Informe del tp 1

Trabajo Practico n°1

Analisis Exploratorio de la aerolínea

Integrantes	Padrón
Marianela Fernanda Gareca Janko	109606
Beltran Malbrán	110036

Hoja de ruta

Link a Drive con los materiales:

https://drive.google.com/drive/folders/1yzugR3R3JMlwtenu3K76ExPgWaO4LZ82?usp=drive_link

Hipótesis 1

¿Existe una relación entre la satisfacción del pasajero y la elección de servicios específicos durante el vuelo? ¿Cómo esta relación varía entre diferentes regiones geográficas?

Archivos usados:

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de airline_data.csv y costumer_airways_data.csv.

Desarrollo:

Para este análisis, se hicieron modificaciones a los archivos para su uso correcto. Se eliminaron las filas duplicadas y se recategorizaron los tipos de datos para optimizar el uso de memoria, mejorando la velocidad de procesamiento. Luego se filtraron aquellos países que no alcanzaban el umbral mínimo de encuestas para ser representativos, es decir, aquellos que contaban con menos del 75% del total de encuestas realizadas. La elección del 75% como umbral mínimo es debido a que este porcentaje representa el tercer cuartil, el cual nos permite obtener los países con mayor cantidad de encuestas, y por ende los mas frecuentes en la aerolínea. Además, solo se tomaron en cuenta las encuestas que tenían la reserva completa.

Una vez hechas las limpiezas de datos correspondientes se analizaron los cinco países con el mejor promedio de puntaje. Podemos observar que a mayor promedio de puntaje las preferencia incrementan, esto es un indicador de que aunque las preferencias varían entre las diferentes categorías de servicios, hay una tendencia hacia puntuaciones más altas en relación con ciertas preferencias, como se muestra en la figura 1.

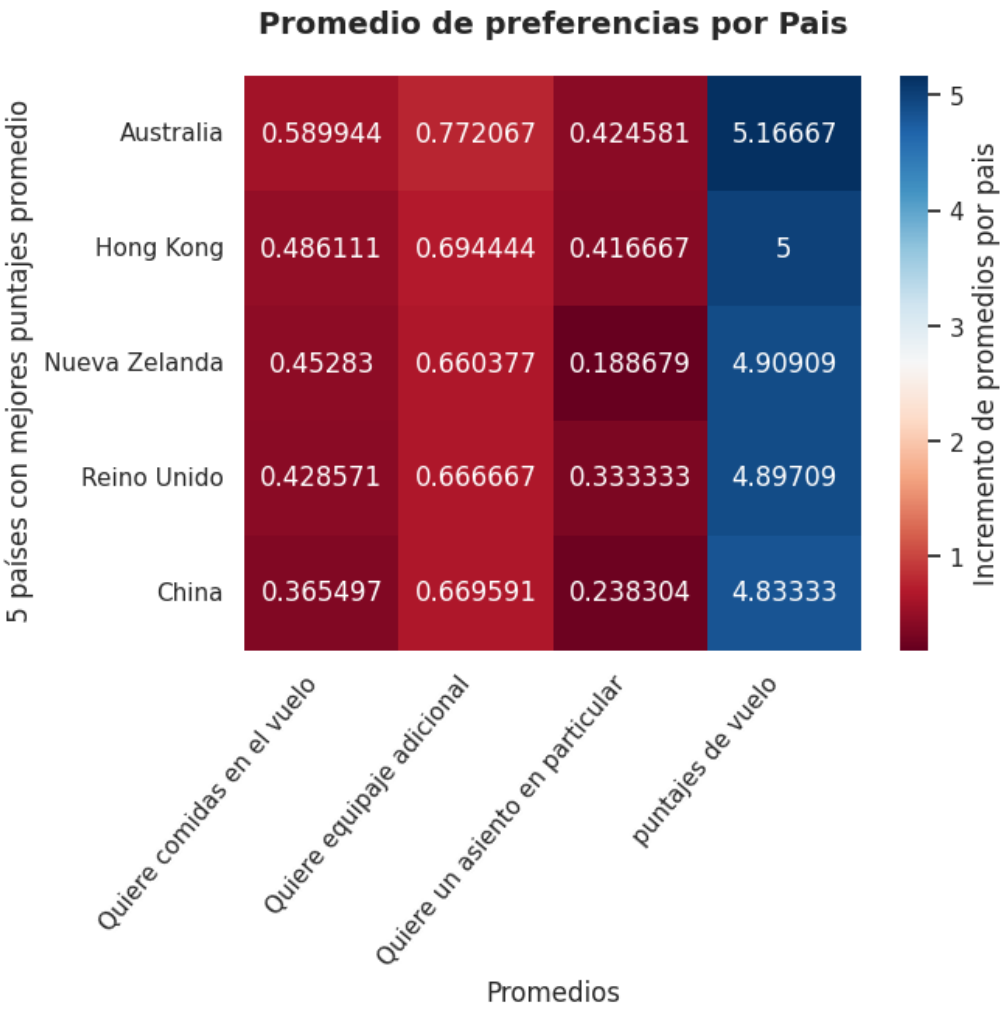

Figura 1— Relación de promedio de preferencias y puntajes de los 5 países con mejor promedio.

Para verificar si la hipótesis está respaldada, analizamos los cinco países con el menor puntaje promedio. En estos se observó que la elección del servicio “Quiere un asiento en particular” fue mayor que en los casos de países con mayor puntaje. Esto podría indicar que existe alguna relación entre la satisfacción del cliente y ciertas preferencias específicas, como las de obtener equipaje adicional o comer en el vuelo pero no con otras, como es el caso de la elección de asientos.

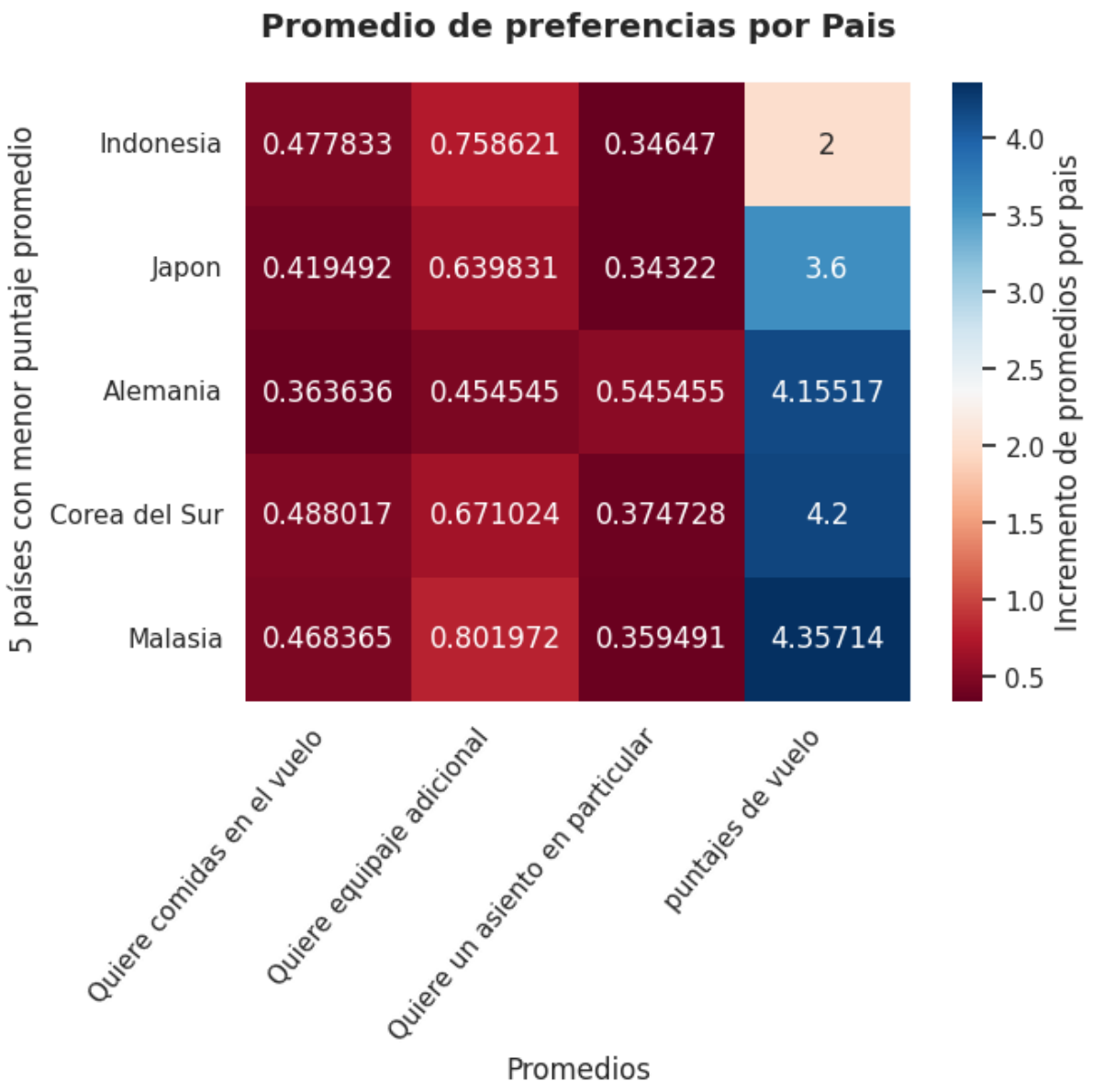

Figura 2 — Relación de promedio de preferencias y puntajes de los 5 países con mejor promedio.

Conclusión

Una vez analizadas las posibles relaciones entre los puntajes y los servicios disponibles durante los vuelos pasamos a realizar un análisis geoespacial de la distribución de los puntajes y de los países con mayor encuestas. Con esto se busca poder visualizar si existe relación entre los países con mayores encuestas realizadas y algún rango de puntaje en específico.

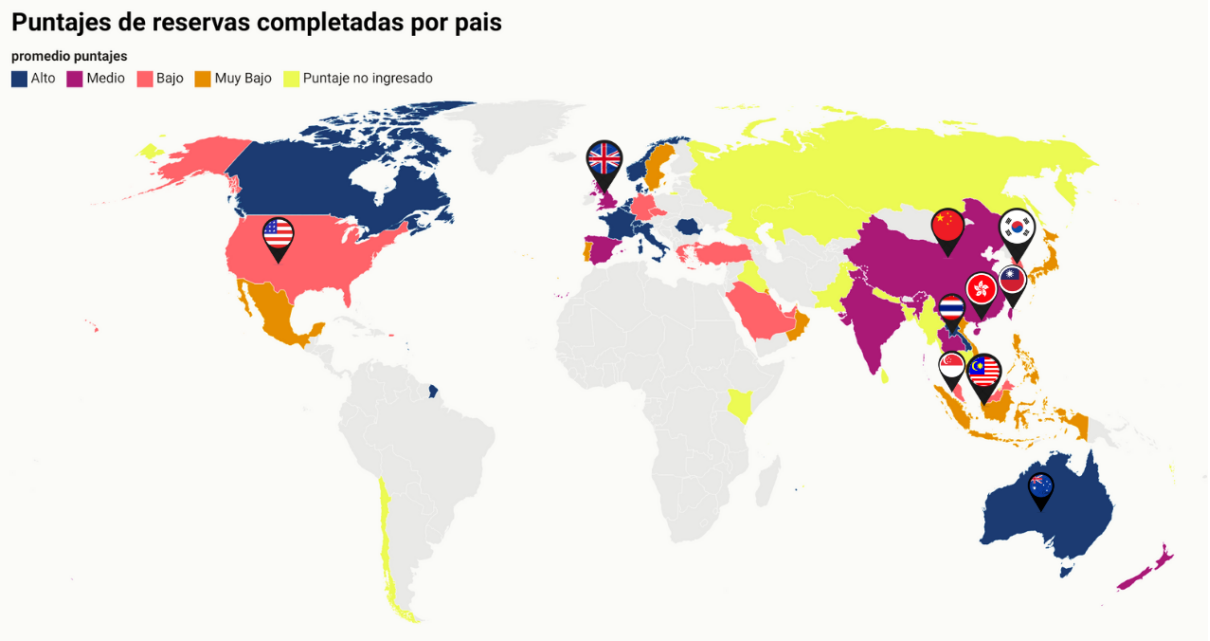

Figura 3 - Promedio de los puntajes de los países que usaron la aerolínea.

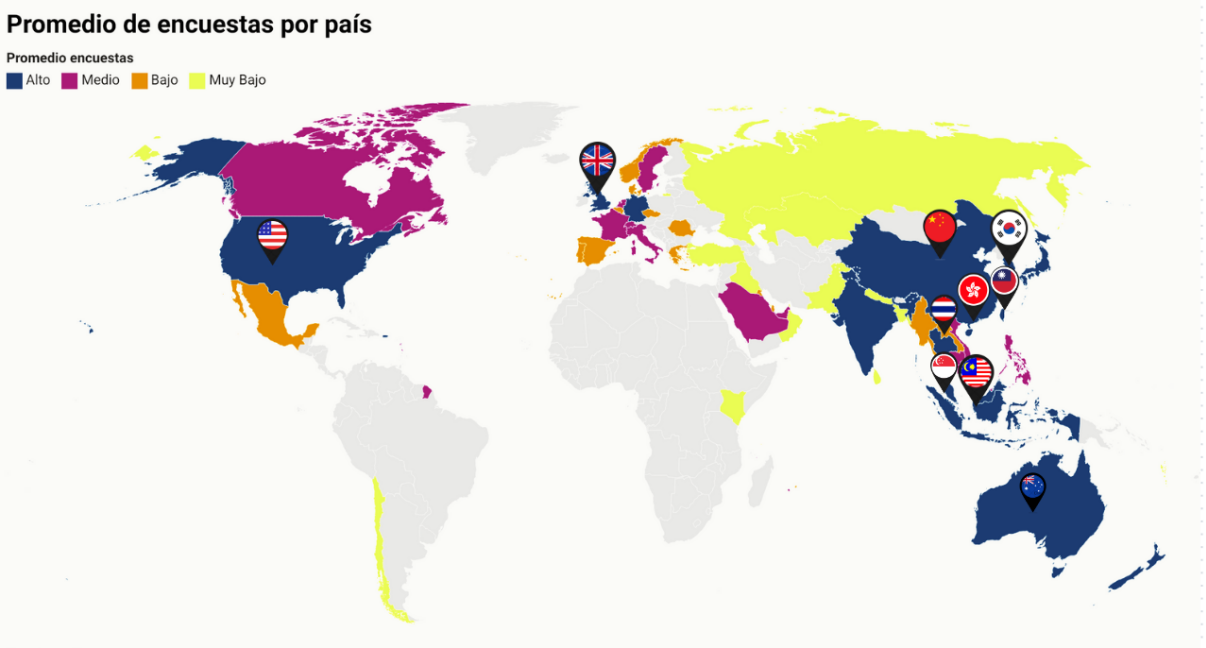

Figura 4- Promedio de encuestas por país.

La figura 3 y 4 nos permiten visualizar una clara relación entre los rangos de encuestas completadas y el puntaje en cada país. En primer lugar se puede observar que los puntajes no ingresados están relacionados casi en su totalidad con los países con un promedio de encuestas muy bajo, por otro lado la mayoría de los países que tuvieron un promedio de puntaje medio y alto, los cuales se encuentran en Europa y Asia, presentan un porcentaje de encuestas ingresadas

superior o igual al 50% del total, es decir están dentro del rango medio y alto, a excepción de los países pertenecientes a la región de Asia oriental los cuales, a pesar de tener un alto porcentaje de encuestas realizadas, presentan valores de puntajes que oscila entre muy bajo y medio.

Basándonos en la información analizada podemos concluir que a mayor cantidad de encuestas los puntajes van incrementando, excepto en la región de Asia. Esta particularidad podría significar que los puntajes están influenciados por los destinos más que por las preferencias en el vuelo o su volumen de encuestas.

Hipótesis 2:

¿Cuáles son las rutas más elegidas? ¿De qué nacionalidad son los viajeros que las toman? ¿Existe alguna relación con los puntajes de estos países?

Archivos usados

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de customer_airways_data.csv y cleaned_reviews.csv

Desarrollo:

En este análisis, buscamos determinar si existe alguna relación entre las rutas más populares y la nacionalidad de los viajeros, así como si estas rutas están vinculadas a los países con mayor puntaje. Para obtener esta información, filtramos los datos para incluir solo las encuestas que indican reservas completas de los clientes verificados, esta selección se hizo para asegurar un análisis preciso y representativo. Identificamos la ruta con más viajeros en cada país y seleccionamos las diez más populares con el fin de ver las rutas más influyentes para la aerolínea, y analizar la posible existencia de relación entre las rutas y las regiones con mayores puntajes, las cuales se pueden observar en la figura 3.

Conclusión:

De este análisis podemos concluir que, tal como se ve en la figura 5, la ruta que va de Nueva Zelanda a Malasia (AKLKUL) es la más elegida por los originarios de Malasia, Reino Unido y de Estados Unidos de América, lo cual la convierte en la ruta la más elegida de la aerolínea por una notable diferencia. Cabe destacar que la mayor cantidad de personas que eligen esta ruta son originarios de Malasia, en segundo lugar vienen los británicos, los cuales transitan por este tramo casi la mitad de veces que los malasios, y en último lugar estan los estadounidenses. Por

otro lado la segunda ruta más elegida es la que va de Vietnam a Australia (SGNSYD) la cual la toman exclusivamente los originarios de Australia.

Figura 5. Relación entre el país de origen del cliente y la ruta que utilizo.

Esta información indica que la mayoría de las personas que utilizan la aerolínea viajan de Nueva Zelanda a Malasia, siendo originarias de este último. Esto permite ampliar la información proporcionada por la hipótesis 1 ya que aunque los viajeros originarios de Nueva Zelanda no expresaron disconformidad con la aerolínea, como se observa en la figura 1, los viajeros malasios y estadounidenses, a pesar de ser de los viajeros más frecuentes como muestra la figura 4, han mostrado su insatisfacción con sus viajes, tal como se puede visualizar en la figura 3.

Hipótesis 3:

¿Cómo afecta el medio de compra de pasajes a los puntajes de los clientes en los diez países con el mayor volumen de compra?

Archivos usados:

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de customer_airways_data.csv y cleaned_reviews.csv

Desarrolló:

Para empezar este análisis realizamos una selección de datos con el fin de evaluar sólo aquellas encuestas que presentaban su reserva completa, esto se realizo para quedarnos con los usuarios que utilizan, y mejor representan, la aerolínea en cuestión. Una vez hecho esto, y teniendo en cuenta las conclusiones de la hipótesis 2, decidimos explorar más en profundidad los diez países que registraron el mayor volumen de compra de pasajes. Seguimos utilizando estos países con el fin de comprender por qué, a pesar de continuar eligiendo la aerolínea solo dos de ellos, Australia y Hong Kong, mostraron un nivel alto de satisfacción.

En primer lugar investigaremos si el proceso de compra del pasaje influye de alguna manera en la variación de puntajes. Para poder realizar esto empezamos viendo cuál fue el medio de compra más usado en todo el mundo, lo cual nos permitió ver que la mayor parte de clientes de la aerolínea utiliza el medio de compra catalogado como “Internet” para reservar su pasaje y el resto utiliza la opción de “Celular” tal como se ve en la Figura 6.

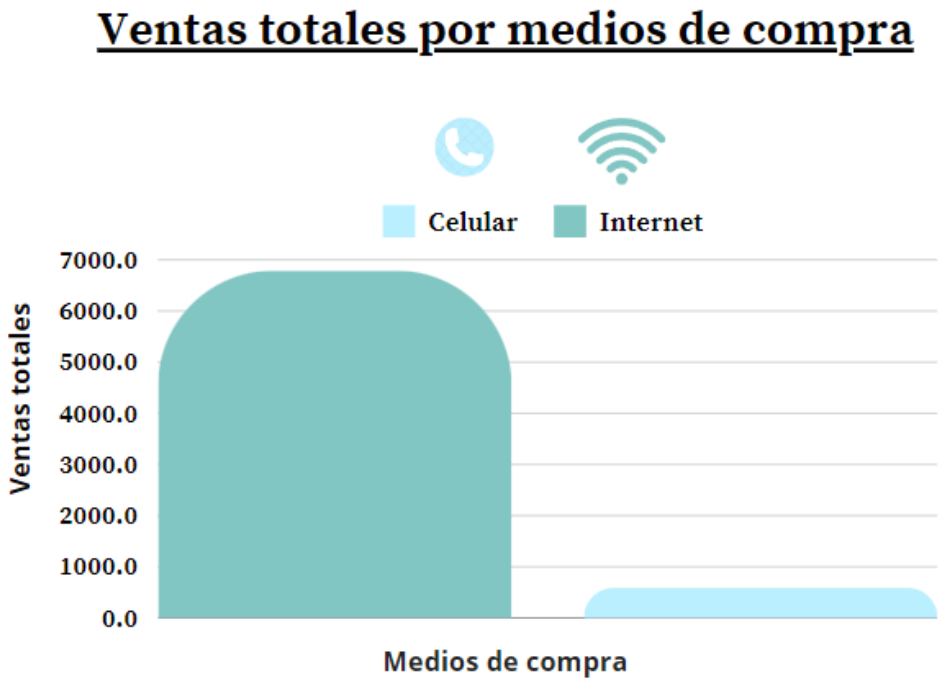

Figura 6 — Reservas totales según el medio de compra.

Conclusión:

Teniendo estos datos nos interesa conocer la preferencia de los diez países mencionados anteriormente, esto nos va a permitir ver si existe algún tipo de relación entre un medio de compra en particular y los países con un promedio de satisfacción bajo. Tal como se ve en la figura 7, podemos ver que no hay una relación directa entre la elección de un medio de compra en específico y su relación en puntajes, sin embargo se descubrió que China, Taiwán y Malasia eligen el celular como medio de compra. Esto es llamativo ya que, tal como se ve en la figura 6, es algo que no suele suceder. Otro dato a tener en cuenta es que Reino Unido no presenta ninguna reserva realizada a través de este.

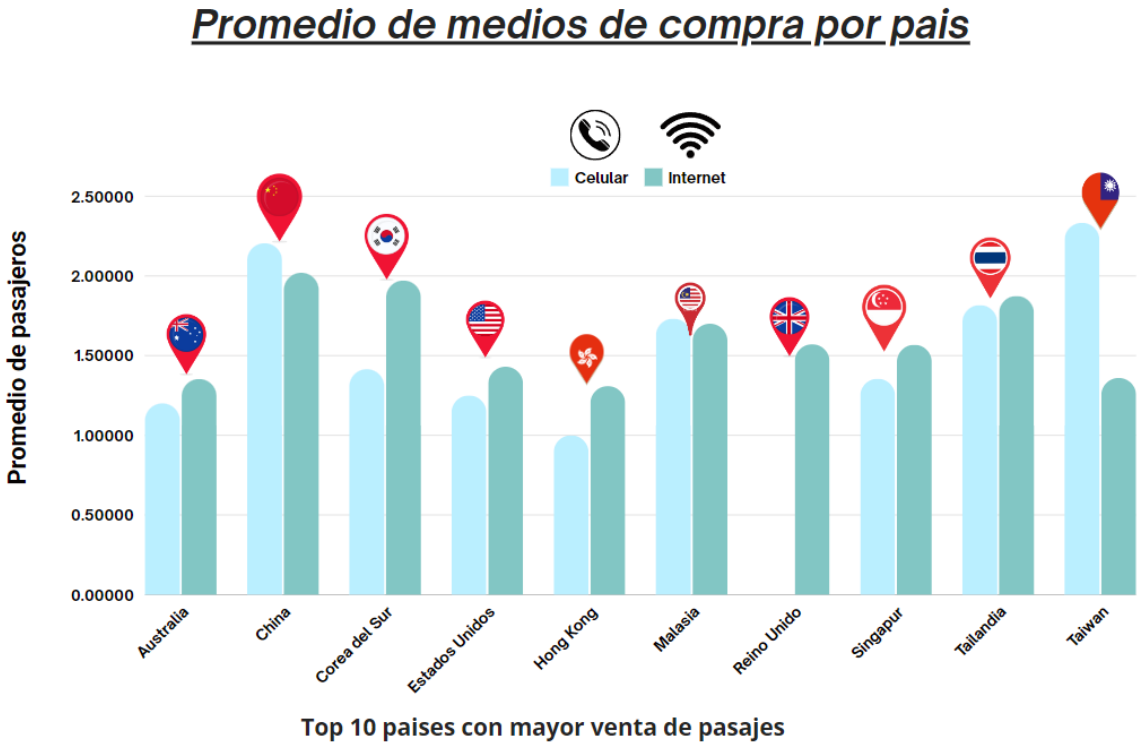

Figura 7 - Promedios de reserva de los 10 países que mas representativos por celular e internet.

Hipótesis 4:

¿Cómo afectan los días y horas de vuelo a la satisfacción de los viajeros en los países con un alto volumen de compra de pasajes?

Archivos usados:

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de customer_airways_data.csv.

Desarrollo:

Si bien la hipótesis 3 no nos presentó datos que respalden el motivo de por que los países con más compra de pasajes no presentan una alta tasa de satisfacción, si nos permitió descartar que está relacionado al medio de compra que utilizan. Teniendo en cuenta esto decidimos ver si la poca cantidad de puntajes altos presentes está relacionada a la hora y día en la cual se realizaron estos vuelos.

Tomando como parámetro que los días y las horas laborables resultan inconvenientes para el viajero, analizaremos el número de pasajeros por día, tal como está indicado en la figura 8, y por hora, como muestra la figura 9. Para poder realizar este análisis conservamos las encuestas que completaron su reserva y sumamos todos los pasajeros de la aerolínea para luego dividirlos por día y hora de partida.

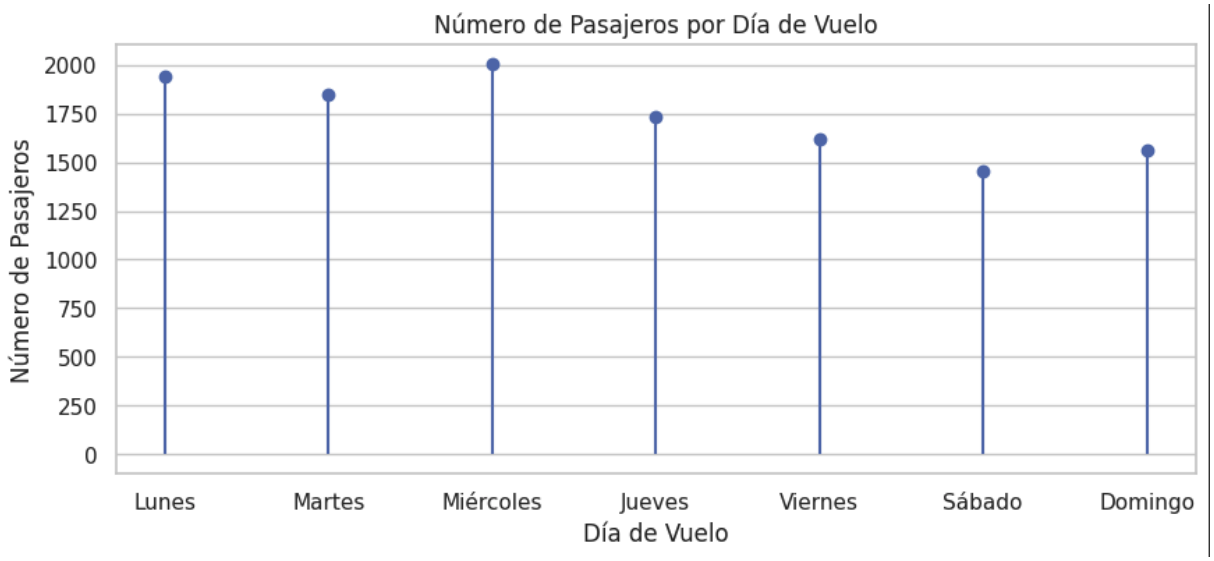

Figura 8 - Flujo de pasajeros durante los días de la semana.

La figura 8 muestra que el flujo más alto de pasajeros ocurre el día miércoles, en segundo lugar el lunes y en tercer lugar el martes. Esto indica que la mayor cantidad de pasajeros viajan en días laborables lo cual sugiere que los bajos puntajes pueden estar relacionados a esto. Para respaldar esta información analizaremos las horas, considerando el intervalo de 09:00 AM A 17:00 PM como horas laborables, y por lo tanto inconvenientes para el cliente.

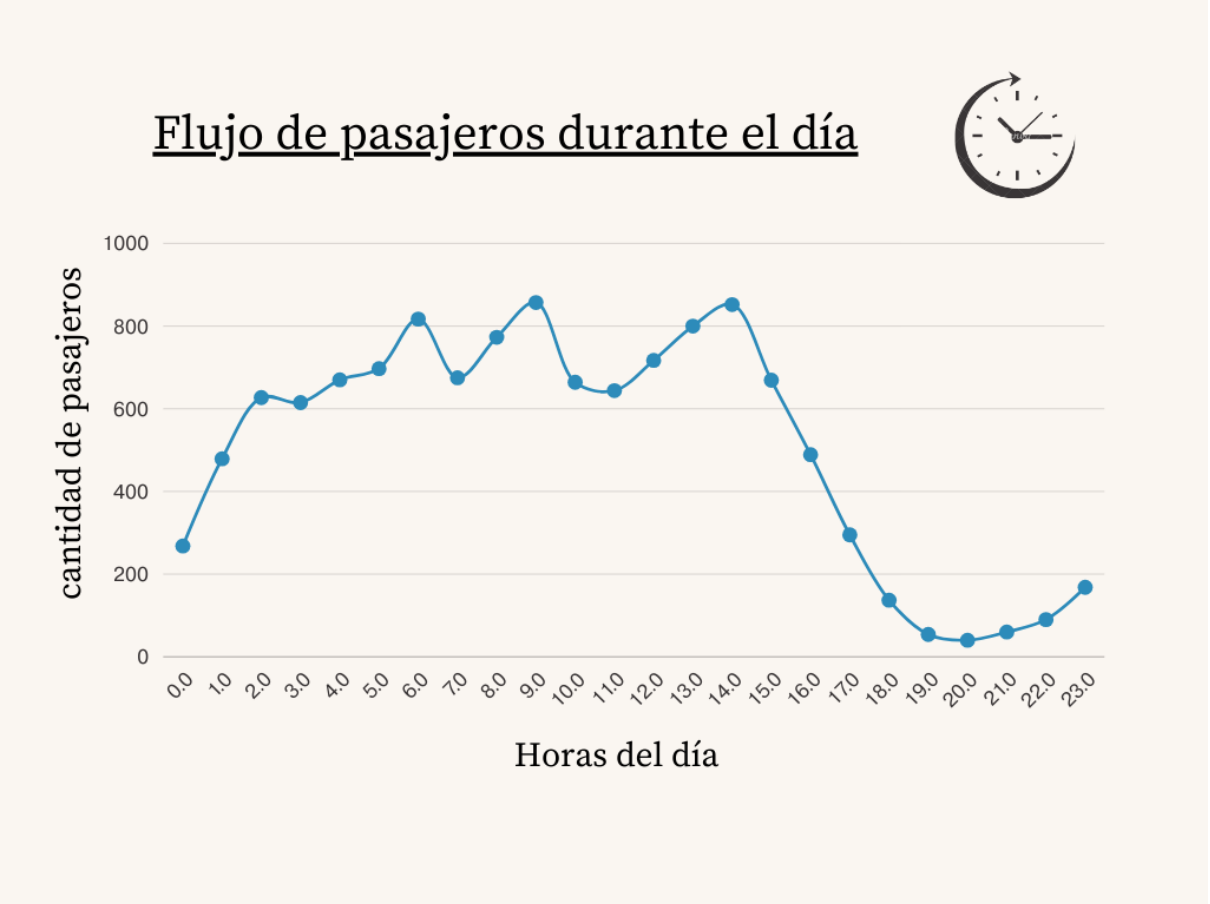

Figura 9 - Flujo de pasajeros durante las horas del día.

Conclusión:

La siguiente figura, referenciada como figura 10, nos permite ver que los días hábiles tienen un mayor flujo de pasajeros que los no hábiles. Esto nos indica que los viajeros, especialmente aquellos originarios de países con un volumen alto de compra de pasajes y que constituyen la mayor parte de los clientes de la aerolínea, viajan los días laborables y en horas laborables. Esto podría significar un quiebre en su rutina, lo cual se ve reflejado en los puntajes.

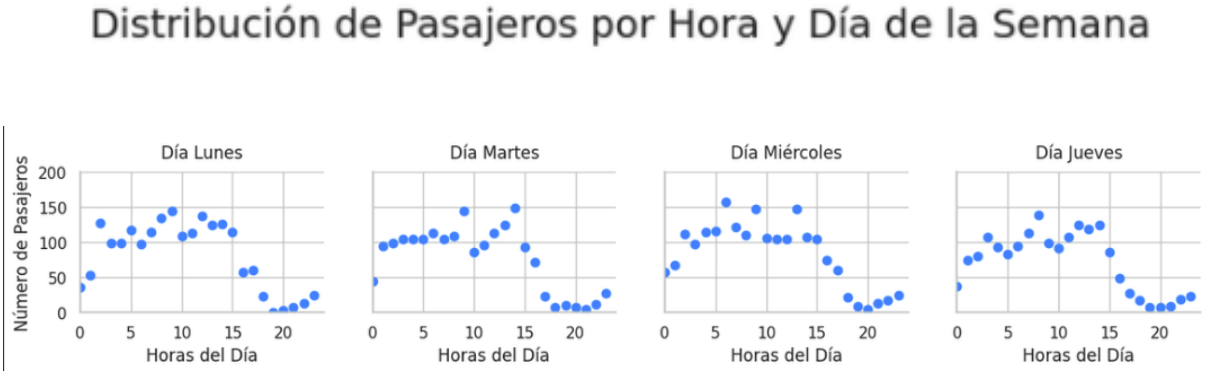

Figura 10 — Cantidad de pasajeros durante la semana y las horas del dia

Figura 10 — Cantidad de pasajeros durante la semana y las horas del día

Hipótesis 5:

¿ El motivo del viaje, determinado por la época del año y el destino, en los puntajes de los clientes de la aerolínea?

Archivos usados

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de customer_airways_data.csv. y cleaned_reviews.csv

Desarrollo:

La figura 8 nos permite ver que el día miércoles fue el día de mayor concurrencia de pasajeros, sumado a esto, como se evidencia en la figura 10, los días laborables suelen ser mucho más concurridos que los no laborables. Esto sugiere que los clientes de la aerolínea se ven perjudicados por estos horarios, aunque existe la posibilidad de que los viajes correspondan a períodos de vacaciones. Para poder comprobar cual de los dos casos es el que mas sucede dividimos el año en cuatro trimestres y elegimos los cinco destinos más visitados. Una vez realizado esto tomamos como parámetro que si el viaje sucede en un trimestre correspondiente a la época con el clima más agradable en el lugar de llegada se asumirá que no fue con fines laborables, en caso contrario se asumirá que si lo fue.

A fin de ver los valores más representativos para la aerolínea se seleccionaron los cinco destinos más visitados dentro del periodo de años, 2018-2023, dados por la cátedra. Esto nos permitirá ver si estos destinos están relacionados con los países con mayor cantidad de compra de pasajes y si el trimestre en el cual se realizó el viaje explica los bajos puntajes de dichos países.

Conclusión

Tal como se puede visualizar en la figura 11, Australia es uno de los destinos más elegidos, principalmente a las ciudades de Gold Coast y Melbourne. La elección de Australia como destino más visitado está respaldada por la figura 3, la cual lo cataloga como un país con un alto índice de satisfacción según los clientes.

Respecto a si los clientes suelen viajar en periodos vacacionales la información presentada indica que hay una mayor cantidad de viajeros, especialmente hacia Australia, durante el tercer trimestre del año en el cual es invierno en esa región, considerado como el mejor clima para conocer dicho país. Por lo cual se podría suponer que su alto puntaje está relacionado a que la mayor parte de sus arribos sucedieron en esa época.

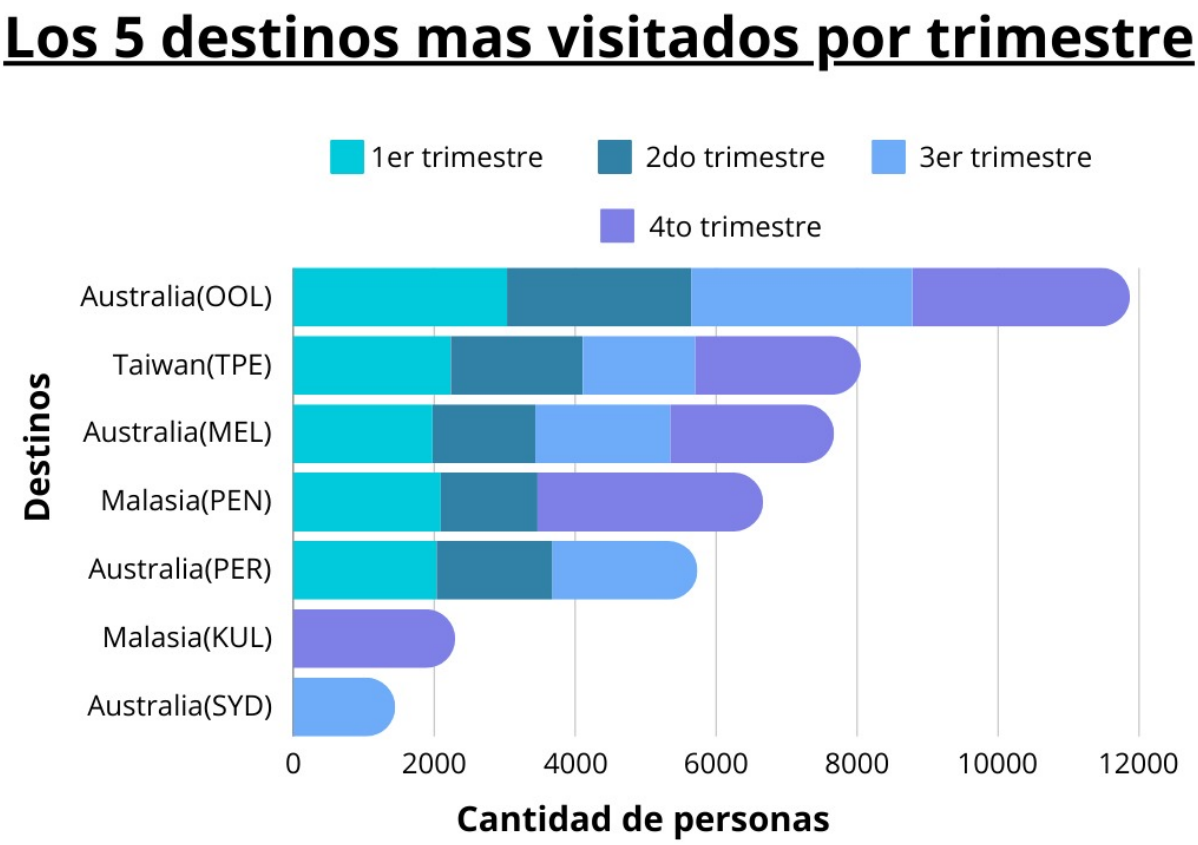

Figura 11 — Destinos más elegidos según los distintos trimestres del año entre (2018-2023).

Hipótesis 6:

¿La verificación y la duración de los vuelos tiene algún impacto en la satisfacción de los clientes en los países con mayor compra de pasajes?

Archivos usados

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de customer_airways_data.csv. y cleaned_reviews.csv

Desarrollo:

Si bien uno de los motivos de la falta de altos puntajes en los países con mayor compra es la época del año en la cual se viaja no podemos suponer que es el único. Por este motivo analizaremos si la cantidad de viajeros verificados y sus respectivas horas de vuelo nos brinda más información.

Para empezar este análisis realizamos un promedio de los puntajes en los usuarios verificados y los no verificados de los diez países con mayor compra de pasajes, tal como se ve en la figura 12 y 13 respectivamente. Estos gráficos nos permiten ver que tanto los clientes verificados como los no verificados presentan, usualmente, puntajes entre 4 y 5.5. Un dato a resaltar es que los pasajeros verificados muestran un rango de puntajes más compacto, entre 2 a 5, mientras que los no verificados presentan todo tipo de puntajes, desde un 1 correspondiente a Malasia hasta un 9 para Taiwán. Esto podría indicar que los pasajeros verificados suelen ser más críticos y constructivos a la hora de dar una respuesta que los que no lo están.

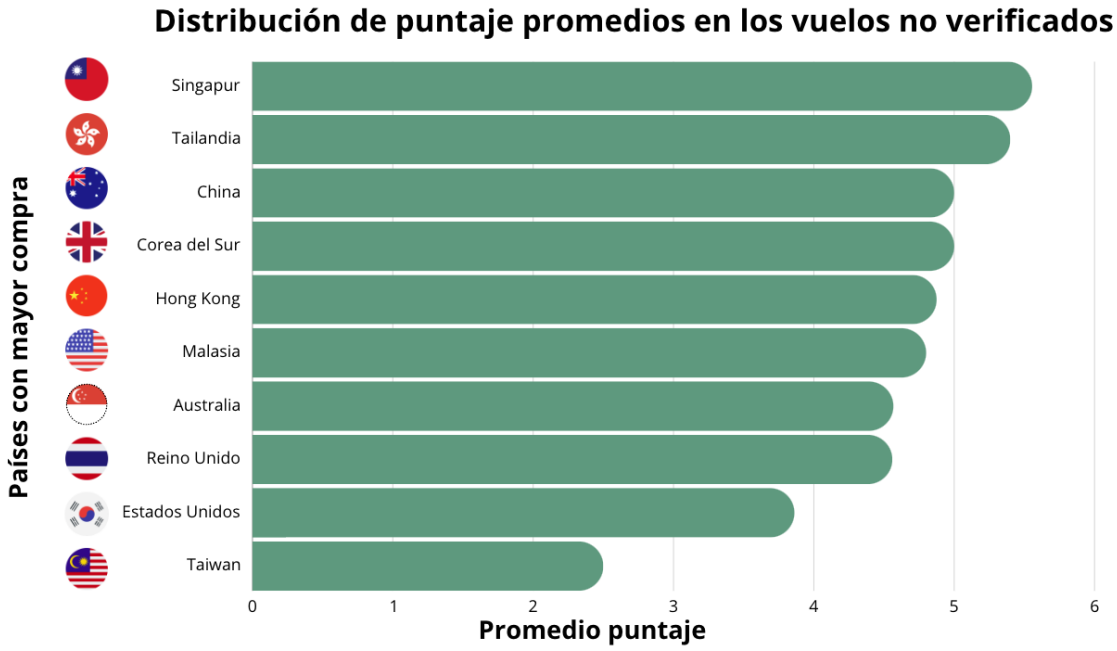

Figura 12 — Relación entre los puntajes promedios de los usuarios verificados y los países con mayor compra

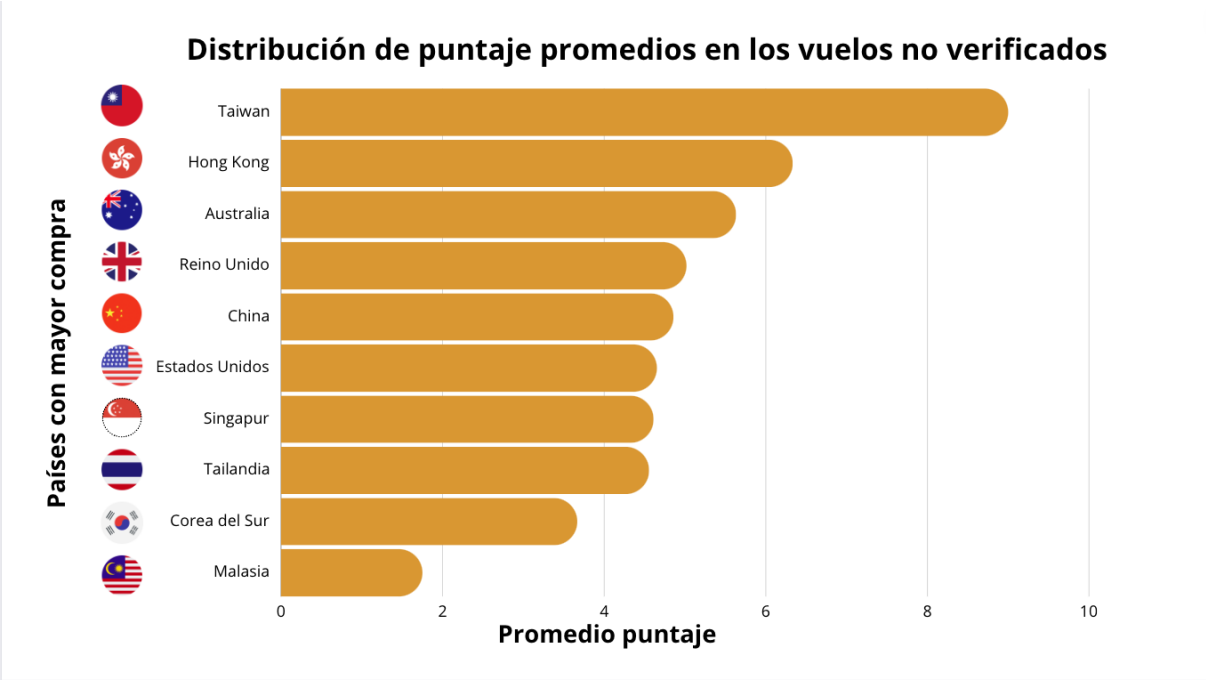

Figura 13— Relación entre los puntajes promedios de los usuarios no verificados y los países con mayor compra

Conclusión

Luego, clasificamos las horas de vuelo en cuatro categorías distintas, para poder hacer esto nos basamos en los cuartiles del promedio de horas de vuelo, presentes en la figura 9. Estas categorías se extienden desde “muy bajo” hasta “alto”, siendo “muy bajo” todo lo menor e igual al primer cuartil, y “alto” lo mayor al tercero.

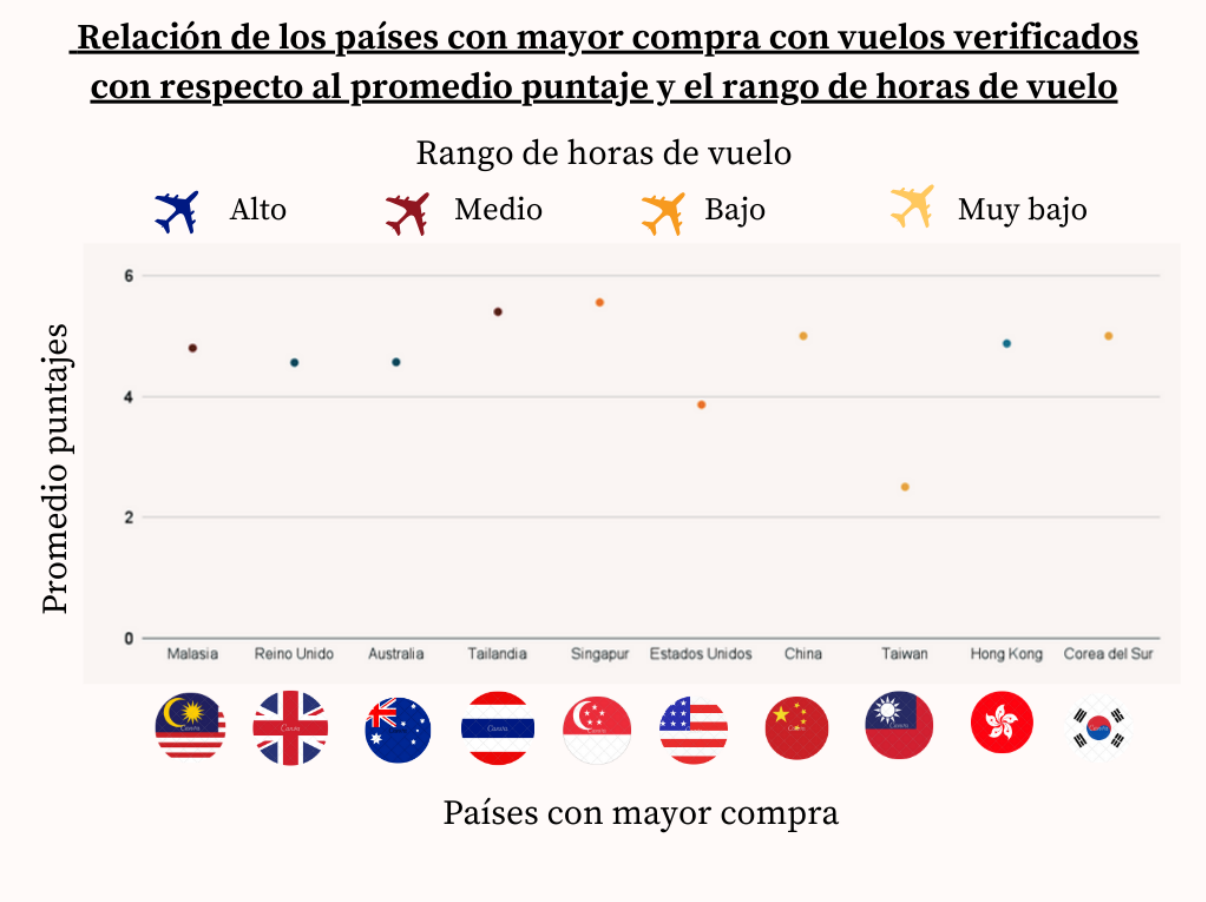

Figura 14 -Relación de los países con mayor compra con respecto al promedio puntaje y el rango de horas de vuelos verificados

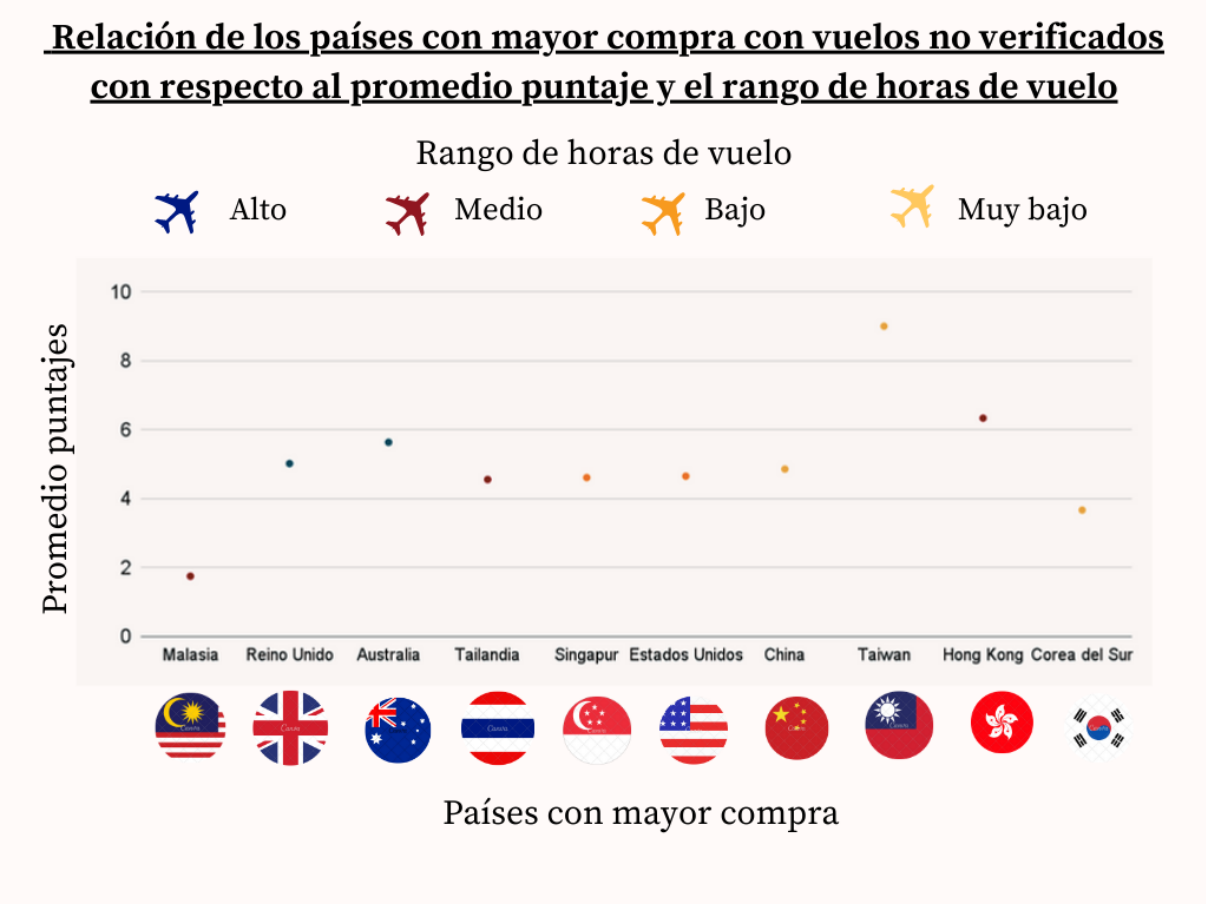

Figura 15— Relación de los países con mayor compra con vuelos no verificados con respecto al promedio puntaje y el rango de horas de vuelo.

Tal como podemos ver en la figura 14, los puntajes promedio están dispersos entre los rangos de horas de vuelo, es importante notar que los rangos muy bajo y bajo incluyen los puntajes más altos, China y Singapur, lo que podría sugerir que vuelos con menos horas de vuelo están relacionados con una alta satisfacción. Sin embargo también hay un puntaje, Taiwán con una calificación de dos y medio, el cual no permite afirmar esto. Los países en el rango "Alto" tienen puntajes de aproximadamente cinco, esto podría indicar que, a pesar de la mayor duración de los vuelos, la satisfacción de los usuarios no decae.

En la figura 15 sucede algo bastante similar, con la salvedad que los puntajes presentes en los rangos “Muy bajo” y “Bajo” sufren una polarización de los puntajes mucho más pronunciada variando desde tres a nueve.

Teniendo en cuenta toda la información presentada este análisis nos indica que a pesar de esperarse que los vuelos más cortos siempre tengan mejores evaluaciones los datos muestran que un alto puntaje puede estar relacionado a esto pero que no es exclusivo a esta condición.

Hipótesis 7:

¿Existe alguna relación entre las reseñas y los puntajes? ¿Las reseñas negativas están vinculadas a los servicios ofrecidos por la aerolínea?

Archivos usados

Para el análisis exploratorio de datos de esta hipótesis utilizamos los archivos de cleaned_reviews.csv y customer_airways_data.csv.

Desarrollo:

Para este análisis, se hicieron modificaciones a los archivos, solo se tomaron en cuenta las encuestas verificadas de los diez países más representativos, mencionados en reiteradas ocasiones en todo el informe. Una vez hecho esto se hizo un análisis de las reseñas, se quitaron las palabras más comunes del lenguaje, también conocidas como Stop-Words, y se llevó cada palabra a su sustantivo. Esto último se logró hacer a través de la lemmatizacion de ellas.

Luego de limpiar y procesar los datos, realizamos un análisis de sentimiento de los mismos. Para el análisis se utilizó los puntajes de polaridad en los sentimientos de cada una de las reseñas, se eligió este método por el motivo que permite cubrir todas las posibilidades de sentimientos, lo cual no sería posible si se utilizara expresiones regulares ya que presentan la limitación del conocimiento de quien las escribe. A partir del análisis recién descrito seleccionamos las positivas, las cuales fueron representadas con el número uno, y las negativas, representadas con el número cero. Los comentarios neutrales fueron descartados debido que solo se busca analizar las reseñas que presentan una opinión informativa de la experiencia del cliente.

Como se puede ver en la figura 16, se realizó un promedio del sentimiento de las reseñas por país en el cual se visualiza que en estas los clientes de la aerolínea fueron más radicales que en los puntajes. Se llega a esta conclusión ya que tanto Malasia como Corea del Sur presentan puntajes “muy bajo” y “bajo” respectivamente pero en las reseñas no presentan comentarios negativos. Lo mismo sucede con Taiwán el cual presenta un puntaje “medio” pero solo posee reseñas negativas. Esto sugiere que la cantidad de reseñas no es lo suficientemente representativa, como sí sucede con los puntajes.

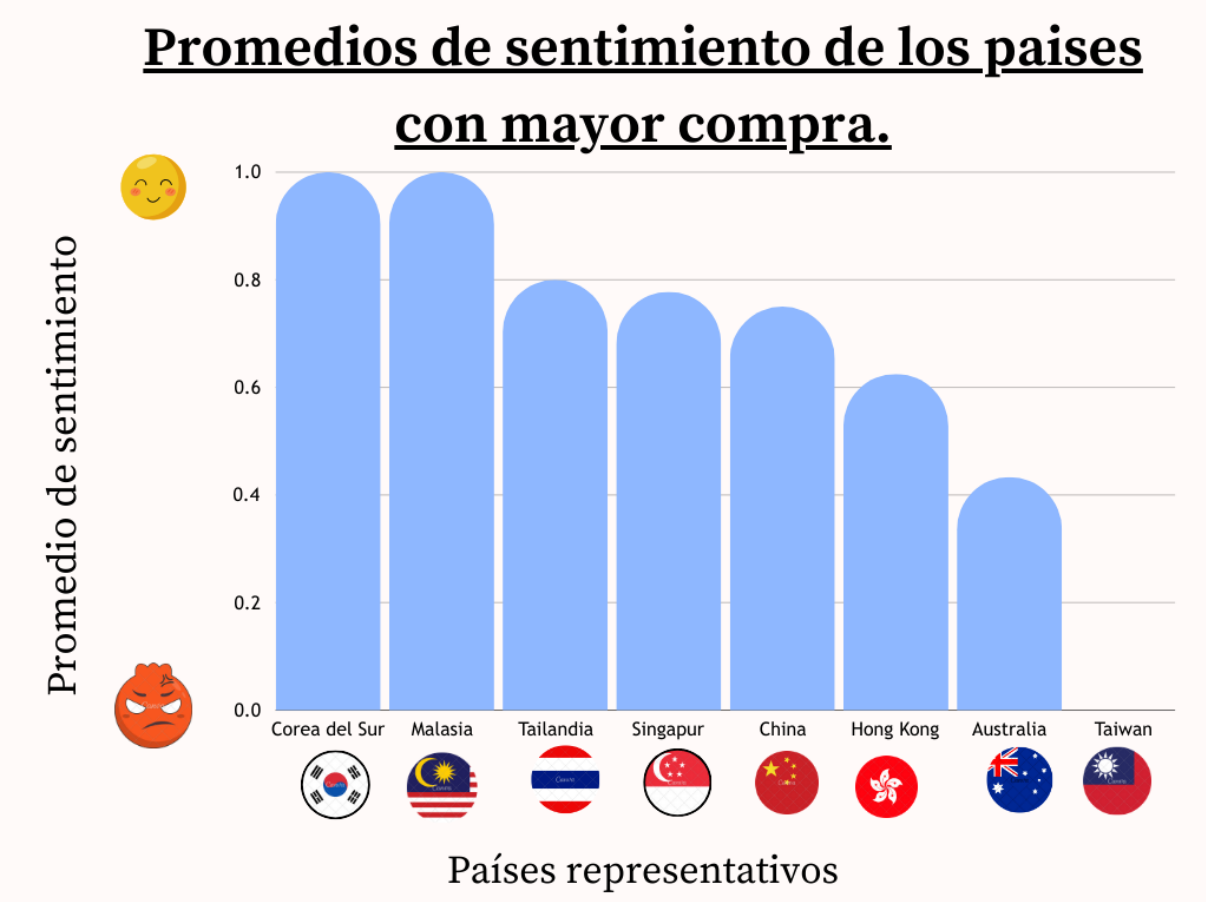

Figura 16 — Promedio del sentimiento en las reseñas por país .

Para verificar si esto es así realizaremos un análisis de distribución de la cantidad de reseñas, mostrado en la figura 17, En este se puede observar que la cantidad presentes en las encuestas son muy reducidas en comparación a la cantidad de puntajes, exceptuando dos casos, representados como Outliers. Logramos identificar a Reino Unido con 553 reseñas y a Estados Unidos con 136 como los Outliers al hacer uso del rango intercuartílico y del tercer cuartil.

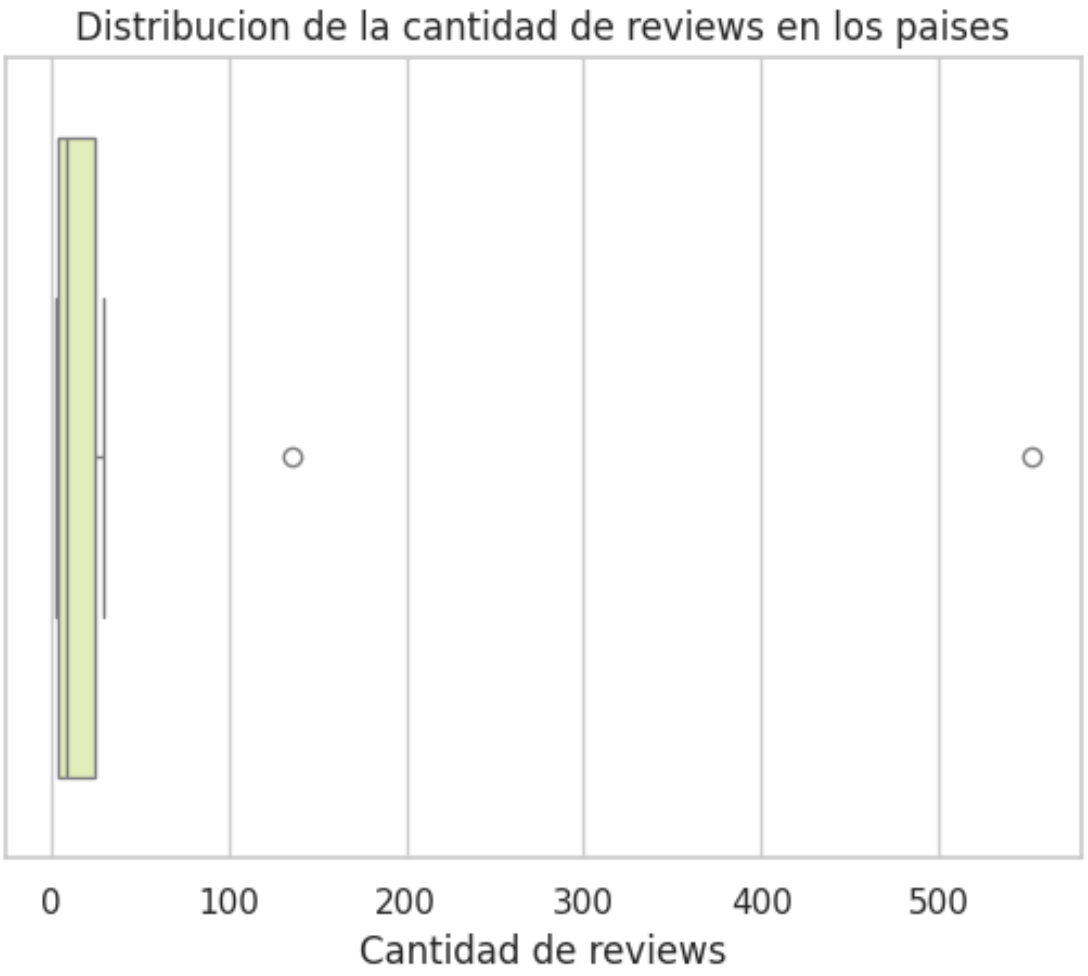

Figura 17— Distribución de reseñas en los países

Dado que los Outliers no permiten ver una representación informativa y concisa de la distribución de las reseñas en las encuestas, se filtraron. Esto nos permitió ver que los países presentan, a lo sumo, treinta reseñas tal como se puede ver en la figura 18. Si bien esto verifica que la cantidad de reseñas presente es considerablemente menor a la cantidad de puntajes ingresados no nos permite ver cuantas reseñas tiene cada país particularmente. Con el fin de poder visualizar esta información y confirmar que la variación en sentimiento entre reseñas y puntajes se debe a la poca cantidad de la primera se realizó la figura 19. En esta figura se volvieron a filtrar los Outliers con fines de poder representar correctamente la cantidad de reseñas de cada uno de los países.

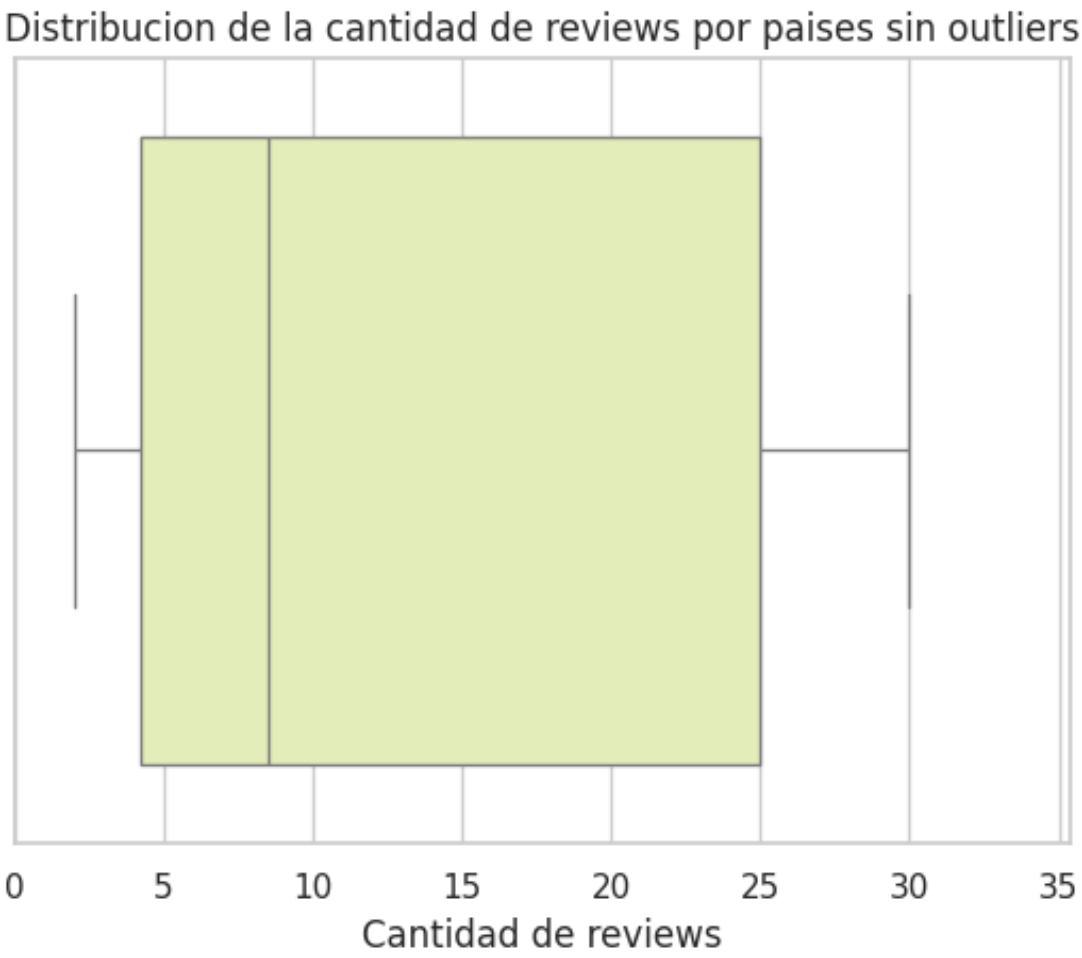

Figura 18 — Distribución de reseñas en los países ,sin los Outliers.

Figura 19 — Cantidad de reseñas mas representativos sin los Outliers.

Teniendo en cuenta la cantidad de reseñas, explicada y detallada anteriormente, pasaremos a analizar solo las negativas. Al seleccionar solo las reseñas negativas resultaron filtrados Malasia y Corea del Sur por el motivo que solo poseen reseñas positivas, como se observa en la figura 16. En primer lugar calculamos un puntaje de negatividad para cada reseña al hacer uso de la polarización de los comentarios, mencionada anteriormente, luego se filtraron la cantidad de reseñas al obtener las primeras 29 con mayor puntaje de negatividad. La elección de este número se hizo con el fin de poder hacer un análisis representativo para todos los países por igual, con esto en mente se seleccionaron, por país, una cantidad de reseñas hasta el 75% del total. El motivo de la elección de este porcentaje, que en este caso equivale a 29,como umbral mínimo ya fue explicado en la hipótesis 1, si bien en dicha hipótesis se usó como umbral mínimo en esta se utiliza como umbral máximo por la presencia de Outliers tan extremos.

Una vez realizado este proceso, se hizo un análisis de palabras, evidenciado en la figura 20, de todas las reseñas de cada uno de los países anteriormente referenciados. Dicho análisis cuenta con la presencia de los Outliers pero con la exclusión de Malasia y Corea del Sur por motivos previamente explicados.

Figura 20 — Palabras en reseñas negativas.

Como se puede visualizar en la figura 20 las reseñas negativas suelen estar relacionadas a demoras, el servicio o el personal de la aerolínea. Para poder tener un análisis más representativo de las quejas en primer lugar obtuvimos sólo las palabras que comunican una experiencia negativa, luego obtuvimos la reseña con el puntaje de negatividad más alto, es decir la que más cantidad de palabras negativas presentaba. Finalmente se utilizó el algoritmo de Count Vectorizer para seleccionar todas las reseñas similares a la más negativa.

Figura 21 — Palabras negativas en las reseñas de la aerolínea.

Conclusión:

Teniendo en cuenta todo lo anteriormente descrito se logró concluir que las opiniones negativas de los clientes, las cuales se pueden visualizar en la figura 21, en conjunto con las palabras presentes en la figura 20, confirman que los bajos puntajes y las reseñas negativas están relacionadas con el servicio, esto se evidencia en relaciones de las palabras como “poor” con “food” o “experience”, “delayed” o “uncomfortable” con “service”, “appalling” con “experience” o “meal”, entre otras relaciones visibles. También se descubrió que una cantidad considerable de quejas son respecto al personal de la aerolínea, tanto en el check-in como durante el vuelo.

En la siguiente visualización se observa la figura original del informe. Dicha visualización está dividida en dos páginas, en la primera se realiza un análisis de las palabras negativas presentes en las reseñas de los países Outliers, esto se hizo en una página aparte bajo el motivo que al tener una cantidad considerable de reseñas es importante resaltar las opiniones de los clientes sobre que mejorar, mientras que en la segunda página se encuentran los cuatro países restantes analizados, cada uno mostrando claramente que palabras utilizaron y su cantidad

https://e.infogram.com/adaf774b-0a8d-4f00-8f6c-3e5b63161836?src=embed"><div

Visualización original del informe