Nivell 1

Exercici 1 y 2:

Resumeix gràficament el data set DelayedFlights.csv

In [7]:
```
import pandas as pd
df= pd.read_csv(r"C:\users\hecto\OneDrive\Documentos\IT Data Science\Sprint2\Sprint2
df.head(3)
```

Out[7]:

	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

Campos de información para el ejercicio 1:

9. UniqueCarrier: unique carrier code

10. FlightNum: flight number

11. TailNum: plane tail number: aircraft registration, unique aircraft identifier

12. ActualElapsedTime: in minutes

13. CRSElapsedTime: in minutes

14. AirTime: in minutes

15. 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 CRS

16. DepDelay: departure delay, in minutes

17. Origin: origin IATA airport code

18. Dest: destination IATA airport code

19. Distance: in miles

Crea Almenys una visualització per:

1.1 Una variable categòrica (UniqueCarrier)

In [98]:
```
# Datos de la variable "UniqueCarrier" : número de registros y etiquetas de la varia
df["UniqueCarrier"]
```

Out[98]:
```
0          WN
1          WN
2          WN
3          WN
4          WN
           ..
1936753    DL
1936754    DL
```

```
1936755    DL
1936756    DL
1936757    DL
Name: UniqueCarrier, Length: 1936758, dtype: object
```

In [100…

```
# Extraemos los vuelos de las 10 compñias principales y el resto de Cías. los agrupa
lineasAereas10 =df["UniqueCarrier"].value_counts()[:10]
sumaLa10 =sum(lineasAereas10)
print ("Total de vuelos de las 10 Líneas Aéreas principales:", suma)
restoLa = 1936757-sumaLa10
print("Total de vuelos del (Resto) de Líeas Aéreas:", restoLa)
lineasAereas10
lineasAereas10.loc["Resto"]=453113
totalVuelos = sum(lineasAereas10)
print ("Total de vuelos del Data Frame:", totalVuelos)
print(lineasAereas10)
print("% de vuelos realizados por las 10 Líneas Aéreas principales", round(sumaLa10/
```

```
Total de vuelos de las 10 Líneas Aéreas principales: 1483644
Total de vuelos del (Resto) de Líeas Aéreas: 453113
Total de vuelos del Data Frame: 1936757
WN       377602
AA       191865
MQ       141920
UA       141426
OO       132433
DL       114238
XE       103663
CO       100195
US        98425
EV        81877
Resto    453113
Name: UniqueCarrier, dtype: int64
% de vuelos realizados por las 10 Líneas Aéreas principales 0.766
```

In [101…

```
print(type(lineasAereas10))
```

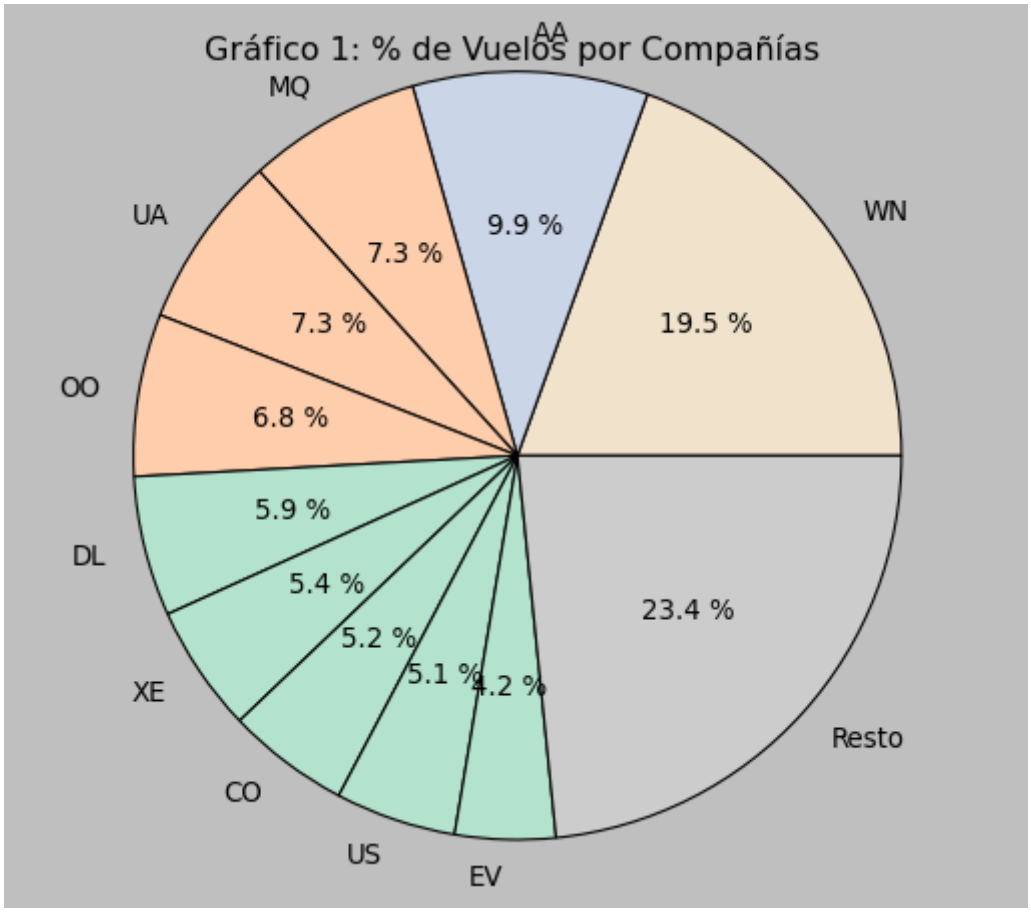
```
<class 'pandas.core.series.Series'>
```

Las 10 primeras Líneas Aéreas concentran el 76,6% de los vuelos del Data Frame

In [240…

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

nombres = ["WN","AA","MQ","UA","OO","DL","XE","CO","US","EV","Resto"]

normdata = colors.Normalize(min(lineasAereas10), max(lineasAereas10))
print(normdata)
colormap = cm.get_cmap("Pastel2")
colores =colormap(normdata(lineasAereas10))
plt.figure(figsize=(8,6))
plt.title("Gráfico 1: % de Vuelos por Compañías ")
plt.pie(lineasAereas10, labels = nombres, autopct="%0.1f %%", colors = colores)
plt.axis("equal")

# Ejercicio 2: Explortar la imagen del Gráfico 1 en formato *.png
plt.savefig("Grafico_1_%_Vuelos_Compañias.png")
plt.show()
```

```
<matplotlib.colors.Normalize object at 0x0000020016638F10>
```

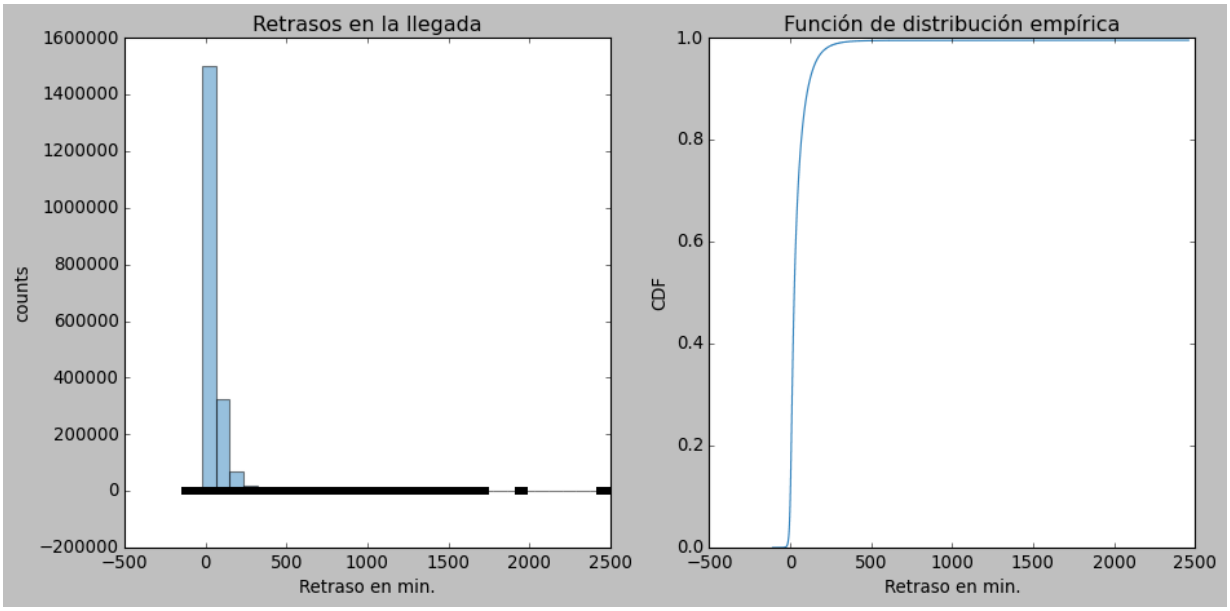


La compaía que más porcentaje de vuelos realiza es : WN con un 19,5% del total de vuelos

1.2 Una variable numérica (ArrDelay)

In [222…

```
# Gráficos distribución observada (empírica): variable ArrDelay

from scipy import stats
import inspect
from statsmodels.distributions.empirical_distribution import ECDF

fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(12, 6))
data = df["ArrDelay"]
# Histograma
axs[0].hist(x=data, bins=30, color="#3182bd", alpha=0.5)
axs[0].plot(data, np.full_like(data, -0.01), '|k', markeredgewidth=9)
axs[0].set_title('Retrasos en la llegada')
axs[0].set_xlabel('Retraso en min.')
axs[0].set_ylabel('counts')

# Función de Distribución Acumulada
# ecdf (empirical cumulative distribution function)
ecdf = ECDF(x=data)
axs[1].plot(ecdf.x, ecdf.y, color="#3182bd")
axs[1].set_title('Función de distribución empírica')
axs[1].set_xlabel('Retraso en min.')
axs[1].set_ylabel('CDF')

plt.savefig("Grafico_2_HistogramaArrDelay_FDE.png")
plt.tight_layout();
```



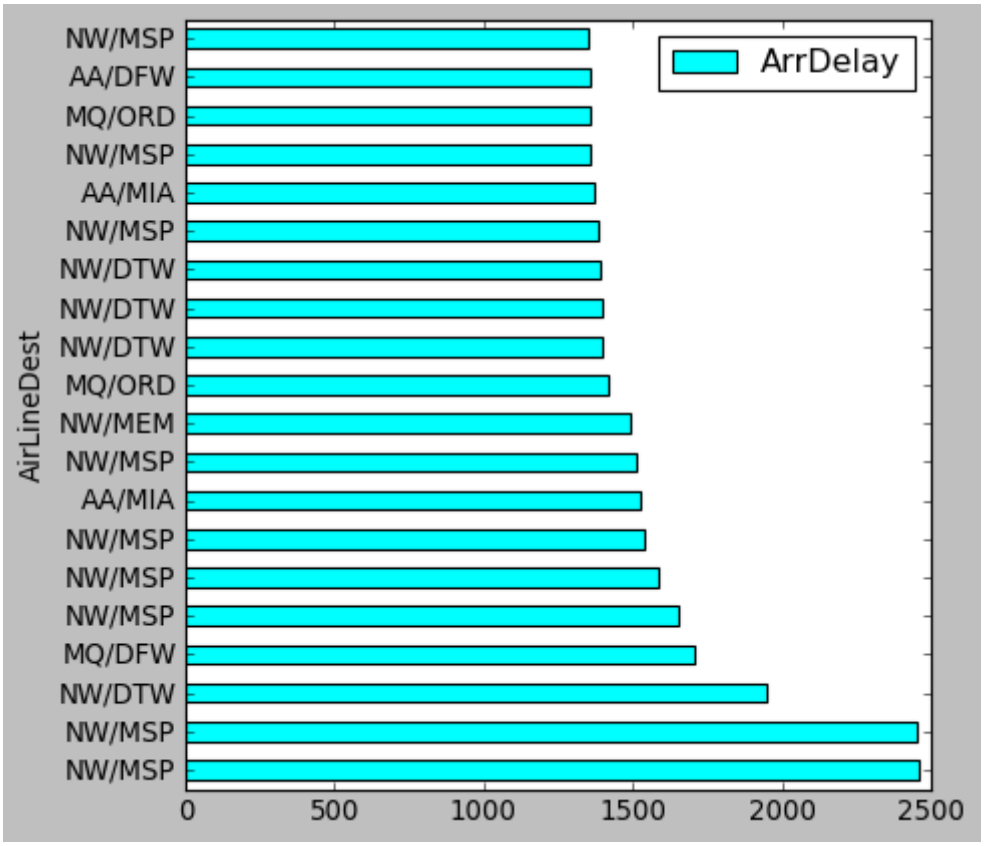
In [169…

```
# Creamos una variable que agrupa la Aerolinia + Aeropuerto de Destino
df["AirLineDest"]= df["UniqueCarrier"].astype(str) + "/" + df["Dest"].astype(str)

llegadaRetraso10 =pd.DataFrame(df, columns=["ArrDelay", "AirLineDest"]).copy()
llegadaRetraso10 = llegadaRetraso10.dropna()
dataRetraso = llegadaRetraso10.sort_values('ArrDelay', ascending=False)[:20]
dataRetraso.set_index("AirLineDest",inplace=True)
print(dataRetraso)
```

```
             ArrDelay
AirLineDest
NW/MSP         2461.0
NW/MSP         2453.0
NW/DTW         1951.0
MQ/DFW         1707.0
NW/MSP         1655.0
NW/MSP         1583.0
NW/MSP         1542.0
AA/MIA         1525.0
NW/MSP         1510.0
NW/MEM         1490.0
MQ/ORD         1417.0
NW/DTW         1395.0
NW/DTW         1395.0
NW/DTW         1392.0
NW/MSP         1382.0
AA/MIA         1370.0
NW/MSP         1359.0
MQ/ORD         1357.0
AA/DFW         1357.0
NW/MSP         1350.0
```

In [246…

```
# Gráfico de Barras de las 20 Aerolineas con mayores retrasos y destinos de los vuel
dataRetraso.plot(kind = 'barh', figsize=(6,6),color= "cyan")

# Ejercicio 2: Exportar la imagen 2 en formato *.png
plt.savefig("Grafico_3_AirLineEst_ArrDelay.png")
```



Lac compañía que más retasos tiene es la NW, con destino a MSP (8 ocasiones de las 20 destacadas)

1.3 Una variable numèrica i una categòrica (ArrDelay i UniqueCarrier)

In [214…

```
# Gráficos de la distribución de los retrasos en las llegadas por Aerolíneas

import seaborn as sns
import plotly.express as px

df_no_outliers = df[df["ArrDelay"].between(df["ArrDelay"].quantile(.25), df["ArrDela
fig = px.box(df_no_outliers, x="UniqueCarrier", y="ArrDelay", color= "UniqueCarrier"
fig.show()
fig.write_html("Grafico4.html")
```

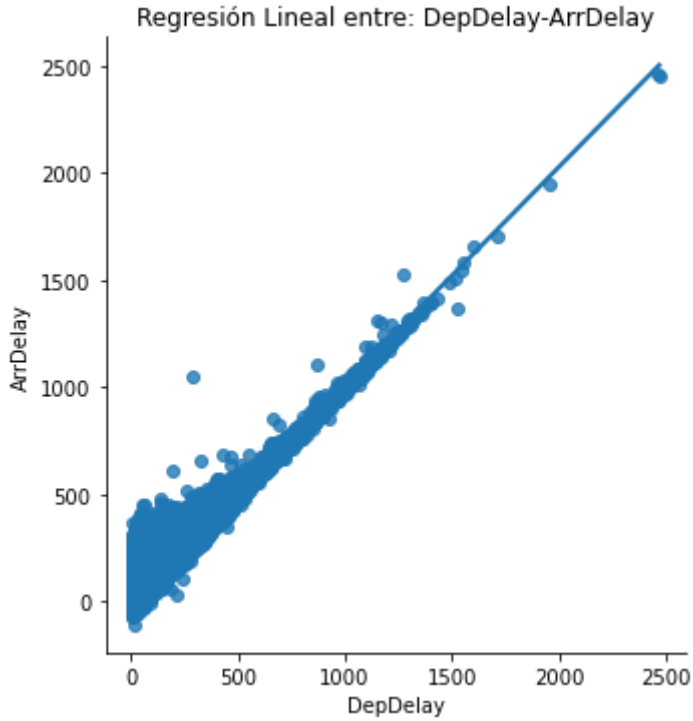
La compañia que acumula menos retrasos en las llegadas al destino en un 75% de sus vuelos es AQ

La compañía que aumula más retrasos en las llegadas al destino en un 75% de sus vuelos es B6

1.4 Dues variables numèriques (ArrDelay i DepDelay)

Salidas y Llegadas con Retrasos por aeropuerto de origen (en min.)

In [16]:
```
import seaborn as sns
fig = sns.lmplot(x='DepDelay',y='ArrDelay',data=df).set(title="Regresión Lineal entr

#Ejercicio 2: Exportar el gráfico
fig.savefig("Grafico5_RegLin.png")
```



Exíste una evidente correlación entre las variables, de forma que cuanto mayor es el retraso en la salidas mayor es la probabilidad de que el avión llegue con retraso al aeropuerto de destino.

1.5 Tres variables (ArrDelay, DepDelay i UniqueCarrier)

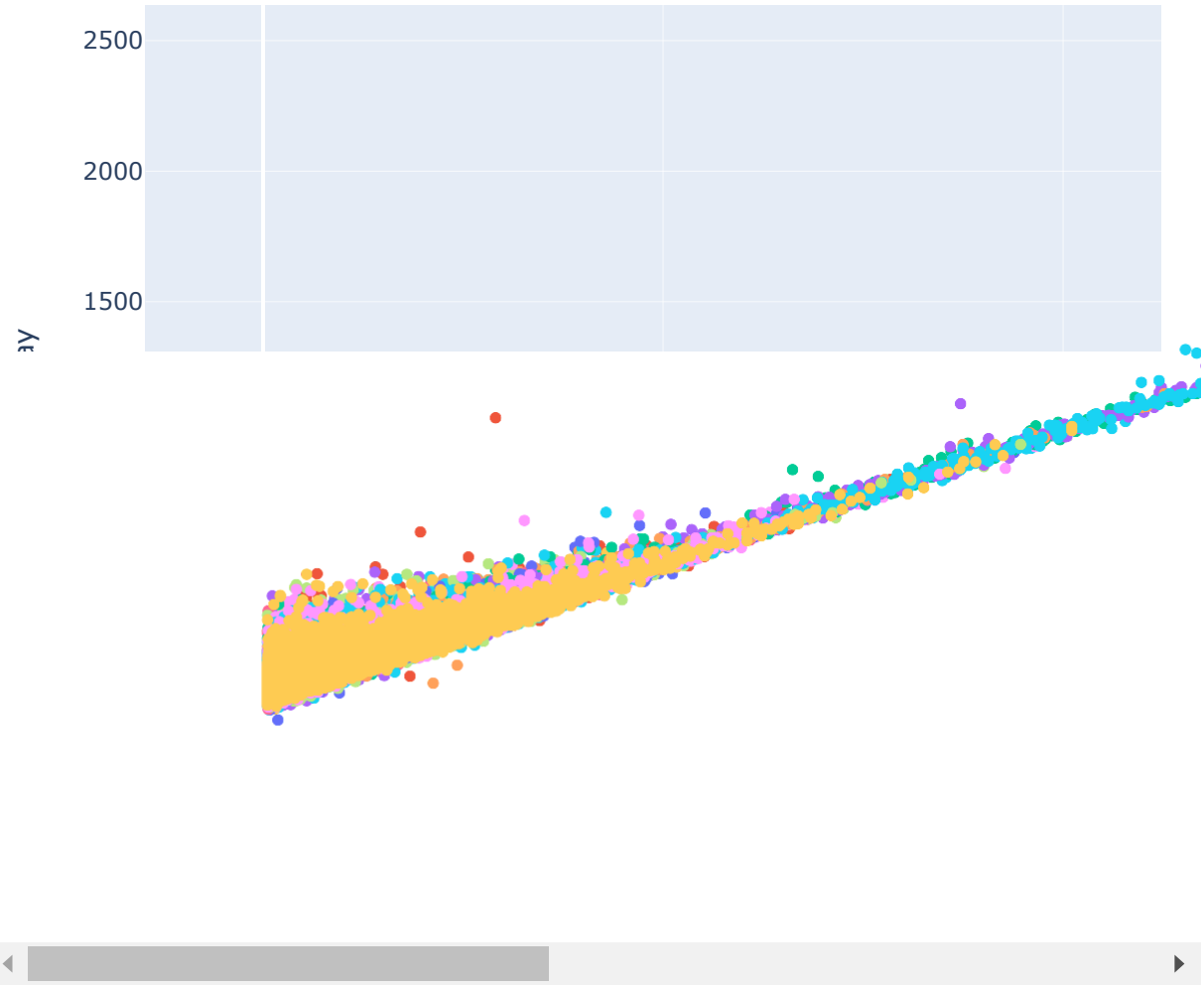
In [23]:
```
df3 = pd.DataFrame(df, columns=["UniqueCarrier","ArrDelay", "DepDelay"])
df3.head()
```

Out[23]:

	UniqueCarrier	ArrDelay	DepDelay
0	WN	-14.0	8.0
1	WN	2.0	19.0
2	WN	14.0	8.0
3	WN	34.0	34.0

	UniqueCarrier	ArrDelay	DepDelay
4	WN	11.0	25.0

In [40]:
```
fig = px.scatter(df3, x="DepDelay", y="ArrDelay", color="UniqueCarrier")
fig.show()
```



In [ ]:
```
# Ejercicio 2: Exportar el grafico en formato HTML
fig.write_html("Grafico_51_3Variables.html")
```

1.6 Més de tres variables (ArrDelay, DepDelay, AirTime i UniqueCarrier)

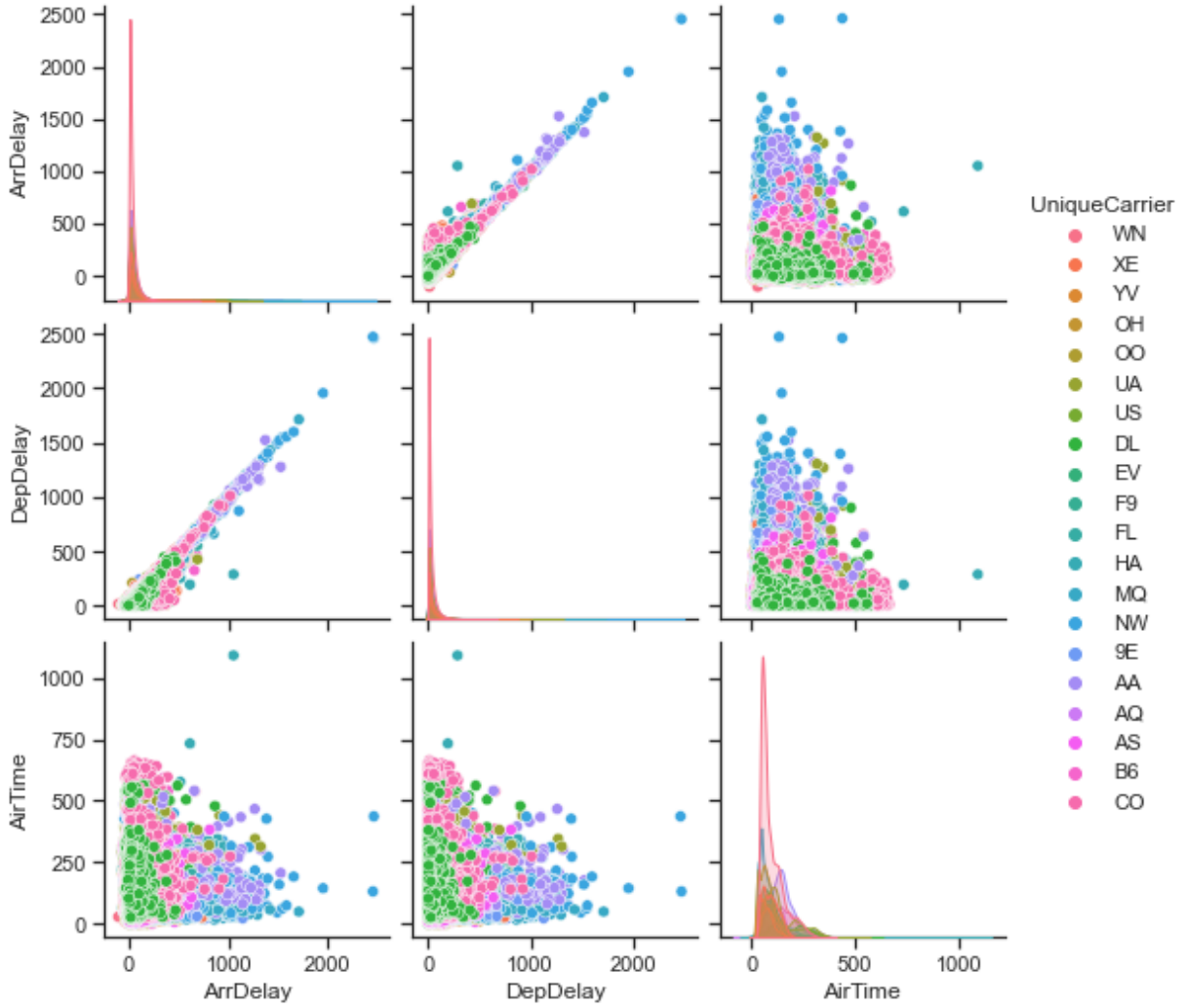
In [19]:
```
df4 = pd.DataFrame(df, columns=["UniqueCarrier","ArrDelay", "DepDelay", "AirTime"])
df4.head()
```

Out[19]:

	UniqueCarrier	ArrDelay	DepDelay	AirTime
0	WN	-14.0	8.0	116.0
1	WN	2.0	19.0	113.0
2	WN	14.0	8.0	76.0
3	WN	34.0	34.0	77.0

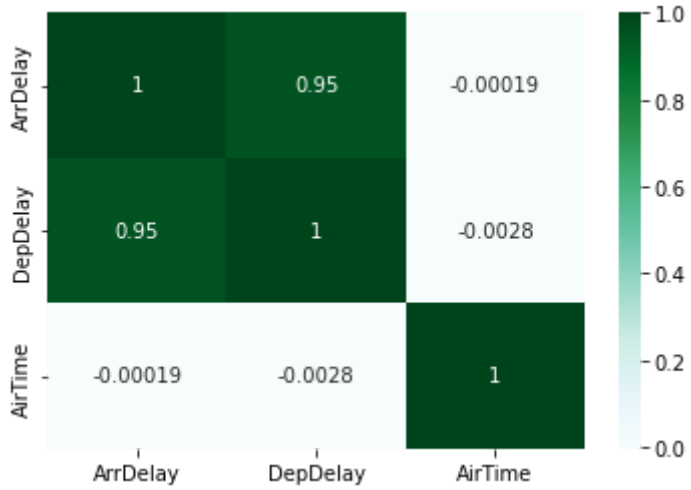
	UniqueCarrier	ArrDelay	DepDelay	AirTime
4	WN	11.0	25.0	87.0

In [21]:
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="ticks", color_codes=True)
fig = sns.pairplot(df4, hue="UniqueCarrier")

# Ejercicio 2: Exportamos las salidas de los graficos de correlación de las 4 variab
plt.savefig("Grafico6_4Variables.png")

plt.show()
```



In [ ]:
```
##### En el gráfico se obseva que existe una fuerte correlación netre las variables
##### Sin embargo no parece existir una fuerte correlación entre las salidas y llega
```

In [23]:
```
# Correlaciones entre las variables ArrDelay, DepDelay y AirTime
import matplotlib.pyplot as plt
fig= sns.heatmap(df4.corr(),cmap='BuGn',annot=True);
plt.savefig("Grafico7_CoeficientesCorr.png")
```



Ejercicio 3

Incorporado en el repositorio en el archivo "Sprint3Ex3.ipnb"

In [ ]: