Análisis de datos (Big Data)

Introducción

En este proyecto se analizan los datos de tráfico del aeropuerto de San Francisco. Para ello he utilizado la librería Dask.

Se realiza un primer análisis para ver qué tipo de datos tenemos y posteriormente un análisis con mayor profundidad para poder sacar conclusiones.

Datos

Los datos están en un csv, que cargamos en un Dataset para poder analizarlos. Este fichero tiene 15007 filas.

La información que tenemos es la siguiente:

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

Esta información se ha obtenido mostrando las propiedades del dataset:

```
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)
```

Podemos ver que la gran mayoría de los datos son categóricos (strings), aunque tenemos cuatro campos numéricos. Para poder analizar bien la información, alguno de los campos categóricos, los representaremos posteriormente con un número.

Análisis Datos

Para ver qué tipo de información tenemos, vamos a realizar algunas consultas:

- ¿Cuántas compañías diferentes aparecen en el fichero?

  Hay 73 compañías

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

  ```
  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

Nota: se ha puesto el “.” como separador decimal

Para continuar con el análisis, eliminamos los registros duplicados por el campo “GEO Región”, manteniendo únicamente aquel con mayor número de pasajeros.

Con eso, obtenemos un Dataset más reducido, con 73 filas y 16 columnas.

Estos resultados los guardamos en un csv diferente al original para no perder información. Los guardamos en:

-Maximo_pasajeros.csv (contiene cada región con el valor mayor de pasajeros).

-Media_pasajeros.csv (contiene cada aerolínea con la media de pasajeros).

Tras este análisis previo, vamos a realizar uno más en profundidad.

Primero vamos a pasar algunas columnas del csv de categóricas a numéricas para que tenga repercusión en nuestros análisis. Para ello, hallaremos los valores únicos para asignarles un número, el cual usaremos de referencia para analizar los resultados.

Vamos a cambiar 5 columnas:

- Price Category Code ['Low Fare' 'Other']
- Activity Type Code ['Deplaned' 'Enplaned' 'Thru / Transit']
- Terminal ['Terminal 1' 'International' 'Terminal 3' 'Other' 'Terminal 2']
- GEO Summary ['Domestic' 'International']
- Adjusted Activity Type Code ['Deplaned' 'Enplaned' 'Thru / Transit * 2']
- Month ['July' 'August' 'September' 'October' 'November' 'April' 'December' 'January' 'February' 'March' 'May' 'June']

Sus respectivos números serán:

- Price Category Code: 0 = low, 1 = other
- Geo Summary: 0 = Domestic, 1 = International
- Adjusted Activity Type Code: 0 = deplaned, 1 = Enplaned, 2 = Thru/Transit * 2
- Terminal: 0 = International, 1 =Terminal 1, 2 = Terminal 2, 3 = Terminal 3, 4 = Other
- A cada mes se le asignara el número real de dicho mes.

Si ahora analizamos el tipo de datos que tenemos, obtenemos:

<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  int64
 6   GEO Region                15007 non-null  object
 7   Activity Type Code        15007 non-null  object
 8   Price Category Code       15007 non-null  int64
 9   Terminal                  15007 non-null  int64
 10  Boarding Area             15007 non-null  object
 11  Passenger Count           15007 non-null  int64
 12  Adjusted Activity Type Code  15007 non-null  int64
 13  Adjusted Passenger Count  15007 non-null  int64
 14  Year                      15007 non-null  int64
 15  Month                     15007 non-null  int64
dtypes: int64(9), object(7)
```

Como se puede ver, ya tenemos más datos numéricos que categóricos.

Análisis Estadístico

Procedemos a hacer algunos análisis estadísticos, como la media, la desviación estándar y la matriz de correlación:

Media:

```
Activity Period           201045.073366
GEO Summary                    0.613714
```

Price Category Code 0.872060

Terminal 1.000200

Passenger Count 29240.521090

Adjusted Activity Type Code 0.590125

Adjusted Passenger Count 29331.917105

Year 2010.385220

Month 6.551343

dtype: float64

Boarding Area	Activity Period	GEO Summary	Price Category Code	Terminal
A	201074.682488	0.847656	0.879809	0.996172
B	200976.570998	0.108379	0.613648	0.000000
C	200992.599349	0.000000	0.820847	0.000000
D	201346.314815	0.185185	0.444444	4.000000
E	200917.158145	0.205707	0.970273	2.000000
F	201035.737110	0.445171	0.955701	2.000000
G	201064.539329	0.931112	0.990982	1.000000
Other	200725.740741	0.074074	1.000000	3.000000

Boarding Area	Passenger Count	Adjusted Activity Type Code
A	11115.767656	0.505837
B	33804.871049	0.639739
C	34423.159609	0.542345
D	105124.197531	0.592593
E	48617.014269	0.689655
F	100600.343500	0.730574
G	14432.325651	0.620240
Other	7.407407	0.666667

Adjusted Passenger Count	Year	Month

Boarding Area

A 11140.662392 2010.681148 6.567656

B 33885.257903 2009.700452 6.525840

C 34444.986156 2009.860749 6.524430

D 105124.197531 2013.398148 6.500000

E 48653.051130 2009.105826 6.575505

F 101086.082789 2010.291939 6.543210

G 14521.331162 2010.579910 6.548347

Other 7.814815 2007.185185 7.222222

Activity Period GEO Summary Price Category Code \

GEO Region

Asia 201046.193706 1.0 0.999389

Australia / Oceania 200993.457259 1.0 0.997286

Canada 201070.151622 1.0 0.981664

Central America 201072.277372 1.0 1.000000

Europe 201073.937769 1.0 0.999043

Mexico 201065.279821 1.0 0.874439

Middle East 201262.476636 1.0 0.990654

South America 201193.266667 1.0 1.000000

US 201018.968432 0.0 0.698810

Terminal Passenger Count Adjusted Activity Type Code \

GEO Region

Asia 1.001833 13435.004583 0.541399

Australia / Oceania 1.000000 6417.016282 0.675712

Canada 1.423836 9777.968265 0.583216

Central America 1.000000 4946.715328 0.500000

Europe 1.000000 12755.652465 0.564864

Mexico 1.131839 7173.620628 0.673543

Middle East 1.000000 8658.612150 0.500000

South America 1.000000 2786.011111 0.500000

US 0.870450 58330.343454 0.610488

GEO Region	Adjusted Passenger Count	Year	Month
Asia	13508.552704	2010.396578	6.535900
Australia / Oceania	6495.104478	2009.869742	6.483039
Canada	9803.791255	2010.635402	6.611425
Central America	4946.715328	2010.656934	6.583942
Europe	12779.055050	2010.673528	6.584969
Mexico	7250.898655	2010.588341	6.445740
Middle East	8658.612150	2012.560748	6.401869
South America	2786.011111	2011.866667	6.600000
US	58485.878385	2010.124030	6.565465

Operating Airline	Activity Period	GEO Summary	Price Category Code	\
ATA Airlines	200586.363636	0.068182	0.0	
Aer Lingus	201151.469388	1.000000	1.0	
Aeromexico	201207.533333	1.000000	1.0	
Air Berlin	201107.500000	1.000000	1.0	
Air Canada	201123.497268	1.000000	1.0	
...	...	...	...	
Virgin Atlantic	201043.744186	1.000000	1.0	
WestJet Airlines	201125.844660	1.000000	1.0	
World Airways	201008.333333	0.666667	1.0	
XL Airways France	201339.096774	1.000000	1.0	
Xtra Airways	200608.000000	0.000000	1.0	

Operating Airline	Terminal	Passenger Count	Adjusted Activity Type Code	\
ATA Airlines	0.340909	8744.636364	0.909091	
Aer Lingus	1.000000	4407.183673	0.500000	

Aeromexico	1.000000	5463.822222	0.500000
Air Berlin	1.000000	2320.750000	0.500000
Air Canada	1.251366	18251.560109	0.500000
...	...	...	...
Virgin Atlantic	1.000000	9847.104651	0.500000
WestJet Airlines	1.000000	5338.155340	0.495146
World Airways	1.000000	261.666667	0.666667
XL Airways France	1.000000	2223.161290	0.548387
Xtra Airways	1.000000	73.000000	0.500000

Operating Airline	Adjusted Passenger Count	Year	Month
ATA Airlines	9661.659091	2005.795455	6.818182
Aer Lingus	4407.183673	2011.448980	6.571429
Aeromexico	5463.822222	2012.011111	6.422222
Air Berlin	2320.750000	2011.000000	7.500000
Air Canada	18251.560109	2011.169399	6.557377
...	...	...	...
Virgin Atlantic	9847.104651	2010.372093	6.534884
WestJet Airlines	5338.155340	2011.184466	7.398058
World Airways	261.666667	2010.000000	8.333333
XL Airways France	2240.129032	2013.322581	6.838710
Xtra Airways	73.000000	2006.000000	8.000000

[77 rows x 9 columns]

Desviación estándar

Activity Period	313.336196
GEO Summary	0.486914
Price Category Code	0.334034

Terminal 0.751869

Passenger Count 58319.509284

Adjusted Activity Type Code 0.603748

Adjusted Passenger Count 58284.182219

Year 3.137589

Month 3.464354

dtype: float64

Boarding Area	Activity Period	GEO Summary	Price Category Code	Terminal
A	314.640882	0.359389	0.325216	0.061756
B	294.255947	0.310937	0.487035	0.000000
C	305.946668	0.000000	0.383637	0.000000
D	151.640665	0.389049	0.497673	0.000000
E	293.286164	0.404458	0.169933	0.000000
F	307.903263	0.497165	0.205834	0.000000
G	313.749343	0.253295	0.094546	0.000000
Other	187.786620	0.266880	0.000000	0.000000

Boarding Area	Passenger Count	Adjusted Activity Type Code
A	13624.028630	0.524673
B	38938.939200	0.627088
C	40149.197576	0.549726
D	62710.950791	0.492112
E	71298.023217	0.676123
F	139056.322983	0.702489
G	16139.631657	0.641480
Other	12.090235	0.620174

Boarding Area	Adjusted Passenger Count	Year	Month

A	13611.953204	3.150962	3.442214
B	38879.405881	2.946662	3.481947
C	40131.604526	3.063467	3.462486
D	62710.950791	1.523438	3.457392
E	71273.692744	2.938169	3.470875
F	138737.780638	3.083225	3.488944
G	16078.628004	3.141470	3.477049
Other	12.171963	1.881837	3.619746

GEO Region	Activity Period	GEO Summary	Price Category Code
Asia	313.677214	0.0	0.024716
Australia / Oceania	298.768639	0.0	0.052058
Canada	320.614235	0.0	0.134209
Central America	324.778464	0.0	0.000000
Europe	316.598582	0.0	0.030934
Mexico	317.481441	0.0	0.331502
Middle East	223.126076	0.0	0.096446
South America	114.173414	0.0	0.000000
US	309.875567	0.0	0.458815

GEO Region	Terminal	Passenger Count	Adjusted Activity Type Code
Asia	0.055253	16188.148776	0.550794
Australia / Oceania	0.000000	2799.840650	0.673141
Canada	0.673248	7833.110588	0.632389
Central America	0.000000	1220.840313	0.500915
Europe	0.000000	8634.076412	0.576296
Mexico	0.748199	5336.223002	0.654476
Middle East	0.000000	2732.719518	0.501172
South America	0.000000	396.758651	0.502801
US	1.085994	84951.316640	0.620626

	Adjusted Passenger Count	Year	Month
GEO Region			
Asia	16147.810667	3.141101	3.491609
Australia / Oceania	2650.383265	2.992176	3.481324
Canada	7805.730644	3.210175	3.414744
Central America	1220.840313	3.251234	3.544882
Europe	8602.128044	3.170094	3.399006
Mexico	5274.346847	3.179553	3.498072
Middle East	2732.719518	2.235764	3.625315
South America	396.758651	1.153402	3.556289
US	84859.991540	3.102956	3.465904

	Activity Period	GEO Summary	Price Category Code	\
Operating Airline				
ATA Airlines	83.311992	0.254972	0.0	
Aer Lingus	331.485075	0.000000	0.0	
Aeromexico	218.109152	0.000000	0.0	
Air Berlin	82.825979	0.000000	0.0	
Air Canada	298.821335	0.000000	0.0	
...	...	...	...	
Virgin Atlantic	312.743907	0.000000	0.0	
WestJet Airlines	299.812711	0.000000	0.0	
World Airways	175.514482	0.577350	0.0	
XL Airways France	113.622872	0.000000	0.0	
Xtra Airways	0.000000	0.000000	0.0	

	Terminal	Passenger Count	Adjusted Activity Type Code	\
Operating Airline				
ATA Airlines	0.525763	8883.122532	0.801690	
Aer Lingus	0.000000	1589.142701	0.502571	
Aeromexico	0.000000	3718.871516	0.501395	

Air Berlin	0.000000	752.846346	0.507093
Air Canada	0.565865	8036.226729	0.500684
...	...	...	...
Virgin Atlantic	0.000000	2019.991756	0.500972
WestJet Airlines	0.000000	2858.033260	0.502421
World Airways	0.000000	8.326664	0.577350
XL Airways France	0.000000	1146.148277	0.567962
Xtra Airways	0.000000	0.000000	0.707107

Operating Airline	Adjusted Passenger Count	Year	Month
ATA Airlines	8595.727324	0.851252	3.718419
Aer Lingus	1589.142701	3.318559	3.470049
Aeromexico	3718.871516	2.184222	3.431623
Air Berlin	752.846346	0.828079	1.732051
Air Canada	8036.226729	2.991997	3.439614
...	...	...	...
Virgin Atlantic	2019.991756	3.131538	3.488829
WestJet Airlines	2858.033260	3.002442	2.219962
World Airways	8.326664	1.732051	2.309401
XL Airways France	1123.862588	1.136870	1.157491
Xtra Airways	0.000000	0.000000	0.000000

Con estos dos análisis conseguimos deducir lo siguiente:

-Con estos resultados podemos ver que hubo más vuelos durante los meses de junio y Julio en 2010. Al igual que volaron alrededor de 30 mil pasajeros en Domestic y en other Category Code.

-La gran mayoría de los pasajeros salen de Bording Area D (low category), en vuelos internacionales y con la compañía American Airline.

Hacemos una matriz de correlación:

	Activity Period	GEO Summary	Price Category Code	Terminal	Passenger Count	Adjusted Activity Type Code	Adjusted Passenger Count	Year	Month
Activity Period	1.000000	0.066100	-0.005754	0.103760	0.060311	-0.052450	0.059336	0.999940	-0.116571
GEO Summary	0.066100	1.000000	0.411498	0.136915	-0.395743	-0.026760	-0.396856	0.066046	-0.003234
Price Category Code	-0.005754	0.411498	1.000000	0.096685	-0.065047	0.001004	-0.064661	-0.005683	-0.005725
Terminal	0.103760	0.136915	0.096685	1.000000	0.260570	0.036001	0.261506	0.103611	0.000776
Passenger Count	0.060311	-0.395743	-0.065047	0.260570	1.000000	-0.071423	0.999941	0.060069	0.014521
Adjusted Activity Type Code	-0.052450	-0.026760	0.001004	0.036001	-0.071423	1.000000	-0.067804	-0.052364	-0.001360
Adjusted Passenger Count	0.059336	-0.396856	-0.064661	0.261506	0.999941	-0.067804	1.000000	0.059096	0.014503
Year	0.999940	0.066046	-0.005683	0.103611	0.060069	-0.052364	0.059096	1.000000	-0.127455

	Activity Period	GEO Summary	Price Category Code	Terminal	Passenger Count	Adjusted Activity Type Code	Adjusted Passenger Count	Year	Month
Month	-0.116571	-0.003234	-0.005725	0.000776	0.014521	-0.001360	0.014503	-0.127455	1.000000

Podemos ver que los datos más relacionados son: Year - GEO Summary, GEO Summary - Price Category Code, GEO Summary - Passerger Count.

Estos resultados nos hacen pensar que influye bastante si el vuelo es Domestic o International.

Conclusiones finales

Después del análisis realizado, podemos deducir lo siguiente:

-Los elementos más significativos de este conjunto de datos son: Month, Year, Passenger Count, Price Category Code, Geo Summary, Adjusted Activity Type Code, Terminal, Boarding Area, Geo Region y Operating Airlines.

-Tenemos que ha habido una mayor concentración de vuelos en Junio y Julio durante 2010.

-Influye bastante si los viajes son Domestic o International y la gran mayoría de los viajeros prefieren coger vuelos internacionales con destino a US, usando American Airlines.

-Los pasajeros que salen del área D suelen tomar vuelos de low price.