# 1 Exercici 1

Resumeix gràficament el data set DelayedFlights.csv

In [1]:

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [2]:

```
df = pd.read_csv("DelayedFlightsNet.csv", index_col = 0)
```

```
C:\Users\Nuria\anaconda3\lib\site-packages\numpy\lib\arraysetops.py:583: FutureWarning: elementwise comparison
stead, but in the future will perform elementwise comparison
  mask |= (ar1 == a)
```

In [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1936758 entries, 3561206 to 4392215
Data columns (total 31 columns):
 #   Column                 Dtype  
---  ------                 -----  
 0   Month                  int64  
 1   DayofMonth             int64  
 2   DayOfWeek              int64  
 3   DepTime                float64
 4   CRSDepTime             int64  
 5   ArrTime                float64
 6   CRSArrTime             int64  
 7   UniqueCarrier          object 
 8   FlightNum              int64  
 9   TailNum                object 
 10  ActualElapsedTime      float64
 11  CRSElapsedTime         float64
 12  AirTime                float64
 13  ArrDelay               float64
 14  DepDelay               float64
 15  Distance               int64  
 16  TaxiIn                 float64
 17  TaxiOut                float64
 18  Diverted               int64  
 19  CarrierDelay           float64
 20  WeatherDelay           float64
 21  NASDelay               float64
 22  SecurityDelay          float64
 23  LateAircraftDelay      float64
 24  DistanceKm             float64
 25  AirTimeH               float64
 26  FlightSpeed            float64
 27  LateLanding            bool   
 28  LateTakeOff            bool   
 29  ElapsedTimeDifference  float64
 30  ArrivalDifference      float64
dtypes: bool(2), float64(19), int64(8), object(2)
memory usage: 447.0+ MB
```

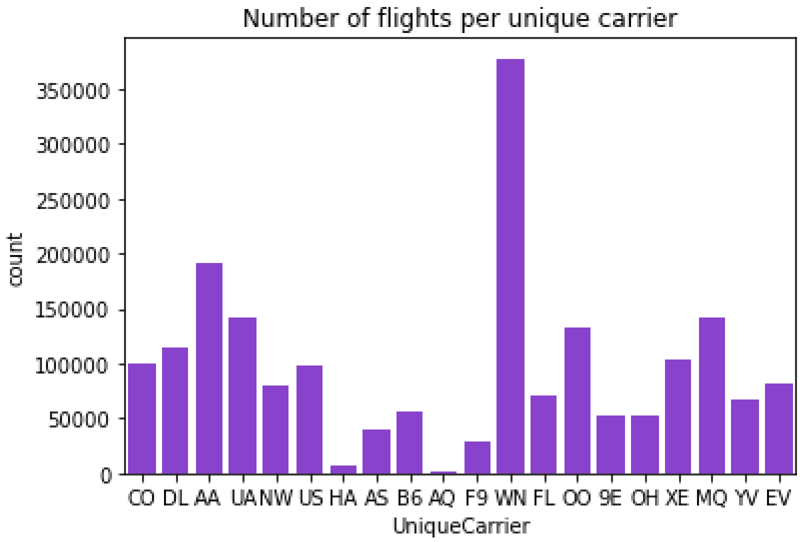
Crea almenys una visualització per:

- Una variable categòrica (UniqueCarrier)

In [16]:

```
sns.countplot( x = "UniqueCarrier", data = df, color = "BlueViolet").set(
    title = "Number of flights per unique carrier")

plt.savefig("flights_carrier.png")
```



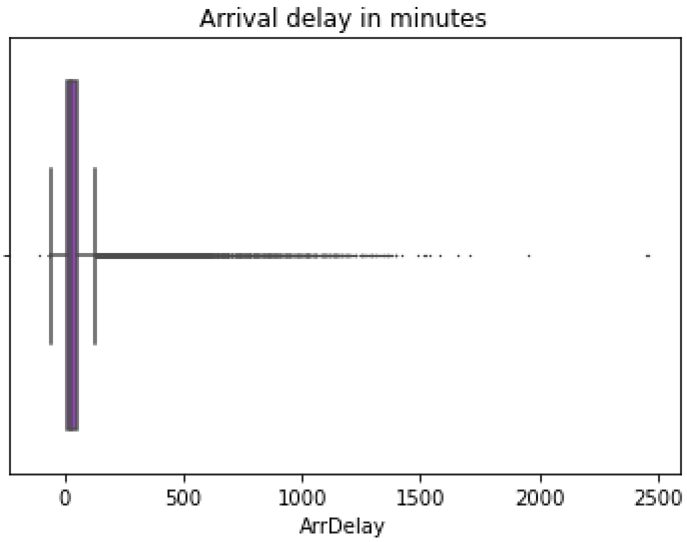
It seems the majority of flights were made by WN (Southwest Airlines), and that would explai amount of delays compared to the other carriers.

- Una variable numèrica (ArrDelay)

In [17]:

```
sns.boxplot(x = "ArrDelay", data = df, color = "DarkOrchid", fliersize = 0.5).set(
    title = "Arrival delay in minutes")

plt.savefig("arrdelay.png")
```

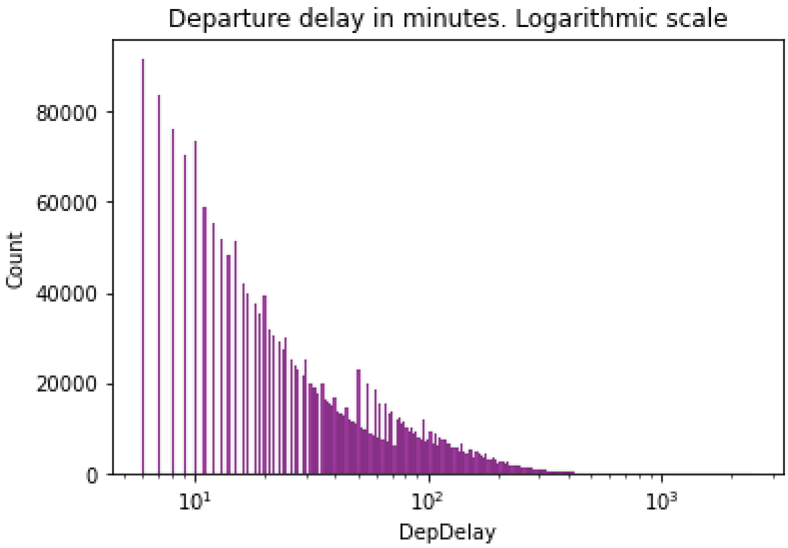


The big majority of arrival delays are comprised in a very small range, but when atypical valu
different from the majority of the data.

In [18]:

```
sns.histplot(df, x = "DepDelay", color = "DarkMagenta", log_scale = True).set(
    title = "Departure delay in minutes. Logarithmic scale")

plt.savefig("depdelay.png")
```



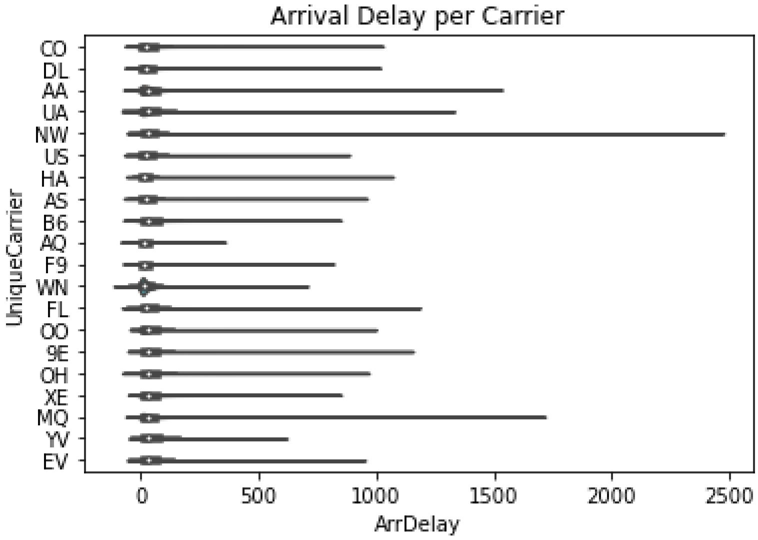
This logarithmic scale plot is made to show all values in a visible scale, since the majority of very small range, same as arrival delays, but there are a small amount of values that are ver

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

In [19]:

```
sns.violinplot(x = "ArrDelay", y = "UniqueCarrier", data = df, scale = "count").se
    title = "Arrival Delay per Carrier")

plt.savefig("arrdelay_carrier.png")
```



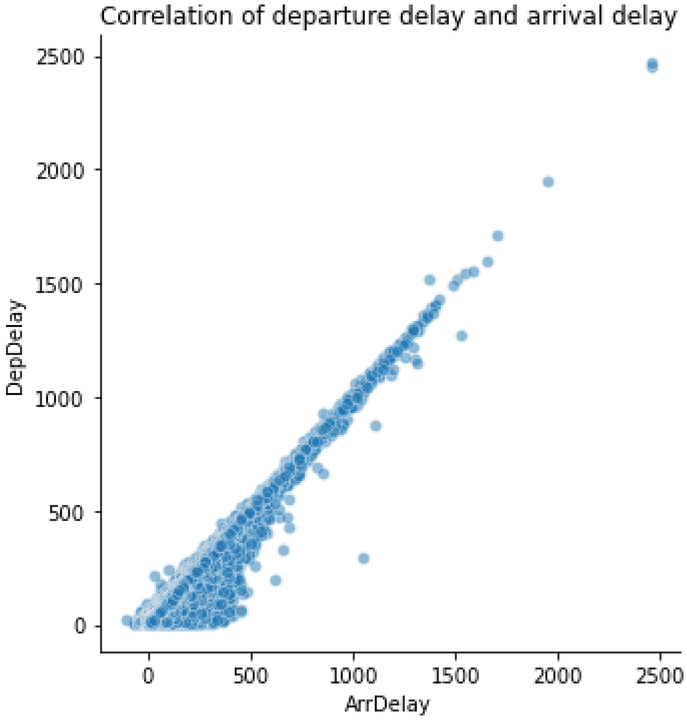
When analyzing arrival delay per carrier, we can see that what was shown in previous visual separated for carriers. The great majority of delays are very close to 0, but there are outliers can see that, although Southwest Airlines had the most flights, it is not the carrier with the bi Northwest Airlines (NW), Envoy Air (MQ), and American Airlines (AA).

- Dues variables numèriques (ArrDelay i DepDelay)

In [3]:

```
sns.relplot(x = "ArrDelay", y = "DepDelay", data = df, alpha = 0.5).set(
    title = "Correlation of departure delay and arrival delay")

plt.savefig("depdelay_arrdelay.png")
```

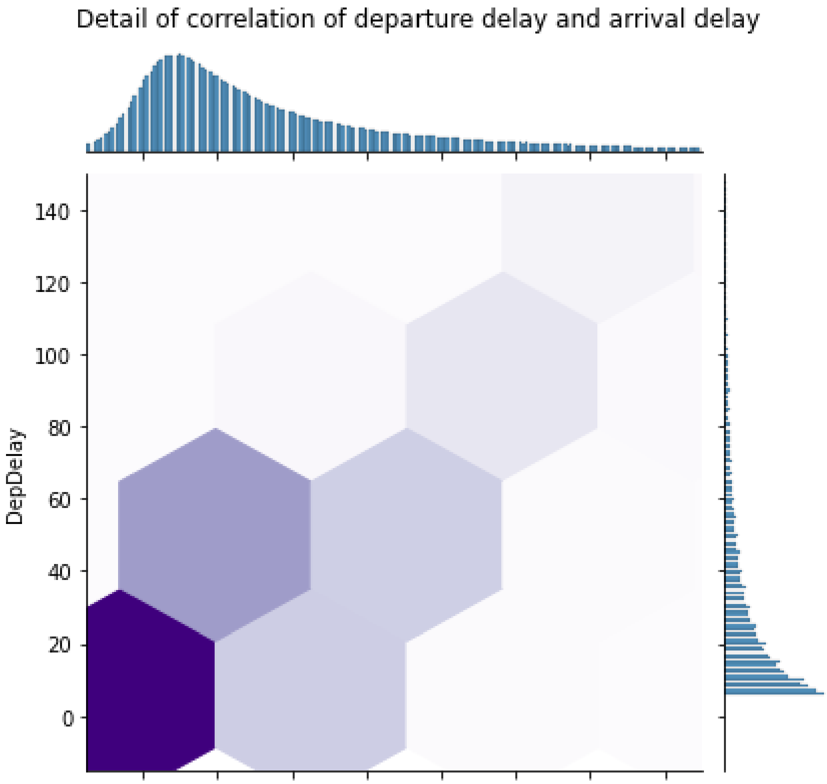


Departure delay and arrival delay are very strongly correlated, although at the base of the pl
be caused by having a greater number of values closer to zero. The more values, the more v

In [4]:

```
plot = sns.jointplot(x = "ArrDelay", y = "DepDelay", data = df, cmap = "Purples",
                          xlim = (-15, 150), ylim = (-15, 150))
plot.fig.suptitle("Detail of correlation of departure delay and arrival delay")
plot.fig.subplots_adjust(top = 0.93)
plt.show()

plt.savefig("detail_depdelay_arrdelay.png")
```



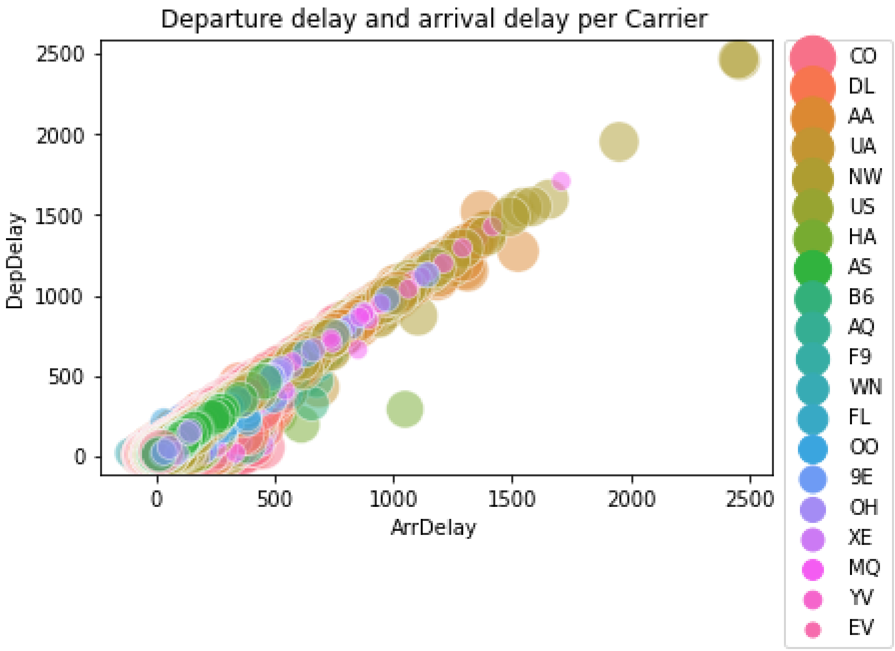
This plot shows only those delays between -15 minutes and 150 minuteses, those majority o this sample, the majority of delays are close to 0. Arrival delays are a bit more spread out be compared to departure delays, which start to go down the farther from 0 they go.

- Tres variables (ArrDelay, DepDelay i UniqueCarrier)

In [5]:

```
sns.scatterplot(data = df, x = "ArrDelay", y = "DepDelay", size = "UniqueCarrier",
                s = 20, sizes = (50, 500), hue = "UniqueCarrier"
               ).set(title = "Departure delay and arrival delay per Carrier")
plt.legend(bbox_to_anchor = (1.02, 1), loc = "upper left", borderaxespad = 0)
plt.show()

plt.savefig("depdelay_arrdelay_carrier.png")
```

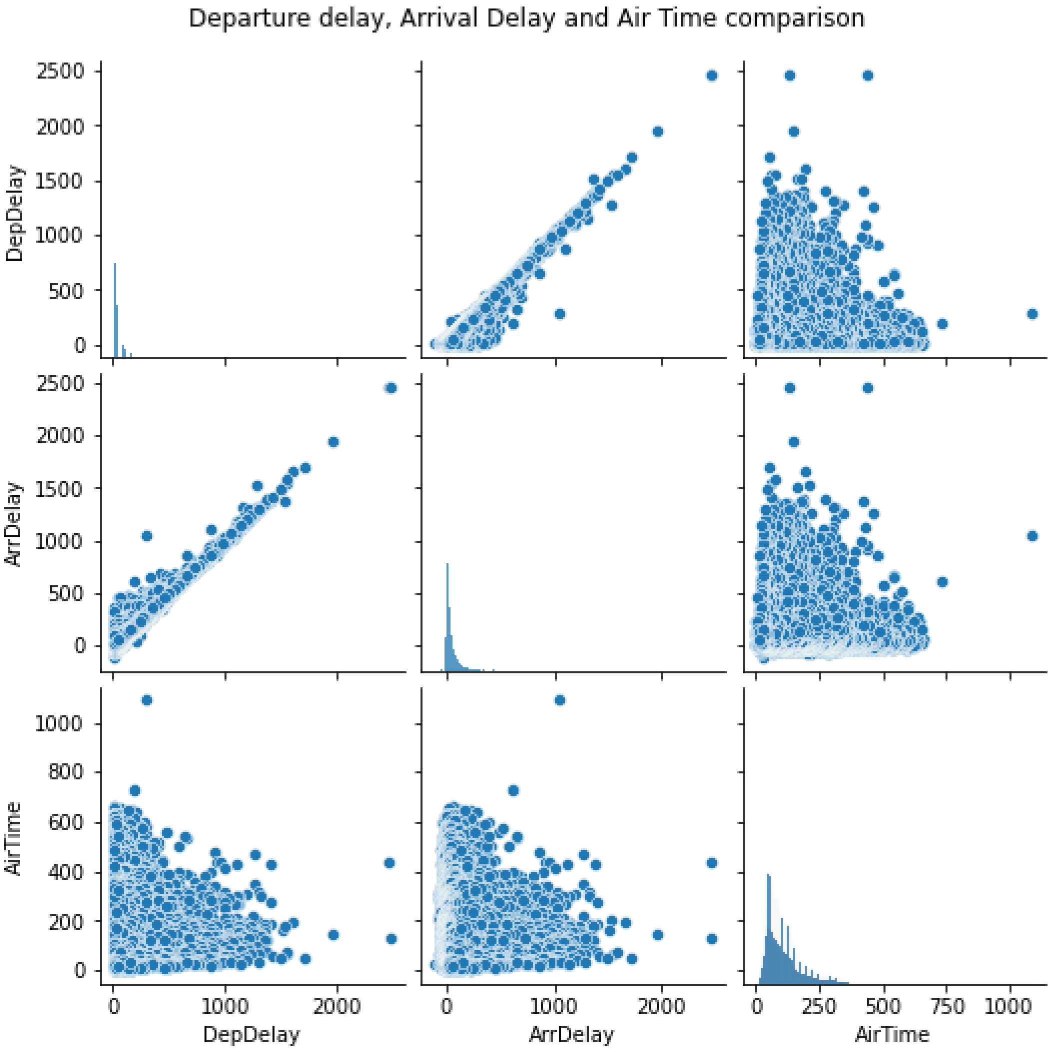


```
<Figure size 432x288 with 0 Axes>
```

Since most of the values are on top of each other, they hardly tell us more information than t outliers.

In [22]:

```
plot = sns.pairplot(df[["DepDelay", "ArrDelay", "AirTime"]])
plot.fig.suptitle("Departure delay, Arrival Delay and Air Time comparison")
plot.fig.subplots_adjust(top = 0.93)
plt.savefig("depdelay_arrdelay_airtime.png")
```



As we can see, air time isn't correlated at all with either arrival or departure delay. Most of th between 0 and 250 minutes.

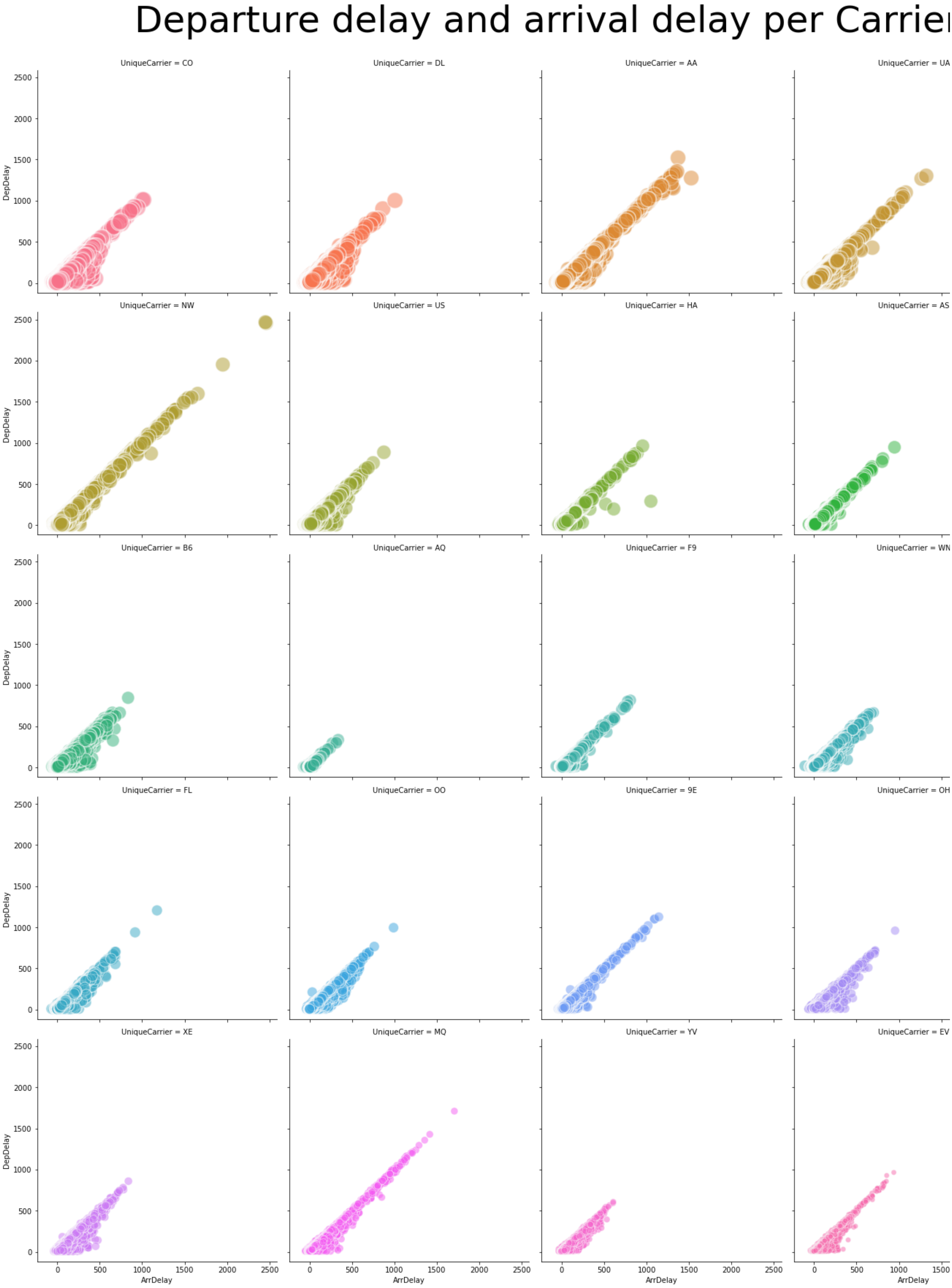
In [19]:

```python
plot = sns.relplot(data = df, x = "ArrDelay", y = "DepDelay", size = "UniqueCarrie
                s = 20, sizes = (50, 500), hue = "UniqueCarrier", col = "UniqueCar

plt.legend(bbox_to_anchor = (1.02, 1), loc = "upper left", borderaxespad = 0)
plot.fig.suptitle("Departure delay and arrival delay per Carrier", fontsize = 50)
plot.fig.subplots_adjust(top = 0.93)
plt.show()

plt.savefig("depdelay_arrdelay_carrier_sep.png")
```

```
No handles with labels found to put in legend.
```



```
<Figure size 432x288 with 0 Axes>
```

When separated for carrier, we can see that the same correlations keeps among the differen

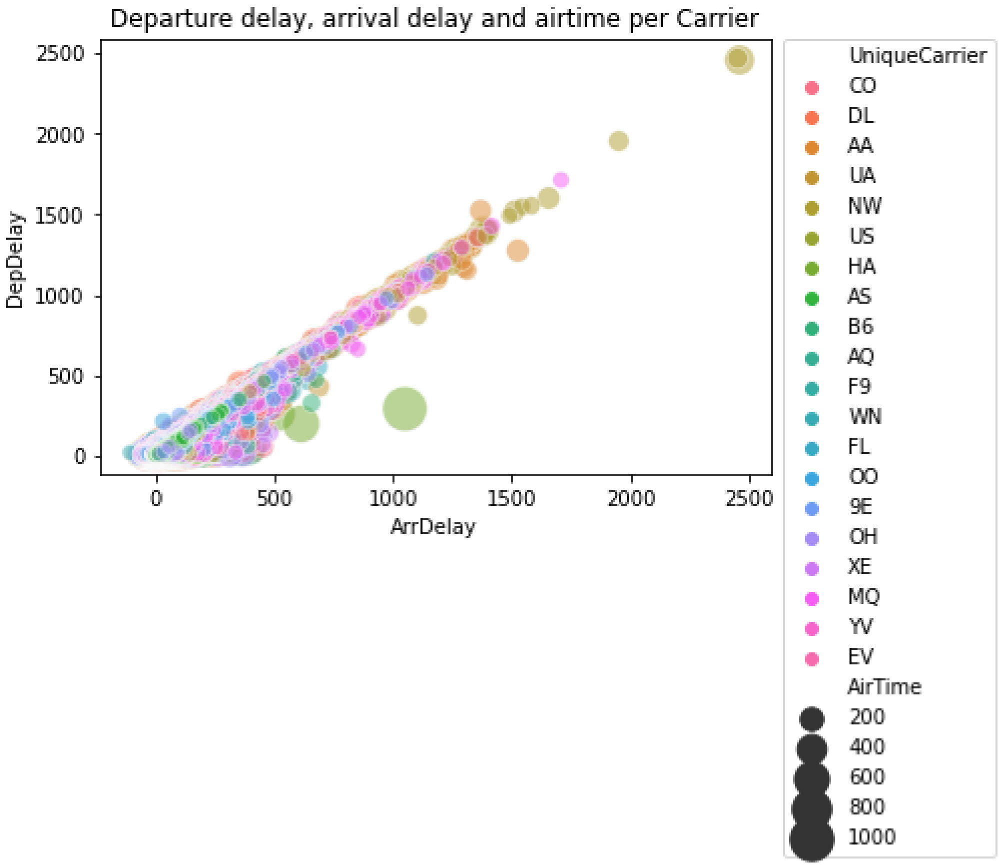
smallest delays are the ones with less flights, with the exeption of Hawaiian Airlines (HA), wh
second fewest flights and has significant delays of up to 1000 minutes, or 17 hours.

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

In [5]:

```
sns.scatterplot(data = df, x = "ArrDelay", y = "DepDelay", hue = "UniqueCarrier",
                alpha = 0.5, s = 20, sizes = (50, 500)
               ).set(title = "Departure delay, arrival delay and airtime per Carri
plt.legend(bbox_to_anchor = (1.02, 1), loc = "upper left", borderaxespad = 0)
plt.show()

plt.savefig("depdelay_arrdelay_carrier_airtime.png")
```



```
<Figure size 432x288 with 0 Axes>
```

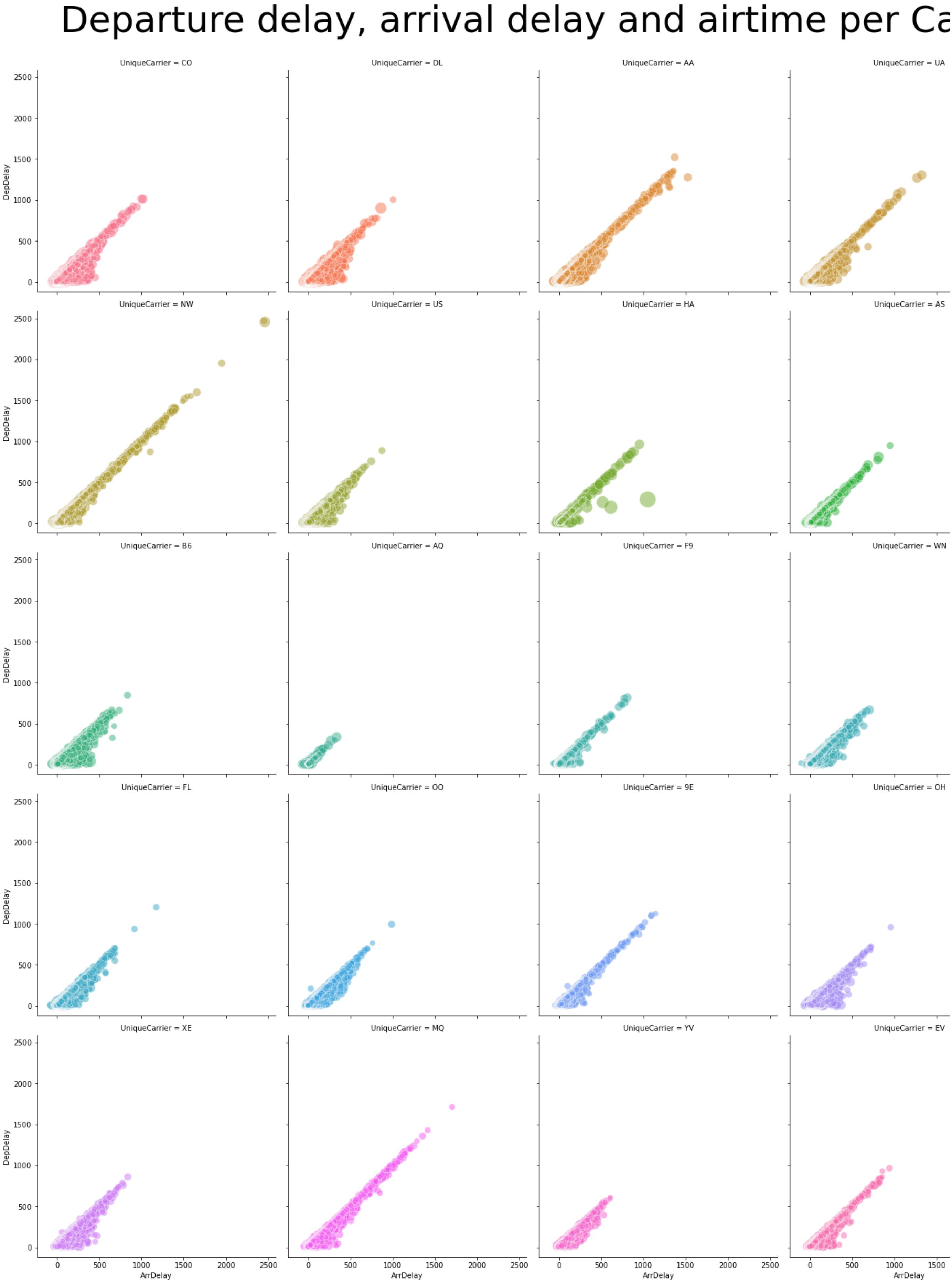
We can see that the majority of flights with great delays have an order of magnitude of delay
means people had to wait up to 10 times the duration of the flight.

In [15]:

```
plot = sns.relplot(data = df, x = "ArrDelay", y = "DepDelay", hue = "UniqueCarrier
            col = "UniqueCarrier", col_wrap = 4, alpha = 0.5, s = 20, sizes = (50,
           )
plot.fig.suptitle("Departure delay, arrival delay and airtime per Carrier", fontsi
plt.legend(bbox_to_anchor = (1.02, 1), loc = "upper left", borderaxespad = 0, font
plot.fig.subplots_adjust(top = 0.93)
plt.show()

plt.savefig("depdelay_arrdelay_carrier_airtime_sep.png")
```

```
No handles with labels found to put in legend.
```



```
<Figure size 432x288 with 0 Axes>
```

The above observation remains when the values are separated per carrier. The great majorit

> the shorter side of duration. That could, however, be explained by those flights being the ma