S04 T01: Transformació Registre Log amb Regular expressions

Descripció: L'anàlisi de registres és una funció important per al control i l'alerta, el compliment de les polítiques de seguretat, l'auditoria i el compliment normatiu, la resposta a incidents de seguretat i fins i tot les investigacions forenses. En analitzar les dades de registre, les empreses poden identificar més fàcilment les possibles amenaces i altres problemes, trobar la causa arrel i iniciar una resposta ràpida per mitigar els riscos.

NIVELL 1

L'analista ha d'assegurar-se que els registres consisteixen en una gamma completa de missatges i s'interpreten segons el context. Els elements de registre han d'estandaritzar-se, utilitzant els mateixos termes o terminologia, per evitar confusions i proporcionar cohesió.

Com Científic de Dades se t'ha proporcionat accés als registres-Logs on queda registrada l'activitat de totes les visites a realitzades a la pàgina web de l'agència de viatges "akumenius.com".

Exercici 1

Estandaritza, identifica i enumera cada un dels atributs / variables de l'estructura de l'arxiu "Web_access_log-akumenius.com" que trobaràs al repositori de GitHub "Data-sources".

Primer de tot importem les llibreries que necessitem i carreguem el document amb el qual hem de treballar. El document l'he arreglat una mica amb un processador de text abans d'utilitzar-lo aquí.

In [1]:
```
import pandas as pd
import numpy as np
import re

dataf = pd.read_table("C:\\Users\\Anna\\DataScience\\SPRINTS\\SPRINT 4\\data.txt", encodir

# els punts de "encoding" i "engine" els he posat per solventar errors que em sortien a l

dataf.loc[10:30,:] #imprimeixo d'aquesta forma perquè així veig files amb localhost i amb
```

Out[1]:

	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (internal dummy connection)" VLOG=-
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (internal dummy connection)" VLOG=-
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

A partir d'ara començaré a treballar sobre la base de dades "data", creant i canviant el nom de les columnes per veure més clarament els atributs amb què treballar, i també estandaritzar tota la informació

In [2]:
```
# comenco per canviar els noms de les columnes que tinc per poder treballar més còmodament

dataf.columns=["pagWeb", "ip", "dataHora", "altres"]
dataf.loc[10:30,:]
```

Out[2]:

	pagWeb	ip	dataHora	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	pagWeb	ip	dataHora	altres
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [3]:
```
# Tipus de dades que trobem a cada columna
dataf.dtypes
```

Out[3]:
```
pagWeb      object
ip          object
dataHora    object
altres      object
dtype: object
```

In [4]:
```
# Petita descripció de cada columna
dataf.describe()
```

Out[4]:

	pagWeb	ip	dataHora	altres
count	261872	261872	261872	246399
unique	408	9276	115882	162295
top	www.akumenius.com	66.249.76.216	□GET /newdesign/libraries/anythingSlider/image...	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
freq	232300	46382	732	13891

In [5]:
```
# Com que en l'anterior punt veiem que la columna altres té menys quantitat de dades que

print(dataf.isnull())

print("_____________________________________________________________________")

print(dataf.count())

print("_____________________________________________________________________")
```

```
print(dataf.isnull().sum())
```

```
        pagWeb     ip  dataHora  altres
0        False  False     False   False
1        False  False     False   False
2        False  False     False   False
3        False  False     False   False
4        False  False     False   False
...        ...    ...       ...     ...
261867   False  False     False   False
261868   False  False     False   False
261869   False  False     False   False
261870   False  False     False   False
261871   False  False     False   False

[261872 rows x 4 columns]
```

```
pagWeb      261872
ip          261872
dataHora    261872
altres      246399
dtype: int64
```

```
pagWeb          0
ip              0
dataHora        0
altres      15473
dtype: int64
```

In [6]:

```
# Eliminem les files que contenen algun valor nul i comprobem que s'han eliminat, de manei

dataf = dataf.dropna(subset=["altres"])

dataf.describe()
```

Out[6]:

	pagWeb	ip	dataHora	altres
count	246399	246399	246399	246399
unique	2	2828	113719	162295
top	www.akumenius.com	66.249.76.216	28/Feb/2014:04:16:25 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
freq	232273	46382	83	13891

In [7]:

```
# Ara vull dividir la columna de "dataHora" en 3 columnes, la data, la de l'hora i la de l

# Creo les columnes buides i les col·loco on vull i amb el nom que vull
dataf.insert(3, "data", "")
dataf.insert(4, "hora1", "")

dataf.loc[10:30,:]
```

Out[7]:

	pagWeb	ip	dataHora	data	hora1	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...

	pagWeb	ip	dataHora	data	hora1	altres
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100			□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100			□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100			□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100			□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100			□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100			□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100			□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100			□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100			□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100			□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100			□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [8]:

```
# Introdueixo la informació amb una funció lambda perquè em permeti agafar el que jo vull

dataf["data"] = dataf["dataHora"].apply(lambda x: x.split(":",1)[0])
dataf["hora1"] = dataf["dataHora"].apply(lambda x: x.split(":",1)[1])


dataf.loc[10:30,:]
```

Out[8]:

	pagWeb	ip	dataHora	data	hora1	altres
10	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
11	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
12	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
13	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
14	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
15	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
16	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
17	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
18	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
19	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
20	localhost	127.0.0.1	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...
21	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:31 +0100	23/Feb/2014	03:10:31 +0100	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...
22	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:33 +0100	23/Feb/2014	03:10:33 +0100	□GET /hoteles-baratos/ofertas-hotel-Metropolis...
23	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:35 +0100	23/Feb/2014	03:10:35 +0100	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...
24	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:38 +0100	23/Feb/2014	03:10:38 +0100	□GET /hoteles-baratos/ofertas-hotel-Kensington...
25	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:39 +0100	23/Feb/2014	03:10:39 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
26	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:40 +0100	23/Feb/2014	03:10:40 +0100	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...
27	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:42 +0100	23/Feb/2014	03:10:42 +0100	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...
28	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:45 +0100	23/Feb/2014	03:10:45 +0100	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...
29	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:46 +0100	23/Feb/2014	03:10:46 +0100	□GET /destinos-baratos/destinosEstrelles/hotel...
30	www.akumenius.com	66.249.76.216	23/Feb/2014:03:10:47 +0100	23/Feb/2014	03:10:47 +0100	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...

In [9]:

```
# Creo un nou dataset per separar l'hora i la UTC de la columna hora

hora = dataf["hora1"].str.split(" ", expand=True)
```

```
# Elimino una columna de més que se m'ha creat
hora.drop(hora.columns[2], axis=1, inplace=True)

# Canvio el nom de les columnes noves
hora.columns = ["hora", "UTC"]

#Ajunto el dataset principal "dataf" amb el dataset nou "hora"
dataf = pd.concat([dataf, hora], axis=1)

# Elimino la coumna "dataHora"
dataf = dataf.drop("dataHora", axis=1)
dataf = dataf.drop("hora1", axis=1)


dataf.loc[10:30,:]
```

Out[9]:

	pagWeb	ip	data	altres	hora	UTC
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100

	pagWeb	ip	data	altres	hora	UTC
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100

In [10]:

```
# Creo la columna que separi el Get i Options fins al "-" de la columna "altres"

dataf.insert(6, "optionGetComplert", "")

# Introdueixo la informació amb una funció lambda perquè em permeti agafar el que jo vull

dataf["optionGetComplert"] = dataf["altres"].map(lambda x: x.split('"-"',1)[0])

dataf.loc[10:30,:]
```

Out[10]:

	pagWeb	ip	data	altres	hora	UTC	option
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS * HT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100	□GET /hoteles-bar hotel-
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100	□GET /hoteles-bar hotel
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100	□GET /hoteles-bar hotel-

	pagWeb	ip	data	altres	hora	UTC	option
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100	□GET /hoteles-bar hotel-
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100	□G baratos/destinosEs
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100	□GET /hoteles-bar hotel-I
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100	□GET /hoteles-bar hote
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100	□GET /hoteles-bar hote
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100	□G baratos/destinosEs
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100	□GET /hoteles-bar hotel

In [11]:

```
# Creo un nou dataset per separar els diferents apartats de la columna "optionGetComplert"

nom = dataf["optionGetComplert"].str.split(" ", expand=True)

# Se m'han creat moltes columnes de més i les elimino utilitzant el drop
nom.drop(nom.columns[5:45], axis=1, inplace=True)

# Canvio el nom de les columnes noves
nom.columns = ["optionGet", "adreçaWeb", "HTTP", "numeroA", "numeroB"]

#Ajunto el dataset principal "dataf" amb el dataset nou "nom"
dataf = pd.concat([dataf, nom], axis=1)

# Elimino la columna "optionGetComplert" perquè ja he separat tota la informació en altres
dataf = dataf.drop("optionGetComplert", axis=1)

dataf.loc[10:30,:]
```

Out[11]:

	pagWeb	ip	data	altres	hora	UTC	optionGet
10	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
11	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
12	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
13	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
14	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
15	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
16	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS
17	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS

	pagWeb	ip	data	altres	hora	UTC	optionGet	
18	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
19	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
20	localhost	127.0.0.1	23/Feb/2014	□OPTIONS * HTTP/1.0" 200 - "-" "Apache (intern...	03:10:31	+0100	□OPTIONS	
21	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Club-&-Hot...	03:10:31	+0100	□GET	/hc
22	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Metropolis...	03:10:33	+0100	□GET	/hc
23	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Faena-Hote...	03:10:35	+0100	□GET	/hc
24	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kensington...	03:10:38	+0100	□GET	/hc
25	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:39	+0100	□GET	barat
26	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Howard-Jho...	03:10:40	+0100	□GET	/hc
27	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Princesa-S...	03:10:42	+0100	□GET	/hc
28	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Kfar-Gilad...	03:10:45	+0100	□GET	/hc
29	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /destinos-baratos/destinosEstrelles/hotel...	03:10:46	+0100	□GET	barat
30	www.akumenius.com	66.249.76.216	23/Feb/2014	□GET /hoteles-baratos/ofertas-hotel-Casual-Hot...	03:10:47	+0100	□GET	/hc

In [12]:

```
#Repeteixo el procés anterior per a l'última part de la columna altres

# Creo un nou dataset per separar l'última part de la columna altres
final = dataf["altres"].str.split('"-"', expand=True)

# Em quedo només amb la columna que vull treballar
final.drop(final.columns[0], axis=1, inplace=True)
final.drop(final.columns[1], axis=1, inplace=True)

# Canvio el nom de les columnes noves
final.columns = ["colFinal"]

print(final)
```

```
                                                colFinal
0                 "Apache (internal dummy connection)" VLOG=-
1                 "Apache (internal dummy connection)" VLOG=-
2                 "Apache (internal dummy connection)" VLOG=-
3                 "Apache (internal dummy connection)" VLOG=-
4                 "Apache (internal dummy connection)" VLOG=-
...                                                     ...
261867    "Mozilla/5.0 (compatible; YandexBot/3.0; +htt...
261868    "Mozilla/5.0+(compatible; UptimeRobot/2.0; ht...
261869            "Apache (internal dummy connection)" VLOG=-
```

```
261870        "Apache (internal dummy connection)" VLOG=-
261871        "Apache (internal dummy connection)" VLOG=-

[246399 rows x 1 columns]
```

In [13]:

```
#Ajunto el dataset principal "dataf" amb el dataset nou "final"
dataf = pd.concat([dataf, final], axis=1)

# Elimino la columna "altres" perquè ja he separat tota la informació en altres columnes
dataf = dataf.drop("altres", axis=1)

dataf.loc[10:30,:]
```

Out[13]:

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb
10	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
11	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
12	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
13	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
14	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
15	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
16	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT
17	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	* HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
18	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
19	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
20	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:31	+0100	□GET	/hoteles-baratos/ofertas-hotel-Club-&-Hotel-Le...	HTT
22	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:33	+0100	□GET	/hoteles-baratos/ofertas-hotel-Metropolis-Hote...	HTT
23	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:35	+0100	□GET	/hoteles-baratos/ofertas-hotel-Faena-Hotel-Bue...	HTT
24	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:38	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kensington-Town...	HTT
25	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:39	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT
26	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:40	+0100	□GET	/hoteles-baratos/ofertas-hotel-Howard-Jhonson-...	HTT
27	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:42	+0100	□GET	/hoteles-baratos/ofertas-hotel-Princesa-Sofia-...	HTT
28	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:45	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kfar-Giladi-en-...	HTT
29	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:46	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
30	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:47	+0100	□GET	/hoteles-baratos/ofertas-hotel-Casual-Hotel-Va...	HTT

In [14]:

```
# Separo en diferents columnes la informació de la columna restant
final1 = dataf["colFinal"].str.split('"', expand=True)

# Elimino les columnes que no tenen informació
final1.drop(final1.columns[0], axis=1, inplace=True)

# Canvio el nom de les columnes noves
final1.columns = ["buscador", "vlog"]

# Elimino la columna "colFinal" perquè ja he separat tota la informació en altres columnes
dataf = dataf.drop("colFinal", axis=1)

#Ajunto el dataset principal "dataf" amb el dataset nou "final1"
dataf = pd.concat([dataf, final1], axis=1)



dataf.loc[10:30,:]
```

Out[14]:

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
10	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
11	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
12	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
13	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
14	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
15	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
16	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
17	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
18	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
19	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
20	localhost	127.0.0.1	23/Feb/2014	03:10:31	+0100	□OPTIONS	*	HTT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:31	+0100	□GET	/hoteles-baratos/ofertas-hotel-Club-&-Hotel-Le...	HTT
22	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:33	+0100	□GET	/hoteles-baratos/ofertas-hotel-Metropolis-Hote...	HTT
23	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:35	+0100	□GET	/hoteles-baratos/ofertas-hotel-Faena-Hotel-Bue...	HTT
24	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:38	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kensington-Town...	HTT
25	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:39	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT
26	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:40	+0100	□GET	/hoteles-baratos/ofertas-hotel-Howard-Jhonson-...	HTT
27	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:42	+0100	□GET	/hoteles-baratos/ofertas-hotel-Princesa-Sofia-...	HTT
28	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:45	+0100	□GET	/hoteles-baratos/ofertas-hotel-Kfar-Giladi-en-...	HTT
29	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:46	+0100	□GET	/destinos-baratos/destinosEstrelles/hoteles-en...	HTT

	pagWeb	ip	data	hora	UTC	optionGet	adreçaWeb	
30	www.akumenius.com	66.249.76.216	23/Feb/2014	03:10:47	+0100	□GET	/hoteles-baratos/ofertas-hotel-Casual-Hotel-Va...	HTT

NIVELL 2

Exercici 2

Neteja, preprocesa, estructura i transforma (dataframe) les dades del registre d'Accés a la web.

En aquest exercici seguiré dividint la informació del dataset anterior en més columnes

In [15]:
```
# Imprimeixo les columnes per tenir els noms més a l'abast
dataf.columns
```

Out[15]:
```
Index(['pagWeb', 'ip', 'data', 'hora', 'UTC', 'optionGet', 'adreçaWeb', 'HTTP',
       'numeroA', 'numeroB', 'buscador', 'vlog'],
      dtype='object')
```

In [16]:
```
# Separo en diferents columnes la informació de la columna data
datasplit = dataf["data"].str.split('/', expand=True)

# Canvio el nom de les columnes noves
datasplit.columns = ["dia", "mes", "any"]

#Ajunto el dataset principal "dataf" amb el dataset nou "datasplit"
dataf = pd.concat([dataf, datasplit], axis=1)

# No elimino la columna data perquè potser m'interessa en un futur
# Si que reordeno l'ordre de les columnes
dataf = dataf.reindex(columns=['pagWeb', 'ip', 'data', 'dia', 'mes',
       'any', 'hora', 'UTC', 'optionGet', 'adreçaWeb', 'HTTP',
       'numeroA', 'numeroB', 'buscador', 'vlog'])

dataf.loc[19:22,:]  #redueixo el resultat que veig per veure més còmode els resultats
```

Out[16]:

	pagWeb	ip	data	dia	mes	any	hora	UTC	optionGet	adreçaWeb	
19	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	+0100	□OPTIONS	*	HT
20	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	+0100	□OPTIONS	*	HT
21	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:31	+0100	□GET	/hoteles-baratos/ofertas-hotel-Club-&-Hotel-Le...	HT
22	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:33	+0100	□GET	/hoteles-baratos/ofertas-hotel-Metropolis-Hote...	HT

In [17]:

```
# Separo en diferents columnes la informació de la columna hora
horasplit = dataf["hora"].str.split(':', expand=True)

# Canvio el nom de les columnes noves
horasplit.columns = ["hores", "minuts", "segons"]

#Ajunto el dataset principal "dataf" amb el dataset nou "horasplit"
dataf = pd.concat([dataf, horasplit], axis=1)

# No elimino la columna hora perquè potser m'interessa en un futur
# Si que reordeno l'ordre de les columnes
dataf = dataf.reindex(columns=['pagWeb', 'ip', 'data', 'dia', 'mes',
       'any', 'hora', "hores", "minuts", "segons", 'UTC', 'optionGet', 'adreçaWeb', 'HTTP'
       'numeroA', 'numeroB', 'buscador', 'vlog'])

dataf.loc[19:22,:]
```

Out[17]:

	pagWeb	ip	data	dia	mes	any	hora	hores	minuts	segons	UTC	optio
19	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□OPTI
20	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□OPTI
21	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□
22	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:33	03	10	33	+0100	□

In [18]:

```
# Separo en diferents columnes la informació de la columna adreçaWeb, separant de tota la

adreçaWebSplit = dataf["adreçaWeb"].str.split('/', expand=True)

# Elimino les columnes que no vull
adreçaWebSplit.drop(adreçaWebSplit.columns[2:20], axis=1, inplace=True)
adreçaWebSplit.drop(adreçaWebSplit.columns[0], axis=1, inplace=True)

# Canvio el nom de les columnes noves
adreçaWebSplit.columns = ["llocWeb"]

#Ajunto el dataset principal "dataf" amb el dataset nou "adreçaWebSplit"
dataf = pd.concat([dataf, adreçaWebSplit], axis=1)

# No elimino la columna adreçaWeb perquè potser m'interessa en un futur
# Si que reordeno l'ordre de les columnes
dataf = dataf.reindex(columns=['pagWeb', 'ip', 'data', 'dia', 'mes',
       'any', 'hora', "hores", "minuts", "segons", 'UTC', 'optionGet', 'adreçaWeb', "llocW
       'numeroA', 'numeroB', 'buscador', 'vlog'])

dataf.loc[19:22,:]
```

Out[18]:

	pagWeb	ip	data	dia	mes	any	hora	hores	minuts	segons	UTC	optio
19	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□OPTI
20	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□OPTI
21	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	+0100	□
22	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:33	03	10	33	+0100	□

In [19]:

```
# Separo en diferents columnes la informació de la columna HTTP

HTTPSplit = dataf["HTTP"].str.split('/', expand=True)

# Elimino les columnes que no vull, o que no m'aporten informació
HTTPSplit.drop(HTTPSplit.columns[2:5], axis=1, inplace=True)

# Canvio el nom de les columnes noves
HTTPSplit.columns = ["HTTP2", "HTTPnum"]

#Ajunto el dataset principal "dataf" amb el dataset nou "HTTPSplit"
dataf = pd.concat([dataf, HTTPSplit], axis=1)

# No elimino la columna adreçaWeb perquè potser m'interessa en un futur
# Si que reordeno l'ordre de les columnes
dataf = dataf.reindex(columns=['pagWeb', 'ip', 'data', 'dia', 'mes',
       'any', 'hora', "hores", "minuts", "segons", 'UTC', 'optionGet', 'adreçaWeb', "llocW
       'numeroA', 'numeroB', 'buscador', 'vlog'])

dataf.loc[19:22,:]
```

Out[19]:

	pagWeb	ip	data	dia	mes	any	hora	hores	minuts	segons	...	optionGet
19	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	...	□OPTIONS
20	localhost	127.0.0.1	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	...	□OPTIONS
21	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	...	□GET

	pagWeb	ip	data	dia	mes	any	hora	hores	minuts	segons	...	optionGet
22	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:33	03	10	33	...	□GET

4 rows × 21 columns

Exercici 3

Geolocalitza les IP's.

Per geolocalitzar les diferents IP, el primer que faré serà eliminar totes les files que representen conexió des d'un localhost, ja que normalment sol ser gent de la propia empresa, per tant, no m'interessa.

In [20]:
```
datafFiltrat = dataf.drop(dataf[dataf["ip"]==" 127.0.0.1 "].index)

datafFiltrat.loc[19:22,:]
```

Out[20]:

	pagWeb	ip	data	dia	mes	any	hora	hores	minuts	segons	...	optionGet
21	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:31	03	10	31	...	□GET
22	www.akumenius.com	66.249.76.216	23/Feb/2014	23	Feb	2014	03:10:33	03	10	33	...	□GET

2 rows × 21 columns

In [21]:
```
print(datafFiltrat.describe())
print(datafFiltrat.dtypes)
```

```
                   pagWeb              ip          data     dia     mes  \
count              232508          232508        232508  232508  232508
unique                  2            2827             8       8       2
top     www.akumenius.com    66.249.76.216    25/Feb/2014      25     Feb
freq               232273           46382         52090   52090  214563

           any      hora   hores  minuts  segons  ... optionGet  \
count   232508    232508  232508  232508  232508  ...    232508
unique       1     62961      24      60      60  ...         5
top       2014  04:16:25      12      50      01  ...      □GET
freq    232508        84   18452    4880    4869  ...    215200

           adreçaWeb  llocWeb       HTTP   HTTP2 HTTPnum numeroA numeroB  \
count          232508   232466     232508  232508  232466  232508  232466
unique          65647      322          6       4       3      11   15080
top     /destinos-get  modules  HTTP/1.1"    HTTP    1.1"     200       -
freq             8115    63447     226612  232458  226612  209952   16160

                                                 buscador     vlog
count                                               84311    84311
```

```
unique                                                   515        1
top     Mozilla/5.0 (compatible; Googlebot/2.1; +http:...   VLOG=-
freq                                                   50827    84311

[4 rows x 21 columns]
pagWeb       object
ip           object
data         object
dia          object
mes          object
any          object
hora         object
hores        object
minuts       object
segons       object
UTC          object
optionGet    object
adreçaWeb    object
llocWeb      object
HTTP         object
HTTP2        object
HTTPnum      object
numeroA      object
numeroB      object
buscador     object
vlog         object
dtype: object
```

In [22]:

```
# Em vull quedr només amb la columna de IP
colIP = datafFiltrat["ip"]

print(colIP)
```

```
21          66.249.76.216
22          66.249.76.216
23          66.249.76.216
24          66.249.76.216
25          66.249.76.216
               ...
261860      66.249.76.216
261861       5.255.253.53
261865       5.255.253.53
261867       5.255.253.53
261868      74.86.158.107
Name: ip, Length: 232508, dtype: object
```

In [23]:

```
# Vull saber els valors que no es repeteixen per reduir la basa de dades
pd.unique(colIP)
```

Out[23]:

```
array([' 66.249.76.216 ', ' 5.255.253.53 ', ' 157.55.35.112 ', ...,
       ' 217.174.248.179 ', ' 188.135.173.80 ', ' 206.198.5.33 '],
      dtype=object)
```

In [24]:

```
# I eliminar els valors que es repeteixen i convertir-los en un nou dataframe

colIP=list(set(colIP))
colIPunique = pd.DataFrame(colIP)
colIPunique.columns = ["IP"]
colIPunique.columns.str.strip()

print(colIPunique)

#Quant l'imprimim veiem que hem passat d'una columna de mida 232508 a una de 2827
```

```
                     IP
0         88.31.131.220
1         77.226.207.26
2       220.181.108.78
3         199.30.20.214
4          5.153.20.133
...                 ...
2822       180.76.5.193
2823        180.76.5.39
2824     186.23.137.154
2825       79.80.215.41
2826      163.247.51.11

[2827 rows x 1 columns]
```

In [25]:

```
# Guardo aquest dataset reduit en format csv i que és el que utilitzaré a IPinfo

colIPunique.to_csv('ips.csv')
```

In [26]:

```
# Importem la llibreria ip2geotools
from ip2geotools.databases.noncommercial import DbIpCity
import time


colIPunique = colIPunique.loc[:50,['IP']]
def IP_info_2(ip):
    try:
        return DbIpCity.get(ip, api_key = 'free').country
    except:
        return np.nan
colIPunique['País'] = colIPunique.apply(IP_info_2)

colIPunique.loc[:50,:]

#no sé perquè no em llegeix el país
```

Out[26]:

	IP	País
0	88.31.131.220	NaN
1	77.226.207.26	NaN
2	220.181.108.78	NaN
3	199.30.20.214	NaN
4	5.153.20.133	NaN
5	190.191.79.143	NaN
6	157.55.32.145	NaN
7	195.55.66.182	NaN
8	185.10.104.196	NaN
9	199.30.20.207	NaN
10	63.147.126.185	NaN
11	213.192.208.181	NaN
12	46.24.158.105	NaN
13	180.76.5.203	NaN

	IP	País
14	89.7.246.98	NaN
15	66.249.75.198	NaN
16	85.152.161.241	NaN
17	220.181.108.99	NaN
18	80.34.64.158	NaN
19	180.76.6.14	NaN
20	81.45.53.165	NaN
21	80.27.103.200	NaN
22	199.30.20.3	NaN
23	202.46.52.29	NaN
24	176.57.141.193	NaN
25	46.39.212.184	NaN
26	66.249.76.213	NaN
27	114.26.127.57	NaN
28	86.221.241.137	NaN
29	95.23.142.9	NaN
30	88.138.178.147	NaN
31	79.108.34.74	NaN
32	31.4.199.114	NaN
33	81.45.52.91	NaN
34	109.211.53.111	NaN
35	180.76.5.195	NaN
36	46.222.103.220	NaN
37	217.12.16.130	NaN
38	89.140.9.2	NaN
39	80.30.133.95	NaN
40	74.80.133.229	NaN
41	202.46.62.29	NaN
42	89.3.95.233	NaN
43	213.37.144.69	NaN
44	87.223.105.21	NaN
45	180.76.6.159	NaN
46	190.88.176.120	NaN
47	157.55.33.124	NaN
48	217.216.120.71	NaN
49	31.4.190.164	NaN

	IP	País
50	217.127.176.170	NaN

Creació del mapa utilitzant ipinfo

https://ipinfo.io/tools/map/786b9832-09e4-44b9-a8cf-70ac4eaa726e

I amb la funció següent ens fa un resum molt més ocmplert de les IPs

https://ipinfo.io/tools/summarize-ips/98232ee3-cb95-40d9-9402-3199c74f1004

NIVELL 3

Exercici 4

Mostra'm la teva creativitat, Sorprèn-me fes un pas més enllà amb l'anàlisi anterior.

In [ ]: