CEU
Universidad
San Pablo

# Análisis de Tráfico Aéreo en el Aeropuerto de San Francisco

## Tecnicas de Big Data y Visualizacion Avanzada

**Caso Final Integrador**
**Maria Angel Lobon Gonzalo**

# Contenido

# Objetivos del Proyecto

- El objetivo del análisis es introducir los conceptos básicos de la ciencia de datos aplicada al tráfico aéreo.
- Se busca explorar y comprender un conjunto de datos con información valiosa sobre aerolíneas, destinos y patrones de tráfico.
- Este análisis permitirá obtener un entendimiento profundo del transporte aéreo y apoyar decisiones informadas en la industria.
- El conjunto de datos incluye múltiples columnas con los siguientes propósitos:
  - **Year** y **Month**: Indican el año y mes de la actividad.
  - **Operating Airline**: Aerolínea operativa.
  - **GEO Summary**: Clasifica el vuelo como nacional o internacional.
  - **GEO Region**: Identifica la región geográfica (p. ej., Estados Unidos, Europa, Asia).
  - **Activity Type Code**: Clasifica la operación como embarque o desembarque.
  - **Passenger Count**: Refleja el número de pasajeros en cada registro.
- Este análisis inicial permitirá identificar tendencias y patrones en el flujo de pasajeros, así como la popularidad de aerolíneas y destinos.
- Los resultados generarán conocimientos útiles para optimizar decisiones en el sector aéreo.

# Técnicas Y Metodología Utilizada

- **Preparación de los datos**: Se comienza con la carga y limpieza del conjunto de datos, eliminando registros duplicados y nulos, y asegurándose de que las variables relevantes estén correctamente formateadas.
- **Filtrado de datos**: Se filtran subconjuntos de datos según criterios específicos, como el año, la aerolínea, el tipo de vuelo (nacional o internacional) y las regiones geográficas de destino.
- **Análisis descriptivo**: Se calculan estadísticas descriptivas para entender las características del tráfico aéreo, como el número promedio de pasajeros por aerolínea y destino, así como la distribución de pasajeros según el mes o la región geográfica.
- **Matriz de correlación**: Se crea una matriz de correlación para identificar relaciones entre variables, como el año y el número de pasajeros, o entre las distintas regiones geográficas y el tráfico de pasajeros.
- **Clustering (Agrupación)**: Se aplica el algoritmo K-Means para agrupar destinos internacionales con características similares, utilizando variables como el número de pasajeros y la región geográfica.
- **Visualización**: Se generan visualizaciones para explorar los resultados de las agrupaciones, mostrando los clusters de destinos internacionales y su relación con el número de pasajeros.
- **Interpretación de resultados**: Se analizan los patrones emergentes de las agrupaciones y correlaciones, lo que permite obtener insights sobre las tendencias de tráfico aéreo y ayudar a la toma de decisiones en la industria.

# Limpieza y Filtrado de Datos

Limpieza

```
Registros antes de eliminar duplicados y nulos: 15007
Activity Period                0
Operating Airline              0
Operating Airline IATA Code    54
Published Airline              0
Published Airline IATA Code    54
GEO Summary                    0
GEO Region                     0
Activity Type Code             0
Price Category Code            0
Terminal                       0
Boarding Area                  0
Passenger Count                0
Adjusted Activity Type Code    0
Adjusted Passenger Count       0
Year                           0
Month                          0
dtype: int64
¿Hay valores nulos en el DataFrame?: True

Verificando valores nulos después de eliminarlos:
Activity Period                0
Operating Airline              0
Operating Airline IATA Code    0
Published Airline              0
Published Airline IATA Code    0
GEO Summary                    0
GEO Region                     0
Activity Type Code             0
Price Category Code            0
Terminal                       0
Boarding Area                  0
Passenger Count                0
Adjusted Activity Type Code    0
Adjusted Passenger Count       0
Year                           0
Month                          0
dtype: int64

¿Hay valores nulos en el DataFrame?: False
```

Filtrado

```
Filtrar para ATA Airlines en 2005
    Operating Airline  Year   Month  Passenger Count
0         ATA Airlines  2005    July            27271
1         ATA Airlines  2005    July            29131
2         ATA Airlines  2005    July             5415
115       ATA Airlines  2005  August            27472
116       ATA Airlines  2005  August            26535

Filtrar vuelos internacionales hacia Europa
   Operating Airline GEO Region  Passenger Count
7          Air France     Europe            12050
8          Air France     Europe            11638
31   BelAir Airlines      Europe              325
32   BelAir Airlines      Europe              545
33   British Airways      Europe            20632

Extraer datos de 2005 a 2020, solo para vuelos nacionales con más de 10,000 pasajeros
    Year  Operating Airline  Passenger Count
0   2005       ATA Airlines            27271
1   2005       ATA Airlines            29131
13  2005    Alaska Airlines            36641
14  2005    Alaska Airlines            39379
23  2005  American Airlines           166577
```

```
Vuelos Internacionales de ATA Airlines en 2005
    Operating Airline GEO Region  Passenger Count
473       ATA Airlines     Canada              284
474       ATA Airlines     Canada              284

Vuelos Internacionales de Air France entre 2005 y 2009
    Operating Airline    GEO Summary  Passenger Count  Year
7           Air France  International            12050  2005
8           Air France  International            11638  2005
122         Air France  International            11230  2005
123         Air France  International            11731  2005
240         Air France  International            10731  2005
```

# Matrices de Correlación

```
Matriz de correlacion entre Año y Pasajeros
                      Year  Passenger Count
Year              1.000000         0.060917
Passenger Count   0.060917         1.000000
```

```
Matriz de correlacion entre Año y Pasajeros
                      Year  Passenger Count
Year              1.000000         0.060917
Passenger Count   0.060917         1.000000

Matriz de correlacion entre cada Mes y Pasajeros
m_February        -0.026258
m_January         -0.016489
m_November        -0.008311
m_September       -0.003156
m_December        -0.002754
m_April           -0.001633
m_March            0.000679
m_October          0.002545
m_May              0.005473
m_June             0.013529
m_July             0.018018
m_August           0.018336
Passenger Count    1.000000
Name: Passenger Count, dtype: float64
```

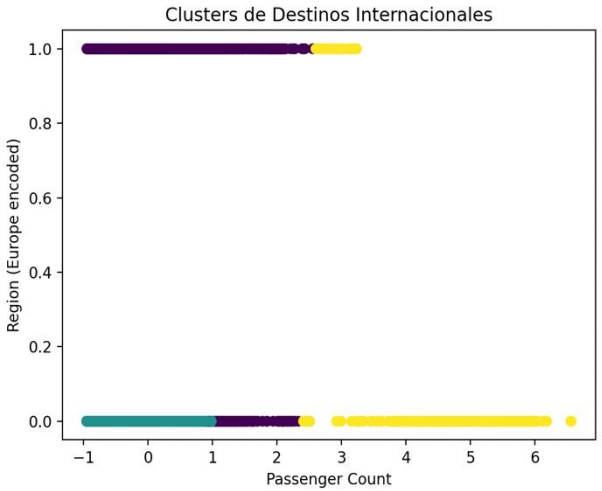
```
Matriz de correlacion entre cada Destino y Pasajeros
GEO_Asia                  -0.144164
GEO_Europe                -0.113674
GEO_Canada                -0.108458
GEO_Mexico                -0.107775
GEO_Australia / Oceania   -0.089400
GEO_Central America       -0.056790
GEO_Middle East           -0.042686
GEO_South America         -0.035392
GEO_US                     0.398198
Passenger Count            1.000000
Name: Passenger Count, dtype: float64
```

El coeficiente de correlación entre el Año y el número de Pasajeros es **0.060917**, lo cual indica una correlación muy baja y positiva. Esto sugiere que el aumento de pasajeros no está fuertemente relacionado con el año de forma lineal.

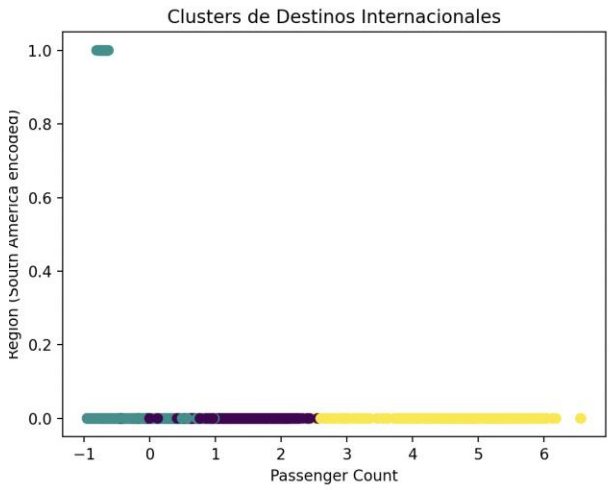
La mayoría de los valores están muy cercanos a 0, lo que indica que no hay una relación lineal significativa entre los meses específicos y el número de pasajeros. Esto sugiere que el volumen de pasajeros no está fuertemente afectado por la temporada o el mes, al menos de manera directa. Algunos meses, como **February (-0.026258)**, **January (-0.016489)** y **November (-0.008311)**, tienen correlaciones negativas muy bajas con el número de pasajeros. Esto podría indicar que en estos meses hay una leve tendencia a tener menos pasajeros en comparación con otros meses.

**GEO_US (0.398198)** tiene una correlación positiva moderada con el número de pasajeros. Esto sugiere que los vuelos hacia destinos dentro de los Estados Unidos tienden a tener un mayor número de pasajeros en comparación con otras regiones. Esto es consistente con la posible alta demanda de vuelos domésticos en EE.UU., debido a su gran tamaño, población y economía. **GEO_Asia (-0.144164)** tiene la correlación negativa más alta. Esto podría indicar que los vuelos hacia Asia representan una proporción menor del volumen total de pasajeros. **GEO_Europe (-0.113674)** y **GEO_Canada (-0.108458)** también tienen correlaciones negativas, aunque algo menores, lo que sugiere que los vuelos hacia estas regiones tienen un impacto más limitado en el volumen total de pasajeros.

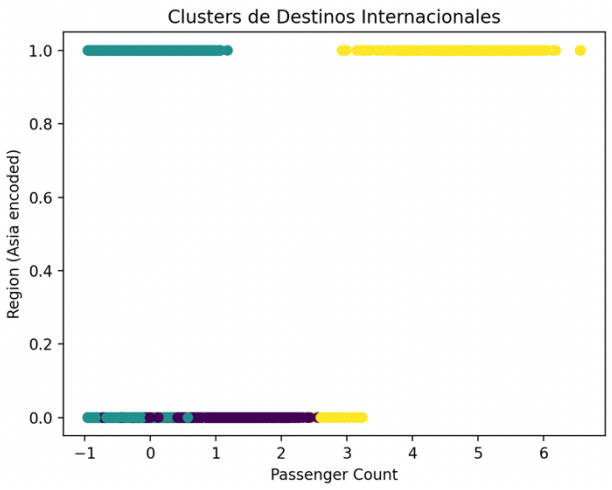
# Clusterización de Vuelos Internacionales




Para las regiones etiquetadas como 1 (Europa), hay un alto número de pasajeros distribuido entre los diferentes grupos, mostrando una varianza significativa. Las regiones etiquetadas como 0 (no europeas) aparecen menos densamente agrupadas en los recuentos bajos de pasajeros. Esta agrupación sugiere que los destinos europeos experimentan una amplia gama de niveles de tráfico de pasajeros, mientras que los destinos no europeos (en este grupo específico) pueden mostrar menos pasajeros o un número más bajo de estos.



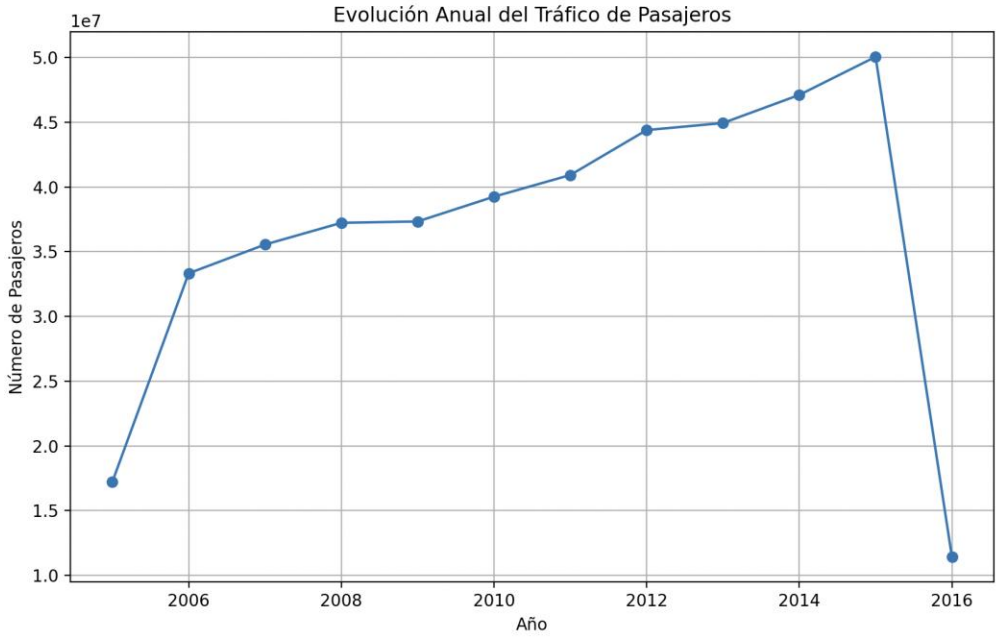

Las regiones de América del Sur (1) tienen menos grupos en general, y el tráfico parece concentrarse en un rango particular de recuentos bajos de pasajeros. Las regiones no sudamericanas (0) muestran una distribución más amplia de recuentos de pasajeros entre los grupos. Los destinos sudamericanos están generalmente asociados con un tráfico de pasajeros más bajo en comparación con los destinos no sudamericanos, que tienen una distribución más amplia y recuentos más altos.




Para las regiones marcadas como 1 (Asia), los recuentos de pasajeros parecen estar concentrados en niveles más altos. Las regiones marcadas como 0 (no asiáticas) tienen grupos tanto en recuentos bajos como altos, lo que sugiere variabilidad. Los destinos asiáticos podrían dominar grupos específicos de alto número de pasajeros. En contraste, las regiones no asiáticas tienen niveles de tráfico mixtos, posiblemente reflejando la diversidad en los patrones de viaje hacia/desde Asia.

# 1. Visualización de Trafico Anual

Este gráfico muestra la evolución anual del tráfico de pasajeros a lo largo del tiempo. Muestra el número de pasajeros en millones en el eje y y el año en el eje x.
El gráfico comienza en 2006 con alrededor de 2.3 millones de pasajeros y muestra un aumento constante en el número de pasajeros a lo largo de los años, alcanzando un pico de casi 5 millones de pasajeros en 2014. Sin embargo, después de 2014, el gráfico muestra una caída pronunciada, descendiendo a poco más de 1 millón de pasajeros en 2016.
Estos datos probablemente representan una métrica importante de la industria del transporte o la aviación, que experimentó un crecimiento significativo seguido de una reciente desaceleración. El gráfico proporciona una representación visual clara de estas tendencias en el tráfico de pasajeros durante el período de 10 años, de 2006 a 2016.



Según el gráfico, hay algunas posibles razones para los aumentos y descensos observados en el tráfico de pasajeros:

El aumento constante de pasajeros de 2006 a 2014 podría atribuirse a factores como el crecimiento económico, el aumento del turismo y los viajes de negocios, la expansión de las rutas aéreas y la capacidad de las aerolíneas, y la creciente demanda de los consumidores de transporte aéreo.

El descenso pronunciado después de 2014 podría haber sido provocado por recesiones económicas, cambios en las preferencias de viaje, interrupciones en la industria de la aviación o cambios en los patrones de transporte. Algunas posibles causas podrían incluir:

Recesiones o desaceleraciones económicas que redujeron los presupuestos para viajes discrecionales

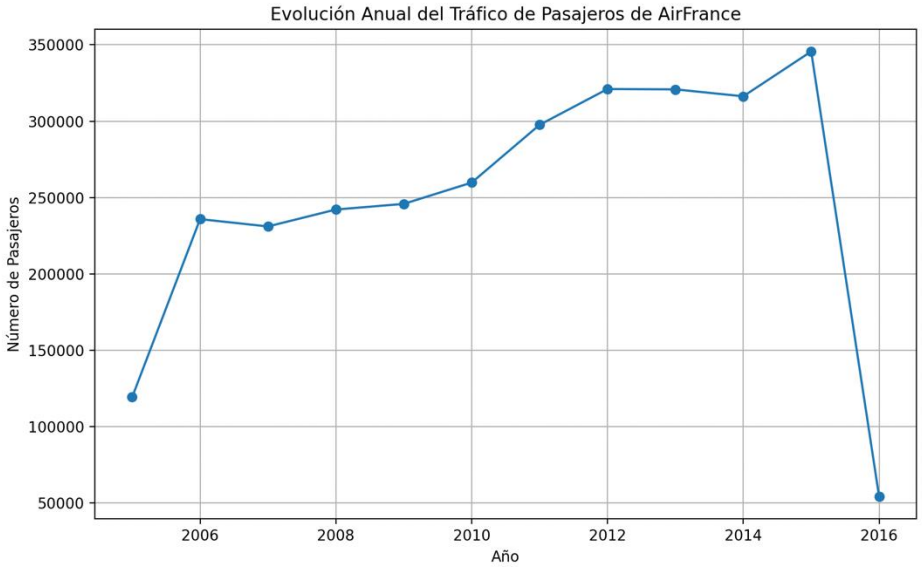
Desplazamientos hacia modos de transporte alternativos como el tren de alta velocidad o la videoconferencia

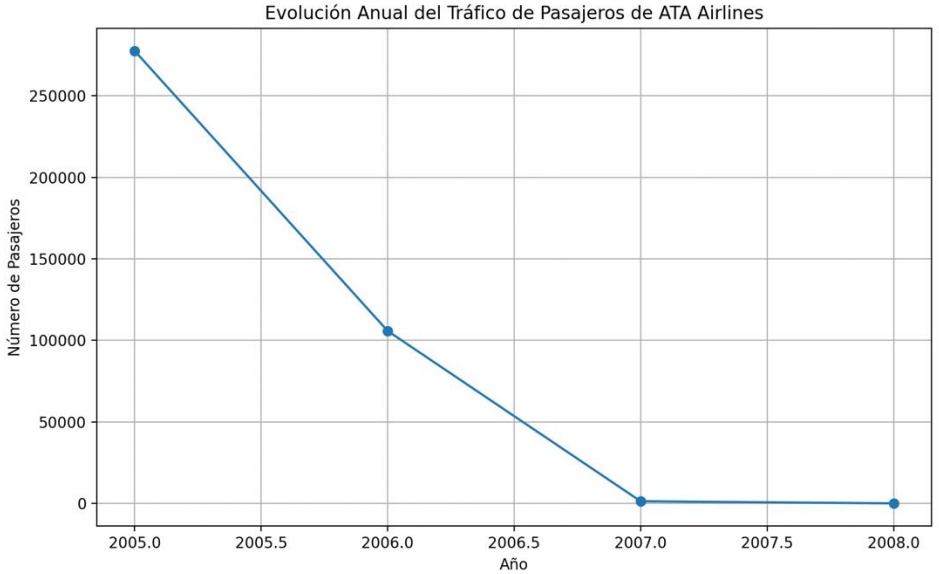
Tensiones geopolíticas, preocupaciones de seguridad o crisis de salud pública que disminuyeron la demanda de viajes aéreos

Desafíos operacionales como quiebras de aerolíneas, consolidaciones o recortes de servicios

Sin más información contextual, es difícil hacer predicciones definitivas. Sin embargo, el gráfico indica una tendencia volátil en los últimos años que podría continuar dependiendo de los desarrollos en el ámbito económico, social y de transporte. Sería necesario un análisis más detallado de los factores industriales y macroeconómicos para hacer pronósticos más informados.

# 1.1 Visualización de Trafico Anual por Aerolineas




Air France presenta una trayectoria de crecimiento sostenido desde 2005 hasta 2014, con un patrón particularmente interesante. La aerolínea experimentó un salto significativo en 2006, pasando de aproximadamente 120,000 a 230,000 pasajeros, seguido de un período de crecimiento más moderado pero constante hasta 2012. Entre 2012 y 2014, el tráfico se mantuvo relativamente estable alrededor de los 320,000 pasajeros, con un pico notable en 2014 alcanzando casi 340,000 pasajeros. Sin embargo, en 2016 se observa una caída dramática hasta aproximadamente 50,000 pasajeros.

Por otro lado, ATA Airlines muestra una historia muy diferente y más breve. La aerolínea experimentó un declive constante y pronunciado desde 2005, donde comenzó con aproximadamente 270,000 pasajeros, hasta su eventual desaparición en 2007. Esta caída fue rápida y sostenida, reflejando probablemente dificultades operativas y financieras significativas que llevaron a su cese de operaciones.

# Visualización de Trafico Anual

Cuando comparamos estas tendencias con el gráfico de tráfico anual general, observamos patrones interesantes:

1. Mientras el tráfico general mostraba un crecimiento constante desde 2005 hasta 2014 (de aproximadamente 17 millones a 50 millones de pasajeros), Air France seguía una tendencia similar aunque a menor escala, sugiriendo que la aerolínea estaba en sintonía con las tendencias generales del mercado.
2. El colapso de ATA Airlines en 2007 no tuvo un impacto visible en la tendencia general del tráfico, lo que sugiere que otras aerolíneas absorbieron rápidamente su cuota de mercado.
3. La caída abrupta en 2016 se observa tanto en el tráfico general como en Air France, indicando que probablemente fue resultado de factores macroeconómicos o eventos que afectaron a toda la industria aérea.

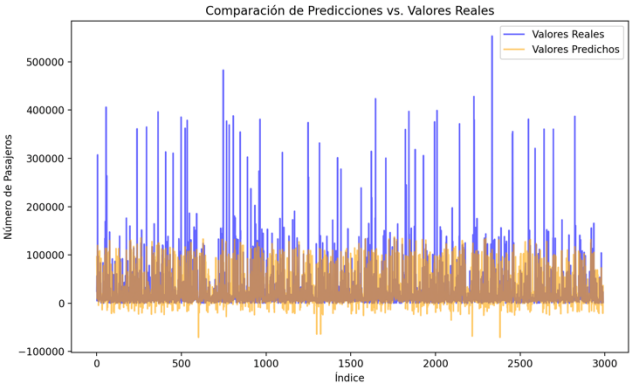
Esta comparación revela cómo las aerolíneas individuales pueden seguir trayectorias muy diferentes dentro de un mismo mercado: mientras algunas prosperan y crecen con el mercado general (como Air France hasta 2014), otras pueden enfrentar dificultades fatales (como ATA Airlines) incluso en períodos de crecimiento general del sector. También demuestra la resistencia del mercado aéreo en general, que puede mantener su crecimiento a pesar de la salida de operadores individuales.

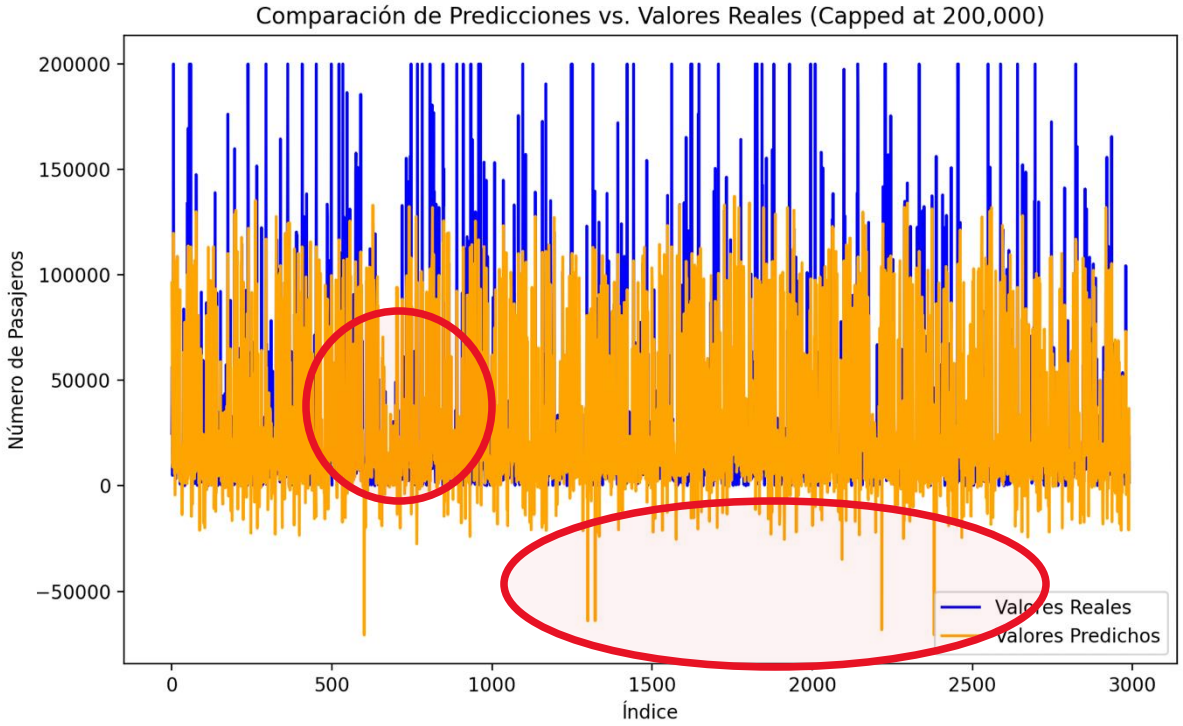
# Comparación de Distintos Modelos

| Predictores del Modelo | Modelo 1 | Modelo 2 | Modelo 3 |
|---|---|---|---|
| Intercepto del Modelo | -1.93 e+16 | 1.85 e+16 | -2129527.6 |
| RMSE | 53778.71 | 48612.47 | 42402.47 |
| MAE | 31165.68 | 48612.47 | 19893.92 |

El primer modelo (Modelo 1) utiliza las variables **'Year'** y **'Month'** como predictoras para estimar el número de pasajeros (**'Passenger Count'**). El segundo modelo (Modelo 2) también utiliza el **'Year'** y **'GEO Region'** como variables predictoras. Finalmente, el tercer modelo (Modelo 3) utiliza las variables **'Year'**, **'Month'**, **'GEO Summary'** (nacional o internacional), y **'Operating Airline'** para predecir el número de pasajeros. El **RMSE es 42,402.47** y el **MAE es 19,893.92**, los cuales son significativamente más bajos que los de los otros dos modelos. Este modelo parece capturar mejor las relaciones entre las variables y los pasajeros, lo que se refleja en su menor error cuadrático medio y error absoluto medio. Además, el intercepto, aunque negativo, es más razonable que en el Modelo 1, lo que sugiere un mejor ajuste.

# Comparación del Modelo de Predicción con Valores Reales




Analizando los gráficos proporcionados, se pueden identificar varios puntos importantes que resaltan áreas de mejora en el modelo de predicción de tráfico de pasajeros. El modelo presenta varias predicciones de valores negativos de pasajeros (representadas por la línea naranja), lo cual es físicamente imposible, ya que el número de pasajeros no puede ser negativo. Estos valores negativos son particularmente notables en ciertos intervalos de índices, como entre los índices 500 y 750, donde se observa una predicción de hasta aproximadamente -70,000 pasajeros; alrededor del índice 1500, donde una predicción baja hasta cerca de -60,000 pasajeros; y entre los índices 2000 y 2500, donde hay varias predicciones negativas significativas.

# Comparación del Modelo de Predicción con Valores Reales

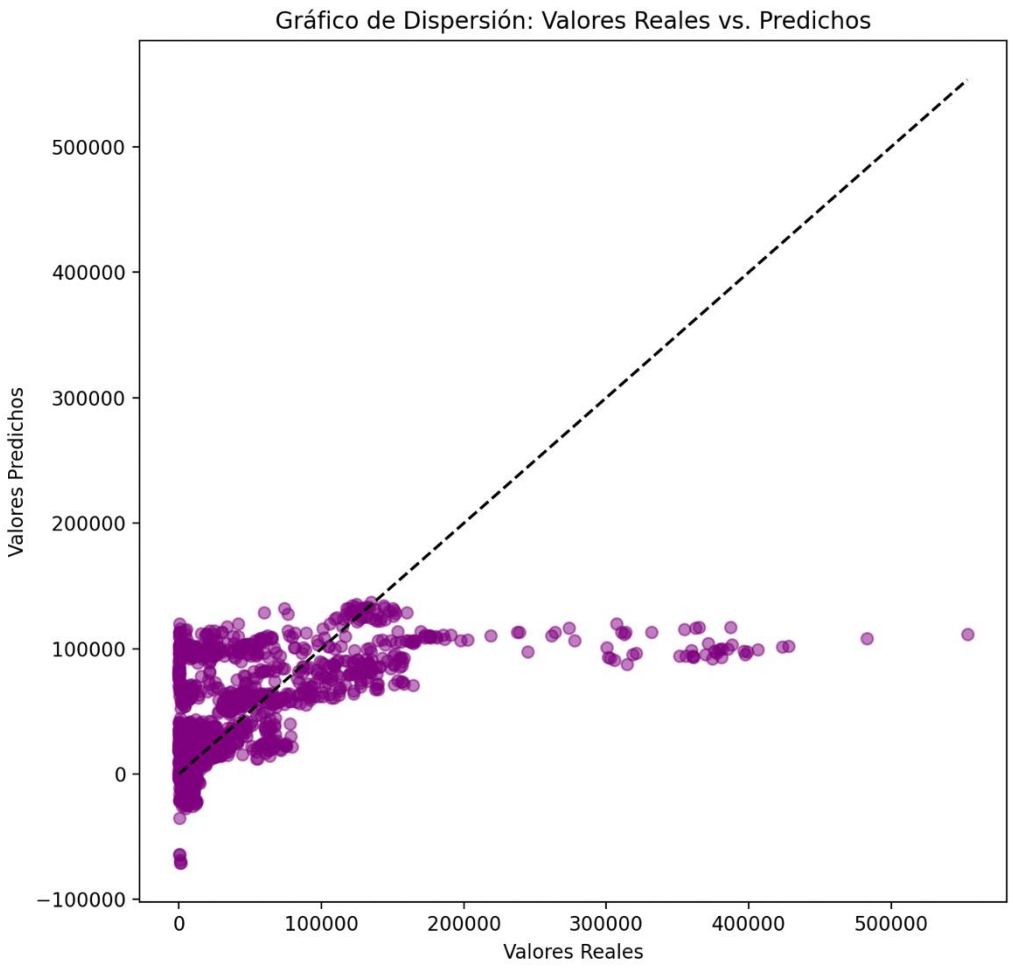
En la región cercana al índice 750, el modelo parece predecir correctamente los valores bajos de pasajeros, y los valores reales (línea azul) y los valores predichos (línea naranja) muestran una mayor concordancia, con fluctuaciones menos extremas y más cercanas a los valores reales. Esto indica que el modelo tiene un mejor rendimiento en este rango específico, a diferencia de otras áreas donde se observan mayores discrepancias.

El segundo gráfico, que muestra un límite superior de 200,000 pasajeros, permite visualizar estos detalles de manera más clara. Es evidente que el modelo enfrenta dificultades para manejar valores extremos, lo que resulta en las predicciones negativas mencionadas.

Para mejorar la precisión y evitar predicciones no realistas, sería necesario considerar ajustes como la normalización de datos para asegurar que los valores de entrada se encuentren en un rango adecuado, la implementación de restricciones en el rango de salida para evitar resultados fuera de los límites físicos, y el uso de funciones de activación que garanticen salidas no negativas. Estas mejoras podrían incrementar la fiabilidad de las predicciones en contextos donde se requiere que los valores sean no negativos.

En resumen, estos puntos sugieren la necesidad de realizar ajustes en el modelo para mejorar su rendimiento y precisión, especialmente en la predicción de valores extremos y la evitación de valores negativos.

# Comparación del Modelo de Predicción con Valores Reales




El gráfico muestra la relación entre los valores reales y predichos, representado como un gráfico de dispersión. La línea discontinua diagonal representa la predicción perfecta (donde los valores reales y predichos serían iguales).

Observaciones clave:

- Hay una gran concentración de puntos en valores más bajos (entre 0 y 100,000)
- La dispersión de los puntos se aleja significativamente de la línea de predicción perfecta, especialmente en valores más altos
- Los valores predichos tienden a subestimar los valores reales cuando estos son altos (más de 200,000)
- Hay algunos valores negativos predichos, lo cual podría indicar un problema en el modelo
- La dispersión se vuelve más pronunciada a medida que aumentan los valores reales

Esta visualización sugiere que el modelo de predicción:

Funciona mejor para valores más bajos

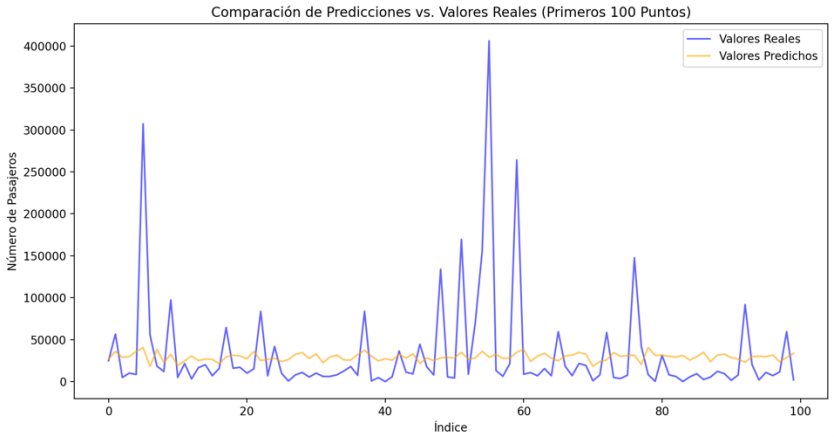
Tiene dificultades para predecir valores extremos

Podría beneficiarse de ajustes para mejorar su precisión en valores más altos

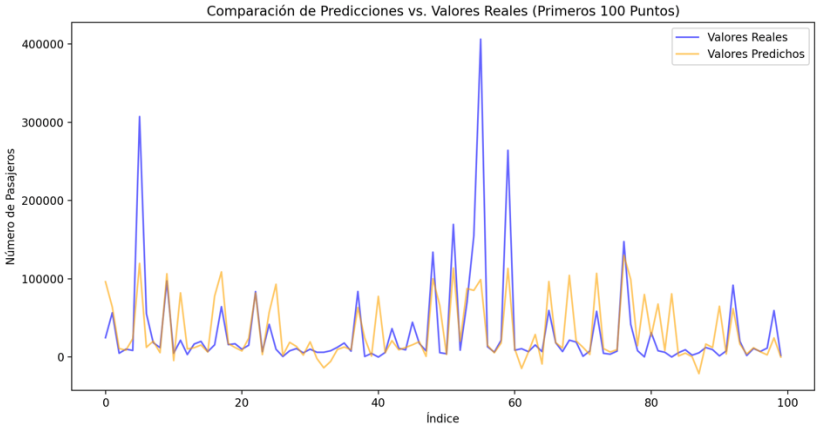
Necesita corrección para evitar predicciones negativas

# Comparación de Distintos Modelos

## Modelo 1




## Modelo 3

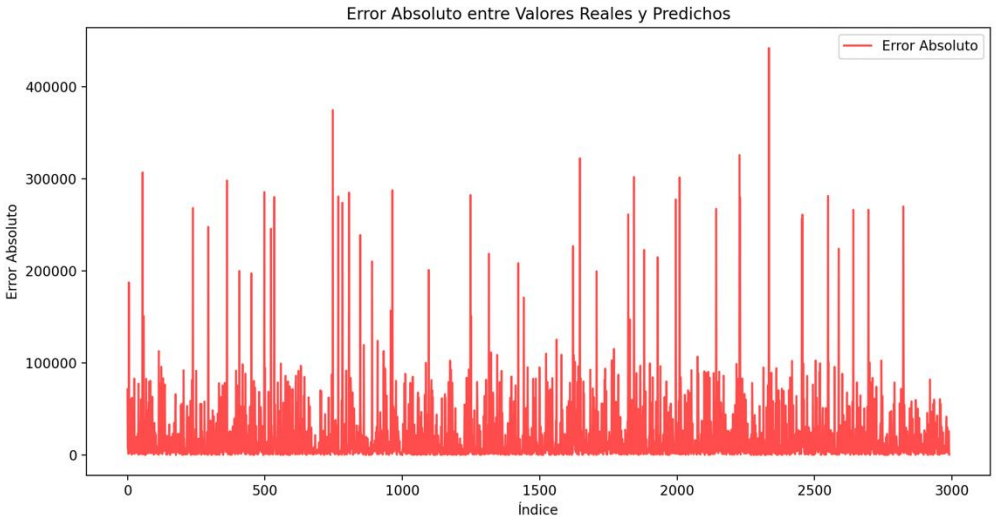



Los modelos predictivos presentados muestran dos enfoques distintos para abordar el complejo desafío de predecir el flujo de pasajeros en un sistema de transporte. El primer modelo exhibe un comportamiento notablemente conservador, manifestando una línea predictiva casi plana que apenas reacciona a las fluctuaciones de los valores reales. Esta aproximación, aunque estable, resulta insuficiente para capturar la rica dinámica del sistema, especialmente en momentos de alta volatilidad o durante eventos que generan picos significativos en el número de pasajeros.

En contraste, el segundo modelo representa una evolución significativa en la capacidad predictiva, demostrando una mayor sensibilidad a los patrones subyacentes en los datos reales. Este modelo logra capturar las tendencias alcistas y bajistas con mayor precisión, y muestra una reactividad más pronunciada ante los cambios en el flujo de pasajeros. Sin embargo, incluso esta versión mejorada enfrenta limitaciones considerables. La más notable es su incapacidad para predecir con exactitud la magnitud de los picos extremos, como se evidencia en el punto donde los valores reales alcanzan los 400,000 pasajeros. Además, el modelo ocasionalmente genera falsos positivos, prediciendo picos que no se materializan en la realidad, y exhibe un ligero retraso temporal en sus predicciones, lo que sugiere una dependencia excesiva de los datos históricos recientes.

A pesar de estas limitaciones, el segundo modelo representa un avance significativo sobre su predecesor, principalmente porque reconoce y trata de adaptarse a la naturaleza inherentemente volátil del sistema que está modelando. Para mejorar aún más su precisión, sería beneficioso considerar la incorporación de variables estacionales, incluir variables explicativas adicionales como eventos especiales o días festivos, implementar modelos más sofisticados capaces de manejar eventos extremos, y desarrollar un sistema robusto de detección de anomalías. Estas mejoras potenciales podrían ayudar a cerrar la brecha entre las predicciones y los valores reales, especialmente en situaciones de alta variabilidad.

# Comparación del Modelo de Predicción con Valores Reales

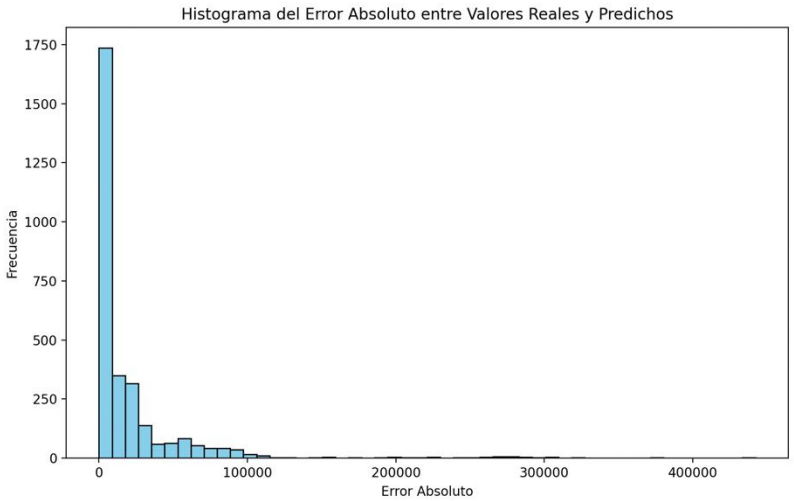



Al analizar los picos, se destaca el más alto alrededor del índice 2200, donde el error alcanza aproximadamente 450,000 unidades. Además, existen otros picos notables que llegan a alrededor de 300,000 a 350,000 unidades. Estos picos representan momentos en los que los valores predichos se desviaron significativamente de los valores reales, sugiriendo la presencia de errores de predicción considerables o valores atípicos en el modelo.

Por otro lado, los valles, que se encuentran cerca de cero, representan momentos en los que las predicciones fueron muy precisas. La mayoría de los puntos de datos se sitúan en estas regiones de bajo error, lo que indica que el modelo generalmente tiene un buen desempeño de base en sus predicciones. La presencia constante de estas áreas de bajo error sugiere que el modelo posee una capacidad sólida para realizar predicciones precisas en condiciones normales.

En cuanto a las observaciones del patrón, los errores parecen estar distribuidos de manera relativamente uniforme a lo largo del rango de índices, sin una tendencia evidente de aumento o disminución del error con el tiempo. Los picos de error son irregulares y parecen impredecibles, lo cual podría indicar que están relacionados con casos atípicos o puntos de datos inusuales específicos, más que con un fallo sistemático del modelo.

# Comparación del Modelo de Predicción con Valores Reales




Analizando el histograma titulado "Histograma del Error Absoluto entre Valores Reales y Predichos", se observa que la distribución muestra una fuerte asimetría positiva hacia la derecha. Esto indica que hay una alta concentración de errores cerca de cero, con una cola que se extiende hacia la derecha, mostrando una frecuencia decreciente a medida que el error absoluto aumenta. La mayor concentración de errores en valores bajos sugiere que el modelo predice con precisión en la mayoría de los casos.

Las características de la distribución reflejan que la barra de mayor frecuencia se encuentra en el extremo izquierdo, cerca de 0, con aproximadamente 1750 ocurrencias. La frecuencia cae drásticamente después de las primeras barras, lo que indica que los errores pequeños son mucho más comunes que los grandes. Sin embargo, existen algunas ocurrencias esporádicas de errores mayores que se extienden hasta alrededor de 400,000, lo cual representa casos en los que el modelo se desvía de manera significativa.

En términos de interpretación del rendimiento del modelo, la alta concentración de errores pequeños sugiere que el modelo es eficaz para la mayoría de las predicciones, manteniendo los errores absolutos cerca de 0. Sin embargo, la presencia de algunos casos atípicos, reflejados en la cola de la distribución, indica que el modelo comete algunos errores grandes en situaciones puntuales. La asimetría positiva de la distribución sugiere además que, cuando el modelo falla, tiende a hacer sobreestimaciones significativas en lugar de cometer errores pequeños y distribuidos de manera uniforme.

# Conclusiones

- En resumen, **el Modelo 3 es el mejor modelo predictivo**, ya que tiene los errores más bajos en ambas métricas (RMSE y MAE), lo que indica que realiza las predicciones con mayor precisión y menor margen de error. Este modelo, al incluir más variables relevantes, parece ser capaz de capturar una mayor cantidad de variabilidad en los datos, lo que lo hace más robusto y confiable para predecir el número de pasajeros en vuelos. A pesar de que el Modelo 2 también muestra una mejora respecto al Modelo 1, el Modelo 3 es claramente superior, lo que sugiere que la inclusión de más características (como el **'Operating Airline'** y el **'GEO Summary'**) en el modelo tiene un impacto positivo en su capacidad predictiva.