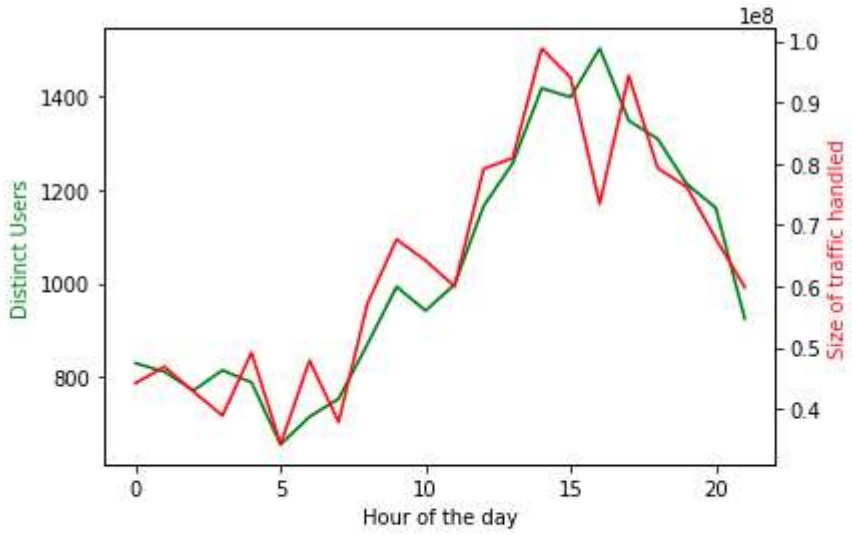
1.

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

ax1.plot(x, hour_grouped['ClientID'].nunique(), 'g-')
ax2.plot(x, hour_grouped['Size'].sum(), 'r-')

ax1.set_xlabel('Hour of the day')
ax1.set_ylabel('Distinct Users', color='g')
ax2.set_ylabel('Size of traffic handled', color='r')
```

```
Text(0,0.5,'Size of traffic handled')
```




```
hour_grouped['ClientID'].nunique().corr(hour_grouped['Size'].sum())
```

```
0.92580977471161385
```

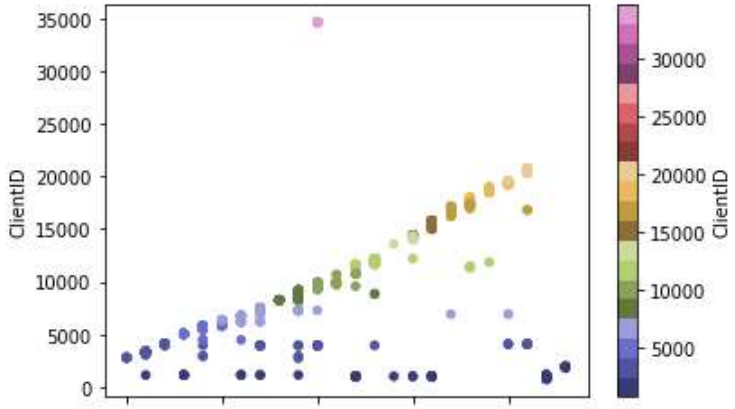
The data is strongly positively correlated, with an r value of 0.926.

2.

```
import numpy as np
unique_clients = log_df['ClientID'].unique()
idxs = np.random.choice(unique_clients.shape[0], 100, replace=False)
clients = unique_clients[idxs]
client_sample = log_df[log_df['ClientID'].isin(clients)]

client_sample.plot(kind='scatter', x='hour', y='ClientID', c='ClientID', cmap='tab20b')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f25162fc0f0>
```



3.

```
log_df_new = pd.read_csv("./wc_day91_1_log.csv",
                     names=['ClientID', 'Date', 'Time', 'URL', 'ResponseCode', 'Size'],
                     na_values=['-'],
                     encoding = "ISO-8859-1")
log_df_new.loc[:,('DateTime')] = pd.to_datetime(log_df_new.apply(lambda row: row['Date'] + ' ' + row['Time'], axis=1))
log_df_new['hour'] = log_df_new.DateTime.dt.hour
log_df_new['URL_type'] = log_df_new['URL'].str.split(".", n = -1, expand = False).str.get(-1)
```

```
/usr/lib/python3/dist-packages/IPython/core/interactiveshell.py:2718: DtypeWarning: Columns (4) have mixed types. Specify dtype
option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

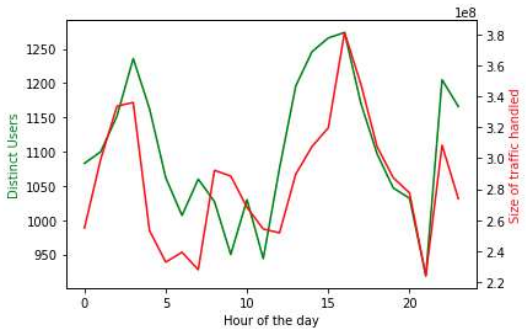
```
hour_grouped_new = log_df_new.groupby(lambda x: log_df_new['DateTime'][x].hour)
```

```
fig, ax1 = plt.subplots()
ax2 = ax1.twinx()
x = hour_grouped_new.size().index

ax1.plot(x, hour_grouped_new['ClientID'].nunique(), 'g-')
ax2.plot(x, hour_grouped_new['Size'].sum(), 'r-')

ax1.set_xlabel('Hour of the day')
ax1.set_ylabel('Distinct Users', color='g')
ax2.set_ylabel('Size of traffic handled', color='r')
```

```
Text(0,0.5,'Size of traffic handled')
```




```
hour_grouped_new['ClientID'].nunique().corr(hour_grouped_new['Size'].sum())
```

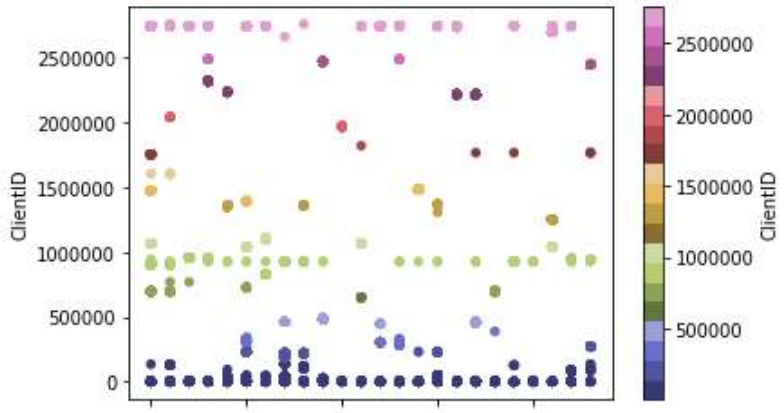
```
0.690702736940218
```

Correlation is still strong and positive, but it is much less strong than the first dataset.

```
unique_clients = log_df_new['ClientID'].unique()
idxs = np.random.choice(unique_clients.shape[0], 100, replace=False)
clients = unique_clients[idxs]
client_sample = log_df_new[log_df_new['ClientID'].isin(clients)]

client_sample.plot(kind='scatter', x='hour', y='ClientID', c='ClientID', cmap='tab20b')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f24fa631780>
```



Overall, there were a few major differences between the two datasets. As mentioned previously, the correlation between traffic size and number of distinct users is much weaker in the second data set, though it is still a strong, positive correlation. Additionally, the first dataset showed a fairly distinct pattern of increasing ClientID numbers as the day progressed. This pattern is nowhere to be found in the second dataset. In this second dataset, there appears to be a few super users who accessed the site throughout the day, but apart from that, access appears to be much more randomized in terms of ClientID's.