In [2]:
```
import pandas as pd
#import requests
#from bs4 import BeautifulSoup

#import time

from pandas import Series, DataFrame

import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib import style


from pandas import set_option
set_option("display.max_rows", 35)

LARGE_FIGSIZE = (12, 8)
style.use('ggplot')
%matplotlib inline
```

In [3]:
```
# file desc: https://www.fec.gov/campaign-finance-data/all-candidates-file-description/
sched_a_df = pd.read_csv('sched_a_2016.txt',sep='|', header=None)
sched_a_df.columns = ['CMTE_ID',
                      'AMNDT_IND',
                      'RPT_TP',
                      'TRANSACTION_PGI',
                      'IMAGE_NUM',
                      'TRANSACTION_TP',
                      'ENTITY_TP',
                      'NAME',
                      'CITY',
                      'STATE',
                      'ZIP_CODE',
                      'EMPLOYER',
                      'OCCUPATION',
                      'TRANSACTION_DT',
                      'TRANSACTION_AMT',
                      'OTHER_ID',
                      'TRAN_ID',
                      'FILE_NUM',
                      'MEMO_CD',
                      'MEMO_TXT',
                      'SUB_ID']

tmp = sched_a_df.head()
tmp
```

```
/Users/sunshine168/.pyenv/versions/anaconda3-4.4.0/lib/python3.6/site-packages/IPython/core/interactiveshell.py:2785: DtypeWarning: Columns (3,5,10,15,18,19) have mixed types. Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Out[3]:

	CMTE_ID	AMNDT_IND	RPT_TP	TRANSACTION_PGI	IMAGE_NUM	TRAN
0	C00572537	N	MY	P	201509169002679475	15
1	C00550087	A	Q3	G	201611189037214001	15
2	C00578013	A	YE	P	201608050200329616	15E
3	C00573758	A	YE	P	201702030200053330	15E
4	C00573758	A	YE	P	201611170200661592	15E

5 rows × 21 columns

In [4]:
```
tmp = tmp.T
tmp
```

Out[4]:

	0	1	
CMTE_ID	C00572537	C00550087	C00578013
AMNDT_IND	N	A	A
RPT_TP	MY	Q3	YE
TRANSACTION_PGI	P	G	P
IMAGE_NUM	201509169002679475	201611189037214001	201608050200329
TRANSACTION_TP	15	15	15E
ENTITY_TP	IND	IND	IND
NAME	CALLAHAN, MICHAEL	LASERSOHN, TOM	MANSON, CONNIE
CITY	MONTREAL	WESTPORT	OLYMPIA
STATE	ZZ	CT	WA
ZIP_CODE	NaN	06880	985063741
EMPLOYER	SELF EMPLOYED	RETIRED	AMERICAN GEOLOGICAL INSTITUTE
OCCUPATION	MEDIA & COMMUNICATIONS	RETIRED	EDITOR/INDEXER
TRANSACTION_DT	3312016	3312016	3172016
TRANSACTION_AMT	19450	1000	50
OTHER_ID	NaN	NaN	C00401224
TRAN_ID	SA11AI.5152	SA11AI.5127	SA0915169616385
FILE_NUM	1024961	1126451	1099259
MEMO_CD	X	X	NaN
MEMO_TXT	NaN	NaN	*EARMARKED CONTRIBUTION: S BELOW
SUB_ID	4091720151253097597	4111820161350974138	103202017003393

In [5]:
```
sched_a_df.shape
```

Out[5]: (14597286, 21)

In [6]:
```
sched_a_df.count()
```

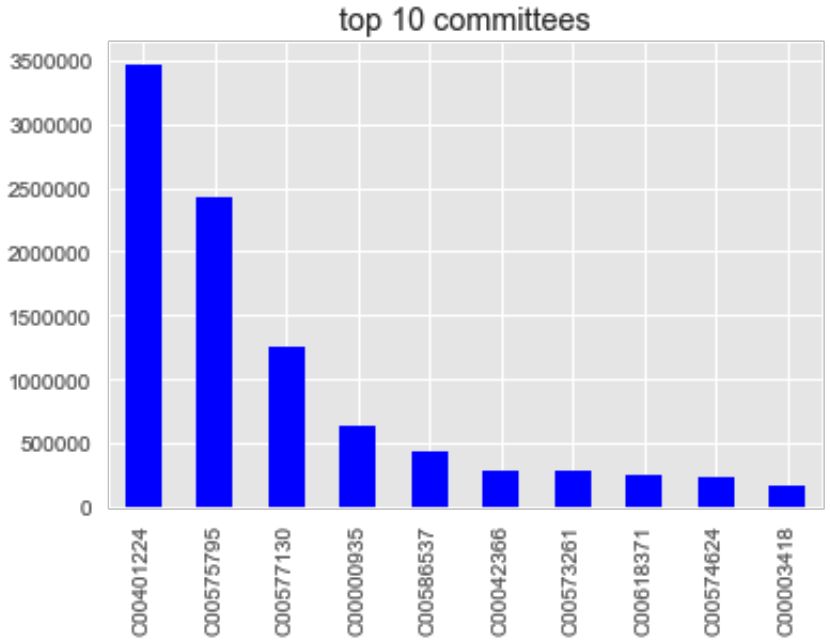
Out[6]:
```
CMTE_ID            14597286
AMNDT_IND          14597286
RPT_TP             14597286
TRANSACTION_PGI    14098549
IMAGE_NUM          14597286
TRANSACTION_TP     14597286
ENTITY_TP          14591652
NAME               14595955
CITY               14590712
STATE              14571224
ZIP_CODE           14574782
EMPLOYER           12600882
OCCUPATION         13970083
TRANSACTION_DT     14597286
TRANSACTION_AMT    14597286
OTHER_ID            5900262
TRAN_ID            14597270
FILE_NUM           14597286
MEMO_CD               76561
MEMO_TXT            7441494
SUB_ID             14597286
dtype: int64
```

In [7]:
```
sched_a_df.CMTE_ID.value_counts()[:10]
```

Out[7]:
```
C00401224    3473182
C00575795    2433232
C00577130    1252557
C00000935     629231
C00586537     432869
C00042366     285671
C00573261     273237
C00618371     245769
C00574624     236054
C00003418     169717
Name: CMTE_ID, dtype: int64
```

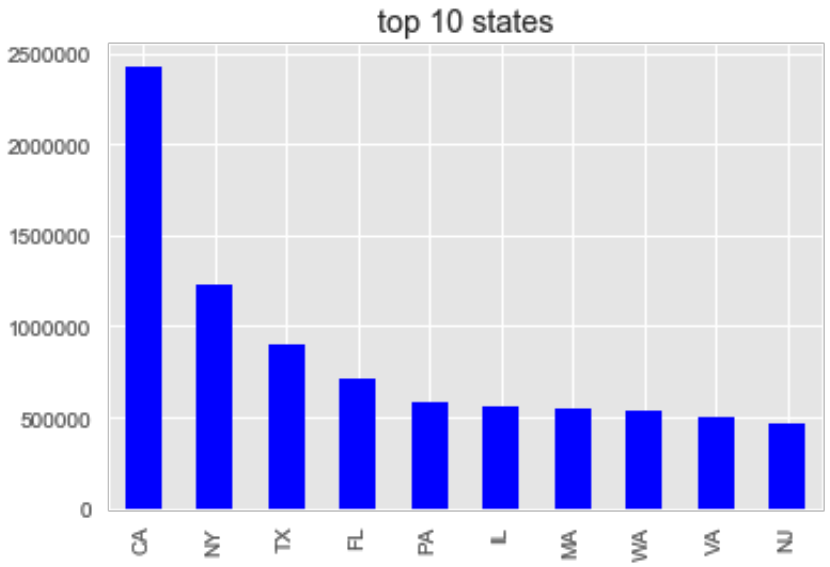
In [8]:
```
# top committee_ids
sched_a_df.CMTE_ID.value_counts()[:10].plot(kind='bar', color='b', tit
le='top 10 committees')
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x11f9779e8>



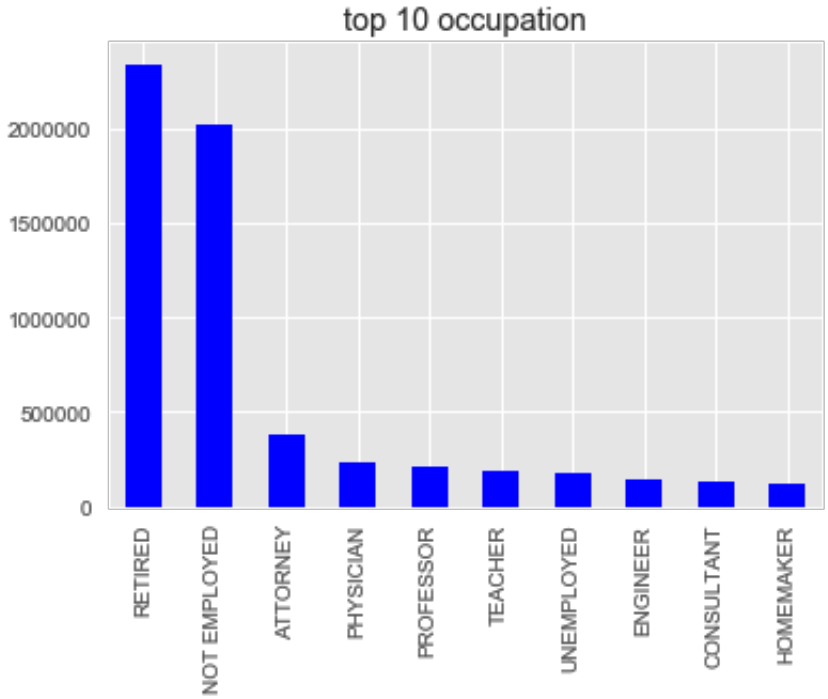
In [9]:
```
# top states
sched_a_df.STATE.value_counts()[:10].plot(kind='bar', color='b', title
='top 10 states')
```

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x11715cd68>



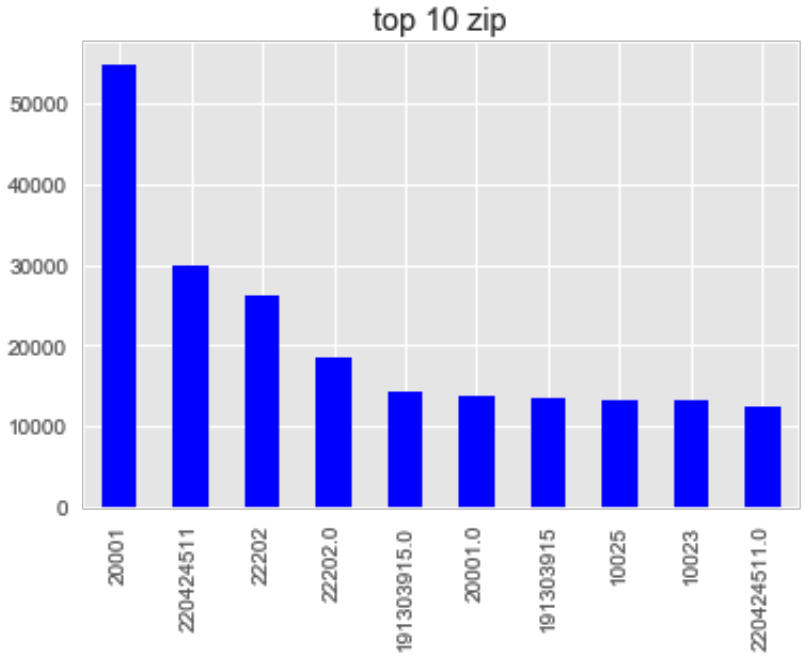
In [10]:
```
# top occupations
sched_a_df.OCCUPATION.value_counts()[:10].plot(kind='bar', color='b',
title='top 10 occupation')
```

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x1190c6668>



In [11]:
```
# top zip code
sched_a_df.ZIP_CODE.value_counts()[:10].plot(kind='bar', color='b', ti
tle='top 10 zip')
```

Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1190fc550>

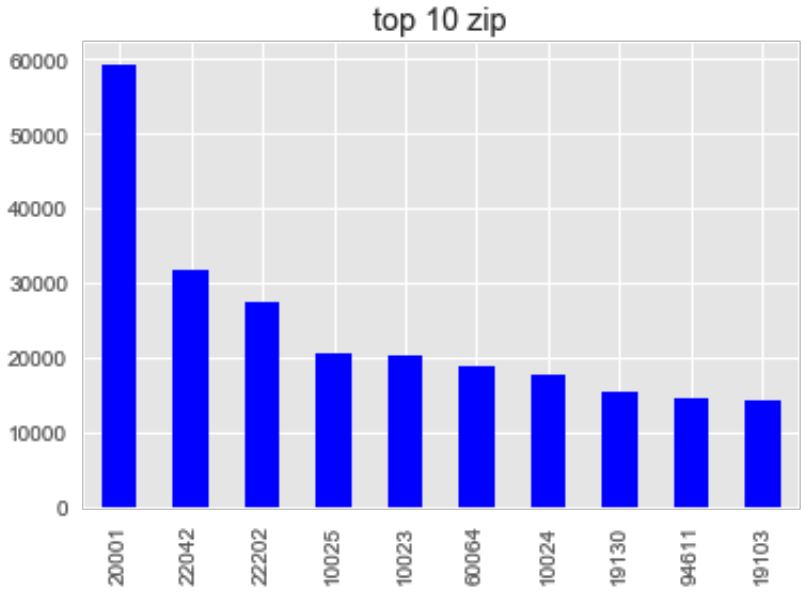


In [12]:
```
sched_a_df.ZIP = sched_a_df.ZIP_CODE.str[:5]
sched_a_df.ZIP[:10]
```

Out[12]:
```
0      NaN
1    06880
2    98506
3    33437
4    33437
5    20814
6    28105
7    80503
8      NaN
9    66221
Name: ZIP_CODE, dtype: object
```

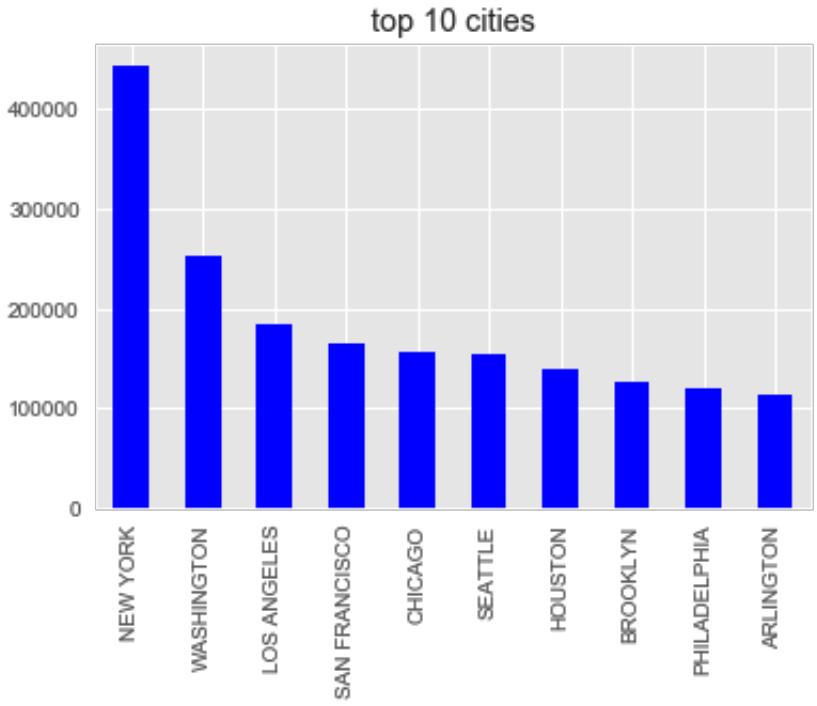
In [13]:
```
# top zips after cleanup
sched_a_df.ZIP.value_counts()[:10].plot(kind='bar', color='b', title='
top 10 zip')
```

Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x14bc8d908>



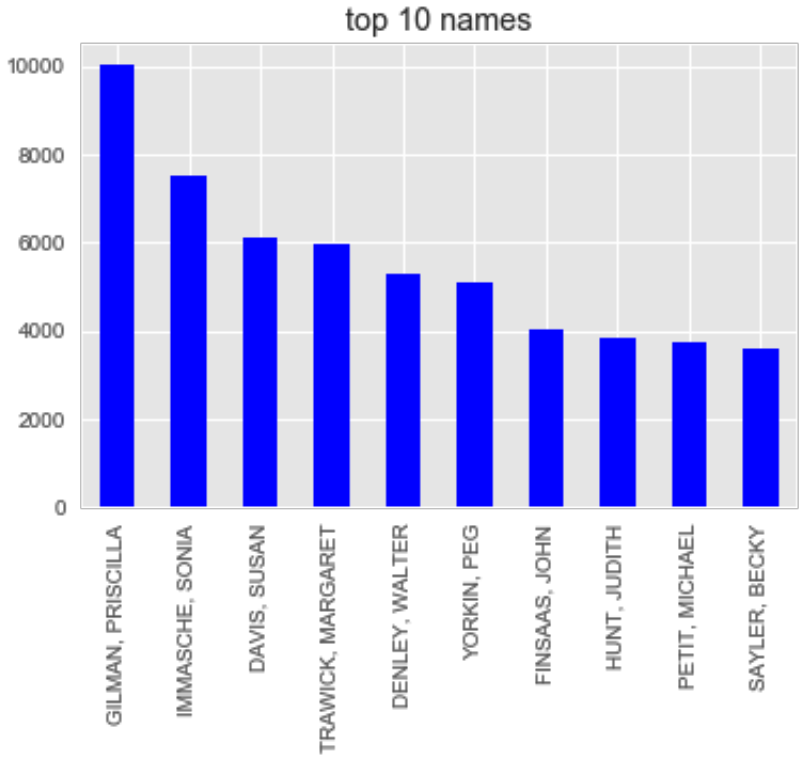
In [14]:
```
sched_a_df.CITY.value_counts()[:10].plot(kind='bar', color='b', title=
'top 10 cities')
```

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x14bc8e2b0>



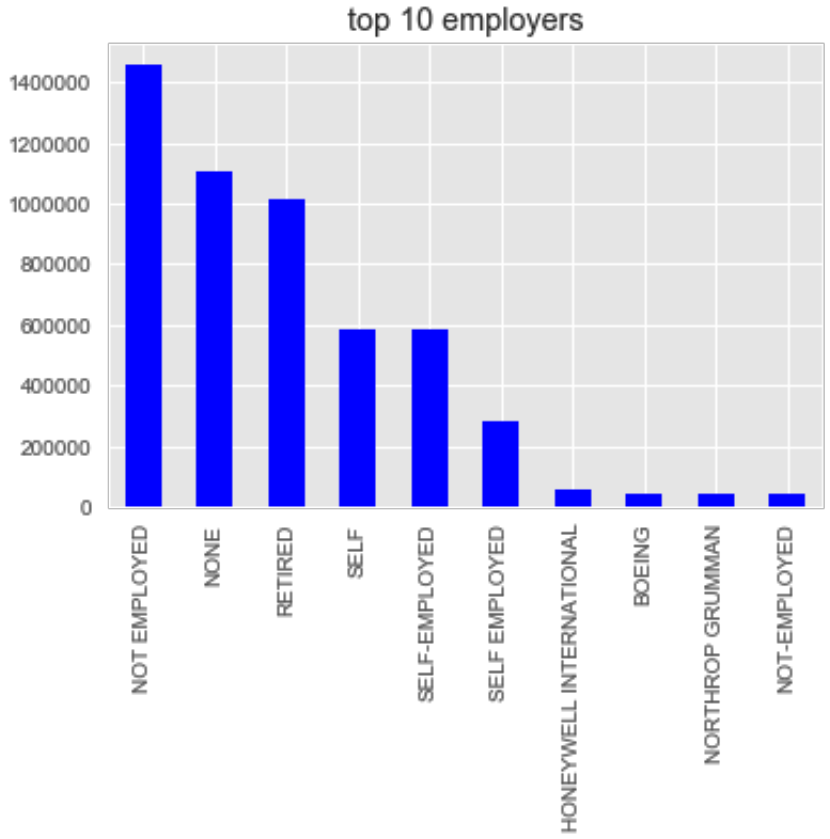
In [15]:
```
sched_a_df.NAME.value_counts()[:10].plot(kind='bar', color='b', title=
'top 10 names')
```

Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x14c3a9c50>



In [16]:
```
sched_a_df.EMPLOYER.value_counts()[:10].plot(kind='bar', color='b', title='top 10 employers')
```

Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x14bf18ef0>



In [17]:
```
sched_a_df['TRANSACTION_AMT'] = sched_a_df['TRANSACTION_AMT'].astype(int)
sched_a_df['TRANSACTION_AMT'].max()
```

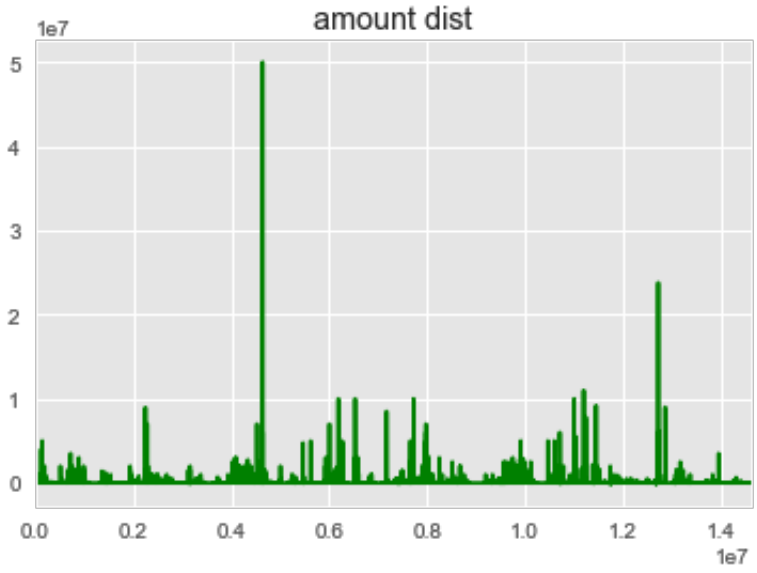
Out[17]: 50000000

In [18]:
```
sched_a_df['TRANSACTION_AMT'].min()
```

Out[18]: -200000

In [19]:
```
sched_a_df['TRANSACTION_AMT'].plot(title='amount dist', color='g')
```

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x14c5f7908>



In [ ]:
```

```