In [19]:
```
import pandas as pd
import matplotlib.pyplot as plt
# plots are displayed inline
%matplotlib inline

recent_grads = pd.read_csv("recent-grads.csv")
```

In [20]:
```
# Default: first 5 rows
recent_grads.head(5)
```

Out[20]:

	Rank	Major_code	Major	Total	Men	Women	Major_category	ShareWomen
0	1	2419	PETROLEUM ENGINEERING	2339.0	2057.0	282.0	Engineering	0.120564
1	2	2416	MINING AND MINERAL ENGINEERING	756.0	679.0	77.0	Engineering	0.101852
2	3	2415	METALLURGICAL ENGINEERING	856.0	725.0	131.0	Engineering	0.153037
3	4	2417	NAVAL ARCHITECTURE AND MARINE ENGINEERING	1258.0	1123.0	135.0	Engineering	0.107313
4	5	2405	CHEMICAL ENGINEERING	32260.0	21239.0	11021.0	Engineering	0.341631

5 rows × 21 columns

In [21]:
```
# First row
recent_grads.iloc[0]
```

Out[21]:
```
Rank                                    1
Major_code                           2419
Major                 PETROLEUM ENGINEERING
Total                                2339
Men                                  2057
Women                                 282
Major_category                Engineering
ShareWomen                       0.120564
Sample_size                            36
Employed                             1976
Full_time                            1849
Part_time                             270
Full_time_year_round                 1207
Unemployed                             37
Unemployment_rate                0.0183805
Median                              110000
P25th                                95000
P75th                               125000
College_jobs                         1534
Non_college_jobs                      364
Low_wage_jobs                         193
Name: 0, dtype: object
```

In [22]:
```
# last five rows
recent_grads.tail()
```

Out[22]:

	Rank	Major_code	Major	Total	Men	Women	Major_category	ShareWomen	S
168	169	3609	ZOOLOGY	8409.0	3050.0	5359.0	Biology & Life Science	0.637293	
169	170	5201	EDUCATIONAL PSYCHOLOGY	2854.0	522.0	2332.0	Psychology & Social Work	0.817099	
170	171	5202	CLINICAL PSYCHOLOGY	2838.0	568.0	2270.0	Psychology & Social Work	0.799859	
171	172	5203	COUNSELING PSYCHOLOGY	4626.0	931.0	3695.0	Psychology & Social Work	0.798746	
172	173	3501	LIBRARY SCIENCE	1098.0	134.0	964.0	Education	0.877960	

5 rows × 21 columns

In [23]:
```
# Estimate calculations for each columns
recent_grads.describe()
```

Out[23]:

	Rank	Major_code	Total	Men	Women	ShareWomen	Sam
count	173.000000	173.000000	172.000000	172.000000	172.000000	172.000000	17
mean	87.000000	3879.815029	39370.081395	16723.406977	22646.674419	0.522223	35
std	50.084928	1687.753140	63483.491009	28122.433474	41057.330740	0.231205	61
min	1.000000	1100.000000	124.000000	119.000000	0.000000	0.000000	
25%	44.000000	2403.000000	4549.750000	2177.500000	1778.250000	0.336026	3
50%	87.000000	3608.000000	15104.000000	5434.000000	8386.500000	0.534024	13
75%	130.000000	5503.000000	38909.750000	14631.000000	22553.750000	0.703299	33
max	173.000000	6403.000000	393735.000000	173809.000000	307087.000000	0.968954	421

In [24]:
```
# Counts the number of rows
raw_data_count = len(recent_grads)
raw_data_count
```

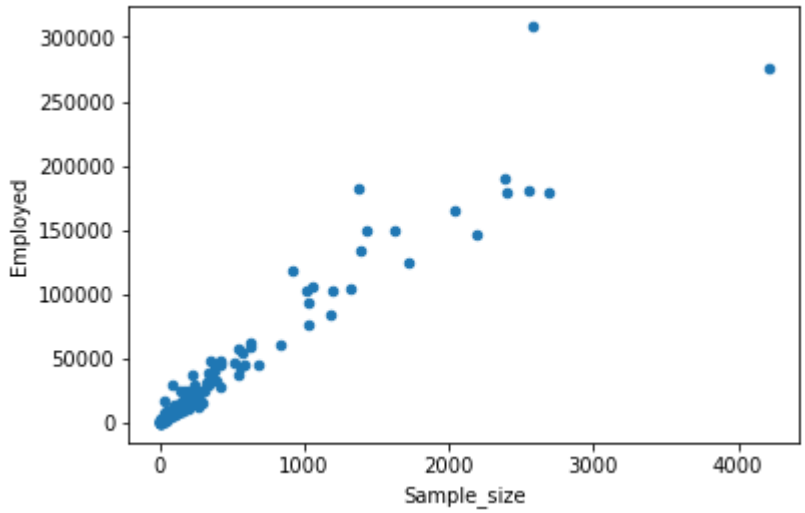
Out[24]: 173

In [25]:
```
# Dropping missing rows
recent_grads = recent_grads.dropna()
cleaned_data_count = len(recent_grads)
cleaned_data_count
```

Out[25]: 172

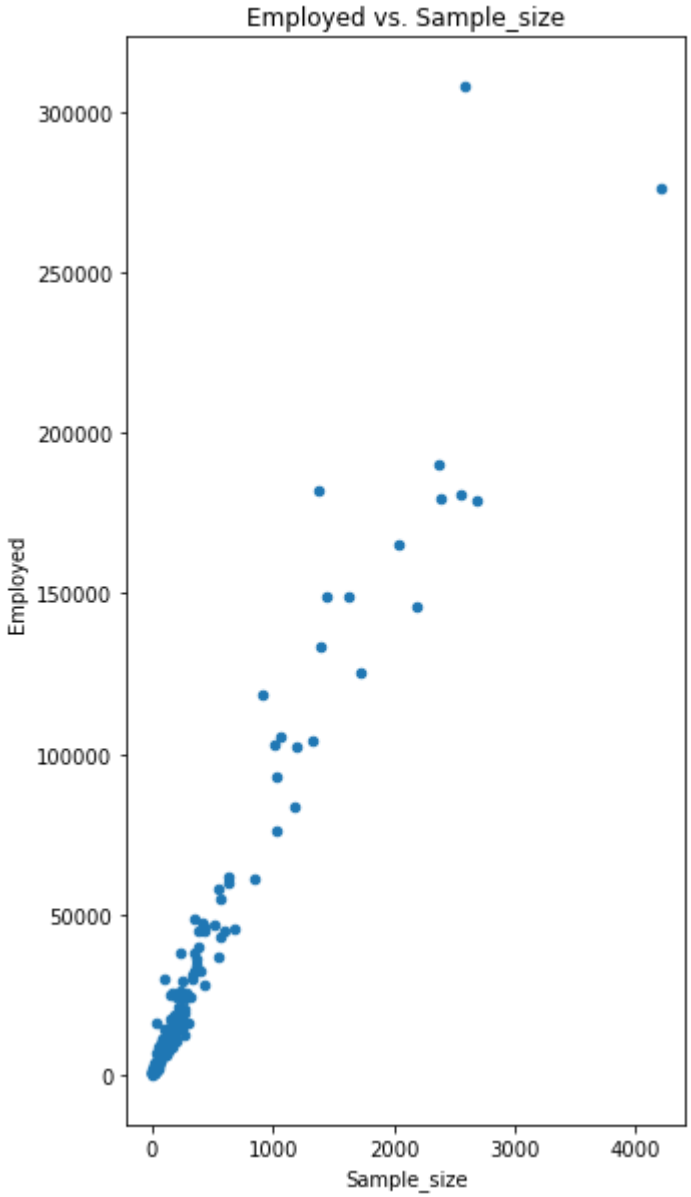
In [26]:
```
recent_grads.plot(x='Sample_size', y='Employed', kind='scatter')
```

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0096a59f98>



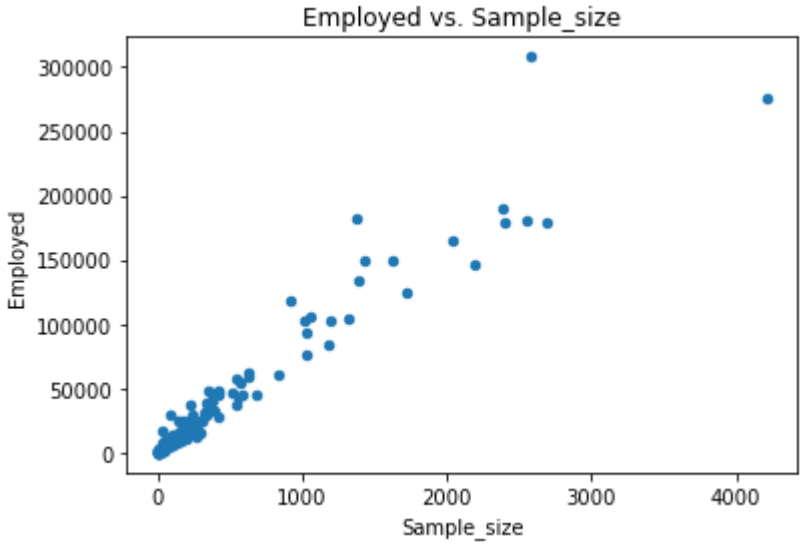
In [27]:
```
recent_grads.plot(x='Sample_size', y='Employed', kind='scatter', title='
```

Out[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f00969f9780>
```



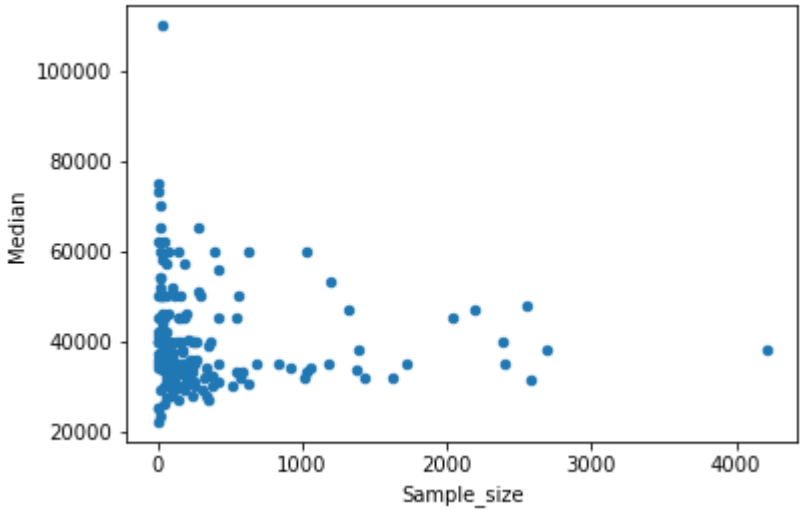
In [28]:
```
ax = recent_grads.plot(x='Sample_size', y='Employed', kind='scatter')
ax.set_title('Employed vs. Sample_size')
```

Out[28]: <matplotlib.text.Text at 0x7f00968e27b8>



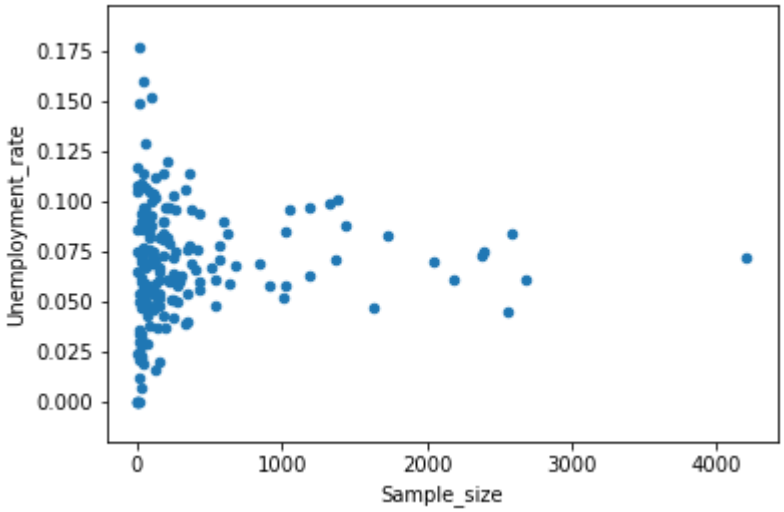
In [29]:
```
recent_grads.plot(x='Sample_size', y='Median', kind='scatter')
```

Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0096a08320>



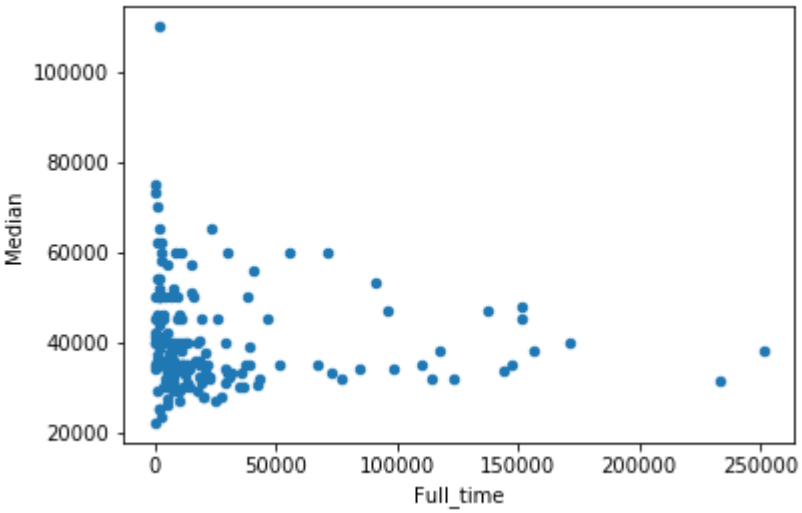
In [30]:
```
recent_grads.plot(x='Sample_size', y='Unemployment_rate', kind='scatter'
```

Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0x7f00968b6518>



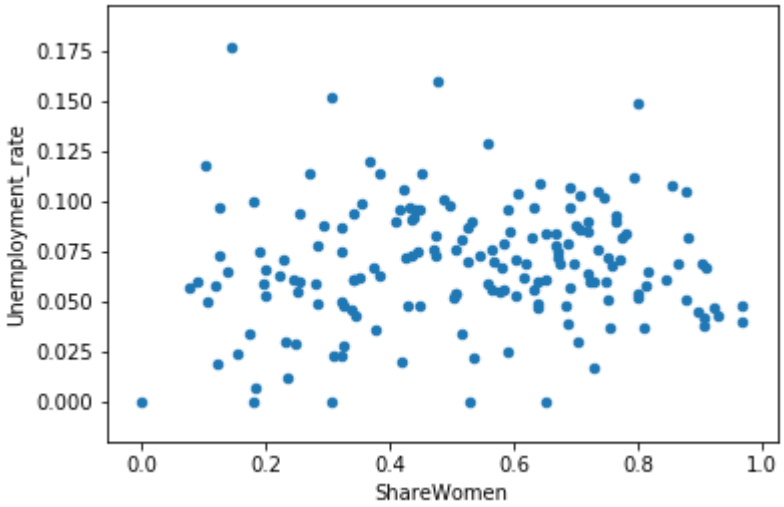
In [31]:
```
recent_grads.plot(x='Full_time', y='Median', kind='scatter')
```

Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x7f009682b9b0>



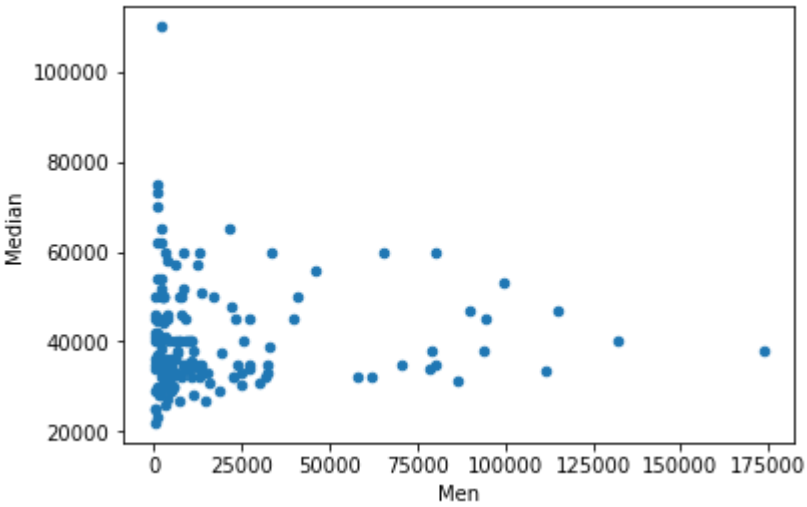
In [32]:
```
recent_grads.plot(x='ShareWomen', y='Unemployment_rate', kind='scatter')
```

Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0096d8d7b8>



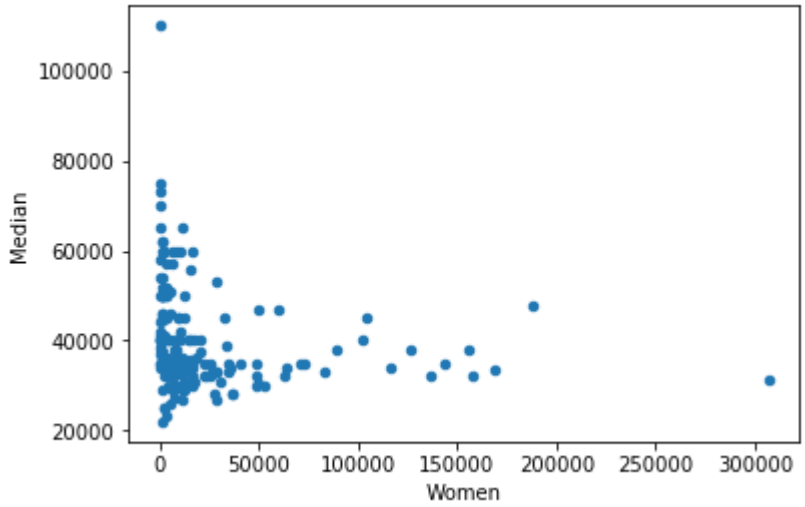
In [33]:
```
recent_grads.plot(x='Men', y='Median', kind='scatter')
```

Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x7f00966de3c8>



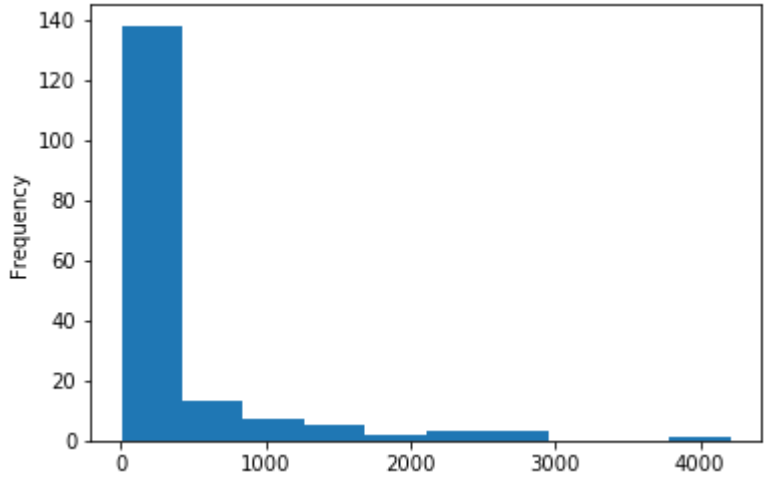
In [34]:
```
recent_grads.plot(x='Women', y='Median', kind='scatter')
```

Out[34]: <matplotlib.axes._subplots.AxesSubplot at 0x7f00967e3b00>



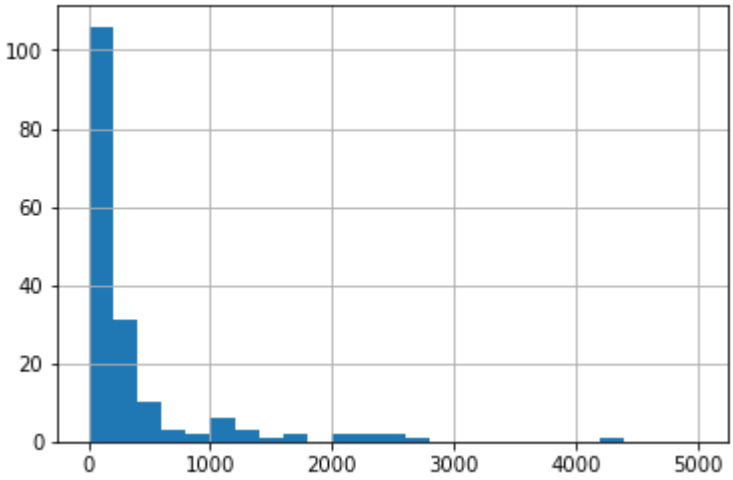
In [37]:
```
# Histograms
recent_grads['Sample_size'].plot(kind='hist')
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0096507be0>



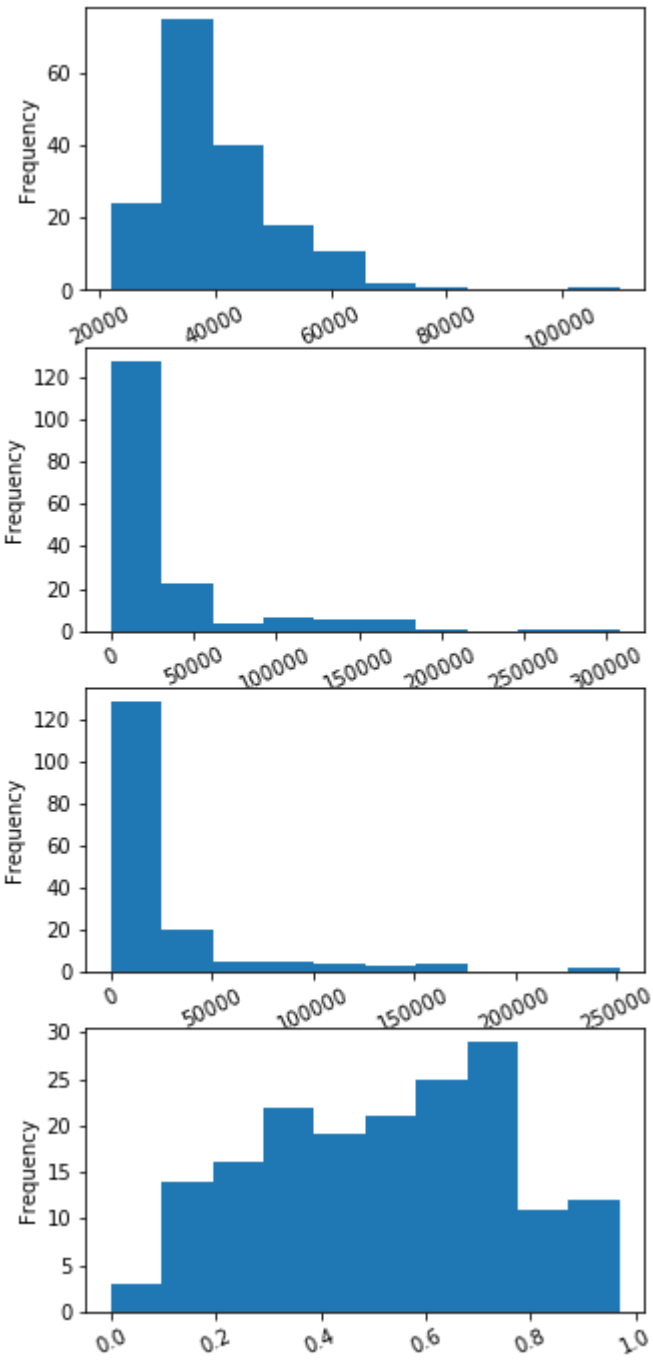
In [38]:
```
recent_grads['Sample_size'].hist(bins=25, range=(0,5000))
```

Out[38]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f00963464e0>
```

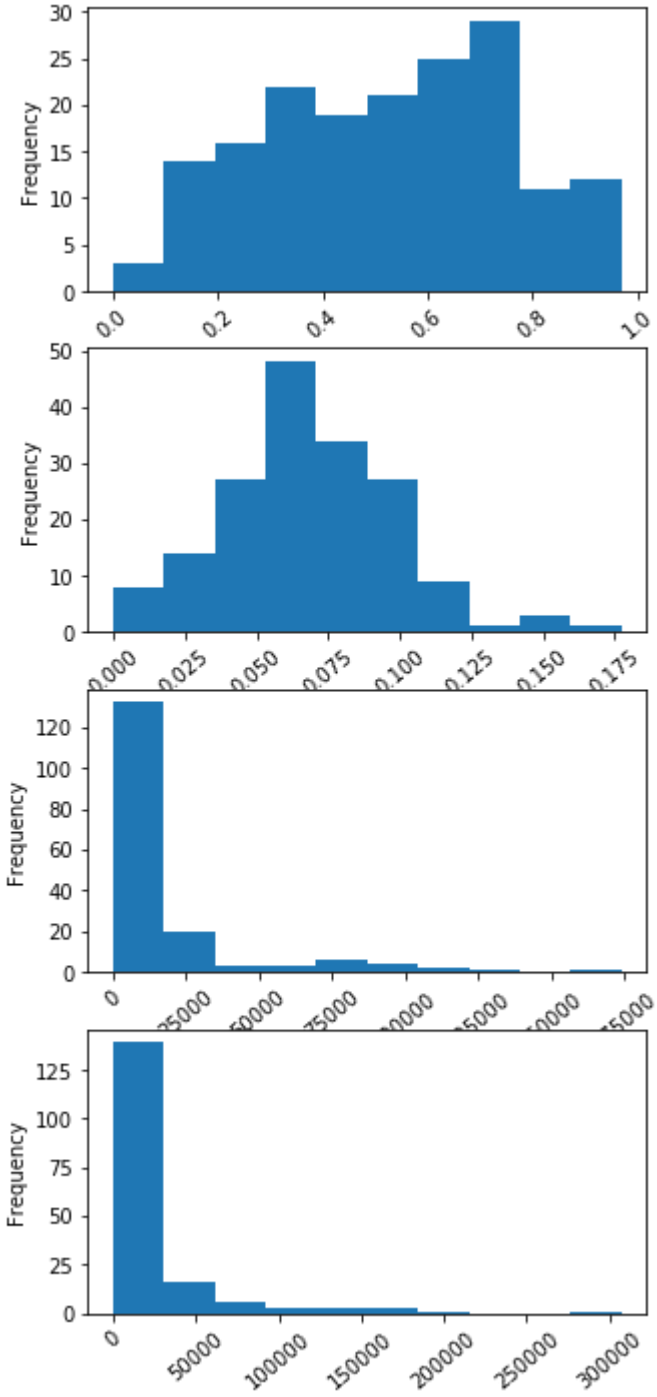


In [43]:
```
cols = ["Sample_size", "Median", "Employed", "Full_time", "ShareWomen",

fig = plt.figure(figsize=(5,12))
for r in range(1,5):
    ax = fig.add_subplot(4,1,r)
    ax = recent_grads[cols[r]].plot(kind='hist', rot=25)
```



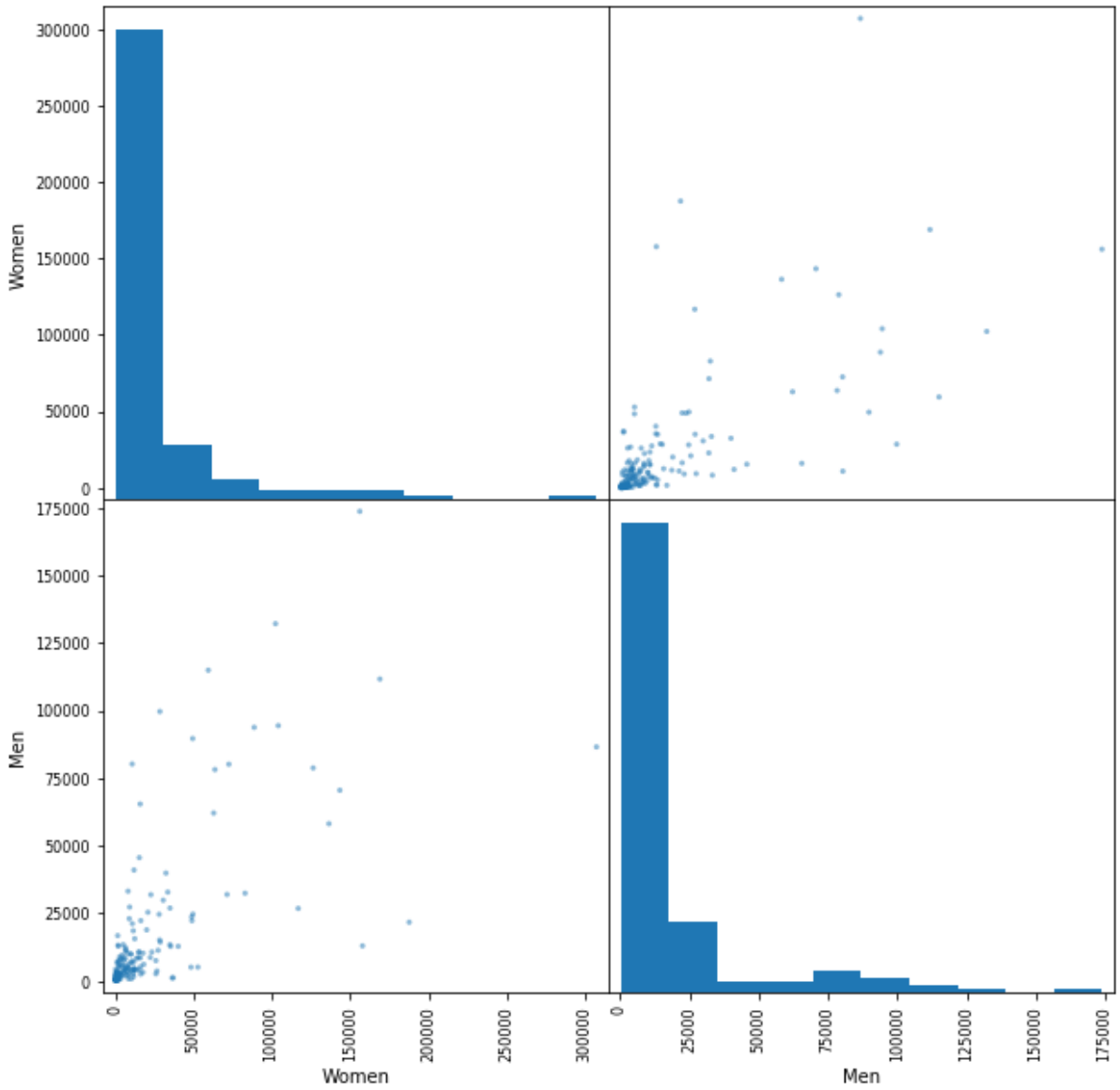
In [46]:
```
fig2 = plt.figure(figsize=(5,12))
for r in range(4,8):
    ax20 = fig2.add_subplot(4,1,r-3)
    ax20 = recent_grads[cols[r]].plot(kind='hist', rot=40)
```



In [49]:
```
# Scatter Matrix Plot
from pandas.tools.plotting import scatter_matrix
scatter_matrix(recent_grads[['Women', 'Men']], figsize=(10,10))
```

```
/home/ernhung92/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:3: FutureWarning: 'pandas.tools.plotting.scatter_matrix' is deprecated, import 'pandas.plotting.scatter_matrix' instead.
  This is separate from the ipykernel package so we can avoid doing imports until
```

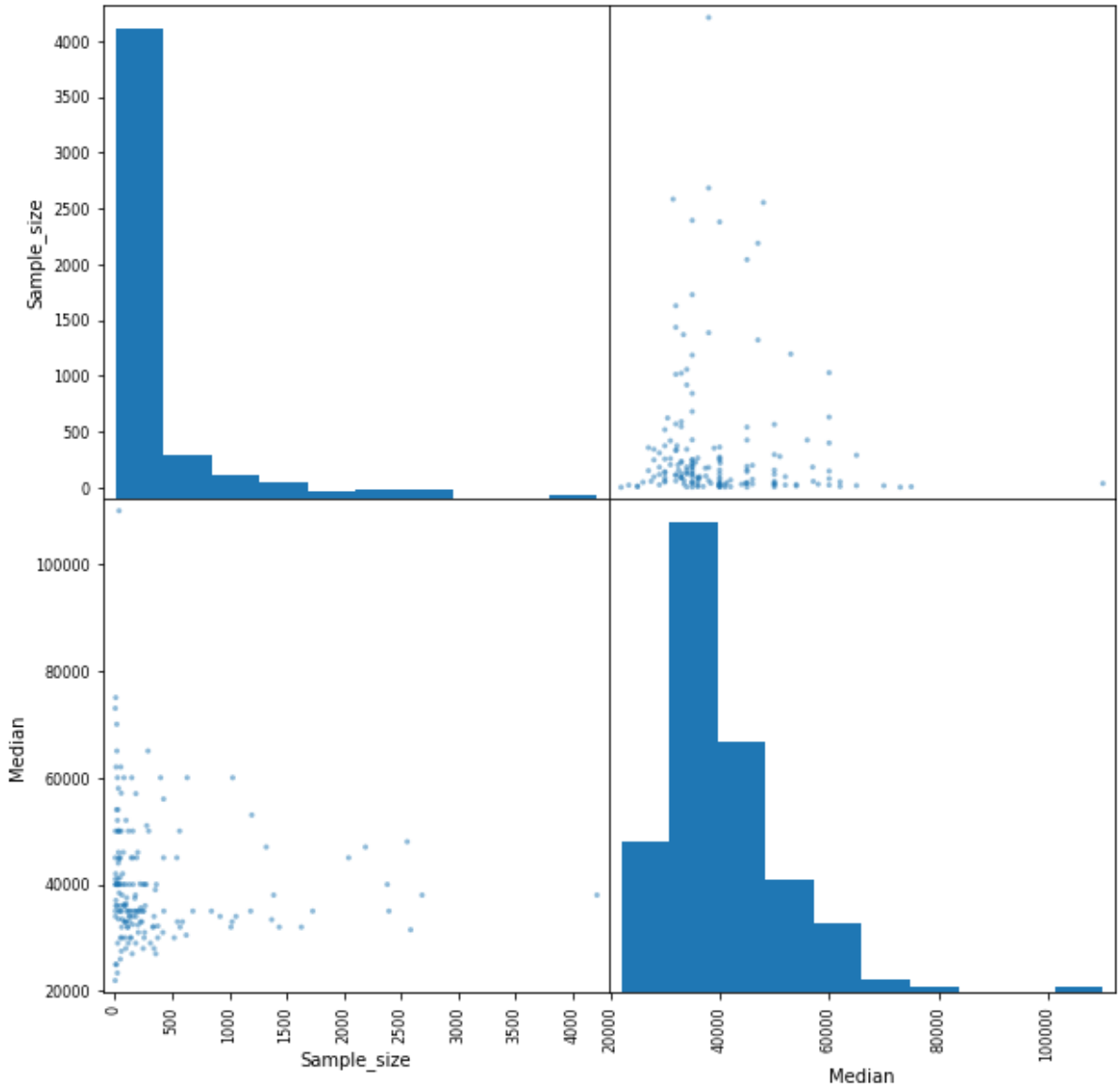
Out[49]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f0095f55080>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0095e7b400>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f0095a5bd30>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0095b2ca90>]], dtype=object)
```



In [51]:
```
scatter_matrix(recent_grads[['Sample_size', 'Median']], figsize=(10,10))
```

```
/home/ernhung92/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1: FutureWarning: 'pandas.tools.plotting.scatter_matrix' is deprecated, import 'pandas.plotting.scatter_matrix' instead.
  """Entry point for launching an IPython kernel.
```

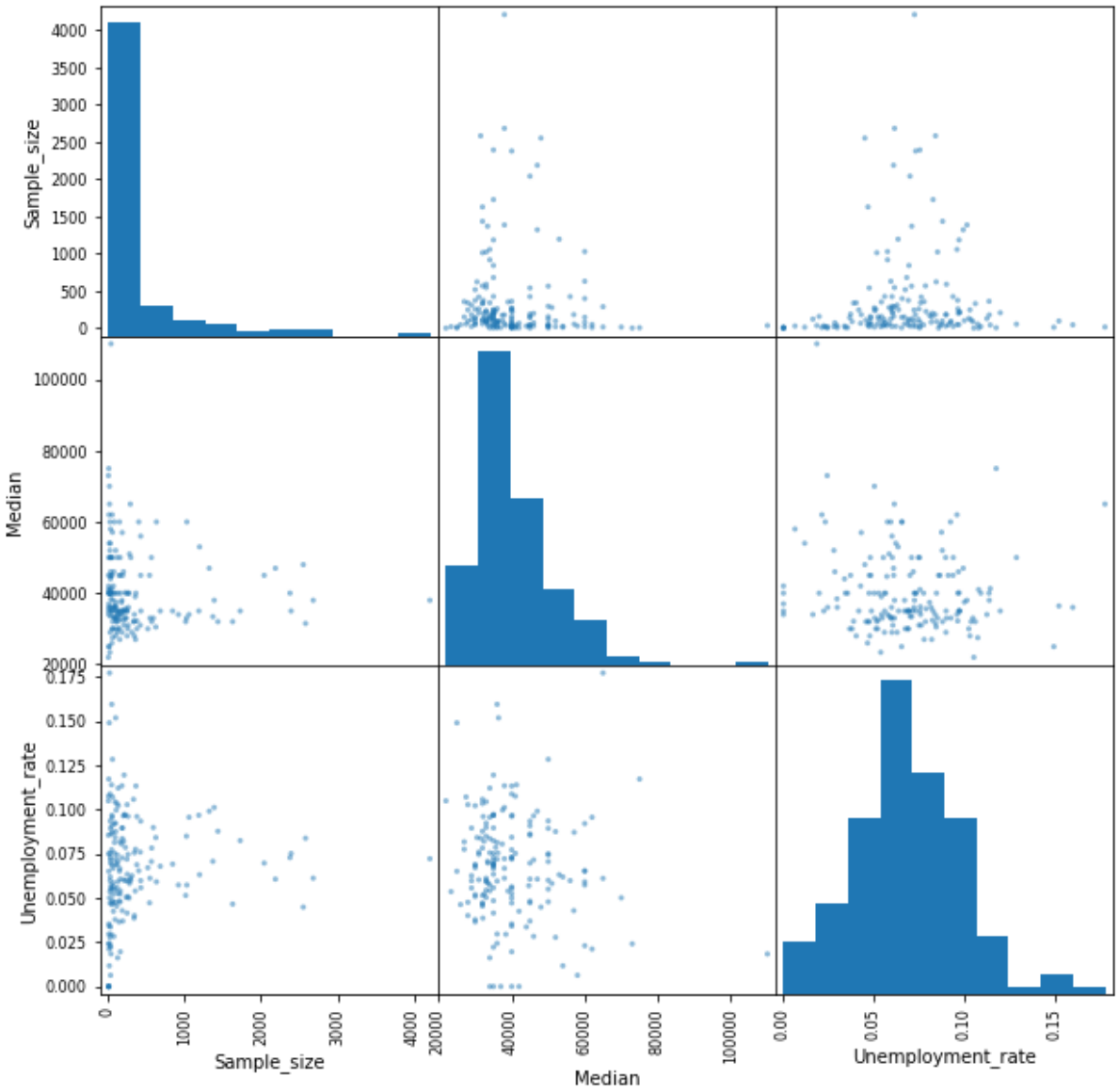
Out[51]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f009544c9e8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f00953f9d68>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f00954248d0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0095190630>]], dtype=object)
```



In [52]:
```
scatter_matrix(recent_grads[['Sample_size', 'Median', 'Unemployment_rate
```

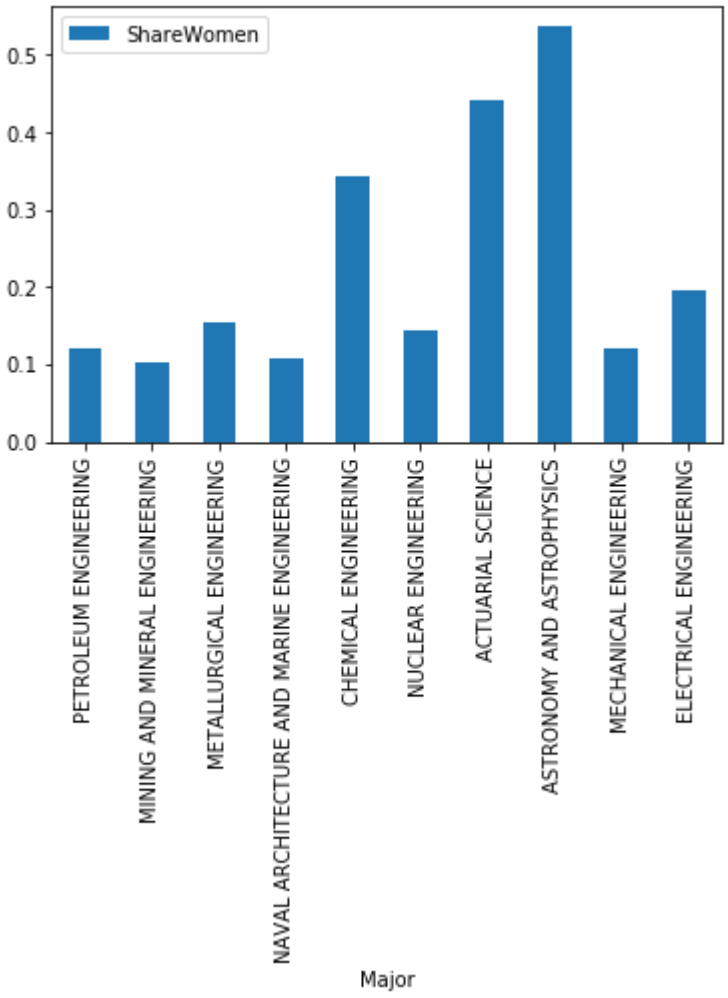
```
/home/ernhung92/anaconda3/lib/python3.6/site-packages/ipykernel_launche
r.py:1: FutureWarning: 'pandas.tools.plotting.scatter_matrix' is deprec
ated, import 'pandas.plotting.scatter_matrix' instead.
  """Entry point for launching an IPython kernel.
```

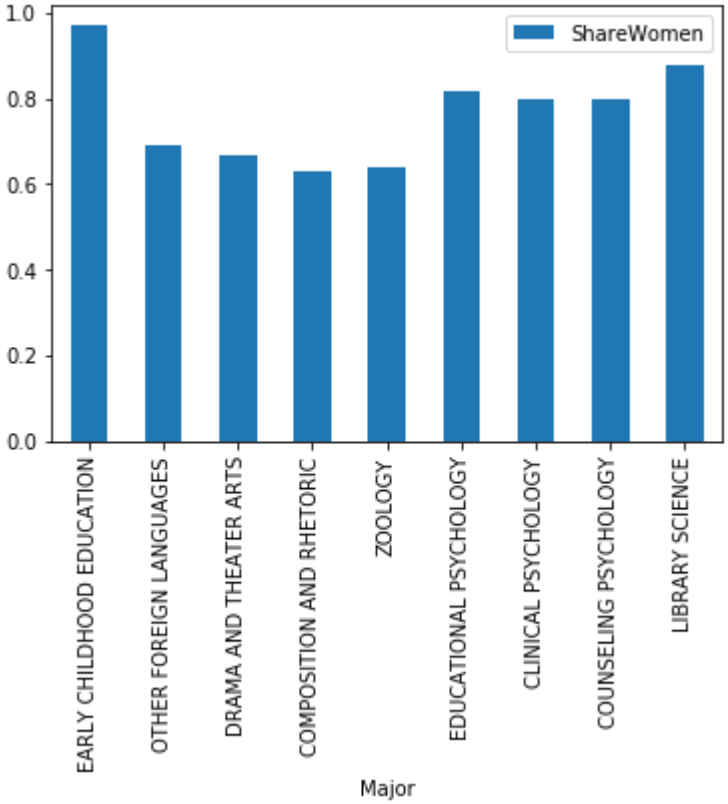
Out[52]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f0094fd63c8
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094f6f2b0
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094f81470
>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f0094ebe3c8
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094e8da90
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094e8dac8
>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f0094d6b4a8
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094ce9160
>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f0094c9a588
>]], dtype=object)
```


Sample_size
Median
Unemployment_rate
0
500
1000
1500
2000
2500
3000
3500
4000
20000
40000
60000
80000
100000
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.00
0.05
0.10
0.15

In [55]:
```
# Bar plots
recent_grads[:10].plot.bar(x='Major', y='ShareWomen', legend=True)
recent_grads[163:].plot.bar(x='Major', y='ShareWomen', legend=True)
```

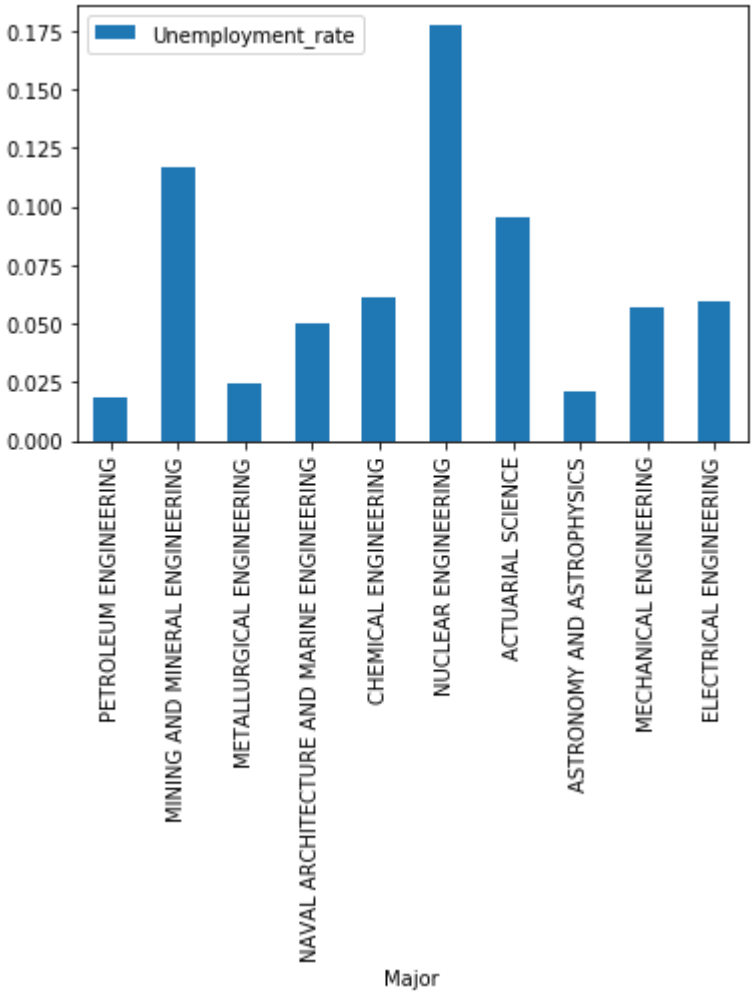
Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x7f008ef1e4a8>

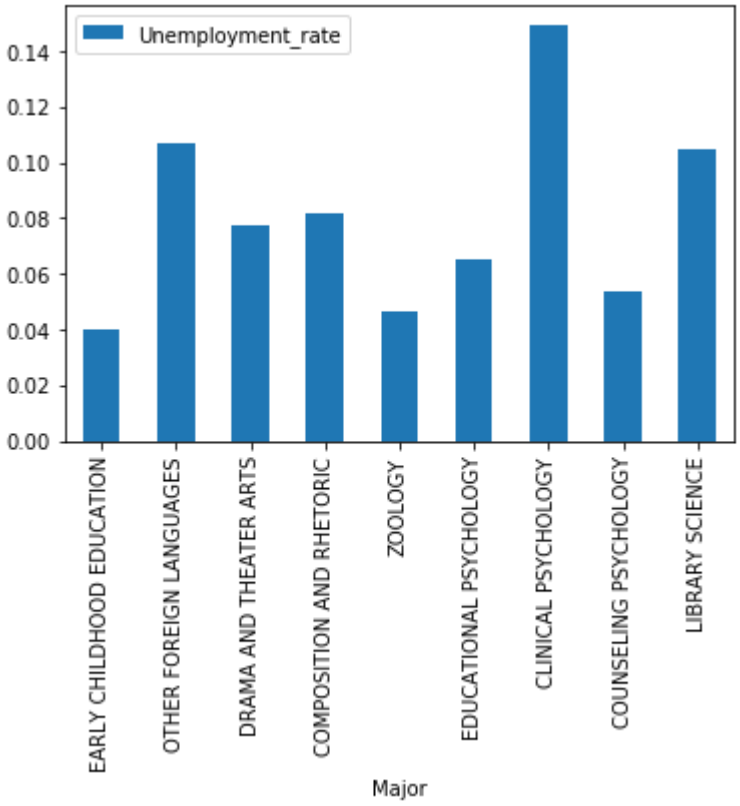



ShareWomen
1.0
0.8
0.6
0.4
0.2
0.0
EARLY CHILDHOOD EDUCATION
OTHER FOREIGN LANGUAGES
DRAMA AND THEATER ARTS
COMPOSITION AND RHETORIC
ZOOLOGY
EDUCATIONAL PSYCHOLOGY
CLINICAL PSYCHOLOGY
COUNSELING PSYCHOLOGY
LIBRARY SCIENCE
Major

In [56]:
```
recent_grads[:10].plot.bar(x='Major', y='Unemployment_rate', legend=True
recent_grads[163:].plot.bar(x='Major', y='Unemployment_rate', legend=Tru
```

Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x7f008ed3c860>




Unemployment_rate
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
EARLY CHILDHOOD EDUCATION
OTHER FOREIGN LANGUAGES
DRAMA AND THEATER ARTS
COMPOSITION AND RHETORIC
ZOOLOGY
EDUCATIONAL PSYCHOLOGY
CLINICAL PSYCHOLOGY
COUNSELING PSYCHOLOGY
LIBRARY SCIENCE
Major

In [58]:
```
import numpy as np

EachGenderMajorCat = recent_grads.pivot_table(index='Major_category', va
EachGenderMajorCat
```

Out[58]:

Major_category	Men	Women
Agriculture & Natural Resources	40357.0	35263.0
Arts	134390.0	222740.0
Biology & Life Science	184919.0	268943.0
Business	667852.0	634524.0
Communications & Journalism	131921.0	260680.0
Computers & Mathematics	208725.0	90283.0
Education	103526.0	455603.0
Engineering	408307.0	129276.0
Health	75517.0	387713.0
Humanities & Liberal Arts	272846.0	440622.0
Industrial Arts & Consumer Services	103781.0	126011.0
Interdisciplinary	2817.0	9479.0
Law & Public Policy	91129.0	87978.0
Physical Sciences	95390.0	90089.0
Psychology & Social Work	98115.0	382892.0
Social Science	256834.0	273132.0

In [64]:
```
EachGenderMajorCat.plot.bar(x=EachGenderMajorCat.index, y=['Men','Women'
```

Out[64]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f008e998e80>
```