TRABAJO FINAL DATA SCIENCE

VIAJES AÉREOS EN ARGENTINA

Alumna: Ninfa M. Pedano
Junio 2024

ABSTRACT

El conjunto de datos analizado proporciona un registro detallado de vuelos comerciales que operan en ciertas rutas durante un período específico. Incluye información sobre la regularidad de vuelos, la capacidad de asientos, la cantidad de pasajeros transportados y datos geográficos como aeropuertos de origen y destino, localidades y provincias.

El objetivo principal es explorar patrones estacionales, identificar correlaciones entre variables como la demanda de pasajeros y la capacidad de vuelo, y desarrollar modelos predictivos para prever la demanda futura de vuelos en diferentes destinos.

MOTIVACIÓN

El análisis de datos de vuelos y pasajeros es importante para poder entender patrones de viajes, demanda de transporte aéreo y comportamientos del consumidor en la industria de la aviación. Proporciona información valiosa para aerolíneas, aeropuertos, agencias de viaje y entidades gubernamentales relacionadas con el transporte y el turismo.

AUDIENCIA

Este análisis está dirigido a profesionales del sector de la aviación, investigadores académicos, autoridades de transporte, analistas de datos y cualquier persona interesada en comprender el movimiento de pasajeros y vuelos a nivel regional o global.

INTRODUCCIÓN

Los datos recopilados incluyen información detallada sobre vuelos comerciales, incluyendo origen, destino, aerolíneas, número de pasajeros y asientos disponibles. Estos datos son esenciales para estudiar la conectividad aérea, la demanda de transporte y los patrones de viaje en diferentes regiones y continentes.

DESCRIPCIÒN DEL CONJUNTO DE DATOS

Los datos fueron obtenidos de la pàgina del Ministerio de Turismo y Deportes de Argentina.
https://datos.yvera.gob.ar/dataset/conectividad-aerea/archivo/aab49234-28c9-48ab-a978-a83485139290

Proporciona un registro detallado de vuelos comerciales que operan en ciertas rutas durante un período específico. Incluye información sobre la regularidad de vuelos, la capacidad de asientos, la cantidad de pasajeros transportados y datos geográficos como aeropuertos de origen y destino, localidades y provincias. Si bien no brinda informaciòn sobre los pasajeros (sòlo cantidad), son datos màs completos y ademas reales.

Aclaraciòn importante: En principio se trabajò con un dataset extraìdo de Kaggle referido a una aerolìnea especìfica, pero con el correr del cursado me di cuenta que era muy pobre la informaciòn del mismo y no podrìa desarrollar de la forma que querìa mi proyecto.

OBJETIVO

El objetivo principal del presente trabajo es explorar patrones estacionales, identificar correlaciones entre variables como la demanda de pasajeros y la capacidad de vuelo, y desarrollar modelos predictivos para prever la demanda futura de vuelos en diferentes destinos.

HIPÓTESIS DEL TRABAJO

- **Hipótesis 1**: Existe una correlación positiva entre el número de vuelos y la cantidad de pasajeros transportados en rutas de alta demanda.
- **Hipótesis 2**: Las aerolíneas experimentan una mayor ocupación de asientos en vuelos durante ciertos períodos del año debido a factores estacionales y eventos específicos.
- **Hipótesis 3**: La regularidad de la ruta aérea influye significativamente en la elección de los pasajeros.

CAMPOS DEL DATASETS

- índice tiempo: Índice de tiempo diario. Fecha de vuelo.
- clasificacion_vuelo: Tipo de conexión.
- clase_vuelo: Regularidad de la ruta aérea.
- aerolínea: Aerolínea comercial.
- origen_aeropuerto: Aeropuerto de origen.
- origen_localidad: Localidad de origen del vuelo.
- origen_provincia: Provincia de origen del vuelo.
- origen_continente: Continente de origen.
- destino_aeropuerto: Aeropuerto de destino.
- destino_localidad: Localidad de destino.
- destino_provincia: Provincia de destino.
- destino_continente: Continente de destino
- pasajeros: Cantidad de pasajeros
- asientos Número decimal: Cantidad de asientos
- vuelos: Cantidad de vuelos

CÁLCULO DE MÉTRICAS

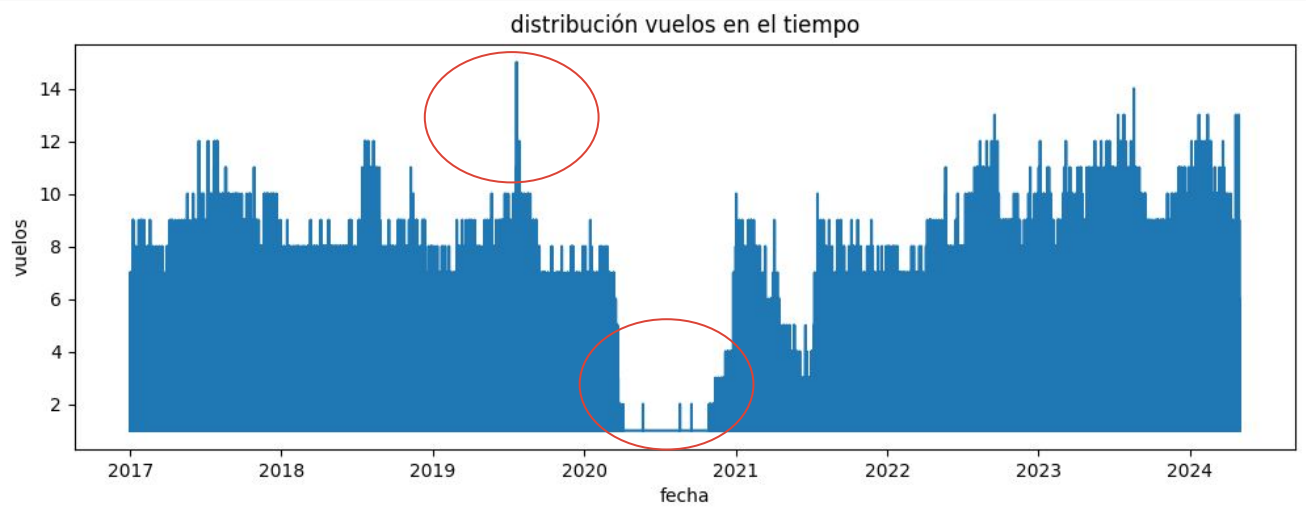
COP=CAPACIDAD OPERATIVA DEL AVIÓN. CANTIDAD DE CANTIDAD DE PASAJEROS/ASIENTOS

CINS=CAPACIDAD INSTALADA. Es la cantidad de asientos posibles de ocupar.(Finalmente no se usó)

PROMAS = PROMEDIO DE ASIENTOS POR AVIÓN. ASIENTOS/VUELOS

```
df['COP'] = df['pasajeros'] / df['asientos']
df['PROMAS'] = df['asientos'] / df['vuelos']
```

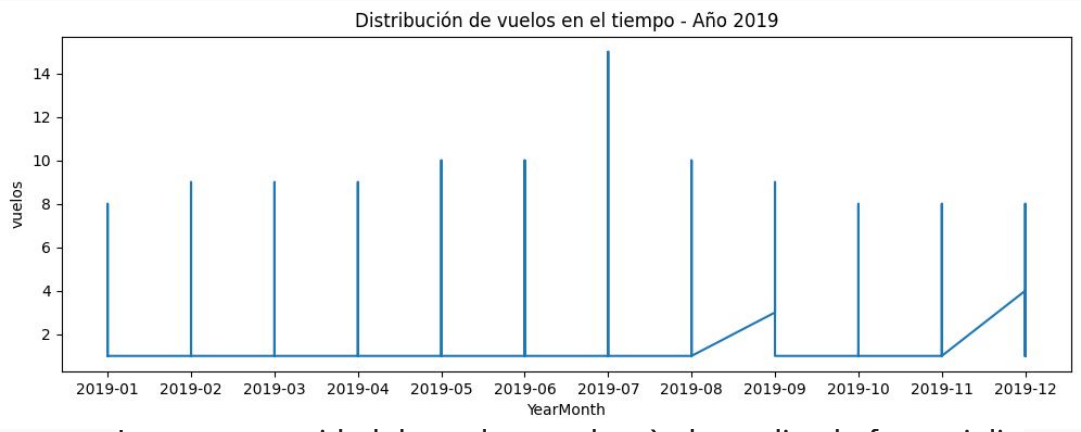
ESTADÍSTICAS GENERALES

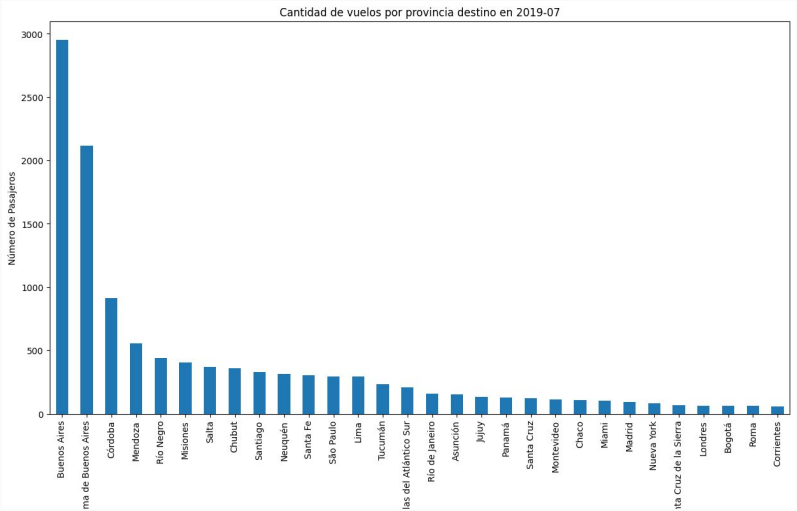


Se observa claramente la caída en los vuelos en el año 2020 producto de la pandemia. Se decide eliminar los datos de ese año para que no generen dispersiones. El año 2024 solo tiene datos de 4 meses.

Llama la atención también en el año 2019 un elevado número de vuelos que màs adelante analizaremos.

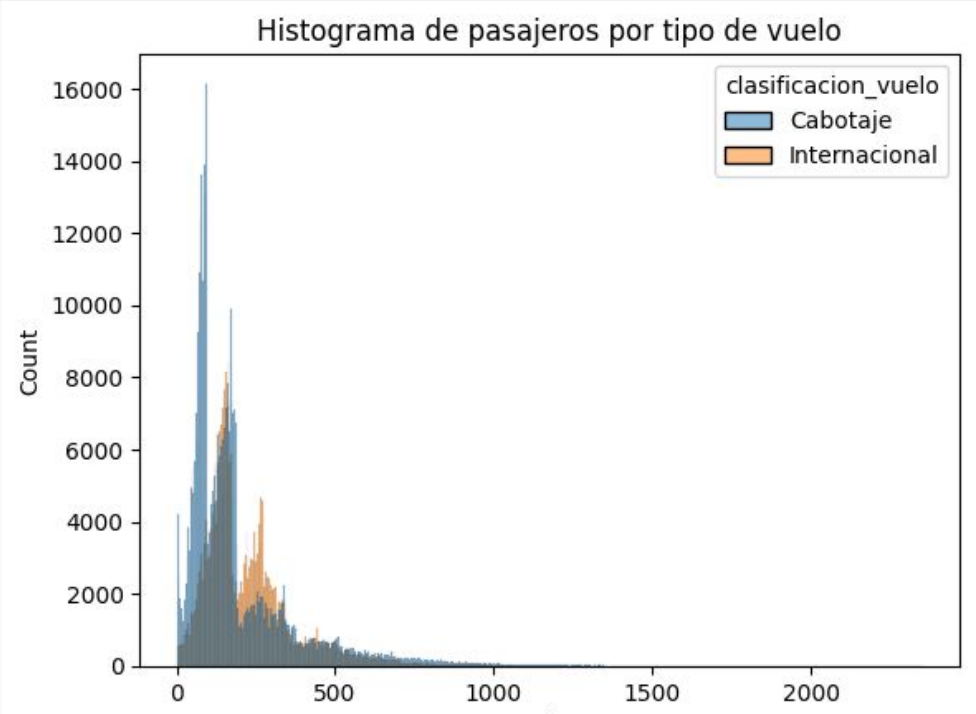
VISUALIZACIONES





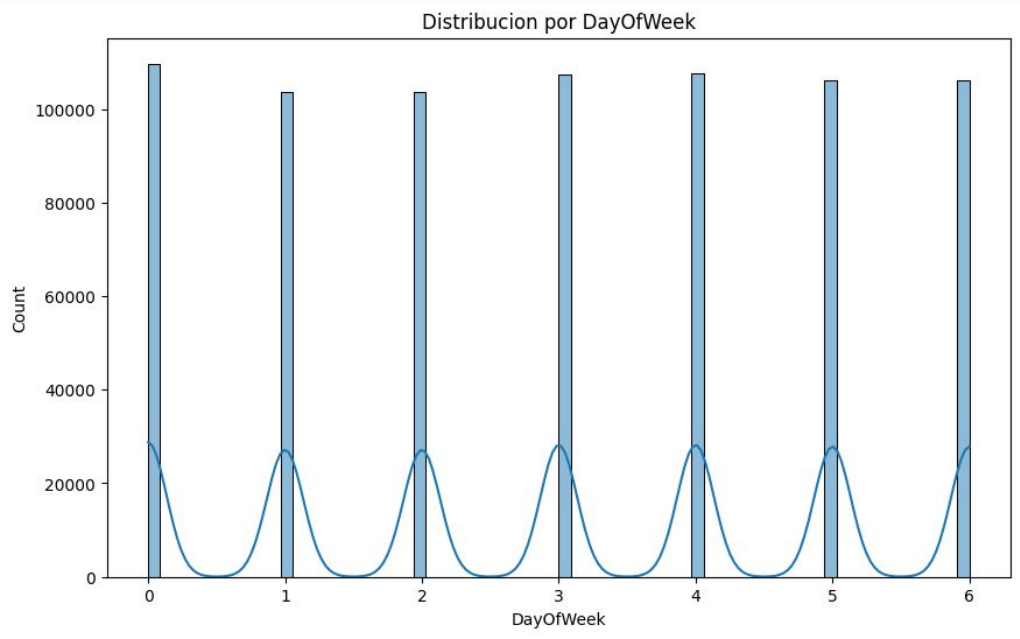
La mayor cantidad de vuelos en el perìodo analizado fue en julio 2019 coincidente con el comienzo de las vacaciones de invierno en Argentina y el destino más frecuente fue Buenos Aires.

HISTOGRAMAS



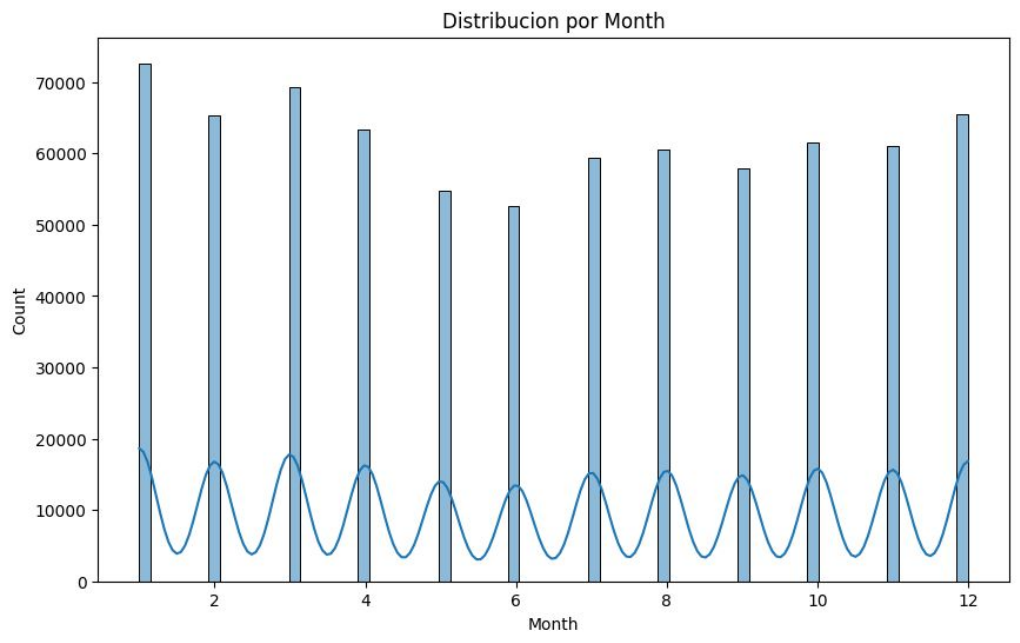
Los vuelos de cabotaje representan mayor índice de frecuencia respecto al tipo Internacional. En el histograma se puede ver la distribución bimodal de dos realidades distintas, mayoritaria de vuelos de cabotaje sobre internacionales y un comportamiento atípico posiblemente relacionado a las aerolíneas y el tipo de avión y capacidad según sea de cabotaje o internacional.

HISTOGRAMAS



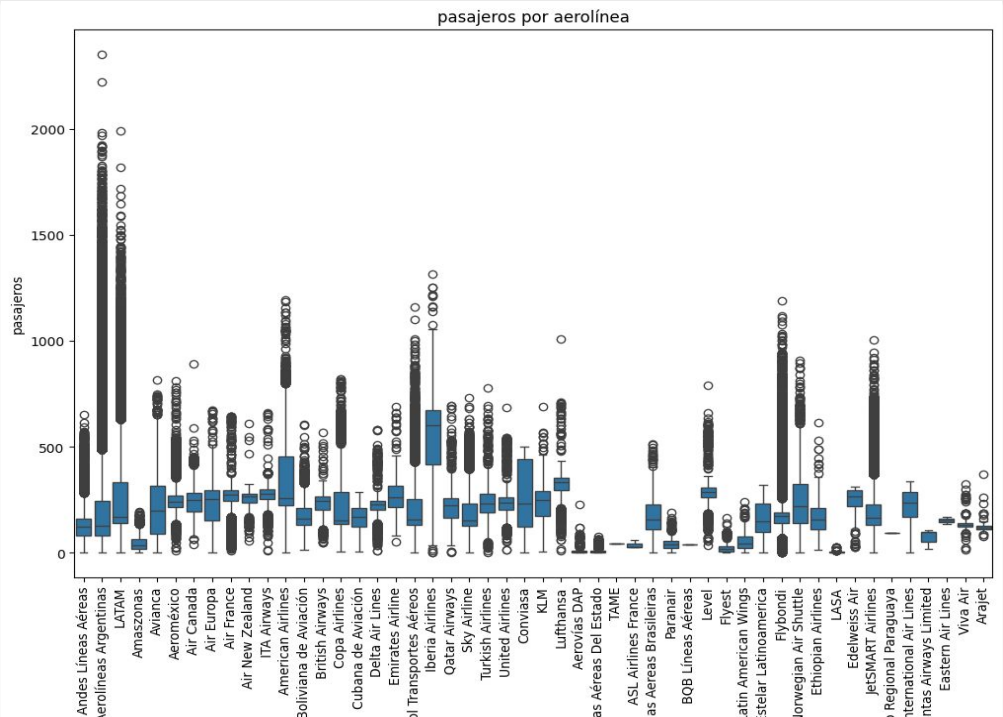
En este histograma podemos ver que los vuelos se distribuyen de forma pareja en los días de la semana, aumentando los días domingos.

HISTOGRAMAS



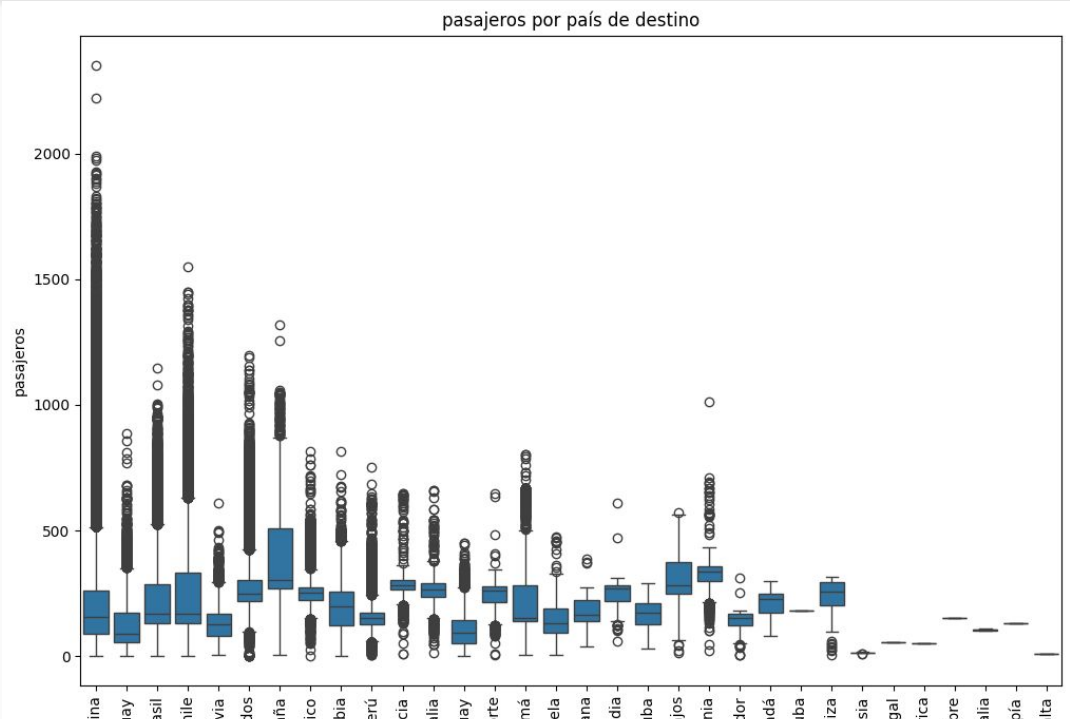
Este gráfico de distribución por mes muestra un patrón repetitivo mes a mes, lo que sugiere una tendencia estacional en los datos. Cada mes tiene picos y valles similares en la distribución de datos.
Los picos más altos son consistentes en los mismos meses, indicando una posible estacionalidad fuerte. En los meses de enero y diciembre se dan las vacaciones de verano en Argentina, lo que muestra la fuerte estacionalidad.

BOXPLOTS



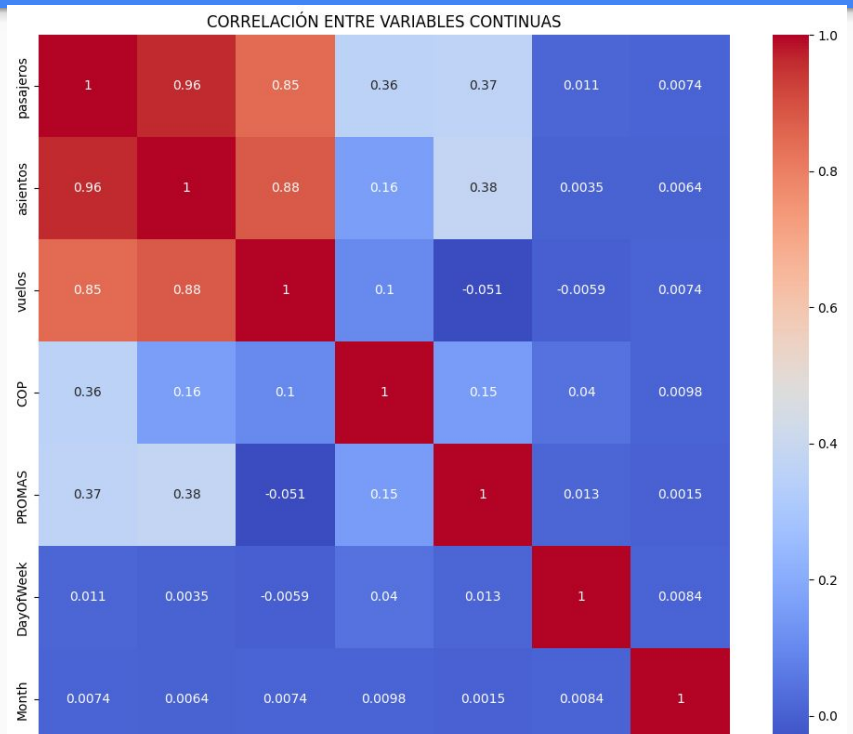
En este boxplot podemos llegar a apreciar la dispersión respecto a cantidades entre la aerolínea “Aerolíneas Argentinas” y las demás.

BOXPLOTS



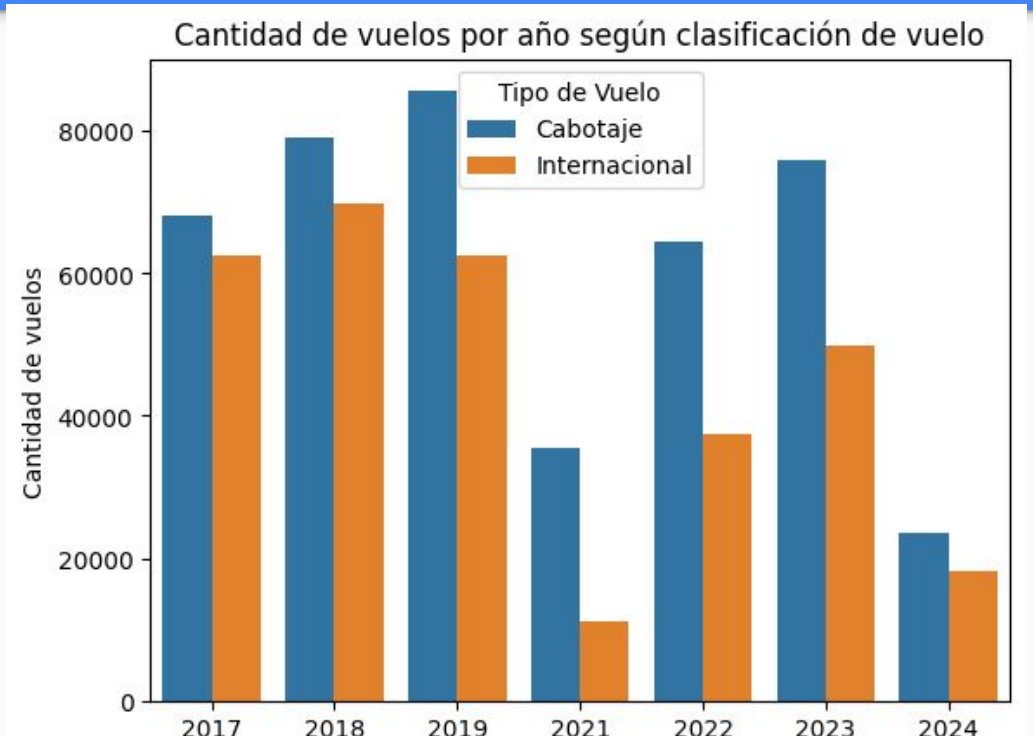
En este boxplot podemos llegar a apreciar la dispersión respecto a cantidades entre pasajeros por país de destino, debido a lo que vimos más arriba, que la mayor parte de vuelos corresponden a vuelos de cabotaje en Argentina.

HEATMAP



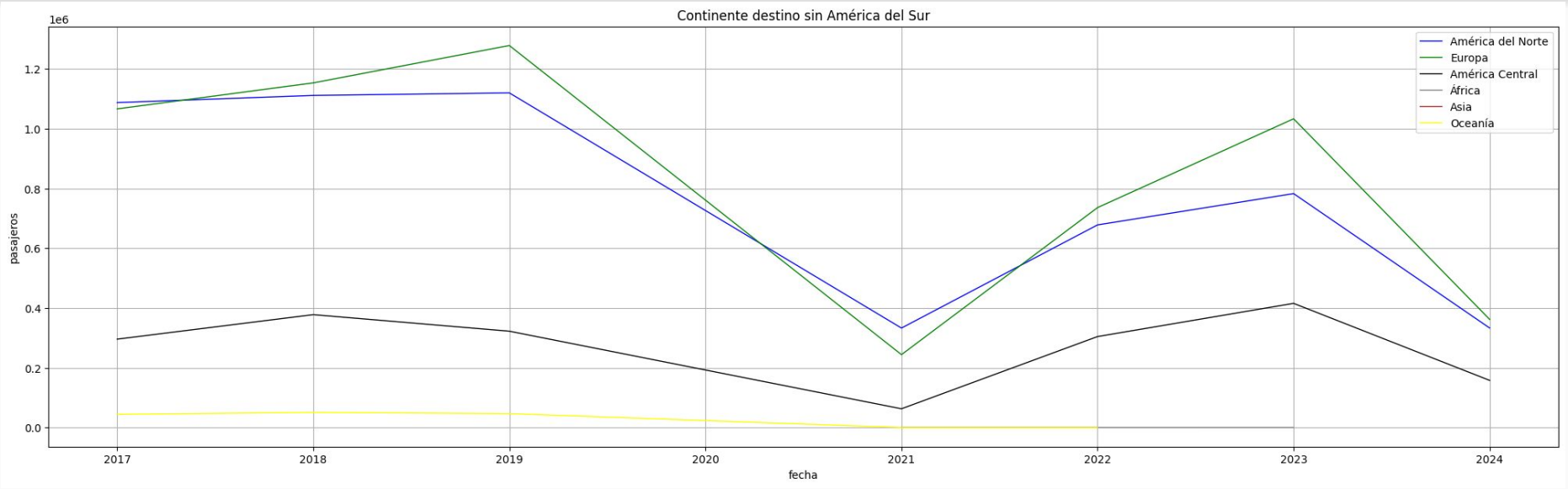
En este Heatmap podemos ver mejor la correlación existente entre asientos y pasajeros, vuelos con asientos y pasajeros.

GRÁFICO DE BARRAS



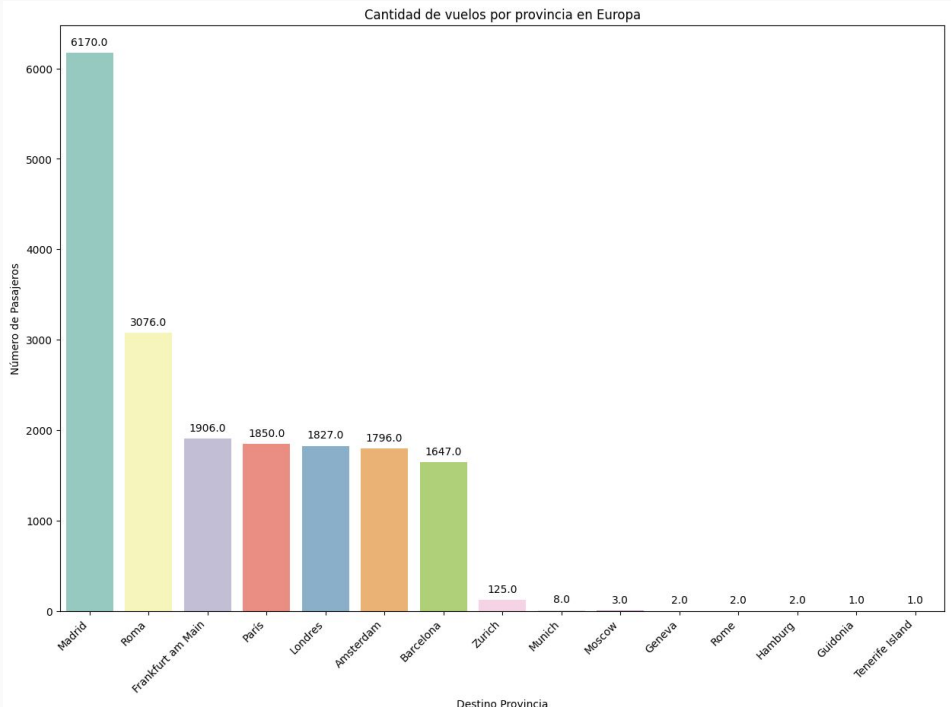
Recordamos que se eliminó el año 2020 del análisis y vemos que en 2021 empezaron los vuelos internacionales fueron muy bajos comparados con los de cabotaje.

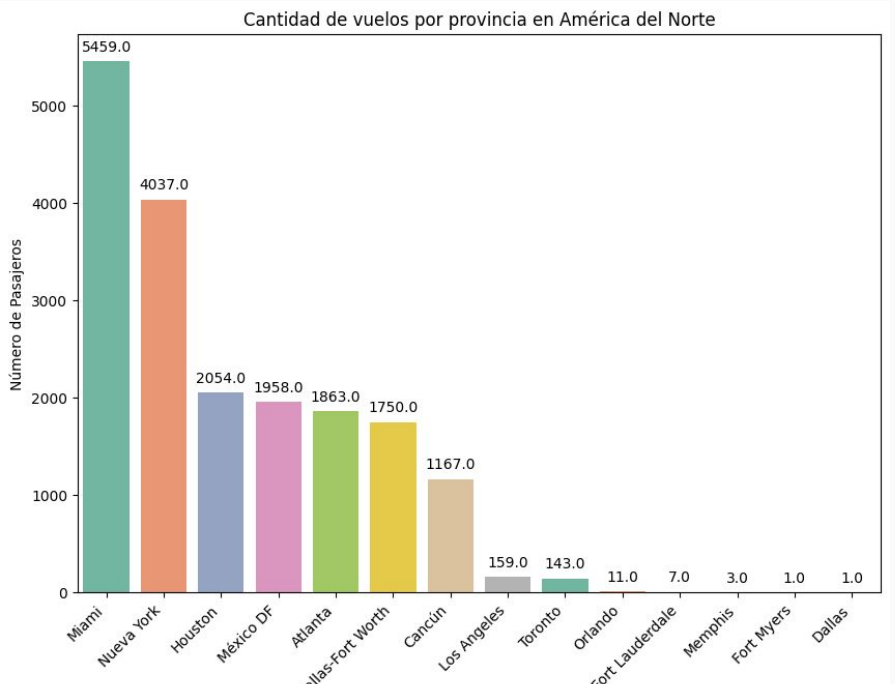
GRÁFICO DE LÍNEAS



Después de América del sur, Europa es el destino más elegido, seguido por América del Norte. En este caso no sirviò eliminar previamente el año 2020 de pandemia.

MÁS GRÁFICOS INFORMATIVOS





SELECCIÓN DE MODELOS

1-Modelo XGBoost

```
Error cuadrático medio (MSE): 0.010243888311194764

Error absoluto medio (MAE): 0.0224120873010958

Coeficiente de determinación (R^2): 0.9939372753914077
```

SELECCIÓN DE MODELOS

2-REGRESIÓN LINEAL

Model	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R2	Test R2
0 Linear	0.209943	0.26838	0.134235	0.154646	0.96362	0.957371

SELECCIÓN DE MODELOS

3-LightGBM

Model	Train RMSE	Test RMSE	Train MAE	Test MAE	Train R2	Test R2
LGBM	0.008162	0.021665	0.000979	0.002079	0.999945	0.999722

COMPARACIÓN DE MODELOS

El modelo de regresión lineal tiene un rendimiento decente, con un R2 alto tanto en el conjunto de entrenamiento como en el de prueba, indicando que el modelo explica bien la variabilidad de los datos.

Sin embargo, tiene errores más altos (RMSE y MAE) en comparación con los modelos XGBoost y LGBM.

El modelo XGBoost tiene un buen rendimiento tanto en el conjunto de entrenamiento como en el de prueba. El RMSE y MAE son bajos, lo que indica que las predicciones están bastante cerca de los valores reales. Además, el R2 es alto, lo que significa que una gran proporción de la variabilidad en los datos de respuesta se explica por el modelo.

COMPARACIÓN DE MODELOS

El modelo LGBM tiene el mejor rendimiento entre todos los modelos evaluados.

Los errores (RMSE y MAE) son los más bajos, tanto en el conjunto de entrenamiento como en el de prueba, indicando una alta precisión.

El R2 es muy alto, casi perfecto, indicando que el modelo LGBM explica casi toda la variabilidad de los datos.

STACKING REGRESSOR

Combinamos las predicciones de los 3 modelos base para producir una predicción final intentando que sea más precisa y robusta que las predicciones individuales de cada modelo.

```
Stacking Regressor Model Metrics:

                Model  Train RMSE  Test RMSE  Train MAE  Test MAE  Train R2  \
0  Stacking Regressor    0.009471   0.022349   0.001237   0.00242  0.999926

    Test R2
0  0.999704
```

CONCLUSIÓN

El Stacking Regressor ha mostrado un rendimiento excelente en términos de RMSE, MAE y R^2 tanto en el conjunto de entrenamiento como en el de prueba, comparado con los modelos individuales. Este resultado subraya la eficacia del Stacking al combinar diferentes modelos base para mejorar la precisión y la robustez del modelo final.

GRACIAS POR SU ATENCIÓN