Análisis trabajo final

En este trabajo hemos realizado las tareas de un analista de Big Data. La empresa Tokio School Viajes quiere estudiar el tráfico que hay en el aeropuerto de San Francisco, para ello nos facilita un conjunto de datos sobre los vuelos que se realizan en esta instalación. Estos datos están disponibles en el archivo air_traffic_data.csv

Los datos que disponemos para hacer el análisis son:

VARIABLE	TIPO
Activity period	INT
Operating Airline	STR
Operating Airline IATA Code	STR
Published Airline IATA Code	STR
GEO Summary	STR
GEO Region	STR
Activity Type Period	STR
Price Category Code	STR
Terminal	STR
Boarding Area	STR
Passenger Count	INT
Adjusted Activity Type Code	STR
Adjusted Passenger Code	STR
Year	INT
Month	STR
Operating Airline	STR

Cargar el conjunto de datos en un dataframe:

```
In [2]: df=dd.read_csv(os.path.join('air_traffic_data.csv')).compute()
        df.head()
```

Out[2]:

	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	Passenger Count	Adju Act
0	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Deplaned	Low Fare	Terminal 1	B	27271	Depl
1	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Enplaned	Low Fare	Terminal 1	B	29131	Enpl
2	200507	ATA Airlines	TZ	ATA Airlines	TZ	Domestic	US	Thru / Transit	Low Fare	Terminal 1	B	5415	T Tra
3	200507	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	Other	Terminal 1	B	35156	Depl
4	200507	Air Canada	AC	Air Canada	AC	International	Canada	Enplaned	Other	Terminal 1	B	34090	Enpl

¿Cómo sabemos de qué tipo es cada variable sin haber visto todos los datos? Utilizamos Pandas para obtener la información de todo el dataset (la base de datos que disponemos) y hemos obtenido el siguiente resultado:

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15007 entries, 0 to 15006
Data columns (total 16 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Activity Period               15007 non-null  int64
 1   Operating Airline             15007 non-null  object
 2   Operating Airline IATA Code   14953 non-null  object
 3   Published Airline             15007 non-null  object
 4   Published Airline IATA Code   14953 non-null  object
 5   GEO Summary                   15007 non-null  object
 6   GEO Region                    15007 non-null  object
 7   Activity Type Code            15007 non-null  object
 8   Price Category Code           15007 non-null  object
 9   Terminal                      15007 non-null  object
 10  Boarding Area                 15007 non-null  object
 11  Passenger Count               15007 non-null  int64
 12  Adjusted Activity Type Code   15007 non-null  object
 13  Adjusted Passenger Count      15007 non-null  int64
 14  Year                          15007 non-null  int64
 15  Month                         15007 non-null  object
dtypes: int64(4), object(12)
memory usage: 1.8+ MB
```

¿Cuántas compañías diferentes utilizan el aeropuerto de San Francisco según nuestros datos?

```
In [4]: a=df.groupby('Operating Airline IATA Code').count()
        company=list(a.index)
        print("hay", len(company), "compañias aereas")
```

```
hay 73 compañias aereas
```

¿Cuántos pasajeros tiene de media cada una de las compañías de las que tenemos datos?

```
In [7]: b=df.groupby('Operating Airline IATA Code')['Passenger Count'].mean()
        print(b)
```

```
Operating Airline IATA Code
4T       312.625000
5Y        34.000000
9W      4280.312500
A8         5.000000
AA    127164.389706
           ...
XE      5631.843750
XJ      2864.727273
XP        73.000000
YV      3710.581197
YX      3883.000000
Name: Passenger Count, Length: 73, dtype: float64
```

Ahora elimino los registros que están repetido en el campo de “GEO Region” dejando aquellos que cuentan con mayor números de pasajeros:

In [8]:
```
df1=df.copy()
c= df.groupby('GEO Region')['Passenger Count'].max()
l=list(c)
l1=list(c.index)

borrar=[]

for i in range(len(df['Passenger Count'])):
    g=l1.index(df['GEO Region'][i])
    if df['Passenger Count'][i] != l[g]:
        borrar.append(i)

df=df.drop(borrar)
```

In [9]:
```
df.reset_index(inplace=True, drop=True)
df.head()
```

Out[9]:

	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	Passenger Count
0	200708	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	Other	Terminal 3	E	39798
1	200708	United Airlines - Pre 07/01/2013	UA	United Airlines - Pre 07/01/2013	UA	International	Asia	Deplaned	Other	International	G	86398
2	201101	LAN Peru	LP	LAN Peru	LP	International	South America	Deplaned	Other	International	A	3685
3	201308	United Airlines	UA	United Airlines	UA	Domestic	US	Deplaned	Other	Terminal 3	F	659837
4	201407	United Airlines	UA	United Airlines	UA	International	Mexico	Deplaned	Other	International	G	29206

Nota: hay mas columnas de las que se aprecia en las imágenes, se pueden consultar todas las columnas en los archivos.

Volcamos los resultados en dos csv nuevos llamados máximos.csv y medias.csv

Una vez realizadas estas nociones básicas vamos a analizar los datos y las relaciones que hay entre ellos más en profundidad.

Para ello necesitamos pasar algunas variables a enteros para poder utilizarlas de manera correcta, he convertido la columna “Month” en números enteros de la forma (Enero=1, Febrero=2, … , Diciembre=12)

además, la columna “Price Category Code” la he pasado también a números enteros.

A continuación, se muestra los resultados de las medias de las variables Boarding Area, GEO Region y Operating Airline:

```
Activity Period             201045.073366
Passenger Count              29240.521090
Adjusted Passenger Count     29331.917105
Year                          2010.385220
dtype: float64
                  Activity Period  Passenger Count  Adjusted Passenger Count  \
Operating Airline
ATA Airlines         200586.363636      8744.636364               9661.659091
Aer Lingus           201151.469388      4407.183673               4407.183673
Aeromexico           201207.533333      5463.822222               5463.822222
Air Berlin           201107.500000      2320.750000               2320.750000
Air Canada           201123.497268     18251.560109              18251.560109
...                            ...              ...                       ...
Virgin Atlantic      201043.744186      9847.104651               9847.104651
WestJet Airlines     201125.844660      5338.155340               5338.155340
World Airways        201008.333333       261.666667                261.666667
XL Airways France    201339.096774      2223.161290               2240.129032
Xtra Airways         200608.000000        73.000000                 73.000000

                            Year
Operating Airline
ATA Airlines        2005.795455
Aer Lingus          2011.448980
Aeromexico          2012.011111
Air Berlin          2011.000000
Air Canada          2011.169399
...                         ...
Virgin Atlantic     2010.372093
WestJet Airlines    2011.184466
World Airways       2010.000000
XL Airways France   2013.322581
Xtra Airways        2006.000000

[77 rows x 4 columns]
```

```
Activity Period             201045.073366
Passenger Count              29240.521090
Adjusted Passenger Count     29331.917105
Year                          2010.385220
dtype: float64
```

GEO Region	Activity Period	Passenger Count	Adjusted Passenger Count	Year
Asia	201046.193706	13435.004583	13508.552704	2010.396578
Australia / Oceania	200993.457259	6417.016282	6495.104478	2009.869742
Canada	201070.151622	9777.968265	9803.791255	2010.635402
Central America	201072.277372	4946.715328	4946.715328	2010.656934
Europe	201073.937769	12755.652465	12779.055050	2010.673528
Mexico	201065.279821	7173.620628	7250.898655	2010.588341
Middle East	201262.476636	8658.612150	8658.612150	2012.560748
South America	201193.266667	2786.011111	2786.011111	2011.866667
US	201018.968432	58330.343454	58485.878385	2010.124030

Boarding Area	Activity Period	Passenger Count	Adjusted Passenger Count	Year
A	201074.682488	11115.767656	11140.662392	2010.681148
B	200976.570998	33804.871049	33885.257903	2009.700452
C	200992.599349	34423.159609	34444.986156	2009.860749
D	201346.314815	105124.197531	105124.197531	2013.398148
E	200917.158145	48617.014269	48653.051130	2009.105826
F	201035.737110	100600.343500	101086.082789	2010.291939
G	201064.539329	14432.325651	14521.331162	2010.579910
Other	200725.740741	7.407407	7.814815	2007.185185

También he obtenido las desviaciones típicas (de las mismas variables) que serán fundamentales en el cálculo de una regresión que se pueda ajustar a los datos.

```
Activity Period              313.336196
Passenger Count            58319.509284
Adjusted Passenger Count   58284.182219
Year                           3.137589
dtype: float64
              Activity Period  Passenger Count  Adjusted Passenger Count  \
Boarding Area
A                  314.640882     13624.028630             13611.953204
B                  294.255947     38938.939200             38879.405881
C                  305.946668     40149.197576             40131.604526
D                  151.640665     62710.950791             62710.950791
E                  293.286164     71298.023217             71273.692744
F                  307.903263    139056.322983            138737.780638
G                  313.749343     16139.631657             16078.628004
Other              187.786620        12.090235                12.171963

                   Year
Boarding Area
A              3.150962
B              2.946662
C              3.063467
D              1.523438
E              2.938169
F              3.083225
G              3.141470
Other          1.881837
                     Activity Period  Passenger Count  \
```

```
GEO Region
Asia                    313.677214     16188.148776
Australia / Oceania     298.768639      2799.840650
Canada                  320.614235      7833.110588
Central America         324.778464      1220.840313
Europe                  316.598582      8634.076412
Mexico                  317.481441      5336.223002
Middle East             223.126076      2732.719518
South America           114.173414       396.758651
US                      309.875567     84951.316640

                    Adjusted Passenger Count      Year
GEO Region
Asia                           16147.810667  3.141101
Australia / Oceania             2650.383265  2.992176
Canada                          7805.730644  3.210175
Central America                 1220.840313  3.251234
Europe                          8602.128044  3.170094
Mexico                          5274.346847  3.179553
Middle East                     2732.719518  2.235764
South America                    396.758651  1.153402
US                             84859.991540  3.102956
                  Activity Period  Passenger Count  Adjusted Passenger Count  \
```

Operating Airline			
ATA Airlines	83.311992	8883.122532	8595.727324
Aer Lingus	331.485075	1589.142701	1589.142701
Aeromexico	218.109152	3718.871516	3718.871516
Air Berlin	82.825979	752.846346	752.846346
Air Canada	298.821335	8036.226729	8036.226729
...	...	...	...
Virgin Atlantic	312.743907	2019.991756	2019.991756
WestJet Airlines	299.812711	2858.033260	2858.033260
World Airways	175.514482	8.326664	8.326664
XL Airways France	113.622872	1146.148277	1123.862588
Xtra Airways	0.000000	0.000000	0.000000

Operating Airline	Year
ATA Airlines	0.851252
Aer Lingus	3.318559
Aeromexico	2.184222
Air Berlin	0.828079
Air Canada	2.991997
...	...
Virgin Atlantic	3.131538
WestJet Airlines	3.002442
World Airways	1.732051
XL Airways France	1.136870
Xtra Airways	0.000000

[77 rows x 4 columns]

Las conclusiones que sacamos con los datos que tenemos hasta ahora son que la mayoría de los clientes compran en la compañía Boarding área B y son vuelos low cost internacionales.

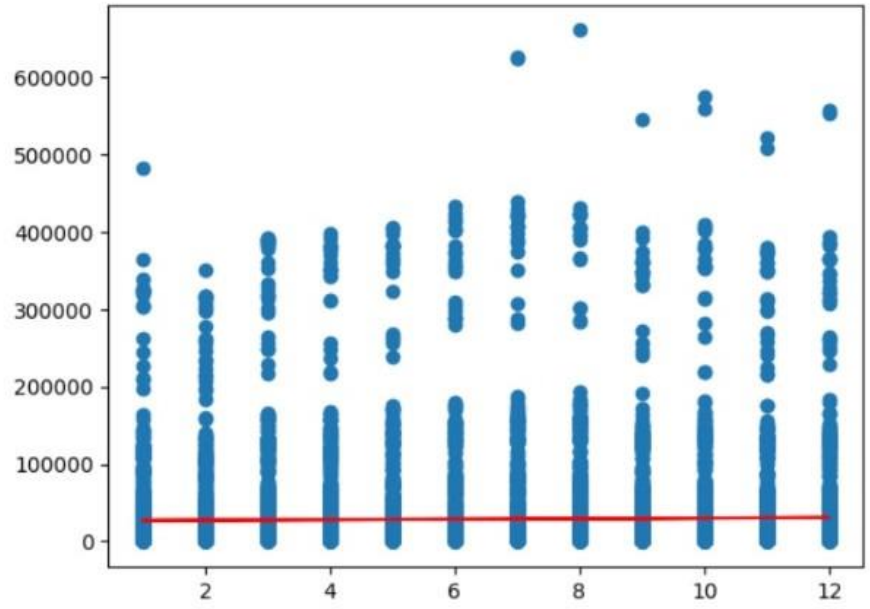
También se ve que los meses donde mas vuelos se realizan son los correspondientes al verano.

He establecido una regresión lineal entre varias variables para obtener información sobre si están relacionadas entre si o si los datos son realmente aleatorios en cuanto a la relación entre las variables

```
#creo el modelo
model = LinearRegression()

#entreno el modelo
model.fit(X_train, y_train)

#grafico los datos
plt.scatter(X_train, y_train)
plt.plot(X_train, model.predict(X_train), color='red')
plt.show()
# calculo el coeficiente r2
print("el coeficiente de correelacion es: ", r2_score(y_test, model.predict(X_test)))
```



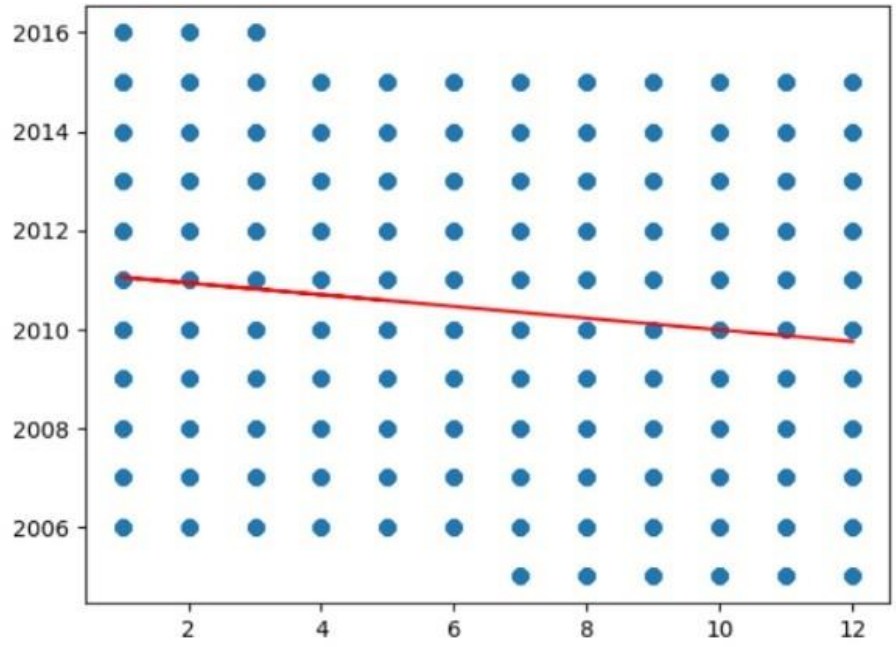

```
el coeficiente de correelacion es:  -0.00126925500044206
```

En este caso las variables que he escogido son Month y Passenger Count, siendo la recta roja la recta de regresion. Tambien he calculado el coeficinte de correlacion que no se acerca ni a 1 ni a -1 luego estas dos variables no tienen un relacion entre ellas(hablamos de relacion lineal)

```python
#creo el modelo
model = LinearRegression()

#entreno el modelo
model.fit(X_train, y_train)

#grafico los datos
plt.scatter(X_train, y_train)
plt.plot(X_train, model.predict(X_train), color='red')
plt.show()
# calculo el coeficiente r2
print("el coeficiente de correelacion es: ", r2_score(y_test, model.predict(X_test)))
```




```
el coeficiente de correelacion es:  0.012949104707880132
```

En este caso las dos variables que he pensado que podían estar relacionadas son Month y Year, sin embargo, obtenemos el mismo resultado, un coeficiente de correlación que no es cercano a 1 ni -1 por lo tanto la conclusión es la misma que en el ejemplo anterior por lo que no hay relación(lineal) entre estas variables.

He probado esto mismo con mas combinaciones de variables y lamentablemente el resultado ha sido el mismo por lo que o no están relacionadas las variables entre si (lo cual nos dificulta el estudio) o la relación que hay es de otro tipo como por ejemplo una logarítmica o una exponencial.