1. Note: It's unclear from the data what "Size" means.
   If "Size" is the number of requests, the answer is 4935

```
log_df.groupby(['Date', 'ResponseCode']).get_group(('30/Apr/1998', 404))['Size'].sum()
```

```
4935.0
```

If size is something else (like the number of bytes of an individual request) the answer is 17

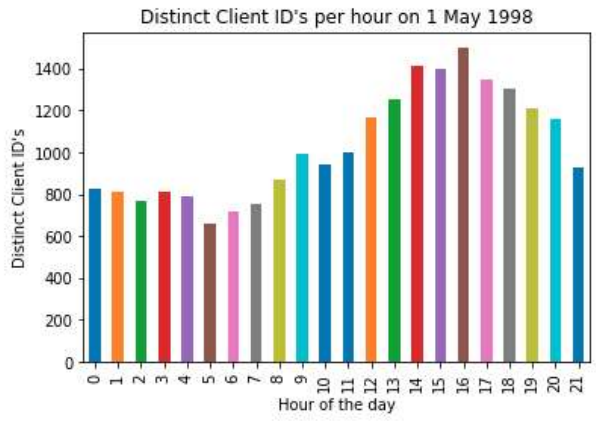
```
log_df.groupby(['Date', 'ResponseCode']).get_group(('30/Apr/1998', 404)).info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17 entries, 275 to 197200
Data columns (total 9 columns):
ClientID        17 non-null int64
Date            17 non-null object
Time            17 non-null object
URL             17 non-null object
ResponseCode    17 non-null float64
Size            17 non-null float64
DateTime        17 non-null datetime64[ns]
hour            17 non-null int64
URL_type        17 non-null object
dtypes: datetime64[ns](1), float64(2), int64(2), object(4)
memory usage: 1.3+ KB
```

2.

```
ax = log_df[log_df['Date'] == '01/May/1998'].groupby('hour')['ClientID'].nunique().plot(kind='bar')
ax.set_ylabel("Distinct Client ID's")
ax.set_xlabel("Hour of the day")
ax.set_title("Distinct Client ID's per hour on 1 May 1998")
```

```
Text(0.5,1,"Distinct Client ID's per hour on 1 May 1998")
```



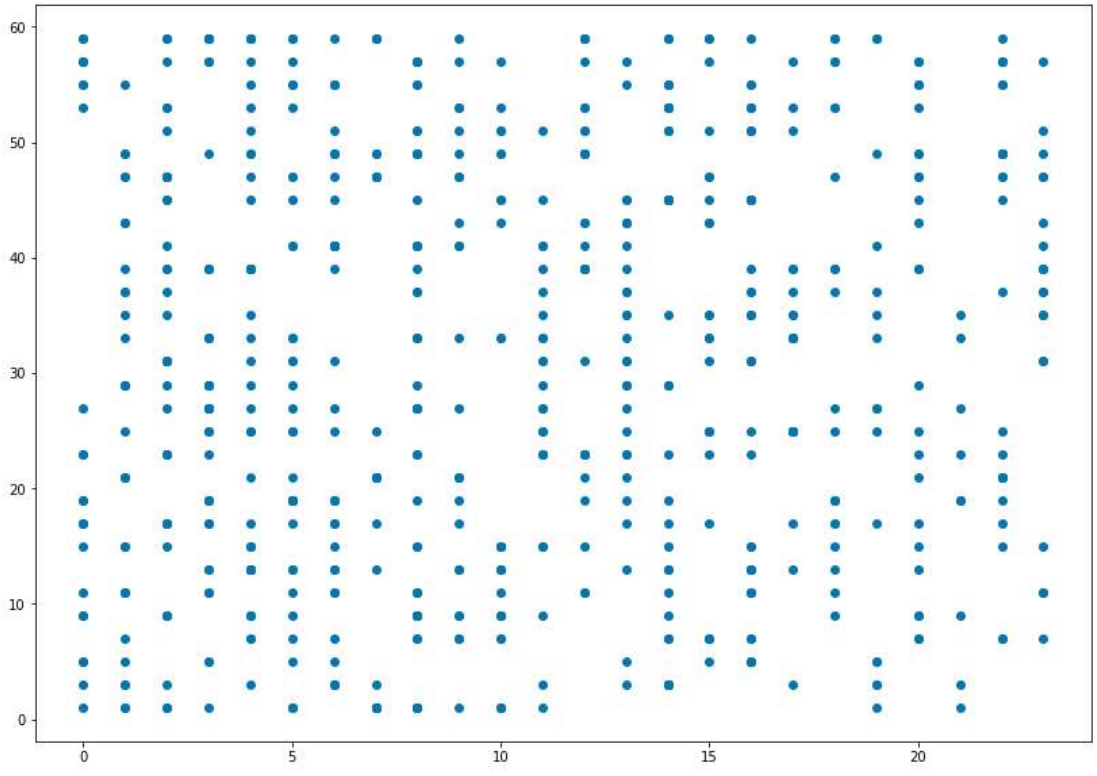
3.

```
# Checking to see if there are any interesting correlations in the data
corr = log_df.corr()
corr.style.background_gradient(cmap='coolwarm').set_precision(2)
```

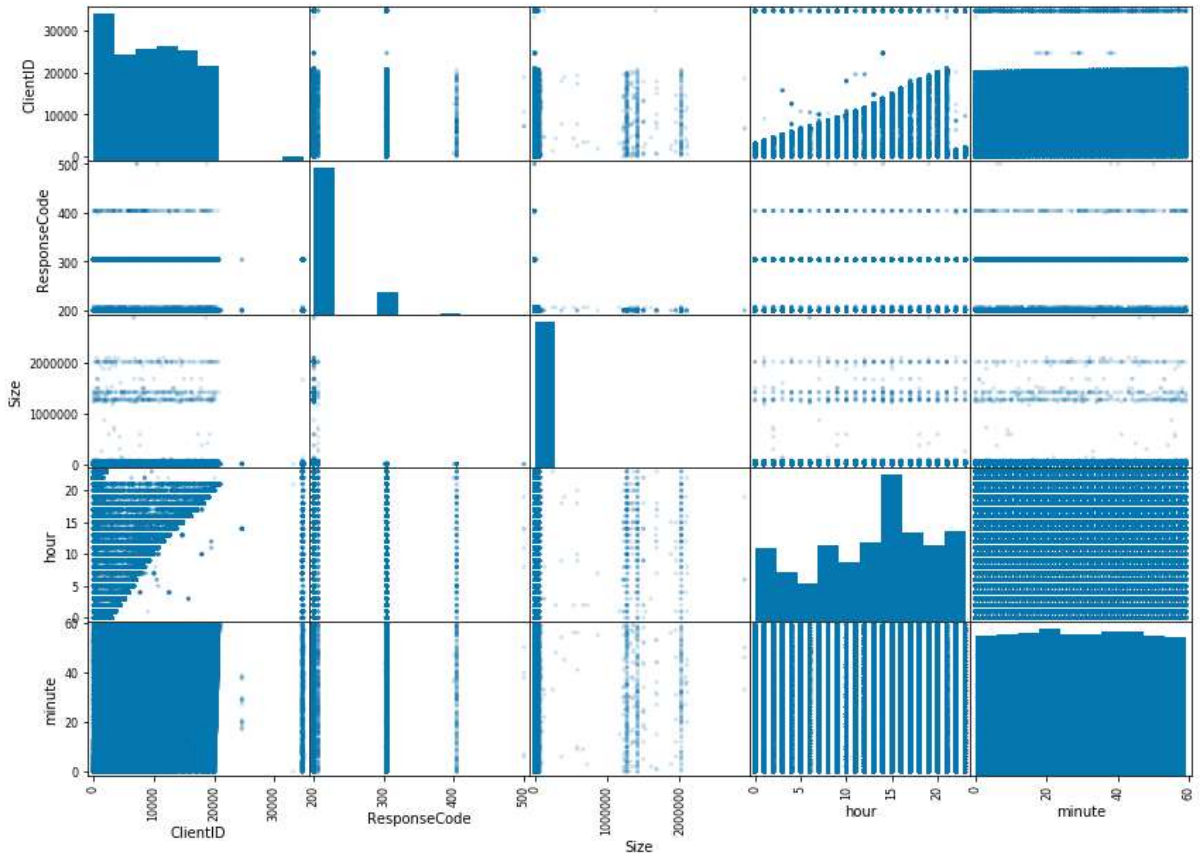
	ClientID	ResponseCode	Size	hour
ClientID	1	-0.057	-0.013	0.41
ResponseCode	-0.057	1	-0.03	-0.0073
Size	-0.013	-0.03	1	-0.0024
hour	0.41	-0.0073	-0.0024	1

```
# Looks like ClientID and hour are decently positively correlated
# (I didn't consider this beforehand, but it does make perfect sense)
# Looking at the plot of when the most popular client ID made requests to see if any patterns emerge

log_df['minute'] = log_df.DateTime.dt.minute
plt.figure(figsize=(14,10))
most_popular = log_df[log_df['ClientID'] == log_df['ClientID'].mode()[0]]
plt.scatter(data=most_popular, x='hour', y='minute')
```

```
<matplotlib.collections.PathCollection at 0x7f36e229e550>
```



```
# Doesn't look like this particular entity had any meaningful patterns
# Taking a step back, I looked at a scatter matrix

pd.plotting.scatter_matrix(log_df, alpha=0.2, figsize=(14,10))
```

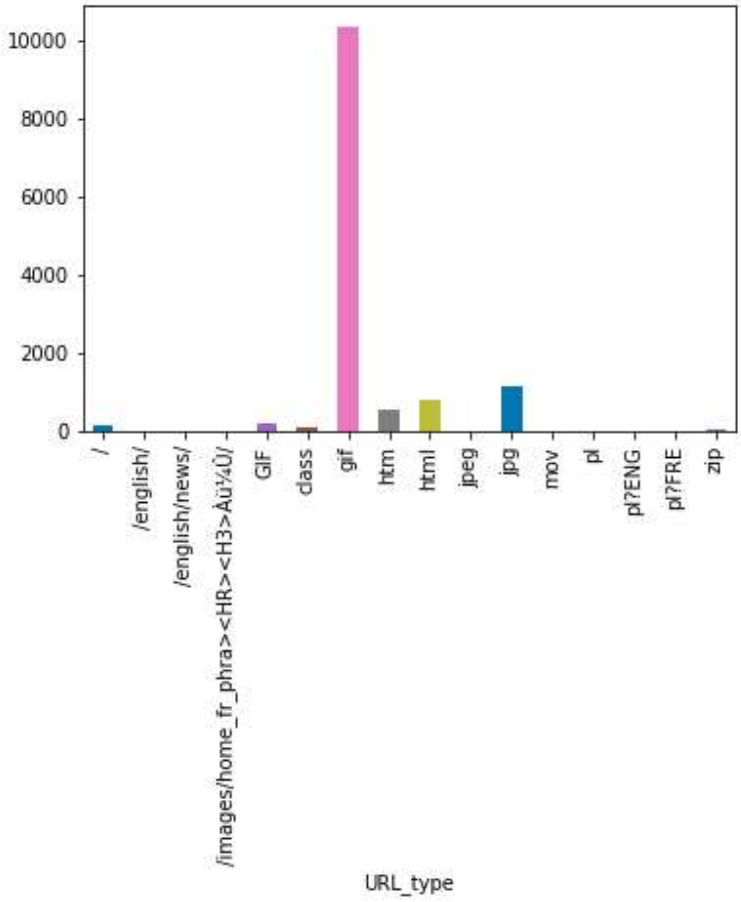


```
# One expected thing that stood out is that client ID's increased as hour increased,
# suggesting that they were issued sequentially
# One unexpected thing is that there looks to be a spike around hour 15

log_df.groupby('hour').size()
```

```
hour
0       6569
1       6103
2       6072
3       6625
4       6019
5       4733
6       4995
7       5094
8       6460
9       7892
10      7465
11      7893
12     10127
13     10225
14     12040
15     12256
16     13367
17     11494
18     11515
19     10386
20      9363
21      6610
22      8658
23      8039
```

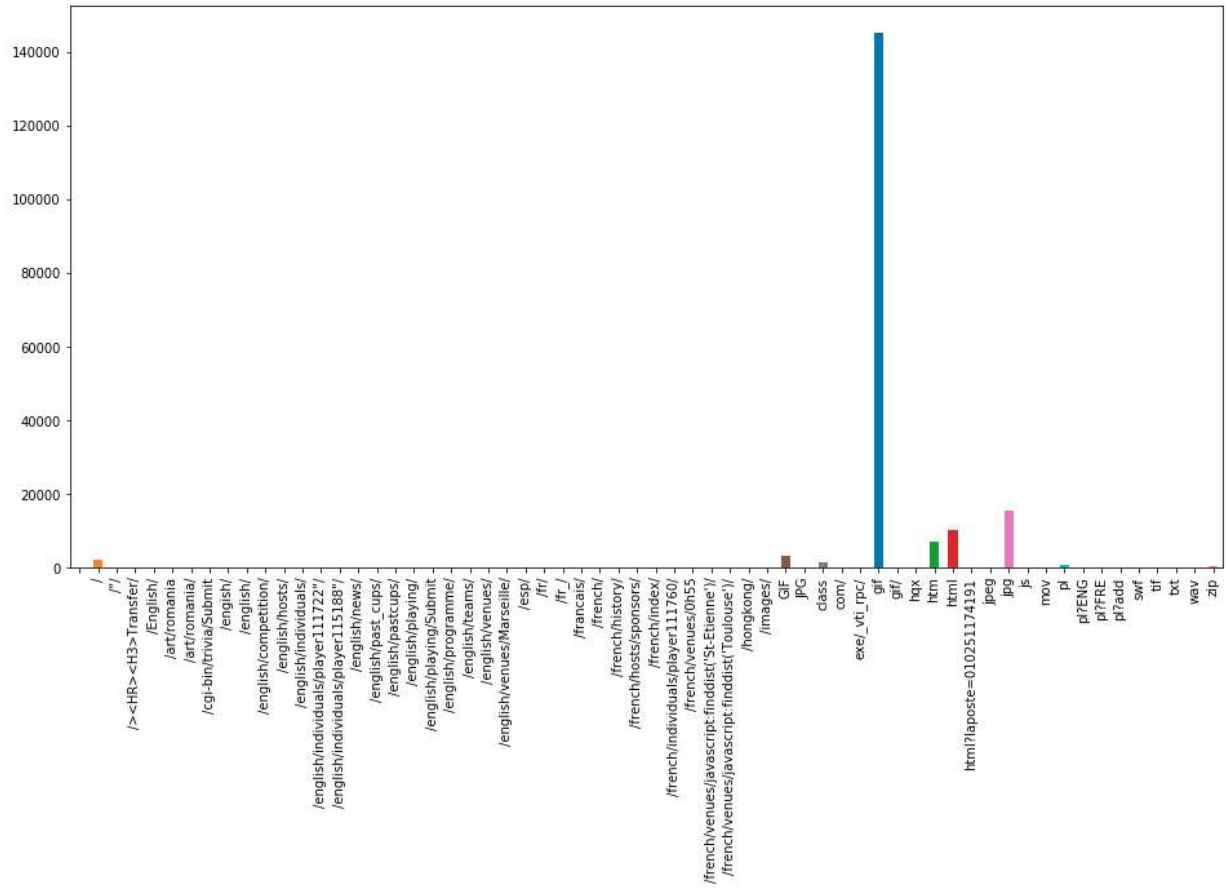
```
# The spike doesn't appear to be as pronounced here as in the chart
# Which is weird
# But just to close this out, lets look at the type of requests during this hour

hr16 = log_df[log_df['hour'] == 16]
hr16.groupby('URL_type').size().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f36dcdeeda0>
```



```
# It looks like gifs were really popular during this hour
# Lets compare that to the rest of the hours

not_hr16 = log_df[log_df['hour'] != 16]
not_hr16.groupby('URL_type').size().plot(kind='bar', figsize=(16,8))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f36dc9dd518>
```




```
# Sadly (at least in terms of interestingness) Hour 16 is a typical hour, just one with a higher volume
# Though if nothing else, it is interesting that traffic was absolutely dominated by gifs
# With only a few other types of requests occuring often enough to even appear on the graph
# So long story short, people really like gifs, especially in the afternoons
```