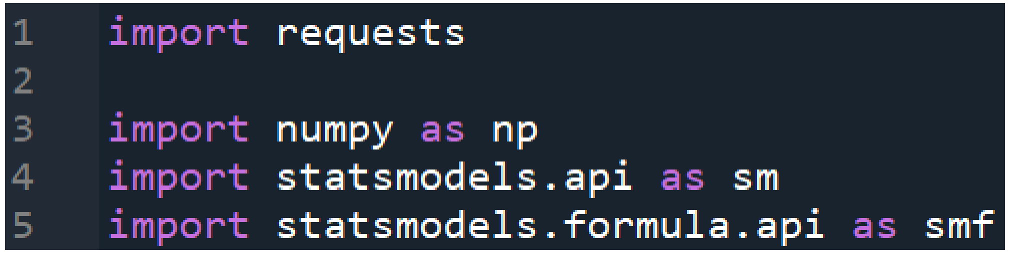
Lecture 17: NumPy and StatsModels

Analysis and model fitting in Python

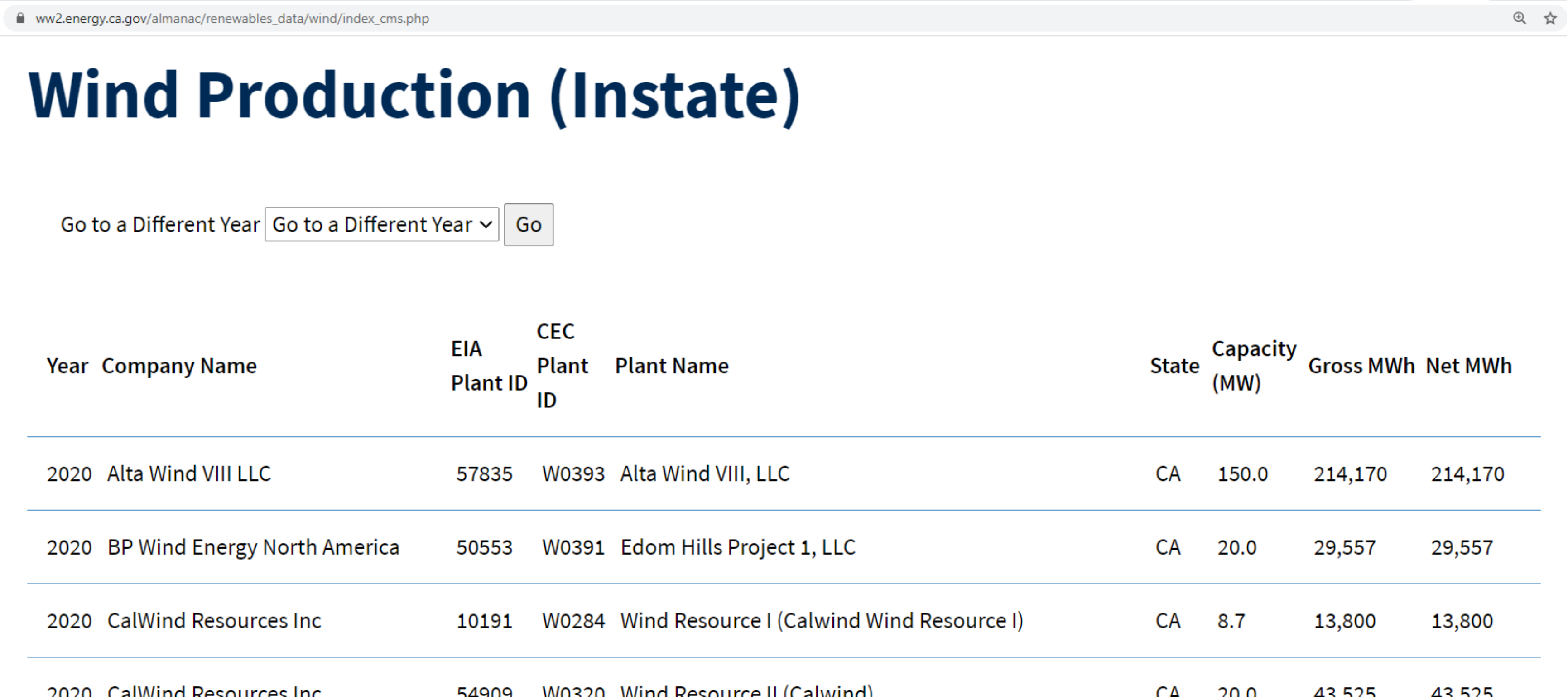
First, the libraries



```python
import requests

import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

Wrap-up from lecture 16: Post Requests

https://ww2.energy.ca.gov/almanac/renewables_data/wind/index_cms.php



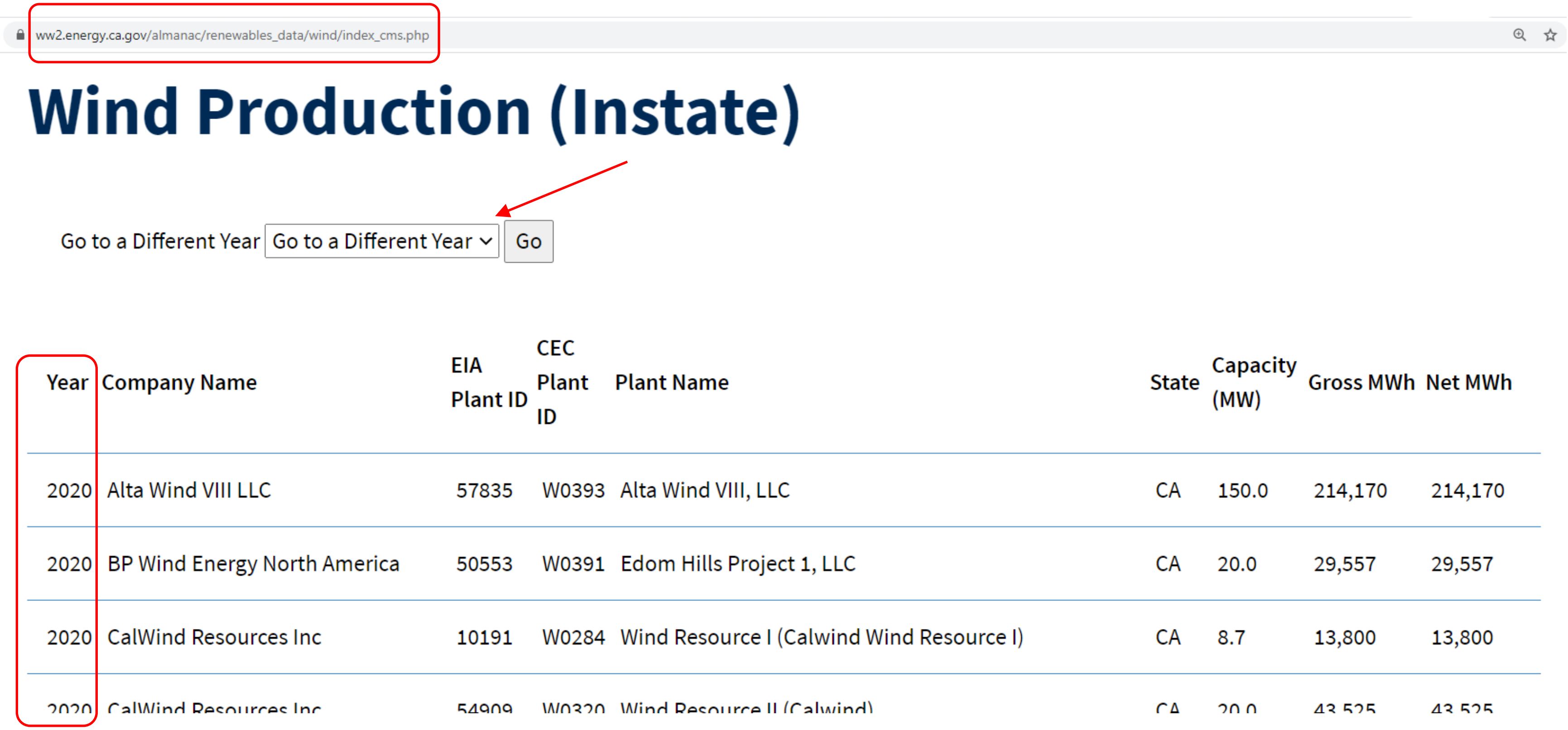
ww2.energy.ca.gov/almanac/renewables_data/wind/index_cms.php

Wind Production (Instate)

Go to a Different Year [Go to a Different Year ⌄] [Go]

Year	Company Name	EIA Plant ID	CEC Plant ID	Plant Name	State	Capacity (MW)	Gross MWh	Net MWh
2020	Alta Wind VIII LLC	57835	W0393	Alta Wind VIII, LLC	CA	150.0	214,170	214,170
2020	BP Wind Energy North America	50553	W0391	Edom Hills Project 1, LLC	CA	20.0	29,557	29,557
2020	CalWind Resources Inc	10191	W0284	Wind Resource I (Calwind Wind Resource I)	CA	8.7	13,800	13,800
2020	CalWind Resources Inc	54909	W0320	Wind Resource II (Calwind)	CA	20.0	43,525	43,525

Wrap-up from lecture 16: Post Requests



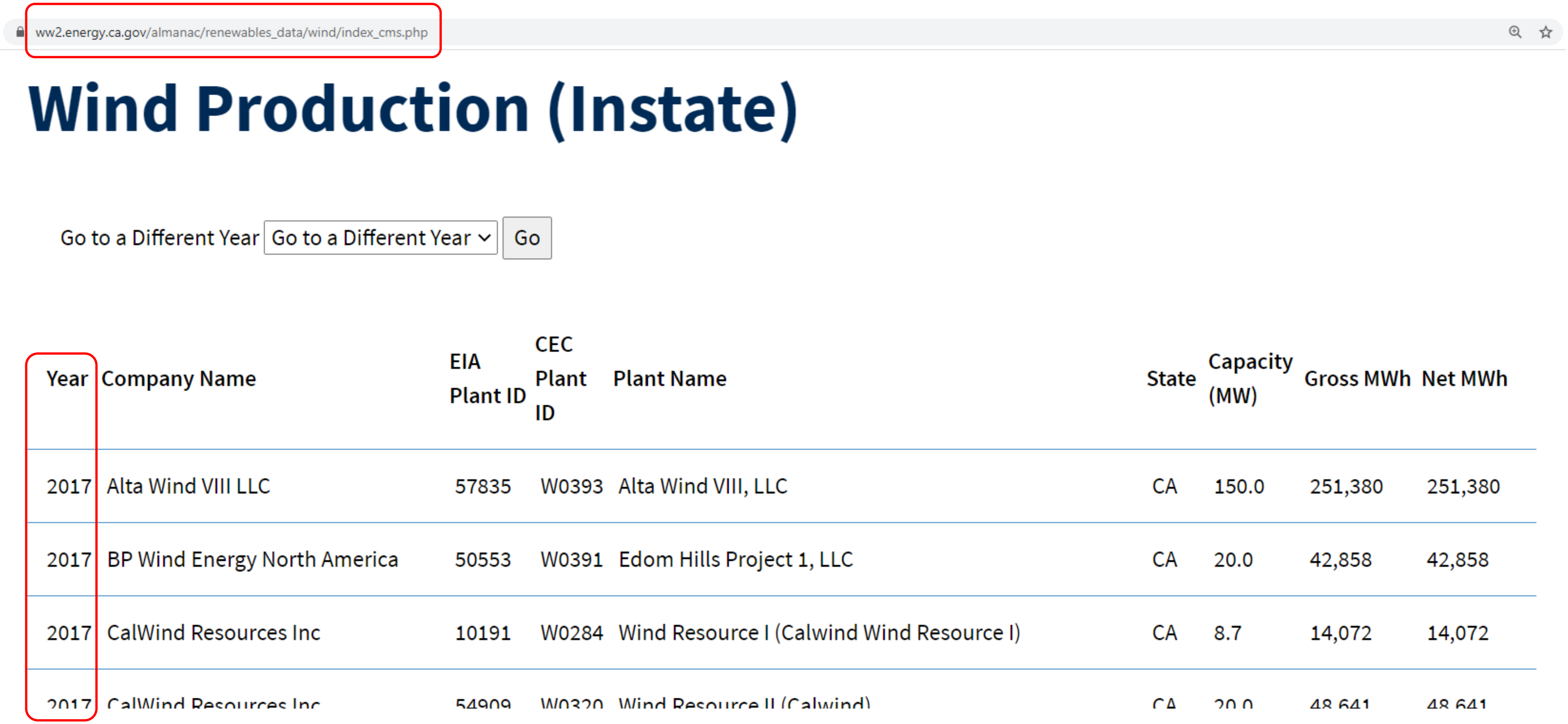
ww2.energy.ca.gov/almanac/renewables_data/wind/index_cms.php

Wind Production (Instate)

Go to a Different Year

Year	Company Name	EIA Plant ID	CEC Plant ID	Plant Name	State	Capacity (MW)	Gross MWh	Net MWh
2020	Alta Wind VIII LLC	57835	W0393	Alta Wind VIII, LLC	CA	150.0	214,170	214,170
2020	BP Wind Energy North America	50553	W0391	Edom Hills Project 1, LLC	CA	20.0	29,557	29,557
2020	CalWind Resources Inc	10191	W0284	Wind Resource I (Calwind Wind Resource I)	CA	8.7	13,800	13,800
2020	CalWind Resources Inc	54909	W0320	Wind Resource II (Calwind)	CA	20.0	43,525	43,525

Wrap-up from lecture 16: Post Requests



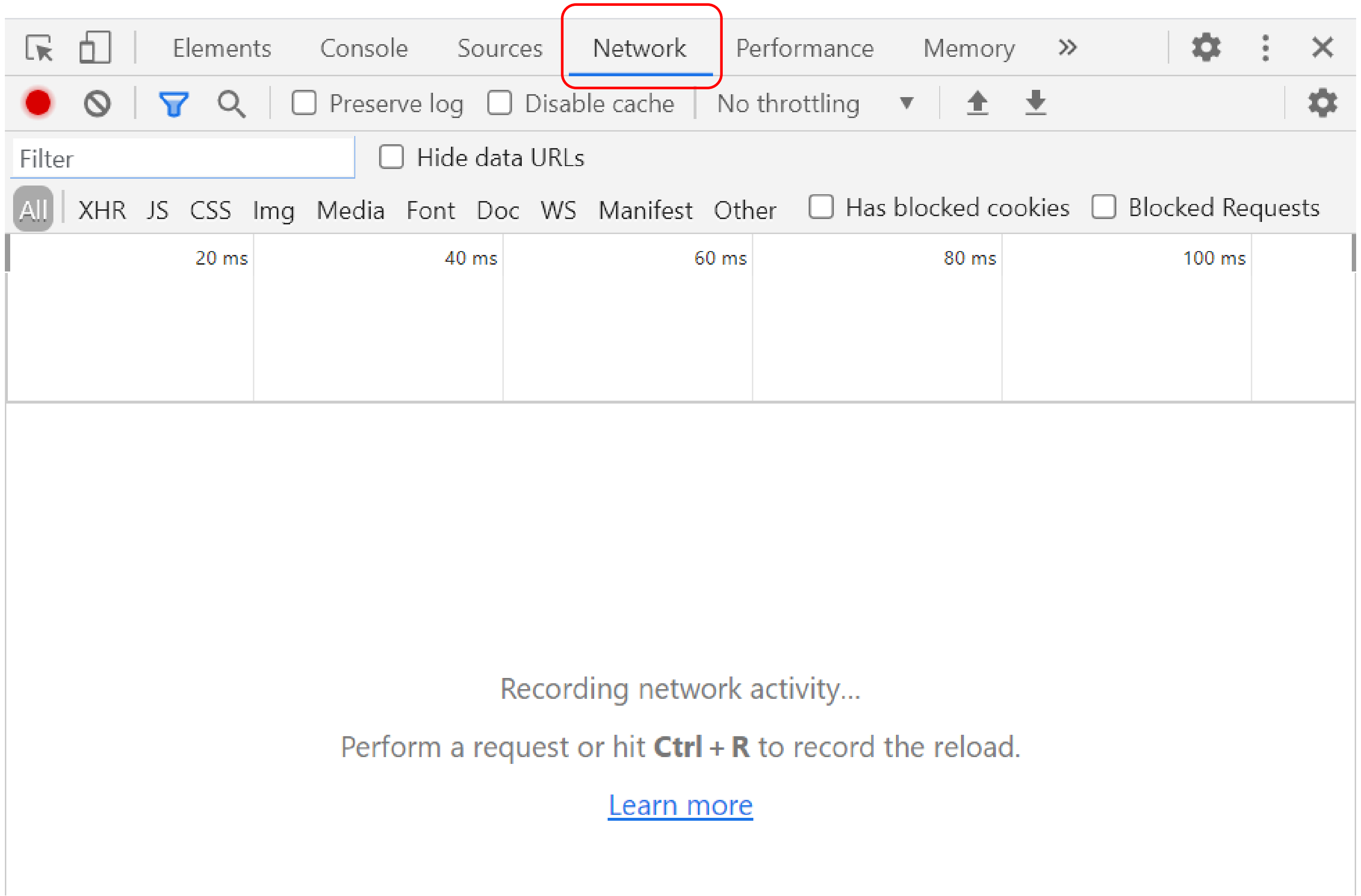
ww2.energy.ca.gov/almanac/renewables_data/wind/index_cms.php

Wind Production (Instate)

Go to a Different Year [Go to a Different Year ▾] [Go]

Year	Company Name	EIA Plant ID	CEC Plant ID	Plant Name	State	Capacity (MW)	Gross MWh	Net MWh
2017	Alta Wind VIII LLC	57835	W0393	Alta Wind VIII, LLC	CA	150.0	251,380	251,380
2017	BP Wind Energy North America	50553	W0391	Edom Hills Project 1, LLC	CA	20.0	42,858	42,858
2017	CalWind Resources Inc	10191	W0284	Wind Resource I (Calwind Wind Resource I)	CA	8.7	14,072	14,072
2017	CalWind Resources Inc	54909	W0320	Wind Resource II (Calwind)	CA	20.0	48,641	48,641

Wrap-up from lecture 16: Post Requests

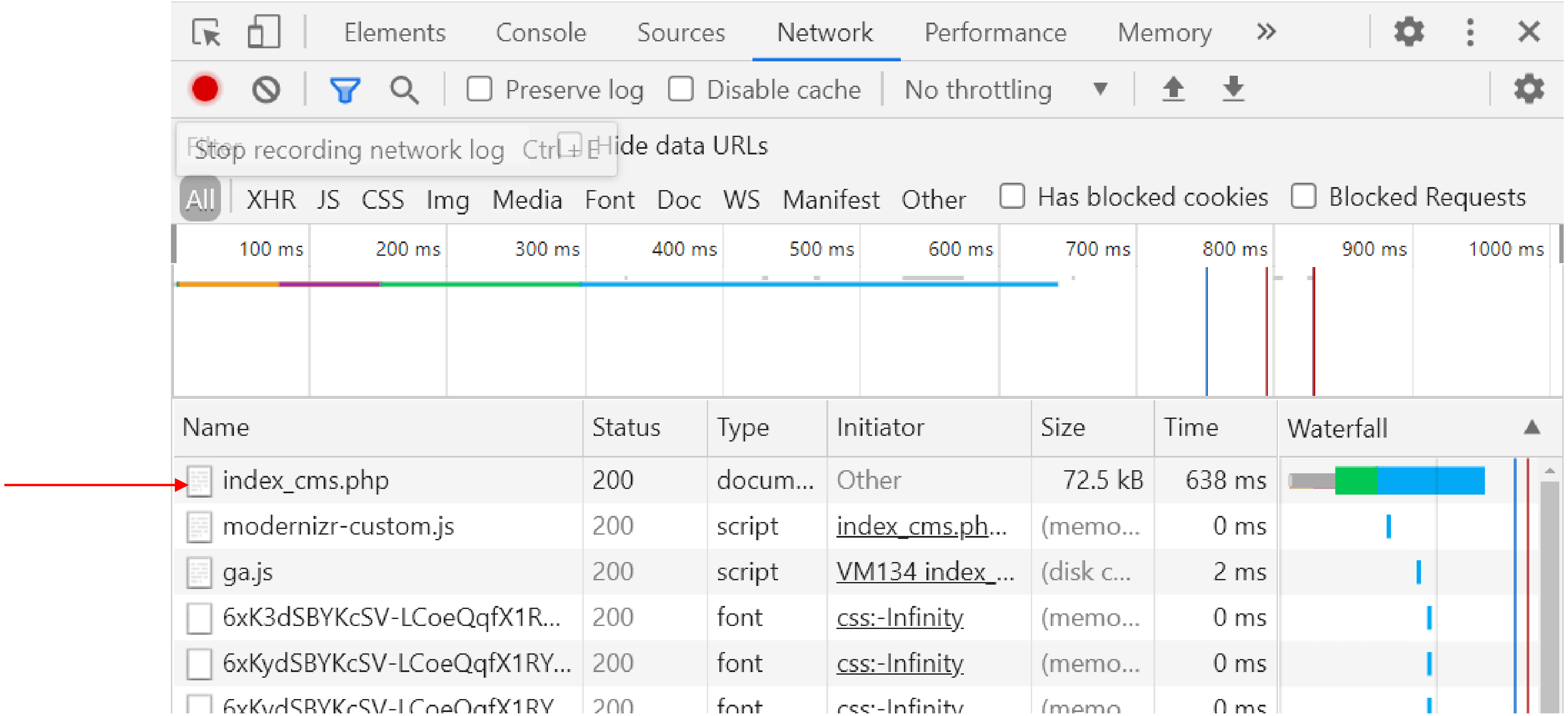


Recording network activity...

Perform a request or hit **Ctrl** + **R** to record the reload.

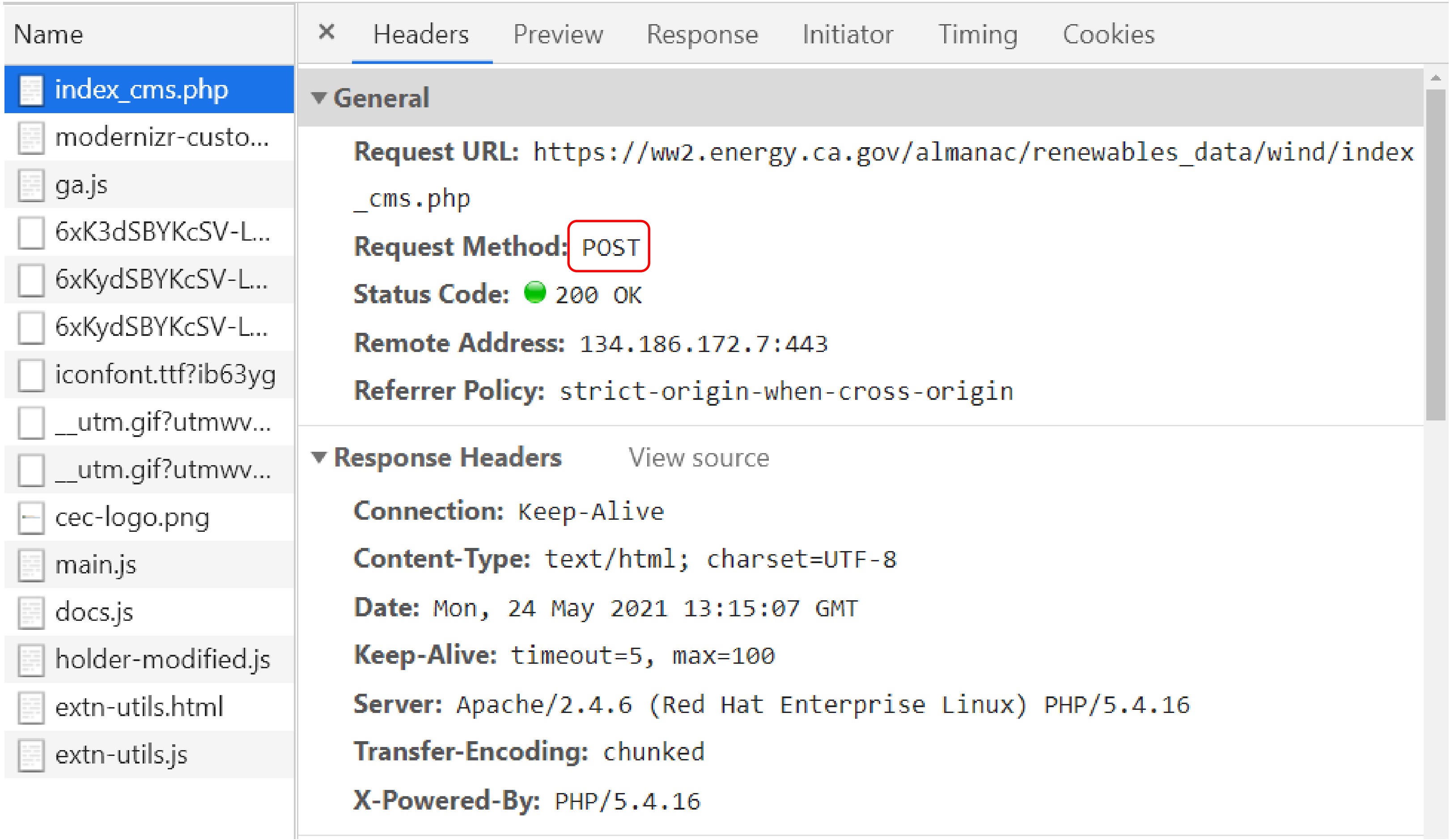
Learn more

Wrap-up from lecture 16: Post Requests



Name	Status	Type	Initiator	Size	Time	Waterfall ▲
index_cms.php	200	docum...	Other	72.5 kB	638 ms	
modernizr-custom.js	200	script	index_cms.ph...	(memo...	0 ms	
ga.js	200	script	VM134 index_...	(disk c...	2 ms	
6xK3dSBYKcSV-LCoeQqfX1R...	200	font	css:-Infinity	(memo...	0 ms	
6xKydSBYKcSV-LCoeQqfX1RY...	200	font	css:-Infinity	(memo...	0 ms	
6xKydSBYKcSV-LCoeQqfX1RY	200	font	css:-Infinity	(memo	0 ms	

Wrap-up from lecture 16: Post Requests



▼ General

Request URL: https://ww2.energy.ca.gov/almanac/renewables_data/wind/index_cms.php

Request Method: POST

Status Code: ● 200 OK

Remote Address: 134.186.172.7:443

Referrer Policy: strict-origin-when-cross-origin

▼ Response Headers View source

Connection: Keep-Alive

Content-Type: text/html; charset=UTF-8

Date: Mon, 24 May 2021 13:15:07 GMT

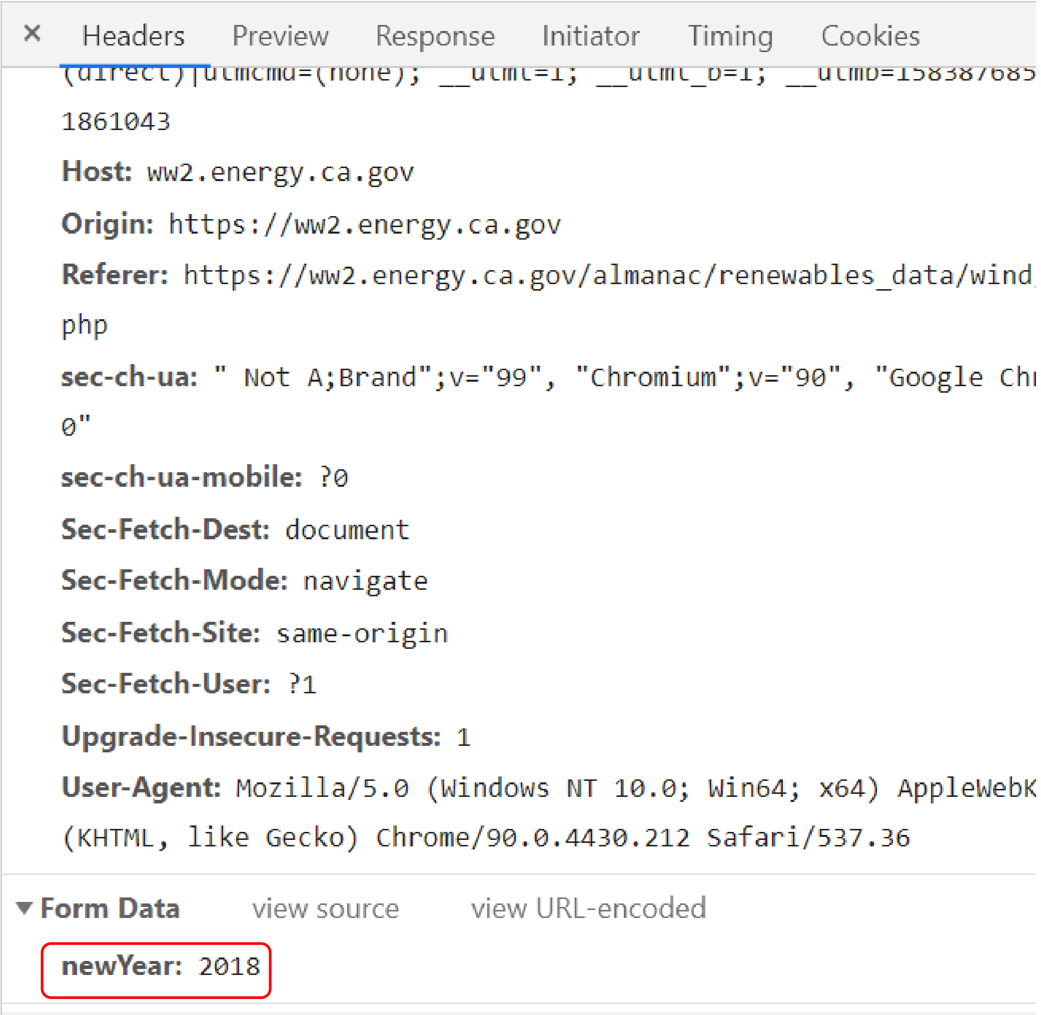
Keep-Alive: timeout=5, max=100

Server: Apache/2.4.6 (Red Hat Enterprise Linux) PHP/5.4.16

Transfer-Encoding: chunked

X-Powered-By: PHP/5.4.16

Wrap-up from lecture 16: Post Requests

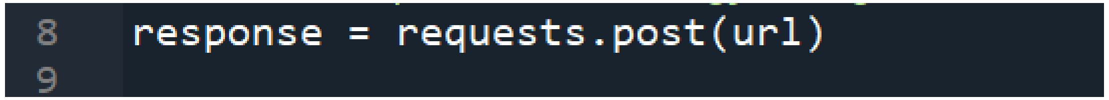


```
(direct)|utmcmd=(none); __utmt=1; __utmt_b=1; __utmb=158387685
1861043
Host: ww2.energy.ca.gov
Origin: https://ww2.energy.ca.gov
Referer: https://ww2.energy.ca.gov/almanac/renewables_data/wind
php
sec-ch-ua: " Not A;Brand";v="99", "Chromium";v="90", "Google Chr
0"
sec-ch-ua-mobile: ?0
Sec-Fetch-Dest: document
Sec-Fetch-Mode: navigate
Sec-Fetch-Site: same-origin
Sec-Fetch-User: ?1
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebK
(KHTML, like Gecko) Chrome/90.0.4430.212 Safari/537.36
```

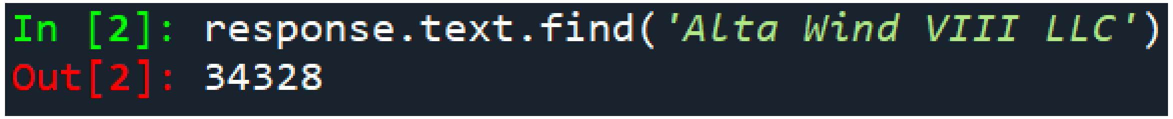
▼ Form Data view source view URL-encoded

newYear: 2018

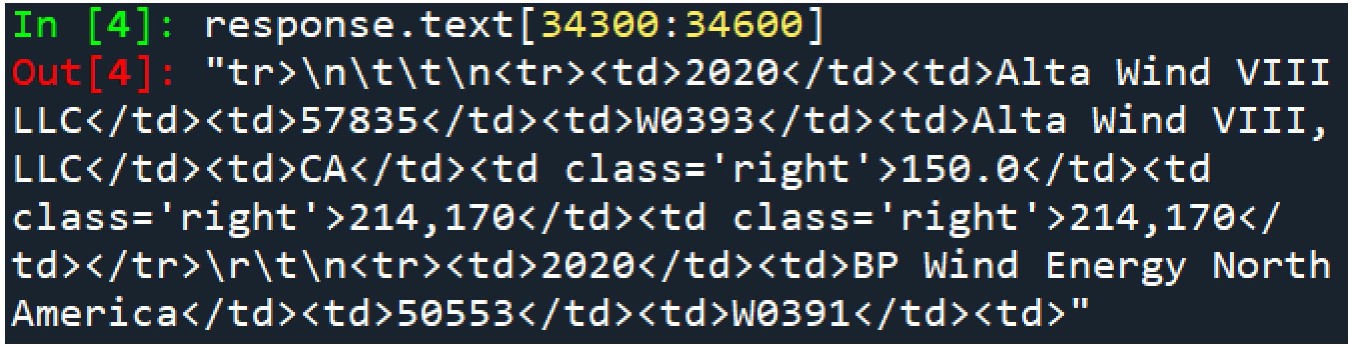
Wrap-up from lecture 16: Post Requests



```
response = requests.post(url)

```

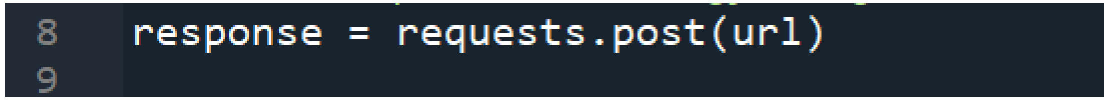


```
In [2]: response.text.find('Alta Wind VIII LLC')
Out[2]: 34328
```

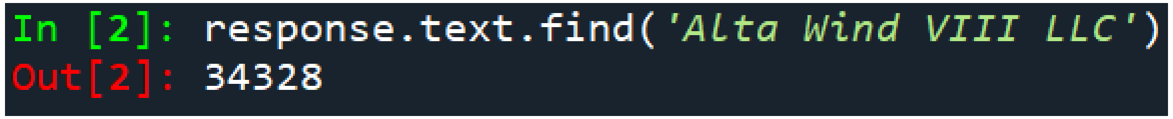


```
In [4]: response.text[34300:34600]
Out[4]: "tr>\n\t\t\n<tr><td>2020</td><td>Alta Wind VIII
LLC</td><td>57835</td><td>W0393</td><td>Alta Wind VIII,
LLC</td><td>CA</td><td class='right'>150.0</td><td
class='right'>214,170</td><td class='right'>214,170</
td></tr>\r\t\n<tr><td>2020</td><td>BP Wind Energy North
America</td><td>50553</td><td>W0391</td><td>"
```

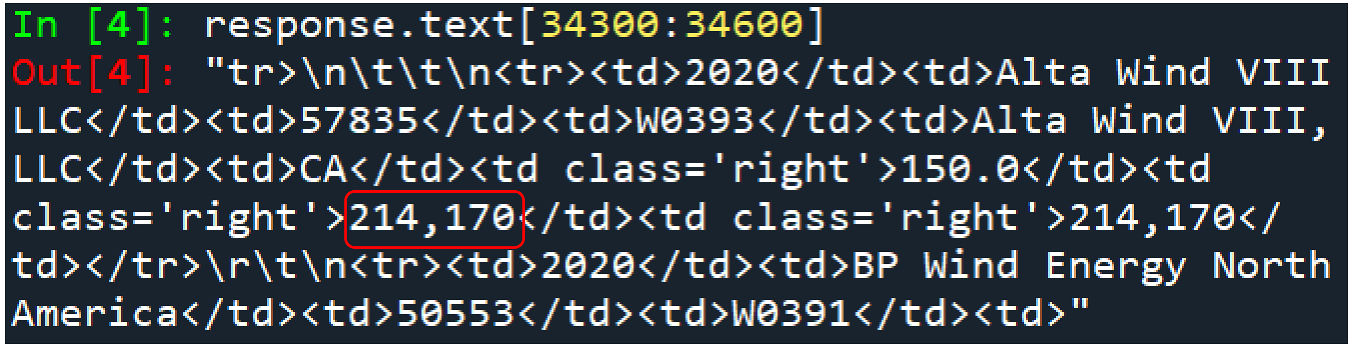
Wrap-up from lecture 16: Post Requests



```
response = requests.post(url)
```

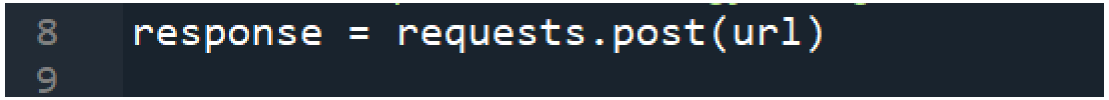


```
In [2]: response.text.find('Alta Wind VIII LLC')
Out[2]: 34328
```

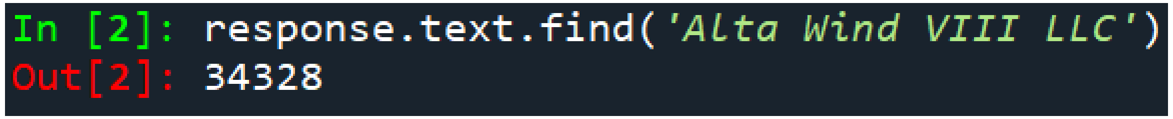


```
In [4]: response.text[34300:34600]
Out[4]: "tr>\n\t\t\n<tr><td>2020</td><td>Alta Wind VIII
LLC</td><td>57835</td><td>W0393</td><td>Alta Wind VIII,
LLC</td><td>CA</td><td class='right'>150.0</td><td
class='right'>214,170</td><td class='right'>214,170</
td></tr>\r\t\n<tr><td>2020</td><td>BP Wind Energy North
America</td><td>50553</td><td>W0391</td><td>"
```

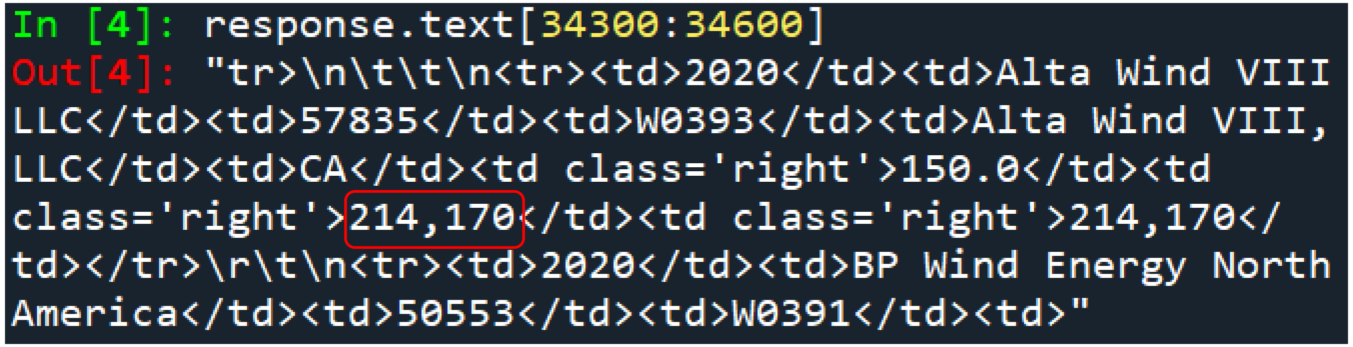
Wrap-up from lecture 16: Post Requests



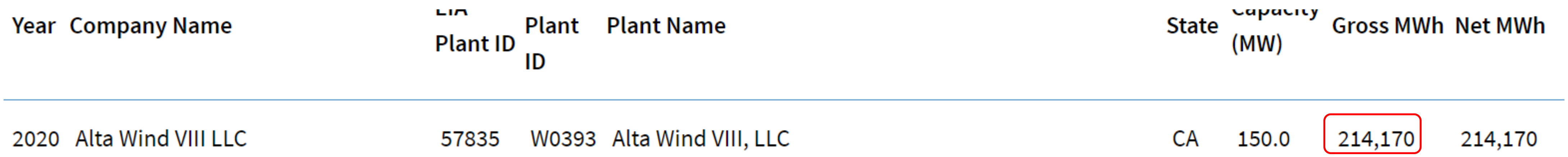
```python
response = requests.post(url)
```



```
In [2]: response.text.find('Alta Wind VIII LLC')
Out[2]: 34328
```

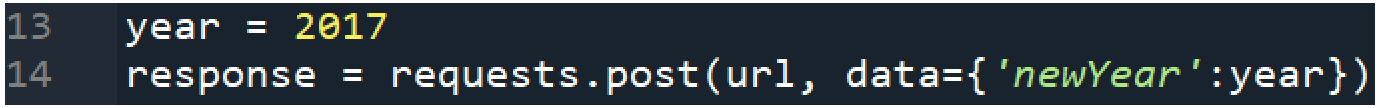


```
In [4]: response.text[34300:34600]
Out[4]: "tr>\n\t\t\n<tr><td>2020</td><td>Alta Wind VIII
LLC</td><td>57835</td><td>W0393</td><td>Alta Wind VIII,
LLC</td><td>CA</td><td class='right'>150.0</td><td
class='right'>214,170</td><td class='right'>214,170</
td></tr>\r\t\n<tr><td>2020</td><td>BP Wind Energy North
America</td><td>50553</td><td>W0391</td><td>"
```

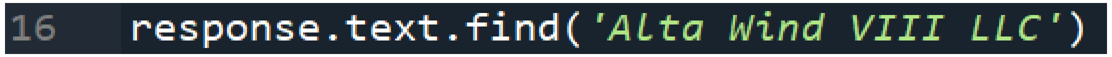


Year	Company Name	EIA Plant ID	Plant ID	Plant Name	State	Capacity (MW)	Gross MWh	Net MWh
2020	Alta Wind VIII LLC	57835	W0393	Alta Wind VIII, LLC	CA	150.0	214,170	214,170

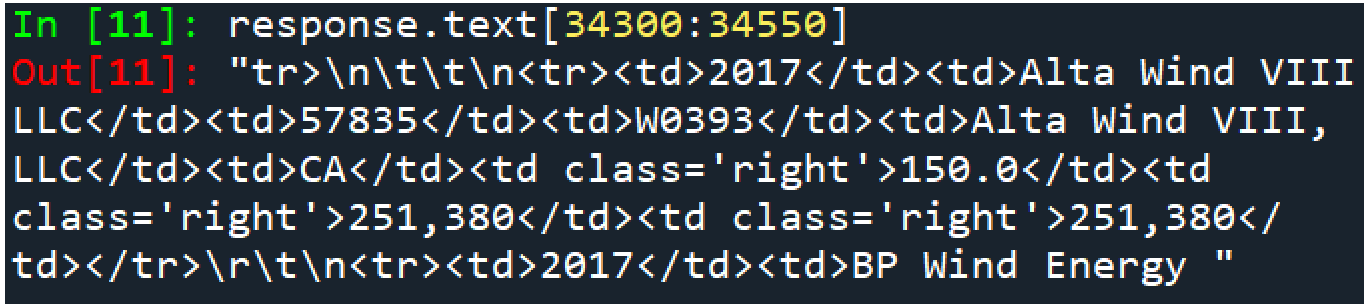
Wrap-up from lecture 16: Post Requests



```python
year = 2017
response = requests.post(url, data={'newYear':year})
```

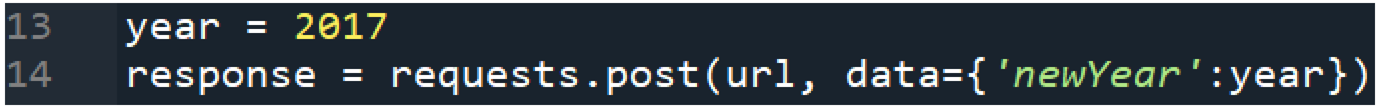


```python
response.text.find('Alta Wind VIII LLC')
```

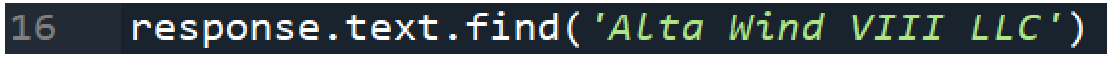


```
In [11]: response.text[34300:34550]
Out[11]: "tr>\n\t\t\n<tr><td>2017</td><td>Alta Wind VIII
LLC</td><td>57835</td><td>W0393</td><td>Alta Wind VIII,
LLC</td><td>CA</td><td class='right'>150.0</td><td
class='right'>251,380</td><td class='right'>251,380</
td></tr>\r\t\n<tr><td>2017</td><td>BP Wind Energy "
```

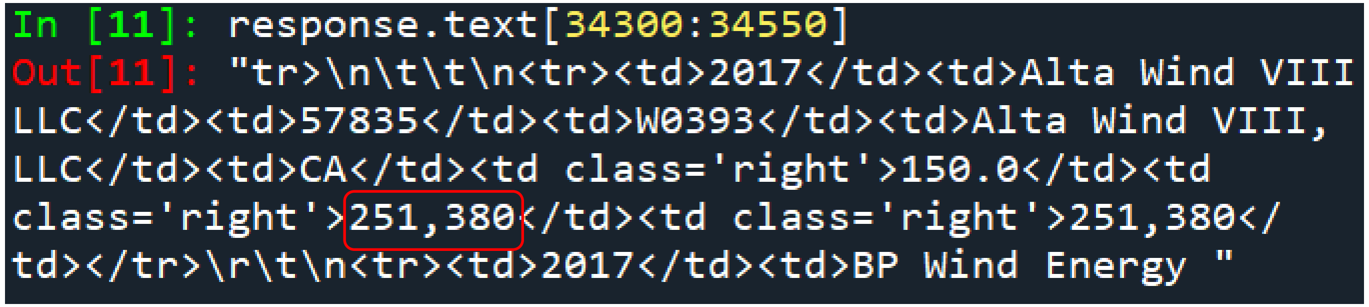
Wrap-up from lecture 16: Post Requests



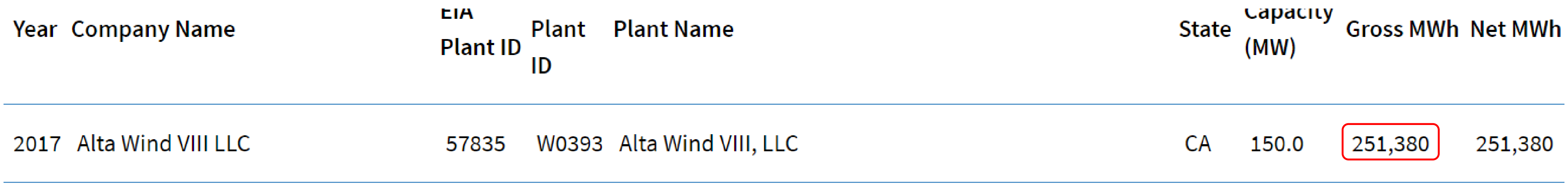
```python
year = 2017
response = requests.post(url, data={'newYear':year})
```



```python
response.text.find('Alta Wind VIII LLC')
```

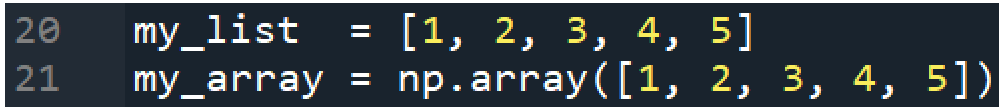


```
In [11]: response.text[34300:34550]
Out[11]: "tr>\n\t\t\n<tr><td>2017</td><td>Alta Wind VIII
LLC</td><td>57835</td><td>W0393</td><td>Alta Wind VIII,
LLC</td><td>CA</td><td class='right'>150.0</td><td
class='right'>251,380</td><td class='right'>251,380</
td></tr>\r\t\n<tr><td>2017</td><td>BP Wind Energy "
```



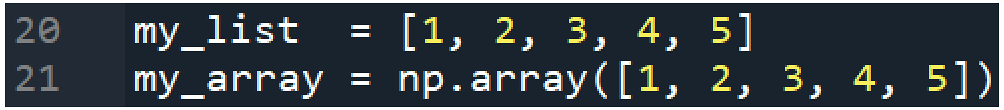
Year	Company Name	EIA Plant ID	Plant ID	Plant Name	State	Capacity (MW)	Gross MWh	Net MWh
2017	Alta Wind VIII LLC	57835	W0393	Alta Wind VIII, LLC	CA	150.0	251,380	251,380

Numpy

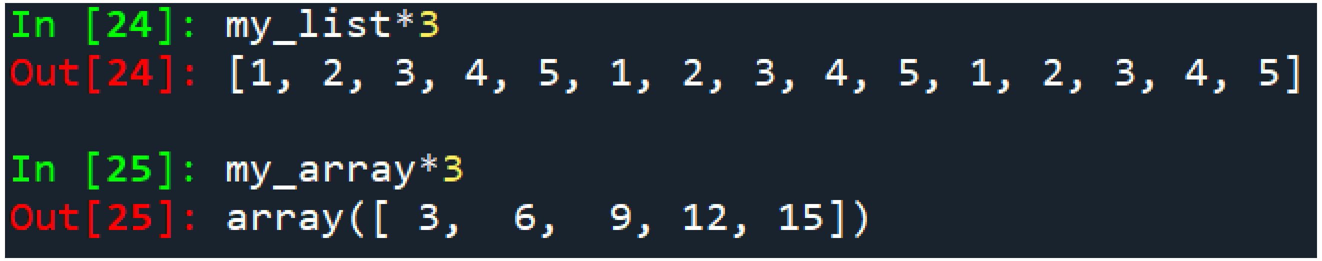


```python
my_list  = [1, 2, 3, 4, 5]
my_array = np.array([1, 2, 3, 4, 5])
```

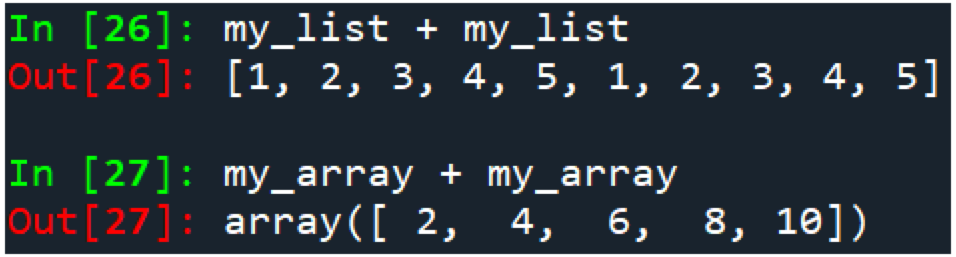
Numpy



```
my_list  = [1, 2, 3, 4, 5]
my_array = np.array([1, 2, 3, 4, 5])
```

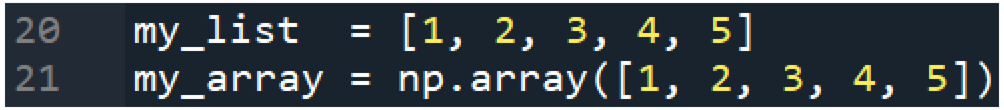


```
In [24]: my_list*3
Out[24]: [1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3, 4, 5]

In [25]: my_array*3
Out[25]: array([ 3,  6,  9, 12, 15])
```

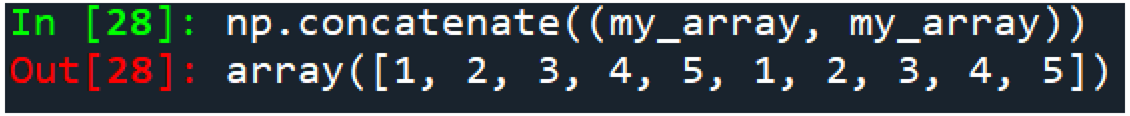


```
In [26]: my_list + my_list
Out[26]: [1, 2, 3, 4, 5, 1, 2, 3, 4, 5]

In [27]: my_array + my_array
Out[27]: array([ 2,  4,  6,  8, 10])
```

Numpy

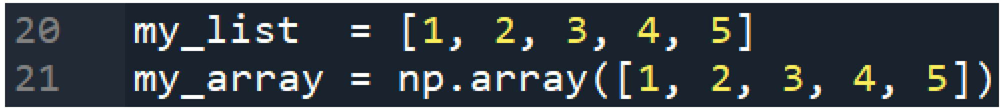


```
my_list  = [1, 2, 3, 4, 5]
my_array = np.array([1, 2, 3, 4, 5])
```

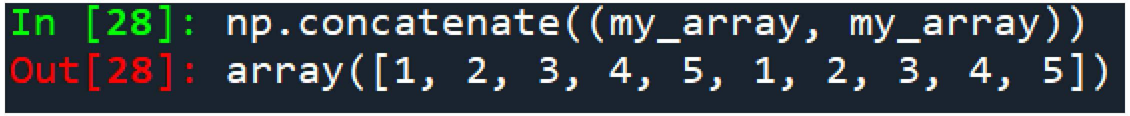


```
In [28]: np.concatenate((my_array, my_array))
Out[28]: array([1, 2, 3, 4, 5, 1, 2, 3, 4, 5])
```

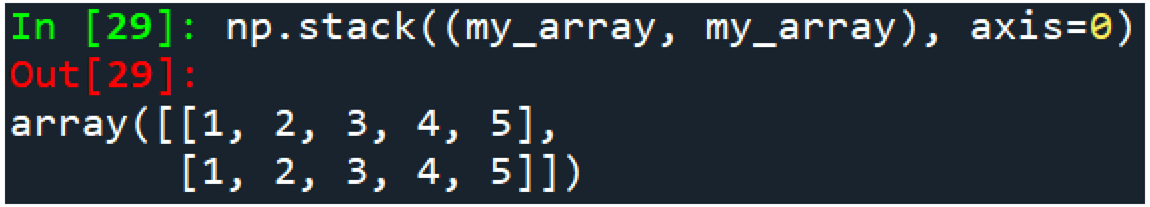
Numpy



```
my_list  = [1, 2, 3, 4, 5]
my_array = np.array([1, 2, 3, 4, 5])
```

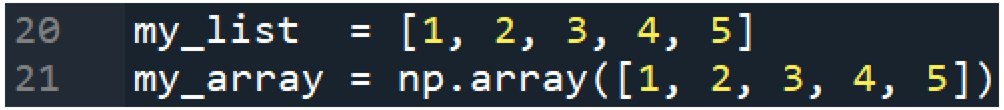


```
In [28]: np.concatenate((my_array, my_array))
Out[28]: array([1, 2, 3, 4, 5, 1, 2, 3, 4, 5])
```

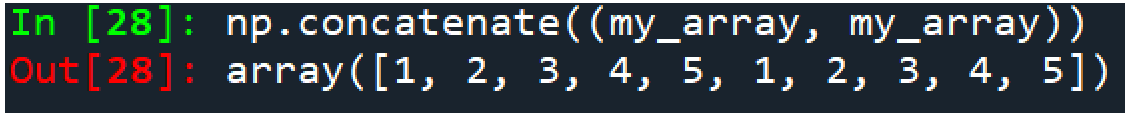


```
In [29]: np.stack((my_array, my_array), axis=0)
Out[29]:
array([[1, 2, 3, 4, 5],
       [1, 2, 3, 4, 5]])
```

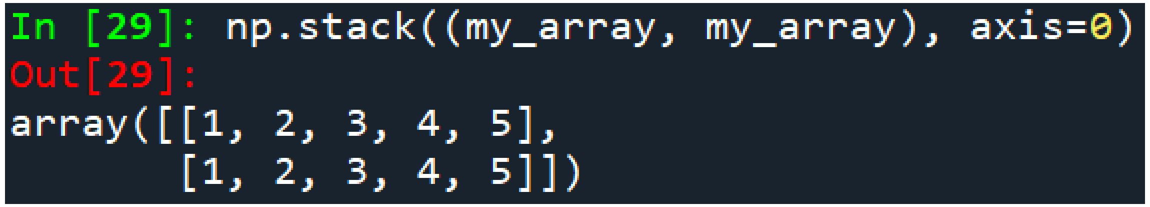
Numpy



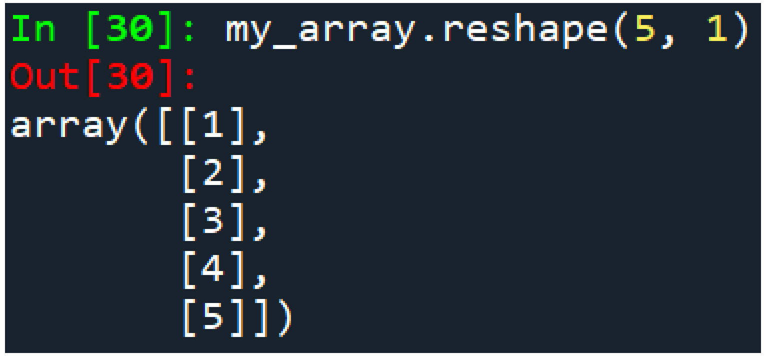
```
my_list  = [1, 2, 3, 4, 5]
my_array = np.array([1, 2, 3, 4, 5])
```



```
In [28]: np.concatenate((my_array, my_array))
Out[28]: array([1, 2, 3, 4, 5, 1, 2, 3, 4, 5])
```

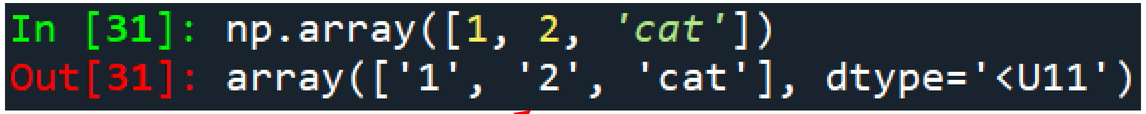


```
In [29]: np.stack((my_array, my_array), axis=0)
Out[29]:
array([[1, 2, 3, 4, 5],
       [1, 2, 3, 4, 5]])
```



```
In [30]: my_array.reshape(5, 1)
Out[30]:
array([[1],
       [2],
       [3],
       [4],
       [5]])
```

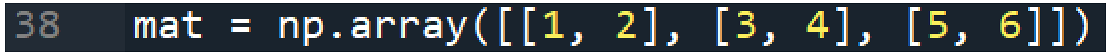
Numpy



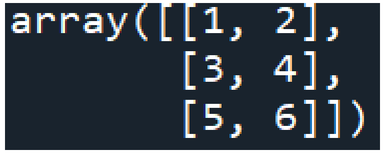
```
In [31]: np.array([1, 2, 'cat'])
Out[31]: array(['1', '2', 'cat'], dtype='<U11')
```

Became strings

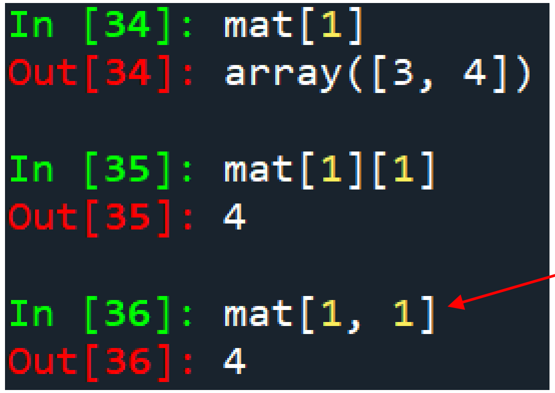
Numpy



```
mat = np.array([[1, 2], [3, 4], [5, 6]])
```



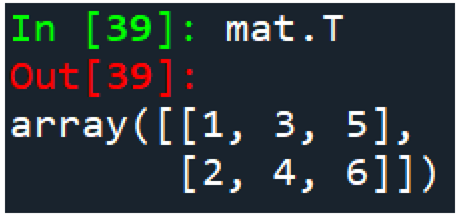
```
array([[1, 2],
       [3, 4],
       [5, 6]])
```



```
In [34]: mat[1]
Out[34]: array([3, 4])

In [35]: mat[1][1]
Out[35]: 4

In [36]: mat[1, 1]
Out[36]: 4
```

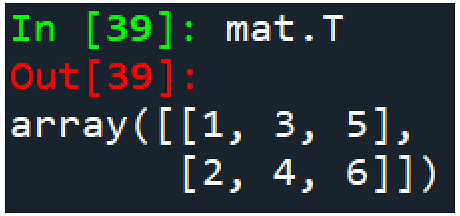
New to Numpy arrays

Numpy: matrix algebra

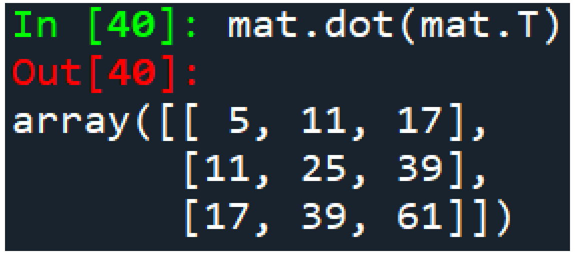


```
In [39]: mat.T
Out[39]:
array([[1, 3, 5],
       [2, 4, 6]])
```

Numpy: matrix algebra

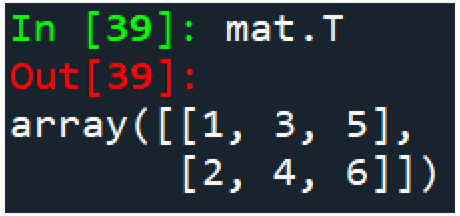


```
In [39]: mat.T
Out[39]:
array([[1, 3, 5],
       [2, 4, 6]])
```

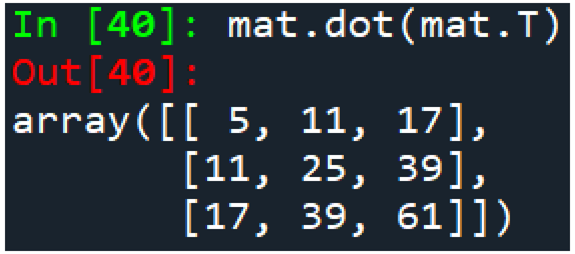


```
In [40]: mat.dot(mat.T)
Out[40]:
array([[ 5, 11, 17],
       [11, 25, 39],
       [17, 39, 61]])
```

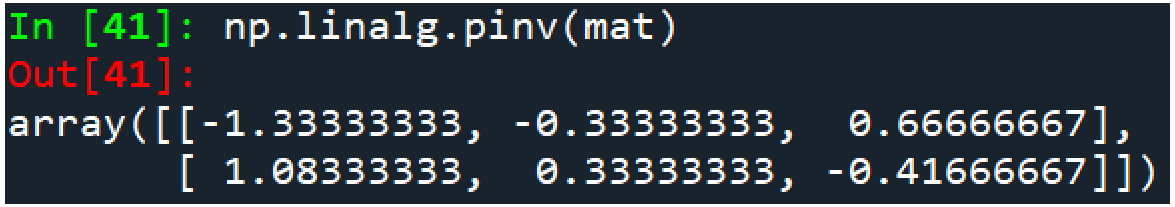
Numpy: matrix algebra



```
In [39]: mat.T
Out[39]:
array([[1, 3, 5],
       [2, 4, 6]])
```

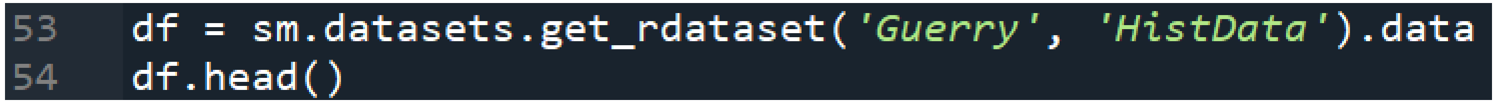


```
In [40]: mat.dot(mat.T)
Out[40]:
array([[ 5, 11, 17],
       [11, 25, 39],
       [17, 39, 61]])
```

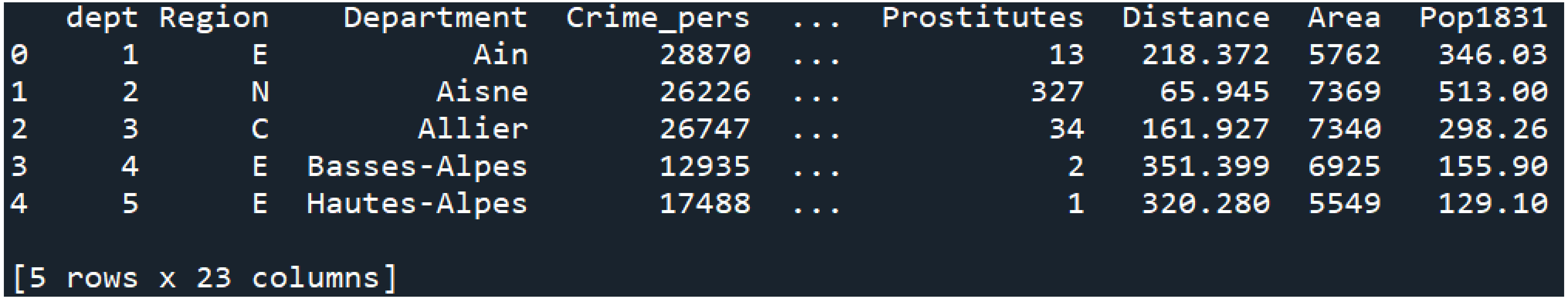


```
In [41]: np.linalg.pinv(mat)
Out[41]:
array([[-1.33333333, -0.33333333,  0.66666667],
       [ 1.08333333,  0.33333333, -0.41666667]])
```

Statsmodels

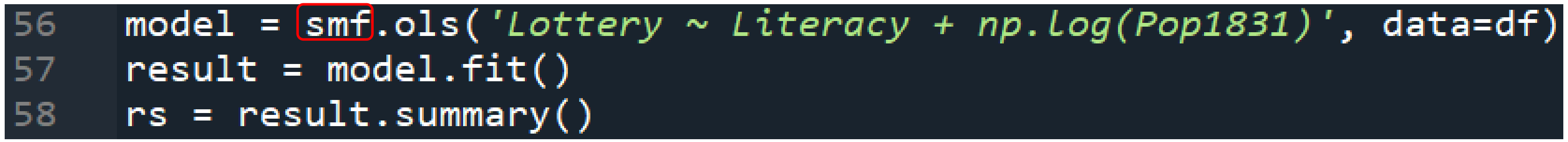


```
df = sm.datasets.get_rdataset('Guerry', 'HistData').data
df.head()
```

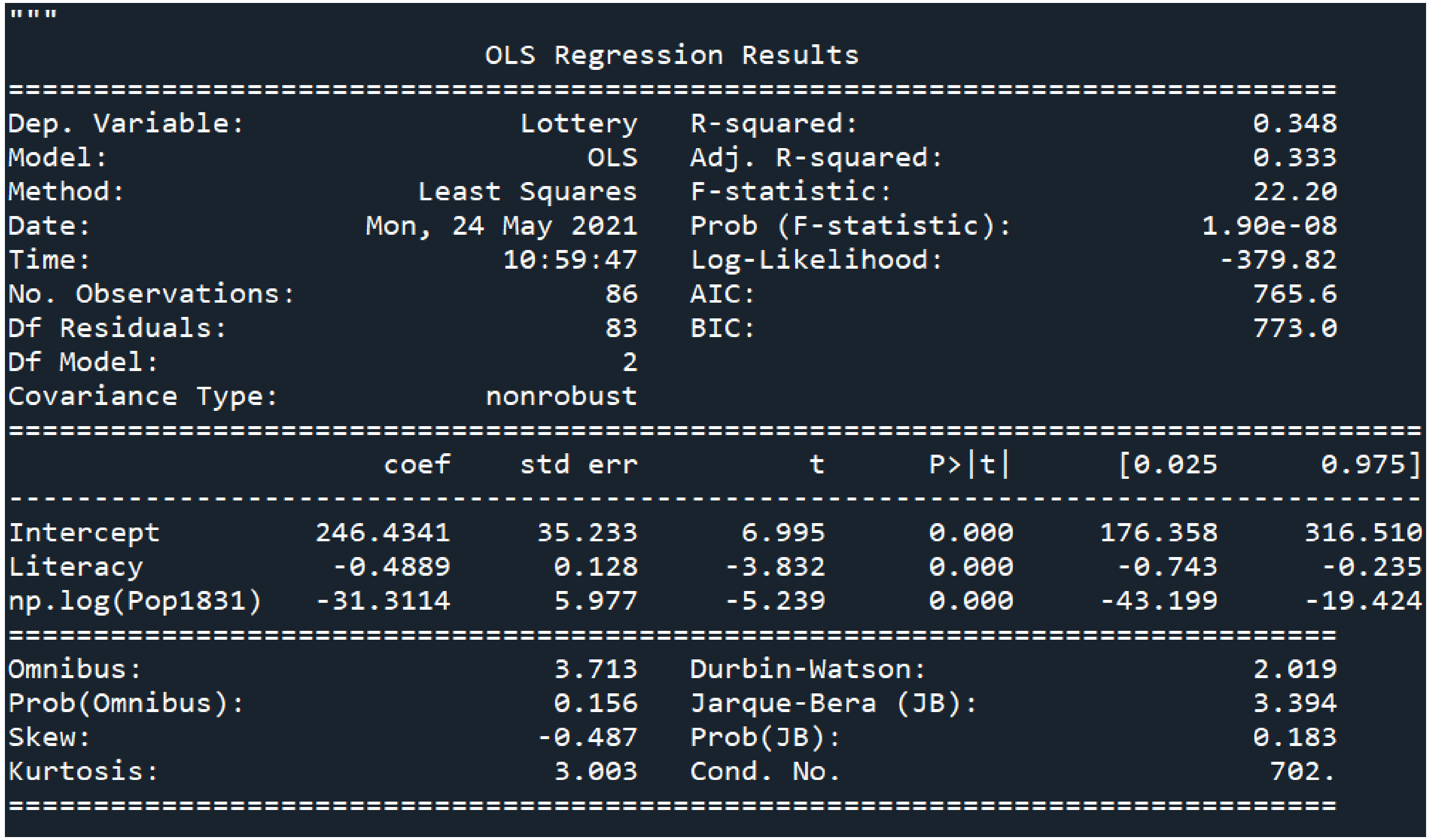


```
   dept Region    Department  Crime_pers  ...  Prostitutes  Distance  Area  Pop1831
0     1      E           Ain       28870  ...           13   218.372  5762   346.03
1     2      N         Aisne       26226  ...          327    65.945  7369   513.00
2     3      C        Allier       26747  ...           34   161.927  7340   298.26
3     4      E  Basses-Alpes       12935  ...            2   351.399  6925   155.90
4     5      E  Hautes-Alpes       17488  ...            1   320.280  5549   129.10

[5 rows x 23 columns]
```

Statsmodels: R-style formulas

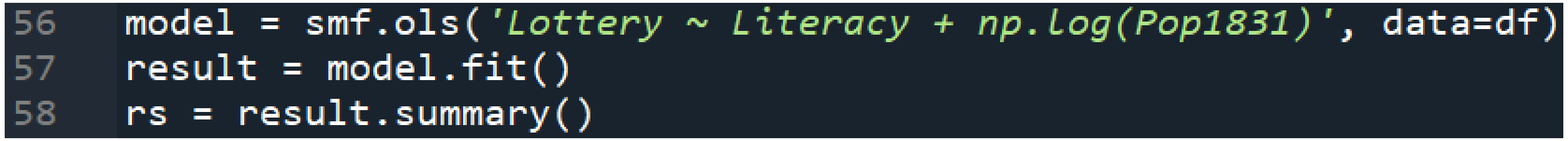


```python
model = smf.ols('Lottery ~ Literacy + np.log(Pop1831)', data=df)
result = model.fit()
rs = result.summary()
```

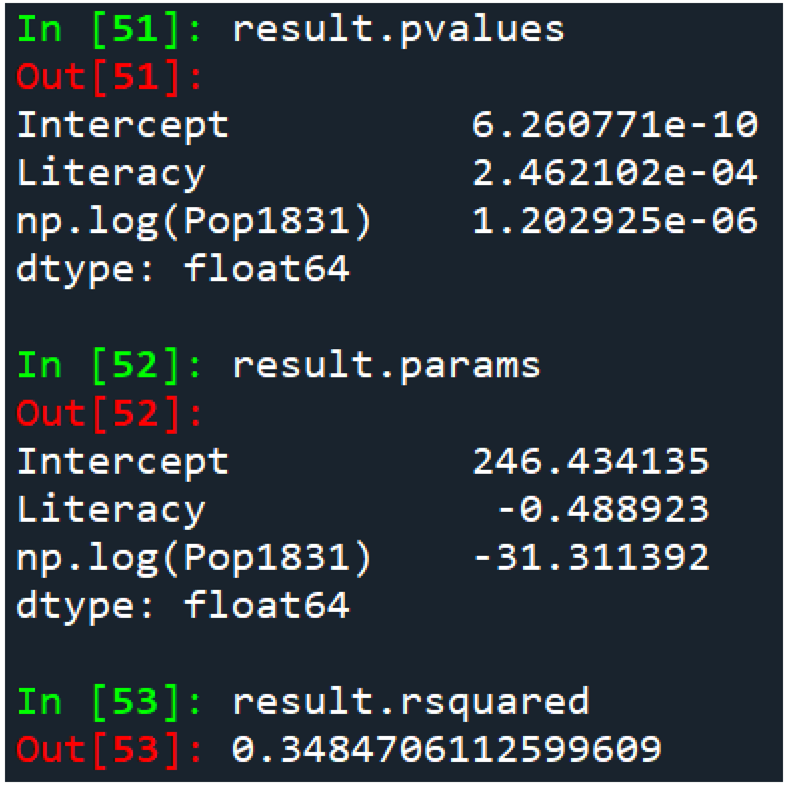


```
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Mon, 24 May 2021   Prob (F-statistic):           1.90e-08
Time:                        10:59:47   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept         246.4341     35.233      6.995      0.000     176.358     316.510
Literacy           -0.4889      0.128     -3.832      0.000      -0.743      -0.235
np.log(Pop1831)   -31.3114      5.977     -5.239      0.000     -43.199     -19.424
==============================================================================
Omnibus:                        3.713   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.156   Jarque-Bera (JB):                3.394
Skew:                          -0.487   Prob(JB):                        0.183
Kurtosis:                       3.003   Cond. No.                         702.
==============================================================================
```

Statsmodels: R-style formulas

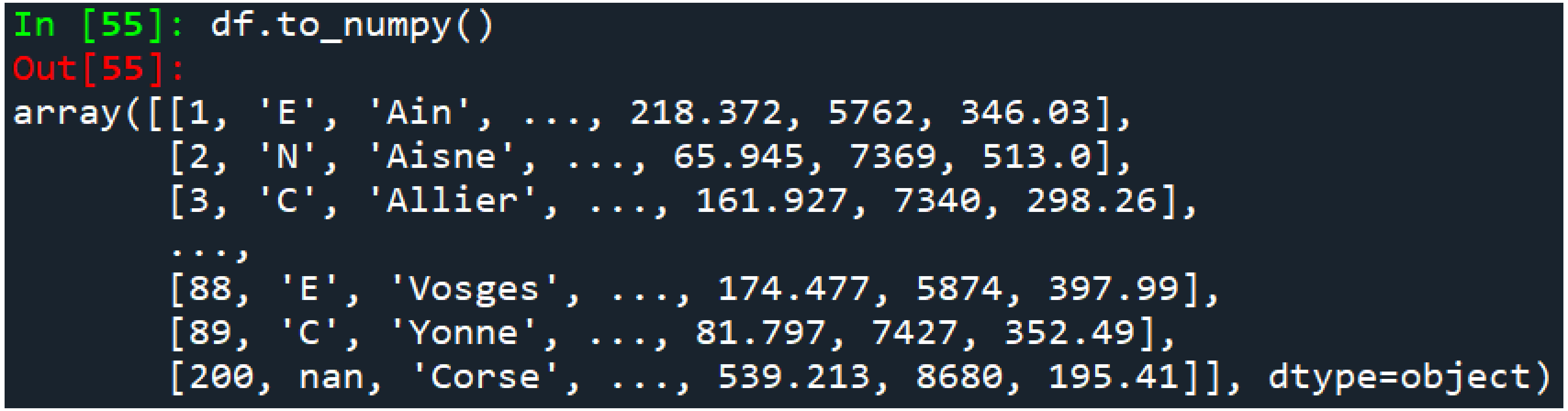


```
model = smf.ols('Lottery ~ Literacy + np.log(Pop1831)', data=df)
result = model.fit()
rs = result.summary()
```



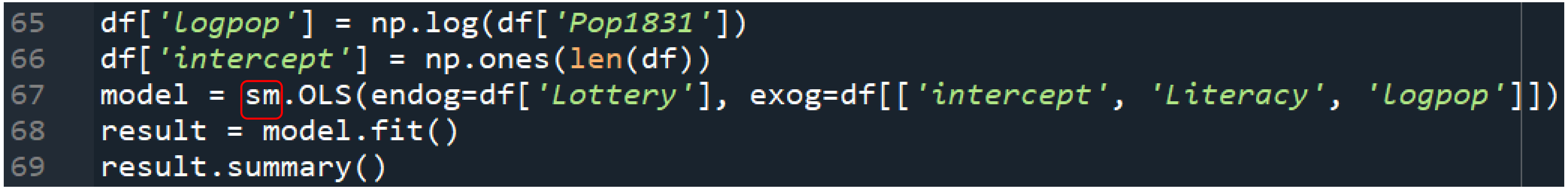
```
In [51]: result.pvalues
Out[51]:
Intercept          6.260771e-10
Literacy           2.462102e-04
np.log(Pop1831)    1.202925e-06
dtype: float64

In [52]: result.params
Out[52]:
Intercept          246.434135
Literacy            -0.488923
np.log(Pop1831)    -31.311392
dtype: float64

In [53]: result.rsquared
Out[53]: 0.348470612599609
```

Statsmodels: matrices

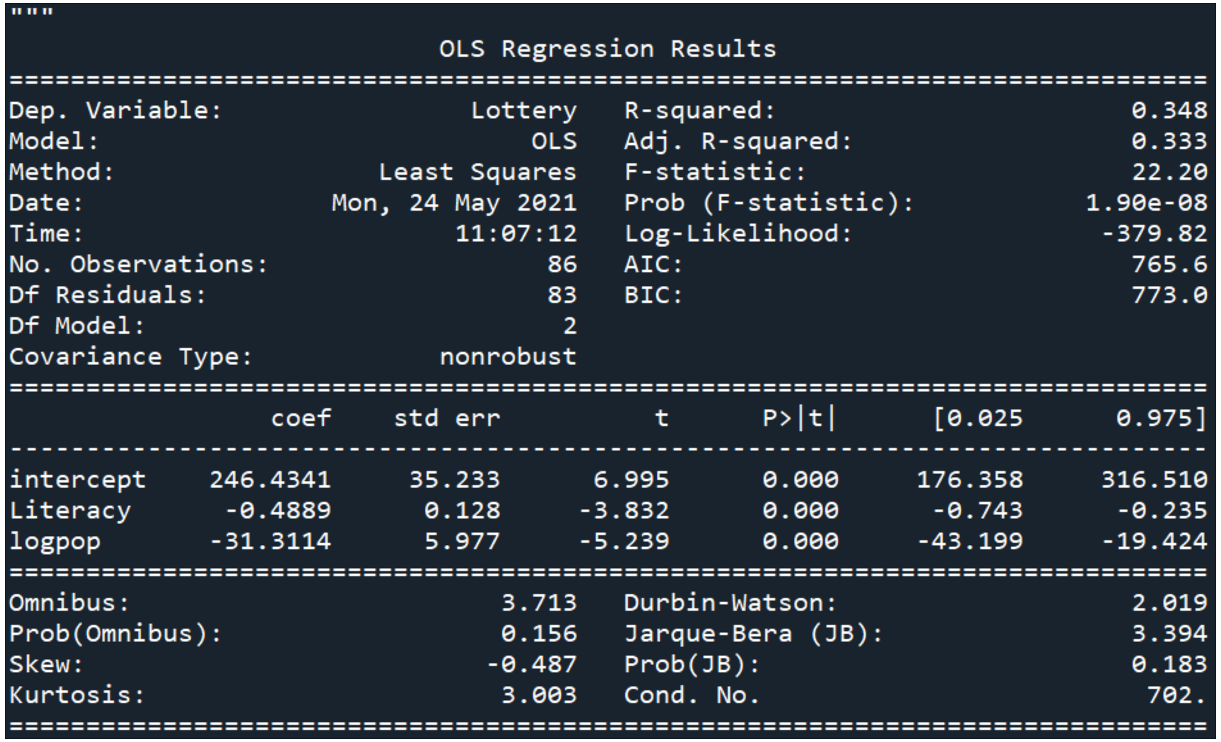


```
In [55]: df.to_numpy()
Out[55]:
array([[1, 'E', 'Ain', ..., 218.372, 5762, 346.03],
       [2, 'N', 'Aisne', ..., 65.945, 7369, 513.0],
       [3, 'C', 'Allier', ..., 161.927, 7340, 298.26],
       ...,
       [88, 'E', 'Vosges', ..., 174.477, 5874, 397.99],
       [89, 'C', 'Yonne', ..., 81.797, 7427, 352.49],
       [200, nan, 'Corse', ..., 539.213, 8680, 195.41]], dtype=object)
```

Statsmodels: matrices



```python
df['Logpop'] = np.log(df['Pop1831'])
df['intercept'] = np.ones(len(df))
model = sm.OLS(endog=df['Lottery'], exog=df[['intercept', 'Literacy', 'Logpop']])
result = model.fit()
result.summary()
```



```
"""
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Mon, 24 May 2021   Prob (F-statistic):           1.90e-08
Time:                        11:07:12   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
intercept    246.4341     35.233      6.995      0.000     176.358     316.510
Literacy      -0.4889      0.128     -3.832      0.000      -0.743      -0.235
logpop       -31.3114      5.977     -5.239      0.000     -43.199     -19.424
==============================================================================
Omnibus:                        3.713   Durbin-Watson:                   2.019
Prob(Omnibus):                  0.156   Jarque-Bera (JB):                3.394
Skew:                          -0.487   Prob(JB):                        0.183
Kurtosis:                       3.003   Cond. No.                         702.
==============================================================================
```