Business Analytics Programming

Lab 1 (Pandas, NY 2016 Fundraising Data)

Dr. Wajahat Gilani

Rutgers Business School

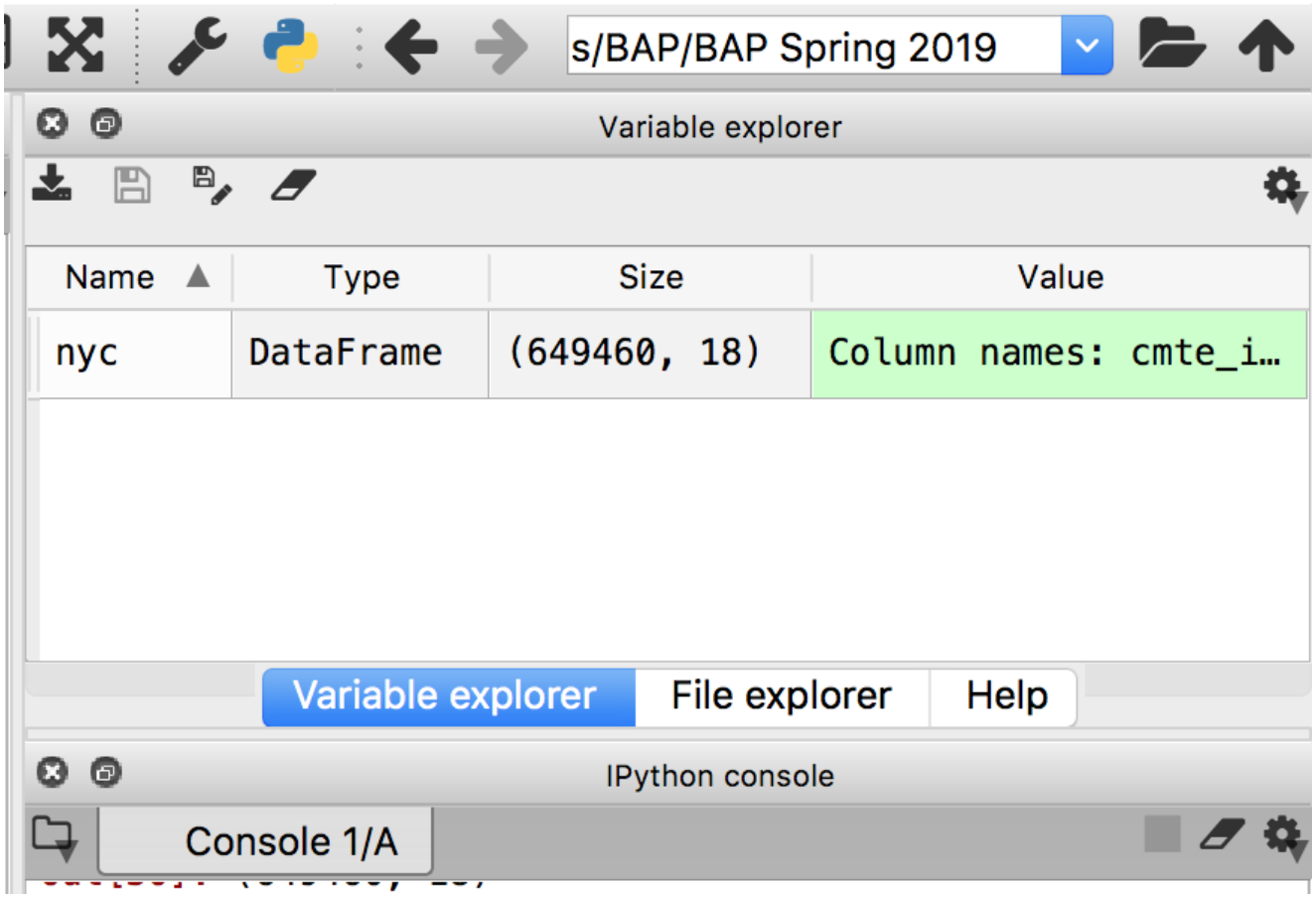
January 31, 2019

The nyc.csv File

The file ny.csv is a list of all the individual donors (no super-pacs) from the state of New York that donated money to the any of the candidates.

Load in nyc.csv File

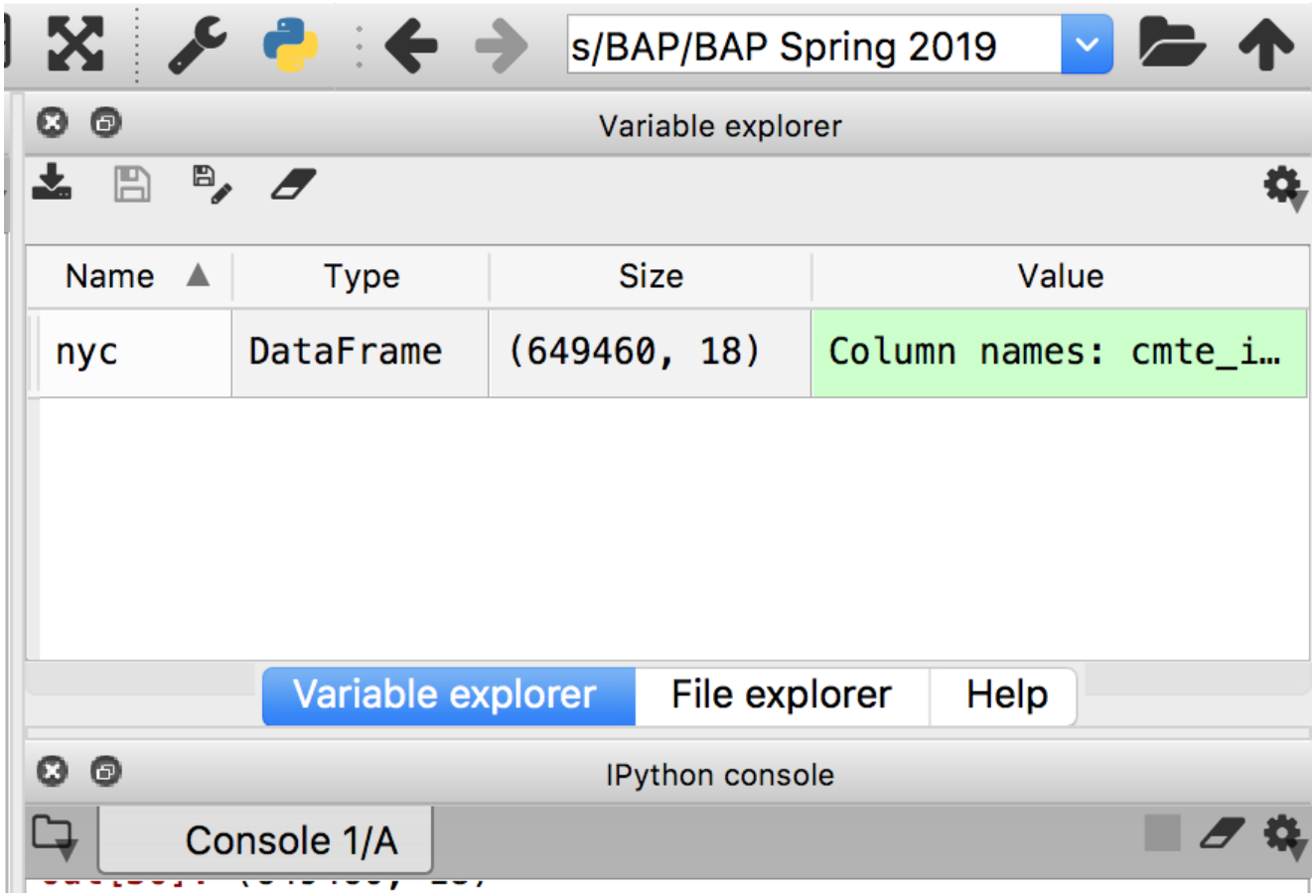
- Download the ny.csv file from BlackBoard
- In spyder, change your folder



- Load the file into a pandas DataFrame

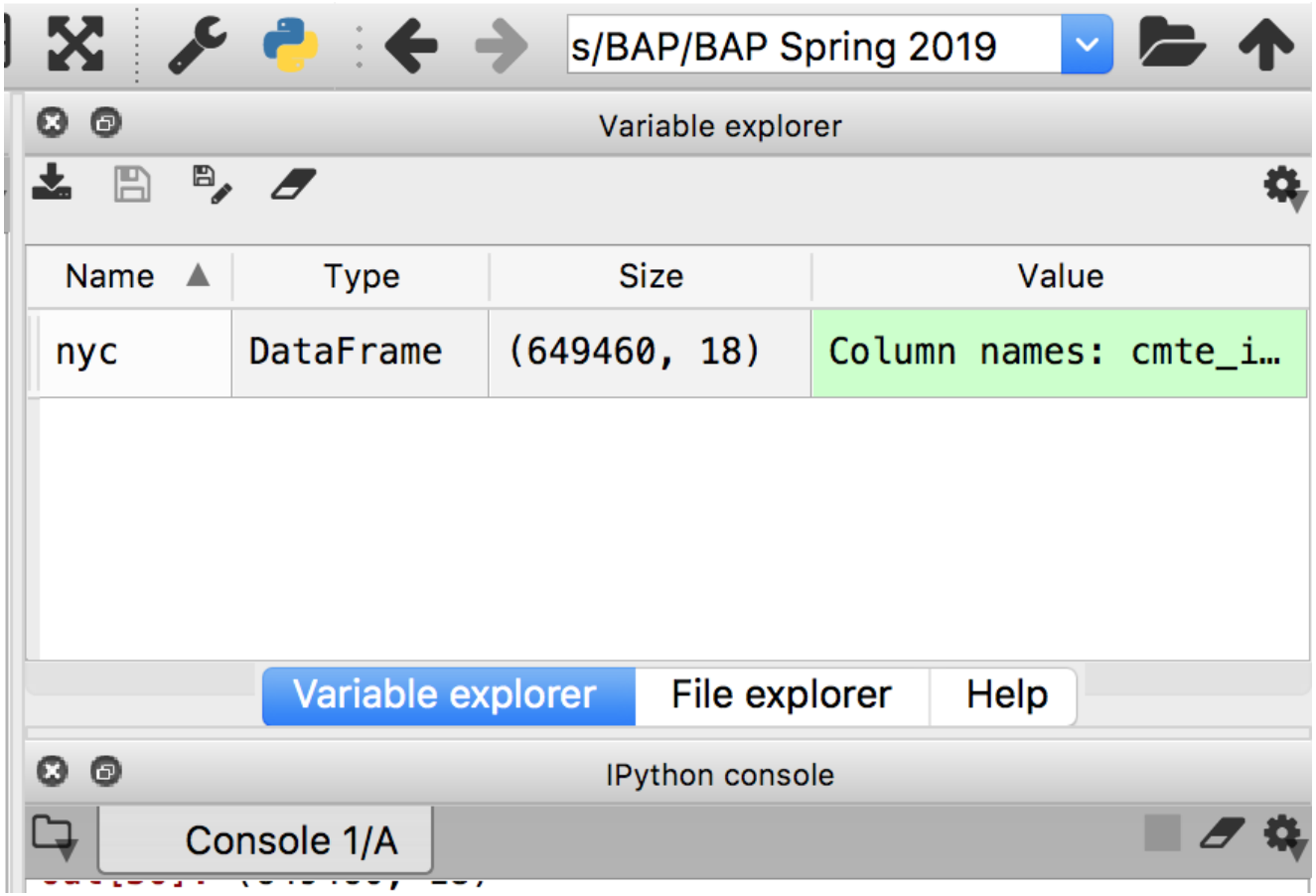
Load in nyc.csv File (Continued)

```
1 import pandas as pd
2 import numpy as np
3
4 nyc = pd.read_csv('ny.csv', index_col=False)
```



Load in nyc.csv File (Continued)

```
1 import pandas as pd
2 import numpy as np
3
4 nyc = pd.read_csv('ny.csv', index_col=False)
```



Dataframes and Series

A Series is a one-dimensional object that can hold any data type such as integers, floats and strings. Let's take a list of items as an input argument and create a Series object for that list.

A DataFrame is a two dimensional object that can have columns with potential different types. Different kind of inputs include dictionaries, lists, series, and even another DataFrame.

Exploring DataFrames

```
1 nyc.shape
```

(649460, 18)

```
1 nyc.columns
```

```
Index(['cmte_id', 'cand_id', 'cand_nm', 'contbr_nm',
'contbr_city',
       'contbr_st', 'contbr_zip', 'contbr_employer',
'contbr_occupation',
       'contb_receipt_amt', 'contb_receipt_dt',
'receipt_desc', 'memo_cd',
       'memo_text', 'form_tp', 'file_num', 'tran_id',
'election_tp'],
      dtype='object')
```

```
1 nyc.dtypes
```

```
cmte_id              object
cand_id              object
cand_nm              object
contbr_nm            object
contbr_city          object
contbr_st            object
contbr_zip           object
contbr_employer      object
contbr_occupation    object
contb_receipt_amt   float64
contb_receipt_dt     object
receipt_desc         object
memo_cd              object
memo_text            object
form_tp              object
file_num              int64
tran_id              object
election_tp          object
```

Exploring DataFrames (Continued)

df.head() - Lets you see the first 5 rows.

df.tail() - Lets you see the last 5 rows.

df.head(n) - You can see the first n rows.

df.tail(n) - You can see the last n rows.

```
1 nyc['cand_nm']
```

List a column of values.

```
1 nyc['cand_nm'].value_counts()
```

```
Clinton, Hillary Rodham      399522
Sanders, Bernard             174564
Trump, Donald J.              36931
Cruz, Rafael Edward 'Ted'     16785
Carson, Benjamin S.            6638
Rubio, Marco                   4813
Bush, Jeb                      2436
Kasich, John R.                1350
Fiorina, Carly                 1218
Paul, Rand                     1141
Stein, Jill                    1001
Johnson, Gary                   782
Christie, Christopher J.        486
Graham, Lindsey O.              362
O'Malley, Martin Joseph         343
Walker, Scott                   265
Huckabee, Mike                  254
Pataki, George E.               182
Lessig, Lawrence                116
McMullin, Evan                  103
Santorum, Richard J.             69
Webb, James Henry Jr.            46
Perry, James R. (Rick)           27
Jindal, Bobby                    21
Gilmore, James S III              5
```

Exploring DataFrames (Continued)

df.head() - Lets you see the first 5 rows.

df.tail() - Lets you see the last 5 rows.

df.head(n) - You can see the first n rows.

df.tail(n) - You can see the last n rows.

```
1 nyc['cand_nm']
```

List a column of values.

```
1 nyc['cand_nm'].value_counts()
```

```
Clinton, Hillary Rodham      399522
Sanders, Bernard             174564
Trump, Donald J.              36931
Cruz, Rafael Edward 'Ted'     16785
Carson, Benjamin S.            6638
Rubio, Marco                   4813
Bush, Jeb                      2436
Kasich, John R.                1350
Fiorina, Carly                 1218
Paul, Rand                     1141
Stein, Jill                    1001
Johnson, Gary                   782
Christie, Christopher J.        486
Graham, Lindsey O.              362
O'Malley, Martin Joseph         343
Walker, Scott                   265
Huckabee, Mike                  254
Pataki, George E.               182
Lessig, Lawrence                116
McMullin, Evan                  103
Santorum, Richard J.             69
Webb, James Henry Jr.            46
Perry, James R. (Rick)           27
Jindal, Bobby                    21
Gilmore, James S III              5
```

Exploring DataFrames (Continued)

```
1 nyc['cand_nm'].value_counts(normalize=True)
```

```
Clinton, Hillary Rodham      0.615160
Sanders, Bernard             0.268783
Trump, Donald J.             0.056864
Cruz, Rafael Edward 'Ted'    0.025845
Carson, Benjamin S.          0.010221
Rubio, Marco                 0.007411
Bush, Jeb                    0.003751
Kasich, John R.              0.002079
Fiorina, Carly               0.001875
Paul, Rand                   0.001757
Stein, Jill                  0.001541
Johnson, Gary                0.001204
Christie, Christopher J.     0.000748
Graham, Lindsey O.           0.000557
O'Malley, Martin Joseph      0.000528
Walker, Scott                0.000408
Huckabee, Mike               0.000391
Pataki, George E.            0.000280
Lessig, Lawrence             0.000179
McMullin, Evan               0.000159
Santorum, Richard J.         0.000106
Webb, James Henry Jr.        0.000071
Perry, James R. (Rick)       0.000042
Jindal, Bobby                0.000032
Gilmore, James S III         0.000008
```

```
1 pd.isnull(nyc.contbr_employer).value_counts()
```

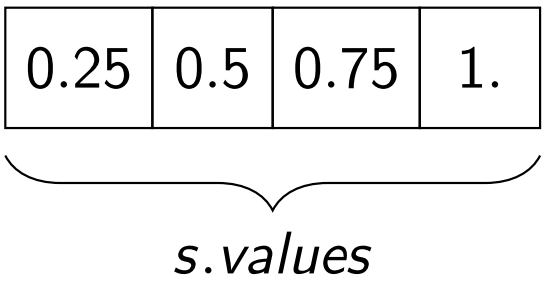
```
False    560658
True      88802
Name: contbr_employer, dtype: int64
```

DataFrames - Series

```
s=pd.Series([.25,.5,.75,1])
```

```
0   0.25
1   0.50
2   0.75
3   1.00
```

```
print(s.values)
```




```
print(s.index)
```

RangeIndex(start=0, stop=4, step=1)

DataFrames - Series (Continued)

```
s[0:2]
```

```
0   0.25
1   0.50
```

```
s[1:3]
```

```
1   0.50
2   0.75
```

```
s1=pd.Series([.25,.5,.75,1],index=['a','b','c','d'])
s2=pd.Series([.5,.75,1,1.25],index=['a','b','c','d'])
df = pd.DataFrame({'s1':s1,'s2':s2})
print(df)
```

```
     s1    s2
a  0.25  0.50
b  0.50  0.75
c  0.75  1.00
d  1.00  1.25
```

DataFrames - Series (Continued)

```
s1=pd.Series([.25,.5,.75,1],index=['a','b','c','d'])
s3=pd.Series([.5,.75,1,1.25],index=['b','c','d','e'])
df = pd.DataFrame({'s1':s1,'s2':s2})
print(df)
```

	s1	s2
a	0.25	0.50
b	0.50	0.75
c	0.75	1.00
d	1.00	1.25

```
s3=pd.Series([.5,.75,1,1.25],index=['b','c','d','e'])
df2 = pd.DataFrame({'s1':s1,'s3':s3})
```

	s1	s3
a	0.25	NaN
b	0.50	0.50
c	0.75	0.75
d	1.00	1.00
e	NaN	1.25

DataFrames - Selecting Rows

```
price = pd.Series({'cherry':2,'berry':1,'orange':3,'apple':4})
qty = pd.Series({'cherry':12,'berry':7,'orange':8,'apple':31})
fruit = pd.DataFrame({'price':price,'qty':qty})
print(fruit)
```

	price	qty
cherry	2	12
berry	1	7
orange	3	8
apple	4	31

```
fruit[1:3]
```

	price	qty
berry	1	7
orange	3	8

DataFrames - Selecting Rows (Continued)

	price	qty
cherry	2	12
berry	1	7
orange	3	8
apple	4	31

```
1 fruit[1:3,0]
```

It will give an error. You cannot select the column directly in the brackets.

```
1 fruit[1:3]['price']
```

berry	1
orange	3

DataFrames - Selecting Rows (iloc method)

iloc allows you more flexibility when you are selecting rows by their positions.

```
1 fruit.iloc[1:3]
```

	price	qty
berry	1	7
orange	3	8

```
1 fruit.iloc[1:3]['price']
```

berry	1
orange	3

```
1 fruit.iloc[1:3,0]
```

berry	1
orange	3

DataFrames - Selecting Rows (loc method)

loc allows you to search rows by their indexes.

```
1 fruit.loc['cherry':'orange']
```

	price	qty
cherry	2	12
berry	1	7
orange	3	8

```
1 fruit.loc['cherry':'orange','price']
```

cherry	2
berry	1
orange	3

Lab 1 - Analyze the NY Fund-Raising Data

You are a data-scientist hired by a political candidate to analyze any possible trends of NY donors. The following questions the campaign wants to know.

- Whether its possible to identify the 'Party' for each candidate (data wrangling)
- Convert the contb_receipt_dt column into an actual date object (data wrangling)
- Using group by, show the number (count) of donations given to each party
- Using group by, show the number of donations given to each party, over time
- Using group by, show the total dollar amount of donations given to each party
- Using group by, show the total dollar amount of donations given to each party, over time
- Which occupations donated the top 5 most money?

Lab 1 - Analyze the NY Fund-Raising Data (Continued)

- Which occupations donated the least 5 amount of money?
- Which employer's employees gave the most money, give the top 5.
- For each candidate, what were the top 5 occupations that donated to their election
- For the 5 candidates that raised the most money, graph their donations by time, in a line graph