PROYECTO FINAL BIG DATA

Alba Bernal Rodríguez

En este documento se expondrán los diversos resultados obtenidos tras el análisis. Así mismo, comentaremos cada uno de los pasos con sus correspondientes conclusiones.

PASO 1: ANÁLISIS

En este primer paso llevé a cabo un análisis del Dataset para ello en primer lugar cargué el Dataset con ayuda de la librería Dask, ya que esta librería nos permite trabajar con grandes volúmenes de datos y además facilita el escalado de librerías como Numpy, Pandas y Scikit-learn. A continuación, consulté los registros del Dataset, los tipos de datos y si había nulos. Como había nulos los eliminé.

Ahora definimos las columnas que son innecesarias para también eliminarlas; las columnas innecesarias son aquellas que son meramente informativas.

Tras todo ello ya tenemos el Dataset limpio y pasamos a la resolución de las preguntas del proyecto.

PASO 2: PREGUNTAS

- Cargamos el conjunto de datos en el dataframe

  CÓDIGO:

```
df_1.to_csv('../data/air_traffic_limpio.csv', index=False)
```

- ¿Cuántas compañías diferentes aparecen en el fichero?

  CÓDIGO:

```
#Vemos cuantas compañías diferentes hay en el dataset
compania_unicas= df_1['Operating Airline'].unique().compute()
print('El número de compañías diferentes es: ', len(compania_unicas))
```

  RESULTADO:

```
El número de compañías diferentes es:  73
```

- ¿Cuántos pasajeros tienen de media los vuelos de cada compañía?

CÓDIGO:

```
media_pasajeros= df_1.groupby('Operating Airline')['Passenger Count'].mean().compute()
print('La media de pasajeros por compañía es: ', media_pasajeros)
```

RESULTADO:

```
La media de pasajeros por compañía es:  Operating Airline
ATA Airlines          8744.636364
Aer Lingus            4407.183673
Aeromexico            5463.822222
Air Berlin            2320.750000
Air Canada           18251.560109
                            ...
Virgin Atlantic       9847.104651
WestJet Airlines      5338.155340
World Airways          261.666667
XL Airways France     2223.161290
Xtra Airways            73.000000
```

- Eliminaremos los registros duplicados por el campo “GEO Región”, manteniendo únicamente aquel con mayor número de pasajeros.

CÓDIGO:

```
#Eliminamos los registros duplicados por el campo 'GEO Region' y mantenemos aquellos con mayor número de pasajeros
df_sin_duplicados = df_1.groupby('GEO Region').apply(lambda x: x.loc[x['Passenger Count'].idxmax()])

# Mostrar el nuevo DataFrame resultante
df_sin_duplicados.compute()
```

RESULTADO:

GEO Region	Activity Period	Operating Airline	GEO Region	Price Category Code	Terminal	Passenger Count	Year	Month
Asia	200708	United Airlines - Pre 07/01/2013	Asia	Other	International	86398	2007	August
Australia / Oceania	201501	Air New Zealand	Australia / Oceania	Other	International	12973	2015	January
Canada	200708	Air Canada	Canada	Other	Terminal 3	39798	2007	August
Central America	201410	TACA	Central America	Other	International	8970	2014	October
Europe	201507	United Airlines	Europe	Other	International	48136	2015	July
Mexico	201407	United Airlines	Mexico	Other	International	29206	2014	July
Middle East	201507	Emirates	Middle East	Other	International	14769	2015	July
South America	201101	LAN Peru	South America	Other	International	3685	2011	January
US	201308	United Airlines	US	Other	Terminal 3	659837	2013	August

- Volcaremos los resultados de los dos puntos anteriores a un CSV

  CÓDIGO:

```
#Volcaremos los datos anteriores a un CSV es decir sin duplicados y una columna donde pondremos la media de pasajeros por compañía
resultados_combinados=dd.merge(df_sin_duplicados, media_pasajeros, on='Operating Airline')
#Lo guardamos en un CSV
resultados_combinados.to_csv('../data/resultados_combinados.csv', single_file = True)
```

PASO 3

En este paso hice un análisis descriptivo de los datos. Para ello se necesita calcular la media y la desviación estándar de cada elemento del conjunto de datos.

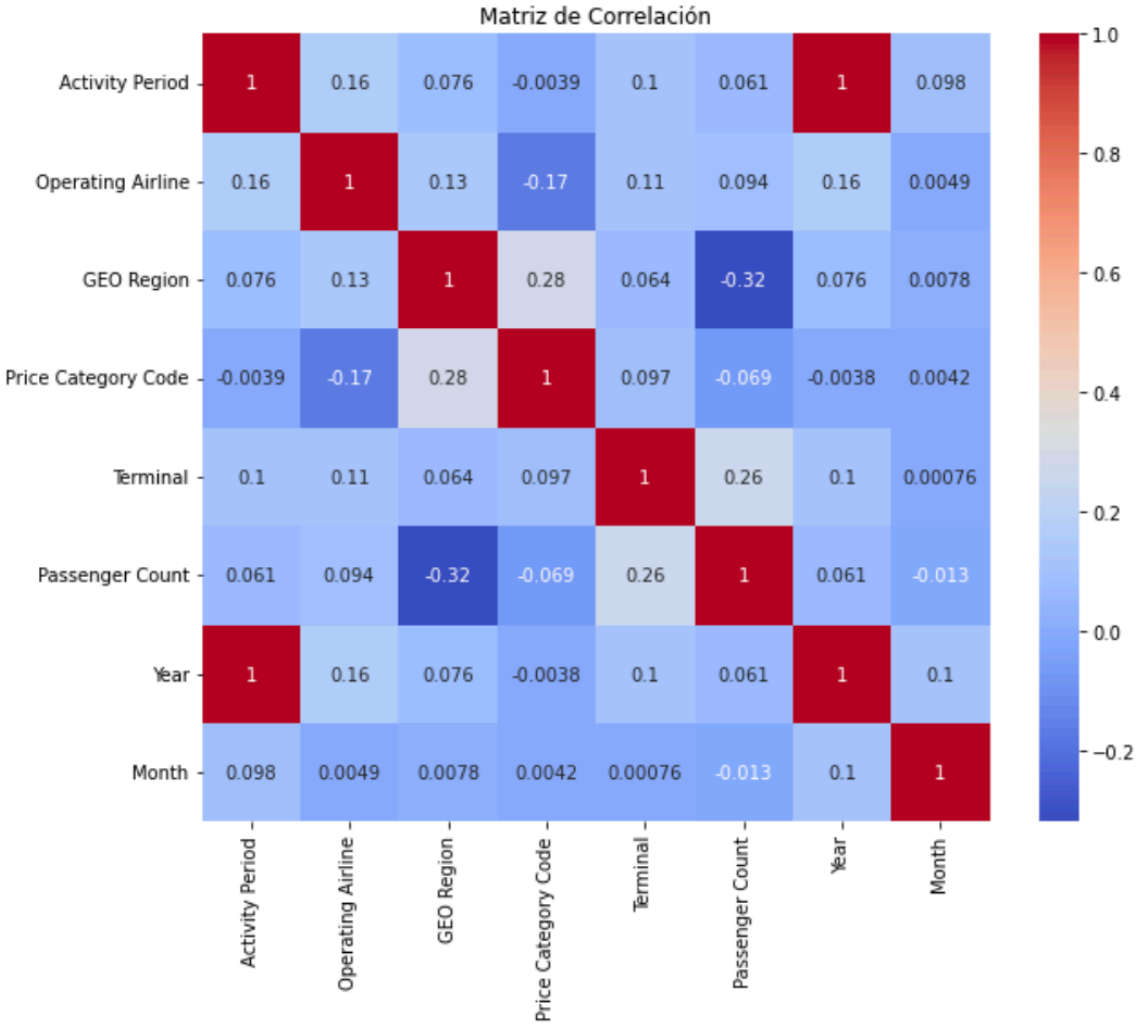
```
# Calcular la media de cada elemento del conjunto de datos
media = df.mean().compute()
# Calcular la desviación estándar de cada elemento del conjunto de datos
desviacion_estandar = df.std().compute()

# Mostrar los resultados
print("Media de cada elemento:")
print(media)
print("\nDesviación estándar de cada elemento:")
print(desviacion_estandar)
```

CONCLUSIÓN:

Como podemos observar los datos anteriores nos brindan información sobre la distribución y variabilidad de los datos. Por ejemplo, podemos ver que tras calcular la media de la columna ``Year`` podemos ver una cierta concentración alrededor del año 2010. Por otro lado, el cálculo de la desviación estándar de la columna ``Passenger_Count`` nos indica que la cantidad de pasajeros por vuelo puede tener una amplia variabilidad, es decir que puede ocurrir que haya vuelos con pocos pasajeros y vuelos con muchos pasajeros.

También analizamos la cantidad de datos del mismo tipo en la columna 'GEO Region' y de ahí observamos que la mayor cantidad de vuelos en US, Asia y Europa.

A continuación, pasamos a calcular la matriz de correlación que nos muestra el grado de relación lineal entre cada par de variables. Para poder construirla en primer lugar necesitamos que los datos sean numéricos, por ello lo convertimos haciendo uso de .categorize(). Una vez así ya podemos construirla.



<u>CONCLUSIONES:</u>

La matriz de correlación proporciona información sobre las relaciones lineales entre las variables en el dataset de vuelos. Aquí hay algunas conclusiones que se pueden extraer de la matriz:

-**Activity Period**: La variable de 'Activity Period' tiene una correlación positiva moderada con las variables de Operating Airline, GEO Region, Terminal, Passenger Count, y Year. Esto sugiere que estos factores tienden a variar en conjunto con el periodo de actividades los vuelos.

-**Operating Airline**: La variable de aerolínea operativa tiene una correlación débil positiva con la Activity period y una correlación débil negativa con Price Category. Esto indica que ciertas aerolíneas pueden tener una mayor presencia en ciertos períodos de actividad y pueden tener diferentes niveles de precios.

-**GEO Region**: La variable 'GEO Region' tiene una correlación débil positiva con Price Category y una correlación débil negativa con Passenger Counts. Esto sugiere que las regiones geográficas pueden influir ligeramente en los precios y en la demanda de pasajeros.

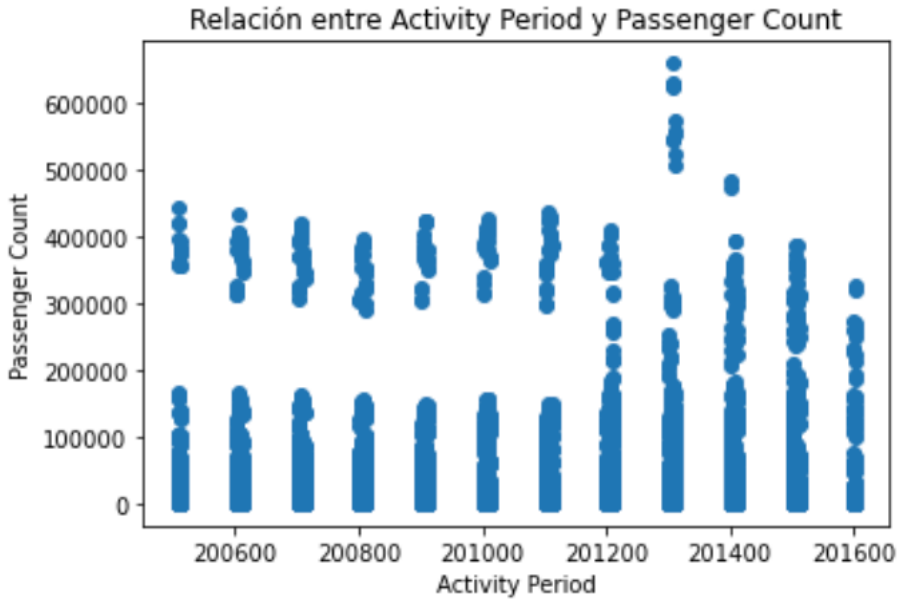
-**Price Category**: La variable de categoría de precio tiene una correlación débil positiva con GEO Region y una correlación débil negativa con Operating Airline. Esto indica que las categorías de precios pueden estar relacionadas con la ubicación geográfica y las aerolíneas específicas.

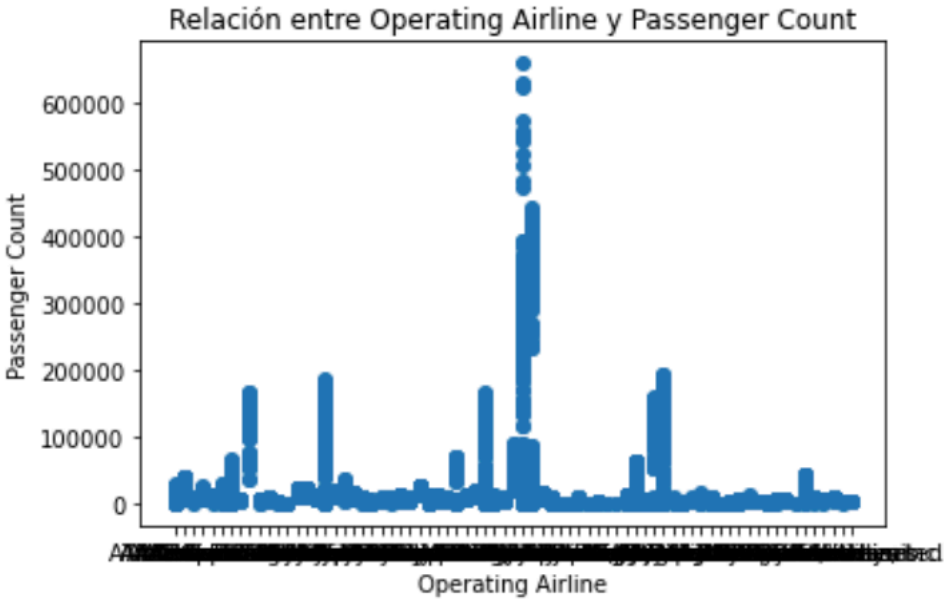
-**Terminal**: La variable de terminal tiene una correlación débil positiva con la actividad period y una correlación débil positiva con Operating Airline. Esto sugiere que los terminales pueden estar asociados con ciertos períodos de actividad y pueden acomodar diferentes volúmenes de pasajeros.

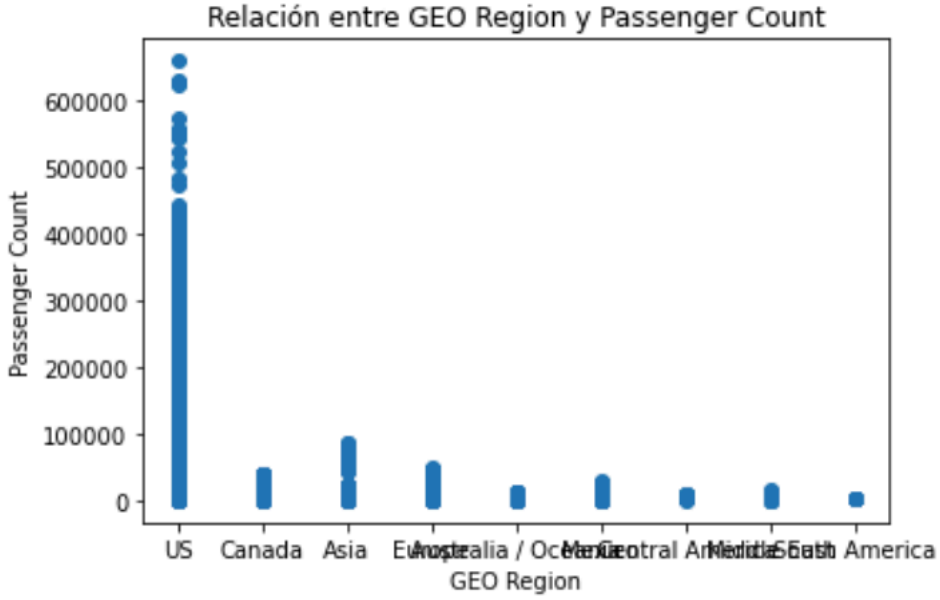
-**Passenger Count**: La variable de recuento de pasajeros tiene una correlación débil positiva con el Terminal y una correlación débil negativa con la GEO Region. Esto indica que el recuento de pasajeros puede estar influenciado por el terminal utilizado y la ubicación geográfica del vuelo.

En resumen, la matriz de correlación proporciona información sobre las relaciones entre las variables del dataset de vuelos. Sin embargo, es importante tener en cuenta que la correlación no implica causalidad, y es necesario realizar un análisis más detallado, que lo haremos a continuación.

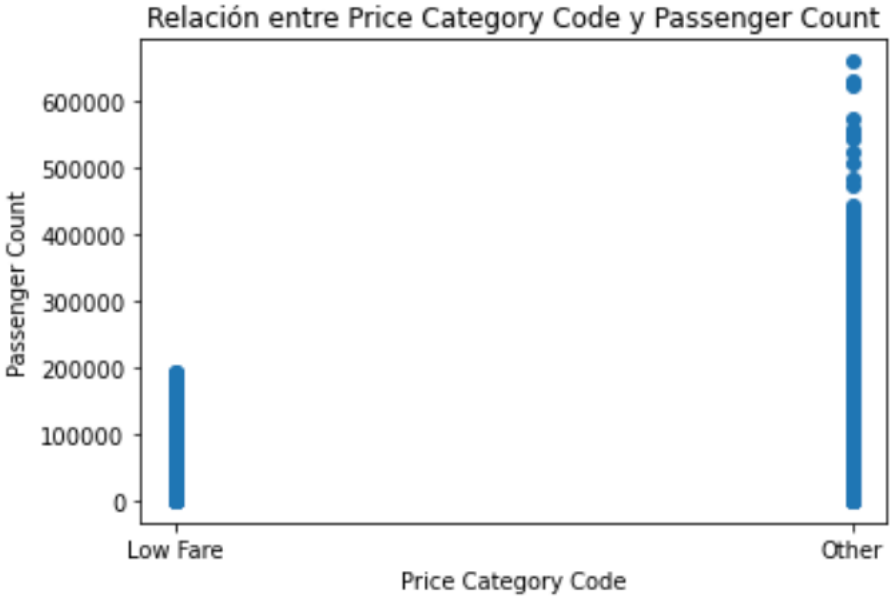
La variable que hemos decidido predecir es Passenger Count por lo que vamos a ver como se relacionan las distintas variables con esta. Tras el análisis obtenemos las siguientes gráficas.

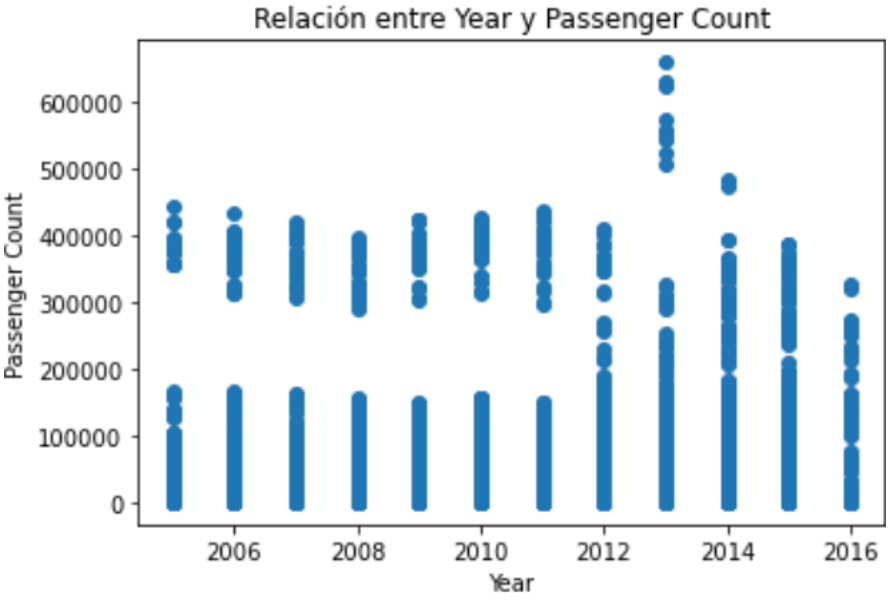


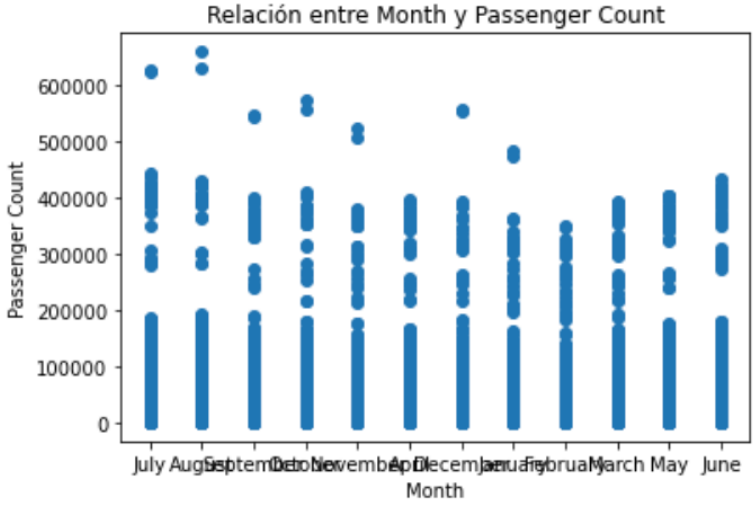

Relación entre Operating Airline y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
Operating Airline


Relación entre GEO Region y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
US
Canada
Asia
GEO Region


Relación entre Price Category Code y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
Low Fare
Other
Price Category Code


Relación entre Price Category Code y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
Low Fare
Other
Price Category Code


Relación entre Year y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
2006
2008
2010
2012
2014
2016
Year


Relación entre Month y Passenger Count
Passenger Count
600000
500000
400000
300000
200000
100000
0
July
March
May
June
Month

CONCLUSIÓN:

Tras analizar las gráficas anteriores vemos que no hay una relación lineal entre las variables y las variables objetivo, lo que quiere decir que no podemos usar un modelo de regresión lineal. Es por ello por lo que usaremos un modelo de regresión logística como el modelo **Decision Tree Regressor** o el modelo de clasificación como **Random Forest Regressor**.

PASO 4

Tras usar ambos modelos obtenemos diversos niveles de predicción.

Decision Tree Regressor: 0.5036912751677852
Random Forest Regressor: 0.8140939597315436

Como vemos con el modelo de clasificación obtenemos mejor porcentaje de predicción.