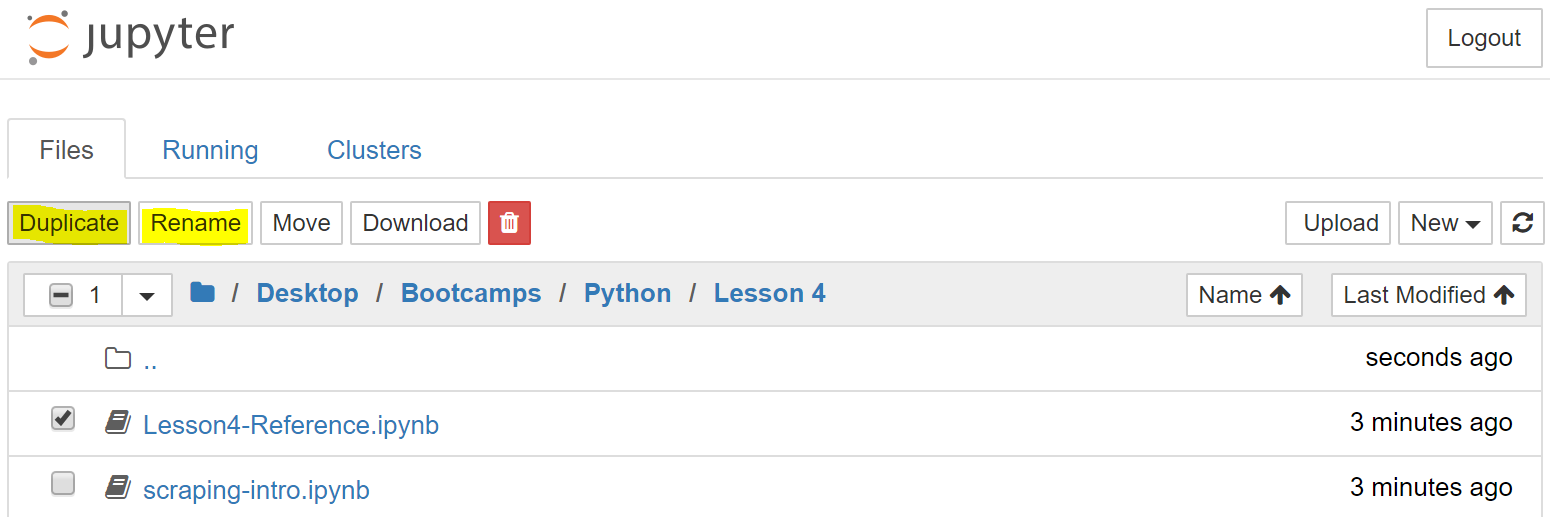
**Lesson 4 – Data-sets & Data-frames**

**Background:** For this boot camp, we will be using a Jupyter Notebook to scrape data from data-sets into data-frames using the ‘pandas’ library. Data scraping is about obtaining data from webpages or other data-sets. There is low level scraping where you parse the data out of the HTML code of a webpage. You can also scrape data from the APIs of certain websites and pre-compiled data.

**Reminders:**

* Make use of short-cuts, like tab completion
* Make use of Google; Resources can help you turn **snippets** into fully functional code
* Make variable names meaningful!!!
  + *(a = b \* c) can be harder to interpret when contextual errors arise*
  + *Using contextually relevant variable names can make it easier for others to jump in and contribute to, or analyze your code when you get stuck.*
  + ***Comment your code for clear logic!***

**Open a Notebook**

1. Launch ‘Jupyter Notebook’ using the shortcut you created in Lesson 1
2. Navigate to your Lesson 4 folder
3. Select & Duplicate the ‘Lesson4-Reference.ipynb’ file
   * *You will be tinkering – things might break – it’s a backup*
4. Open the new duplicate file
5. Rename it: ‘scraping-intro’

**1: Import the Required Libraries**

* Run the import snippet to ensure there are no errors
* Make note of the libraries being imported, and the namespaces used
* References for each library can be found in the **Code-References.gdoc**
  + *These documents are in development (Google is your friend…)*

**2: Read the User Data**

* Read over the code to ensure you understand the logic involved
* Pandas has a function to read csv files and turn them into tables
  + *What happens if you change the column name?*
  + *What happens if you change the separator?*
  + **How do you return more than 5 users data at a time?**

**Note:** *Try your best to answer any questions in* ***bold****, other questions are for speculation and should help guide you to understand each portion of the lesson.*

**3: Read the Movie Ratings**

* Read over the code to ensure you understand the logic involved
  + *Why is the separator ‘\t’ as opposed to ‘|’ like the previous data?*
  + *How can you find this information?*
    - *This situation is unique as we have been provided a pre-compiled data-set with a* ***resource*** *for how the data is assembled*
    - [*http://files.grouplens.org/datasets/movielens/ml-100k*](http://files.grouplens.org/datasets/movielens/ml-100k)
    - *Open the ‘README’ file and navigate to the ‘u.user’ section*
    - *Notice the classification in the 4th column*

**4: Read the Movie Information**

* Read over the code to ensure you understand the logic involved
  + *Why is their encoding for this data, but not the others?*
    - *What happens when you remove it?*
  + *What happens when the column range does not match the number of column names provided?*

**5: Get Information About the Movie Data**

* Read over the code to ensure you understand the logic involved
  + *The functions being used are from the ‘pandas’ library*
  + *Just by looking at the output, what kind of information do these functions return?*
  + *Why are only two columns returned for the ‘.describe( )’function?*
    - *Can you think of ways it might be used to provide more advanced functionality or statistical analysis to a program?*

**6: Selecting Data**

* Read over the code to ensure you understand the logic involved
  + *A DataFrame is a group of Series’ with a shared index*
  + *A single DataFrame column holds a single Series*
* Why is the output no longer formatted like section 4 and prior?
* What does the ‘.iloc[ ]’ function do?

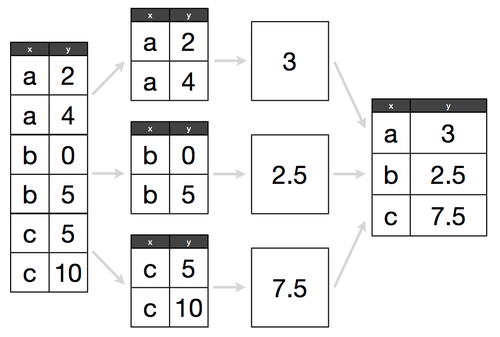
**7: Filtering Data**

* Read over the code to ensure you understand the logic involved
  + Selecting users older than 25 from the user data
    - Creating a new sub-DataFrame with that data
  + Notice how the column name can be referenced via the data-set name
    - This is only possible after it has been processed into a table via pandas

**8: Overview Quiz 1**

* show users aged 40 and male
* get/show the mean age of female programmers

**9: Split-Apply-Combine**

* splitting the data into groups based on some specified criteria
* applying a function to each group of criteria independently of each other
* combining the results into a data structure

**Find Diligent Users**

* split data per user ID
* count ratings
* combine the result

**10: Overview Quiz 2**

* get the average rating per movie
* get the movie titles with the highest average rating
* get the number of ratings per movie

**11: Passing a Function**

* Read over the code to ensure you understand the logic involved

**12: Overview Quiz 3**

* get the average rating per user
* list all occupations and if they are male dominant
* get the total number of male and female users

**13: Pandas Wrap-up**

* Have some fun with this section, test your understanding of the content by putting it to practice!
* **Do your best to implement your own versions of the following using the different data-sets provided above:**
  + Create data frames
  + Get sub-frames
  + Filter data
  + Use group-by
  + Apply a user defined function