In [2]:
```
import pylab
import pandas as pd
```

In [206]:
```
log_df = pd.read_csv("wc_day6_1_sample.csv",
                     names=['ClientID', 'Date', 'Time', 'URL', 'Respon
seCode', 'Size'],
                     na_values=['-'])

log_df['Size'].median()
```

Out[206]:
```
914.0
```

In [207]:
```
grouped = log_df.groupby('ResponseCode')
```

In [208]:
```
%matplotlib inline
```

In [16]:
```
import matplotlib.pyplot as pp
```

In [210]:
```
may1_df = log_df[log_df['Date'] == '01/May/1998']
may1_df['DateTime'] = pd.to_datetime(may1_df.apply(lambda row: row['Da
te'] + ' ' + row['Time'], axis=1))
hour_grouped = may1_df.groupby(lambda x: may1_df['DateTime'][x].hour)
```

In [211]:
```
hour_grouped.size()
```

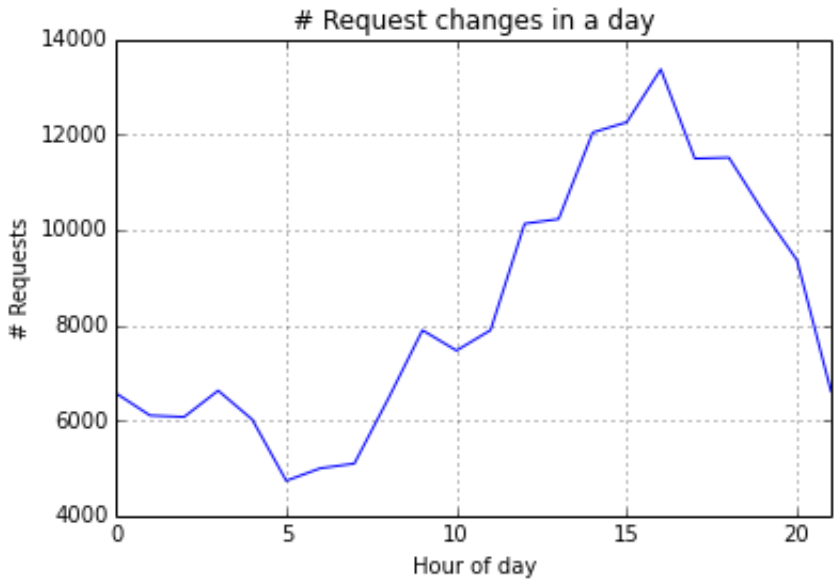
Out[211]:
```
0      6569
1      6103
2      6072
3      6625
4      6019
5      4733
6      4995
7      5094
8      6460
9      7892
10     7465
11     7893
12    10127
13    10225
14    12040
15    12256
16    13367
17    11494
18    11515
19    10386
20     9363
21     6610
dtype: int64
```

In [212]:
```
hour_grouped['Size'].sum()
```

Out[212]:
```
0     44166352
1     46857868
2     42803283
3     38868040
4     49190470
5     34184105
6     47877742
7     37838488
8     57224306
9     67645841
10    64193518
11    59961757
12    79150391
13    80907946
14    98825640
15    94044070
16    73413868
17    94389754
18    79264404
19    76209823
20    67784666
21    59834046
Name: Size, dtype: float64
```

In [213]:
```
ax = hour_grouped.size().plot()
ax.set_ylabel("# Requests")
ax.set_xlabel("Hour of day")
ax.set_title("# Request changes in a day")
```

Out[213]:
```
<matplotlib.text.Text at 0x7f1ff600b790>
```



In [214]:
```
multi_grouped = log_df.groupby(['ResponseCode', 'Date'])
group = multi_grouped.get_group((404,'30/Apr/1998'))
```

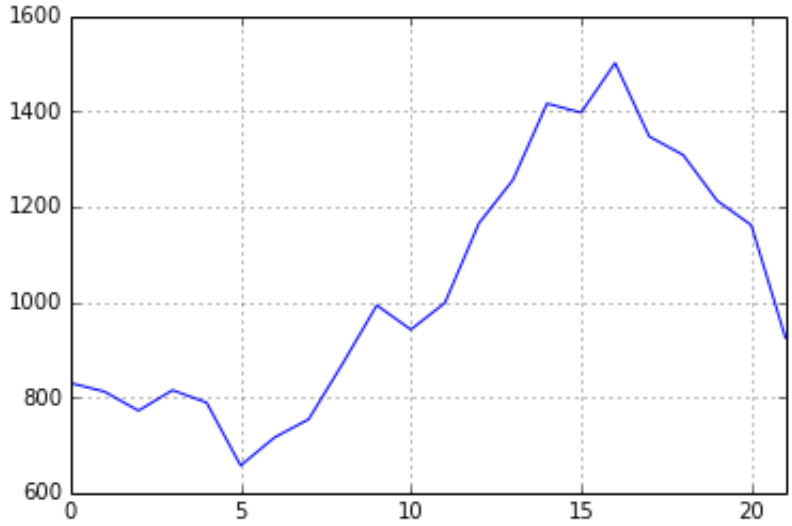
In [215]:
```
################### Question 1
len(group.index)
```

Out[215]: 17

In [216]:
```
##### Finding Number of Unique Users
uniqid = hour_grouped['ClientID'].nunique()
```

In [217]:
```
###################### Question 2 :- Plotting unique number of user for every hour on "1/May/1998"
uniqid.plot()
```

Out[217]: <matplotlib.axes.AxesSubplot at 0x7f1ff6063650>



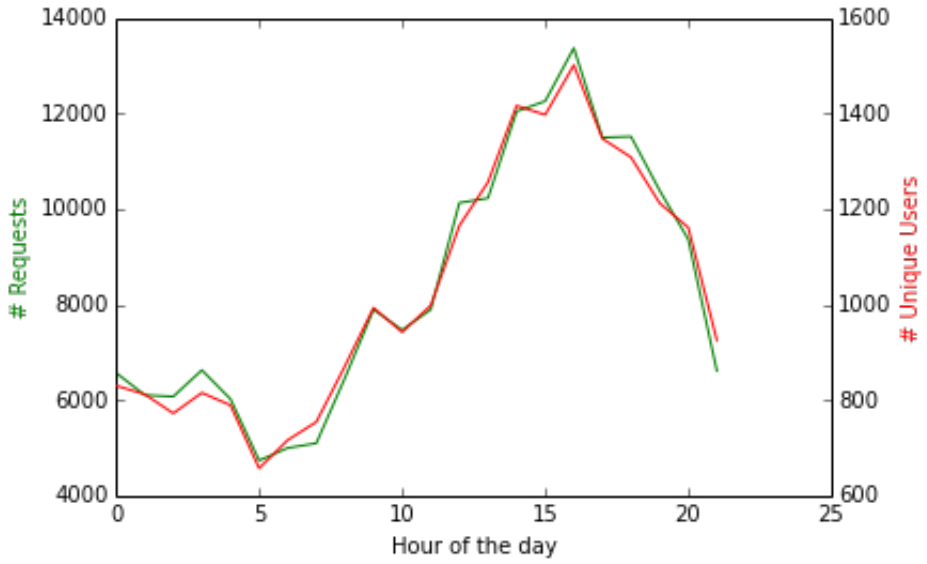
In [218]:
```
######## Question 3 :- Correlation graph of number of Requests to numb
er of Unique users
##### It is clear through the graph that there is a strong postive cor
elation between
#####the users and the Requests i.e As the number of
##### users increase during the day, the number of requests also incre
ases.

fig, ax1 = pp.subplots()
ax2 = ax1.twinx()
x = hour_grouped.size().index

ax1.plot(x, hour_grouped.size(), 'g-')
ax2.plot(x, uniqid, 'r-')

ax1.set_xlabel('Hour of the day')
ax1.set_ylabel('# Requests', color='g')
ax2.set_ylabel('# Unique Users', color='r')
```

Out[218]: <matplotlib.text.Text at 0x7f1ff5efa7d0>

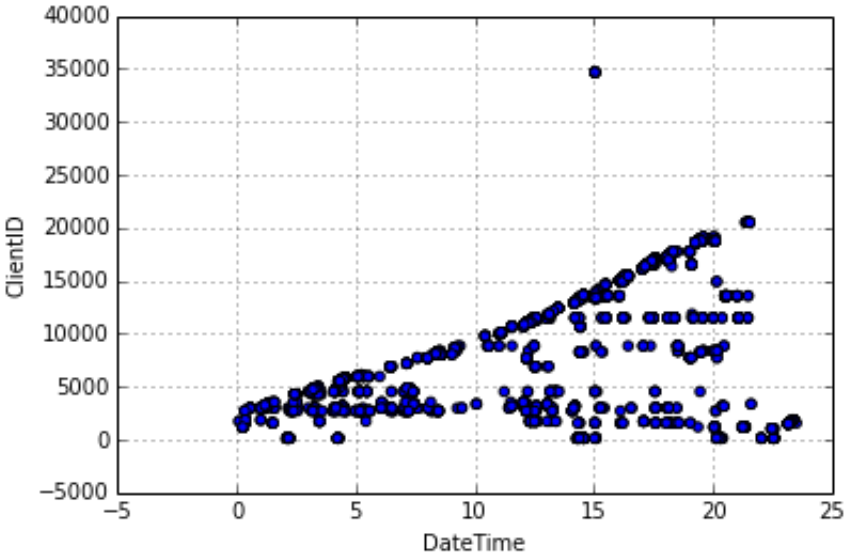


In [6]:
```
log_df = pd.read_csv("wc_day6_1_sample.csv",
                     names=['ClientID', 'Date', 'Time', 'URL', 'Respon
seCode', 'Size'],
                     na_values=['-'])
uniqueClientIds = log_df['ClientID'].unique()[0:100]
newDF = log_df[log_df['ClientID'].isin(uniqueClientIds)]
```

In [7]:
```
########## Question 4 :- From the below graph it is visible that there
is corelation between the client-id
##########and the time at which they Visit the websites. As the hours
of the day progress we
####### see that users with higher client-id starting to visit the web
site. This
###### gives a useful insight about theviewing habbits of the users.

% matplotlib inline
newDF['DateTime'] = newDF.apply(lambda row: float(row['Time'].split(":
")[0] + "." + row['Time'].split(":")[1]), axis=1)
newDF.plot(kind='scatter',x='DateTime', y='ClientID')
```

Out[7]: <matplotlib.axes.AxesSubplot at 0x7fdbdad74ed0>



In [9]:
```
log_df = pd.read_csv("wc_day91_2.csv",
                     names=['ClientID', 'Date', 'Time', 'URL', 'Respon
seCode', 'Size'],
                     na_values=['-'])
```

In [10]:
```
############ Making necessary corrections to the data and producing th
e data which would be useful
log_df = log_df.drop("Time",1)
log_df['Time'] = log_df.apply(lambda row: row['Date'].split(":")[1] +
":" +  row['Date'].split(":")[2]  + ":"+ row['Date'].split(":")[3], ax
is=1)
```

In [11]:
```
log_df['Date'] = log_df.apply(lambda row: row['Date'].split(":")[0].re
place("[","") , axis=1)
log_df['URL'] =  log_df.apply(lambda row: row['URL'].replace(","," ")
, axis=1)
```

In [225]:
```
log_df
```

Out[225]:

	ClientID	Date	URL	Response

0	2743832	24/Jul/1998	GET /english/history/body.html HTTP/1.1	200
1	2572248	24/Jul/1998	GET / HTTP/1.0	200
2	31798	24/Jul/1998	GET /french/competition/maincomp.htm HTTP/1.0	200
3	1848501	24/Jul/1998	GET / HTTP/1.0	200
4	248	24/Jul/1998	GET /images/home_intro.anim.gif HTTP/1.0	200
5	2742956	24/Jul/1998	GET /french/history/images/history_hm_nav.gif ...	304
6	299067	24/Jul/1998	GET /english/images/news_btn_part_off.gif HTTP...	304
7	2033693	24/Jul/1998	GET /french/images/nav_venue_off.gif HTTP/1.0	200
8	2560	24/Jul/1998	GET /french/images/hm_top_stories_head.gif HTT...	200
9	65455	24/Jul/1998	GET /images/hm_ligne1_col2.gif HTTP/1.1	200
10	2033693	24/Jul/1998	GET /french/images/nav_team_off.gif HTTP/1.0	200
11	415336	24/Jul/1998	GET /images/home_intro.anim.gif HTTP/1.1	200
12	65455	24/Jul/1998	GET /images/hm_ligne1_col3.gif HTTP/1.1	200
13	2630107	24/Jul/1998	GET /images/acc_welcome_f.gif HTTP/1.0	200
14	65455	24/Jul/1998	GET /images/hm_ligne2_col1.gif HTTP/1.1	200
15	2150066	24/Jul/1998	GET /images/comp_bg2_hm.gif HTTP/1.0	404
16	65455	24/Jul/1998	GET /images/hm_hola.gif HTTP/1.1	200
17	2630107	24/Jul/1998	GET /images/acc_anime.gif HTTP/1.0	200
18	2630107	24/Jul/1998	GET /images/acc_welcome_e.gif HTTP/1.0	200
19	520440	24/Jul/1998	GET /images/102383s.gif HTTP/1.0	200
20	2743826	24/Jul/1998	GET /english/individuals/playerphoto75952_1.ht...	200
21	73721	24/Jul/1998	GET /images/s102329.gif HTTP/1.0	200
22	310517	24/Jul/1998	GET /english/images/nav_field_off.gif HTTP/1.1	200
23	2743832	24/Jul/1998	GET /english/history/images/history_hm_bg2.jpg...	200
24	2743832	24/Jul/1998	GET /images/space.gif HTTP/1.1	200

25	2743832	24/Jul/1998	GET /english/history/images/history_hm_3094.gi...	200
26	531	24/Jul/1998	GET / HTTP/1.0	200

In [12]:
```
may1_df = log_df[log_df['Date'] == '24/Jul/1998']
may1_df['DateTime'] = pd.to_datetime(may1_df.apply(lambda row: row['Da
te'] + ' ' + row['Time'], axis=1))
hour_grouped = may1_df.groupby(lambda x: may1_df['DateTime'][x].hour)

may2_df = log_df[log_df['Date'] == '25/Jul/1998']
may2_df['DateTime'] = pd.to_datetime(may2_df.apply(lambda row: row['Da
te'] + ' ' + row['Time'], axis=1))
hour_grouped2 = may2_df.groupby(lambda x: may2_df['DateTime'][x].hour)
```

In [20]:
```
######Getting Unique users
uniqid = hour_grouped['ClientID'].nunique()
uniqid2 = hour_grouped2['ClientID'].nunique()
```

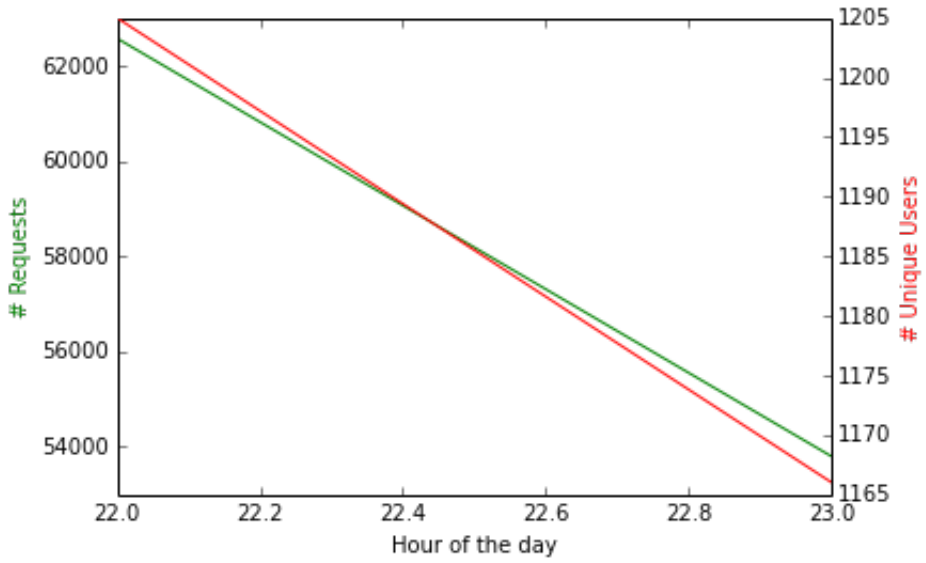
In [14]:
```
% matplotlib inline
```

In [17]:
```
#######Question 5 :- Part 1 :- a. 24th July Here also we can see a pos
itive corelation between
####### the number of unique users and the number of requests being
##### generated per hour. As the number of users decrease we can see t
hat the number of requests also decrease linearly.
fig, ax1 = pp.subplots()
ax2 = ax1.twinx()
x = hour_grouped.size().index

ax1.plot(x, hour_grouped.size(), 'g-')
ax2.plot(x, uniqid, 'r-')

ax1.set_xlabel('Hour of the day')
ax1.set_ylabel('# Requests', color='g')
ax2.set_ylabel('# Unique Users', color='r')
```

Out[17]: <matplotlib.text.Text at 0x7fdbdbf1e0d0>

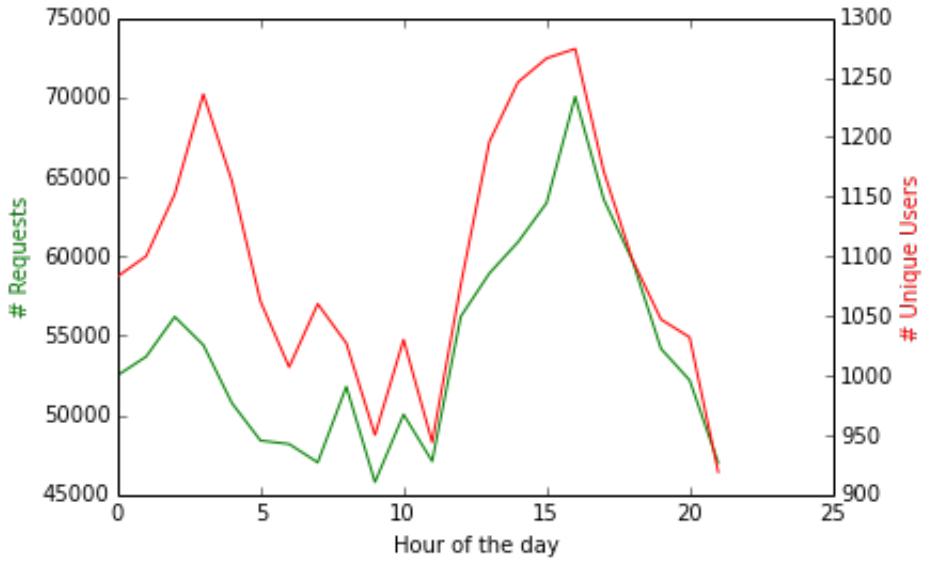


In [21]:
```
####### Question 5 :- Part 1 :- a. 25th July Here also we can see a po
sitive corelation between the number of unique users
####### and the number of requests beingvgenerated per hour. As the nu
mber of users increase and decrease during the day
####### we can see that the number of requests also increassedecrease
in propotion.
fig, ax1 = pp.subplots()
ax2 = ax1.twinx()
x = hour_grouped2.size().index

ax1.plot(x, hour_grouped2.size(), 'g-')
ax2.plot(x, uniqid2, 'r-')

ax1.set_xlabel('Hour of the day')
ax1.set_ylabel('# Requests', color='g')
ax2.set_ylabel('# Unique Users', color='r')
```

Out[21]:
```
<matplotlib.text.Text at 0x7fdbdaf76110>
```



In [230]:
```
uniqueClientIds = log_df['ClientID'].unique()[0:100]
newDf = log_df[log_df['ClientID'].isin(uniqueClientIds)]
```

In [232]:
```
###### Question 5 Part 2 :-  Here we can see that activity patterns ba
sed on the id's of the users. This we can
##### corelate with the fact users with certain group of id's tend to
browse in certain specific patterns during
##### the day. The results of quesstion 4 and 5 show how that there is
a positive corelation in the graphs in both
##### cases.The number of users
##### tend to increase at certain hours of the day which leads to ince
arsed server activity.

% matplotlib inline
newDF['DateTime'] = newDF.apply(lambda row: float(row['Time'].split(":
")[0] + "." + row['Time'].split(":")[1]), axis=1)
newDF.plot(kind='scatter',x='DateTime', y='ClientID')
```

Out[232]: <matplotlib.axes.AxesSubplot at 0x7f1ff53551d0>