S09_T01_SkLearn_Train_Test

May 1, 2022

1 S09 T01: Practicant amb training i test sets

[48]:
```
import random
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statistics
import scipy.stats
from scipy.stats import norm
from scipy import stats
from scipy.stats import t
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score
from numpy.polynomial.polynomial import polyfit
from sklearn.decomposition import PCA
import matplotlib.cm as cm
```

1.1 Exercici 1. Parteix el conjunt de dadesDelayedFlights.csv en train i test. Estudia els dos conjunts per separat, a nivell descriptiu

Importem el fitxer i li fem una ullada a la informació que conté.

[49]:
```
vols = pd.read_csv('//home/rusi/Escritorio/rubenIT/DataSources/DelayedFlights.
↪csv')#importem i li assignem un nom de dataframe
```

[50]:
```
print(vols.info())
print(vols.describe())
print(vols.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 30 columns):
```

```
 #   Column             Dtype
---  ------             -----
 0   Unnamed: 0         int64
 1   Year               int64
 2   Month              int64
 3   DayofMonth         int64
 4   DayOfWeek          int64
 5   DepTime            float64
 6   CRSDepTime         int64
 7   ArrTime            float64
 8   CRSArrTime         int64
 9   UniqueCarrier      object
 10  FlightNum          int64
 11  TailNum            object
 12  ActualElapsedTime  float64
 13  CRSElapsedTime     float64
 14  AirTime            float64
 15  ArrDelay           float64
 16  DepDelay           float64
 17  Origin             object
 18  Dest               object
 19  Distance           int64
 20  TaxiIn             float64
 21  TaxiOut            float64
 22  Cancelled          int64
 23  CancellationCode   object
 24  Diverted           int64
 25  CarrierDelay       float64
 26  WeatherDelay       float64
 27  NASDelay           float64
 28  SecurityDelay      float64
 29  LateAircraftDelay  float64
dtypes: float64(14), int64(11), object(5)
memory usage: 443.3+ MB
None
         Unnamed: 0       Year         Month    DayofMonth     DayOfWeek  \
count  1.936758e+06  1936758.0  1.936758e+06  1.936758e+06  1.936758e+06
mean   3.341651e+06     2008.0  6.111106e+00  1.575347e+01  3.984827e+00
std    2.066065e+06        0.0  3.482546e+00  8.776272e+00  1.995966e+00
min    0.000000e+00     2008.0  1.000000e+00  1.000000e+00  1.000000e+00
25%    1.517452e+06     2008.0  3.000000e+00  8.000000e+00  2.000000e+00
50%    3.242558e+06     2008.0  6.000000e+00  1.600000e+01  4.000000e+00
75%    4.972467e+06     2008.0  9.000000e+00  2.300000e+01  6.000000e+00
max    7.009727e+06     2008.0  1.200000e+01  3.100000e+01  7.000000e+00

            DepTime    CRSDepTime       ArrTime    CRSArrTime     FlightNum  \
count  1.936758e+06  1.936758e+06  1.929648e+06  1.936758e+06  1.936758e+06
mean   1.518534e+03  1.467473e+03  1.610141e+03  1.634225e+03  2.184263e+03
```

```
std    4.504853e+02  4.247668e+02  5.481781e+02  4.646347e+02  1.944702e+03
min    1.000000e+00  0.000000e+00  1.000000e+00  0.000000e+00  1.000000e+00
25%    1.203000e+03  1.135000e+03  1.316000e+03  1.325000e+03  6.100000e+02
50%    1.545000e+03  1.510000e+03  1.715000e+03  1.705000e+03  1.543000e+03
75%    1.900000e+03  1.815000e+03  2.030000e+03  2.014000e+03  3.422000e+03
max    2.400000e+03  2.359000e+03  2.400000e+03  2.400000e+03  9.742000e+03

       …      Distance        TaxiIn       TaxiOut     Cancelled  \
count  …  1.936758e+06  1.929648e+06  1.936303e+06  1.936758e+06
mean   …  7.656862e+02  6.812975e+00  1.823220e+01  3.268348e-04
std    …  5.744797e+02  5.273595e+00  1.433853e+01  1.807562e-02
min    …  1.100000e+01  0.000000e+00  0.000000e+00  0.000000e+00
25%    …  3.380000e+02  4.000000e+00  1.000000e+01  0.000000e+00
50%    …  6.060000e+02  6.000000e+00  1.400000e+01  0.000000e+00
75%    …  9.980000e+02  8.000000e+00  2.100000e+01  0.000000e+00
max    …  4.962000e+03  2.400000e+02  4.220000e+02  1.000000e+00

           Diverted  CarrierDelay  WeatherDelay      NASDelay  SecurityDelay  \
count  1.936758e+06  1.247488e+06  1.247488e+06  1.247488e+06   1.247488e+06
mean   4.003598e-03  1.917940e+01  3.703571e+00  1.502164e+01   9.013714e-02
std    6.314722e-02  4.354621e+01  2.149290e+01  3.383305e+01   2.022714e+00
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
50%    0.000000e+00  2.000000e+00  0.000000e+00  2.000000e+00   0.000000e+00
75%    0.000000e+00  2.100000e+01  0.000000e+00  1.500000e+01   0.000000e+00
max    1.000000e+00  2.436000e+03  1.352000e+03  1.357000e+03   3.920000e+02

       LateAircraftDelay
count       1.247488e+06
mean        2.529647e+01
std         4.205486e+01
min         0.000000e+00
25%         0.000000e+00
50%         8.000000e+00
75%         3.300000e+01
max         1.316000e+03

[8 rows x 25 columns]
   Unnamed: 0  Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  \
0           0  2008      1           3          4   2003.0        1955
1           1  2008      1           3          4    754.0         735
2           2  2008      1           3          4    628.0         620
3           4  2008      1           3          4   1829.0        1755
4           5  2008      1           3          4   1940.0        1915

   ArrTime  CRSArrTime UniqueCarrier  …  TaxiIn TaxiOut  Cancelled  \
0   2211.0        2225            WN  …     4.0     8.0          0
1   1002.0        1000            WN  …     5.0    10.0          0
```

```
2    804.0         750            WN  …     3.0    17.0          0
3   1959.0        1925            WN  …     3.0    10.0          0
4   2121.0        2110            WN  …     4.0    10.0          0

   CancellationCode  Diverted  CarrierDelay  WeatherDelay NASDelay  \
0                 N         0           NaN           NaN      NaN
1                 N         0           NaN           NaN      NaN
2                 N         0           NaN           NaN      NaN
3                 N         0           2.0           0.0      0.0
4                 N         0           NaN           NaN      NaN

  SecurityDelay  LateAircraftDelay
0           NaN                NaN
1           NaN                NaN
2           NaN                NaN
3           0.0               32.0
4           NaN                NaN

[5 rows x 30 columns]
```

[51]:
```
vols.columns
```

[51]:
```
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')
```

[52]:
```
vols.describe(include = "all")
```

[52]:
```
          Unnamed: 0       Year         Month    DayofMonth     DayOfWeek  \
count   1.936758e+06  1936758.0  1.936758e+06  1.936758e+06  1.936758e+06
unique           NaN        NaN           NaN           NaN           NaN
top              NaN        NaN           NaN           NaN           NaN
freq             NaN        NaN           NaN           NaN           NaN
mean    3.341651e+06     2008.0  6.111106e+00  1.575347e+01  3.984827e+00
std     2.066065e+06        0.0  3.482546e+00  8.776272e+00  1.995966e+00
min     0.000000e+00     2008.0  1.000000e+00  1.000000e+00  1.000000e+00
25%     1.517452e+06     2008.0  3.000000e+00  8.000000e+00  2.000000e+00
50%     3.242558e+06     2008.0  6.000000e+00  1.600000e+01  4.000000e+00
75%     4.972467e+06     2008.0  9.000000e+00  2.300000e+01  6.000000e+00
max     7.009727e+06     2008.0  1.200000e+01  3.100000e+01  7.000000e+00

             DepTime    CRSDepTime       ArrTime    CRSArrTime UniqueCarrier  \
count   1.936758e+06  1.936758e+06  1.929648e+06  1.936758e+06       1936758
```

unique	NaN	NaN	NaN	NaN	20
top	NaN	NaN	NaN	NaN	WN
freq	NaN	NaN	NaN	NaN	377602
mean	1.518534e+03	1.467473e+03	1.610141e+03	1.634225e+03	NaN
std	4.504853e+02	4.247668e+02	5.481781e+02	4.646347e+02	NaN
min	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	NaN
25%	1.203000e+03	1.135000e+03	1.316000e+03	1.325000e+03	NaN
50%	1.545000e+03	1.510000e+03	1.715000e+03	1.705000e+03	NaN
75%	1.900000e+03	1.815000e+03	2.030000e+03	2.014000e+03	NaN
max	2.400000e+03	2.359000e+03	2.400000e+03	2.400000e+03	NaN

	…	TaxiIn	TaxiOut	Cancelled	CancellationCode	\
count	…	1.929648e+06	1.936303e+06	1.936758e+06	1936758	
unique	…	NaN	NaN	NaN	4	
top	…	NaN	NaN	NaN	N	
freq	…	NaN	NaN	NaN	1936125	
mean	…	6.812975e+00	1.823220e+01	3.268348e-04	NaN	
std	…	5.273595e+00	1.433853e+01	1.807562e-02	NaN	
min	…	0.000000e+00	0.000000e+00	0.000000e+00	NaN	
25%	…	4.000000e+00	1.000000e+01	0.000000e+00	NaN	
50%	…	6.000000e+00	1.400000e+01	0.000000e+00	NaN	
75%	…	8.000000e+00	2.100000e+01	0.000000e+00	NaN	
max	…	2.400000e+02	4.220000e+02	1.000000e+00	NaN	

	Diverted	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	\
count	1.936758e+06	1.247488e+06	1.247488e+06	1.247488e+06	1.247488e+06	
unique	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	
mean	4.003598e-03	1.917940e+01	3.703571e+00	1.502164e+01	9.013714e-02	
std	6.314722e-02	4.354621e+01	2.149290e+01	3.383305e+01	2.022714e+00	
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
50%	0.000000e+00	2.000000e+00	0.000000e+00	2.000000e+00	0.000000e+00	
75%	0.000000e+00	2.100000e+01	0.000000e+00	1.500000e+01	0.000000e+00	
max	1.000000e+00	2.436000e+03	1.352000e+03	1.357000e+03	3.920000e+02	

	LateAircraftDelay
count	1.247488e+06
unique	NaN
top	NaN
freq	NaN
mean	2.529647e+01
std	4.205486e+01
min	0.000000e+00
25%	0.000000e+00
50%	8.000000e+00

```
75%           3.300000e+01
max           1.316000e+03

[11 rows x 30 columns]
```

Observem que a totes les columnes tenim Nan que caldria tractar.

[53]:
```
#analitzem el percentatge de NaN per cada un dels camps
(vols.isnull().sum())*100 / len(vols)
```

[53]:
```
Unnamed: 0            0.000000
Year                  0.000000
Month                 0.000000
DayofMonth            0.000000
DayOfWeek             0.000000
DepTime               0.000000
CRSDepTime            0.000000
ArrTime               0.367108
CRSArrTime            0.000000
UniqueCarrier         0.000000
FlightNum             0.000000
TailNum               0.000258
ActualElapsedTime     0.433043
CRSElapsedTime        0.010223
AirTime               0.433043
ArrDelay              0.433043
DepDelay              0.000000
Origin                0.000000
Dest                  0.000000
Distance              0.000000
TaxiIn                0.367108
TaxiOut               0.023493
Cancelled             0.000000
CancellationCode      0.000000
Diverted              0.000000
CarrierDelay         35.588855
WeatherDelay         35.588855
NASDelay             35.588855
SecurityDelay        35.588855
LateAircraftDelay    35.588855
dtype: float64
```

Comprovem que el percentatge de NaN no és significatiu per a tots els camps, excepte els "*Delay". Decidim prescindir dels primers, i assignem el valor 0 als "*Delay" amb la funció ***fillna()***. En aquest últim cas, podríem haber optat també per substituir el valor NaN per la mitja de retards.

[54]:
```
#Assignem valor 0 a les 4 columnes "*Delay"
vols02=vols.iloc[:,25:30].fillna(0)
print(vols02.describe(include="all"))
print(vols02.head())
```

```
       CarrierDelay  WeatherDelay      NASDelay  SecurityDelay  \
count  1.936758e+06  1.936758e+06  1.936758e+06   1.936758e+06
mean   1.235367e+01  2.385512e+00  9.675607e+00   5.805836e-02
std    3.613493e+01  1.734036e+01  2.808958e+01   1.623934e+00
min    0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
50%    0.000000e+00  0.000000e+00  0.000000e+00   0.000000e+00
75%    1.000000e+01  0.000000e+00  6.000000e+00   0.000000e+00
max    2.436000e+03  1.352000e+03  1.357000e+03   3.920000e+02

       LateAircraftDelay
count       1.936758e+06
mean        1.629374e+01
std         3.585904e+01
min         0.000000e+00
25%         0.000000e+00
50%         0.000000e+00
75%         1.800000e+01
max         1.316000e+03
   CarrierDelay  WeatherDelay  NASDelay  SecurityDelay  LateAircraftDelay
0           0.0           0.0       0.0            0.0                0.0
1           0.0           0.0       0.0            0.0                0.0
2           0.0           0.0       0.0            0.0                0.0
3           2.0           0.0       0.0            0.0               32.0
4           0.0           0.0       0.0            0.0                0.0
```

Eliminem les columnes "*Delay" i les tornem a unir al dataframe

[55]:
```
#eliminem les columnes
vols03=vols.drop(vols.iloc[:,25:30],axis=1)
vols03.describe(include="all")
```

[55]:

```
          Unnamed: 0       Year         Month    DayofMonth     DayOfWeek  \
count   1.936758e+06  1936758.0  1.936758e+06  1.936758e+06  1.936758e+06
unique           NaN        NaN           NaN           NaN           NaN
top              NaN        NaN           NaN           NaN           NaN
freq             NaN        NaN           NaN           NaN           NaN
mean    3.341651e+06     2008.0  6.111106e+00  1.575347e+01  3.984827e+00
std     2.066065e+06        0.0  3.482546e+00  8.776272e+00  1.995966e+00
min     0.000000e+00     2008.0  1.000000e+00  1.000000e+00  1.000000e+00
25%     1.517452e+06     2008.0  3.000000e+00  8.000000e+00  2.000000e+00
50%     3.242558e+06     2008.0  6.000000e+00  1.600000e+01  4.000000e+00
75%     4.972467e+06     2008.0  9.000000e+00  2.300000e+01  6.000000e+00
```

```
max     7.009727e+06     2008.0  1.200000e+01  3.100000e+01  7.000000e+00

             DepTime    CRSDepTime       ArrTime    CRSArrTime UniqueCarrier  \
count   1.936758e+06  1.936758e+06  1.929648e+06  1.936758e+06       1936758
unique           NaN           NaN           NaN           NaN            20
top              NaN           NaN           NaN           NaN            WN
freq             NaN           NaN           NaN           NaN        377602
mean    1.518534e+03  1.467473e+03  1.610141e+03  1.634225e+03           NaN
std     4.504853e+02  4.247668e+02  5.481781e+02  4.646347e+02           NaN
min     1.000000e+00  0.000000e+00  1.000000e+00  0.000000e+00           NaN
25%     1.203000e+03  1.135000e+03  1.316000e+03  1.325000e+03           NaN
50%     1.545000e+03  1.510000e+03  1.715000e+03  1.705000e+03           NaN
75%     1.900000e+03  1.815000e+03  2.030000e+03  2.014000e+03           NaN
max     2.400000e+03  2.359000e+03  2.400000e+03  2.400000e+03           NaN

        …      ArrDelay      DepDelay   Origin     Dest      Distance  \
count   …  1.928371e+06  1.936758e+06  1936758  1936758  1.936758e+06
unique  …           NaN           NaN      303      304           NaN
top     …           NaN           NaN      ATL      ORD           NaN
freq    …           NaN           NaN   131613   108984           NaN
mean    …  4.219988e+01  4.318518e+01      NaN      NaN  7.656862e+02
std     …  5.678472e+01  5.340250e+01      NaN      NaN  5.744797e+02
min     … -1.090000e+02  6.000000e+00      NaN      NaN  1.100000e+01
25%     …  9.000000e+00  1.200000e+01      NaN      NaN  3.380000e+02
50%     …  2.400000e+01  2.400000e+01      NaN      NaN  6.060000e+02
75%     …  5.600000e+01  5.300000e+01      NaN      NaN  9.980000e+02
max     …  2.461000e+03  2.467000e+03      NaN      NaN  4.962000e+03

             TaxiIn       TaxiOut     Cancelled CancellationCode  \
count   1.929648e+06  1.936303e+06  1.936758e+06          1936758
unique           NaN           NaN           NaN                4
top              NaN           NaN           NaN                N
freq             NaN           NaN           NaN          1936125
mean    6.812975e+00  1.823220e+01  3.268348e-04              NaN
std     5.273595e+00  1.433853e+01  1.807562e-02              NaN
min     0.000000e+00  0.000000e+00  0.000000e+00              NaN
25%     4.000000e+00  1.000000e+01  0.000000e+00              NaN
50%     6.000000e+00  1.400000e+01  0.000000e+00              NaN
75%     8.000000e+00  2.100000e+01  0.000000e+00              NaN
max     2.400000e+02  4.220000e+02  1.000000e+00              NaN

            Diverted
count   1.936758e+06
unique           NaN
top              NaN
freq             NaN
mean    4.003598e-03
```

```
std     6.314722e-02
min     0.000000e+00
25%     0.000000e+00
50%     0.000000e+00
75%     0.000000e+00
max     1.000000e+00

[11 rows x 25 columns]
```

[56]:
```
#les afegim sense NaN al DataFrame a on les havíem eliminat
vols04 = vols03.merge(vols02, how='inner', left_index=True, right_index=True)
```

[57]:
```
vols04.describe(include="all")
```

[57]:
```
          Unnamed: 0       Year         Month    DayofMonth     DayOfWeek  \
count   1.936758e+06  1936758.0  1.936758e+06  1.936758e+06  1.936758e+06
unique           NaN        NaN           NaN           NaN           NaN
top              NaN        NaN           NaN           NaN           NaN
freq             NaN        NaN           NaN           NaN           NaN
mean    3.341651e+06     2008.0  6.111106e+00  1.575347e+01  3.984827e+00
std     2.066065e+06        0.0  3.482546e+00  8.776272e+00  1.995966e+00
min     0.000000e+00     2008.0  1.000000e+00  1.000000e+00  1.000000e+00
25%     1.517452e+06     2008.0  3.000000e+00  8.000000e+00  2.000000e+00
50%     3.242558e+06     2008.0  6.000000e+00  1.600000e+01  4.000000e+00
75%     4.972467e+06     2008.0  9.000000e+00  2.300000e+01  6.000000e+00
max     7.009727e+06     2008.0  1.200000e+01  3.100000e+01  7.000000e+00

             DepTime    CRSDepTime       ArrTime    CRSArrTime UniqueCarrier  \
count   1.936758e+06  1.936758e+06  1.929648e+06  1.936758e+06       1936758
unique           NaN           NaN           NaN           NaN            20
top              NaN           NaN           NaN           NaN            WN
freq             NaN           NaN           NaN           NaN        377602
mean    1.518534e+03  1.467473e+03  1.610141e+03  1.634225e+03           NaN
std     4.504853e+02  4.247668e+02  5.481781e+02  4.646347e+02           NaN
min     1.000000e+00  0.000000e+00  1.000000e+00  0.000000e+00           NaN
25%     1.203000e+03  1.135000e+03  1.316000e+03  1.325000e+03           NaN
50%     1.545000e+03  1.510000e+03  1.715000e+03  1.705000e+03           NaN
75%     1.900000e+03  1.815000e+03  2.030000e+03  2.014000e+03           NaN
max     2.400000e+03  2.359000e+03  2.400000e+03  2.400000e+03           NaN

        …        TaxiIn       TaxiOut     Cancelled  CancellationCode  \
count   …  1.929648e+06  1.936303e+06  1.936758e+06           1936758
unique  …           NaN           NaN           NaN                 4
top     …           NaN           NaN           NaN                 N
freq    …           NaN           NaN           NaN           1936125
mean    …  6.812975e+00  1.823220e+01  3.268348e-04               NaN
std     …  5.273595e+00  1.433853e+01  1.807562e-02               NaN
```

```
min     …  0.000000e+00  0.000000e+00  0.000000e+00              NaN
25%     …  4.000000e+00  1.000000e+01  0.000000e+00              NaN
50%     …  6.000000e+00  1.400000e+01  0.000000e+00              NaN
75%     …  8.000000e+00  2.100000e+01  0.000000e+00              NaN
max     …  2.400000e+02  4.220000e+02  1.000000e+00              NaN

            Diverted  CarrierDelay  WeatherDelay      NASDelay SecurityDelay  \
count   1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
unique           NaN           NaN           NaN           NaN           NaN
top              NaN           NaN           NaN           NaN           NaN
freq             NaN           NaN           NaN           NaN           NaN
mean    4.003598e-03  1.235367e+01  2.385512e+00  9.675607e+00  5.805836e-02
std     6.314722e-02  3.613493e+01  1.734036e+01  2.808958e+01  1.623934e+00
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%     0.000000e+00  1.000000e+01  0.000000e+00  6.000000e+00  0.000000e+00
max     1.000000e+00  2.436000e+03  1.352000e+03  1.357000e+03  3.920000e+02

        LateAircraftDelay
count        1.936758e+06
unique                NaN
top                   NaN
freq                  NaN
mean         1.629374e+01
std          3.585904e+01
min          0.000000e+00
25%          0.000000e+00
50%          0.000000e+00
75%          1.800000e+01
max          1.316000e+03

[11 rows x 30 columns]
```

[58]:
```
vols04=vols04.fillna(0)
```

[59]:
```
#analitzem el percentatge de NaN per cada un dels camps i comprovem
 ↪l'efectivitat dels passos seguits
(vols04.isnull().sum())*100 / len(vols04)
```

[59]:
```
Unnamed: 0          0.0
Year                0.0
Month               0.0
DayofMonth          0.0
DayOfWeek           0.0
DepTime             0.0
CRSDepTime          0.0
```

```
ArrTime              0.0
CRSArrTime           0.0
UniqueCarrier        0.0
FlightNum            0.0
TailNum              0.0
ActualElapsedTime    0.0
CRSElapsedTime       0.0
AirTime              0.0
ArrDelay             0.0
DepDelay             0.0
Origin               0.0
Dest                 0.0
Distance             0.0
TaxiIn               0.0
TaxiOut              0.0
Cancelled            0.0
CancellationCode     0.0
Diverted             0.0
CarrierDelay         0.0
WeatherDelay         0.0
NASDelay             0.0
SecurityDelay        0.0
LateAircraftDelay    0.0
dtype: float64
```

[60]:
```
vols04.describe(include="all")
```

[60]:

	Unnamed: 0	Year	Month	DayofMonth	DayOfWeek
count	1.936758e+06	1936758.0	1.936758e+06	1.936758e+06	1.936758e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	3.341651e+06	2008.0	6.111106e+00	1.575347e+01	3.984827e+00
std	2.066065e+06	0.0	3.482546e+00	8.776272e+00	1.995966e+00
min	0.000000e+00	2008.0	1.000000e+00	1.000000e+00	1.000000e+00
25%	1.517452e+06	2008.0	3.000000e+00	8.000000e+00	2.000000e+00
50%	3.242558e+06	2008.0	6.000000e+00	1.600000e+01	4.000000e+00
75%	4.972467e+06	2008.0	9.000000e+00	2.300000e+01	6.000000e+00
max	7.009727e+06	2008.0	1.200000e+01	3.100000e+01	7.000000e+00

	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier
count	1.936758e+06	1.936758e+06	1.936758e+06	1.936758e+06	1936758
unique	NaN	NaN	NaN	NaN	20
top	NaN	NaN	NaN	NaN	WN
freq	NaN	NaN	NaN	NaN	377602
mean	1.518534e+03	1.467473e+03	1.604230e+03	1.634225e+03	NaN
std	4.504853e+02	4.247668e+02	5.557685e+02	4.646347e+02	NaN

```
min     1.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00           NaN
25%     1.203000e+03  1.135000e+03  1.313000e+03  1.325000e+03           NaN
50%     1.545000e+03  1.510000e+03  1.714000e+03  1.705000e+03           NaN
75%     1.900000e+03  1.815000e+03  2.030000e+03  2.014000e+03           NaN
max     2.400000e+03  2.359000e+03  2.400000e+03  2.400000e+03           NaN

        …        TaxiIn       TaxiOut     Cancelled  CancellationCode  \
count   …  1.936758e+06  1.936758e+06  1.936758e+06           1936758
unique  …           NaN           NaN           NaN                 4
top     …           NaN           NaN           NaN                 N
freq    …           NaN           NaN           NaN           1936125
mean    …  6.787964e+00  1.822792e+01  3.268348e-04               NaN
std     …  5.280008e+00  1.433957e+01  1.807562e-02               NaN
min     …  0.000000e+00  0.000000e+00  0.000000e+00               NaN
25%     …  4.000000e+00  1.000000e+01  0.000000e+00               NaN
50%     …  5.000000e+00  1.400000e+01  0.000000e+00               NaN
75%     …  8.000000e+00  2.100000e+01  0.000000e+00               NaN
max     …  2.400000e+02  4.220000e+02  1.000000e+00               NaN

            Diverted  CarrierDelay  WeatherDelay      NASDelay SecurityDelay  \
count   1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
unique           NaN           NaN           NaN           NaN           NaN
top              NaN           NaN           NaN           NaN           NaN
freq             NaN           NaN           NaN           NaN           NaN
mean    4.003598e-03  1.235367e+01  2.385512e+00  9.675607e+00  5.805836e-02
std     6.314722e-02  3.613493e+01  1.734036e+01  2.808958e+01  1.623934e+00
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%     0.000000e+00  1.000000e+01  0.000000e+00  6.000000e+00  0.000000e+00
max     1.000000e+00  2.436000e+03  1.352000e+03  1.357000e+03  3.920000e+02

        LateAircraftDelay
count        1.936758e+06
unique                NaN
top                   NaN
freq                  NaN
mean         1.629374e+01
std          3.585904e+01
min          0.000000e+00
25%          0.000000e+00
50%          0.000000e+00
75%          1.800000e+01
max          1.316000e+03

[11 rows x 30 columns]
```

Ara ja tenim el dataframe sense NaN, i a priori, net. Partim els valors en Train i Test. Com a columna objectiu (y=target column) escollim "Distance".

- Donarem un pes a train de 2/3 de la mostra, i 1/3 per test.
- Fixarem el random_state per tal de reproduir els resultats.

[61]:
```
print(vols04.info())
print(vols04.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 30 columns):
 #   Column             Dtype
---  ------             -----
 0   Unnamed: 0         int64
 1   Year               int64
 2   Month              int64
 3   DayofMonth         int64
 4   DayOfWeek          int64
 5   DepTime            float64
 6   CRSDepTime         int64
 7   ArrTime            float64
 8   CRSArrTime         int64
 9   UniqueCarrier      object
 10  FlightNum          int64
 11  TailNum            object
 12  ActualElapsedTime  float64
 13  CRSElapsedTime     float64
 14  AirTime            float64
 15  ArrDelay           float64
 16  DepDelay           float64
 17  Origin             object
 18  Dest               object
 19  Distance           int64
 20  TaxiIn             float64
 21  TaxiOut            float64
 22  Cancelled          int64
 23  CancellationCode   object
 24  Diverted           int64
 25  CarrierDelay       float64
 26  WeatherDelay       float64
 27  NASDelay           float64
 28  SecurityDelay      float64
 29  LateAircraftDelay  float64
dtypes: float64(14), int64(11), object(5)
memory usage: 443.3+ MB
None
   Unnamed: 0  Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  \
```

```
0           0  2008      1           3           4   2003.0        1955
1           1  2008      1           3           4    754.0         735
2           2  2008      1           3           4    628.0         620
3           4  2008      1           3           4   1829.0        1755
4           5  2008      1           3           4   1940.0        1915

   ArrTime  CRSArrTime UniqueCarrier  …  TaxiIn TaxiOut  Cancelled  \
0   2211.0        2225            WN  …     4.0     8.0          0
1   1002.0        1000            WN  …     5.0    10.0          0
2    804.0         750            WN  …     3.0    17.0          0
3   1959.0        1925            WN  …     3.0    10.0          0
4   2121.0        2110            WN  …     4.0    10.0          0

   CancellationCode  Diverted  CarrierDelay  WeatherDelay NASDelay  \
0                 N         0           0.0           0.0      0.0
1                 N         0           0.0           0.0      0.0
2                 N         0           0.0           0.0      0.0
3                 N         0           2.0           0.0      0.0
4                 N         0           0.0           0.0      0.0

  SecurityDelay  LateAircraftDelay
0           0.0                0.0
1           0.0                0.0
2           0.0                0.0
3           0.0               32.0
4           0.0                0.0

[5 rows x 30 columns]
```

Prescindim de les columnes que no són numèriques.

[62]:
```
x=vols04.drop(columns=["Unnamed:␣
  ↪0","Distance","UniqueCarrier","TailNum","Origin","Dest","CancellationCode"])
y=vols04.Distance
print(x.describe())
print(y.describe())
```

```
             Year         Month    DayofMonth     DayOfWeek       DepTime  \
count  1936758.0  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
mean      2008.0  6.111106e+00  1.575347e+01  3.984827e+00  1.518534e+03
std          0.0  3.482546e+00  8.776272e+00  1.995966e+00  4.504853e+02
min       2008.0  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
25%       2008.0  3.000000e+00  8.000000e+00  2.000000e+00  1.203000e+03
50%       2008.0  6.000000e+00  1.600000e+01  4.000000e+00  1.545000e+03
75%       2008.0  9.000000e+00  2.300000e+01  6.000000e+00  1.900000e+03
max       2008.0  1.200000e+01  3.100000e+01  7.000000e+00  2.400000e+03

         CRSDepTime       ArrTime    CRSArrTime     FlightNum  \
```

```
count  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
mean   1.467473e+03  1.604230e+03  1.634225e+03  2.184263e+03
std    4.247668e+02  5.557685e+02  4.646347e+02  1.944702e+03
min    0.000000e+00  0.000000e+00  0.000000e+00  1.000000e+00
25%    1.135000e+03  1.313000e+03  1.325000e+03  6.100000e+02
50%    1.510000e+03  1.714000e+03  1.705000e+03  1.543000e+03
75%    1.815000e+03  2.030000e+03  2.014000e+03  3.422000e+03
max    2.359000e+03  2.400000e+03  2.400000e+03  9.742000e+03

       ActualElapsedTime  …      DepDelay        TaxiIn       TaxiOut  \
count       1.936758e+06  …  1.936758e+06  1.936758e+06  1.936758e+06
mean        1.327286e+02  …  4.318518e+01  6.787964e+00  1.822792e+01
std         7.243471e+01  …  5.340250e+01  5.280008e+00  1.433957e+01
min         0.000000e+00  …  6.000000e+00  0.000000e+00  0.000000e+00
25%         8.000000e+01  …  1.200000e+01  4.000000e+00  1.000000e+01
50%         1.160000e+02  …  2.400000e+01  5.000000e+00  1.400000e+01
75%         1.650000e+02  …  5.300000e+01  8.000000e+00  2.100000e+01
max         1.114000e+03  …  2.467000e+03  2.400000e+02  4.220000e+02

          Cancelled      Diverted  CarrierDelay  WeatherDelay      NASDelay  \
count  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
mean   3.268348e-04  4.003598e-03  1.235367e+01  2.385512e+00  9.675607e+00
std    1.807562e-02  6.314722e-02  3.613493e+01  1.734036e+01  2.808958e+01
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%    0.000000e+00  0.000000e+00  1.000000e+01  0.000000e+00  6.000000e+00
max    1.000000e+00  1.000000e+00  2.436000e+03  1.352000e+03  1.357000e+03

       SecurityDelay  LateAircraftDelay
count   1.936758e+06       1.936758e+06
mean    5.805836e-02       1.629374e+01
std     1.623934e+00       3.585904e+01
min     0.000000e+00       0.000000e+00
25%     0.000000e+00       0.000000e+00
50%     0.000000e+00       0.000000e+00
75%     0.000000e+00       1.800000e+01
max     3.920000e+02       1.316000e+03

[8 rows x 23 columns]
count    1.936758e+06
mean     7.656862e+02
std      5.744797e+02
min      1.100000e+01
25%      3.380000e+02
50%      6.060000e+02
75%      9.980000e+02
max      4.962000e+03
```

```
Name: Distance, dtype: float64
```

[63]:
```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33,␣
↪random_state=4)
```

[64]:
```
print(x_train.describe())
print(x_train.head())
```

```
              Year         Month    DayofMonth     DayOfWeek       DepTime  \
count  1297627.0  1.297627e+06  1.297627e+06  1.297627e+06  1.297627e+06
mean      2008.0  6.109843e+00  1.574720e+01  3.984289e+00  1.518756e+03
std          0.0  3.481609e+00  8.776006e+00  1.996141e+00  4.504804e+02
min       2008.0  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
25%       2008.0  3.000000e+00  8.000000e+00  2.000000e+00  1.203000e+03
50%       2008.0  6.000000e+00  1.600000e+01  4.000000e+00  1.545000e+03
75%       2008.0  9.000000e+00  2.300000e+01  6.000000e+00  1.900000e+03
max       2008.0  1.200000e+01  3.100000e+01  7.000000e+00  2.400000e+03

         CRSDepTime       ArrTime    CRSArrTime     FlightNum  \
count  1.297627e+06  1.297627e+06  1.297627e+06  1.297627e+06
mean   1.467621e+03  1.604393e+03  1.634281e+03  2.183296e+03
std    4.248092e+02  5.558326e+02  4.647430e+02  1.943757e+03
min    0.000000e+00  0.000000e+00  0.000000e+00  1.000000e+00
25%    1.135000e+03  1.313000e+03  1.325000e+03  6.100000e+02
50%    1.510000e+03  1.714000e+03  1.706000e+03  1.543000e+03
75%    1.815000e+03  2.030000e+03  2.015000e+03  3.422000e+03
max    2.359000e+03  2.400000e+03  2.400000e+03  9.741000e+03

       ActualElapsedTime  …      DepDelay        TaxiIn       TaxiOut  \
count       1.297627e+06  …  1.297627e+06  1.297627e+06  1.297627e+06
mean        1.326897e+02  …  4.318557e+01  6.790388e+00  1.822460e+01
std         7.241072e+01  …  5.345342e+01  5.280226e+00  1.432080e+01
min         0.000000e+00  …  6.000000e+00  0.000000e+00  0.000000e+00
25%         8.000000e+01  …  1.200000e+01  4.000000e+00  1.000000e+01
50%         1.160000e+02  …  2.400000e+01  5.000000e+00  1.400000e+01
75%         1.650000e+02  …  5.300000e+01  8.000000e+00  2.100000e+01
max         7.760000e+02  …  2.457000e+03  2.250000e+02  4.220000e+02

          Cancelled      Diverted  CarrierDelay  WeatherDelay      NASDelay  \
count  1.297627e+06  1.297627e+06  1.297627e+06  1.297627e+06  1.297627e+06
mean   3.398511e-04  3.962618e-03  1.232975e+01  2.399529e+00  9.690718e+00
std    1.843193e-02  6.282451e-02  3.609058e+01  1.746325e+01  2.813189e+01
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%    0.000000e+00  0.000000e+00  9.000000e+00  0.000000e+00  6.000000e+00
max    1.000000e+00  1.000000e+00  1.951000e+03  1.352000e+03  1.357000e+03
```

```
       SecurityDelay  LateAircraftDelay
count   1.297627e+06       1.297627e+06
mean    5.720211e-02       1.629101e+01
std     1.593506e+00       3.589830e+01
min     0.000000e+00       0.000000e+00
25%     0.000000e+00       0.000000e+00
50%     0.000000e+00       0.000000e+00
75%     0.000000e+00       1.800000e+01
max     3.920000e+02       1.316000e+03

[8 rows x 23 columns]
         Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  \
84384    2008      1           7          1   1603.0        1555   1823.0
1424846  2008      8           6          3   1830.0        1735   2131.0
1923886  2008     12          25          4   2019.0        2005   2136.0
383942   2008      3          12          3   1653.0        1630   1753.0
1811287  2008     12          13          6   1857.0        1752   2239.0

         CRSArrTime  FlightNum  ActualElapsedTime  …  DepDelay  TaxiIn  \
84384          1841        777              140.0  …       8.0     6.0
1424846        2026        716              181.0  …      55.0    18.0
1923886        2138        487              137.0  …      14.0    10.0
383942         1735         47               60.0  …      23.0     2.0
1811287        2127        167              402.0  …      65.0     4.0

         TaxiOut  Cancelled  Diverted  CarrierDelay  WeatherDelay  NASDelay  \
84384       12.0          0         0           0.0           0.0       0.0
1424846     24.0          0         0           2.0           0.0      10.0
1923886      9.0          0         0           0.0           0.0       0.0
383942      10.0          0         0           0.0           0.0       0.0
1811287     17.0          0         0          65.0           0.0       7.0

         SecurityDelay  LateAircraftDelay
84384              0.0                0.0
1424846            0.0               53.0
1923886            0.0                0.0
383942             0.0               18.0
1811287            0.0                0.0

[5 rows x 23 columns]
```

[65]:
```
print(x_test.describe())
print(x_test.head())
```

```
            Year          Month     DayofMonth      DayOfWeek        DepTime  \
count  639131.0  639131.000000  639131.000000  639131.000000  639131.000000
mean     2008.0       6.113672      15.766200       3.985920    1518.084307
std         0.0       3.484451       8.776805       1.995613     450.495201
```

```
min      2008.0       1.000000       1.000000       1.000000       1.000000
25%      2008.0       3.000000       8.000000       2.000000    1203.000000
50%      2008.0       6.000000      16.000000       4.000000    1545.000000
75%      2008.0       9.000000      23.000000       6.000000    1900.000000
max      2008.0      12.000000      31.000000       7.000000    2400.000000

          CRSDepTime        ArrTime     CRSArrTime      FlightNum  \
count  639131.000000  639131.000000  639131.000000  639131.000000
mean     1467.170564    1603.897769    1634.111117    2186.226299
std       424.680779     555.638756     464.415184    1946.620637
min         1.000000       0.000000       0.000000       1.000000
25%      1135.000000    1313.000000    1325.000000     610.000000
50%      1510.000000    1713.000000    1705.000000    1544.000000
75%      1815.000000    2030.000000    2014.000000    3422.000000
max      2359.000000    2400.000000    2400.000000    9742.000000

       ActualElapsedTime  …       DepDelay         TaxiIn        TaxiOut  \
count      639131.000000  …  639131.000000  639131.000000  639131.000000
mean          132.807546  …      43.184372       6.783043      18.234664
std            72.483393  …      53.299020       5.279566      14.377615
min             0.000000  …       6.000000       0.000000       0.000000
25%            80.000000  …      12.000000       4.000000      10.000000
50%           116.000000  …      24.000000       5.000000      14.000000
75%           165.000000  …      53.000000       8.000000      21.000000
max          1114.000000  …    2467.000000     240.000000     386.000000

          Cancelled       Diverted   CarrierDelay   WeatherDelay  \
count  639131.00000  639131.000000  639131.000000  639131.000000
mean        0.00030       0.004087      12.402227       2.357054
std         0.01733       0.063797      36.224780      17.088121
min         0.00000       0.000000       0.000000       0.000000
25%         0.00000       0.000000       0.000000       0.000000
50%         0.00000       0.000000       0.000000       0.000000
75%         0.00000       0.000000      10.000000       0.000000
max         1.00000       1.000000    2436.000000    1148.000000

           NASDelay  SecurityDelay  LateAircraftDelay
count  639131.000000  639131.000000      639131.000000
mean        9.644929       0.059797          16.299302
std        28.003472       1.684020          35.779226
min         0.000000       0.000000           0.000000
25%         0.000000       0.000000           0.000000
50%         0.000000       0.000000           0.000000
75%         6.000000       0.000000          18.000000
max      1207.000000     357.000000        1254.000000

[8 rows x 23 columns]
        Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  \
```

```
700956   2008      4          24          4   1204.0        1150   1328.0
876909   2008      5          26          1   2147.0        2130     20.0
64689    2008      1          28          1   1639.0        1556   1844.0
563728   2008      3           8          6    332.0        2005    610.0
1728398  2008     11          11          2   2152.0        2055     31.0

         CRSArrTime  FlightNum  ActualElapsedTime  …  DepDelay  TaxiIn  \
700956         1300       1456               84.0  …      14.0     7.0
876909           27        354              153.0  …      17.0     5.0
64689          1731       3802              125.0  …      43.0     8.0
563728         2253       1449              158.0  …     447.0    11.0
1728398        2221        442              159.0  …      57.0     7.0

         TaxiOut  Cancelled  Diverted  CarrierDelay  WeatherDelay  NASDelay  \
700956      26.0          0         0           0.0           0.0      14.0
876909      18.0          0         0           0.0           0.0       0.0
64689       36.0          0         0          73.0           0.0       0.0
563728      16.0          0         0           0.0           0.0     437.0
1728398     95.0          0         0           0.0           7.0      73.0

         SecurityDelay  LateAircraftDelay
700956             0.0               14.0
876909             0.0                0.0
64689              0.0                0.0
563728             0.0                0.0
1728398            0.0               50.0

[5 rows x 23 columns]
```

[66]:
```
print(y_train.describe())
print(y_train.head())
```

```
count    1.297627e+06
mean     7.653645e+02
std      5.744378e+02
min      1.100000e+01
25%      3.380000e+02
50%      6.060000e+02
75%      9.980000e+02
max      4.962000e+03
Name: Distance, dtype: float64
84384       920
1424846     946
1923886     853
383942      293
1811287    2611
Name: Distance, dtype: int64
```

[67]:
```
print(y_test.describe())
print(y_test.head())
```

```
count    639131.000000
mean        766.339311
std         574.564604
min          30.000000
25%         340.000000
50%         607.000000
75%         998.000000
max        4962.000000
Name: Distance, dtype: float64
700956      236
876909     1024
64689       532
563728     1024
1728398     429
Name: Distance, dtype: int64
```

[68]:
```
#Fem el recompte de files per una columna qualsevol ("Year"), i per␣
↪"UniqueCarrier"
print("Files x_train: " + str(x_train.Year.count()))
print("Files y_train: " + str(y_train.count()))
print("Files x_test: " + str(x_test.Year.count()))
print("Files y_test: " + str(y_test.count()))
a=y_train.count()/(y_train.count()+y_test.count())
print("Percentatge train: " + str(a*100))
print("Percentatge test: " + str(100-a*100))
print("Total Files: " + str(vols04.Year.count()))
```

```
Files x_train: 1297627
Files y_train: 1297627
Files x_test: 639131
Files y_test: 639131
Percentatge train: 66.99995559589789
Percentatge test: 33.00004440410211
Total Files: 1936758
```

També podem extreure similar informació amb *.shape*.

[69]:
```
print(x_train.shape, x_test.shape, y_train.shape, y_test.shape)
print(x_train.head())
print(y_train.head())
```

```
(1297627, 23) (639131, 23) (1297627,) (639131,)
         Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  \
84384    2008      1           7          1   1603.0        1555   1823.0
1424846  2008      8           6          3   1830.0        1735   2131.0
```

```
1923886  2008     12          25          4   2019.0        2005   2136.0
383942   2008      3          12          3   1653.0        1630   1753.0
1811287  2008     12          13          6   1857.0        1752   2239.0

         CRSArrTime  FlightNum  ActualElapsedTime  …  DepDelay  TaxiIn  \
84384          1841        777              140.0  …       8.0     6.0
1424846        2026        716              181.0  …      55.0    18.0
1923886        2138        487              137.0  …      14.0    10.0
383942         1735         47               60.0  …      23.0     2.0
1811287        2127        167              402.0  …      65.0     4.0

         TaxiOut  Cancelled  Diverted  CarrierDelay  WeatherDelay  NASDelay  \
84384       12.0          0         0           0.0           0.0       0.0
1424846     24.0          0         0           2.0           0.0      10.0
1923886      9.0          0         0           0.0           0.0       0.0
383942      10.0          0         0           0.0           0.0       0.0
1811287     17.0          0         0          65.0           0.0       7.0

         SecurityDelay  LateAircraftDelay
84384              0.0                0.0
1424846            0.0               53.0
1923886            0.0                0.0
383942             0.0               18.0
1811287            0.0                0.0

[5 rows x 23 columns]
84384       920
1424846     946
1923886     853
383942      293
1811287    2611
Name: Distance, dtype: int64
```

1.2 Exercici 2. Aplica algun procés de transformació (estandarditzar les dades numèriques, crear columnes dummies, polinomis...)

Observem les columnes que tenim i la informació que té el nostre dataframe.

[70]:
```
vols04.columns
```

[70]:
```
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')
```

[71]:
```
vols04.describe(include="all")
```

[71]:

	Unnamed: 0	Year	Month	DayofMonth	DayOfWeek
count	1.936758e+06	1936758.0	1.936758e+06	1.936758e+06	1.936758e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	3.341651e+06	2008.0	6.111106e+00	1.575347e+01	3.984827e+00
std	2.066065e+06	0.0	3.482546e+00	8.776272e+00	1.995966e+00
min	0.000000e+00	2008.0	1.000000e+00	1.000000e+00	1.000000e+00
25%	1.517452e+06	2008.0	3.000000e+00	8.000000e+00	2.000000e+00
50%	3.242558e+06	2008.0	6.000000e+00	1.600000e+01	4.000000e+00
75%	4.972467e+06	2008.0	9.000000e+00	2.300000e+01	6.000000e+00
max	7.009727e+06	2008.0	1.200000e+01	3.100000e+01	7.000000e+00

	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier
count	1.936758e+06	1.936758e+06	1.936758e+06	1.936758e+06	1936758
unique	NaN	NaN	NaN	NaN	20
top	NaN	NaN	NaN	NaN	WN
freq	NaN	NaN	NaN	NaN	377602
mean	1.518534e+03	1.467473e+03	1.604230e+03	1.634225e+03	NaN
std	4.504853e+02	4.247668e+02	5.557685e+02	4.646347e+02	NaN
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	NaN
25%	1.203000e+03	1.135000e+03	1.313000e+03	1.325000e+03	NaN
50%	1.545000e+03	1.510000e+03	1.714000e+03	1.705000e+03	NaN
75%	1.900000e+03	1.815000e+03	2.030000e+03	2.014000e+03	NaN
max	2.400000e+03	2.359000e+03	2.400000e+03	2.400000e+03	NaN

	…	TaxiIn	TaxiOut	Cancelled	CancellationCode
count	…	1.936758e+06	1.936758e+06	1.936758e+06	1936758
unique	…	NaN	NaN	NaN	4
top	…	NaN	NaN	NaN	N
freq	…	NaN	NaN	NaN	1936125
mean	…	6.787964e+00	1.822792e+01	3.268348e-04	NaN
std	…	5.280008e+00	1.433957e+01	1.807562e-02	NaN
min	…	0.000000e+00	0.000000e+00	0.000000e+00	NaN
25%	…	4.000000e+00	1.000000e+01	0.000000e+00	NaN
50%	…	5.000000e+00	1.400000e+01	0.000000e+00	NaN
75%	…	8.000000e+00	2.100000e+01	0.000000e+00	NaN
max	…	2.400000e+02	4.220000e+02	1.000000e+00	NaN

	Diverted	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay
count	1.936758e+06	1.936758e+06	1.936758e+06	1.936758e+06	1.936758e+06
unique	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN
mean	4.003598e-03	1.235367e+01	2.385512e+00	9.675607e+00	5.805836e-02

```
std     6.314722e-02  3.613493e+01  1.734036e+01  2.808958e+01  1.623934e+00
min     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
50%     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
75%     0.000000e+00  1.000000e+01  0.000000e+00  6.000000e+00  0.000000e+00
max     1.000000e+00  2.436000e+03  1.352000e+03  1.357000e+03  3.920000e+02

        LateAircraftDelay
count        1.936758e+06
unique                NaN
top                   NaN
freq                  NaN
mean         1.629374e+01
std          3.585904e+01
min          0.000000e+00
25%          0.000000e+00
50%          0.000000e+00
75%          1.800000e+01
max          1.316000e+03

[11 rows x 30 columns]
```

Apliquem Estandarització en dos atributs: "Distance" i "AirTime". Després comrpovem el valor de la desviació que sigui 1.

[72]:
```
ss=StandardScaler()
Xstd=ss.fit_transform(vols04[["Distance","AirTime"]].values)
print(Xstd[1:20])
```

```
[[ 0.07713737  0.07539358]
 [-0.43637094 -0.46191427]
 [-0.43637094 -0.44739244]
 [-0.13522877 -0.3021741 ]
 [ 1.43662887  1.77444812]
 [ 0.10847009 -0.02625926]
 [ 0.10847009 -0.01173742]
 [-1.0508402  -1.02826578]
 [ 1.25907684  1.52757695]
 [ 1.25907684  1.41140228]
 [ 0.12587715  0.03182808]
 [-0.94987925 -0.85400378]
 [-0.94987925 -0.88304745]
 [-0.94987925 -0.85400378]
 [-0.94987925 -0.88304745]
 [-0.94987925 -0.83948195]
 [ 0.56975721  0.51104859]
 [ 0.56975721  0.68531059]
 [ 0.35913177  0.38035209]]
```

[73]:
```
print("Mitjana: " + str(Xstd.mean()))
print("Desviació estàndard: " + str(Xstd.std()))
```

```
Mitjana: 2.3479823028301114e-19
Desviació estàndard: 0.9999999999999998
```

Convertim en Dummie una columna. Veiem que podria ser "UniqueCarrier" o "CancellationCode", amb 20 i 4 valors únics. Els visualitzem agrupats abans de prendre una decisió.

[74]:
```
#Llistem les dades úniques per a cada columna
a=vols04.groupby("UniqueCarrier")["UniqueCarrier"].nunique()
print(a)
b=vols04.groupby("CancellationCode")["CancellationCode"].nunique()
print(b)
```

```
UniqueCarrier
9E    1
AA    1
AQ    1
AS    1
B6    1
CO    1
DL    1
EV    1
F9    1
FL    1
HA    1
MQ    1
NW    1
OH    1
OO    1
UA    1
US    1
WN    1
XE    1
YV    1
Name: UniqueCarrier, dtype: int64
CancellationCode
A    1
B    1
C    1
N    1
Name: CancellationCode, dtype: int64
```

Per tal de no fer una taula molt gran, **normalitzem a dummie** tan sols CancellationCode.

[75]:
```
dummy_cancellation=pd.get_dummies(vols04["CancellationCode"])
print(dummy_cancellation.info())
print(dummy_cancellation.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 4 columns):
 #   Column  Dtype
---  ------  -----
 0   A       uint8
 1   B       uint8
 2   C       uint8
 3   N       uint8
dtypes: uint8(4)
memory usage: 7.4 MB
None
   A  B  C  N
0  0  0  0  1
1  0  0  0  1
2  0  0  0  1
3  0  0  0  1
4  0  0  0  1
```

[76]:
```
#Afegim les noves columnes al df.
vols05=vols04.merge(dummy_cancellation,left_index=True,right_index=True)
print(vols05.info())
print(vols05.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 34 columns):
 #   Column             Dtype
---  ------             -----
 0   Unnamed: 0         int64
 1   Year               int64
 2   Month              int64
 3   DayofMonth         int64
 4   DayOfWeek          int64
 5   DepTime            float64
 6   CRSDepTime         int64
 7   ArrTime            float64
 8   CRSArrTime         int64
 9   UniqueCarrier      object
 10  FlightNum          int64
 11  TailNum            object
 12  ActualElapsedTime  float64
 13  CRSElapsedTime     float64
 14  AirTime            float64
 15  ArrDelay           float64
 16  DepDelay           float64
 17  Origin             object
 18  Dest               object
```

```
 19  Distance           int64
 20  TaxiIn             float64
 21  TaxiOut            float64
 22  Cancelled          int64
 23  CancellationCode   object
 24  Diverted           int64
 25  CarrierDelay       float64
 26  WeatherDelay       float64
 27  NASDelay           float64
 28  SecurityDelay      float64
 29  LateAircraftDelay  float64
 30  A                  uint8
 31  B                  uint8
 32  C                  uint8
 33  N                  uint8
dtypes: float64(14), int64(11), object(5), uint8(4)
memory usage: 450.7+ MB
None
   Unnamed: 0  Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  \
0           0  2008      1           3          4   2003.0        1955
1           1  2008      1           3          4    754.0         735
2           2  2008      1           3          4    628.0         620
3           4  2008      1           3          4   1829.0        1755
4           5  2008      1           3          4   1940.0        1915

   ArrTime  CRSArrTime UniqueCarrier  …  Diverted CarrierDelay  \
0   2211.0        2225            WN  …         0          0.0
1   1002.0        1000            WN  …         0          0.0
2    804.0         750            WN  …         0          0.0
3   1959.0        1925            WN  …         0          2.0
4   2121.0        2110            WN  …         0          0.0

   WeatherDelay  NASDelay  SecurityDelay  LateAircraftDelay  A  B  C  N
0           0.0       0.0            0.0                0.0  0  0  0  1
1           0.0       0.0            0.0                0.0  0  0  0  1
2           0.0       0.0            0.0                0.0  0  0  0  1
3           0.0       0.0            0.0               32.0  0  0  0  1
4           0.0       0.0            0.0                0.0  0  0  0  1

[5 rows x 34 columns]
```

1.3 Exercici 3. Resumeix les noves columnes generades de manera estadística i gràfica

Comencem per la descripció estadística de les noves columnes "A", "B", "C" i "N" que provenen de "CancellationCode".

[77]:
```
#Informació estadística
vols06=vols05.iloc[:,-4:]
print(vols06.describe())
```

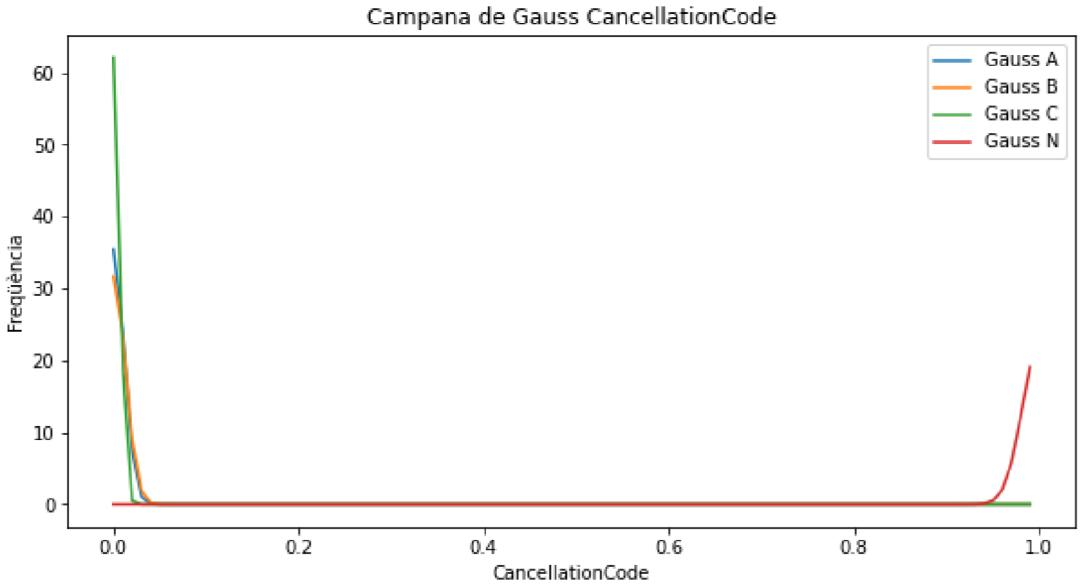
```
                A             B             C             N
count  1.936758e+06  1.936758e+06  1.936758e+06  1.936758e+06
mean   1.270164e-04  1.585123e-04  4.130614e-05  9.996732e-01
std    1.126944e-02  1.258917e-02  6.426854e-03  1.807562e-02
min    0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
25%    0.000000e+00  0.000000e+00  0.000000e+00  1.000000e+00
50%    0.000000e+00  0.000000e+00  0.000000e+00  1.000000e+00
75%    0.000000e+00  0.000000e+00  0.000000e+00  1.000000e+00
max    1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00
```

Visualitzem les dades gràficament. - Provem d'exposar-les en una Campana de Gauss.

[78]:
```
fig = plt.figure(figsize=(10,5))

sd_A=vols06["A"].std(ddof=1)
mean_A=vols06["A"].mean()
sd_B=vols06["B"].std(ddof=1)
mean_B=vols06["B"].mean()
sd_C=vols06["C"].std(ddof=1)
mean_C=vols06["C"].mean()
sd_N=vols06["N"].std(ddof=1)
mean_N=vols06["N"].mean()

#centrem la campana
maxim=vols06["A"].max()
print("Mx: " + str(maxim))
minim=vols06["A"].min()
print("Mn: " + str(minim))
meitat=maxim-(maxim-minim)/2
print("Meitat: " + str(meitat))

x_axis = np.arange(minim, maxim, 0.01)

ax1=plt.plot(x_axis, norm.pdf(x_axis, mean_A, sd_A),label="Gauss A")
ax2=plt.plot(x_axis, norm.pdf(x_axis, mean_B, sd_B),label="Gauss B")
ax3=plt.plot(x_axis, norm.pdf(x_axis, mean_C, sd_C),label="Gauss C")
ax4=plt.plot(x_axis, norm.pdf(x_axis, mean_N, sd_N),label="Gauss N")


plt.legend(loc="upper right")
plt.xlabel("CancellationCode")
plt.ylabel("Freqüència")
plt.title("Campana de Gauss CancellationCode")
```

```
plt.show()
```

```
Mx: 1
Mn: 0
Meitat: 0.5
```



Lògicament, la distribució de les dades es troba en valors de 0 i 1, tenint el codi de cancel · lació N la més gran freqüència (1).

[79]:
```
vols05.columns
```

[79]:
```
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', 'A',
       'B', 'C', 'N'],
      dtype='object')
```

[80]:
```
vols07 = vols05.groupby(["UniqueCarrier"],as_index=False).sum()
print(vols07)
print(vols07.info())
```

```
  UniqueCarrier     Unnamed: 0       Year     Month  DayofMonth  DayOfWeek  \
0            9E   170397455960  104185080    298844      799455     206043
1            AA   660814379301  385264920   1140719     3037390     761654
```

2	AQ	575980219	1506000	1059	9420	3264
3	AS	147184353425	78900344	251762	620656	156574
4	B6	206523247090	111072520	351710	884032	222842
5	CO	370422377133	201191560	627552	1592486	391816
6	DL	431524410816	229389904	752229	1802098	461910
7	EV	282876566605	164409016	511686	1285873	327395
8	F9	97269682486	56764152	175339	433999	115375
9	FL	251000877298	143138272	449552	1138477	291925
10	HA	32891833965	15039920	58146	124897	32190
11	MQ	476136413706	284975360	846195	2228599	549894
12	NW	262123968006	158848864	462861	1238320	314501
13	OH	166823805479	105735256	317356	822973	204157
14	OO	429324092956	265925464	810921	2085748	526834
15	UA	447837841451	283983408	836108	2197254	566733
16	US	328023899762	197637400	603853	1546359	392108
17	WN	1181175784119	758224816	2317794	5979985	1520711
18	XE	305701310959	208155304	596307	1627533	401693
19	YV	223341319823	134662504	425741	1055106	270027

	DepTime	CRSDepTime	ArrTime	CRSArrTime	…	Diverted	\
0	76292505.0	72848735	81787863.0	81542617	…	258	
1	293058658.0	282629998	309814772.0	320554945	…	909	
2	1068264.0	1049014	1172254.0	1184636	…	6	
3	58514212.0	56997753	61970536.0	62920684	…	272	
4	85935593.0	85674697	83241633.0	90766789	…	380	
5	150904920.0	146183960	161985024.0	169890739	…	426	
6	172800373.0	167774196	185957994.0	190098600	…	489	
7	122103677.0	116413005	131230168.0	129293422	…	86	
8	42560539.0	42205053	45671449.0	48341051	…	43	
9	113851192.0	111604430	117438846.0	120703214	…	308	
10	10870156.0	10558814	11812668.0	12042148	…	15	
11	208883975.0	199388073	224538293.0	221701314	…	593	
12	118390413.0	114566851	127171133.0	129532670	…	249	
13	80134355.0	76363743	85420366.0	85725010	…	192	
14	196135360.0	187329241	207186113.0	206757427	…	564	
15	213106974.0	205065095	222124577.0	228513440	…	475	
16	147433428.0	145990809	156121366.0	160436188	…	392	
17	590684393.0	572644345	620531890.0	630086305	…	1386	
18	157587169.0	150010334	167906991.0	170130085	…	470	
19	100716943.0	96841237	103920708.0	104876363	…	241	

	CarrierDelay	WeatherDelay	NASDelay	SecurityDelay	LateAircraftDelay	\
0	908509.0	149270.0	391601.0	1553.0	905944.0	
1	2821907.0	418677.0	2115126.0	6626.0	3334548.0	
2	8342.0	589.0	195.0	89.0	5350.0	
3	481815.0	47431.0	232822.0	7086.0	591343.0	
4	676126.0	55219.0	978731.0	2093.0	1293092.0	
5	996161.0	200282.0	1515500.0	11392.0	1254212.0	

```
6      1411220.0      167124.0  1216198.0         1365.0          1600158.0
7      1686899.0      565167.0   879783.0         3519.0           640078.0
8       248286.0       21017.0   265336.0          391.0           186316.0
9       438646.0       36745.0   743773.0            0.0          1769271.0
10      146051.0        4450.0      700.0          194.0            81932.0
11     1705271.0      417169.0  1458650.0         1172.0          2604077.0
12     1478247.0      309087.0   706658.0         4516.0           857781.0
13     1072800.0      671702.0   707604.0         1251.0           163662.0
14     1643741.0      297394.0  1546483.0        11091.0          2265110.0
15     1720150.0      214132.0  1466669.0         1661.0          3231417.0
16     1096887.0      101309.0   915826.0        10663.0          1374091.0
17     2261002.0      510665.0  1308052.0        25821.0          6611132.0
18     1218009.0      247486.0  1613685.0        13173.0          1962578.0
19     1906001.0      185245.0   675918.0         8789.0           824946.0

       A     B     C         N
0   27.0  24.0   7.0   51827.0
1   25.0  18.0   3.0  191819.0
2    0.0   0.0   0.0     750.0
3    4.0   7.0   0.0   39282.0
4    7.0   3.0   0.0   55305.0
5   26.0  11.0   1.0  100157.0
6   11.0   7.0   3.0  114217.0
7   19.0   9.0   1.0   81848.0
8    0.0   2.0   0.0   28267.0
9    3.0   4.0   0.0   71277.0
10   3.0   0.0   0.0    7487.0
11  13.0  58.0  33.0  141816.0
12   4.0   7.0   5.0   79092.0
13   3.0   8.0   1.0   52645.0
14  28.0  57.0   4.0  132344.0
15  31.0  16.0   0.0  141379.0
16  15.0   6.0   5.0   98399.0
17   5.0  10.0   0.0  377587.0
18   4.0  35.0   7.0  103617.0
19  18.0  25.0  10.0   67010.0

[20 rows x 30 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 30 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   UniqueCarrier      20 non-null     object
 1   Unnamed: 0         20 non-null     int64
 2   Year               20 non-null     int64
 3   Month              20 non-null     int64
 4   DayofMonth         20 non-null     int64
```

```
 5   DayOfWeek          20 non-null     int64
 6   DepTime            20 non-null     float64
 7   CRSDepTime         20 non-null     int64
 8   ArrTime            20 non-null     float64
 9   CRSArrTime         20 non-null     int64
 10  FlightNum          20 non-null     int64
 11  ActualElapsedTime  20 non-null     float64
 12  CRSElapsedTime     20 non-null     float64
 13  AirTime            20 non-null     float64
 14  ArrDelay           20 non-null     float64
 15  DepDelay           20 non-null     float64
 16  Distance           20 non-null     int64
 17  TaxiIn             20 non-null     float64
 18  TaxiOut            20 non-null     float64
 19  Cancelled          20 non-null     int64
 20  Diverted           20 non-null     int64
 21  CarrierDelay       20 non-null     float64
 22  WeatherDelay       20 non-null     float64
 23  NASDelay           20 non-null     float64
 24  SecurityDelay      20 non-null     float64
 25  LateAircraftDelay  20 non-null     float64
 26  A                  20 non-null     float64
 27  B                  20 non-null     float64
 28  C                  20 non-null     float64
 29  N                  20 non-null     float64
dtypes: float64(18), int64(11), object(1)
memory usage: 4.8+ KB
None
```

Com que "N" té molts més valors que "A","B" i "C", els normalitzem per tenir una visió percentual i a escala de tots ells en un gràfic.

[82]:
```
vols07["A_N"]=(vols07.A - vols07.A.min()) / ( vols07.A.max() - vols07.A.min())
vols07["B_N"]=(vols07.B - vols07.B.min()) / ( vols07.B.max() - vols07.B.min())
vols07["C_N"]=(vols07.C - vols07.C.min()) / ( vols07.C.max() - vols07.C.min())
vols07["N_N"]=(vols07.N - vols07.N.min()) / ( vols07.N.max() - vols07.N.min())
```

[83]:
```
print(vols07.info())
print(vols07.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 34 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   UniqueCarrier      20 non-null     object
 1   Unnamed: 0         20 non-null     int64
 2   Year               20 non-null     int64
```

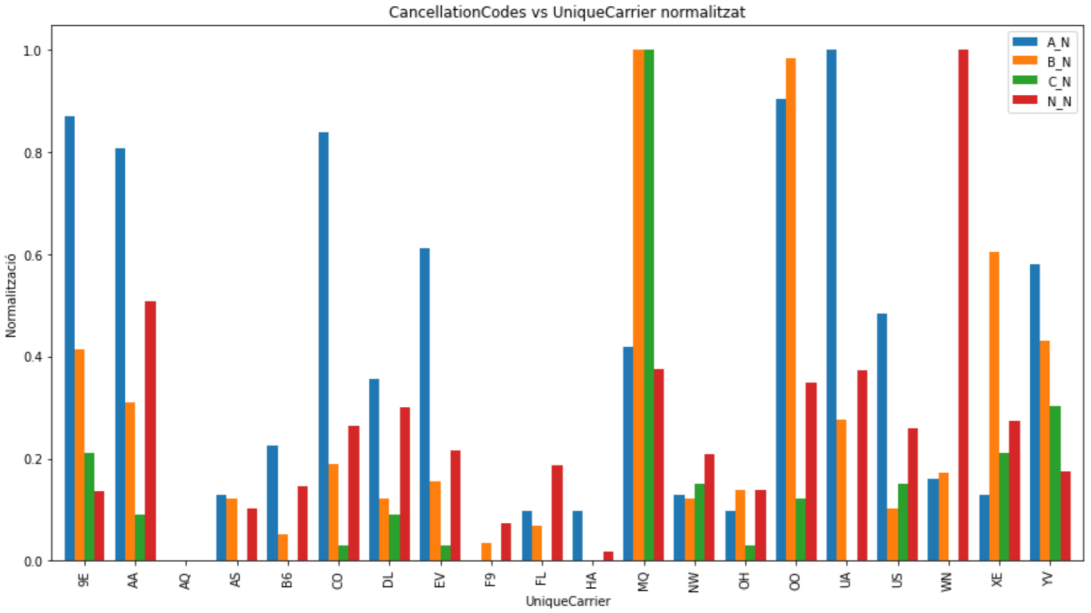
```
 3   Month              20 non-null     int64
 4   DayofMonth         20 non-null     int64
 5   DayOfWeek          20 non-null     int64
 6   DepTime            20 non-null     float64
 7   CRSDepTime         20 non-null     int64
 8   ArrTime            20 non-null     float64
 9   CRSArrTime         20 non-null     int64
 10  FlightNum          20 non-null     int64
 11  ActualElapsedTime  20 non-null     float64
 12  CRSElapsedTime     20 non-null     float64
 13  AirTime            20 non-null     float64
 14  ArrDelay           20 non-null     float64
 15  DepDelay           20 non-null     float64
 16  Distance           20 non-null     int64
 17  TaxiIn             20 non-null     float64
 18  TaxiOut            20 non-null     float64
 19  Cancelled          20 non-null     int64
 20  Diverted           20 non-null     int64
 21  CarrierDelay       20 non-null     float64
 22  WeatherDelay       20 non-null     float64
 23  NASDelay           20 non-null     float64
 24  SecurityDelay      20 non-null     float64
 25  LateAircraftDelay  20 non-null     float64
 26  A                  20 non-null     float64
 27  B                  20 non-null     float64
 28  C                  20 non-null     float64
 29  N                  20 non-null     float64
 30  A_N                20 non-null     float64
 31  B_N                20 non-null     float64
 32  C_N                20 non-null     float64
 33  N_N                20 non-null     float64
dtypes: float64(22), int64(11), object(1)
memory usage: 5.4+ KB
None
  UniqueCarrier    Unnamed: 0       Year    Month  DayofMonth  DayOfWeek  \
0            9E  170397455960  104185080   298844      799455     206043
1            AA  660814379301  385264920  1140719     3037390     761654
2            AQ     575980219    1506000     1059        9420       3264
3            AS  147184353425   78900344   251762      620656     156574
4            B6  206523247090  111072520   351710      884032     222842

       DepTime  CRSDepTime      ArrTime  CRSArrTime  …  SecurityDelay  \
0   76292505.0    72848735   81787863.0    81542617  …         1553.0
1  293058658.0   282629998  309814772.0   320554945  …         6626.0
2    1068264.0     1049014    1172254.0     1184636  …           89.0
3   58514212.0    56997753   61970536.0    62920684  …         7086.0
4   85935593.0    85674697   83241633.0    90766789  …         2093.0
```

```
   LateAircraftDelay     A     B    C         N       A_N       B_N       C_N  \
0           905944.0  27.0  24.0  7.0   51827.0  0.870968  0.413793  0.212121
1          3334548.0  25.0  18.0  3.0  191819.0  0.806452  0.310345  0.090909
2             5350.0   0.0   0.0  0.0     750.0  0.000000  0.000000  0.000000
3           591343.0   4.0   7.0  0.0   39282.0  0.129032  0.120690  0.000000
4          1293092.0   7.0   3.0  0.0   55305.0  0.225806  0.051724  0.000000

        N_N
0  0.135541
1  0.507034
2  0.000000
3  0.102251
4  0.144771

[5 rows x 34 columns]
```

[115]:
```
#Escollim les columnes que en interessen per fer el gràfic
vols08=vols07.iloc[:,[0,30,31,32,33]]
print(vols08.head())
```

```
  UniqueCarrier       A_N       B_N       C_N       N_N
0            9E  0.870968  0.413793  0.212121  0.135541
1            AA  0.806452  0.310345  0.090909  0.507034
2            AQ  0.000000  0.000000  0.000000  0.000000
3            AS  0.129032  0.120690  0.000000  0.102251
4            B6  0.225806  0.051724  0.000000  0.144771
```

[122]:
```
vols08.plot(x = "UniqueCarrier",kind = "bar",stacked = False,width=0.8,␣
↪mark_right = True,figsize=(15,8))

plt.title("CancellationCodes vs UniqueCarrier normalitzat",fontsize=12)
plt.xlabel("UniqueCarrier")
plt.ylabel("Normalització")


plt.show()
```



[125]:
```
vols08.describe(include="all")
```

[125]:

	UniqueCarrier	A_N	B_N	C_N	N_N
count	20	20.000000	20.000000	20.000000	20.000000
unique	20	NaN	NaN	NaN	NaN
top	9E	NaN	NaN	NaN	NaN
freq	1	NaN	NaN	NaN	NaN
mean	NaN	0.396774	0.264655	0.121212	0.254901
std	NaN	0.338988	0.292931	0.225699	0.216518
min	NaN	0.000000	0.000000	0.000000	0.000000
25%	NaN	0.120968	0.094828	0.000000	0.137169
50%	NaN	0.290323	0.146552	0.030303	0.211550
75%	NaN	0.661290	0.336207	0.151515	0.313129
max	NaN	1.000000	1.000000	1.000000	1.000000

[128]:
```
vols08[0:20]
```

[128]:

	UniqueCarrier	A_N	B_N	C_N	N_N
0	9E	0.870968	0.413793	0.212121	0.135541
1	AA	0.806452	0.310345	0.090909	0.507034
2	AQ	0.000000	0.000000	0.000000	0.000000
3	AS	0.129032	0.120690	0.000000	0.102251
4	B6	0.225806	0.051724	0.000000	0.144771
5	CO	0.838710	0.189655	0.030303	0.263793
6	DL	0.354839	0.120690	0.090909	0.301104
7	EV	0.612903	0.155172	0.030303	0.215207

8	F9	0.000000	0.034483	0.000000	0.073021
9	FL	0.096774	0.068966	0.000000	0.187155
10	HA	0.096774	0.000000	0.000000	0.017878
11	MQ	0.419355	1.000000	1.000000	0.374342
12	NW	0.129032	0.120690	0.151515	0.207894
13	OH	0.096774	0.137931	0.030303	0.137712
14	OO	0.903226	0.982759	0.121212	0.349207
15	UA	1.000000	0.275862	0.000000	0.373183
16	US	0.483871	0.103448	0.151515	0.259128
17	WN	0.161290	0.172414	0.000000	1.000000
18	XE	0.129032	0.603448	0.212121	0.272975
19	YV	0.580645	0.431034	0.303030	0.175832

[132]:
```
#prescindim de "N" perquè no és una cancel·lació
print(vols07.iloc[:,[0,26,27,28,30,31,32]])
```

	UniqueCarrier	A	B	C	A_N	B_N	C_N
0	9E	27.0	24.0	7.0	0.870968	0.413793	0.212121
1	AA	25.0	18.0	3.0	0.806452	0.310345	0.090909
2	AQ	0.0	0.0	0.0	0.000000	0.000000	0.000000
3	AS	4.0	7.0	0.0	0.129032	0.120690	0.000000
4	B6	7.0	3.0	0.0	0.225806	0.051724	0.000000
5	CO	26.0	11.0	1.0	0.838710	0.189655	0.030303
6	DL	11.0	7.0	3.0	0.354839	0.120690	0.090909
7	EV	19.0	9.0	1.0	0.612903	0.155172	0.030303
8	F9	0.0	2.0	0.0	0.000000	0.034483	0.000000
9	FL	3.0	4.0	0.0	0.096774	0.068966	0.000000
10	HA	3.0	0.0	0.0	0.096774	0.000000	0.000000
11	MQ	13.0	58.0	33.0	0.419355	1.000000	1.000000
12	NW	4.0	7.0	5.0	0.129032	0.120690	0.151515
13	OH	3.0	8.0	1.0	0.096774	0.137931	0.030303
14	OO	28.0	57.0	4.0	0.903226	0.982759	0.121212
15	UA	31.0	16.0	0.0	1.000000	0.275862	0.000000
16	US	15.0	6.0	5.0	0.483871	0.103448	0.151515
17	WN	5.0	10.0	0.0	0.161290	0.172414	0.000000
18	XE	4.0	35.0	7.0	0.129032	0.603448	0.212121
19	YV	18.0	25.0	10.0	0.580645	0.431034	0.303030

[142]:
```
print("Total cancel·lacions per A: " + str(vols07["A"].sum()))
print("Total cancel·lacions per B: " + str(vols07["B"].sum()))
print("Total cancel·lacions per C: " + str(vols07["C"].sum()))

#print (df.groupby(by=['Fruit','Date']).sum().groupby(level=[0]).cumsum())
```

```
Total cancel·lacions per A: 246.0
Total cancel·lacions per B: 307.0
Total cancel·lacions per C: 80.0
```

From the following website we can explain CancellationCodes:

https://www.kaggle.com/code/adveros/flight-delay-eda-exploratory-data-analysis/notebook
https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations

CancellationCode reason for cancellation (A = carrier, B = weather, C = NAS National Aviation System, D = security)

How are these categories defined?

- Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.
- Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

Conclusions: - Podem veure que el codi CancellationCode quan és "N", significa que no és va cancel · lar. - També observem que per raons de seguretat no n'hi hagut cap cancel · lació perquè "D" no ha aparegut. - Si busquem els 1 a cada un dels motius, podrem veure quina és la companyia que més cancel · la ordenada per motiu (pel A-Carrier és UA i OO, per B-weather és MQ i OO). Per C-NAS amb molta diferència respecte el segon és MQ; al gràfic, es pot veure com la barra de color verd és la més llarga, i que totes les altres són molt menors. - Les barres més altes que es repeteixen són les blaves A-Carrier. Podríem afirmar que és el motiu de cancel · lació més freqüent en totes les companyies, però fent la suma per atribut veim que és B, pel clima, amb 307 cancel · lacions.

[ ]: