BIG DATA

PROYECTO FINAL

Atanas Turlakov

Tokio.

Introduccion

Este proyecto se inscribe en el contexto del negocio de viajes y turismo, centrando su atención en el análisis del tráfico aéreo desde San Francisco hacia Tokio y otras importantes ciudades de Japón. Ante el creciente interés y la significativa demanda de viajes entre estos destinos, nuestro estudio se propone examinar detalladamente las tendencias de tráfico aéreo para identificar patrones, picos de demanda, y otros factores críticos que influyen en el flujo de pasajeros.

El objetivo principal de esta investigación es proporcionar un análisis exhaustivo que sirva de base para la toma de decisiones estratégicas futuras, tanto para mejorar la oferta de servicios como para optimizar las operaciones de las aerolíneas y agencias de viajes involucradas. A través de este estudio, buscamos ofrecer insights valiosos que contribuyan a la planificación y desarrollo de estrategias más efectivas en el sector de viajes y turismo, mejorando así la experiencia de los viajeros y potenciando el crecimiento del mercado.

Metodologia

- Enfoque del analizis - Nuestro análisis se ha estructurado en varias etapas clave para garantizar una comprensión profunda y detallada del tráfico aéreo entre San Francisco y Japón. Las fases incluyen:
- Análisis Exploratorio: Mediante técnicas de análisis descriptivo, exploramos las tendencias generales y patrones en los datos. Esto incluyó la evaluación de volúmenes de tráfico por temporada, análisis de frecuencias y distribuciones de vuelos.
- Análisis Cuantitativo: Profundizamos en el estudio mediante análisis estadísticos y modelado predictivo para entender los factores que influyen en el tráfico aéreo y predecir tendencias futuras.
- Preparación de Datos: Iniciamos con la recopilación y limpieza de datos de múltiples fuentes, asegurando que la información sea precisa y esté actualizada. Este proceso implicó la eliminación de registros duplicados, el manejo de valores faltantes y la estandarización de formatos.

Metodologia

- **Visualización de Datos: Para facilitar la interpretación de nuestros hallazgos, desarrollamos una serie de visualizaciones, incluyendo gráficos de tendencias, mapas de calor y diagramas de dispersión.**

  Herramientas y tecnologias: La realización de este análisis fue posible gracias al uso de avanzadas herramientas y tecnologías de big data y análisis de datos, incluyendo: PySpark, Python, Bash, Docker, Cassandran y otros

Desarollo

Para empesar con el estudio trabajaremos con datos desde un archivo .csv. Primero vamos a categorizar los datos en el archivo en una tabla para poder planificar los pasos siguientes. Podemos ver claramente el tipo de datos con los que podemos trabajar. Tenemos dos tipos de datos de texto y numericos los que usaremos.

Nombre del Campo	Tipo de Dato
Activity Period	Numérico
Operating Airline	Categórico
Operating Airline IATA Code	Categórico
Published Airline	Categórico
Published Airline IATA Code	Categórico
GEO Summary	Categórico
GEO Region	Categórico
Activity Type Code	Categórico
Price Category Code	Categórico
Terminal	Categórico
Boarding Area	Categórico
Passenger Count	Numérico
Adjusted Activity Type Code	Categórico
Adjusted Passenger Count	Numérico
Year	Numérico
Month	Categórico

Para profundizar en nuestro análisis, realizamos consultas especializadas para recuperar registros específicos de las aerolíneas 'Air China' y 'AirBerlín'. Estas consultas nos ayudaran a tener una vision mas clara sobre el trafico generado por estas dos companias. Tenemos en la primera consulta todos los registros de “Air China” y en la segunda todos los registros de “Air Berlin” embarcados por la puerta G

Proyecto Final Big Data

```
##############################   Recuperar todos los registros de la aerolínea “Air China”   ##############################
```

Published Airline	Published Airline IATA Code	Terminal	Boarding Area	Passenger Count	Month	Year	GEO Region	Price Category Code	Activity Type Code
Air China	CA	International	G	7187	March	2011	Asia	Other	Deplaned
Air China	CA	International	G	7154	May	2007	Asia	Other	Deplaned
Air China	CA	International	G	2767	February	2006	Asia	Other	Enplaned
Air China	CA	International	G	8187	September	2014	Asia	Other	Deplaned
Air China	CA	International	G	7467	December	2012	Asia	Other	Enplaned
Air China	CA	International	G	7836	March	2015	Asia	Other	Enplaned
Air China	CA	International	G	5596	November	2008	Asia	Other	Deplaned
Air China	CA	International	G	7659	March	2014	Asia	Other	Deplaned
Air China	CA	International	G	7188	November	2011	Asia	Other	Deplaned
Air China	CA	International	G	6693	November	2011	Asia	Other	Enplaned
Air China	CA	International	G	7653	August	2012	Asia	Other	Enplaned
Air China	CA	International	G	6883	March	2012	Asia	Other	Enplaned
Air China	CA	International	G	3078	January	2006	Asia	Other	Enplaned
Air China	CA	International	G	7484	December	2011	Asia	Other	Enplaned
Air China	CA	International	G	7078	April	2012	Asia	Other	Deplaned
Air China	CA	International	G	2862	February	2006	Asia	Other	Deplaned
Air China	CA	International	G	7446	October	2014	Asia	Other	Enplaned
Air China	CA	International	G	8771	July	2013	Asia	Other	Deplaned
Air China	CA	International	G	3800	February	2009	Asia	Other	Enplaned

- En esta tabla nos podemos centrar en el volumen de pasajeros, destrubucion temporal y el precio. Estas variables nos pueden indicar como cambian los valores. Los datos muestran variaciones en “Passenger Count” lo que indica dependecnias estacionales sobre la demanda. Se ve claramente que todos los vuelos estan embarcados por la puerta G lo que puede indicar que es una area asignada especificamente a “Air China”

  Nuestro análisis de los registros de Air China revela variaciones significativas en el tráfico de pasajeros que podrían estar ligadas a factores estacionales o a cambios en las preferencias de los viajeros.

Proyecto Final Big Data

```
##############################  Recuperar todos los vuelos de la compañía “AirBerlín”embarcados porla puerta “G” ##############################
```

Published Airline	Published Airline IATA Code	Terminal	Boarding Area	Passenger Count	Month	Year	GEO Region	Price Category Code	Activity Type Code
Air Berlin	AB	International	G	2357	September	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2620	July	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2085	August	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2343	September	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	1455	October	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2261	July	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	972	May	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2294	August	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	1337	May	2010	Europe	Other	Deplaned
Air Berlin	AB	International	G	2548	June	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	1689	October	2010	Europe	Other	Enplaned
Air Berlin	AB	International	G	2158	June	2010	Europe	Other	Deplaned

```
nasko@desktop:~$
```

Tokio.

La consistencia en la utilización de la puerta 'G' para los vuelos de Air Berlin sugiere una operación optimizada que podría influir en la satisfacción del cliente y la eficiencia operacional. El uso significativo tanto de “Air Berlin” como de “Air China”

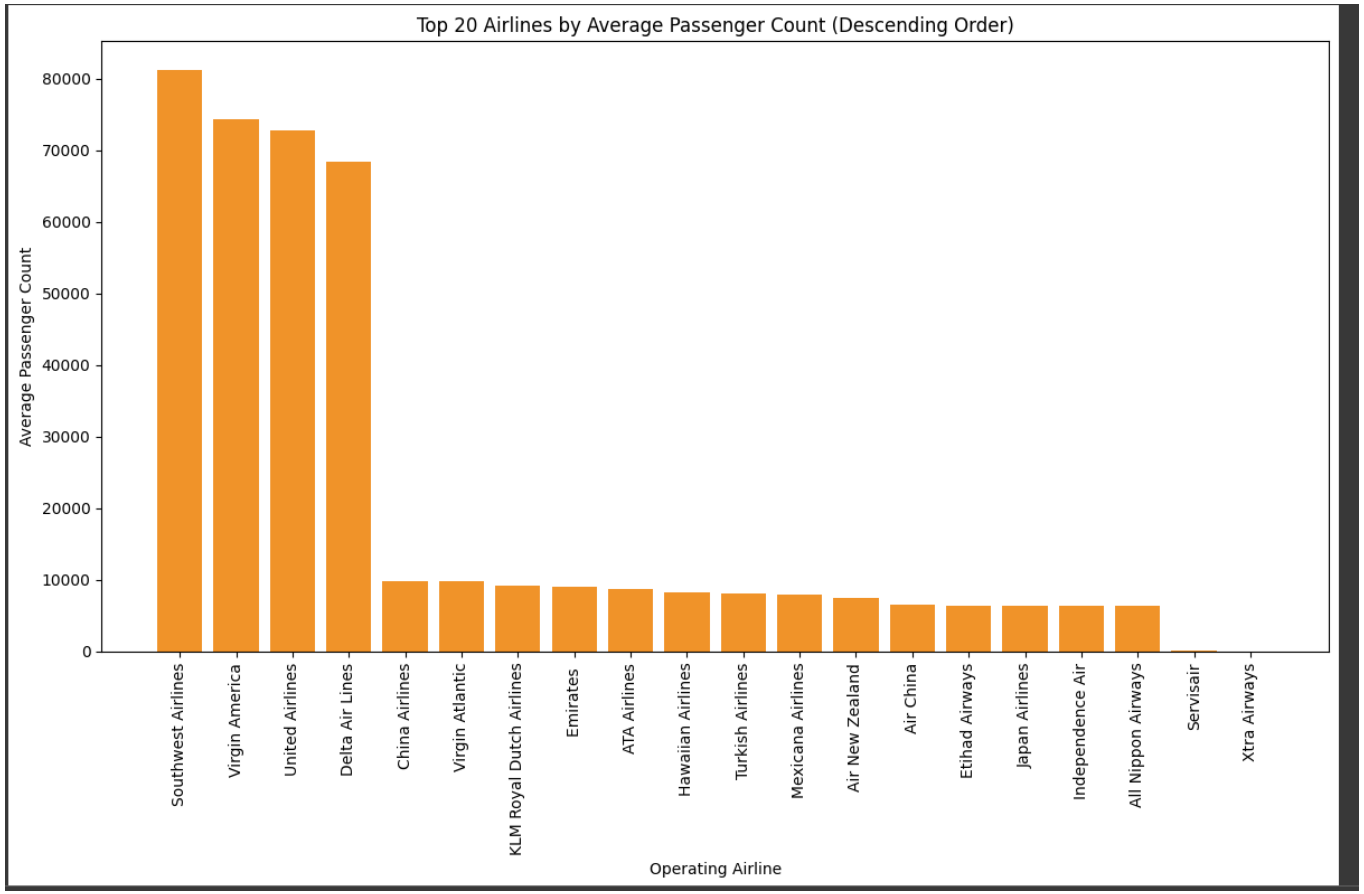
En la siguente consulta utilizaremos la tecnologia PySpark para visualizar las diferentes companias operando en el aeropuerto de San Francisco. El resultado muestra los primeros 20 de 77

Operating Airline
Icelandair
Ameriflight
Cathay Pacific
Aeromexico
Etihad Airways
Philippine Airlines
United Airlines -...
Turkish Airlines
Swiss International
Independence Air
Miami Air Interna...
Air France
Japan Airlines
Midwest Airlines
Atlas Air, Inc
JetBlue Airways
China Eastern
Mexicana Airlines
Air Canada
Allegiant Air

Tokio.

Proyecto Final Big Data

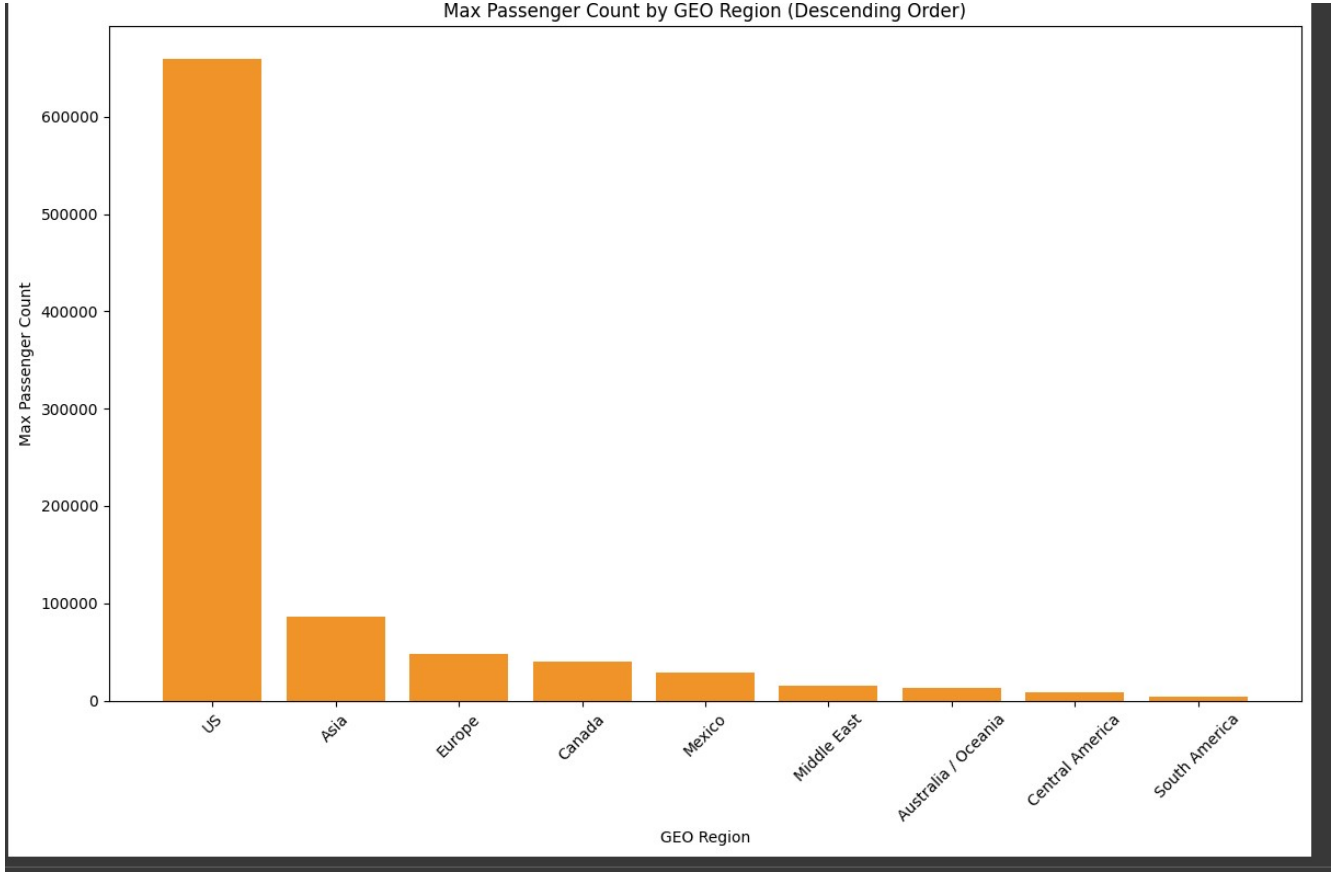
Aqui tenemos un grafico que representa la media de pasajeros por cada compania de los primeros 20. Se puede ver un claro contraste entre las companias americanas y el resto debido a la ubicación.



Tokio.

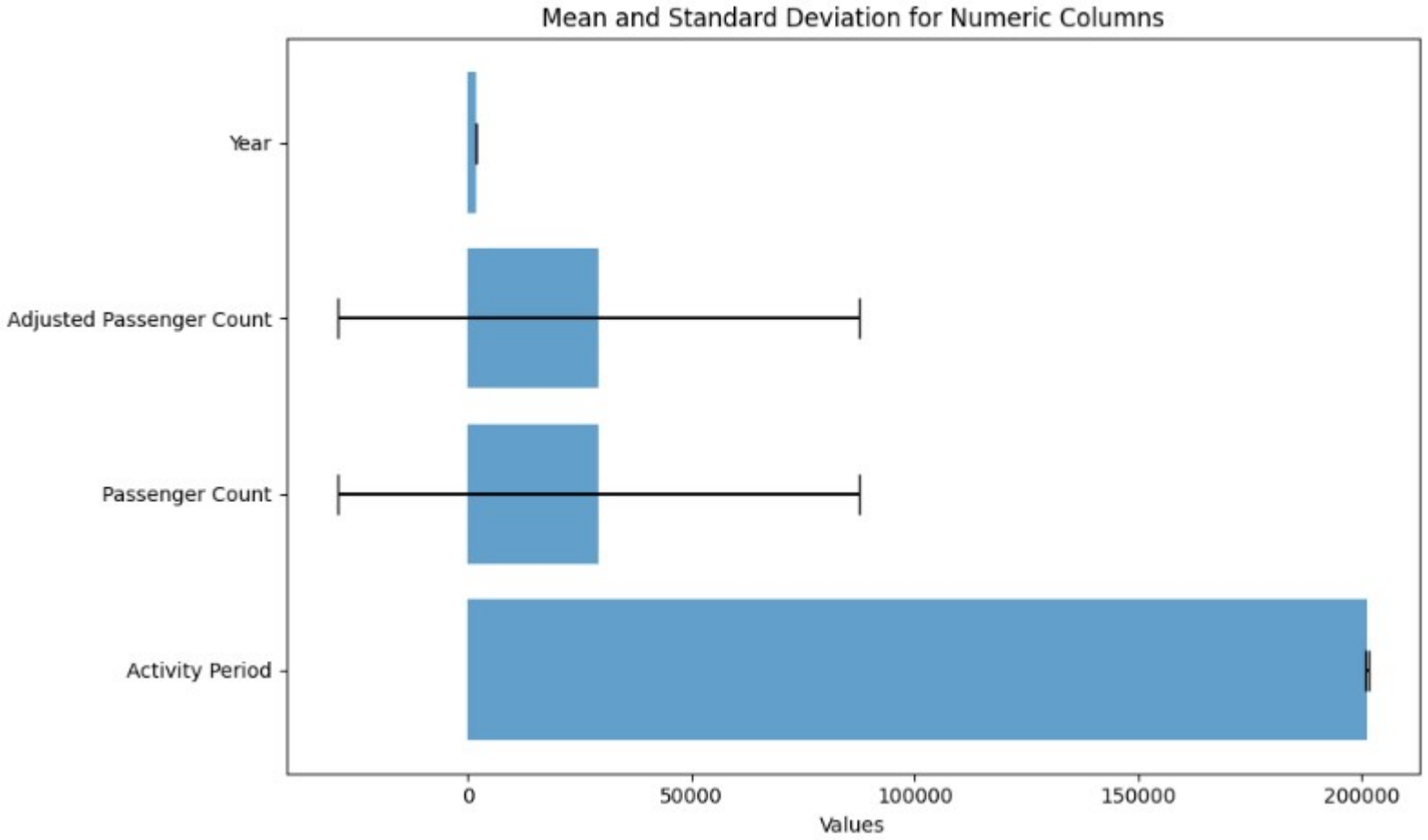
Proyecto Final Big Data

En el siguiente grafico podemos observar el numero de pasajeros dividido por regiones. Podemos observar tambien que la region con mas pasajeros es US por la distancia y numero de vuelos pero el segundo es Asia probablermente debido a la poblacion ya que es casi 6 vesez mas que Europa



Tokio.

Analizis estadistico: Realizamos un análisis estadístico para comprender mejor las tendencias y variabilidades en los datos de tráfico aéreo. Para el proposito calculamos la media y la desviación estándar para diferentes variables numéricas del conjunto de datos



Tokio.

Este gráfico nos muestra la media y la variabilidad de cuatro variables clave en nuestro conjunto de datos. Nos permite ver no solo el promedio de cada variable sino también qué tan dispersos están los datos alrededor de ese promedio, lo cual es indicado por las barras de error que representan la desviación estándar

Activity Period: La variable 'Activity Period' muestra una media alta con una desviación estándar relativamente pequeña, lo que sugiere que los datos de este período están concentrados cerca del promedio.

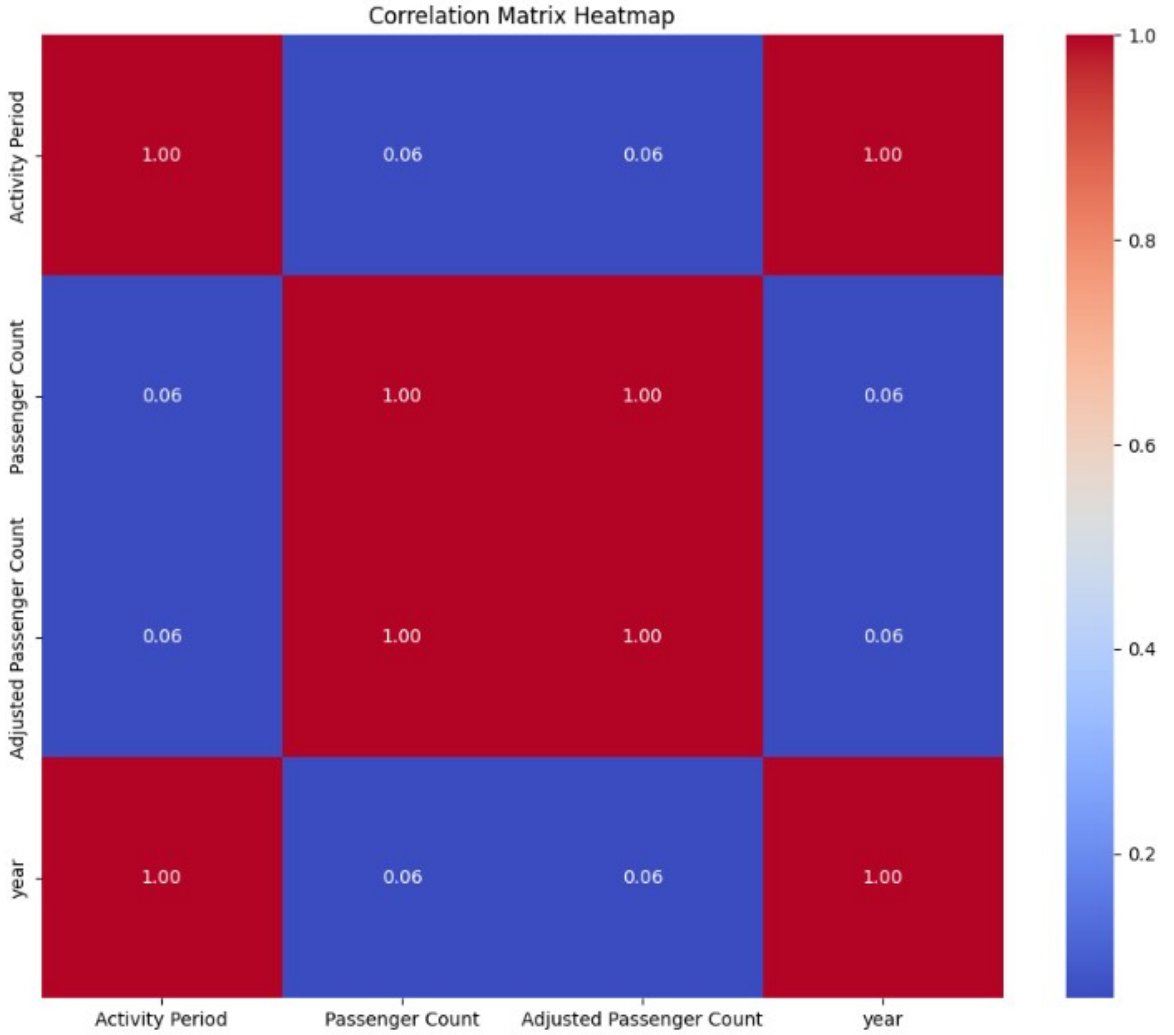
Passenger Count: En cuanto al 'Passenger Count', la media es menor que la del 'Activity Period', pero con una desviación estándar más amplia, indicando una mayor variabilidad en el número de pasajeros.

Adjusted Passenger Count: La 'Adjusted Passenger Count' sigue un patrón similar al 'Passenger Count', lo que puede reflejar fluctuaciones en los números ajustados de pasajeros a lo largo del tiempo.

Year: La variable 'year' tiene una escala diferente, lo que se refleja en la desviación estándar extremadamente pequeña en comparación con su media, indicando que los datos de 'year' son muy consistentes y varían poco.

Conclusión: Estas estadísticas son esenciales para comprender la estabilidad y las tendencias en el tráfico aéreo. Por ejemplo, la baja variabilidad en 'year' muestra consistencia a lo largo del tiempo, mientras que la alta variabilidad en el 'Passenger Count' podría señalar diferentes factores que afectan el volumen de pasajeros, como la temporada, eventos especiales o cambios en la industria del turismo.

Visualización de la Matriz de Correlación para Datos de Tráfico Aéreo: El mapa de calor de la matriz de correlación nos permite ver la relación entre diferentes variables numéricas de nuestro conjunto de datos. Los valores cercanos a 1 o -1 indican una fuerte correlación positiva o negativa, respectivamente, mientras que valores cercanos a 0 indican una falta de correlación



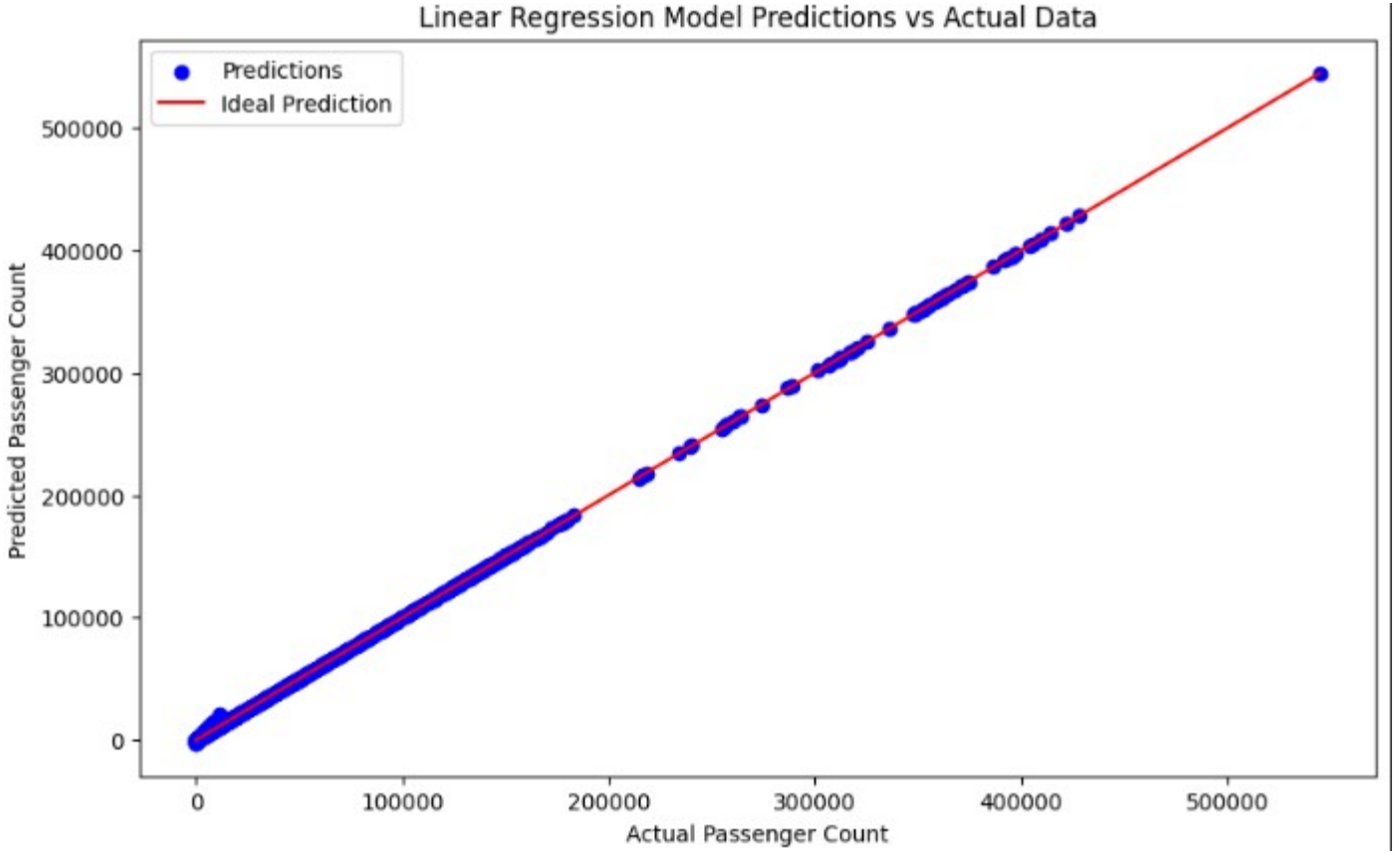
Tokio.

Analizar las variables:
Activity Period y Year: Se observa una correlación casi perfecta (1.00), lo que sugiere que estas dos variables están altamente relacionadas, posiblemente debido a que el 'Activity Period' está directamente relacionado con el 'Year'

Passenger Count y Adjusted Passenger Count: También muestran una correlación casi perfecta (1.00), indicando que cualquier ajuste realizado a la cuenta de pasajeros conserva la relación subyacente en los datos

Correlaciones Bajas: Las correlaciones entre 'Activity Period' y 'Passenger Count', así como entre 'Year' y 'Passenger Count' son bajas (0.06), lo que implica que no hay una relación lineal fuerte entre estas variables.

Evaluacion del modelo de prediccion lineal: Precision del modelo de regrecion. El gráfico muestra un alto nivel de precisión en las predicciones del modelo. La mayoría de los puntos azules (predicciones) se alinean con la línea roja (ideal), lo que indica un ajuste casi perfecto.



Rendimiento del modelo:

- **RMSE (Raíz del Error Cuadrático Medio): 504.152**

  El RMSE mide la diferencia entre los valores predichos y los valores reales. Un RMSE de 504.152, dependiendo del rango de los valores de 'Passenger Count', puede considerarse bajo, lo que sugiere que las predicciones del modelo son generalmente precisas.

- **R^2 (Coeficiente de Determinación): 0.9999**

  Un valor de R^2 tan cercano a 1 indica que el modelo puede explicar casi toda la variabilidad de los datos de conteo de pasajeros. Esto demuestra una eficacia excepcional en las predicciones del modelo.

El modelo de regresión lineal muestra un rendimiento sobresaliente, con predicciones que coinciden muy de cerca con los datos reales. Esto se evidencia tanto visualmente en el gráfico como numéricamente en las métricas de rendimiento.