In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
comcast = pd.read_csv(r'C:\Users\ashok\Downloads\PG\PGP Data Science\Data_Science_With_Python_2020\Projects\Comcast_telecom_complaints_data.csv')
```

In [3]:
```
comcast.head()
```

Out[3]:

	Ticket #	Customer Complaint	Date	Date_month_year	Time	Received Via	City	State	Zip code	Status	Fili Bel Son
0	250635	Comcast Cable Internet Speeds	22-04-15	22-Apr-15	3:53:50 PM	Customer Care Call	Abingdon	Maryland	21009	Closed	
1	223441	Payment disappear - service got disconnected	04-08-15	04-Aug-15	10:22:56 AM	Internet	Acworth	Georgia	30102	Closed	
2	242732	Speed and Service	18-04-15	18-Apr-15	9:55:47 AM	Internet	Acworth	Georgia	30101	Closed	
3	277946	Comcast Imposed a New Usage Cap of 300GB that ...	05-07-15	05-Jul-15	11:59:35 AM	Internet	Acworth	Georgia	30101	Open	
4	307175	Comcast not working and no service to boot	26-05-15	26-May-15	1:25:26 PM	Internet	Acworth	Georgia	30101	Solved	

In [4]:
```
comcast.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
```

In [5]:
```
comcast['Date']=pd.to_datetime(comcast['Date'])
```

In [6]:
```
comcast.head()
```

Out[6]:

	Ticket #	Customer Complaint	Date	Date_month_year	Time	Received Via	City	State	Zip code	Status	Fil Be So
0	250635	Comcast Cable Internet Speeds	2015-04-22	22-Apr-15	3:53:50 PM	Customer Care Call	Abingdon	Maryland	21009	Closed	
1	223441	Payment disappear - service got disconnected	2015-04-08	04-Aug-15	10:22:56 AM	Internet	Acworth	Georgia	30102	Closed	
2	242732	Speed and Service	2015-04-18	18-Apr-15	9:55:47 AM	Internet	Acworth	Georgia	30101	Closed	
3	277946	Comcast Imposed a New Usage Cap of 300GB that ...	2015-05-07	05-Jul-15	11:59:35 AM	Internet	Acworth	Georgia	30101	Open	
4	307175	Comcast not working and no service to boot	2015-05-26	26-May-15	1:25:26 PM	Internet	Acworth	Georgia	30101	Solved	

In [7]:
```
comcast['Date1']=pd.to_datetime(comcast['Date_month_year'])
```

In [8]:
```
comcast.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2224 entries, 0 to 2223
Data columns (total 12 columns):
 #   Column                       Non-Null Count  Dtype         
---  ------                       --------------  -----         
 0   Ticket #                     2224 non-null   object        
 1   Customer Complaint           2224 non-null   object        
 2   Date                         2224 non-null   datetime64[ns]
 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        
 11  Date1                        2224 non-null   datetime64[ns]
dtypes: datetime64[ns](2), int64(1), object(9)
memory usage: 208.6+ KB
```

In [9]:
```
# group by date
comcast.groupby('Date1')['Ticket #'].count()
```

Out[9]:
```
Date1
2015-01-04    18
2015-01-05    12
2015-01-06    25
2015-02-04    27
2015-02-05     7
              ..
2015-11-05    12
2015-11-06    21
2015-12-04    15
2015-12-05     7
2015-12-06    43
Name: Ticket #, Length: 91, dtype: int64
```

In [10]:
```
# convert into df
Dailycomplaint=pd.DataFrame(comcast.groupby('Date1')['Ticket #'].count().reset_index())
Dailycomplaint
```

Out[10]:

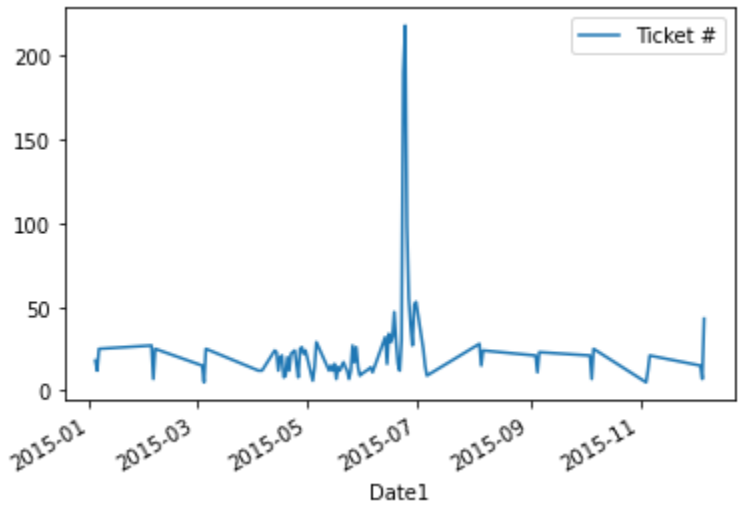
	Date1	Ticket #
0	2015-01-04	18
1	2015-01-05	12
2	2015-01-06	25
3	2015-02-04	27
4	2015-02-05	7
...	...	...
86	2015-11-05	12
87	2015-11-06	21
88	2015-12-04	15
89	2015-12-05	7
90	2015-12-06	43

91 rows × 2 columns

In [11]:
```
plt.figure(figsize=(24,12))
Dailycomplaint.plot(kind='line', x='Date1', y='Ticket #')
```

Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x23576340040>

<Figure size 1728x864 with 0 Axes>
```



In [12]:
```
import datetime
comcast['Month']=comcast['Date'].dt.month
```

/

In [13]:
```
comcast.head()
```

Out[13]:

	Ticket #	Customer Complaint	Date	Date_month_year	Time	Received Via	City	State	Zip code	Status	Fil Be Sol
0	250635	Comcast Cable Internet Speeds	2015-04-22	22-Apr-15	3:53:50 PM	Customer Care Call	Abingdon	Maryland	21009	Closed	
1	223441	Payment disappear - service got disconnected	2015-04-08	04-Aug-15	10:22:56 AM	Internet	Acworth	Georgia	30102	Closed	
2	242732	Speed and Service	2015-04-18	18-Apr-15	9:55:47 AM	Internet	Acworth	Georgia	30101	Closed	
3	277946	Comcast Imposed a New Usage Cap of 300GB that ...	2015-05-07	05-Jul-15	11:59:35 AM	Internet	Acworth	Georgia	30101	Open	
4	307175	Comcast not working and no service to boot	2015-05-26	26-May-15	1:25:26 PM	Internet	Acworth	Georgia	30101	Solved	

In [14]:
```
Monthlycomplaint=pd.DataFrame(comcast.groupby('Month')['Ticket #'].count().reset_index
())
Monthlycomplaint
```

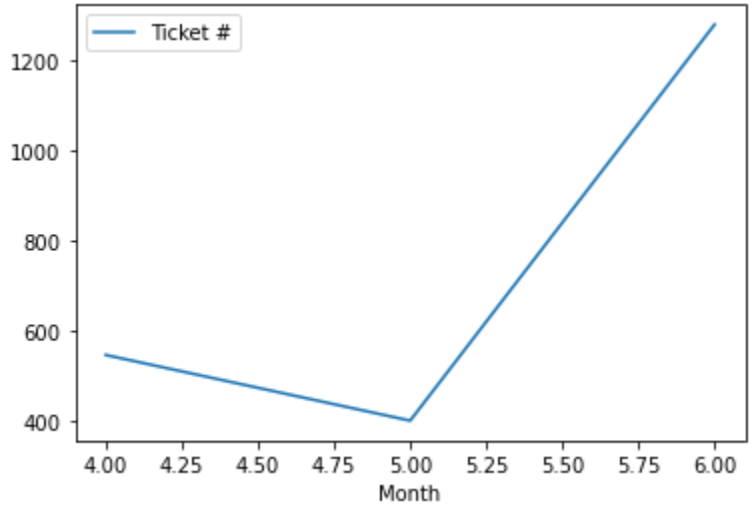
Out[14]:

	Month	Ticket #
0	4	545
1	5	399
2	6	1280

In [15]:
```
plt.figure(figsize=(9,9))
Monthlycomplaint.plot(kind='line', x='Month', y='Ticket #')
```

Out[15]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x235764bf0a0>

<Figure size 648x648 with 0 Axes>
```



In [16]:
```
#which complaint is maximum
comcast['Customer Complaint']
```

Out[16]:
```
0                            Comcast Cable Internet Speeds
1              Payment disappear - service got disconnected
2                                           Speed and Service
3         Comcast Imposed a New Usage Cap of 300GB that ...
4                 Comcast not working and no service to boot
                               ...
2219                                     Service Availability
2220             Comcast Monthly Billing for Returned Modem
2221                                complaint about comcast
2222                  Extremely unsatisfied Comcast customer
2223                    Comcast, Ypsilanti MI Internet Speed
Name: Customer Complaint, Length: 2224, dtype: object
```

In [ ]:
```
```

In [17]:
```
# Document frequency : a word comes in all the documents
# term frequency : number of times a term comes in a document

from sklearn.feature_extraction.text import CountVectorizer
```

In [18]:
```
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ashok\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[18]:
```
True
```

/

In [19]:
```
count_vec=CountVectorizer(stop_words='english', max_df=0.8, min_df=0.2)

#max_df = when creating the bag of words, remove those words which have document freq > 0.8
#min_df = when creating the bag of words, remove those words which have document freq < 0.2
```

In [20]:
```
count_vec_matrix = count_vec.fit_transform(comcast['Customer Complaint'].values.astype('U'))
```

In [21]:
```
count_vec_matrix
```

Out[21]:
```
<2224x2 sparse matrix of type '<class 'numpy.int64'>'
        with 1775 stored elements in Compressed Sparse Row format>
```

In [22]:
```
# LDA

from sklearn.decomposition import LatentDirichletAllocation
```

In [33]:
```
# n_components =  no. of topics to be created
lda = LatentDirichletAllocation()
```

In [34]:
```
lda
```

Out[34]:
```
LatentDirichletAllocation()
```

In [35]:
```
# use lda class on the count vectorizer class
lda.fit(count_vec_matrix)
```

Out[35]:
```
LatentDirichletAllocation()
```

In [36]:
```
lda.components_
```

Out[36]:
```
array([[3.36154809e+01, 3.36016911e+01],
       [1.00001598e-01, 2.99067649e+02],
       [3.27585501e+02, 1.00000345e-01],
       [4.06072432e+01, 4.25978796e+01],
       [3.86233150e+01, 3.86097512e+01],
       [3.49543545e+02, 1.00000692e-01],
       [3.36060020e+02, 1.00000371e-01],
       [3.91187236e+01, 3.91052417e+01],
       [4.11179747e+01, 4.11051282e+01],
       [4.06281956e+01, 4.16126577e+01]])
```

In [39]:
```
count_vec.get_feature_names()[1]
```

Out[39]:
```
'internet'
```

In [40]:
```
#to get what all topics
lda.components_
```

Out[40]:
```
array([[3.36154809e+01, 3.36016911e+01],
       [1.00001598e-01, 2.99067649e+02],
       [3.27585501e+02, 1.00000345e-01],
       [4.06072432e+01, 4.25978796e+01],
       [3.86233150e+01, 3.86097512e+01],
       [3.49543545e+02, 1.00000692e-01],
       [3.36060020e+02, 1.00000371e-01],
       [3.91187236e+01, 3.91052417e+01],
       [4.11179747e+01, 4.11051282e+01],
       [4.06281956e+01, 4.16126577e+01]])
```

In [41]:
```
first_topic = lda.components_[0]
first_topic
```

Out[41]:
```
array([33.61548086, 33.60169105])
```

In [42]:
```
#indcies of the words which are coming in first topic

first_topic_words = first_topic.argsort()[-8:0]
first_topic_words
```

Out[42]:
```
array([], dtype=int64)
```

In [43]:
```
for i in first_topic_words:
    print(count_vec.get_feature_names()[i])
```

/

In [44]:
```
for i, topic in enumerate(lda.components_):
    print([count_vec.get_feature_names()[i]
for i in topic.argsort()[-10:]])
    print("\n")
```

```
['internet', 'comcast']


['comcast', 'internet']


['internet', 'comcast']


['comcast', 'internet']


['internet', 'comcast']


['internet', 'comcast']


['internet', 'comcast']


['internet', 'comcast']


['internet', 'comcast']


['comcast', 'internet']
```

In [100]:
```
# categorize into open/close
comcast['open/close'] = comcast['Status'].replace('Solved', 'Closed').replace('Pending',
'Open')
```

In [101]:
```
comcast.head(10)
```

Out[101]:

	Ticket #	Customer Complaint	Date	Date_month_year	Time	Received Via	City	State	Zip code	Status	F B S
0	250635	Comcast Cable Internet Speeds	2015-04-22	22-Apr-15	3:53:50 PM	Customer Care Call	Abingdon	Maryland	21009	Closed	
1	223441	Payment disappear - service got disconnected	2015-04-08	04-Aug-15	10:22:56 AM	Internet	Acworth	Georgia	30102	Closed	
2	242732	Speed and Service	2015-04-18	18-Apr-15	9:55:47 AM	Internet	Acworth	Georgia	30101	Closed	
3	277946	Comcast Imposed a New Usage Cap of 300GB that ...	2015-05-07	05-Jul-15	11:59:35 AM	Internet	Acworth	Georgia	30101	Open	
4	307175	Comcast not working and no service to boot	2015-05-26	26-May-15	1:25:26 PM	Internet	Acworth	Georgia	30101	Solved	
5	338519	ISP Charging for arbitrary data limits with ov...	2015-06-12	06-Dec-15	9:59:40 PM	Internet	Acworth	Georgia	30101	Solved	
6	361148	Throttling service and unreasonable data caps	2015-06-24	24-Jun-15	10:13:55 AM	Customer Care Call	Acworth	Georgia	30101	Pending	
7	359792	Comcast refuses to help troubleshoot and corre...	2015-06-23	23-Jun-15	6:56:14 PM	Internet	Adrian	Michigan	49221	Solved	
8	318072	Comcast extended outages	2015-06-01	06-Jan-15	11:46:30 PM	Customer Care Call	Alameda	California	94502	Closed	
9	371214	Comcast Raising Prices and Not Being Available...	2015-06-28	28-Jun-15	6:46:31 PM	Customer Care Call	Alameda	California	94501	Open	

In [107]:
```
#statewise stacked bar chart
#statewisecomplaint=pd.DataFrame(comcast.groupby('State')['open/close'].count().reset_in
dex())
#statewisecomplaint

statewisecomplaint = pd.crosstab(comcast['State'], comcast['open/close'])
statewisecomplaint
```

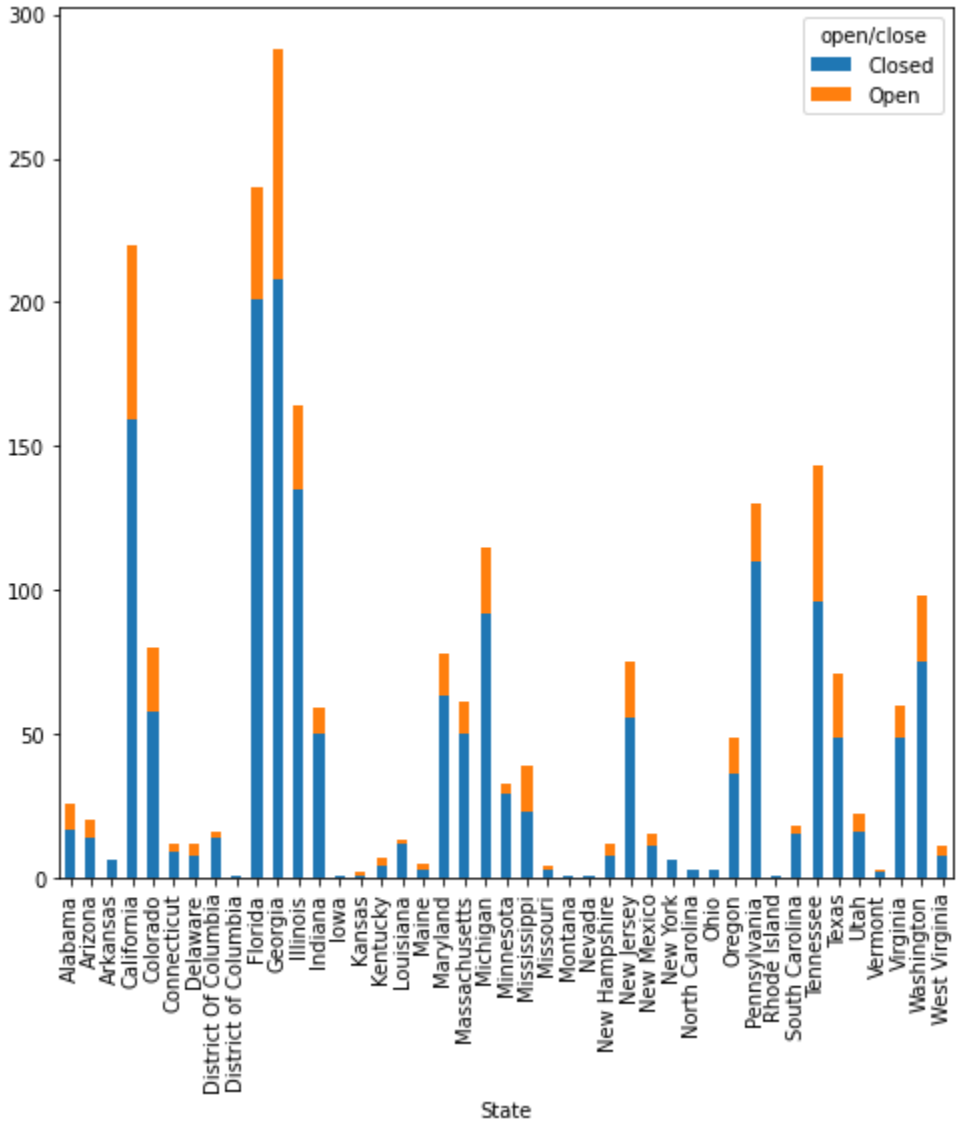
Out[107]:

open/close	Closed	Open
State		
Alabama	17	9
Arizona	14	6
Arkansas	6	0
California	159	61
Colorado	58	22
Connecticut	9	3
Delaware	8	4
District Of Columbia	14	2
District of Columbia	1	0
Florida	201	39
Georgia	208	80
Illinois	135	29
Indiana	50	9
Iowa	1	0
Kansas	1	1
Kentucky	4	3
Louisiana	12	1
Maine	3	2
Maryland	63	15
Massachusetts	50	11
Michigan	92	23
Minnesota	29	4
Mississippi	23	16
Missouri	3	1
Montana	1	0
Nevada	1	0
New Hampshire	8	4
New Jersey	56	19
New Mexico	11	4
New York	6	0
North Carolina	3	0
Ohio	3	0
Oregon	36	13
Pennsylvania	110	20
Rhode Island	1	0
South Carolina	15	3

open/close	Closed	Open
State		
Tennessee	96	47
Texas	49	22
Utah	16	6
Vermont	2	1
Virginia	49	11
Washington	75	23
West Virginia	8	3

In [108]:
```
statewisecomplaint.plot(kind='bar',figsize=(8,8),stacked=True)
```

Out[108]: <matplotlib.axes._subplots.AxesSubplot at 0x2357cc08550>



In [120]:
```
#maxcomplaint
max = statewisecomplaint.max()
print(max)
#Georgia
```

```
open/close
Closed    208
Open       80
dtype: int64
```

In [124]:
```
# % unresloved
statewisecomplaint['unresolvedpct']= (statewisecomplaint['Open']/(statewisecomplaint['Closed']+statewisecomplaint['Open']))*100
```

In [127]:
```
statewisecomplaint
```

Out[127]:

open/close State	Closed	Open	unresolvedpct
Alabama	17	9	34.615385
Arizona	14	6	30.000000
Arkansas	6	0	0.000000
California	159	61	27.727273
Colorado	58	22	27.500000
Connecticut	9	3	25.000000
Delaware	8	4	33.333333
District Of Columbia	14	2	12.500000
District of Columbia	1	0	0.000000
Florida	201	39	16.250000
Georgia	208	80	27.777778
Illinois	135	29	17.682927
Indiana	50	9	15.254237
Iowa	1	0	0.000000
Kansas	1	1	50.000000
Kentucky	4	3	42.857143
Louisiana	12	1	7.692308
Maine	3	2	40.000000
Maryland	63	15	19.230769
Massachusetts	50	11	18.032787
Michigan	92	23	20.000000
Minnesota	29	4	12.121212
Mississippi	23	16	41.025641
Missouri	3	1	25.000000
Montana	1	0	0.000000
Nevada	1	0	0.000000
New Hampshire	8	4	33.333333
New Jersey	56	19	25.333333
New Mexico	11	4	26.666667
New York	6	0	0.000000
North Carolina	3	0	0.000000
Ohio	3	0	0.000000
Oregon	36	13	26.530612
Pennsylvania	110	20	15.384615
Rhode Island	1	0	0.000000
South Carolina	15	3	16.666667

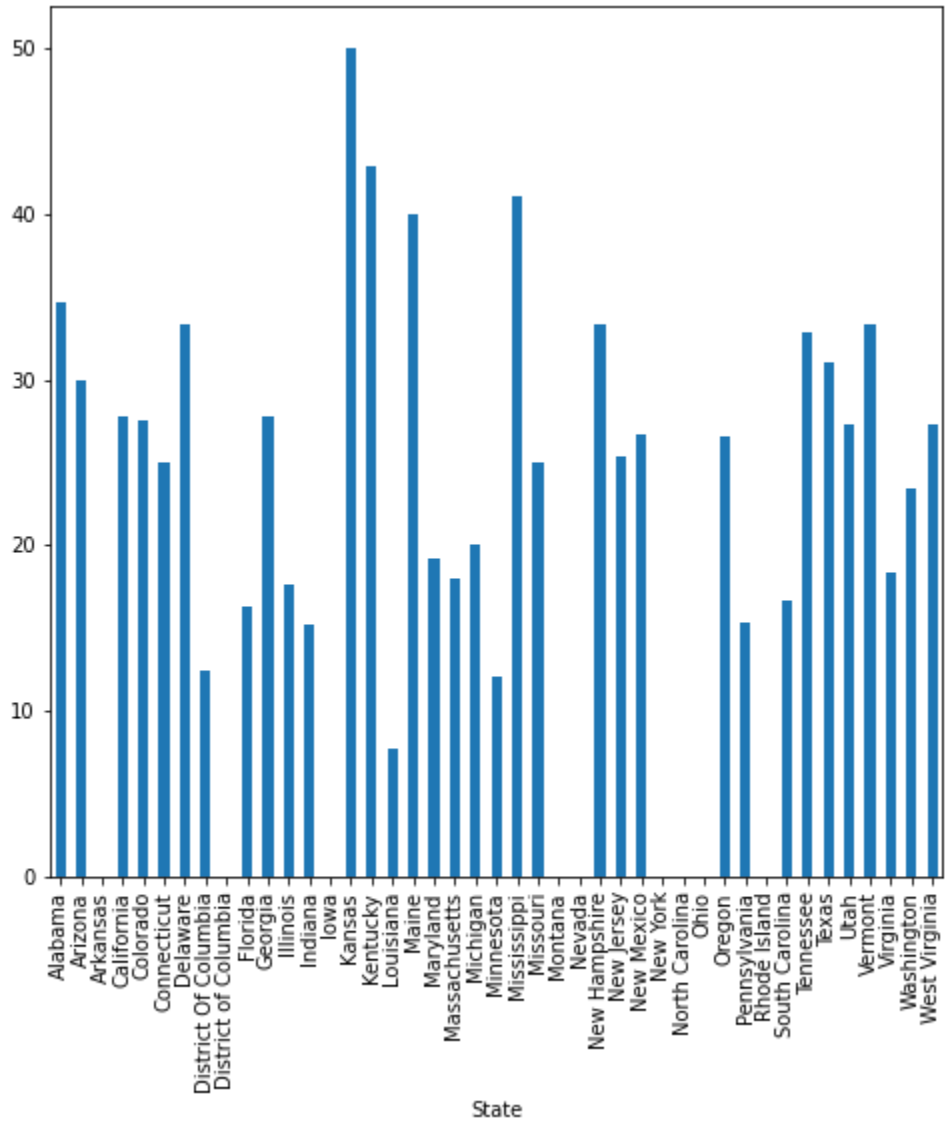
open/close State	Closed	Open	unresolvedpct
Tennessee	96	47	32.867133
Texas	49	22	30.985915
Utah	16	6	27.272727
Vermont	2	1	33.333333
Virginia	49	11	18.333333
Washington	75	23	23.469388
West Virginia	8	3	27.272727

In [126]:
```
statewisecomplaint['unresolvedpct'].max()
```

Out[126]: 50.0

In [128]:
```
statewisecomplaint['unresolvedpct'].plot(kind='bar',figsize=(8,8))
```

Out[128]: <matplotlib.axes._subplots.AxesSubplot at 0x2357c312dc0>



In [129]:
```
# Kansas has max % unresolved complaint
```

In [136]:
```
relsolvedpct = (statewisecomplaint['Closed'].sum()/(statewisecomplaint['Closed'].sum() +
statewisecomplaint['Open'].sum()))*100
```

In [137]:
```
# % complaint resolved
relsolvedpct
```

Out[137]: 76.75359712230215

In [ ]:

In [ ]:

In [ ]: