Importing Data in Pandas

Reading in Flat Files

What are flat files?

- Flat files usually refer to plain text files that contain records: that is table data
  - Rows are records / observations
  - Columns are variables / features
- Common flat files are
  - comma delimited files (csv)
  - delimited .txt files (common delimiters are tabs (`\t`) or spaces)

Reading in flat files

- It is possible to read in flat file with NumPy if all the data is numeric

- However it is much more standard to use pandas because of its ability to handle multiple data types

- Common pandas functions for reading flat file:
  ```
  .read_csv()
  .read_table()
  .read_fwf()
  ```

pd.read_csv (https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html)

- pd.read_csv('filename.csv')
- There are a ton of other arguments. Here are some common ones:
  - sep: delimiter to use (default is `','` for csv and `'\t'` for table)
  - header: row number to use as column names
  - index_col: column to use as row labels
  - skiprows: line numbers to skip at start of file
  - nrows: number of rows to read
  - na_values: additional strings to recognize as NA/NaN
  - encoding: encoding to use for UTF when reading files

Other Types of Files

Data comes in a lot of formats

(not just flat files)

Excel Files

```
# import first sheet of an excel file to a pandas DataFrame
df = pd.read_excel('filename.xlsx')

#import any sheet of an excel file to a pandas DataFrame
df = pd.read_excel('filename.xlsx', sheet_name = <index/name of sheet>)

#import All sheets into an Ordered Dictionary where key is the
#sheet name and the value is the data
df = pd.read_excel('file.name.xlsx', sheet_name = None)

#import all sheets into a pandas Excel File
xl = pd.ExcelFile('filename.xlsx')

#see sheet names in file
xl.sheet_names

#read a sheet from pandas Excel file into a DataFrame
df = xl.parse('sheet_name')
```

Data comes in a lot of formats

(not just flat files)

Python Pickle Format

```
# read a pickle formatted file
df = pd.read_pickle('filename.pkl')

#write a pickle formatted file
pd.to_pickle(object, path)
```

Data comes in a lot of formats

(not just flat files)

HTML Tables

```
# read HTML string/file/url into a list of DataFrames
#  (easiest form of webscraping)

df = pd.read_html(url)
```

Data comes in a lot of formats

(not just flat files)

- Stata
- SAS
- JSON
- SPSS
- MATLAB
- Feather Format
- and many more
    - There is probably documentation, a blog post, and/or an online question service (i.e., OverStack) that can show you how to import just about any type of file

Reading plain text files with **`open()`**

Reading lines from a Plain Text File

- Plain Text generally refers to a document or file that contains only text

- To open a plain text file, use Python's "open" function to open a connection to the file

"r" is to read only
(and is the default)

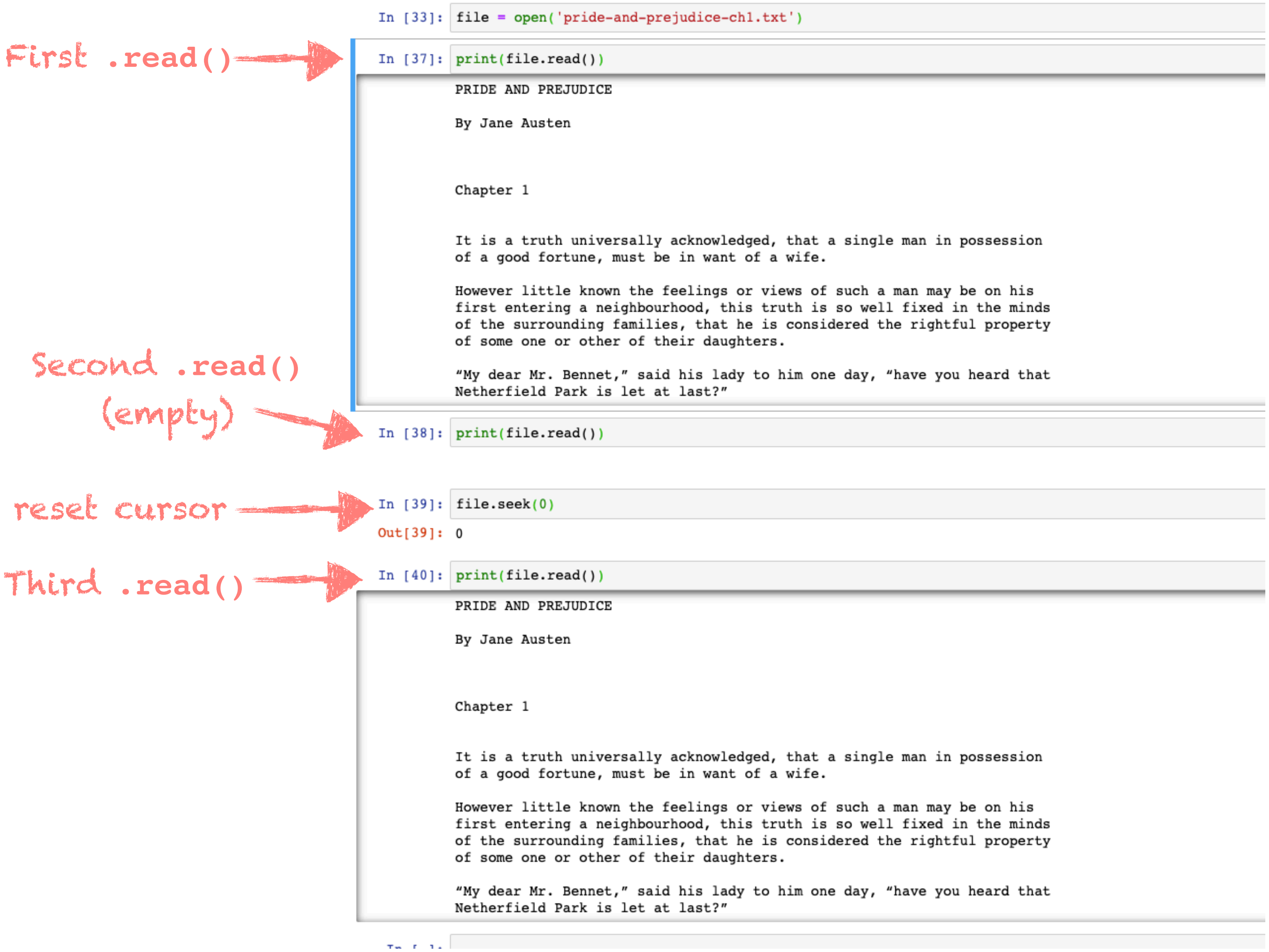
```
file = open('filename.txt', mode = 'r')
text = file.read()
file.close()
```

(you can write to a file by using "mode = 'w'")

it is good practice to always close the connection so you don't forget later and get errors somewhere

Reading lines from a Plain Text File

- Note that the command file.read() actually goes through and "reads" the file from start to finish
  - The cursor starts and the beginning of the file and finishes at the end the end of the file so if you call file.read() again, empty text will be returned because there is nothing more to read.
- To reset the cursor back to the beginning using the command file.seek(0)
- See example on the next slide



```
In [33]: file = open('pride-and-prejudice-ch1.txt')
```

```
In [37]: print(file.read())
```

```
         PRIDE AND PREJUDICE

         By Jane Austen



         Chapter 1


         It is a truth universally acknowledged, that a single man in possession
         of a good fortune, must be in want of a wife.

         However little known the feelings or views of such a man may be on his
         first entering a neighbourhood, this truth is so well fixed in the minds
         of the surrounding families, that he is considered the rightful property
         of some one or other of their daughters.

         “My dear Mr. Bennet,” said his lady to him one day, “have you heard that
         Netherfield Park is let at last?”
```

```
In [38]: print(file.read())
```

```
In [39]: file.seek(0)
Out[39]: 0
```

```
In [40]: print(file.read())
```

```
         PRIDE AND PREJUDICE

         By Jane Austen



         Chapter 1


         It is a truth universally acknowledged, that a single man in possession
         of a good fortune, must be in want of a wife.

         However little known the feelings or views of such a man may be on his
         first entering a neighbourhood, this truth is so well fixed in the minds
         of the surrounding families, that he is considered the rightful property
         of some one or other of their daughters.

         “My dear Mr. Bennet,” said his lady to him one day, “have you heard that
         Netherfield Park is let at last?”
```

```
In [ ]:
```

Making a list of lines

- Along with `.read()`, another common method is `.readlines()` which will read each line in as a separate item in a list

- A new line is indicated with `"\n"`

```
file = open('filename.txt', mode = 'r')
text_list = file.readlines()
file.close()
```

Using a context manager

- We said earlier that once a file is opened, it should always be closed

- In order to avoid the mistake of forgetting to close the file, we can open text files with a context manager using `with`

```
with open('filename.txt', mode = 'r') as myfile:
    text_list = myfile.readlines()
```

- The file won't be open outside the context manager