones de notas y los máximos y los mínimos. Si es cierto como detalle que aquellos que contratan tanto equipaje extra como asiento preferencial serían de todos los cliente que contratan algo los que quedan menos satisfechos.
Esto lo que deja en evidencia es que parecería que de forma general no hay un servicio extra que sea superior a los demás en cuanto a dejar al cliente más satisfecho. Todos siguen la misma tendencia en cuanto a calificaciones dadas.

Servicios extra contratados

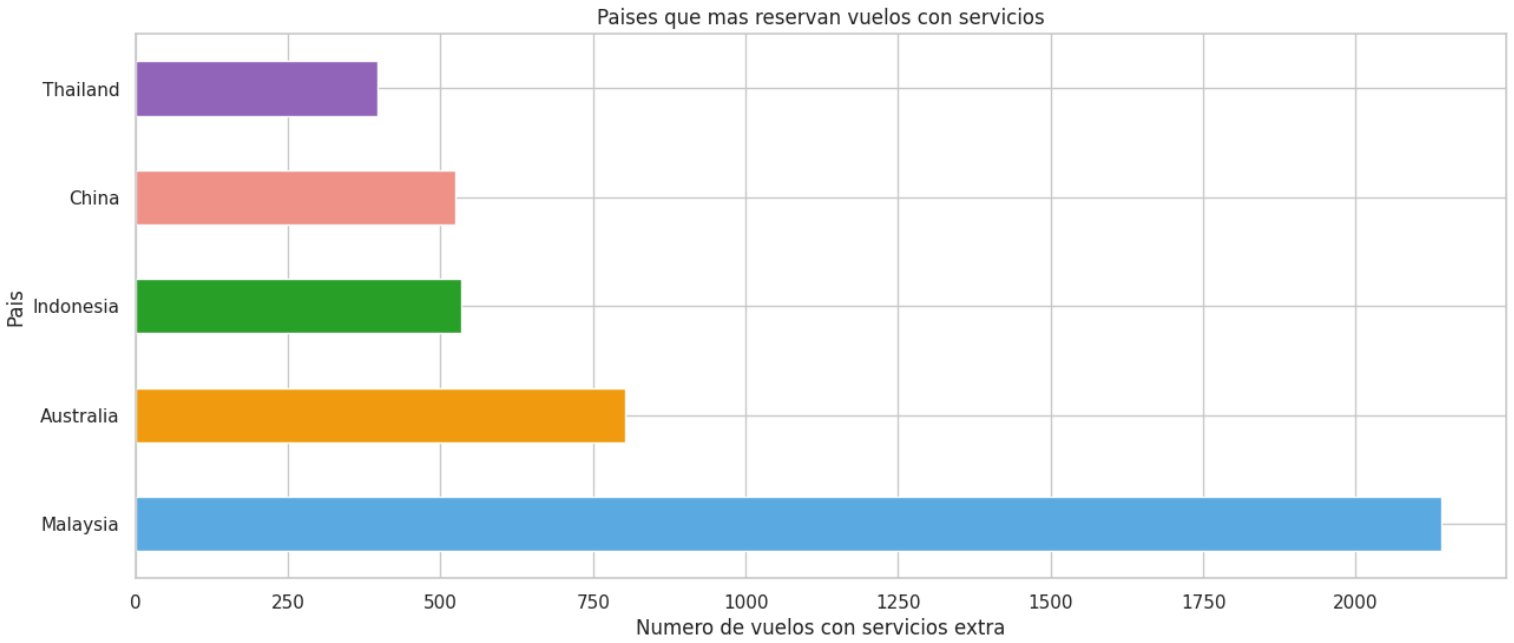
Algo que queríamos investigar en más profundidad era saber cuales son los servicios extras más populares entre los pasajeros.



En el gráfico de pastel, podemos ver que el servicio más popular, como era de esperarse, es el "Equipaje extra". Seguido por bastante lejos de la "Comida en vuelo" y el "Asiento preferencial", los cuales comparten entre sí porcentajes más cercanos.
Como algo ampliamente positivo para la aerolínea de todas las reservas de vuelos solo en el 8.8% no se contrató ninguno de los los tres servicios extra ofrecidos.
Algo que podemos plantear a partir de este plot es que la aerolínea podría beneficiarse de estrategias de marketing que promuevan estos otros dos servicios a través de ofertas especiales o descuentos y de esa forma tratar de que los mismos reemplacen el lugar ocupado por los vuelos en los que no se contrata nada.

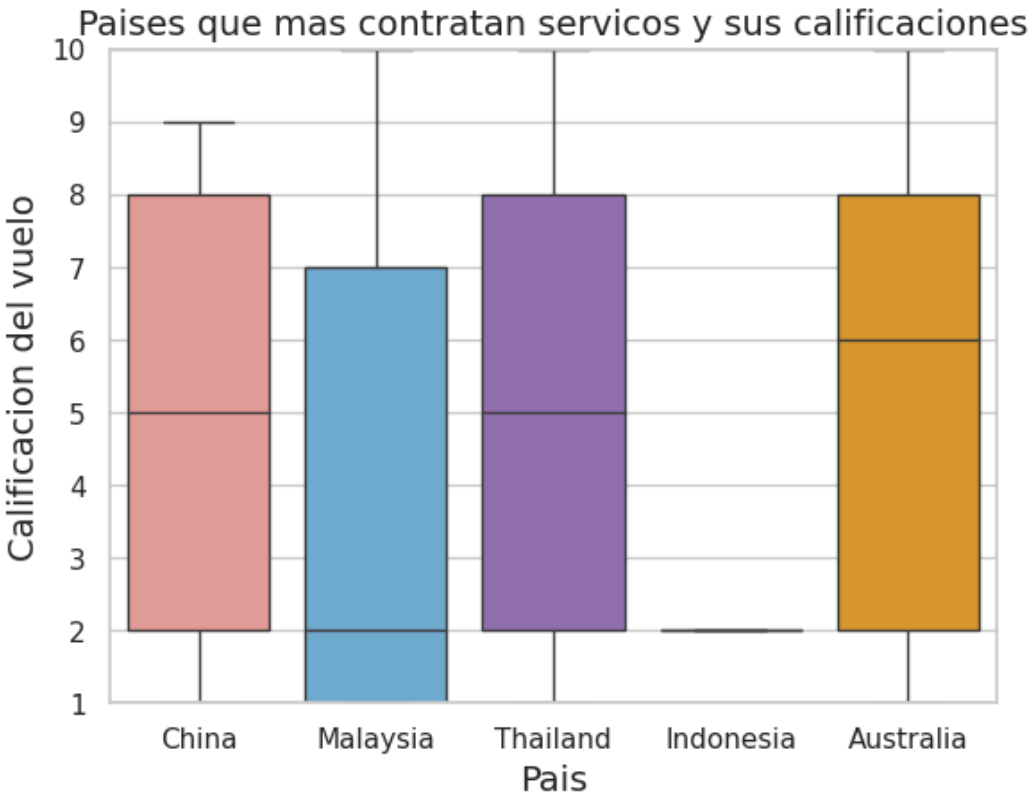
Países que más reservan vuelos con servicios

Para seguir entrando aún más en detalle, nos propusimos investigar cuales son los cinco países desde los que se realizan el mayor número de reservas de vuelos con algún tipo de servicio adicional. Esto lo hacemos con la finalidad de descubrir cuales son los mercados en donde los clientes gastan más plata en la aerolínea y así la empresa pueda poner el foco en captar un número de clientes potenciales aún mayor. También para detectar qué servicios están contratando, en qué cantidad y cuales son sus calificaciones.



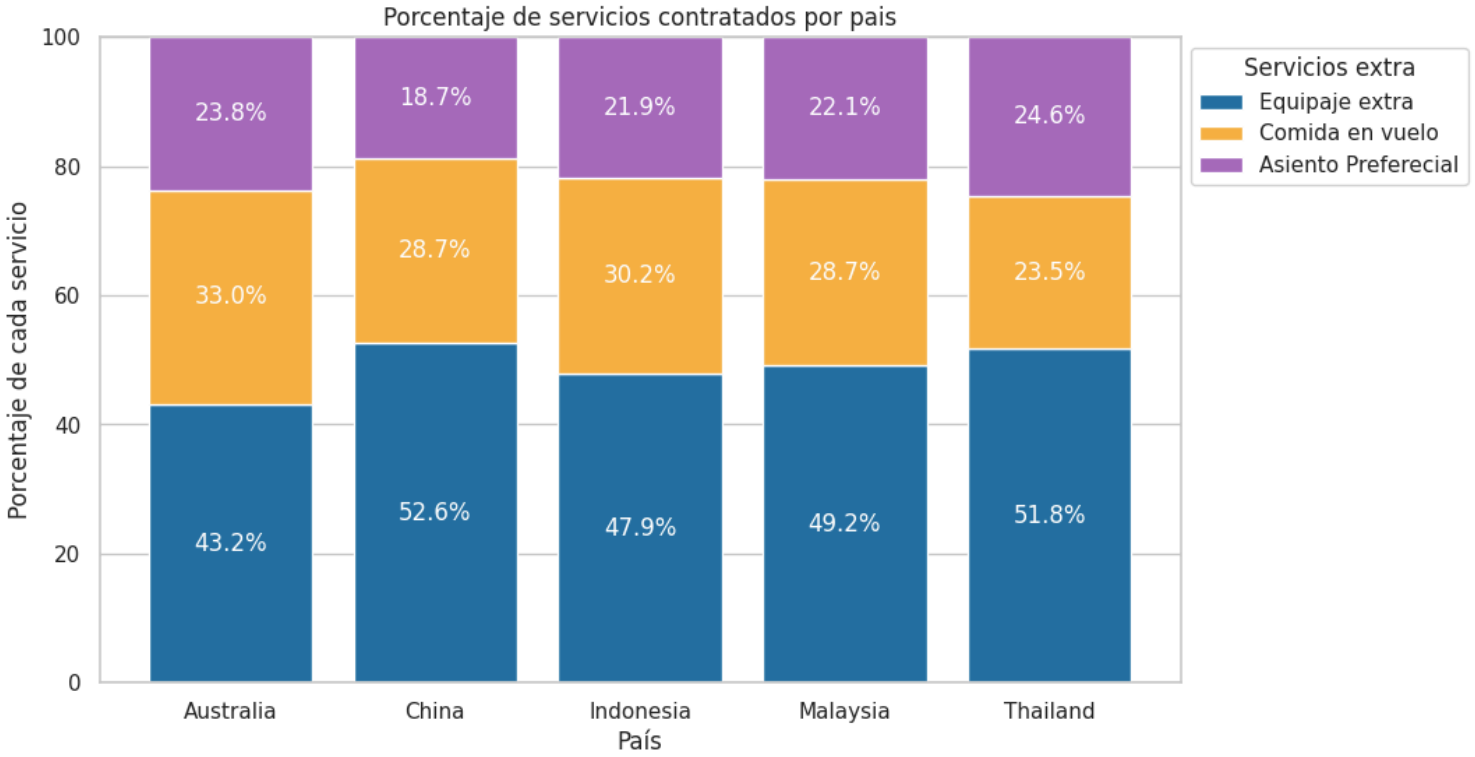
Podemos observar que la distribución geográfica de los países que más reservan vuelos con servicios provienen de la región del continente Asiatico, con Malasia siendo por una amplia diferencia el país con más reservas. Por otro lado tenemos a Australia en el segundo puesto representando a Oceanía. Las aerolíneas pueden utilizar esta información para aprovechar oportunidades de mercado en estos países donde la demanda de servicios adicionales es alta, y personalizar las estrategias de marketing.

Analizamos también qué notas están dando estos países.



Acá podemos ver varias cosas.
Hay variaciones en las calificaciones promedio de vuelo entre los diferentes países. Por ejemplo, Australia tiene las calificaciones más altas con un promedio de 6 (seis) por los que serían los que tienen la mejor experiencia. Por otra parte, Malasia tiene las calificaciones más bajas en promedio. Luego tanto China como Tailandia tienen un promedio similar en un punto medio. Por último para Indonesia ni se pudo generar el boxplot, lo que indica que no hay datos de este país en cuanto a calificaciones.
Varias cosas pueden concluirse de este plot. Por un lado es un punto muy positivo que Australia siendo el segundo país que más servicios contrata tenga una nota bastante alta, marca que los clientes en ese país están teniendo una buena experiencia.
Por otro lado sería recomendable para la compañía, teniendo en cuenta que Malasia es el mercado más grande en cuanto a servicios, ver cual es la situación de la sucursal de aquel país y que es lo que está causando las calificaciones tan bajas.
Por último sería muy beneficioso incentivar a los clientes de indonesia, siendo el tercer mercado más importante, a calificar más sus experiencia de vuelo a través de alguna campaña. Ya que creemos que sus opiniones pueden ser de gran ayuda a la hora de mejorar la calidad del servicio en general.

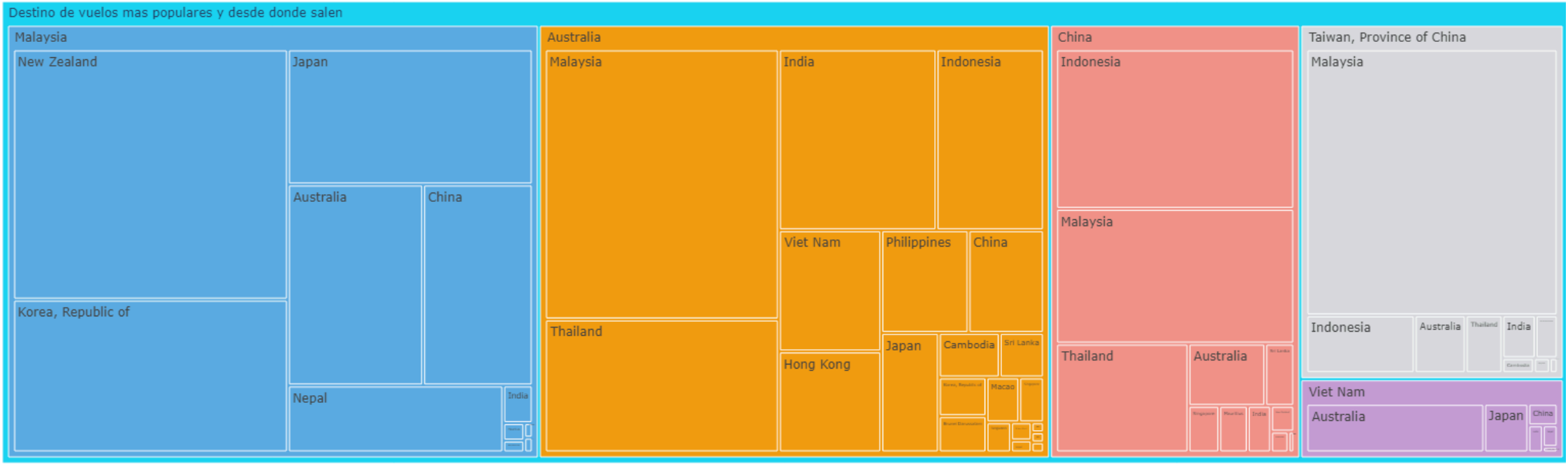
Para cerrar el capítulo de los países que más contrata servicios extra en los vuelos quisimos para c/u de los mismos ver los distintos porcentajes que ocupan los distintos servicios en el total de vuelos con servicios adicionales.



Se puede observar que en todos los casos el servicio siempre más contratado es el equipaje extra, seguido por la comida en vuelo y por último el asiento preferencial. Se sigue de forma evidente la misma tendencia mostrada en el pie plot que se encuentra un par de hojas más arriba.
Si hay algo que se puede destacar es el hecho de que de todos los países Australia es por al menos con una diferencia de 4 puntos el país que más comida en vuelo contrata relativo a la cantidad de vuelos con servicios totales. Coincidentemente también es como vimos más arriba el país con mejores calificaciones promedio y el segundo más grande en cuanto a volumen de vuelos con servicios extra.
Esto podría indicar que respecto a los otros países el contratar comida durante el vuelo tiene un mayor impacto positivo en la experiencia de vuelo. También es cierto que de todos los países es el que menos contrata equipaje extra por al menos 4 puntos también, lo que podría marcar que los demás países no estarían teniendo quizá la mejor de las experiencias usando este servicio.

Destino de vuelos más populares y desde donde salen

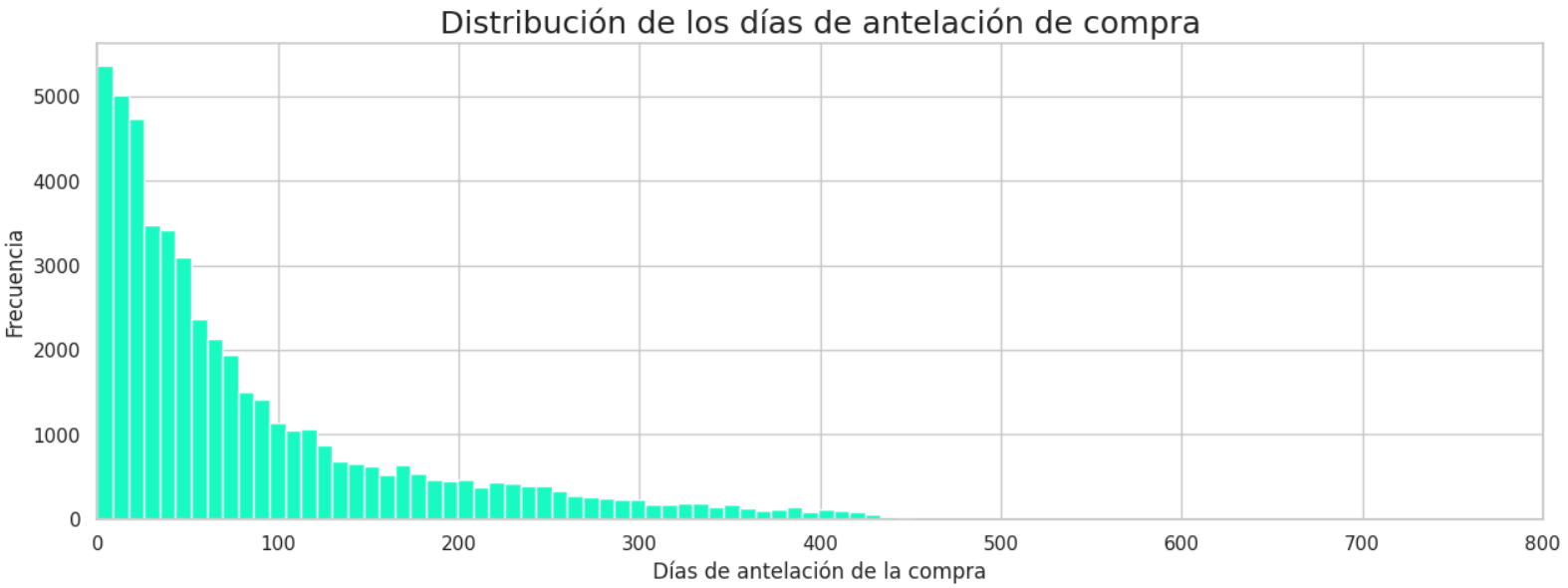
Para visualizar datos sobre los destinos de vuelos más populares y desde donde salen utilizamos un treemap para mostrar esta información de manera jerárquica. Los países más grandes en el gráfico representan aquellos destinos a los que más personas viajan, en este caso, Australia, Malasia, Vietnam, Taiwán y China destacan como destinos de interés. También podemos observar desde qué países salen la mayoría de los vuelos que van hacia estos destinos, mientras más grande el país de origen más vuelos salen de ese país.



Podemos ver cómo a lo largo de todo el treemap hay un amplio flujo de vuelos entre países del continente asiatico, siendo solo la excepción a la regla Australia. Creemos que este treemap puede ser una buena manera de ver los mercados en los que potencialmente se puedan seguir expandiendo la contratación de vuelos con servicios adicionales y profundizar aún más en aquellos países donde el servicio está más establecido como China, Australia, etc.

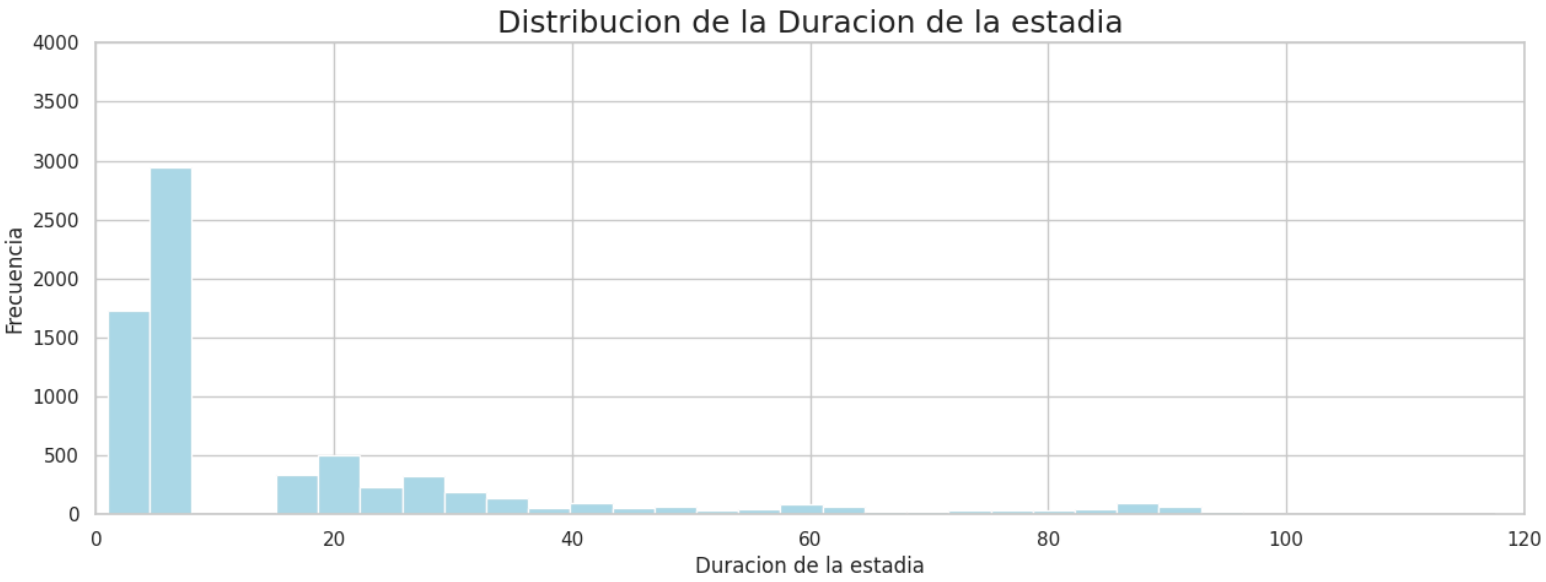
Distribución de los días de antelación de compra

Nuestra hipótesis se basa en que los pasajeros que compran boletos de avión con poco tiempo de antelación suele ser la mayoría de los casos. Para ello armamos un histograma para ver como es la distribución.



Nuestra hipótesis se confirmó como se puede ver. Podemos observar que los patrones de reserva llegan a un pico pocos días antes del vuelo y a partir de los 50 días aproximadamente en adelante la cantidad de clientes baja abruptamente llegando al punto de que casi nadie reserva por ejemplo con 400 días de antelación.
Esto es bueno para la empresa porque los clientes pueden llegar estar más apurados y tener menos tiempo para organizar su viaje, lo que podría llevarlos a optar por servicio adicionales, principalmente equipaje extra para hacer su experiencia de viaje más conveniente y cómoda olvidándose de tener que medir la cantidad de equipaje que llevan.

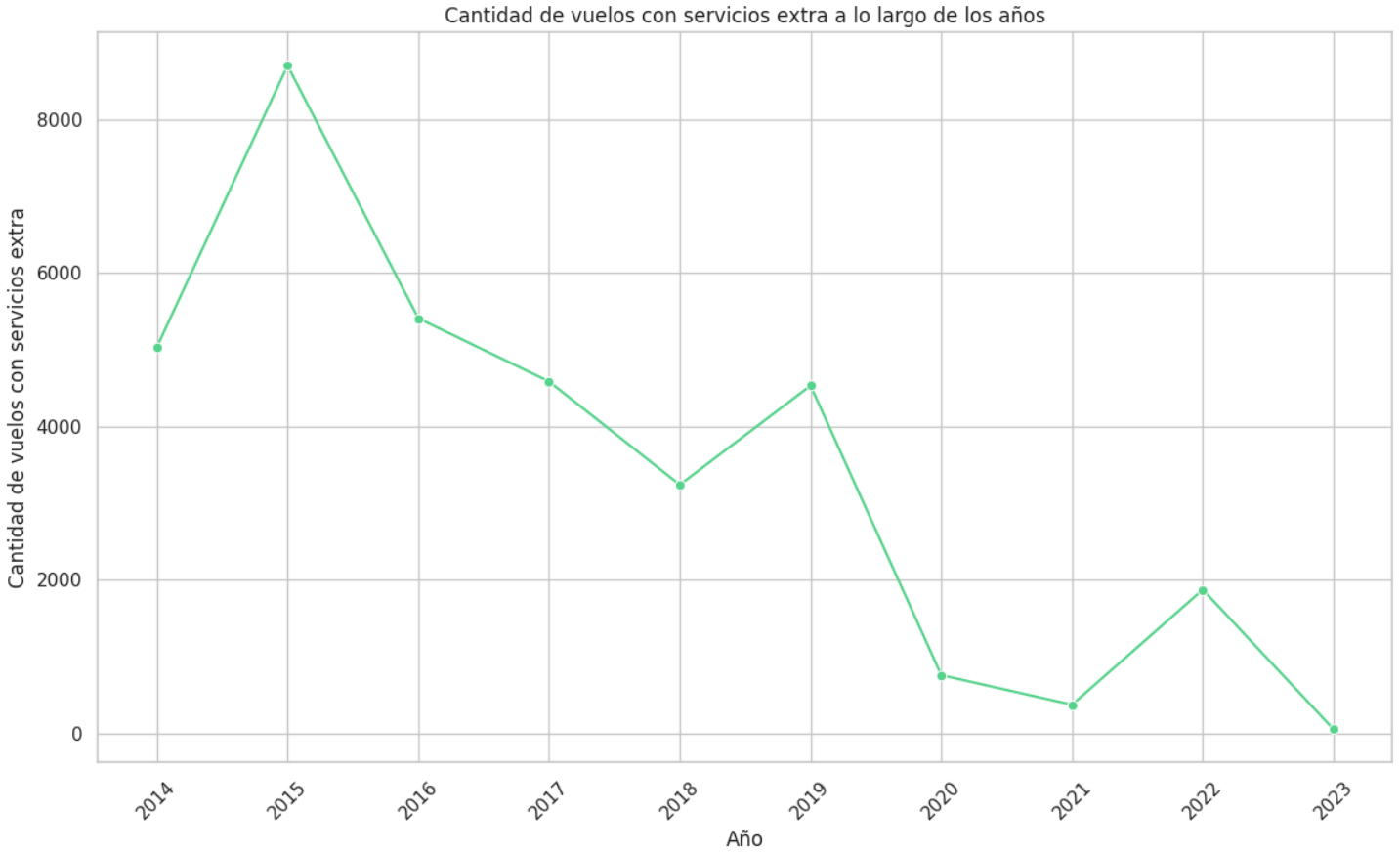
Hábitos de viaje



En las figuras se muestra por un lado, en el treemap, la distribución de vuelos según el día de la semana y en el histograma la distribución de las distintas duraciones de la estadía en un rango específico.
La alta actividad los Miércoles, Lunes y Martes podría indicar una concentración de viajes de negocios o viajes relacionados con el trabajo. La alta concentración de la duración de estadía en el rango de 0-15 días aprox. podría respaldar esta hipótesis, ya que los viajes de negocios tienden a ser más cortos, mientras que los fines de semana (Viernes, Sábado y Domingo) podrían ser más populares para los viajes de turismo u ocio.

Cantidad de vuelos con servicios extra a lo largo de los años

Queríamos ver cómo fue evolucionando a lo largo de los años el número de vuelos con algún tipo de servicio adicional contratado.



El análisis de la cantidad de vuelos con servicios extra a lo largo de los años revela una tendencia significativamente afectada por la pandemia de COVID-19. Se observa un pico notable en 2015, seguido de una disminución gradual en los años siguientes. Una posible explicación para este pico inicial podría ser un período de crecimiento económico o cambios en las políticas de la aerolínea que llevaron a más beneficios para los pasajeros lo que se tradujo en una mayor cantidad de vuelos con servicios extra contratados.

Luego por alguna razón empezó a haber una caída que se vio acelerada por la pandemia, con una marcada reducción en la cantidad de vuelos con servicios adicionales. Al comparar los períodos antes y después del inicio de la pandemia, es evidente el cambio sustancial en la cantidad de vuelos con servicios extra.

Aunque sí se puede destacar que luego del 2022 se ve una recuperación en la cantidad, la caída en 2023 es tan solo porque no se cuenta con datos suficientes de ese año.

Palabras más frecuentes en reviews

Primero, nos interesaba saber qué proporción de comentarios era de cada tipo para tener una idea general del sentimiento de los usuarios, los dividimos en dos grupos: los positivos y los negativos. Para determinar si el review tendía a ser positivo o negativo hicimos uso de la librería externa TextBlob (cuya documentación se encuentra en el sig. link: TextBlob).

Lo primero fue asignarle a cada review un puntaje de acuerdo al sentimiento en el comentario. Un puntaje menor o igual a cero se considera como negativo, caso contrario es tomado como positivo.

Después, realizamos un análisis de frecuencia de palabras en todos los reviews para identificar las palabras más comunes. Notamos que había muchas palabras irrelevantes, así que las sacamos de la lista.

Finalmente, generamos wordclouds separadas para los reviews positivos y negativos, lo que nos permitió visualizar las palabras más frecuentes asociadas con cada sentimiento

Para los reviews positivos consideramos aquellos que tengan una calificación mayor igual a 5 y un sentimiento positivo. Esto lo hacemos ya que aunque TextBlob fue de ayuda encontramos casos incongruentes donde el sentimiento era positivo pero la nota era 2.

Para los negativos consideramos aquellos reviews que tuvieran tanto sentimiento negativo como nota menor a 5.

Palabras más frecuentes en reviews positivos

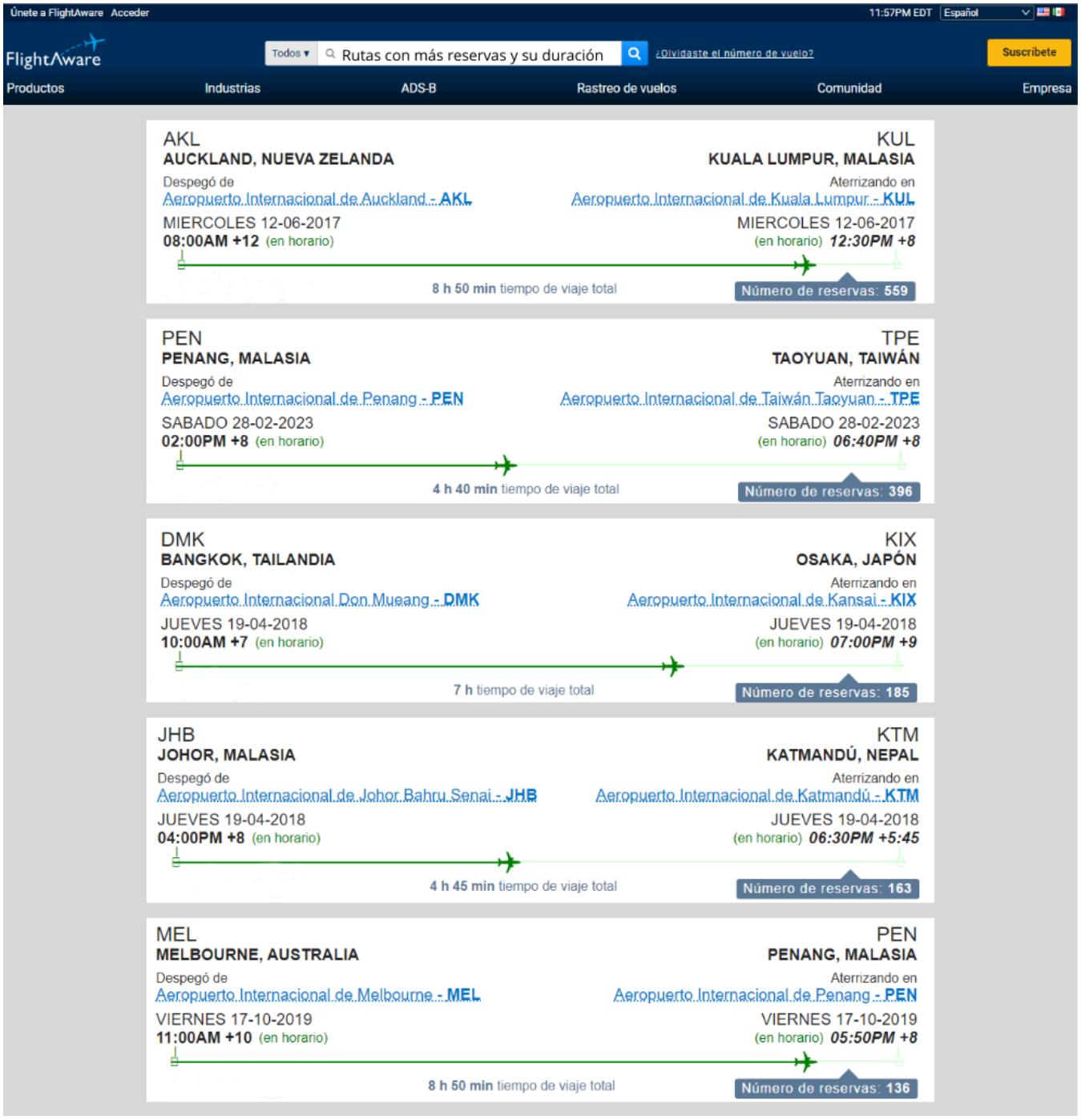
Los pasajeros que dejaron comentarios positivos es posible que tuvieron experiencias satisfactorias con el servicio, la tripulación, la comida, los tiempos de espera y la cabina.

Palabras más frecuentes en reviews negativos

Los pasajeros pueden haber experimentado problemas con la calidad del servicio, la comida, los tiempos de espera, los asientos y la actitud del personal.

Las palabras comunes tales la calidad del servicio, la comida a bordo, la comodidad de los asientos, la puntualidad y la actitud del personal en los comentarios negativos y positivos nos brindan una visión de que son aspectos críticos que afectan la experiencia del cliente y su percepción de una aerolínea. Como insight, la aerolínea puede utilizar esta información para identificar áreas específicas que requieren mejoras y para priorizar acciones correctivas.

Visualización final



La imagen simulada de la página de FlightAware nos muestra las rutas más populares, aquellas que conectan ciudades como Auckland, Kuala Lumpur, Penang, Bangkok, Osaka y Melbourne. Estas rutas suelen implicar vuelos de larga distancia, con duraciones significativas.

Al observar los datos sobre los países que más reservan vuelos con servicios adicionales, como equipaje extra, comidas a bordo o asientos preferenciales, notamos que estos países coinciden con los destinos de las rutas más largas y populares. Por ejemplo, países como Malasia, Australia y Tailandia son destinos comunes en estas

rutas.
Como vimos los pasajeros que viajan a destinos más lejanos y con vuelos más largos tienen más probabilidades de contratar servicios adicionales para hacer su viaje más cómodo y conveniente. Por lo tanto, la visualización de las rutas más reservadas y su duración destaca el hecho de que para beneficio de la empresa la mayoría de reservas se dan en estas rutas largas.
También al ver que entre las rutas con más reservas hay algunas con una duración corta en comparación seria una buena idea el hecho de en las mismas también incentivar a la compra de servicios adicionales.
Otra opción sería que la aerolínea podría focalizarse y solo destacar en servicios adicionales disponibles en las rutas de larga distancia y resaltando como estos pueden mejorar la experiencia de viaje de los pasajeros.