Estructures_Dataframe_Sprint2_Parte5

- Exercici 1

Descarrega el data set Airlines Delay: Airline on-time statistics and delay causes i carrega'l a un pandas Dataframe. Explora les dades que conté, i queda't únicament amb les columnes que consideris rellevants.

In [1]:
```
import pandas as pd
import numpy as np
import os
```

In [2]:
```
#'APPLESSD:\Users\sandychiereghin\Downloads\archive\DelayedFlights.csv'
#pd.read_csv('read.csv')

DelayedFlights = pd.read_csv('DelayedFlights.csv')

DelayedFlights
```

Out[2]:

	Unnamed: 0	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	...	TaxiIn	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay	I
0	0	2008	1	3	4	2003.0	1955	2211.0	2225	WN	...	4.0	8.0	0	N	0	NaN	NaN	
1	1	2008	1	3	4	754.0	735	1002.0	1000	WN	...	5.0	10.0	0	N	0	NaN	NaN	
2	2	2008	1	3	4	628.0	620	804.0	750	WN	...	3.0	17.0	0	N	0	NaN	NaN	
3	4	2008	1	3	4	1829.0	1755	1959.0	1925	WN	...	3.0	10.0	0	N	0	2.0	0.0	
4	5	2008	1	3	4	1940.0	1915	2121.0	2110	WN	...	4.0	10.0	0	N	0	NaN	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1936753	7009710	2008	12	13	6	1250.0	1220	1617.0	1552	DL	...	9.0	18.0	0	N	0	3.0	0.0	
1936754	7009717	2008	12	13	6	657.0	600	904.0	749	DL	...	15.0	34.0	0	N	0	0.0	57.0	
1936755	7009718	2008	12	13	6	1007.0	847	1149.0	1010	DL	...	8.0	32.0	0	N	0	1.0	0.0	
1936756	7009726	2008	12	13	6	1251.0	1240	1446.0	1437	DL	...	13.0	13.0	0	N	0	NaN	NaN	
1936757	7009727	2008	12	13	6	1110.0	1103	1413.0	1418	DL	...	8.0	11.0	0	N	0	NaN	NaN	

1936758 rows × 30 columns

In [3]:
```
DelayedFlights.columns
```

Out[3]:
```
Index(['Unnamed: 0', 'Year', 'Month', 'DayofMonth', 'DayOfWeek', 'DepTime',
       'CRSDepTime', 'ArrTime', 'CRSArrTime', 'UniqueCarrier', 'FlightNum',
       'TailNum', 'ActualElapsedTime', 'CRSElapsedTime', 'AirTime', 'ArrDelay',
       'DepDelay', 'Origin', 'Dest', 'Distance', 'TaxiIn', 'TaxiOut',
       'Cancelled', 'CancellationCode', 'Diverted', 'CarrierDelay',
       'WeatherDelay', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay'],
      dtype='object')
```

In [4]:
```
NewDelayedFlights = DelayedFlights.drop(columns=["DayOfWeek","TaxiIn","TaxiOut","CarrierDelay","WeatherDelay","NASDelay","SecurityDelay","LateAircraftDelay","Origin","D
NewDelayedFlights
```

Out[4]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	C
0	0	2008	1	3	2003.0	1955	2211.0	2225	WN	335	N712SW	128.0	150.0	116.0	-14.0	8.0	810	
1	1	2008	1	3	754.0	735	1002.0	1000	WN	3231	N772SW	128.0	145.0	113.0	2.0	19.0	810	
2	2	2008	1	3	628.0	620	804.0	750	WN	448	N428WN	96.0	90.0	76.0	14.0	8.0	515	
3	4	2008	1	3	1829.0	1755	1959.0	1925	WN	3920	N464WN	90.0	90.0	77.0	34.0	34.0	515	
4	5	2008	1	3	1940.0	1915	2121.0	2110	WN	378	N726SW	101.0	115.0	87.0	11.0	25.0	688	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1936753	7009710	2008	12	13	1250.0	1220	1617.0	1552	DL	1621	N938DL	147.0	152.0	120.0	25.0	30.0	906	
1936754	7009717	2008	12	13	657.0	600	904.0	749	DL	1631	N3743H	127.0	109.0	78.0	75.0	57.0	481	
1936755	7009718	2008	12	13	1007.0	847	1149.0	1010	DL	1631	N909DA	162.0	143.0	122.0	99.0	80.0	689	
1936756	7009726	2008	12	13	1251.0	1240	1446.0	1437	DL	1639	N646DL	115.0	117.0	89.0	9.0	11.0	533	
1936757	7009727	2008	12	13	1110.0	1103	1413.0	1418	DL	1641	N908DL	123.0	135.0	104.0	-5.0	7.0	874	

1936758 rows × 20 columns

Exercici 2

Fes un informe complet del data set:.

- Resumeix estadísticament les columnes d'interès
- Troba quantes dades faltants hi ha per columna
- Crea columnes noves (velocitat mitjana del vol, si ha arribat tard o no...)
- Taula de les aerolínies amb més endarreriments acumulats
- Quins són els vols més llargs? I els més endarrerits? Etc.

In [5]:
```
NewDelayedFlights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 20 columns):
 #   Column             Dtype  
---  ------             -----  
 0   Unnamed: 0         int64  
 1   Year               int64  
 2   Month              int64  
 3   DayofMonth         int64  
 4   DepTime            float64
 5   CRSDepTime         int64  
 6   ArrTime            float64
 7   CRSArrTime         int64  
 8   UniqueCarrier      object 
 9   FlightNum          int64  
 10  TailNum            object 
 11  ActualElapsedTime  float64
 12  CRSElapsedTime     float64
 13  AirTime            float64
 14  ArrDelay           float64
 15  DepDelay           float64
 16  Distance           int64  
 17  Cancelled          int64  
 18  CancellationCode   object 
 19  Diverted           int64  
dtypes: float64(7), int64(10), object(3)
memory usage: 295.5+ MB
```

In [38]:
```
NewDelayedFlights['Year'].info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 1936758 entries, 0 to 1936757
Series name: Year
Non-Null Count    Dtype
--------------    -----
1936758 non-null  int64
dtypes: int64(1)
memory usage: 14.8 MB
```

In [6]:
```
NewDelayedFlightsFiltrado = NewDelayedFlights.fillna (0)
NewDelayedFlightsFiltrado
```

Out[6]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	C
0	0	2008	1	3	2003.0	1955	2211.0	2225	WN	335	N712SW	128.0	150.0	116.0	-14.0	8.0	810	
1	1	2008	1	3	754.0	735	1002.0	1000	WN	3231	N772SW	128.0	145.0	113.0	2.0	19.0	810	
2	2	2008	1	3	628.0	620	804.0	750	WN	448	N428WN	96.0	90.0	76.0	14.0	8.0	515	
3	4	2008	1	3	1829.0	1755	1959.0	1925	WN	3920	N464WN	90.0	90.0	77.0	34.0	34.0	515	
4	5	2008	1	3	1940.0	1915	2121.0	2110	WN	378	N726SW	101.0	115.0	87.0	11.0	25.0	688	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1936753	7009710	2008	12	13	1250.0	1220	1617.0	1552	DL	1621	N938DL	147.0	152.0	120.0	25.0	30.0	906	
1936754	7009717	2008	12	13	657.0	600	904.0	749	DL	1631	N3743H	127.0	109.0	78.0	75.0	57.0	481	
1936755	7009718	2008	12	13	1007.0	847	1149.0	1010	DL	1631	N909DA	162.0	143.0	122.0	99.0	80.0	689	
1936756	7009726	2008	12	13	1251.0	1240	1446.0	1437	DL	1639	N646DL	115.0	117.0	89.0	9.0	11.0	533	
1936757	7009727	2008	12	13	1110.0	1103	1413.0	1418	DL	1641	N908DL	123.0	135.0	104.0	-5.0	7.0	874	

1936758 rows × 20 columns

In [40]:
```
NewDelayedFlightsFiltrado["Cancelled"]
```

Out[40]:
```
0          0
1          0
2          0
3          0
4          0
          ..
1936753    0
1936754    0
1936755    0
1936756    0
1936757    0
Name: Cancelled, Length: 1936758, dtype: int64
```

In [7]:
```
#iloc filtro por indice
#loc filtro por identificador de cada fila y con las columnas que se quieren seleccionar.
# NewDelayedFlightsFiltrado.loc[[0:10],["Year","Dest"]]

NewDelayedFlightsFiltrado.iloc[10:21]
```

Out[7]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	Cancell
10	17	2008	1	3	754.0	745	940.0	955	WN	1144	N778SW	226.0	250.0	205.0	-15.0	9.0	1489	
11	18	2008	1	3	1323.0	1255	1526.0	1510	WN	4	N674AA	123.0	135.0	110.0	16.0	28.0	838	
12	19	2008	1	3	1416.0	1325	1512.0	1435	WN	54	N643SW	56.0	70.0	49.0	37.0	51.0	220	
13	21	2008	1	3	1657.0	1625	1754.0	1735	WN	623	N724SW	57.0	70.0	47.0	19.0	32.0	220	
14	22	2008	1	3	1900.0	1840	1956.0	1950	WN	717	N786SW	56.0	70.0	49.0	6.0	20.0	220	
15	23	2008	1	3	1039.0	1030	1133.0	1140	WN	1244	N714CB	54.0	70.0	47.0	-7.0	9.0	220	
16	25	2008	1	3	1520.0	1455	1619.0	1605	WN	2553	N394SW	59.0	70.0	50.0	14.0	25.0	220	
17	26	2008	1	3	1422.0	1255	1657.0	1610	WN	188	N215WN	155.0	195.0	143.0	47.0	87.0	1093	
18	27	2008	1	3	1954.0	1925	2239.0	2235	WN	1754	N243WN	165.0	190.0	155.0	4.0	29.0	1093	
19	30	2008	1	3	2107.0	1945	2334.0	2230	WN	362	N798SW	147.0	165.0	134.0	64.0	82.0	972	
20	33	2008	1	3	1312.0	1300	1546.0	1550	WN	1397	N247WN	154.0	170.0	140.0	-4.0	12.0	972	

In [8]:
```
NewDelayedFlightsFiltrado[NewDelayedFlightsFiltrado["Cancelled"]>0]
# and = & , or = |
```

Out[8]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	C
1542406	5463024	2008	10	27	1622.0	1420	0.0	1520	WN	27	N601WN	0.0	60.0	0.0	0.0	122.0	276	
1546593	5484245	2008	10	25	1323.0	1255	0.0	1442	XE	2347	N26549	0.0	107.0	0.0	0.0	28.0	529	
1547161	5486876	2008	10	22	1825.0	1815	0.0	1927	XE	2819	N12946	0.0	72.0	0.0	0.0	10.0	351	
1547178	5486924	2008	10	22	1733.0	1715	0.0	1818	XE	2890	N16944	0.0	63.0	0.0	0.0	18.0	253	
1548271	5491819	2008	10	15	1943.0	1745	0.0	1857	XE	2117	N26545	0.0	72.0	0.0	0.0	118.0	295	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1934590	7002526	2008	12	7	1526.0	1444	0.0	1654	DL	1743	N958DL	0.0	130.0	0.0	0.0	42.0	712	
1935491	7006018	2008	12	10	1431.0	1422	0.0	1527	DL	1405	N906DL	0.0	125.0	0.0	0.0	9.0	689	
1935651	7006289	2008	12	10	1459.0	1447	0.0	1650	DL	1706	N914DN	0.0	123.0	0.0	0.0	12.0	712	
1935876	7006809	2008	12	11	1026.0	955	0.0	1219	DL	892	N928DL	0.0	144.0	0.0	0.0	31.0	760	
1936470	7008584	2008	12	12	703.0	630	0.0	734	DL	1372	N908DE	0.0	64.0	0.0	0.0	33.0	185	

633 rows × 20 columns

In [9]:
```
NewDelayedFlightsFiltrado[(NewDelayedFlightsFiltrado["Diverted"]>0)]
```

Out[9]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	C
1280	1763	2008	1	3	922.0	915	0.0	1050	WN	1069	N630WN	0.0	95.0	0.0	0.0	7.0	480	
1372	1911	2008	1	3	2325.0	1900	0.0	2030	WN	2092	N302SW	0.0	90.0	0.0	0.0	265.0	447	
1776	2651	2008	1	4	1949.0	1905	0.0	1910	WN	1403	N504SW	0.0	65.0	0.0	0.0	44.0	335	
1831	2726	2008	1	4	737.0	705	0.0	825	WN	178	N718SW	0.0	80.0	0.0	0.0	32.0	358	
2244	3672	2008	1	4	1849.0	1630	0.0	1755	WN	239	N636WN	0.0	85.0	0.0	0.0	139.0	345	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1934369	7001470	2008	12	7	1928.0	1645	29.0	2032	DL	133	N3764D	0.0	407.0	0.0	0.0	163.0	2586	
1934921	7004192	2008	12	9	1957.0	1905	22.0	2013	DL	792	N3739P	0.0	128.0	0.0	0.0	52.0	606	
1935596	7006200	2008	12	10	714.0	640	1153.0	859	DL	1610	N956DL	0.0	79.0	0.0	0.0	34.0	341	
1935716	7006401	2008	12	11	1355.0	1106	7.0	1950	DL	26	N3747D	0.0	344.0	0.0	0.0	169.0	2475	
1935978	7007034	2008	12	11	1527.0	1520	2106.0	1708	DL	1102	N924DL	0.0	108.0	0.0	0.0	7.0	533	

7754 rows × 20 columns

In [10]:
```
NewDelayedFlightsFiltrado[(NewDelayedFlightsFiltrado["ArrDelay"]>0)]
```

Out[10]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	C
1	1	2008	1	3	754.0	735	1002.0	1000	WN	3231	N772SW	128.0	145.0	113.0	2.0	19.0	810	
2	2	2008	1	3	628.0	620	804.0	750	WN	448	N428WN	96.0	90.0	76.0	14.0	8.0	515	
3	4	2008	1	3	1829.0	1755	1959.0	1925	WN	3920	N464WN	90.0	90.0	77.0	34.0	34.0	515	
4	5	2008	1	3	1940.0	1915	2121.0	2110	WN	378	N726SW	101.0	115.0	87.0	11.0	25.0	688	
5	6	2008	1	3	1937.0	1830	2037.0	1940	WN	509	N763SW	240.0	250.0	230.0	57.0	67.0	1591	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1936752	7009709	2008	12	13	1552.0	1520	1735.0	1718	DL	1620	N905DE	43.0	58.0	27.0	17.0	32.0	151	
1936753	7009710	2008	12	13	1250.0	1220	1617.0	1552	DL	1621	N938DL	147.0	152.0	120.0	25.0	30.0	906	
1936754	7009717	2008	12	13	657.0	600	904.0	749	DL	1631	N3743H	127.0	109.0	78.0	75.0	57.0	481	
1936755	7009718	2008	12	13	1007.0	847	1149.0	1010	DL	1631	N909DA	162.0	143.0	122.0	99.0	80.0	689	
1936756	7009726	2008	12	13	1251.0	1240	1446.0	1437	DL	1639	N646DL	115.0	117.0	89.0	9.0	11.0	533	

1723415 rows × 20 columns

In [11]:
```
vuelosCancelados = 633
vuelosDesviados = 7754
vuelosRetrasados = 1247488
vuelosTotales = 1936758

pVuelosCancelados = round(100*vuelosCancelados)/vuelosTotales

print("El pergentaje de vuelos cancelados es: ", pVuelosCancelados)

pVuelosDesviados = round(100*vuelosDesviados)/vuelosTotales

print("El pergentaje de vuelos desviados es: ", pVuelosDesviados)

pVuelosRetrasados = round(100*vuelosRetrasados)/vuelosTotales

print("El pergentaje de vuelos cancelados es: ", pVuelosRetrasados)
```

```
El pergentaje de vuelos cancelados es:  0.03268348446217855
El pergentaje de vuelos desviados es:  0.40035977649246834
El pergentaje de vuelos cancelados es:  64.41114481003822
```

In [12]:
```
def definirRetraso (ArrDelay):
    if ArrDelay > 0 :
        return "Delayed"


NewDelayedFlightsFiltrado["DelayLabel"] = NewDelayedFlightsFiltrado["ArrDelay"].apply(definirRetraso)

NewDelayedFlightsFiltrado
```

Out[12]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	Cancelle
0	0	2008	1	3	2003.0	1955	2211.0	2225	WN	335	...	128.0	150.0	116.0	-14.0	8.0	810	
1	1	2008	1	3	754.0	735	1002.0	1000	WN	3231	...	128.0	145.0	113.0	2.0	19.0	810	
2	2	2008	1	3	628.0	620	804.0	750	WN	448	...	96.0	90.0	76.0	14.0	8.0	515	
3	4	2008	1	3	1829.0	1755	1959.0	1925	WN	3920	...	90.0	90.0	77.0	34.0	34.0	515	
4	5	2008	1	3	1940.0	1915	2121.0	2110	WN	378	...	101.0	115.0	87.0	11.0	25.0	688	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1936753	7009710	2008	12	13	1250.0	1220	1617.0	1552	DL	1621	...	147.0	152.0	120.0	25.0	30.0	906	
1936754	7009717	2008	12	13	657.0	600	904.0	749	DL	1631	...	127.0	109.0	78.0	75.0	57.0	481	
1936755	7009718	2008	12	13	1007.0	847	1149.0	1010	DL	1631	...	162.0	143.0	122.0	99.0	80.0	689	
1936756	7009726	2008	12	13	1251.0	1240	1446.0	1437	DL	1639	...	115.0	117.0	89.0	9.0	11.0	533	
1936757	7009727	2008	12	13	1110.0	1103	1413.0	1418	DL	1641	...	123.0	135.0	104.0	-5.0	7.0	874	

1936758 rows × 21 columns

In [13]:
```
NewDelayedFlightsFiltrado["Speed/h"] = (NewDelayedFlightsFiltrado["Distance"] / NewDelayedFlightsFiltrado["AirTime"])*60

NewDelayedFlightsFiltrado.head()
```

Out[13]:

	Unnamed: 0	Year	Month	DayofMonth	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	...	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Distance	Cancelled	CancellationCode	Dive
0	0	2008	1	3	2003.0	1955	2211.0	2225	WN	335	...	150.0	116.0	-14.0	8.0	810	0	N	
1	1	2008	1	3	754.0	735	1002.0	1000	WN	3231	...	145.0	113.0	2.0	19.0	810	0	N	
2	2	2008	1	3	628.0	620	804.0	750	WN	448	...	90.0	76.0	14.0	8.0	515	0	N	
3	4	2008	1	3	1829.0	1755	1959.0	1925	WN	3920	...	90.0	77.0	34.0	34.0	515	0	N	
4	5	2008	1	3	1940.0	1915	2121.0	2110	WN	378	...	115.0	87.0	11.0	25.0	688	0	N	

5 rows × 22 columns

In [14]:
```
#vuelos más largos
NewDelayedFlightsFiltrado['AirTime'].nlargest(n=5)
```

Out[14]:
```
1488690    1091.0
1367047     733.0
362529      664.0
556381      655.0
556379      654.0
Name: AirTime, dtype: float64
```

In [15]:
```
#vuelos más cortos
NewDelayedFlightsFiltrado['AirTime'].nsmallest(n=5)
```

Out[15]:
```
1280    0.0
1372    0.0
1776    0.0
1831    0.0
2244    0.0
Name: AirTime, dtype: float64
```

In [22]:
```
grouped = NewDelayedFlightsFiltrado.groupby('UniqueCarrier')
print (grouped['ArrDelay'].agg(np.sum))
```

```
UniqueCarrier
9E     2420468.0
AA     8889066.0
AQ       15814.0
AS     1406735.0
B6     3025749.0
CO     4045932.0
DL     4535644.0
EV     3888131.0
F9      788549.0
FL     3100150.0
HA      255613.0
MQ     6396704.0
NW     3462075.0
OH     2675993.0
OO     5978936.0
UA     6733013.0
US     3571867.0
WN    11319092.0
XE     5176042.0
YV     3691461.0
Name: ArrDelay, dtype: float64
```

In [30]:
```
file_name = 'NewDelayedFlightsFiltrado_sheet.xlsx'
NewDelayedFlightsFiltrado.head().to_excel(file_name)
```

In [32]:
```
NewDelayedFlightsFiltrado.shape
```

Out[32]:
```
(1936758, 22)
```

In [ ]: