Pandas Cheat Sheet

- What is the purpose of a pandas library? Pandas is used to work with datasets and allows you to analyze, explore, and manipulate the data.
- What is different about a pandas dataframe in comparison to a numpy array? The major difference between the two is that pandas dataframes can store heterogeneous data types whereas numpy arrays can only store homogenous data.
- What is the index of a dataframe? The index of a dataframe is a series of labels that identify each row. The labels can be integers, strings, etc. In my code, I like to make the index a datetime group. This is different than other columns because it is a label rather than a variable.
- How to setup a pandas dataframe by reading a file: This can be done by using pd.read_table OR pd.read_csv(). This tells it to read the data file and put it into a dataframe format specified by the content of the read_table. A path to the file must be included along with other arguments you might need.

```
Ex. df = pd.read_table(filepath, sep='\t', skiprows=30,
                       names = ['agency_cd', 'site_no', 'datetime', 'flow', 'code'],
                       parse_dates = ['datetime'])
```

- How to set the index of a pandas dataframe: pandas.DataFrame.set_index(args) This allows you to set the index using existing columns.
  ex. In my streamflow code, we define a column as 'datetime' I later do:

```
dataframe_index = dataframe.set_index('datetime')
```

  This makes my index labels be the datetime group.
- How to slice a pandas dataframe:
  - .loc : accesses a group of rows or columns by labels
    ex. I like this example from our class exercises

```
data = np.ones((7,3))
data_frame=pd.DataFrame(data, columns = ['data1','data2','data3'],
                        index = ['a','b','c','d','e','f','g'])
data_frame.loc[['a','e']] = 3
```

output:

	data1	data2	data3
a	3.0	3.0	3.0
b	1.0	1.0	1.0
c	1.0	1.0	1.0
d	1.0	1.0	1.0
e	3.0	3.0	3.0
f	1.0	1.0	1.0
g	1.0	1.0	1.0

  - .iloc : integer based slicing
    ex. using the same example from above

```
dataframe.iloc[:4,:] = dataframe.iloc[:4,:]*7
```

output:

	data1	data2	data3
a	21	21	21
b	7	7	7
c	7	7	7
d	7	7	7
e	3	3	3
f	1	1	1
g	1	1	1

- slicing by column name : (also .loc)

ex. site_flow = streamflow[["site_no", "flow"]]

	site_no	flow
0		
1		
2		
⋮		

- slicing by column number : (also .iloc)

ex. site_flow = streamflow[:, 1]

	site_no
0	
1	
2	
⋮	

- Pandas functions:

1. pd.head() : returns top n (5 by default) values of the dataframe
2. pd.info() : generates the summary of the dataframe
3. pd.describe() : returns descriptive statistics about the data such as mean, min, and max
4. pd.sort_values() : sorts the dataframe is ascending or descending order of passed column
5. pd.nlargest() : used to get n largest values from a dataframe or series.