Sprint02_Tasca05

April 1, 2022

1 Sprint 02. Tasca 05

1.1 By José Manuel Castaño

1.2 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.

[20]:
```
import pandas as pd

airlines = pd.read_csv('DelayedFlights.csv', index_col =0)
pd.set_option('display.max_columns', None)
airlines.head(10)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\numpy\lib\arraysetops.py:583:
FutureWarning: elementwise comparison failed; returning scalar instead, but in
the future will perform elementwise comparison
  mask |= (ar1 == a)
```

[20]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	\
0	2008	1	3	4	2003.0	1955	2211.0	
1	2008	1	3	4	754.0	735	1002.0	
2	2008	1	3	4	628.0	620	804.0	
4	2008	1	3	4	1829.0	1755	1959.0	
5	2008	1	3	4	1940.0	1915	2121.0	
6	2008	1	3	4	1937.0	1830	2037.0	
10	2008	1	3	4	706.0	700	916.0	
11	2008	1	3	4	1644.0	1510	1845.0	
15	2008	1	3	4	1029.0	1020	1021.0	
16	2008	1	3	4	1452.0	1425	1640.0	

	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	\
0	2225	WN	335	N712SW	128.0	
1	1000	WN	3231	N772SW	128.0	
2	750	WN	448	N428WN	96.0	
4	1925	WN	3920	N464WN	90.0	
5	2110	WN	378	N726SW	101.0	

```
6         1940            WN        509  N763SW              240.0
10         915            WN        100  N690SW              130.0
11        1725            WN       1333  N334SW              121.0
15        1010            WN       2272  N263WN               52.0
16        1625            WN        675  N286WN              228.0

    CRSElapsedTime  AirTime  ArrDelay  DepDelay Origin Dest  Distance  TaxiIn  \
0            150.0    116.0     -14.0       8.0    IAD  TPA       810     4.0
1            145.0    113.0       2.0      19.0    IAD  TPA       810     5.0
2             90.0     76.0      14.0       8.0    IND  BWI       515     3.0
4             90.0     77.0      34.0      34.0    IND  BWI       515     3.0
5            115.0     87.0      11.0      25.0    IND  JAX       688     4.0
6            250.0    230.0      57.0      67.0    IND  LAS      1591     3.0
10           135.0    106.0       1.0       6.0    IND  MCO       828     5.0
11           135.0    107.0      80.0      94.0    IND  MCO       828     6.0
15            50.0     37.0      11.0       9.0    IND  MDW       162     6.0
16           240.0    213.0      15.0      27.0    IND  PHX      1489     7.0

    TaxiOut  Cancelled CancellationCode  Diverted  CarrierDelay  WeatherDelay  \
0       8.0          0                N         0           NaN           NaN
1      10.0          0                N         0           NaN           NaN
2      17.0          0                N         0           NaN           NaN
4      10.0          0                N         0           2.0           0.0
5      10.0          0                N         0           NaN           NaN
6       7.0          0                N         0          10.0           0.0
10     19.0          0                N         0           NaN           NaN
11      8.0          0                N         0           8.0           0.0
15      9.0          0                N         0           NaN           NaN
16      8.0          0                N         0           3.0           0.0

    NASDelay  SecurityDelay  LateAircraftDelay
0        NaN            NaN                NaN
1        NaN            NaN                NaN
2        NaN            NaN                NaN
4        0.0            0.0               32.0
5        NaN            NaN                NaN
6        0.0            0.0               47.0
10       NaN            NaN                NaN
11       0.0            0.0               72.0
15       NaN            NaN                NaN
16       0.0            0.0               12.0
```

[2]:
```
airlines.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1936758 entries, 0 to 7009727
Data columns (total 29 columns):
```

```
 #   Column             Dtype
---  ------             -----
 0   Year               int64
 1   Month              int64
 2   DayofMonth         int64
 3   DayOfWeek          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  Origin             object
 17  Dest               object
 18  Distance           int64
 19  TaxiIn             float64
 20  TaxiOut            float64
 21  Cancelled          int64
 22  CancellationCode   object
 23  Diverted           int64
 24  CarrierDelay       float64
 25  WeatherDelay       float64
 26  NASDelay           float64
 27  SecurityDelay      float64
 28  LateAircraftDelay  float64
dtypes: float64(14), int64(10), object(5)
memory usage: 443.3+ MB
```

[14]:
```
#Mirem els vols cancel.lats
airlines['Cancelled'].value_counts()
```

[14]:
```
0    1936125
1        633
Name: Cancelled, dtype: int64
```

[6]:
```
airlines[airlines['Cancelled'] ==1]
```

[6]:

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	\
5463024	2008	10	27	1	1622.0	1420	NaN	
5484245	2008	10	25	6	1323.0	1255	NaN	
5486876	2008	10	22	3	1825.0	1815	NaN	

```
5486924  2008     10          22          3   1733.0        1715      NaN
5491819  2008     10          15          3   1943.0        1745      NaN
…        …    …          …         …       …         …      …
7002526  2008     12           7          7   1526.0        1444      NaN
7006018  2008     12          10          3   1431.0        1422      NaN
7006289  2008     12          10          3   1459.0        1447      NaN
7006809  2008     12          11          4   1026.0         955      NaN
7008584  2008     12          12          5    703.0         630      NaN

         CRSArrTime UniqueCarrier  FlightNum TailNum  ActualElapsedTime  \
5463024        1520            WN         27  N601WN                NaN
5484245        1442            XE       2347  N26549                NaN
5486876        1927            XE       2819  N12946                NaN
5486924        1818            XE       2890  N16944                NaN
5491819        1857            XE       2117  N26545                NaN
…               …             …        …     …                    …
7002526        1654            DL       1743  N958DL                NaN
7006018        1527            DL       1405  N906DL                NaN
7006289        1650            DL       1706  N914DN                NaN
7006809        1219            DL        892  N928DL                NaN
7008584         734            DL       1372  N908DE                NaN

         CRSElapsedTime  AirTime  ArrDelay  DepDelay Origin Dest  Distance  \
5463024            60.0      NaN       NaN     122.0    HOU  HRL       276
5484245           107.0      NaN       NaN      28.0    CLT  EWR       529
5486876            72.0      NaN       NaN      10.0    JAN  IAH       351
5486924            63.0      NaN       NaN      18.0    IAH  BTR       253
5491819            72.0      NaN       NaN     118.0    IAH  HRL       295
…                   …       …        …        …     …  …       …
7002526           130.0      NaN       NaN      42.0    BUF  ATL       712
7006018           125.0      NaN       NaN       9.0    ATL  IAH       689
7006289           123.0      NaN       NaN      12.0    ATL  BUF       712
7006809           144.0      NaN       NaN      31.0    ATL  JFK       760
7008584            64.0      NaN       NaN      33.0    LGA  BOS       185

         TaxiIn  TaxiOut  Cancelled CancellationCode  Diverted  CarrierDelay  \
5463024     NaN     19.0          1                A         0           NaN
5484245     NaN      NaN          1                B         0           NaN
5486876     NaN      NaN          1                C         0           NaN
5486924     NaN      NaN          1                B         0           NaN
5491819     NaN      NaN          1                B         0           NaN
…            …       …         …                …        …           …
7002526     NaN      NaN          1                A         0           NaN
7006018     NaN      NaN          1                C         0           NaN
7006289     NaN     37.0          1                A         0           NaN
7006809     NaN      NaN          1                A         0           NaN
7008584     NaN     33.0          1                B         0           NaN
```

```
         WeatherDelay  NASDelay  SecurityDelay  LateAircraftDelay
5463024           NaN       NaN            NaN                NaN
5484245           NaN       NaN            NaN                NaN
5486876           NaN       NaN            NaN                NaN
5486924           NaN       NaN            NaN                NaN
5491819           NaN       NaN            NaN                NaN
…                 …         …              …                  …
7002526           NaN       NaN            NaN                NaN
7006018           NaN       NaN            NaN                NaN
7006289           NaN       NaN            NaN                NaN
7006809           NaN       NaN            NaN                NaN
7008584           NaN       NaN            NaN                NaN

[633 rows x 29 columns]
```

1.2.1 Exploració de dades

Observem 2 grans blocs d'anal.lisi: la de vols amb el seu temps, endarreriments, etc i la de vols cancel.lats amb les seves causes. En aquest exercisi analitzarem el primer bloc.

Veiem que ActualElapsedTime = AirTime + TaxiIn + TaxiOut. Year és sempre 2008.

Per tal d'estudiar temps de vol, endarreriments, velocitats, etc, seleccionem els camps més rellevants

[21]:
```
#Nova taula flights (copia) amb els camps rellevants i sense els vols cancel.
 ↪lats
flights = airlines[airlines['Cancelled']␣
 ↪==0][['Month','DayofMonth','DayOfWeek',␣
 ↪'UniqueCarrier','FlightNum','CRSDepTime','AirTime' ,'ActualElapsedTime',␣
 ↪'Distance','ArrDelay' ]].copy()
flights.head(10)
```

[21]:
```
    Month  DayofMonth  DayOfWeek UniqueCarrier  FlightNum  CRSDepTime  \
0       1           3          4            WN        335        1955
1       1           3          4            WN       3231         735
2       1           3          4            WN        448         620
4       1           3          4            WN       3920        1755
5       1           3          4            WN        378        1915
6       1           3          4            WN        509        1830
10      1           3          4            WN        100         700
11      1           3          4            WN       1333        1510
15      1           3          4            WN       2272        1020
16      1           3          4            WN        675        1425

    AirTime  ActualElapsedTime  Distance  ArrDelay
0     116.0              128.0       810     -14.0
1     113.0              128.0       810       2.0
```

```
2      76.0                96.0       515      14.0
4      77.0                90.0       515      34.0
5      87.0               101.0       688      11.0
6     230.0               240.0      1591      57.0
10    106.0               130.0       828       1.0
11    107.0               121.0       828      80.0
15     37.0                52.0       162      11.0
16    213.0               228.0      1489      15.0
```

1.3 - 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.
```

[54]:
```
#Resum estadístic de les columnes amb més interès
flights[['AirTime','ActualElapsedTime','Distance','ArrDelay']].describe()
```

[54]:
```
             AirTime  ActualElapsedTime      Distance      ArrDelay
count   1.928371e+06       1.928371e+06  1.936125e+06  1.928371e+06
mean    1.082771e+02       1.333059e+02  7.657387e+02  4.219988e+01
std     6.864261e+01       7.206007e+01  5.744840e+02  5.678472e+01
min     0.000000e+00       1.400000e+01  1.100000e+01 -1.090000e+02
25%     5.800000e+01       8.000000e+01  3.380000e+02  9.000000e+00
50%     9.000000e+01       1.160000e+02  6.060000e+02  2.400000e+01
75%     1.370000e+02       1.650000e+02  9.980000e+02  5.600000e+01
max     1.091000e+03       1.114000e+03  4.962000e+03  2.461000e+03
```

[55]:
```
#Dades que falten per columna
flights.isnull().sum()
```

[55]:
```
Month                   0
DayofMonth              0
DayOfWeek               0
UniqueCarrier           0
FlightNum               0
CRSDepTime              0
AirTime              7754
ActualElapsedTime    7754
Distance                0
ArrDelay             7754
dtype: int64
```

[56]:
```
flights[flights['AirTime'].isnull()]
```

[56]:

	Month	DayofMonth	DayOfWeek	UniqueCarrier	FlightNum	CRSDepTime	AirTime	ActualElapsedTime	Distance	ArrDelay
1763	1	3	4	WN	1069	915	NaN	NaN	480	NaN
1911	1	3	4	WN	2092	1900	NaN	NaN	447	NaN
2651	1	4	5	WN	1403	1905	NaN	NaN	335	NaN
2726	1	4	5	WN	178	705	NaN	NaN	358	NaN
3672	1	4	5	WN	239	1630	NaN	NaN	345	NaN
…	…	…	…	…	…	…	…	…	…	…
7001470	12	7	7	DL	133	1645	NaN	NaN	2586	NaN
7004192	12	9	2	DL	792	1905	NaN	NaN	606	NaN
7006200	12	10	3	DL	1610	640	NaN	NaN	341	NaN
7006401	12	11	4	DL	26	1106	NaN	NaN	2475	NaN
7007034	12	11	4	DL	1102	1520	NaN	NaN	533	NaN

[7754 rows x 10 columns]

[57]:
```
#Com veiem que els registres nulls coincideixen i són els únics en els tres␣
 ↪camps, crec que lo millor és eliminar els registres
flights = flights.dropna(how='any')
#Comprovo que no queden registres nuls
flights.isnull().sum()
```

[57]:
```
Month                0
DayofMonth           0
DayOfWeek            0
UniqueCarrier        0
FlightNum            0
CRSDepTime           0
AirTime              0
ActualElapsedTime    0
Distance             0
ArrDelay             0
dtype: int64
```

[17]:
```
#Afegeixo velocitat mitjana, delayed/on time,
flights['vmed'] = flights['Distance'] / (flights['AirTime']/60)
flights['delay'] = flights['ArrDelay'].apply(lambda x: 'Delayed' if x>0 else␣
↪'On time')
flights.head(10)
```

[17]:

	Month	DayofMonth	DayOfWeek	UniqueCarrier	FlightNum	CRSDepTime	\
0	1	3	4	WN	335	1955	
1	1	3	4	WN	3231	735	
2	1	3	4	WN	448	620	
4	1	3	4	WN	3920	1755	
5	1	3	4	WN	378	1915	
6	1	3	4	WN	509	1830	
10	1	3	4	WN	100	700	
11	1	3	4	WN	1333	1510	
15	1	3	4	WN	2272	1020	
16	1	3	4	WN	675	1425	

	AirTime	ActualElapsedTime	Distance	ArrDelay	vmed	delay
0	116.0	128.0	810	-14.0	418.965517	On time
1	113.0	128.0	810	2.0	430.088496	Delayed
2	76.0	96.0	515	14.0	406.578947	Delayed
4	77.0	90.0	515	34.0	401.298701	Delayed
5	87.0	101.0	688	11.0	474.482759	Delayed
6	230.0	240.0	1591	57.0	415.043478	Delayed
10	106.0	130.0	828	1.0	468.679245	Delayed
11	107.0	121.0	828	80.0	464.299065	Delayed
15	37.0	52.0	162	11.0	262.702703	Delayed
16	213.0	228.0	1489	15.0	419.436620	Delayed

[18]:
```
#Aerolinies amb més endarreriments acumulats i el promig d'endarreriment
#No acumula els que han arribat abans del temps previst)
endarreriments = flights[flights['ArrDelay']>0].
↪groupby(flights['UniqueCarrier'])['ArrDelay'].agg(['sum', 'mean'])
endarreriments.sort_values(by ='sum', ascending = False)
```

[18]:

UniqueCarrier	sum	mean
WN	11609347.0	35.752200
AA	9007400.0	52.308693
UA	6850031.0	55.247086
MQ	6443938.0	49.323276
OO	6019322.0	49.362172
XE	5227263.0	55.424629
DL	4620911.0	45.786501
CO	4159659.0	49.729324
EV	3920352.0	52.153146

```
YV              3706402.0  58.563131
US              3678122.0  44.175278
NW              3498782.0  48.329056
FL              3129740.0  48.143921
B6              3080816.0  63.947859
OH              2696559.0  54.915261
9E              2443157.0  52.097343
AS              1438977.0  42.101202
F9               798332.0  31.053835
HA               257923.0  35.827615
AQ                17134.0  26.198777
```

[13]:
```
#Vols més llargs
vols_mes_llargs = flights.groupby(['UniqueCarrier','FlightNum'])['Distance'].
 ↪agg(max)
vols_mes_llargs.sort_values(ascending=False).head(10)
```

[13]:
```
UniqueCarrier  FlightNum
CO             14           4962
               15           4962
DL             851          4502
               1560         4502
               1282         4502
               1273         4502
               850          4502
               1561         4502
UA             1410         4243
AA             73           4243
Name: Distance, dtype: int64
```

[19]:
```
#Vols més endarrerits
vols_mes_endarrerits = flights[flights['ArrDelay']>0].
 ↪groupby(['UniqueCarrier','FlightNum'])['ArrDelay'].agg(max)
vols_mes_endarrerits.sort_values(ascending=False).head(10)
```

[19]:
```
UniqueCarrier  FlightNum
NW             808          2461.0
               1699         2453.0
               1107         1951.0
MQ             3538         1707.0
NW             357          1655.0
               512          1583.0
               1472         1542.0
AA             2398         1525.0
NW             804          1510.0
               1743         1490.0
Name: ArrDelay, dtype: float64
```

1.4 - Exercici 3

Exporta el data set net i amb les noves columnes a Excel.

[22]:
```
# Excel no permet importacions superiors a 1,048,576 registres. Per tant,␣
↪partim el DataFrame
flights1 = flights.iloc[0:1000000]
flights2 = flights.iloc[1000000:]

flights1.to_excel('flights_excel1.xlsx', sheet_name='Sheet1')
flights2.to_excel('flights_excel2.xlsx', sheet_name='Sheet2')
```

[ ]:
```

```