In [ ]:

```
import pandas as pd
import numpy as np
df = pd.read_csv('DelayedFlights.csv')

# obtenemos el numero total de registros
num_rows = df.shape[0]

# generar una seleccion aleatoria de registros
keep_rows = np.random.choice([True, False], size=num_rows, p=[0.3, 0.7])
num_rows_to_keep = min(sum(keep_rows), 200000)

# descartando registros
df_reduced = df.loc[keep_rows][:num_rows_to_keep]

# creando nuevo archivo de trabajo
df_reduced.to_csv('EstadisticaVuelosReducido.csv', index=False)
```

In [5]:

```
import pandas as pd
import numpy as np

#mostrando nuevo archivo de trabajo reducido
df = pd.read_csv('EstadisticaVuelosReducido.csv')
df

print("Nombre de columnas:", len(df.columns))
print("Nombre de columnas:", list(df.columns))

print("Primeros 3D Objects/ registros:")
print(df.head(3))
```

```
Nombre de columnas: 30
Nombre de columnas: ['Unnamed: 0', 'Year', 'Month', 'DayofMonth', 'DayOfWeek', 'DepTim
e', 'CRSDepTime', 'ArrTime', 'CRSArrTime', 'UniqueCarrier', 'FlightNum', 'TailNum', 'Act
ualElapsedTime', 'CRSElapsedTime', 'AirTime', 'ArrDelay', 'DepDelay', 'Origin', 'Dest',
'Distance', 'TaxiIn', 'TaxiOut', 'Cancelled', 'CancellationCode', 'Diverted', 'CarrierDe
lay', 'WeatherDelay', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay']
Primeros 3D Objects/ registros:
   Unnamed: 0  Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  \
0           2  2008      1           3          4    628.0         620
1           4  2008      1           3          4   1829.0        1755
2          11  2008      1           3          4   1644.0        1510

   ArrTime  CRSArrTime UniqueCarrier  ...  TaxiIn TaxiOut  Cancelled  \
0    804.0         750            WN  ...     3.0    17.0          0
1   1959.0        1925            WN  ...     3.0    10.0          0
2   1845.0        1725            WN  ...     6.0     8.0          0

   CancellationCode  Diverted  CarrierDelay  WeatherDelay NASDelay  \
0                 N         0           NaN           NaN      NaN
1                 N         0           2.0           0.0      0.0
2                 N         0           8.0           0.0      0.0

  SecurityDelay  LateAircraftDelay
0           NaN                NaN
1           0.0               32.0
2           0.0               72.0

[3 rows x 30 columns]
```

In [11]:

```
# borrando columnas por no tener datos relevantes como el año, ya que se refiere a un so
#para efectos de practica borrare algunas columnas sin datos
import pandas as pd

df = pd.read_csv('EstadisticaVuelosReducido.csv')
```

```
# eliminar les columnes 'col2' i 'col4'
df = df.drop(['Unnamed: 0','FlightNum','TaxiIn','TaxiOut'], axis=1)

# mostrar les primeres tres files del dataset
print(df.head(3))
print("Nombre de columnas:", len(df.columns))
print("Nombre de columnas:", list(df.columns))
```

```
   Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  ArrTime  \
0  2008      1           3          4    628.0         620    804.0
1  2008      1           3          4   1829.0        1755   1959.0
2  2008      1           3          4   1644.0        1510   1845.0

   CRSArrTime UniqueCarrier TailNum  ...  Dest  Distance  Cancelled  \
0         750            WN  N428WN  ...   BWI       515          0
1        1925            WN  N464WN  ...   BWI       515          0
2        1725            WN  N334SW  ...   MCO       828          0

   CancellationCode  Diverted CarrierDelay WeatherDelay  NASDelay  \
0                 N         0          NaN          NaN       NaN
1                 N         0          2.0          0.0       0.0
2                 N         0          8.0          0.0       0.0

   SecurityDelay LateAircraftDelay
0            NaN               NaN
1            0.0              32.0
2            0.0              72.0

[3 rows x 26 columns]
Nombre de columnas: 26
Nombre de columnas: ['Year', 'Month', 'DayofMonth', 'DayOfWeek', 'DepTime', 'CRSDepTim
e', 'ArrTime', 'CRSArrTime', 'UniqueCarrier', 'TailNum', 'ActualElapsedTime', 'CRSElapse
dTime', 'AirTime', 'ArrDelay', 'DepDelay', 'Origin', 'Dest', 'Distance', 'Cancelled', 'C
ancellationCode', 'Diverted', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDela
y', 'LateAircraftDelay']
```

In [20]:

```
import pandas as pd
import numpy as np


df = pd.read_csv('EstadisticaVuelosReducido.csv')
import math

print('TOP 10 AEROLINEAS CON EL MAYOR RETRASO')
total_delays = df.groupby(['UniqueCarrier'])['ArrDelay'].sum().reset_index()
top_delays = total_delays.sort_values(by=['ArrDelay'], ascending=False)
print(top_delays.head(10))

print('TOP DE LOS VUELOS MAS LARGOS')
df['FlightTime'] = df['AirTime'] + df['ArrDelay']
top_flights = df[['Origin', 'Dest', 'FlightNum', 'FlightTime']].sort_values(by=['FlightT
print(top_flights)

print('TOP DE LOS VUELOS MAS ATRASADOS')
df['TotalDelay'] = abs(df['ArrDelay'].fillna(0)) + abs(df['DepDelay'].fillna(0))
top_delayed_flights = df[['FlightNum', 'Origin', 'Dest', 'TotalDelay']].sort_values(by=[
print(top_delayed_flights)

print('CREANDO NUEVAS COLUMNAS Y GENERANDO NUEVA INFORMACION')
# Calcular la distancia en kilómetros
df["Distancia_km"] = df["Distance"] * 1.60934

# Calcular el tiempo en horas
df["Tiempo_h"] = df["AirTime"] / 60
```

```python
# Calcular la velocidad media en km/h
df["Velocidad_media"] = df["Distancia_km"] / df["Tiempo_h"]

# Imprimir el resultado
print(df.head())
```

```
TOP 10 AEROLINEAS CON EL MAYOR RETRASO
   UniqueCarrier   ArrDelay
17            WN  1291447.0
1             AA   799227.0
15            UA   778834.0
14            OO   735318.0
11            MQ   634128.0
18            XE   597398.0
7             EV   426335.0
19            YV   394998.0
16            US   371855.0
6             DL   362355.0
TOP DE LOS VUELOS MAS LARGOS
       Origin  Dest FlightNum  FlightTime
156456    BNA  MEM       1743      1537.0
95743     PDX  MSP        218      1519.0
97349     LAS  DTW       1192      1482.0
94741     VPS  ORD       4477      1476.0
105075    RSW  STL       2233      1453.0
154614    TPA  MSP        443      1404.0
154158    HNL  PDX        218      1361.0
39606     FLL  DTW        243      1327.0
104735    SDF  DFW       1965      1317.0
39539     HNL  PDX        218      1279.0
TOP DE LOS VUELOS MAS ATRASADOS
        FlightNum Origin  Dest TotalDelay
156456       1743    BNA  MEM      2980.0
94741        4477    VPS  ORD      2714.0
95743         218    PDX  MSP      2698.0
105075       2233    RSW  STL      2628.0
97349        1192    LAS  DTW      2514.0
154614        443    TPA  MSP      2449.0
163678       2228    HDN  DFW      2354.0
104735       1965    SDF  DFW      2313.0
94856        4513    SHV  ORD      2285.0
39606         243    FLL  DTW      2266.0
CREANDO NUEVAS COLUMNAS Y GENERANDO NUEVA INFORMACION
   Unnamed: 0  Year  Month  DayofMonth  DayOfWeek  DepTime  CRSDepTime  \
0           2  2008      1           3          4    628.0         620   
1           4  2008      1           3          4   1829.0        1755   
2          11  2008      1           3          4   1644.0        1510   
3          15  2008      1           3          4   1029.0        1020   
4          16  2008      1           3          4   1452.0        1425   

   ArrTime  CRSArrTime UniqueCarrier  ...  CarrierDelay WeatherDelay  \
0    804.0         750            WN  ...           NaN          NaN   
1   1959.0        1925            WN  ...           2.0          0.0   
2   1845.0        1725            WN  ...           8.0          0.0   
3   1021.0        1010            WN  ...           NaN          NaN   
4   1640.0        1625            WN  ...           3.0          0.0   

   NASDelay  SecurityDelay  LateAircraftDelay  FlightTime  TotalDelay  \
0       NaN            NaN                NaN        90.0        22.0   
1       0.0            0.0               32.0       111.0        68.0   
2       0.0            0.0               72.0       187.0       174.0   
3       NaN            NaN                NaN        48.0        20.0   
4       0.0            0.0               12.0       228.0        42.0   

  Distancia_km  Tiempo_h  Velocidad_media
0     828.81010  1.266667       654.323763
```

```
1    828.81010  1.283333       645.826052
2   1332.53352  1.783333       747.215058
3    260.71308  0.616667       422.777968
4   2396.30726  3.550000       675.016130

[5 rows x 35 columns]
```

```
In [ ]:
```