

Conjunto de datos de Aerolíneas

Alumno: Militello Gastón
TP : Entrega Final

Abstract

El conjunto de datos de aerolíneas proporciona una fuente valiosa de información para analizar y optimizar las operaciones de las aerolíneas y mejorar la experiencia del cliente. Este conjunto de datos abarca una amplia gama de variables, incluyendo datos demográficos de los pasajeros, información sobre rutas de vuelo y estados de vuelo. A continuación, se resumen los aspectos más destacados de este conjunto de datos:

1. VARIABLES

ID de pasajero :** identificador único para cada pasajero

Nombre - Nombre del pasajero

Apellido - Apellido del pasajero

Género - Género del pasajero

Edad - Edad del pasajero

Nacionalidad - Nacionalidad del pasajero

Nombre del Aeropuerto- Nombre del aeropuerto donde
abordó el pasajero

Código de país del aeropuerto :** código de país de la ubicación del aeropuerto

Nombre del país :** nombre del país en el que está ubicado el aeropuerto.

Continente del aeropuerto :** continente donde está situado el aeropuerto.

Continentes** - Continentes involucrados en la ruta del vuelo.

Fecha de salida** - Fecha de salida del vuelo

Aeropuerto de llegada** - Aeropuerto de destino del vuelo

Nombre del piloto :** nombre del piloto que opera el vuelo.

Estado del vuelo :** estado actual del vuelo (p. ej., puntual, retrasado, cancelado)

2. Preguntas de Investigación:

¿Como se contribuyen los vuelos según su estado?

¿Cuál es la nacionalidad más común entre los pasajeros en vuelos internacionales?

¿Se pueden identificar patrones de preferencia de vuelo basados en el género de los pasajeros?

¿Pilotos con mas vuelos demorados?

¿Los dos países mas elegidos y su distribución por genero Genero?

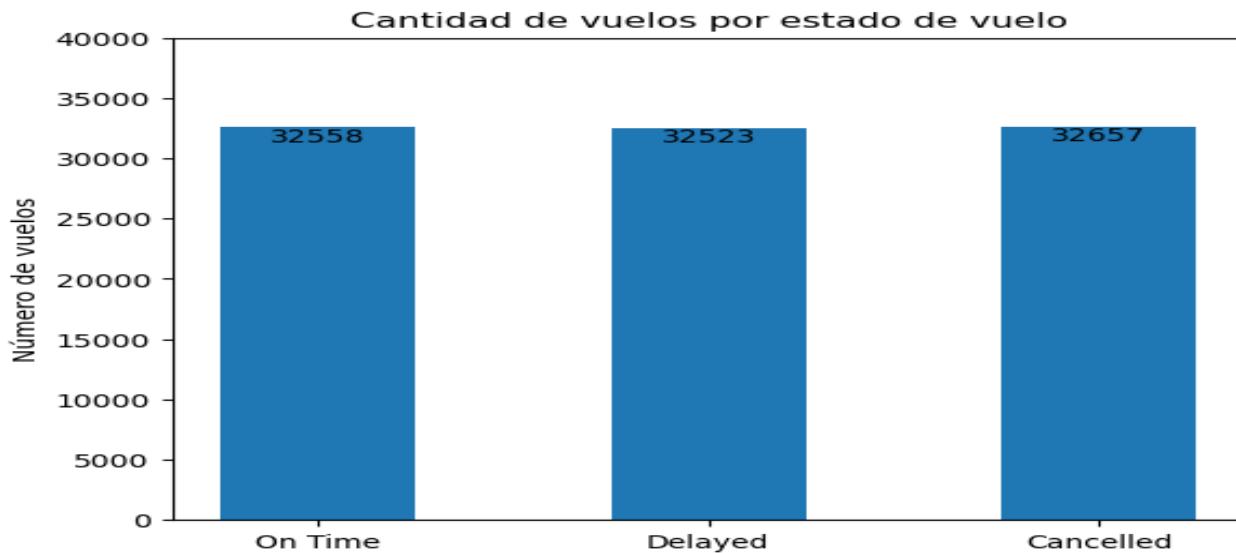
2. Hipótesis de Investigación:

Hipótesis nula: No hay una diferencia significativa en la tasa de cancelación de vuelos entre aeropuertos de diferentes continentes. Hipótesis alternativa: La tasa de cancelación de vuelos varía significativamente según el continente de ubicación del aeropuerto.

Hipótesis nula: No hay diferencia en las preferencias de vuelo entre géneros. Hipótesis alternativa: Los pasajeros de género masculino y femenino tienen preferencias de vuelo significativamente diferentes.

Hipótesis nula: La nacionalidad de los pasajeros es independiente de la ruta de vuelo. Hipótesis alternativa: Algunas rutas de vuelo tienen una alta proporción de pasajeros de una nacionalidad específica.

ANÁLISIS EXPLORATORIO



ANÁLISIS EXPLORATORIO

Las 5 nacionalidades más comunes entre los pasajeros en vuelos internacionales son:

China 18160

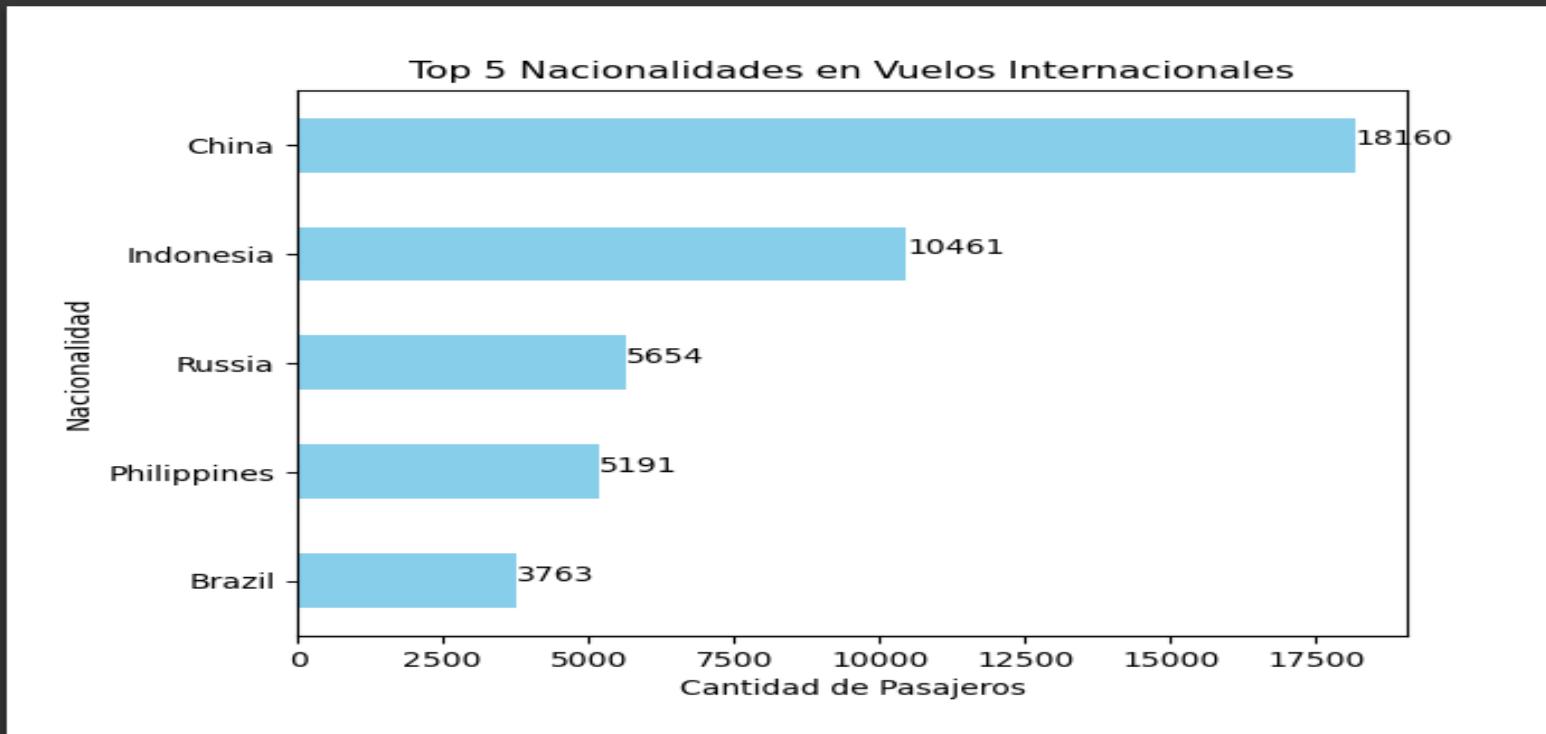
Indonesia 10461

Russia 5654

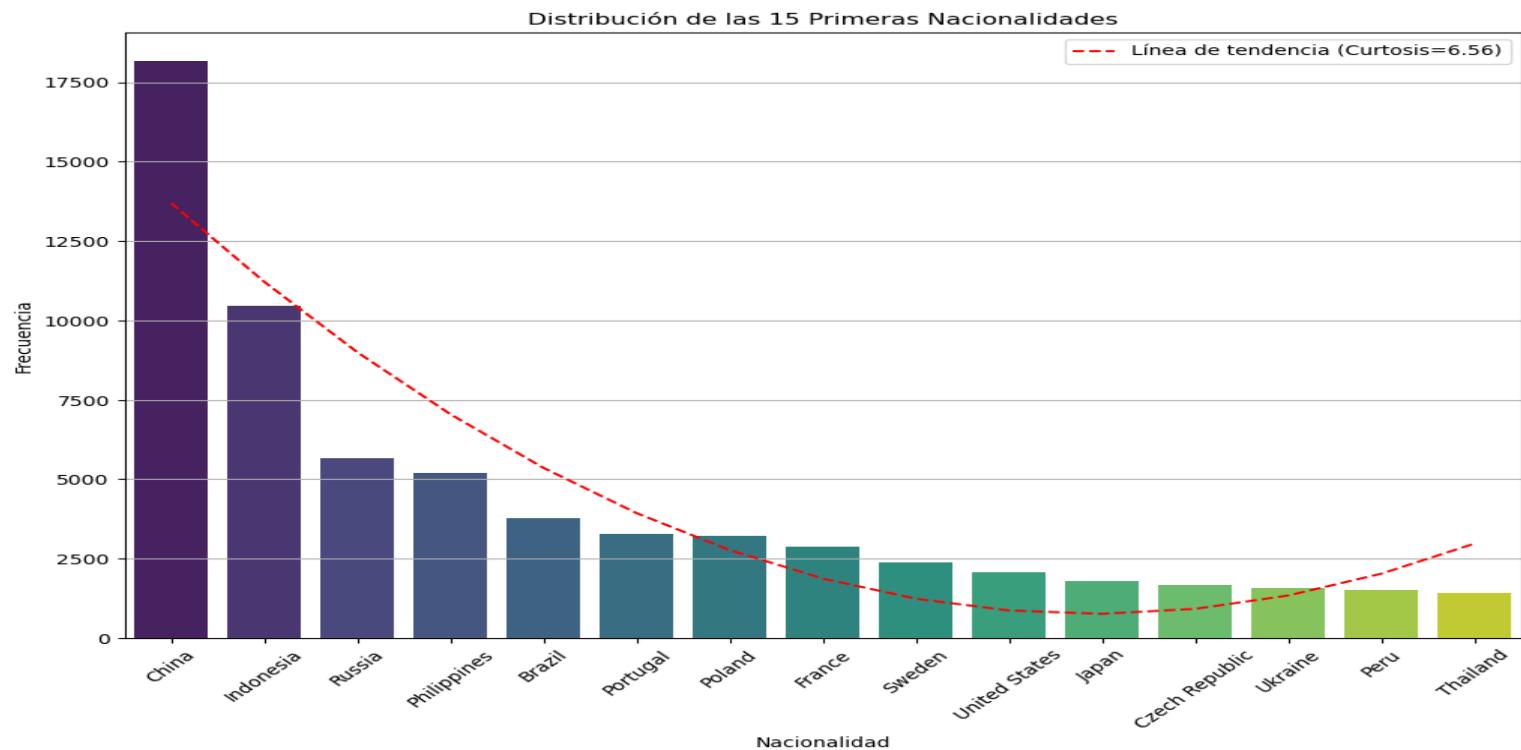
Philippines 5191

Brazil 3763

ANÁLISIS EXPLORATORIO



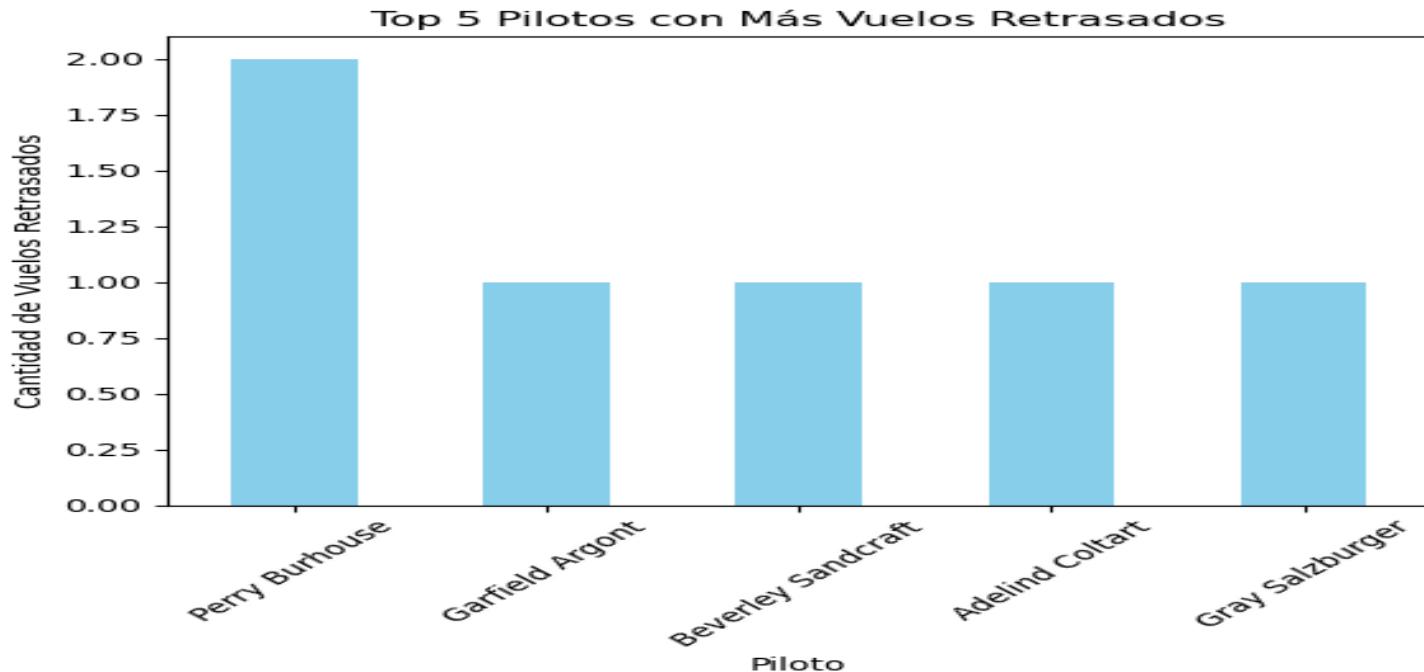
ANÁLISIS EXPLORATORIO



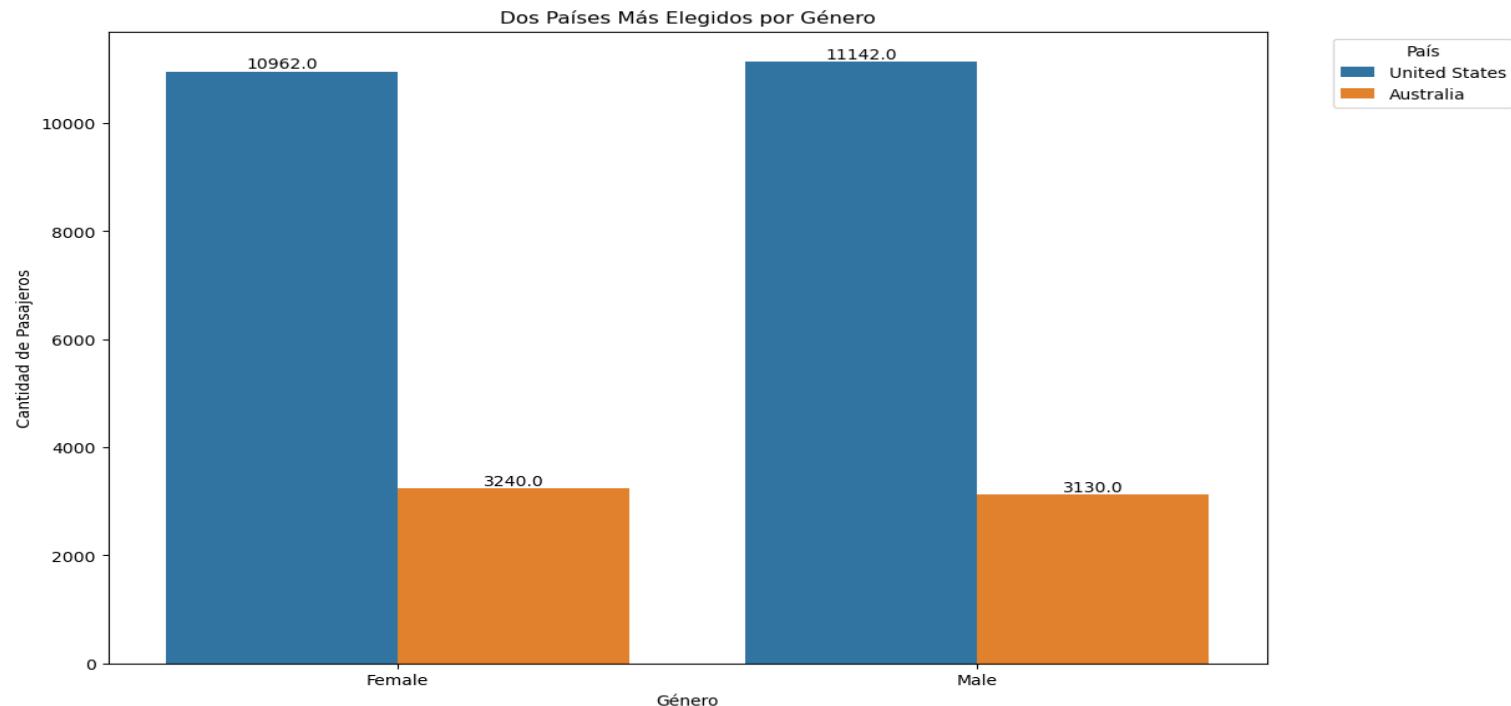
ANÁLISIS EXPLORATORIO

Una curtosis de 6.56 indica que la distribución de frecuencia de las nacionalidades tiene colas más pesadas y es más puntiaguda que la distribución normal, lo que sugiere la presencia de valores extremos o una mayor concentración de datos alrededor de la media con colas largas en comparación con una distribución normal.

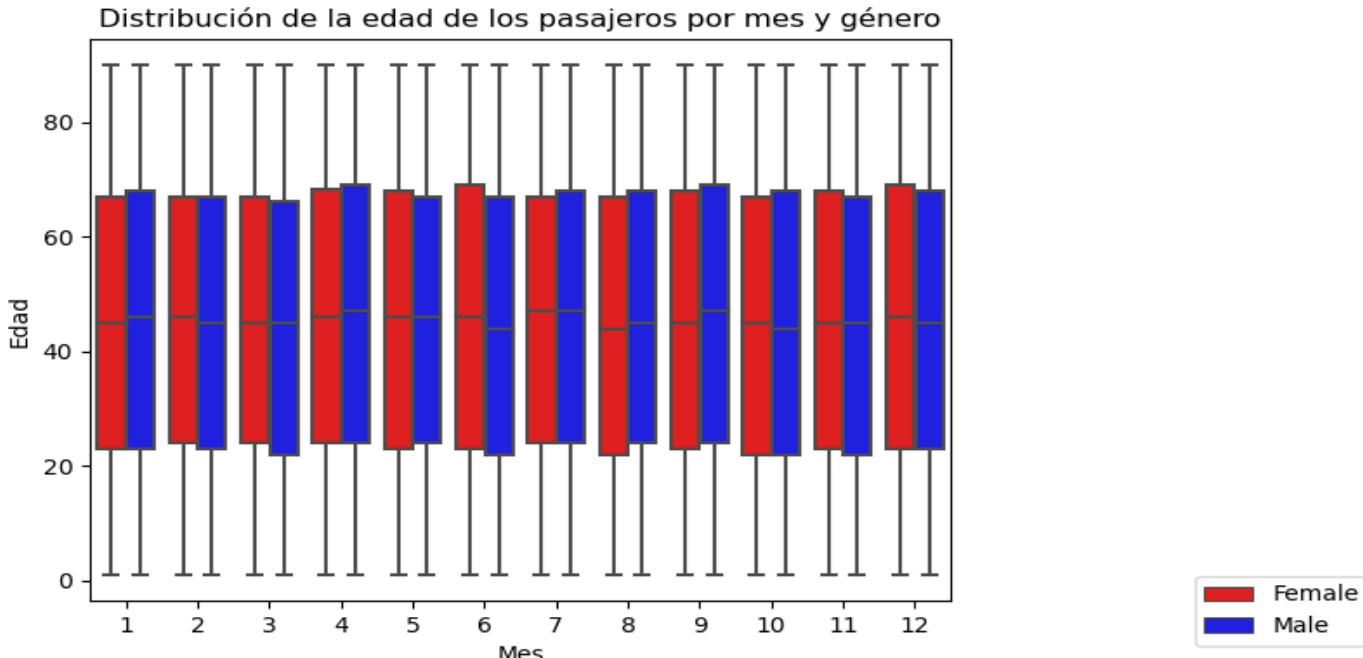
ANÁLISIS EXPLORATORIO



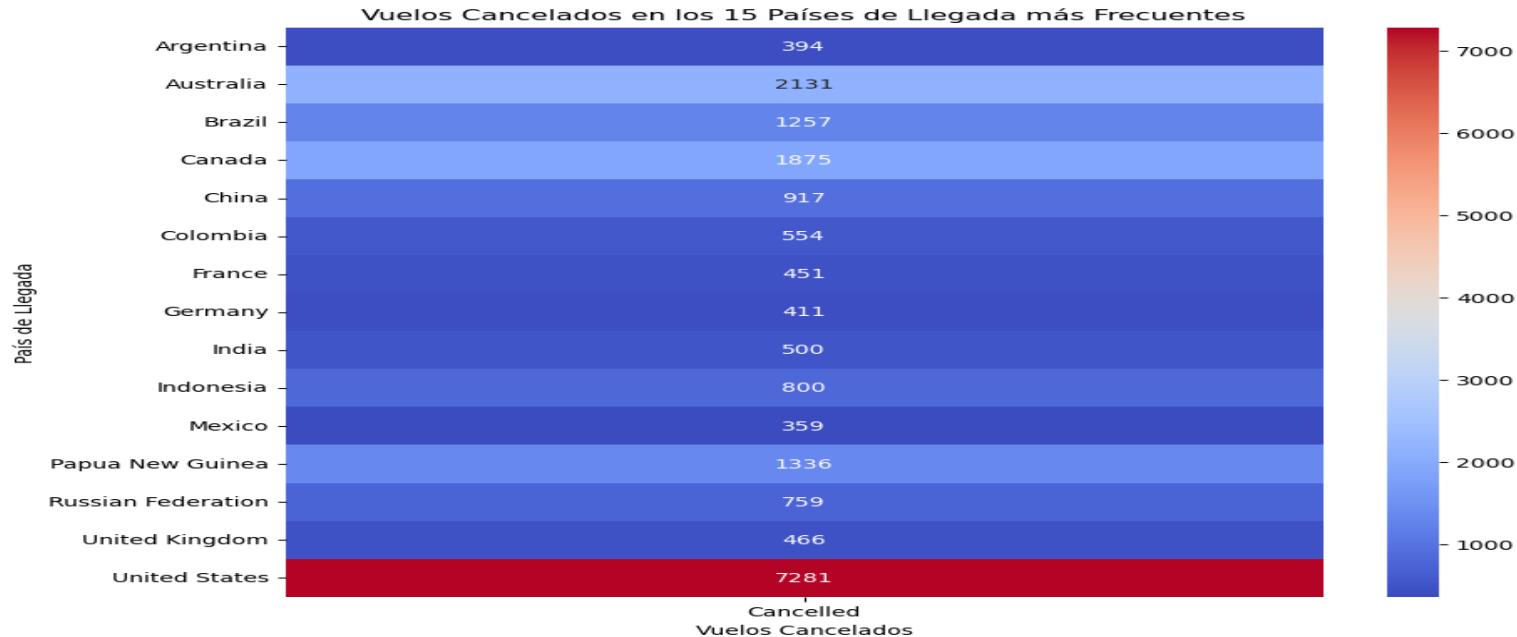
ANÁLISIS EXPLORATORIO



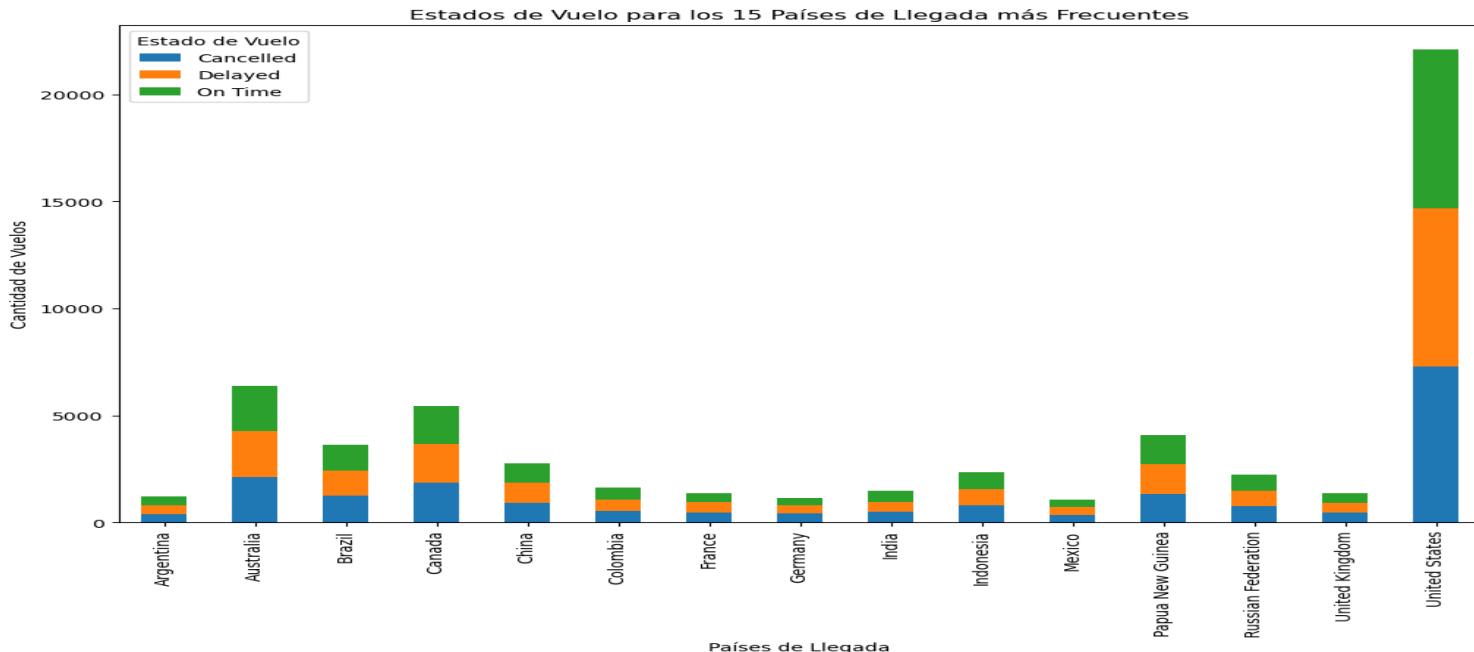
ANÁLISIS EXPLORATORIO

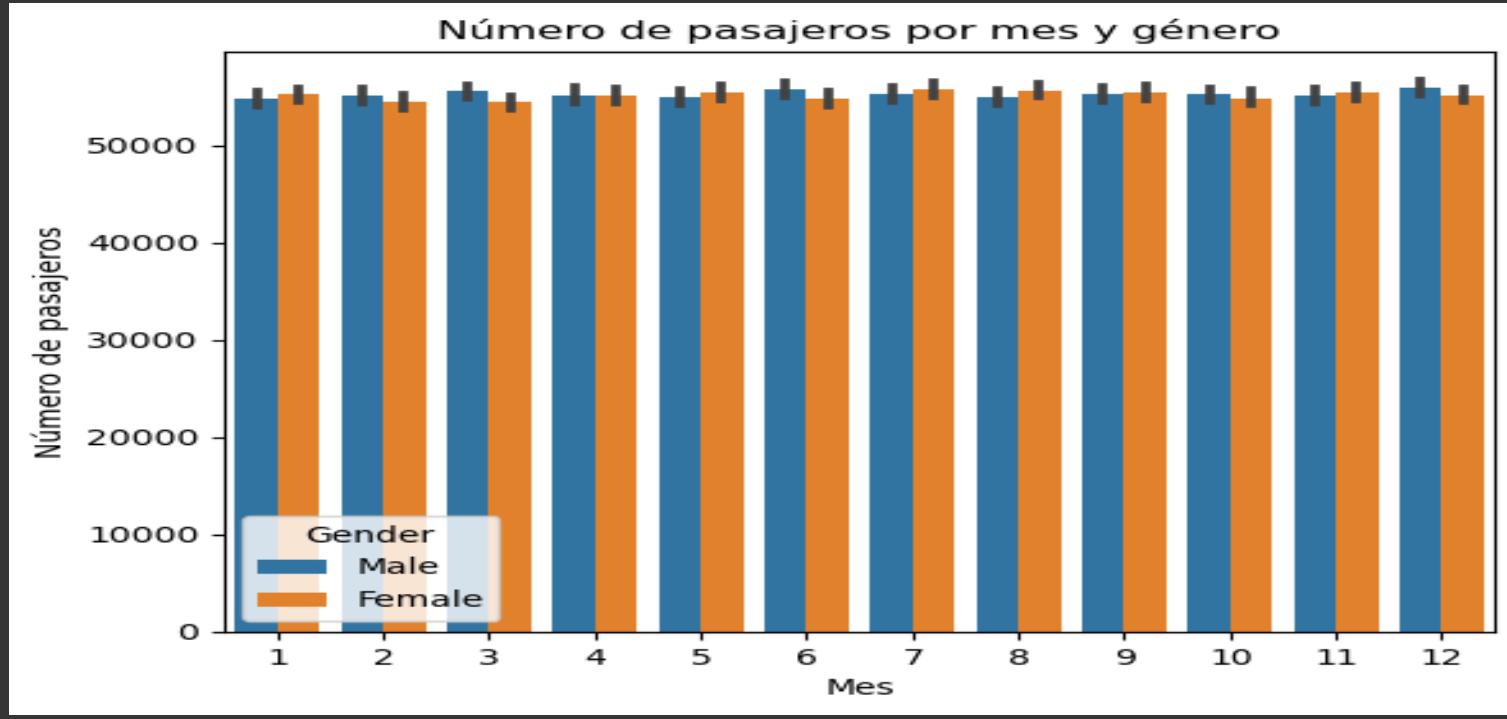


ANÁLISIS EXPLORATORIO

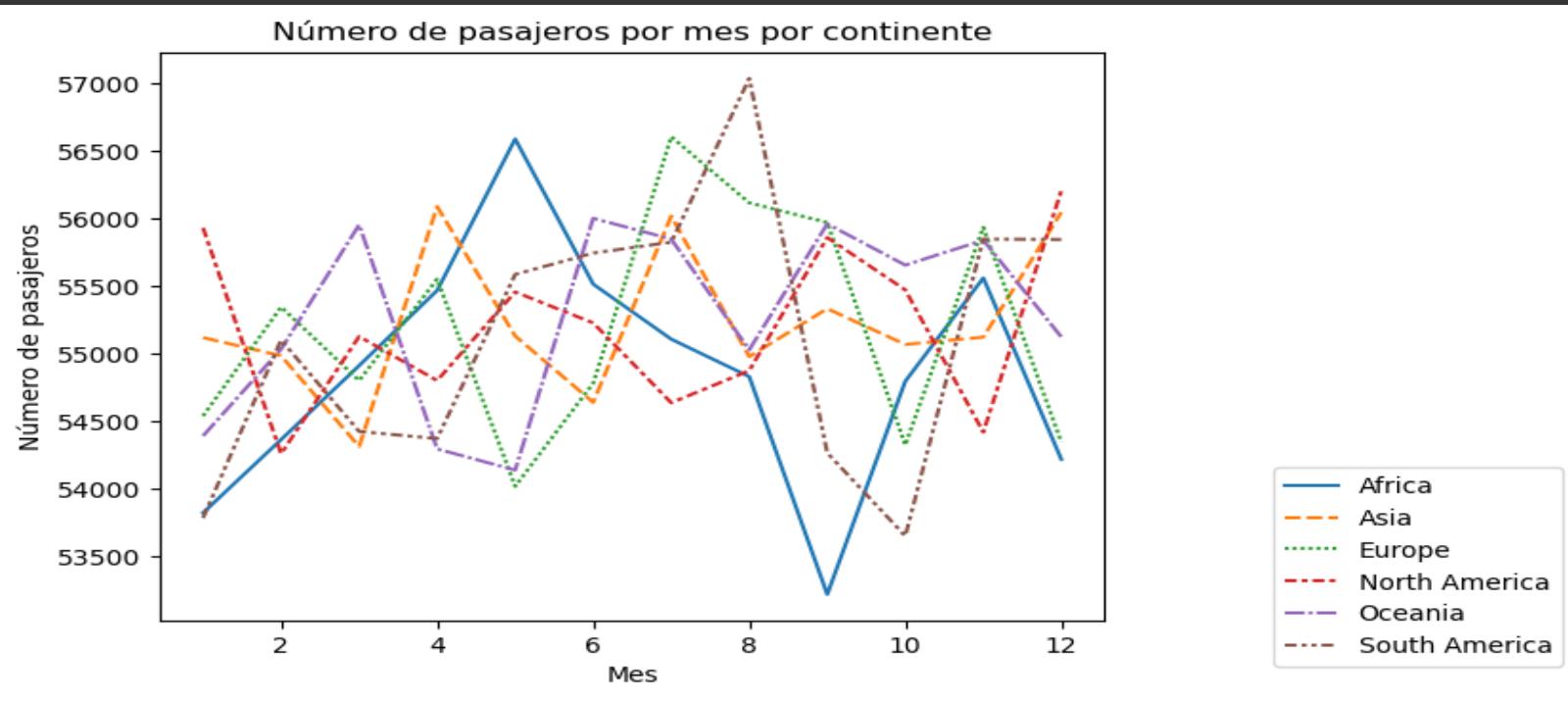


ANÁLISIS EXPLORATORIO





No vemos diferencia entre Géneros en los distintos Meses del Año



Este grafico muestra la cantidad de pasajeros por mes en los distintos continentes podemos ver que en todos los continentes hay mucha variación en los distintos meses

ANÁLISIS EXPLORATORIO

Hipótesis nula: No hay una diferencia significativa en la tasa de cancelación de vuelos entre aeropuertos de diferentes continentes. Hipótesis alternativa: La tasa de cancelación de vuelos varía significativamente según el continente de ubicación del aeropuerto.

Dado que el valor p (0.918) es mucho mayor que el nivel de significancia típico de 0.05, no hay suficiente evidencia para rechazar la hipótesis nula. En otras palabras, no hay una diferencia significativa en la tasa de cancelación de vuelos entre aeropuertos de diferentes continentes, según los datos proporcionados

ANÁLISIS EXPLORATORIO

Hipótesis nula: La nacionalidad de los pasajeros es independiente de la ruta de vuelo.

Hipótesis alternativa: Algunas rutas de vuelo tienen una alta proporción de pasajeros de una nacionalidad específica.

Un valor p de 0.3211 indica que no hay suficiente evidencia para rechazar la hipótesis nula a un nivel de significancia del 0.05. En otras palabras, no hay suficiente evidencia para afirmar que la nacionalidad de los pasajeros está asociada de manera significativa con la ruta de vuelo (o la combinación de aeropuertos de salida y llegada) .

Podemos observar que no existe diferencia notoria en los distintos estados de los vuelos, esto debe llevar a un plan de acción para reducir la cantidad de vuelos demorados y cancelados.

Las 5 nacionalidades más comunes entre los pasajeros en vuelos internacionales son:

China 18317

Indonesia 10559

Russia 5693

Philippines 5239

Brazil 3791

Las 5 nacionalidades menos comunes entre los pasajeros en vuelos internacionales son:

Romania 2

British Virgin Islands 2

Jersey 1

Norfolk Island 1

Sint Maarten 1

También tenemos países más elegidos y su distribución por género.

Esta información es muy valiosa para el departamento de marketing .

Informamos los Pilotos con mayor cantidad de vuelos cancelados, no es un número significativo , pero se debe prestar atención.

Las hipótesis planteadas , nos muestran que debemos obtener más datos para poder tener resultados más favorables.

INSIGHTS & RECOMENDACIONES

Países de Llegada Más Frecuentes: Identificamos los 15 países de llegada más frecuentes, lo que nos permite focalizar nuestros análisis en estos destinos principales.

Estado de los Vuelos: Observamos la distribución de los estados de vuelo ('On Time', 'Delayed', 'Cancelled') para estos países de llegada más frecuentes. Esto nos brinda información sobre la puntualidad y la frecuencia de los vuelos en cada destino.

Vuelos Cancelados por País: Mediante el análisis de los vuelos cancelados en los 15 países de llegada más frecuentes, podemos identificar los destinos con mayor incidencia de cancelaciones. Esto podría indicar posibles problemas operativos o condiciones adversas en esos destinos.

Total de Vuelos Arribados: Además, al incluir el total de vuelos arribados en el gráfico, obtenemos una perspectiva completa de la cantidad de vuelos que llegan a estos destinos y cómo se comparan con los vuelos cancelados.

Relación entre Países de Llegada y Estado de los Vuelos: El análisis de la relación entre los países de llegada y el estado de los vuelos nos ayuda a comprender mejor la puntualidad y la eficiencia operativa en diferentes destinos, lo que puede ser útil para planificar y optimizar rutas y operaciones de vuelo.

Estos insights nos brindan una comprensión más profunda de la dinámica de los vuelos hacia los países de llegada más frecuentes y pueden ser útiles para tomar decisiones informadas en la gestión y planificación de operaciones aéreas.

Conclusiones y recomendaciones:

Podemos observar que no existe diferencia notoria en los distintos estados de los vuelos, esto debe llevar a un plan de acción para reducir la cantidad de vuelos demorados y cancelados.

Las 5 nacionalidades más comunes entre los pasajeros en vuelos internacionales son: China 18317 Indonesia 10559 Russia 5693 Philippines 5239 Brazil 3791 .

Las 5 nacionalidades menos comunes entre los pasajeros en vuelos internacionales son: Romania 2 British Virgin Islands 2 Jersey 1 Norfolk Island 1 Sint Maarten 1

También tenemos países más elegidos y su distribución por género.

Esta información es muy valiosa para el departamento de marketing .

Informamos los Pilotos con mayor cantidad de vuelos cancelados, no es un número significativo , pero se debe prestar atención. Las hipótesis planteadas , nos muestran que debemos obtener más datos para poder tener resultados más favorables.

**SE RECOMIENDA PRESTAR MUCHA ATENCIÓN Y CORREGIR
LOS VUELOS CANCELADOS Y DEMORADOS.**

Entrenamiento y optimización de modelos de Machine Learning

Luego del Análisis del dataset y ver las características del Negocio , propusimos poder clasificar por estado del vuelo. Para esto se compararon los algoritmos que hemos visto para saber cual es el de mayor rendimiento, con nuestro dataset.

Model	Precision (Cancelled)	Precision (Delayed)	Precision (On Time)	Recall (Cancelled)	Recall (Delayed)	Recall (On Time)	F1-score (Cancelled)	F1-score (Delayed)	F1-score (On Time)	Accuracy
Logistic Regression	0.34	0.35	0.34	0.35	0.34	0.33	0.34	0.35	0.34	0.34
Random Forest	0.33	0.34	0.34	0.35	0.33	0.33	0.34	0.33	0.33	0.33
K-Nearest Neighbors	0.33	0.34	0.34	0.47	0.32	0.21	0.39	0.33	0.26	0.33

Los tres algoritmos deben ser mejorados , con mejores técnicas u obtener una mayor cantidad de datos.

Conclusión :

Se utilizaron técnicas para buscar hiperparametros que permitan un mejor resultado.

RandomizedSearchCV – Este algoritmo utiliza muchos recursos y su tiempo de ejecución es muy alto.

Otra opción fue RandomizedSearchCV , aunque en principio parece mas rápido , También Requiere muchos recursos para el tamaño del dateset.

La última prueba realizada en busca de los mejores hiperparametros y mejor algoritmo nos da que son :

```
model = RandomForestClassifier(max_depth=30, min_samples_leaf=4,  
min_samples_split=10, n_estimators=200, random_state=42)
```

Precisión del modelo: 0.3327706159197872

Este algoritmo nos da resultados no muy buenos, por lo tanto llegamos a la conclusión que se debe mejorar el dataset obteniendo mas datos o agregando otras columnas que ayuden con el objetivo buscado.