```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

url = "https://raw.githubusercontent.com/sarabjeet050/Data-Science-
Project/00f16b0c80e9b2bd0b5d84ab4661a931d3c0d761/
Comcast_telecom_complaints_data.csv"
df = pd.read_csv(url)

df.head()
```

```
   Ticket #                          Customer Complaint
Date  \
0   250635                  Comcast Cable Internet Speeds  22-04-
15
1   223441      Payment disappear - service got disconnected  04-08-
15
2   242732                            Speed and Service  18-04-
15
3   277946  Comcast Imposed a New Usage Cap of 300GB that ...  05-07-
15
4   307175          Comcast not working and no service to boot  26-05-
15

   Date_month_year          Time        Received Via      City      State
\
0       22-Apr-15   3:53:50 PM  Customer Care Call  Abingdon  Maryland

1       04-Aug-15  10:22:56 AM            Internet   Acworth   Georgia

2       18-Apr-15   9:55:47 AM            Internet   Acworth   Georgia

3       05-Jul-15  11:59:35 AM            Internet   Acworth   Georgia

4       26-May-15   1:25:26 PM            Internet   Acworth   Georgia


   Zip code  Status Filing on Behalf of Someone
0     21009  Closed                          No
1     30102  Closed                          No
2     30101  Closed                         Yes
3     30101    Open                         Yes
4     30101  Solved                          No
```

```python
df.isna().sum()
```

```
Ticket #                     0
Customer Complaint           0
```

```
Date                           0
Date_month_year                0
Time                           0
Received Via                   0
City                           0
State                          0
Zip code                       0
Status                         0
Filing on Behalf of Someone    0
dtype: int64
```

df.describe(include='all').T

	count	unique	top	freq
Ticket #	2224	2224	250635	1
Customer Complaint	2224	1841	Comcast	83
Date	2224	91	24-06-15	218
Date_month_year	2224	91	24-Jun-15	218
Time	2224	2190	12:41:14 PM	2
Received Via	2224	2	Customer Care Call	1119
City	2224	928	Atlanta	63
State	2224	43	Georgia	288
Zip code	2224.0	NaN	NaN	NaN
Status	2224	4	Solved	973
Filing on Behalf of Someone	2224	2	No	2021

	mean	std	min	25%
Ticket #	NaN	NaN	NaN	NaN
Customer Complaint	NaN	NaN	NaN	NaN
Date	NaN	NaN	NaN	NaN
Date_month_year	NaN	NaN	NaN	NaN
Time	NaN	NaN	NaN	NaN
Received Via	NaN	NaN	NaN	

NaN			
City	NaN	NaN	NaN
NaN			
State	NaN	NaN	NaN
NaN			
Zip code	47994.393435	28885.279427	1075.0
30056.5			
Status	NaN	NaN	NaN
NaN			
Filing on Behalf of Someone	NaN	NaN	NaN
NaN			

	50%	75%	max
Ticket #	NaN	NaN	NaN
Customer Complaint	NaN	NaN	NaN
Date	NaN	NaN	NaN
Date_month_year	NaN	NaN	NaN
Time	NaN	NaN	NaN
Received Via	NaN	NaN	NaN
City	NaN	NaN	NaN
State	NaN	NaN	NaN
Zip code	37211.0	77058.75	99223.0
Status	NaN	NaN	NaN
Filing on Behalf of Someone	NaN	NaN	NaN

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2224 entries, 0 to 2223
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Ticket #                    2224 non-null   object
 1   Customer Complaint          2224 non-null   object
 2   Date                        2224 non-null   object
 3   Date_month_year             2224 non-null   object
 4   Time                        2224 non-null   object
 5   Received Via                2224 non-null   object
 6   City                        2224 non-null   object
 7   State                       2224 non-null   object
 8   Zip code                    2224 non-null   int64
 9   Status                      2224 non-null   object
 10  Filing on Behalf of Someone  2224 non-null   object
dtypes: int64(1), object(10)
memory usage: 191.2+ KB
```

```
df.shape
```

```
(2224, 11)
```

```python
# Provide the trend chart for the number of complaints at monthly and
daily granularity levels.

df['Date_month_year'] = df['Date_month_year'].apply(pd.to_datetime)

df = df.set_index('Date_month_year')

df.head()
```

```
                 Ticket #                            Customer
Complaint  \
Date_month_year

2015-04-22        250635                Comcast Cable Internet
Speeds
2015-08-04        223441      Payment disappear - service got
disconnected
2015-04-18        242732                            Speed and
Service
2015-07-05        277946  Comcast Imposed a New Usage Cap of 300GB
that ...
2015-05-26        307175      Comcast not working and no service to
boot

                     Date       Time        Received Via
City  \
Date_month_year

2015-04-22        22-04-15   3:53:50 PM  Customer Care Call   Abingdon

2015-08-04        04-08-15  10:22:56 AM            Internet    Acworth

2015-04-18        18-04-15   9:55:47 AM            Internet    Acworth

2015-07-05        05-07-15  11:59:35 AM            Internet    Acworth

2015-05-26        26-05-15   1:25:26 PM            Internet    Acworth


                   State  Zip code  Status Filing on Behalf of
Someone
Date_month_year

2015-04-22        Maryland     21009  Closed
No
2015-08-04         Georgia     30102  Closed
No
2015-04-18         Georgia     30101  Closed
Yes
2015-07-05         Georgia     30101    Open
Yes
```

```
2015-05-26        Georgia      30101  Solved
No
```

```python
#Plotting monthly trend chart

plt.figure(figsize=(9,7))
month = df.groupby(pd.Grouper(freq="M")).size().plot()
plt.xlabel("MONTH")
plt.ylabel("FREQUENCY")
plt.title("MONTHLY TREND CHART")
```

```
Text(0.5, 1.0, 'MONTHLY TREND CHART')
```
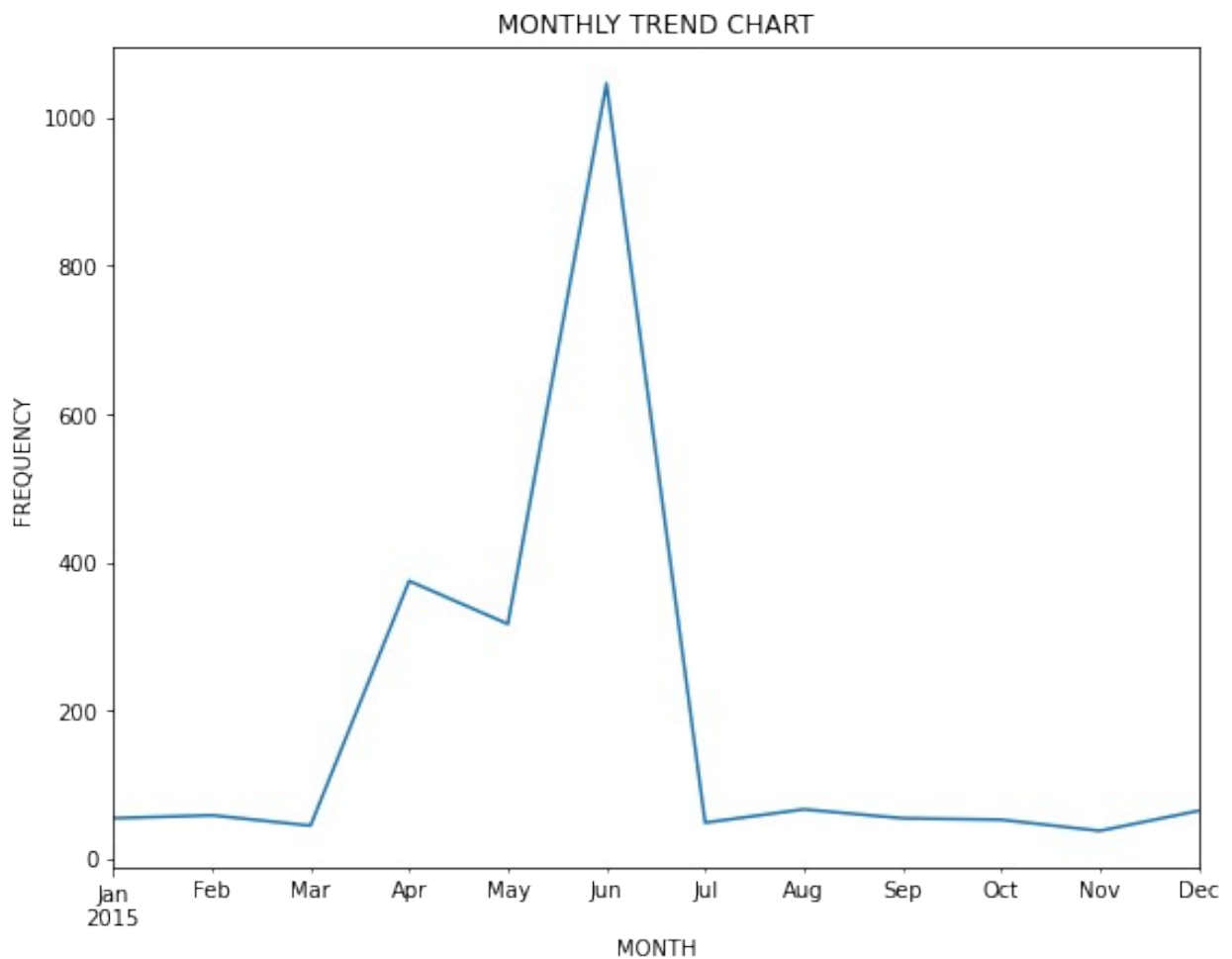


```python
df['Date'].value_counts()[:10]
```

```
24-06-15     218
23-06-15     190
25-06-15      98
26-06-15      55
30-06-15      53
```

```
29-06-15        51
18-06-15        47
06-12-15        43
27-06-15        39
15-06-15        34
Name: Date, dtype: int64
```
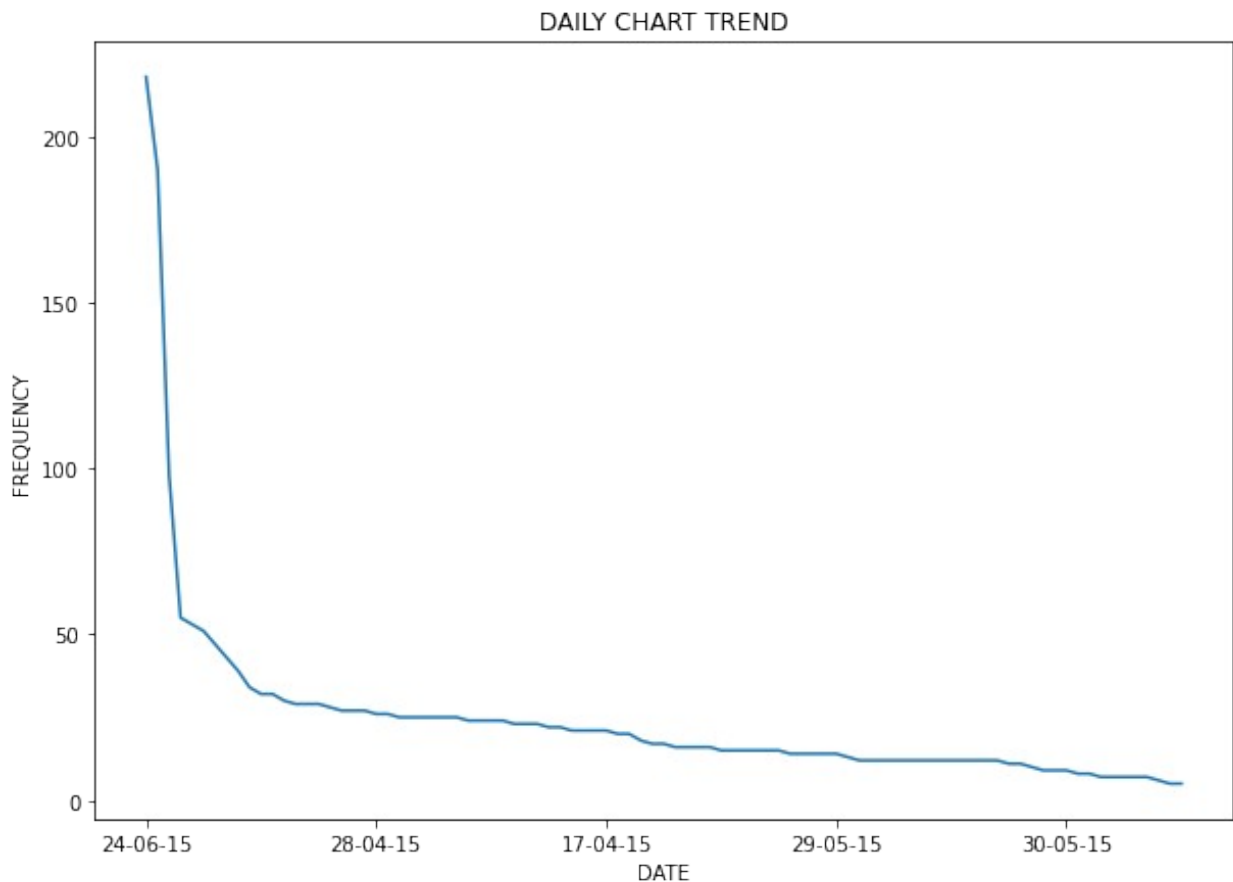
```python
#Plotting Daily trend chart

plt.figure(figsize=(10,7))
df = df.sort_values(by='Date')
df['Date'].value_counts().plot()
plt.xlabel("DATE")
plt.ylabel("FREQUENCY")
plt.title("DAILY CHART TREND")
```

```
Text(0.5, 1.0, 'DAILY CHART TREND')
```



```python
# Provide a table with the frequency of complaint types.

df['Customer Complaint'].value_counts()[:10]
```

```
Comcast                       83
Comcast Internet              18
Comcast Data Cap              17
comcast                       13
Data Caps                     11
Comcast Billing               11
Comcast Data Caps             11
Unfair Billing Practices       9
Comcast data caps              8
Comcast/Xfinity                8
Name: Customer Complaint, dtype: int64
```

```python
df['Customer Complaint'].value_counts()[:10].plot.bar()
```

```
<AxesSubplot:>
```
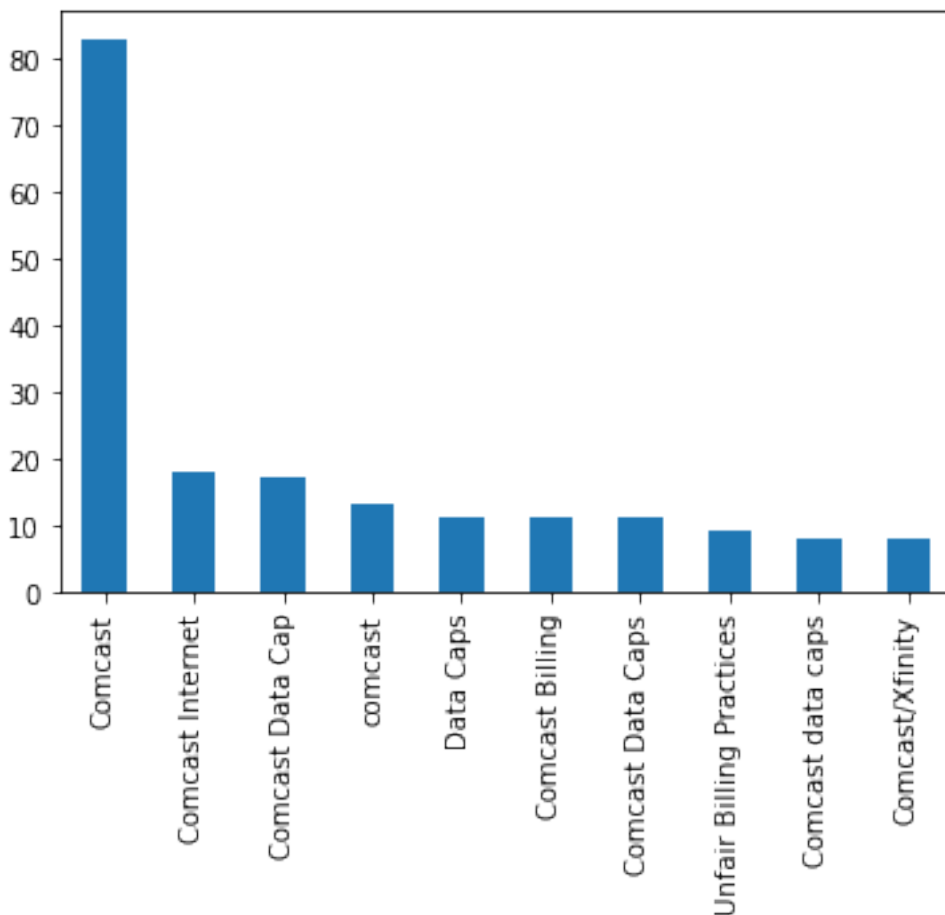


```python
# Which complaint types are maximum i.e., around internet, network
issues, or across any other domains.

df['Customer Complaint'].values
```

```
array(['Fraudulent claims reported to collections agency',
       'Comcast refusal of service', 'Comcast Cable', ..., 'n/a (b)
(6)',
       'Complaint against Comcast for incredibly bad service',
       'Questionable internet slowdown'], dtype=object)

internet_issues1 = df[df['Customer
Complaint'].str.contains('speed')].count()

internet_issues2 = df[df['Customer
Complaint'].str.contains('network')].count()

internet_issues3 = df[df['Customer
Complaint'].str.contains('data')].count()

internet_issues4 = df[df['Customer
Complaint'].str.contains('internet')].count()

billing_issues1 = df[df['Customer
Complaint'].str.contains('billing')].count()

billing_issues2 = df[df['Customer
Complaint'].str.contains('charges')].count()

service_issues1 = df[df['Customer
Complaint'].str.contains('service')].count()

service_issues2 = df[df['Customer
Complaint'].str.contains('customer')].count()

total_internet_issues = internet_issues1 + internet_issues2 +
internet_issues3 + internet_issues4

total_billing_issues = billing_issues1 + billing_issues2

total_service_issues = service_issues1 + service_issues2

other_issues = df.count() - (total_internet_issues +
total_billing_issues  + total_service_issues)

total_internet_issues
```

```
Ticket #                       374
Customer Complaint             374
Date                           374
Time                           374
Received Via                   374
City                           374
State                          374
Zip code                       374
Status                         374
```

```
Filing on Behalf of Someone     374
dtype: int64

total_billing_issues

Ticket #                        169
Customer Complaint              169
Date                            169
Time                            169
Received Via                    169
City                            169
State                           169
Zip code                        169
Status                          169
Filing on Behalf of Someone     169
dtype: int64

total_service_issues

Ticket #                        360
Customer Complaint              360
Date                            360
Time                            360
Received Via                    360
City                            360
State                           360
Zip code                        360
Status                          360
Filing on Behalf of Someone     360
dtype: int64

other_issues

Ticket #                        1321
Customer Complaint              1321
Date                            1321
Time                            1321
Received Via                    1321
City                            1321
State                           1321
Zip code                        1321
Status                          1321
Filing on Behalf of Someone     1321
dtype: int64
```

```python
#Create a new categorical variable with value as Open and Closed. Open
& Pending is to be categorized as Open and Closed & Solved is to be
categorized as Closed.

df['Status'].unique()
```

```
array(['Closed', 'Open', 'Solved', 'Pending'], dtype=object)

df["New Status"] = ["Open" if status == "Open" or status == "Pending"
else "Closed" for status in df["Status"]]

df.sample(10)
```

```
                  Ticket #                                Customer
Complaint  \
Date_month_year

2015-08-04        223179  Comcast bandwidth every evening drops to 10%
o...
2015-06-24        361980                        Throttled Internet
Speeds
2015-06-20        353984                        Connecting the
service
2015-05-13        287000            Xfinity Throttling Apple TV
bitrate
2015-06-25        363966                                    Bill
Flucuation
2015-06-24        361123                                  Comcast
Billing
2015-04-24        255808
availabilty
2015-06-22        356202
Internet cap
2015-06-16        343049          comcast fraudulent pricing and
practices
2015-05-19        297456  Excessive early termination fees due
immediate...

                    Date        Time        Received Via
City  \
Date_month_year

2015-08-04        04-08-15   1:04:24 AM            Internet
Angels Camp
2015-06-24        24-06-15   1:44:55 PM            Internet
Roseville
2015-06-20        20-06-15   3:28:21 PM  Customer Care Call
Bethesda
2015-05-13        13-05-15   3:15:58 AM  Customer Care Call  Colorado
Springs
2015-06-25        25-06-15   9:15:11 AM  Customer Care Call
Manvel
2015-06-24        24-06-15  10:05:42 AM            Internet
Tallahassee
2015-04-24        24-04-15   3:56:45 PM            Internet
Kent
```

```
2015-06-22        22-06-15    3:07:00 PM  Customer Care Call       Fort
Campbell
2015-06-16        16-06-15   11:55:45 AM  Customer Care Call
Hobart
2015-05-19        19-05-15    3:45:08 PM                Internet
Stephens City

                         State  Zip code   Status Filing on Behalf of
Someone  \
Date_month_year

2015-08-04       California     95222    Closed
No
2015-06-24       California     95661   Pending
No
2015-06-20         Maryland     20817    Solved
No
2015-05-13         Colorado     80918      Open
No
2015-06-25            Texas     77578    Solved
No
2015-06-24          Florida     32311    Solved
Yes
2015-04-24       Washington     98030    Closed
No
2015-06-22         Kentucky     42223    Solved
No
2015-06-16          Indiana     46342    Solved
No
2015-05-19         Virginia     22655    Solved
No

                 New Status
Date_month_year
2015-08-04          Closed
2015-06-24            Open
2015-06-20          Closed
2015-05-13            Open
2015-06-25          Closed
2015-06-24          Closed
2015-04-24          Closed
2015-06-22          Closed
2015-06-16          Closed
2015-05-19          Closed
```

```python
# Provide state wise status of complaints in a stacked bar chart. Use
the categorized variable from Q3. Provide insights on:

state_complain = df.groupby(['State','New Status']).size().unstack()
```

```
state_complain

New Status          Closed  Open
State
Alabama              17.0    9.0
Arizona              14.0    6.0
Arkansas              6.0    NaN
California          159.0   61.0
Colorado             58.0   22.0
Connecticut           9.0    3.0
Delaware              8.0    4.0
District Of Columbia 14.0    2.0
District of Columbia  1.0    NaN
Florida             201.0   39.0
Georgia             208.0   80.0
Illinois            135.0   29.0
Indiana              50.0    9.0
Iowa                  1.0    NaN
Kansas                1.0    1.0
Kentucky              4.0    3.0
Louisiana            12.0    1.0
Maine                 3.0    2.0
Maryland             63.0   15.0
Massachusetts        50.0   11.0
Michigan             92.0   23.0
Minnesota            29.0    4.0
Mississippi          23.0   16.0
Missouri              3.0    1.0
Montana               1.0    NaN
Nevada                1.0    NaN
New Hampshire         8.0    4.0
New Jersey           56.0   19.0
New Mexico           11.0    4.0
New York              6.0    NaN
North Carolina        3.0    NaN
Ohio                  3.0    NaN
Oregon               36.0   13.0
Pennsylvania        110.0   20.0
Rhode Island          1.0    NaN
South Carolina       15.0    3.0
Tennessee            96.0   47.0
Texas                49.0   22.0
Utah                 16.0    6.0
Vermont               2.0    1.0
Virginia             49.0   11.0
Washington           75.0   23.0
West Virginia         8.0    3.0


state_complain.plot.bar(figsize=(10,7),stacked=True)
```

```
<AxesSubplot:xlabel='State'>
```
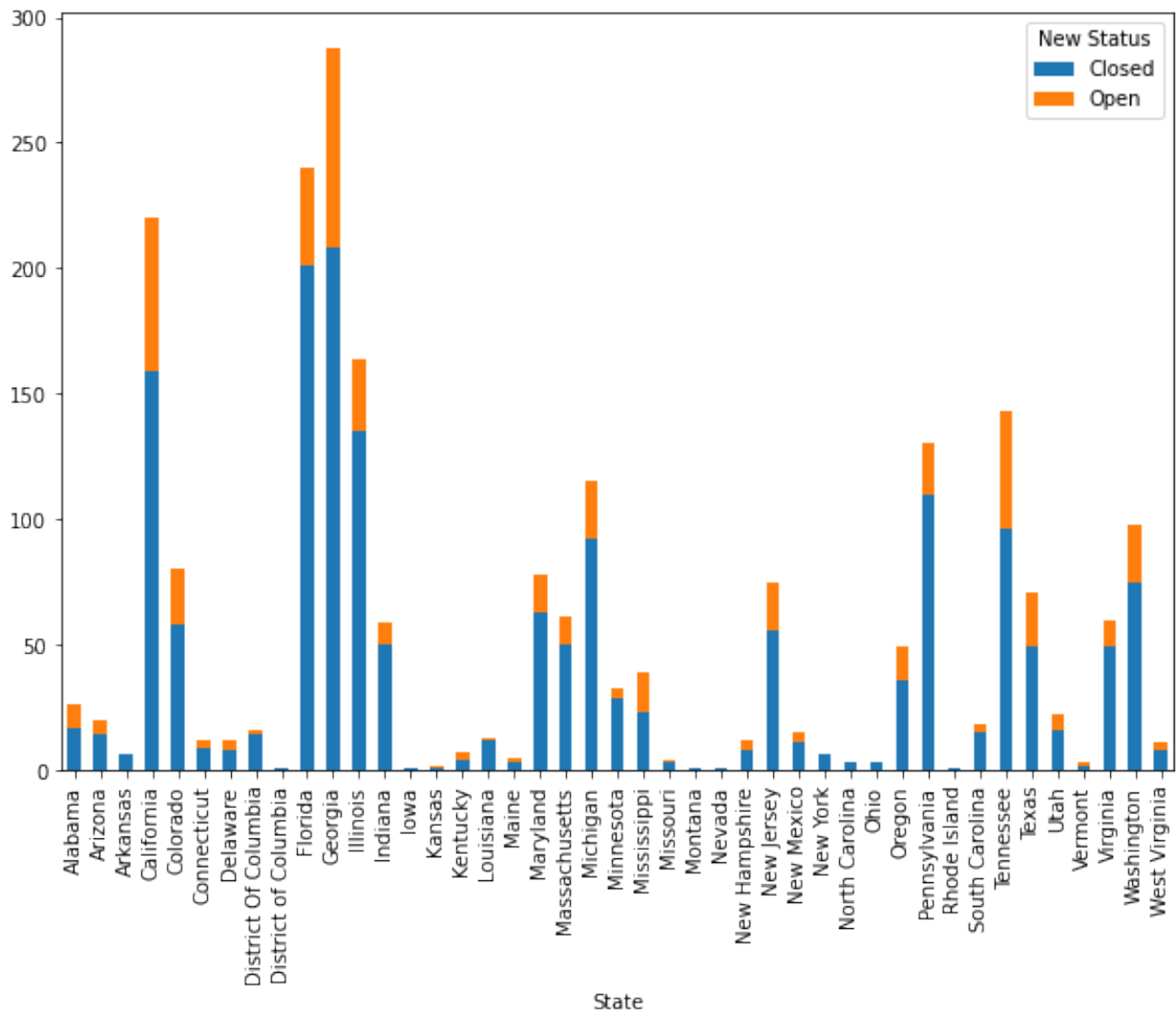


```python
# Which state has the maximum complaints

df.groupby('State').size().sort_values(ascending=False)[:5]

State
Georgia        288
Florida        240
California     220
Illinois       164
Tennessee      143
dtype: int64

# Which state has the highest percentage of unresolved complaints
```

```python
unresolved_data = df.groupby(['State', 'New
Status']).size().unstack().fillna(0).sort_values(by="Open",ascending =
False)

unresolved_data['Unresolved Cmpln Prcnt'] =
unresolved_data['Open']/unresolved_data['Open'].sum()*100

unresolved_data
```

New Status	Closed	Open	Unresolved Cmpln Prcnt
State			
Georgia	208.0	80.0	15.473888
California	159.0	61.0	11.798839
Tennessee	96.0	47.0	9.090909
Florida	201.0	39.0	7.543520
Illinois	135.0	29.0	5.609284
Washington	75.0	23.0	4.448743
Michigan	92.0	23.0	4.448743
Colorado	58.0	22.0	4.255319
Texas	49.0	22.0	4.255319
Pennsylvania	110.0	20.0	3.868472
New Jersey	56.0	19.0	3.675048
Mississippi	23.0	16.0	3.094778
Maryland	63.0	15.0	2.901354
Oregon	36.0	13.0	2.514507
Virginia	49.0	11.0	2.127660
Massachusetts	50.0	11.0	2.127660
Alabama	17.0	9.0	1.740812
Indiana	50.0	9.0	1.740812
Utah	16.0	6.0	1.160542
Arizona	14.0	6.0	1.160542
New Hampshire	8.0	4.0	0.773694
New Mexico	11.0	4.0	0.773694
Minnesota	29.0	4.0	0.773694
Delaware	8.0	4.0	0.773694
West Virginia	8.0	3.0	0.580271
Connecticut	9.0	3.0	0.580271
Kentucky	4.0	3.0	0.580271
South Carolina	15.0	3.0	0.580271
Maine	3.0	2.0	0.386847
District Of Columbia	14.0	2.0	0.386847
Kansas	1.0	1.0	0.193424
Vermont	2.0	1.0	0.193424
Missouri	3.0	1.0	0.193424
Louisiana	12.0	1.0	0.193424
Montana	1.0	0.0	0.000000
Rhode Island	1.0	0.0	0.000000
Ohio	3.0	0.0	0.000000
District of Columbia	1.0	0.0	0.000000
North Carolina	3.0	0.0	0.000000

```
New York                          6.0    0.0                    0.000000
Nevada                            1.0    0.0                    0.000000
Arkansas                          6.0    0.0                    0.000000
Iowa                              1.0    0.0                    0.000000
```

```
unresolved_data.plot()
```

```
<AxesSubplot:xlabel='State'>
```
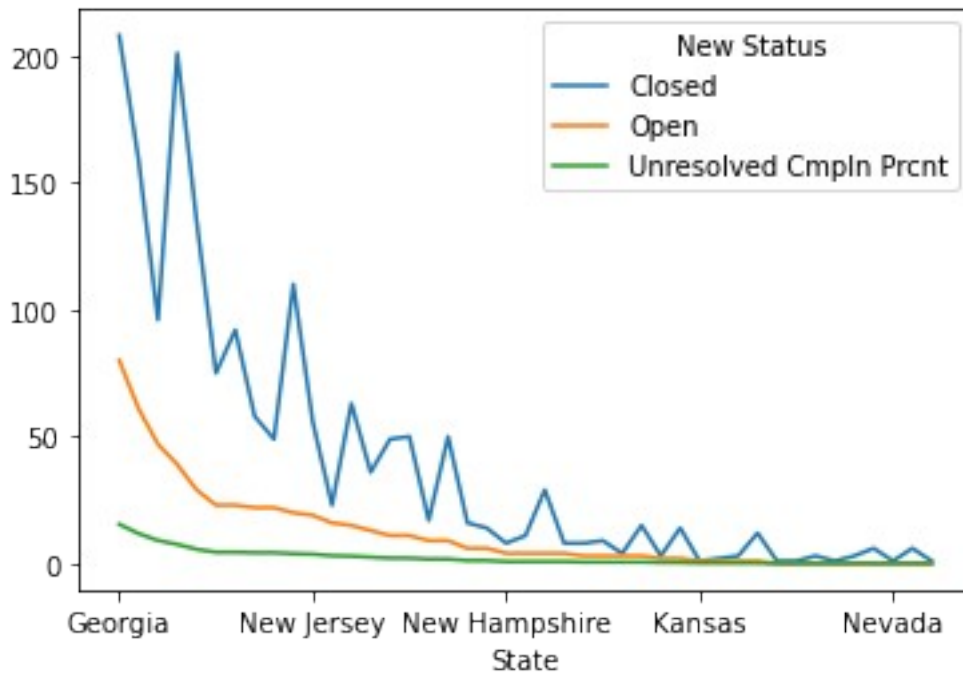


```
# Provide the percentage of complaints resolved till date, which were
received through the Internet and customer care calls.
```

```
df.head()
```

```
                 Ticket #                                   Customer
Complaint  \
Date_month_year

2015-01-04        211976  Fraudulent claims reported to collections
agency
2015-01-04        211677                          Comcast refusal of
service
2015-01-04        212507                                     Comcast
Cable
2015-01-04        213120                                        Data
Overages
2015-01-04        211478
Comcast
```

```
                       Date        Time        Received Via
City  \
Date_month_year

2015-01-04      04-01-15    1:26:53 PM   Customer Care Call
Atlanta
2015-01-04      04-01-15   12:01:06 PM   Customer Care Call
Wayne
2015-01-04      04-01-15    3:54:43 PM              Internet
Franklin
2015-01-04      04-01-15    8:05:57 PM              Internet
Savannah
2015-01-04      04-01-15   10:47:35 AM              Internet   North
Huntingdon


                         State   Zip code   Status Filing on Behalf of
Someone  \
Date_month_year

2015-01-04            Georgia      30312   Closed
No
2015-01-04       Pennsylvania      19087   Closed
No
2015-01-04          Tennessee      37067   Closed
No
2015-01-04            Georgia      31406   Closed
No
2015-01-04       Pennsylvania      15642   Closed
No


                   New Status
Date_month_year
2015-01-04             Closed
2015-01-04             Closed
2015-01-04             Closed
2015-01-04             Closed
2015-01-04             Closed
```

```python
resolved_data = df.groupby(['Received Via','New
Status']).size().unstack().fillna(0)

resolved_data.head()
```

```
New Status            Closed  Open
Received Via
Customer Care Call       864   255
Internet                 843   262
```

```python
resolved_data['Resolved Cmpln Prcnt'] =
resolved_data['Closed']/resolved_data['Closed'].sum()*100
```

```
resolved_data.plot(kind='bar',figsize=(10,7))
```

<AxesSubplot:xlabel='Received Via'>