https://colab.research.google.com/drive/1RKZOXnmoweH_oPwv3MSDJP6cln-712kB

Enlace para acceder al Google Colab.

A continuación, vamos a analizar una base de datos contenida en un csv con diversas columnas y diferentes registros.

Se trata de una base de datos de diferentes compañías aéreas que nos brinda información sobre sus vuelos y sus políticas de vuelo, además del país del que proceden.

Para empezar a analizar la base de datos usaremos la librería pandas para pasar el csv a una tabla con la que trabajaremos más fácilmente.

	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	Price Category Code	Terminal	Boarding Area
0	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Deplaned	Low Fare	Terminal 1	B
1	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Enplaned	Low Fare	Terminal 1	B
2	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Thru / Transit	Low Fare	Terminal 1	B
3	200507	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	Other	Terminal 1	B
4	200507	Air Canada	AC	Air Canada	AC	International	Canada	Enplaned	Other	Terminal 1	B

A continuación, sacaremos algunos datos que nos puedan interesar, como la cantidad de compañías distintas que figuran en la tabla y cuantas veces aparece cada una (1), además de cuantas pertenecen a cada país (2).

1.

```
United Airlines - Pre 07/01/2013    2154
SkyWest Airlines                     963
United Airlines                      892
Alaska Airlines                      751
Delta Air Lines                      386
                                     ...
Evergreen International Airlines       2
Atlas Air, Inc                         2
Xtra Airways                           2
Pacific Aviation                       2
Boeing Company                         1
Name: Operating Airline, Length: 77, dtype: int64
```

2.

```
GEO Region
Asia                  20
Australia / Oceania    4
Canada                10
Central America        3
Europe                19
Mexico                 7
Middle East            2
South America          1
US                    36
Name: Operating Airline, dtype: int64
```

Vemos en que año comenzó cada compañía y a que país pertenece:

Operating Airline	GEO Region	Year
ATA Airlines	Canada	2005
Frontier Airlines	US	2005
Hawaiian Airlines	US	2005
Horizon Air	US	2005
Icelandair	Europe	2005
...	...	...
Swissport USA	Asia	2014
Air India Limited	Asia	2015
COPA Airlines, Inc.	Central America	2015
Sun Country Airlines	Mexico	2015
Turkish Airlines	Europe	2015

Una vez visto esto pasaremos a limpiar un poco nuestra tabla para eliminar los datos redundantes y los que aportan información irrelevante para crear una nueva tabla que contenga solo la información relevante llama df_aire2:

	Operating Airline	GEO Summary	GEO Region	Price Category Code	Adjusted Activity Type Code	Adjusted Passenger Count	Year	Month
0	ATA Airlines	Domestic	US	Low Fare	Deplaned	27271	2005	July
1	ATA Airlines	Domestic	US	Low Fare	Enplaned	29131	2005	July
2	ATA Airlines	Domestic	US	Low Fare	Thru / Transit * 2	10830	2005	July
3	Air Canada	International	Canada	Other	Deplaned	35156	2005	July
4	Air Canada	International	Canada	Other	Enplaned	34090	2005	July
...	...	...	...	...	...	...	...	...
15002	Virgin America	Domestic	US	Low Fare	Enplaned	194636	2016	March
15003	Virgin America	International	Mexico	Low Fare	Deplaned	4189	2016	March
15004	Virgin America	International	Mexico	Low Fare	Enplaned	4693	2016	March
15005	Virgin Atlantic	International	Europe	Other	Deplaned	12313	2016	March
15006	Virgin Atlantic	International	Europe	Other	Enplaned	10696	2016	March

Donde podemos ver que cuenta con muchos menos campos, pero mantiene toda la información relevante.

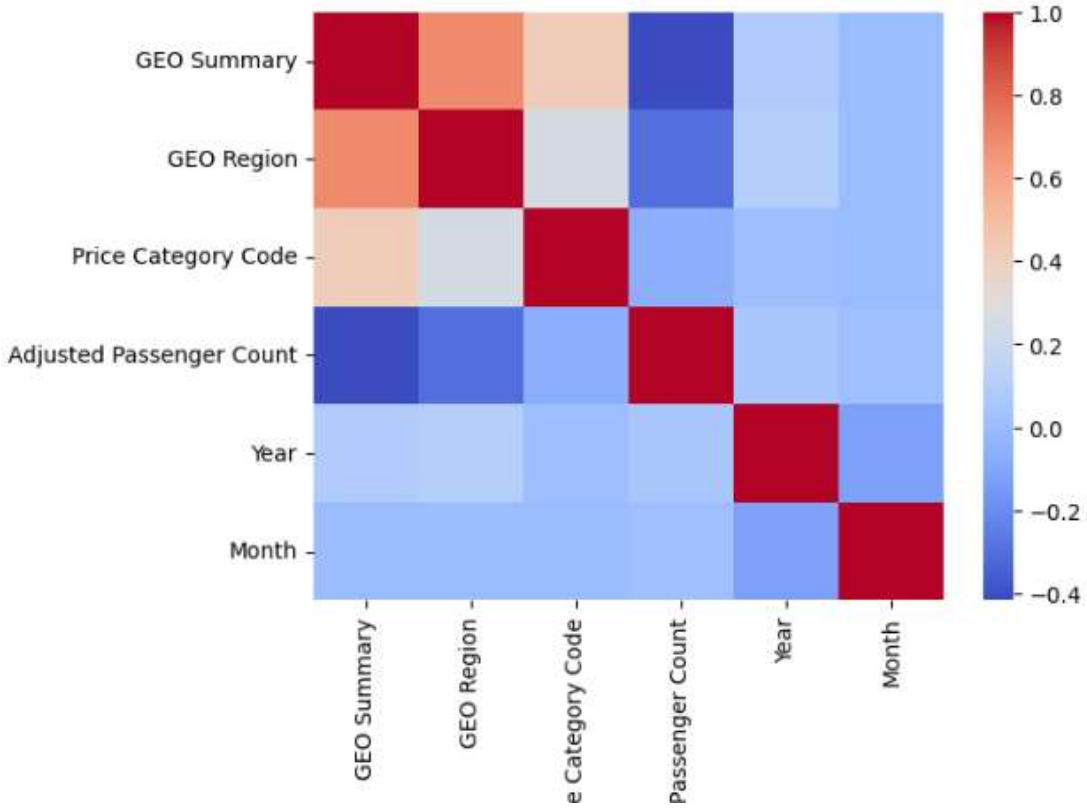
También quitamos las filas cuyo Type Code sea Enplaned, ya que es información redundante y con menos credibilidad que Deplaned, lo cual nos reduce la tabla a la mitad.

Para poder analizar nuestra tabla y sacar gráficas, además de modelos de predicción deberemos pasar nuestras variables categóricas a variables numéricas. Para ello asignamos un valor numérico para cada variable categórica distinta y quedaría así:

	Operating Airline	GEO Summary	GEO Region	Price Category Code	Adjusted Passenger Count	Year	Month
0	ATA Airlines	0	0	0	27271	2005	7
3	Air Canada	1	3	1	35156	2005	7
5	Air China	1	1	1	6263	2005	7
7	Air France	1	5	1	12050	2005	7
9	Air New Zealand	1	2	1	4998	2005	7
...	...	...	...	...	...	...	...
14995	United Airlines	1	5	1	20876	2016	3
14997	United Airlines	1	6	1	25660	2016	3
15001	Virgin America	0	0	0	186464	2016	3
15003	Virgin America	1	6	0	4189	2016	3
15005	Virgin Atlantic	1	5	1	12313	2016	3

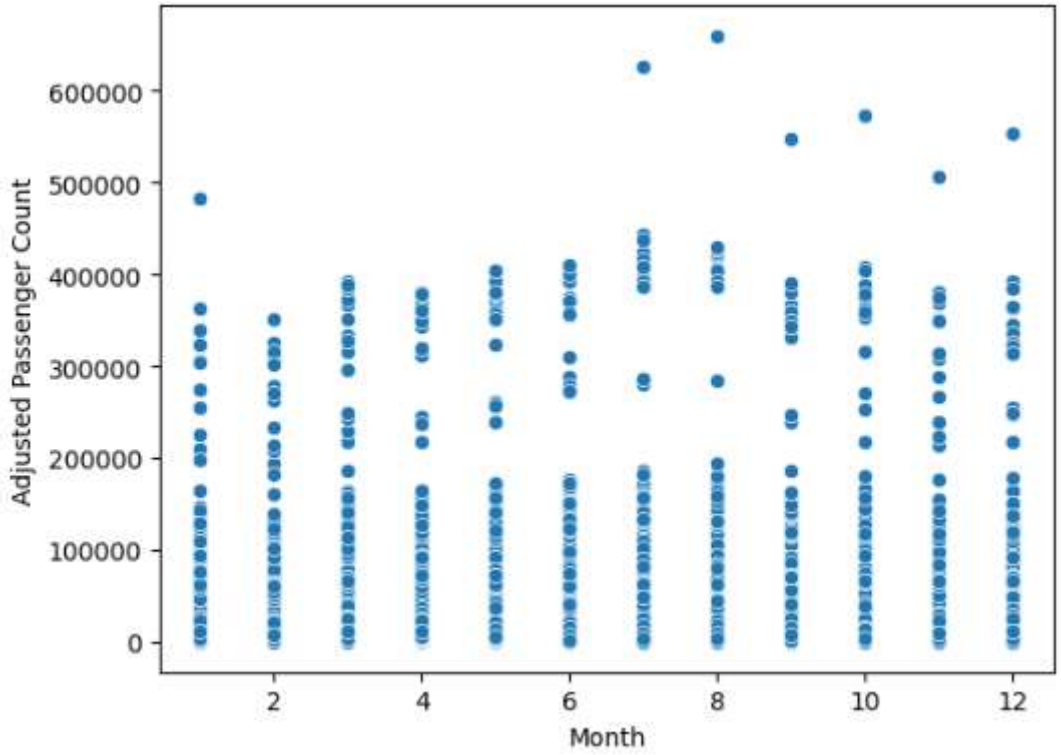
De momento no hemos pasado a numérica la columna Operating Airline pero lo haremos más adelante.

A continuación, sacamos una matriz de correlación para ver como influyen unas variables con otras:

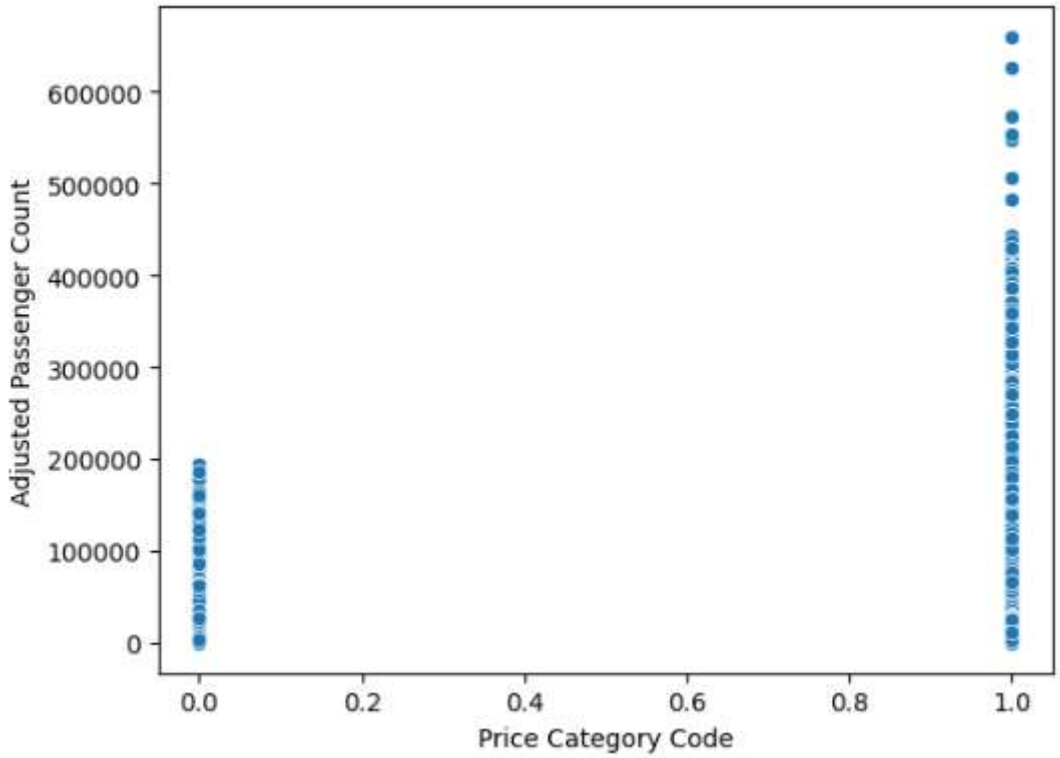


De la cual podemos ver que no hay apenas relación de las variables entre sí, lo que nos indica que nuestro modelo de predicción va a ser bastante malo, ya que la correlación entre las variables es baja. Donde más correlación hay entre Geo Summary y Geo Region, lo cual indica únicamente que en ciertos países predominan mayoritariamente o vuelos internacionales o vuelos domésticos.

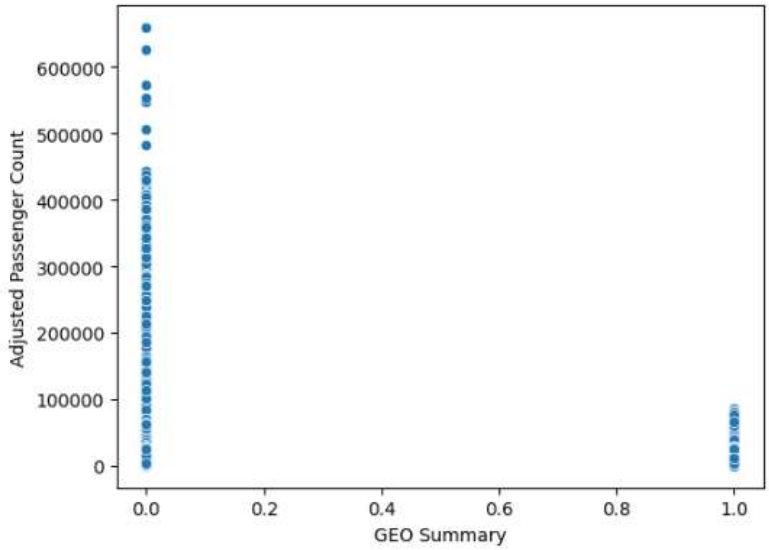
A continuación, iremos sacando graficas que relacionen unas variables con otras para ver que podemos deducir de estas.



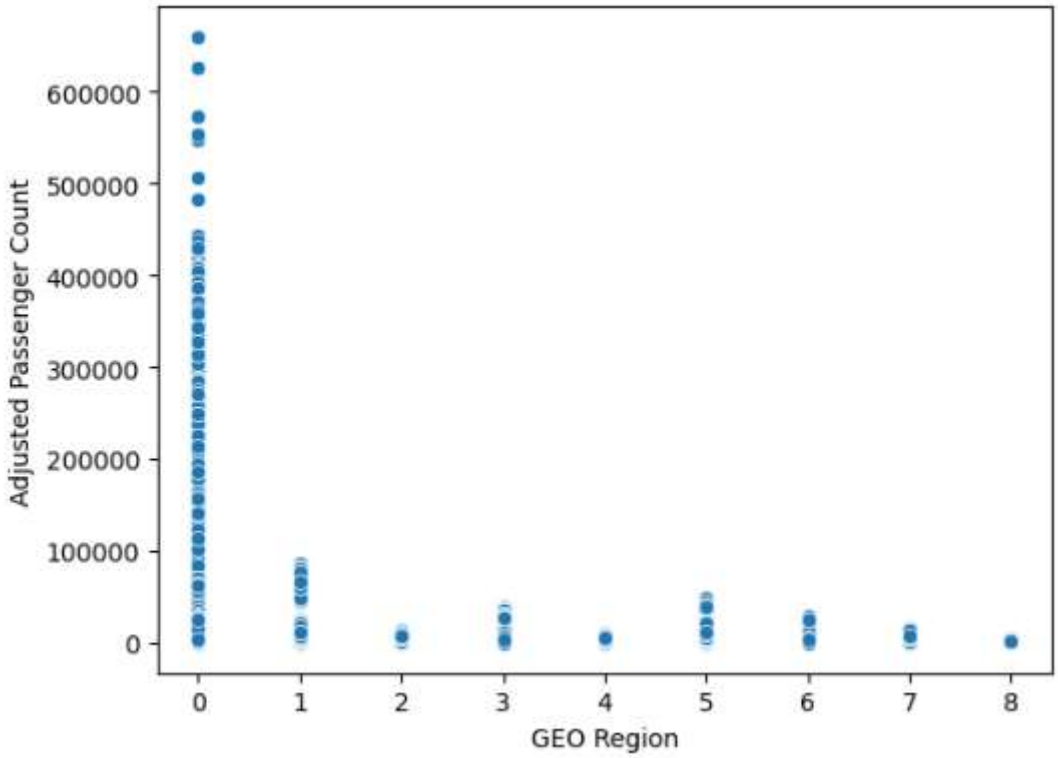
podemos ver que, aunque la gráfica sea muy regular, hay un ligero aumento de la gente que viaja en los meses de verano y que en invierno disminuye como era de esperar, seguramente por el clima y porque son periodos de vacaciones. En octubre también hay una ligera subida, y otra en diciembre seguramente debida a las vacaciones de navidad.



De esta otra grafica podemos deducir que la gente no coge tantos vuelos low price como esperaríamos o que no hay tantos como vuelos por un precio estándar o caro. Tampoco podemos afirmar mucho ya que solo hay dos categorías que son low price y otros, por lo que otros engloban desde vuelos muy caros a vuelos por un precio estándar, pero aun así impacta ver que el precio no es tan relevante a la hora de calcular cuantos pasajeros va a tener un vuelo.

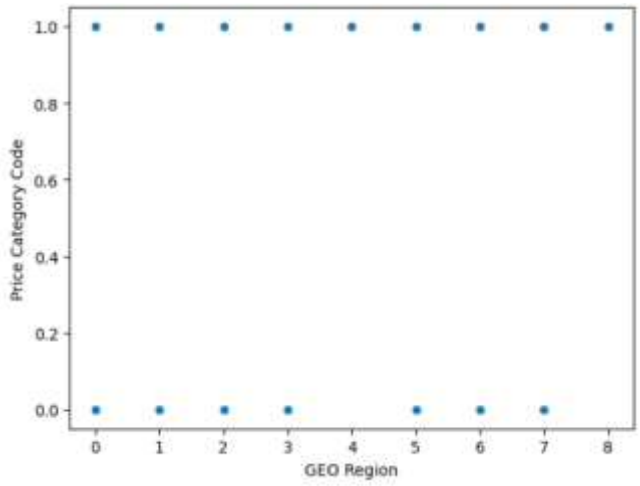


De esta otra grafica sí que podemos afirmar que son mucho más comunes los vuelos dentro de un mismo país que los internacionales, seguramente por el precio y porque viajar a otro país implica además tener tiempo de vacaciones para poder amortizar el viaje y pasar suficiente tiempo en el otro país.

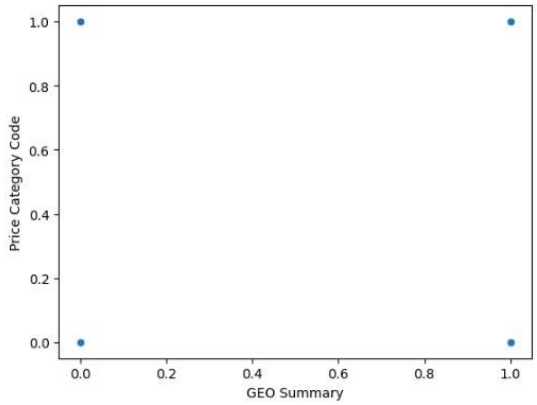


En orden tenemos 0 = US,1 = 'Asia',2= "Australia / Oceania",3= "Canada",4="Central America",5="Europe",6="Mexico",7="Middle East",8="South América".

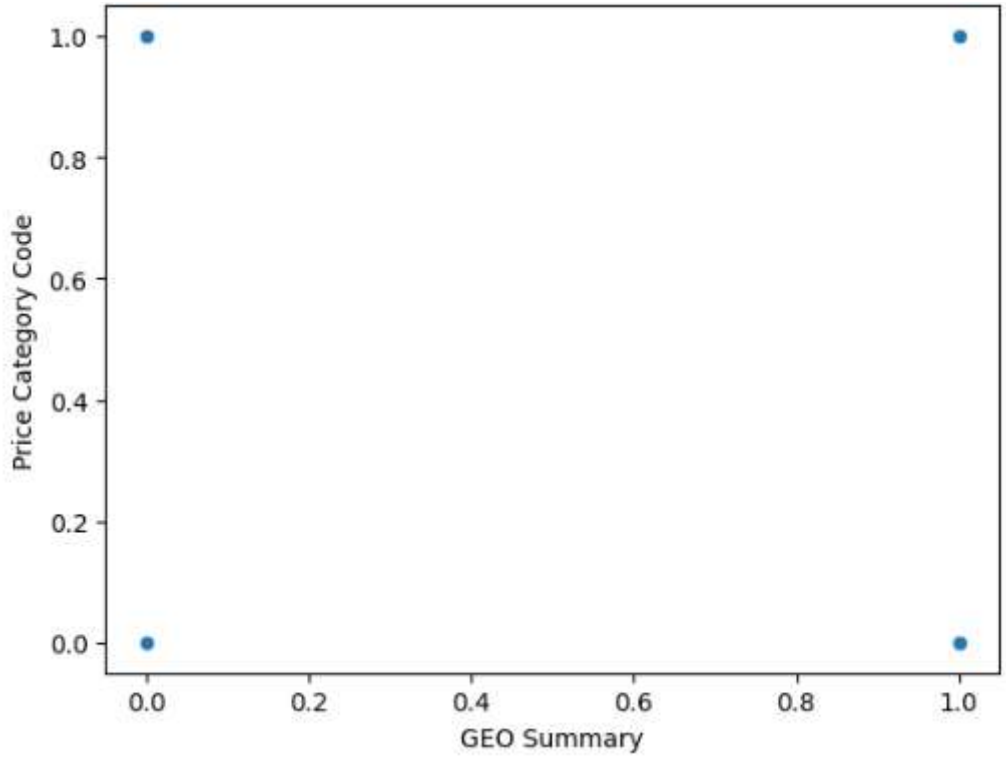
Podemos ver que la gente de estados unidos viaja mucho más que la gente del resto del mundo en avión, en Sudamérica, Australia, centro América es muy poco frecuente, seguramente por su nivel de vida, ya que son países por lo general con menos riqueza. Seguramente también se deba a que cuenta con más compañías que el resto de los países y estas compañías tienen más aviones por lo que figuran más en la tabla.



De esta grafica únicamente podemos observar que en centro América y Sudamérica no hay vuelos baratos, por eso en parte es muy probable que esos países tengan menos viajeros que el resto.

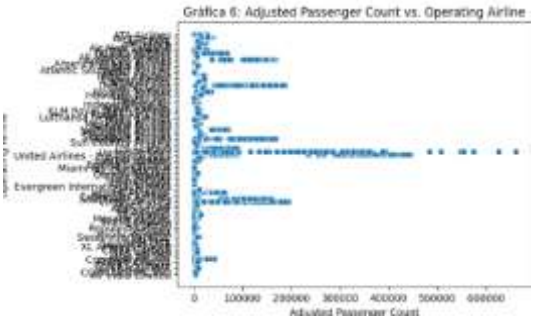


No podemos ver ninguna relación en que los vuelos internacionales o "domésticos" sean más o menos caros debido a esto.



Aquí podemos ver que en 2013 hay un pico en cuanto a que es el año en el que más personas han viajado en avión, aunque no sabemos porque, quizás se deba a que aparentemente se crearon más aviones para llevar cantidades más variadas de personas, ya que de 2005 a 2011 aparentemente los aviones dan un salto grande en cuanto a la cantidad de pasajeros que viajaban, seguramente porque hubiera menos modelos de avión y estos o eran masivos, con mucha gente, o eran pequeños/normales, con una capacidad de gente más baja.

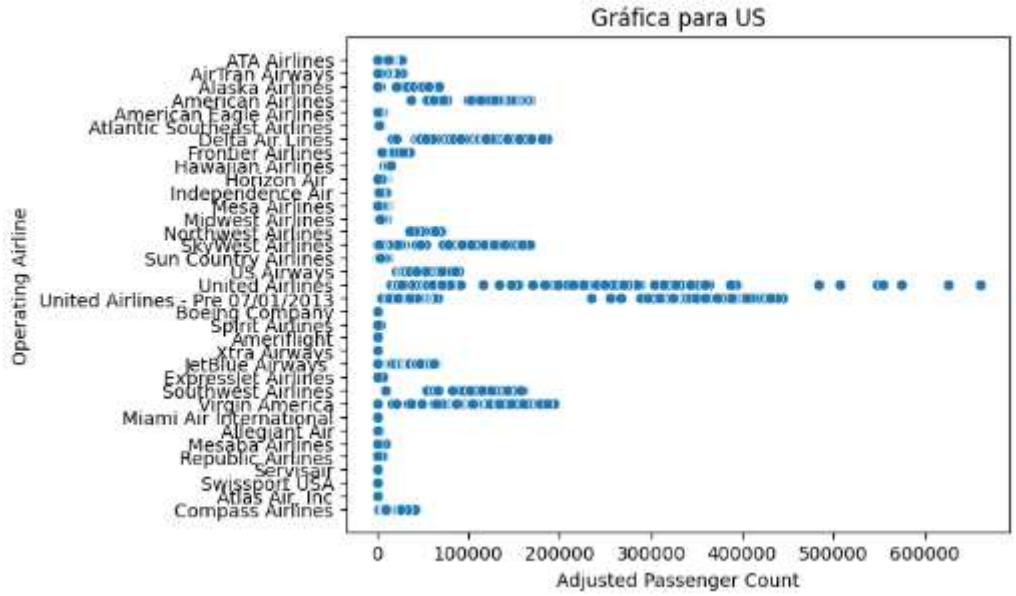
A continuación, sacamos las graficas que comparan la compañía con el resto de las variables para ver cuales nos pueden dar información y cuales no:

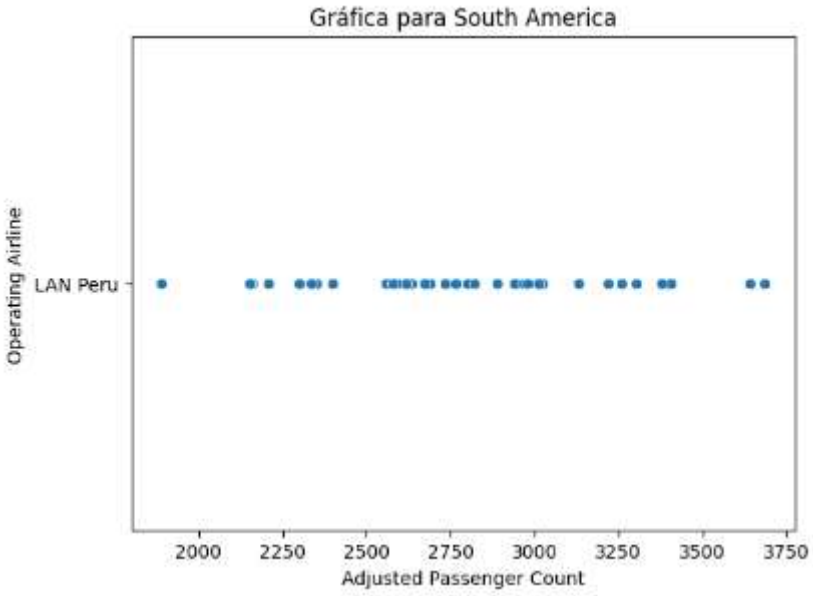


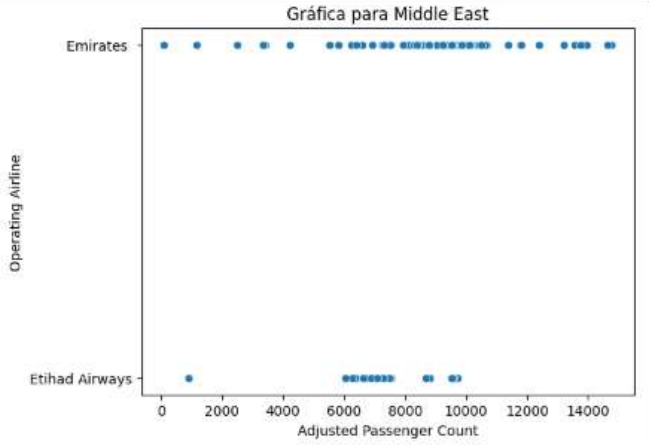
Estas parecen ser las interesantes, ya que el resto o son irrelevantes o son información ya analizada previamente.

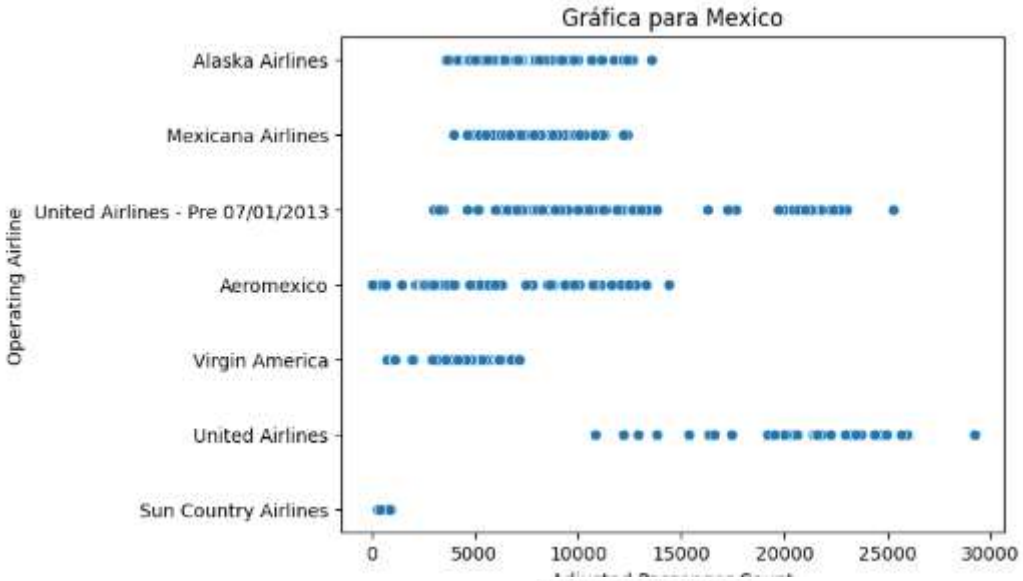
De analizar las otras lo único que podemos deducir es que hay compañías que tienen una política de vuelos doméstico, internacional o ambas.

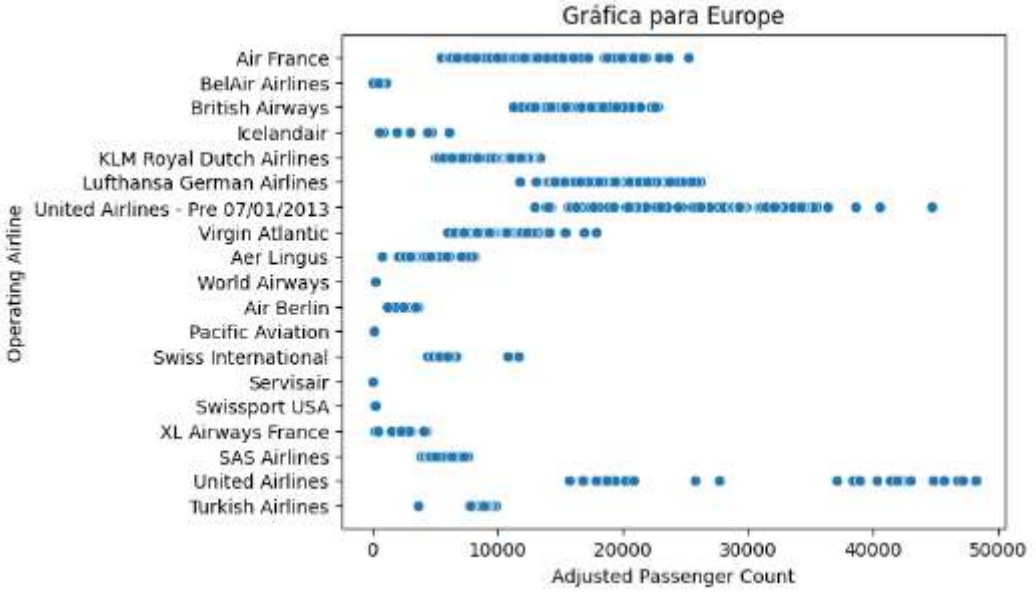
Bien, analicemos ahora las gráficas que comparan compañía con cantidad de personas, y hagámoslo por países ya que seguramente influya como vimos antes, e intentemos sacar conclusiones.

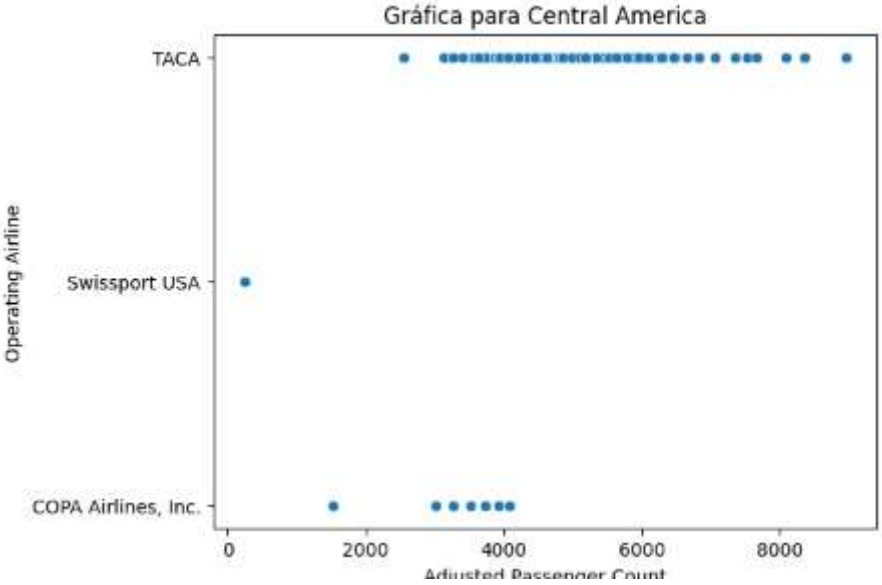


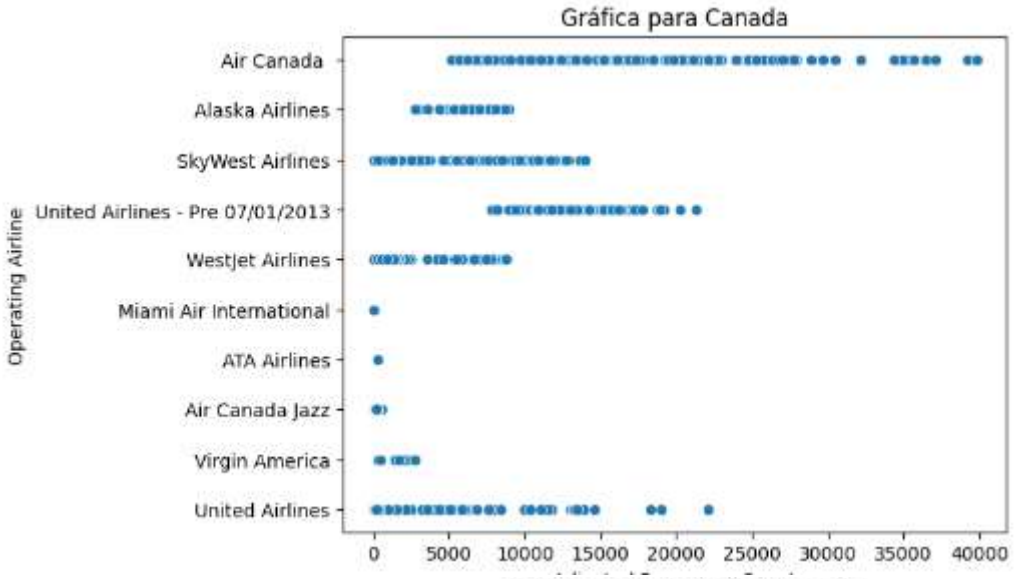

Gráfica para South America
Operating Airline
LAN Peru
2000
2250
2500
2750
3000
3250
3500
3750
Adjusted Passenger Count

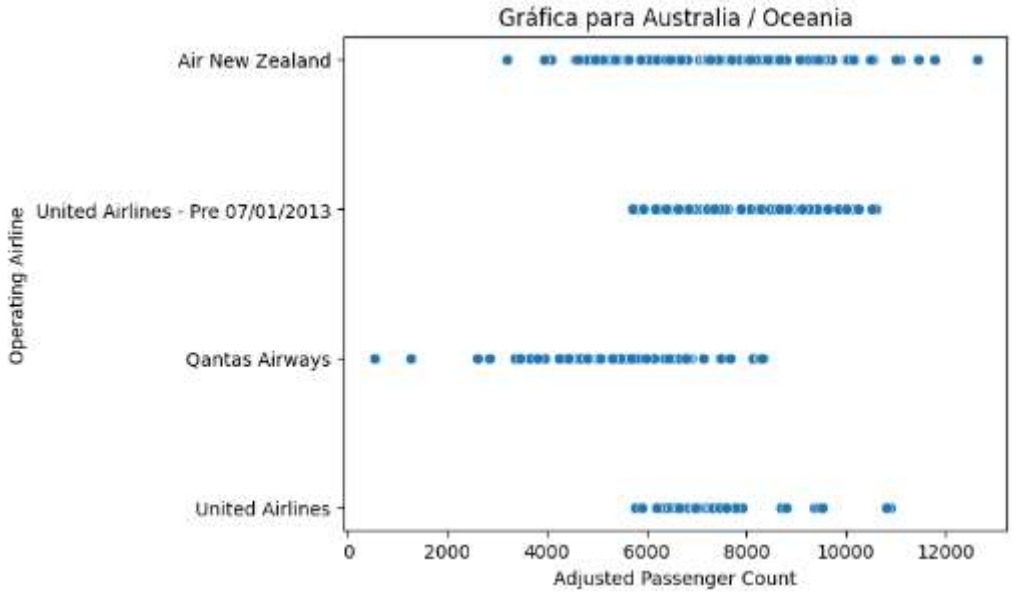

Gráfica para Middle East
Emirates
Operating Airline
Etihad Airways
0
2000
4000
6000
8000
10000
12000
14000
Adjusted Passenger Count

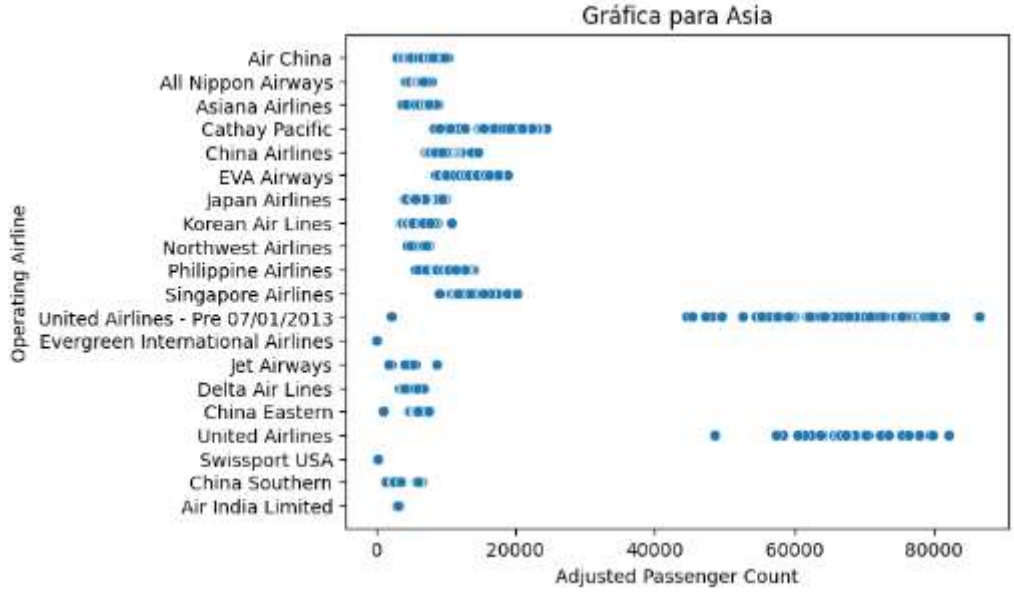

Gráfica para Mexico
Alaska Airlines
Mexicana Airlines
United Airlines - Pre 07/01/2013
Aeromexico
Virgin America
United Airlines
Sun Country Airlines
Operating Airline
0
5000
10000
15000
20000
25000
30000


Gráfica para Europe
Air France
BelAir Airlines
British Airways
Icelandair
KLM Royal Dutch Airlines
Lufthansa German Airlines
United Airlines - Pre 07/01/2013
Virgin Atlantic
Aer Lingus
World Airways
Air Berlin
Pacific Aviation
Swiss International
Servisair
Swissport USA
XL Airways France
SAS Airlines
United Airlines
Turkish Airlines
Operating Airline
0
10000
20000
30000
40000
50000
Adjusted Passenger Count


Gráfica para Central America
TACA
Swissport USA
COPA Airlines, Inc.
Operating Airline
0
2000
4000
6000
8000


Gráfica para Canada
Operating Airline
Air Canada
Alaska Airlines
SkyWest Airlines
United Airlines - Pre 07/01/2013
WestJet Airlines
Miami Air International
ATA Airlines
Air Canada Jazz
Virgin America
United Airlines
0
5000
10000
15000
20000
25000
30000
35000
40000


Gráfica para Australia / Oceania
Operating Airline
Air New Zealand
United Airlines - Pre 07/01/2013
Qantas Airways
United Airlines
0
2000
4000
6000
8000
10000
12000
Adjusted Passenger Count


Gráfica para Asia
Operating Airline
Air China
All Nippon Airways
Asiana Airlines
Cathay Pacific
China Airlines
EVA Airways
Japan Airlines
Korean Air Lines
Northwest Airlines
Philippine Airlines
Singapore Airlines
United Airlines - Pre 07/01/2013
Evergreen International Airlines
Jet Airways
Delta Air Lines
China Eastern
United Airlines
Swissport USA
China Southern
Air India Limited
0
20000
40000
60000
80000
Adjusted Passenger Count

Podemos ver que en Asia y en estados unidos son los países en los que más personas viajan seguidos de Europa y Canadá, en el resto de países viajan bastantes menos personas. Seguramente se deba a que existe mucha más población en esos países y que tienen mejores condiciones de vida. Podemos ver también que las compañías que ofrecen vuelos internacionales y domésticos y las que ofrecen vuelos baratos y caros, es decir las compañías con más variedad de servicios son las que más viajeros llevan a bordo, seguramente debido a que acogen a un público muy amplio y no a un tipo más concreto de persona. También podemos ver que estas compañías son brutalmente más grandes (a nivel económico y de recursos) que el resto de las compañías y por lo tanto tienen más aviones y más gente volando.

A continuación, entrenaremos con sklearn nuestro dataset para ver si con una serie de datos de entrada es capaz de predecir el número de pasajeros.

Ahora si transformamos en números la columna empresas, ya que en teoría es una de las más importantes para predecir los pasajeros de un vuelo.

Normalizamos los datos para que no pete cuando haga el entrenamiento con los datos y probamos diferentes modelos.

Probando anteriormente hemos podido comprobar que eliminando de x test la variable mes y la variable geo summary aumenta la precisión, lo que quiere decir que esas variables solo confunden a nuestro aprendizaje supervisado

Modelo de regresión logística:

```
Precisión del modelo: 0.0
Coeficiente de determinación R^2: -0.16724425864535952
6437
```

6437 es la predicción del primer valor que como vemos está lejos del valor real que es 27000.

Vemos que el modelo logístico no es nada útil ya que tiene un R2 muy lejano a 1 o -1 y que la precisión es de 0. Podemos ver que la solución que nos ha dado también dista bastante de la real

Modelo RandomForest:

```
Precisión del modelo: 0.0014134275618374558
```

Como vemos es inútil, aunque un mínimo mejor que el anterior

Modelo KNeighborsClassifier:

```
Precisión del modelo: 0.0014134275618374558
```

Tiene la misma precisión que el modelo anterior por lo que también es irrelevante.

Modelo PLSRegression:

```
Coeficiente de determinación (R²): 0.16882290621999185
```

vemos que con este modelo tenemos un R2 un poco mejor pero aun así no es suficiente como para decir que las variables mes, región, sumario y compañía sean relevantes a la hora de calcular las personas que va a tener el vuelo.

después de probar varios modelos más vemos que todos los resultados son bastante malos, por lo que vamos a probar a eliminar los registros cuya desviación estándar sea bastante alta, de manera que no afecten a nuestro modelo de predicción.

Pese a ello probando otra vez un par de modelos mas:

Modelo KNeighborsClassifier:

```
Precisión del modelo: 0.0014134275618374558
```

Modelo RandomForest:

```
Precisión del modelo: 0.0014134275618374558
```

Vemos que no varia pese a ello por lo que podemos concluir entonces con que no existe mucha relación aparente entre los valores de la columna adjusted passenger count y el resto de columnas, es decir los valores de esta columna no dependen del resto de columnas y por lo cual es muy difícil predecir cuánta gente va a coger un vuelo pese a saber el resto de datos, aunque esto es un poco contradictorio con las primeras graficas que obtuvimos que nos mostraban como algunos países y algunas compañías tenían más viajeros que el resto, aunque esto, visto lo visto, seguramente se debe a que de estos países salen más vuelos y las compañías tienen más vuelos, pero no por ello llevan más gente en cada vuelo, si no que al tener más vuelos pues llevan más gente, aunque el promedio de gente por vuelo sea relativamente parecido al resto de compañías y países.
