Sprint 2 : S02 T05: Exploració de les dades

Nivell I

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 [215...
```
import pandas as pd
df= pd.read_csv(r"C:\users\hecto\OneDrive\Documentos\IT Data Science\Sprint2\Sprint2
df.head(3)
```

Out[215...

	Unnamed: 0	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTim
0	0	2008	1	3	4	2003.0	1955	2211.0	222
1	1	2008	1	3	4	754.0	735	1002.0	100
2	2	2008	1	3	4	628.0	620	804.0	75

3 rows × 30 columns

""" El Data Frame contiene los siguientes campos de información:

1. Unnamed: id
2. Year: 2008
3. Month: 1-12
4. DayofMonth: 1-31
5. DayOfWeek: 1 (Monday) - 7 (Sunday)
6. DepTime: actual departure time (local, hhmm)
7. CRSDepTime: scheduled departure time (local, hhmm)
8. ArrTime: actual arrival time (local, hhmm)
9. CRSArrTime: scheduled arrival time (local, hhmm)
10. UniqueCarrier: unique carrier code
11. FlightNum: flight number
12. TailNum: plane tail number: aircraft registration, unique aircraft identifier
13. ActualElapsedTime: in minutes
14. CRSElapsedTime: in minutes
15. AirTime: in minutes
16. ArrDelay arrival delay, in minutes: A flight is counted as "on time" if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS)
17. DepDelay: departure delay, in minutes
18. Origin: origin IATA airport code

19. Dest: destination IATA airport code
20. Distance: in miles
21. TaxiIn: taxi in time, in minutes
22. TaxiOut: taxi out time in minutes
23. Cancelled: *was the flight cancelled
24. CancellationCode: reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
25. Diverted: 1 = yes, 0 = no
26. CarrierDelay: in minutes. Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly
passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.
27. WeatherDelay: in minutes. Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.
28. NASDelay: in minutes. Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
29. SecurityDelay: in minutes. Security delay is 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.
30. LateAircraftDelay: in minutes. Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

"""

In [216…

```
# descripcion estadistica las variables numéricas (int o float) y objectos
# quito la columna "Unname"
df.drop('Unnamed: 0',axis=1,inplace=True)
df.describe(include="all").round(1)
```

Out[216…

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSAr
count	1936758.0	1936758.0	1936758.0	1936758.0	1936758.0	1936758.0	1929648.0	1936
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
mean	2008.0	6.1	15.8	4.0	1518.5	1467.5	1610.1	1
std	0.0	3.5	8.8	2.0	450.5	424.8	548.2	
min	2008.0	1.0	1.0	1.0	1.0	0.0	1.0	
25%	2008.0	3.0	8.0	2.0	1203.0	1135.0	1316.0	1

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSAr
50%	2008.0	6.0	16.0	4.0	1545.0	1510.0	1715.0	1
75%	2008.0	9.0	23.0	6.0	1900.0	1815.0	2030.0	2
max	2008.0	12.0	31.0	7.0	2400.0	2359.0	2400.0	2

11 rows × 29 columns

Exercici 2

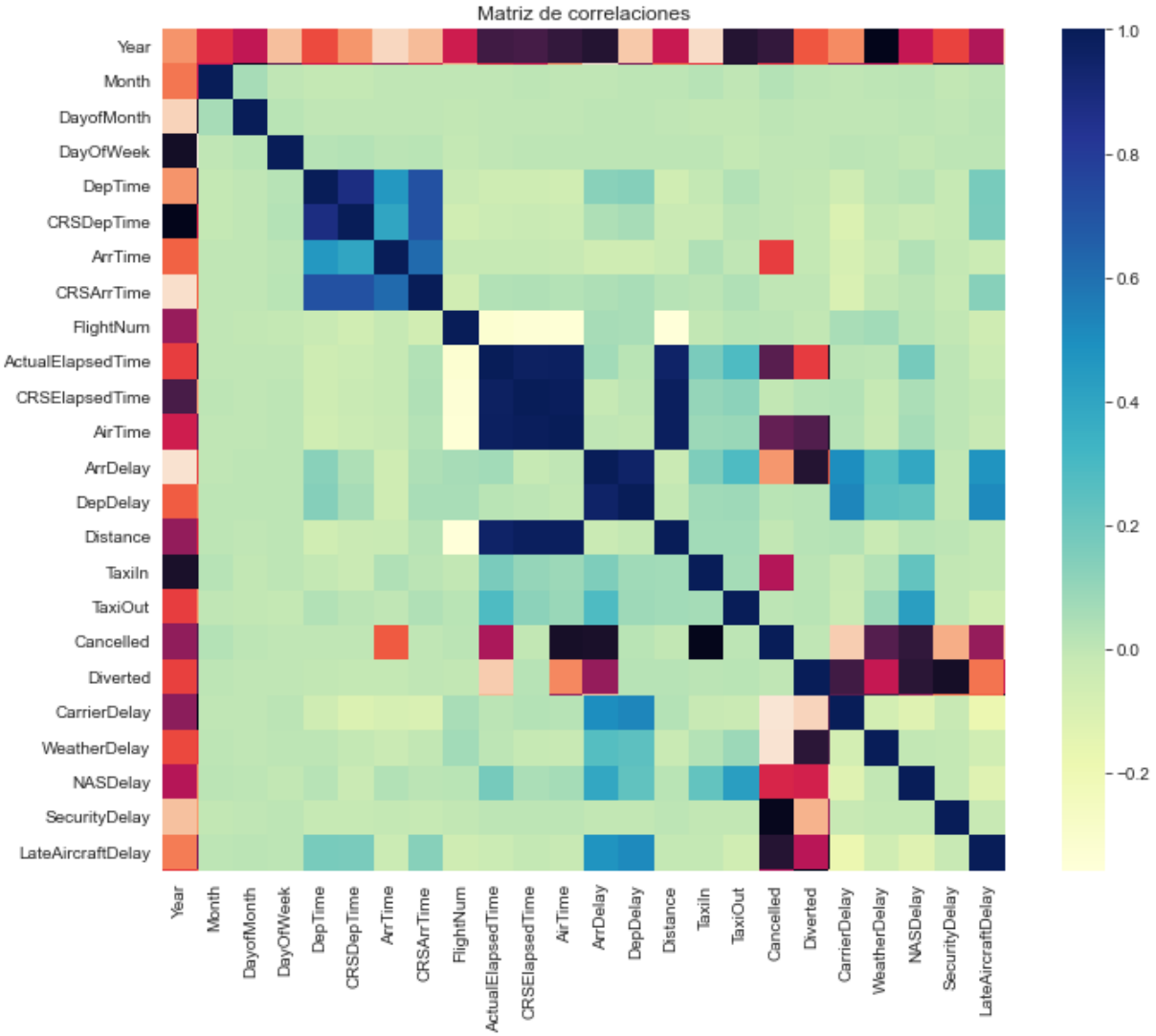
Fes un informe complet del data set:

1. Resumeix estadísticament les columnes d'interès

1.1 Matriz de correlaciones entre todas las variables del Data Frame

In [126…

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib.colors import LogNorm

# matriz de correlaciones
corrmat = df.corr()
#print(corrmat)

# dimensión de la tabla de colores definida a partir de una módulo aleatorio
Z = np.random.rand(30, 30)

#fig, ax0 = plt.subplots()
f, ax = plt.subplots(figsize=(12,9))
ax.pcolor(Z)

ax.set_title('Matriz de correlaciones')
#x,y = corrmat

#sns.heatmap(corrmat, vmax=10, square=True);
sns.heatmap(corrmat, cmap="YlGnBu", square = True)

plt.show()
```



""" Observaciones sobre los datos:

1. A través de la matriz de correlación se observa que algunas variables del conjunto de datos, presentan multicolinealidad, es decir, se pueden predecir linealmente a partir de otras.
2. Solo cuando el retraso en la llegada es superior a 15 minutos, se informa sobre la causa del retraso.
3. El retraso de llegada es: CarrierDelay + WeatherDelay + NASDelay + LateAircraftDelay.
4. En los casos de cancelación o desvío no hay datos relacionados con las causas del retraso.
5. La mayoría de ocasiones, los aeropuertos y los transportistas asignan un CRSElapsedTime superior al tiempo real empleado en las operaciones, Taxi In + Taxi out + Airtime (Tiempo transcurrido real). Por este motivo cuando los aviones despegan a tiempo, pueden aterrizar antes de la la hora prevista y les permite absorber retrasos por vuelos encadenados.

"""

1.2 Minutos de retraso en la llegada del avión

In [133…

```
print(np.round(df['ArrDelay'].describe()))
```

```
count    1928371.0
mean          42.0
std           57.0
min         -109.0
25%            9.0
50%           24.0
```

```
75%           56.0
max         2461.0
Name: ArrDelay, dtype: float64
```

1.3 Minutos de retraso en la salida del avión

In [137…

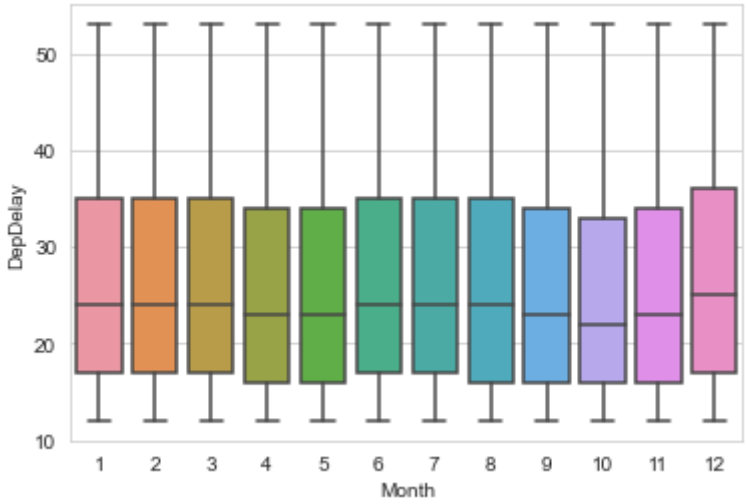
```
print(np.round(df['DepDelay'].describe()))
```

```
count    1936758.0
mean          43.0
std           53.0
min            6.0
25%           12.0
50%           24.0
75%           53.0
max         2467.0
Name: DepDelay, dtype: float64
```

1.4 Graficos de Retrasos en las Salidas (en minutos) por Meses

In [247…

```
# Se han eliminado los datos del quartile >0.75 porque desvirtuaban la escala del bo

df_no_outliers = df[df["DepDelay"].between(df["DepDelay"].quantile(.25), df["DepDela
plot1= sns.boxplot(data=df_no_outliers.sort_values(by="DepDelay",ascending = False),
plot1.figure.savefig("Dep_delay-Month_plot.png")
```



2. Troba quantes dades faltants hi ha per columna

In [230…

```
#tipo de datos y cuántos nulos hay
df.info(show_counts = True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 31 columns):
 #   Column             Non-Null Count    Dtype
---  ------             --------------    -----
 0   Year               1936758 non-null  int64
 1   Month              1936758 non-null  int64
 2   DayofMonth         1936758 non-null  int64
 3   DayOfWeek          1936758 non-null  int64
 4   DepTime            1936758 non-null  float64
 5   CRSDepTime         1936758 non-null  int64
 6   ArrTime            1929648 non-null  float64
 7   CRSArrTime         1936758 non-null  int64
 8   UniqueCarrier      1936758 non-null  object
```

```
 9   FlightNum          1936758 non-null  int64
 10  TailNum            1936753 non-null  object
 11  ActualElapsedTime  1928371 non-null  float64
 12  CRSElapsedTime     1936560 non-null  float64
 13  AirTime            1928371 non-null  float64
 14  ArrDelay           1928371 non-null  float64
 15  DepDelay           1936758 non-null  float64
 16  Origin             1936758 non-null  object
 17  Dest               1936758 non-null  object
 18  Distance           1936758 non-null  int64
 19  TaxiIn             1929648 non-null  float64
 20  TaxiOut            1936303 non-null  float64
 21  Cancelled          1936758 non-null  int64
 22  CancellationCode   1936758 non-null  object
 23  Diverted           1936758 non-null  int64
 24  CarrierDelay       1247488 non-null  float64
 25  WeatherDelay       1247488 non-null  float64
 26  NASDelay           1247488 non-null  float64
 27  SecurityDelay      1247488 non-null  float64
 28  LateAircraftDelay  1247488 non-null  float64
 29  TotalDelay         1928371 non-null  float64
 30  FlightCode         1936758 non-null  object
dtypes: float64(15), int64(10), object(6)
memory usage: 458.1+ MB
```

In [129…

```
# dónde están los valores nulos
df.isnull().sum()
```

Out[129…

```
Year                    0
Month                   0
DayofMonth              0
DayOfWeek               0
DepTime                 0
CRSDepTime              0
ArrTime              7110
CRSArrTime              0
UniqueCarrier           0
FlightNum               0
TailNum                 5
ActualElapsedTime    8387
CRSElapsedTime        198
AirTime              8387
ArrDelay             8387
DepDelay                0
Origin                  0
Dest                    0
Distance                0
TaxiIn               7110
TaxiOut               455
Cancelled               0
CancellationCode        0
Diverted                0
CarrierDelay       689270
WeatherDelay       689270
NASDelay           689270
SecurityDelay      689270
LateAircraftDelay  689270
dtype: int64
```

In [130…

```
# porcentaje de valores nulos / total
(df.isnull().sum()/len(df)*100).round(1)
```

```
Year                 0.0
```

Out[130…
```
Month                0.0
DayofMonth           0.0
DayOfWeek            0.0
DepTime              0.0
CRSDepTime           0.0
ArrTime              0.4
CRSArrTime           0.0
UniqueCarrier        0.0
FlightNum            0.0
TailNum              0.0
ActualElapsedTime    0.4
CRSElapsedTime       0.0
AirTime              0.4
ArrDelay             0.4
DepDelay             0.0
Origin               0.0
Dest                 0.0
Distance             0.0
TaxiIn               0.4
TaxiOut              0.0
Cancelled            0.0
CancellationCode     0.0
Diverted             0.0
CarrierDelay        35.6
WeatherDelay        35.6
NASDelay            35.6
SecurityDelay       35.6
LateAircraftDelay   35.6
dtype: float64
```

3. Crea columnes noves (velocitat mitjana del vol, si ha arribat tard o no...)

3.1 Columna de Velocidad Media del Vuelo

In [150…
```
# unidad de medida de la Velocidad Media (millas/horas)
df['Velocidad_Med']= df['Distance'].mean()*60/df['CRSElapsedTime'].mean()
df[:3]
```

Out[150…

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueC
0	2008	1	3	4	2003.0	1955	2211.0	2225	
1	2008	1	3	4	754.0	735	1002.0	1000	
2	2008	1	3	4	628.0	620	804.0	750	

3 rows × 30 columns

3.2 Columna de LLegada con Retraso o en Tiempo

In [159…
```
df['Llegada_retraso']=df.ArrDelay>=0
df[:5]
```

Out[159…

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueC
0	2008	1	3	4	2003.0	1955	2211.0	2225	
1	2008	1	3	4	754.0	735	1002.0	1000	
2	2008	1	3	4	628.0	620	804.0	750	

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueC
3	2008	1	3	4	1829.0	1755	1959.0	1925	
4	2008	1	3	4	1940.0	1915	2121.0	2110	

5 rows × 32 columns

4. Taula de les aerolínies amb més endarreriments acumulats

In [217…

```
df["TotalDelay"]=df['DepDelay']+df['ArrDelay']
df[:3]
```

Out[217…

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueC
0	2008	1	3	4	2003.0	1955	2211.0	2225	
1	2008	1	3	4	754.0	735	1002.0	1000	
2	2008	1	3	4	628.0	620	804.0	750	

3 rows × 30 columns

In [218…

```
retraso = df[['UniqueCarrier','TotalDelay']].copy()
retraso_ok = retraso.dropna().sort_values("TotalDelay", ascending=False)
retraso_ok = retraso_ok.groupby('UniqueCarrier').aggregate(sum)
retraso_ok.sort_values("TotalDelay", ascending=False)
```

Out[218…

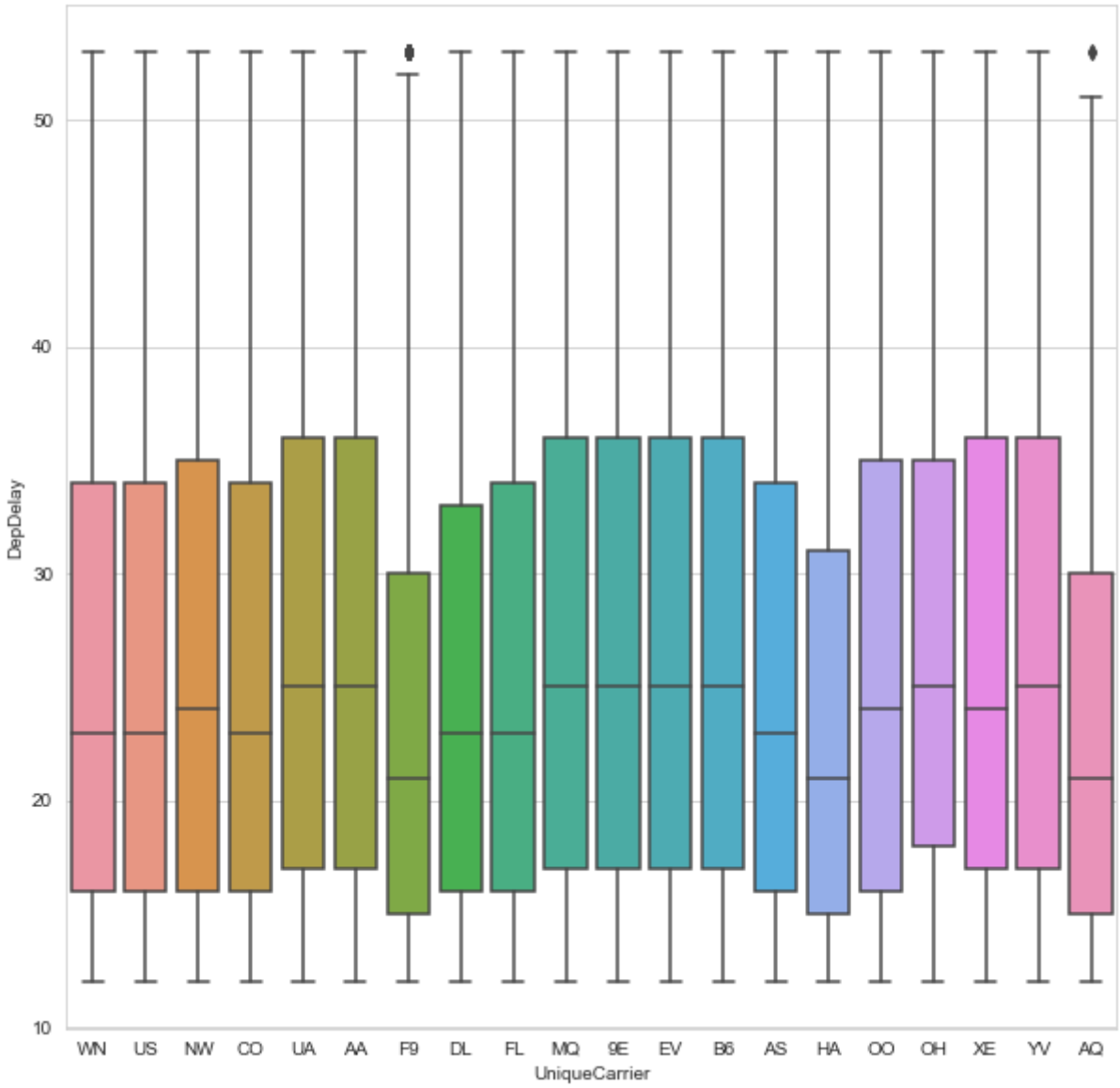
UniqueCarrier	TotalDelay
WN	24331347.0
AA	17746439.0
UA	13764664.0
MQ	12554319.0
OO	11869335.0
XE	10329576.0
DL	8971757.0
CO	8340506.0
EV	7834335.0
YV	7387293.0
US	7370623.0
NW	6715503.0
FL	6115528.0
B6	6043070.0
OH	5241678.0

UniqueCarrier	TotalDelay
9E	4862296.0
AS	2888170.0
F9	1569572.0
HA	502618.0
AQ	35176.0

4.2 Grafico de los Retrasos en las Salidas por Compañías Aéreas

In [249…

```
df_no_outliers = df[df["DepDelay"].between(df["DepDelay"].quantile(.25), df["DepDela
plt.figure(figsize=(10,10))
plot1= sns.boxplot(data=df_no_outliers.sort_values(by="DepDelay",ascending=True), x=
plot1.figure.savefig("Dep_delay_UniqueCarrier_plot.png")
```



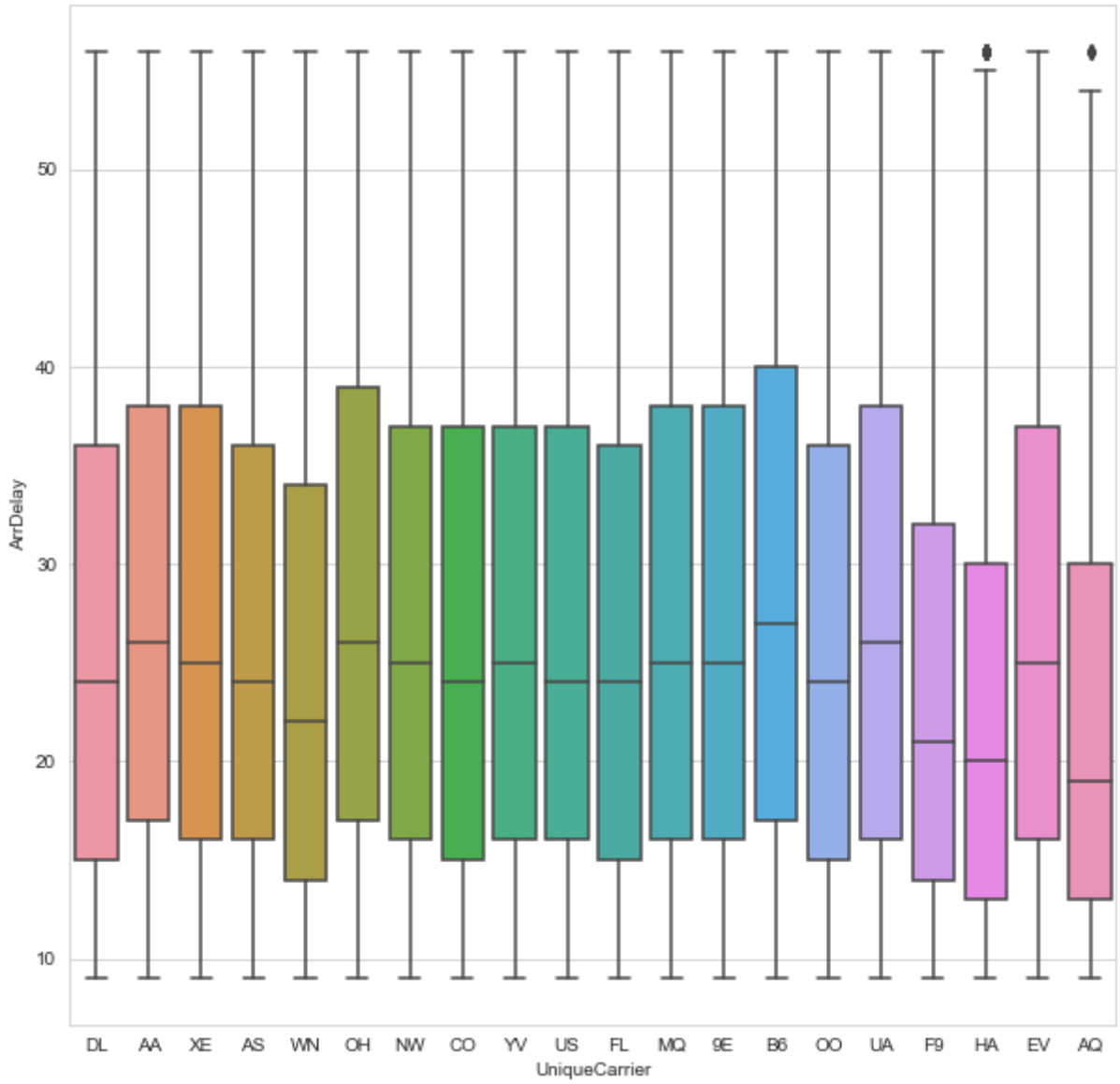
Las aerolínias con menores retrasos son: F9, HA y AQ

4.3 Grafico de los Retrasos en las Llegadas por Compañías Aéreas

In [250…

```
df_no_outliers = df[df["ArrDelay"].between(df["ArrDelay"].quantile(.25), df["ArrDela
plt.figure(figsize=(10,10))
```

```
plot1= sns.boxplot(data=df_no_outliers.sort_values(by="ArrDelay",ascending=True), x=
plot1.figure.savefig("Arr_delay_UniqueCarrier_plot.png")
```



Las aerolíneas con menores retrasos en las llegadas son: HA, AQ y F9

5. Quins són els vols més llargs? I els més endarrerits?

5.1 Más tiempo en Aire

In [221…

```
# Creamos un Data Frame para clasificar los códigos de los vuelos en un mismo campo
df["FlightCode"] = df["UniqueCarrier"].astype(str) + "/" + df["FlightNum"].astype(st
vuelosLargos = pd.DataFrame(df, columns=["AirTime","FlightCode"]).copy()
vuelosLargos_ok = vuelosLargos.dropna()
vuelosLargos_ok[:5]
```

Out[221…

	AirTime	FlightCode
0	116.0	WN/335
1	113.0	WN/3231
2	76.0	WN/448
3	77.0	WN/3920
4	87.0	WN/378

In [222...
```
##### Ordenamos de forma ascendente los Vuelos más Largos en función de "AirTime"
vuelosLargos_ok.sort_values("AirTime", ascending=False).head()
```

Out[222...

	AirTime	FlightCode
1488690	1091.0	HA/21
1367047	733.0	HA/28
362529	664.0	CO/15
556381	655.0	CO/15
556385	654.0	CO/15

5.2 Vuelos más Largos

In [223...
```
df["FlightCode"] = df["UniqueCarrier"].astype(str) + "/" + df["FlightNum"].astype(st
vuelosLargos = pd.DataFrame(df, columns=["ActualElapsedTime","FlightCode"]).copy()
vuelosLargos_ok = vuelosLargos.dropna()
vuelosLargos_ok[:5]
```

Out[223...

	ActualElapsedTime	FlightCode
0	128.0	WN/335
1	128.0	WN/3231
2	96.0	WN/448
3	90.0	WN/3920
4	101.0	WN/378

In [224...
```
vuelosLargos_ok.sort_values("ActualElapsedTime", ascending=False).head()
```

Out[224...

	ActualElapsedTime	FlightCode
1488690	1114.0	HA/21
1926817	790.0	CO/15
1173580	776.0	DL/151
1418032	750.0	CO/15
1367047	750.0	HA/28

5.3 Vuelos más retrasados

In [225...
```
retrasados = pd.DataFrame(df, columns=["ArrDelay","FlightCode"]).copy()
retrasados_ok = Retrasados.dropna()
retrasados_ok.sort_values('ArrDelay', ascending=False).head()
```

Out[225...

	ArrDelay	FlightCode
322516	2461.0	NW/808
686014	2453.0	NW/1699
839306	1951.0	NW/1107

	ArrDelay	FlightCode
1009553	1707.0	MQ/3538
1881639	1655.0	NW/357

5.4 Los 5 aeropuertos con más retraso en la salida

In [226...
```
airDelays = df[["DepDelay","Origin"]].copy()
airportDelays = airDelays.groupby("Origin").aggregate(sum)
airportDelays.sort_values("DepDelay", ascending=False)[:5]
```

Out[226...

Origin	DepDelay
ORD	6365866.0
ATL	5382082.0
DFW	3658231.0
DEN	2801893.0
EWR	2669013.0

Exercici 3

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

In [227...
```
path = "C:\\Users\hecto\\Downloads\\DelayedFlightsVams.csv"
df.to_csv(path, index = False)
```

In [251...
```
df.info(show_counts = True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1936758 entries, 0 to 1936757
Data columns (total 31 columns):
 #   Column             Non-Null Count    Dtype
---  ------             --------------    -----
 0   Year               1936758 non-null  int64
 1   Month              1936758 non-null  int64
 2   DayofMonth         1936758 non-null  int64
 3   DayOfWeek          1936758 non-null  int64
 4   DepTime            1936758 non-null  float64
 5   CRSDepTime         1936758 non-null  int64
 6   ArrTime            1929648 non-null  float64
 7   CRSArrTime         1936758 non-null  int64
 8   UniqueCarrier      1936758 non-null  object
 9   FlightNum          1936758 non-null  int64
 10  TailNum            1936753 non-null  object
 11  ActualElapsedTime  1928371 non-null  float64
 12  CRSElapsedTime     1936560 non-null  float64
 13  AirTime            1928371 non-null  float64
 14  ArrDelay           1928371 non-null  float64
 15  DepDelay           1936758 non-null  float64
 16  Origin             1936758 non-null  object
 17  Dest               1936758 non-null  object
 18  Distance           1936758 non-null  int64
 19  TaxiIn             1929648 non-null  float64
```

```
 20  TaxiOut            1936303 non-null  float64
 21  Cancelled          1936758 non-null  int64
 22  CancellationCode   1936758 non-null  object
 23  Diverted           1936758 non-null  int64
 24  CarrierDelay       1247488 non-null  float64
 25  WeatherDelay       1247488 non-null  float64
 26  NASDelay           1247488 non-null  float64
 27  SecurityDelay      1247488 non-null  float64
 28  LateAircraftDelay  1247488 non-null  float64
 29  TotalDelay         1928371 non-null  float64
 30  FlightCode         1936758 non-null  object
dtypes: float64(15), int64(10), object(6)
memory usage: 458.1+ MB
```