```python
In [60]: #2012 Election Mit and Barack from HuffPost Pollster> polls data
         import pandas as pd
         from pandas import Series, DataFrame
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         sns.set_style('whitegrid')
         %matplotlib inline
         #from __future__ import division
         import requests # API to gather web info
         from io import StringIO # for reading data
```

```python
In [61]: url='https://elections.huffingtonpost.com/pollster/2012-general-election-romney-vs-obama.csv'
         source=requests.get(url).text
         poll_data=StringIO(source)
         poll_data
```

```
Out[61]: <_io.StringIO at 0x27216625558>
```

```python
In [62]: #poll_df=pd.read_csv(poll_data, header=None,error_bad_lines=False)
         poll_df=pd.read_csv(poll_data)
         poll_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 586 entries, 0 to 585
Data columns (total 17 columns):
Pollster                586 non-null object
Start Date              586 non-null object
End Date                586 non-null object
Entry Date/Time (ET)    586 non-null object
Number of Observations  564 non-null float64
Population              586 non-null object
Mode                    586 non-null object
Obama                   586 non-null float64
Romney                  586 non-null float64
Undecided               423 non-null float64
Other                   202 non-null float64
Pollster URL            586 non-null object
Source URL              584 non-null object
Partisan                586 non-null object
Affiliation             586 non-null object
Question Text           0 non-null float64
Question Iteration      586 non-null int64
dtypes: float64(6), int64(1), object(10)
memory usage: 78.0+ KB
```

```python
In [63]: poll_df.head()
```

Out[63]:

	Pollster	Start Date	End Date	Entry Date/Time (ET)	Number of Observations	Population	Mode	Obama	Romney	Undecided	Other	
0	Politico/GWU/Battleground	2012-11-04	2012-11-05	2012-11-06T08:40:26Z	1000.0	Likely Voters	Live Phone	47.0	47.0	6.0	NaN	http://elections.
1	YouGov/Economist	2012-11-03	2012-11-05	2012-11-26T15:31:23Z	740.0	Likely Voters	Internet	49.0	47.0	3.0	NaN	http://elections.
2	Gravis Marketing	2012-11-03	2012-11-05	2012-11-06T09:22:02Z	872.0	Likely Voters	Automated Phone	48.0	48.0	4.0	NaN	http://elections.
3	IBD/TIPP	2012-11-03	2012-11-05	2012-11-06T08:51:48Z	712.0	Likely Voters	Live Phone	50.0	49.0	NaN	1.0	http://elections.
4	Rasmussen	2012-11-03	2012-11-05	2012-11-06T08:47:50Z	1500.0	Likely Voters	Automated Phone	48.0	49.0	NaN	NaN	http://elections.

```
In [64]: sns.factorplot('Number of Observations', data=poll_df) #, hue='Partisan')
```

Out[64]: `<seaborn.axisgrid.FacetGrid at 0x272168fc438>`
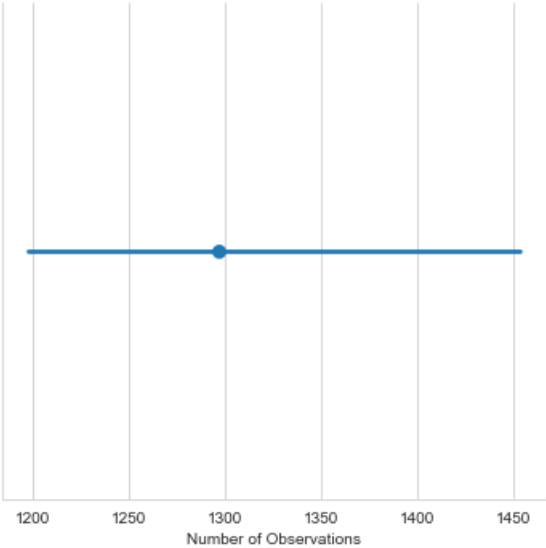

```
In [65]: avg=pd.DataFrame(poll_df.mean())
         avg.drop('Number of Observations', axis=0, inplace=True)
         avg
```

Out[65]:

|	0
Obama	46.805461
Romney	44.614334
Undecided	6.550827
Other	3.376238
Question Text	NaN
Question Iteration	1.000000

```
In [66]: std=pd.DataFrame(poll_df.std())
         std.drop('Number of Observations', axis=0, inplace=True)
         std
```

Out[66]:

	0
Obama	2.422058
Romney	2.906180
Undecided	3.701754
Other	2.692726
Question Text	NaN
Question Iteration	0.000000

```
In [67]: avg.plot(yerr=std, kind='bar', legend=False)
```

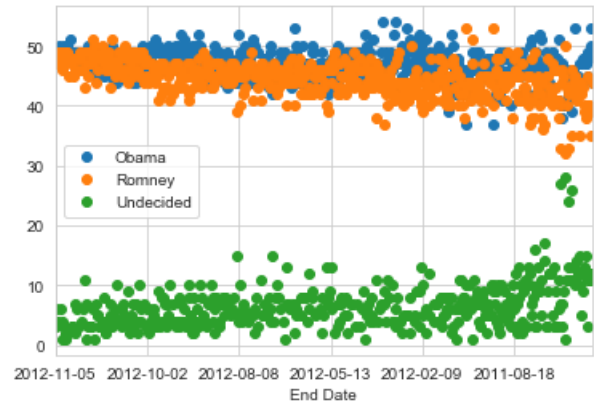Out[67]: <matplotlib.axes._subplots.AxesSubplot at 0x272181037f0>
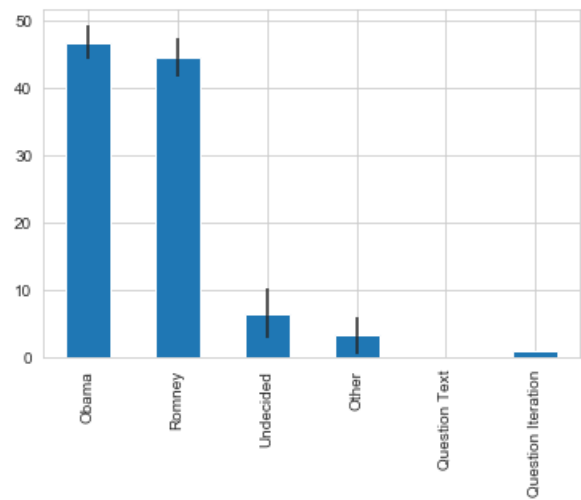

```
In [68]: poll_avg=pd.concat([avg,std], axis=1)
         poll_avg.columns=['Average', 'Std']
         poll_avg
```

Out[68]:

	Average	Std
Obama	46.805461	2.422058
Romney	44.614334	2.906180
Undecided	6.550827	3.701754
Other	3.376238	2.692726
Question Text	NaN	NaN
Question Iteration	1.000000	0.000000

```
In [69]: poll_df.plot(x='End Date', y=['Obama', 'Romney', 'Undecided'], linestyle=' ', marker='o')
```

Out[69]: <matplotlib.axes._subplots.AxesSubplot at 0x2721816f668>

```
In [70]: from datetime import datetime
         poll_df['Difference']=(poll_df.Obama-poll_df.Romney)/100 # negative difference --> Romney lead
         poll_df.head()
```

Out[70]:

	Pollster	Start Date	End Date	Entry Date/Time (ET)	Number of Observations	Population	Mode	Obama	Romney	Undecided	Other	
0	Politico/GWU/Battleground	2012-11-04	2012-11-05	2012-11-06T08:40:26Z	1000.0	Likely Voters	Live Phone	47.0	47.0	6.0	NaN	http://elections.
1	YouGov/Economist	2012-11-03	2012-11-05	2012-11-26T15:31:23Z	740.0	Likely Voters	Internet	49.0	47.0	3.0	NaN	http://elections.
2	Gravis Marketing	2012-11-03	2012-11-05	2012-11-06T09:22:02Z	872.0	Likely Voters	Automated Phone	48.0	48.0	4.0	NaN	http://elections.
3	IBD/TIPP	2012-11-03	2012-11-05	2012-11-06T08:51:48Z	712.0	Likely Voters	Live Phone	50.0	49.0	NaN	1.0	http://elections.
4	Rasmussen	2012-11-03	2012-11-05	2012-11-06T08:47:50Z	1500.0	Likely Voters	Automated Phone	48.0	49.0	NaN	NaN	http://elections.

```
In [71]: poll_df=poll_df.groupby(['Start Date'], as_index=False).mean()
         poll_df
```

Out[71]:

	Start Date	Number of Observations	Obama	Romney	Undecided	Other	Question Text	Question Iteration	Difference
0	2009-03-13	1403.000000	44.00	44.00	12.000000	NaN	NaN	1	0.0000
1	2009-04-17	686.000000	50.00	39.00	11.000000	NaN	NaN	1	0.1100
2	2009-05-14	1000.000000	53.00	35.00	12.000000	NaN	NaN	1	0.1800
3	2009-06-12	638.000000	48.00	40.00	12.000000	NaN	NaN	1	0.0800
4	2009-07-15	577.000000	49.00	40.00	11.000000	NaN	NaN	1	0.0900
...	...	...	...	...	...	...	...	...	...
352	2012-10-31	10420.250000	48.00	46.50	4.000000	2.0	NaN	1	0.0150
353	2012-11-01	1989.000000	49.00	47.25	2.400000	1.5	NaN	1	0.0175
354	2012-11-02	1041.750000	49.25	48.50	2.000000	1.0	NaN	1	0.0075
355	2012-11-03	1310.833333	48.50	48.00	4.333333	1.0	NaN	1	0.0050
356	2012-11-04	1000.000000	47.00	47.00	6.000000	NaN	NaN	1	0.0000

357 rows × 9 columns

```
In [72]: poll_df.plot('Start Date', 'Difference', figsize=(12,4), marker='o', linestyle='-', color='blue')
```

Out[72]: <matplotlib.axes._subplots.AxesSubplot at 0x27218103a58>
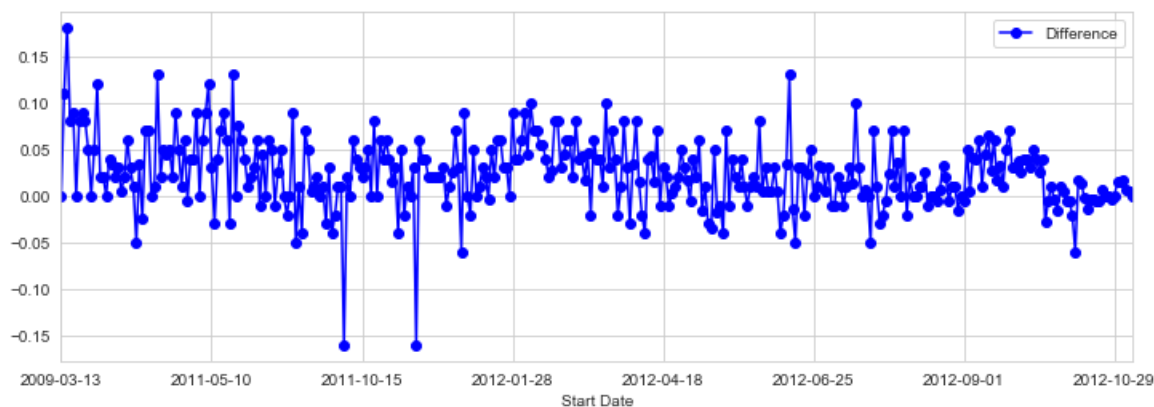
```
In [73]:   #Find data in oct 2012
           row_in=0
           xlimit=[]
           for data in poll_df['Start Date']:
               if data[0:7]=='2012-10':
                   xlimit.append(row_in)
                   row_in+=1
               else:
                   row_in+=1
           print (min(xlimit))
           print (max(xlimit))


           325
           352
```
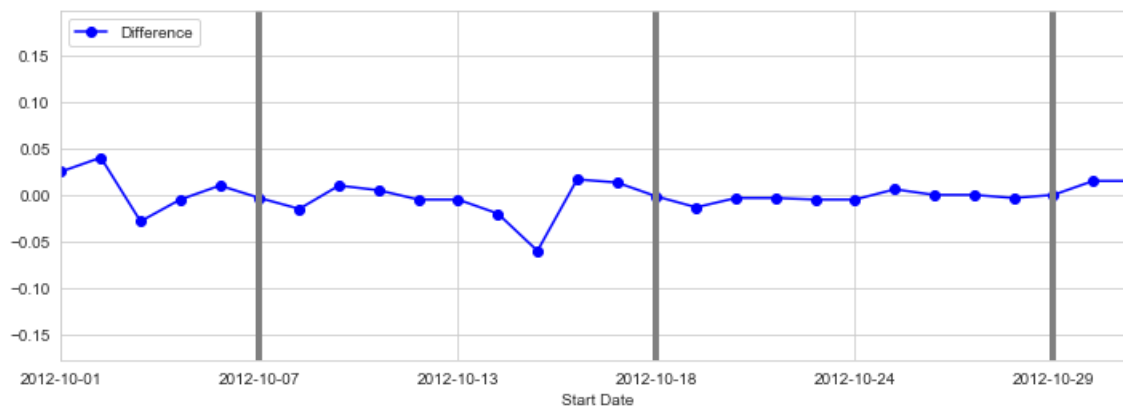
```
In [74]:   poll_df.plot('Start Date', 'Difference', xlim=(325,352) ,figsize=(12,4), marker='o', linestyle='-', color='blue')
           #axvline() plot a vertical line
           plt.axvline(x=325+5, linewidth=4, color='grey')
           plt.axvline(x=325+15, linewidth=4, color='grey')
           plt.axvline(x=325+25, linewidth=4, color='grey')
```

Out[74]:   <matplotlib.lines.Line2D at 0x2721827c710>


```
In [75]:   donor_df=pd.read_csv('Election_Donor_Data.csv')
           donor_df['contb_receipt_amt'].value_counts()
```

```
Out[75]:   100.00     178188
           50.00      137584
           25.00      110345
           250.00      91182
           500.00      57984
                       ...
           97.15           1
           122.32          1
           188.65          1
           122.40          1
           132.12          1
           Name: contb_receipt_amt, Length: 8079, dtype: int64
```

```
In [76]:   #!pwd
           #cwd
           import os
           os.getcwd()
```

Out[76]:   'C:\\Users\\Lyci\\DrillsLG'

```
In [77]:   don_mean=donor_df['contb_receipt_amt'].mean()
           don_std=donor_df['contb_receipt_amt'].std()
           print ("The average donation was %.2f with a std deviation %.2f" %(don_mean, don_std))


           The average donation was 298.24 with a std deviation 3749.67
```

```
In [78]: #big difference between avg and std ...find outliers
         top_donor=donor_df['contb_receipt_amt'].copy()
         top_donor.sort_values() # gives -ve values/refunds...need to be removed
```

Out[78]: 114604   -30800.00
         226986   -25800.00
         101356   -7500.00
         398429   -5500.00
         250737   -5455.00
                       ...
         319478   526246.17
         344419   1511192.17
         344539   1679114.65
         326651   1944042.43
         325136   2014490.51
         Name: contb_receipt_amt, Length: 1001731, dtype: float64

```
In [79]: top_donor=top_donor[top_donor>0]
         top_donor.sort_values()
```

Out[79]: 335573   0.01
         335407   0.01
         335352   0.01
         324596   0.01
         329896   0.01
                       ...
         319478   526246.17
         344419   1511192.17
         344539   1679114.65
         326651   1944042.43
         325136   2014490.51
         Name: contb_receipt_amt, Length: 991475, dtype: float64

```
In [80]: top_donor.value_counts().head(10)
```

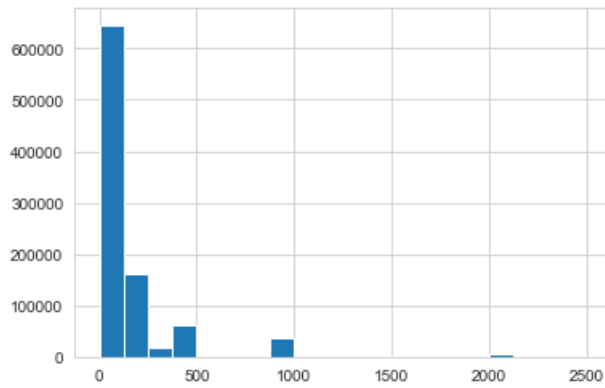Out[80]: 100.0   178188
         50.0   137584
         25.0   110345
         250.0   91182
         500.0   57984
         2500.0   49005
         35.0   37237
         1000.0   36494
         10.0   33986
         200.0   27813
         Name: contb_receipt_amt, dtype: int64

```
In [81]: com_don=top_donor[top_donor<2500]
         com_don
```

Out[81]: 0   250.0
         1   50.0
         2   250.0
         3   250.0
         4   300.0
                      ...
         1001722   100.0
         1001723   100.0
         1001724   500.0
         1001728   500.0
         1001729   500.0
         Name: contb_receipt_amt, Length: 940511, dtype: float64
```

```
In [82]: com_don.hist(bins=20)
```

Out[82]: <matplotlib.axes._subplots.AxesSubplot at 0x2722455c160>


```
In [83]: candidates=donor_df.cand_nm.unique()
         candidates # see Obama among republcan list, that need a party affiliation column
```

Out[83]: array(['Bachmann, Michelle', 'Romney, Mitt', 'Obama, Barack',
               "Roemer, Charles E. 'Buddy' III", 'Pawlenty, Timothy',
               'Johnson, Gary Earl', 'Paul, Ron', 'Santorum, Rick',
               'Cain, Herman', 'Gingrich, Newt', 'McCotter, Thaddeus G',
               'Huntsman, Jon', 'Perry, Rick'], dtype=object)
```

```
In [84]: donor_df
```

Out[84]:

	cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_employer	contbr_occupation	contb_r
0	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RETIRED	RETIRED	
1	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RETIRED	RETIRED	
2	C00410118	P20002978	Bachmann, Michelle	SMITH, LANIER	LANETT	AL	3.68633e+08	INFORMATION REQUESTED	INFORMATION REQUESTED	
3	C00410118	P20002978	Bachmann, Michelle	BLEVINS, DARONDA	PIGGOTT	AR	7.24548e+08	NONE	RETIRED	
4	C00410118	P20002978	Bachmann, Michelle	WARDENBURG, HAROLD	HOT SPRINGS NATION	AR	7.19016e+08	NONE	RETIRED	
...	...	...	...	...	...	...	...	...	...	
1001726	C00500587	P20003281	Perry, Rick	GORMAN, CHRIS D. MR.	INFO REQUESTED	XX	99999	INFORMATION REQUESTED PER BEST EFFORTS	INFORMATION REQUESTED PER BEST EFFORTS	
1001727	C00500587	P20003281	Perry, Rick	DUFFY, DAVID A. MR.	INFO REQUESTED	XX	99999	DUFFY EQUIPMENT COMPANY INC.	BUSINESS OWNER	
1001728	C00500587	P20003281	Perry, Rick	GRANE, BRYAN F. MR.	INFO REQUESTED	XX	99999	INFORMATION REQUESTED PER BEST EFFORTS	INFORMATION REQUESTED PER BEST EFFORTS	
1001729	C00500587	P20003281	Perry, Rick	TOLBERT, DARYL MR.	INFO REQUESTED	XX	99999	T.A.C.C.	LONGWALL MAINTENANCE FOREMAN	
1001730	C00500587	P20003281	Perry, Rick	ANDERSON, MARILEE MRS.	INFO REQUESTED	XX	99999	INFORMATION REQUESTED PER BEST EFFORTS	INFORMATION REQUESTED PER BEST EFFORTS	

1001731 rows × 16 columns

```
In [85]: '''for i in range(0, len(donor_df)):
         if donor_df['cand_nm'][i]=='Obama, Barack':
             donor_df['cand_party'][i]='Democrat'
         else:
             donor_df['cand_party'][i]='Republican' '''
```

Out[85]: "for i in range(0, len(donor_df)):\n   if donor_df['cand_nm'][i]=='Obama, Barack':\n   donor_df['cand_party'][i]='Democrat'\n   else:\n   donor_df['cand_party'][i]='Republican' "

```
In [86]: # Dictionary of party affiliation
         party_map = {'Bachmann, Michelle': 'Republican','Cain, Herman': 'Republican', 'Gingrich, Newt': 'Republican', 'Huntsm
         # Now map the party with candidate
         donor_df['Party'] = donor_df.cand_nm.map(party_map)
```

```
In [87]: '''for i in range(0, len(donor_df)):
             if donor_df['cand_nm'][i]=='Obama, Barack':
                 donor_df['Party'][i]='Democrat'
             else:
                 donor_df['Party'][i]='Republican' '''
         donor_df.head()
```

Out[87]:

	cmte_id	cand_id	cand_nm	contbr_nm	contbr_city	contbr_st	contbr_zip	contbr_employer	contbr_occupation	contb_receipt_a
0	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RETIRED	RETIRED	250
1	C00410118	P20002978	Bachmann, Michelle	HARVEY, WILLIAM	MOBILE	AL	3.6601e+08	RETIRED	RETIRED	50
2	C00410118	P20002978	Bachmann, Michelle	SMITH, LANIER	LANETT	AL	3.68633e+08	INFORMATION REQUESTED	INFORMATION REQUESTED	250
3	C00410118	P20002978	Bachmann, Michelle	BLEVINS, DARONDA	PIGGOTT	AR	7.24548e+08	NONE	RETIRED	250
4	C00410118	P20002978	Bachmann, Michelle	WARDENBURG, HAROLD	HOT SPRINGS NATION	AR	7.19016e+08	NONE	RETIRED	300

```
In [88]: donor_df.groupby('cand_nm')['contb_receipt_amt'].count()
```

```
Out[88]: cand_nm
         Bachmann, Michelle              13140
         Cain, Herman                    20107
         Gingrich, Newt                  47679
         Huntsman, Jon                    4156
         Johnson, Gary Earl               1234
         McCotter, Thaddeus G               74
         Obama, Barack                  593746
         Paul, Ron                      143757
         Pawlenty, Timothy                4555
         Perry, Rick                     13575
         Roemer, Charles E. 'Buddy' III   5920
         Romney, Mitt                   107229
         Santorum, Rick                  46559
         Name: contb_receipt_amt, dtype: int64
```

```
In [89]: cand_amount=donor_df.groupby('cand_nm')['contb_receipt_amt'].sum()
         cand_amount
```

```
Out[89]: cand_nm
         Bachmann, Michelle             2.677435e+06
         Cain, Herman                   7.047265e+06
         Gingrich, Newt                 1.201183e+07
         Huntsman, Jon                  3.204350e+06
         Johnson, Gary Earl             5.669616e+05
         McCotter, Thaddeus G           3.703000e+04
         Obama, Barack                  1.335026e+08
         Paul, Ron                      2.072257e+07
         Pawlenty, Timothy              4.255054e+06
         Perry, Rick                    1.841151e+07
         Roemer, Charles E. 'Buddy' III 3.674575e+05
         Romney, Mitt                   8.521925e+07
         Santorum, Rick                 1.072809e+07
         Name: contb_receipt_amt, dtype: float64
```

```
In [90]: i=0
         for donor in cand_amount:
             print('The candidate %s raise %0.0f dollars'% (cand_amount.index[i], donor))
             print ('\n')
             i+=1
```

The candidate Bachmann, Michelle raise 2677435 dollars

The candidate Cain, Herman raise 7047265 dollars

The candidate Gingrich, Newt raise 12011832 dollars

The candidate Huntsman, Jon raise 3204350 dollars

The candidate Johnson, Gary Earl raise 566962 dollars

The candidate McCotter, Thaddeus G raise 37030 dollars

The candidate Obama, Barack raise 133502591 dollars

The candidate Paul, Ron raise 20722567 dollars

The candidate Pawlenty, Timothy raise 4255054 dollars

The candidate Perry, Rick raise 18411512 dollars

The candidate Roemer, Charles E. 'Buddy' III raise 367458 dollars

The candidate Romney, Mitt raise 85219249 dollars

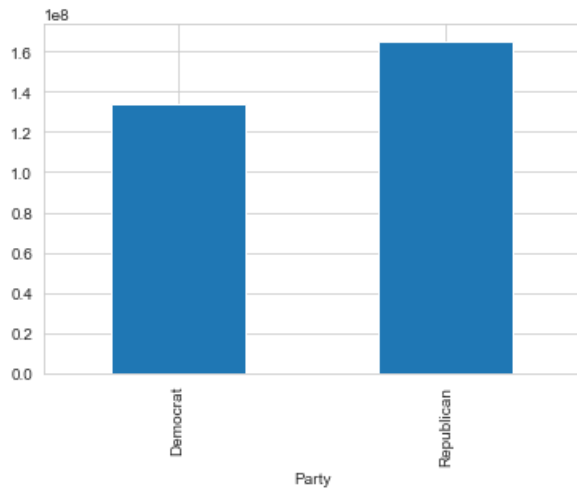The candidate Santorum, Rick raise 10728090 dollars


```
In [91]: cand_amount.plot(kind='bar')
```

Out[91]: <matplotlib.axes._subplots.AxesSubplot at 0x27224570d68>
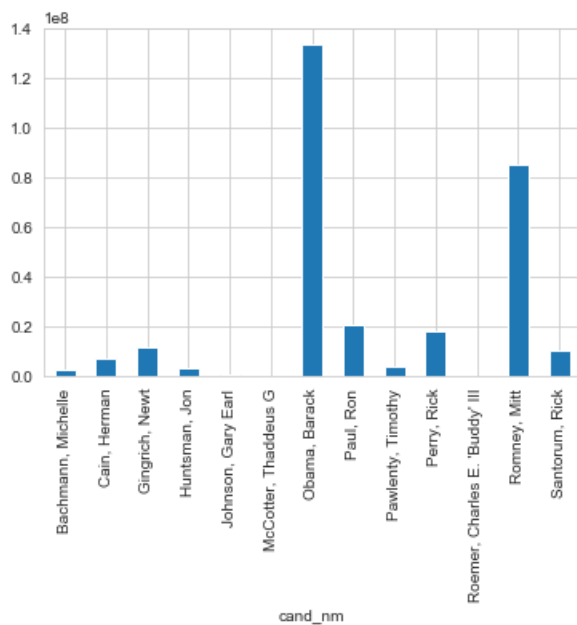
```
In [92]: donor_df.groupby('Party')['contb_receipt_amt'].sum().plot(kind='bar')
```

Out[92]: <matplotlib.axes._subplots.AxesSubplot at 0x27218f2c630>


```
In [93]: occupation_df = donor_df.pivot_table('contb_receipt_amt',
                                               index='contbr_occupation',
                                               columns='Party', aggfunc='sum')

         occupation_df.tail()
```

Out[93]:

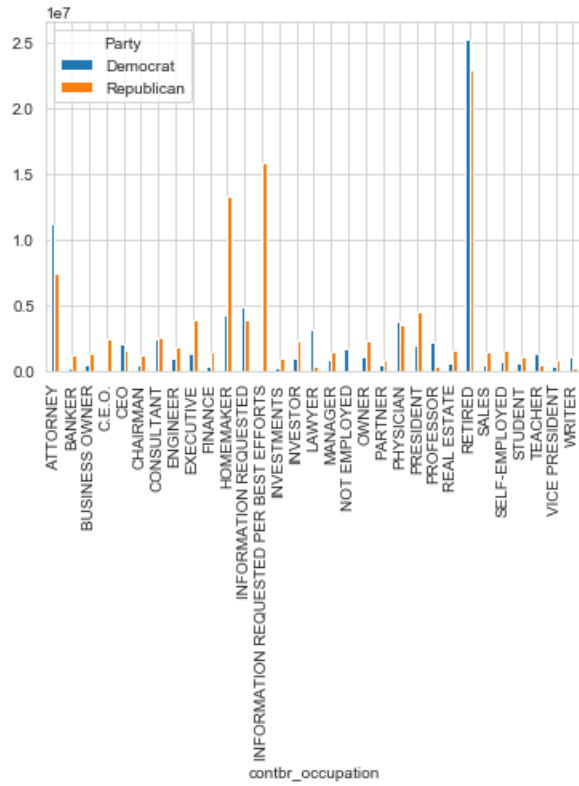| Party	Democrat	Republican
contbr_occupation		
ZOOKEEPER	35.0	NaN
ZOOLOGIST	400.0	NaN
ZOOLOGY EDUCATION	25.0	NaN
\NONE\	NaN	250.0
~	NaN	75.0

```
In [94]: occupation_df.shape
```

Out[94]: (45073, 2)

In [95]:
```python
occupation_df= occupation_df[occupation_df.sum(1)>1000000]

occupation_df.plot(kind='bar')
```
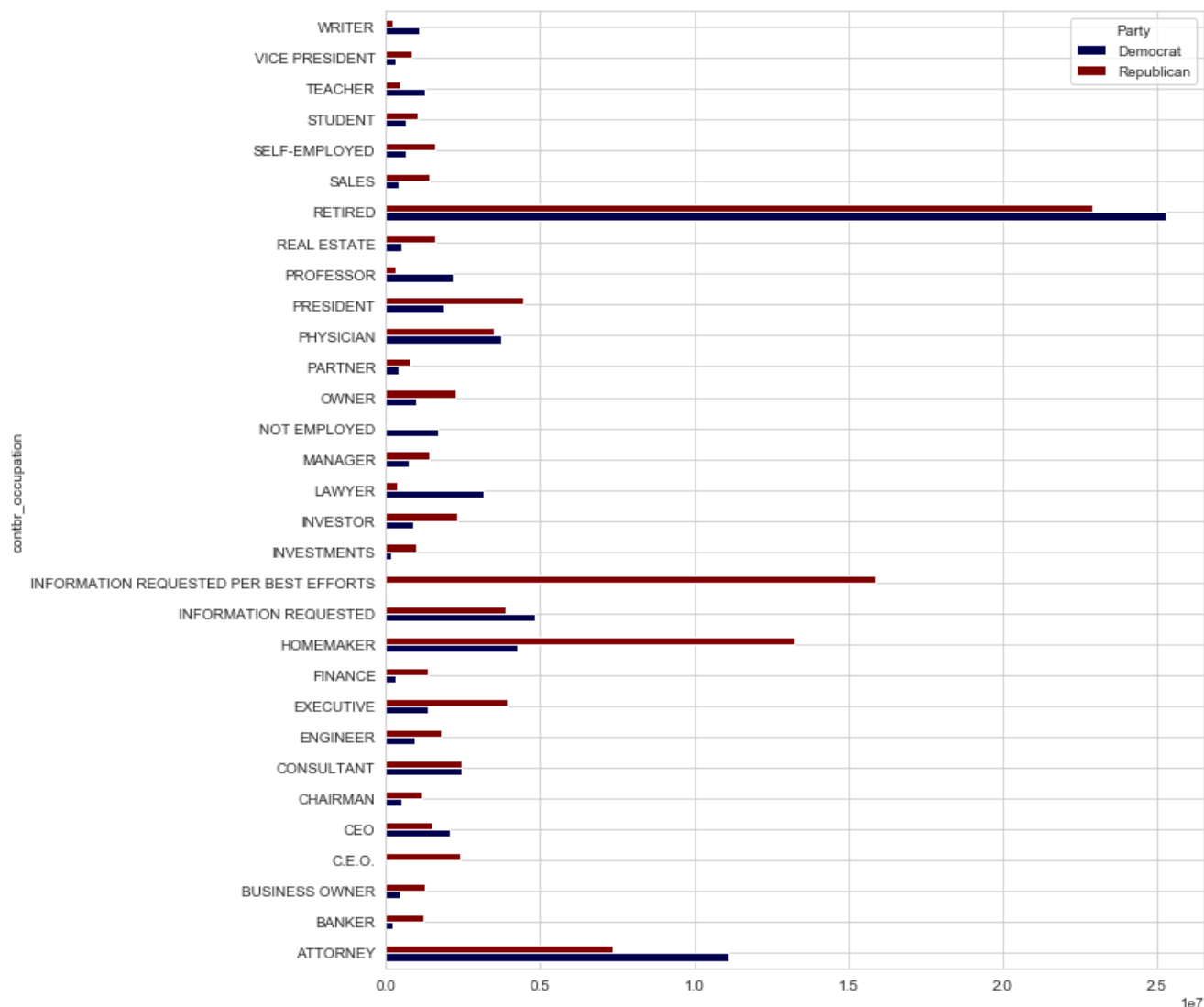
Out[95]: <matplotlib.axes._subplots.AxesSubplot at 0x27218f956d8>

```
In [96]: occupation_df.plot(kind='barh',figsize=(10,12),cmap='seismic')
```

```
Out[96]: <matplotlib.axes._subplots.AxesSubplot at 0x272190c78d0>
```



```
In [104]: # combine CEO and C.E.O , remove information requested...
          #occupation_df.drop(['INFORMATION REQUESTED PER BEST EFFORTS'], axis=0, inplace=True)
```
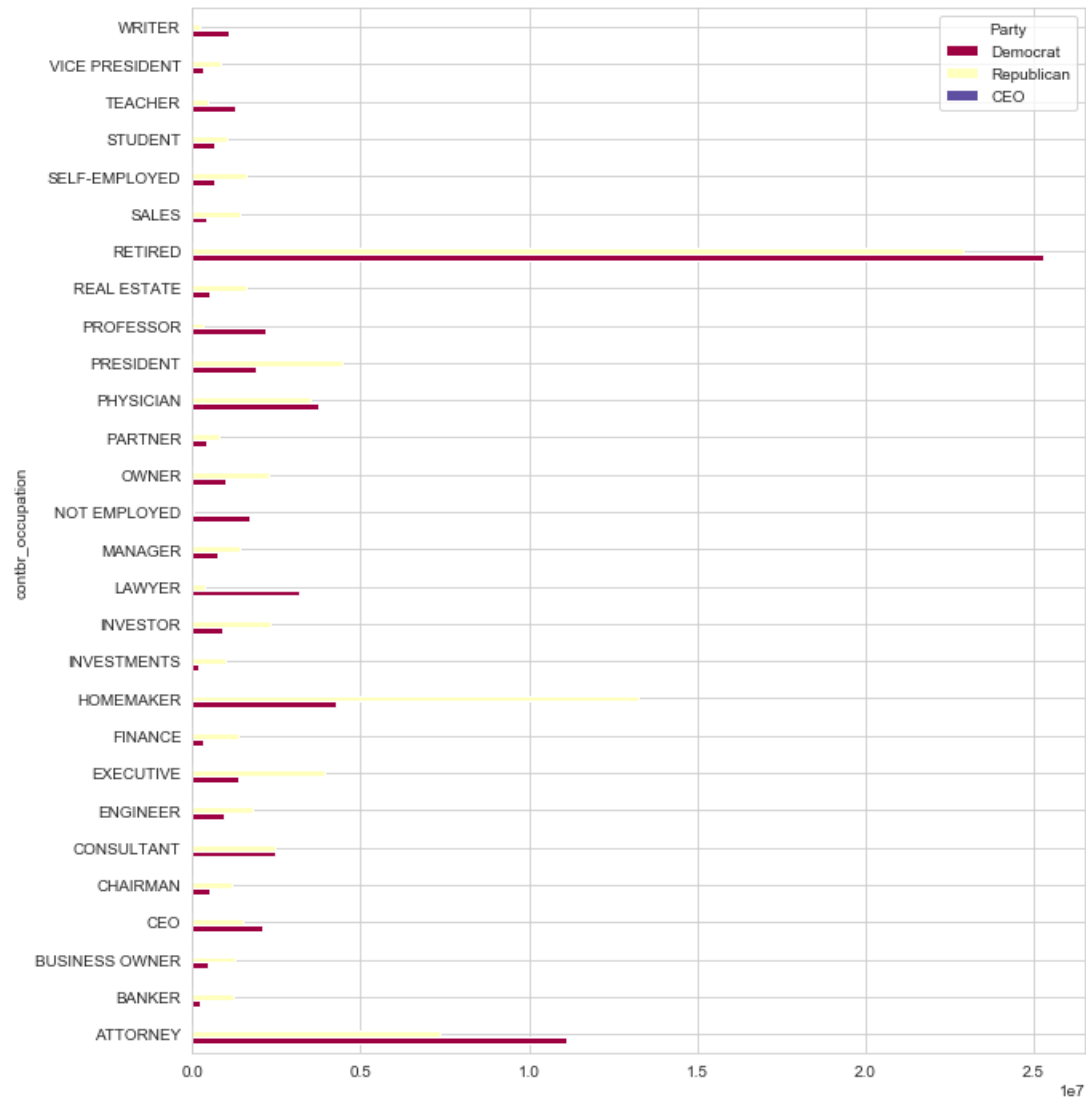
```
In [106]: #occupation_df['CEO']=occupation_df.loc['CEO']+occupation_df.loc['C.E.O.']
```

```
In [108]: #occupation_df.drop('C.E.O.', inplace=True)
          #occupation_df.plot(kind='barh',figsize=(10,12),cmap='seismic')
```

```
In [109]: occupation_df.drop(['INFORMATION REQUESTED'], axis=0, inplace=True)
```

`occupation_df.plot(kind='barh',figsize=(10,12),cmap='Spectral')`

`<matplotlib.axes._subplots.AxesSubplot at 0x272193c00f0>`

In [ ]: