BIG DATA

PROYECTO FINAL

Carlos Ramírez Martín

03- Científico

New Technology School

Tokio.

Índice

Introducción

Presentación y Objetivos

Este documento contiene los resultados y conclusiones correspondiente a las tareas realizadas

El proyecto se centra en el análisis de un conjunto de datos relacionados con el tráfico aéreo en el aeropuerto de San Francisco, California.

Para obtener una comprensión completa del proyecto, se recomienda explorar la siguiente documentación:

- **01 - Configuración** *01_configuracion_carlos_ramirez_martin.pdf*
  - Documento donde se detalla el proceso de configuración e instalación de las dependencias necesarias para el proyecto.

- **02 - Técnico:** *02_tecnico_carlos_ramirez_martin.pdf*
  - Documento técnico donde se recoge el código fuente empleado para la resolución de cada una de las tareas solicitadas

- **04 - Presentación:** *04_presentacion_carlos_ramirez_martin.pdf*
  - Presentación que contiene los resultados del proyecto usando técnicas de storytelling.

Spark

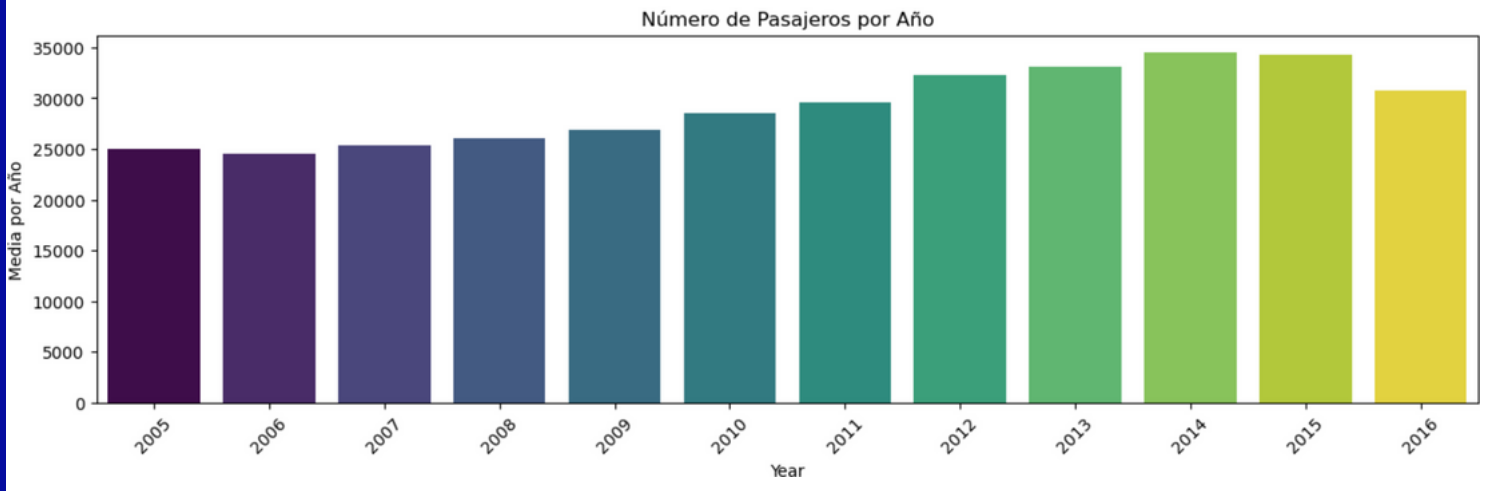
Análisis Descriptivo

Tras analizar las métricas estadísticas de las columnas **Passenger Count** y **Adjusted Passenger Count**, **se destaca una ligera discrepancia entre ambos conjuntos de valores**, donde los valores de **Adjusted Passenger Count** son ligeramente superiores. Es plausible que la **columna Passenger Count** represente el número de pasajeros que adquirieron el billete, mientras que **Adjusted Passenger Count** podría referirse al número de pasajeros que efectivamente tomaron el vuelo.
La media general de pasajeros se sitúa en **29,437.34** (sistema americano).

Mean Passenger Count	Std Passenger Count	Mean Adjusted Passenger Count	Std Adjusted Passenger Count
29345.62	58398.45	29437.34	58362.88

Al examinar los datos por año, **se observa un incremento constante en la media de pasajeros desde el 2005, sin interrupciones hasta el año 2014**. Sin embargo, en 2015 se registra una ligera disminución en el volumen de pasajeros, pasando de una media de **34,470.10** en **2014** a **34,292.53 en 2015**. El mayor descenso se identifica en **2016**, donde la media se sitúa en **30,811.85**, marcando así un punto notable de contracción en la tendencia ascendente observada en los años previos.

Se evidencia una distribución bastante uniforme a lo largo de todos los años, con excepción de 2005 y 2016. Es probable que esta disparidad se deba a que la recopilación de datos para estos periodos no sea completa. El año 2005 representa el primer año de recopilación de datos, y 2016 no abarca el año completo. *Sería pertinente evaluar la conveniencia de excluir los datos correspondientes a 2005 y 2016 con el fin de analizar periodos anuales completos.*

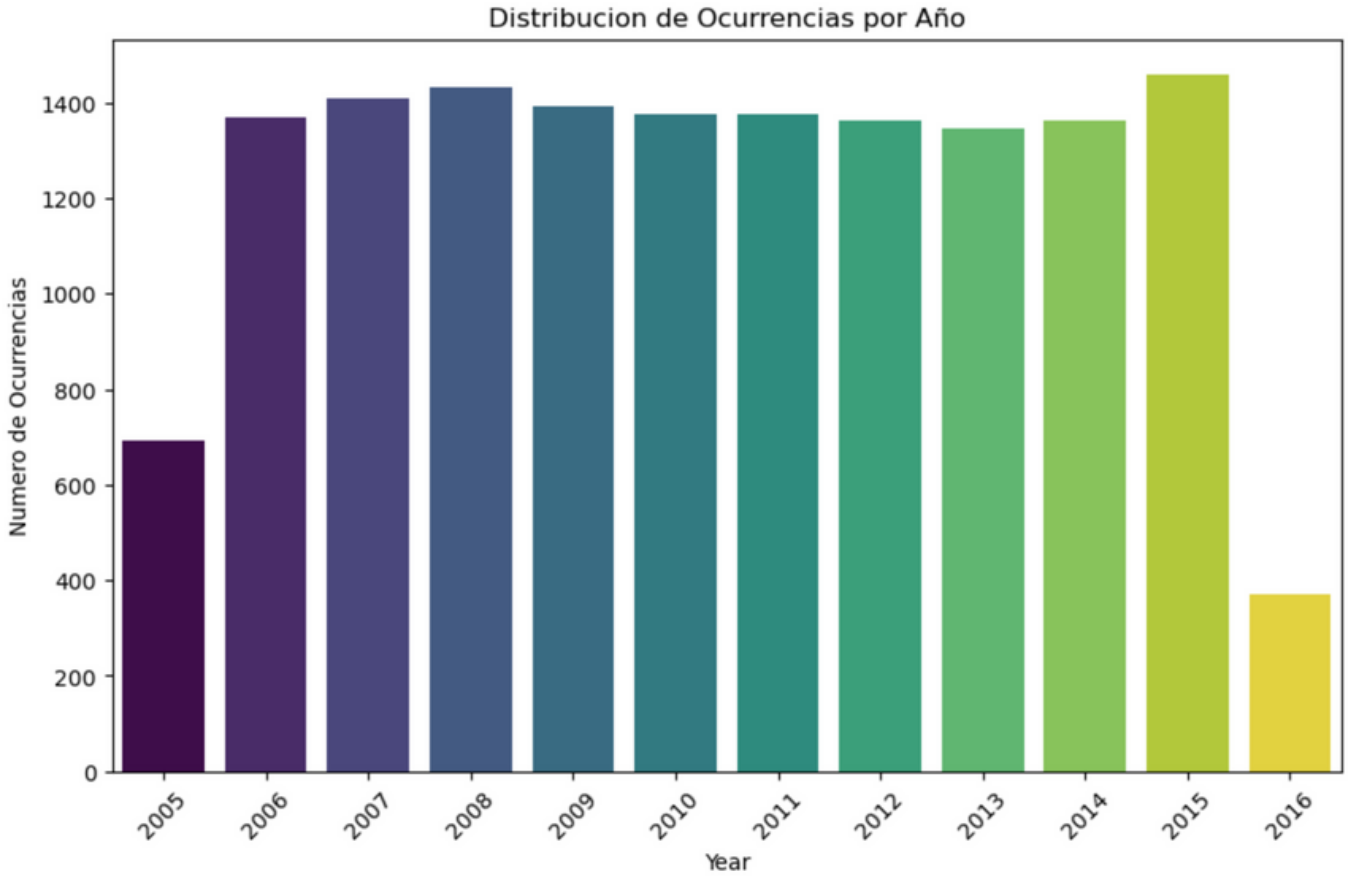


Al examinar el número de vuelos anuales, se observa que los extremos, periodo 2005 y 2016, coinciden con los años que presentan menos ocurrencias. Esta discrepancia posiblemente se deba a la falta de datos completos en ambos extremos temporales.

Es relevante resaltar que, entre los años 2008 y 2013, se observa una disminución en el número de vuelos. Contrariamente, el número de pasajeros durante ese mismo periodo ha experimentado un aumento significativo. Esta tendencia sugiere un incremento en la rentabilidad por vuelo a lo largo de los años.

Sin embargo, en el año 2015, se observa un aumento en el número de vuelos, acompañado de una ligera disminución en el número de pasajeros. Este fenómeno podría indicar un cambio en la dinámica del mercado o factores específicos que afectaron la demanda de vuelos durante ese periodo.

Durante 2016 desciende el numero de vuelos hasta el 2.48% pero el descenso en volumen de pasajeros no es tan pronunciado

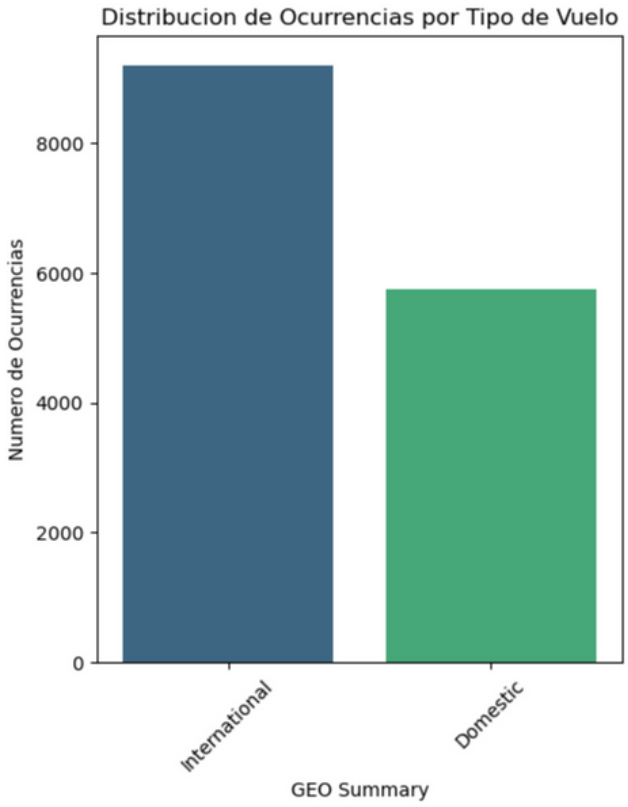


Las observaciones realizadas revelan l**a presencia de un total de 73 aerolíneas** en nuestro conjunto de datos. Entre ellas, **American Airlines** destaca con una impresionante media de pasajeros de **127,164.39**, seguida por **Southwest Airlines** con **81,223.35**, **Virgin America** con **74,405.35** y **United Airlines** con **72,827.22**. Estas aerolíneas son líderes en volumen de pasajeros, todas superando la media de **70,000** pasajeros.

En el extremo opuesto de la escala, encontramos a **Evergreen International Airlines 2**, **Ameriflight 5.36**, **Atlas Air, Inc 35.5**, y **Xtra Airways 73**. Ninguna de estas aerolíneas logra alcanzar **la media de 100 pasajeros**, situándose en una categoría de menor volumen en comparación con sus contrapartes líderes.

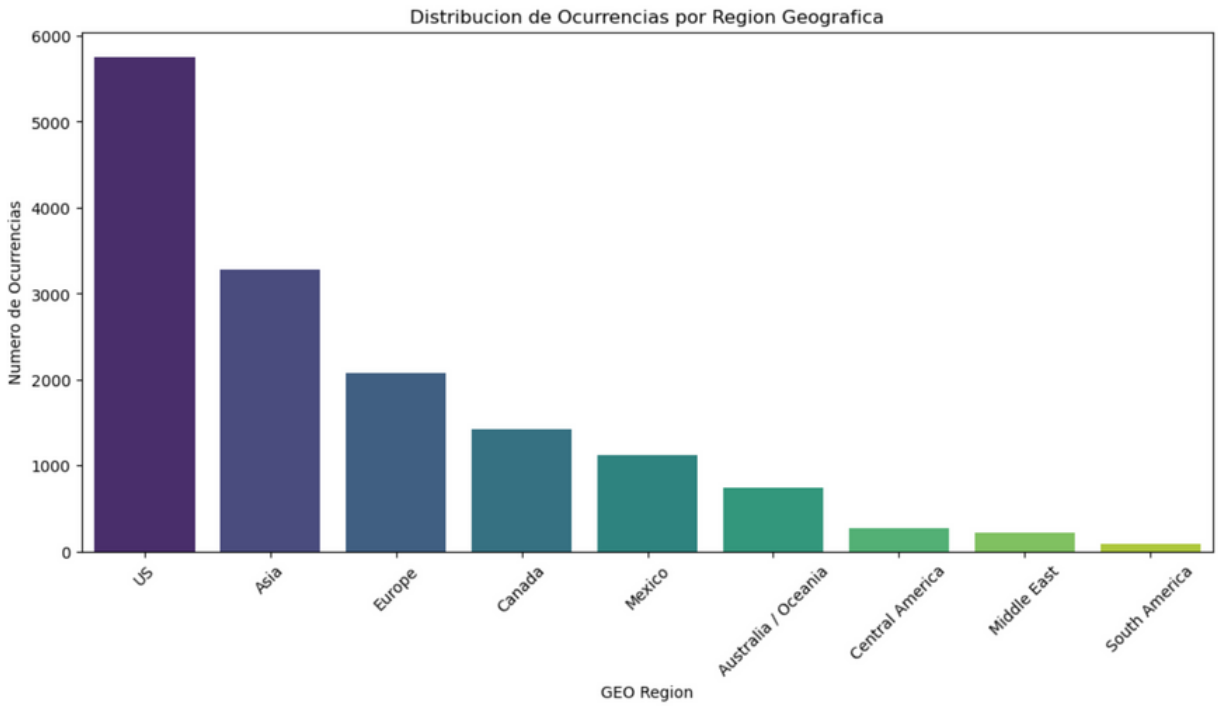
Operating Airline	Mean Passenger Count	Std Passenger Count	Mean Adjusted Passenger Count	Std Adjusted Passenger Count
American Airlines	127164.39	22044.24	127164.39	22044.24
Southwest Airlines	81188.16	60358.22	81223.35	60310.95
Virgin America	74405.35	68539.61	74405.35	68539.61
United Airlines	72732.06	111407.61	72827.22	111353.45
Delta Air Lines	68498.5	52441.71	68515.42	52420.02
US Airways	55317.82	17368.96	55317.82	17368.96
United Airlines -...	48915.47	101345.43	49365.52	101159.05
SkyWest Airlines	37083.84	47114.39	37083.88	47114.36
JetBlue Airways	35261.14	15781.56	35261.14	15781.56
Northwest Airlines	26109.25	23299.2	26205.5	23201.11
Compass Airlines	23358.56	13643.17	23359.84	13640.96
Lufthansa German ...	19301.97	3158.42	19301.97	3158.42
Air Canada	18251.56	8036.23	18251.56	8036.23
Frontier Airlines	17787.68	4894.09	17787.68	4894.09
British Airways	17625.12	2490.01	17625.12	2490.01
Alaska Airlines	17251.64	16964.72	17564.68	16729.12
Cathay Pacific	17121.33	4000.22	17121.33	4000.22
Singapore Airlines	14746.65	1969.16	14746.65	1969.16

Indicar que el **61.5%** de los vuelos registrados son considerados como vuelos **internacionales**, frente al **38.5% de procedencia doméstica.**



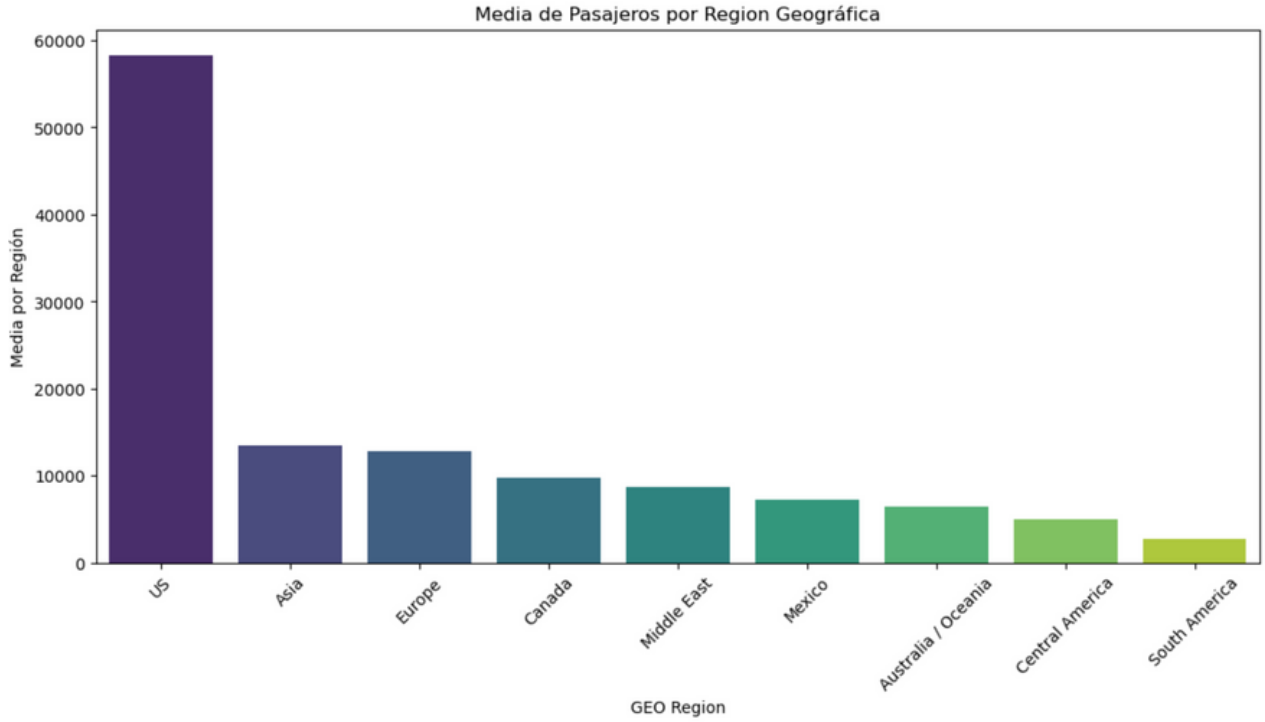
Los datos revelan que las compañías aéreas que operan en **Estados Unidos registran la mayor cantidad de vuelos**, alcanzando un total de **5,757**, lo que representa el **38.50% de todo el conjunto de dato**s. **Asia y Europa ocupan el segundo y tercer lugar respectivamente**, con **3,272 y 2,078** vuelos cada uno, constituyendo el **21.88% y el 13.90**%.

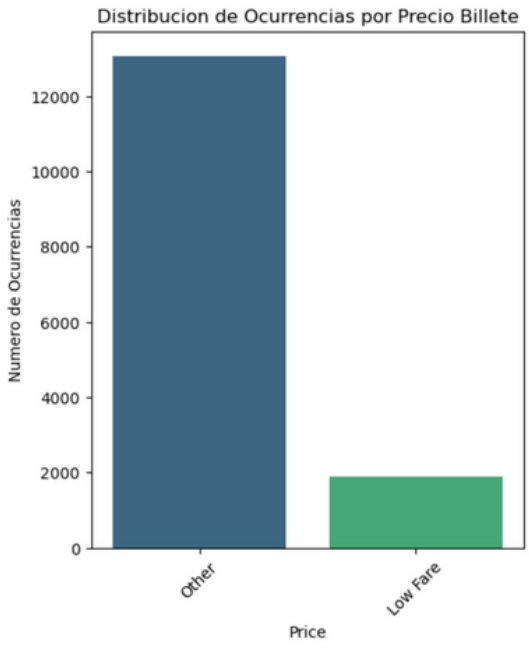
En contraste, **América del Sur, Oriente Medio y América Central presentan una menor frecuencia**, con **90, 214 y 272** vuelos respectivamente. Estas cuatro zonas geográficas combinadas **no alcanzan ni siquiera el 4% del total de los datos recopilados en el conjunto de datos.**



Como era de anticipar, **la correlación entre el número de vuelos y el volumen de pasajeros es positiva**. En consecuencia, la zona de **Estados Unidos** exhibe una media de pasajeros de **58,330.34**, seguida por **Asia con 13,435** y **Europa con 12,755.65**. Este patrón sugiere una relación directa entre la cantidad de vuelos y el volumen de pasajeros, donde regiones con más vuelos tienden a tener mayores medias de pasajeros.

Un dato que me ha sorprendido es la distribución en función del costo del billete. **Se destaca que el 87.4% de los vuelos realizados han sido abonados con una tarifa denominada como other**, la cual entendemos que hace referencia a tarifas sin descuento o a la tarificación de billetes en clases preferentes. Los billetes más asequibles, comúnmente conocidos como **low cost**, representan el **12.6%**, evidenciando así una marcada predominancia de la categoría other en la estructura de tarifas.





Spark

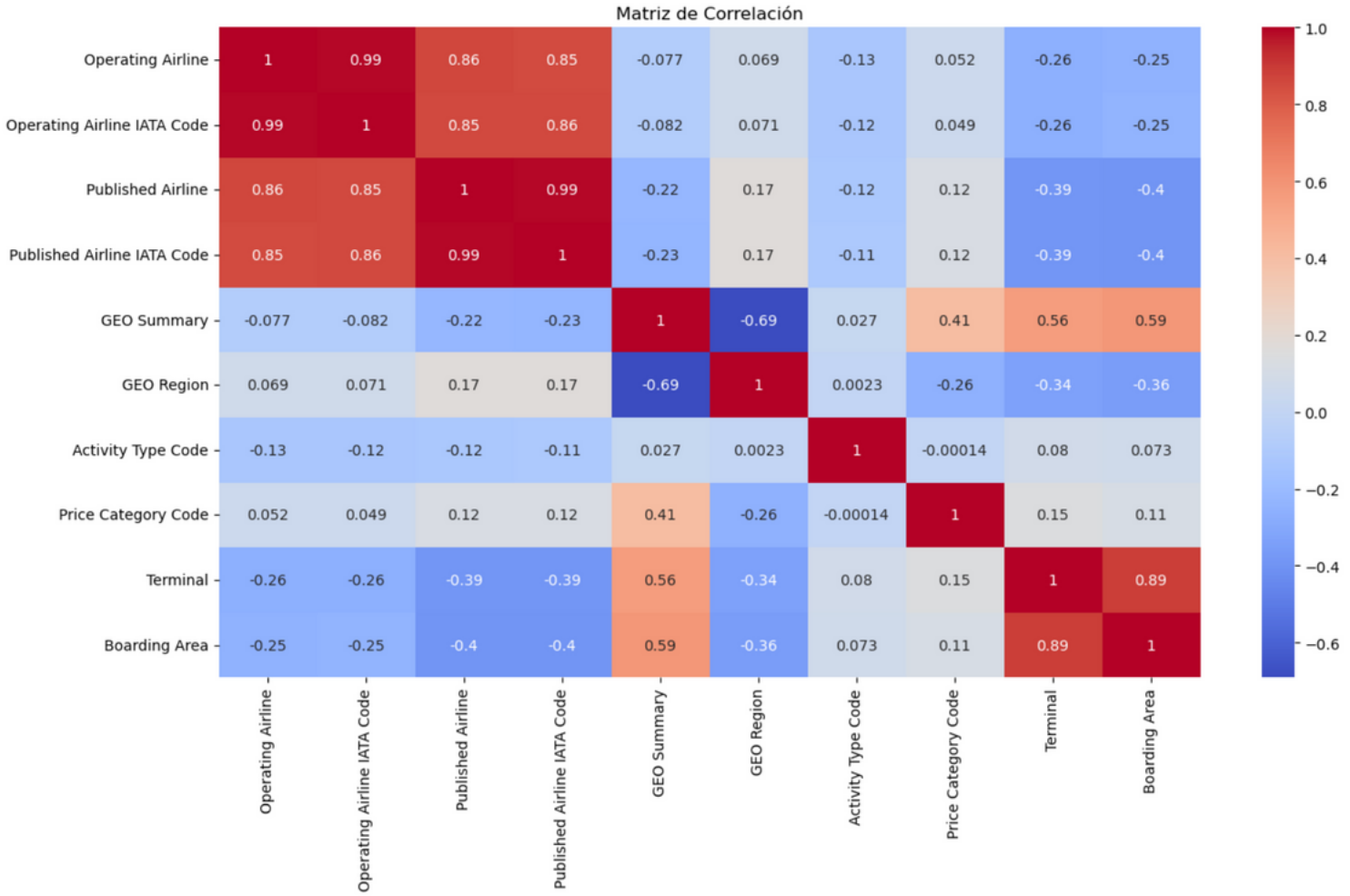
Análisis Correlativo

La matriz de correlación proporciona una visión detallada de las relaciones lineales entre las variables categóricas convertidas en índices numéricos. A continuación comentamos algunas observaciones clave derivadas de esta matriz:

- **Correlación Positiva Fuerte entre Aerolíneas:**
  Las columnas relacionadas con aerolíneas, como **Operating Airline**, **Operating Airline IATA Code**, **Published Airline**, y **Published Airline IATA Code**, exhiben correlaciones positivas cercanas a 1. Esto indica una fuerte asociación entre estas variables, lo cual es razonable ya que representan información similar.
  Así mismo observamos una alta correlación entre la **Terminal** y **Boargind Area**, esta es una observación lógica ya que entendemos que el aeropuerto tiene asignado una áreas específicas de embarque a cada terminal.

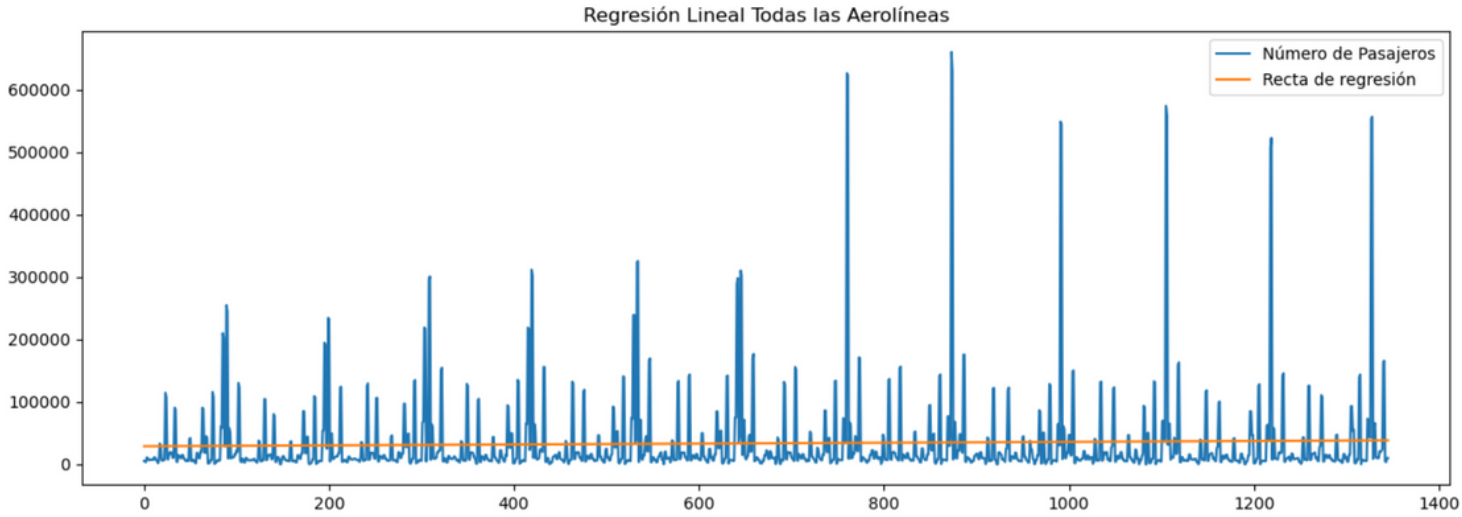
- **Correlaciones Negativas con GEO Summary**:
  La columna **GEO Summary** muestra correlaciones negativas con algunas de las variables relacionadas con aerolíneas. Este resultado sugiere que **ciertas zonas geográficas pueden estar asociados inversamente con ciertos aspectos de la actividad de las aerolíneas**. Es posible que algunas áreas geográficas tengan una mayor actividad de aerolíneas domesticas, mientras que otras estén más orientadas a vuelos internacionales. Esto también puede ser debido a los acuerdos políticos entre paises como restricciones de vuelo o acuerdos comerciales que puedan afectar a las operaciones de las aerolíneas con un perfil más internacional.

- **Variables con Correlaciones Cercanas a Cero:**
  Algunas de las correlaciones entre variables son cercanas a cero, indicando una falta de relación lineal. Por ejemplo, la correlación entre **GEO Summary y Activity Type Code** es relativamente baja, sugiriendo que estas dos variables pueden ser independientes entre sí.



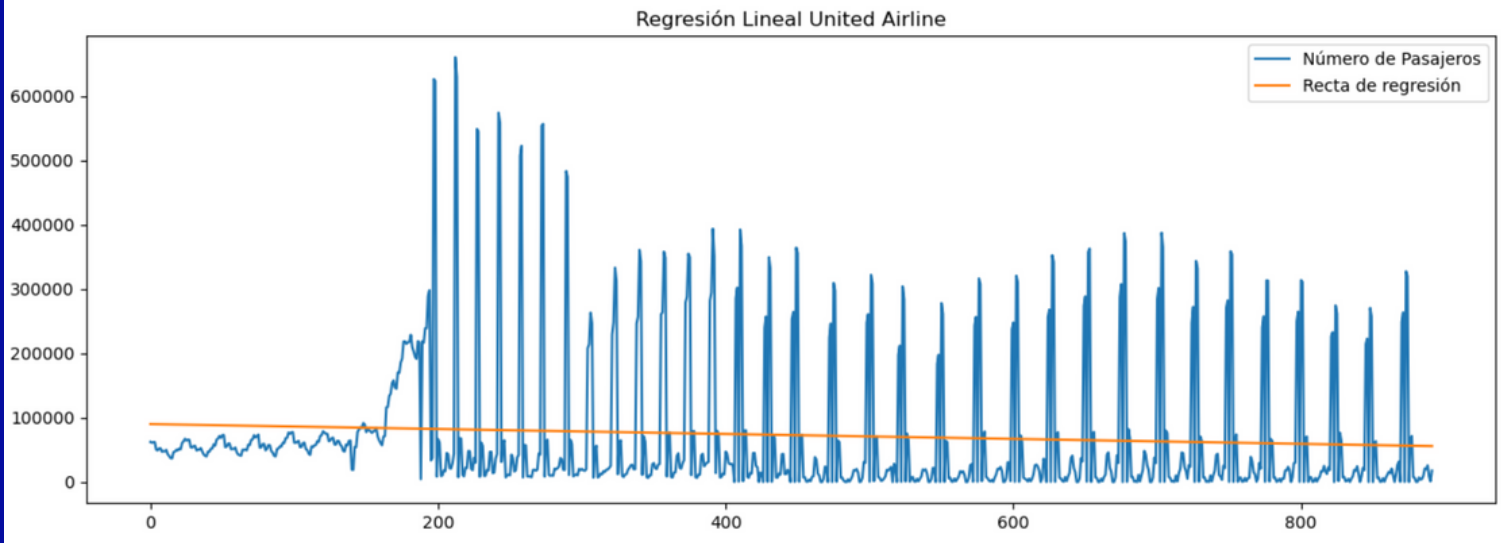
Realizamos un análisis de la correlación lineal considerando: Todas las aerolíneas; United Airlines y Air France.

Comparamos el volumen de pasajeros entre el primer y segundo periodo del año. El objetivo es examinar si, durante el primer periodo, caracterizado por una menor incidencia de festividades y periodos vacaciones, la cifra de pasajeros tiende a disminuir. En contraste, evaluamos si, en el segundo periodo del año, la correlación cambia y se observa un aumento en el número de pasajeros.

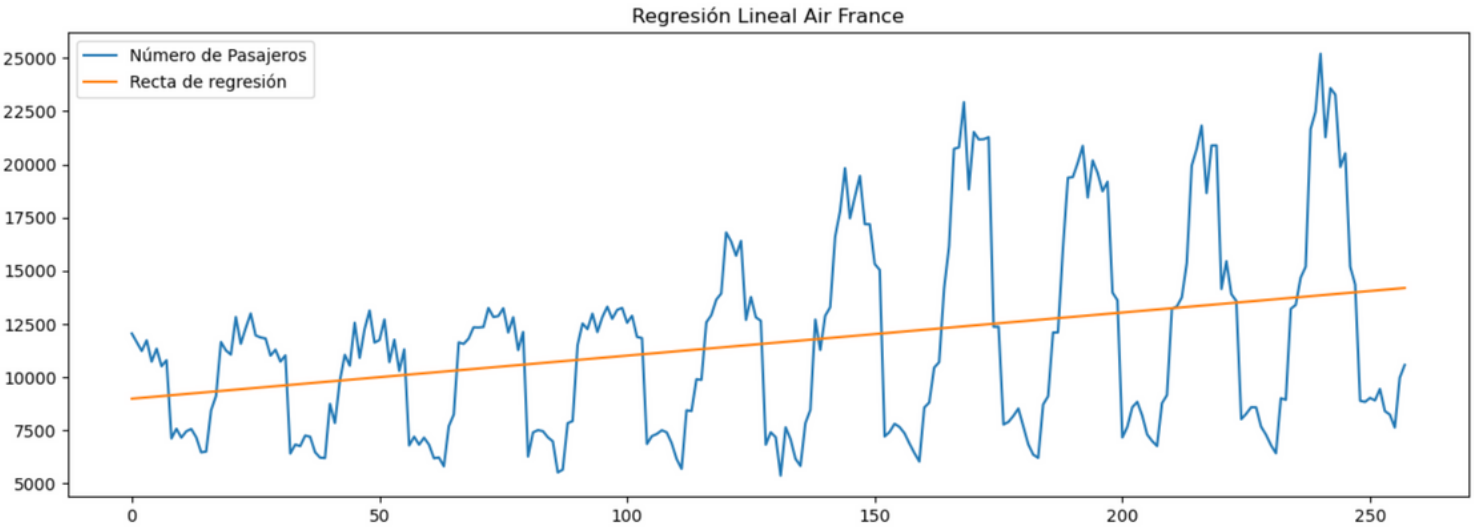


En cuanto a la correlación teniendo en cuenta todas las aerolíneas, observamos que la correlación entre ambos periodos es muy cercana a cero, concretamente 0.14 puntos.

Realizando el mismo ejercicio pero esta vez, solo sobre la aerolínea con más actividad, United Airline, tampoco encontramos correlación.



En cambio, al analizar los datos de Air France, si observamos una correlación negativa, esto nos indica que siempre observaremos un periodo con más pasajeros que el otro



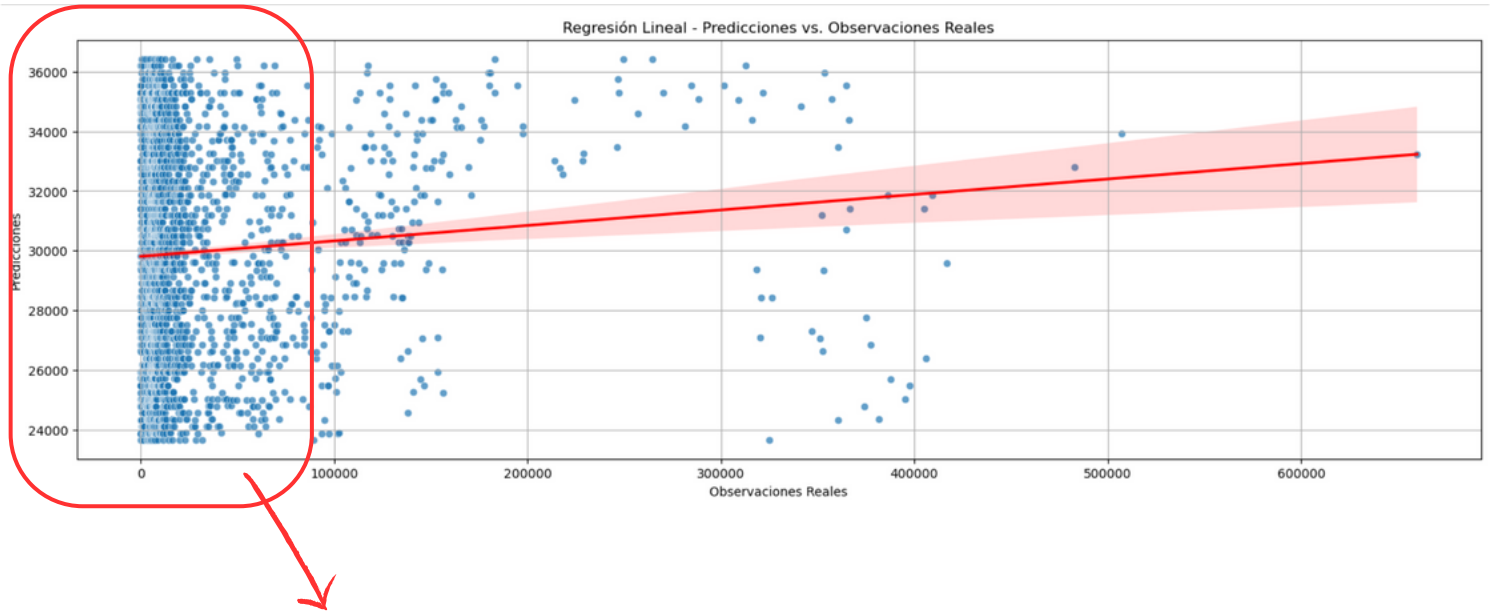
Spark

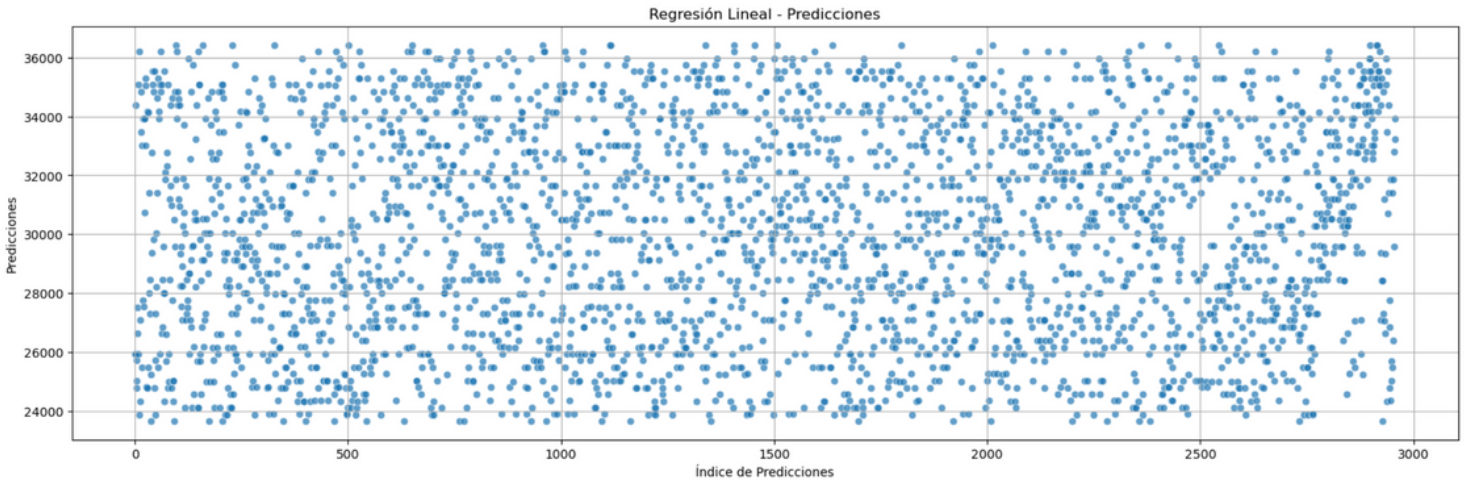
Algoritmo de Regresión Lineal

Planteamos el uso de un algoritmo de regresión lineal con el objeto de predecir la tendencia de pasajeros a lo largo del tiempo.
El modelo fue construido utilizando las variables **Year y Month como predictores** para la variable **objetivo Adjusted Passenger Count.**

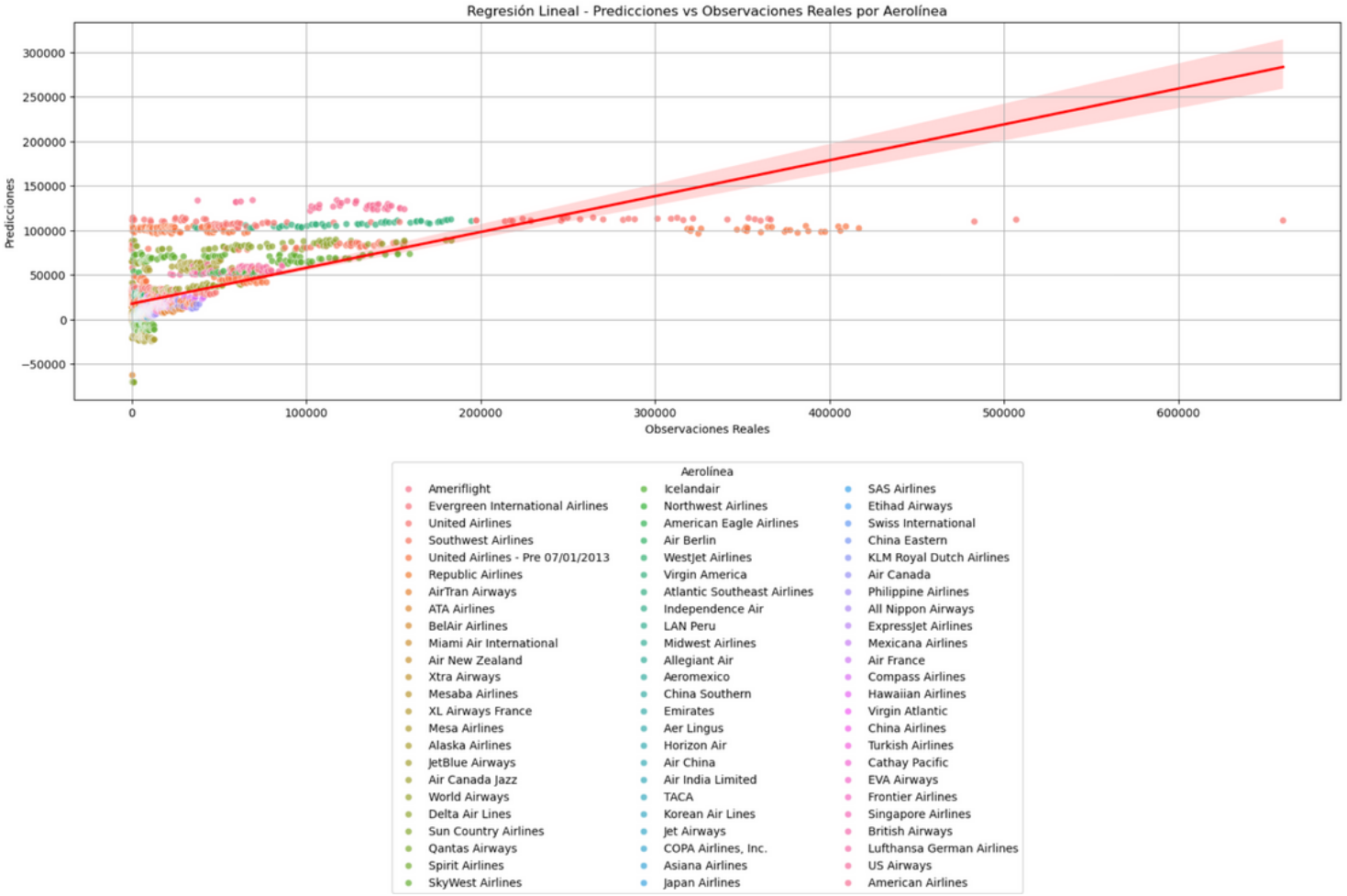
El modelo de regresión lineal ha sido entrenado utilizando un conjunto de datos de entrenamiento que representa el 80% de los datos originales, mientras que el 20% restante se utilizó como conjunto de prueba para evaluar la capacidad predictiva del modelo. A continuación, se presentan los resultados de las predicciones realizadas sobre el conjunto de prueba.

Las predicciones obtenidas revelan una dispersión considerable entre los valores, indicando la necesidad de incorporar más variables correlacionadas para refinar las estimaciones.

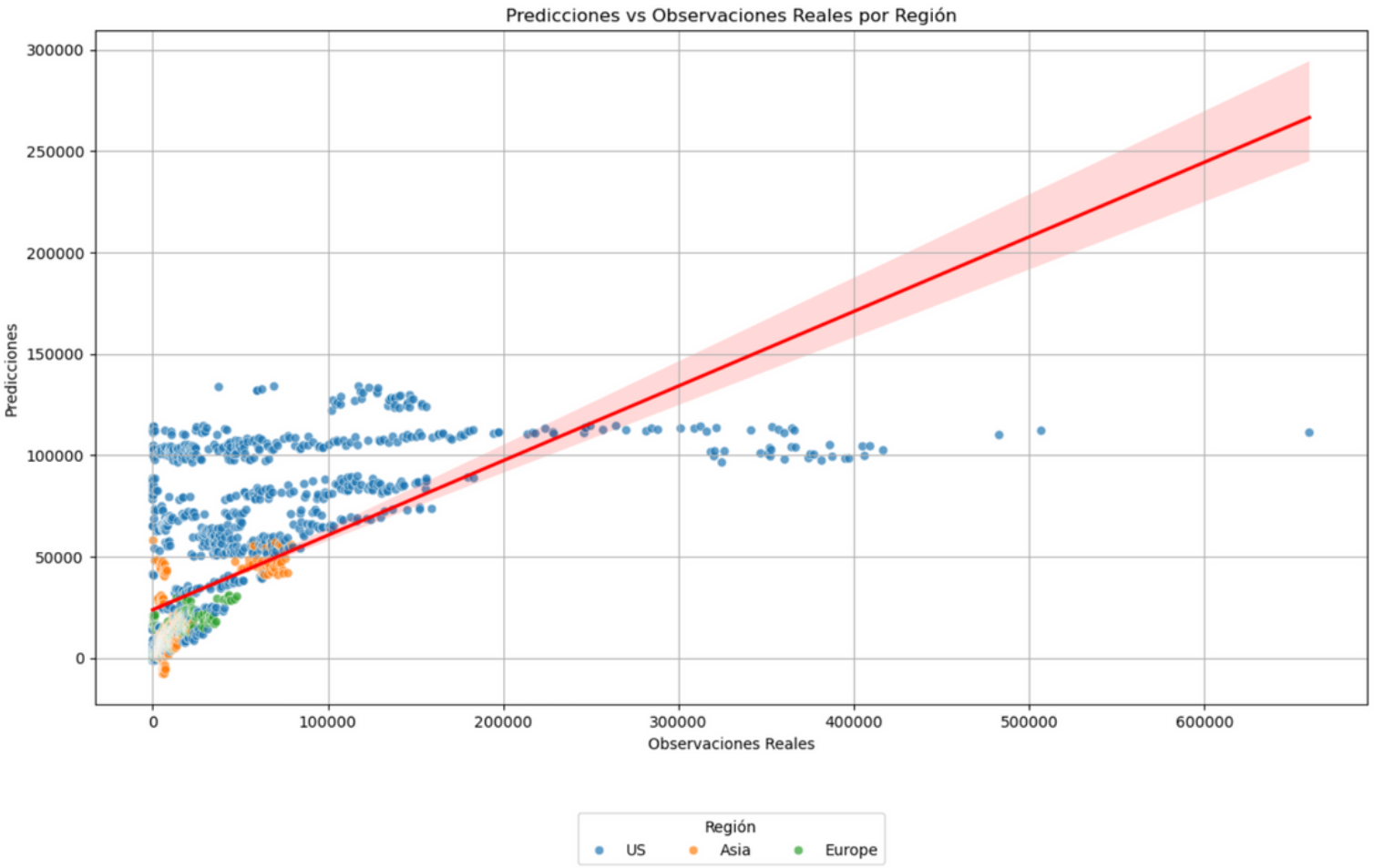




Para solventarlo añadimos los datos referentes a la región, ya que la variable mostró una correlación durante el análisis. No obstante, los resultados no fueron concluyentes. Aunque la dispersión no es tan marcada, no pudimos identificar correlaciones entre aerolíneas que compartan los mismos destinos.



Por último, en este tercer escenario, capacitamos al algoritmo para prever las tendencias futuras en las regiones más populares, Estados Unidos, Asia y Europa, LA dispersión de los datos en este caso no es tan notable en escenarios anteriores. Por ejemplo, para Europa, pudimos observa una relación lineal dévil entre las variables predictoras y la variable de respuesta


Predicciones vs Observaciones Reales por Región
Predicciones
Observaciones Reales
300000
250000
200000
150000
100000
50000
0
0
100000
200000
300000
400000
500000
600000
Región
US
Asia
Europe